# On Lower Complexity Bounds for Large-Scale Smooth Convex Optimization[*]

Cristóbal Guzmán[†] and Arkadi Nemirovski[‡]

H. Milton Stewart School of Industrial and Systems Engineering,

Georgia Institute of Technology, Atlanta, GA, USA.

**Abstract**

In this note we present tight lower bounds on the information-based complexity of large-scale smooth convex minimization problems. We demonstrate, in particular, that the $k$-step Conditional Gradient (a.k.a. Frank-Wolfe) algorithm as applied to minimizing smooth convex functions over the $n$-dimensional box with $n \geq k$ is optimal, up to an $O(\ln n)$-factor, in terms of information-based complexity.

## 1 Introduction

In this note, we deal with information-based complexity of convex optimization problems, and we start with outlining the associated notions. We are interested in solving optimization programs 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 Euclidean space $E$, and $f$ is known to belong to a given family $\mathcal{F}$ of continuous convex functions on $E$. 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 $x \in E$ taking values in some *information space* $\mathcal{I}$; when solving $(P_{f,X})$, an algorithm at every step can input to the oracle a query point $x \in E$ and gets back the value $\mathcal{O}(f,x)$. In the sequel, we always assume the oracle to be *local*, meaning that whenever $x \in E$ and $f, g \in \mathcal{F}$ are such that $f(\cdot) = g(\cdot)$ in some neighbourhood of $x$, we have $\mathcal{O}(f,x) = \mathcal{O}(g,x)$.

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

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

where the *search rules* $X_t(\cdot)$ are deterministic functions of their arguments; we can identify $\mathcal{M}$ as a collection of these rules. Thus, $x_1$ is defined 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 $x_k = x_k(\mathcal{M}, f)$ as the approximate solution generated by $k$-step solution method $\mathcal{M}$ as 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 $k = 1, 2, ...$ defined by

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

where the right hand side infinum is taken over all $k$-step solution algorithms $\mathcal{M}$ utilizing oracle $\mathcal{O}$. The inverse to the risk function

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

[*] The research was supported by the NSF grant CMMI-1232623.

[†] cguzman@gatech.edu

[‡] arkadi.nemirovski@isye.gatech.edu

for $\varepsilon > 0$ is called the *information-based complexity of the family* $\mathcal{F}_X$ taken with respect to oracle $\mathcal{O}$. The standard reference on information-based complexity of various broad families of convex programs is [4]; for some recent developments, see [8, 1] and references therein.

This note is primarily motivated by recently renewed interest in the *Conditional Gradient* (a.k.a. Frank-Wolfe) method, originating from [3], for solving problems $(P_{f,X})$ with smooth convex objectives $f$. This method utilizes the standard first order oracle ($\mathcal{O}(f,x) = (f(x), \nabla f(x))$ and possesses two remarkable features:

1. *Dimension-independent sublinear convergence rate* depending solely on $t$ and of the properly measured smoothness parameters of $f$. Specifically, assuming w.l.o.g. that $X$ linearly spans $E$, the set $\frac{1}{2}[X - X]$ is the unit ball of certain, depending solely on $X$, norm on $E$; we denote this norm $\|\cdot\|_X$, and its conjugate norm by $\|\cdot\|_{X,*}$. Assuming that the objective $f$ in $(P_{f,X})$ is convex and $(\kappa, L)$ *smooth on* $X$, meaning that $f$ has Hölder continuous, with parameters $\kappa \in (1,2], L < \infty$, gradient on $X$:

$$\|\nabla f(x) - \nabla f(y)\|_{X,*} \leq L\|x - y\|_X^{\kappa-1} \ \forall x, y \in X, \tag{1}$$

   the conditional algorithm with $t = 1, 2, \ldots$ steps ensures that

$$f(x_t) - \mathrm{Opt}(f) \leq 8L/t^{\kappa-1}, \ t = 2, 3, \ldots \tag{2}$$

2. *Possibility to work with "complex geometry" sets* $X$. A nice fact about Conditional Gradient methods (which is exactly the reason of the renewed interest in them, primarily coming from nuclear norm minimization arising in Machine Learning) is that *these methods are applicable when* $X$ *is represented by a not too computationally expensive Linear Minimization oracle* (one capable of minimizing linear objectives over $X$), which is essentially less restrictive than the ability, required by the majority of other first order algorithms, to deal with much more computationally demanding *prox mappings*.

