Федеральное государственное автономное образовательное учреждение высшего образования
«Московский физико-технический институт (национальный исследовательский университет)»
(МФТИ, Физтех)

На правах рукописи

Новицкий Василий Геннадьевич

Новые оценки для стохастических безградиентных методов с одноточечным оракулом

Специальность 1.2.3.
Теоретическая информатика, кибернетика

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук, профессор
Гасников Александр Владимирович

Москва — 2024

Moscow Institute of Physics and Technology
(MIPT, Phystech)

Manuscript

Novitskii Vasilii Gennadievich

New Bounds for One-point Stochastic Gradient-free Methods

Specialty 1.2.3.
Theoretical Informatics, Cybernetics

Dissertation for the degree of
Candidate of Physical-Mathematical Sciences

Scientific supervisor:
Doctor of Physical-Mathematical Sciences, Professor
Gasnikov Alexander Vladimirovich

Moscow — 2024

Contents

Introduction

We focus on the problem of zero-order stochastic minimization:

$$f(x) \to \min_{x \in Q}, \tag{1}$$

in which the aim is to minimize an unknown convex or strongly convex function f where no gradient realization is given but a function value is available (zero-order oracle or so called black-box oracle) at each iteration with some additive noise ξ. The noise can arise from rounding error or from the case when the function value cannot be given exactly but its estimate with the noise ξ can be given, moreover, the noise ξ can be adversarial. These problems have received significant attention in the literature (see [1—12]) and are fundamental for many applications where the derivative of function is not available or it is hard to calculate derivatives. In this case we can approximate the gradient g as follows:

$$\frac{n}{2\tau}(f(x+\tau e) + \dot{\xi} - f(x-\tau e) - \ddot{\xi})e, \tag{2}$$

where e is uniformly distributed on the unit Euclidean sphere, $\dot{\xi}$ is the noise at the point $x+\tau e$ and $\ddot{\xi}$ is the noise at the point $x-\tau e$. This estimation of the gradient allows us to build gradient-like zeroth-order algorithms for the minimization problem (1).

In the estimation of gradient (2) the two possibilities are usually considered. The first one is to obtain a function value with some noise in one point $\frac{n}{\tau}(f(x+\tau e)+\xi)e$ [5; 13; 14]. The second one is to observe function values in two points with the noise at each iteration exactly as in (2). The use of three and more points does not make dramatic difference to the results for two points [15].

With two-point feedback usually two opportunities are considered: where the two noisy evaluations are obtained with the same noise and the noisy functions are Lipschitz or 2-smooth [15; 16] or the noises $\dot{\xi}$ and $\ddot{\xi}$ are independent zero-mean random variables [2; 8; 17—23].

Note that despite our algorithm gets two function values for iteration , they are obtained with different noise $\dot{\xi}$ and $\ddot{\xi}$, our approach allows adversarial noise (no independence or zero-mean assumption, no Lipschitz assumption), so it is correct to regard estimation algorithms using the estimation (2) of the gradient one-point and to compare it with one-point algorithms.

We study minimization problem (1) for non-smooth case and for higher order smooth case. In non-smooth case we assume that f is Lipschitz continious. For non-smooth objective f we use randomized smoothing function

$$f_\tau(x) = \mathbb{E}_u f(x+\tau u),$$

where u is random vector uniformly distributed on the Euclidean ball. This approach goes back to 1970s [13; 24] and helps to use first-order-like methods in non-smooth case.

In higher order smooth case we consider functions f satisfying the generalized Hölder condition with parameter $\beta > 2$ (see inequality (1.2)). Informally, it means that our function is smoother than the Lipschitz-continuous function (functions with Lipschitz gradient). Lipschitz-continuous function satisfies the generalized Hölder condition with parameter $\beta = 2$.

To exploit higher order smoothness we use Kernel smoothing proposed by Katkovnik in [25] and one-point algorithm first proposed by Granichin in [8] and later independently by Polyak and Tsybakov in [18].

We also address the saddle-point or min-max problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \varphi(x,y). \tag{3}$$

This problem is more complex than the minimization one and has many applications. These are the classic matrix game and Nash equilibrium, robust optimization [26; 27], signal processing [28], as well as modern machine learning problems: Generative Adversarial Networks (GANs) [29], Reinforcement Learning (RL) [30].

For the saddle-point problem one can use gradient-like methods with the following direction of the descent: $(-\nabla_x \varphi(x,y), \nabla_y \varphi(x,y))$. We develop zeroth-order methods exploiting higher-order smoothness for saddle-point problems.

The dissertation topic is highly relevant as it addresses a crucial areas of optimization. The first area is a usage of high-order-smoothness methods in minimization problems and saddle-point problems. These methods have the potential to solve problems faster than usual methods exploiting functional or gradient smoothness.

The second area is a problem of non-smooth optimization. It is highly relevant due to the fact that a lot of modern machine learning problems are non-smooth, and for some of them analysis, based on first-order methods, is not applicable. However, using smoothing technique, we can apply first-order methods for non-smooth problems. Currently, there are numerous applications that require efficient optimization methods, especially in machine learning, signal processing, image processing, and other fields where data exhibits complex structures.

The aim of the work can be summarized as follows:

- To develop zeroth-order **minimization** algorithm exploiting **higher order smoothness** of the target function with one-point oracle under quasi-adversarial noise assumption. To estimate theoretically its convergence in convex and strongly-convex case.
- To develop zeroth-order **saddle-point** algorithm exploiting **higher order smoothness** of the target function with one-point oracle under quasi-adversarial noise assumption. To estimate theoretically its convergence in convex-concave and strongly-convex-strongly-concave case.
- To develop zeroth-order minimization algorithm for **non-smooth** convex and strongly-convex case based on smoothing technique under adversarial noise assumption. To estimate theoretically its convergence in convex and strongly-convex case. To obtain the upper bounds on the maximum level of adversarial noise that does not affect the convergence of the minimization algorithm.

Scientific novelty:

- We propose the algorithm for minimization problem in the case the target function is higher order smooth and provide the upper bounds for its optimization error in convex and strongly-convex case. We focus on one-point oracle case that is more challenging and less studied compared to two-point case with quasi-adversarial noise.

– We also propose the algorithm for saddle-point problem in the case the target function is higher order smooth and provide the upper bounds for its optimization error in convex-concave and strongly-convex-strongly-concave case with quasi-adversarial noise. We also focus on one-point oracle case. The exploiting of higher-order smoothness is unusual for the saddle-point problems.
– We propose the new algorithm for non-smooth minimization problem under adversarial noise and upper bounds for it and generic approach (smoothing scheme) that, based on optimal first-order methods, allows to obtain zeroth-order algorithms for non-smooth convex optimization problems. The proposed generic approach allows us to construct an algorithm with the best iteration complexity among the algorithms that have optimal oracle complexity.

  The special focus is made to the calculation of the maximum level of adversarial noise that does not spoil the convergence rate. The bounds for complexity of presented algorithm and for the maximum level of adversarial noise are new and best known, moreover both bounds coincide with the lower bounds in certain regimes.

  Presented algorithm also allows to obtain iteration complexity similar to first-order methods, which allows to exploit parallel computations to accelerate the convergence of presented algorithms.

All the results of this work are new and extend the community's knowledge in the field of research.

Theoretical and practical value: the new upper bounds of optimization error of the proposed algorithms for minimization problem and the saddle-point problem when the target function is higher-order smooth and when the noise is quasi-adversarial. We deal with one-point oracle that is practically realistic and more fair compared to two-point oracle. We provide the bounds for convex and strongly-convex case for minimization problem and the bounds for convex-concave and strongly-convex-strongly-concave case for saddle-point problem. All the bounds are new.

We also propose a smoothing scheme for non-smooth minimization problems that can be used to obtain zeroth-order algorithms based on optimal first-order algorithms. We provide theoretical bounds under adversarial noise assumption for minimization problem in convex and strongly-convex case for the algorithm based on smoothing scheme applied to batched Accelerated gradient method. Also we estimate the admissible limitations for the adversarial noise in such way that the noise do not worsen the convergence of the algorithm in both convex and strongly-convex cases.

From practical point of view, we conducted a numerical experiment for the minimization algorithm using higher order smoothness of the target function to show its advantages compared to the similar algorithm allowing only Lipschitz continuity of the gradient.

Proposition for the defence:

– New algorithm and new upper bounds for solving convex and strongly convex derivative-free minimization problems using one-point feedback in higher order smoothness case under quasi-adversarial noise.

– New algorithm and new upper bounds for solving convex-concave and and strongly-convex-strongly-concave derivative-free saddle-point problems using one-point feedback in higher order smoothness case under quasi-adversarial noise.
– New algorithm and new upper bounds for non-smooth convex and strongly-convex derivative-free minimization problems using one-point feedback under adversarial noise assumption. New upper bounds for the maximum level of adversarial noise.

Reliability of the work of the research results is ensured by:
– Strictness and correctness of mathematical proofs and reasoning;
– Numerical experiments;
– Coincidence of theoretical results with experimental data;
– Qualified validation at international and Russian scientific conferences and seminars. The reliability is further substantiated by the publication of research results in peer-reviewed scientific publications.

Probation of the work was conducted at the following scientific conferences and seminars:
– ICML Workshop «The power of first-order smooth optimization for black-box non-smooth problem», 20 July 2022, Baltimore, USA;
– The International Conference «Mathematical Optimization Theory and Operations Research», 5-10 July 2021, Irkutsk;
– The International Conference «Quasilinear Equations, Inverse Problems and Their Applications», 23-29 August 2021, Sochi;
– 63-rd Russian Scientific Conference of MIPT, 23-29 November 2020, Dolgoprudny;

Personal contribution.
– Chapter 1. This part is made by myself.
– Chapter 2. This part is made by myself.
– Chapter 3. I worked on the proofs of the theorems under adversarial noise and on the calculations of maximum level of adversarial noise.

Structure of the work. The dissertation consists of introduction, 3 chapters, conclusion and 4 appendices. The total size of the dissertation is 85 pages, including 2 figures and 5 tables. List of references consists of 50 references.

Chapter 1. Exploiting Higher Order Smoothness in Derivative-free Minimization Problems

1.1 Introduction

We study the problem of zero-order stochastic optimization in which the aim is to minimize an unknown convex or strongly convex function where no gradient realization is given but a function value is available at each iteration with some additive noise ξ. We also study a closely related problem of continuous stochastic bandits. These problems have received significant attention in the literature (see [1—7; 9—12]) and are fundamental for many application where the derivative of function is not available or it is hard to calculate derivatives.

The goal is to exploit higher order smoothness of the function to improve the performance of projected gradient-like algorithms. Our approach is outlined in Algorithm 1, in which a sequential algorithm gets at each iteration two function values under some noise. At each iteration the algorithm gets function values at points $x_k + \delta_k$ and $x_k - \delta_k$, where $\delta_k = \tau_k r_k e_k$. Here r_k is uniformly distributed random variable, e_k is uniformly distributed on the Euclidean sphere, τ_k – is tunable parameter of the algorithm, the smaller τ_k is, the smaller approximation error of the gradient $\|\widetilde{g_k} - \nabla f(x_k)\|_2$ is (in Chapters 1 and 2 we use only the Euclidean norm $\|\cdot\|_2$) but the bigger variance of $\|\widetilde{g_k}\|_2$ is, so the trade-off between these terms is needed. Our approach uses kernel smoothing technique proposed by Katkovnik in [25] and one-point algorithm proposed first by Granichin in [8] and later independently by Polyak and Tsybakov in [18], this helps to exploit higher order smoothness.

Algorithm 1 Zero-order Stochastic Projected Gradient for Minimization Problem

Requires: Kernel $K : [-1, 1] \to \mathbb{R}$, step size $\alpha_k > 0$, parameters τ_k.

for $k = 1, \ldots, N$ **do**

1. Generate scalar r_k uniformly on [−1,1] and vector e_k uniformly on the Euclidean unit sphere $S_n = \{e \in \mathbb{R}^n : \|e\|_2 = 1\}$.
2. $y_k := f(x_k + \tau_k r_k e_k) + \dot{\xi}_k$, $y'_k := f(x_k - \tau_k r_k e_k) + \ddot{\xi}_k$
3. Define $\widetilde{g_k} := \frac{n}{2\tau_k}(y_k - y'_k)e_k K(r_k)$
4. Update $x_{k+1} := \Pi_Q(x_k - \alpha_k \widetilde{g_k})$

end for

Output: $\{x_k\}_{k=1}^N$.

Note that despite our algorithm gets two function values for iteration, they are obtained with different noise $\dot{\xi}_k$ and $\ddot{\xi}_k$ and no i.i.d or zero-mean of the noise is assumed so it is correct to regard Algorithm 1 as one-point and to compare it with one-point algorithms.

Here we study functions satisfying the generalized Hölder condition with parameter $\beta > 2$ (see inequality (1.2) below).

We address the question: what is the performance of Algorithm 1, namely the explicit dependency of the convergence rate on the main parameters n (dimension), N, μ (strong convexity parameter for strongly convex functions), β. To handle this task, we prove an upper bound for Algorithm 1.

1.2 Contributions

Our main contributions can be summarized as follows:

1. For strongly-convex case: under quasi-adversarial noise assumption (see Assumption 1) we establish for all $\beta > 2$ the upper bound of order $O\left(\frac{n^{2-\frac{1}{\beta}}}{\mu N^{\frac{\beta-1}{\beta}}}\right)$ for the optimization error of Algorithm 1 for strongly convex case.
2. For convex case: under quasi-adversarial noise assumption (see Assumption 1) we establish for all $\beta > 2$ that after $N(\varepsilon) = O\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$ iterations of Algorithm 1 for the regularized function $f_\mu(x) := f(x) + \frac{\varepsilon}{2R^2}\|x - x_0\|_2^2$ we achieve the optimization error less than or equal to ε.

1.3 Preliminaries

In this section we give the necessary notation, definitions and assumptions.

1.3.1 Notation

Let $\langle \cdot, \cdot \rangle$ and $\|\cdot\|_2$ be the standard inner product and Euclidean norm on $\mathbb{R}^n$ respectively. For every closed convex set $Q \subset \mathbb{R}^n$ and for every $x \in \mathbb{R}^n$, let $\Pi_Q(x)$ denote the Euclidean projection of x on Q. $\dot{\xi}_k$ means the noise at the point $x_k + \tau_k r_k e_k$ and $\ddot{\xi}_k$ means the noise at the point $x_k - \tau_k r_k e_k$ in the Algorithm 1. We denote $\xi_k = (\dot{\xi}_k, \ddot{\xi}_k)$.

1.3.2 Problem

We address the conditional minimization problem

$$f(x) \to \min_{x \in Q}, \tag{1.1}$$

where $f : U_{\varepsilon_0}(Q) \to \mathbb{R}$ – function (convex or strongly convex), $Q \subset \mathbb{R}^n$ – convex compact set (Euclidean metrics).

This assumption on the availability of the objective values f in a small neighbourhood $U_{\varepsilon_0}(Q)$ of the Q is quite common in the literature, see, e.g., [31; 32] and can be established in two ways. The first one is changing the set Q in problem (1.1) to a slightly smaller set $\widetilde{Q}$ such that $U_{\varepsilon_0}(\widetilde{Q}) \subseteq Q$, see, e.g., [33]. The second one is the extension of f to the whole space $\mathbb{R}^n$ with preserving the (μ-strong) convexity, Lipschitz continuity and higher order smoothness of the function f [34], more precisely, by changing the objective to $f_{new}(x) := f\left(\Pi_Q(x)\right) + \alpha \min_{y \in Q} \|x - y\|_2$.

The minimization problem (1.1) can be formulated as follows: find the sequence $\{x_k\}_{k=1}^N \subset Q$ minimizing the average regret:

$$\frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[f(x_k) - f(x^*)\right].$$

If the average regret is less than or equal to ε, then the optimization error of averaged estimator $\overline{x}_N = \frac{1}{N} \sum\limits_{k=1}^{N} x_k$ is also less than or equal to ε:

$$\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leqslant \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[f(x_k) - f(x^*)\right] \leqslant \varepsilon.$$

1.3.3 Noise

The function values $f(x_k + \tau_k r_k e_k)$ and $f(x_k - \tau_k r_k e_k)$ are given with additive noise $\dot{\xi}_k$ and $\ddot{\xi}_k$ respectively (see Algorithm 1). Recall that the Algorithm 1 is randomized: for all $k = 1, 2, \ldots, N$ it holds that scalar r_k is distributed uniformly on [−1,1] conditionally given the current point x_k and for all $k = 1, 2, \ldots, N$ it holds that vector e_k is distributed uniformly on the Euclidean unit sphere $S_n = \{e \in \mathbb{R}^n : \|e\|_2 = 1\}$ conditionally given the current point x_k. Moreover, the noise $\xi_k = (\dot{\xi}_k, \ddot{\xi}_k)$ can have a random nature.

Assumption 1. *For all $k = 1, 2, \ldots, N$ it holds that*

1. $\mathbb{E}_{e_k, r_k, \xi_k}[\dot{\xi}_k^2 | x_k] \leqslant \Delta^2$ *and* $\mathbb{E}_{e_k, r_k, \xi_k}[\ddot{\xi}_k^2 | x_k] \leqslant \Delta^2$ *where* $\Delta \geqslant 0$;
2. *the random variables $\dot{\xi}_k$ and $\ddot{\xi}_k$ are independent from e_k and r_k, the random variables e_k and r_k are independent.*

Remark. Note, that as $\dot{\xi}_k$ and $\ddot{\xi}_k$ are independent from e_k and r_k, then Assumption 1 implies $\mathbb{E}_{e_k,r_k,\xi_k}[\dot{\xi}_k^2|x_k] = \mathbb{E}_{\xi_k}[\dot{\xi}_k^2|x_k] \leqslant \Delta^2$ and $\mathbb{E}_{e_k,r_k,\xi_k}[\ddot{\xi}_k^2|x_k] = \mathbb{E}_{\xi_k}[\ddot{\xi}_k^2|x_k] \leqslant \Delta^2$.

We do not assume here neither zero-mean of $\dot{\xi}_k$ and $\ddot{\xi}_k$ nor i.i.d of $\{\dot{\xi}_k\}_{k=1}^N$ and $\{\ddot{\xi}_k\}_{k=1}^N$ as condition 2 from Assumption 1 allows to avoid that. Without the condition 2 the Assumption 1 can be regarded as adversarial. With the condition 2 the Assumption 1 can be regarded as quasi-adversarial. In Chapter 1 and 2 we consider quasi-adversarial setup, and in Chapter 3 we consider adversarial setup, see Assumption 13.

Let us define what higher-order smoothness is.

Definition 2. *Let l denote maximal integer number strictly less than β. Let $\mathcal{F}_\beta(L_\beta)$ denote the set of all functions $f : \mathbb{R}^n \to \mathbb{R}$ which are differentiable l times and for all x,$z \in U_{\varepsilon_0}(Q)$ satisfy Hölder condition:*

$$\left| f(z) - \sum_{0 \leqslant |m| \leqslant l} \frac{1}{m!} D^m f(x)(z-x)^m \right| \leqslant L_\beta \|z-x\|_2^\beta, \tag{1.2}$$

where $L_\beta > 0$, the sum is over multi-index $m = (m_1, \ldots, m_n) \in \mathbb{N}^n$, we use the notation $m! = m_1! \cdot \ldots \cdot m_n!$, $|m| = m_1 + \cdots + m_n$ and we defined

$$D^m f(x) z^m = \frac{\partial^{|m|} f(x)}{\partial^{m_1} x_1 \ldots \partial^{m^n} x_n} z_1^{m_1} \cdot \ldots \cdot z_n^{m_n}, \ \forall z = (z_1, \ldots, z_n) \in \mathbb{R}^n.$$

Let $\mathcal{F}_{\mu,\beta}(L_\beta)$ denote the set of μ-strongly convex functions $f \in \mathcal{F}_\beta(L_\beta)$. Recall that f is called μ-strongly convex for some $\mu > 0$ if for all $x, z \in \mathbb{R}^n$ it holds that $f(z) \geqslant f(x) + \langle \nabla f(x), z - x \rangle + \frac{\mu}{2}\|x - z\|_2^2$.

1.3.4 Kernel

For the gradient estimator $\widetilde{g_k}$ we use the kernel

$$K : [-1, 1] \to \mathbb{R},$$

satisfying

$$\mathbb{E}[K(r)] = 0, \ \mathbb{E}[rK(r)] = 1, \ \mathbb{E}[r^j K(r)] = 0, j = 2, \ldots, l, \ \mathbb{E}\left[|r|^\beta |K(r)|\right] \leqslant \infty, \tag{1.3}$$

where r is a uniformly distributed on $[-1, 1]$ random variable. This helps us to get better bounds on the gradient bias $\|\widetilde{g_k} - \nabla f(x_k)\|_2$ (see Theorem 3 for details).

A weighted sum of Legendre polynoms is an example of such kernels:

$$K_\beta(r) := \sum_{m=0}^{l(\beta)} p'_m(0) p_m(r), \tag{1.4}$$

where $l(\beta)$ is maximal integer number strictly less than β and $p_m(r) = \sqrt{2m+1}\mathcal{L}_m(r)$, $\mathcal{L}_m(u)$ is Legendre polynom. We have

$$\mathbb{E}\left[p_m p_{m'}\right] = \delta(m - m').$$

As $\{p_m(r)\}_{m=0}^{j}$ is a basis for polynoms of degree less than or equal to j we can represent $u^j := \sum\limits_{m=0}^{j} b_m p_m(r)$ for some integers $\{b_m\}_{m=0}^{j}$ (they depend on j).

Let's calculate the expectation:

$$\mathbb{E}\left[r^j K_\beta(r)\right] = \sum_{m=0}^{j} b_m p_m'(0) = (r^j)'|_{r=0} = \delta(j-1),$$

here $\delta(0) = 1$ and $\delta(x) = 1$ if $x \neq 0$. We proved that the presented $K_\beta(r)$ satisfies (1.3). We have the following kernels for different betas (see Figure 1.1):

$$\begin{aligned} K_\beta(r) &= 3r, && \beta \in [2,3], \\ K_\beta(r) &= \frac{15r}{4}(5-7r^2), && \beta \in (3,5], \\ K_\beta(r) &= \frac{105r}{64}(99r^4 - 126r^2 + 35), && \beta \in (5,7]. \end{aligned}$$

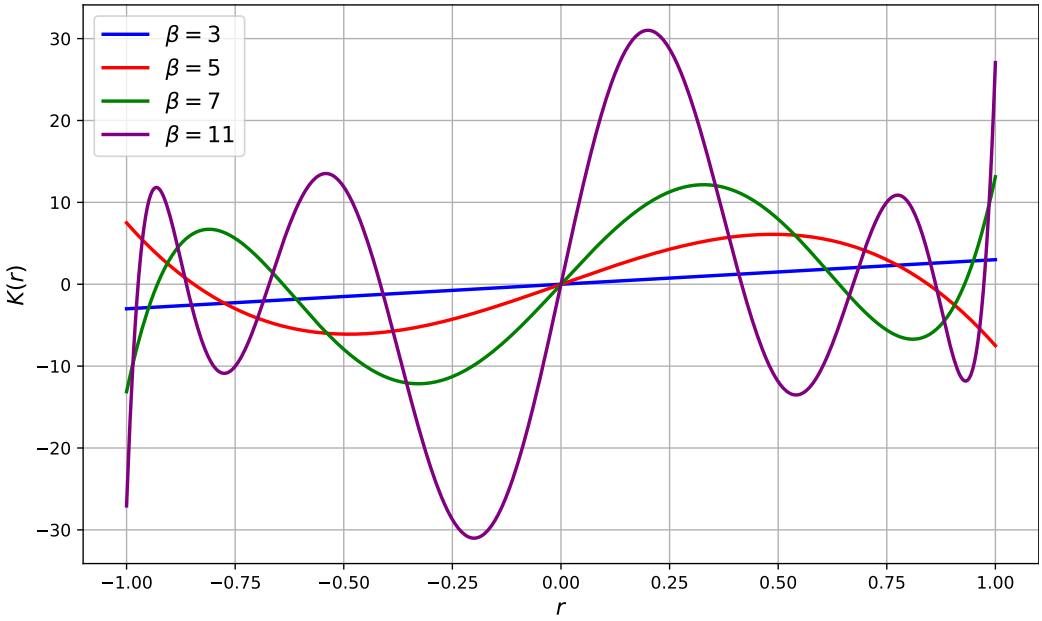

Figure 1.1 — Examples of kernels from (1.4)

For Theorem 3 and Theorem 4 we need to introduce the constants

$$\kappa_\beta = \int_{-1}^{1} |u|^\beta |K(u)|\, du \tag{1.5}$$

and

$$\kappa = \int_{-1}^{1} K^2(u)\, du. \tag{1.6}$$

It is proved in [2] that κ_β and κ do not depend on n, they depend only on β:

$$\kappa_\beta \leqslant 2\sqrt{2}(\beta - 1), \tag{1.7}$$

$$\kappa \leqslant \sqrt{3}\beta^{3/2}. \tag{1.8}$$

1.4 Main result

In this section we prove upper bounds on the optimization error of Algorithm 1 for the problem of minimization of strongly convex function (Theorem 3) and of convex function (Theorem 4).

Theorem 3. *Let $f \in \mathcal{F}_{\mu,\beta}(L_\beta)$ with μ, $L_\beta > 0$ and $\beta > 2$. Let Assumption 1 hold and let Q be a convex compact subset of $\mathbb{R}^n$. Let f be M_2-Lipschitz on the Euclidean τ_1-neighborhood of Q.*

Then the optimization error of averaged estimator $\overline{x}_N = \frac{1}{N}\sum\limits_{k=1}^{N} x_k$ where the points x_k are given by Algorithm 1 with parameters

$$\tau_k = \left(\frac{3\kappa\Delta^2 n}{2(\beta-1)(\kappa_\beta L_\beta)^2}\right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}, \quad \alpha_k = \frac{2}{\mu k}, \quad k = 1, \ldots, N$$

satisfies

$$\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leqslant \frac{1}{\mu}\left(n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N}\right),$$

where $A_1 = 3\beta(\kappa\Delta^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L_\beta)^{\frac{2}{\beta}}$, $A_2 = 3c\kappa M_2^2$, κ_β and κ are constants depending only on β, see (1.5) and (1.6), $c < 15000$.

See the proof in the Appendix B.1.

We emphasize that the usage of kernel smoothing technique, measure concentration inequalities and the assumption that $\dot{\xi}_k$ is independent from e_k or r_k (Assumption 1) lead to the results better than the state-of-the-art ones for $\beta > 2$ (see Table 2 and Table 3). The last assumption also allows us not to assume neither zero-mean of $\dot{\xi}_k$ and $\ddot{\xi}_k$ nor i.i.d of $\{\dot{\xi}_k\}_{k=1}^N$ and $\{\ddot{\xi}_k\}_{k=1}^N$.

Theorem 4. *Let $f \in \mathcal{F}_\beta(L_\beta)$ with μ, $L_\beta > 0$ and $\beta > 2$. Let Assumption 1 hold and let Q be a convex compact subset of $\mathbb{R}^n$. Let f be M_2-Lipschitz on the Euclidean τ_1-neighborhood of Q. Let $\overline{x}_N$ denote $\frac{1}{N}\sum\limits_{k=1}^{N} x_k$.*

Then we achieve the optimization error $\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leqslant \varepsilon$ after $N(\varepsilon)$ steps of Algorithm 1 with settings from Theorem 3 for the regularized function: $f_\mu(x) := f(x) + \frac{\mu}{2}\|x - x_0\|_2^2$, where $\mu \leqslant \frac{\varepsilon}{R^2}$, $R = \|x_0 - x^*\|_2$, $x_0 \in Q$ – arbitrary point.

$$N(\varepsilon) = \max\left\{\left(R\sqrt{2A_1}\right)^{\frac{2\beta}{\beta-1}} \frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \left(R\sqrt{2c'A_2}\right)^{2(1+\rho)} \frac{n^{1+\rho}}{\varepsilon^{2(1+\rho)}}\right\},$$

where $A_1 = 3\beta(\kappa\Delta^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L_\beta)^{\frac{2}{\beta}}$, $A_2 = 3c\kappa M_2^2$ – constants from Theorem 3, $\rho > 0$ – arbitrarily small positive number.

See the proof in the Appendix B.2.

1.5 Numerical experiment

In our experiment we compare the Algorithm 1 (with $\beta = 3$ and $\beta = 5$) proposed here with Gasnikov's one-point method for the minimization problem.

We consider the problem of the minimization of the following function

$$f(x) = \frac{1}{2}x^T Ax + \frac{1}{10}\sum_{k=1}^{50} |x_k|^4$$

on the Euclidean ball $Q = \{x \in \mathbb{R}^{50} : \|x\|_2 \leqslant 1\}$.

The starting point is x_0 with $\|x_0\|_2 = 1/2$. The dependency of $f(\overline{x}_N) - f(x^*)$ (optimization error) on N (iteration number) is presented on Figure 1.2. The optimization error has its mean and 0.95-confidence interval. As the constant L_β for Algorithm 1 with $\beta = 5$ is equal to zero, we choose $L_\beta = 0.001$.

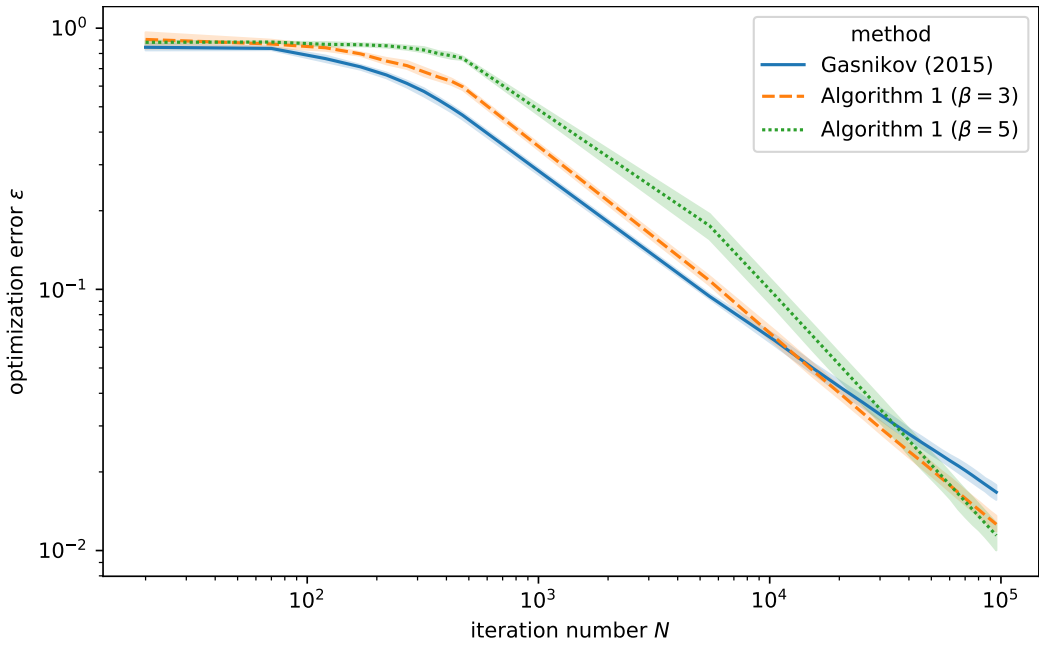

Figure 1.2 — The dependence of optimization error ε of Algorithm 1 on iteration number N

We see on Figure 1.2 that the usage of higher-order smoothness by Algorithm 1 helps to overcome the methods which do not use this.

Moreover, from Table 1 we see that the dependence of ε on N is better when we use higher-order smoothness.

Table 1 — The dependence of optimization error ε on iteration number N

	theory	experiment
Gasnikov, $\beta = 2$, 2015	$\varepsilon \sim N^{-0.5}$	$\varepsilon \sim N^{-0.61}$
Algorithm 1, $\beta = 3$, 2020	$\varepsilon \sim N^{-2/3}$	$\varepsilon \sim N^{-0.73}$
Algorithm 1, $\beta = 5$, 2020	$\varepsilon \sim N^{-4/5}$	$\varepsilon \sim N^{-0.91}$

1.6 Conclusion

For clarity we compare our results with state-of-the-art ones in Table 2 (dependence of optimization error ε on the number of iteration N, dimension n and β, μ) and Table 3 (dependence of the number of iteration N on the optimization error ε, dimension n and β, μ). To summarize the results we use $\tilde{O}()$, where $\tilde{O}()$ coincides with $O()$ up to the logarithmic factor.

