Proceedings of Machine Learning Research vol 99:1–26, 2019 32nd Annual Conference on Learning Theory

Lower Bounds for Parallel and Randomized Convex Optimization

Jelena Diakonikolas JELENA.D@BERKELEY.EDU
University of California, Berkeley

Cristóbal Guzmán CRGUZMANP@MAT.UC.CL
Millennium Nucleus Center for the Discovery of Structures in Complex Data
Pontificia Universidad Católica de Chile

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study the question of whether parallelization in the exploration of the feasible set can be used to speed up convex optimization, in the local oracle model of computation. We show that the answer is negative for both deterministic and randomized algorithms applied to essentially any of the interesting geometries and nonsmooth, weakly-smooth, or smooth objective functions. In particular, we show that it is not possible to obtain a polylogarithmic (in the sequential complexity of the problem) number of parallel rounds with a polynomial (in the dimension) number of queries per round. In the majority of these settings and when the dimension of the space is polynomial in the inverse target accuracy, our lower bounds match the oracle complexity of sequential convex optimization, up to at most a logarithmic factor in the dimension, which makes them (nearly) tight. Prior to our work, lower bounds for parallel convex optimization algorithms were only known in a small fraction of the settings considered in this paper, mainly applying to Euclidean (ℓ_2) and ℓ_∞ spaces. Our work provides a more general and streamlined approach for proving lower bounds in the setting of parallel convex optimization.

Keywords: Lower bounds, convex optimization, parallel algorithms, randomized algorithms

1. Introduction

Given the scale of modern datasets resulting in extremely large problem instances, an attractive approach to reducing the time required for performing computational tasks is via parallelization. Indeed, many classical problems in computer science are well-known to be solvable in polylogarithmic number of rounds of parallel computation, with polynomially-bounded number of processors.

When it comes to convex optimization, parallelization is in general highly beneficial in computing local function information (at a single point from the feasible set), such as its gradient or Hessian, and can generally be exploited to improve the performance of optimization algorithms. However, a natural barrier for further speedups is parallelizing the exploration of the feasible set. This leads to the following question:

Is it possible to improve the oracle complexity of convex optimization via parallelization?

Here, oracle complexity is defined as the number of adaptive rounds an algorithm needs to query an arbitrary oracle providing local information about the function, such as, e.g., its value, gradient, Hessian, or a Taylor approximation at the queried point from the feasible set, before reaching a solution with a specified accuracy. Most of the commonly used optimization methods, such as, e.g., gradient descent, mirror descent, Newton's method, the ellipsoid method, Frank-Wolfe, and Nesterov's accelerated method, all work in this local oracle model.

The study of parallel oracle complexity of convex optimization was initiated by Nemirovski (1994). In this work, it was shown that for deterministic nonsmooth Lipschitz-continuous optimization over the ℓ_∞ ball, it is not possible to attain polylogarithmic parallel round complexity with polynomially many processors. Since the work of Nemirovski (1994) and until very recently, there has been no further progress in obtaining lower bounds for other settings, such as, e.g., the setting of randomized algorithms and weakly/strongly smooth optimization over more general feasible sets.

Very recently, motivated by the applications in online learning, local differential privacy, and adaptive data analysis, lower bounds for parallel convex optimization *over the Euclidean space* have been obtained in Smith et al. (2017); Balkanski and Singer (2018); Woodworth et al. (2018); Duchi et al. (2018). Our main result shows that it is not possible to improve the oracle complexity of convex optimization via parallelization, for deterministic or randomized algorithms, different levels of smoothness, and essentially all interesting geometries – general ℓ_p spaces for $p \in [1, \infty]$, together with their matrix spectral analogues, known as Schatten spaces, Sch_p. The resulting lower bounds are robust to enlargements of the feasible set, and thus apply in the unconstrained case as well. This is a much more general setting than previously addressed in the literature. The general ℓ_p settings considered in this paper are of fundamental interest. For example, ℓ_1-setups naturally appear in sparsity-oriented learning applications; Sch_1 (a.k.a. nuclear norm) appears in matrix completion problems Nesterov and Nemirovski (2013); finally, smooth ℓ_∞-setups have been used in the design of fast algorithms for network flow problems Lee et al. (2013); Kelner et al. (2014).

1.1. Our Results

Our results rule out the possibility of improvements by parallelization, showing that, in high dimensions, sequential methods are already optimal for any amount of parallelization that is polynomial in the dimension.[1] Our approach is to provide a generic lower bound for parallel oracle algorithms and use reductions between different classes of optimization problems. Below, $\varepsilon > 0$ is the target accuracy, K is the number of parallel queries per round, and d is the dimension.

Main Theorem *(Informal) Unless K is exponentially large in the dimension d, any (possibly randomized) algorithm working in the local oracle model and querying up to K points per round, when applied to the following classes of convex optimization problems over ℓ_p balls and Sch_p balls:*

- *Nonsmooth (Lipschitz-continuous) minimization for $1 < p < \infty$ and $d = \Omega(\mathrm{poly}(\frac{1}{\varepsilon^{p+p/(p-1)}}))$;*
- *Smooth (Lipschitz-continuous gradient) minimization for $2 \leq p \leq \infty$ and $d = \Omega(\mathrm{poly}(\frac{1}{\varepsilon}))$;*
- *Weakly-smooth (Hölder-continuous gradient) minimization for $2 \leq p \leq \infty$ and $d = \Omega(\mathrm{poly}(\frac{1}{\varepsilon}))$*

takes asymptotically at least as many rounds to reach an ε-approximate solution as it would take without any parallelization, up to, at most, a $1/\ln(d)$ factor.

As mentioned before, our result easily extends to unconstrained optimization over ℓ_p normed spaces. The small subset of the possible cases not included in the theorem are off by small factors and are

[1] Ruling out parallelization via an exponential number of queries is unlikely, since such a high number of queries would, in general, allow an algorithm to construct an ε-net of the feasible set and choose the best point from it.

Function class	$p = 1$	$1 < p < 2$	$2 \leq p < \infty$	$p = \infty$
Nonsmooth ($\kappa = 0$)	$\Omega\big(\frac{1}{\varepsilon^{2/3}}\big)$	$\Omega\big(\frac{1}{\varepsilon^{2}}\big)$	$\Omega\big(\frac{1}{\varepsilon^{p}}\big)$	$\Omega\Big(\big(\frac{\varepsilon^2 d}{\ln(dK/\gamma)}\big)^{1/3}\Big)$ $(*)$
Smooth ($\kappa = 1$)	$\Omega\Big(\frac{1}{\ln(d)\varepsilon^{2/5}}\Big)$	$\Omega\Big(\frac{1}{\ln(d)\varepsilon^{2/5}}\Big)$	$\Omega\Big(\frac{1}{\min\{p,\ln(d)\}\varepsilon^{p/(p+2)}}\Big)$	$\Omega\Big(\frac{1}{\ln(d)\varepsilon}\Big)$

Table 1: High probability lower bounds for parallel convex optimization, in the ℓ_p^d and Sch_p^d setups. Here, d is the dimension, ε is the accuracy, K is the number of parallel queries per round, and $1-\gamma$ is the confidence. Except for $(*)$, the high dimensional regime requires $d = \Omega(\mathrm{poly}(1/\varepsilon, \ln(K/\gamma)))$.

still very informative: they rule out the possibility of any significant improvement in the round complexity via parallelization (see Table 1 and the discussions in Sections 1.2 and 3).

To present the results in a unified manner, we use the definition of weakly-smooth functions, i.e., functions with κ-Hölder-continuous gradient, which interpolates between the classes of nonsmooth ($\kappa = 0$) and smooth functions ($\kappa = 1$) (see Section 1.4 for a precise definition). These two special cases are summarized in Table 1. For the precise statements encompassing the weakly-smooth cases ($\kappa \in (0, 1)$) as well as the specific *high-dimensional* regime for d, see Section 3.

The largest gap obtained by our results is in the nonsmooth ℓ_1-setup. Here, the $\Omega(1/\varepsilon^{2/3})$ bound comes from a reduction from the ℓ_∞ case, which explains the discontinuity in the first row of the table. We also consider a non-standard setting of ℓ_p-Lipschitz nonsmooth optimization for $p \in [1, 2)$ over an ℓ_2 ball inscribed in the unit ℓ_p ball. Even in this smaller domain the complexity is $\Omega(1/\varepsilon^2)$, which provides a strong evidence of higher complexity for the ℓ_1-setting.

1.2. Overview of the Techniques

Most of the lower bounds for large-scale convex optimization in the literature (e.g., Nemirovsky and Yudin (1983); Guzmán and Nemirovski (2015); Woodworth et al. (2018); Balkanski and Singer (2018)) are based on the construction of a hard problem instance defined as the maximum of affine functions. Each affine function f_i is defined by its direction vector $\mathbf{z}^i$ and offset δ_i. Examples of the direction vectors that are typically used in these works include signed orthant vectors, signed Hadamard bases, uniform vectors from the unit sphere and scaled Rademacher sequences. At an intuitive level, a careful choice of these affine functions prevents any algorithm from learning more than one direction vector $\mathbf{z}^i$ per adaptive round. At the same time, an appropriately chosen set of affine functions ensures that the algorithm needs to learn *all* of the vectors $\mathbf{z}^i$ before being able to construct an ε-approximate solution. We take the same approach in this paper.

Most relevant to our work are the recent lower bounds for parallel convex optimization over Euclidean (ℓ_2) spaces Balkanski and Singer (2018); Woodworth et al. (2018), which are tight in the large-scale regime. In these works, the argument about learning one vector $\mathbf{z}^i$ at a time is derived by an appropriate concentration inequality, while the upper bound on the optimal objective value is obtained from a good candidate solution, built as a combination of the random vectors. However, there is no obvious way of generalizing the lower bounds for the Euclidean setting to the more general ℓ_p geometries. For example, in the ℓ_p-setup for $p > 2$, these arguments only lead to a lower bound of $\Omega(1/\varepsilon^2)$, which is far from the sequential complexity $\Theta(1/\varepsilon^p)$. On the other hand, the use of relationships between the ℓ_p norms leads to uninformative lower bounds. In particular, for $p \in [1, 2)$, the appropriate application of inequalities relating ℓ_p norms needs to be done for both feasible sets (relating $\|\cdot\|_p$ and $\|\cdot\|_2$) and the Lipschitz (or smoothness) constants (relating $\|\cdot\|_{p^*}$

and $\|\cdot\|_2$, where $p^* = \frac{p}{p-1}$). Unless $p \approx 2$, this approach leads to a degradation in the lower bound by a polynomial factor in d. For example, if ℓ_2 case is used to infer a lower bound for the ℓ_1 setup, the resulting lower bound would be of the order $1/(d\varepsilon^2)$ and $1/(d\sqrt{\varepsilon})$, for nonsmooth and smooth cases, respectively. Such quantities are are far from the sequential lower bounds $\Omega(\ln(d)/\varepsilon^2)$ and $\Omega(1/\sqrt{\varepsilon})$ applying to the nonsmooth and smooth settings, respectively.

Our lower bounds are based on families of random vectors $\mathbf{z}^1, \ldots, \mathbf{z}^M$ which: (i) satisfy concentration along their marginals, so the "learning one vector per round" argument applies; and (ii) lead to a large negative optimal value via a minimax duality argument. Each particular regime will require different constructions of the random vectors, that we describe in Section 3. However, all lower bounds will be obtained from a general result (Theorem 2) that shows that (i) and (ii) suffice to get a lower bound for parallel convex optimization and completely streamlines the analysis.

1.3. Related Work

As mentioned earlier, until very recently, the literature on black-box parallel convex optimization was extremely scarce. Here we summarize the main lines of work.

Worst-case Lower Bounds for Sequential Convex Optimization. Classical theory of (sequential) oracle complexity in optimization was developed by Nemirovsky and Yudin (1983). This work provides sharp worst-case lower bounds for nonsmooth optimization, and a suboptimal (and rather technical) lower bound for randomized algorithms, for ℓ_p settings, where $1 \leq p \leq \infty$. Smooth convex optimization in this work is addressed by lower bounding the oracle complexity of convex quadratic optimization, which only applies to deterministic algorithms and the ℓ_2 setup. Nearly-tight lower bounds for deterministic non-Euclidean smooth convex optimization were obtained only recently Guzmán and Nemirovski (2015), mostly by the use of a smoothing of hard nonsmooth families. It is worth mentioning that none of these lower bounds are robust to parallelization. Further, prior to our work, there were no known lower bounds against sequential ($K = 1$) randomized algorithms in the general setting of weakly and strongly smooth minimization over ℓ_p spaces. The only exception is the lower bound for (strongly) smooth minimization over the Euclidean (ℓ_2) spaces, due to Woodworth and Srebro (2016).

Parallel Convex Optimization. The study of parallel oracle complexity in convex optimization was initiated in Nemirovski (1994), proving a worst-case lower bound $\Omega\big((\frac{d}{\ln(2Kd)})^{1/3}\ln(1/\varepsilon)\big)$ on the complexity in the ℓ_∞-setup. The argument from Nemirovski (1994) is based on a sequential use of the probabilistic method to generate the subgradients of a hard instance and applies to an arbitrary dimension beyond a fixed constant. The author conjectured that this lower bound is suboptimal, which still remains an open problem.

More recently, several lower bounds have been obtained for various settings of parallel convex optimization, but all applying only to *either box (ℓ_∞-ball) or ℓ_2-ball constrained Euclidean spaces*. In particular, Smith et al. (2017) showed that poly-log in $1/\varepsilon$ oracle complexity is not possible with polynomially-many in d parallel queries for nonsmooth Lipschitz-continuous minimization. This bound was further improved by Duchi et al. (2018), in the context of stochastic minimization with either Lipschitz-continuous or smooth and strongly convex objectives.

Tight lower bounds in the Euclidean setup have been obtained in Woodworth et al. (2018) and Balkanski and Singer (2018). Both of these works provide a tight lower bound $\Omega(1/\varepsilon^2)$ for randomized algorithms and nonsmooth Lipschitz objectives, when the dimension is sufficiently high (polynomial in $1/\varepsilon$, which is similar to our setting). The work in Balkanski and Singer (2018) further considers strongly convex Lipschitz objectives. While this setting is not considered in our work, we note that it is possible to incorporate it in our framework using the ideas from Srebro and Sridharan (2012). To obtain lower bounds that apply against randomized algorithms, Balkanski and Singer (2018) uses an intricate adaptivity argument. Our lower bound is based on a more direct application of the probabilistic method, and is arguably simpler.