In spite of the just outlined attractive features of Conditional Gradient methods, these algorithms usually are *not* optimal from the viewpoint of information-based complexity theory. For example when $X$ is the unit $n$-dimensional Euclidean ball and $f$ has Lipschitz continuous gradient (i.e., (1) holds true, with some $L$, for $\kappa = 2$, (2) exhibits the rate of convergence $O(1)L/t$, while the "true" optimal dimension-independent rate of convergence in this case is $O(1)L/t^2$ [1]. The only interesting situation where the existing knowledge did *not* predict non-optimality of Conditional Gradient algorithm in terms of information-based complexity, was the one where $X$ is a high-dimensional unit box. This paper was primarily motivated by the desire to understand what is the information-based complexity of smooth convex minimization over the $n$-dimensional unit box $X$. The main contribution of this note is in demonstrating that *when minimizing convex functions with Lipschitz continuous, with constant* $L$ *w.r.t.* $\|\cdot\|_X = \|\cdot\|_\infty$, *gradient over* $n$*-dimensional unit box* $X$, *the risk, for all* $t \leq n$, *is lower-bounded by* $O(1)/(t \ln n)$, *so that the Conditional Gradient algorithm is in this case nearly* (up to an $O(\ln n)$ factor) *optimal in terms of information-based complexity.*

In fact, in what follows we provide tight lower bounds on information-based complexity if minimizing $(\kappa, L)$-smooth functions on $\|\cdot\|_p$-balls, where $2 \leq p \leq \infty$, so that the just mentioned lower complexity bound for smooth convex minimization over the unit box is the special case $p = \infty$ of the bounds to be presented. It should be mentioned that for the case of $p < \infty$ these bounds, obtained by the second author of this note, were announced in [5, 2]; however, aside of the very special case of $p = 2$, the highly technical original proofs of the bounds were never published. Motivated as explained above, we recently have revisited the original proofs and were able to simplify them dramatically, thus making them publishable.

The rest of this note is organized as follows. In section 2 we restate in full details the problem we are interested in. Section 3 is devoted to the main component of our construction – a Moreau-type scheme for approximating a convex Lipschitz continuous function by a convex function with Lipschitz continuous gradients; 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]. Section 4 contains the main result of this note – a lower bound on information-based complexity of smooth convex minimization.

[1] For the $O(1)L/t^2$ lower risk bound, see [4]; an $O(1)L/t^2$ upper risk bound in the case in question is achieved by the celebrated *Nesterov's optimal algorithm for smooth convex minimization* [6, 7].

[2] It well may happen that the extensions of the classical Moreau results which we present in section 3 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.

# 2 The problem

Let $E$ be an $n$-dimensional Euclidean space, and $\|\cdot\|$ be a norm on $E$ (not necessary the Euclidean one). Let, further, $X$ be a nonempty closed and bounded convex set in $E$. Given a positive real $L$ and $\kappa \in (1,2]$, consider the family $\mathcal{F}_{\|\cdot\|}(\kappa, L)$ of all continuously differentiable convex functions $f : E \to \mathbf{R}$ satisfying the inequality

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|^{\kappa-1} \quad \forall x, y \in E, \tag{3}$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$. We associate with $\|\cdot\|, X, \kappa, L$ the family of convex optimization problems $\mathcal{P} = \mathcal{P}(\mathcal{F}_{\|\cdot\|}(\kappa, L), X)$; recall that this is the family of all problems of the form $\mathrm{Opt}(f) = \min_{x\in X} f(x)$ with $f \in \mathcal{F}_{\|\cdot\|}(\kappa, L)$.

We equip the family $\mathcal{F}_{\|\cdot\|}(\kappa, L)$ 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* $n \geq t$, and the reason is as follows: it is known [4] that when $t \gg n$, $\mathrm{Risk}(t)$ "basically forgets the details specifying $\mathcal{P}$" and goes to 0, as $t \to \infty$, as $O(\exp\{-O(1)t/n\})$; the data $\|\cdot\|, X, L, \kappa$ participating in the description of $\mathcal{P}$ affect only the factor hidden in the outer $O(\cdot)$ and thus become irrelevant when $t \gg n$. In contrast to this, the risk $\mathrm{Risk}(t)$ in the "large-scale regime" $t \leq n$ is (at least in the cases we are about to consider) nearly independent of $n$ and just *sublinearly* decreases as $t$ grows, and its behavior in this range heavily depends on $\mathcal{P}$.

