INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2023IPPAG001

Thèse de doctorat

# Derivative-free stochastic optimization, online learning and fairness

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 3 février 2023, par

**Aria Akhavanfoomani**

Composition du Jury :

| | |
|---|---|
| Vianney Perchet<br>Professeur, CREST, ENSAE, IP Paris | Président |
| Anatoli Juditsky<br>Professeur, Université Grenoble Alpes | Rapporteur |
| Aurélien Garivier<br>Professeur, Ecole Normale Supérieure de Lyon (UMPA) | Rapporteur |
| Gábor Lugosi<br>Professeur, Pompeu Fabra University and Barcelona School of Economics | Examinateur |
| Alexandra Carpentier<br>Professeur, Universität Potsdam | Examinatrice |
| Nicolas Flammarion<br>Professeur, École Polytechnique Fédérale de Lausanne | Examinateur |
| Alexandre B. Tsybakov<br>Professeur, CREST, ENSAE, IP Paris | Directeur de thèse |
| Massimiliano Pontil<br>Professeur, Istituto Italiano di Tecnologia and University College London | Co-directeur de thèse |

# Acknowledgment

I wish to thank my supervisors Alexandre Tsybakov and Massimiliano Pontil, for giving me the chance to work with them, for all the times that they trusted me in my ups and downs, and for all the encouragement that they gave me to grow independent. Sasha, you are my scientific idol. It is a great honor for me to be your student. Benefiting from your limitless experience and knowledge is a lifetime achievement for me. Massi, you made the Genovese lab like a home for me and your support from our first encounter has sheltered me against all of my mind's shadows. Thank you for all the succour and helping me build my self confidence.

Many thanks to Anatoli Judistky and Aurelien Garivier for the time that they spent reading my manuscript and for their detailed and constructive reports. I also want to express my gratitude to Gabor Lugosi, Alexandra Carpentier, Vianney Perchet, and Nicolas Flammarion for accepting to be members of my jury. It is an honor and a pleasure to present my work in your presence.

Evgenii Chzhen, to describe and picture your role in my life, I couldn't find any word in any language that I'm familiar with. Thanks to Lu, I found out that there is a word in Chinese that explains it extremely well. To my only 师兄(shixiong): for all the seconds that you spent teaching me, for all the moments that you spent motivating me, and finally for all the hours that we laughed and laughed and laughed out loud, "I hate you"!

I would like to thank my co-authors Riccardo Grazzi and Davit Gogolashvili. Riccardo, if it was not because of your extreme obsession for the problem, we would never have managed to carry on and finish the work. Davit, thanks for all the insightful scientific discussions that we had while playing backgammon and eating focaccia. Thank you both for being amazing friends and astonishing collaborators.

My extreme appreciation to Saverio Salzo, Arnak Dalalyan, Cristina Butucea, Victor Emanuel Brunel, Simo Ndaoud, and Jaouad Mourtada for all the time that they spent to teach me, consult me, and for their absolute support.

Amir and Avo, my Iranian and Armenian officemates who were initially a great remedy for my homesickness and who became my best friends: thanks for always being there for me.

From the genovese lab, Dimitri, Isak, Leonardo, and from ENSAE, Arshak, Flore, Julien, and Solenne: all the moments that we hanged out and discussed were a great pleasure. Your company created an extremely welcoming and friendly atmosphere.

I would like to thank my parents Artemis and Mehdi, my cousins, Alireza, Amir Hosein, and Hengameh, and my friends outside of the academia, Shiva, Jalal, Samad, Siavash, and Hedieh. I tried to demonstrate my gratitude deservingly, but I can not express my feelings. I can only say thank you for all of the momentos.

Finally, I wish to express my gratitude to Niloufar. In our eleven years epic friendship, in all of my life's extreme points, you were there for me. None of these, literally none of these, would be possible without you by my side. Thank you for everything.

# Chapter 1

# Introduction

## 1.1 Optimization

Optimization is a branch of study where the goal is to estimate an extremal quantity associated with some function. Examples of such extremal quantities include a minimizer, a maximizer, a saddle point, and other. In this section, we introduce and study one of the most common models where the extremal quantity is a minimizer of a function $f : \mathbb{R}^d \to \mathbb{R}$ in a closed, convex subset $\Theta$ of $\mathbb{R}^d$. Namely, we are interested in estimating

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \Theta} f(\mathbf{x}) \enspace , \tag{1.1}$$

by some potentially random $\hat{\mathbf{x}}$, which guarantees a small *optimization error*, i.e.

$$\mathbf{E}\left[f(\hat{\mathbf{x}})\right] - f(\mathbf{x}^*) \enspace . \tag{1.2}$$

where the expectation in (1.2) is with respect to the probability distribution of $\hat{\mathbf{x}}$. Of course, without specifying the type of information that we can access about the function $f$, the above stated problem is hopeless. Below, we present a rather general query model, which includes many popular observation schemes.

Consider an iterative procedure, such that at each time $t \geq 1$, for any choice $\mathbf{x} \in \mathbb{R}^d$ of the learner, the nature outputs a noisy information about the function, encoded by the link function $F(\mathbf{x}, \xi(\mathbf{x}))$, where $\xi(\mathbf{x})$ is a measurement noise. Formally, there exists $F : \mathbb{R}^d \times \mathbb{R}^{\ell_1} \to \mathbb{R}^{\ell_2}$, for some positive integers $\ell_1, \ell_2$, such that for every vector $\mathbf{x} \in \mathbb{R}^d$ selected by the learner, the nature samples noise variable $\xi(\mathbf{x})$ and returns $F(\mathbf{x}, \xi(\mathbf{x}))$ to the learner.

The above framework is rather abstract and the concrete problem and estimation strategy will depend on the form of the link function $F$. Let us provide some examples of link functions $F$ and relate them to the well-known settings in the optimization literature. First, we provide the definition for the sub-gradient of a function.

**Definition 1.1.1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$, and fix $\boldsymbol{x} \in \mathbb{R}^d$. We call $\partial f(\boldsymbol{x}) \subseteq \mathbb{R}^d$ the set of sub-gradients (or sub-differentials) of $f$ at point $\boldsymbol{x}$, if for any $\boldsymbol{g} \in \partial f(\boldsymbol{x})$, and $\boldsymbol{y} \in \mathbb{R}^d$, we have*

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \leq \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{y} \rangle \enspace .$$

If $f$ is a convex function, the above definition is a generalization of the gradient of $f$. Particularly, if $f$ is convex and differentiable at $\mathbf{x} \in \mathbb{R}^d$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$ (see Lemma 1.2.2).

**Example 1.1.2** (First-order optimization)**.** *We call an optimization problem a first-order problem if the link function $F : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ and the evaluation at point $\boldsymbol{x} \in \mathbb{R}^d$ contains information about a sub-gradient of the function at point $\boldsymbol{x}$. The simplest case is when $F(\boldsymbol{x}, \xi(\boldsymbol{x})) \in \partial f(\boldsymbol{x})$.*

*In the stochastic setting with a differentiable objective function $f$, the most well-know case is $\mathbf{E}[F(\boldsymbol{x}, \xi(\boldsymbol{x}))] = \nabla f(\boldsymbol{x})$ (Robbins and Monro, 1951). A particular example is the additive noise model, where $F(\boldsymbol{x}, \xi(\boldsymbol{x})) = \nabla f(\boldsymbol{x}) + \xi(\boldsymbol{x})$.*

**Example 1.1.3** (Zero-order optimization)**.** *An optimization problem is called a zero-order problem if for any $\boldsymbol{x} \in \mathbb{R}^d$, the link function $F : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ provides information about the function values. As an example one can consider the additive noise model $F(\boldsymbol{x}, \xi(\boldsymbol{x})) = f(\boldsymbol{x}) + \xi(\boldsymbol{x})$. This is the main case studied below. Throughout this thesis, whenever we mention that the learner has access to zero-order information we are referring to $F(\boldsymbol{x}, \xi(\boldsymbol{x})) = f(\boldsymbol{x}) + \xi(\boldsymbol{x})$.*

In the following three sections, we provide a brief background on convex analysis and convex optimization.

**Notation and conventions.** We let $\langle \cdot, \cdot \rangle$ be the standard inner product in $\mathbb{R}^d$, and for $q \in [1, \infty]$, we denote by $\|\cdot\|_q$ the $\ell_q$-norm. For $k \geq 1$, we let $[k]$ the set of all the positive integers that are less or equal to $k$. For $q \in [1, \infty]$ we introduce the open $\ell_q$-ball and $\ell_q$-sphere respectively as

$$\mathcal{B}_q^d \triangleq \left\{\mathbf{x} \in \mathbb{R}^d \ : \ \|\mathbf{x}\|_q < 1\right\} \qquad \text{and} \qquad \partial B_q^d \triangleq \left\{\mathbf{x} \in \mathbb{R}^d \ : \ \|\mathbf{x}\|_q = 1\right\} \enspace .$$

For $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we write $\mathbf{x} \geq 0$ if $x_i \geq 0$, for all $i \in [d]$. Furthermore, for $d$-dimensional multi-index $\boldsymbol{m} = (m_1, \ldots, m_d)$, where $m_j \geq 0$ are integers, we define $|\boldsymbol{m}| = m_1 + \ldots + m_d$, $\boldsymbol{m}! = m_1! \ldots m_d!$, and for any $\boldsymbol{u} \in \mathbb{R}^d$, let $\boldsymbol{u}^{\boldsymbol{m}} = u_1^{m_1} \ldots u_d^{m_d}$. We denote the differentiation operator as $D^{\boldsymbol{m}} = \frac{\partial^{|\boldsymbol{m}|}}{\partial u_1^{m_1} \ldots \partial u_d^{m_d}}$. Also, throughout this manuscript we adopt the convention that $1/\infty = 0$, and $0 \log(0) = 0$.

## 1.2 Some facts from convex analysis

In this section, we briefly recall some basic facts from convex analysis.

**Definition 1.2.1.** *We call $f : \mathbb{R}^d \to \mathbb{R}$ a convex function, if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\lambda \in [0, 1]$, it satisfies*

$$f\left(\lambda \boldsymbol{x} + (1 - \lambda)\, \boldsymbol{y}\right) \leq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y}) \enspace .$$

One of the main reasons why convex functions are interesting from an algorithmic point of view is that a local property of these functions can be translated to global phenomena. We illustrate this bridge between local and global properties of convex functions in the two following lemmas.

**Lemma 1.2.2.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. This is true if and only if, for any $\boldsymbol{x} \in \mathbb{R}^d$, $\partial f(\boldsymbol{x}) \neq \emptyset$. Furthermore, if $f$ is differentiable at $\boldsymbol{x}$, then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$.*

*Proof.* The first and second statement of lemma can be found, respectively, in (Bubeck, 2015, Proposition 1.1) and (Polyak, 1987, Chapter 5, Lemma 5). □

We will distinguish between unconstrained minimization which corresponds to $\Theta = \mathbb{R}^d$ in (1.1), and constrained minimization when $\Theta$ in (1.1) is a compact and convex set. One of the most useful properties of a convex function in the setting of unconstrained optimization is the fact that a local minimum is the global minimum. Also, note that for any convex function $f$, we have $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, if and only if $0 \in \partial f(\mathbf{x}^*)$. Now, assume that $f$ is a convex and differentiable function. Then by Lemma 1.2.2, we deduce that

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \quad \text{if and only if} \quad \nabla f(\mathbf{x}^*) = 0 \enspace .$$

A general version of *optimality condition* is stated in the following lemma, which is valid for a closed and convex $\Theta$.

**Lemma 1.2.3.** *[(Nesterov, 2018, Theorem 3.1.24)] Let $\Theta$ be a closed and convex subset of $\mathbb{R}^d$. Then, for convex function $f : \mathbb{R}^d \to \mathbb{R}$ we have*

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

*if and only if*

$$\partial f(\boldsymbol{x}^*) \cap \mathcal{C}_\Theta(\boldsymbol{x}^*) \neq \emptyset \ ,$$

*where we introduced the cone* $\mathcal{C}_\Theta(\boldsymbol{x}^*) = \{\boldsymbol{g} \in \mathbb{R}^d : \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}^* \rangle \geq 0, \ \textit{for all } \boldsymbol{y} \in \Theta\}$.

## 1.3 Convex optimization: sub-gradient descent

First, we introduce the *sub-gradient descent* algorithm, which is the most fundamental algorithm in convex optimization. An initial version of *gradient descent* can be traced back to (Cauchy, 1847), which had been defined for unconstrained optimization of differentiable functions. Here, we consider a more general (projected and with sub-gradients) version of the method. Starting from an initial point $\mathbf{x}_1$, *sub-gradient descent* is the following iterative procedure

$$\mathbf{x}_{t+1} = \mathrm{Proj}_\Theta\left(\mathbf{x}_t - \eta_t \mathbf{g}_t\right), \qquad t = 1, 2, \ldots \ , \tag{1.3}$$

where $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$, $\eta_t > 0$[1]. Here and in what follows $\Theta$ is a closed and convex set and we introduce $\mathrm{Proj}_\Theta(\mathbf{x}) = \arg\min_{\mathbf{y} \in \Theta} \|\mathbf{x} - \mathbf{y}\|_2$. If $f$ is differentiable and convex, by Lemma 1.2.2 we have $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$ and one can show that for the updates in (1.3), we have $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$. The word *descent* in the name of the algorithm is due to this property. However, in the current formulation of the algorithm where we use $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$, there is no such guarantee and the function values may increase (see (Orabona, 2019, Example 6.1.)). In this thesis, our main object of interest is a generalization of (1.3), where the updates are as follows

$$\mathbf{x}_{t+1} = \mathrm{Proj}_\Theta\left(\mathbf{x}_t - \eta_t \tilde{\mathbf{g}}_t\right), \qquad t = 1, 2, \ldots \ , \tag{1.4}$$

and $\tilde{\mathbf{g}}_t$ is a random vector in $\mathbb{R}^d$ that the learner forms at round $t$ based on zero or first-order information as an alternative for $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$. In the following lemma, we state a bound on the optimization error of Algorithm (1.4) when $f$ is a convex function.

**Lemma 1.3.1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and assume that $\boldsymbol{x}_t$ is generated by* (1.4) *Then, for $t \geq 1$, we have*

$$\begin{aligned}\mathbf{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)\right] \leq \frac{1}{2\eta_t}\left(\mathbf{E}\left[\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2^2\right] - \mathbf{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|_2^2\right]\right) + \frac{\eta_t}{2}\mathbf{E}\left[\|\tilde{\boldsymbol{g}}_t\|_2^2\right] \\ + \mathbf{E}\left[\inf_{\boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)} \langle \boldsymbol{g}_t - \mathbf{E}\left[\tilde{\boldsymbol{g}}_t | \boldsymbol{x}_t\right], \boldsymbol{x}_t - \boldsymbol{x}^* \rangle\right] \ .\end{aligned} \tag{1.5}$$

[1] Throughout the introduction, we adopt the convention that $\eta_t$ is non-random. However, this convention is not valid in Chapter 4, in which we study an adaptive scheme in the context of online learning.

*Proof.* Since $f$ is a convex function, for any $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ we have

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle \\ &= \langle \mathbf{g}_t - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle + \langle \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle \ . \end{aligned} \tag{1.6}$$

Furthermore, for the term $\langle \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle$ by the definition of (1.4), we have

$$\begin{aligned} \langle \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle &= \frac{1}{\eta_t} \langle \mathbf{x}_t - (\mathbf{x}_t - \eta_t \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right]), \mathbf{x}_t - \mathbf{x}^* \rangle \\ &= \frac{1}{2\eta_t} \mathbf{E}\left[ \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 \|\tilde{\mathbf{g}}_t\|_2^2 - \|(\mathbf{x}_t - \eta_t \tilde{\mathbf{g}}_t) - \mathbf{x}^*\|_2^2 \, | \mathbf{x}_t \right] \ , \end{aligned} \tag{1.7}$$

where the last display is obtained by the inequality $2\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a} - \boldsymbol{b}\|^2$, for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$. On the other hand, by the contraction property of the Euclidean projection $\mathrm{Proj}_\Theta(\cdot)$, we get

$$\|(\mathbf{x}_t - \eta_t \tilde{\mathbf{g}}_t) - \mathbf{x}^*\|_2^2 \geq \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \ . \tag{1.8}$$

By combining (1.6), (1.7) and (1.8) we deduce that

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq &\frac{1}{2\eta_t} \left( \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \mathbf{E}\left[ \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \, | \mathbf{x}_t \right] \right) \\ &+ \leq \langle \mathbf{g}_t - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle + \frac{\eta_t}{2} \mathbf{E}\left[ \|\tilde{\mathbf{g}}_t\|_2^2 \, | \mathbf{x}_t \right] . \end{aligned} \tag{1.9}$$

Recall that (1.9) is valid for any $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$. Taking infimum over all $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$, and total expectation of both sides of the inequality yields

$$\begin{aligned} \mathbf{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq &\frac{1}{2\eta_t} \left( \mathbf{E}\left[ \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \right] - \mathbf{E}\left[ \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \right] \right) + \frac{\eta_t}{2} \mathbf{E}\left[ \|\tilde{\mathbf{g}}_t\|_2^2 \right] \\ &+ \mathbf{E}\left[ \inf_{\mathbf{g}_t \in \partial f(\mathbf{x}_t)} \langle \mathbf{g}_t - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle \right] . \end{aligned}$$

□

As an application of the bound provided in (1.5), let us consider the case where at each round $t \in [T]$ the learner has access to a random vector $\tilde{\mathbf{g}}_t \in \mathbb{R}^d$, such that

$$\textbf{i)}\ \mathbf{E}[\tilde{\mathbf{g}}_t | \mathbf{x}_t] \in \partial f(\mathbf{x}_t) \text{ almost surely} \ , \quad \text{and} \quad \textbf{ii)}\ \mathbf{E}[\|\tilde{\mathbf{g}}_t\|_2^2] \leq L^2, \text{ for } L > 0 \ . \tag{1.10}$$

Then, it follows from (1.5) that

$$\mathbf{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{1}{2\eta_t} \left( \mathbf{E}\left[ \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \right] - \mathbf{E}\left[ \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \right] \right) + \frac{\eta_t}{2} L^2 \ . \tag{1.11}$$

Moreover, assume that $\eta_1 = \cdots = \eta_T = \eta$, and $\mathbf{x}_1$ is non-random. By summing up both sides of (1.11) from $1$ to $T$ we get

$$\sum_{t=1}^{T} \mathbf{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta}{2}L^2 \ .$$

In the case of constrained optimization, we have $\|\mathbf{x}_1 - \mathbf{x}^*\|_2 \leq R$, where $R = \max_{\mathbf{x},\mathbf{y}\in\Theta} \|\mathbf{x} - \mathbf{y}\|_2$. By letting $\eta = \frac{R}{L\sqrt{T}}$, we get

$$\mathbf{E}\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*)\right] \leq \frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{RL}{\sqrt{T}} \ ,$$

where $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$, and we have used the convexity of $f$. Similarly, in the case of unconstrained optimization by letting $\eta = \frac{1}{L\sqrt{T}}$, we get

$$\mathbf{E}\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*)\right] \leq \frac{L}{2\sqrt{T}}\left(\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 + 1\right) \ .$$

We note that here we take an averaged version $\bar{\mathbf{x}}_T$ of sub-gradient descent. Such averaged techniques were first introduced by Polyak and Juditsky (1992), and they are widely used now.

Now, assume that at each round $t \in [T]$, the learner has access to a non-random vector $\tilde{\mathbf{g}}_t \in \partial f(\mathbf{x}_t)$. Then condition **i**) in obviously (1.10) holds. However, condition **ii**) in (1.10) is a further assumption on the objective function $f$.

**Definition 1.3.2.** *For $L > 0$, $q \in [1, \infty]$, we call $f : \mathbb{R}^d \to \mathbb{R}$ a $L$-Lipschitz function with respect to $\|\cdot\|_q$, if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, it satisfies*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_q \ .$$

One can conclude that if $f$ is a $L$-Lipschitz function, with respect to $\|\cdot\|_q$, then for any $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{g} \in \partial f(\mathbf{x})$, we have $\|\mathbf{g}\|_{q^*} \leq L$, where $1/q^* + 1/q = 1$. Therefore, assuming that $f$ is a $L$-Lipschitz function with respect to $\|\cdot\|_2$ is a sufficient condition that ensures property **ii**) in (1.10). We summarize these observations in the following corollary.

**Corollary 1.3.3.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and $L$-Lipschitz function with respect to $\|\cdot\|_2$. Assume that at each round $t \in [T]$ the learner has access to a deterministic vector $\boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)$. Let $\boldsymbol{x}_t$ to be generated by the updates in (1.4), with $\eta_t = \frac{1}{L\sqrt{T}}$, for $t \in [T]$. Then,*

$$f(\bar{\boldsymbol{x}}_T) - f(\boldsymbol{x}^*) \leq \frac{L}{2\sqrt{T}}\left(\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|_2^2 + 1\right) \ .$$

*Moreover, in the case of constrained optimization with $R = \max_{\boldsymbol{x},\boldsymbol{y}\in\Theta} \|\boldsymbol{x} - \boldsymbol{y}\|$, by letting $\eta_t =$*

$\frac{R}{L\sqrt{T}}$ *we get*

$$f(\bar{\boldsymbol{x}}_T) - f(\boldsymbol{x}^*) \leq \frac{RL}{\sqrt{T}} \quad .$$

Let us pursue our analysis with a more restrictive assumption on the curvature of the objective function $f$.

**Definition 1.3.4.** *For $\alpha > 0$, we call $f : \mathbb{R}^d \to \mathbb{R}$ an $\alpha$-strongly convex function if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\boldsymbol{g} \in \partial f(\boldsymbol{x})$ it satisfies*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) - \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\alpha}{2} \left\| \boldsymbol{x} - \boldsymbol{y} \right\|^2 \quad .$$

Clearly, any $\alpha$-strongly convex function is convex. The following lemma is a result similar to Lemma 1.3.1, which gives a tighter upper bound for the optimization error when $f$ is strongly convex.

**Lemma 1.3.5.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an $\alpha$-strongly convex function. Assume that $\boldsymbol{x}_t$ is generated by (1.4). Then, we have*

$$\begin{aligned}\mathbf{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)\right] \leq \frac{1}{2\eta_t} &\left(\left(1 - \frac{\eta_t \alpha}{2}\right) \mathbf{E}\left[\left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|_2^2\right] - \mathbf{E}\left[\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\right\|_2^2\right]\right) \\ &+ \frac{\eta_t}{2}\mathbf{E}\left[\left\|\tilde{\boldsymbol{g}}_t\right\|_2^2\right] + \frac{1}{\alpha}\mathbf{E}\left[\inf_{\boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)} \left\|\boldsymbol{g}_t - \mathbf{E}\left[\tilde{\boldsymbol{g}}_t | \boldsymbol{x}_t\right]\right\|_2^2\right] \quad .\end{aligned} \tag{1.12}$$

*Proof.* Since $f$ is an $\alpha$-strongly function, for any $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ we can write

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle - \frac{\alpha}{2} \left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 \\ &= \langle \mathbf{g}_t - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle + \langle \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle - \frac{\alpha}{2} \left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 \\ &\leq \frac{1}{\alpha} \left\| \mathbf{g}_t - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right] \right\|_2^2 + \langle \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle - \frac{\alpha}{4} \left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 \quad , \end{aligned}$$

where the last inequality follows from the fact that

$$\langle \mathbf{g}_t - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^* \rangle \leq \frac{1}{\alpha} \left\| \mathbf{g}_t - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right] \right\|_2^2 + \frac{\alpha}{4} \left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 \quad .$$

Using (1.7), and a similar argument as the one leading to (1.9) we find:

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{1}{2\eta_t} \left(\left(1 - \frac{\eta_t \alpha}{2}\right) \left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \mathbf{E}\left[\left\| \mathbf{x}_{t+1} - \mathbf{x}^* \right\|_2^2 | \mathbf{x}_t\right]\right) + \frac{1}{\alpha} \left\| \mathbf{g}_t - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right] \right\|_2^2 + \\ + \frac{\eta_t}{2} \mathbf{E}\left[\left\| \tilde{\mathbf{g}}_t \right\|_2^2 | \mathbf{x}_t\right] \quad . \end{aligned}$$

By taking infimum over $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ and expectations of both sides we conclude the proof. □

Again as an example, assume that at each round the learner has access to a vector

$\tilde{\mathbf{g}}_t \in \mathbb{R}^d$ such that (1.10) holds. Then, by (1.12) we can write

$$\mathbf{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{1}{2\eta_t}\left(\left(1 - \frac{\eta_t \alpha}{2}\right)\mathbf{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] - \mathbf{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right]\right) + \frac{\eta_t}{2}L^2 \ .$$

Assigning $\eta_t = \frac{4}{\alpha(t+1)}$, and multiplying both sides by $t$ yields

$$\mathbf{E}\left[t\left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right)\right] \leq \frac{\alpha}{8}\left(t(t-1)\mathbf{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] - t(t+1)\mathbf{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right]\right) + \frac{2L^2}{\alpha} \ .$$

Summing both sides from $1$ to $T$ and dividing by $\frac{T(T+1)}{2}$ implies

$$\mathbf{E}\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*)\right] \leq \frac{2}{T(T+1)}\sum_{t=1}^{T}\mathbf{E}\left[t\left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right)\right] \leq \frac{4L^2}{\alpha(T+1)} \ ,$$

where we introduced the weighted average estimator $\bar{\mathbf{x}}_T = \frac{2}{T(T+1)}\sum_{t=1}^{T} t\mathbf{x}_t$. From the above discussion, we conclude that for Algorithm (1.4), the optimization error of a convex Lipschitz in $\|\cdot\|_2$ function is of the order $1/\sqrt{T}$ and for a strongly convex Lipschitz in $\|\cdot\|_2$ function it scales as $1/T$. These are classical results in optimization. We refer to (Bubeck, 2015, Chapter 4) for references and historical review.

In the next section, we present a generalization of the algorithm that is provided in (1.3), which is adaptive to the geometry induced by the function $f$.

## 1.4 Convex optimization: mirror descent and Nesterov's dual averaging

First, we place ourselves in the setting of first order optimization where at each round the learner has access to a deterministic vector which is a sub-gradient of convex function $f$ at the current update. Namely, at round $t$ the learner observes a deterministic vector $F(\mathbf{x}_t, \xi(\mathbf{x}_t)) \in \partial f(\mathbf{x}_t)$ and defines $\tilde{\mathbf{g}}_t = F(\mathbf{x}_t, \xi(\mathbf{x}_t))$. Note that the bounds that we derived in (1.5) and (1.12) contain the term $\mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_2^2\right]$. As we outlined in Lemma 1.3.3, if the objective function $f$ is $L$-Lipschitz the term $\|\tilde{\mathbf{g}}_t\|_2^2$ can be uniformly bounded. Now, assume that $f$ is $L$-Lipschitz with respect to the $\ell_1$-norm, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_1 \ .$$

It implies that $\|\tilde{\mathbf{g}}_t\|_\infty \leq L$ and consequently the only guarantee that we can get for $\|\tilde{\mathbf{g}}_t\|_2$ is that it is uniformly bounded by $\sqrt{d}L$, which leads to a sub-optimal rate for the optimization error in terms of dependency on the dimension $d$. In other words, the sub-gradient descent algorithm (1.3) cannot be adopted to the underlying geometry that is induced by the objective function $f$, and it is suitable only if $f$ is $L$-Lipschitz with respect to $\ell_2$-norm. In this section we introduce the *mirror descent* algorithm that is initially proposed by Nemirovsky and Yudin

(1983) to overcome this issue. From now on in this chapter, we assume that the objective function $f$ is $L$-Lipschitz with respect to the $\ell_q$-norm, for some $q \in [1, \infty]$.

Let $\tilde{\Theta} \subseteq \mathbb{R}^d$ be an open set such that $\Theta \subseteq \tilde{\Theta}$. Assume that function $V : \tilde{\Theta} \to \mathbb{R}$ is such that the following conditions hold.

1. $V$ is differentiable on $\Theta$.
2. $V$ is a $1$-strongly convex function on $\Theta$ with respect to the $\|\cdot\|_q$ norm, for some $q \in [1, \infty]$ i.e., for any $\mathbf{x}, \mathbf{y} \in \Theta$, we have

$$V(\mathbf{x}) \geq V(\mathbf{y}) + \langle \nabla V(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_q^2 \ .$$

We define the Bergman divergence function with respect to $V$ as $B_V : \Theta \times \Theta \to \mathbb{R}$ such that

$$B_V(\mathbf{x}; \mathbf{y}) = V(\mathbf{x}) - V(\mathbf{y}) - \langle \nabla V(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ .$$

Note that since we assume that $V$ is $1$-strongly convex with respect to the $\|\cdot\|_q$ norm we have

$$B_V(\mathbf{x}, \mathbf{y}) \geq \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_q^2 \ .$$

The mirror descent algorithm is the following iterative procedure

$$\mathbf{x}_{t+1} = \left( \operatorname*{arg\,min}_{\mathbf{x} \in \Theta} \eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + B_V(\mathbf{x}; \mathbf{x}_t) \right), \qquad t \in [T] \ . \tag{1.13}$$

For example, let $V(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|_2^2$. Then, $B_V(\mathbf{x}; \mathbf{y}) = \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$, and for (1.13) we can write

$$\begin{aligned} \mathbf{x}_{t+1} &= \operatorname*{arg\,min}_{\mathbf{x} \in \Theta} \left( \eta_t \langle \mathbf{g}_t, \mathbf{x} \rangle + \frac{1}{2} \|\mathbf{x}_t - \mathbf{x}\|_2^2 \right) \\ &= \operatorname*{arg\,min}_{\mathbf{x} \in \Theta} \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}\|_2^2 = \mathrm{Proj}_{\Theta}\left(\mathbf{x}_t - \eta_t \mathbf{g}_t\right) \ . \end{aligned}$$

Therefore, we conclude that the sub-gradient descent algorithm (1.3) is a special case of mirror descent (1.13). Now again assume that, at reach round $t \in [T]$, the learner has access to a random vector $\tilde{\mathbf{g}}_t \in \mathbb{R}^d$, and consider the following updates

$$\mathbf{x}_{t+1} = \operatorname*{arg\,min}_{\mathbf{x} \in \Theta} \left( \eta_t \langle \tilde{\mathbf{g}}_t, \mathbf{x} \rangle + B_V(\mathbf{x}; \mathbf{x}_t) \right) \ . \tag{1.14}$$

In the following lemma we outline the performance of (1.14) when $f$ is a convex function.

**Lemma 1.4.1.** *Assume that $f$ is a convex function, and for $t \in [T]$ let $\mathbf{x}_t$ be generated by*

(1.14) *with $\eta_t > 0$ such that $\eta_{t+1} \le \eta_t$. Then, we have*

$$\sum_{t=1}^{T} \mathbf{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)\right] \le \max_{t\in[T]} \frac{\mathbf{E}\left[B_V(\boldsymbol{x}^*;\boldsymbol{x}_t)\right]}{\eta_T} + \sum_{t=1}^{T} \mathbf{E}\left[\inf_{\boldsymbol{g}_t\in\partial f(\boldsymbol{x}_t)} \langle \boldsymbol{g}_t - \mathbf{E}\left[\tilde{\boldsymbol{g}}_t|\boldsymbol{x}_t\right], \boldsymbol{x}_t - \boldsymbol{x}^*\rangle\right]$$
$$+ \frac{1}{2}\sum_{t=1}^{T} \eta_t \mathbf{E}\left[\|\tilde{\boldsymbol{g}}_t\|_{q^*}^2\right] \quad , \tag{1.15}$$

*where $q^*$ is such that $1/q^* + 1/q = 1$. Furthermore, if $\eta_1 = \cdots = \eta_T = \eta$, and $\boldsymbol{x}_1$ is non-random, then*

$$\sum_{t=1}^{T} \mathbf{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)\right] \le \frac{B_V(\boldsymbol{x}^*;\boldsymbol{x}_1)}{\eta} + \sum_{t=1}^{T} \mathbf{E}\left[\inf_{\boldsymbol{g}_t\in\partial f(\boldsymbol{x}_t)} \langle \boldsymbol{g}_t - \mathbf{E}\left[\tilde{\boldsymbol{g}}_t|\boldsymbol{x}_t\right], \boldsymbol{x}_t - \boldsymbol{x}^*\rangle\right]$$
$$+ \frac{\eta}{2}\sum_{t=1}^{T} \mathbf{E}\left[\|\tilde{\boldsymbol{g}}_t\|_{q^*}^2\right] \quad . \tag{1.16}$$

*Proof.* Since $f$ is convex we can write

$$\sum_{t=1}^{T} \mathbf{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \le \sum_{t=1}^{T} \mathbf{E}\left[\langle \mathbf{E}\left[\tilde{\mathbf{g}}_t|\mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^*\rangle\right] + \sum_{t=1}^{T} \mathbf{E}\left[\inf_{\mathbf{g}_t\in\partial f(\mathbf{x}_t)} \langle \mathbf{g}_t - \mathbf{E}\left[\tilde{\mathbf{g}}_t|\mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}^*\rangle\right] \quad . \tag{1.17}$$

We conclude the proof by using (Orabona, 2019, Theorem 6.8.). □

Note that in the above bound, we get $\mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_{q^*}^2\right]$ instead of $\mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_2^2\right]$, that appeared in (1.5). Thus, for any $q \in [1,\infty]$, if $f$ is a $L$-Lipschitz function with respect to $\|\cdot\|_q$, it is reasonable to take $V$ that is strongly convex with respect to the $\|\cdot\|_q$-norm. Some classical examples of functions $V$ are as follows.

**Example 1.4.2.** *Let $\Theta$ be a convex subset of $\mathbb{R}^d$, and $q \in (1,2]$. Then, $V(\boldsymbol{x}) = \frac{1}{2(q-1)}\|\boldsymbol{x}\|_q^2$ is $1$-strongly convex with respect to $\|\cdot\|_q$ and differentiable on $\Theta$.*

**Example 1.4.3.** *Let $\Theta = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_1 = 1, \quad \boldsymbol{x} \ge 0\}$. Then, $V(\boldsymbol{x}) = \sum_{i=1}^d x_i \log(x_i)$ is $1$-strongly convex with respect to $\|\cdot\|_1$ and differentiable on $\Theta$. Moreover, for any $\boldsymbol{x}, \boldsymbol{y} \in \Theta$, we have $B_V(\boldsymbol{x};\boldsymbol{y}) = \sum_{i=1}^d x_i \log(\frac{x_i}{y_i})$.*

If $\arg\min_{\mathbf{x}\in\tilde{\Theta}} V(\mathbf{x}) \in \Theta$, then by assigning $\mathbf{x}_1 = \arg\min_{\mathbf{x}\in\tilde{\Theta}} V(\mathbf{x})$, the term $B_V(\mathbf{x}^*;\mathbf{x}_1)$ in (1.16) becomes equal to $V(\mathbf{x}^*) - \min_{\mathbf{x}\in\Theta} V(\mathbf{x})$. Then, in the setting of Example 1.4.2 the term $B_V(\mathbf{x}^*;\mathbf{x}_1) = V(\mathbf{x}^*) - \min_{\mathbf{x}\in\Theta} V(\mathbf{x})$ is bounded by a constant independent of the dimension, and in Example 1.4.3, we have $B_V(\mathbf{x}^*;\mathbf{x}_1) = V(\mathbf{x}^*) - \min_{\mathbf{x}\in\Theta} V(\mathbf{x}) \le \log(d)$. However, Algorithm (1.14) has a drawback in the case of varying $\eta_t$. In Example 1.4.3 there is no guarantee that $\mathbf{x}_t$ is bounded away from the vertices of $\Theta$, which means that $B_V(\mathbf{x}^*;\mathbf{x}_t)$ can explode. Therefore, (1.14) is not useful for this example. On the other hand, in the setting of Example 1.4.2, if $\Theta$ is not bounded, once more the term in $\max_{t\in[T]} \mathbf{E}\left[B_V(\mathbf{x}^*;\mathbf{x}_t)\right]$ in (1.15) might be big as there is no uniform upper bound on the norm of $\mathbf{x}_t$. This drawback is crucial in the schemes where the learner is not aware of the parameter $T$ or the Lipschitz constant $L$ since in these schemes one typically needs to use varying $\eta_t$. Particularly, in the context of online learning (Orabona and Pál, 2016, Section 4) proved that mirror descent with varying $\eta_T$ can achieve a linear regret for the aforementioned examples. To overcome this issue, we introduce Nesterov's dual averaging algorithm (see (Orabona, 2019, Section 7.13) for a historical overview) that shares the spirit of mirror descent, in the sense of adaptivity to the geometry of $f$.

Let $V : \Theta \to \mathbb{R}$ satisfy the following conditions.

1. $V$ is semi-lower continuous.

2. $V$ is a $1$-strongly convex function on $\Theta$ with respect to $\|\cdot\|_q$, for $q \in [1, \infty]$.

Equipped with such function $V$ consider the following iterative procedure

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\Theta}{\arg\min} \left( \eta_t \sum_{k=1}^{t} \langle \tilde{\mathbf{g}}_t, \mathbf{x} \rangle + V(\mathbf{x}) \right) \quad , \tag{1.18}$$

where $\eta_t > 0$ and $\tilde{\mathbf{g}}_t$ is a random vector that is received by the learner at round $t$. In the following lemma we state a bound for the optimization error of Algorithm (1.18).

**Lemma 1.4.4.** *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function. Then, for Algorithm (1.18) with $\eta_t > 0$ such that $\eta_{t+1} \le \eta_t$ and non-random $\boldsymbol{x}_1$ we have*

$$\sum_{t=1}^{T} \mathbf{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)\right] \le \frac{V(\boldsymbol{x}^*) - \min_{\boldsymbol{x}\in\Theta} V(\boldsymbol{x})}{\eta_T} + \frac{1}{2}\sum_{t=1}^{T} \eta_t \mathbf{E}\left[\|\tilde{\boldsymbol{g}}_t\|_{q^*}^2\right] \tag{1.19}$$

$$+ \sum_{t=1}^{T} \mathbf{E}\left[\inf_{\boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)} \langle \boldsymbol{g}_t - \mathbf{E}\left[\tilde{\boldsymbol{g}}_t | \boldsymbol{x}_t\right], \boldsymbol{x}_t - \boldsymbol{x}^* \rangle\right] \quad ,$$

*where $1/q^* + 1/q = 1$.*

This lemma follows from (1.17) and (Orabona, 2019, Corollary 7.9.). Note that here we used the fact that $\eta_t$ is non-random, for $t \in [T]$.

## 1.5 Zero-order optimization of Lipschitz functions

In this section, we assume that the objective function $f$ is convex, and the link function provides only noisy zero-order information. Particularly, we assume that for any $\mathbf{x} \in \mathbb{R}^d$ we have $F(\mathbf{x}, \xi(\mathbf{x})) = f(\mathbf{x}) + \xi$, where and $\mathbf{E}[\xi^2] \le \sigma^2$ for $\sigma > 0$. Consider the updates in (1.18), and assume that for any $h > 0$, there exists a *surrogate function* $\mathsf{f}_h(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$, such that $\mathsf{f}_h$ is convex and differentiable, and for any $\mathbf{x} \in \mathbb{R}^d$ we have

$$0 \le \mathsf{f}_h(\mathbf{x}) - f(\mathbf{x}) \le \Delta, \quad \text{and} \quad \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right] = \nabla \mathsf{f}_h(\mathbf{x}_t) \text{ almost surely} \;,$$

where $\Delta > 0$. Then, by using the bound that is provided by (1.19), we get

$$\sum_{t=1}^{T} \mathbf{E}\left[\mathsf{f}_h(\mathbf{x}_t) - \mathsf{f}_h(\mathbf{x}^*)\right] \leq \frac{V(\mathbf{x}^*) - \min_{\mathbf{x}\in\Theta} V(\mathbf{x})}{\eta_T} + \frac{1}{2}\sum_{t=1}^{T} \eta_t \mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_{q^*}^2\right] \ .$$

and accordingly

$$\sum_{t=1}^{T} \mathbf{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{V(\mathbf{x}^*) - \min_{\mathbf{x}\in\Theta} V(\mathbf{x})}{\eta_T} + \frac{1}{2}\sum_{t=1}^{T} \eta_t \mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_{q^*}^2\right] + T\Delta \ .$$

Thus, if $\Delta$ is small enough a meaningful bound for the optimization error of $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$ follows if we provide an adequately tight upper bound for $\mathbf{E}[\|\tilde{\mathbf{g}}_t\|_{q^*}^2]$, and assign properly the step sizes $\eta_t$ for $t \in [T]$. The idea of using a particular surrogate function $\mathsf{f}_h(\cdot)$ is first proposed by (Nemirovsky and Yudin, 1983, Chapter 9.3.2) as an exercise. Below we provide two examples of $\tilde{\mathbf{g}}_t$ and $\mathsf{f}_h(\cdot)$, where in the construction of $\tilde{\mathbf{g}}_t$ we only use zero-order information.

**Gradient estimator based on $\ell_2$-randomization.** At round $t \in [T]$, let $h_t > 0$, and let $\zeta_t^\circ$ be a random vector uniformly distributed on $\partial B_2^d$. Assume that we receive the link function's feedback at the two following points

$$y_t = F(\mathbf{x}_t - h_t\zeta_t^\circ, \xi_t), \quad \text{and} \quad y_t' = F(\mathbf{x}_t + h_t\zeta_t^\circ, , \xi_t') \ ,$$

where $\xi_t$ and $\xi_t'$ are noises. Construct the gradient estimator

$$\tilde{\mathbf{g}}_t = \frac{d}{2h_t}\left(y_t - y_t'\right)\zeta_t^\circ \ . \tag{1.20}$$

An initial version of this estimator, that is constructed only based on one point feedback is proposed by Nemirovsky and Yudin (1983) and further studied by Flaxman et al. (2005). The fact that $\mathbf{E}[\tilde{\mathbf{g}}_t|\mathbf{x}_t] = \nabla \mathsf{f}_h(\mathbf{x}_t)$, where such that for any $\mathbf{x} \in \mathbb{R}^d$,

$$\mathsf{f}_h(\mathbf{x}) = \mathbf{E}\left[f(\mathbf{x} + h\boldsymbol{U})\right],$$

and $\boldsymbol{U}$ is uniformly distributed in $B_2^d$ is stated without proof in Nemirovsky and Yudin (1983) and Flaxman et al. (2005). It is referred to Stokes' theorem in Flaxman et al. (2005). The precise version of Stokes' theorem needed for the estimator to work is stated and proved in Chapter 4. The current form of (1.20) is introduced by Agarwal et al. (2010), and it is further analyzed by Duchi et al. (2015); Novitskii and Gasnikov (2021); Shamir (2017). Furthermore, for $q = 1, 2$, the bound on $\mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_{q^*}^2\right]$, with $\tilde{\mathbf{g}}_t$ given in (1.20) is obtained by (Shamir, 2017, Corollaries 2 and 3) for the functions $f$ that are $L$-Lipschitz with respect to $\|\cdot\|_q$.

**Gradient estimator based on $\ell_1$-randomization.** At round $t \in [T]$, let $h_t > 0$, and let $\zeta_t^\diamond$ be a random vector uniformly distributed on $\partial B_1^d$. Assume that we receive the link function's feedback at the two following points

$$y_t = F(\mathbf{x}_t - h_t\boldsymbol{\zeta}_t^{\diamond}, \xi_t), \quad \text{and} \quad y_t' = F(\mathbf{x}_t + h_t\boldsymbol{\zeta}_t^{\diamond}, \xi_t') \ ,$$

where $\xi_t$ and $\xi_t'$ are noises. Construct the gradient estimator

$$\tilde{\mathbf{g}}_t = \frac{d}{2h_t}\left(y_t - y_t'\right)\operatorname{sign}(\boldsymbol{\zeta}_t^{\diamond}) \ , \tag{1.21}$$

where we introduce $\operatorname{sign} : \mathbb{R}^d \to [-1, 1]^d$ as the component-wise sign function (defined at $0$ as $1$). Consider the function $\mathsf{f}_h : \mathbb{R}^d \to \mathbb{R}$, such that for any $\mathbf{x} \in \mathbb{R}^d$

$$\mathsf{f}_h(\mathbf{x}) = \mathbf{E}\left[f(\mathbf{x} + h\boldsymbol{V})\right],$$

where $\boldsymbol{V}$ is uniformly distributed on $B_1^d$. This estimator is proposed and analyzed in Chapter 4. Similar to the case of gradient estimator based on $\ell_2$-randomization, we use Stokes' theorem to show that $\mathbf{E}[\tilde{\mathbf{g}}_t|\mathbf{x}_t] = \nabla \mathsf{f}_h(\mathbf{x}_t)$ (see Lemma 4.6.1). Moreover, due to the simple form of the algorithm we are able to provide a desired bound for the term $\mathbf{E}[\|\tilde{\mathbf{g}}_t\|_{q^*}^2]$, for any $q \in [1, \infty]$, and any $f$ that is $L$-Lipschitz with respect to $\|\cdot\|_q$. Our analysis is based on a novel weighted Poincaré type inequality, which is proposed in Section 4.6. For the case $q = 1$, the latter estimator outperforms $\ell_2$-randomization estimator in the optimization error, up to a $\sqrt{\log(d)}$ factor, and if $q = 2$ they both lead to the optimal upper bound on the optimization error(see Theorem 4.4.1). Moreover, in terms of required memory, we only need $d$ bits and $1$ float to store (1.21), which is more economic compared to the $\ell_2$-randomization method that needs to store $d$ floats.

In the end of this section, we wish to mention the **gradient estimator based on Gaussian randomization**, which is introduced by Nesterov (2011) and plays an important role in the literature of zero-order optimization. At round $t \in [T]$, let $h_t > 0$, and let $\boldsymbol{\zeta}_t^G$ be a standard Gaussian vector. Assume that we receive the link function's feedback at the two following points

$$y_t = F(\mathbf{x}_t - h_t\boldsymbol{\zeta}_t^{G}, \xi_t), \quad \text{and} \quad y_t' = F(\mathbf{x}_t + h_t\boldsymbol{\zeta}_t^{G}, \xi_t') \ ,$$

where $\xi_t$ and $\xi_t'$ are noises. Construct the gradient estimator

$$\tilde{\mathbf{g}}_t = \frac{1}{2h_t}\left(y_t - y_t'\right)\boldsymbol{\zeta}_t^{G} \ ,$$

and consider the function $\mathsf{f}_h : \mathbb{R}^d \to \mathbb{R}$, such that for any $\mathbf{x} \in \mathbb{R}^d$

$$\mathsf{f}_h(\mathbf{x}) = \mathbf{E}\left[f(\mathbf{x} + h\boldsymbol{\zeta}_t^G)\right].$$

However, since this estimator is beyond the scope of the thesis we do not establish its properties and we refer the reader to Nesterov (2011); Ghadimi and Lan (2013); Nesterov and Spokoiny (2017); Balasubramanian and Ghadimi (2021) and references therein.

## 1.6 Zero-order optimization of highly smooth functions

In this section, we assume that the objective function $f$ is $\alpha$-strongly convex and differentiable. Moreover, for any $\mathbf{x} \in \mathbb{R}^d$, we assume that $F(\mathbf{x}, \xi(\mathbf{x})) = f(\mathbf{x}) + \xi$, where $\xi$ is independent from $\mathbf{x}$, and $\mathbf{E}\left[\xi^2\right] \leq \sigma^2$, for $\sigma > 0$. Let $\tilde{\mathbf{g}}_t$ be a gradient estimator that employs only zero-order information. In order to obtain a bound on the optimization error of sub-gradient descent we need to provide control on the terms

$$\mathbf{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right]\|_2^2\right], \quad \text{and} \quad \mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_2^2\right] \quad . \tag{1.22}$$

Indeed, it follows from (1.3.5) and the fact that $f$ is a convex and differentiable function, by Lemma 1.2.2 we have $\partial f(\mathbf{x}_t) = \{\nabla f(\mathbf{x}_t)\}$.

Constructing an estimator $\tilde{\mathbf{g}}_t$ that gives a small squared bias term $\mathbf{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right]\|_2^2\right]$ can be viewed as a problem of non-parametric estimation of the gradient $\nabla f(\cdot)$ at certain point $\mathbf{x}_t$. From the intuition that comes from the literature on non-parametric statistics, one may wonder if the local function approximation (in this case estimating $\nabla f(\mathbf{x}_t)$) could benefit from a *higher order smoothness* assumption on the objective function $f$. To characterize the notion of *higher order smoothness*, we use the well-known $\beta$-Hölder class of functions, namely $\mathcal{F}_\beta(L)$ (defined in (2.1)), where we restrict our attention to the case $\beta \geq 2$.

In what follows we assume that $f \in \mathcal{F}_\beta(L)$, and consider the following two settings.

**Active scheme.** This is the usual setting in optimization and online learning, where at round $t$ the learner chooses $\mathbf{z} = \mathbf{x}_t$ based on $\{\mathbf{x}_1, F(\mathbf{x}_1, \xi(\mathbf{x}_1)), \cdots, \mathbf{x}_{t-1}, F(\mathbf{x}_{t-1}, \xi(\mathbf{x}_{t-1}))\}$ and observes $F(\mathbf{z}, \xi(\mathbf{z}))$ where $\mathbf{z} \in \mathbb{R}^d$ (or, more restrictively, $\mathbf{z} \in \Theta$). We discuss this setting in Chapters 2, 3, and 5.

**Passive scheme.** We study this setting in Chapter 6, where we are not allowed to choose the query points and has only access to the whole set $\{\mathbf{z}_1, F(\mathbf{z}_1, \xi(\mathbf{z}_1)), \cdots, \mathbf{z}_n, F(\mathbf{z}_n, \xi(\mathbf{z}_n))\}$, with $\mathbf{z}_i \in \mathbb{R}^d$ for $i \in [n]$.

**Gradient estimators for active scheme.** Exploiting higher order smoothness in the active scheme can be traced back to Polyak and Tsybakov (1990), where the authors first suggested to use a smoothing kernel to construct the gradient estimator. Later on, this approach was developed by Dippon (2003a) and Bach and Perchet (2016). In what follows, we study the gradient estimator that is proposed by Bach and Perchet (2016) (Chapters 1 and 5), along with two more examples of gradient estimators that are introduced and analyzed in Chapters 2 and 5. Before introducing these estimators we define an entity that plays a crucial role in the structure of estimators that take advantage of higher order smoothness property, namely, the kernel function $K : [-1, 1] \to \mathbb{R}$, such that

$$\int K(u)\,\mathrm{d}u = 0, \int uK(u)\,\mathrm{d}u = 1, \int u^j K(u)\,\mathrm{d}u = 0, \text{ for } j = 2, \cdots, \ell, \int |u|^\beta |K(u)|\,\mathrm{d}u < \infty \ ,$$

where $\ell$ is the largest integer smaller than $\beta$. For examples of such kernel functions we refer the reader to Polyak and Tsybakov (1990) and Bach and Perchet (2016). After introducing the kernel function $K$, we are ready to present the gradient estimators.

**Smooth gradient estimator based on $\ell_2$-randomization.** At round $t \in [T]$, let $h_t > 0$, let $\boldsymbol{\zeta}_t^\circ$ be a random vector uniformly distributed on $\partial B_2^d$, and $r_t$ be uniformly distributed on the interval $[-1, 1]$. Then, we receive the link function's feedback at two points as follows:

$$y_t = F(\mathbf{x}_t - h_t r_t \boldsymbol{\zeta}_t^\circ, \xi_t), \quad \text{and} \quad y_t' = F(\mathbf{x}_t + h_t r_t \boldsymbol{\zeta}_t^\circ, \xi_t') \ ,$$

where $\xi_t, \xi_t'$ are noises. Define the gradient estimator (see Bach and Perchet (2016)) as follows:

$$\tilde{\mathbf{g}}_t = \frac{d}{2h_t}\left(y_t - y_t'\right)\boldsymbol{\zeta}_t^\circ K(r_t) \ . \tag{1.23}$$

In Chapter 1, we investigate the optimization error of Algorithm (1.4) with $\tilde{\mathbf{g}}_t$ as in (1.23). The analysis provided in Chapter 1 corresponds to the paper Akhavan et al. (2020). It is further refined in a follow-up work by Novitskii and Gasnikov (2021).

An estimator that is based on $\ell_1$ randomization (smooth version of (1.21)) is introduced in Chapter 5.

**Smooth gradient estimator based on $\ell_1$-randomization.** At round $t \in [T]$, let $h_t > 0$, let $\boldsymbol{\zeta}_t^\diamond$ be a random vector uniformly distributed on $\partial B_1^d$, and $r_t$ be uniformly distributed on the interval $[-1, 1]$. Then, we receive the link function's feedback at two points as follows:

$$y_t = F(\mathbf{x}_t - h_t r_t \boldsymbol{\zeta}_t^\diamond, \xi_t), \quad \text{and} \quad y_t' = F(\mathbf{x}_t + r_t h_t \boldsymbol{\zeta}_t^\diamond, \xi_t') \ .$$

where $\xi_t, \xi_t'$ are noises. Define the gradient estimator as follows:

$$\tilde{\mathbf{g}}_t = \frac{d}{2h_t}\left(y_t - y_t'\right)\operatorname{sign}\left(\boldsymbol{\zeta}_t^\diamond\right)K(r_t) \ . \tag{1.24}$$

In Chapter 5 we show that the performance of (1.24) is comparable to (1.23). However, the same discussion that we provided on the comparison between (1.21) and (1.20) is valid here. We only need $d$ bits and $1$ float to store (1.24), which is more economic compared to (1.23).

The next estimator might be the most intuitive one as it estimates the gradient coordinate wise. It is introduced in Chapter 3.

**Smooth gradient estimator based on coordinate-wise differences.** For $i \in [d]$, let $\mathbf{e}_i \in \mathbb{R}^d$ be the $i$-th canonical basis in $\mathbb{R}^d$. Assume that $T = dT_0$, for a positive integer $T_0$. At round $t \in [T_0]$, let $h_t > 0$, and $r_t$ be uniformly distributed on the unit interval $[-1, 1]$. Then, we receive the link function's feedback at $2d$ points as follows:

$$y_{t,i} = F(\mathbf{x}_t - h_t r_t \mathbf{e}_i, \xi_{t,i}), \quad \text{and} \quad y'_{t,i} = F(\mathbf{x}_t + h_t r_t \mathbf{e}_i, \xi'_{t,i}) \ ,$$

where $\xi_{t,i}, \xi'_{t,i}$ are noises for $i \in [d]$. We define the gradient estimator with components

$$\tilde{g}_{t,i} = \frac{1}{2h_t} \left(y_{t,i} - y'_{t,i}\right) \mathbf{e}_i K(r_t) \ , \tag{1.25}$$

and let $\tilde{\mathbf{g}}_t = (\tilde{g}_{t,1}, \cdots, \tilde{g}_{t,d})$. In Chapter 3 we show that the performance of smooth gradient estimator based on finite differences is analogous to (1.23), as a function of $d$ and $T$. However, there are certain computational benefits in (1.25). For a fixed number of function queries $2T$, estimators (1.23) and (1.24) require $T$ calls to generate $\zeta_t^\circ$ (or $\zeta_t^\diamond$) and $T$ calls to generate $r_t$, for $t \in [T]$, but (1.25) only needs $T/d$ calls to generate $r_t$.

**Gradient estimator for passive scheme.** In the passive scheme, unlike the active scheme, the learner is not allowed to explore the domain of the function. There is a given set of random points $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ and the noisy function evaluation at these points, namely

$$F(\mathbf{z}_1, \xi(\mathbf{z}_1)), \cdots, F(\mathbf{z}_n, \xi(\mathbf{z}_n))$$

that are revealed to the learner and the construction of $\tilde{\mathbf{g}}_t$ is due to this information. Recall that in the beginning of this section we assumed that $F(\mathbf{z}_i, \xi(\mathbf{z}_i)) = f(\mathbf{z}_i) + \xi_i$, where $\xi_i$ is independent from $\mathbf{z}_i$, with $\mathbf{E}[\|\xi_i\|^2] \leq \sigma^2$, for $i \in [n]$ and $\sigma > 0$. However, for optimization in passive scheme we need a more restrictive assumption on the noise. Particularly, we need to assume that $\xi_i$s are mutually independent and $\mathbf{E}\left[\xi_i\right] = 0$. If we use sub-gradient descent with estimated gradient we need to obtain a small upper bound for the terms in (1.22). To get a small first term in (1.22) one may consider non-parametric estimation of $\nabla f(\cdot)$ at point $\mathbf{x}_t$. The case $d = 1$ with a deterministic set of points is studied by Müller (1985, 1989) and an extension to the multivariate case is analyzed by Facer and Müller (2003). Härdle and Nixdorf (1987) considered the i.i.d. stochastic design setting for $d = 1$ by proposing a sequential procedure that is based on non-parametric kernel estimation of $f$. Tsybakov (1990a) considered the problem in a general dimension $d$ using local polynomial rather that kernel estimator, and proved the asymptotic minimax optimality of the proposed algorithm for estimating $\mathbf{x}^*$. The main drawback of the gradient estimator introduced by Tsybakov (1990a) is that it requires the values of the input density function at the design points $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, which is inaccessible in many practical situations. In Chapter 6, as an estimator for the gradient $\tilde{\mathbf{g}}_t$ we propose a regularized version of local polynomial estimators that does not require the values of the density function, and we obtain a non-asymptotic convergence rate for the optimization error, which is minimax optimal up to a $\log$ factor. However, similar to the aforementioned references, we still need the assumption that the density function of the i.i.d. observations $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$, should be bounded away from zero on an open neighbourhood of $\Theta$.

We now give the definition of our gradient estimator. First, we define a smoothing kernel. Let the smoothing kernel $K : \mathbb{R}^d \to \mathbb{R}$ with a compact support be such that

$$\forall \boldsymbol{u} \in \mathbb{R}^d : \quad K(\boldsymbol{u}) \geq 0, \quad \text{and} \quad \sup_{\boldsymbol{u} \in \mathbb{R}^d} K(\boldsymbol{u}) < \infty \ .$$

Moreover, we assume that $K$ is $L_K$-Lipschitz with respect to $\|\cdot\|_2$. Denote by $S$ the cardinality of the set $\{\boldsymbol{m} : |\boldsymbol{m}| \leq \ell\}$ where $\boldsymbol{m}$ is a $d$-dimensional multi index. Let $h > 0$. For any $\mathbf{z} \in \Theta$, condider

$$\boldsymbol{\theta}(\mathbf{z}) = \left(h^{|\boldsymbol{m}^{(1)}|} D^{\boldsymbol{m}^{(1)}} f(\mathbf{z}), \ldots, h^{|\boldsymbol{m}^{(S)}|} D^{\boldsymbol{m}^{(S)}} f(\mathbf{z})\right)^{\top} \ ,$$

where $\boldsymbol{m}^(1) = \mathbf{0}$, and $\boldsymbol{m}^{(i+1)} = \mathbf{e}_i$, for $i \in [d]$ and let $A \in \mathbb{R}^{d \times S}$ be a matrix with entries

$$A_{i,j} = \begin{cases} 1, & \text{if} \quad j = i+1 \\ 0, & \text{otherwise} \ , \end{cases}$$

for $i \in [d]$ and $j \in [S]$. Note that with the above structure, for any $\mathbf{z} \in \mathbb{R}^d$ we have

$$\nabla f(\mathbf{z}) = h^{-1} A \boldsymbol{\theta}(\mathbf{z}) \ .$$

At each point $\mathbf{z} \in \mathbb{R}^d$, define the following estimator for $\boldsymbol{\theta}(\mathbf{z})$. Denote by $U : \mathbb{R}^d \to \mathbb{R}^S$ a function such that

$$\forall \boldsymbol{u} \in \mathbb{R}^d : \quad U(\boldsymbol{u}) = \left(\frac{\boldsymbol{u}^{\boldsymbol{m}^{(1)}}}{\boldsymbol{m}^{(1)}!}, \ldots, \frac{\boldsymbol{u}^{\boldsymbol{m}^{(S)}}}{\boldsymbol{m}^{(S)}!}\right)^{\top} \ .$$

For any $\mathbf{z} \in \mathbb{R}^d$, $t \in [n]$, we propose a regularized version of local polynomial estimator for $\boldsymbol{\theta}(\mathbf{z})$ as follows (for a classical version of local polynomial estimators see (Tsybakov, 2009, Section 1.6)):

$$\hat{\boldsymbol{\theta}}_t(\mathbf{z}) = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^S} \sum_{i=1}^{t} \left[y_i - \langle \boldsymbol{\theta}, U\left(\frac{\mathbf{z}_i - \mathbf{z}}{h_t}\right)\rangle\right]^2 K\left(\frac{\mathbf{z}_i - \mathbf{z}}{h_t}\right) + \frac{\lambda_t}{2} \|\boldsymbol{\theta}\|_2^2 \ ,$$

where $\lambda_t, h_t > 0$ and $y_i = F(\mathbf{z}_i, \xi(\mathbf{z}_i))$. Now, consider an iterative procedure with the updates introduced in (1.4), where $\tilde{\mathbf{g}}_t$ is constructed as follows:

**Smooth gradient estimator based on regularized local polynomial estimator.** For $t \in [n]$, let $\lambda_t, h_t > 0$, and construct the gradient estimator

$$\tilde{\mathbf{g}}_t = h_t^{-1} A \hat{\boldsymbol{\theta}}_t(\mathbf{z}_t) \ . \tag{1.26}$$

In Chapter 6, we analyze the error of estimator $\mathbf{x}^*$ in the $\|\cdot\|_2$-norm of algorithm (1.4) with $\tilde{\mathbf{g}}_t$ as in (1.26), and we show that it is optimal up to a logarithmic factor. Moreover, for any

$\mathbf{z} \in \Theta$, with careful assignments for $h_n, \lambda_n > 0$, we prove that $\tilde{\mathbf{g}}(\mathbf{z}) = h_n^{-1} A \hat{\boldsymbol{\theta}}_n(\mathbf{z})$ is an optimal estimator for $\nabla f(\mathbf{z})$, with respect to the number of observations $n$.

## 1.7 Distributed optimization

In this section, we consider a generalization of problem (1.2), in which the objective function $f$ is shattered in $m$ pieces and each piece is given to an individual. Namely, we assume that

$$f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} f_i(\mathbf{x}) \enspace ,$$

with $f_i : \mathbb{R}^d \to \mathbb{R}$, for $i \in [m]$, and each $f_i$ is associated to an agent $i$. At each round, for any $\mathbf{x} \in \mathbb{R}^d$, agent $i$ has access to the noisy values of $f_i$, encoded by $F_i(\mathbf{x}, \xi(\mathbf{x})) = f_i(\mathbf{x}) + \xi(\mathbf{x})$. However, the exchange of information between the agents is limited to a prescribed network of connections. We characterize this network by the undirected and connected graph, $\mathcal{G} = (V, E)$, where $V = [m]$ is the set of nodes and each node corresponds to an agent. Also, $E \subseteq V \times V$ is the set of edges that induces the notion of neighbourhood in the network. Namely, agents $i \neq j \in [m]$ are neighbours if and only if $(i, j) \in E$, and the exchange of information is only possible between the neighbouring agents. Consider an iterative procedure, at round $t$ let $\mathbf{x}_t^i$ be the update of agent $i$, which only depends on the outputs of $F_i$ and the information that she perceives from her neighboring agents. In this framework, the goal is to obtain a small optimization error for the updates of each agent after $T$ rounds. The problem of interest can be formulated as controlling the following error term

$$\max_{i \in [n]} \mathbf{E}\left[f(\mathbf{x}_T^i) - f(\mathbf{x}^*)\right] \enspace , \tag{1.27}$$

where $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \Theta} f(\mathbf{x})$. Now, consider a critical case where $\mathcal{G}$ is a complete graph that is, the agents have complete information on each other. Then, at each round, each agent $i$ has access to the zero-order information that is provided by each $F_i$, and one can expect that the hardness of problem (1.27) is equivalent to that of the usual problem in order to have optimization error (1.2). On the contrary, if $\mathcal{G}$ is a sparse graph the agents get less information and problem (1.27) gets harder than the one in (1.2). Now, let us to propose an iterative algorithm, and formalize this intuition in practice.

**Definition 1.7.1.** *We call $W \in \mathbb{R}^{m \times m}$ a consensus matrix associated with $\mathcal{G}$ if for $i, j \in [m]$ it satisfies*

$$W_{i,j} \geq 0, \quad W_{i,j} \neq 0 \quad \textit{if} \quad (i,j) \in E \quad \textit{or} \quad i = j, \quad \textit{and} \quad \sum_{j=1}^{m} W_{i,j} = 1 \enspace ,$$

In this scheme, based on a given $\mathcal{G}$ we are allowed to construct an associated consensus matrix $W$ which is given to the agents. Accordingly, for a given $W$ we are ready to present an iterative procedure in order to have a (1.27). Let $\eta_t > 0$, and let $\mathbf{x}_1^1, \ldots, \mathbf{x}_n^1$ be non-random vectors in $\mathbb{R}^d$. Consider the following updates

$$\mathbf{x}_{t+1}^i = \sum_{j=1}^m W_{i,j} \mathrm{Proj}_\Theta \left( \mathbf{x}_t^j - \eta_t \mathbf{g}_t^j \right) \quad , \tag{1.28}$$

where $\mathbf{g}_t^j$ is based on the information perceived by $F_j$. Note that at each round, the each agent $j$ provides a local update $\mathrm{Proj}_\Theta \left( \mathbf{x}_t^j - \eta_t \mathbf{g}_t^j \right)$. Then, using the consensus matrix $W$, agent $i$ takes a weighted sum of the local updates of her neighbouring agents.

In Chapter 3, we study Problem (1.27), for a strongly convex objective function $f$. Moreover, we assume that each $f_i$ is $\beta$-smooth and we study the performance of (1.28), in which we use Estimator (1.25) as a candidate for $\mathbf{g}_t^i$. In order to investigate how the distributed nature of the problem plays its role in our analysis, consider the quantity $\rho = \left\| W - m^{-1}\mathbf{1}\mathbf{1}^\top \right\|_{\mathsf{op}}$, where $\|\cdot\|_{\mathsf{op}}$ is the operator norm and $\mathbf{1}$ is a vector in $\mathbb{R}^d$ with the coordinates all equal to one. Note that by the definition of $W$, we have $\rho \leq 1$. Moreover, we note that for any connected graph $\mathcal{G}$ one can construct an associated $W$, in which $\rho < 1$ see the example that is provided by (3.2). Therefore, without loss of generality, let us assume that $\rho < 1$. For a 2-smooth objective function $f$, our analysis outlined in Corollary 3.6.3 exhibits the fact that the rate of convergence of (1.27) for the algorithm (1.28) depends on $\rho$, and it is of the order $(1-\rho)^{-1}$. If $\mathcal{G}$ is a complete graph it is reasonable to take $W = m^{-1}\mathbf{1}\mathbf{1}^\top$, which gives $\rho = 0$, and, as we expect, there is no trace of the distributed nature of the problem. However, if $\mathcal{G}$ is a sparse graph then $W$ is sparse and $\rho$ is close to 1 that causes an explosion of the term $(1-\rho)^{-1}$.

## 1.8 Contextual bandits and fairness

Consider a sequential decision making problem where at each time-step an employer has to select one candidate from a pool to hire for a job. The employer does not know how well a candidate will perform if hired, but they can learn it over time by measuring the performance of previously selected similar candidates. This scenario can be formalized as a (linear) contextual bandit problem (see (Auer, 2002; Chu et al., 2011; Lattimore and Szepesvári, 2020) and references therein), where each candidate is represented by a context vector, and after the employer (or agent) chooses a candidate, it receives a reward, i.e. a scalar value measuring the true performance of the candidate, which depends (linearly) on the context.

In the above framework, the typical objective is to find a policy for the employer to select candidates with the highest rewards (Abbasi-Yadkori et al., 2011; Auer, 2002; Auer et al., 2002; Lattimore and Szepesvári, 2020). However, in some important scenarios this objective may not be appropriate; if candidates belong to different sensitive groups (e.g. based on ethnicity, gender, etc.) the resulting policy might discriminate or even exclude some groups completely in the selection process. This may happen when some groups have lower expected reward than others, e.g. because they acquired less skills due lower financial support. Another example arises when each candidate in the pool, if selected, will perform a different kind of job, and the associated reward is job-specific. For instance, if the employer is a university and each candidate is a researcher in a different discipline, then the rewards associated to different disciplines will be substantially different and incomparable, e.g. citations counts greatly vary among different subjects; see (Kearns et al., 2017) for a discussion. In both of the above scenarios, it is unfair to directly compare rewards of candidates belonging to different groups.

A simple way to deal with this issue would be to select the candidate to hire uniformly at random. This policy satisfies a notion of fairness called *demographic parity* (see Calders et al. (2009); Mehrabi et al. (2021) and references therein), which requires the probability of selecting a candidate from a certain group to be equal for all groups. However, as it is apparent, this approach completely ignores the employer's goal of selecting good candidates and is also unfair to candidates who spent effort acquiring credentials for the job. In this work, we provide a fair way of comparing candidates from different groups via the *relative rank*, that is the rank (or quantile) of the reward w.r.t. the rewards distribution of the candidate's group. We call a policy *group meritocratic fair* (GMF) if it always selects a candidate with the highest *relative rank*. Such a policy is meritocratic but only in terms of the within-group performance. A closely related idea has been introduced in Kearns et al. (2017) for settings where the candidates' rewards are available before the selection, while we are not aware of a similar notion in the multi-armed bandits literature.

A GMF policy requires the knowledge of the relative rank of each candidate which is not directly observed by the agent and depends on the underlying reward model and on the distributions of rewards.

In recent years algorithmic fairness has received a lot of attention, becoming a large area of machine learning research. The potential for learning algorithms to amplify pre-existing bias and cause harm to human beings has triggered researchers to study solutions to mitigate or remove unfairness of the learned predictor, see Barocas et al. (2018); Calmon et al. (2017); Chierichetti et al. (2017); Donini et al. (2018); Dwork et al. (2018); Hardt et al. (2016) and references therein. Fairness in sequential decision problems (see Zhang and Liu (2021) for a survey) is usually divided into two categories: group fairness (GF) and individual fairness. We give an overview of these notions below.

GF requires some statistical measure to be (approximately) equal across different sensitive groups. A prominent example relevant to this work is *demographic parity*, which requires that the probability that the policy selects a candidate from a certain group should be the same for all groups. A similar notion is used by Chen et al. (2020); Patil et al. (2020), where the probability that the policy selects a candidate has to always be greater than a given threshold for all candidates. Li et al. (2019) impose a weaker requirement concerning the expected fraction of candidates selected from each group. Other examples of GF in sequential decision problems are *equal opportunity* (Bechavod et al., 2019) and *equalized odds* (Blum et al., 2018). Under some assumptions on the distributions of the contexts, the GMF policy and greedy policy that we propose satisfy variants of demographic parity at each round.

Individual fairness can be divided in two categories: fairness through awareness (FA) (Liu et al., 2017; Wang et al., 2021) and meritocratic fairness (MF) (Joseph et al., 2018, 2016). FA is based on the idea that similar individuals should be treated similarly and is designed to avoid "winner takes all" scenarios where some individuals cannot be selected when they have a lower reward than others in the pool, even if the difference between rewards is very small. For example, Wang et al. (2021) propose a policy where the probability of selecting a context over another is lower when the context has a lower reward, but is never zero. MF instead requires that less qualified individual should not be favored over more qualified ones, which could happen during the learning process. For example Joseph et al. (2016) proposes an algorithm based on confidence intervals, where if the uncertainty between the best arms is too high the arm is selected uniformly at random. This guarantees meritocratic fairness at each round but comes at a cost in terms of regret.

Our definition of fairness falls between group and meritocratic fairness. It is meritocratic because it states that a candidate with a worse relative rank than another should never be selected. It is also based on groups since the relative ranks directly depend on the distribution of rewards of each group. A similar idea of fairness based on relative rank has been introduced in Kearns et al. (2017), which study the problem of selecting candidates from different groups based on their scalar-valued score when the scores between groups are incomparable (e.g. number of citations in different research areas). Contrary to our work, where the (noisy) rewards are observed only for the selected candidates, in Kearns et al. (2017) the noiseless scores for all candidates can be accessed before the selection. This difference makes the estimation of the relative rank simpler in Kearns et al. (2017), as the rewards CDFs can be estimated more efficiently.

## 1.9 Résumé en français

### Optimisation

L'optimisation est une branche de l'étude où le but est d'estimer une quantité extrême associée à une certaine fonction. Des exemples de telles quantités extrémales incluent un minimiseur, un maximiseur, un point de selle, et autres. Dans cette section, nous présentons et étudions l'un des modèles les plus courants où la quantité extrême est un minimiseur d'une fonction $f : \mathbb{R}^d \to \mathbb{R}$ dans un sous-ensemble fermé et convexe $\Theta$ de $\mathbb{R}^d$. Plus précisément, nous sommes intéressés par l'estimation de En particulier, le problème qui nous intéresse est d'estimer

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\Theta} f(\mathbf{x}) \ ,$$

par une certaine $\hat{\mathbf{x}}$ potentiellement aléatoire, ce qui garantit une petite *erreur d'optimisation*, c'est-à-dire que

$$\mathbf{E}\left[f(\hat{\mathbf{x}})\right] - f(\mathbf{x}^*) \ . \tag{1.29}$$

où l'espérance dans (1.29) est par rapport à la distribution de probabilité de $\hat{\mathbf{x}}$. Bien entendu, sans spécifier le type d'information auquel nous pouvons accéder sur la fonction $f$, le problème énoncé ci-dessus est sans espoir. Nous présentons ci-dessous un modèle d'interrogation assez général, qui inclut de nombreux schémas d'observation populaires.

Considérons une procédure itérative, telle que à chaque temps $t \geq 1$, pour tout choix $\mathbf{x} \in \mathbb{R}^d$ de l'apprenant, la nature produit une information bruyante sur la fonction, codée par la fonction de liaison $F(\mathbf{x}, \xi(\mathbf{x}))$, où $\xi(\mathbf{x})$ est un bruit de mesure. Formellement, il existe $F : \mathbb{R}^d \times \mathbb{R}^{\ell_1} \to \mathbb{R}^{\ell_2}$, pour quelques entiers positifs $\ell_1, \ell_2$, de sorte que pour chaque vecteur $\mathbf{x} \in \mathbb{R}^d$ sélectionné par l'apprenant, la nature échantillonne la variable de bruit $\xi(\mathbf{x})$ et renvoie $F(\mathbf{x}, \xi(\mathbf{x}))$ à l'apprenant.

Le cadre ci-dessus est plutôt abstrait et le problème concret et la stratégie d'estimation dépendront de la forme de la fonction de liaison $F$. Fournissons quelques exemples de fonctions de liaison $F$ et mettons-les en relation avec les paramètres bien connus de la littérature sur l'optimisation. Tout d'abord, nous donnons la définition du sous-gradient d'une fonction.

**Définition 1.9.1.** *Soit $f : \mathbb{R}^d \to \mathbb{R}$, et fixons $\boldsymbol{x} \in \mathbb{R}^d$. Nous appelons $\partial f(\boldsymbol{x}) \subseteq \mathbb{R}^d$ l'ensemble des sous-gradients (ou sous-différentielles) de $f$ au point $\boldsymbol{x}$, si pour tout $\boldsymbol{g}$ $dans \partial f(\boldsymbol{x})$, et $\boldsymbol{y} dans \mathbb{R}^d$, nous avons*

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \leq \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{y} \rangle.$$

Si $f$ est une fonction convexe, la définition ci-dessus est une généralisation du gradient de $f$. En particulier, si $f$ est convexe et différentiable en $\mathbf{x}$ $dans \mathbb{R}^d$, alors $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$ (voir le Lemme 1.2.2).

**Exemple 1.9.2** (Optimisation du premier ordre)**.** *Nous appelons un problème d'optimisation un problème du premier ordre si la fonction de liaison $F : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ et l'évaluation au point $\boldsymbol{x} \in \mathbb{R}^d$ contiennent des informations sur un sous-gradient de la fonction au point $\boldsymbol{x}$. Le cas le plus simple est celui où $F(\boldsymbol{x}, \xi(\boldsymbol{x})) \in \partial f(\boldsymbol{x})$.*

*Dans le cadre stochastique avec une fonction objectif différentiable $f$, le cas le plus connu est $\mathbf{E}[F(\boldsymbol{x}, \xi(\boldsymbol{x}))] = \nabla f(\boldsymbol{x})$ (Robbins and Monro, 1951). Un exemple particulier est le modèle de bruit additif, où $F(\boldsymbol{x}, \xi(\boldsymbol{x})) = \nabla f(\boldsymbol{x}) + \xi(\boldsymbol{x})$.*

**Exemple 1.9.3** (Optimisation d'ordre zéro)**.** *Un problème d'optimisation est appelé un problème d'ordre zéro si pour tout $\boldsymbol{x}$ $dans \mathbb{R}^d$, la fonction de liaison $F : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ fournit des informations sur les valeurs de la fonction. À titre d'exemple, on peut considérer le modèle de bruit additif $F(\boldsymbol{x}, \xi(\boldsymbol{x})) = f(\boldsymbol{x}) + \xi(\boldsymbol{x})$. C'est le cas principal étudié ci-dessous. Tout au long de cette thèse, chaque fois que nous mentionnerons que l'apprenant a accès à des informations d'ordre zéro, nous ferons référence à $F(\boldsymbol{x}, \xi(\boldsymbol{x})) = f(\boldsymbol{x}) + \xi(\boldsymbol{x})$.*

Dans les sections suivantes, nous présentons un bref historique de l'analyse convexe et de l'optimisation convexe.

### Optimisation d'ordre zéro de fonctions Lipschitz

Dans cette section, nous supposons que la fonction objectif $f$ est convexe, et que la fonction de liaison ne fournit que des informations bruyantes d'ordre zéro. En particulier, nous supposons que pour toute $\mathbf{x} \in \mathbb{R}^d$ nous avons $F(\mathbf{x}, \xi(\mathbf{x})) = f(\mathbf{x}) + \xi$, où et $\mathbf{E}[\xi^2] \leq \sigma^2$ pour $\sigma > 0$. Considérons les mises à jour dans (1.18), et supposons que pour tout $h > 0$, il existe une *fonction de substitution* $\mathsf{f}_h(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$, telle que $\mathsf{f}_h$ est convexe et différentiable, et pour tout $\mathbf{x} \in \mathbb{R}^d$ nous avons

$$0 \leq \mathsf{f}_h(\mathbf{x}) - f(\mathbf{x}) \leq \Delta, \quad \text{et} \quad \mathbf{E}\left[\tilde{\mathbf{g}}_t | \mathbf{x}_t\right] = \nabla \mathsf{f}_h(\mathbf{x}_t) \quad \text{presque sûrement} \;,$$

où $\Delta > 0$. Ensuite, en utilisant la limite fournie par (1.19), nous obtenons

$$\sum_{t=1}^{T} \mathbf{E}\left[\mathsf{f}_h(\mathbf{x}_t) - \mathsf{f}_h(\mathbf{x}^*)\right] \leq \frac{V(\mathbf{x}^*) - \min_{\mathbf{x} \in \Theta} V(\mathbf{x})}{\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \eta_t \mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_{q^*}^2\right] \;.$$

et par conséquent

$$\sum_{t=1}^{T} \mathbf{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{V(\mathbf{x}^*) - \min_{\mathbf{x} \in \Theta} V(\mathbf{x})}{\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \eta_t \mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_{q^*}^2\right] + T\Delta \;.$$

Ainsi, si $\Delta$ est suffisamment petit, une limite significative pour l'erreur d'optimisation de $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$ suit si nous fournissons une limite supérieure suffisamment serrée pour $\mathbf{E}[\|\tilde{\mathbf{g}}_t\|_{q^*}^2]$, et attribuons correctement les tailles de pas $\eta_t$ pour $t \in [T]$. L'idée d'utiliser une fonction de substitution particulière $\mathsf{f}_h(\cdot)$ est proposée pour la première fois par (Nemirovsky and Yudin, 1983, Chapitre 9.3.2) comme un exercice. Nous fournissons ci-dessous deux exemples de $\tilde{\mathbf{g}}_t$ et de $\mathsf{f}_h(\cdot)$, où dans la construction de $\tilde{\mathbf{g}}_t$ nous n'utilisons que des informations d'ordre zéro.

**Estimateur de gradient basé sur la randomisation de $\ell_2$.** Au tour $t \in [T]$, laissez $h_t > 0$, et laissez $\zeta_t^\circ$ être un vecteur aléatoire uniformément distribué sur $\partial B_2^d$. Supposons que nous recevions la rétroaction de la fonction de liaison aux deux points suivants

$$y_t = F(\mathbf{x}_t - h_t \zeta_t^\circ, \xi_t), \quad \text{et} \quad y_t' = F(\mathbf{x}_t + h_t \zeta_t^\circ, , \xi_t') \;,$$

où $\xi_t$ et $\xi_t'$ sont des bruits. Construire l'estimateur du gradient

$$\tilde{\mathbf{g}}_t = \frac{d}{2h_t}\left(y_t - y_t'\right) \zeta_t^\circ \;. \tag{1.30}$$

Une version initiale de cet estimateur, qui n'est construite que sur la base d'une rétroaction ponctuelle, est proposée par Nemirovsky and Yudin (1983) et étudiée plus en détail par Flaxman et al. (2005). Le fait que $\mathbf{E}[\tilde{\mathbf{g}}_t | \mathbf{x}_t] = \mathsf{f}_h(\mathbf{x}_t)$, où tel que pour tout $\mathbf{x}\; dans \mathbb{R}^d$,

$$\mathsf{f}_h(\mathbf{x}) = \mathbf{E}\left[f(\mathbf{x} + hU)\right],$$

et que $\boldsymbol{U}$ est uniformément distribué dans $B_2^d$ est énoncé sans preuve dans Nemirovsky and Yudin (1983) et Flaxman et al. (2005). Il est fait référence au théorème de Stokes dans Flaxman et al. (2005). La version précise du théorème de Stokes nécessaire au fonctionnement de l'estimateur est énoncée et prouvée au Chapitre 4. La forme actuelle de (1.30) est introduite par Agarwal et al. (2010), et elle est analysée plus en détail par Duchi et al. (2015); Novitskii and Gasnikov (2021); Shamir (2017). De plus, pour $q = 1, 2$, la borne sur $\mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_{q^*}^2\right]$, avec $\tilde{\mathbf{g}}_t$ donné dans (1.30) est obtenue par (Shamir, 2017, Corollaires 2 et 3) pour les fonctions $f$ qui sont $L$-Lipschitz par rapport à $\|\cdot\|_q$.

**Estimateur de gradient basé sur la randomisation de $\ell_1$.** Au tour $t \in [T]$, laissez $h_t > 0$, et laissez $\zeta_t^\diamond$ être un vecteur aléatoire uniformément distribué sur $\partial B_1^d$. Supposons que nous recevions la rétroaction de la fonction de liaison aux deux points suivants

$$y_t = F(\mathbf{x}_t - h_t\boldsymbol{\zeta}_t^\diamond, \xi_t), \quad \text{et} \quad y_t' = F(\mathbf{x}_t + h_t\boldsymbol{\zeta}_t^\diamond, \xi_t') \enspace ,$$

où $\xi_t$ et $\xi_t'$ sont des bruits. Construire l'estimateur du gradient

$$\tilde{\mathbf{g}}_t = \frac{d}{2h_t}\left(y_t - y_t'\right) \operatorname{sign}(\boldsymbol{\zeta}_t^\diamond) \enspace , \tag{1.31}$$

où nous introduisons $\operatorname{sign} : \mathbb{R}^d \to [-1, 1]^d$ comme la fonction de signe par composante (définie à $0$ comme $1$). Considérons la fonction $\mathsf{f}_h : \mathbb{R}^d \to \mathbb{R}$, telle que pour tout $\mathbf{x} \in \mathbb{R}^d$

$$\mathsf{f}_h(\mathbf{x}) = \mathbf{E}\left[f(\mathbf{x} + h\boldsymbol{V})\right],$$

où $\boldsymbol{V}$ est uniformément distribué sur $B_1^d$. Cet estimateur est proposé et analysé au Chapitre 4. Comme dans le cas de l'estimateur du gradient basé sur la randomisation $\ell_2$, nous utilisons le théorème de Stokes pour montrer que $\mathbf{E}[\tilde{\mathbf{g}}_t|\mathbf{x}_t] = \mathsf{f}_h(\mathbf{x}_t)$ (voir le lemme 4.6.1). De plus, grâce à la forme simple de l'algorithme, nous sommes en mesure de fournir une limite souhaitée pour le terme $\mathbf{E}[\|\tilde{\mathbf{g}}_t\|_{q^*}^2]$, pour tout $q \in [1, \infty]$, et tout $f$ qui est $L$-Lipschitz par rapport à $\|\cdot\|_q$. Notre analyse est basée sur une nouvelle inégalité pondérée de type Poincaré, qui est proposée dans la Section 4.6. Pour le cas $q = 1$, ce dernier estimateur surpasse l'estimateur par randomisation $\ell_2$ en ce qui concerne l'erreur d'optimisation, jusqu'à un facteur $\sqrt{\log(d)}$, et si $q = 2$, ils conduisent tous deux à la limite supérieure optimale de l'erreur d'optimisation (voir le Théorème 4.4.1). De plus, en termes de mémoire requise, nous n'avons besoin que de $d$ bits et $1$ flottants pour stocker (1.31), ce qui est plus économique par rapport à la méthode de randomisation $\ell_2$ qui nécessite de stocker $d$ flottants.

À la fin de cette section, nous souhaitons mentionner l'**estimateur de gradient basé sur la randomisation gaussienne**, qui est introduit par Nesterov (2011) et joue un rôle important dans la littérature de l'optimisation d'ordre zéro. Au tour $t \in [T]$, laissez $h_t > 0$, et laissez $\zeta_t^G$ être un vecteur gaussien standard. Supposons que nous recevions la rétroaction de la fonction de liaison aux deux points suivants

$$y_t = F(\mathbf{x}_t - h_t\boldsymbol{\zeta}_t^G, \xi_t), \quad \text{et} \quad y'_t = F(\mathbf{x}_t + h_t\boldsymbol{\zeta}_t^G, \xi'_t) \ ,$$

où $\xi_t$ et $\xi'_t$ sont des bruits. Construire l'estimateur du gradient

$$\tilde{\mathbf{g}}_t = \frac{1}{2h_t}\left(y_t - y'_t\right)\boldsymbol{\zeta}_t^G \ ,$$

et considérons la fonction $\mathsf{f}_h : \mathbb{R}^d \to \mathbb{R}$, telle que pour tout $\mathbf{x} \in \mathbb{R}^d$, on obtient

$$\mathsf{f}_h(\mathbf{x}) = \mathbf{E}\left[f(\mathbf{x} + h\boldsymbol{\zeta}_t^G)\right] .$$

Cependant, comme cet estimateur dépasse le cadre de la thèse, nous n'établissons pas ses propriétés et nous renvoyons le lecteur aux documents suivants : Nesterov (2011) ; Ghadimi and Lan (2013) ; Nesterov and Spokoiny (2017) ; Balasubramanian and Ghadimi (2021) et leurs références.

### Optimisation d'ordre zéro de fonctions hautement lisses

Dans cette section, nous supposons que la fonction objectif $f$ est $\alpha$-fortement convexe et différentiable. De plus, pour toute $\mathbf{x} \in \mathbb{R}^d$, nous supposons que $F(\mathbf{x}, \xi(\mathbf{x})) = f(\mathbf{x}) + \xi$, où $\xi$ est indépendant de $\mathbf{x}$, et $Exp\left[\xi^2\right] \leq \sigma^2$, pour $\sigma > 0$. Soit $\tilde{\mathbf{g}}_t$ un estimateur de gradient qui n'utilise que l'information d'ordre zéro. Afin d'obtenir une limite sur l'erreur d'optimisation de la descente par sous-gradient, nous devons fournir un contrôle sur les termes

$$\mathbf{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{E}\left[\tilde{\mathbf{g}}_t|\mathbf{x}_t\right]\|_2^2\right], \quad \text{and} \quad \mathbf{E}\left[\|\tilde{\mathbf{g}}_t\|_2^2\right] \quad . \tag{1.32}$$

En effet, il découle de (1.3.5) et du fait que $f$ est une fonction convexe et différentiable, par le Lemme 1.2.2 nous avons $\partial f(\mathbf{x}_t) = \{\nabla f(\mathbf{x}_t)\}$. Construire un estimateur $\tilde{\mathbf{g}}_t$ qui donne un petit terme de biais au carré $\mathbf{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{E}\left[\tilde{\mathbf{g}}_t|\mathbf{x}_t\right]\|_2^2\right]$ peut être vu comme un problème d'estimation non paramétrique du gradient $\nabla f(\cdot)$ en un certain point $\mathbf{x}_t$. D'après l'intuition issue de la littérature sur les statistiques non paramétriques, on peut se demander si approximation locale de la fonction (dans ce cas, l'estimation de $\nabla f(\mathbf{x}_t)$) ne pourrait pas bénéficier d'une hypothèse de *lissé d'ordre supérieur* sur la fonction objectif $f$. Pour caractériser la notion de *lissé d'ordre supérieur*, nous utilisons la classe de fonctions de Hölder bien connue de $\beta$, à savoir $\mathcal{F}_\beta(L)$ (définie dans (2.1)), où nous limitons notre attention au cas $\beta \geq 2$.

Dans ce qui suit, nous supposons que $f \in \mathcal{F}_\beta(L)$, et considérons les deux paramètres suivants.

**Schéma actif.** Il s'agit du paramètre habituel en optimisation et en apprentissage en ligne, où, au tour $t$, l'apprenant choisit $\mathbf{z} = \mathbf{x}_t$ en fonction de

$$\{\mathbf{x}_1, F(\mathbf{x}_1, \xi(\mathbf{x}_1)), \cdots, \mathbf{x}_{t-1}, F(\mathbf{x}_{t-1}, \xi(\mathbf{x}_{t-1}))\}$$

et observe $F(\mathbf{z}, \xi(\mathbf{z}))$ où $\mathbf{z} \in \mathbb{R}^d$ (ou, de manière plus restrictive, $\mathbf{z} \in \Theta$). Nous discutons de ce paramètre dans les Chapitres 2, 3, et 5.

**Schéma passif.** Nous étudions ce paramètre dans le Chapitre 6, où nous ne sommes pas autorisés à choisir les points de requête et n'avons accès qu'à l'ensemble

$$\{\mathbf{z}_1, F(\mathbf{z}_1, \xi(\mathbf{z}_1)), \cdots, \mathbf{z}_n, F(\mathbf{z}_n, \xi(\mathbf{z}_n))\}$$

, avec $\mathbf{z}_i \in \mathbb{R}^d$ pour $i \in [n]$.

**Estimateurs de gradient pour le schéma actif.** L'exploitation du lissage d'ordre supérieur dans le schéma actif remonte à Polyak and Tsybakov (1990), où les auteurs ont d'abord suggéré d'utiliser un noyau de lissage pour construire l'estimateur du gradient. Plus tard, cette approche a été développée par Dippon (2003a) et Bach and Perchet (2016). Dans ce qui suit, nous étudions l'estimateur du gradient qui est proposé par Bach and Perchet (2016) (Chapitres 1 et 5), ainsi que deux autres exemples d'estimateurs de gradient qui sont présentés et analysés dans les Chapitres 2 et 5. Avant de présenter ces estimateurs, nous définissons une entité qui joue un rôle crucial dans la structure des estimateurs qui tirent parti de la propriété de lissage d'ordre supérieur, à savoir la fonction noyau $K : [-1, 1] \to \mathbb{R}$, telle que

$$\int K(u)\,\mathrm{d}u = 0, \int uK(u)\,\mathrm{d}u = 1, \int u^j K(u)\,\mathrm{d}u = 0, \text{ pour } j = 2, \cdots, \ell, \int |u|^\beta |K(u)|\,\mathrm{d}u < \infty \enspace ,$$

où $\ell$ est le plus grand entier inférieur à $\beta$. Pour des exemples de telles fonctions à noyau, nous renvoyons le lecteur à Polyak and Tsybakov (1990) et Bach and Perchet (2016). Après avoir introduit la fonction noyau $K$, nous sommes prêts à présenter les estimateurs du gradient.

**Estimateur de gradient lisse basé sur la randomisation $\ell_2$.** Au tour $t \in [T]$, laissez $h_t > 0$, laissez $\boldsymbol{\zeta}_t^\circ$ être un vecteur aléatoire uniformément distribué sur $\partial B_2^d$, et $r_t$ être uniformément distribué sur l'intervalle $[-1, 1]$. Ensuite, nous recevons le retour de la fonction de liaison en deux points comme suit :

$$y_t = F(\mathbf{x}_t - h_t r_t \boldsymbol{\zeta}_t^\circ, \xi_t), \quad \text{et} \quad y_t' = F(\mathbf{x}_t + h_t r_t \boldsymbol{\zeta}_t^\circ, \xi_t') \enspace ,$$

où $\xi_t, \xi_t'$ sont des bruits. Définissez l'estimateur du gradient (voir Bach and Perchet (2016)) comme suit :

$$\tilde{\mathbf{g}}_t = \frac{d}{2h_t} \left(y_t - y_t'\right) \boldsymbol{\zeta}_t^\circ K(r_t) \enspace . \tag{1.33}$$

Dans le Chapitre 1, nous étudions l'erreur d'optimisation de l'algorithme (1.4) avec $\tilde{\mathbf{g}}_t$ comme dans (1.33). L'analyse fournie dans le Chapitre 1 correspond à l'article Akhavan et al. (2020).

Elle est affinée dans un travail de suivi par Novitskii and Gasnikov (2021).

Un estimateur basé sur la randomisation $\ell_1$ (version lisse de (1.31)) est présenté au Chapitre 5.

**Estimateur de gradient lisse basé sur la randomisation $\ell_1$.** Au tour $t \in [T]$, laissez $h_t > 0$, laissez $\boldsymbol{\zeta}_t^{\diamond}$ être un vecteur aléatoire uniformément distribué sur $\partial B_1^d$, et $r_t$ être uniformément distribué sur l'intervalle $[-1, 1]$. Ensuite, nous recevons le retour de la fonction de liaison en deux points comme suit :

$$y_t = F(\mathbf{x}_t - h_t r_t \boldsymbol{\zeta}_t^{\diamond}, \xi_t), \quad \text{et} \quad y_t' = F(\mathbf{x}_t + r_t h_t \boldsymbol{\zeta}_t^{\diamond}, \xi_t') \ .$$

où $\xi_t, \xi_t'$ sont des bruits. Définissez l'estimateur du gradient comme suit :

$$\tilde{\mathbf{g}}_t = \frac{d}{2h_t} \left(y_t - y_t'\right) \operatorname{sign}\left(\boldsymbol{\zeta}_t^{\diamond}\right) K(r_t) \ . \tag{1.34}$$

Au Chapitre 5, nous montrons que les performances de (1.34) sont comparables à celles de (1.33). Cependant, la même discussion que nous avons fournie sur la comparaison entre (1.31) et (1.30) est valable ici. Nous n'avons besoin que de $d$ bits et $1$ entier pour stocker (1.34), ce qui est plus économique comparé à (1.33).

L'estimateur suivant est peut-être le plus intuitif car il estime le gradient par coordonnées. Il est présenté au Chapitre 3.

**Estimateur de gradient lisse basé sur les différences entre coordonnées.** Pour $i \in [d]$, laissez $\mathbf{e}_i \in \mathbb{R}^d$ être la $i$ième base canonique dans $\mathbb{R}^d$. Supposons que $T = dT_0$, pour un entier positif $T_0$. Au tour $t \in [T_0]$, laissons $h_t > 0$, et $r_t$ être uniformément distribués sur l'intervalle unitaire $[-1, 1]$. Ensuite, nous recevons le retour de la fonction de liaison à $2d$ points comme suit :

$$y_{t,i} = F(\mathbf{x}_t - h_t r_t \mathbf{e}_i, \xi_{t,i}), \quad \text{et} \quad y_{t,i}' = F(\mathbf{x}_t + h_t r_t \mathbf{e}_i, \xi_{t,i}') \ ,$$

où $\xi_{t,i}, \xi_{t,i}'$ sont des bruits pour $i \in [d]$. Nous définissons l'estimateur du gradient avec des composantes

$$\tilde{g}_{t,i} = \frac{1}{2h_t} \left(y_{t,i} - y_{t,i}'\right) \mathbf{e}_i K(r_t) \ , \tag{1.35}$$

et soit $\tilde{\mathbf{g}}_t = (\tilde{g}_{t,1}, \cdots, \tilde{g}_{t,d})$. Au Chapitre 3, nous montrons que les performances de l'estimateur du gradient lisse basé sur les différences finies sont analogues à celles de (1.33), en fonction de $d$ et $T$. Cependant, l'(1.35) présente certains avantages en termes de calcul. Pour un nombre fixe de requêtes de fonctions $2T$, les estimateurs (1.33) et (1.34) nécessitent $T$ d'appels pour générer $\zeta_t^{\circ}$ (ou $\zeta_t^{\diamond}$) et $T$ d'appels pour générer $r_t$, pour $t \in [T]$, mais (1.35) ne nécessite que $T/d$ d'appels pour générer $r_t$.

**Estimateur de gradient pour le schéma passif.** Dans le schéma passif, contrairement au schéma actif, l'apprenant n'est pas autorisé à explorer le domaine de la fonction. Il existe un ensemble donné de points aléatoires $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ et l'évaluation bruyante de la fonction à ces points, à savoir

$$F(\mathbf{z}_1, \xi(\mathbf{z}_1)), \cdots, F(\mathbf{z}_n, \xi(\mathbf{z}_n))$$

qui sont révélés à l'apprenant et la construction de $\tilde{\mathbf{g}}_t$ est due à ces informations. Rappelons qu'au début de cette section, nous avons supposé que $F(\mathbf{z}_i, \xi(\mathbf{z}_i)) = f(\mathbf{z}_i) + \xi_i$, où $\xi_i$ est indépendant de $\mathbf{z}_i$, avec $\mathbf{E}[\|\xi_i\|^2] \leq \sigma^2$, pour $i \in [n]$ et $\sigma > 0$. Cependant, pour l'optimisation dans le schéma passif, nous avons besoin d'une hypothèse plus restrictive sur le bruit. En particulier, nous devons supposer que les $\xi_i$s sont mutuellement indépendants et que $\mathbf{E}[\xi_i] = 0$. Si nous utilisons la descente par sous-gradient avec gradient estimé, nous devons obtenir une petite limite supérieure pour les termes de (1.32). Pour obtenir un petit premier terme dans (1.32), on peut considérer une estimation non paramétrique de $\nabla f(\cdot)$ au point $\mathbf{x}_t$. Le cas $d = 1$ avec un ensemble déterministe de points est étudié par Müller (1985, 1989) et une extension au cas multivarié est analysée par Facer and Müller (2003). Härdle and Nixdorf (1987) ont considéré le cadre du plan stochastique i.i.d. pour $d = 1$ en proposant une procédure séquentielle qui est basée sur l'estimation non-paramétrique du noyau de $f$. Tsybakov (1990a) a considéré le problème dans une dimension générale $d$ en utilisant un polynôme local plutôt qu'un estimateur à noyau, et a prouvé l'optimalité asymptotique minimax de l'algorithme proposé pour estimer $\mathbf{x}^*$. Le principal inconvénient de l'estimateur du gradient introduit par Tsybakov (1990a) est qu'il nécessite les valeurs de la fonction de densité d'entrée aux points de conception $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, ce qui est inaccessible dans de nombreuses situations pratiques. Dans le Chapitre 6, nous proposons, comme estimateur du gradient $\tilde{\mathbf{g}}_t$, une version régularisée des estimateurs polynomiaux locaux qui ne nécessite pas les valeurs de la fonction de densité, et nous obtenons un taux de convergence non asymptotique pour l'erreur d'optimisation, qui est minimax optimal jusqu'à un facteur $\log$. Cependant, comme dans les références susmentionnées, nous avons toujours besoin de l'hypothèse selon laquelle la fonction de densité des observations i.i.d. $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$, doit être bornée loin de zéro dans un voisinage ouvert de $\Theta$.

Nous donnons maintenant la définition de notre estimateur de gradient. Tout d'abord, nous définissons un noyau de lissage. Soit le noyau de lissage $K : \mathbb{R}^d \to \mathbb{R}$ avec un support compact tel que

$$\forall \boldsymbol{u} \in \mathbb{R}^d : \quad K(\boldsymbol{u}) \geq 0, \quad \text{et} \quad \sup_{\boldsymbol{u} \in \mathbb{R}^d} K(\boldsymbol{u}) < \infty \ .$$

De plus, nous supposons que $K$ est $L_K$-Lipschitz par rapport à $\|\cdot\|_2$. Dénotons par $S$ la cardinalité de l'ensemble $\{\boldsymbol{m} : |\boldsymbol{m}| \leq \ell\}$ où $\boldsymbol{m}$ est un multi-indice à $d$ dimensions. Soit $h > 0$. Pour tout $\mathbf{z} \in \Theta$, condensons

$$\boldsymbol{\theta}(\mathbf{z}) = \left(h^{|\boldsymbol{m}^{(1)}|} D^{\boldsymbol{m}^{(1)}} f(\mathbf{z}), \ldots, h^{|\boldsymbol{m}^{(S)}|} D^{\boldsymbol{m}^{(S)}} f(\mathbf{z})\right)^\top \quad ,$$

où $\boldsymbol{m}^{(1)} = \mathbf{0}$, et $\boldsymbol{m}^{(i+1)} = \mathbf{e}_i$, pour $i \in [d]$ et laissons $A \in \mathbb{R}^{d \times S}$ être une matrice avec des entrées

$$A_{i,j} = \begin{cases} 1, & \text{if} \quad j = i+1 \\ 0, & \text{sinon} \end{cases} ,$$

pour $i$
$in[d]$ et $j \in [S]$. Notez qu'avec la structure ci-dessus, pour tout $\mathbf{z} \in \mathbb{R}^d$ nous avons

$$\nabla f(\mathbf{z}) = h^{-1} A \boldsymbol{\theta}(\mathbf{z}) \enspace .$$

En chaque point $\mathbf{z} \in \mathbb{R}^d$, définissez l'estimateur suivant pour $\boldsymbol{\theta}(\mathbf{z})$. On désigne par $U : \mathbb{R}^d \to \mathbb{R}^S$ une fonction telle que

$$\forall \boldsymbol{u} \in \mathbb{R}^d : \quad U(\boldsymbol{u}) = \left( \frac{\boldsymbol{u}^{\boldsymbol{m}^{(1)}}}{\boldsymbol{m}^{(1)}!}, \ldots, \frac{\boldsymbol{u}^{\boldsymbol{m}^{(S)}}}{\boldsymbol{m}^{(S)}!} \right)^{\top} \enspace .$$

Pour toute $\mathbf{z} \in \mathbb{R}^d$, $t \in [n]$, nous proposons une version régularisée de l'estimateur polynomial local pour $\boldsymbol{\theta}(\mathbf{z})$ comme suit (pour une version classique des estimateurs polynomiaux locaux, voir (Tsybakov, 2009, Section 1.6)) :

$$\hat{\boldsymbol{\theta}}_t(\mathbf{z}) = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^S} \sum_{i=1}^{t} \left[ y_i - \left\langle \boldsymbol{\theta}, U\left( \frac{\mathbf{z}_i - \mathbf{z}}{h_t} \right) \right\rangle \right]^2 K\left( \frac{\mathbf{z}_i - \mathbf{z}}{h_t} \right) + \frac{\lambda_t}{2} \|\boldsymbol{\theta}\|_2^2 \enspace ,$$

où $\lambda_t, h_t > 0$ et $y_i = F(\mathbf{z}_i, \xi(\mathbf{z}_i))$. Considérons maintenant une procédure itérative avec les mises à jour introduites dans (1.4), où $\tilde{\mathbf{g}}_t$ est construit comme suit :

**Estimateur de gradient lisse basé sur un estimateur polynomial local régularisé.** Pour $t \in [n]$, laissons $\lambda_t, h_t > 0$, et construisons l'estimateur du gradient

$$\tilde{\mathbf{g}}_t = h_t^{-1} A \hat{\boldsymbol{\theta}}_t(\mathbf{z}_t) \enspace . \tag{1.36}$$

Au Chapitre 6, nous analysons l'erreur de l'estimateur $\mathbf{x}^*$ dans la norme $\|\cdot\|_2$ de l'algorithme (1.4) avec $\tilde{\mathbf{g}}_t$ comme dans (1.36), et nous montrons qu'il est optimal jusqu'à un facteur logarithmique. De plus, pour toute $\mathbf{z} \in \Theta$, avec des affectations prudentes pour $h_n, \lambda_n > 0$, nous prouvons que $\tilde{\mathbf{g}}(\mathbf{z}) = h_n^{-1} A \hat{\boldsymbol{\theta}}_n(\mathbf{z})$ est un estimateur optimal pour $\nabla f(\mathbf{z})$, par rapport au nombre d'observations $n$.

### Optimisation distribuée

Dans cette section, nous considérons une généralisation du problème (1.29), dans lequel la fonction objectif $f$ est brisée en $m$ morceaux et chaque morceau est donné à un individu. Plus précisément, nous supposons que

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}) \ ,$$

avec $f_i : \mathbb{R}^d \to \mathbb{R}$, pour $i \in [m]$, et chaque $f_i$ est associé à un agent $i$. A chaque tour, pour tout $\mathbf{x} \in \mathbb{R}^d$, l'agent $i$ a accès aux valeurs bruitées de $f_i$, codées par $F_i(\mathbf{x}, \xi(\mathbf{x})) = f_i(\mathbf{x}) + \xi(\mathbf{x})$. Cependant, l'échange d'informations entre les agents est limité à un réseau prescrit de connexions. Nous caractérisons ce réseau par le graphe non dirigé et connecté, $\mathcal{G} = (V, E)$, où $V = [m]$ est l'ensemble des nœuds et chaque nœud correspond à un agent. De plus, $E \subseteq V \times V$ est l'ensemble des arêtes qui induit la notion de voisinage dans le réseau. A savoir, les agents $i \neq j \in [m]$ sont voisins si et seulement si $(i, j) \in E$, et l'échange d'informations n'est possible qu'entre les agents voisins. Considérons une procédure itérative, au tour $t$ let $\mathbf{x}_t^i$ être la mise à jour de l'agent $i$, qui ne dépend que des sorties de $F_i$ et des informations qu'elle perçoit de ses agents voisins. Dans ce cadre, le but est d'obtenir une petite erreur d'optimisation pour les mises à jour de chaque agent après $T$ rounds. Le problème d'intérêt peut être formulé comme le contrôle du terme d'erreur suivant

$$\max_{i \in [n]} \mathbf{E}\left[f(\mathbf{x}_T^i) - f(\mathbf{x}^*)\right] \ , \tag{1.37}$$

où $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \Theta} f(\mathbf{x})$. Maintenant, considérons un cas critique où $\mathcal{G}$ est un graphe complet, c'est-à-dire que les agents ont une information complète les uns sur les autres. Alors, à chaque tour, chaque agent $i$ a accès à l'information d'ordre zéro qui est fournie par chaque $F_i$, et on peut s'attendre à ce que la dureté du problème (1.37) soit équivalente à celle du problème habituel pour avoir une erreur d'optimisation (1.29). Au contraire, si $\mathcal{G}$ est un graphe clairsemé, les agents obtiennent moins d'informations et le problème (1.37) devient plus dur que celui de (1.29). Maintenant, proposons un algorithme itératif, et formalisons cette intuition en pratique.

**Definition 1.9.4.** *Nous appelons $W \in \mathbb{R}^{m \times m}$ une matrice de consensus associée à $\mathcal{G}$ si pour $i, j \in [m]$ elle satisfait à*

$$W_{i,j} \geq 0, \quad W_{i,j} \neq 0 \quad \textit{si} \quad (i,j) \in E \quad \textit{ou} \quad i = j, \quad \textit{et} \quad \sum_{j=1}^{m} W_{i,j} = 1 \ ,$$

Dans ce schéma, sur la base d'une $\mathcal{G}$ donnée, nous sommes autorisés à construire une matrice de consensus associée $W$ qui est donnée aux agents. En conséquence, pour une $W$ donnée, nous sommes prêts à présenter une procédure itérative afin d'obtenir une (1.37). Soit $\eta_t > 0$, et que $\mathbf{x}_1^1, \ldots, \mathbf{x}_n^1$ soient des vecteurs non aléatoires dans $\mathbb{R}^d$. Considérons les mises à jour suivantes

$$\mathbf{x}_{t+1}^{i} = \sum_{j=1}^{m} W_{i,j} \mathrm{Proj}_{\Theta}\left(\mathbf{x}_t^j - \eta_t \mathbf{g}_t^j\right) \quad , \tag{1.38}$$

où $\mathbf{g}_t^j$ est basé sur l'information perçue par $F_j$. Notez qu'à chaque tour, chaque agent $j$ fournit une mise à jour locale $\mathrm{Proj}_{\Theta}\left(\mathbf{x}_t^j - \eta_t \mathbf{g}_t^j\right)$. Ensuite, en utilisant la matrice de consensus $W$, l'agent $i$ prend une somme pondérée des mises à jour locales de ses agents voisins.

Dans le Chapitre 3, nous étudions le problème (1.37), pour une fonction objectif fortement convexe $f$. De plus, nous supposons que chaque $f_i$ est $\beta$-smooth et nous étudions la performance de (1.38), dans lequel nous utilisons l'estimateur (1.35) comme candidat pour $\mathbf{g}_t^i$. Afin d'étudier comment la nature distribuée du problème joue son rôle dans notre analyse, considérons la quantité $\rho = \left\|W - m^{-1}\mathbf{1}\mathbf{1}^\top\right\|_{\mathrm{op}}$, où $\|\cdot\|_{\mathrm{op}}$ est la norme de l'opérateur et $\mathbf{1}$ est un vecteur dans $\mathbb{R}^d$ avec les coordonnées toutes égales à un. Notons que par la définition de $W$, nous avons $\rho \leq 1$. De plus, on note que pour tout graphe connecté $\mathcal{G}$ on peut construire un $W$ associé, dans lequel $\rho < 1$ voir l'exemple qui est fourni par (3.2). Par conséquent, sans perte de généralité, supposons que $\rho < 1$. Pour une fonction objective $f$ lisse à 2, notre analyse décrite dans le Corollaire 3.6.3 montre le fait que le taux de convergence de (1.37) pour l'algorithme (1.38) dépend de $\rho$, et qu'il est de l'ordre de $(1-\rho)^{-1}$. Si $\mathcal{G}$ est un graphe complet, il est raisonnable de prendre $W = m^{-1}\mathbf{1}\mathbf{1}^\top$, ce qui donne $\rho = 0$, et, comme nous nous y attendons, il n'y a aucune trace de la nature distribuée du problème. Cependant, si $\mathcal{G}$ est un graphe clairsemé alors $W$ est clairsemé et $\rho$ est proche de $1$ ce qui provoque une explosion du terme $(1-\rho)^{-1}$.

### Bandits contextuels et équité

Considérons un problème de prise de décision séquentielle où, à chaque étape temporelle, un employeur doit sélectionner un candidat dans une réserve pour l'embaucher pour un poste. L'employeur ne connaît pas les performances d'un candidat s'il est embauché, mais il peut l'apprendre au fil du temps en mesurant les performances de candidats similaires sélectionnés précédemment. Ce scénario peut être formalisé sous la forme d'un problème de bandit contextuel (linéaire) (voir (Auer, 2002; Chu et al., 2011; Lattimore and Szepesvári, 2020) et les références y afférentes), où chaque candidat est représenté par un vecteur de contexte, et après que l'employeur (ou l'agent) ait choisi un candidat, il reçoit une récompense, c'est-à-dire une valeur scalaire mesurant la performance réelle du candidat, qui dépend (linéairement) du contexte.

Dans le cadre ci-dessus, l'objectif typique est de trouver une politique permettant à employeur de sélectionner les candidats ayant les récompenses les plus élevées (Abbasi-Yadkori et al., 2011; Auer, 2002; Auer et al., 2002; Lattimore and Szepesvári, 2020). Cependant, dans certains scénarios importants, cet objectif peut ne pas être approprié ; si les candidats appartiennent à différents groupes sensibles chaque candidat appartient à un groupe sensible différent (par exemple, sur la base de l'appartenance ethnique, du sexe, etc.), la politique qui en résulte peut entraîner une discrimination, voire l'exclusion totale de certains groupes dans le processus de sélection. Il en résulte une injustice sociale. Les personnes qui, en raison de l'injustice sociale, ont soit une rémunération inférieure à celle des autres, par exemple parce qu'elles ont acquis moins de compétences en raison d'une rémunération inférieure à celle des autres. Cela peut se produire lorsque certains groupes ont une récompense attendue plus faible que d'autres, par exemple parce qu'ils ont acquis moins de compétences en raison d'un soutien financier plus faible. d'un soutien financier. ou lorsqu'ils sont évalués injustement, par exemple en raison de préjugés raciaux ou sexistes dans le processus d'évaluation. Nous supposons que les récompenses que nous recevons sont correctes, et non qu'elles sont fausses/biaisées. Dans notre contexte, l'injustice vient strictement de l'impossibilité de comparer les candidats entre eux en raison de circonstances externes, et non d'un biais dans la mesure de la récompense. Un autre exemple se produit lorsque le candidat, s'il est sélectionné, effectuera un travail différent, par exemple lorsque l'employeur est une université et que chaque candidat est un chercheur dans une matière différente (chimie, mathématiques, etc.) ; (Kearns et al., 2017). se produit lorsque chaque candidat du pool, s'il est sélectionné, effectuera un type de travail différent, et que la récompense associée est spécifique au travail. Par exemple , si l'employeur est une université et que chaque candidat est un chercheur dans une discipline différente, alors les récompenses associées à différentes disciplines seront substantiellement différentes et incomparables, par exemple : le nombre de citations varie considérablement d'une discipline à l'autre ; voir (Kearns et al., 2017) pour une discussion. Dans les deux scénarios ci-dessus, il est injuste de comparer directement les récompenses de candidats appartenant à des groupes différents.

