Journal of Complexity 31 (2015) 1–14

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

On lower complexity bounds for large-scale smooth convex optimization ☆

Cristóbal Guzmán *, Arkadi Nemirovski

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive NW Atlanta, GA 30332-0205, USA

ARTICLE INFO

Article history:
Received 18 March 2014
Accepted 4 August 2014
Available online 13 August 2014

Keywords:
Smooth convex optimization
Lower complexity bounds
Optimal algorithms

ABSTRACT

We derive lower bounds on the black-box oracle complexity of large-scale smooth convex minimization problems, with emphasis on minimizing smooth (with Hölder continuous, with a given exponent and constant, gradient) convex functions over high-dimensional $\|\cdot\|_p$-balls, $1 \le p \le \infty$. Our bounds turn out to be tight (up to logarithmic in the design dimension factors), and can be viewed as a substantial extension of the existing lower complexity bounds for large-scale convex minimization covering the nonsmooth case and the "Euclidean" smooth case (minimization of convex functions with Lipschitz continuous gradients over Euclidean balls). As a byproduct of our results, we demonstrate that the classical Conditional Gradient algorithm is near-optimal, in the sense of Information-Based Complexity Theory, when minimizing smooth convex functions over high-dimensional $\|\cdot\|_\infty$-balls and their matrix analogies – spectral norm balls in the spaces of square matrices.

1. Introduction

Huge sizes of convex optimization problems arising in some modern applications (primarily, in big-data-oriented signal processing and machine learning) are beyond the "practical grasp" of the state-of-the-art Interior Point Polynomial Time methods with their computationally demanding iterations.

☆ This research was supported by the NSF grant CMMI-1232623.

* Corresponding author.
E-mail addresses: cguzman@gatech.edu (C. Guzmán), arkadi.nemirovski@isye.gatech.edu (A. Nemirovski).

http://dx.doi.org/10.1016/j.jco.2014.08.003

Indeed, aside of rare cases of problems with "extremely favourable" structure, the arithmetic cost of an interior point iteration is at least cubic in the design dimension n of the instance; with n in the range of 10^4–10^6, as is the case in the outlined applications, this makes a single iteration "last forever". The standard techniques for handling large-scale convex problems – those beyond the practical grasp of Interior Point methods – are First Order methods (FOM's). Under favourable circumstances, iterations of FOM's are much cheaper than those of interior point methods, and the convergence rate, although just sublinear, is fully or nearly dimension-independent, which makes FOM's the methods of choice when medium-accuracy solutions to large-scale convex programs are sought. Now, as a matter of fact, all known FOM's are "black-box-oriented"—they "learn" the problem being solved solely via the local information (values and (sub)gradients of the objective and the constraints) accumulated along the search points generated by the algorithm. As a result, "limits of performance" of FOM's are governed by Information-Based Complexity Theory. Some basic results in this direction have been established in the literature [17]; in particular, we know well enough what is the Information-Based Complexity of natural families of convex minimization problems $\min_{x\in X} f(x)$ with *nonsmooth* Lipschitz continuous objectives f and how the complexity depends on the geometry and the dimension of the domain X. In the smooth case, our understanding is somehow limited; essentially, tight *lower* complexity bounds are known only in the case when X is Euclidean ball and f is convex function with Lipschitz continuous gradient. Lower bounds here come from least-squares problems [13,14], and the underlying techniques for generating "hard instances" heavily utilize the rotational invariance of a Euclidean ball.

In this paper, we derive tight lower bounds on information-based complexity of families of convex minimization problems $\{\min_{x\in X} f(x) : f \in \mathcal{F}\}$, where X is n-dimensional $\|\cdot\|_p$-ball, $1 \le p \le \infty$, and $\mathcal{F}$ is the family of all continuously differentiable convex objectives with given smoothness parameters (Hölder exponent and constant). We believe that these bounds could be of interest in some modern applications, like ℓ_1 and nuclear norm minimization in Compressed Sensing, where one seeks to minimize a smooth, most notably, quadratic convex function over a high-dimensional ℓ_1-ball in $\mathbf{R}^n$ or nuclear norm ball in the space of $n \times n$ matrices. Another instructive application of our results is establishing the near-optimality, in the sense of information-based complexity, of Conditional Gradient (a.k.a. Frank–Wolfe) algorithm as applied to minimizing smooth convex functions over large-scale boxes (or unit balls of spectral norm on the space of matrices).[1]

1.1. Contributions

Our first contribution is a unified framework to prove lower bounds for a variety of domains and different smoothness parameters of the objective with respect to a norm (for consistency we use the norm induced by the domain). 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 [8] of a Lipschitz continuous convex function f and the smoothing kernel is smooth, and its local behaviour depends only on the local behaviour 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.[2]

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 [17], which now are properly smoothed by the inf-convolution.

1 Originating from [4], Conditional Gradient algorithm was intensively studied in 1970s (see [3,21] 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 [7,9,10,6,2] and references therein.

2 It well may happen that the extensions of the classical Moreau results which we present in Section 2 are known, so that the material in this section does not pretend to be novel. This being said, at this point in time we do not have at our disposal references to the results on smoothing we need, and therefore we decided to augment these simple results with their proofs, in order to make our presentation self-contained.

With this general result, we are able to provide a unified analysis for lower bounds for smooth convex minimization over n-dimensional $\|\cdot\|_p$-balls, $1 \le p \le \infty$. We show that in the large-scale case, our lower complexity bounds match, within at worst a logarithmic in n factor, the upper complexity bounds associated with Nesterov's fast gradient algorithms [16,11]. When $p = \infty$, this result implies near optimality of the Conditional Gradient algorithm.

As a final application, we point out how our lower bounds extend to matrix optimization under Schatten norm constraints.

1.2. Related work

Oracle complexity: The analysis of convex optimization algorithms via oracle complexity and lower complexity bounds were first studied in [17]. Other standard references are [15,18]. The oracle complexity of smooth convex optimization over Euclidean domains was studied in [17,13,14].

For optimal methods under non-Euclidean domains for smooth spaces and p-norms, where $2 \le p < \infty$, we refer to [11] (for the case $p = 2$ there is an interesting new algorithm that adapts itself to the smoothness parameter in the objective [19]).

