ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 9, pp. 817–830.

SPECIAL ISSUE

Gradient-Free Stochastic Optimization for Additive Models

A. Akhavan[*,a] **and A.B. Tsybakov**[**,b]

[*] *University of Oxford, United Kingdom*
[**] *CREST, ENSAE, IP Paris, France*
e-mail: [a] *arya.akhavan@stats.ox.ac.uk,* [b] *alexandre.tsybakov@ensae.fr*

Received March 3, 2025
Revised May 25, 2025
Accepted June 27, 2025

Abstract—We address the problem of zero-order optimization from noisy observations for an objective function satisfying the Polyak–Łojasiewicz or the strong convexity condition. Additionally, we assume that the objective function has an additive structure and satisfies a higher-order smoothness property, characterized by the Hölder family of functions. The additive model for Hölder classes of functions is well-studied in the literature on nonparametric function estimation, where it is shown that such a model benefits from a substantial improvement of the estimation accuracy compared to the Hölder model without additive structure. We study this established framework in the context of gradient-free optimization. We propose a randomized gradient estimator that, when plugged into a gradient descent algorithm, allows one to achieve minimax optimal optimization error of the order $dT^{-(\beta-1)/\beta}$, where d is the dimension of the problem, T is the number of queries and $\beta \geqslant 2$ is the Hölder degree of smoothness. We conclude that, in contrast to nonparametric estimation problems, no substantial gain of accuracy can be achieved when using additive models in gradient-free optimization.

Keywords: additive model, gradient-free optimization, minimax optimality, Polyak-Łojasiewicz condition

DOI: 10.31857/S0005117925090026

1. INTRODUCTION

Additive modeling is a popular approach to dimension reduction in nonparametric estimation problems [12, 28, 30]. It consists of considering that the unknown function $f : \mathbb{R}^d \to \mathbb{R}$ to be estimated from the data has the form $f(\boldsymbol{x}) = \sum_{j=1}^d f_j(x_j)$, where x_j's are the coordinates of $\boldsymbol{x} \in \mathbb{R}^d$ and f_j's are unknown functions of one variable. The main property proved in the literature on additive models in nonparametric regression can be summarized as follows. If each of the functions f_j is β-Hölder (see Definition 1 below) then the minimax rate of estimation of f, pointwise or in L_2-norm, is of the order $n^{-\beta/(2\beta+1)}$, where n is the number of observations [28]. This is in contrast with the problem of estimating β-Hölder functions on $\mathbb{R}^d$ without any additive structure, since for such functions the minimax rate is known to be $n^{-\beta/(2\beta+d)}$ [13, 14, 26, 27]. Thus, there is a substantial improvement in the rate of estimation when passing from general to additive models in nonparametric regression setting.

In the present paper, we show that such a dimension reduction property fails to hold in the context of gradient-free optimization. We consider additive modeling in the problem of minimizing an unknown function $f : \mathbb{R}^d \to \mathbb{R}$ when only sequential evaluations of values of f are available, cor-

rupted with noise. We assume that f is either strongly convex or satisfies the Polyak–Łojasiewicz (PL) condition [18, 22] and admits an additive representation as described above, where the components f_j are β-Hölder.

The setting that we consider belongs to the family of gradient-free (or zero-order) stochastic optimization problems, for which a rich literature is now available, see [1, 2, 4–6, 8, 10, 15, 17, 19–21, 23, 25] and the references therein. These papers did not assume any additive structure of f. It was proved in [23] that the minimax optimal rate of the optimization error, when f is β-Hölder with $\beta \geqslant 2$ and satisfies the quadratic growth condition, is of the order $T^{-(\beta-1)/\beta}$ as function of the number of sequential queries T, to within an unspecified factor depending on the dimension d. Further developments were devoted to exploring the dependency of the minimax rate on d assuming that f is β-Hölder with $\beta \geqslant 2$ and is either strongly convex [1, 4, 15, 21, 25] or satisfies the PL condition [1, 9]. In the PL case, unconstrained minimization was studied while the strongly convex case was analyzed both in constrained and unconstrained settings. A considerable progress was achieved though the complete solution valid for all $\beta \geqslant 2$ is not yet available. There exists a minimax lower bound for the class of β-Hölder and strongly convex functions, which scales as $dT^{-(\beta-1)/\beta}$ (proved in [25] for $\beta = 2$ and Gaussian noise and in [2] for all $\beta \geqslant 2$ and more general noise; see also [1] for a yet more general lower bound). Moreover, the same rate is attained for $\beta = 2$ under general conditions on the noise (no independence or zero-mean assumption), see [2]. Thus, for $\beta = 2$ we know that the minimax rate is of the order $d\sqrt{T}$. For $\beta > 2$, the literature provides different dependencies of the upper bounds on d determined by the geometry of the β-Hölder condition. Thus, for the Hölder classes defined by pointwise Taylor approximation the best known upper bound is of the order $d^{2-1/\beta}T^{-(\beta-1)/\beta}$ when strongly convex functions are considered [21]. On the other hand, for the Hölder classes defined by tensor-type conditions one can achieve the rate $d^{2-2/\beta}T^{-(\beta-1)/\beta}$ both for strongly convex functions and for PL functions [1]. Finally, the recent paper [31], assuming again strong convexity, deals with the class of functions that admit a Lipschitz Hessian. This represents a type of Hölder condition for $\beta = 3$ and the paper [31] derives the upper rate $dT^{-2/3}$. We note that the lower bound of [2] with the rate $dT^{-(\beta-1)/\beta}$ is valid for all the above mentioned Hölder classes since this lower bound is obtained on additive functions that belong to all of them. Thus, the rate $dT^{-(\beta-1)/\beta}$ appears to be minimax optimal not only for $\beta = 2$ but also for $\beta = 3$ under a suitable definition of 3-Hölder class of strongly convex functions.

The main result of the present paper is to establish an upper bound with the rate $dT^{-(\beta-1)/\beta}$ for the optimization error in the noisy gradient-free setting over the class of β-Hölder functions satisfying the additive model and either the PL condition or the strong convexity condition. Together with the lower bound proved in [2], it implies that $dT^{-(\beta-1)/\beta}$ is the minimax optimal rate of the optimization error in this setting for all $\beta \geqslant 2$. This conclusion is quite surprising as it goes against the intuition acquired from the classical results on nonparametric estimation mentioned above. Indeed, it means that, at least for $\beta \in \{2, 3\}$, there is no improvement in the rate neither in T nor in d when passing from the general β-Hölder model to the additive β-Hölder model. If any, an improvement can be obtained only in a factor independent of T and d. Such a property may be explained by the fact that the optimization setting is easier than nonparametric function estimation in the sense that it aims at estimating a specific functional of the unknown f (namely, its minimizer) rather than f as a whole object. We also refer to a another somewhat similar fact in the gradient-free stochastic optimization setting. Specifically, there is no dramatic difference between the complexity of minimizing a strongly convex function with Lipschitz gradient, which corresponds to the case $\beta = 2$ discussed above with the minimax rate $d\sqrt{T}$, and the complexity of minimizing a convex function with no additional properties, for which one can construct an algorithm converging with the rate between $d^{1.5}\sqrt{T}$ and $d^{1.75}\sqrt{T}$ (up to a logarithmic factor) as recently shown in [7].

2. PROBLEM SETUP

Let Θ be a closed convex subset of $\mathbb{R}^d$. We consider the problem of minimizing an unknown function $f : \mathbb{R}^d \to \mathbb{R}$ over the set Θ based on noisy evaluations of f at query points that can be chosen sequentially depending on the past observations. Specifically, we assume that at round $t \in \{1, \dots, T\}$ we can observe two noisy evaluations of f at points $\boldsymbol{z}_t, \boldsymbol{z}'_t \in \mathbb{R}^d$, i.e.,

$$y_t = f(\boldsymbol{z}_t) + \xi_t, \qquad y'_t = f(\boldsymbol{z}'_t) + \xi'_t,$$

where ξ_t, ξ'_t are scalar noise variables and the query points $\boldsymbol{z}_t, \boldsymbol{z}'_t$ can be chosen depending on $\{\boldsymbol{z}_i, \boldsymbol{z}'_i, y_i, y'_i\}_{i=1}^{t-1}$ and on a suitable randomization.