Une manière simple de traiter ce problème serait de sélectionner le candidat à embaucher de manière uniforme et aléatoire. Cette politique satisfait une notion d'équité appelée *parité démographique* (voir Calders et al. (2009); Mehrabi et al. (2021) et les références y afférentes), qui exige que la probabilité de sélectionner un candidat d'un certain groupe soit égale pour tous les groupes. Malgré sa simplicité séduisante, cette approche n'est pas du tout satisfaisante. Cependant, comme on peut le constater, cette approche ne tient absolument pas compte de l'objectif de l'employeur, qui est de sélectionner de bons candidats, et elle est également injuste pour les candidats qui ont consacré des efforts à l'acquisition de compétences pour le poste. Dans ce travail, nous proposons une manière équitable de comparer les candidats de différents groupes via le *rang relatif*, c'est-à-dire le rang (ou le quantile) de la récompense par rapport à la distribution des récompenses du groupe du candidat. Nous appelons une politique *group meritocratic fair* (GMF) si elle sélectionne toujours un candidat avec le *rang relatif* le plus élevé. Une telle politique est méritocratique mais uniquement en termes de performance au sein du groupe. Une idée très proche a été introduite dans Kearns et al. (2017) pour les situations où les récompenses des candidats sont disponibles avant la sélection, alors que nous ne connaissons pas de notion similaire dans la littérature sur les bandits à bras multiples.

Une politique GMF nécessite la connaissance du rang relatif de chaque candidat, qui n'est pas directement observé par l'agent. qui n'est pas directement observé par l'agent et dépend du modèle de récompense sous-jacent et de la distribution des récompenses.

Ces dernières années, l'équité algorithmique a fait l'objet d'une grande attention, devenant un vaste domaine de recherche en apprentissage automatique. Le risque que les algorithmes d'apprentissage amplifient les préjugés préexistants et causent du tort aux êtres humains a incité les chercheurs à étudier des solutions pour atténuer ou supprimer l'injustice du prédicteur appris, voir Barocas et al. (2018); Calmon et al. (2017); Chierichetti et al. (2017); Donini et al. (2018); Dwork et al. (2018); Hardt et al. (2016) et les références qui y figurent. L'équité dans les problèmes de décision séquentielle (voir Zhang and Liu (2021) pour une étude) est généralement divisée en deux catégories : l'équité de groupe (GF) et l'équité individuelle. Nous donnons ci-dessous un aperçu de ces notions.

L'équité de groupe exige qu'une certaine mesure statistique soit (approximativement) égale entre les différents groupes sensibles. Un exemple important et pertinent pour ce travail est la *parité démographique*, qui exige que la probabilité que la politique sélectionne un candidat d'un certain groupe soit la même pour tous les groupes. Une notion similaire est utilisée par Chen et al. (2020); Patil et al. (2020), où la probabilité que la politique sélectionne un candidat doit toujours être supérieure à un seuil donné pour tous les candidats. Li et al. (2019) imposent une exigence plus faible concernant la fraction attendue de candidats sélectionnés dans chaque groupe. D'autres exemples de GF dans les problèmes de décision séquentielle sont : *chances égales* (Bechavod et al., 2019) et *chances égales*. (Blum et al., 2018). Sous certaines hypothèses sur les distributions des contextes, la politique GMF et la politique avide que nous proposons satisfont des variantes de la parité démographique à chaque tour.

L'équité individuelle peut être divisée en deux catégories : l'équité par la sensibilisation (FA) (Liu et al., 2017; Wang et al., 2021) et l'équité méritocratique (MF) (Joseph et al., 2018, 2016). L'équité méritocratique repose sur l'idée que des individus similaires doivent être traités de manière similaire et est conçue pour éviter les scénarios du type "le gagnant prend tout", dans lesquels certains individus ne peuvent pas être sélectionnés lorsqu'ils ont une récompense inférieure à celle des autres membres du pool, même si la différence entre les récompenses est très faible. Par exemple, Wang et al. (2021) propose une politique où la probabilité de sélectionner un contexte plutôt qu'un autre est plus faible lorsque le contexte a une récompense inférieure, mais n'est jamais nulle. La MF exige plutôt que les individus moins qualifiés ne soient pas favorisés par rapport aux plus qualifiés, ce qui pourrait se produire pendant le processus d'apprentissage. Par exemple, Joseph et al. (2016) propose un algorithme basé sur les intervalles de confiance, où si l'incertitude entre les meilleurs bras est trop élevée, le bras est sélectionné uniformément au hasard. Cela garantit une équité méritocratique à chaque tour mais a un coût en termes de regret.

Une notion d'équité notable qui ne relève pas de GF et IF est Gillen et al. (2018), où l'idée est que l'on peut seulement "reconnaître l'injustice quand on la voit". Ils supposent qu'il existe un oracle qui peut dire à l'agent pour quels candidats choisis le choix était injuste.

Notre définition de l'équité se situe entre l'équité de groupe et l'équité méritocratique. Elle est méritocratique car elle stipule qu'un candidat ayant un rang relatif plus mauvais qu'un autre ne devrait jamais être sélectionné. Elle est également basée sur les groupes puisque les rangs relatifs dépendent directement de la distribution des récompenses de chaque groupe. Une idée similaire d'équité basée sur le rang relatif a été introduite dans Kearns et al. (2017), qui étudie le problème de la sélection de candidats de différents groupes sur la base de leur score à valeur scalaire lorsque les scores entre les groupes sont incomparables (par exemple, le nombre de citations dans différents domaines de recherche). Contrairement à notre travail, où les récompenses (bruyantes) ne sont observées que pour les candidats sélectionnés, dans Kearns et al. (2017) les scores non bruyants de tous les candidats sont accessibles avant la sélection. Cette différence rend l'estimation du rang relatif plus simple dans Kearns et al. (2017), car les CDF des récompenses peuvent être estimées plus efficacement.

Par exemple, Wang et al. (2021) propose une politique où la probabilité de sélectionner un contexte plutôt qu'un autre est plus grande lorsque le contexte a une plus grande récompense estimée, mais n'est jamais nulle. Cependant, cette probabilité peut être très faible si deux armes ont des récompenses qui sont toujours très éloignées. Un moyen sans doute plus direct d'éviter cet effet serait de s'appuyer sur des notions d'équité de groupe. Par exemple, puisque chaque bras correspond à un groupe différent, une politique satisfait à la *parité démographique* si elle sélectionne chaque bras avec une probabilité égale. Sans hypothèses supplémentaires, la seule politique qui satisfait cette propriété est celle qui sélectionne chaque bras uniformément au hasard à chaque tour, indépendamment des contextes.

## 1.10 Preview of the contributions

### Chapter 2: Exploiting higher order smoothness in derivative-free optimization and continuous bandits

This chapter is based on the paper "Exploiting higher order smoothness in derivative-free optimization and continuous" Akhavan et al. (2020), by Arya Akhavan, Massimiliano Pontil, and Alexandre Tsybakov, in Advances in Neural Information Processing Systems, 33.

We study the problem of zero-order optimization of a strongly convex function. The goal is to find the minimizer of the function by a sequential exploration of its values, under measurement noise. We study the impact of higher order smoothness properties of the function on the optimization error and on the cumulative regret. To solve this problem we consider a randomized approximation of the projected gradient descent algorithm. The gradient is estimated by a randomized procedure involving two function evaluations and a smoothing kernel. We derive upper bounds for this algorithm both in the constrained and unconstrained settings and prove minimax lower bounds for any sequential search method. Our results imply that the zero-order algorithm is nearly optimal in terms of sample complexity and the problem parameters. Based on this algorithm, we also propose an estimator of the minimum value of the function achieving almost sharp oracle behavior. We compare our results with the state-of-the-art, highlighting a number of key improvements.

### Chapter 3: Distributed zero-order optimization under adversarial noise

This chapter is based on the paper "Distributed Zero-Order Optimization under Adversarial Noise" Akhavan et al. (2021), by Arya Akhavan, Massimiliano Pontil, and Alexandre Tsybakov, in Advances in Neural Information Processing Systems, 34.

We study the problem of distributed zero-order optimization for a class of strongly convex functions. They are formed by the average of local objectives, associated to different nodes in a prescribed network. We propose a distributed zero-order projected gradient descent algorithm to solve the problem. Exchange of information within the network is permitted only between neighbouring nodes. An important feature of our procedure is that it can query only function values, subject to a general noise model, that does not require zero mean or independent errors. We derive upper bounds for the average cumulative regret and optimization error of the algorithm which highlight the role played by a network connectivity parameter, the number of variables, the noise level, the strong convexity parameter, and smoothness properties of the local objectives. The bounds indicate some key improvements of our method over the state-of-the-art, both in the distributed and standard zero-order optimization settings. We also comment on lower bounds and observe that the dependency over certain function parameters in the bound is nearly optimal.

### Chapter 4: A gradient estimator via L1-randomization for online zero-order optimization with two point feedback

This chapter is based on the paper "A gradient estimator via L1-randomization for online zero-order optimization with two point feedback", by Arya Akhavan, Evgenii Chzhen, Massimiliano Pontil, and Alexandre Tsybakov, Akhavan et al. (2022a) to appear in Advances in Neural Information Processing Systems, 35.

This chapter studies online zero-order optimization of convex and Lipschitz functions. We present a novel gradient estimator based on two function evaluations and randomization on the $\ell_1$-sphere. Considering different geometries of feasible sets and Lipschitz assumptions we analyse online dual averaging algorithm with our estimator in place of the usual gradient. We consider two types of assumptions on the noise of the zero-order oracle: Canceling noise and adversarial noise. We provide an anytime and completely data-driven algorithm, which is adaptive to all parameters of the problem. In the case of canceling noise that was previously studied in the literature, our guarantees are either comparable or better than state-of-the-art bounds obtained by Duchi et al. (2015) and Shamir (2017) for non-adaptive algorithms. Our analysis is based on deriving a new weighted weighted Poincaré type inequality for the uniform measure on the $\ell_1$-sphere with explicit constants, which may be of independent interest.

### Chapter 5: Zero-order optimization of highly smooth functions: improved analysis and a new algorithm

This chapter is based on a joint work with Evgenii Chzhen, Massimiliano Pontil, and Alexandre Tsybakov.

This chapter studies minimization problems with zero-order noisy oracle information under the assumption that the objective function is highly smooth and possibly satisfies additional properties. We consider two kinds of zero-order projected gradient descent algorithms, which differ in the form of the gradient estimator. The first algorithm uses a gradient estimator based on randomization on the $\ell_2$ sphere, and smoothing kernel due to Bach and Perchet (2016) and it has been used for zero-order optimization of strongly convex functions. We present an improved analysis of this algorithm for the same class of functions and we derive rates of convergence for more general function classes. In particular, we consider functions which satisfy the Polyak-Łojasiewicz condition instead of strong convexity, and the larger class of highly smooth non-convex functions. The second algorithm is based on $\ell_1$-type randomization. We show that this novel algorithm enjoys similar theoretical guarantees as the first one and, in the case of noiseless oracle, it enjoys better bounds. The improvements are achieved by new bounds on bias and variance for both algorithms, which are obtained via Poincaré type inequalities for uniform distributions on $\ell_1$ or $\ell_2$ spheres. The optimality of the upper bounds is discussed and a slightly more general lower bound than in Chapter 2 is presented.

### Chapter 6: Zero-order optimization of highly smooth functions in a passive scheme

This chapter is based on the paper "Estimating the minimizer and the minimum value of a regression function under passive design", by Arya Akhavan, Davit Gogolashvili, and Alexandre Tsybakov, Akhavan et al. (2022b), submitted to Electronic Journal of Statistics.

We propose a new method for estimating the minimizer $\mathbf{x}^*$ and the minimum value $f^*$ of a smooth and strongly convex regression function $f$ from the observations contaminated by random noise. Our estimator $\mathbf{z}_n$ of the minimizer $\mathbf{x}^*$ is based on a version of the projected gradient descent with the gradient estimated by a regularized local polynomial algorithm. Next, we propose a two-stage procedure for estimation of the minimum value $f^*$ of regression function $f$. At the first stage, we construct an accurate enough estimator of $\mathbf{x}^*$, which can be, for example, $\mathbf{z}_n$. At the second stage, we estimate the function value by at the point obtained at the first stage using a rate optimal nonparametric procedure. We derive non-asymptotic upper bounds for the quadratic risk and optimization error of $\mathbf{z}_n$, and for the risk of estimating $f^*$. We establish minimax lower bounds showing that, under certain choice of parameters, the proposed algorithms achieve the minimax optimal rates of convergence on the class of smooth and strongly convex functions.

### Chapter 7: Group meritocratic fairness in linear contextual bandits

This chapter is based on the paper "Group meritocratic fairness in linear contextual bandits", by Riccardo Grazzi, Arya Akhavan, John Isak Texas Falk, Leonardo Cella, and Massimiliano Pontil, Grazzi et al. (2022) to appear in Advances in Neural Information Processing Systems, 35.

We study the linear contextual bandit problem where an agent has to select one candidate from a pool and each candidate belongs to a sensitive group. In this setting, candidates' rewards may not be directly comparable between groups, for example when the agent is an employer hiring candidates from different ethnic groups and some groups have a lower reward due to discriminatory bias and/or social injustice. We propose a notion of fairness that states that the agent's policy is fair when it selects a candidate with highest relative rank, which measures how good the reward is when compared to candidates from the same group. This is a very strong notion of fairness, since the relative rank is not directly observed by the agent and depends on the underlying reward model and on the distribution of rewards. Thus we study the problem of learning a policy which approximates a fair policy under the condition that the contexts are independent between groups and the distribution of rewards of each group is absolutely continuous. In particular, we design a greedy policy which at each round constructs a ridge regression estimator from the observed context-reward pairs, and then computes an estimate of the relative rank of each candidate using the empirical cumulative distribution function. We prove that the greedy policy achieves, after $T$ rounds, up to log factors and with high probability, a fair pseudo-regret of order $\sqrt{dT}$, where $d$ is the dimension of the context vectors. The policy also satisfies demographic parity at each round when averaged over all possible information available before the selection. We finally show with a proof of concept simulation and experiments on the US Census data that our policy achieves sub-linear fair pseudo-regret also in practice.

# Chapter 2

# Exploiting higher order smoothness in derivative-free optimization and continuous bandits

We study the problem of zero-order optimization of a strongly convex function. The goal is to find the minimizer of the function by a sequential exploration of its values, under measurement noise. We study the impact of higher order smoothness properties of the function on the optimization error and on the cumulative regret. To solve this problem we consider a randomized approximation of the projected gradient descent algorithm. The gradient is estimated by a randomized procedure involving two function evaluations and a smoothing kernel. We derive upper bounds for this algorithm both in the constrained and unconstrained settings and prove minimax lower bounds for any sequential search method. Our results imply that the zero-order algorithm is nearly optimal in terms of sample complexity and the problem parameters. Based on this algorithm, we also propose an estimator of the minimum value of the function achieving almost sharp oracle behavior. We compare our results with the state-of-the-art, highlighting a number of key improvements.

## 2.1 Introduction

We study the problem of zero-order stochastic optimization, in which we aim to minimize an unknown strongly convex function via a sequential exploration of its function values, under measurement error, and a closely related problem of continuous (or continuum-armed) stochastic bandits. These problems have received significant attention in the literature, see Agarwal et al. (2010, 2011); Bach and Perchet (2016); Bartlett et al. (2019); Belloni et al. (2015); Bubeck and Cesa-Bianchi (2012); Bubeck et al. (2017); Duchi et al. (2015); Dvurechensky et al. (2018); Flaxman et al. (2005); Hu et al. (2016a,b); Jamieson et al. (2012); Locatelli and Carpentier (2018); Malherbe and Vayatis (2017); Nesterov and Spokoiny (2017); Rakhlin et al. (2012); Saha and Tewari (2011); Shalev-Shwartz (2012); Shamir (2013, 2017); Wang et al. (2018b), and are fundamental for many applications in which the derivatives of the function are either too expensive or impossible to compute. A principal goal of this paper is to exploit higher order smoothness properties of the underlying function in order to improve the performance of search algorithms. We derive upper bounds on the estimation error for a class of projected gradient-like algorithms, as well as close matching lower bounds, that characterize the role played by the number of iterations, the strong convexity parameter, the smoothness parameter, the number of variables, and the noise level.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be the function that we wish to minimize over a closed convex subset $\Theta$ of $\mathbb{R}^d$. Our approach, outlined in Algorithm 1, builds upon previous work in which a sequential algorithm queries at each iteration a pair of function values, under a general noise model. Specifically, at iteration $t$ the current guess $x_t$ for the minimizer of $f$ is used to build two perturbations $x_t + \delta_t$ and $x_t - \delta_t$, where the function values are queried subject to additive measurement errors $\xi_t$ and $\xi_t'$, respectively. The values $\delta_t$ can be chosen in different ways. In this paper, we set $\delta_t = h_t r_r \zeta_t$ (Line 1), where $h_t > 0$ is a suitably chosen small parameter, $r_t$ is random and uniformly distributed on $[-1, 1]$, and $\zeta_t$ is uniformly distributed on the unit sphere. The estimate for the gradient is then computed at Line 2 and used inside a projected gradient method scheme to compute the next exploration point. We introduce a suitably chosen kernel $K$ that allows us to take advantage of higher order smoothness of $f$.

The idea of using randomized procedures for derivative-free stochastic optimization can be traced back to Nemirovski and Yudin (Nemirovsky and Yudin, 1983, Sec. 9.3) who suggested an algorithm with one query per step at point $x_t + h_t\zeta_t$, with $\zeta_t$ uniform on the unit sphere. Its versions with one, two or more queries were studied in several papers including Agarwal et al. (2010); Bach and Perchet (2016); Flaxman et al. (2005); Shamir (2017). Using two queries per step leads to better performance bounds as emphasized in Agarwal et al. (2010); Bach and Perchet (2016); Duchi et al. (2015); Flaxman et al. (2005); Polyak and Tsybakov (1990); Shamir (2017). Randomizing sequences other than uniform on the sphere were also explored: $\zeta_t$ uniformly distributed on a cube Polyak and Tsybakov (1990), Gaussian $\zeta_t$ Nesterov (2011); Nesterov and Spokoiny (2017), $\zeta_t$ uniformly distributed on the vertices of a cube Shamir (2013) or satisfying some general assumptions Dippon (2003a); Duchi et al. (2015). Except for Bach

**Algorithm 1** Zero-Order Stochastic Projected Gradient

```
Requires   Kernel $K : [-1,1] \to \mathbb{R}$, step size $\eta_t > 0$ and parameter $h_t$, for $t = 1, \dots, T$
Initialization   Generate scalars $r(1), \dots, r(T)$ uniformly on the interval $[-1,1]$, vectors $\zeta(1), \dots, \zeta(T)$ uniformly distributed on the unit sphere $S_d = \{\zeta \in \mathbb{R}^d : \|\zeta\| = 1\}$, and choose $x(1) \in \Theta$
For $t = 1, \dots, T$
    1. Let $y(t) = f(x(t) + h_t r(t)\zeta(t)) + \xi(t)$ and $y'(t) = f(x(t) - h_t r(t)\zeta(t)) + \xi'(t)$,
    2. Define $\hat{g}(t) = \frac{d}{2h_t}(y(t) - y'(t))\zeta(t)K(r(t))$
    3. Update $x_{t+1} = x_t - \eta_t \hat{g}(t)$
Return   $(x(t))_{t=1}^T$
```

and Perchet (2016); Dippon (2003a); Polyak and Tsybakov (1990), these works study settings with low smoothness of $f$ (2-smooth or less) and do not invoke kernels $K$ (i.e. $K(\cdot) \equiv 1$ and $r_t \equiv 1$ in Algorithm 1). The use of randomization with smoothing kernels was proposed by Polyak and Tsybakov Polyak and Tsybakov (1990) and further developed by Dippon Dippon (2003a), and Bach and Perchet Bach and Perchet (2016) to whom the current form of Algorithm 1 is due.

In this paper we consider higher order smooth functions $f$ satisfying the generalized Hölder condition with parameter $\beta \geq 2$, cf. inequality (2.1) below. For integer $\beta$, this parameter can be roughly interpreted as the number of bounded derivatives. Furthermore, we assume that $f$ is $\alpha$-strongly convex. For such functions, we address the following two main questions:

(a) What is the performance of Algorithm 1 in terms of the cumulative regret and optimization error, namely what is the explicit dependency of the rate on the main parameters $d, T, \alpha, \beta$?

(b) What are the fundamental limits of any sequential search procedure expressed in terms of minimax optimization error?

To handle task (a), we prove upper bounds for Algorithm 1, and to handle (b), we prove minimax lower bounds for any sequential search method.

**Contributions.** Our main contributions can be summarized as follows: **i)** Under an adversarial noise assumption (cf. Assumption 2.2.1 below), we establish for all $\beta \geq 2$ upper bounds of the order $\frac{d^2}{\alpha}T^{-\frac{\beta-1}{\beta}}$ for the optimization risk and $\frac{d^2}{\alpha}T^{\frac{1}{\beta}}$ for the cumulative regret of Algorithm 1, both for its constrained and unconstrained versions; **ii)** In the case of independent noise satisfying some natural assumptions (including the Gaussian noise), we prove a minimax lower bound of the order $\frac{d}{\alpha}T^{-\frac{\beta-1}{\beta}}$ for the optimization risk when $\alpha$ is not very small. This shows that to within the factor of $d$ the bound for Algorithm 1 cannot be improved for all $\beta \geq 2$; **iii)** We show that, when $\alpha$ is too small, below some specified threshold, higher order smoothness does not help to improve the convergence rate. We prove that in this regime the rate cannot be faster than $d/\sqrt{T}$, which is not better (to within the dependency on $d$) than for derivative-free minimization of simply convex functions Agarwal et al. (2011); Hu et al. (2016b); **iv)** For $\beta = 2$, we obtain a bracketing of the optimal rate between $O(d/\sqrt{\alpha T})$ and $\Omega(d/(\max(1,\alpha)\sqrt{T}))$. In a special case when $\alpha$ is a fixed numerical constant, this validates a conjecture in Shamir (2013) (claimed there as proved fact) that the optimal rate for $\beta = 2$ scales as $d/\sqrt{T}$; **v)** We propose a simple algorithm of estimation of the value $\min_x f(x)$ requiring three queries per step and attaining the optimal rate $1/\sqrt{T}$ for all $\beta \geq 2$. The best previous work on this problem Belitser et al. (2012) suggested a method with exponential complexity and proved a bound of the order $c(d,\alpha)/\sqrt{T}$ for $\beta > 2$ where $c(d,\alpha)$ is an unspecified constant.

**Notation.** Throughout the paper we use the following notation. We let $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ be the standard inner product and Euclidean norm on $\mathbb{R}^d$, respectively. For every close convex set $\Theta \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$ we denote by $\mathrm{Proj}_\Theta(x) = \mathrm{argmin}\{\|z - x\| : z \in \Theta\}$ the Euclidean projection of $x$ to $\Theta$. We assume everywhere that $T \geq 2$. We denote by $\mathcal{F}_\beta(L)$ the class of functions with Hölder smoothness $\beta$ (inequality (2.1) below). Recall that $f$ is $\alpha$-strongly convex for some $\alpha > 0$ if, for any $x, y \in \mathbb{R}^d$ it holds that $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|x - y\|^2$. We further denote by $\mathcal{F}_{\alpha,\beta}(L)$ the class of all $\alpha$-strongly convex functions belonging to $\mathcal{F}_\beta(L)$.

**Organization.** We start in Section 2.2 with some preliminary results on the gradient estimator. Section 2.3 presents our upper bounds for Algorithm 1, both in the constrained and unconstrained case. In Section 2.4 we observe that a slight modification of Algorithm 1 can be used to estimated the minimum value (rather than the minimizer) of $f$. Section 2.4 presents improved upper bounds in the case $\beta = 2$. In Section 2.6 we establish minimax lower bounds. Finally, Section 2.7 contrasts our results with previous work in the literature and discusses future directions of research.

## 2.2 Preliminaries

In this section, we give the definitions, assumptions and basic facts that will be used throughout the paper. For $\beta > 0$, let $\ell$ be the greatest integer strictly less than $\beta$. We denote by $\mathcal{F}_\beta(L)$ the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that are $\ell$ times differentiable and satisfy, for all $x, z \in \Theta$ the Hölder-type condition

$$\left| f(z) - \sum_{0 \leq |m| \leq \ell} \frac{1}{m!} D^m f(x)(z - x)^m \right| \leq L\|z - x\|^\beta, \tag{2.1}$$

where $L > 0$, the sum is over the multi-index $m = (m_1, ..., m_d) \in \mathbb{N}^d$, we used the notation $m! = m_1! \cdots m_d!$, $|m| = m_1 + \cdots + m_d$, and we defined

$$D^m f(x)\nu^m = \frac{\partial^{|m|} f(x)}{\partial^{m_1} x_1 \cdots \partial^{m_d} x_d} \nu_1^{m_1} \cdots \nu_d^{m_d}, \quad \forall \nu = (\nu_1, \ldots, \nu_d) \in \mathbb{R}^d.$$

In this paper, we assume that the gradient estimator defined by Algorithm 1 uses a kernel function $K : [-1, 1] \to \mathbb{R}$ satisfying

$$\int K(u)du = 0, \int uK(u)du = 1, \int u^j K(u)du = 0, \ j = 2, \dots, \ell, \ \int |u|^\beta |K(u)|du < \infty.$$

Examples of such kernels obtained as weighted sums of Legendre polynomials are given in Polyak and Tsybakov (1990) and further discussed in Bach and Perchet (2016).

**Assumption 2.2.1.** *It holds, for all $t \in \{1, \dots, T\}$, that: (i) the random variables $\xi_t$ and $\xi_t'$ are independent from $\zeta_t$ and from $r_t$, and the random variables $\zeta_t$ and $r_t$ are independent; (ii) $\mathbb{E}[\xi_t^2] \le \sigma^2$, and $\mathbb{E}[(\xi_t')^2] \le \sigma^2$, where $\sigma \ge 0$.*

Note that we do not assume $\xi_t$ and $\xi_t'$ to have zero mean. Moreover, they can be non-random and no independence between noises on different steps is required, so that the setting can be considered as adversarial. Having such a relaxed set of assumptions is possible because of randomization that, for example, allows the proofs go through without assuming the zero mean noise.

We will also use the following assumption.

**Assumption 2.2.2.** *Function $f : \mathbb{R}^d \to \mathbb{R}$ is 2-smooth, that is, differentiable on $\mathbb{R}^d$ and such that $\|\nabla f(x) - \nabla f(x')\| \le \bar{L}\|x - x'\|$ for all $x, x' \in \mathbb{R}^d$, where $\bar{L} > 0$.*

It is easy to see that this assumption implies that $f \in \mathcal{F}_2(\bar{L}/2)$. The following lemma gives a bound on the bias of the gradient estimator.

**Lemma 2.2.3.** *Let $f \in \mathcal{F}_\beta(L)$, with $\beta > 1$ and let Assumption 2.2.1 hold. Let $\hat{g}_t$ and $x_t$ be defined by Algorithm 1 and let $\kappa_\beta = \int |u|^\beta |K(u)|du$. Then*

$$\|\mathbb{E}[\hat{g}_t \,|\, x_t] - \nabla f(x_t)\| \le \kappa_\beta L d h_t^{\beta-1}.$$

If $K$ be a weighted sum of Legendre polynomials, $\kappa_\beta \le 2\sqrt{2}\beta$, with $\beta \ge 1$ (see e.g., (Bach and Perchet, 2016, Appendix A.3)).

The next lemma provides a bound on the stochastic variability of the estimated gradient by controlling its second moment.

**Lemma 2.2.4.** *Let Assumption 2.2.1 hold, let $\hat{g}_t$ and $x_t$ be defined by Algorithm 1 and set $\kappa = \int K^2(u)du$. Then*

*(i) If $\Theta \subseteq \mathbb{R}^d$, $\nabla f(x^*) = 0$ and Assumption 2.2.2 holds,*

$$\mathbb{E}[\|\hat{g}_t\|^2 \,|\, x_t] \le 9\kappa \bar{L}^2 \left( d\|x_t - x^*\|^2 + \frac{d^2 h_t^2}{8} \right) + \frac{3\kappa d^2 \sigma^2}{2h_t^2},$$

*(ii) If $f \in \mathcal{F}_2(L)$ and $\Theta$ is a closed convex subset of $\mathbb{R}^d$ such that $\max_{x\in\Theta}\|\nabla f(x)\| \leq G$, then*

$$\mathbb{E}[\|\hat{g}_t\|^2 \,|\, x_t] \leq 9\kappa\left(G^2 d + \frac{L^2 d^2 h_t^2}{2}\right) + \frac{3\kappa d^2\sigma^2}{2h_t^2}.$$

## 2.3 Upper bounds

In this section, we provide upper bounds on the cumulative regret and on the optimization error of Algorithm 1, which are defined as

$$\sum_{t=1}^{T} \mathbb{E}[f(x_t) - f(x)],$$

and

$$\mathbb{E}[f(\hat{x}_T) - f(x^*)],$$

respectively, where $x \in \Theta$ and $\hat{x}_T$ is an estimator after $T$ queries. Note that the provided upper bound for cumulative regret is valid for any $x \in \Theta$.

First we consider Algorithm 1 when the convex set $\Theta$ is bounded (constrained case).

**Theorem 2.3.1.** (Upper Bound, Constrained Case.) *Let $f \in \mathcal{F}_{\alpha,\beta}(L)$ with $\alpha, L > 0$ and $\beta \geq 2$. Let Assumptions 2.2.1 and 2.2.2 hold and let $\Theta$ be a convex compact subset of $\mathbb{R}^d$. Assume that $\max_{x\in\Theta}\|\nabla f(x)\| \leq G$. If $\sigma > 0$ then the cumulative regret of Algorithm 1 with*

$$h_t = \left(\frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2}\right)^{\frac{1}{2\beta}} t^{-\frac{1}{2\beta}}, \quad \eta_t = \frac{2}{\alpha t}, \quad t = 1, \dots, T$$

*satisfies*

$$\forall x \in \Theta : \sum_{t=1}^{T} \mathbb{E}[f(x_t) - f(x)] \leq \frac{1}{\alpha}\left(d^2\left(A_1 T^{1/\beta} + A_2\right) + A_3 d \log T\right), \tag{2.2}$$

*where $A_1 = 3\beta(\kappa\sigma^2)^{\frac{\beta-1}{\beta}}(\kappa_\beta L)^{\frac{2}{\beta}}$, $A_2 = \bar{c}\bar{L}^2(\sigma/L)^{\frac{2}{\beta}} + 9\kappa G^2/d$ with constant $\bar{c} > 0$ depending only on $\beta$, and $A_3 = 9\kappa G^2$. The optimization error of averaged estimator $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ satisfies*

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{1}{\alpha}\left(d^2\left(\frac{A_1}{T^{\frac{\beta-1}{\beta}}} + \frac{A_2}{T}\right) + A_3\frac{d\log T}{T}\right), \tag{2.3}$$

*where $x^* = \arg\min_{x\in\Theta} f(x)$. If $\sigma = 0$, then the cumulative regret and the optimization error of Algorithm 1 with any $h_t$ chosen small enough and $\eta_t = \frac{2}{\alpha t}$ satisfy the bounds (2.2) and (2.3), respectively, with $A_1 = 0, A_2 = 9\kappa G^2/d$ and $A_3 = 10\kappa G^2$.*

*Proof sketch.* We use the definition of Algorithm 1 and strong convexity of $f$ to obtain an upper bound for $\sum_{t=1}^{T}\mathbb{E}[f(x_t) - f(x)\,|\,x_t]$, which depends on the bias term $\sum_{t=1}^{T}\|\mathbb{E}[\hat{g}_t\,|\,x_t] - \nabla f(x_t)\|$ and on the stochastic error term $\sum_{t=1}^{T}\mathbb{E}[\|\hat{g}_t\|^2]$. By substituting $h_t$ (that is derived from balancing the two terms) and $\eta_t$ in Lemmas 2.2.3 and 2.2.4 we obtain upper bounds for

$\sum_{t=1}^T \|\mathbb{E}[\hat{g}_t\,|\,x_t]-\nabla f(x_t)\|$ and $\sum_{t=1}^T \mathbb{E}[\|\hat{g}_t\|^2]$ that imply the desired upper bound for $\sum_{t=1}^T \mathbb{E}[f(x_t)-f(x)\,|\,x_t]$ due to a recursive argument in the spirit of Bartlett et al. (2008). □

In the non-noisy case ($\sigma = 0$) we get the rate $\frac{d}{\alpha}\log T$ for the cumulative regret, and $\frac{d}{\alpha}\frac{\log T}{T}$ for the optimization error. In what concerns the optimization error, this rate is not optimal since one can achieve much faster rate under strong convexity Nesterov and Spokoiny (2017). However, for the cumulative regret in our derivative-free setting it remains an open question whether the result of Theorem 2.3.1 can be improved. Previous papers on derivative-free online methods with no noise (Agarwal et al., 2010; Duchi et al., 2015; Flaxman et al., 2005) provide slower rates than $(d/\alpha)\log T$. The best known so far is $(d^2/\alpha)\log T$, cf. (Agarwal et al., 2010, Corollary 5). We may also notice that the cumulative regret bounds of Theorem 2.3.1 trivially extend to the case when we query functions $f_t$ depending on $t$ rather than a single $f$. Another immediate fact is that on the r.h.s. of inequalities (2.2) and (2.3) we can take the minimum with $GBT$ and $GB$, respectively, where $B$ is the Euclidean diameter of $\Theta$. Finally, the factor $\log T$ in the bounds for the optimization error can be eliminated by considering averaging from $T/2$ to $T$ rather than from 1 to $T$, in the spirit of Rakhlin et al. (2012). We refer to Section 2.8 for the details and proofs of these facts.

We now study the performance of Algorithm 1 when $\Theta = \mathbb{R}^d$. In this case we make the following choice for the parameters $h_t$ and $\eta_t$ in Algorithm 1:

$$\begin{aligned} h_t &= T^{-\frac{1}{2\beta}}, \quad \eta_t = \frac{1}{\alpha T}, \quad t = 1,\dots,T_0,\\ h_t &= t^{-\frac{1}{2\beta}}, \quad \eta_t = \frac{2}{\alpha t}, \quad t = T_0+1,\dots,T, \end{aligned} \tag{2.4}$$

where $T_0 = \max\left\{k \geq 0 : C_1\bar{L}^2 d > \alpha^2 k/2\right\}$ and $C_1$ is a positive constant[1] depending only on the kernel $K(\cdot)$ (this is defined in the proof of Theorem 2.3.2 in Section 2.8) and recall $\bar{L}$ is the Lipschitz constant on the gradient $\nabla f$. Finally, define the estimator

$$\bar{x}_{T_0,T} = \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} x_t. \tag{2.5}$$

**Theorem 2.3.2.** (Upper Bounds, Unconstrained Case.) *Let $f \in \mathcal{F}_{\alpha,\beta}(L)$ with $\alpha, L > 0$ and $\beta \geq 2$. Let Assumptions 2.2.1 and 2.2.2 hold. Assume also that $\alpha > \sqrt{C_* d/T}$, where $C_* > 72\kappa\bar{L}^2$. Let $x_t$'s be the updates of Algorithm 1 with $\Theta = \mathbb{R}^d$, $h_t$ and $\eta_t$ as in (2.4) and a non-random $x_1 \in \mathbb{R}^d$. Then the estimator defined by (2.5) satisfies*

$$\mathbb{E}[f(\bar{x}_{T_0,T}) - f(x^*)] \leq C\kappa\bar{L}^2\frac{d}{\alpha T}\|x_1 - x^*\|^2 + C\frac{d^2}{\alpha}\left((\kappa_\beta L)^2 + \kappa\left(\bar{L}^2+\sigma^2\right)\right)T^{-\frac{\beta-1}{\beta}}$$

*where $C > 0$ is a constant depending only on $\beta$ and $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$.*

[1] If $T_0 = 0$ the algorithm does not use (2.4). Assumptions of Theorem 2.3.2 are such that condition $T > T_0$ holds.

*Proof sketch.* As in the proof of Theorem 2.3.1, we apply Lemmas 2.2.3 and 2.2.4. But we can only use Lemma 2.2.4(i) and not Lemma 2.2.4(ii) and thus the bound on the stochastic error now involves $\|x_t - x^*\|^2$. So, after taking expectations, we need to control an additional term containing $r_t = \mathbb{E}[\|x_t - x^*\|^2]$. However, the issue concerns only small $t$ ($t \leq T_0 \sim d^2/\alpha$) since for bigger $t$ this term is compensated due to the strong convexity with parameter $\alpha > \sqrt{C_* d/T}$. This motivates the method where we use the first $T_0$ iterations to get a suitably good (but not rate optimal) bound on $r_{T_0+1}$ and then proceed analogously to Theorem 2.3.1 for iterations $t \geq T_0 + 1$. □

## 2.4 Estimation of $f(x^*)$

In this section, we apply the above results to estimation of the minimum value $f(x^*) = \min_{x \in \Theta} f(x)$ for functions $f$ in the class $\mathcal{F}_{\alpha,\beta}(L)$. The literature related to this problem assumes that $x_t$'s are either i.i.d. with density bounded away from zero on its support Tsybakov (1990a) or $x_t$'s are chosen sequentially Belitser et al. (2012); Mokkadem and Pelletier (2007). In the fist case, from the results in Tsybakov (1990a) one can deduce that $f(x^*)$ cannot be estimated better than at the slow rate $T^{-\beta/(2\beta+d)}$. For the second case, which is our setting, the best result so far is obtained in Belitser et al. (2012). The estimator of $f(x^*)$ in Belitser et al. (2012) is defined via a multi-stage procedure whose complexity increases exponentially with the dimension $d$ and it is shown to achieve (asymptotically, for $T$ greater than an exponent of $d$) the $c(d,\alpha)/\sqrt{T}$ rate for functions in $\mathcal{F}_{\alpha,\beta}(L)$ with $\beta > 2$. Here, $c(d,\alpha)$ is some constant depending on $d$ and $\alpha$ in an unspecified way.

Observe that $f(\bar{x}_T)$ is not an estimator since it depends on the unknown $f$, so Theorem 2.3.1 does not provide a result about estimation of $f(x^*)$. In this section, we show that using the computationally simple Algorithm 1 and making one more query per step (that is, having three queries per step in total) allows us to achieve the $1/\sqrt{T}$ rate for all $\beta \geq 2$ with no dependency on the dimension in the main term. Note that the $1/\sqrt{T}$ rate cannot be improved. Indeed, one cannot estimate $f(x^*)$ with a better rate even using the ideal but non-realizable oracle that makes all queries at point $x^*$. That is, even if $x^*$ is known and we sample $T$ times $f(x^*) + \xi_t$ with independent centered variables $\xi_t$, the error is still of the order $1/\sqrt{T}$.

In order to construct our estimator, at any step $t$ of Algorithm 1 we make along with $y_t$ and $y_t'$ the third query $y_t'' = f(x_t) + \xi_t''$, where $\xi_t''$ is some noise and $x_t$ are the updates of Algorithm 1. We estimate $f(x^*)$ by $\hat{M} = \frac{1}{T}\sum_{t=1}^{T} y_t''$. The properties of estimator $\hat{M}$ are summarized in the next theorem, which is an immediate corollary of Theorem 2.3.1.

**Theorem 2.4.1.** *Let the assumptions of Theorem 2.3.1 be satisfied. Let $\sigma > 0$ and assume that $(\xi_t'')_{t=1}^T$ are independent random variables with $\mathbb{E}[\xi_t''] = 0$ and $\mathbb{E}[(\xi_t'')^2] \leq \sigma^2$ for $t =$*

$1, \ldots, T$. *If $f$ attains its minimum at point $x^* \in \Theta$, then*

$$\mathbb{E}|\hat{M} - f(x^*)| \le \frac{\sigma}{T^{\frac{1}{2}}} + \frac{1}{\alpha}\left(d^2\left(\frac{A_1}{T^{\frac{\beta-1}{\beta}}} + \frac{A_2}{T}\right) + A_3 \frac{d \log T}{T}\right).$$

**Remark 2.4.2.** *With three queries per step, the risk (error) of the oracle that makes all queries at point $x^*$ does not exceed $\sigma/\sqrt{3T}$. Thus, for $\beta > 2$ the estimator $\hat{M}$ achieves asymptotically as $T \to \infty$ the oracle risk up to a numerical constant factor. We do not obtain such a sharp property for $\beta = 2$, in which case the remainder term in Theorem 2.4.1 accounting for the accuracy of Algorithm 1 is of the same order as the main term $\sigma/\sqrt{T}$.*

Note that in Theorem 2.4.1 the noises $(\xi_t'')_{t=1}^T$ are assumed to be independent and zero mean random variables, which is essential to obtain the $1/\sqrt{T}$ rate. Nevertheless, we do not require independence between the noises $(\xi_t'')_{t=1}^T$ and the noises in the other two queries $(\xi_t)_{t=1}^T$ and $(\xi_t')_{t=1}^T$. Another interesting point is that for $\beta = 2$ the third query is not needed and $f(x^*)$ is estimated with the $1/\sqrt{T}$ rate either by $\hat{M} = \frac{1}{T}\sum_{t=1}^T y_t$ or by $\hat{M} = \frac{1}{T}\sum_{t=1}^T y_t'$. This is an easy consequence of the above argument, the property (2.16) – see Lemma 2.8.3 in Section 2.8 – which is specific for the case $\beta = 2$, and the fact that the optimal choice of $h_t$ is of order $t^{-1/4}$ for $\beta = 2$.

## 2.5 Improved bounds for $\beta = 2$

In this section, we consider the case $\beta = 2$ and obtain improved bounds that scale as $d$ rather than $d^2$ with the dimension in the constrained optimization setting analogous to Theorem 2.3.1. First note that for $\beta = 2$ we can simplify the algorithm. The use of kernel $K$ is redundant when $\beta = 2$, and therefore in this section we define the approximate gradient as

$$\hat{g}_t = \frac{d}{2h_t}(y_t - y_t')\zeta_t, \tag{2.6}$$

where $y_t = f(x_t + h_t\zeta_t) + \xi_t$ and $y_t' = f(x_t - h_t\zeta_t) + \xi_t'$. A well-known observation that goes back to Nemirovsky and Yudin (1983) consists in the fact that $\hat{g}_t$ defined in (2.6) is an unbiased estimator of the gradient at point $x_t$ of the surrogate function $\hat{f}_t$ defined by

$$\hat{f}_t(x) = \mathbb{E} f(x + h_t\tilde{\zeta}), \quad \forall x \in \mathbb{R}^d,$$

where the expectation $\mathbb{E}$ is taken with respect to the random vector $\tilde{\zeta}$ uniformly distributed on the unit ball $B_d = \{u \in \mathbb{R}^d : \|u\| \le 1\}$. The properties of the surrogate $\hat{f}_t$ are described in Lemmas 2.8.2 and 2.8.3 presented in Section 2.8.

The improvement in the rate that we get for $\beta = 2$ is due to the fact that we can consider Algorithm 1 with $\hat{g}_t$ defined in (2.6) as the SGD for the surrogate function. Then the bias of approximating $f$ by $\hat{f}_t$ scales as $h_t^2$, which is smaller than the squared bias of approximating the gradient arising in the proof of Theorem 2.3.1 that scales as $d^2 h_t^{2(\beta-1)} = d^2 h_t^2$ when $\beta = 2$. On the other hand, the stochastic variability terms are the same for both methods of proof. This explains the gain in dependency on $d$. However, this technique does not work for $\beta > 2$ since then the error of approximating $f$ by $\hat{f}_t$, which is of the order $h_t^\beta$ (with $h_t$ small), becomes too large compared to the bias $d^2 h_t^{2(\beta-1)}$ of Theorem 2.3.1.

**Theorem 2.5.1.** *Let $f \in \mathcal{F}_{\alpha,2}(L)$ with $\alpha, L > 0$. Let Assumption 2.2.1 hold and let $\Theta$ be a convex compact subset of $\mathbb{R}^d$. Assume that $\max_{x\in\Theta} \|\nabla f(x)\| \le G$. If $\sigma > 0$ then for the updates $x_t$ as in item 3 of Algorithm 1 with $\hat{g}_t$ defined in (2.6) and parameters $h_t = \left(\frac{3d^2\sigma^2}{4L\alpha t + 9L^2 d^2}\right)^{1/4}$ and $\eta_t = \frac{1}{\alpha t}$ we have*

$$\forall x \in \Theta: \quad \mathbb{E}\sum_{t=1}^{T} \big(f(x_t) - f(x)\big) \le \min\left(GBT, 2\sqrt{3L}\sigma\frac{d}{\sqrt{\alpha}}\sqrt{T} + A_4\frac{d^2}{\alpha}\log T\right), \tag{2.7}$$

*where $B$ is the Euclidean diameter of $\Theta$ and $A_4 = 6.5L\sigma + 22G^2/d$. Moreover, if $x^* = \arg\min_{x\in\Theta} f(x)$ the optimization error of averaged estimator $\bar{x}_T = \frac{1}{T}\sum_{t=1}^T x_t$ is bounded as*

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \le \min\left(GB, 2\sqrt{3L}\sigma\frac{d}{\sqrt{\alpha T}} + A_4\frac{d^2}{\alpha}\frac{\log T}{T}\right). \tag{2.8}$$

*Finally, if $\sigma = 0$, then the cumulative regret of the same procedure $x_t$ with any $h_t > 0$ chosen small enough and $\eta_t = \frac{1}{\alpha t}$ and the optimization error of its averaged version are of the order $\frac{d^2}{\alpha}\log T$ and $\frac{d^2}{\alpha}\frac{\log T}{T}$, respectively.*

Note that the terms $\frac{d^2}{\alpha}\log T$ and $\frac{d^2}{\alpha}\frac{\log T}{T}$ appearing in these bounds can be improved to $\frac{d}{\alpha}\log T$ and $\frac{d}{\alpha}\frac{\log T}{T}$ at the expense of assuming that the norm $\|\nabla f\|$ is uniformly bounded by $G$ not only on $\Theta$ but also on a large enough Euclidean neighborhood of $\Theta$. Moreover, the $\log T$ factor in the bounds for the optimization error can be eliminated by considering averaging from $T/2$ to $T$ rather than from 1 to $T$ in the spirit of Rakhlin et al. (2012). We refer to Section 2.8 for the details and proofs of these facts. A major conclusion is that, when $\sigma > 0$ and we consider the optimization error, those terms are negligible with respect to $d/\sqrt{\alpha T}$ and thus an attainable rate is $\min(1, d/\sqrt{\alpha T})$.

We close this section by noting, in connection with the bandit setting, that the bound (2.7) extends straightforwardly (up to a change in numerical constants) to the cumulative regret of the form $\mathbb{E}\sum_{t=1}^T \big(f_t(x_t \pm h_t\zeta_t) - f_t(x)\big)$, where the losses are measured at the query points and $f$ depends on $t$. This fact follows immediately from the proof of Theorem 2.5.1 presented in Section 2.8 and the property (2.16), see Lemma 2.8.3 in Section 2.8.

## 2.6 Lower bound

In this section we prove a minimax lower bound on the optimization error over all sequential strategies that allow the query points depend on the past. For $t = 1, \dots, T$, we assume that

$y_t = f(z_t) + \xi_t$ and we consider strategies of choosing the query points such that $z_1 \in \mathbb{R}^d$ is a random variable and $z_t = \Phi_t(z_1, y_1, \dots, z_{t-1}, y_{t-1}, \boldsymbol{\zeta}_t)$ for $t \geq 2$, where $\Phi_t$'s are measurable functions with values in $\mathbb{R}^d$, and $\{\boldsymbol{\zeta}_t\}$ is a sequence of random variables with values in some measurable space $(\mathcal{Z}, \mathcal{U})$ (a randomizing sequence) satisfying the condition that $\boldsymbol{\zeta}_t$ is independent of $(z_1, y_1, \dots, z_{t-1}, y_{t-1})$. We denote by $\Pi_T$ the set of all such strategies. The noises $\xi_1, \dots, \xi_T$ are assumed in this section to be independent with cumulative distribution function $F$ satisfying the condition

$$\int \log\big(dF(u)/dF(u+v)\big)dF(u) \leq I_0 v^2, \quad |v| < v_0, \tag{2.9}$$

for some $0 < I_0 < \infty$, $0 < v_0 \leq \infty$, and such that $\xi_t$ is independent of $(z_1, y_1, \dots, z_{t-1}, y_{t-1}, \boldsymbol{\zeta}_t)$. Using the second order expansion of the logarithm w.r.t. $v$, one can verify that this assumption is satisfied when $F$ has a smooth enough density with finite Fisher information. For example, for Gaussian distribution $F$ this condition holds with $v_0 = \infty$. Note that the class $\Pi_T$ includes the sequential strategy of Algorithm 1 that corresponds to taking $T$ as an even number, and choosing $z_t = x_t + \zeta_t r_t$ and $z_t = x_t - \zeta_t r_t$ for even $t$ and odd $t$, respectively.

**Theorem 2.6.1.** *Let $\Theta = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$. For $\alpha, L > 0, \beta \geq 2$, let $\mathcal{F}'_{\alpha,\beta}$ denote the set of functions $f$ that attain their minimum over $\mathbb{R}^d$ in $\Theta$ and belong to $\mathcal{F}_{\alpha,\beta}(L) \cap \{f : \max_{x\in\Theta} \|\nabla f(x)\| \leq G\}$, where $G > 2\alpha$. Then for any strategy in the class $\Pi_T$ we have*

$$\sup_{f \in \mathcal{F}'_{\alpha,\beta}} \mathbb{E}\big[f(z_T) - \min_x f(x)\big] \geq C \min\Big(\max(\alpha, T^{-1/2+1/\beta}), \frac{d}{\sqrt{T}}, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\Big), \tag{2.10}$$

*and*

$$\sup_{f \in \mathcal{F}'_{\alpha,\beta}} \mathbb{E}\big[\, \|z_T - x^*(f)\|^2 \,\big] \geq C \min\Big(1, \frac{d}{T^{\frac{1}{\beta}}}, \frac{d}{\alpha^2} T^{-\frac{\beta-1}{\beta}}\Big), \tag{2.11}$$

*where $C > 0$ is a constant that does not depend of $T, d$, and $\alpha$, and $x^*(f)$ is the minimizer of $f$ on $\Theta$.*

The proof is given in Section 2.8. It extends the proof technique of Polyak and Tsybakov [28], by applying it to more than two probe functions. The proof takes into account dependency on the dimension $d$, and on $\alpha$. The final result is obtained by applying Assouad's Lemma, see e.g. Tsybakov (2009).

We stress that the condition $G > 2\alpha$ in this theorem is necessary. It should always hold if the intersection $\mathcal{F}_{\alpha,\beta}(L) \cap \{f : \max_{x\in\Theta} \|\nabla f(x)\| \leq G\}$ is not empty. Notice also that the threshold $T^{-1/2+1/\beta}$ on the strong convexity parameter $\alpha$ plays an important role in bounds (2.10) and (2.11). Indeed, for $\alpha$ below this threshold, the bounds start to be independent of $\alpha$. Moreover, in this regime, the rate of (2.10) becomes $\min(T^{1/\beta}, d)/\sqrt{T}$, which is asymptotically $d/\sqrt{T}$ and thus not better as function of $T$ than the rate attained for zero-order minimization of simply convex functions Agarwal et al. (2011); Belloni et al. (2015). Intuitively, it seems reasonable that $\alpha$-strong convexity should be of no added value for very small $\alpha$. Theorem 2.6.1

allows us to quantify exactly how small such $\alpha$ should be. Also, quite naturally, the threshold becomes smaller when the smoothness $\beta$ increases.
Finally note that for $\beta = 2$ the lower bounds (2.10) and (2.11) are, in the interesting regime of large enough $T$, of order $d/(\max(\alpha, 1)\sqrt{T})$ and $d/(\max(\alpha^2, 1)\sqrt{T})$, respectively. This highlights the near minimax optimal properties of Algorithm 1 in the setting of Theorem 2.5.1.

## 2.7 Discussion and related work

There is a great deal of attention to zero-order feedback stochastic optimization and convex bandits problems in the recent literature. Several settings are studied: (i) deterministic in the sense that the queries contain no random noise and we query functions $f_t$ depending on $t$ rather than $f$ where $f_t$ are Lipschitz or 2-smooth Agarwal et al. (2010); Flaxman et al. (2005); Nesterov (2011); Nesterov and Spokoiny (2017); Saha and Tewari (2011); Shamir (2017); (ii) stochastic with two-point feedback where the two noisy evaluations are obtained with the same noise and the noisy functions are Lipschitz or 2-smooth Duchi et al. (2015); Nesterov (2011); Nesterov and Spokoiny (2017) (this setting does not differ much from (i) in terms of the analysis and the results); (iii) stochastic, where the noises $\xi_i$ are independent zero-mean random variables Agarwal et al. (2011); Bach and Perchet (2016); Bartlett et al. (2019); Dippon (2003a); Fabian (1967a); Jamieson et al. (2012); Locatelli and Carpentier (2018); Polyak and Tsybakov (1990); Shamir (2013). In this paper, we considered a setting, which is more general than (iii) by allowing for adversarial noise (no independence or zero-mean assumption in contrast to (iii), no Lipschitz assumption in contrast to settings (i) and (ii)), which are both covered by our results when the noise is set to zero.

One part of our results are bounds on the cumulative regret, cf. (2.2) and (2.7). We emphasize that they remain trivially valid if the queries are from $f_t$ depending on $t$ instead of $f$, and thus cover the setting (i). To the best of our knowledge, there were no such results in this setting previously, except for Bach and Perchet (2016) that gives bounds with suboptimal dependency on $T$ in the case of classical (non-adversarial) noise. In the non-noisy case, we get bounds on the cumulative regret with faster rates than previously known for the setting (i). It remains an open question whether these bounds can be improved.

The second part of our results dealing with the optimization error $\mathbb{E}[f(\bar{x}_T) - f(x^*)]$ is closely related to the work on derivative-free stochastic optimization under strong convexity and smoothness assumptions initiated in Fabian (1967a); Polyak and Tsybakov (1990) and more recently developed in Bach and Perchet (2016); Dippon (2003a); Jamieson et al. (2012); Shamir (2013). It was shown in Polyak and Tsybakov (1990) that the minimax optimal rate for $f \in \mathcal{F}_{\alpha,\beta}(L)$ scales as $c(\alpha, d)T^{-(\beta-1)/\beta}$, where $c(\alpha, d)$ is an unspecified function of $\alpha$ and $d$ (for $d = 1$ an upper bound of the same order was earlier established in Fabian (1967a)). The issue of establishing non-asymptotic fundamental limits as function of the main parameters of the problem ($\alpha$, $d$ and $T$) was first addressed in Jamieson et al. (2012) giving a lower bound

$\Omega(\sqrt{d/T})$ for $\beta = 2$. This was improved to $\Omega(d/\sqrt{T})$ when $\alpha \asymp 1$ by Shamir Shamir (2013) who conjectured that the rate $d/\sqrt{T}$ is optimal for $\beta = 2$, which indeed follows from our Theorem 2.5.1 (although Shamir (2013) claims the optimality as proved fact by referring to results in Agarwal et al. (2010), such results cannot be applied in setting (iii) because the noise cannot be considered as Lipschitz). A result similar to Theorem 2.5.1 is stated without proof in Bach and Perchet (Bach and Perchet, 2016, Proposition 7) but not for the cumilative regret and with a suboptimal rate in the non-noisy case. For integer $\beta \geq 3$, Bach and Perchet Bach and Perchet (2016) present explicit upper bounds as functions of $\alpha$, $d$ and $T$ with, however, suboptimal dependency on $T$ except for their Proposition 8 that is problematic (see Section 2.8 for the details). Finally, by slightly modifying the proof of Theorem 2.3.1 we get that the estimation risk $\mathbb{E}\big[\,\|\bar{x}_T - x^*\|^2\,\big]$ is $O((d^2/\alpha^2)T^{-(\beta-1)/\beta})$, which is to within factor $d$ of the main term in the lower bound (2.11) (see Section 2.8 for details).

The lower bound in Theorem 2.6.1 is, to the best of our knowledge, the first result providing non-asymptotic fundamental limits under general configuration of $\alpha$, $d$ and $T$. The known lower bounds Jamieson et al. (2012); Polyak and Tsybakov (1990); Shamir (2013) either give no explicit dependency on $\alpha$ and $d$, or treat the special case $\beta = 2$ and $\alpha \asymp 1$. Moreover, as an interesting consequence of our lower bound we find that, for small strong convexity parameter $\alpha$ (namely, below the $T^{-1/2+1/\beta}$ threshold), the best achievable rate cannot be substantially faster than for simply convex functions, at least for moderate dimensions. Indeed, for such small $\alpha$, our lower bound is asymptotically $\Omega(d/\sqrt{T})$ independently of the smoothness index $\beta$ and on $\alpha$, while the achievable rate for convex functions is shown to be $d^{16}/\sqrt{T}$ in Agarwal et al. (2011) and improved to $d^{3.75}/\sqrt{T}$ in Belloni et al. (2015) (both up to log-factors). The gap here is only in the dependency on the dimension. Our results imply that for $\alpha$ above the $T^{-1/2+1/\beta}$ threshold, the gap between upper and lower bounds is much smaller. Thus, our upper bounds in this regime scale as $(d^2/\alpha)T^{-(\beta-1)/\beta}$ while the lower bound of Theorem 2.6.1 is of the order $\Omega\big((d/\alpha)T^{-(\beta-1)/\beta}\big)$; moreover for $\beta = 2$, upper and lower bounds match in the dependency on $d$.

We hope that our work will stimulate further study at the intersection of zero-order optimization and convex bandits in machine learning. An important open problem is to study novel algorithms which match our lower bound simultaneously in all main parameters. For example a class of algorithms worth exploring are those using memory of the gradient in the spirit of Nesterov accelerated method. Yet another important open problem is to study lower bounds for the regret in our setting. Finally, it would be valuable to study extensions of our work to locally strongly convex functions.

## 2.8 Proofs and additional results

In Section 2.8 we provide some auxiliary results, including those stated in Section 2.2 above. In Section 2.8 we give proofs of the results which were only stated or whose proof was only sketched in the paper. For reader's convenience all such results are restated below. Section 2.8 contains some comments on previous results in Bach and Perchet (2016). Finally, in Section 2.8 we present refined versions of Theorems 2.3.1 and 2.5.1.

## Auxiliary results

*Proof of Lemma 2.2.3.* To lighten the presentation and without loss of generality we drop the lower script "$t$" in all quantities. Using the Taylor expansion we have

$$f(x+hr\zeta) = f(x) + \langle \nabla f(x), hr\zeta\rangle + \sum_{2\le|m|\le\ell} \frac{(rh)^{|m|}}{m!} D^{(m)} f(x)\zeta^m + R(hr\zeta),$$

where by assumption $|R(hr\zeta)| \le L\|hr\zeta\|^\beta = L|r|^\beta h^\beta$. Thus,

$$\mathbb{E}[\hat{g}|x] = \frac{d}{h}\mathbb{E}\Big[\Big(\langle \nabla f(x), hr\zeta\rangle + \sum_{2\le|m|\le\ell, |m| \text{ odd}} \frac{(rh)^{|m|}}{m!} D^{(m)} f(x)\zeta^m + \frac{R(hr\zeta) - R(-hr\zeta)}{2}\Big)\zeta K(r)\Big].$$

Since $\zeta$ is uniformly distributed on the unit sphere we have $\mathbb{E}[\zeta\zeta^\top] = (1/d)I_{d\times d}$, where $I_{d\times d}$ is the identity matrix. Therefore,

$$\mathbb{E}\Big[\frac{d}{h}\langle \nabla f(x), h\zeta\rangle\zeta\Big] = \nabla f(x).$$

As $\int r^{|m|}K(r)dr = 0$ for $2 \le |m| \le \ell$ and $\int rK(r)dr = 1$ we conclude that

$$\begin{aligned}\|\mathbb{E}[\hat{g}\,|x] - \nabla f(x)\| &= \frac{d}{2h}\|\mathbb{E}\big[\big(R(hr\zeta) - R(-hr\zeta)\big)\zeta K(r)\big]\|\\ &\le \frac{d}{2h}\mathbb{E}\big[|R(hr\zeta) - R(-hr\zeta)|\,|K(r)|\big] \le \kappa_\beta L d h^{\beta-1}.\end{aligned}$$

□

*Proof of Lemma 2.2.4.* We have

$$\begin{aligned}\|\hat{g}\|^2 &= \frac{d^2}{4h^2}\big\|\big(f(x+hr\zeta) - f(x-hr\zeta) + \xi - \xi'\big)\zeta K(r)\big\|^2\\ &= \frac{d^2}{4h^2}\big(f(x+hr\zeta) - f(x-hr\zeta) + \xi - \xi'\big)^2 K^2(r).\end{aligned}$$

Using the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ we get

$$\mathbb{E}[\|\hat{g}\|^2\,|x] \le \frac{3d^2}{4h^2}\Big(\mathbb{E}\big[\big(f(x+hr\zeta) - f(x-hr\zeta)\big)^2 K^2(r)\big] + 2\kappa\sigma^2\Big). \qquad (2.12)$$

Here,

$$\begin{aligned}\big(f(x+hr\zeta)-f(x-hr\zeta)\big)^2 &= \big(f(x+hr\zeta)-f(x-hr\zeta)\pm f(x)\pm 2\langle\nabla f(x),hr\zeta\rangle\big)^2\\ &\le 3\Big\{\Big(f(x+hr\zeta){-}f(x)-\langle\nabla f(x),hr\zeta\rangle\Big)^2\\ &+\Big(f(x-hr\zeta)-f(x)-\langle\nabla f(x),-hr\zeta\rangle\Big)^2+4\langle\nabla f(x),hr\zeta\rangle^2\Big\}\\ &\le 3\left(\frac{\bar{L}^2}{2}\|hr\zeta\|^4+4\langle\nabla f(x),hr\zeta\rangle^2\right), \qquad (2.13)\end{aligned}$$

where the last inequality follows from standard properties of convex functions with Lipschitz continuous gradient, see e.g., (Bubeck, 2015, Lemma 3.4). Taking the expectation and using the fact that $\mathbb{E}[\zeta\zeta^\top]=(1/d)I_{d\times d}$ we obtain

$$\mathbb{E}[(f(x+hr\zeta)-f(x-hr\zeta))^2K^2(r)]\le 3\kappa\left(\frac{\bar{L}^2h^4}{2}+\frac{4h^2}{d}\|\nabla f(x)\|^2\right). \qquad (2.14)$$

To prove part (i) of the lemma, it is enough to combine (2.12), (2.14) and the inequality $\|\nabla f(x)\|\le\bar{L}\|x-x^*\|$ that follows from the Lipschitz gradient assumption and the fact that $\nabla f(x^*)=0$. Next, under the assumptions of part (ii) of the lemma we get analogously to (2.13) that

$$\big(f(x+hr\zeta)-f(x-hr\zeta)\big)^2\le 3\left(2L^2\|hr\zeta\|^4+4\langle\nabla f(x),hr\zeta\rangle^2\right).$$

This yields inequality (2.14) with the only difference that $\bar{L}^2/2$ is replaced by $2L^2$. Together with (2.12), it implies the result. □

**Lemma 2.8.1.** *Let $f$ be Lipschitz continuous with constant $G>0$ in a Euclidean $h_t$-neighborhood of the set $\Theta$, and let Assumption 2.2.1 (i) hold. Let $\hat{g}_t$ and $x_t$ be defined by Algorithm 1. Then*

$$\mathbb{E}[\|\hat{g}_t\|^2\,|\,x_t]\le\kappa\Big(C^*G^2d+\frac{3d^2}{2h_t^2}\sigma^2\Big),$$

*where $C^*>0$ is a numerical constant and $\kappa=\int K^2(u)du$.*

*Proof.* We have

$$\begin{aligned}\|\hat{g}\|^2&=\frac{d^2}{4h^2}\|(f(x+hr\zeta)-f(x-hr\zeta)+\xi-\xi')\zeta K(r)\|^2\\ &=\frac{d^2}{4h^2}(f(x+hr\zeta)-f(x-hr\zeta)+\xi-\xi')^2K^2(r).\end{aligned}$$

Using the inequality $(a+b+c)^2\le 3(a^2+b^2+c^2)$ we get

$$\mathbb{E}[\|\hat{g}\|^2\,|\,x]\le\frac{3d^2}{4h^2}\left(\mathbb{E}[(f(x+hr\zeta)-f(x-hr\zeta))^2K^2(r)]+2\kappa\sigma^2\right).$$

The lemma now follows by using (Shamir, 2017, Lemma 10), which shows by a concentration argument that if $x \in \Theta$, $r \in [-1, 1]$ are fixed, $\zeta$ is uniformly distributed on the unit sphere and $f$ is Lipschitz continuous with constant $G > 0$ in a Euclidean $h$-neighborhood of the set $\Theta$, then

$$\mathbb{E}[(f(x + hr\zeta) - f(x - hr\zeta))^2] \le c\frac{(hr)^2G^2}{d},$$

where $c > 0$ is a numerical constant. □

**Lemma 2.8.2.** *Let $f(\cdot)$ be a convex function on $\mathbb{R}^d$ and $h_t > 0$. Then the following holds.*

*(i) Function $\hat{f}_t(\cdot)$ is convex on $\mathbb{R}^d$.*

*(ii) $\hat{f}_t(x) \ge f(x)$ for all $x \in \mathbb{R}^d$.*

*(iii) Function $\hat{f}_t(\cdot)$ is differentiable on $\mathbb{R}^d$ and for the conditional expectation given $x_t$ we have*

$$\mathbb{E}[\hat{g}_t | x_t] = \nabla \hat{f}_t(x_t).$$

*Proof.* Item (i) is straightforward. To prove item (ii), consider $g_t \in \partial f(x)$. Then,

$$\hat{f}_t(x) \ge \mathbb{E}\big[f(x) + h_t\langle g_t, \tilde{\zeta}\rangle\big] = f(x) + h_t\langle g_t, \mathbb{E}[\tilde{\zeta}]\rangle = f(x).$$

For item (iii) we refer to (Nemirovsky and Yudin, 1983, pg. 350), or Flaxman et al. (2005). It is based on the fact that for any $x \in \mathbb{R}^d$ using Stokes formula we have

$$\begin{aligned}\nabla \hat{f}_t(x) &= \frac{1}{V(B_d)h_t^d}\int_{\|v\|=h_t} f(x+v)\frac{v}{\|v\|}\mathrm{d}s_{h_t}(v) = \frac{d}{V(S_d)h_t}\int_{\|u\|=1} f(x+h_tu)u\,\mathrm{d}s_1(u)\\ &= \frac{d}{V(S_d)h_t}\int_{\|u\|=1} f(x+h_tu)u\,\mathrm{d}s_1(u) = \mathbb{E}\Big[\frac{d}{h_t}f(x+h_t\zeta_t)\zeta_t\Big]\end{aligned}$$

where $V(B_d)$ is the volume of the unit ball $B_d$, $\mathrm{d}s_r(\cdot)$ is the element of spherical surface of raduis $r$ in $\mathbb{R}^d$, and $V(S_d) = dV(B_d)$ is the surface area of the unit sphere in $\mathbb{R}^d$. Since $f(x + h_t\zeta_t)\zeta_t$ has the same distribution as $f(x - h_t\zeta_t)(-\zeta_t)$ we also get

$$\mathbb{E}\Big[\frac{d\big(f(x+h_t\zeta_t) - f(x-h_t\zeta_t)\big)\zeta_t}{2h_t}\Big] = \nabla \hat{f}_t(x).$$

□

**Lemma 2.8.3.** *If $f$ is $\alpha$-strongly convex then $\hat{f}_t$ is $\alpha$-strongly convex. If $f \in \mathcal{F}_2(L)$ then for any $x \in \mathbb{R}^d$ and $h_t > 0$ we have*

$$|\hat{f}_t(x) - f(x)| \le Lh_t^2. \tag{2.15}$$

*and*

$$|\mathbb{E}f(x \pm h_t\zeta_t) - f(x)| \le Lh_t^2. \tag{2.16}$$

*Proof.* Using the fact that $\mathbb{E}[\tilde{\zeta}] = 0$ we have

$$|\mathbb{E}\big[f(x + h_t\tilde{\zeta}) - f(x)\big]| = |\mathbb{E}\big[f(x + h_t\tilde{\zeta}) - f(x) - \langle \nabla f(x), h_t\tilde{\zeta}\rangle\big]| \leq Lh_t^2\mathbb{E}[\|\tilde{\zeta}\|^2] \leq Lh_t^2.$$

Thus, (2.15) follows. The proof of (2.16) is analogous. The $\alpha$-strong convexity of $\hat{f}_t$ is equivalent to the relation

$$\langle \nabla \hat{f}_t(x) - \nabla \hat{f}_t(x'), x - x'\rangle \geq \alpha \left\|x - x'\right\|^2, \quad \forall x, x' \in \mathbb{R}^d,$$

which is proved as follows:

$$\begin{aligned}\langle \nabla \hat{f}_t(x) - \nabla \hat{f}_t(x'), x - x'\rangle &= \langle \mathbb{E}\big[\nabla f(x + h_t\tilde{\zeta}) - \nabla f(x' + h_t\tilde{\zeta})\big], x - x'\rangle \\ &= \mathbb{E}\big[\langle \nabla f(x + h_t\tilde{\zeta}) - \nabla f(x' + h_t\tilde{\zeta}), (x + h_t\tilde{\zeta}) - (x' + h_t\tilde{\zeta})\rangle\big] \\ &\geq \alpha \left\|x - x'\right\|^2, \quad \forall x, x' \in \mathbb{R}^d,\end{aligned}$$

due to the $\alpha$-strong convexity of $f$. □

## Proofs

*Proof of Theorem 2.3.1.* Fix an arbitrary $x \in \Theta$. By the definition of the algorithm, we have $\|x_{t+1} - x\|^2 \leq \|x_t - \eta_t\hat{g}_t - x\|^2$, which is equivalent to

$$\langle \hat{g}_t, x_t - x\rangle \leq \frac{\|x_t - x\|^2 - \|x_{t+1} - x]\|^2}{2\eta_t} + \frac{\eta_t}{2}\|\hat{g}_t\|^2. \tag{2.17}$$

By the strong convexity assumption we have

$$f(x_t) - f(x) \leq \langle \nabla f(x_t), x_t - x\rangle - \frac{\alpha}{2}\|x_t - x\|^2.$$

Combining the last two displays and setting $a_t = \|x_t - x\|^2$ we obtain

$$\begin{aligned}\mathbb{E}[f(x_t) - f(x)\,|\,x_t] &\leq \|\mathbb{E}[\hat{g}_t\,|\,x_t] - \nabla f(x_t)\|\|x_t - x\| + \frac{1}{2\eta_t}\mathbb{E}[a_t - a_{t+1}\,|\,x_t] \\ &+ \frac{\eta_t}{2}\mathbb{E}[\|\hat{g}_t\|^2\,|\,x_t] - \frac{\alpha}{2}\mathbb{E}[a_t\,|\,x_t] \\ &\leq \kappa_\beta L d h_t^{\beta-1}\|x_t - x\| + \frac{1}{2\eta_t}\mathbb{E}[a_t - a_{t+1}\,|\,x_t] \\ &+ \frac{\eta_t}{2}\mathbb{E}[\|\hat{g}_t\|^2\,|\,x_t] - \frac{\alpha}{2}\mathbb{E}[a_t\,|\,x_t],\end{aligned} \tag{2.18}$$

where the second inequality follows from Lemma 2.2.3. As $2ab \leq a^2 + b^2$ we have

$$dh_t^{\beta-1}\|x_t - x\| \leq \frac{1}{2}\Big(\frac{2\kappa_\beta L}{\alpha}d^2h_t^{2(\beta-1)} + \frac{\alpha}{2\kappa_\beta L}\|x_t - x\|^2\Big). \tag{2.19}$$

We conclude, taking the expectations and letting $r_t = \mathbb{E}[a_t]$, that

$$\mathbb{E}[f(x_t) - f(x)] \le \frac{r_t - r_{t+1}}{2\eta_t} - \frac{\alpha}{4} r_t + (\kappa_\beta L)^2 \frac{d^2}{\alpha} h_t^{2(\beta-1)} + \frac{\eta_t}{2}\mathbb{E}[\|\hat{g}_t\|^2]$$

Summing both sides over $t$ gives

$$\sum_{t=1}^{T} \mathbb{E}[f(x_t) - f(x)] \le \frac{1}{2}\sum_{t=1}^{T}\left(\frac{r_t - r_{t+1}}{\eta_t} - \frac{\alpha}{2} r_t\right) + \sum_{t=1}^{T}\left((\kappa_\beta L)^2 \frac{d^2}{\alpha} h_t^{2(\beta-1)} + \frac{\eta_t}{2}\mathbb{E}[\|\hat{g}_t\|^2]\right).$$

The first sum on the r.h.s. is smaller than 0 for our choice of $\eta_t = \frac{2}{\alpha t}$. Indeed,

$$\sum_{t=1}^{T}\left(\frac{r_t - r_{t+1}}{\eta_t} - \frac{\alpha}{2} r_t\right) \le r_1\Big(\frac{1}{\eta_1} - \frac{\alpha}{2}\Big) + \sum_{t=2}^{T} r_t\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \frac{\alpha}{2}\right) = 0.$$

From this remark and Lemma 2.2.4(ii) (where we use that Assumption 2.2.2 implies $f \in \mathcal{F}_2(\bar{L}/2)$) we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[f(x_t) - f(x)] &\le \frac{1}{\alpha}\sum_{t=1}^{T}\left((\kappa_\beta L)^2 d^2 h_t^{2(\beta-1)} + \frac{1}{t}\mathbb{E}[\|\hat{g}_t\|^2]\right) \\
&\le \frac{1}{\alpha}\sum_{t=1}^{T}\left((\kappa_\beta L)^2 d^2 h_t^{2(\beta-1)} + \frac{1}{t}\Big[9\kappa\Big(G^2 d + \frac{\bar{L}^2 d^2 h_t^2}{8}\Big) + \frac{3\kappa d^2 \sigma^2}{2h_t^2}\Big]\right) \\
&\le \frac{d^2}{\alpha}\sum_{t=1}^{T}\left[\left\{(\kappa_\beta L)^2 h_t^{2(\beta-1)} + \frac{3}{2}\frac{\kappa\sigma^2}{h_t^2 t}\right\} + \frac{9\kappa \bar{L}^2 h_t^2}{8t}\right] + \frac{9\kappa G^2}{\alpha} d(\log T + 1). \qquad (2.20)
\end{aligned}$$

If $\sigma > 0$ then our choice of $h_t = \left(\frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2}\right)^{\frac{1}{2\beta}} t^{-\frac{1}{2\beta}}$ is the minimizer of the main term (in curly brackets in (2.20)). Plugging this $h_t$ in (2.20) and using the fact that $\sum_{t=1}^{T} t^{-1+1/\beta} \le \beta T^{1/\beta}$ for $\beta \ge 2$ we get (2.2). Inequality (2.3) follows from (2.2) in view of the convexity of $f$. If $\sigma = 0$ the stochastic variability term in (2.20) disappears and one can choose $h_t$ as small as desired, in particular, such that the sum in (2.20) is smaller than $\frac{\kappa G^2}{\alpha} d \log T$. This yields the bounds for $\sigma = 0$. □

*Proof of Theorem 2.4.1.* We have

$$\begin{aligned}
\mathbb{E}|\hat{M} - f(x^*)| &\le \mathbb{E}\Big|\frac{1}{T}\sum_{t=1}^{T}\xi_t''\Big| + \mathbb{E}\Big|\frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f(x^*))\Big| \\
&= \mathbb{E}\Big|\frac{1}{T}\sum_{t=1}^{T}\xi_t''\Big| + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f(x_t) - f(x^*)] \\
&\le \frac{\sigma}{T^{\frac{1}{2}}} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f(x_t) - f(x^*)]
\end{aligned}$$

and the theorem follows by using (2.2). □

*Proof of Theorem 2.3.2.* We start as in the proof of Theorem 2.3.1 to get (2.18). Then, using the strong convexity of $f$ and the fact that $x^*$ is the minimizer of $f$ we get analogously to (2.19) that

$$dh_t^{\beta-1}\|x_t-x^*\| \leq \frac{1}{2}\Big(\frac{2\kappa_\beta L}{\alpha}d^2h_t^{2(\beta-1)}+\frac{\alpha}{2\kappa_\beta L}\|x_t-x^*\|^2\Big) \leq \frac{\kappa_\beta L}{\alpha}d^2h_t^{2(\beta-1)}+\frac{f(x_t)-f(x^*)}{2\kappa_\beta L}.$$

Combining the last display and (2.18), using Lemma 2.2.4 and letting $r_t = \mathbb{E}[\|x_t-x^*\|^2]$ we get

$$\mathbb{E}[f(x_t)-f(x^*)] \leq \frac{r_t-r_{t+1}}{\eta_t}-\alpha r_t+2(\kappa_\beta L)^2\frac{d^2}{\alpha}h_t^{2(\beta-1)}+\kappa\eta_t\left[9\bar{L}^2\left(dr_t+\frac{d^2h_t^2}{8}\right)+\frac{3d^2\sigma^2}{2h_t^2}\right]. \tag{2.21}$$

For $t=1,\dots,T_0$, since $h_t = T^{-\frac{1}{2\beta}}$ and $\eta_t = (\alpha T)^{-1}$ we have the following consequence of (2.21)

$$r_{t+1} \leq r_t\left(1-\frac{1}{T}+\frac{9\kappa\bar{L}^2}{(\alpha T)^2}d\right)+b_T \leq r_t\left(1+\frac{9\kappa\bar{L}^2}{(\alpha T)^2}d\right)+b_T \tag{2.22}$$

where

$$\begin{aligned} b_T &= \frac{d^2}{\alpha^2 T}\left(2(\kappa_\beta L)^2T^{-\frac{\beta-1}{\beta}}+\frac{9}{8}\kappa\bar{L}^2T^{-\frac{\beta+1}{\beta}}+\frac{3}{2}\kappa\sigma^2T^{-\frac{\beta-1}{\beta}}\right) \leq \\ &\leq \frac{d^2}{\alpha^2 T}\left(2(\kappa_\beta L)^2+\frac{9}{8}\kappa\bar{L}^2+\frac{3}{2}\kappa\sigma^2\right)T^{-\frac{\beta-1}{\beta}}. \end{aligned} \tag{2.23}$$

Letting $C_3 = 9\kappa\bar{L}^2$, inequality (2.22) is of the form $r_{t+1} \leq r_t q + b_T$, with $q = (1+\frac{C_3 d}{(\alpha T)^2})$. Then

$$r_{T_0+1} \leq r_1 q^{T_0} + b_T\sum_{j=1}^{T_0-1}q^j \leq r_1 q^{T_0}+b_T\frac{q^{T_0}}{q-1} \leq \left(r_1+\frac{(\alpha T)^2}{C_3 d}b_T\right)q^{T_0}.$$

Now, assuming

$$T_0 = \left\lfloor\frac{4C_3 d}{\alpha^2}\right\rfloor \tag{2.24}$$

we obtain

$$\begin{aligned} q^{T_0} &= \exp\left[T_0\log\left(1+\frac{C_3 d}{(\alpha T)^2}\right)\right] \\ &\leq \exp\left[\frac{4C_3 d}{\alpha^2}\log\left(1+\frac{C_3 d}{(\alpha T)^2}\right)\right] \\ &\leq \exp\left(\frac{4C_3^2 d^2}{\alpha^4 T^2}\right) \leq \exp\left(\frac{4C_3^2}{C_*^2}\right) =: C_4 \end{aligned}$$

where in the last inequality we have used the assumption that, for $C_* > 0$ large enough,

$$\alpha > \sqrt{\frac{C_* d}{T}}. \tag{2.25}$$

As we shall see, this also guarantees that $T_0 < T$. In conclusion, we obtain

$$\begin{aligned} r_{T_0+1} &\leq C_4\left(r_1 + \frac{(\alpha T)^2}{C_3 d} b_T\right) \\ &\leq C_4\left(r_1 + \frac{(\alpha T)^2}{C_3 d}\frac{d^2}{\alpha^2 T}\left(2(\kappa_\beta L)^2 + \frac{9}{8}\kappa\bar{L}^2 + \frac{3}{2}\kappa\sigma^2\right)T^{-\frac{\beta-1}{\beta}}\right) \\ &= C_4\left(r_1 + \frac{d}{C_3}\left(2(\kappa_\beta L)^2 + \frac{9}{8}\kappa\bar{L}^2 + \frac{3}{2}\kappa\sigma^2\right)T^{\frac{1}{\beta}}\right). \end{aligned} \tag{2.26}$$

We now go back to inequality (2.21). Recalling the definition of $\bar{x}_{T_0,T}$ and the fact that $h_t = t^{-\frac{1}{2\beta}}$ and $\eta_t = \frac{2}{\alpha t}$ for $t \in \{T_0+1, \ldots T\}$, we deduce from (2.21) that

$$\begin{aligned} (T-T_0)\mathbb{E}[f(\bar{x}_{T_0,T}) - f(x^*)] \quad \leq \quad & \sum_{t=T_0+1}^{T} (r_t - r_{t+1})\frac{\alpha t}{2} - \alpha r_t + 18\kappa\frac{\bar{L}^2}{\alpha t} d r_t \\ & + \frac{d^2}{\alpha}\sum_{t=T_0+1}^{T}\left(2(\kappa_\beta L)^2 t^{-\frac{\beta-1}{\beta}} + \frac{9}{4}\kappa\bar{L}^2 t^{-\frac{\beta+1}{\beta}} + 3\kappa\sigma^2 t^{-\frac{\beta-1}{\beta}}\right). \end{aligned}$$

Since $9\kappa\bar{L}^2 = C_3$ condition (2.24) implies that $\frac{18\kappa\bar{L}^2}{\alpha t}d \leq \frac{\alpha}{2}$ for $t \geq T_0 + 1$. Thus

$$(T-T_0)\mathbb{E}[f(\bar{x}_{T_0,T}) - f(x^*)] \leq \frac{\alpha}{2}\sum_{t=T_0+1}^{T}\Big[(r_t - r_{t+1})t - r_t\Big] + U_T,$$

where

$$U_T = \frac{d^2}{\alpha}\left(2(\kappa_\beta L)^2 + \frac{9}{4}\kappa\bar{L}^2 + 3\kappa\sigma^2\right)\sum_{t=T_0}^{T} t^{-\frac{\beta-1}{\beta}} \leq \frac{d^2}{\alpha}\left(2(\kappa_\beta L)^2 + \frac{9}{4}\kappa\bar{L}^2 + 3\kappa\sigma^2\right)\beta T^{\frac{1}{\beta}}.$$

On the other hand

$$\sum_{t=T_0+1}^{T}\Big[(r_t - r_{t+1})t - r_t\Big] \leq r_{T_0+1}(T_0 + 1 - 1) + \sum_{t=T_0+2}^{T} r_t(t - (t-1) - 1) = T_0 r_{T_0+1}.$$

Using inequality (2.26) and condition (2.24) we get

$$\begin{aligned} \frac{\alpha T_0}{2} r_{T_0+1} \quad &\leq \quad \frac{2C_3C_4 d}{\alpha}\left(r_1 + \frac{d}{C_3}\left(2(\kappa_\beta L)^2 + \frac{9}{8}\kappa\bar{L}^2 + \frac{3}{2}\kappa\sigma^2\right)T^{\frac{1}{\beta}}\right) \\ &= \quad 2C_4\left(9\kappa\bar{L}^2\frac{d}{\alpha}r_1 + \frac{d^2}{\alpha}\left(2(\kappa_\beta L)^2 + \frac{9}{8}\kappa\bar{L}^2 + \frac{3}{2}\kappa\sigma^2\right)T^{\frac{1}{\beta}}\right). \end{aligned}$$

These bounds imply

$$(T-T_0)\mathbb{E}[f(\bar{x}_{T_0,T}) - f(x^*)] \leq 18C_4\kappa\bar{L}^2\frac{d}{\alpha}r_1 + (2C_4 + \beta)\frac{d^2}{\alpha}\left(2(\kappa_\beta L)^2 + \frac{9}{4}\kappa\bar{L}^2 + 3\kappa\sigma^2\right)T^{\frac{1}{\beta}}.$$

Since $C_* > 8C_3 = 72\kappa\bar{L}^2$ it follows from (2.24) and (2.25) that $T \geq 2T_0$. Thus

$$\mathbb{E}[f(\bar{x}_{T_0,T}) - f(x^*)] \leq 36C_4\kappa\bar{L}^2\frac{d}{\alpha T}r_1 + \Big(4C_4 + 2\beta\Big)\frac{d^2}{\alpha}\left(2(\kappa_\beta L)^2 + \frac{9}{4}\kappa\bar{L}^2 + 3\kappa\sigma^2\right)T^{-\frac{\beta-1}{\beta}}.$$

□

*Proof of Theorem 2.5.1.* Fix $x \in \Theta$. Due to the $\alpha$-strong convexity of $\hat{f}_t$ (cf. Lemma 2.8.3) we have

$$\hat{f}_t(x_t) - \hat{f}_t(x) \leq \langle\nabla\hat{f}_t(x_t), x_t - x\rangle - \frac{\alpha}{2}\|x_t - x\|^2.$$

Using (2.15) and Lemma 2.8.2(ii) we obtain

$$f(x_t) - f(x) \leq Lh_t^2 + \langle\nabla\hat{f}_t(x_t), x_t - x\rangle - \frac{\alpha}{2}\|x_t - x\|^2.$$

Using this property and exploiting inequality (2.17) we find, with an argument similar to the proof of Theorem 2.3.1, that

$$\forall x \in \Theta: \qquad \mathbb{E}\big[f(x_t) - f(x)\big] \leq Lh_t^2 + \frac{r_t - r_{t+1}}{2\eta_t} - \frac{\alpha}{2}r_t + \frac{\eta_t}{2}\mathbb{E}[\|\hat{g}_t\|^2].$$

By assumption, $\eta_t = \frac{1}{\alpha t}$. Summing up from $t = 1$ to $T$ and reasoning again analogously to the proof of Theorem 2.3.1 we obtain

$$\forall x \in \Theta: \qquad \mathbb{E}\sum_{t=1}^{T}\big(f(x_t) - f(x)\big) \leq \sum_{t=1}^{T}\Big(Lh_t^2 + \frac{1}{2\alpha t}\mathbb{E}[\|\hat{g}_t\|^2]\Big).$$

Now, inspection of the proof of Lemma 2.2.4 shows that it remains valid with $\kappa = 1$ when $K(\cdot) \equiv 1$ in Algorithm 1. This yields

$$\mathbb{E}[\|\hat{g}_t\|^2] \leq 9\left(G^2d + \frac{L^2d^2h_t^2}{2}\right) + \frac{3d^2\sigma^2}{2h_t^2}.$$

Thus,

$$\forall x \in \Theta: \qquad \mathbb{E}\sum_{t=1}^{T}\big(f(x_t) - f(x)\big) \leq \sum_{t=1}^{T}\Big[\Big(L + \frac{9L^2d^2}{4\alpha t}\Big)h_t^2 + \frac{3d^2\sigma^2}{4h_t^2\alpha t} + \frac{9G^2d}{2\alpha t}\Big].$$

The chosen value $h_t = \left(\frac{3d^2\sigma^2}{4L\alpha t+9L^2d^2}\right)^{1/4}$ minimizes the r.h.s. and yields

$$\begin{aligned}\forall x \in \Theta : \quad \mathbb{E}\sum_{t=1}^{T}\big(f(x_t)-f(x)\big) &\leq \frac{3}{2}\sum_{t=1}^{T}\frac{d^2\sigma^2}{\alpha t}\left(\frac{4L\alpha t+9L^2d^2}{3d^2\sigma^2}\right)^{1/2}+\frac{9G^2}{2}\frac{d}{\alpha}(1+\log T)\\ &\leq \sum_{t=1}^{T}\sqrt{3}\Big[\frac{d\sigma\sqrt{L}}{\sqrt{\alpha t}}+\frac{3Ld^2\sigma}{2\alpha t}\Big]+9G^2\frac{d}{\alpha}(1+\log T)\\ &\leq 2\sqrt{3L}\sigma\frac{d}{\sqrt{\alpha}}\sqrt{T}+\Big(\frac{3\sqrt{3}}{2}\sigma L+\frac{9G^2}{d}\Big)\frac{d^2}{\alpha}(1+\log T).\end{aligned}$$

As $1+\log T \leq ((\log 2)^{-1}+1)\log T$ for any $T \geq 2$, we obtain (2.7). On the other hand, we have the straightforward bound

$$\forall x \in \Theta : \qquad \mathbb{E}\sum_{t=1}^{T}\big(f(x_t)-f(x)\big) \leq GBT. \tag{2.27}$$

The remaining part of the proof follows the same lines as in Theorem 2.3.1. □

*Proof of Theorem 2.6.1.* We use the fact that $\sup_{f\in\mathcal{F}'_{\alpha,\beta}}$ is bigger than the maximum over a finite family of functions in $\mathcal{F}'_{\alpha,\beta}$. We choose this finite family in a way that its members cannot be distinguished from each other with positive probability but are separated enough from each other to guarantee that the maximal optimization error for this family is of the order of the desired lower bound.

We first assume that $\alpha \geq T^{-1/2+1/\beta}$.

Let $\eta_0 : \mathbb{R} \to \mathbb{R}$ be an infinitely many times differentiable function such that

$$\eta_0(x) = \begin{cases} = 1 & \text{if } |x| \leq 1/4,\\ \in (0,1) & \text{if } 1/4 < |x| < 1,\\ = 0 & \text{if } |x| \geq 1.\end{cases}$$

Set $\eta(x) = \int_{-\infty}^{x}\eta_0(\tau)d\tau$. Let $\Omega = \big\{-1,1\big\}^d$ be the set of binary sequences of length $d$. Consider the finite set of functions $f_\omega : \mathbb{R}^d \to \mathbb{R}, \omega \in \Omega$, defined as follows:

$$f_\omega(u) = \alpha(1+\delta)\left\|u\right\|^2/2+\sum_{i=1}^{d}\omega_i r h^\beta \eta(u_i h^{-1}), \qquad u=(u_1,\dots,u_d),$$

where $\omega_i \in \{-1,1\}$, $h = \min\big((\alpha^2/d)^{\frac{1}{2(\beta-1)}}, T^{-\frac{1}{2\beta}}\big)$ and $r > 0, \delta > 0$ are fixed numbers that will be chosen small enough.

Let us prove that $f_\omega \in \mathcal{F}'_{\alpha,\beta}$ for $r > 0$ and $\delta > 0$ small enough. It is straightforward to check that if $r$ is small enough the functions $f_\omega$ are $\alpha$-strongly convex and belong to $\mathcal{F}_\beta(L)$.

Next, the components of the gradient $\nabla f_\omega$ have the form

$$(\nabla f_\omega(u))_i = \alpha(1+\delta)u_i + \omega_i r h^{\beta-1}\eta_0(u_i h^{-1}).$$

Thus,

$$\|\nabla f_\omega(u)\|^2 \leq 2\alpha^2(1+\delta)^2 \|u\|^2 + 2r^2\alpha^2$$

and the last expression can be rendered smaller than $G^2$ uniformly in $u \in \Theta$ by the choice of $\delta$ and $r$ small enough since $G^2 > 4\alpha^2$.

Finally, we check that the minimizers of functions $f_\omega$ belong to $\Theta$. Notice that we can choose $r$ small enough to have $\alpha^{-1}(1+\delta)^{-1}rh^{\beta-2} < 1/4$ and that under this condition the equation $\nabla f_\omega(x) = 0$ has the solution

$$x^*_\omega = (x^*(\omega_1), \ldots, x^*(\omega_d)),$$

where $x^*(\omega_i) = -\omega_i\alpha^{-1}(1+\delta)^{-1}rh^{\beta-1}$. Using the definition of $h$ we obtain

$$\|x^*_\omega\| \leq d^{1/2}\alpha^{-1}(1+\delta)^{-1}rh^{\beta-1} \leq d^{1/2}\alpha^{-1}(1+\delta)^{-1}r(\alpha^2/d)^{1/2} \leq (1+\delta)^{-1}r < 1$$

for $r > 0$ small enough, which means that $x^*_\omega$ belongs to the interior of $\Theta$.

Combining all the above remarks we conclude that the family of functions $\{f_\omega, \omega \in \Omega\}$ is a subset of $\mathcal{F}'_{\alpha,\beta}$ for $r > 0$ and $\delta > 0$ small enough.

Set for brevity $(z_i, y_i)_{i=1}^t = (z_1, y_1, \ldots, z_t, y_t)$, $(\boldsymbol{\zeta}_i)_{i=1}^t = (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_t)$. For any fixed $\omega \in \Omega$, we denote by $\mathbf{P}_{\omega,T}$ the probability measure corresponding to the joint distribution of $((z_i, y_i)_{i=1}^T, (\boldsymbol{\zeta}_i)_{i=1}^T)$ where $y_t = f_\omega(z_t) + \xi_t$ with independent identically distributed $\xi_t$'s such that (2.9) holds, $\xi_t$ is independent of $(z_1, y_1, \ldots, z_{t-1}, y_{t-1}, \boldsymbol{\zeta}_t)$ for each $t$, and $z_t$'s chosen by a sequential strategy in $\Pi_T$. We have

$$d\mathbf{P}_{\omega,T}((z_i, y_i)_{i=1}^T, (\boldsymbol{\zeta}_i)_{i=1}^T) = dF\big(y_1 - f_\omega(z_1)\big)\prod_{t=2}^T dF\Big(y_t - f_\omega\big(\Phi_t((z_i, y_i)_{i=1}^{t-1}, \boldsymbol{\zeta}_t)\big)\Big)d\mathbb{P}_t(\boldsymbol{\zeta}_t),$$

where $\mathbb{P}_t$ is the probability measure corresponding to the distribution of $\boldsymbol{\zeta}_t$. Without loss of generality, we omit here the dependence of $\Phi_i$ on $z_2, \ldots, z_{i-1}$ since $z_i, i \geq 2$, is a Borel function of $z_1, y_1, \ldots, y_{i-1}$. Let $\mathbf{E}_{\omega,T}$ denote the expectation w.r.t. $\mathbf{P}_{\omega,T}$.

Consider the statistic

$$\hat{\omega} \in \arg\min_{\omega\in\Omega} \|z_T - x^*_\omega\|.$$

Since $\|x^*_{\hat{\omega}} - x^*_\omega\| \leq \|z_T - x^*_\omega\| + \|z_T - x^*_{\hat{\omega}}\| \leq 2\|z_T - x^*_\omega\|$ for all $\omega \in \Omega$ we obtain

$$\begin{aligned}\mathbf{E}_{\omega,T}\big[\,\|z_T - x^*_\omega\|^2\,\big] &\geq \frac{1}{4}\mathbf{E}_{\omega,T}\big[\,\|x^*_\omega - x^*_{\hat{\omega}}\|^2\,\big] \\ &= \alpha^{-2}r^2h^{2\beta-2}\mathbf{E}_{\omega,T}\rho(\hat{\omega}, \omega),\end{aligned}$$

where $\rho(\hat{\omega},\omega) = \sum_{i=1}^{d} \mathbb{I}(\hat{\omega}_i \neq \omega_i)$ is the Hamming distance between $\hat{\omega}$ and $\omega$. Taking the maximum over $\Omega$ and then the minimum over all statistics $\hat{\omega}$ with values in $\Omega$ we obtain

$$\max_{\omega\in\Omega} \mathbf{E}_{\omega,T}\big[\, \|z_T - x^*_\omega\|^2 \,\big] \geq \alpha^{-2} r^2 h^{2\beta-2} \inf_{\hat{\omega}} \max_{\omega\in\Omega} \mathbf{E}_\omega \rho(\hat{\omega},\omega).$$

By (Tsybakov, 2009, Theorem 2.12), if for some $\gamma > 0$ and all $\omega, \omega' \in \Omega$ such that $\rho(\omega,\omega') = 1$ we have $KL(\mathbf{P}_{\omega,T}, \mathbf{P}_{\omega',T}) \leq \gamma$, where $KL(\cdot,\cdot)$ denotes the Kullback-Leibler divergence, then

$$\inf_{\hat{\omega}} \max_{\omega\in\Omega} \mathbf{E}_{\omega,T} \rho(\hat{\omega},\omega) \geq \frac{d}{4}\exp(-\gamma).$$

Now for all $\omega, \omega' \in \Omega$ such that $\rho(\omega,\omega') = 1$ we have

$$\begin{aligned}
KL(\mathbf{P}_{\omega,T}, \mathbf{P}_{\omega',T}) &= \int \log\Big(\frac{d\mathbf{P}_{\omega,T}}{d\mathbf{P}_{\omega',T}}\Big) d\mathbf{P}_{\omega,T} \\
&= \int \Bigg[ \log\Big(\frac{dF(y_1 - f_\omega(z_1))}{dF(y_1 - f_{\omega'}(z_1))}\Big) + \\
&\qquad + \sum_{t=2}^{T} \log\left(\frac{dF(y_t - f_\omega(\Phi_t((z_i,y_i)_{i=1}^{t-1}, \boldsymbol{\zeta}_t)))}{dF(y_t - f_{\omega'}(\Phi_t((z_i,y_i)_{i=1}^{t-1}, \boldsymbol{\zeta}_t)))}\right)\Bigg] \\
&\qquad\qquad dF\big(y_1 - f_\omega(z_1)\big) \prod_{t=2}^{T} dF\Big(y_t - f_\omega\big(\Phi_t((z_i,y_i)_{i=1}^{t-1}, \boldsymbol{\zeta}_t)\big)\Big) d\mathbb{P}_t(\boldsymbol{\zeta}_t) \\
&\leq T I_0 \max_{u\in\mathbb{R}} |f_\omega(u) - f_{\omega'}(u)|^2 = I_0 r^2 \eta^2(1),
\end{aligned}$$

where the last inequality is granted if $r < v_0/\eta(1)$ due to (2.9). Assuming in addition that $r$ satisfies $r^2 \leq (\log 2)/\big(I_0 \eta^2(1)\big)$ we obtain $KL(\mathbf{P}_{\omega,T}, \mathbf{P}_{\omega',T}) \leq \log 2$. Therefore, we have proved that if $\alpha \geq T^{-1/2+1/\beta}$ then there exist $r > 0$ and $\delta > 0$ small enough such that

$$\max_{\omega\in\Omega} \mathbf{E}_{\omega,T}\big[\, \|z_T - x^*_\omega\|^2 \,\big] \geq \frac{1}{8} d\alpha^{-2} r^2 h^{2\beta-2} = \frac{r^2}{8} \min\Big(1, \frac{d}{\alpha^2} T^{-\frac{\beta-1}{\beta}}\Big). \tag{2.28}$$

This implies (2.11) for $\alpha \geq T^{-1/2+1/\beta}$. In particular, if $\alpha = \alpha_0 := T^{-1/2+1/\beta}$ the bound (2.28) is of the order $\min\Big(1, dT^{-\frac{1}{\beta}}\Big)$. Then for $0 < \alpha < \alpha_0$ we also have the bound of this order since the classes $\mathcal{F}'_{\alpha,\beta}$ are nested: $\mathcal{F}'_{\alpha_0,\beta} \subset \mathcal{F}'_{\alpha,\beta}$. This completes the proof of (2.11).

We now prove (2.10). From (2.28) and $\alpha$-strong convexity of $f$ we get that, for $\alpha \geq T^{-1/2+1/\beta}$,

$$\max_{\omega\in\Omega} \mathbf{E}_{\omega,T}\big[f(z_T) - f(x^*_\omega)\big] \geq \frac{r^2}{16} \min\Big(\alpha, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\Big).$$

This implies (2.10) in the zone $\alpha \geq T^{-1/2+1/\beta}$ since for such $\alpha$ we have

$$\min\Big(\alpha, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\Big) = \min\Big(\max(\alpha, T^{-1/2+1/\beta}), \frac{d}{\sqrt{T}}, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\Big).$$

On the other hand,

$$\min\left(\alpha_0, \frac{d}{\alpha_0}T^{-\frac{\beta-1}{\beta}}\right) = \min\left(T^{-1/2+1/\beta}, \frac{d}{\sqrt{T}}\right),$$

and the same lower bound holds for $0 < \alpha < \alpha_0$ by the nestedness argument that we used to prove (2.11) in the zone $0 < \alpha < \alpha_0$. Thus, (2.10) follows.

□

### Comments on Bach and Perchet (2016)

In this section we comment on issues with some claims in the paper of Bach and Perchet Bach and Perchet (2016), which presents a number of valuable results and provides a motivation for our work. We wish to clarify such issues for the sake of understanding, as otherwise a comparison to the results presented here would be misleading.

Bach and Perchet Bach and Perchet (2016) introduce Algorithm 1 in the current form and provide upper bounds for its optimisation error and online regret when $f \in \mathcal{F}_\beta(L)$ with integer $\beta$. The setting where $f$ is strongly convex is considered in Propositions 4,6-8 and 9 of that paper. Propositions 4, 6,9 give the rates decaying in $T$ not faster than $T^{-\frac{\beta-1}{\beta+1}}$, which is slower than the optimal rate $T^{-\frac{\beta-1}{\beta}}$. Proposition 8 dealing with asymptotic results is problematic. It is stated as bounds on $\|x_N - x^*\|$ but the authors presumably mean bounds on $\mathbb{E}\,\|x_N - x^*\|^2$. A dependence of the bound on the initial value of the algorithm is missing in the part of Proposition 8 entitled "unconstrained optimization of strongly convex mappings". This remark also concerns Proposition 7.

### Additional results

In this section, we provide refined versions of Theorems 2.3.1 and 2.5.1. First we state a non-asymptotic version of Chung's lemma (Chung, 1954, Lemma 1). It allows us to obtain in Theorem 2.8.5 upper bounds for $\mathbb{E}\{\|x_t - x^*\|^2\}$, where $x_t$ is generated by a constrained version of Algorithm 1 (i.e., with compact $\Theta$) under the assumptions of Theorems 2.3.1 and 2.5.1. By using this result and considering averaging from $\lfloor T/2 \rfloor + 1$ to $T$ rather than from 1 to $T$, in Theorems 2.8.6 and 2.8.7 we provide finer upper bounds for the optimization error than in Theorems 2.3.1 and 2.5.1. The refinement consists in the fact that we get rid of the logarithmic factors appearing in (2.3) and (2.8). Finally, in Theorem 2.8.8 we show that the term $\frac{d^2}{\alpha}\log T$ in the bound on the cumulative regret in Theorem 2.5.1 can be improved to $\frac{d}{\alpha}\log T$ under a slightly more restrictive assumption (we assume that the norm $\|\nabla f\|$ is uniformly bounded by $G$ on a large enough Euclidean neighborhood of $\Theta$ rather than only on $\Theta$).

**Lemma 2.8.4.** *Let $\{b_t\}$ be a sequence of real numbers such that for all integers $t \geq 2$,*

$$b_{t+1} < \left(1 - \frac{1}{t}\right) b_t + \sum_{i=1}^{N} \frac{a_i}{t^{p_i+1}}, \tag{2.29}$$

*where* $0 < p_i < 1$ *and* $a_i \geq 0$ *for* $1 \leq i \leq N$*. Then for* $t \geq 2$ *we have*

$$b_t < \frac{2b_2}{t} + \sum_{i=1}^{N} \frac{a_i}{(1-p_i)t^{p_i}}. \tag{2.30}$$

*Proof.* For any fixed $t > 0$ the convexity of the mapping $u \mapsto g(u) = (t+u)^{-p}$ implies that $g(1) - g(0) \geq g'(0)$, i.e.,

$$\frac{1}{t^p} - \frac{1}{(t+1)^p} \leq \frac{p}{t^{p+1}}.$$

Thus,

$$\frac{a_i}{t^{p+1}} \leq \frac{a_i}{1-p}\left(\frac{1}{(t+1)^p} - \left(1 - \frac{1}{t}\right)\frac{1}{t^p}\right). \tag{2.31}$$

Using (2.29), and (2.31) and rearranging terms we get

$$b_{t+1} - \sum_{i=1}^{N} \frac{a_i}{(1-p_i)(t+1)^{p_i}} \leq \left(1 - \frac{1}{t}\right)\left[b_t - \sum_{i=1}^{N} \frac{a_i}{(1-p_i)t^{p_i}}\right].$$

Letting $\tau_t = b_t - \sum_{i=1}^{N} \frac{a_i}{(1-p_i)t^{p_i}}$ we have $\tau_{t+1} \leq (1 - \frac{1}{t})\tau_t$. Now, if $\tau_2 \leq 0$ then $\tau_t \leq 0$ for any $t \geq 2$ and thus (2.30) holds. Otherwise, if $\tau_2 > 0$ then for $t \geq 3$ we have

$$\tau_t \leq \tau_2 \prod_{i=2}^{t-1}\left(1 - \frac{1}{i}\right) \leq \frac{2\tau_2}{t} \leq \frac{2b_2}{t},$$

where we have used the inequalities $\sum_{i=2}^{t-1} \log\left(1 - \frac{1}{i}\right) \leq -\sum_{i=2}^{t-1} \frac{1}{i} \leq -\log(t-1) \leq \log(2/t)$. Thus, (2.30) holds in this case as well. □

**Theorem 2.8.5.** *Let* $f \in \mathcal{F}_{\alpha,\beta}(L)$ *with* $\beta \geq 2$, $\alpha, L > 0$, $\sigma > 0$*, and let Assumption 2.2.1 hold. Consider Algorithm 1 where* $\Theta$ *is a convex compact subset of* $\mathbb{R}^d$ *and assume that* $\max_{x\in\Theta} \|\nabla f(x)\| \leq G$.

*(i) If Assumption 2.2.2 holds,* $h_t = \left(\frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2}\right)^{\frac{1}{2\beta}} t^{-\frac{1}{2\beta}}$ *and* $\eta_t = \frac{2}{\alpha t}$ *then for* $t \geq 1$ *we have*

$$\mathbb{E}\left[\,\|x_t - x^*\|^2\,\right] < \frac{2G^2}{\alpha^2 t} + A_5 \frac{d^2}{\alpha^2} t^{-\frac{\beta-1}{\beta}} \tag{2.32}$$

*where* $x^* = \arg\min_{x\in\Theta} f(x)$ *and* $A_5 > 0$ *is a constant that does not depend on* $d, \alpha, t$.

*(ii) If* $\beta = 2$, $h_t = \left(\frac{3d^2\sigma^2}{4L\alpha t + 9L^2d^2}\right)^{1/4}$ *and* $\eta_t = \frac{1}{\alpha t}$ *then for* $t \geq 1$ *we have that*

$$\mathbb{E}\left[\,\|x_t - x^*\|^2\,\right] < \frac{2G^2}{\alpha^2 t} + A_6 \frac{d}{\alpha^{\frac{3}{2}} t^{\frac{1}{2}}} + A_7 \frac{d^2}{\alpha^2 t}, \tag{2.33}$$

*where* $A_6, A_7 > 0$ *are constants that do not depend on* $d, \alpha, t$.

*Proof.* Let $r_t = \mathbb{E}\,\|x_t - x^*\|^2$. To prove the theorem, we will show that under the assumptions of the theorem $\{r_t\}$ satisfies (2.29) with suitable $a_i$ and $p_i$, and then use Lemma 2.8.4.

We start by noticing that, in view of the $\alpha$-strong convexity of $f$ and the fact that $f$ is Lipschitz continuous with constant $G$ in $\Theta$ for any $t \geq 1$ we have

$$\|x_t - x^*\|^2 \leq \frac{G^2}{\alpha^2}. \tag{2.34}$$

Thus, (2.32) and (2.33) hold for $t = 1$ and it suffices to prove the theorem for $t \geq 2$. The definition of Algorithm 1 gives that, for $t \geq 1$,

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - 2\eta_t\langle \hat{g}_t, x_t - x^*\rangle + \eta_t^2 \|\hat{g}_t\|^2 .$$

Taking conditional expectation of both sides of this inequality given $x_t$ we obtain

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 |x_t] \quad \leq \quad \|x_t - x^*\|^2 - 2\eta_t\langle \mathbb{E}[\hat{g}_t|x_t], x_t - x^*\rangle + \eta_t^2\mathbb{E}[\|\hat{g}_t\|^2 |x_t].$$

Using this inequality and Lemmas 2.2.3 and 2.2.4(ii) we find

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x^*\|^2 |x_t] \quad &\leq \quad \|x_t - x^*\|^2 - 2\eta_t\alpha \|x_t - x^*\|^2 + 2\eta_t\kappa_\beta L d h_t^{\beta-1}\|x_t - x^*\| + \\ &+\eta_t^2 \left[\left(9\kappa\left(G^2 d + \frac{L^2 d^2 h_t^2}{2}\right) + \frac{3\kappa d^2\sigma^2}{2h_t^2}\right)\right].\end{aligned} \tag{2.35}$$

On the other hand, for $\lambda > 0$, we have

$$d h_t^{\beta-1} \|x_t - x^*\| \leq \frac{1}{2}\left(\frac{\kappa_\beta L}{\alpha\lambda} d^2 h_t^{2(\beta-1)} + \frac{\alpha\lambda}{\kappa_\beta L}\|x_t - x^*\|^2\right). \tag{2.36}$$

Combining (2.36) and (2.35) we get

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x^*\|^2 |x_t] \quad &\leq \quad (1 - (2-\lambda)\eta_t\alpha) \|x_t - x^*\|^2 + \tfrac{(\kappa_\beta L)^2}{\alpha\lambda}\eta_t d^2 h_t^{2(\beta-1)} + \\ &+\eta_t^2 \left[\left(9\kappa\left(G^2 d + \tfrac{L^2 d^2 h_t^2}{2}\right) + \tfrac{3\kappa d^2\sigma^2}{2h_t^2}\right)\right].\end{aligned} \tag{2.37}$$

Substituting $h_t = \left(\frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2}\right)^{\frac{1}{2\beta}} t^{-\frac{1}{2\beta}}$, $\eta_t = \frac{2}{\alpha t}$, $\lambda = \frac{3}{2}$ in (2.37), and taking the expectation over $x_t$ we obtain

$$\begin{aligned}r_{t+1} \quad &\leq \quad \left(1 - \frac{1}{t}\right) r_t + \frac{4(\kappa_\beta L)^2}{3\alpha^2} d^2 \left(\frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2}\right)^{\frac{\beta-1}{\beta}} t^{-\frac{2\beta-1}{\beta}} + \\ &+\frac{18\kappa L^2 d^2}{\alpha^2}\left(\frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2}\right)^{\frac{1}{\beta}} t^{-\frac{2\beta+1}{\beta}} + \frac{36\kappa}{\alpha^2 t^2} G^2 d + \\ &+\frac{6\kappa d^2\sigma^2}{\alpha^2}\left(\frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2}\right)^{-\frac{1}{\beta}} t^{-\frac{2\beta-1}{\beta}}.\end{aligned}$$

Thus, we have

$$r_{t+1} \quad < \quad \Big(1 - \frac{1}{t}\Big) r_t + C \frac{d^2}{\alpha^2} t^{-\frac{2\beta-1}{\beta}},$$

where

$$\begin{aligned} C \quad = \quad & \frac{4(\kappa_\beta L)^2}{3} \left( \frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2} \right)^{\frac{\beta-1}{\beta}} + 18\kappa L^2 \left( \frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2} \right)^{\frac{1}{\beta}} + \\ & + \frac{36\kappa}{d} G^2 + 6\kappa\sigma^2 \left( \frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2} \right)^{-\frac{1}{\beta}}. \end{aligned}$$

This is a particular instance of (2.29). Therefore, we can apply Lemma 2.8.4, which yields that, for all $t \geq 2$,

$$r_t < \frac{2G^2}{\alpha^2 t} + \beta C \frac{d^2}{\alpha^2} t^{-\frac{\beta-1}{\beta}}.$$

Thus, (2.32) follows.