# 3 Local Smoothing

In this section we introduce the main tool 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 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.

## 3.1 Smoothing Functional

Let $\|\cdot\|$ be a norm on a Euclidean space $E$, and $\mathcal{C}_{\|\cdot\|}$ be the set of all Lipschitz continuous, with constant 1 w.r.t. $\|\cdot\|$, convex functions on $E$. Let also $\phi(h)$ be a twice continuously differentiable convex function defined on an open convex set $\mathrm{Dom}\,\phi \subset E$ 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 $x \in \partial G$.
Note that A and B imply that whenever $f \in \mathcal{C}_{\|\cdot\|}$, the function $f(x) + \phi(x)$ attains its minimum on the set $\mathrm{int}\,G$. Indeed, for every $x \in \partial G$ 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 $\mathrm{int}\,G$. As a result, for every $f \in \mathcal{C}_{\|\cdot\|}$ and $x \in \mathbf{R}^n$ one has
$$\min_{h\in\mathrm{Dom}\phi}[f(x+h) + \phi(h)] = \min_{h\in\mathrm{int}G}[f(x+h) + \phi(h)], \tag{4}$$
and the right hand side minimum is achieved.

C. For some $M_\phi < \infty$ we have
$$e^T[\nabla^2\phi(h)]e \leq M_\phi\|e\|^2 \quad \forall(e \in E, h \in G). \tag{5}$$

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

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

as to the *smoothing* of $f$. Observe that by our assumptions on $\phi$ we have

1. $\mathcal{S}[f](x) = f(x + h(x)) + \phi(h(x))$, where $h(x) \in \mathrm{int}\,G$ is such that
$$f'(x + h(x)) + \phi'(h(x)) = 0 \tag{6}$$
for properly selected $f'(x + h(x)) \in \partial f(x + h(x))$;

2. $f(x) \geq \mathcal{S}[f](x) \geq f(x) - R_{\|\cdot\|}(G)$, where
$$R_{\|\cdot\|}(G) = \max_{h\in G} \|h\|;$$
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
$$\|\nabla\mathcal{S}[f](x) - \nabla\mathcal{S}[f](y)\|_* \leq M_\phi\|x - y\| \quad \forall(x, y \in E, f \in \mathcal{C}). \tag{7}$$

To prove the last claim, by the standard approximation argument, it suffices to consider the case when, in addition to the inclusion $f \in \mathcal{C}_{\|\cdot\|}$ and our previous assumptions A – C on $\phi$, $f$ and $\phi$ are $\mathrm{C}^\infty$ smooth and $\phi$ is strongly convex. As we remember,
$$\mathcal{S}[f](x) = f(x + h(x)) + \phi(h(x)), \tag{8}$$
where $h(x) : E \to G$ is well defined and solves the nonlinear system of equations
$$F(x, h(x)) = 0,\ F(x, h) := f'(x + h) + \phi'(h) = 0. \tag{9}$$
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 (9), we get
$$\begin{array}{l}\underbrace{f''(x + h(x))}_{P}[I + h'(x)] + \underbrace{\phi''(h(x))}_{Q} h'(x) = 0 \quad \Leftrightarrow \quad P + (P + Q)h'(x) = 0\\ \Rightarrow h'(x) = -[P + Q]^{-1}P = -[P + Q]^{-1}[P + Q - Q] = [P + Q]^{-1}Q - I.\end{array}$$
On the other hand, differentiating (8), we get
$$\begin{array}{l}\langle\nabla\mathcal{S}[f](x), e\rangle = \langle f'(x + h(x)), e + h'(x)e\rangle + \langle\phi'(h(x)), h'(x)e\rangle\\ = \langle f'(x + h(x)), e\rangle + \langle\underbrace{f'(x + h(x)) + \phi'(h(x))}_{=0}, h'(x)e\rangle = -\langle\phi'(h(x)), e\rangle\end{array}$$
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,
$$e^T\nabla^2\mathcal{S}[f](x)e = -[h'(x)e]^T\phi''(h(x))e = -e^T[[P+Q]^{-1}Q-I]^TQe = e^TQe-e^TQ[P+Q]^{-1}Qe \leq e^TQe \leq M_\phi\|e\|^2,$$
and (7) follows.