It should be mentioned that for the case $2 \le p < \infty$ the lower bounds in this paper were announced in [16,11] (and proved by the second author of this paper); however, aside of the very special case of $p = 2$, the highly technical original proofs of the bounds were never published. For this reason, we recently have revisited the original proofs and were able to simplify them dramatically, thus making them publishable.

The Conditional Gradient algorithm and complexity under Linear Optimization oracles: The recent body of work on the Conditional Gradient algorithm is enormous. For upper bounds on its complexity we refer to [1,7,10,12,5]. Interestingly, the last two references include results on linear convergence of the Conditional Gradient method for the strongly convex case, accelerated methods based on Linear Optimization oracles, and applications to stochastic and online convex programming.

Besides these accuracy upper bounds, there are some interesting lower bounds for algorithms based on a Linear Optimization oracle (whose only assumption is that the Linear Optimization oracle returns a solution that is a vertex of the domain): some of these contributions can be found in [10,12]. Observe that a Linear Optimization oracle is in general less powerful than an arbitrary local oracle (in particular the first-order one) considered in our paper, and thus their lower bounds do not imply ours. However, our result for $p = \infty$ improves on their lower bounds (disregarding logarithmic factors).

1.3. Notation and preliminaries

Algorithms and complexity: In the black-box oracle complexity model for convex optimization we are interested in solving problems of the form

$$\mathrm{Opt}(f) = \min_{x \in X} f(x) \qquad (P_{f,X})$$

where X is a given convex compact subset of a normed space $(\mathbf{E}, \|\cdot\|)$, and f is known to belong to a given family $\mathcal{F}$ of continuous convex functions on $\mathbf{E}$. This defines the family of problems $\mathcal{P}(\mathcal{F}, X)$ comprised of problems $(P_{f,X})$ with $f \in \mathcal{F}$. We assume that the family $\mathcal{F}$ is equipped with an *oracle* $\mathcal{O}$ which, formally, is a function $\mathcal{O}(f, x)$ of $f \in \mathcal{F}$ and $x \in \mathbf{E}$ taking values in some *information space* $\mathcal{I}$; when solving $(P_{f,X})$, an algorithm at every step can sequentially call the oracle at a query point $x \in \mathbf{E}$, obtaining the value $\mathcal{O}(f, x)$. In the sequel, we always assume the oracle to be *local*, meaning that for all $x \in \mathbf{E}$ and $f, g \in \mathcal{F}$ such that $f(\cdot) = g(\cdot)$ in a neighbourhood of x, we have $\mathcal{O}(f, x) = \mathcal{O}(g, x)$. The most common example of oracle is the first-order oracle, which returns the value and a subgradient of f at x. However, observe that when the subdifferential is not a singleton not every such oracle satisfies the local property, and we need to further restrict it to satisfy locality.

A T-step algorithm $\mathcal{M}$, utilizing oracle $\mathcal{O}$, for the family $\mathcal{P}(\mathcal{F}, X)$ is a procedure as follows. As applied to a problem $(P_{f,X})$ with $f \in \mathcal{F}$, $\mathcal{M}$ generates a sequence $x_t = x_t(\mathcal{M}, f)$, $1 \le t \le T$ of *search points* according to the recurrence

$$x_t = X_t(\{x_\tau, \mathcal{O}(f, x_\tau)\}_{\tau=1}^{t-1}),$$

where the *search rules* $X_t(\cdot)$ are deterministic functions of their arguments; we can identify $\mathcal{M}$ with the collection of these rules. Thus, x_1 is specified by $\mathcal{M}$ and is independent of f, and all subsequent search points are deterministic functions of the preceding search points and the information on f provided by $\mathcal{O}$ when queried at these points. We treat $x_T = x_T(\mathcal{M}, f)$ as the approximate solution generated by the T-step solution method $\mathcal{M}$ applied to $(P_{f,X})$, and define the *minimax risk* associated with the family $\mathcal{P}(\mathcal{F}, X)$ and oracle $\mathcal{O}$ as the function of T defined by

$$\mathrm{Risk}_{\mathcal{F},X,\mathcal{O}}(T) = \inf_{\mathcal{M}} \sup_{f\in\mathcal{F}} \left[f(x_T(\mathcal{M}, f)) - \mathrm{Opt}(f)\right],$$

where the right hand side infimum is taken over all T-step solution algorithms $\mathcal{M}$ utilizing oracle $\mathcal{O}$ and such that $x_T(\mathcal{M}, f) \in X$ for all $f \in \mathcal{F}$. The inverse to the risk function

$$\mathcal{C}_{\mathcal{F},X,\mathcal{O}}(\varepsilon) = \min\left\{k : \mathrm{Risk}_{\mathcal{F},X,\mathcal{O}}(k) \le \varepsilon\right\}$$

for $\varepsilon > 0$ is called the *information-based (or oracle) complexity of the family* $\mathcal{P}(\mathcal{F}, X)$ with respect to oracle $\mathcal{O}$.

Geometry and smoothness: Let $\mathbf{E}$ be an n-dimensional Euclidean space, and $\|\cdot\|$ be a norm on $\mathbf{E}$ (not necessarily the Euclidean one). Let, further, X be a nonempty closed and bounded convex set in $\mathbf{E}$. Given a positive real L and $\kappa \in (1, 2]$, consider the family $\mathcal{F}_{\|\cdot\|}(\kappa, L)$ of all continuously differentiable convex functions $f : \mathbf{E} \to \mathbf{R}$ which are (κ, L)-smooth w.r.t. $\|\cdot\|$, i.e. satisfy the relation

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|^{\kappa-1} \quad \forall x, y \in E, \tag{1}$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$. We associate with $\|\cdot\|, X, \kappa, L$ the family of convex optimization problems $\mathcal{P} = \mathcal{P}(\mathcal{F}_{\|\cdot\|}(\kappa, L), X)$.

We assume the family $\mathcal{F}_{\|\cdot\|}(\kappa, L)$ is equipped with a local oracle $\mathcal{O}$. To avoid extra words, we assume that this oracle is at least as powerful as the First Order oracle, meaning that $f(x)$, $\nabla f(x)$ is a component of $\mathcal{O}(f, x)$.