We now prove (2.33). Since $\beta = 2$, using Lemmas 2.2.3, 2.2.4(ii), and 2.8.3 we obtain

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \, | x_t] \leq (1 - \eta_t \alpha) \|x_t - x^*\|^2 + 2\eta_t L h_t^2 + \eta_t^2 \left[ \left( 9 \left( G^2 d + \frac{L^2 d^2 h_t^2}{2} \right) + \frac{3 d^2 \sigma^2}{2 h_t^2} \right) \right].$$

Setting here $h_t = \left( \frac{3d^2\sigma^2}{4L\alpha t + 9L^2 d^2} \right)^{1/4}$, $\eta_t = \frac{1}{\alpha t}$, and taking the expectation over $x_t$ we get

$$\begin{aligned} r_{t+1} \quad & \leq \quad \Big(1 - \frac{1}{t}\Big) r_t + \left( \frac{(4L\alpha t + 9L^2 d^2)^{1/2}}{\alpha^2} \right) \frac{\sqrt{3} d\sigma}{t^2} + \frac{9G^2 d}{\alpha^2 t^2} \\ & \leq \quad \Big(1 - \frac{1}{t}\Big) r_t + A_6' \frac{d}{\alpha^{\frac{3}{2}} t^{\frac{3}{2}}} + A_7' \frac{d^2}{\alpha^2 t^2}, \end{aligned}$$

where $A_6' = 2\sqrt{3L}\sigma$ and $A_7' = 3\sqrt{3L}\sigma + \frac{9G^2}{d}$. Applying Lemma 2.8.4 for $t \geq 2$ we get

$$r_t < \frac{2G^2}{\alpha^2 t} + 2A_6' \frac{d}{\alpha^{\frac{3}{2}} t^{\frac{1}{2}}} + 2A_7' \frac{d^2}{\alpha^2 t}.$$

□

Consider the estimator

$$\hat{x}_T = \frac{1}{T - \lfloor T/2 \rfloor} \sum_{t = \lfloor T/2 \rfloor + 1}^{T} x_t. \tag{2.38}$$

The following two theorems provide bounds on the optimization error of this estimator.

**Theorem 2.8.6.** *Let $f \in \mathcal{F}_{\alpha,\beta}(L)$ with $\beta \geq 2$, $\alpha, L > 0$, $\sigma > 0$, and let Assumptions 2.2.1 and 2.2.2 hold. Consider Algorithm 1 where $\Theta$ is a convex compact subset of $\mathbb{R}^d$ and assume that*

*$\max_{x\in\Theta}\|\nabla f(x)\| \le G$. If $h_t = \left(\frac{3\kappa\sigma^2}{2(\beta-1)(\kappa_\beta L)^2}\right)^{\frac{1}{2\beta}} t^{-\frac{1}{2\beta}}$ and $\eta_t = \frac{2}{\alpha t}$ then the optimization error of the estimator (2.38) satisfies*

$$\mathbb{E}[f(\hat{x}_T) - f(x^*)] \le \qquad \min\left(GB, \frac{1}{\alpha}\left(d^2\Big(\frac{A_1'}{T^{\frac{\beta-1}{\beta}}} + \frac{A_2'}{T}\Big) + \frac{A_3' d}{T}\right)\right),$$

*where $x^* = \arg\min_{x\in\Theta} f(x)$. Here $A_1', A_2'$ and $A_3'$ are positive constants that do not depend on $d, \alpha, T$, and $B$ is the Euclidean diameter of $\Theta$.*

*Proof.* With the same steps as in the proof of Theorem 2.3.1 (see (2.20)) but taking now the sum over $t = \lfloor T/2\rfloor + 1, \dots, T$ rather than over $t = 1, \dots, T$ we obtain

$$\begin{aligned}\sum_{t=\lfloor T/2\rfloor+1}^{T} \mathbb{E}[f(x_t) - f(x^*)] \le r_{\lfloor T/2\rfloor+1}\frac{\lfloor T/2\rfloor\alpha}{2} + \frac{1}{\alpha}\sum_{t=\lfloor T/2\rfloor+1}^{T} &\Big((\kappa_\beta L)^2 d^2 h_t^{2(\beta-1)} + \\ &+ \frac{1}{t}\Big[9\kappa\Big(G^2 d + \frac{\bar{L}^2 d^2 h_t^2}{8}\Big) + \frac{3\kappa d^2\sigma^2}{2h_t^2}\Big]\Big) \\ \le r_{\lfloor T/2\rfloor+1}\frac{\lfloor T/2\rfloor\alpha}{2} + \frac{9\kappa G^2 d}{\alpha} &\sum_{t=\lfloor T/2\rfloor+1}^{T}\frac{1}{t} \\ &+ \frac{1}{\alpha}\sum_{t=1}^{T}\Big((\kappa_\beta L)^2 d^2 h_t^{2(\beta-1)} + \frac{\bar{L}^2 d^2 h_t^2}{8t} + \frac{3\kappa d^2\sigma^2}{2h_t^2 t}\Big).\end{aligned}$$

For the last sum here, we use exactly the same bound as in the proof of Theorem 2.3.1. Moreover, it follows from Theorem 2.8.5 that

$$r_{\lfloor T/2\rfloor+1} < \frac{4G^2}{\alpha^2 T} + A_5'\frac{d^2}{\alpha^2}T^{-\frac{\beta-1}{\beta}},$$

where $A_5' = 2^{(\beta-1)/\beta}A_5$. Combining these remarks and using the fact that $\sum_{t=\lfloor T/2\rfloor+1}^{T}\frac{1}{t} \le \log(T/\lfloor T/2\rfloor) \le 2$ for all $T \ge 2$ (recall that we assume $T \ge 2$ throughout the paper), as well as the the convexity of $f$ we get

$$\mathbb{E}[f(\hat{x}_T) - f(x^*)] \;\; \le \;\; \frac{1}{\alpha}\left(d^2\Big(\frac{A_1'}{T^{\frac{\beta-1}{\beta}}} + \frac{A_2'}{T}\Big) + \frac{A_3' d}{T}\right),$$

where $A_1' = 2A_1 + \frac{A_5'}{2}$, $A_2' = 2\bar{c}\bar{L}^2(\sigma/L)^{\frac{2}{\beta}}$ with constant $\bar{c}$ as in Theorem 2.3.1 and $A_3' = 2G^2(18\kappa + 1/d)$. On the other hand we have the straightforward bound

$$\mathbb{E}[f(\hat{x}_T) - f(x^*)] \le GB.$$

□

**Theorem 2.8.7.** *Let $f \in \mathcal{F}_{\alpha,2}(L)$ with $\alpha, L > 0$, $\sigma > 0$, and let Assumption 2.2.1 hold. Consider the version of Algorithm 1 as in Theorem 2.5.1 where $\Theta$ is a convex compact subset of $\mathbb{R}^d$ and* *assume that* $\max_{x\in\Theta}\|\nabla f(x)\| \le G$. *If* $h_t = \left(\frac{3d^2\sigma^2}{4L\alpha t+9L^2d^2}\right)^{1/4}$ *and* $\eta_t = \frac{1}{\alpha t}$ *then the optimization error of the estimator (2.38) satisfies*

$$\mathbb{E}[f(\hat{x}_T) - f(x^*)] \le \min\left(GB, A_8\frac{d}{\sqrt{\alpha T}} + A_9\frac{d^2}{\alpha T}\right),$$

*where* $x^* = \arg\min_{x\in\Theta} f(x)$. *Here* $A_8$ *and* $A_9$ *are positive constants that do not depend on* $d, \alpha, T$*, and* $B$ *is the Euclidean diameter of* $\Theta$.

*Proof.* Arguing as in the proof of Theorem 2.5.1 but taking the sum over $\lfloor T/2\rfloor + 1, \dots, T$ rather than over $1, \dots, T$ we obtain

$$\begin{aligned}
\sum_{t=\lfloor T/2\rfloor+1}^{T} \mathbb{E}\big[f(x_t) - f(x^*)\big] &\le r_{\lfloor T/2\rfloor+1}\frac{\lfloor T/2\rfloor\alpha}{2} + \sum_{t=\lfloor T/2\rfloor+1}^{T}\left[\Big(L + \frac{9L^2d^2}{4\alpha t}\Big)h_t^2 + \frac{3d^2\sigma^2}{4h_t^2\alpha t} + \frac{9G^2d}{2\alpha t}\right]\\
&\le r_{\lfloor T/2\rfloor+1}\frac{\lfloor T/2\rfloor\alpha}{2} + \sum_{t=\lfloor T/2\rfloor+1}^{T}\left[\sqrt{3}\frac{d\sigma\sqrt{L}}{\sqrt{\alpha t}} + \frac{3\sqrt{3}Ld^2\sigma}{2\alpha t} + \frac{9G^2d}{2\alpha t}\right]\\
&\le r_{\lfloor T/2\rfloor+1}\frac{\lfloor T/2\rfloor\alpha}{2} + 2\sqrt{3L}\sigma\frac{d}{\sqrt{\alpha}}\sqrt{T} + \frac{3d}{2\alpha}(\sqrt{3}Ld\sigma + 3G^2)\sum_{t=\lfloor T/2\rfloor+1}^{T}\frac{1}{t}\\
&\le r_{\lfloor T/2\rfloor+1}\frac{\lfloor T/2\rfloor\alpha}{2} + 2\sqrt{3L}\sigma\frac{d}{\sqrt{\alpha}}\sqrt{T} + \frac{3d}{\alpha}(\sqrt{3}Ld\sigma + 3G^2),
\end{aligned}$$

where we have used the inequality $\sum_{t=\lfloor T/2\rfloor+1}^{T}\frac{1}{t} \le \log(T/\lfloor T/2\rfloor) \le 2$ for all $T \ge 2$ (recall that we assume $T \ge 2$ throughout the paper). It follows from Theorem 2.8.5, that

$$r_{\lfloor T/2\rfloor+1} < \frac{4G^2}{\alpha^2 T} + \sqrt{2}A_6\frac{d}{\alpha^{\frac{3}{2}}T^{\frac{1}{2}}} + 2A_7\frac{d^2}{\alpha^2 T}.$$

Combining the last two displays yields

$$\sum_{t=\lfloor T/2\rfloor+1}^{T} \mathbb{E}\big[f(x_t) - f(x^*)\big] \le \frac{G^2}{\alpha} + A_6\frac{d}{2\sqrt{2}\sqrt{\alpha}}\sqrt{T} + A_7\frac{d^2}{2\alpha} + 2\sqrt{3L}\sigma\frac{d}{\sqrt{\alpha}}\sqrt{T} + \frac{3d}{\alpha}(\sqrt{3}Ld\sigma + 3G^2).$$

From this inequality, using the fact that $f$ is a convex function, we obtain

$$\mathbb{E}[f(\hat{x}_T) - f(x^*)] \le A_8\frac{d}{\sqrt{\alpha T}} + A_9\frac{d^2}{\alpha T},$$

where $A_8 = \frac{A_6}{\sqrt{2}} + 4\sqrt{3L}\sigma$ and $A_9 = A_7 + 2(3\sqrt{3}L\sigma + (9d+1)G^2/d^2)$. □

**Theorem 2.8.8.** *Let* $f \in \mathcal{F}_{\alpha,2}(L)$ *with* $\alpha, L > 0$*, and let Assumption 2.2.1 hold. Consider the version of Algorithm 1 as in Theorem 2.5.1 where* $\Theta$ *is a convex compact subset of* $\mathbb{R}^d$*, and* $h_t = \left(\frac{3d^2\sigma^2}{4L\alpha t}\right)^{\frac{1}{4}}$, $\eta_t = \frac{1}{\alpha t}$. *If* $f$ *is Lipschitz continuous with Lipschitz constant* $G$ *on the Euclidean*

*$h_1$-neighborhood of $\Theta$, then for $\sigma > 0$ we have the following bound for the cumulative regret:*

$$\forall x \in \Theta : \sum_{t=1}^{T} \mathbb{E}[f(x_t) - f(x)] \le \min\left(GBT, 2\sqrt{3L}\sigma \frac{d}{\sqrt{\alpha}}\sqrt{T} + \frac{C^* G^2}{2}\frac{d}{\alpha}(1+\log T)\right), \quad (2.39)$$

*where $B$ is the Euclidean diameter of $\Theta$.*

*If $\sigma = 0$, then the cumulative regret for any $h_t$ chosen small enough and $\eta_t = \frac{1}{\alpha t}$ satisfies*

$$\forall x \in \Theta : \sum_{t=1}^{T} \mathbb{E}[f(x_t) - f(x)] \le \min\left(GBT, C^* G^2 \frac{d}{\alpha}(1+\log T)\right)$$

*Proof.* The argument is analogous to the proof of Theorem 2.5.1. The difference is only in the bound on $\mathbb{E}[\|\hat{g}_t\|^2]$. To evaluate this term, we now use Lemma 2.8.1 (noticing that when $K(\cdot) \equiv 1$ this lemma is satisfied with $\kappa = 1$). This yields

$$\forall x \in \Theta : \qquad \mathbb{E}\sum_{t=1}^{T}\big(f(x_t) - f(x)\big) \le \sum_{t=1}^{T}\left[Lh_t^2 + \frac{1}{2\alpha t}\left(C^* G^2 d + \frac{3d^2\sigma^2}{2h_t^2}\right)\right].$$

The chosen value $h_t = \left(\frac{3d^2\sigma^2}{4L\alpha t}\right)^{\frac{1}{4}}$ minimizes the r.h.s. and together with (2.27) yields (2.39). The remaining part of the proof follows the same lines as in Theorem 2.3.1. □

**Lemma 2.8.9.** *Let $f_t \in \mathcal{F}_\beta(L)$ where $\beta \in [1,2]$ and $L > 0$. Then for any $x \in \mathbb{R}^d$ and $h_t > 0$ we have*

$$|\hat{f}_t(x) - f_t(x)| \le Lh_t^\beta, \quad (2.40)$$

*and*

$$|\mathbb{E}f_t(x \pm h_t\zeta_t) - f_t(x)| \le Lh_t^\beta. \quad (2.41)$$

*Proof.* For $\beta = 1$ the proof of (2.40) is obvious since $\left\|\tilde{\zeta}\right\| \le 1$. For $1 < \beta \le 2$ using the fact that $\mathbb{E}[\tilde{\zeta}] = 0$ we have

$$|\mathbb{E}\big[f_t(x + h_t\tilde{\zeta}) - f_t(x)\big]| = |\mathbb{E}\big[f_t(x + h_t\tilde{\zeta}) - f_t(x) - \langle \nabla f_t(x), h_t\tilde{\zeta}\rangle\big]| \le Lh_t^\beta \mathbb{E}[\left\|\tilde{\zeta}\right\|^\beta] \le Lh_t^\beta.$$

Thus, (2.40) follows. The proof of (2.41) is analogous. □

# Chapter 3

# Distributed zero-order optimization under adversarial noise

We study the problem of distributed zero-order optimization for a class of strongly convex functions. They are formed by the average of local objectives, associated to different nodes in a prescribed network. We propose a distributed zero-order projected gradient descent algorithm to solve the problem. Exchange of information within the network is permitted only between neighbouring nodes. An important feature of our procedure is that it can query only function values, subject to a general noise model, that does not require zero mean or independent errors. We derive upper bounds for the average cumulative regret and optimization error of the algorithm which highlight the role played by a network connectivity parameter, the number of variables, the noise level, the strong convexity parameter, and smoothness properties of the local objectives. The bounds indicate some key improvements of our method over the state-of-the-art, both in the distributed and standard zero-order optimization settings. We also comment on lower bounds and observe that the dependency over certain function parameters in the bound is nearly optimal.

## 3.1 Introduction

We study the problem of distributed optimization where each node (or agent) has an objective function $f_i : \mathbb{R}^d \to \mathbb{R}$ and exchange of information is limited between neighbouring agents within a prescribed network of connections. The goal is to minimize the average of these objectives on a closed bounded convex set $\Theta \subset \mathbb{R}^d$,

$$\min_{x \in \Theta} f(x) \quad \text{where} \quad f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{3.1}$$

Distributed optimization has been widely studied in the literature, we refer to Boyd et al. (2011); Duchi et al. (2012); Jakovetić (2019); Jakovetić et al. (2014); Kia et al. (2015); Lobel et al. (2011); Nedic and Ozdaglar (2009); Nedic et al. (2010); Pu et al. (2021); Scaman et al. (2019); Shi et al. (2014); Tsitsiklis et al. (1986) and references therein. This problem has broad applications such as multi-agent target seeking Liu et al. (2017), distributed learning Kraska et al. (2013), and wireless networks Park et al. (2021), among others.

We address problem (3.1) from the perspective of zero-order distributed optimization. That is we assume that only function values can be queried by the algorithm, subject to measurement noise. During the optimization procedure, each agent maintains a local copy of the variables which are sequentially updated based on local and neighboring functions' queries. We wish to devise such optimization procedures which are efficient in bounding the average optimization error and cumulative regret in terms of the functions' properties and network topology.

**Contributions** Our principal contribution is a distributed zero-order optimization algorithm, introduced in Section 3.2, which we show to achieve tight rates of convergence under certain assumptions on the objective functions, outlined in Section 3.3. Specifically, we consider that the local objectives $f_i$ are $\beta$-Hölder and the average objective $f$ is $\alpha$-strongly convex. The algorithm relies on a novel zero-order gradient estimator, presented in Section 3.4. Although conceptually very simple, this estimator, when employed within our algorithm, allows us to obtain an $O(d^2)$ computational gain as well as improved error rates than previous state-of-the-art zero-order optimization procedures Akhavan et al. (2020); Bach and Perchet (2016), in the special case of standard (undistributed) setting. Another key advantage of our approach is due to the general noise model presented in Section 3.5, under which function values are queried. The noise variables do not need to be zero mean or independently sampled, and thus they include "adversarial" noise. In Section 3.6, we derive the rates of convergence for the cumulative regret and the optimization error of the proposed algorithm, and in Section 3.7 we consider the special case of 2-smooth functions. The rates highlight the dependency with respect to the number of variables $d$, the number of function queries $T$, the spectral gap of the network matrix $1 - \rho$, and the parameters $n$, $\alpha$ and $\beta$. The bounds enjoy a better dependency on $1 - \rho$ than previous bounds on zero-order distributed optimization Qu and Li

(2018); Tang et al. (2019); Yu et al. (2019). We also compare our bounds to related lower bounds in Akhavan et al. (2020) for undistributed setting, observing that our rates are optimal either with respect to $T$ and $\alpha$, or with respect to $T$ and $d$.

**Previous Work** We briefly comment on previous related work and defer to Section 3.8 for a more in depth discussion and comparison. For both deterministic and stochastic scenarios of problem (3.1), a large body of literature is devoted to first-order gradient based methods with a consensus scheme (see the papers cited above and references therein). On the other hand, the study of zero-order methods was started only recently Hajinezhad et al. (2019); Qu and Li (2018); Sahu et al. (2018a,b); Tang et al. (2019); Yu et al. (2019). The works Qu and Li (2018); Tang et al. (2019); Yu et al. (2019) are dealing with zero-order distributed methods in noise-free settings while the noisy setting is developed in Hajinezhad et al. (2019); Sahu et al. (2018a,b). Namely, Hajinezhad et al. (2019) considers 2-point zero-order methods with stochastic queries for non-convex optimization but assume that the noise is the same for both queries, which makes the problem analogous to noise-free scenario in terms of optimization rates. Papers Sahu et al. (2018a,b) study zero-order distributed optimization for strongly convex and $\beta$-smooth functions $f_i$ with $\beta \in \{2, 3\}$. They derive bounds on the optimization error, though without providing closed form expressions.

**Notation** Throughout we denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ be the standard inner product and Euclidean norm on $\mathbb{R}^d$, respectively, and by $\| \cdot \|_*$ the spectral norm of a matrix. The notation $\mathbb{I}$ is used for the $n$-dimensional identity matrix and $\mathbf{1}$ for the vector in $\mathbb{R}^n$ with all entries equal to 1. We denote by $e_j$ the $j$-th canonical basis vector in $\mathbb{R}^d$. For any set $A$, the number of elements in $A$ is denoted by $|A|$. For $x \in \mathbb{R}$, the value $\lfloor x \rfloor$ is the maximal integer less than $x$. For every closed convex set $\Theta \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$ we denote by $\mathrm{Proj}_\Theta(x) = \mathrm{argmin}\{\|z - x\| : z \in \Theta\}$ the Euclidean projection of $x$ onto $\Theta$. We denote by $\mathrm{diam}(\Theta)$ the Euclidean diameter of $\Theta$. Finally we let $U[-1, 1]$ be the uniform distribution on $[-1, 1]$.

## 3.2 The Problem

Let $n$ be the number of agents and let $\mathcal{G} = (V, E)$ be an undirected graph, where $V = \{1, \ldots, n\}$ is the set of nodes and $E \subseteq V \times V$ is the set of edges. The adjacency matrix of $\mathcal{G}$ is the symmetric matrix $(A_{ij})_{i,j=1}^n$ defined as $A_{ij} = 1$, if $(i, j) \in E$ and zero otherwise. We consider the following sequential learning framework, where each agent $i$ gets values of function $f_i$ corrupted by noise and shares information with other agents. At step $t$, agent $i$ acts as follows:

- makes queries and gets noisy values of $f_i$,
- provides a local output $u^i(t)$ based on these queries and on the past information,
- broadcasts $u^i(t)$ to neighboring agents,

- updates its local variable using information from other agents as follows:

$$x^i(t{+}1){=}\sum_{j=1}^{n} W_{ij}u^j(t),$$

where $W = (W_{ij})_{i,j=1}^n$ is a given matrix called the consensus matrix.

Below we use the following condition on the consensus matrix.

**Assumption 3.2.1.** *Matrix $W$ is symmetric, doubly stochastic, and $\rho := \left\|W - n^{-1}\mathbf{1}\mathbf{1}^\top\right\|_* < 1$.*

Matrix $W$ accounts for the connectivity properties of the network. If $W_{ij} = 0$ the agents $i$ and $j$ are not connected (do not exchange information). Often $W$ is defined as a doubly stochastic matrix function of the adjacency matrix $A$ of the graph. One popular example is as follows:

$$W_{ij} = \begin{cases} \frac{A_{ij}}{\gamma \max\{d(i),d(j)\}} & \text{if } i \neq j, \\ 1 - \sum\limits_{k:k\neq i} \frac{A_{ki}}{\gamma \max\{d(i),d(k)\}} & \text{if } i = j, \end{cases} \tag{3.2}$$

where $d(i) = \sum_{j=1}^n A_{ij}$ is the degree of node $i$ and $\gamma > 0$ is a constant. Then, clearly, $W = (W_{ij})$ is a symmetric and doubly stochastic matrix, and $W_{ij} = 0$ if agents $i$ and $j$ are not connected. Moreover, we have $\rho < 1 - c/n^2$ for a constant $c > 0$ (see Olshevsky (2014); Qu and Li (2018)). Values of spectral gaps $\rho$ for some other $W$ reflecting different network topologies can be found in Duchi et al. (2012). Typically, $\rho < 1 - a_n$, where $a_n = \Omega(n^{-1})$ or $a_n = \Omega(n^{-2})$. Parameter $\rho$ can be viewed as a measure of difference between the distributed problem and a standard optimization problem. If the graph of communication is a complete graph a natural choice is $W = n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$ and then $\rho = 0$. For more examples of consensus matrices $W$, see Duchi et al. (2012); Olshevsky and Tsitsiklis (2009) and references therein.

The local outputs $u^i$ can be defined in different ways. Our approach is outlined in Algorithm 2. At Step 1, an estimate of the gradient of the local objective $f_i$ at $x^i(t)$ is constructed. This involves a randomized procedure that we describe and justify in Section 3.4. The local output $u^i$ is defined as an update of the projected gradient algorithm with such an estimated gradient. At Step 2 of the algorithm, each agent computes the next point by a local consensus gradient descent step, which uses local and neighbor information. Step 2 of the algorithm is known as gossip method, see e.g., Boyd et al. (2006)), which was initially introduced as an approach for the networks with the imposed connection between the nodes changing by time. We also refer to Sayin et al. (2017) for similar algorithms in the context of distributed stochastic first-order gradient methods.

**Algorithm 2** Distributed Zero-Order Gradient

```
Input  Communication matrix $(W_{ij})_{i,j=1}^n$, step sizes $(\eta_t > 0)_{t=1}^{T_0-1}$
Initialization  Choose initial vectors $x^1(1) = \cdots = x^n(1) \in \mathbb{R}^d$
For $t = 1, \ldots, T_0 - 1$
  For $i = 1, \ldots, n$
    1. Build an estimate $g^i(t)$ of the gradient $\nabla f_i(x^i(t))$ using noisy evaluations of $f_i$
    2. Update $x^i(t+1) = \sum_{k=1}^n W_{ik}\mathrm{Proj}_\Theta(x^k(t) - \eta_t g^k(t))$
  End
End
Output  Approximate minimizer $\bar{x}(T_0) = \frac{1}{n}\sum_{i=1}^n x^i(T_0)$ of the average objective $f = \frac{1}{n}\sum_{i=1}^n f_i$
```

## 3.3 Assumptions on Local Objectives

In this section, we give some definitions and introduce our assumptions on the local objective functions $f_1, \ldots, f_n$.

**Definition 3.3.1.** *Denote by $\mathcal{F}_\beta(L)$ the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that are $\ell = \lfloor \beta \rfloor$ times differentiable and satisfy, for all $x, z \in \mathbb{R}^d$ the Hölder-type condition*

$$\left| f(z) - \sum_{0 \le |m| \le \ell} \frac{1}{m!} D^m f(x)(z-x)^m \right| \le L\|z - x\|^\beta,$$

*where $L > 0$, the sum is over the multi-index $m = (m_1, ..., m_d) \in \mathbb{N}^d$, we used the notation $m! = m_1! \cdots m_d!$, $|m| = m_1 + \cdots + m_d$, and we defined, for every $\nu = (\nu_1, \ldots, \nu_d) \in \mathbb{R}^d$,*

$$D^m f(x)\nu^m = \frac{\partial^{|m|} f(x)}{\partial^{m_1} x_1 \cdots \partial^{m_d} x_d} \nu_1^{m_1} \cdots \nu_d^{m_d}.$$

*Elements of the class $\mathcal{F}_\beta(L)$ are referred to as $\beta$-Hölder functions.*

**Definition 3.3.2.** *Function $f : \mathbb{R}^d \to \mathbb{R}$ is called 2-smooth if it is differentiable on $\mathbb{R}^d$ and there exists $\bar{L} > 0$ such that, for every $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$, it holds that*

$$\|\nabla f(x) - \nabla f(x')\| \le \bar{L}\|x - x'\|.$$

**Definition 3.3.3.** *Let $\alpha > 0$. Function $f : \mathbb{R}^d \to \mathbb{R}$ is called $\alpha$-strongly convex if $f$ is differentiable on $\mathbb{R}^d$ and*

$$f(x) - f(x') \ge \langle \nabla f(x'), x - x' \rangle + \frac{\alpha}{2}\left\|x - x'\right\|^2, \ \forall x, x' \in \mathbb{R}^d.$$

**Assumption 3.3.4.** *Functions $f_1, \ldots, f_n$: (i) belong to the class $\mathcal{F}_\beta(L)$, for some $\beta \ge 2$, and (ii) are 2-smooth.*

In Section 3.6 we will analyse the convergence properties of Algorithm 2 when the objective function $f$ in 3.1 is $\alpha$-strongly convex. We stress that we do not need the functions

**Algorithm 3** Gradient Estimator with $2d$ Queries

```
Input   Function $F : \mathbb{R}^d \to \mathbb{R}$ and point $x \in \mathbb{R}^d$
Requires Kernel $K : [-1, 1] \to \mathbb{R}$, parameter $h > 0$
Initialization   Generate random $r$ from uniform distribution on $[-1, 1]$
For $j = 1, \dots, d$
    1. Obtain noisy values $y_j = F(x + hre_j) + \xi_j$ and $y'_j = F(x - hre_j) + \xi'_j$
    2. Compute $g_j = \frac{1}{2h}(y_j - y'_j)K(r)$
End
Output   $g = (g_j)_{j=1}^d \in \mathbb{R}^d$ estimator of $\nabla F(x)$
```

$f_1, \dots, f_n$, to be as well $\alpha$-strongly convex. It is enough to make such an assumption on the compound function $f$, while the local functions $f_i$ only need to satisfy the smoothness conditions stated in Assumption 3.3.4 above.

## 3.4 Gradient Estimator

In this section, we detail our choice of gradient estimators $g^i(t)$ used at Step 1 of Algorithm 2. We consider Algorithm 3. For any function $F : \mathbb{R}^d \to \mathbb{R}$ and any point $x$, the vector $g$ returned by Algorithm 3 is an estimate of $\nabla F(x)$ based on noisy observations of $F$ at randomized points. The estimator is computed for every node $i$ at each step $t$, thus giving the vectors $g = g^i(t)$ in Algorithm 2. The gradient estimator crucially requires a kernel function $K : [-1, 1] \to \mathbb{R}$ that allows us to take advantage of possible higher order smoothness properties of $f$. Specifically, in what follows we assume that

$$\int uK(u)du = 1, \int u^j K(u)du = 0, \ j = 0, 2, 3, \dots, \ell, \text{ and } \kappa_\beta \equiv \int |u|^\beta |K(u)|du < \infty, \quad (3.3)$$

for given $\beta \geq 2$ and $\ell = \lfloor \beta \rfloor$. In Polyak and Tsybakov (1990) such kernels can be constructed as weighted sums of Legendre polynomials, in which case $\kappa_\beta \leq 2\sqrt{2}\beta$ with $\beta \geq 1$; see also Appendix A.3 in Bach and Perchet (2016) for a derivation.

The gradient estimator in Algorithm 3 differs from the standard $2d$-point Kiefer-Wolfowitz type estimator in that it uses multiplication by a random variable $K(r)$ with a well-chosen kernel $K$. On the other hand, it is also different from the previous kernel-based estimators in zero-order optimization literature Akhavan et al. (2020); Bach and Perchet (2016); Polyak and Tsybakov (1990) in that it needs $2d$ function queries per step, whereas those estimators require only one or two queries; see, in particular, Algorithm 1 in Akhavan et al. (2020) for a comparison. At first sight, this seems a big drawback of the estimator proposed here, however we will show below that thanks to this estimator we achieve both a more efficient optimization procedure and better rate of convergences for the optimization error.

When the estimator in Algorithm 3 is used at the $t$-th outer step of Algorithm 2, it should be intended as a random variable that depends on the randomization used during the current estimation at the given node, as well as on the randomness of the past iterations, inducing the $\sigma$-algebra $\mathcal{F}_t$ (see Section 3.5 for the definition). Bounds for the bias of this estimator conditional on the past and for its second moment play an important role below, in our analysis of the convergence rates. These bounds are presented in the next two lemmas, whose proofs are presented in Section 3.9. We state them in the simpler setting of Algorithm 3, with no reference to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$.

**Lemma 3.4.1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function in $\mathcal{F}_\beta(L)$, $\beta \geq 2$, and let the random variables $\xi_1, \dots, \xi_d$ and $\xi'_1, \dots, \xi'_d$ be independent of $r$ and satisfy $\mathbb{E}[|\xi_j|] < \infty$, $\mathbb{E}[|\xi'_j|] < \infty$, for $j = 1, \dots, d$. Let the kernel satisfy conditions (3.3). If the gradient estimator $g$ of $f$ given by Algorithm 3 then, for all $x \in \mathbb{R}^d$,*

$$\|\mathbb{E}[g] - \nabla f(x)\| \leq L\kappa_\beta \sqrt{d} h^{\beta-1}.$$

It is straightforward to see that the bound of Lemma 3.4.1 holds when the estimators are build recursively during the execution of Algorithm 1 and the expectation is taken conditionally on $\mathcal{F}_t$. This will be used in the proofs.

**Lemma 3.4.2.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $2$-smooth and let $\max_{x\in\Theta} \|\nabla f(x)\| \leq G$, $\kappa \equiv \int K^2(u)du < \infty$. Let the random variables $\xi_1, \dots, \xi_d$ and $\xi'_1, \dots, \xi'_d$ be independent of $r$ and $\mathbb{E}[\xi_j^2] \leq \sigma^2$, $\mathbb{E}[(\xi'_j)^2] \leq \sigma^2$ for $j = 1, \dots, d$. If $g$ is defined by Algorithm 3, where $x$ is a random variable with values in $\Theta$ independent of $r$ and depending on $\xi_1, \dots, \xi_d$ and $\xi'_1, \dots, \xi'_d$ in an arbitrary way, then*

$$\mathbb{E}\|g\|^2 \leq \frac{3d\kappa}{2}\left(\frac{\sigma^2}{h^2} + \frac{3\bar{L}^2}{4}h^2\right) + 9G^2\kappa.$$

## 3.5 Noise Model

Algorithm 3 is called to compute estimators of gradients of the local functions $f_i, i = 1, \dots n$, at each iteration $t$ of Algorithm 2. Thus, we assume that agent $i$ at iteration $t$ generates a uniform random variable $r_i(t) \sim U[-1, 1]$ and gets $2d$ noisy observations, defined, for $j = 1, \dots, d$

$$\begin{aligned} y_{i,j}(t) &= f(x^i(t) + h_t r_i(t) e_j) + \xi_{i,j}(t) \\ y'_{i,j}(t) &= f(x^i(t) + h_t r_i(t) e_j) + \xi'_{i,j}(t) \end{aligned}$$

where the parameters $h_t > 0$ will be specified later.

In what follows, we denote by $\mathcal{F}_t$ the $\sigma$-algebra generated by the random variables $x^i(t)$, for $i = 1, \dots, n$. In order to meet the conditions of Lemmas 3.4.1 and 3.4.2 for each $(i, t)$, we impose the following assumption on the collection of random variables $(r_i(t), \xi_{i,j}(t), \xi'_{i,j}(t))$.

**Assumption 3.5.1.** *For all integers $t$ and $i \in \{1, \dots, n\}$ the following properties hold.*

*(i) The random variables $r_i(t) \sim U[-1, 1]$ are independent of $\xi_{i,1}(t), \dots \xi_{i,d}(t)$, $\xi'_{i,1}(t), \dots, \xi'_{i,d}(t)$ and from the $\sigma$-algebra $\mathcal{F}_t$,*

*(ii)* $\mathbb{E}[(\xi_{i,j}(t))^2] \le \sigma^2$, $\mathbb{E}[(\xi'_{i,j}(t))^2] \le \sigma^2$ *for* $j = 1, \dots, d$, *and some* $\sigma \ge 0$.

Assumption 3.5.1 is very mild. Indeed, its part (i) occurs as a matter of course since it is unnatural to assume dependence between the random environment noise and artificial random variables $r_i(t)$ generated by the agents. We state (i) only for the purpose of formal rigor. Remarkably, we do not assume the noises $\xi_{i,j}(t)$ and $\xi'_{i,j}(t)$ to have zero mean. What is more, these variables can be deterministic and no independence between them for different $i, j, t$ is required, so we consider an adversarial environment. Having such a relaxed assumption on the noise is possible because of the multiplication by the zero-mean variable $K(r)$ in Algorithm 3. This and the fact that all components of the vectors are treated separately allows the proofs go through without the zero-mean assumption and under arbitrary dependence between the noises.

## 3.6 Main Results

In this section, we provide upper bounds on the performance of the proposed algorithms. Recall that $T_0$ is the number of outer iterations in Algorithm 2. Let $T$ be the total number of times that we observed noisy values of each $f_i$. At each iteration of Algorithm 3 we make $2d$ queries. Thus, to keep the total budget equal to $T$ we need to make $T_0 = T/(2d)$ steps of Algorithm 2 (assuming that $T/(2d)$ is an integer). We compare our results to lower bounds for any algorithm with the total budget of $T$ queries.

For given $\beta \ge 2$, we choose the tuning parameters $\eta_t$ and $h = h_t$ in Algorithms 1 and 3 as

$$\eta_t = \frac{2}{\alpha t}, \qquad \text{and} \qquad h_t = t^{-\frac{1}{2\beta}}. \tag{3.4}$$

Inspection of the proofs in Section 3.9 shows that these values of $\eta_t$ and $h_t$ lead to the best rates minimizing the bounds. As one can expect, there are two contributions to the bounds, one representing the usual stochastic optimization error, while the second one accounts for the distributed character of the problem. This second contribution to the bounds is driven by the following quantity that we call the mean discrepancy: $\Delta(t) \equiv n^{-1}\sum_{i=1}^{n}\mathbb{E}[\|x^i(t) - \bar{x}(t)\|^2]$. It plays an important role in our argument and may be of interest by itself, cf. Tang et al. (2019). The next lemma gives a control of the mean discrepancy.

**Lemma 3.6.1.** *Let Assumptions 3.2.1, 3.3.4, and 3.5.1 hold. Let* $\Theta$ *be a convex compact subset of* $\mathbb{R}^d$*. Assume that diam*$(\Theta) \le \mathcal{K}$ *and* $\max_{x\in\Theta}\|\nabla f(x)\| \le G$*. If the updates* $x^i(t), \bar{x}(t)$ *are defined by Algorithm 1, in which the gradient estimators for* $i$*-th agent are defined by Algorithm 3 with* $F = f_i$, $i = 1, \dots, n$, *and parameters (3.4) then*

$$\Delta(t) \le \mathcal{A}\left(\frac{\rho}{1-\rho}\right)^2 \frac{d}{\alpha^2} t^{-\frac{2\beta-1}{\beta}}, \tag{3.5}$$

*where $\mathcal{A}$ is a constant independent of $t, d, \alpha, n, \rho$. The explicit value of $\mathcal{A}$ can be found in the proof.*

*Proof Sketch.* Let $V(t) = \sum_{i=1}^n \left\|x^i(t) - \bar{x}(t)\right\|^2$, and $z^i(t) = \mathrm{Proj}_\Theta\left(x^i(t) - \eta_t g^i(t)\right) - \left(x^i(t) - \eta_t g^i(t)\right)$. The first step is to show that, due to the definition of the algorithm and Assumptions 3.2.1 on matrix $W$, we have

$$V(t+1) \leq \rho^2 \sum_{i=1}^n \left\|x^i(t) - \bar{x}(t) - \eta_t(g^i(t) - \bar{g}(t)) + z^i(t) - \bar{z}(t)\right\|^2, \tag{3.6}$$

where $\bar{g}(t)$ and $\bar{z}(t)$ denote the averages of $g^i(t)$'s and $z^i(t)$'s over the agents $i$. From (3.6), by using the fact that $\|z^i(t)\| \leq \eta_t \|g^i(t)\|$, applying Lemma 3.4.1 conditionally on $\mathcal{F}_t$, taking expectations and then applying Lemma 3.4.2 we deduce the recursion

$$\Delta(t+1) \leq \rho\Delta(t) + \mathcal{A}_1 \frac{\rho^2}{1-\rho} \cdot \frac{d}{\alpha^2} t^{-\frac{2\beta-1}{\beta}},$$

where $\mathcal{A}_1 > 0$ is a constant. The initialization of Algorithm 1 is chosen so that $\Delta(1) = 0$. It follows that $\Delta(t)$ is bounded by a discrete convolution that can be carefully evaluated leading to (3.5). □

Using Lemma 3.6.1 we obtain the following theorem.

**Theorem 3.6.2.** *Let $f$ be an $\alpha$-strongly convex function and let the assumptions of Lemma 3.6.1 be satisfied. Then for any $x \in \Theta$ the cumulative regret satisfies*

$$\sum_{t=1}^{T_0} \mathbb{E}\left[f(\bar{x}(t)) - f(x)\right] \leq \frac{d}{\alpha(1-\rho)} T_0^{\frac{1}{\beta}} \left(\mathcal{B}_1 + \mathcal{B}_2\rho^2\right) + \frac{\mathcal{B}_3}{\alpha(1-\rho)}(\log(T_0) + 1),$$

*where the positive constants $\mathcal{B}_i$ are independent of $T_0, d, \alpha, n, \rho$. The explicit values of these constants can be found in the proof. Furthermore, if $x^*$ is the minimizer of $f$ over $\Theta$ the optimization error of the averaged estimator $\hat{x}(T_0) = \frac{1}{T_0}\sum_{t=1}^{T_0} \bar{x}(t)$ satisfies*

$$\mathbb{E}[f(\hat{x}(T_0)) - f(x^*)] \leq \frac{d}{\alpha(1-\rho)} T_0^{-\frac{\beta-1}{\beta}} \left(\mathcal{B}_1 + \mathcal{B}_2\rho^2\right) + \frac{\mathcal{B}_3}{\alpha(1-\rho)}\left(\frac{\log(T_0)+1}{T_0}\right). \tag{3.7}$$

*Proof sketch.* Note first that, due to the definition of Algorithm 1 and to the properties of matrix $W$ we have $\bar{x}(t+1) = \bar{x}(t) - \eta_t \bar{g}(t) + \bar{z}(t)$. This resembles the usual recursion of the gradient algorithm with an additional term $\bar{z}(t) = n^{-1}\sum_{i=1}^n z^i(t)$, where $\|z^i(t)\| \leq \eta_t\|g^i(t)\|$. Using this bound and $\alpha$-strong convexity of $f$, analyzing the recursion in the standard way and taking conditional expectations we obtain that, for any $x \in \Theta$,

$$f(\bar{x}(t)) - f(x) \leq \frac{1}{2\eta_t}\mathbb{E}\big[a_t - a_{t+1}|\mathcal{F}_t\big] - \frac{\alpha a_t}{2} + \frac{2\eta_t}{n}\sum_{i=1}^{n}\mathbb{E}\big[\left\|g^i(t)\right\|^2|\mathcal{F}_t\big] \\ + \underbrace{\left\|\mathbb{E}\big[\bar{g}(t)|\mathcal{F}_t\big] - \nabla f(\bar{x}(t))\right\|\|\bar{x}(t) - x\|}_{\text{Bias1}} + \underbrace{\frac{1}{\eta_t}\mathbb{E}\big[\langle \bar{z}(t), \bar{x}(t) - x\rangle|\mathcal{F}_t\big]}_{\text{Bias2}}, \tag{3.8}$$

where $a_t = \|\bar{x}(t) - x\|^2$. Here, the term $\text{Bias2}$ is entirely induced by the distributed nature of the problem. Using the properties of Euclidean projection and some algebra, it can be bounded as

$$\text{Bias2} \leq \frac{3\eta_t}{2(1-\rho)n}\sum_{i=1}^{n}\mathbb{E}\big[\left\|g^i(t)\right\|^2|\mathcal{F}_t\big] + \frac{1-\rho}{2n\eta_t}\sum_{i=1}^{n}\left\|x^i(t) - \bar{x}(t)\right\|^2.$$

On the other hand, $\text{Bias1}$ accumulates two contributions, the first due to the gradient approximation (cf. Lemma 3.4.1) and the second due to the distributed nature of the problem:

$$\text{Bias1} \leq \kappa_\beta L\sqrt{d}h_t^{\beta-1}\|\bar{x}(t) - x\| + \frac{\bar{L}}{n}\sum_{i=1}^{n}\left\|x^i(t) - \bar{x}(t)\right\|\|\bar{x}(t) - x\| \\ \leq \left(\frac{(\kappa_\beta L)^2}{\alpha}dh_t^{2(\beta-1)} + \frac{\alpha a_t}{4}\right) + \left(\frac{\bar{L}t\alpha(1-\rho)}{n}\sum_{i=1}^{n}\left\|x^i(t) - \bar{x}\right\|^2 + \frac{\bar{L}\mathcal{K}^2}{4t\alpha(1-\rho)}\right). \tag{3.9}$$

Next, we combine inequalities (3.8)–(3.9), take expectations of both sides of the resulting inequality, and use Lemmas 3.4.2 and 3.6.1 to bound the second moments $\mathbb{E}\big[\left\|g^i(t)\right\|^2\big]$ and the mean discrepancy. The final result is obtained by summing up from $t = 1$ to $t = T_0$ and recalling that $\eta_t = \frac{2}{\alpha t}$, $h_t = t^{-\frac{1}{2\beta}}$. □

Due to $\alpha$-strong convexity of $f$, Theorem 3.6.2 immediately implies a bound on the estimation error $\mathbb{E}[\|\hat{x}(T_0) - x^*\|^2]$. The bound is of the order of the right-hand side of (3.7) divided by $\alpha$. Furthermore, we get the following result about local estimators, which follows from a slight modification of Lemma 3.6.1 and Theorem 3.6.2.

**Corollary 3.6.3.** *Let Assumptions 3.2.1, 3.3.4, and 3.5.1 hold. Let $\Theta$ be a convex compact subset of $\mathbb{R}^d$. Assume that diam$(\Theta) \leq \mathcal{K}$ and $\max_{x\in\Theta}\|\nabla f(x)\| \leq G$. If the updates $x^i(t)$ are defined by Algorithm 1, in which the gradient estimators for $i$-th agent are defined by Algorithm 3 with $F = f_i$, $i = 1, \dots, n$, and parameters $\eta_t = \frac{4}{\alpha(t+1)}, h_t = t^{-\frac{1}{2\beta}}$ then the local average estimator $\hat{x}^i(T_0) = \frac{2}{T_0(T_0+1)}\sum_{t=1}^{T_0} t x^i(t)$ satisfies*

$$\mathbb{E}[\|\hat{x}^i(T_0) - x^*\|^2] \leq \mathcal{C}\min\left\{1, \frac{d}{\alpha^2(1-\rho)}T_0^{-\frac{\beta-1}{\beta}}\left(1 + \frac{n\rho^2}{(1-\rho)T_0}\right)\right\}, \quad i = 1, \dots, n,$$

*where $\mathcal{C} > 0$ is a positive constant independent of $T_0, d, \alpha, n, \rho$.*

We now state a corollary of Theorem 3.6.2 for an algorithm with total budget of $T$ queries.

Assume that $T_0 = T/(2d)$ is an integer. As our algorithm makes $2d$ queries per step the estimator $\hat{x}(T/(2d))$ uses the total budget of $T$ queries. Combining Theorem 3.6.2 with the trivial bound $\mathbb{E}[f(\hat{x}(T/(2d)) - f(x^*)] \le G\mathcal{K}$ we get the following result.

**Corollary 3.6.4.** *Let $T \ge 2d$ and let the assumptions of Theorem 3.6.2 be satisfied. Then we have*

$$\mathbb{E}[f(\hat{x}(T/(2d)) - f(x^*)] \le \mathcal{C} \min\left\{1, \frac{d^{2-1/\beta}}{\alpha(1-\rho)} T^{-\frac{\beta-1}{\beta}}\right\},$$

*where $\mathcal{C} > 0$ is a positive constant independent of $T, d, \alpha, n, \rho$.*

We now state several important implications of our results.

**Remark 3.6.5.** *Previous bounds on zero-order distributed optimization Qu and Li (2018); Tang et al. (2019); Yu et al. (2019) contain a dependency of $(1-\rho)^{-2}$ in the "connectivity" parameter $\rho$. While Theorem 3.6.2 covers a more difficult noisy setting, our bound displays a better dependency of $(1-\rho)^{-1}$. Since most common values of $1-\rho$ are of the order $n^{-2}$ (or $n^{-1}$), this represents a substantial gain.*

**Remark 3.6.6.** *The case $n = 1$, $\rho = 0$ corresponds to usual (undistributed) zero-order stochastic optimization. Then Corollary 3.6.4 gives a bound of order $\min\left(1, \frac{d^{2-1/\beta}}{\alpha} T^{-\frac{\beta-1}{\beta}}\right)$. This improves upon the bound[1] $\min\left(1, \frac{d^2}{\alpha} T^{-\frac{\beta-1}{\beta}}\right)$ obtained under the same assumptions in Akhavan et al. (2020). Still our bound does not match the minimax lower bound established in Akhavan et al. (2020) and equal to*

$$\min\left(\max(\alpha, T^{-1/2+1/\beta}), \frac{d}{\sqrt{T}}, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\right). \tag{3.10}$$

*For $\alpha \asymp 1$ the lower bound (3.10) scales as $\min\left(1, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\right)$. It has the same behavior in the interesting regime of $\alpha$ not too small ($\alpha \ge T^{-1/2+1/\beta}$) and $T \ge d$. Note, however, that the lower bound (3.10) is obtained for the setting with i.i.d. noise, while our upper bound is valid under adversarial noise. Therefore, it may seem rather surprising that the ratio is only $d^{1-1/\beta}$.*

**Remark 3.6.7.** *With the same budget of queries $T$, the $2d$-point method in Algorithm 3 is computationally simpler than the methods with one or two queries per step Akhavan et al. (2020); Bach and Perchet (2016); Polyak and Tsybakov (1990) previously suggested for the same setting. For example, the method in Akhavan et al. (2020); Bach and Perchet (2016) prescribes, at each step $t = 1, \ldots, T$, to generate a random variable uniformly distributed on the unit sphere in $\mathbb{R}^d$. This requires of order $d$ calls of one-dimensional random variable generator. Overall, in $T$ steps, the number of calls is of order $dT$. For our method with the*

[1] The recent work Novitskii and Gasnikov (2021) obtains the same improvement using the gradient estimator of Akhavan et al. (2020). However, as we explain in Remark 3.6.7 that estimator is less appealing from the computational point of view.

*same budget $T$, we make  of order $T_0 = T/(2d)$ steps and at each step we need to call the generator only once in order to get $r \sim U[-1, 1]$. Thus, with the same budget of queries, Algorithm 3 needs $\sim 1/d^2$ less calls of random variable generator than the gradient estimator in Akhavan et al. (2020); Bach and Perchet (2016).*

In Section 3.9, we present a numerical comparison between the algorithm proposed in this paper and that in Akhavan et al. (2020). The results confirm our theoretical findings. The algorithm of this paper converges faster and the advantage is more pronounced as $d$ increases.

## 3.7 Improved Bounds for $\beta = 2$

In this section we provide improved upper bounds for the case $\beta = 2$ in Corollary 3.6.4, where we relax the dependency over $d$, from $d^{3/2}$ to $d$.

Following the literature on undistributed zero-order optimization, we use a standard 2-point method with elements of the analysis developed in Agarwal et al. (2010); Akhavan et al. (2020); Duchi et al. (2015); Flaxman et al. (2005); Shamir (2013, 2017) among others. Specifically, we define

$$g^i(t) = \frac{d}{2h_t}(y_i(t) - y_i'(t))\zeta_i(t) \tag{3.11}$$

$$\text{where} \;\; y_i(t) = f_i(x^i(t) + h_t\zeta_i(t)) + \xi_i(t), \;\; y_i'(t) = f_i(x^i(t) - h_t\zeta_i(t)) + \xi_i'(t),$$

with the random variables $\zeta_i(t)$, $1 \leq i \leq n$, $1 \leq t \leq T$, that are i.i.d. uniformly distributed on the unit Euclidean sphere in $\mathbb{R}^d$. We make the following assumption on the noise analogous to Assumption 3.5.1.

**Assumption 3.7.1.** *For all integers $t$ and all $i \in \{1, \dots, n\}$ the following properties hold.*

*(i) The random variables $\zeta_i(t)$ are independent of $\xi_i(t)$, $\xi_i'(t)$ and from the $\sigma$-algebra $\mathcal{F}_t$,*

*(ii) $\mathbb{E}[(\xi_i(t))^2] \leq \sigma^2$, $\mathbb{E}[(\xi_i'(t))^2] \leq \sigma^2$ for some $\sigma \geq 0$.*

**Theorem 3.7.2.** *Let $f$ be an $\alpha$-strongly convex function. Let Assumptions 3.2.1, 3.3.4, and 3.7.1 hold with $\beta = 2$. Let $\Theta$ be a convex compact subset of $\mathbb{R}^d$, and assume that diam$(\Theta) \leq \mathcal{K}$. Assume that $\max_{x\in\Theta} \|\nabla f_i(x)\| \leq G$, for $1 \leq i \leq n$. Let the updates $x^i(t), \bar{x}(t)$ be defined by Algorithm 1, in which the gradient estimator for $i$-th agent is defined by (3.11), and $\eta_t = \frac{1}{\alpha t}$, $h_t = \left(\frac{3d^2\sigma^2}{2L\alpha t + 9L^2d^2}\right)^{1/4}$. Then for the estimator $\tilde{x}(T) = \frac{1}{T - \lfloor T/2 \rfloor}\sum_{t=\lfloor T/2\rfloor+1}^{T} \bar{x}(t)$ we have*

$$\mathbb{E}[f(\tilde{x}(T)) - f(x^*)] \leq \frac{\mathcal{B}}{1-\rho}\left(\frac{d}{\sqrt{\alpha T}} + \frac{d^2}{\alpha T}\right),$$

*where $\mathcal{B} > 0$ is a constant independent of $T, d, \alpha, n, \rho$.*

The main idea of the proof is to use surrogate functions $\hat{f}_t^i(x)$, for $1 \leq i \leq n$, defined, for every $x \in \mathbb{R}^d$, as $\hat{f}_t^i(x) = \mathbb{E} f_i(x + h_t \tilde{\zeta})$, where the expectation with respect to the random vector $\tilde{\zeta}$ uniformly distributed on the unit ball $B_d = \{u \in \mathbb{R}^d : \|u\| \leq 1\}$. A result, which can be traced back to Nemirovsky and Yudin (1983) implies the fact that $g^i(t)$ is an unbiased estimator of the gradient of the surrogate function $\hat{f}_t^i$ at $x^i(t)$. Thus, we can consider Algorithm 1 as a gradient descent for the surrogate function. Then replacing $f_i$ and $f$ by the surrogate functions with the cost of the order $h_t^2$, we can recover the initial problem. This method does not work for $\beta > 2$ since the error of approximation by surrogate function becomes of bigger order than the optimal rate $T^{-\frac{\beta-1}{\beta}}$. The results that we implement as tools for this section are given in Section 3.9.

Combining Theorem 3.7.2 with the obvious bound $\mathbb{E}[f(\tilde{x}(T)) - f(x^*)] \leq G\mathcal{K}$ we obtain

$$\mathbb{E}[f(\tilde{x}(T)) - f(x^*)] \leq \frac{\mathcal{B}'}{1-\rho} \min\left(1, \frac{d}{\sqrt{\alpha T}}\right), \tag{3.12}$$

where $\mathcal{B}' > 0$ is a constant independent of $T, d, \alpha, n, \rho$. By comparing this upper bound with the minimax lower bound (3.10) for $\beta = 2$, one can note that (3.12) is optimal with respect to the parameters $T$ and $d$ when $\alpha \asymp 1$.

## 3.8 Discussion

We expand our discussion on previous related work, comparing our results to the state-of-the-art distributed and undistributed zero-order optimization settings, and highlight few key open problems.

**Comparison to Zero-Order Distributed Settings** Distributed opimization with noisy functions' queries was considered in detail in Sahu et al. (2018a,b), where the setting differs from ours in some key aspects: the updates are obtained not as in Step 2 of Algorithm 1 but rather via decentralized techniques, matrix $W$ is random, the noise is zero-mean random rather than adversarial, and 2-point gradient estimator is used. Papers Sahu et al. (2018a,b) provide, for $\beta = 2$ and $\beta = 3$, bounds on $\mathbb{E}[\|x^i(T) - x^*\|^2]$ of the order at least $\frac{n^{3/2}}{(1-\rho)^2} T^{-1/2}$ and $\frac{n^{3/2}}{(1-\rho)^2} T^{-2/3}$, respectively, as functions of $n$, $\rho$ and $T$. Their bounds contain uncontrolled terms of the form $\mathbb{E}[\|x^i(k_0) - x^*\|^2]$ for some large enough $k_0 = k_0(n, \alpha, d)$ leaving unclear the resulting rate. Paper Hajinezhad et al. (2019) considers 2-point methods with stochastic queries but assume that the noise is the same for both queries and deal with non-convex optimization. Noisy-free zero-order distributed optimization is studied by Qu and Li (2018); Tang et al. (2019); Yu et al. (2019). From these, Tang et al. (2019) is the closest to our work as it builds on the updates as at Step 2 of Algorithm 1 (though without projections). The bounds obtained therein are of the order $(1-\rho)^{-2}$ considered as functions of $\rho$, although they hold for the larger class of gradient dominant functions. As noted in Remark 3.6.5 the bound of Theorem 3.6.2 scales only as $(1-\rho)^{-1}$ and this bound holds true, in particular, for noisy-free setting, which is its special case corresponding to $\sigma = 0$. Since most common values of $1 - \rho$ are of the order $n^{-2}$ (or $n^{-1}$), this represents a substantial gain. Moreover, Theorem 3.6.2 covers a difficult noise setting as we deal with adversarial noise. It is also worthwhile to note that the first-order distributed optimization exhibits much better dependency on $\rho$ since bounds that scale as $(1-\rho)^{-1/2}$ can be achieved, see (Duchi et al., 2015; Scaman et al., 2019). Some of the references mentioned above considered unconstrained optimization while our results deal with constrained optimization. Note that the only difference in the proofs of the upper bounds for constrained and unconstrained cases is in the presence of an additional term proportional to the second moment of the gradient at the update (see Lemma 2.4 in Akhavan et al. (2020) for a similar argument). Since this additional term is bounded independently of $\rho$ the overall dependency on $\rho$ remains the same.

**Computational and Statistical Advantage of the Proposed Gradient Estimator** As we highlighted in Section 3.6 the gradient estimator in Algorithm 3 requires $2d$ function queries. At first sight this seems problematic when the dimension $d$ is high, as they need at least $T = 2d$ queries. However, the lower bounds in Akhavan et al. (2020); Shamir (2013) reported in (3.10) above indicate that no estimator can achieve nontrivial convergence rate for zero-order optimization when $T \lesssim d^{\frac{\beta}{\beta-1}}$. Thus, having the total budget of $T \gg d$ queries is a necessary condition for success of any zero-order stochastic optimization method. Algorithms with one or two queries per step can, of course, be realized for $T \lesssim d$ but in this case they do not enjoy any nontrivial error behavior. Moreover, by Remark 3.6.7, with the same total budget of queries $T$, the gradient estimator from Algorithm 3 is computationally more efficient[2] than the estimators in Akhavan et al. (2020); Bach and Perchet (2016); Polyak and Tsybakov (1990). Indeed, with the same budget of queries, it needs $1/d^2$ less calls of random variable generator than it would be for the gradient estimator in Akhavan et al. (2020); Bach and Perchet (2016). At the same time, as detailed in Remark 3.6.6 the proposed gradient estimator yields better rates on the optimization error. We conclude that the proposed zero-order optimization procedure provides both a computational and statistical improvement over the state-of-the-art methods in Akhavan et al. (2020).

**Limitations and Future Work** A main problem, which remains open, is to study whether the dependency of $(1-\rho)^{-1}$ in the upper bounds in Corollary 3.6.4 and Theorem 3.7.2 is minimax optimal. Moreover, in the standard (undistributed) setting it remains an open problem to design a zero-order optimization procedure that meets the minimax lower bound (3.10) with respect to all problem parameters ($T, d, \beta$ and $\alpha$). Further directions of research include the analysis of disturbed zero-order algorithms for larger classes of functions, such as $\alpha$-gradient dominant ones, as well as extension of our results to stochastic updates or asynchronous activation schemes.

[2] One may object that the computation bottleneck in zero-order optimization is in function evaluation; however such costs are *external* to the optimization procedure, for example they may be performed by black-box software running on external machines or devices. Thus such costs should not be taken into account in evaluating the procedure itself. In this sense our computational speedup is important for high dimensional settings.

## 3.9 Proofs and numerical illustration

### Auxiliary Lemma

**Lemma 3.9.1.** *Let $W$ be a matrix satisfying Assumption 3.2.1 and let $x^i = \sum_{j=1}^n W_{i,j}u^j$ for $i = 1, \dots, n$, where $u^1, \dots, u^n$ are some vectors in $\mathbb{R}^d$. Set $\bar{x} = n^{-1}\sum_{i=1}^n x^i$, $\bar{u} = n^{-1}\sum_{i=1}^n u^i$. Then*

$$\sum_{i=1}^n \left\|x^i - \bar{x}\right\|^2 \le \rho^2 \sum_{i=1}^n \left\|u^i - \bar{u}\right\|^2 .$$

*Proof.* Introduce the matrices $X^\top = (x^1, \dots, x^n) \in \mathbb{R}^{d\times n}$, $U^\top = (u^1, \dots, u^n) \in \mathbb{R}^{d\times n}$ and the centering matrix $H = \mathbb{I} - \frac{1}{n}\mathbb{1}\mathbb{1}^\top \in \mathbb{R}^{n\times n}$. Notice that $\sum_{i=1}^n \left\|x^i - \bar{x}\right\|^2 = \mathrm{Tr}(\Sigma)$, where $\mathrm{Tr}(\Sigma)$ is the trace of the matrix

$$\Sigma = \sum_{i=1}^n (x^i - \bar{x})(x^i - \bar{x})^\top = \sum_{i=1}^n x^i(x^i)^\top - \bar{x}\bar{x}^\top = X^\top H X .$$

It is not hard to check that $\mathrm{Tr}(\Sigma) = \mathrm{Tr}(U^\top WHWU)$. Moreover, as $W$ is symmetric and $W\mathbb{1} = \mathbb{1}$ we have $HW = W - \frac{1}{n}\mathbb{1}\mathbb{1}^\top := \overline{W} = WH$. Thus, $WHW = WH^2W = H\overline{W}^2H$ and

$$\mathrm{Tr}(\Sigma) = \mathrm{Tr}(U^\top H\overline{W}^2HU) \le \|\overline{W}^2\|_* \,\mathrm{Tr}(U^\top H^2U) \le \rho^2\mathrm{Tr}(U^\top HU) = \rho^2\sum_{i=1}^n \left\|u^i - \bar{u}\right\|^2 .$$

□

### Proofs for Section 3.4

*Proof of Lemma 3.4.1.* By Taylor expansion we have

$$\frac{f(x{+}hre_j){-}f(x{-}hre_j)}{2h} = \frac{\partial f(x)}{\partial x_j}r + \frac{1}{h}\sum_{2\le m\le \ell, m\ \text{odd}} \frac{(rh)^m}{m!}\frac{\partial^m f(x)}{\partial x_j^m} + \frac{R(hre_j){-}R(-hre_j)}{2h},$$

where $|R(\pm hre_j)| \le L\|hre_j\|^\beta = L|r|^\beta h^\beta$. Using (3.3) it follows that

$$\left|\mathbb{E}[g_j] - \frac{\partial f(x)}{\partial x_j}\right| = \left|\mathbb{E}\left[\frac{f(x+hre_j) - f(x-hre_j)}{2h}K(r)\right] - \frac{\partial f(x)}{\partial x_j}\right| \le L\kappa_\beta h^{\beta-1},$$

which implies the result. □

*Proof of Lemma 3.4.2.* Fix $j \in 1, \dots, d$. Using the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ and the independence between $r$ and $(\xi_j, \xi_j')$ we have

$$\begin{aligned} \mathbb{E}[g_j^2] &= \frac{1}{4h^2}\mathbb{E}\left[(f(x+hre_j) - f(x-hre_i) + \xi_i - \xi_i')^2K^2(r)\right] \\ &\le \frac{3}{4h^2}\mathbb{E}\left[\left((f(x+hre_j) - f(x-hre_j))^2 + 2\sigma^2\right)K^2(r)\right]. \end{aligned} \tag{3.13}$$

The same calculations as in the proof of Lemma 2.4 in Akhavan et al. (2020) yield

$$\big(f(x+hre_j)-f(x-hre_j)\big)^2 \quad\leq\quad 3\left(\frac{\bar{L}^2}{2}\|hre_j\|^4+4\langle\nabla f(x),hre_j\rangle^2\right),$$

Finally, we combine this inequality with (3.13) to obtain

$$\mathbb{E}[g_j^2]\leq\frac{3}{2}\kappa\left(\frac{\sigma^2}{h^2}+\frac{3\bar{L}^2}{4}h^2\right)+9\kappa\mathbb{E}[\langle\nabla f(x),e_i\rangle^2],$$

which immediately implies the lemma. □

## Proofs for Section 3.6

Recall the notation $\Delta(t) = n^{-1}\sum_{i=1}^n\mathbb{E}[\left\|x^i(t)-\bar{x}(t)\right\|^2]$, $\bar{g}(t) = \frac{1}{n}\sum_{i=1}^n g^i(t)$, and $z^i(t) = \mathrm{Proj}_\Theta\Big(x^i(t)-\eta_t g^i(t)\Big)-(x^i(t)-\eta_t g^i(t))$. We also set $\bar{z}(t)=\frac{1}{n}\sum_{i=1}^n z^i(t)$.

*Proof of Lemma 3.6.1.* Set $V(t) = \sum_{i=1}^n\left\|x^i(t)-\bar{x}(t)\right\|^2$. The definition of Algorithm 2 and Lemma 3.9.1 imply:

$$V(t+1)\leq\rho^2\sum_{i=1}^n\left\|x^i(t)-\bar{x}(t)-\eta_t(g^i(t)-\bar{g}(t))+z^i(t)-\bar{z}(t)\right\|^2.$$

The result is immediate if $\rho=0$. Therefore, in rest of the proof we assume that $\rho>0$. We have

$$\begin{aligned}V(t+1)\leq\rho^2\sum_{i=1}^n\Big[V(t)&+\eta_t^2\left\|g^i(t)-\bar{g}(t)\right\|^2+\left\|z^i(t)-\bar{z}(t)\right\|^2 && (3.14)\\ &-2\eta_t\Big\langle x^i(t)-\bar{x}(t),g^i(t)-\bar{g}(t)\Big\rangle && (3.15)\\ &-2\eta_t\Big\langle g^i(t)-\bar{g}(t),z^i(t)-\bar{z}(t)\Big\rangle && (3.16)\\ &+2\Big\langle x^i(t)-\bar{x}(t),z^i(t)-\bar{z}(t)\Big\rangle\Big]. && (3.17)\end{aligned}$$

For any $z\in\mathbb{R}^d$, we have $\sum_{i=1}^n\left\|g^i(t)-\bar{g}(t)\right\|^2\leq\sum_{i=1}^n\left\|g^i(t)-z\right\|^2$, so that

$$\eta_t^2\sum_{i=1}^n\mathbb{E}\big[\left\|g^i(t)-\bar{g}(t)\right\|^2|\mathcal{F}_t\big]\leq\eta_t^2\sum_{i=1}^n\mathbb{E}\big[\left\|g^i(t)\right\|^2|\mathcal{F}_t\big].$$

Next, from the definition of the projection,

$$\begin{aligned}\left\|z^i(t)\right\| &= \left\|\mathrm{Proj}_\Theta\Big(x^i-\eta_t g^i(t)\Big)-(x^i-\eta_t g^i(t))\right\|\\ &\leq\left\|x^i-(x^i-\eta_t g^i(t))\right\|=\eta_t\left\|g^i(t)\right\|. && (3.18)\end{aligned}$$

Therefore, for the term containing $\left\|z^i(t)-\bar{z}(t)\right\|^2$ in (3.14) we obtain

$$\sum_{i=1}^n \mathbb{E}[\left\|z^i(t)-\bar{z}(t)\right\|^2|\mathcal{F}_t] \le \sum_{i=1}^n \mathbb{E}[\left\|z^i(t)\right\|^2|\mathcal{F}_t] \le \eta_t^2 \sum_{i=1}^n \mathbb{E}\Big[\left\|g^i(t)\right\|^2|\mathcal{F}_t\Big].$$

For the expression in (3.15), by decoupling we get

$$-2\eta_t \sum_{i=1}^n \mathbb{E}\Big[\Big\langle x^i(t)-\bar{x}(t), g^i(t)-\bar{g}(t)\Big\rangle|\mathcal{F}_t\Big] \le \lambda V(t) + \frac{\eta_t^2}{\lambda}\sum_{i=1}^n \mathbb{E}\Big[\left\|g^i(t)\right\|^2|\mathcal{F}_t\Big],$$

where $\lambda > 0$ is a value to be chosen later. For the expression in (3.16), we have

$$\begin{aligned}-2\eta_t \sum_{i=1}^n \mathbb{E}\Big[\Big\langle g^i(t)-\bar{g}(t), z^i(t)-\bar{z}(t)\Big\rangle|\mathcal{F}_t\Big] &\le \eta_t^2\sum_{i=1}^n \mathbb{E}\Big[\left\|g^i(t)-\bar{g}(t)\right\|^2|\mathcal{F}_t\Big] + \sum_{i=1}^n \mathbb{E}\Big[\left\|z^i(t)-\bar{z}(t)\right\|^2|\mathcal{F}_t\Big]\\ &\le 2\eta_t^2\sum_{i=1}^n \mathbb{E}\Big[\left\|g^i(t)\right\|^2|\mathcal{F}_t\Big].\end{aligned}$$

Similarly, for the expression in (3.17), using the Cauchy–Schwarz inequality we get

$$\begin{aligned}2\sum_{i=1}^n \mathbb{E}\Big[\Big\langle x^i(t)-\bar{x}(t), z^i(t)-\bar{z}(t)\Big\rangle|\mathcal{F}_t\Big] &\le 2\sum_{i=1}^n \mathbb{E}\Big[\left\|x^i(t)-\bar{x}(t)\right\|\left\|z^i(t)-\bar{z}(t)\right\||\mathcal{F}_t\Big]\\ &\le \lambda V(t) + \frac{1}{\lambda}\sum_{i=1}^n \mathbb{E}\Big[\left\|z^i(t)-\bar{z}(t)\right\|^2|\mathcal{F}_t\Big]\\ &\le \lambda V(t) + \frac{\eta_t^2}{\lambda}\sum_{i=1}^n \mathbb{E}\Big[\left\|g^i(t)\right\|^2|\mathcal{F}_t\Big].\end{aligned}$$

Combining the above inequalities yields

$$\mathbb{E}[V(t+1)|\mathcal{F}_t] \le \rho^2(1+2\lambda)V(t) + \rho^2\Big(4+\frac{2}{\lambda}\Big)\eta_t^2\sum_{i=1}^n \mathbb{E}\Big[\left\|g^i(t)\right\|^2|\mathcal{F}_t\Big]. \tag{3.19}$$

Taking expectations in (3.19) and applying Lemma 3.4.2 we obtain

$$\Delta(t+1) \le \rho^2(1+2\lambda)\Delta(t) + \rho^2\Big(4+\frac{2}{\lambda}\Big)\eta_t^2\Big(9\kappa G^2 + d\Big(\frac{9h_t^2\kappa\bar{L}^2}{8} + \frac{3\kappa\sigma^2}{2h_t^2}\Big)\Big).$$

Choose here $\lambda = \frac{1-\rho}{2\rho}$. Then, using the fact that $\eta_t = \frac{2}{\alpha t}$, $h_t = t^{-\frac{1}{2\beta}}$ we find

$$\Delta(t+1) \le \rho\Delta(t) + \mathcal{A}_1\frac{\rho^2}{1-\rho}\cdot\frac{d}{\alpha^2}t^{-\frac{2\beta-1}{\beta}}, \tag{3.20}$$

where $\mathcal{A}_1 = \frac{144\kappa G^2}{d} + 18\kappa\bar{L}^2 + 24\kappa\sigma^2$. Due to the recursion in (3.20) we have, for any $t \geq 3$,

$$\begin{aligned}\Delta(t+1) &\leq \rho^t\Delta(1) + \mathcal{A}_1\frac{\rho^2}{1-\rho}\cdot\frac{d}{\alpha^2}\sum_{s=1}^{t} s^{-\frac{2\beta-1}{\beta}}\rho^{t-s} \\ &\leq \mathcal{A}_1\frac{\rho^2}{1-\rho}\cdot\frac{d}{\alpha^2}\Big(\frac{1}{\lfloor\frac{t}{2}\rfloor}\sum_{s=1}^{\lfloor\frac{t}{2}\rfloor} s^{-\frac{2\beta-1}{\beta}}\sum_{k=t-\lfloor\frac{t}{2}\rfloor}^{t-1}\rho^k + \frac{1}{\lfloor\frac{t}{2}\rfloor}\sum_{s=\lfloor\frac{t}{2}\rfloor+1}^{t} s^{-\frac{2\beta-1}{\beta}}\sum_{k=0}^{t-\lfloor\frac{t}{2}\rfloor-1}\rho^k\Big), \qquad (3.21)\end{aligned}$$

where $\Delta(1) = 0$ by the choice of initial values and the last inequality uses the fact that if the function $\phi_1(\cdot)$ is monotone decreasing and $\phi_2(\cdot)$ is monotone increasing then

$$\frac{1}{S}\sum_{s=1}^{S}\phi_1(s)\phi_2(s) \leq \left(\frac{1}{S}\sum_{s=1}^{S}\phi_1(s)\right)\left(\frac{1}{S}\sum_{s=1}^{S}\phi_2(s)\right),$$

see, e.g., (Devroye et al., 1996, Theorem A.19). The sums in (3.21) satisfy

$$\sum_{s=1}^{\lfloor\frac{t}{2}\rfloor} s^{-\frac{2\beta-1}{\beta}} \leq 1+\int_1^\infty s^{-\frac{2\beta-1}{\beta}} = \frac{2\beta-1}{\beta-1}, \qquad \sum_{s=\lfloor\frac{t}{2}\rfloor+1}^{t} s^{-\frac{2\beta-1}{\beta}} \leq \frac{t}{2}\left(\frac{t}{2}\right)^{-\frac{2\beta-1}{\beta}} = 2^{\frac{\beta-1}{\beta}}t^{-\frac{\beta-1}{\beta}},$$

$$\sum_{k=0}^{t-\lfloor\frac{t}{2}\rfloor-1}\rho^k \leq \frac{1}{1-\rho}, \qquad \sum_{k=t-\lfloor\frac{t}{2}\rfloor}^{t-1}\rho^k \leq \sum_{k=\lfloor\frac{t}{2}\rfloor}^{t-1}\rho^k \leq t\rho^{\lfloor\frac{t}{2}\rfloor}/2 \leq \frac{8}{\log(1/\rho)t},$$

where the last inequality follows from the fact that $\rho^k \leq \frac{1}{\log(1/\rho)k^2}$ for any positive integer $k$. Plugging the above inequalities in (3.21) gives

$$\begin{aligned}\Delta(t+1) &\leq \mathcal{A}_1\frac{\rho^2}{1-\rho}\frac{d}{\alpha^2}\Big(\frac{24}{\log(1/\rho)t^2}\frac{2\beta-1}{\beta-1} + 3(2^{\frac{\beta-1}{\beta}})\frac{t^{-\frac{2\beta-1}{\beta}}}{1-\rho}\Big) \\ &\leq \mathcal{A}_2\frac{\rho^2}{(1-\rho)^2}\frac{d}{\alpha^2}t^{-\frac{2\beta-1}{\beta}},\end{aligned}$$

where $\mathcal{A}_2 = \left(24\frac{2\beta-1}{\beta-1} + 3(2^{\frac{\beta-1}{\beta}})\right)\mathcal{A}_1$. Therefore, setting $\mathcal{A} := 2\mathcal{A}_2$ we conclude that, for $t \geq 3$,

$$\Delta(t) \leq \mathcal{A}\frac{\rho^2}{(1-\rho)^2}\frac{d}{\alpha^2}t^{-\frac{2\beta-1}{\beta}}.$$

For $t \in \{1, 2\}$ the bound of the lemma holds trivially since $\bar{x}$ and all $x^i$ belong to the compact $\Theta$.

□

*Proof of Theorem 3.6.2.* From the definition of Algorithm 2 and (3.18) we obtain

$$\begin{aligned}\|\bar{x}(t+1)-x\|^2 &= \|\bar{x}(t)-x\|^2 + \|\bar{z}(t)\|^2 + \eta_t^2\,\|\bar{g}(t)\|^2\\ &\quad - 2\eta_t\langle \bar{g}(t), \bar{x}(t)-x\rangle + 2\langle \bar{z}(t), \bar{x}(t)-x\rangle - 2\eta_t\langle \bar{z}(t), \bar{g}(t)\rangle\\ &\le \|\bar{x}(t)-x\|^2 - 2\eta_t\langle \bar{g}(t), \bar{x}(t)-x\rangle + 2\langle \bar{z}(t), \bar{x}(t)-x\rangle + \frac{4\eta_t^2}{n}\sum_{i=1}^{n}\left\|g^i(t)\right\|^2.\end{aligned}$$

It follows that

$$\langle \bar{g}(t), \bar{x}(t)-x\rangle \le \frac{\|\bar{x}(t)-x\|^2 - \|\bar{x}(t+1)-x\|^2}{2\eta_t} + \frac{1}{\eta_t}\langle \bar{z}(t), \bar{x}(t)-x\rangle + \frac{2\eta_t}{n}\sum_{i=1}^{n}\left\|g^i(t)\right\|^2.$$

The strong convexity assumption implies

$$f(\bar{x}(t)) - f(x) \le \langle \nabla f(\bar{x}(t)), \bar{x}(t)-x\rangle - \frac{\alpha}{2}\|\bar{x}(t)-x\|^2.$$

Combining the last two displays and taking conditional expectations from both sides we get

$$\begin{aligned}\mathbb{E}\big[f(\bar{x}(t)) - f(x)|\mathcal{F}_t\big] &\le \left\|\mathbb{E}\big[\bar{g}(t)|\mathcal{F}_t\big] - \nabla f(\bar{x}(t))\right\|\|\bar{x}(t)-x\| + \frac{1}{2\eta_t}\mathbb{E}\big[a_t - a_{t+1}|\mathcal{F}_t\big]\\ &\quad + \frac{2\eta_t}{n}\sum_{i=1}^{n}\mathbb{E}\big[\left\|g^i(t)\right\|^2|\mathcal{F}_t\big] - \frac{\alpha}{2}a_t + \frac{1}{\eta_t}\mathbb{E}\big[\langle \bar{z}(t), \bar{x}(t)-x\rangle|\mathcal{F}_t\big], \qquad (3.22)\end{aligned}$$

where $a_t = \|\bar{x}(t)-x\|^2$.

The first term in right hand side of (3.22) is bounded as follows

$$\begin{aligned}\left\|\mathbb{E}\big[\bar{g}(t)|\mathcal{F}_t\big] - \nabla f(\bar{x}(t))\right\|\|\bar{x}(t)-x\| &\le \Bigg[\left\|\mathbb{E}\big[\bar{g}(t)|\mathcal{F}_t\big] - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x^i(t))\right\|\\ &\quad + \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x^i(t)) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\bar{x}(t))\right\|\Bigg]\|\bar{x}(t)-x\|\\ &\le \kappa_\beta L\sqrt{d}h_t^{\beta-1}\|\bar{x}(t)-x\| + \frac{\bar{L}}{n}\sum_{i=1}^{n}\left\|x^i(t)-\bar{x}(t)\right\|\|\bar{x}(t)-x\|, \qquad (3.23)\end{aligned}$$

where the last inequality is due to Lemma 3.4.1 and Assumption 3.3.4(ii). We now decouple the terms in (3.23) using the fact that $ab \le \frac{a^2}{v} + \frac{vb^2}{4}, \forall a,b \ge 0, v > 0$. Thus, we obtain

$$\kappa_\beta L\sqrt{d}h_t^{\beta-1}\|\bar{x}(t)-x\| \le \frac{(\kappa_\beta L)^2}{\alpha}dh_t^{2(\beta-1)} + \frac{\alpha}{4}\|\bar{x}(t)-x\|^2 \qquad (3.24)$$

and, taking $v = t\alpha(1-\rho)$,

$$\frac{\bar{L}}{n}\sum_{i=1}^{n}\left\|x^i(t)-\bar{x}(t)\right\|\|\bar{x}(t)-x\| \le \frac{\bar{L}t\alpha(1-\rho)}{n}\sum_{i=1}^{n}\left\|x^i(t)-\bar{x}\right\|^2 + \frac{\bar{L}\mathcal{K}^2}{4t\alpha(1-\rho)}. \qquad (3.25)$$

The bound (3.25) brings us to the quantity $\sum_{i=1}^n \left\|x^i(t)-\bar{x}(t)\right\|^2$ that can be controlled in expectation via Lemma 3.6.1. Note that the choice of $v = t\alpha(1-\rho)$ here is motivated by the fact that, once Lemma 3.6.1 is applied (see the end of this proof), it minimizes the final bound in $\rho$ and $\alpha$. We could have kept $v$ in the form $v = v_0 t$ (with an arbitrary parameter $v_0 > 0$) until the application of Lemma 3.6.1 and then optimize over $v_0$. However, we prefer to insert the optimal value $v_0 = \alpha(1-\rho)$ already at this stage.

Combining (3.24) and (3.25) with (3.23) gives

$$\begin{aligned}\left\|\mathbb{E}\left[\bar{g}(t)|\mathcal{F}_t\right]-\nabla f(\bar{x}(t))\right\|\|\bar{x}(t)-x\| \leq & \frac{(\kappa_\beta L)^2}{\alpha} d h_t^{2(\beta-1)}+\frac{\alpha}{4}\|\bar{x}(t)-x\|^2+ \\ & +\frac{\bar{L} t\alpha(1-\rho)}{n} \sum_{i=1}^n\left\|x^i(t)-\bar{x}(t)\right\|^2+\frac{\bar{L}\mathcal{K}^2}{4t\alpha(1-\rho)}. \end{aligned} \tag{3.26}$$

Next, we have

$$\begin{aligned}\frac{1}{\eta_t}\langle\bar{z}(t), \bar{x}(t)-x\rangle &= \frac{1}{n\eta_t} \sum_{i=1}^n\langle z^i(t), \bar{x}(t)-x\rangle \\ &\leq \frac{1}{n\eta_t} \sum_{i=1}^n\langle z^i(t), \bar{x}(t)-\left(x^i(t)-\eta_t g^i(t)\right)\rangle+\langle z^i(t),\left(x^i(t)-\eta_t g^i(t)\right)-x\rangle. \end{aligned} \tag{3.27}$$

Since $\text{Proj}_\Theta(\cdot)$ is the Euclidean projection on the convex set $\Theta$, for any $w\in\mathbb{R}^d, x\in\Theta$ we have $\langle \text{Proj}_\Theta(w)-w, \text{Proj}_\Theta(w)-x\rangle \leq 0$, which implies

$$\langle \text{Proj}_\Theta(w)-w, w-x\rangle = -\left\|\text{Proj}_\Theta(w)-w\right\|^2+\langle \text{Proj}_\Theta(w)-w, \text{Proj}_\Theta(w)-x\rangle \leq 0.$$

Therefore,

$$\langle z^i(t), x^i-\eta_t g^i(t)-x\rangle = \langle \text{Proj}_\Theta(x^i(t)-\eta_t g^i(t))-(x^i(t)-\eta_t g^i(t)), x^i(t)-\eta_t g^i(t)-x\rangle \leq 0.$$

Applying this inequality in (3.27) and using (3.18) we find

$$\begin{aligned}\frac{1}{\eta_t}\langle\bar{z}(t), \bar{x}(t)-x\rangle &\leq \frac{1}{n\eta_t} \sum_{i=1}^n\langle z^i(t),\left(\bar{x}(t)-x^i(t)\right)+\eta_t g^i(t)\rangle \\ &\leq \frac{1}{n\eta_t} \sum_{i=1}^n\left\|z^i(t)\right\|\left\|x^i(t)-\bar{x}(t)\right\|+\frac{1}{n} \sum_{i=1}^n\left\|z^i(t)\right\|\left\|g^i(t)\right\| \\ &\leq \frac{1}{2n\eta_t} \sum_{i=1}^n\left[\frac{\eta_t^2\left\|g^i(t)\right\|^2}{1-\rho}+(1-\rho)\left\|x^i-\bar{x}(t)\right\|^2\right]+\frac{\eta_t}{n} \sum_{i=1}^n\left\|g^i(t)\right\|^2 \\ &\leq \frac{3\eta_t}{2(1-\rho)n} \sum_{i=1}^n\left\|g^i(t)\right\|^2+\frac{1-\rho}{2n\eta_t} \sum_{i=1}^n\left\|x^i(t)-\bar{x}(t)\right\|^2. \end{aligned} \tag{3.28}$$

Inserting (3.28) and (3.26) in (3.22) and using the fact that $\eta_t = \frac{2}{\alpha t}$ we get

$$\begin{aligned}\mathbb{E}[f(\bar{x}(t)) - f(x)|\mathcal{F}_t] \leq &\frac{1}{2\eta_t}\mathbb{E}[a_t - a_{t+1}|\mathcal{F}_t] - \frac{\alpha}{4}a_t \\ &+ \frac{(1+4\bar{L})t\alpha(1-\rho)}{4n}\sum_{i=1}^{n}\left\|x^i - \bar{x}(t)\right\|^2 + \\ &+ \frac{7\eta_t}{2(1-\rho)n}\sum_{i=1}^{n}\mathbb{E}[\left\|g^i(t)\right\|^2|\mathcal{F}_t] + \frac{(\kappa_\beta L)^2}{\alpha}dh_t^{2(\beta-1)} + \frac{\bar{L}\mathcal{K}^2}{4t\alpha(1-\rho)}.\end{aligned}$$

where the last inequality follows from. Taking the expectations, setting $r_t := \mathbb{E}[a_t]$ and applying Lemma 3.4.2 we get

$$\begin{aligned}\mathbb{E}\big[f(\bar{x}(t)) - f(x)\big] \leq &\frac{r_t - r_{t+1}}{2\eta_t} - \frac{\alpha}{4}r_t + \frac{(1+4\bar{L})t\alpha(1-\rho)}{4}\Delta(t) + \\ &+ \frac{7}{\alpha(1-\rho)t}\Big(9\kappa G^2 + d\Big(\frac{9h_t^2\kappa\bar{L}^2}{8} + \frac{3\kappa\sigma^2}{2h_t^2}\Big)\Big) \\ &+ \frac{(\kappa_\beta L)^2}{\alpha}dh_t^{2(\beta-1)} + \frac{\bar{L}\mathcal{K}^2}{4t\alpha(1-\rho)}.\end{aligned} \tag{3.29}$$

Notice that for our choice of $\eta_t = \frac{2}{\alpha t}$ we have

$$\sum_{t=1}^{T_0}\left(\frac{r_t - r_{t+1}}{2\eta_t} - \frac{\alpha}{4}r_t\right) \leq 0.$$

Recall that $h_t = t^{-\frac{1}{2\beta}}$. We can see now that this choice of $h_t$ is the minimizer of the main term depending on $h_t$ on the right hand side of (3.29), which is (up to multiplicative constant) of the order of $h_t^{2(\beta-1)} + \frac{1}{th_t^2}$. By substituting this $h_t$ in (3.29) and summing over $t$ we get

$$\sum_{t=1}^{T_0}\mathbb{E}\big[f(\bar{x}(t)) - f(x)\big] \leq \frac{(1+4\bar{L})\alpha(1-\rho)}{4}\sum_{t=1}^{T_0}t\Delta(t) + \mathcal{B}_1\frac{d}{\alpha(1-\rho)}T_0^{\frac{1}{\beta}} + \frac{\bar{L}\mathcal{K}^2}{4\alpha(1-\rho)}\big(\log(T_0) + 1\big),$$

where $\mathcal{B}_1 = 7\beta\Big(\frac{9\kappa G^2}{d} + (\frac{9\kappa\bar{L}^2}{8} + \frac{3\kappa\sigma^2}{2})\Big) + \beta(\kappa_\beta L)^2$. Finally, using Lemma 3.6.1 we obtain

$$\sum_{t=1}^{T_0}\mathbb{E}\big[f(\bar{x}(t)) - f(x)\big] \leq \mathcal{B}_1\frac{d}{\alpha(1-\rho)}T_0^{\frac{1}{\beta}} + \mathcal{B}_2\frac{\rho^2}{1-\rho}\frac{d}{\alpha}T_0^{\frac{1}{\beta}} + \frac{\mathcal{B}_3}{\alpha(1-\rho)}\big(\log(T_0) + 1\big),$$

where $\mathcal{B}_2 = \frac{\beta(1+4\bar{L})}{4}\mathcal{A}$, and $\mathcal{B}_3 = \bar{L}\mathcal{K}^2$. This proves the first bound of the theorem. The second bound (3.7) follows immediately by the convexity of $f$. □

*Proof of Corollary 3.6.3.* In contrast to the previous proofs, now we have $\eta_t = \frac{4}{\alpha(t+1)}$ rather than $\eta_t = \frac{2}{\alpha t}$.

1°. Inspection of the proof of Lemma 3.6.1 immediately yields that Lemma 3.6.1 remains valid with $\eta_t = \frac{4}{\alpha(t+1)}$ instead of $\eta_t = \frac{2}{\alpha t}$, up to a change in constant $\mathcal{A}$. Thus,

$$\Delta(t) \leq \bar{\mathcal{A}} \left(\frac{\rho}{1-\rho}\right)^2 \frac{d}{\alpha^2} t^{-\frac{2\beta-1}{\beta}}, \tag{3.30}$$

$$\mathbb{E}\big[\|\hat{x}^i(t) - \bar{x}(t)\|^2\big] \leq \bar{\mathcal{A}} n \left(\frac{\rho}{1-\rho}\right)^2 \frac{d}{\alpha^2} t^{-\frac{2\beta-1}{\beta}}, \quad i = 1, \dots, n, \tag{3.31}$$

where $\bar{\mathcal{A}} > 0$ is a constant independent of $t, d, \alpha, n, \rho$.

2°. Next, we show that, up to changes in constants $\mathcal{B}_i$, the bound (3.7) of Theorem 3.6.2 remains valid with $\eta_t = \frac{4}{\alpha(t+1)}$ instead of $\eta_t = \frac{2}{\alpha t}$ if we replace $\hat{x}(T_0)$ in (3.7) by the estimator

$$\hat{x}_\star(T_0) := \frac{2}{T_0(T_0+1)} \sum_{t=1}^{T_0} t\bar{x}(t).$$

Indeed, repeating the proof of Theorem 3.6.2 until (3.29), multiplying both sides of (3.29) by $t$, summing up from $t = 1$ to $T_0$ and using the fact that

$$\sum_{t=1}^{T_0} \left(\frac{t(r_t - r_{t+1})}{2\eta_t} - \frac{\alpha}{4} t r_t\right) \leq 0 \qquad \text{if } \eta_t = \frac{4}{\alpha(t+1)},$$

we find that, for all $x \in \Theta$,

$$\sum_{t=1}^{T_0} t\, \mathbb{E}\big[f(\bar{x}(t)) - f(x)\big] \leq \frac{(1+4\bar{L})\alpha(1-\rho)}{4} \sum_{t=1}^{T_0} t^2 \Delta(t) + \bar{\mathcal{B}}_1 \frac{d}{\alpha} T_0^{1+\frac{1}{\beta}} + \frac{\bar{L}\mathcal{K}^2}{4\alpha(1-\rho)} T_0,$$

where $\bar{\mathcal{B}}_1$ is a positive constant independent of $T_0, d, \alpha, n, \rho$. Using (3.30) we get, for all $x \in \Theta$,

$$\frac{2}{T_0(T_0+1)} \sum_{t=1}^{T_0} t\, \mathbb{E}\big[f(\bar{x}(t)) - f(x)\big] \leq \bar{\mathcal{B}}_2 \frac{d}{\alpha(1-\rho)} T_0^{-1+\frac{1}{\beta}},$$

where $\bar{\mathcal{B}}_2$ is a positive constant independent of $T_0, d, \alpha, n, \rho$. In view of the convexity of $f$, it follows that

$$\mathbb{E}\big[f(\hat{x}_\star(T_0)) - f(x^*)\big] \leq \bar{\mathcal{B}}_2 \frac{d}{\alpha(1-\rho)} T_0^{-1+\frac{1}{\beta}}.$$

As $f$ is strongly convex we also have

$$\mathbb{E}\big[\|\hat{x}_\star(T_0) - x^*\|^2\big] \leq 2\bar{\mathcal{B}}_2 \frac{d}{\alpha^2(1-\rho)} T_0^{-1+\frac{1}{\beta}}. \tag{3.32}$$

On the other hand, convexity of function $\|\cdot\|^2$ implies that

$$\|\hat{x}^i(T_0) - \hat{x}_\star(T_0)\|^2 = \left\|\frac{2}{T_0(T_0+1)}\sum_{t=1}^{T_0} t(x^i(t) - \bar{x}(t))\right\|^2$$

$$\leq \frac{2}{T_0(T_0+1)}\sum_{t=1}^{T_0} t\|x^i(t) - \bar{x}(t)\|^2. \tag{3.33}$$

Combining (3.31) and (3.33) we obtain

$$\mathbb{E}\left[\|\hat{x}^i(T_0) - \hat{x}_\star(T_0)\|^2\right] \leq \bar{\mathcal{C}}n\left(\frac{\rho}{1-\rho}\right)^2 \frac{d}{\alpha^2}T_0^{-\frac{2\beta-1}{\beta}}, \tag{3.34}$$

where $\bar{\mathcal{C}} > 0$ is a constant independent of $T_0, d, \alpha, n, \rho$. The desired result now follows from (3.32), (3.34) and the fact that $\|\hat{x}^i(T_0) - x^*\|$ is trivially bounded by the diameter of $\Theta$.

□

## Proofs for Section 3.7

We first restate the following three lemmas from Akhavan et al. (2020).

**Lemma 3.9.2.** *Let for $\beta = 2$, Assumptions 3.3.4 and 3.7.1 hold. Let $\bar{g}(t)$ be the average of gradient estimators for $n$ agents defined each by (3.11), and $h = h_t$. If $\max_{x\in\Theta}\|\nabla f_i(x)\| \leq G$, for $1 \leq i \leq n$, then*

$$\mathbb{E}[\|\bar{g}(t)\|^2] \leq 9\kappa\Big(G^2 d + \frac{L^2d^2h_t^2}{2}\Big) + \frac{3\kappa d^2\sigma^2}{2h_t^2}.$$

Introduce the notation

$$\hat{f}_t(x) = \mathbb{E}f(x + h_t\tilde{\zeta}), \qquad \forall x \in \mathbb{R}^d,$$

and

$$\hat{f}_t^i(x) = \mathbb{E}f_i(x + h_t\tilde{\zeta}), \qquad \forall x \in \mathbb{R}^d.$$

**Lemma 3.9.3.** *Suppose $f_i$ is differentiable. For the conditional expectation given $\mathcal{F}_t$, we have*

$$\mathbb{E}[g^i(t)|\mathcal{F}_t] = \nabla\hat{f}_t^i(x^i(t)).$$

**Lemma 3.9.4.** *If $f$ is $\alpha$-strongly convex then $\hat{f}_t$ is $\alpha$-strongly convex. If $f \in \mathcal{F}_2(L)$, for any $x \in \mathbb{R}^d$ and $h_t > 0$, we have*

$$|\hat{f}_t(x) - f(x)| \leq Lh_t^2,$$

*and*

$$|\mathbb{E}f(x \pm h_t\zeta_t) - f(x)| \leq Lh_t^2.$$

**Lemma 3.9.5.** *Let Assumptions 3.2.1, 3.3.4, and 3.7.1 hold with $\beta = 2$. Let $\Theta$ be a convex compact subset of $\mathbb{R}^d$, and assume that diam$(\Theta) \leq \mathcal{K}$. Assume that $\max_{x\in\Theta} \|\nabla f_i(x)\| \leq G$, for $1 \leq i \leq n$. Let the updates $x^i(t), \bar{x}(t)$ be defined by Algorithm 2, in which the gradient estimator for $i$-th agent is defined by (3.11), and $\eta_t = \frac{1}{\alpha t}$, $h_t = \left(\frac{3d^2\sigma^2}{2L\alpha t + 9L^2d^2}\right)^{1/4}$. Then*

$$\Delta(t) \leq \Big(\frac{\rho}{1-\rho}\Big)^2\Big(\mathcal{A}'_1\frac{d}{\alpha^{3/2}}t^{-\frac{3}{2}} + \mathcal{A}'_2\frac{d^2}{\alpha^2}t^{-2}\Big),$$

*where $\mathcal{A}'_1$ and $\mathcal{A}'_2$ are positive constants independent of $T, d, \alpha, n, \rho$.*

*Proof.* Similarly to Lemma 3.6.1 we obtain

$$\mathbb{E}[V(t+1)|\mathcal{F}_t] \leq \rho^2(1+2\lambda)V(t) + \rho^2(4+\frac{2}{\lambda})\eta_t^2\sum_{i=1}^{n}\mathbb{E}[\left\|g^i(t)\right\|^2|\mathcal{F}_t].$$

Choosing $\lambda = \frac{1-\rho}{2\rho}$ and using Lemma 3.9.2 we get

$$\mathbb{E}[V(t+1)|\mathcal{F}_t] \leq \rho V(t) + \frac{4\rho^2}{1-\rho}\eta_t^2\Big(9(G^2d + \frac{L^2d^2h_t^2}{2}) + \frac{3d^2\sigma^2}{2h_t^2}\Big).$$

Taking here the expectations and setting $\eta_t = \frac{1}{\alpha t}$ and $h_t = \left(\frac{3d^2\sigma^2}{2L\alpha t + 9L^2d^2}\right)^{1/4}$ yields

$$\Delta(t+1) \leq \rho\Delta(t) + \frac{\rho^2}{1-\rho}\Big(\mathcal{A}'_3\frac{d}{\alpha^{3/2}t^{3/2}} + \mathcal{A}'_4\frac{d^2}{\alpha^2t^2}\Big)$$

with $\mathcal{A}'_3 = 2\sqrt{6}L\sigma$, and $\mathcal{A}'_4 = 12\sqrt{3}L\sigma + \frac{36G^2}{d}$. On the other hand, by recursion we have

$$\Delta(t+1) \leq \rho^t\Delta(1) + \frac{\rho^2}{1-\rho}\frac{d}{\alpha^{3/2}}\Bigg(\mathcal{A}'_3\sum_{s=1}^{t}s^{-\frac{3}{2}}\rho^{t-s} + \mathcal{A}'_4\frac{d}{\alpha^{1/2}} + \sum_{s=1}^{t}s^{-2}\rho^{t-s}\Bigg).$$

Here $\Delta(1) = 0$ due to the initialization. The sums on right hand side can be estimated by using an argument, which is quite analogous to what was done in the proof of Lemma 3.6.1, after equation (3.21), leading to the result of the lemma. □

**Lemma 3.9.6.** *Let the assumptions of Lemma 3.9.5 hold and let $f$ be an $\alpha$-strongly convex function. Then*

$$\mathbb{E}[\|\bar{x}(t) - x^*\|^2] \leq \frac{\mathcal{C}}{1-\rho}\left(\frac{d}{t^{1/2}\alpha^{3/2}} + \frac{d^2}{t\alpha^2}\right),$$

*where $\mathcal{C} > 0$ is a constant independent of $T, d, \alpha, n, \rho$.*

*Proof.* First note that due to the strong convexity assumption we have

$$\|\bar{x}(1) - x^*\|^2 \leq \frac{G^2}{\alpha^2}.$$

Therefore, for $t = 1$ the result holds. For $t \geq 2$, by the definition of the algorithm we have

$$\begin{aligned}\|\bar{x}(t+1) - x^*\|^2 \leq \|\bar{x}(t) - x^*\|^2 + \eta_t^2 \|\bar{g}(t)\|^2 + \|\bar{z}(t)\|^2 - 2\eta_t\langle\bar{g}(t), \bar{z}(t)\rangle - \\ - 2\eta_t\langle\bar{g}(t), \bar{x}(t) - x^*\rangle + 2\langle\bar{x}(t) - x^*, \bar{z}(t)\rangle.\end{aligned}$$

Taking conditional expectations we get

$$\mathbb{E}[a_{t+1}|\mathcal{F}_t] \leq a_t + \frac{2\eta_t^2}{n}\sum_{i=1}^n \mathbb{E}[\left\|g^i(t)\right\|^2|\mathcal{F}_t] - 2\eta_t\mathbb{E}[\langle\bar{g}(t), \bar{z}(t)\rangle|\mathcal{F}_t] - \tag{3.35}$$

$$- 2\eta_t\mathbb{E}[\langle\bar{g}(t), \bar{x}(t) - x^*\rangle|\mathcal{F}_t] + 2\mathbb{E}[\langle\bar{x}(t) - x^*, \bar{z}(t)\rangle|\mathcal{F}_t], \tag{3.36}$$

where we used the fact that $\left\|z^i(t)\right\| \leq \eta_t \left\|g^i(t)\right\|$ for $1 \leq i \leq n$.

For the term $-2\eta_t\mathbb{E}[\langle\bar{g}(t), \bar{x}(t) - x^*\rangle|\mathcal{F}_t]$ in (3.35), we have

$$-2\eta_t\mathbb{E}[\langle\bar{g}(t), \bar{x}(t) - x^*\rangle|\mathcal{F}_t] \leq -\frac{2\eta_t}{n}\sum_{i=1}^n \Big(\mathbb{E}[\langle g^i(t) - \nabla\hat{f}_t^i(x^i(t)), \bar{x}(t) - x^*\rangle|\mathcal{F}_t] + \tag{3.37}$$

$$+ \langle\nabla\hat{f}_t^i(x^i(t)) - \nabla\hat{f}_t^i(\bar{x}(t)), \bar{x}(t) - x^*\rangle + \tag{3.38}$$

$$+ \langle\nabla\hat{f}_t(\bar{x}(t)), \bar{x}(t) - x^*\rangle\Big) \tag{3.39}$$

For the term in (3.37), by Lemma 3.9.3 we have

$$-\frac{2\eta_t}{n}\sum_{i=1}^n \mathbb{E}[\langle g^i(t) - \nabla\hat{f}_t^i(x^i(t)), \bar{x}(t) - x^*\rangle|\mathcal{F}_t] = 0.$$

For the term in (3.38), decoupling yields

$$-\frac{2\eta_t}{n}\sum_{i=1}^n\langle\nabla\hat{f}_t^i(x^i(t)) - \nabla\hat{f}_t^i(\bar{x}(t)), \bar{x}(t) - x^*\rangle \leq \frac{\eta_t t\alpha}{n}(1-\rho)V(t) + \frac{\bar{L}^2\eta_t}{t\alpha}\frac{1}{1-\rho}a_t.$$

Next, we use the strong convexity (cf. Lemma 3.9.4) to handle (3.39):

$$-2\eta_t\langle\nabla\hat{f}_t(\bar{x}(t)), \bar{x}(t) - x^*\rangle \leq -2\eta_t\alpha a_t.$$

Finally, for the term containing $2\langle\bar{x}(t) - x^*, \bar{z}(t)\rangle$ in (3.36) we obtain similarly to (3.28) that

$$2\mathbb{E}[\langle\bar{x}(t) - x^*, \bar{z}(t)\rangle|\mathcal{F}_t] \leq \frac{3\eta_t^2}{(1-\rho)n}\sum_{i=1}^n \mathbb{E}[\left\|g^i(t)\right\|^2|\mathcal{F}_t] + \frac{1-\rho}{n}V(t).$$

Combining the above inequalities yields

$$\mathbb{E}[a_{t+1}|\mathcal{F}_t] \leq (1-2\eta_t\alpha)a_t + \frac{2\eta_t^2}{n}\sum_{i=1}^n \mathbb{E}[\|\bar{g}(t)\|^2|\mathcal{F}_t] - 2\eta_t\mathbb{E}[\langle\bar{g}(t), \bar{z}(t)\rangle|\mathcal{F}_t] + \frac{\eta_t\bar{L}^2\mathcal{K}^2}{t\alpha(1-\rho)} + \\ + \frac{\eta_t t\alpha+1}{n}(1-\rho)V(t) + \frac{3\eta_t^2}{(1-\rho)n}\sum_{i=1}^n \mathbb{E}[\|g^i(t)\|^2|\mathcal{F}_t].$$

Now, recalling that $\eta_t = \frac{1}{t\alpha}$, $h_t = \left(\frac{3d^2\sigma^2}{2L\alpha t+9L^2d^2}\right)^{1/4}$, taking the expectations and applying Lemma 3.9.2 we find

$$r_{t+1} \leq \left(1-\frac{2}{t}\right)r_t + 2(1-\rho)\Delta(t) + \frac{C}{(1-\rho)}\left(\frac{d}{t^{3/2}\alpha^{3/2}} + \frac{d^2}{t^2\alpha^2}\right), \tag{3.40}$$

where $r_t = \mathbb{E}[a_t]$, and $C > 0$ is a constant independent of $T, d, \alpha, n, \rho$. Using Lemma 3.9.5 to bound $\Delta(t)$ in (3.40) we get

$$r_{t+1} \leq \left(1-\frac{2}{t}\right)r_t + \frac{C'}{(1-\rho)}\left(\frac{d}{t^{3/2}\alpha^{3/2}} + \frac{d^2}{t^2\alpha^2}\right),$$

where $C' > 0$ is a constant independent of $T, d, \alpha, n, \rho$. The desired result follows from this recursion by applying (Akhavan et al., 2020, Lemma D.1). □

*Proof of Theorem 3.7.2.* Fix $x \in \Theta$. Due to the $\alpha$-strong convexity of $\hat{f}_t$, we have

$$\hat{f}_t(\bar{x}(t)) - \hat{f}_t(x^*) \leq \langle\nabla\hat{f}_t(\bar{x}(t)), \bar{x}(t)-x^*\rangle - \frac{\alpha}{2}\|\bar{x}(t)-x^*\|^2.$$

Thus, by Lemma 3.9.4 we get

$$f(\bar{x}(t)) - f(x^*) \leq 2Lh_t^2 + \langle\nabla\hat{f}_t(\bar{x}(t)), \bar{x}(t)-x^*\rangle - \frac{\alpha}{2}\|\bar{x}(t)-x^*\|^2.$$

Let $a_t = \|\bar{x}(t)-x^*\|^2$. Taking conditional expectations and applying Lemma 3.9.3 we obtain

$$\begin{aligned}\mathbb{E}[f(\bar{x}(t)) - f(x^*)|\mathcal{F}_t] &\leq 2Lh_t^2 + \frac{1}{n}\sum_{i=1}^n\langle\nabla\hat{f}_t^i(\bar{x}(t)) - \nabla\hat{f}_t^i(x^i(t)), \bar{x}(t)-x^*\rangle - \frac{\alpha}{2}a_t \\ &\quad + \mathbb{E}[\langle\bar{g}(t), \bar{x}(t)-x^*\rangle|\mathcal{F}_t] \\ &\leq 2Lh_t^2 + \frac{1}{n}\sum_{i=1}^n\mathbb{E}[\langle\nabla\hat{f}_t^i(\bar{x}(t)) - \nabla\hat{f}_t^i(x^i(t)), \bar{x}(t)-x^*\rangle|\mathcal{F}_t] \\ &\quad - \frac{\alpha}{2}a_t + \frac{a_t - \mathbb{E}[a_{t+1}|\mathcal{F}_t]}{2\eta_t} \\ &\quad + \frac{1}{\eta_t}\mathbb{E}[\langle\bar{z}(t), \bar{x}(t)-x^*\rangle|\mathcal{F}_t] + \frac{2\eta_t}{n}\sum_{i=1}^n\mathbb{E}[\|g^i(t)\|^2|\mathcal{F}_t],\end{aligned}$$

where the last inequality uses the definition of the algorithm. Now, by decoupling we find

$$\frac{1}{n}\sum_{i=1}^{n}\langle\nabla\hat{f}_t^i(\bar{x}(t))-\nabla\hat{f}_t^i(x^i(t)),\bar{x}(t)-x^*\rangle\leq\frac{t\alpha}{2n}(1-\rho)V(t)+\frac{1}{2(1-\rho)}\frac{\bar{L}^2}{t\alpha}\mathcal{K}^2,$$

while similarly to (3.28) we also have

$$\frac{1}{\eta_t}\mathbb{E}[\langle\bar{z}(t),\bar{x}(t)-x^*\rangle|\mathcal{F}_t]\leq\frac{1}{1-\rho}\frac{3\eta_t}{2n}\sum_{i=1}^{n}\mathbb{E}[\left\|g^i(t)\right\|^2|\mathcal{F}_t]+(1-\rho)\frac{1}{2n\eta_t}V(t).$$

Combining the above inequalities and applying Lemma 3.9.2 yields

$$\begin{aligned}\mathbb{E}[f(\bar{x}(t))-f(x^*)|\mathcal{F}_t]\leq\Big(\frac{1}{\eta_t}+t\alpha\Big)\frac{1-\rho}{2n}V(t)+\frac{1}{2(1-\rho)}\frac{\bar{L}^2}{t\alpha}\mathcal{K}^2-\frac{\alpha}{2}a_t+\frac{a_t-\mathbb{E}[a_{t+1}|\mathcal{F}_t]}{2\eta_t}+\\+2Lh_t^2+\left(2+\frac{3}{2(1-\rho)}\right)\frac{\eta_t}{n}\sum_{i=1}^{n}\mathbb{E}[\left\|g^i(t)\right\|^2|\mathcal{F}_t].\end{aligned}\tag{3.41}$$

Let $r_t=\mathbb{E}[a_t]$. Using the fact that $\eta_t=\frac{1}{\alpha t}$, $h_t=\left(\frac{3d^2\sigma^2}{2L\alpha t+9L^2d^2}\right)^{1/4}$, taking the expectations in (3.41) and applying Lemma 3.9.2 we find

$$\mathbb{E}[f(\bar{x}(t))-f(x^*)]\leq t\alpha\Big(\frac{r_t-r_{t+1}}{2}\Big)-\frac{\alpha}{2}r_t+(1-\rho)\alpha t\Delta(t)+\frac{C_1}{1-\rho}\Big(\frac{d}{\sqrt{\alpha t}}+\frac{d^2}{\alpha t}\Big),$$

where $C_1>0$ is a constant independent of $T,d,\alpha,n,\rho$. Summing up both sides over $t$ gives

$$\sum_{t=\lfloor\frac{T}{2}\rfloor+1}^{T}\mathbb{E}[f(\bar{x}(t))-f(x^*)]\leq r_{\lfloor\frac{T}{2}\rfloor+1}\frac{\lfloor\frac{T}{2}\rfloor\alpha}{2}+(1-\rho)\alpha\sum_{t=\lfloor\frac{T}{2}\rfloor+1}^{T}t\Delta(t)+\frac{C_2}{1-\rho}\Big(\frac{d\sqrt{T}}{\sqrt{\alpha}}+\frac{d^2}{\alpha}\Big)$$

where $C_2>0$ is a constant independent of $T,d,\alpha,n,\rho$. We now apply Lemma 3.9.5 to bound $\Delta(t)$ and Lemma 3.9.6 to bound $r_{\lfloor\frac{T}{2}\rfloor+1}$. It follows that

$$\sum_{t=\lfloor\frac{T}{2}\rfloor+1}^{T}\mathbb{E}[f(\bar{x}(t))-f(x^*)]\leq\frac{C_3}{1-\rho}\Big(\frac{d\sqrt{T}}{\sqrt{\alpha}}+\frac{d^2}{\alpha}\Big),$$

where $C_3>0$ is a constant independent of $T,d,\alpha,n,\rho$. The desired bound for $\mathbb{E}[f(\tilde{x}(T))-f(x^*)]$ follows from this inequality by the convexity of $f$.

□

### Numerical Experiments

In this section we present a numerical comparison between the proposed method and the zero-order method in Akhavan et al. (2020) based on 2-point gradient estimator. Since the goal is to study the effect of the new gradient estimator, we consider the standard (undistributed) setting.

We wish to minimize the following function $f : \mathbb{R}^d \to \mathbb{R}$,

$$f(x) = \frac{\alpha}{2} x^\top A x + L h^3 \sum_{i=1}^{d} \psi(h^{-1} x_i), \tag{3.42}$$

where $\alpha, L, h$ are positive parameters, $A$ is a positive definite matrix in $\mathbb{R}^{d\times d}$ with smallest eigenvalue equal to $1$, and $\psi(x) = \int_{-\infty}^{x}\int_{-\infty}^{z} \phi(t)dtdz$, with

$$\phi(x) = \begin{cases} 0 & \text{if } x < -a \\ \frac{2}{a}x + 2 & \text{if } -a \le x < -\frac{a}{2} \\ -\frac{2}{a}x & \text{if } -\frac{a}{2} \le x \le \frac{a}{2} \\ \frac{2}{a}x - 2 & \text{if } \frac{a}{2} \le x \le a \\ 0 & \text{if } a < x, \end{cases}$$

where $a > 0$. A direct computation gives that

$$\psi(x) = \begin{cases} 0 & \text{if } x < -a \\ \frac{x^3}{3a} + ax^2 + ax + \frac{a^2}{3} & \text{if } -a \le x < -\frac{a}{2} \\ -\frac{x^3}{3a} + \frac{a}{2}x + \frac{a^2}{4} & \text{if } -\frac{a}{2} \le x \le \frac{a}{2} \\ \frac{x^3}{3a} - ax^2 + ax + \frac{a^2}{6} & \text{if } \frac{a}{2} \le x \le a \\ \frac{a^2}{2} & \text{if } a < x. \end{cases}$$

Let $\Theta = \{x \in \mathbb{R}^d : \|x\| \le 1, \text{ and } x_i \le 0, \text{ for } 1 \le i \le d\}$. Since for any $x \in \Theta$, $\phi(x) \ge 0$, then $\psi$ is convex on $\Theta$, which implies $\alpha$-strong convexity of $f$ on $\Theta$. Also, the second derivative of $Lh^3\psi(h^{-1}x)$ is Lipschitz continuous with Lipschitz constant equal to $\frac{2L}{a}$. Therefore $f$ is $\beta$-Hölder with $\beta = 3$. We choose the kernel function, $K : [-1, 1] \to \mathbb{R}$, such that $K(x) = \frac{15}{8}x(5 - 7x^3)$. For each iteration $t$, we fix $h_t = t^{-\frac{1}{6}}$, and $\eta_t = \frac{2}{\alpha t}$. Function evaluations at a fixed point $x \in \mathbb{R}^d$ are obtained in the form $f(x) + \zeta$ where $\zeta$ is a random variable uniformly distributed in $[-5, 5]$.

In this implementation we assign $\alpha = 2$, $h = 10^{-3}$, $L = 10^{7.5}$, $a = 10$. We also let $A = B + \mathbb{I}$, where $B$ is a randomly generated sparse positive definite matrix in $\mathbb{R}^{d\times d}$ and $\mathbb{I}$ is the $d$-dimensional identity matrix. For the initialization, we generate a $d$-dimensional Gaussian random variable and project it on $\Theta$.