Table 2 — The dependence of optimization error ε on N (number of iterations), n (dimension), μ, β

	strongly convex	convex
lower bound [1]	$O\left(\min\left(\frac{n}{\mu N^{\frac{\beta-1}{\beta}}}, \frac{n}{\sqrt{N}}\right)\right)$	$O\left(\min\left(\frac{\sqrt{n}}{N^{\frac{\beta-1}{2\beta}}}, \frac{n}{\sqrt{N}}\right)\right)^*$
this work (2020)	$\tilde{O}\left(\frac{n^{2-\frac{1}{\beta}}}{\mu N^{\frac{\beta-1}{\beta}}}\right)$	$\tilde{O}\left(\frac{n^{1-\frac{1}{2\beta}}}{N^{\frac{\beta-1}{2\beta}}}\right)$
Akhavan, Pontil, Tsybakov (2020) [1]	$\tilde{O}\left(\frac{n^2}{\mu N^{\frac{\beta-1}{\beta}}}\right)$	$\tilde{O}\left(\frac{n}{N^{\frac{\beta-1}{2\beta}}}\right)^*$
Bach, Perchet (2016) [2]	$O\left(\frac{n^{2-\frac{2}{\beta+1}}}{(\mu N)^{\frac{\beta-1}{\beta+1}}}\right)$	$O\left(\frac{n^{1-\frac{1}{\beta+1}}}{N^{\frac{\beta-1}{2(\beta+1)}}}\right)$
Gasnikov and al. (2015), $\beta = 2$, [5]	$\tilde{O}\left(\frac{n}{\sqrt{\mu N}}\right)$	$\tilde{O}\left(\frac{\sqrt{n}}{N^{1/4}}\right)$
Akhavan, Pontil, Tsybakov (2020), special case $\beta = 2$ [1]	$\tilde{O}\left(\frac{n}{\sqrt{\mu N}}\right)$	$\tilde{O}\left(\frac{\sqrt{n}}{N^{1/4}}\right)^*$
Zhang and al. (2020) [6]	$O\left(\frac{n}{\sqrt{\mu N}}\right)^*$	$O\left(\frac{\sqrt{n}}{N^{1/4}}\right)$

Table 3 — The dependence of N (number of iterations) on ε, n (dimension), μ, β

	strongly convex	convex
lower bound [1]	$O\left(\min\left(\frac{n^{1+\frac{1}{\beta-1}}}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}}, \frac{n^2}{\varepsilon^2}\right)\right)$	$O\left(\min\left(\frac{n^{1+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \frac{n^2}{\varepsilon^2}\right)\right)^*$
this work (2020)	$\tilde{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Akhavan, Pontil, Tsybakov (2020) [1]	$\tilde{O}\left(\frac{n^{2+\frac{2}{\beta-1}}}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{O}\left(\frac{n^{2+\frac{2}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)^*$
Bach, Perchet (2016) [2]	$O\left(\frac{n^{2+\frac{2}{\beta-1}}}{\mu\varepsilon^{\frac{\beta+1}{\beta-1}}}\right)$	$O\left(\frac{n^{2+\frac{2}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Gasnikov and al. (2015), $\beta = 2$ [5]	$\tilde{O}\left(\frac{n^2}{\mu\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^2}{\varepsilon^3}\right)$
Akhavan, Pontil, Tsybakov (2020), special case $\beta = 2$ [1]	$\tilde{O}\left(\frac{n^2}{\mu\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^2}{\varepsilon^3}\right)^*$
Zhang and al. (2020) [6]	$O\left(\frac{n^2}{\mu\varepsilon^2}\right)^*$	$O\left(\frac{n^2}{\varepsilon^3}\right)$

Comments on Table 2 and Table 3.

1. Note that in Table 2 and Table 3 the right column equals to the central one by $\mu \sim \varepsilon$.
2. Note that the results of this work have better dependency $\varepsilon(N)$ or $N(\varepsilon)$ than Gasnikov's one-point method only if $\beta > 2$ else another technique in Theorem 3 is better (see [5] or Theorem 5.1 in [1]). The result in this work is achieved using both kernel smoothing technique and measure concentration inequalities.
3. The lower bound for strongly convex case is got under conditions $\mu \geqslant N^{-1/2+1/\beta}$ (otherwise it is better to use convex methods) and (see [1]) $2\mu \leqslant \max\limits_{x \in Q} \|\nabla f(x)\|_2$.
4. The bounds marked with an asterisk * are not given in corresponding articles but they can be got.
5. Too optimistic bounds $O\left(\frac{n^{2-\frac{4}{\beta+1}}}{(\mu N)^{\frac{\beta-1}{\beta+1}}}\right)$ and $O\left(\frac{n^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$ were claimed in [2] instead of $O\left(\frac{n^{2-\frac{2}{\beta+1}}}{(\mu N)^{\frac{\beta-1}{\beta+1}}}\right)$ and $O\left(\frac{n^{2+\frac{2}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$, but Akhavan, Pontil and Tsybakov [1] found error in Lemma 2 in [2] where factor d of dimension (n in our notation) is missing.

Chapter 2. Exploiting Higher Order Smoothness in Derivative-free Saddle-Point Problems

2.1 Introduction

Recently GANs and Reinforcement Learning caused a big interest for saddle-point problems, see [29; 30; 35]. So in this section, we generalize the results for minimization problems to saddle-point problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\varphi(x,y), \tag{2.1}$$

where $\varphi(\cdot,y)$ is convex function defined on compact convex set $\mathcal{X}\subset\mathbb{R}^{n_x}$, $\varphi(x,\cdot)$ is concave function defined on compact convex set $\mathcal{Y}\subset\mathbb{R}^{n_y}$. For convenience, we denote $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ and then $z\in\mathcal{Z}\subset\mathbb{R}^n$ means $z\overset{\text{def}}{=}(x,y)$, where $x\in\mathcal{X}$, $y\in\mathcal{Y}$. When we use $\varphi(z)$, we mean $\varphi(z)=\varphi(x,y)$.

Function $\varphi(z)$ is convex-concave on $\mathcal{Z}$ if $\varphi(\cdot,y)$ is convex on $\mathcal{X}$ for all $y\in\mathcal{Y}$ and $\varphi(x,\cdot)$ is concave on $\mathcal{Y}$ for all $x\in\mathcal{X}$. Function $\varphi(z)$ is μ-strongly-convex-strongly-concave in $\mathcal{Z}$ with $\mu>0$ if $\varphi(\cdot,y)$ is μ-strongly-convex on $\mathcal{X}$ for all $y\in\mathcal{Y}$ and $\varphi(x,\cdot)$ is μ-strongly-concave on $\mathcal{Y}$ for all $x\in\mathcal{X}$.

In this chapter we study higher-order smooth functions φ functions satisfying so called generalized Hölder condition with parameter $\beta>2$, see (2). Let $\Phi_\beta(L_\beta)$ denote the set of all functions $\varphi:\mathbb{R}^n\to\mathbb{R}$ which satisfy Hölder condition (1.2) Let $\Phi_{\mu,\beta}(L_\beta)$ denote the set of μ-strongly-convex-strongly-concave functions $\varphi\in\Phi_\beta(L_\beta)$.

For the saddle-point problem (3) we propose to use Algorithm 2 which is a slightly modified version of Algorithm 1.

Algorithm 2 Zero-order Stochastic Projected Gradient for Saddle-Point Problems

Requires: Kernel $K:[-1,1]\to\mathbb{R}$, step size $\alpha_k>0$, parameters τ_k.

for $k=1,\dots,N$ **do**

1. Generate scalar r_k uniformly on $[-1,1]$ and vector e_k uniformly on the Euclidean unit sphere $S_n=\{e\in\mathbb{R}^n:\|e\|_2=1\}$.
2. $u_k:=\varphi(z_k+\tau_k r_k e_k)+\dot{\xi}_k$, $u_k':=\varphi(z_k-\tau_k r_k e_k)+\ddot{\xi}_k$
3. Define $\widetilde{g}_k:=\frac{n}{2\tau_k}(u_k-u_k')\begin{pmatrix}(e_k)_x\\-(e_k)_y\end{pmatrix}K(r_k)$
4. Update $z_{k+1}:=\Pi_Z(z_k-\alpha_k\widetilde{g}_k)$

end for

Output: $\{z_k\}_{k=1}^N$.

2.2 Contributions

Our main contributions can be summarized as follows:

1. For strongly-convex-strongly-concave case: under quasi-adversarial noise assumption (see Assumption 1) we establish for all $\beta > 2$ the upper bound of order $O\left(\frac{n^{2-\frac{1}{\beta}}}{\mu N^{\frac{\beta-1}{\beta}}}\right)$ for the optimization error of Algorithm 2 for strongly convex case.
2. For convex-concave case: under quasi-adversarial noise assumption (see Assumption 1) we establish for all $\beta > 2$ that after $N(\varepsilon) = O\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$ iterations of Algorithm 2 for the regularized function for the regularized function: $\varphi_\mu(z) := \varphi(z) + \frac{\mu}{2}\|x - x_0\|_2^2 - \frac{\mu}{2}\|y - y_0\|_2^2$, we achieve the optimization error less than or equal to ε.

2.3 Main Result

We prove upper bounds on the optimization error of Algorithm 2 for the saddle-point problem (3) for μ-strongly-convex-strongly-concave casefunction (Theorem 5) and of convex-concave function (Theorem 6).

Theorem 5. *Let $\varphi \in \Phi_{\mu,\beta}(L_\beta)$ with μ, $L_\beta > 0$ and $\beta > 2$. Let Assumption 1 hold and let $\mathcal{Z}$ be a convex compact subset of $\mathbb{R}^n$. Let φ be M_2-Lipschitz on the Euclidean τ_1-neighborhood of $\mathcal{Z}$ (see τ_k below).*

Then the rate of convergence is given by Algorithm 2 with parameters

$$\tau_k = \left(\frac{3\kappa\Delta^2 n}{2(\beta-1)(\kappa_\beta L_\beta)^2}\right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}, \quad \alpha_k = \frac{2}{\mu k}, \quad k = 1, \ldots, N$$

satisfies

$$\begin{aligned}\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] &\leqslant \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \\ &\leqslant \frac{1}{\mu}\left(n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N}\right),\end{aligned}$$

where $\overline{z}_N = \frac{1}{N}\sum\limits_{k=1}^{N} z_k$, $A_1 = 3\beta(\kappa\Delta^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L_\beta)^{\frac{2}{\beta}}$, $A_2 = 3c\kappa M_2^2$, κ_β and κ are constants depending only on β, see (1.5) and (1.6), $c < 15000$.

See the proof in the Appendix C.1.

Theorem 6. *Let $\varphi \in \Phi_\beta(L_\beta)$ with $L_\beta > 0$ and $\beta > 2$. Let Assumption 1 hold and let $\mathcal{Z}$ be a convex compact subset of $\mathbb{R}^n$. Let φ be M_2-Lipschitz on the Euclidean τ_1-neighborhood of $\mathcal{Z}$ (τ_k is*

parameter from Theorem 5 for the regularized function $\varphi_\mu(z)$ *whose description is given below). Let* $\overline{z}_N$ *denote* $\frac{1}{N}\sum\limits_{k=1}^{N} z_k$.

Let us define $N(\varepsilon)$*:*

$$N(\varepsilon) = \max\left\{\left(R\sqrt{2A_1}\right)^{\frac{2\beta}{\beta-1}} \frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \left(R\sqrt{2c'A_2}\right)^{2(1+\rho)} \frac{n^{1+\rho}}{\varepsilon^{2(1+\rho)}}\right\},$$

where $A_1 = 3\beta(\kappa\Delta^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L_\beta)^{\frac{2}{\beta}}$, $A_2 = 3c\kappa M_2^2$ *– constants from Theorem 5,* $\rho > 0$ *– arbitrarily small positive number,* c' *is a constant which depends on* ρ.

Then the rate of convergence is given by the following expression:

$$\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] \leqslant \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \leqslant \varepsilon \tag{2.2}$$

after $N(\varepsilon)$ *steps of Algorithm 2 with settings from Theorem 5 for the regularized function:* $\varphi_\mu(z) := \varphi(z) + \frac{\mu}{2}\|x - x_0\|_2^2 - \frac{\mu}{2}\|y - y_0\|_2^2$, *where* $\mu \leqslant \frac{\varepsilon}{R^2}$, $R = \|z_0 - z^*\|_2$, $z_0 \in \mathcal{Z}$ *– arbitrary point.*

See the proof in the Appendix C.2.

2.4 Conclusion

We compare our results for saddle-point problems (2.1) with state-of-the-art ones for minimization problems (1.1) in Table 4. We use $\tilde{O}()$, where $\tilde{O}()$ coincides with $O()$ up to the logarithmic factors.

Table 4 — Comparison of oracle complexity of one-point zeroth-order methods for higher order smooth *convex/strongly-convex* minimization (Min) and *convex-concave/strongly-convex-strongly-concave* saddle-point (SP) problems

	strongly convex or strongly-convex-strongly-concave	convex or convex-concave
lower bound [1]	$O\left(\min\left(\frac{n^{1+\frac{1}{\beta-1}}}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}}, \frac{n^2}{\varepsilon^2}\right)\right)$	$O\left(\min\left(\frac{n^{1+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \frac{n^2}{\varepsilon^2}\right)\right)^*$
Min. problem, Chapter 1, [36]	$\tilde{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
SP problem, Chapter 2, [37]	$\tilde{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{O}\left(\frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$

Comments on Table 4.

1. The bounds marked with an asterisk * are not given in corresponding articles but they can be got.
2. Although the saddle-point problems are more complicated compared to minimization problems, we see that the results for saddle-point problems in higher order smoothness case coincide with the results for minimization problems.

Chapter 3. Non-smooth Minimization Problem

3.1 Introduction

In this chapter we return to the minimization problem (1) and (1.1) but now $f : U_{\varepsilon_0}(Q) \to \mathbb{R}$ is *non-smooth convex* function, $Q \subset \mathbb{R}^n$ – convex compact set. We remind that the function is need to be defined in the small neighbourhood of the set Q but this is common in literature, see Section 1.3.2 for the details.

We assume the function f is Lipschitz-continuous with constant M, i.e. $|f(y) - f(x)| \leqslant M\|y - x\|_p$ on $U_{\varepsilon_0}(Q)$, where $p \in [1,2]$ and $\|\cdot\|_p$ is the p-norm (non-Euclidean setup). We use the notation M_2 for the Lipschitz constant with respect to Euclidean norm ($p = 2$). If $p \leqslant 2$ then $\|x\|_2 \leqslant \|x\|_p \leqslant n^{1/p-1/2}\|x\|_2$, that implies $M \leqslant M_2 \leqslant Mn^{1/p-1/2}$. We write q-norm for the dual of p-norm, so that $1/p + 1/q = 1$.

In this chapter we describe the *smoothing scheme* that allows us to develop batch-parallel gradient-free methods for non-smooth convex problems based on batched-gradient algorithms for smooth stochastic convex problems. Under adversarial noise assumption we apply smoothing scheme to batched Accelerated gradient method for the convex case. We also generalize smoothing scheme to strongly-convex optimization problems. A special focus is made to the calculation of the upper bounds for the admissible noise in the convex and strongly-convex case.

3.2 Contributions

Our main contributions can be summarized as follows:

1. For convex case: under adversarial noise assumption (Assumption 13) we prove Theorem 17 and Corollary 18 for zeroth order one-point Algorithm 3. This Algorithm shows optimal both oracle complexity and iteration complexity for Euclidean case ($p = 2$) and optimal oracle complexity for $p = 1$.
2. For strongly-convex case: under adversarial noise assumption (Assumption 13) we prove Theorem 19 for the restarting scheme applied to the Algorithm 3, that shows optimal both oracle complexity and iteration complexity for Euclidean case ($p = 2$) and optimal oracle complexity for $p = 1$.
3. We estimate the maximum level of adversarial noise that does not affect the rate of convergence for the convex and strongly convex case, see section 3.10. This estimate coincides with the lower bounds for the maximum noise level in certain regimes.

3.3 Notation

We denote standard scalar product $\langle x,y\rangle = \sum\limits_{i=1}^{n} x_i y_i$.

We denote the diameter D_p of the set Q in p-norm: $D_p = \max\limits_{x,y\in Q} \|x-y\|_p$.

We use $\tilde{O}(g(n))$ and $\tilde{\Omega}(h(n))$, which means $O(g(n))$ and $\Omega(h(n))$ respectively up to logarithmical on n factors.

For random variable u we denote centered random variable $\mathring{u} \stackrel{\text{def}}{=} u - \mathbb{E}[u]$. For any measurable function $f(u)$ we denote $\mathring{f}(u) \stackrel{\text{def}}{=} f(u) - \mathbb{E}[f(u)]$.

We denote conditional expectation $\mathbb{E}_x[f(x,y)|y] = \int_x x p_f(x,y) dx$, where $p_f(x,y)$ is a generalized probability density function of random variable f.

For any vector x and number k, we denote $x^{\bar{k}} = \{x^1, \ldots, x^k\}$.

3.4 Preliminaries

In this section we give the necessary notation, definitions and assumptions.

3.4.1 Distance generating function and Bregman divergence

We say that $d : Q \to \mathbb{R}$ is a distance generating function with respect respect to p-norm if it is continuously differentiable and 1-strongly convex w.r.t. p-norm, i.e.

$$d(y) \geqslant d(x) + \langle \nabla d(x), y - x\rangle + \frac{1}{2}\|y-x\|_p^2 \ \ \forall x,y \in Q.$$

Bregman divergence assosiated with distance generating function d is a function $V(x,y)$ that satisfies:

$$V(x,y) = d(y) - (d(x) + \langle \nabla d(x), y - x\rangle).$$

As $d(x)$ is 1-strongly convex w.r.t. p-norm, we obtain that $V(x,y) \geqslant \frac{1}{2}\|x-y\|_p^2$.

In this chapter we use the following distance generating function for p-norm:

$$d(x) = \begin{cases} \frac{\|x-x_1\|_p^2}{2(p-1)}, & a \leqslant p \leqslant 2, \\ \frac{e\|x-x_1\|_a^2}{2(a-1)}, & 1 \leqslant p \leqslant a, \end{cases} \tag{3.1}$$

where

$$a = \frac{2\ln n}{2\ln n - 1}, \tag{3.2}$$

and x_1 is a fixed point (starting point for the Algorithm 3 in Section 3.7).

This distance generating function is continuously differentiable as $a > 1$. The next Lemma shows that $d(x)$ from (3.1) is 1-strongly convex w.r.t. p-norm.

Lemma 7. *Let distance generating function $d(x)$ be defined as follows:*

$$d(x) = \begin{cases} \frac{\|x-x_1\|_p^2}{2(p-1)}, & a \leqslant p \leqslant 2, \\ \frac{e\|x-x_1\|_a^2}{2(a-1)}, & 1 \leqslant p < a, \end{cases} \tag{3.3}$$

where $a = \frac{2\ln n}{2\ln n - 1}$. Then $d(x)$ is 1-strongly convex w.r.t p-norm on $\mathbb{R}^n$.

See the proof in Appendix D.1.1.

We have the lower bound for Bregman divergence $V(x,y) \geqslant \frac{1}{2}\|x-y\|_p^2$, but for the future results we also need upper bound for Bregman divergence $V(x_1, x)$ between fixed point x_1 and arbitrary point x. The next Lemma describes the upper bound for $V(x_1, x)$ if $\|x - x_1\|_p \leqslant 1$:

Lemma 8. *Let distance generating function $d(x)$ be defined as follows:*

$$d(x) = \begin{cases} \frac{\|x-x_1\|_p^2}{2(p-1)}, & a \leqslant p \leqslant 2, \\ \frac{e\|x-x_1\|_a^2}{2(a-1)}, & 1 \leqslant p < a, \end{cases} \tag{3.4}$$

where $a = \frac{2\ln n}{2\ln n - 1}$. Then

$$2\max\{V(x_1, x) : \|x - x_1\|_p \leqslant 1\} \leqslant r_{p,n}^2, \tag{3.5}$$

where

$$r_{p,n}^2 \overset{def}{=} \begin{cases} \frac{1}{p-1}, & a \leqslant p \leqslant 2, \\ e(2\ln n - 1), & 1 \leqslant p < a. \end{cases}$$

See the proof in Appendix D.1.2.

If $\|x - x_1\|_p > 1$, then using norm properties and the properties of $V(x_1, x)$:

$$V(x_1, x) = V\left(x_1, x_1 + \frac{x - x_1}{\|x - x_1\|_p}\right)\|x - x_1\|_p^2 \leqslant \frac{r_{p,n}^2\|x - x_1\|_p^2}{2}.$$

In the case $p = 1$: $V(x_1, x) \leqslant O(\ln n)\|x - x_1\|_p^2$, that is $O(\ln n)$ times worser than for $p = 2$. The choice of another $d(x)$, for example, $d(x) = \sum_i x_i \ln x_i$ for $p = 1$ does not help to remove the logarithmical on dimension multiplier. However, the use of $p < 2$ is still efficient for certain cases.

Let us define prox mapping $\mathrm{Prox}_x : \mathbb{R}^n \to Q$ associated with distance generating function d and point x:

$$\mathrm{Prox}_x(y) = \arg\min_{z \in Q} \left(\langle y, z - x\rangle + V(x, z)\right).$$

For Euclidean case $d(x) = \frac{1}{2}\|x\|_2^2$, then $V(x,z) = \frac{1}{2}\|x - z\|_2^2$ and $\mathrm{Prox}_x(y) = \Pi_Q(x - y)$, where $\Pi_Q(z) = \arg\min_{v \in Q} \|z - v\|_2$ is Euclidean projection of z on Q, which is used in Algorithms 1 and 2.

3.4.2 Smoothing function

The main element of the smoothing scheme is the randomized *smoothing function* $f_\tau(x)$ for non-smooth objective f. This approach goes back to 1970s [13; 24]. For a small $\tau > 0$, the *smoothing function* for the function f is defined as follows:

$$f_\tau(x) = \mathbb{E}_u f(x + \tau u), \tag{3.6}$$

where u is random vector uniformly distributed on the unit Euclidean ball.

For a unit random vector e uniformly distributed on the unit Euclidean sphere with center at zero, let

$$g(x,e) = \frac{n}{2\tau}\left(f(x + \tau e) - f(x - \tau e)\right) e. \tag{3.7}$$

For the following results we need a bound on expectation $\mathbb{E}\|e\|_q^m$ for $m = 2$, $m = 4$, $m = 8$. The next Lemma 9 provides such a bound.

Lemma 9. *Let e be a random unit vector uniformly distributed on the unit Euclidean sphere with the zero center, so $\|e\|_2 = 1$. Assume $q > 2$ and $n \geqslant 2$. Then for any number $m : 1 \leqslant m \leqslant 8$ the expectation $\mathbb{E}\|e\|_q^m \leqslant a_{q,n}^m$, where $a_{q,n}^2 = \min\{4q - 1, 5\ln n\} n^{\frac{2}{q}-1}$. For the case $q = 2$, we can take $a_{2,n}^2 = 1$.*

See the proof in Appendix D.1.3.

Lemma 10 (Lemma 1 from [14]) proves that $g(x,e)$ is the unbiased estimator of $f_\tau(x)$.

Lemma 10. $\nabla f_\tau(x) = \mathbb{E}_e g(x,e)$.

See the proof in Appendix D.1.4.

If the function f is M_2-Lipschitz, then using the properties of distribution of Lipschitz functions on Euclidean sphere and concentration measure theory from [38] we can bound the second and the fourth moment of the $\|g(x,e)\|_q$.

Lemma 11. $\mathbb{E}_e\|g(x,e)\|_q^2 \leqslant \sqrt{\mathbb{E}_e\|g(x,e)\|_q^4} = O(a_{q,n}^2 n M^2)$, *where $a_{q,n}^2$ is defined in Lemma 9.*

See the proof in Appendix D.1.5.

The following lemma describes the properties of f_τ for non-Euclidean setup.

Lemma 12 (properties of f_τ)**.** *For all $x,y \in Q$, we have*

– *the inequality*

$$f(x) \leqslant f_\tau(x) \leqslant f(x) + \tau M_2; \tag{3.8}$$

– *$f_\tau(x)$ is M-Lipschitz:*

$$|f_\tau(y) - f_\tau(x)| \leqslant M\|y - x\|_p;$$

– $f_\tau(x)$ *has* $L = \dfrac{\sqrt{n}M}{\tau}$*-Lipschitz gradient:*

$$\|\nabla f_\tau(y) - \nabla f_\tau(x)\|_q \leqslant L\|y - x\|_p, \tag{3.9}$$

where q *us such that* $1/p + 1/q = 1$*;*

– $f_\tau(x)$ *inherits (strong) convexity of* $f(x)$*: if* $f(x)$ *is (*μ*-strongly) convex on* $U_{\varepsilon_0}(Q)$*, then* $f_\tau(x)$ *is (*μ*-strongly) convex on* Q *if* $\tau \leqslant \varepsilon_0$*.*

See the proof in Appendix D.1.6.

3.4.3 Gradient approximation

The function values are used in zeroth order algorithms to estimate the gradient. We assume that the function values are given with the additive noise $\xi = (\dot{\xi}, \ddot{\xi})$.

We use the gradient approximation similar to the one from Algorithm 1 but without kernel smoothing:

$$\widetilde{g}(x,e,\xi) = \frac{n}{2\tau}\left(f(x+\tau e) + \dot{\xi} - f(x-\tau e) - \ddot{\xi}\right)e, \tag{3.10}$$

where e is a random vector uniformly distributed on the Euclidean unit sphere in $\mathbb{R}^n$ (conditionally given x).

As in Assumption 1 we do not assume here neither zero-mean of $\dot{\xi}$ and $\ddot{\xi}$. Moreover, in contrast to the Assumption 1 we do not assume even the independence of noise variables $\dot{\xi}$ and $\ddot{\xi}$ from the random vector e.

Assumption 13 (noise assumptions)**.** *For noise* $\dot{\xi}$ *and* $\ddot{\xi}$ *from* (3.10) *and for* $\Delta \geqslant 0$ *it holds:*

$$\mathbb{E}_{e,\xi}\left[\dot{\xi}^2 \middle| x\right] \leqslant \Delta^2 \text{ and } \mathbb{E}_{e,\xi}\left[\ddot{\xi}^2 \middle| x\right] \leqslant \Delta^2,\ p = 2 \text{ (Euclidean case)}, \tag{3.11}$$

$$\mathbb{E}_{e,\xi}\left[\dot{\xi}^4 \middle| x\right] \leqslant \Delta^4 \text{ and } \mathbb{E}_{e,\xi}\left[\ddot{\xi}^4 \middle| x\right] \leqslant \Delta^4,\ 1 \leqslant p < 2 \text{ (non-Euclidean case)} \tag{3.12}$$

In this thesis, we use less restrictive noise assumptions $\mathbb{E}_{e,\xi}[\dot{\xi}^4|x] \leqslant \Delta^4$ for $p < 2$ and $\mathbb{E}_{e,\xi}[\dot{\xi}^2|x] \leqslant \Delta^2$ for $p = 2$ than the assumption $|\dot{\xi}| \leqslant \Delta$ always surely that is usually used in literature, see [39; 40]. The same is true for the noise $\ddot{\xi}$.

The next Lemma (14) describes the properties of the gradient estimator (3.10):

Lemma 14 (properties of $\widetilde{g}(x,e,\xi)$)**.** *For all* $x \in Q$*, under the Assumption 13 we have*

– *Bias estimate: for a fixed vector* $s \in \mathbb{R}^n$

$$|\mathbb{E}_{e,\xi}\left[\langle \widetilde{g}(x,e,\xi) - \nabla f_\tau(x), s\rangle \,|\, x\right]| \leqslant \frac{\sqrt{n}\Delta\|s\|_2}{\tau}. \tag{3.13}$$

– $\widetilde{g}(x,e)$ *has bounded second moment:*

$$\mathbb{E}_{e,\xi}\left[\|\widetilde{g}(x,e,\xi)\|_q^2 | x\right] = O\left(a_{q,n}^2 \cdot \left(nM_2^2 + \frac{n^2\Delta^2}{\tau^2}\right)\right), \tag{3.14}$$

where $1/p + 1/q = 1$ *and* $a_{q,n}^2$ *is defined in Lemma 9.*

See the proof in Appendix D.1.7.

3.5 Smoothing scheme

Now we are ready to describe general approach called *Smoothing scheme.*

Assume that we have some robust (it means the the bias does not accumulate over iterations) batched algorithm $\mathbf{A}(L,\sigma^2)$ that solves problem (1) under the assumption that f is smooth and Lipschitz-continuous with constant L, i.e. satisfies

$$\|\nabla f(y) - \nabla f(x)\|_q \leqslant L\|y - x\|_p, \ \forall x,y \in U_{\varepsilon_0}(Q). \tag{3.15}$$

Assume that this algorithm uses a stochastic *first order* oracle that depends on a random variable η and returns at a point x an stochastic gradient $\nabla_x f(x,\eta)$ which satisfies:

$$\mathbb{E}_\eta \left[\|\nabla_x f(x,\eta) - \nabla f(x)\|_q^2\right] \leqslant \sigma^2. \tag{3.16}$$

We assume that to reach ε-suboptimality in expectation, this algorithm requires $N(L,\varepsilon)$ successive iterations and $T(L,\sigma^2,\varepsilon)$ stochastic first order oracle calls, i.e. $\mathbf{A}(L,\sigma^2)$ allows batch parallelization with the batch size $B(L,\sigma^2,\varepsilon) = T(L,\sigma^2,\varepsilon)/N(L,\varepsilon)$.

Smoothing scheme. Apply $\mathbf{A}(L,\sigma^2)$ to the smoothed problem

$$\min_{x \in Q \subseteq \mathbb{R}^n} f_\tau(x) \tag{3.17}$$

with

$$\tau = \frac{\varepsilon}{2M_2} \tag{3.18}$$

and stochastic gradient estimator $\widetilde{g}(x,e,\xi)$, where $\varepsilon > 0$ is the desired accuracy for solving problem (1) in terms of the expectation.

According to (3.8) from Lemma 12, an $(\varepsilon/2)$-solution to (3.17) is an ε-solution to the initial problem (1). The algorithm requires the Lipschitz constant $L = \frac{\sqrt{n}M}{\tau}$ (3.9) of the $\nabla f_\tau(x)$. As $\tau = \frac{\varepsilon}{2M_2}$ (3.18), we have

$$L = \frac{2\sqrt{n}MM_2}{\varepsilon}. \tag{3.19}$$

Thus, we obtain that first order $\mathbf{A}(L,\sigma^2)$ with stochastic gradient estimator (3.10) becomes a zeroth order method for solving non-smooth problem (1).

We underline that this approach is flexible and generic as we can take different algorithms as $\mathbf{A}(L,\sigma^2)$. For example, we can take batched Accelerated gradient method [41—45] which is robust (as the third term in the RHS of (3.22) does not increase with the increase of N).

3.6 Batching

To reduce the variance (3.14) of gradient estimator we can use batched gradient estimator arised from (3.10):

$$\widetilde{g}_B\left(x, \left\{e^i\right\}_{i=1}^B, \left\{\xi^i\right\}_{i=1}^B\right) = \frac{1}{B}\left(\sum_{i=1}^B \widetilde{g}\left(x, e^i, \xi^i\right)\right) \tag{3.20}$$

where $\{e^i\}_{i=1}^B$ are random i.i.d. vectors uniformly distributed (conditionally given x) on the Euclidean unit sphere with zero center in $\mathbb{R}^n$, and for every i from 1 to B: ξ^i satisfies the adversarial noise Assumption 13 with corresponding e^i.

The well known fact for the variance of the mean of independent vectors in the Euclidean metrics is the following:

$$\mathbb{E}_{e^{\bar{B}}}\left[\left\|\frac{1}{B}\left(\sum_{i=1}^B \mathring{g}(x, e^i)\right)\right\|_2^2 \middle| x\right] \leqslant \frac{1}{B^2}\sum_{i=1}^B \mathbb{E}_{e^i}\left[\left\|\mathring{g}(x, e^i)\right\|_2^2 \middle| x\right],$$

where we used non-noisy ($\Delta = 0$) centered gradient estimator $\mathring{g}(x, e^i)$ instead of noisy $\mathring{\widetilde{g}}(x, e^i, \xi^i)$ to ensure independence. In the next lemma we prove the bound for the variance of non-noisy ($\Delta = 0$) batched gradient estimator in p-norm.

Lemma 15. *Let g_i be independent random variables with $\mathbb{E}\left\|\mathring{g}_i\right\|_q^4 = \sigma_i^4 < \infty$ for all i from 1 to B. Then it holds:*

$$\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^B \mathring{g}_i\right)\right\|_q^2 \leqslant 4r_{p,n}^2 \frac{1}{B^2}\sum_{i=1}^B \sigma_i^2,$$

where $r_{p,n}^2$ is defined in Lemma 8:

$$r_{p,n}^2 = \begin{cases} \frac{1}{p-1}, & \frac{2\ln n}{2\ln n - 1} \leqslant p \leqslant 2, \\ 2e\ln n - e, & 1 \leqslant p < \frac{2\ln n}{2\ln n - 1}. \end{cases}$$

If additionally $\sigma_i = \sigma$ for all i from 1 to B, then we can simplify the inequation above:

$$\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^B \mathring{g}_i\right)\right\|_q^2 \leqslant 4r_{p,n}^2 \frac{\sigma^2}{B}.$$

See the proof in Appendix D.1.8.