## 3.2 Approximating a function by smoothing

For $\chi > 0$ and $f \in \mathcal{C}_{\|\cdot\|}$, let
$$\mathcal{S}_\chi[f](x) = \min_{h\in\chi\mathrm{Dom}\,\phi}[f(x) + \chi\phi(h/\chi)].$$
Observe that $\mathcal{S}[f]_\chi(\cdot)$ can be obtained as follows:

- We associate with $f \in \mathcal{C}_{\|\cdot\|}$ the function $F(x) = \chi^{-1}f(\chi x)$; observe that this function belongs to $\mathcal{C}_{\|\cdot\|}$ along with $f$;
- We pass from $F$ to its smoothing
$$\begin{array}{l}\mathcal{S}[F](x) = \min_{g\in\mathrm{Dom}\,\phi}\left[F(x + g) + \phi(g)\right] = \min_{g\in\mathrm{Dom}\,\phi}\left[\chi^{-1}f(\chi x + \chi g) + \phi(g)\right]\\ = \chi^{-1}\min_{h\in\chi\mathrm{Dom}\,\phi}\left[f(\chi x + h) + \chi\phi(h/\chi)\right]\\ = \chi^{-1}\mathcal{S}_\chi[f](\chi x).\end{array}$$

It follows that

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

The latter relation combines with (7) to imply that

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

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

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

S.2. $\sup_{x\in E} |f(x) - \mathcal{S}_\chi[f](x)| \leq \chi R_{\|\cdot\|}(G)$. Moreover, $f(x) \geq \mathcal{S}_\chi[f](x) \geq f(x) - \chi R_{\|\cdot\|}(G)$.

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

### 3.3 Example: $p$-norm Smoothing

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

$$\begin{array}{rcl} \phi(x) &=& \phi_{r,\theta}(x) = 2\left(\sum_{j=1}^n |x_j|^r\right)^{2\theta/r}, \\ G &=& \{x \in \mathbf{R}^n : \|x\|_p \leq 1\}. \end{array} \tag{10}$$

Observe that $\phi$ is twice continuously differentiable on $\operatorname{Dom}\phi = \mathbf{R}^n$ function satisfying A. Besides this, $r \leq p$ ensures that $\sum_j |x_j|^r \geq 1$ whenever $\|x\|_p = 1$, so that $\phi(x) > \|x\|_p$ when $x \in \partial G$, which implies B. Besides this, we have

$$\begin{array}{rcll} e^T[\nabla^2\phi(x)]e &=& 4r\theta(2\theta/r - 1)(\sum_j |x_j|^r)^{2\theta/r-2}\left[\sum_j |x_j|^{r-1}\text{sign}(x_j)e_j\right]^2 & \\ && +4\theta(r-1)(\sum_j |x_j|^r)^{2\theta/r-1}\sum_j |x_j|^{r-2}e_j^2 & \\ &\leq& 4\theta(r-1)(\sum_j |x_j|^r)^{2\theta/r-1}\sum_j |x_j|^{r-2}e_j^2 & (11) \\ &\leq& 4\theta(r-1)\left[\|x\|_p^r\, n^{1-r/p}\right]^{2\theta/r-1}\left[\sum_j |x_j|^{\frac{(r-2)p}{p-2}}\right]^{\frac{p-2}{p}}\left[\sum_j |e_j|^p\right]^{\frac{2}{p}} & (12) \\ &\leq& 4\theta(r-1)\left[\|x\|_p^r\, n^{1-r/p}\right]^{2\theta/r-1}\left[\|x\|_p^{\frac{(r-2)p}{p-2}} n^{1-\frac{r-2}{p-2}}\right]^{1-2/p}\|e\|_p^2 & (13) \\ &\leq& 4\theta(r-1)\|x\|_p^{2\theta-2}\, n^{\frac{2\theta(p-r)}{pr}}\|e\|_p^2, & \end{array}$$

Note we used that $2\theta/r < 1$ in (11), 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 (12), (13), and the Hölder inequality in (12).