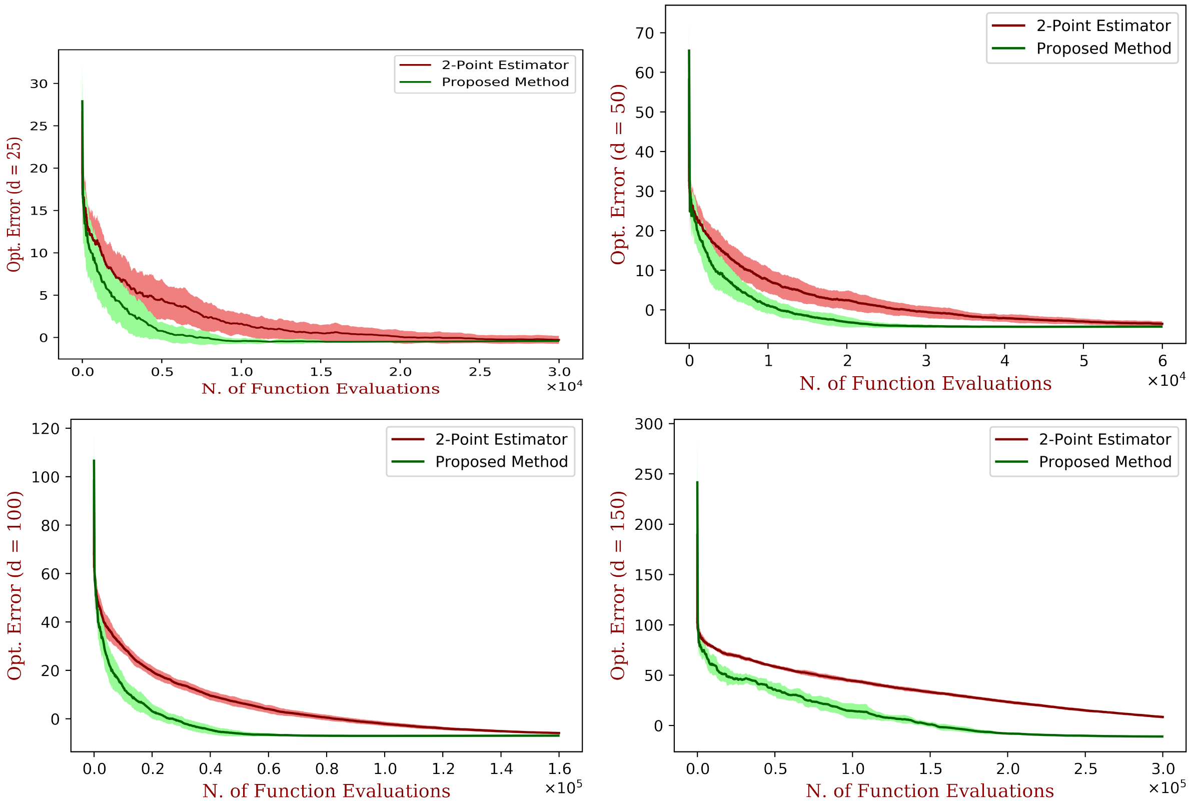

Figure 3.1: Optimization error vs. number of function evaluations for the 2-Point Estimator in Akhavan et al. (2020) and our method, run on function (3.42) for different number of variables ($d = 25, 50, 100, 150$ clockwise from top-left).

The design of $f$ in (3.42) is inspired by the function that has been used in the proof of the lower bound in Akhavan et al. (2020). It is a quadratic function plus the perturbation $Lh^3 \sum_{i=1}^{d} \psi(h^{-1}x_i)$, which adds difficulty to estimation of the minimizer. We have chosen this worst case function to provide a comparison between two algorithms in a long run and growing dimension. In Figure 3.1 we display the average optimization error of the method proposed in this paper and that of the 2-Point estimator from Akhavan et al. (2020) versus the total number of function evaluations, for different dimensions $d$. This result is averaged over $40$ trials, corresponding to different random initialization, noisy function evaluations and randomization in the optimization procedures. We would like to emphasize that both methods are considered with the same budget of function evaluations, which means that the number of iterations for the two algorithms differ. Thus, if $T$ is the total number of function evaluations, the 2-point estimator makes $T/2$ iterations, while the proposed method makes only $T/(2d)$ iterations.

## Chapter 4

# A gradient estimator via L1-randomization for online zero-order optimization with two point feedback

This chapter studies online zero-order optimization of convex and Lipschitz functions. We present a novel gradient estimator based on two function evaluations and randomization on the $\ell_1$-sphere. Considering different geometries of feasible sets and Lipschitz assumptions we analyse online dual averaging algorithm with our estimator in place of the usual gradient. We consider two types of assumptions on the noise of the zero-order oracle: canceling noise and adversarial noise. We provide an anytime and completely data-driven algorithm, which is adaptive to all parameters of the problem. In the case of canceling noise that was previously studied in the literature, our guarantees are either comparable or better than state-of-the-art bounds obtained by Duchi et al. (2015) and Shamir (2017) for non-adaptive algorithms. Our

analysis is based on deriving a new weighted Poincaré type inequality for the uniform measure on the $\ell_1$-sphere with explicit constants, which may be of independent interest.

## 4.1 Introduction

In this work we study the problem of convex online zero-order optimization with two-point feedback, in which adversary fixes a sequence $f_1, f_2, \ldots : \mathbb{R}^d \to \mathbb{R}$ of convex functions and the goal of the learner is to minimize the cumulative regret with respect to the best action in a prescribed convex set $\Theta \subseteq \mathbb{R}^d$. This problem has received significant attention in the context of continuous bandits and online optimization (see e.g., Agarwal et al., 2010; Akhavan et al., 2020; Bubeck and Cesa-Bianchi, 2012; Bubeck et al., 2017; Flaxman et al., 2005; Gasnikov et al., 2016; Lattimore and Gyorgy, 2021; Novitskii and Gasnikov, 2021; Saha and Tewari, 2011; Shamir, 2017, and references therein).

We consider the following protocol: at each round $t = 1, 2, \ldots$ the algorithm chooses $\mathbf{x}'_t, \mathbf{x}''_t \in \mathbb{R}^d$ (that can be queried outside of $\Theta$) and the adversary reveals

$$f_t(\mathbf{x}'_t) + \xi'_t \qquad \text{and} \qquad f_t(\mathbf{x}''_t) + \xi''_t \enspace ,$$

where $\xi'_t, \xi''_t \in \mathbb{R}$ are the noise variables (random or not) to be specified. Based on the above information and the previous rounds, the learner outputs $\mathbf{x}_t \in \Theta$ and suffers loss $f_t(\mathbf{x}_t)$. The goal of the learner is to minimize the cumulative regret

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \Theta} \sum_{t=1}^{T} f_t(\mathbf{x}) \enspace .$$

At the core of our approach is a novel zero-order gradient estimator based on two function evaluations outlined in Algorithm 4. A key novelty of our estimator is that it employs a randomization step over the $\ell_1$ sphere. This is in contrast to most of the prior work (see e.g., Agarwal et al., 2010; Akhavan et al., 2020, 2021; Bach and Perchet, 2016; Duchi et al., 2015; Flaxman et al., 2005; Gasnikov et al., 2017; Nemirovsky and Yudin, 1983; Polyak and Tsybakov, 1990; Shamir, 2013) that was employing $\ell_2$ or $\ell_\infty$ type randomizations to define $\mathbf{x}'_t, \mathbf{x}''_t$. We use the proposed estimator within an online dual averaging procedure to tackle the zero-order online convex optimization problem, matching or improving the state-of-the-art results. Duchi et al. (2015) and Shamir (2017) have studied instances of the above problem under the assumption that $\xi'_t = \xi''_t$, which we will further refer to as canceling noise assumption. Specifically, Duchi et al. (2015) considered the stochastic optimization framework where $f_t = f$, for every $t$, and obtained bounds on the optimization error rather than on cumulative regret, while Shamir (2017) analyzed the case $\xi'_t = \xi''_t = 0$. The results in Duchi et al. (2015); Shamir (2017) are obtained for the objective functions that are Lipschitz with respect to the $\ell_q$-norm for $q = 1$ and $q = 2$, although, with extra derivations it is possible to extend the above mentioned results beyond such cases. The proposed method allows us to improve upon these results in several aspects.

**Contributions.** The contributions of the present paper can be summarized as follows. **1)** We present a new randomized zero-order gradient estimator and study its statistical properties, both under canceling noise and under adversarial noise (see Lemma 4.6.1 and Lemma 4.6.5); **2)** In the canceling noise case ($\xi'_t = \xi''_t$) in Theorem 4.4.1 we show that dual averaging based on our gradient estimator either improves or matches the state-of-the-art bounds Duchi et al. (2015); Shamir (2017). We derive the results for Lipschitz functions with respect to all $\ell_q$-norms, $q \in [1, \infty]$. In particular, when $q = 1$ and $\Theta$ is the probability simplex, our bound is better by a $\sqrt{\log(d)}$ factor than that of Duchi et al. (2015); Shamir (2017); **3)** We propose a completely data-driven and anytime version of the algorithm, which is adaptive to all parameters of the problem. We show that it achieves analogous performance as the non-adaptive algorithm in the case of canceling noise and only slightly worse performance under adversarial noise. To the best of our knowledge, no adaptive algorithms were developed for zero-order online problems in our setting so far; **4)** As a key element of our analysis, we derive in Lemma 4.6.3 a weighted Poincaré type inequality (following the terminology of Bobkov and Ledoux, 2009) with explicit constants for the uniform measure on $\ell_1$-sphere. This result may be of independent interest.

**Notation.** Throughout the paper we use the following notation. We denote by $\|\cdot\|_p$ the $\ell_p$-norm in $\mathbb{R}^d$. For any $\mathbf{x} \in \mathbb{R}^d$ we denote by $\mathbf{x} \mapsto \operatorname{sign}(\mathbf{x})$ the component-wise sign function (defined at $0$ as $1$). We let $\langle \cdot, \cdot \rangle$ be the standard inner product in $\mathbb{R}^d$. For $p \in [1, \infty]$ we introduce the open $\ell_p$-ball and $\ell_p$-sphere respectively as

$$B^d_p \triangleq \left\{\mathbf{x} \in \mathbb{R}^d \,:\, \|\mathbf{x}\|_p < 1\right\} \qquad \text{and} \qquad \partial B^d_p \triangleq \left\{\mathbf{x} \in \mathbb{R}^d \,:\, \|\mathbf{x}\|_p = 1\right\} \ .$$

For two $a, b \in \mathbb{R}$, we denote by $a \wedge b$ (*resp.* $a \vee b$) the minimum (*resp.* the maximum) between $a$ and $b$. We denote by $\Gamma : (0, \infty) \to \mathbb{R}$, the gamma function. In what follows, $\log$ always stands for the natural logarithm and $e$ is Euler's number.

## 4.2 The algorithm

Let $\Theta$ be a closed convex subset of $\mathbb{R}^d$ and let $V : \Theta \to \mathbb{R}$ be a convex function. The procedure that we propose in this paper is summarized in Algorithm 4.

**Intuition behind the gradient estimate.** The form of gradient estimator $\mathbf{g}_t$ in Algorithm 1 is explained by Stokes' theorem (see Theorem 4.9.1 in Section 4.9 and the discussion that follows). Stokes' theorem provides a connection between the gradient of a function $f$ (first order information) and $f$ itself (zero order information). Under some regularity conditions, it

**Algorithm 4** Zero-Order $\ell_1$-randomized online dual averaging

```
**Requires** Convex function $V(\cdot)$, step size $\eta_1 > 0$, and parameters $h_t$, for $t = 1, 2, \dots,$
**Initialization** Generate independently vectors $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \dots,$ uniformly distributed on $\partial B_1^d$, set $\mathbf{z}_1 = \mathbf{0}$
**For** $t = 1, \dots,$ **do**
    1. $\mathbf{x}_t = \arg\max_{\mathbf{x}\in\Theta}\{\eta_t\langle \mathbf{z}_t, \mathbf{x}\rangle - V(\mathbf{x})\}$
    2. $y_t' = f_t(\mathbf{x}_t + h_t\boldsymbol{\zeta}_t) + \xi'$ and $y_t'' = f_t(\mathbf{x}_t + h_t\boldsymbol{\zeta}_t) + \xi''$
    3. $\mathbf{g}_t = \frac{d}{h_t}\left(y_t' - y_t''\right)\operatorname{sign}\left(\boldsymbol{\zeta}_t\right)$
    4. $\mathbf{z}_{t+1} = \mathbf{z}_t - \mathbf{g}_t$
    5. update the step-size $\eta_{t+1}$
```

establishes that

$$\int_D \nabla f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\partial D} f(\mathbf{x})\boldsymbol{n}(\mathbf{x})\,\mathrm{d}S(\mathbf{x}) \enspace ,$$

where $\partial D$ is the boundary of $D$, $\boldsymbol{n}$ is the outward normal vector to $\partial D$, and $\mathrm{d}S(\mathbf{x})$ denotes the surface measure. Introducing $\boldsymbol{U}^D$ and $\boldsymbol{\zeta}^{\partial D}$ distributed uniformly on $D$ and $\partial D$ respectively, we can rewrite the above identity as

$$\mathbf{E}[\nabla f(\boldsymbol{U}^D)] = \frac{\operatorname{Vol}_{d-1}(\partial D)}{\operatorname{Vol}_d(D)} \cdot \mathbf{E}[f(\boldsymbol{\zeta}^{\partial D})\boldsymbol{n}(\boldsymbol{\zeta}^{\partial D})] \enspace ,$$

where $\operatorname{Vol}_{d-1}(\partial D)$ is the surface area of $D$ and $\operatorname{Vol}_d(D)$ is its volume. In what follows we consider the special case $D = \mathcal{B}_1^d$. For this choice of $D$ we have $\boldsymbol{n}(\mathbf{x}) = \frac{1}{\sqrt{d}} \cdot \operatorname{sign}(\mathbf{x})$ with $\operatorname{Vol}_{d-1}(\partial D)/\operatorname{Vol}_d(D) = d^{3/2}$ leading to our gradient estimate for the two-point feedback setup.

**Computational aspects.** Let us highlight two appealing practical features of the $\ell_1$-randomized gradient estimator $\mathbf{g}_t$ in Algorithm 4. First, we can easily evaluate any $\ell_p$-norm of $\mathbf{g}_t$. Indeed, it holds that $\|\mathbf{g}_t\|_p = (d^{1+1/p}/2h_t)|y_t' - y_t''|$, i.e., computing $\|\mathbf{g}_t\|_p$ only requires $O(1)$ elementary operations. Second, this gradient estimator is very economic in terms of the required memory: in order to store $\mathbf{g}_t$ we only need $d$ bits and $1$ float. None of these properties is inherent to the popular alternatives based on the randomization over the $\ell_2$-sphere (see e.g., Bach and Perchet, 2016; Flaxman et al., 2005; Nemirovsky and Yudin, 1983) or on Gaussian randomization (see e.g., Ghadimi and Lan, 2013; Nesterov, 2011; Nesterov and Spokoiny, 2017).

To compute $\mathbf{g}_t$ one needs to generate $\boldsymbol{\zeta}_t$ distributed uniformly on $\partial B_1^d$. The most straightforward way to do it consists in first generating a $d$-dimensional vector of i.i.d. centered scaled Laplace random variables and then normalizing this vector by its $\ell_1$-norm. The result is guaranteed to follow the uniform distribution on $\partial B_d^1$ (see e.g., Schechtman and Zinn, 1990, Lemma 1). Furthermore, to sample from the centered scaled Laplace distribution one can simply use inverse transform sampling. Indeed, if $U$ is distributed uniformly on $(0, 1)$, then $\log(2U)\mathbf{1}\,(U > 1/2) - \log(2 - 2U)\mathbf{1}\,(U \geq 1/2)$ follows centered scaled Laplace distribution.

## 4.3 Assumptions

We say that the convex function $V(\cdot)$ is $1$-strongly convex with respect to the $\ell_p$-norm on $\Theta$ if

$$V(\mathbf{x}') \geq V(\mathbf{x}) + \left\langle \boldsymbol{w}\, , \mathbf{x}' - \mathbf{x} \right\rangle + \frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|_p^2 \enspace ,$$

for all $\mathbf{x}, \mathbf{x}' \in \Theta$ and all $\boldsymbol{w} \in \partial V(\mathbf{x})$, where $\partial V(\mathbf{x})$ is the subdifferential of $V$ at point $\mathbf{x}$.

Throughout the paper, we assume that $p, q \in [1, \infty]$, $d \geq 3$, and set $p^*, q^* \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{p^*} = 1$ and $\frac{1}{q} + \frac{1}{q^*} = 1$, with the usual convention $1/\infty = 0$. We will use the following assumptions.

**Assumption 4.3.1.** *The following conditions hold:*

1. *The set $\Theta \subset \mathbb{R}^d$ is compact and convex.*
2. *There exists $V : \Theta \to \mathbb{R}$, which is lower semi-continuous, $1$-strongly convex on $\Theta$ w.r.t. the $\ell_p$-norm and such that*

   $$\sup_{\boldsymbol{x} \in \Theta} V(\boldsymbol{x}) - \inf_{\boldsymbol{x} \in \Theta} V(\boldsymbol{x}) \leq R^2$$

   *for some constant $R > 0$.*
3. *Each function $f_t : \mathbb{R}^d \to \mathbb{R}$ is convex on $\mathbb{R}^d$ for all $t \geq 1$.*
4. *For all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$, and all $t \geq 1$ we have $|f_t(\boldsymbol{x}) - f_t(\boldsymbol{x}')| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_q$ for some constant $L > 0$.*

Assumption 4.3.1 is rather standard in the study of dual averaging-type algorithms and have been previously considered in the context of zero-order problems in Duchi et al. (2015); Shamir (2017). We assume that $\Theta$ is compact as we are interested in the worst-case regret, which ensures that $R < +\infty$. We discuss extensions of our results to the case of unbounded $\Theta$ in Section 4.8. Note that the constant $R > 0$ is not necessarily dimension independent. Below we provide two classical examples of $V$ (see e.g., Shalev-Shwartz, 2012, Section 2).

**Example 4.3.2.** *Let $\Theta$ be any convex subset of $\mathbb{R}^d$ and $p \in (1, 2]$. Then, $V(\boldsymbol{x}) = \frac{1}{2(p-1)}\|\boldsymbol{x}\|_p^2$ is $1$-strongly convex on $\Theta$ w.r.t. the $\ell_p$-norm.*

**Example 4.3.3.** *Let $\Theta = \left\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_1 = 1\, ,\ \ \boldsymbol{x} \geq 0\right\}$. Then*[1]*, $V(\boldsymbol{x}) = \sum_{j=1}^d x_j \log(x_j)$ is $1$-strongly convex on $\Theta$ w.r.t. the $\ell_1$-norm and $R^2 \leq \log(d)$.*

**Assumptions on the noise.** We consider two different assumptions on the noises $\xi_t', \xi_t''$. The first noise assumption is common in the stochastic optimization context (see e.g., Duchi et al., 2015; Ghadimi and Lan, 2013; Nesterov, 2011; Nesterov and Spokoiny, 2017; Shamir, 2017).

**Assumption 4.3.4** (Canceling noise)**.** *For all $t = 1, 2, \ldots$, it holds that $\xi_t' = \xi_t''$ almost surely.*

[1] We use the convention that $0 \log(0) = 0$.

Formally, Assumption 4.3.4 permits noisy evaluations of function values. However, due to the fact that we are allowed to query $f_t$ at two points, taking difference of $y'_t$ and $y''_t$ in the estimator of the gradient effectively erases the noise. It results in a smaller variance of our gradient estimator. Importantly, Assumption 4.3.4 covers the case of no noise, that is, the classical online optimization setting as defined, e.g., in Shalev-Shwartz (2012).

Second, we consider an adversarial noise assumption, which is essentially equivalent to the assumptions used in Akhavan et al. (2020, 2021).

**Assumption 4.3.5** (Adversarial noise). *For all $t = 1, 2, \dots$, it holds that: (i) $\mathbf{E}[(\xi'_t)^2] \leq \sigma^2$ and $\mathbf{E}[(\xi''_t)^2] \leq \sigma^2$; (ii) $(\xi'_t)_{t\geq 1}$ and $(\xi''_t)_{t\geq 1}$ are independent of $(\boldsymbol{\zeta}_t)_{t\geq 1}$*

Assumption 4.3.5 allows for stochastic $\xi'_t$ and $\xi''_t$ that are not necessarily zero-mean or independent over the trajectory. Furthermore, it permits bounded non-stochastic adversarial noises. Part (ii) of Assumption 4.3.5 is always satisfied. Indeed, $\xi'_t$'s and $\xi''_t$'s are coming from the environment and are unknown to the learner while $\boldsymbol{\zeta}_t$'s are artificially generated by the learner. We mention part (ii) only for formal mathematical rigor.

Note that, since the choice of function $V$ belongs to the learner and $\Theta$ is given, it is always reasonable to assume that parameter $R$ is known. At the same time, parameters $L$ and $\sigma$ may be either known or unknown. We will study both cases in the next sections.

## 4.4 Upper bounds on the regret

In this section, we present the main convergence results for Algorithm 4 when $L, \sigma, T$ are known to the learner. The case when they are unknown is analyzed in Section 4.5, where we develop fully adaptive versions of Algorithm 4.

To state our results in a unified way, we introduce the following sequence that depends on the dimension $d$ and on the norm index $q \geq 1$:

$$\mathrm{b}_q(d) \triangleq \frac{1}{d+1} \cdot \begin{cases} qd^{\frac{1}{q}} & \text{if } q \in [1, \log(d)), \\ e\log(d) & \text{if } q \geq \log(d). \end{cases}$$

The value $\mathrm{b}_q(d)$ will explicitly influence the choice of the step size $\eta > 0$ and of the discretization parameter $h > 0$.

The first result of this section establishes the convergence guarantees under the canceling noise assumption. This case was previously considered by Duchi et al. (2015) and Shamir (2017).

**Theorem 4.4.1.** *Let Assumptions 4.3.1 and 4.3.4 be satisfied. Then, Algorithm 4 with the parameters*

$$\eta = \frac{AR}{L}\sqrt{\frac{d^{-1-\frac{2}{q\wedge 2}+\frac{2}{p}}}{T}} \quad \textit{and any} \quad h \leq \frac{7R}{100\mathrm{b}_q(d)\sqrt{T}} d^{\frac{1}{2}+\frac{1}{q\wedge 2}-\frac{1}{p}} \enspace ,$$

*where* $\mathtt{A} = (\sqrt{6}+\sqrt{12})^{-1}$*, satisfies, for any* $\boldsymbol{x} \in \Theta$*,*

$$\mathbf{E}\left[\sum_{t=1}^{T}\left(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x})\right)\right] \leq 11.9 \cdot RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}}\ .$$

Note that, as in other related works Duchi et al. (2015); Hu et al. (2016b); Nesterov (2011); Nesterov and Spokoiny (2017); Shamir (2017), under the canceling noise (or no noise) assumption the discretization parameter $h > 0$ can be chosen arbitrary small. This is due to the fact that, under the canceling noise assumption, the variance of the gradient estimate $\mathbf{g}_t$ is bounded by a constant independent of $h$. It is no longer the case under the adversarial noise assumption as exhibited in the next theorem.

**Theorem 4.4.2.** *Let Assumptions 4.3.1 and 4.3.5 be satisfied. Then Algorithm 4 with the parameters*

$$\eta = \frac{R}{\sqrt{TL}}\left(\frac{\sigma \mathrm{b}_q(d)}{\sqrt{2}R}\sqrt{Td^{4-\frac{2}{p}}}+\mathtt{A}Ld^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}\right)^{-\frac{1}{2}} \quad \textit{and} \quad h = \left(\frac{\sqrt{2}R\sigma}{L\mathrm{b}_q(d)}\right)^{\frac{1}{2}} T^{-\frac{1}{4}}d^{1-\frac{1}{2p}}\ ,$$

*where* $\mathtt{A} = 6(1+\sqrt{2})^2$*, satisfies, for any* $\boldsymbol{x} \in \Theta$*,*

$$\begin{aligned}\mathbf{E}\left[\sum_{t=1}^{T}\left(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x})\right)\right] &\leq 11.9 \cdot RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}} \\ &\quad + 2.4 \cdot \sqrt{RL\sigma}T^{\frac{3}{4}} \cdot \begin{cases}\sqrt{qd^{1+\frac{1}{q}-\frac{1}{p}}} & \textit{if } q \in [1, \log(d)), \\ \sqrt{e\log(d)d^{1-\frac{1}{p}}} & \textit{if } q \geq \log(d).\end{cases}\end{aligned}$$

**Comparison to state-of-the-art bounds.** We provide two examples of $p, q, \Theta$ and compare results for our new method to those of Duchi et al. (2015); Shamir (2017) where only the canceling noise Assumption 4.3.4 and $q \in \{1, 2\}$ were considered.

**Corollary 4.4.3.** *Let* $p = q = 2$ *and* $\Theta = \mathcal{B}_2^d$*. Then under Assumption 4.3.4, Algorithm 4 with* $V : \Theta \to \mathbb{R}$ *defined in Example 4.3.2, satisfies*

$$\mathbf{E}\left[\sum_{t=1}^{T}\left(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x})\right)\right] \leq 11.9 \cdot L\sqrt{dT}\ .$$

In the setup of Corollary 4.4.3, Duchi et al. (2015) obtain $O(L\sqrt{dT\log(d)})$ rate and Shamir (2017) exhibits $O(L\sqrt{dT})$, which is the optimal rate. Both results do not specify the leading absolute constants.

**Corollary 4.4.4.** *Let* $p = q = 1$ *and* $\Theta = \left\{\boldsymbol{x} \in \mathbb{R}^d\ :\ \boldsymbol{x} \geq 0,\ \|\boldsymbol{x}\|_1 = 1\right\}$*. Then under Assump-*

*tion 4.3.4, Algorithm 4 with $V : \Theta \to \mathbb{R}$ defined in Example 4.3.3, satisfies*

$$\mathbf{E}\left[\sum_{t=1}^{T}\big(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x})\big)\right] \le 11.9 \cdot L\sqrt{dT\log(d)}\ .$$

In the setup of Corollary 4.4.4, Shamir (2017) proves the rate $O(L\sqrt{dT}\log(d))$ for the method with $\ell_2$-randomization. On the other hand, Duchi et al. (2015) derived a lower bound $\Omega(\sqrt{dT/\log(d)})$. Thus, our algorithm further reduces the gap between the upper and the lower bounds.

Finally, note that in the case $p = 1$, $q = 2$ with $V : \Theta \to \mathbb{R}$ defined in Example 4.3.3 the bound of Theorem 4.4.1 is of the order $O(\sqrt{T\log(d)})$. This case was handled by an algorithm with $\ell_1$-randomization slightly different from ours in Gasnikov et al. (2016) leading to the suboptimal rate $O(\sqrt{dT\log(d)})$.

## 4.5 Adaptive algorithms

Theorems 4.4.1 and 4.4.2 used the step size $\eta$ and the discretization parameter $h$ that depend on the potentially unknown quantities $L$, $\sigma$, and the optimization horizon $T$. In this section, we show that, under the canceling noise Assumption 4.3.4, adaptation to unknown $L$ comes with nearly no price. On the other hand, under the adversarial noise Assumption 4.3.5, our adaptive rate has a slightly worse dependence on $L$ and $\sigma$ in the dominant term. The proof is based on combining the adaptive scheme for online dual averaging (see Section 7.13 in Orabona, 2019, for an overview) with our bias and variance evaluations, cf. Section 4.6 below.

**Theorem 4.5.1.** *Let Assumptions 4.3.1 and 4.3.4 be satisfied. Then, Algorithm 4 with the parameters*[2]

$$\eta_t = \frac{R}{\sqrt{2.75 \cdot \sum_{k=1}^{t-1}\|\boldsymbol{g}_k\|_{p^*}^2}} \quad \textit{and any} \quad h_t \le \frac{7R}{200\mathrm{b}_q(d)\sqrt{t}}d^{\frac{1}{2}+\frac{1}{q\wedge 2}-\frac{1}{p}}\ ,$$

*satisfies for any $\boldsymbol{x} \in \Theta$*

$$\mathbf{E}\left[\sum_{t=1}^{T}\big(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x})\big)\right] \le 110.6 \cdot RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}}\ .$$

The above result gives, up to an absolute constant, the same convergence rate as that of the non-adaptive Theorem 4.4.1. In other words, the price for adaptive algorithm does not depend on the parameters of the problem. Finally, we derive an adaptive algorithm under Assumption 4.3.5.

[2] We adopt the convention that $\eta_1 = 1$ and $1/0 = 1$ in the definition of $\eta_t$.

**Theorem 4.5.2.** *Let Assumptions 4.3.1 and 4.3.5 be satisfied. Then, Algorithm 4 with the parameters*

$$\eta_t = \frac{R}{\sqrt{2.75 \cdot \sum_{k=1}^{t-1} \|\boldsymbol{g}_k\|_{p^*}^2}} \quad \textit{and any} \quad h_t = \left(6.65\sqrt{6} \cdot \frac{R}{\mathrm{b}_q(d)}\right)^{\frac{1}{2}} t^{-\frac{1}{4}} d^{1-\frac{1}{2p}} \ ,$$

*satisfies for any $\boldsymbol{x} \in \Theta$*

$$\mathbf{E}\left[\sum_{t=1}^{T} \left(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x})\right)\right] \leq 110.6 \cdot RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}}$$
$$+ 5.9 \cdot \sqrt{R}\,(\sigma{+}L)\,T^{\frac{3}{4}} \cdot \begin{cases} \sqrt{qd^{1+\frac{1}{q}-\frac{1}{p}}} & \textit{if } q \in [1, \log(d)) \\ \sqrt{e\log(d)d^{1-\frac{1}{p}}} & \textit{if } q \geq \log(d) \end{cases} .$$

Note that the bound of Theorem 4.5.2 has a less advantageous dependency on $\sigma$ and $L$ compared to Theorem 4.4.2, where we had $\sqrt{\sigma L}$ instead of $\sigma + L$. We remark that if $\sigma$ is known but $L$ is unknown, one can recover the $\sqrt{\sigma L}$ dependency by selecting $h_t$ depending on $\sigma$. We do not state this result that can be derived in a similar way and favor here only the fully adaptive version.

## 4.6 Elements of proofs

In this section, we outline major ingredients for the proofs of Theorems 4.4.1 – 4.5.2. The full proofs can be found in Section 4.9. Here, we only focus on novel elements without reproducing the general scheme of online dual averaging analysis (see e.g., Orabona, 2019; Shalev-Shwartz, 2012). Namely, we highlight two key facts, which are the smoothing lemma (Lemma 4.6.1) and the weighted Poincaré type inequality for the uniform measure on $\partial B_1^d$ (Lemma 4.6.3) used to control the variance.

### Bias and smoothing lemma

First, as in the prior work that was using smoothing ideas (see e.g., Flaxman et al., 2005; Nemirovsky and Yudin, 1983; Shamir, 2017), we show that our gradient estimate $\mathbf{g}_t$ is an unbiased estimator of a surrogate version of $f_t$ and establish its approximation properties.

**Lemma 4.6.1** (Smoothing lemma)**.** *Fix $h > 0$ and $q \in [1, \infty]$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an $L$-Lipschitz function w.r.t. the $\ell_q$-norm. Let $\boldsymbol{U}$ be distributed uniformly on $\mathcal{B}_1^d$ and $\zeta$ be distributed uniformly on $\partial B_d^1$. Let $\mathsf{f}_h(\boldsymbol{x}) \triangleq \mathbf{E}[f(\boldsymbol{x} + h\boldsymbol{U})]$ for $\boldsymbol{x} \in \mathbb{R}^d$. Then $\mathsf{f}_h$ is differentiable and*

$$\mathbf{E}\left[\frac{d}{2h}\left(f(\boldsymbol{x} + h\boldsymbol{\zeta}) - f(\boldsymbol{x} - h\boldsymbol{\zeta})\right)\operatorname{sign}(\boldsymbol{\zeta})\right] = \nabla \mathsf{f}_h(\boldsymbol{x}) \ .$$

*Furthermore, we have for all $d \geq 3$ and all $\boldsymbol{x} \in \mathbb{R}^d$,*

$$|\mathsf{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \mathrm{b}_q(d) L h \ . \tag{4.1}$$

*Finally, if $\Theta \subset \mathbb{R}^d$ is convex, $f$ is convex in $\Theta + h\mathcal{B}_1^d$, then $\mathsf{f}_h$ is convex in $\Theta$ and $\mathsf{f}_h(\boldsymbol{x}) \geq f(\boldsymbol{x})$ for $\boldsymbol{x} \in \Theta$.*

*Proof.* There are three claims to prove. For the first one, we notice that $\boldsymbol{\zeta}$ has the same distribution as $-\boldsymbol{\zeta}$, hence,

$$\mathbf{E}\left[\frac{d}{2h}\big(f(\mathbf{x} + h\boldsymbol{\zeta}) - f(\mathbf{x} - h\boldsymbol{\zeta})\big)\operatorname{sign}(\boldsymbol{\zeta})\right] = \mathbf{E}\left[\frac{d}{h} f(\mathbf{x} + h\boldsymbol{\zeta})\operatorname{sign}(\boldsymbol{\zeta})\right] \ ,$$

and the first claim follows from Theorem 4.9.3 in Section 4.9 (a version of Stokes', or divergence, theorem) applied to $g(\cdot) = f(\mathbf{x} + h\cdot)$ with observation that $\nabla g(\cdot) = h\nabla f(\mathbf{x} + h\cdot)$ where $\nabla f$ is the gradient defined almost everywhere and whose existence is ensured by the Rademacher theorem.

We now prove the approximation property (4.1). Assuming $d \geq 3$ grants that $\log(d) \geq 1$. Since $f$ is $L$-Lipschitz w.r.t. the $\ell_q$-norm we get that, for any $\mathbf{x} \in \mathbb{R}^d$,

$$|\mathsf{f}_h(\mathbf{x}) - f(\mathbf{x})| \leq Lh\,\mathbf{E}\|\boldsymbol{U}\|_q \ . \tag{4.2}$$

If $q \in [1, \log(d))$ then (4.1) follows from Lemma 4.6.2. If $q \geq \log(d)$ then using again Lemma 4.6.2 we find

$$\mathbf{E}\|\boldsymbol{U}\|_q \leq \mathbf{E}\|\boldsymbol{U}\|_{\log(d)} \leq \frac{\log(d) d^{\frac{1}{\log(d)}}}{d+1} = \frac{e\log(d)}{d+1} \ ,$$

which together with (4.2) yields the desired bound.

Finally, if $f$ is convex in $\Theta + h\mathcal{B}_1^d$, then for all $\mathbf{x}, \mathbf{x}' \in \Theta$ and $\alpha \in [0,1]$ we have

$$\mathsf{f}_h(\alpha\mathbf{x} + (1-\alpha)\mathbf{x}') = \mathbf{E}\Big[f\big(\alpha(\mathbf{x} + h\boldsymbol{U}) + (1-\alpha)(\mathbf{x}' + h\boldsymbol{U})\big)\Big] \leq \alpha\mathsf{f}_h(\mathbf{x}) + (1-\alpha)\mathsf{f}_h(\mathbf{x}') \ .$$

Thus $\mathsf{f}_h$ is indeed convex on $\Theta$. Furthermore, again by convexity of $f$, we deduce that for any $\mathbf{x} \in \Theta$

$$\mathsf{f}_h(\mathbf{x}) = \mathbf{E}[f(\mathbf{x} + h\boldsymbol{U})] \geq \mathbf{E}\left[f(\mathbf{x}) + \langle \boldsymbol{w}, h\boldsymbol{U}\rangle\right] = f(\mathbf{x}) \quad \text{where } \boldsymbol{w} \in \partial f(\mathbf{x}) \ .\square$$

The proof of Lemma 4.6.1 relies on the control of the $\ell_q$-norm of random vector $\boldsymbol{U}$ established in the next result.

**Lemma 4.6.2.** *Let $q \in [1, \infty)$ and let $\boldsymbol{U}$ be distributed uniformly on $\mathcal{B}_1^d$. Then $\mathbf{E}\|\boldsymbol{U}\|_q \leq \frac{qd^{\frac{1}{q}}}{d+1}$.*

*Proof.* Let $W_1, \dots, W_d, W_{d+1}$ be i.i.d. random variables having the Laplace distribution with mean $0$ and scale parameter $1$. Set $\boldsymbol{W} = (W_1, \dots, W_d)$. Then, following (Barthe et al., 2005,

Theorem 1) we have

$$\boldsymbol{U} \overset{d}{=} \frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_1 + |W_{d+1}|} \enspace ,$$

where $\overset{d}{=}$ denotes equality in distribution. Furthermore, (Schechtman and Zinn, 1990, Lemma 1) states that the random variables

$$\frac{(\boldsymbol{W}, |W_{d+1}|)}{\|\boldsymbol{W}\|_1 + |W_{d+1}|} \qquad \text{and} \qquad \|\boldsymbol{W}\|_1 + |W_{d+1}| \enspace ,$$

are independent. Hence, for any $q \in [1, \infty)$, it holds that

$$\mathbf{E}\|\boldsymbol{U}\|_q = \frac{\mathbf{E}\|\boldsymbol{W}\|_q}{\mathbf{E}\|(\boldsymbol{W}, W_{d+1})\|_1} = \frac{1}{d+1}\mathbf{E}\|\boldsymbol{W}\|_q \overset{(a)}{\leq} \frac{1}{d+1}\left(\mathbf{E}\|\boldsymbol{W}\|_q^q\right)^{\frac{1}{q}} = \frac{d^{\frac{1}{q}}\Gamma^{\frac{1}{q}}(q+1)}{d+1} \overset{(b)}{\leq} \frac{qd^{\frac{1}{q}}}{d+1} \enspace ,$$

where $(a)$ follows from Jensen's inequality and $(b)$ uses the fact that $\Gamma^{1/q}(q+1) \leq q$ for $q \geq 1$. □

### fr:variance and weighted Poincaré type inequality

We additionally need to control the squared $\ell_{p^*}$-norm of each gradient estimator $\mathbf{g}_t$. This is where we get the main improvement of our procedure compared to previously proposed methods. To derive the result, we first establish the following lemma of independent interest, which allows us to control the variance of Lipschitz functions on $\partial B_1^d$. The proof of this lemma is given in the Section 4.9.

**Lemma 4.6.3.** *Let $d \geq 3$. Assume that $G : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function, and $\zeta$ is distributed uniformly on $\partial B_1^d$. Then*

$$\mathrm{Var}(G(\boldsymbol{\zeta})) \leq \frac{4}{d(d-2)}\mathbf{E}\left[\|\nabla G(\boldsymbol{\zeta})\|_2^2\left(1 + \sqrt{d}\|\boldsymbol{\zeta}\|_2\right)^2\right] \enspace .$$

*Furthermore, if $G : \mathbb{R}^d \to \mathbb{R}$ is an $L$-Lipschitz function w.r.t. the $\ell_2$-norm then*

$$\mathrm{Var}(G(\boldsymbol{\zeta})) \leq \frac{4L^2}{d(d-2)}\left(1 + \sqrt{\frac{2d}{d+1}}\right)^2 \enspace .$$

**Remark 4.6.4.** *Since $d^2/(d(d-2)) \leq 3$ for all $d \geq 3$, the last inequality of Lemma 4.6.3 implies that*

$$\mathrm{Var}(G(\boldsymbol{\zeta})) \leq 12\big(1 + \sqrt{2}\big)^2\big(L/d\big)^2 \enspace , \qquad \forall d \geq 3 \enspace . \tag{4.3}$$

We can now deduce the following bound on the squared $\ell_{p^*}$-norm of $\mathbf{g}_t$.

**Lemma 4.6.5.** *Let $p \in [1, \infty]$ and $p^* = \frac{p}{p-1}$. Assume that $f_t$ is $L$-Lipschitz w.r.t. the $\ell_q$-norm. Then, for all $d \geq 3$,*

$$\mathbf{E}\|\boldsymbol{g}_t\|_{p^*}^2 \leq 12(1+\sqrt{2})^2 L^2 d^{1+\frac{2}{q \wedge 2}-\frac{2}{p}} + \begin{cases} 0 & \textit{under canceling noise Assumption 4.3.4,} \\ \dfrac{d^{4-\frac{2}{p}}\sigma^2}{h^2} & \textit{under adversarial noise Assumption 4.3.5.} \end{cases}$$

*Proof.* Using the definition of $\mathbf{g}_t$ we get

$$\begin{aligned} \mathbf{E}[\|\mathbf{g}_t\|_{p^*}^2 \mid \mathbf{x}_t] &= \frac{d^2}{4h^2}\mathbf{E}[(f_t(\mathbf{x}_t + h\boldsymbol{\zeta}_t) - f_t(\mathbf{x}_t - h\boldsymbol{\zeta}_t) + \xi'_t - \xi''_t)^2 \|\operatorname{sign}(\boldsymbol{\zeta}_t)\|_{p^*}^2 \mid \mathbf{x}_t] \\ &= \frac{d^{4-\frac{2}{p}}}{4h^2}\mathbf{E}[(f_t(\mathbf{x}_t + h\boldsymbol{\zeta}_t) - f_t(\mathbf{x}_t - h\boldsymbol{\zeta}_t) + \xi'_t - \xi''_t)^2 \mid \mathbf{x}_t] \ . \end{aligned}$$

Let $G(\boldsymbol{\zeta}) \triangleq f_t(\mathbf{x}_t + h\boldsymbol{\zeta}) - f_t(\mathbf{x}_t - h\boldsymbol{\zeta})$. First, observe that $\mathbf{E}[G(\boldsymbol{\zeta}_t) \mid \mathbf{x}_t] = 0$ and under both Assumption 4.3.4 and Assumption 4.3.5(ii) it holds that $\mathbf{E}[G(\boldsymbol{\zeta}_t)(\xi'_t - \xi''_t) \mid \mathbf{x}_t] = 0$. Using these remarks and the fact that under adversarial noise Assumption 4.3.5, $\mathbf{E}[(\xi'_t - \xi''_t)^2 \mid \mathbf{x}_t] \leq 4\sigma^2$, we find:

$$\mathbf{E}[\|\mathbf{g}_t\|_{p^*}^2 \mid \mathbf{x}_t] \leq \frac{d^{4-\frac{2}{p}}}{4h^2}\left(\operatorname{Var}(G(\boldsymbol{\zeta}_t) \mid \mathbf{x}_t) + \begin{cases} 0 & \text{under cancelling noise Assumption 4.3.4} \\ 4\sigma^2 & \text{under adversarial noise Assumption 4.3.5} \end{cases}\right) .$$

Furthermore, since $f_t$ is $L$-Lipschitz, w.r.t. the $\ell_q$-norm, the map $\boldsymbol{\zeta} \mapsto G(\boldsymbol{\zeta})$ is $\big(2Lhd^{\frac{1}{q\wedge 2}-\frac{1}{2}}\big)$-Lipschitz w.r.t. the $\ell_2$-norm. Applying (4.3) to bound $\operatorname{Var}(G(\boldsymbol{\zeta}_t) \mid \mathbf{x}_t)$, yields the desired result. □

Note that under adversarial noise Assumption 4.3.5, the bound on squared $\ell_{p^*}$-norm of $\mathbf{g}_t$ gets an additional term $d^{4-\frac{2}{p}}\sigma^2 h^{-2}$. In contrast to the case of canceling noise Assumption 4.3.4, this does not allow us to take $h$ arbitrary small hence inducing the bias-variance trade-off.

## 4.7 Numerical illustration

In this section, we provide a numerical comparison of our algorithm with the method based on $\ell_2$-randomization from Shamir (2017) (see Section 4.9 for the definition). We consider the no noise model and $f_t = f$, $\forall t$, with the function $f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f(\mathbf{x}) = \|\mathbf{x} - \boldsymbol{c}\|_2 + \|\mathbf{x} - 0.1 \cdot \boldsymbol{c}\|_1 \quad ,$$

where $\boldsymbol{c} = (c_1, \ldots, c_d)^\top \in \mathbb{R}^d$ such that $c_j = \exp(j)/\sum_{i=1}^d \exp(i)$ for $j = 1, \ldots, d$. We choose

$$\Theta = \left\{\mathbf{x} \in \mathbb{R}^d \ : \ \|\mathbf{x}\|_1 = 1, \, \mathbf{x} \geq 0\right\} \quad \text{and} \quad V(\mathbf{x}) = \sum_{j=1}^d x_j \log(x_j) \ .$$

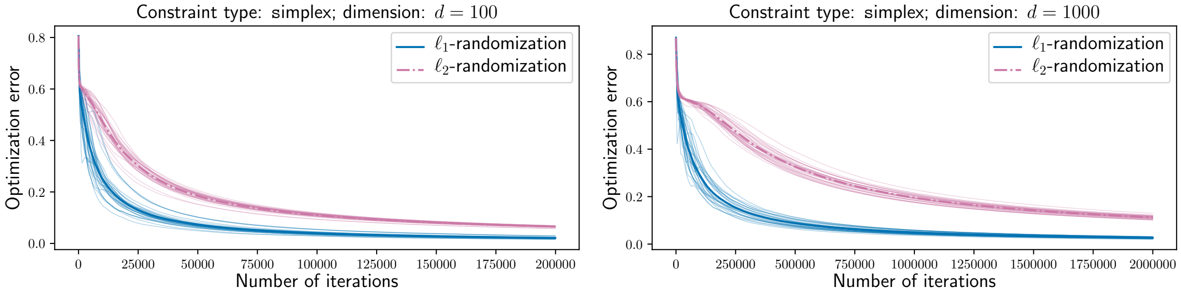

Figure 4.1: Opt. error vs. number of iterations for $\ell_2$-randomization (as in Shamir (2017)) and our method.

As stated in Example 4.3.3, $V$ is $1$-strongly convex on $\Theta$ w.r.t. the $\ell_1$-norm and $R \leq \sqrt{\log(d)}$. Moreover, $f$ is a Lipschitz function w.r.t. the $\ell_1$-norm. We deploy the adaptive parameterization that appears in Theorem 4.5.1. In Figure 4.1 we present the optimization error of the algorithms, which is defined as

$$f\left(\frac{1}{t}\sum_{i=1}^{t}\mathbf{x}_i\right) - \min_{\mathbf{x}\in\Theta} f(\mathbf{x}) \enspace .$$

The results are reported over $30$ trials. We plot all the $30$ runs alongside the average performance. One can observe that the $\ell_1$-randomization method behaves significantly better than the $\ell_2$-randomization algorithm. The theoretical bound for our method in this setup has a $\sqrt{\log d}$ gain in the rate.

## 4.8 Discussion and comparison to prior work

We introduced and analyzed a novel estimator for the gradient based on randomization over the $\ell_1$-sphere. We established guarantees for the online dual averaging algorithm with the gradient replaced by the proposed estimator. We provided an anytime and completely data-driven algorithm, which is adaptive to all parameters of the problem. Our analysis is based on deriving a weighted Poincaré type inequality for the uniform measure on the $\ell_1$-sphere that may be of independent interest.

Under the *canceling noise assumption* and $q \in \{1, 2\}$, our setting is analogous to Duchi et al. (2015); Shamir (2017). For the case $q = p = 2$ and canceling noise, we show that the performance of our method is the same as in (Shamir, 2017, Corollary 2) up to absolute constants that were not made explicit in Shamir (2017). For the case of $q = p = 1$ and canceling noise, we improved the bound (Shamir, 2017, Corollary 3) by a $\sqrt{\log(d)}$ factor. For the case $q = 2$, $p \geq 1$, comparing with the lower bound in (Duchi et al., 2015, Proposition 1), shows that the result of Theorem 4.4.1 is minimax optimal. For the case $q = p = 1$, (Duchi et al., 2015, Proposition 2) shows that our result in Theorem 4.4.1 is optimal up to a $\log(d)$ factor.

Under the *adversarial noise assumption*, Theorem 4.4.2 provides the rate $O(T^{3/4})$, that is, we get an additional $T^{1/4}$ factor compared to the canceling noise case. It remains unclear whether it is optimal under adversarial noise – this question deserves further investigation. Note that, under sub-Gaussian $\mathrm{i.i.d.}$ noise assumption and $q = p = 2$, one can achieve the rate $\tilde{O}(d^a\sqrt{T})$ with a relatively big $a > 0$ Agarwal et al. (2011); Belloni et al. (2015); Bubeck et al. (2017); Lattimore and Gyorgy (2021). In particular, with an ellipsoid type method Lattimore and Gyorgy (2021) obtains the rate $\mathcal{O}(d^{4.5}\sqrt{T}\log(T)^2)$ for the cumulative regret.

Finally, let us discuss the compactness of $\Theta$. It is straightforward to extend the results of Theorems 4.4.1, 4.4.2 to any closed convex $\Theta$ considering the regret against a fixed action $\mathbf{x} \in \Theta$. Indeed, using (Orabona, 2019, Corollary 7.9), one only needs to replace $R$ appearing in both Theorems 4.4.1, 4.4.2 by an upper bound on $\sqrt{V(\mathbf{x}) - \inf_{\mathbf{x}' \in \Theta} V(\mathbf{x}')}$. The adaptive case is more complicated. One way to tackle this case is to use (Orabona and Pál, 2016, Theorem 1) requiring a control of $\mathbf{E}\max_{t=1,\dots,T}\|\mathbf{g}_t\|_{p^*}$. This term can be controlled under the canceling noise Assumption 4.3.4 using the Lipschitzness of $f_t$'s, so that Theorem 4.5.1 extends to unbounded $\Theta$. However, without the canceling noise assumption, following the approach outlined above, one needs to control $\mathbf{E}\max_{t=1,\dots,T}\frac{|\xi'_t - \xi''_t|}{h_t}$. The adversarial noise Assumption 4.3.5 is not sufficient to reasonably control this term, so that extending Theorem 4.5.2 to unbounded $\Theta$ is not possible without further assumptions.

## 4.9 Proofs

This section contains the proofs and results omitted from the main body. In Section 4.9 we recall the appropriate version of the Stokes' theorem and discuss its applicability for Lipschitz functions on $\mathcal{B}_1^d$. In Section 4.9 we provide the proof of Lemma 4.6.3. Finally, in Section 4.9 we provide the proofs of Theorems 4.4.1, 4.4.2, 4.5.1, 4.5.2.

**Additional notation** For two functions $g, \eta : \mathbb{R}^d \to \mathbb{R}$, we denote by $\eta \star g$ their convolution defined point-wise for $\mathbf{x} \in \mathbb{R}^d$ as

$$(\eta \star g)(\mathbf{x}) = \int_{\mathbb{R}^d} \eta(\mathbf{x} - \mathbf{x}')g(\mathbf{x}')\,\mathrm{d}\mathbf{x}' \enspace .$$

The standard mollifier $\eta_\epsilon : \mathbb{R}^d \to \mathbb{R}$ is defined as $\eta_\epsilon(\mathbf{x}) = \epsilon^{-d}\eta_1(\mathbf{x}/\epsilon)$ for $\epsilon > 0$ and $\mathbf{x} \in \mathbb{R}$, where $\eta_1 : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\eta_1(\mathbf{x}) = \begin{cases} C\exp\left(\frac{1}{\|\mathbf{x}\|_2^2 - 1}\right) & \text{if } \|\mathbf{x}\|_2 \leq 1 \\ 0 & \text{otherwise} \end{cases} ,$$

with $C$ chosen so that $\int_{\mathbb{R}^d}\eta_1(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1$.

### Integration by parts

We first recall the following result that can be found in (Zorich, 2016, Section 13.3.5, Exercise 14a).

**Theorem 4.9.1** (Integration by parts in a multiple integral)**.** *Let $D$ be an open connected subset of $\mathbb{R}^d$ with a piecewise smooth boundary $\partial D$ oriented by the outward unit normal $\boldsymbol{n} = (n_1, \dots, n_d)^\top$. Let $g$ be a continuously differentiable function in $D \cup \partial D$. Then*

$$\int_D \nabla g(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} = \int_{\partial D} g(\boldsymbol{\zeta}) \boldsymbol{n}(\boldsymbol{\zeta}) \, \mathrm{d}S(\boldsymbol{\zeta}) \enspace .$$

**Remark 4.9.2.** *We refer to (Zorich, 2016, Section 12.3.2, Definitions 4 and 5) for the definition of piecewise smooth surfaces and their orientations respectively.*

The idea of using the instance of Theorem 4.9.1 (also called Stokes' theorem) with $D = \mathcal{B}_2^d$ to obtain $\ell_2$-randomized estimators of the gradient belongs to Nemirovsky and Yudin (1983). It was further used in several papers (Bach and Perchet, 2016; Flaxman et al., 2005; Shalev-Shwartz, 2012; Shamir, 2017) to mention just a few. Those papers were referring to Nemirovsky and Yudin (1983) but Nemirovsky and Yudin (1983) did not provide an exact statement of the result (nor a reference) and only tossed the idea in a discussion. However, the classical analysis formulation as presented in Theorem 4.9.1 does not apply to Lipschitz continuous functions that were considered in (Bach and Perchet, 2016; Flaxman et al., 2005; Shalev-Shwartz, 2012; Shamir, 2017). We are not aware of whether its extension to Lipschitz continuous functions, though rather standard, is proved in the literature.

In this paper, we apply Theorem 4.9.1 with the $\ell_1$-ball $D = \mathcal{B}_1^d$. Our aim in this section is to provide a variant of Theorem 4.9.1 applicable to a Lipschitz continuous function $g : \mathbb{R}^d \to \mathbb{R}$, which is not necessarily continuously differentiable on $D \cup \partial D = \mathcal{B}_1^d \cup \partial B_1^d$. To this end, we will go through the argument of approximating $g$ by $C^\infty(\Omega)$ functions, where $\Omega \subset \mathbb{R}^d$ is an open bounded connected subset of $\mathbb{R}^d$ such that $D \cup \partial D \subset \Omega$. Let $g_n = \eta_{1/n} \star g$, where $\eta_{1/n}$ is the standard mollifier. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a function satisfying the Lipschitz condition w.r.t. the $\ell_1$-norm: $|g(\boldsymbol{u}) - g(\boldsymbol{u}')| \leq L\|\boldsymbol{u} - \boldsymbol{u}'\|_1$. Since $g$ is continuous in $\Omega$ and, by construction $D \cup \partial D \subset \Omega$, then using basic properties of mollification (see e.g., Evans and Gariepy, 2018, Theorem 4.1 (ii)) we have

$$g_n \longrightarrow g$$

uniformly on $D \cup \partial D$ (in particular, uniformly on $\partial D$). Furthermore, let $\nabla g$ be the gradient of $g$, which by Rademacher theorem (see e.g., Evans and Gariepy, 2018, Theorem 3.2) is well defined almost everywhere w.r.t. the Lebesgue measure and

$$\|\nabla g(\boldsymbol{u})\|_\infty \leq L \qquad \text{a.e.}$$

It follows that $\frac{\partial g}{\partial u_j}$ is absolutely integrable on $\Omega$ for any $j \in [d]$. Furthermore, since

$$\frac{\partial g_n}{\partial u_j} = \eta_{1/n} \star \left(\frac{\partial g}{\partial u_j}\right) \quad ,$$

we can apply (Evans and Gariepy, 2018, Theorem 4.1 (iii)) that yields

$$\int_D \|\nabla g_n(\boldsymbol{u}) - \nabla g(\boldsymbol{u})\|_2 \,\mathrm{d}\boldsymbol{u} \longrightarrow 0 \enspace .$$

Combining the above remarks we obtain that the result of Theorem 4.9.1 is valid for functions $g$ that are Lipschitz continuous w.r.t. the $\ell_1$-norm. Thus, it is also valid when the Lipschitz condition is imposed w.r.t. any $\ell_q$-norm with $q \in [1, \infty]$. Specifying this conclusion for the particular case $D = B_1^d$, we obtain the following theorem.

**Theorem 4.9.3.** *Let the function $g : \mathbb{R}^d \to \mathbb{R}$ be Lipschitz continuous w.r.t. the $\ell_q$-norm with $q \in [1, \infty]$. Then*

$$\int_{B_1^d} \nabla g(\boldsymbol{u}) \,\mathrm{d}\boldsymbol{u} = \frac{1}{\sqrt{d}} \int_{\partial B_1^d} g(\boldsymbol{\zeta}) \operatorname{sign}(\boldsymbol{\zeta}) \,\mathrm{d}S(\boldsymbol{\zeta}) \enspace ,$$

*where $\nabla g(\cdot)$ is defined up to a set of zero Lebesgue measure by the Rademacher theorem.*

### Proof of Lemma 4.6.3

To prove Lemma 4.6.3, we first recall the weighted Poincaré inequality for the univariate exponential measure (mean $0$ and scale parameter $1$ Laplace distribution).

**Lemma 4.9.4** (Lemma 2.1 in Bobkov and Ledoux (1997))**.** *Let $W$ be mean $0$ and scale parameter $1$ Laplace random variable. Let $g : \mathbb{R} \to \mathbb{R}$ be continuous almost everywhere differentiable function such that*

$$\mathbf{E}[|g(W)|] < \infty \quad \textit{and} \quad \mathbf{E}[|g'(W)|] < \infty \quad \textit{and} \quad \lim_{|w| \to \infty} g(w) \exp(-|w|) = 0 \enspace ,$$

*then,*

$$\mathbf{E}[(g(W) - \mathbf{E}[g(W)])^2] \leq 4\mathbf{E}[(g'(W))^2].$$

We are now in a position to prove Lemma 4.6.3. The proof is inspired by (Barthe and Wolff, 2009, Lemma 2).

*Proof of Lemma 4.6.3.* Throughout the proof, we assume without loss of generality that $\mathbf{E}[G(\boldsymbol{\zeta})] = 0$. Indeed, if it is not the case, we use the result for the centered function $\tilde{G}(\boldsymbol{\zeta}) = G(\boldsymbol{\zeta}) - \mathbf{E}[G(\boldsymbol{\zeta})]$, which has the same gradient.

First, consider the case of continuously differentiable $G$. Let $\boldsymbol{W} = (W_1, \dots, W_d)$ be a vector of i.i.d. mean $0$ and scale parameter $1$ Laplace random variables and define $\boldsymbol{T}(\boldsymbol{w}) = \boldsymbol{w}/\left\|\boldsymbol{w}\right\|_1$. Introduce the notation

$$F(\boldsymbol{w}) \triangleq \|\boldsymbol{w}\|_1^{1/2} G(\boldsymbol{T}(\boldsymbol{w})) \ .$$

Lemma 1 in Schechtman and Zinn (1990) asserts that, for $\zeta$ uniformly distributed on $\partial B_1^d$,

$$\boldsymbol{T}(\boldsymbol{W}) \overset{d}{=} \zeta \quad \text{and} \quad \boldsymbol{T}(\boldsymbol{W}) \text{ is independent of } \|\boldsymbol{W}\|_1 \ . \tag{4.4}$$

In particular,

$$\mathrm{Var}(F(\boldsymbol{W})) = d\, \mathrm{Var}(G(\boldsymbol{\zeta})) \ .$$

Using the Efron-Stein inequality (see e.g., Boucheron et al., 2013, Theorem 3.1) we obtain

$$\mathrm{Var}(F(\boldsymbol{W})) \leq \sum_{i=1}^{d} \mathbf{E}\left[\mathrm{Var}_i(F)\right] \ ,$$

where

$$\mathrm{Var}_i(F) = \mathbf{E}\Big[\big(F(\boldsymbol{W}) - \mathbf{E}[F(\boldsymbol{W}) \mid \boldsymbol{W}^{-i}]\big)^2 \mid \boldsymbol{W}^{-i}\Big]$$

with $\boldsymbol{W}^{-i} \triangleq (W_1, \dots, W_{i-1}, W_{i+1}, \dots, W_d)$. Note that on the event $\{\boldsymbol{W}^{-i} \neq \mathbf{0}\}$ (whose complement has zero measure), the function

$$w \mapsto F(W_1, \dots, W_{i-1}, w, W_{i+1}, \dots, W_d) \ ,$$

satisfies the assumptions of Lemma 4.9.4. Thus,

$$d\, \mathrm{Var}(G(\boldsymbol{\zeta})) = \mathrm{Var}(F(\boldsymbol{W})) \leq 4 \sum_{j=1}^{d} \mathbf{E}\left[\left(\frac{\partial F}{\partial w_j}(\boldsymbol{W})\right)^2\right] = 4\mathbf{E}\|\nabla F(\boldsymbol{W})\|_2^2 \ . \tag{4.5}$$

In order to compute $\nabla F(\boldsymbol{W})$, we observe that for every $i \neq j \in [d]$ we have for all $\boldsymbol{w} \neq \mathbf{0}$ such that $w_i, w_j \neq 0$

$$\frac{\partial T_i}{\partial w_j}(\boldsymbol{w}) = -\frac{w_i \, \mathrm{sign}(w_j)}{\|\boldsymbol{w}\|_1^2} \quad \text{and} \quad \frac{\partial T_i}{\partial w_i}(\boldsymbol{w}) = \frac{1}{\|\boldsymbol{w}\|_1} - \frac{w_i \, \mathrm{sign}(w_i)}{\|\boldsymbol{w}\|_1^2} \ .$$

Thus, the Jacobi matrix of $\boldsymbol{T}(\boldsymbol{w})$ has the form

$$\mathbf{J}_{\boldsymbol{T}}(\boldsymbol{w}) = \frac{\mathbf{I}}{\|\boldsymbol{w}\|_1} - \frac{\boldsymbol{w}(\mathrm{sign}(\boldsymbol{w}))^\top}{\|\boldsymbol{w}\|_1^2} = \frac{1}{\|\boldsymbol{w}\|_1}\Big(\mathbf{I} - \boldsymbol{T}(\boldsymbol{w})\big(\mathrm{sign}(\boldsymbol{w})\big)^\top\Big) \ .$$

It follows that almost surely

$$\nabla F(\boldsymbol{W}) = \frac{1}{2\|\boldsymbol{W}\|_1^{1/2}} G(\boldsymbol{T}(\boldsymbol{W}))\,\mathrm{sign}(\boldsymbol{W}) + \frac{1}{\|\boldsymbol{W}\|_1^{1/2}}\Big(\mathbf{I} - \boldsymbol{T}(\boldsymbol{W})\big(\,\mathrm{sign}(\boldsymbol{W})\big)^\top\Big)\nabla G(\boldsymbol{T}(\boldsymbol{W})) \quad .$$

Observe that since $\langle \mathrm{sign}(\boldsymbol{W})\,, \boldsymbol{T}(\boldsymbol{W})\rangle = 1$ almost surely, we have

$$\big(\,\mathrm{sign}(\boldsymbol{W})\big)^\top \Big(\mathbf{I} - \boldsymbol{T}(\boldsymbol{W})\big(\,\mathrm{sign}(\boldsymbol{W})\big)^\top\Big)\nabla G(\boldsymbol{T}(\boldsymbol{W})) = 0 \quad \text{almost surely} \ .$$

The above two equations imply that almost surely

$$\begin{aligned}4\|\nabla F(\boldsymbol{W})\|_2^2 &= \frac{d}{\|\boldsymbol{W}\|_1} G^2(\boldsymbol{T}(\boldsymbol{W})) + \frac{4}{\|\boldsymbol{W}\|_1}\left\|\Big(\mathbf{I} - \boldsymbol{T}(\boldsymbol{W})\big(\,\mathrm{sign}(\boldsymbol{W})\big)^\top\Big)\nabla G(\boldsymbol{T}(\boldsymbol{W}))\right\|_2^2 \\ &\le \frac{d}{\|\boldsymbol{W}\|_1} G^2(\boldsymbol{T}(\boldsymbol{W})) + \frac{4}{\|\boldsymbol{W}\|_1}\|\nabla G(\boldsymbol{T}(\boldsymbol{W}))\|_2^2\,(1+\sqrt{d}\|\boldsymbol{T}(\boldsymbol{W})\|_2)^2 \ ,\end{aligned}$$

where we used the fact that the operator norm of $\mathbf{I} - \boldsymbol{a}\boldsymbol{b}^\top$ is not greater than $1+\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2$. Combining the above bound with (4.5), and using the facts that $\mathbf{E}[\|\boldsymbol{W}\|_1^{-1}] = \frac{1}{d-1}$, $\mathbf{E}[G(\boldsymbol{T}(\boldsymbol{W}))] = \mathbf{E}[G(\boldsymbol{\zeta})] = 0$ and the independence of $\|\boldsymbol{W}\|_1$ and $\boldsymbol{T}(\boldsymbol{W})$ (cf. (4.4)) yields

$$d\left(1 - \frac{1}{d-1}\right)\mathrm{Var}(G(\boldsymbol{\zeta})) \le \frac{4}{d-1}\mathbf{E}\left[\|\nabla G(\boldsymbol{T}(\boldsymbol{W}))\|_2^2(1+\sqrt{d}\|\boldsymbol{T}(\boldsymbol{W})\|_2)^2\right] \ .$$

Rearranging, we deduce the first claim of the lemma since $\boldsymbol{T}(\boldsymbol{W}) \overset{d}{=} \boldsymbol{\zeta}$.

To prove the second statement of the lemma regarding Lipschitz functions, it is sufficient to apply the first one to $G_n$—the sequence of smoothed versions of $G$ such that $G_n \in C^\infty(\mathbb{R})$ and

$$G_n \longrightarrow G \ ,$$

uniformly on every compact subset, and $\sup_{n\ge1}\|\nabla G_n(\mathbf{x})\|_2 \le L$ for almost all $\mathbf{x} \in \mathbb{R}^d$. A sequence $G_n$ satisfying these properties can be constructed by standard mollification due to the fact that $G$ is Lipschitz continuous (see e.g., Evans and Gariepy, 2018, Theorem 4.2). Finally, to obtain the value $\mathbf{E}\|\boldsymbol{T}(\boldsymbol{W})\|_2^2 = \mathbf{E}\|\boldsymbol{\zeta}\|_2^2$ we use Lemma 4.9.5 below. □

**Lemma 4.9.5.** *Let $\boldsymbol{\zeta}$ be distributed uniformly on $\partial B_1^d$. Then,* $\mathbf{E}\,\|\boldsymbol{\zeta}\|_2^2 = \frac{2}{d+1}$.

*Proof.* We use the same tools as in the proof of Lemma 4.6.2. Let $\boldsymbol{W} = (W_1, \dots, W_d)$ be a vector of i.i.d. random variables following the Laplace distribution with mean $0$ and scale parameter $1$. By (4.4) we have that $\boldsymbol{\zeta} \overset{d}{=} \frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_1}$ and $\boldsymbol{\zeta}$ is independent of $\|\boldsymbol{W}\|_1$. Therefore,

$$\mathbf{E}\|\boldsymbol{\zeta}\|_2^2 = \frac{\mathbf{E}\,\|\boldsymbol{W}\|_2^2}{\mathbf{E}\,\|\boldsymbol{W}\|_1^2} \ . \tag{4.6}$$

Here,

$$\mathbf{E}\,\|\boldsymbol{W}\|_2^2 = \sum_{j=1}^{d} \mathbf{E}[W_j^2] = d\mathbf{E}[W_1^2] = 2d \ .$$

Furthermore, $\|\boldsymbol{W}\|_1$ follows the Erlang distribution with parameters $(d, 1)$, which implies

$$\mathbf{E}\,\|\boldsymbol{W}\|_1^2 = \frac{1}{\Gamma(d)} \int_0^{\infty} x^{d+1} \exp(-x)\,\mathrm{d}x = \frac{\Gamma(d+2)}{\Gamma(d)} \ . \tag{4.7}$$

The lemma follows by combining (4.6) – (4.7). □

## Upper bounds

The proofs of Theorems 4.4.1, 4.4.2, 4.5.1, 4.5.2 resemble each other. They only differ in the ways of handling the variance terms depending on $\|\mathbf{g}_t\|_{p^*}^2$ and in the choice of parameters. For this reason, we suggest the interested reader to follow the proofs in a linear manner starting from the next paragraph.

**Common part of the proofs of Theorems 4.4.1, 4.4.2.** We start with the part of the proofs that is common for Theorems 4.4.1, 4.4.2. Fix some $\mathbf{x} \in \Theta$. Due to Assumption 4.3.1, we can use Lemma 4.6.1, which implies

$$\mathbf{E}\left[\sum_{t=1}^{T} \langle \mathbf{E}\left[\mathbf{g}_t \mid \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}\rangle\right] = \mathbf{E}\left[\sum_{t=1}^{T} \langle \nabla \mathsf{f}_{t,h}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}\rangle\right] \geq \mathbf{E}\left[\sum_{t=1}^{T} \left(\mathsf{f}_{t,h}(\mathbf{x}_t) - \mathsf{f}_{t,h}(\mathbf{x})\right)\right] ,$$

where $\mathsf{f}_{t,h}(\mathbf{x}) = \mathbf{E}[f_t(\mathbf{x} + h\boldsymbol{U})]$ with $\boldsymbol{U}$ uniformly distributed on $B_1^d$. Furthermore, by the approximation property derived in Lemma 4.6.1 and the standard bound on the cumulative regret of dual averaging algorithm (see e.g., Orabona, 2019, Corollary 7.9.) we deduce that

$$\begin{aligned}\mathbf{E}\left[\sum_{t=1}^{T} \left(f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right)\right] &\leq \mathbf{E}\left[\sum_{t=1}^{T} \langle \mathbf{E}\left[\mathbf{g}_t | \mathbf{x}_t\right], \mathbf{x}_t - \mathbf{x}\rangle\right] + L\mathrm{b}_q(d)\sum_{t=1}^{T} h_t \\ &\leq \frac{R^2}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \mathbf{E}\,\|\mathbf{g}_t\|_{p^*}^2 + L\mathrm{b}_q(d)\sum_{t=1}^{T} h_t \ ,\end{aligned} \tag{4.8}$$

where in the last inequality we used the identity $\eta_1 = \ldots = \eta_T = \eta$. The results of Theorems 4.4.1, 4.4.2 follow from the bound (4.8) as detailed below.

*Proof of Theorem 4.4.1.* Here $h_1 = \ldots = h_T = h$, and we work under Assumption 4.3.4. In this case, bounding $\mathbf{E}\|\mathbf{g}_t\|_{p^*}$ in (4.8) via Lemma 4.6.5 yields

$$\mathbf{E}\left[\sum_{t=1}^{T} \left(f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right)\right] \leq \frac{R^2}{\eta} + 6(1+\sqrt{2})^2 L^2 \cdot \eta T d^{1+\frac{2}{q \wedge 2} - \frac{2}{p}} + LhT\mathrm{b}_q(d) \ .$$

Minimizing the the right hand side of the above inequality over $\eta > 0$ and substituting $\eta = \frac{R}{L(\sqrt{6}+\sqrt{12})}\sqrt{\frac{d^{-1-\frac{2}{q\wedge 2}+\frac{2}{p}}}{T}}$ we deduce that

$$\mathbf{E}\left[\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right)\right] \leq 2\left(\sqrt{6}+\sqrt{12}\right) RLd^{\frac{1}{2}+\frac{1}{q\wedge 2}-\frac{1}{p}}\sqrt{T} + LhT\mathrm{b}_q(d) \enspace .$$

Taking $h \leq \frac{7R}{100\mathrm{b}_q(d)\sqrt{T}} d^{\frac{1}{2}+\frac{1}{q\wedge 2}-\frac{1}{p}}$ makes negligible the second summand in the above bound. This concludes the proof. □

*Proof of Theorem 4.4.2.* Here again $h_1 = \ldots = h_T = h$, but we work under Assumption 4.3.5. Then, bounding $\mathbf{E}\|\mathbf{g}_t\|_{p^*}$ in (4.8) via Lemma 4.6.5 yields

$$\mathbf{E}\left[\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right)\right] \leq \frac{R^2}{\eta} + \eta T\left(\frac{d^{4-\frac{2}{p}}\sigma^2}{h^2} + 6\left(1+\sqrt{2}\right)^2 L^2 d^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}\right) + LhT\mathrm{b}_q(d) \enspace .$$

Minimizing the right hand side of the above inequality over $\eta > 0$ and substituting the optimal value

$$\eta = \frac{R}{\sqrt{T}}\left(\frac{d^{4-\frac{2}{p}}\sigma^2}{2h^2} + 6\left(1+\sqrt{2}\right)^2 L^2 d^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}\right)^{-\frac{1}{2}} \enspace ,$$

results in the following upper bound on the regret

$$\begin{aligned}\mathbf{E}\left[\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right)\right] &\leq 2R\sqrt{T}\left(\frac{d^{4-\frac{2}{p}}\sigma^2}{2h^2} + 6\left(1+\sqrt{2}\right)^2 L^2 d^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}\right)^{\frac{1}{2}} + LhT\mathrm{b}_q(d)\\ &\leq 2\left(\sqrt{6}+\sqrt{12}\right) RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}} + \sqrt{2}R\sqrt{T}\frac{d^{2-\frac{1}{p}}\sigma}{h} + LhT\mathrm{b}_q(d) \enspace ,\end{aligned}$$

where for the last inequality we used the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. Minimizing over $h > 0$ the last expression and substituting the optimal value $h = \left(\frac{\sqrt{2}R\sigma}{L\mathrm{b}_q(d)}\right)^{\frac{1}{2}} T^{-\frac{1}{4}} d^{1-\frac{1}{2p}}$ we get

$$\mathbf{E}\left[\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right)\right] \leq 11.9RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}} + 2.4\sqrt{RL\sigma}T^{\frac{3}{4}}\sqrt{\mathrm{b}_q(d)}d^{\frac{1}{2}-\frac{1}{2p}}.\square$$

**Common part of the proofs of Theorems 4.5.1, 4.5.2.** Here, we state the common parts of the proofs for Theorems 4.5.1, 4.5.2. Similar to the first inequality in (4.8), we have

$$\mathbf{E}\left[\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right)\right] \leq \mathbf{E}\left[\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}\rangle\right] + L\mathrm{b}_q(d)\sum_{t=1}^{T} h_t \enspace .$$

Note that without loss of generality, we can assume that $\sum_{k=1}^{t}\|\mathbf{g}_k\|_{p^*}^2 \neq 0$, for all $t \geq 1$. This is a consequence of the fact that if $\sum_{k=1}^{t}\|\mathbf{g}_k\|_{p^*}^2 = 0$, then the first term on the r.h.s. of the above inequality will be zero up to round $t$. Thus, we can erase these iterates from the cumulative regret, only paying the bias term for those rounds. In what follows we essentially use (Orabona and Pál, 2016, Corollary 1), which we re-derive for the sake of clarity. Assume that $\eta_t = \frac{\lambda}{\sqrt{\sum_{k=1}^{t-1} \|\mathbf{g}_k\|_{p^*}^2}}$ for $t \in \{2, \dots, T\}$ and $\lambda > 0$. Then, applying (Orabona and Pál, 2016, Theorem 1) we deduce that

$$\mathbf{E}\left[\sum_{t=1}^{T} (f_t(\mathbf{x}_t) - f_t(\mathbf{x}))\right] \leq \left(\frac{R^2}{\lambda} + 2.75 \cdot \lambda\right) \mathbf{E}\left[\sqrt{\sum_{t=1}^{T} \|\mathbf{g}_t\|_{p^*}^2}\right] \\ + 3.5D \cdot \mathbf{E}[\max_{t\in[T]} \|\mathbf{g}_t\|_{p^*}] + L\mathrm{b}_q(d) \sum_{t=1}^{T} h_t \enspace ,$$

where we introduced $D = \sup_{\boldsymbol{u},\boldsymbol{w}\in\Theta} \|\boldsymbol{u} - \boldsymbol{w}\|_p$. By (Orabona and Pál, 2016, Proposition 1), we have $D \leq \sqrt{8}R$. Moreover, by Jensen's inequality, using the rough bound $\mathbf{E}[\max_{t\in[T]} \|\mathbf{g}_t\|_{p^*}] \leq \sqrt{\sum_{t=1}^{T} \mathbf{E}\left[\|\mathbf{g}_t\|_{p^*}^2\right]}$, and substituting $\lambda = \frac{R}{\sqrt{2.75}}$, we deduce that

$$\mathbf{E}\left[\sum_{t=1}^{T} (f_t(\mathbf{x}_t) - f_t(\mathbf{x}))\right] \leq \left(2\sqrt{2.75} + 3.5\sqrt{8}\right) R\sqrt{\sum_{t=1}^{T} \mathbf{E}\left[\|\mathbf{g}_t\|_{p^*}^2\right]} + L\mathrm{b}_q(d) \sum_{t=1}^{T} h_t \enspace . \tag{4.9}$$

Proofs of Theorems 4.5.1, 4.5.2 provided below follow from the above inequality by properly selecting $h_t > 0$.

*Proof of Theorem 4.5.1.* The bound of Lemma 4.6.5 under Assumption 4.3.4 applied to (4.9) yields

$$\begin{aligned}\mathbf{E}\left[\sum_{t=1}^{T} (f_t(\mathbf{x}_t) - f_t(\mathbf{x}))\right] &\leq 2\left(2\sqrt{2.75} + 3.5\sqrt{8}\right)\left(\sqrt{3} + \sqrt{6}\right) RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}} + L\mathrm{b}_q(d) \sum_{t=1}^{T} h_t \\ &\leq 110.53 \cdot RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}} + L\mathrm{b}_q(d) \sum_{t=1}^{T} h_t \enspace .\end{aligned}$$

Taking $h_t \leq \frac{7R}{200\mathrm{b}_q(d)\sqrt{t}} d^{\frac{1}{2}+\frac{1}{q\wedge 2}-\frac{1}{p}}$ makes negligible the last summand in the above bound. This concludes the proof. □

*Proof of Theorem 4.5.2.* Using (4.9), the bound of Lemma 4.6.5 under Assumption 4.3.5 and the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we deduce that

$$\begin{aligned}\mathbf{E}\left[\sum_{t=1}^{T}\big(f_t(\mathbf{x}_t) - f_t(\mathbf{x})\big)\right] &\leq \left(2\sqrt{2.75} + 3.5\sqrt{8}\right) R \left(\sum_{t=1}^{T} \frac{d^{4-\frac{2}{p}}\sigma^2}{h_t^2} + 12(1+\sqrt{2})^2 L^2 T \cdot d^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}\right)^{\frac{1}{2}} \\ &\quad + L\mathrm{b}_q(d)\sum_{t=1}^{T} h_t \\ &\leq 110.6 \cdot RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}} + 13.3R \cdot d^{2-\frac{1}{p}}\sigma \left(\sum_{t=1}^{T}\frac{1}{h_t^2}\right)^{\frac{1}{2}} \\ &\quad + L\mathrm{b}_q(d)\sum_{t=1}^{T} h_t \enspace .\end{aligned}$$

Since $h_t = \left(6.65\sqrt{6} \cdot \frac{R}{\mathrm{b}_q(d)}\right)^{\frac{1}{2}} t^{-\frac{1}{4}} d^{1-\frac{1}{2p}}$ and $\sum_{t=1}^{T} t^{\frac{1}{2}} \leq \frac{2}{3}T^{\frac{3}{2}}$ and $\sum_{t=1}^{T} t^{-\frac{1}{4}} \leq \frac{4}{3}T^{\frac{3}{4}}$, we get

$$\mathbf{E}\left[\sum_{t=1}^{T}\big(f_t(\mathbf{x}_t) - f_t(\mathbf{x})\big)\right] \leq 110.6 \cdot RL\sqrt{Td^{1+\frac{2}{q\wedge 2}-\frac{2}{p}}} + 5.9 \cdot \sqrt{R}\,(\sigma + L)\,T^{\frac{3}{4}}\sqrt{\mathrm{b}_q(d)}d^{\frac{1}{2}-\frac{1}{2p}} \enspace . \square$$

## Definition of $\ell_2$-randomized estimator

In this section we recall the algorithm of Shamir (2017). Let $\boldsymbol{\zeta}^{\circ} \in \mathbb{R}^d$ be distributed uniformly on $\partial B_2^d$. Instead of the gradient estimator that we introduce in Algorithm 4, at a each step $t \geq 1$, Shamir (2017) uses

$$\mathbf{g}_t^{\circ} \triangleq \frac{d}{2h}(y_t' - y_t'')\boldsymbol{\zeta}_t^{\circ} \enspace ,$$

where $y_t' = f_t(\mathbf{x}_t + h_t\boldsymbol{\zeta}^{\circ})$, $y_t'' = f_t(\mathbf{x}_t - h_t\boldsymbol{\zeta}_t^{\circ})$, and $\boldsymbol{\zeta}_t^{\circ}$'s are independent random variables with the same distribution as $\boldsymbol{\zeta}^{\circ}$.

# Chapter 5

# Zero-order optimization of highly smooth functions: improved analysis and a new algorithm

This chapter studies minimization problems with zero-order noisy Oracle information under the assumption that the objective function is highly smooth and possibly satisfies additional properties. We consider two kinds of zero-order projected gradient descent algorithms, which differ in the form of the gradient estimator. The first algorithm uses a gradient estimator based on randomization on the $\ell_2$ sphere. The precise form that we consider is due to Bach and Perchet (2016) and it has been used for zero-order optimization of strongly convex functions. We present an improved analysis of this algorithm for the same class of functions and we derive rates of convergence for more general function classes. In particular, we consider functions which satisfies the Polyak-Łojasiewicz condition instead of strong convexity, and the larger class of highly smooth non-convex functions. The second algorithm is based on $\ell_1$-type randomization, and it extends the recently proposed algorithm of Akhavan et al. (2022a) who dealt with Lipschitz convex functions. We show that this novel algorithm enjoys similar theoretical guarantees than the first one and, in the case of noiseless Oracle, it enjoys better bounds. The improvements are achieved by new bounds on bias and variance for both algorithms, which are obtained via Poincaré type inequalities for uniform distributions on $\ell_1$ or $\ell_2$ spheres. The optimality of the upper bounds is discussed and a slightly more general lower bound than the state-of-the art bound in Akhavan et al. (2020) is presented.

## 5.1 Introduction

In this work, we study the problem of zero-order optimization for certain types of smooth functions. Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\Theta \subset \mathbb{R}^d$, we are interested in solving the following optimization problem

$$f^\star \triangleq \inf_{\mathbf{x} \in \Theta} f(\mathbf{x}) \enspace ,$$

and we assume that $f^\star$ is finite. One main theme of this paper is to exploit higher order smoothness properties of the underlying function $f$ in order to improve the performance of the optimization algorithm. We consider that the algorithm has access to a zero-order stochastic Oracle, which, given a point $\mathbf{x} \in \mathbb{R}^d$ returns a noisy value of $f(\mathbf{x})$, under a general noise model.

We study two kinds of zero-order projected gradient descent algorithms, which differ in the form of the gradient estimator. Both algorithms can be written as an iterative update of the form

$$\mathbf{x}_1 \in \mathbb{R}^d \qquad \text{and} \qquad \mathbf{x}_{t+1} = \mathrm{Proj}_{\Theta}(\mathbf{x}_t - \eta_t \mathbf{g}_t) \qquad t \geq 1 \enspace ,$$

where $\mathbf{g}_t$ is a gradient estimator at the point $\mathbf{x}_t$, $\eta_t$ is a step-size, and $\mathrm{Proj}_{\Theta}(\cdot)$ is the Euclidean projection operator onto the set $\Theta$. In either case, the gradient estimator is built from two noisy function values, that are queried at two random perturbations of the current guess for the solution, and it involves an additional randomization step. The first algorithm uses a form of $\ell_2$-randomization, and it has been used previously in the literature (see e.g. Akhavan et al., 2020; Bach and Perchet, 2016; Novitskii and Gasnikov, 2021; Polyak and Tsybakov, 1990). The precise form of the gradient estimator that we consider here has been introduced by Bach and Perchet (2016) and it has been used for zero-order optimization of strongly convex functions. The second algorithm is a modification of the approach proposed and analysed in Akhavan et al. (2022a) for online minimization of Lipschitz convex functions. It is based on an alternative randomization scheme, which uses $\ell_1$-geometry in place of the $\ell_2$ one.

A principal goal of this paper is to derive sharper upper bounds for both algorithms under different assumptions on the underlying function $f$ that we wish to minimize. These assumptions are used to set the step size in the algorithms and the perturbation parameter used inside the gradient estimator. Previous works considered mostly the strongly convex case (Akhavan et al., 2020; Bach and Perchet, 2016; Novitskii and Gasnikov, 2021; Polyak and Tsybakov, 1990) and in this paper we provide a refined analysis, improving the dependency on the dimension derived by (Akhavan et al., 2020; Novitskii and Gasnikov, 2021). Furthermore, we complement these results by considering the cases of smooth $f$ (not necessary convex), and smooth $f$, which additionally satisfies the gradient dominance condition. For the new algorithm, we establish similar results discussed above and highlight the improvement in the noiseless case.

### Summary of the upper bounds

In this subsection, we give a high-level overview of the main contribution of this work. Apart from the improved guarantees for the previously studied function classes, one of the main novelties of our work is the analysis in the case of a non-convex smoothness objective function $f$, for which we provide a convergence rate to a stationary point. Furthermore, we study the case of $\alpha$-gradient dominant $f$—a popular relaxation of strong convexity, which includes non-convex functions. To the best of our knowledge, the analysis of stochastic zero-order optimization in these two cases is novel. In Section 5.5 we derive lower bounds and discuss the (sub)-optimality of our convergence rates.

In the following we highlight the guarantee that we derive for the two analysed algorithms. Each of the guarantees differs in the dependency on the main parameters of the problem, which is a consequence of the different types of available properties of the objective function. Let us also mention that we mainly deal with the unconstrained optimization case, $\Theta = \mathbb{R}^d$. This is largely due to the fact that the Polyak-Łojasiewicz inequality is mainly used in the unconstrained case and the exertion of this condition to the constrained case is still an active area of research (see e.g., Balashov et al., 2020, and references therein). Meanwhile, for the strongly convex case, as in previous works (Akhavan et al., 2020; Bach and Perchet, 2016; Novitskii and Gasnikov, 2021), we additionally treat the constrained optimization. In this section we only sketch our results in the case $\Theta = \mathbb{R}^d$.

**Rate of convergence under only smoothness assumption.** Assume that $f$ is $\beta$-Hölder with Lipschitz continuous gradient. Then, after $2T$ Oracle queries both considered algorithms provide a point $\mathbf{x}_S$ satisfying

$$\mathbf{E}\left[\|\nabla f(\mathbf{x}_S)\|^2\right] \lesssim \left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}} \text{ under the assumption that } T \geq d^{\frac{1}{\beta}} \enspace ,$$

where $S$ is a random variable with values in $\{1, \cdots, T\}$, and $\lesssim$ conceals the multiplicative constants that do not depend on $T$ and $d$. To the best of our knowledge, this result is the first convergence guarantee for the zero order stochastic optimization under the considered noise model. Balasubramanian and Ghadimi (2021); Ghadimi and Lan (2013) consider zero-order optimization of non-convex objective function with Lipschitz gradient. They allow querying two function values with identical noises, effectively reducing the convergence analysis to the non-stochastic case. Carmon et al. (2017) study deterministic optimization of highly smooth functions assuming that the higher order derivatives are observed, and Arjevani et al. (2022)

consider stochastic optimization with first order Oracle. Thus, a direct comparison of our results with theirs is not possible.

**Rate of convergence under smoothness and Polyak-Łojasiewicz assumptions.** Assume that $f$ is $\beta$-Hölder with Lipschitz continuous gradient and satisfies $\alpha$-Polyak-Łojasiewicz inequality. Then, after $2T$ Oracle queries, both considered algorithms provide a point $\mathbf{x}_T$ for which the expected optimization error is upper bounded as

$$\mathbf{E}[f(\mathbf{x}_T) - f^\star] \lesssim \frac{d}{\alpha T} + \frac{1}{\alpha \wedge \alpha^2}\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{\beta}} \quad \text{under the assumption that } T \geq d^{2-\frac{\beta}{2}} \enspace ,$$

where $\lesssim$ conceals the multiplicative constants that do not depend on $T$, $d$ and $\alpha$. The Polyak-Łojasiewicz assumption, was considered in the context of first order optimization by Polyak (1963), who derived linear convergence of the gradient descent algorithm. Years later, this condition received attention in the machine learning and optimization community following the work of Karimi et al. (2016). To the best of our knowledge, zero-order optimization under the considered noise model with the Polyak-Łojasiewicz assumption has not previously studied. Very recently Rando et al. (2022) studied a related problem under the Polyak-Łojasiewicz assumption. Unlike our work, they deploy a per-coordinate (random) gradient estimator in the spirit of Akhavan et al. (2021), treat $\alpha$ as constant, and do not consider high-order smoothness.

**Rate of convergence under smoothness and strong convexity.** Assume that $f$ is $\beta$-Hölder with Lipschitz continuous gradient and satisfies $\alpha$-strong convexity. Then, after $2T$ Oracle queries, both considered algorithms provide a point $\mathbf{x}_T$ such that

$$\mathbf{E}[f(\mathbf{x}_T) - f^\star] \lesssim \frac{d}{\alpha T} + \frac{1}{\alpha}\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{\beta}} \quad \text{under the assumption that } T \geq d^{2-\frac{\beta}{2}} \enspace ,$$

where $\lesssim$ conceals the multiplicative constants that do not depend on $T$, $d$ and $\alpha$. The closest result to ours is that Akhavan et al. (2020) who split the proof into two cases: $\beta = 2$ (Lipschitz continuous gradient) and $\beta > 2$ (higher order smoothness). In the former case, they obtain optimal dependency (linear in $d$) on the dimension, while in the latter case they get $d^2$. Later, Novitskii and Gasnikov (2021) and Akhavan et al. (2021), for the case $\beta > 2$, improved this dependency to $d^{2-1/\beta}$, which still does not match with the linear dependency for $\beta = 2$. In contrast, we provide a unified analysis leading to $d^{(2\beta\,-\,2)/\beta}$ dependency for any $\beta \geq 2$; the improvement is both in the rate and in the proof technique.

## Notation

Throughout the paper, we use the following notation. For any $k \in \mathbb{N}$ we denote by $[k]$, the set of first $k$ positive integers. For any $\mathbf{x} \in \mathbb{R}^d$ we denote by $\mathbf{x} \mapsto \operatorname{sign}(\mathbf{x})$ the component-wise sign function (defined at $0$ as $1$). We let $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ be the standard inner product and Euclidean norm on $\mathbb{R}^d$, respectively. For every close convex set $\Theta \subset \mathbb{R}^d$ and $\mathbf{x} \in \mathbb{R}^d$ we denote by $\mathrm{Proj}_\Theta(\mathbf{x}) = \mathrm{argmin}\{\|\mathbf{z} - \mathbf{x}\| \,:\, \mathbf{z} \in \Theta\}$ the Euclidean projection of $\mathbf{x}$ onto $\Theta$. For any $p \in [1, +\infty]$ we let $\|\cdot\|_p$ be the $\ell_p$-norm in $\mathbb{R}^d$ and introduce the open $\ell_p$-ball and $\ell_p$-sphere respectively as

$$\mathcal{B}_p^d \triangleq \left\{\mathbf{x} \in \mathbb{R}^d \,:\, \|\mathbf{x}\|_p < 1\right\} \qquad \text{and} \qquad \partial B_p^d \triangleq \left\{\mathbf{x} \in \mathbb{R}^d \,:\, \|\mathbf{x}\|_p = 1\right\} \enspace .$$

For any $\beta \geq 2$ we let $\lfloor \beta \rfloor$ the largest integer which is strictly less than $\beta$. Given multi-index $\boldsymbol{m} = (m_1, \ldots, m_d) \in \mathbb{N}^d$, we set $\boldsymbol{m}! \triangleq m_1! \cdots m_d!$, $|\boldsymbol{m}| \triangleq m_1 + \cdots + m_d$.

### Structure of the paper

The paper, is organized as following. In Section 5.2, we recall some preliminaries and introduce the classes of functions considered throughout. In Section 5.3, we presents the two algorithms that are studied in the paper. In Section 5.4, we present the upper bounds for both algorithms, and in each of the considered function classes. In Section 5.5, we establish minimax lower bounds for the zero-order optimization problem. Finally in Section 5.6, we discuss our results and compare them to previous related work. The proof of most of the results are presented in Section 5.7.

## 5.2 Preliminaries

For any multi-index $\boldsymbol{m} \in \mathbb{N}^d$, any $|\boldsymbol{m}|$-times continuous differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, and every $\boldsymbol{h} = (h_1, \ldots, h_d)^\top \in \mathbb{R}^d$ we define

$$D^{\boldsymbol{m}} f(\mathbf{x}) \triangleq \frac{\partial^{|\boldsymbol{m}|} f(\mathbf{x})}{\partial^{m_1} x_1 \cdots \partial^{m_d} x_d} \,, \qquad \boldsymbol{h}^{\boldsymbol{m}} \triangleq h_1^{m_1} \cdots h_d^{m_d} \enspace .$$

For any $k$-linear form $A : \left(\mathbb{R}^d\right)^k \to \mathbb{R}$ define its norm as

$$\|A\| \triangleq \sup \left\{|A[\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k]| \,:\, \|\boldsymbol{h}_j\| \leq 1,\ j \in [k]\right\} \enspace .$$

Whenever $\boldsymbol{h}_1 = \ldots = \boldsymbol{h}_k = \boldsymbol{h}$ we write $A[\boldsymbol{h}]^k$ to denote $A[\boldsymbol{h}, \ldots, \boldsymbol{h}]$. Given a $k$-times continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$ we denote by $f^{(k)}(\mathbf{x}) : \left(\mathbb{R}^d\right)^k \to \mathbb{R}$ the following $k$-linear form

$$f^{(k)}(\mathbf{x})[\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k] = \sum_{|\boldsymbol{m}_1| = \cdots = |\boldsymbol{m}_k| = 1} D^{\boldsymbol{m}_1 + \cdots + \boldsymbol{m}_k} f(\mathbf{x}) \boldsymbol{h}_1^{\boldsymbol{m}_1} \cdots \boldsymbol{h}_k^{\boldsymbol{m}_k} \,, \quad \forall \boldsymbol{h}_1, \ldots, \boldsymbol{h}_k \in \mathbb{R}^d \enspace ,$$

where $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k \in \mathbb{N}^d$. We note that since $f$ is $k$-times continuously differentiable in $\mathbb{R}^d$, then $f^{(k)}(\mathbf{x})$ is symmetric for all $\mathbf{x} \in \mathbb{R}^d$.

### Classes of functions

We start this section by stating all the relevant definitions and assumptions related to the target function $f$. Following (Nemirovski, 2000, Section 1.3) we recall the definition of high order Hölder smoothness, which was also considered by Bach and Perchet (2016).

**Definition 5.2.1** (Higher order smoothness)**.** *Fix some $\beta \geq 2$ and $L > 0$. Denote by $\mathcal{F}_\beta(L)$ the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that are $\ell = \lfloor \beta \rfloor$ times continuously differentiable and satisfy, for all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$ the Hölder-type condition*

$$\left\| f^{(\ell)}(\boldsymbol{x}) - f^{(\ell)}(\boldsymbol{z}) \right\| \leq L \left\| \boldsymbol{x} - \boldsymbol{z} \right\|^{\beta - \ell} \enspace .$$

**Remark 5.2.2** (On the definition of the class)**.** *Akhavan et al. (2020) consider a slightly different definition of higher order smoothness. Namely, they consider a class $\mathcal{F}'_\beta(L')$ defined as $\ell$-times continuously differentiable functions $f$ satisfying for all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$*

$$|f(\boldsymbol{x}) - T^{\ell}_{\boldsymbol{z}}(\boldsymbol{x})| \leq L' \left\| \boldsymbol{x} - \boldsymbol{z} \right\|^{\beta} \enspace ,$$

*where $T^{\ell}_{\boldsymbol{z}}(\cdot)$ is the Taylor polynomial of order $\ell$ of $f$ around $\boldsymbol{z}$. In Section 5.7 we show that if $f \in \mathcal{F}_\beta(L)$, then $f \in \mathcal{F}'_\beta(L/\ell!)$. That is to say, the functional class considered by Akhavan et al. (2020) is not smaller. Note however that if $f$ is convex and $\beta = 2$, then our class coincides with that of Akhavan et al. (2020)—the class of functions with Lipschitz continuous gradient.*

Since we study the minimization of highly smooth functions, in what follows, we will always assume that $f$ belongs to $\mathcal{F}_\beta(L)$ for some $\beta \geq 2$ and $L > 0$. We additionally require that the function $f \in \mathcal{F}_2(\bar{L})$ for some $\bar{L} > 0$, that is, its gradient is Lipschitz continuous.

**Assumption 5.2.3.** *The function $f \in \mathcal{F}_\beta(L) \cap \mathcal{F}_2(\bar{L})$ for some $\beta \geq 2$ and $L, \bar{L} > 0$.*

We will start our analysis by providing rates of convergence to a stationary point of $f$ under Assumption 5.2.3. The first additional assumption that we consider is the Polyak-Łojasiewicz condition, which we refer to as $\alpha$-gradient dominance. This condition became rather popular since it leads to linear convergence of the gradient descent algorithm, without convexity (see, *e.g.*, Karimi et al., 2016; Polyak, 1963).

**Definition 5.2.4** ($\alpha$-gradient dominance)**.** *Let $\alpha > 0$. Function $f : \mathbb{R}^d \to \mathbb{R}$ is called $\alpha$-gradient dominant on $\mathbb{R}^d$, if $f$ is differentiable on $\mathbb{R}^d$ and satisfies Polyak-Łojasiewicz inequality,*

$$2\alpha(f(\boldsymbol{x}) - f^\star) \leq \left\| \nabla f(\boldsymbol{x}) \right\|^2 , \qquad \forall \boldsymbol{x} \in \mathbb{R}^d \enspace .$$

Finally, we consider the second additional condition, which is the $\alpha$-strong convexity.

**Definition 5.2.5** ($\alpha$-strong convexity)**.** *Let $\alpha > 0$. Function $f : \mathbb{R}^d \to \mathbb{R}$ is called $\alpha$-strongly convex on $\mathbb{R}^d$, if it is differentiable on $\mathbb{R}^d$ and satisfies*

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \left\langle \nabla f(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \right\rangle + \frac{\alpha}{2} \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|^2 , \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d \enspace .$$

Under this condition and Assumption 5.2.3, we are in the framework previously considered in Akhavan et al. (2020); Bach and Perchet (2016); Novitskii and Gasnikov (2021); Polyak and Tsybakov (1990). We will provide bounds improving upon these results.

Note that an important example of family of functions satisfying the $\alpha$-dominance condition is given by composing strongly convex functions with a linear transformation. Let $n \in \mathbb{N}$, $\mathbf{A} \in \mathbb{R}^{n\times d}$ and define

$$\mathcal{F}(\mathbf{A}) = \left\{f \,:\, f(\mathbf{x}) = g(\mathbf{A}\mathbf{x}),\ \text{g is } \alpha\text{-strongly convex}\right\} \enspace .$$

Note that if $\mathbf{A}^\top\mathbf{A}$ is not invertible then the functions in $\mathcal{F}(\mathbf{A})$ are not necessarily strongly convex. However, it can be shown that any $f \in \mathcal{F}(\mathbf{A})$ is an $\alpha\gamma$-gradient dominant function, where $\gamma$ is the smallest non-zero singular value of $A$ (see, *e.g.,* Karimi et al., 2016). Alternatively, we can consider the following family of functions

$$\mathcal{F}'(\mathbf{A}) = \left\{f \,:\, f(\mathbf{x}) = g(\mathbf{A}\mathbf{x}), \quad g \in C^2(\mathbb{R}^d), \quad g \text{ strictly convex}\right\} \enspace ,$$

which is a set of $\alpha$-gradient dominant functions on any compact subset of $\mathbb{R}^d$, for some $\alpha > 0$. A popular example of such a function, appearing in machine learning applications, is the logistic loss, defined as $g(\mathbf{A}\mathbf{x}) = \sum_{i=1}^m \log(1 + \exp(\boldsymbol{a}_i^\top\mathbf{x}))$, where for $1 \leq i \leq n$, $\boldsymbol{a}_i$ is $i$-th row of $\mathbf{A}$, and $\mathbf{x} \in \mathbb{R}^d$. For this and more examples, see e.g. (Garrigos et al., 2022) and references therein.

In what follows we will consider three different scenarios: *i)* the case of only smoothness assumption on $f$ *ii)* additional $\alpha$-gradient dominance *iii)* additional $\alpha$-strong convexity. Let $\hat{\mathbf{x}}$ be an output of any algorithm. For the first scenario we study stationary point guarantee and bound for $\mathbf{E}\left\|\nabla f(\hat{\mathbf{x}})\right\|^2$. For the second and the third we will consider optimization error: $\mathbf{E}[f(\hat{\mathbf{x}}) - f^\star]$. Note that under strong convexity (as long as $\nabla f(\mathbf{x}^*) = 0$) as well as under $\alpha$-dominance gradient (see, *e.g.,* Karimi et al., 2016, Appendix A.), for any $\mathbf{x} \in \mathbb{R}^d$

$$f(\mathbf{x}) - f^\star \geq \frac{\alpha}{2}\left\|\mathbf{x} - \mathbf{x}_p^*\right\|^2 \enspace ,$$

where $\mathbf{x}_p^*$ is the Euclidean projection of $\mathbf{x}$ onto the set $\arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$—solution set of the considered optimization problem—which is a singleton in case of strong convexity. Thus, for the last two scenarios, our upper bounds on $\mathbf{E}[f(\hat{\mathbf{x}}) - f^\star]$ imply immediately upper bounds for $\left\|\mathbf{x} - \mathbf{x}_p^*\right\|^2$ multiplying the former by $2/\alpha$.

## 5.3 Algorithms

Given a closed and convex set $\Theta \subseteq \mathbb{R}^d$, we consider the following optimization scheme

$$\mathbf{x}_1 \in \Theta \qquad \text{and} \qquad \mathbf{x}_{t+1} = \operatorname{Proj}_{\Theta}(\mathbf{x}_t - \eta_t\mathbf{g}_t) \qquad t \geq 1 \enspace , \tag{5.1}$$

where $\mathbf{g}_t$ is an update direction, approximating the gradient direction $\nabla f(\mathbf{x}_t)$ and $\eta_t > 0$ is a step-size. Allowing one to perform two function evaluations per step, we consider two gradient estimators $\mathbf{g}_t$ which are based on different randomization schemes. They both employ a smoothing kernel $K : [-1, 1] \to \mathbb{R}$ which we assume to satisfy, for $\beta \geq 2$ and $\ell = \lfloor \beta \rfloor$, the conditions

$$\int K(r)\,\mathrm{d}r{=}0,\int rK(r)\,\mathrm{d}r{=}1,\int r^j K(r)\,\mathrm{d}r{=}0,\ j{=}2,\dots,\ell,\ \kappa_\beta{\triangleq}\int |r|^\beta |K(r)|\,\mathrm{d}r < \infty\ .$$

In (Polyak and Tsybakov, 1990) it was suggested to construct such kernels employing Legendre polynomials, in which case $\kappa_\beta \leq 2\sqrt{2}\beta$, cf. Bach and Perchet, 2016, Appendix A.3.

We are now in a position to introduce the two estimators. Similarly to earlier works on zero-order stochastic optimization (see e.g., Akhavan et al., 2022a; Bach and Perchet, 2016; Flaxman et al., 2005; Nemirovsky and Yudin, 1983; Novitskii and Gasnikov, 2021) we use gradient estimators based on a result, which is sometimes referred to as Stokes' theorem. A general form of this result can be found in Akhavan et al., 2022a, Appendix A.

**Gradient estimator based on $\ell_2$-randomization.** At time $t \geq 1$, let $\zeta_t^\circ$ be distributed uniformly on $\partial B_2^d$, $r_t$ uniformly distributed on $[-1, 1]$, and $h_t > 0$. Query two points:

$$y_t = f(\mathbf{x}_t + h_t r_t \zeta_t^\circ) + \xi_t \qquad \text{and} \qquad y_t' = f(\mathbf{x}_t - h_t r_t \zeta_t^\circ) + \xi_t'\ .$$

Using the above feedback, define the gradient estimate as

$$(\ell_2\texttt{-randomization}) \qquad \mathbf{g}_t^\circ \triangleq \frac{d}{2h_t}(y_t - y_t')\zeta_t^\circ K(r_t)\ . \tag{5.2}$$

We use the superscript $\circ$ to emphasize the fact that $\mathbf{g}_t^\circ$ is based on the $\ell_2$-randomization.

**Gradient estimator based on $\ell_1$-randomization.** At time $t \geq 1$, let $\zeta_t^\diamond$ be distributed uniformly on $\partial B_1^d$, $r_t$ uniformly distributed on $[-1, 1]$, and $h_t > 0$. Query two points:

$$y_t = f(\mathbf{x}_t + h_t r_t \zeta_t^\diamond) + \xi_t \qquad \text{and} \qquad y_t' = f(\mathbf{x}_t - h_t r_t \zeta_t^\diamond) + \xi_t'\ .$$

Using the above feedback, define the gradient estimate as

$$(\ell_1\texttt{-randomization}) \qquad \mathbf{g}_t^\diamond \triangleq \frac{d}{2h_t}(y_t - y_t')\,\mathrm{sign}(\zeta_t^\diamond)K(r_t)\ . \tag{5.3}$$

Similarly to the previous algorithm, we use the superscript $\diamond$ to emphasize the fact that $\mathbf{g}_t^\diamond$ is based on the $\ell_1$-randomization. We refer to Akhavan et al. (2022a) who highlighted the potential computational and memory gains of this gradient estimator. A related estimator was previously studied by Gasnikov et al. (2016). Their analysis is not sufficiently refined to establish any advantage of the $\ell_1$-randomization.

We impose the assumption used by Akhavan et al. (2020) over the random variables that we generate in the estimators (5.2) and (5.3), which intuitively forces the Oracle to select noise variables before observing the current query points.

**Assumption 5.3.1.** *For all $t \in \{1, \ldots, T\}$, it holds that:*

*(i) the random variables $\xi_t$ and $\xi'_t$ are independent from $\zeta^\circ_t$* (resp. *$\zeta^\diamond_t$) and from $r_t$ conditionally on $\boldsymbol{x}_t$, and the random variables $\zeta^\circ_t$* (resp. *$\zeta^\diamond_t$) and $r_t$ are independent;*

*(ii) $\mathbf{E}[\xi_t^2] \leq \sigma^2$ and $\mathbf{E}[(\xi'_t)^2] \leq \sigma^2$, where $\sigma \geq 0$.*

Note that we do not assume $\xi_t$ and $\xi'_t$ to have zero mean. Moreover, they can be non-random and no independence between noises on different time steps is required, so that the setting can be considered as *almost* adversarial. In particular, the first part of the assumption does not permit a *completely* adversarial setup—the Oracle is not allowed to choose the noise variable depending on the current query points (i.e. the two perturbations of $\mathbf{x}_t$). However, Assumption 5.3.1 encompasses the following protocol: before running the algorithm, the Oracle fixes an arbitrary bounded (by $\sigma$) sequence $(\xi_t, \xi'_t)_{t=1}^T$ of "noise" pairs, possibly with full knowledge of the algorithm employed by the statistician, and reveals this sequence query by query.

In the next two subsections we study the bias and variance of the two estimators. As we shall see, $\ell_1$-randomization can be more advantageous in the noiseless case than its $\ell_2$-counterpart (cf. Remark 5.3.6).

### Bias and variance of $\ell_2$-randomization

The next results allows us to control the bias and the second moment of gradient estimators $\mathbf{g}^\circ_1, \ldots, \mathbf{g}^\circ_T$, and play a crucial role in our analysis.

**Lemma 5.3.2** (Bias of $\ell_2$-randomization)**.** *Let Assumption 5.3.1 be fulfilled. Suppose that $f \in \mathcal{F}_\beta(L)$ for some $\beta \geq 2$ and $L > 0$. Let $\boldsymbol{x}_t$ and $\boldsymbol{g}^\circ_t$ be defined in* (5.2) *at time $t \geq 1$. Let $\ell = \lfloor \beta \rfloor$. Then,*

$$\|\mathbf{E}[\boldsymbol{g}^\circ_t \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \leq \kappa_\beta \frac{L}{(\ell-1)!} \cdot \frac{d}{d+\beta-1} h_t^{\beta-1} \enspace .$$

Intuitively, the smaller $h_t$ is, the more accurately $\mathbf{g}_t$ estimates the gradient. Lemma 5.3.2 was claimed in (Bach and Perchet, 2016, second inequality of Lemma 2) but the proof was not provided. The proof of Lemma 5.3.2, presented in Section 5.7, relies on the fact that $\mathbf{g}^\circ_t$ is an unbiased estimator of some surrogate function, which is strongly related to the original $f$. It should be noted that for $\beta > 2$ the bounds on the bias derived by Akhavan et al. (2020) and by Novitskii and Gasnikov (2021), who obtained $d$ and $\sqrt{d}$ dependency respectively, cannot be directly compared to our result. This is due to our Remark 5.2.2, which emphasizes that both of the aforementioned groups of authors work under a slightly different notion of smoothness.

Nevertheless, if $f$ is convex and $\beta = 2$ our result improves upon those in (Akhavan et al., 2020) and (Novitskii and Gasnikov, 2021) by factors of $d$ and $\sqrt{d}$ respectively, since both smoothness classes coincide.

The next lemma emphasizes the trade-off between the bias and the variance term which does not permit taking $h_t$ arbitrary small.

**Lemma 5.3.3** (Variance of $\ell_2$-randomization)**.** *If Assumption 5.3.1 holds true, $f \in \mathcal{F}_2(\bar{L})$ and $d \geq 2$, then*

$$\mathbf{E}\|\boldsymbol{g}_t^\circ\|^2 \leq \frac{d^2\kappa}{d-1}\mathbf{E}\left[\left(\|\nabla f(\boldsymbol{x}_t)\| + \bar{L}h_t\right)^2\right] + \frac{d^2\sigma^2\kappa}{h_t^2} \enspace ,$$

*where* $\kappa = \int_{-1}^{1} K^2(r)\,\mathrm{d}r$.

The result of Lemma 5.3.3 can be further simplified as

$$\mathbf{E}\|\mathbf{g}_t^\circ\|^2 \leq 4d\kappa\mathbf{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 + 4d\kappa\bar{L}^2h_t^2 + \frac{d^2\sigma^2\kappa}{h_t^2}, \qquad d \geq 2 \enspace .$$

Let us provide some remarks about this result. First, the leading term of order $d^2h_t^{-2}$ in the above bound is the same as in (Akhavan et al., 2020, Lemma 2.4) and in (Bach and Perchet, 2016, Appendix C1, beginning of the proof of Proposition 3), but we obtain a better constant. The main improvement *w.r.t.* to both works lies in the lower order term, unlike the aforementioned references, the term $h_t^2$ is multiplied by $d$ instead of $d^2$. On the first sight mild, this improvement is crucial for our guarantees and, in particular, for the condition $T \geq d^{2-\beta/2}$ (we would have had $T \geq d^3$ with the previously known versions of the variance bounds (Akhavan et al., 2020; Bach and Perchet, 2016). The proof relies on the Poincaré inequality for the uniform distribution on $\partial B_2^d$, which was exploited by Akhavan et al. (2022a) for the $\ell_1$-randomization.

*Proof of Lemma 5.3.3.* For simplicity we drop the subscript $t$ from all the quantities. By Assumption 5.3.1

$$\begin{aligned}\mathbf{E}\|\mathbf{g}^\circ\|^2 &= \frac{d^2}{4h^2}\mathbf{E}\left[\left(f(\mathbf{x}+hr\boldsymbol{\zeta}^\circ) - f(\mathbf{x}-hr\boldsymbol{\zeta}^\circ) + (\xi-\xi')\right)^2 K^2(r)\right] \\ &\leq \frac{d^2}{4h^2}\left(\mathbf{E}\left[(f(\mathbf{x}+hr\boldsymbol{\zeta}^\circ) - f(\mathbf{x}-hr\boldsymbol{\zeta}^\circ)^2K^2(r)\right] + 4\kappa\sigma^2\right) \enspace .\end{aligned} \tag{5.4}$$

In what follows, all appearing expectations should be understood conditionally on $\mathbf{x}_t$. Note that since $\mathbf{E}[f(\mathbf{x}+hr\boldsymbol{\zeta}^\circ) - f(\mathbf{x}-hr\boldsymbol{\zeta}^\circ) \mid r] = 0$ and $f \in \mathcal{F}_2(\bar{L})$, then using Wirtinger-Poincaré inequality (see, *e.g.,* Beckner, 1989; Osserman, 1978, Eq. (3.1) or Theorem 2, respectively) we deduce

$$\mathbf{E}\left[(f(\mathbf{x}+hr\boldsymbol{\zeta}^\circ) - f(\mathbf{x}-hr\boldsymbol{\zeta}^\circ))^2 \mid r\right] \leq \frac{h^2}{d-1}\mathbf{E}\left[\|\nabla f(\mathbf{x}+hr\boldsymbol{\zeta}^\circ) + \nabla f(\mathbf{x}-hr\boldsymbol{\zeta}^\circ)\|^2 \mid r\right] \enspace . \tag{5.5}$$

Since $f \in \mathcal{F}_2(\bar{L})$, then the triangle inequality further implies that

$$\mathbf{E}\left[\|\nabla f(\mathbf{x} + hr\boldsymbol{\zeta}^\circ) + \nabla f(\mathbf{x} - hr\boldsymbol{\zeta}^\circ)\|^2 \;\middle|\; r\right] \leq 4\left(\|\nabla f(\mathbf{x})\| + \bar{L}h\right)^2 \quad . \tag{5.6}$$

We conclude by plugging the above bound into Eq. (5.5) and taking into account Eq. (5.4). □

Note that Eq. (5.6) combined with Eq. (5.5) can bee seen as a version of (Shamir, 2017, Lemma 9), who relied on a concentration argument and obtained non-explicit constants. Indeed, Lemma 9 of Shamir (2017), is implied by Eqs. (5.5)–(5.6).