The work in Woodworth et al. (2018) further considers an extension to stochastic and smooth objectives. However, the "statistical term" in Woodworth et al. (2018) comes from a typical minimax estimation bound, and its accuracy can, in fact, be reduced by parallelization at a rate $1/\sqrt{N}$, where N is the total number of queries. Their construction of subgradients for the hard function is based on random vectors from the unit sphere; our use of Rademacher sequences makes the analysis simpler and more broadly applicable. On the other hand, Woodworth et al. (2018) also provides lower bounds for (non-local) prox oracles, which are not considered in this paper.

1.4. Notation and Preliminaries

Vector Spaces and Classes of Functions. Let $(\mathbf{E}, \|\cdot\|)$ be a d-dimensional normed vector space, where $d < \infty$. We denote vectors in this space by bold letters, e.g., $\boldsymbol{x}, \boldsymbol{y}$, etc., and by $(\mathbf{E}^*, \|\cdot\|_*)$ its dual space. We use the bracket notation $\langle \mathbf{z}, \boldsymbol{x}\rangle$ to denote the evaluation of the linear functional $\mathbf{z} \in \mathbf{E}^*$ at a point $\boldsymbol{x} \in \mathbf{E}$; in particular, $\|\mathbf{z}\|_* = \sup_{\|x\|\leq 1} \langle \mathbf{z}, \boldsymbol{x}\rangle$. We denote the ball of $\mathbf{E}$ centered at $\boldsymbol{x}$ and of radius r by $\mathcal{B}_{\|\cdot\|}(\boldsymbol{x}, r)$, and the unit ball by $\mathcal{B}_{\|\cdot\|} := \mathcal{B}_{\|\cdot\|}(0,1)$. Our most important case of study is the space $\ell_p^d = (\mathbb{R}^d, \|\cdot\|_p)$, where $1 \leq p \leq \infty$. For simplicity, in this case we use the notation $\mathcal{B}_p^d(\boldsymbol{x}, r) := \mathcal{B}_{\|\cdot\|_p}(\boldsymbol{x}, r)$. The dual space of ℓ_p^d is isometrically isomorphic to $\ell_{p^*}^d$, where $p^* = p/(p-1)$; in this case, the bracket is just the standard inner product in $\mathbb{R}^d$. Other important example is the case of Schatten spaces: $\text{Sch}_p^d = (\mathbb{R}^{d\times d}, \|\cdot\|_{\text{Sch},p})$. Here, for any $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, $\|\boldsymbol{X}\|_{\text{Sch},p} = (\sum_{i=1}^d \sigma_i(\boldsymbol{X})^p)^{1/p}$, where $\sigma_1(\boldsymbol{X}), \ldots, \sigma_d(\boldsymbol{X})$ are the singular values of $\boldsymbol{X}$.

Given $\kappa \geq 0$, we use $\mathcal{F}^{\kappa}_{(\mathbf{E},\|\cdot\|)}(\mu)$ to denote the class of convex functions $f : \mathbf{E} \to \mathbb{R}$ such that

$$\left\|D^{\lfloor \kappa+1\rfloor} f(\boldsymbol{y}) - D^{\lfloor \kappa+1\rfloor}(\boldsymbol{x})\right\|_{\text{op}} \leq \mu \|\boldsymbol{y} - \boldsymbol{x}\|^{\kappa+1-\lfloor \kappa+1\rfloor} \qquad (\forall \boldsymbol{x},\, \boldsymbol{y} \in \mathbf{E}), \tag{1}$$

where D^t is the t^{th} derivative operator and $\|A\|_{\text{op}} := \sup_{\|\boldsymbol{h}\|\leq 1} |A[\boldsymbol{h}; \ldots; \boldsymbol{h}]|$ is the induced operator norm on symmetric multilinear forms w.r.t. $\|\cdot\|$.

To clarify this definition, let us provide some useful examples: (i) $\kappa = 0$ corresponds to bounded variation of subgradients, $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_* \leq \mu$. This class contains all $\mu/2$-Lipschitz convex functions, but is also invariant under affine perturbations;[2] (ii) $\kappa \in (0,1)$ corresponds to Hölder continuous gradients, $\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_* \leq \mu \|\boldsymbol{y} - \boldsymbol{x}\|^{\kappa}$; (iii) $\kappa = 1$ corresponds to Lipschitz-continuity of the gradient, $\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_* \leq \mu \|\boldsymbol{y} - \boldsymbol{x}\|$; and (iv) $\kappa = 2$ corresponds to Lipschitz-continuous Hessian, $\|Hf(\boldsymbol{y}) - Hf(\boldsymbol{x})\|_{\text{op}} \leq \mu \|\boldsymbol{y} - \boldsymbol{x}\|$.

Optimization Problems, Algorithms, and Oracles. We consider convex programs of the form

$$\min\{f(\boldsymbol{x}) : \ \boldsymbol{x} \in \mathcal{X}\},$$

where $f : \mathbf{E} \to \mathbb{R}$ is a convex function from a given class of objectives $\mathcal{F}$ (such as the ones described above), and $\mathcal{X} \subseteq \mathbf{E}$ is convex and closed. We denote by f^* the optimal value of the problem. Our goal is, given an accuracy $\varepsilon > 0$, to find an ε-solution; i.e., an $\boldsymbol{x} \in \mathcal{X}$ such that $f(\boldsymbol{x}) - f^* \leq \varepsilon$.

[2] Our lower bounds for nonsmooth optimization are in fact given by classes of Lipschitz convex functions, but to keep the notation unified we use (1) instead.

We study complexity of convex optimization in the oracle model of computation. In this model, the algorithm queries points from the feasible set $\mathcal{X}$, and it obtains partial information about the objective via a *local oracle* $\mathcal{O}$. Given objective $f \in \mathcal{F}$, and a query $\boldsymbol{x} \in \mathcal{X}$, we denote the oracle answer by $\mathcal{O}_f(\boldsymbol{x})$ (when f is clear from the context we omit it from the notation). We say that an oracle $\mathcal{O}$ is local if given two functions $f, g : \mathbf{E} \to \mathbb{R}$ such that $f \equiv g$ in the neighborhood of some point $\boldsymbol{x} \in \mathcal{X}$, it must be that $\mathcal{O}_f(\boldsymbol{x}) = \mathcal{O}_g(\boldsymbol{x})$. Notable examples of local oracles are the gradient over the class $\mathcal{F}^{\kappa}_{\|\cdot\|}(\mu)$, with $\kappa > 0$;[3] and a κ^{th}-order Taylor expansion over the class $\mathcal{F}^{\kappa}_{\|\cdot\|}(\mu)$, with κ being a non-negative integer.

In the K-parallel setting of convex optimization Nemirovski (1994), an algorithm works in rounds. At every round, it performs a batch of queries $X^t = \{\boldsymbol{x}^t_1, \ldots, \boldsymbol{x}^t_K\}$, for $\boldsymbol{x}^t_k \in \mathcal{X}$ $(\forall k \in [K])$, where we have used the shorthand notation $k \in [K]$ to denote $k \in \{1, \ldots, K\}$. Given the queries, the local oracle $\mathcal{O}$ replies with a batch of answers: $\mathcal{O}_f(X^t) := (\mathcal{O}_f(\boldsymbol{x}^t_1), \ldots, \mathcal{O}_f(\boldsymbol{x}^t_K))$.

The algorithm may work adaptively over rounds: every batch of queries may depend on queries and answers from previous rounds:

$$X^{t+1} = U^{t+1}(X^1, \mathcal{O}_f(X^1), \ldots, X^t, \mathcal{O}_f(X^t)) \qquad (\forall t \geq 1), \tag{2}$$

where the first round of queries $X^1 = U^1(\emptyset)$ is instance-independent (the algorithm has no specific information about f at the beginning). Functions $(U^t)_{t\geq 1}$, may be deterministic or randomized, and this would characterize the deterministic or randomized nature of the algorithm. We are interested in the effect of parallelization on the complexity of convex optimization in the described oracle model. Notice that $K = 1$ corresponds to the traditional notion of (sequential) oracle complexity.

Notion of Complexity. Let $\mathcal{O}$ be a local oracle for a class of functions $\mathcal{F}$, and let $\mathcal{A}^K(\mathcal{O})$ be the class of K-parallel deterministic algorithms interacting with oracle $\mathcal{O}$. Given $\varepsilon > 0$, $f \in \mathcal{F}$, and $A \in \mathcal{A}^K(\mathcal{O})$, define the running time $T(A, f, \varepsilon)$ as the minimum number of rounds before algorithm A finds an ε-solution. The notion of complexity used in this work is known as the *high probability* complexity, defined as:

$$\text{Compl}^{\gamma}_{\text{HP}}(\mathcal{F}, \mathcal{X}, K, \varepsilon) = \sup_{F \in \Delta(\mathcal{F})} \inf_{A \in \mathcal{A}^K(\mathcal{O})} \inf\{\tau : \mathbb{P}_{f\sim F}[T(A, f, \varepsilon) \leq \tau] \geq \gamma\},$$

where $\gamma \in (0, 1)$ is a confidence parameter and $\Delta(\mathcal{F})$ is the set of probability distributions over the class of functions $\mathcal{F}$. The high probability complexity subsumes other well-known notions of complexity, including distributional, randomized, and worst-case, in the local oracle model. More details about the relationship between these different notions of complexity are provided in Appendix A.1 and can also be found in Braun et al. (2017).

Additional Background. Additional background and statements of several useful definitions and facts that are important for our analysis are provided in Appendix A.

1.5. Organization of the Paper

Next section provides a general lower bound that is the technical backbone of all the results in this paper. Section 3 then overviews the applications of this result in the general ℓ_p setups. Omitted proofs from Sections 2 and 3 are provided in Appendices B and C, respectively. We conclude in Section 4 with a discussion of obtained results and directions for future work.

[3] When $\kappa = 0$, not every subgradient oracle is local. However, this is a reasonable assumption for black-box algorithms (e.g, when we cannot access a dual formulation, or a smoothing of the objective).

2. General Complexity Bound

To prove the claimed complexity results from the introduction, we will work with a suitably chosen class of random nonsmooth Lipschitz-continuous problem instances. The results for the classes of problems with higher order of smoothness will be established (mostly) through the use of smoothing maps. In particular, we will make use of the following definition of locally smoothable spaces:

Definition 1 *A space $(\mathbf{E}, \|\cdot\|)$ is (κ, η, r, μ)-*locally smoothable *if there exists a mapping*

$$\begin{array}{cccc} \mathcal{S}: & \mathcal{F}^0_{(\mathbf{E},\|\cdot\|)}(1) & \to & \mathcal{F}^\kappa_{(\mathbf{E},\|\cdot\|)}(\mu) \\ & f & \mapsto & \mathcal{S}f \end{array},$$

referred to as the local smoothing, such that: (i) $\|f - \mathcal{S}f\|_\infty \leq \eta$; and (ii) if $f, g \in \mathcal{F}^0_{\|\cdot\|}(1)$ and $\boldsymbol{x} \in \mathbf{E}$ are such that $f|_{B_{\|\cdot\|}(\boldsymbol{x},2r)} \equiv g|_{B_{\|\cdot\|}(\boldsymbol{x},2r)}$ then $\mathcal{S}f|_{\mathcal{B}_{\|\cdot\|}(\boldsymbol{x},r)} \equiv \mathcal{S}g|_{\mathcal{B}_{\|\cdot\|}(\boldsymbol{x},r)}$.

Namely, a space is (κ, η, r, μ)-locally smoothable if there exists a mapping $\mathcal{S}$ that maps all nonsmooth functions to functions in $\mathcal{F}^\kappa_{\|\cdot\|}(\mu)$, such that a function f and its map $\mathcal{S}f$ do not differ by more than η when evaluated at any point from the space, and the map preserves the equivalence of functions over sufficiently small neighborhoods of points from the space. This last property is crucial to argue about the behavior of a local oracle.

The following theorem is the backbone of all the results from this paper: all complexity bounds will be obtained as its applications.

Theorem 2 *Let $(\mathbf{E}, \|\cdot\|)$ be a normed space and $\mathcal{X} \supseteq \mathcal{B}_{\|\cdot\|}$ be a closed and convex subset of $\mathbf{E}$. Suppose there exist a positive integer M, independent random vectors $\mathbf{z}^1, \ldots, \mathbf{z}^M$ supported on $\mathcal{B}_{\|\cdot\|^*}$, $\varepsilon > 0$, $\alpha > 0$, and $0 < \gamma < 1/2$, such that, if we define $\bar{\delta} = 16\sqrt{\frac{\ln(MK/\gamma)}{\alpha}}$, we have:*

- *(a) $(\mathbf{E}, \|\cdot\|)$ is (κ, η, r, μ)-locally smoothable, with $\mu > 0$, $0 < r \leq \bar{\delta}/8$, and $\eta \leq \varepsilon\mu/4$;*
- *(b) $\mathbb{P}\big[\inf_{\boldsymbol{\lambda}\in\Delta_M} \big\|\sum_{i\in[M]} \lambda_i \mathbf{z}^i\big\|_* \leq 4\mu\varepsilon\big] \leq \gamma$;*
- *(c) For any $i \in [M]$, $\boldsymbol{x} \in \mathcal{B}_{\|\cdot\|}$, and $\delta > 0$*

$$\mathbb{P}[\langle \mathbf{z}^i, \boldsymbol{x}\rangle \geq \delta] \leq \exp\{-\alpha\delta^2\} \quad \text{and} \quad \mathbb{P}[\langle \mathbf{z}^i, \boldsymbol{x}\rangle \leq -\delta] \leq \exp\{-\alpha\delta^2\};$$

- *(d) $\bar{\delta} \leq \mu\varepsilon/M$.*

Then, the high probability complexity of class $\mathcal{F}^\kappa_{(\mathbf{E},\|\cdot\|)}(1)$ on $\mathcal{X}$ satisfies

$$\mathrm{Compl}^{2\gamma}_{\mathrm{HP}}(\mathcal{F}^\kappa_{(\mathbf{E},\|\cdot\|)}(1), \mathcal{X}, K, \varepsilon) \geq M.$$