Our goal is to establish lower bounds on the risk $\mathrm{Risk}(T)$, taken w.r.t. the oracle $\mathcal{O}$, of the just defined family of problems $\mathcal{P}$. In the sequel, we focus solely on the *'large-scale' case* $T \le n$, and the reason is as follows: it is known [17] that when $T \gg n$, $\mathrm{Risk}(T)$ "basically forgets the details specifying $\mathcal{P}$" and is upper-bounded by $O(\exp\{-CT/n\})$, where C is an absolute constant, with the data $\|\cdot\|, X, L, \kappa$ of $\mathcal{P}$ affecting only the hidden factor in the outer $O(\cdot)$ and thus irrelevant when $T \gg n$. In contrast to this, in the large-scale regime $T \le n$, $\mathrm{Risk}(T)$ is (at least in the cases we are about to consider) nearly independent of n and goes to 0 *sublinearly* as T grows, and its behaviour in this range heavily depends on $\mathcal{P}$. In what follows, we focus solely on the large-scale regime.

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 behaviour 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.

2.1. Smoothing kernel

Let $(\mathbf{E}, \langle\cdot,\cdot\rangle)$ be a finite-dimensional Euclidean space, $\|\cdot\|$ be a norm on $\mathbf{E}$ (not necessarily induced by $\langle\cdot,\cdot\rangle$), and $\mathcal{C}_{\|\cdot\|}$ be the set of all Lipschitz continuous, with constant 1 w.r.t. $\|\cdot\|$, convex functions on $\mathbf{E}$. Let also $\phi(\cdot)$ ("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

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

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

$$\min_{h\in\text{Dom}\phi}[f(x+h)+\phi(h)] = \min_{h\in\text{int}G}[f(x+h)+\phi(h)], \tag{3}$$

and the right hand side minimum is achieved.

Given a function $f \in \mathcal{C}_{\|\cdot\|}$, we refer to the function

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

as to the *smoothing* of f; it clearly belongs to $\mathcal{C}_{\|\cdot\|}$. Observe that by our assumptions on ϕ we have

1. $\mathcal{S}[f](x) = f(x+h(x))+\phi(h(x))$, where $h(x) \in \text{int}\,G$ is such that
$$f'(x+h(x))+\phi'(h(x)) = 0 \tag{4}$$
for properly selected $f'(x+h(x)) \in \partial f(x+h(x))$;
2. $f(x) \geq \mathcal{S}[f](x) \geq f(x) - \rho_{\|\cdot\|}(G)$, where
$$\rho_{\|\cdot\|}(G) = \max_{h\in G}\|h\|;$$
indeed, 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 $f \in \mathcal{C}_{\|\cdot\|}$), while $h(x) \in G$.
3. We have that for all $f \in \mathcal{C}_{\|\cdot\|}$
$$\|\nabla\mathcal{S}[f](x) - \nabla\mathcal{S}[f](y)\|_* \leq M_\phi\|x-y\| \quad \forall x,y \in E. \tag{5}$$

For a proof of (5) see Appendix A.

2.2. Approximating a function by smoothing

For $\chi > 0$ and $f \in \mathcal{C}_{\|\cdot\|}$, let

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

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

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

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

It follows that

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

The latter relation combines with (5) to imply that

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

As bottom-line, if we can find a function ϕ as described above we have that for any convex function $f : \mathbf{E} \to \mathbf{R}$ with Lipschitz constant 1 w.r.t. $\|\cdot\|$ 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. $\|\cdot\|$:

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

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 $x + \chi G$.

2.3. Example: p-norm smoothing

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

$$\phi(x) = \phi_{r,\theta}(x) = 2\left(\sum_{j=1}^n |x_j|^r\right)^{2\theta/r}, \qquad G = \{x \in \mathbf{R}^n : \|x\|_p \le 1\}. \tag{6}$$

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

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

Applying the results of the previous section, we get

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

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

(ii) $\|\nabla \mathcal{S}_\chi[f](x) - \nabla \mathcal{S}_\chi[f](y)\|_{\frac{p}{p-1}} \le O(1)\min[p, \ln n]\chi^{-1}\|x-y\|_p$ *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*

$$B^p_\chi(x) = \{y : \|y - x\|_p \le \chi\}.$$

3. Lower complexity bounds for smooth convex minimization

In this section we utilize Proposition 1 to prove our main result, namely, a general lower bound on the oracle complexity of smooth convex minimization, and then specialize this result for the case of minimization over $\|\cdot\|_p$ balls, where $2 \le p \le \infty$.

Proposition 2. *Let*

I. *$\|\cdot\|$ be a norm on $\mathbf{R}^n$ and X be a nonempty convex set containing the unit ball of $(\mathbf{R}^n, \|\cdot\|)$;*

II. *T be a positive integer and Δ be a positive real with the following property:*
There exist T linear forms $\langle \omega_i, \cdot\rangle$ on $\mathbf{R}^n$, $1 \le i \le T$, such that
(a) $\|\omega_i\|_* \le 1$ *for $i \le T$, and*
(b) *for every collection $\xi^T = (\xi_1, \ldots, \xi_T)$ with $\xi_i \in \{-1, 1\}$, it holds*

$$\min_{x\in X} \max_{1\le i\le T} \xi_i \langle \omega_i, x\rangle \le -\Delta; \tag{7}$$

[3] From now on, $O(1)$'s stand for appropriate positive absolute constants.

III. *M and ρ be positive reals such that for properly selected convex twice continuously differentiable on an open convex set* $\mathrm{Dom}\,\phi \subset \mathbf{R}^n$ *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 2.1 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 $\mathcal{M}$ associated with this oracle there exists a problem $(P_{f,X})$ with $f \in \mathcal{F}_{\|\cdot\|}(\kappa, L)$ such that

$$f(x_T(\mathcal{M}, f)) - \mathrm{Opt}(f) \ge \frac{\Delta^\kappa}{2^{\kappa+1}(\rho M)^{\kappa-1}} \cdot \frac{L}{T^{\kappa-1}}. \tag{8}$$

Proof. 1^0. Let us set

$$\delta = \frac{\Delta}{2T}, \qquad \chi = \frac{\delta}{2\rho} = \frac{\Delta}{4T\rho}, \qquad \beta = \frac{L\chi^{\kappa-1}}{2^{2-\kappa}M^{\kappa-1}} = \frac{L\Delta^{\kappa-1}}{2^\kappa(T\rho M)^{\kappa-1}}. \tag{9}$$

