INFORMATION, COMPLEXITY AND STRUCTURE IN CONVEX OPTIMIZATION

A Thesis
Presented to
The Academic Faculty

by

Cristóbal A. Guzmán Paredes

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Algorithms, Combinatorics, and Optimization

School of Industrial and Systems Engineering
Georgia Institute of Technology
May 2015

INFORMATION, COMPLEXITY AND STRUCTURE IN CONVEX OPTIMIZATION

Approved by:

Professor Arkadi Nemirovski,
Co-advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Alexandre d'Aspremont
Département d'Informatique
École Normale Supérieure

Professor Sebastian Pokutta,
Co-advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Santosh Vempala
School of Computer Science
Georgia Institute of Technology

Professor Shabbir Ahmed
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: 30 March 2015

ACKNOWLEDGEMENTS

I am indebted to a number of remarkable people that supported the present work in different forms. First and foremost, I would like to thank my advisors, Dr. Arkadi Nemirovski and Dr. Sebastian Pokutta, for their guidance, intellectual insights, and everlasting patience that were crucial for my academic maturation. I would also like to thank all the members of the thesis committe: Dr. Shabbir Ahmed, Dr. Alexandre d'Aspremont, and Dr. Santosh Vempala, for their interest in this work.

I have been extremely fortunate to collaborate with outstanding researchers during my Ph.D. years. In this regard, I warmly thank Dr. Alexandre d'Aspremont, Dr. Gabor Braun, Dr. Roberto Cominetti, Dr. Vitaly Feldman, Dr. Frank Vallentin, and Dr. Jan Vondrak. I would also like to extend my gratitude to the Institute of Applied Mathematics at TU Delft, Centrum Wiskunde & Informatica, and the Theory Group at the IBM Almaden Research Center for the opportunity to work with them during summer internships and visits.

I want to thank (current and former) faculty at the ACO program, particularly Dr. Nina Balcan, Dr. William J. Cook, Dr. Santanu Dey, Dr. Christian Houdré, Dr. Vladimir Koltchinskii, Dr. Marco Molinaro, Dr. Prasad Tetali, and Dr. Robin Thomas; for excellent courses, seminars, and the vibrant academic environment at Georgia Tech. Furthermore, I would also like to thank my classmates: Abhishek Banerjee, Niao He, Arefin Huq, Arindam Khan, Chun-Hung Liu, Nolan Leung, Aurko Roy and Daniel Zink; for valuable discussions, friendship, and the organization of the ACO Student Seminar. Finally, I would like to thank all the staff of ISyE, specially Pam Morrison, Mark Reese and Yvonne Smith.

Life during these five years wouldn't have been the same without the warming company of great people. Specially during my first year, my adaptation in Atlanta was smooth and exciting thanks to Gustavo Angulo, Rodolfo Carvajal, Tamara Gutierrez, Diego Morán, Pamela Rosas and Timothy Sprock. I would also like to thank Felipe Castillo, Natalia Castillo, Andrés Iroumé, Guido Lagos, Álvaro Lorca, Alejandra Parrao, Ana Rojas, Daniel Silva, and all the Chilean and Latin American community in Georgia Tech and Alpharetta.

Finally, I would like to thank my parents: Ester Paredes, Patricio Guzmán, and my brothers Patricio and Francisco; thank you for being there, despite the distance. I want to specially dedicate this thesis to my wife, Lisset Manzano: Thank you for your wisdom, patience and love. Without you this wouldn't have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

Symbol	Description
$\mathbf{E}$	Real vector space/subspace
$\|\cdot\|$	Norm
$(\mathbf{E}, \|\cdot\|)$	Normed (real) vector space
$\|\cdot\|_p$	p-norm: $\|x\|_p = (\sum_i \|x_i\|^p)^{1/p}$
ℓ_p^n	Real vector space $(\mathbb{R}^n, \|\cdot\|_p)$
X	Convex closed set
$B_{\|\cdot\|}(x, R)$	Ball of radius R centered at x of space $(\mathbf{E}, \|\cdot\|)$
$B_{\|\cdot\|}$	Unit ball of a space $(\mathbf{E}, \|\cdot\|)$
B_p^n	Unit ball of ℓ_p^n
$x = (x_1, \ldots, x_n)$	n-dimensional vector
$x^1, \ldots, x^T$	Sequence of n-dimensional query points
$[n]$	Set of numbers $\{1, \ldots, n\}$
T	Number of iterations
t	Iterate
x	feasible vector
f	instance, or objective function
x^*	optimal solution
$f^*, \mathrm{Opt}(f)$	optimal value
$\mathcal{F}_{\|\cdot\|}(\kappa, L)$	Class of (κ, L)-smooth functions w.r.t. norm $\|\cdot\|$
$\mathcal{F}_p(\kappa, L),\ \mathcal{F}_p^n(\kappa, L)$	Class of (κ, L)-smooth functions w.r.t. ℓ_p^n-norm

$\mathcal{P} = (\mathcal{F}, X)$	Class of minimization problems over domain X, where $f \in \mathcal{F}$
$I(E)$	Indicator function of event E
$\mathbb{H}[X]$	Entropy of random variable X
$\mathbb{H}[X \mid Y]$	Conditional entropy of X given Y
$\mathbb{I}[X;Y]$	Mutual information between X and Y
$\mathbb{I}[X;Y \mid Z]$	Conditional mutual information between X and Y given Z
$s^{\oplus(i)}$	String obtained by flipping the i-th bit and removing all following bits of string s
$s \sqsubseteq t$	Relation string s is a prefix of string t
$s \parallel t$	Relation neither string s or t is a prefix of the other
$s\vert_l$	The prefix of s consisting of its first l bits
$s0$, $s1$	Strings obtained by appending a 0 or 1 to s, respectively
$\perp$	Empty string

SUMMARY

This thesis is focused on the limits of performance of large-scale convex optimization algorithms. Classical theory of oracle complexity, first proposed by Nemirovski and Yudin in 1983, successfully established the worst-case behavior of methods based on local oracles (a generalization of first-order oracle for smooth functions) for nonsmooth convex minimization, both in the large-scale and low-scale regimes; and the complexity of approximately solving linear systems of equations (equivalent to convex quadratic minimization) over Euclidean balls, under a matrix-vector multiplication oracle.

Our work extends the applicability of lower bounds in two directions:

- **Worst-Case Complexity of Large-Scale Smooth Convex Optimization:** We generalize lower bounds on the complexity of first-order methods for convex optimization, considering classes of convex functions with Hölder continuous gradients. Our technique relies on the existence of a *smoothing kernel*, which defines a smooth approximation for any convex function via infimal convolution. As a consequence, we derive lower bounds for ℓ_p/ℓ_q-setups, where $1 \leq p, q \leq \infty$, and extend to its matrix analogue: Smooth convex minimization (with respect to the Schatten q-norm) over matrices with bounded Schatten p-norm.

  The major consequences of this result are the near-optimality of the Conditional Gradient method over box-type domains ($p = q = \infty$), and the near-optimality of Nesterov's accelerated method over the cross-polytope ($p = q = 1$).

- **Distributional Complexity of Nonsmooth Convex Optimization:** In this

work, we prove average-case lower bounds for the complexity of nonsmooth convex optimization. We introduce an information-theoretic method to analyze the complexity of oracle-based algorithms solving a random instance, based on the *reconstruction principle*.

Our technique shows that all known lower bounds for nonsmooth convex optimization can be derived by an *emulation procedure* from a common *String-Guessing Problem*, which is combinatorial in nature. The derived average-case lower bounds extend to hold with high probability, and for algorithms with bounded probability error, via Fano's inequality.

Finally, from the proposed technique we establish the equivalence (up to constant factors) of distributional, randomized, and worst-case complexity for black-box convex optimization. In particular, there is no gain from randomization in this setup.

CHAPTER I

INTRODUCTION

First-order algorithms are the methods of choice when solving extremely large-scale convex optimization problems, the reasons being twofold. First, in the large scale case, an iteration of a first order method is typically much computationally cheaper than the iteration of (the only, as far as constrained problems are concerned) competitors – Interior Point methods. Second, under favorable circumstances, first order methods exhibit *dimension-independent* rate for convergence. Albeit sublinear, this dimension-independent rate makes first order algorithms well suited for large scale convex optimization, provided medium accuracy solutions are sought.

Given the practical importance of first order algorithms, understanding theoretical limits of their performance is a truly important avenue of research. A commonly adopted way to pose this question is offered by Information-Based Complexity Theory and is based on *local oracle* model of solution algorithms. In this model, we consider gradient information as provided by an oracle, and we want to design algorithms capable to generate approximate solution of a desired quality via a minimax optimal number of oracle calls (minimum over solution algorithms, maximum over problem instances from a given family), disregarding other computational aspects (such as the computational expenses of processing oracle's answers). The main features of this model responsible for meaningful results are

• *locality* of the oracle, meaning that the oracle's answer is uniquely defined by the behavior of the objective and constraints of the queried instance in an "infinitesimally small" neighborhood of the query point, and

• focusing on wide enough families of instances, in order to avoid the situation when a small number of calls to the oracle allows to identify the instance we are solving within the family[1]

While both these requirements by themselves are rather restrictive[2], they are well suited to investigating first order algorithms, since these algorithms by their nature are oriented at local *first order* oracles[3]; as a matter of fact, over the years Information-Based Complexity has made a significant impact on convex optimization techniques.[4]

In this thesis we develop novel techniques for analyzing the oracle complexity of convex optimization, with emphasis on deriving *lower* complexity bounds for families of problems with *non-Euclidean geometry*, where the smoothness of the objectives and sizes/geometries of feasible domains are quantified w.r.t. $\|\cdot\|_p$-norms (and their "noncommutative" matrix analogies) on the argument space, with emphasis on the cases of $p = 1$ and $p = \infty$ rather than on the Euclidean case of $p = 2$. Note that the case of non-Euclidean geometry is of primary importance for state-of-the-art applications in signal processing and machine learning.

The rest of the Introduction is structured as follows: First, in Section 1.1 we detail the precise objectives of our work. Next, in Section 1.2 we outline the required basics on optimization algorithms and oracle complexity, and summarize known upper and lower bounds on complexity of convex optimization. Finally, in Section 1.3 we describe the specific contributions of this work.

[1] Indeed, since the approach in question ignores computational cost of processing the acquired information, the number of oracle calls needed to solve an instance to arbitrary accuracy can be only smaller than the number of calls needed to identify the instance in the family.

[2] in convex programming, we usually possess complete information on problem instance from the very beginning – how else could we know that the instances are convex?

[3] i.e., oracles returning values and gradients of the objective at the constraints at query points.

[4] As a most striking example, note that the discovery of Nesterov's Fast Gradient algorithms was stimulated by the desire to "bridge" lower and upper oracle complexity bounds for smooth convex optimization.

1.1 Outline of Goals

I. *Derive tight lower complexity bounds and identify nearly-optimal algorithms for convex optimization:* The most important objective of complexity analysis is understanding limits of performance of optimization algorithms and identifying algorithms with theoretically optimal, or nearly so, performance. In this thesis we establish results of this type, some of them surprising, for several well-known algorithms in setups motivated by modern applications.

II. *Gain insights for algorithm design:* Whenever a lower complexity bound is far off the complexity bounds of known algorithms, one may wonder whether this is because the bound is too weak, or the existing algorithms are far from being optimal. The second possibility over the years proved to be, and still is a powerful stimulus for algorithmic design. In this thesis we identify several regimes where presumably there is room for algorithmic improvements.

III. *Obtain stronger guarantees of hardness for problems:* From a practical point of view, there are situations were worst-case analysis might be too conservative to give a realistic view on complexity. For example, this is the case in learning applications, where an algorithm can use *a priori* knowledge about problem parameters, and further update such priors with new information. In the present thesis we provide novel tools to analyze hardness in such situations, by a *distributional complexity* viewpoint. Our techniques lead to a better understanding of average-case analysis and the role of randomization in convex optimization.

1.2 Background

The main object of study in this thesis are convex optimization problems of the form

$$\mathrm{Opt}(f) = \min_{x \in X} f(x) \tag{1}$$

where X is a compact convex subset of a (finite-dimensional) Euclidean space $\mathbf{E}$, and $f : \mathbf{E} \to \mathbb{R}$ is a convex function from some family, usually specified by smoothness parameters of the objectives f (Hölder exponent of ∇f and the corresponding constant) taken with respect to some norm $\|\cdot\|$ on $\mathbf{E}$.[5] We start with introducing the families $\mathcal{F}$ we intend to work with.

1.2.1 Normed Euclidean spaces

The design variables in our optimization problems run trough a Euclidean space $\mathbf{E}$ equipped with some norm $\|\cdot\|$ (not necessarily the Euclidean one); thus, the first component of our setup is a pair $(\mathbf{E}, \|\cdot\|)$ comprised of a Euclidean space and a norm on this space. The inner product on $\mathbf{E}$ will be denoted $\langle \cdot, \cdot \rangle$, and this inner product allows to identify $\mathbf{E}$ and its dual space, so that the norm dual (a.k.a. conjugate) to $\|\cdot\|$ turns out to be a norm on $\mathbf{E}$:

$$\|\xi\|_* = \max_{x \in \mathbf{E}, \|x\| \le 1} \langle \xi, x \rangle : \mathbf{E} \to \mathbb{R}. \tag{2}$$

We will be especially interested in the standard finite-dimensional L_p-spaces $\ell_p^n = (\mathbb{R}^n, \|\cdot\|_p)$, $1 \le p \le \infty$, where

$$\|x\|_p = \begin{cases} \left(\sum_i |x_i|^p\right)^{1/p}, & 1 \le p < \infty \\ \max_i |x_i|, & p = \infty \end{cases} \qquad [x = [x_1; ...; x_n] \in \mathbb{R}^n]$$

[5] The spaces $\mathbf{E}$ we work with are finite dimensional, so that the property of the feasible set X to be convex and compact is independent of the particular choice of the norm. The same is true for the property ∇f to be Hölder continuous with some exponent; this choice, however, affects the value of the Hölder constant of ∇f.

the norm conjugate to $\|\cdot\|_p$ is $\|\cdot\|_{p_*}$, with

$$p_* = \frac{p}{p-1}.$$

Another family of finite-dimensional normed spaces we will be interested in is the family of *Schatten p-spaces* $\mathrm{Sch}_p^n = (\mathbb{R}^{n\times n}, \|\cdot\|_{\mathrm{Sch},p})$, where the *Schatten p-norm* of an $n \times n$ real matrix x is $\|\sigma(x)\|_p$, $\sigma(x)$ being the vector of singular values of x. The norm dual to $\|\cdot\|_{\mathrm{Sch},p}$ is $\|\cdot\|_{\mathrm{Sch},p_*}$, with the same p_* as above.

1.2.2 Families of convex functions

Let $(\mathbf{E}, \|\cdot\|)$ be a Euclidean normed space. Given a real $L > 0$, we denote by $\mathrm{Lip}_{\mathbf{E},\|\cdot\|}(L)$ the family of all *convex* Lipschitz continuous, with constant L w.r.t. the norm $\|\cdot\|$, functions on $\mathbf{E}$:

$$\mathrm{Lip}_{\mathbf{E},\|\cdot\|}(L) = \{f : \mathbf{E} \to \mathbb{R} : \ |f(x) - f(y)| \le L\|x - y\| \ \forall (x, y) \in \mathbf{E} \ \& \ f \text{ is convex}\}.$$

Given $L > 0$ and $\kappa \in (1, 2]$, we denote by $\mathcal{F}_{\mathbf{E},\|\cdot\|}(\kappa, L)$ the family of all differentiable convex functions on E with Hölder continuous, with exponent $(\kappa - 1)$ and constant L w.r.t. $\|\cdot\|$, gradient ∇f:

$$\mathcal{F}_{\mathbf{E},\|\cdot\|}(\kappa, L) = \left\{ f : \mathbf{E} \to \mathbb{R} : \begin{array}{c} \|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|^{\kappa-1} \ \forall (x, y) \in \mathbf{E} \\ \& \ f \text{ is convex} \end{array} \right\}.$$

when $(\mathbf{E}, \|\cdot\|) = \ell_p^n$, we simplify the notation $\mathrm{Lip}_{\mathbf{E},\|\cdot\|}(L)$, $\mathcal{F}_{\mathbf{E},\|\cdot\|}(\kappa, L)$ to $\mathrm{Lip}_p^n(L)$, $\mathcal{F}_p^n(\kappa, L)$, and further omit superscript n when the dimension of $\mathbf{E}$ is clear from the context.

The introduced spaces of convex Lipschitz continuous functions, and of convex smooth functions – of various degrees of smoothness – are exactly the families of objectives for problems (1) we intend to consider.

1.2.3 Oracle complexity of optimization

The goal of this section is to present the notion of *oracle*, or *information based*, complexity of of a class of computational problems. The general theory of oracle algorithms and their complexity is known as *Information-Based Complexity Theory* (IBCT). Some standard references in this area are [44], [37]. For the specific case of oracle complexity of convex optimization, the standard reference is [32]; as a matter of fact, most of the results we will review in this section were proved there, and we closely follow that presentation. More modern developments dealt with stochastic oracles, online (regret) minimization, and exploiting specific problems' structures, without much progress in understanding of the traditional setting. In full accordance with the needs of this thesis, we define complexity-related notions in the context of optimization problems (1) we are interested in.

1.2.3.1 Classes of optimization problems, oracles, solution algorithms

Classes. (a.k.a families) of optimization problems we are about to consider are comprised of problems of the form (1) – *instances* – which share common feasible domain X, which by default will be a a nonempty compact convex set in Euclidean normed space $(\mathbf{E}, \|\cdot\|)$. Since the feasible domain is common for all instances, an instance can be identified with its objective f, and a class of problems – with a pair $\mathcal{P} = (\mathcal{F}, X)$, where $\mathcal{F}$ is a family of objectives f. It is convenient to assume that all these objectives are defined on the entire $\mathbf{E}$, so that $\mathcal{F}$ is just a particular family of real-valued functions on $\mathbf{E}$; by default, *all functions from $\mathcal{F}$ are convex* (and thus continuous). Usually the feasible domain X of instances from the class under consideration will be fixed by the context, and we shall refer to an instance with objective f as to *problem f*. Note that under our default assumptions, which include compactness of X and imply continuity of f, all instances are solvable.

Approximate solutions and their accuracy. For wide enough classes of optimization problems we are about to consider, it does not make sense to require from a solution algorithm finding an exactly optimal solution in finite time, thus we should speak about complexity of finding *approximate* solutions of a given quality and should therefore agree how to quantify the quality of a candidate solution. As explained in [32] (see also [31]), there are deep reasons to quantify the quality of a candidate solution by its *residual in terms of the objective*; specifically, given $\varepsilon > 0$, we say that a candidate solution $x \in \mathbf{E}$ is ε-solution to (1), if this solution is feasible – $x \in X$ – and

$$f(x) - \mathrm{Opt}(f) \leq \varepsilon.$$

The accuracy $\varepsilon(x|f, X)$ of a candidate solution to problem (1) is defined as $+\infty$ when $x \notin X$ and as $f(x) - \mathrm{Opt}(f)$ otherwise, i.e., as the smallest ε for which x is an ε-solution to the problem.

Solution algorithms and oracles. In IBCT, when speaking about algorithms for solving problems from a given class $(\mathcal{F}, X)$, it is assumed that the algorithm knows the class in advance, but does not know the objective f of a particular instance the algorithm is applied to; thus, in order to build a solution of a prescribed accuracy ε, the algorithm should "learn" f, specifically, by calls to an *oracle* $\mathcal{O}$. An oracle is defined as a mapping

$$\mathbf{E} \times \mathcal{F} \ni (x, f) \to \mathcal{O}_f(x) \in \mathcal{I},$$

where $\mathcal{I}$ is some *information space*. When processing problem f, an algorithm $\mathcal{B}$ queries the oracle at subsequent *search points* $x^1, x^2, \ldots$, with $x^t \in \mathbf{E}$ depending solely on the information returned by the oracle as queried at the previous search points. Thus, x^1 depends solely on an algorithm $\mathcal{B}$; x^2 is some function of x^1 and $\mathcal{O}_f(x^1)$; x^3 is some function of x^1, $\mathcal{O}_f(x^1)$, x^2, $\mathcal{O}_f(x^2)$, etc. Thus, an algorithm $\mathcal{B}$ is

a collection of *search rules*

$$\mathcal{B}^t : (\mathbf{E} \times \mathcal{I})^{t-1} \to \mathbf{E},$$

and

$$x^t = \mathcal{B}^t(x^1, \mathcal{O}_f(x^1), \ldots, x^{t-1}, \mathcal{O}_f(x^{t-1})) \tag{3}$$

When speaking about deterministic algorithms (which is by default in the sequel), the search rules are deterministic functions of their arguments; in the IBCT framework, there are no restrictions on the computational complexity of these rules.

For complexity-oriented purposes, it suffices to restrict ourselves with *T-step algorithms*, where $T = 1, 2, \ldots$; a T-step algorithm $\mathcal{B}^T$ runs recurrence (3) for $t = 1, ..., T$, and the last search point $x^T = x^T[\mathcal{B}^T, f]$ is considered as the approximate solution generated by the algorithm $\mathcal{B}^T$ as applied to problem f.

In some cases, we will be interested in *randomized algorithms*, where the right hand sides in the search rules (3), aside of the arguments shown in (3), depend on random parameters s^t which without loss of generality can be assumed to be independent over t and uniformly distributed on $[0, 1]$, see [32]. When speaking about randomized algorithms, it makes sense to allow for a random and instance-dependent number of steps (i.e., oracle calls), rather than to think of this number as instance-independent and fixed by the description of the algorithm[6]. To this end, we augment the search rules by *termination rules* which, depending on the trajectory $x^1, \mathcal{O}_f(x^1), ..., x^{t-1}, \mathcal{O}_f(x^{t-1})$ and on s^t, determine when to terminate the search process and how to generate the resulting approximate solution. Thus, a randomized algorithm $\mathcal{B}$ is a collection of search *and termination* rules parameterized by step (iteration) number $t \in \mathbb{N}$, with the rules corresponding to step t

[6] A deterministic algorithm also could be allowed to generate approximate solution in course of an instance-dependent number of steps; this option, however, does not affect the worst-case risk and complexity, as defined below, which are the entities we are ultimately interested in, and we lose nothing by treating the number of steps as an instance-independent parameter of a solution algorithm.

being deterministic functions of $x^1, \mathcal{O}_f(x^1), \ldots, x^{t-1}, \mathcal{O}_f(x^{t-1}), s^t$ taking values in $\mathbf{E}$ (search rules) and in $\{\text{"stop"}, \text{"continue"}\} \times \mathbf{E}$ (termination rules); the algorithm terminates at the first step t for which the "stop"/"continue" component of the output of the termination rule is "stop", and the resulting approximate solution is the $\mathbf{E}$-component of this output. For a randomized algorithm $\mathcal{B}$ and $f \in \mathcal{F}$, we denote by $T[\mathcal{B}, f]$ and $x[\mathcal{B}, f]$ the (random) number of steps and approximate solution generated by $\mathcal{B}$ as applied to problem f.

Local oracles. An oracle $\mathcal{O}$ for a family $\mathcal{F}$ of functions on $\mathbf{E}$ is called *local*, if, when queried at a point x about $f \in \mathcal{F}$, the information returned by the oracle is fully determined by the behavior of f in an "infinitesimal neighborhood of x," meaning that whenever $x \in \mathbf{E}$ and $f, g \in \mathcal{F}$ are such that $f \equiv g$ in some neighborhood (perhaps depending on f, g, x) of x, one has

$$\mathcal{O}_f(x) = \mathcal{O}_g(x).$$

The most important for us local oracle is *the first order oracle* which, as queried about f at a point x, returns the value $f(x)$ and the subdifferential $\partial f(x)$ of f at x (recall that all our families are comprised of convex real-valued functions on $\mathbf{E}$, and thus any such f has nonempty subdifferential at every point). Along with *the* first order oracle, we can consider oracles which, as queried about f at x, return the value $f(x)$ and *a* subgradient $f'(x) \in \partial f(x)$ of f at x. When $\mathcal{F}$ is comprised of continuously differentiable functions, all oracles of this type coincide with *the* first order oracle as defined above; in contrast, when $\mathcal{F}$ contains nonsmooth functions, there are many first order oracles for $\mathcal{F}$, and not all of them are local (since the selection of the subgradient $f'(x)$ in the subdifferential $\partial f(x)$ could not necessarily depend solely on the local behavior of f at x).

It is worthy of mentioning that for every family $\mathcal{F}$ of functions on $\mathbf{E}$, there exists the "most powerful" local oracle, the *universal* one; as queried about $f \in \mathcal{F}$

at $x \in \mathbf{E}$, the universal oracle returns the equivalence class of f with respect to the equivalence relation on $\mathcal{F}$ given by

$$f \sim g \Leftrightarrow \text{ there exists a neighborhood } V \text{ of } x \text{ such that } f \equiv g \text{ on } V.$$

Clearly, the universal oracle for $\mathcal{F}$ can emulate any local oracle, meaning that the answers of a given local oracle $\mathcal{O}$ as queried about $f \in \mathcal{F}$ at $x \in \mathbf{E}$ can be obtained by deterministic transformation (possibly depending on x) from the answer of the universal oracle, as queried about f at x.

1.2.3.2 Risk and complexity

Given a problem class $\mathcal{P} = (\mathcal{F}, X)$ and an oracle $\mathcal{O}$ for $\mathcal{F}$, IBCT defines the *minimax T-risk* of the class w.r.t. the oracle as the function

$$\mathrm{Risk}(T) = \inf_{\mathcal{B}^T} \sup_{f \in \mathcal{F}} \varepsilon(x^T[\mathcal{B}, f], f) : \mathbf{N} \to \mathbb{R}_+,$$

where the infinum is taken over all T-step algorithms utilizing oracle $\mathcal{O}$. Thus, $\mathrm{Risk}(T) = \varepsilon$ means that, first, whenever $\varepsilon' > \varepsilon$, there exists a T-step algorithm, utilizing oracle $\mathcal{O}$, which, as applied to every problem instance from the class $\mathcal{P}$, in T steps generates an ε'-solution to the instance; and that, second, for every $\varepsilon' < \varepsilon$ and every T-step algorithm $\mathcal{B}_T$ utilizing oracle $\mathcal{O}$, there exists a "hard" instance $f \in \mathcal{F}$, meaning that for the approximate solution $x^T = x^T[\mathcal{B}^T, f]$ generated by $\mathcal{B}^T$ as applied to the instance we have $\varepsilon(x^T) > \varepsilon'$.

The inverse of the risk is called the *complexity* (full name: "ε-complexity of problem class $\mathcal{P}$ w.r.t. oracle $\mathcal{O}$"). This is the function $\mathrm{Compl}(\varepsilon)$ defined as

$$\mathrm{Compl}(\varepsilon) = \min \left\{ T : \begin{array}{c} \text{there exists } T\text{-step algorithm } \mathcal{B}^T \text{ utilizing oracle } \mathcal{O} \\ \text{and such that } \varepsilon(x^T[\mathcal{B}^T, f], f) \le \varepsilon \text{ for all } f \in \mathcal{F} \end{array} \right\}.$$

Thus, the claim that the complexity, as evaluated at some ε, of a class $\mathcal{P}$ w.r.t. an oracle $\mathcal{O}$ is equal to some T means that there exists a T-step method, utilizing oracle

$\mathcal{O}$, capable to solve every instance from the class within accuracy ε, and there is no $(T-1)$-step method with the same property.

1.2.3.3 *Oracle complexity: extensions*

In this thesis we will be interested in broader notions of complexity than the just defined worst-case one. We have already introduced the notion of a randomized algorithm, and our first task is to define complexity in this framework; what we intend to do is to use the minimax value (min over algorithms, max over problem instances) of the *expected* number of steps before an approximate solution of a required quality is achieved. Specifically, given problem class $\mathcal{P} = (\mathcal{F}, X)$, oracle $\mathcal{O}$, and $\varepsilon > 0$, we define the *randomized ε-complexity* $\mathrm{Compl}_{\mathcal{R}}(\varepsilon)$ of $\mathcal{P}$ w.r.t. $\mathcal{O}$ as follows. We define $\mathfrak{B}[\varepsilon]$ as the family of all randomized algorithms $\mathcal{B}$, utilizing oracle $\mathcal{O}$, for which the approximate solution $x[\mathcal{B}, f]$ is, for every $f \in \mathcal{F}$, and ε-solution to problem f:

$$\varepsilon(x[\mathcal{B}, f], f) \leq \varepsilon \quad \forall f \in \mathcal{F}$$

The *randomized complexity* is then defined as

$$\mathrm{Compl}_{\mathcal{R}}(\varepsilon) = \inf_{\mathcal{B} \in \mathfrak{B}[\varepsilon]} \sup_{f \in \mathcal{F}} \mathbf{E}\{T(\mathcal{B}, f)\}, \tag{4}$$

where $T(\mathcal{B}, f)$ is the (random) number of steps of $\mathcal{B}$, as applied to problem f, before termination, and the expectation is taken over the realizations of $\mathcal{B}$ as applied to f.

Another important measure we will study is the *distributional complexity* (full name: "ε-distributional complexity of problem class $\mathcal{P}$ w.r.t. oracle $\mathcal{O}$") defined as follows:

$$\mathrm{Compl}_{\mathcal{D}}(\varepsilon) = \sup_{\mathbb{P} \in \mathfrak{P}} \inf_{\mathcal{B} \in \mathfrak{B}[\varepsilon]} \int_{\mathcal{F}} \mathbf{E}\{T[\mathcal{B}, f]\} \mathbb{P}(df), \tag{5}$$

where $\mathfrak{P}$ is the family of all probability distributions $\mathbb{P}$ on $\mathcal{F}$, and $\mathfrak{B}[\varepsilon]$, same as

above, is the family of randomized algorithms guaranteed to terminate with ε-solutions, for arbitrary instance from the class $\mathcal{B}$. The notion of distributional complexity corresponds to the case where we believe that "in reality" the problems from the class $\mathcal{P}$ (i.e., the objectives f from $\mathcal{F}$) are generated at random according to some probability distribution $\mathbb{P}$, and we can adjust a solution algorithm to this distribution; the outer supremum reflects the fact that we want to end up with a characteristic of $\mathcal{P}$ (and $\mathcal{O}$), and not with something depending on the distribution of instances.