Remark 3 *Theorem 2 is stated for domains containing the unit ball and function class $\mathcal{F}^\kappa_{(\mathbf{E},\|\cdot\|)}(1)$. Handling arbitrary radius $R > 0$ and regularity constant μ can be achieved by a simple rescaling and change of variables, which we omit for space considerations. The result is that if the lower bound for $R = \mu = 1$ is $M(\varepsilon)$, then the lower bound for arbitrary $R, \mu > 0$ would be $M(\varepsilon/(\mu R^{\kappa+1}))$.*

Remark 4 *Even though Theorem 2 is stated for the standard setting, in which $\mathcal{X}$ contains the unit ball w.r.t. the norm of the space, $\|\cdot\|$, it is possible to extend it in a generic way to non-standard settings in which these two norms do not agree. For an example of such a setting, see Theorem 9.*

To prove Theorem 2, we need to build a distribution over $\mathcal{F}^{\kappa}_{(\mathbf{E},\|\cdot\|)}(1)$ such that any K-parallel deterministic algorithm interacting with a local oracle on $\mathcal{X}$ needs M rounds to reach an ε solution, with probability $1-2\gamma$. We propose a family of objectives as follows. Given $\mathbf{z}^1,\ldots,\mathbf{z}^M$ as in the theorem, consider the problem (P) $\min\{F(\boldsymbol{x}) : \boldsymbol{x}\in\mathcal{X}\}$, where:

$$F(\boldsymbol{x}) := \frac{1}{\mu}\mathcal{S}\Big(\max\Big\{\frac{1}{2}\max_{i\in[M]}\big[\langle\mathbf{z}^i,\cdot\rangle - i\bar{\delta}\big],\ \|\cdot\| - \frac{1}{2}(3(1+r)+M\bar{\delta})\Big\}\Big)(\boldsymbol{x}), \tag{3}$$

By construction, $F\in\mathcal{F}^{\kappa}_{\|\cdot\|}(1)$ surely. Observe that, since $\|\mathbf{z}^i\|_*\leq 1$, for all i:

- (O_1) When $\|\boldsymbol{x}\|\leq 1+2r$, it must be $1/2\max_{i\in[M]}\big[\langle\mathbf{z}^i,\boldsymbol{x}\rangle - i\bar{\delta}\big]\geq\|\boldsymbol{x}\|-1/2(3(1+r)+M\bar{\delta})$; i.e., within the unit ball, F is only determined by its left term (and not the norm term).
- (O_2) When $\|\boldsymbol{x}\|\geq 3(1+r)+(M-1)\bar{\delta}$, it must be $1/2\max_{i\in[M]}\big[\langle\mathbf{z}^i,\boldsymbol{x}\rangle - i\bar{\delta}\big]\leq\|\boldsymbol{x}\|-1/2(3(1+r)-M\bar{\delta})$; i.e., outside the ball of radius $3(1+r)+(M-1)\bar{\delta}\leq 4$,[4] F is only determined by the norm term (and not by $\mathbf{z}^1,\ldots,\mathbf{z}^M$).

We claim that any K-parallel deterministic algorithm that works in M rounds, with probability $1-2\gamma$, will fail to query a point with optimality gap less than ε. This suffices to prove the theorem. The proof consists of three main parts: (i) establishing an upper bound on the minimum value F^* of (3), which holds with probability $1-\gamma$, (ii) establishing a lower bound on the value of the algorithm's output $\min\{F(\boldsymbol{x}) : \boldsymbol{x}\in\bigcup_{t\in[M]}X^t\}$, which holds with probability $1-\gamma$, and (iii) combining the first two parts to show that the optimality gap $\min\{F(\boldsymbol{x})-F^* : \boldsymbol{x}\in\bigcup_{t\in[M]}X^t\}$ of the best solution found by the algorithm after M rounds is higher than ε, with probability $1-2\gamma$. The full proof is provided in Appendix B.

3. Lower Bounds for Parallel Convex Optimization over ℓ_p Balls

In this section, we show how the general complexity bound from Theorem 2 can be applied to obtain several lower bounds for parallel convex optimization. Our main case of study will be ℓ_p^d spaces.

Remark 5 *In what follows, we will prove several lower bounds for ℓ_p-setups. Interestingly, we can obtain analog lower bounds for Schatten spaces. This can be obtained by simply noting that the restriction of the Schatten norm to diagonal matrices coincides with $\|\cdot\|_p$, and therefore we can embed $\mathcal{B}_p^d$, as well as $\mathcal{F}^{\kappa}_{\ell_p^d}(1)$ through this restriction (for more details, we refer the reader to Guzmán and Nemirovski (2015)). This embedding has a quadratic cost in the large-scale regime; in particular, it remains polynomial in $1/\varepsilon$ and $\ln(K/\gamma)$.*

3.1. Nonsmooth Optimization

To apply Theorem 2 in the nonsmooth case, we do not need to apply any smoothing at all. This is formally stated as "any normed space is $(0,0,0,1)$-locally smoothable," and its consequence is that Property (a) of the theorem is automatically satisfied. Thus, it suffices to construct a probability distribution over $\mathbf{z}^i$'s that under suitable constraints on α and the number of rounds M satisfies Assumptions (b) and (c) from the theorem. Assumption (d) simply constrains M by $M\leq\frac{\varepsilon}{\delta}$.

[4] From (b) we may assume that $4\mu\varepsilon\leq 1$, and then using the bounds on r and $\bar{\delta}$ from (a) and (d), we get the bound.

Let $\mathbf{r}^i$ denote an independent (over i) d-dimensional vector of independent Rademacher entries (i.e., a vector whose entries take values ± 1 w.p. $1/2$, independently of each other). Let $\mathbf{I}_L^i$ denote the $d \times d$ diagonal matrix, whose $L \leq d$ diagonal entries take value 1, while the remaining entries are zero. The positions of the non-zero entries on the diagonal of $\mathbf{I}_L^i$ will, in general, depend on i, and will be specified later. Given $p \geq 1$, vectors $\mathbf{z}^i \in \mathcal{B}_{p^*}^d$ are then defined as:

$$\mathbf{z}^i = \frac{1}{L^{1/p^*}} \mathbf{I}_L^i \mathbf{r}^i. \tag{4}$$

3.1.1. Bounds for $1 \leq p \leq 2$

When $p \in [1, 2]$, it suffices to choose $L = d$, so that $\mathbf{z}^i = d^{-1/p^*}\mathbf{r}^i$. We start by proving a lower bound that applies in the regime when $d = \Omega(\mathrm{poly}(\log(K/\gamma), 1/\varepsilon^{p^*}))$. Hence the bound deteriorates as p tends to one, and, in particular, does not apply to the case when $p = 1$. However, we will also show that it is possible to derive a lower bound for a restricted feasible set: the lower bound will apply to Lipschitz-continuous nonsmooth minimization over an ℓ_2 ball inscribed in the unit ℓ_p ball and it will apply in the regime of $d = \Omega(\mathrm{poly}(\log(K/\gamma), 1/\varepsilon))$. This provides a strong indication that obtaining speedups from parallelizing convex optimization is not any easier when p is close to 1 than in other regimes of p. The following lemma gives a sufficient condition for assumption (b) from Theorem 2 to hold. Its proof is provided in Appendix C.

Lemma 6 *Let $1 < p \leq 2$ and let $\mathbf{z}^1, \ldots, \mathbf{z}^M$ be chosen according to Eq. (4), where*

$$M \leq \min\left\{\frac{1}{200\varepsilon^2}, \frac{d/12 - \ln(1/\gamma)}{\ln(3/\varepsilon)}\right\},$$

then for all $\gamma \in (1/\mathrm{poly}(d), 1) : \mathbb{P}[\min_{\boldsymbol{\lambda} \in \Delta_M} \|\sum_{i \in [M]} \lambda_i \mathbf{z}^i\|_{p^*} \leq 4\varepsilon] \leq \gamma$.

To obtain the claimed lower bound for the nonsmooth case, we only need to establish the concentration of inner products within the feasible domain. When $p > 1$, this is obtained as a simple application of Hoeffding's Inequality. These two facts provide the claimed lower bound.

Theorem 7 *Let $1 < p \leq 2$ and $\mathcal{X} \supseteq \mathcal{B}_p^d$. Let $\varepsilon \in (0, 1/2)$ and $\gamma \in (1/\mathrm{poly}(d), 1)$. Then:*

$$\mathrm{Compl}_{\mathrm{HP}}^{\gamma}(\mathcal{F}_{\ell_p^d}^0(1), \mathcal{X}, K, \varepsilon) \geq M := \min\left\{\frac{1}{200\varepsilon^2},\ \frac{\varepsilon d^{1/p^*}}{32\sqrt{\ln(MK/\gamma)}}\right\}.$$

Proof We verify the conditions of Theorem 2. Recall that in the nonsmooth case condition (a) is automatically satisfied. For (b), by a direct application of Hoeffding's Inequality, for all $x \in \mathcal{B}_p^d$

$$\mathbb{P}[\langle \mathbf{z}^i, \boldsymbol{x}\rangle > \delta] = \mathbb{P}[\langle \mathbf{r}^i, \boldsymbol{x}\rangle > d^{1/p^*}\delta] \leq \exp\{-d^{2/p^*}\delta^2\}.$$

In particular, we have that $\alpha = d^{2/p^*}$ suffices to satisfy (b). Property (c) is obtained from Lemma 6, which requires bounding M according to the lemma. Assumption (d) holds as long as $M \leq \varepsilon/\bar{\delta}$. As $\bar{\delta} = 16\sqrt{\frac{\ln(MK/\gamma)}{\alpha}}$, it is sufficient to require: $M \leq \frac{\varepsilon d^{1/p^*}}{32} \frac{1}{\sqrt{\ln(MK/\gamma)}}$. ■

Remark 8 *Even though M is implicitly defined in Theorem 7, an explicit definition for M can be obtained by using a looser bound $\ln(dK/\gamma)$ instead of $\ln(MK/\gamma)$. We keep this definition to highlight the large scale regime for d. In particular, the high-dimensional regime is determined by solving for d the inequality $\frac{\varepsilon d^{1/p^*}}{32\sqrt{\ln(MK/\gamma)}} \geq M$, where $M = \frac{1}{200\varepsilon^2}$.*

We can conclude from Theorem 7 that as long as d is "sufficiently large" (namely, as long as $d = \Omega(\big(\sqrt{\ln(K/(\varepsilon\gamma))}/\varepsilon^3\big)^{p^*}))$, any ε-approximate K-parallel algorithm takes $\Omega(1/\varepsilon^2)$ iterations, which is asymptotically optimal – this bound is tight in the sequential case (when $K = 1$) and is thus unimprovable Nemirovsky and Yudin (1983). Unfortunately, this lower bound becomes uninformative when $p^* = \Omega(\ln d)$; in particular, when $p = 1$.

A Lower Bound for a Nonstandard Setting. As we mention above, none of the techniques of this paper is able to provide a $\Omega(1/\varepsilon^2)$ lower bound for the nonsmooth ℓ_1-Lipschitz optimization over a unit ℓ_1 ball. However, we can show a slightly weaker result: Namely, that ℓ_1-Lipschitz convex optimization over a subset of the $\mathcal{B}_1^d$-ball has parallel complexity $\Omega(1/\varepsilon^2)$. In fact, this result holds more generally for ℓ_p-Lipschitz convex optimization, where $p \in [1, 2]$,[5] over an ℓ_2 ball inscribed in the unit ℓ_p ball. The proof is provided in Appendix C.

Theorem 9 *Let $\varepsilon \in (0, 1/2)$, $\gamma \in (1/\mathrm{poly}(d), 1)$, and $p \in [1, 2]$. Then:*

$$\mathrm{Compl}_{\mathrm{HP}}^{\gamma}(\mathcal{F}^0_{\ell_p^d}(1), \mathcal{B}_2^d(1/d^{1/p-1/2}), K, \varepsilon) \geq M := \min\Big\{\frac{1}{200\varepsilon^2},\ \frac{\varepsilon d^{1/2}}{32\sqrt{\ln(MK/\gamma)}}\Big\}.$$

3.1.2. Bounds for $p \geq 2$

It is possible to extend Lemma 6 to the case of $p \geq 2$. However, due to the upper bound on M from Lemma 6, the best dimension-independent lower bound on the number of queries we could obtain in this setting would be of the order $1/\varepsilon^2$. Given that in the sequential setting the best dimension-independent lower bound is $\Omega(1/\varepsilon^p)$, we need a stronger result than what we obtained in Lemma 6.

This is achieved through a different construction of $\mathbf{z}^i$'s, where these vectors are no longer supported on all d coordinates, but only on $L < d$ of them; moreover, we will choose their supports to be disjoint. The construction is as follows. Let $\{J_i\}_{i=1}^M$ be a collection of subsets of $\{1, \dots, d\}$ such that $|J_i| = L$ and $J_i \cap J_{i'} = \emptyset$, $\forall i \neq i'$ (here, we assume that $d \geq ML$). Set $\mathbf{I}_L^i = \mathrm{diag}(\mathbb{1}_{J_i})$, i.e., the (j, j) element of the diagonal matrix $\mathbf{I}_L^i$ is 1 if $j \in J_i$ and 0 otherwise. As before (see (4)), $\mathbf{z}^i$ is defined as $\mathbf{z}^i = \frac{1}{L^{1/p^*}}\mathbf{I}_L^i \mathbf{r}^i$, where $(r_j^i)_{i\in[M], j\in[d]}$ is an independent Rademacher sequence.

Our next result addresses the nonsmooth $p \geq 2$ case, by a direct application of Theorem 2 to our construction above. More details are provided in Appendix C.

Theorem 10 *Let $p \geq 2$, $\mathcal{X} \supseteq \mathcal{B}_p^d$, and $\varepsilon \in (0, 1/2)$, $\gamma \in (1/\mathrm{poly}(d), 1)$. Then:*

$$\mathrm{Compl}_{\mathrm{HP}}^{\gamma}(\mathcal{F}^{\kappa}_{\ell_p^d}(1), \mathcal{X}, K, \varepsilon) \geq M := \min\Big\{\frac{1}{(4\varepsilon)^p},\ \frac{\varepsilon^{2/3}}{8}\Big(\frac{d}{\ln(MK/\gamma)}\Big)^{1/3}\Big\}.$$

In particular we have that the required number of queries to reach an ε-approximate solution is $\Omega(\frac{1}{\varepsilon^p})$, as long as $d = \Omega(\frac{\ln(K/\gamma)+p\ln(1/\varepsilon)}{\varepsilon^{3p+2}})$. When $p \to \infty$, the right term in the definition of M dominates, and we have $M = \Omega\Big(\varepsilon^{2/3}\big(\frac{d}{\ln(dK/\gamma)}\big)^{1/3}\Big)$, which, for constant ε, matches the best known bound for deterministic algorithms in this setting, due to Nemirovski (1994).