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

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

Observe that all these functions belong to $\mathcal{C}_{\|\cdot\|}$ due to $\|\omega_j\|_* \le 1$, for $j \le T$, so that the smoothed functions

$$f^{\sigma(\cdot),\xi^T}(x) = \beta\mathcal{S}_\chi[g^{\sigma(\cdot),\xi^T}](x) \tag{10}$$

(see Section 2.2) are well defined continuously differentiable convex functions on $\mathbf{R}^n$ which, by item S.1 in Section 2.2, satisfy that for all x, y in X

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

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

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

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

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

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

3^0. Given a local oracle $\mathcal{O}$ and an associated T-step method $\mathcal{M}$, let us define a sequence $x_1, \dots, x_T$ of points in $\mathbf{R}^n$, a permutation $\sigma(\cdot)$ of $\{1, \dots, T\}$ and a collection $\xi^T \in \{-1, 1\}^T$ by the following T-step recurrence:

- *Step* 1: x_1 is the first point of the trajectory of $\mathcal{M}$ (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, $1 \le i \le T$, that maximizes $|\langle\omega_i, x_1\rangle|$, and specify $\xi_1 \in \{-1, 1\}$ in such a way that $\xi_1\langle\omega_{\sigma(1)}, x_1\rangle = |\langle\omega_{\sigma(1)}, x_1\rangle|$. We set
$$g^1(x) = \xi_1\langle\omega_{\sigma(1)}, x\rangle, \qquad f^1(x) = \beta\mathcal{S}_\chi[g^1](x).$$
- *Step* t, $2 \le t \le T$: At the beginning of this step, we have at our disposal the already built points $x_\tau \in \mathbf{R}^n$, distinct from each other integers $\sigma(\tau) \in \{1, \dots, T\}$ and quantities $\xi_\tau \in \{-1, 1\}$, for $1 \le \tau < t$. At step t, we build $x_t, \sigma(t), \xi_t$, as follows. We set
$$g^{t-1}(x) = \max_{1\le\tau<t} \left[\xi_\tau\langle\omega_{\sigma(\tau)}, x\rangle - (\tau-1)\delta\right],$$
thus getting a function from $\mathcal{C}_{\|\cdot\|}$, and define its smoothing $f^{t-1}(x) = \beta\mathcal{S}_\chi[g^{t-1}](x)$ which, same as above, belongs to $\mathcal{F}_{\|\cdot\|}(\kappa, L)$. We further define
– x_t as the t-th point of the trajectory of $\mathcal{M}$ as applied to f^{t-1},

– $\sigma(t)$ as the index i that maximizes $|\langle \omega_i, x_t \rangle|$, over $i \le T$ *distinct from* $\sigma(1), \dots, \sigma(t-1)$,
– $\xi_t \in \{-1, 1\}$ such that $\xi_t \langle \omega_{\sigma(t)}, x_t \rangle = |\langle \omega_{\sigma(t)}, x_t \rangle|$
thus completing step t.

After T steps of this recurrence, we get at our disposal a sequence $x_1, \dots, x_T$ of points from $\mathbf{R}^n$, a permutation $\sigma(\cdot)$ of indexes $1, \dots, T$ and a collection $\xi^T = (\xi_1, \dots, \xi_T) \in \{-1, 1\}^T$; these entities define the functions

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

4^0. We claim that $x_1, \dots, x_T$ is the trajectory of $\mathcal{M}$ as applied to f^T. By construction, x_1 indeed is the first point of the trajectory of $\mathcal{M}$ 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 $t \le s \le T$

$$\xi_s \langle \omega_{\sigma(s)}, x_{t-1} \rangle \le |\langle \omega_{\sigma(t-1)}, x_{t-1} \rangle| = \xi_{t-1} \langle \omega_{\sigma(t-1)}, x_{t-1} \rangle. \tag{11}$$

Also

$$g^T(x) = \max\Big[g^{t-1}(x), \underbrace{\max_{t \le s \le T} [\xi_s \langle \omega_{\sigma(s)}, x \rangle - (s-1)\delta]}_{=:g_t(x)}\Big], \tag{12}$$

and

$$g^{t-1}(x_{t-1}) \ge \xi_{t-1} \langle \omega_{\sigma(t-1)}, x_{t-1} \rangle - (t-2)\delta.$$

Invoking (11), we get

$$\begin{aligned} t \le s \le T \;\Rightarrow\; & g^{t-1}(x_{t-1}) \ge [\xi_s \langle \omega_{\sigma(s)}, x_{t-1} \rangle - (s-1)\delta] + \delta \\ \Rightarrow\; & g^{t-1}(x_{t-1}) \ge g_t(x_{t-1}) + \delta. \end{aligned}$$

Since both g^{t-1} and g_t belong to $\mathcal{C}_{\|\cdot\|}$, it follows that $g^{t-1}(x) \ge g_t(x)$ in the $\|\cdot\|$-ball B of radius $\delta/2$ centred at x_{t-1}, whence, by (12),

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

From $\chi\rho = \delta/2$ we have that $g^{t-1} \in \mathcal{C}_{\|\cdot\|}$ and $g^T \in \mathcal{C}_{\|\cdot\|}$ coincide on the set $x_{t-1} + \chi G$, whence, as we know from item S.3 in Section 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 neighbourhood of x_{t-1}, as claimed.