We will also consider the *high-probability oracle complexity* of $\mathcal{P}$ w.r.t. $\mathcal{O}$ defined as

$$\mathrm{Compl}^{\beta}_{\mathcal{HP}}(\varepsilon) = \sup_{\mathbb{P}\in\mathfrak{P}} \inf_{\mathcal{B}\in\mathfrak{B}[\varepsilon]} \inf\{\tau : \mathbb{P}\{f : T[\mathcal{B}, f] \le \tau\} \ge 1-\beta\}$$

with the same as above $\mathfrak{P}$ and $\mathfrak{B}[\varepsilon]$, and $\beta \in (0,1)$ is "reliability parameter;" this is what we get from the distributional complexity when passing from quantifying the "typical," w.r.t. a distribution $\mathbb{P}$ on instances, running time $T[\mathcal{B}, f]$ of an algorithm $\mathcal{B}$ by the upper $(1-\beta)$-quantile of this time rather than by its expected value.

For $\mathcal{P}$ and $\mathcal{O}$ fixed, we clearly have

$$\mathrm{Compl}_{\mathcal{D}}(\cdot) \le \mathrm{Compl}_{\mathcal{R}}(\cdot) \le \mathrm{Compl}(\cdot),$$

implying that a lower bound on a weaker notion of complexity in this chain automatically lower-bounds a stronger notion of complexity.

1.2.3.4 Oracle complexity: motivation and impact

Information-Based Complexity Theory is a well established area of theoretical research with rich body of diverse and highly nontrivial results answering challenging and natural theoretical questions. As far as optimization is concerned, the oracle model for algorithms has a wide scope capturing all known "broad scope" methods of Nonlinear Optimization. In particular, this model is well suited for first

order optimization methods which are in the focus of this thesis. This being said, when defining oracle complexity, one ignores the computational effort needed to process the oracle answers and run the optimization process. As a result, the oracle complexity $\mathrm{Compl}(\cdot)$ is a *lower bound*, potentially highly biased, on the true computational effort of solving problems from a given class within a given accuracy. It should be stressed that, as a matter of fact, this is the only known source of nontrivial lower complexity bounds in Nonlinear Optimization – when passing to more adequate complexity models, like counting the total number of real arithmetic operations needed to solve problems from a given class within desired accuracy, no nontrivial lower complexity bounds are known, and this is so even for problems as simple as minimizing quadratic forms, not speaking about Linear and Semidefinite Optimization. On the other hand, lower complexity bounds, even as biased as those of oracle complexity, are extremely valuable in algorithmic design. Indeed, when the lower bound on oracle complexity matches an upper complexity bound associated with a particular algorithm (this indeed is the case for many problem classes), this is a strong argument in favor of the search strategy utilized in the algorithm; whenever this is case, one can safely focus on implementation, with the goal to reduce the computational effort per iteration. On the other hand, significant gap between the best known lower bounds on oracle complexity of a particular class and complexity of existing solution algorithms suggests that the search strategies in question are far from being optimal and they should be improved – this is exactly what happened with the discovery of Fast Gradient Methods [34] (Nesterov, 1983) which now form the backbone of large-scale smooth and composite convex minimization and form the main component in the optimization toolbox for signal processing. This discovery was stimulated by the gap between the best known risk lower bound $O(1/T^2)$ [7] in smooth large-scale convex optimization and the $O(1/T)$

[7] From now on, every use of $O(1)$ denotes a positive absolute constant.

rate of convergence of the best algorithms known at the time; this gap suggested severe non-optimality of traditional methods and stimulated systematic research on their improvement culminating in Nesterov's discovery of $O(1/T^2)$-converging algorithms.

We believe that this discussion motivates sufficiently well our research agenda, that is, investigating worst case oriented oracle complexity of large-scale smooth convex optimization problems and with non-Euclidean geometry and distributional complexity of nonsmooth convex optimization.

1.2.4 Oracle complexity of convex optimization: overview of known results

This thesis is focused on deriving novel lower bounds on oracle complexity of convex optimization; to put our results in proper perspective allowing to judge on novelty and tightness of our results, we summarize here the results on the oracle complexity of convex optimization known from the literature. As a matter of fact, all these results deal with worst case oriented complexity, and this is what we call "complexity" in this section. In accordance with the subject of this thesis, we restrict our summary to the results on complexity of problem classes associated with broad families of convex objectives described in Section 1.2.2, specifically, the families $\mathrm{Lip}_{\mathbf{E},\|\cdot\|}(L)$ of convex Lipschitz continuous, and the families $\mathcal{F}_{\mathbf{E},\|\cdot\|}(\kappa, L)$ of smooth convex objectives on Euclidean normed space $(\mathbf{E}, \|\cdot\|)$; we refer to these two cases as to *nonsmooth* and *smooth* ones, respectively.

It turns out that the risk of our classes of convex minimization problems exhibit different behavior depending on whether the number of steps T is large as compared to the dimension n of the space of variables, namely, $T \geq O(1)n\ln(n)$ ("Low-scale regime") or it is not the case ("Large-scale regime" $T \leq O(1)n\ln(n)$).

1.2.4.1 Low-Scale Regime

In the low-scale regime, the basic complexity results can be summarized as follows:

Theorem 1.2.1. [32] *Let* $(\mathbf{E}, \|\cdot\|)$ *be a normed Euclidean space,* n *be the dimension of* $\mathbf{E}$*, and* X *be a convex compact subset of* $\mathbf{E}$*. Given* $L > 0$*, let* $\mathcal{P}$ *be the class of all convex problems* (9) *with Lipschitz continuous, with constant* L *w.r.t.* $\|\cdot\|$*, objectives* $f : \mathbf{E} \to \mathbb{R}$*, and let* $R = R_{\|\cdot\|}(X)$ *be the* $\|\cdot\|$*-diameter of* X*:*

$$R_{\|\cdot\|}(X) = \max_{x,y \in X} \|x - y\|.$$

Then

(i) *The complexity of* $\mathcal{P}$ *w.r.t. every first order oracle can be upper bounded as*

$$0 < \varepsilon \quad \Rightarrow \quad \mathrm{Compl}(\varepsilon) \leq O(1) n \ln\left(\frac{LR + 2\varepsilon}{\varepsilon}\right). \tag{6}$$

(ii) *When* X *contains* $\|\cdot\|$*-ball of radius* $r > 0$*, the complexity of* $\mathcal{P}$ *w.r.t. every local oracle can be lower bounded as*

$$0 < \varepsilon \leq \widehat{\varepsilon} \quad \Rightarrow \quad \mathrm{Compl}(\varepsilon) \geq O(1) n \ln\left(\frac{Lr}{\varepsilon}\right), \tag{7}$$

with properly selected $\widehat{\varepsilon} \geq O(1) n^{-1}$ *depending on* X *and* $\|\cdot\|$*.*

Several comments are in order:

(i) The upper bound (6) is yielded by the Center of Gravity method [25], [36]; this method, however, of purely academic interest, since its implementation requires computing centers of gravity of general type convex sets (say, general polytopes, if X is a polytope), which is a computationally intractable task, at least as far as deterministic computations are concerned. The first implementable method with polynomial in n and $\ln(1/\varepsilon)$ upper complexity bound $O(1) n^2 \ln\left(\frac{LRn+2\varepsilon}{\varepsilon}\right)$) was the Ellipsoid algorithm (Nemirovski & Yudin, 1976 [32]; Shor, 1977 [41]). An algorithm with complexity bound (6) and polynomial (provided X is a polytope) arithmetic complexity of iteration – the *Inscribed Ellipsoid Method* – was proposed in [23].

(ii) In terms of risk rather than complexity, (6) reads

$$\text{Risk}(t) \leq O(1)LR\exp\{-O(1)t/n\}, \quad t = 1, 2, ... \tag{8}$$

thus exhibiting linear convergence with the *non-asymptotic* convergence ratio $\exp\{-O(1)/n\}$.

(iii) When the *relative accuracy* $\omega := \varepsilon/(LR)$ is small, upper and lower complexity bounds in Theorem 1.2.1 are within absolute constant factor of each other. What small means, it depends on how "close" to a $\|\cdot\|$-ball is X. For example, when $R/r = O(1)$, then $\omega \leq O(1)/n$ is small. When X is a $\|\cdot\|$-ball, the range of values of ω where the bounds are within an absolute constant of each other depends on the geometry of the norm $\|\cdot\|$; say, when $\|\cdot\| = \|\cdot\|_p$, this range is $\omega \leq O(1)n^{-1/p}$.

(iv) Passing from nonsmooth to smooth case does not affect the asymptotic, $\omega \to +0$, behavior $O(n\ln(1/\omega))$ of complexity, it affects only the range of the values of the accuracy where this behavior indeed takes place.

1.2.4.2 Large-Scale Regime

We have seen that with a first order oracle available, the complexity of convex minimization, in smooth and nonsmooth cases alike, can be upper-bounded by $O(1)n\ln(A/\varepsilon)$, where the scale factor A is readily given by the description of the problem class, and that the asymptotic, $\varepsilon \to +0$, of the complexity is $O(1)n\ln(1/\varepsilon)$. The good news here is that *unless A is astronomically large* (which normally is not the case), *we can generate high accuracy solutions in a "polynomial time" fashion, with $O(\dim \mathbf{E})$ calls to the first order oracle per accuracy digit*, and this is the best we can act, as far as broad problem classes in question are concerned. A clear downside of this fact is that according to the upper bound, the cost of an accuracy digit, whether the first or the thousandth, is $O(n)$ oracle calls; were it indeed the true complexity,

our practical possibilities to solve in reasonable time large-scale problems from the classes under consideration would be nonexisting. Fortunately, it turns out that convex problems with *favorable geometry* can be solved within *moderate* accuracy with *dimension-independent*, or nearly so, complexity. Moreover, in contrast to what happens when $\varepsilon \to 0$, the complexity of finding medium-accuracy solutions is the smaller the better are smoothness properties of the objective. The summary of the related complexity results is as follows.

Preliminaries: smooth Euclidean normed spaces. Let $(\mathbf{E}, \|\cdot\|)$ be a Euclidean normed space, and let $\chi \geq 1$ and $r \in (1, 2]$. Following [22, section 2.3.1], we say that $(\mathbf{E}, \|\cdot\|)$ is (χ, r)-*smooth*, if there exists a convex continuously differentiable function $W(\cdot) : \mathbf{E} \to \mathbb{R}$ such that $W(0) = 0$, $W(\xi) \geq r^{-1}\|\xi\|_*^r$ for all ξ, and

$$\forall(\xi, \eta \in \mathbf{E}) : W(\xi + \eta) \leq W(\xi) + \langle \nabla W(\xi), \eta \rangle + \chi r^{-1}\|\eta\|_*^r.$$

For example, it is well known that ℓ_p-spaces ℓ_p^n is (χ, r)-smooth with the parameters χ, r and functions W as follows:

$$\begin{array}{lll} (a) & \text{for } 1 \leq p \leq 2: & r = 2,\ \chi = \min\left[\frac{1}{p-1}, 2\mathrm{e}\ln(n)\right],\ W(\xi) = \frac{1}{2}\|\xi\|_s^2 \\ & & \left[s = \min\left[\frac{p}{p-1}, \ln(n) + 1\right]\right]; \\ (b) & \text{for } 2 \leq p \leq \infty: & r = \max\left[\frac{p}{p-1}, \frac{\ln(n)+1}{\ln(n)}\right],\ \chi = 2\mathrm{e},\ W(\xi) = \frac{\mathrm{e}}{r}\|\xi\|_r^r. \end{array} \tag{9}$$

From the results of [20] if follows that the Schatten p-space Sch_p^n in the range $1 \leq p \leq 2$ is $(\chi, 2)$ smooth with the value of χ coinciding, within an absolute constant factor, with the one indicated in (9.*a*).

Situation. Let us fix a normed Euclidean space $(\mathbf{E}, \|\cdot\|)$ along with a convex compact and nonempty subset X in $\mathbf{E}$ and smoothness parameters $\kappa \in (1, 2]$, $L > 0$, and consider the class $\mathcal{P} = (\mathcal{F}, X)$ of problems (1) associated with X and the family of objectives $\mathcal{F} = \mathcal{F}_{\mathbf{E},\|\cdot\|}(\kappa, L)$, see Section 1.2.2. We equip the family $\mathcal{F}$ with the

first order oracle which, queried about $f \in \mathcal{F}$ at a point $x \in X$, returns $f(x)$ and $\nabla f(x)$.

Upper complexity bounds. It is known that the complexity of $\mathcal{P}$ can be *upper-bounded* in terms of smoothness parameters of $(\mathbf{E}, \|\cdot\|)$, $\|\cdot\|$-diameter of X and smoothness parameters κ, L of the objectives we are minimizing. It is convenient to express the bounds in terms of risk rather than complexity. The best known so far upper risk bounds for large-scale smooth convex optimization are given by the following result.

Theorem 1.2.2. [22, Section 2.3] *Let $(\mathbf{E}, \|\cdot\|)$ be (χ, r)-smooth, let $\kappa \in (1, 2]$, $L > 0$, $R > 0$, and let $X \subset \mathbf{E}$ be a convex compact set of $\|\cdot\|$-diameter not exceeding R. Then for the problem class $\mathcal{P}$ we have specified it holds for all $T \geq 1$*

$$\mathrm{Risk}(T) \leq O(1)[r_* \chi]^{\kappa/r} \frac{LR^\kappa}{T^{\kappa(1+1/r_*)-1}}. \tag{10}$$

The right hand side in (10) is the efficiency estimate of a (slightly modified) Nesterov's Fast Gradient method; the modification in question and the derivation of its efficiency estimate[8] (10) can be found in [22, Section 2.3].

Theorem 1.2.2 combines with the above information on the smoothness parameters of ℓ_p- and Schatten p-spaces to yield the following

Corollary 1.2.3. *Let $p \in [1, \infty]$, and consider the space $(\mathbf{E}, \|\cdot\|) = \ell_p^n$. Let $\kappa \in (1, 2]$, $L > 0$, $R > 0$, and let $X \subset \mathbf{E}$ be a convex compact set of $\|\cdot\|_p$-diameter not exceeding R. Then for the class $\mathcal{P}$ of problems of minimizing over X smooth convex functions from the*

[8] efficiency estimate of a T-step method $\mathcal{B}^T$ is an upper bound on the worst-case, over the problems f from the family in question, inaccuracy $\varepsilon(x^T[\mathcal{B}^T, f], f)$ of approximate solutions generated by the method.

family $\mathcal{F}_{\|\cdot\|_p}(\kappa, L)$ equipped with the first order oracle it holds for all $T \geq 1$:

$$\begin{array}{ll}(a) & \textit{in the range } 1 \leq p \leq 2\textit{:} \quad \mathrm{Risk}(T) \leq O(1)\underbrace{\left(\min\left[\frac{1}{p-1}, \ln(n)\right]\right)^{\kappa/2}}_{C(p)} \frac{LR^\kappa}{T^{\frac{3}{2}\kappa-1}}; \\ (b) & \textit{in the range } 2 \leq p \leq \infty\textit{:} \quad \mathrm{Risk}(T) \leq O(1)\underbrace{(\min[p, \ln(n)])^\kappa}_{C(p)} \frac{LR^\kappa}{T^{\kappa(1+\frac{1}{\min[p,\ln(n)]})-1}}. \end{array} \tag{11}$$

In the range $1 \leq p \leq 2$ of values of p, the same upper bound holds true when $(\mathbf{E}, \|\cdot\|)$ is Schatten p-space rather than ℓ_p^n.

As far as the normed spaces of primary interest in this thesis – the spaces ℓ_p^n – are concerned, the above results say the following (we use notation from Corollary 1.2.3):

(i) Bound (11) holds true for all $T = 1, 2, ...$ and thus holds true in both low- and large-scale regime. However, in the low-scale regime $T \geq O(1)n\ln(n)$ this bound is progressively, as T grows, outperformed by bound (8) and as such is of no interest.

(ii) As far as the dependence of the bound (11) on L and R is concerned, it is proportional to LR^κ, as it should be by homogeneity and scaling reasons[9] With this in mind, *in the rest of this discussion we consider the normalized situation where $L = R = 1$.*

(iii) When p is bounded away from 1 and from ∞, the risk of the family $\mathcal{P}$ admits *uniform in the dimension n* upper bound. Moreover, the factor $C(p)$ in (11) is "nearly uniformly bounded" in the entire range $[1, \infty]$ of values of p – we always have $C_p \leq \ln^2(n)$. Thus, the essence of the matter is the rate at which

[9] We always can scale the variables and the objective to enforce $L = R = 1$, preserving the value of κ, and accuracy ε in terms of the objective in the scaled problem, in full accordance with (11), corresponds to accuracy εLR^κ in the original problem.

the right hand side in (11) decreases as T grows. As a function of T, the bound is (proportional to) $T^{-\mu}$ with

$$\mu = \begin{cases} \frac{3}{2}\kappa - 1, & 1 \le p \le 2 \\ \kappa - 1 + \frac{\kappa}{\min[p, \ln(n)]}, & 2 \le p \le \infty. \end{cases} \tag{12}$$

(iv) From (12) we see that in the range of $1 \le p \le 2$, the bound (11) decreases, as $T \to \infty$, at the rate $T^{-[\frac{3}{2}\kappa-1]}$ depending solely on the smoothness modulus κ of the objective and *completely independent of* p; when κ varies from 1 (nonsmooth case) to 2 (fully smooth case), the rate improves from $O(T^{-1/2})$ to $O(T^{-2})$. In contrast to this, *in the range* $2 \le p \le \infty$ *the rate of convergence is heavily affected by* p – as p grows from 2 to ∞, μ decreases from $\frac{3}{2}\kappa - 1$ to $\kappa - 1 + \Delta$, $\Delta = \kappa/\ln(n)$ (in fact, the worst – the smallest – value of μ is achieved already at $p = \ln(n)$. Note that in fact in the complexity context $\mu = \kappa - 1 + \Delta$ is basically as bad as $\mu = \kappa - 1$; indeed, when T does not exceed a polynomial in n, say, n^2, $T^{-[\kappa-1+\kappa/\ln(n)]}$ and $T^{-[\kappa-1]}$ coincide within an absolute constant factor, and absolute constant factors traditionally "go beyond the resolution" of Information-Based complexity bounds. The case of $\ln(T) \gg \ln(n)$, where the component $\kappa/\ln(n)$ in μ indeed is important, is by itself of no interest – this is what was called "low-scale" regime, and it was already explained that in this regime there exist methods with much better efficiency estimates $O(\exp\{-O(1)T/n\})$.

(v) The bound (12) admits passing to limit as $\kappa \to +1$, which suggests (and this suggestion indeed is true) that the algorithm underlying the bound is capable to solve convex problems with Lipschitz continuous objectives; the efficiency estimate of the resulting algorithm as applied to convex objective f with Lipschitz constant, taken w.r.t. $\|\cdot\|$, $L/2$ is obtained from the bound (12) by setting $\kappa = 1$.

An alternative way to get the same upper complexity bounds for nonsmooth convex optimization (i.e., the counterpart of $\mathcal{P}$ where the family of objectives is $\mathrm{Lip}_{\mathbf{E},\|\cdot\|}(L)$) under a first order oracle reporting the value and a subgradient of the objective at the query point is to use the Mirror Descent algorithm [32]. For the sake of convenience, we provide these bounds:

$$\begin{array}{ll}(a) & \text{in the range } 1 \le p \le 2\text{:} \quad \mathrm{Risk}(T) \le O(1)\sqrt{\min\left[\frac{1}{p-1}, \ln(n)\right]}\frac{LR}{\sqrt{T}};\\ (b) & \text{in the range } 2 \le p \le \infty\text{:} \quad \mathrm{Risk}(T) \le O(1)\min[p, \ln(n)]\frac{LR}{T^{1/p}}\end{array} \tag{13}$$

(one can replace $T^{1/p}$ with $T^{1/\min[p,\ln(n)]}$, but in the large-scale regime it does not make any difference).

(vi) The validity of upper bound (11) for Schatten p-spaces with $p > 2$ is not known. However, in the particular case $p = \infty$ (and in fact $p \ge \ln n$ suffices) there is an alternative method – Conditional Gradient – that achieves an upper complexity bound

$$\mathrm{Risk}(T) \le O(1)\frac{LR^\kappa}{T^{\kappa-1}},$$

which, up to logarithmic in the dimension terms, coincides with (11) for $p \ge \ln n$. This algorithm applies for both ℓ_p^n and Sch_p^n, and –to the best of our knowledge – it is not known to be optimal in the local oracle model.

Lower complexity bounds, nonsmooth case. It is known [32] that in the case of $(\mathbf{E}, \|\cdot\|) = \ell_p^n$, $\|\cdot\|_p$-ball of radius $R/2$ in the role of X and $\mathrm{Lip}_p^n(L)$ in the role of $\mathcal{F}$, the risk of the problem class $\mathcal{P} = (\mathcal{F}, X)$ equipped with the universal local oracle *in the range* $1 \le T \le n/4$ (which is slightly smaller than the large-scale range) can be *lower-bounded* as follows:

$$\begin{array}{ll}(a) & \text{in the range } 1 \le p \le 2\text{:} \quad \mathrm{Risk}(T) \ge O(1)\frac{LR}{\sqrt{T}};\\ (b) & \text{in the range } 2 \le p \le \infty\text{:} \quad \mathrm{Risk}(T) \ge O(1)\frac{LR}{T^{1/p}}\end{array} \tag{14}$$

Note that these lower risk bounds fit the upper bounds (13) within a factor which is

– just a constant when p is bounded away from 1 and from ∞ (the constant depends solely on the endpoints of such a range, and

– does not exceed $O(1)\ln(n)$ in the entire range $[1,\infty]$ of values of p.

Lower complexity bounds, smooth case. To the best of our knowledge, the only known lower complexity bounds for large-scale smooth convex minimization deal with the case of problems with Euclidean geometry and Lipschitz continuous gradient of the objective. The corresponding result is as follows:

Theorem 1.2.4. [32, 28, 29] *Let $(\mathbf{E}, \|\cdot\|) = \ell_2^n$, $L > 0$, $R > 0$, and let X be $\|\cdot\|_2$-ball of diameter R in $\mathbf{E} = \mathbb{R}^n$. Consider the family $\mathcal{F}_{2,L}$ of all convex quadratic forms $f(x) = \frac{1}{2}x^TAx - b^Tx + c : \mathbb{R}^n \to \mathbb{R}$ with positive semidefinite matrices A of spectral norm not exceeding L (i.e., convex quadratic forms from $\mathcal{F}_{\ell_2^n}(2, L)$), and let $\mathcal{P}$ be the family of convex minimization problems $\min_{x\in X} f(x)$ with objectives $f \in \mathcal{F}_{2,L}$). Then in the range $1 \le T \le n/4$ the risk, taken w.r.t. the first order oracle, of the family $\mathcal{P}$ can be lower-bounded by $O(1)\frac{LR^2}{T^2}$.*

As a result, in the case of $p = 2$, $\kappa = 2$ and in the large-scale regime $n \ge 4T$, the risk, taken w.r.t. the first order oracle, of the family $\mathcal{P}$ can be lower-bounded as

$$\mathrm{Risk}(T) \ge O(1)\frac{LR^2}{T^2}. \tag{15}$$

Note that the lower risk bound (15) is within absolute constant factor of the upper bound (11) corresponding to the case of $(\mathbf{E}, \|\cdot\|) = \ell_2^n$, $p = 2$, $\kappa = 2$ and Euclidean ball of diameter R in the role of X; thus, in this case *and in the large-scale regime $n \ge 4T$* both upper and lower risk bounds are tight. As it was already explained, large-scale regime is important here.

One of the main goals of this thesis is to build lower risk bounds for smooth convex minimization which fit the upper bounds (11) *in the entire range of situations*

covered by Corollary 1.2.3, provided that X is "as massive" as it is allowed by the Corollary (specifically, X contains a $\|\cdot\|_p$-ball of diameter R [10]).

1.2.5 Information theory

We introduce some basics of Information Theory that will be needed in Chapter 3. For a thorough presentation on the subject we refer to [6]. From now on, $\log(.)$ denotes the binary logarithm and capital letters will typically represent random variables or events. We can describe an event E as a random variable by the indicator function $I(E)$, which takes value 1 if E happens, and 0 otherwise.

The *entropy* of a discrete random variable A is

$$\mathbb{H}[A] := -\sum_{a \in \text{range}(A)} \mathbb{P}[A = a] \log \mathbb{P}[A = a].$$

This definition extends naturally to *conditional entropy* $\mathbb{H}[A \mid B]$ by using the corresponding conditional distribution and taking expectation, i.e.,

$$\mathbb{H}[A \mid B] = \sum_b \mathbb{P}[B = b] \, \mathbb{H}[A \mid B = b].$$

Fact 1.2.5 (Properties of entropy)**.**

Bounds $0 \leq \mathbb{H}[A] \leq \log |\text{range}(A)|$

$\mathbb{H}[A] = \log |\text{range}(A)|$ if and only if A is uniformly distributed.

Monotonicity $\mathbb{H}[A] \geq \mathbb{H}[A \mid B]$;

The notion of *mutual information* defined as $\mathbb{I}[A; B] := \mathbb{H}[A] - \mathbb{H}[A \mid B]$ of two random variables A and B captures how much information about a 'hidden' A is leaked by observing B. Sometimes A and B are a collection of variables, then a

[10] "massiveness" is indeed important – Corollary 1.2.3 allows for X to be "long and narrow," say, to be just a segment of $\|\cdot\|_p$-diameter R; clearly, in this case the true risk of $\mathcal{P}$ (which now is in fact a family of univariate convex minimization problems) is incomparably smaller than the bound (11).

comma is used to separate the components of A or B, and a semicolon to separate A and B themselves: e.g., $\mathbb{I}[A_1, A_2; B] = \mathbb{I}[(A_1, A_2); B]$. Mutual information is a symmetric quantity and naturally extends to *conditional mutual information* $\mathbb{I}[A; B \mid C]$ as in the case of entropy. Clearly, $\mathbb{H}[A] = \mathbb{I}[A; A]$.

Fact 1.2.6 (Properties of mutual information)**.**

Bounds If A is a discrete variable, then $0 \leq \mathbb{I}[A; B] \leq \mathbb{H}[A]$

Chain rule $\mathbb{I}[A_1, A_2; B] = \mathbb{I}[A_1; B] + \mathbb{I}[A_2; B \mid A_1]$.

Symmetry $\mathbb{I}[A; B] = \mathbb{I}[B; A]$.

Independent variables The variables A and B are independent if and only if $\mathbb{I}[A; B] = 0$.

A simple but very powerful result in information theory is *Fano's inequality*, which allows us to lower bound the probability of guessing the value of a random variable X from information of a correlated random variable Y.

Theorem 1.2.7 (Fano's Inequality, [6])**.** *Let X be a random variable taking values on a finite set $\mathcal{X}$. For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$ is a Markov chain (i.e., X and $\hat{X}$ are conditionally independent given Y), if we define $P_e = \mathbb{P}\left[X \neq \hat{X}\right]$, we have*

$$\mathbb{H}[P_e] + P_e \log |\mathcal{X}| \geq \mathbb{H}\left[X \,\middle|\, \hat{X}\right] \geq \mathbb{H}[X \mid Y].$$

This inequality can be weakened to

$$P_e \log |\mathcal{X}| \geq \frac{\mathbb{H}[X \mid Y] - 1}{\log |\mathcal{X}|}.$$

1.3 Outline of Results

In this Section we present the major contributions of this thesis, together with the key ideas that allow us to derive these results.

1.3.1 Chapter II: Worst-Case Oracle Complexity of Large-Scale Smooth Convex Optimization

We study the problem of minimization of smooth convex functions. We generalize the analysis of the nonsmooth case by a *local smoothing* of hard instances; when such a local smoothing is possible, we derive a general lower bound on the complexity. In cases where local smoothing is not directly applicable, we provide alternative proofs based on convex geometry, specifically on random projections of the feasible domain.

First, we introduce the notion of a *smoothing kernel*, which is a smooth convex function ϕ with "nice" properties (see A, B, C in Section 2.2.1 in Chapter 2). Under the existence of such function, we can construct an smooth approximation for arbitrary $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ that we call *local smoothing*, and is given by the following expression

$$\mathcal{S}_\chi[f](x) = \min_{h \in \chi \mathrm{Dom}\,\phi} [f(x+h) + \chi\phi(h/\chi)].$$

Note that this construction extends the classical Moreau smoothing [26, 27, 45], corresponding to the Euclidean case where $(\mathbf{E}, \|\cdot\|)$ is $\mathbf{R}^n$ with the standard Euclidean norm, and the smoothing kernel $\phi(x) = \frac{1}{2}\|x\|_2^2$. The following results can be considered as the non-Euclidean extension of the classical result by Moreau:

Theorem (2.2.2). *Let $(\mathbf{E}, \|\cdot\|)$ be a (finite-dimensional) normed space such that there exists a smoothing kernel ϕ. Then for any convex function $f : \mathbf{E} \to \mathbb{R}$ such that $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ and every $\chi > 0$ there exists a smooth (i.e., with Lipschitz continuous gradient) approximation $\mathcal{S}_\chi[f]$ that satisfies:*

S.1. $\mathcal{S}_\chi[f]$ is convex and Lipschitz continuous with constant 1 w.r.t. $\|\cdot\|$ and has a Lipschitz continuous gradient, with constant M_ϕ/χ, w.r.t. $\|\cdot\|$:

$$\|\nabla\mathcal{S}_\chi[f](x) - \nabla\mathcal{S}_\chi[f](y)\|_* \le \chi^{-1} M_\phi \|x - y\| \ \forall x, y;$$

S.2. $\sup_{x\in E}|f(x)-\mathcal{S}_\chi[f](x)|\le\chi\rho_{\|\cdot\|}(G)$. *Moreover,* $f(x)\ge\mathcal{S}_\chi[f](x)\ge f(x)-\chi\rho_{\|\cdot\|}(G)$.

S.3. $\mathcal{S}_\chi[f]$ *depends on* f *in a local fashion: the value and the derivative of* $\mathcal{S}_\chi[f]$ *at* x *depend only on the restriction of* f *onto the set* $x+\chi G$.

The main result in this Chapter, Proposition 2.3.1, establishes a general lower bound for the complexity of smooth minimization.