The bound for p-norm is similar to Euclidean case ($p = 2$) but with the additional condition on the boundness of fourths moments and the worser constant in the RHS.

For the noisy batched gradient estimator (3.20) we introduce σ_B^2 for the following expectation.

Lemma 16 (variation of batched noisy gradient estimator)**.**

$$\sigma_B^2 \stackrel{def}{=} \sup_{x \in Q} \mathbb{E}_{e^{\bar{B}}, \xi^{\bar{B}}}\left[\left\|\frac{1}{B}\left(\sum_{i=1}^B \widetilde{g}(x, e^i, \xi^i)\right) - \nabla f_\tau(x)\right\|_q^2 \middle| x\right] \leqslant O\left(a_{q,n}^2 \cdot \left(\frac{n r_{p,n}^2 M_2^2}{B} + \frac{n^2\Delta^2}{\tau^2}\right)\right),$$

where $1/p + 1/q = 1$, $a_{q,n}^2$ is defined in Lemma 9, $r_{p,n}^2$ is defined in Lemma 8.

See the proof in Appendix D.1.9.

3.7 Batched Smoothed accelerated Gradient method

In this section we apply Smoothing scheme to Accelerated Gradient Method from [42] (AC-SA assuming $h(x)$ from [42] is equal to zero) with the batched stochastic gradient estimator (3.20). The result is given in Algorithm 3.

Algorithm 3 Smoothed Batched Accelerated Gradient Method

Requires: Step sizes $\{\beta_k\}_{k=1}^N$ and $\{\gamma_k\}_{k=1}^N$, scalar parameter τ, batch size B.

Initialization: Let the initial point be $x_1^{ag} = x_1$ and distance generating function $d(x)$ from (3.1).

for $k = 1, \ldots, N$ **do**

1. Set $x_k^{md} = \beta_k^{-1} x_k + (1 - \beta_k^{-1}) x_k^{ag}$
2. Generate i.i.d. vectors $e_k^1, \ldots, e_k^B$ uniformly on the Euclidean unit sphere S_n with zero center and call the batched gradient approximation:

$$\widetilde{G}_k(x_k^{md}) = \widetilde{g}_B\left(x_k^{md}, \left\{e_k^i\right\}_{i=1}^B, \left\{\xi_k^i\right\}_{i=1}^B\right) = \frac{1}{B}\left(\sum_{i=1}^B \widetilde{g}_k\left(x_k^{md}, e_k^i, \xi_k^i\right)\right)$$

3. Update

$$x_{k+1} = \mathrm{Prox}_{x_k}(\gamma_k \widetilde{G}_k(x_k^{md})),$$
$$x_{k+1}^{ag} = \beta_k^{-1} x_{k+1} + (1 - \beta_k^{-1}) x_k^{ag}$$

end for

Output: x_{N+1}^{ag}.

3.8 Main Result

Let us define γ^* for the next theorem proving the convergence rate of the Algorithm 3:

$$\gamma^* = \min\left\{\frac{1}{2L}, \frac{2\sqrt{3} r_{p,n} R_p}{(N+2)^{3/2} \cdot \sigma_B}\right\}, \tag{3.21}$$

where σ_B^2 is defined in Lemma 16, $r_{p,n}^2$ is defined in Lemma 8, $R_p = \|x_1 - x_\tau^*\|_p$ is a distance in p-borm between the starting point x_1 and any optimal solution of smoothed problem $x_\tau^* \in \arg\min_{x \in Q} f_\tau(x)$.

Define distance generating function for the Algorithm 3 from (3.1), where x_1 is a starting point. Remember, that distance generating function (3.1) depends on starting point x_1.

The next theorem proves the dependence of the expectation of functional suboptimality $\mathbb{E}\left[f(x_{N+1}^{ag}) - f(x^*)\right]$ on the iteration number N:

Theorem 17. *Assume that f is convex and M-Lipschitz w.r.t. p-norm and M_2-Lipschitz w.r.t. Euclidean norm. Let us define step sizes of Algorithm 3 as $\beta_k = \frac{(k+1)}{2}$, $\gamma_k = \frac{(k+1)\gamma^*}{2}$ for all k from 1 to N, where γ^* is defined from* (3.21)*. Then under the Assumption 13:*

$$\mathbb{E}\left[f(x_{N+1}^{ag}) - f(x^*)\right] \leqslant \frac{8Lr_{p,n}^2R_p^2}{N^2} + \frac{4\sigma_B r_{p,n}R_p}{\sqrt{N}} + \frac{\Delta\sqrt{n}D_2}{\tau} + M_2\tau, \tag{3.22}$$

where $L = \frac{\sqrt{n}M_2}{\tau}$ (3.9)*, σ_B is defined in Lemma 16, $r_{p,n}$ is defined in Lemma 8, $D_2 = \max\limits_{x,y\in Q}\|x-y\|_2$ is the Euclidean diameter of the set Q, $R_p = \|x_1 - x_\tau^*\|_p$ is the distance from the starting point x_1 to the minimizer x_τ^* of the function $f_\tau(x)$ on the set Q: $x_\tau^* \in \arg\min_{x\in Q} f_\tau(x)$.*

See the proof in Appendix D.2.1.

We can substitute the following bounds for τ, L and σ_B^2 into (3.22):

$$\tau \overset{(3.18)}{=} \frac{\varepsilon}{2M_2}, \tag{3.23}$$

$$L \overset{(3.9)}{\leqslant} \frac{\sqrt{n}M}{\tau} \simeq \frac{\sqrt{n}MM_2}{\varepsilon}, \tag{3.24}$$

$$\sigma_B^2 \overset{Lemma\ 16}{\lesssim} a_{q,n}^2 \cdot \left(\frac{nr_{p,n}^2M_2^2}{B} + \frac{n^2\Delta^2}{\tau^2}\right). \tag{3.25}$$

If the noise level Δ is sufficiently small (see Section 3.10 for details), then to achieve ε-accuracy in terms of expectation of functional suboptimality $\mathbb{E}\left[f(x_{N+1}^{ag}) - f(x^*)\right] \lesssim \varepsilon$, we need

$$\frac{\sqrt{n}MM_2r_{p,n}^2R_p^2}{\varepsilon N^2} + \frac{a_{q,n}r_{p,n}^2R_p\sqrt{n}M_2}{\sqrt{N\cdot B}} \lesssim \varepsilon.$$

Corollary 18. *Based on the batched Accelerated gradient method, the Smoothing scheme applied to non-smooth problem* (1)*, provides a gradient-free method with*

$$N(\varepsilon) = O\left(\frac{n^{1/4}\sqrt{M_2M}r_{p,n}R_p}{\varepsilon}\right) = \begin{cases} O\left(\frac{n^{1/4}M_2R_2}{\varepsilon}\right), & p = 2\ (q=2), \\ O\left(\frac{(\ln n)^{1/2}n^{1/4}\sqrt{M_2M}R_1}{\varepsilon}\right), & p = 1\ (q=\infty) \end{cases}$$

successive iterations and

$$T(\varepsilon) = N(\varepsilon)\cdot B(\varepsilon) = O\left(\frac{a_{q,n}^2nM_2^2r_{p,n}^4R_p^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{nM_2^2D_2^2}{\varepsilon^2}\right), & p = 2\ (q=2), \\ O\left(\frac{(\ln n)^3M_2^2R_1^2}{\varepsilon^2}\right), & p = 1\ (q=\infty) \end{cases}$$

zeroth-order oracle calls, where $1/p+1/q = 1$, $a_{q,n}^2$ is defined in Lemma 9, $r_{p,n}^2$ is defined in Lemma 8; $D_2 = \max\limits_{x,y\in Q}\|x-y\|_2$ is the Euclidean diameter of the set Q, $R_p = \|x_1 - x_\tau^*\|_p$ is the distance from the start point x_1 to the minimizer x_τ^* of the function $f_\tau(x)$ on the set Q: $x_\tau^* \in \arg\min_{x\in Q} f_\tau(x)$.

See the proof in Appendix D.2.2.

3.9 Strongly Convex Problems

By using the restart technique [13; 46] we can prove a counterpart of Theorem 17 for the case when f is μ-strongly convex w.r.t. the p-norm for some $p \in [1,2]$ and $\mu \geqslant \frac{\varepsilon}{2R_p^2}$ (otherwise, it is better to use the algorithm for convex but not strongly convex functions), where $R_p = \|x_{start} - x_\tau^*\|_p$ with $x_\tau^* = \arg\min_{x \in Q} f_\tau(x)$, $\tau = \frac{\mu R_p^2}{4M_2}$.

Theorem 19. *Based on the batched Accelerated gradient method, the Smoothing scheme applied to non-smooth and strongly convex problem* (1)*, provides a gradient-free method with*

$$O\left(\frac{n^{1/4}\sqrt{M_2 M} r_{p,n}}{\sqrt{\mu\varepsilon}}\right) = \begin{cases} O\left(\frac{n^{1/4} M_2}{\sqrt{\mu\varepsilon}}\right), & p = 2 \ (q = 2), \\ O\left(\frac{(\ln n)^{1/2} n^{1/4} \sqrt{M_2 M}}{\sqrt{\mu\varepsilon}}\right), & p = 1 \ (q = \infty) \end{cases}$$

successive iterations and

$$O\left(\frac{a_{q,n}^2 n M_2^2 r_{p,n}^4}{\mu\varepsilon}\right) = \begin{cases} O\left(\frac{n M_2^2}{\mu\varepsilon}\right), & p = 2 \ (q = 2), \\ O\left(\frac{(\ln n)^3 M_2^2}{\mu\varepsilon}\right), & p = 1 \ (q = \infty) \end{cases}$$

zeroth-order oracle calls, where $1/p + 1/q = 1$, $a_{q,n}^2$ *is defined in Lemma 9,* $r_{p,n}^2$ *is defined in Lemma 8.*

See the proof in Appendix D.2.3.

3.10 Noise

In this section we estimate the maximum level of noise that does not impact the rate of convergence of Algorithm 3. The detailed calculations are presented in Appendix D.3.

We consider here only the convex case, for the strongly convex case, the maximum noise level coincides with convex case, see the last paragraph in Appendix D.3.

For convex case, we use the Theorem 17 and substitute (3.23), (3.24) and (3.25) into (3.22) to obtain:

$$\mathbb{E}\left[f(x_{N+1}^{ag}) - f(x^*)\right] \lesssim \frac{\sqrt{n} M M_2 r_{p,n}^2 R_p^2}{\varepsilon N^2} + \frac{a_{q,n} r_{p,n} R_p}{\sqrt{N}} \left(\frac{\sqrt{n} r_{p,n} M_2}{\sqrt{B}} + \frac{n \Delta M_2}{\varepsilon}\right) + \frac{\Delta M_2 D_2 \sqrt{n}}{\varepsilon} + \varepsilon. \tag{3.26}$$

To preserve the rate of convergence of Algorithm 3, the noise containing terms in (3.26) must satisfy:

$$\frac{\Delta M_2 D_2 \sqrt{n}}{\varepsilon} + \frac{a_{q,n} r_{p,n} R_p}{\sqrt{N}} \cdot \frac{n \Delta M_2}{\varepsilon} \lesssim \varepsilon.$$

Thus, substituting $N(\varepsilon) = O\left(\frac{n^{1/4}\sqrt{M_2 M} r_{p,n} R_p}{\varepsilon}\right)$ from Corollary 18, we obtain:

$$\Delta = O\left(\min\left\{\frac{\varepsilon^2}{\sqrt{n} M_2 D_2}, \frac{M^{1/4}\varepsilon^{3/2}}{a_{q,n}\sqrt{r_{p,n}}\sqrt{n} M_2^{3/4}\sqrt{R_p}}\right\}\right) = O(\min\{\Delta_1, \Delta_2\}),$$

where $1/p + 1/q = 1$, $a_{q,n}^2$ is defined in Lemma 9, $r_{p,n}^2$ is defined in Lemma 8 and

$$\Delta_1 = \frac{\varepsilon^2}{\sqrt{n} M_2 D_2}, \Delta_2 = \frac{M^{1/4}\varepsilon^{3/2}}{a_{q,n}\sqrt{r_{p,n}}\sqrt{n} M_2^{3/4}\sqrt{R_p}}.$$

Noise level Δ_1 restricts the bias term $\frac{\Delta\sqrt{n} D_2}{\tau}$ in (3.22) and noise level Δ_2 arises from the second moment σ_B^2 of the batched gradient estimator in q-norm (see Lemma 16).

The detailed calculation (see (D.97) in Appendix D.3) shows, that $\Delta_1 \lesssim \Delta_2$, so it is sufficient for noise to be:

$$\Delta = O\left(\frac{\varepsilon^2}{\sqrt{n} M_2 D_2}\right).$$

Note that our upper bound coincides with the first part of the lower bound for the maximum noise level is $\Delta = \tilde{\Omega}\left(\max\left\{\frac{\varepsilon^2}{\sqrt{n}}, \frac{\varepsilon}{n}\right\}\right)$, which is obtained for the class of algorithms using polynomial $\text{Poly}\left(n, \frac{1}{\varepsilon}\right)$ noisy function calculations $f(x) + \xi$, see [34]. Here $\tilde{\Omega}()$ means $\Omega()$ up to logarithmical $\text{PolyLog}\left(n, \frac{1}{\varepsilon}\right)$ factors.

3.11 Conclusion

We compare our results for the convex case with state-of-the-art ones for minimization problems (1.1) in Table 5.

Table 5 — Comparison of iteration and oracle complexity of one-point zeroth-order methods for non-smooth convex minimization problems

	iteration complexity $N(\varepsilon)$	oracle complexity $T(\varepsilon)$
lower bound, $p = 1$, [15; 47]	$\tilde{O}\left(\min\left\{n^{1/4}\varepsilon^{-1}, n^{1/3}\varepsilon^{-2/3}\right\}\right)$	$\tilde{O}\left(n \cdot \varepsilon^{-2}\right)$
Algorithm 3, $p = 1$, Chapter 3, [39]	$\tilde{O}\left(\left(\frac{M_2}{M}\right)^{1/2} \cdot n^{1/4}\varepsilon^{-1}\right)$	$\tilde{O}\left(\frac{M_2^2}{M^2} \cdot \varepsilon^{-2}\right)$
lower bound, $p = 2$, [15; 47; 48]	$\tilde{O}\left(\min\left\{n^{1/4}\varepsilon^{-1}, n^{1/3}\varepsilon^{-2/3}\right\}\right)$	$\tilde{O}\left(n\varepsilon^{-2}\right)$
Algorithm 3, $p = 2$, Chapter 3, [39]	$O\left(n^{1/4}\varepsilon^{-1}\right)$	$O\left(n\varepsilon^{-2}\right)$
Scaman et al., 2019, $p = 2$, [49], first-order method	$O\left(n^{1/4}\varepsilon^{-1}\right)$	$O\left(n\varepsilon^{-2}\right)^*$
Bubeck et al., $p = 2$, 2019, [48]	$O\left(n^{1/3}\varepsilon^{-2/3}\right)$	$O\left(n^{4/3}\varepsilon^{-8/3}\right)$

Comments on Table 5.

1. In [49] the authors used first-order method, using $O(\varepsilon^{-2})$ calculations of true gradient $\nabla f(x)$, so we marked with an asterisk * their result for oracle complexity. Our Algorithm 3 is zeroth-order and shows the similar performance.
2. Lower bound is obtained for a wider class of algorithms that allow Poly (n,ε^{-1}) oracle calls per iteration.
3. For Euclidean case ($p = 2$), both iteration and oracle complexities coincide with the lower bound in certain regime: $n \gtrsim \left(\frac{MR_p}{\varepsilon}\right)^4$.
4. For the case $p = 1$, oracle complexity coincides with the lower bound as $M_2 \leqslant \sqrt{n}M$ and in the worst case $M_2 = \sqrt{n}M$. Iteration complexity is $\sqrt[4]{n}$ times worser up to a logarithmical factor.
5. For the Euclidean case ($p = 2$), the authors of [48] obtain a better in some regimes bound $N \sim n^{1/3}/\varepsilon^{2/3}$ for the number of iterations. On the contrary, they have significantly worser oracle complexity $T(\varepsilon)$.

Conclusion

In this work, we have addressed a broad spectrum of topics in the field of numerical optimization, with a particular emphasis on leveraging higher-order smoothness for minimization and saddle-point problems. These advanced methods offer significant improvements over traditional approaches that rely solely on functional or gradient smoothness, potentially leading to faster convergence.

A primary focus of this dissertation is on developing minimization and saddle-point algorithms that exploit higher-order smoothness of the target function. By incorporating the concept of higher-order smoothness, our methods have the potential to outperform conventional algorithms that rely on first-order or gradient-based information.

The dissertation also delves into non-smooth optimization, addressing the challenge of adversarial noise. We have shown that it is possible to maintain optimal iteration complexity without sacrificing oracle complexity, even in the presence of adversarial noise. This finding is significant because it implies that effective optimization can be achieved without compromising the quality of the solutions or the efficiency of the algorithms. Our results demonstrate that by employing smoothing techniques, it is feasible to adapt first-order methods to non-smooth problems, thereby expanding their applicability to a wider range of real-world scenarios where traditional methods may fall short.

Furthermore, a crucial aspect of our work is the evaluation of the maximum noise level that can be tolerated in non-smooth optimization problems. We have derived bounds that quantify how much adversarial noise can be accommodated while still achieving optimal performance. This evaluation is critical for understanding the robustness of our algorithms and for ensuring their practical applicability.

Our exploration covers both theoretical and practical aspects of the discussed problems. The theoretical contributions include the development of new algorithms and convergence bounds for higher-order smoothness in minimization and saddle-point problems. These results offer valuable insights into how these advanced techniques can be applied and provide a solid foundation for further research. On the practical side, our numerical experiment validates the effectiveness of the proposed algorithms, showcasing their advantages over existing methods in optimization tasks.

A promising direction for future research is to extend our results to obtain large-probability bounds for optimization errors in the higher-order smoothness case. This would involve generalizing our current findings to account for probabilistic guarantees, thereby providing a more comprehensive understanding of the performance of our algorithms. Such extensions would further enhance the robustness and applicability of the proposed methods, offering a broader scope for their use in diverse practical applications.

Overall, this dissertation makes a significant contribution to the field of optimization by introducing new methods that improve theoretical understanding and offer practical solutions applicable in machine learning, signal processing, and other advanced applications.

References

1. Akhavan, A. Exploiting Higher Order Smoothness in Derivative-free Optimization and Continuous Bandits / A. Akhavan, M. Pontil, A. B. Tsybakov // arXiv preprint arXiv:2006.07862. — 2020. — URL: https://arxiv.org/abs/2006.07862 (дата обр. 22.08.2024).
2. Bach, F. Highly-smooth zero-th order online optimization / F. Bach, V. Perchet // Conference on Learning Theory. — 2016. — С. 257—283.
3. Gasnikov, A. Stochastic gradient methods with inexact oracle / A. Gasnikov, P. Dvurechensky, Y. Nesterov // arXiv preprint arXiv:1411.4218. — 2014. — URL: https://arxiv.org/abs/1411.4218 (дата обр. 22.08.2024).
4. Gasnikov, A. Gradient and gradient-free methods for stochastic convex optimization with inexact oracle / A. Gasnikov, P. Dvurechensky, D. Kamzolov // arXiv preprint arXiv:1502.06259. — 2015. — URL: https://arxiv.org/abs/1502.06259 (дата обр. 22.08.2024).
5. Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case / A. V. Gasnikov [и др.] // Automation and remote control. — 2017. — Т. 78, № 2. — С. 224—234.
6. Boosting One-Point Derivative-Free Online Optimization via Residual Feedback / Y. Zhang [и др.] // IEEE Transactions on Automatic Control. — 2024. — С. 1—8.
7. Bubeck, S. Kernel-based methods for bandit convex optimization / S. Bubeck, Y. T. Lee, R. Eldan // Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing. — 2017. — С. 72—85.
8. Granichin, O. N. A stochastic recursive procedure with correlated noises in the observation, that employs trial perturbations at the input / O. N. Granichin // Vestnik of the Leningrad University. Mathematics. — 1989. — № 1. — С. 27—31.
9. Granichin, O. Randomized Algorithms of an Estimation and Optimization Under Almost Arbitrary Noises / O. Granichin, B. Polyak. — М.: Nauka, 2003.
10. Larson, J. Derivative-free optimization methods / J. Larson, M. Menickelly, S. M. Wild // Acta Numerica. — 2019. — Т. 28. — С. 287—404.
11. Conn, A. R. Introduction to Derivative-Free Optimization / A. R. Conn, K. Scheinberg, L. N. Vicente. — Society for Industrial, Applied Mathematics, 2009.
12. Spall, J. C. Introduction to Stochastic Search and Optimization / J. C. Spall. — 1-е изд. — New York, NY, USA : John Wiley & Sons, Inc., 2003.
13. Nemirovsky, A. Problem Complexity and Method Efficiency in Optimization.-J. Wiley & Sons, New York / A. Nemirovsky, D. Yudin. — 1983.
14. Flaxman, A. D. Online convex optimization in the bandit setting: gradient descent without a gradient / A. D. Flaxman, A. T. Kalai, H. B. McMahan // Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms. — 2005. — С. 385—394.

15. Optimal rates for zero-order convex optimization: The power of two function evaluations / J. C. Duchi [и др.] // IEEE Transactions on Information Theory. — 2015. — Т. 61, № 5. — С. 2788—2806.

16. Nesterov, Y. Random gradient-free minimization of convex functions / Y. Nesterov, V. Spokoiny // Foundations of Computational Mathematics. — 2017. — Т. 17, № 2. — С. 527—566.

17. Fabian, V. Stochastic approximation of minima with improved asymptotic speed / V. Fabian // The Annals of Mathematical Statistics. — 1967. — С. 191—200.

18. Polyak, B. T. Optimal order of accuracy of search algorithms in stochastic optimization / B. T. Polyak, A. B. Tsybakov // Problemy Peredachi Informatsii. — 1990. — Т. 26, № 2. — С. 45—53.

19. Dippon, J. Accelerated randomized stochastic optimization / J. Dippon // The Annals of Statistics. — 2003. — Т. 31, № 4. — С. 1260—1281.

20. Stochastic convex optimization with bandit feedback / A. Agarwal [и др.] // Advances in Neural Information Processing Systems. — 2011. — Т. 24.

21. Shamir, O. On the complexity of bandit and derivative-free stochastic convex optimization / O. Shamir // Conference on Learning Theory. — PMLR. 2013. — С. 3—24.

22. Bartlett, P. L. A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption / P. L. Bartlett, V. Gabillon, M. Valko // Algorithmic Learning Theory. — PMLR. 2019. — С. 184—206.

23. Locatelli, A. Adaptivity to smoothness in x-armed bandits / A. Locatelli, A. Carpentier // Conference on Learning Theory. — PMLR. 2018. — С. 1463—1492.

24. Ermoliev, Y. Stochastic programming methods / Y. Ermoliev. — 1976.

25. Katkovnik, V. Y. Linear Bounds and Stochastic Optimization Problems: Method of Parametric Averaging Operators / V. Y. Katkovnik. — Nauka, 1976.

26. Ben-Tal, A. Robust optimization. Т. 28 / A. Ben-Tal, L. El Ghaoui, A. Nemirovski. — Princeton university press, 2009.

27. Nesterov, Y. Smooth minimization of non-smooth functions / Y. Nesterov // Mathematical programming. — 2005. — Т. 103. — С. 127—152.

28. Convex optimization, game theory, and variational inequality theory / G. Scutari [и др.] // IEEE Signal Processing Magazine. — 2010. — Т. 27, № 3. — С. 35—49.

29. Goodfellow, I. NIPS 2016 tutorial: Generative adversarial networks / I. Goodfellow // arXiv preprint arXiv:1701.00160. — 2016. — URL: https://arxiv.org/abs/1701.00160 (дата обр. 22.08.2024).

30. Jin, Y. Efficiently Solving MDPs with Stochastic Mirror Descent / Y. Jin, A. Sidford // Proceedings of the 37th International Conference on Machine Learning. Т. 119 / под ред. H. D. III, A. Singh. — PMLR, 13–18 Jul.2020. — С. 4890—4900. — (Proceedings of Machine Learning Research).

31. Yousefian, F. On stochastic gradient and subgradient methods with adaptive steplength sequences / F. Yousefian, A. Nedić, U. V. Shanbhag // Automatica. — 2012. — Т. 48, № 1. — С. 56—67. — URL: https://www.sciencedirect.com/science/article/pii/S0005109811004833.

32. Duchi, J. C. Randomized smoothing for stochastic optimization / J. C. Duchi, P. L. Bartlett, M. J. Wainwright // SIAM Journal on Optimization. — 2012. — Т. 22, № 2. — С. 674—701.

33. Beznosikov, A. Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem / A. Beznosikov, A. Sadiev, A. Gasnikov // International Conference on Mathematical Optimization Theory and Operations Research. — Springer. 2020. — С. 105—119.

34. Risteski, A. Algorithms and matching lower bounds for approximately-convex optimization / A. Risteski, Y. Li // Advances in Neural Information Processing Systems. — 2016. — Т. 29. — С. 4745—4753.

35. Zeroth-Order Algorithms for Smooth Saddle-Point Problems / A. Sadiev [и др.] // arXiv preprint arXiv:2009.09908. — 2020. — URL: https://arxiv.org/abs/2009.09908 (дата обр. 22.08.2024).

36. Novitskii, V. Improved exploitation of higher order smoothness in derivative-free optimization / V. Novitskii, A. Gasnikov // Optimization Letters. — 2022. — Т. 16, № 7. — С. 2059—2071.

37. Beznosikov, A. One-point gradient-free methods for smooth and non-smooth saddle-point problems / A. Beznosikov, V. Novitskii, A. Gasnikov // International Conference on Mathematical Optimization Theory and Operations Research. — Springer. 2021. — С. 144—158.

38. Ledoux, M. The concentration of measure phenomenon / M. Ledoux. — American Mathematical Soc., 2001.

39. The power of first-order smooth optimization for black-box non-smooth problems / A. Gasnikov [и др.] // Proceedings of the 39th International Conference on Machine Learning. Т. 162 / под ред. K. Chaudhuri [и др.]. — PMLR, 17–23 Jul.2022. — С. 7241—7265. — (Proceedings of Machine Learning Research).

40. Beznosikov, A. Derivative-free method for composite optimization with applications to decentralized distributed optimization / A. Beznosikov, E. Gorbunov, A. Gasnikov // IFAC-PapersOnLine. — 2020. — Т. 53, № 2. — С. 4038—4043.

41. Better Mini-Batch Algorithms via Accelerated Gradient Methods / A. Cotter [и др.] // Advances in Neural Information Processing Systems. — 2011. — Т. 24. — С. 1647—1655.

42. Lan, G. An optimal method for stochastic composite optimization / G. Lan // Mathematical Programming. — 2012. — Т. 133, № 1. — С. 365—397.

43. Devolder, O. Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization : дис. ... канд. / Devolder Olivier. — PhD thesis, 2013.

44. Dvurechensky, P. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle / P. Dvurechensky, A. Gasnikov // Journal of Optimization Theory and Applications. — 2016. — Т. 171, № 1. — С. 121—145.

45. Gorbunov, E. Optimal decentralized distributed algorithms for stochastic convex optimization / E. Gorbunov, D. Dvinskikh, A. Gasnikov // arXiv preprint arXiv:1911.07363. — 2019. — URL: https://arxiv.org/abs/1911.07363 (дата обр. 22.08.2024).

46. Juditsky, A. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization / A. Juditsky, Y. Nesterov // Stochastic Systems. — 2014. — Т. 4, № 1. — С. 44—80.

47. Diakonikolas, J. Lower Bounds for Parallel and Randomized Convex Optimization. / J. Diakonikolas, C. Guzmán // J. Mach. Learn. Res. — 2020. — Т. 21. — С. 5—1.

48. Complexity of highly parallel non-smooth convex optimization / S. Bubeck [и др.] // Advances in neural information processing systems. — 2019.

49. Optimal convergence rates for convex distributed optimization in networks / K. Scaman [и др.] // Journal of Machine Learning Research. — 2019. — Т. 20. — С. 1—31.

50. Gorbunov, E. 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. / E. Gorbunov, E. A. Vorontsova, A. V. Gasnikov // Mathematical Notes. — 2019. — Т. 106.

List of figures

List of tables

Appendix A

Appendix: Basic Facts

Lemma 20 (three point lemma)**.** $V(x,z) = V(x,y) + \langle \nabla d(y) - \nabla d(x), z - y\rangle + V(y,z)$.

Proof. By definition $V(x,y) = d(y) - (d(x) + \langle \nabla d(x), y - x\rangle)$. Then

$$\begin{aligned} V(x,z) - V(x,y) - V(y,z) &= -\langle \nabla d(x), z - x\rangle + \langle \nabla d(x), y - x\rangle + \langle \nabla d(y), z - y\rangle \\ &= -\langle \nabla d(x), z - y\rangle + \langle \nabla d(y), z - y\rangle = \langle \nabla d(y) - \nabla d(x), z - y\rangle. \end{aligned}$$

□

Lemma 21. *Let $x_{n+1} = Prox_g(x_n)$. Then for any $x \in Q$:*

$$\langle g, x_{n+1} - x\rangle \leqslant V(x_n, x) - V(x_{n+1}, x) - V(x_n, x_{n+1}).$$

Proof.
By definition of prox mapping:

$$x_{n+1} = \arg\min_{x \in Q} \left(\langle g, x - x_n\rangle + V(x_n, x)\right).$$

Differentiating (we use, that $\nabla_x V(x_n,x) = \nabla d(x) - \nabla d(x_n)$), we obtain for any x in Q:

$$\langle g + \nabla d(x_{n+1}) - \nabla d(x_n), x - x_{n+1}\rangle \geqslant 0$$

$V(x,z) = V(x,y) + \langle \nabla d(y) - \nabla d(x), z - y\rangle + V(y,z)$.
Rearranging the terms and combining with three point lemma (Lemma 20), we get:

$$\langle g, x_{n+1} - x\rangle \leqslant \langle \nabla d(x_{n+1}) - \nabla d(x_n), x - x_{n+1}\rangle = V(x_n, x) - V(x_n, x_{n+1}) - V(x_{n+1}, x).$$

□

Lemma 22 (relation between p-norms)**.** *Let $p_2 > p_1 \geqslant 1$. Then $\|x\|_{p_2} \leqslant \|x\|_{p_1} \leqslant n^{\frac{1}{p_1} - \frac{1}{p_2}} \|x\|_{p_2}$.*

Proof.
For the right inequality, we use Holder inequality for $\frac{1}{r}$ and $1 - \frac{1}{r}$, $r > 1$:

$$\sum_{i=1}^{n} |a_i||b_i| \leqslant \left(\sum_{i=1}^{n} |a_i|^r\right)^{\frac{1}{r}} \left(\sum_{i=1}^{n} |b_i|^{\frac{r}{r-1}}\right)^{1-\frac{1}{r}}.$$

Apply it to the case $|a_i| = |x_i|^{p_1}$, $|b_i| = 1$ and $r = \frac{p_2}{p_1} > 1$:

$$\sum_{i=1}^{n} |x_i|^{p_1} \leqslant \left(\sum_{i=1}^{n} |x_i|^{p_2}\right)^{\frac{p_1}{p_2}} \cdot n^{1 - \frac{p_1}{p_2}}.$$

Then:

$$\|x\|_{p_1} = \left(\sum_{i=1}^{n} |x_i|^{p_1}\right)^{\frac{1}{p_1}} \leqslant \left(\sum_{i=1}^{n} |x_i|^{p_2}\right)^{\frac{p_1}{p_2}\cdot\frac{1}{p_1}} \cdot n^{\frac{1}{p_1}-\frac{1}{p_2}} = \|x\|_{p_2} \cdot n^{\frac{1}{p_1}-\frac{1}{p_2}}.$$

For the left inequality, without loss of generality, we can take $y = \frac{x}{\|x\|_{p_1}}$. As $\|y\|_{p_1} = 1$, the same holds for all coordinates absolute values $|y_i| \leqslant 1$, and positive degree $\frac{p_2}{p_1}$ does not increase them $|y_i|^{\frac{p_2}{p_1}} \leqslant |y_i|$, the same is $|y_i|^{p_2} \leqslant |y_i|^{p_1}$. Then

$$\|y\|_{p_2} = \left(\sum_{i=1}^{n} |y_i|^{p_2}\right)^{\frac{1}{p_2}} \leqslant \left(\sum_{i=1}^{n} |y_i|^{p_1}\right)^{\frac{1}{p_2}} = \left(\|y\|_{p_1}^{p_1}\right)^{\frac{1}{p_2}} = 1^{\frac{1}{p_2}} = 1,$$

consequently,

$$\|x\|_{p_2} = \|x\|_{p_1}\|y\|_{p_2} \leqslant \|x\|_{p_1} \cdot 1 = \|x\|_{p_1}.$$

□

Lemma 23 (tails of standard normal distribution)**.** *Let $\eta \sim \mathcal{N}(0,1)$ be a standard normal random variable. Then for any $t \geqslant 0$*

$$\mathbb{P}(\eta > t) \leqslant \frac{1}{2}e^{-\frac{t^2}{2}}.$$

Proof.