### Bias and variance of $\ell_1$-randomization

This section analyses the gradient estimate based on the $\ell_1$-randomization. The results below display a very different bias and variance behavior compared to the $\ell_2$-randomization.

**Lemma 5.3.4** (Bias of $\ell_1$-randomization)**.** *Let Assumption 5.3.1 be fulfilled. Suppose that $f \in \mathcal{F}_\beta(L)$ for some $\beta \geq 2$ and $L > 0$. Let $\boldsymbol{x}_t$ and $\boldsymbol{g}_t^\diamond$ be defined in (5.3) at time $t \geq 1$. Let $\ell = \lfloor \beta \rfloor$. Then,*

$$\|\mathbf{E}[\boldsymbol{g}_t^\diamond \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \leq L\kappa_\beta h_t^{\beta-1}\ell^{\beta-\ell}d^{\frac{1-\beta}{2}} \quad .$$

We emphasize that both Lemma 5.3.4 and Lemma 5.3.2 give identical dependency on the discretization parameter $h_t$. However, note that unlike the bias bound derived for the $\ell_2$ case, which was *dimension* independent, the result of Lemma 5.3.4 actually depends on the dimension in a *favourable* way. In particular, the bias is controlled by a decreasing function of the ambient dimension and this dependency becomes more and more favorable for smoother functions. Yet, the price for such a favorable control of bias is an inflated bound on the variance, which is established below.

**Lemma 5.3.5** (Variance of $\ell_1$-randomization)**.** *Let Assumption 5.3.1 be fulfilled. Assume that $f \in \mathcal{F}_2(\bar{L})$ and $d \geq 3$, then*

$$\mathbf{E}\|\boldsymbol{g}_t^\diamond\|^2 \leq \frac{\bar{\mathtt{C}}_{d,1}d^2\kappa}{d-2}\mathbf{E}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{\bar{\mathtt{C}}_{d,2}d^2\kappa\bar{L}^2h_t^2}{(d-2)(d+1)} + \frac{d^3\sigma^2\kappa}{h_t^2} \quad ,$$

*where $\bar{\mathtt{C}}_{d,1} \leq 8\left(1 + \sqrt{\frac{2d}{d+1}}\right)^2$ and $\bar{\mathtt{C}}_{d,2} = 16\big(3 + \sqrt{8 + \frac{592}{(d+3)^2}} + \frac{22}{d}\big)$.*

Let us first discuss the constants $\bar{\mathtt{C}}_{d,1}$ and $\bar{\mathtt{C}}_{d,2}$. First of all, it is important to observe that both $\bar{\mathtt{C}}_{d,1}$ and $\bar{\mathtt{C}}_{d,2}$ can be computed in practice as they only depend on the dimension. This fact is important for the eventual choice of $\eta_t$ and $h_t$ for the estimator (5.3) . Note that, asymptotically we have

$$\lim_{d\to\infty} \bar{\mathtt{C}}_{d,1} = 8(1+\sqrt{2})^2 \leq 46.63 \qquad \text{and} \qquad \lim_{d\to\infty} \bar{\mathtt{C}}_{d,2} = 16(3+\sqrt{8}) \leq 93.26 \quad .$$

Furthermore, $\bar{\mathtt{C}}_{d,1}$ is increasing (hence, remains upper-bounded by $8(1+\sqrt{2})^2$) with the growth of $d$. Meanwhile, $\bar{\mathtt{C}}_{d,2}$ decreases with the growth of $d$, which implies that for all $d \geq 3$ it holds that $\bar{\mathtt{C}}_{d,2} \leq \bar{\mathtt{C}}_{3,2} \leq 244.5$. Hence, both $\bar{\mathtt{C}}_{d,1}$ and $\bar{\mathtt{C}}_{d,2}$ are of constant order. In view of the above and combined with the fact that $a - 2 \geq a/3$ for all $a \geq 3$, the inequality of Lemma 5.3.5 can be further simplified as

$$\mathbf{E}\|\mathbf{g}_t^{\diamond}\|^2 \leq d\kappa A_1 \mathbf{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 + \kappa \bar{L}^2 A_2 h_t^2 + \frac{d^3\sigma^2\kappa}{h_t^2}\ , \qquad d \geq 3\ , \tag{5.7}$$

with $A_1 = 139.9, A_2 = 733.5$. Yet, since both $\bar{\mathtt{C}}_{d,1}$ and $\bar{\mathtt{C}}_{d,2}$ are known explicitly, they can be implemented in practice and used directly for the choice of the discretization $h_t$ and the step-size $\eta_t$. For this reason, we keep the derived bound on the variance as is and do not rely on its simplified version from Eq. (5.7).

Modulo such absolute constants, the leading term *w.r.t.* $h_t$ in Lemma 5.3.5 is the same as for $\ell_2$-randomization in Lemma 5.3.3. However, for the $\ell_2$-randomization this term (in $h_t$) involves only a quadratic dependency on the dimension $d$, while in the case of $\ell_1$ randomization this dependency is cubic. Interestingly, the constant in front of the negligible term $h_t^2$ does not grow with the growth of dimension. In contrast, the corresponding term in Lemma 5.3.3 involves linear dependency on the dimension. We summarize these observations in the following remark which considers the noiseless case.

*Proof of Lemma 5.3.5.* For simplicity we drop subscript index $t$ from all the quantities. Similarly to the proof of Lemma 5.3.3, using Assumption 5.3.1, we deduce that

$$\mathbf{E}\|\mathbf{g}^{\diamond}\|^2 \leq \frac{d^3}{4h^2}\left(\mathbf{E}[(f(\mathbf{x}+hr\boldsymbol{\zeta}^{\diamond}) - f(\mathbf{x}-hr\boldsymbol{\zeta}^{\diamond}))^2 K^2(r)] + 4\sigma^2\kappa\right)\ . \tag{5.8}$$

Consider $G : \mathbb{R}^d \to \mathbb{R}$ defined for all $\boldsymbol{u} \in \mathbb{R}^d$ as $G(\boldsymbol{u}) = f(\mathbf{x}+hr\boldsymbol{u}) - f(\mathbf{x}-hr\boldsymbol{u})$. Using the fact that $f \in \mathcal{F}_2(\bar{L})$ we obtain for all $\boldsymbol{u} \in \mathbb{R}^d$

$$\|\nabla G(\boldsymbol{u})\|^2 \leq 8h^4\bar{L}^2\left\|\boldsymbol{u}\right\|^2 + 8h^2\left\|\nabla f(\mathbf{x})\right\|^2\ .$$

In what follows, all the expectations appearing should be understood conditionally on $\mathbf{x}_t$. Applying (Akhavan et al., 2022a, Lemma 3) to the function $G$ defined above, we deduce that

$$\mathbf{E}\left[(G(\boldsymbol{\zeta}^{\diamond}))^2 \mid r\right] \leq \frac{32h^2}{d(d-2)}\mathbf{E}\left[\left(h^2\bar{L}^2\left\|\boldsymbol{\zeta}^{\diamond}\right\|^2 + \left\|\nabla f(\mathbf{x})\right\|^2\right)(1+\sqrt{d}\left\|\boldsymbol{\zeta}^{\diamond}\right\|)^2\right]\ .$$

Lemma 5.7.7, provided in Section 5.7, gives upper bounds on the expectations appearing in the above inequality for all $d \geq 3$ and its application yields

$$\mathbf{E}\left[\left(f(\mathbf{x}+hr\boldsymbol{\zeta}^{\diamond}) - f(\mathbf{x}-hr\boldsymbol{\zeta}^{\diamond})\right)^2 \mid r\right] \leq \frac{\mathtt{C}_{d,1}h^2}{d(d-2)}\left\|\nabla f(\mathbf{x})\right\|^2 + \frac{\mathtt{C}_{d,2}\bar{L}^2h^4}{d(d-2)(d+1)}\ ,$$

where $\mathtt{C}_{d,1} = 32\left(1+\sqrt{\frac{2d}{d+1}}\right)^2$ and $\mathtt{C}_{d,2} = 64\left(3+\sqrt{8+\frac{592}{(d+3)^2}}+\frac{22}{d}\right)$. We conclude by substituting the above bound into the r.h.s. of Eq. (5.8). □

**Remark 5.3.6** (On the advantage of $\ell_1$-randomization). *In the noiseless case ($\sigma = 0$) both bias and variance of the $\ell_1$-randomization are strictly smaller than that of $\ell_2$-randomization. Indeed, if $\sigma = 0$*

$$\begin{cases} \|\mathbf{E}[\boldsymbol{g}_t^{\circ} \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \lesssim h_t^{\beta-1} \\ \mathbf{E}\|\boldsymbol{g}_t^{\circ}\|^2 \lesssim d\mathbf{E}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 + dh_t^2 \end{cases} \quad \text{and} \quad \begin{cases} \|\mathbf{E}[\boldsymbol{g}_t^{\diamond} \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \lesssim \left(\frac{h_t}{\sqrt{d}}\right)^{\beta-1} \\ \mathbf{E}\|\boldsymbol{g}_t^{\diamond}\|^2 \lesssim d\mathbf{E}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 + h_t^2 \end{cases},$$

*where $\lesssim$ hides multiplicative constants that do not depend on $h_t$ and $d$. As a thought experiment, substitute $d = 10^6$ and $\beta = 2$.*

## 5.4 Upper bounds

In this section, we present the convergence guarantees for the two considered gradient estimators for various classes of objective functions $f$. Each of the following subsections is structured similarly: first, we define the choice of $\eta_t$ and $h_t$ involved in both algorithms, and then, for each class of the objective functions, we state the corresponding convergence guarantees.

In Section 5.4 we consider the case that $f$ has higher order derivatives, $f \in \mathcal{F}_2(\bar{L}) \cap \mathcal{F}_\beta(L)$ for $\beta \geq 2$, and establish the guarantee for the stationary point. In Section 5.4 we *additionally* assume that $f$ is $\alpha$-gradient dominant and provide guarantees on the optimization error. In Section 5.4 we analyze the case of strongly convex target function (both constrained and unconstrained cases), improving the upper bound derived by Akhavan et al. (2020) and Novitskii and Gasnikov (2021). Unless stated otherwise, the convergence guarantees presented in this section hold under the assumption that the number of queries $T$ is known before running the algorithms.

### Only smoothness assumptions

In this part we only assume that the objective function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumption 5.2.3. In particular, since there is no guarantee of the unicity (or existence) of the minimizer, our goal is modest—we only want to obtain a nearly stationary point.

The guarantee of this section will be stated on a randomly sampled point along the trajectory. The distribution on the trajectory is chosen carefully, to guarantee the desired convergence. The distribution that we are going to eventually use is defined in the following lemma.

**Lemma 5.4.1.** *Consider the iterative algorithm defined in Eq. (5.1) with $\Theta = \mathbb{R}^d$. Assume that there exist two positive sequences $b_t, v_t : \mathbb{N} \to [0, \infty)$ and $m \geq 0$ such that for all $t \geq 1$ it holds* *almost surely that*

$$\|\mathbf{E}[\boldsymbol{g}_t \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \leq b_t \qquad \text{and} \qquad \mathbf{E}\,\|\boldsymbol{g}_t\|^2 \leq v_t + m\mathbf{E}\,\|\nabla f(\boldsymbol{x}_t)\|^2 \enspace .$$

*Assume that $\eta_t$ in Eq. (5.1) is chosen to satisfy $\bar{L}\eta_t m < 1$ and that $f^\star > -\infty$. Let $S$ be a random variable with values in $[T]$, which is independent from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T, \boldsymbol{g}_1, \ldots, \boldsymbol{g}_T$ and such that*

$$\mathbf{P}(S = t) = \frac{\eta_t\left(1 - \bar{L}\eta_t m\right)}{\sum_{t=1}^T \eta_t\left(1 - \bar{L}\eta_t m\right)} \enspace .$$

*Then, it holds that*

$$\mathbf{E}\,\|\nabla f(\boldsymbol{x}_S)\|^2 \leq \frac{2(\mathbf{E}[f(\boldsymbol{x}_1)] - f^\star) + \sum_{t=1}^T \eta_t\left(b_t^2 + \bar{L}\eta_t v_t\right)}{\sum_{t=1}^T \eta_t\left(1 - \bar{L}\eta_t m\right)} \enspace .$$

The above lemma, as well as its proof, is similar to the techniques used by Ghadimi and Lan (2013) in the context of zero-order non-convex optimization. However, in the work of Ghadimi and Lan (2013) the noise model is different—for them, both $b_t$ and $v_t$ decrease when the discretization parameter $h_t$ ($\mu$ in their notation) decreases. Note that the distribution of $S$ in Lemma 5.4.1 depends on our choice of $\eta_t$ and $m$. In the following result, we are going to specify the exact values of $\eta_t$. Meanwhile, the value of $m$ is obtained either from Lemma 5.3.3 or from Lemma 5.3.5, depending on the choice of the algorithm.

The next result, as with all other results of this section, requires a definition of algorithm-dependent parameters, which are needed as an input to our algorithms

$$(\mathfrak{y}, \mathfrak{h}) = \begin{cases} \left(\frac{1}{8\kappa\bar{L}},\, d^{\frac{1}{2\beta-1}}\right) & \text{for } \ell_2\text{-randomization} \\ \left(\frac{d-2}{2\bar{L}\bar{\mathtt{C}}_{d,1}d},\, d^{\frac{2\beta+1}{4\beta-2}}\right) & \text{for } \ell_1\text{-randomization} \end{cases} \enspace . \tag{5.9}$$

We recall that $\bar{\mathtt{C}}_{d,1}$ is the constant which appears and is explicitly defined in Lemma 5.3.5.

**Theorem 5.4.2.** *Let Assumptions 5.2.3 and 5.3.1 hold, for $\beta \geq 2$. Consider gradient estimators (5.2), (5.3) used in (5.1), with parameterization $\eta_t$ and $h_t$ defined for $t = 1, \ldots, T$ as*

$$\eta_t = \min\left(\frac{\mathfrak{y}}{d},\, d^{-\frac{2(\beta-1)}{2\beta-1}} T^{-\frac{\beta}{2\beta-1}}\right) \qquad \text{and} \qquad h_t = \mathfrak{h} \cdot T^{-\frac{1}{2(2\beta-1)}} \enspace .$$

*where the pair of constants $(\mathfrak{y}, \mathfrak{h})$ is defined in Eq. (5.9) depending on the algorithm. Assume that $\Theta = \mathbb{R}^d$, $\boldsymbol{x}_1$ is deterministic, and $T \geq d^{\frac{1}{\beta}}$, then for $\boldsymbol{x}_S$ defined in Lemma 5.4.1 we have*

$$\mathbf{E}\,\|\nabla f(\boldsymbol{x}_S)\|^2 \leq \left(\mathtt{A}_1(f(\boldsymbol{x}_1) - f^\star) + \mathtt{A}_2\right) \cdot \left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}},$$

*where the constants $\mathtt{A}_1, \mathtt{A}_2 > 0$ depend only on $\sigma, L, \bar{L}$, $\beta$, and the choice of the algorithm.*

### Smoothness and $\alpha$-gradient dominance

| | $b$ | $\mathsf{V}_1$ | $\mathsf{V}_2$ | $\mathsf{V}_3$ |
|---|---|---|---|---|
| $\ell_2$-randomization | $\frac{\kappa_\beta}{(\ell-1)!} \cdot \frac{d}{d+\beta-1}$ | $4d\kappa$ | $4d\kappa$ | $d^2\kappa$ |
| $\ell_1$-randomization | $\kappa_\beta \ell^{\beta-\ell} d^{\frac{1-\beta}{2}}$ | $139.9d\kappa$ | $733.5\kappa$ | $d^3\kappa$ |

Table 5.1: Bias and variance of both gradient estimators to be used in Theorems 5.4.4 and 5.4.6.

In the context of deterministic first-order optimization, the $\alpha$-gradient dominance allows to obtain rates of convergence of gradient descent algorithm, which are similar to the case of the strongly convex objective function with Lipschitz gradient (this rate is often called linear in the optimization literature (see e.g., Karimi et al., 2016)). Hence, it is natural to expect that in our context the $\alpha$-gradient dominance leads to the same convergence rates as $\alpha$-strong convexity. We show in Theorem 5.4.4 that the rates are only inflated by a multiplicative factor $\alpha^{-1}$ compared to the strongly convex case that will be discussed in the next section.

**Assumption 5.4.3.** *Assume that the sequence of* $(\boldsymbol{g}_t)_{t\geq 1}$ *used in Algorithm 5.1 satisfies*

$$\|\mathbf{E}[\boldsymbol{g}_t \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \leq bLh_t^{\beta-1} \quad \text{and} \quad \mathbf{E}\|\boldsymbol{g}_t\|^2 \leq \mathsf{V}_1\mathbf{E}\|\nabla f(\boldsymbol{x}_t)\|^2 + \mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2} \ .$$

Note that Assumption 5.4.3 holds for gradient estimators (5.2), (5.3), with the values that are indicated in Table 5.1, see Lemmas 5.3.2–5.3.5.

**Theorem 5.4.4.** *Let $f$ be an $\alpha$-gradient dominant function and Assumptions 5.2.3 and 5.3.1 hold, for $\beta \geq 2$. Consider an iterative procedure defined in Eq. (5.1) and assume that Assumption 5.4.3 is satisfied. Set*

$$\eta_t = \min\left(\frac{1}{2\bar{L}\mathsf{V}_1}, \frac{4}{\alpha t}\right) \qquad \text{and} \qquad h_t = \left(4\bar{L}\cdot\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{2\beta}} \cdot \begin{cases} t^{-\frac{1}{2\beta}} & \text{if } \eta_t = {}^4\!/\alpha t \\ T^{-\frac{1}{2\beta}} & \text{if } \eta_t = {}^1\!/2\bar{L}\mathsf{V}_1 \end{cases} .$$

*Assume that $\Theta = \mathbb{R}^d$, and $\boldsymbol{x}_1$ is deterministic, then*

$$\mathbf{E}[f(\boldsymbol{x}_T) - f^\star] \leq \mathtt{A}_1 \cdot \frac{\mathsf{V}_1}{\alpha T}(f(\boldsymbol{x}_1) - f^\star) + \frac{\mathtt{A}_2}{\alpha \wedge \alpha^2}\left(\mathsf{V}_3\left\{\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}} + \mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}T^{-\frac{2}{\beta}}\right\}T^{-\frac{\beta-1}{\beta}}\right. ,$$

*where $\mathtt{A}_1, \mathtt{A}_2 > 0$ depend only on $\sigma^2, L, \bar{L}, \beta$.*

The above theorem states a general result for any gradient estimator that satisfies Assumption 5.4.3. Given the values provided in Table 5.1, we can state the following corollary for both considered estimators.

**Corollary 5.4.5.** *Let $f$ be an $\alpha$-gradient dominant function and Assumptions 5.2.3 and 5.3.1 hold, for $\beta \geq 2$. Consider gradient estimators (5.2), (5.3) used in (5.1), with parameterization*

*$\eta_t$ and $h_t$ that are outlined in Theorem 5.4.4, where $b, \mathsf{V}_1, \mathsf{V}_2, \mathsf{V}_3$ are assigned based on Table 5.1 for each algorithm, respectively. Assume that $\Theta = \mathbb{R}^d$, $\boldsymbol{x}_1$ is deterministic, and $T \geq d^{2-\frac{\beta}{2}}$. Then*

$$\mathbf{E}[f(\boldsymbol{x}_T) - f^\star] \leq \mathtt{A}_1 \cdot \frac{d}{\alpha T}\left(f(\boldsymbol{x}_1) - f^\star\right) + \frac{\mathtt{A}_2}{\alpha \wedge \alpha^2}\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{\beta}} \quad ,$$

*where $\mathtt{A}_1, \mathtt{A}_2 > 0$ depend only on $\sigma^2, L, \bar{L}, \beta$, and the choice of the algorithm.*

## Smoothness and strong convexity

In this section, we provide the guarantee on the average point $\bar{\mathbf{x}}_T$ along the trajectory of the algorithm, that is,

$$\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t \quad .$$

Note that instead of taking the average along the trajectory one may simply sample uniformly at random one point along the trajectory and our guarantees still hold. However, to avoid additional randomness, we stick to the deterministic averaging.

### Unconstrained optimization

In this part we consider the case of $\Theta = \mathbb{R}^d$ and, as in the previous parts, we assume that $T$ is known beforehand. Similar to the previous section, first we state a general result that can be applied to any gradient estimator that satisfies Assumption 5.4.3.

**Theorem 5.4.6.** *Let $f$ be an $\alpha$-strongly convex function and Assumptions 5.2.3 and 5.3.1 hold, for $\beta \geq 2$. Consider an iterative procedure defined in Eq. (5.1) and assume that Assumption 5.4.3 is satisfied. Set*

$$\eta_t = \min\left(\frac{\alpha}{4\bar{L}^2\mathsf{V}_1}, \frac{4}{\alpha t}\right) \qquad \textit{and} \qquad h_t = \left(\frac{2\mathsf{V}_3}{b^2}\right)^{\frac{1}{2\beta}} \cdot \begin{cases} t^{-\frac{1}{2\beta}} & \textit{if } \eta_t = {}^4/_{\alpha t} \\ T^{-\frac{1}{2\beta}} & \textit{if } \eta_t = {}^1/_{2\bar{L}\mathsf{V}_1} \end{cases} \quad .$$

*Assume that $\Theta = \mathbb{R}^d$, $\boldsymbol{x}_1$ is deterministic, then*

$$\mathbf{E}[f(\bar{\boldsymbol{x}}_T) - f^*] \leq \mathtt{A}_1\frac{\mathsf{V}_1}{\alpha T}\left\|\boldsymbol{x}_1 - \boldsymbol{x}^*\right\|^2 + \left\{\mathtt{A}_2 b^{\frac{2}{\beta}}\mathsf{V}_3^{\frac{\beta-1}{\beta}} + \mathtt{A}_3\mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}\left(\mathsf{V}_1^{-\frac{2}{\beta}} + T^{-\frac{2}{\beta}}\right)\right\}\frac{T^{-\frac{\beta-1}{\beta}}}{\alpha} \quad ,$$

*where the constants $\mathtt{A}_1, \mathtt{A}_2, \mathtt{A}_3 > 0$ depend only on $\sigma, L, \bar{L}, \beta$.*

Subsequently, in Corollary 5.4.7, we customize the above theorem for gradient estimators (5.2) and (5.3), with assignments of $\eta_t, h_t$ that are again selected based on Table 5.1.

**Corollary 5.4.7.** *Let $f$ be an $\alpha$-strongly convex function and Assumptions 5.2.3 and 5.3.1 hold, for $\beta \geq 2$. Consider gradient estimators (5.2), (5.3) used in (5.1), with parameterization $\eta_t$ and $h_t$ that are precised in Theorem 5.4.6, where $b, \mathsf{V}_1, \mathsf{V}_2, \mathsf{V}_3$ are set according to Table 5.1 for each algorithm, respectively. Assume that $\Theta = \mathbb{R}^d$, $\boldsymbol{x}_1$ is deterministic, and $T \geq d^{2-\frac{\beta}{2}}$. Then*

$$
\mathbf{E}[f(\boldsymbol{x}_T) - f^\star] \leq \mathtt{A}_1 \cdot \frac{d}{\alpha T} \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \frac{\mathtt{A}_2}{\alpha}\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{\beta}} \quad ,
$$

*where $\mathtt{A}_1, \mathtt{A}_2 > 0$ depend only on $\sigma^2, L, \bar{L}, \beta$, and the choice of the algorithm.*

With an slightly different notion of smoothness, (Akhavan et al., 2020, Theorem 3.2) derived a similar result, that is comparable to Corollary 5.4.7. However, in the latter the authors imposed an additional condition on $\alpha$ (i.e., $\alpha \gtrsim \sqrt{d/T}$), which is necessary for their convergence guarantee.

### Constrained optimization

We now assume that $\Theta \subset \mathbb{R}^d$ is a compact convex set. Consequently, since $f$ is continuously differentiable, its gradient is bounded on $\Theta$. This fact allows us to develop any-time results, that is, in the present part, we do not require the knowledge of the optimization horizon $T$. The results of this section are essentially corollaries of the following general result.

**Lemma 5.4.8.** *Let $\Theta \subset \mathbb{R}^d$ be a compact convex set. Assume that $f$ is $\alpha$-strongly convex on $\Theta$ and $\sup_{\boldsymbol{x}\in\Theta} \|\nabla f(\boldsymbol{x})\|_2 \leq G$. Consider the iterative algorithm in Eq. (5.1). Assume that there exist two positive sequences $b_t, v_t, : \mathbb{N} \to [0, \infty)$ and $m \in [0, \infty)$ such that for all $t \geq 1$ it holds almost surely that*

$$
\|\mathbf{E}[\boldsymbol{g}_t \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \leq b_t \qquad \text{and} \qquad \mathbf{E}\|\boldsymbol{g}_t\|^2 \leq v_t + m\mathbf{E}\|\nabla f(\boldsymbol{x}_t)\|^2 \quad .
$$

*Then, for $\bar{\boldsymbol{x}}_T = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t$ we have*

$$
\mathbf{E}[f(\bar{\boldsymbol{x}}_T) - f^\star] \leq \frac{mG^2 \log(eT)}{\alpha T} + \frac{1}{\alpha T}\sum_{t=1}^{T}\left(\frac{v_t}{t} + b_t^2\right) \quad .
$$

Using the bounds on the variance and bias derived in Section 5.3, we can deduce the following two guarantees for gradient estimators (5.2), (5.3) used in (5.1), by analysing the resulting recursive relations.

**Theorem 5.4.9.** *Let $f$ be an $\alpha$-strongly convex function and Assumptions 5.2.3 and 5.3.1 hold, for $\beta \geq 2$. Consider gradient estimators (5.2), (5.3) used in (5.1), with parametrization $\eta_t = \frac{2}{\alpha t}$ and $h_t = \mathfrak{h} \cdot t^{-1/2\beta}$, where the constant $\mathfrak{h}$ equals to $d^{\frac{1}{\beta}}$ for (5.2) and to $d^{\frac{2+\beta}{2\beta}}$ for (5.3). Assume* *that $\Theta$ is a convex and closed subset of $\mathbb{R}^d$, with $\sup_{\boldsymbol{x}\in\Theta}\|\nabla f(\boldsymbol{x})\|_2 \leq G$. Then,*

$$
\mathbf{E}[f(\bar{\boldsymbol{x}}_T) - f^\star] \leq \mathtt{A}_1 \cdot \frac{dG^2\log(eT)}{\alpha T} + \mathtt{A}_2 \cdot \frac{1}{\alpha}\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{\beta}} + \mathtt{A}_3 \cdot \frac{d^{1+\frac{2}{\beta}}}{\alpha T} \enspace ,
$$

*with $\mathtt{A}_1, \mathtt{A}_2, \mathtt{A}_3 > 0$ depend only on $\sigma^2, L, \bar{L}, \beta$ and the choice of the algorithm.*

The above result is nearly identical to Corollary 5.4.7—they only differ in the exact values of constants $\mathtt{A}_1, \mathtt{A}_2, \mathtt{A}_3$. Theorem 5.4.9 should be compared to (Akhavan et al., 2020, Theorem 3.1 and Theorem 5.1) and with (Novitskii and Gasnikov, 2021, Theorem 1). However, we again recall that both Akhavan et al. (2020); Novitskii and Gasnikov (2021) work with a slightly different, compared to ours, notion of smoothness (both coincide in case $\beta = 2$, see our Remark 5.2.2). The term $\frac{d^{1+2/\beta}}{T}$, appearing in both results is negligible as long as $T \geq d^{4-\beta}$, which, in the worst case of $\beta = 2$ means that $T \geq d^2$.

## 5.5 Lower bounds

In this section we prove a minimax lower bound on the optimization error over all sequential strategies that allow the query points depend on the past. The established lower bound is similar to that of Akhavan et al. (2020). However, we work with the Hellinger distance instead of the Kullback–Leibler divergence, which allows us to encompass a larger family of noises. For $t = 1, \ldots, T$, we assume that $y_t = f(\boldsymbol{z}_t) + \xi_t$ and we consider strategies of choosing the query points as $\boldsymbol{z}_t = \Phi_t(\boldsymbol{z}_1, y_1, \cdots, \boldsymbol{z}_{t-1}, y_{t-1}, \boldsymbol{\tau}_t)$ where $\Phi_t$'s are Borel functions and $\boldsymbol{z}_1 \in \mathbb{R}^d$ is any random variable, and $\{\boldsymbol{\tau}_t\}$ is a sequence of random variables in a measurable space $(\mathcal{Z}, \mathcal{U})$, such that $\boldsymbol{\tau}_t$ is independent of $(\boldsymbol{z}_1, y_1, \cdots, \boldsymbol{z}_{t-1}, y_{t-1})$. We denote by $\Pi_T$ the set of all such strategies. First, as mentioned above, we establish a generalization of (Akhavan et al., 2020, Eq. (13))—a requirement of bounded Kullback–Leibler divergence, that we replace by the Hellinger distance.

**Lemma 5.5.1.** *For any $f : \mathbb{R}^d \to \mathbb{R}$ and any sequential strategy $\boldsymbol{z}_t = \Phi_t(\boldsymbol{z}_1, y_1, \ldots, y_{t-1}, \boldsymbol{\tau}_t)$ with $y_t = f(\boldsymbol{z}_t) + \xi_t$ for $t = 1, \ldots, T$, denote by $\mathbf{P}_f$ the joint distribution of $((\boldsymbol{z}_i, y_i)_{i=1}^T, (\boldsymbol{\tau}_i)_{i=1}^T)$. Assume that $\xi_1, \ldots, \xi_T$ are* i.i.d. *with cumulative distribution function $F : \mathbb{R} \to \mathbb{R}$ satisfying*

$$
\int \left(\sqrt{\mathrm{d}F(u)} - \sqrt{\mathrm{d}F(u+v)}\right)^2 \mathrm{d}u \leq I_0 v^2\,, \qquad |v| \leq v_0 \enspace , \tag{5.10}
$$

*for some $0 < I_0 < \infty$, $0 < v_0 \leq \infty$, and such that $\xi_t$ is independent of $(\boldsymbol{z}_1, y_1, \ldots, \boldsymbol{z}_{t-1}, y_{t-1}, \boldsymbol{\tau})$. Furthermore, assume that for all $t = 1, \ldots, T$ it holds that $z_t \in \Theta \subset \mathbb{R}^d$, then for any $f, f' : \mathbb{R}^d \to \mathbb{R}$ such that $\max_{\boldsymbol{u}\in\Theta}|f(\boldsymbol{u}) - f'(\boldsymbol{u})| \leq B \leq v_0$ it holds that*

$$
\frac{1}{2}H^2(\mathbf{P}_f, \mathbf{P}_{f'}) \leq 1 - \left(1 - \frac{I_0}{2}B\right)^T \enspace ,
$$

*where $H^2(\cdot,\cdot)$ is the Hellinger distance, defined for two probability distributions $\mathbf{P},\mathbf{P}'$ as*

$$H^2(\mathbf{P},\mathbf{P}') \triangleq \int (\sqrt{\mathrm{d}\mathbf{P}} - \sqrt{\mathrm{d}\mathbf{P}'})^2 \ .$$

Our construction of the lower bound relies heavily on Lemma 5.5.1. In particular, this construction is built upon i.i.d. noise satisfying the condition in Eq. (5.10). The condition in Eq. (5.10) is not restrictive—for example, for Gaussian distribution $F$ it is satisfied with $v_0 = \infty$. As it is noted by Akhavan et al. (2020), the class $\Pi_T$ includes the sequential strategy of Algorithm (5.1) with either of the considered estimators. Indeed, taking $T$ as an even number, and choosing $\mathbf{z}_t = \mathbf{x}_t + h_t\boldsymbol{\zeta}_t r_t$ and $\mathbf{z}_t = \mathbf{x}_t - h_t\boldsymbol{\zeta}_t r_t$ or even $t$ and odd $t$, respectively.

Let us also explain the improvement upon (Akhavan et al., 2020, Eq. (13)), where a similar bound is required for the Kullback–Leibler divergence instead of the Hellinger distance. First, from purely quantitative point of view, any upper bound for the Kullback–Leibler divergence implies an upper bound for the squared Hellinger distance, hence (Akhavan et al., 2020, Eq. (13)) is a stronger condition in comparison with Eq. (5.10). More qualitatively, Eq. (5.10) encompasses a larger family of noises satisfies. In particular, in order to use Kullback–Leibler divergence between two distributions, we need one of them to be absolutely continuous with respect to the other one, while the Hellinger distance does not require such a restricted condition. As an example, one can consider $F$ to be a bounded support distribution. Then, the Kullback–Leibler divergence between $F(\cdot)$ and $F(\cdot + v)$ is unbounded. However, the Hellinger distance remains bounded and $F$ can be used as the distribution of $\xi_1, \dots, \xi_T$ in the lower bound.

**Theorem 5.5.2.** *lowerB Let $\Theta = \{\boldsymbol{x} \in \mathbb{R}^d \ : \ \|\boldsymbol{x}\| \leq 1\}$. For $\alpha, L > 0$, $\beta \geq 2$, let $\mathcal{F}'_{\alpha,\beta}$ denote the set of functions $f$ that attain their minimum over $\mathbb{R}^d$ in $\Theta$ and belong to $\mathcal{F}_{\alpha,\beta}(L) \cap \{f \ : \ \max_{\boldsymbol{x}\in\Theta} \|\nabla f(\boldsymbol{x})\| \leq G\}$, where $G > 2\alpha$. Then for any strategy in the class $\Pi_T$ we have*

$$\sup_{f\in\mathcal{F}'_{\alpha,\beta}} \mathbf{E}\big[f(\boldsymbol{z}_T) - f^\star\big] \geq C \min\left(\max\left(\alpha,\, T^{-1/2+1/\beta}\right), \frac{d}{\sqrt{T}}, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\right) \ , \tag{5.11}$$

*and*

$$\sup_{f\in\mathcal{F}'_{\alpha,\beta}} \mathbf{E}\,\|\boldsymbol{z}_T - \boldsymbol{x}^*(f)\|^2 \geq C \min\left(1,\, \frac{d}{T^{\frac{1}{\beta}}}, \frac{d}{\alpha^2} T^{-\frac{\beta-1}{\beta}}\right) \ , \tag{5.12}$$

*where $C > 0$ is a constant that does not depend of $T, d$, and $\alpha$, and $\boldsymbol{x}^*(f)$ is the minimizer of $f$ on $\Theta$.*

For the family of $\alpha$-strongly convex functions, all the discussions provided by Akhavan et al. (2020) after their Theorem 6.1 are applicable to our case. Furthermore, since any $\alpha$-strongly convex function is $\alpha$-gradient dominant, the outlined result in Eq. (5.11) is a valid lower bound for the family of $\alpha$-gradient dominant functions. This fact, highlights the minimax optimality of our proposed algorithms (see Corollary 5.4.5), for $\alpha$-gradient dominant functions, with respect to $T$ (and $d$, if $\beta = 2$).

The following theorem is a simple and direct corollary of Theorem 5.5.2, which provides an ad-hoc lower bound for the gradient of the family of smooth non-convex functions.

**Theorem 5.5.3.** *Consider the class, for $\beta \geq 2$, $L > 0$, $\bar{L} > 0$: $\tilde{\mathcal{F}}_\beta(L, \bar{L}) = \{f \in \mathcal{F}_\beta(L), f$ is $\bar{L}$-smooth$\}$. Let $\{\boldsymbol{z}_t\}_{t=1}^T$ be any algorithm belonging to the same class of sequential strategies as Theorem 5.5.2. Let $S$ be any random random variable taking values in $\{1, \ldots, T\}$, independent of other sources of randomness. Then, under the assumptions of Theorem 5.5.2 we have*

$$\sup_{f \in \tilde{\mathcal{F}}_\beta(L, \bar{L})} \mathbf{E} \|\nabla f(\boldsymbol{z}_S)\|^2 \geq C d T^{-\frac{\beta-1}{\beta}} \enspace ,$$

*where $C > 0$ does not depend on $d, T$, and $\beta$.*

## 5.6 Discussion

We have provided an improved analysis of the algorithm of Bach and Perchet (2016) and introduced a new algorithm based on the $\ell_1$-randomization. The new algorithm enjoys similar guarantees as the previously known one. Note that each of the considered cases involves different points along the trajectory: randomly sampled for the non-convex case; the last point under the Polyak-Łojasiewicz condition; averaged for the strongly convex case. Hence, a natural question for future research is: can we devise the same guarantees for *the same* point (or randomized in the same way). Another promising direction for future works is a numerical and theoretical justification of the benefits brought by the $\ell_1$-randomization, possibly, relying on our Remark 5.3.6. Finally, the question of adaptivity for zero-order optimization with adversarial noise setting remains largely open.

## 5.7 Proofs

In this section we first provide some auxiliary results and then prove the results stated in the main body of the paper.

**Additional notation** Let $\boldsymbol{W}_1, \boldsymbol{W}_2$ be two random variables, we write $\boldsymbol{W}_1 \overset{d}{=} \boldsymbol{W}_2$ to denote their equality in distribution. We also denote by $\Gamma : \mathbb{R}_+ \to \mathbb{R}_+$ the gamma function defined, for every $z > 0$, as $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) \,\mathrm{d}x$.

### Consequences of smoothness assumption

Let us first provide some immediate consequences of the smoothness assumption that we consider.

**Remark 5.7.1.** *For all $k \in \mathbb{N} \setminus \{0\}$ and all $\boldsymbol{h} \in \mathbb{R}^d$ it holds that*

$$f^{(k)}(\boldsymbol{x})[\boldsymbol{h}]^k = \sum_{|\boldsymbol{m}_1| = \cdots = |\boldsymbol{m}_k| = 1} D^{\boldsymbol{m}_1 + \cdots + \boldsymbol{m}_k} f(\boldsymbol{x}) \boldsymbol{h}^{\boldsymbol{m}_1 + \cdots + \boldsymbol{m}_k} = \sum_{|\boldsymbol{m}| = k} \frac{k!}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\boldsymbol{x}) \boldsymbol{h}^{\boldsymbol{m}} \ .$$

*Proof.* The first equality of the remark follows from the definition. For the second one it is sufficient to show that for each $\boldsymbol{m} = (m_1, \dots, m_d)^\top \in \mathbb{N}^d$ with $|\boldsymbol{m}| = k$ there exist exactly $k!/\boldsymbol{m}!$ distinct choices of $(\boldsymbol{m}_1, \dots, \boldsymbol{m}_k) \in (\mathbb{N}^d)^k$ with $|\boldsymbol{m}_1| = \ldots = |\boldsymbol{m}_k| = 1$ and $\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_k = \boldsymbol{m}$. To see this, we map $\boldsymbol{m} \in \mathbb{N}^d$ into a *word* containing letters from $\{a_1, a_2, \dots, a_d\}$ as

$$\boldsymbol{m} \mapsto W(\boldsymbol{m}) \triangleq \underbrace{a_1 \dots a_1}_{m_1-\text{times}} \underbrace{a_2 \dots a_2}_{m_2-\text{times}} \dots \underbrace{a_d \dots a_d}_{m_d-\text{times}} \ .$$

By construction, each letter $a_j$ is repeated exactly $m_j$-times in $W(\boldsymbol{m})$. Furthermore, if $|\boldsymbol{m}| = k$, then $W(\boldsymbol{m})$ contains exactly $k$ letters. From now on, fix an arbitrary $\boldsymbol{m} \in \mathbb{N}$ with $|\boldsymbol{m}| = k$. Given $(\boldsymbol{m}_1, \dots, \boldsymbol{m}_k) \in (\mathbb{N}^d)^k$ such that $|\boldsymbol{m}_1| = \ldots = |\boldsymbol{m}_k| = 1$ and $\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_k = \boldsymbol{m}$, define[1]

$$(\boldsymbol{m}_1, \dots, \boldsymbol{m}_k) \mapsto W(\boldsymbol{m}_1) + W(\boldsymbol{m}_2) + \ldots + W(\boldsymbol{m}_k) \ .$$

We observe that the condition $\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_k = \boldsymbol{m}$, implies that the word $W(\boldsymbol{m}_1) + W(\boldsymbol{m}_2) + \ldots + W(\boldsymbol{m}_k)$ is a permutation of $W(\boldsymbol{m})$. A standard combinatorial fact states that the number of distinct permutations of $W(\boldsymbol{m})$ is given by the multinomial coefficient, i.e., by $k!/\boldsymbol{m}!$. Since the mapping $(\boldsymbol{m}_1, \dots, \boldsymbol{m}_k) \mapsto W(\boldsymbol{m}_1) + W(\boldsymbol{m}_2) + \ldots + W(\boldsymbol{m}_k)$ is invertible, we conclude. □

**Lemma 5.7.2.** *Assume that $f \in \mathcal{F}_\beta(L)$ for some $\beta \geq 2$ and $L > 0$. Let $\boldsymbol{v} \in \mathbb{R}^d$ with $\|\boldsymbol{v}\| = 1$ and defined the function $g_{\boldsymbol{v}} : \mathbb{R}^d \to \mathbb{R}$ as $g_{\boldsymbol{v}}(x) \equiv \langle \boldsymbol{v} \, , \nabla f(x) \rangle$, $x \in \mathbb{R}^d$. Then $g_{\boldsymbol{v}} \in \mathcal{F}_{\beta - 1}(L)$.*

*Proof.* Set $\ell \triangleq \lfloor \beta \rfloor$. Note that since $f$ is $\ell$ times continuously differentiable, then $g_{\boldsymbol{v}}$ is $\ell - 1$ times continuously differentiable. Furthermore, for any $\boldsymbol{h}^1, \dots, \boldsymbol{h}^{\ell-1} \in \mathbb{R}^d$

$$\begin{aligned}
g_{\boldsymbol{v}}^{(\ell-1)}(\mathbf{x})[\boldsymbol{h}^1, \dots, \boldsymbol{h}^{\ell-1}] &= \sum_{|\boldsymbol{m}_1| = \ldots = |\boldsymbol{m}_{\ell-1}| = 1} D^{\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_{\ell-1}} g_{\boldsymbol{v}}(\mathbf{x}) h_1^{\boldsymbol{m}_1} \cdot \ldots \cdot \boldsymbol{h}_{\ell-1}^{\boldsymbol{m}_{\ell-1}} \\
&= \sum_{|\boldsymbol{m}_1| = \ldots = |\boldsymbol{m}_\ell| = 1} D^{\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_\ell} f(\mathbf{x}) h_1^{\boldsymbol{m}_1} \cdot \ldots \cdot \boldsymbol{h}_{\ell-1}^{\boldsymbol{m}_{\ell-1}} \boldsymbol{v}^{\boldsymbol{m}_\ell} \\
&= f^{(\ell)}(\mathbf{x})[\boldsymbol{h}^1, \dots, \boldsymbol{h}^{\ell-1}, \boldsymbol{v}] \ .
\end{aligned}$$

[1] The summation of words is defined as concatenation.

Hence, for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$ we can write by definition of the norm of a $\ell-1$-linear form

$$\begin{aligned}
&\left\| g_{\boldsymbol{v}}^{(\ell-1)}(\mathbf{x}) - g_{\boldsymbol{v}}^{(\ell-1)}(\mathbf{z}) \right\| \\
&\quad = \sup\left\{ |g_v^{(\ell-1)}(\mathbf{x})[\boldsymbol{h}^1, \dots, \boldsymbol{h}^{\ell-1}] - g_{\boldsymbol{v}}^{(\ell-1)}(\mathbf{z})[\boldsymbol{h}^1, \dots, \boldsymbol{h}^{\ell-1}]| \,:\, \|\boldsymbol{h}^j\| = 1 \; j \in [\ell-1] \right\} \\
&\quad = \sup\left\{ |f^{(\ell)}(\mathbf{x})[\boldsymbol{h}^1, \dots, \boldsymbol{h}^{\ell-1}, \boldsymbol{v}] - f^{(\ell)}(\mathbf{z})[\boldsymbol{h}^1, \dots, \boldsymbol{h}^{\ell-1}, \boldsymbol{v}]| \,:\, \|\boldsymbol{h}^j\| = 1 \; j \in [\ell-1] \right\} \\
&\quad \le \left\| f^{(\ell)}(\mathbf{x}) - f^{(\ell)}(\mathbf{z}) \right\| \le L \|\mathbf{x} - \mathbf{z}\|^{\beta-\ell} \enspace .
\end{aligned}$$

□

**Lemma 5.7.3.** *Fix some real $\beta \ge 2$ and assume that $f \in \mathcal{F}_\beta(L)$. Then, for all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$*

$$\left| f(\boldsymbol{x}) - \sum_{0 \le |\boldsymbol{m}| \le \ell} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{z})^{\boldsymbol{m}} \right| \le \frac{L}{\ell!} \|\boldsymbol{x} - \boldsymbol{z}\|^\beta \enspace .$$

*Proof.* Fix some $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$. By Taylor's theorem there exists $c \in (0, 1)$ such that

$$f(\mathbf{x}) = \sum_{0 \le |\boldsymbol{m}| \le \ell-1} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\mathbf{z})(\mathbf{x} - \mathbf{z})^{\boldsymbol{m}} + \sum_{|\boldsymbol{m}| = \ell} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\mathbf{z} + c(\mathbf{x} - \mathbf{z}))(\mathbf{x} - \mathbf{z})^{\boldsymbol{m}} \enspace .$$

Thus, invoking Remark 5.7.1 and the fact that $f \in \mathcal{F}_\beta(L)$, we can write

$$\begin{aligned}
\left| f(\mathbf{x}) - \sum_{|\boldsymbol{m}| \le \ell} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\mathbf{z})(\mathbf{x} - \mathbf{z})^{m} \right| &= |\sum_{|\boldsymbol{m}| = \ell} \frac{1}{\boldsymbol{m}!} \left( D^{\boldsymbol{m}} f(\mathbf{z} + c(\mathbf{x} - \mathbf{z})) - D^{\boldsymbol{m}} f(\mathbf{z}) \right) (\mathbf{x} - \mathbf{z})^{\boldsymbol{m}}| \\
&= \frac{1}{\ell!} |f^{(\ell)}(\mathbf{z} + c(\mathbf{x} - \mathbf{z}))[\mathbf{x} - \mathbf{z}]^\ell - f^{(\ell)}(\mathbf{z})[\mathbf{x} - \mathbf{z}]^\ell| \\
&\le \frac{L}{\ell!} \|\mathbf{x} - \mathbf{z}\|^\ell \, \|c(\mathbf{x} - \mathbf{z})\|^{\beta-\ell} \le \frac{L}{\ell!} \|\mathbf{x} - \mathbf{z}\|^\beta \enspace .
\end{aligned}$$

□

## On biases and variances

### $\ell_2$-randomization

In this subsection we study the bias and variance of $\ell_2$-randomization algorithm. It is split into two parts, with the first one focused on the bias and the second one focused on the variance.

#### Control of the bias

**Lemma 5.7.4.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. Let $r, \boldsymbol{U}^\circ, \boldsymbol{\zeta}^\circ$ be uniformly distributed on $[-1, 1], \mathcal{B}_2^d$, and $\partial B_2^d$, respectively. Then, for any $h > 0$, we have*

$$\mathbf{E}[\nabla f(\boldsymbol{x} + hr\boldsymbol{U}^\circ) r K(r)] = \frac{d}{h} \mathbf{E}[f(\boldsymbol{x} + hr\boldsymbol{\zeta}^\circ)\boldsymbol{\zeta}^\circ K(r)] \enspace .$$

*Proof.* Fix $r \in [-1,1] \setminus \{0\}$. Define $\phi : \mathbb{R}^d \to \mathbb{R}$ as $\phi(\boldsymbol{u}) = f(\mathbf{x} + hr\boldsymbol{u})K(r)$ and note that $\nabla\phi(\boldsymbol{u}) = hr\nabla f(\mathbf{x} + hr\boldsymbol{u})K(r)$. Hence, we have

$$\begin{aligned}\mathbf{E}[\nabla f(\mathbf{x} + hr\boldsymbol{U}^\circ)K(r) \mid r] = \frac{1}{hr}\mathbf{E}[\nabla\phi(\boldsymbol{U}^\circ) \mid r] &= \frac{d}{hr}\mathbf{E}[\phi(\boldsymbol{\zeta}^\circ)\boldsymbol{\zeta}^\circ \mid r]\\ &= \frac{d}{hr}K(r)\mathbf{E}[f(\mathbf{x} + hr\boldsymbol{\zeta}^\circ)\boldsymbol{\zeta}^\circ \mid r] \enspace ,\end{aligned}$$

where the second equality is obtained from a version of Stokes' theorem (see e.g., Zorich, 2016, Section 13.3.5, Exercise 14a). Multiplying by $r$ from both sides, using the fact that $r$ follows continuous distribution, and taking total expectation concludes the proof. □

*Proof of Lemma 5.3.2.* Using Lemma 5.7.4, the fact that $\int_{-1}^{1} rK(r)\,\mathrm{d}r = 1$, and the variational representation of the Euclidean norm, we can write

$$\|\mathbf{E}[\mathbf{g}_t^\circ \mid \mathbf{x}_t] - \nabla f(\mathbf{x}_t)\| = \sup_{\boldsymbol{v} \in \partial B_2^d} \mathbf{E}[\big(\nabla_{\boldsymbol{v}} f(\mathbf{x} + h_t r_t \boldsymbol{U}^\circ) - \nabla_{\boldsymbol{v}} f(\mathbf{x})\big) r_t K(r_t)] \enspace , \tag{5.13}$$

where we recall that $\boldsymbol{U}^\circ$ is uniformly distributed on $\mathcal{B}_2^d$. Lemma 5.7.2 asserts that for any $\boldsymbol{v} \in \partial B_2^d$ the directional gradient $\nabla_{\boldsymbol{v}} f(\cdot)$ is $(\beta - 1, L)$-Hölder. Thus, thanks to Lemma 5.7.3, the following Taylor's expansion holds

$$\nabla_{\boldsymbol{v}} f(\mathbf{x}_t + h_t r_t \boldsymbol{U}^\circ) = \nabla_{\boldsymbol{v}} f(\mathbf{x}_t) + \sum_{1 \leq |\boldsymbol{m}| \leq \ell - 1} \frac{(r_t h_t)^{|\boldsymbol{m}|}}{\boldsymbol{m}!} D^{\boldsymbol{m}} \nabla_{\boldsymbol{v}} f(\mathbf{x}_t)(\boldsymbol{U}^\circ)^{\boldsymbol{m}} + R(h_t r_t \boldsymbol{U}^\circ) \enspace , \tag{5.14}$$

where the residual term $R(\cdot)$ satisfies $|R(\mathbf{x})| \leq \frac{L}{(\ell-1)!}\|\mathbf{x}\|^{\beta-1}$.

Substituting Eq. (5.14) into Eq. (5.13) and using the “zeroing-out” properties of the kernel $K$, we deduce that

$$\|\mathbf{E}[\mathbf{g}_t^\circ \mid \mathbf{x}_t] - \nabla f(\mathbf{x}_t)\| \leq \kappa_\beta h_t^{\beta-1} \frac{L}{(\ell-1)!} \mathbf{E}\|\boldsymbol{U}^\circ\|^{\beta-1} = \kappa_\beta h_t^{\beta-1} \frac{L}{(\ell-1)!} \frac{d}{d + \beta - 1} \enspace ,$$

where the last equality is obtained from the fact that $\mathbf{E}\|\boldsymbol{U}^\circ\|^q = \frac{d}{d+q}$, for any $q \geq 0$. □

### $\ell_1$-randomization

In this subsection we study bias and variance of $\ell_1$-randomization algorithm.

Let $\zeta$ be a real valued random variable with $\mathbf{E}[\zeta^2] \leq 4\sigma^2$ and $\boldsymbol{\zeta}^\diamond$ be distributed uniformly on $\partial B_1^d$. Assume that both are independent from each other. In this section analyze the estimator

$$\mathbf{g}_{\mathbf{x},h}^\diamond = \frac{d}{2h}\left(f(\mathbf{x} + hr\boldsymbol{\zeta}^\diamond) - f(\mathbf{x} - hr\boldsymbol{\zeta}^\diamond) + \zeta\right) \operatorname{sign}(\boldsymbol{\zeta}^\diamond) K(r) \enspace , \tag{5.15}$$

which coincides with the estimator in (5.3) at time $t = 1, \ldots, T$ with $\zeta = \xi_t - \xi_t'$.

### Control of bias

**Lemma 5.7.5.** *Let $\boldsymbol{U}^\diamond$ be uniformly distributed on $\mathcal{B}_1^d$ and $\boldsymbol{\zeta}^\diamond$ be uniformly distributed on $\partial B_1^d$. Fix some $\boldsymbol{x} \in \mathbb{R}^d$ and $h > 0$, and let Assumption 5.3.1 be fulfilled, then the estimator in Eq. (5.15) satisfies*

$$\mathbf{E}[\boldsymbol{g}^\diamond_{\boldsymbol{x},h}] = \mathbf{E}[\nabla f(\boldsymbol{x} + hr\boldsymbol{U}^\diamond)rK(r)] \enspace .$$

*Proof.* The proof is analogous to that of Lemma 5.7.4 using (Akhavan et al., 2022a, Theorem 6). □

In order to control the bias of the estimator in Eq. (5.15), we need the following result, which controls the moments of the Euclidean norm of $\boldsymbol{U}^\diamond$.

**Lemma 5.7.6.** *Let $\boldsymbol{U}^\diamond \in \mathbb{R}^d$ be distributed uniformly on $\mathcal{B}_1^d$, then for any $\beta \geq 2$ it holds that*

$$\mathbf{E}\left\|\boldsymbol{U}^\diamond\right\|^\beta \leq \frac{d^{\frac{\beta}{2}}\Gamma(\beta+1)\Gamma(d+1)}{\Gamma(d+\beta+1)} \enspace .$$

*Proof.* Let $\boldsymbol{W} = (W_1, \dots, W_d), W_{d+1}$ be $\mathrm{i.i.d.}$ random variables following Laplace distribution with mean $0$ and scale parameter $1$. Then, following (Barthe et al., 2005, Theorem 1) we have

$$\boldsymbol{U}^\diamond \overset{d}{=} \frac{\boldsymbol{W}}{\left\|\boldsymbol{W}\right\|_1 + |W_{d+1}|} \enspace ,$$

where the above equality holds in distribution. Furthermore, (Barthe et al., 2005, Theorem 2) (see also Rachev and Ruschendorf (1991); Schechtman and Zinn (1990)) states that

$$\frac{(\boldsymbol{W}, |W_{d+1}|)}{\left\|\boldsymbol{W}\right\|_1 + |W_{d+1}|} \qquad \text{and} \qquad \left\|\boldsymbol{W}\right\|_1 + |W_{d+1}| \enspace ,$$

are independent. Hence, we can write that $\mathbf{E}\left\|\boldsymbol{U}^\diamond\right\|^\beta$ equals to

$$\mathbf{E}\left[\left(\frac{\sum_{j=1}^d W_j^2}{(\left\|\boldsymbol{W}\right\|_1 + |W_{d+1}|)^2}\right)^{\beta/2}\right] = \frac{\mathbf{E}\left\|\boldsymbol{W}\right\|^\beta}{\mathbf{E}\left\|(\boldsymbol{W}, W_{d+1})\right\|_1^\beta} \enspace , \tag{5.16}$$

where the equality follows by the independence recalled above. Note that for any $j = 1, \dots, d$ it holds that $|W_j|$ is $\exp(1)$ random variable. Thus, since $\beta \geq 2$, we can write by Jensen's inequality

$$\mathbf{E}\left\|\boldsymbol{W}\right\|^\beta = d^{\frac{\beta}{2}}\mathbf{E}\left(\frac{1}{d}\sum_{j=1}^d W_j^2\right)^{\beta/2} \leq d^{\frac{\beta}{2}-1}\sum_{j=1}^d \mathbf{E}[W_j^\beta] = d^{\frac{\beta}{2}}\mathbf{E}[W_1^\beta] = d^{\frac{\beta}{2}}\Gamma(\beta+1) \enspace . \tag{5.17}$$

It remains to provide a suitable expression (or lower bound) for $\mathbf{E}\left\|(\boldsymbol{W}, W_{d+1})\right\|_1^\beta$. We observe that $\left\|(\boldsymbol{W}, W_{d+1})\right\|_1$ follows Erlang distribution with parameters $(d+1, 1)$ (as a sum of $d+1$

i.i.d. $\exp(1)$ random variables). Hence, recalling the expression for the density of the Erlang distribution

$$\mathbf{E}\,\|(\boldsymbol{W}, W_{d+1})\|_1^{\beta} = \frac{1}{\Gamma(d+1)}\int_0^\infty x^{d+\beta}\exp(-x)\,\mathrm{d}x = \frac{\Gamma(d+\beta+1)}{\Gamma(d+1)}\ . \tag{5.18}$$

Substituting Eqs. (5.17)–(5.18) into Eq. (5.16) we conclude. □

We are in position to derive an upper bound on the bias of the gradient estimator in Eq. (5.15).

*Proof of Lemma 5.3.4.* Invoking Lemma 5.7.5 and following the same lines as in the proof of Lemma 5.3.2, we deduce that

$$\|\mathbf{E}[\mathbf{g}_t^\diamond \mid \mathbf{x}_t] - \nabla f(\mathbf{x}_t)\| \le \kappa_\beta h_t^{\beta-1}\frac{L}{(\ell-1)!}\mathbf{E}\,\|\boldsymbol{U}^\diamond\|^{\beta-1} \le \kappa_\beta h_t^{\beta-1}\frac{L}{(\ell-1)!}\frac{d^{\frac{\beta-1}{2}}\Gamma(\beta)\Gamma(d+1)}{\Gamma(d+\beta)}\ ,$$

where the last inequality is thanks to Lemma 5.7.6. Recall the following property of Gamma function: for any $z > 0$ we have $\Gamma(z+1) = z\Gamma(z)$. Therefore, applying this property iteratively and recalling the definition of $\ell$, we deduce that

$$\frac{\Gamma(d+1)}{\Gamma(d+\beta)} = \frac{\Gamma(d+1)}{\Gamma\big(d+\underbrace{(\beta-\ell)}_{\in(0,1]}\big)\prod_{i=1}^{\ell}(d+\beta-i)} \le \frac{(d+\beta-\ell)^{1-(\beta-\ell)}}{\prod_{i=1}^{\ell}(d+\beta-i)} \le \frac{1}{d^{\beta-1}}\ ,$$

where the first inequality is obtained from (Feng, 2010, Remark 2.1.1). Finally, for the term $\frac{\Gamma(\beta)}{(\ell-1)!}$, we proceed analogously and deduce that it is bounded by $\ell^{\beta-\ell}$. Combining the last three displays concludes the proof. □

### Control of variance

We now address how to control the variance of the $\ell_1$-randomized estimator.

**Lemma 5.7.7.** *For all $d \ge 3$ it holds that*

$$\mathbf{E}\left[(1+\sqrt{d}\,\|\boldsymbol{\zeta}^\diamond\|)^2\right] \le \left(1+\sqrt{\frac{2d}{d+1}}\right)^2, \qquad \mathbf{E}\left[(1+\sqrt{d}\,\|\boldsymbol{\zeta}^\diamond\|)^2\,\|\boldsymbol{\zeta}^\diamond\|^2\right] \le \frac{2}{d+1}\left(3+\sqrt{8+\frac{592}{(d+3)^2}}+\frac{22}{d}\right).$$

*Proof.* In what follows we will make use of the following expression for the moments of Dirichlet distribution

$$\mathbf{E}[(\boldsymbol{\zeta}^\diamond)^{\boldsymbol{m}}] = \frac{\Gamma(d)}{\Gamma(d+|\boldsymbol{m}|)}\prod_{i=1}^{d}\Gamma(m_i+1) = \frac{(d-1)!\boldsymbol{m}!}{(d-1+|\boldsymbol{m}|)!}\ , \tag{5.19}$$

for any multi-index $\boldsymbol{m} = (m_1, \dots, m_d) \in \mathbb{N}^d$ with even coordinates.

**First bound.** The expression for second moments of Dirichlet distribution in Eq. (5.19) yields

$$\mathbf{E}[(1+\sqrt{d}\,\|\boldsymbol{\zeta}^\diamond\|)^2] \le 1 + 2\sqrt{d\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^2} + d\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^2 \le 1 + 2\sqrt{\frac{2d}{d+1}} + \frac{2d}{d+1} = \left(1+\sqrt{\frac{2d}{d+1}}\right)^2 \quad .$$

The proof of the first claimed bound is completed.

**Second bound.** We repeat similar argument for $\mathbf{E}[(1+\sqrt{d}\,\|\boldsymbol{\zeta}^\diamond\|)^2\,\|\boldsymbol{\zeta}^\diamond\|^2]$. By Jensen's inequality, it holds that

$$\mathbf{E}[(1+\sqrt{d}\,\|\boldsymbol{\zeta}^\diamond\|)^2\,\|\boldsymbol{\zeta}^\diamond\|^2] \le \mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^2 + 2\sqrt{d\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^6} + d\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^4 \quad .$$

We already know that $\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^2 = 2/(d+1)$ and it remains to evaluate $\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^6$ and $\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^4$. Using multinomial identity and the expression for the moments in Eq. (5.19), we deduce

$$\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^6 = \sum_{|\boldsymbol{m}|=3} \frac{6}{\boldsymbol{m}!}\mathbf{E}[(\boldsymbol{\zeta}^\diamond)^{2\boldsymbol{m}}] = \sum_{|\boldsymbol{m}|=3} \frac{6}{\boldsymbol{m}!}\cdot\frac{(d-1)!(2\boldsymbol{m})!}{(d+5)!} = 6\frac{(d-1)!}{(d+5)!}\sum_{|\boldsymbol{m}|=3}\frac{(2\boldsymbol{m})!}{\boldsymbol{m}!} \quad ,$$

and

$$\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^4 = \sum_{|\boldsymbol{m}|=2} \frac{2}{\boldsymbol{m}!}\mathbf{E}[(\boldsymbol{\zeta}^\diamond)^{2\boldsymbol{m}}] = \sum_{|\boldsymbol{m}|=3} \frac{2}{\boldsymbol{m}!}\cdot\frac{(d-1)!(2\boldsymbol{m})!}{(d+3)!} = 2\frac{(d-1)!}{(d+3)!}\sum_{|\boldsymbol{m}|=2}\frac{(2\boldsymbol{m})!}{\boldsymbol{m}!} \quad .$$

Direct calculations show that $\sum_{|\boldsymbol{m}|=3}\frac{(2\boldsymbol{m})!}{\boldsymbol{m}!} = \frac{4}{3}d((d+3)^2+74)$ and $\sum_{|\boldsymbol{m}|=2}\frac{(2\boldsymbol{m})!}{\boldsymbol{m}!} = 2d(d+11)$. Thus, after some basic algebraic simplifications, we get the following expressions for all $d \ge 1$

$$\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^6 = \frac{8d!(d+3)^2}{(d+5)!}\left(1+\frac{74}{(d+3)^2}\right) \qquad \text{and} \qquad \mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^4 = \frac{4d!d}{(d+3)!}\left(1+\frac{11}{d}\right) \quad .$$

Therefore, we have the following upper bound

$$\begin{aligned}
\mathbf{E}[(1+\sqrt{d}\,\|\boldsymbol{\zeta}^\diamond\|)^2\,\|\boldsymbol{\zeta}^\diamond\|^2] &\le \mathbf{E}[\|\boldsymbol{\zeta}^\diamond\|^2] + 2\sqrt{d\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^6} + d\mathbf{E}\,\|\boldsymbol{\zeta}^\diamond\|^4 \\
&= \frac{2}{d+1} + 2\sqrt{\frac{8d!d(d+3)^2}{(d+5)!}\left(1+\frac{74}{(d+3)^2}\right)} + \frac{4d!d^2}{(d+3)!}\left(1+\frac{11}{d}\right) \\
&\le \frac{2}{d+1} + 2\sqrt{\frac{8}{(d+1)^2}\left(1+\frac{74}{(d+3)^2}\right)} + \frac{4}{d+1}\left(1+\frac{11}{d}\right)
\end{aligned}$$

The proof of the second claimed bound follows by rearranging the right-hand-side of the above inequality. □

### A technical lemma

In our proofs we will often need to deal with various recursive relations. In this section we provide a result, which are used extensively in nearly every single proof, it is a direct adaptations of (Akhavan et al., 2020, Lemma D.1).

**Lemma 5.7.8.** *Let $\{\delta_t\}_{t\geq 1}$ be a sequence of real numbers such that for all integers $t > t_0 \geq 1$,*

$$\delta_{t+1} \leq \left(1 - \frac{2}{t}\right)\delta_t + \sum_{i=1}^{N} \frac{a_i}{t^{p_i+1}} \quad , \tag{5.20}$$

*where $p_i \in (0, 2)$ and $a_i \geq 0$ for $i \in [N]$. Then for $t \geq t_0 \geq 3$, we have*

$$\delta_t \leq \frac{2(t_0 - 1)\delta_{t_0}}{t} + \sum_{i=1}^{N} \frac{a_i}{(2 - p_i)t^{p_i}} \quad . \tag{5.21}$$

*Proof.* For any fixed $t > 0$ the convexity of the mapping $u \mapsto g(u) = (t + u)^{-p}$ implies that $g(1) - g(0) \geq g'(0)$, i.e., $\frac{1}{t^p} - \frac{1}{(t+1)^p} \leq \frac{p}{t^{p+1}}$. Thus, using the fact that $\frac{1}{t^p} - \frac{p}{t^{p+1}} = \frac{(2-p)+(t-2)}{t^{p+1}} \leq \frac{1}{(t+1)^p}$,

$$\frac{a_i}{t^{p+1}} \leq \frac{a_i}{2 - p}\left\{\frac{1}{(t + 1)^p} - \left(1 - \frac{2}{t}\right)\frac{1}{t^p}\right\} \quad . \tag{5.22}$$

Using Eq. (5.20) and Eq. (5.22) and rearranging terms, for any $t \geq t_0$ we get

$$\delta_{t+1} - \sum_{i=1}^{N} \frac{a_i}{(2 - p_i)(t + 1)^{p_i}} \leq \left(1 - \frac{2}{t}\right)\left\{\delta_t - \sum_{i=1}^{N} \frac{a_i}{(2 - p_i)t^{p_i}}\right\} \quad .$$

Letting $\tau_t = \delta_t - \sum_{i=1}^{N} \frac{a_i}{(2-p_i)t^{p_i}}$ we have $\tau_{t+1} \leq (1 - \frac{2}{t})\tau_t$. Now, if $\tau_{t_0} \leq 0$ then $\tau_t \leq 0$ for any $t \geq t_0$ and thus (5.21) holds. Otherwise, if $\tau_{t_0} > 0$ then for $t \geq t_0 + 1$ we have

$$\tau_t \leq \tau_{t_0} \prod_{i=t_0}^{t-1}\left(1 - \frac{2}{i}\right) \leq \tau_{t_0} \prod_{i=t_0}^{t-1}\left(1 - \frac{1}{i}\right) \leq \frac{(t_0 - 1)\tau_{t_0}}{t} \leq \frac{2(t_0 - 1)\delta_{t_0}}{t} \quad .$$

Thus, (5.21) holds in this case as well. □

### Upper bounds: only smoothness. Proof of Theorem 5.4.2

In this section we provide the proof of Theorem 5.4.2. The proof will be split into two parts: for (5.2) and (5.3) respectively. But first we start with the part of the proof that is common for both Algorithms. Both of these proofs follow from Lemma 5.4.1, which, we recall, states that

$$\mathbf{E}\left[\|\nabla f(\mathbf{x}_S)\|^2\right] \leq \frac{2\delta_1 + \sum_{t=1}^{T} \eta_t \left(b_t^2 + \bar{L}\eta_t v_t\right)}{\sum_{t=1}^{T} \eta_t \left(1 - \bar{L}\eta_t m\right)} \quad , \tag{5.23}$$

where $\delta_1 = \mathbf{E}[f(\mathbf{x}_1)] - f^\star$. In particular, using corresponding bounds on the variance and bias, we substitute these values in the above inequality and focus on deriving an upper bound for the resulting sequences. Further unifying the proof, we introduce the following short-hand notation

$$\Xi_T \triangleq d^{-\frac{2(\beta-1)}{2\beta-1}} T^{-\frac{\beta}{2\beta-1}} \enspace .$$

Using this notation, algorithm 5.1 with either (5.2) or (5.3) have the following initialization of parameters

$$\eta_t = \min\left(\frac{\mathfrak{y}}{d}, \Xi_T\right) \qquad \text{and} \qquad h_t = \mathfrak{h} \cdot T^{-\frac{1}{2(2\beta-1)}} \enspace ,$$

where we recall that

$$(\mathfrak{y}, \mathfrak{h}) = \begin{cases} \left(\frac{1}{8\kappa\bar{L}},\, d^{\frac{1}{2\beta-1}}\right) & \text{for estimator (5.2)} \\ \left(\frac{d-2}{2\bar{L}\bar{\mathtt{C}}_{d,1}d},\, d^{\frac{2\beta+1}{4\beta-2}}\right) & \text{for estimator (5.3)} \end{cases} \enspace .$$

Furthermore, in the notation of Lemma 5.4.1 and, in particular, of Eq. (5.23) above, the bounds in Lemma 5.3.3 and Lemma 5.3.5 imply that the choice of $\eta_t$ for both Algorithms ensures that

$$1 - \bar{L}\eta_t m \leq \frac{1}{2} \enspace .$$

Thus, as a consequence of the above argument and Eq. (5.23), both Algorithms satisfy

$$\mathbf{E}\left[\|\nabla f(\mathbf{x}_S)\|^2\right] \leq \left(\sum_{t=1}^{T}\eta_t\right)^{-1}\left(4\delta_1 + 2\sum_{t=1}^{T}\eta_t b_t^2 + 2\bar{L}\sum_{t=1}^{T}\eta_t^2 v_t\right) \enspace . \tag{5.24}$$

Furthermore, since $\eta_t = \min(\mathfrak{y}/d, \Xi_T)$, then in either case we have

$$\left(\sum_{t=1}^{T}\eta_t\right)^{-1} = \max\left(\frac{d}{T\mathfrak{y}}, \frac{1}{T\Xi_T}\right) \leq \frac{d}{T\mathfrak{y}} + \frac{1}{T\Xi_T} \enspace .$$

Substituting the above into Eq. (5.24) we deduce that

$$\mathbf{E}\left[\|\nabla f(\mathbf{x}_S)\|^2\right] \leq \left(\frac{d}{T\mathfrak{y}} + \frac{1}{T\Xi_T}\right)\left(4\delta_1 + 2\sum_{t=1}^{T}\eta_t b_t^2 + 2\bar{L}\sum_{t=1}^{T}\eta_t^2 v_t\right) \enspace .$$

Finally, by the definition of $\eta_t$ we have $\eta_t \leq \Xi_T$ for all $t = 1, \ldots, T$. Thus, the above can be further bounded as

$$\mathbf{E}\left[\|\nabla f(\mathbf{x}_S)\|^2\right] \leq \left(\frac{d}{\mathfrak{y}} + \frac{1}{\Xi_T}\right)\frac{4\delta_1}{T} + 2\left(\frac{d\Xi_T}{T\mathfrak{y}} + \frac{1}{T}\right)\sum_{t=1}^{T}\left\{b_t^2 + \bar{L}\Xi_T v_t\right\} \enspace . \tag{5.25}$$

In the rest of the proof we use the algorithm specific bounds on $b_t$ and $v_t$ as well as the particular choice of $\mathfrak{y}$ in order to further bound the above inequality.

### Part I: for estimator 5.2

Lemma 5.3.2 (for the bias) and Lemma 5.3.3 (for the variance) in the notation of Eq. (5.23) read as

$$b_t^2 \leq \left(\frac{\kappa_\beta L}{(\ell-1)!}\right)^2 h_t^{2(\beta-1)} \quad \text{and} \quad v_t = 4d\kappa \bar{L}^2 h_t^2 + \frac{d^2\sigma^2\kappa}{2h_t^2}\,, \quad \text{and} \quad m = 4d\kappa\ .$$

Substituting the above into Eq. (5.25), we deduce that

$$\begin{aligned}\mathbf{E}\left\|\nabla f(\mathbf{x}_S)\right\|^2 \leq &\left(\frac{d}{\mathfrak{y}} + \Xi_T^{-1}\right)\frac{4\delta_1}{T} \\ &+ \left(\frac{d\Xi_T}{T\mathfrak{y}} + \frac{1}{T}\right)\sum_{t=1}^{T}\left\{\mathtt{A}_3 h_t^{2(\beta-1)} + \Xi_T d^2\big(\mathtt{A}_4 d^{-1}h_t^2 + \mathtt{A}_5 h_t^{-2}\big)\right\} \\ \leq &\left(\frac{d}{\mathfrak{y}} + \Xi_T^{-1}\right)\frac{4\delta_1}{T} + \frac{d\Xi_T + 1}{T}\sum_{t=1}^{T}\left\{\mathtt{A}_6 h_t^{2(\beta-1)} + \mathtt{A}_7 d^2\Xi_T\big(d^{-1}h_t^2 + h_t^{-2}\big)\right\}\end{aligned} \tag{5.26}$$

where $\mathtt{A}_3 = \left(\frac{\kappa_\beta L}{(\ell-1)!}\right)^2$, $\mathtt{A}_4 = 4\kappa\bar{L}^3$, $\mathtt{A}_5 = \frac{\kappa\sigma^2\bar{L}}{2}$, and $\mathtt{A}_6 = 2\mathtt{A}_3\big(\mathfrak{y}^{-1}+1\big)$, $\mathtt{A}_7 = 2\big(\mathfrak{y}^{-1}+1\big)\big(\mathtt{A}_4+\mathtt{A}_5\big)$. Since $h_t = h_T$ for $t = 1,\dots,T$, then Eq. (5.26) reads as

$$\mathbf{E}\left\|\nabla f(\mathbf{x}_S)\right\|^2 \leq \left(\frac{d}{\mathfrak{y}} + \Xi_T^{-1}\right)\frac{4\delta_1}{T} + (d\Xi_T + 1)\Big(\mathtt{A}_6 h_T^{2(\beta-1)} + \mathtt{A}_7 d^2\Xi_T\big(d^{-1}h_T^2 + h_T^{-2}\big)\Big). \tag{5.27}$$

Substituting the expressions for $\Xi_T$ and $h_T$ into the above bound, the right hand side of Eq. (5.27) reduces to

$$\frac{4d}{T\mathfrak{y}}\delta_1 + \left\{4\delta_1 + \left(\left(\frac{d}{T^\beta}\right)^{\frac{1}{2\beta-1}} + 1\right)\left(\mathtt{A}_6 + \mathtt{A}_7\left(1 + d^{\frac{5-2\beta}{2\beta-1}}T^{-\frac{2}{2\beta-1}}\right)\right)\right\}\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}}\ .$$

To conclude, we note that the assumption $T \geq d^{\frac{1}{\beta}}$, implies that for all $\beta \geq 2$ we have $d^{\frac{5-2\beta}{2\beta-1}}T^{-\frac{2}{2\beta-1}} \leq 1$ and $({}^{d}/_{T^\beta})^{\frac{1}{2\beta-1}} \leq 1$. Therefore, the final bound reads as

$$\mathbf{E}\left[\left\|\nabla f(\mathbf{x}_S)\right\|^2\right] \leq \frac{4d}{T\mathfrak{y}}\delta_1 + \Big(4\delta_1 + 2\big(\mathtt{A}_6 + 2\mathtt{A}_7\big)\Big)\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}} \leq \big(\mathtt{A}_1\delta_1 + \mathtt{A}_2\big)\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}},$$

where we introduced $\mathtt{A}_1 = 4(\mathfrak{y}^{-1}+1)$ and $\mathtt{A}_2 = 2\big(\mathtt{A}_6 + 2\mathtt{A}_7\big)$.

### Part II: for estimator 5.3

Lemma 5.3.4 (for the the bias) and Lemma 5.3.5 (more precisely, Eq. (5.7) for the variance) imply that

$$b_t^2 \leq (\kappa_\beta \ell L)^2 h_t^{2(\beta-1)} d^{1-\beta}, \qquad v_t = \frac{\bar{\mathtt{C}}_{d,2} d^2 \kappa \bar{L}^2 h_t^2}{(d-2)(d+1)} + \frac{d^3 \sigma^2 \kappa}{2h_t^2}, \qquad \text{and} \qquad m = \frac{\bar{\mathtt{C}}_{d,1} d^2 \kappa}{d-2} \ ,$$

with $\ell = \lfloor \beta \rfloor$. Similarly to the previous paragraph, from Eq. (5.25) and the above bounds on $b_t, v_t, m$ we deduce that

$$\begin{aligned} \mathbf{E} \left\| \nabla f(\mathbf{x}_S) \right\|^2 \leq (d\Xi_T{+}1) \left( \mathtt{A}_6 d^{1-\beta} h_T^{2(\beta-1)} {+} \Xi_T \left( \mathtt{A}_7 h_T^2 {+} \mathtt{A}_8 d^3 h_T^{-2} \right) \right) \\ + \left( \frac{d}{\mathfrak{y}} {+} \Xi_T^{-1} \right) \frac{4\delta_1}{T} \ , \end{aligned} \tag{5.28}$$

where the constants are defined as

$$\mathtt{A}_6 = 2(\kappa_\beta \ell L)^2 \big( \mathfrak{y}^{-1} + 1 \big), \quad \mathtt{A}_7 = \frac{2\bar{\mathtt{C}}_{d,2} d^2 \kappa \bar{L}^3}{(d-2)(d+1)} \big( \mathfrak{y}^{-1} + 1 \big), \quad \mathtt{A}_8 = \bar{L} \sigma^2 \kappa \big( \mathfrak{y}^{-1} + 1 \big) \ .$$

Substituting $\Xi_T$ and $h_T$ into Eq. (5.28), we deduce that

$$\mathbf{E} \left\| \nabla f(\mathbf{x}_S) \right\|^2 \leq \frac{4d}{T\mathfrak{y}} \delta_1 + \left\{ 4\delta_1 + \left( \left( \frac{d}{T^\beta} \right)^{\frac{1}{2\beta-1}} + 1 \right) \left( \mathtt{A}_6 + \mathtt{A}_8 + \mathtt{A}_7 \left( \frac{d^{5-2\beta}}{T^2} \right)^{\frac{1}{2\beta-1}} \right) \right\} \left( \frac{d^2}{T} \right)^{\frac{\beta-1}{2\beta-1}} \ .$$

Finally, we assumed that $T \geq d^{\frac{1}{\beta}}$, which implies that both $\frac{d}{T^\beta}$ and $\frac{d^{5-2\beta}}{T^2}$ are upper bounded by one. Thus, the final bound reads as

$$\mathbf{E} \left\| \nabla f(\mathbf{x}_S) \right\|^2 \leq (\mathtt{A}_1 \delta_1 + \mathtt{A}_2) \left( \frac{d^2}{T} \right)^{\frac{\beta-1}{2\beta-1}} \ ,$$

where $\mathtt{A}_1 = 4(\mathfrak{y}^{-1} + 1)$, and $\mathtt{A}_2 = 2\left( \mathtt{A}_6 + \mathtt{A}_7 + \mathtt{A}_8 \right)$.

### Master lemmas for gradient dominant and strongly convex case

*Proof of Theorem 5.4.4.* For compactness we write $\mathbf{E}_t[\cdot]$ in place of $\mathbf{E}[\cdot \mid \mathbf{x}_t]$. Using Lipschitz continuity of $\nabla f$ (see e.g. Bubeck, 2015, Lemma 3.4) and the update of the algorithm in Eq. (5.1) we can write

$$\begin{aligned} \mathbf{E}_t[f(\mathbf{x}_{t+1})] &\leq f(\mathbf{x}_t) - \eta_t \left\langle \nabla f(\mathbf{x}_t) \, , \mathbf{E}_t[\mathbf{g}_t] \right\rangle + \frac{\bar{L}\eta_t^2}{2} \mathbf{E}_t \left[ \left\| \mathbf{g}_t \right\|^2 \right] \\ &\leq f(\mathbf{x}_t) - \eta_t \left\| \nabla f(\mathbf{x}_t) \right\|^2 + \eta_t \left\| \nabla f(\mathbf{x}_t) \right\| \left\| \mathbf{E}_t[\mathbf{g}_t] - \nabla f(\mathbf{x}_t) \right\| + \frac{\bar{L}\eta_t^2}{2} \mathbf{E}_t \left[ \left\| \mathbf{g}_t \right\|^2 \right] \ . \end{aligned}$$

Furthermore, invoking the assumption of the bias and the variance of $\mathbf{g}_t$ and using the fact that $2ab \leq a^2 + b^2$ we deduce for any iterative procedure expressed in Eq. (5.1) satisfies

$$\delta_{t+1} \leq \delta_t - \frac{\eta_t}{2}(1 - \bar{L}\eta_t \mathsf{V}_1)\mathbf{E}[\|\nabla f(\mathbf{x}_t)\|^2] + \frac{\eta_t}{2}\left(b^2L^2h_t^{2(\beta-1)} + \bar{L}\eta_t\left(\mathsf{V}_2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right)\right) \enspace ,$$

where $\delta_t = \mathbf{E}[f(\mathbf{x}_t) - f^\star]$. Furthermore, our choice of the step-size $\eta_t$ ensures that $1 - \bar{L}\eta_t\mathsf{V}_1 \geq \frac{1}{2}$. Then, since $f$ is $\alpha$-gradient dominant we deduce that

$$\delta_{t+1} \leq \delta_t\left(1 - \frac{\eta_t\alpha}{2}\right) + \frac{\eta_t}{2}\left(b^2L^2h_t^{2(\beta-1)} + \bar{L}\eta_t\left(\mathsf{V}_2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right)\right) \enspace . \tag{5.29}$$

In what follows we will analyze the above recursion. Let us set $T_0 = \lfloor \frac{8\bar{L}\mathsf{V}_1}{\alpha} \rfloor$—the moment when $\eta_t$ switches its regime. In the first part of the proof we suppose that $T > T_0$, then we analyse the case of $T \geq T_0$.

**The first part:** $T > T_0$**.** In this part the recursion (5.29) has two different regimes, depending on the value of $\eta_t$.

In the first regime, for any $t = T_0 + 1, \ldots, T$, we have $\eta_t = \frac{4}{\alpha t}$ and (5.29) can be written as

$$\delta_{t+1} \leq \delta_t\left(1 - \frac{2}{t}\right) + 2b^2L^2 \cdot \frac{h_t^{2(\beta-1)}}{\alpha t} + \frac{8\bar{L}}{\alpha^2t^2}\left(\mathsf{V}_2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right) \enspace . \tag{5.30}$$

Additionally in this regime of $t$, we have $h_t = \left(\frac{4\bar{L}\mathsf{V}_3}{b^2t}\right)^{\frac{1}{2\beta}}$. Thus, substituting this expression for $h_t$ into Eq. (5.30), we deduce that

$$\delta_{t+1} \leq \delta_t\left(1 - \frac{2}{t}\right) + \frac{\mathtt{A}_3}{\alpha \wedge \alpha^2} \cdot \mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}} t^{-\frac{2\beta-1}{\beta}} + \frac{\mathtt{A}_4}{\alpha^2} \cdot \mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}} t^{-\frac{2\beta+1}{\beta}} \enspace ,$$

where $\mathtt{A}_3 = 2^{3-\frac{2}{\beta}}(L^2 + \sigma^2)\bar{L}^{1-\frac{1}{\beta}}$, and $\mathtt{A}_4 = 2^{3+\frac{2}{\beta}}\bar{L}^{1+\frac{1}{\beta}}$. Applying Lemma 5.7.8 to the above recursion we obtain

$$\delta_T \leq \frac{2T_0}{T}\delta_{T_0+1} + \frac{\beta\mathtt{A}_3}{(\beta{+}1)(\alpha{\wedge}\alpha^2)} \cdot \mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}} T^{-\frac{\beta-1}{\beta}} + \frac{\beta\mathtt{A}_4}{(3\beta{+}1)\alpha^2} \cdot \mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}} T^{-\frac{\beta+1}{\beta}}. \tag{5.31}$$

In the second regime of $t \in [1, T_0]$, we have $h_t = \left(\frac{4\bar{L}\mathsf{V}_3}{b^2T}\right)^{\frac{1}{2\beta}}$, $\eta_t = \frac{1}{2\bar{L}\mathsf{V}_1}$, and $\frac{4}{(T_0+1)\alpha} \leq \eta_t \leq \frac{4}{T_0\alpha}$. Substituting $h_t$ and $\eta_t$ into Eq. (5.29), for $1 \leq t \leq T_0$

$$\begin{aligned}\delta_{t+1} \leq \delta_t\left(1 - \frac{2}{T_0 + 1}\right) + \frac{2^{3-\frac{2}{\beta}}\bar{L}^{1-\frac{1}{\beta}}}{T_0(\alpha \wedge \alpha^2)} \cdot \mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}}\left(L^2T^{-\frac{\beta-1}{\beta}} + \sigma^2\frac{T^{\frac{1}{\beta}}}{T_0}\right)\\ + \frac{2^{3+\frac{2}{\beta}}\bar{L}^{1+\frac{1}{\beta}}}{\alpha^2T_0^2} \cdot \mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}} T^{-\frac{1}{\beta}} \enspace .\end{aligned}$$

Using a rough bound $1 - \frac{2}{T_0+1} \leq 1$ and unfolding the above recursion, we obtain from the above

$$\delta_{T_0+1} \leq \delta_1 + \frac{2^{3-\frac{2}{\beta}}\bar{L}^{1-\frac{1}{\beta}}}{\alpha\wedge\alpha^2}\cdot \mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}}\left(L^2T^{-\frac{\beta-1}{\beta}} + \sigma^2\frac{T^{\frac{1}{\beta}}}{T_0}\right) + \frac{2^{3+\frac{2}{\beta}}\bar{L}^{1+\frac{1}{\beta}}}{\alpha^2 T_0}\cdot\mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}T^{-\frac{1}{\beta}} \ .$$

Taking into account the above derived inequality, the definition of $T_0$, and the fact that $T_0 \leq T$, the right hand side of the last equation can be bounded as

$$\frac{2T_0}{T}\delta_{T_0+1} \leq \frac{16\bar{L}\mathsf{V}_1}{\alpha T}\delta_1 + \frac{2\mathtt{A}_3}{\alpha\wedge\alpha^2}\cdot\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}}T^{-\frac{\beta-1}{\beta}} + \frac{2\mathtt{A}_4}{\alpha^2}\cdot\mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}T^{-\frac{\beta+1}{\beta}} \ . \tag{5.32}$$

The combination of Eq. (5.31) and Eq. (5.32) implies

$$\delta_T \leq \mathtt{A}_1\cdot\frac{\mathsf{V}_1}{\alpha T}\delta_1 + \frac{\mathtt{A}_2}{\alpha\wedge\alpha^2}\cdot\left(\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}} + \mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}T^{-\frac{2}{\beta}}\right)T^{-\frac{\beta-1}{\beta}} \ ,$$

where $\mathtt{A}_1 = 16\bar{L}$ and $\mathtt{A}_2 = \left(2+\frac{\beta}{\beta+1}\right)\mathtt{A}_3 + \left(2+\frac{\beta}{3\beta+1}\right)\mathtt{A}_4$.

**The second part: handling the case $T \leq T_0$.** The above analysis was performed under the assumption that $T > T_0$, to conclude, we need to complete the above derived bound for the case when $T \leq T_0$. Eq (5.29) and the fact that in this regime $h_t = (\frac{4\bar{L}\mathsf{V}_3}{b^2T})^{\frac{1}{2\beta}}$ imply

$$\begin{aligned}\delta_{T+1} \leq \delta_1&\left(1-\frac{2}{T_0+1}\right)^T + \frac{2^{3-\frac{2}{\beta}}\bar{L}^{1-\frac{1}{\beta}}}{T_0(\alpha\wedge\alpha^2)}\cdot\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}}\sum_{t=1}^{T}\left(L^2T^{-\frac{\beta-1}{\beta}} + \sigma^2\frac{T^{\frac{1}{\beta}}}{T_0}\right)\\ &+\frac{2^{3+\frac{2}{\beta}}\bar{L}^{1+\frac{1}{\beta}}}{\alpha^2T_0^2}\cdot\mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}\sum_{t=1}^{T}T^{-\frac{1}{\beta}}\\ \leq&\delta_1\left(1-\frac{2}{T_0+1}\right)^T + \frac{\mathtt{A}_3}{\alpha\wedge\alpha^2}\cdot\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}}T^{-\frac{\beta-1}{\beta}} + \frac{\mathtt{A}_4}{\alpha^2}\cdot\mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}T^{-\frac{\beta+1}{\beta}} \ .\end{aligned}$$

Recall that for any $\rho, T > 0$, we have $(1-\rho)^T \leq \exp(-\rho T) \leq \frac{1}{\rho T}$ and hence for $\rho = \frac{2}{T_0+1}$ we can write

$$\delta_{T+1} \leq \mathtt{A}_1\frac{\mathsf{V}_1}{\alpha(T+1)}\delta_1 + \frac{\mathtt{A}_2}{\alpha\wedge\alpha^2}\left(\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2}\right)^{-\frac{1}{\beta}} + \mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}T^{-\frac{2}{\beta}}\right)(T+1)^{-\frac{\beta-1}{\beta}} \ ,$$

where the last inequality follows by the definition of $T_0$ and the fact that $T+1 \leq 2T$, for $T \geq 1$. □

**Lemma 5.7.9.** *Consider the iterative algorithm defined in Eq. (5.1). Assume that $f$ is $\alpha$-strongly convex on $\mathbb{R}^d$ and Assumption 5.4.3 is satisfied. Then, we have for all $\boldsymbol{x} \in \Theta$*

$$\mathbf{E}[f(\boldsymbol{x}_t) - f(\boldsymbol{x})] \leq \frac{r_t - r_{t+1}}{2\eta_t} - r_t\left(\frac{\alpha}{4} - \frac{\eta_t}{2}\bar{L}^2\mathsf{V}_1\right) + \frac{(bLh_t^{\beta-1})^2}{\alpha} + \frac{\eta_t}{2}\left(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right) \ , \tag{5.33}$$

*where $r_t = \mathbf{E}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2$.*

*Proof.* Recall the notation $\mathbf{E}_t[\cdot] = \mathbf{E}[\cdot \mid \mathbf{x}_t]$. For any $\mathbf{x} \in \Theta$, by the definition of projection,

$$\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 = \left\|\mathrm{Proj}_\Theta\big(\mathbf{x}_t - \eta_t \mathbf{g}_t\big) - \mathbf{x}\right\|^2 \leq \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}\|^2 \quad , \tag{5.34}$$

where the above inequality is obtained from the definition of Euclidean projection on the set $\Theta$ and the fact that $\mathbf{x} \in \Theta$. Expanding squares and rearranging the above inequality, we deduce that Eq. (5.34) is equivalent to

$$\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}\rangle \leq \frac{\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2}{2\eta_t} + \frac{\eta_t}{2}\|\mathbf{g}_t\|^2 \quad . \tag{5.35}$$

On the other hand, since $f$ is a $\alpha$-strongly function on $\Theta$, we have

$$f(\mathbf{x}_t) - f(\mathbf{x}) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}\rangle - \frac{\alpha}{2}\|\mathbf{x}_t - \mathbf{x}\|^2 \quad . \tag{5.36}$$

Combining Eq. (5.35) with Eq. (5.36) and denoting $a_t = \|\mathbf{x}_t - \mathbf{x}\|^2$, we deduce that

$$\begin{aligned}\mathbf{E}_t[f(\mathbf{x}_t) - f(\mathbf{x})] &\leq \|\mathbf{E}_t[\mathbf{g}_t] - \nabla f(\mathbf{x}_t)\|\,\|\mathbf{x}_t - \mathbf{x}\| + \frac{1}{2\eta_t}\mathbf{E}_t[a_t - a_{t+1}] + \frac{\eta_t}{2}\mathbf{E}_t\|\mathbf{g}_t\|^2 - \frac{\alpha}{2}\mathbf{E}_t[a_t] \\ &\leq bLh_t^{\beta-1}\|\mathbf{x}_t - \mathbf{x}\| + \frac{1}{2\eta_t}\mathbf{E}_t[a_t - a_{t+1}] \\ &\quad + \frac{\eta_t}{2}\Big(\mathsf{V}_1\mathbf{E}\|\nabla f(\mathbf{x}_t)\|^2 + \mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\Big) - \frac{\alpha}{2}\mathbf{E}_t[a_t] \quad .\end{aligned} \tag{5.37}$$

Since $2ab \leq a^2 + b^2$, we can write

$$bLh_t^{\beta-1}\|\mathbf{x}_t - \mathbf{x}\| \leq \frac{(bLh_t^{\beta-1})^2}{\alpha} + \frac{\alpha}{4}\|\mathbf{x}_t - \mathbf{x}\|^2 \quad . \tag{5.38}$$

Substituting Eq. (5.38) in Eq. (5.37), setting $r_t = \mathbf{E}[a_t]$, and taking total expectation from both sides of Eq. (5.37), yield

$$\mathbf{E}[f(\mathbf{x}_t) - f(\mathbf{x})] \leq \frac{r_t - r_{t+1}}{2\eta_t} - r_t\left(\frac{\alpha}{4} - \frac{\eta_t}{2}\bar{L}^2\mathsf{V}_1\right) + \frac{(bLh_t^{\beta-1})^2}{\alpha} + \frac{\eta_t}{2}\left(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right) \quad ,$$

where the last display is obtained from $\alpha$ strong convexity of $f$. □

*Proof of Theorem 5.4.6.* Since, by definition, $\eta_t \leq \frac{\alpha}{4\bar{L}^2\mathsf{V}_1}$, then we have $\frac{\alpha}{4} - \frac{\eta_t}{2}\bar{L}^2\mathsf{V}_1 \geq \frac{\alpha}{8}$ and (5.33) is simplified as

$$\mathbf{E}[f(\mathbf{x}_t) - f^\star] \leq \frac{r_t - r_{t+1}}{2\eta_t} - \frac{\alpha}{8}r_t + \frac{(bLh_t^{\beta-1})^2}{\alpha} + \frac{\eta_t}{2}(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}) \quad . \tag{5.39}$$

Recall that $T_0 = \lfloor\frac{16\bar{L}^2\mathsf{V}_1}{\alpha^2}\rfloor$—the moment when $\eta_t$ changes its behaviour. Let us first assume that $T_0 < T$ and provide the proof for this case. The case of $T \leq T_0$ will be treated separately, in the end of the present proof. The case distinguishing is slightly different for this proof compared to the previous proofs. By convexity of $f$, we have

$$f(\bar{\mathbf{x}}_T) - f^\star \le \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{x}_t) - f^\star) = \underbrace{\frac{1}{T}\sum_{t=T_0+1}^{T}(f(\mathbf{x}_t) - f^\star)}_{\text{case 1}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T_0}(f(\mathbf{x}_t) - f^\star)}_{\text{case 2}} \ . \tag{5.40}$$

**The first part:** $T > T_0$**.** For any $T_0 + 1 \le t \le T$, we have $\eta_t = \frac{4}{\alpha t}$ and $h_t = \left(\frac{2\mathsf{V}_3}{b^2 t}\right)^{\frac{1}{2\beta}}$. Summing both side of Eq. (5.39) from $T_0 + 1$ to $T$, and substituting $\eta_t$ and $h_t$, we deduce that

$$\sum_{t=T_0+1}^{T}\mathbf{E}[f(\mathbf{x}_t) - f^\star] \le \underbrace{\frac{\alpha}{8}\sum_{t=T_0+1}^{T}\left((r_t - r_{t+1})\,t - r_t\right)}_{=:\text{I}} + \underbrace{\frac{\mathtt{A}_4}{\alpha} b^{\frac{2}{\beta}} \mathsf{V}_3^{\frac{\beta-1}{\beta}} \sum_{t=T_0+1}^{T} t^{-\frac{\beta-1}{\beta}}}_{=:\text{II}}$$
$$+ \underbrace{\frac{\mathtt{A}_5}{\alpha}\mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}\sum_{t=T_0+1}^{T} t^{-\frac{1}{\beta}-1}}_{=:\text{III}} \ ,$$

where we defined I, II, and III, with $\mathtt{A}_4 = 2^{\frac{\beta-1}{\beta}}(L^2 + \sigma^2)$ and $\mathtt{A}_5 = 2^{\frac{\beta+1}{\beta}}\bar{L}^2$.

It is straightforward to see that I $\le \frac{\alpha}{8}T_0 r_{T_0+1}$ (the summation, involved in I, is telescoping). Furthermore, we have II $\le \frac{\mathtt{A}_4}{\alpha} b^{\frac{2}{\beta}} \mathsf{V}_3^{\frac{\beta-1}{\beta}} \sum_{t=1}^{T} t^{-\frac{\beta-1}{\beta}} \le \frac{\beta \mathtt{A}_4}{\alpha} b^{\frac{2}{\beta}} \mathsf{V}_3^{\frac{\beta-1}{\beta}} T^{\frac{1}{\beta}}$. Finally, for the term III we have

$$\text{III} \le \frac{\mathtt{A}_5}{\alpha}\mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}}\sum_{t=T_0+1}^{T} t^{-\frac{1}{\beta}-1} \le \frac{\mathtt{A}_5}{\alpha}\mathsf{V}_2\left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}} T_0^{-\frac{2}{\beta}} \sum_{t=1}^{T} t^{-\frac{\beta-1}{\beta}} \le \frac{\mathtt{A}_6}{\alpha}\mathsf{V}_2\left(\frac{\mathsf{V}_3}{\mathsf{V}_1^2 b^2}\right)^{\frac{1}{\beta}} T^{\frac{1}{\beta}} \ ,$$

where $\mathtt{A}_6 = 2^{-\frac{8}{\beta}}\beta \cdot \mathtt{A}_5$, and the last two inequalities are obtained from the fact that $\alpha \le \bar{L}$. Combining the bounds on I, II, and III we obtain

$$\sum_{t=T_0+1}^{T}\mathbf{E}[f(\mathbf{x}_t) - f^\star] \le \frac{\alpha}{8}T_0 r_{T_0+1} + \left(\beta \mathtt{A}_4 b^{\frac{2}{\beta}}\mathsf{V}_3^{\frac{\beta-1}{\beta}} + \mathtt{A}_6 \mathsf{V}_2\left(\frac{\mathsf{V}_3}{\mathsf{V}_1^2 b^2}\right)^{\frac{1}{\beta}}\right)\frac{T^{\frac{1}{\beta}}}{\alpha} \ . \tag{5.41}$$

For the term $r_{T_0+1}$, Eq. (5.39) implies that for $1 \le t \le T_0$

$$r_{t+1} \le r_t + 2\eta_t \cdot \frac{(bLh_t^{\beta-1})^2}{\alpha} + \eta_t^2\left(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right) \ .$$

In addition, for $1 \le t \le T_0$ we have $\eta_t = \frac{\alpha}{4\bar{L}^2\mathsf{V}_1}$ and $\eta_t \le \frac{4}{\alpha T_0}$. Therefore, unfolding the above recursion we get

$$r_{T_0+1} \le r_1 + \sum_{t=1}^{T_0}\left(\frac{8}{\alpha^2 T_0}\left(bLh_t^{\beta-1}\right)^2 + \frac{16}{\alpha^2T_0^2}\left(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right)\right) \ .$$

Plugging in $h_t = \left(\frac{2\mathsf{V}_3}{b^2 T}\right)^{\frac{1}{2\beta}}$, yields

$$r_{T_0+1} \leq r_1 + 8\left(\mathtt{A}_4 b^{\frac{2}{\beta}} \mathsf{V}_3^{\frac{\beta-1}{\beta}} + \mathtt{A}_5 \mathsf{V}_2 \left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}}\right) \frac{T^{\frac{1}{\beta}}}{\alpha^2 T_0} \ .$$

Accordingly, the right hand side of Eq. (5.41) can be bounded as

$$\frac{\alpha}{8} T_0 r_{T_0+1} \leq \frac{2\bar{L}^2 \mathsf{V}_1}{\alpha} r_1 + \left(\mathtt{A}_4 b^{\frac{2}{\beta}} \mathsf{V}_3^{\frac{\beta-1}{\beta}} + \mathtt{A}_5 \mathsf{V}_2 \left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}}\right) \frac{T^{\frac{1}{\beta}}}{\alpha} \ . \tag{5.42}$$

Substituting the bound in Eq. (5.42) into Eq. (5.41), we deduce that

$$\begin{aligned} \sum_{t=T_0+1}^{T} \mathbf{E}[f(\mathbf{x}_t) - f^\star] \leq & \left((\beta{+}1)\, \mathtt{A}_4 b^{\frac{2}{\beta}} \mathsf{V}_3^{\frac{\beta-1}{\beta}} + (\mathtt{A}_5{+}\mathtt{A}_6)\, \mathsf{V}_2 \left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}} \left(\mathsf{V}_1^{-\frac{2}{\beta}} + T^{-\frac{2}{\beta}}\right)\right) \frac{T^{\frac{1}{\beta}}}{\alpha} \\ & + \frac{2\bar{L}^2 \mathsf{V}_1}{\alpha} r_1 \ . \end{aligned} \tag{5.43}$$

In the second regime of $t \in [1, T_0]$, note that $\frac{4}{\alpha(T_0+1)} \leq \eta_t$. Then, summing (5.39) from $1$ to $T_0$ we have

$$\sum_{t=1}^{T_0} \mathbf{E}[f(\mathbf{x}_t) - f^\star] \leq \frac{\alpha(T_0+1)}{8} r_1 + \left(\mathtt{A}_4 b^{\frac{2}{\beta}} \mathsf{V}_3^{\frac{\beta-1}{\beta}} + \mathtt{A}_5 \mathsf{V}_2 \left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}}\right) \frac{T^{\frac{1}{\beta}}}{\alpha} \ . \tag{5.44}$$

As indicated in Eq. (5.40) at the beginning of the proof, we sum up the two studied cases—Eq. (5.43) and Eq. (5.44)—to deduce that

$$\mathbf{E}[f(\bar{\mathbf{x}}_T) - f^\star] \leq \mathtt{A}_1 \frac{\mathsf{V}_1}{\alpha T} r_1 + \left(\mathtt{A}_2 b^{\frac{2}{\beta}} \mathsf{V}_3^{\frac{\beta-1}{\beta}} + \mathtt{A}_3 \mathsf{V}_2 \left(\frac{\mathsf{V}_3}{b^2}\right)^{\frac{1}{\beta}} \left(\mathsf{V}_1^{-\frac{2}{\beta}} + T^{-\frac{2}{\beta}}\right)\right) \frac{T^{-\frac{\beta-1}{\beta}}}{\alpha} \ ,$$

where $\mathtt{A}_1 = \frac{49}{16}\bar{L}^2$ (for this bound we used $T_0 \geq 16$), $\mathtt{A}_2 = (\beta+2)\mathtt{A}_4$, and $\mathtt{A}_3 = 2\mathtt{A}_5 + \mathtt{A}_6$.

**The second part: handling the case $T \leq T_0$.** At the end, we state the proof for the case $T \leq T_0$. By Eq. (5.39), we can write

$$\sum_{t=1}^{T} \mathbf{E}[f(\mathbf{x}_t) - f^\star] \leq \frac{r_1}{2\eta_1} + \left(\mathtt{A}_4 b^{\frac{2}{\beta}} \mathsf{V}_3^{\frac{\beta-1}{\beta}} T^{\frac{1}{\beta}} + \mathtt{A}_5 \mathsf{V}_2 \left(\frac{\mathsf{V}_3}{\mathsf{b}^2}\right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}}\right) \frac{T^{\frac{1}{\beta}}}{\alpha} \ .$$

We conclude by convexity of $f$. □

### Constrained optimization: the strongly convex-case

*Proof of Lemma 5.4.8.* Recalling that $\sup_{\mathbf{x}\in\Theta}\|\nabla f(\mathbf{x})\| \leq G$, by Lemma 5.7.9 we have for any $t = 1, \ldots, T$

$$0 \leq \mathbf{E}[f(\mathbf{x}_t) - f^\star] \leq \frac{r_t - r_{t+1}}{2\eta_t} - \frac{\alpha}{4}r_t + \frac{b_t^2}{\alpha} + \frac{\eta_t}{2}\big(v_t + mG^2\big) \enspace .$$

Summing up the above inequalities over $t = 1, \ldots, T$, we get

$$\sum_{t=1}^{T}\mathbf{E}[f(\mathbf{x}_t) - f^\star] \leq \frac{1}{2}\sum_{t=1}^{T}\left(\frac{r_t - r_{t+1}}{\eta_t} - \frac{\alpha}{2}r_t\right) + \sum_{t=1}^{T}\left(\frac{\eta_t}{2}(v_t + mG^2) + \frac{b_t^2}{\alpha}\right) \enspace . \tag{5.45}$$

Since we set $\eta_t = \frac{2}{\alpha t}$, then the first term on the r.h.s. of Eq. (5.45) is non-positive. Substitution of $\eta_t = \frac{2}{\alpha t}$ into Eq. (5.45) in conjunction with the non-positivity of the first term in (5.45), implies

$$\sum_{t=1}^{T}\mathbf{E}[f(\mathbf{x}_t) - f^\star] \leq \frac{1}{\alpha}\sum_{t=1}^{T}\left(\frac{1}{t}(v_t + mG^2) + b_t^2\right) \enspace .$$

The proof now follows by the standard bound on the partial sum of the harmonic series and the convexity of $f$. □

*Proof of Theorem 5.4.9 .* The proof is devided in two parts, one for estimator 5.2 and the other one for estimator 5.3. In this result we set $\eta_t = \frac{2}{\alpha t}$ and $h_t = \mathfrak{h}\cdot t^{-1/2\beta}$, where $\mathfrak{h}$ equals to $d^{\frac{1}{\beta}}$ for estimator 5.2 and to $d^{\frac{2+\beta}{2\beta}}$ for estimator 5.3. Analysis of both algorithms starts with Lemma 5.4.8, which states that

$$\mathbf{E}[f(\bar{\mathbf{x}}_T) - f^\star] \leq \frac{mG^2\left(\log(T) + 1\right)}{\alpha T} + \frac{1}{\alpha T}\sum_{t=1}^{T}\left(\frac{v_t}{t} + b_t^2\right) \enspace ,$$

for any algorithm encompassed by Eq. 5.1. We recall that for the application of the above inequality it is assumed that $f$ is $\alpha$-strongly convex; $\Theta$ is a convex and closed subset of $\mathbb{R}^d$, with $\sup_{\mathbf{x}\in\Theta}\|\nabla f(\mathbf{x})\|_2 \leq G$.

**Part I: for estimator 5.2** By the bias and variance bounds in Lemmas 5.3.2 and 5.3.3 respectively, we have

$$\mathbf{E}[f(\bar{\mathbf{x}}_T) - f^\star] \leq 4\kappa G^2\log(eT)\frac{d}{\alpha T} + \frac{1}{\alpha T}\sum_{t=1}^{T}\left(\mathbb{A}_4 h_t^{2(\beta-1)} + \frac{1}{t}d\left(\mathbb{A}_5 h_t^2 + \mathbb{A}_6 d h_t^{-2}\right)\right) \enspace ,$$

where $\mathtt{A}_4 = (\kappa L)^2$, $\mathtt{A}_5 = 4\kappa\bar{L}^2$, and $\mathtt{A}_6 = \frac{\sigma^2\kappa}{2}$. Substituting $h_t = \left(\frac{d^2}{t}\right)^{\frac{1}{2\beta}}$, we deduce that

$$\mathbf{E}[f(\bar{\mathbf{x}}_T) - f^\star] \leq \frac{4d\kappa G^2 \log(eT)}{\alpha T} + \frac{1}{\alpha T}\sum_{t=1}^{T}\left((\mathtt{A}_4 + \mathtt{A}_6)\left(\frac{d^2}{t}\right)^{\frac{\beta-1}{\beta}} + \mathtt{A}_5 d^{1+\frac{2}{\beta}} t^{-\frac{\beta+1}{\beta}}\right) \ . \tag{5.46}$$

It remains to bound the partial sum appearing in the above inequality. It holds that $\sum_{t=1}^{T} t^{-\frac{\beta-1}{\beta}} \leq \beta T^{\frac{1}{\beta}}$ and $\sum_{t=1}^{T} t^{-\frac{1}{\beta}-1} \leq 1+\beta$. Therefore, Eq. (5.46) can be further bounded as

$$\mathbf{E}[f(\bar{\mathbf{x}}_T) - f^\star] \leq \mathtt{A}_1\left(\log(T)+1\right)\frac{d}{\alpha T} + \frac{\mathtt{A}_2}{\alpha T} d^{\frac{2(\beta-1)}{\beta}} T^{\frac{1}{\beta}} + \frac{\mathtt{A}_3}{\alpha T} d^{1+\frac{2}{\beta}} \ ,$$

where $\mathtt{A}_1 = 4\kappa G^2$, $\mathtt{A}_2 = \beta\left(\mathtt{A}_3 + \mathtt{A}_5\right)$, and $\mathtt{A}_3 = (\beta+1)\mathtt{A}_5$.

**Part II: for estimator 5.3** Using Lemma 5.3.4 (bound on the bias) and Lemma 5.3.5 (bound on the variance), we get

$$\mathbf{E}[f(\bar{\mathbf{x}}_T) - f^\star] \leq \mathtt{A}_1\left(\log(T)+1\right)\frac{d}{\alpha T} + \frac{1}{\alpha T}\sum_{t=1}^{T}\left(\mathtt{A}_4 h_t^{2(\beta-1)} d^{1-\beta} + \frac{1}{t}\left(\mathtt{A}_5 h_t^2 + \mathtt{A}_6 d^3 h_t^{-2}\right)\right) \ ,$$

where $\mathtt{A}_1 = \frac{\bar{\mathtt{C}}_{d,1} d\kappa}{d-2}$, $\mathtt{A}_4 = (\kappa_\beta \ell L)^2$, $\mathtt{A}_5 = \frac{\bar{\mathtt{C}}_{d,2} d^2 \kappa \bar{L}^2}{(d-2)(d+1)}$, and $\mathtt{A}_6 = \frac{\sigma^2\kappa}{2}$. Plugging in $h_t = d^{\frac{2+\beta}{2\beta}} t^{-\frac{1}{2\beta}}$, implies

$$\mathbf{E}[f(\bar{\mathbf{x}}_T) - f^\star] \leq \mathtt{A}_1\left(\log(T)+1\right)\frac{d}{\alpha T} + \frac{1}{\alpha T}\sum_{t=1}^{T}\left((\mathtt{A}_4 + \mathtt{A}_6)\left(\frac{d^2}{t}\right)^{\frac{\beta-1}{\beta}} + \mathtt{A}_5 d^{1+\frac{2}{\beta}} t^{-\frac{\beta+1}{\beta}}\right) \ .$$

With a similar argument as in the previous paragraph, we dedeuce that

$$\mathbf{E}[f(\bar{\mathbf{x}}_T) - f^\star] \leq \mathtt{A}_1\left(\log(T)+1\right)\frac{d}{\alpha T} + \frac{\mathtt{A}_2}{\alpha T} d^{\frac{2(\beta-1)}{\beta}} T^{\frac{1}{\beta}} + \frac{\mathtt{A}_3}{\alpha T} d^{1+\frac{2}{\beta}} \ .$$

where we assigned $\mathtt{A}_2 = \beta\left(\mathtt{A}_4 + \mathtt{A}_6\right)$, and $\mathtt{A}_3 = (\beta+1)\,\mathtt{A}_5$. □

### Lower bounds

*Proof of Lemma 5.5.1.* Observe that since the noise $\xi_1, \ldots, \xi_T$ is assumed to be independent, and $\xi_t$ is independent of $(\mathbf{z}_1, y_1, \ldots, \mathbf{z}_{t-1}, y_{t-1}, \boldsymbol{\tau}_t)$ for each $t$, the following decomposition holds

$$\mathrm{d}\mathbf{P}_f = \mathrm{d}F(y_1 - f(\mathbf{z}_1))\prod_{t=2}^{T}\mathrm{d}F\Big(y_t - f\Big(\Phi_t(\mathbf{z}_1, y_1, \ldots, y_{t-1})\Big)\Big)\,\mathrm{d}\mathbb{P}_t(\boldsymbol{\tau}_t) \ ,$$

where $\mathbb{P}_t$ is the probability measure corresponding to the distribution of $\boldsymbol{\tau}_t$. Introduce for compactness $\mathrm{d}F_{f,t} \triangleq \mathrm{d}F\big(y_t - f\big(\Phi_t(\mathbf{z}_1, y_1, \ldots, y_{t-1})\big)\big)\,\mathrm{d}\mathbb{P}_t(\boldsymbol{\tau}_t)$, then in this notation we have

$\mathrm{d}\mathbf{P}_f = \prod_{t=1}^T \mathrm{d}F_{f,t}$. Analogously, the same holds for $\mathbf{P}_{f'}$. Using the definition of the Hellinger distance we can write

$$1-\frac{1}{2}H^2(\mathbf{P}_f,\mathbf{P}_{f'}) = \int \sqrt{\mathrm{d}\mathbf{P}_f\,\mathrm{d}\mathbf{P}_{f'}} = \prod_{t=1}^T \int \sqrt{\mathrm{d}F_{f,t}}\sqrt{\mathrm{d}F_{f',t}} = \prod_{t=1}^T \left(1 - \frac{H^2\left(\mathrm{d}F_{f,t},\, \mathrm{d}F_{f',t}\right)}{2}\right) \enspace .$$

Finally, invoking the assumption on the cumulative distribution of the noise, we get

$$\begin{aligned}\prod_{t=1}^T \left(1 - \frac{H^2\left(\mathrm{d}F_{f,t},\, \mathrm{d}F_{f',t}\right)}{2}\right) &\geq \min_{1\leq t\leq T}\left(1 - \frac{H^2\left(\mathrm{d}F_{f,t},\, \mathrm{d}F_{f',t}\right)}{2}\right)^T\\ &\geq \left(1 - \max_{\boldsymbol{u}\in\Theta}\frac{I_0|f(\boldsymbol{u}) - f'(\boldsymbol{u})|}{2}\right)^T \enspace .\end{aligned}$$

Substituting into the penultimate equality and rearranging we conclude. □

*Proof of Theorem 5.5.2.* The proof closely follows the lower bound established in Akhavan et al. (2020). As mentioned in the main body of the paper, the main improvement of our lower bound is in the use of the Hellinger distance in place of the KL-divergence. Hence, for this proof we only briefly recall the construction of Akhavan et al. (2020). We first assume that $\alpha \geq T^{-1/2+1/\beta}$.