Throughout the paper we assume that f follows the additive model

$$f(\boldsymbol{x}) = \sum_{j=1}^{d} f_j(x_j),$$

where x_j's are the coordinates of $\boldsymbol{x} \in \mathbb{R}^d$ and f_j's are unknown functions of one variable.

We assume that each of the functions $f_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \dots, d$, belongs to the class of β-Hölder functions specified by the following definition, with $\beta \geqslant 2$.

Definition 1. For $\beta > 0$, $L > 0$, denote by $\mathcal{F}_\beta(L)$ the set of all functions $f : \mathbb{R} \to \mathbb{R}$ that are $\ell = \lfloor \beta \rfloor$ times differentiable and satisfy, for all $x, z \in \mathbb{R}$, the condition

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

where $f^{(m)}$ is the mth derivative of f and $\lfloor \beta \rfloor$ is the largest integer less than β. Elements of the class $\mathcal{F}_\beta(L)$ are referred to as **β-Hölder functions**.

If $\beta > 2$ the fact that $f_j \in \mathcal{F}_\beta(L)$ does not imply that $f_j \in \mathcal{F}_2(L)$, however we will need the latter condition as well. It will be convenient to use it in a slightly different form given by the next definition.

Definition 2. The function $f : \mathbb{R} \to \mathbb{R}$ is called $\bar{L}$-smooth if it is differentiable on $\mathbb{R}$ and there exists $\bar{L} > 0$ such that, for every $x, x' \in \mathbb{R}$, it holds that

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

The class of all $\bar{L}$-smooth functions will be denoted by $\mathcal{F}'_2(\bar{L})$.

We also assume that f is either an α-strongly convex function or an α-PL function as stated in the next two definitions.

Definition 3. Let $\alpha > 0$. The function $f : \mathbb{R}^d \to \mathbb{R}$ is called an α-Polyak–Łojasiewicz function (shortly, α-PL function) if f is differentiable on $\mathbb{R}^d$ and

$$2\alpha \left(f(\boldsymbol{x}) - \min_{\boldsymbol{z} \in \mathbb{R}^d} f(\boldsymbol{z}) \right) \leqslant \|\nabla f(\boldsymbol{x})\|^2 \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^d,$$

where $\|\cdot\|$ denotes the Euclidean norm.

Functions satisfying the PL condition are not necessarily convex. The PL condition is a useful tool in optimization problems since it leads to linear convergence of the gradient descent algorithm without convexity as shown by Polyak [22]. For more details and discussion on the PL condition see [16].

Algorithm 1: Zero-Order Stochastic Projected Gradient

Input: Constraint set Θ, kernel function $K : [-1, 1] \to \mathbb{R}$, step size $\eta_t > 0$ and perturbation parameter $h_t > 0$, for $t = 1, \dots, T$

Initialization: Generate vectors $\boldsymbol{r}_t = (r_{t,1}, \dots, r_{t,d}) \in \mathbb{R}^d$ for $t = 1, \dots, T$, where the components $r_{t,i}$ are independent and distributed uniformly from the interval $[-1, 1]$, and choose $\boldsymbol{x}_1 \in \Theta$

for $t = 1, \dots, T$ **do**

  Observe $y_t = f(\boldsymbol{x}_t + h_t \boldsymbol{r}_t) + \xi_t$ and $y' = f(\boldsymbol{x}_t - h_t \boldsymbol{r}_t) + \xi'_t$; `// query`

  for $j = 1, \dots, d$ **do**

    $g_{t,j} = \frac{d}{2h_t}(y_t - y'_t)K(r_{t,j})$

  Let $\boldsymbol{g}_t = (g_{t,1}, \dots, g_{t,d})$; `// gradient estimator`

  $\boldsymbol{x}_{t+1} = \mathrm{Proj}_\Theta(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t)$; `// update`

Definition 4. Let $\alpha > 0$. The function $f : \mathbb{R}^d \to \mathbb{R}$ is called α-strongly convex if f is differentiable on $\mathbb{R}^d$ and

$$f(\boldsymbol{x}) - f(\boldsymbol{x}') \leqslant \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}' \rangle - \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2 \qquad \text{for all } \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d.$$

In order to minimize f, we apply a version of projected gradient descent with estimated gradient presented in Algorithm 1. Let $\{\eta_t\}_{t=1}^T$ be a sequence of positive numbers and let $\{\boldsymbol{g}_t\}_{t=1}^T$ be a sequence of random vectors. Consider any fixed $\boldsymbol{x}_1 \in \mathbb{R}^d$ and let the vectors $\boldsymbol{x}_t$ for $t = 2, \dots, T$ be defined by the recursion

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

where $\mathrm{Proj}_\Theta(\cdot)$ is the Euclidean projection over Θ.

In this paper, the gradient estimator $\boldsymbol{g}_t = (g_{t,1}, \dots, g_{t,d}) \in \mathbb{R}^d$ at round $t \in \{1, \dots, T\}$ of the algorithm is defined as follows. For given $\beta \geqslant 2$ and $\ell = \lfloor \beta \rfloor$, let $K : [-1, 1] \to \mathbb{R}$ be a function such that

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

Assume that $\kappa := \int K^2(r)dr$ is finite. Functions K satisfying these conditions are not hard to construct. In particular, a construction based on Legendre polynomials can be used, see, for example, [4, 23, 29].

At each round t of the algorithm, we generate a random vector $\boldsymbol{r}_t = (r_{t,1}, \dots, r_{t,d}) \in \mathbb{R}^d$, where the components $r_{t,j}$ are independent and distributed uniformly on $[-1, 1]$. For $h_t > 0$, we draw two noisy evaluations

$$y_t = f(\boldsymbol{x}_t + h_t \boldsymbol{r}_t) + \xi_t, \qquad y'_t = f(\boldsymbol{x}_t - h_t \boldsymbol{r}_t) + \xi'_t,$$

and, for $j \in \{1, \dots, d\}$, we define

$$g_{t,j} = \frac{1}{2h_t}(y_t - y'_t)K(r_{t,j}). \tag{4}$$

We consider the gradient estimator $\boldsymbol{g}_t = (g_{t,1}, \dots, g_{t,d})$. Note that other choices of gradient estimator can lead to similar results as those that we obtain below, namely estimators based on finite difference approximations taking into account higher order smoothness. In contrast to (4), such higher order finite difference schemes have a complicated form and require many queries per step of the algorithm.

We assume that the noise variables ξ_t, ξ'_t and the randomizing variables $r_{t,j}$ satisfy the following.

Assumption 1. There exists $\sigma^2 > 0$ such that for all $t \in \{1, \dots, T\}$ the following holds.

(i) The random variables $r_{t,j} \sim U[-1, 1]$, $j = 1, \dots, d$, are independent of $\boldsymbol{x}_t$ and conditionally independent of ξ_t, ξ'_t given $\boldsymbol{x}_t$,

(ii) $\mathbb{E}[\xi_t^2] \leqslant \sigma^2$, $\mathbb{E}[(\xi'_t)^2] \leqslant \sigma^2$.

Assumption 1 (i) can be considered not as a restriction but as a part of the definition of the algorithm dealing with the choice of the randomizing variables $r_{t,j}$. Randomizations are naturally chosen independent of all other sources of randomness. For the proofs we need even a weaker property that we state here as an assumption in order to refer to it in what follows. Note also that Assumption 1 does not require the noises ξ_t, ξ'_t to have zero mean. Moreover, they can be non-random and we do not assume independence between these noises on different rounds of the algorithm. The fact that convergence of gradient-free algorithms can be achieved under general conditions on the noise of similar type, not requiring independence and zero means, dates back to [11, 24].