Proposition (2.3.1). *Let*

I. $\|\cdot\|$ *be a norm on* $\mathbb{R}^n$ *and* X *be a nonempty convex set;*

II. T *be a positive integer and* Γ *be a positive real with the following property:*

There exist T *linear forms* $\langle\xi^t,\cdot\rangle$ *on* $\mathbb{R}^n$, $1\le t\le T$, *such that*

(a) $\|\xi^t\|_*\le 1$ *for* $t\le T$, *and*

(b) for every collection $s=(s_1,...,s_T)$ *with* $s_t\in\{-1,1\}$, *it holds*

$$\min_{x\in X}\max_{1\le t\le T}s_t\langle\xi^t,x\rangle\le-\Gamma;$$

III. M *and* ρ *be positive reals such that for properly selected convex twice continuously differentiable on an open convex set* $\mathrm{Dom}\,\phi\subset\mathbb{R}^n$ *function* ϕ *and a convex compact subset* $G\subset\mathrm{Dom}\,\phi$ *the triple* $(\phi,G,M_\phi=M)$ *satisfies properties A, B, C from Section 2.2.1 and* $\rho_{\|\cdot\|}(G)\le\rho$.

Then for every $L>0$, $\kappa\in(1,2]$, *every local oracle* $\mathcal{O}$ *and every* T*-step method* $\mathcal{A}$ *associated with this oracle there exists a problem* $(P_{f,X})$ *with* $f\in\mathcal{F}_{\|\cdot\|}(\kappa,L)$ *such that*

$$f(x_T(\mathcal{A},f))-\mathrm{Opt}(f)\ge\frac{\Gamma^\kappa}{2^{\kappa+1}(\rho M)^{\kappa-1}}\cdot\frac{L}{T^{\kappa-1}}.$$

We are interested in the specific case of ℓ_p/ℓ_q-settings, where $\mathbf{E}=\mathbb{R}^n$, $\|\cdot\|=\|\cdot\|_q$, and X contains the unit p-ball B_p^n. By application of Proposition 2.3.1 and

tools from high-dimensional convex geometry, we obtain lower bounds in the complexity. We outline our new lower bounds in Table 1.3.1. In the table, $\mathcal{R}(T)$ is defined as the ratio between the best-known algorithm in the given range, and our lower complexity bound, and $\tilde{O}(\cdot)$ omits factors that are at most cubic in the logarithm of the dimension.

Table 2: Worst-case risk lower bounds for (κ, L)-smooth convex optimization in the ℓ_p/ℓ_q-setting.

Range q	Range p	Risk Lower bound	$\mathcal{R}(T)$
$1 \le q \le 2$	$p \le q$	$O\left(\frac{1}{[\ln n]^{\kappa-1}} \frac{LR^\kappa}{T^{\kappa[\frac{3}{2}+\frac{1}{p}-\frac{1}{q}]-1}}\right)$	$\tilde{O}\left(T^{\kappa[\frac{1}{p}-\frac{1}{q}]}\right)$
	$p > q$	$O\left(\frac{n^{\kappa(\frac{1}{q}-\frac{1}{p})}}{[\ln n]^{\kappa-1}} \frac{LR^\kappa}{T^{\frac{3\kappa}{2}-1}}\right)$	$\tilde{O}(1)$
$2 < q \le \infty$	$p \le q$	$O\left(\frac{1}{[\min\{q, \ln n\}]^{\kappa-1}} \frac{LR^\kappa}{T^{\kappa[1+\frac{1}{p}]-1}}\right)$	$\tilde{O}\left(T^{\kappa[\frac{1}{p}-\frac{1}{q}]}\right)$
	$p > q$	$O\left(\frac{n^{\kappa(\frac{1}{q}-\frac{1}{p})}}{[\min\{q, \ln n\}]^{\kappa-1}} \frac{LR^\kappa}{T^{\kappa[1+\frac{1}{q}]-1}}\right)$	$\tilde{O}(1)$

Some comments are in order.

A. We see that as far as large-scale case is concerned, *in the range $p \ge q$, our lower complexity bounds are tight within logarithmic in n factor*. This, in particular, implies near-optimality in this range of Nesterov's Fast Gradient algorithm (in its version, developed in [22, Section 2.3], adjusted to convex objectives with Hölder continuous gradients and smoothness quantification taken w.r.t. $\|\cdot\|_q$-norms, $1 \le q \le \infty$).

B. In the case of $p = q = \infty$, the upper complexity bounds (46) can be achieved not only with the aforementioned Nesterov's type algorithm; they are nothing but

the standard upper complexity bounds of the classical Conditional Gradient algorithm originating from [11], see also [39, 10, 19, 14] and references therein. This algorithm recently has attracted a lot of attention, primarily in the Machine Learning community, due to its ability to work with "difficult geometry domains" (like nuclear norm or total variation norm balls) where the *proximal* first order algorithms (which form the vast majority of first order methods) become too computationally expensive in terms of Information-Based Complexity Theory.

C. In contrast to what happens in the range $p \geq q$, in the range $1 \leq p < q \leq \infty$ there is a substantial gap $\mathcal{R}(T) \approx T^{\kappa(1/p-1/q)}$ between the upper and the lower complexity bounds. Our *guess* is that the "guilty party" here is the *upper* bound (see Open Problem 2.5.1), the motivation being as follows. The upper complexity bound (46) in the range $p < q$ is just independent of p; were this bound tight, it would mean that with fixed degree of smoothness (quantified w.r.t. the norm $\|\cdot\|_q$), minimizing smooth objectives over the unit $\|\cdot\|_q$-ball is basically as difficult as minimizing these objectives over the unit $\|\cdot\|_p$-ball with $p < q$, in spite of the fact that the second ball is "incomparably smaller" than the first one when n is large.

D. *Matrix case:* The results obtained for ℓ_p/ℓ_q-setups can be extended to matrix (spectral) setups, for optimization of $\|\cdot\|_{\text{Sch},q}$-smooth functions of a matrix decision variable $X \in \mathbb{R}^{n\times n}$, with a p-Schatten norm constraint: $\|X\|_{\text{Sch},p} \leq R$. The lower bounds are essentially the same, with the caveat that they only hold for $T \leq n$, which is the square-root of the ambient dimension.

On the other hand, our understanding of upper bounds is more restricted: Only for $1 \leq q \leq 2$ we have upper bounds analogous to the ℓ_p/ℓ_q-case.

1.3.2 Chapter III: Distributional Oracle Complexity of Convex Optimization

We propose an information-theoretic framework to analyze the oracle complexity of convex optimization. Our method is based on distributional complexity (also known as average-case analysis). We remind the reader that the *ε-distributional complexity of problem class $\mathcal{P}$ w.r.t. oracle $\mathcal{O}$* is defined by

$$\mathrm{Compl}_{\mathcal{D}}(\varepsilon) = \sup_{\mathbb{P}\in\mathfrak{P}} \inf_{\mathcal{B}\in\mathfrak{B}[\varepsilon]} \int_{\mathcal{F}} \mathbf{E}\{T[\mathcal{B}, f]\}\mathbb{P}(df), \tag{16}$$

where $\mathfrak{P}$ is the family of all probability distributions $\mathbb{P}$ on $\mathcal{F}$, and $\mathfrak{B}[\varepsilon]$ is the family of randomized algorithms guaranteed to terminate with ε-solutions, for arbitrary instances from the class $\mathcal{B}$. Our techniques can be easily extended to the *high probability oracle complexity* of $\mathcal{P}$ w.r.t. $\mathcal{O}$ defined as

$$\mathrm{Compl}^{\beta}_{\mathcal{HP}}(\varepsilon) = \sup_{\mathbb{P}\in\mathfrak{P}} \inf_{\mathcal{B}\in\mathfrak{B}[\varepsilon]} \inf\{\tau : \mathbb{P}\{f : T[\mathcal{B}, f] \leq \tau\} \geq 1-\beta\}$$

For the notions above we can moreover consider *algorithms with bounded error*, where there is a probability $P_e > 0$ that the algorithm does not output an ε-solution. This potentially gives the algorithm the freedom to discard expensive instances, as done for Monte-Carlo algorithms.

The notions considered above define weak notions of complexity, since the following chain holds

$$\mathrm{Compl}_{\mathcal{D}} \leq \mathrm{Compl}_{\mathcal{R}} \leq \mathrm{Compl}\,.$$

The reader should observe that the first inequality is not necessarily an equality, since Yao's minimax principle [1] (or equivalently, Von Neumann/Sion minimax theorem) does not apply, since both instances and algorithms are defined by infinite families.

Our main result is that for nonsmooth convex optimization, the three measures of complexity in the chain above – worst-case, randomized, distributional – coincide up to a constant factor. Furthermore, for fixed $\beta \in (0, 1]$, high probability

oracle complexity also coincides, up to constant factor, with the measures of complexity above. Specifically, our work contributions can be summarized as follows:

Information-theoretic framework. Our work is the first to provide an information-theoretic analysis for deterministic local oracles. This analysis is based on the *reconstruction principle*, and is given by the following

Lemma (3.2.1). *Let F be a random variable with finite range $\mathcal{F}$. For a given algorithm determining F via querying an oracle, with error probability bounded by P_e, suppose that the information gain from each oracle answer is bounded, i.e., for some constant C*

$$\mathbb{I}\left[F; A_t \mid \Pi_{<t}, Q_t, T \geq t\right] \leq C, \qquad t \geq 0. \tag{17}$$

Then, the distributional oracle complexity of the algorithm is lower bounded by

$$\mathbb{E}\left[T\right] \geq \frac{\mathbb{H}\left[F\right] - \mathbb{H}\left[P_e\right] - P_e \log |\mathcal{F}|}{C}.$$

Moreover, for all t we have

$$\mathbb{P}\left[T < t\right] \leq \frac{\mathbb{H}\left[P_e\right] + P_e \log |\mathcal{F}| + Ct}{\mathbb{H}\left[F\right]}.$$

In particular, if F is uniformly distributed, then for $t = \frac{\beta \log |\mathcal{F}| - \log 2}{C}$, $\mathbb{P}\left[T \geq t\right] = 1 - P_e - \beta$.

Therefore, if we can bound the information gain extracted from the oracle by a constant C, then we obtain a lower bound $\mathbb{E}\left[T\right] \geq \mathbb{H}\left[F\right]/C$, together with bounds for high-probability and for algorithms with bounded error.

It is important to observe that the setup described above is not constrained to convex optimization, and it is well-suited for any situation where we want to determine a random instance from oracle information. We are able to relate this notion with convex optimization by finding families of functions with a *packing property*, which implies that *optimization amounts for determining the instance*.

Definition (3.1.1). *A function family $\mathcal{F}$ satisfies the packing property for an accuracy level $\varepsilon > 0$, if for every different members $f, g \in \mathcal{F}$, we have $\mathcal{S}_\varepsilon(f) \cap \mathcal{S}_\varepsilon(g) = \emptyset$, where*

$$\mathcal{S}_\varepsilon(f) := \{x \in X : \ f(x) < f^* - \varepsilon\}.$$

Common source of hardness. From the general framework proposed above we establish the complexity of a *String-Guessing problem*. The description of this problem is the following

Oracle (3.3.1. String Guessing Oracle $\mathcal{O}_S$**).**

Query: *A string $s \in \{0,1\}^{\leq M}$ and an injective function $\sigma \colon [|s|] \to [M]$.*

Answer: *Smallest $k \in \mathbb{N}$ so that $S_{\sigma(k)} \neq s_k$ if it exists, otherwise* `EQUAL`*.*

We establish an $O(M)$ lower bound for this problem, for distributional, high probability, and bounded error complexity.

Proposition (3.3.2)**.** *Let M be a positive integer, and S be a uniformly random binary string of length M. Let $\mathcal{O}_S$ be the String Guessing Oracle (Oracle 3.3.1). Then for any bounded error algorithm having access to S only through $\mathcal{O}_S$, the expected number of queries required to identify S with error probability at most P_e is at least $[(1 - P_e)M - 1]/2$. Moreover, given $\beta > 0$ if we let $t = \frac{\beta M - \log 2}{2}$, then $\mathbb{P}[T \geq t] = 1 - P_e - \beta$, where T is the number of queries.*

This core problem is then utilized to derive lower bounds for convex optimization algorithms under a specific subgradient oracle. We make this connection explicit by studying families of problems indexed by strings: for the large-scale case, function instances have the form

$$f_s(x) = \max_{t \in [M]} s_t \langle \xi^t, x \rangle,$$

and for the low-scale case, there is a more involved recursive definition (see Section 3.5). The local oracle under study is the one that provides the bit s_t such that t

is a maximizer coordinate for query x (when not unique we choose the first one in some prescribed order). Note that although the explicit information of the oracle is a single bit, there is implicit information on other bits from the magnitudes of the coordinates of the query $|\langle \xi^t, x\rangle|$.

The reader may observe the functions above define the same hard family studied in the previous Chapter (without perturbations), but where we are specifying the oracle in use.

For both low-scale and large-scale settings, we derive lower complexity lower bounds analogous to the ones derived by Nemirovski and Yudin [32] in Theorems 3.5.2 and 3.6.1. The novelty here is the way we derive these lower bounds from *oracle emulation*, a procedure to compare the complexity of problems, translating oracle information of one problem into the one of the other

Definition (3.4.1). *Let $\mathcal{O}_1 \colon Q_1 \to R_1$ and $\mathcal{O}_2 \colon Q_2 \to R_2$ be two oracles for the same problem. An* emulation *of $\mathcal{O}_1$ by $\mathcal{O}_2$ consists of*

(i) a query emulation function $q \colon Q_1 \to Q_2$ (translating queries of $\mathcal{O}_1$ for $\mathcal{O}_2$),

(ii) an answer emulation function $a \colon Q_1 \times R_2 \to R_1$ (translating answers back)

such that $\mathcal{O}_1(x) = a(x, \mathcal{O}_2(q(x)))$ for all $x \in Q_1$.

In Lemma 3.4.2 is it proved that if $\mathcal{O}_1$ can be emulated by $\mathcal{O}_2$, then the oracle complexity of $\mathcal{O}_1$ is lower bounded by the one of $\mathcal{O}_2$, for any measure of complexity (since it holds pointwise).

First lower bounds for distributional and high-probability complexity for all local oracles. Our results for arbitrary local oracles, established in Section 3.7 and further extended in Section 3.8, lead to the following lower complexity bounds in the ℓ_p/ℓ_q-setting. We remind the reader that in this setting, the class of functions

is given by convex Lipschitz continuous functions with constant $L > 0$ w.r.t. $\|\cdot\|_q$, and the optimization domain X is given by a ball of radius $R > 0$, $B_p^n(R)$.

First, for the unit ℓ_∞-ball (i.e., $p = \infty$), as well as the *low-scale regime* (see the Table 1.3.2 below for the opposite range, i.e., large-scale), we have a distributional complexity lower bound of $O\left(n \log \frac{LR}{\varepsilon}\right)$ for algorithms with zero error probability. For algorithms with error probability P_e, the lower bound is $O\left((1 - P_e)n \log \frac{LR}{\varepsilon}\right)$. On the other hand, the high probability complexity of level β for algorithms with error bounded by P_e is $O\left(\delta(1 - P_e)n \log \frac{LR}{\varepsilon}\right)$

The large-scale lower bounds are summarized in Table 1.3.2 . In particular, for the standard setting, where $p = q$, we obtain lower bounds for nonsmooth convex optimization matching [32]. For $1 \leq p \leq 2$ and $n \geq 1/\varepsilon^2$, the distributional oracle complexity is lower bounded by $O(LR/\varepsilon^2)$; for $2 < p < \infty$ and $n \geq 1/\varepsilon^p$, the distributional oracle complexity is lower bounded by $O(LR/\varepsilon^p)$. For arbitrary $1 \leq p < \infty$, in the case of algorithms with error probability bounded by P_e, if we let $r := \max\{2, p\}$, the distributional complexity is lower bounded by $O\left((1 - P_e)LR/\varepsilon^r\right)$, and the high-probability complexity of level δ is lower bounded by $O\left(\delta(1 - Pe)LR/\varepsilon^r\right)$.

Close the gap between randomized and worst-case complexity. As a byproduct, our lower bounds for distributional complexity close the logarithmic gap between randomized and worst-case complexity, established in [32]. In other words, for black-box convex optimization, there is no gain from randomization. This is in stark contrast with specific problems where randomization can show a dramatic speed up [33], [21].

Table 3: Distributional complexity lower bounds for nonsmooth convex optimization in the ℓ_p/ℓ_q-setting.

Range q	Large-scale range	Range p	Lower bound
$1 \le q \le 2$	$n \ge \frac{1}{\varepsilon^{(\frac{1}{p}-\frac{1}{q}+\frac{1}{2})^{-1}}}$	$p \le q$	$O\left(\frac{LR}{\varepsilon^{(\frac{1}{p}-\frac{1}{q}+\frac{1}{2})^{-1}}}\right)$
		$p > q$	$O\left(\frac{LR\, n^{2(\frac{1}{q}-\frac{1}{p})}}{\varepsilon^2}\right)$
$2 < q < \infty$	$n \ge \frac{1}{\varepsilon^p}$	$p \le q$	$O\left(\frac{LR}{\varepsilon^p}\right)$
		$p > q$	$O\left(\frac{LR\, n^{(1-\frac{q}{p})}}{\varepsilon^q}\right)$

CHAPTER II

THE WORST-CASE ORACLE COMPLEXITY OF LARGE-SCALE SMOOTH CONVEX OPTIMIZATION

2.1 Introduction

The theory of oracle complexity in convex optimization was quite successful on establishing tight limits of performance for *nonsmooth* convex minimization problems [32]. From this it is known that for large-scale instances variants of Mirror-Descent provide optimal convergence rates for domains given by ℓ_p-balls, where $1 \leq p < \infty$.

However, in the smooth case, our understanding is limited; essentially, tight *lower* complexity bounds are known only in the case when the domain X is an Euclidean ball and the objective f is a convex function with Lipschitz continuous gradient (w.r.t. $\|\cdot\|_2$). In this case, lower bounds are obtained from least-squares problems [28, 29], and the underlying techniques for generating a hard family of instances heavily utilize the rotational invariance of the Euclidean ball.

In this chapter, we derive lower bounds on the oracle complexity of classes of convex minimization problems beyond the nonsmooth case. In the terminology and notation of the Introduction, specifically in Section 1.2.2, we consider classes of problems $\mathcal{P} = (\mathcal{F}, X)$ for ℓ_p/ℓ_q-settings (where $1 \leq p, q \leq \infty$), i.e., where X is an n-dimensional $\|\cdot\|_p$-ball, and $\mathcal{F} = \mathcal{F}_q(\kappa, L)$ is the family of all continuously differentiable convex objectives with given smoothness parameters (Hölder exponent κ, and constant L) w.r.t. $\|\cdot\|_q$. These bounds are a substantial extension of the existing lower complexity bounds for large-scale convex minimization covering the nonsmooth case and the Euclidean smooth case. Moreover, the lower bounds

derived for vector optimization can be easily translated to their matrix analogies – domains given by Schatten norm balls in the space of square matrices.

Our results are nearly tight for what we called the *standard case*, where $p = q$. The main motivation for these results is the connection of minimization algorithms with some modern applications, such as ℓ_1 and nuclear norm minimization in Compressed Sensing [35], where one seeks to minimize a smooth, most notably, quadratic convex function over a high-dimensional ℓ_1-ball in $\mathbb{R}^n$ or nuclear norm ball in the space of $n \times n$ matrices. In this case, our results indicate that modifications of Nesterov's fast gradient method [22, Section 2.3] are nearly optimal over the class of black-box methods for smooth convex minimization.

Another instructive application of our results is establishing the near-optimality of the conditional gradient (a.k.a. Frank-Wolfe) algorithm as applied to minimizing smooth convex functions over large-scale boxes, and spectral norm unit balls in the space of matrices (correspoinding to $p = q = \infty$). This algorithm, first proposed in [11], was intensively studied in 1970's (see [10, 39] and references therein); recently, there is a significant burst of interest in this technique, due to its ability to handle smooth large-scale convex programs on "difficult geometry" domains, see [15, 5, 17, 18, 14, 7, 24, 12] and references therein.

Our results go far beyond to the general – *non-standard case* – where $p \neq q$. A motivation for this general case is in linear regression models, where we search for a linear predictor within a set X, which is assumed to be norm-bounded, e.g. $X = B_p^n$; and measure the performance of a predictor by a loss function arising from random samples $(a_1, b_1), \ldots, (a_m, b_m) \in B_{q_*}^n \times [-1, 1]$. Thus the empirical risk minimization problem we obtain, $\min\{\frac{1}{m}\sum_{j=1}^m (a_j^T x - b_j)^2 : \|x\|_p \leq 1\}$, is a particular case of the ℓ_p/ℓ_q-setting. In modern applications, the way the predictor space (with parameter p) and the distribution bound (with parameter q) are chosen do not necessarily coincide.

Our lower complexity bounds are nearly tight for most ranges of p and q. Remarkably, in the case $p \geq q$ our nearly tight bounds turn out to be *dimension-dependent*, a phenomenon that – to the best of our knowledge – is new in lower bounds on the oracle complexity of convex optimization. It is worth mentioning that on the upper bound side, dimension-dependent complexity bounds have been systematically observed, e.g. [9, 3, 8]; our work justifies the near-optimality of these methods.

Surprisingly, our lower bounds are not tight is when $p < q$. We regard the open problem of improving (upper or lower) complexity bounds in this range as a major one; e.g., it includes minimization algorithms for Compressed Sensing with Fourier measurements (for $p = 1$ and $q = 2$) [40] and, more broadly, any family of linear measurements with vectors of bounded Euclidean norm. In this case, our results show room for potential acceleration over standard methods. We finish this discussion by stressing the fact that if our open problem is resolved by improving upper complexity bounds, that implies that even in the "favorable geometry" case, i.e., where $1 \leq p, q \leq 2$, Nesterov's Accelerated method might not be a universally optimal method; casting some doubt on common belief within the optimization community.

2.1.1 The approach

In order to construct hard instances for lower bounds we need the normed space under consideration to satisfy a "smoothing property." Namely, we need the existence of a *smoothing kernel* – a convex function with Lipschitz continuous gradient and "fast growth." These properties guarantee that the inf-convolution [16] of a Lipschitz continuous convex function f and the smoothing kernel is smooth, and its local behavior depends only on the local behavior of f. A novelty here, if any, stems from the fact that we need Lipschitz continuity of the gradient w.r.t. a given,

not necessarily Euclidean, norm, while the standard Moreau envelope technique is adjusted to the case of the Euclidean norm.

We establish lower bounds on complexity of smooth convex minimization for general spaces satisfying the smoothing property. Our proof mimics the construction of hard instances for nonsmooth convex minimization [32], which now are properly smoothed by the inf-convolution.

From local smoothing we derive lower complexity bounds for convex optimization in the so called ℓ_p/ℓ_q-setting, i.e., minimization of functions in the class $\mathcal{F}_{\|\cdot\|_q}(\kappa, L)$ over domain $X \subset \mathbb{R}^n$ containing a unit p-ball B_p^n. It is worth mentioning that local smoothing can only be used directly when $q \geq 2$, which is a limitation ultimately related to Banach space geometry [2]. When $1 \leq q < 2$ we follow a different path: it turns out that random projections of the feasible domain contain sets for which we can derive lower complexity bounds. By taking a hard family of functions on this set, and *lifting* those instances to the whole feasible domain, we obtain new lower bounds for the complexity.

2.2 *Local Smoothing*

In this section we introduce the main component of our technique, a Moreau-type approximation of a nonsmooth convex function f by a smooth one. The main feature of this smoothing, instrumental for our ultimate goals, is that it is local – the local behavior of the approximation at a point depends solely on the restriction of f onto a neighbourhood of the point, the size of the neighbourhood being under our full control.

2.2.1 Smoothing kernel

Let $(\mathbf{E}, \langle\cdot,\cdot\rangle)$ be a finite-dimensional Euclidean space, and $\|\cdot\|$ be a norm on $\mathbf{E}$ (not necessarily induced by $\langle\cdot,\cdot\rangle$). Let also $\phi(\cdot)$ (the *smoothing kernel*) be a twice continuously differentiable convex function defined on an open convex set $\mathrm{Dom}\,\phi \subset \mathbf{E}$

with the following properties:

A. $0 \in \mathrm{Dom}\,\phi$ and $\phi(0) = 0$, $\phi'(0) = 0$;

B. There exists a compact convex set $G \subseteq \mathrm{Dom}\,\phi$ such that $0 \in \mathrm{int}\,G$ and $\phi(x) > \|x\|$ for all x from the boundary of G.

C. For some $M_\phi < \infty$ we have

$$\langle e, \nabla^2\phi(h)e\rangle \le M_\phi\|e\|^2 \quad \forall(e \in \mathbf{E}, h \in G). \tag{18}$$

Note that A and B imply that for all $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$, the function $f(x) + \phi(x)$ attains its minimum on the set $\mathrm{int}\,G$. Indeed, for every x from the boundary of G we have $f(x) + \phi(x) \ge f(0) - \|x\| + \phi(x) > f(0) + \phi(0)$, so that the (clearly existing) minimizer of $f + \phi$ on G is a point from $\mathrm{int}\,G$. As a result, for every $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ and $x \in \mathbf{E}$ one has

$$\min_{h\in\mathrm{Dom}\phi}[f(x+h) + \phi(h)] = \min_{h\in\mathrm{int}G}[f(x+h) + \phi(h)], \tag{19}$$

and the right hand side minimum is achieved.

Given a function $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$, we refer to the function

$$\mathcal{S}[f](x) = \min_{h\in\mathrm{Dom}\,\phi}[f(x+h) + \phi(h)] = \min_{h\in G}[f(x+h) + \phi(h)] \tag{20}$$

as to the *smoothing* of f. The properties of this function are summarized in the following

Lemma 2.2.1. *Let $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ and let $\mathcal{S}[f](x)$ be given by* (20)*. The following properties are satisfied*

(0) $\mathcal{S}[f](x) \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$;

(i) $\mathcal{S}[f](x) = f(x + h(x)) + \phi(h(x))$, *where* $h(x) \in \mathrm{int}\,G$ *is such that*

$$f'(x + h(x)) + \phi'(h(x)) = 0 \tag{21}$$

for properly selected $f'(x + h(x)) \in \partial f(x + h(x))$;

(ii) $f(x) \geq \mathcal{S}[f](x) \geq f(x) - \rho_{\|\cdot\|}(G)$, *where*

$$\rho_{\|\cdot\|}(G) = \max_{h \in G} \|h\|;$$

(iii) We have that for all $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$

$$\|\nabla \mathcal{S}[f](x) - \nabla \mathcal{S}[f](y)\|_* \leq M_\phi \|x - y\| \quad \forall x, y \in E. \tag{22}$$

Proof. We prove (i) first, as it is needed for the rest.

(i) Indeed, (22) corresponds to the first-order optimality conditions for optimization problem (20), whose optimum $h(x)$ lies in the interior of G.

(0) First observe that for all $x \in \operatorname{int} X$, $\partial f(x) \subseteq B_{\|\cdot\|_*}(1)$. Now, from (22),

$$\begin{aligned} \nabla S[f](x) &= f'(x + h(x)) + h'(x)[\underbrace{f'(x + h(x)) + \phi(h(x))}_{=0}] \\ &= f'(x + h(x)) \in B_{\|\cdot\|_*}(1). \end{aligned}$$

Finally, by convexity of $\mathcal{S}[f](\cdot)$

$$\mathcal{S}[f](x) - \mathcal{S}[f](y) \leq \langle \nabla \mathcal{S}[f](x), x - y \rangle \leq \|\mathcal{S}[f](x)\|_* \|x - y\| \leq \|x - y\|.$$

The reverse inequality can be analogously bounded, obtaining that $\mathcal{S}[f] \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$.

(ii) By A we have $\phi(h) \geq \phi(0) = 0$, so that $f(x) = f(x) + \phi(0) \geq \mathcal{S}[f](x) = f(x + h(x)) + \phi(h(x)) \geq f(x + h(x)) \geq f(x) - \|h(x)\|$ (recall that $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$), while $h(x) \in G$.

(iii) By the standard approximation argument, it suffices to establish this relation in the case when, in addition to the inclusion $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ and the assumptions A – C on ϕ, f and ϕ are $\mathcal{C}^\infty$ smooth and ϕ is strongly convex. By (21),

$$\mathcal{S}[f](x) = f(x + h(x)) + \phi(h(x)), \tag{23}$$

where $h : E \to G$ is well defined and solves the nonlinear system of equations

$$F(x, h(x)) = 0, \ F(x,h) := f'(x+h) + \phi'(h). \tag{24}$$

We have $\frac{\partial F(x,h)}{\partial h} = f''(x+h) + \phi''(h) \succ 0$, implying by the Implicit Function Theorem that $h(x)$ is smooth. Differentiating the identity $F(x, h(x)) \equiv 0$, we get

$$\begin{aligned} \underbrace{f''(x+h(x))}_{P}[I + h'(x)] + \underbrace{\phi''(h(x))}_{Q} h'(x) = 0 \\ \Leftrightarrow \quad P + (P+Q)h'(x) = 0 \end{aligned}$$

$$\Rightarrow h'(x) = -[P+Q]^{-1}P = [P+Q]^{-1}Q - I.$$

On the other hand, differentiating (23), we get

$$\begin{aligned} \langle \nabla \mathcal{S}[f](x), e \rangle &= \langle f'(x+h(x)), e + h'(x)e \rangle + \langle \phi'(h(x)), h'(x)e \rangle \\ &= \langle f'(x+h(x)), e \rangle + \langle \underbrace{f'(x+h(x)) + \phi'(h(x))}_{=0}, h'(x)e \rangle \\ &= -\langle \phi'(h(x)), e \rangle, \end{aligned}$$

that is,

$$\nabla \mathcal{S}[f](x) = -\phi'(h(x)).$$

As a result, for all e, x, we have, taking into account that P, Q are symmetric positive definite,

$$\begin{aligned} \langle e, \nabla^2 \mathcal{S}[f](x)e \rangle &= -\langle h'(x)e, \phi''(h(x))e \rangle \\ &= -\langle [[P+Q]^{-1}Q - I]e, Qe \rangle \\ &= \langle e, Qe \rangle - \langle e, Q[P+Q]^{-1}Qe \rangle \\ &\leq \langle e, Qe \rangle \leq M_\phi \|e\|^2, \end{aligned}$$

and (22) follows.