Let $\eta_0 : \mathbb{R} \to \mathbb{R}$ be an infinitely many times differentiable function such that

$$\eta_0(x) = \begin{cases} = 1 & \text{if } |x| \leq 1/4,\\ \in (0,1) & \text{if } 1/4 < |x| < 1,\\ = 0 & \text{if } |x| \geq 1.\end{cases}$$

Set $\eta(x) = \int_{-\infty}^x \eta_0(\tau)d\tau$. Let $\Omega = \big\{-1,1\big\}^d$ be the set of binary sequences of length $d$. Consider the finite set of functions $f_\omega : \mathbb{R}^d \to \mathbb{R}, \boldsymbol{\omega} = (\omega_1,\dots,\omega_d) \in \Omega$, defined as follows:

$$f_{\boldsymbol{\omega}}(\boldsymbol{u}) = \alpha(1+\delta)\left\|\boldsymbol{u}\right\|^2/2 + \sum_{i=1}^d \omega_i r h^\beta \eta(u_i h^{-1}), \qquad \boldsymbol{u} = (u_1,\dots,u_d),$$

where $\omega_i \in \{-1,1\}$, $h = \min\left((\alpha^2/d)^{\frac{1}{2(\beta-1)}}, T^{-\frac{1}{2\beta}}\right)$ and $r > 0, \delta > 0$ are fixed numbers that will be chosen small enough.

Akhavan et al. (2020) showed that $f_{\boldsymbol{\omega}} \in \mathcal{F}'_{\alpha,\beta}$ for $r > 0$ and $\delta > 0$ small enough. In particular, the show that minimizers of functions $f_{\boldsymbol{\omega}}$ belong to $\Theta$ and are of the form

$$\mathbf{x}^*_{\boldsymbol{\omega}} = (x^*(\omega_1),\dots,x^*(\omega_d)) \enspace ,$$

where $x^*(\omega_i) = -\omega_i\alpha^{-1}(1+\delta)^{-1}rh^{\beta-1}$.

For any fixed $\boldsymbol{\omega} \in \Omega$, we denote by $\mathbf{P}_{\boldsymbol{\omega},T}$ the probability measure corresponding to the joint distribution of $((\mathbf{z}_i, y_i)_{i=1}^T, (\boldsymbol{\tau}_i)_{i=1}^T)$ where $y_t = f_{\boldsymbol{\omega}}(\mathbf{z}_t) + \xi_t$ with independent identically distributed $\xi_t$'s such that (2.9) holds, $\xi_t$ is independent of $(\mathbf{z}_1, y_1, \dots, \mathbf{z}_{t-1}, y_{t-1}, \boldsymbol{\tau}_t)$ for each $t$, and $\mathbf{z}_t$'s chosen by a sequential strategy in $\Pi_T$. Consider the statistic

$$\hat{\boldsymbol{\omega}} \in \arg\min_{\boldsymbol{\omega}\in\Omega} \|\mathbf{z}_T - \mathbf{x}^*_{\boldsymbol{\omega}}\| \quad .$$

Classical triangle inequality based arguments yield

$$\max_{\boldsymbol{\omega}\in\Omega} \mathbf{E}_{\boldsymbol{\omega},T}\big[\, \|\mathbf{z}_T - \mathbf{x}^*_{\boldsymbol{\omega}}\|^2 \,\big] \geq \alpha^{-2} r^2 h^{2\beta-2} \inf_{\hat{\boldsymbol{\omega}}} \max_{\boldsymbol{\omega}\in\Omega} \mathbf{E}_{\boldsymbol{\omega},T}\big[\rho(\hat{\boldsymbol{\omega}}, \boldsymbol{\omega})\big] \quad .$$

Note that for all $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$ such that $\rho(\boldsymbol{\omega}, \boldsymbol{\omega}') = 1$ we have

$$\max_{\boldsymbol{u}\in\mathbb{R}^d} |f_{\boldsymbol{\omega}}(\boldsymbol{u}) - f_{\boldsymbol{\omega}'}(\boldsymbol{u})| \leq 2rh^{\beta}\eta(1) \leq 2rT^{-1/2}\eta(1) \quad .$$

Thus, letting $2r < \big(v_0/\eta(1)\big)T^{1/2}$ to ensure that $2rT^{-1/2}\eta(1) \leq v_0$ we apply Lemma 5.5.1 and deduce for such $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$ that

$$H^2(\mathbf{P}_{\boldsymbol{\omega},T}, \mathbf{P}_{\boldsymbol{\omega}',T}) \leq 2\Big(1 - \big(1 - T^{-1}\big)^T\Big) \leq 2\left(1 - \frac{1}{4}\right) = 3/2 \quad .$$

Applying (Tsybakov, 2009, Theorem 2.12) we deduce that

$$\inf_{\hat{\boldsymbol{\omega}}} \max_{\boldsymbol{\omega}\in\Omega} \mathbf{E}_{\boldsymbol{\omega},T}[\rho(\hat{\boldsymbol{\omega}}, \boldsymbol{\omega})] \geq 0.01 \cdot d \quad .$$

Therefore, we have proven that if $\alpha \geq T^{-\frac{\beta+2}{2\beta}}$ then there exist $r > 0$ and $\delta > 0$ such that

$$\max_{\boldsymbol{\omega}\in\Omega} \mathbf{E}_{\boldsymbol{\omega},T}\big[\, \|\mathbf{z}_T - \mathbf{x}^*_{\boldsymbol{\omega}}\|^2 \,\big] \geq 0.001 \cdot d\alpha^{-2} r^2 h^{2\beta-2} = 0.01 \times r^2 \min\left(1,\, \frac{d}{\alpha^2} T^{-\frac{\beta-1}{\beta}}\right) \quad . \tag{5.47}$$

This implies (5.12) for $\alpha \geq T^{-\frac{\beta+2}{2\beta}}$, by the inclusion of the functional classes, we have the bound of this order for all $\alpha$.

We now prove (5.11). From (5.47) and $\alpha$-strong convexity of $f$ we get that, for $\alpha \geq T^{-\frac{\beta+2}{2\beta}}$,

$$\max_{\boldsymbol{\omega}\in\Omega} \mathbf{E}_{\boldsymbol{\omega},T}\big[f(\mathbf{z}_T) - f(x^*_{\boldsymbol{\omega}})\big] \geq 0.005 \cdot r^2 \min\left(\alpha,\, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\right) \quad .$$

This implies (5.11) in the zone $\alpha \geq T^{-\frac{\beta+2}{2\beta}}$ since for such $\alpha$ we have

$$\min\left(\alpha,\, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\right) = \min\left(\max(\alpha, T^{-\frac{\beta+2}{2\beta}}),\, \frac{d}{\sqrt{T}},\, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\right) \quad .$$

On the other hand, $\min\big(\alpha_0,\, \frac{d}{\alpha_0} T^{-\frac{\beta-1}{\beta}}\big) = \min\big(T^{-\frac{\beta+2}{2\beta}}, d/\sqrt{T}\big)$, and the same lower bound holds for $0 < \alpha < \alpha_0$ by the nestedness argument that we used to prove (5.12) in the zone $0 < \alpha < \alpha_0$. Thus, (5.11) follows. □

*Proof of Theorem 5.5.3.* Clearly, $\mathcal{F}'_{1,\beta} \subset \mathcal{F}_{\beta}$, where $\mathcal{F}'_{1,\beta}$ is the class of functions considered in Theorem 5.5.2. As the functions in $\mathcal{F}'_{1,\beta}$ are 1-strongly convex, for $t \geq 1$, $\|\nabla f(\mathbf{z}_t)\|^2 \geq 2(f(\mathbf{z}_t) - f^\star)$. Using Theorem 5.5.2, for $T \geq 1$, we have

$$\sup_{f \in \mathcal{F}'_{1,\beta}} \mathbf{E}[f(\mathbf{z}_t) - f^\star] \geq C' d T^{-\frac{\beta-1}{\beta}}, \tag{5.48}$$

where $C' > 0$ does not depend on $d, T$, and $\beta$. We have for any random variable $S$ that we consider $\mathbf{E}[\|\nabla f(\mathbf{z}_S)\|^2] = \sum_{t=1}^{T} p_t \mathbf{E}[\|\nabla f(\mathbf{z}_t)\|^2]$, where $(p_1, \ldots, p_T)$ is a probability vector: $p_t \geq 0$, and $\sum_{t=1}^{T} p_t = 1$. Thus,

$$\mathbf{E}[\|\nabla f(\mathbf{z}_S)\|^2] \geq 2 \sum_{t=1}^{T} p_t \mathbf{E}[f(\mathbf{z}_t) - f^\star] \geq 2\mathbf{E}\left[ f\left( \sum_{t=1}^{T} p_t \mathbf{z}_t \right) - f^\star \right] . \tag{5.49}$$

Note that $\tilde{\mathbf{z}}_T = \sum_{t=1}^{T} p_t \mathbf{z}_t$ is an estimator depending only on the past, that is in the same class of estimators as $\mathbf{z}_T$. Therefore, the bound (5.48) holds for $\tilde{\mathbf{z}}_T$ as well. Combining (5.48) and (5.49) gives (5.5.2). □

# Chapter 6

# Zero-order optimization of highly smooth functions in a passive scheme

We propose a new method for estimating the minimizer $\mathbf{x}^*$ and the minimum value $f^*$ of a smooth and strongly convex regression function $f$ from the observations contaminated by random noise. Our estimator $\mathbf{z}_n$ of the minimizer $\mathbf{x}^*$ is based on a version of the projected gradient descent with the gradient estimated by a regularized local polynomial algorithm. Next, we propose a two-stage procedure for estimation of the minimum value $f^*$ of regression function $f$. At the first stage, we construct an accurate enough estimator of $\mathbf{x}^*$, which can be, for example, $\mathbf{z}_n$. At the second stage, we estimate the function value by at the point obtained at the first stage using a rate optimal nonparametric procedure. We derive non-asymptotic upper bounds for the quadratic risk and optimization error of $\mathbf{z}_n$, and for the risk of estimating $f^*$. We establish minimax lower bounds showing that, under certain choice of parameters, the proposed algorithms achieve the minimax optimal rates of convergence on the class of smooth and strongly convex functions.

## 6.1 Introduction

Estimating the minimum value and the minimizer of an unknown function from observation of its noisy values on a finite set of points is a key problem in many applications. Let $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$ be a design set and let $\Theta$ be a compact and convex subset of $\mathbb{R}^d$. Assume that we observe noisy values of an unknown regression function $f : \mathbb{R}^d \to \mathbb{R}$ at points of the design set:

$$y_i = f(\mathbf{x}_i) + \xi_i, \quad i = 1, \dots, n, \tag{6.1}$$

where $\xi_i$'s are independent zero mean errors with $\mathbf{E}[\xi_i^2] \leq \sigma^2$. Our goal is to estimate the minimum value of the regression function $f^* = \min_{x \in \Theta} f(x)$ and its location $\mathbf{x}^* = \arg\min_{x \in \Theta} f(x)$ when $\mathbf{x}^*$ is unique. As accuracy measures of an estimator $\hat{\mathbf{x}}_n$ of $\mathbf{x}^*$ we consider the expected optimization error $\mathbf{E}(f(\hat{\mathbf{x}}_n) - f^*)$ and the quadratic risk $\mathbf{E}\,\|\hat{\mathbf{x}}_n - \mathbf{x}^*\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. The accuracy of an estimator $T_n$ of $f^*$ will be measured by the risk $\mathbf{E}|T_n - f^*|$. We will assume that $f$ belongs to the class of $\beta$-Hölder smooth and strongly convex functions with $\beta \geq 2$ (see Section 6.2 for the definitions).

The existing literature considers two different assumptions on the choice of the design. Under the *passive* design setting, the points $\mathbf{x}_i$ are sampled independently from some probability distribution. Under the *active* (or sequential) design setting, for each $i$ the statistician can plan the experiment by selecting the point $\mathbf{x}_i$ depending on the previous queries and the corresponding responses $\mathbf{x}_1, y_1, \dots, \mathbf{x}_{i-1}, y_{i-1}$. The accuracy of estimation under the active design is at least as good as under the passive design but it can be strictly better, which is the case for the problems considered here.

**Active design, estimation of $\mathbf{x}^*$.** Active (or sequential) scheme has a long history starting at least from the seminal work of Kiefer and Wolfowitz (1952) where an analog of the Robins-Monro algorithm was introduced to estimate the minimizer $\mathbf{x}^*$ of a univariate function $f$. The idea of the Kiefer-Wolfowitz (KW) method is to approximate the derivative of $f$ using first order differences of $y_i$'s and plug this estimator in the gradient algorithm. Kiefer and Wolfowitz (1952) proved convergence in probability of the KW algorithm under some regularity conditions on the regression function. A multivariate extension of the KW algorithm was proposed by Blum (1954). Convergence rates of the KW algorithm for $d = 1$ were investigated in Dupač (1957) proving an upper bound on the quadratic risk of the order $n^{-2/3}$ for $\beta = 3$. By using suitably chosen linear combinations of first order differences to approximate the gradient, Fabian (1967b) proved the existence of a method that attains, for odd integers $\beta \geq 3$, the quadratic risk of the order $n^{-(\beta-1)/\beta}$ for functions $f$ with bounded $\beta$th partial derivatives. The method of Fabian (1967b) uses $(\beta-1)/2$ evaluations $y_i$ at every step of the algorithm in order to approximate the gradient. Chen (1988) and Polyak and Tsybakov (1990) have established minimax lower bounds for the estimation risk on the class of $\beta-$Hölder smooth and strongly convex functions $f$, for all $\beta \geq 2$. For the quadratic risk, these bounds are of the order $n^{-(\beta-1)/\beta}$. Polyak and Tsybakov (1990) proposed a new class of methods using smoothing kernels and randomization to approximate the gradient. This constitutes an alternative to the earlier used deterministic schemes derived from finite differences. Polyak and Tsybakov (1990) proved that such randomized methods attain the minimax optimal rate $n^{-(\beta-1)/\beta}$ on the above classes for all $\beta \geq 2$ and not only for odd integers $\beta \geq 3$. An additional advantage over Fabian's algorithm is the computational simplicity of these methods. In particular, they require at each step only one or two evaluations of the function. For subsequent developments on similar methods, we refer to Akhavan et al. (2020, 2021); Bach and Perchet (2016); Dippon (2003b), where one can find further references.

**Active design, estimation of $f^*$.** The problem of estimating $f^*$ under the active scheme was first considered by Mokkadem and Pelletier (2007) who suggested a recursive estimator and proved its asymptotic normality with $\sqrt{n}$ scaling. Belitser et al. (2012) defined an estimator of $f^*$ via a multi-stage procedure whose complexity increases exponentially with the dimension $d$, and showed that this estimator achieves (asymptotically, for $n$ greater than an exponent of $d$) the $O_p(1/\sqrt{n})$ rate when $f$ is $\beta$-Hölder and strongly convex with $\beta > 2$. Akhavan et al. (2020) improved upon this result by constructing a simple computationally feasible estimator $\hat{f}_n$ such that $\mathbf{E}|\hat{f}_n - f^*| = O(1/\sqrt{n})$ for $\beta \geq 2$. It can be easily shown that the rate $1/\sqrt{n}$ cannot be further improved when estimating $f^*$. Indeed, using the oracle that puts all the queries at the unknown true minimizer $\mathbf{x}^*$ one cannot achieve better rate under the Gaussian noise.

**Passive design, estimation of $\mathbf{x}^*$.** The problem of estimating the minimizer $\mathbf{x}^*$ under the i.i.d. passive design was probably first studied in Härdle and Nixdorf (1987), where some consistency and asymptotic normality results were discussed. Tsybakov (1990a) proposed to estimate $\mathbf{x}^*$ by a recursive procedure using local polynomial approximations of the gradient. Considering the class of strongly convex and $\beta$-Hölder ($\beta \geq 2$) regression functions $f$, Tsybakov (1990a) proves that the minimax optimal rate of estimating $\mathbf{x}^*$ on the above class of functions is $n^{-(\beta-1)/(2\beta+d)}$, and shows that the proposed estimator attains this optimal rate. However, in order to define this estimator, one needs to know of the marginal density of the design points that may be inaccessible in practice.

There was also some work on estimating $\mathbf{x}^*$ in different passive design settings. Several papers are analyzing estimation of $\mathbf{x}^*$ in a passive scheme, where $\mathbf{x}_i$'s are given non-random points in $[0, 1]$ (Müller (1985, 1989)) or in $[0, 1]^d$ (Facer and Müller (2003)). Another line of work (Härdle and Nixdorf (1987); Nazin et al. (1989, 1992); Tsybakov (1990a)) is to consider the problem of estimating the zero of a nonparametric regression function under i.i.d. design, also called passive stochastic approximation when recursive algorithms are used. Nazin et al. (1989, 1992); Tsybakov (1990a) establish minimax optimal rates for this problem and propose passive stochastic approximation algorithms attaining these rates. Application to transfer learning is recently developed in Krishnamurthy and Yin (2022), where one can find further references on passive stochastic approximation.

**Passive design, estimation of $f^*$.** To the best of our knowledge, the problem of estimating $f^*$ under i.i.d. passive design was not studied. However, there was some work on a related and technically slightly easier problem of estimating the maximum of a function observed under the Gaussian white noise model in dimension $d = 1$ (Ibragimov and Khas'minskii (1982); Lepski (1993)). Extrapolating these results to the regression model and general $d$ suggests that the optimal rate of convergence for estimating $f^*$ on the class of $\beta$-Hölder regression functions $f$ is of the order $(n/\log n)^{-\beta/(2\beta+d)}$. It is stated as a conjecture in Belitser et al. (2021) for the passive model with equidistant deterministic design. It remains unclear whether this conjecture is true since, for higher dimensions, the effect of the equidistant grid induces an additional bias. However, we prove below that, under the i.i.d. random design, the minimax optimal rate on the class of $\beta$-Hölder functions (without strong convexity) is indeed $(n/\log n)^{-\beta/(2\beta+d)}$. We are not aware of any results on estimation of $f^*$ on the class of $\beta$-Hölder and strongly convex regression functions $f$, which is the main object of study in the current work.

Finally, we review some results on a related problem of estimating the mode of a probability density function. There exists an extensive literature on this problem. In the univariate case, Parzen (1962) proposed the maximizer of kernel density estimator (KDE) as an estimator for the mode. Direct estimate of the mode based on order statistics was proposed by Grenander (1965), where the consistency of the proposed method was shown. Other estimators of the mode in the univariate case were considered by (Chernoff, 1964; Dalenius, 1965; Venter, 1967). The minimax rate of mode estimation on the class of $\beta$-Hölder densities that are strongly concave near the maximum was shown to be $n^{-(\beta-1)/(2\beta+d)}$ in Tsybakov (1990b), where the optimal recursive algorithm was introduced. It generalizes an earlier result of Khas'minskii (1979) who considered the special case $d = 1, \beta = 2$ and derived the minimax lower bound of the order $n^{-1/5}$ matching the upper rate provided by Parzen (1962). Klemelä (2005) proposed to use the maximizer of KDE with the smoothing parameter chosen by the Lepski method (Lepskii, 1991), and showed that this estimator achieves optimal adaptive rate of convergence. Dasgupta and Kpotufe (2014) proposed minimax optimal estimators of the mode based on $k$-nearest neighbor density estimators, emphasizing the implementation ease of the method. Computational complexity of mode estimation was investigated by Arias-Castro et al. (2022) showing the impossibility of a minimax optimal algorithm with sublinear computational complexity. It was shown that the maximum of a histogram, with a proper choice of bandwidth, achieves the minimax rate while running in linear time. Bayesian approach to the mode estimation was developed by Yoo and Ghosal (2019).

**Contributions.** In this paper, we consider the model described at the beginning of this section under the i.i.d. passive observation scheme. The contributions of the present work can be summarized as follows.

- Assuming that $f$ belongs to the class of $\beta$-Hölder and strongly convex regression functions we construct a recursive estimator of the minimizer $\mathbf{x}^*$ adaptive to the unknown marginal density of $\mathbf{x}_i$'s and achieving the minimax optimal rate $n^{-(\beta-1)/(2\beta+d)}$, for $\beta \geq 2$, up to a logarithmic factor.

- We show that the minimax optimal rate for the problem of estimating the minimum value $f^*$ of function $f$ on the above class of functions scales as $n^{-\beta/(2\beta+d)}$, and we propose an algorithm achieving this optimal rate for $\beta > 2$.
- We prove that the minimax optimal rate of estimating $f^*$ on the class of $\beta$-Hölder functions (without strong convexity) is of the order $(n/\log n)^{-\beta/(2\beta+d)}$. Thus, dropping the assumption of strong convexity causes a deterioration of the minimax rate only by a logarithmic factor. It suggests that strong convexity is not a crucial advantage in estimation of the minimum value of a function under the passive design.

Given our results, we have the following table summarizing the minimax optimal rates for estimation under the active and passive design. We note that the convergence rates for the passive scheme suffer from the curse of dimensionality, while the rates for the active scheme are independent of the dimension.

| | rate of quadratic risk, estimation of $\mathbf{x}^*$ | rate of estimating $f^*$ |
|---|---|---|
| `passive scheme` | $n^{-\frac{2(\beta-1)}{2\beta+d}}$ | $n^{-\frac{\beta}{2\beta+d}}$ |
| `active scheme` | $n^{-\frac{\beta-1}{\beta}}$ | $n^{-\frac{1}{2}}$ |

Table 6.1: Comparisons between the rates of convergence for passive and active schemes

**Notation.** In all the theorems, where the rates contain $\log(n)$, we assume that $n \geq 2$. We denote by $\mathbf{E}_f$ the expectation with respect to the distribution of $(\mathbf{x}_i, y_i)_{i=1}^n$ satisfying the model (6.1); we also abbreviate this notation to $\mathbf{E}$ when there is no ambiguity. Vectors are represented by bold symbols while uppercase English letters are used to denote matrices. We denote by $\|\cdot\|$ the Euclidean norm, and by $\|\cdot\|_{\mathsf{op}}$ the operator norm, i.e., for a matrix $A$ we have $\|A\|_{\mathsf{op}} = \sup_{\|\boldsymbol{u}\|\leq 1} \|A\boldsymbol{u}\|$. We denote the smallest eigenvalue of a square matrix $U$ by $\lambda_{\min}(U)$. For any $m \in \mathbb{N}$, we denote by $[n]$ the set that contains all positive integers $k$, such that $1 \leq k \leq m$. For $\beta \in \mathbb{R}_+$, let $\lfloor \beta \rfloor$ be the biggest integer smaller than $\beta$. Let $S$ denote the number elements in the set $\{\boldsymbol{m} : |\boldsymbol{m}| \leq \ell\}$, where $\boldsymbol{m}$ is a d-dimensional multi-index. For $\boldsymbol{u} \in \mathbb{R}^d$, let $U(\boldsymbol{u}) = \left(\frac{\boldsymbol{u}^{\boldsymbol{m}^{(1)}}}{\boldsymbol{m}^{(1)}}, \dots, \frac{\boldsymbol{u}^{\boldsymbol{m}^{(S)}}}{\boldsymbol{m}^{(S)}}\right)^\top$, where the numbering is such that $\boldsymbol{m}^{(1)} = (0, \dots, 0), \boldsymbol{m}^{(2)} = (1, 0, \dots, 0), \dots, \boldsymbol{m}^{(d+1)} = (0, \dots, 0, 1)$. For $d$-dimensional multi-index $\boldsymbol{m} = (m_1, \dots, m_d)$, where $m_j \geq 0$ are integers, we define the absolute value $|\boldsymbol{m}| = m_1 + \ldots + m_d$, the factorial $\boldsymbol{m}! = m_1! \ldots m_d!$, the power $\boldsymbol{u}^{\boldsymbol{m}} = u_1^{m_1} \ldots u_d^{m_d}$ and the differentiation operator $D^{\boldsymbol{m}} = \frac{\partial^{|\boldsymbol{m}|}}{\partial u_1^{m_1} \ldots \partial u_d^{m_d}}$.

## 6.2 Definitions and assumptions

We first introduce the class of $\beta$-Hölder functions that will be used throughout the paper. For $\beta, L > 0$, by $\mathcal{F}_\beta(L)$ we denote the class of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$

satisfying the the following inequality

$$\left| f(\mathbf{x}) - \sum_{|\boldsymbol{m}| \leq l} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\mathbf{x})(\mathbf{x} - \mathbf{x}')^{\boldsymbol{m}} \right| \leq L \|\mathbf{x} - \mathbf{x}'\|^{\beta}, \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d.$$

Our estimators will be based on kernels satisfying the following assumption.

**Assumption 6.2.1.** *The kernel $K : \mathbb{R}^d \to \mathbb{R}$ has a compact support $\mathrm{Supp}(K)$ contained in the unit Euclidean ball, and satisfies the following conditions*

$$K(\boldsymbol{u}) \geq 0, \qquad \int K(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} = 1, \qquad \sup_{\boldsymbol{u} \in \mathbb{R}^d} K(\boldsymbol{u}) < \infty \ .$$

*Furthermore, for special requirements of our analysis, we assume that $K$ is a $L_K$-Lipschitz function, i.e. for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have*

$$|K(\boldsymbol{x}) - K(\boldsymbol{y})| \leq L_K \, \|\boldsymbol{x} - \boldsymbol{y}\| \ .$$

**Assumption 6.2.2.** *It holds for all $i, i' \in [n]$, that: (i) $\xi_i$ and $\boldsymbol{x}_{i'}$ are independent; (ii) $\mathbf{E}[\xi_i] = 0$; (iii) $\xi_i$ is sub-Gaussian random variable, i.e., there exists $\sigma > 0$ such that for any $t \geq 0$ it satisfies $\mathbf{P}\left[|\xi_i| \geq t\right] \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right)$.*

**Assumption 6.2.3.** *We consider the model (6.1) with $f : \mathbb{R}^d \to \mathbb{R}$ satisfying the following assumptions*

*(i) The function $f$ attains its minimum at $\boldsymbol{x}^* \in \Theta$.*

*(ii) The function $f$ belongs to Hölder functional class $\mathcal{F}_\beta(L)$ with $\beta \geq 2$.*

*(iii) There exists $\alpha > 0$ such that the function $f$ is $\alpha$-strongly convex on $\Theta$ i.e. for any $\boldsymbol{x}, \boldsymbol{y} \in \Theta$, it satisfies*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \ .$$

*(iv) The function $f$ is uniformly bounded on the set $\Theta' = \{\boldsymbol{x} + \boldsymbol{y} : \boldsymbol{x} \in \Theta \quad \text{and} \quad \|\boldsymbol{y}\| \leq 1\}$ such that $\sup_{\boldsymbol{x} \in \Theta'} f(\boldsymbol{x}) \leq M$.*

By $\mathcal{F}_{\beta,\alpha}(L)$ we denote the class of regression functions $f$ satisfying Assumption 6.2.3. Next, we introduce an assumption on the distribution of $\mathbf{x}_i$'s.

**Assumption 6.2.4.** *The random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are i.i.d. with distribution admitting a density $p(\cdot)$ with respect to the Lebesgue measure such that*

$$0 < p_{\min} \leq p(\boldsymbol{x}) \leq p_{\max} < \infty, \quad \forall \boldsymbol{x} \in \Theta'.$$

Throughout this paper, we call $\mathtt{A} > 0$ a numerical constant, if $\mathtt{A}$ can only depend on $d$, $\Theta$, $\beta$, $L$, $M$, $p_{\max}$, $p_{\min}$, $K$, and $\sigma$, where the dependence on $d$ is at most of a polynomial order with the degree of polynomial only depending on $\beta$. We note that dependence on the strong convexity parameter $\alpha$ is not included in the numerical constant since we specify it explicitly in our upper bounds.

## 6.3 Estimating the minimizer

We estimate the minimizer $\mathbf{x}^*$ via an approximation of the gradient algorithm, where we replace the gradient by its local polynomial estimator. The objective function $f \in \mathcal{F}_\beta(L)$ in model (6.1) can be well approximated by its Taylor polynomial of order $\ell$ in the neighbourhood of the target point $\mathbf{z}$,

$$f(\mathbf{x}) \approx \sum_{|\boldsymbol{m}|\leq \ell} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\mathbf{z})(\mathbf{z}-\mathbf{x})^{\boldsymbol{m}} = \boldsymbol{\theta}^\top(\mathbf{z})\boldsymbol{U}\left(\frac{\mathbf{x}-\mathbf{z}}{h}\right) \ ,$$

where $\mathbf{x}$ is sufficiently close to $\mathbf{z}$ and, for $h > 0$,

$$\boldsymbol{U}(\boldsymbol{u}) = \left(\frac{\boldsymbol{u}^{\boldsymbol{m}^{(1)}}}{\boldsymbol{m}^{(1)}!}, \dots, \frac{\boldsymbol{u}^{\boldsymbol{m}^{(S)}}}{\boldsymbol{m}^{(S)}!}\right)^\top, \quad \boldsymbol{\theta}(\mathbf{z}) \quad = \left(h^{|\boldsymbol{m}^{(1)}|} D^{\boldsymbol{m}^{(1)}} f(\mathbf{z}), \dots, h^{|\boldsymbol{m}^{(S)}|} D^{\boldsymbol{m}^{(S)}} f(\mathbf{z})\right)^\top \ .$$

After approximating the objective function by its Taylor expansion, we define the *local polynomial estimator* of $\boldsymbol{\theta}(\mathbf{z})$ (see e.g. (Tsybakov, 2009, Section 1.6)) as follows

$$\hat{\boldsymbol{\theta}}_k(\mathbf{z}) = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbf{R}^S} \sum_{i=1}^k \left[y_i - \boldsymbol{\theta}^\top \boldsymbol{U}\left(\frac{\mathbf{x}_i-\mathbf{z}}{h}\right)\right]^2 K\left(\frac{\mathbf{x}_i-\mathbf{z}}{h}\right) \ ,$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel satisfying Assumption 6.2.1. Let the matrix $B_k(\mathbf{z})$ and the vector $\boldsymbol{D}_k(\mathbf{z})$ be defined as

$$B_k(\mathbf{z}) = \frac{1}{kh^d}\sum_{i=1}^k \boldsymbol{U}\left(\frac{\mathbf{x}_i-\mathbf{z}}{h}\right)\boldsymbol{U}^\top\left(\frac{\mathbf{x}_i-\mathbf{z}}{h}\right) K\left(\frac{\mathbf{x}_i-\mathbf{z}}{h}\right) \ ,$$

$$\boldsymbol{D}_k(\mathbf{z}) = \frac{1}{kh^d}\sum_{i=1}^k y_i \boldsymbol{U}\left(\frac{\mathbf{x}_i-\mathbf{z}}{h}\right) K\left(\frac{\mathbf{x}_i-\mathbf{z}}{h}\right) \ .$$

Since $\hat{\boldsymbol{\theta}}_k(\mathbf{z})$ is a weighted least squares estimator we can write it in the form

$$\hat{\boldsymbol{\theta}}_k(\mathbf{z}) = B_k(\mathbf{z})^{-1}\boldsymbol{D}_k(\mathbf{z}) \ ,$$

provided the matrix $B_k(\mathbf{z})$ is invertible. The above remarks suggest to define an estimator for $\nabla f(\mathbf{z})$ as

$$\mathbf{g}_k(\mathbf{z}) = \frac{1}{h} A\hat{\boldsymbol{\theta}}_k(\mathbf{z}) \ , \tag{6.2}$$

**Algorithm 5** Passive Zero-Order Stochastic Projected Gradient

```
**Requires** Kernel $K : \mathbb{R}^d \to \mathbb{R}$, step-sizes $\eta_k > 0$, parameters $h_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{1}{2\beta+d}}$ and $\lambda_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{\beta}{2\beta+d}}$, for $k \in [n]$.
**Initialization** Choose $\mathbf{z}_1 \in \Theta$, and assign $\eta_k = \frac{1}{\alpha k}$, for $k \in [n]$.
**For** $k \in [n]$
    1. Let $\mathbf{g}_{k,\lambda}(\mathbf{z}_k) = h_k^{-1}\left(AB_{k,\lambda}^{-1}(\mathbf{z}_k)\boldsymbol{D}_k(\mathbf{z}_k)\right)$ ,
    2. Update $\mathbf{z}_{k+1} = \mathrm{Proj}_{\Theta}\left(\mathbf{z}_k - \eta_k \mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right)$ .
**Return** $(\mathbf{z}_k)_{k=1}^n$
```

where $A$ is the matrix with elements

$$A_{i,j} = \begin{cases} 1, & \text{if} \quad j = i+1 \\ 0, & \text{otherwise} \end{cases} ,$$

for $i \in [d]$, and $j \in [S]$.

Since $B_k(\mathbf{z})$ is not necessarily invertible, instead of using the estimator (6.2) we consider its regularized version. Namely, we add a regularization constant $\lambda > 0$ to the diagonal entries of $B_k(\mathbf{z})$ and define $B_{k,\lambda}(\mathbf{z}) = B_k(\mathbf{z}) + \lambda\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This leads to the following regularized estimator of the gradient

$$\mathbf{g}_{k,\lambda}(\mathbf{z}) = \frac{1}{h}A\hat{\boldsymbol{\theta}}_{k,\lambda}(\mathbf{z}) := \frac{1}{h}A(B_k(\mathbf{z}) + \lambda\mathbf{I})^{-1}\boldsymbol{D}_k(\mathbf{z}) \ . \tag{6.3}$$

The corresponding approximate gradient descent procedure is presented as Algorithm 5. It outputs $\mathbf{z}_k$ that will be used as an estimator of $\mathbf{x}^*$. At round $k$ of Algorithm 5, the matrix $B_{k,\lambda}(\mathbf{z}_k) = B_k(\mathbf{z}_k) + \lambda\mathbf{I}$ and the vector $\boldsymbol{D}_k(\mathbf{z}_k)$ can be computed recursively based on the first $k$ observations.

For any $f \in \mathcal{F}_\beta(L)$, one can show that, under a suitable choice of parameters $h$ and $\lambda$, the estimation error for (6.3) $\mathbf{E}\left\|\mathbf{g}_{n,\lambda}(\mathbf{x}) - \nabla f(\mathbf{x})\right\|$, is of the order $n^{-(\beta-1)/(2\beta+d)}$. The rate $n^{-(\beta-1)/(2\beta+d)}$ is known to be minimax optimal for estimating the gradient on the class $\mathcal{F}_\beta(L)$, cf. Stone (1982). This is true without the strong convexity assumption. However, the bounds provided by Stone (1982) are asymptotic in $n$.

In the following theorem, we study the performance of Algorithm 5 for estimating $\mathbf{x}^*$.

**Theorem 6.3.1.** *Let Assumptions 6.2.2–6.2.4 hold. Then, for $\boldsymbol{z}_n$ generated by Algorithm 5, we have*

$$\mathbf{E}\left\|\boldsymbol{z}_n - \boldsymbol{x}^*\right\|^2 \leq \mathtt{A}\min\left(1, \left(\frac{\log(n)}{n}\right)^{\frac{2(\beta-1)}{2\beta+d}}\alpha^{-2}\right) \quad , \tag{6.4}$$

*where $\mathtt{A} > 0$ is a numerical constant.*

*Proof outline.* We use the definition of Algorithm 5 and strong convexity of $f$ to obtain an upper bound for $\mathbf{E}[\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 | \mathbf{z}_k]$, which depends on the bias term $\left\|\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{z}_k)|\mathbf{z}_k] - \nabla f(\mathbf{z}_k)\right\|$ and on the stochastic error term $\mathbf{E}\left[\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right\|^2 |\mathbf{z}_k\right]$. We control the bias and the stochastic error terms uniformly over $\mathbf{z}_k \in \Theta$, which is granted by Lemmas 6.7.3 and 6.7.4. This uniformity is the reason why the bound (6.4) includes an extra logarithmic factor compared to the optimal rate $n^{-(\beta-1)/(2\beta+d)}$ proved in Tsybakov (1990a). □

We consider the logarithmic factor appearing in (6.4) as a price to pay for the fact that our algorithm is adaptive to the marginal density of $\mathbf{x}_i$'s. Indeed, Tsybakov (1990a) considered estimators that can depend on the marginal density of $\mathbf{x}_i$'s and achieve the optimal rate $n^{-(\beta-1)/(2\beta+d)}$, while Algorithm 5 is free of such dependence. Note also that our algorithm can be realized in online mode with the data that arrive progressively. We conjecture that the extra logarithmic factor can be eliminated if we estimate $\mathbf{x}^*$ by the minimizer of the local polynomial estimator of $f$. However, such a method needs the whole sample and cannot be realized in online mode. It remains an open question whether there exists an algorithm combining all the three advantages, that is, online realization, adaptivity to the marginal density and convergence with the sharp optimal rate $n^{-(\beta-1)/(2\beta+d)}$.

In the following theorem, we provide a bound on the optimization error $\mathbf{E}\left[f(\bar{\mathbf{z}}_n) - f^*\right]$, where $\bar{\mathbf{z}}_n$ is the average of the outputs of Algorithm 5 throughout $n$ iterations.

**Theorem 6.3.2.** *Let Assumptions 6.2.2–6.2.4 hold. Then, for $\mathbf{z}_n$ is generated by Algorithm 5, we have*

$$\mathbf{E}\left[f(\bar{\mathbf{z}}_n) - f^*\right] \leq \mathtt{A}\min\left(1, \left(\frac{\log(n)}{n}\right)^{\frac{2(\beta-1)}{2\beta+d}} \alpha^{-1}\right) ,$$

*where $\bar{\mathbf{z}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{z}_k$, and $\mathtt{A} > 0$ is a numerical constant.*

Note that inequality (6.8) below and the strong convexity of $f$ imply a minimax lower bound for the optimization error with the rate $n^{-2(\beta-1)/(2\beta+d)}$. Hence, Theorem 6.3.2 shows that $\bar{\mathbf{z}}_n$ achieves the minimax optimal rate with respect to the optimization error to within a logarithmic factor. It is interesting to compare this result with the optimal rates for the optimization error in the case of active design. As already discussed in the introduction, under the active design for the same class of functions $f$ as in Theorem 6.3.2, the dimension disappears from the optimal rate – it upgrades to $n^{-(\beta-1)/\beta}$, cf. Akhavan et al. (2020); Polyak and Tsybakov (1990). On the other hand, under active design and the class of $\beta$-Hölder functions (without strong convexity) the optimal rate for the optimization error deteriorates substantially and becomes $(n/\log n)^{-\beta/(2\beta+d)}$, cf. Wang et al. (2018a). For all $\beta > 2$, this is worse as the rate under passive design and strong convexity obtained in Theorem 6.3.2.

**Algorithm 6** Estimating the Minimum Value

```
Requires  Algorithm 5, kernel $K : [-1,1]^d \to \mathbb{R}$, parameters $h_n = n^{-\frac{1}{2\beta+d}}$ and $\lambda_n = n^{-\frac{\beta}{2\beta+d}}$.

1. Randomly split the data $D$ in two equal parts $D_1$ and $D_2$
2. From the subsample $D_1$, using the updates of Algorithm 5, construct $\bar{\mathbf{z}}_m$.
3. Based on the second subsample $D_2$, construct the estimator $f_n(\bar{\mathbf{z}}_m) = \boldsymbol{U}^\top(0)\hat{\boldsymbol{\theta}}_{m:n,\lambda}(\bar{\mathbf{z}}_m)$
Return  $f_n(\bar{\mathbf{z}}_m)$
```

## 6.4 Estimating the minimum value of the regression function $f$

In this section, we apply the above results to estimate the minimum value $f^* = \min_{\mathbf{x}\in\Theta} f(\mathbf{x})$ of function $f$ that belongs to the class $\mathcal{F}_{\beta,\alpha}(L)$. Note that $f(\bar{\mathbf{z}}_n)$, which is analyzed in Theorem 6.3.2 is not an estimator for $f^*$, because it depends on the unknown $f$. The estimation of $f^*$ proceeds by estimating the minimizer and the value of the function separately on two subsamples of equal size. Throughout this section, we assume that $n$ is an even positive integer and we denote $m = n/2$. First, we split the data into two subsamples $D_1 = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ and $D_2 = \{(\mathbf{x}_{m+1}, y_{m+1}) \dots, (\mathbf{x}_n, y_n)\}$. Then, we supply Algorithm 5 with $D_1$ as the input, and we construct $\bar{\mathbf{z}}_m = \frac{1}{m}\sum_{k=1}^m \mathbf{z}_k$, where $\mathbf{z}_k$ is the update of Algorithm 5, at round $k \in [m]$. At the next step, we estimate $f(\bar{\mathbf{z}}_n)$ for fixed $\bar{\mathbf{z}}_n$ to obtain an estimator for $f^*$. At this step we can use any rate optimal estimator. To be specific, we take a regularized local polynomial estimator defined in the same spirit as the estimator of the gradient (6.2). For $\lambda, h > 0$ we define $\hat{\boldsymbol{\theta}}_{m:n,\lambda}(\mathbf{z}) = (B_{m:n}(\mathbf{z}) + \lambda\mathbf{I})^{-1}\boldsymbol{D}_{m:n}(\mathbf{z})$, where we introduced

$$B_{m:n}(\mathbf{z}) = \frac{2}{nh^d}\sum_{i=m+1}^{n} \boldsymbol{U}\left(\frac{\mathbf{x}_i - \mathbf{z}}{h}\right)\boldsymbol{U}^\top\left(\frac{\mathbf{x}_i - \mathbf{z}}{h}\right)K\left(\frac{\mathbf{x}_i - \mathbf{z}}{h}\right) \quad ,$$

$$\boldsymbol{D}_{m:n}(\mathbf{z}) = \frac{2}{nh^d}\sum_{i=m+1}^{n} y_i\boldsymbol{U}\left(\frac{\mathbf{x}_i - \mathbf{z}}{h}\right)K\left(\frac{\mathbf{x}_i - \mathbf{z}}{h}\right) \quad .$$

*Regularized local polynomial estimator* of the function $f$ at point $\mathbf{z}$ is defined as

$$f_n(\mathbf{z}) = \boldsymbol{U}^\top(0)\hat{\boldsymbol{\theta}}_{m:n,\lambda}(\mathbf{z}) \quad . \tag{6.5}$$

Minimum value estimation by local polynomial estimator is outlined in Algorithm 6.

Local polynomial estimator is known to be optimal (Stone, 1982), however, to the best of our knowledge, all the rates considered in the literature are asymptotic and hold only for sufficiently big $n$. Below we provide an upper bound for (6.5) which is non-asymptotic.

**Theorem 6.4.1.** *Under Assumptions 6.2.2, 6.2.3, and 6.2.4, for any $\boldsymbol{x} \in \Theta$, we have*

$$\mathbf{E}\left[\left(f_n(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right] \le \left(B_{\mathit{bias}}^2 + B_{\mathit{var}}\right) n^{-\frac{2\beta}{2\beta+d}} \quad .$$

Note that this theorem may be of independent interest since, to the best of our knowledge, the non-asymptotic rates of convergence of a regularized local polynomial estimator have not been studied in the literature.

The following theorem gives the rate of convergence for the estimator $T_n = f_n(\bar{\mathbf{z}}_m)$.

**Theorem 6.4.2.** *Assume that $f$ satisfies Assumptions 6.2.2, 6.2.3, and 6.2.4, and $\beta \geq 2$. Then, we have*

$$\mathbf{E}\left|T_n - f(\boldsymbol{x}^*)\right| \leq \max\left(1, \alpha^{-1}\right) \cdot \begin{cases} c_1(\log(n)/n)^{\frac{2}{4+d}} & \text{if } \beta = 2 \ , \\ c_2 n^{-\frac{\beta}{2\beta+d}} & \text{if } \beta > 2 \ , \end{cases} \tag{6.6}$$

*where $c_1, c_2 > 0$ are numerical constants.*

*Proof.* Using the fact that, for any fixed $\mathbf{x}$ the estimator $f_n(\mathbf{x})$ is measurable with respect to the second half of the sample and $\bar{\mathbf{z}}_m$ is measurable with respect to its first half we get

$$\begin{aligned} \mathbf{E}\left|T_n - f(\mathbf{x}^*)\right| &\leq \mathbf{E}|f_n(\bar{\mathbf{z}}_m) - f(\bar{\mathbf{z}}_m)| + \mathbf{E}|f(\bar{\mathbf{z}}_m) - f(\mathbf{x}^*)| \\ &\leq \mathbf{E}\left[\left(\mathbf{E}\left[\left(f_n(\bar{\mathbf{z}}_m) - f(\bar{\mathbf{z}}_m)\right)^2 |\bar{\mathbf{z}}_m\right]\right)^{\frac{1}{2}}\right] + \mathbf{E}|f(\bar{\mathbf{z}}_m) - f(\mathbf{x}^*)| \ . \end{aligned}$$

By using the fact that $f_n(\cdot)$ and $\bar{\mathbf{z}}_m$ are independent, and by Theorems 6.3.2 and 6.4.1, we deduce

$$\begin{aligned} \mathbf{E}\left|T_n - f(\mathbf{x}^*)\right| &\leq \max\left(1, \alpha^{-1}\right)\left(c_3 n^{-\frac{\beta}{2\beta+d}} + c_4\left(\frac{\log(n)}{n}\right)^{\frac{2(\beta-1)}{2\beta+d}}\right) \\ &\leq \max\left(1, \alpha^{-1}\right) \cdot \begin{cases} c_1(\log(n)/n)^{\frac{2}{4+d}} & \text{if } \beta = 2 \ , \\ c_2 n^{-\frac{\beta}{2\beta+d}} & \text{if } \beta > 2 \ , \end{cases} \end{aligned}$$

where $c_1, c_2, c_3, c_4 > 0$ are numerical constants. □

Theorem 6.4.2 shows that estimation of $f^*$ for smooth and strongly convex functions under passive design is realized with the same rate as function estimation. The lower bound (6.9) below shows that the slow rate in (6.6) cannot be improved in a minimax sense and it corresponds to the rate of a smooth function estimation at a *fixed* point. We show below that the rate $(n/\log n)^{-\beta/(2\beta+d)}$ is optimal for $\beta$-smooth regression functions without strong convexity assumption. It corresponds to the rates of function estimation in supremum norm. The strong convexity assumption allows us to reduce the global function reconstruction problem to a simpler, point estimation, leading to the rates without extra logarithmic factor. Note that the rate $n^{-\beta/(2\beta+d)}$ cannot be improved even when $\mathbf{x}^*$ is known as the function estimation at the point of minimum is still required.

Note that, for $\beta > 2$, the convergence rate of Algorithm 1 used at the first stage to estimate the minimizer is more than needed to achieve the rate (6.6). The optimal estimate of $f^*$ can be obtained by estimating the minimizer at a slower rate, namely, $n^{-\beta/(2\beta+d)}$ for the optimization risk. Therefore, it is not necessary to have $\mathbf{z}_n$ as an estimator at the first step - it can be replaced by some suboptimal estimators. This could be beneficial considering the fact that suboptimal algorithms may be computationally less costly.

In an active design setting, much faster rate can be obtained, see Table 6.1. Specifically, $f^*$ can be estimated with the parametric rate $Cn^{-1/2}$ where $C > 0$ is a constant, which is independent of the dimension $d$ and smoothness $\beta$ for any $\beta > 2$ and all $n$ large enough (Akhavan et al., 2020). Clearly, the rate $n^{-1/2}$ cannot be improved even by using the ideal but non-realizable oracle that makes all queries at point $\mathbf{x}^*$.

## 6.5 Lower bounds

The following theorem provides lower bounds for the minimax risks of arbitrary estimators on the class $\mathcal{F}_{\beta,\alpha}(L)$. Let $w(\cdot)$ be a monotone non-decreasing function on $[0,\infty)$ such that $w(0) = 0$ and $w \not\equiv 0$.

**Theorem 6.5.1.** *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be i.i.d. random vectors with a bounded Lebesgue density on $\mathbb{R}^d$. Assume that the random variables $\xi_i$ are i.i.d. having a density $p_\xi(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}$ such that*

$$\exists I_* > 0, v_0 > 0 : \quad \int \left(\sqrt{p_\xi(u)} - \sqrt{p_\xi(u+v)}\right)^2 \mathrm{d}u \leq I_* v^2 \ , \tag{6.7}$$

*for $|v| \leq v_0$. Then, for any $\beta, \alpha, L > 0$ we have*

$$\inf_{\boldsymbol{x}_n} \sup_{f \in \mathcal{F}_{\beta,\alpha}(L)} \mathbf{E}_f w(n^{\frac{\beta-1}{2\beta+d}} \|\boldsymbol{x}_n - \boldsymbol{x}^*\|) \geq c_1, \tag{6.8}$$

*and*

$$\inf_{f_n} \sup_{f \in \mathcal{F}_{\beta,\alpha}(L)} \mathbf{E}_f w(n^{\frac{\beta}{2\beta+d}} |f_n - f^*|) \geq c_1', \tag{6.9}$$

*where $\inf_{\boldsymbol{x}_n}$ and $\inf_{f_n}$ denote the infimum over all estimators of the minimizer and over all estimators of the minimum value of $f$, respectively, and $c_1 > 0, c_1' > 0$ are constants that depend only on $\beta, \alpha, L, \Theta, I_*, v_0$, and $w(\cdot)$.*

Condition (6.7) is rather general. It is satisfied, for example, for the Gaussian distribution and also for a large class of regular densities, cf. Ibragimov and Khas'minskii (1981). The lower bound (6.8) was proved in Tsybakov (1990a) under a more restrictive condition on the density $p_\xi$.

The proof of Theorem 6.5.1 is given in Section 6.7. It is based on a reduction to the problem of testing two hypotheses.

Considering the bounds (6.9), (6.8) with $w(u) = u^2$ and $w(u) = u$, respectively, and combining them with Theorems 6.3.1 and 6.4.2 we obtain that the estimator $T_n$ is minimax optimal for $f^*$, and $\mathbf{z}_n$ is minimax optimal up to a logarithmic factor for $\mathbf{x}^*$ on the class of functions $\mathcal{F}_{\beta,\alpha}(L)$.

In the next theorem, we provide a minimax lower bound on estimation of $f^*$ over the class of $\beta$-Hölder functions $\mathcal{F}_\beta(L)$ when there is no strong convexity assumption.

**Theorem 6.5.2.** *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be i.i.d. random vectors with a bounded Lebesgue density on $\mathbb{R}^d$, and let $\xi_i$ be i.i.d. Gaussian random variables with zero mean and variance $\sigma^2$. Assume that $\Theta$ contains an open subset of $\mathbb{R}^d$. Then, for any $\beta > 0, L > 0,$ we have*

$$\inf_{f_n} \sup_{f \in \mathcal{F}_\beta(L)} \mathbf{E}_f w \left( \left( \frac{n}{\log n} \right)^{\frac{\beta}{2\beta+d}} |f_n - f^*| \right) \geq c_3,$$

*where $\inf_{f_n}$ denotes the infimum over all estimators of the minimum value of $f$ and $c_3 > 0$ is a constant that depends only on $\beta, \alpha, L, \Theta, \sigma^2$, and $w(\cdot)$.*

Theorem 6.5.2 implies that $\left(\frac{n}{\log n}\right)^{-\frac{\beta}{2\beta+d}}$ is the minimax rate of estimating the minimum value $f^*$ on the class $\mathcal{F}_\beta(L)$. Indeed, the matching upper bound with the rate $\left(\frac{n}{\log n}\right)^{-\frac{\beta}{2\beta+d}}$ is obtained in a trivial way if we estimate $f^*$ by the minimum of any rate optimal (in supremum norm) nonparametric estimator of $f$, for example, by the local polynomial estimator as in Stone (1982).

Thus, if we drop the assumption of strong convexity, the minimax rate deteriorates only by a logarithmic factor. It suggests that strong convexity is not a crucial advantage in estimation of the minimum value of a function under the passive design.

## 6.6 Conclusion

In this paper, we have considered the problem of estimating the minimizer and the minimum value of the regression function from the i.i.d data with a special focus on highly smooth and strongly convex regression functions. We provide upper bounds for the proposed algorithms. We show that the rates of estimation of the minimizer is the same as the rate for estimating the gradient of the regression function. To estimate the minimum value we have used two-stage procedure where in the first step we estimate the location of the minimum followed by the estimation of the function value at the estimated in the first step point. We obtain optimal nonparametric rates of convergence for our two-stage procedure.

An interesting open question is to make our algorithms adaptive to the unknown smoothness $\beta$, that is, to develop a data-driven choice of the smoothing parameter $h$ and of the regularization parameter $\lambda$. When considering adaptation to the unknown smoothness of function $f$, the optimal rates for estimation of $f^*$ will be presumably slower than the minimax rates by a logarithmic factor.

## 6.7 Proofs

In this section, we provide the proofs of Theorems 6.3.1, 6.4.1 and 6.5.1. Section 6.7 is devoted to the proof of Theorem 6.3.1 on the upper bound for the Algorithm 1. Section 6.7 provides the proof of Theorem 6.4.1. In section 6.7, we prove Theorem 6.5.1 on the lower bounds.

### Proof of Theorem 6.3.1

For the proof of Theorem 6.3.1 we need some preliminary lemmas.

**Lemma 6.7.1.** *For $k \in [n]$, let $\boldsymbol{g}_{k,\lambda}$ be defined by (6.3). Under Assumptions 6.2.2, 6.2.3, and 6.2.4, for any $\boldsymbol{x} \in \Theta$ the following upper bound holds*

$$\left\|\mathbf{E}[\boldsymbol{g}_{k,\lambda}(\boldsymbol{x})] - \nabla f(\boldsymbol{x})\right\| \leq \mathtt{A}_{\textit{bias}} \left(\frac{\log(k+1)}{k}\right)^{\frac{\beta-1}{2\beta+d}} \quad ,$$

*where $\mathtt{A}_{\textit{bias}} > 0$ is a numerical constant.*

*Proof.* We introduce a shorter notation. For any $k \in [n]$, and $i \in [k]$, let

$$M_{i,k}(\mathbf{x}) = \boldsymbol{U}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right)\boldsymbol{U}^\top\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right)K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) \quad , \quad \text{and} \quad \boldsymbol{R}_{i,k}(\mathbf{x}) = \boldsymbol{U}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right)K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) \quad .$$

Also, we use the notation $\boldsymbol{C}_k(\mathbf{x}) = \frac{1}{kh_k^d}\sum_{i=1}^k \boldsymbol{R}_{i,k}(\mathbf{x})f(\mathbf{x}_i)$, $\boldsymbol{D}_k(\mathbf{x}) = \frac{1}{kh_k^d}\sum_{i=1}^k \boldsymbol{R}_{i,k}(\mathbf{x})y_i$, and note that $\mathbf{E}\left[\boldsymbol{C}_k(\mathbf{x})\right] = \mathbf{E}\left[\boldsymbol{D}_k(\mathbf{x})\right]$.

To shorten the notation, set $B = \mathbf{E}[B_k(\mathbf{x})]$. By letting $\phi_k = \mathbf{g}_{k,\lambda}(\mathbf{x}) - h_k^{-1}\left(AB^{-1}B_k(\mathbf{x})\boldsymbol{c}_k(f,\mathbf{x})\right)$, we can write

$$\mathbf{E}\left[\phi_k\right] = \mathbf{E}\left[\mathbf{g}_{k,\lambda}(\mathbf{x})\right] - h_k^{-1}A\boldsymbol{c}_k(f,\mathbf{x}) = \mathbf{E}\left[\mathbf{g}_{k,\lambda}(\mathbf{x})\right] - \nabla f(\mathbf{x}) \ ,$$

where $\boldsymbol{c}_k(f,\mathbf{x}) = \left(h_k^{|\boldsymbol{m}^{(1)}|}D^{\boldsymbol{m}^{(1)}}f(\mathbf{x}), \ldots, h_k^{|\boldsymbol{m}^{(S)}|}D^{\boldsymbol{m}^{(S)}}f(\mathbf{x})\right)^\top$. Also, note that by Assumption 6.2.2, $\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{x})] = \mathbf{E}\left[h_k^{-1}\left(AB_{k,\lambda}^{-1}(\mathbf{x})\boldsymbol{C}_k(\mathbf{x})\right)\right]$. To conclude the proof, we need to provide an upper bound for the term $\|\mathbf{E}\left[\phi_k\right]\|$. Let

$$\begin{aligned}
\boldsymbol{\psi}_{1,k} &= h_k^{-1}\left(AB^{-1}\boldsymbol{C}_k(\mathbf{x})\right) \quad , \\
\boldsymbol{\psi}_{2,k} &= h_k^{-1}\left(A(B+\lambda_k\mathbf{I})^{-1}\boldsymbol{C}_k(\mathbf{x})\right) \quad .
\end{aligned}$$

Then we have

$$\begin{aligned}
\|\mathbf{E}[\boldsymbol{\phi}_k]\| &= \left\|\mathbf{E}\left[\left(\boldsymbol{\psi}_{1,k} - h_k^{-1}\left(AB^{-1}B_k(\mathbf{x})\boldsymbol{c}_k(f,\mathbf{x})\right)\right) + \left(\boldsymbol{\psi}_{2,k} - \boldsymbol{\psi}_{1,k}\right) + \left(\mathbf{g}_{k,\lambda}(\mathbf{x}) - \boldsymbol{\psi}_{2,k}\right)\right]\right\| \\
&\leq \underbrace{\left\|\mathbf{E}[\boldsymbol{\psi}_{1,k} - h_k^{-1}\left(AB^{-1}B_k(\mathbf{x})\boldsymbol{c}_k(f,\mathbf{x})\right)]\right\|}_{\text{term I}} + \underbrace{\left\|\mathbf{E}[\boldsymbol{\psi}_{2,k} - \boldsymbol{\psi}_{1,k}]\right\|}_{\text{term II}} + \underbrace{\left\|\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{x}) - \boldsymbol{\psi}_{2,k}]\right\|}_{\text{term III}} \quad .
\end{aligned}$$

We provide adequately tight upper bounds for each of the terms above separately. For term I, we can write

$$\begin{aligned}\text{term I} &= h_k^{-1}\left\|AB^{-1}\mathbf{E}\left[\frac{1}{kh_k^d}\sum_{i=1}^{k}\boldsymbol{R}_{i,k}(\mathbf{x})\left(f(\mathbf{x}_i)-\boldsymbol{U}^\top\left(\frac{\mathbf{x}_i-\mathbf{x}}{h_k}\right)\boldsymbol{c}_k(f,\mathbf{x})\right)\right]\right\|\\ &\leq h_k^{-1}\left\|AB^{-1}\right\|_{\text{op}}\left\|\mathbf{E}\left[\frac{1}{kh_k^d}\sum_{i=1}^{k}\boldsymbol{R}_{i,k}(\mathbf{x})\left(f(\mathbf{x}_i)-\boldsymbol{U}^\top\left(\frac{\mathbf{x}_i-\mathbf{x}}{h_k}\right)\boldsymbol{c}_k(f,\mathbf{x})\right)\right]\right\| \ .\end{aligned}$$

Since $\|A\|_{\text{op}} \leq 1$, by Lemma 6.7.8(iii), we deduce that $\left\|AB^{-1}\right\|_{\text{op}} \leq \lambda_{\min}^{-1}$. Then we can write

$$\text{term I} \leq h_k^{-1}\lambda_{\min}^{-1}\left(\frac{1}{kh_k^d}\sum_{i=1}^{k}\mathbf{E}\left[\left\|\boldsymbol{R}_{i,k}(\mathbf{x})\left(f(\mathbf{x}_i)-\boldsymbol{U}^\top\left(\frac{\mathbf{x}_i-\mathbf{x}}{h_k}\right)\boldsymbol{c}_k(f,\mathbf{x})\right)\right\|\right]\right) \ .$$

Since by Assumption 6.2.3, $f \in \mathcal{F}_\beta(L)$ for any $i \in [k]$ we have

$$\left|f(\mathbf{x}_i)-\boldsymbol{U}^\top\left(\frac{\mathbf{x}_i-\mathbf{x}}{h_k}\right)\boldsymbol{c}_k(f,\mathbf{x})\right| \leq L\left\|\mathbf{x}-\mathbf{x}_i\right\|^\beta \ ,$$

and we can write

$$\begin{aligned}\text{term I} &\leq Lh_k^{-1}\lambda_{\min}^{-1}\left(\frac{1}{kh_k^d}\sum_{i=1}^{k}\mathbf{E}\left[\left\|\boldsymbol{R}_{i,k}(\mathbf{x})\right\|\left\|\mathbf{x}-\mathbf{x}_i\right\|^\beta\right]\right)\\ &= Lh_k^{-d-1}\lambda_{\min}^{-1}\int_{\mathbb{R}^d}\|\mathbf{x}-\boldsymbol{u}\|^\beta\left\|\boldsymbol{U}\left(\frac{\boldsymbol{u}-\mathbf{x}}{h_k}\right)K\left(\frac{\boldsymbol{u}-\mathbf{x}}{h_k}\right)\right\|p(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\\ &= Lh_k^{\beta-1}\lambda_{\min}^{-1}\int_{\mathbb{R}^d}\|\boldsymbol{w}\|^\beta\left\|U(\boldsymbol{w})K(\boldsymbol{w})\right\|p(\mathbf{x}+h_k\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} \leq \mathtt{A}_1h_k^{\beta-1} \ ,\end{aligned}$$

where we introduced $\mathtt{A}_1 = L\lambda_{\min}^{-1}p_{\max}\kappa_\beta$, and $\kappa_\beta = \int_{\mathbb{R}^d}\|\boldsymbol{u}\|^\beta\left\|U(\boldsymbol{u})K(\boldsymbol{u})\right\|\,\mathrm{d}\boldsymbol{u}$. For term II, we deduce that

$$\begin{aligned}\text{term II} &= h_k^{-1}\left\|A\left((B+\lambda_k\mathbf{I})^{-1}-B^{-1}\right)\mathbf{E}\left[\boldsymbol{C}_k(\mathbf{x})\right]\right\|\\ &\leq h_k^{-1}\lambda_k\left\|A\right\|_{\text{op}}\left\|B^{-1}\right\|_{\text{op}}\left\|(B+\lambda_k\mathbf{I})^{-1}\right\|_{\text{op}}\mathbf{E}\left[\left\|\boldsymbol{C}_k(\mathbf{x})\right\|\right] \ .\end{aligned}$$

By Assumption 6.2.3(iii), we have $\sup_{\mathbf{x}\in\Theta'} f(\mathbf{x}) \leq M$. Also, $\mathbf{E}\left[\left\|\boldsymbol{C}_k(\mathbf{x})\right\|\right] \leq \mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{C}_k(\mathbf{x})\right\|\right]$, where by Lemma 6.7.8(ii) we get $\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{C}_k(\mathbf{x})\right\|\right] \leq Mp_{\max}\nu_{1,1}$. Moreover, by Lemma 6.7.8(iii), we can write $\left\|B^{-1}\right\|_{\text{op}}\left\|(B+\lambda_k\mathbf{I})^{-1}\right\|_{\text{op}} \leq \lambda_{\min}^{-2}$. Therefore, we deduce that

$$\text{term II} \leq \mathtt{A}_2h_k^{-1}\lambda_k \ ,$$

with $\mathtt{A}_2 = Mp_{\max}\nu_{1,1}\lambda_{\min}^{-2}$. Finally, we need to bound term III

$$\begin{aligned}\text{term III} &\leq h_k^{-1}\left\|\mathbf{E}\left[A\left(B_{k,\lambda}^{-1}(\mathbf{x}) - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right)(\boldsymbol{C}_k(\mathbf{x}) - \mathbf{E}\left[\boldsymbol{C}_k(\mathbf{x})\right])\right]\right\| \\ &\quad + h_k^{-1}\left\|\mathbf{E}\left[A\left(B_{k,\lambda}^{-1}(\mathbf{x}) - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right)\mathbf{E}\left[\boldsymbol{C}_k(\mathbf{x})\right]\right]\right\| \\ &\leq h_k^{-1}\mathbf{E}\left[\left\|B_{k,\lambda}^{-1}(\mathbf{x}) - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right\|_{\mathsf{op}}\|\boldsymbol{C}_k(\mathbf{x}) - \mathbf{E}\left[\boldsymbol{C}_k(\mathbf{x})\right]\|\right] \\ &\quad + h_k^{-1}\mathbf{E}\left[\left\|B_{k,\lambda}^{-1}(\mathbf{x}) - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right\|_{\mathsf{op}}\right]\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|\boldsymbol{C}_k(\mathbf{x})\|\right] \quad .\end{aligned}$$

For the first term on the r.h.s., we use Lemma 6.7.14, and we get

$$\text{term III} \leq \mathtt{A}_3 k^{-1}h_k^{-d-1} + h_k^{-1}\mathbf{E}\left[\left\|B_{k,\lambda}(\mathbf{x})^{-1} - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right\|_{\mathsf{op}}\right]\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|\boldsymbol{C}_k(\mathbf{x})\|\right] \quad ,$$

where $\mathtt{A}_3 > 0$ is the numerical constant that appears in Lemma 6.7.14. By invoking Lemma 6.7.8(i), the second term on the r.h.s. can be bounded by the following expression

$$\begin{aligned}&\mathtt{A}_4 h_k^{-1}\mathbf{E}\left[\|B_{k,\lambda}(\mathbf{x})\|_{\mathsf{op}}^{-1}\|B_{k,\lambda}(\mathbf{x}) - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])\|_{\mathsf{op}}\right] \leq \\ &\qquad \mathtt{A}_4 h_k^{-1}\left(\mathbf{E}\left[\|B_{k,\lambda}(\mathbf{x})\|_{\mathsf{op}}^{-2}\right]\mathbf{E}\left[\|B_{k,\lambda}(\mathbf{x}) - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])\|_{\mathsf{op}}^{2}\right]\right)^{\frac{1}{2}} \quad ,\end{aligned}$$

where for the last display we used the Cauchy-Schwarz inequality and we introduced $\mathtt{A}_4 = Mp_{\max}\nu_{1,1}\lambda_{\min}^{-1}$. Now, by using Jensen's inequality and Lemma 6.7.10, we get

$$\text{term III} \leq \mathtt{A}_3 k^{-1}h_k^{-d-1} + \mathtt{A}_4 k^{-\frac{1}{2}}h_k^{-\frac{d}{2}-1} \leq \mathtt{A}_5 k^{-\frac{1}{2}}h_k^{-\frac{d}{2}-1},$$

where $\mathtt{A}_5 = 3\mathtt{A}_3 + \mathtt{A}_4$. By combining all of these bounds we obtain

$$\left\|\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{x})] - \nabla f(\mathbf{x})\right\| \leq \mathtt{A}_6\left(h_k^{\beta-1} + h_k^{-1}\lambda_k + h_k^{-1-\frac{d}{2}}k^{-\frac{1}{2}}\right) \quad , \tag{6.10}$$

where $\mathtt{A}_6 = \max(\mathtt{A}_1, \mathtt{A}_2, \mathtt{A}_5)$. Finally, by substituting $h_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{1}{2\beta+d}}$, and $\lambda_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{\beta}{2\beta+d}}$, we deduce that

$$\left\|\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{x})] - \nabla f(\mathbf{x})\right\| \leq \mathtt{A}_{\mathsf{bias}}\left(\frac{\log(k+1)}{k}\right)^{\frac{\beta-1}{2\beta+d}} \quad ,$$

where $\mathtt{A}_{\mathsf{bias}} = 3\mathtt{A}_6$. □

The following lemma provides a bound of the variance uniformly over $\Theta$.

**Lemma 6.7.2.** *Let $\mathbf{g}_{k,\lambda}$ be defined by Algorithm 1, and let Assumptions 6.2.2, 6.2.3, and 6.2.4* *hold. Then, we have*

$$\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\left\|\boldsymbol{g}_{k,\lambda}(\boldsymbol{x})-\mathbf{E}\left[\boldsymbol{g}_{k,\lambda}(\boldsymbol{x})\right]\right\|^2\right]\leq \mathtt{A}_{\textit{var}}\left(\frac{\log(k+1)}{k}\right)^{\frac{2(\beta-1)}{2\beta+d}} \quad ,$$

*where* $\mathtt{A}_{\textit{var}}>0$ *is a numerical constant.*

*Proof.* Let $\boldsymbol{G}_k(\mathbf{x})=\frac{1}{kh_k^d}\sum_{i=1}^k \boldsymbol{R}_{i,k}(\mathbf{x})\xi_i$, and recall that $\boldsymbol{C}_k(\mathbf{x})=\frac{1}{kh_k^d}\sum_{i=1}^k \boldsymbol{R}_{i,k}(\mathbf{x})f(\mathbf{x}_i)$. Then, we have

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\mathbf{g}_{k,\lambda}(\mathbf{x})-\mathbf{E}\left[\mathbf{g}_k(\mathbf{x})\right]\right\|^2\right]\leq \underbrace{2\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|h_k^{-1}AB_{k,\lambda}^{-1}(\mathbf{x})\boldsymbol{G}_k(\mathbf{x})\right\|^2\right]}_{\text{term I}}$$
$$+\underbrace{2\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|h_k^{-1}AB_{k,\lambda}^{-1}(\mathbf{x})\boldsymbol{C}_k(\mathbf{x})-\mathbf{E}\left[h_k^{-1}AB_{k,\lambda}^{-1}(\mathbf{x})\boldsymbol{C}_k(\mathbf{x})\right]\right\|^2\right]}_{\text{term II}} \quad .$$

For term I, we have

$$\text{term I}\leq \underbrace{4h_k^{-2}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|A\left(B_{k,\lambda}^{-1}(\mathbf{x})-(\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right)\boldsymbol{G}_k(\mathbf{x})\right\|^2\right]}_{\text{term III}}$$
$$+\underbrace{4h_k^{-2}\left\|(\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right\|_{\text{op}}^2\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{G}_k(\mathbf{x})\right\|^2\right]}_{\text{term IV}} \quad .$$

For term III, by using the property of Assumption 6.2.2, we can write

$$\text{term III}\leq 4\sigma^2\lambda_{\min}^{-2}\lambda_k^{-2}h_k^{-2}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|B_{k,\lambda}(\mathbf{x})-\mathbf{E}[B_{k,\lambda}(\mathbf{x})]\right\|_{\text{op}}^2\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{G}(\mathbf{x})\right\|^2\right] \quad .$$

Now, by invoking the Cauchy-Schwarz inequality we get

$$\text{term III}\leq 4\sigma^2\lambda_{\min}^{-2}\lambda_k^{-2}h_k^{-2}\left(\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|B_{k,\lambda}(\mathbf{x})-\mathbf{E}[B_{k,\lambda}(\mathbf{x})]\right\|_{\text{op}}^4\right]\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{G}(\mathbf{x})\right\|^4\right]\right)^{\frac{1}{2}}$$

Using Lemmas 6.7.9 and 6.7.11, we obtain

$$\text{term III}\leq \mathtt{A}_1\lambda_k^{-2}k^{-2}h_k^{-2d-2}\log(k+1)^2 \quad ,$$

where $\mathtt{A}_1>0$ is a numerical constant. Now, by using the inequality $k\geq\lambda_k^{-2}h_k^{-d}\log(k+1)$, we can write

$$\text{term III}\leq \mathtt{A}_1k^{-1}h_k^{-d-2}\log(k+1) \quad .$$

For term IV, we have

$$\text{term IV} \leq 4\sigma^2\lambda_{\min}^{-2}k^{-1}h_k^{-2d-2}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|R_{1,k}\|^2\right] \leq \mathtt{A}_2 k^{-1}h_k^{-d-2} \ ,$$

where the last inequality is obtained by Lemma 6.7.8(i), with $\mathtt{A}_2 = 4\sigma^2\lambda_{\min}^{-2}p_{\max}\nu_{1,2}$. Therefore, we deduce that

$$\text{term I} \leq \mathtt{A}_3 k^{-1}h_k^{-d-2} \ ,$$

with $\mathtt{A}_3 = \mathtt{A}_1 + \mathtt{A}_2$. We continue the proof by providing an adequately tight upper bound for term II.

$$\begin{aligned}\text{term II} \leq\ & \underbrace{4h_k^{-2}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|B_{k,\lambda}^{-1}(\mathbf{x}) - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right\|_{\text{op}}^2 \|\boldsymbol{C}_k(\mathbf{x})\|^2\right]}_{\text{term V}} \\ & + \underbrace{4h_k^{-2}\left\|(\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right\|_{\text{op}}^2 \mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|\boldsymbol{C}_k(\mathbf{x}) - \mathbf{E}[\boldsymbol{C}_k(\mathbf{x})]\|^2\right]}_{\text{term VI}} \ .\end{aligned}$$

Similar to term III, for term V we have

$$\text{term V} \leq 4M^2\lambda_{\min}^{-2}\lambda_k^{-2}k^{-1}h_k^{-d-2}\log(k+1)\left(k^{-3}h_k^{-3d}p_{\max}\nu_{1,4} + k^{-2}h_k^{-2d}p_{\max}^2\nu_{1,2}^2\right)^{\frac{1}{2}} \leq \mathtt{A}_4 k^{-1}h_k^{-d-2} \ ,$$

where $\mathtt{A}_4 = 4M^2\lambda_{\min}^{-2}(p_{\max}\nu_{1,4} + p_{\max}^2\nu_{1,2}^2)^{\frac{1}{2}}$. Finally, for term VI, by Lemma 6.7.13 we can write

$$\text{term VI} \leq \mathtt{A}_5 k^{-1}h_k^{-d-2}\log(k+1) \ ,$$

where $\mathtt{A}_5 > 0$ is a numerical constant. Thus, we deduce that

$$\text{term II} \leq \mathtt{A}_6 k^{-1}h_k^{-d-2}\log(k+1) \ ,$$

with $\mathtt{A}_6 = \mathtt{A}_4 + \mathtt{A}_5$. We conclude the proof by letting $\mathtt{A}_{\mathsf{var}} = \mathtt{A}_3 + \mathtt{A}_6$, and substituting the parameters $h_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{1}{2\beta+d}}$, and $\lambda_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{\beta}{2\beta+d}}$. □

**Lemma 6.7.3.** *Let $\boldsymbol{g}_{k,\lambda}$ be defined by Algorithm 1, and let Assumptions 6.2.2, 6.2.3, and 6.2.4 hold. Then, we have*

$$\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\left\|\boldsymbol{g}_{k,\lambda}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^2\right] \leq \mathtt{A}_{\mathit{error}}\left(\frac{\log(k+1)}{k}\right)^{\frac{2(\beta-1)}{2\beta+d}} \ ,$$

*where $\mathtt{A}_{\mathit{error}} > 0$ is a numerical constant.*

*Proof.* We can write

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\mathbf{g}_{k,\lambda}(\mathbf{x})-\nabla f(\mathbf{x})]\right\|^2\right]\leq\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\mathbf{g}_{k,\lambda}(\mathbf{x})-\mathbf{E}\left[g_{k,\lambda}(\mathbf{x})\right]\right\|^2\right]+\sup_{\mathbf{x}\in\Theta}\left\|\mathbf{E}\left[g_{k,\lambda}(\mathbf{x})\right]-\nabla f(\mathbf{x})\right\|^2 \quad .$$

We conclude the proof by using Lemmas 6.7.1 and 6.7.2, and letting $\mathtt{A}_{\text{error}}=\mathtt{A}_{\text{bias}}^2+\mathtt{A}_{\text{var}}$. □

**Lemma 6.7.4.** *Let $\boldsymbol{g}_{k,\lambda}$ be defined by Algorithm 1, and let Assumptions 6.2.2, 6.2.3, and 6.2.4 hold. Then, we have*

$$\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\left\|\boldsymbol{g}_{k,\lambda}(\boldsymbol{x})\right\|^2\right]\leq \mathit{A}_{\mathit{sm}}k^{\frac{2+d}{2\beta+d}}\log(k+1)^{\frac{2(\beta-1)}{2\beta+d}} \quad ,$$

*where $\mathit{A}_{\mathit{sm}}>0$ is a numerical constant.*

*Proof.* Let $\boldsymbol{G}_k(\mathbf{x})=\frac{1}{kh_k^d}\sum_{i=1}^k\boldsymbol{R}_{i,k}(\mathbf{x})\xi_i$, and recall that $\boldsymbol{C}_k(\mathbf{x})=\frac{1}{kh_k^d}\sum_{i=1}^k\boldsymbol{R}_{i,k}(\mathbf{x})f(\mathbf{x}_i)$. By the definition of $\mathbf{g}_{k,\lambda}$, we can write

$$\begin{aligned}\mathbf{E}[\sup_{\mathbf{x}\in\Theta}\|\mathbf{g}_k(\mathbf{x})\|^2]&\leq h_k^{-2}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})\|_{\text{op}}^{-2}\left\|\boldsymbol{D}_k(\mathbf{x})\right\|^2\right]\\&\leq\underbrace{2h_k^{-2}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})\|_{\text{op}}^{-2}\left\|\boldsymbol{C}_k(\mathbf{x})\right\|^2\right]}_{\text{term I}}+\underbrace{2h_k^{-2}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})\|_{\text{op}}^{-2}\left\|\boldsymbol{G}_k(\mathbf{x})\right\|^2\right]}_{\text{term II}} \quad ,\end{aligned}$$

where the last inequality uses the fact that $(u+v)^2\leq 2u^2+2v^2$, for any $u,v\geq 0$. Now, for term I, we can write

$$\begin{aligned}\text{term I}\leq&\underbrace{4h_k^{-2}\left(\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})\|_{\text{op}}^{-2}\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{C}_k(\mathbf{x})-\mathbf{E}\left[\boldsymbol{C}_k(\mathbf{x})\right]\right\|^2\right]\right)}_{\text{term III}}\\&+\underbrace{4h_k^{-2}\left(\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})\|_{\text{op}}^{-2}\right]\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{C}_k(\mathbf{x})\right\|^2\right]\right)}_{\text{term IV}} \quad ,\end{aligned}$$

To provide an upper bound for term III, we use the Cauchy-Schwarz inequality, which yields

$$\text{term III}\leq 4h_k^{-2}\left(\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})\|_{\text{op}}^{-4}\right]\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{C}_k(\mathbf{x})-\mathbf{E}\left[\boldsymbol{C}_k(\mathbf{x})\right]\right\|^4\right]\right)^{\frac{1}{2}}\leq\mathtt{A}_1k^{-1}h_k^{-d-2}\log(k+1) \quad ,$$

where we have used Lemmas 6.7.10 and 6.7.13, with $\mathtt{A}_1>0$ being a numerical constant. For term IV, by invoking Lemmas 6.7.8(i) and 6.7.10, we deduce that

$$\text{term IV}\leq\mathtt{A}_2h_k^{-d-2} \quad ,$$

where $\mathtt{A}_2>0$ is a numerical constant. Finally, it is enough to provide an upper bound for term II. We have

$$\text{term II} \leq 2h_k^{-2}\left(\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})\|_{\text{op}}^{-4}\right]\mathbf{E}\left[\|\boldsymbol{G}_k(\mathbf{x})\|^4\right]\right)^{\frac{1}{2}} \quad ,$$

where the last inequality is due to Cauchy-Schwarz. Thanks to Lemmas 6.7.9 and 6.7.11, we can write

$$\text{term II} \leq \mathtt{A}_3 k^{-1}h_k^{-d-2}\log(k+1) \quad .$$

Now it is straightforward to see that the sum of the terms is dominated by $\mathtt{A}_{\text{sm}}h_k^{-d-2}\log(k+1)$, where $\mathtt{A}_{\text{sm}} = \mathtt{A}_1 + \mathtt{A}_2 + \mathtt{A}_3$. By substituting $h_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{1}{2\beta+d}}$, we conclude the proof. □

Now, we are ready to proof Theorem 6.3.1.

*Proof of Theorem 6.3.1.* By the definition of the algorithm and the contracting property of the Euclidean projection, for any $k \in [n]$, we have

$$\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{z}_k - \mathbf{x}^*\|^2 - \frac{4}{\alpha k}\underbrace{(\mathbf{z}_k - \mathbf{x}^*)^\top \mathbf{E}\left[\mathbf{g}_{k,\lambda}(\mathbf{z}_k)|\mathbf{z}_k\right]}_{\text{term I}} + \frac{4}{\alpha^2k^2}\mathbf{E}\left[\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right\|^2|\mathbf{z}_k\right] \quad .$$

By adding and subtracting $\nabla f(\mathbf{z}_n)$, for term I we get

$$\begin{aligned}\mathbf{E}[\delta_{k+1}|\mathbf{z}_k] \leq \delta_k - \frac{4}{\alpha k}\langle \mathbf{z}_k - x^*, \nabla f(\mathbf{z}_k)\rangle + \frac{4}{\alpha k}\|\mathbf{z}_k - \mathbf{x}^*\|\left\|\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{z}_k)|\mathbf{z}_k] - \nabla f(\mathbf{z}_k)\right\| \\ + \frac{4}{\alpha^2k^2}\mathbf{E}\left[\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right\|^2|\mathbf{z}_k\right] \quad ,\end{aligned} \tag{6.11}$$

where $\delta_k = \|\mathbf{z}_k - \mathbf{x}^*\|^2$. Since $f$ is an $\alpha$-strongly function, we have

$$\alpha\delta_k \leq \langle \mathbf{z}_k - \mathbf{x}^*, \nabla f(\mathbf{z}_k)\rangle \quad . \tag{6.12}$$

Combining (6.11) and (6.12), yields

$$\mathbf{E}[\delta_{k+1}|\mathbf{z}_k] \leq \left(1 - \frac{4}{k}\right)\delta_k + \frac{4}{\alpha k}\underbrace{\|\mathbf{z}_k - \mathbf{x}^*\|\left\|\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{z}_k)|\mathbf{z}_k] - \nabla f(\mathbf{z}_k)\right\|}_{\text{term II}} + \frac{4}{\alpha^2k^2}\mathbf{E}\left[\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right\|^2|\mathbf{z}_k\right] \quad . \tag{6.13}$$

Note that for any $a, b \in \mathbb{R}$ and $\gamma > 0$, we have $2a \cdot b \leq \gamma a^2 + \frac{b^2}{\gamma}$. For term II in (6.13), we can write

$$\|\mathbf{z}_k - \mathbf{x}^*\|\left\|\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{z}_k)|\mathbf{z}_k] - \nabla f(\mathbf{z}_k)\right\| \leq \frac{3\alpha}{4}\delta_k + \frac{1}{3\alpha}\left\|\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{z}_k)|\mathbf{z}_k] - \nabla f(\mathbf{z}_k)\right\|^2 \quad .$$

Plugging in the above upper bound for term II, and taking the total expectation yields

$$\begin{aligned}\tilde{\delta}_{k+1} &\leq \left(1-\frac{1}{k}\right)\tilde{\delta}_k + \frac{1}{3\alpha^2 k}\mathbf{E}\left[\left\|\mathbf{E}[\mathbf{g}_{k,\lambda}(\mathbf{z}_k)|\mathbf{z}_k] - \nabla f(\mathbf{z}_k)\right\|^2\right] + \frac{4}{\alpha^2 k^2}\mathbf{E}\left[\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right\|^2\right] \\ &\leq \left(1-\frac{1}{k}\right)\tilde{\delta}_k + \frac{1}{3\alpha^2 k}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\mathbf{g}_{k,\lambda}(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2\right] + \frac{4}{\alpha^2 k^2}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\mathbf{g}_{k,\lambda}(\mathbf{x})\right\|^2\right] \quad ,\end{aligned}$$

where $\tilde{\delta}_k = \mathbf{E}[\delta_k]$, and form first to second inequality we used Jensen's inequality. By invoking Lemmas 6.7.3 and 6.7.4, we deduce that

$$\tilde{\delta}_{k+1} \leq \left(1-\frac{1}{k}\right)\tilde{\delta}_k + \mathtt{A}_1 k^{-1-\frac{2(\beta-1)}{2\beta+1}}\log(k+1)^{\frac{2(\beta-1)}{2\beta+d}}\alpha^{-2} \quad ,$$

where $\mathtt{A}_1 = \mathtt{A}_{\text{error}}/3 + 4\mathtt{A}_{\text{sm}}$, is a numerical constant. Finally, by using (Akhavan et al., 2020, Lemma D.1.) we conclude the proof:

$$\mathbf{E}\left\|\mathbf{z}_n - \mathbf{x}^*\right\|^2 \leq \left(\frac{4\text{diam}(\Theta)}{n} + \mathtt{A}_2 n^{-\frac{2(\beta-1)}{2\beta+d}}\alpha^{-2}\right)\log(n)^{\frac{2(\beta-1)}{2\beta+d}} \quad ,$$

where $\mathtt{A}_2 = \frac{4\beta+4}{d+2}\mathtt{A}_1$, $\text{diam}(\Theta) = \sup_{\mathbf{x},\mathbf{y}\in\Theta}\left\|\mathbf{x}-\mathbf{y}\right\|^2$, and in order to obtain the last inequality we used the fact that $\log(n+1) \leq \log(n)$ for $n \geq 2$. □

*Proof of Theorem 6.3.2.* By the definition of Algorithm 1, we have $\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 \leq \left\|\mathbf{z}_k - \eta_k \mathbf{g}_{k,\lambda}(\mathbf{z}_k) - \mathbf{x}^*\right\|^2$. Therefore, we can write

$$\langle \mathbf{g}_{k,\lambda}(\mathbf{z}_k), \mathbf{z}_k - \mathbf{x}^*\rangle \leq \frac{\|\mathbf{z}_k - \mathbf{x}^*\|^2 - \|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2}{2\eta_k} + \frac{\eta_k}{2}\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right\|^2 \quad . \tag{6.14}$$

On the other hand, by Assumption 6.2.3(iii) we have

$$f(\mathbf{z}_k) - f^* \leq \langle \nabla f(\mathbf{z}_k), \mathbf{z}_k - \mathbf{x}^*\rangle - \frac{\alpha}{2}\|\mathbf{z}_k - \mathbf{x}^*\|^2 \quad . \tag{6.15}$$

Combining (6.14) and (6.15) gives

$$\begin{aligned}\mathbf{E}\left[f(\mathbf{z}_k) - f^*|\mathbf{z}_k\right] \leq \left\|\mathbf{E}\left[\mathbf{g}_{k,\lambda}(\mathbf{z}_k) - \nabla f(\mathbf{z}_k)|\mathbf{z}_k\right]\right\|\|\mathbf{z}_k - \mathbf{x}^*\| + \frac{1}{2\eta_k}\mathbf{E}\left[a_k - a_{k+1}|\mathbf{z}_k\right] \\ + \frac{\eta_k}{2}\mathbf{E}\left[\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right\|^2|\mathbf{z}_k\right] - \frac{\alpha}{2}a_k \quad ,\end{aligned}$$

where $a_k = \|\mathbf{z}_k - \mathbf{x}^*\|^2$. Using the inequality $ab \leq a^2 + b^2$ implies

$$\begin{aligned}\mathbf{E}\left[f(\mathbf{z}_k) - f^*|\mathbf{z}_k\right] &\leq \frac{2}{\alpha}\left\|\mathbf{E}\left[\mathbf{g}_{k,\lambda}(\mathbf{z}_k) - \nabla f(\mathbf{z}_k)|\mathbf{z}_k\right]\right\|^2 + \frac{1}{2\eta_k}\mathbf{E}\left[a_k - a_{k+1}|\mathbf{z}_k\right] \\ &\quad + \frac{\eta_k}{2}\mathbf{E}\left[\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right\|^2|\mathbf{z}_k\right] - \frac{\alpha}{4}a_k \\ &\leq \frac{2}{\alpha}\mathbf{E}\left[\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k) - \nabla f(\mathbf{z}_k)\right\|^2|\mathbf{z}_k\right] + \frac{1}{2\eta_k}\mathbf{E}\left[a_k - a_{k+1}|\mathbf{z}_k\right] \\ &\quad + \frac{\eta_k}{2}\mathbf{E}\left[\left\|\mathbf{g}_{k,\lambda}(\mathbf{z}_k)\right\|^2|\mathbf{z}_k\right] - \frac{\alpha}{4}a_k \quad ,\end{aligned}$$

where the last inequality is due to Jensen's inequality. Taking total expectation from both sides of the above inequality and setting $r_k = \mathbf{E}\left[\|\mathbf{z}_k - \mathbf{x}^*\|^2\right]$ gives

$$\mathbf{E}\left[f(\mathbf{z}_k) - f^*\right] \le \frac{2}{\alpha}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\mathbf{g}_{k,\lambda}(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2\right] + \frac{1}{2\eta_k}\left(r_k - r_{k+1}\right)$$
$$+ \frac{\eta_k}{2}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|\mathbf{g}_{k,\lambda}(\mathbf{x})\right\|^2\right] - \frac{\alpha}{4}r_k \enspace ,$$

Substituting $\eta_k = \frac{2}{\alpha k}$ and using Lemmas 6.7.3 and 6.7.4 implies

$$\mathbf{E}\left[f(\mathbf{z}_k) - f^*\right] \le \frac{\alpha k}{4}\left(r_k - r_{k+1}\right) - \frac{\alpha}{4}r_k + \mathtt{A}_1\left(\frac{\log(k+1)}{k}\right)^{\frac{2(\beta-1)}{2\beta+d}}\alpha^{-1} \enspace ,$$

where $\mathtt{A} = 4\mathtt{A}_{\mathsf{error}} + \mathtt{A}_{\mathsf{sm}}$. Summing both sides form $1$ to $n$ yields

$$\sum_{k=1}^{n}\mathbf{E}\left[f(\mathbf{z}_k) - f^*\right] \le \mathtt{A}n^{\frac{2+d}{2\beta+d}}\log(n)^{\frac{2(\beta-1)}{2\beta+d}}\alpha^{-1} \enspace ,$$

where $\mathtt{A} = \frac{4\beta+2d}{2+d}\mathtt{A}_1$. In order to obtain the last inequality we used the fact that $\sum_{k=1}^{n}k^{-\frac{2(\beta-1)}{2\beta+d}} \le \frac{2\beta+d}{2+d}n^{\frac{2+d}{2\beta+d}}$, and $\log(n+1) \le 2\log(n)$ for $n \ge 2$. We conclude the proof by using the convexity of $f$. □

### Proof of Theorem 6.4.1

**Lemma 6.7.5.** *Under Assumption 6.2.2, 6.2.3, and 6.2.4, for any $\boldsymbol{x} \in \Theta$ we have*

$$\left|\mathbf{E}\left[f_n(\boldsymbol{x})\right] - f(\boldsymbol{x})\right| \le \mathtt{B}_{\mathsf{bias}}n^{-\frac{\beta}{2\beta+d}} \enspace ,$$

*where $\mathtt{B}_{\mathsf{bias}} > 0$ is a numerical constant.*

*Proof.* Let $B = \mathbf{E}[B_{m:n}(\mathbf{x})]$, and $\boldsymbol{\phi}_n = f_n(\mathbf{x}) - \boldsymbol{U}^\top(0)B^{-1}B_{m:n}(\mathbf{x})\boldsymbol{c}_n(f,\mathbf{x})$. It is straightforward to see that

$$\mathbf{E}\left[\boldsymbol{\phi}_n\right] = \mathbf{E}\left[f_n(\mathbf{x})\right] - \boldsymbol{U}^\top(0)\boldsymbol{c}_n(f,\mathbf{x}) = \mathbf{E}\left[f_n(\mathbf{x})\right] - f(\mathbf{x}) \enspace .$$

Therefore, we need to provide an upper bound for the term $\left|\mathbf{E}\left[\boldsymbol{\phi}_n\right]\right|$. Let

$$\boldsymbol{\psi}_{1,n} = \boldsymbol{U}^\top(0)B^{-1}\boldsymbol{C}_{m:n}(\mathbf{x}) \enspace ,$$
$$\boldsymbol{\psi}_{2,n} = \boldsymbol{U}^\top(0)\left(B + \lambda_n\mathbf{I}\right)^{-1}\boldsymbol{C}_{m:n}(\mathbf{x}) \enspace ,$$

where $\boldsymbol{C}_{m:n}(\mathbf{x}) = \frac{2}{nh_n^d}\sum_{k=m+1}^{n}\boldsymbol{R}_k(\mathbf{x})f(\mathbf{x}_k)$. Now, we can write

$$|\mathbf{E}[\boldsymbol{\phi}_n| \leq \underbrace{|\mathbf{E}[\boldsymbol{\psi}_{1,n} - h_n^{-1}\left(\boldsymbol{U}^\top(0)B^{-1}B_{m:n}(\mathbf{x})\boldsymbol{c}_n(f,\mathbf{x})\right)]|}_{\text{term I}} + \underbrace{|\mathbf{E}[\boldsymbol{\psi}_{2,n} - \boldsymbol{\psi}_{1,n}|}_{\text{term II}} + \underbrace{|\mathbf{E}[f_n(\mathbf{x}) - \boldsymbol{\psi}_{2,n}]|}_{\text{term III}} \ .$$

By following similar steps as in the proof of Lemma 6.7.1, we get

$$\text{term I} \leq \mathtt{B}_1 h_n^{\beta} \ , \quad \text{term II} \leq \mathtt{B}_2\lambda_n \quad \text{and} \quad \text{term III} \leq \mathtt{B}_3 h_n^{-\frac{d}{2}} n^{-\frac{1}{2}} \ ,$$

where $\mathtt{B}_1, \mathtt{B}_2, \mathtt{B}_3 > 0$, are numerical constants. Therefore, we deduce that

$$|\mathbf{E}[\boldsymbol{\phi}_n]| \leq \mathtt{B}_4\left(h_n^{\beta} + \lambda_n + h_n^{-\frac{d}{2}} n^{-\frac{1}{2}}\right) \ ,$$

with $\mathtt{B}_4 = \max(\mathtt{B}_1, \mathtt{B}_2, \mathtt{B}_3)$. We conclude the proof by substituting $h_n = n^{-\frac{1}{2\beta+d}}$, and $\lambda_n = n^{-\frac{\beta}{2\beta+d}}$. □

**Lemma 6.7.6.** *Let Assumptions 6.2.2, 6.2.3, and 6.2.4 hold. Then, for any $\boldsymbol{x} \in \Theta$ we have*

$$\mathbf{E}\left[\left(f_n(\boldsymbol{x}) - \mathbf{E}\left[f_n(\boldsymbol{x})\right]\right)^2\right] \leq B_{\mathit{var}} n^{-\frac{2\beta}{2\beta+d}} \ ,$$

*where $B_{\mathit{var}} > 0$ is a numerical constant.*

*Proof.* Similar to the proof of Lemma 6.7.4, let $\boldsymbol{G}_n(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{k=1}^{n}\boldsymbol{R}_k(\mathbf{x})\xi_k$, and $\boldsymbol{C}_n(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{k=1}^{n}\boldsymbol{R}_k(\mathbf{x})f(\mathbf{x}_k)$. Then, we have

$$\mathbf{E}[(f_n(\mathbf{x}) - \mathbf{E}\left[f_n(\mathbf{x})\right])^2] \leq \underbrace{2\mathbf{E}\left[\left\|B_{n,\lambda}^{-1}(\mathbf{x})\boldsymbol{G}_n(\mathbf{x})\right\|^2\right]}_{\text{term I}} + \underbrace{2\mathbf{E}\left[\left\|B_{n,\lambda}^{-1}(\mathbf{x})\boldsymbol{C}_n(\mathbf{x}) - \mathbf{E}\left[B_{n,\lambda}^{-1}(\mathbf{x})\boldsymbol{C}_n(\mathbf{x})\right]\right\|^2\right]}_{\text{term II}} \ .$$

For term I, we can write

$$\text{term I} \leq \underbrace{4\mathbf{E}\left[\left\|\left(B_{n,\lambda}^{-1}(\mathbf{x}) - (\mathbf{E}[B_{n,\lambda}])^{-1}\right)\boldsymbol{G}_n(\mathbf{x})\right\|^2\right]}_{\text{term III}} + \underbrace{4\left\|\mathbf{E}[B_{n,\lambda}(\mathbf{x})]\right\|^{-2}\mathbf{E}\left[\|\boldsymbol{G}_n(\mathbf{x})\|^2\right]}_{\text{term IV}} \ .$$

By Assumption 6.2.2, for term III we get

$$\text{term III} \leq 4\sigma^2\lambda_{\min}^{-2}\lambda_n^{-2}\mathbf{E}\left[\|B_{n,\lambda}(\mathbf{x}) - \mathbf{E}[B_{n,\lambda}(\mathbf{x})]\|_{\mathsf{op}}^2\left(n^{-2}h_n^{-2d}\sum_{k=1}^{n}\|\boldsymbol{R}_k(\mathbf{x})\|^2\right)\right] \ .$$

Now, by using the Cauchy-Schwarz inequality, we can write

$$\text{term III} \leq 4\sigma^2\lambda_{\min}^{-2}\lambda_n^{-2}\left(\mathbf{E}\left[\|B_{n,\lambda}(\mathbf{x}) - \mathbf{E}[B_{n,\lambda}(\mathbf{x})]\|_{\mathsf{op}}^4\right]\mathbf{E}\left[\left(n^{-2}h_n^{-2d}\sum_{k=1}^{n}\|\boldsymbol{R}_k(\mathbf{x})\|^2\right)^2\right]\right)^{\frac{1}{2}} \ .$$

Using Lemma 6.10 implies

$$\begin{aligned}\text{term III} &\leq 4\sigma^2\lambda_{\min}^{-2}\lambda_n^{-2}n^{-1}h_n^{-d}\left(\mathbf{E}\left[\left(n^{-2}h_n^{-2d}\sum_{k=1}^{n}\|\boldsymbol{R}_k(\mathbf{x})\|^2\right)^2\right]\right)^{\frac{1}{2}}\\ &\leq 4\sigma^2\lambda_{\min}^{-2}\lambda_n^{-2}n^{-1}h_n^{-d}\Bigg(n^{-4}h_n^{-4d}\bigg(\sum_{k=1}^{n}\mathbf{E}[\|\boldsymbol{R}_k(\mathbf{x})\|^4]+\sum_{j,k=1}^{n}\mathbf{E}[\|\boldsymbol{R}_j(\mathbf{x})\|^2]\mathbf{E}[\|\boldsymbol{R}_k(\mathbf{x})\|^2]\bigg)\Bigg)^{\frac{1}{2}} .\end{aligned}$$

By invoking Lemma 6.7.8(i) and the fact that $1 \leq nh_n$ we deduce that

$$\text{term III} \leq 4\sigma^2\lambda_{\min}^{-2}\lambda_n^{-2}n^{-2}h_n^{-2d}\left(p_{\max}\nu_{1,4}+(p_{\max}\nu_{1,2})^2\right)^{\frac{1}{2}} .$$

Now, by using the inequality $n \geq \lambda_n^{-2}h_n^{-d}$ we get

$$\text{term III} \leq \mathtt{B}_1 n^{-1}h_n^{-d} ,$$

where $\mathtt{B}_1 = 4\sigma^2\lambda_{\min}^{-2}(p_{\max}\nu_{1,4}+(p_{\max}\nu_{1,2})^2)^{\frac{1}{2}}$. For term IV, we can write

$$\text{term IV} \leq 4\sigma^2\lambda_{\min}^{-2}n^{-1}h_n^{-2d}\mathbf{E}\left[\|\boldsymbol{R}_n(\mathbf{x})\|^2\right] \leq \mathtt{B}_2 n^{-1}h_n^{-d} ,$$

where the last inequality is a result of Lemma 6.7.8(i) with $\mathtt{B}_2 = 4\sigma^2\lambda_{\min}^{-2}p_{\max}\nu_{1,2}$ as a numerical constant. For term II, we have

$$\begin{aligned}\text{term II} \leq{}& \underbrace{4\mathbf{E}\left[\left\|B_{n,\lambda}^{-1}(\mathbf{x})-(\mathbf{E}[B_{n,\lambda}(\mathbf{x})])^{-1}\right\|_{\mathsf{op}}^2\|\boldsymbol{C}_n(\mathbf{x})\|^2\right]}_{\text{term V}}\\ &+\underbrace{4\left\|(\mathbf{E}[B_{n,\lambda}(\mathbf{x})])^{-1}\right\|_{\mathsf{op}}^2\mathbf{E}\left[\|\boldsymbol{C}_n(\mathbf{x})-\mathbf{E}[\boldsymbol{C}_n(\mathbf{x})]\|^2\right]}_{\text{term VI}} .\end{aligned}$$

$$\text{term V} \leq 4M^2\lambda_{\min}^{-2}\lambda_n^{-2}k^{-1}h_n^{-d}\left(n^{-3}h_n^{-3d}p_{\max}\nu_{1,4}+n^{-2}h_n^{-2d}p_{\max}^2\nu_{1,2}^2\right)^{\frac{1}{2}} \leq \mathtt{B}_3 n^{-1}h_n^{-d} ,$$

where $\mathtt{B}_3 = 4n^2\lambda_{\min}^{-2}(p_{\max}\nu_{1,2}+(p_{\max}\nu_{1,2})^2)^{\frac{1}{2}}$. For term VI, by using Lemma 6.7.12, we have

$$\text{term VI} \leq \mathtt{B}_4 n^{-1}h_n^{-d} ,$$

with $\mathtt{B}_4 > 0$ as a numerical constant. Finally, by combing all of these bounds, we get

$$\mathbf{E}\left[(f_n(\mathbf{x})-\mathbf{E}[f_n(\mathbf{x})])^2\right] \leq \mathtt{B}_{\mathsf{var}} n^{-1}h_n^{-d} ,$$

where we introduced $\mathtt{B}_{\mathsf{var}} = \mathtt{B}_1+\mathtt{B}_2+\mathtt{B}_3+\mathtt{B}_4$. We conclude the proof by substituting $h_n = n^{-\frac{1}{2\beta+d}}$. □

*Proof of Theorem 6.4.1.* We have

$$\mathbf{E}\left[(f_n(\mathbf{x}) - f(\mathbf{x}))^2\right] = (\mathbf{E}\left[f_n(\mathbf{x})\right] - f(\mathbf{x}))^2 + \mathbf{E}\left[(f_n(\mathbf{x}) - \mathbf{E}\left[f_n(\mathbf{x})\right])^2\right] \ .$$

We conclude the proof by using Lemmas 6.7.5 and 6.7.6. □

## Proof of Theorem 6.5.1

We first prove (6.9). We apply the scheme of proving lower bounds for estimation of functionals described in Section 2.7.4 in Tsybakov (2009). Moreover, we use its basic form when the problem is reduced to testing two simple hypotheses (that is, the mixture measure $\mu$ from Section 2.7.4 in Tsybakov (2009) is the Dirac measure). The functional we are estimating is $F(f) = f^* = \min_{\mathbf{x}\in\Theta} f(\mathbf{x})$, where $\Theta$ is a sufficiently large Euclidean ball centered at 0. We choose the two hypotheses as the probability measures $P_1^{\otimes n}$ and $P_2^{\otimes n}$, where $P_j$ stands for the distribution of a pair $(\mathbf{x}_i, y_i)$ satisfying (6.1) with $f = f_j$, $j = 1, 2$. For $r > 0$, $\delta > 0$, we set

$$f_1(\mathbf{x}) = \alpha(1+\delta)\|\mathbf{x}\|^2/2, \quad f_2(\mathbf{x}) = f_1(\mathbf{x}) + rh_n^{\beta}\Phi\left(\frac{\mathbf{x} - \mathbf{x}^{(n)}}{h_n}\right),$$

where $h_n = n^{-1/(2\beta+d)}$, $\mathbf{x}^{(n)} = (h_n/8, 0, \ldots, 0) \in \mathbb{R}^d$ and $\Phi(\mathbf{x}) = \prod_{i=1}^d \Psi(x_i)$ with

$$\Psi(t) = \int_{-\infty}^{t} (\eta(y + 1/2) - \eta(y))\, \mathrm{d}y,$$

where $\eta(\cdot)$ is an infinitely many times differentiable function on $\mathbb{R}^1$ such that

$$\eta(x) \geq 0, \quad \eta(x) = \begin{cases} 0, & x \notin [0, 1/2] \\ 1, & x \in [1/8, 3/8] \end{cases} .$$

It is shown in Tsybakov (1990a) that if $r$ is small enough the functions $f_1$ and $f_2$ are $\alpha$-strongly convex and belong to $\mathcal{F}_\beta(L)$. Thus, $f_j \in \mathcal{F}_{\beta,\alpha}(L), j = 1, 2$. It is also not hard to check (cf. Tsybakov (1990a)) that for the function $\eta_1(y) = \eta(y + 1/2) - \eta(y)$ we have

$$\eta_1\left(-\frac{r\Psi^{d-1}(0)h_n^{\beta-2}}{\alpha(1+\delta)} - \frac{1}{8}\right) = 1$$

when $r < \alpha(1+\delta)/4$. Using this remark we get that the minimizers $\mathbf{x}_j^* = \arg\min_{\mathbf{x}\in\Theta} f_j(\mathbf{x})$ have the form

$$\mathbf{x}_1^* = (0, 0, \ldots, 0) \quad \text{and} \quad \mathbf{x}_2^* = \left(-\frac{r\Psi^{d-1}(0)h_n^{\beta-1}}{\alpha(1+\delta)}, 0, \ldots, 0\right).$$

The values of the functional $F$ on $f_1$ and $f_2$ are $F(f_1) = 0$ and

$$\begin{aligned}
F(f_2) &= f_2(\mathbf{x}_2^*) \\
&= \frac{r^2\Psi^{2(d-1)}(0)}{2\alpha(1+\delta)} h_n^{2(\beta-1)} + r\Psi^{d-1}(0)\Psi\left(-\frac{r\Psi^{d-1}(0)h_n^{\beta-2}}{\alpha(1+\delta)} - \frac{1}{8}\right) h_n^{\beta} \\
&\geq \frac{r^2\Psi^{2(d-1)}(0)}{2\alpha(1+\delta)} h_n^{2(\beta-1)} + r\Psi^{d-1}(0)\Psi(-1/4)h_n^{\beta} \quad \text{(for } r \text{ small enough)} \\
&\geq r\Psi^{d-1}(0)\Psi(-1/4)h_n^{\beta}.
\end{aligned}$$

Here, $\Psi(0) = \int_{-\infty}^{\infty} \eta(y)\,\mathrm{d}y > 0$ and $\Psi(-1/4) = \int_{-\infty}^{1/4} \eta(y)\,\mathrm{d}y > 0$.

Note that assumption (i) of Theorem 2.14 in Tsybakov (2009) is satisfied with $\beta_0 = \beta_1 = 0$, $c = 0$ and $s = r\Psi^{d-1}(0)\Psi(-1/4)h_n^{\beta}/2$. Therefore, by Theorem 2.15 (ii) in Tsybakov (2009), (6.9) will be proved if we show that

$$\mathrm{H}^2\left(P_1^{\otimes n}, P_2^{\otimes n}\right) \leq a < 2, \tag{6.16}$$

where $\mathrm{H}^2(P,Q)$ denotes the Hellinger distance between the probability measures $P$ and $Q$. Using assumption (6.7) we obtain

$$\begin{aligned}
\mathrm{H}^2\left(P_1^{\otimes n}, P_2^{\otimes n}\right) &= 2\left(1 - \left(1 - \frac{\mathrm{H}^2(P_1,P_2)}{2}\right)^n\right) \\
&\leq n\,\mathrm{H}^2(P_1,P_2) \quad (\text{as } (1-x)^n \geq 1 - xn,\ x \in [0,1]) \\
&= n\int\left(\sqrt{p_\xi(y)} - \sqrt{p_\xi\left(y + (f_1(\mathbf{x}) - f_2(\mathbf{x}))\right)}\right)^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}y \\
&\leq nI_* \int \left(f_1(\mathbf{x}) - f_2(\mathbf{x})\right)^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= nI_* r^2 h_n^{2\beta+d} \int \Phi^2(\boldsymbol{u}) p\left(\mathbf{x}^{(n)} + \boldsymbol{u}h_n\right)\,\mathrm{d}\boldsymbol{u} \\
&\leq p_{\max} I_* r^2 \int \Phi^2(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}, \quad \text{for } r \leq v_0,
\end{aligned}$$

where $p_{\max}$ is the maximal value of the density $p(\cdot)$ of $\mathbf{x}_i$. Choosing $r \leq \sqrt{a/\left(p_{\max} I_* \int \Phi^2(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\right)}$, with $a < 2$ we obtain (6.16). This completes the proof of (6.9).

In order to prove (6.8), it suffices to use the same construction of two hypotheses as above, apply the Hellinger version of Theorem 2.2 from Tsybakov (2009), and to notice that $\|\mathbf{x}_1^* - \mathbf{x}_2^*\| \geq cn^{-(\beta-1)/(2\beta+d)}$, where $c > 0$ is a constant.

### Proof of Theorem 6.5.2

We apply again the scheme of proving lower bounds for estimation of functionals from Section 2.7.4 in Tsybakov (2009). However, we use a different construction of the hypotheses. Without loss of generality, assume that $n \geq 2$, that $\Theta$ contains the cube $[0,1]^d$. Define

$h_n = (n/\log(n))^{-1/(2\beta+d)}$, $N = (1/h_n)^d$, and assume without loss of generality that $N$ is an integer. For $r > 0$, we set

$$f_j(\mathbf{x}) = -rh_n^{\beta}\Phi\left(\frac{\mathbf{x} - \mathbf{t}^{(j)}}{h_n}\right), \quad j = 1, \dots, N,$$

where $\Phi(\mathbf{x}) = \prod_{i=1}^{d} \Psi(x_i)$, where $\Psi(\cdot)$ is an infinitely many times differentiable function on $\mathbb{R}$ taking positive values on its support $[-1/2, 1/2]$, and we denote by $\mathbf{t}^{(1)}, \dots, \mathbf{t}^{(N)}$ the $N$ points of the equispaced grid on $[0,1]^d$ with step $h_n$ over each coordinate, such that the supports of all $f_j$'s are included in $[0,1]^d$ and are disjoint. It is not hard to check that for $r$ small enough all the functions $f_j$, $j = 1, \dots, N$, belong to $\mathcal{F}_\beta(L)$.

We consider the product probability measures $P_0^{\otimes n}$ and $P_1^{\otimes n}, \dots P_N^{\otimes n}$, where $P_0$ stands for the distribution of a pair $(\mathbf{x}_i, y_i)$ satisfying (6.1) with $f \equiv 0$, and $P_j$ stands for the distribution of $(\mathbf{x}_i, y_i)$ satisfying (6.1) with $f = f_j$. Consider the mixture probability measure $\mathbb{P}_\mu = \frac{1}{N}\sum_{j=1}^{N} P_j^{\otimes n}$, where $\mu$ denotes the uniform distribution on $\{1, \dots, N\}$.

Note that, for each $j = 1, \dots, N$, we have $F(f_j) = -rh_n^\beta \Phi_{\max}$, where $F(f) = f^* = \min_{\mathbf{x}\in\Theta} f(\mathbf{x})$, and $\Phi_{\max} > 0$ denotes the maximal value of function $\Phi(\cdot)$. Let

$$\chi^2(P', P) = \int (\mathrm{d}P'/\,\mathrm{d}P)^2\,\mathrm{d}P - 1$$

denote the chi-square divergence between two mutually absolutely continuous probability measures $P'$ and $P$. We will use the following lemma, which is a special case of Theorem 2.15 in Tsybakov (2009).