3. STATEMENT OF THE RESULTS

In this section, we provide upper bounds on the optimization error of the algorithm defined in Section 2. First, we assume that function f represented by the additive model satisfies the α-PL condition. Note that imposing this condition on the sum f implies that all the components f_j are PL functions. When dealing with PL functions we consider the problem of unconstrained minimization and we introduce the notation

$$f^* = \min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}).$$

In what follows, for any positive integer n we denote by $[n]$ the set of positive integers less than or equal to n.

Theorem 1. *Let $\alpha > 0$, $\beta \geqslant 2$ and $\bar{L}, L > 0$. For $j \in [d]$, let $f_j : \mathbb{R} \to \mathbb{R}$ be such that $f_j \in \mathcal{F}_2'(\bar{L}) \cap \mathcal{F}_\beta(L)$. Assume that $f : \mathbb{R}^d \to \mathbb{R}$ has the form $f(\boldsymbol{x}) = \sum_{j=1}^d f_j(x_j)$, and that f is an α-PL function. Let Assumption 1 hold, and let $\{\boldsymbol{x}_t\}_{t=1}^T$ be the updates of Algorithm 1 with $\Theta = \mathbb{R}^d$, and*

$$\eta_t = \min\Big(\frac{4}{\alpha t}, \frac{1}{18\bar{L}d\kappa}\Big),$$

$$h_t = \left(\frac{3\bar{L}}{\alpha} \cdot \frac{\kappa\sigma^2}{L^2\kappa_\beta^2}\right)^{\frac{1}{2\beta}} \cdot \begin{cases} t^{-\frac{1}{2\beta}} & \text{if } \eta_t = \dfrac{4}{\alpha t}, \\ T^{-\frac{1}{2\beta}} & \text{if } \eta_t = \dfrac{1}{18\bar{L}d\kappa}. \end{cases}$$

Then,

$$\mathbf{E}\left[f(\boldsymbol{x}_T) - f^*\right] \leqslant \frac{144\bar{L}d\kappa}{\alpha T}(f(\boldsymbol{x}_1) - f^*) + \frac{d}{\alpha}\left(\mathsf{A}_1\left(\frac{\bar{L}}{\alpha T}\right)^{\frac{\beta-1}{\beta}} + \mathsf{A}_2\frac{\bar{L}^3 d}{T}\left(\frac{\bar{L}}{\alpha T L^2}\right)^{\frac{1}{\beta}}\right),$$

where $\mathsf{A}_1, \mathsf{A}_2 > 0$ depend only on β and σ^2.

Corollary 1. *Under the conditions of Theorem 1, if $T \geqslant (\bar{L}/\alpha)((\bar{L}^2 d\alpha)^{\frac{\beta}{2}}/L)$, then*

$$\mathbf{E}\left[f(\boldsymbol{x}_T) - f^*\right] \leqslant \frac{144\bar{L}d\kappa}{\alpha T}(f(\boldsymbol{x}_1) - f^*) + \mathsf{A}\frac{d}{\alpha}\left(\frac{\bar{L}}{\alpha T}\right)^{\frac{\beta-1}{\beta}},$$

where $\mathsf{A} > 0$ depends only on β and σ^2.

The bounds of Theorem 1 and Corollary 1 show how the rate of convergence depends on the parameters T, d and α but also on the aspect ratio $\bar{L}/\alpha$. Moreover, in Corollary 1 the condition $T \geqslant Cd^{\beta/2}$, where $C > 0$ is a constant, does not bring an additional restriction when $2 \leqslant \beta \leqslant 3$ since the condition is weaker than giving the range of T, for which the bound makes sense. Indeed, for $2 \leqslant \beta \leqslant 3$ the bound is greater than a constant independent of T, d if $T \leqslant Cd^{\beta/2}$.

Next, we study the optimization error of the algorithm defined in Section 2 under the assumption that the objective function is α-strongly convex. Unlike the case of PL functions, we consider here the problem of constrained minimization

$$\min_{\boldsymbol{x}\in\Theta} f(\boldsymbol{x}),$$

where Θ is a compact convex subset of $\mathbb{R}^d$.

Theorem 2. *Let $\alpha > 0$, $\beta \geqslant 2$ and $\bar{L}, L > 0$. For $j \in [d]$, let $f_j : \mathbb{R} \to \mathbb{R}$ be such that $f_j \in \mathcal{F}_2'(\bar{L}) \cap \mathcal{F}_\beta(L)$. Assume that $f : \mathbb{R}^d \to \mathbb{R}$ has the form $f(\boldsymbol{x}) = \sum_{j=1}^d f_j(x_j)$, and that f is an α-strongly convex function on a compact convex subset Θ of $\mathbb{R}^d$ such that $\max_{\boldsymbol{x}\in\Theta} \|\nabla f(\boldsymbol{x})\| \leqslant G$. Let Assumption 1 hold, and let $\{\boldsymbol{x}_t\}_{t=1}^T$ be the updates of Algorithm 1 with*

$$\eta_t = \frac{4}{\alpha(T+1)}, \qquad h_t = \left(\frac{3}{2t}\frac{\kappa\sigma^2}{\kappa_\beta^2 L^2}\right)^{\frac{1}{2\beta}}.$$

Consider the weighted estimator

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

Then, for any $\boldsymbol{x} \in \Theta$ we have

$$\mathbf{E}\left[f(\bar{\boldsymbol{x}}_T) - f(\boldsymbol{x})\right] \leqslant \frac{18G^2 d}{\alpha T} + \frac{d}{\alpha}\left(\mathsf{A}_1 L^{\frac{2}{\beta}} + \mathsf{A}_2 d\bar{L}^2 (LT)^{-\frac{2}{\beta}}\right) T^{-\frac{\beta-1}{\beta}},$$

where $\mathsf{A}_1, \mathsf{A}_2 > 0$ depend only on β and σ^2.

Corollary 2. *Under the conditions of Theorem 2, if $T \geqslant (d\bar{L})^{\beta/2} L^{-2}$ then*

$$\mathbf{E}\Big[f(\bar{\boldsymbol{x}}_T) - \min_{\boldsymbol{x}\in\Theta} f(\boldsymbol{x})\Big] \leqslant \mathsf{A}\frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}},$$

where $\mathsf{A} > 0$ does not depend on T, d and α.

Similarly to Corollary 1, we may note that that for $2 \leqslant \beta \leqslant 3$ the condition $T \geqslant Cd^{\beta/2}$ in Corollary 2, where $C > 0$ is a constant, does not bring an additional restriction but indicates the meaningful range of T since the bound is greater than a constant when $T \leqslant Cd^{\beta/2}$.

Remark 1. In view of strong convexity, Theorem 2 and Corollary 2 immediately imply the corresponding bounds on the estimation error $\mathbf{E}[\|\bar{\boldsymbol{x}}_T - \boldsymbol{x}^*\|^2]$, where $\boldsymbol{x}^*$ is the minimizer of f on Θ. Thus, under the assumptions of Corollary 2 we have

$$\mathbf{E}[\|\bar{\boldsymbol{x}}_T - \boldsymbol{x}^*\|^2] \leqslant 2\mathsf{A}\frac{d}{\alpha^2} T^{-\frac{\beta-1}{\beta}},$$

where $\mathsf{A} > 0$ is the constant from Corollary 2.

It follows from Corollary 2 and the proofs of the lower bounds in [1, 2] that, under non-restrictive conditions on the parameters of the problem, the rate $\frac{d}{\alpha}T^{-\frac{\beta-1}{\beta}}$ in Corollary 2 is minimax optimal on the class of additive functions f satisfying the assumptions of Theorem 2. The lower bound that we need is not explicitly stated in [1, 2] but follows immediately from the proofs in those papers since the lower bounds in [1, 2] are obtained on additive functions. For completeness, we provide here the statement of the lower bound for additive functions based on [1].