We see that setting $r = \min[p, 3\ln n]$ 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

$$x \in G \quad \Rightarrow \quad e^T[\nabla^2\phi(x)]e \leq O(1)\min[p, \ln n]\|e\|_p^2 \;\; \forall e \in \mathbf{R}^n, \tag{14}$$

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

Up to now, we have assumed that $p > 2$. In the case of $p = 2$, 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 $M_\phi = 1$.

Applying the results of the previous section, we get

**Proposition 3.1.** *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) *One has $f(x) \geq \mathcal{S}_\chi[f](x) \geq f(x) - \chi$ for all $x$.*

(ii) $\|\nabla \mathcal{S}_\chi[f](x) - \nabla \mathcal{S}_\chi[f](y)\|_{\frac{p}{p-1}} \leq O(1)\min[p, \ln n]\chi^{-1}\|x-y\|_p$ *for all $x, y$, with absolute constant $O(1)$;*

(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*

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

# 4 Lower Bounds for the Information-Based Complexity of Smooth Convex Minimization

In this section we utilize Proposition 3.1 to prove our main result, namely, a general lower bound on the information-based complexity of smooth convex minimization, and then specify this result for the case of minimization over $\|\cdot\|_p$ balls, $2 \leq p \leq \infty$.

**Proposition 4.1.** *Let*

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

II. *$k$ be a positive integer and $\Delta$ be a positive real with the following property:*
*One can point out $k$ linear forms $\langle\omega_i, \cdot\rangle$ on $\mathbf{R}^n$, $1 \leq i \leq k$, such that*

*(a) $\|\omega_i\|_* \leq 1$ for $i \leq k$, and*

*(b) for every collection $\xi^k = (\xi_1, ..., \xi_k)$ with $\xi_i \in \{1; -1\}$, it holds*

$$\min_{x \in X} \max_{1 \leq i \leq k} \xi_i \langle\omega_i, x\rangle \leq -\Delta; \tag{15}$$

III. *$M$ and $R$ be positive reals such that for properly selected convex twice continuously differentiable on an open convex set $\operatorname{Dom}\phi \subset \mathbf{R}^n$ function $\phi$ and a convex compact subset $G \subset \operatorname{Dom}\phi$ the triple $(\phi, G, M_\phi = M)$ satisfies properties A, B, C from section 3.1 and $R_{\|\cdot\|}(G) \leq R$.*

*Then for every $L > 0$, $\kappa \in (1, 2]$, every local oracle $\mathcal{O}$ and every $k$-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*

$$f(x_k(\mathcal{M}, f)) - \operatorname{Opt}(f) \geq \frac{\Delta^\kappa}{2^{\kappa+1}(RM)^{\kappa-1}} \cdot \frac{L}{k^{\kappa-1}}. \tag{16}$$

**Proof.** $1^0$. Let us set

$$\delta = \frac{\Delta}{2k}, \ \chi = \frac{\delta}{2R} = \frac{\Delta}{4kR}, \ \beta = \frac{L\chi^{\kappa-1}}{2^{2-\kappa}M^{\kappa-1}} = \frac{L\Delta^{\kappa-1}}{2^\kappa(kRM)^{\kappa-1}}. \tag{17}$$

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

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

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

$$f^{\sigma(\cdot),\xi^k}(x) = \beta \mathcal{S}_\chi[g^{\sigma(\cdot),\xi^k}](x) \tag{18}$$

(see section 3.2) are well defined continuously differentiable convex functions on $\mathbf{R}^n$ which, by item S.1 in section 3.2, satisfy the relation

$$\|\nabla f^{\sigma(\cdot),\xi^k}(x) - \nabla f^{\sigma(\cdot),\xi^k}(y)\|_* \leq \beta \min[\chi^{-1}M\|x-y\|, 2] \ \forall x, y,$$

whence, as it is immediately seen, $\|\nabla f^{\sigma(\cdot),\xi^k}(x) - \nabla f^{\sigma(\cdot),\xi^k}(y)\|_* \leq \beta 2^{2-\kappa}(\chi^{-1}M)^{\kappa-1}\|x-y\|^{\kappa-1}$ for all $x, y$. Recalling the definition of $\beta$, we conclude that $f^{\sigma(\cdot),\xi^k}(\cdot) \in \mathcal{F}_{\|\cdot\|}(\kappa, L)$.