[5] When $p = 2$, the inscribed ℓ_2 ball is exactly the unit ℓ_p ball.

3.2. Smooth and Weakly Smooth Optimization

To apply Theorem 2 and obtain lower bounds for (weakly) smooth classes of functions, we need to design an appropriate local smoothing. This is indeed possible for $p \geq 2$, as we show below.

Remark 11 *Here we list some known local smoothings from the literature.*

1. *Let $2 \leq p \leq \infty$, $d \in \mathbb{N}$, and $0 \leq \kappa \leq 1$. Then, for any $\eta > 0$, the space $\ell_p^d = (\mathbb{R}^d, \|\cdot\|_p)$ is $(\kappa, \eta, \eta, \mu)$-locally smoothable when $\mu = 2^{1-\kappa}(\min\{p, \ln d\}/\eta)^{\kappa}$. We prove this in the Appendix A, following Guzmán and Nemirovski (2015).*
2. *Let $d, \kappa \in \mathbb{N}$ and $\eta > 0$. Then ℓ_2^d is $(\kappa, \kappa\eta, \kappa\eta, (d/\eta)^{\kappa})$-locally smoothable. This is achieved by a sequential integral convolution w.r.t. the uniform kernel on the ball of radius η Agarwal and Hazan (2018). They also show that for $1 \leq L \leq d$, the restriction of $\mathcal{S}$ to the set:*

$$\big\{f : \mathbb{R}^d \to \mathbb{R} :\; f \in \mathcal{F}^0_{\ell_2^d}(1),\; (\exists\, \Gamma \textit{ subspace of dim. } L)\, (\forall \boldsymbol{y} \in \Gamma^{\perp})\; f(\boldsymbol{x}) = f(\boldsymbol{x} + \boldsymbol{y})\big\},$$

satisfies an improved $(\kappa, \kappa\eta, \kappa\eta, (L/\eta)^{\kappa})$ local smoothing property.

Our next result addresses the smooth ℓ_p^d-setup when $p \geq 2$. Its proof is provided in Appendix C.

Theorem 12 *Let $p \geq 2$, $\mathcal{X} \supseteq \mathcal{B}_p^d$, and $\varepsilon \in (0, 1/2)$, $\gamma \in (1/\mathrm{poly}(d), 1)$. Then:*

$$\mathrm{Compl}^{\gamma}_{\mathrm{HP}}(\mathcal{F}^{\kappa}_{\ell_p^d}(1), \mathcal{X}, K, \varepsilon) \geq M := \min\left\{ \left(\frac{1}{2^{3+4\kappa}\, \varepsilon\, (\min\{p, \ln(d)\})^{\kappa}}\right)^{\frac{p}{1+\kappa(1+p)}}, \right.$$
$$\left. \frac{d}{2^9 \ln(MK/\gamma)} \left(2^{\frac{1+3p+2\kappa(1+p)}{1+p}} \min\{p, \ln(d)\}^{\kappa} \varepsilon\right)^{\frac{2(1+p)}{1+\kappa(1+p)}} \right\}.$$

The bound from Theorem 12 may be difficult to read, so let us point out a few notable special cases:

- When $\kappa = 0$, $p \to \infty$, the bound is uninformative, and one should instead use Theorem 10. This is a consequence of the particular choice of L in the proof, and its dependence on κ.
- When $\kappa \in (0, 1]$, $p \to \infty$, if $d = \Omega\left((\ln(\frac{K}{\gamma}) + \frac{1}{\kappa}\ln(\frac{1}{\varepsilon}))(\frac{1}{\varepsilon})^{\frac{3}{\kappa}}\right)$, then $M = \frac{1}{\ln(d)}(\frac{1}{2^{3+4\kappa}\varepsilon})^{1/\kappa}$, which is tight up to a factor $\frac{1}{\ln(d)}$ and achieved for $K = 1$ by Frank and Wolfe (1956) method.
- When $\kappa = 0$, $p < \infty$, and $d = \Omega\left((\ln(K/\gamma) + p\ln(1/\varepsilon))(\frac{1}{\varepsilon})^{3p+2}\right)$, then $M = (\frac{1}{8\varepsilon})^p$, which is achieved for $K = 1$ by the Mirror-Descent method Nemirovsky and Yudin (1983).
- When $\kappa = 1$, $p < \infty$, and $d = \Omega\left(\max\{(\ln(K/\gamma) + \ln(1/\varepsilon))(\frac{1}{\varepsilon})^3, \exp(p)\}\right)$, then $M = (\frac{1}{128p\varepsilon})^{\frac{p}{p+2}}$. These bounds are unimprovable and are achieved for $K = 1$ by the Nemirovski-Nesterov accelerated method Nemirovskii and Nesterov (1985); d'Aspremont et al. (2018).

Remark 13 *The proof strategy of Theorem 12 for $p = 2$ can also be used to obtain lower bounds for higher-order smooth convex optimization, following Agarwal and Hazan (2018). Namely, using the sequential integral convolution smoothing from Remark 11, we can obtain analog lower bounds as in Agarwal and Hazan (2018), that also apply to parallel algorithms. We defer the details of this simple corollary to the full version of the paper.*

Unfortunately, the smoothing approach is not immediately applicable when $1 \le p < 2$, due to the fact that there are no known regularizers for an infimal convolution smoothing. This is related to the fact that these spaces are not 2-uniformly smooth (see, e.g., Ball et al. (1994)) which leads to a natural barrier for the approach. However, this difficulty has been circumvented by Guzmán and Nemirovski (2015), where lower bounds in this regime are shown by a reduction from the $p = \infty$ case, specifically through a linear embedding of problem classes. We follow the same approach, and for the sake of brevity, we only provide a proof sketch in Appendix C.

Theorem 14 *Let $1 \le p < 2$, $0 < \kappa \le 1$, $\mathcal{X} \supseteq \mathcal{B}_p^d$, $\varepsilon \in (0, 1/2)$, $\gamma \in (1/\text{poly}(d), 1)$. Then, there exist constants $\nu, c(\kappa) > 0$, such that if $d \ge \frac{1}{\nu}\left\lceil 2(\ln(\nu dK/\gamma))^{\frac{2\kappa}{3+2\kappa}} \left(\frac{1}{\varepsilon}\right)^{\frac{6}{3+2\kappa}} \right\rceil$, then:*

$$\text{Compl}_{\text{HP}}^{\gamma}(\mathcal{F}_{\ell_p^d}^{\kappa}(1), \mathcal{X}, K, \varepsilon) \ge M := \frac{c_\kappa}{\ln(1/\varepsilon) + \kappa \ln\ln(dK/\gamma)} \left(\frac{1}{\varepsilon}\right)^{\frac{2}{3+2\kappa}}.$$

Let us consider some special cases of the bound from Theorem 14. Suppose that d is sufficiently high-dimensional so that the theorem applies (note that $d = \Omega(\ln(dK/\gamma)\varepsilon^{-2})$ suffices). When $\kappa = 1$, then $M = \Omega(\frac{1}{\ln(1/\varepsilon)+\ln\ln(dK/\gamma)}(\frac{1}{\varepsilon})^{2/5})$. This bound does not match the sequential complexity $\Theta(1/\sqrt{\varepsilon})$ of this problem – apart from the logarithmic factors, the exponent in $1/\varepsilon$ is off by $1/10$. This is a direct consequence of the right term in Theorem 12 not being large enough for $p \to \infty$, as the bound in Theorem 14 is obtained from this case. Further improvements of this term would also improve the bound for the nonsmooth ℓ_∞ case of Nemirovski (1994) for, at least, some regimes of ε. Similarly, when $\kappa = 0$, the exponent in $1/\varepsilon$ is $2/3$, which is off by additive $4/3$ from the sequential complexity of this setting. This is aligned with the intuition that smooth lower bounds have a milder high-dimensional regime than nonsmooth ones (which holds in the sequential case). This way, the embedding approach is stronger on higher levels of smoothness.

The main difficulty in obtaining tighter bounds in these regimes (ℓ_∞ and its implications on smooth and weakly-smooth $p \in [1, 2)$ settings) is in relaxing Assumption (d) from Theorem 2. It seems unlikely that this would be possible without completely changing the hard instance used in its proof (as Assumption (d) is crucially used in bounding below the optimality gap), and would likely require a fundamentally different approach from the one used here, as well as in the related work.

4. Conclusion

This paper rules out the possibility of significantly improving the complexity of convex optimization via parallelization in the exploration of the feasible set with polynomially-bounded in the dimension number of queries per round, for essentially all interesting geometries and classes of functions with different levels of smoothness. Most of the obtained lower bounds match the sequential complexity of these problems, up to, at most, a logarithmic factor in the dimension, and are, thus, (nearly) tight.

However, our bounds only apply to the high-dimensional setting, where $d = \Omega(1/\text{poly}(\varepsilon))$. In the low-dimensional setting, the only bound we are aware of is in terms of worst-case complexity (for deterministic algorithms) for nonsmooth optimization over the ℓ_∞ ball, due to Nemirovski (1994). The bound is $\Omega((\frac{d}{\ln(dK)})^{1/3}\ln(1/\varepsilon))$. It was conjectured in Nemirovski (1994) that the correct bound for nonsmooth optimization over the ℓ_∞ ball should be $\Omega(\frac{d}{\ln(K)}\ln(1/\varepsilon))$. Our analysis recovers a bound similar to Nemirovski's result in the stronger high probability complexity model, but *only for constant ε*. We conjecture that in the low-dimensional setting of both (weakly-)smooth and nonsmooth optimization the correct answer should be $\Omega(\frac{d}{\ln(K/\gamma)}\ln(1/\varepsilon))$.

Acknowledgments

Part of this work was done while JD was a Microsoft Research Fellow at the Simons Institute for the Theory of Computing, for the program on Foundations of Data Science, and while she was a postdoctoral researcher at Boston University. JD was partially supported by the NSF grant #CCF-1740855. CG was partially supported by the FONDECYT project 11160939, and the Millennium Science Initiative of the Ministry of Economy, Development, and Tourism, grant "Millennium Nucleus Center for the Discovery of Structures in Complex Data."

We thank Adam Smith for pointing out the importance of lower bounds in parallel convex optimization from the local differential privacy perspective, for his useful comments and insights that led to this work, and for many useful discussions. We also thank Ilias Diakonikolas for sharing his expertise on anticoncentration, as well as Yossi Arjevani, Nicolas Casabianca, Vitaly Feldman, and José Verschae for valuable discussions and comments regarding this work. An earlier version of this paper made an incorrect use of the orthogonal splittings of ℓ_p spaces, which has since been corrected. We thank Boris Kashin for pointing out this issue. We also thank an anonymous reviewer who pointed out imprecise conditioning in the proof of Lemma 22, which has since been corrected.

References

Naman Agarwal and Elad Hazan. Lower bounds for higher-order convex optimization. In *Proc. COLT'18*, 2018.

Eric Balkanski and Yaron Singer. Parallelization does not accelerate convex optimization: Adaptivity lower bounds for non-smooth convex minimization. *arXiv preprint arXiv:1808.03880*, 2018.

K. Ball, E. Carlen, and E.H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994.

Gábor Braun, Cristobal Guzman, and Sebastian Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Trans. Information Theory*, 63 (7):4709–4724, 2017.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *arXiv preprint arXiv:1710.11606*, 2017.

A. d'Aspremont, C. Guzmán, and M. Jaggi. Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405, 2018.

John Duchi, Feng Ruan, and Chulhee Yun. Minimax bounds on stochastic batched convex optimization. In *Proc. COLT'18*, 2018.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1 – 14, 2015.

Uffe Haagerup. The best constants in the Khintchine inequality. *Studia Mathematica*, 70:231–283, 1981.

Jonathan A. Kelner, Yin Tat Lee, Lorenzo Orecchia, and Aaron Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *Proc. ACM-SIAM SODA'14*, 2014.

Yin Tat Lee, Satish Rao, and Nikhil Srivastava. A new approach to computing maximum flows using electrical flows. In *Proc. ACM STOC'13*, 2013.

A. Nemirovski. On parallel complexity of nonsmooth convex optimization. *Journal of Complexity*, 10(4):451–463, 1994.

A. Nemirovskii and Y. Nesterov. Optimal methods of smooth convex optimization *(in Russian)*. *Zh. Vychisl. Mat. i Mat. Fiz.*, 25(3):356–369, 1985.

A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983. ISBN 0 471 10345 4.

Y. Nesterov and A. Nemirovski. On First-Order Algorithms for ℓ_1/Nuclear Norm Minimization. *Acta Numerica*, 22, 4 2013.

G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, 1 edition, 1989.

Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *Proc. IEEE SP'17*, 2017.

N. Srebro and K. Sridharan. On convex optimization, fat shattering and learning. Unpublished, 2012. URL http://ttic.uchicago.edu/~karthik/optfat.pdf.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2019.

Blake Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, and Nati Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. In *Proc. NeurIPS'18*, 2018.

Blake E. Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Proc. NIPS'16*, 2016.

Appendix A. Additional Background

For completeness, this section provides additional background and statements of some known facts that are used in the proofs of our lower bounds.

A.1. Notions of Complexity in the Local Oracle Model

The *worst-case* oracle complexity is defined as:

$$\mathrm{Compl}_{\mathrm{WC}}(\mathcal{F},\mathcal{X},K,\varepsilon)=\inf_{A\in\mathcal{A}^K(\mathcal{O})}\sup_{f\in\mathcal{F}}T(A,f,\varepsilon).$$

For the case of randomized algorithms, it can be shown Nemirovsky and Yudin (1983) that their complexity is equivalent to the one obtained from the expected running time over mixtures of deterministic algorithms. That means that we can define the *randomized* oracle complexity as:

$$\mathrm{Compl}_{\mathrm{R}}(\mathcal{F},\mathcal{X},K,\varepsilon)=\inf_{R\in\Delta(\mathcal{A}^K(\mathcal{O}))}\sup_{f\in\mathcal{F}}\mathbb{E}_{A\sim R}[T(A,f,\varepsilon)],$$

where $\Delta(\mathcal{B})$ is the set of probability distributions on the set $\mathcal{B}$.