**Lemma 6.7.7.** *Assume that there exist $v > 0, b > 0$ such that $F(f_j) = -2v$ for $j = 1, \dots, N$ and $\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \le b$, Then*

$$\inf_{\hat{f}_n} \sup_{j=0,1,\dots,N} P_j^{\otimes n}\big(|\hat{f}_n - F(f_j)| \ge v\big) \ge \frac{1}{4}\exp(-b),$$

*where $\inf_{\hat{f}_n}$ denotes the infimum over all estimators.*

In our case, the first condition of this lemma is satisfied with $v = rh_n^\beta \Phi_{\max}/2$. We now check that the second condition $\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \le b$ holds with some constant $b > 0$ independent of $n$. Using a standard representation of the chi-square divergence of a Gaussian mixture from the pure Gaussian noise measure (see, for example, Lemma 8 in Carpentier et al. (2019)) we obtain

$$\begin{aligned}
\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) &= \frac{1}{N^2}\sum_{j,j'=1}^{N} \mathbf{E}\exp\left(\frac{\sum_{i=1}^n f_j(\mathbf{x}_i) f_{j'}(\mathbf{x}_i)}{\sigma^2}\right) - 1 \\
&= \frac{1}{N^2}\sum_{j,j'=1}^{N} \mathbf{E}\exp\left(\frac{\sum_{i=1}^n f_j(\mathbf{x}_i) f_{j'}(\mathbf{x}_i)}{\sigma^2}\right) - 1 \\
&= \frac{1}{N^2}\sum_{j=1}^{N} \mathbf{E}\exp\left(\frac{\sum_{i=1}^n f_j^2(\mathbf{x}_i)}{\sigma^2}\right) + \frac{N(N-1)}{N^2} - 1 \\
&\le \frac{1}{N^2}\sum_{j=1}^{N} \mathbf{E}\exp\left(\frac{\sum_{i=1}^n f_j^2(\mathbf{x}_i)}{\sigma^2}\right) \\
&= \frac{1}{N^2}\sum_{j=1}^{N} \left[\mathbf{E}\exp\left(\frac{f_j^2(\mathbf{x}_1)}{\sigma^2}\right)\right]^n,
\end{aligned}$$

where the equality in the third line is due to the fact that if $j \neq j'$ then $f_j$ and $f_{j'}$ have disjoint supports and thus $f_j(\mathbf{x}_i) f_{j'}(\mathbf{x}_i) = 0$. Note that $\max_{\mathbf{x}\in\mathbb{R}^d} f_j^2(\mathbf{x}) \le r^2 \Phi_{\max}^2$, for all $j = 1, \dots, N$. Choose $r$ such that $r \le \sigma/\Phi_{\max}$. Then $\frac{f_j^2(\mathbf{x}_1)}{\sigma^2} \le 1$, and using the elementary inequality $\exp(u) \le 1 + 2u, u \in [0,1]$, we obtain that $\exp\left(\frac{f_j^2(\mathbf{x}_1)}{\sigma^2}\right) \le 1 + \frac{2 f_j^2(\mathbf{x}_1)}{\sigma^2}$ for all $j = 1, \dots, N$. Substituting this bound in the last display and noticing that $\mathbf{E}(f_j^2(\mathbf{x}_1)) = \int f_j^2(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \le p_{\max} r^2 h_n^{2\beta+d} \int \Phi^2(\mathbf{x})\, \mathrm{d}\mathbf{x} = c_* \frac{\log n}{n}$, where $c_* = p_{\max} r^2 \int \Phi^2(\mathbf{x})\, \mathrm{d}\mathbf{x}$, we obtain:

$$\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \le \frac{1}{N}\left[1 + \frac{2\mathbf{E}(f_j^2(\mathbf{x}_1))}{\sigma^2}\right]^n \le \frac{1}{N}\left[1 + \frac{2c_* \log n}{\sigma^2 n}\right]^n \le \frac{1}{N}\exp\left(\frac{2c_* \log n}{\sigma^2}\right) = \frac{n^{c_0}}{N},$$

where $c_0 = 2c_*/\sigma^2 = 2p_{\max} r^2 \int \Phi^2(\mathbf{x})\, \mathrm{d}\mathbf{x}/\sigma^2$. Since $N = (n/\log n)^{\frac{d}{2\beta+d}}$ we finally get

$$\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \le n^{c_0 - \frac{d}{2\beta+d}} (\log n)^{\frac{d}{2\beta+d}}.$$

By choosing $r$ small enough to have $c_0 \le \frac{d}{2(2\beta+d)}$ we obtain that $\chi^2(\mathbb{P}_\mu, P_0^{\otimes n}) \le \left(\frac{\log n}{\sqrt{n}}\right)^{\frac{d}{2\beta+d}} \le \left(\frac{\log 2}{\sqrt{2}}\right)^{\frac{d}{2\beta+d}} := b$. Thus, the second condition of Lemma 6.7.7 holds if $r$ is chosen as a small enough constant. Notice that, in Lemma 6.7.7, the rate $v$ is of the desired order $(n/\log n)^{-\frac{\beta}{2\beta+d}}$. The result of the theorem now follows from Lemma 6.7.7 and the standard argument to obtain the lower bounds, see Section 2.7.4 in Tsybakov (2009).

## Proofs of auxiliary lemmas

Recall that $\Theta' = \{\mathbf{x} + \mathbf{y} : \mathbf{x} \in \Theta \quad \text{and} \quad \|\mathbf{y}\| \le 1\} \supseteq \{\mathbf{x} + \mathbf{y} : \mathbf{x} \in \Theta \quad \text{and} \quad \mathbf{y} \in \mathrm{Supp}(K)\}$.

**Lemma 6.7.8.** *For any $q \geq 1$, let*

$$\nu_{1,q} = \int_{\mathbb{R}^d} \|\boldsymbol{U}(\boldsymbol{u})K(\boldsymbol{u})\|^q \,\mathrm{d}\boldsymbol{u} \ , \qquad \nu_{2,q} = \int_{\mathbb{R}^d} \left\|\boldsymbol{U}\left(\boldsymbol{u}\right)\boldsymbol{U}^\top\left(\boldsymbol{u}\right)K\left(\boldsymbol{u}\right)\right\|_{op}^q \mathrm{d}\boldsymbol{u} \ ,$$

*and $p_{\max} = \max_{\boldsymbol{y} \in \Theta'} p(\boldsymbol{y})$. Then, under Assumption 6.2.4, for any $\boldsymbol{x} \in \Theta$, $k \in [n]$, and $i \in [k]$, we have*

*(i)* $h_k^{-d}\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\|\boldsymbol{R}_{i,k}(\boldsymbol{x})\|^q\right] \leq p_{\max}\nu_{1,q}$ .

*(ii)* $h_k^{-d}\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\|M_{i,k}(\boldsymbol{x})\|_{op}^q\right] \leq p_{\max}\nu_{2,q}$ .

*(iii) There exists $\lambda_{\min} > 0$, such that $\inf_{\boldsymbol{x}\in\Theta}\lambda_{\min}\left(\mathbf{E}\left[B_k(\boldsymbol{x})\right]\right) \geq \lambda_{\min}$.*

*Proof.* We have

$$\begin{aligned} h_k^{-d}\mathbf{E}\left[\left\|\sup_{\mathbf{x}\in\Theta}\boldsymbol{R}_{i,k}(\mathbf{x})\right\|^q\right] &= h_k^{-d}\int_{\mathbb{R}^d}\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{U}\left(\frac{\mathbf{y}-\mathbf{x}}{h_k}\right)K\left(\frac{\mathbf{y}-\mathbf{x}}{h_k}\right)\right\|^q p(\mathbf{y})\,\mathrm{d}\mathbf{y} \\ &= \int_{\mathbb{R}^d}\|\boldsymbol{U}(\boldsymbol{u})K(\boldsymbol{u})\|^q\sup_{\mathbf{x}\in\Theta}p(\mathbf{x}+h_k\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \leq p_{\max}\nu_{1,q} \ . \end{aligned}$$

For (ii) we can write

$$\begin{aligned} h_k^{-d}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|M_{i,k}(\mathbf{x})\|_{\mathrm{op}}^q\right] &= h_k^{-d}\int_{\mathbb{R}^d}\sup_{\mathbf{x}\in\Theta}\left\|\boldsymbol{U}\left(\frac{\mathbf{y}-\mathbf{x}}{h_k}\right)\boldsymbol{U}^\top\left(\frac{\mathbf{y}-\mathbf{x}}{h_k}\right)K\left(\frac{\mathbf{y}-\mathbf{x}}{h_k}\right)\right\|_{\mathrm{op}}^q p(\mathbf{y})\,\mathrm{d}\mathbf{y} \\ &= \int_{\mathbb{R}^d}\left\|\boldsymbol{U}\left(\boldsymbol{u}\right)\boldsymbol{U}^\top\left(\boldsymbol{u}\right)K\left(\boldsymbol{u}\right)\right\|_{\mathrm{op}}^q\sup_{\mathbf{x}\in\Theta}p(\mathbf{x}+h_k\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \leq p_{\max}\nu_{2,q} \ . \end{aligned}$$

Similarly, for (iii) we get

$$\begin{aligned} \mathbf{E}\left[B_k(\mathbf{x})\right] &= h_k^{-d}\mathbf{E}\left[\boldsymbol{U}\left(\frac{\mathbf{x}_1-\mathbf{x}}{h_k}\right)\boldsymbol{U}^\top\left(\frac{\mathbf{x}_1-\mathbf{x}}{h_k}\right)K\left(\frac{\mathbf{x}_1-\mathbf{x}}{h_k}\right)\right] \\ &= \int_{\mathbb{R}^d}\boldsymbol{U}\left(\boldsymbol{u}\right)\boldsymbol{U}^\top\left(\boldsymbol{u}\right)K\left(\boldsymbol{u}\right)p(\mathbf{x}+h_k\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \ . \end{aligned}$$

Introducing the notation $H = \int_{\mathbb{R}^d}\boldsymbol{U}\left(\boldsymbol{u}\right)\boldsymbol{U}^\top\left(\boldsymbol{u}\right)K\left(\boldsymbol{u}\right)$ we deduce that $\inf_{\mathbf{x}\in\Theta}\lambda_{\min}\left(\mathbf{E}\left[B_k(\mathbf{x})\right]\right) \geq p_{\min}\lambda_{\min}\left(H\right)$. By (Tsybakov, 1986, Lemma 1), we have $\lambda_{\min}\left(H\right) > 0$. We conclude the proof by letting $\lambda_{\min} = p_{\min}\lambda_{\min}\left(H\right)$. □

**Lemma 6.7.9.** *Let $k \in [n]$, and $h_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{1}{2\beta+d}}$. Let Assumption 6.2.4 hold. Then, for any $\boldsymbol{x} \in \Theta$, we have*

$$\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\|B_{k,\lambda}(\boldsymbol{x}) - \mathbf{E}\left[B_{k,\lambda}(\boldsymbol{x})\right]\|_{op}^4\right] \leq A_1 h_k^{-2d}k^{-2}\log(k+1)^2 \ . \tag{6.17}$$

*Furthermore, for $k \geq \lambda_k^{-2}h_k^{-d}\log(k+1)$, we have $\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\|B_{k,\lambda}(\boldsymbol{x})\|_{op}^{-4}\right] \leq A_2\lambda_{\min}^{-4}$ , where $A_1, A_2 > 0$ are numerical constants.*

*Proof.* In order to prove (6.17), we first show that $\left\|B_{k,\lambda}(\mathbf{x}) - \mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\right\|_{\text{op}}$ is bounded from above by a Lipschitz function. Let $Q_{i,k}(\mathbf{x}) = h_k^{-d} M_{i,k}(\mathbf{x}) - h_k^{-d}\mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]$.

**1. Proving a Lipschitz upper bound:** First note that

$$\left\|B_{k,\lambda}(\mathbf{x}) - \mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\right\|_{\text{op}} \leq \sum_{s=1}^{S} \left| k^{-1} h_k^{-d} \sum_{i=1}^{k} \left(Q_{i,k}(\mathbf{x}) - \mathbf{E}\left[Q_{i,k}(\mathbf{x})\right]\right)_s \right| \quad ,$$

where $\left(Q_{i,k}(\mathbf{x}) - \mathbf{E}\left[Q_{i,k}(\mathbf{x})\right]\right)_s$ is the $(s,s)$-entry of the matrix $Q_{i,k}(\mathbf{x}) - \mathbf{E}\left[Q_{i,k}(\mathbf{x})\right]$. Recall that the kernel function $K$ is $L_K$-Lipschitz. Furthermore, for $s \in [S]$ let $G^{(s)} : \mathbb{R}^d \to \mathbb{R}$, such that

$$G^{(s)}(\boldsymbol{u}) = \Big(\boldsymbol{U}\left(\boldsymbol{u}\right)\boldsymbol{U}^{\top}\left(\boldsymbol{u}\right)\Big)_s \quad ,$$

it is straightforward to check that $G^{(s)}$ is a continuously differentiable function. Let $\Omega$ be a convex and compact subset of $\mathbb{R}^d$, such that $\mathrm{Supp}(K) \subseteq \Omega$, and let $L_G^{(s)} = \max_{\boldsymbol{u}\in\Omega} \left\|\nabla G^{(s)}(\boldsymbol{u})\right\|$, and $L_G = \max_{s\in[S]} L_s$. Now, it is clear to see that for any $s \in [S]$, $G^{(s)}$ is a $L_G$-Lipschitz function on $\mathrm{Supp}(K)$. Moreover, for any $s \in [S]$, and $\mathbf{x}, \mathbf{y} \in \Theta$, we can write

$$\begin{aligned}\left| k^{-1} h_k^{-d} \sum_{i=1}^{k} \left( (Q_{i,k}(\mathbf{x}) - \mathbf{E}[Q_{i,k}(\mathbf{x})])_s - (Q_{i,k}(\mathbf{y}) - \mathbf{E}\left[Q_{i,k}(\mathbf{y})\right])_s \right) \right| &\leq k^{-1} h_k^{-d} \underbrace{\sum_{i=1}^{k} \left|(Q_{i,k}(\mathbf{x}))_s - (Q_{i,k}(\mathbf{y}))_s\right|}_{\text{term I}} \\ &+ k^{-1} h_k^{-d} \underbrace{\sum_{i=1}^{k} \mathbf{E}\left[\left|(Q_{i,k}(\mathbf{x}))_s - (Q_{i,k}(\mathbf{y}))_s\right|\right]}_{\text{term II}} \quad .\end{aligned}$$

For term I if $h_k^{-1}(\mathbf{x}_i - \mathbf{x}), h_k^{-1}(\mathbf{x}_i - \mathbf{y}) \in \mathrm{Supp}(K)$, we have

$$\begin{aligned}\text{term I} &= \sum_{i=1}^{k} \left| G^{(s)}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) - G^{(s)}\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_k}\right) K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_k}\right) \right| \\ &= \sum_{i=1}^{k} \left| \left( G^{(s)}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) - G^{(s)}\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_k}\right) \right) K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) \right. \\ &\qquad \left. + \left( K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) - K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_k}\right) \right) G^{(s)}\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_k}\right) \right| \\ &\leq k h_k^{-1} \mathbb{A}_3 \left\|\mathbf{x} - \mathbf{y}\right\| \quad ,\end{aligned}$$

where $\mathbb{A}_3 = \max_{\boldsymbol{u}\in\mathrm{Supp}(K)} L_G K(\boldsymbol{u}) + \max_{s\in[S], i\in[k], \boldsymbol{u}\in\mathrm{Supp}(K)} L_K G^s(\boldsymbol{u})$. The scenarios when either one or both of the points $h_k^{-1}(\mathbf{x}_i - \mathbf{x}), h_k^{-1}(\mathbf{x}_i - \mathbf{y})$ do not belong to $\mathrm{Supp}(K)$, can be treated similarly. For term II, with exactly the same steps, we can write

$$\text{term II} \leq k h_k^{-1} \mathbb{A}_3 \left\|\mathbf{x} - \mathbf{y}\right\| \quad .$$

By combining all these bounds we deduce that

$$\sum_{s=1}^{S}\left|k^{-1}h_k^{-d}\sum_{i=1}^{k}\left(\left(Q_{i,k}(\mathbf{x})-\mathbf{E}\left[Q_{i,k}(\mathbf{x})\right]\right)_s-\left(Q_{i,k}(\mathbf{y})+\mathbf{E}\left[Q_{i,k}(\mathbf{y})\right]\right)_s\right)\right|\leq \mathbb{A}_{\mathrm{Lip}}h_k^{-d-1}\left\|\mathbf{x}-\mathbf{y}\right\| \quad ,$$

where $\mathbb{A}_{\mathrm{Lip}}=2S\mathbb{A}_3$.

**2. Providing an upper bound for the probability:** For any $t\geq 0$, we can write

$$\begin{aligned}\mathbf{P}\left[\sup_{\mathbf{x}\in\Theta}\left\|B_{k,\lambda}(\mathbf{x})-\mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\right\|_{\mathsf{op}}\geq t\right] &\leq \mathbf{P}\left[\sum_{s=1}^{S}\sup_{\mathbf{x}\in\Theta}\left|k^{-1}h_k^{-d}\sum_{i=1}^{k}F_i^{(s)}(\mathbf{x})\right|\geq t\right] \\ &\leq \underbrace{\sum_{s=1}^{S}\mathbf{P}\left[\sup_{\mathbf{x}\in\Theta}\left|k^{-1}h_k^{-d}\sum_{i=1}^{k}F_i^{(s)}(\mathbf{x})\right|\geq \frac{t}{S}\right]}_{\text{term III}} \quad ,\end{aligned} \tag{6.18}$$

where we defined $F_i^{(s)}(\mathbf{x})=\left(Q_{i,k}(\mathbf{x})-\mathbf{E}\left[Q_{i,k}(\mathbf{x})\right]\right)_s$. From now on, we focus on providing an upper bound for term III. For $\epsilon>0$, consider an $\epsilon$-net of $\Theta$, namely $\mathcal{N}$, with cardinality $\mathcal{N}\left(\Theta,\epsilon\right)$. Therefore, for any $\mathbf{x}\in\Theta$, there exists $\mathbf{y}\in\mathcal{N}$, such that $\|\mathbf{x}-\mathbf{y}\|<\epsilon$, and we can write

$$\begin{aligned}\text{term III} &\leq \sum_{s=1}^{S}\mathcal{N}\left(\Theta,\epsilon\right)\sup_{\mathbf{x}\in\mathcal{N}}\mathbf{P}\left[\left|k^{-1}h_k^{-d}\sum_{i=1}^{k}F_i^{(s)}(\mathbf{x})\right|\geq\frac{t}{S}-\mathbb{A}_{\mathrm{Lip}}h_k^{-d-1}\epsilon\right] \\ &\leq \sum_{s=1}^{S}\left(\frac{\mathsf{diam}(\Theta)}{\epsilon}+1\right)^{d}\sup_{\mathbf{x}\in\mathcal{N}}\mathbf{P}\left[\left|k^{-1}h_k^{-d}\sum_{i=1}^{k}F_i^{(s)}(\mathbf{x})\right|\geq\frac{t}{S}-\mathbb{A}_{\mathrm{Lip}}h_k^{-d-1}\epsilon\right] \quad .\end{aligned}$$

where $\mathsf{diam}(\Theta)=\max_{\mathbf{x},\mathbf{y}\in\Theta}\|\mathbf{x}-\mathbf{y}\|$, and we used the fact that $\mathcal{N}\left(\Theta,\epsilon\right)\leq\left(\frac{\mathsf{diam}(\Theta)}{\epsilon}+1\right)^{d}$. By assigning $\epsilon=\frac{t}{2\mathbb{A}_{\mathrm{Lip}}S}h_k^{d+1}$, we get

$$\text{term III}\leq\sum_{s=1}^{S}\left(\frac{2\mathbb{A}_{\mathrm{Lip}}S\mathsf{diam}(\Theta)}{t}\cdot h_k^{-d-1}+1\right)^{d}\sup_{\mathbf{x}\in\mathcal{N}}\mathbf{P}\left[\left|k^{-1}h_k^{-d}\sum_{i=1}^{k}F_i^{(s)}(\mathbf{x})\right|\geq\frac{t}{2S}\right] \quad .$$

Invoking Bernstein's inequality we deduce that

$$\text{term III}\leq\sum_{s=1}^{S}\left(\frac{2\mathbb{A}_{\mathrm{Lip}}S\mathsf{diam}(\Theta)}{t}\cdot h_k^{-d-1}+1\right)^{d}\mathbf{P}\left[-\frac{1}{2}\cdot\min\left(\frac{kt^2}{S^2v^2},\frac{kt}{S\omega}\right)\right] \quad , \tag{6.19}$$

where

$$v^2=\sup_{\mathbf{x}\in\mathcal{N},s\in[S]}\mathbf{E}\left[\left|h_k^{-d}F_1^{(s)}(\mathbf{x})\right|^2\right] \quad , \quad \text{and} \quad \omega=\sup_{\mathbf{x}\in\mathcal{N},s\in[S],i\in[k]}h_k^{-d}\left|\left(Q_{i,k}(\mathbf{x})-\mathbf{E}\left[Q_{i,k}(\mathbf{x})\right]\right)_s\right| \quad .$$

We continue the proof by providing upper bounds for the terms $\upsilon$ and $\omega$. For $\upsilon$, we can write

$$\begin{aligned}\upsilon^2 &= \sup_{\mathbf{x}\in\mathcal{N}, s\in[S]} h_k^{-2d}\mathbf{E}\left[\left|\left(Q_{1,k}(\mathbf{x}) - \mathbf{E}\left[Q_{1,k}(\mathbf{x})\right]\right)_s\right|^2\right] \\ &\leq \sup_{\mathbf{x}\in\mathcal{N}, s\in[S]} h_k^{-2d}\mathbf{E}\left[\left|\left(Q_{1,k}(\mathbf{x})\right)_s\right|^2\right] \\ &\leq \sup_{\mathbf{x}\in\mathcal{N}, s\in[S]} h_k^{-d}\int\left|\left(\boldsymbol{U}(\boldsymbol{u})\,\boldsymbol{U}^\top(\boldsymbol{u})\,K(\boldsymbol{u})\right)_s\right|^2 p(\mathbf{x}+h_k\boldsymbol{u}) \leq h_k^{-d}\mathtt{A}_4 \ ,\end{aligned}$$

where $\mathtt{A}_4 = p_{\max}\sup_{s\in[S]}\int\left|\left(\boldsymbol{U}(\boldsymbol{u})\,\boldsymbol{U}^\top(\boldsymbol{u})\,K(\boldsymbol{u})\right)_s\right|^2 \mathrm{d}\boldsymbol{u}$. Similarly, for $\omega$, we have

$$\omega \leq \sup_{\mathbf{x}\in\mathcal{N}, s\in[S], i\in[k]} h_k^{-d}\left(\left|\left(Q_{i,k}(\mathbf{x})\right)_s\right| + \mathbf{E}\left[\left|\left(Q_{i,k}(\mathbf{x})\right)_s\right|\right]\right) \leq h_k^{-d}\mathtt{A}_5 \ ,$$

where $\mathtt{A}_5 = 2\sup_{\boldsymbol{u}\in\mathrm{Supp}(K), s\in[S]}\kappa_{\max}\left|\left(\boldsymbol{U}(\boldsymbol{u})\,\boldsymbol{U}^\top(\boldsymbol{u})\right)_s\right|$. By substituting these bounds in (6.19), we get

$$\begin{aligned}\text{term III} &\leq \sum_{s=1}^{S}\left(\frac{\mathtt{A}_7}{t}h_k^{-d-1}+1\right)^d \exp\left(-\mathtt{A}_6\cdot\min\left(kt^2h_k^d, kth_k^d\right)\right) \qquad (6.20)\\ &= S\exp\left(-\mathtt{A}_6\cdot\min\left(kt^2h_k^d, kth_k^d\right) + d\log\left(\frac{\mathtt{A}_7}{t}h_k^{-d-1}+1\right)\right) \ ,\end{aligned}$$

where $\mathtt{A}_6 = \min\left(\frac{1}{2S^2\mathtt{A}_2}, \frac{1}{2S\mathtt{A}_3}\right)$, and $\mathtt{A}_7 = 2\mathtt{A}_{\mathrm{Lip}}S\mathrm{diam}(\Theta)$. Finlay, by replacing (6.20) in (6.18), we get

$$\mathbf{P}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x}) - \mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{\mathrm{op}} \geq t\right] \leq S\exp\left(-\mathtt{A}_6\cdot\min\left(kt^2h_k^d, kth_k^d\right) + d\log\left(\frac{\mathtt{A}_7}{t}h_k^{-d-1}+1\right)\right) \ .$$

**3. Getting the final upper bound:** For any $a\geq 0$, we can write

$$\begin{aligned}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x}) - \mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{\mathrm{op}}^4\right] &= \int_{t=0}^{\infty}4t^3\mathbf{P}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x}) - \mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{\mathrm{op}} \geq t\right]\mathrm{d}t \\ &\leq a^4 + S\int_{t=a}^{\infty}4t^3\exp\left(-\underbrace{\mathtt{A}_6\cdot\min\left(kt^2h_k^d, kth_k^d\right)}_{\text{term IV}} + \underbrace{d\log\left(\frac{\mathtt{A}_8}{t}h_k^{-d-1}+1\right)}_{\text{V}}\right)\mathrm{d}t \ ,\end{aligned}$$

where $\mathtt{A}_8 = \max(\mathtt{A}_7, 1, \mathtt{A}_6^{-2})$. Note that term IV is an increasing and term V is a decreasing function of $t$. Now we wish to assign $a$ large enough to ensure that term IV dominates term V. Let

$$a = 4d\frac{2\beta+d}{\beta}\sqrt[3]{\frac{\mathtt{A}_8}{\mathtt{A}_6}}\sqrt{\log(k+1)}h_k^{-\frac{d}{2}}k^{-\frac{1}{2}} \ .$$

We have two possibilities. First assume that $a<1$, then we have

$$\text{term IV} = \mathtt{A}_6ka^2h_k^d \geq 16d^2\mathtt{A}_6^{\frac{1}{3}}\mathtt{A}_8^{\frac{2}{3}}\log(k+1) \ .$$

Since $\mathtt{A}_8 \geq 1$, we have

$$\text{term V} \leq d \log\left(\frac{2\mathtt{A}_8}{a} h_k^{-d-1}\right) \leq d \log\left(\frac{\mathtt{A}_6^{\frac{1}{3}} \mathtt{A}_8^{\frac{2}{3}}}{2d} \left(\frac{k}{\log(k+1)}\right)^{\frac{\beta+d+1}{2\beta+d}}\right) \leq d \log\left(\frac{\mathtt{A}_6^{\frac{1}{3}} \mathtt{A}_8^{\frac{2}{3}}}{d} \frac{k}{2\log(k+1)}\right) \ ,$$

where the last inequality is obtained from the fact that $k/\log(k+1) \geq 1$, and $\frac{\beta+d+1}{2\beta+d} \leq 1$. Since $2\log(k+1) \geq 1$, we deduce that

$$\text{term V} \leq d \log\left(\frac{\mathtt{A}_6^{\frac{1}{3}} \mathtt{A}_8^{\frac{2}{3}}}{d} k\right) \ ,$$

which implies

$$2 \cdot \text{term V} \leq \text{term IV} \ .$$

Now, assume that $a \geq 1$. Then we have

$$\begin{aligned} \text{term IV} = \mathtt{A}_6 k a h_k^d &\geq 4d \frac{2\beta+d}{\beta} \mathtt{A}_6^{\frac{2}{3}} \mathtt{A}_8^{\frac{1}{3}} \sqrt{\log(k+1)} h_k^{\frac{d}{2}} k^{\frac{1}{2}} \\ &\geq 4d \frac{2\beta+d}{\beta} \mathtt{A}_6^{\frac{2}{3}} \mathtt{A}_8^{\frac{1}{3}} k^{\frac{\beta}{2\beta+d}} \ , \end{aligned}$$

and

$$\begin{aligned} \text{term V} &\leq d \log\left(\frac{\mathtt{A}_6^{\frac{2}{3}} \mathtt{A}_8^{\frac{1}{3}}}{4d} \left(\frac{k}{\log(k+1)}\right)^{\frac{\beta+d+1}{2\beta+d}} + 1\right) \\ &\leq d \log\left(\frac{\mathtt{A}_6^{\frac{2}{3}} \mathtt{A}_8^{\frac{1}{3}}}{2d} k + 1\right) \\ &\leq d \frac{2\beta+d}{\beta} \log\left(\left(\mathtt{A}_6^{\frac{2}{3}} \mathtt{A}_8^{\frac{1}{3}} k + 1\right)^{\frac{\beta}{2\beta+d}}\right) \ , \end{aligned}$$

Since $\mathtt{A}_6^{\frac{2}{3}} \mathtt{A}_8^{\frac{1}{3}} \geq 1$ and $\beta/(2\beta+d) \leq 1$, we have

$$\left(\mathtt{A}_6^{\frac{2}{3}} \mathtt{A}_8^{\frac{1}{3}} k + 1\right)^{\frac{\beta}{2\beta+d}} \leq \left(\mathtt{A}_6^{\frac{2}{3}} \mathtt{A}_8^{\frac{1}{3}} k^{\frac{\beta}{2\beta+d}}\right) + 1 \ .$$

Furthermore, using the fact that $\log(x+1) \leq x$, for all $x > 0$, we find that term V $\leq d\frac{2\beta+d}{\beta} \cdot \mathtt{A}_6^{\frac{2}{3}} \mathtt{A}_8^{\frac{1}{3}} k^{\frac{\beta}{2\beta+d}}$, and consequently that

$$2 \cdot \text{term V} \leq \text{term IV} \ .$$

Therefore, we deduce that

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})-\mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{\mathsf{op}}^4\right] \le \mathtt{A}_9 k^{-2}h_k^{-2d}\log(k+1)^2 + \underbrace{S\int_{t=0}^{\infty}4t^3\exp\left(-\mathtt{A}_{10}\min\left(kt^2h_k^d, kth_k^d\right)\right)\,\mathrm{d}t}_{\text{term VI}} ,$$

where $\mathtt{A}_9 = \left(4d^{\frac{2\beta+d}{\beta}}\sqrt[3]{\frac{\mathtt{A}_8}{\mathtt{A}_6}}\right)^4$ and $\mathtt{A}_{10} = \frac{\mathtt{A}_6}{2}$. To conclude the proof it is enough to provide an upper bound for term VI. In order to calculate the integral in term VI, we proceed with similar steps as in the proof of Lemma 6.7.10 and we obtain

$$\text{term VI} \le \mathtt{A}_{11}k^{-2}h_k^{-2d} ,$$

where $\mathtt{A}_{11} > 0$ only depends on $\mathtt{A}_{10}$ and S. We conclude the first part of the proof by letting $\mathtt{A}_1 = \mathtt{A}_9 + \mathtt{A}_{11}$. For the second part of the proof, similar to the proof of Lemma 6.7.10, we can write

$$\begin{aligned}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|B_{k,\lambda}(\mathbf{x})^{-1}\right\|_{\mathsf{op}}^4\right] &\le 4\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|B_{k,\lambda}(\mathbf{x})^{-1}-\left(\mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\right)^{-1}\right\|_{\mathsf{op}}^4\right] + 4\sup_{\mathbf{x}\in\Theta}\left\|\left(\mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\right)^{-1}\right\|_{\mathsf{op}}^4 \\ &\le 4\lambda_{\min}^{-4}\lambda_k^{-4}\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})-\mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{\mathsf{op}}^4\right] + 4\lambda_{\min}^{-4} .\end{aligned}$$

By the first part of the proof we have $\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})-\mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{\mathsf{op}}^4\right] \le \mathtt{A}_1k^{-2}h_k^{-2}\log(k+1)^2$, which gives

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|B_{k,\lambda}(\mathbf{x})^{-1}\right\|_{\mathsf{op}}^4\right] \le 4\mathtt{A}_1\lambda_{\min}^{-4}k^{-2}h_k^{-2d}\log(k+1)^2\lambda_k^{-4} + 4\lambda_{\min}^{-4} .$$

Since $k \ge \lambda_k^2h_k^{-d}\log(k+1)$, we deduce that

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|B_{k,\lambda}(\mathbf{x})^{-1}\right\|_{\mathsf{op}}^4\right] \le 4(\mathtt{A}_1+1)\lambda_{\min}^{-4} .$$

We finish the proof by assigning $\mathtt{A}_2 = 4(\mathtt{A}_1+1)$. □

**Lemma 6.7.10.** *Let $k \in [n]$, with $1 \le kh_k^d$, and let Assumption 6.2.4 hold. Then, for any $\mathbf{x} \in \Theta$, we have*

$$\mathbf{E}\left[\|B_{k,\lambda}(\mathbf{x})-\mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{op}^4\right] \le \mathtt{A}_1h_k^{-2d}k^{-2} .$$

*Furthermore, for $k \ge \lambda_k^{-2}h_k^{-d}$, we have $\mathbf{E}\left[\|B_{k,\lambda}(\mathbf{x})\|_{op}^{-4}\right] \le \mathtt{A}_2\lambda_{\min}^{-4}$ , where $\mathtt{A}_1, \mathtt{A}_2 > 0$ are numerical constants.*

*Proof.* Let $Q_{i,k}(\mathbf{x}) = h_k^{-d} M_{i,k}(\mathbf{x}) - h_k^{-d}\mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]$. We introduce

$$\omega = \sup_{\mathbf{x}\in\Theta}\max_{1\le i\le n}\|Q_{i,k}(\mathbf{x})\|_{\mathsf{op}}, \quad \text{and} \quad \upsilon^2 = \sup_{\mathbf{x}\in\Theta}\left\|\sum_{i=1}^{n} \mathbf{E}Q_{i,k}^{\top}(\mathbf{x})Q_{i,k}(\mathbf{x})\right\|_{\mathsf{op}} . \tag{6.21}$$

Note that for any $i \in [k]$ and $\mathbf{x} \in \mathbb{R}^d$, $Q_{i,k}(\mathbf{x}) \in \mathbb{R}^{S\times S}$. Then by (Vershynin, 2019, Theorem 5.4.1), for any $t \geq 0$, we have

$$\mathbf{P}\left[\left\|\sum_{i=1}^{k} Q_{i,k}(\mathbf{x})\right\|_{\mathsf{op}} \geq t\right] \leq 2S\exp\left(-c\min\left(\frac{t^2}{\upsilon^2}, \frac{t}{\omega}\right)\right),$$

where $c > 0$ is a numerical constant.

$$\begin{aligned}\mathbf{E}\left[\left\|\sum_{i=1}^{k} Q_{i,k}(\mathbf{x})\right\|_{\mathsf{op}}^{4}\right] &= \int_0^\infty 4t^3\mathbf{P}\left[\left\|\sum_{i=1}^{k} \boldsymbol{Q}_{i,k}(\mathbf{x})\right\|_{\mathsf{op}} \geq t\right] \mathrm{d}t \\ &= \underbrace{4S\int_0^{\frac{\upsilon^2}{K}} t^3\exp\left(-c\frac{t^2}{\upsilon^2}\right)\mathrm{d}t}_{\text{term I}} + \underbrace{4S\int_{\frac{\upsilon^2}{K}}^{\infty} t^3\exp\left(-c\frac{t}{\omega}\right)\mathrm{d}t}_{\text{term II}} .\end{aligned}$$

We provide upper bounds for the terms I and II, separately.

$$\begin{aligned}\text{term I} &= 2S\frac{\upsilon^2}{c}\left(-t^2\exp\left(-c\frac{t^2}{\upsilon^2}\right)\right)\Bigg|_{t=0}^{\frac{\upsilon^2}{\omega}} + 4S\frac{\upsilon^2}{c}\int_0^{\frac{\upsilon^2}{\omega}} t\exp\left(-c\frac{t^2}{\upsilon^2}\right)\mathrm{d}t \\ &\leq 4S\frac{\upsilon^2}{c}\int_0^{\frac{\upsilon^2}{\omega}} t\exp\left(-c\frac{t^2}{\upsilon^2}\right)\mathrm{d}t \\ &= 2S\frac{\upsilon^4}{c^2}\int_0^{\frac{\upsilon^2}{\omega}} \frac{2ct}{\upsilon^2}\exp\left(-c\frac{t^2}{\upsilon^2}\right)\mathrm{d}t \\ &= -2S\frac{\upsilon^4}{c^2}\exp(-c\frac{t^2}{\upsilon^2})\Bigg|_{t=0}^{\frac{\upsilon^2}{\omega}} \leq 2s\frac{\upsilon^4}{c^2} .\end{aligned}$$

Similarly, for term II we can write

$$\begin{aligned}
\text{term II} &= -4S\frac{\omega}{c}t^3 \exp\left(-c\frac{t}{\omega}\right)\Bigg|_{\frac{\upsilon^2}{\omega}}^{\infty} + 12S\frac{\omega}{c}\int_{\frac{\upsilon^2}{\omega}}^{\infty} t^2 \exp\left(-c\frac{t}{\omega}\right)\,\mathrm{d}t \\
&= 4S\frac{\upsilon^6}{c\omega^2}\exp\left(-c\frac{\upsilon^2}{\omega^2}\right) + 12S\frac{\omega}{c}\int_{\frac{\upsilon^2}{\omega}}^{\infty} t^2 \exp\left(-c\frac{t}{\omega}\right)\,\mathrm{d}t \\
&\leq 4S\frac{\upsilon^4}{c^2} + 12S\frac{\omega}{c}\int_{\frac{\upsilon^2}{\omega}}^{\infty} t^2 \exp\left(-c\frac{t}{\omega}\right)\,\mathrm{d}t \\
&\leq 4S\frac{\upsilon^4}{c^2} - 12S\frac{\omega^2}{c^2}t^2 \exp\left(-c\frac{t}{\omega}\right)\Bigg|_{\frac{\upsilon^2}{\omega}}^{\infty} + 24S\frac{\omega^2}{c^2}\int_{\frac{\upsilon^2}{\omega}}^{\infty} t\exp(-c\frac{t}{\omega})\,\mathrm{d}t \\
&\leq 4S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^4}{c^2}\exp\left(-c\frac{\upsilon^2}{\omega^2}\right) + 24S\frac{\omega^2}{c^2}\int_{\frac{\upsilon^2}{\omega}}^{\infty} t\exp(-c\frac{t}{\omega})\,\mathrm{d}t \\
&= 4S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^2\omega^2}{c^3} - 24S\frac{\omega^3}{c^3}t\exp\left(-c\frac{t}{\omega}\right)\Bigg|_{\frac{\upsilon^2}{\omega}}^{\infty} + 24S\frac{\omega^3}{c^3}\int_{\frac{\upsilon^2}{\omega}}^{\infty} \exp\left(-c\frac{t}{\omega}\right)\,\mathrm{d}t \\
&\leq 4S\frac{\upsilon^4}{c^2} + 12s\frac{\upsilon^2\omega^2}{c^3} + 24S\frac{\omega^4}{c^4} + 24s\frac{\omega^3}{c^3}\int_{\frac{\upsilon^2}{\omega}}^{\infty} \exp\left(-c\frac{t}{\omega}\right)\,\mathrm{d}t \\
&\leq 4S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^2\omega^2}{c^3} + 24S\frac{\omega^4}{c^4} - 24S\frac{\omega^4}{c^4}\exp\left(-c\frac{t}{\omega}\right)\Bigg|_{\frac{\upsilon^2}{\omega}}^{\infty} \\
&\leq 4S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^2\omega^2}{c^3} + 24S\frac{\omega^4}{c^4} + 24S\frac{\omega^6}{c^5\upsilon^2} \quad .
\end{aligned}$$

By combining the provided bounds for the terms I and II, we deduce that

$$\mathbf{E}\left[\left\|\sum_{i=1}^{k} Q_{i,k}(\mathbf{x})\right\|_{\text{op}}^{4}\right] \leq 6S\frac{\upsilon^4}{c^2} + 12S\frac{\upsilon^2\omega^2}{c^3} + 24S\frac{\omega^4}{c^4} + 24S\frac{\omega^6}{c^5\upsilon^2} \quad . \tag{6.22}$$

To conclude the first part of the proof, it is enough to bound from above the terms $K$ and $\sigma$ defined in (6.21). For $K$, we can write

$$\omega \leq \sup_{\mathbf{x}\in\Theta}\max_{i\in[k]} 2h_k^{-d}\left\|U\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right)U\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right)^{\top} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right)\right\|_{\text{op}} \leq \mathbb{A}_4 h_n^{-d} \quad ,$$

where $\mathbb{A}_4 = \max_{\boldsymbol{u}\in\mathrm{Supp}(K)}\left\|U\left(\boldsymbol{u}\right)U\left(\boldsymbol{u}\right)^{\top} K\left(\boldsymbol{u}\right)\right\|_{\text{op}}$. For $\upsilon^2$, by Lemma 6.7.8(ii), we have

$$\begin{aligned}
\upsilon^2 = \sup_{\mathbf{x}\in\Theta}\sum_{i=1}^{k}\left\|\mathbf{E}Q_{i,k}^{\top}(\mathbf{x})Q_{i,k}(\mathbf{x})\right\|_{\text{op}} &\leq kh_k^{-2d}\sup_{\mathbf{x}\in\Theta}\mathbf{E}\left[\left\|U\left(\frac{\mathbf{x}_1 - \mathbf{x}}{h_k}\right)U\left(\frac{\mathbf{x}_1 - \mathbf{x}}{h_k}\right)^{\top} K\left(\frac{\mathbf{x}_1 - \mathbf{x}}{h_k}\right)\right\|_{\text{op}}^{2}\right] \\
&\leq p_{\max}\nu_{2,2}kh_k^{-d} \quad .
\end{aligned}$$

Substituting the above bounds in (6.22) we get

$$\mathbf{E}\left[\left\|\sum_{i=1}^{k} Q_{i,k}(\mathbf{x})\right\|_{\mathsf{op}}^{4}\right] \leq \mathtt{A}_4\left[k^2 h_k^{-2d} + k h_n^{-3d} + h_k^{-4d}\right] \quad ,$$

where $\mathtt{A}_4 > 0$ is a numerical constant. Since $\frac{1}{k}\sum_{i=1}^{k} Q_{i,k}(\mathbf{x}) = B_{k,\lambda}(\mathbf{x}) - \mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]$ we deduce that

$$\mathbf{E}\left[\|B_{k,\lambda}(\mathbf{x}) - \mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{\mathsf{op}}^{4}\right] \leq \mathtt{A}_3\left[k^{-2} h_k^{-2d} + k^{-3} h_k^{-3d} + k^{-4} h_k^{-4d}\right] \quad .$$

Since $1 \leq k h_k^d$ we get

$$\mathbf{E}\left[\|B_{k,\lambda}(\mathbf{x}) - \mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{\mathsf{op}}^{4}\right] \leq \mathtt{A}_1 h_k^{-2d} k^{-2} \quad ,$$

with $\mathtt{A}_1 = 3\mathtt{A}_3$. For the second part the proof, we can write

$$\begin{aligned}\mathbf{E}\left[\left\|B_{k,\lambda}(\mathbf{x})^{-1}\right\|_{\mathsf{op}}^{4}\right] &\leq 4\mathbf{E}\left[\left\|B_{k,\lambda}(\mathbf{x})^{-1} - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right\|_{\mathsf{op}}^{4}\right] + \frac{4}{\lambda_{\min}^{4}} \\ &\leq \frac{4\mathtt{A}_1}{\lambda^4 \lambda_{\min}^{4}} h_k^{-4d} k^{-2} + \frac{4}{\lambda_{\min}^{4}} \quad .\end{aligned}$$

So, for any $k \geq \lambda_k^{-2} h_k^{-d}$, we have

$$\mathbf{E}\left[\left\|B_{k,\lambda}(\mathbf{x})^{-2}\right\|_{\mathsf{op}}\right] \leq \frac{\mathtt{A}_2}{\lambda_{\min}^{4}},$$

where $\mathtt{A}_2 = 4\mathtt{A}_1 + 4$. □

**Lemma 6.7.11.** *Let $k \in [n]$, with $h_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{1}{2\beta+d}}$, and let Assumptions 6.2.2 and 6.2.4 hold. Then, we have*

$$\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\left\|k^{-1} h_k^{-d}\sum_{i=1}^{k} \boldsymbol{R}_{i,k}(\boldsymbol{x})\xi_i\right\|^{4}\right] \leq \mathtt{A} k^{-2} h_k^{-2d}\log(k+1)^2 \quad .$$

*Proof.* Let $\boldsymbol{G}_k(\mathbf{x}) = k^{-1} h_k^{-d}\sum_{i=1}^{k} \boldsymbol{R}_{i,k}(\mathbf{x})\xi_i$. The objective of this proof is to provide a control for the term $\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|\boldsymbol{G}_k(\mathbf{x})\|^4\right]$. First, we show that $\boldsymbol{G}(\cdot)$ is upper bounded by a Lipschitz function.

**Providing a Lipschitz upper bound.** Note that we can write

$$\|\boldsymbol{G}(\mathbf{x})\| \leq \sum_{s=1}^{S}\left|k^{-1} h_k^{-d}\sum_{i=1}^{k}\left(\boldsymbol{U}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right)\right)_s \xi_i\right| \quad ,$$

where $\left(\boldsymbol{U}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h_k}\right) K\left(\frac{\mathbf{x}_i-\mathbf{x}}{h_k}\right)\right)_s$ is the $s$-th coordinate of the vector $\boldsymbol{U}\left(\frac{\mathbf{x}_i-\mathbf{x}}{h_k}\right) K\left(\frac{\mathbf{x}_i-\mathbf{x}}{h_k}\right)$, for $s \in [S]$. With similar steps of deduction as in the proof of Lemma 6.7.9, we can see that there exists $\mathbb{A}_{\mathsf{Lip}} > 0$, such that for any $\mathbf{x}, \mathbf{y} \in \Theta$

$$\sum_{s=1}^{S} \left| k^{-1} h_k^{-d} \sum_{i=1}^{k} \left( \left( \boldsymbol{U}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right)\right)_s - \left( \boldsymbol{U}\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_k}\right) K\left(\frac{\mathbf{x}_i - \mathbf{y}}{h_k}\right)\right)_s \right) \xi_i \right| \le \mathbb{A}_{\mathsf{Lip}} h_k^{-d-1} \left\| \mathbf{x} - \mathbf{y} \right\| \ .$$

Let $F_i^{(s)}(\mathbf{x}) = \left( \boldsymbol{U}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right) K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h_k}\right)\right)_s \xi_i$, for $i \in [k]$. Thus, we can write

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta} \|\boldsymbol{G}(\mathbf{x})\|\right] \le \sum_{s=1}^{S} \mathbf{E}\left[\sup_{\mathbf{x}\in\Theta} \left| k^{-1} h_k^{-d} \sum_{i=1}^{k} F_i^{(s)}(\mathbf{x}) \right| \right] \ ,$$

Now, for $\epsilon > 0$, consider an $\epsilon$-net of $\Theta$, namely $\mathcal{N}$, with cardinality $\mathcal{N}(\Theta, \epsilon)$. Then, we have

$$\begin{aligned} \mathbf{E}\left[\sup_{\mathbf{x}\in\Theta} \|\boldsymbol{G}(\mathbf{x})\|^4\right] &\le \sum_{s=1}^{S} 4\mathbf{E}\left[\sup_{\mathbf{x}\in\mathcal{N}} \left| k^{-1} h_k^{-d} \sum_{i=1}^{k} F_i^{(s)}(\mathbf{x}) \right|^4 \right] \\ &\quad + \sum_{s=1}^{S} 4\mathbf{E}\left[\sup_{\mathbf{x},\boldsymbol{w}:\|\mathbf{x}-\boldsymbol{w}\|\le\epsilon} k^{-4} h_k^{-4d} \left| \sum_{i=1}^{k} \left( F_i^{(s)}(\mathbf{x}) - F_i^{(s)}(\boldsymbol{w}) \right) \right|^4 \right] \\ &\le \underbrace{\sum_{s=1}^{S} \mathbf{E}\left[\sup_{\mathbf{x}\in\mathcal{N}} \left| k^{-1} h_k^{-d} \sum_{i=1}^{k} F_i^{(s)}(\mathbf{x}) \right|^4 \right]}_{\text{term I}} + \mathbb{A}_1 h_k^{-4d-4} \epsilon^4 \ , \end{aligned} \tag{6.23}$$

where we introduced $\mathbb{A}_1 = 32\mathbb{A}_{\mathsf{Lip}}^4 \sigma^4$. Now, we wish to provide an upper bound for term I above.

$$\begin{aligned} \mathbb{P}\left[\sup_{\mathbf{x}\in\mathcal{N}} \left| k^{-1} h_k^{-d} \sum_{i=1}^{k} F_i^{(s)}(\mathbf{x}) \right| \ge t \right] &\le \mathcal{N}(\Theta, \epsilon) \sup_{\mathbf{x}\in\mathcal{N}} \mathbb{P}\left[\left| k^{-1} h_k^{-d} \sum_{i=1}^{k} F_i^{(s)}(\mathbf{x}) \right| \ge t \right] \\ &\le \left(\frac{\mathsf{diam}(\Theta)}{\epsilon} + 1\right)^d \sup_{\mathbf{x}\in\mathcal{N}} \mathbb{P}\left[\left| k^{-1} h_k^{-d} \sum_{i=1}^{k} F_i^{(s)}(\mathbf{x}) \right| \ge t \right] \end{aligned}$$

**The term $k^{-1} h_k^{-d} \sum_{i=1}^{k} F_i^{(s)}(\mathbf{x})$ is sub-Gaussian.** For any $t \ge 0$, and $\mathbf{x} \in \mathcal{N}$ we can write

$$\mathbf{E}\left[\exp\left(t\left(k^{-1} h_k^{-d} \sum_{i=1}^{k} F_i^{(s)}(\mathbf{x})\right)^2\right)\right] = \sum_{q=0}^{\infty} \frac{t^q k^{-2q} h_k^{-2qd} \mathbf{E}\left[\left(\sum_{i=1}^{k} F_i^{(s)}(\mathbf{x})\right)^{2q}\right]}{q!} \ . \tag{6.24}$$

Now, for any non negative integer $q$, we provide an upper bound for the term $\mathbf{E}\left[k^{-2q}h_k^{-2qd}\left(\sum_{i=1}^k F_i^{(s)}(\mathbf{x})\right)^{2q}\right]$.

$$\begin{aligned}\mathbf{E}\left[\left(\left(\boldsymbol{U}(\frac{\mathbf{x}_i-\mathbf{x}}{h_k})K(\frac{\mathbf{x}_i-\mathbf{x}}{h_k})\right)_s\right)^{2q}\right] &= \int\left(\left(\boldsymbol{U}(\frac{\mathbf{x}_i-\mathbf{x}}{h_k})K(\frac{\mathbf{x}_i-\mathbf{x}}{h_k})\right)_s\right)^{2q} p(\mathbf{x}_i)\,\mathrm{d}\mathbf{x}_i \\ &= h_k^d\int\left((\boldsymbol{U}(\boldsymbol{u})K(\boldsymbol{u}))_s\right)^{2q}p(h_k\boldsymbol{u}+\mathbf{x})\,\mathrm{d}\boldsymbol{u}\\ &\le h_k^d L_{\max}^{2q}\ ,\end{aligned}$$

where we introduced $L_{\max}=\max_{s\in[S]}\max_{\boldsymbol{u}\in\mathsf{Supp}(K)}|\,(\boldsymbol{U}(\boldsymbol{u})K(\boldsymbol{u}))_s\,|$.

$$\begin{aligned}k^{-2q}h_k^{-2qd}\mathbf{E}\left[\left(\sum_{i=1}^k F_i^{(s)}(\mathbf{x})\right)^{2q}\right] &\le (2\sigma L_{\max})^{2q}q!k^{-2q}h_k^{-2qd}\left(kh_k^d+k^2h_k^{2d}\binom{q-1}{1}+\dots+k^qh_k^{qd}\binom{q-1}{q-1}\right)\\ &\le (\mathtt{A}_2k^{-1}h_k^{-d})^q q!\ ,\end{aligned}$$

where $\mathtt{A}_2=4\sigma^2L_{\max}^2$, and we used the fact that $kh_k^d\ge 1$. By letting $t_0=2\mathtt{A}_2k^{-1}h_k^{-d}$ in (6.24), we can write

$$\mathbf{E}\left[\exp\left(t_0\left(k^{-1}h_k^{-d}\sum_{i=1}^k F_i^{(s)}(\mathbf{x})\right)^2\right)\right]\le 2\ .$$

Let $\epsilon=\mathsf{diam}(\Theta)h_k^{\frac{3d}{4}+1}k^{-1}$. Since $k^{-1}h_k^{-d}\sum_{i=1}^k F_i^{(s)}$ is sub-Gaussian, we have

$$\begin{aligned}\mathbb{P}\left[\sup_{\mathbf{x}\in\mathcal{N}}\left|k^{-1}h_k^{-d}\sum_{i=1}^k F_i^{(s)}(\mathbf{x})\right|\ge t\right] &\le 2\exp\left(-\frac{t^2}{4\mathtt{A}_2k^{-1}h_k^{-d}}+d\log\left(\frac{\mathsf{diam}(\Theta)}{\epsilon}+1\right)\right)\\ &\le 2\exp\left(-\frac{t^2}{4\mathtt{A}_2k^{-1}h_k^{-d}}+\mathtt{A}_3\log\left(k\right)\right)\ ,\end{aligned}$$

where $\mathtt{A}_3=d\left(\frac{3d+4}{4(2\beta+d)}+1\right)$.

**The final bound.** For any $a\ge 0$, we have

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\mathcal{N}}\|\boldsymbol{G}(\mathbf{x})\|^4\right]\le Sa^4+8S\int_a^\infty t^3\exp\left(-\frac{t^2}{4\mathtt{A}_2k^{-1}h_k^{-d}}+\mathtt{A}_3\log\left(k\right)\right)\,\mathrm{d}t\ .$$

Take $a=(8\mathtt{A}_2\mathtt{A}_3k^{-1}h_k^{-d}\log(k))^{\frac{1}{2}}$, then we have

$$\begin{aligned}\mathbf{E}\left[\sup_{\mathbf{x}\in\mathcal{N}}\|\boldsymbol{G}(\mathbf{x})\|^4\right] &\le 64S\mathtt{A}_2^2\mathtt{A}_3^2k^{-2}h_k^{-2d}\log(k)^2+8S\int_a^\infty t^3\exp\left(-\frac{t^2}{8\mathtt{A}_2k^{-1}h_k^{-d}}\right)\,\mathrm{d}t\\ &\le \mathtt{A}_4k^{-2}h_k^{-2d}\log(k+1)^2\ ,\end{aligned}\tag{6.25}$$

where $\mathtt{A}_4>0$ is a numerical constant. By using (6.25), and substituting $\epsilon=\mathsf{diam}(\Theta)h_k^{\frac{3d}{4}+1}k^{-1}$

in (6.23), we get

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|\boldsymbol{G}(\mathbf{x})\|^4\right] \leq \mathtt{A}k^{-2}h_k^{-2d}\log(k+1)^2 \ ,$$

where we introduced $\mathtt{A} = \mathtt{A}_1 + \mathtt{A}_4$. □

**Lemma 6.7.12.** *Let $k \in [n]$, with $1 \leq kh_k^d$, and let Assumptions 6.2.3(iv) and 6.2.4 hold. Then, for any $\boldsymbol{x} \in \Theta$, we have*

$$\mathbf{E}\left[\|\boldsymbol{C}_k(\boldsymbol{x}) - \mathbf{E}\left[\boldsymbol{C}_k(\boldsymbol{x})\right]\|^4\right] \leq \mathtt{A}h_k^{-2d}k^{-2} \ .$$

We omit the proof of this lemma since it follows the same lines as the proof of Lemma 6.7.10.

**Lemma 6.7.13.** *Let $k \in [n]$, and $h_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{1}{2\beta+d}}$. Let Assumptions 6.2.3(iv) and 6.2.4 hold. Then, for any $\boldsymbol{x} \in \Theta$, we have*

$$\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\|\boldsymbol{C}_k(\boldsymbol{x}) - \mathbf{E}\left[\boldsymbol{C}_k(\boldsymbol{x})\right]\|^4\right] \leq \mathtt{A}h_k^{-2d}k^{-2}\log(k+1)^2 \ .$$

We omit the proof of this lemma since it follows the same lines as the proof of Lemma 6.7.9.

**Lemma 6.7.14.** *Let $k \in [n]$, and $h_k = \left(\frac{\log(k+1)}{k}\right)^{\frac{1}{2\beta+d}}$. Let Assumption 6.2.4 hold, and $k \geq \lambda_k^{-2}h_k^{-d}\log(k+1)$. Then, for any $\boldsymbol{x} \in \Theta$, we have*

$$\mathbf{E}\left[\sup_{\boldsymbol{x}\in\Theta}\left\|B_{k,\lambda}(\boldsymbol{x})^{-1} - (\mathbf{E}[B_{k,\lambda}(\boldsymbol{x})])^{-1}\right\|_{op}\sup_{\boldsymbol{x}\in\Theta}\|\boldsymbol{C}_k(\boldsymbol{x}) - \mathbf{E}\left[\boldsymbol{C}_k(\boldsymbol{x})\right]\|\right] \leq \mathtt{A}h_k^{-d}k^{-1}\log(k+1) \ ,$$

*where $\mathtt{A} > 0$ is a numerical constant.*

*Proof.* In view of the Cauchy-Schwarz inequality, it is enough to provide an upper bound for the following terms:

$$\left(\underbrace{\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\left\|B_{k,\lambda}(\mathbf{x})^{-1} - (\mathbf{E}[B_{k,\lambda}(\mathbf{x})])^{-1}\right\|_{\text{op}}^2\right]}_{\text{term I}}\underbrace{\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|\boldsymbol{C}_k(\mathbf{z}_n) - \mathbf{E}\left[\boldsymbol{C}_k(\mathbf{x})\right]\|^2\right]}_{\text{term II}}\right)^{\frac{1}{2}} \ .$$

For term I, we use the Cauchy-Schwarz inequality once more and we get

$$\text{term I} \leq \lambda_{\min}^{-2}\left(\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x})\|_{\text{op}}^{-4}\right]\mathbf{E}\left[\sup_{\mathbf{x}\in\Theta}\|B_{k,\lambda}(\mathbf{x}) - \mathbf{E}\left[B_{k,\lambda}(\mathbf{x})\right]\|_{\text{op}}^4\right]\right)^{\frac{1}{2}} \leq \mathtt{A}_1h_k^{-d}k^{-1}\log(k+1) \ ,$$

where the last inequality is obtained by Lemma 6.7.9, , and $\mathtt{A}_1 > 0$ is a numerical constant.

For term II, Lemma 6.7.13, yields

$$\text{term II} \leq \left(\mathbf{E}\left[\|\boldsymbol{C}_k(\mathbf{x}) - \mathbf{E}\left[\boldsymbol{C}_k(\mathbf{x})\right]\|^2\right]\right) \leq \mathtt{A}_2 h_k^{-d} k^{-1} \log(k+1) \ ,$$

where $\mathtt{A}_2 > 0$ is the numerical constant that appears in Lemma 6.7.13. We conclude the proof by combining the above bounds. □

# Chapter 7

# Group meritocratic fairness in linear contextual bandits

We study the linear contextual bandit problem where an agent has to select one candidate from a pool and each candidate belongs to a sensitive group. In this setting, candidates' rewards may not be directly comparable between groups, for example when the agent is an employer hiring candidates from different ethnic groups and some groups have a lower reward due to discriminatory bias and/or social injustice. We propose a notion of fairness that states that the agent's policy is fair when it selects a candidate with highest relative rank, which measures how good the reward is when compared to candidates from the same group. This is a very strong notion of fairness, since the relative rank is not directly observed by the agent and depends on the underlying reward model and on the distribution of rewards. Thus we study the problem of learning a policy which approximates a fair policy under the condition that the contexts are independent between groups and the distribution of rewards of each group is absolutely continuous. In particular, we design a greedy policy which at each round constructs a ridge regression estimate from the observed context-reward pairs, and then computes an estimate of the relative rank of each candidate using the empirical cumulative distribution function. We prove that, despite its simplicity and the lack of an initial exploration phase, the greedy policy achieves, up to log factors and with high probability, a fair pseudo-regret of order $\sqrt{dT}$ after $T$ rounds, where $d$ is the dimension of the context vectors. The policy also satisfies demographic parity at each round when averaged over all possible information available before the selection. Finally, we use simulated settings and experiments on the US census data to show that our policy achieves sub-linear fair pseudo-regret also in practice.

## 7.1 Introduction

We consider the linear contextual bandit setup (Auer, 2002) where at each round $t \in [T]$, an agent receives a set of feature vectors $\{X_{t,a}\}_{a=1}^{K}$ with $X_{t,a} \subset \mathbb{R}^d$ sampled from the environment, one for each arm $a \in [K]$. We assume that context (or candidate) $X_{t,a}$ has an associated reward $\langle \mu^*, X_{t,a} \rangle$ where $\mu^* \in \mathbb{R}^d$ is unknown to the agent. After the agent selects the arm $a_t$, it receives the noisy reward equal to $r_{t,a_t} = \langle \mu^*, X_{t,a_t} \rangle + \eta_t$, where $\eta_t$ is some scalar noise (formally specified later). In addition, we assume that each arm represents a fixed sensitive group (e.g. based on ethnicity, gender, etc.). The latter assumption simplifies the presentation but implies that at each round the agent receives exactly one candidate for each group. This can be too restrictive e.g. when candidates are sampled i.i.d. together with their group and/or some groups are minorities. However, our results can be easily adapted to more realistic settings without such assumption, as we show in Section 7.5 and more rigorously in Section 7.7. Excluding these sections, we use arm and group interchangeably in all that follows.

Usually, the goal of the agent is to maximise the expected cumulative reward $\sum_{t=1}^{T} \langle \mu^*, X_{t,a_t} \rangle$. Since as we previously explained, this objective might be unfair to some of the sensitive groups, we instead use a different kind of reward which measures the relative performance of a candidate compared to others of the same arm/group. First, we additionally assume, for each group $a$, that $\{X_{t,a}\}_{t=1}^{T}$ are i.i.d and have the same distribution of $X_a$, which we define to be a random variable with unknown distribution. We call the distribution of $\langle \mu^*, X_a \rangle$ the reward distribution of arm $a$ and denote with $\mathcal{F}_a$ its CDF, i.e. $\mathcal{F}_a(r) = \mathbb{P}(\langle \mu^*, X_a \rangle \leq r)$ for every $r \in \mathbb{R}$. Then, we introduce the *relative rank* of candidate $X_{t,a}$ as $\mathcal{F}_a(\langle \mu^*, X_{t,a} \rangle)$, that is the probability that a sample from the reward distribution of arm $a$ is lower than the reward of $X_{t,a}$. We argue that the relative rank, allows to have a fair way of comparing candidates from different groups and introduce the following fairness definition.

**Definition 7.1.1** (Group Meritocratic Fairness). *A policy $\{a_t^*\}_{t=1}^{\infty}$ is group meritocratic fair (GMF) if for all $t \in \mathbb{N}, a \in [K]$*

$$\mathcal{F}_{a_t^*}(\langle \mu^*, X_{t,a_t^*} \rangle) \geq \mathcal{F}_a(\langle \mu^*, X_{t,a} \rangle) \ .$$

A GMF policy chooses candidates with the highest reward compared to candidates from the same group. This is a strong definition of fairness which is impossible to satisfy at each round for a learned policy. As in standard linear contextual bandits, $\mu^*$ is unknown and must be learned. In this setting however, we have the additional challenge of learning the CDF for the rewards of each arm, $\mathcal{F}_a$. Thus, we will focus on how to learn a GMF policy by introducing the following regret definition.

**Definition 7.1.2** (Fair Pseudo-Regret). *Let $T \in \mathbb{N}$, $\{a_t\}_{t=1}^T$ be the evaluated policy and $\{a_t^*\}_{t=1}^T$ be a GMF policy. Then we denote by (cumulative) fair pseudo-regret the quantity*

$$R_F(T) := \sum_{t=1}^T \mathcal{F}_{a_t^*}(\langle \mu^*, X_{t,a_t^*}\rangle) - \mathcal{F}_{a_t}(\langle \mu^*, X_{t,a_t}\rangle) \ .$$

The goal of the learned policy will be to minimize the fair pseudo-regret, since a policy with sublinear fair pseudo-regret will get closer and closer to a GMF fair policy over time.

**Remark 7.1.3.** *The fair pseudo-regret resembles the standard pseudo-regret defined as*

$$R(T) := \sum_{t=1}^T \langle \mu^*, X_{t,a_t^{\mathrm{opt}}}\rangle - \langle \mu^*, X_{t,a_t}\rangle \quad \textit{with} \quad a_t^{\mathrm{opt}} \in \arg\max_{a\in[K]} \langle \mu^*, X_{t,a}\rangle \ ,$$

*where rewards are replaced by relative ranks and $a_t^{\mathrm{opt}}$ by the GMF policy $a_t^*$. Furthermore, since the CDF restricted to the support is strictly increasing, when the reward distributions are the same for each arm, i.e. $\mathcal{F}_a = \mathcal{F}_{a'}$ for all $a, a' \in [K]$, then a policy minimizing the fair pseudo-regret also minimizes the standard pseudo-regret and vice versa. This is not true in the general case, where fair and standard pseudo-regrets are often competing objectives. For example, when $\{\langle \mu^*, X_a\rangle\}_{a=1}^K$ are independent and absolutely continuous and there exists $\hat{a}$ such that $\langle \mu^*, X_{\hat{a}}\rangle > \langle \mu^*, X_a\rangle$ for every $a \neq \hat{a}$, then for every $t$, $a_t^{\mathrm{opt}} = \hat{a}$, while as we will show in Proposition 7.1.4, $a_t^*$ selects each arm with equal probability. Thus, with non-zero probability $a_t^{\mathrm{opt}}$ has a linear fair pseudo-regret while $a_t^*$ has a linear standard pseudo-regret. Moreover, in Section 7.7, for $K = 2$, we show that if $\langle \mu^*, X_1\rangle$ and $\langle \mu^*, X_2\rangle$ are independent, absolutely continuous, but not identically distributed, then the GMF policy has a linear standard regret and $\{a_t^{opt}\}_{t=1}^\infty$ has a linear fair regret with positive probability.*

Learning a GMF policy brings several challenges. The relative rank is not directly observed by the agent, which receives instead only the noisy reward. This implies that the agent has to estimate $\mathcal{F}_a$, which in general might not even be Lipschitz continuous. This is the main reason why we restrict our analysis to the case where the rewards $\{\langle \mu^*, X_a\rangle\}_{a=1}^K$ are independent and absolutely continuous. In particular, for any $t \geq 0$, let $\mathcal{H}_t^- := \cup_{i=1}^t \left\{\{X_{i,a}\}_{a=1}^K, r_{i,a_i}, a_i\right\}$ with $\mathcal{H}_0^- = \varnothing$ and $\mathcal{H}_t := \mathcal{H}_t^- \cup \{\{X_{t+1,a}\}_{a=1}^K\}$ be respectively the history and the information available for the decision at round $t+1$, then the following holds.

**Proposition 7.1.4** (GMF policy satisfies *history-agnostic demographic parity*). *Let $\{\langle \mu^*, X_a\rangle\}_{a=1}^K$ be independent and absolutely continuous and for every $a \in [K], t \in \mathbb{N}$, let $X_{t,a}$ be an i.i.d. copy of $X_a$. Then for every $t \in \mathbb{N}$, $\{\mathcal{F}_a(\langle \mu^*, X_{t,a}\rangle)\}_{a=1}^K$ are i.i.d. uniform on $[0,1]$ and*

$$\mathbb{P}(a_t^* = a \,|\, \mathcal{H}_{t-1}^-) = \frac{1}{K} \qquad \forall a \in [K], \tag{7.1}$$

*for any GMF policy $\{a_t^*\}_{t=1}^{\infty}$. Note, the randomness lies exclusively in the current contexts $\{X_{t,a}\}_{a=1}^{K}$.*

*Proof.* Let $\psi_a := \mathcal{F}_a(\langle \mu^*, X_{t,a}\rangle)$. From the assumptions $\{\psi_a\}_{a=1}^K$ are i.i.d random variables, independent from $\mathcal{H}_{t-1}^-$, with uniform distribution on $[0,1]$ (see (Casella and Berger, 2021, Theorem 2.1.10)). Hence $\forall a_1, a_2 \in [K]$: $\mathbb{P}(\psi_{a_1} = \psi_{a_2}) = 0$, $\mathbb{P}(a_t^* = a \,|\, \mathcal{H}_{t-1}^-) = \mathbb{P}(a_t^* = a)$ and

$$\mathbb{P}(a_t^* = a_1) = \mathbb{P}(\psi_{a_1} > \psi_{a'}, \forall a' \neq a_1) = \mathbb{P}(\psi_{a_2} > \psi_{a'}, \forall a' \neq a_2) = \mathbb{P}(a_t^* = a_2) = 1/K \ .$$

□

We call property (7.1) history-agnostic demographic parity since it states that, at each round, the policy selects all groups with equal probability regardless of the history. Recall that in our setup each arm corresponds to a sensitive group. Proposition 7.1.4 ensures that a GMF policy will keep exploring regardless of the history. This fact plays a key role in the design of our policy, which is greedy without the need of an exploration phase.

**Remark 7.1.5.** *Note that in the standard linear contextual bandit setting, the optimal policy $a_t^{\mathrm{opt}}$ does not necessarily satisfy Equation (7.1) even when we assume that $\{\langle \mu^*, X_a\rangle\}_{a=1}^K$ are independent and absolutely continuous. This is true since when the rewards of one arm are always lower than at least one of the other arms, that arm will never be selected by the optimal policy.*

In the following, we state and discuss the assumptions made for the analysis of our greedy policy.

**Assumption 7.1.6.** *Let $\mu^* \in \mathbb{R}^d$ be the underlying reward model. We assume that:*

(i) *The noise random variable $\eta_t$ is zero mean $R$-subgaussian, conditioned on $\mathcal{H}_{t-1}$.*

(ii) *Let $X_a$ be a random variable with values in $\mathbb{R}^d$ and such that $\|X_a\|_2 \leq L$ almost surely. For any $a \in [K]$, $\{X_{i,a}\}_{i=1}^T$ are i.i.d. copies of $X_a$.*

(iii) *The random variables $\{X_a\}_{a=1}^K$ are mutually independent.*

(iv) *For every $a \in [K]$, there exist $d_a \geq 1$, an absolutely continuous random variable $Y_a$ with values in $\mathbb{R}^{d_a}$ admitting a density $f_a$, $B_a \in \mathbb{R}^{d \times d_a}$ and $c_a \in \mathbb{R}^d$ such that $B_a^\top B_a = \mathbb{I}_{d_a}$,*

$$X_a = B_a Y_a + c_a \quad \text{and} \quad {\mu^*}^\top B_a \neq 0 \ .$$

Assumption 7.1.6(i) is a standard assumption on the noise in stochastic bandits. 7.1.6(ii) implies that the actions taken by the policy do not affect future contexts. This is needed to allow the learning of the distribution of rewards for each group and is also used in Chen et al. (2020); Li et al. (2019). 7.1.6(iv) implies that $\langle \mu^*, X_a\rangle$ is absolutely continuous and is satisfied when $X_a$ is absolutely continuous in a subspace of $\mathbb{R}^d$ which is not orthogonal to $\mu^*$ [1]. This fact combined with 7.1.6(iii) ensures that Proposition 7.1.4 holds. Assumptions 7.1.6(iii)-(iv) are specific to our setting and a current limitation of the analysis. Notice however, that 7.1.6(iii) is reasonable when the groups are sufficiently isolated, e.g. each context is sourced from a different country/group, while assuming that the rewards $\langle \mu^*, X_a \rangle$ are absolutely continuous is natural when the contexts contain continuous attributes. Furthermore 7.1.6(iv) allows $\mu^*$ to act differently on each group, similarly to the case when there is a different reward vector for each sensitive group. An example of this is showed in the simulation experiment in Section 7.4.

## 7.2 The fair-greedy policy

If Proposition 7.1.4 holds, then there is no arm with relative rank always strictly worse than the others and any learned policy with sub-linear fair pseudo-regret will select all arms with equal probability in the limit when the number of rounds goes to infinity. Hence, using confidence intervals will not help in decreasing the probability that one arm is selected. Furthermore, estimating the relative ranks $\{\mathcal{F}_a(\langle \mu^*, X_{t,a} \rangle)\}_{a=1}^K$ is challenging, since they are not directly observed and using the past noisy rewards $\{r_{i,a_i}\}_{i=1}^{t-1}$ to construct the empirical CDF for each group, similarly to Kearns et al. (2017), can be inaccurate due to the presence of noise.

For the reasons above, we propose the greedy approach in Algorithm 7, which uses the following two-stage procedure at each round $t$. First it assembles the previously selected contexts and corresponding rewards from iterate $1$ up to $\tilde{t} = \lfloor (t-1)/2 \rfloor$ (line 4) in order to construct an estimate $\mu_{\tilde{t}}$ of $\mu^*$ (line 5), which is a noisy version of the ridge regression estimate. Secondly, for each arm $a$, our policy computes an estimate of the relative rank $\mathcal{F}_a(\langle \mu^*, X_{t,a} \rangle)$, namely $\hat{\mathcal{F}}_{t,a}(\langle \mu_{\tilde{t}}, X_{t,a} \rangle)$, which is the empirical CDF value of $\langle \mu_{\tilde{t}}, X_{t,a} \rangle$ and is constructed using $\mu_{\tilde{t}}$ and the contexts from round $\tilde{t}+1$ up to $t$ (line 6). Lastly, it selects $a_t$ uniformly at random among the arms maximizing the relative rank estimate (line 7).

Fair-Greedy has two hyperparameters $\lambda$ and $\rho$, although the latter can be set arbitrarily small without affecting the regret. Moreover, it is greedy as at each time $t$, it always selects from the arms the one with the highest currently estimated relative rank. However, contrary to standard greedy approaches in bandits, Fair-Greedy does not require an initial exploration phase because it naturally explores all arms, as the following lemma and remark show.

**Lemma 7.2.1** (Fair-Greedy satisfies *information averaged demographic parity*)**.** *Let $a_t$ be the action taken by Fair-Greedy at time $t$ and let Assumption 7.1.6 be satisfied. Then, for all $t \geq 1$ we have*

$$\mathbb{P}(a_t = a) = \frac{1}{K} \quad . \tag{7.2}$$

[1] E.g. $X_a$ cannot be sum of random variables that are independent and absolutely continuous in orthogonal subspaces of $\mathbb{R}^d$.

**Algorithm 7** Fair-Greedy

```
Requires regularization parameter $\lambda > 0$ and noise magnitude $\rho \in (0,1]$ .
for $t = 1 \dots T$ do
    Receive contexts $\{X_{t,a}\}_{a=1}^{K}$
    Set $\tilde{t} = \lfloor (t-1)/2 \rfloor$, $X_{1:\tilde{t}} = (X_{1,a_1}, \dots, X_{\tilde{t},a_{\tilde{t}}})^\top$, $r_{1:\tilde{t}} = (r_{1,a_1}, \dots, r_{\tilde{t},a_{\tilde{t}}})$.
    If $\tilde{t} = 0$ set $\mu_{\tilde{t}} = 0$, else let $V_{\tilde{t}} := X_{1:\tilde{t}}^\top X_{1:\tilde{t}} + \lambda \mathbb{I}_d$, generate $\gamma_{\tilde{t}} \sim \mathcal{N}(0, \mathbb{I}_d)$ and compute
        $\mu_{\tilde{t}} := V_{\tilde{t}}^{-1} X_{1:\tilde{t}}^\top r_{1:\tilde{t}} + \frac{\rho}{d\sqrt{\tilde{t}}} \cdot \gamma_{\tilde{t}}$ .
    For each $a \in [K]$ compute
        $\hat{\mathcal{F}}_{t,a}(\langle \mu_{\tilde{t}}, X_{t,a} \rangle) := (t - 1 - \tilde{t})^{-1} \sum_{s=\tilde{t}+1}^{t-1} \mathbb{1}\left\{ \langle \mu_{\tilde{t}}, X_{s,a} \rangle \leq \langle \mu_{\tilde{t}}, X_{t,a} \rangle \right\}$ .
    Sample action
        $a_t \sim \mathcal{U}\big[ \arg\max_{a \in [K]} \hat{\mathcal{F}}_{t,a}(\langle \mu_{\tilde{t}}, X_{t,a} \rangle) \big]$ .
    Observe noisy reward $r_{t,a_t} = \langle \mu, X_{t,a_t} \rangle + \eta_t$.
end for
```

*Proof sketch (proof in Section 7.7).* The noise term in $\mu_{\tilde{t}}$ ensures that $\mu_{\tilde{t}}$ is absolutely continuous and hence $\mu_{\tilde{t}}^\top B_a \neq 0$ almost surely. Combining this with Assumption 7.1.6(iv) we obtain that $\langle \mu_{\tilde{t}}, X_a \rangle$ is also absolutely continuous (see Lemma 7.7.1). Moreover, thanks to Assumption 7.1.6(ii)(iii) we can show that the random variables in $\{\hat{\mathcal{F}}_{t,a}(\langle \mu_{\tilde{t}}, X_{t,a} \rangle)\}_{a=1}^{K}$ are i.i.d. when conditioned on $\mu_{\tilde{t}}$. Note that $a_t$ is sampled uniformly form the argmax of i.i.d. random variables, when conditioned on $\mu_{\tilde{t}}$, which implies $\mathbb{P}(a_t = a \,|\, \mu_{\tilde{t}}) = 1/K$. The statement follows by taking the expectation over $\mu_{\tilde{t}}$. □

**Remark 7.2.2.** *It is easy to verify (through Lemma 7.2.1) that at any number of rounds $T$, the Fair-Greedy policy selects in expectation $T/K$ candidates from every group, i.e.* $\mathbb{E}\big[ \sum_{t=1}^{T} \mathbb{1}\{a_t = a\} \big] = \frac{T}{K}$ *for every $a \in [K]$. This also holds for the GMF policy and the one selecting arms uniformly at random.*

Since $\mathbb{P}(a_t = a) = \mathbb{E}_{\mathcal{H}_{t-1}}[\mathbb{P}(a_t = a \,|\, \mathcal{H}_{t-1})]$, with $\mathcal{H}_{t-1}$ being the information available to the policy before making a decision at round $t$, we call the property in (7.2) information-averaged demographic parity, which is weaker than history-agnostic demographic parity (in (7.1)). However, our analysis still requires a lower bound on $\mathbb{P}(a_t = a \,|\, \mathcal{H}_{t-1}^{-})$ which is presented in the next section.

**Remark 7.2.3** (Computational cost of Fair-Greedy)**.** *Compared to common linear contextual bandits approaches based on ridge regression, Algorithm 7 has an higher computational and memory cost which grow linearly with $t$. $\mu_{\tilde{t}}$ requires us to compute the product of $V_{\tilde{t}}^{-1}$ and $X_{1:\tilde{t}}^\top r_{1:\tilde{t}}$, which can be stored using $d^2$ and $d$ values respectively and updated online (via sherman-morrison (Hager, 1989)). However, Algorithm 7 also requires, at each round $t$, to* *keep in memory $K(t-1-\tilde{t})$ $d$-dimensional contexts and to compute the same number of scalar products to construct the empirical CDF for all $K$ groups.*

## 7.3 Regret analysis

In this section we present the analysis leading to the high probability $\tilde{O}(K^3 + \sqrt{dT})$ upper bound on the fair pseudo-regret of the greedy policy in Algorithm 7. We start by showing two key properties of CDF functions in the following lemma (proof in Section 7.7). Recall that for a continuous random variable $Z$ we denote by $f_Z$ the associated probability density function (PDF).

**Lemma 7.3.1.** *Let Assumption 7.1.6(iv) hold and set $\forall a \in [K]$, $Z_a := \langle \mu^*, X_a \rangle$ so that $\mathcal{F}_a = \mathcal{F}_{Z_a}$ and $M := \max_{a\in[K], z\in\mathbb{R}} f_{Z_a}(z) < +\infty$ as the maximum PDF value of the rewards of all groups. Then, the following two statements are true.*

(i) *$\mathcal{F}_a$ is Lipschitz continuous for every $a \in [K]$, and in particular for any $r, r' \in \mathbb{R}$ we have*

$$\sup_{a\in[K]} |\mathcal{F}_a(r) - \mathcal{F}_a(r')| \leq M|r - r'| \enspace .$$

(ii) *For every $a \in [K]$, let $\mu \in \mathbb{R}^d$, $\tilde{Z}_a := \langle \mu, X_a \rangle$. Then we have*

$$\sup_{a\in[K], r\in\mathbb{R}} |\mathcal{F}_a(r) - \mathcal{F}_{\tilde{Z}_a}(r)| \leq 2M \left\| \mu^* - \mu \right\| \left\| x_{\max} \right\|_* \enspace ,$$

*for any norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$ , where $\|x_{\max}\|_* := \sup_{x\in\cup_{a=1}^K \mathrm{Supp}(X_a)} \|x\|_*$ and $\mathrm{Supp}(X_a)$ is the support of the random variable $X_a$ .*

Lemma 7.3.1(i) bounds the Lipschitz constant of $\mathcal{F}_a$ and its derivation is straightforward. Lemma 7.3.1(ii) is needed since we only have access to an estimate of $\mu^*$, which will take the role of $\mu$. Its derivation is more subtle and could be of independent interest. By using Lemma 7.3.1 and the Dvoretzky–Kiefer–Wolfowitz-Massart (DKWM) inequality Dvoretzky et al. (1956); Massart (1990) to bound the gap between CDF and empirical CDF, we obtain the following result.

**Lemma 7.3.2** (Instant regret bound)**.** *Let Assumption 7.1.6(ii)(iv) hold and $a_t$ to be generated by Algorithm 7. Then with probability at least $1-\delta/4$, for all $t$ such that $3 \leq t \leq T$ we have*

$$\mathcal{F}_{a_t^*}(\langle \mu^*, X_{t,a_t^*} \rangle) - \mathcal{F}_{a_t}(\langle \mu^*, X_{t,a_t} \rangle) \leq 6M \left\| \mu^* - \mu_{\tilde{t}} \right\|_{V_{\tilde{t}}} \left\| x_{\max} \right\|_{V_{\tilde{t}}^{-1}} + 2\sqrt{\frac{\log(8KT/\delta)}{t-1}} \enspace ,$$

*where $\|x_{\max}\|_{V_{\tilde{t}}^{-1}} := \sup_{x\in\cup_{a=1}^K \mathrm{Supp}(X_a)} \|x\|_{V_{\tilde{t}}^{-1}}$.*

*Proof.* Let $Z_t := \langle \mu_{\tilde{t}}, X_{a_t} \rangle$, $Z_t^* := \langle \mu_{\tilde{t}}, X_{a_t^*} \rangle$ and $\mathcal{F}_{Z_t}$, $\mathcal{F}_{Z_t^*}$ be their CDF conditioned on $\mu_{\tilde{t}}$, $a_t$, and $a_t^*$. Let also $R_{\text{inst}}(t) := \mathcal{F}_{a_t^*}(\langle \mu^*, X_{t,a_t^*}\rangle) - \mathcal{F}_{a_t}(\langle \mu^*, X_{t,a_t}\rangle)$ . Then we can write

$$\begin{aligned} R_{\text{inst}}(t) = &\underbrace{\mathcal{F}_{a_t^*}(\langle \mu^*, X_{t,a_t^*}\rangle) - \mathcal{F}_{a_t^*}(\langle \mu_{\tilde{t}}, X_{t,a_t^*}\rangle)}_{\text{(I)}} + \underbrace{\mathcal{F}_{a_t^*}(\langle \mu_{\tilde{t}}, X_{t,a_t^*}\rangle) - \mathcal{F}_{Z_t^*}(\langle \mu_{\tilde{t}}, X_{t,a_t^*}\rangle)}_{\text{(II)}} \\ &+ \underbrace{\mathcal{F}_{Z_t^*}(\langle \mu_{\tilde{t}}, X_{t,a_t^*}\rangle) - \hat{\mathcal{F}}_{t,a_t^*}(\langle \mu_{\tilde{t}}, X_{t,a_t^*}\rangle)}_{\text{(III)}} + \underbrace{\hat{\mathcal{F}}_{t,a_t^*}(\langle \mu_{\tilde{t}}, X_{t,a_t^*}\rangle) - \hat{\mathcal{F}}_{t,a_t}(\langle \mu_{\tilde{t}}, X_{t,a_t}\rangle)}_{\text{(IV)}} \\ &+ \underbrace{\hat{\mathcal{F}}_{t,a_t}(\langle \mu_{\tilde{t}}, X_{t,a_t}\rangle) - \mathcal{F}_{Z_t}(\langle \mu_{\tilde{t}}, X_{t,a_t}\rangle)}_{\text{(V)}} + \underbrace{\mathcal{F}_{Z_t}(\langle \mu_{\tilde{t}}, X_{t,a_t}\rangle) - \mathcal{F}_{a_t}(\langle \mu_{\tilde{t}}, X_{t,a_t}\rangle)}_{\text{(VI)}} \\ &+ \underbrace{\mathcal{F}_{a_t}(\langle \mu_{\tilde{t}}, X_{t,a_t}\rangle) - \mathcal{F}_{a_t}(\langle \mu^*, X_{t,a_t}\rangle)}_{\text{(VII)}} \ . \end{aligned}$$

Since $a_t$ is chosen greedily in Algorithm 7 we have (IV) $\leq 0$. Then, applying Lemma 7.3.1(i), Cauchy-Schwarz and $\|X_{t,a}\|_* \leq \|x_{\max}\|_*$ for (I) and (VII) and Lemma 7.3.1(ii) for (II) and (VI), we obtain

$$\text{(I) + (VII)} \leq 2M \|\mu^* - \mu_{\tilde{t}}\| \|x_{\max}\|_* \quad , \qquad \text{(II) + (VI)} \leq 4M \|\mu^* - \mu_{\tilde{t}}\| \|x_{\max}\|_* \quad .$$

By noticing that $\hat{\mathcal{F}}_{t,a}(\cdot)$ is the empirical CDF of the random variable $\langle \mu_{\tilde{t}}, X_a\rangle$ conditioned to $\mu_{\tilde{t}}$, we can bound (III) and (V) directly using the DKWM inequality (see Lemma 7.7.4), which gives that with probability at least $1 - \delta/4$ and for all $t$ such that $3 \leq t \leq T$ we have

$$\text{(III) + (V)} \leq 2\sqrt{\frac{\log(8KT/\delta)}{t-1}} \quad .$$

We conclude the proof by combining the previous bounds and setting $\|\cdot\| = \|\cdot\|_{V_{\tilde{t}}}$ . □

We proceed by controlling the term $\|\mu^* - \mu_{\tilde{t}}\|_{V_{\tilde{t}}} \|x_{\max}\|_{V_{\tilde{t}}^{-1}}$ in Lemma 7.3.2. The quantity $\|\mu^* - \mu_{\tilde{t}}\|_{V_{\tilde{t}}}$ can be bounded using the OFUL confidence bounds (Abbasi-Yadkori et al., 2011, Theorem 2), since the noise term in $\mu_{\tilde{t}}$ decreases at an appropriate rate. Controlling $\|x_{\max}\|_{V_{\tilde{t}}^{-1}}$ requires instead different results than the ones in Abbasi-Yadkori et al. (2011), since it depends on the distributions of $\{X_a\}_{a=1}^K$ and not only on previous contexts and rewards. Hence, to provide an upper bound for $\|x_{\max}\|_{V_{\tilde{t}}^{-1}}$ which decreases with $t$, we also rely on Assumption 7.1.6(iii) and the structure of algorithm 7, which enable the following history-agnostic lower bound on the probability of selecting one arm.

**Proposition 7.3.3.** *Let Assumption 7.1.6 hold, $a_t$ be generated by Algorithm 7 and $c \in [0,1)$. Then with probability at least $1-\delta/4$, for all $a \in [K]$ and all $t \geq 3 + 8\log^{3/2}(5K\,\mathrm{e}/\delta)/(1 - \sqrt[K]{c})^3$ we have*

$$\mathbb{P}(a_t = a \,|\, \mathcal{H}_{t-1}^-) \geq \frac{c}{K} \quad ,$$

*where we recall that $\mathcal{H}_t^- = \cup_{i=1}^t \left\{\{X_{i,a}\}_{a=1}^K, r_{i,a_i}, a_i\right\}$.*

*Proof sketch (proof in Section 7.7).* For any $a \in [K]$, let $\hat{r}_{t,a} = \langle \mu_{\tilde{t}}, X_{t,a} \rangle$ be the estimated reward for arm $a$ at round $t$, denote with $\mathcal{F}_{\hat{r}_{t,a}}$ the CDF of $\hat{r}_{t,a}$ conditioned on $\mu_{\tilde{t}}$, and let

$$\phi_{t,a} := \mathcal{F}_{\hat{r}_{t,a}}(\hat{r}_{t,a}) \ , \quad \text{and} \quad \hat{\phi}_{t,a} := \hat{\mathcal{F}}_{t,a}(\hat{r}_{t,a}) \ ,$$

where $\hat{\mathcal{F}}_{t,a}(\hat{r}_{t,a})$ is defined in line 6 of Algorithm 7. Now, by the definition of $a_t$ (line 7 of Algorithm 7), we have

$$\mathbb{P}(a_t = a \,|\, \mathcal{H}_{t-1}^{-}) = \sum_{m=1}^{K} \frac{1}{m} \mathbb{P}(a \in C_t, |C_t| = m \,|\, \mathcal{H}_{t-1}^{-}) \ ,$$

where we introduced $C_t := \arg\max_{a \in [K]} \hat{\phi}_{t,a}$. Let $\epsilon_t > 0$ and continue the analysis conditioning on the events where $\sup_{a \in [K]} |\phi_{t,a} - \hat{\phi}_{t,a}| \leq \epsilon_t$. Then, we can write

$$\mathbb{P}(a_t = a \,|\, \mathcal{H}_{t-1}^{-}) \geq \mathbb{P}(\hat{\phi}_{t,a} > \hat{\phi}_{t,a'} \,,\, \forall a' \neq a \,|\, \mathcal{H}_{t-1}^{-}) \geq \mathbb{P}(\phi_{t,a} > \phi_{t,a'} + 2\epsilon_t \,,\, \forall a' \neq a \,| \mathcal{H}_{t-1}^{-}) \ ,$$

where in the first inequality we considered the case when $a \in C_t$ and $|C_t| = 1$, and in the second inequality we considered the worst case scenario where $\hat{\phi}_{t,a} = \phi_{t,a} - \epsilon_t$ and $\hat{\phi}_{t,a'} = \phi_{t,a'} + \epsilon_t$. Assumption 7.1.6(iv) and the additive noise in $\mu_{\tilde{t}}$ imply that $\langle \mu_{\tilde{t}}, X_a \rangle$ is an absolutely continuous random variable for each $a \in [K]$, which yields that $\{\phi_{t,a}\}_{a \in [K]}$ is uniformly distributed on $[0, 1]$. Furthermore, $\{\phi_{t,a}\}_{a \in [K]}$ are also independent due to Assumption 7.1.6(iii). Thus we have

$$\mathbb{P}(a_t = a \,|\, \mathcal{H}_{t-1}^{-}) \geq \int_0^1 \big(\mathbb{P}(\phi_{t,a'} < \mu - 2\epsilon_t)\big)^{K-1} \, \mathrm{d}\mu = \int_{2\epsilon_t}^1 (\mu - 2\epsilon_t)^{K-1} \, \mathrm{d}\mu = \frac{(1 - 2\epsilon_t)^K}{K} \ .$$

Finally, thanks to Assumption 7.1.6(ii) we can invoke the DKWM inequality to appropriately bound $\epsilon_t$ in high probability for all $t$ sufficiently large. □

The property in Proposition 7.3.3 guarantees that, for sufficiently large $t$, the policy can get arbitrarily close to satisfy history-agnostic demographic parity in (7.2). In particular this allows us to control $\|x_{\max}\|_{V_{\tilde{t}}^{-1}}$ by using a standard matrix concentration inequality (Tropp, 2011, Theorem 3.1) on a special decomposition of $V_{\tilde{t}}$, thereby enabling the following result (proof in Section 7.7).

**Lemma 7.3.4.** *Let Assumption 7.1.6 hold, $a_t$ be generated by Algorithm 7, $\tau_1 = 32K^3 \log^{3/2}\big(5K\,\mathrm{e}/\delta\big)$, $\tau_2 = \frac{54L^2}{\lambda_{\min}^{+}(\Sigma)} \log(\frac{4d}{\delta})$ and $\tau = 4\max(\tau_1, \tau_2) + 3$. Then, with probability at least $1 - \frac{3\delta}{4}$, for all $t \geq \tau$ we have*

$$\|\mu^* - \mu_{\tilde{t}}\|_{V_{\tilde{t}}} \|x\|_{V_{\tilde{t}}^{-1}} \leq \frac{8L}{\sqrt{\lambda_{\min}^{+}(\Sigma) \cdot t}} \Big[ b_1 \sqrt{d \log((8 + 4t\max(L^2/\lambda, 1))/\delta)} + \lambda^{\frac{1}{2}} \|\mu^*\|_2 \Big] \ ,$$

*where $b_1 = \lambda^{\frac{1}{2}} + R + L$, $\Sigma := K^{-1} \sum_{a=1}^{K} \mathbb{E}\big[X_a X_a^\top\big]$ and $\lambda_{\min}^{+}(\Sigma)$ is its smallest nonzero eigenvalue.*

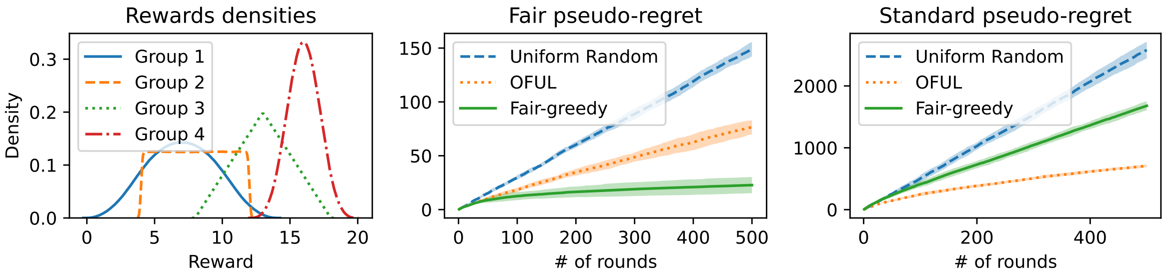

Figure 7.1: **Simulation Results**. First image is a density plot of the reward distributions while the second and third plot show the standard and fair pseudo-regrets, with mean (solid lines) ± standard deviation (shaded region) over 10 runs. To approximate the true reward CDF for each group we use the empirical CDF with $10^7$ samples.

Finally we obtain the desired high probability regret bound by combining Lemma 7.3.2 with Lemma 7.3.4 and summing over the $T$ rounds (see Section 7.7 for a proof).

**Theorem 7.3.5.** *Let Assumption 7.1.6 hold and $a_t$ be generated by Algorithm 7. Then, with probability at least $1-\delta$, for any $T \geq 1$ we have*

$$R_{\mathsf{F}}(T) \leq \frac{96ML}{\sqrt{\lambda^{+}_{\min}(\Sigma)}}\left[(\lambda^{\frac{1}{2}} + R + L)\sqrt{dT\log((8+4T\max(L^2/\lambda,1))/\delta)} + \sqrt{\lambda T}\left\|\mu^*\right\|_2\right]$$
$$+ 8\sqrt{\frac{T\log(8KT/\delta)}{3}} + \tau \enspace ,$$

*with $\tau$ defined in Lemma 7.3.4. Hence $R_{\mathsf{F}}(T) = O(K^3\log^{3/2}(K/\delta) + \sqrt{dT\log(KT/\delta)})$.*

The regret bound in Theorem 7.3.5 has two terms. The $O(K^3\log^{3/2}(K))$ term describes the rounds needed to satisfy Proposition 7.3.3 with $c = 1/2$. The remaining part, which is of order $O(\sqrt{dT\log(KT)})$ is instead associated to the convergence of the empirical CDF and to the bandit performance. Indeed, it recalls the standard regret bound holding for finite-action linear contextual bandits Auer (2002); Chu et al. (2011); Lattimore and Szepesvári (2020).

## 7.4 Simulation with diverse reward distributions

We present an illustrative proof of concept experiment which simulates groups with diverse reward distributions. We set $K = 4$, $\eta_t = 2\xi_t$, where $\xi_t$ has standard normal distribution, $X_a = B_aY_a + c_a$ where each coordinate of $Y_a \in \mathbb{R}^4$ is an independent sample from the uniform distribution on $[0,1]$, $B_a \in R^{(4K+1)\times 4}$ is such that $X_a$ contains $Y_a$ starting from the $4a$-th coordinate and $c_a$ has all the coordinates set to zero except for the last which is set to $3a$ to simulate a group bias. In this setup $\mu^*$ acts differently on each group, in particular, we note that $\mu^* \in R^{4K+1}$ has its last coordinate multiplying the group bias in $c_a$, which we set to $1$, and $4$ group-specific coordinates, which we set to manually picked values between $0$ and $9$. Results are shown in Figure 7.1, where we compare our greedy policy in Algorithm 7 with OFUL Abbasi-Yadkori et al. (2011), both with regularization parameter set to $0.1$, and with the

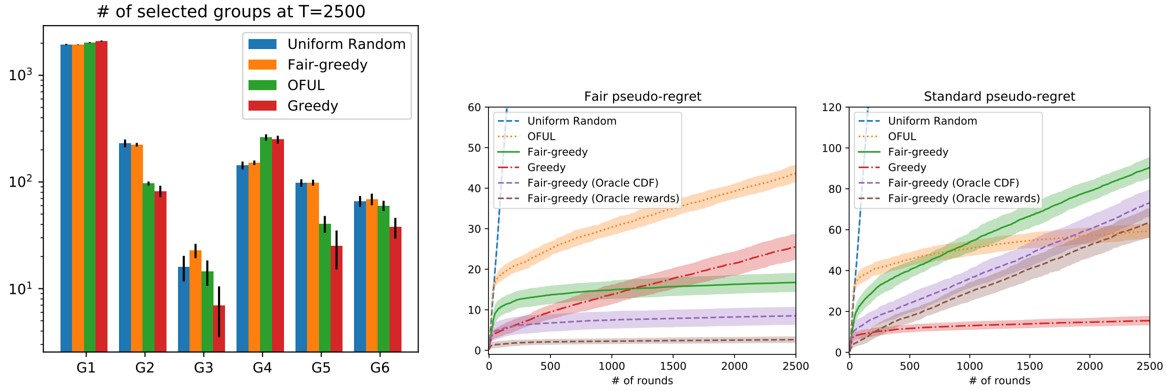

Figure 7.2: **US Census Results. Group** $=$ **Ethnicity**. First image shows mean (colored bars) and std (thinner black bars), while the other two show the mean (solid lines) $\pm$ standard deviation (shaded region) over 10 runs. To compute the reward CDF for each group we use the empirical CDF on $5K$ samples from $D2$. Percentage of selected groups is computed by dividing the number of candidates of a given group selected by the policy by the total number of candidates of that group received by the agent. G$X$ with $X \in \{1, \dots, 6\}$, stands for group $X$.

Uniform Random policy. We observe that, as expected from our analysis, our policy achieves sublinear fair pseudo-regret, while also having better-than random, although linear, standard regret. Additional details and an experiment on US census data with gender as the sensitive group are in Section 7.7.

## 7.5 Multiple candidates for each group

In this section, we analyze the more realistic case where contexts from a given arm do not necessarily belong to the same group. The complete analysis is presented in Section 7.7. In particular, we assume that at each round $t$, the agent receives $\{(X_{t,a}, s_{t,a})\}_{a=1}^{K}$, which are $K$ i.i.d. random variables where $s_{t,a} \in [G]$ is the sensitive group of the context $X_{t,a} \in \mathbb{R}^d$ and $G$ is the total number of groups. This setting can model for example a hiring scenario where at each round the employer has to choose among candidates belonging to different ethnic groups, some of which are minorities and hence have a small probability $\mathbb{P}(s_{t,a} = i)$ of being in the pool of received candidates. By naturally adapting the definition of fair-regret $R_{\mathsf{F}}(T)$, the Fair-Greedy policy and Assumption 7.1.6 to this setting, with probability $1 - \delta$ we obtain the following regret bound (see Corollary 7.7.16 in Section 7.7).

$$R_{\mathsf{F}}(T) = O\left(\frac{G\log(GT/\delta)}{Kq_{\min}} + \frac{(KG)^{3/2}\log^{3/2}(G/\delta)}{q_{\min}^{3/2}} + \sqrt{\frac{dT\log\left(GT/\delta\right)}{(1+K/G)q_{\min}}}\right) \tag{7.3}$$

where $q_{\min} = \min_{i\in[G]} \mathbb{P}(s_{t,a} = i)G$, so that $q_{\min} = 1$ if and only if each group has equal probability of being sampled and $q_{\min} > 0$ without loss of generality. (7.3) is similar to Theorem 7.3.5, having the same dependency on $\delta$ and $T$ but an improved dependency on the number of arms $K$ when $K > G$, since contexts from all arms can be used to estimate the

CDF of each group. The first term in (7.3) comes from the application of the Chernoff bound to lower bound the number of candidates in each group received by the agent, which is now random.

**US Census experiments. Group = Ethnicity.** We test this setting in practice by simulating the hiring scenario discussed above with data from the US Census containing the income and other useful indicators of several individuals in the United States. This data is accessed via the FolkTables library (Ding et al., 2021). In particular, at each round, we sample $K = 10$ candidates at random from the population containing the $G = 6$ largest ethnic groups[2], the reward is a previously computed linear estimate of the income, while the noisy reward is the true reward plus some small gaussian noise. We compare the Fair-Greedy Policy with OFUL (Abbasi-Yadkori et al., 2011), Greedy (selects the candidate with the best estimated reward) and Uniform Random in Figure 7.2. Similarly to the synthetic experiment in Section 7.4, the Fair-greedy policy achieves the best fair pseudo-regret and standard regret better than Uniform Random. Note that Greedy outperforms OFUL, which is too conservative in this scenario. Furthermore, the Fair-Greedy policy selects approximately the same percentage of candidates from each group, similarly to Uniform Random, while OFUL and Greedy select smaller percentages from G2, G3, G5 and G6. In Section 7.7 we provide more details and a comparison with two oracle fair policies which shows that knowing $\mu^*$ plays a more important role than knowing the true reward CDFs of each group.

## 7.6 Conclusions and future work

We introduced the concept of group meritocratic fairness in linear contextual bandits, which states that a fair policy should select, at each round, the candidate with the highest relative rank in the pool. This allows us to compare candidates coming from different sensitive groups, but it is hard to satisfy since the relative rank is not directly observed and depends on both the underlying reward model and on the rewards distribution for each group. After defining an appropriate fair pseudo-regret we analyzed a greedy policy and proved that its fair pseudo-regret is sublinear with high probability.

This result was possible since we restricted the analysis to the case where the contexts of different groups are independent random variables and the rewards are absolutely continuous. Relaxing these assumptions is a challenging avenue for future work. In particular, without the independence of contexts across arms, different approaches relying on confidence intervals might be necessary. Other two interesting directions are (i) to study the optimality of the proposed results and establishing lower bounds for any algorithm which minimises the fair pseudo-regret and (ii) to design a learning policy which aims at achieving a tradeoff between group meritocratic fairness and reward maximization.

[2] We remove groups with less than $5$K individuals to compute accurately the true CDFs for the fair regret.

## 7.7 Proofs and additional results

### Auxiliary lemmas

**Lemma 7.7.1.** *Let $n \in \mathbb{N}$, and assume that $Y, \nu$ are independent random variables in $\mathbb{R}^n$, such that $Y$ is absolutely continuous and $\nu \neq 0$, almost surely. Then, $\nu^\top Y$ is an absolutely continuous random variable.*

*Proof.* It is enough to show that for any $A \subset \mathbb{R}$ with zero Lebesgue measure, $\mathbb{P}(\nu^\top Y \in A) = 0$. Let $A \subseteq \mathbb{R}$, then we can write

$$\mathbb{P}(\nu^\top Y \in A) = \mathbb{E}[\mathbb{P}(\nu^\top Y \in A \,|\, \nu)] \ .$$

We proceed the proof by controlling the term $\mathbb{P}(\nu^\top Y \in A \,|\, \nu)$. We know that $\nu \neq 0$ almost surely. Now, since $Y$ and $\nu$ are independent, let $\nu = w$ for a fixed $w \in \mathbb{R}^n$ such that $w \neq 0$, then we have that

$$\mathbb{P}(w^\top Y \in A) = \int_{y \in \mathbb{R}^n} \mathbb{1}\left\{ \frac{w^\top}{\|w\|_2} Y \in A' \right\} f_Y(y) \, \mathrm{d}y \ ,$$

where we defined $A' := \left\{ \frac{x}{\|w\|_2} : x \in A \right\}$. Now consider the change of basis matrix $R = (v_1, \dots, v_n)^\top$, such that $v_1 = \frac{w}{\|w\|_2}$, with $RR^\top = \mathbb{I}_n$. By assigning $\hat{Y} = R^\top Y$, we can write

$$\mathbb{P}(w^\top Y \in A) = \int_{\hat{y} \in \mathbb{R}^n} \mathbb{1}\left\{ \hat{y}_1 \in A' \right\} f_Y(R\hat{y}) \, \mathrm{d}\hat{y} \ .$$

Since we assume that $Y$ is an absolutely continuous random variable, there exists $M_Y > 0$, such that $\sup_{y \in \mathbb{R}^n} f_Y(y) \leq M_Y$ almost surely, which allows us to write

$$\mathbb{P}(w^\top Y \in A) \leq M_Y \int_{\hat{y} \in \mathbb{R}^n} \mathbb{1}\left\{ \hat{y}_1 \in A' \right\} \, \mathrm{d}\hat{y} \ .$$

Finally, it is straightforward to check that if $A$ has a zero Lebesgue measure, then $A'$ also has a a zero Lebesgue measure, which gives $\mathbb{P}(w^\top Y \in A) = 0$. □

**Lemma 7.7.2** (Lipschitz CDF). *Let $n \in \mathbb{N}$, $\nu \in \mathbb{R}^n / \{0\}$ and $b \in \mathbb{R}$. Let also $Y$ be an absolutely continuous random variable with values in $\mathbb{R}^n$, with probability density function $f_Y$. Then the CDF of $Z = \langle \nu, Y \rangle + b$, namely $\mathcal{F}_Z$, is Lipschitz continuous. More specifically*

$$|\mathcal{F}_Z(r) - \mathcal{F}_Z(r')| \leq M' |r - r'| \qquad \forall r, r' \in \mathbb{R} \ ,$$

*where $M' = \max_{z \in \mathbb{R}} f_Z(z)$.*

*Proof.* Since $\nu \neq 0$ and $Y$ is absolutely continuous, $Z$ is also absolutely continuous with probability density $f_Z$ (see Lemma 7.7.1). Furthermore, if $r' \leq r$, we can write

$$\mathcal{F}_Z(r) - \mathcal{F}_Z(r') = \int_{-\infty}^{r} f_Z(t)\,\mathrm{d}t - \int_{-\infty}^{r'} f_Z(t)\,\mathrm{d}t = \int_{r'}^{r} f_Z(t)\,\mathrm{d}t \leq M'(r - r').$$

Applying the same reasoning to the case when $r \leq r'$ concludes the proof. □

**Lemma 7.7.3.** *Let $\{X_a\}_{a=1}^K$ be $K$ random variables with values in $\mathbb{R}^d$ and such that they are all $0$ with probability strictly less than one. Define $\Sigma = K^{-1}\sum_{a=1}^K \mathbb{E}[X_a X_a^\top]$ and let $\Sigma = USU^\top$ be its compact eigenvalue decomposition with $U \in \mathbb{R}^{d\times r}$, $S \in \mathbb{R}^{r\times r}$ with $1 \leq r \leq d$. Assume that $S$ is invertible. Then, for any $y \in \cup_{a=1}^K \mathrm{Supp}(X_a)$, we have $UU^\top y = y$ and $\lambda_{\min}^+(\Sigma)\left\|y\right\|_2^2 \leq y^\top \Sigma y$, where $\lambda_{\min}^+(\Sigma)$ is the smallest non-zero eigenvalue of the matrix $\Sigma$.*

*Proof.* Let $X$ be a random variable with the distribution $\mathbb{P}(X) = K^{-1}\sum_{a=1}^K \mathbb{P}(X_a)$. It is straightforward to check that $\Sigma = \mathbb{E}[XX^\top]$, and $y \in \mathrm{Supp}(X)$. We can also write $y = y_1 + y_2$ where $y_2 \in \mathrm{Ker}(\Sigma) := \{z \in \mathbb{R}^d \,:\, \Sigma z = 0\}$ and $y_1 \in \mathrm{Ker}(\Sigma)^\perp := \{z \in \mathbb{R}^d \,:\, \langle z, x\rangle = 0, \forall x \in \mathrm{Ker}(\Sigma)\}$. This implies that

$$y_2^\top \Sigma y_2 = \mathbb{E}[y_2^\top XX^\top y_2] = 0 \ .$$

Now, let $f(x) = (y_2^\top x)^2$. Then, $f(x) \geq 0$, for any $x \in \mathbb{R}^d$ and $f(y) = \|y_2\|_2^4$. Furthermore, since, $f(x)$ is a continuous function there exists $\epsilon > 0$, such that for any $z \in B(y,\epsilon) = \{x \in \mathbb{R}^d : \left\|x - y\right\|_2 < \epsilon\}$, $f(z) \geq \frac{\|y_2\|_2^4}{2}$. On the other hand, since $y \in \mathrm{Supp}(X)$, $\mathbb{P}(X \in B(y,\epsilon)) > 0$. Hence, we can write

$$0 = y_2^\top \Sigma y_2 = \mathbb{E}[f(X)] \geq \mathbb{E}[f(X)\mathbb{1}\left\{X \in B(y,\epsilon)\right\}] \geq \frac{\|y_2\|_2^4}{2}\mathbb{P}(X \in B(y,\epsilon)) \ ,$$

therefore $y_2 = 0$ which implies that $y \in \mathrm{Ker}(\Sigma)^\perp$. Since $UU^\top y$ is the orthogonal projection of $y$ onto $\mathrm{Ker}(\Sigma)^\perp$ we conclude that $y = UU^\top y$, $y^\top = y^\top UU^\top$ and

$$y^\top \Sigma y = y^\top USU^\top y \geq \lambda_{\min}^+(\Sigma) y^\top UU^\top y = \lambda_{\min}^+(\Sigma)\left\|y\right\|_2^2 \ .$$

□

## Proof of Lemma 7.2.1

*Proof.* If $t < 3$ then $\mu_{\tilde{t}} = 0$ and $a_t \sim \mathcal{U}[[K]]$ and the statement follows. If $t \geq 3$, Let $\mu \in \mathcal{S} := \{\mu' \in \mathbb{R}^d \,:\, \mu'^\top B_a \neq 0 \, \forall a \in [K]\}$, $\hat{r}_{i,a} = \langle \mu, X_{i,a}\rangle$ and $t' = t - \tilde{t} - 1$. Then by Lemma 7.7.1 $\hat{r}_{i,a}$ is absolutely continuous. Given a permutation of indices $j = (j_1, \ldots, j_{t'})$ where $j_i \in \{\tilde{t}+1, \ldots, t\}$, for $i \in [t']$. Let $\Omega_a$ be the set of the events of $\{X_{i,a}\}_{i=\tilde{t}+1}^t$ and $P$ be the set of all permutations of the indices $\{\tilde{t}+1, \ldots, t\}$. Consider the event

$$E_{a,j} = \{\omega \in \Omega_a \,:\, \hat{r}_{j_1,a} < \cdots < \hat{r}_{j_{t'},a}\,\} \ .$$

Since $\{\hat{r}_{i,a}\}_{i=\tilde{t}+1}^{t}$ are absolutely continuous, we have for all $k \neq i$, $\mathbb{P}(\hat{r}_{j_i,a} = \hat{r}_{j_k,a}) = 0$ and this yields $\Omega_a = \cup_{j \in P} E_{a,j}$ and $E_{a,j} \cap E_{a,j'} = \emptyset$ for all $j \neq j'$. Furthermore, since $\{\hat{r}_{i,a}\}_{i=\tilde{t}+1}^{t}$ are i.i.d. we have that $p_a := \mathbb{P}(E_{a,j}) = \mathbb{P}(E_{a,j'})$ for all $j \neq j'$. In particular, since $|P| = t'!$ we have $p_a = 1/(t'!)$.

Let $\phi_a = (t'-1)^{-1} \sum_{i=\tilde{t}+1}^{t-1} \mathbb{1}\{\hat{r}_{i,a} < \hat{r}_{t,a}\}$. Let $b \in \{0, \dots, t-1\}$ and let $P_b = \{j \in P : j_{b+1} = t\}$. We have that $|P_b| = (t'-1)!$ and

$$\mathbb{P}(\phi_a = b/(t'-1)) = \sum_{j \in P_b} \mathbb{P}(E_{a,j}) = (t'-1)! p_a = \frac{1}{t'} \ .$$

As a consequence, for all $a \in [K]$, $\phi_a$ is uniform over $\{0, 1/(t'-1), \dots, 1\}$. Since $\{r_{i,a}\}_{i \in [t+1], a \in [K]}$ are mutually independent we have that $\{\phi_a\}_{a \in [K]}$ are i.i.d. discrete uniform random variables. As a consequence, let $\hat{a} = \mathcal{U}\left[\arg\max_{a' \in [K]} \hat{\phi}_a\right]$ we have that $\mathbb{P}(\hat{a} = a) = 1/K$. Using the definition of $\hat{a}$ we have

$$\mathbb{P}(a_t = a) = \frac{1}{K}\mathbb{P}(\mu_{\tilde{t}} \in \mathcal{S}) + \mathbb{P}(a_t = a \,|\, \mu_{\tilde{t}} \in \mathcal{S}^c)\mathbb{P}(\mu_{\tilde{t}} \in \mathcal{S}^c) = \frac{1}{K} \ ,$$

where the last equality is derived by the fact that by the construction of $\mu_{\tilde{t}}$, $\mathbb{P}(\mu_{\tilde{t}} \in \mathcal{S}) = 1$. □

## Proofs of results in Section 7.3

The following result is used in the Proof of Lemma 7.3.2. Its proof is obtained by using the Dvoretzky–Kiefer–Wolfowitz-Massart inequality Dvoretzky et al. (1956); Massart (1990) combined with a union bound.