□

2.2.2 Approximating a function by smoothing

For $\chi > 0$ and $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$, let

$$\mathcal{S}_\chi[f](x) = \min_{h \in \chi \mathrm{Dom}\,\phi} [f(x+h) + \chi\phi(h/\chi)]. \tag{25}$$

Observe that $\mathcal{S}_\chi[f](\cdot)$ can be obtained as follows:

- We associate with $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ the function $f_\chi(x) = \chi^{-1} f(\chi x)$; observe that this function belongs to $\mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ along with f;
- We pass from f_χ to its smoothing

$$\begin{array}{rcl} \mathcal{S}[f_\chi](x) & = & \min_{g \in \mathrm{Dom}\,\phi} [f_\chi(x+g) + \phi(g)] \\ & = & \min_{g \in \mathrm{Dom}\,\phi} \left[\chi^{-1} f(\chi x + \chi g) + \phi(g)\right] \\ & = & \chi^{-1} \min_{h \in \chi \mathrm{Dom}\,\phi} [f(\chi x + h) + \chi\phi(h/\chi)] \\ & = & \chi^{-1} \mathcal{S}_\chi[f](\chi x). \end{array}$$

It follows that

$$\mathcal{S}_\chi[f](x) = \chi \mathcal{S}[f_\chi](\chi^{-1} x).$$

The latter relation combines with (22) to imply that

$$\|\nabla \mathcal{S}_\chi[f](x) - \nabla \mathcal{S}_\chi[f](y)\|_* \le \chi^{-1} M_\phi \|x - y\| \ \forall x, y.$$

As bottom-line, we have proved the following

Theorem 2.2.2. *Let $(\mathbf{E}, \|\cdot\|)$ be a (finite-dimensional) normed space such that there exists a smoothing kernel ϕ. Then for any convex function $f : \mathbf{E} \to \mathbb{R}$ such that $f \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ and every $\chi > 0$ there exists a smooth (i.e., with Lipschitz continuous gradient) approximation $\mathcal{S}_\chi[f]$ that satisfies:*

S.1. $\mathcal{S}_\chi[f]$ *is convex and Lipschitz continuous with constant 1 w.r.t.* $\|\cdot\|$ *and has a Lipschitz continuous gradient, with constant* M_ϕ/χ*, w.r.t.* $\|\cdot\|$*:*

$$\|\nabla\mathcal{S}_\chi[f](x) - \nabla\mathcal{S}_\chi[f](y)\|_* \le \chi^{-1}M_\phi\|x-y\| \ \forall x,y;$$

S.2. $\sup_{x\in E}|f(x) - \mathcal{S}_\chi[f](x)| \le \chi\rho_{\|\cdot\|}(G)$*. Moreover,* $f(x) \ge \mathcal{S}_\chi[f](x) \ge f(x) - \chi\rho_{\|\cdot\|}(G)$*.*

S.3. $\mathcal{S}_\chi[f]$ *depends on* f *in a local fashion: the value and the derivative of* $\mathcal{S}_\chi[f]$ *at* x *depend only on the restriction of* f *onto the set* $x + \chi G$*.*

2.2.3 Example: q-norm smoothing

Let $n > 1$ and $q \in [2,\infty]$, and consider the case of $\mathbf{E} = \mathbb{R}^n$, endowed with the standard inner product, and $\|\cdot\| = \|\cdot\|_q$. Assume for a moment that $q > 2$, and let r be a real such that $2 < r \le q$. Let also $\theta > 1$ be such that $2\theta/r < 1$. Let us set

$$\begin{array}{rcl} \phi(x) & = & \phi_{r,\theta}(x) = 2\left(\sum_{j=1}^n |x_j|^r\right)^{2\theta/r}, \\ G & = & \{x \in \mathbb{R}^n : \|x\|_q \le 1\}. \end{array} \tag{26}$$

Observe that ϕ is twice continuously differentiable on $\mathrm{Dom}\,\phi = \mathbb{R}^n$ function satisfying A. Besides this, $r \le q$ ensures that $\sum_j |x_j|^r \ge 1$ whenever $\|x\|_q = 1$, so that $\phi(x) > \|x\|_q$ when x is a boundary point of G, which implies B. Besides, by choosing $r = \min[q, 3\ln n]$ and selecting $\theta > 1$ close enough to 1, C is satisfied for $M_\phi = O(1)\min[q, \ln n]$.

Let us prove the latter statement. By definition of ϕ:

$$\begin{aligned}
& \langle e, [\nabla^2\phi(x)]e\rangle && (27)\\
= \;& 4r\theta(2\theta/r-1)(\sum_j |x_j|^r)^{2\theta/r-2}\left[\sum_j |x_j|^{r-1}\mathrm{sign}(x_j)e_j\right]^2 \\
& +4\theta(r-1)(\sum_j |x_j|^r)^{2\theta/r-1}\sum_j |x_j|^{r-2}e_j^2 \\
\leq \;& 4\theta(r-1)(\sum_j |x_j|^r)^{2\theta/r-1}\sum_j |x_j|^{r-2}e_j^2 && (28)\\
\leq \;& 4\theta(r-1)\left[\|x\|_q^r\, n^{1-r/q}\right]^{2\theta/r-1}\left[\sum_j |x_j|^{\frac{(r-2)q}{q-2}}\right]^{\frac{q-2}{q}}\left[\sum_j |e_j|^q\right]^{\frac{2}{q}} && (29)\\
\leq \;& 4\theta(r-1)\left[\|x\|_q^r\, n^{1-r/q}\right]^{2\theta/r-1}\left[\|x\|_q^{\frac{(r-2)q}{q-2}} n^{1-\frac{r-2}{q-2}}\right]^{1-2/q}\|e\|_q^2 && (30)\\
\leq \;& 4\theta(r-1)\|x\|_q^{2\theta-2}\, n^{\frac{2\theta(q-r)}{qr}}\|e\|_q^2,
\end{aligned}$$

Note we used that $2\theta/r < 1$ in (28), the inequality $\sum_{j=1}^n |a_j|^u \leq (\sum_i |a_i|^v)^{u/v} n^{1-u/v}$ (for $0 < u \leq v \leq \infty$ and $u < \infty$) in (29), (30), and the Hölder inequality in (29).

We see that setting $r = \min[q, 3\ln n]$ and choosing $\theta > 1$ close to 1, we ensure the postulated inequalities $2 < r \leq q$, $\theta > 1$, and $2\theta/r < 1$, as well as the relation

$$x \in G \quad \Rightarrow \quad \langle e, [\nabla^2\phi(x)]e\rangle \leq O(1)\min[q, \ln n]\|e\|_q^2 \;\; \forall e \in \mathbb{R}^n, \tag{31}$$

expressing the fact that ϕ, G satisfy assumption C with $M_\phi = O(1)\min[q, \ln n]$.

For the case of $q = 2$, we can set $\phi(x) = 2\|x\|_2^2$ and, as above, $G = \{x : \|x\|_2 \leq 1\}$, clearly ensuring A, B, and the validity of C with $M_\phi = 2$.

Applying the results on smoothing, we get the following

Proposition 2.2.3. *Let $q \in [2, \infty]$ and $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz continuous, with constant 1 w.r.t. the norm $\|\cdot\|_q$, convex function. For every $\chi > 0$, there exists a convex continuously differentiable function $\mathcal{S}_\chi[f](x) : \mathbb{R}^n \to \mathbb{R}$ with the following properties:*

(i) $f(x) \geq \mathcal{S}_\chi[f](x) \geq f(x) - \chi$, *for all* x;

(ii) $\|\nabla \mathcal{S}_\chi[f](x) - \nabla \mathcal{S}_\chi[f](y)\|_{q_*} \leq O(1) \min[q, \ln n] \chi^{-1} \|x - y\|_q$ *for all* x, y;

(iii) *For every* x, *the restriction of* $\mathcal{S}_\chi[f](\cdot)$ *on a small enough neighbourhood of* x *depends solely on the restriction of* f *on the set* $B_q(x, \chi)$.

2.3 Main Result: Lower Complexity Bounds from Local Smoothing

In this section we use the local smoothing developed in Section 2.2.1 to prove our main result, namely, a general lower bound on the oracle complexity of smooth convex minimization.

Proposition 2.3.1. *Let*

I. $\|\cdot\|$ *be a norm on* $\mathbb{R}^n$ *and* X *be a nonempty convex set;*

II. T *be a positive integer and* Γ *be a positive real with the following property:*

There exist T *linear forms* $\langle \xi^i, \cdot \rangle$ *on* $\mathbb{R}^n$, $1 \leq i \leq T$, *such that*

(a) $\|\xi^i\|_* \leq 1$ *for* $i \leq T$, *and*

(b) for every collection $s^T = (s_1, \ldots, s_T)$ *with* $s_i \in \{-1, 1\}$, *it holds*

$$\min_{x \in X} \max_{1 \leq i \leq T} s_i \langle \xi^i, x \rangle \leq -\Gamma; \tag{32}$$

III. M *and* ρ *be positive reals such that for properly selected convex twice continuously differentiable on an open convex set* $\mathrm{Dom}\, \phi \subset \mathbb{R}^n$ *function* ϕ *and a convex compact subset* $G \subset \mathrm{Dom}\, \phi$ *the triple* $(\phi, G, M_\phi = M)$ *satisfies properties A, B, C from Section 2.2.1 and* $\rho_{\|\cdot\|}(G) \leq \rho$.

Then for every $L > 0$, $\kappa \in (1, 2]$, *every local oracle* $\mathcal{O}$ *and every* T*-step method* $\mathcal{A}$ *associated with this oracle for the class of problems* $\mathcal{F}_{\mathbf{E}, \|\cdot\|}(\kappa, L)$, *there exists an instance* $f \in \mathcal{F}_{\|\cdot\|, \mathbf{E}}(\kappa, L)$ *such that*

$$f(x^T(\mathcal{A}, f)) - \mathrm{Opt}(f) \geq \frac{\Gamma^\kappa}{2^{\kappa+1}(\rho M)^{\kappa-1}} \cdot \frac{L}{T^{\kappa-1}}. \tag{33}$$

Proof. 1^0. Let us set

$$\delta = \frac{\Gamma}{2T},\ \chi = \frac{\delta}{2\rho} = \frac{\Gamma}{4T\rho},\ \beta = \frac{L\chi^{\kappa-1}}{2^{2-\kappa}M^{\kappa-1}} = \frac{L\Gamma^{\kappa-1}}{2^{\kappa}(T\rho M)^{\kappa-1}}. \tag{34}$$

2^0. Given a permutation $i \mapsto \sigma(i)$ of $\{1,...,T\}$ and a collection $s^T \in \{-1,1\}^T$, we associate with these data the functions

$$g^{\sigma(\cdot),s^T}(x) = \max_{1\le i\le T}\left[s_i\langle \xi^{\sigma(i)}, x\rangle - (i-1)\delta\right].$$

Observe that all these functions belong to $\mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ due to $\|\xi^j\|_* \le 1$, for $j \le T$, so that the smoothed functions

$$f^{\sigma(\cdot),s^T}(x) = \beta\mathcal{S}_\chi[g^{\sigma(\cdot),s^T}](x) \tag{35}$$

(see Section 2.2.2) are well defined continuously differentiable convex functions on $\mathbb{R}^n$ which, by item S.1 in Section 2.2.2, satisfy that for all x, y in X

$$\|\nabla f^{\sigma(\cdot),s^T}(x) - \nabla f^{\sigma(\cdot),s^T}(y)\|_* \le \beta\chi^{-1}M\|x-y\|.$$

On the other hand, since $f^{\sigma(\cdot),s^T}$ is Lipschitz continuous with constant β w.r.t. $\|\cdot\|$ (see S.1), for all $x, y \in X$

$$\|\nabla f^{\sigma(\cdot),s^T}(x) - \nabla f^{\sigma(\cdot),s^T}(y)\|_* \le 2\beta.$$

Combining these two inequalities, we obtain that for all x, y

$$\|\nabla f^{\sigma(\cdot),s^T}(x) - \nabla f^{\sigma(\cdot),s^T}(y)\|_* \le \beta 2^{2-\kappa}(\chi^{-1}M)^{\kappa-1}\|x-y\|^{\kappa-1}.$$

Recalling the definition of β, we conclude that $f^{\sigma(\cdot),s^T}(\cdot) \in \mathcal{F}_{\mathbf{E},\|\cdot\|}(\kappa, L)$.

3^0. Given a local oracle $\mathcal{O}$ and an associated T-step method $\mathcal{A}$, let us define a sequence $x^1, \dots, x^T$ of points in $\mathbb{R}^n$, a permutation $\sigma(\cdot)$ of $\{1,...,T\}$ and a collection $s^T \in \{-1,1\}^T$ by the following T-step recurrence:

- *Step 1:* x^1 is the first point of the trajectory of $\mathcal{A}$ (this point depends solely on the method and is independent of the problem the method is applied to). We define $\sigma(1)$ as the index i, $1 \leq i \leq T$, that maximizes $|\langle \xi^i, x^1 \rangle|$, and specify $s_1 \in \{-1, 1\}$ in such a way that $s_1 \langle \xi^{\sigma(1)}, x^1 \rangle = |\langle \xi^{\sigma(1)}, x^1 \rangle|$. We set

$$g^1(x) = s_1 \langle \xi^{\sigma(1)}, x \rangle, \; f^1(x) = \beta \mathcal{S}_\chi[g^1](x).$$

- *Step t, $2 \leq t \leq T$:* At the beginning of this step, we have at our disposal the already built points $x^\tau \in \mathbb{R}^n$, distinct from each other integers $\sigma(\tau) \in \{1, ..., T\}$ and quantities $s_\tau \in \{-1, 1\}$, for $1 \leq \tau < t$. At step t, we build x^t, $\sigma(t)$, s_t, as follows. We set

$$g^{t-1}(x) = \max_{1 \leq \tau < t} \left[s_\tau \langle \xi^{\sigma(\tau)}, x \rangle - (\tau - 1)\delta \right],$$

  thus getting a function from $\mathrm{Lip}_{\mathbf{E}, \|\cdot\|}(1)$, and define its smoothing $f^{t-1}(x) = \beta \mathcal{S}_\chi[g^{t-1}](x)$ which, same as above, belongs to $\mathcal{F}_{\mathbf{E}, \|\cdot\|}(\kappa, L)$. We further define

  - x^t as the t-th point of the trajectory of $\mathcal{A}$ as applied to f^{t-1},
  - $\sigma(t)$ as the index i that maximizes $|\langle \xi^i, x^t \rangle|$, over $i \leq T$ *distinct from* $\sigma(1), \ldots, \sigma(t-1)$,
  - $s_t \in \{-1, 1\}$ such that $s_t \langle \xi^{\sigma(t)}, x^t \rangle = |\langle \xi^{\sigma(t)}, x^t \rangle|$

  thus completing step t.

After T steps of this recurrence, we get at our disposal a sequence $x^1, \ldots, x^T$ of points from $\mathbb{R}^n$, a permutation $\sigma(\cdot)$ of indexes $1, \ldots, T$ and a collection $s^T = (s_1, \ldots, s_T) \in \{-1, 1\}^T$; these entities define the functions

$$g^T = g^{\sigma(\cdot), s^T}, \; f^T = \beta \mathcal{S}_\chi[g^{\sigma(\cdot), s^T}].$$

4^0. We claim that $x^1, \ldots, x^T$ is the trajectory of $\mathcal{A}$ as applied to f^T. By construction, x^1 indeed is the first point of the trajectory of $\mathcal{A}$ as applied to f^T. In view of

this fact, taking into account the definition of x^t and the locality of the oracle $\mathcal{O}$, all we need to support our claim is to verify that for every t, $2 \le t \le T$, the functions f^T and f^{t-1} coincide in some neighbourhood of x^{t-1}. By construction, we have that for $t \le r \le T$

$$s_r\langle \xi^{\sigma(r)}, x^{t-1}\rangle \le |\langle \xi^{\sigma(t-1)}, x^{t-1}\rangle| = s_{t-1}\langle \xi^{\sigma(t-1)}, x^{t-1}\rangle. \tag{36}$$

Also

$$g^T(x) = \max\big[g^{t-1}(x), \underbrace{\max_{t\le r\le T}[s_r\langle \xi^{\sigma(r)}, x\rangle - (r-1)\delta]}_{=:g_t(x)}\big], \tag{37}$$

and

$$g^{t-1}(x^{t-1}) \ge s_{t-1}\langle \xi^{\sigma(t-1)}, x^{t-1}\rangle - (t-2)\delta.$$

Invoking (36), we get

$$\begin{aligned} t \le r \le T \quad &\Rightarrow \quad g^{t-1}(x^{t-1}) \ge [s_r\langle \xi^{\sigma(r)}, x^{t-1}\rangle - (r-1)\delta] + \delta \\ &\Rightarrow \quad g^{t-1}(x^{t-1}) \ge g_t(x^{t-1}) + \delta. \end{aligned}$$

Since both g^{t-1} and g_t belong to $\mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$, it follows that $g^{t-1}(x) \ge g_t(x)$ in the $\|\cdot\|$-ball B of radius $\delta/2$ centered at x^{t-1}, whence, by (37),

$$x \in B \quad \Rightarrow \quad g^T(x) = g^{t-1}(x).$$

From $\chi\rho = \delta/2$ we have that $g^{t-1} \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ and $g^T \in \mathrm{Lip}_{\mathbf{E},\|\cdot\|}(1)$ coincide on the set $x^{t-1} + \chi G$, whence, as we know from item S.3 in Section 2.2.2, $f^{t-1}(\cdot) = \beta\mathcal{S}_\chi[g^{t-1}](\cdot)$ and $f^T(\cdot) = \beta\mathcal{S}_\chi[g^T](\cdot)$ coincide in a neighborhood of x^{t-1}, as claimed.

5^0. We have

$$\begin{aligned} g^T(x^T) &\ge s_T\langle \xi^{\sigma(T)}, x^T\rangle - (T-1)\delta \\ &= |\langle \xi^{\sigma(T)}, x^T\rangle| - (T-1)\delta \\ &\ge -(T-1)\delta, \end{aligned}$$

whence, by item S.2 in Section 2.2.2, $\mathcal{S}_\chi[g^T](x^T) \geq -(T-1)\delta - \chi\rho \geq -T\delta = -\Gamma/2$, implying that

$$f^T(x^T) \geq -\beta\Gamma/2.$$

On the other hand, by (32) there exists $x^* \in X$ such that

$$g^T(x^*) \leq \max_{1\leq i\leq T} s_i\langle\xi^{\sigma(i)}, x^*\rangle \leq -\Gamma,$$

whence $\mathcal{S}_\chi[g^T](x^*) \leq g^T(x^*) \leq -\Gamma$ and thus $\mathrm{Opt}(f^T) \leq f^T(x^*) \leq -\beta\Gamma$. Since, as we have seen, $x^1, \ldots, x^T$ is the trajectory of $\mathcal{A}$ as applied to f^T, x^T is the approximate solution generated by $\mathcal{A}$ as applied to f^T, and we see that the inaccuracy of this solution, in terms of the objective, is at least $\frac{\beta\Gamma}{2} = \frac{\Gamma^\kappa}{2^{\kappa+1}(\rho M)^{\kappa-1}} \cdot \frac{L}{T^{\kappa-1}}$, as required. Besides this, f^T is of the form $f^{\sigma(\cdot),s^T}$, and we have seen that all these functions belong to $\mathcal{F}_{\mathbf{E},\|\cdot\|}(\kappa, L)$. □

As an immediate consequence, we can establish a lower complexity bound for domains with arbitrary radius

Corollary 2.3.2. *Let $R > 0$. Under the notation and assumptions of Proposition 2.3.1, the minimax risk of the problem class $\mathcal{P} = (\mathcal{F}_{\mathbf{E},\|\cdot\|}(\kappa, L), RX)$ satisfies the lower bound*

$$\mathrm{Risk}_{\mathcal{F},RX,\mathcal{O}}(T) \geq \frac{\Gamma^\kappa}{2^{\kappa+1}(\rho M)^{\kappa-1}} \cdot \frac{LR^\kappa}{T^{\kappa-1}}. \tag{38}$$

Proof. We observe that re-scaling the domain by R leads to the modified bound (32)

$$\Gamma_{RX} \geq \Gamma R.$$

Therefore, by Proposition 2.3.1 we obtain the desired lower bound

$$\mathrm{Risk}_{\mathcal{F},X,\mathcal{O}}(T) \geq \frac{\Gamma^\kappa}{2^{\kappa+1}(\rho M)^{\kappa-1}} \cdot \frac{LR^\kappa}{T^{\kappa-1}}.$$

□

2.4 *The ℓ_p/ℓ_q Setting*

In this Section we study the particular case when X contains a unit p-ball B_p^n and $\|\cdot\| = \|\cdot\|_q$, for $1 \leq p, q \leq \infty$. Therefore, for notational convenience we will denote $\mathcal{F} = \mathcal{F}_q(\kappa, L)$ and $X = B_p^n$.[1]

From Proposition 2.2.3 we have a direct way to construct smooth convex functions when $2 \leq q \leq \infty$. First, we use this result to find explicit lower bounds on the complexity. Next, we provide alternative techniques when $1 \leq q < 2$.

2.4.1 Case $q \geq 2$

(i) **Standard basis construction:** Let $T \leq n$, and consider the set of linear forms over $\mathbb{R}^n$ given by $\xi^t := e_t$ with $t = 1, \ldots, T$, i.e., the first T canonical vectors. It is easy to see that $\|\xi^t\|_{q_*} = 1$ and

$$\Gamma_{\text{std}} = -\min_{x \in B_p^n} \max_{t \in [T]} s_t \langle \xi^t, x \rangle \geq 1/T^{1/p}.$$

We can readily use this family of linear functionals in Proposition 2.3.1 leading to the following lower bound

$$\text{Risk}_{\mathcal{F}, B_p^n, \mathcal{O}}(T) \geq \frac{O(1)}{[\min\{q, \ln n\}]^{\kappa-1}} \frac{1}{T^{\kappa(1+1/p)-1}}.$$

(ii) **Partitioned basis construction:** Let $T \leq n$, and consider a subpartition of $[n]$ into T disjoint subsets $A_1, \ldots, A_T$, each of them having size $m = \lfloor n/T \rfloor$. It is easy to see that $n/T \geq m \geq n/(2T)$; therefore, the n-dimensional vectors

$$\xi^t = m^{-1/q_*} \mathbb{1}_{A_t}$$

are such that $\|\xi^t\|_{q_*} = 1$. Now let $x \in \mathbb{R}^n$ be a vector such that on each subset A_t its components have value $-\text{sign}(s_t)/n^{1/p}$, therefore $\|x\|_p = 1$. With this

[1] However, the reader must observe that the lower bounds we obtain will still hold for any X containing B_p^n.

choice, we obtain that

$$s_t\langle \xi^t, x\rangle \;=\; -mn^{-1/p}m^{-1/q_*} \le -\left(\frac{n}{2T}\right)^{1/q} n^{-1/p} = -\frac{1}{2^{1/q}}\frac{n^{1/q-1/p}}{T^{1/q}}.$$

This proves that

$$\Gamma_{\text{prt}} := -\min_{x\in B_p^n}\max_{t\in[T]} s_t\langle \xi^t, x\rangle \ge \frac{n^{1/q-1/p}}{2^{1/q}T^{1/q}}.$$

Using this family of linear functionals ξ^t in Proposition 2.3.1 we obtain a lower bound

$$\text{Risk}_{\mathcal{F},B_p^n,\mathcal{O}}(T) \ge \frac{O(1)}{[\min\{q,\ln n\}]^{\kappa-1}}\frac{n^{\kappa(1/q-1/p)}}{T^{\kappa(1+1/q)-1}}.$$

The two lower bounds stated above are valid when $q \ge 2$. Let us now see which one is tighter depending on p. For this, observe that for determining which lower bound is better we only need to compare the value of Γ for both constructions. This way,

$$\frac{\Gamma_{\text{std}}(T)}{\Gamma_{\text{prt}}(T)} = \tilde{\Theta}(1)\left(\frac{T}{n}\right)^{\frac{1}{q}-\frac{1}{p}}.$$

Therefore, for $T \le n$, if $q \ge p$ we obtain a tighter lower bound by the standard basis construction; otherwise, we should use the partitioned basis construction.

This leads to the following

Corollary 2.4.1. *Let $2 \le q \le \infty$, $1 \le p \le \infty$, $\kappa \in (1,2]$, $L > 0$, and let $X \subset \mathbb{R}^n$ be a convex set containing the ball $B_p(R)$. Then, for $T \le n$ and every local oracle $\mathcal{O}$, the minimax risk of the family of problems $\mathcal{P} = (\mathcal{F}, X)$ with $\mathcal{F} = \mathcal{F}_q(\kappa, L)$ is given by*

(i) If $p \ge q$

$$\text{Risk}_{\mathcal{F},X,\mathcal{O}}(T) \ge O(1)\frac{n^{\kappa(1/q-1/p)}}{[\min\{q,\ln n\}]^{\kappa-1}}\frac{LR^\kappa}{T^{\kappa(1+1/q)-1}} \tag{39}$$

(ii) If $q > p$

$$\text{Risk}_{\mathcal{F},X,\mathcal{O}}(T) \ge O(1)\frac{1}{[\min\{q,\ln n\}]^{\kappa-1}}\frac{LR^\kappa}{T^{\kappa(1+1/p)-1}}. \tag{40}$$

2.4.2 Case $q < 2$

Local smoothing provides lower bounds when $q \geq 2$. It is known that finding smoothing kernels for further ranges of q is unlikely, which is ultimately related to results in Banach space geometry [2]. In this subsection we study this case by alternative methods based on ideas from convex geometry.

Specifically, our proofs are based on studying random projections of the feasible domain. These bodies will contain a domain for which we know lower complexity bounds, and by lifting those families of instances to the whole domain we can use the results from the previous subsection to derive lower complexity bounds.

Finally, for the case $1 \leq p < q \leq 2$, we obtain a lower bound based only on the ideas of the previous subsection. However, it turns out that this lower bound does not match upper bounds obtained by existing algorithms. This leads to an interesting open question, regarding optimal algorithms in this range.

2.4.2.1 Case $p \geq q$

Our first construction is based on random projection of the feasible domain. For this we will need a well known Lemma on random projections; for the sake of completeness, we present the proof.

Lemma 2.4.2. *There exists an absolute constant $0 < \alpha < 1$ such that for all $n \geq 1/\alpha$ and all T, $1 \leq T \leq \alpha n$, a Gaussian random matrix $G \in \mathbb{R}^{T \times n}$ i.e., a matrix with iid $\mathcal{N}(0,1)$ entries), satisfies with probability $\geq 1/2$ the relation*

$$\alpha n B_2^T \subseteq G B_\infty^n.$$

Proof. From now on c_i stand for appropriate positive absolute constants.

1^0. Let us consider an arbitrary, but fixed, $y \in \mathbb{R}^T$ with $\|y\|_2 = 1$. We will first prove that there exist $c_1, c_2, c_3 > 0$ such that

$$\mathbb{P}[\|G^T y\|_1 > c_1 n] \leq \exp(-c_3 n) \quad \& \quad \mathbb{P}[\|G^T y\|_1 < c_2 n] \leq \exp(-c_3 n). \tag{41}$$

To prove this, observe that given y, the vector Gy has iid $\mathcal{N}(0,1)$ coordinates ξ_i, $i = 1, \ldots, n$, and thus

$$\mathbb{E}[\exp(\|G^T y\|_1)] = \exp(c_4 n), \quad \mathbb{E}[\exp(-\|G^T y\|_1)] = \exp(-c_5 n).$$

Setting $c_1 = 2c_4$, we obtain by Markov's inequality

$$\mathbb{P}[\|G^T y\|_1 > c_1 n] \leq \exp(-c_1 n)\mathbb{E}[\exp(\|G^T y\|_1)] \leq \exp(c_4 n - c_1 n) = \exp(-c_4 n).$$

Similarly, if $c_2 = c_5/2$ we get

$$\begin{aligned} \mathbb{P}[\|G^T y\|_1 < c_2 n] &= \mathbb{P}[-\|G^T y\|_1 > -c_2 n] \leq \exp(c_2 n)\mathbb{E}[\exp(-\|G^T y\|)] \\ &\leq \exp(c_2 n - c_5 n) \qquad = \exp(-c_2 n). \end{aligned}$$

Finally, choosing $c_3 := \min\{c_2, c_4\}$ we obtain the desired bound (41).

2^0. Next, we want to generalize our result for fixed y to hold uniformly on the unit sphere. For this, consider a minimal cardinality ε-net in $\|\cdot\|_2$, on the $\|\cdot\|_2$-unit T-dimensional sphere, that we call Γ^T. It is a well-known fact that $|\Gamma^T| \leq (c_7\varepsilon)^{-T}$.

Let $Z = GB_\infty^n$, and

$$\begin{aligned} \theta(y) &:= \max_{z \in Z} \langle y, z \rangle : \mathbb{R}^T \to \mathbb{R}, \\ M &:= \max_{\|y\|_2 = 1} \theta(y), \\ \mu &:= \min_{\|y\|_2 = 1} \theta(y). \end{aligned}$$

Note that $\theta(\cdot)$ is Lipschitz continuous w.r.t $\|\cdot\|_2$ with constant M on B_2^T. We consider the event E defined as

$$\forall y \in \Gamma_{1/2}^T : \quad \|G^T y\|_1 \leq c_1 n.$$

Since $|\Gamma_{1/2}^T| \leq \exp(c_8 T)$, by the union bound

$$\mathbb{P}(E) \geq 1 - |\Gamma^T| \exp(-c_3 n) \geq 1 - \exp(c_8 T - c_3 n).$$

From this, it is easy to see that on event E we have $M \leq 2c_1 n$. For this, let $\bar{y}$ be a unit vector such that $\theta(\bar{y}) = M$, and $\tilde{y}$ a point from Γ^T with $\|\tilde{y} - \bar{y}\|_2 \leq 1/2$. By Lipschitz continuity of $\theta(\cdot)$, we have

$$\theta(\tilde{y}) \geq \theta(\bar{y}) - M\|\bar{y} - \tilde{y}\|_2 \geq M - M/2 = M/2,$$

thus, when E takes place, we have

$$M \leq 2\theta(\tilde{y}) = 2\max_{z \in Z} \tilde{y}^T z = 2\max_{w \in B_\infty^n} \langle w, G^T y \rangle = 2\|G^T y\|_1 \leq 2c_1 n.$$

Now let $c_9 < 1$ be such that $c_2 - 2c_9 c_1 \geq c_2/2$, and let F be the event

$$\forall y \in \Gamma_{c_9}^T : \|G^T y\|_1 \geq c_2 n.$$

Since $|\Gamma_{c_9}^T| \leq \exp(c_{10} T)$ for properly selected c_{10}, by the union bound

$$\mathbb{P}[F] \geq 1 - |\Gamma_{c_9}^T| \exp(-c_3 n) \geq 1 - \exp(c_{10} T - c_3 n).$$

We prove now that when event F takes place, we have $\mu \geq c_2 n - c_9 M$. Let $\bar{y}$ of unit norm be such that $\theta(\bar{y}) = \mu$, and let $\tilde{y} \in \Gamma_{c_9}^T$ satisfying $\|\bar{y} - \tilde{y}\|_2 \leq c_9$. By Lipschitz continuity of $\theta(\cdot)$, we obtain

$$\theta(\tilde{y}) \leq \theta(\bar{y}) + M\|\bar{y} - \tilde{y}\|_2 \leq \mu + c_9 M,$$

which implies

$$\mu + c_9 M \geq \theta(\tilde{y}) = \max_{z \in Z} \langle \tilde{y}, z \rangle = \max_{w \in B_\infty^n} \langle \tilde{y}, Gw \rangle = \max_{w \in B_\infty^n} \langle w, G^T \tilde{y} \rangle = \|G^T y\|_1 \geq c_2 n.$$

Observe that when both E and F take place, which happens with probability $1 - \exp(c_8 T - c_3 n) - \exp(c_{10} T - c_3 n)$, the inequalities $M \leq 2c_1 n$ and $\mu \geq c_2 n - c_9 M \geq c_2 n - 2c_1 c_9 n \geq c_2 n/2$ are satisfied, and therefore

$$\mathbb{P}\left[\frac{c_2 n}{2} \leq \max_{z \in G B_\infty^n} \langle y, z \rangle \leq 2c_1 n, \quad \forall \|y\|_2 = 1\right] \geq 1 - \exp(c_8 T - c_3 n) - \exp(c_{10} T - c_3 n).$$

Selecting $\alpha \leq c_2/2$, we have that for any $T \leq \alpha n$ the probability above is lower bounded by $1 - 2\exp(-c_{12}n)$, while the inequalities in the description of the event imply that

$$\frac{c_2 n}{2} B_2^T \subseteq GB_\infty^n \subseteq 2c_1 n B_2^T,$$

so that $\alpha n B_2^T \subset GB_\infty^n$ with probability $\geq 1 - 2\exp\{-c_{12}n\}$, which is $\geq 1/2$ for large enough values of n. Reducing, if necessary, α (but keeping it positive absolute constant), we can ensure that all $n \geq 1/\alpha$ (which indeed is so whenever $1 \leq T \leq \alpha n$) are "large enough." □

The next result provides lower complexity bounds for ℓ_p/ℓ_q-settings, and crucially relies on the Lemma above.