We may consider an even weaker notion of *distributional* oracle complexity, defined as

$$\mathrm{Compl}_{\mathrm{D}}(\mathcal{F},\mathcal{X},K,\varepsilon)=\sup_{F\in\Delta(\mathcal{F})}\inf_{A\in\mathcal{A}^K(\mathcal{O})}\mathbb{E}_{f\sim F}[T(A,f,\varepsilon)].$$

In this case, it is important to note that lower bounds cannot be obtained from adversarial choices of f, as the probability distribution on instances F must be set before the algorithm is chosen. It is easily seen that:

$$\mathrm{Compl}_{\mathrm{D}}(\mathcal{F},\mathcal{X},K,\varepsilon)\leq\mathrm{Compl}_{\mathrm{R}}(\mathcal{F},\mathcal{X},K,\varepsilon)\leq\mathrm{Compl}_{\mathrm{WC}}(\mathcal{F},\mathcal{X},K,\varepsilon).$$

Finally, given a confidence parameter $0<\gamma<1$, *high probability* complexity is defined as:

$$\mathrm{Compl}_{\mathrm{HP}}^{\gamma}(\mathcal{F},\mathcal{X},K,\varepsilon)=\sup_{F\in\Delta(\mathcal{F})}\inf_{A\in\mathcal{A}^K(\mathcal{O})}\inf\{\tau:\mathbb{P}_{f\sim F}[T(A,f,\varepsilon)\leq\tau]\geq\gamma\}.$$

Notice that a lower bound on the high probability complexity with confidence parameter γ gives a lower bound on the distributional complexity, by the law of total probability

$$\mathrm{Compl}_{\mathrm{D}}(\mathcal{F},\mathcal{X},K,\varepsilon)\geq(1-\gamma)\mathrm{Compl}_{\mathrm{HP}}^{\gamma}(\mathcal{F},\mathcal{X},K,\varepsilon).$$

All lower bounds in this work are for high probability complexity, with $\gamma=1/\mathrm{poly}(d)$.

A.2. Geometry of ℓ_p Spaces

In the proof of Theorem 14, we make use Dvoretzky's Theorem, on the existence of nearly Euclidean sections of the $\|\cdot\|_p$ ball. Its full description and proof may be found in (Pisier, 1989, Theorem 4.15). Here we state a concise version with what is needed for our results.

Theorem 15 (Dvoretzky) *There exists a universal constant $0<\alpha<1$ such that for any $d>1$, there exists a subspace $F\subseteq\mathbb{R}^d$ of dimension at most αd and an ellipsoid $\mathcal{E}\subseteq F$ such that*

$$\frac{1}{2}\mathcal{E}\subseteq\mathcal{B}_p^d\cap F\subseteq\mathcal{E}.$$

A.3. Smoothings

Claim 16 *Let $2\leq p\leq\infty$, $d\in\mathbb{N}$, and $0\leq\kappa\leq1$. Then, for any $\eta>0$, the space $\ell_p^d=(\mathbb{R}^d,\|\cdot\|_p)$ is (κ,η,η,μ)-locally smoothable when $\mu=2^{1-\kappa}(\min\{p,\ln d\}/\eta)^{\kappa}$.*

Proof First, we use the fact from (Guzmán and Nemirovski, 2015, Proposition 1) that ℓ_p^d is $(1, \eta, \eta, \mu)$-locally smoothable with parameter $\tilde{\mu} = \min\{p, \ln d\}/\eta$. This can be achieved by infimal convolution smoothing

$$\mathcal{S}f(\boldsymbol{x}) = \inf_{h \in \mathcal{B}_p(0,\eta)} [f(\boldsymbol{x} + \boldsymbol{h}) + \phi(\boldsymbol{h})] \qquad (\forall \boldsymbol{x} \in \mathbb{R}^d),$$

where $\phi(\boldsymbol{x}) = 2\|\boldsymbol{x}\|_r^2$ with $r = \min\{p, 3\ln d\}$ as a regularizer. Furthermore, in this reference it is proved that if f is a 1-Lipschitz function, then not only $\mathcal{S}f \in \mathcal{F}^1_{\ell_p^d}(\mu)$ but also $\mathcal{S}f$ is 1-Lipschitz as well; therefore, the following two inequalities hold for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$

$$\begin{aligned} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_*^{1-\kappa} &\leq 2^{1-\kappa} \\ \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_*^{\kappa} &\leq \tilde{\mu}^{\kappa}, \end{aligned}$$

and multiplying these inequalities, we obtain $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_* \leq 2^{1-\kappa}\tilde{\mu}^{\kappa} = \mu$. ■

A.4. Deviation Bounds

Here we state some specific probabilistic deviation bounds that we need for our results. The first one is the left-sided Bernstein inequality, which may be found in (Wainwright, 2019, Chapter 2).

Theorem 17 (Left-Sided Bernstein Inequality) *Let $Y_1, \dots, Y_n$ be nonnegative independent random variables, with finite second moment. Then, for any $\delta > 0$,*

$$\mathbb{P}\Big[\sum_{k=1}^{n}(Y_k - \mathbb{E}[Y_k]) \leq -n\delta\Big] \leq \exp\Big\{ - \frac{n\delta^2}{\frac{2}{n}\sum_{k=1}^{n}\mathbb{E}[Y_k^2]}\Big\}.$$

We also remind the reader of the Khintchine inequality, which provides bounds for L^p moments of Rademacher sequences (see, e.g., Haagerup (1981)).

Theorem 18 (Khintchine) *Let $0 < p < \infty$. There exist constants $c_p, c_p' > 0$ such that for any $x_1, \dots, x_L \in \mathbb{R}$, and $r_1, \dots, r_L$ a Rademacher sequence*

$$c_p\|\boldsymbol{x}\|_2 \leq \Big(\mathbb{E}\Big|\sum_{i=1}^{L} r_i x_i\Big|^p\Big)^{1/p} \leq c_p'\|\boldsymbol{x}\|_2.$$

A.5. Packings and Cardinality of ε-Nets

To show that it is possible to satisfy the assumption of Lemma 20 in the proof of Theorem 2, we will frequently rely on the following simple lemma, which follows by constructing an (ε/M)-net w.r.t. ℓ_∞^M of the simplex, Δ_M.

Lemma 19 *If, $\forall \boldsymbol{\lambda} \in \Delta_M$, $\mathbb{P}\big[\big\|\sum_{i=1}^M \lambda_i \mathbf{z}^i\big\|_* \leq (c+1)\varepsilon\big] \leq \gamma'$ for $\varepsilon \in (0,1)$, $c > 0$, and $\gamma' \in (0,1)$, then:*

$$\mathbb{P}\big[\min_{\boldsymbol{\lambda} \in \Delta_M} \big\|\sum_{i=1}^{M} \lambda_i \mathbf{z}^i\big\|_* \leq c\varepsilon\big] \leq \Big(\frac{3}{\varepsilon}\Big)^M \gamma'.$$

Proof The proof follows by constructing an (ε/M)-net Γ w.r.t. the ℓ_∞ norm. In particular, let Γ be a discrete set of points from Δ_M. To apply the argument, we need to establish that:

$$\left| \inf_{\boldsymbol{\lambda}\in\Delta_M} \Big\| \sum_{i=1}^{M} \lambda_i \mathbf{z}^i \Big\|_* - \inf_{\boldsymbol{\lambda}'\in\Gamma} \Big\| \sum_{i=1}^{M} \lambda_i' \mathbf{z}^i \Big\|_* \right| \le \varepsilon. \tag{5}$$

For (5) to hold, it suffices to show that for every $\boldsymbol{\lambda} \in \Delta_M$, there exists $\boldsymbol{\lambda}' \in \Gamma$ such that

$$\Big\| \sum_{i=1}^{M} \lambda_i' \mathbf{z}^i \Big\|_* \le \Big\| \sum_{i=1}^{M} \lambda_i \mathbf{z}^i \Big\|_* + \varepsilon.$$

By the triangle inequality,

$$\begin{aligned} \| \sum_{i=1}^{M} \lambda_i' \mathbf{z}^i \|_* - \| \sum_{i=1}^{M} \lambda_i \mathbf{z}^i \|_* &\le \| \sum_{i=1}^{M} (\lambda_i' - \lambda_i) \mathbf{z}^i \|_* \\ &\le M \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_\infty \max_{i\in[M]} \|\mathbf{z}^i\|_* \\ &\le M \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_\infty, \end{aligned}$$

as $\mathbf{z}^i \in \mathcal{B}_{\|\cdot\|_*}$. Hence, it suffices to have $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_\infty \le \varepsilon/M$.

Define the discrete set ((ε/M)-net) Γ to be the set of vectors $\boldsymbol{\lambda}'$ such that $\forall j \in \{1, ..., M\}$: $\lambda_j' = n_j \left\lceil \frac{M}{\varepsilon} \right\rceil^{-1}$, where $n_j \ge 0$, $\forall M$, and $\sum_{j=1}^{M} n_j = \left\lceil \frac{M}{\varepsilon} \right\rceil$. Clearly, for any $\boldsymbol{\lambda} \in \Delta_M$, we can choose $\boldsymbol{\lambda}' \in \Gamma$ such that $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_\infty \le \varepsilon/M$. Applying the union bound over $\boldsymbol{\lambda}' \in \Gamma$ and using the lemma assumption:

$$\mathbb{P}\Big[\inf_{\boldsymbol{\lambda}'\in\Gamma} \Big\| \sum_{i=1}^{M} \lambda_i' \mathbf{z}^i \Big\|_* \le (c+1)\varepsilon \Big] \le |\Gamma| \gamma'.$$

The size of the ε-net Γ can be bounded by $|\Gamma| = \binom{\lceil \frac{M}{\varepsilon} \rceil + M}{M} \le \left(\frac{3}{\varepsilon}\right)^M$ using the standard stars and bars combinatorial argument. To complete the proof, it remains to apply the bound from Eq. (5). ■

Appendix B. Proof of Theorem 2

B.1. Upper Bound on the Optimum.

The upper bound on F^* is obtained based on the assumptions from Part (b) of Theorem 2, as follows.

Lemma 20 *If* $\mathbb{P}\Big[\inf_{\boldsymbol{\lambda}\in\Delta_M} \Big\| \sum_{i\in M} \lambda_i \mathbf{z}^i \Big\|_* \le 4\mu\varepsilon\Big] \le \gamma$, *then*

$$\mathbb{P}[F^* \le -2\varepsilon + (\eta - \bar{\delta}/2)/\mu] \ge 1 - \gamma,$$

where $F^* = \min_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$ *for* $F(\boldsymbol{x})$ *defined in* (3), *and* $\mathcal{S}$ *is a smoothing map that satisfies the assumptions from Theorem* 2.

Proof Observe first that:

$$
\begin{aligned}
F^* &\leq \frac{1}{\mu}\min_{\boldsymbol{x}\in\mathcal{X}} \mathcal{S}\Big(\max\Big\{\frac{1}{2}\max_{i\in[M]}\big[\left\langle \mathbf{z}^i, \cdot\right\rangle - i\bar{\delta}\big], \|\cdot\| - \frac{1}{2}(3(1+r)+M\bar{\delta})\Big\}\Big)(\boldsymbol{x})\\
&\leq \frac{1}{\mu}\Big(\min_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}}\Big(\max\Big\{\frac{1}{2}\max_{i\in[M]}\big[\left\langle \mathbf{z}^i, \boldsymbol{x}\right\rangle - i\bar{\delta}\big], \|\boldsymbol{x}\| - \frac{1}{2}(3(1+r)+M\bar{\delta})\Big\} + \eta\Big)\\
&\leq \frac{1}{2\mu}\Big(\min_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}}\max_{i\in[M]}\left\langle \mathbf{z}^i, \boldsymbol{x}\right\rangle\Big) + \frac{\eta-\bar{\delta}/2}{\mu},
\end{aligned}
$$

where we have used Property (i) from the definition of local smoothing, and property (O_1) (to assert that the maximum is achieved by the left term).

The rest of the proof is a simple corollary of minimax duality. In particular, as

$$
\max_{i\in[M]}\left\langle \mathbf{z}^i, \boldsymbol{x}\right\rangle = \max_{\boldsymbol{\lambda}\in\Delta_M}\sum_{i\in[M]}\lambda_i\left\langle \mathbf{z}^i, \boldsymbol{x}\right\rangle,
$$

we have that: $\min_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}}\max_{i\in[M]}\left\langle \mathbf{z}^i, \boldsymbol{x}\right\rangle = \max_{\boldsymbol{\lambda}\in\Delta_M}\min_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}}\sum_{i\in[M]}\lambda_i\left\langle \mathbf{z}^i, \boldsymbol{x}\right\rangle$. Finally:

$$
\min_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}}\sum_{i\in[M]}\lambda_i\left\langle \mathbf{z}^i, \boldsymbol{x}\right\rangle = -\max_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}}\sum_{i\in[M]}\lambda_i\left\langle \mathbf{z}^i, \boldsymbol{x}\right\rangle = -\Big\|\sum_{i=1}^{m}\lambda_i\mathbf{z}^i\Big\|_*,
$$

by the (standard) definition of the dual norm $\|\cdot\|_*$. Hence:

$$
F^* \leq -\frac{1}{2\mu}\left(\min_{\boldsymbol{\lambda}\in\Delta_M}\Big\|\sum_{i=1}^{m}\lambda_i\mathbf{z}^i\Big\|_*\right) + \frac{2\eta-\bar{\delta}}{2\mu},
$$

and it remains to apply the assumption from the statement of the lemma. ■

B.2. Lower Bound on the Algorithm's Output.

Lower bound on the algorithm's output requires more technical work and is based on showing that, at every round t, w.h.p., the algorithm can only learn $\mathbf{z}^1, \ldots, \mathbf{z}^t$ and (aside from implicit bounds) has no information about $\mathbf{z}^{t+1}, \ldots \mathbf{z}^M$. Then, due to Part (c) of Theorem 2, w.h.p., none of the queried points up to round M can align well with vector $\mathbf{z}^M$, which will allow us to show that for all the queried points $\boldsymbol{x}$ up to round M, $F(\boldsymbol{x})$ is $\Omega(\varepsilon)$-far from the optimum F^*.