Consider all strategies of choosing the query points as $\boldsymbol{z}_t = \Phi_t\big((\boldsymbol{z}_i, y_i)_{i=1}^{t-1}, (\boldsymbol{z}'_i, y'_i)_{i=1}^{t-1}, \tau_t\big)$ and $\boldsymbol{z}'_t = \Phi'_t\big((\boldsymbol{z}_i, y_i)_{i=1}^{t-1}, (\boldsymbol{z}'_i, y'_i)_{i=1}^{t-1}, \tau_t\big)$ for $t \geqslant 2$, 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 $\{\tau_t\}$ is a sequence of random variables with values in a measurable space $(\mathcal{Z}, \mathcal{U})$, such that τ_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 the algorithm of Section 2 with the gradient estimator (4). In this case, $\tau_t = \boldsymbol{r}_t$, $\boldsymbol{z}_t = \boldsymbol{x}_t + h_t \boldsymbol{r}_t$ and $\boldsymbol{z}'_t = \boldsymbol{x}_t - h_t \boldsymbol{r}_t$.

The lower bound of [1] that we are using here is proved under the following assumption on the noises (ξ, ξ'_t). Let $H^2(\cdot, \cdot)$ be the squared Hellinger distance 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 2. For every $t \geqslant 1$, 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

$$H^2(P_{F_t(\cdot,\cdot)}, P_{F_t(\cdot+v,\cdot+v)}) \leqslant I_0 v^2, \qquad |v| \leqslant v_0, \tag{5}$$

  for some $0 < I_0 < \infty$, $0 < v_0 \leqslant \infty$. 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 $((\boldsymbol{z}_i, y_i)_{i=1}^{t-1}, (\boldsymbol{z}'_i, y'_i)_{i=1}^{t-1}, \tau_t)$.

Let $\Theta = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leqslant 1\}$. Fix $\alpha, L, \bar{L} > 0, G > \alpha$, $\beta \geqslant 2$, and denote by $\mathcal{F}$ the set of all functions f that satisfy the assumptions of Theorem 2 and attain their minimum over $\mathbb{R}^d$ in Θ.

Theorem 3. *Let $\Theta = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leqslant 1\}$ and let Assumption 2 hold. Assume that $\alpha > T^{-1/2+1/\beta}$ and $T \geqslant d^\beta$. 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*

$$\sup_{f \in \mathcal{F}} \mathbf{E}\big[f(\tilde{\boldsymbol{x}}_T) - \min_{\boldsymbol{x} \in \Theta} f(\boldsymbol{x})\big] \geqslant C \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}, \tag{6}$$

where $C > 0$ is a constant that does not depend of T, d, and α.

Theorem 3 follows immediately from the proof of Theorem 22 in [1] since the family of functions used there belongs to the class $\mathcal{F}$. The lower bound of Theorem 22 in [1] has the form

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

which reduces to $C\frac{d}{\alpha}T^{-\frac{\beta-1}{\beta}}$ under the assumptions on T, d and α used in Theorem 3.

Remark 2. Since the strong convexity and the PL property hold for each additive component of f, another possible approach would be to run the procedure component-wise (minimize separately each component f_j). However, it leads to a worse result. Indeed, in this case we need to make at each step $2d$ queries in parallel (two queries for each component) and thus can make only $\sim T/d$

steps if the total budget of queries is T. At the end, applying Theorems 1 or 2 in the one-dimensional case, for each component we obtain the error in T and d of the order $(T/d)^{-(\beta-1)/\beta}$. This rate cannot be improved as follows from the one-dimensional instance of Theorem 3. Summing up over the d components, the overall error will be of the order $d(T/d)^{-(\beta-1)/\beta} = d^{2-1/\beta}T^{-(\beta-1)/\beta}$, that is, the error will depend on d in a sub-optimal way.

4. PROOFS

We start by proving some auxiliary lemmas.

Lemma 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function such that $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\| \leqslant \bar{L}\|\boldsymbol{x} - \boldsymbol{x}'\|$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$, where $\bar{L} > 0$. Let Assumption 1 hold and let $\{\boldsymbol{x}_t\}_{t=1}^T$ be the updates of Algorithm 1 with $\Theta = \mathbb{R}^d$. Then, for all $t \in [T]$ we have*

$$\mathbf{E}\left[f(\boldsymbol{x}_{t+1})|\boldsymbol{x}_t\right] \leqslant f(\boldsymbol{x}_t) - \frac{\eta_t}{2}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{\eta_t}{2}\|\mathbf{E}\left[\boldsymbol{g}_t|\boldsymbol{x}_t\right] - \nabla f(\boldsymbol{x}_t)\|^2 + \frac{\bar{L}\eta_t^2}{2}\mathbf{E}\left[\|\boldsymbol{g}_t\|^2|\boldsymbol{x}_t\right]. \tag{7}$$

Proof. The assumption on f implies that

$$f(\boldsymbol{x}_{t+1}) = f(\boldsymbol{x}_t - \eta_t\boldsymbol{g}_t) \leqslant f(\boldsymbol{x}_t) - \eta_t\langle\nabla f(\boldsymbol{x}_t), \boldsymbol{g}_t\rangle + \frac{\bar{L}\eta_t^2}{2}\|\boldsymbol{g}_t\|^2.$$

By adding and subtracting $\eta_t\|\nabla f(\boldsymbol{x}_t)\|^2$ we obtain

$$f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t) - \eta_t\|\nabla f(\boldsymbol{x}_t)\|^2 - \eta_t\langle\nabla f(\boldsymbol{x}_t),\ \boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t)\rangle + \frac{\bar{L}\eta_t^2}{2}\|\boldsymbol{g}_t\|^2.$$

Taking the conditional expectation gives

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

The lemma follows by using the inequality $2ab \leqslant a^2 + b^2$, $\forall a, b \in \mathbb{R}$.

Lemma 2. *For $j \in [d]$, let $f_j : \mathbb{R} \to \mathbb{R}$ be such that $f_j \in \mathcal{F}_\beta(L)$, where $\beta \geqslant 2$ and $L > 0$. Assume that $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the additive model $f(\boldsymbol{x}) = \sum_{j=1}^d f_j(x_j)$. Let Assumption 1(i) hold. Then, for all $t \in [T]$ we have*

$$\|\mathbf{E}\left[\boldsymbol{g}_t|\boldsymbol{x}_t\right] - \nabla f(\boldsymbol{x}_t)\| \leqslant \kappa_\beta L\sqrt{d}h_t^{\beta-1}.$$