5^0. We have

$$\begin{aligned} g^T(x_T) &\ge \xi_T \langle \omega_{\sigma(T)}, x_T \rangle - (T-1)\delta \\ &= |\langle \omega_{\sigma(T)}, x_T \rangle| - (T-1)\delta \\ &\ge -(T-1)\delta, \end{aligned}$$

whence, by item S.2 in Section 2.2, $\mathcal{S}_\chi [g^T](x_T) \ge -(T-1)\delta - \chi\rho \ge -T\delta = -\Delta/2$, implying that

$$f^T(x_T) \ge -\beta\Delta/2.$$

On the other hand, by (7) there exists $x_* \in X$ such that $g^T(x_*) \le \max_{1 \le i \le T} \xi_i \langle \omega_{\sigma(i)}, x_* \rangle \le -\Delta$, whence $\mathcal{S}_\chi [g^T](x_*) \le g^T(x_*) \le -\Delta$ and thus $\mathrm{Opt}(f^T) \le f^T(x_*) \le -\beta\Delta$. Since, as we have seen, $x_1, \dots, x_T$ is the trajectory of $\mathcal{M}$ as applied to f^T, x_T is the approximate solution generated by $\mathcal{M}$ as applied to f^T, and we see that the inaccuracy of this solution, in terms of the objective, is at least $\frac{\beta\Delta}{2} = \frac{\Delta^\kappa}{2^{\kappa+1}(\rho M)^{\kappa-1}} \cdot \frac{L}{T^{\kappa-1}}$, as required. Besides this, f^T is of the form $f^{\sigma(\cdot), \xi^T}$, and we have seen that all these functions belong to $\mathcal{F}_{\|\cdot\|}(\kappa, L)$. $\square$

Remark 1. Note that the previous result immediately implies the lower bound

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

on the complexity of the family $\mathcal{P}(\mathcal{F}_{\|\cdot\|}(\kappa, L), X)$, provided X contains the unit $\|\cdot\|$-ball. Note that this bound is independent of the local oracle $\mathcal{O}$.

The case where X contains a $\|\cdot\|$-ball of radius $R > 0$ instead of the unit $\|\cdot\|$ ball can be reduced to the latter case by scaling instances $f(\cdot) \mapsto f(\cdot/R)$, which corresponds to the transformation $(\kappa, L) \mapsto (\kappa, \bar{L} := LR^{\kappa})$ of the smoothness parameters. Thus, assuming that X contains $\|\cdot\|$-ball of radius R, we have

$$\mathrm{Risk}_{\mathcal{F},R\cdot X,\mathcal{O}}(T) \geq \frac{\Delta^{\kappa}}{2^{\kappa+1}(\rho M)^{\kappa-1}} \cdot \frac{LR^{\kappa}}{T^{\kappa-1}}. \quad (13)$$

4. Case of $\|\cdot\| = \|\cdot\|_p$

In this section we provide lower complexity bounds for smooth convex optimization over $\|\cdot\|_p$-balls for the case when $\|\cdot\| = \|\cdot\|_p$. In Section 4.1 we show that Proposition 2 implies nearly tight optimal complexity bounds for the range $2 \leq p \leq \infty$; moreover, for fixed and finite p, the bound is tight within a factor depending solely on p. For the case $p = \infty$, our lower bound matches the approximation guarantees of the Conditional Gradient algorithm, up to a logarithmic factor, proving near-optimality of the algorithm.

In Section 4.2 we study the range $1 \leq p < 2$. Here we prove nearly optimal complexity bounds by using nearly-Euclidean sections of the $\|\cdot\|_p$-ball, together with the $p = \infty$ lower bound.

4.1. *Smooth convex minimization over $\|\cdot\|_p$-balls, $2 \leq p \leq \infty$*

Consider the case when $\|\cdot\|$ is the norm $\|\cdot\|_p$ on $\mathbf{R}^n$, $2 \leq p \leq \infty$. Given positive integer $T \leq n$, let us specify ω_i, $1 \leq i \leq T$, as the first T standard basis vectors, so that for every collection $\xi^T \in \{-1, 1\}^T$ one clearly has

$$\min_{\|x\|_p \leq 1} \max_{1 \leq i \leq T} \xi_i \langle \omega_i, x \rangle \leq -T^{-1/p}. \quad (14)$$

Invoking the results from Section 2.3 (cf. Proposition 1), we see that when $X \subset \mathbf{R}^n$ is a convex set containing the unit $\|\cdot\|_p$-ball, Assumptions II and III in Proposition 2 are satisfied with $M = O(1)\min[p, \ln n]$, $\rho = 1$ and $\Delta = T^{-1/p}$. Applying Proposition 2, we arrive at

Corollary 1. *Let $2 \leq p \leq \infty$, $\kappa \in (1, 2]$, $L > 0$, and let $X \subset \mathbf{R}^n$ be a convex set containing the unit ball w.r.t. $\|\cdot\|_p$. Then, for every $T \leq n$ and every local oracle $\mathcal{O}$, the minimax risk of the family of problems $\mathcal{P}(\mathcal{F}, X)$ with $\mathcal{F} = \mathcal{F}_{\|\cdot\|_p}(\kappa, L)$ admits the lower bound*

$$\mathrm{Risk}_{\mathcal{F},X,\mathcal{O}}(T) \geq \frac{O(1)}{[\min[p, \ln n]]^{\kappa-1}} \frac{L}{T^{\kappa+\frac{\kappa}{p}-1}}, \quad (15)$$

independent of the local oracle $\mathcal{O}$ in use.

Let us discuss some interesting consequences of the above result.

A. *Complexity of smooth minimization over the box:* Corollary 1 implies that when X is the unit $\|\cdot\|_\infty$-ball in $\mathbf{R}^n$, the T-step minimax risk $\mathrm{Risk}_{\mathcal{F},X,\mathcal{O}}(T)$ of minimizing over X of objectives from the family $\mathcal{F} = \mathcal{F}_{\|\cdot\|_\infty}(\kappa, L)$ in the range $T \leq n$ is lower-bounded by $O(1)L/(T^{\kappa-1}\ln n)$. On the other hand, from the standard efficiency estimate of Conditional Gradient algorithm (see, e.g., [3,21,2]) it follows that when applying the method to minimizing a function $f \in \mathcal{F}_{\|\cdot\|_\infty}(\kappa, L)$ over a convex compact domain X of $\|\cdot\|_\infty$-diameter $2R$, the inaccuracy after $T = 1, 2, \ldots$ steps does not exceed

$$O(1)\frac{LR^{\kappa}}{T^{\kappa-1}}.$$

We see that when X is in between two $\|\cdot\|_\infty$-balls with ratio of sizes θ, the lower complexity bound coincides with the upper one within the factor $O(1)\theta^\kappa \ln^{\kappa-1}(n)$. In particular, *when minimizing functions $f \in \mathcal{F}_{\|\cdot\|_\infty}(\kappa, L)$ over n-dimensional unit box X, the performance of the Conditional Gradient algorithm, as expressed by its minimax risk, cannot be improved by more than $O(\ln^{\kappa-1}(n))$ factor, for any local oracle in use.* In fact, the same conclusion remains true when $\|\cdot\|_\infty$ and the unit box X are replaced with $\|\cdot\|_p$ and the unit $\|\cdot\|_p$-ball with "large" p, specifically, $p = O(1)\ln n$.