In the following, we denote the history of the algorithm-oracle interaction until round $t-1$ as $\Pi^{<t} := (X^s, \mathcal{O}_F(X^s))_{s<t}$. We also define the following events

$$
\mathcal{E}^t(\boldsymbol{x}) := \Big\{\left\langle \mathbf{z}^t, \boldsymbol{x}\right\rangle > -\frac{\bar{\delta}}{4}\Big\} \cap \Big\{\left\langle \mathbf{z}^i, \boldsymbol{x}\right\rangle < \frac{\bar{\delta}}{4}\ (\forall i > t)\Big\}, \text{ and } \boldsymbol{\mathcal{E}}^t := \bigcap_{\boldsymbol{x}\in\overline{X}^t}\mathcal{E}^t(\boldsymbol{x}),
$$

where $\bar{\delta}$ is defined as in Theorem 2. Furthermore, we define the "good history" events by:

$$
\boldsymbol{\mathcal{E}}^{<t} := \bigcap\{\boldsymbol{\mathcal{E}}^s : s < t\}.
$$

To avoid making vacuous statements, we take $\boldsymbol{\mathcal{E}}^{<1}$ to be the entire probability space, so that $\mathbb{P}\big[\boldsymbol{\mathcal{E}}^{<1}\big] = 1$. We remind the reader that, based on Property (O_2), when we prove our claim, it suffices to focus on vectors within the ball of radius 4. For this reason, given a batch of queries $X^t = \{\boldsymbol{x}_1^t, \ldots, \boldsymbol{x}_K^t\}$, we define its *relevant queries* as $\overline{X}^t = X^t \cap \mathcal{B}_{\|\cdot\|}(0,4)$.

We first prove that, conditionally on event $\boldsymbol{\mathcal{E}}^{<t}$, X^t is a deterministic function of $\{\mathbf{z}^i\}_{i<t}$.

Proposition 21 *Let $t \in [M-1]$ and suppose event $\boldsymbol{\mathcal{E}}^t$ holds. Then, $\forall \boldsymbol{x} \in \overline{X}^t + \mathcal{B}_{\|\cdot\|}(0,r)$:*

$$F(\boldsymbol{x}) = \frac{1}{\mu}\mathcal{S}\Big(\max\Big\{\frac{1}{2}\max_{i\in[t]}\big[\langle \mathbf{z}^i, \cdot\rangle - i\bar{\delta}\big], \|\cdot\| - \frac{1}{2}(3(1+r) + M\bar{\delta})\Big\}\Big)(\boldsymbol{x}).$$

Moreover, conditionally on $\boldsymbol{\mathcal{E}}^{<t}$, X^t is a deterministic function of $\{\mathbf{z}^i\}_{i<t}$.

Proof Let $f(\boldsymbol{x}) = \max_{i\in[M]}\big[\langle \mathbf{z}^i, \boldsymbol{x}\rangle - i\bar{\delta}\big]$. We will show that for any $\boldsymbol{x}_k^t \in \overline{X}^t$ and $\boldsymbol{x}$ such that $\|\boldsymbol{x} - \boldsymbol{x}_k^t\| \leq 2r$, we have $f(\boldsymbol{x}) = g(\boldsymbol{x})$, where $g(\boldsymbol{x}) = \max_{i\in[t]}\big[\langle \mathbf{z}^i, \boldsymbol{x}\rangle - i\bar{\delta}\big]$ (notice that g only includes $\mathbf{z}^i$ for $i \in [t]$). The first part of the proposition is then obtained from Part (ii) of Definition 1.

To prove the claim, notice that since $\|\mathbf{z}^i\|_* \leq 1$, we have:

$$\langle \mathbf{z}^i, \boldsymbol{x}\rangle \leq \langle \mathbf{z}^i, \boldsymbol{x}_k^t\rangle + \|\boldsymbol{x} - \boldsymbol{x}_k^t\| \cdot \|\mathbf{z}^i\|_* \leq \langle \mathbf{z}^i, \boldsymbol{x}_k^t\rangle + 2r.$$

Similarly, $\langle \mathbf{z}^i, \boldsymbol{x}_k^t\rangle \leq \langle \mathbf{z}^i, \boldsymbol{x}\rangle + \|\boldsymbol{x} - \boldsymbol{x}_k^t\| \cdot \|\mathbf{z}^i\|_* \leq \langle \mathbf{z}^i, \boldsymbol{x}\rangle + 2r$.

Further, by the definition of $\boldsymbol{\mathcal{E}}^t$, and since $2r \leq \bar{\delta}/4$ (by Theorem 2, Assumption (a)),

$$\begin{aligned} \langle \mathbf{z}^i, \boldsymbol{x}\rangle - i\bar{\delta} \;&\leq\; \langle \mathbf{z}^i, \boldsymbol{x}_k^t\rangle + 2r - (t+1)\bar{\delta} \;<\; \frac{\bar{\delta}}{2} - (t+1)\bar{\delta} \\ &<\; \langle \mathbf{z}^t, \boldsymbol{x}_k^t\rangle - 2r - t\delta \;\leq\; \langle \mathbf{z}^t, \boldsymbol{x}\rangle - t\bar{\delta}. \end{aligned}$$

For the second part of the proposition, first observe that $X^t = U^t(\Pi^{<t})$, where U^t is a deterministic function; thus it suffices to prove that, conditionally on $\boldsymbol{\mathcal{E}}^{<t}$, $\Pi^{<t}$ is a deterministic function of $\{\mathbf{z}^i\}_{i<t}$. We prove the last claim by induction on t. For the base case, $\Pi^{<1}$ is empty, thus the property trivially holds. For the inductive step, suppose that conditionally on $\boldsymbol{\mathcal{E}}^{<t}$, $\Pi^{<t}$ is a deterministic function of $\{\mathbf{z}^i\}_{i<t}$. Now notice that $X^t = U^t(\Pi^{<t})$, thus it is a deterministic function of $\{\mathbf{z}^i\}_{i<t}$. On the other hand, the first part of the proposition guarantees that under $\boldsymbol{\mathcal{E}}^t$, $F|_{\overline{X}^t + \mathcal{B}_{\|\cdot\|}(0,r)}$ is a deterministic function of $\{\mathbf{z}^i\}_{i\leq t}$; this proves that $(X^t, \mathcal{O}_F(X^t))$ is a deterministic function of $\{\mathbf{z}^i\}_{i\leq t}$. Finally, combining this with the induction hypothesis, $\Pi^{<t+1} = (\Pi^{<t}, (X^t, \mathcal{O}_F(X^t)))$ is a deterministic function of $\{\mathbf{z}^i\}_{i<t+1}$, proving the inductive step, and thus the result. ■

The last result shows that, under $\boldsymbol{\mathcal{E}}^{<t}$, X^t is predictable w.r.t. $\{\mathbf{z}^i\}_{i<t}$. This means that conditionally on the history and event $\boldsymbol{\mathcal{E}}^{<t}$, X^t is fixed. This is key to leverage the randomness of $\{\mathbf{z}^i\}_{i\geq t}$ for the t-th batch of queries. However, there is still a problem: Conditionally on $\boldsymbol{\mathcal{E}}^{<t}$, the distribution of $\{\mathbf{z}^i\}_{i\geq t}$ is different than when there is no conditioning (recall that $\boldsymbol{\mathcal{E}}^{<t}$ itself depends on $\{\mathbf{z}^i\}_{i\geq t}$). In the next lemma we show that, similar as in Carmon et al. (2017); Woodworth et al. (2018), even after sequential conditioning, the distribution of $\{\mathbf{z}^i\}_{i\geq t}$ remains sufficiently well-concentrated to carry out the lower bound strategy.

Lemma 22 *Under Assumptions (a) and (c) from Theorem 2, we have:*

$$\mathbb{P}\Big[\bigcap_{t\in[M]}\boldsymbol{\mathcal{E}}^t\Big] \geq 1-\gamma.$$

Proof First observe that, for any $1 \leq t \leq M$, by the law of total probability:

$$\mathbb{P}[(\boldsymbol{\mathcal{E}}^t)^c|\,\boldsymbol{\mathcal{E}}^{<t}] \;\;=\;\; \int_{\boldsymbol{\xi}} \mathbb{P}[(\boldsymbol{\mathcal{E}}^t)^c|\,\boldsymbol{\mathcal{E}}^{<t},\, \{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}]\,\mathrm{d}\mathbb{P}\big[\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}|\boldsymbol{\mathcal{E}}^{<t}\big]$$

On the other hand, by the previous proposition, X^t is a deterministic function of $\{\mathbf{z}^i\}_{i<t}$, conditionally on $\boldsymbol{\mathcal{E}}^{<t}$. Recall that:

$$(\boldsymbol{\mathcal{E}}^t)^c = \Big\{\exists \boldsymbol{x} \in \overline{X}^t : \big\langle \mathbf{z}^t, \boldsymbol{x}\big\rangle < -\bar{\delta}/4 \text{ or } (\exists i > t)\,\big\langle \mathbf{z}^i, \boldsymbol{x}\big\rangle > \bar{\delta}/4\Big\}.$$

To simplify the notation, denote:

$$(\boldsymbol{\mathcal{E}}^t)^c_{\{\overline{X}^t\to X\}} = \Big\{\exists \boldsymbol{x} \in X : \big\langle \mathbf{z}^t, \boldsymbol{x}\big\rangle < -\bar{\delta}/4 \text{ or } (\exists i > t)\,\big\langle \mathbf{z}^i, \boldsymbol{x}\big\rangle > \bar{\delta}/4\Big\}.$$

Therefore, we further have:

$$\mathbb{P}[(\boldsymbol{\mathcal{E}}^t)^c|\,\boldsymbol{\mathcal{E}}^{<t}] \leq \int_{\boldsymbol{\xi}} \sup_{\substack{X\subseteq\mathcal{B}_{\|\cdot\|}(0,4),\\ |X|\leq K}} \mathbb{P}\Big[(\boldsymbol{\mathcal{E}}^t)^c_{\{\overline{X}^t\to X\}}\Big|\boldsymbol{\mathcal{E}}^{<t}, \{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\Big]\mathrm{d}\mathbb{P}\big[\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}|\boldsymbol{\mathcal{E}}^{<t}\big],$$

where we have used that $\overline{X}^t$ is conditionally deterministic. Now that X is fixed, we can use the union bound as follows:

$$\begin{aligned}
&\mathbb{P}[(\boldsymbol{\mathcal{E}}^t)^c|\,\boldsymbol{\mathcal{E}}^{<t}]\\
&\leq K\int_{\boldsymbol{\xi}} \sup_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}(0,4)} \mathbb{P}\Big[\big\langle \mathbf{z}^t, \boldsymbol{x}\big\rangle < -\bar{\delta}/4\Big|\boldsymbol{\mathcal{E}}^{<t}, \{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\Big]\mathrm{d}\mathbb{P}[\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}|\boldsymbol{\mathcal{E}}^{<t}]\\
&\quad + (M-1)K\int_{\boldsymbol{\xi}} \sup_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}(0,4)} \max_{j>t}\mathbb{P}\Big[\big\langle \mathbf{z}^j, \boldsymbol{x}\big\rangle > \bar{\delta}/4\Big|\boldsymbol{\mathcal{E}}^{<t}, \{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\Big]\mathrm{d}\mathbb{P}[\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}|\boldsymbol{\mathcal{E}}^{<t}].
\end{aligned}$$

Observe for the first integral in the last expression that we can write:

$$\begin{aligned}
&\int_{\boldsymbol{\xi}} \sup_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}(0,4)} \mathbb{P}\Big[\big\langle \mathbf{z}^t, \boldsymbol{x}\big\rangle < -\bar{\delta}/4\Big|\boldsymbol{\mathcal{E}}^{<t}, \{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\Big]\mathrm{d}\mathbb{P}[\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}|\boldsymbol{\mathcal{E}}^{<t}]\\
&\qquad = \int_{\boldsymbol{\xi}} \sup_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}(0,4)} \frac{\mathbb{P}\big[\big\langle \mathbf{z}^t, \boldsymbol{x}\big\rangle < -\bar{\delta}/4, \boldsymbol{\mathcal{E}}^{<t}|\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\big]}{\mathbb{P}\big[\boldsymbol{\mathcal{E}}^{<t}|\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\big]}\mathrm{d}\mathbb{P}\big[\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}|\boldsymbol{\mathcal{E}}^{<t}\big]\\
&\qquad = \int_{\boldsymbol{\xi}} \sup_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}(0,4)} \frac{\mathbb{P}\big[\big\langle \mathbf{z}^t, \boldsymbol{x}\big\rangle < -\bar{\delta}/4, \boldsymbol{\mathcal{E}}^{<t}|\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\big]}{\mathbb{P}\big[\boldsymbol{\mathcal{E}}^{<t}\big]}\mathrm{d}\mathbb{P}\big[\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\big]\\
&\qquad \leq \int_{\boldsymbol{\xi}} \sup_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}(0,4)} \frac{\mathbb{P}\big[\big\langle \mathbf{z}^t, \boldsymbol{x}\big\rangle < -\bar{\delta}/4|\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\big]}{\mathbb{P}\big[\boldsymbol{\mathcal{E}}^{<t}\big]}\mathrm{d}\mathbb{P}\big[\{\mathbf{z}^i\}_{i<t} = \boldsymbol{\xi}\big]\\
&\qquad = \frac{\sup_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}(0,4)}\mathbb{P}\big[\big\langle \mathbf{z}^t, \boldsymbol{x}\big\rangle < -\bar{\delta}/4\big]}{\mathbb{P}\big[\boldsymbol{\mathcal{E}}^{<t}\big]},
\end{aligned}$$

where we have used the Bayes rule in the second equality. Applying the same arguments to the second integral, we finally have:

$$\begin{aligned}\mathbb{P}\big[(\boldsymbol{\mathcal{E}}^t)^c \big| \, \boldsymbol{\mathcal{E}}^{<t}\big] &\leq \frac{MK \sup_{\boldsymbol{x}\in\mathcal{B}_{\|\cdot\|}(0,4)} \max\big\{\mathbb{P}\big[\langle \mathbf{z}^t, \boldsymbol{x}\rangle < -\bar{\delta}/4\big], \, \max_{j>t}\mathbb{P}\big[\langle \mathbf{z}^j,\boldsymbol{x}\rangle > \bar{\delta}/4\big]\big\}}{\mathbb{P}[\boldsymbol{\mathcal{E}}^{<t}]} \\ &\leq \frac{MK\exp\{-\alpha\bar{\delta}^2/256\}}{\mathbb{P}[\boldsymbol{\mathcal{E}}^{<t}]} \leq \frac{\gamma}{\mathbb{P}[\boldsymbol{\mathcal{E}}^{<t}]}.\end{aligned}$$

Inductively, each $\boldsymbol{\mathcal{E}}^{<t}$ happens with non-zero probability, as $\mathbb{P}\big[\boldsymbol{\mathcal{E}}^{<1}\big] = 1$ and $\gamma < 1$.

