Journal of Machine Learning Research 25 (2024) 1-50 Submitted 6/23; Published 11/24

Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm

Arya Akhavan ARIA.AKHAVANFOOMANI@IIT.IT
CSML, Istituto Italiano di Tecnologia
CMAP, École Polytechnique, IP Paris

Evgenii Chzhen EVGENII.CHZHEN@CNRS.FR
CNRS, LMO, Université Paris-Saclay

Massimiliano Pontil MASSIMILIANO.PONTIL@IIT.IT
CSML, Istituto Italiano di Tecnologia
University College London

Alexandre B. Tsybakov ALEXANDRE.TSYBAKOV@ENSAE.FR
CREST, ENSAE, IP Paris

Editor: Krishnakumar Balasubramanian

Abstract

This work 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 over the ℓ_2 sphere due to Bach and Perchet (2016). We present an improved analysis of this algorithm on the class of highly smooth and strongly convex functions studied in the prior work, and we derive rates of convergence for two more general classes of non-convex functions. Namely, we consider highly smooth functions satisfying the Polyak-Łojasiewicz condition and the class of highly smooth functions with no additional property. The second algorithm is based on randomization over the ℓ_1 sphere, and it extends to the highly smooth setting the algorithm that was recently proposed for Lipschitz convex functions in Akhavan et al. (2022). We show that, in the case of noiseless oracle, this novel algorithm enjoys better bounds on bias and variance than the ℓ_2 randomization and the commonly used Gaussian randomization algorithms, while in the noisy case both ℓ_1 and ℓ_2 algorithms benefit from similar improved theoretical guarantees. The improvements are achieved thanks to a new proof techniques based on Poincaré type inequalities for uniform distributions on the ℓ_1 or ℓ_2 spheres. The results are established under weak (almost adversarial) assumptions on the noise. Moreover, we provide minimax lower bounds proving optimality or near optimality of the obtained upper bounds in several cases.

Keywords: smooth optimization, zero-order oracle, gradient-free optimization, stochastic zero-order algorithms, minimax optimality, Polyak-Łojasiewicz condition

1. Introduction

In this work, we study the problem of gradient-free optimization for certain types of smooth functions. Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\Theta \subseteq \mathbb{R}^d$. We are interested in solving the following optimization problem

$$f^\star \triangleq \inf_{\boldsymbol{x}\in\Theta} f(\boldsymbol{x}) \,,$$

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 $\boldsymbol{x} \in \mathbb{R}^d$ returns a noisy value of $f(\boldsymbol{x})$, 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

$$\boldsymbol{x}_1 \in \Theta \qquad \text{and} \qquad \boldsymbol{x}_{t+1} = \mathrm{Proj}_\Theta(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) \qquad t = 1, 2, \dots \,,$$

where $\boldsymbol{g}_t$ is a gradient estimator at the point $\boldsymbol{x}_t$, $\eta_t > 0$ is a step size, and $\mathrm{Proj}_\Theta(\cdot)$ is the Euclidean projection operator onto the set Θ. 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. Also, both algorithms invoke smoothing kernels in order to take advantage of higher order smoothness, following the approach initially suggested in (Polyak and Tsybakov, 1990). The first algorithm uses a form of ℓ_2 randomization introduced by Bach and Perchet (2016) and it has been studied in the context of gradient-free optimization of strongly convex functions in (Akhavan et al., 2020; Novitskii and Gasnikov, 2022). The second algorithm is an extension, by incorporating smoothing kernels, of the approach proposed and analysed in Akhavan et al. (2022) for online minimization of Lipschitz convex functions. It is based on an alternative randomization scheme, which uses ℓ_1-geometry in place of the ℓ_2 one.

A principal goal of this paper is to derive upper bounds on the expected optimization error of both algorithms under different assumptions on the underlying function f. These assumptions are used to set the step size in the algorithms and the perturbation parameter used in the gradient estimator. Previous works on gradient-free optimization of highly smooth functions considered mostly the strongly convex case (Polyak and Tsybakov, 1990; Bach and Perchet, 2016; Akhavan et al., 2020, 2021; Novitskii and Gasnikov, 2022). In this paper, we provide a refined analysis of the strongly convex case, improving the dependence on the dimension d and the strong convexity parameter α for the algorithm with ℓ_2 randomization, and showing analogous results for the new method with ℓ_1 randomization. For the special case of strongly convex functions with Lipschitz gradient, we find the minimax optimal dependence of the bounds on all the three parameters of the problem (namely, the horizon T, the dimension d and α) and we show that both algorithms attain the minimax rate, which equals $\alpha^{-1}d/\sqrt{T}$. This finalizes the line of work starting from (Polyak and Tsybakov, 1990), where it was proved that optimal dependence on T is of the order $1/\sqrt{T}$, and papers proving that optimal dependence on d and T scales as $d/\sqrt{T}$ (Shamir (2013) establishing a lower bound and Akhavan et al. (2020) giving a complete proof, see the discussion in Section 5). Furthermore, we complement these results by considering highly smooth but not necessary convex functions f, and highly smooth functions f, which additionally satisfy the gradient dominance (Polyak-Łojasiewicz) condition. To this end, we develop unified tools that can cover a variety of gradient estimators, and then apply them to the algorithm with ℓ_2 randomization and to our new algorithm based on ℓ_1 randomization. We show that, in the case of noiseless oracle, this novel algorithm enjoys better bounds on bias and variance than its ℓ_2 counterpart, while in the noisy case, both algorithms benefit from similar theoretical guarantees. The improvements in the analysis are achieved thanks to a new method of evaluating the bias and variance of both algorithms based on Poincaré type inequalities for uniform distributions on the ℓ_1 or ℓ_2 spheres. Moreover, we establish all our upper bounds under very weak (almost adversarial) assumptions on the noise.

1.1 Summary of the upper bounds

In this subsection, we give a high-level overview of the main contributions 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 highly smooth objective function, for which we provide a convergence rate to a stationary point. Furthermore, we study the case of α-gradient dominant f, a popular relaxation of strong convexity, which includes non-convex functions. To the best of our knowledge, the analysis of noisy zero-order optimization in these two cases is novel. In Section 5 we derive minimax lower bounds and discuss the optimality or near optimality of our convergence rates.

In the following we highlight the guarantees 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 (Bach and Perchet, 2016; Akhavan et al., 2020; Novitskii and Gasnikov, 2022), we additionally treat the constrained optimization. In this section, we only sketch our results in the case $\Theta = \mathbb{R}^d$.

We would like to emphasize that we do not assume the measurement noise to have zero mean. Moreover, the noise 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.

Rate of convergence under only smoothness assumption. Assume that f is a β-Hölder function with Lipschitz continuous gradient, where $\beta \geq 2$. Then, after $2T$ oracle queries both considered algorithms provide a point $\boldsymbol{x}_S$ satisfying

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \lesssim \left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}} \text{ under the assumption that } T \geq d^{\frac{1}{\beta}} ,$$

where S is a random variable with values in $\{1, \dots, T\}$, $\|\cdot\|$ denotes the Euclidean norm, and the sign $\lesssim$ conceals a multiplicative constant that does 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. In a related development, Ghadimi and Lan (2013); Balasubramanian and Ghadimi (2021) study zero-order optimization of non-convex objective function with Lipschitz gradient, which corresponds to $\beta = 2$. They assume querying two function values with identical noises, in which case the analysis and the convergence rates are essentially analogous to the particular case of our setting with no noise (see the discussion in Section 2.2 below). The work of Carmon et al. (2020) studies noiseless optimization of highly smooth functions assuming that the derivatives up to higher order are observed, and Arjevani et al. (2022) consider stochastic optimization with first order oracle. These papers cannot be directly compared with our work as the settings are different.

Rate of convergence under smoothness and Polyak-Łojasiewicz assumptions. Assume that f is a β-Hölder function with Lipschitz continuous gradient, with a Lipschitz constant $\bar{L}$. Additionally, let $\beta \geq 2$, and suppose that f satisfies the Polyak-Łojasiewicz inequality with a constant α. Then, after $2T$ oracle queries, both considered algorithms provide a point $\boldsymbol{x}_T$, for which the expected optimization error satisfies

$$\mathbf{E}[f(\boldsymbol{x}_T) - f^\star] \lesssim \frac{1}{\alpha}\left(\frac{\mu d^2}{T}\right)^{\frac{\beta-1}{\beta}} \quad \text{under the assumption that } T \gtrsim d^{2-\frac{\beta}{2}} \,,$$

where $\mu = \bar{L}/\alpha$, and the signs $\lesssim$ and $\gtrsim$ conceal multiplicative constants that do not depend on T, d and α. The Polyak-Łojasiewicz assumption was introduced in the context of first order optimization by Polyak (1963) who showed that it implies 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 was not previously studied. Very recently Rando et al. (2022) studied a related problem under the Polyak-Łojasiewicz assumption when querying two function values with identical noises, which can be compared with the analysis in the particular case of our setting with no noise (see the discussion in Section 2.2). Unlike our work, Rando et al. (2022) do not deal with higher order smoothness and do not derive the dependency of the bounds on d, μ and α.

Rate of convergence under smoothness and strong convexity. Assume that f is a β-Hölder function with Lipschitz continuous gradient, where $\beta \geq 2$, and satisfies α-strong convexity condition. Then, after $2T$ oracle queries, both considered algorithms provide a point $\hat{\boldsymbol{x}}_T$ such that

$$\mathbf{E}[f(\hat{\boldsymbol{x}}_T) - f^\star] \lesssim \frac{1}{\alpha}\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{\beta}} \quad \text{under the assumption that } T \geq d^{2-\frac{\beta}{2}} \,,$$

where $\lesssim$ conceals a multiplicative constant that does not depend on T, d and α. The closest result to ours is obtained in Akhavan et al. (2020) and it splits into two cases: $\beta = 2$ (Lipschitz continuous gradient) and $\beta > 2$ (higher order smoothness). For $\beta = 2$, Akhavan et al. (2020) deal with a compact Θ and prove a bound with optimal dependence on the dimension (linear in d) but sub-optimal in α, while for $\beta > 2$ they derive (for $\Theta = \mathbb{R}^d$ and for compact Θ) the rate with sub-optimal dimension factor d^2. Later, Akhavan et al. (2021) and Novitskii and Gasnikov (2022) improved the dimension factor to $d^{2-1/\beta}$ for $\beta > 2$, which still does not match the linear dependence as $\beta \to 2$. In contrast, by considering a slightly different definition of smoothness, we provide below a unified analysis leading to the dimension factor $d^{2-2/\beta}$ for any $\beta \geq 2$, under constrained and unconstrained Θ; the improvement is both in the rate and in the proof technique.

1.2 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 $\boldsymbol{x} \in \mathbb{R}^d$ we denote by $\boldsymbol{x} \mapsto \operatorname{sign}(\boldsymbol{x})$ 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 $\boldsymbol{x} \in \mathbb{R}^d$ we denote by $\operatorname{Proj}_\Theta(\boldsymbol{x}) = \operatorname{argmin}\{\|\boldsymbol{z} - \boldsymbol{x}\| \,:\, \boldsymbol{z} \in \Theta\}$ the Euclidean projection of $\boldsymbol{x}$ onto Θ. For any $p \in [1, +\infty]$ we let $\|\cdot\|_p$ be the ℓ_p-norm in $\mathbb{R}^d$ and we introduce the open ℓ_p-ball and ℓ_p-sphere, respectively, as

$$B_p^d \triangleq \left\{\boldsymbol{x} \in \mathbb{R}^d \,:\, \|\boldsymbol{x}\|_p < 1\right\} \qquad \text{and} \qquad \partial B_p^d \triangleq \left\{\boldsymbol{x} \in \mathbb{R}^d \,:\, \|\boldsymbol{x}\|_p = 1\right\} .$$

For any $\beta \geq 2$ we let $\lfloor \beta \rfloor$ be the largest integer strictly less than β. Given a multi-index $\boldsymbol{m} = (m_1, \dots, m_d) \in \mathbb{N}^d$, we set $\boldsymbol{m}! \triangleq m_1! \cdots m_d!$, $|\boldsymbol{m}| \triangleq m_1 + \cdots + m_d$.

1.3 Structure of the paper

The paper is organized as follows. In Section 2, we recall some preliminaries and introduce the classes of functions considered throughout. In Section 3, we present the two algorithms that are studied in the paper. In Section 4, we present the upper bounds for both algorithms, and in each of the considered function classes. In Section 5, we establish minimax lower bounds for the zero-order optimization problem. The proofs of most of the results are presented in the appendix.

2. Preliminaries

For any multi-index $\boldsymbol{m} \in \mathbb{N}^d$, any $|\boldsymbol{m}|$ times continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, and every $\boldsymbol{h} = (h_1, \dots, h_d)^\top \in \mathbb{R}^d$ we define

$$D^{\boldsymbol{m}} f(\boldsymbol{x}) \triangleq \frac{\partial^{|\boldsymbol{m}|} f(\boldsymbol{x})}{\partial^{m_1} x_1 \cdots \partial^{m_d} x_d}\,, \qquad \boldsymbol{h}^{\boldsymbol{m}} \triangleq h_1^{m_1} \cdots h_d^{m_d}\,.$$

For any k-linear form $A : \left(\mathbb{R}^d\right)^k \to \mathbb{R}$, we define its norm as

$$\|A\| \triangleq \sup\left\{|A[\boldsymbol{h}_1, \dots, \boldsymbol{h}_k]| \,:\, \|\boldsymbol{h}_j\| \leq 1,\; j \in [k]\right\} .$$

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 $\boldsymbol{x} \in \mathbb{R}^d$ we denote by $f^{(k)}(\boldsymbol{x}) : \left(\mathbb{R}^d\right)^k \to \mathbb{R}$ the following k-linear form:

$$f^{(k)}(\boldsymbol{x})[\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k] = \sum_{|\boldsymbol{m}_1| = \cdots = |\boldsymbol{m}_k| = 1} D^{\boldsymbol{m}_1 + \cdots + \boldsymbol{m}_k} f(\boldsymbol{x}) \boldsymbol{h}_1^{\boldsymbol{m}_1} \cdots \boldsymbol{h}_k^{\boldsymbol{m}_k}\,, \qquad \forall \boldsymbol{h}_1, \ldots, \boldsymbol{h}_k \in \mathbb{R}^d\,,$$

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 $f^{(k)}(\boldsymbol{x})$ is symmetric for all $\boldsymbol{x} \in \mathbb{R}^d$.

2.1 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 higher order Hölder smoothness.

Definition 1 (Higher order smoothness) *Fix some $\beta \geq 2$ and $L > 0$. We 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*

$$\left\| f^{(\ell)}(\boldsymbol{x}) - f^{(\ell)}(\boldsymbol{z}) \right\| \leq L \left\| \boldsymbol{x} - \boldsymbol{z} \right\|^{\beta - \ell}\,.$$

Remark 2 *Definition 1 of higher order smoothness was used by Bach and Perchet (2016) who considered only integer β, while Polyak and Tsybakov (1990); Akhavan et al. (2020, 2021) use a slightly different definition. Namely, they consider a class $\mathcal{F}'_\beta(L)$ defined as the set of all ℓ times continuously differentiable functions f satisfying for all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$, the condition*

$$|f(\boldsymbol{x}) - T^\ell_{\boldsymbol{z}}(\boldsymbol{x})| \leq L \left\| \boldsymbol{x} - \boldsymbol{z} \right\|^\beta\,,$$

where $T^\ell_{\boldsymbol{z}}(\cdot)$ is the Taylor polynomial of order $\ell = \lfloor \beta \rfloor$ of f around $\boldsymbol{z}$. If $f \in \mathcal{F}_\beta(L)$, then $f \in \mathcal{F}'_\beta(L/\ell!)$ (cf. Appendix A). Thus, the results obtained for classes $\mathcal{F}'_\beta$ hold true for $f \in \mathcal{F}_\beta$ modulo a change of constant. Moreover, if f is convex and $\beta = 2$ (Lipschitz continuous gradient) the properties defining the two classes are equivalent to within constants, cf. (Nesterov, 2018, Theorem 2.1.5).

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 $f \in \mathcal{F}_2(\bar{L})$ for some $\bar{L} > 0$, that is, the gradient of f is Lipschitz continuous.

Assumption A *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 A. The first additional assumption that we consider is the Polyak-Łojasiewicz condition, which is also referred to as α-gradient dominance. This condition became rather popular since it leads to linear convergence of the gradient descent algorithm without convexity as shown by Polyak (1963) and further discussed by Karimi et al. (2016).

Definition 3 **(α-gradient dominance)** *Let $\alpha > 0$. Function $f : \mathbb{R}^d \to \mathbb{R}$ is called α-gradient dominant on $\mathbb{R}^d$, if f is differentiable on $\mathbb{R}^d$ and satisfies Polyak-Łojasiewicz inequality,*

$$2\alpha(f(\boldsymbol{x}) - f^\star) \leq \|\nabla f(\boldsymbol{x})\|^2 \ , \qquad \forall \boldsymbol{x} \in \mathbb{R}^d \, . \tag{1}$$

Finally, we consider the second additional condition, which is the α-strong convexity.

Definition 4 **(α-strong convexity)** *Let $\alpha > 0$. Function $f : \mathbb{R}^d \to \mathbb{R}$ is called α-strongly convex on $\mathbb{R}^d$, if it is differentiable on $\mathbb{R}^d$ and satisfies*

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

We recall that α-strong convexity implies (1) (see, *e.g.,* Nesterov, 2018, Theorem 2.1.10), and thus it is a more restrictive property than α-gradient dominance.

An important example of family of functions satisfying the α-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

$$\mathcal{F}(\mathbf{A}) = \left\{ f \ : \ f(\boldsymbol{x}) = g(\mathbf{A}\boldsymbol{x}), \ \text{ g is } \alpha\text{-strongly convex} \right\} .$$

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 γ 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

$$\mathcal{F}'(\mathbf{A}) = \left\{ f \ : \ f(\boldsymbol{x}) = g(\mathbf{A}\boldsymbol{x}), \quad g \in C^2(\mathbb{R}^d), \quad g \ \text{ is strictly convex} \right\} ,$$

which is a set of α-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 $g(\mathbf{A}\boldsymbol{x}) = \sum_{i=1}^n \log(1 + \exp(\boldsymbol{a}_i^\top \boldsymbol{x}))$, where for $1 \leq i \leq n$, $\boldsymbol{a}_i$ is i-th row of $\mathbf{A}$, and $\boldsymbol{x} \in \mathbb{R}^d$. For this and more examples, see e.g. (Garrigos et al., 2023) and references therein.

In what follows, we consider three different scenarios: *(i)* the case of only smoothness assumption on f, *(ii)* smoothness and α-gradient dominance assumptions, *(iii)* smoothness and α-strong convexity assumptions. Let $\tilde{\boldsymbol{x}}$ be an output of any algorithm. For the first scenario, we obtain stationary point guarantee, that is, a bound on $\mathbf{E}[\|\nabla f(\tilde{\boldsymbol{x}})\|^2]$. For the second and the third scenarios, we provide bounds for the optimization error $\mathbf{E}[f(\tilde{\boldsymbol{x}}) - f^\star]$.

Remark 5 *Note that under α-strong convexity and the fact that $\nabla f(\boldsymbol{x}^\star) = 0$, as well as under α-gradient dominance (see,* e.g., *Karimi et al., 2016, Appendix A), for any $\boldsymbol{x} \in \mathbb{R}^d$ we have*

$$f(\boldsymbol{x}) - f^\star \geq \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \ , \tag{2}$$

where $\boldsymbol{x}^*$ is the Euclidean projection of $\boldsymbol{x}$ onto the solution set $\arg\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$ of the considered optimization problem, which is a singleton in case of strong convexity. Thus, our upper bounds on $\mathbf{E}[f(\tilde{\boldsymbol{x}}) - f^\star]$ obtained under α-strong convexity or α-gradient dominance imply immediately upper bounds for $\mathbf{E}[\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|^2]$ with an extra factor $2/\alpha$.

2.2 Classical stochastic optimization versus our setting

The classical zero-order stochastic optimization (CZSO) setting considered by Nemirovsky and Yudin (1983); Nesterov (2011); Ghadimi and Lan (2013); Duchi et al. (2015); Nesterov and Spokoiny (2017); Balasubramanian and Ghadimi (2021); Rando et al. (2022) among others assumes that there is a function $F : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ such that our target function is its expectation over the second argument,

$$f(\boldsymbol{x}) = \mathbf{E}[F(\boldsymbol{x}, \xi)],$$

where ξ is a random variable. In order to find a minimizer of f, at each step t of the algorithm one makes two queries and gets outputs of the form $(F(\boldsymbol{x}_t, \xi_t), F(\boldsymbol{x}_t + h_t\boldsymbol{\zeta}_t, \xi_t))$, $t = 1, 2, \ldots$, where ξ_t's are independent identically distributed (iid) realizations of random variable ξ, and $\boldsymbol{x}_t$, $\boldsymbol{x}_t + h_t\boldsymbol{\zeta}_t$ are query points at step t. Here, $h_t > 0$ is a perturbation parameter and $\boldsymbol{\zeta}_t$ is a random or deterministic perturbation. We emphasize two features of this setting:

(a) the two queries are obtained with the same random variable ξ_t;

(b) the random variables ξ_t are iid over t[1].

Both (a) and (b) are not assumed in our setting. On the other hand, we assume additive noise structure. That is, at step t, we can only observe the values $F(\boldsymbol{z}_t, \xi_t) = f(\boldsymbol{z}_t) + \xi_t$ for any choice of query points $\boldsymbol{z}_t$ depending only on the observations at the previous steps, but we do not assume that two queries are obtained with the same noise ξ_t or are iid. We deal with almost adversarial noise, see Assumption B below. In particular, we do not assume that the noise is zero-mean. Thus, in general, we can have $\mathbf{E}[F(\boldsymbol{x}, \xi_t)] \neq f(\boldsymbol{x})$.