Theorem 2.4.3. *Let $1 \leq q \leq 2$ and $1 \leq p \leq \infty$, and let domain $X \subset \mathbb{R}^n$ contain the p-ball $B_p^n(R)$ of radius $R > 0$. There exists an absolute constant $\alpha \in (0,1)$ such that for all n, T with $1 \leq T \leq \alpha n$ and arbitrary local oracle $\mathcal{O}$ for the family $\mathcal{F} = \mathcal{F}_q(\kappa, L)$, the minimax risk, taken w.r.t. $\mathcal{O}$, of the class of problems $\mathcal{P} = (\mathcal{F}, X)$ satisfies*

$$\mathrm{Risk}_{\mathcal{F},X,\mathcal{O}}(T) \geq O(1) \frac{n^{\kappa(1/q-1/p)}}{[\ln n]^{\kappa-1}} \frac{LR^\kappa}{T^{\frac{3\kappa}{2}-1}}. \tag{42}$$

Proof. First, note that by Corollary 2.3.2, it suffices to prove the result for the particular case when $L = R = 1$, which we assume from now on.

$\mathbf{1}^0$ By Lemma 2.4.2, for properly selected absolute constant $\alpha \in (0,1)$ and all T, $1 \leq \alpha n$, random Gaussian $T \times n$ matrix G with probability $\geq 1/2$ ensures that $\alpha n B_2^T \subset GB_\infty^n$. Since $B_\infty^n \subset n^{1/p} B_p^n$, we conclude that

$$\alpha n^{1/p_*} B_2^T \subseteq GB_p^n. \tag{43}$$

with probability $\geq 1/2$. Further, denoting by g_t^T rows of G, we have

$$\|G\|_{q\to\infty} = \max_{t\leq T} \|g_t\|_{q_*} \leq n^{1/q_*} \max_{i,t} |G_{it}|,$$

whence, setting

$$u = \sqrt{2\ln(8Tn)},$$

we get

$$\mathbb{P}\{\|G\|_{q\to\infty} > n^{1/q_*}u\} \le \mathbb{P}\{\max_{i,t}|G_{ij}| > u\} \le 2Tn\tfrac{1}{\sqrt{2\pi}}\int_u^\infty \exp\{-s^2/2\}ds$$
$$\le 2Tn\exp\{-u^2/2\} \le 1/4.$$

We see that for Gaussian matrix G relation (43) hods true with probability at least $1/2$, while $\|G\|_{q\to\infty} \le n^{1/q_*}u$ with probability at least $3/4$. We conclude that for properly selected absolute constants $\alpha \in (0,1)$ and n_0, for every $n \ge n_0$ and every $T \le \alpha n$ there exists a matrix $\bar{G}$ such that

$$\|\bar{G}\|_{q\to\infty} \le n^{1/q^*}u \quad \& \quad \alpha n^{1/p_*}B_2^T \subset \bar{G}B_p^n \subset \bar{G}X \qquad [u = \sqrt{2\ln(8Tn)}]$$

(recall that X contains $\|\cdot\|_p$-ball of radius R and that we are under normalization $R = L = 1$). Setting $r = \frac{\alpha n^{1/p_*}}{T^{1/2}}$ and observing that $rB_\infty^T \subset \alpha n^{1/p_*}B_2^T$, we conclude that

$$\begin{gathered} rB_\infty^T \subset \bar{G}X \ \&\ \|G\|_{q\to\infty} \le n^{1/q_*}u \\ [r = \tfrac{\alpha n^{1/p_*}}{T^{1/2}}, u = \sqrt{2\ln(8Tn)}] \end{gathered} \tag{44}$$

2^0 Now let

$$\Upsilon = \frac{1}{\left[n^{1/q_*}u\right]^\kappa},$$

and let $\mathcal{F}'$ be the family of functions on $\mathbb{R}^n$ given by

$$\mathcal{F}' = \{\tilde{f}(x) = f(\bar{G}x) : f \in \mathcal{F}_\infty^T(\kappa,\Upsilon)\}.$$

It is immediately seen that $\mathcal{F}'$ is contained in $\mathcal{F}_q(\kappa,1)$.

> Indeed, it suffices to note that if $(\mathbf{E},\|\cdot\|_\mathbf{E})$, $(\mathbf{F},\|\cdot\|_\mathbf{F})$ are two normed Euclidean spaces, and $x \mapsto Hx$ is a linear map from $\mathbf{E}$ to $\mathbf{F}$, and g is convex smooth function on $\mathbf{F}$ with smoothness parameters (κ,M) w.r.t.

$\|\cdot\|_{\mathbf{F}}$, then the function $h(x) := g(Hx)$ is convex with smoothness parameters $(\kappa, M\|G\|_{\|\cdot\|_{\mathbf{E}}\mapsto\|\cdot\|_{\mathbf{F}}})$:

$$\|\nabla h(x) - \nabla h(y)\|_{\mathbf{E},*} = \|H^*[\nabla g(Hx) - \nabla g(Hy)]\|_{\mathbf{E},*}$$
$$\leq \underbrace{\|H^*\|_{\|\cdot\|_{\mathbf{F},*}\to\|\cdot\|_{\mathbf{E},*}}}_{=\|H\|_{\|\cdot\|_{\mathbf{E}}\to\|\cdot\|_{\mathbf{F}}}} M\|Hx - Hy\|_{\mathbf{F}}^{\kappa-1} \leq \|H\|^{\kappa}_{\|\cdot\|_{\mathbf{E}}\to\|\cdot\|_{\mathbf{F}}} M\|x-y\|_{\mathbf{E}}^{\kappa}.$$

Specifying H as $\bar{G}$, $(\mathbf{E}, \|\cdot\|_{\mathbf{E}})$ as ℓ_q^n and $(\mathbf{F}, \|\cdot\|_{\mathbf{F}})$ as ℓ_∞^T, we conclude that whenever $f \equiv g \in \mathcal{F}_\infty^T(\kappa, \mathrm{Y})$, it holds

$$\tilde{f}(x) = f(\bar{G}x) \in \mathcal{F}_q(\kappa, \|\bar{G}\|_{q\to\infty}\mathrm{Y}),$$

while $\|\bar{G}\|^{\kappa}_{q\to\infty}\mathrm{Y} \leq 1$ by (44) and the definition of Y.

By Corollary 2.4.1 as applied to the class $\mathcal{Q}$ of problems $\min_{u\in\bar{G}X} f(u)$, $f \in \mathcal{F}_\infty(\kappa, \mathrm{Y})$ equipped with the universal local oracle, taking into account that $\bar{G}X$ contains the box rB_∞^T, the T-step risk of $\mathcal{Q}$ admits the lower bound

$$\mathrm{Risk}^{\mathcal{Q}}(T) \geq O(1)\frac{\mathrm{Y}r^\kappa}{[\ln T]^{\kappa-1}T^{\kappa-1}} = O(1)\frac{n^{\kappa(\frac{1}{q}-\frac{1}{p})}}{[\ln n]^{\kappa-1}T^{\frac{3\kappa}{2}-1}} \qquad (*)$$

It is intuitive (we prove this claim in the next paragraph) that the T-step risk, taken w.r.t. arbitrary local oracle, when solving problems from the class $(\mathcal{F}', X)$, that is, problems which are obtained from problems belonging to $\mathcal{Q}$ by lifting, cannot be less than the T-step risk of $\mathcal{Q}$ taken w.r.t. the maximal local oracle; since the class of interest $(\mathcal{F}, X)$ is only larger than $(\mathcal{F}', X)$ due to $\mathcal{F}' \subset \mathcal{F}$, we conclude that the right hand side in $(*)$ lower bounds the T-step risk of the class $(\mathcal{F}, X)$, exactly as stated in (42) in the normalized case $L = R = 1$. The proof, modulo the above "intuitive claim," is completed.

$\mathbf{3}^0$ It remains to prove our claim about risk's behavior under lifting. Observe, first, that the claim we intend to justify indeed needs a justification: we cannot just argue that solving "lifted" problems – those with objectives from the family $\mathcal{F}'$ - cannot be simpler than solving problems from $\mathcal{Q}$ due to the fact that the problems

from the latter family can be reduced to those from the former one; we should specify the local oracles associated with the families in question, and to ensure that lifting does not simplify problems because, e.g., the oracle for the lifted family is more informative than the oracle for the original family.

The justification here is as follows. Consider the universal local oracle (see Section 1.2.3 in the Introduction) for family $\mathcal{Q}$. It is easy to see that this oracle induces the universal local oracle for the lifted family $\mathcal{F}'$; with this in mind, it is immediately seen that any universal-oracle-based T-step method A' for solving problems from the family $\mathcal{F}'$ induces a universal-oracle-based T-step method A for solving problems from the family $\mathcal{Q}$ in such a way that the trajectory $y^1, y^2, \ldots$ of A on a problem $\min_{u \in \bar{G}X} f(u)$ is linked to the trajectory $x^1, x^2, \ldots$ of A' on the problem $\min_{x \in X} f(\bar{G}x)$ by the relation $y^t = Gx^t$. Consequently, when the universal oracles are used, a lower bound on the T-step minimax risk of $\mathcal{Q}$ automatically is a lower bound on the same quantity for the class $(\mathcal{F}', X)$. Finally, since the lower bound holds for the universal oracle associated with $(\mathcal{F}', X)$, it holds when the universal oracle is replaced with any local one (since the universal oracle mimics any other local oracle). □

2.4.2.2 *Improvements in the case $p \leq q \leq 2$*

While Theorem 2.4.3 covers the case of $q \leq 2$ and all $p \in [1, \infty]$, in the range $1 \leq p \leq q \leq 2$ it can be improved:

Theorem 2.4.4. *Let $1 \leq q \leq 2$ and $1 \leq p \leq q$, and let domain $X \subset \mathbb{R}^n$ contain the p-ball $B_p^n(R)$ of radius $R > 0$. With the same absolute constant α as in Theorem 2.4.3, for all n, T such that $1 \leq T \leq \alpha n$, and for every local oracle $\mathcal{O}$ for the family $\mathcal{F} = \mathcal{F}_q(\kappa, L)$, the minimax risk, taken w.r.t. $\mathcal{O}$, of the class of problems $\mathcal{P} = (\mathcal{F}, X)$ satisfies*

$$\mathrm{Risk}_{\mathcal{F},X,\mathcal{O}}(T) \geq O(1) \frac{LR^\kappa}{[\ln n]^{\kappa-1} T^{\kappa[\frac{3}{2}+\frac{1}{p}-\frac{1}{q}]-1}}. \tag{45}$$

Proof. Under the premise of Theorem, let $\bar{n} = \lfloor T/\alpha \rfloor$, so that $1 \leq T \leq \alpha\bar{n}$ and $\bar{n} \leq n$, let $\bar{X}$ be the orthogonal projection of X on the plane of the first $\bar{n}$ variables, and let $\bar{\mathcal{F}}$ be the family of all functions from $\mathcal{F}_q(\kappa, L)$ which depend on the first $\bar{n}$ variables only. Clearly, the T-step risk of the class of problems $\bar{\mathcal{P}} = (\bar{\mathcal{F}}, \bar{X})$, taken w.r.t. the universal oracle, can be only smaller than the risk in the left hand side of (45). On the other hand, the former risk clearly is exactly the same as the T-step risk of the problem class considered in Theorem 2.4.3, with $\bar{n}$ in the role of n, $\bar{X}$ in the role of X, and all other entities in the premise of Theorem 2.4.3 being the same as for Theorem 2.4.4. In view of this observation and Theorem 2.4.3, the right hand side in (42), with $\bar{n}$ replacing n, lower-bounds the risk in (45); as it is immediately seen, the bound in question is nothing but (45) (note that by construction T is within an absolute constant factor of $\bar{n}$). □

Note that in the context of Theorem 2.4.3, replacing n with $O(1)T$ is legitimate independently of whether $p \leq q$ or $p \geq q$; this modification, however, improves the bound only when $p \leq q$.

2.5 Consequences

In this section we compare our new lower bounds on complexity of large-scale smooth convex optimization with upper bounds provided by existing algorithms in the ℓ_p/ℓ_q setting, assuming that the feasible domain X of the instances in question is $\|\cdot\|_p$-ball of radius R. Thus, we will speak about the class $\mathcal{P}^n_{p,q}$ of all convex optimization problems of the form

$$\min_{x \in \mathbb{R}^n, \|x\|_p \leq R} f(x)$$

with objectives f running through the family $\mathcal{F}_q(\kappa, L)$ on $\mathbb{R}^n$, for some fixed $\kappa \in (1, 2]$ and $L > 0$. We assume that the class is equipped with local oracle $\mathcal{O}$ which is "at least as powerful" as the first order oracle; this assumption is important as

far as upper complexity bounds are concerned). We also will be interested in the large-scale regime – in the bounds for T-step risk $\mathrm{Risk}(T)$ in the range

$$1 \leq T \leq \alpha n,$$

where $\alpha \in (0,1)$ is the absolute constant from Lemma 2.4.2

For reader's convenience, we reproduce here upper complexity bounds from Corollary 1.2.3, taking into account the parameter responsible for smoothness (now called q) in Corollary 1.2.3 was called p, and that what was called R in Corollary 1.2.3 – the diameter of the feasible domain in the norm responsible for smoothness quantification is in our current situation the diameter of $\|\cdot\|_p$-ball of radius R in $\mathbb{R}^n$ taken w.r.t. the $\|\cdot\|_q$-norm, that is, the diameter is $2Rn^{\max[1/q-1/p,0]}$. With this in mind, and taking into account that we are in the case of $T \leq O(1)n$, the upper complexity bounds from Corollary 1.2.3 now read

$$\begin{aligned}
&(a)\ 1 \leq q \leq 2 \Rightarrow \ \mathrm{Risk}(T) \leq O(1)\left(\min\left[\tfrac{1}{q-1}, \ln(n)\right]\right)^{\kappa/2} \frac{n^{\kappa\max[1/q-1/p,0]} L R^\kappa}{T^{\frac{3}{2}\kappa-1}};\\
&(b)\ 2 \leq q \leq \infty \Rightarrow \mathrm{Risk}(T) \leq O(1)\left(\min[q, \ln(n)]\right)^{\kappa} \frac{n^{\kappa\max[1/q-1/p,0]} L R^\kappa}{T^{\kappa(1+\frac{1}{q})-1}}.
\end{aligned} \tag{46}$$

2.5.1 Tightness of bounds

We are about to compare these upper complexity bounds with the lower bounds derived in this Chapter, specifically, the bounds given by Corollary 2.4.1 and Theorems 2.4.3, 2.4.4. We shall quantify the tightness by the ratios $\mathcal{R}(T)$ of the upper bounds (46) on the T-step risk to our lower bounds on this risk, *skipping factors polynomial in* $\ln(n)$ and using $\preceq$, $\approx$ and $\succeq$ instead of $\geq$, $=$ and $\leq$ to stress that the relations are within factors depending solely on n and polynomial in $\ln(n)$; note in all $\approx$ and $\succeq$ to follow, the degrees of the hidden polynomial in $\ln(n)$ factors in fact do not exceed 3.

Case of $p \le q$, $q \le 2$. Here the lower bound on T-risk is given by Theorem 2.4.4 and reads

$$\text{Risk}(T) \succeq \frac{LR^{\kappa}}{T^{\kappa[\frac{3}{2}+\frac{1}{p}-\frac{1}{q}]-1}}$$

and the tightness of the bound satisfies

$$\mathcal{R}(T) \preceq T^{\kappa(1/p-1/q)}.$$

Case of $p \le q$, $q \ge 2$. Here the lower bound on T-risk is given by Corollary 2.4.1 and reads

$$\text{Risk}(T) \succeq \frac{LR^{\kappa}}{T^{\kappa[1+\frac{1}{p}]-1}}$$

and the tightness of the bound satisfies

$$\mathcal{R}(T) \preceq T^{\kappa(1/p-1/q)}.$$

Case of $p \ge q$, $q \le 2$. Here the lower bound on T-risk is given by Theorem 2.4.3 and reads

$$\text{Risk}(T) \succeq \frac{n^{\kappa[\frac{1}{q}-\frac{1}{p}]}LR^{\kappa}}{T^{\frac{3}{2}\kappa-1}}$$

and the bound is nearly tight

$$\mathcal{R}(T) \approx 1.$$

Case of $p \ge q$, $q \ge 2$. Here the lower bound on T-risk is given by Corollary 2.4.1 and reads

$$\text{Risk}(T) \succeq \frac{n^{\kappa[\frac{1}{q}-\frac{1}{p}]}LR^{\kappa}}{T^{\kappa[1+\frac{1}{q}]-1}}$$

and the bound is nearly tight:

$$\mathcal{R}(T) \approx 1.$$

2.5.2 Comments

Some comments are in order.

A. We see that as far as large-scale case is concerned, *in the range $p \geq q$, our lower complexity bounds are tight within logarithmic in n factor*. This, in particular, implies near-optimality in this range of Nesterov's Fast Gradient algorithm (in its version, developed in [22, Section 2.3], adjusted to convex objectives with Hölder continuous gradients and smoothness quantification taken w.r.t. $\|\cdot\|_q$-norms, $1 \leq q \leq \infty$).

B. In the case of $p = q = \infty$, the upper complexity bounds (46) can be achieved not only with the aforementioned Nesterov's type algorithm; they are nothing but the standard upper complexity bounds of the classical Conditional Gradient algorithm originating from [11], see also [39, 10, 19, 14] and references therein. This algorithm recently has attracted a lot of attention, primarily in the Machine Learning community, due to its ability to work with "difficult geometry domains" (like nuclear norm or total variation norm balls) where the *proximal* first order algorithms (which form the vast majority of first order methods) become too computationally expensive[2]. According to our complexity results, *Conditional Gradient is nearly-optimal when minimizing smooth convex functions over high-dimensional boxes* (provided that the smoothness is quantified w.r.t. $\|\cdot\|_\infty$-norm). To the best of our knowledge, this is the first result on near-optimality of Conditional Gradient algorithm in terms of Information-Based Complexity Theory.

C. In contrast to what happens in the range $p \geq q$, in the range $1 \leq p < q \leq \infty$ there is a substantial gap $\mathcal{R}(T) \approx T^{\kappa(1/p-1/q)}$ between the upper and the lower complexity bounds. Our *guess* is that the "guilty party" here is the *upper* bound,

[2] a proximal algorithm requires at every step solving an auxiliary problem of minimizing over problem's domain the sum of a linear function and a "simple" *nonlinear* function (e.g., squared Euclidean norm); in contrast, the auxiliary problems arising in Conditional Gradient algorithm require minimizing over problem's domain just a linear function. The latter problems never are more difficult (and in the "difficult geometry case" are much simpler) than the former ones.

the motivation being as follows. The upper complexity bound (46) in the range $p < q$ is just independent of p; were this bound tight, it would mean that with fixed degree of smoothness (quantified w.r.t. the norm $\|\cdot\|_q$), minimizing smooth objectives over the unit $\|\cdot\|_q$-ball is basically as difficult as minimizing these objectives over the unit $\|\cdot\|_p$-ball with $p < q$, in spite of the fact that the second ball is "incomparably smaller" than the first one when n is large. In any case, we believe that the outlined "complexity gap" deserves in-depth investigation and pose the following

Open Problem 2.5.1. Given n, $p < q$, $\kappa \in (1, 2]$, and $L > 0$, what is the worst-case oracle complexity of minimizing objectives from the family $\mathcal{F}_q(\kappa, L)$ over the unit ball of ℓ_p^n in the large-scale regime? Is it true that under the circumstances, the upper complexity bound (46) can be significantly improved?

A positive answer to the last question would be not just an academic achievement. Consider, e.g., problems of the form

$$\begin{array}{c}\mathrm{Opt}(A) := \min_{x\in\mathbb{R}^n:\|x\|_1\le 1} \frac{1}{2}\|Ax-b\|_2^2 \\ [A \in \mathbb{R}^{m\times n}]\end{array} \tag{47}$$

which play central role in state-of-the-art sparsity-oriented signal processing. For every $q \ge 1$, the objectives here clearly belong to the family

$$\mathcal{F}_q(2, \|A\|_{1,q_*}^2),$$

where $\|A\|_{1,q_*} = \max_{x\in\mathbf{R}\|x\|_1\le 1} \|Ax\|_{q_*}$, $q_* = \frac{q}{q-1}$. Restricting ourselves to the case of $q \le 2$ and invoking (46), we see that for every $T \le n$, an appropriate T-step first order method ensures that

$$\|Ax^T - b\|_2^2 - \mathrm{Opt}(A) \preceq \frac{\|A\|_{1,q_*}^2}{T^2}, \tag{48}$$

where $x^T = x^T(A, b)$ is the approximate solution generated by the method as applied to the instance with the data (A, b); here the logarithmic in n factor hidden

in $\preceq$ is independent of T and does not exceed $O(1)\ln n$). In the large-scale regime, this is the best known so far efficiency estimate for solving problems (47) by first order algorithms. According to this estimate, our abilities to solve problems (47) are the same whether we restrict the matrices A to have the spectral norm at most 1 (i.e., restrict the objectives in (47) to belong to $\mathcal{F}_2(2,1)$), or to have magnitudes of all entries not exceeding 1 (i.e., restrict the objectives in (47) to reside in $\mathcal{F}_1(2,1)$). Indeed, in both these cases, the only *guaranteed* upper bound on $\|A\|_{1,q_*}$ is 1, that is, the best known efficiency estimate in both cases is

$$\|Ax^T - b\|_2^2 - \mathrm{Opt}(A) \preceq \frac{1}{T^2}. \tag{49}$$

Were our guess expressed in the above Open Problem indeed true, it would mean that with spectral norm of A in (47), the efficiency estimate (49) is essentially non-optimal. For example, were the "true complexity" in the ℓ_1/ℓ_2 setting be similar to our lower bound (which, for $p = 1$, $q = 2$, $\kappa = 2$ and $L = R = 1$, is $\approx T^{-3}$), the T-step risk achievable when solving problems (47) with the spectral norm of A not exceeding 1 would be $\preceq T^{-3}$, which is much better than (49).

Whichever could be the answer to the above Open Problem, the problem seems to be extremely challenging, especially when the "true answer" is the one we guess, since in the existing literature, to the best of our knowledge, there is no hint on how Fast Gradient methods (and these are the methods underlying the upper bounds (46)) could be accelerated.

D. Finally, taking into account that ℓ_r^n, as a normed space, can be viewed as the subspace of the Schatten space Sch_r^n comprised of the diagonal matrices, and the restriction of an $n \times n$ matrix on its diagonal, considered as a mapping from Sch_r^n onto ℓ_r^n, is of norm 1, it is immediately seen that our lower complexity bounds for ℓ_p/ℓ_r setting remain true when passing to the matrix analogy $\mathrm{Sch}_p/\mathrm{Sch}_q$ of this setting – the one with Sch_p^n and Sch_q^n in the roles of ℓ_p^n and ℓ_q^n, respectively. As stated

in Corollary 1.2.3, *in the range of smoothness parameter* $p \leq 2$, the upper risk bounds from the Corollary hold true for the Schatten spaces; taking into account that the Schatten norms obey the same inequalities

$$1 \leq r \leq s \leq \infty \Rightarrow \|x\|_{\mathrm{Sch},s} \leq \|x\|_{\mathrm{Sch},r} \leq n^{1/r-1/s}\|x\|_{\mathrm{Sch},s} \ \forall x \in \mathbb{R}^{n\times n}$$

as the their vector counterparts $\|\cdot\|_r$, $\|\cdot\|_s$, it follows that *in the range* $q \leq 2$, the upper risk bounds (46) hold true in the $\mathrm{Sch}_p/\mathrm{Sch}_q$ setting, we conclude that *our lower complexity bounds remain intact when passing from* ℓ_p/ℓ_q *setting to its matrix analogy* Sch_p/Sch_q*, and when* $q \leq 2$*, the same can be said about the above "tightness analysis" of the bounds.*

CHAPTER III

DISTRIBUTIONAL ORACLE COMPLEXITY OF CONVEX OPTIMIZATION

3.1 Introduction

Lower complexity bounds studied in the previous chapter, together with most known results in the literature, are based on the technique of *resisting oracles*, namely, for a sequence of adaptive queries an adversary continuously selects a function to provide the less informative consistent answers. In this chapter we take an alternative approach, based on *average-case analysis* of algorithms via distributional complexity.

Our interest in this model is twofold: On the one hand, the lower bounds obtained for distributional complexity are considerably stronger than worst-case lower bounds. Conceptually, this model of computation allows algorithms to exploit the distribution of instances for accelerating average running time. However, our lower bounds show that this extra power does not give improvement over worst-case behavior. On the other hand, the distributional approach allows us to bring insights from information theory for a systematic study of oracle complexity, unifying previous approaches.

3.1.1 The approach

Our general approach to study distributional lower bounds is based on the *reconstruction principle*: First, an algorithm that solves problems over a distribution of instances must be capable of extracting the information of the instance solely via the oracle. Second, by the so-called chain rule we can split the information gain of

the algorithm among iterations. We establish in Lemma 3.2.1 that if the information gain is bounded by a constant $C > 0$ throughout iterations, then the expected number of iterations is lower bounded by the entropy of the instance divided by C.

We apply the above methodology to study a *String-Guessing Problem* (SGP). This is a problem which is combinatorial in nature, and can be easily analyzed with our result, which is done in Proposition 3.3.2. Next, in Section 3.4 we provide a method to compare the complexity of different oracles for a given family of instances. This *oracle-emulation* procedure will allow us to derive distributional complexity lower bounds for local oracles for convex optimization.

From oracle emulation we provide a unified approach for the complexity of nonsmooth convex optimization over domains given by ℓ_p-balls, for both low-scale and large-scale regimes. Our techniques are presented in Sections 3.5, 3.6, and 3.7. The first two show explicit emulations for a particular type of oracle – so called single-coordinate – for the low-scale and large-scale regimes, respectively. Finally, by a random perturbation of instances we can use the lower bounds for single-coordinate oracles to derive a lower bound for *arbitrary local oracles*.

In the case of the box as well as the ℓ_p-ball for $1 \leq p \leq \infty$, our bounds show that all four complexity measures coincide, namely, high-probability, distributional, randomized, and worst-case complexity. This not only simplifies the proofs in [32] for randomized complexity, but also improves a gap between worst-case and randomized complexity, first studied by Nemirovski & Yudin ([32, 4.4.3 Proposition2]).

3.1.2 Notation and preliminaries

For a given oracle-based (not necessarily minimization) algorithm, we record the communication between the algorithm and the oracle. Let Q_t be the t-th query

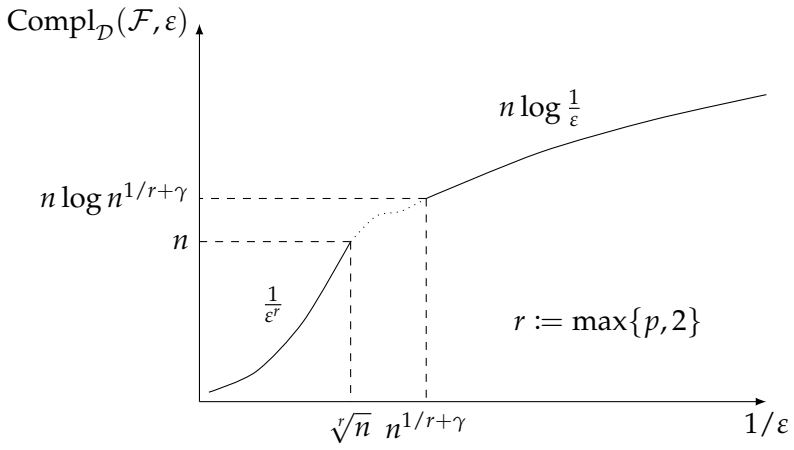

Figure 1: Distributional complexity as a function of $1/\varepsilon$ for the ℓ_p-ball, $1 \leq p < \infty$.

of the algorithm and A_t be the t-th oracle answer. Thus $\Pi_t := (Q_t, A_t)$ is the t-th query-answer pair. The full transcript of the communication is denoted by $\Pi = (\Pi_1, \Pi_2, \dots)$, and for given $t \geq 0$ partial transcripts are defined as $\Pi_{\leq t} := (\Pi_1, \dots, \Pi_t)$ and $\Pi_{<t} := (\Pi_1, \dots, \Pi_{t-1})$. By convention, $\Pi_{<1}$ and $\Pi_{\leq 0}$ are empty sequences.