Let $g(t)$ denote $\mathbb{P}(\eta > t)$, $t \geqslant 0$, then $g(t) = \int\limits_{x=t}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$. Consider function $h(t) = \frac{1}{2}e^{-\frac{t^2}{2}} - g(t)$. $h(0) = \frac{1}{2} - \frac{1}{2} = 0$, and at $t \to +\infty$: $h(t) \to 0$. Let us calculate the derivative of $h(t)$:

$$h'(t) = \left(\frac{1}{\sqrt{2\pi}} - \frac{t}{2}\right) e^{-\frac{t^2}{2}}.$$

We see that for $t \in \left[0, \frac{\sqrt{2}}{\sqrt{\pi}}\right]$: $h'(t) \geqslant 0$, so for $t \in \left[0, \frac{\sqrt{2}}{\sqrt{\pi}}\right]$: $h(t) \geqslant h(0) = 0$.

For $t \geqslant \frac{\sqrt{2}}{\sqrt{\pi}}$: $h'(t) \leqslant 0$, so $t \geqslant \frac{\sqrt{2}}{\sqrt{\pi}}$: $h(t) \geqslant h(+\infty) = 0$.

We proved that $\frac{1}{2}e^{-\frac{t^2}{2}} - g(t) = h(t) \geqslant 0$ for all $t \geqslant 0$, that means that $\mathbb{P}(\eta > t) = g(t) \leqslant e^{-\frac{t^2}{2}}$.

□

Lemma 24. *For any $r > 0$ and $s > 0$:*

$$\left(\sum_{i=1}^{n} |a_i|^r |b_i|^s\right)^{r+s} \leqslant \left(\sum_{i=1}^{n} |a_i|^{r+s}\right)^{r} \left(\sum_{i=1}^{n} |b_i|^{r+s}\right)^{s}.$$

Proof.

Set $x_i = |a_i|^r$ and $y_i = |b_i|^s$. Then the statement of the Lemma becomes:

$$\left(\sum_{i=1}^{n} x_i y_i\right)^{r+s} \leqslant \left(\sum_{i=1}^{n} x_i^{\frac{r+s}{r}}\right)^{r} \left(\sum_{i=1}^{n} y_i^{\frac{r+s}{s}}\right)^{s},$$

which is the same as the following inequality:

$$\sum_{i=1}^{n} x_i y_i \leqslant \left(\sum_{i=1}^{n} x_i^{\frac{r+s}{r}}\right)^{\frac{r}{r+s}} \left(\sum_{i=1}^{n} y_i^{\frac{r+s}{s}}\right)^{\frac{s}{r+s}},$$

which is equal to Holder inequality for $p = \frac{r+s}{r}$ and $q = \frac{r+s}{s}$.

Lemma 25. *For any function f that is M_2-Lipschitz w.r.t. 2-norm, it holds that if e is distributed uniformly on the unit Euclidean sphere, then for some numerical constant $c < 15000$, that does not depend on f and n, the following inequation holds:*

$$\sqrt{\mathbb{E}\left[(f(e) - \mathbb{E}[f(e)])^4\right]} \leqslant \sqrt[4]{\mathbb{E}\left[(f(e) - \mathbb{E}[f(e)])^8\right]} \leqslant \frac{cM_2^2}{n}.$$

Proof.

The left inequation in this Lemma follows from $\mathbb{E}[A] \leqslant \sqrt{\mathbb{E}[A^2]}$. For the right inequation we use the proposition 2.10 in [38]) as a result of the concentration of Lipschitz functions on Euclidean sphere:

$$\mathbb{P}\left(|f(e) - \mathbb{E}f(e)| > t\right) \leqslant 2\exp\left(-\frac{m_n t^2}{32 \cdot 9\pi^2 M_2^2}\right) + 4\exp\left(-\frac{m_n t^2}{8 \cdot 9\pi^2 M_2^2}\right),$$

where m_n is the median of chi-squared distribution χ_n^2, i.e. $\mathbb{P}(\chi_n^2 \leqslant m_n) = \frac{1}{2}$. By calculation one can obtain that for $n \geqslant 2$ the median $m_n \geqslant \frac{n}{2}$. Thus,

$$\mathbb{P}\left(|f(e) - \mathbb{E}f(e)| > t\right) \leqslant 2\exp\left(-\frac{nt^2}{576\pi^2 M_2^2}\right) + 4\exp\left(-\frac{nt^2}{144\pi^2 M_2^2}\right).$$

Exploiting formula $\mathbb{E}[X] = \int\limits_{t=0}^{+\infty} \mathbb{P}(X > t)dt$ for $X \geqslant 0$, we obtain:

$$\begin{aligned}
\mathbb{E}\left[(f(e) - \mathbb{E}[f(e)])^8\right] &= \int\limits_{t=0}^{+\infty} \mathbb{P}\left(|f(e) - \mathbb{E}f(e)|^8 > t\right) dt = \int\limits_{t=0}^{+\infty} \mathbb{P}\left(|f(e) - \mathbb{E}f(e)| > t^{1/8}\right) dt \\
&\leqslant \int\limits_{t=0}^{+\infty} \left(2\exp\left(-\frac{nt^{1/4}}{576\pi^2 M_2^2}\right) + 4\exp\left(-\frac{nt^{1/4}}{144\pi^2 M_2^2}\right)\right) dt \\
&= 48\left(\frac{576\pi^2 M_2^2}{n}\right)^4 + 96\left(\frac{144\pi^2 M_2^2}{n}\right)^4 = \left(\frac{288\sqrt[4]{774}\pi^2 M_2^2}{n}\right)^4,
\end{aligned}$$

where in the last equation the integral calculation $\int\limits_{z=0}^{+\infty} \exp(-\sqrt[4]{z})dz = 24$ was used.

Taking fourth root of the both sides and taking $c = 288\sqrt[4]{774}\pi^2 < 15000$, we obtain for $n \geqslant 2$:

$$\sqrt{\mathbb{E}\left[(f(e) - \mathbb{E}[f(e)])^4\right]} \leqslant \frac{cM_2^2}{n}.$$

For $n = 1$: $e = +1$ or $e = -1$. Define $m = \frac{f(1)+f(-1)}{2}$ and $a = f(1) - m$. Then $f(1) = m + a$ and $f(-1) = m - a$, $|f(e) - \mathbb{E}[f(e)]| = |f(e) - m| = |a|$. As f is M_2-Lipschitz, then $|a| \leqslant M_2$, consequently:

$$\sqrt[4]{\mathbb{E}\left[f(e) - \mathbb{E}[f(e)]\right]^8} = a^2 \leqslant M_2^2 = \frac{1 \cdot M_2^2}{1} \leqslant \frac{cM_2^2}{n}.$$

□

Lemma 26. *For $n \geqslant 2$ and $\kappa(n)$ defined as*

$$\kappa(n) = \begin{cases} \frac{2}{\pi}, & n \text{ is even}, \\ 1, & n \text{ is odd}, \end{cases}$$

it holds that

$$\kappa(n)\frac{n!!}{(n-1)!!} < \sqrt{n}.$$

Proof. We need to prove that a_n defined from

$$a_n = \frac{\kappa(n)}{\sqrt{n}} \cdot \frac{n!!}{(n-1)!!}$$

is less than 1 for any $n \geqslant 2$.

$a_2 = \frac{2}{\pi\sqrt{2}} \cdot \frac{2}{1} = \frac{2\sqrt{2}}{\pi} < 1$ and $a_3 = \frac{1}{\sqrt{3}} \cdot \frac{3}{2} = \frac{\sqrt{3}}{2} < 1$.

It holds that $a_{n+2} < a_n$ for any n:

$$a_{n+2} = \frac{\kappa(n+2)}{\sqrt{n+2}} \cdot \frac{n+2!!}{(n+1)!!} = \frac{\kappa(n)}{\sqrt{n+2}} \cdot \frac{n!!}{(n-1)!!} \cdot \frac{n+2}{(n+1)} = \frac{a_n\sqrt{n}\sqrt{n+2}}{n+1} < a_n,$$

in the last inequation we used $n(n+2) < (n+1)^2$. Consequently, for even n:

$$a_n \leqslant a_2 = \frac{2}{\pi\sqrt{2}} \cdot \frac{2}{1} = \frac{2\sqrt{2}}{\pi} < 1,$$

and for odd n:

$$a_n \leqslant a_3 = \frac{1}{\sqrt{3}} \cdot \frac{3}{2} = \frac{\sqrt{3}}{2} < 1.$$

□

Appendix B

Appendix for Chapter 1

B.1 Proof of Theorem 3

Theorem 3. *Let $f \in \mathcal{F}_{\mu,\beta}(L_\beta)$ with μ, $L_\beta > 0$ and $\beta > 2$. Let Assumption 1 hold and let Q be a convex compact subset of $\mathbb{R}^n$. Let f be M_2-Lipschitz on the Euclidean τ_1-neighborhood of Q.*

Then the optimization error of averaged estimator $\overline{x}_N = \frac{1}{N}\sum_{k=1}^{N} x_k$ where the points x_k are given by Algorithm 1 with parameters

$$\tau_k = \left(\frac{3\kappa\Delta^2 n}{2(\beta-1)(\kappa_\beta L_\beta)^2}\right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}, \quad \alpha_k = \frac{2}{\mu k}, \quad k = 1, \dots, N$$

satisfies

$$\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leqslant \frac{1}{\mu}\left(n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N}\right),$$

where $A_1 = 3\beta(\kappa\Delta^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L_\beta)^{\frac{2}{\beta}}$, $A_2 = 3c\kappa M_2^2$, κ_β and κ are constants depending only on β, see (1.5) and (1.6), $c < 15000$.

Step 1.

Fix an arbitrary $x \in Q$. As x_{k+1} is the Euclidean projection we have $\|x_{k+1} - x\|_2^2 \leqslant \|x_k - \alpha_k \widetilde{g}_k - x\|_2^2$ which is equivalent to

$$\langle \widetilde{g}_k, x_k - x\rangle \leqslant \frac{\|x_k - x\|_2^2 - \|x_{k+1} - x\|_2^2}{2\alpha_k} + \frac{\alpha_k}{2}\|\widetilde{g}_k\|_2^2. \tag{B.1}$$

By the strong convexity assumption we have

$$f(x_k) - f(x) \leqslant \langle \nabla f(x_k), x_k - x\rangle - \frac{\mu}{2}\|x_k - x\|_2^2. \tag{B.2}$$

Combining the last two inequations we obtain

$$\begin{aligned} f(x_k) - f(x) \leqslant & \langle \nabla f(x_k) - \widetilde{g}_k, x_k - x\rangle + \frac{\|x_k - x\|_2^2 - \|x_{k+1} - x\|_2^2}{2\alpha_k} \\ & + \frac{\alpha_k}{2}\|\widetilde{g}_k\|_2^2 - \frac{\mu}{2}\|x_k - x\|_2^2. \end{aligned} \tag{B.3}$$

Taking conditional expectation given x_k with respect to r_k, ξ_k and e_k we obtain

$$\begin{aligned} f(x_k) - f(x) \leqslant & \langle \nabla f(x_k) - \mathbb{E}\left[\widetilde{g}_k | x_k\right], x_k - x\rangle + \frac{\alpha_k}{2}\mathbb{E}\left[\|\widetilde{g}_k\|_2^2 | x_k\right] \\ & + \frac{\|x_k - x\|_2^2 - \mathbb{E}\left[\|x_{k+1} - x\|_2^2 | x_k\right]}{2\alpha_k} - \frac{\mu}{2}\|x_k - x\|_2^2. \end{aligned} \tag{B.4}$$

Step 2 (Bounding bias term)

Our aim is to bound the first term in (B.4), namely $\langle \nabla f(x_k) - \mathbb{E}\left[\widetilde{g}_k | x_k\right], x_k - x\rangle$. Using the Taylor expansion we have

$$\begin{aligned} f\left(x_k + \tau_k r_k e_k\right) =& f(x_k) + \langle \nabla f(x_k), \tau_k r_k e_k\rangle \\ &+ \sum_{2 \leqslant |m| \leqslant l} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)} f(x_k) e_k^m + R(\tau_k r_k e_k), \end{aligned} \tag{B.5}$$

where from Higher order smoothness of f it holds that $|R(\tau_k r_k e_k)| \leqslant L_\beta \|\tau_k r_k e_k\|_2^\beta = L_\beta (\tau_k \cdot |r_k|)^\beta$. Thus,

$$\begin{aligned} \widetilde{g}_k =& \Big(\langle \nabla f(x_k), \tau_k r_k e_k\rangle + \sum_{2 \leqslant |m| \leqslant l, |m| \text{ odd}} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)} f(x_k) e_k^m \\ &+ \frac{1}{2} R(\tau_k r_k e_k) - \frac{1}{2} R(-\tau_k r_k e_k) + \dot{\xi}_k - \ddot{\xi}_k \Big) \frac{n}{\tau_k} K(r_k) e_k. \end{aligned} \tag{B.6}$$

Using the properties of the smoothing kernel K, independence of e_k and r_k (Assumption 1) and the fact that $\mathbb{E}\left[e_k e_k^T\right] = \frac{1}{n}\mathbb{I}_{n\times n}$ we obtain

$$\mathbb{E}_{e_k, r_k}\left[\langle \nabla f(x_k), \tau_k r_k e_k\rangle \frac{n}{\tau_k} K(r_k) e_k \Big| x_k \right] = \nabla f(x_k). \tag{B.7}$$

Using the fact that $\mathbb{E}\left[r_k^{|m|} K(r_k)\right] = 0$ if $2 \leqslant |m| \leqslant l$ or $|m| = 0$, independence between noises $\dot{\xi}_k$, $\ddot{\xi}_k$ and the pair (e_k, r_k) from Assumption 1, and the fact that $\mathbb{E}[e_k] = 0$, we have

$$\mathbb{E}_{e_k, r_k, \xi_k}\left[\Big(\sum_{2 \leqslant |m| \leqslant l, |m| \text{ odd}} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)} f(x_k) e_k^m + \dot{\xi}_k - \ddot{\xi}_k \Big) \frac{n}{\tau_k} K(r_k) e_k \right] = 0. \tag{B.8}$$

Combining (B.6), (B.7) and (B.8) and using the definition of κ_β we obtain

$$\begin{aligned} &|\langle \nabla f(x_k) - \mathbb{E}_{e_k, r_k, \xi_k}\left[\widetilde{g}_k | x_k\right], x_k - x\rangle| = \\ &\quad = \left| \mathbb{E}_{e_k, r_k, \xi_k}\left[\left(\frac{1}{2} R(\tau_k r_k e_k) - \frac{1}{2} R(-\tau_k r_k e_k) \right) \frac{n}{\tau_k} K(r_k) \langle e_k, x_k - x\rangle \Big| x_k \right] \right| \\ &\quad \leqslant L_\beta \tau_k^{\beta - 1} \cdot \mathbb{E}_{r_k}\left[|r_k|^\beta K(r_k) \right] \cdot n \left| \mathbb{E}_{e_k}\left[\langle e_k, x_k - x\rangle | x_k \right] \right| \\ &\quad \leqslant \kappa_\beta L_\beta \sqrt{n} \tau_k^{\beta - 1} \|x_k - x\|_2, \end{aligned} \tag{B.9}$$

where in the last inequality the fact that $\left|\mathbb{E}_e\left[\langle e, s\rangle\right]\right|^2 \leqslant \mathbb{E}_e\left[\langle e, s\rangle^2\right] = \frac{\|s\|_2^2}{n}$ was used (the fact from concentration measure theory). Applying the inequality $ab \leqslant 1/2(a^2 + b^2)$ to the last expression in (B.9) we finally get

$$|\langle \nabla f(x_k) - \mathbb{E}_{e_k, r_k, \xi_k}\left[\widetilde{g}_k | x_k\right], x_k - x\rangle| \leqslant \frac{(\kappa_\beta L_\beta)^2}{\mu} n \tau_k^{2(\beta - 1)} + \frac{\mu}{4} \|x_k - x\|_2^2. \tag{B.10}$$

Step 3 (Bounding second moment of gradient estimator)

Our aim is to estimate $\mathbb{E}\left[\|\widetilde{g}_k\|_2^2|x_k\right]$ which is the second term in (B.4). The expectation here is with respect to r_k, $\dot{\xi}_k$ and $\ddot{\xi}_k$. To lighten the presentation and withous loss of generality we drop the lower script k in all quantities.

We have

$$\begin{aligned}\|\widetilde{g}\|_2^2 =&\frac{n^2}{4\tau^2}\|(f(x+\tau re)-f(x-\tau re)+\dot{\xi}-\ddot{\xi})K(r)e\|_2^2\\ =&\frac{n^2}{4\tau^2}\left((f(x+\tau re)-f(x-\tau re)+\dot{\xi}-\ddot{\xi})\right)^2K^2(r).\end{aligned} \tag{B.11}$$

Using the inequality $(a+b+c)^2\leqslant 3(a^2+b^2+c^2)$ and Assumption 1 we get

$$\mathbb{E}_{e,r,\xi}\left[\|\widetilde{g}\|_2^2|x\right]\leqslant\frac{3n^2}{4\tau^2}\left(\mathbb{E}\left[(f(x+\tau re)-f(x-\tau re))^2K^2(r)\big|x\right]+2\kappa\Delta^2\right). \tag{B.12}$$

Lemma 25 states that for any function f which is M_2-Lipschitz with respect to 2-norm, it holds that if e is uniformly distributed on the Euclidean unit sphere, then

$$\sqrt{\mathbb{E}_e\left[(f(e)-\mathbb{E}[f(e)])^4\right]}\leqslant\frac{cM_2^2}{n}, \tag{B.13}$$

where $c<15000$ is a positive numerical constant.

Using (B.13), symmetry of Euclidean unit sphere and the inequality $(a+b)^2\leqslant 2(a^2+b^2)$ we obtain (all the expectations are given x):

$$\begin{aligned}\mathbb{E}\left[(f(x+e)-f(x-e))^2\big|x\right]&=\mathbb{E}_e\left[(f(x+e)-f(x-e))^2\right]\\ &\leqslant\mathbb{E}_e\left[((f(x+e)-\mathbb{E}_e[f(x+e)])-(f(x-e)-\mathbb{E}_e[f(x-e)]))^2\right]\\ &\leqslant 2\mathbb{E}_e\left[(f(x+e)-\mathbb{E}_e[f(x+e)])^2\right]+2\mathbb{E}_e\left[(f(x-e)-\mathbb{E}_e[f(x-e)])^2\right]\\ &\leqslant 2\sqrt{\mathbb{E}_e\left[(f(x+e)-\mathbb{E}_e[f(x+e)])^4\right]}+2\sqrt{\mathbb{E}_e\left[(f(x-e)-\mathbb{E}_e[f(x-e)])^4\right]}\\ &\leqslant\frac{4cM_2^2}{n},\end{aligned} \tag{B.14}$$

so we have

$$\mathbb{E}_e\left[(f(x+\tau re)-f(x-\tau re))^2\big|x\right]\leqslant\frac{4c(\tau r)^2M_2^2}{n}\leqslant\frac{4c\tau^2M_2^2}{n}. \tag{B.15}$$

By substituting (B.15) into (B.12), using independence of e and r and returning the lower script k we finally get

$$\mathbb{E}_{e_k,r_k,\xi_k}\left[\|\widetilde{g}_k\|_2^2|x\right]\leqslant\kappa\left(3cnM_2^2+\frac{3(n\Delta)^2}{2\tau_k^2}\right). \tag{B.16}$$

Step 4

Let ρ_k^2 denote full expectation $\mathbb{E}[\|x_k - x\|_2^2]$. Substituting (B.10) and (B.16) into (B.4), taking full expectation and summing over k we obtain

$$\begin{aligned}\sum_{k=1}^{N} \mathbb{E}[f(x_k) - f(x)] \leqslant & \sum_{k=1}^{N} \left(\frac{(\kappa_\beta L_\beta)^2}{\mu} n \tau_k^{2(\beta-1)} + \frac{\alpha_k}{2} \kappa \left(3cnM_2^2 + \frac{3(n\Delta)^2}{2\tau_k^2} \right) \right) \\ & + \sum_{k=1}^{N} \left(\frac{\rho_k^2 - \rho_{k+1}^2}{2\alpha_k} - \left(\frac{\mu}{2} - \frac{\mu}{4} \right) \rho_k^2 \right).\end{aligned} \tag{B.17}$$

Let $\rho_{N+1}^2 = 0$. Then setting $\alpha_k = \dfrac{2}{\mu k}$ yields

$$\begin{aligned}\sum_{k=1}^{N} \left(\frac{\rho_k^2 - \rho_{k+1}^2}{2\alpha_k} - \frac{\mu}{4} \rho_k^2 \right) &\leqslant \rho_1^2 \left(\frac{1}{2\alpha_1} - \frac{\mu}{4} \right) + \sum_{k=2}^{N+1} \rho_k^2 \left(\frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}} - \frac{\mu}{4} \right) \\ &= \rho_1^2 \left(\frac{\mu}{4} - \frac{\mu}{4} \right) + \sum_{k=2}^{N+1} \rho_k^2 \left(\frac{\mu}{4} - \frac{\mu}{4} \right) = 0.\end{aligned} \tag{B.18}$$

Substituting (B.18) into (B.17) with $\alpha_k = \frac{2}{\mu k}$ we obtain

$$\begin{aligned}\sum_{k=1}^{N} \mathbb{E}[f(x_k) - f(x)] &\leqslant \frac{1}{\mu} \sum_{k=1}^{N} \left((\kappa_\beta L_\beta)^2 n \tau_k^{2(\beta-1)} + \kappa \left(3cnM_2^2 + \frac{3(n\Delta)^2}{2\tau_k^2} \right) \frac{1}{k} \right) \\ &= \frac{1}{\mu} \sum_{k=1}^{N} \left(\left[n \cdot (\kappa_\beta L_\beta)^2 \tau_k^{2(\beta-1)} + n^2 \cdot \frac{3\kappa\Delta^2}{2k\tau_k^2} \right] + \frac{3c\kappa n M_2^2}{k} \right).\end{aligned} \tag{B.19}$$

If $\Delta > 0$ then $\tau_k = \left(\dfrac{3\kappa\Delta^2 n}{2(\beta-1)(\kappa_\beta L_\beta)^2} \right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}$ is the minimizer of square brackets. Plugging this τ_k in (B.19) and using two inequalities: for the expression in square brackets $\sum\limits_{k=1}^{N} k^{-1+1/\beta} \leqslant \beta N^{1/\beta}$ (if $\beta > 2$) and for the term after square brackets $\sum\limits_{k=1}^{N} \frac{1}{k} \leqslant 1 + \ln N$ we get

$$\sum_{k=1}^{N} \mathbb{E}[f(x_k) - f(x)] \leqslant \frac{1}{\mu} \left(n^{2-\frac{1}{\beta}} A_1 N^{\frac{1}{\beta}} + A_2 n(1 + \ln N) \right) \tag{B.20}$$

with A_1 and A_2 from the formulation of Theorem 3. Due to the convexity of f we finally prove the theorem

$$\mathbb{E}\left[f(\overline{x}_N) - f(x^*) \right] \leqslant \frac{1}{\mu} \left(n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1 + \ln N)}{N} \right). \tag{B.21}$$

□

B.2 Proof of Theorem 4

Theorem 4. *Let $f \in \mathcal{F}_\beta(L_\beta)$ with μ, $L_\beta > 0$ and $\beta > 2$. Let Assumption 1 hold and let Q be a convex compact subset of $\mathbb{R}^n$. Let f be M_2-Lipschitz on the Euclidean τ_1-neighborhood of Q. Let $\overline{x}_N$ denote $\frac{1}{N}\sum\limits_{k=1}^{N} x_k$.*

Then we achieve the optimization error $\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leqslant \varepsilon$ after $N(\varepsilon)$ steps of Algorithm 1 with settings from Theorem 3 for the regularized function: $f_\mu(x) := f(x) + \frac{\mu}{2}\|x - x_0\|_2^2$, where $\mu \leqslant \frac{\varepsilon}{R^2}$, $R = \|x_0 - x^*\|_2$, $x_0 \in Q$ – arbitrary point.

$$N(\varepsilon) = \max\left\{\left(R\sqrt{2A_1}\right)^{\frac{2\beta}{\beta-1}} \frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \left(R\sqrt{2c'A_2}\right)^{2(1+\rho)} \frac{n^{1+\rho}}{\varepsilon^{2(1+\rho)}}\right\},$$

where $A_1 = 3\beta(\kappa\Delta^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L_\beta)^{\frac{2}{\beta}}$, $A_2 = 3c\kappa M_2^2$ – constants from Theorem 3, $\rho > 0$ – arbitrarily small positive number.

Step 1

Let x^* and x_μ^* denote $\arg\min\limits_{x\in Q} f(x)$ and $\arg\min\limits_{x\in Q} f_\mu(x)$ respectively. Setting $\mu = \frac{\varepsilon}{R^2}$ and using the inequality $f_\mu(x_\mu^*) \leqslant f_\mu(x^*)$ we obtain

$$\begin{aligned} f(\overline{x}_N) - f(x^*) &= f_\mu(\overline{x}_N) - f_\mu(x^*) - \frac{\mu}{2}\|\overline{x}_N - x_0\|_2^2 + \frac{\mu}{2}\|x^* - x_0\|_2^2 \\ &\leqslant f_\mu(\overline{x}_N) - f_\mu(x^*) + \frac{\mu}{2}\|x^* - x_0\|_2^2 \\ &\leqslant f_\mu(\overline{x}_N) - f_\mu(x_\mu^*) + \frac{\varepsilon}{2}. \end{aligned} \tag{B.22}$$

Step 2

Now we apply Theorem 3 for $f_\mu(x)$ and bound RHS by $\frac{\varepsilon}{2}$:

$$\mathbb{E}\left[f_\mu(\overline{x}_N) - f_\mu(x^*)\right] \leqslant \frac{1}{\mu}\left(n^{2-\frac{1}{\beta}}\frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2\frac{n(1+\ln N)}{N}\right) \leqslant \frac{\varepsilon}{2}. \tag{B.23}$$

The inequality (B.23) is done if ($\mu = \frac{\varepsilon}{R^2}$)

$$\max\left\{n^{2-\frac{1}{\beta}}\frac{A_1}{N^{\frac{\beta-1}{\beta}}}, A_2\frac{n(1+\ln N)}{N}\right\} \leqslant \frac{\mu\varepsilon}{2} = \frac{\varepsilon^2}{2R^2}. \tag{B.24}$$

It is true that $1 + \ln N \leqslant c' N^{\frac{\rho}{\rho+1}}$ for some $c' > 0$. So the inequality (B.24) holds if

$$N \geqslant \max\left\{\left(R\sqrt{2A_1}\right)^{\frac{2\beta}{\beta-1}} \frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \left(R\sqrt{2c'A_2}\right)^{2(1+\rho)} \frac{n^{1+\rho}}{\varepsilon^{2(1+\rho)}}\right\}. \tag{B.25}$$

The inequalities (B.22) and (B.23) yield $\mathbb{E}\left[f(\overline{x}_N) - f(x^*)\right] \leqslant \varepsilon$.

□

Appendix C

Appendix for Chapter 2

C.1 Proof of Theorem 5

Theorem 5. *Let $\varphi \in \Phi_{\mu,\beta}(L_\beta)$ with μ, $L_\beta > 0$ and $\beta > 2$. Let Assumption 1 hold and let $\mathcal{Z}$ be a convex compact subset of $\mathbb{R}^n$. Let φ be M_2-Lipschitz on the Euclidean τ_1-neighborhood of $\mathcal{Z}$ (see τ_k below).*

Then the rate of convergence is given by Algorithm 2 with parameters

$$\tau_k = \left(\frac{3\kappa\Delta^2 n}{2(\beta-1)(\kappa_\beta L_\beta)^2}\right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}, \quad \alpha_k = \frac{2}{\mu k}, \quad k = 1, \dots, N$$

satisfies

$$\begin{aligned}\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] &\leqslant \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \\ &\leqslant \frac{1}{\mu}\left(n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N}\right),\end{aligned}$$

where $\overline{z}_N = \frac{1}{N}\sum\limits_{k=1}^{N} z_k$, $A_1 = 3\beta(\kappa\Delta^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L_\beta)^{\frac{2}{\beta}}$, $A_2 = 3c\kappa M_2^2$, κ_β and κ are constants depending only on β, see (1.5) and (1.6), $c < 15000$.