In the CZSO setting, the values $F(\boldsymbol{x}_t, \xi_t)$ and $F(\boldsymbol{x}_t + h_t\boldsymbol{\zeta}_t, \xi_t)$ are used to obtain gradient approximations involving the divided differences $(F(\boldsymbol{x}_t + h_t\boldsymbol{\zeta}_t, \xi_t) - F(\boldsymbol{x}_t, \xi_t))/h_t$. A popular choice is the gradient estimator with Gaussian randomization suggested by Nesterov (2011):

$$\boldsymbol{g}_t^G \triangleq \frac{1}{h_t}(F(\boldsymbol{x}_t + h_t\boldsymbol{\zeta}_t, \xi_t) - F(\boldsymbol{x}_t, \xi_t))\boldsymbol{\zeta}_t\,, \quad \text{with} \quad \boldsymbol{\zeta}_t \sim \mathcal{N}(0, I_d)\,, \tag{3}$$

where $\mathcal{N}(0, I_d)$ denotes the standard Gaussian distribution in $\mathbb{R}^d$. In the case of additive noise, the divided differences are equal to $(f(\boldsymbol{x}_t + h_t\boldsymbol{\zeta}_t) - f(\boldsymbol{x}_t))/h_t$, that is, the analysis of these algorithms reduces to that of noiseless (deterministic) optimization setting. When ξ_t's are not additive, the assumptions that are often made in the literature on CZSO are such that the rates of convergence are the same as in the additive noise case, which is equivalent to noiseless case due to the above remark.

We can summarize this discussion as follows:

- the results obtained in the literature on CZSO, as well as some tools (e.g., averaging in Algorithm 1 of Balasubramanian and Ghadimi, 2021), do not apply in our framework;

[1] Some papers, for example, Ghadimi and Lan (2013); Gasnikov et al. (2016) relax this assumption.

- our upper bounds on the expected optimization error in the case of no noise ($\sigma = 0$) imply identical bounds in the CZSO setting when the noise is additive. Under some assumptions made in the CZSO literature (e.g., under Assumption B of Duchi et al., 2015), these bounds for $\sigma = 0$ also extend to the general CZSO setting, with possible changes only in constants and not in the rates. In other cases the rates in CZSO setting can only be slower than ours (e.g., under Assumptions A1b, A2 of Ghadimi and Lan, 2013).

3. Algorithms

Given a closed convex set $\Theta \subseteq \mathbb{R}^d$, we consider the following optimization scheme

$$\boldsymbol{x}_1 \in \Theta \qquad \text{and} \qquad \boldsymbol{x}_{t+1} = \mathrm{Proj}_\Theta(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) \qquad t \geq 1\,, \tag{4}$$

where $\boldsymbol{g}_t$ is an update direction, approximating the gradient direction $\nabla f(\boldsymbol{x}_t)$ and $\eta_t > 0$ is a step-size. Allowing one to perform two function evaluations per step, we consider two gradient estimators $\boldsymbol{g}_t$ 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

$$\int K(r)\,\mathrm{d}r{=}0, \int rK(r)\,\mathrm{d}r{=}1, \int r^j K(r)\,\mathrm{d}r{=}0,\ j{=}2,\ldots,\ell,\ \kappa_\beta{\triangleq} \int |r|^\beta |K(r)|\,\mathrm{d}r < \infty\,. \tag{5}$$

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 dealing with ℓ_2-randomized methods (see e.g., Nemirovsky and Yudin, 1983; Flaxman et al., 2005; Bach and Perchet, 2016; Akhavan et al., 2020) we use gradient estimators based on a result, which is sometimes referred to as Stokes' theorem. A general form of this result, not restricted to the ℓ_2 geometry, can be found in Akhavan et al., 2022, Appendix A.

Gradient estimator based on ℓ_2 randomization. At time $t \geq 1$, let $\boldsymbol{\zeta}_t^\circ$ be distributed uniformly on the ℓ_2-sphere ∂B_2^d, let r_t be uniformly distributed on $[-1, 1]$, and $h_t > 0$. Query two points:

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

where ξ_t, ξ_t' are noises. Using the above feedback, define the gradient estimator as

$$(\ell_2\ \texttt{randomization}) \qquad \boldsymbol{g}_t^\circ \triangleq \frac{d}{2h_t}(y_t - y_t')\boldsymbol{\zeta}_t^\circ K(r_t)\,. \tag{6}$$

We use the superscript $\circ$ to emphasize the fact that $\boldsymbol{g}_t^\circ$ is based on the ℓ_2 randomization.

Gradient estimator based on ℓ_1 randomization. At time $t \geq 1$, let $\boldsymbol{\zeta}_t^\diamond$ be distributed uniformly on the ℓ_1-sphere ∂B_1^d, let r_t be uniformly distributed on $[-1, 1]$, and $h_t > 0$. Query two points:

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

Using the above feedback, define the gradient estimator as

$$\texttt{(}\ell_1\texttt{ randomization)} \qquad \boldsymbol{g}_t^{\diamond} \triangleq \frac{d}{2h_t}(y_t - y_t')\operatorname{sign}(\boldsymbol{\zeta}_t^{\diamond})K(r_t)\,. \tag{7}$$

We use the superscript $\diamond$ reminiscent of the form of the ℓ_1-sphere in order to emphasize the fact that $\boldsymbol{g}_t^{\diamond}$ is based on the ℓ_1 randomization. The idea of using an ℓ_1 randomization (different from (7)) was probably first invoked by Gasnikov et al. (2016). We refer to Akhavan et al. (2022) who highlighted the potential computational and memory gains of another ℓ_1 randomization gradient estimator compared to its ℓ_2 counterpart, as well as its advantages in theoretical guarantees. The estimator of Akhavan et al. (2022) differs from (7) as it does not involve the kernel K but the same computational and memory advantages remain true for the estimator (7).

Throughout the paper, we impose the following assumption on the noises ξ_t, ξ_t' and on the random variables that we generate in the estimators (6) and (7).

Assumption B *For all $t \in \{1, \dots, T\}$, it holds that:*

(i) the random variables ξ_t and ξ_t' are independent from $\boldsymbol{\zeta}_t^{\circ}$ (resp. *$\boldsymbol{\zeta}_t^{\diamond}$) and from r_t conditionally on $\boldsymbol{x}_t$, and the random variables $\boldsymbol{\zeta}_t^{\circ}$* (resp. *$\boldsymbol{\zeta}_t^{\diamond}$) and r_t are independent;*

(ii) $\mathbf{E}[\xi_t^2] \le \sigma^2$ and $\mathbf{E}[(\xi_t')^2] \le \sigma^2$, where $\sigma \ge 0$.

Let us emphasize that we do not assume ξ_t and ξ_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. Nevertheless, the first part of Assumption B does not permit a completely adversarial setup. Indeed, the oracle is not allowed to choose the noise variable depending on the current randomization, i.e., $\boldsymbol{\zeta}_t^{\circ}$ (resp. $\boldsymbol{\zeta}_t^{\diamond}$) and r_t. However, Assumption B encompasses the following protocol: at each round, the oracle generates pairs of noise (ξ_t, ξ_t') with a second moment bounded by σ. This is done possibly with full knowledge of the algorithm employed by the learner, the previous actions, and the past information received by the learner.

In the next two subsections, we study the bias and variance of the two estimators. As we shall see, the ℓ_1 randomization can be more advantageous in the noiseless case than its ℓ_2 counterpart (cf. Remark 11).

3.1 Bias and variance of ℓ_2 randomization

The next results allow us to control the bias and the second moment of gradient estimators $\boldsymbol{g}_1^{\circ}, \dots, \boldsymbol{g}_T^{\circ}$, and play a crucial role in our analysis.

Lemma 6 (Bias of ℓ_2 randomization) *Let Assumption B be fulfilled. Suppose that $f \in \mathcal{F}_\beta(L)$ for some $\beta \ge 2$ and $L > 0$. Let $\boldsymbol{g}_t^{\circ}$ be defined in (6) at time $t \ge 1$. Let $\ell = \lfloor \beta \rfloor$. Then,*

$$\|\mathbf{E}[\boldsymbol{g}_t^{\circ} \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \le \kappa_\beta \frac{L}{(\ell-1)!} \cdot \frac{d}{d+\beta-1} h_t^{\beta-1}\,. \tag{8}$$

Intuitively, the smaller h_t is, the more accurately $\boldsymbol{g}_t$ estimates the gradient. A result analogous to Lemma 6 with a bigger constant was claimed in (Bach and Perchet, 2016, Lemma 2). The proof of Lemma 6 is presented in the appendix. It relies on the fact that $\boldsymbol{g}_t^\circ$ is an unbiased estimator of some surrogate function, which is strongly related to the original f. The factor in front of $h_t^{\beta-1}$ in Lemma 6 is $O(1)$ as function of d. It should be noted that for $\beta > 2$ the bounds on the bias obtained in Akhavan et al. (2020) and Novitskii and Gasnikov (2022), where the factors scale as $O(d)$ and $O(\sqrt{d})$, respectively, cannot be directly compared to Lemma 6. This is due to the fact that those bounds are proved under a different notion of smoothness, cf. Remark 2. Nevertheless, if f is convex and $\beta = 2$ both notions of smoothness coincide, and Lemma 6 improves upon the bounds in (Akhavan et al., 2020) and (Novitskii and Gasnikov, 2022) by factors of order d and $\sqrt{d}$, respectively.

Lemma 7 (Variance of ℓ_2 randomization) *Let Assumption B hold and $f \in \mathcal{F}_2(\bar{L})$ for some $\bar{L} > 0$. Then, for any $d \geq 2$,*

$$\mathbf{E}[\|\boldsymbol{g}_t^\circ\|^2] \leq \frac{d^2\kappa}{d-1}\mathbf{E}\left[\left(\|\nabla f(\boldsymbol{x}_t)\| + \bar{L}h_t\right)^2\right] + \frac{d^2\sigma^2\kappa}{h_t^2}\,,$$

where $\kappa = \int_{-1}^{1} K^2(r)\,\mathrm{d}r$.

The result of Lemma 7 can be simplified as

$$\mathbf{E}[\|\boldsymbol{g}_t^\circ\|^2] \leq 4d\kappa\mathbf{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + 4d\kappa\bar{L}^2h_t^2 + \frac{d^2\sigma^2\kappa}{h_t^2}\,, \qquad d \geq 2\,. \tag{9}$$

Let us provide some remarks about this inequality. First, the leading term of order $d^2h_t^{-2}$ in (9) 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. Indeed, unlike in those papers, the term h_t^2 is multiplied by d instead of d^2. This improvement is crucial for the condition $T \geq d^{2-\beta/2}$, under which we obtain our main results below. In particular, there is no condition on the horizon T whenever $\beta \geq 4$. On the contrary, we would need $T \geq d^3$ if we would used the previously known versions of the variance bounds (Bach and Perchet, 2016; Akhavan et al., 2020). The proof of Lemma 7 presented below relies on the Poincaré inequality for the uniform distribution on ∂B_2^d.

Proof of Lemma 7 For simplicity we drop the subscript t from all the quantities. Using Assumption B and the fact that

$$\mathbf{E}[f(\boldsymbol{x} + hr\boldsymbol{\zeta}^\circ) - f(\boldsymbol{x} - hr\boldsymbol{\zeta}^\circ) \mid \boldsymbol{x}, r] = 0 \tag{10}$$

we obtain

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{g}^\circ\|^2] &= \frac{d^2}{4h^2}\mathbf{E}\left[\left(f(\boldsymbol{x} + hr\boldsymbol{\zeta}^\circ) - f(\boldsymbol{x} - hr\boldsymbol{\zeta}^\circ) + (\xi - \xi')\right)^2 K^2(r)\right] \\ &\leq \frac{d^2}{4h^2}\left(\mathbf{E}\left[(f(\boldsymbol{x} + hr\boldsymbol{\zeta}^\circ) - f(\boldsymbol{x} - hr\boldsymbol{\zeta}^\circ)^2 K^2(r)\right] + 4\kappa\sigma^2\right)\,.\end{aligned} \tag{11}$$

Since $f \in \mathcal{F}_2(\bar{L})$ and (10) holds, then using Wirtinger-Poincaré inequality (see, *e.g.*, Osserman, 1978, (3.1)) we get

$$\mathbf{E}\left[(f(\boldsymbol{x}+hr\boldsymbol{\zeta}^\circ)-f(\boldsymbol{x}-hr\boldsymbol{\zeta}^\circ))^2 \,\middle|\, \boldsymbol{x}, r\right] \le \frac{h^2}{d-1}\mathbf{E}\left[\|\nabla f(\boldsymbol{x}+hr\boldsymbol{\zeta}^\circ)+\nabla f(\boldsymbol{x}-hr\boldsymbol{\zeta}^\circ)\|^2 \,\middle|\, \boldsymbol{x}, r\right]. \tag{12}$$

The fact that $f \in \mathcal{F}_2(\bar{L})$ and the triangle inequality imply that

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}+hr\boldsymbol{\zeta}^\circ)+\nabla f(\boldsymbol{x}-hr\boldsymbol{\zeta}^\circ)\|^2 \,\middle|\, \boldsymbol{x}, r\right] \le 4\left(\|\nabla f(\boldsymbol{x})\| + \bar{L}h\right)^2 . \tag{13}$$

Combining (11) – (13) proves the lemma. ■

Note that Eqs. (12)–(13) imply, in particular, Lemma 9 of Shamir (2017). Our method based on Poincaré's inequality yields explicitly the constants in the bound. In this aspect, we improve upon Shamir (2017), where a concentration argument leads only to non-specified constants.

3.2 Bias and variance of ℓ_1 randomization

This section analyses the gradient estimate based on the ℓ_1 randomization. The results below display a very different bias and variance behavior compared to the ℓ_2 randomization.

Lemma 8 (Bias of ℓ_1 randomization) *Let Assumption B be fulfilled and $f \in \mathcal{F}_\beta(L)$ for some $\beta \ge 2$ and $L > 0$. Let $\boldsymbol{g}_t^\diamond$ be defined in (7) at time $t \ge 1$. Let $\ell = \lfloor \beta \rfloor$. Then,*

$$\|\mathbf{E}[\boldsymbol{g}_t^\diamond \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \le L c_\beta \kappa_\beta \ell^{\beta-\ell} d^{\frac{1-\beta}{2}} h_t^{\beta-1} . \tag{14}$$

When $2 \le \beta < 3$, then $c_\beta = 2^{\frac{\beta-1}{2}}$, and if $\beta \ge 3$ we have $c_\beta = 1$.

Notice that Lemma 8 gives the same dependence on the discretization parameter h_t as Lemma 6. However, unlike the bias bound in Lemma 6, which is dimension independent, the result of Lemma 8 depends on the dimension in a favorable way. In particular, the bias is controlled by a decreasing function of the dimension and this dependence becomes more and more favorable for smoother functions. Yet, the price for such a favorable control of the bias is an inflated bound on the variance, which is established below.

Lemma 9 (Variance of ℓ_1 randomization) *Let Assumption B be fulfilled and $f \in \mathcal{F}_2(\bar{L})$ for some $\bar{L} > 0$. Then, for any $d \ge 3$,*

$$\mathbf{E}[\|\boldsymbol{g}_t^\diamond\|^2] \le \frac{8d^3\kappa}{(d+1)(d-2)}\mathbf{E}\left[\left(\|\nabla f(\boldsymbol{x}_t)\| + \bar{L}h_t\sqrt{\frac{2}{d}}\right)^2\right] + \frac{d^3\sigma^2\kappa}{h_t^2} ,$$

where $\kappa = \int_{-1}^{1} K^2(r)\,\mathrm{d}r$.

Combined with the facts that $a^2/((a-2)(a+1)) \leq 2.25$ for all $a \geq 3$ and $(a+b)^2 \leq 2a^2 + 2b^2$, the inequality of Lemma 9 can be further simplified as

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{g}_t^{\diamond}\|^2] &\leq \frac{16d^3\kappa}{(d+1)(d-1)}\mathbf{E}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 + \frac{32d^2\kappa\bar{L}^2h_t^2}{(d+1)(d-1)} + \frac{d^3\sigma^2\kappa}{h_t^2} \\ &\leq 36d\kappa\mathbf{E}\left\|\nabla f(\boldsymbol{x}_t)\right\|^2 + 72\kappa\bar{L}^2h_t^2 + \frac{d^3\sigma^2\kappa}{h_t^2}\,, \qquad d \geq 3\,.\end{aligned} \tag{15}$$

The proof of Lemma 9 is based on the following lemma, which gives a version of Poincaré's inequality improving upon the previously derived in (Akhavan et al., 2022, Lemma 3). We provide sharper constants and an easier to use expression.

Lemma 10 *Let $d \geq 3$. Assume that $G : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function, and $\boldsymbol{\zeta}^{\diamond}$ is distributed uniformly on ∂B_1^d. Then*

$$\mathrm{Var}(G(\boldsymbol{\zeta}^{\diamond})) \leq \frac{4}{d-2}\mathbf{E}\left[\|\nabla G(\boldsymbol{\zeta}^{\diamond})\|^2\,\|\boldsymbol{\zeta}^{\diamond}\|^2\right]\,.$$

Furthermore, if $G : \mathbb{R}^d \to \mathbb{R}$ is an L-Lipschitz function w.r.t. the ℓ_2-norm then

$$\mathrm{Var}(G(\boldsymbol{\zeta}^{\diamond})) \leq \frac{8L^2}{(d+1)(d-2)} \leq \frac{18L^2}{d^2}\,.$$

The proof of Lemma 10 is given in the Appendix.

Proof of Lemma 9 For simplicity we drop the subscript t from all the quantities. Similarly to the proof of Lemma 7, using Assumption B we deduce that

$$\mathbf{E}[\|\boldsymbol{g}^{\diamond}\|^2] \leq \frac{d^3}{4h^2}\left(\mathbf{E}[(f(\boldsymbol{x}+hr\boldsymbol{\zeta}^{\diamond}) - f(\boldsymbol{x}-hr\boldsymbol{\zeta}^{\diamond}))^2K^2(r)] + 4\sigma^2\kappa\right)\,. \tag{16}$$

Consider $G : \mathbb{R}^d \to \mathbb{R}$ defined for all $\boldsymbol{u} \in \mathbb{R}^d$ as $G(\boldsymbol{u}) = f(\boldsymbol{x}+hr\boldsymbol{u}) - f(\boldsymbol{x}-hr\boldsymbol{u})$. Using the fact that $f \in \mathcal{F}_2(\bar{L})$ we obtain for all $\boldsymbol{u} \in \mathbb{R}^d$

$$\|\nabla G(\boldsymbol{u})\|^2 \leq 4h^2\left(\|\nabla f(\boldsymbol{x})\| + \bar{L}h\|\boldsymbol{u}\|\right)^2\,.$$

Applying Lemma 10 to the function G defined above we deduce that

$$\mathbf{E}\left[(G(\boldsymbol{\zeta}^{\diamond}))^2 \mid \boldsymbol{x}, r\right] \leq \frac{16h^2}{d-2}\mathbf{E}\left[\left(\|\nabla f(\boldsymbol{x})\| + \bar{L}h\|\boldsymbol{\zeta}^{\diamond}\|\right)^2\|\boldsymbol{\zeta}^{\diamond}\|^2\right]\,.$$

Lemma 31 provided in the Appendix gives upper bounds on the expectations appearing in the above inequality for all $d \geq 3$. Its application yields:

$$\mathbf{E}\left[\left(f(\boldsymbol{x}+hr\boldsymbol{\zeta}^{\diamond}) - f(\boldsymbol{x}-hr\boldsymbol{\zeta}^{\diamond})\right)^2 \mid \boldsymbol{x}, r\right] \leq \frac{32h^2}{(d+1)(d-2)}\left(\|\nabla f(\boldsymbol{x})\| + \bar{L}h\sqrt{\frac{2}{d}}\right)^2\,.$$

We conclude the proof by combining this bound with (16). ■

Modulo absolute constants, the leading term *w.r.t.* h_t in Lemma 9 is the same as for the ℓ_2 randomization in Lemma 7. However, for the ℓ_2 randomization this term involves only a quadratic dependence on the dimension d, while in the case of ℓ_1 randomization the dependence is cubic. Looking at the h_t^2 term that essentially matters only when $\sigma = 0$, we note that the factor in front of this term in Lemma 9 is bounded as the dimension grows. In contrast, the corresponding term in Lemma 7 depends linearly on the dimension. We summarize these observations in the following remark focusing on the noiseless case.

Remark 11 (On the advantage of ℓ_1 randomization) *In the noiseless case ($\sigma = 0$) both bias and variance under the ℓ_1 randomization are strictly smaller than under the ℓ_2 randomization. Indeed, if $\sigma = 0$*

$$\begin{cases} \|\mathbf{E}[\boldsymbol{g}_t^{\circ} \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \lesssim h_t^{\beta-1} \\ \mathbf{E}[\|\boldsymbol{g}_t^{\circ}\|^2] \lesssim d\mathbf{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + (\sqrt{d}h_t)^2 \end{cases} \qquad \textit{and} \qquad \begin{cases} \|\mathbf{E}[\boldsymbol{g}_t^{\diamond} \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \lesssim \left(\frac{h_t}{\sqrt{d}}\right)^{\beta-1} \\ \mathbf{E}[\|\boldsymbol{g}_t^{\diamond}\|^2] \lesssim d\mathbf{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + h_t^2 \end{cases} ,$$

where the signs $\lesssim$ hide multiplicative constants that do not depend on h_t and d. Notice that given an h_t for ℓ_2 randomization, $\sqrt{d}h_t$ used with ℓ_1 randomization gives the same order of magnitude for both bias and variance. This is especially useful in the floating-point arithmetic, where very small values of h_t (on the level of machine precision) can result in high rounding errors. Thus, in the absence of noise, ℓ_1 randomization can be seen as a more numerically stable alternative to the ℓ_2 randomization.