Proof. Using Assumption 1(i) for any $j \in [d]$ and $t \in [T]$ we have $\mathbf{E}(K(r_{t,j})) = 0$, $\mathbf{E}\left[\xi_t K(r_{t,j})|\boldsymbol{x}_t\right] = 0$ and $\mathbf{E}\left[\xi_t' K(r_{t,j})|\boldsymbol{x}_t\right] = 0$. Thus,

$$\begin{aligned}\mathbf{E}\left[g_{t,j}|\boldsymbol{x}_t\right] &= \frac{1}{2h_t}\mathbf{E}\left[(f_j(x_{t,j} + h_t r_{t,j}) - f_j(x_{t,j} - h_t r_{t,j}))K(r_{t,j})|\boldsymbol{x}_t\right] \\ &\quad + \frac{1}{2h_t}\sum_{m\neq j}\mathbf{E}\left[(f_m(x_{t,m} + h_t r_{t,m}) - f_m(x_{t,m} - h_t r_{t,m}))K(r_{t,j})|\boldsymbol{x}_t\right] \\ &= \frac{1}{2h_t}\mathbf{E}\left[(f_j(x_{t,j} + h_t r_{t,j}) - f_j(x_{t,j} - h_t r_{t,j}))K(r_{t,j})|\boldsymbol{x}_t\right],\end{aligned}$$

where we have used the fact that $\mathbf{E}(K(r_{t,j})) = 0$ and $r_{t,j}$ is independent of $r_{t,m}$, for $m \neq j$, and of $\boldsymbol{x}_t$. From Taylor expansion we obtain that

$$\begin{aligned}\frac{1}{2h_t}\left(f_j(x_{t,j} + h_t r_{t,j}) - f_j(x_{t,j} - h_t r_{t,j})\right) &= f_j'(x_{t,j})r_{t,j} + \frac{1}{h_t}\sum_{1\leqslant m\leqslant \ell,\, m \text{ odd}}\frac{h_t^m}{m!}f_j^{(m)}(x_{t,j})r_{t,j}^m \\ &\quad + \frac{R(h_t r_{t,j}) - R(-h_t r_{t,j})}{2h_t},\end{aligned}$$

where $|R(-h_t r_{t,j})|, |R(h_t r_{t,j})| \leqslant L|r_{t,j}|^\beta h_t^\beta$. Multiplying both sides of this inequality by $K(r_{t,j})$ and taking the conditional expectation implies

$$|\mathbf{E}\left[g_{t,j}|\boldsymbol{x}_t\right] - f_j'(x_{t,j})| \leqslant L\mathbf{E}\left[|r_{t,j}|^\beta K(r_{t,j})\right] h_t^{\beta-1} = \kappa_\beta L h_t^{\beta-1}.$$

The result of the lemma follows from this inequality and the fact that

$$\|\mathbf{E}\left[\boldsymbol{g}_t|\boldsymbol{x}_t\right] - \nabla f(\boldsymbol{x}_t)\| \leqslant \sqrt{d}\max_{j=1,\dots,d}|\mathbf{E}\left[g_{t,j}|\boldsymbol{x}_t\right] - f_j'(x_{t,j})|.$$

Lemma 3. *For $j \in [d]$, let $f_j : \mathbb{R} \to \mathbb{R}$ be such that $f_j \in \mathcal{F}_2'(\bar{L})$, where $\bar{L} > 0$. Assume that $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the additive model $f(\boldsymbol{x}) = \sum_{j=1}^d f_j(x_j)$. Let Assumption 1 hold. Then, for all $t \in [T]$ we have*

$$\mathbf{E}\left[\|\boldsymbol{g}_t\|^2|\boldsymbol{x}_t\right] \leqslant \frac{3}{2}\kappa d\left(\frac{3}{4}\left(d\bar{L}^2h_t^2 + 8\|\nabla f(\boldsymbol{x}_t)\|^2\right) + \frac{\sigma^2}{h_t^2}\right).$$

Proof. For $i \in [d]$ we define

$$G_i = f_i(x_{t,i} + h_t r_{t,i}) - f_i(x_{t,i} - h_t r_{t,i}).$$

We have

$$\mathbf{E}\left[g_{t,j}^2|\boldsymbol{x}_t\right] = \frac{1}{4h_t^2}\mathbf{E}\left[\left(\sum_{i=1}^d G_i + \xi_t - \xi_t'\right)^2 K^2(r_{t,j})|\boldsymbol{x}_t\right].$$

Note that $r_{t,i}$ and $-r_{t,i}$ have the same distribution. Therefore, $\mathbf{E}[G_i|\boldsymbol{x}_t] = 0$ and we can write

$$\begin{aligned}
\mathbf{E}\left[g_{t,j}^2|\boldsymbol{x}_t\right] &\leqslant \frac{3}{4h_t^2}\mathbf{E}\left[\left(\left(\sum_{i=1}^d G_i\right)^2 + \xi_t^2 + (\xi_t')^2\right)K^2(r_{t,j})|\boldsymbol{x}_t\right] \\
&\leqslant \frac{3}{4h_t^2}\mathbf{E}\left[\sum_{i=1}^d G_i^2K^2(r_{t,j}) + \sum_{i,k=1,i\neq k}^d G_iG_kK^2(r_{t,j})|\boldsymbol{x}_t\right] + \frac{3\sigma^2\kappa}{2h_t^2} \\
&= \frac{3}{4h_t^2}\mathbf{E}\left[\sum_{i=1}^d G_i^2K^2(r_{t,j})|\boldsymbol{x}_t\right] + \frac{3\sigma^2\kappa}{2h_t^2}.
\end{aligned}$$

Since $f_i \in \mathcal{F}_2'(\bar{L})$ we have that, for all $i \in [d]$,

$$\begin{aligned}
G_i^2 &= \left((f_i(x_{t,i} + h_t r_{t,i}) - f(x_{t,i}) - f_i'(x_{t,i})h_t r_{t,i}) - (f_i(x_{t,i} - h_t r_{t,i}) - f(x_{t,i}) + f_i'(x_{t,i})h_t r_{t,i})\right. \\
&\quad \left. + 2f_i'(x_{t,i})h_t r_{t,i}\right)^2 \\
&\leqslant 3\Big((f_i(x_{t,i} + h_t r_{t,i}) - f(x_{t,i}) - f_i'(x_{t,i})h_t r_{t,i})^2 + (f_i(x_{t,i} - h_t r_{t,i}) - f(x_{t,i}) + f_i'(x_{t,i})h_t r_{t,i})^2 \\
&\quad + 4\left(f_i'(x_{t,i})h_t r_{t,i}\right)^2\Big) \\
&\leqslant 3\left(\frac{\bar{L}^2}{2}h_t^4 + 4\left(f_i'(x_{t,i})\right)^2 h_t^2\right).
\end{aligned}$$

Thus,

$$\mathbf{E}\left[g_{t,j}^2|\boldsymbol{x}_t\right] \leqslant \frac{9}{4}\kappa\left(\frac{d\bar{L}^2}{2}h_t^2 + 4\sum_{i=1}^d(f_i'(x_{t,i}))^2\right) + \frac{3\sigma^2\kappa}{2h_t^2},$$

which implies the lemma.

Lemma 4. *For $j \in [d]$, let $f_j : \mathbb{R} \to \mathbb{R}$ be such that $f_j \in \mathcal{F}_2'(\bar{L}) \cap \mathcal{F}_\beta(L)$, where $\beta \geqslant 2$ and $\bar{L}, L > 0$. Assume that $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the additive model $f(\boldsymbol{x}) = \sum_{j=1}^d f_j(x_j)$. Let Assumption 1 hold. Assume that $\{\boldsymbol{x}_t\}_{t=1}^T$ are updates of Algorithm 1 with $\Theta = \mathbb{R}^d$ and with the gradient estimators defined by (4). Then, for all $t \in [T]$ we have*

$$\mathbf{E}\left[f(\boldsymbol{x}_{t+1})|\boldsymbol{x}_t\right] \leqslant f(\boldsymbol{x}_t) - \frac{\eta_t}{2}\left(1 - 9\bar{L}d\kappa\eta_t\right)\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{\eta_t}{2}d\left(L\kappa_\beta h_t^{\beta-1}\right)^2 + \frac{3\bar{L}\eta_t^2}{4}\kappa d\left(\frac{\sigma^2}{h_t^2} + \frac{3\bar{L}^2 d}{4}h_t^2\right).$$

Proof. The result follows by combining Lemmas 1, 2 and 3.