Step 1

Fix an arbitrary $z \in \mathcal{Z}$. As z_{k+1} is the Euclidean projection we have $\|z_{k+1} - z\|_2^2 \leqslant \|z_k - \alpha_k \widetilde{g}_k - z\|_2^2$ which is equivalent to

$$\langle \widetilde{g}_k, z_k - z\rangle \leqslant \frac{\|z_k - z\|_2^2 - \|z_{k+1} - z\|_2^2}{2\alpha_k} + \frac{\alpha_k}{2}\|\widetilde{g}_k\|_2^2. \tag{C.1}$$

Using the strong convexity-concavity and combining x and y parts of the argument z together we have

$$\begin{aligned}\varphi(x_k, y) - \varphi(x, y_k) =& \varphi(x_k, y) - \varphi(x_k, y_k) + \varphi(x_k, y_k) - \varphi(x, y_k) \\ \leqslant& \langle -\nabla_y \varphi(x_k, y_k), y_k - y\rangle - \frac{\mu}{2}\|y_k - y\|_2^2 \\ &+ \langle -\nabla_x \varphi(x_k, y_k), x_k - x\rangle - \frac{\mu}{2}\|x_k - x\|_2^2 \\ =& \langle \widetilde{\nabla}\varphi(z_k), z_k - z\rangle - \frac{\mu}{2}\|z_k - z\|_2^2.\end{aligned} \tag{C.2}$$

Combining the last two inequations we obtain

$$\begin{aligned}\varphi(x_k, y) - \varphi(x, y_k) \leqslant &\langle \widetilde{\nabla}\varphi(z_k) - \widetilde{g_k}, z_k - z\rangle + \frac{\|z_k - z\|_2^2 - \|z_{k+1} - z\|_2^2}{2\alpha_k} \\ &+ \frac{\alpha_k}{2}\|\widetilde{g_k}\|_2^2 - \frac{\mu}{2}\|z_k - z\|_2^2.\end{aligned} \tag{C.3}$$

Taking conditional expectation given z_k with respect to r_k, e_k and ξ_k we obtain

$$\begin{aligned}\varphi(x_k, y) - \varphi(x, y_k) \leqslant &\langle \widetilde{\nabla}\varphi(z_k) - \mathbb{E}\left[\widetilde{g_k}|z_k\right], z_k - z\rangle + \frac{\alpha_k}{2}\mathbb{E}\left[\|\widetilde{g_k}\|_2^2|z_k\right] \\ &+ \frac{\|z_k - z\|_2^2 - \mathbb{E}\left[\|z_{k+1} - z\|_2^2|z_k\right]}{2\alpha_k} - \frac{\mu}{2}\|z_k - z\|_2^2.\end{aligned} \tag{C.4}$$

Step 2 (Bounding bias term)

Our aim is to bound the first term in (C.4), namely $\langle \widetilde{\nabla}\varphi(z_k) - \mathbb{E}\left[\widetilde{g_k}|z_k\right], z_k - z\rangle$. Using the Taylor expansion we have

$$\begin{aligned}\varphi\left(z_k + \tau_k r_k e_k\right) =&\varphi(z_k) + \langle \nabla\varphi(z_k), \tau_k r_k e_k\rangle \\ &+ \sum_{2\leqslant|m|\leqslant l} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)}\varphi(z_k)e_k^m + R(\tau_k r_k e_k),\end{aligned} \tag{C.5}$$

where by assumption $|R(\tau_k r_k e_k)| \leqslant L_\beta\|\tau_k r_k e_k\|_2^\beta = L_\beta(\tau_k \cdot |r_k|)^\beta$. Thus,

$$\begin{aligned}\widetilde{g_k} =&\Big(\langle \nabla\varphi(z_k), \tau_k r_k e_k\rangle + \sum_{2\leqslant|m|\leqslant l,|m|\text{ odd}} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)}\varphi(z_k)e_k^m \\ &+ \frac{1}{2}R(\tau_k r_k e_k) - \frac{1}{2}R(-\tau_k r_k e_k) + \dot{\xi}_k - \ddot{\xi}_k\Big)\frac{n}{\tau_k}K(r_k)\begin{pmatrix}(e_k)_x \\ -(e_k)_y\end{pmatrix}.\end{aligned} \tag{C.6}$$

Using the properties of the smoothing kernel K, independence of e_k and r_k (Assumption 1) and the fact that $\mathbb{E}\left[e_k e_k^T\right] = \frac{1}{n}\mathbb{I}_{n\times n}$ we obtain

$$\mathbb{E}_{e_k, r_k}\left[\langle \nabla\varphi(z_k), \tau_k r_k e_k\rangle \frac{n}{\tau_k}K(r_k)\begin{pmatrix}(e_k)_x \\ -(e_k)_y\end{pmatrix}\middle| z_k\right] = \widetilde{\nabla}\varphi(z_k). \tag{C.7}$$

Using the fact that $\mathbb{E}\left[r_k^{|m|}K(r_k)\right] = 0$ if $2 \leqslant |m| \leqslant l$ or $|m| = 0$, independence between noises $\dot{\xi}_k$, $\ddot{\xi}_k$ and the pair (e_k, r_k) from Assumption 1, and the fact that $\mathbb{E}[e_k] = 0$, we have

$$\mathbb{E}_{e_k, r_k, \xi_k}\left[\Big(\sum_{2\leqslant|m|\leqslant l,|m|\text{ odd}} \frac{(\tau_k r_k)^{|m|}}{m!} D^{(m)}\varphi(z_k)e_k^m + \dot{\xi}_k - \ddot{\xi}_k\Big)\frac{n}{\tau_k}K(r_k)\begin{pmatrix}(e_k)_x \\ -(e_k)_y\end{pmatrix}\middle| x_k\right] = 0. \tag{C.8}$$

Substituting (C.6), (C.7) and (C.8) in the first term in (C.4) and using the definition of κ_β (see (1.5)) we obtain

$$\begin{aligned}\left|\langle\widetilde{\nabla}\varphi(z_k)-\mathbb{E}_{e_k,r_k,\xi_k}\left[\widetilde{g}_k|z_k\right],z_k-z\rangle\right| &= \\ =\left|\mathbb{E}_{e_k,r_k,\xi_k}\left[\left(\frac{1}{2}R(\tau_k r_k e_k)-\frac{1}{2}R(-\tau_k r_k e_k)\right)\frac{n}{\tau_k}K(r_k)\left\langle\begin{pmatrix}(e_k)_x\\-(e_k)_y\end{pmatrix},z_k-z\right\rangle\middle|z_k\right]\right| & \\ \leqslant L_\beta\tau_k^{\beta-1}\cdot\mathbb{E}_{r_k}\left[|r_k|^\beta K(r_k)\right]\cdot n\left|\mathbb{E}_{e_k}\left[\langle e_k,z_k-z\rangle|z_k\right]\right| & \\ \leqslant \kappa_\beta L_\beta\sqrt{n}\tau_k^{\beta-1}\|z_k-z\|_2, &\end{aligned} \tag{C.9}$$

where in the last two inequalities the symmetry of Euclidean sphere and the fact from concentration measure theory that $\left|\mathbb{E}_e\left[\langle e,s\rangle\right]\right|^2\leqslant\mathbb{E}_e\left[\langle e,s\rangle^2\right]=\frac{\|s\|_2^2}{n}$ were used. Applying the inequality $ab\leqslant 1/2(a^2+b^2)$ to the last expression in (C.9) we finally get

$$\left|\langle\widetilde{\nabla}\varphi(z_k)-\mathbb{E}_{e_k,r_k,\xi_k}\left[\widetilde{g}_k|z_k\right],z_k-z\rangle\right|\leqslant\frac{\left(\kappa_\beta L_\beta\right)^2}{\mu}n\tau_k^{2(\beta-1)}+\frac{\mu}{4}\|z_k-z\|_2^2. \tag{C.10}$$

Step 3 (Bounding second moment of gradient estimator)

Our aim is to estimate $\mathbb{E}\left[\|\widetilde{g}_k\|_2^2|z_k\right]$ which is the second term in (C.4). The expectation here is with respect to r_k, $\dot{\xi}_k$ and $\ddot{\xi}_k$. To lighten the presentation and without loss of generality we drop the lower script k in all quantities.

We have

$$\begin{aligned}\|\widetilde{g}\|_2^2=&\frac{n^2}{4\tau^2}\left\|(\varphi(z+\tau re)-\varphi(z-\tau re)+\dot{\xi}-\ddot{\xi})K(r)\begin{pmatrix}e_x\\-e_y\end{pmatrix}\right\|_2^2 \\ =&\frac{n^2}{4\tau^2}\left((\varphi(z+\tau re)-\varphi(z-\tau re)+\dot{\xi}-\ddot{\xi})\right)^2K^2(r).\end{aligned} \tag{C.11}$$

Using the inequality $(a+b+c)^2\leqslant 3(a^2+b^2+c^2)$ and Assumption 1 we get

$$\mathbb{E}\left[\|\widetilde{g}\|_2^2|z\right]\leqslant\frac{3n^2}{4\tau^2}\left(\mathbb{E}\left[(\varphi(z+\tau re)-\varphi(z-\tau re))^2K^2(r)\middle|z\right]+2\kappa\Delta^2\right). \tag{C.12}$$

Using the symmetry of Euclidean unit sphere and the inequality $(a+b)^2\leqslant 2(a^2+b^2)$ we obtain

$$\begin{aligned}\mathbb{E}\left[(\varphi(z+e)-\varphi(z-e))^2\middle|z\right]=\mathbb{E}_e\left[(\varphi(z+e)-\varphi(z-e))^2\right] & \\ \leqslant\mathbb{E}_e\left[((\varphi(z+e)-\mathbb{E}_e[\varphi(z+e)])-(\varphi(z-e)-\mathbb{E}_e[\varphi(z-e)]))^2\right] & \\ \leqslant 2\mathbb{E}_e\left[(\varphi(z+e)-\mathbb{E}_e[\varphi(z+e)])^2\right]+2\mathbb{E}_e\left[(\varphi(z-e)-\mathbb{E}_e[\varphi(z-e)])^2\right] & \\ \leqslant 2\sqrt{\mathbb{E}_e\left[(\varphi(z+e)-\mathbb{E}_e[\varphi(z+e)])^4\right]}+2\sqrt{\mathbb{E}_e\left[(\varphi(z-e)-\mathbb{E}_e[\varphi(z-e)])^4\right]} & \\ \leqslant\frac{4cM_2^2}{n}, &\end{aligned} \tag{C.13}$$

where in the last step the concentration inequality (B.13) was used, $c < 15000$ is a numerical constant. So we have

$$\mathbb{E}\left[(\varphi(z+\tau re)-\varphi(z-\tau re))^2\big|z\right] \leqslant \frac{4c(\tau r)^2 M_2^2}{n} \leqslant \frac{4c\tau^2 M_2^2}{n}. \tag{C.14}$$

By substituting (C.14) into (C.12), using independence of e and r and returning the lower script k we finally get

$$\mathbb{E}\left[\|\widetilde{g}_k\|_2^2|z_k\right] \leqslant \kappa\left(3cnM_2^2 + \frac{3(n\Delta)^2}{2\tau_k^2}\right). \tag{C.15}$$

Step 4

Let ρ_k^2 denote full expectation $\mathbb{E}[\|z_k - z\|_2^2]$. Substituting (C.10) and (C.15) into (C.4), taking full expectation we obtain

$$\begin{aligned}\mathbb{E}[\varphi(x_k,y)-\varphi(x,y_k)] \leqslant & \frac{(\kappa_\beta L_\beta)^2}{\mu} n\tau_k^{2(\beta-1)} + \frac{\alpha_k}{2}\kappa\left(3cnM_2^2 + \frac{3(n\Delta)^2}{2\tau_k^2}\right) \\ & + \frac{\rho_k^2-\rho_{k+1}^2}{2\alpha_k} - \left(\frac{\mu}{2}-\frac{\mu}{4}\right)\rho_k^2.\end{aligned} \tag{C.16}$$

Using the convexity-concavity of φ and (C.16) we have

$$\begin{aligned}\mathbb{E}\left[\varphi\left(\overline{x}_N,y\right)-\varphi\left(x,\overline{y}_N\right)\right] &\leqslant \frac{1}{N}\sum_{k=1}^N \varphi\left(x_k,y\right) - \frac{1}{N}\sum_{k=1}^N \varphi\left(x,y_k\right) \\ &\leqslant \frac{1}{N}\sum_{k=1}^N\left(\frac{(\kappa_\beta L_\beta)^2}{\mu} n\tau_k^{2(\beta-1)} + \frac{\alpha_k}{2}\kappa\left(3cnM_2^2+\frac{3(n\Delta)^2}{2\tau_k^2}\right)\right) \\ &\quad + \frac{1}{N}\sum_{k=1}^N\left(\frac{\rho_k^2-\rho_{k+1}^2}{2\alpha_k} - \frac{\mu}{4}\rho_k^2\right).\end{aligned} \tag{C.17}$$

Let $\rho_{N+1}^2 = 0$. Then setting $\alpha_k = \dfrac{2}{\mu k}$ yields

$$\begin{aligned}\sum_{k=1}^N\left(\frac{\rho_k^2-\rho_{k+1}^2}{2\alpha_k} - \frac{\mu}{4}\rho_k^2\right) &\leqslant \rho_1^2\left(\frac{1}{2\alpha_1}-\frac{\mu}{4}\right) + \sum_{k=2}^{N+1}\rho_k^2\left(\frac{1}{2\alpha_k}-\frac{1}{2\alpha_{k-1}}-\frac{\mu}{4}\right) \\ &= \rho_1^2\left(\frac{\mu}{4}-\frac{\mu}{4}\right) + \sum_{k=2}^{N+1}\rho_k^2\left(\frac{\mu}{4}-\frac{\mu}{4}\right) = 0.\end{aligned} \tag{C.18}$$

Substituting (C.18) into (C.16) with $\alpha_k = \frac{2}{\mu k}$ we obtain

$$\begin{aligned}&\mathbb{E}[\varphi\left(\overline{x}_N,y\right)-\varphi\left(x,\overline{y}_N\right)] \\ &\qquad \leqslant \frac{1}{\mu N}\sum_{k=1}^N\left((\kappa_\beta L_\beta)^2 n\tau_k^{2(\beta-1)} + \kappa\left(3cnM_2^2+\frac{3(n\Delta)^2}{2\tau_k^2}\right)\frac{1}{k}\right) \\ &\qquad = \frac{1}{\mu N}\sum_{k=1}^N\left(\left[n\cdot(\kappa_\beta L_\beta)^2\tau_k^{2(\beta-1)} + n^2\cdot\frac{3\kappa\Delta^2}{2k\tau_k^2}\right] + \frac{3c\kappa nM_2^2}{k}\right).\end{aligned} \tag{C.19}$$

If $\Delta > 0$ then $\tau_k = \left(\dfrac{3\kappa\Delta^2 n}{2(\beta-1)(\kappa_\beta L_\beta)^2}\right)^{\frac{1}{2\beta}} k^{-\frac{1}{2\beta}}$ is the minimizer of square brackets. Plugging this τ_k in (C.19) and using two inequalities: for the expression in square brackets $\sum\limits_{k=1}^{N} k^{-1+1/\beta} \leqslant \beta N^{1/\beta}$ (if $\beta > 2$) and for the term after square brackets $\sum\limits_{k=1}^{N} \frac{1}{k} \leqslant 1 + \ln N$ we get

$$\mathbb{E}[\varphi(\overline{x}_N, y) - \varphi(x, \overline{y}_N)] \leqslant \frac{1}{\mu}\left(n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N}\right).$$

with A_1 and A_2 from the formulation of Theorem 5.

Taking the minimum over x and the maximum over y we finally obtain

$$\begin{aligned}\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] &\leqslant \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \\ &\leqslant \frac{1}{\mu}\left(n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N}\right).\end{aligned}$$

□

C.2 Proof of Theorem 6

Theorem 6. *Let $\varphi \in \Phi_\beta(L_\beta)$ with $L_\beta > 0$ and $\beta > 2$. Let Assumption 1 hold and let $\mathcal{Z}$ be a convex compact subset of $\mathbb{R}^n$. Let φ be M_2-Lipschitz on the Euclidean τ_1-neighborhood of $\mathcal{Z}$ (τ_k is parameter from Theorem 5 for the regularized function $\varphi_\mu(z)$ whose description is given below). Let $\overline{z}_N$ denote $\frac{1}{N}\sum\limits_{k=1}^{N} z_k$.*

Let us define $N(\varepsilon)$:

$$N(\varepsilon) = \max\left\{\left(R\sqrt{2A_1}\right)^{\frac{2\beta}{\beta-1}} \frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \left(R\sqrt{2c'A_2}\right)^{2(1+\rho)} \frac{n^{1+\rho}}{\varepsilon^{2(1+\rho)}}\right\},$$

where $A_1 = 3\beta(\kappa\Delta^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L_\beta)^{\frac{2}{\beta}}$, $A_2 = 3c\kappa M_2^2$ – constants from Theorem 5, $\rho > 0$ – arbitrarily small positive number, c' is a constant which depends on ρ.

Then the rate of convergence is given by the following expression:

$$\mathbb{E}\left[\varphi(\overline{x}_N, y^*) - \varphi(x^*, \overline{y}_N)\right] \leqslant \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \leqslant \varepsilon \tag{C.20}$$

after $N(\varepsilon)$ steps of Algorithm 2 with settings from Theorem 5 for the regularized function: $\varphi_\mu(z) := \varphi(z) + \frac{\mu}{2}\|x - x_0\|_2^2 - \frac{\mu}{2}\|y - y_0\|_2^2$, where $\mu \leqslant \frac{\varepsilon}{R^2}$, $R = \|z_0 - z^*\|_2$, $z_0 \in \mathcal{Z}$ – arbitrary point.

Step 1

Let $z^* = (x^*, y^*)$ and $z^*_\mu = (x^*_\mu, y^*_\mu)$ denote the solutions of the saddle-point problems for functions $\varphi(z)$ and $\varphi_\mu(z)$ respectively. Let $\mathring{\overline{x}}_N$ denote $\overline{x}_N - x_0$, $\mathring{\overline{y}}_N$ denote $\overline{y}_N - y_0$ respectively. Let $\mathring{x}$ denote $x - x_0$, $\mathring{y}$ denote $y - y_0$ and $\mathring{z}$ denote $z - z_0$, where $z = (x,y)$, $z_0 = (x_0,y_0)$ and so on.

Setting $\mu = \frac{\varepsilon}{R^2}$ and using the inequality $\varphi_\mu(\overline{x}_N, y^*) - \varphi_\mu(x^*, \overline{y}_N) \leqslant \varphi_\mu\left(\overline{x}_N, y^*_\mu\right) - \varphi_\mu\left(x^*_\mu, \overline{y}_N\right)$ we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\varphi(\overline{x}_N, y^*)\right] - \mathbb{E}\left[\varphi(x^*, \overline{y}_N)\right] \leqslant \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi(x, \overline{y}_N)\right] \\
&= \max_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathbb{E}\left[\varphi_\mu(\overline{x}_N, y) - \varphi_\mu(x, \overline{y}_N) - \frac{\mu\mathring{\overline{x}}_N^2}{2} + \frac{\mu\mathring{y}^2}{2} + \frac{\mu\mathring{x}^2}{2} - \frac{\mu\mathring{\overline{y}}_N^2}{2}\right] \\
&\leqslant \max_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathbb{E}\left[\varphi_\mu(\overline{x}_N, y) - \varphi_\mu(x, \overline{y}_N) + \frac{\mu\mathring{z}^2}{2}\right] \\
&\leqslant \max_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathbb{E}\left[\varphi_\mu(\overline{x}_N, y) - \varphi_\mu(x, \overline{y}_N)\right] + \frac{\varepsilon}{2} \\
&= \max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi_\mu(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi_\mu(x, \overline{y}_N)\right] + \frac{\varepsilon}{2}
\end{aligned}
\tag{C.21}
$$

Step 2

Now we apply Theorem 5 for $\varphi_\mu(z)$ until function error is not greater than $\frac{\varepsilon}{2}$:

$$
\max_{y\in\mathcal{Y}} \mathbb{E}\left[\varphi_\mu(\overline{x}_N, y)\right] - \min_{x\in\mathcal{X}} \mathbb{E}\left[\varphi_\mu(x, \overline{y}_N)\right] \leqslant \frac{1}{\mu}\left(n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}} + A_2 \frac{n(1+\ln N)}{N}\right) \leqslant \frac{\varepsilon}{2}. \tag{C.22}
$$

Using that $\mu = \frac{\varepsilon}{R^2}$ the inequality (C.22) is done if

$$
\max\left\{n^{2-\frac{1}{\beta}} \frac{A_1}{N^{\frac{\beta-1}{\beta}}}, A_2 \frac{n(1+\ln N)}{N}\right\} \leqslant \frac{\mu\varepsilon}{2} = \frac{\varepsilon^2}{2R^2}. \tag{C.23}
$$

It is true that $1 + \ln N \leqslant c' N^{\frac{\rho}{\rho+1}}$ for some $c' > 0$. So the inequality (C.23) holds if

$$
N \geqslant \max\left\{\left(R\sqrt{2A_1}\right)^{\frac{2\beta}{\beta-1}} \frac{n^{2+\frac{1}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}, \left(R\sqrt{2c'A_2}\right)^{2(1+\rho)} \frac{n^{1+\rho}}{\varepsilon^{2(1+\rho)}}\right\}. \tag{C.24}
$$

The inequalities (C.21) and (C.22) yield (2.2).

□

Appendix D

Appendix for Chapter 3

D.1 Lemmas

D.1.1 Proof of Lemma 7

Lemma 7. *Let distance generating function $d(x)$ be defined as follows:*

$$d(x) = \begin{cases} \frac{\|x-x_1\|_p^2}{2(p-1)}, & a \leqslant p \leqslant 2, \\ \frac{e\|x-x_1\|_a^2}{2(a-1)}, & 1 \leqslant p < a, \end{cases} \tag{D.1}$$

where $a = \frac{2\ln n}{2\ln n - 1}$. Then $d(x)$ is 1-strongly convex w.r.t p-norm on $\mathbb{R}^n$.

Proof.

For simplicity assume that starting point x_1 in the definition (D.1) of $d(x)$ is equal to zero. This does not affect the proof.

Further, x_i means i-th coordinate of the n-dimensional vector x.

Case $p \geqslant a$.

Consider strictly positive ortant $(\mathbb{R}^+)^n \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : x_i > 0 \text{ for all } i \text{ from 1 to n}\}$. The distance generating function $d(x)$ is twice continuously differentiable on $(\mathbb{R}^+)^n$. For such a function, 1-strong convexity w.r.t. p-norm holds true only and only if

$$D^2 d(x)[h,h] \geqslant \|h\|_p^2 \text{ for all } x \in (\mathbb{R}^+)^n,\ h \in \mathbb{R}^n.$$

Let us calculate the differentials:

$$Dd(x)[h] = \frac{1}{p-1}\left(\sum_{i=1}^n |x_i|^p\right)^{\frac{2}{p}-1}\sum_{i=1}^n |x_i|^{p-1}\text{sign}(x_i)h_i$$

$$D^2d(x)[h,h] = \frac{1}{p-1}\underbrace{\left(\frac{2}{p}-1\right)}_{\geqslant 0}\left(\sum_{i=1}^n |x_i|^p\right)^{\frac{2}{p}-2}\left(\sum_{i=1}^n |x_i|^{p-1}\text{sign}(x_i)h_i\right)^2$$

$$+\frac{1}{p-1}\left(\sum_{i=1}^n |x_i|^p\right)^{\frac{2}{p}-1}\sum_{i=1}^n (p-1)|x_i|^{p-2}h_i^2$$

$$\geqslant \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{2}{p}-1}\left(\sum_{i=1}^n |x_i|^{p-2}h_i^2\right). \tag{D.2}$$

Using Lemma 24 for $r = \frac{2}{p} - 1 > 0$ and $s = 1$, $\left(r + s = \frac{2}{p}\right)$; $a_i = |x_i|^{\frac{p^2}{2}}$ and $b_i = (|x_i|^{p-2}h_i^2)^{\frac{p}{2}}$:

$$\left(\sum_{i=1}^n |a_i|^r|b_i|^s\right)^{r+s} \leqslant \left(\sum_{i=1}^n |a_i|^{r+s}\right)^r\left(\sum_{i=1}^n |b_i|^{r+s}\right)^s,$$

$$\left(\sum_{i=1}^n |x_i|^{\frac{p^2 r}{2}}\left(|x_i|^{p-2}h_i^2\right)^{\frac{ps}{2}}\right)^{r+s} \leqslant \left(\sum_{i=1}^n |x_i|^{\frac{p^2(r+s)}{2}}\right)^r\left(\sum_{i=1}^n \left(|x_i|^{p-2}h_i^2\right)^{\frac{p(r+s)}{2}}\right)^s,$$

$$\left(\sum_{i=1}^n |x_i|^{\frac{p^2 r}{2}+\frac{(p-2)ps}{2}}|h_i|^{ps}\right)^{\frac{2}{p}} \leqslant \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{2}{p}-1}\left(\sum_{i=1}^n \left(|x_i|^{p-2}h_i^2\right)\right).$$

Calculating the degree of $|x_i|$ in the LHS of the last inequation:

$$\frac{p^2 r}{2} + \frac{(p-2)ps}{2} = \frac{p^2(r+s)}{2} - ps = p\cdot 1 - p\cdot 1 = 0$$

and the degree of $|h_i|$: $ps = p$, we obtain:

$$\left(\sum_{i=1}^n |h_i|^p\right)^{\frac{2}{p}} \leqslant \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{2}{p}-1}\left(\sum_{i=1}^n \left(|x_i|^{p-2}h_i^2\right)\right). \tag{D.3}$$

Substituting (D.3) into (D.2) we obtain:

$$D^2d(x)[h,h] \geqslant \left(\sum_{i=1}^n |h_i|^p\right)^{\frac{2}{p}} = \|h\|_p^2,$$

that proves 1-strong convexity of $d(x)$ w.r.t. p-norm on $(\mathbb{R}^+)^n$.

By symmetry of $d(x)$ we prove that $d(x)$ is 1-strongly convex w.r.t. p-norm on any open ortant

$$\underbrace{\mathbb{R}^{+,-}\times\mathbb{R}^{+/-}\times\cdots\times\mathbb{R}^{+/-}}_{n \text{ times}}, \tag{D.4}$$

where $\mathbb{R}^{+/-}$ means either $\mathbb{R}^+$ or $\mathbb{R}^-$ independently.

We use induction on the dimension n to prove the 1-strong convexity of $d(x)$ w.r.t. p-norm on the whole $\mathbb{R}^n$. The base of induction is $n = 1$, in this case we have $d(x) = \frac{1}{2(p-1)}x^2$, which is

$\frac{1}{p-1}$-strongly convex on $\mathbb{R}$. As $p \leqslant 2$, we have $\frac{1}{p-1} \geqslant 1$, consequently $d(x)$ is 1-strongly convex on $\mathbb{R}$. The base of induction is proved.

Now assume that for the dimension $n-1$ the distance generating function is 1-strongly convex w.r.t. p-norm.

We need to prove that

$$d(y) \geqslant d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\|y - x\|_p^2 \ \ \forall x,y \in \mathbb{R}^n. \tag{D.5}$$

For every $x \neq y$, consider the parametrization of closed interval $[x,y]$: $x_t = x + t(y - x)$, $t \in [0,1]$, and function of one-dimensional argument $f_{(x,y)}(t) = d(x_t)$, $t \in [0,1]$. Its derivative is $f'_{(x,y)}(t) = \langle \nabla d(x_t), y - x \rangle$.

For any $t_1, t_2 \in [0,1]$ the inequality (D.5) holds for x_{t_1} and x_{t_2}:

$$d(x + t_2(y - x)) \geqslant d(x + t_1(y - x)) + \langle \nabla d(x + t_1(y - x)), y - x \rangle (t_2 - t_1) + \frac{1}{2}(t_2 - t_1)^2 \|y - x\|_p^2,$$

or, in terms of $f(t)$:

$$f_{(x,y)}(t_2) \geqslant f_{(x,y)}(t_1) + f'_{(x,y)}(t_1)(t_2 - t_1) + \frac{1}{2}(t_2 - t_1)^2 \|y - x\|_p^2, \ \ \forall t_1,t_2 \in [0,1]; \forall x,y \in \mathbb{R}^n. \tag{D.6}$$

In fact, (D.6) means $\|y - x\|_p^2$-strong convexity of $f_{(x,y)}(t)$ on $t \in [0,1]$. The inequation (D.6) followed from (D.5). Moreover, if (D.6) is done, then (D.5) holds true by taking $t_1 = 0$ and $t_2 = 1$. We proved that (D.5) and (D.6) are equivalent.

Fix $x,y \in \mathbb{R}^n$, $x \neq y$. If the i-th coordinate $x_i = y_i = 0$ for some i, then $[\nabla d(x)]_i = [\nabla d(y)]_i = 0$, and (D.5) falls into $n-1$-dimensional case, for which (D.5) is done by the induction assumption.

If for all i from 1 to n either $x_i \neq 0$ or $y_i \neq 0$, we consider the parametrization of closed interval $[x,y]$: $x_t = x + t(y - x)$, $t \in [0,1]$, and function of one-dimensional argument $f(t) = d(x_t)$, $t \in [0,1]$. Its derivative is $f'(t) = \langle \nabla d(x_t), y - x \rangle$. As for all i from 1 to n either $x_i \neq 0$ or $y_i \neq 0$, then each coordinate of x_t: $[x_t]_i = x_i + t(y_i - x_i)$ can be equal to zero at maximum one point t, and the total amount of points t for which at lease one coordinate of x_t is equal to zero is less or equal to n. We conclude, that there are no more than $K+1$ intervals $\{(a_i, a_{i+1})\}_{i=0}^K$, $K \leqslant n$, such that $0 \stackrel{\text{def}}{=} a_0 < a_1 < \ldots < a_{K+1} \stackrel{\text{def}}{=} 1$, and at each open interval $t \in (a_i, a_{i+1})$ the point x_t has only non-zero coordinates and all its coordinates do not change their signs.

For each interval $t \in (a_i, a_{i+1})$ the corresponding interval $(x_{t_i}, x_{t_{i+1}})$ belongs to one of ortants (D.4), on which the distance generating function $d(x)$ is 1-strongly convex w.r.t. p-norm, that is equivalent to the $\|y - x\|_p^2$-strong convexity of $f_{(x,y)}(t)$ on (a_i, a_{i+1}):

$$f_{(x,y)}(t_2) \geqslant f_{(x,y)}(t_1) + f'_{(x,y)}(t_1)(t_2 - t_1) + \frac{1}{2}(t_2 - t_1)^2 \|y - x\|_p^2, \ \ \forall t_1,t_2 \in (a_i, a_{i+1}).$$

The function $f_{(x,y)}(t)$ of one-dimensional argument is $\|y - x\|_p^2$-strongly convex on the set T if and only if the function $g_{(x,y)}(t) = f_{(x,y)}(t) - \frac{\|y-x\|_p^2}{2} t^2$ is convex on T. Thus, $g_{(x,y)}(t)$ is convex on each open interval (a_i, a_{i+1}).

For continuously differentiable function $g_{(x,y)}(t)$ the convexity is equivalent to monotonically non-decrease of $g'_{(x,y)}(t)$. Thus, $g'_{(x,y)}(t)$ is monotonically non-decreasing on open intervals

$\{(a_i, a_{i+1})\}_{i=0}^K$. As $g'_{(x,y)}(t)$ is a continuous on [0,1] function, $g'_{(x,y)}(t)$ is monotonically non-decreasing on closing of open intervals $\{[a_i, a_{i+1}]\}_{i=0}^K$, and consequently, $g'_{(x,y)}(t)$ is monotonically non-decreasing on $[a_0, a_{K+1}] = [0,1]$, that implies convexity of $g_{(x,y)}(t)$ on [0,1]. The convexity of $g_{(x,y)}(t)$ on [0,1] is equivalent to $\|y-x\|_p^2$-strong convexity of the function $f_{(x,y)}(t)$ on the set [0,1].

We fixed any x,$y \in \mathbb{R}^n$ and proved that for these x,y the function $f_{(x,y)}(t)$ is $\|y-x\|_p^2$-strongly convex on [0,1], so we proved (D.6) which is equivalent to (D.5).

Case $p < a$.

As $d(x)$ coincides with $d(x)$ for $p = a$ up to a constant multiplier e, we obtain that $d(x)$ is e-strongly convex on w.r.t. a-norm. We need to prove, that $d(x)$ is 1-strongly convex on w.r.t. p-norm. To do this, it is sufficient to show that

$$\frac{e}{2}\|y-x\|_a^2 \geqslant \frac{1}{2}\|y-x\|_p^2. \tag{D.7}$$

We make a substitution $z = y - x$ and recall that $a = \frac{2\ln n}{2\ln n - 1}$. As $1 \leqslant p < a$, we obtain:

$$\begin{aligned}\frac{\|z\|_p^2}{\|z\|_a^2} \leqslant \left(n^{\frac{1}{p}-\frac{1}{a}}\right)^2 &= \exp\left(\frac{2(a-p)}{pa}\ln n\right) = \exp\left(\frac{2(a-p)(\ln n)(2\ln n - 1)}{2p\ln n}\right)\\ &= \exp\left(\left(\frac{a}{p}-1\right)(2\ln n - 1)\right) \geqslant \exp((a-1)(2\ln n - 1))\\ &= \exp(2\ln n - (2\ln n - 1)) = e,\end{aligned}$$

that proves (D.7) and completes the proof of this Lemma.

□

D.1.2 Proof of Lemma 8

Lemma 8. *Let distance generating function $d(x)$ be defined as follows:*

$$d(x) = \begin{cases} \frac{\|x-x_1\|_p^2}{2(p-1)}, & a \leqslant p \leqslant 2,\\ \frac{e\|x-x_1\|_a^2}{2(a-1)}, & 1 \leqslant p < a,\end{cases}$$

where $a = \frac{2\ln n}{2\ln n - 1}$. Then

$$2\max\{V(x_1, x) : \|x - x_1\|_p \leqslant 1\} \leqslant r_{p,n}^2, \tag{D.8}$$

where

$$r_{p,n}^2 \overset{def}{=} \begin{cases} \frac{1}{p-1}, & a \leqslant p \leqslant 2,\\ e(2\ln n - 1), & 1 \leqslant p < a.\end{cases}$$

Proof.

By the definition of the Bregman divergence we have $V(x_1, x) = d(x)$, thus for p such that $a \leqslant p \leqslant 2$:

$$2\max_{x\in Q}\{d(x) : \|x - x_1\|_p \leqslant 1\} = \frac{2}{2(p-1)}\max_{x\in Q}\{\|x - x_1\|_p^2 : \|x - x_1\|_p \leqslant 1\} \leqslant \frac{1}{p-1},$$

and for p such that $1 \leqslant p \leqslant a$:

$$2\max_{x\in Q}\{d(x) : \|x - x_1\|_p \leqslant 1\} = \frac{2e}{2(a-1)}\max_{x\in Q}\{\|x - x_1\|_a^2 : \|x - x_1\|_p \leqslant 1\}$$
$$\leqslant \frac{e}{a-1}\max_{x\in Q}\{\|x - x_1\|_p^2 : \|x - x_1\|_p \leqslant 1\} \leqslant \frac{e}{a-1} = \frac{e(2\ln n - 1)}{2\ln n - (2\ln n - 1)} = e(2\ln n - 1).$$

□

D.1.3 Proof of Lemma 9

Lemma 9. *Let e be a random unit vector uniformly distributed on the unit Euclidean sphere with the zero center, so $\|e\|_2 = 1$. Assume $q > 2$ and $n \geqslant 2$. Then for any number $m : 1 \leqslant m \leqslant 8$ the expectation $\mathbb{E}\|e\|_q^m \leqslant a_{q,n}^m$, where $a_{q,n}^2 = \min\{4q - 1, 5\ln n\}n^{\frac{2}{q}-1}$. For the case $q = 2$, we can take $a_{2,n}^2 = 1$.*

Proof.

For the case $q = 2$ (Euclidean case): $\|e\|_2 = 1$, thus, $\mathbb{E}\|e\|_2^m = 1$. Next, assume $q > 2$.