For comparison, the corresponding bounds for gradient estimators with Gaussian randomization defined in (3) are proved for $\beta = 2$ and have the form (cf. Nesterov (2011) or (Ghadimi and Lan, 2013, Theorem 3.1)):

$$\begin{cases} \left\|\mathbf{E}[\boldsymbol{g}_t^{G} \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\right\| \lesssim d^{3/2}h_t \\ \mathbf{E}[\|\boldsymbol{g}_t^{G}\|^2] \lesssim d\mathbf{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + d^3h_t^2, \end{cases} \tag{17}$$

where the signs $\lesssim$ hide multiplicative constants that do not depend on h_t and d. Setting $\beta = 2$ in Remark 11 we see that the dependence on h_t in (17) is of the same order as for ℓ_1 and ℓ_2 randomizations with $\beta = 2$, but the dimension factors are substantially bigger. Also, the bounds in (17) for the Gaussian randomization are tight. Thus, the Gaussian randomization is less efficient than its ℓ_1 and ℓ_2 counterparts in the noiseless setting.

4. Upper bounds

In this section, we present convergence guarantees for the two considered gradient estimators and for three classes of objective functions f. Each of the following subsections is structured similarly:

first, we define the choice of η_t and h_t involved in both algorithms and then, for each class of the objective functions, we state the corresponding convergence guarantees.

Throughout this section, we assume that $f \in \mathcal{F}_2(\bar{L}) \cap \mathcal{F}_\beta(L)$ for some $\beta \geq 2$. Under this assumption, in Section 4.1 we establish a guarantee for the stationary point. In Section 4.2 we additionally assume that f is α-gradient dominant and provide upper bounds on the optimization error. In Section 4.3 we additionally assume that f is α-strongly convex and provide upper bounds on the optimization error for both constrained and unconstrained cases. 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.

4.1 Only smoothness assumptions

In this subsection, we only assume that the objective function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumption A. In particular, since there is no guarantee of the existence of the minimizer, our goal is modest – we only want to obtain a nearly stationary point.

The plan of our study is as follows. We first obtain guarantees for algorithm (4) with gradient estimator $\boldsymbol{g}_t$ satisfying some general assumption, and then concretize the results for the gradient estimators (6) and (7). We use the following assumption.

Assumption C *Assume that there exist two positive sequences* $b_t, v_t : \mathbb{N} \to [0, \infty)$ *and* $\mathsf{V}_1 \geq 0$ *such that for all* $t \geq 1$ *it holds almost surely that*

$$\|\mathbf{E}[\boldsymbol{g}_t \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \leq b_t \qquad \textit{and} \qquad \mathbf{E}[\|\boldsymbol{g}_t\|^2] \leq \mathsf{V}_1 \mathbf{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + v_t\,.$$

Note that Assumption C holds for the gradient estimators (6) and (7) with b_t, v_t and V_1 specified in Lemmas 6–9 (see also Assumption D and Table 1 below).

The results of this subsection will be stated on a randomly sampled point along the trajectory of the algorithm. The distribution over the trajectory is chosen carefully, in order to guarantee the desired convergence. The distribution that we are going to use is defined in the following lemma.

Lemma 12 *Let* $f \in \mathcal{F}_2(\bar{L})$ *for some* $\bar{L} > 0$, $\Theta = \mathbb{R}^d$ *and* $f^\star > -\infty$. *Let* $\boldsymbol{x}_t$ *be defined by algorithm* (4) *with* $\boldsymbol{g}_t$ *satisfying Assumption* C. *Assume that* η_t *in* (4) *is chosen such that* $\bar{L}\eta_t\mathsf{V}_1 < 1$. *Let* S *be a random variable with values in* $[T]$, *which is independent from* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T, \boldsymbol{g}_1, \dots, \boldsymbol{g}_T$ *and distributed with the law*

$$\mathbf{P}(S = t) = \frac{\eta_t\left(1 - \bar{L}\eta_t\mathsf{V}_1\right)}{\sum_{t=1}^T \eta_t\left(1 - \bar{L}\eta_t\mathsf{V}_1\right)}, \quad t \in [T]\,.$$

Then,

$$\mathbf{E}[\|\nabla f(\boldsymbol{x}_S)\|^2] \leq \frac{2(\mathbf{E}[f(\boldsymbol{x}_1)] - f^\star) + \sum_{t=1}^T \eta_t\left(b_t^2 + \bar{L}\eta_t v_t\right)}{\sum_{t=1}^T \eta_t\left(1 - \bar{L}\eta_t\mathsf{V}_1\right)}\,.$$

Lemma 12 is obtained by techniques similar to Ghadimi and Lan (2013). However, the paper Ghadimi and Lan (2013) considers only a particular choice of $\boldsymbol{g}_t$ defined via a Gaussian randomization, and a different setting (cf. the discussion in Section 2.2), under which v_t does not increase as the discretization parameter h_t (μ in the notation of Ghadimi and Lan (2013)) decreases. In our setting, this situation happens only when there is no noise ($\sigma = 0$), while in the noisy case v_t increases as h_t tends to 0.

Note that the distribution of S in Lemma 12 depends on the choice of η_t and V_1. In the following results, we are going to specify the exact values of η_t. We also provide the values of V_1 for the gradient estimators (6) and (7). Regarding these two estimators, it will be convenient to use the following instance of Assumption C.

Assumption D *There exist positive numbers $b, \mathsf{V}_1, \mathsf{V}_2, \mathsf{V}_3$ such that for all $t \geq 1$ the gradient estimators $\boldsymbol{g}_t$ satisfy almost surely the inequalities*

$$\|\mathbf{E}[\boldsymbol{g}_t \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \leq bLh_t^{\beta-1} \quad \textit{and} \quad \mathbf{E}[\|\boldsymbol{g}_t\|^2] \leq \mathsf{V}_1\mathbf{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}.$$

It follows from Lemmas 6–9 that Assumption D holds for gradient estimators (6) and (7) with the values that are indicated in Table 1. Note that the bounds for the variance in those lemmas do not cover the case $d = 1$ for the ℓ_2 randomization and $d = 1, 2$ for the ℓ_1 randomization. Nevertheless, it is straightforward to check that in these cases Assumption D remains valid with V_j's given in Table 1.

Estimator	b	V_1	V_2	V_3
ℓ_2 `randomization`	$\frac{\kappa_\beta}{(\ell-1)!} \cdot \frac{d}{d+\beta-1}$	$4d\kappa$	$4d\kappa$	$d^2\kappa$
ℓ_1 `randomization`	$c_\beta\kappa_\beta\ell^{\beta-\ell}d^{\frac{1-\beta}{2}}$	$36d\kappa$	72κ	$d^3\kappa$

Table 1: Factors in the bounds for bias and variance of both gradient estimators, $\ell = \lfloor\beta\rfloor$, $d \geq 1$.

The next theorem requires a definition of algorithm-dependent parameters, which are needed as an input to our algorithms. We set

$$(\mathfrak{y}, \mathfrak{h}) = \begin{cases} \left((8\kappa\bar{L})^{-1},\, d^{\frac{1}{2\beta-1}}\right) & \text{for } \ell_2 \text{ randomization,} \\ \left((72\kappa\bar{L})^{-1},\, d^{\frac{2\beta+1}{4\beta-2}}\right) & \text{for } \ell_1 \text{ randomization.} \end{cases} \tag{18}$$

Theorem 13 *Let Assumptions A and B hold, and $\Theta = \mathbb{R}^d$. Let $\boldsymbol{x}_t$ be defined by algorithm (4) with gradient estimator (6) or (7), where the parameters η_t and h_t are set for $t = 1, \dots, T$, as*

$$\eta_t = \min\left(\frac{\mathfrak{y}}{d},\, d^{-\frac{2(\beta-1)}{2\beta-1}}T^{-\frac{\beta}{2\beta-1}}\right) \qquad \textit{and} \qquad h_t = \mathfrak{h}T^{-\frac{1}{2(2\beta-1)}}\,,$$

and the constants $\mathfrak{y}$ and $\mathfrak{h}$ are given in (18)*. Assume that $\boldsymbol{x}_1$ is deterministic and $T \geq \mathrm{d}^{\frac{1}{\beta}}$. Then, for the random variable S defined in Lemma 12, we have*

$$\mathbf{E}[\|\nabla f(\boldsymbol{x}_S)\|^2] \leq \Big(\mathsf{A}_1(f(\boldsymbol{x}_1) - f^\star) + \mathsf{A}_2\Big)\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}} ,$$

where the constants $\mathsf{A}_1, \mathsf{A}_2 > 0$ depend only on $\sigma, L, \bar{L}$, β, and on the choice of the gradient estimator.

In the case $\sigma = 0$, the result of this theorem can be improved. As explained in Section 2.2, this case is analogous to the CZSO setting, and it is enough to assume that $\beta = 2$ since higher order smoothness does not lead to improvement in the main term of the rates. Due to Remark 11 (or Assumption D) one can set h_t for both methods as small as one wishes, and thus sufficiently small to make the sum over t in the numerator of the inequality of Lemma 12 less than an absolute constant. Then, choosing $\eta_t = (2\mathsf{V}_1\bar{L})^{-1}$ and recalling that, for both algorithms, V_1 scales as d up to a multiplicative constant (cf. Table 1) we get the following result.

Theorem 14 *Let f be a function belonging to $\mathcal{F}_2(\bar{L})$ for some $\bar{L} > 0$, $\Theta = \mathbb{R}^d$, and let Assumptions A and B hold with $\sigma = 0$. Let $\boldsymbol{x}_t$ be defined by algorithm* (4) *with deterministic $\boldsymbol{x}_1$, and gradient estimators* (6) *or* (7) *for $\beta = 2$, where $\eta_t = (2\mathsf{V}_1\bar{L})^{-1}$ and h_t is chosen sufficiently small. Then we have*

$$\mathbf{E}[\|\nabla f(\boldsymbol{x}_S)\|^2] \leq \mathsf{A}(f(\boldsymbol{x}_1) - f^\star + 1)\frac{\bar{L}d}{T} ,$$

where $\mathsf{A} > 0$ is an absolute constant depending only the choice of the gradient estimator.

The rate $O(d/T)$ in Theorem 14 coincides with the rate derived in (Nesterov and Spokoiny, 2017, inequality (68)) for $\beta = 2$ under the classical zero-order stochastic optimization setting, where the authors were using Gaussian rather than ℓ_1 or ℓ_2 randomization. In a setting with non-additive noise, Ghadimi and Lan (2013) exhibit a slower rate of $O(\sqrt{d/T})$.

4.2 Smoothness and α-gradient dominance

We now provide the analysis of our algorithms under smoothness and α-gradient dominance (Polyak-Łojasiewicz) conditions.

Theorem 15 *Let f be an α-gradient dominant function, $\Theta = \mathbb{R}^d$, and let Assumptions A and B hold, with $\sigma > 0$. Let $\boldsymbol{x}_t$ be defined by algorithm* (4) *with $\boldsymbol{g}_t$ satisfying Assumption D, deterministic $\boldsymbol{x}_1$ and*

$$\eta_t = \min\left((2\bar{L}\mathsf{V}_1)^{-1}, \frac{4}{\alpha t}\right), \qquad h_t = \left(\frac{4\bar{L}\sigma^2\mathsf{V}_3}{b^2L^2\alpha}\right)^{\frac{1}{2\beta}} \cdot \begin{cases} t^{-\frac{1}{2\beta}} & \textit{if } \eta_t = \frac{4}{\alpha t} \\ T^{-\frac{1}{2\beta}} & \textit{if } \eta_t = \frac{1}{2\bar{L}\mathsf{V}_1} \end{cases} .$$

Then

$$
\begin{aligned}
\mathbf{E}[f(\boldsymbol{x}_T) - f^\star] \le \mathtt{A}_1 \frac{\bar{L}\mathsf{V}_1}{\alpha T}(f(\boldsymbol{x}_1) - f^\star) \\
+ \frac{\mathtt{A}_2}{\alpha}\left(\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}} + \mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{\frac{1}{\beta}}\left(\frac{\alpha T}{\bar{L}\sigma^2}\right)^{-\frac{2}{\beta}}\sigma^{-2}\right)\left(\frac{\alpha T}{\bar{L}\sigma^2}\right)^{-\frac{\beta-1}{\beta}},
\end{aligned}
$$

where $\mathtt{A}_1, \mathtt{A}_2 > 0$ depend only on β.

Theorem 15 provides a general result for any gradient estimator that satisfies Assumption D. By taking the values V_j from Table 1 we immediately obtain the following corollary for our ℓ_1- and ℓ_2-randomized gradient estimators.

Corollary 16 *Let f be an α-gradient dominant function, $\Theta = \mathbb{R}^d$, and let Assumptions A and B hold, with $\sigma > 0$. Let $\boldsymbol{x}_t$ be defined by algorithm (4) with deterministic $\boldsymbol{x}_1$ and gradient estimators (6) or (7). Set the parameters η_t and h_t as in Theorem 15, where $b, \mathsf{V}_1, \mathsf{V}_2, \mathsf{V}_3$ are given in Table 1 for each gradient estimator, respectively. Then for any $T \ge d^{2-\frac{\beta}{2}}\frac{\sigma^2}{\alpha L^2}$ we have*

$$
\mathbf{E}[f(\boldsymbol{x}_T) - f^\star] \le \mathtt{A}_1 \frac{\bar{L}d}{\alpha T}(f(\boldsymbol{x}_1) - f^\star) + \left(\mathtt{A}_2 + \mathtt{A}_3\frac{\bar{L}^2}{\sigma^2}\right)\left(\frac{\bar{L}\sigma^2 d^2}{\alpha T}\right)^{\frac{\beta-1}{\beta}}\frac{L^{\frac{2}{\beta}}}{\alpha},
$$

where $\mathtt{A}_1, \mathtt{A}_2, \mathtt{A}_3 > 0$ depend only on β and on the choice of the gradient estimator.

Note that here we consider σ and L as numerical constants. The condition $T \gtrsim d^{2-\frac{\beta}{2}}/\alpha$ mentioned in Corollary 16 is satisfied in all reasonable cases since it is weaker than the condition $T \gtrsim d^2/\alpha$ guaranteeing non-triviality of the bounds.

Recall that, in the context of deterministic optimization with first order oracle, the α-gradient dominance allows one to obtain the rates of convergence of gradient descent algorithm, which are similar to the case of strongly convex objective function with Lipschitz gradient (Polyak (1963); Karimi et al. (2016)). A natural question is whether the same property holds in our setting of stochastic optimization with zero-order oracle and higher order smoothness. Theorem 15 shows the rates are only inflated by a multiplicative factor $\mu^{(\beta-1)/\beta}$, where $\mu = \bar{L}/\alpha$, compared to the α-strongly convex case that will be considered in Section 4.3.

Consider now the case $\sigma = 0$, which is analogous to the CZSO setting as explained in Section 2.2. In this case, we assume that $\beta = 2$ since higher order smoothness does not lead to improvement in the main term of the rates. We set the parameters η_t, h_t as follows:

$$
\eta_t = \min\left((2\bar{L}\mathsf{V}_1)^{-1}, \frac{4}{\alpha t}\right), \qquad h_t \le \left(\frac{\bar{L}\vee 1}{\alpha\wedge 1}T\left(2b^2\bar{L} + \frac{8\bar{L}^2\mathsf{V}_2}{\alpha}\right)\right)^{-\frac{1}{2}}. \tag{19}
$$

Theorem 17 *Let f be an α-gradient dominant function belonging to $\mathcal{F}_2(\bar{L})$ for some $\bar{L} > 0$, $\Theta = \mathbb{R}^d$, and let Assumptions A and B hold with $\sigma = 0$. Let $\boldsymbol{x}_t$ be defined by algorithm (4) with deterministic $\boldsymbol{x}_1$, and gradient estimators (6) or (7) for $\beta = 2$. Set the parameters η_t and h_t as in (19). Then we have*

$$\mathbf{E}[f(\boldsymbol{x}_T) - f^\star] \leq \mathtt{A}_1 \frac{\bar{L}d}{\alpha T}\left((f(\boldsymbol{x}_1) - f^\star) + \mathtt{A}_2\right) ,$$

where $\mathtt{A}_1, \mathtt{A}_2 > 0$ are absolute constants depending only the choice of the gradient estimator.

Note that, in the CZSO setting, Rando et al. (2022) proved the rate $O(T^{-1})$ for the optimization error under α-gradient dominance by using an ℓ_2 randomization gradient estimator. However, unlike Theorem 17 the bound obtained in that paper does not provide the dependence on the dimension d and on variables $\bar{L}, \alpha$.

4.3 Smoothness and strong convexity

In this subsection, we additionally assume that f is a strongly convex function and denote by $\boldsymbol{x}^\star$ its unique minimizer. We provide a guarantee on the weighted average point $\hat{\boldsymbol{x}}_T$ along the trajectory of the algorithm defined as

$$\hat{\boldsymbol{x}}_T = \frac{2}{T(T+1)}\sum_{t=1}^{T} t\boldsymbol{x}_t.$$

We consider separately the cases of unconstrained and constrained optimization.

4.3.1 UNCONSTRAINED OPTIMIZATION

In this part we assume that $\Theta = \mathbb{R}^d$ and the horizon T is known to the learner. Similar to Section 4.2, we first state a general result that can be applied to any gradient estimator satisfying Assumption D.

Theorem 18 *Let f be an α-strongly convex function, $\Theta = \mathbb{R}^d$, and let Assumptions A and B hold. Let $\boldsymbol{x}_t$ be defined by algorithm (4) with $\boldsymbol{g}_t$ satisfying Assumption D, deterministic $\boldsymbol{x}_1$ and*

$$\eta_t = \min\left(\frac{\alpha}{8\bar{L}^2\mathsf{V}_1}, \frac{4}{\alpha(t+1)}\right), \qquad h_t = \left(\frac{4\sigma^2\mathsf{V}_3}{b^2L^2}\right)^{\frac{1}{2\beta}} \cdot \begin{cases} t^{-\frac{1}{2\beta}} & \text{if } \eta_t = \frac{8}{\alpha(t+1)} \\ T^{-\frac{1}{2\beta}} & \text{if } \eta_t = \frac{\alpha}{4\bar{L}^2\mathsf{V}_1} \end{cases} .$$

Then

$$\mathbf{E}[f(\hat{\boldsymbol{x}}_T) - f^\star] \leq \mathtt{A}_1 \frac{\bar{L}^2\mathsf{V}_1}{\alpha T}\left\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\right\|^2 + \Bigg\{\mathtt{A}_2 (bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}} + \mathtt{A}_3\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}}\Bigg\}\frac{T^{-\frac{\beta-1}{\beta}}}{\alpha} ,$$

where the constants $\mathtt{A}_1, \mathtt{A}_2, \mathtt{A}_3 > 0$ depend only on β.

Subsequently, in Corollary 19, we customize the above theorem for gradient estimators (6) and (7), with assignments of η_t, h_t that are again selected based on Table 1. We also include a bound for $\mathbf{E}[\|\hat{\boldsymbol{x}}_T - \boldsymbol{x}^\star\|^2]$, which comes as an immediate consequence due to (2).

Corollary 19 *Let f be an α-strongly convex function, $\Theta = \mathbb{R}^d$, and let Assumptions A and B hold. Let $\boldsymbol{x}_t$ be defined by algorithm (4) with gradient estimator (6) or (7), and parameters η_t, h_t as in Theorem 18, where $b, \mathsf{V}_1, \mathsf{V}_2, \mathsf{V}_3$ are given in Table 1 for each gradient estimator, respectively. Let $\boldsymbol{x}_1$ be deterministic. Then for any $T \geq d^{2-\frac{\beta}{2}}\frac{\sigma^2}{L^2}$ we have*

$$\mathbf{E}[f(\hat{\boldsymbol{x}}_T) - f^\star] \leq \mathtt{A}_1 \frac{\bar{L}^2 d}{\alpha T}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \left(\mathtt{A}_2 + \mathtt{A}_3\frac{\bar{L}^2}{\sigma^2}\right)\left(\frac{d^2\sigma^2}{T}\right)^{\frac{\beta-1}{\beta}} L^{\frac{2}{\beta}}\alpha^{-1}, \tag{20}$$

$$\mathbf{E}[\|\hat{\boldsymbol{x}}_T - \boldsymbol{x}^\star\|^2] \leq 2\mathtt{A}_1 \frac{\bar{L}^2 d}{\alpha^2 T}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + 2\left(\mathtt{A}_2 + \mathtt{A}_3\frac{\bar{L}^2}{\sigma^2}\right)\left(\frac{d^2\sigma^2}{T}\right)^{\frac{\beta-1}{\beta}} L^{\frac{2}{\beta}}\alpha^{-2}, \tag{21}$$

where $\mathtt{A}_1, \mathtt{A}_2, \mathtt{A}_3 > 0$ depend only on β and on the choice of the gradient estimator.

With a slightly different definition of smoothness class (which coincides with ours for $\beta = 2$, cf. Remark 2), a result comparable to Corollary 19 is derived in (Akhavan et al., 2020, Theorem 3.2). However, that result imposes an additional condition on α (i.e., $\alpha \gtrsim \sqrt{d/T}$) and provides a bound with the dimension factor d^2 rather than $d^{2-2/\beta}$ in Corollary 19. We also note that earlier Bach and Perchet (2016) analyzed the case ℓ_2-randomized gradient estimator with integer $\beta > 2$ and proved a bound with a slower (suboptimal) rate $T^{-\frac{\beta-1}{\beta+1}}$.

4.3.2 CONSTRAINED OPTIMIZATION

We now assume that $\Theta \subset \mathbb{R}^d$ is a compact convex set. In the present part, we do not need the knowledge of the horizon T to define the updates $\boldsymbol{x}_t$. We first state the following general theorem valid when $\boldsymbol{g}_t$ is any gradient estimator satisfying Assumption D.

Theorem 20 *Let $\Theta \subset \mathbb{R}^d$ be a compact convex set. Assume that f is an α-strongly convex function, Assumptions A and B hold, and $\max_{\boldsymbol{x}\in\Theta}\|\nabla f(\boldsymbol{x})\| \leq G$. Let $\boldsymbol{x}_t$ be defined by algorithm (4) with gradient estimator $\boldsymbol{g}_t$ satisfying Assumption D and $\eta_t = \frac{4}{\alpha(t+1)}$, $h_t = \left(\frac{\sigma^2\mathsf{V}_3}{b^2L^2t}\right)^{\frac{1}{2\beta}}$. Then*

$$\mathbf{E}[f(\hat{\boldsymbol{x}}_t) - f^\star] \leq \frac{4\bar{L}^2\mathsf{V}_1G^2}{\alpha T} + \frac{\mathtt{A}_1}{\alpha}\left(\mathsf{V}_3\sigma^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{-\frac{1}{\beta}} + \mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}T^{-\frac{2}{\beta}}\right)T^{-\frac{\beta-1}{\beta}},$$

where the constant $\mathtt{A}_1 > 0$ depends only on β.