We conclude the proof by conditioning:

$$\mathbb{P}\Big[\bigcap_{t\in[M]}\boldsymbol{\mathcal{E}}^t\Big] = \frac{\mathbb{P}\Big[\bigcap_{t\in[M]}\boldsymbol{\mathcal{E}}^t\Big]}{\mathbb{P}\Big[\bigcap_{t<M}\boldsymbol{\mathcal{E}}^t\Big]}\mathbb{P}\Big[\bigcap_{t<M}\boldsymbol{\mathcal{E}}^t\Big] = \mathbb{P}\Big[\boldsymbol{\mathcal{E}}^M\Big|\boldsymbol{\mathcal{E}}^{<M}\Big]\,\mathbb{P}\Big[\boldsymbol{\mathcal{E}}^{<M}\Big] \geq 1-\gamma.$$

■

Finally, Lemma 22 and Proposition 21 imply the following lower bound on the algorithm's output:

$$\mathbb{P}\Big[\min_{t\in[M],\,k\in[K]} F(\boldsymbol{x}_k^t) \geq -\frac{\bar{\delta}}{2\mu}\Big(\frac{1}{4}+M+\frac{2\eta}{\bar{\delta}}\Big)\Big] \geq 1-\gamma, \tag{6}$$

as, when all events $\{\boldsymbol{\mathcal{E}}^t : t\in[M]\}$ hold simultaneously (and, in particular, when event $\boldsymbol{\mathcal{E}}^M$ holds), we have, by the definitions of these events and the random problem instance (3), that:

$$\min_{t\in[M],\,k\in[K]} F(\boldsymbol{x}_k^t) \geq \frac{1}{2\mu}\min\Big\{\big\langle \mathbf{z}^M, \boldsymbol{x}\big\rangle - M\bar{\delta} : \boldsymbol{x}\in\cup_{t\in[M]}\overline{X}^t\Big\} - \frac{\eta}{\mu} \geq -\frac{\bar{\delta}}{8\mu} - M\frac{\bar{\delta}}{2\mu} - \frac{\eta}{\mu}.$$

B.3. Bounding the Optimality Gap.

To complete the proof of Theorem 2, it remains to combine the results from the previous two subsections and argue that, w.p. $1-\gamma$, the optimality gap of any solution output by the algorithm is higher than ε.

Remaining Proof of Theorem 2 Applying Lemma 20, with probability $1-\gamma$, $F^* \leq -2\varepsilon + (\eta - \bar{\delta}/2)/\mu$. From Eq. (6), w.p. $1-\gamma$, $\min_{t\in[M],\,k\in[K]} F(\boldsymbol{x}_k^t) \geq -\frac{\bar{\delta}}{2\mu}\Big(\frac{1}{4}+M+\frac{2\eta}{\bar{\delta}}\Big)$. Hence, with probability $1-2\gamma$,

$$\min_{t\in[M],\,k\in[K]} F(\boldsymbol{x}_k^t) - F^* \geq 2\varepsilon - \frac{\bar{\delta}}{2\mu}\Big(M-\frac{3}{4}\Big) - \frac{2\eta}{\mu} > \varepsilon,$$

as, by the theorem assumptions, $\bar{\delta} \leq \varepsilon\mu/M$ and $\eta \leq \varepsilon\mu/4$. ■

Appendix C. Omitted Proofs from Section 3

C.1. Nonsmooth Optimization for $1 \leq p \leq 2$

Lemma 6 *Let $1 < p \leq 2$ and let $\mathbf{z}^1, \dots, \mathbf{z}^M$ be chosen according to Eq. (4), where*

$$M \leq \min\Big\{\frac{1}{200\varepsilon^2}, \frac{d/12 - \ln(1/\gamma)}{\ln(3/\varepsilon)}\Big\},$$

then for all $\gamma \in (1/\mathrm{poly}(d), 1) : \mathbb{P}[\min_{\boldsymbol{\lambda} \in \Delta_M} \| \sum_{i\in[M]} \lambda_i \mathbf{z}^i \|_{p^*} \leq 4\varepsilon] \leq \gamma$.

Proof By the choice of $\mathbf{z}^i$'s, $\| \sum_{i\in[M]} \lambda_i \mathbf{z}^i \|_{p^*}^{p^*} = \frac{1}{d}\sum_{j\in[d]} \Big|\sum_{i\in[M]} \lambda_i r_j^i\Big|^{p^*}$. Hence, using Lemma 19, it suffices to show that:

$$\mathbb{P}\Big[\frac{1}{d}\sum_{j\in[d]}\Big|\sum_{i\in[M]} \lambda_i r_j^i\Big|^{p^*} \leq (\varepsilon')^{p^*}\Big] \leq \gamma',$$

for $\varepsilon' = 5\varepsilon$ and sufficiently small γ' (namely, for $\gamma' \leq (\frac{\varepsilon}{3})^M\gamma$).

Let $Y_j := \Big|\sum_{i\in[M]} \lambda_i r_j^i\Big|^{p^*}$, for $j \in [d]$, and notice that Y_j's are nonnegative and i.i.d. Moreover, by Khintchine's Inequality, there exist constants c, c' such that:

$$\begin{aligned}
\mathbb{E}[Y_1] &= \mathbb{E}\Big|\sum_{i\in[M]} \lambda_i r_j^i\Big|^{p^*} \geq c\Big(\sum_{i\in[M]} \lambda_i^2\Big)^{p^*/2} = c\|\boldsymbol{\lambda}\|_2^{p^*} \\
\mathbb{E}[Y_1^2] &= \mathbb{E}\Big|\sum_{i\in[M]} \lambda_i r_j^i\Big|^{2p^*} \leq c'\Big(\sum_{i\in[M]} \lambda_i^2\Big)^{2p^*/2} = c'\|\boldsymbol{\lambda}\|_2^{2p^*}.
\end{aligned}$$

In particular, when $1 \leq p \leq 2$, $c' = 1$ and $c \geq 1/\sqrt{2}$ Haagerup (1981). Therefore, by the left-sided Bernstein's Inequality (Theorem 17) for any $0 < \eta < c$:

$$\mathbb{P}\Big[\frac{1}{d}\sum_{j\in[d]} Y_j \leq (c-\eta)\|\boldsymbol{\lambda}\|_2^{p^*}\Big] \quad \leq \quad \exp\left(-\frac{d\eta^2\|\boldsymbol{\lambda}\|_2^{2p^*}}{2c'\|\boldsymbol{\lambda}\|_2^{2p^*}}\right) = \exp\left(-\frac{d\eta^2}{2c'}\right).$$

As $\boldsymbol{\lambda} \in \Delta_M$, it must be $\|\boldsymbol{\lambda}\|_2 \geq 1/\sqrt{M}$. Choosing $\eta = c/2$, we have $(c-\eta)\|\boldsymbol{\lambda}\|_2 \geq \frac{1}{2\sqrt{2M}} \geq 5\varepsilon = \varepsilon'$, and it follows that:

$$\mathbb{P}\Big[\frac{1}{d}\sum_{j\in[d]}\Big|\sum_{i\in[M]} \lambda_i r_j^i\Big|^{p^*} \leq (\varepsilon')^{p^*}\Big] \leq \exp\Big(-\frac{dc^2}{8c'}\Big) \leq \exp\Big(-\frac{d}{8\sqrt{2}}\Big).$$

To complete the proof, it suffices to have $d \geq 8\sqrt{2}\,(\ln(1/\gamma) + M\ln(3/\varepsilon))$. This is clearly satisfied for $M \leq \frac{d/12 - \ln(1/\gamma)}{\ln(3/\varepsilon)}$ from the lemma's assumptions. ■

Theorem 9 *Let $\varepsilon \in (0, 1/2)$, $\gamma \in (1/\mathrm{poly}(d), 1)$, and $p \in [1, 2]$. Then:*

$$\mathrm{Compl}_{\mathrm{HP}}^{\gamma}(\mathcal{F}_{\ell_p^d}^0(1), \mathcal{B}_2^d(1/d^{1/p-1/2}), K, \varepsilon) \geq M := \min\Big\{\frac{1}{200\varepsilon^2},\ \frac{\varepsilon d^{1/2}}{32\sqrt{\ln(MK/\gamma)}}\Big\}.$$

Proof As before, we will prove this result as an application of Theorem 2. Let $\mathbf{r}^i$ be an independent standard Rademacher sequence in $\mathbb{R}^d$ and $\mathbf{z}^i = \frac{1}{d^{1/p^*}}\mathbf{r}^i$. Assumption (a) is automatically satisfied since $\kappa = 0$. On the other hand, and since we are working on a feasible set $\mathcal{X} \neq \mathcal{B}_p^d$, we need to adapt the upper bound on the optimum provided in Assumption (b). By standard duality arguments:

$$\begin{aligned}\min_{\boldsymbol{x}\in\mathcal{B}_2(1/d^{1/p-1/2})}\max_{i\in[M]}\{\langle \mathbf{z}^i,\boldsymbol{x}\rangle - i\delta\} &\leq \frac{1}{d^{1/p-1/2}}\min_{\boldsymbol{x}\in\mathcal{B}_2(1)}\max_{\boldsymbol{\lambda}\in\Delta_M}\Big\langle \sum_{i\in[M]}\lambda_i\mathbf{z}^i,\boldsymbol{x}\Big\rangle - \delta\\ &= -\frac{1}{d^{1/p-1/2}}\min_{\boldsymbol{\lambda}\in\Delta_M}\Big\|\sum_{i\in[M]}\lambda_i\mathbf{z}^i\Big\|_2 - \delta.\end{aligned}$$

Therefore, we replace Property (b) by the following probabilistic guarantee

$$\mathbb{P}\Big[\min_{\boldsymbol{\lambda}\in\Delta_M}\Big\|\sum_{i\in[M]}\lambda_i\frac{\mathbf{z}^i}{d^{1/p-1/2}}\Big\|_2 \leq 4\varepsilon\Big] = \mathbb{P}\Big[\min_{\boldsymbol{\lambda}\in\Delta_M}\Big\|\sum_{i\in[M]}\lambda_i\frac{\mathbf{r}^i}{\sqrt{d}}\Big\|_2 \leq 4\varepsilon\Big] \leq \gamma$$

for any $M \leq \min\left\{\frac{1}{200\varepsilon^2}, \frac{d/12-\ln(1/\gamma)}{\ln(3/\varepsilon)}\right\}$, which holds by Lemma 6. On the other hand, the concentration required in Assumption (c) is satisfied for any $\boldsymbol{x}\in\mathcal{B}_2^d(1/d^{1/p-1/2})$ by Hoeffding

$$\mathbb{P}[\langle\mathbf{z}^i,\boldsymbol{x}\rangle > \delta] \leq \exp\{-\delta^2 d^{2/p^*}/\|\boldsymbol{x}\|_2\} \leq \exp\{-d\delta^2\}.$$

In particular, we can choose $\alpha = d$. Finally, Assumption (d) is satisfied for $M \leq \frac{\varepsilon}{\delta} = \frac{\varepsilon}{16}\sqrt{\frac{d}{\ln(MK/\gamma)}}$, completing the proof. ■

C.2. Smooth, Weakly Smooth, and Nonsmooth Optimization for $p \geq 2$

We start by showing that, under suitable constraints on M and L, Assumptions (b) and (c) from Theorem 2 are satisfied. This will suffice to apply Theorem 2 in the case of nonsmooth optimization (i.e., for $\mathcal{S}$ being a $(0,0,0,1)$-local smoothing). To obtain results in the smooth and weakly smooth settings, we will then show how to satisfy the remaining assumptions for a suitable local smoothing.