As we will index functions by strings, let us introduce the necessary string operations. Let $s \in \{0,1\}^*$ be a binary string, then $s^{\oplus(i)}$ denotes the string obtained from s by flipping the i-th bit and deleting all bits following the i-th one. Let $s \sqsubseteq t$ denote that s is a prefix of t and $s \parallel t$ denote that neither is a prefix of the other. As a shorthand let $s|_l$ be the prefix of s consisting of the first l bits. We shall write $s0$ and $s1$ for the strings obtained by appending a 0 and 1 to s, respectively. Furthermore, the empty string is denoted by $\perp$.

An important feature of families of functions that we will study in this chapter is the *packing property*. For a function $f : X \to \mathbb{R}$, we define the set of ε-minima as

$$\mathcal{S}_\varepsilon(f) := \{x \in X : \ f(x) < f^* - \varepsilon\}.$$

Definition 3.1.1 (Packing property)**.** A function family $\mathcal{F}$ satisfies the packing property for an accuracy level $\varepsilon > 0$, if no two different members $f, g \in \mathcal{F}$ have common ε-minima, i.e., $\mathcal{S}_\varepsilon(f) \cap \mathcal{S}_\varepsilon(g) = \emptyset$.

Note that for a family $\mathcal{F}$ satisfying the packing property with accuracy ε, minimization of an unknown instance is equivalent to instance identification. This fact will allow us to use the reconstruction principle.

3.2 Information-Theoretic Lower Bounds from the Reconstruction Principle

We consider an unknown instance F that is randomly chosen from a finite family $\mathcal{F}$ of instances. For a given algorithm querying an oracle $\mathcal{O}$, let T be the number of queries the algorithm asks to determine the instance. Of course, the number T may depend on the instance, as algorithms can adapt their queries according to the oracle answers (see the Introduction, Section 1.2.3, for the specific setup). However, we assume that $T < \infty$ almost surely, i.e., we require algorithms to almost always terminate (this is a mild assumption as $\mathcal{F}$ is finite).

Algorithms are allowed to have an error probability bounded by P_e, i.e., the algorithm is only required to return the correct answer with probability $1 - P_e$ *across all instances*. The latter statement is important as both, being perfectly correct on a $1 - P_e$ fraction of the input and outputting garbage in P_e cases, as well as providing the correct answer for *each instance* with probability $1 - P_e$, are admissible here.

For bounded-error algorithms, the high-probability complexity is the required number of queries to produce a correct answer with probability $1 - P_e - \beta$. This adjustment is justified, as a wrong answer is allowed with probability P_e.

Lemma 3.2.1. *Let F be a random variable with finite range $\mathcal{F}$. For a given algorithm determining F via querying an oracle, with error probability bounded by P_e, suppose that*

the information gain from the each oracle answer is bounded, i.e., for some constant C

$$\mathbb{I}\left[F; A_t \mid \Pi_{<t}, Q_t, T \geq t\right] \leq C, \qquad t \geq 0. \tag{50}$$

Then, the distributional oracle complexity of the algorithm is lower bounded by

$$\mathbb{E}\left[T\right] \geq \frac{\mathbb{H}\left[F\right] - \mathbb{H}\left[P_e\right] - P_e \log|\mathcal{F}|}{C}.$$

Moreover, for all t we have

$$\mathbb{P}\left[T < t\right] \leq \frac{\mathbb{H}\left[P_e\right] + P_e \log|\mathcal{F}| + Ct}{\mathbb{H}\left[F\right]}.$$

In particular, if F is uniformly distributed, then for $t = \frac{\beta \log|\mathcal{F}| - \log 2}{C}$, $\mathbb{P}\left[T \geq t\right] = 1 - P_e - \beta$.

Proof. By induction on t we will first prove the following claim

$$\mathbb{I}\left[F; \Pi\right] = \sum_{i=1}^{t} \mathbb{I}\left[F; \Pi_i \mid \Pi_{<i}, T \geq i\right] \mathbb{P}\left[T \geq i\right] + \mathbb{I}\left[F; \Pi \mid \Pi_{\leq t}, T \geq t\right] \mathbb{P}\left[T \geq t\right]. \tag{51}$$

The case $t = 0$ is obvious. For $t > 0$, note that the event $T = t$ is independent of F given $\Pi_{\leq t}$, as at step t the algorithm has to decide whether to continue based solely on the previous oracle answers and private random sources. If the algorithm stops, then $\Pi = \Pi_{\leq t}$. Therefore,

$$\begin{aligned}
&\mathbb{I}\left[F; \Pi \mid \Pi_{\leq t}, T \geq t\right] \\
&= \mathbb{I}\left[F; \Pi, I(T = t) \mid \Pi_{\leq t}, T \geq t\right] \\
&= \underbrace{\mathbb{I}\left[F; I(T = t) \mid \Pi_{\leq t}, T \geq t\right]}_{=0} + \mathbb{I}\left[F; \Pi \mid \Pi_{\leq t}, I(T = t), T \geq t\right] \\
&= \underbrace{\mathbb{I}\left[F; \Pi \mid \Pi_{\leq t}, T = t\right]}_{=0, \text{ as } \Pi_{\leq t} = \Pi} \mathbb{P}\left[T = t \mid T \geq t\right] + \mathbb{I}\left[F; \Pi \mid \Pi_{\leq t}, T \geq t+1\right] \mathbb{P}\left[T \geq t+1 \mid T \geq t\right] \\
&= \left(\mathbb{I}\left[F; \Pi_{t+1} \mid \Pi_{<t+1}, T \geq t+1\right] + \mathbb{I}\left[F; \Pi \mid \Pi_{\leq t+1}, T \geq t+1\right]\right) \mathbb{P}\left[T \geq t+1 \mid T \geq t\right],
\end{aligned}$$

obtaining the identity

$$\begin{aligned}
&\mathbb{I}\left[F; \Pi \mid \Pi_{\leq t}, T \geq t\right] \mathbb{P}\left[T \geq t\right] \\
&= \left(\mathbb{I}\left[F; \Pi_{t+1} \mid \Pi_{<t+1}, T \geq t+1\right] + \mathbb{I}\left[F; \Pi \mid \Pi_{\leq t+1}, T \geq t+1\right]\right) \mathbb{P}\left[T \geq t+1\right],
\end{aligned}$$

from which the induction follows.

Now, in (51) by letting t go to infinity, $\mathbb{P}[T \geq t]$ will converge to 0, while $\mathbb{I}[F; \Pi \mid \Pi_{\leq t}, T \geq t]$ is bounded by $\mathbb{H}[F]$, proving that

$$\mathbb{I}[F; \Pi] = \sum_{i=1}^{\infty} \mathbb{I}[F; \Pi_i \mid \Pi_{<i}, T \geq i] \mathbb{P}[T \geq i]. \tag{52}$$

Note that Q_i is chosen solely based on $\Pi_{<i}$, and is conditionally independent of F. Therefore, by the chain rule, $\mathbb{I}[F; \Pi_i \mid \Pi_{<i}, T \geq i] = \mathbb{I}[F; A_i \mid \Pi_{<i}, Q_i, T \geq i]$. Plugging this equation into (52), we obtain

$$\mathbb{I}[F; \Pi] = \sum_{i=1}^{\infty} \mathbb{I}[F; A_i \mid \Pi_{<i}, Q_i, T \geq i] \mathbb{P}[T \geq i] \leq C \sum_{i=0}^{\infty} \mathbb{P}[T \geq i] = C \cdot \mathbb{E}[T].$$

Finally, as the algorithm determines F with error probability at most P_e, Fano's inequality [6, Theorem 2.10.1] applies

$$\mathbb{H}[F \mid \Pi] \leq \mathbb{H}[P_e] + P_e \log |\mathcal{F}|. \tag{53}$$

We therefore obtain

$$\mathbb{H}[F] = \mathbb{H}[F \mid \Pi] + \mathbb{I}[F; \Pi] \leq \mathbb{H}[P_e] + P_e \log |\mathcal{F}| + C \cdot \mathbb{E}[T],$$

and therefore

$$\mathbb{E}[T] \geq \frac{\mathbb{H}[F] - \mathbb{H}[P_e] - P_e \log |\mathcal{F}|}{C},$$

as claimed.

We will now establish concentration for the number of required queries. For this we reuse (51), the split-up of information up to query t:

$$\begin{aligned}
\mathbb{I}[F; \Pi] &= \sum_{i=1}^{t} \mathbb{I}[F; \Pi_i \mid \Pi_{<i}, T \geq i] \mathbb{P}[T \geq i] + \mathbb{I}[F; \Pi \mid \Pi_{\leq t}, T \geq t] \mathbb{P}[T \geq t] \\
&= \sum_{i=1}^{t} \underbrace{\mathbb{I}[F; A_i \mid \Pi_{<i}, Q_i, T \geq i]}_{\leq C} \mathbb{P}[T \geq i] + \underbrace{\mathbb{I}[F; \Pi \mid \Pi_{\leq t}, T \geq t]}_{\leq \mathbb{H}[F]} \mathbb{P}[T \geq t] \\
&\leq Ct + \mathbb{H}[F] \mathbb{P}[T \geq t],
\end{aligned}$$

which we combine with (53):

$$\mathbb{H}[F] = \mathbb{H}[F \mid \Pi] + \mathbb{I}[F;\Pi] \leq \mathbb{H}[P_e] + P_e \log|\mathcal{F}| + Ct + \mathbb{H}[F]\,\mathbb{P}[T \geq t],$$

and therefore

$$\mathbb{P}[T < t] \leq \frac{\mathbb{H}[P_e] + P_e \log|\mathcal{F}| + Ct}{\mathbb{H}[F]}.$$

Specializing to uniform distributions provides the last claim of the Lemma. □

3.3 String-Guessing Problem (SGP)

For a fixed length M we consider the problem of identifying a hidden string $S \in \{0,1\}^M$ picked uniformly at random. The oracle $\mathcal{O}_S$ accepts queries for any part of the string. Formally, a query is a pair (s, σ), where s is a string of length at most M, and $\sigma\colon [|s|] \to [M]$ is an embedding indicating an order of preference. The intent is to ask whether $S_{\sigma(k)} = s_k$ for all k. The oracle will reveal the smallest k so that $S_{\sigma(k)} \neq s_k$ if such a k exists or will assert correctness of the guessed part of the string. More formally we have:

Oracle 3.3.1 (String Guessing Oracle $\mathcal{O}_S$)**.**
Query: A string $s \in \{0,1\}^{\leq M}$ and an injective function $\sigma\colon [|s|] \to [M]$.
Answer: Smallest $k \in \mathbb{N}$ so that $S_{\sigma(k)} \neq s_k$ if it exists, otherwise `EQUAL`.

From Lemma 3.2.1 we can establish an expectation and high probability lower bound on the number of queries, even for bounded error algorithms. The key is that the oracle does not reveal any information about the bits after the first wrongly guessed bit, not even involuntarily.

Proposition 3.3.2 (String Guessing Problem)**.** *Let M be a positive integer, and S be a uniformly random binary string of length M. Let $\mathcal{O}_S$ be the String Guessing Oracle (Oracle 3.3.1). Then for any bounded error algorithm having access to S only through $\mathcal{O}_S$, the expected number of queries required to identify S with error probability at most*

P_e is at least $[(1-P_e)M-1]/2$. Moreover, given $\beta > 0$ if we let $t = \frac{\beta M - \log 2}{2}$, then $\mathbb{P}[T \geq t] = 1 - P_e - \beta$, where T is the number of queries.

Proof. We will prove the following claim by induction: At any step t, given the partial transcript $\Pi_{<t}$, some bits of S are totally determined, and the remaining ones are still uniformly distributed. The claim is obvious for $t = 0$. Now suppose that the claim holds for some $t - 1 \geq 0$. The next query $Q_t := (s; \sigma)$ is independent of S given $\Pi_{<t}$. Let us fix $\Pi_{<t}$ and $(s; \sigma)$, and implicitly condition on them until stated otherwise. We differentiate two cases.

CASE 1: *The oracle answer is* `EQUAL`. This is the case if and only if $s_\ell = S_{\sigma(\ell)}$ for all $\ell \in [|s|]$. Thus the oracle answer reveals the bits $\{S_{\sigma(\ell)} \mid \ell \in [|s|]\}$, actually determining them.

CASE 2: *The oracle answer is k.* This is the case if and only if $s_j = S_{\sigma(j)}$ for all $j < k$ and $s_k \neq S_{\sigma(k)}$. Thus the oracle answer reveals $\{S_{\sigma(\ell)} \mid \ell \in [k]\}$ (the k-th bit by flipping), determining them.

In both cases, the answer is independent of the other bits, therefore the ones among them, which are not determined by previous oracle answers, remain uniformly distributed and mutually independent. This establishes the claim for Π_t, finishing the induction.

We extend the analysis to estimate the mutual information of S and the oracle answer A_t. We keep $\Pi_{<t}$ and Q_t fixed, and implicitly assume $T \geq t$, as otherwise Q_t and A_t don't exist. For readability, we drop the conditions in the computations below; all quantities are to be considered conditioned on $\Pi_{<t}$, Q_t provided $T \geq t$.

Let $m := \mathbb{H}[S]$ be the number of undetermined bits just before query t. Let K be the number of additionally determined bits due to query Q_t and oracle answer A_t, hence obviously

$$\mathbb{H}[S \mid A_t] = \mathbb{E}[m - K].$$

The analysis above shows that for all $k \geq 1$, a necessary condition for $K \geq k$ is that $s_j = S_{\sigma(j)}$ for the $k-1$ smallest j with $S_{\sigma(j)}$ not determined before query t and that these $k-1$ smallest j really exist. The probability of this condition is $1/2^{k-1}$ (or 0 if there are not sufficiently many j) and so in any case we have

$$\mathbb{P}[K \geq k] \leq \frac{1}{2^{k-1}}, \qquad k \geq 1.$$

Combining these statements we see that,

$$\begin{aligned} \mathbb{I}[S; A_t] &= \mathbb{H}[S] - \mathbb{H}[S \mid A_t] = m - \mathbb{E}[m - K] \\ &= \mathbb{E}[K] = \sum_{i \in [m]} \mathbb{P}[K \geq i] \leq \sum_{i \in [\infty]} \frac{1}{2^{i-1}} = 2, \end{aligned}$$

with $\Pi_{<t}, Q_t$ still fixed.

Now we re-add the conditionals, vary $\Pi_{<t}, Q_t$, and take expectation still assuming $T \geq t$, obtaining

$$\mathbb{I}[S; A_t \mid \Pi_{<t}, Q_t, T \geq t] \leq 2$$

where T is the number of queries. By Lemma 3.2.1 applies we obtain $\mathbb{E}[T] \geq [(1-P_e)M - \mathbb{H}[P_e]]/2 \geq [(1-P_e)M - 1]/2$ (the binary entropy is upper bounded by 1) and $\mathbb{P}[T \geq t] = 1 - P_e - \beta$, as claimed. □

3.4 Oracle Emulation

In this section we introduce *oracle emulation*, which is a special type of reduction from one oracle to another, both for the same family of instances. This reduction allows to transform algorithms based on one oracle to the other preserving their oracle complexity, i.e, the number of queries asked. The crucial result is Lemma 3.4.2, which we will apply to emulations of various convex optimization oracles by the String Guessing Oracle $\mathcal{O}_S$.

Definition 3.4.1 (Oracle emulation)**.** Let $\mathcal{O}_1 \colon Q_1 \to R_1$ and $\mathcal{O}_2 \colon Q_2 \to R_2$ be two oracles for the same problem. An *emulation* of $\mathcal{O}_1$ by $\mathcal{O}_2$ consists of

(i) a query emulation function $q\colon Q_1 \to Q_2$ (translating queries of $\mathcal{O}_1$ for $\mathcal{O}_2$),

(ii) an answer emulation function $a\colon Q_1 \times R_2 \to R_1$ (translating answers back)

such that $\mathcal{O}_1(x) = a(x, \mathcal{O}_2(q(x)))$ for all $x \in Q_1$.

An emulation leads to a reduction, since emulating oracles are at least as complex as the oracles they emulate.

Lemma 3.4.2. *If there is an emulation of $\mathcal{O}_1$ by $\mathcal{O}_2$, then the oracle complexity of $\mathcal{O}_1$ is at least that of $\mathcal{O}_2$. Here oracle complexity can be worst-case, randomized, distributional, and high probability; all even for bounded-error algorithms.*

Proof. Let A_1 be an algorithm using $\mathcal{O}_1$, and let $\mathcal{O}_2$ emulate $\mathcal{O}_1$. Let q and a be the query emulation function and the answer emulation function, respectively. We define an algorithm A_2 for $\mathcal{O}_2$ simulating A_1 as follows: Whenever A_1 asks a query x to oracle $\mathcal{O}_1$, oracle $\mathcal{O}_2$ is queried with $q(x)$, and the simulated A_1 receives as answer $a(x, \mathcal{O}_2(q(x)))$ (which is $\mathcal{O}_1(x)$ by definition of the emulation). Finally, the return value of the simulated A_1 is returned.

Obviously, A_2 makes the same number of queries as A_1 for every input, and therefore the two algorithms have the same oracle complexity. This proves that the oracle complexity of $\mathcal{O}_1$ is at least that of $\mathcal{O}_2$. □

3.5 Single-Coordinate Oracle Complexity for the Box

In the following we will analyze a simple class of oracles, called 'single-coordinate', closely mimicking the String Guessing Oracle. Later, all results will be carried over to general local oracles via perturbation in Section 3.7.

Definition 3.5.1 (Single-coordinate oracle)**.** A first-order oracle $\tilde{\mathcal{O}}$ is *single-coordinate* if for all $x \in X$ the subgradient $\nabla f(x)$ in its answer is the one supported on the least coordinate axis; i.e., $\nabla f(x) = \lambda e_i$ for the smallest $1 \le i \le n$ with some $\lambda \in \mathbb{R}$.

Choosing the smallest possible i corresponds to choosing the first wrong bit by the String Guessing Oracle. Not all function families possess a single-coordinate oracle, but maximum of coordinate functions do, and single-coordinate oracles are a natural choice for them. From now on, we will denote single-coordinate oracles exclusively by $\tilde{\mathcal{O}}$.

We establish a lower bound on the *distributional* and *high probability oracle complexity* for nonsmooth convex optimization over $[-R, +R]^n$, for single-coordinate oracles.

Theorem 3.5.2. *Let $L, R > 0$. There exists a finite family $\mathcal{F}$ of Lipschitz continuous convex functions on the ℓ_∞^n-ball $B_\infty(0, R)$ with Lipschitz constant L in the ℓ_∞^n norm, and a single-coordinate local oracle $\tilde{\mathcal{O}}$, such that both the distributional and the high-probability oracle complexity for finding an ε-minimum of a uniformly random instance is $\Omega\left(n \log \frac{LR}{\varepsilon}\right)$.*

For bounded-error algorithms with error bound P_e, the distributional complexity is $\Omega\left((1 - P_e) n \log \frac{LR}{\varepsilon}\right)$, and the high-probability complexity of level β is $\Omega\left(\beta n \log \frac{LR}{\varepsilon}\right)$.

In the following we will restrict ourselves to the case $L = R = 1$, as the theorem reduces to it via an easy scaling argument. We start with the one dimensional case in Section 3.5.1 for a simpler presentation of the main ideas. We generalize to multiple dimensions in Section 3.5.2 by considering maxima of coordinate functions, thereby using the different coordinates to represent different portions of a string.

3.5.1 One dimensional case

Let $X := [-1, 1]$, we define recursively a function family $\mathcal{F}$ on X, which is inspired by the one in [30, Lemma 1.1.1]. For an interval $I = [a, b]$, let $I(t) := a + (1 + t)(b - a)/2$ denote the t-point on I for $-1 \leq t \leq 1$, e.g., $I(-1)$ is the left end point of I, and $I(+1)$ is the right end point, and $I(0)$ is the midpoint. Let $I[t_1, t_2]$ denote the subinterval $[I(t_1), I(t_2)]$. The family $\mathcal{F} = \{f_s\}_s$ will be indexed by binary strings

s of length M, where $M \in \mathbb{N}$ depends on the accuracy and will be chosen later. It is convenient to define f_s also for shorter strings, as we proceed by recursion on the length of s. We also define intervals I_s and breakpoints b_l of the range of the functions satisfying the following properties:

(F-1) The interval I_s has length $2 \cdot (1/4)^{|s|}$.

Motivation: allow a strictly nesting family.

(F-2) If $s \parallel t$, then $\operatorname{int}(I_s) \cap \operatorname{int}(I_t) = \emptyset$. If $t \sqsubseteq s$, we have $I_s \subseteq I_t$ (I_s are nested intervals).

Motivation: instances can be distinguished by its associated intervals. Captures packing property.

(F-3) $f_s \geq f_{s|_l}$ with $f_s(x) = f_{s|_l}(x)$ if $x \in [-1,1] \setminus \operatorname{int}\left(I_{s|_l}\right)$.

Motivation: long prefix determines much of the function.

(F-4) The function f_s restricted to the interval I_s is of the form

$$f_s(x) = b_{|s|} - 2^{-3|s|} + 2^{-|s|}|x - I_s(0)| \qquad x \in I_s,$$

where $b_{|s|} = f_s(I_s(-1)) = f_s(I_s(+1))$ is the function value on the endpoints of I_s. This is symmetric on I_s as $I_s(0)$ is the midpoint of I_s.

Motivation: recursive structure: repeat absolute value function on small intervals.

(F-5) For $t \sqsubseteq s$, we have $f_s(x) < b_{|t|}$ if and only if $x \in \operatorname{int}(I_t)$.

Motivation: level sets encode substrings.

Construction of the function family

We start with the empty string $\perp$, and define $f_\perp(x) := |x|$ and $I_\perp := [-1,1]$. In particular, $b_0 = 1$. The further b_k we define via the recursion $b_{k+1} := b_k - 2 \cdot (1/4)^{k+1} \cdot 2^{-k}$.

Given f_s and I_s, we define f_{s0} and I_{s0} to be the *right modification* of f_s via

$$I_{s0} := I_s\left[-\frac{1}{2}, 0\right]$$

$$f_{s0}(x) := \begin{cases} b_{|s|+1} - 2^{-3(|s|+1)} + 2^{-|s|-1}\left|x - I_s\left(-\frac{1}{4}\right)\right|, & \text{if } x \in I_s\left[-\frac{1}{2}, 1\right] \\ f_s(x), & \text{otherwise.} \end{cases}$$

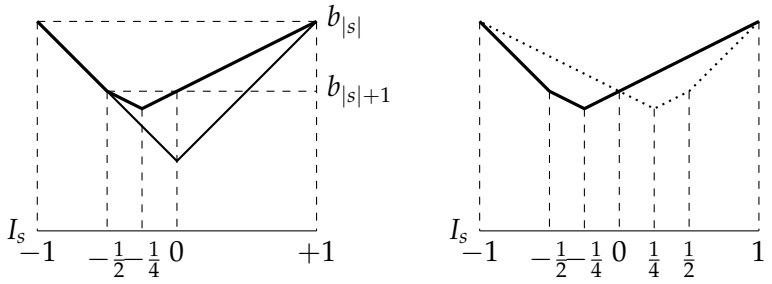

Figure 2: Right modification on the left side: the solid normal line is before the modification, the solid thick line after it. On the right side: right modification is the solid thick line; left modification is the dotted line.

Similarly, the *left modification* f_{s1} of f_s is the reflection of f_{s0} with respect to $I_s(0)$, and I_{s1} is the reflection of I_{s0} with respect to $I_s(0)$, i.e.,

$$I_{s1} := I_s\left[0, \frac{1}{2}\right]$$

$$f_{s1}(x) := \begin{cases} b_{|s|+1} - 2^{-(3|s|+1)} + 2^{-|s|-1}\left|x - I_s\left(\frac{1}{4}\right)\right|, & \text{if } x \in I_s\left[-1, \frac{1}{2}\right] \\ f_s(x), & \text{otherwise.} \end{cases}$$

Observe that $I_{s0}, I_{s1} \subseteq I_s$ and $\operatorname{int}(I_{s0}) \cap \operatorname{int}(I_{s1}) = \emptyset$.

This finishes the definition of the f_s. Clearly, these functions are convex and Lipschitz continuous with Lipschitz constant 1, satisfying (F-1)–(F-5).

We establish the packing property for $\mathcal{F}$.

Lemma 3.5.3. *The family $\mathcal{F}$ satisfies the packing property for $M = \lfloor (1/3)\log(1/\varepsilon) \rfloor$.*

Proof. Note that f_S has its minimum at the midpoint of I_S, and the function value at the endpoints of I_S are at least $(1/2)^{3M} \geq \varepsilon$ larger than the value at the midpoint.

Therefore every ε-optimal solution lies in the interior of I_S, i.e., $\mathcal{S}_\varepsilon(f_S) \subseteq \text{int}\,(I_S)$. Therefore by (F-2), the $\mathcal{S}_\varepsilon(f_S)$ are pairwise disjoint. □

In the following $F \in \mathcal{F}$ will be an instance picked uniformly at random. The random variable S will be the associated string of length M so that $F = f_S$ and S is also distributed uniformly.

Reduction to the String Guessing Problem

We will now provide an oracle for family $\mathcal{F}$ that can be emulated by the String Guessing Oracle. As a first step, we relate the query point x with the indexing strings of the functions. At a high level, the lemma below shows the existence of a prefix of the unknown string determining most of the local behavior of the function at a given query point. From this we will prove in Lemma 3.5.5 that the oracle answer only reveals this prefix.

Lemma 3.5.4. *Let $x \in [-1, +1]$ be a query point. Then there is a non-empty binary string s with $l := |s| \leq M$ with the following properties.*

(i) $f_{s^{\oplus(1)}}(x) \geq b_1 > f_{s^{\oplus(2)}}(x) \geq \cdots \geq b_{l-1} > f_{s^{\oplus(l)}}(x) \geq f_s(x)$. *If $l < M$ then also $f_s(x) \geq b_l$.*

(ii) Every binary string t of length M has a unique prefix p from $\{s^{\oplus(1)}, \ldots, s^{\oplus(l)}, s\}$. Moreover, $f_t(x) = f_p(x)$.

Proof. Let s_0 be the longest binary string of length less than M, such that x lies in the interior of I_{s_0}. We choose s to be the one of the two extensions of s_0 by 1 bit, for which f_s has the smaller function value at x (if the two values are equal, then either extension will do). Let $l := |s|$, thus $f_{s^{\oplus(l)}}(x) \geq f_s(x)$.

Note that by the choice of s_0, the point x is not an interior point of I_s unless $l = M$. By (F-2), the point x is neither an interior point of any of the $I_{s^{\oplus(1)}}, \ldots, I_{s^{\oplus(l)}}$.

To prove (ii), let t be any binary string of length M. The existence and uniqueness of a prefix p of t from the set $\{s^{\oplus(1)}, \ldots, s^{\oplus(l)}, s\}$ is clear. In particular, unless

$p = t = s$ and $l = M$, the point x is not an interior point of I_p, hence $f_t(x) = f_p(x)$ follows from (F-3). When $p = t$, then $f_t(x) = f_p(x)$ is obviously true.

Now we prove (i). Recall that $f_{s^{\oplus(l)}}(x) \geq f_s(x)$ by the choice of s. First, if $l < M$ then $x \notin \operatorname{int}(I_s)$ by choice, hence $f_s(x) \geq b_l$ by (F-5). Second, let us prove $f_{s^{\oplus(i)}}(x) \geq b_i > f_{s^{\oplus(i+1)}}(x)$ for all $i \leq l$. As $x \notin \operatorname{int}\left(I_{s^{\oplus(i)}}\right)$, by (F-5) we have $f_{s^{\oplus(i)}}(x) \geq b_i$. Finally, since $x \in \operatorname{int}\left(I_{s|_i}\right)$ and $s|_i \sqsubseteq s^{\oplus(i+1)}$, again by (F-5) we get $f_{s^{\oplus(i+1)}}(x) < b_i$. □

Our construction of instances encodes prefixes in level sets of the instance. The previous lemma indicates that algorithms in this case need to identify a random string, where the oracle reveals prefixes of such string. The following lemma formally shows an emulation by the String Guessing Oracle.

Lemma 3.5.5. *There is a single-coordinate local oracle $\tilde{\mathcal{O}}$ for the family $\mathcal{F}$ above, which is emulated by the String Guessing Oracle $\mathcal{O}_S$ on strings of length M.*

Proof. We define the emulation functions first, as they determine the emulated oracle $\tilde{\mathcal{O}}$. Let $x \in [-1, 1]$ and s the string from Lemma 3.5.4. We define the query emulation function as $q(x) := (s, \mathrm{id})$. Moreover, let $l = |s|$.

Now we need to emulate the oracle answer. From Lemma 3.5.4 (ii) there exists a prefix P of S such that $f_S(x) = f_P(x)$. We define the following function p of the $\mathcal{O}_S$ oracle answer

$$p(x, \texttt{EQUAL}) := s,$$
$$p(x, k) := s^{\oplus(k)}.$$

Note that $P = p(x, \mathcal{O}_S(q(x)))$. We claim that p depends on f_S only locally around x. First, if $f_s(x) < f_{s^{\oplus(l)}}(x)$ then by Lemma 3.5.4 (i) $f_S(x)$ determines P (and thus p). Otherwise, depending on whether f_S is increasing or decreasing around x, we can determine if $P_l = s_l$.

Since $f_S(x) = f_P(x)$ and $f_S \geq f_P$, a valid oracle answer to the query point x is $f_P(x)$ as function value and a subgradient $\nabla f_P(x)$ of f_P at x as $\nabla f_S(x)$. Therefore we define the answer emulation as $a(x, R) := (f_{p(x,R)}(x), \nabla f_{p(x,R)}(x))$. This provides a single-coordinate local oracle $\tilde{\mathcal{O}}$ for the family $\mathcal{F}$ (the single-coordinate condition is trivially satisfied when $n = 1$) that can be emulated by the String Guessing Oracle $\mathcal{O}_S$. □

The previous lemma together with Lemma 3.4.2 leads to a straightforward proof of Theorem 3.5.2 in the one dimensional case.

Proof of Theorem 3.5.2 for $n = 1$. Let A be a black box optimization algorithm for $\mathcal{F}$ accessing the oracle $\tilde{\mathcal{O}}$. As $\mathcal{F}$ satisfies the packing property by Lemma 3.5.3, in order to find an ε-minimum the algorithm A has to identify the string s defining the function $f = f_s$ (and from an ε-minimum the string s can be recovered).