Lemma 5. *For $\alpha > 0$ assume that f is an α-strongly convex function. Assume that $\{\boldsymbol{x}_t\}_{t=1}^T$ are the updates of Algorithm 1. Then, for all $t \in [T]$ and $\boldsymbol{x} \in \Theta$, we have*

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}) \leqslant (2\eta_t)^{-1}\left(\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \mathbf{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2|\boldsymbol{x}_t\right]\right) + \frac{1}{\alpha}\|\nabla f(\boldsymbol{x}_t) - \mathbf{E}\left[\boldsymbol{g}_t|\boldsymbol{x}_t\right]\|^2 + \frac{\eta_t}{2}\mathbf{E}\left[\|\boldsymbol{g}_t\|^2|\boldsymbol{x}_t\right] - \frac{\alpha}{4}\|\boldsymbol{x}_t - \boldsymbol{x}\|^2.$$

Proof. Fix $\boldsymbol{x} \in \Theta$. Then, by the contraction property of the Euclidean projection we have $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2 \leqslant \|\boldsymbol{x}_t - \eta_t\boldsymbol{g}_t - \boldsymbol{x}\|^2$. Equivalently,

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leqslant (2\eta_t)^{-1}(\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|^2) + \frac{\eta_t}{2}\|\boldsymbol{g}_t\|^2.$$

Let $a_t = \|\boldsymbol{x}_t - \boldsymbol{x}\|^2$. Since f is an α-strongly convex function we have

$$\begin{aligned} f(\boldsymbol{x}_t) - f(\boldsymbol{x}) &\leqslant \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}\rangle - \frac{\alpha}{2}a_t \\ &= \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle + \langle \nabla f(\boldsymbol{x}_t) - \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle - \frac{\alpha}{2}a_t \\ &\leqslant (2\eta_t)^{-1}(a_t - a_{t+1}) + \langle \nabla f(\boldsymbol{x}_t) - \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle + \frac{\eta_t}{2}\|\boldsymbol{g}_t\|^2 - \frac{\alpha}{2}a_t. \end{aligned}$$

Taking the conditional expectation given $\boldsymbol{x}_t$ and using the inequality $ab \leqslant a^2/\lambda + \lambda b^2/4$ valid for all $a, b \in \mathbb{R}$ and $\lambda > 0$ we deduce that

$$\begin{aligned} f(\boldsymbol{x}_t) - f(\boldsymbol{x}) &\leqslant (2\eta_t)^{-1}(a_t - \mathbf{E}\left[a_{t+1}|\boldsymbol{x}_t\right]) + \langle \nabla f(\boldsymbol{x}_t) - \mathbf{E}\left[\boldsymbol{g}_t|\boldsymbol{x}_t\right], \boldsymbol{x}_t - \boldsymbol{x}\rangle + \frac{\eta_t}{2}\mathbf{E}\left[\|\boldsymbol{g}_t\|^2|\boldsymbol{x}_t\right] - \frac{\alpha}{2}a_t \\ &\leqslant (2\eta_t)^{-1}(a_t - \mathbf{E}\left[a_{t+1}|\boldsymbol{x}_t\right]) + \|\nabla f(\boldsymbol{x}_t) - \mathbf{E}\left[\boldsymbol{g}_t|\boldsymbol{x}_t\right]\|\,\|\boldsymbol{x}_t - \boldsymbol{x}\| + \frac{\eta_t}{2}\mathbf{E}\left[\|\boldsymbol{g}_t\|^2|\boldsymbol{x}_t\right] - \frac{\alpha}{2}a_t \\ &\leqslant (2\eta_t)^{-1}(a_t - \mathbf{E}\left[a_{t+1}|\boldsymbol{x}_t\right]) + \frac{1}{\alpha}\|\nabla f(\boldsymbol{x}_t) - \mathbf{E}\left[\boldsymbol{g}_t|\boldsymbol{x}_t\right]\|^2 + \frac{\eta_t}{2}\mathbf{E}\left[\|\boldsymbol{g}_t\|^2|\boldsymbol{x}_t\right] - \frac{\alpha}{4}a_t. \end{aligned}$$

Proof of Theorem 1. Since $\eta_t \leqslant 1/(18\bar{L}d\kappa)$ from Lemma 4 we have

$$\mathbf{E}\left[f(\boldsymbol{x}_{t+1})|\boldsymbol{x}_t\right] \leqslant f(\boldsymbol{x}_t) - \frac{\eta_t}{4}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{\eta_t}{2}d\left(L\kappa_\beta h_t^{\beta-1}\right)^2 + \frac{3\bar{L}\eta_t^2}{4}\kappa d\left(\frac{\sigma^2}{h_t^2} + \frac{3\bar{L}^2 d}{4}h_t^2\right).$$

Taking the expectation from both sides of this inequality and using the fact that f is an α-PL function we obtain

$$\delta_{t+1} \leqslant \delta_t\left(1 - \frac{\eta_t\alpha}{2}\right) + \frac{\eta_t}{2}d\left(L\kappa_\beta h_t^{\beta-1}\right)^2 + \frac{3\bar{L}\eta_t^2}{4}\kappa d\left(\frac{\sigma^2}{h_t^2} + \frac{3\bar{L}^2 d}{4}h_t^2\right), \tag{8}$$

where $\delta_t = \mathbf{E}\left[f(\boldsymbol{x}_t) - f^*\right]$. Let $T_0 = \lfloor 72\bar{L}d\kappa/\alpha \rfloor$. Note that

$$\eta_t = \begin{cases} \dfrac{1}{18\bar{L}d\kappa} & \text{if } t < T_0, \\ \dfrac{4}{\alpha t} & \text{if } t \geqslant T_0. \end{cases}$$

We consider separately the cases $T > T_0$ and $T \leqslant T_0$.

If $T > T_0$, then for $t \geqslant T_0$ we have $\eta_t = 4/(\alpha t)$ and we can write

$$\delta_{t+1} \leqslant \delta_t \left(1 - \frac{2}{t}\right) + \frac{d}{\alpha t}\left(2\left(L\kappa_\beta h_t^{\beta-1}\right)^2 + \frac{3\bar{L}}{\alpha t}\kappa\left(\frac{\sigma^2}{h_t^2} + \frac{3\bar{L}^2 d}{4}h_t^2\right)\right).$$

Substituting here $h_t = (\frac{3\bar{L}}{\alpha t} \cdot \frac{\kappa\sigma^2}{2L^2\kappa_\beta^2})^{\frac{1}{2\beta}}$ gives

$$\delta_{t+1} \leqslant \delta_t \left(1 - \frac{2}{t}\right) + \frac{d}{\alpha t}\left(\mathsf{A}_3\left(\frac{\bar{L}}{\alpha t}\right)^{\frac{\beta-1}{\beta}} + \mathsf{A}_4\frac{\bar{L}^3 d}{t}\left(\frac{\bar{L}}{\alpha t L^2}\right)^{\frac{1}{\beta}}\right),$$

where $\mathsf{A}_3 = 6\left((2\kappa_\beta^2/3)(\kappa\sigma^2)^{\beta-1}\right)^{1/\beta}$ and $\mathsf{A}_4 = 9\left(3\kappa\sigma^2/(2\kappa_\beta^2)\right)^{\frac{1}{\beta}}/4$. Since $T \geqslant T_0$, applying [1, Lemma 32] leads to the bound

$$\delta_T \leqslant \frac{2T_0}{T}\delta_{T_0+1} + \frac{d}{\alpha}\left(\frac{\beta\mathsf{A}_1}{\beta+1}\left(\frac{\bar{L}}{\alpha T}\right)^{\frac{\beta-1}{\beta}} + \frac{\beta\mathsf{A}_2}{2\beta+1}\frac{\bar{L}^3 d}{T}\left(\frac{\bar{L}}{\alpha T L^2}\right)^{\frac{1}{\beta}}\right).$$