Using the bounds on the variance and bias of gradient estimators (6) and (7) from Section 3, Remark 5 and the trivial bounds $\mathbf{E}[f(\hat{\boldsymbol{x}}_T) - f^\star] \leq GB$, $\mathbf{E}[\|\hat{\boldsymbol{x}}_T - \boldsymbol{x}^\star\|^2] \leq B^2$, where B is the Euclidean diameter of Θ, we immediately obtain the following corollary.

Corollary 21 *Let $\Theta \subset \mathbb{R}^d$ be a compact convex set. Assume that f is an α-strongly convex function, Assumptions A and B hold, and $\max_{\boldsymbol{x}\in\Theta}\|\nabla f(\boldsymbol{x})\| \leq G$. Let $\boldsymbol{x}_t$ be defined by algorithm (4) with gradient estimator (6) or (7), and parameters η_t, h_t as in Theorem 20, where $b, \mathsf{V}_1, \mathsf{V}_2, \mathsf{V}_3$ are given in Table 1 for each gradient estimator, respectively. Then for any $T \geq d^{2-\frac{\beta}{2}}\frac{\sigma^2}{L^2}$ we have*

$$\mathbf{E}[f(\hat{\boldsymbol{x}}_T) - f^\star] \leq \min\left(GB, \frac{4\bar{L}^2\mathsf{V}_1 G^2}{\alpha T} + \left(\mathtt{A}_1 + \mathtt{A}_2\frac{\bar{L}^2}{\sigma^2}\right)\left(\frac{d^2\sigma^2}{T}\right)^{\frac{\beta-1}{\beta}} L^{\frac{2}{\beta}}\alpha^{-1}\right), \tag{22}$$

$$\mathbf{E}[\|\hat{\boldsymbol{x}}_T - \boldsymbol{x}^\star\|^2] \leq \min\left(B^2, \frac{2GB}{\alpha}, \frac{8\bar{L}^2\mathsf{V}_1 G^2}{\alpha^2 T} + 2\left(\mathtt{A}_1 + \mathtt{A}_2\frac{\bar{L}^2}{\sigma^2}\right)\left(\frac{d^2\sigma^2}{T}\right)^{\frac{\beta-1}{\beta}} L^{\frac{2}{\beta}}\alpha^{-2}\right), \tag{23}$$

where B is the Euclidean diameter of Θ, and $\mathtt{A}_1, \mathtt{A}_2 > 0$ depends only on β and on the choice of the gradient estimator.

In a similar setting, but assuming independent zero-mean ξ_t's, Bach and Perchet (2016) considered the case of ℓ_2 randomization and proved, for integer $\beta > 2$, a bound with suboptimal rate $T^{-\frac{\beta-1}{\beta+1}}$. Corollary 21 can be also compared to Akhavan et al. (2020, 2021) (ℓ_2 randomization and coordinate-wise radomization) and, for $\beta > 2$, to Novitskii and Gasnikov (2022) (ℓ_2 randomization). However, those papers use a slightly different definition of β-smoothness class (both definitions coincide if $\beta = 2$, see Remark 2). Their bounds guarantee the rate $O\big(\frac{d^{2-1/\beta}}{\alpha T}\big)$ for $\beta > 2$ (Akhavan et al., 2021, Corollary 6), (Novitskii and Gasnikov, 2022, Theorem 1) and $O\big(\frac{d}{\sqrt{\alpha T}}\big)$ for $\beta = 2$ (Akhavan et al., 2020, Theorem D.4) by using two different approaches for the two cases. In contrast, Corollary 21 yields $O\big(\frac{d^{2-2/\beta}}{\alpha T}\big)$ and $O\big(\frac{d}{\alpha\sqrt{T}}\big)$, respectively, and obtains these rates by a unified approach for all $\beta \geq 2$, and simultaneously under ℓ_1 and ℓ_2 randomizations. Note that, under the condition $T \geq d^2$ guaranteeing non-triviality of the bound, and $\alpha \geq 1$ the rate $O\big(\frac{d}{\alpha\sqrt{T}}\big)$ that we obtain in Corollary 21 for $\beta = 2$ matches the minimax lower bound (cf. Theorem 22 below) as a function of all the three parameters T, d, and α.

5. Lower bounds

In this section we prove minimax lower bounds on the optimization error over all sequential strategies with two-point feedback that allow the query points depend on the past. For $t = 1, \dots, T$, we assume that the values $y_t = f(\boldsymbol{z}_t) + \xi_t$ and $y_t' = f(\boldsymbol{z}_t') + \xi_t'$ are observed, where (ξ, ξ_t') are random noises, and $(\boldsymbol{z}_t, \boldsymbol{z}_t')$ are query points. We consider all strategies of choosing the query points as $\boldsymbol{z}_t = \Phi_t\big((\boldsymbol{z}_i, y_i)_{i=1}^{t-1}, (\boldsymbol{z}_i', y_i')_{i=1}^{t-1}, \boldsymbol{\tau}_t\big)$ and $\boldsymbol{z}_t' = \Phi_t'\big((\boldsymbol{z}_i, y_i)_{i=1}^{t-1}, (\boldsymbol{z}_i', y_i')_{i=1}^{t-1}, \boldsymbol{\tau}_t\big)$ for $t \geq 2$, where Φ_t's and Φ_t''s are measurable functions, $\boldsymbol{z}_1, \boldsymbol{z}_1' \in \mathbb{R}^d$ are any random variables, and $\{\boldsymbol{\tau}_t\}$ is a sequence of random variables with values in a measurable space $(\mathcal{Z}, \mathcal{U})$, such that $\boldsymbol{\tau}_t$ is independent of $\big((\boldsymbol{z}_i, y_i)_{i=1}^{t-1}, (\boldsymbol{z}_i', y_i')_{i=1}^{t-1}\big)$. We denote by Π_T the set of all such strategies of choosing query points up to $t = T$. The class Π_T includes the sequential strategy of Algorithm (4) with either of the two considered gradient estimators (6) and (7). In this case, $\boldsymbol{\tau}_t = (\boldsymbol{\zeta}_t, r_t)$, $\boldsymbol{z}_t = \boldsymbol{x}_t + h_t\boldsymbol{\zeta}_t r_t$ and $\boldsymbol{z}_t' = \boldsymbol{x}_t - h_t\boldsymbol{\zeta}_t r_t$, where $\boldsymbol{\zeta}_t = \boldsymbol{\zeta}_t^\circ$ or $\boldsymbol{\zeta}_t = \boldsymbol{\zeta}_t^\diamond$.

To state our assumption on the noises (ξ, ξ_t'), we introduce the squared Hellinger distance $H^2(\cdot,\cdot)$ defined for two probability measures $\mathbf{P}, \mathbf{P}'$ on a measurable space $(\Omega, \mathcal{A})$ as

$$H^2(\mathbf{P}, \mathbf{P}') \triangleq \int (\sqrt{\mathrm{d}\mathbf{P}} - \sqrt{\mathrm{d}\mathbf{P}'})^2 \,.$$

Assumption E *For every $t \geq 1$, the following holds:*

- *The cumulative distribution function $F_t : \mathbb{R}^2 \to \mathbb{R}$ of random variable (ξ_t, ξ_t') is such that*

$$H^2(P_{F_t(\cdot,\cdot)}, P_{F_t(\cdot+v,\cdot+v)}) \leq I_0 v^2 \,, \qquad |v| \leq v_0 \,, \tag{24}$$

  for some $0 < I_0 < \infty$, $0 < v_0 \leq \infty$. Here, $P_{F(\cdot,\cdot)}$ denotes the probability measure corresponding to the c.d.f. $F(\cdot,\cdot)$.

- *The random variable (ξ_t, ξ_t') is independent of $((\boldsymbol{z}_i, y_i)_{i=1}^{t-1}, (\boldsymbol{z}_i', y_i')_{i=1}^{t-1}, \boldsymbol{\tau}_t)$.*

Condition (24) is not restrictive and encompasses a large family of distributions. It is satisfied with small enough v_0 for distributions that correspond to regular statistical experiments, (see *e.g.,* Ibragimov and Khas'minskii, 1982, Chapter 1). If F_t is a Gaussian c.d.f. condition (24) is satisfied with $v_0 = \infty$.

To state the lower bounds, we consider a subset of the classes of functions, for which we obtained the upper bounds in Section 4. Let $\Theta = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq 1\}$. For $\alpha, L, \bar{L} > 0$, $\beta \geq 2$, let $\mathcal{F}_{\alpha,\beta}$ denote the set of all α-strongly convex functions f that satisfy Assumption A, attain their minimum over $\mathbb{R}^d$ in Θ and such that $\max_{\boldsymbol{x}\in\Theta} \|\nabla f(\boldsymbol{x})\| \leq G$, and the condition $G > \alpha$ is satisfied [2].

Theorem 22 *Let $\Theta = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq 1\}$ and let Assumption E hold. Then, for any estimator $\tilde{\boldsymbol{x}}_T$ based on the observations $((\boldsymbol{z}_t, y_t), (\boldsymbol{z}_t', y_t'), t = 1, \dots, T)$, where $((\boldsymbol{z}_t, \boldsymbol{z}_t'), t = 1, \dots, T)$ are obtained by any strategy in the class Π_T we have*

$$\sup_{f \in \mathcal{F}_{\alpha,\beta}} \mathbf{E}\big[f(\tilde{\boldsymbol{x}}_T) - f^\star\big] \geq C \min\left(\max\left(\alpha,\, T^{-1/2+1/\beta}\right),\, \frac{d}{\sqrt{T}},\, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\right), \tag{25}$$

and

$$\sup_{f \in \mathcal{F}_{\alpha,\beta}} \mathbf{E}[\|\boldsymbol{z}_T - \boldsymbol{x}^*(f)\|^2] \geq C \min\left(1,\, \frac{d}{T^{\frac{1}{\beta}}},\, \frac{d}{\alpha^2} T^{-\frac{\beta-1}{\beta}}\right), \tag{26}$$

where $C > 0$ is a constant that does not depend of T, d, and α, and $\boldsymbol{x}^\star(f)$ is the minimizer of f on Θ.

[2] The condition $G \geq \alpha$ is necessary for the class $\mathcal{F}_{\alpha,\beta}$ to be non-empty. Indeed, due to (1) and (2), for all $f \in \mathcal{F}_{\alpha,\beta}$ and $x \in \Theta$ we have $\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq G/\alpha$, and thus $2G/\alpha \geq \mathrm{diam}(\Theta) = 2$.

Some remarks are in order here. First, note that the threshold $T^{-1/2+1/\beta}$ on the strong convexity parameter α plays an important role in bounds (25) and (26). Indeed, for α below this threshold, the bounds start to be independent of α. Intuitively, it seems reasonable that α-strong convexity should be of no added value for small α. Theorem 22 allows us to quantify exactly how small such α should be, namely, $\alpha \lesssim T^{-1/2+1/\beta}$. In particular, for $\beta = 2$ the threshold occurs at $\alpha \asymp 1$. Also, quite naturally, the threshold becomes smaller when the smoothness β increases.

In the regime below the $T^{-1/2+1/\beta}$ threshold, the rate of (25) becomes $\min(T^{1/\beta}, d)/\sqrt{T}$, which is asymptotically $d/\sqrt{T}$ independently of the smoothness index β and on α. Thus, we obtain that $d/\sqrt{T}$ is a lower bound over the class of simply convex functions. On the other hand, the achievable rate for convex functions is shown to be $d^{16}/\sqrt{T}$ in Agarwal et al. (2011) and improved to $d^{4.5}/\sqrt{T}$ in Lattimore and György (2021) (both results are up to poly-logarithmic factors, and under sub-Gaussian noise ξ_t). The gap between our lower bound $d/\sqrt{T}$ and these upper bounds is only in the dependence on the dimension, but this gap is substantial. In the regime where α is above the $T^{-1/2+1/\beta}$ threshold, our results imply that the gap between upper and lower bounds is much smaller. Thus, our upper bounds in this regime scale as $\frac{d^{2-2/\beta}}{\alpha}T^{-\frac{\beta-1}{\beta}}$ while the lower bound of Theorem 22 is of the order $\frac{d}{\alpha}T^{-\frac{\beta-1}{\beta}}$.

Consider now the case $\beta = 2$. Then the lower bounds (25) and (26) are of order $d/(\max(\alpha, 1)\sqrt{T})$ and $d/(\max(\alpha^2, 1)\sqrt{T})$, respectively, under the condition $T \geq d^2$ guaranteeing non-triviality of the rates. If, in addition, $\alpha \gtrsim 1$ (meaning that α is above the threshold $\alpha \asymp 1$) we obtain the lower rates $d/(\alpha\sqrt{T})$ and $d/(\alpha^2\sqrt{T})$, respectively. Comparing this remark to Corollary 21 we obtain the following result.

Corollary 23 *Let $\beta = 2$ and let the assumptions of Theorem 22 and Corollary 21 hold. If $\alpha \geq 1$ and $T \geq \max(d^2, \bar{L}^2 d, \bar{L}^4 G^4)$ then there exist positive constants c, C that do not depend on T, d, and α such that we have the following bounds on the minimax risks:*

$$c\frac{d}{\alpha\sqrt{T}} \leq \inf_{\tilde{\boldsymbol{x}}_T} \sup_{f \in \mathcal{F}_{\alpha,\beta}} \mathbf{E}\big[f(\tilde{\boldsymbol{x}}_T) - f^\star\big] \leq C\frac{d}{\alpha\sqrt{T}}, \tag{27}$$

and

$$c\frac{d}{\alpha^2\sqrt{T}} \leq \inf_{\tilde{\boldsymbol{x}}_T} \sup_{f \in \mathcal{F}_{\alpha,\beta}} \mathbf{E}[\|\tilde{\boldsymbol{x}}_T - \boldsymbol{x}^*(f)\|^2] \leq C\frac{d}{\alpha^2\sqrt{T}}, \tag{28}$$

where $\boldsymbol{x}^\star(f)$ is the minimizer of f on Θ, and the infimum is over all estimators $\tilde{\boldsymbol{x}}_T$ based on query points obtained via strategies in the class Π_T. The minimax rates in (27) and (28) are attained by the estimator $\tilde{\boldsymbol{x}}_T = \hat{\boldsymbol{x}}_T$ with parameters as in Corollary 21.

Thus, the weighted average estimator $\hat{\boldsymbol{x}}_T$ as in Corollary 21 is minimax optimal with respect to all the three parameters T, d, and α, both in the optimization error and in the estimation risk. Note that we introduced the condition $T \geq \max(\bar{L}^2 d, \bar{L}^4 G^4)$ in Corollary 23 to guarantee that the upper bounds are of the required order, cf. Corollary 21. Thus, G is allowed to be a function of T that grows not too fast. Since $G > \alpha$ this condition also prevents α from being too large, that is, Corollary 23 does not hold if $\alpha \gtrsim T^{1/4}$.

The issue of finding the minimax optimal rates in gradient-free stochastic optimization under strong convexity and smoothness assumptions has a long history. It was initiated in Fabian (1967); Polyak and Tsybakov (1990) and more recently developed in Dippon (2003); Jamieson et al. (2012); Shamir (2013); Bach and Perchet (2016); Akhavan et al. (2020, 2021). It was shown in Polyak and Tsybakov (1990) that the minimax optimal rate on the class of α-strong convex and β-Hölder functions scales as $c(\alpha, d)T^{-(\beta-1)/\beta}$ for $\beta \geq 2$, where $c(\alpha, d)$ is an unspecified function of α and d (for $d = 1$ and integer $\beta \geq 2$ an upper bound of the same order was earlier derived in Fabian (1967)). The issue of establishing non-asymptotic fundamental limits as function of the main parameters of the problem (α, d and T) was first addressed in Jamieson et al. (2012) giving a lower bound $\Omega(\sqrt{d/T})$ for $\beta = 2$, without specifying the dependency on α. This was improved to $\Omega(d/\sqrt{T})$ when $\alpha \asymp 1$ by Shamir (2013) who also claimed that the rate $d/\sqrt{T}$ is optimal for $\beta = 2$ referring to an upper bound in Agarwal et al. (2010). However, invoking Agarwal et al. (2010) in the setting with random noise ξ_t (for which the lower bound of Shamir (2013) was proved) is not legitimate because in Agarwal et al. (2010) the observations are considered as a Lipschitz function of t. The complete proof of minimax optimality of the rate $d/\sqrt{T}$ for $\beta = 2$ under random noise was later provided in Akhavan et al. (2020). However, the upper and the lower bounds in Akhavan et al. (2020) still differ in their dependence on α. Corollary 23 completes this line of work by establishing the minimax optimality as a function of the whole triplet (T, d, α) for $\beta = 2$.

The main lines of the proof of Theorem 22 follow Akhavan et al. (2020). However, in Akhavan et al. (2020) the assumptions on the noise are much more restrictive – the random variables ξ_t are assumed iid and instead of (24) a much stronger condition is imposed, namely, a bound on the Kullback–Leibler divergence. In particular, in order to use the Kullback–Leibler divergence between two distributions we need one of them to be absolutely continuous with respect to the other. Using the Hellinger distance allows us to drop this restriction. For example, if F_t is a distribution with bounded support then the Kullback–Leibler divergence between $F_t(\cdot)$ and $F_t(\cdot + v)$ is $+\infty$ while the Hellinger distance is finite.

Acknowledgment: The work of A. Akhavan and A.B. Tsybakov was supported by the grant of French National Research Agency (ANR) "Investissements d'Avenir" LabEx Ecodec/ANR-11-LABX-0047.

References

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proc. 23rd International Conference on Learning Theory*, pages 28–40, 2010.

A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 25, pages 1035–1043, 2011.

A. Akhavan, M. Pontil, and A.B. Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. In *Advances in Neural Information Processing Systems 33*, 2020.

A. Akhavan, M. Pontil, and A. B. Tsybakov. Distributed zero-order optimization under adversarial noise. In *Advances in Neural Information Processing Systems 34*, 2021.

A. Akhavan, E. Chzhen, M. Pontil, and A.B. Tsybakov. A gradient estimator via l1-randomization for online zero-order optimization with two point feedback. In *Advances in Neural Information Processing Systems 35*, 2022.

Y. Arjevani, Y. Carmon, J. Duchi, D. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 2022.

F. Bach and V. Perchet. Highly-smooth zero-th order online optimization. In *Proc. 29th Annual Conference on Learning Theory*, 2016.

M. V. Balashov, B. T. Polyak, and A. A. Tremba. Gradient projection and conditional gradient methods for constrained nonconvex minimization. *Numerical Functional Analysis and Optimization*, 41(7):822–849, 2020.

K. Balasubramanian and S. Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42, 2021.

F. Barthe, O. Guédon, S. Mendelson, and A. Naor. A probabilistic approach to the geometry of the L_p^n ball. *The Annals of Probability*, 33(2):480–513, 2005.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–257, 2015.

Y. Carmon, J. Duchi, O Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184:71–120, 2020.

J. Dippon. Accelerated randomized stochastic optimization. *Ann. Statist.*, 31(4):1260–1281, 2003.

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

V. Fabian. Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, 38(1):191–200, 1967.

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proc. 16th Annual ACM-SIAM Symposium on Discrete algorithms (SODA)*, 2005.

G. Garrigos, L. Rosasco, and S. Villa. Convergence of the forward-backward algorithm: Beyond the worst case with the help of geometry. *Mathematical Programming*, 198:937–996, 2023.

A. Gasnikov, A. Lagunovskaya, I. Usmanova, and F. Fedorenko. Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control*, 77(11):2018–2034, 2016.

S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

I. A. Ibragimov and R. Z. Khas'minskii. Estimation of the maximum value of a signal in gaussian white noise. *Mat. Zametki*, 32(4):746–750, 1982.

K. G. Jamieson, R. Nowak, and B. Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, volume 26, pages 2672–2680, 2012.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, 2016.

T. Lattimore and A. György. Improved regret for zeroth-order stochastic convex bandits. In *Advances in Neural Information Processing Systems 34*, 2021.

A. Nemirovski. Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour 28*, 2000.

A. S. Nemirovsky and D. B Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley & Sons, 1983.

Y. Nesterov. Random gradient-free minimization of convex functions. Technical Report 2011001, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2011.

Y. Nesterov. *Lectures on Convex Optimization*. Springer, 2018.

Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17:527—566, 2017.

V. Novitskii and A. Gasnikov. Improved exploitation of higher order smoothness in derivative-free optimization. *Optimization Letters*, 16:2059–2071, 2022.

R. Osserman. The isoperimetric inequality. *Bulletin of the American Mathematical Society*, 84(6): 1182–1238, 1978.

B. T. Polyak and A. B. Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problems of Information Transmission*, 26(2):45–53, 1990.

B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3:864–878, 1963.

F. Qi and Q.-M. Luo. Bounds for the ratio of two gamma functions: from Wendel's asymptotic relation to Elezović-Giordano-Pečarić's theorem. *Journal of Inequalities and Applications*, 2013.

S. T. Rachev and L. Ruschendorf. Approximate independence of distributions on spheres and their stability properties. *The Annals of Probability*, 19(3):1311 – 1337, 1991.

M. Rando, C. Molinari, S. Villa, and L. Rosasco. Stochastic zeroth order descent with structured directions. *arXiv:2206.05124*, 2022.

G. Schechtman and J. Zinn. On the volume of the intersection of two L_p^n balls. *Proceedings of the American Mathematical Society*, 110(1):217–224, 1990.

O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proc. 30th Annual Conference on Learning Theory*, pages 1–22, 2013.

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

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

V. A. Zorich. *Mathematical analysis II*. Springer, 2016.

Appendix

In this appendix 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 $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x)\,\mathrm{d}x$.

Appendix A. Consequences of the smoothness assumption

Let us first provide some immediate consequences of the smoothness assumption that we consider.