In terms of Assumption (b), we can in fact obtain a much stronger result than needed in Theorem 2:

Lemma 23 *Let $p \geq 2$, $\varepsilon \in (0,1)$, $\mu > 0$, $\mathbf{z}^i$'s chosen as described in Section 3.1.2 and:*

$$M \leq \Big(\frac{1}{4\mu\varepsilon}\Big)^p$$

then:

$$\mathbb{P}\Big[\min_{\boldsymbol{\lambda}\in\Delta_M}\Big\|\sum_{i\in[M]}\lambda_i\mathbf{z}^i\Big\|_{p^*} \leq 4\mu\varepsilon\Big] = 0.$$

Proof Let $\boldsymbol{\lambda}\in\Delta_M$ be fixed. Observe that, since $\mathbf{z}^i$'s have disjoint support (each $\mathbf{z}^i$ is supported on J_i such that $|J_i| = L$ and $J_i \cap J_{i'} = \emptyset$ for all $i \neq i'$), vector $\sum_{i\in[M]}\lambda_i\mathbf{z}^i$ is such that its coordinates indexed by $j \in J_i$ (L of them) are equal to $\lambda_i z_j^i$, $\forall i \in [M]$. Therefore, using the definition of $\mathbf{z}^i$ (Equation 4):

$$\Big\|\sum_{i\in[M]}\lambda_i\mathbf{z}^i\Big\|_{p^*}^{p^*} = \sum_{i\in[M]}\Big(L\cdot\Big(\lambda_i\frac{1}{L^{1/p^*}}\Big)^{p^*}\Big) = \|\boldsymbol{\lambda}\|_{p^*}^{p^*}.$$

By the relationship between ℓ_p norms and the definition of $\boldsymbol{\lambda}$, we have that $1 = \|\boldsymbol{\lambda}\|_1 \leq M^{1/p}\|\boldsymbol{\lambda}\|_{p^*}$. Hence:

$$\Big\| \sum_{i\in[M]} \lambda_i \mathbf{z}^i \Big\|_{p^*} = \|\boldsymbol{\lambda}\|_{p^*} \geq M^{-1/p} \geq 4\mu\varepsilon.$$

Since this holds for all $\boldsymbol{\lambda} \in \Delta_M$ surely, the proof is complete. ■

For Assumption (c), we have the following (simple) lemma:

Lemma 24 *Let $p \geq 2$ and $\mathbf{z}^i$'s chosen as described in Section 3.1.2, then:*

$$\mathbb{P}[\langle \mathbf{z}^i, \boldsymbol{x}\rangle \geq \delta] = \mathbb{P}[\langle \mathbf{z}^i, \boldsymbol{x}\rangle \leq -\delta] \leq \exp\Big(-\frac{L\delta^2}{2}\Big) \qquad (\forall \boldsymbol{x} \in \mathcal{B}_p^d).$$

Proof By the definition of $\mathbf{z}^i$ and Hoeffding's Inequality, $\forall \boldsymbol{x} \in \mathcal{X}$:

$$\mathbb{P}[\langle \mathbf{z}^i, \boldsymbol{x}\rangle > \delta] = \mathbb{P}[\langle \mathbf{z}^i, \boldsymbol{x}\rangle < -\delta] = \mathbb{P}\Big[\sum_{j\in J_i} r_j^i x_j > \delta L^{1/p^*}\Big]$$
$$\leq \exp\Big(-\frac{L^{2/p^*}\delta^2}{2\sum_{j\in J_i} {x_j}^2}\Big).$$

As $|J_i| = L$, by the relations between ℓ_p norms, $(\sum_{j\in J_i} {x_j}^2)^{1/2} \leq L^{1/2-1/p}(\sum_{j\in J_i} {x_j}^p)^{1/p} \leq L^{1/2-1/p}$. Thus, it follows that:

$$\mathbb{P}[\langle \mathbf{z}^i, \boldsymbol{x}\rangle > \delta] = \mathbb{P}[\langle \mathbf{z}^i, \boldsymbol{x}\rangle < -\delta] \leq \Big(-\frac{L^{2/p^*}\delta^2}{2L^{1-2/p}}\Big) = \exp\Big(-\frac{L\delta^2}{2}\Big),$$

as claimed. ■

To obtain the result for the nonsmooth case, we can take $\mu = 1$ and apply Theorem 2, as follows.

Theorem 10 *Let $p \geq 2$, $\mathcal{X} \supseteq \mathcal{B}_p^d$, and $\varepsilon \in (0, 1/2)$, $\gamma \in (1/\mathrm{poly}(d), 1)$. Then:*

$$\mathrm{Compl}_{\mathrm{HP}}^{\gamma}(\mathcal{F}_{\ell_p^d}^{\kappa}(1), \mathcal{X}, K, \varepsilon) \geq M := \min\Big\{\frac{1}{(4\varepsilon)^p}, \frac{\varepsilon^{2/3}}{8}\Big(\frac{d}{\ln(MK/\gamma)}\Big)^{1/3}\Big\}.$$

Proof For Lemma 23 to apply, it suffices to have $M \leq \frac{1}{(4\varepsilon)^p}$, as in the nonsmooth case $\mu = 1$. Lemma 24 implies that it suffices to set $\alpha = L/2 = d/(2M)$. As $\bar{\delta} = 16\sqrt{\frac{\ln(MK/\gamma)}{\alpha}}$, to satisfy Assumption (d) from Theorem 2 (which requires $\bar{\delta} \leq \varepsilon/M$), it suffices to have:

$$M \leq \frac{\varepsilon}{16}\sqrt{\frac{d}{2M\ln(MK/\gamma)}},$$

or, equivalently: $M \leq \frac{\varepsilon^{2/3}}{8}\Big(\frac{d}{\ln(MK/\gamma)}\Big)^{1/3}$, as claimed. ■

To obtain lower bounds for the κ-weakly smooth case (where $\kappa \in [0, 1]$; $\kappa = 0$ is the nonsmooth case from the above and $\kappa = 1$ is the standard notion of smoothness), we need to, in addition to using Lemmas 23 and 24, choose an appropriate local smoothing that satisfies the remaining conditions from Theorem 2. By doing so, we can obtain the following result.

Theorem 12 *Let $p \geq 2$, $\mathcal{X} \supseteq \mathcal{B}_p^d$, and $\varepsilon \in (0, 1/2)$, $\gamma \in (1/\mathrm{poly}(d), 1)$. Then:*

$$\mathrm{Compl}_{\mathrm{HP}}^{\gamma}(\mathcal{F}_{\ell_p^d}^{\kappa}(1), \mathcal{X}, K, \varepsilon) \geq M := \min\Bigg\{\Big(\frac{1}{2^{3+4\kappa}\,\varepsilon\,(\min\{p, \ln(d)\})^{\kappa}}\Big)^{\frac{p}{1+\kappa(1+p)}}, \\ \frac{d}{2^9 \ln(MK/\gamma)}\left(2^{\frac{1+3p+2\kappa(1+p)}{1+p}} \min\{p, \ln(d)\}^{\kappa}\varepsilon\right)^{\frac{2(1+p)}{1+\kappa(1+p)}}\Bigg\}.$$

Proof From Remark 11 we have that ℓ_p^d is $(\kappa, \eta, \eta, \mu)$-locally smoothable (observe here that $r = \eta$) for any $0 \leq \kappa \leq 1$, as long as $\mu = 2^{1-\kappa}(\min\{p, \ln(d)\}/\eta)^{\kappa}$.

Let $\eta = \bar{\delta}/8$. Denote $\bar{\mu} = 2^{1-\kappa}(8\min\{p, \ln(d)\})^{\kappa}$, so that $\mu = \frac{\bar{\mu}}{\bar{\delta}^{\kappa}}$. To satisfy Assumptions (a) and (d), we need to have $\bar{\delta} \leq \min\{2\varepsilon\mu, \varepsilon\mu/M\}$, and it suffices to enforce $M \leq \frac{\mu\varepsilon}{\bar{\delta}} = \frac{\varepsilon\bar{\mu}}{\bar{\delta}^{1+\kappa}}$. To satisfy Assumption (c), by Lemma 24 we can choose $\alpha = \frac{L}{2}$, which leads to the following bound on M:

$$M \leq \frac{\varepsilon\bar{\mu}}{2^{4(1+\kappa)}}\Big(\frac{L}{2\ln(MK/\gamma)}\Big)^{\frac{1+\kappa}{2}}. \tag{7}$$

To satisfy the remaining assumption from Theorem 2 (Assumption (b), using Lemma 23), we need to impose the following constraint on M:

$$M \leq \left(\frac{1}{4\mu\varepsilon}\right)^p = \left(\frac{\bar{\delta}^{\kappa}}{4\varepsilon\bar{\mu}}\right)^p = \Big(\frac{2^{2(2\kappa-1)}}{\varepsilon\bar{\mu}}\Big)^p\Big(\frac{L}{2\ln(MK/\gamma)}\Big)^{-\frac{p\kappa}{2}} \tag{8}$$

The right-hand sides of the inequalities in Equations (7) and (8) are equal when

$$L = 2^9\ln(MK/\gamma)\cdot\left(\frac{4}{(4\bar{\mu}\varepsilon)^{p+1}}\right)^{2/[1+\kappa(p+1)]}$$

and, thus, we make this choice for L. As $d \geq ML$, we also need to satisfy $M \leq d/L$, finally leading to the claimed bound:

$$M \leq \min\left\{\left(\frac{1}{4^{1+\kappa}\bar{\mu}\varepsilon}\right)^{\frac{p}{1+\kappa(1+p)}}, \frac{d}{2^9\ln(MK/\gamma)}\left(4^{\frac{p}{1+p}}\bar{\mu}\varepsilon\right)^{\frac{2(1+p)}{1+\kappa(1+p)}}\right\}$$

The rest of the proof follows by plugging $\bar{\mu} = 2^{1+2\kappa}(\min\{p, \ln(d)\})^{\kappa}$ in the last equation. ■

C.3. Smooth and Weakly Smooth Optimization for $1 \leq p < 2$

Theorem 14 *Let $1 \leq p < 2$, $0 < \kappa \leq 1$, $\mathcal{X} \supseteq \mathcal{B}_p^d$, $\varepsilon \in (0, 1/2)$, $\gamma \in (1/\mathrm{poly}(d), 1)$. Then, there exist constants $\nu, c(\kappa) > 0$, such that if $d \geq \frac{1}{\nu}\left\lceil 2(\ln(\nu dK/\gamma))^{\frac{2\kappa}{3+2\kappa}}\left(\frac{1}{\varepsilon}\right)^{\frac{6}{3+2\kappa}}\right\rceil$, then:*

$$\mathrm{Compl}_{\mathrm{HP}}^{\gamma}(\mathcal{F}_{\ell_p^d}^{\kappa}(1), \mathcal{X}, K, \varepsilon) \geq M := \frac{c_{\kappa}}{\ln(1/\varepsilon) + \kappa\ln\ln(dK/\gamma)}\Big(\frac{1}{\varepsilon}\Big)^{\frac{2}{3+2\kappa}}.$$

Proof Sketch By Dvoretzky's Theorem (see Appendix A), there exists a universal constant $\nu > 0$ such that for any $T \leq \nu d$ there exists a subspace $F \subseteq \mathbb{R}^d$ of dimension T, and a centered ellipsoid $\mathcal{E} \subseteq F$, such that

$$\frac{1}{2}\mathcal{E} \subseteq F \cap \mathcal{B}_p^d \subseteq \mathcal{E}. \tag{9}$$

By an application of the Hahn-Banach theorem, we can certify that there exist vectors $\boldsymbol{g}^1, \ldots, \boldsymbol{g}^T \in \mathcal{B}_{p^*}^d$, such that $\mathcal{E} = \{\boldsymbol{x} \in F : \sum_{i\in[T]} \langle \boldsymbol{g}^i, \boldsymbol{x} \rangle^2 \leq 1\}$.

Consider now linear mapping $G : (\mathbb{R}^d, \|\cdot\|_p) \mapsto (\mathbb{R}^T, \|\cdot\|_\infty)$ such that $G\boldsymbol{x} := (\langle \boldsymbol{g}^1, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{g}^T, \boldsymbol{x} \rangle)$, and notice that by the previous paragraph the operator norm of G is upper bounded by 1. We observe that:

- For any $f \in \mathcal{F}_{\ell_\infty^T}^\kappa(\mu)$, function $\tilde{f} := f \circ G$ belongs to $\mathcal{F}_{\ell_p^d}^\kappa(\mu)$. In other words, the whole function class $\mathcal{F}_{\ell_\infty^T}^\kappa(\mu)$ can be obtained from $\mathcal{F}_{\ell_p^d}^\kappa(\mu)$ through the linear embedding G.
- We claim that any local oracle for the class $\{\tilde{f} : f \in \mathcal{F}_{\ell_\infty^T}^\kappa(\mu)\}$ can be obtained from a local oracle for the class $\mathcal{F}_{\ell_p^d}^\kappa(\mu)$ (for a proof of this claim, see (Guzmán and Nemirovski, 2015, Appendix C)).
- From (9), the set $\mathcal{Y} = G\mathcal{B}_p^d$ is such that $\frac{1}{2\sqrt{T}}\mathcal{B}_\infty^T \subseteq \frac{1}{2}\mathcal{B}_2^T \subseteq \mathcal{Y} \subseteq \mathcal{B}_2^T$.

From these facts, we can conclude that the oracle complexity over $\mathcal{X}$ with function class $\mathcal{F}_{\ell_p^d}^\kappa(1)$ is at least the one obtained in the embedded space $\mathcal{Y}$ with the respective embedded function class $\mathcal{F}_{\ell_\infty^T}^\kappa(1)$, thus

$$\begin{aligned}\mathrm{Compl}_{\mathrm{HP}}^\gamma(\mathcal{F}_{\ell_p^d}^\kappa(1), \mathcal{X}, K, \varepsilon) &\geq \mathrm{Compl}_{\mathrm{HP}}^\gamma(\mathcal{F}_{\ell_\infty^T}^\kappa(1), \mathcal{Y}, K, \varepsilon)\\ &\geq \mathrm{Compl}_{\mathrm{HP}}^\gamma(\mathcal{F}_{\ell_\infty^T}^\kappa(1), \mathcal{B}_\infty^T(0, 1/[2\sqrt{T}]), K, \varepsilon)\end{aligned}$$

Denote $\varepsilon' = 2\varepsilon\sqrt{T}$. By Theorem 12 applied to $p = \infty$, together with Remark 3, we get that it is sufficient to require, as long as $T \leq \nu d$, that:

$$\begin{aligned} M &= \min\left\{\frac{1}{\ln(T)}\Big(\frac{1}{2^{3+4\kappa}\varepsilon'}\Big)^{1/\kappa}, \frac{T\ln^2(T)}{2^9\ln(\nu dK/\gamma)}\big(2^{3+2\kappa}\varepsilon'\big)^{2/\kappa}\right\}\\ &= \min\left\{\frac{1}{\ln(T)}\Big(\frac{1}{2^{4(1+\kappa)}\varepsilon\sqrt{T}}\Big)^{1/\kappa}, \frac{T\ln^2(T)}{2^9\ln(\nu dK/\gamma)}\big(2^{2(2+\kappa)}\varepsilon\sqrt{T}\big)^{2/\kappa}\right\}.\end{aligned}$$

In the last expression, the left term in the minimum is lower whenever:

$$T\ln^2 T \geq (2^9\ln(dK/\gamma))^{\frac{2\kappa}{3+2\kappa}}\left(\frac{1}{2}\right)^{\frac{8(2+3\kappa)}{3+2\kappa}}\left(\frac{1}{\varepsilon}\right)^{\frac{6}{3+2\kappa}},$$

and it suffices to choose:

$$T = \left\lceil 2(\ln(\nu dK/\gamma))^{\frac{2\kappa}{3+2\kappa}}\left(\frac{1}{\varepsilon}\right)^{\frac{6}{3+2\kappa}}\right\rceil.$$

Under this choice, as long as $d \geq T/\nu$, the oracle complexity is lower bounded by:

$$M = \frac{c_\kappa}{\ln(1/\varepsilon) + \kappa\ln\ln(dK/\gamma)}\left(\frac{1}{\varepsilon}\right)^{\frac{2}{3+2\kappa}},$$

where c_κ is an absolute constant that only depends on κ, as claimed. ■