If $T_0 = 0$ then the proof is complete for the case $T > T_0$. Otherwise, for any $t \leqslant T_0$, we have $\eta_t = 1/(18\bar{L}d\kappa)$ and $h_t = (\frac{3\bar{L}}{\alpha T} \cdot \frac{\kappa\sigma^2}{2L^2\kappa_\beta^2})^{\frac{1}{2\beta}}$. Note that

$$\frac{4}{(T_0+1)\alpha} \leqslant \eta_t \leqslant \frac{4}{T_0\alpha},$$

and from (8) we deduce that, for all $t \leqslant T_0$,

$$\delta_{t+1} \leqslant \delta_t\left(1 - \frac{2}{T_0+1}\right) + \frac{2d}{\alpha T_0}\left(\mathsf{A}_3\left(\frac{\bar{L}}{\alpha}\right)^{\frac{\beta-1}{\beta}}\left(T^{-\frac{\beta-1}{\beta}} + \frac{T^{\frac{1}{\beta}}}{T_0}\right) + \mathsf{A}_4\frac{\bar{L}^3 d}{T_0}\left(\frac{\bar{L}}{\alpha T L^2}\right)^{\frac{1}{\beta}}\right).$$

Note that $1 - 2/(T_0+1) \leqslant 1$. Thus, by taking into account the definition of T_0 and the fact that $T > T_0$ we get

$$\frac{2T_0}{T}\delta_{t+1} \leqslant \frac{144\bar{L}d\kappa}{\alpha T}\delta_1 + \frac{4d}{\alpha}\left(2\mathsf{A}_3\left(\frac{\bar{L}}{\alpha T}\right)^{\frac{\beta-1}{\beta}} + \mathsf{A}_4\frac{\bar{L}^3 d}{T}\left(\frac{\bar{L}}{\alpha T L^2}\right)^{\frac{1}{\beta}}\right).$$

Therefore,

$$\delta_T \leqslant \frac{144\bar{L}d\kappa}{\alpha T}\delta_1 + \frac{d}{\alpha}\left(\mathsf{A}_1\left(\frac{\bar{L}}{\alpha T}\right)^{\frac{\beta-1}{\beta}} + \mathsf{A}_2\frac{\bar{L}^3 d}{T}\left(\frac{\bar{L}}{\alpha T L^2}\right)^{\frac{1}{\beta}}\right),$$

where $\mathsf{A}_1 = \mathsf{A}_3\left(\beta/(\beta+1) + 8\right)$ and $\mathsf{A}_2 = \mathsf{A}_4\left(\beta/(2\beta+1) + 4\right)$.

Consider now the case $T \leqslant T_0$. Then from (8) we get that, for all $t \leqslant T$,

$$\delta_{t+1} \leqslant \delta_t \left(1 - \frac{2}{T_0+1}\right) + \frac{2d}{\alpha T_0}\left(\mathsf{A}_3 \left(\frac{\bar{L}}{\alpha}\right)^{\frac{\beta-1}{\beta}} \left(T^{-\frac{\beta-1}{\beta}} + \frac{T^{\frac{1}{\beta}}}{T_0}\right) + \mathsf{A}_4 \frac{\bar{L}^3 d}{T_0}\left(\frac{\bar{L}}{\alpha T L^2}\right)^{\frac{1}{\beta}}\right).$$

Thus,

$$\delta_{T+1} \leqslant \delta_1 \left(1 - \frac{2}{T_0+1}\right)^T + \frac{2d}{\alpha}\left(2\mathsf{A}_3 \left(\frac{\bar{L}}{\alpha T}\right)^{\frac{\beta-1}{\beta}} + \mathsf{A}_4 \frac{\bar{L}^3 d}{T}\left(\frac{\bar{L}}{\alpha T L^2}\right)^{\frac{1}{\beta}}\right).$$

Since $(1-\lambda)^T \leqslant \exp(-\lambda T) \leqslant 1/(\lambda T)$ for all $T, \lambda > 0$, we deduce from the previous display that

$$\delta_{T+1} \leqslant \frac{T_0+1}{2T}\delta_1 + \frac{2d}{\alpha}\left(2\mathsf{A}_3 \left(\frac{\bar{L}}{\alpha T}\right)^{\frac{\beta-1}{\beta}} + \mathsf{A}_4 \frac{\bar{L}^3 d}{T}\left(\frac{\bar{L}}{\alpha T L^2}\right)^{\frac{1}{\beta}}\right).$$

By using the fact that $\lambda + 1 \leqslant 2\lambda$ for all $\lambda \geqslant 1$ we finally get

$$\begin{aligned}\delta_{T+1} &\leqslant \frac{2T_0}{T+1}\delta_1 + \frac{2d}{\alpha}\left(2\mathsf{A}_3 \left(\frac{2\bar{L}}{\alpha(T+1)}\right)^{\frac{\beta-1}{\beta}} + \mathsf{A}_4 \frac{2\bar{L}^3 d}{T+1}\left(\frac{2\bar{L}}{\alpha(T+1)L^2}\right)^{\frac{1}{\beta}}\right) \\ &\leqslant \frac{144\bar{L}d\kappa}{\alpha(T+1)}\delta_1 + \frac{d}{\alpha}\left(\mathsf{A}_1 \left(\frac{\bar{L}}{\alpha(T+1)}\right)^{\frac{\beta-1}{\beta}} + \mathsf{A}_2 \frac{\bar{L}^3 d}{T+1}\left(\frac{\bar{L}}{\alpha(T+1)L^2}\right)^{\frac{1}{\beta}}\right),\end{aligned}$$

where $\mathsf{A}_1 = 2^{\frac{3\beta-1}{\beta}}\mathsf{A}_3$ and $\mathsf{A}_2 = 2^{\frac{2\beta+1}{\beta}}\mathsf{A}_4$.

Proof of Theorem 2. Fix $\boldsymbol{x} \in \Theta$. By Lemma 5 we have

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}) \leqslant \frac{a_t - \mathbf{E}\left[a_{t+1}|\boldsymbol{x}_t\right]}{2\eta_t} + \frac{1}{\alpha}\|\nabla f(\boldsymbol{x}_t) - \mathbf{E}\left[\boldsymbol{g}_t|\boldsymbol{x}_t\right]\|^2 + \frac{\eta_t}{2}\mathbf{E}\left[\|\boldsymbol{g}_t\|^2|\boldsymbol{x}_t\right] - \frac{\alpha}{4}a_t,$$

where $a_t = \|\boldsymbol{x}_t - \boldsymbol{x}\|^2$. Using Lemmas 2 and 3 and the assumption that $\max_{\boldsymbol{x}\in\Theta}\|\nabla f(\boldsymbol{x})\| \leqslant G$ we obtain

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}) \leqslant \frac{a_t - \mathbf{E}\left[a_{t+1}|\boldsymbol{x}_t\right]}{2\eta_t} + \frac{d}{\alpha}(\kappa_\beta L h_t^{\beta-1})^2 + \frac{3\eta_t}{4}\kappa d\left(\frac{3}{4}\left(d\bar{L}^2 h_t^2 + 8G^2\right) + \frac{\sigma^2}{h_t^2}\right) - \frac{\alpha}{4}a_t.$$

Let $b_t = \mathbf{E}\left[\|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right]$. Using the definition $\eta_t = 4/(\alpha(t+1))$ and taking the expectation we obtain

$$\mathbf{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x})\right] \leqslant \frac{\alpha}{4}\left(t\left(b_t - b_{t+1}\right) - b_t\right) + \frac{d}{\alpha}(\kappa_\beta L h_t^{\beta-1})^2 + \frac{3}{2\alpha(t+1)}\kappa d\left(\frac{3}{4}\left(d\bar{L}^2 h_t^2 + 8G^2\right) + \frac{\sigma^2}{h_t^2}\right).$$

Summing up both sides of this inequality form 1 to T and using the fact that

$$\sum_{t=1}^{T} t\left(\left(t+1\right)\left(b_t - b_{t+1}\right) - b_t\right) \leqslant 0$$

we find

$$\sum_{t=1}^{T} t\mathbf{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x})\right] \leqslant \frac{d}{\alpha}\sum_{t=1}^{T}\left(t(\kappa_\beta L h_t^{\beta-1})^2 + \frac{3t}{2(t+1)}\kappa\left(\frac{3}{4}\left(d\bar{L}^2 h_t^2 + 8G^2\right) + \frac{\sigma^2}{h_t^2}\right)\right).$$