Remark 24 *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 $\boldsymbol{m} = (m_1, \ldots, m_d)^\top \in \mathbb{N}^d$ with $|\boldsymbol{m}| = k$ there exist exactly $k!/\boldsymbol{m}!$ distinct choices of $(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k) \in (\mathbb{N}^d)^k$ 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 $\{a_1, a_2, \ldots, a_d\}$ as

$$\boldsymbol{m} \mapsto W(\boldsymbol{m}) \triangleq \underbrace{a_1 \ldots a_1}_{m_1-\text{times}} \underbrace{a_2 \ldots a_2}_{m_2-\text{times}} \ldots \underbrace{a_d \ldots a_d}_{m_d-\text{times}} .$$

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 $\boldsymbol{m} \in \mathbb{N}^d$ with $|\boldsymbol{m}| = k$. Given $(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k) \in (\mathbb{N}^d)^k$ such that $|\boldsymbol{m}_1| = \ldots = |\boldsymbol{m}_k| = 1$ and $\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_k = \boldsymbol{m}$, define[3]

$$(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k) \mapsto W(\boldsymbol{m}_1) + W(\boldsymbol{m}_2) + \ldots + W(\boldsymbol{m}_k) .$$

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 $(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k) \mapsto W(\boldsymbol{m}_1) + W(\boldsymbol{m}_2) + \ldots + W(\boldsymbol{m}_k)$ is invertible, we conclude. ■

Lemma 25 *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 ℓ times continuously differentiable, then $g_{\boldsymbol{v}}$ is $\ell - 1$ times continuously differentiable. Furthermore, for any $\boldsymbol{h}^1, \ldots, \boldsymbol{h}^{\ell-1} \in \mathbb{R}^d$

$$\begin{aligned}
g_{\boldsymbol{v}}^{(\ell-1)}(\boldsymbol{x})[\boldsymbol{h}^1, \ldots, \boldsymbol{h}^{\ell-1}] &= \sum_{|\boldsymbol{m}_1| = \ldots = |\boldsymbol{m}_{\ell-1}| = 1} D^{\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_{\ell-1}} g_{\boldsymbol{v}}(\boldsymbol{x}) h_1^{\boldsymbol{m}_1} \cdot \ldots \cdot \boldsymbol{h}_{\ell-1}^{m_{\ell-1}} \\
&= \sum_{|\boldsymbol{m}_1| = \ldots = |\boldsymbol{m}_\ell| = 1} D^{\boldsymbol{m}_1 + \ldots + \boldsymbol{m}_\ell} f(\boldsymbol{x}) h_1^{\boldsymbol{m}_1} \cdot \ldots \cdot \boldsymbol{h}_{\ell-1}^{\boldsymbol{m}_{\ell-1}} \boldsymbol{v}^{\boldsymbol{m}_\ell} \\
&= f^{(\ell)}(\boldsymbol{x})[\boldsymbol{h}^1, \ldots, \boldsymbol{h}^{\ell-1}, \boldsymbol{v}] .
\end{aligned}$$

Hence, for any $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$ we can write by definition of the norm of a $\ell-1$-linear form

$$\begin{aligned}
&\left\| g_{\boldsymbol{v}}^{(\ell-1)}(\boldsymbol{x}) - g_{\boldsymbol{v}}^{(\ell-1)}(\boldsymbol{z}) \right\| \\
&\quad = \sup\left\{ \left| g_{\boldsymbol{v}}^{(\ell-1)}(\boldsymbol{x})[\boldsymbol{h}^1, \ldots, \boldsymbol{h}^{\ell-1}] - g_{\boldsymbol{v}}^{(\ell-1)}(\boldsymbol{z})[\boldsymbol{h}^1, \ldots, \boldsymbol{h}^{\ell-1}] \right| \,:\, \|\boldsymbol{h}^j\| = 1 \; j \in [\ell-1] \right\} \\
&\quad = \sup\left\{ \left| f^{(\ell)}(\boldsymbol{x})[\boldsymbol{h}^1, \ldots, \boldsymbol{h}^{\ell-1}, \boldsymbol{v}] - f^{(\ell)}(\boldsymbol{z})[\boldsymbol{h}^1, \ldots, \boldsymbol{h}^{\ell-1}, \boldsymbol{v}] \right| \,:\, \|\boldsymbol{h}^j\| = 1 \; j \in [\ell-1] \right\} \\
&\quad \leq \left\| f^{(\ell)}(\boldsymbol{x}) - f^{(\ell)}(\boldsymbol{z}) \right\| \leq L\|\boldsymbol{x} - \boldsymbol{z}\|^{\beta-\ell} .
\end{aligned}$$

[3] The summation of words is defined as concatenation.

■

Lemma 26 *Fix some real $\beta \geq 2$ and assume that $f \in \mathcal{F}_\beta(L)$. Then, for all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$*

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

Proof Fix some $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$. Taylor expansion yields that, for some $c \in (0,1)$,

$$f(\boldsymbol{x}) = \sum_{0 \leq |\boldsymbol{m}| \leq \ell-1} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{z})^{\boldsymbol{m}} + \sum_{|\boldsymbol{m}|=\ell} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\boldsymbol{z}+c(\boldsymbol{x}-\boldsymbol{z}))(\boldsymbol{x}-\boldsymbol{z})^{\boldsymbol{m}} .$$

Thus, using Remark 24 and the fact that $f \in \mathcal{F}_\beta(L)$, we get

$$\begin{aligned}
\left| f(\boldsymbol{x}) - \sum_{|\boldsymbol{m}| \leq \ell} \frac{1}{\boldsymbol{m}!} D^{\boldsymbol{m}} f(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{z})^{m} \right| &= \left| \sum_{|\boldsymbol{m}|=\ell} \frac{1}{\boldsymbol{m}!} \left(D^{\boldsymbol{m}} f(\boldsymbol{z}+c(\boldsymbol{x}-\boldsymbol{z})) - D^{\boldsymbol{m}} f(\boldsymbol{z}) \right) (\boldsymbol{x}-\boldsymbol{z})^{\boldsymbol{m}} \right| \\
&= \frac{1}{\ell!} \left| f^{(\ell)}(\boldsymbol{z}+c(\boldsymbol{x}-\boldsymbol{z}))[\boldsymbol{x}-\boldsymbol{z}]^{\ell} - f^{(\ell)}(\boldsymbol{z})[\boldsymbol{x}-\boldsymbol{z}]^{\ell} \right| \\
&\leq \frac{L}{\ell!} \|\boldsymbol{x}-\boldsymbol{z}\|^{\ell} \|c(\boldsymbol{x}-\boldsymbol{z})\|^{\beta-\ell} \leq \frac{L}{\ell!} \|\boldsymbol{x}-\boldsymbol{z}\|^{\beta} .
\end{aligned}$$

■

Appendix B. Bias and variance of the gradient estimators

B.1 Gradient estimator with ℓ_2 randomization

In this section we prove Lemma 6 that provides a bound on the bias of the gradient estimator with ℓ_2 randomization. The variance of this estimator is evaluated in Lemma 7 in the main body of the paper.

We will need the following auxiliary lemma.

Lemma 27 *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 $[-1,1]$, B_2^d, and ∂B_2^d, respectively. Then, for any $h > 0$, we have*

$$\mathbf{E}[\nabla f(\boldsymbol{x} + hr\boldsymbol{U}^\circ) r K(r)] = \frac{d}{h} \mathbf{E}[f(\boldsymbol{x} + hr\boldsymbol{\zeta}^\circ)\boldsymbol{\zeta}^\circ K(r)] .$$

Proof Fix $r \in [-1, 1] \setminus \{0\}$. Define $\phi : \mathbb{R}^d \to \mathbb{R}$ as $\phi(\boldsymbol{u}) = f(\boldsymbol{x} + hr\boldsymbol{u})K(r)$ and note that $\nabla\phi(\boldsymbol{u}) = hr\nabla f(\boldsymbol{x} + hr\boldsymbol{u})K(r)$. Hence, we have

$$\begin{aligned}\mathbf{E}[\nabla f(\boldsymbol{x} + hr\boldsymbol{U}^\circ)K(r) \mid r] = \frac{1}{hr}\mathbf{E}[\nabla\phi(\boldsymbol{U}^\circ) \mid r] &= \frac{d}{hr}\mathbf{E}[\phi(\boldsymbol{\zeta}^\circ)\boldsymbol{\zeta}^\circ \mid r]\\ &= \frac{d}{hr}K(r)\mathbf{E}[f(\boldsymbol{x} + hr\boldsymbol{\zeta}^\circ)\boldsymbol{\zeta}^\circ \mid r]\,,\end{aligned}$$

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 a continuous distribution, and taking the total expectation concludes the proof. ■

Proof of Lemma 6 Using Lemma 27, the fact that $\int_{-1}^{1} rK(r)\,\mathrm{d}r = 1$, and the variational representation of the Euclidean norm, we can write

$$\|\mathbf{E}[\boldsymbol{g}_t^\circ \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| = \sup_{\boldsymbol{v}\in\partial B_2^d} \mathbf{E}\big[\big(\nabla_{\boldsymbol{v}} f(\boldsymbol{x} + h_t r_t\boldsymbol{U}^\circ) - \nabla_{\boldsymbol{v}} f(\boldsymbol{x})\big)r_t K(r_t)\big]\,, \tag{29}$$

where we recall that $\boldsymbol{U}^\circ$ is uniformly distributed on B_2^d. Lemma 25 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, due to Lemma 26 we have the following Taylor expansion:

$$\nabla_{\boldsymbol{v}} f(\boldsymbol{x}_t + h_t r_t\boldsymbol{U}^\circ) = \nabla_{\boldsymbol{v}} f(\boldsymbol{x}_t) + \sum_{1\le|\boldsymbol{m}|\le\ell-1} \frac{(r_t h_t)^{|\boldsymbol{m}|}}{\boldsymbol{m}!} D^{\boldsymbol{m}}\nabla_{\boldsymbol{v}} f(\boldsymbol{x}_t)(\boldsymbol{U}^\circ)^{\boldsymbol{m}} + R(h_t r_t\boldsymbol{U}^\circ)\,, \tag{30}$$

where the residual term $R(\cdot)$ satisfies $|R(\boldsymbol{x})| \le \frac{L}{(\ell-1)!}\|\boldsymbol{x}\|^{\beta-1}$.

Substituting (30) in (29) and using the "zeroing-out" properties of the kernel K, we deduce that

$$\|\mathbf{E}[\boldsymbol{g}_t^\circ \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \le \kappa_\beta h_t^{\beta-1}\frac{L}{(\ell-1)!}\mathbf{E}\,\|\boldsymbol{U}^\circ\|^{\beta-1} = \kappa_\beta h_t^{\beta-1}\frac{L}{(\ell-1)!}\frac{d}{d+\beta-1}\,,$$

where the last equality is obtained from the fact that $\mathbf{E}\,\|\boldsymbol{U}^\circ\|^q = \frac{d}{d+q}$, for any $q \ge 0$. ■

B.2 Gradient estimator with ℓ_1 randomization

In this section, we prove Lemma 8 that gives a bound on the bias of our gradient estimator with ℓ_1 randomization, and Lemma 10, that provides a Poincaré type inequality crucial for the control of its variance.

Let ζ be a real valued random variable with $\mathbf{E}[\zeta^2] \le 4\sigma^2$ and let $\boldsymbol{\zeta}^\diamond$ be distributed uniformly on ∂B_1^d. Assume that ζ and $\boldsymbol{\zeta}^\diamond$ are independent from each other and from the random variable r, which is uniformly distributed on $[-1, 1]$. In order to control the bias and variance of the gradient estimator (7) for any fixed t, it is sufficient to do it for the random vector

$$\boldsymbol{g}_{\boldsymbol{x},h}^\diamond = \frac{d}{2h}\left(f(\boldsymbol{x} + hr\boldsymbol{\zeta}^\diamond) - f(\boldsymbol{x} - hr\boldsymbol{\zeta}^\diamond) + \zeta\right)\mathrm{sign}(\boldsymbol{\zeta}^\diamond)K(r), \tag{31}$$

where $\zeta = \xi_t - \xi_t'$.

B.2.1 CONTROL OF THE BIAS

Lemma 28 *Let $\boldsymbol{U}^\diamond$ be uniformly distributed on B_1^d and $\boldsymbol{\zeta}^\diamond$ be uniformly distributed on ∂B_1^d. Fix some $\boldsymbol{x} \in \mathbb{R}^d$ and $h > 0$, and let Assumption B be fulfilled, then the estimator in (31) satisfies*

$$\mathbf{E}[\boldsymbol{g}^\diamond_{\boldsymbol{x},h}] = \mathbf{E}[\nabla f(\boldsymbol{x} + hr\boldsymbol{U}^\diamond) rK(r)]\,.$$

Proof The proof is analogous to that of Lemma 27 using (Akhavan et al., 2022, Theorem 6). ■

In order to obtain a bound on the bias of the estimator in (31) we need the following result, which controls the moments of the Euclidean norm of $\boldsymbol{U}^\diamond$.

Lemma 29 *Let $\boldsymbol{U}^\diamond \in \mathbb{R}^d$ be distributed uniformly on B_1^d. Then for any $\beta \geq 1$ it holds that*

$$\mathbf{E}[\|\boldsymbol{U}^\diamond\|^\beta] \leq \frac{c_{\beta+1} d^{\frac{\beta}{2}} \Gamma(\beta+1)\Gamma(d+1)}{\Gamma(d+\beta+1)}\,,$$

where $c_{\beta+1} = 2^{\beta/2}$ for $1 \leq \beta < 2$ and $c_{\beta+1} = 1$ for $\beta \geq 2$.

Proof Let $\boldsymbol{W} = (W_1, \ldots, W_d), W_{d+1}$ be 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

$$\boldsymbol{U}^\diamond \overset{d}{=} \frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_1 + |W_{d+1}|}\,,$$

where the sign $\overset{d}{=}$ stands for equality in distribution. Furthermore, it follows from (Barthe et al., 2005, Theorem 2) (see also Rachev and Ruschendorf (1991); Schechtman and Zinn (1990)) that

$$\frac{(\boldsymbol{W}, |W_{d+1}|)}{\|\boldsymbol{W}\|_1 + |W_{d+1}|} \qquad \text{and} \qquad \|\boldsymbol{W}\|_1 + |W_{d+1}|\,,$$

are independent. Hence,

$$\mathbf{E}[\|\boldsymbol{U}^\diamond\|^\beta] = \mathbf{E}\left[\left(\frac{\sum_{j=1}^d W_j^2}{(\|\boldsymbol{W}\|_1 + |W_{d+1}|)^2}\right)^{\beta/2}\right] = \frac{\mathbf{E}[\|\boldsymbol{W}\|^\beta]}{\mathbf{E}[\|(\boldsymbol{W}, W_{d+1})\|_1^\beta]}\,, \tag{32}$$

where the equality follows from the independence recalled above. Note that $|W_j|$ is $\exp(1)$ random variable for any $j = 1, \ldots, d$. Thus, if $1 \leq \beta < 2$ by Jensen's inequality we can write

$$\mathbf{E}[\|\boldsymbol{W}\|^\beta] = \mathbf{E}\left(\sum_{j=1}^d W_j^2\right)^{\frac{\beta}{2}} \leq \left(\sum_{j=1}^d \mathbf{E}[W_j^2]\right)^{\frac{\beta}{2}} = d^{\frac{\beta}{2}} \left(\mathbf{E}[W_1^2]\right)^{\frac{\beta}{2}} = d^{\frac{\beta}{2}} \Gamma(3)^{\frac{\beta}{2}}\,. \tag{33}$$

If $\beta \geq 2$, again by Jensen's inequality we have

$$\mathbf{E}[\|\boldsymbol{W}\|^{\beta}] = d^{\frac{\beta}{2}}\mathbf{E}\left(\frac{1}{d}\sum_{j=1}^{d} W_j^2\right)^{\frac{\beta}{2}} \leq d^{\frac{\beta}{2}-1}\sum_{j=1}^{d}\mathbf{E}[W_j^{\beta}] = d^{\frac{\beta}{2}}\mathbf{E}[W_1^{\beta}] = d^{\frac{\beta}{2}}\Gamma(\beta+1)\,. \tag{34}$$

It remains to provide a suitable expression for $\mathbf{E}[\|(\boldsymbol{W}, W_{d+1})\|_1^{\beta}]$. We observe that $\|(\boldsymbol{W}, W_{d+1})\|_1$ follows the Erlang distribution with parameters $(d+1, 1)$ (as a sum of $d+1$ i.i.d. $\exp(1)$ random variables). Hence, using the expression for the density of the Erlang distribution we get

$$\mathbf{E}[\|(\boldsymbol{W}, W_{d+1})\|_1^{\beta}] = \frac{1}{\Gamma(d+1)}\int_0^{\infty} x^{d+\beta}\exp(-x)\,\mathrm{d}x = \frac{\Gamma(d+\beta+1)}{\Gamma(d+1)}\,. \tag{35}$$

Combining (32)–(35) proves the lemma. ■

Proof of Lemma 8 Using Lemma 28 and following the same lines as in the proof of Lemma 6 we deduce that

$$\|\mathbf{E}[\boldsymbol{g}_t^{\diamond} \mid \boldsymbol{x}_t] - \nabla f(\boldsymbol{x}_t)\| \leq \kappa_\beta h_t^{\beta-1}\frac{L}{(\ell-1)!}\mathbf{E}\,\|\boldsymbol{U}^{\diamond}\|^{\beta-1} \leq \kappa_\beta h_t^{\beta-1}\frac{L}{(\ell-1)!}\frac{c_\beta d^{\frac{\beta-1}{2}}\Gamma(\beta)\Gamma(d+1)}{\Gamma(d+\beta)}\,,$$

where the last inequality is due to Lemma 29. Next, recall that the Gamma function satisfies $\Gamma(z+1) = z\Gamma(z)$ for any $z > 0$. Applying this relation iteratively and using the fact that $\ell = \lfloor \beta \rfloor$ we get:

$$\frac{\Gamma(d+1)}{\Gamma(d+\beta)} = \frac{\Gamma(d+1)}{\Gamma\big(d+\underbrace{(\beta-\ell)}_{\in(0,1]}\big)\prod_{i=1}^{\ell}\big(d+\beta-i\big)} \leq \frac{(d+\beta-\ell)^{1-(\beta-\ell)}}{\prod_{i=1}^{\ell}\big(d+\beta-i\big)} \leq \frac{1}{d^{\beta-1}}\,,$$

where the first inequality is obtained from (Qi and Luo, 2013, Remark 1). Proceeding analogously we obtain that $\frac{\Gamma(\beta)}{(\ell-1)!} \leq \ell^{\beta-\ell}$. Combining this bound with the two preceding displays yields the lemma. ■

B.2.2 POINCARÉ INEQUALITY FOR THE CONTROL OF THE VARIANCE

We now prove the Poincaré inequality of Lemma 10 used to control the variance of the ℓ_1-randomized estimator.

Proof of Lemma 10 The beginning of the proof is the same as in (Akhavan et al., 2022, Lemma 3). In particular, without loss of generality we assume that $\mathbf{E}[G(\boldsymbol{\zeta})] = 0$, and consider first the case of continuously differentiable G. Let $\boldsymbol{W} = (W_1, \ldots, W_d)$ be a vector such that the components

W_j are i.i.d. Laplace random variables with mean 0 and scale parameter 1. Set $\boldsymbol{T}(\boldsymbol{w}) = \boldsymbol{w}/\left\|\boldsymbol{w}\right\|_1$. Lemma 1 in Schechtman and Zinn (1990) asserts that, for $\boldsymbol{\zeta}$ uniformly distributed on ∂B_1^d,

$$\boldsymbol{T}(\boldsymbol{W}) \overset{d}{=} \boldsymbol{\zeta} \quad \text{and} \quad \boldsymbol{T}(\boldsymbol{W}) \text{ is independent of } \|\boldsymbol{W}\|_1 . \tag{36}$$

Furthermore, in the proof of Lemma 3 in Akhavan et al. (2022), it is shown that

$$\operatorname{Var}(G(\boldsymbol{\zeta})) \leq \frac{4}{d(d-2)}\mathbf{E}\left[\left\|\mathbf{I} - \boldsymbol{T}(\boldsymbol{W})\big(\operatorname{sign}(\boldsymbol{W})\big)^\top\right\|^2 \|\nabla G(\boldsymbol{T}(\boldsymbol{W}))\|^2\right], \tag{37}$$

where $\mathbf{I}$ is the identity matrix, $\|\cdot\|$ applied to matrices denotes the spectral norm, and $\operatorname{sign}(\cdot)$ applied to vectors denotes the vector of signs of the coordinates.