Step 1

As for any number $m : 1 \leqslant m \leqslant 8$ and random variable $v \geqslant 0$ from Holder inequality for $\frac{8}{m}$ and $\frac{8}{8-m}$ we obtain:

$$(\mathbb{E}v)^{\frac{8}{m}} \leqslant \left(\left(\mathbb{E}v^{8/m}\right)^{\frac{m}{8}}\left(\mathbb{E}1^{\frac{8}{8-m}}\right)^{1-\frac{m}{8}}\right)^{\frac{8}{m}} = \left(\mathbb{E}v^{8/m}\right)^{\frac{m}{8}\cdot\frac{8}{m}} = \mathbb{E}v^{8/m},$$

and for $v = \|e\|_q^m$:

$$\left(\mathbb{E}\|e\|_q^m\right)^{\frac{8}{m}} \leqslant \mathbb{E}\|e\|_q^8,$$

so it remains to prove $\left(\mathbb{E}\|e\|_q^m\right) \leqslant a_{q,n}^m$ only for $m = 8$, as for $1 \leqslant m < 8$:

$$\mathbb{E}\|e\|_q^m \leqslant \left(\mathbb{E}\|e\|_q^8\right)^{\frac{m}{8}} \leqslant \left(a_{q,n}^8\right)^{\frac{m}{8}} = a_{q,n}^m.$$

Step 2

Recall that for $a = \frac{2}{q} \leqslant 1$ the function x^a is concave and $\mathbb{E}[x^a] \leqslant (\mathbb{E}[x])^a$, so

$$\mathbb{E}[\|e\|_q^8] = \mathbb{E}\left[\left(\sum_{k=1}^n |e_k|^q\right)^{4\cdot\frac{2}{q}}\right] \leqslant \left(\mathbb{E}\left[\left(\sum_{k=1}^n |e_k|^q\right)^4\right]\right)^{\frac{2}{q}} \leqslant (n\mathbb{E}[|e_2|^q])^{\frac{2}{q}}, \tag{D.9}$$

The convexity of the function x^2 implies $\left(\frac{1}{n}\left(\sum |x_i|\right)\right)^2 \leqslant \frac{1}{n}\sum |x_i|^2$, the same is $\left(\sum |x_i|\right)^2 \leqslant n \cdot \sum |x_i|^2$. Taking squared the last inequality and again applying the convexity of x^2 we get $\left(\sum |x_i|\right)^4 \leqslant n^2 \cdot \left(\sum |x_i|^2\right)^2 \leqslant n^3 \sum |x_i|^4$. Now we can simplify (D.9):

$$\mathbb{E}[\|e\|_q^8] \leqslant \left(\mathbb{E}\left[\left(\sum_{k=1}^n |e_k|^q\right)^4\right]\right)^{\frac{2}{q}} \leqslant \left(n^3\sum_{k=1}^n \mathbb{E}[|e_k|^{4q}]\right)^{\frac{2}{q}} = \left(n^4\mathbb{E}[|e_1|^{4q}]\right)^{\frac{2}{q}}, \tag{D.10}$$

where in the last equation we used the symmetry of the distribution of e.

The distribution of e on Euclidean sphere coincides with the distribution of $\frac{X}{\|X\|_2}$, where $X \sim \mathcal{N}(0,I_n)$ is a standard Gaussian random vector.

The exact calculation and upper bound for $\mathbb{E}|e_1|^\alpha$ were done in [50]:

$$\mathbb{E}|e_1|^\alpha = \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2})\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{\alpha+n}{2})} \leqslant \left(\frac{\alpha-1}{n}\right)^{\frac{\alpha}{2}},$$

where $\Gamma(x)$ is a gamma-function. Using the last bound we can calculate (D.10):

$$\mathbb{E}[\|e\|_q^8] \leqslant \left(n^4\mathbb{E}[|e_1|^{4q}]\right)^{\frac{2}{q}} \leqslant n^{\frac{8}{q}}\left(\frac{4q-1}{n}\right)^{2q\cdot\frac{2}{q}} = n^{\frac{8}{q}-4}(4q-1)^4,$$

which is the same as:

$$\left(\mathbb{E}[\|e\|_q^8]\right)^{\frac{1}{4}} \leqslant (4q-1)n^{\frac{2}{q}-1} \tag{D.11}$$

Step 3

We are going to prove the bound for $q = \infty$. For a standard normal vector $u \sim \mathcal{N}(0, I_n)$ and any $\lambda > 0$ it holds that

$$\begin{aligned}\mathbb{P}\left(\|u\|_2^2 < \frac{n}{2}\right) &= \mathbb{P}\left(e^{-\frac{\lambda}{2}\|u\|_2^2} > e^{-\frac{\lambda n}{4}}\right) \leqslant e^{\frac{\lambda n}{4}}\mathbb{E}e^{-\frac{\lambda}{2}\|u\|_2^2} \\ &= e^{\frac{\lambda n}{4}}\int\limits_{\mathbb{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{\lambda}{2}\|x\|_2^2 - \frac{1}{2}\|x\|_2^2} dx = e^{\frac{\lambda n}{4}}(1+\lambda)^{-\frac{n}{2}} = e^{\frac{\lambda n}{4} - \frac{n}{2}\ln(1+\lambda)},\end{aligned} \tag{D.12}$$

where Markov's inequality $\mathbb{P}(X > a) \leqslant \frac{\mathbb{E}X}{a}$ for positive random variable $e^{-\frac{\lambda}{2}\|u\|_2^2}$ was applied. Minimizing (D.12) over $\lambda > 0$, we have that the minimum is obtained at $\lambda = 1$ and

$$\mathbb{P}\left(\|u\|_2^2 < \frac{n}{2}\right) \leqslant e^{\left(\frac{1}{4} - \frac{\ln 2}{2}\right)n}. \tag{D.13}$$

We recall the fact that the distribution of e on the unit Euclidean is identical to $\frac{u}{\|u\|_2}$, where u is a standard Gaussian vector, so using the Law of total probability and the fact from Lemma 22 that $\|\cdot\|_\infty \leqslant \|\cdot\|_2$ we obtain:

$$\begin{aligned}
\mathbb{E}\left[\|e\|_\infty^8\right] &= \mathbb{E}\left[\frac{\|u\|_\infty^8}{\|u\|_2^8}\right] = \mathbb{P}\left(\|u\|_2^2 \leqslant \frac{n}{2}\right)\mathbb{E}\left[\frac{\|u\|_\infty^8}{\|u\|_2^8}\middle|\|u\|_2^2 \leqslant \frac{n}{2}\right] + \mathbb{P}\left(\|u\|_2^2 > \frac{n}{2}\right)\mathbb{E}\left[\frac{\|u\|_\infty^8}{\|u\|_2^8}\middle|\|u\|_2^2 > \frac{n}{2}\right] \\
&\leqslant e^{\left(\frac{1}{4} - \frac{\ln 2}{2}\right)n} \cdot 1 + \mathbb{P}\left(\|u\|_2^2 > \frac{n}{2}\right) \cdot \mathbb{E}\left[\frac{\|u\|_\infty^8}{\left(\frac{n}{2}\right)^4}\middle|\|u\|_2^2 > \frac{n}{2}\right] \\
&= e^{\left(\frac{1}{4} - \frac{\ln 2}{2}\right)n} + \left(\frac{2}{n}\right)^4 \mathbb{P}\left(\|u\|_2^2 > \frac{n}{2}\right)\mathbb{E}\left[\|u\|_\infty^8\middle|\|u\|_2^2 > \frac{n}{2}\right] \\
&\leqslant e^{\left(\frac{1}{4} - \frac{\ln 2}{2}\right)n} + \left(\frac{2}{n}\right)^4 \mathbb{E}\left[\|u\|_\infty^8\right],
\end{aligned} \tag{D.14}$$

where in the last inequality we used the formula for full expectation of non-negative function $f(x)$ via conditionals: $\mathbb{E}[f(x)|A]\mathbb{P}(A) \leqslant \mathbb{E}[f(x)|A]\mathbb{P}(A) + \mathbb{E}[f(x)|\bar{A}]\mathbb{P}(\bar{A}) = \mathbb{E}[f(x)]$.

Using the fact that all the coordinates u_k of standard normal vector u are independent, the Bernoulli inequation ($(1+x)^n \geqslant 1 + xn$, for $x \geqslant -1$) and Lemma 23 for the distribution of tails of coordinates u_k we obtain:

$$\begin{aligned}
\mathbb{P}(\|u\|_\infty \leqslant t) = \mathbb{P}\left(\bigcap_{k=1}^n \{|u_k| \leqslant t\}\right) &= \prod_{k=1}^n \mathbb{P}(|u_k| \leqslant t) = (1 - \mathbb{P}(|u_1| > t))^n \\
&\geqslant 1 - n\mathbb{P}(|u_1| > t) = 1 - 2n\mathbb{P}(u_1 > t) \geqslant 1 - ne^{-\frac{t^2}{2}},
\end{aligned}$$

which is the same as $\mathbb{P}(\|u\|_\infty > t) \leqslant ne^{-\frac{t^2}{2}}$.

Now we are ready to calculate $\mathbb{E}\|u\|_\infty^8$:

$$\begin{aligned}
\mathbb{E}\|u\|_\infty^8 &= \int_0^{+\infty} \mathbb{P}(\|u\|_\infty^8 > t)dt = \int_0^{+\infty} \mathbb{P}(\|u\|_\infty > t^{\frac{1}{8}})dt \leqslant r^4 + \int_{r^4}^{+\infty} \mathbb{P}(\|u\|_\infty > t^{\frac{1}{8}})dt \\
&\leqslant r^4 + \int_{r^4}^{+\infty} n\exp\left(-\frac{t^{1/4}}{2}\right)dt = r^4 + \int_r^{+\infty} 4n\exp\left(-\frac{z}{2}\right)z^3 dz \\
&= r^4 + 32ne^{-\frac{r}{2}}\left(r^3 + 6r^2 + 24r + 48\right).
\end{aligned}$$

Taking $r = 2\ln n$, we obtain:

$$\begin{aligned}
\mathbb{E}\|u\|_\infty^8 &\leqslant \ln^4 n\left(16 + 256\ln^{-1} n + 768\ln^{-2} n + 1536\ln^{-3} n + 1536\ln^{-4} n\right) \\
&\leqslant \ln^4 n\left(16 + 256\ln^{-1} 2 + 768\ln^{-2} 2 + 1536\ln^{-3} 2 + 1536\ln^{-4} 2\right) \\
&< 11\ln^4 n.
\end{aligned}$$

Applying the last bound for (D.14) we get:

$$\mathbb{E}\left[\|e\|_\infty^8\right] \leqslant e^{\left(\frac{1}{4}-\frac{\ln 2}{2}\right)n} + \left(\frac{2}{n}\right)^4 \mathbb{E}\left[\|u\|_\infty^8\right] \leqslant e^{\left(\frac{1}{4}-\frac{\ln 2}{2}\right)n} + \frac{176\ln^4 n}{n^4}$$
$$= \frac{\ln^4 n}{n^4}\left(176 + \frac{n^4 e^{\left(\frac{1}{4}-\frac{\ln 2}{2}\right)n}}{\ln^4 n}\right) \leqslant \frac{\ln^4 n}{n^4}\left(176+350\right) < \left(\frac{5\ln n}{n}\right)^4. \tag{D.15}$$

Step 4

We bring all the bounds together. Using Lemma 22 and (D.15) we obtain

$$\left(\mathbb{E}\left[\|e\|_q^8\right]\right)^{\frac{1}{4}} \leqslant \left(n^{\frac{8}{q}}\mathbb{E}\left[\|e\|_\infty^8\right]\right)^{\frac{1}{4}} \leqslant (5\ln n)n^{\frac{2}{q}-1}. \tag{D.16}$$

Putting together (D.11) and (D.16) we obtain

$$\left(\mathbb{E}\left[\|e\|_q^8\right]\right)^{\frac{1}{4}} \leqslant \min\{4q-1, 5\ln n\}n^{\frac{2}{q}-1}.$$

We proved the Lemma for $m = 8$, and the **Step 1** proves this Lemma for any number m such that $1 \leqslant m < 8$.

□

D.1.4 Proof of Lemma 10

Lemma 10. $\nabla f_\tau(x) = \mathbb{E}_e g(x,e)$.

Proof.

It was defined in (3.6) that the smoothing function $f_\tau(x)$ is:

$$f_\tau(x) = \mathbb{E}_u f(x+\tau u),$$

where u is random vector uniformly distributed on the unit Euclidean ball with zero center. It was defined in (3.7) that the gradient estimator $g(x,e)$ is:

$$g(x,e) = \frac{n}{2\tau}\left(f(x+\tau e) - f(x-\tau e)\right)e,$$

where e is random vector uniformly distributed on the unit Euclidean sphere with zero center.

Substitute $z = \tau u$, then, according the definition of $f_\tau(x)$

$$f_\tau(x) = \frac{1}{V\left(B_2^n(\tau)\right)} \int\limits_{\|z\|_2 \leqslant \tau} f(x+z)dz, \tag{D.17}$$

where $B_2^n(\tau)$ is a Euclidean ball with zero center and radius τ, $V\left(B_2^n(\tau)\right)$ is the volume of this ball.

If $n = 1$, then by fundamental theorem of calculus:

$$\frac{d}{dx}\int\limits_{-\tau}^{\tau} f(x+z)dz = f(x+\tau) - f(x-\tau).$$

For n-dimensional generalization, it follows from the Stokes theorem that

$$\nabla \int\limits_{\|z\|_2 \leqslant \tau} f(x+z)dz = \int\limits_{\|z\|_2=\tau} f(x+z)\frac{z}{\|z\|_2}dS_\tau(z), \tag{D.18}$$

where $dS_\tau(z)$ is an element of a spherical surface of radius τ.

Combining (D.17) and (D.18), we obtain:

$$\nabla f_\tau(x) = \frac{1}{V(B_2^n(\tau))}\int\limits_{\|z\|_2=\tau} f(x+z)\frac{z}{\|z\|_2}dS_\tau(z) = \frac{A(\partial B_2^n(\tau))}{V\left(B_2^n(\tau)\right)}\mathbb{E}_e[f(x+\tau e)e],$$

where $\partial B_2^n(\tau)$ – is a surface of Euclidean ball with zero center and radius τ, $A(\partial B_2^n(\tau))$ is the area of this surface. In n-dimensional space the formula for the ratio of the spherical surface and the volume of the Euclidean ball with radius τ is:

$$\frac{A(\partial B_2^n(\tau))}{V\left(B_2^n(\tau)\right)} = \frac{n}{\tau},$$

thus,

$$\nabla f_\tau(x) = \frac{n}{\tau}\mathbb{E}_e[f(x+\tau e)e] = \mathbb{E}_e\left[\frac{n}{\tau}f(x+\tau e)e\right].$$

Since e has the same distribution as $-e$, we also obtain:

$$\nabla f_\tau(x) = \mathbb{E}_e\left[\frac{n\left(f(x+\tau e) - f(x-\tau e)\right)}{2\tau}e\right] = \mathbb{E}_e\left[g(x,e)\right].$$

□

D.1.5 Proof of Lemma 11

Lemma 11. $\mathbb{E}_e\|g(x,e)\|_q^2 \leqslant \sqrt{\mathbb{E}_e\|g(x,e)\|_q^4} \leqslant ca_{q,n}^2 nM^2$, *where* $a_{q,n}^2$ *is defined in Lemma 9.*

Proof.

Recall that

$$g(x,e) = \frac{n(f(x+\tau e) - f(x-\tau e))}{2\tau}e.$$

Since $f(x+\tau e)$ has the same distribution on e as $f(x-\tau e)$, we obtain $\mathbb{E}_e[f(x+\tau e)] = \mathbb{E}_e[f(x-\tau e)] \stackrel{\text{def}}{=} m$ and $\mathbb{E}_e[(f(x+\tau e)-m)^8] = \mathbb{E}_e[(f(x-\tau e)-m)^8]$. Thus, using inequations $(A+B)^4 \leqslant 8A^4+8B^4$ and $\mathbb{E}[AB] \leqslant \sqrt{\mathbb{E}[A^2]\mathbb{E}[B^2]}$ we obtain bound:

$$\begin{aligned}
\mathbb{E}_e\left[(f(x+\tau e)-f(x-\tau e))^4 \|e\|_q^4\right] &\leqslant 8\mathbb{E}_e\left[(f(x+\tau e)-m)^4 \|e\|_q^4\right] + 8\mathbb{E}_e\left[(f(x-\tau e)-m)^4 \|e\|_q^4\right] \\
&\leqslant 8\sqrt{\mathbb{E}_e\left[(f(x+\tau e)-m)^8\right]} \cdot \sqrt{\mathbb{E}_e\left[\|e\|_q^8\right]} \\
&+ 8\sqrt{\mathbb{E}_e\left[(f(x-\tau e)-m)^8\right]} \cdot \sqrt{\mathbb{E}_e\left[\|e\|_q^8\right]} \\
&\overset{\text{Lemma 9}}{\leqslant} 16a_{q,n}^4\sqrt{\mathbb{E}_e\left[(f(x+\tau e)-m)^8\right]} \\
&\overset{\text{Lemma 25}}{\leqslant} \frac{16a_{q,n}^4 c^2\tau^4 M_2^4}{n^2},
\end{aligned} \tag{D.19}$$

where in the last inequation Lemma 25 was used for τM_2-Lipschitz function w.r.t. 2-norm: $h(e) = f(x+\tau e)$.

Finally, using $\mathbb{E}[A] \leqslant \sqrt{\mathbb{E}[A^2]}$:

$$\mathbb{E}_e\|g(x,e)\|_q^2 \leqslant \sqrt{\mathbb{E}_e\|g(x,e)\|_q^4} \overset{\text{(D.19)}}{\leqslant} \left(\frac{n}{2\tau}\right)^2 \cdot \sqrt{\frac{16a_{q,n}^4 c^2\tau^4 M_2^4}{n^2}} = a_{q,n}^2 cnM_2^2. \tag{D.20}$$

□

D.1.6 Proof of Lemma 12

Lemma 12 (properties of f_τ)**.** *For all $x,y \in Q$, we have*

– *the inequality*

$$f(x) \leqslant f_\tau(x) \leqslant f(x) + \tau M_2; \tag{D.21}$$

– *$f_\tau(x)$ is M-Lipschitz:*

$$|f_\tau(y) - f_\tau(x)| \leqslant M\|y-x\|_p; \tag{D.22}$$

– *$f_\tau(x)$ has $L = \dfrac{\sqrt{n}M}{\tau}$-Lipschitz gradient:*

$$\|\nabla f_\tau(y) - \nabla f_\tau(x)\|_q \leqslant L\|y-x\|_p, \tag{D.23}$$

where q us such that $1/p + 1/q = 1$;

– *$f_\tau(x)$ inherits (strong) convexity of $f(x)$: if $f(x)$ is (μ-strongly) convex on $U_{\varepsilon_0}(Q)$, then $f_\tau(x)$ is (μ-strongly) convex on Q if $\tau \leqslant \varepsilon_0$.*

Proof.

– For the first inequality in (D.21), we use the convexity of function $f(x)$

$$f_\tau(x) = \mathbb{E}_u\left[f(x+\tau u)\right] \geqslant \mathbb{E}_u\left[f(x) + \langle \nabla f(x), \tau u\rangle\right] = \mathbb{E}_u\left[f(x)\right] = f(x).$$

For the second inequality in (D.21), we use the fact that f is M_2-Lipschitz:

$$|f_\tau(x) - f(x)| = |\mathbb{E}_u\left[f(x+\tau u)\right] - f(x)| \leqslant \mathbb{E}_u\left[|f(x+\tau u) - f(x)|\right] \leqslant \mathbb{E}_u\left[M_2 \cdot \|\tau u\|_2\right] \leqslant \tau M_2.$$

– For (D.22), we have:

$$|f_\tau(y) - f_\tau(x)| = |\mathbb{E}_u\left[f(y+\tau u) - f(x+\tau u)\right]| \leqslant \mathbb{E}_u|f(y+\tau u) - f(x+\tau u)| \leqslant M\|y-x\|_p.$$

– For (D.23), we apply Lemma 11 from [32]:

$$\begin{aligned}\|\nabla f_\tau(y) - \nabla f_\tau(x)\|_q &= \left\|\nabla \mathbb{E}_{Z\sim B_2^n(\tau)}\left[f(y+Z)\right] - \nabla \mathbb{E}_{Z\sim B_2^n(\tau)}\left[f(x+Z)\right]\right\|_q \\ &= \left\|\mathbb{E}_{Z\sim B_2^n(\tau)}\left[\nabla f(y+Z)\right] - \mathbb{E}_{Z\sim B_2^n(\tau)}\left[\nabla f(x+Z)\right]\right\|_q \\ &\leqslant M \underbrace{\int_{B_2^n(\tau)} |\mu(z-y) - \mu(z-x)| dz}_{I_1},\end{aligned}$$

where $\mu(x) = \dfrac{1}{V(B_2^n(\tau))} \cdot \mathbb{I}\left(x \in B_2^n(\tau)\right)$. Note that $f(x)$ is not assumed to be differentiable but the Lebesgue measure of the set where the convex function is not differentiable is equal to zero.

Using the bound for Integral I_1 from Lemma 8 in [31] we obtain

$$\|\nabla f_\tau(y) - \nabla f_\tau(x)\|_q \leqslant \kappa(n)\frac{n!!}{(n-1)!!}\frac{M}{\tau}\|y-x\|_2,$$

where

$$\kappa(n) = \begin{cases} \frac{2}{\pi}, & n \text{ is even;} \\ 1, & n \text{ is odd.} \end{cases}$$

.

Since $\|y-x\|_2 \leqslant \|y-x\|_p$ for $p \leqslant 2$ and $\kappa(n)\dfrac{n!!}{(n-1)!!} < \sqrt{n}$ (see Lemma 26), we obtain:

$$\|\nabla f_\tau(y) - \nabla f_\tau(x)\|_q \leqslant \frac{\sqrt{n}M}{\tau}\|y-x\|_p.$$

– For the last item of this Lemma, we assume that $f(x)$ is μ-strongly convex on $U_{\varepsilon_0}(Q)$. When $f(x)$ is convex only, set $\mu = 0$. Then for any x,$y \in U_{\varepsilon_0}(Q)$ and for any $\alpha \in (0,1)$:

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha)f(y) - \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|_p^2.$$

Fix arbitrary x,$y \in Q$ and arbitrary $\alpha \in (0,1)$. For vector e such that $\|e\|_2 = 1$, it holds that $x + \tau e \in U_\tau(Q) \subseteq U_{\varepsilon_0}(Q)$ and $y + \tau e \in U_\tau(Q) \subseteq U_{\varepsilon_0}(Q)$. Then, applying μ-strong convexity of f to the points $x + \tau e$ and $y + \tau e$:

$$f(\alpha(x+\tau e) + (1-\alpha)(y+\tau e)) \leqslant \alpha f(x+\tau e) + (1-\alpha)f(y+\tau e) - \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|_p^2.$$

Taking expectation on e we obtain the μ-strong convexity of f_τ on Q:

$$f_\tau(\alpha x + (1-\alpha)y) \leqslant \alpha f_\tau(x) + (1-\alpha)f_\tau(y) - \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|_p^2.$$

□

D.1.7 Proof of Lemma 14

Lemma 14 (properties of $\widetilde{g}(x,e)$)**.** *For all $x \in Q$, under the Assumption 13 we have*
– Bias estimate: for a fixed vector $s \in \mathbb{R}^n$

$$|\mathbb{E}_{e,\xi}\left[\langle \widetilde{g}(x,e,\xi) - \nabla f_\tau(x), s\rangle \,|\, x\right]| \leqslant \frac{\sqrt{n}\Delta\|s\|_2}{\tau}. \tag{D.24}$$

– $\widetilde{g}(x,e)$ has bounded second moment:

$$\mathbb{E}_{e,\xi}\left[\|\widetilde{g}(x,e,\xi)\|_q^2 | x\right] = O\left(a_{q,n}^2 \cdot \left(nM_2^2 + \frac{n^2\Delta^2}{\tau^2}\right)\right), \tag{D.25}$$

where $1/p + 1/q = 1$ and $a_{q,n}^2$ is defined in Lemma 9.

Proof.

All expectations is this proof are conditional over e and ξ given x: $\mathbb{E}[\cdot] \leftarrow \mathbb{E}_{e,\xi}[\cdot|x]$.

For the first item of this Lemma (D.24) we note that $\widetilde{g}(x,e) = g(x,e) + \dfrac{n}{2\tau}\left(\dot{\xi} - \ddot{\xi}\right)e$, where $\mathbb{E}g(x,e) = f_\tau(x)$ according to Lemma 10. Thus, $\mathbb{E}\langle g(x,e),s\rangle = \langle f_\tau(x), s\rangle$ and

$$\begin{aligned}
|\mathbb{E}\langle \widetilde{g}(x,e) - \nabla f_\tau(x), s\rangle| &= |\mathbb{E}\langle \widetilde{g}(x,e) - g(x,e), s\rangle + \mathbb{E}\langle g(x,e) - \nabla f_\tau(x), s\rangle| \\
&= |\mathbb{E}\langle \widetilde{g}(x,e) - g(x,e), s\rangle| = \left|\mathbb{E}\left[\frac{n}{2\tau}\left(\dot{\xi} - \ddot{\xi}\right)\langle e,s\rangle\right]\right| \\
&\leqslant \mathbb{E}\left[\frac{n}{2\tau}\left|\left(\dot{\xi} - \ddot{\xi}\right)\langle e,s\rangle\right|\right] \leqslant \frac{n}{2\tau}\sqrt{\mathbb{E}\left[\left(\dot{\xi} - \ddot{\xi}\right)^2\right]}\sqrt{\mathbb{E}\left[\langle e,s\rangle^2\right]}.
\end{aligned} \tag{D.26}$$

Substituting the bound following from Assumption 13 for noise, we have:

$$\mathbb{E}\left[\left(\dot{\xi} - \ddot{\xi}\right)^2\right] \leqslant 2\mathbb{E}[\dot{\xi}^2] + 2\mathbb{E}[\ddot{\xi}^2] \leqslant 4\Delta^2.$$

Since $\mathbb{E}[ee^T] = \frac{1}{n}I_n$, where I_n is the identity matrix of size n, we obtain the equation for fixed vector s:

$$\mathbb{E}\left[\langle e,s\rangle^2\right] = \mathbb{E}\left[s^T e e^T s\right] = s^T\mathbb{E}\left[ee^T\right]s = s^T\frac{I_n}{n}s = \frac{\|s\|_2^2}{n}.$$

Substituting the last two bounds into (D.26), we obtain:

$$|\mathbb{E}\langle \widetilde{g}(x,e) - \nabla f_\tau(x), s\rangle| \leqslant \frac{n \cdot 2\Delta\|s\|_2}{2\tau\sqrt{n}} = \frac{\sqrt{n}\Delta\|s\|_2}{\tau}.$$

For the second item of this Lemma (D.25) we use $(A+B)^2 \leqslant 2A^2 + 2B^2$ and $\mathbb{E}[AB] \leqslant \sqrt{\mathbb{E}[A^2]\mathbb{E}[B^2]}$ to derive:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(f(x+\tau e) + \dot{\xi} - f(x-\tau e) - \ddot{\xi}\right)e\right\|_q^2\right] \\
&\leqslant 2\mathbb{E}\left[\left(f(x+\tau e) - f(x-\tau e)\right)^2\|e\|_q^2\right] + 2\mathbb{E}\left[\left(\dot{\xi} - \ddot{\xi}\right)^2\|e\|_q^2\right] \\
&\overset{(D.19)}{\leqslant} 2a_{q,n}^2\frac{4c\tau^2M_2^2}{n} + 2\mathbb{E}\left[\left(\dot{\xi} - \ddot{\xi}\right)^2\|e\|_q^2\right].
\end{aligned} \tag{D.27}$$

For the second term in (D.27) we have an alternative depending on p. For Euclidean case ($p = q = 2$) under Assumption 13 we have $\mathbb{E}\left[\dot{\xi}^2\right] \leqslant \Delta^2$ and $\mathbb{E}\left[\ddot{\xi}^2\right] \leqslant \Delta^2$, then using $(A+B)^2 \leqslant 2A^2 + 2B^2$ and $\mathbb{E}[AB] \leqslant \sqrt{\mathbb{E}[A^2]\mathbb{E}[B^2]}$ we can obtain:

$$\mathbb{E}\left[(\dot{\xi} - \ddot{\xi})^2 \|e\|_2^2\right] = \mathbb{E}\left[(\dot{\xi} - \ddot{\xi})^2\right] \leqslant 2\mathbb{E}\left[\dot{\xi}^2\right] + 2\mathbb{E}\left[\ddot{\xi}^2\right] \leqslant 4\Delta^2. \tag{D.28}$$

For non-Euclidean case ($p < 2$, $q > 2$) under Assumption 13 we have $\mathbb{E}\left[\dot{\xi}^4\right] \leqslant \Delta^4$ and $\mathbb{E}\left[\ddot{\xi}^4\right] \leqslant \Delta^4$, then using the inequalities $(A+B)^2 \leqslant 2A^2 + 2B^2$ and $\mathbb{E}[AB] \leqslant \sqrt{\mathbb{E}[A^2]\mathbb{E}[B^2]}$ we can obtain:

$$\begin{aligned}\mathbb{E}\left[(\dot{\xi} - \ddot{\xi})^2 \|e\|_q^2\right] &\leqslant 2\mathbb{E}\left[\dot{\xi}^2 \|e\|_q^2\right] + 2\mathbb{E}\left[\ddot{\xi}^2 \|e\|_q^2\right] \leqslant 2\sqrt{\mathbb{E}\left[\dot{\xi}^4\right]\mathbb{E}\left[\|e\|_q^4\right]} + 2\sqrt{\mathbb{E}\left[\ddot{\xi}^4\right]\mathbb{E}\left[\|e\|_q^4\right]} \\ &\overset{Lemma\ 9}{\leqslant} 4\Delta^2 a_{q,n}^2. \end{aligned} \tag{D.29}$$

Since $a_{2,n} = 1$, we can combine (D.28) and (D.29) to obtain that under Assumption 13:

$$\mathbb{E}\left[(\dot{\xi} - \ddot{\xi})^2 \|e\|_q^2\right] \leqslant 4\Delta^2 a_{q,n}^2. \tag{D.30}$$

Substituting (D.30) into (D.27), we have:

$$\mathbb{E}\left[\left\|\left(f(x+\tau e) + \dot{\xi} - f(x-\tau e) - \ddot{\xi}\right) e\right\|_q^2\right] \leqslant 8a_{q,n}^2 \left(\frac{c\tau^2 M_2^2}{n} + \Delta^2\right).$$

Finally, we prove (D.25):

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{g}(x,e)\|_q^2\right] &= \mathbb{E}\left[\frac{n}{\tau}\left\|\left(f(x+\tau e) + \dot{\xi} - f(x-\tau e) - \ddot{\xi}\right) e\right\|_q^2\right] \\ &\leqslant 8a_{q,n}^2 \left(cnM_2^2 + \frac{n^2\Delta^2}{\tau^2}\right) \\ &= O\left(a_{q,n}^2 \left(nM_2^2 + \frac{n^2\Delta^2}{\tau^2}\right)\right).\end{aligned}$$

Note that the constant $c > 0$ is defined in Lemma 25 and does not depend on function f or dimension n.

□

D.1.8 Proof of Lemma 15

Lemma 15. *Let g_i be independent random variables with $\mathbb{E}\,\|\mathring{g}_i\|_q^4 = \sigma_i^4 < \infty$ for all i from 1 to B. Then it holds:*

$$\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^{B} \mathring{g}_i\right)\right\|_q^2 \leqslant 4r_{p,n}^2 \frac{1}{B^2} \sum_{i=1}^{B} \sigma_i^2,$$

where $r^2_{p,n}$ is defined in Lemma 8.

If additionally $\sigma_i = \sigma$ for all i from 1 to B, then we can simplify the inequation above:

$$\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^{B}\mathring{g}_i\right)\right\|_q^2 \leqslant 4r^2_{p,n}\frac{\sigma^2}{B}.$$

Proof.

Let $U = B_p^n(1)$ be a ball with zero center and radius 1 w.r.t. p-norm: $U = \{u \in \mathbb{R}^n : \|u\|_p \leqslant 1$. Set standard distance generating function $d(x)$ with $d(0) = 0$:

$$d(u) = \begin{cases} \frac{\|u\|_p^2}{2(p-1)}, & a \leqslant p \leqslant 2, \\ \frac{e\|u\|_a^2}{2(a-1)}, & 1 \leqslant p < a, \end{cases}$$

where $a = \frac{2\ln n}{2\ln n - 1}$. Then, according to Lemma 8:

$$\max_{u\in U} V(0,u) \leqslant \frac{r^2_{p,n}}{2}, \tag{D.31}$$

where

$$r^2_{p,n} = \begin{cases} \frac{1}{p-1}, & a \leqslant p \leqslant 2, \\ e(2\ln n - 1), & 1 \leqslant p < a. \end{cases}$$

Fix a parameter $\gamma > 0$. Set a sequence $\{u_i\}_{i=1}^{B+1}$: $u_1 \stackrel{\text{def}}{=} 0$ and for $i \geqslant 1$:

$$u_{i+1} = \operatorname{prox}_{\gamma\mathring{g}_i}(u_i) = \arg\min_{u\in U}\left(\langle\gamma\mathring{g}_i, u - u_i\rangle + V(u_i,u)\right).$$

We note that in general the sequence $\{u_i\}_{i=1}^{B+1}$ depends on γ.