B. *Tightness:* In fact, in the case of $2 \le p < \infty$ the lower complexity bounds for smooth convex minimization over $\|\cdot\|_p$-balls established in Corollary 1, are tight: it is shown in [16], see also [11, Section 2.3] that a properly modified Nesterov's algorithm $\mathcal{N}$ for smooth convex optimization via the first-order oracle, as applied to problems of minimizing functions f from $\mathcal{F}_{\|\cdot\|_p}(\kappa, L)$ over the n-dimensional unit $\|\cdot\|_p$-ball X, for any number $T \ge 1$ of steps ensures that

$$f(x_T(\mathcal{N}, f)) - \min_{x\in X} f(x) \le C(p)\frac{L}{T^{\kappa+\frac{\kappa}{p}-1}},$$

with $C(p)$ depending solely on p, which is in full accordance with (15).

4.2. Smooth convex minimization over $\|\cdot\|_p$-balls, $1 \le p \le 2$

We have obtained lower complexity bounds for smooth convex minimization over $\|\cdot\|_p$-balls, where $2 \le p \le \infty$. Now we consider the case $1 \le p < 2$. We will build nearly tight bounds by reducing to the case of $p = \infty$.

Proposition 3. *Let $1 \le p \le 2$, $\kappa \in (1, 2]$, $L > 0$, and let $X \subset \mathbf{R}^n$ be a convex set containing the unit $\|\cdot\|_p$-ball. For properly selected absolute constant $\alpha \in (0, 1)$ and for every $T \le \alpha n$, the minimax risk of the family of problems $\mathcal{P}(\mathcal{F}, X)$ with $\mathcal{F} = \mathcal{F}_{\|\cdot\|_p}(\kappa, L)$ admits the lower bound*

$$\mathrm{Risk}_{\mathcal{F},X,\mathcal{O}}(T) \ge O(1)\left(\frac{L}{\ln^{\kappa-1}(T+1)T^{\frac{3\kappa}{2}-1}}\right), \tag{16}$$

independent of the local oracle $\mathcal{O}$ in use.

Proof. 1^0. By Dvoretzky's Theorem for the $\|\cdot\|_p$-ball [20, Theorem 4.15], there exists an absolute constant $\alpha \in (0, 1)$, such that for any $p \in [1, 2]$ and positive integers n, T satisfying $T \le \alpha n$ there is a subspace $M \subseteq \mathbf{R}^n$ of dimension T, and a centred at the origin ellipsoid $E \subseteq M$, such that

$$\frac{1}{2}E \subseteq B_M := \left\{x \in M : \|x\|_p \le 1\right\} \subseteq E. \tag{17}$$

Let $\{\gamma_i(\cdot) : i = 1, \ldots, T\}$ be linear forms on M such that $E = \{y \in M : \sum_{i=1}^T \gamma_i^2(y) \le 1\}$. By the second inclusion in (17), for every i, the maximum of the linear form $\gamma_i(\cdot)$ over B_M does not exceed 1, whence, by the Hahn–Banach Theorem, the form $\gamma_i(\cdot)$ can be extended from M to a linear form on the entire $\mathbf{R}^n$ to have the maximum over $B := \{x : \|x\|_p \le 1\}$ not exceeding 1. In other words, we can point out vectors $g_i \in \mathbf{R}^n$, $1 \le i \le T$ such that $\gamma_i(y) = \langle g_i, y\rangle$ for every $y \in M$ and $\|g_i\|_{\frac{p}{p-1}} \le 1$, for all $1 \le i \le T$. Now consider the linear mapping

$$x \mapsto Gx := [\langle g_1, x\rangle; \ldots; \langle g_T, x\rangle] : \mathbf{R}^n \to \mathbf{R}^T.$$

By the above, the operator norm of this mapping induced by the norms $\|\cdot\|_p$ on the argument and $\|\cdot\|_\infty$ on the image spaces does not exceed 1. As a result, when $f : \mathbf{R}^T \to \mathbf{R}$ belongs to $\mathcal{F}^T_{\|\cdot\|_\infty}(\kappa, L)$, the function $f_+ : \mathbf{R}^n \to \mathbf{R}$ defined by $f_+(x) = f(Gx)$, for $x \in \mathbf{R}^n$, belongs to the family $\mathcal{F}^n_{\|\cdot\|_p}(\kappa, L)$.[4] Setting $Y = GX$, we get a convex compact set in $\mathbf{R}^T$.

[4] To avoid abuse of notation, we have added to our usual notation $\mathcal{F}_{\|\cdot\|}(\cdot,\cdot)$ for families of smooth convex functions superscript indicating the argument dimension of the functions in question.

2^0. Observe that an optimization problem of the form

$$\min_{y\in Y} f(y) \qquad (P_{f,Y})$$

can be naturally reduced to the problem

$$\min_{x\in X} f_+(x), \qquad (P_{f_+,X})$$

and when the objective of the former problem belongs to $\mathcal{F}^T := \mathcal{F}^T_{\|\cdot\|_\infty}(\kappa, L)$, the objective of the latter problem belongs to $\mathcal{F}^n = \mathcal{F}^n_{\|\cdot\|_p}(\kappa, L)$. It is intuitively clear that the outlined reducibility implies that the complexity of solving problems from the family $\Phi_n := \{(P_{f,X}) : f \in \mathcal{F}^n\}$ cannot be smaller than the complexity of solving problems from the family $\Phi_T := \{(P_{f,Y}) : f \in \mathcal{F}^T\}$. Taking this claim for granted (for a proof, see Appendix C), let us derive from it the desired result. To this end, observe that from the first inclusion in (17) it follows that Y contains the centred at the origin $\|\cdot\|_\infty$-ball of radius $R = \frac{1}{2\sqrt{T}}$ (indeed, by construction this ball is already contained in the image of $\frac{1}{2}E \subset X$). By Corollary 1 as applied to $p = \infty$ and to Y in the role of X, the worst-case, w.r.t. problems from the family Φ_T, inaccuracy of any T-step method based on a local oracle is at least

$$\frac{O(1)}{\ln^{\kappa-1}(T+1)} \frac{R^\kappa L}{T^{\kappa-1}} = O(1)\left(\frac{1}{\ln^{\kappa-1}(T+1)} \frac{L}{T^{\frac{3\kappa}{2}-1}}\right), \qquad (18)$$

see (13). According to our claim, the latter quantity lower-bounds the worst-case, w.r.t. problems from the family Φ_n, inaccuracy of any T-step method based on a local oracle, and (16) follows. □

Finally, we remark that the lower complexity bound stated in Proposition 3 in the smooth case $\kappa > 1$ is, to the best of our knowledge, new (the nonsmooth case $\kappa = 1$ was considered already in [17]). This lower bound matches, up to logarithmic in n factors, the upper complexity bound for the family in question, see [11].