From this point, the proof diverges from that of Lemma 3 in of Akhavan et al. (2022). Instead of bounding the spectral norm of $\mathbf{I} - \boldsymbol{a}\boldsymbol{b}^\top$ by $1 + \|\boldsymbol{a}\|\,\|\boldsymbol{b}\|$ as it was done in that paper, we compute it exactly, which leads to the main improvement. Namely, Lemma 30 proved below gives

$$\left\|\mathbf{I} - \boldsymbol{T}(\boldsymbol{W})\big(\operatorname{sign}(\boldsymbol{W})\big)^\top\right\|^2 = d\,\|\boldsymbol{T}(\boldsymbol{W})\|^2 .$$

Combining this equality with (37) we obtain the first bound of the lemma. The second bound of the lemma (regarding Lipschitz functions G) is deduced from the first one by the same argument as in Akhavan et al. (2022). ■

Lemma 30 *Let $\boldsymbol{a} \in \mathbb{R}^d$ be such that $\|\boldsymbol{a}\|_1 = 1$. Then,*

$$\left\|\mathbf{I} - \boldsymbol{a}\big(\operatorname{sign}(\boldsymbol{a})\big)^\top\right\| = \sqrt{d}\,\|\boldsymbol{a}\| .$$

Proof of Lemma 30 Let $\boldsymbol{u} = \boldsymbol{a}/\left\|\boldsymbol{a}\right\|$, $\boldsymbol{v} = \operatorname{sign}(\boldsymbol{a})/\sqrt{d}$, and $\gamma = \sqrt{d}\,\|\boldsymbol{a}\|$. Then, since $1 = \|\boldsymbol{a}\|_1 = \langle \boldsymbol{a}\,, \operatorname{sign}(\boldsymbol{a})\rangle$ we have $\langle \boldsymbol{u}\,, \boldsymbol{v}\rangle = 1/\gamma$. Consider the matrix $\mathbf{Q} = [\boldsymbol{u}, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_d]$, such that $\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$. Let $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top$. For any matrix $\mathbf{B} \in \mathbb{R}^{d\times d}$ we have $\|\mathbf{B}\| = \left\|\mathbf{Q}^\top\mathbf{B}\mathbf{Q}\right\|$ and $\|\mathbf{B}\|^2 = \left\|\mathbf{B}\mathbf{B}^\top\right\|$. Using these remarks and the fact that $\|\mathbf{Q}\boldsymbol{v}\|^2 = 1$, $\mathbf{Q}^\top\boldsymbol{u} = \boldsymbol{e}_1$, we deduce that

$$\begin{aligned}\left\|\mathbf{I} - \boldsymbol{a}\big(\operatorname{sign}(\boldsymbol{a})\big)^\top\right\|^2 &= \left\|(\mathbf{I} - \gamma\boldsymbol{e}_1(\mathbf{Q}^\top\boldsymbol{v})^\top)(\mathbf{I} - \gamma\boldsymbol{e}_1(\mathbf{Q}^\top\boldsymbol{v})^\top)^\top\right\| \\ &= \left\|\mathbf{I} - \gamma\boldsymbol{e}_1(\mathbf{Q}^\top\boldsymbol{v})^\top - \gamma(\mathbf{Q}^\top\boldsymbol{v})\boldsymbol{e}_1 + \gamma^2\boldsymbol{e}_1\boldsymbol{e}_1^\top\right\| = \|\mathbf{A}\| ,\end{aligned}$$

where

$$\mathbf{A} = \left[\begin{array}{c|c} \gamma^2 - 1 & -\gamma\bar{\boldsymbol{v}}^\top \\ \hline -\gamma\bar{\boldsymbol{v}} & \mathbf{I} \end{array}\right] ,$$

with $(\bar{\boldsymbol{v}})_j = \left\langle \boldsymbol{q}_{j+1}\,, \boldsymbol{v}\right\rangle$ for $j = 1, \ldots, d-1$. Let us find the eigenvalues of $\mathbf{A}$. For any $\lambda \in \mathbb{R}$, using the expression for the determinant of a block matrix we get

$$\det(\mathbf{A} - \lambda \mathbf{I}) = (1-\lambda)^{d-1}\left(\gamma^2 - 1 - \lambda - \frac{\gamma^2 \left\|\bar{\boldsymbol{v}}\right\|^2}{1-\lambda}\right)\,.$$

Note that $1 = \left\|\mathbf{Q}\boldsymbol{v}\right\|^2 = \frac{1}{\gamma^2} + \left\|\bar{\boldsymbol{v}}\right\|^2$. Hence,

$$\det(\mathbf{A} - \lambda \mathbf{I}) = (1-\lambda)^{d-2}\big((1-\lambda)(\gamma^2 - 1 - \lambda) - (\gamma^2 - 1)\big) = (1-\lambda)^{d-2}(\lambda - \gamma^2)\lambda\,.$$

Thus, $\|\mathbf{I} - \boldsymbol{a}\big(\operatorname{sign}(\boldsymbol{a})\big)^\top\| = \max\{\gamma, 1\} = \max\{\sqrt{d}\left\|\boldsymbol{a}\right\|, 1\}$. We conclude the proof by observing that $\sqrt{d}\left\|\boldsymbol{a}\right\| \geq \left\|\boldsymbol{a}\right\|_1 = 1$. ■

Finally, we provide the following auxiliary lemma used in the proof of Lemma 9.

Lemma 31 *For all $d \geq 3$ and all $\boldsymbol{c} \in \mathbb{R}^d, h > 0$ it holds that*

$$\mathbf{E}(\|\boldsymbol{c}\| + h\left\|\boldsymbol{\zeta}^\diamond\right\|)^2 \left\|\boldsymbol{\zeta}^\diamond\right\|^2 \leq \frac{2}{d+1}\left(\|\boldsymbol{c}\| + h\sqrt{\frac{2}{d}}\right)^2\,.$$

Proof Observe that the vector $|\boldsymbol{\zeta}^\diamond| \triangleq (|\zeta_1^\diamond|, \ldots, |\zeta_d^\diamond|)^\top$ follows the Dirichlet distribution (i.e., the uniform distribution of the probability simplex on d atoms). In what follows we will make use of the following expression for the moments of the Dirichlet distribution:

$$\mathbf{E}[(\boldsymbol{\zeta}^\diamond)^{\boldsymbol{m}}] = \frac{\Gamma(d)}{\Gamma(d + |\boldsymbol{m}|)}\prod_{i=1}^{d}\Gamma(m_i + 1) = \frac{(d-1)!\boldsymbol{m}!}{(d-1+|\boldsymbol{m}|)!}\,, \tag{38}$$

for any multi-index $\boldsymbol{m} = (m_1, \ldots, m_d) \in \mathbb{N}^d$ with even coordinates.

Using (38) we get

$$\mathbf{E}\left\|\boldsymbol{\zeta}^\diamond\right\|^2 = \frac{2}{d+1}\,. \tag{39}$$

Furthermore, using the multinomial identity and the expression for the moments in (38) we find

$$\mathbf{E}\left\|\boldsymbol{\zeta}^\diamond\right\|^4 = \sum_{|\boldsymbol{m}|=2}\frac{2}{\boldsymbol{m}!}\mathbf{E}[(\boldsymbol{\zeta}^\diamond)^{2\boldsymbol{m}}] = \sum_{|\boldsymbol{m}|=2}\frac{2}{\boldsymbol{m}!}\cdot\frac{(d-1)!(2\boldsymbol{m})!}{(d+3)!} = 2\frac{(d-1)!}{(d+3)!}\sum_{|\boldsymbol{m}|=2}\frac{(2\boldsymbol{m})!}{\boldsymbol{m}!}\,.$$

Direct calculations show that $\sum_{|\boldsymbol{m}|=2}\frac{(2\boldsymbol{m})!}{\boldsymbol{m}!} = 2d(d+5)$. Hence, we deduce that, for all $d \geq 1$,

$$\mathbf{E}\left\|\boldsymbol{\zeta}^\diamond\right\|^4 = \frac{4d!(d+5)}{(d+3)!}\,.$$

Note that $\frac{d(d+5)}{(d+2)(d+3)} \leq 1$ for all $d \geq 1$. Thus,

$$\mathbf{E}\left\|\boldsymbol{\zeta}^{\diamond}\right\|^4 = \frac{4d!(d+5)}{(d+3)!} = \frac{4(d+5)}{(d+1)(d+2)(d+3)} \leq \frac{4}{d(d+1)}\,. \tag{40}$$

Finally, observe that by the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbf{E}(\|\boldsymbol{c}\| + h\left\|\boldsymbol{\zeta}^{\diamond}\right\|)^2\left\|\boldsymbol{\zeta}^{\diamond}\right\|^2 &\leq h^2\mathbf{E}\left\|\boldsymbol{\zeta}^{\diamond}\right\|^4 + 2h\|\boldsymbol{c}\|\sqrt{\mathbf{E}\left\|\boldsymbol{\zeta}^{\diamond}\right\|^2\mathbf{E}\left\|\boldsymbol{\zeta}^{\diamond}\right\|^4} + \|\boldsymbol{c}\|^2\mathbf{E}\left\|\boldsymbol{\zeta}^{\diamond}\right\|^2\\ &= \left(\|\boldsymbol{c}\|\sqrt{\mathbf{E}\left\|\boldsymbol{\zeta}^{\diamond}\right\|^2} + h\sqrt{\mathbf{E}\left\|\boldsymbol{\zeta}^{\diamond}\right\|^4}\right)^2\,.\end{aligned}$$

Combining this bound with (39) and (40) concludes the proof. ■

Appendix C. A technical lemma

In this section, we provide a lemma, which will be useful to handle recursive relations in the main proofs. It is a direct extension of (Akhavan et al., 2020, Lemma D.1).

Lemma 32 *Let $\{\delta_t\}_{t\geq 1}$ be a sequence of real numbers such that for all integers $t > t_0 \geq 1$,*

$$\delta_{t+1} \leq \left(1 - \frac{c}{t}\right)\delta_t + \sum_{i=1}^{N}\frac{a_i}{t^{p_i+1}}\,, \tag{41}$$

where $c \geq 1$, $p_i \in (0, c)$ and $a_i \geq 0$ for $i \in [N]$. Then for $t \geq t_0 \geq c+1$, we have

$$\delta_t \leq \frac{2(t_0-1)\delta_{t_0}}{t} + \sum_{i=1}^{N}\frac{a_i}{(c-p_i)t^{p_i}}\,. \tag{42}$$

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 $\frac{1}{t^p} - \frac{p}{t^{p+1}} = \frac{(c-p)+(t-c)}{t^{p+1}} \leq \frac{1}{(t+1)^p}$,

$$\frac{a_i}{t^{p+1}} \leq \frac{a_i}{c-p}\left\{\frac{1}{(t+1)^p} - \left(1 - \frac{c}{t}\right)\frac{1}{t^p}\right\}\,. \tag{43}$$

Using (41), (43) and rearranging terms, for any $t \geq t_0$ we get

$$\delta_{t+1} - \sum_{i=1}^{N}\frac{a_i}{(c-p_i)(t+1)^{p_i}} \leq \left(1 - \frac{c}{t}\right)\left\{\delta_t - \sum_{i=1}^{N}\frac{a_i}{(c-p_i)t^{p_i}}\right\}\,.$$

Letting $\tau_t = \delta_t - \sum_{i=1}^N \frac{a_i}{(c-p_i)t^{p_i}}$ we have $\tau_{t+1} \le (1-\frac{c}{t})\tau_t$. Now, if $\tau_{t_0} \le 0$ then $\tau_t \le 0$ for any $t \ge t_0$ and thus (42) holds. Otherwise, if $\tau_{t_0} > 0$ then for $t \ge t_0 + 1$ we have

$$\tau_t \le \tau_{t_0} \prod_{i=t_0}^{t-1}\left(1 - \frac{c}{i}\right) \le \tau_{t_0} \prod_{i=t_0}^{t-1}\left(1 - \frac{1}{i}\right) \le \frac{(t_0-1)\tau_{t_0}}{t} \le \frac{2(t_0-1)\delta_{t_0}}{t} .$$

Thus, (42) holds in this case as well. ■

Appendix D. Upper bounds

D.1 Upper bounds: Only smoothness assumption

Proof of Lemma 12 For brevity we write $\mathbf{E}_t[\cdot]$ in place of $\mathbf{E}[\cdot \mid \boldsymbol{x}_t]$. Using Lipschitz continuity of ∇f and the definition of the algorithm in (4) we can write

$$\begin{aligned}\mathbf{E}_t[f(\boldsymbol{x}_{t+1})] &\le f(\boldsymbol{x}_t) - \eta_t \left\langle \nabla f(\boldsymbol{x}_t), \mathbf{E}_t[\boldsymbol{g}_t]\right\rangle + \frac{\bar{L}\eta_t^2}{2}\mathbf{E}_t\left[\|\boldsymbol{g}_t\|^2\right]\\ &\le f(\boldsymbol{x}_t) - \eta_t\|\nabla f(\boldsymbol{x}_t)\|^2 + \eta_t\|\nabla f(\boldsymbol{x}_t)\|\,\|\mathbf{E}_t[\boldsymbol{g}_t] - \nabla f(\boldsymbol{x}_t)\| + \frac{\bar{L}\eta_t^2}{2}\mathbf{E}_t\left[\|\boldsymbol{g}_t\|^2\right] .\end{aligned}$$

Furthermore, invoking the assumption on the bias and the variance of $\boldsymbol{g}_t$ and using the fact that $2ab \le a^2 + b^2$ we deduce

$$\begin{aligned}\mathbf{E}_t[f(\boldsymbol{x}_{t+1})] - f(\boldsymbol{x}_t) &\le -\eta_t\|\nabla f(\boldsymbol{x}_t)\|^2 + \eta_t b_t\|\nabla f(\boldsymbol{x}_t)\| + \frac{\bar{L}\eta_t^2}{2}\left(v_t + \mathsf{V}_1\|\nabla f(\boldsymbol{x}_t)\|^2\right)\\ &\le -\eta_t\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{\eta_t}{2}\left(b_t^2 + \|\nabla f(\boldsymbol{x}_t)\|^2\right) + \frac{\bar{L}\eta_t^2}{2}\left(v_t + \mathsf{V}_1\|\nabla f(\boldsymbol{x}_t)\|^2\right)\\ &= -\frac{\eta_t}{2}\left(1 - \bar{L}\eta_t\mathsf{V}_1\right)\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{\eta_t}{2}\left(b_t^2 + \bar{L}\eta_t v_t\right) .\end{aligned} \tag{44}$$

Let S be a random variable with values in $\{1,\dots,T\}$, which is independent from $\boldsymbol{x}_1,\dots,\boldsymbol{x}_T,\boldsymbol{g}_1,\dots,\boldsymbol{g}_T$ and such that

$$\mathbf{P}(S=t) = \frac{\eta_t\left(1 - \bar{L}\eta_t\mathsf{V}_1\right)}{\sum_{t=1}^T \eta_t\left(1 - \bar{L}\eta_t\mathsf{V}_1\right)} .$$

Assume that η_t in (4) is chosen to satisfy $\bar{L}\eta_t m < 1$ and that $f^\star > -\infty$. Taking total expectation in (44) and summing up these inequalities for $t \le T$, combined with the fact that $f(\boldsymbol{x}_{T+1}) \ge f^\star$, we deduce that

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \le \frac{2(\mathbf{E}[f(\boldsymbol{x}_1)] - f^\star) + \sum_{t=1}^T \eta_t\left(b_t^2 + \bar{L}\eta_t v_t\right)}{\sum_{t=1}^T \eta_t\left(1 - \bar{L}\eta_t\mathsf{V}_1\right)} .$$

■

Proof of Theorem 13 The proof will be split into two parts: for gradient estimators (6) and (7), respectively. Both of these proofs follow from Lemma 12, which states that

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq \frac{2\delta_1 + \sum_{t=1}^T \eta_t \left(b_t^2 + \bar{L}\eta_t v_t\right)}{\sum_{t=1}^T \eta_t \left(1 - \bar{L}\eta_t \mathsf{V}_1\right)}\,, \tag{45}$$

where $\delta_1 = \mathbf{E}[f(\boldsymbol{x}_1)] - f^\star$. Using the corresponding bounds on the bias b_t and variance v_t, we substitute these values in the above inequality with η_t and h_t obtained by optimizing the obtained expressions.

We start with the part of the proof that is common for both gradient estimators. Introduce the notation

$$\Xi_T := d^{-\frac{2(\beta-1)}{2\beta-1}} T^{-\frac{\beta}{2\beta-1}}\,.$$

Using this notation, we consider algorithm (4) with gradient estimators (6) or (7) such that

$$\eta_t = \min\left(\frac{\mathfrak{y}}{d}, \Xi_T\right) \qquad \text{and} \qquad h_t = \mathfrak{h} T^{-\frac{1}{2(2\beta-1)}}\,,$$

where

$$(\mathfrak{y}, \mathfrak{h}) = \begin{cases} \left((8\kappa\bar{L})^{-1},\, d^{\frac{1}{2\beta-1}}\right) & \text{for estimator (6)} \\ \left((72\kappa\bar{L})^{-1},\, d^{\frac{2\beta+1}{4\beta-2}}\right) & \text{for estimator (7)} \end{cases}.$$

Given the values of V_1 in Table 1, the choice of η_t for both algorithms ensures that

$$\frac{1}{2} \leq 1 - \bar{L}\eta_t \mathsf{V}_1\,.$$

Thus we get from (45) that both algorithms satisfy

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq \left(\sum_{t=1}^T \eta_t\right)^{-1} \left(4\delta_1 + 2\sum_{t=1}^T \eta_t b_t^2 + 2\bar{L}\sum_{t=1}^T \eta_t^2 v_t\right)\,. \tag{46}$$

Furthermore, since $\eta_t = \min(\mathfrak{y}/d,\, \Xi_T)$, then in both cases we have

$$\left(\sum_{t=1}^T \eta_t\right)^{-1} = \max\left(\frac{d}{T\mathfrak{y}}, \frac{1}{T\Xi_T}\right) \leq \frac{d}{T\mathfrak{y}} + \frac{1}{T\Xi_T}\,.$$

Using this bound in (46) we deduce that

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq \left(\frac{d}{T\mathfrak{y}} + \frac{1}{T\Xi_T}\right)\left(4\delta_1 + 2\sum_{t=1}^{T}\eta_t b_t^2 + 2\bar{L}\sum_{t=1}^{T}\eta_t^2 v_t\right) .$$

Finally, by the definition of η_t we have $\eta_t \leq \Xi_T$ for all $t = 1, \dots, T$, which yields

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq \left(\frac{d}{\mathfrak{y}} + \frac{1}{\Xi_T}\right)\frac{4\delta_1}{T} + 2\left(\frac{d\Xi_T}{T\mathfrak{y}} + \frac{1}{T}\right)\sum_{t=1}^{T}\left\{b_t^2 + \bar{L}\Xi_T v_t\right\} . \tag{47}$$

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 $\mathfrak{y}$ and $\mathfrak{h}$ in order to get the final results.

D.2 Bounds for the gradient estimator (6) - ℓ_2 randomization

Lemma 6 for the bias and Lemma 7 for the variance imply that

$$b_t^2 \leq \left(\frac{\kappa_\beta L}{(\ell-1)!}\right)^2 h_t^{2(\beta-1)} \quad \text{and} \quad v_t = 4d\kappa\bar{L}^2 h_t^2 + \frac{d^2\sigma^2\kappa}{2h_t^2}\,, \quad \text{and} \quad \mathsf{V}_1 = 4d\kappa\,.$$

Using these bounds in (47) we get

$$\begin{aligned}
\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] &\leq \left(\frac{d}{\mathfrak{y}} + \Xi_T^{-1}\right)\frac{4\delta_1}{T} \\
&\quad + \left(\frac{d\Xi_T}{T\mathfrak{y}} + \frac{1}{T}\right)\sum_{t=1}^{T}\left\{\mathtt{A}_3 h_t^{2(\beta-1)} + \Xi_T d^2\left(\mathtt{A}_4 d^{-1}h_t^2 + \mathtt{A}_5 h_t^{-2}\right)\right\} \\
&\leq \left(\frac{d}{\mathfrak{y}} + \Xi_T^{-1}\right)\frac{4\delta_1}{T} + \frac{d\Xi_T+1}{T}\sum_{t=1}^{T}\left\{\mathtt{A}_6 h_t^{2(\beta-1)} + \mathtt{A}_7 d^2\Xi_T\left(d^{-1}h_t^2 + h_t^{-2}\right)\right\}
\end{aligned} \tag{48}$$

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 $t = 1, \dots, T$, inequality (48) has the form

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq \left(\frac{d}{\mathfrak{y}} + \Xi_T^{-1}\right)\frac{4\delta_1}{T} + (d\Xi_T+1)\Big(\mathtt{A}_6 h_T^{2(\beta-1)} + \mathtt{A}_7 d^2\Xi_T\big(d^{-1}h_T^2 + h_T^{-2}\big)\Big). \tag{49}$$

After substituting the expressions for Ξ_T and h_T into the above bound, the right hand side of (49) 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 $\big(d/T^{\beta}\big)^{\frac{1}{2\beta-1}} \leq 1$. Therefore, the final bound takes the form

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq \frac{4d}{T\mathfrak{y}}\delta_1 + \Big(4\delta_1 + 2\big(\mathtt{A}_6 + 2\mathtt{A}_7\big)\Big)\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}} \leq \Big(\mathtt{A}_1\delta_1 + \mathtt{A}_2\Big)\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}},$$

where $\mathtt{A}_1 = 4(\mathfrak{y}^{-1}+1)$ and $\mathtt{A}_2 = 2\big(\mathtt{A}_6 + 2\mathtt{A}_7\big)$.

D.3 Bounds for the gradient estimator (7) - ℓ_1 randomization

Lemma 8 for the the bias and the bound (15) for the variance imply that

$$b_t^2 \leq (c_\beta \kappa_\beta \ell L)^2 h_t^{2(\beta-1)} d^{1-\beta}, \qquad v_t = 72\kappa \bar{L}^2 h_t^2 + \frac{d^3\sigma^2\kappa}{h_t^2}, \qquad \text{and} \qquad \mathsf{V}_1 = 36 d\kappa\,,$$

with $\ell = \lfloor \beta \rfloor$. Using these bounds in (47) we get

$$\begin{aligned}\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq (d\Xi_T{+}1)\left(\mathtt{A}_6 d^{1-\beta} h_T^{2(\beta-1)}{+}\Xi_T\left(\mathtt{A}_7 h_T^2{+}\mathtt{A}_8 d^3 h_T^{-2}\right)\right)& \\ + \left(\frac{d}{\mathfrak{y}}{+}\Xi_T^{-1}\right)\frac{4\delta_1}{T}\,,&\end{aligned} \tag{50}$$

where the constants are defined as

$$\mathtt{A}_6 = 2(c_\beta\kappa_\beta\ell L)^2\big(\mathfrak{y}^{-1}+1\big), \quad \mathtt{A}_7 = 144\kappa\bar{L}^3\big(\mathfrak{y}^{-1}+1\big), \quad \mathtt{A}_8 = 2\bar{L}\sigma^2\kappa\big(\mathfrak{y}^{-1}+1\big)\,.$$

Substituting the expressions for Ξ_T and h_T in (50), we deduce that

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq \frac{4d}{T\mathfrak{y}}\delta_1 + \left\{4\delta_1 + \left(\left(\frac{d}{T^\beta}\right)^{\frac{1}{2\beta-1}} + 1\right)\left(\mathtt{A}_6 + \mathtt{A}_8 + \mathtt{A}_7\left(\frac{d^{5-2\beta}}{T^2}\right)^{\frac{1}{2\beta-1}}\right)\right\}\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}}.$$

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 less than or equal to one. Thus, we have

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq (\mathtt{A}_1\delta_1 + \mathtt{A}_2)\left(\frac{d^2}{T}\right)^{\frac{\beta-1}{2\beta-1}},$$

where $\mathtt{A}_1 = 4(\mathfrak{y}^{-1}+1)$, and $\mathtt{A}_2 = 2\left(\mathtt{A}_6 + \mathtt{A}_7 + \mathtt{A}_8\right)$.