Applying Lemma 21, we obtain

$$\langle\gamma\mathring{g}_i, u_{i+1} - u\rangle \leqslant V(u_i,u) - V(u_{i+1},u) - V(u_i,u_{i+1}).$$

Adding $\langle\gamma\mathring{g}_i, u_i - u_{i+1}\rangle$ to the both sides, using $V(u_i,u_{i+1}) \geqslant \frac{1}{2}\|u_i - u_{i+1}\|_p^2$ and the fact that $A,B - \frac{B^2}{2} \leqslant \frac{A^2}{2}$:

$$\begin{aligned}\langle\gamma\mathring{g}_i, u_i - u\rangle &\leqslant V(u_i,u) - V(u_{i+1},u) - V(u_i,u_{i+1}) + \langle\gamma\mathring{g}_i, u_i - u_{i+1}\rangle \\ &\leqslant V(u_i,u) - V(u_{i+1},u) - \frac{1}{2}\|u_i - u_{i+1}\|_p^2 + \gamma\|\mathring{g}_i\|_q\|u_i - u_{i+1}\|_p \\ &\leqslant V(u_i,u) - V(u_{i+1},u) + \frac{1}{2}\gamma^2\|\mathring{g}_i\|_q^2.\end{aligned}$$

Summing and using $V(u_1,u) = V(0,u) \leqslant \frac{1}{2}r^2_{p,n}$ (D.31) and $V(u_{B+1},u) \geqslant 0$ we obtain that for any $u \in U$:

$$\begin{aligned}\gamma\sum_{i=1}^{B}\langle\mathring{g}_i, u_i - u\rangle &\leqslant V(u_1,u) - V(u_{B+1},u) + \frac{1}{2}\gamma^2\sum_{i=1}^{B}\|\mathring{g}_i\|_q^2 \\ &\leqslant \frac{1}{2}r^2_{p,n} - \frac{1}{2}\gamma^2\sum_{i=1}^{B}\|\mathring{g}_i\|_q^2,\end{aligned}$$

or for any $u \in U$:

$$\gamma \left\langle \sum_{i=1}^{B} \mathring{g}_i, -u \right\rangle \leqslant \frac{1}{2} r_{p,n}^2 - \frac{1}{2}\gamma^2 \sum_{i=1}^{B} \|\mathring{g}_i\|_q^2 - \gamma \sum_{i=1}^{B} \langle \mathring{g}_i, u_i \rangle.$$

As for any $u \in U$: $-u \in U$, we obtain:

$$\gamma \left\langle \sum_{i=1}^{B} \mathring{g}_i, u \right\rangle \leqslant \frac{1}{2} r_{p,n}^2 - \frac{1}{2}\gamma^2 \sum_{i=1}^{B} \|\mathring{g}_i\|_q^2 - \gamma \sum_{i=1}^{B} \langle \mathring{g}_i, u_i \rangle. \tag{D.32}$$

Combining with the definition of conjugate norm:

$$\left\| \sum_{i=1}^{B} \mathring{g}_i \right\|_q = \sup_{u \in B} \left\langle \sum_{i=1}^{B} \mathring{g}_i, u \right\rangle \leqslant \frac{1}{2\gamma} r_{p,n}^2 - \frac{1}{2}\gamma \sum_{i=1}^{B} \|\mathring{g}_i\|_q^2 - \sum_{i=1}^{B} \langle \mathring{g}_i, u_i \rangle.$$

Squaring, then using $(\frac{A}{2} + \frac{B}{2} + C)^2 \leqslant A^2 + B^2 + 2C^2$ and after that taking expectation, we obtain:

$$\mathbb{E} \left\| \sum_{i=1}^{B} \mathring{g}_i \right\|_q^2 \leqslant \frac{1}{\gamma^2} r_{p,n}^4 + \gamma^2 \mathbb{E} \left[\left(\sum_{i=1}^{B} \|\mathring{g}_i\|_q^2 \right)^2 \right] + 2\mathbb{E} \left[\left(\sum_{i=1}^{B} \langle \mathring{g}_i, u_i \rangle \right)^2 \right]. \tag{D.33}$$

For the expectation in the second term in (D.33):

$$\mathbb{E} \left[\left(\sum_{i=1}^{B} \|\mathring{g}_i\|_q^2 \right)^2 \right] \leqslant \left(\sum_{i=1}^{B} \sigma_i^2 \right)^2, \tag{D.34}$$

as $\mathbb{E}\|\mathring{g}_i\|_q^4 \leqslant \sigma_i^4$ and for $i \neq j$: $\mathbb{E}\left[\|\mathring{g}_i\|_q^2 \|\mathring{g}_j\|_q^2\right] \leqslant \sqrt{\mathbb{E}\|\mathring{g}_i\|_q^4}\sqrt{\mathbb{E}\|\mathring{g}_j\|_q^4} \leqslant \sigma_i^2\sigma_j^2$.

For the expectation in the third term in (D.33):

$$\begin{aligned}
\mathbb{E} \left[\left(\sum_{i=1}^{B} \langle \mathring{g}_i, u_i \rangle \right)^2 \right] &= \sum_{i=1}^{B} \mathbb{E}\left[\langle \mathring{g}_i, u_i \rangle^2 \right] + 2 \sum_{i=1}^{B} \sum_{j=i+1}^{B} \mathbb{E}\left[\langle \mathring{g}_i, u_i \rangle \langle \mathring{g}_j, u_j \rangle \right] \\
&\leqslant \sum_{i=1}^{B} \mathbb{E}\left[\|\mathring{g}_i\|_q^2 \right] + 2 \sum_{i=1}^{B} \sum_{j=i+1}^{B} \mathbb{E}\left[\langle \mathring{g}_i, u_i \rangle \langle \mathring{g}_j, u_j \rangle \right] \\
&\leqslant \sum_{i=1}^{B} \sigma_i^2 + 0,
\end{aligned} \tag{D.35}$$

as for fixed γ and $j > i$: $\mathring{g}_j$ depends only on e_j and the product $\langle \mathring{g}_i, u_i \rangle u_j$ depends only on history $\{e_m\}_{m=1}^{j-1}$ and does not depend on e_j, moreover, $\mathbb{E}_{e_j}[\mathring{g}_j] = 0$. Denote $H = \{e_m\}_{m=1}^{j-1}$. Using independence of $\{e_m\}_{m=1}^{j}$, consequently independence of H and e_j, we obtain:

$$\mathbb{E}_{H,e_j}\left[\langle \mathring{g}_i, u_i \rangle \langle \mathring{g}_j, u_j \rangle \right] = \langle \mathbb{E}_{e_j}[\mathring{g}_j], \mathbb{E}_H\left[\langle \mathring{g}_i, u_i \rangle u_j \right] \rangle = 0.$$

Summing up, we substitute (D.34) and (D.35) into (D.33) and choose

$$\gamma^2 = \frac{r_{p,n}^2}{\sum_{i=1}^{B} \sigma_i^2}$$

to have:

$$\mathbb{E}\left\|\sum_{i=1}^{B}\mathring{g}_i\right\|_q^2 \leqslant \frac{1}{\gamma^2}r_{p,n}^4+\gamma^2\left(\sum_{i=1}^{B}\sigma_i^2\right)^2+2\sum_{i=1}^{B}\sigma_i^2$$
$$\leqslant 2r_{p,n}^2\left(\sum_{i=1}^{B}\sigma_i^2\right)+2\left(\sum_{i=1}^{B}\sigma_i^2\right).$$
$$=2\left(r_{p,n}^2+1\right)\left(\sum_{i=1}^{B}\sigma_i^2\right)\leqslant 4r_{p,n}^2\left(\sum_{i=1}^{B}\sigma_i^2\right),$$

where in the last inequation we used $r_{p,n}^2 \geqslant 1$ (follows from Lemma 8). Dividing by B^2 we prove the Lemma for arbitrary σ_i^2. The case $\sigma_i = \sigma$ is obvious.

□

D.1.9 Proof of Lemma 16

Lemma 16 (variation of batched noisy gradient estimator)**.**

$$\sigma_B^2 \overset{def}{=} \sup_{x\in Q}\mathbb{E}_{e^{\bar{B}},\xi^{\bar{B}}}\left\|\frac{1}{B}\left(\sum_{i=1}^{B}\widetilde{g}(x,e^i,\xi^i)\right)-\nabla f_\tau(x)\right\|_q^2 \leqslant O\left(a_{q,n}^2\cdot\left(\frac{nr_{p,n}^2M_2^2}{B}+\frac{n^2\Delta^2}{\tau^2}\right)\right),$$

where $1/p+1/q=1$, $a_{q,n}^2$ *is defined in Lemma 9 and* $r_{p,n}^2$ *is defined in Lemma 8.*

Proof.

We represent noisy gradient estimator as the sum of noise-free and consequently unbiased (Lemma 10) estimator and the noise term:

$$\widetilde{g}(x,e^i) = g(x,e^i)+\delta(x,e^i),$$

where

$$g(x,e^i)=\frac{n}{2\tau}\left(f(x+\tau e^i)-f(x-\tau e^i)\right)e^i,$$
$$\delta(x,e^i)=\frac{n}{2\tau}\left(\dot{\xi}^i-\ddot{\xi}^i\right)e^i,$$

and $\mathbb{E}[g(x,e^i,\xi^i)] = \nabla f_\tau(x)$ (Lemma 10).

For brevity, we use following notation: $\widetilde{g}_i \overset{\text{def}}{=} \widetilde{g}(x,e^i)$, $g_i \overset{\text{def}}{=} g(x,e^i)$, $\delta_i \overset{\text{def}}{=} \delta(x,e^i)$ and $m \overset{\text{def}}{=} \nabla f_\tau(x)$. In this proof $\mathbb{E}[\cdot]$ means $\mathbb{E}_{e^{\bar{B}},\xi^{\bar{B}}}[\cdot|x]$.

It follows from (D.20) that $\mathbb{E}\|g_i\|_q^2 \leqslant \sqrt{\mathbb{E}\|g_i\|_q^4} \leqslant \sigma^2$, where

$$\sigma^2 = a_{q,n}^2 cnM_2^2. \tag{D.36}$$

We represent the average of noisy gradient estimators:

$$\left(\frac{1}{B}\sum_{i=1}^{B}\widetilde{g}_i\right)-m=\left(\frac{1}{B}\sum_{i=1}^{B}(g_i-m)\right)+\frac{1}{B}\sum_{i=1}^{B}\delta_i \overset{Lemma\ 10}{=} \left(\frac{1}{B}\sum_{i=1}^{B}\mathring{g}_i\right)+\frac{1}{B}\sum_{i=1}^{B}\delta_i. \tag{D.37}$$

Applying Lemma 15 to the first term of (D.37), we get:

$$\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^{B}\mathring{g}_i\right)\right\|_q^2 \leqslant 2(r_{p,n}^2+1)\frac{\sigma^2}{B}. \tag{D.38}$$

For the second term of (D.37) we have use the convexity of square function

$$\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2 \leqslant \frac{1}{n}\sum_{i=1}^{n}X_i^2$$

to obtain:

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^{B}\delta_i\right)\right\|_q^2 &= \frac{1}{B^2}\mathbb{E}\left\|\sum_{i=1}^{B}\delta_i\right\|_q^2 \leqslant \frac{1}{B^2}\mathbb{E}\left(\sum_{i=1}^{B}\|\delta_i\|_q\right)^2 \leqslant \frac{1}{B}\mathbb{E}\left(\sum_{i=1}^{B}\|\delta_i\|_q^2\right)\\ &= \frac{1}{B}\sum_{i=1}^{B}\mathbb{E}\|\delta_i\|_q^2.\end{aligned} \tag{D.39}$$

For estimating $\mathbb{E}\|\delta_i\|_q^2$ we have an alternative depending on p. For Euclidean case ($p=q=2$), under Assumption 13 we have $\mathbb{E}\left[(\dot{\xi}^i)^2\right] \leqslant \Delta^2$ and $\mathbb{E}\left[(\ddot{\xi}^i)^2\right] \leqslant \Delta^2$, then using $(A+B)^2 \leqslant 2A^2+2B^2$ we can obtain:

$$\mathbb{E}\|\delta_i\|_2^2 = \mathbb{E}\left[\left(\frac{n}{2\tau}\right)^2(\dot{\xi}^i-\ddot{\xi}^i)^2\|e^i\|_2^2\right] = \frac{n^2}{4\tau^2}\mathbb{E}\left[(\dot{\xi}^i-\ddot{\xi}^i)^2\right] \leqslant \frac{n^2}{4\tau^2}\left(2\mathbb{E}\left[(\dot{\xi}^i)^2\right]+2\mathbb{E}\left[(\ddot{\xi}^i)^2\right]\right) \leqslant \frac{n^2\Delta^2}{\tau^2}. \tag{D.40}$$

For non-Euclidean case ($p<2$, $q>2$), under Assumption 13 we have $\mathbb{E}\left[(\dot{\xi}^i)^4\right] \leqslant \Delta^4$ and $\mathbb{E}\left[(\ddot{\xi}^i)^4\right] \leqslant \Delta^4$, then using the inequalities $(A+B)^2 \leqslant 2A^2+2B^2$ and $\mathbb{E}[AB] \leqslant \sqrt{\mathbb{E}[A^2]\mathbb{E}[B^2]}$ we can obtain

$$\begin{aligned}\mathbb{E}\|\delta_i\|_q^2 &= \mathbb{E}\left[\left(\frac{n}{2\tau}\right)^2(\dot{\xi}^i-\ddot{\xi}^i)^2\|e^i\|_q^2\right] = \frac{n^2}{4\tau^2}\mathbb{E}\left[(\dot{\xi}^i-\ddot{\xi}^i)^2\|e^i\|_q^2\right]\\ &\leqslant \frac{n^2}{4\tau^2}\left(2\mathbb{E}\left[(\dot{\xi}^i)^2\|e^i\|_q^2\right]+2\mathbb{E}\left[(\ddot{\xi}^i)^2\|e^i\|_q^2\right]\right)\\ &\leqslant \frac{n^2}{4\tau^2}\left(2\sqrt{\mathbb{E}\left[(\dot{\xi}^i)^4\right]\mathbb{E}\left[\|e^i\|_q^4\right]}+2\sqrt{\mathbb{E}\left[(\ddot{\xi}^i)^4\right]\mathbb{E}\left[\|e^i\|_q^4\right]}\right)\\ &\overset{Lemma\ 9}{\leqslant} a_{q,n}^2\frac{n^2\Delta^2}{\tau^2}.\end{aligned} \tag{D.41}$$

Since $a_{2,n}=1$, we can combine (D.40) and (D.41) to obtain that under Assumption 13:

$$\mathbb{E}\|\delta_i\|_q^2 \leqslant a_{q,n}^2\frac{n^2\Delta^2}{\tau^2} \tag{D.42}$$

and substituting (D.42) to (D.39):

$$\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^{B}\delta_i\right)\right\|_q^2 \leqslant \frac{1}{B}\sum_{i=1}^{B}\mathbb{E}\|\delta_i\|_q^2 = a_{q,n}^2\frac{n^2\Delta^2}{\tau^2}. \tag{D.43}$$

Combining (D.37), (D.38), (D.43), (D.36) and using $\|A+B\|_q^2 \leqslant (\|A\|_q + \|B\|_q)^2 \leqslant 2\|A\|_q^2 + 2\|B\|_q^2$, we obtain for all $x \in Q$:

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^{B}\widetilde{g}_i\right) - m\right\|_q^2 &\overset{(D.37)}{\leqslant} 2\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^{B}\mathring{g}_i\right) - m\right\|_q^2 + 2\mathbb{E}\left\|\frac{1}{B}\left(\sum_{i=1}^{B}\delta_i\right)\right\|_q^2 \\
&\overset{(D.38),(D.43)}{\leqslant} \frac{4\sigma^2}{B}(r_{p,n}^2+1) + 2a_{q,n}^2\frac{n^2\Delta^2}{\tau^2} \\
&\overset{(D.36)}{=} \frac{4a_{q,n}^2 cnM_2^2}{B}\left(r_{p,n}^2+1\right) + 2a_{q,n}^2\frac{n^2\Delta^2}{\tau^2} \\
&= 2a_{q,n}^2\left(\frac{4cr_{p,n}^2nM_2^2}{B} + \frac{n^2\Delta^2}{\tau^2}\right) \\
&= O\left(a_{q,n}^2\cdot\left(\frac{nr_{p,n}^2M_2^2}{B} + \frac{n^2\Delta^2}{\tau^2}\right)\right).
\end{aligned} \tag{D.44}$$

□

D.2 Theorems

D.2.1 Proof of Theorem 17

Theorem 17. *Assume that f is convex and M-Lipschitz w.r.t. p-norm and M_2-Lipschitz w.r.t. Euclidean norm. Let us define step sizes of Algorithm 3 as $\beta_k = \frac{(k+1)}{2}$, $\gamma_k = \frac{(k+1)\gamma^*}{2}$ for all k from 1 to N, where γ^* is defined from (3.21). Then under the Assumption 13:*

$$\mathbb{E}\left[f(x_{N+1}^{ag}) - f(x^*)\right] \leqslant \frac{8Lr_{p,n}^2R_p^2}{N^2} + \frac{4\sigma_B r_{p,n}R_p}{\sqrt{N}} + \frac{\Delta\sqrt{n}D_2}{\tau} + M_2\tau,$$

where $L = \frac{\sqrt{n}M_2}{\tau}$, σ_B is defined in Lemma 16, $r_{p,n}$ is defined in Lemma 8, $D_2 = \max\limits_{x,y\in Q}\|x-y\|_2$ is the Euclidean diameter of the set Q, $R_p = \|x_1 - x_\tau^*\|_p$ is the distance from the starting point x_1 to the minimizer x_τ^* of the function $f_\tau(x)$ on the set Q: $x_\tau^* \in \arg\min_{x\in Q} f_\tau(x)$.

Proof.

Recall the Algorithm 3. Initialization: set

$$\begin{aligned}
\beta_k &= \frac{k+1}{2}, \\
\gamma_k &= \frac{(k+1)\gamma^*}{2},
\end{aligned}$$

where

$$\gamma^* = \min\left\{\frac{1}{2L}, \frac{2\sqrt{3}r_{p,n}R_p}{(N+2)^{3/2}\cdot\sigma_B}\right\}, \tag{D.45}$$

$$L \overset{(3.9)}{=} \frac{\sqrt{n}M_2}{\tau} \tag{D.46}$$

$$r_{p,n}^2 \overset{Lemma\ 8}{=} \begin{cases} \frac{1}{p-1}, & a \leqslant p \leqslant 2,\ a = \frac{2\ln n}{2\ln n - 1} \\ e(2\ln n - 1), & 1 \leqslant p < a. \end{cases} \tag{D.47}$$

$$\sigma_B^2 \overset{(D.44)}{=} 2a_{q,n}^2\left(\frac{4cr_{p,n}^2 nM_2^2}{B} + \frac{n^2\Delta^2}{\tau^2}\right). \tag{D.48}$$

Iteration:

1. Set
$$x_k^{md} = \beta_k^{-1}x_k + (1-\beta_k^{-1})x_k^{ag} \tag{D.49}$$
2. Generate vectors $e_k^1, \ldots, e_k^B$ uniformly on the Euclidean unit sphere S_n with zero center and call the batched gradient approximation:
$$\widetilde{G}_k = \frac{1}{B}\left(\sum_{i=1}^{B}\widetilde{g}_k\left(x_k^{md}, e_k^i, \xi_k^i\right)\right) \tag{D.50}$$
3. Update
$$x_{k+1} = \mathrm{Prox}_{x_k}(\gamma_k\widetilde{G}_k(x_k^{md})), \tag{D.51}$$
$$x_{k+1}^{ag} = \beta_k^{-1}x_{k+1} + (1-\beta_k^{-1})x_k^{ag} \tag{D.52}$$

Notation

For brevity, we use the following notation:

$$\widetilde{g}_k^i \overset{\text{def}}{=} \widetilde{g}_k\left(x_k^{md}, e_k^i, \xi_k^i\right)$$

and for the batched values:

$$\widetilde{G}_k = \frac{1}{B}\sum_{i=1}^{B}\widetilde{g}_k^i. \tag{D.53}$$

We also define the bias ζ_k between noisy batched gradient estimator and the true gradient of smoothed function f_τ:

$$\zeta_k = \widetilde{G}_k - \nabla f_\tau(x_k^{md}). \tag{D.54}$$

Denote by x^* any minimizer of $f(x)$:

$$x^* \in \arg\min_{x\in Q} f(x)$$

and by x_τ^* any minimizer of $f_\tau(x)$:

$$x_\tau^* \in \arg\min_{x\in Q} f_\tau(x).$$

Preliminaries

In this proof $\mathbb{E}[\cdot]$ means full expectation over $\{e_k^{\bar{B}}, \xi_k^{\bar{B}}\}_{k=1}^N$.
The convexity of continuous differentiable $f_\tau(x)$ on Q:

$$f_\tau(y) \overset{Lemma\ 12}{\geqslant} f_\tau(x) + \langle \nabla f_\tau(x), y - x\rangle. \tag{D.55}$$

$f_\tau(x)$ has $L = \dfrac{\sqrt{n}M_2}{\tau}$-Lipschitz gradient w.r.t. p-norm:

$$f_\tau(y) \overset{Lemma\ 12}{\leqslant} f_\tau(x) + \langle \nabla f_\tau(x), y - x\rangle + \frac{L}{2}\|x - y\|_p^2. \tag{D.56}$$

The estimate for Bregman divergence:

$$V(x,y) \overset{Section\ 3.4.1}{\geqslant} \frac{1}{2}\|x - y\|_p^2. \tag{D.57}$$

The boundness of batched gradient estimator second moment:

$$\mathbb{E}\left\|\widetilde{G}_k - \nabla f_\tau(x_k^{md})\right\|_q^2 \overset{Lemma\ 16}{\leqslant} \sigma_B^2, \tag{D.58}$$

where full expectation bound follows from conditional expectation bound in Lemma 16.

Bias estimate:

$$\mathbb{E}\left[\langle \widetilde{G}_k - \nabla f_\tau(x_k^{md}), x_\tau^* - x_k\rangle\right] \leqslant \frac{\Delta\sqrt{n}D_2}{\tau} \tag{D.59}$$

Proof. Define $e_j^{\bar{B}} \overset{\text{def}}{=} \{e_j^1, \ldots, e_j^B\}$, $\xi_j^{\bar{B}} \overset{\text{def}}{=} \{\xi_j^1, \ldots, \xi_j^B\}$. Fix $\{e_1^{\bar{B}}, \ldots, e_{k-1}^{\bar{B}}\}$ and $\{\xi_1^{\bar{B}}, \ldots, \xi_{k-1}^{\bar{B}}\}$ and denote history $H = \{e_j^{\bar{B}}, \xi_j^{\bar{B}}\}_{j=1}^{k-1}$. Then the point x_k^{md} is fixed as a linear combination of x_1 and gradients $\widetilde{G}_1, \ldots, \widetilde{G}_{k-1}$ depending on H and not depending on $e_k^{\bar{B}}$, $\xi_k^{\bar{B}}$ (see Algorithm 3), consequently, $\{e_k^i\}_{i=1}^B$ are distributed independently and uniformly on Euclidean unit sphere. The vector $x_\tau^* - x_k$ is also fixed. Applying the first part (3.13) of Lemma 14 we obtain:

$$\begin{aligned}\mathbb{E}_{e_k^{\bar{B}}, \xi_k^{\bar{B}}}\Big[\langle \widetilde{G}_k - \underbrace{\nabla f_\tau(x_k^{md}), x_\tau^* - x_k}_{fixed}\rangle \mid H\Big] &\overset{(D.53)}{=} \frac{1}{B}\sum_{i=1}^B \langle \mathbb{E}_{e_k^{\bar{B}}, \xi_k^{\bar{B}}}\left[\widetilde{g}_k^i \mid H\right] - \nabla f_\tau(x_k^{md}), x_\tau^* - x_k\rangle \\ &\overset{(3.13)}{\leqslant} \frac{1}{B}\sum_{i=1}^B \frac{\sqrt{n}\Delta\|x_\tau^* - x_k\|_2}{\tau} = \frac{\sqrt{n}\Delta\|x_\tau^* - x_k\|_2}{\tau}.\end{aligned}$$

$$\mathbb{E}\left[\langle \widetilde{G}_k - \nabla f_\tau(x_k^{md}), x_\tau^* - x_k\rangle\right] \leqslant \frac{\sqrt{n}\Delta\mathbb{E}\|x_\tau^* - x_k\|_2}{\tau} \leqslant \frac{\sqrt{n}\Delta D_2}{\tau}.$$

Step 1

Recall that

$$x_{k+1}^{ag} - x_k^{md} \overset{(D.52),(D.49)}{=} \beta_k^{-1}(x_{k+1} - x_k) \tag{D.60}$$

Using L-Lipschitz continuity of $\nabla f_\tau(x)$ and the properties of Bregman divergence:

$$\begin{aligned}
\beta_k\gamma_k\left[f_\tau(x_{k+1}^{ag}) - f_\tau(x)\right] &\overset{(D.56)}{\leqslant} \beta_k\gamma_k\left[f_\tau(x_k^{md}) + \langle\nabla f_\tau(x_k^{md}), x_{k+1}^{ag} - x_k^{md}\rangle + \frac{L}{2}\|x_{k+1}^{ag} - x_k^{md}\|_p^2\right] \\
&\overset{(D.60)}{\leqslant} \beta_k\gamma_k\left[f_\tau(x_k^{md}) + \langle\nabla f_\tau(x_k^{md}), x_{k+1}^{ag} - x_k^{md}\rangle\right] + \frac{\gamma_k L}{2\beta_k}\|x_{k+1} - x_k\|_p^2 \\
&\overset{(D.57)}{\leqslant} \beta_k\gamma_k\left[f_\tau(x_k^{md}) + \langle\nabla f_\tau(x_k^{md}), x_{k+1}^{ag} - x_k^{md}\rangle\right] \\
&\quad + \frac{1}{2}V(x_k, x_{k+1}) - \frac{\beta_k - \gamma_k L}{2\beta_k}\|x_{k+1} - x_k\|_p^2
\end{aligned} \tag{D.61}$$

Using the convexity of $f_\tau(x)$:

$$\begin{aligned}
&\beta_k\gamma_k\left[f_\tau(x_k^{md}) + \langle\nabla f_\tau(x_k^{md}), x_{k+1}^{ag} - x_k^{md}\rangle\right] \\
&\overset{(D.52)}{=} \beta_k\gamma_k\left[f_\tau(x_k^{md}) + \langle\nabla f_\tau(x_k^{md}), \beta_k^{-1}x_{k+1} + (1-\beta_k^{-1})x_k^{ag} - x_k^{md}\rangle\right] \\
&= (\beta_k - 1)\gamma_k\left[f_\tau(x_k^{md}) + \langle\nabla f_\tau(x_k^{md}), x_k^{ag} - x_k^{md}\rangle\right] + \gamma_k\left[f_\tau(x_k^{md}) + \langle\nabla f_\tau(x_k^{md}), x_{k+1} - x_k^{md}\rangle\right] \\
&\overset{(D.55)}{\leqslant} (\beta_k - 1)\gamma_k f_\tau(x_k^{ag}) + \gamma_k\left[f_\tau(x_k^{md}) + \langle\widetilde{G}_k, x_{k+1} - x_k^{md}\rangle - \langle\zeta_k, x_{k+1} - x_k^{md}\rangle\right] \\
&= (\beta_k - 1)\gamma_k f_\tau(x_k^{ag}) + \gamma_k\left[f_\tau(x_k^{md}) + \langle\widetilde{G}_k, x_{k+1} - x_k^{md}\rangle\right] \\
&- \gamma_k\langle\zeta_k, x_k - x_k^{md}\rangle - \gamma_k\langle\zeta_k, x_{k+1} - x_k\rangle \\
&\leqslant (\beta_k - 1)\gamma_k f_\tau(x_k^{ag}) + \gamma_k\left[f_\tau(x_k^{md}) + \langle\widetilde{G}_k, x_{k+1} - x_k^{md}\rangle\right] \\
&- \gamma_k\langle\zeta_k, x_k - x_k^{md}\rangle - \gamma_k\|\zeta_k\|_q\|x_{k+1} - x_k\|_p
\end{aligned} \tag{D.62}$$

Substituting (D.62) into (D.61) and using $-AB - \frac{B^2}{2} \leqslant \frac{A^2}{2}$ we obtain:

$$\begin{aligned}
\beta_k\gamma_k f_\tau(x_{k+1}^{ag}) &\leqslant (\beta_k - 1)\gamma_k f_\tau(x_k^{ag}) + \gamma_k\left[f_\tau(x_k^{md}) + \langle\widetilde{G}_k, x_{k+1} - x_k^{md}\rangle\right] + \frac{1}{2}V(x_k, x_{k+1}) \\
&\quad - \frac{\beta_k - \gamma_k L}{2\beta_k}\|x_{k+1} - x_k\|_p^2 - \gamma_k\langle\zeta_k, x_k - x_k^{md}\rangle - \gamma_k\|\zeta_k\|_q\|x_{k+1} - x_k\|_p \\
&\leqslant (\beta_k - 1)\gamma_k f_\tau(x_k^{ag}) + \gamma_k\left[f_\tau(x_k^{md}) + \langle\widetilde{G}_k, x_{k+1} - x_k^{md}\rangle\right] + \frac{1}{2}V(x_k, x_{k+1}) \\
&\quad - \gamma_k\langle\zeta_k, x_k - x_k^{md}\rangle + \frac{\|\zeta_k\|_q^2\gamma_k^2\beta_k}{2(\beta_k - \gamma_k L)}.
\end{aligned} \tag{D.63}$$

Using Lemma 21, the convexity of f and the definition of ζ_k:

$$\begin{aligned}
&\gamma_k\left[f_\tau(x_k^{md})+\langle\widetilde{G}_k,x_{k+1}-x_k^{md}\rangle\right]+\frac{1}{2}V(x_k,x_{k+1})\\
&=\gamma_k\left[f_\tau(x_k^{md})+\langle\widetilde{G}_k,x-x_k^{md}\rangle\right]+\gamma_k\langle\widetilde{G}_k,x_{k+1}-x\rangle+\frac{1}{2}V(x_k,x_{k+1})\\
&\overset{Lemma\ 21}{\leqslant}\gamma_k\left[f_\tau(x_k^{md})+\langle\widetilde{G}_k,x-x_k^{md}\rangle\right]+V(x_k,x)-V(x_{k+1},x)\underbrace{-V(x_k,x_{k+1})+\frac{1}{2}V(x_k,x_{k+1})}_{\leqslant 0}\\
&\overset{(D.54)}{\leqslant}\gamma_k\left[f_\tau(x_k^{md})+\langle\nabla f_\tau(x_k^{md}),x-x_k^{md}\rangle\right]+\gamma_k\langle\zeta_k,x-x_k^{md}\rangle+V(x_k,x)-V(x_{k+1},x)\\
&\overset{(D.55)}{\leqslant}\gamma_k f_\tau(x)+\gamma_k\langle\zeta_k,x-x_k^{md}\rangle+V(x_k,x)-V(x_{k+1},x).
\end{aligned}\tag{D.64}$$

Substituting (D.64) into (D.63) we obtain:

$$\begin{aligned}
\beta_k\gamma_k f_\tau(x_{k+1}^{ag})&\leqslant(\beta_k-1)\gamma_k f_\tau(x_k^{ag})+\gamma_k f_\tau(x)+\gamma_k\langle\zeta_k,x-x_k^{md}\rangle+V(x_k,x)-V(x_{k+1},x)\\
&\quad-\gamma_k\langle\zeta_k,x_k-x_k^{md}\rangle+\frac{\|\zeta_k\|_q^2\gamma_k^2\beta_k}{2(\beta_k-\gamma_k L)}\\
&=(\beta_k-1)\gamma_k f_\tau(x_k^{ag})+\gamma_k f_\tau(x)+V(x_k,x)-V(x_{k+1},x)\\
&\quad+\frac{\|\zeta_k\|_q^2\gamma_k^2\beta_k}{2(\beta_k-\gamma_k L)}+\gamma_k\langle\zeta_k,x-x_k\rangle
\end{aligned}\tag{D.65}$$