4.3. *Matrix case*

We have proved lower bounds for smooth optimization over $\|\cdot\|_p$-balls for all $1 \le p \le \infty$. Now we show how these bounds can be used for proving lower complexity bounds on smooth convex minimization over Schatten norm balls in the spaces of matrices. Recall that the Schatten p-norm $\|x\|_{\mathrm{Sch},p}$ of an $n \times n$ matrix x is, by definition the p-norm of the vector of singular values of x. The problems we are interested in now are of the form

$$\min_{x\in\mathbf{R}^{n\times n}} \{f(x) : \|x\|_{\mathrm{Sch},p} \le 1\}$$

where $f \in \mathcal{F}_{\|\cdot\|_{\mathrm{Sch},p}}(\kappa, L)$.

Observe that Corollary 1 remains true when replacing in it the embedding space $\mathbf{E} = \mathbf{R}^n$ of X with the space $\mathbf{E} = \mathbf{R}^{n\times n}$ of $n \times n$ matrices, the norm $\|\cdot\|_p$ on $\mathbf{R}^n$ with the Schatten norm $\|\cdot\|_{\mathrm{Sch},p}$, and the requirement "$X \subset \mathbf{R}^n$ is a convex set containing the unit ball of $\|\cdot\|_p$" with the requirement "$X \subset \mathbf{R}^{n\times n}$ is a convex set containing the unit ball of $\|\cdot\|_{\mathrm{Sch},p}$". This claim is an immediate consequence of the fact that when restricting an $n \times n$ matrix onto its diagonal, we get a linear mapping of $\mathbf{R}^{n\times n}$ onto $\mathbf{R}^n$, and the factor norm on $\mathbf{R}^n$ induced, via this mapping, by $\|\cdot\|_{\mathrm{Sch},p}$ is nothing but the usual $\|\cdot\|_p$-norm. Consequently, minimizing a function from $\mathcal{F}_{\|\cdot\|_p}(\kappa, L)$ over the unit $\|\cdot\|_p$ ball X of $\mathbf{R}^n$ reduces to minimizing a convex function of exactly the same smoothness, as measured w.r.t. $\|\cdot\|_{\mathrm{Sch},p}$, over the unit Schatten p-norm ball X^+ of $\mathbf{R}^{n\times n}$. As a result, every universal (i.e., valid for every local oracle) lower bound on the minimax risk for the problem class $\mathcal{P}(\mathcal{F}_{\|\cdot\|_p}(\kappa, L), X)$ automatically is a universal lower bound on the minimax risk for the problem class $\mathcal{P}(\mathcal{F}_{\|\cdot\|_{\mathrm{Sch},p}}(\kappa, L), X^+)$.

Note, however, that the "matrix extension" of our lower complexity bounds is not "completely costless"—the resulting bounds are applicable when $T \le n$ ($p \ge 2$) or $T = O(1)n$ ($1 \le p \le 2$), and n now is the square root of the actual dimension of x. Thus, in the matrix case our lower complexity bounds are applicable in relatively more narrow range of values of T.

Appendix A. Justification of (5)

In order to prove (5), by the standard approximation argument, it suffices to establish this relation in the case when, in addition to the inclusion $f \in \mathcal{C}_{\|\cdot\|}$ and the assumptions A–C on ϕ, f and ϕ are C^∞ smooth and ϕ is strongly convex. By (4),

$$\mathcal{S}[f](x) = f(x + h(x)) + \phi(h(x)), \tag{A.1}$$

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

$$F(x, h(x)) = 0, \qquad F(x, h) := f'(x + h) + \phi'(h). \tag{A.2}$$

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

$$\begin{aligned} &\underbrace{f''(x + h(x))}_{P}[I + h'(x)] + \underbrace{\phi''(h(x))}_{Q} h'(x) = 0 \\ &\Leftrightarrow P + (P + Q)h'(x) = 0 \\ &\Rightarrow h'(x) = -[P + Q]^{-1}P = [P + Q]^{-1}Q - I. \end{aligned}$$

On the other hand, differentiating (A.1), 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,

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

and (5) follows.

Appendix B. Proof for Section 2.1

We have:

$$\begin{aligned} \langle e, [\nabla^2 \phi(x)]e\rangle &= 4r\theta(2\theta/r - 1)\left(\sum_j |x_j|^r\right)^{2\theta/r-2}\left[\sum_j |x_j|^{r-1}\text{sign}(x_j)e_j\right]^2 \\ &\quad + 4\theta(r - 1)\left(\sum_j |x_j|^r\right)^{2\theta/r-1}\sum_j |x_j|^{r-2}e_j^2 \\ &\le 4\theta(r - 1)\left(\sum_j |x_j|^r\right)^{2\theta/r-1}\sum_j |x_j|^{r-2}e_j^2 \qquad \text{(B.1)} \\ &\le 4\theta(r - 1)\left[\|x\|_p^r\, n^{1-r/p}\right]^{2\theta/r-1}\left[\sum_j |x_j|^{\frac{(r-2)p}{p-2}}\right]^{\frac{p-2}{p}}\left[\sum_j |e_j|^p\right]^{\frac{2}{p}} \qquad \text{(B.2)} \end{aligned}$$

$$\leq 4\theta(r-1)\left[\|x\|_p^r n^{1-r/p}\right]^{2\theta/r-1}\left[\|x\|_p^{\frac{(r-2)p}{p-2}} n^{1-\frac{r-2}{p-2}}\right]^{1-2/p}\|e\|_p^2 \qquad (B.3)$$

$$\leq 4\theta(r-1)\|x\|_p^{2\theta-2} n^{\frac{2\theta(p-r)}{pr}}\|e\|_p^2,$$

Note we used that $2\theta/r < 1$ in (B.1), 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 (B.2), (B.3), and the Hölder inequality in (B.2).