$\blacksquare$

Proof of Theorem 14 As in the proof of Theorem 13 we use Lemma 12, cf. (45):

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq \frac{2\delta_1 + \sum_{t=1}^T \eta_t(b_t^2 + \bar{L}\eta_t v_t)}{\sum_{t=1}^T \eta_t(1 - \bar{L}\eta_t \mathsf{V}_1)}.$$

From this inequality and the fact that, by assumption, $\eta_t = (2\bar{L}\mathsf{V}_1)^{-1}$ we obtain:

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq \frac{8\bar{L}\mathsf{V}_1\delta_1 + 2\sum_{t=1}^{T}(b_t^2 + (2\mathsf{V}_1)^{-1}v_t)}{T}.$$

Since the gradient estimators (6) and (7) satisfy Assumption D and we consider the case $\sigma = 0$, $\beta = 2$, the values $b_t = bLh_t$ and $v_t = \mathsf{V}_2\bar{L}^2h_t^2$ can be made as small as possible by choosing h_t small enough. Thus, we can take h_t sufficiently small to have $\sum_{t=1}^{T}(b_t^2 + (2\mathsf{V}_1)^{-1}v_t) \leq \bar{L}\mathsf{V}_1$. Under this choice of h_t,

$$\mathbf{E}\left[\|\nabla f(\boldsymbol{x}_S)\|^2\right] \leq (8\delta_1 + 2)\frac{\bar{L}\mathsf{V}_1}{T}\,.$$

Using the values of V_1 for the gradient estimators (6) and (7) (see Table 1) we obtain the result. ■

D.4 Upper bounds: Smoothness and α-gradient dominance

Proof of Theorem 15 For brevity, we write $\mathbf{E}_t[\cdot]$ in place of $\mathbf{E}[\cdot \mid \boldsymbol{x}_t]$. Using Lipschitz continuity of ∇f (see e.g. Bubeck, 2015, Lemma 3.4) and the definition of the algorithm in (4) with $\Theta = \mathbb{R}^d$ we have

$$\begin{aligned}\mathbf{E}_t[f(\boldsymbol{x}_{t+1})] &\leq f(\boldsymbol{x}_t) - \eta_t\left\langle \nabla f(\boldsymbol{x}_t)\,, \mathbf{E}_t[\boldsymbol{g}_t]\right\rangle + \frac{\bar{L}\eta_t^2}{2}\mathbf{E}_t\left[\|\boldsymbol{g}_t\|^2\right]\\ &\leq f(\boldsymbol{x}_t) - \eta_t\|\nabla f(\boldsymbol{x}_t)\|^2 + \eta_t\|\nabla f(\boldsymbol{x}_t)\|\,\|\mathbf{E}_t[\boldsymbol{g}_t] - \nabla f(\boldsymbol{x}_t)\| + \frac{\bar{L}\eta_t^2}{2}\mathbf{E}_t\left[\|\boldsymbol{g}_t\|^2\right].\end{aligned}$$

Next, invoking Assumption D on the bias and variance of $\boldsymbol{g}_t$ and using the elementary inequality $2ab \leq a^2 + b^2$ we get that, for the iterative procedure (4) with $\Theta = \mathbb{R}^d$,

$$\delta_{t+1} \leq \delta_t - \frac{\eta_t}{2}(1 - \bar{L}\eta_t\mathsf{V}_1)\mathbf{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \frac{\eta_t}{2}\left(b^2L^2h_t^{2(\beta-1)} + \bar{L}\eta_t\left(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right)\right),$$

where $\delta_t = \mathbf{E}[f(\boldsymbol{x}_t) - f^\star]$. Furthermore, our choice of the step size η_t ensures that $1 - \bar{L}\eta_t\mathsf{V}_1 \geq \frac{1}{2}$. Using this inequality and the fact that f is α-gradient dominant we deduce that

$$\delta_{t+1} \leq \delta_t\left(1 - \frac{\eta_t\alpha}{2}\right) + \frac{\eta_t}{2}\left(b^2L^2h_t^{2(\beta-1)} + \bar{L}\eta_t\left(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right)\right). \tag{51}$$

We now analyze this recursion according to the cases $T > T_0$ and $T \leq T_0$, where $T_0 := \left\lfloor \frac{8\bar{L}\mathsf{V}_1}{\alpha} \right\rfloor$ is the value of t, where η_t switches its regime.

First case: $T > T_0$**.** In this case, the recursion (51) has two different regimes, depending on the value of η_t. In the first regime, for any $t = T_0+1, \dots, T$, we have $\eta_t = \frac{4}{\alpha t}$ and (51) takes the form

$$\delta_{t+1} \le \delta_t\left(1-\frac{2}{t}\right) + 2b^2L^2\cdot\frac{h_t^{2(\beta-1)}}{\alpha t} + \frac{8\bar{L}}{\alpha^2 t^2}\Big(\mathsf{V}_2\bar{L}^2 h_t^2 + \mathsf{V}_3\sigma^2 h_t^{-2}\Big). \tag{52}$$

Additionally in this regime of t, we have $h_t = \big(\frac{4\bar{L}\sigma^2\mathsf{V}_3}{b^2L^2\alpha t}\big)^{\frac{1}{2\beta}}$. Using this expression for h_t in (52) we obtain that

$$\begin{aligned}\delta_{t+1} \le \delta_t\left(1-\frac{2}{t}\right) &+ \mathtt{A}_3\cdot\frac{1}{\alpha}\left(\frac{\bar{L}\sigma^2}{\alpha}\right)^{1-\frac{1}{\beta}}\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}} t^{-\frac{2\beta-1}{\beta}}\\ &+ \mathtt{A}_4\cdot\frac{1}{\alpha}\left(\frac{\bar{L}}{\alpha}\right)^{1+\frac{1}{\beta}}\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}} t^{-\frac{2\beta+1}{\beta}},\end{aligned} \tag{53}$$

where $\mathtt{A}_3 = 2^{4-\frac{2}{\beta}}$, and $\mathtt{A}_4 = 2^{3+\frac{2}{\beta}}$. Applying Lemma 32 to the above recursion we get

$$\begin{aligned}\delta_T \le \frac{2T_0}{T}\delta_{T_0+1} &+ \frac{\beta\mathtt{A}_3}{(\beta{+}1)\alpha}\cdot\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}}\left(\frac{\alpha T}{\bar{L}\sigma^2}\right)^{-\frac{\beta-1}{\beta}}\\ &+ \frac{\beta\mathtt{A}_4}{(3\beta{+}1)\alpha}\cdot\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}\left(\frac{\alpha T}{\bar{L}}\right)^{-\frac{\beta+1}{\beta}}.\end{aligned} \tag{54}$$

If $T_0 = 0$, we conclude the proof for the case $T > T_0$. Otherwise, we consider the second regime that corresponds to $t \in [1, T_0]$. In this regime, we have $h_t = \big(\frac{4\bar{L}\sigma^2\mathsf{V}_3}{b^2L^2\alpha T}\big)^{\frac{1}{2\beta}}$, $\eta_t = \frac{1}{2\bar{L}\mathsf{V}_1}$, and $\frac{4}{(T_0+1)\alpha} \le \eta_t \le \frac{4}{T_0\alpha}$. Using these expressions for h_t and η_t in (51) we get that, for $1 \le t \le T_0$,

$$\begin{aligned}\delta_{t+1} \le \delta_t\left(1-\frac{2}{T_0+1}\right) &+ \frac{2^{4-\frac{2}{\beta}}}{T_0}\cdot\frac{1}{\alpha}\left(\frac{\bar{L}\sigma^2}{\alpha}\right)^{1-\frac{1}{\beta}}\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}}\left(T^{-\frac{\beta-1}{\beta}} + \frac{T^{\frac{1}{\beta}}}{T_0}\right)\\ &+ \frac{2^{3+\frac{2}{\beta}}}{T_0^2}\cdot\frac{1}{\alpha}\left(\frac{\bar{L}}{\alpha}\right)^{1+\frac{1}{\beta}}\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}T^{-\frac{1}{\beta}}.\end{aligned}$$

Using the rough bound $1-\frac{2}{T_0+1} \le 1$ and unfolding the above recursion we obtain:

$$\begin{aligned}\delta_{T_0+1} \le \delta_1 &+ 2^{4-\frac{2}{\beta}}\cdot\frac{1}{\alpha}\left(\frac{\bar{L}\sigma^2}{\alpha}\right)^{1-\frac{1}{\beta}}\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}}\left(T^{-\frac{\beta-1}{\beta}} + \frac{T^{\frac{1}{\beta}}}{T_0}\right)\\ &+ \frac{2^{3+\frac{2}{\beta}}}{T_0}\cdot\frac{1}{\alpha}\left(\frac{\bar{L}}{\alpha}\right)^{1+\frac{1}{\beta}}\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}T^{-\frac{1}{\beta}}.\end{aligned}$$

Taking into account the definition of T_0, and the fact that $T_0 \le T$ we further obtain:

$$\begin{aligned}\frac{2T_0}{T}\delta_{T_0+1} \le \frac{16\bar{L}\mathsf{V}_1}{\alpha T}\delta_1 &+ \frac{2\mathtt{A}_3}{\alpha}\cdot \mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}}\left(\frac{\alpha T}{\bar{L}\sigma^2}\right)^{-\frac{\beta-1}{\beta}}\\ &+ \frac{2\mathtt{A}_4}{\alpha}\cdot\left(\frac{\bar{L}}{\alpha}\right)^{1+\frac{1}{\beta}}\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}\left(\frac{\alpha T}{\bar{L}}\right)^{-\frac{\beta+1}{\beta}}\quad .\end{aligned}\tag{55}$$

Finally, combining (54) and (55) yields:

$$\delta_T \le \mathtt{A}_1\cdot\frac{\bar{L}\mathsf{V}_1}{\alpha T}\delta_1 + \frac{\mathtt{A}_2}{\alpha}\cdot\left(\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}} + \mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{\frac{1}{\beta}}\left(\frac{\alpha T}{\bar{L}\sigma^2}\right)^{-\frac{2}{\beta}}\sigma^{-2}\right)\left(\frac{\alpha T}{\bar{L}\sigma^2}\right)^{-\frac{\beta-1}{\beta}}\quad ,$$

where $\mathtt{A}_1 = 16$ and $\mathtt{A}_2 = \left(2 + \frac{\beta}{\beta+1}\right)\mathtt{A}_3 + \left(2 + \frac{\beta}{3\beta+1}\right)\mathtt{A}_4$.

Second case: $T \le T_0$. In this case, we have $h_t = \left(\frac{4\bar{L}\sigma^2\mathsf{V}_3}{b^2L^2\alpha T}\right)^{\frac{1}{2\beta}}$ and thus (51) takes the form

$$\begin{aligned}\delta_{T+1} &\le \delta_1\left(1 - \frac{2}{T_0+1}\right)^T + \frac{2^{4-\frac{2}{\beta}}}{T_0}\cdot\frac{1}{\alpha}\left(\frac{\bar{L}\sigma^2}{\alpha}\right)^{1-\frac{1}{\beta}}\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}}\sum_{t=1}^{T}\left(T^{-\frac{\beta-1}{\beta}} + \frac{T^{\frac{1}{\beta}}}{T_0}\right)\\ &\quad + \frac{2^{3+\frac{2}{\beta}}}{T_0^2}\cdot\frac{1}{\alpha}\left(\frac{\bar{L}}{\alpha}\right)^{1+\frac{1}{\beta}}\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}\sum_{t=1}^{T}T^{-\frac{1}{\beta}}\\ &\le \delta_1\left(1 - \frac{2}{T_0+1}\right)^T + \frac{\mathtt{A}_3}{\alpha}\cdot\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}}\left(\frac{\alpha T}{\bar{L}\sigma^2}\right)^{-\frac{\beta-1}{\beta}} + \frac{\mathtt{A}_4}{\alpha}\cdot\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}\left(\frac{\alpha T}{\bar{L}}\right)^{-\frac{\beta+1}{\beta}}\quad .\end{aligned}$$

Note that, for any $\rho, T > 0$, we have $(1-\rho)^T \le \exp(-\rho T) \le \frac{1}{\rho T}$. Using this inequality for $\rho = \frac{2}{T_0+1}$, the definition of T_0 and the fact that $T + 1 \le 2T$ we obtain:

$$\begin{aligned}\delta_{T+1} &\le \mathtt{A}_1\frac{\bar{L}\mathsf{V}_1}{\alpha(T+1)}\delta_1\\ &\quad + \frac{\mathtt{A}_2}{\alpha}\left(\mathsf{V}_3\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{-\frac{1}{\beta}} + \mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3}{b^2L^2}\right)^{\frac{1}{\beta}}\left(\frac{\alpha(T+1)}{\bar{L}\sigma^2}\right)^{-\frac{2}{\beta}}\sigma^{-2}\right)\left(\frac{\alpha(T+1)}{\bar{L}\sigma^2}\right)^{-\frac{\beta-1}{\beta}}\quad .\end{aligned}$$

■

Proof of Theorem 17 As in the proof of Theorem 15, we consider separately the cases $T > T_0$ and $T \le T_0$, where $T_0 := \left\lfloor\frac{8\bar{L}\mathsf{V}_1}{\alpha}\right\rfloor$.

The case $T > T_0$. First, consider the algorithm at steps $t = T_0, \dots, T$, where we have $\eta_t = \frac{4}{\alpha t}$. Since $\sigma = 0$ and $\beta = 2$, from (51) we have

$$\delta_{t+1} \leq \delta_t \left(1 - \frac{2}{t}\right) + \left(\frac{2b^2 L^2}{\alpha t} + \frac{8\bar{L}^3}{\alpha^2 t^2}\mathsf{V}_2\right) h_t^2 \,. \tag{56}$$

Since $\beta = 2$, we have $L = \bar{L}$. Thus, using the assumption that $h_t \leq \left(\frac{\bar{L}\vee 1}{\alpha \wedge 1} T \left(2b^2 \bar{L} + \frac{8\bar{L}^2 \mathsf{V}_2}{\alpha}\right)\right)^{-\frac{1}{2}}$ we deduce from (56) that

$$\delta_{t+1} \leq \delta_t \left(1 - \frac{2}{t}\right) + \frac{\bar{L}}{\alpha t^2} \,. \tag{57}$$

Applying Lemma 32 to the above recursion gives

$$\delta_T \leq \frac{2T_0}{T}\delta_{T_0+1} + \frac{\bar{L}}{\alpha T} \,. \tag{58}$$

If $T_0 = 0$, we conclude the proof for the case $T > T_0$. Otherwise, we consider the algorithm at steps $t = 1, \dots, T_0$, where $\eta_t = \frac{1}{2\bar{L}\mathsf{V}_1}$ and $\frac{4}{(T_0+1)\alpha} \leq \eta_t \leq \frac{4}{T_0 \alpha}$. From (51) with $\sigma = 0$ and $\beta = 2$ we obtain

$$\delta_{t+1} \leq \delta_t \left(1 - \frac{2}{T_0+1}\right) + \frac{\bar{L}}{\alpha T_0}\left(2b^2 \bar{L} + \frac{8}{\alpha}\bar{L}^2 \mathsf{V}_2\right) h_t^2 \,. \tag{59}$$

Using here the assumption that $h_t \leq \left(\frac{\bar{L}\vee 1}{\alpha \wedge 1} T \left(2b^2 \bar{L} + \frac{8\bar{L}^2 \mathsf{V}_2}{\alpha}\right)\right)^{-\frac{1}{2}}$ and a rough bound $1 - \frac{2}{T_0+1} \leq 1$ and summing up both sides of the resulting inequality from $t = 1$ to $t = T_0$ we get

$$\delta_{T_0+1} \leq \delta_1 + \frac{\bar{L}(\alpha \wedge 1)}{\alpha(\bar{L} \vee 1)T} \,.$$

Combining this inequality with (58) and using the definition of T_0 and the fact that $T_0 \leq T$ we obtain the bound

$$\delta_T \leq \frac{16\bar{L}\mathsf{V}_1}{\alpha T}\delta_1 + \frac{16\bar{L}\mathsf{V}_1}{\alpha T^2} + \frac{\bar{L}}{\alpha T} \,.$$

It remains to note that $\mathsf{V}_1 \leq 36 d\kappa$, cf. Table 1. This implies the theorem for the case $T < T_0$ with $\mathtt{A}_1 = 576\kappa$ and $\mathtt{A}_2 = 1/T + 1/\mathtt{A}_1 d$.

Second case: $T \leq T_0$**.** Using the fact that $h_t \leq \left(\frac{T}{\alpha \wedge 1}\left(2b^2 \bar{L} + \frac{8\bar{L}^2 \mathsf{V}_2}{\alpha}\right)\right)^{-\frac{1}{2}}$ and unfolding the recursion in (59) gives

$$\delta_T \leq \delta_1 \left(1 - \frac{2}{T_0+1}\right)^T + \frac{\bar{L}}{\alpha T} \,.$$

From the elementary inequality $(1-\rho)^T \leq \frac{1}{\rho T}$, which is valid for all $\rho, T > 0$, we obtain with $\rho = \frac{2}{T_0+1}$ that

$$\delta_T \leq \frac{T_0+1}{2T}\delta_1 + \frac{\bar{L}}{\alpha T} \leq \frac{T_0}{T}\delta_1 + \frac{\bar{L}}{\alpha T} \leq \mathtt{A}_1 \frac{\bar{L}d}{\alpha T}\left(\delta_1 + \mathtt{A}_2\right) \,,$$

where $\mathtt{A}_1 = 288\kappa$, $\mathtt{A}_2 = 1/\mathtt{A}_1 d$, and the last inequality follows from the facts that $T_0 \leq \frac{8\bar{L}\mathsf{V}_1}{\alpha}$ and $\mathsf{V}_1 \leq 36d\kappa$, cf. Table 1. ■

D.5 Smoothness and α-strong convexity: unconstrained minimization

We will use the following basic lemma.

Lemma 33 *Consider the iterative algorithm defined in (4). Let f be α-strongly convex on $\mathbb{R}^d$, let Assumption D be satisfied. Let the minimizer $\boldsymbol{x}^\star$ of f on Θ be such that $\nabla f(\boldsymbol{x}^\star) = 0$. Then we have*

$$\mathbf{E}[f(\boldsymbol{x}_t) - f^\star] \leq \frac{r_t - r_{t+1}}{2\eta_t} - r_t\left(\frac{\alpha}{4} - \frac{\eta_t}{2}\bar{L}^2\mathsf{V}_1\right) + \frac{(bLh_t^{\beta-1})^2}{\alpha} + \frac{\eta_t}{2}\left(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right) \,, \tag{60}$$

where $r_t = \mathbf{E}[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2]$.

Proof Recall the notation $\mathbf{E}_t[\cdot] = \mathbf{E}[\cdot \mid \boldsymbol{x}_t]$. For any $\boldsymbol{x} \in \Theta$, by the definition of projection,

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2 = \left\|\mathrm{Proj}_\Theta\big(\boldsymbol{x}_t - \eta_t\boldsymbol{g}_t\big) - \boldsymbol{x}\right\|^2 \leq \|\boldsymbol{x}_t - \eta_t\boldsymbol{g}_t - \boldsymbol{x}\|^2 \,. \tag{61}$$

Expanding the squares and rearranging the above inequality, we deduce that (61) is equivalent to

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leq \frac{\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2}{2\eta_t} + \frac{\eta_t}{2}\|\boldsymbol{g}_t\|^2 \,. \tag{62}$$

On the other hand, since f is a α-strongly convex function on Θ, we have

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}) \leq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle - \frac{\alpha}{2}\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 \,. \tag{63}$$

Combining (62) with (63) and introducing the notation $a_t = \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2$ we deduce that

$$\mathbf{E}_t[f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)] \leq \|\mathbf{E}_t[\boldsymbol{g}_t] - \nabla f(\boldsymbol{x}_t)\|\,\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \frac{1}{2\eta_t}\mathbf{E}_t[a_t - a_{t+1}] + \frac{\eta_t}{2}\mathbf{E}_t\|\boldsymbol{g}_t\|^2 - \frac{\alpha}{2}\mathbf{E}_t[a_t] \tag{64}$$

$$\begin{aligned} &\leq bLh_t^{\beta-1}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \frac{1}{2\eta_t}\mathbf{E}_t[a_t - a_{t+1}] \\ &\quad + \frac{\eta_t}{2}\Big(\mathsf{V}_1\mathbf{E}[\|\nabla f(\boldsymbol{x}_t)\|^2] + \mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\Big) - \frac{\alpha}{2}\mathbf{E}_t[a_t] \,. \end{aligned} \tag{65}$$

Using the elementary inequality $2ab \le a^2 + b^2$ we have

$$bLh_t^{\beta-1}\left\|\boldsymbol{x}_t - \boldsymbol{x}^\star\right\| \le \frac{(bLh_t^{\beta-1})^2}{\alpha} + \frac{\alpha}{4}\left\|\boldsymbol{x}_t - \boldsymbol{x}^\star\right\|^2 . \tag{66}$$

Substituting (66) in (65), setting $r_t = \mathbf{E}[a_t]$, using the fact that

$$\|\nabla f(\boldsymbol{x}_t)\|^2 \le \bar{L}^2\left\|\boldsymbol{x}_t - \boldsymbol{x}^\star\right\|^2 = \bar{L}^2 a_t^2$$

and taking the total expectation from both sides of the resulting inequality yields the lemma. ■

Proof of Theorem 18 By definition, $\eta_t \le \frac{\alpha}{4\bar{L}^2\mathsf{V}_1}$, so that $\frac{\alpha}{4} - \frac{\eta_t}{2}\bar{L}^2\mathsf{V}_1 \ge \frac{\alpha}{8}$ and (60) implies that

$$\mathbf{E}[f(\boldsymbol{x}_t) - f^\star] \le \frac{r_t - r_{t+1}}{2\eta_t} - \frac{\alpha}{8}r_t + \frac{(bLh_t^{\beta-1})^2}{\alpha} + \frac{\eta_t}{2}(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2})\,. \tag{67}$$

Set $T_0 := \max\left(\left\lfloor \frac{32\bar{L}^2\mathsf{V}_1}{\alpha^2} - 1\right\rfloor, 0\right)$. This is the value of t, where η_t switches its regime. We analyze the recursion (67) separately for the cases $T > T_0$ and $T \le T_0$. We will use the fact that, by the convexity of f and Jensen's inequality,

$$f(\hat{\boldsymbol{x}}_T) - f^\star \le \frac{2}{T(T+1)}\sum_{t=1}^{T} t(f(\boldsymbol{x}_t) - f^\star). \tag{68}$$

First case: $T > T_0$. In this case, we decompose the sum in (68) into the sum over $t \in [T_0+1, T]$ and the sum over $t \in [1, T_0]$.