Subtracting $\gamma_k f_\tau(x)$ from both sides, we obtain:

$$\begin{aligned}
\beta_k\gamma_k\left[f_\tau(x_{k+1}^{ag})-f_\tau(x)\right]&\leqslant V(x_k,x)-V(x_{k+1},x)+(\beta_k-1)\gamma_k\left[f_\tau(x_k^{ag})-f_\tau(x)\right]+\\
&\quad+\frac{\|\zeta_k\|_q^2\gamma_k^2\beta_k}{2(\beta_k-\gamma_k L)}+\gamma_k\langle\zeta_k,x-x_k\rangle
\end{aligned}\tag{D.66}$$

Step 2

Note that for $\beta_k=\frac{k+1}{2}$, $\gamma_k=\frac{(k+1)\gamma^*}{2}$ it holds that:

$$\beta_1=1;\tag{D.67}$$

$$0<(\beta_{k+1}-1)\gamma_{k+1}<\beta_k\gamma_k,\ \forall k;\tag{D.68}$$

$$2\gamma_k L\leqslant\beta_k,\ \forall k.\tag{D.69}$$

Define η_k:

$$\eta_k=\frac{\|\zeta_k\|_q^2\gamma_k^2\beta_k}{2(\beta_k-\gamma_k L)}+\gamma_k\langle\zeta_k,x_\tau^*-x_k\rangle\tag{D.70}$$

Then

$$
\begin{aligned}
(\beta_{N+1}-1)\gamma_{N+1}\left[f_\tau(x^{ag}_{N+1})-f_\tau(x^*_\tau)\right] &\overset{(D.68)}{\leqslant} \beta_N\gamma_N\left[f_\tau(x^{ag}_{N+1})-f_\tau(x^*_\tau)\right]\\
&\leqslant V(x_N,x)-V(x_{N+1},x)+(\beta_N-1)\gamma_N\left[f_\tau(x^{ag}_N)-f_\tau(x)\right]\\
&+\eta_N \leqslant \ldots \leqslant \sum_{k=1}^{N}(V(x_k,x^*_\tau)-V(x_{k+1},x^*_\tau))+\sum_{k=1}^{N}\eta_k\\
&+\underbrace{(\beta_1-1)}_{=0}\gamma_1\left[f_\tau(x^{ag}_1)-f_\tau(x)\right]\\
&\overset{(D.67)}{\leqslant} V(x_1,x^*_\tau)\underbrace{-V(x_{N+1},x^*_\tau)}_{\leqslant 0}+\sum_{k=1}^{N}\eta_k\\
&\overset{(D.72)}{\leqslant} r^2_{p,n}\|x_1-x^*_\tau\|_p^2+\sum_{k=1}^{N}\eta_k, \qquad (D.71)
\end{aligned}
$$

where in the last inequation the fact that $d(x_1)=0$ and $\nabla d(x_1)=0$ for our distance generating function (3.1) was used, consequently

$$
V(x_1,x^*_\tau)=\|x_1-x^*_\tau\|_p^2 V\left(x_1,x_1+\frac{x^*_\tau-x_1}{\|x_1-x^*_\tau\|_p}\right)\overset{(D.67)}{\leqslant} r^2_{p,n}\|x_1-x^*_\tau\|_p^2. \qquad (D.72)
$$

Estimate full expectation $\mathbb{E}[\eta_k]$:

$$
\begin{aligned}
\mathbb{E}[\eta_k] &= \frac{\mathbb{E}\left[\|\zeta_k\|_q^2\right]\gamma_k^2\beta_k}{2(\beta_k-\gamma_k L)}+\gamma_k\mathbb{E}\left[\langle\zeta_k,x^*_\tau-x_k\rangle\right]\\
&\overset{(D.59),(D.58)}{\leqslant} \frac{\sigma_B^2(k+1)^2(\gamma^*)^2}{8(1-\gamma^* L)}+\gamma^*\frac{k+1}{2}\frac{\Delta\sqrt{n}D_2}{\tau}
\end{aligned}
$$

Summing and using $\sum\limits_{k=1}^{N}(k+1)^2 \leqslant \int_1^{N+1}(u+1)^2du \leqslant \frac{(N+2)^3}{3}$ and $\sum\limits_{k=1}^{N}(k+1)=\frac{N(N+2)}{2}$ we obtain:

$$
\begin{aligned}
\sum_{k=1}^{N}\mathbb{E}\eta_k &\leqslant \sum_{k=1}^{N}\frac{\sigma_B^2(k+1)^2(\gamma^*)^2}{8(1-\gamma^* L)}+\sum_{k=1}^{N}\gamma^*\frac{k+1}{2}\frac{\Delta\sqrt{n}D_2}{\tau}\\
&\leqslant \frac{(N+2)^3}{24}\frac{\sigma_B^2(\gamma^*)^2}{1-\gamma^* L}+\frac{N(N+2)}{4}\gamma^*\frac{\Delta\sqrt{n}D_2}{\tau} \qquad (D.73)
\end{aligned}
$$

Substituting $\beta_{N+1}=\frac{N+2}{2}$, $\gamma_{N+1}=\frac{(N+2)\gamma^*}{2}$ and (D.73) into (D.71):

$$
\begin{aligned}
\mathbb{E}\left[f_\tau(x^{ag}_{N+1})-f_\tau(x^*_\tau)\right] &\leqslant \frac{4r^2_{p,n}\|x_1-x^*_\tau\|_p^2}{N(N+2)\gamma^*}+\frac{(N+2)^3}{24}\frac{4}{N(N+2)}\frac{\sigma_B^2\gamma^*}{1-\gamma^* L}+\frac{\Delta\sqrt{n}D_2}{\tau}\\
&= \frac{4r^2_{p,n}\|x_1-x^*_\tau\|_p^2}{N(N+2)\gamma^*}+\frac{(N+2)^2}{6N}\frac{\sigma_B^2\gamma^*}{1-\gamma^* L}+\frac{\Delta\sqrt{n}D_2}{\tau}.
\end{aligned}
$$

Substituting γ^*:

$$
\gamma^*\overset{(D.45)}{=}\min\left\{\frac{1}{2L},\frac{2\sqrt{3}r_{p,n}R_p}{(N+2)^{3/2}\cdot\sigma_B}\right\}.
$$

and using $N \leqslant N+2 \leqslant 3N$ we obtain:

$$\begin{aligned}\mathbb{E}\left[f_\tau(x_{N+1}^{ag}) - f_\tau(x_\tau^*)\right] &\leqslant \frac{4r_{p,n}^2\|x_1 - x_\tau^*\|_p^2}{N(N+2)\gamma^*} + \frac{(N+2)^2}{6N}\frac{\sigma_B^2\gamma^*}{1-\gamma^* L} + \frac{\Delta\sqrt{n}D_2}{\tau} \\ &\leqslant \frac{4r_{p,n}^2\|x_1 - x_\tau^*\|_p^2}{N(N+2)\gamma^*} + \frac{(N+2)^2\sigma_B^2\gamma^*}{3N} + \frac{\Delta\sqrt{n}D_2}{\tau} \\ &\overset{(D.45)}{\leqslant} \frac{8Lr_{p,n}^2\|x_1 - x_\tau^*\|_p^2}{N(N+2)} + \frac{4\sigma_B r_{p,n}\|x_1 - x_\tau^*\|_p}{\sqrt{N}} + \frac{\Delta\sqrt{n}D_2}{\tau} \\ &\leqslant \frac{8Lr_{p,n}^2R_p^2}{N^2} + \frac{4\sigma_B r_{p,n}R_p}{\sqrt{N}} + \frac{\Delta\sqrt{n}D_2}{\tau}. \end{aligned} \tag{D.74}$$

Step 3

As $f_\tau(x_\tau^*) \leqslant f_\tau(x^*)$:

$$\begin{aligned} f(x_{N+1}^{ag}) - f(x^*) &\overset{(3.8)}{\leqslant} f_\tau(x_{N+1}^{ag}) - f(x^*) \overset{(3.8)}{\leqslant} f_\tau(x_{N+1}^{ag}) - f_\tau(x^*) + M_2\tau \\ &\leqslant f_\tau(x_{N+1}^{ag}) - f_\tau(x_\tau^*) + M_2\tau. \end{aligned}$$

Taking expectation:

$$\mathbb{E}\left[f(x_{N+1}^{ag}) - f(x^*)\right] \overset{(D.74)}{\leqslant} \frac{8Lr_{p,n}^2R_p^2}{N^2} + \frac{4\sigma_B r_{p,n}R_p}{\sqrt{N}} + \frac{\Delta\sqrt{n}D_2}{\tau} + M_2\tau, \tag{D.75}$$

where

$$\begin{aligned} L &\overset{Lemma\ 14}{=} \frac{\sqrt{n}M_2}{\tau}, \\ \sigma_B^2 &\overset{(D.48)}{=} 2a_{q,n}^2\left(\frac{4cr_{p,n}^2nM_2^2}{B} + \frac{n^2\Delta^2}{\tau^2}\right). \end{aligned} \tag{D.76}$$

□

D.2.2 Proof of Corollary 18

Corollary 18. *Based on the batched Accelerated gradient method, the Smoothing scheme applied to non-smooth problem* (1)*, provides a gradient-free method with*

$$N(\varepsilon) = O\left(\frac{n^{1/4}\sqrt{M_2 M}r_{p,n}R_p}{\varepsilon}\right) = \begin{cases} O\left(\frac{n^{1/4}M_2R_2}{\varepsilon}\right), & p = 2\ (q=2), \\ O\left(\frac{(\ln n)^{1/2}n^{1/4}\sqrt{M_2 M}R_1}{\varepsilon}\right), & p = 1\ (q=\infty) \end{cases}$$

successive iterations and

$$T(\varepsilon) = N(\varepsilon)\cdot B(\varepsilon) = O\left(\frac{a_{q,n}^2nM_2^2r_{p,n}^4R_p^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{nM_2^2R_2^2}{\varepsilon^2}\right), & p = 2\ (q=2), \\ O\left(\frac{(\ln n)^3M_2^2R_1^2}{\varepsilon^2}\right), & p = 1\ (q=\infty) \end{cases}$$

zeroth-order oracle calls, where $1/p+1/q=1$, $a^2_{q,n}$ *is defined in Lemma 9,* $r^2_{p,n}$ *is defined in Lemma 8;* $D_2 = \max\limits_{x,y\in Q}\|x-y\|_2$ *is the Euclidean diameter of the set* Q, $R_p = \|x_1 - x^*_\tau\|_p$ *is the distance from the start point* x_1 *to the minimizer* x^*_τ *of the function* $f_\tau(x)$ *on the set* Q*:* $x^*_\tau \in \arg\min_{x\in Q} f_\tau(x)$.

Proof.

We have

$$\tau = \frac{\varepsilon}{2M_2}, \tag{D.77}$$

$$L \overset{Lemma\ 3.9}{=} \frac{\sqrt{n}M}{\tau} \overset{(D.77)}{=} \frac{2\sqrt{n}MM_2}{\varepsilon}, \tag{D.78}$$

$$\sigma_B^2 \overset{Lemma\ 16}{\leqslant} 2a^2_{q,n}\left(\frac{4cr^2_{p,n}nM_2^2}{B} + \frac{n^2\Delta^2}{\tau^2}\right). \tag{D.79}$$

We also have (D.75) from Theorem 17:

$$\mathbb{E}\left[f(x^{ag}_{N+1}) - f(x^*)\right] \leqslant \frac{8Lr^2_{p,n}R_p^2}{N^2} + \frac{4\sigma_B r_{p,n}R_p}{\sqrt{N}} + \frac{\Delta\sqrt{n}D_2}{\tau} + M_2\tau. \tag{D.80}$$

Substituting (D.78), (D.77) and (D.79) to (D.80) and using $\sqrt{A+B} \leqslant \sqrt{A}+\sqrt{B}$, we obtain:

$$\begin{aligned}\mathbb{E}\left[f(x^{ag}_{N+1}) - f(x^*)\right] \leqslant{}& \frac{16\sqrt{n}MM_2r^2_{p,n}R_p^2}{\varepsilon N^2} + \frac{4\sqrt{2}a_{q,n}r_{p,n}R_p}{\sqrt{N}}\left(\frac{2\sqrt{c}\sqrt{n}r_{p,n}M_2}{\sqrt{B}} + \frac{2n\Delta M_2}{\varepsilon}\right)\\ &+ \frac{\Delta M_2 D_2\sqrt{n}}{\varepsilon} + \frac{\varepsilon}{2},\end{aligned} \tag{D.81}$$

where c is a constant from Lemma 25.

Assume, that noise level Δ is sufficiently small (see Appendix D.3 for details) so that all noise-containing terms are less than $\varepsilon/4$:

$$\frac{4\sqrt{2}a_{q,n}r_{p,n}R_p}{\sqrt{N}} \cdot \frac{2n\Delta M_2}{\varepsilon} + \frac{\Delta M_2 D_2\sqrt{n}}{\varepsilon} \leqslant \frac{\varepsilon}{4}. \tag{D.82}$$

Then it holds that:

$$\mathbb{E}\left[f(x^{ag}_{N+1}) - f(x^*)\right] \leqslant \frac{16\sqrt{n}MM_2r^2_{p,n}R_p^2}{\varepsilon N^2} + \frac{8\sqrt{2}a_{q,n}\sqrt{c}r^2_{p,n}\sqrt{n}M_2R_p}{\sqrt{NB}} + \frac{3\varepsilon}{4},$$

To have $\mathbb{E}\left[f(x^{ag}_{N+1}) - f(x^*)\right] \leqslant \varepsilon$ we need:

$$\frac{16\sqrt{n}MM_2r^2_{p,n}R_p^2}{\varepsilon N^2} \leqslant \frac{\varepsilon}{8} \tag{D.83}$$

and

$$\frac{8\sqrt{2}a_{q,n}\sqrt{c}r^2_{p,n}\sqrt{n}M_2R_p}{\sqrt{NB}} \leqslant \frac{\varepsilon}{8}. \tag{D.84}$$

From (D.83) we obtain the total number of iterations:

$$\begin{aligned}N = \frac{8\sqrt{2}n^{1/4}\sqrt{M_2M}r_{p,n}R_p}{\varepsilon} &= O\left(\frac{n^{1/4}\sqrt{M_2M}r_{p,n}R_p}{\varepsilon}\right)\\ &= \begin{cases} O\left(\frac{n^{1/4}M_2D_2}{\varepsilon}\right), & p=2\ (q=2),\\ O\left(\frac{(\ln n)^{1/2}n^{1/4}\sqrt{M_2M}R_1}{\varepsilon}\right), & p=1\ (q=\infty).\end{cases}\end{aligned} \tag{D.85}$$

and from (D.84) and (D.85) we obtain the batch size:

$$
\begin{aligned}
B &\overset{\text{(D.84)}}{=} \max\left\{1, \frac{8192ca_{q,n}^2 nM_2^2 r_{p,n}^4 R_p^2}{\varepsilon^2 N}\right\} \\
&\overset{\text{(D.85)}}{=} \max\left\{1, \frac{512\sqrt{2}ca_{q,n}^2 n^{3/4} M_2^{3/2} r_{p,n}^3 R_p}{\varepsilon M^{1/2}}\right\} \\
&= O\left(\max\left\{1, \frac{a_{q,n}^2 n^{3/4} M_2^{3/2} r_{p,n}^3 R_p}{\varepsilon M^{1/2}}\right\}\right) \\
&= \begin{cases} O\left(\max\left\{1, \frac{M_2 D_2}{\varepsilon}\right\}\right), & p = 2 \ (q = 2), \\ O\left(\max\left\{1, \frac{(\ln n)^{5/2} n^{-1/4} M_2^{3/2} R_1}{\varepsilon M^{1/2}}\right\}\right), & p = 1 \ (q = \infty). \end{cases}
\end{aligned}
$$

We also obtain the total number of oracle calls T:

$$
\begin{aligned}
T = N \cdot B &= \max\left\{\frac{8\sqrt{2} n^{1/4}\sqrt{M_2 M} r_{p,n} R_p}{\varepsilon}, \frac{8192ca_{q,n}^2 nM_2^2 r_{p,n}^4 R_p^2}{\varepsilon^2}\right\} \\
&= O\left(\max\left\{\frac{n^{1/4}\sqrt{M_2 M} r_{p,n} R_p}{\varepsilon}, \frac{a_{q,n}^2 nM_2^2 r_{p,n}^4 R_p^2}{\varepsilon^2}\right\}\right) \\
&= \begin{cases} O\left(\max\left\{\frac{n^{1/4} M_2 D_2}{\varepsilon}, \frac{nM_2^2 D_2^2}{\varepsilon^2}\right\}\right), & p = 2 \ (q = 2), \\ O\left(\max\left\{\frac{(\ln n)^{1/2} n^{1/4}\sqrt{M_2 M} R_1}{\varepsilon}, \frac{(\ln n)^3 M_2^2 R_1^2}{\varepsilon^2}\right\}\right), & p = 1 \ (q = \infty). \end{cases}
\end{aligned} \tag{D.86}
$$

□

D.2.3 Proof of Theorem 19

Theorem 19. *Based on the batched Accelerated gradient method, the Smoothing scheme applied to non-smooth and strongly convex problem* (1)*, provides a gradient-free method with*

$$
O\left(\frac{n^{1/4}\sqrt{M_2 M} r_{p,n}}{\sqrt{\mu\varepsilon}}\right) = \begin{cases} O\left(\frac{n^{1/4} M_2}{\sqrt{\mu\varepsilon}}\right), & p = 2 \ (q = 2), \\ O\left(\frac{(\ln n)^{1/2} n^{1/4}\sqrt{M_2 M}}{\sqrt{\mu\varepsilon}}\right), & p = 1 \ (q = \infty) \end{cases}
$$

successive iterations and

$$
O\left(\frac{a_{q,n}^2 nM_2^2 r_{p,n}^4}{\mu\varepsilon}\right) = \begin{cases} O\left(\frac{nM_2^2}{\mu\varepsilon}\right), & p = 2 \ (q = 2), \\ O\left(\frac{(\ln n)^3 M_2^2}{\mu\varepsilon}\right), & p = 1 \ (q = \infty) \end{cases}
$$

zeroth-order oracle calls, where $1/p + 1/q = 1$, $a_{q,n}^2$ *is defined in Lemma 9,* $r_{p,n}^2$ *is defined in Lemma 8.*

Proof: Below we use *restarts scheme* described in [46].

Set $z_1 = x_{start}$ and denote:

$$\rho_1 = R_p = \|z_1 - x^*_{\tau_1}\|_p,$$

where $x^*_{\tau_1} = \arg\min_{x\in Q} f_{\tau_1}(x)$ and $\tau_1 = \frac{\mu R_p^2}{4M_2}$. If we can not calculate R_p, we can take $\rho_1 = R_p \leqslant D_p$, where D_p is a diameter of Q w.r.t. p-norm.

Assume that $\mu \geqslant \frac{\varepsilon}{2R_p^2}$ (otherwise, it is better to use the algorithm for convex but not strongly convex functions).

We do K runs of the Algorithm 3 in the following way. Let us start with the point z_1. Suggest we already did $k-1$ runs of the Algorithm 3 and now we start the k-th run. We set the start point $x_1 = z_k$, the corresponding distance generating function with center at $x_1 = z_k$ (3.1). $\varepsilon_k = \frac{\mu R^2}{2}4^{-k}$ and run $N_k = N(\varepsilon_k)$ iterations (D.85) and $T_k = T(\varepsilon_k)$ oracle calls (D.86). Denote the final point $x^{ag}_{N_k+1}$ of the k-th run as the start point z_{k+1} for the $(k+1)$-th run and so on. After K runs we obtain the final point z_K.

If the noise level Δ_k is sufficiently small (see (D.99) in Appendix D.3):

$$\Delta_k \leqslant \frac{\varepsilon_k^2}{160\sqrt{n}M_2D_2}, \tag{D.87}$$

then from (D.85) and (D.86) we have that to achieve the error ε_k, we need

$$N_{\varepsilon_k} = \frac{8\sqrt{2}n^{1/4}\sqrt{M_2M}r_{p,n}R_k}{\varepsilon_k}$$

iterations and

$$T(\varepsilon_k) = \frac{8192ca^2_{q,n}nM_2^2r^4_{p,n}R_k^2}{\varepsilon_k^2}.$$

oracle calls, where $R_k = \|z_k - x^*_{\tau_k}\|_p$, $x^*_{\tau_k} = \arg\min\limits_{x\in Q} f_{\tau_k}(x)$, $\tau_k = \frac{\varepsilon_k}{2M_2}$

We take

$$\varepsilon_k = \frac{\mu R^2}{2}4^{-(k-1)}, \tag{D.88}$$

$$\varepsilon_K = \varepsilon \implies K = 1 + \frac{\ln\left(\frac{\mu R^2}{2\varepsilon}\right)}{\ln 4}. \tag{D.89}$$

The function f_τ is continuously differentiable and μ-strongly convex on Q: (Lemma 12):

$$f_\tau(z_k) \geqslant f_\tau(x^*_{\tau_k}) + \frac{\mu}{2}\|z_k - x^*_\tau\|_p^2. \tag{D.90}$$

Now we can estimate R_k for Algorithm 3 in strongly convex case:

$$R_k^2 = \|z_k - x^*_{\tau_k}\|_p^2 \overset{(D.90)}{\leqslant} \frac{2}{\mu}\left[f_\tau(z_k) - f_\tau(x^*_{\tau_k})\right] \overset{(3.8)}{\leqslant} \frac{2}{\mu}\left[f(z_k) - f(x^*_{\tau_k}) + \tau_k M_2\right].$$

Taking expectation and using $\tau_k = \frac{\varepsilon_k}{2M_2}$, $\mathbb{E}[f(z_k) - f(x^*_{\tau_k})] \leqslant \varepsilon_{k-1}$ we obtain (for $k \geqslant 2$):

$$\mathbb{E}[R_k^2] \leqslant \frac{2}{\mu}\left[\varepsilon_{k-1} + \frac{\varepsilon_k}{2}\right] \overset{(D.88)}{\leqslant} \frac{2}{\mu}\left(4\varepsilon_k + \frac{1}{2}\varepsilon_k\right) = \frac{9\varepsilon_k}{\mu} \overset{(D.88)}{=} \frac{9}{2}R^2 4^{-(k-1)}. \tag{D.91}$$

For $k = 1$ the estimate (D.91) holds true as $R^2 \leqslant \frac{9}{8}R^2$.

We obtain the number of iterations and the k^{th} restart

$$N_k = N(\varepsilon_k) = \frac{8\sqrt{2}n^{1/4}\sqrt{M_2 M}r_{p,n}R_k}{\varepsilon_k} \overset{\text{(D.88),(D.91)}}{=} \frac{24n^{1/4}\sqrt{M_2 M}r_{p,n}}{\mu R} \cdot 2^{k-1}$$

and the number of oracle calls

$$T_k = T(\varepsilon_k) = \frac{8192ca_{q,n}^2 nM_2^2 r_{p,n}^4 R_k^2}{\varepsilon_k^2} \overset{\text{(D.88),(D.91)}}{=} \frac{73728ca_{q,n}^2 nM_2^2 r_{p,n}^4}{\mu^2 R^2} \cdot 4^{k-1}.$$

We obtain the total number of iterations:

$$\begin{aligned} N = \sum_{k=1}^{K} N_k &\leqslant 2^K \cdot \frac{24n^{1/4}\sqrt{M_2 M}r_{p,n}}{\mu R} = 2\sqrt{\frac{\mu R^2}{2\varepsilon}} \cdot \frac{24n^{1/4}\sqrt{M_2 M}r_{p,n}}{\mu R} \\ &= \frac{24\sqrt{2}n^{1/4}\sqrt{M_2 M}r_{p,n}}{\sqrt{\mu\varepsilon}} = O\left(\frac{n^{1/4}\sqrt{M_2 M}r_{p,n}}{\sqrt{\mu\varepsilon}}\right) \\ &= \begin{cases} O\left(\frac{n^{1/4}M_2}{\sqrt{\mu\varepsilon}}\right), & p = 2 \ (q = 2), \\ O\left(\frac{(\ln n)^{1/2}n^{1/4}\sqrt{M_2 M}}{\sqrt{\mu\varepsilon}}\right), & p = 1 \ (q = \infty). \end{cases} \end{aligned}$$

The total number of oracle calls:

$$\begin{aligned} T = \sum_{k=1}^{K} T_k &\leqslant \frac{2}{4} \cdot 4^K \cdot \frac{73728ca_{q,n}^2 nM_2^2 r_{p,n}^4}{\mu^2 R^2} = 2 \cdot \frac{\mu R^2}{2\varepsilon} \cdot \frac{73728ca_{q,n}^2 nM_2^2 r_{p,n}^4}{\mu^2 R^2} \\ &= \frac{73728ca_{q,n}^2 nM_2^2 r_{p,n}^4}{\mu\varepsilon} = O\left(\frac{a_{q,n}^2 nM_2^2 r_{p,n}^4}{\mu\varepsilon}\right) \\ &= \begin{cases} O\left(\frac{nM_2^2}{\mu\varepsilon}\right), & p = 2 \ (q = 2), \\ O\left(\frac{(\ln n)^3 M_2^2}{\mu\varepsilon}\right), & p = 1 \ (q = \infty). \end{cases} \end{aligned}$$

□

D.3 Noise calculation

In this section we estimate in details the maximum level of noise Δ for the Section 3.10 that does not impact the rate of convergence of Algorithm 3.

To preserve the rate of convergence of Algorithm 3, the noise containing terms in (D.81) must satisfy:

$$\frac{4\sqrt{2}a_{q,n}r_{p,n}R_p}{\sqrt{N}} \cdot \frac{2n\Delta M_2}{\varepsilon} + \frac{\Delta M_2 D_2\sqrt{n}}{\varepsilon} \overset{\text{(D.82)}}{\leqslant} \frac{\varepsilon}{4}. \tag{D.92}$$

Thus, substituting $N(\varepsilon) = \frac{8\sqrt{2}n^{1/4}\sqrt{M_2 M}r_{p,n}R_p}{\varepsilon}$ from (D.85), we obtain:

$$\left(\frac{2^{7/4}a_{q,n}\sqrt{r_{p,n}}\sqrt{R_p}}{\sqrt{n}{M_2}^{1/4}M^{1/4}} \cdot \frac{\sqrt{n}\sqrt{\varepsilon}}{D_2} + 1\right)\Delta M_2 D_2\sqrt{n} \leqslant \frac{\varepsilon^2}{4}.$$

Define

$$z = \frac{a_{q,n}\sqrt{r_{p,n}}\sqrt{R_p}}{D_2 {M_2}^{1/4} M^{1/4}}\sqrt{\varepsilon}. \tag{D.93}$$

Let us get upper bound on z. We recall, that $D_2 = \max\limits_{x,y\in Q}\|x-y\|_2$ is the Euclidean diameter of the set Q, $R_p = \|x_1 - x_\tau^*\|_p$ is the distance from the start point x_1 to the minimizer x_τ^* of the function $f_\tau(x)$ on the set Q: $x_\tau^* \in \arg\min_{x\in Q} f_\tau(x)$, consequently $R_p \leqslant D_p$, where is the diameter of the set Q in p-norm.

We note that $\varepsilon \leqslant MR_p$ for any $p \in [1,2]$ otherwise we can use the starting point x_1 as a solution and not to implement the Algorithm 3:

$$f(x_1) - f(x^*) \leqslant f(x_1) - f(x_\tau^*) \leqslant M\|x_1 - x_\tau^*\|_p = MR_p.$$

Thus, using the last inequation, $R_p \leqslant D_p$ and Lemma 22, we obtain:

$$\left(\frac{R_p^2\varepsilon^2}{D_2^4 M_2 M}\right)^{1/4} \leqslant \left(\frac{R_p^2 M R_p M_2 R_2}{D_2^4 M_2 M}\right)^{1/4} \leqslant \left(\frac{D_p}{D_2}\right)^{3/4} \leqslant n^{\frac{3}{4p}-\frac{3}{8}}.$$

Substituting the last bound into (D.93) and using $\frac{1}{p}+\frac{1}{q} = 1 \Leftrightarrow (p-1)(q-1) = 1$, we obtain:

$$\begin{aligned} z &\leqslant a_{q,n}\sqrt{r_{p,n}} n^{\frac{3}{4p}-\frac{3}{8}} \\ &\overset{Lemma\ 8,9}{\leqslant} \sqrt{\min\{4q-1{,}5\ln n\}} n^{\frac{1}{q}-\frac{1}{2}} \left(\min\left\{\frac{1}{p-1}, e(2\ln n - 1)\right\}\right)^{\frac{1}{4}} n^{\frac{3}{4p}-\frac{3}{8}} \\ &\leqslant \sqrt{\min\{4q-1{,}5\ln n\}} \left(\min\{q-1, e(2\ln n-1)\}\right)^{\frac{1}{4}} n^{\frac{1}{q}-\frac{1}{2}+\frac{3}{4}\left(1-\frac{1}{q}\right)-\frac{3}{8}} \\ &= \sqrt{\min\{4q-1{,}5\ln n\}} \left(\min\{q-1, e(2\ln n-1)\}\right)^{\frac{1}{4}} n^{\frac{1}{4q}-\frac{1}{8}}. \end{aligned} \tag{D.94}$$

Case $2 \leqslant q \leqslant 4$**.** In this case we use $n^\alpha \leqslant 1$ for $\alpha = \frac{1}{4q} - \frac{1}{8} \leqslant 0$:

$$z \leqslant \sqrt{4q-1}\sqrt[4]{q-1}\cdot 1 \leqslant \sqrt{15}\sqrt[4]{3} = \sqrt[4]{675} < 6. \tag{D.95}$$

Case $q > 4$**.** In this case we use $n^{\frac{1}{4q}-\frac{1}{8}} \leqslant n^{-\frac{1}{16}}$, the substitution $t = n^{\frac{1}{4}}$ and the fact that for any $x > 0$ it holds that $e\ln x \leqslant x$:

$$\begin{aligned} z &\leqslant \sqrt{5\ln n}\sqrt[4]{2e\ln n}\cdot n^{-\frac{1}{16}} = \left(50e(\ln n)^3 n^{-\frac{1}{4}}\right)^{\frac{1}{4}} = \left(50e\cdot 64\frac{(\ln t)^3}{t}\right)^{\frac{1}{4}} \\ &= (3200e)^{\frac{1}{4}}\cdot\left(\frac{3\ln t^{\frac{1}{3}}}{t^{\frac{1}{3}}}\right)^{\frac{3}{4}} \leqslant (3200e)^{\frac{1}{4}}\cdot\left(\frac{3}{e}\right)^{\frac{3}{4}} = \left(\frac{3200\cdot 27}{e^2}\right)^{\frac{1}{4}} < 11. \end{aligned} \tag{D.96}$$

Combining (D.95) and (D.96), we obtain $z \leqslant 11$ for $p \in [1,2]$ ($q \geqslant 2$).

Remark for Section 3.10. It is easy to see from the definitions of z, Δ_1, Δ_2 that

$$\frac{\Delta_1}{\Delta_2} = z \leqslant 11, \tag{D.97}$$

where Δ_1 and Δ_2 are defined in Section 3.10, z id defined in (D.93).

Now we can take the maximum noise level:

$$\Delta_{\max} = \frac{\varepsilon^2}{160\sqrt{n}M_2 D_2} \tag{D.98}$$

and check that if $\Delta \leqslant \Delta_{\max}$, then (D.92) holds true:

$$\frac{4\sqrt{2}a_{q,n}r_{p,n}R_p}{\sqrt{N}} \cdot \frac{2n\Delta M_2}{\varepsilon} + \frac{\Delta M_2 D_2\sqrt{n}}{\varepsilon} \overset{\text{(D.93)}}{=} \left(1 + 2^{\frac{7}{4}}z\right)\frac{\Delta M_2 D_2\sqrt{n}}{\varepsilon} < 40\frac{\Delta_{\max} M_2 D_2\sqrt{n}}{\varepsilon}$$
$$\overset{\text{(D.98)}}{\leqslant} 40\frac{\varepsilon^2}{160\sqrt{n}M_2D_2} \cdot \frac{M_2D_2\sqrt{n}}{\varepsilon} = \frac{\varepsilon}{4}.$$

where we used $2^{\frac{7}{4}} \cdot z + 1 < 2^{\frac{7}{4}} \cdot 11 + 1 < 40$.

Finally, we obtained the maximum noise level

$$\Delta \leqslant \Delta_{\max} \overset{\text{(D.98)}}{\leqslant} \frac{\varepsilon^2}{160\sqrt{n}M_2D_2} = O\left(\frac{\varepsilon^2}{\sqrt{n}M_2D_2}\right). \tag{D.99}$$

For strongly-convex case, at the k-th start of Algorithm 3: $\varepsilon_k \geqslant \varepsilon$ (see (D.89), (D.88)), consequently, the most restrictive noise level is attained at the last iteration and is equal to (D.99).

□