**Lemma 7.7.4.** *Let Assumption 7.1.6(ii) hold and $\hat{\mathcal{F}}_{t,a}(r)$, and $\mu_{\tilde{t}}$ to be generated by Algorithm 7. Let $Z_a := \langle \mu_{\tilde{t}}, X_a \rangle$ and denote with $\mathcal{F}_{Z_a}$ its CDF, conditioned on $\mu_{\tilde{t}}$. Then, with probability at least $1-\delta$ we have that for all $3 \leq t \leq T$*

$$\sup_{a \in [K], r \in \mathbb{R}} |\hat{\mathcal{F}}_{t,a}(r) - \mathcal{F}_{Z_a}(r)| \leq \sqrt{\frac{\log(2KT/\delta)}{t-1}} \ .$$

*Proof.* Let $3 \leq t \in T$ and recall that $\tilde{t} = \lfloor (t-1)/2 \rfloor$. Note that from Assumption 7.1.6(ii), for all $a \in [K]$, $\{\langle \mu_{\tilde{t}}, X_{i,a} \rangle\}_{i=\tilde{t}+1}^{t-1}$ are i.i.d copies of $Z_a$, conditioned on $\mu_{\tilde{t}}$. Since $\hat{\mathcal{F}}_{t,a}(r)$ is the empirical CDF of $Z_a$ conditioned on $\mu_{\tilde{t}}$, we can apply the Dvoretzky–Kiefer–Wolfowitz-Massart inequality Dvoretzky et al. (1956); Massart (1990) to obtain

$$\mathbb{P}\left(\sup_{r \in \mathbb{R}} |\hat{\mathcal{F}}_{t,a}(r) - \mathcal{F}_{Z_a}(r)| \geq \sqrt{\frac{\log(2/\delta')}{2(t-1-\tilde{t})}}\right) \leq \delta' \ .$$

Therefore, since $\tilde{t} \leq (t-1)/2$ we deduce that $\mathbb{P}\left[E_{t,a}\right] \leq \delta'$ , where

$$E_{t,a} = \left\{ \{X_{i,a}\}_{i=\tilde{t}+1}^{t-1}, \mu_{\tilde{t}} \, : \, \sup_{r \in \mathbb{R}} |\hat{\mathcal{F}}_{t,a}(r) - \mathcal{F}_{Z_a}(r)| \geq \sqrt{\frac{\log(2/\delta)}{t-1}} \right\} \ .$$

Consequently, by applying a union bound we obtain

$$\mathbb{P}\left[\cup_{i=1}^{T} \cup_{a=1}^{K} E_{t,a}\right] \le \sum_{i=1}^{T} \sum_{a=1}^{K} \mathbb{P}(E_{t,a}) \le KT\delta' \ ,$$

Finally, by substituting $\delta' = \delta/(KT)$ and computing the probability of the complement of $\cup_{i=1}^{T} \cup_{a=1}^{K} E_{t,a}$, we obtain the desired result. □

## Proof of Lemma 7.3.1

*Proof.* For every $a \in [K]$, by Assumption 7.1.6(iv), we have that $X_a = B_a Y_a + c_a$ where $Y_a \in \mathbb{R}^{d_a}$ is absolutely continuous with density $f_a$. Let $\nu^* := \mu^{*\top} B_a$, $\nu := \mu^\top B_a$, $b^* := \langle \mu^*, c_a \rangle$, $b := \langle \mu, c_a \rangle$. Then we have

$$Z_a = \langle \mu^*, X_a \rangle = \langle \nu^*, Y_a \rangle + b^* \ , \quad \text{and} \quad \tilde{Z}_a = \langle \mu, X_a \rangle = \langle \nu, Y_a \rangle + b \ .$$

From Assumption 7.1.6(iv) we also have that $\nu^* \neq 0$, hence, by applying Lemma 7.7.2 with $\nu = \nu^*$ and $Y = Y_a$ and by taking the maximum over $a \in [K]$, the statement (i) follows.

We now prove (ii). Since $Y_a$ is absolutely continuous we can write for any $r \in \mathbb{R}$

$$\begin{aligned} |\mathcal{F}_a(r) - \mathcal{F}_{\tilde{Z}_a}(r)| &= |\mathcal{F}_{\langle \nu^*, Y_a \rangle + b^*}(r) - \mathcal{F}_{\langle \nu, Y_a \rangle + b}(r)| \\ &\le \int_{y \in \mathbb{R}^{d_a}} \left| \mathbb{1}\left\{\langle \nu^*, y \rangle + b^* \le r\right\} - \mathbb{1}\left\{\langle \nu, y \rangle + b \le r\right\} \right| f_a(y) \, \mathrm{d}y \ . \end{aligned}$$

Now, by adding and subtracting $q(y) := \langle \nu^* - \nu, y \rangle + b^* - b$ and letting $r' := r - b^*$ we have

$$\begin{aligned} |\mathcal{F}_a(r) - \mathcal{F}_{\tilde{Z}_a}(r)| &\le \int_{y \in \mathbb{R}^{d_a}} \left| \mathbb{1}\left\{\langle \nu^*, y \rangle \le r'\right\} - \mathbb{1}\left\{\langle \nu^*, y \rangle \le r' + q(y)\right\} \right| f_a(y) \, \mathrm{d}y \\ &\le \int_{y \in \mathbb{R}^{d_a}} \mathbb{1}\left\{r' - |q(y)| \le \langle \nu^*, y \rangle \le r' + |q(y)|\right\} f_a(y) \, \mathrm{d}y \ . \end{aligned}$$

By Cauchy-Schwarz inequality, for any $y \in \mathrm{Supp}(Y_a)$, we get

$$|q(y)| \le |\langle \nu^* - \nu, y \rangle + b^* - b| \le |\langle \mu^* - \mu, B_a y + c_a \rangle| \le \|\mu^* - \mu\| \, \|x_{\max}\|_* \ ,$$

where we defined $\|x_{\max}\|_* := \max_{x \in \cup_{a=1}^{K} \text{Supp}(X_a)} \|x\|_* = \max_{y \in \cup_{a=1}^{K} \text{Supp}(Y_a)} \|B_a y + c_a\|_*$. We now let $\kappa = \|\mu^* - \mu\| \, \|x_{\max}\|_*$, and note that

$$|\mathcal{F}_a(r) - \mathcal{F}_{\tilde{Z}_a}(r)| \le \int_{y \in \mathbb{R}^{d_a}} \mathbb{1}\left\{r' - \kappa \le \langle \nu^*, y \rangle \le r' + \kappa\right\} f_a(y) \, \mathrm{d}y \ .$$

To control the above integral, we provide a proper change of variables. To this end, since by the assumption $\nu^* \neq 0$, we let $\{v_1, \ldots, v_{d_a}\}$ be an orthonormal basis of $\mathbb{R}^{d_a}$, with $v_1 = \nu^* / \|\nu^*\|_2$. Moreover, let $R = (v_1, \ldots, v_{d_a})$ to be the corresponding change of basis matrix. Then, for all $y \in \mathbb{R}^{d_a}$, we can always write $y = R\hat{y}$, where $\hat{y}_i = \langle y, v_i \rangle$, with $\hat{y}_1 = \frac{\langle \nu^*, y \rangle}{\|\nu^*\|_2}$. Hence we denote with $\hat{Y}_a = R^\top Y_a$ which now has the first coordinate parallel to $\nu^*$. Using the change of variables formula for multivariate integrals and noting that we are applying a rotation and hence $|\text{det}(R)| = 1$ and $f_{R\hat{Y}_a}(R\hat{y}) = f_{\hat{Y}_a}(\hat{y})/|\text{det}(R)| = f_{\hat{Y}_a}(\hat{y})$ we get

$$|\mathcal{F}_a(r) - \mathcal{F}_{\tilde{Z}_a}(r)| \leq \int_{\hat{y} \in \mathbb{R}^{d_a}} \mathbb{1}\left\{\frac{r' - \kappa}{\|\nu^*\|_2} \leq \hat{y}_1 \leq \frac{r' + \kappa}{\|\nu^*\|_2}\right\} f_{\hat{Y}_a}(\hat{y})\,\mathrm{d}\hat{y} \ .$$

Let $z = (\hat{y}_2, \dots, \hat{y}_{d_a})$. By Fubini's Theorem, and with the convention that $f_{\hat{Y}_a}(\hat{y}_1, z) = f_{\hat{Y}}(\hat{y}_1, z_1, \dots, z_{d_a - 1})$ we have

$$\begin{aligned}|\mathcal{F}_a(r) - \mathcal{F}_{\tilde{Z}_a}(r)| &\leq \int_{\hat{y}_1 \in \mathbb{R}} \mathbb{1}\left\{\frac{r' - \kappa}{\|\nu^*\|_2} \leq \hat{y}_1 \leq \frac{r' + \kappa}{\|\nu^*\|_2}\right\} \int_{z \in \mathbb{R}^{d_a - 1}} f_{\hat{Y}_a}(\hat{y}_1, z)\,\mathrm{d}z\,\mathrm{d}\hat{y}_1 \\ &= \int_{\hat{y}_1 \in \mathbb{R}} \mathbb{1}\left\{\frac{r' - \kappa}{\|\nu^*\|_2} \leq \hat{y}_1 \leq \frac{r' + \kappa}{\|\nu^*\|_2}\right\} f_{\hat{Y}_1}(\hat{y}_1)\,\mathrm{d}\hat{y}_1 \ ,\end{aligned}$$

where $f_{\hat{Y}_1}(\hat{y}_1) := \int_{z \in \mathbb{R}^{d_a - 1}} f_{\hat{Y}}(\hat{y}_1, z_1, \dots, z_{d_a - 1})\,\mathrm{d}z$ is the marginal density of $\hat{Y}_1 = \frac{\langle \nu^*, Y_a \rangle}{\|\nu^*\|_2}$, and we highlight that $\hat{Y}_1 = (Z_a - b^*)/\,\|\nu^*\|_2$. Finally note that

$$\max_{y_1 \in \mathbb{R}} f_{\hat{Y}_1}(y_1) = \|\nu^*\|_2 \max_{y_1 \in \mathbb{R}} f_{Z_a}(y_1) = \|\nu^*\|_2\, M \ ,$$

which yields

$$|\mathcal{F}_a(r) - \mathcal{F}_{\tilde{Z}_a}(r)| \leq 2\kappa M \ ,$$

and (ii) follows by substituting the definition of $\kappa$. □

### Proof of Proposition 7.3.3

*Proof.* Recall the definition of Algorithm 7. For any $a \in [K]$ let $\hat{r}_{t,a} = \mu_{\tilde{t}}^\top X_{t,a}$, which is the estimated reward for arm $a$, at round $t$. Note that $\mu_{\tilde{t}}$ and $X_{t,a}$ are independent random variables. Furthermore, denote with $\mathcal{F}_{\hat{r}_{t,a}}$ the CDF of $\hat{r}_{t,a}$ conditioned on $\mu_{\tilde{t}}$, and let

$$\phi_{t,a} := \mathcal{F}_{\hat{r}_{t,a}}(\hat{r}_{t,a}) \ , \quad \text{and} \quad \hat{\phi}_{t,a} := \hat{\mathcal{F}}_{t,a}(\hat{r}_{t,a}) \ .$$

Now, by the definition of the algorithm, we have

$$\mathbb{P}(a_t = a \,|\, \mathcal{H}_{t-1}^-) = \sum_{m=1}^{K} \frac{1}{m} \mathbb{P}(a \in C_t, |C_t| = m \,|\, \mathcal{H}_{t-1}^-) \ ,$$

where we introduced $C_t := \arg\max_{a\in[K]} \hat{\phi}_{t,a}$. Let $\epsilon_t > 0$ and continue the analysis conditioning on the events where $\sup_{a\in[K]} |\phi_{t,a} - \hat{\phi}_{t,a}| \leq \epsilon_t$. Then, we can write

$$\mathbb{P}(a_t = a \,|\, \mathcal{H}_{t-1}^-) \geq \mathbb{P}(\hat{\phi}_{t,a} > \hat{\phi}_{t,a'}\,,\, \forall a' \neq a \,|\, \mathcal{H}_{t-1}^-) \geq \mathbb{P}(\phi_{t,a} > \phi_{t,a'} + 2\epsilon_t\,,\, \forall\, a' \neq a \,|\mathcal{H}_{t-1}^-) \ ,$$

where in the first inequality we considered the case when $a \in C_t$ and $|C_t| = 1$. In the second inequality we considered the worst case scenario where $\hat{\phi}_{t,a} = \phi_{t,a} - \epsilon_t$ and $\hat{\phi}_{t,a'} = \phi_{t,a'} + \epsilon_t$. Recall that by the construction of the algorithm $\mu_{\tilde{t}} = V_{\tilde{t}}^{-1} X_{1:\tilde{t}}^\top r_{1:\tilde{t}} + (1/\sqrt{d\tilde{t}}) \cdot \gamma_{\tilde{t}}$. for all $a \in [K]$, the additive noise $(1/\sqrt{d\tilde{t}})\gamma_{\tilde{t}}$ assures that $\mu_{\tilde{t}}^\top B_a \neq 0$, almost surely. Therefore, by Lemma 7.7.1 $\hat{r}_{t,a} = \langle \mu_{\tilde{t}}, X_{t,a}\rangle$ conditioned on $\mu_{\tilde{t}}$ is absolutely continuous.

assumption 7.1.6(iii) and (Casella and Berger, 2021, Theorem 2.1.10) yield that $\{\phi_{t,a}\}_{a\in[K]}$ are independent and uniformly distributed on $[0,1]$ and in turn that

$$\mathbb{P}(a_t = a \,|\, \mathcal{H}_{t-1}^-) \geq \int_0^1 \big(\mathbb{P}(\phi_{t,a'} < \mu - 2\epsilon_t)\big)^{K-1} \,\mathrm{d}\mu = \int_{2\epsilon_t}^1 (\mu - 2\epsilon_t)^{K-1} \,\mathrm{d}\mu = \frac{(1-2\epsilon_t)^K}{K}. \tag{7.4}$$

We continue by computing an $\epsilon_t$ for which $\sup_{a\in[K]} |\phi_{t,a} - \hat{\phi}_{t,a}| \leq \epsilon_t$ holds with high probability. Observing that, conditioned on $\mu_{\tilde{t}}$, $\hat{\mathcal{F}}_{t,a}$ is the empirical CDF of $\mathcal{F}_{\hat{r}_{t,a}}$, we can use the Dvoretzky–Kiefer–Wolfowitz-Massart inequality to obtain, for any $a \in [K]$, $t \geq 3$, and $s \geq 0$

$$\mathbb{P}\left(|\phi_{t,a} - \hat{\phi}_{t,a}| \geq s\right) \leq 2\exp\left(-2s^2(t - \tilde{t} - 1)\right) \ .$$

Now, let $\tau_0 := 3 + 8\log^{3/2}\left(5K\,\mathrm{e}/\delta\right)/\left(1 - \sqrt[K]{c}\right)^3$. By applying the union bound, we can write

$$\begin{aligned}\mathbb{P}\left(\sup_{t\geq\tau_0, a\in[K]} |\phi_{t,a} - \hat{\phi}_{t,a}| \geq s\right) &\leq K \sum_{t=\tau_0}^{\infty} \mathbb{P}\left(|\phi_{t,a} - \hat{\phi}_{t,a}| \geq s\right) \\ &\leq 2K \sum_{t=\tau_0}^{\infty} \exp\left(-2s^2(t - \tilde{t} - 1)\right).\end{aligned}$$

Since $\tilde{t} = \lfloor \frac{t-1}{2} \rfloor$, it is straightforward to check that

$$\mathbb{P}\left(\sup_{t\geq\tau_0, a\in[K]} |\phi_{t,a} - \hat{\phi}_{t,a}| \geq s\right) \leq 2K \int_{t=\tau_0-1}^{\infty} \exp\left(-s^2 t\right)\,\mathrm{d}t \leq 2K s^{-2} \exp\left(-s^2(\tau_0 - 1)\right) \ .$$

Now, for any $\delta \in (0,1)$, by assigning $s = \sqrt{\frac{\log(4K(\tau_0-1)/\delta)}{\tau_0 - 1}}$, we get

$$\mathbb{P}\left(\sup_{t\geq\tau_0, a\in[K]} |\phi_{t,a} - \hat{\phi}_{t,a}| \geq \sqrt{\frac{\log(4K(\tau_0-1)/\delta)}{\tau_0 - 1}}\right) \leq \frac{\delta}{2\log\left(4K(\tau_0-1)/\delta\right)} \leq \frac{\delta}{4} \ , \tag{7.5}$$

where from $\tau_0 \geq 3, \delta < 1 \implies 4K(\tau_0 - 1)/\delta \geq 8 \geq e^2 \implies \log\left(4K(\tau_0-1)/\delta\right) \geq 2$ we obtain the last inequality. From (7.4), it follows that

$$\inf_{t\geq\tau_0, a\in[K]} \mathbb{P}\left(a_t = a|\mathcal{H}_{t-1}^-\right) \geq \frac{(1-2\sup_{t\geq\tau}\epsilon_t)^K}{K} \ .$$

Moreover, form (7.5), by letting $\epsilon_t = \sqrt{\frac{\log(4K(\tau_0-1)/\delta)}{\tau_0-1}}$, with probability at least $1-\frac{\delta}{4}$, we have

$$\inf_{t\geq\tau, a\in[K]} \mathbb{P}\left(a_t = a|\mathcal{H}_{t-1}^-\right) \geq \frac{1}{K}\left(1-2\sqrt{\underbrace{\frac{\log(4K(\tau_0-1)/\delta)}{\tau_0-1}}_{\text{(I)}}}\right)^K \ . \tag{7.6}$$

For the term (I) in the above, using $\log(x) \leq \log(5\,\mathrm{e}/4)x^{1/3}$ and $x \geq x^{2/3}$ for any $x \geq 1$ we deduce that

$$\text{(I)} = \frac{\log(4K/\delta)+\log(\tau_0-1)}{\tau_0-1} \leq \frac{\log(4K/\delta)+\log(5\,\mathrm{e}/4)}{(\tau_0-1)^{2/3}} = \frac{\log(5K\,\mathrm{e}/\delta)}{(\tau_0-1)^{2/3}} \ .$$

Now, by substituting $\tau_0 = 3+8\log^{3/2}\left(5K\,\mathrm{e}/\delta\right)/\left(1-\sqrt[K]{c}\right)^3$, we get that (I) $\leq \frac{1}{4}\left(1-\sqrt[K]{c}\right)^2$ and conclude the proof by plugging this inequality in (7.6). □

## Proof of Lemma 7.3.4

We start by establishing some required lemmas.

**Lemma 7.7.5.** *Let $\Sigma$, $\tau_1$, $\tau_2$ be defined in Lemma 7.3.4, $\tau_3 = \max(\tau_1,\tau_2)$, $\Sigma = USU^\top$ be the compact eigenvalue decomposition of $\Sigma$, with $U \in \mathbb{R}^{d\times r}$, $S \in \mathbb{R}^{r\times r}$ is a diagonal matrix with non-zero diagonal elements, and $U^\top U = \mathbb{I}_r$. Denote $\hat{S}_{t_0} = \sum_{i=1}^{\tilde{t}} U^\top X_{i,a_i}X_{i,a_i}^\top U$, where for $i \in [\tilde{t}]$, $a_i$ is given by Algorithm 7. Then with probability at least $1-\frac{\delta}{2}$, for any $t \geq 2\tau_3+3$ we have*

$$\lambda_{\min}\left(\hat{S}_{t_0}\right) \geq \frac{(\tilde{t}-\tau_3)\lambda_{\min}^+(\Sigma)}{4} \ .$$

*Proof.* Let $\tilde{S}_{t_0} := \sum_{i=1}^{\tilde{t}} \mathbb{E}\left[U^\top X_{i,a_i}X_{i,a_i}^\top U|\mathcal{H}_{i-1}^-\right]$. First, note that for any $\tau_3 \leq i \leq \tilde{t}$, we can write

$$\begin{aligned}\mathbb{E}\left[U^\top X_{i,a_i}X_{i,a_i}^\top U|\mathcal{H}_{i-1}^-\right] &= \sum_{a=1}^K \mathbb{E}\left[U^\top X_{i,a_i}X_{i,a_i}^\top U|\mathcal{H}_{i-1}^-, a_i = a\right]\mathbb{P}\left(a_i = a\,|\,\mathcal{H}_{i-1}^-\right) \\ &= \sum_{a=1}^K \mathbb{E}\left[U^\top X_{i,a}X_{i,a}^\top U\right]\mathbb{P}\left(a_i = a\,|\,\mathcal{H}_{i-1}^-\right) \ ,\end{aligned}$$

where the last equality holds based on the fact that $X_{i,a_i}X_{i,a_i}^\top$ conditioned on $a_i$, is independent from $\mathcal{H}_{i-1}^-$. Then, since $t \geq 3+64K^3\log^{3/2}\left(5K\,\mathrm{e}/\delta\right)$, by utilizing Proposition 7.3.3 with $c = \frac{1}{2}$

and noting that $1/(1-\sqrt[K]{1/2}) \leq 2K$ for all $K \geq 1$, with probability at least $1-\frac{\delta}{4}$, we have $\mathbb{P}(a_i = a \,|\, \mathcal{H}_{i-1}^{-}) \geq \frac{1}{2K}$. Therefore, with probability at least $1-\frac{\delta}{4}$, we obtain

$$\lambda_{\min}\left(\mathbb{E}\left[U^\top X_{i,a_i} X_{i,a_i}^\top U \,|\, \mathcal{H}_{i-1}^{-}\right]\right) \geq \frac{1}{2}\lambda_{\min}\left(K^{-1}\sum_{a=1}^{K} U^\top \mathbb{E}\left[X_a X_a^\top\right] U\right) = \frac{\lambda_{\min}^{+}(\Sigma)}{2} \ ,$$

and consequently, with probability at least $1-\frac{\delta}{4}$ we have

$$\lambda_{\min}(\tilde{S}_{\tilde{t}}) \geq \sum_{i=1}^{\tilde{t}} \lambda_{\min}(U^\top \mathbb{E}[X_{i,a_i} X_{i,a_i}^\top \,|\, \mathcal{H}_{i-1}^{-}] U) \geq (\tilde{t}-\tau_3)\cdot \frac{\lambda_{\min}^{+}(\Sigma)}{2} \ , \tag{7.7}$$

where in the last two displays we used the concavity attribute of the function $\lambda_{\min}(\cdot)$. Note that $\{X_{i,a_i}\}_{i=1}^{\infty}$, is an adaptive sequence with respect to the filtration $\{\mathcal{H}_i^{-}\}_{i=0}^{\infty}$, with

$$\left\|U^\top X_{i,a_i} X_{i,a_i}^\top U\right\|_{\text{op}} \leq \|X_{i,a_i}\|_2^2 \leq L^2 \ ,$$

for any $i \in [\tilde{t}]$. Let $\iota = \tilde{t} - \tau_3$. Now, by invoking (Tropp, 2011, Theorem 3.1) (with $\delta = \frac{1}{2}$ and $\mu = \lambda_{\min}^{+}(\Sigma)/2$, where $\delta, \mu$ are constants that appear in the latter theorem), we have

$$\mathbb{P}\left(\lambda_{\min}\left(\hat{S}_{t_0}\right) \leq \frac{\iota\lambda_{\min}^{+}(\Sigma)}{4} \quad \text{and} \quad \lambda_{\min}\left(\tilde{S}_{t_0}\right) \geq \frac{\iota\lambda_{\min}^{+}(\Sigma)}{2}\right) \leq d\cdot\left(\frac{e^{-\frac{1}{2}}}{\frac{1}{2}^{\frac{1}{2}}}\right)^{\frac{\iota\lambda_{\min}^{+}(\Sigma)}{4L^2}} \leq q \ ,$$

where we introduced $q = d\cdot\exp(-\frac{\iota\lambda_{\min}^{+}(\Sigma)}{27L^2})$, and we used the inequality $\mathrm{e}^{-\frac{1}{2}}\cdot\frac{1}{2}^{-\frac{1}{2}} \leq \mathrm{e}^{-\frac{4}{27}}$. Note that since $\tilde{t} \geq \tau_3 = \frac{54L^2}{\lambda_{\min}^{+}(\Sigma)}\log(\frac{4d}{\delta})$, we have $q \leq \frac{\delta}{4}$. Let $p = \mathbb{P}[\lambda_{\min}(\tilde{S}_{t_0}) \geq \frac{\iota\lambda_{\min}^{+}(\Sigma)}{2}]$, then we can write

$$\mathbb{P}\left(\lambda_{\min}\left(\hat{S}_{t_0}\right) \leq \frac{\iota\lambda_{\min}^{+}(\Sigma)}{4} \,\middle|\, \lambda_{\min}\left(\tilde{S}_{t_0}\right) \geq \frac{\iota\lambda_{\min}^{+}(\Sigma)}{2}\right) \leq \frac{\delta}{4p} \ ,$$

and accordingly

$$\mathbb{P}\left(\lambda_{\min}\left(\hat{S}_{t_0}\right) \geq \frac{\iota\lambda_{\min}^{+}(\Sigma)}{4} \quad \text{and} \quad \lambda_{\min}\left(\tilde{S}_{t_0}\right) \geq \frac{\iota\lambda_{\min}^{+}(\Sigma)}{2}\right) \geq 1-\frac{\delta}{2} \ ,$$

where we used $p \geq 1-\frac{\delta}{4}$, which follows from (7.7). Substituting $\iota = \tilde{t}-\tau_3$ gives the final result. □

**Lemma 7.7.6.** *Let $x \in \cup_{a=1}^{K}\mathrm{Supp}(X_a)$ and $\tau_3$ be defined in Lemma 7.7.5, then with probability at least $1-\frac{\delta}{2}$, for all $t \geq 2\tau_3+3$ we have*

$$\|x\|_{V_{\tilde{t}}^{-1}} \leq \frac{2L}{\sqrt{\lambda_{\min}^{+}(\Sigma)(\tilde{t}-\tau_3)}} \ .$$

*Proof.* Note that if $x = 0$ it is straightforward to check that the statement holds. So without loss of generality we assume that $x \in \mathfrak{S}$, where $\mathfrak{S} = \cup_{a=1}^{K} \mathrm{Supp}(X_{t,a}) - \{0\}$. Consider the compact singular value decomposition $\Sigma = USU^\top$ where $U \in \mathbb{R}^{d\times r}$, $S \in \mathbb{R}^{r\times r}$ is a diagonal matrix with non-zero diagonal elements (due to Assumption 7.1.6(iv)) and $U^\top U = \mathbb{I}_r$. Denote $\hat{S}_{\tilde{t}} = U^\top \hat{\Sigma}_{\tilde{t}} U$. For any $x \in \mathfrak{S}$ we have from Lemma 7.7.3 that $UU^\top x = x$, and $x^\top UU^\top = x^\top$. First, we claim that

$$U^\top (\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d)^{-1} U = (\hat{S}_{\tilde{t}} + \lambda \mathbb{I}_r)^{-1} \ .$$

To prove the above claim, it is enough to show that

$$(\hat{S}_{\tilde{t}} + \lambda \mathbb{I}_r) U^\top (\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d)^{-1} U = U^\top (\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d)^{-1} U (\hat{S}_{\tilde{t}} + \lambda \mathbb{I}_r) = \mathbb{I}_r \ .$$

Note that

$$\begin{aligned}
(\hat{S}_{\tilde{t}} + \lambda \mathbb{I}_r) U^\top (\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d)^{-1} U &= \left(U^\top \hat{\Sigma}_{\tilde{t}} U + \lambda \mathbb{I}_r\right) U^\top (\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d)^{-1} U \\
&= U^\top \left(\hat{\Sigma}_{\tilde{t}} UU^\top + \lambda \mathbb{I}_d\right) \left(\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d\right)^{-1} U \\
&= U^\top \left(\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d\right) \left(\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d\right)^{-1} U = \mathbb{I}_r \ .
\end{aligned}$$

With similar steps one can show that $U^\top (\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d)^{-1} U (\hat{S}_{\tilde{t}} + \lambda \mathbb{I}_r) = \mathbb{I}_r$, and therefore $U^\top (\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d)^{-1} U = (\hat{S}_{\tilde{t}} + \lambda \mathbb{I}_r)^{-1}$. By exploiting this fact, we can write

$$\begin{aligned}
\|x\|^2_{V_{\tilde{t}}^{-1}} &= \|x\|_2^2 \left(\frac{x}{\|x\|_2}^\top (\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d)^{-1} \frac{x}{\|x\|_2}\right) \\
&= \|x\|_2^2 \left(\frac{x}{\|x\|_2}^\top UU^\top (\hat{\Sigma}_{\tilde{t}} + \lambda \mathbb{I}_d)^{-1} UU^\top \frac{x}{\|x\|_2}\right) \\
&= \|x\|_2^2 \left(\frac{x}{\|x\|_2}^\top U (\hat{S}_{\tilde{t}} + \lambda \mathbb{I}_r)^{-1} U^\top \frac{x}{\|x\|_2}\right) \\
&= \|x\|_2^2 \left(\frac{x^\top U}{\|x^\top U\|_2} (\hat{S}_{\tilde{t}} + \lambda \mathbb{I}_r)^{-1} \frac{U^\top x}{\|U^\top x\|_2}\right) \ ,
\end{aligned}$$

where the second and last equations are results of Lemma 7.7.3, and consequently

$$\|x\|^2_{V_{\tilde{t}}^{-1}} \leq \frac{L^2}{\lambda_{\min}(\hat{S}_{\tilde{t}})} \ . \tag{7.8}$$

On the other hand, from Lemma 7.7.5, with probability at least $1 - \frac{\delta}{2}$, we have

$$\lambda_{\min}\left(\hat{S}_{t_0}\right) \geq \frac{(\tilde{t} - \tau_3)\lambda^+_{\min}(\Sigma)}{4} \ . \tag{7.9}$$

Finally, by combining (7.8) and (7.9) with probability at least $1-\frac{\delta}{2}$ we have

$$\|x\|^2_{V_{\tilde{t}}^{-1}} \leq \frac{4L^2}{\lambda^+_{\min}(\Sigma)(\tilde{t}-\tau_3)} \ .$$

□

**Lemma 7.7.7.** *With probability at least* $1-\frac{\delta}{4}$*, for all* $t \geq 3$ *we have*

$$\|\mu^* - \mu_{\tilde{t}}\|_{V_{\tilde{t}}} \leq (\lambda^{\frac{1}{2}} + R + L)\sqrt{d\log((8+8\tilde{t}\max(L^2/\lambda,1))/\delta)} + \lambda^{\frac{1}{2}}\|\mu^*\|_2 \ .$$

*Proof.* Recall that by the definition of Algorithm 7, we have $\mu_{\tilde{t}} = V_{\tilde{t}}^{-1}X_{1:\tilde{t}}^\top r_{1:\tilde{t}} + (1/d\sqrt{\tilde{t}})\cdot\gamma_{\tilde{t}}$. Therefore, we can write

$$\|\mu^* - \mu_{\tilde{t}}\|_{V_{\tilde{t}}} \leq \underbrace{\left\|\mu^* - V_{\tilde{t}}^{-1}X_{1:\tilde{t}}^\top r_{1:\tilde{t}}\right\|_{V_{\tilde{t}}}}_{\text{(I)}} + \underbrace{\frac{\rho}{d\sqrt{\tilde{t}}}\|\gamma_{\tilde{t}}\|_{V_{\tilde{t}}}}_{\text{(II)}} \ .$$

We proceed the proof by providing upper bounds for (I) and (II). For (I), by invoking (Abbasi-Yadkori et al., 2011, Theroem 2), with probability at least $1-\frac{\delta}{8}$, for all $t \geq 3$, which implies $\tilde{t} \geq 1$ we have

$$\begin{aligned}\text{(I)} &\leq R\sqrt{d\log((8+8\tilde{t}L^2/\lambda)/\delta)} + \lambda^{\frac{1}{2}}\|\mu^*\|_2 \\ &\leq R\sqrt{d\log((8+8\tilde{t}\max(L^2/\lambda,1))/\delta)} + \lambda^{\frac{1}{2}}\|\mu^*\|_2 \ .\end{aligned}$$

On the other hand, since $\rho \leq 1$, for term (II) we have

$$\text{(II)} \leq \frac{1}{d\sqrt{\tilde{t}}}\|V_{\tilde{t}}\|^{\frac{1}{2}}_{\text{op}}\|\gamma_{\tilde{t}}\|_2 \leq \frac{L+\lambda^{\frac{1}{2}}}{d}\|\gamma_{\tilde{t}}\|_2 \ .$$

$$\begin{aligned}\mathbb{P}\big(\text{(II)} \geq (L+\lambda^{\frac{1}{2}})\sqrt{\log(8d/\delta)}\big) &\leq \mathbb{P}\big(\|\gamma_{\tilde{t}}\|_2 \geq d\sqrt{\log(8d/\delta)}\big) \\ &\leq d\mathbb{P}\big(|\gamma_{1,\tilde{t}}| \geq \sqrt{\log(8d/\delta)}\big) \leq \frac{\delta}{8} \ .\end{aligned}$$

Thus, by applying the union bound with probability at least $1-\frac{\delta}{8}$, for all $t \geq 3$ we have

$$\text{(II)} \leq (L+\lambda^{\frac{1}{2}})\sqrt{\log(8\tilde{t}d/\delta)} \leq (L+\lambda^{\frac{1}{2}})\sqrt{d\log((8+8\tilde{t}\max(L^2/\lambda,1))/\delta)} \ .$$

□

*Proof of Lemma 7.3.4.* Recall that $\tau_3 = \max(\tau_1, \tau_2)$. From Lemma 7.7.6, with probability at least $1 - \frac{\delta}{2}$ for all $t \geq 2\tau_3 + 3$ we have

$$\|x\|_{V_{\tilde{t}}^{-1}} \leq \frac{2L}{\sqrt{\lambda_{\min}^{+}(\Sigma)(\tilde{t} - \tau_3)}} \quad .$$

From Lemma 7.7.7, with probability at least $1 - \frac{\delta}{4}$ for all $t \geq 3$

$$\|\mu^* - \mu_{\tilde{t}}\|_{V_{\tilde{t}}} \leq (\lambda^{\frac{1}{2}} + R + L)\sqrt{d \log((8 + 8\tilde{t} \max(L^2/\lambda, 1))/\delta)} + \lambda^{\frac{1}{2}} \|\mu^*\|_2 \quad .$$

Thus, combining Lemmas 7.7.6 and 7.7.7, with probability at least $1 - \frac{3\delta}{4}$ for all $t \geq 2\tau_3 + 3$ we have

$$\|\mu^* - \mu_{\tilde{t}}\|_{V_{\tilde{t}}} \|x\|_{V_t^{-1}} \leq \\ \frac{2L}{\sqrt{\lambda_{\min}^{+}(\Sigma)(\tilde{t} - \tau_3)}} \left( (\lambda^{\frac{1}{2}} + R + L)\sqrt{d \log((8 + 8\tilde{t} \max(L^2/\lambda, 1))/\delta)} + \lambda^{\frac{1}{2}} \|\mu^*\|_2 \right) \quad .$$

By the fact that $t \geq 4\tau_3 + 3$, we have $\tilde{t} \geq 2\tau_3$, which implies $\frac{1}{\sqrt{\tilde{t} - \tau_3}} \leq \sqrt{\frac{2}{\tilde{t}}}$. We conclude the proof by using the inequality $\tilde{t} \geq \frac{t-3}{2} \geq \frac{t}{8}$, for all $t \geq 4$. □

### Proof of Theorem 7.3.5

*Proof.* Combining Lemma 7.3.2 with Lemma 7.3.4 and using $1/(t-1) \leq 3/(4t)$ for all $t \geq 4$ we obtain, with probability at least $1 - \delta$ and for all $\tau \leq t \leq T$

$$\mathcal{F}_{a_t^*}(\langle \mu^*, X_{t,a_t^*} \rangle) - \mathcal{F}_{a_t}(\langle \mu^*, X_{t,a_t} \rangle) \leq 4\sqrt{\frac{\log(8KT/\delta)}{3t}} \\ + \frac{48ML}{\sqrt{\lambda_{\min}^{+}(\Sigma)t}} \left( (\lambda^{\frac{1}{2}} + R + L)\sqrt{d \log((8 + 4t \max(L^2/\lambda, 1))/\delta)} + \lambda^{\frac{1}{2}} \|\mu^*\|_2 \right) \quad .$$

By summing up the last inequality, with probability at least $1 - \delta$ we get

$$\sum_{t=\tau}^{T} \left[ \mathcal{F}_{a_t^*}(\langle \mu^*, X_{t,a_t^*} \rangle) - \mathcal{F}_{a_t}(\langle \mu^*, X_{t,a_t} \rangle) \right] \leq 8\sqrt{\frac{T \log(8KT/\delta)}{3}} \\ + \frac{96ML}{\sqrt{\lambda_{\min}^{+}(\Sigma)}} \left( (\lambda^{\frac{1}{2}} + R + L)\sqrt{dT \log((8 + 4T \max(L^2/\lambda, 1))/\delta)} + \sqrt{\lambda T} \|\mu^*\|_2 \right) . \tag{7.10}$$

where the last display is obtained by the inequality $\sum_{t=1}^{T} t^{-\frac{1}{2}} \leq 2T^{\frac{1}{2}}$. On the other hand, for $t \in [T]$, $\mathcal{F}_{a_t^*}(\langle \mu^*, X_{t,a_t^*} \rangle) - \mathcal{F}_{a_t}(\langle \mu^*, X_{t,a_t} \rangle) \leq 1$, and we can write

$$\sum_{t=1}^{\tau} \left[ \mathcal{F}_{a_t^*}(\langle \mu^*, X_{t,a_t^*} \rangle) - \mathcal{F}_{a_t}(\langle \mu^*, X_{t,a_t} \rangle) \right] \leq \tau \quad . \tag{7.11}$$

By combining (7.10) and (7.11), we conclude the proof. □

## Experiments

In this section we include additional details on the simulation experiments in Section 7.4 and an experiment on the US census data.

### Additional details on the simulation

We use the following value for the underlying linear model used in Figure 7.1.

$$\mu^* = (\underbrace{4, 3, 7, 0,}_{\text{Group 1}} \underbrace{8, 0, 0, 0,}_{\text{Group 2}} \underbrace{5, 5, 0, 0,}_{\text{Group 3}} \underbrace{2, 2, 2, 2,}_{\text{Group 4}} 1) \enspace .$$

Each slice of 4 coordinates of $\mu^*$ affects a different group. Furthermore, since each coordinate of $Y_a$ follows a standard uniform distribution, the resulting reward distributions for each group follow weighted variants of the Irwin-Hall distribution (Hall, 1927).

### Experiments on US census data

In this section, we present an experiment performed using the US Census data and the Falk-Tables library[3] Ding et al. (2021). In particular we construct a dataset with features similar to the UCI Adult dataset but where the target is the person's income instead of the binary variable indicating if the income is more or less than $50K$ dollars. We use this target as a possibly inaccurate proxy for how well a candidate will perform on the job, hence it will be used as the noisy reward for the bandit problem.

**Setup and Preprocessing.** To setup the bandit problem, we construct 2 datasets, namely $D1$ and $D2$, by selecting $500K$ males and $500K$ females random samples first from the $2017$ US Census Survey, to assemble $D1$, and then from the $2018$ survey to assemble $D2$. We use $D1$ to find mean and standard deviation for each feature and also for the target. After that we normalize features and target from $D2$ by subtracting the mean and dividing by the standard deviation previously computed on $D1$. We then construct $\mu^*$ as a ridge regression estimate on the samples from $D2$ with the regularization parameter equal to $10^{-8}$. The regression vector $\mu^*$ will be used to compute the (true) rewards for the samples. We construct the bandit problem with $K = 2$ arms/groups which correspond to the gender identities male and female. At each round, the context vectors of one male and one female candidate are sampled from $D2$ and after one of the two is selected by the policy, its corresponding noisy reward (i.e. its income) is received by the agent.

**Baselines.** We compare our method, namely *Fair-greedy* (Algorithm 7), with the following baselines.

[3] https://github.com/zykls/folktables

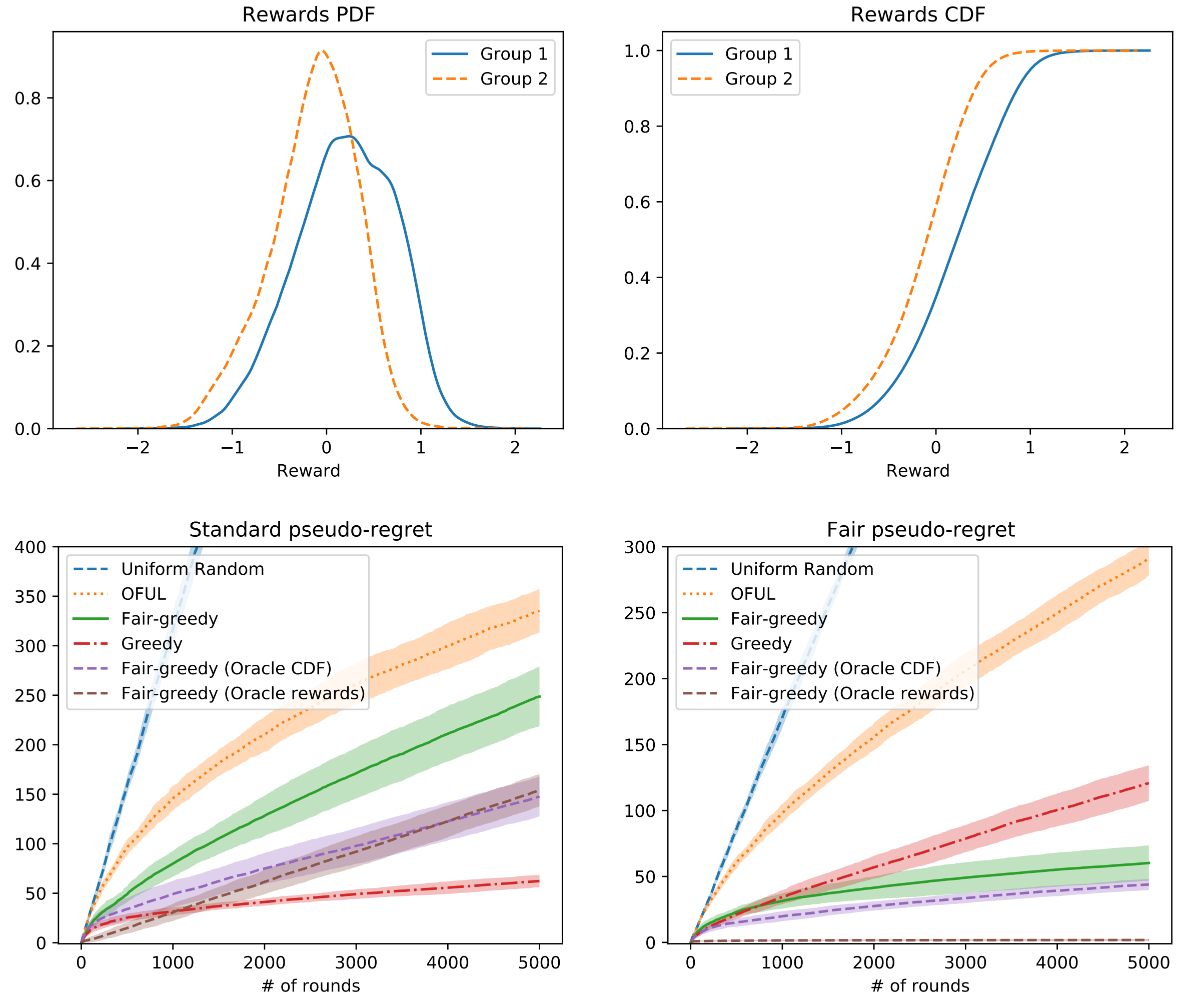

Figure 7.3: **US Census Results**. Top two images are density and CDF plots of the reward distributions while the bottom two plots are the standard and fair pseudo-regrets, with mean (solid lines) ± standard deviation (shaded region) over 10 runs. To compute the reward PDF and CDF for each group we use the empirical CDF on all $500K$ samples from $D2$.

- *Uniform Random*, which selects an arm uniformly at random at each round.
- *OFUL* Abbasi-Yadkori et al. (2011), with exploration parameter set to $0.1$.
- *Greedy*, which computes the ridge regression estimate for the reward vector using all the selected contexts and noisy rewards in the history and then selects the arm maximising the estimated reward.
- *Fair-greedy (Oracle CDF)*, which is a variant of *Fair-greedy* where all the selected contexts and noisy rewards in the history are used to compute the ridge regression estimate and the empirical CDF of each group is replaced by the true CDF.
- *Fair-greedy (Oracle rewards)*, which is another variant of *Fair-greedy* where the ridge regression estimate is replaced by the true reward model $\mu^*$ and all contexts in the

history are used to compute the empirical CDF for each group.

Note that the last two methods are *oracle methods* because they rely either on the true CDF of the rewards for each group or on $\mu^*$, which are unknown to the agent. All methods using a ridge regression estimate have the regularization parameter set to $0.1$. We observed that varying this parameter did not affect much the relative performance of the methods.

**Results.** The results and the reward distributions are illustrated in Figure 7.3. We note that in this case, *Greedy* performs much better than OFUL, which appears to be too conservative for this problem. In particular, the standard pseudo-regret of *Greedy* is unrivaled after $1000$ rounds. Furthermore, since there is a large overlap in the distributions of rewards, our *Fair-greedy* policy performs much better than the *Uniform Random* policy even in terms of standard pseudo-regret, while it outperforms all non-oracle methods in terms of fair pseudo-regret. As expected, the oracle methods both achieve a lower fair pseudo-regret than *Fair-greedy*, and we note that knowing only the underlying model $\mu^*$ is significantly more advantageous than knowing only the CDF for each group.

### Multiple candidates for each group

This section contains a rigorous treatement of the content in Section 7.5. We consider the more realistic case where contexts from a given arm do not necessarily belong to the same group. In particular, we assume that at each round $t$, the agent receives tuples $\{(X_{t,a}, s_{t,a})\}_{a=1}^K$, where $s_{t,a} \in [G]$ is the sensitive group of the context $X_{t,a} \in \mathbb{R}^d$ and $G$ is the total number of groups. After that the agent selects action $a_t$ and subsequently receives the noisy reward $\langle \mu^*, X_{t,a}\rangle + \eta_t$

Note that we recover the original setting discussed in Section 7.1 when $G = K$ and $s_{t,a} = a$ for every $a \in [K]$, $t \in \mathbb{N}$. A more realistic scenario is when $\{(X_{t,a}, s_{t,a})\}_{a=1}^K$ are i.i.d., and the distribution represents e.g. the underlying population of candidates, where $\mathbb{P}(s_{t,a} = i)$ is the same for all $a \in [K]$ and can be small when the group $i$ is a minority. The following analysis applies to both cases.

We impose the following assumption, which is a natural extension of Assumption 7.1.6.

**Assumption 7.7.8.** *Let $\mu^* \in \mathbb{R}^d$ be the underlying reward model. We assume that:*

(i) *The noise random variable $\eta_t$ is zero mean $R$-subgaussian, conditioned on $\mathcal{H}_{t-1}$.*

(ii) *For any $a \in [K]$, let $(X_a, s_a)$ be a random variable with values in $\mathbb{R}^d \times [G]$ and $\|X_a\|_2 \leq L$ almost surely. $\{(X_{t,a}, s_{t,a})\}_{t=1}^T$ are i.i.d. copies of $(X_a, s_a)$. $X_a$ conditioned to $s_a = i$ is a copy of the random variable $\hat{X}_i$ which is independent on the arm, for every $a \in [K]$.*

(iii) *For every $a \in [K]$ $X_a$ conditioned to $s_a$ is independent from $(X_{a'}, s_{a'})$ for any $a' \neq a$.*

(iv) *For every $i \in [G]$, then there exist $d_i \geq 1$, an absolutely continuous random variable $Y_i$ with values in $\mathbb{R}^{d_i}$ admitting a density $f_i$, $B_i \in \mathbb{R}^{d \times d_i}$ and $c_i \in \mathbb{R}^d$ such that $B_i^\top B_i = \mathbb{I}_{d_i}$,*

$$\hat{X}_i = B_i Y_i + c_i \quad \text{and} \quad {\mu^*}^\top B_i \neq 0 \ .$$

We define $\mathcal{F}(r, i) = \mathbb{P}(\langle \mu^*, \hat{X}_i \rangle \leq r) = \mathbb{P}(\langle \mu^*, X_a \rangle \leq r \,|\, s_a = i)$ for any $r \in \mathbb{R}$, $i \in G$. Hence we can extend the definition of group meritocratic fairness as follows.

**Definition 7.7.9** (GMF policy). *a policy $\{a_t^*\}_{t=1}^\infty$ is group meritocratic fair (GMF) if for all $t \in \mathbb{N}, a \in [K]$ it satisfies*

$$\mathcal{F}(\langle \mu^*, X_{t,a_t^*} \rangle, s_{a_t^*}) \geq \mathcal{F}(\langle \mu^*, X_{t,a} \rangle, s_{t,a_t}) \ .$$

The fair pseudo-regret is now defined as

$$R_F(T) = \sum_{t=1}^{T} \mathcal{F}(\langle \mu^*, X_{t,a_t^*} \rangle, s_{a_t^*}) - \mathcal{F}(\langle \mu^*, X_{t,a_t} \rangle, s_{a_t})$$

We can adapt Proposition 7.1.4 to this setting as follows.

**Proposition 7.7.10** (GMF policy satisfies *history-agnostic demographic parity*). *Let $\{\langle \mu^*, X_a \rangle\}_{a=1}^K$ conditioned to $\{s_a\}_{a=1}^K$ be independent and absolutely continuous and for every $a \in [K], t \in \mathbb{N}$, let $(X_{t,a}, s_{t,a})$ be an i.i.d. copy of $(X_a, s_a)$. Then for every $t \in \mathbb{N}$, $\{\mathcal{F}(\langle \mu^*, X_{t,a} \rangle, s_{t,a})\}_{a=1}^K$ conditioned to $\{s_{t,a}\}_{a=1}^K$ are i.i.d. uniform on $[0, 1]$ and*

$$\mathbb{P}(a_t^* = a \,|\, s_{t,a}, \mathcal{H}_{t-1}^-) = 1/K \qquad \forall a \in [K],$$

*for any GMF policy $\{a_t^*\}_{t=1}^\infty$.*

*Proof.* Let $\psi_a := \mathcal{F}(\langle \mu^*, X_{t,a} \rangle, s_{t,a})$. From the assumptions $\{\psi_a\}_{a=1}^K$ conditioned to $\{s_{t,a}\}_{a=1}^K$ are i.i.d random variables, independent from $\mathcal{H}_{t-1}^-$, with uniform distribution on $[0, 1]$ (see (Casella and Berger, 2021, Theorem 2.1.10)). Let $\tilde{\mathbb{P}} = \mathbb{P}(\cdot \,|\, \{s_{t,a}\}_{a=1}^K, \mathcal{H}_{t-1}^-)$, we have that $\forall a_1, a_2 \in [K]$: $\tilde{\mathbb{P}}(\psi_{a_1} = \psi_{a_2}) = 0$, $\tilde{\mathbb{P}}(a_t^* = a \,|\, \mathcal{H}_{t-1}^-) = \tilde{\mathbb{P}}(a_t^* = a)$ and

$$\tilde{\mathbb{P}}(a_t^* = a_1) = \tilde{\mathbb{P}}(\psi_{a_1} > \psi_{a'}, \forall a' \neq a_1) = \tilde{\mathbb{P}}(\psi_{a_2} > \psi_{a'}, \forall a' \neq a_2) = \tilde{\mathbb{P}}(a_t^* = a_2) = 1/K \ .$$

Let $S_{t,\not{a}} = \{s_{t,a} : a \in [K]/a\}$, then the statement follows from

$$\mathbb{P}(a_t^* = a \,|\, s_{t,a}, \mathcal{H}_{t-1}^-) = \mathbb{E}_{S_{t,\not{a}}}[\tilde{\mathbb{P}}(a_t^* = a)] \ .$$

□

Proposition 7.7.10 states that the probability of selecting an arm does not change based on group membership. Fair-Greedy V2 in Algorithm 8 is the extension of the fair-greedy policy to this new setting.

**Algorithm 8** Fair-Greedy V2

```
1: **Requires** regularization parameter $\lambda > 0$, and noise magnitude $\rho \in (0,1]$
2: **for** $t = 1 \dots T$ **do**
3:     Receive $\{(X_{t,a}, s_{t,a})\}_{a=1}^{K}$
4:     Set $\tilde{t} = \lfloor (t-1)/2 \rfloor$, $X_{1:\tilde{t}} = (X_{1,a_1}, \dots, X_{\tilde{t},a_{\tilde{t}}})^\top$, $r_{1:\tilde{t}} = (r_{1,a_1}, \dots, r_{\tilde{t},a_{\tilde{t}}})$.
5:     **If** $\tilde{t} = 0$ set $\mu_{\tilde{t}} = 0$, **else** let $V_{\tilde{t}} := X_{1:\tilde{t}}^\top X_{1:\tilde{t}} + \lambda \mathbb{I}_d$, generate $\gamma_{\tilde{t}} \sim \mathcal{N}(0, \mathbb{I}_d)$ and compute
           $\mu_{\tilde{t}} := V_{\tilde{t}}^{-1} X_{1:\tilde{t}}^\top r_{1:\tilde{t}} + \frac{\rho}{d\sqrt{\tilde{t}}} \cdot \gamma_{\tilde{t}} \ .$
6:     For each $a \in [K]$, let $i := s_{t,a}$ and $N_{t,i} = \sum_{j=\tilde{t}+1}^{t-1} \sum_{a'=1}^{K} \mathbb{1}\left\{s_{j,a'} = i\right\}$, compute
           $\hat{\mathcal{F}}_t(\langle \mu_{\tilde{t}}, X_{t,a} \rangle, i) := N_{t,i}^{-1} \sum_{j=\tilde{t}+1}^{t-1} \sum_{a'=1}^{K} \mathbb{1}\left\{\langle \mu_{\tilde{t}}, X_{j,a'} \rangle \leq \langle \mu_{\tilde{t}}, X_{t,a} \rangle\right\} \mathbb{1}\left\{s_{j,a'} = i\right\} \ .$
7:     Sample action
           $a_t \sim \mathcal{U}\big[\operatorname*{arg\,max}_{a \in [K]} \hat{\mathcal{F}}_t(\langle \mu_{\tilde{t}}, X_{t,a} \rangle, s_{t,a})\big] \ .$
8:     Observe noisy reward $r_{t,a_t} = \langle \mu, X_{t,a_t} \rangle + \eta_t$.
9: **end for**
```

Notice that the number of contexts used for the CDF approximation for group $i \in [G]$ is now the random variable $N_{t,i}$. Furthermore, we are now using contexts from all the arms to estimate the CDFs, which as we will see it can improve the dependency on $K$ in the fair pseudo-regret bound. We observe that the information averaged demographic parity property of Lemma 7.2.1 does not transfer directly to Fair-Greedy V2, because at each round, there can be a different number of candidates for each group. However, as we will see, the regret is still similar to the original case.

The following Lemma establishes an high probability lower bound on $N_{t,i}$.

**Lemma 7.7.11.** *Let* $q_K := \min_{i \in [G]} \sum_{a=1}^{K} \mathbb{P}(s_a = i)$ *and let*

$$\mathcal{R} = \mathbb{1}\left\{\exists a \in [K] \text{ such that } \forall i \in [G]\, \mathbb{P}(s_a = i) < 1\right\} \ .$$

$\mathcal{R} = 1$ *means that the sensitive attribute is random for at least one arm, while is deterministic if* $\mathcal{R} = 0$. *Then, let* $\alpha = \mathcal{R}b + (1 - \mathcal{R})$ *with* $b \in (0,1)$ *and* $t_N := 3 + \mathcal{R}\lceil \frac{2}{(1-\alpha)^2 q_K} \log(GT/\delta) \rceil$, *with* $N_{t,i}$ *defined at Line 6 of Algorithm 8 and,without loss of generality,* $q_K > 0$. *For simiplicity we let* $\mathcal{R}x = 0$ *when* $\mathcal{R} = 0, x = \infty$. *We have that with probablity at least* $1 - \mathcal{R}\delta$, *for every* $t \in \{t_N, \dots, T\}$

$$\min_{i \in [G]} N_{t,i} \geq (t - 1 - \tilde{t})\alpha q_K$$

*Proof.* If $\mathcal{R} = 0$, then $\mathbb{P}(s_a = i) = \mathbb{1}\left\{s_a = i\right\}$ and the result follows.

If $\mathcal{R} = 1$ instead, note that for every $i \in [G]$ we have that

$$\mathbb{E}[N_{t,i}] = \sum_{j=\tilde{t}-1}^{t-1} \sum_{a=1}^{K} \mathbb{P}(s_a = i) \geq (t - 1 - \tilde{t}) q_K \ .$$

Applying the Chernoff bound we have that with probability at least $1 - \delta$, for all $t > t_N$

$$N_{t,i} \geq \alpha \mathbb{E}[N_{t,i}] \geq (t - 1 - \tilde{t}) \alpha q_K \ ,$$

and the statement follows □

Let $S_T(t_N, \alpha) := \{\{\{s_{t,a}\}_{a=1}^{K}\}_{t=1}^{T} \ : \ \min_i N_{t,i} \geq (t - \tilde{t} - 1)\alpha q_K$ for all $t_N \leq t \leq T\}$ be the event when Lemma 7.7.11 is satisfied. We can then proceed the analysis assuming that $S_T(t_N, \alpha)$ holds. Noticing that the maximum number of approximate CDFs to be computed at each round is $G$ we can adapt Lemma 7.7.4 as follows.

**Lemma 7.7.12.** *Let Assumption 7.7.8(ii) hold and $\hat{\mathcal{F}}_t(r, i)$, and $\mu_{\tilde{t}}$ to be generated by Algorithm 8. Let $Z_i := \langle \mu_{\tilde{t}}, \hat{X}_i \rangle$ and denote with $\mathcal{F}_{Z_i}(\cdot)$ its CDF, conditioned on $\mu_{\tilde{t}}$. Then, if the event $S_T(t_N, \alpha)$ is satisfied, with probability at least $1 - \delta$ we have that for all $t_N \leq t \leq T$*

$$\sup_{i \in [G], r \in \mathbb{R}} |\hat{\mathcal{F}}_t(r, i) - \mathcal{F}_{Z_i}(r)| \leq \sqrt{\frac{\log(2GT/\delta)}{(t-1)\alpha q_K}} \ .$$

Then, following the steps in Lemma 7.3.2, we obtain the following bound on the instantaneous regret.

**Lemma 7.7.13** (Instant regret bound)**.** *Let Assumption 7.7.8(ii)(iv) hold and $a_t$ to be generated by Algorithm 8. Then, if the event $S_T(t_N, \alpha)$ is satisfied, with probability at least $1 - \delta/4$, for all $t$ such that $t_N \leq t \leq T$ we have*

$$\mathcal{F}(\langle \mu^*, X_{t,a_t^*} \rangle, s_{a_t^*}) - \mathcal{F}(\langle \mu^*, X_{t,a_t} \rangle, s_{a_t}) \leq 6M \left\| \mu^* - \mu_{\tilde{t}} \right\|_{V_{\tilde{t}}} \left\| x_{\max} \right\|_{V_{\tilde{t}}^{-1}} + 2\sqrt{\frac{\log(16GT/\delta)}{(t-1)\alpha q_K}} \ ,$$

*where* $\|x_{\max}\|_{V_{\tilde{t}}^{-1}} := \sup_{x \in \cup_{i=1}^{G} \mathrm{Supp}(\hat{X}_i)} \|x\|_{V_{\tilde{t}}^{-1}}$.

*Proof sketch.* Uses the decomposition in the proof of Lemma 7.3.2, then Lemma 7.7.12 and a version of Lemma 7.3.1 adapted to this more general setting. □

We can bound $\|\mu^* - \mu_{\tilde{t}}\|_{V_{\tilde{t}}}$ using the confidence bounds in OFUL (Abbasi-Yadkori et al., 2011). To bound $\|x_{\max}\|_{V_{\tilde{t}}^{-1}}$ instead, we first provide an adaptation of Proposition 7.3.3, which guarantees sufficient exploration of all arms. The proof is very similar to that of Proposition 7.3.3 and we report it here for completeness.

**Proposition 7.7.14.** *Let Assumption 7.1.6 hold, $a_t$ be generated by Algorithm 7 and $c \in [0, 1)$. Then, if $S_T(t_N, \alpha)$ is satisfied, with probability at least $1 - \delta/4$, for all $a \in [K]$ and all $t \geq$*

$\max\left(t_N, 3 + 8\log^{3/2}\left(5G\,\mathrm{e}/\delta\right)\left(1 - \sqrt[K]{c}\right)^{-3}(q_K\alpha)^{-3/2}\right)$ *we have*

$$\mathbb{P}(a_t = a \,|\, s_{t,a}, \mathcal{H}_{t-1}^-) \geq \frac{c}{K} \quad ,$$

*where we recall that* $\mathcal{H}_t^- = \cup_{i=1}^t \left\{\{(X_{i,a}, s_{i,a})\}_{a=1}^K, r_{i,a_i}, a_i\right\}$.

*Proof.* Recall the definition of Algorithm 8. For any $a \in [K]$ let $\hat{r}_{t,a} = \mu_{\tilde{t}}^\top X_{t,a}$, which is the estimated reward for arm $a$, at round $t$. Note that $\mu_{\tilde{t}}$ and $X_{t,a}$ are independent random variables. Furthermore, denote with $\mathcal{F}_{\hat{r}_{t,a}}(\cdot, s_{t,a})$ the CDF of $\hat{r}_{t,a}$ conditioned on $\mu_{\tilde{t}}$ and $s_{t,a}$, and let

$$\phi_{t,a} := \mathcal{F}_{\hat{r}_{t,a}}(\hat{r}_{t,a}, s_{t,a}) \quad , \qquad \text{and} \qquad \hat{\phi}_{t,a} := \hat{\mathcal{F}}_t(\hat{r}_{t,a}, s_{t,a}) \quad .$$

Let $C_t := \arg\max_{a\in[K]} \hat{\phi}_{t,a}$. Now, by the definition of the algorithm, we have

$$\mathbb{P}(a_t = a \,|\, \{s_{t,a}\}_{a=1}^K, \mathcal{H}_{t-1}^-) = \sum_{m=1}^K \frac{1}{m} \mathbb{P}(a \in C_t, |C_t| = m \,|\, \mathcal{H}_{t-1}^-) \quad ,$$

Let $\epsilon_t > 0$ and $\tilde{\mathbb{P}}(\cdot) = \mathbb{P}(\cdot \,|\, \{s_{t,a}\}_{a=1}^K, \mathcal{H}_{t-1}^-, \sup_{a\in[K]} |\phi_{t,a} - \hat{\phi}_{t,a}| \leq \epsilon_t)$. Then, we can write

$$\tilde{\mathbb{P}}(a_t = a) \geq \tilde{\mathbb{P}}(\hat{\phi}_{t,a} > \hat{\phi}_{t,a'}\,, \forall a' \neq a) \geq \tilde{\mathbb{P}}(\phi_{t,a} > \phi_{t,a'} + 2\epsilon_t\,, \forall\, a' \neq a) \quad ,$$

where in the first inequality we considered the case when $a \in C_t$ and $|C_t| = 1$. In the second inequality we considered the worst case scenario where $\hat{\phi}_{t,a} = \phi_{t,a} - \epsilon_t$ and $\hat{\phi}_{t,a'} = \phi_{t,a'} + \epsilon_t$. Recall that by the construction of the algorithm $\mu_{\tilde{t}} = V_{\tilde{t}}^{-1} X_{1:\tilde{t}}^\top r_{1:\tilde{t}} + (1/\sqrt{d\tilde{t}}) \cdot \gamma_{\tilde{t}}$. for all $i \in [G]$, the additive noise $(1/\sqrt{d\tilde{t}})\gamma_{\tilde{t}}$ assures that $\mu_{\tilde{t}}^\top B_i \neq 0$, almost surely. Therefore, by Lemma 7.7.1 $\hat{r}_{t,a} = \langle \mu_{\tilde{t}}, X_{t,a}\rangle$ conditioned on $\mu_{\tilde{t}}$ is absolutely continuous.

assumption 7.1.6(iii) and (Casella and Berger, 2021, Theorem 2.1.10) yield that $\{\phi_{t,a}\}_{a\in[K]}$ conditioned to $\{s_{t,a}\}_{a=1}^K$ are independent and uniformly distributed on $[0,1]$ and in turn that

$$\tilde{\mathbb{P}}(a_t = a) \geq \int_0^1 \left(\mathbb{P}(\phi_{t,a'} < \mu - 2\epsilon_t)\right)^{K-1} \mathrm{d}\mu = \int_{2\epsilon_t}^1 (\mu - 2\epsilon_t)^{K-1} \mathrm{d}\mu = \frac{(1-2\epsilon_t)^K}{K}. \tag{7.12}$$

We continue by computing an $\epsilon_t$ for which $\sup_{a\in[K]} |\phi_{t,a} - \hat{\phi}_{t,a}| \leq \epsilon_t$ holds with high probability. Observing that, conditioned on $\mu_{\tilde{t}}$ and $\{s_{t,a}\}_{a=1}^K$, $\hat{\mathcal{F}}_{t,a}(\cdot, s_{t,a})$ is the empirical CDF of $\mathcal{F}_{\hat{r}_{t,a}}(, s_{t,a})$, we can use Lemma 7.7.11 and the Dvoretzky–Kiefer–Wolfowitz-Massart inequality to obtain, for any $a \in [K]$, $t \geq t_N$, and $s \geq 0$

$$\mathbb{P}\left(|\phi_{t,a} - \hat{\phi}_{t,a}| \geq s\right) \leq 2\exp\left(-2s^2(t - \tilde{t} - 1)(\alpha q_K)\right) \quad .$$

Now, let $\tau_0 := \max\left(t_N, 3 + 8\log^{3/2}\left(5G\,\mathrm{e}/\delta\right)\left(1 - \sqrt[K]{c}\right)^{-3}(\alpha q_K)^{-3/2}\right)$. By applying the union bound and noticing that we have max of $G$ CDFs and approximate CDFs, we can write

$$\begin{aligned}\mathbb{P}\left(\sup_{t\geq\tau_0, a\in[K]} |\phi_{t,a}-\hat{\phi}_{t,a}|\geq s\right) &\leq G\sum_{t=\tau_0}^{\infty}\mathbb{P}\left(|\phi_{t,a}-\hat{\phi}_{t,a}|\geq s\right)\\ &\leq 2G\sum_{t=\tau_0}^{\infty}\exp\left(-2s^2(t-\tilde{t}-1)(\alpha q_K)\right).\end{aligned}$$

Since $\tilde{t}=\lfloor\frac{t-1}{2}\rfloor$, it is straightforward to check that

$$\begin{aligned}\mathbb{P}\left(\sup_{t\geq\tau_0, a\in[K]} |\phi_{t,a}-\hat{\phi}_{t,a}|\geq s\right) &\leq 2G\int_{t=\tau_0-1}^{\infty}\exp\left(-s^2\alpha q_K t\right)\,\mathrm{d}t\\ &\leq \frac{2G}{\alpha q_K s^2}\exp\left(-s^2\alpha q_K(\tau_0-1)\right)\quad.\end{aligned}$$

Now, for any $\delta\in(0,1)$, by assigning $s=\sqrt{\frac{\log(4G(\tau_0-1)/\delta)}{(\tau_0-1)\alpha q_K}}$, we get

$$\mathbb{P}\left(\sup_{t\geq\tau_0, a\in[K]} |\phi_{t,a}-\hat{\phi}_{t,a}|\geq \sqrt{\frac{\log(4G(\tau_0-1)/\delta)}{(\tau_0-1)\alpha q_K}}\right)\leq\frac{\delta}{2\log\left(4G(\tau_0-1)/\delta\right)}\leq\frac{\delta}{4}\ , \tag{7.13}$$

where from $\tau_0\geq 3, \delta<1 \implies 4G(\tau_0-1)/\delta\geq 8\geq e^2 \implies \log\left(4G(\tau_0-1)/\delta\right)\geq 2$ we obtain the last inequality. From (7.12), it follows that

$$\inf_{t\geq\tau_0, a\in[K]}\mathbb{P}\left(a_t=a|\{s_{t,a}\}_{a=1}^K,\mathcal{H}_{t-1}^-\right)\geq\frac{(1-2\sup_{t\geq\tau}\epsilon_t)^K}{K}\quad.$$

Moreover, form (7.13), by letting $\epsilon_t=\sqrt{\frac{\log(4G(\tau_0-1)/\delta)}{(\tau_0-1)\alpha q_K}}$, with probability at least $1-\frac{\delta}{4}$, we have

$$\inf_{t\geq\tau, a\in[K]}\mathbb{P}\left(a_t=a|\{s_{t,a}\}_{a=1}^K,\mathcal{H}_{t-1}^-\right)\geq\frac{1}{K}\left(1-2\sqrt{\underbrace{\frac{\log(4G(\tau_0-1)/\delta)}{(\tau_0-1)\alpha q_K}}_{\text{(I)}}}\right)^K\quad. \tag{7.14}$$

For the term (I) in the above, using $\log(x)\leq\log(5\,\mathrm{e}/4)x^{1/3}$ and $x\geq x^{2/3}$ for any $x\geq 1$ we deduce that

$$\text{(I)}=\frac{\log(4G/\delta)+\log(\tau_0-1)}{(\tau_0-1)\alpha q_K}\leq\frac{\log(4G/\delta)+\log(5\,\mathrm{e}/4)}{(\tau_0-1)^{2/3}\alpha q_K}=\frac{\log(5G\,\mathrm{e}/\delta)}{(\tau_0-1)^{2/3}\alpha q_K}\quad.$$

Now, since $\tau_0\geq 3+8\log^{3/2}\left(5G\,\mathrm{e}/\delta\right)\left(1-\sqrt[K]{c}\right)^{-3}(\alpha q_K)^{-3/2}$, we get that (I) $\leq\frac{1}{4}\left(1-\sqrt[K]{c}\right)^2$ and conclude the proof by plugging this inequality in (7.14). □

Furthermore, for fixed $t$, let $\tilde{\mathbb{E}}=\mathbb{E}[\cdot\,|\,\mathcal{H}_{t-1}^-]$ and $\tilde{\mathbb{P}}=\mathbb{P}[\cdot\,|\,\mathcal{H}_{t-1}^-]$. Note that if the assumptions of Proposition 7.7.14 are satisfied, then

$$\begin{aligned}\tilde{\mathbb{E}}[X_{t,a_t}X_{t,a_t}^\top] &= \sum_{i=1}^{G} \mathbb{E}[\hat{X}_i\hat{X}_i^\top]\tilde{\mathbb{P}}(s_{t,a_t} = i) \\ &= \sum_{i=1}^{G} \mathbb{E}[\hat{X}_i\hat{X}_i^\top]\sum_{a=1}^{K} \tilde{\mathbb{P}}(a_t = a \,|\, s_{t,a} = i)\tilde{\mathbb{P}}(s_{t,a} = i) \\ &\geq cK^{-1}\sum_{i=1}^{G} \mathbb{E}[\hat{X}_i\hat{X}_i^\top]q_K = c\frac{q_K G}{K}\frac{1}{G}\sum_{i=1}^{G}\mathbb{E}[\hat{X}_i\hat{X}_i^\top] \qquad (7.15)\end{aligned}$$

where we applied Proposition 7.7.14 in the last line. We can bound $\|x_{\max}\|_{V_{\tilde{t}}^{-1}}$ in Lemma 7.7.13 in the same way as in Lemma 7.7.6 using (7.15) with $c = 1/2$ when needed in the proof of Lemma 7.7.5. Combining the previous results we obtain the following regret bound.

**Theorem 7.7.15.** *Let Assumption 7.7.8 hold, $a_t$ be generated by Algorithm 8 and $\Sigma := G^{-1}\sum_{i=1}^{G}\mathbb{E}[\hat{X}_i\hat{X}_i^\top]$ Then, with probability at least $1-\delta$, for any $T \geq 1$ we have*

$$\begin{aligned}R_F(T) \leq & \frac{96ML\sqrt{K}}{\sqrt{\lambda_{\min}^{+}(\Sigma)q_K G}}\left[(\lambda^{\frac{1}{2}} + R + L)\sqrt{dT\log((8+4T\max(L^2/\lambda, 1))/\delta_1)} + \sqrt{\lambda T}\left\|\mu^*\right\|_2\right] \\ & + 8\sqrt{\frac{T\log(8GT/\delta_1)}{3\alpha q_K}} + \tau \ ,\end{aligned}$$

*where $\delta = \delta_1 + \mathcal{R}\delta_2$, $\tau = 4\max(\tau_1, \tau_2, \mathcal{R}\tau_3) + 3$ and*

$$\tau_1 = \frac{32K^3}{(\alpha q_K)^{3/2}}\log^{3/2}\left(5G\,\mathrm{e}/\delta_1\right), \quad \tau_2 = \frac{54L^2}{\lambda_{\min}^{+}(\Sigma)}\log(4d/\delta_1), \quad \tau_3 = \frac{2}{(1-\alpha)^2 q_K}\log(GT/\delta_2),$$

*where $q_K$, $\mathcal{R}$ and $\alpha$ are defined in Lemma 7.7.11 and we use the convention $\mathcal{R}\tau_3 = 0$ if $\mathcal{R} = 0, \tau_3 = \infty$. Hence*

$$R_F(T) = O\left(\mathcal{R}\frac{\log(GT/\delta_2)}{q_K} + \frac{K^3\log^{3/2}(G/\delta_1)}{q_K^{3/2}} + \sqrt{\frac{dT\log\left(GT/\delta_1\right)}{(1+G/K)q_K)}}\right) .$$

*Proof Sketch.* First, assume $S_T(t_N, \alpha)$ holds and use a similar strategy of Theorem 7.3.5 to get a bound w.p. at least $1-\delta_1$. Then combine this result with Lemma 7.7.11. □

Notice that in the case where each arm corresponds to a different sensitive group, i.e. when $G = K$, $s_a = a$ and therefore $q_K = 1$, $\mathcal{R} = 0$ and $\alpha = 1$, we recover Theorem 7.3.5. Moreover, we have the following corollary which shows an advantage for higher number of arms compared to the bound in Theorem 7.3.5 when $\{(X_a, s_a)\}_{a=1}^{K}$ are i.i.d..

**Corollary 7.7.16.** *Let $\{(X_a, s_a)\}_{a=1}^{K}$ be i.i.d. and $q_{\min} := \min_{i\in[G]}\mathbb{P}(s_a = i)G$. If Assumption 7.7.8 holds and $a_t$ is generated via Algorithm 8 we have that with probability at least $1-\delta$,* *for any $T \geq 1$ and $\alpha \in (0,1)$ we have that*

$$R_F(T) \leq \frac{96ML}{\sqrt{\lambda_{\min}^{+}(\Sigma) q_{\min}}} \left[ (\lambda^{\frac{1}{2}} + R + L) \sqrt{dT \log((8 + 4T \max(L^2/\lambda, 1))/\delta_1)} + \sqrt{\lambda T} \left\| \mu^* \right\|_2 \right]$$
$$+ 8 \sqrt{\frac{TG \log(8GT/\delta_1)}{3K\alpha q_{\min}}} + \tau \ ,$$

*where $\delta = \delta_1 + \delta_2$, $\tau = 4 \max(\tau_1, \tau_2, \tau_3) + 3$ and*

$$\tau_1 = \frac{32(KG)^{3/2}}{(\alpha q_{\min})^{3/2}} \log^{3/2} \left( \frac{5G\,\mathrm{e}}{\delta_1} \right), \ \tau_2 = \frac{54L^2}{\lambda_{\min}^{+}(\Sigma)} \log \left( \frac{4d}{\delta_1} \right), \ \tau_3 = \frac{2G}{(1-\alpha)^2 K q_{\min}} \log \left( \frac{GT}{\delta_2} \right).$$

*Hence*

$$R_F(T) = O \left( \frac{G \log(GT/\delta)}{K q_{\min}} + \frac{(KG)^{3/2} \log^{3/2}(G/\delta)}{q_{\min}^{3/2}} + \sqrt{\frac{dT \log(GT/\delta)}{(1 + K/G) q_{\min}}} \right)$$

Note that in Corollary 7.7.16, $q_{\min} > 0$ without loss of generality and $q_{\min} = 1$ if and only if each group has the same probability of being sampled. Furthermore $q_{\min}/G$ is the probability that a context belongs to the less common group, which depends on the problem at hand. Note that there is an advantage compared to Theorem 7.3.5 in terms of number of arms when $K > G$. This is because context coming from all arms can be use to estimate the CDF of a given group.

## Additional details on the US census experiments

This experiment is introduced in Section 7.5 and similarly to that of Section 7.7, is performed using the US Census data. However, candidates are sampled from the original dataset at random together with their sensitive group (their ethnicity). Hence, we use Fair-greedy V2 (Algorithm 8). Differently from Section 7.7 where we use the target income as noisy reward, here we add artificial noise with standard deviation $0.2$ directly to the true reward.

**Setup and Preprocessing.** To setup the bandit problem, we construct two datasets: $D1$ and $D2$. We first load all the data from the $2017$ US Census Survey to assemble $D1$, and then from the $2018$ survey to assemble $D2$. Then we retain only candidates from $6$ ethnic groups containing at least $5 \times 10^3$ candidates, in order to accurately compute the true CDF for each group. We use $D1$ to find mean and standard deviation for each feature and also for the target. After that we normalize features and target of $D2$ by subtracting the mean and dividing by the standard deviation previously computed on $D1$. We then construct $\mu^*$ as a ridge regression estimate on the samples from $D2$ with the regularization parameter equal to $10^{-8}$. The regression vector $\mu^*$ will be used to compute the (true) rewards for the samples. We construct the bandit problem as follows. At each round, the context vectors of $K = 10$ individual are sampled from $D2$ and after one is selected by the policy, its corresponding noisy

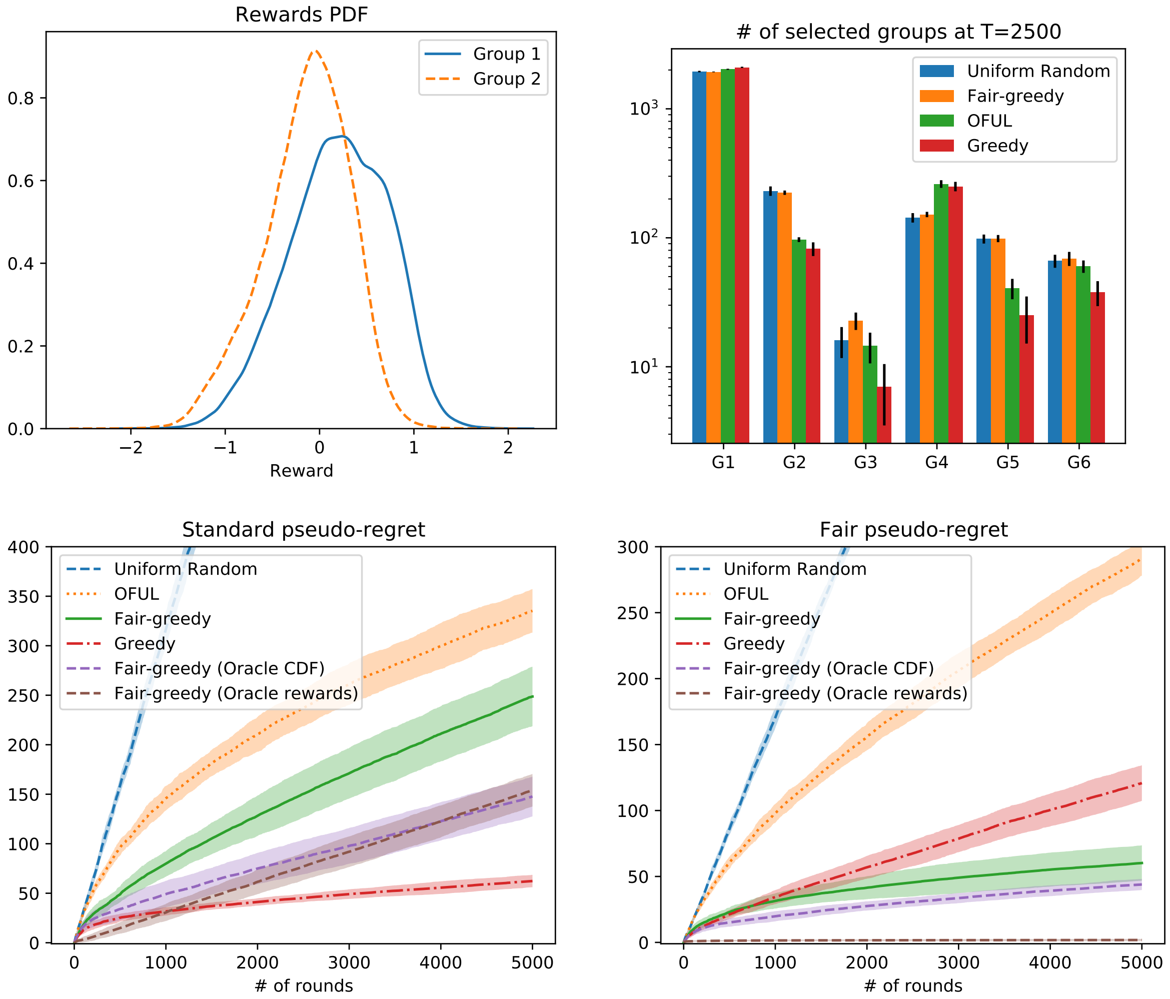

Figure 7.4: **US Census Results. Group $=$ Ethnicity**. First image is the density plots of the reward distributions, the second image is the number candidates (in log scale) from each group which are selected by each policy (mean and std over 10 runs), while the bottom two plots are the standard and fair pseudo-regrets, with mean (solid lines) $\pm$ standard deviation (shaded region) over 10 runs. To compute the reward CDF for each group we use the empirical CDF on $5K$ samples from $D2$.

reward, obtained by adding gaussian noise with standard deviation $0.2$ to the true reward, is received by the agent.

**Baselines.** We compare our method with the same baselines as in Section 7.7, where the two oracle policies are now variants of Fair-Greedy V2. Moreover, we set the regularization parameter for all policies using a ridge estimate to 0.1 and the exploration parameter of OFUL to 0.01.

**Results (Figure 7.4).** We draw similar conclusions as in Section 7.7. In particular, Greedy performing better than OFUL and the Fair-Greedy policy achieving sublinear fair pseudo-regret, but worse than Oracle methods. Additionaly we can see that knowing $\mu^*$ plays a more important role than knowing the true reward CDFs. In this case, the gap between the Uniform random policy and the others is even larger since $K = 10$. Moreover, as expected, Fair-greedy selects more candidates from underperforming (in terms of reward) minority groups, when compared with OFUL and Greedy.

### Trade off between fairness and reward maximization

In this section, we show for which problems the GMF policy and the optimal policy have competing goals. in particular, for the case of $K = 2$, when the rewards are absolutely continuous and independent across arms, whenever they are not identically distributed, the GMF policy achieves linear standard pseudo-regret with nonzero probability. The following theorem proves this result.

**Theorem 7.7.17.** *Let Assumption 7.1.6 hold with $K = 2$, and assume that $\mathcal{F}_1 \neq \mathcal{F}_2$. Let $\bar{r}_{t,a} := \langle \mu^*, X_{t,a} \rangle$, $\{a_t^*\}_{t=1}^T$ be the GMF policy (see Definition 7.1.1) and $\{a_t^{\textit{opt}}\}_{t=1}^T$ be the optimal policy (see Remark 7.1.3). Then, there exists $\epsilon > 0$, such that*

$$p = \int_0^1 \left[\max(\mathcal{F}_1(\mathcal{F}_2^{-1}(y) - \epsilon) - y, 0) + \max(y - \mathcal{F}_1(\mathcal{F}_2^{-1}(y) + \epsilon), 0)\right] \mathrm{d}y > 0 \ .$$

*Furthermore with probability at least $\frac{\epsilon p}{4L\|\mu^*\|}$, for any $T > 0$, we have*

$$T \cdot \frac{\epsilon p}{2} \leq \sum_{t=1}^{T} \left[\bar{r}_{t,a_t^{\textit{opt}}} - \bar{r}_{t,a_t^*}\right] \ .$$

*Proof.* Let $\bar{r}_a := \langle \mu^*, X_a \rangle$, $q_a = \mathcal{F}_a(r_a)$ be the CDF value of $r_a$ and $\mathcal{F}_a^{-1}$ be the quantile function, i.e. such that $\mathcal{F}_a^{-1}(x) = \inf\{y \in \mathbb{R} : x \leq \mathcal{F}_a(y)\}$. For $\epsilon > 0$ consider the set $E^\epsilon := E_1^\epsilon \cup E_2^\epsilon$ where

$$\begin{aligned} E_1^\epsilon &:= \{(x, y) \in [0,1]^2 : x > y, \mathcal{F}_1^{-1}(x) < \mathcal{F}_2^{-1}(y) - \epsilon\} \ , \\ E_2^\epsilon &:= \{(x, y) \in [0,1]^2 : x < y, \mathcal{F}_1^{-1}(x) > \mathcal{F}_2^{-1}(y) + \epsilon\} \ . \end{aligned}$$

Note that we can write

$$\begin{aligned} E_1^\epsilon &= \{(x, y) \in [0,1]^2 : y < x < \mathcal{F}_1(\mathcal{F}_2^{-1}(y) - \epsilon)\} \ , \\ E_2^\epsilon &= \{(x, y) \in [0,1]^2 : \mathcal{F}_1(\mathcal{F}_2^{-1}(y) + \epsilon) < x < y\} \ . \end{aligned}$$

Now, let $g_{1,2}(y, \epsilon) = \mathcal{F}_1(\mathcal{F}_2^{-1}(y) + \epsilon)$. Since from Assumption 7.1.6(ii)(iv), $q_1$ an $q_2$ are $i.i.d$

uniform on $[0,1]$ we have that

$$\begin{aligned}\mathbb{P}((q_1,q_2)\in E^{\epsilon}) &= \mathbb{P}((q_1,q_2)\in E_1^{\epsilon}) + \mathbb{P}((q_1,q_2)\in E_2^{\epsilon})\\ &= \int_0^1\int_y^{g_{1,2}(y,-\epsilon)} \text{d}x\text{d}y + \int_0^1\int_{g_{1,2}(y,\epsilon)}^{y} \text{d}x\text{d}y\\ &= \int_0^1 \max(g_{1,2}(y,-\epsilon)-y,0)\text{d}y + \int_0^1 \max(y-g_{1,2}^{\epsilon}(y,\epsilon),0)\text{d}y\\ &= \int_0^1 \left[\max(\mathcal{F}_1(\mathcal{F}_2^{-1}(y)-\epsilon)-y,0)+\max(y-\mathcal{F}_1(\mathcal{F}_2^{-1}(y)+\epsilon),0)\right]\text{d}y\ .\end{aligned}$$

Since $\mathcal{F}_1\neq\mathcal{F}_2$, and $\mathcal{F}_1,\mathcal{F}_2$ are absolutely continuous, there exists $\epsilon'>0$, such that $\mathcal{F}_2^{-1}(y)-\mathcal{F}_1^{-1}(y)>\epsilon'$, or $\mathcal{F}_2^{-1}(y)-\mathcal{F}_1^{-1}(y)<\epsilon'$ for y inside a closed interval, and hence $\mathbb{P}((q_1,q_2)\in E^{\epsilon'})>0$. This concludes the proof by letting $\epsilon=\epsilon'$, and $p=\mathbb{P}((q_1,q_2)\in E^{\epsilon})$.

Now, let $q_{t,a}=\mathcal{F}_a(\bar{r}_{t,a})$, then for the expected value of the instantaneous standard regret, at round $t$, we can write

$$\mathbb{E}\big[\bar{r}_{t,a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}\big]\geq\int_{(x,y)\in E^{\epsilon}}|\mathcal{F}_2^{-1}(y)-\mathcal{F}_1^{-1}(x)|\,\mathrm{d}x\,\mathrm{d}y\geq\epsilon\mathbb{P}((q_{t,1},q_{t,2})\in E^{\epsilon})=\epsilon p>0\ ,$$

and for the standard regret, we have

$$\sum_{t=1}^{T}\mathbb{E}\big[\bar{r}_{t,a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}\big]\geq T\cdot\epsilon p>0\ .$$

Finally, let $\Omega$ be the event that $\frac{1}{2}\cdot\sum_{t=1}^{T}\mathbb{E}\big[\bar{r}_{t,a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}\big]\leq\sum_{t=1}^{T}[\bar{r}_{t,a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}]$. Considering the fact that $\sum_{t=1}^{T}[\bar{r}_{t,a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}]\leq 2L\left\|\mu^*\right\|T$, we deduce

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[r_{a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}] &= \sum_{t=1}^{T}\Big[\mathbb{E}[\bar{r}_{t,a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}\,|\,\Omega]\mathbb{P}(\Omega)+\mathbb{E}[\bar{r}_{t,a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}\,|\,\Omega^c]\mathbb{P}(\Omega^c)\Big]\\ &\leq 2L\left\|\mu^*\right\|T\mathbb{P}(\Omega)+\sum_{t=1}^{T}\mathbb{E}[\bar{r}_{t,a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}]/2\ ,\end{aligned}$$

and we get $\frac{\epsilon p}{4L\|\mu^*\|}\leq\sum_{t=1}^{T}\mathbb{E}[\bar{r}_{t,a_t^{\text{opt}}}-\bar{r}_{t,a_t^*}]/(4L\left\|\mu^*\right\|T)\leq\mathbb{P}(\Omega)$, which finishes the proof. □

**Remark 7.7.18.** *In Theorem 7.7.17, $\epsilon\leq 2L\left\|\mu^*\right\|$, otherwise $p=0$. On the other hand, by the definition $p\leq 1$, and accordingly $\frac{\epsilon p}{4L\|\mu^*\|}\leq 1/2$.*

**Remark 7.7.19.** *With similar reasoning as in the proof of Theorem 7.7.17, we can show that if $\mathcal{F}_1\neq\mathcal{F}_2$ the optimal policy (see Remark 7.1.3) has linear fair pseudo-regret with positive probability, that is independent of $T$. In particular, there exist $c,c'>0$, such that for any $T>0$, $\mathbb{P}(T\cdot c'\leq\sum_{t=1}^{T}[\mathcal{F}_{a_t^*}(\bar{r}_{t,a_t^*})-\mathcal{F}_{a_t^{\text{opt}}}(\bar{r}_{t,a_t^{\text{opt}}})])>c$.*

**Example 7.7.20** (Disjoint supports)**.** *As an example consider the case when $K=2$ and $\bar{r}_{t,1}-\bar{r}_{t,2}\geq\epsilon>0$, for all $t\geq 1$, almost surely. Then, $\mathcal{F}_1(\mathcal{F}_2^{-1}(y)-\epsilon)=\mathcal{F}_1(\mathcal{F}_2^{-1}(y)+\epsilon)=0$ for* *every* $y \in [0, 1]$. *Hence we have*

$$p = \int_0^1 \left[\max(\mathcal{F}_1(\mathcal{F}_2^{-1}(y) - \epsilon) - y, 0) + \max(y - \mathcal{F}_1(\mathcal{F}_2^{-1}(y) + \epsilon), 0)\right] \, \mathrm{d}y = 1/2 \ \ .$$

*Then by Theorem 7.7.17, with probability at least* $\frac{\epsilon}{8L\|\mu^*\|}$*, for any* $T > 0$*, we have* $\sum_{t=1}^{T}[\bar{r}_{t,a_t^{\mathit{opt}}} - \bar{r}_{t,a_t^*}] \geq \frac{T\epsilon}{4}$.

# Bibliography

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320.

Agarwal, A., Dekel, O., and Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proc. 23rd International Conference on Learning Theory*, pages 28–40.

Agarwal, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Rakhlin, A. (2011). Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 25, pages 1035–1043.

Akhavan, A., Chzhen, E., Pontil, M., and Tsybakov, A. (2022a). A gradient estimator via l1-randomization for online zero-order optimization with two point feedback. *arXiv:2205.13910*.

Akhavan, A., Gogolashvili, D., and Alexandre B., T. (2022b). Estimating the minimizer and the minimum value of a regression function under passive design. *arXiv preprint arXiv:2211.16457*.

Akhavan, A., Pontil, M., and Tsybakov, A. (2020). Exploiting higher order smoothness in derivative-free optimization and continuous bandits. In *Advances in Neural Information Processing Systems 33*.

Akhavan, A., Pontil, M., and Tsybakov, A. B. (2021). Distributed zero-order optimization under adversarial noise. *arXiv preprint arXiv:2102.01121*.

Arias-Castro, E., Qiao, W., and Zheng, L. (2022). Estimation of the global mode of a density: Minimaxity, adaptation, and computational complexity. *Electronic Journal of Statistics*, 16(1):2774–2795.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2022). Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

Bach, F. and Perchet, V. (2016). Highly-smooth zero-th order online optimization. In *Proc. 29th Annual Conference on Learning Theory*, pages 1–27.

Balashov, M., Polyak, B., and Tremba, A. (2020). Gradient projection and conditional gradient methods for constrained nonconvex minimization. *Numerical Functional Analysis and Optimization*, 41(7):822–849.

Balasubramanian, K. and Ghadimi, S. (2021). Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42.

Barocas, S., Hardt, M., and Narayanan, A. (2018). *Fairness and Machine Learning*. fairmlbook.org.

Barthe, F., Guédon, O., Mendelson, S., and Naor, A. (2005). A probabilistic approach to the geometry of the Lpn-ball. *The Annals of Probability*, 33(2):480 – 513.

Barthe, F. and Wolff, P. (2009). Remarks on non-interacting conservative spin systems: the case of gamma distributions. *Stochastic processes and their applications*, 119(8):2711–2723.

Bartlett, P. L., Gabillon, V., and Valko, M. (2019). A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. In *Proc. 30th International Conference on Algorithmic Learning Theory*, pages 184–206.

Bartlett, P. L., Hazan, E., and Rakhlin, A. (2008). Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20*, pages 65–72.

Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, S. Z. (2019). Equal opportunity in online classification with partial feedback. *Advances in Neural Information Processing Systems*, 32.

Beckner, W. (1989). A generalized poincaré inequality for gaussian measures. *Proceedings of the American Mathematical Society*, 105(2):397–400.

Belitser, E., Ghosal, S., and van Zanten, H. (2012). Optimal two-stage procedures for estimating location and size of the maximum of a multivariate regression function. *The Annals of Statistics*, 40(6):2850–2876.

Belitser, E., Ghosal, S., and van Zanten, H. (2021). Correction note: "Optimal two-stage procedures for estimating location and size of the maximum of a multivariate regression function". *Ann. Statist.*, 49(1):612–613.

Belloni, A., Liang, T., Narayanan, H., and Rakhlin, A. (2015). Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Proc. 28th Annual Conference on Learning Theory*, pages 240–265.

Blum, A., Gunasekar, S., Lykouris, T., and Srebro, N. (2018). On preserving non-discrimination when combining expert advice. *Advances in Neural Information Processing Systems*, 31.

Blum, J. R. (1954). Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744.

Bobkov, S. and Ledoux, M. (1997). Poincaré's inequalities and talagrand's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107(3):383–400.

Bobkov, S. G. and Ledoux, M. (2009). Weighted Poincaré-type inequalities for Cauchy and other convex measures. *The Annals of Probability*, 37(2):403 – 427.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. (2006). Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–257.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Bubeck, S., Lee, Y. T., and Eldan, R. (2017). Kernel-based methods for bandit convex optimization. In *Proc. 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Neural Information Processing Systems*.

Carmon, Y., Duchi, J., Hinder, O., and Sidford, A. (2017). Lower bounds for finding stationary points i. *Mathematical Programming*, 184.

Carpentier, A., Collier, O., Comminges, L., Tsybakov, A., and Wang, Y. (2019). Minimax rate of testing in sparse linear regression. *Automation and Remote Control*, 80:1817–1834.

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Cauchy, A. (1847). Methode generale pour la resolution des systemes d'equations simultanees. *C.R. Acad. Sci. Paris*, 25:536–538.

Chen, H. (1988). Lower rate of convergence for locating a maximum of a function. *The Annals of Statistics*, pages 1330–1334.

Chen, Y., Cuellar, A., Luo, H., Modi, J., Nemlekar, H., and Nikolaidis, S. (2020). Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pages 181–190. PMLR.

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16(1):31–41.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. In *Neural Information Processing Systems*.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214.

Chung, K. L. (1954). On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483.

Dalenius, T. (1965). The mode–a neglected statistical parameter. *Journal of the Royal Statistical Society. Series A (General)*, pages 110–117.

Dasgupta, S. and Kpotufe, S. (2014). Optimal rates for k-nn density and mode estimation. *Advances in Neural Information Processing Systems*, 27.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

Dippon, J. (2003a). Accelerated randomized stochastic optimization. *Ann. Statist.*, 31(4):1260–1281.

Dippon, J. (2003b). Accelerated randomized stochastic optimization. *The Annals of Statistics*, 31(4):1260–1281.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems*.

Duchi, J. C., Agarwal, A., and Wainwright, M. J. (2012). Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606.

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

Dupač, V. (1957). O kiefer-wolfowitzově aproximační methodě. *Časopis pro pěstování matematiky*, 82(1):47–75.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669.

Dvurechensky, P., Gasnikov, A., and Gorbunov, E. (2018). An accelerated method for derivative-free smooth stochastic convex optimization. *arXiv preprint arXiv:1802.09022*.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. D. M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*.

Evans, L. C. and Gariepy, R. F. (2018). *Measure theory and fine properties of functions*. Routledge.

Fabian, V. (1967a). Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, 38(1):191–200.

Fabian, V. (1967b). Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, pages 191–200.

Facer, M. R. and Müller, H.-G. (2003). Nonparametric estimation of the location of a maximum in a response surface. *Journal of Multivariate Analysis*, 87(1):191–217.

Feng, Q. (2010). *Bounds for the ratio of two gamma functions*. Journal of Inequalities and Applications.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proc. 16th Annual ACM-SIAM Symposium on Discrete algorithms (SODA)*, pages 385—-394.

Garrigos, G., Rosasco, L., and Villa, S. (2022). Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry. *Mathematical Programming*, pages 1–60.

Gasnikov, A., Krymova, E., Lagunovskaya, A., Usmanova, I., and Fedorenko, F. (2017). Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case. *Automation and Remote Control*, 78(2):224–234.

Gasnikov, A., Lagunovskaya, A., Usmanova, I., and Fedorenko, F. (2016). Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control*, 77(11):2018–2034.

Ghadimi, S. and Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Gillen, S., Jung, C., Kearns, M., and Roth, A. (2018). Online learning with an unknown fairness metric. *Advances in neural information processing systems*, 31.

Grazzi, R., Akhavan, A., Falk, I., Cella, L., and Pontil, M. (2022). Group meritocratic fairness in linear contextual bandits. *arXiv:2206.03150*.

Grenander, U. (1965). Some direct estimates of the mode. *The Annals of Mathematical Statistics*, pages 131–138.

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM review*, 31(2):221–239.

Hajinezhad, D., Hong, M., and Garcia, A. (2019). Zeroth order nonconvex multi-agent optimization over networks. *IEEE Transactions on Automatic Control*, 64(10):3995–4010.

Hall, P. (1927). The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika*, pages 240–245.

Härdle, W. and Nixdorf, R. (1987). Nonparametric sequential estimation of zeros and extrema of regression functions. *IEEE transactions on information theory*, 33(3):367–372.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Neural Information Processing Systems*.

Hu, X., Prashanth, L. A., György, A., and Szepesvári, C. (2016a). (Bandit) convex optimization with biased noisy gradient oracles. In *Proc. 10th International Conference on Artificial Intelligence and Statistics*, pages 819–828.

Hu, X., Prashanth, L. A., György, A., and Szepesvári, C. (2016b). Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Proc. 10th International Conference on Artificial Intelligence and Statistics*, pages 819–828.

Ibragimov, I. A. and Khas'minskii, R. Z. (1981). *Statistical Estimation, Asymptotic Theory*. Springer, New York.

Ibragimov, I. A. and Khas'minskii, R. Z. (1982). Estimation of the maximum value of a signal in gaussian white noise. *Mat. Zametki*, 32(4):746–750.

Jakovetić, D. (2019). A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):31–46.

Jakovetić, D., Xavier, J., and Moura, J. M. F. (2014). Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59(5):1131–1146.

Jamieson, K. G., Nowak, R., and Recht, B. (2012). Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, volume 26, pages 2672–2680.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2018). Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. (2016). Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811.

Kearns, M., Roth, A., and Wu, Z. S. (2017). Meritocratic fairness for cross-population selection. In *International Conference on Machine Learning*, pages 1828–1836. PMLR.

Khas'minskii, R. Z. (1979). Lower bound for the risks of nonparametric estimates of the mode. *Contributions to statistics*, pages 91–97.

Kia, S., Cortés, J., and Martínez, S. (2015). Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. *Autom.*, 55:254–264.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466.

Klemelä, J. (2005). Adaptive estimation of the mode of a multivariate density. *Journal of Nonparametric Statistics*, 17(1):83–105.

Kraska, T., Talwalkar, A., Duchi, J. C., Griffith, R., Franklin, M., and Jordan, M. I. (2013). MLbase: A distributed machine-learning system. In *CIDR*.

Krishnamurthy, V. and Yin, G. (2022). Multikernel passive stochastic gradient algorithms and transfer learning. *IEEE Trans. Automat. Control*, 67:1792–1805.

Lattimore, T. and Gyorgy, A. (2021). Improved regret for zeroth-order stochastic convex bandits. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 2938–2964. PMLR.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lepski, O. V. (1993). Estimation of the maximum of a nonparametric signal up to a constant. *Theory Probab. Appl.*, 38:152–158.

Lepskii, O. (1991). On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466.

Li, F., Liu, J., and Ji, B. (2019). Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813.

Liu, L., Luo, C., and Shen, F. (2017). Multi-agent formation control with target tracking and navigation. In *2017 IEEE International Conference on Information and Automation (ICIA)*, pages 98–103.

Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., and Parkes, D. C. (2017). Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*.

Lobel, I., Ozdaglar, A., and Feijer, D. (2011). Distributed multi-agent optimization with state-dependent communication. *Mathematical Programming*, 129:255—284.

Locatelli, A. and Carpentier, A. (2018). Adaptivity to smoothness in x-armed bandits. In *Proc. 31st Annual Conference on Learning Theory*, pages 1–30.

Malherbe, C. and Vayatis, N. (2017). Global optimization of lipschitz functions. In *Proc. 34th International Conference on Machine Learning*, pages 2314–2323.

Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Mokkadem, A. and Pelletier, M. (2007). A companion for the kiefer–wolfowitz–blum stochastic approximation algorithm. *The Annals of Statistics*, 35(4):1749–1772.

Müller, H.-G. (1985). Kernel estimators of zeros and of location and size of extrema of regression functions. *Scandinavian journal of statistics*, pages 221–232.

Müller, H.-G. (1989). Adaptive nonparametric peak estimation. *The Annals of Statistics*, pages 1053–1069.

Nazin, A., Polyak, B., and Tsybakov, A. (1989). Passive stochastic approximation. *Automat. Remote Control*, 50:1563–1569.

Nazin, A. V., Polyak, B. T., and Tsybakov, A. B. (1992). Optimal and robust algorithms of passive stochastic approximation. *IEEE Transactions on Information Theory*, 38(5):1577–1583.

Nedic, A. and Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61.

Nedic, A., Ozdaglar, A., and Parrilo, P. (2010). Constrained consensus and optimization in multi-agent networks. *Automatic Control, IEEE Transactions on*, 55:922–938.

Nemirovski, A. (2000). Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85.

Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization.* Wiley & Sons.

Nesterov, Y. (2011). Random gradient-free minimization of convex functions. Technical Report 2011001, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain.

Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer Optimization and Its Applications. Springer.

Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17:527—566.

Novitskii, V. and Gasnikov, A. (2021). Improved exploiting higher order smoothness in derivative-free optimization and continuous bandit. *arXiv preprint arXiv:2101.03821*.

Olshevsky, A. (2014). Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control. *arXiv: Optimization and Control*.

Olshevsky, A. and Tsitsiklis, J. (2009). Convergence speed in distributed consensus and control. *SIAM Journal on Control and Optimization*, 48(1):33–55.

Orabona, F. (2019). A modern introduction to online learning. *ArXiv*, abs/1912.13213.

Orabona, F. and Pál, D. (2016). Scale-free online learning. *Theoretical Computer Science*, 716.

Osserman, R. (1978). The isoperimetric inequality. *Bulletin of the American Mathematical Society*, 84(6):1182 – 1238.

Park, J., Samarakoon, S., Elgabli, A., Kim, J., Bennis, M., Kim, S.-L., and Debbah, M. (2021). Communication-efficient and distributed learning over wireless networks: Principles and applications. *Proceedings of the IEEE*, 109(5):796–819.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Patil, V., Ghalme, G., Nair, V., and Narahari, Y. (2020). Achieving fairness in the stochastic multi-armed bandit problem. In *AAAI*, pages 5379–5386.

Polyak, B. (1963). Gradient methods for the minimisation of functionals. *Ussr Computational Mathematics and Mathematical Physics*, 3:864–878.

Polyak, B. and Juditsky, A. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855.

Polyak, B. T. (1987). Introduction to optimization. *Optimization Software, Inc, New York*.

Polyak, T. B. and Tsybakov, A. B. (1990). Optimal order of accuracy of search algorithms in stochastic optimization. *Problems of Information Transmission*, 26(2):45–53.

Pu, S., Shi, W., Xu, J., and Nedić, A. (2021). Push-pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, 66(1):1–16.

Qu, G. and Li, N. (2018). Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network System*, 5(5):1245–1260.

Rachev, S. T. and Ruschendorf, L. (1991). Approximate Independence of Distributions on Spheres and Their Stability Properties. *The Annals of Probability*, 19(3):1311 – 1337.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proc. 29th Int. Conf. on Machine Learning*, pages 1571–1578.

Rando, M., Molinari, C., Villa, S., and Rosasco, L. (2022). Stochastic zeroth order descent with structured directions. *arXiv preprint arXiv:2206.05124*.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407.

Saha, A. and Tewari, A. (2011). Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proc. 14th International Conference on Artificial Intelligence and Statistics*, pages 636–642.

Sahu, A., Jakovetic, D., Bajovic, D., and Kar, S. (2018a). Communication-efficient distributed strongly convex stochastic optimization: Non-asymptotic rates. *arXiv:1809.02920*.

Sahu, A. K., Jakovetic, D., Bajovic, D., and Kar, S. (2018b). Distributed zeroth order optimization over random networks: A kiefer-wolfowitz stochastic approximation approach. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 4951–4958. IEEE.

Sayin, M. O., Vanli, N. D., Kozat, S. S., and Başar, T. (2017). Stochastic subgradient algorithms for strongly convex optimization over distributed networks. *IEEE Transactions on Network Science and Engineering*, 4(4):248–260.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. (2019). Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20:1–31.

Schechtman, G. and Zinn, J. (1990). On the volume of the intersection of two $L_p^n$ balls. *Proc. Amer. Math. Soc.*, 110(1):217–224.

Shalev-Shwartz, S. (2012). *Online learning and online convex optimization*, volume 4. Now Publishers, Inc.

Shamir, O. (2013). On the complexity of bandit and derivative-free stochastic convex optimization. In *Proc. 30th Annual Conference on Learning Theory*, pages 1–22.

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

Shi, W., Wu, G., and Yin, W. (2014). Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, 10:1040–1053.

Tang, Y., Zhang, J., and Li, N. (2019). Distributed zero-order algorithms for nonconvex multi-agent optimization. *arXiv preprint arXiv:1908.11444v3*.

Tropp, J. A. (2011). User-friendly tail bounds for matrix martingales. Technical report, California Institute of Technology.

Tsitsiklis, J., Bertsekas, D., and Athans, M. (1986). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812.

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York.

Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, pages 69–84.

Tsybakov, A. B. (1990a). Locally-polynomial algorithms of passive stochastic approximation. *Problems of Control and Information Theory*, 19:181–195.

Tsybakov, A. B. (1990b). Recursive estimation of the mode of a multivariate distribution. *Problems of Information Transmission*, 26:31–37.

Venter, J. (1967). On estimation of the mode. *The Annals of Mathematical Statistics*, pages 1446–1455.

Vershynin, R. (2019). High-dimensional probability. *Cambridge University Press*.

Wang, L., Bai, Y., Sun, W., and Joachims, T. (2021). Fairness of exposure in stochastic bandits. *arXiv preprint arXiv:2103.02735*.

Wang, Y., Balakrishnan, S., and Singh, A. (2018a). Optimization of smooth functions with noisy observations: Local minimax rates. *Advances in Neural Information Processing Systems*, 31.

Wang, Y., Du, S., Balakrishnan, S., and Singh, A. (2018b). Stochastic zeroth-order optimization in high dimensions. In *Proc. 21st International Conference on Artificial Intelligence and Statistics*, pages 1356–1365.

Yoo, W. W. and Ghosal, S. (2019). Bayesian mode and maximum estimation and accelerated rates of contraction. *Bernoulli*, 25(3):2330–2358.

Yu, Z., Ho, D. W. C., and Yuan, D. (2019). Distributed randomized gradient-free mirror descent algorithm for constrained optimization. *arXiv preprint arXiv:1903.04157*.

Zhang, X. and Liu, M. (2021). Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*, pages 525–555. Springer.

Zorich, V. A. (2016). *Mathematical analysis II*. Springer.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

**Titre:** Optimisation stochastique sans dérivation, apprentissage en ligne et équité

**Mots clés:** Optimisation, optimisation d'ordre zéro, apprentissage en ligne, schéma passif, bandits contextuels, équité

**Résumé:** Dans cette thèse, nous étudions d'abord le problème de l'optimisation d'ordre zéro dans le cadre actif pour des fonctions lisses et trois classes différentes de fonctions : i) les fonctions qui satisfont la condition de Polyak-Łojasiewicz, ii) les fonctions fortement convexes, et iii) la classe plus large des fonctions non convexes fortement lisses. De plus, nous proposons un nouvel algorithme basé sur la randomisation de type $\ell_1$, et nous étudions ses propriétés pour les fonctions convexes Lipschitz dans un cadre d'optimisation en ligne. Notre analyse est due à la dérivation d'une nouvelle inégalité de type Poincaré pour la mesure uniforme sur la sphère $\ell_1$ avec des constantes explicites.

Ensuite, nous étudions le problème d'optimisation d'ordre zéro dans les schémas passifs. Nous proposons une nouvelle méthode pour estimer le minimiseur et la valeur minimale d'une fonction de régression lisse et fortement convexe $f$. Nous dérivons des limites supérieures pour cet algorithme et prouvons des limites inférieures minimax pour un tel cadre.

Enfin, nous étudions le problème du bandit contextuel linéaire sous contraintes d'équité où un agent doit sélectionner un candidat dans un pool, et où chaque candidat appartient à un groupe sensible. Nous proposons une nouvelle notion d'équité qui est pratique dans l'exemple susmentionné. Nous concevons une politique avide qui calcule une estimation du rang relatif de chaque candidat en utilisant la fonction de distribution cumulative empirique, et nous prouvons sa propriété optimale.

**Title:** Derivative-free stochastic optimization, online learning and fairness

**Keywords:** Optimization, zero-order optimization, online learning, passive scheme, contextual bandits, fairness

**Abstract:** In this thesis, we first study the problem of zero-order optimization in the active setting for smooth and three different classes of functions: i) the functions that satisfy the Polyak-Łojasiewicz condition, ii) strongly convex functions, and iii) the larger class of highly smooth non-convex functions. Furthermore, we propose a novel algorithm that is based on $\ell_1$-type randomization, and we study its properties for Lipschitz convex functions in an online optimization setting. Our analysis is due to deriving a new Poincaré type inequality for the uniform measure on the $\ell_1$-sphere with explicit constants.

Then, we study the zero-order optimization problem in the passive schemes. We propose a new method for estimating the minimizer and the minimum value of a smooth and strongly convex regression function $f$. We derive upper bounds for this algorithm and prove minimax lower bounds for such a setting.

In the end, we study the linear contextual bandit problem under fairness constraints where an agent has to select one candidate from a pool, and each candidate belongs to a sensitive group. We propose a novel notion of fairness which is practical in the aforementioned example. We design a greedy policy that computes an estimate of the relative rank of each candidate using the empirical cumulative distribution function, and we proved its optimal property.

**Institut Polytechnique de Paris**
91120 Palaiseau, France