We first evaluate the sum $\sum_{t=T_0+1}^{T} t\mathbf{E}[f(\boldsymbol{x}_t) - f^\star]$. For any $t \in [T_0+1, T]$, we have $\eta_t = \frac{8}{\alpha(t+1)}$ and $h_t = \left(\frac{4\sigma^2\mathsf{V}_3}{b^2L^2t}\right)^{\frac{1}{2\beta}}$. Using in (67) these values of η_t and h_t, multiplying by t, and summing both sides of the resulting inequality from T_0+1 to T we deduce that

$$\sum_{t=T_0+1}^{T} t\mathbf{E}[f(\boldsymbol{x}_t) - f^\star] \le \underbrace{\frac{\alpha}{16}\sum_{t=T_0+1}^{T} t\left((r_t - r_{t+1})(t+1) - 2r_t\right)}_{=:\mathsf{I}} + \underbrace{\frac{\mathbb{A}_4}{\alpha}(bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}}\sum_{t=T_0+1}^{T} t^{\frac{1}{\beta}}}_{=:\mathsf{II}}$$
$$+ \underbrace{\frac{\mathbb{A}_5}{\alpha}\bar{L}^2\mathsf{V}_2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}\sum_{t=T_0+1}^{T} t^{-\frac{1}{\beta}}}_{=:\mathsf{III}},$$

where $\mathbb{A}_4 = 2^{\frac{3\beta-2}{\beta}}$, $\mathbb{A}_5 = 2^{\frac{2\beta+2}{\beta}}$ and we defined the terms I, II, and III that will be evaluated separately.

It is not hard to check that $\mathsf{I} \leq \frac{\alpha}{16} T T_0 r_{T_0+1}$ since the summation in term I is telescoping. Next, we have

$$\mathsf{II} \leq \frac{\mathtt{A}_4}{\alpha}(bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}}\sum_{t=1}^{T} t^{\frac{1}{\beta}} \leq \frac{\mathtt{A}_6}{\alpha}(bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}} T^{\frac{1}{\beta}+1}.$$

Finally,

$$\mathsf{III} \leq \frac{\mathtt{A}_5}{\alpha}\bar{L}^2\mathsf{V}_2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}\sum_{t=1}^{T} t^{-\frac{1}{\beta}} \leq \frac{\mathtt{A}_7}{\alpha}\bar{L}^2\mathsf{V}_2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}} T^{1-\frac{1}{\beta}}\,,$$

where $\mathtt{A}_6 = \frac{\beta+1}{\beta}\mathtt{A}_4$ and $\mathtt{A}_7 = \frac{\beta-1}{\beta}\mathtt{A}_5$. Combining these bounds on I, II, and III we obtain

$$\begin{aligned}\sum_{t=T_0+1}^{T} t\mathbf{E}[f(\boldsymbol{x}_t) - f^\star] &\leq \frac{\alpha}{16} T T_0 r_{T_0+1} \\ &\quad + \left(\mathtt{A}_6(bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}} + \mathtt{A}_7\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}}\right)\frac{T^{\frac{1}{\beta}+1}}{\alpha}\,.\end{aligned} \tag{69}$$

If $T_0 = 0$ then combining (68) and (69) proves the theorem. If $T_0 \geq 1$ we need additionally to control the value r_{T_0+1} on the right hand side of (69). It follows from (67) that, for $1 \leq t \leq T_0$,

$$r_{t+1} \leq r_t + 2\eta_t\frac{(bLh_t^{\beta-1})^2}{\alpha} + \eta_t^2\left(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right).$$

Moreover, for $1 \leq t \leq T_0$ we have $\eta_t = \frac{\alpha}{4\bar{L}^2\mathsf{V}_1}$ and $\eta_t \leq \frac{8}{\alpha(T_0+1)}$. Therefore, unfolding the above recursion we get

$$r_{T_0+1} \leq r_1 + \sum_{t=1}^{T_0}\left(\frac{16}{\alpha^2T_0}\left(bLh_t^{\beta-1}\right)^2 + \frac{64}{\alpha^2T_0^2}\left(\mathsf{V}_2\bar{L}^2h_t^2 + \mathsf{V}_3\sigma^2h_t^{-2}\right)\right).$$

For $1 \leq t \leq T_0$ we have $h_t = \left(\frac{4\sigma^2\mathsf{V}_3}{b^2L^2T}\right)^{\frac{1}{2\beta}}$, which yields

$$r_{T_0+1} \leq r_1 + 16\left(\mathtt{A}_4(bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}} + \mathtt{A}_5\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}}\right)\frac{T^{\frac{1}{\beta}}}{\alpha^2T_0}\,,$$

so that

$$\frac{\alpha}{16} T T_0 r_{T_0+1} \leq \frac{2\bar{L}^2T\mathsf{V}_1}{\alpha}r_1 + \left(\mathtt{A}_4(bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}} + \mathtt{A}_5\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}}\right)\frac{T^{\frac{1}{\beta}+1}}{\alpha}\,. \tag{70}$$

It follows from (69) and (70) that

$$\begin{aligned}\sum_{t=T_0+1}^{T} t\mathbf{E}[f(\boldsymbol{x}_t)-f^\star] \le \frac{2\bar{L}^2 T\mathsf{V}_1}{\alpha} r_1 + \Bigg\{&(\mathtt{A}_4+\mathtt{A}_6)(bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}} \\ &+ (\mathtt{A}_5+\mathtt{A}_7)\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}}\Bigg\}\frac{T^{\frac{1}{\beta}+1}}{\alpha}\,,\end{aligned} \tag{71}$$

We now evaluate the sum $\sum_{t=1}^{T_0} t\mathbf{E}[f(\boldsymbol{x}_t)-f^\star]$. Recall that for $t\in[1,T_0]$ the parameters h_t,η_t take constant values: $h_t=\left(\frac{4\sigma^2\mathsf{V}_3}{b^2L^2T}\right)^{\frac{1}{2\beta}}$ and $\eta_t=\frac{\alpha}{4\bar{L}^2\mathsf{V}_1}$. Omitting in (67) the term $-\alpha r_t/8$, summing the resulting recursion from 1 to T_0 and using the inequality $\eta_t\le\frac{8}{\alpha(T_0+1)}$ we obtain

$$\begin{aligned}\sum_{t=1}^{T_0} t\mathbf{E}[f(\boldsymbol{x}_t)-f^\star] &\le T\sum_{t=1}^{T_0}\mathbf{E}[f(\boldsymbol{x}_t)-f^\star] \\ &\le \frac{Tr_1}{2\eta_1} + T\sum_{t=1}^{T_0}\left(\frac{(bLh_t^{\beta-1})^2}{\alpha}+\frac{\eta_t}{2}(\mathsf{V}_2\bar{L}^2h_t^2+\mathsf{V}_3\sigma^2h_t^{-2})\right) \\ &\le \frac{Tr_1}{2\eta_1} + T^2\frac{(bLh_1^{\beta-1})^2}{\alpha}+\frac{2T}{\alpha}(\mathsf{V}_2\bar{L}^2h_1^2+\mathsf{V}_3\sigma^2h_1^{-2}) \\ &= \frac{2\bar{L}^2T\mathsf{V}_1}{\alpha}r_1 + \left(\mathtt{A}_4(bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}} + \mathtt{A}_5\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}T^{-\frac{2}{\beta}}\right)\frac{T^{\frac{1}{\beta}+1}}{\alpha}\,.\end{aligned} \tag{72}$$

Summing up (71) and (72) and using (68) we obtain the bound of the theorem:

$$\mathbf{E}[f(\hat{\boldsymbol{x}}_T)-f^\star] \le \mathtt{A}_1\frac{\bar{L}^2\mathsf{V}_1}{\alpha T}r_1 + \left(\mathtt{A}_2(bL)^{\frac{2}{\beta}}(\mathsf{V}_3\sigma^2)^{\frac{\beta-1}{\beta}} + \mathtt{A}_3\mathsf{V}_2\bar{L}^2\left(\frac{\mathsf{V}_3\sigma^2}{b^2L^2}\right)^{\frac{1}{\beta}}T^{-\frac{2}{\beta}}\right)\frac{T^{-\frac{\beta-1}{\beta}}}{\alpha}\,,$$

where $\mathtt{A}_1=8$, $\mathtt{A}_2=4\mathtt{A}_4+2\mathtt{A}_6$, and $\mathtt{A}_3=4\mathtt{A}_5+2\mathtt{A}_7$.

Second case: $T\le T_0$. In this scenario, the summation $\sum_{t=1}^{T} t\mathbf{E}[f(\boldsymbol{x}_t)-f^\star]$ is treated as in (72), with the only difference that T_0 is replaced by T. As a result, we obtain the same bound as in (72). ■

D.6 Smoothness and α-strong convexity: constrained minimization

Proof of Theorem 20 Since $\sup_{\boldsymbol{x}\in\Theta}\|\nabla f(\boldsymbol{x})\|\le G$ we get from (65) that, for any $t=1,\dots,T$,

$$\mathbf{E}[f(\boldsymbol{x}_t)-f^\star] \le \frac{r_t-r_{t+1}}{2\eta_t}-\frac{\alpha}{4}r_t+\frac{\left(bLh_t^{\beta-1}\right)^2}{\alpha}+\frac{\eta_t}{2}\left(\mathsf{V}_1G^2+\mathsf{V}_2\bar{L}^2h_t^2+\mathsf{V}_3\sigma^2h_t^{-2}\right). \tag{73}$$

Multiplying both sides of (73) by t, summing up from $t = 1$ to T and using the fact that

$$\sum_{t=1}^{T} \left(\frac{t(r_t - r_{t+1})}{2\eta_t} - \frac{\alpha}{4} t r_t \right) \leq 0 \qquad \text{if} \quad \eta_t = \frac{4}{\alpha(t+1)}$$

we find that

$$\sum_{t=1}^{T} t\mathbf{E}[f(\boldsymbol{x}_t) - f^\star] \leq \frac{1}{\alpha} \sum_{t=1}^{T} \left[t(bLh_t^{\beta-1})^2 + \frac{2t}{t+1} \big(\mathsf{V}_1 G^2 + \mathsf{V}_2 \bar{L}^2 h_t^2 + \mathsf{V}_3 \sigma^2 h_t^{-2}\big) \right].$$

Since $h_t = \left(\frac{\sigma^2 \mathsf{V}_3}{b^2 L^2 t}\right)^{\frac{1}{2\beta}}$ we obtain

$$\sum_{t=1}^{T} t\mathbf{E}[f(\boldsymbol{x}_t) - f^\star] \leq \frac{2\mathsf{V}_1 G^2 T}{\alpha} + \frac{\mathtt{A}_3}{\alpha} \left(\mathsf{V}_3 \sigma^2 \left(\frac{\mathsf{V}_3 \sigma^2}{b^2 L^2} \right)^{-\frac{1}{\beta}} + \mathsf{V}_2 \bar{L}^2 \left(\frac{\mathsf{V}_3 \sigma^2}{b^2 L^2} \right)^{\frac{1}{\beta}} T^{-\frac{2}{\beta}} \right) T^{1+\frac{1}{\beta}}, \tag{74}$$

where $\mathtt{A}_3 = 2$. To complete the proof, we multiply both sides of (74) by $\frac{2}{T(T+1)}$ and use (68). ■

Appendix E. Proof of the lower bounds

Set for brevity $A_t = \{(\boldsymbol{z}_i, y_i)_{i=1}^t, (\boldsymbol{z}'_i, y'_i)_{i=1}^t\}$ for $t \geq 1$. Without loss of generality, we assume that $\boldsymbol{z}_1$ and $\boldsymbol{z}'_1$ are fixed and we prove the result for any fixed $(\boldsymbol{z}_1, \boldsymbol{z}'_1)$. For any $f : \mathbb{R}^d \to \mathbb{R}$ and any sequential strategy in the class Π_T, such that $\boldsymbol{z}_t = \Phi_t(A_{t-1}, \boldsymbol{\tau}_t)$ with $y_t = f(\boldsymbol{z}_t) + \xi_t$ and $\boldsymbol{z}'_t = \Phi'_t(A_{t-1}, \boldsymbol{\tau}_t)$ for $t \geq 2$, with $y'_t = f(\boldsymbol{z}'_t) + \xi'_t$ for $t = 1, \dots, T$, we will denote by $\mathbf{P}_f$ the joint distribution of $(A_T, (\boldsymbol{\tau}_i)_{i=2}^T)$.

We start with the following lemma that will be used in the proof of Theorem 22.

Lemma 34 *Let Assumption E be satisfied. Then, for any functions $f, f' : \mathbb{R}^d \to \mathbb{R}$ such that $\|f - f'\|_\infty = \max_{\boldsymbol{u} \in \mathbb{R}^d} |f(\boldsymbol{u}) - f'(\boldsymbol{u})| \leq v_0$ it holds that*

$$\frac{1}{2} H^2(\mathbf{P}_f, \mathbf{P}_{f'}) \leq 1 - \left(1 - \frac{I_0}{2} \|f - f'\|_\infty^2 \right)^T.$$

Proof Since, for each $t \geq 2$,the noise (ξ_t, ξ'_t) is independent of $(A_{t-1}, (\boldsymbol{\tau}_i)_{i=1}^t)$ and $\boldsymbol{\tau}_t$ is independent of A_{t-1} we have

$$\mathrm{d}\mathbf{P}_f = \mathrm{d}F_1(y_1 - f(\boldsymbol{z}_1), y'_1 - f(\boldsymbol{z}'_1)) \prod_{t=2}^{T} \mathrm{d}F_t\Big(y_t - f\Big(\Phi_t(A_{t-1}, \boldsymbol{\tau}_t)\Big), y'_t - f\Big(\Phi'_t(A_{t-1}, \boldsymbol{\tau}_t)\Big)\Big) \mathrm{d}\mathbb{P}_t(\boldsymbol{\tau}_t)$$

where $\mathbb{P}_t$ is the probability measure corresponding to the distribution of $\boldsymbol{\tau}_t$. Set for brevity $\mathrm{d}F_{f,1} \triangleq \mathrm{d}F_1\big(y_1 - f(\boldsymbol{z}_1), y'_1 - f(\boldsymbol{z}'_1)\big)$ and $\mathrm{d}F_{f,t} \triangleq \mathrm{d}F_t\big(y_t - f(\boldsymbol{z}_t), y'_t - f(\boldsymbol{z}'_t)\big) \mathrm{d}\mathbb{P}_t(\boldsymbol{\tau}_t)$, $t \geq 2$.

With this notation, we have $\mathrm{d}\mathbf{P}_f = \prod_{t=1}^{T} \mathrm{d}F_{f,t}$. Using the definition of Hellinger distance we obtain

$$1-\frac{1}{2}H^2(\mathbf{P}_f, \mathbf{P}_{f'}) = \int \sqrt{\mathrm{d}\mathbf{P}_f \, \mathrm{d}\mathbf{P}_{f'}} = \prod_{t=1}^{T} \int \sqrt{\mathrm{d}F_{f,t}}\sqrt{\mathrm{d}F_{f',t}} = \prod_{t=1}^{T} \left(1 - \frac{H^2\left(\mathrm{d}F_{f,t}, \, \mathrm{d}F_{f',t}\right)}{2}\right).$$

Finally, invoking Assumption E, we get

$$\begin{aligned}\prod_{t=1}^{T} \left(1 - \frac{H^2\left(\mathrm{d}F_{f,t}, \, \mathrm{d}F_{f',t}\right)}{2}\right) &\geq \min_{1\leq t\leq T} \left(1 - \frac{H^2\left(\mathrm{d}F_{f,t}, \, \mathrm{d}F_{f',t}\right)}{2}\right)^T \\ &\geq \left(1 - \frac{I_0 \left\|f - f'\right\|_\infty^2}{2}\right)^T,\end{aligned}$$

which implies the lemma. ■

Proof of Theorem 22 The proof follows the general lines given in Akhavan et al. (2020), so that we omit some details that can be found in that paper. 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.

It is shown in Akhavan et al. (2020) that if $\alpha \geq T^{-1/2+1/\beta}$ then $f_{\boldsymbol{\omega}} \in \mathcal{F}'_{\alpha,\beta}$ for $r > 0$ and $\delta > 0$ small enough, and the minimizers of functions $f_{\boldsymbol{\omega}}$ belong to Θ and are of the form

$$\boldsymbol{x}^*_{\boldsymbol{\omega}} = \left(x^\star(\omega_1), \dots, x^\star(\omega_d)\right),$$

where $x^\star(\omega_i) = -\omega_i \alpha^{-1}(1+\delta)^{-1} r h^{\beta-1}$.

For any fixed $\boldsymbol{\omega} \in \Omega$, we denote by $\mathbf{P}_{\boldsymbol{\omega},T}$ the probability measure corresponding to the joint distribution of $(A_T, (\boldsymbol{\tau}_i)_{i=2}^T)$ where $y_t = f_{\boldsymbol{\omega}}(\boldsymbol{z}_t) + \xi_t$ and $y'_t = f_{\boldsymbol{\omega}}(\boldsymbol{z}'_t) + \xi'_t$ with (ξ_t, ξ'_t)'s satisfying Assumption E, and $(\boldsymbol{z}_t, \boldsymbol{z}'_t)$'s chosen by a sequential strategy in Π_T. Consider the statistic

$$\hat{\boldsymbol{\omega}} \in \arg\min_{\boldsymbol{\omega}\in\Omega} \left\|\tilde{\boldsymbol{x}}_T - \boldsymbol{x}^*_{\boldsymbol{\omega}}\right\|.$$

Classical triangle inequality based arguments yield

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

Note that for all $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$ such that $\rho(\boldsymbol{\omega}, \boldsymbol{\omega}') = 1$ we have

$$\max_{\boldsymbol{u}\in\mathbb{R}^d} |f_{\boldsymbol{\omega}}(\boldsymbol{u}) - f_{\boldsymbol{\omega}'}(\boldsymbol{u})| \leq 2rh^{\beta}\eta(1) \leq 2rT^{-1/2}\eta(1).$$

Thus, choosing r small enough to satisfy $2r\eta(1) < \min(v_0, I_0^{-1/2})$ we ensure $2rT^{-1/2}\eta(1) \leq v_0$ to apply Lemma 34 and deduce for the considered $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$ that

$$H^2(\mathbf{P}_{\boldsymbol{\omega},T}, \mathbf{P}_{\boldsymbol{\omega}',T}) \leq 2\Big(1 - \big(1 - (2T)^{-1}\big)^T\Big) \leq 1,$$

where we have used the fact that $1 - x \geq 4^{-x}$ for $0 < x \leq 1/2$. Applying (Tsybakov, 2009, Theorem 2.12) we deduce that

$$\inf_{\hat{\boldsymbol{\omega}}} \max_{\boldsymbol{\omega}\in\Omega} \mathbf{E}_{\boldsymbol{\omega},T}[\rho(\hat{\boldsymbol{\omega}}, \boldsymbol{\omega})] \geq 0.3\, d.$$

Therefore, we have proved that if $\alpha \geq T^{-1/2+1/\beta}$ there exist $r > 0$ and $\delta > 0$ such that

$$\max_{\boldsymbol{\omega}\in\Omega} \mathbf{E}_{\boldsymbol{\omega},T}\big[\,\|\tilde{\boldsymbol{x}}_T - \boldsymbol{x}^*_{\boldsymbol{\omega}}\|^2\,\big] \geq 0.3\, d\alpha^{-2} r^2 h^{2\beta-2} = 0.3\, r^2 \min\Big(1,\, \frac{d}{\alpha^2} T^{-\frac{\beta-1}{\beta}}\Big). \tag{75}$$

This implies (26) for $\alpha \geq T^{-1/2+1/\beta}$. In particular, if $\alpha = \alpha_0 := T^{-1/2+1/\beta}$ the bound (75) 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 (26).

We now prove (25). From (75) and α-strong convexity of f we get that, for $\alpha \geq T^{-\frac{\beta+2}{2\beta}}$,

$$\max_{\boldsymbol{\omega}\in\Omega} \mathbf{E}_{\boldsymbol{\omega},T}\big[f(\tilde{\boldsymbol{x}}_T) - f(x^*_{\boldsymbol{\omega}})\big] \geq 0.15\, r^2 \min\Big(\alpha,\, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}\Big)\,. \tag{76}$$

This implies (25) in the zone $\alpha \geq T^{-\frac{\beta+2}{2\beta}} = \alpha_0$ since for such α we have

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

On the other hand, $\min\big(\alpha_0,\, \frac{d}{\alpha_0} T^{-\frac{\beta-1}{\beta}}\big) = \min\big(T^{-\frac{\beta+2}{2\beta}}, d/\sqrt{T}\big)$. The same lower bound holds for $0 < \alpha < \alpha_0$ by the nestedness argument that we used to prove (26) in the zone $0 < \alpha < \alpha_0$. Thus, (25) follows. ■