Substituting here $h_t = (\frac{3}{2t}\frac{\kappa\sigma^2}{\kappa_\beta^2 L^2})^{\frac{1}{2\beta}}$ yields

$$\sum_{t=1}^{T} t\mathbf{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x})\right] \leqslant \frac{9G^2dT}{\alpha} + \frac{d}{\alpha}\sum_{t=1}^{T}\left(\mathsf{A}_3(L^2t)^{\frac{1}{\beta}} + \mathsf{A}_4 d\bar{L}^2\left(L^2t\right)^{-\frac{1}{\beta}}\right),$$

where $\mathsf{A}_3 = 2(3\kappa\sigma^2/2)^{\frac{\beta-1}{\beta}}\kappa_\beta^{\frac{2}{\beta}}$ and $\mathsf{A}_4 = (1/2)\,(3/2)^{\frac{2\beta+1}{\beta}}\,\kappa^{\frac{\beta+1}{\beta}}\,(\sigma/\kappa_\beta)^{\frac{2}{\beta}}$. Using the inequality $\sum_{t=1}^{T} t^{-\frac{1}{\beta}} \leqslant (\beta/(\beta-1))T^{\frac{\beta-1}{\beta}}$ we find

$$\sum_{t=1}^{T} t\mathbf{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x})\right] \leqslant \frac{9G^2dT}{\alpha} + \frac{d}{\alpha}\left(\mathsf{A}_3 L^{\frac{2}{\beta}} + \frac{\beta\mathsf{A}_4}{\beta-1}d\bar{L}^2(LT)^{-\frac{2}{\beta}}\right)T^{\frac{\beta+1}{\beta}}.$$

Dividing both sides of this inequality by $T(T+1)/2$ and applying Jensen's inequality we finally get

$$\mathbf{E}\left[f(\bar{\boldsymbol{x}}_T) - f(\boldsymbol{x})\right] \leqslant \frac{18G^2d}{\alpha T} + \frac{d}{\alpha}\left(\mathsf{A}_1 L^{\frac{2}{\beta}} + \mathsf{A}_2 d\bar{L}^2(LT)^{-\frac{2}{\beta}}\right)T^{-\frac{\beta-1}{\beta}},$$

where $\mathsf{A}_1 = 2\mathsf{A}_3$ and $\mathsf{A}_2 = 2\beta\mathsf{A}_4/(\beta-1)$.

FUNDING

The research of Arya Akhavan was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number EP/Y028333/1].

REFERENCES

1. Akhavan, A., Chzhen, E., Pontil, M., and Tsybakov, A.B., Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm, *J. Mach. Learn. Res.*, 2024, vol. 25, no. 370, pp. 1–50.
2. Akhavan, A., Pontil, M., and Tsybakov, A., Exploiting higher order smoothness in derivative-free optimization and continuous bandits, *Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 9017–9027.
3. Akhavan, A., Pontil, M., and Tsybakov, A., Distributed zero-order optimization under adversarial noise, *Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 10209–10220.
4. Bach, F. and Perchet, V., Highly-smooth zero-th order online optimization, *Proc. 29th Annu. Conf. Learn. Theory*, 2016, pp. 1–27.
5. Balasubramanian, K. and Ghadimi, S., Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points, *Found. Comput. Math.*, 2021, pp. 1–42.
6. Fabian, V., Stochastic approximation of minima with improved asymptotic speed, *Ann. Math. Stat.*, 1967, vol. 38, no. 1, pp. 191–200.
7. Fokkema, H., van der Hoeven, D., Lattimore, T., and Mayo, J.J., Online Newton method for bandit convex optimisation, 2024, https://arxiv.org/abs/2406.06506.
8. Gasnikov, A., Lagunovskaya, A., Usmanova, I., and Fedorenko, F., Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex, *Autom. Remote Control*, 2016, vol. 77, no. 11, pp. 2018–2034.
9. Gasnikov, A., Lobanov, A.V., and Stonyakin, F.S., Highly smooth zeroth-order methods for solving optimization problems under the PL condition, *Comput. Math. Math. Phys.*, 2024, vol. 64, pp. 739–770.
10. Ghadimi, S. and Lan, G., Stochastic first- and zeroth-order methods for nonconvex stochastic programming, *SIAM J. Optim.*, 2013, vol. 23, no. 4, pp. 2341–2368.
11. Granichin, O.N., A stochastic approximation procedure with disturbance at the input, *Autom. Remote Control*, 1992, vol. 53, no. 2, pp. 232–237.

12. Hastie, T. and Tibshirani, R., Generalized additive models, *Stat. Sci.*, 1986, vol. 1, no. 3, pp. 297–310.

13. Ibragimov, I.A. and Khas'minskii, R.Z., Bounds for the risks of non-parametric regression estimates, *Theory Probab. Appl.*, 1982, vol. 27, pp. 84–99.

14. Ibragimov, I.A. and Khas'minskii, R.Z., *Statistical Estimation: Asymptotic Theory*, Springer, 1981.

15. Jamieson, K.G., Nowak, R., and Recht, B., Query complexity of derivative-free optimization, *Adv. Neural Inf. Process. Syst.*, 2012, vol. 26, pp. 2672–2680.

16. Karimi, H., Nutini, J., and Schmidt, M., Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition, *Mach. Learn. Knowl. Discov. Databases*, 2016, pp. 795–811.

17. Kiefer, J. and Wolfowitz, J., Stochastic estimation of the maximum of a regression function, *Ann. Math. Stat.*, 1952, vol. 23, pp. 462–466.

18. Lojasiewicz, S., A topological property of real analytic subsets, *Coll. CNRS, Équ. Dériv. Partielles*, 1963, vol. 117, pp. 87–89.

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

20. Nesterov, Yu. and Spokoiny, V., Random gradient-free minimization of convex functions, *Found. Comput. Math.*, 2017, vol. 17, no. 2, pp. 527–566.

21. Novitskii, V. and Gasnikov, A., Improved exploitation of higher order smoothness in derivative-free optimization, *Optim. Lett.*, 2022, vol. 16, pp. 2059–2071.

22. Polyak, B.T., Gradient methods for minimizing functionals, *Comput. Math. Math. Phys.*, 1963, vol. 3, no. 4, pp. 864–878.

23. Polyak, B.T. and Tsybakov, A.B., Optimal order of accuracy of search algorithms in stochastic optimization, *Probl. Inf. Transm.*, 1990, vol. 26, no. 2, pp. 45–53.

24. Polyak, B.T. and Tsybakov, A.B., On stochastic approximation with arbitrary noise (the KW-case), *Adv. Sov. Math.*, Amer. Math. Soc., 1992, vol. 12, pp. 107–113.

25. Shamir, O., On the complexity of bandit and derivative-free stochastic convex optimization, *Proc. 30th Annu. Conf. Learn. Theory*, 2013, pp. 1–22.

26. Stone, C.J., Optimal rates of convergence for nonparametric estimators, *Ann. Stat.*, 1980, vol. 8, pp. 1348–1360.

27. Stone, C.J., Optimal global rates of convergence for nonparametric regression, *Ann. Stat.*, 1982, vol. 10, pp. 1040–1053.

28. Stone, C.J., Additive regression and other nonparametric models, *Ann. Stat.*, 1985, vol. 13, pp. 689–705.

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

30. Wood, S.N., *Generalized Additive Models*, Boca Raton, Chapman and Hall/CRC, 2017.

31. Yu, Q., Wang, Y., Huang, B., Lei, Q., and Lee, J.D., Stochastic zeroth-order optimization under strongly convexity and Lipschitz Hessian: Minimax sample complexity, *Adv. Neural Inf. Process. Syst.*, 2024, vol. 37, pp. 99564–99600.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board