Let $F = f_S$ be the random instance chosen with uniform distribution. Together with the emulation defined in Lemma 3.5.5, algorithm A solves the String Guessing Problem for strings of length M, hence requiring at least $[(1 - P_e)M - 1]/2$ queries in expectation with error probability at most P_e by Proposition 3.3.2. Moreover, with probability $1 - P_e - \beta$, the number of queries is at least $\Omega(\beta M)$. This proves the theorem for $n = 1$ by the choice of M. □

3.5.2 Multidimensional case

Construction of function family

In the general n-dimensional case the main difference is using a larger indexing string. Therefore we choose $M = \lfloor (1/3) \log(1/\varepsilon) \rfloor$, and consider n-tuples $s_1, \ldots, s_n$ of binary strings of length M as indexing set for the function family $\mathcal{F}$, and define the member functions via

$$f_{s_1,\ldots,s_n}(x_1, \ldots, x_n) := \max_{i \in [n]} f_{s_i}(x_i), \tag{54}$$

where the f_{s_i} are the functions from the one dimensional case. This way, the size of $\mathcal{F}$ is 2^{nM}. Note that as the f_{s_i} are 1-Lipschitz, the $f_{s_1,\ldots,s_n}$ are 1-Lipschitz in the ℓ_∞ norm, too. We prove that $\mathcal{F}$ satisfies the packing property.

Lemma 3.5.6. *The family $\mathcal{F}$ above satisfies the packing property for $M = \lfloor (1/3)\log(1/\varepsilon) \rfloor$.*

Proof. As the minimum values of all the one dimensional f_{s_i} coincide, obviously the set of ε-minima of $f_{s_1,\ldots,s_n}$ is the product of its components:

$$\mathcal{S}_\varepsilon(f_{s_1,\ldots,s_n}) = \prod_{i\in[n]} \mathcal{S}_\varepsilon(f_{s_i}).$$

Hence the claim reduces to the one dimensional case, proved in Lemma 3.5.3. □

Let $S = (S_1, \ldots, S_n)$ denote the tuple of strings indexing the actual instance, hence the S_i are mutually independent uniform binary strings; and let $F = f_{S_1,\ldots,S_n}$.

Reduction to the String Guessing Problem

We argue as in the one dimensional case, but now the string for the String Guessing Oracle is the concatenation of the strings $S_1, \ldots, S_n$, and therefore has length nM.

Lemma 3.5.7. *There is a single-coordinate oracle $\tilde{\mathcal{O}}$ for family $\mathcal{F}$ that can be emulated by the String Guessing Oracle $\mathcal{O}_S$ with associated string the concatenation of the $S_1, \ldots, S_n$.*

Before proving the result, let us motivate our choice for the first-order oracle. The general case arises from an interleaving of the case $n = 1$. As we have seen in the proof of Lemma 3.5.5, for $n = 1$ querying the first-order oracle leads to querying prefixes. By (F-3), if S is the string defining the function f_S, then for any prefix S' of S we have $f_{S'} \leq f_S$; this gives a lower bound on the unknown instance. By querying a point x we obtain such a prefix with the additional property $f_{S'}(x) = f_S(x)$, which localizes the minimizer in an interval, and thus provides an upper bound on its value.

Now, for general n we want to upper bound the maximum as well by prefixes of the hidden strings. In particular, there is no use to querying any potential prefixes u for coordinate i such that $f_u(x_i)$ is strictly smaller than the candidate maximum; they are not revealed by the oracle.

The query string for the String Guessing Oracle now arises by interleaving the query strings for each coordinate. In particular, if we restrict the query string to the substring consisting only of prefixes for a specific coordinate i, then these substrings should be ordered by $\sqsubseteq$, which is precisely the ordering we used for the case $n = 1$ as a necessary condition. Thus, a natural way of interleaving these query strings is by their objective function value. Moreover, refining this order by the lexicographic order on coordinates will induce a single-coordinate oracle.

Proof. Let $x = (x_1, \ldots, x_n)$ be a query point. For a family of strings $\{S_i\}_i$ regard S as their concatenation, and for notational convenience let $S_{i,h}$ denote the h-th bit of S_i. Applying Lemma 3.5.4 to each coordinate $i \in [n]$, there is a number l_i and a string s_i of length l_i associated to the point x_i.

We define the confidence order $\prec$ of labels (i, h) with $i \in [n]$ and $h \in [l_i]$ as the one induced by the lexicographic order on the pairs $(-f_{s_i^{\oplus(h)}}(x_i), i)$ i.e.,

$$(i_1, h_1) \prec (i_2, h_2) \iff \begin{cases} f_{s_{i_1}^{\oplus(h_1)}}(x_{i_1}) > f_{s_{i_2}^{\oplus(h_2)}}(x_{i_2}) & \text{or} \\ f_{s_{i_1}^{\oplus(h_1)}}(x_{i_1}) = f_{s_{i_2}^{\oplus(h_2)}}(x_{i_2}) \ \wedge \ i_1 \leq i_2. \end{cases} \tag{55}$$

We restrict to the labels (i, h) with $f_{s_i^{\oplus(h)}}(x_i) \geq \max_{j \in [n]} f_{s_j}(x_j)$ (there is no use to query the rest of labels, as pointed out above). Let $(i_1, h_1), \ldots, (i_k, h_k)$ be the sequence of these labels in $\prec$-increasing confidence order. Let t be the string of length k with $t_m = s_{i_m, h_m}$ for all $m \in [k]$. We define the query emulation as $q(x) = (t, \sigma)$ with $\sigma_m := (i_m, h_m)$.

Now, when queried with this string the String Guessing Oracle returns the index of the first mismatch. This string corresponds to a prefix of S (in the order

given by $\prec$). To be precise, we define a coordinate j and a prefix p as helper functions in x and the oracle answer for the answer emulation a (with the intent of having $f_S(x) = f_p(x_j)$ and p a prefix of S_j). If the oracle answer is `EQUAL`, then we choose $j = i_k$, and set $p := s_j|_{h_k}$. If the oracle answer is a number m then we set $j := i_m$ and $p := s_{i_m}^{\oplus(h_m)}$.

Analogously as in the proof of Lemma 3.5.5, both p and j depend only on x and on the local behavior of f_S around x. Moreover, it is easy to see that $f_S(x) = f_p(x_j)$ and $f_S(y) \geq f_{S_j}(y_j) \geq f_p(y_j)$for all y, which means that $\nabla f_p(x_j)e_j$ is a subgradient of f_S at x.

We now define the answer emulation

$$a(x, R) = (f_{p(x,R)}(x_{j(x,R)}), \nabla f_{p(x,R)}(x_{j(x,R)})e_{j(x,R)}),$$

and thus the oracle $\tilde{\mathcal{O}}(x) = a(x, \mathcal{O}_S(q(x)))$ is a first-order local oracle for the family $\mathcal{F}$ that can be emulated by the String Guessing Oracle. Finally, the single-coordinate condition is satisfied from the confidence order of the queries, which proves our result. □

We are ready to prove Theorem 3.5.2.

Proof of Theorem 3.5.2. The proof is analogous to the case $n = 1$. However, by the emulation via Lemma 3.5.7 we solve the String Guessing Problem for strings of length nM. Thus by Proposition 3.3.2 we obtain the claimed bounds the same way as in the case $n = 1$. □

3.6 Single-Coordinate Oracle Complexity for ℓ_p-Balls

In this section we examine the complexity of convex nonsmooth optimization on the unit ball $B_p(0, 1)$ in the ℓ_p^n norm for $1 \leq p < \infty$. Again, we restrict our analysis to the case of single-coordinate oracles. We distinguish the large-scale case (i.e., $\varepsilon \geq 1/n^{\max\{p,2\}}$), and low-scale case (i.e., $\varepsilon \leq n^{-1/\max\{p,2\}-\gamma}$, for fixed $\gamma > 0$).

3.6.1 Large-scale case

Theorem 3.6.1. *Let $1 \leq p < \infty$ and $\varepsilon \geq 1/\sqrt[p]{n}$. There exists a finite family $\mathcal{F}$ of convex Lipschitz continuous functions in the ℓ_p^n norm with Lipschitz constant 1 on the n-dimensional unit ball $B_p(0,1)$, and a single-coordinate local oracle $\tilde{\mathcal{O}}$ for $\mathcal{F}$, such that both the distributional and the high-probability oracle complexity of finding an ε-minimum under the uniform distribution are $\Omega\left(1/\varepsilon^{\max\{p,2\}}\right)$.*

For bounded-error algorithms with error probability at most P_e, the distributional complexity is
$\Omega\left((1-P_e)/\varepsilon^{\max\{p,2\}}\right)$*, while the high probability complexity of level β is* $\Omega\left(\beta/\varepsilon^{\max\{p,2\}}\right)$.

Remark 3.6.2 (The case $p = 1$). For $p = 1$, the lower bound can be improved to $\Omega\left(\frac{\ln n}{\varepsilon^2}\right)$ by a nice probabilistic argument, see [32, Section 4.4.5.2].

As in the previous section, we will construct a single-coordinate oracle that can be emulated by the String Guessing Oracle. As the lower bound does not depend on the dimension, we shall restrict our attention to the first $M = \Omega(1/\varepsilon^{\max\{p,2\}})$ coordinates. For these coordinates, it will be convenient to work in an orthogonal basis of vectors with maximal ratio of ℓ_p norm and ℓ_2 norm, to efficiently pack functions in the ℓ_p-ball. For $p \geq 2$ the standard basis vectors e_i already have maximal ratio, but for $p < 2$ it requires a basis of vectors with all coordinates of all vectors being ± 1, see Figure 3. In particular, in our working basis the ℓ_p norm might look different than in the standard basis. We shall present the two cases

Figure 3: Unit vectors of maximal ℓ_p norm together with the unit Euclidean ball in gray and the unit ℓ_p-ball in black.

uniformly, keeping the differences to a bare minimum.

The exact setup is as follows. Let $r := \max\{p, 2\}$ for simplicity. We define M and the working basis for the first M coordinates, such that the coordinates as functions will have Lipschitz constant at most 1.

CASE 1: $2 \leq p < \infty$. We let $M := \left\lfloor \frac{1}{\varepsilon^p} \right\rfloor - 1$. The working basis is chosen to be the standard basis.

CASE 2: $1 \leq p < 2$. Let l be the largest integer with $1/\varepsilon^2 > 2^l$, and define $M := 2^l$. Since $\varepsilon \geq 1/n^2$, obviously $M < 1/\varepsilon^2 \leq n$. In the standard basis, the space $\mathbb{R}^2$ has an orthogonal basis of ± 1 vectors, e.g., $(1, 1)$ and $(1, -1)$. Taking l-fold tensor power, we obtain an orthogonal basis of $\mathbb{R}^M$ consisting of ± 1 vectors ν_i in the standard basis. We shall work in the scaled orthogonal basis $\xi_i := \nu_i / \sqrt[p_*]{M}$. Note that the coordinate functions have Lipschitz constant at most 1, as $\langle \xi_i, x \rangle \leq \|\xi_i\|_{p_*} \|x\|_p$ for all x, and $\|\xi_i\|_{p_*} = 1$.

Clearly in both cases, $M \leq n$ and $M = \Omega(1/\varepsilon^r)$, but $M < 1/\varepsilon^r$. From now on, we shall use $\|\cdot\|_p$ for the p norm in the original basis, and $\|\cdot\|_2$ for the 2 norm in the working basis. Note that $\|x\|_p \leq \|x\|_2$ if $p < 2$.

Construction of function family

We define our functions $f_s \colon B_p(0, 1) \to \mathbb{R}$ as maximum of (linear) coordinate functions:

$$f_s(x) = \max_{i \in [M]} s_i x_i, \tag{56}$$

where the x_i are the coordinates of x in our working basis.

We parameterize the family $\mathcal{F} = \{f_s : s \in \{-1, +1\}^M\}$ via sequences $s = (s_1, \ldots, s_M)$ of signs ± 1 of length M. By the above this family satisfies the requirements of Theorem 3.6.1. We establish the packing property for $\mathcal{F}$.

Lemma 3.6.3. *The family $\mathcal{F}$ satisfies the packing property.*

Proof. Let $x = (x_1, \ldots, x_n)$ be an ε-minimum of f_s. We compare it with

$$x^* := \left(-\frac{s_1}{\sqrt[r]{M}}, \ldots, -\frac{s_M}{\sqrt[r]{M}}, 0, \ldots, 0 \right).$$

Recall that $r = \max\{p, 2\}$. The vector x^* lies in the unit L^p-ball. This is obvious for $p \geq 2$, while for $p < 2$ this follows from $\|x^*\|_p \leq \|x^*\|_2 = 1$.

Therefore, as $M < 1/\varepsilon^r$, we obtain for all $i \in [M]$

$$s_i x_i \leq f_s(x_1, \ldots, x_n) \leq f_s^* + \varepsilon \leq f_s(x^*) + \varepsilon = -\frac{1}{\sqrt[r]{M}} + \varepsilon < 0,$$

i.e., $s_i = -\text{sign} x_i$. Hence every ε-minimum x uniquely determines s, proving the packing property. □

Let $F \in \mathcal{F}$ be chosen uniformly at random, and let S be the associated string of length M so that $F = f_S$ and thus $S \in \{-1, +1\}^M$ is uniformly distributed.

Reduction to the String Guessing Problem

The main idea is that the algorithm learns solely some entries S_i of the string S from an oracle answer.

Lemma 3.6.4. *There is a single-coordinate local oracle $\tilde{\mathcal{O}}$ that can be emulated by the String Guessing Oracle $\mathcal{O}_S$.*

Proof. To better suit the present problem, we now use ± 1 for the values of bits of strings.

Given a query x, we introduce an ordering $\prec$ on the set of coordinates $\{1, 2, \ldots, M\}$: we map each coordinate i to the pair $(-|x_i|, i)$, and take the lexicographic order on these pairs, i.e.,

$$i_1 \prec i_2 \iff \begin{cases} |x_{i_1}| > |x_{i_2}| & \text{or} \\ |x_{i_1}| = |x_{i_2}| \ \wedge \ i_1 \leq i_2. \end{cases}$$

Let $\sigma(1), \ldots, \sigma(k)$ be the indices $i \in [M]$ put into $\prec$-increasing order with k the minimum between M and the $\prec$-first i s.t. $x_i = 0$. Let s be the string of length k

with $s_j = -\text{sign} x_{\sigma(j)}$. If $x_{\sigma(k)} = 0$, we put $s_k = +1$. (The value -1 would also do.) The query emulation q is defined via $q(x) := (s, \sigma)$.

We now define helper functions J and p in x and a query of $\mathcal{O}_S$. We set

$$J(x, \texttt{EQUAL}) := k, \qquad p(x, \texttt{EQUAL}) := s_k, \qquad J(x, j) := j, \qquad p(x, j) := -s_j.$$

For the remainder of the proof we drop the arguments of these functions and simply write J and p instead of $J(x, \mathcal{O}_S(q(x)))$ and $p(x, \mathcal{O}_S(q(x)))$, respectively to ease readability.

Actually, J is the $\prec$-smallest index j with $f_S(x) = S_j x_j$. If $j \neq \sigma(k)$ then $p = S_j$; in the case $J = \sigma(k)$, the value of p is $+1$ if f_S is partially locally increasing in x in the J-th coordinate, and it is -1 if it is decreasing. In other words, J and p are local. Moreover, $f_S(x) = px_J$ and $f_S(y) \geq py_J$ for all y, therefore pe_J is a subgradient of f_S at x.

We define the query emulation a via $a(x, R) := (p(x, R)x_{J(x,R)}, p(x, R)e_{J(x,R)})$. Oracle $\tilde{\mathcal{O}}$ is defined by the emulation $\tilde{\mathcal{O}}(x) = a(x, \mathcal{O}_S(q(x)))$, which is clearly single-coordinate. Thus $\tilde{\mathcal{O}}(x)$ is a valid answer to query x. □

We are ready to prove Theorem 3.6.1

Proof of Theorem 3.6.1. The proof is analogous to the proof of Theorem 3.5.2. Given the oracle $\mathcal{O}$ in Lemma 3.6.4, every black box algorithm A having access to this oracle solves the String Guessing Problem for strings of length $M = \Theta(1/\varepsilon^{\max\{p,2\}})$ using the String Guessing Oracle only. Hence the claimed lower bounds are obtained by Proposition 3.3.2. □

3.6.2 The low-scale case: reduction to the box case

We show that for small accuracies, the ℓ_p-ball lower bound follows from Theorem 3.5.2. Before we establish this result, let us observe that for technical reasons the optimal lower bound when $1 \leq p < 2$ will be postponed until Section 3.7.

Proposition 3.6.5. *Let $1 \leq p < \infty$, and $\varepsilon \leq n^{-\frac{1}{p}-\gamma}$ with $\gamma > 0$. There exists a family $\mathcal{F}$ of convex Lipschitz continuous functions in the ℓ_p norm with Lipschitz constant 1 on the n-dimensional unit Euclidean ball $B_p(0,1)$, and a single-coordinate oracle for family $\mathcal{F}$, such that both the distributional and the high-probability oracle complexity of level β of finding an ε-minimum under the uniform distribution is $\Omega\left(\beta n \log \frac{1}{\varepsilon}\right)$.*

For algorithms with error probability at most P_e, the distributional complexity is $\Omega\left((1-P_e) n \log \frac{1}{\varepsilon}\right)$ and the high probability complexity of level β is $\Omega\left(\beta n \log \frac{1}{\varepsilon}\right)$.

Proof. The proof is based on a rescaling argument.

We have $[-\frac{1}{\sqrt[p]{n}}, \frac{1}{\sqrt[p]{n}}]^n \subseteq B_p(0,1)$ and thus by Theorem 3.5.2 there exists a family of convex Lipschitz continuous functions with Lipschitz constant 1 (in the ℓ_∞ norm, therefore also in the ℓ_p norm), and a single-coordinate oracle for $\mathcal{F}$, with both distributional oracle complexity and high-probability oracle complexity $\Omega\left(n \log \frac{1}{\varepsilon \sqrt[p]{n}}\right) = \Omega\left(n \log \frac{1}{\varepsilon}\right)$ for large n, where the last equality follows from the fact that for $\varepsilon \leq n^{-1/p-\gamma}$ with $\gamma > 0$ we have $\varepsilon \sqrt[p]{n} \leq \varepsilon^{\frac{\gamma}{1/p+\gamma}}$.

□

For the case of the ℓ_p-ball with $1 \leq p < \infty$, we thus close the gap exhibited in Figure 1 for arbitrary small but fixed $\gamma > 0$.

Remark 3.6.6 (Understanding the dimensionless speed up in terms of entropy). The observed (dimensionless) performance for the ℓ_p-ball, for $2 \leq p < \infty$, has a nice interpretation when comparing the total entropy of the function families. Whereas in the unit box we could pack up to roughly $2^{n \log \frac{1}{\varepsilon}}$ instances with nonintersecting ε-solutions, we can only pack roughly $2^{1/\varepsilon^p}$ into the ℓ_p-ball. This drop in entropy alone can explain the observed speed up.

We give some intuition by comparing the volume of the unit box with the volume of the inscribed unit ℓ_p-ball. Suppose that there are $K_n \approx 2^{n \log 1/\varepsilon}$ 'equidistantly' packed instances in the box; this number is roughly the size of the function

family used above. Intersecting with the ℓ_p-ball, see Figure 4 for an illustration,

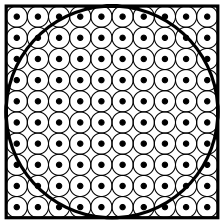

Figure 4: Equidistantly packed points with a neighbourhood in a ball and a box. The number of points in each is proportional to its volume.

we end up with roughly $K_n V_n$ instances, where $V_n = (2\Gamma(1/p+1))^n/\Gamma(n/p+1)$ is the volume of the unit ball. For the boundary case $\varepsilon = 1/\sqrt[p]{n}$:

$$\begin{aligned}\mathbb{H}[F] &\approx \log K_n V_n \\ &\approx n \log n^{1/p} + n\left(1 + \frac{1}{p}\log\frac{1}{p} - \frac{1}{p} + \log\sqrt{2\pi}\right) - \left(\frac{n}{p}\log\frac{n}{p} - \frac{n}{p} + \log\sqrt{2\pi}\right) \\ &\approx n\left(1 + \log\sqrt{2\pi}\right) \approx \frac{1+\log\sqrt{2\pi}}{\varepsilon^p},\end{aligned}$$

i.e., the entropy of the function family in the ball drops significantly, being in line with the existence of fast methdods in this case.

3.7 Lower Complexity Bounds for Arbitrary Local Oracles

We extend our results in Sections 3.5 and 3.6 to arbitrary local oracles. The key observation is that for query points where the instance is locally linear any local oracle reduces to the the single-coordinate oracle studied in previous sections. Thus, we can prove lower bounds by perturbing our instances in such a way that we avoid singular[1] query points with probability one.

We present full proofs for expectation (distributional) lower bounds, however

[1] In our framework, singular points are defined as the ones where a subgradient depends on more than one bit encoding the instance. This coincides with points of nonsmoothness in the large-scale case, but in the box case there is a more subtle property, see Lemma 3.7.5.

observe that lower bounds w.h.p. (and with bounded error) follow analogous arguments by averaging on conditional probabilities, instead of conditional expectations.

Before going into the explicit constructions, we show a result showing that the *universal oracle* (see Introduction, Section 1.2.3) emulates any local oracle. We remind the reader that the universal oracle $\mathcal{O}$ is defined by the property: for query $x \in \mathbb{R}^n$, $\mathcal{O}_f(x)$ is the family of functions $g \in \mathcal{F}$ such that there exists a neighborhood around x (possibly depending on g) where $f = g$. From the following lemma, it suffices to show lower bounds on $\mathcal{O}$ to deduce lower bounds for arbitrary local oracles.

Lemma 3.7.1. *Let $\mathcal{F}$ be a finite family of functions. Then the universal oracle $\mathcal{O}$ is such that any local oracle $\mathcal{O}'$ can be emulated by $\mathcal{O}$*

Proof. Let $\mathcal{O}'$ be any local oracle, and x be a query point. Let the query emulation be the identity. Now, for the answer emulation, by definition, for instances $f, g \in \mathcal{F}$, we have $\mathcal{O}_f(x) = \mathcal{O}_g(x)$ if and only if $f = g$ around x. Therefore the function $a(x, \mathcal{O}_f(x)) = \mathcal{O}'_f(x)$ is well-defined; this defines an oracle emulation of $\mathcal{O}'$ by $\mathcal{O}$, proving the result. □

For the rest of the section, let $\tilde{\mathcal{O}}$ be the single-coordinate oracle studied in previous sections, and let $\mathcal{O}$ be the universal oracle. Note that we state the theorems below for $\mathcal{O}$ an arbitrary local oracle, but from Lemma 3.7.1 w.l.o.g. we may choose for the proofs $\mathcal{O}$ to be the universal oracle.

3.7.1 Large-scale complexity for ℓ_p-Balls

Recall that in Section 3.6.1, different function families were used for the case $1 \leq p < 2$ and $2 \leq p < \infty$. However, the proof below is agnostic to which family is used, by following the notation from (56).

Theorem 3.7.2. *Let $1 \leq p < \infty$, $\varepsilon \geq 1/n^{\max\{p,2\}}$, $\mathcal{F} = \mathcal{F}_p^n(1)$, and let X to be the n-dimensional unit ball B_p. Then, for error probability P_e, the distributional oracle complexity of problem class $\mathcal{P} = (\mathcal{F}, X)$ is $\Omega\left((1-P_e)/\varepsilon^{\max\{p,2\}}\right)$, and the high probability complexity of level β is $\Omega\left(\beta/\varepsilon^{\max\{p,2\}}\right)$.*

Before proving this theorem let us introduce the hard function family, which is a perturbed version of the hard instances in Section 3.6.1.

Construction of function family

Let $1 \leq p < \infty$, $\varepsilon \geq 1/n^{\max\{p,2\}}$, and $X := B_p$. Let M and f_s be defined as in the proof of Theorem 3.6.1, and $\bar{\delta} := \varepsilon/(KM)$, where $K > 0$ is a constant. Consider the infinite family $\mathcal{F} := \left\{f_{s,\delta}(x) : s \in \{-1,+1\}^M,\ \delta \in [0,\bar{\delta}]^M\right\}$, where

$$f_{s,\delta}(x) = f_s(x+\delta).$$

Finally, we consider the random variable $F = f_{S,\Delta}$ on $\mathcal{F}$ where $S \in \{-1,1\}^M$ and $\Delta \in [0,\bar{\delta}]^M$ are chosen independently and uniformly at random.

Proof. The proof requires two steps: first, showing that the subfamily of instances with a fixed perturbation δ is as hard as the unperturbed one for the single-coordinate oracle. Second, by properly averaging over δ we obtain the expectation lower bound.

Lower bound for fixed perturbation under oracle $\tilde{\mathcal{O}}$ Let $\delta \in [0,\bar{\delta}]^M$ be a fixed vector, and $\tilde{\mathcal{F}} = \{f_{s,\delta} \colon s \in \{-1,+1\}^M\}$. Since $f_{s,\delta}(x) = f_s(x+\delta)$, for a fixed perturbation the subfamily of instances is just a re-centering of the unperturbed ones. We claim that the complexity of this family under $\tilde{\mathcal{O}}$ is lower bounded as follows, $\mathbb{E}[T] \geq \frac{M(1-\varepsilon/K)}{2}$.

In fact, consider the ball $B_p(-\delta, r)$, where $r = 1-\varepsilon/K$. Let $x \in B_p(-\delta, r)$, then

$$\|x\|_p \leq \|x+\delta\|_p + M\bar{\delta} \leq 1 - \varepsilon/K + \varepsilon/K = 1,$$

so $x \in B_p(0,1)$. Therefore, $B_p(-\delta, r) \subseteq B_p(0,1)$, and thus the complexity of $\tilde{\mathcal{F}}$ over $B_p(0,1)$ can be lower bounded by the complexity of the same family over $B_p(-\delta, r)$ (optimization on a subset is easier in terms of oracle complexity). Now observe that the problem of minimizing $\tilde{\mathcal{F}}$ over $B_p(-\delta, r)$ under $\tilde{\mathcal{O}}$ is equivalent to the problem studied in Section 3.6.1, only with the radius scaled by r. This re-scaled problem has the same complexity as the original one, only with an extra r factor. Thus,

$$\mathbb{E}[T] \geq \frac{Mr}{2} = \frac{M(1-\varepsilon/K)}{2} \quad \forall \delta \in [0, \bar{\delta}]^M.$$

Lower Bounds for $\mathcal{F}$ under oracle $\mathcal{O}$ To conclude our proof, we need to argue that oracle $\mathcal{O}$ does not provide more information than $\tilde{\mathcal{O}}$ with probability 1. Let A be an algorithm and T the number of queries it requires to determine S (which is a random variable in both S and Δ).

We will show first that throughout its trajectory $(X^1, \ldots, X^T)$, algorithm A queries singular points of $f_{S,\Delta}$ with probability zero. Formally, we have

Lemma 3.7.3 (on unpredictability, large-scale case)**.** *For an $\mathcal{O}$-based algorithm solving family F with queries $X^1, \ldots, X^T$ we define, for $t \geq 0$, the set of maximizer coordinates as*

$$I^t := \{i \in [M] : S_i(X_i^t + \Delta_i) = f_{S,\Delta}(X^t)\}$$

if $t \leq T$, and $I^t = \emptyset$ otherwise, and let us consider the event E where the set of maximizers include at most one new coordinate at each iteration

$$E := \left\{ \left| I^t \setminus \bigcup_{s<t} I^s \right| \leq 1, \quad \forall t \leq T \right\}.$$

Then $\mathbb{P}[E] = 1$.

Proof. We prove by induction that before every query $t \geq 1$ the set of 'unseen' coordinates $I_c^t := [M] \setminus (\cup_{s<t} I^s)$ is such that perturbations $(\Delta_i)_{i \in I_c^t}$ are absolutely

continuous (w.r.t. the Lebesgue measure). Moreover, from this we can prove simultaneously that

$$\mathbb{P}\left[\left|I^t \setminus \bigcup_{s<t} I^s\right| > 1 \,\middle|\, \Pi_{<t}\right] = 0.$$

We start from the base case $t = 0$, which is evident since the distribution on Δ is uniform. Indeed, since singular points (for all possible realizations of S) lie in a smaller dimensional manifold, then $|I^1| = 1$ almost surely. In the inductive step, suppose the claim holds up to t and consider the $(t+1)$-th query. Then what the transcript provides for coordinates in I_c^{t+1} are upper bounds for the perturbations Δ_i given S_i. In fact, from the $(t+1)$-th oracle answer all we obtain are S_j and Δ_j, where j is such that $f_{S,\Delta}(X^{t+1}) = S_j X_j^{t+1} + \Delta_j$; note that such j is almost surely unique among $j \in I_c^t$, by induction. For the rest of the coordinates $i \neq j$ we implicitly know

$$S_i X_i^{t+1} + \Delta_i \leq S_j X_j^{t+1} + \Delta_j,$$

i.e., $\Delta_i \leq D_{i,+}$ if $S_i = 1$, and $\Delta_i \leq D_{i,-}$ if $S_i = -1$; where $D_{i,\pm}$ are constants depending on $(X^0, \ldots, X^{t+1})$, S_j, Δ_j, but not depending on any of the other unknowns. Thus, at every iteration we obtain for non-maximizer coordinates upper bounds on the perturbation Δ_i, conditionally on the sign of S_i. These bounds are such that $\Delta_i = D_{i,\pm}$ with probability zero, as the distribution on $(\Delta_i)_{i \in I_c^t}$ (conditionally on the transcript), which is the one described above, is absolutely continuous. Moreover, by absolute continuity,

$$\mathbb{P}\left[\left|I^{t+1} \setminus \bigcup_{s\leq t} I^s\right| > 1 \,\middle|\, \Pi_{\leq t}\right] = 0,$$

proving the inductive step.

Finally, by the union bound

$$\mathbb{P}\left[\bar{E}\right] \leq \sum_{t=1}^{M} \mathbb{P}\left[\left|I^{t+1} \setminus \bigcup_{s\leq t} I^s\right| > 1\right].$$

And from the previous argument,

$$\mathbb{P}\left[\left|I^{t+1} \setminus \bigcup_{s\le t} I^s\right| > 1\right] = \mathbb{E}_{\Pi_{\le t}}\left[\mathbb{P}\left[\left|I^{t+1} \setminus \bigcup_{s\le t} I^s\right| > 1 \,\middle|\, \Pi_{\le t}\right]\right] = 0.$$

□

With the Lemma on unpredictability the proof becomes straightforward. We claim that on event E, the oracle answer provided by $\mathcal{O}$ can be emulated by the answer provided by $\tilde{\mathcal{O}}$ on the same point; thus, the trajectory of A is equivalent to the trajectory of some algorithm querying $\tilde{\mathcal{O}}$.