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

$$x \in G \Rightarrow \langle e, [\nabla^2\phi(x)]e\rangle \leq O(1)\min[p, \ln n]\|e\|_p^2 \quad \forall e \in \mathbf{R}^n, \qquad (B.4)$$

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

Appendix C. Item 2^0 of the proof of Proposition 3

Observe, first, that the claim we intend to justify indeed needs a justification: we cannot just argue that solving "lifted" problems – those from the family $\Phi_n^+ = \{(P_{f^+,X}) : f \in \mathcal{F}^T_{\|\cdot\|_\infty}(\kappa, L)\} \subset \Phi_n$ – cannot be simpler than solving problems from Φ_T 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 just because the oracle for the "lifted" family is more informative than the oracle for the original family. The justification here is as follows: observe that among local oracles for families of real-valued functions on $\mathbf{R}^m$ there is the "most informative" one, let us call it *maximal*; when queried about a function f at a point x, the maximal oracle returns the class $\widetilde{f}$ of f w.r.t. the equivalence relation "f is equivalent to g if and only if f and g coincide with each other in some (perhaps depending on f and g) neighbourhood of x". Clearly the maximal oracle allows to mimic any other local oracle, so that for every family of problems, the lower complexity bounds valid for the maximal oracle are valid for any other local oracle. Now, it is easily seen that the maximal oracle for the family of functions $\mathcal{F}^T$ induces the maximal oracle for the lifted family $\{f^+ : f \in \mathcal{F}^T\}$; with this in mind, it is immediately seen that any maximal-oracle-based T-step method $\mathcal{M}^+$ for solving problems from the family Φ_n^+ induces a maximal-oracle-based T-step method $\mathcal{M}$ for solving problems from the family Φ^T in such a way that the trajectory $x_1, x_2, \ldots$ of $\mathcal{M}$ on a problem $(P_{f,Y})$ is linked to the trajectory $x_1^+, x_2^+, \ldots$ of $\mathcal{M}^+$ on $(P_{f_+,X})$ by the relation $x_t = Gx_t^+$. Consequently, when the maximal oracles are used, a lower bound on the T-step minimax risk of Φ_T automatically is a lower bound on the same quantity for the family $\Phi_n^+ = \{(P_{f_+,X}) : f \in \mathcal{F}^T\}$, and therefore for the larger family Φ_n. In particular, the quantity (18), which by Corollary 1 lower-bounds the maximal-oracle-based T-step minimax risk when solving problems from Φ_T, lower-bounds the similar quantity for Φ_n, and thus—the T-step minimax risk of Φ_n taken w.r.t. any local oracle, as claimed.

References

[1] K.L. Clarkson, Coresets, sparse greedy approximation and the Frank–Wolfe algorithm, in: SODA 2008, 2008, pp. 922–931.
[2] B. Cox, A. Juditsky, A. Nemirovski, Dual subgradient algorithms for large-scale nonsmooth learning problems, Math. Program. Ser. B (2013) 1–38. http://dx.doi.org/10.1007/s10107-013-0725-1.
[3] V. Demyanov, A. Rubinov, Approximate Methods in Optimization Problems, American Elsevier, 1970.
[4] M. Frank, P. Wolfe, An algorithm for quadratic programming, Nav. Res. Logist. Q. 3 (1956) 95–110.
[5] D. Garber, E. Hazan, Playing non-linear games with linear oracles, in: 2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2013, pp. 420–428.
[6] Z. Harchaoui, A. Juditsky, A. Nemirovski, Conditional gradient algorithms for norm-regularized smooth convex optimization, Math. Program. Ser. A (2014) 1–38. http://dx.doi.org/10.1007/s10107-014-0778-9.
[7] E. Hazan, Sparse approximate solutions to semidefinite programs, in: LATIN 2008, 2008, pp. 306–316.
[8] J.-B. Hiriart-Urruty, C. Lemaréchal, Fundamentals of Convex Analysis, Springer Verlag, Heidelberg, 2001.
[9] M. Jaggi, Sparse convex optimization methods for machine learning (Ph.D. thesis), ETH Zurich, October 2011, http://dx.doi.org/10.3929/ethz-a-007050453.
[10] M. Jaggi, Revisiting Frank–Wolfe: projection-free sparse convex optimization, in: ICML 2013, Vol. 28, 2013, pp. 427–435.
[11] L. Khachiyan, A. Nemirovski, Y. Nesterov, Optimal methods for the solution of large-scale convex programming problems, in: K.-H. Elster (Ed.), Modern Mathematical Methods in Optimization, Academie Verlag, Berlin, 1993.

[12] G. Lan, The complexity of large-scale convex programming under a linear optimization oracle, 2013. arXiv:1309.5550.
[13] A. Nemirovski, On optimality of Krylov's information when solving linear operator equations, J. Complexity 7 (2) (1991) 121–130.
[14] A. Nemirovski, Information-based complexity of linear operator equations, J. Complexity 8 (2) (1992) 153–175.
[15] A. Nemirovski, Efficient methods in convex programming, 1994. http://www2.isye.gatech.edu/~nemirovs/.
[16] A. Nemirovskii, Y. Nesterov, Optimal methods of smooth convex optimization, Zh. Vychisl. Mat. Mat. Fiz. 25 (3) (1985) 356–369. (in Russian).
[17] A. Nemirovski, D. Yudin, Problem Complexity and Method Efficiency in Optimization, first ed., Wiley-Interscience, 1983.
[18] Y. Nesterov, Introductory Lectures on Convex Optimization: A Basic Course, first ed., Springer, Netherlands, 2004.
[19] Y. Nesterov, Universal gradient methods for convex optimization problems, Math. Program. Ser. A (2014) 1–24. http://dx.doi.org/10.1007/s10107-014-0790-0.
[20] G. Pisier, The Volume of Convex Bodies and Banach Space Geometry, first ed., Cambridge University Press, 1989.
[21] B. Pshenichnyj, Y. Danilin, Numerical Methods in Extremal Problems, Mir, 1978.