To prove our claim, let $\mathcal{O}$ be the universal oracle for the family of perturbed instances $\mathcal{F}$. We observe that on event E, oracle $\tilde{\mathcal{O}}$ is as powerful as $\mathcal{O}$, since the oracle answer of $\mathcal{O}$ for instance $f_{s,\delta}$ is the set $\{f_{r,\gamma} \colon r_j = s_j,\ \gamma_j = \delta_j\}$, where j is the unique maximizer coordinate of $f_{s,\delta}$ on x. Note that this oracle answer can be trivially emulated from the answer by $\tilde{\mathcal{O}}$, which is essentially (s_j, δ_j).

By the claim we conclude that for all δ excluding the measure zero set $\bar{E}$, $\mathbb{E}[T \mid \Delta = \delta] \ge \frac{M(1-\varepsilon/K)}{2}$. By averaging over δ we obtain $\mathbb{E}[T] \ge \frac{M(1-\varepsilon/K)}{2}$. By choosing $K > 0$ arbitrarily large we obtain the desired lower bound. □

3.7.2 Complexity for the box

For the box case we will first introduce the family construction, which turns out to be slightly more involved than the one in Section 3.5. Similarly as in the large-scale case, we first analyze the perturbed family for a fixed perturbation under the single-coordinate oracle, and then we prove the Lemma on unpredictability. With this the rest of the proof is analogous to the large-scale case and thus left as an exercise.

Theorem 3.7.4. *Let $L, R > 0$, $\mathcal{F} = \mathcal{F}_\infty^n(L)$, and let $X = B_\infty(0, R)$. Then, for error probability P_e, the distributional complexity of problem class $\mathcal{P} = (\mathcal{F}, X)$ is $\Omega\left((1-P_e)n \log \frac{LR}{\varepsilon}\right)$, and the high-probability oracle complexity of level β is $\Omega\left(\beta n \log \frac{LR}{\varepsilon}\right)$.*

As in Section 3.5, w.l.o.g. we prove the Theorem for $L = R = 1$, and recall that w.l.o.g. $\mathcal{O}$ is the universal oracle.

One dimensional construction of function family

First we define the perturbed instances for the one dimensional family. The multidimensional family will be defined simply as the maximum of one dimensional functions, as in (54).

We will utilize different perturbations for each level (in the recursive definition) of the function. For this reason, in order to preserve convexity, and in order to not reveal the behavior of lower levels through perturbations, we need to patch the perturbations of consecutive levels in a consistent way.

Given $0 < \varepsilon \leq 1$, let $M := \lfloor \frac{1}{3-\ln\alpha} \ln(1/\varepsilon) \rfloor$ and $\bar{\delta} := \frac{1-\alpha}{4}(\frac{\alpha}{8})^M$, where $\alpha := 1 - 8\varepsilon/(5KM)$, and K is a large constant. Note that for K large enough $\alpha > 1/e$, independently of the values $\varepsilon \in (0, 1]$ and $M \geq 1$; this way, we guarantee that $M \geq \lfloor \frac{1}{4} \ln(1/\varepsilon) \rfloor$. Once we have defined our function family we justify our choice for these parameters.

Let us recall from Section 3.5 the recursive definition of intervals $(I_s)_{s\in\{0,1\}^M}$ and properties (F-1)–(F-5). We will prove there exists a family $\tilde{\mathcal{F}} = \{f_{s,\delta} \colon [-1,1] \to \mathbb{R} : \ s \in \{0,1\}^l,\ 0 < \delta_i \leq \bar{\delta},\ i = 1,\dots,M\}$, satisfying properties (F-1), (F-2), and the analogues of (F-3)–(F-5) described below

(G-3) $f_{s,\delta} \geq f_{s|_l,\delta}$ with $f_{s,\delta}(x) = f_{s|_l,\delta}(x)$ if and only if $x \in [-1,1] \setminus \operatorname{int}\left(I^{\delta}_{s|_l}\right)$, where

$$I^{\delta}_{s|_l} := I_{s|_l}\left[-1 + \left(\frac{2}{\alpha}\right)^l \frac{\delta_{l+1}}{1-\alpha}, 1 - \left(\frac{2}{\alpha}\right)^l \frac{\delta_{l+1}}{1-\alpha/2}\right].$$

(G-4) The function $f_{s,\delta}$ restricted to the interval I_s is of the form

$$f_{s,\delta}(x) = b_{|s|,\delta} - \left(\frac{\alpha}{8}\right)^{|s|} + \left(\frac{\alpha}{2}\right)^{|s|} |x - I_s(0)| \qquad x \in I_s,$$

where $b_{|s|,\delta} = f_{s,\delta}(I_s(-1)) = f_{s,\delta}(I_s(+1))$ is the function value on the endpoints of I_s (defined inductively on $|s|$ and δ_i, $i \leq |s|$).

(G-5) For $t \sqsubseteq s$, we have $f_{s,\delta}(x) < b_{|t|,\delta}$ if and only if $x \in \text{int}\,(I_t)$.

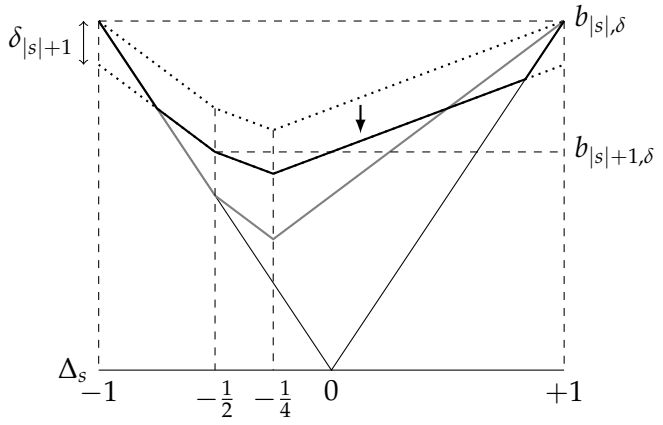

Figure 5: Comparison between instance from Section 3.5.1 (grey line) and perturbed one (thick line).

We construct our instance inductively, the case $|s| = 0$ being trivial ($f_\perp(x) = |x|$; note this function does not depend on the perturbations δ). Moreover, let $b_{0,\delta} = 1$, and inductively $b_{l+1,\delta} := b_{l,\delta} - \frac{\alpha}{2}\left(\frac{\alpha}{8}\right)^l - \delta_{l+1}$. Suppose now $|s| = l$ and $\delta \in [0, \bar{\delta}]^M$, and for simplicity let $s_{l+1} = 0$ (the case $s_{l+1} = 1$ is analogous). By inductive hypothesis $f_{s,\delta}(I_s(-1)) = f_{s,\delta}(I_s(+1)) = b_{|s|,\delta}$. We consider the perturbed extension given by

$$g_{s0,\delta}(x) := \begin{cases} b_{l+1,\delta} - \left(\frac{\alpha}{8}\right)^{l+1} + \left(\frac{\alpha}{2}\right)^{l+1} \left| x - I_s\left(-\frac{1}{4}\right)\right|, & \text{if } x \in I_s\left[-\frac{1}{2}, 1\right] \\ b_{l,\delta} + \alpha\left[f_{s,\delta}(x) - b_{l,\delta}\right] - \delta_{l+1}, & \text{otherwise.} \end{cases}$$

We define the new perturbed instance as follows

$$f_{s0,\delta}(x) = \max\{g_{s0,\delta}(x), f_{s,\delta}(x)\} \qquad x \in [-1, 1].$$

Note for example that at $x = I_s(-1/2)$ the function $g_{s0,\delta}$ is continuous, and moreover $g_{s0,\delta}(x) = b_{l+1,\delta} > b_{l,\delta} - \frac{1}{2}\left(\frac{\alpha}{8}\right)^l = f_{s,\delta}(x)$, where the strict inequality holds by definition of $\bar{\delta}$; similarly, for $x = I_s(0)$, $g_{s0,\delta}(x) = b_{l+1,\delta} > f_{s,\delta}(x)$. This way, we guarantee that at the interval I_{s0} the maximum defining $f_{s0,\delta}$ is only achieved by $g_{s0,\delta}$.

The key property of the perturbed instances is the following: Since $\delta_{l+1} > 0$ then $f_{s0,\delta}$ is smooth at $I_s(-1)$ and $I_s(+1)$, and its local behavior does not depend on $\delta_{l+1}, \ldots, \delta_M$. Furthermore, for all $x \in [-1,1] \setminus \text{int}\left(I_s^\delta\right)$, we have $f_{s0,\delta}(x) = f_{s,\delta}(x)$, from which is easy to prove (G-3).

Finally, observe that properties (F-1), (F-2), (G-4) and (G-5) are straightforward to verify. This proves the existence of our family. Moreover, by construction, the function defined above is convex, continuous, and has Lipschitz constant bounded by 1.

To finish our discussion, let us explain the role of these perturbations, and the choice of parameters. First observe that the definition of $g_{s,\delta}$ is obtained by applying two operations to the extension used in Section 3.5: first we reduce the slope of the extension by a factor α, and then we 'push-down' the function values by an additive perturbation $\delta_{|s|+1}$ (see Figure 5). The motivation for the perturbed family is to provide instances with similar structure than in Section 3.5; in particular, we preserve the nesting property of level sets. The main difference with the perturbed instance is the smoothness at $I_s(-1)$, $I_s(+1)$: by doing this we hide the behavior (in particular the perturbations) of deeper level sets from its behavior outside the interior of this level set, for any local oracle. In the multidimensional construction the perturbations will have a similar role than in the large-scale case, making the maximizer term unique w.p.1. for any oracle query, as perturbations in different coordinates will be conditionally independent. This process will continue throughout iterations, and the independence of perturbations for deeper level sets is crucial for this to happen.

Multidimensional construction of the family

As in the unperturbed case, the obvious multidimensional extension is to consider the maximum among all coordinates of the one dimensional instance, namely, for a concatenation of (nM)-dimensional strings $\{s_i : i \in [n]\}$, s, and concatenation

of (nM)-dimensional vectors $\{\delta_i : i \in [n]\}$, δ, let

$$f_{s,\delta}(x) := \max_{i\in[n]} f_{s_i,\delta_i}(x). \tag{57}$$

Lower bound for fixed perturbation under oracle $\tilde{\mathcal{O}}$. Note that from (F-1) and (G-5) the packing property is satisfied when $M = \lfloor \frac{1}{3-\ln\alpha} \ln(1/\varepsilon) \rfloor$. Next, emulation by the String Guessing Problem comes from analogous results to Lemmas 3.5.4 and 3.5.7, considering the obvious modifications due to the perturbations, and whose proofs are thus omitted. This establishes the lower bound $\Omega(n \log(1/\varepsilon))$.

Lower Bounds for $\mathcal{F}$ under oracle $\mathcal{O}$. Similarly as in the large-scale case, the fundamental task is to prove that w.p. 1 at every iteration the information provided by $\mathcal{O}$ can be emulated by the single-coordinate oracle $\tilde{\mathcal{O}}$ studied earlier.

For this, we will analyze the oracle answer, showing that for any nontrivial query the maximizer in (57) is unique w.p. 1. The role of perturbations is crucial for this analysis. With this in hand, the lower bound comes from an averaging argument analogous the large-scale case.

Lemma 3.7.5 (On unpredictability, box case)**.** *For an $\mathcal{O}$-based algorithm solving family $\mathcal{F}$ with queries $X^1, \dots, X^T$ let the set of maximizer coordinates be*

$$J^t := \{(i,l) :\ f_{S,\Delta}(X^t) = f_{S_i,\Delta_i}(X^t),\ b_{l+1,\delta} < f_{S,\Delta}(X^t) \le b_{l,\delta}\}$$

for $t \le T$, and $J^t = \emptyset$ otherwise. For a query $t \le T$ let the i-th depth l_i be such that (i, l_i) is $\prec$-maximal among elements of J^{t-1} with first coordinate i. Finally, let $J^t_c := \{(i,l) : (i,l) \succ (i,l_i)\}$.

Then the distribution of $(\Delta_{i,h})_{J^t_c}$ conditionally on $(\Pi_{<t}, Q_t)$ is absolutely continuous. Moreover, after the oracle answer A_t, with probability 1 either we only obtain (inexact) lower bounds on some of the $\Delta_{i,h}$, or J^t is a singleton.

Proof. For $t < T$, let the active set be defined as

$$\mathcal{I}^t := \operatorname{int}\left(\prod_{i=1}^{n} I^{\Delta_i}_{s_i|_{l_i+1}}\right).$$

We prove the lemma by induction on t. The case $t = 1$ clearly satisfies that $(I_{i,l})_{(i,l)\in[n]\times[M]}$ is absolutely continuous. Next, after the first oracle call, there are two cases: first, if the query lies outside the active set $\mathcal{I}^1$, then after the oracle answer all what is learnt are lower bounds on the perturbations (this since the instance behaves as an absolute value function of the maximizer coordinates); by absolute continuity these lower bounds are inexact w.p. 1. If the query lies in $\mathcal{I}^1$ then since the perturbations are absolutely continuous, and since (for all possible realizations of S) the set where the maximizer is not unique is a smaller dimensional manifold, the maximizer in $f_{S,\Delta}$ is unique w.p. 1. In this case all bits preceding this maximizer in the $\prec$-order are learnt, and potentially some perturbations for these bits as well.

Next, let $t \geq 1$, and suppose the lemma holds up to query t. Then we know that $(\Delta_{i,h})_{J^t_c}$ is absolutely continuous, conditionally on $(\Pi_{<t}, Q_t)$, and that the oracle answer A_t is such that w.p. 1 either we only obtain (inexact) lower bounds on some $\Delta_{i,h}$, or J^t is a singleton. In the first case, $(\Delta_{i,h})_{J^t_c}$ remains absolutely continuous (since lower bounds are inexact), so clearly the statement holds true for $t+1$. In the case J^t is a singleton, note that $(\Delta_{i,h})_{(i,h)\in J^{t+1}_c}$ remains independent and uniform by construction of the function family. This way, by performing the same analysis as in the base case over the set $\prod_{i=1}^{n} I_{s_i|_{l_i+1}}$ we conclude that the lemma holds for $t+1$. □

Let us define the set

$$E := \bigcap_{t\leq T} \{(\Delta_{i,h})_{J^t_c} \text{ is absolutely continuous } \vee\, |J^t| \leq 1\}.$$

By the previous Lemma, $\mathbb{P}[E] = 1$. It is clear that on event E, oracle $\mathcal{O}$ can be emulated by $\tilde{\mathcal{O}}$ by following an analogous approach as in Section 3.7.1. It is left

as exercise to derive from this the lower complexity bound $\Omega(n \log(1/\varepsilon))$, and its variants for expectation, high probability, and bounded error algorithms.

3.7.2.1 *The low-scale case: reduction to the box when* $1 \leq p < 2$

Finally, as a consequence of our strong lower bounds for arbitrary oracles on the box we derive optimal lower complexity bounds for low-scale optimization over ℓ_p balls for $1 \leq p < 2$

Proposition 3.7.6. *Let* $1 \leq p < 2$*, and* $\varepsilon \leq n^{-\frac{1}{2}-\gamma}$ *with* $\gamma > 0$*. There exists a family* $\mathcal{F}$ *of convex Lipschitz continuous functions in the* ℓ_p *norm with Lipschitz constant* 1 *on the n-dimensional unit Euclidean ball* B_p *such that for any local oracle for family* $\mathcal{F}$ *and error probability* P_e*, the distributional oracle complexity of problem class* $\mathcal{P} = (\mathcal{F}, X)$ *is* $\Omega\left((1 - P_e)n \log \frac{1}{\varepsilon}\right)$ *and the high probability complexity of level* β *is* $\Omega\left(\beta n \log \frac{1}{\varepsilon}\right)$*.*

Proof. This proof is based on convex geometry and it is inspired by [13].

Let $\varepsilon \leq 1/n^{1/2+\gamma}$ and $X \coloneqq B_p$. By Dvoretzky's Theorem on the ℓ_p-ball [38, Theorem 4.15], there exists a universal constant $\alpha \in (0, 1)$ (i.e., independent of p and n), such that for $k = \lfloor \alpha n \rfloor$ there exists a subspace $L \subseteq \mathbb{R}^n$ of dimension k, and a centered ellipsoid $E \subseteq L$ such that

$$\frac{1}{2}E \subseteq X \cap L \subseteq E. \tag{58}$$

Let $\{\gamma_i(\cdot) : i = 1, \ldots, k\}$ be linear forms on L such that $E = \{y \in L : \sum_{i=1}^k \gamma_i^2(y) \leq 1\}$. By the second inclusion above, for every $i \in [k]$ the maximum of γ_i over $X \cap L$ does not exceed 1, whence, by the Hahn-Banach Theorem, the linear form $\gamma_i(\cdot)$ can be extended from L to $\mathbb{R}^n$ with its maximum over X not exceeding 1. In other words, there exist k vectors $g_i \in \mathbb{R}^n$, $1 \leq i \leq k$, such that $\gamma_i(y) = \langle g_i, y \rangle$ for every $y \in L$ and $\|g_i\|_{p_*} \leq 1$, for all $1 \leq i \leq k$. Now consider the linear mapping

$$x \mapsto Gx \coloneqq (\langle g_1, x \rangle, \ldots, \langle g_k, x \rangle) \colon \mathbb{R}^n \to \mathbb{R}^k.$$

The operator norm of this mapping induced by the norms $\|\cdot\|_p$ on the domain and $\|\cdot\|_\infty$ on the codomain does not exceed 1. Therefore, for any Lipschitz continuous function $f\colon \mathbb{R}^k \to \mathbb{R}$ with Lipschitz constant 1 in the ℓ_∞ norm, the function $\tilde{f}\colon \mathbb{R}^n \to \mathbb{R}$ defined by $\tilde{f}(x) = f(Gx)$ is Lipschitz continuous with constant 1 in the ℓ_p norm. We claim (postponing its proof) that the complexity of Lipschitz continuous functions in the ℓ_p norm on $X \subseteq \mathbb{R}^n$ is lower bounded by the complexity of Lipschitz continuous functions in the ℓ_∞ norm on $B_\infty\left(0, \frac{1}{2\sqrt{k}}\right) \subseteq \mathbb{R}^k$ (as $G\left(B_\infty\left(0, \frac{1}{2\sqrt{k}}\right)\right) \subseteq \frac{1}{2}E \subseteq X$). We conclude that the distributional and high probability oracle complexity of the former family is lower bounded by

$$\Omega\left(k \log \frac{1}{2\sqrt{k}\varepsilon}\right) = \Omega\left(n \log \frac{1}{\varepsilon\sqrt{n}}\right) = \Omega\left(n \log \frac{1}{\varepsilon}\right),$$

for large n, since for $\varepsilon \leq n^{-1/2-\gamma}$ with $\gamma > 0$ we have $\varepsilon\sqrt{n} \leq \varepsilon^{\frac{\gamma}{1/2+\gamma}}$.

We finish the proof by proving the claim: let $\mathcal{G}$ be the subfamily of Lipschitz continuous functions with constant 1 for the ℓ_∞^k norm given by (57), defined on the box $B_\infty^k(0, 1/(2\sqrt{k}))$, and let $\mathcal{F}$ be the respective family of 'lifted' instances $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$, which are Lipschitz continuous functions with constant 1 for the ℓ_p^n norm, defined on the unit ball $B_p^n(0, 1)$.

Observe that the universal oracle $\mathcal{O}$ on $\mathcal{G}$ induces the universal oracle for family $\mathcal{F}$. Namely, if we let $\tilde{\mathcal{O}}$ be the oracle for family $\mathcal{F}$ defined by $\tilde{\mathcal{O}}_{\tilde{f}}(x) = \tilde{\mathcal{O}}_{\tilde{g}}(x)$ if and only if $\mathcal{O}_f(Gx) = \mathcal{O}_g(Gx)$, then it is easy to see that $\tilde{\mathcal{O}}$ is the universal oracle for $\mathcal{F}$. This way, any oracle for $\mathcal{F}$ can be emulated by an oracle for $\mathcal{G}$, and thus by Lemma 3.4.2 lower bounds for $\mathcal{G}$ also hold for $\mathcal{F}$.

□

3.8 Final Comments

Unification of lower bounds and randomized complexity. Our results unify the classical analysis of oracle complexity of nonsmooth convex optimization. In particular, since lower bounds for distributional complexity coincide up to a constant

factor with the worst-case one, we conclude that the complexity of randomized algorithms is lower bounded by worst-case complexity up to a constant factor, closing a logarithmic gap in Nemirovski & Yudin [32].

Beyond the standard setting. Another interesting extension of the provided results is related to the non-standard ℓ_p/ℓ_q-setting studied in the previous chapter. First, notice that since low-scale complexity bounds are the of the same order as worst-case complexity, the only regime where the non-standard setting indeed makes a difference is the large-scale regime.

Construction of hard families in the previous chapter involved a combination of designing families of linear functionals with large gap Γ, and in some situations liftings by random sections/projections of the domain, which preserve lower bounds (see, e.g., the proof of Proposition 3.7.6). These observations lead to the following

Corollary 3.8.1. *The distributional oracle complexity of minimization of the class of functions* $\mathrm{Lip}_q^n(L)$ *over the ball* $B_p^n(R)$ *is lower bounded by*

(i) If $p \le q$, *then* $\mathrm{Compl}_{\mathcal{D}}(\varepsilon) = \Omega\left(\frac{LR}{\varepsilon^{1/\mu}}\right)$, *where* $\mu := \frac{1}{p} - [\frac{1}{q} - \frac{1}{2}]_+$;

(ii) If $p > q$, *then* $\mathrm{Compl}_{\mathcal{D}}(\varepsilon) = \Omega\left(n^{\frac{1}{q}-\frac{1}{p}} \frac{LR}{\varepsilon^{\max\{2,q\}}}\right)$.

Finally, observe that these lower bounds coincide with worst-case upper bounds (see (46) for $\kappa = 1$), up to a constant factor.

Distributional complexity and fat-shattering numbers. In the nonsmooth case we can also consider a very general framework, where the minimization domain is given by a symmetric convex body $X \subseteq \mathbb{R}^n$, and the family of functions $\mathcal{F}$ is the class of Lipschitz continuous w.r.t. a norm $\|\cdot\|$, i.e., for every $f \in \mathcal{F}$ and $x \in \mathbb{R}^n$, $\partial f(x) \subseteq B_{\|\cdot\|_*}$. In this case, it was established by Srebro and Sridharan [42] that

the ε-worst-case oracle complexity can be lower bounded by the *fat-shattering dimension* of the class of linear functionals $B_{\|\cdot\|_*}$ at scale 2ε. We propose the following open problem, which is a generalization of results from this chapter (which were first proposed in [4]) and [42] for the so-called *universal case*.

Open Problem 3.8.2. Under the notation above, does distributional complexity of problem class $\mathcal{P} = (\mathcal{F}, X)$ satisfy the lower bound below?

$$\mathrm{Compl}_{\mathcal{D}}(\varepsilon) = \Omega(\mathrm{fat}_{2\varepsilon}(\mathrm{lin}(X, B_{\|\cdot\|_*}))).$$

In this respect, it is worth noticing that bounds in Corollary 3.8.1 are consistent with this conjecture, and also that this conjecture has not only been verified for worst-case oracle complexity, but also for online optimization [43].

REFERENCES

[1] ARORA, S. and BARAK, B., *Computational complexity*. Cambridge: Cambridge University Press, 2009.

[2] BALL, K., CARLEN, E., and LIEB, E., "Sharp uniform convexity and smoothness inequalities for trace norms," *Inventiones mathematicae*, vol. 115, no. 1, pp. 463–482, 1994.

[3] BOYER, C., WEISS, P., and BIGOT, J., "An algorithm for variable density sampling with block-constrained acquisition," *SIAM J. Imaging Sciences*, pp. 1080–1107, 2014.

[4] BRAUN, G., GUZMÁN, C., and POKUTTA, S., "Lower Bounds on the Oracle Complexity of Convex Optimization Via Information Theory." arXiv:1407.5144, 2014.

[5] CLARKSON, K. L., "Coresets, sparse greedy approximation and the Frank-Wolfe algorithm," in *SODA 2008*, pp. 922–931, 2008.

[6] COVER, T. and THOMAS, J., *Elements of information theory*. Wiley-interscience, 2006.

[7] COX, B., JUDITSKY, A., and NEMIROVSKI, A., "Dual subgradient algorithms for large-scale nonsmooth learning problems," *Mathematical Programming Series B*, pp. 1–38, 2013.

[8] D'ASPREMONT, A., GUZMÁN, C., and JAGGI, M., "An optimal affine invariant smooth minimization algorithm." arXiv:1301.0465, 2014.

[9] D'ASPREMONT, A., "Smooth optimization with approximate gradient.," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1171–1183, 2008.

[10] DEMYANOV, V. and RUBINOV, A., *Approximate Methods in Optimization Problems*. American Elsevier, 1970.

[11] FRANK, M. and WOLFE, P., "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, vol. 3, pp. 95–110, 1956.

[12] GARBER, D. and HAZAN, E., "Playing non-linear games with linear oracles," in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pp. 420–428, IEEE, 2013.

[13] GUZMÁN, C. and NEMIROVSKI, A., "On lower complexity bounds for large-scale smooth convex optimization," *Journal of Complexity*, vol. 31, no. 1, pp. 1 – 14, 2015.

[14] HARCHAOUI, Z., JUDITSKY, A., and NEMIROVSKI, A., "Conditional gradient algorithms for norm-regularized smooth convex optimization," *Mathematical Programming Series A*, pp. 1–38, 2014.

[15] HAZAN, E., "Sparse approximate solutions to semidefinite programs," in *LATIN 2008*, pp. 306–316, 2008.

[16] HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C., *Fundamentals of Convex Analysis*. Springer Verlag, Heidelberg, 2001.

[17] JAGGI, M., *Sparse Convex Optimization Methods for Machine Learning*. PhD thesis, ETH Zurich, Oct. 2011.

[18] JAGGI, M., "Revisiting Frank-Wolfe: projection-free sparse convex optimization," in *ICML 2013*, vol. 28, pp. 427–435, 2013.

[19] JAGGI, M. and SULOVSKY, M., "A simple algorithm for nuclear norm regularized problems," in *ICML*, 2010.

[20] JUDITSKY, A. and NEMIROVSKI, A., "Large deciations of vector-values martingales in 2-smooth normed spaces." arXiv:0809.0813, 2008.

[21] JUDITSKY, A., NEMIROVSKI, A., and TAUVEL, C., "Solving variational inequalities with stochastic mirror-prox algorithm," *Stoch. Syst.*, vol. 1, no. 1, pp. 17–58, 2011.

[22] KHACHIYAN, L., NEMIROVSKI, A., and NESTEROV, Y., "Optimal methods for the solution of large-scale convex programming problems," in *Modern Mathematical Methods in Optimization* (ELSTER, K.-H., ed.), Academie Verlag, Berlin, 1993.

[23] KHACHIYAN, L., TARASOV, S., and ERLIKH, A., "The Inscribed Ellipsoid Method," *Soviet Math. Doklady*, vol. 37, pp. 226–230, 1988.

[24] LAN, G., "The complexity of large-scale convex programming under a Linear Optimization oracle." arXiv:1309.5550, 2013.

[25] LEVIN, A., "On an algorithm for the minimization of convex functions.," *Sov. Math., Dokl.*, vol. 6, pp. 268–290, 1965.

[26] MOREAU, J.-J., "Fonctions convexes duales et points proximaux dans un espace hilbertien," *CR Acad. Sci. Paris Sér. A Math*, vol. 255, pp. 2897–2899, 1962.

[27] MOREAU, J.-J., "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société mathématique de France*, vol. 93, pp. 273–299, 1965.

[28] NEMIROVSKI, A., "On optimality of Krylov's information when solving linear operator equations," *Journal of Complexity*, vol. 7, no. 2, pp. 121–130, 1991.

[29] NEMIROVSKI, A., "Information-Based Complexity of linear operator equations," *Journal of Complexity*, vol. 8, no. 2, pp. 153–175, 1992.

[30] NEMIROVSKI, A., "Efficient Methods in Convex Programming." `http://www2.isye.gatech.edu/~nemirovs/`, 1994.

[31] NEMIROVSKI, A., ONN, S., and ROTHBLUM, U., "Accuracy certificates for computational problems with convex structure," *Mathematics of Operations Research*, vol. 35, no. 1, pp. 52–78, 2010.

[32] NEMIROVSKI, A. and YUDIN, D., *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1 ed., 1983.

[33] NEMIROVSKI, A., "Discussion on: "why is resorting to fate wise? a critical look at randomized algorithms in systems and control"," *European Journal of Control*, vol. 16, no. 5, pp. 432 – 436, 2010.

[34] NESTEROV, Y., "A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$," *Soviet Math. Dokl.*, vol. 27:2, pp. 372–376, 1983.

[35] NESTEROV, Y. and NEMIROVSKI, A., "On First-Order Algorithms for ℓ_1/Nuclear Norm Minimization," *Acta Numerica*, vol. 22, pp. 509–575, 4 2013.

[36] NEWMAN, D. J., "Location of the maximum on unimodal surfaces," *J. ACM*, vol. 12, pp. 395–398, July 1965.

[37] PACKEL, E. W., "Complexity and information by Joseph Traub and A. G. Werschulz," *Complexity*, vol. 4, no. 5, pp. 39–40, 1999.

[38] PISIER, G., *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, 1 ed., 1989.

[39] PSHENICHNYJ, B. and DANILIN, Y., *Numerical Methods in Extremal Problems*. Mir, 1978.

[40] RUDELSON, M. and VERSHYNIN, R., "On sparse reconstruction from Fourier and Gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

[41] SHOR, N., "Cut-off method with space extension in convex programming problems," *Cybernetics*, vol. 13, no. 1, pp. 94–96, 1977.

[42] SREBRO, N. and SRIDHARAN, K., "On convex optimization, fat shattering and learning." `http://ttic.uchicago.edu/~karthik/optfat.pdf`, 2012.

[43] SREBRO, N., SRIDHARAN, K., and TEWARI, A., "On the universality of online mirror descent," in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pp. 2645–2653, 2011.

[44] TRAUB, J. F. J. F., WASILKOWSKI, G. W., and WOŹNIAKOWSKI, H., *Information-based complexity*. Computer science and scientific computing, Boston: Academic Press, 1988. Includes indexes.

[45] YOSIDA, K., *Functional analysis*. Springer Verlag, Berlin, 1964.