$3^0$. Given a local oracle $\mathcal{O}$ and associated $k$-step method $\mathcal{M}$, let us define a sequence $x_1, ..., x_k \in \mathbf{R}^n$, a permutation $\sigma(\cdot)$ of $\{1, ..., k\}$ and a collection $\xi^k \in \{-1; 1\}^k$ by the following $k$-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 of the largest in magnitude of the quantities $\langle\omega_i, x_1\rangle$ 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
$$g^1(x) = \xi_1\langle\omega_{\sigma(1)}, x\rangle,\ f^1(x) = \beta\mathcal{S}_\chi[g^1](x).$$
- *Step $t$, $2 \leq t \leq k$:* 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 $\sigma(\tau) \in \{1, ..., k\}$ and quantities $\xi_\tau \in \{-1;1\}$, $1 \leq \tau < t$. At step $t$, we build $x_t$, $\sigma(t)$, $\xi_t$, namely, as follows. We set
$$g^{t-1}(x) = \max_{1\leq\tau<t}\left[\xi_\tau\langle\omega_{\sigma(\tau)}, x\rangle - (\tau-1)\delta\right],$$
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 $\mathcal{F}_{\|\cdot\|}(2, L)$. 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 of the largest in magnitude of the quantities $\langle\omega_i, x_t\rangle$ with $i$ varying in $\{1, ..., k\}$ *and distinct from* $\sigma(1), ..., \sigma(t-1)$,
  - $\xi_t \in \{-1;1\}$ such that $\xi_t\langle\omega_{\sigma(t)}, x_t\rangle = |\langle\omega_{\sigma(t)}, x_t\rangle|$

  thus completing step $t$.

After $k$ steps of this recurrence, we get at our disposal a sequence $x_1, ..., x_k$ of points form $\mathbf{R}^n$, a permutation $\sigma(\cdot)$ of indexes $1, ..., k$ and a collection $\xi^k = (\xi_1, ..., \xi_k) \in \{-1;1\}^n$; these entities define the functions
$$g^k = g^{\sigma(\cdot),\xi^k},\ f^k = \beta\mathcal{S}_\chi[g^{\sigma(\cdot),\xi^k}].$$

$4^0$. We claim that $x_1, ..., x_k$ is the trajectory of $\mathcal{M}$ as applied to $f^k$. By construction, $x_1$ indeed is the first point of the trajectory of $\mathcal{M}$ as applied to $f^k$. 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 $t$, $2 \leq t \leq k$, the functions $f^k$ and $f^{t-1}$ coincide in some neighbourhood of $x_{t-1}$. By construction, we have
$$t \leq s \leq k \quad\Rightarrow\quad \xi_s\langle\omega_{\sigma(s)}, x_{t-1})\rangle \leq |\langle\omega_{\sigma(t-1)}, x_{t-1}\rangle| = \xi_{t-1}\langle\omega_{\sigma(t-1)}, x_{t-1}\rangle, \tag{19}$$
$$g^k(x) = \max\left[g^{t-1}(x), \underbrace{\max_{t\leq s\leq k}\left[\xi_s\langle\omega_\sigma(s), x\rangle - (s-1)\delta\right]}_{g_t(x)}\right] \tag{20}$$
and
$$g^{t-1}(x_{t-1}) \geq \xi_{t-1}\langle\omega_{\sigma(t-1)}, x_{t-1}\rangle - (t-2)\delta.$$
Invoking (19), we get
$$\begin{array}{rl} t \leq s \leq k & \Rightarrow\ g^{t-1}(x_{t-1}) \geq [\xi_s\langle\omega_\sigma(s), x\rangle - (s-1)\delta] + \delta \\ & \Rightarrow\ g^{t-1}(x_{t-1}) \geq g_t(x_{t-1}) + \delta. \end{array}$$
Since both $g^{t-1}$ and $g_t$ belong to $\mathcal{C}_{\|\cdot\|}$, it follows that $g^{t-1}(x) \geq g_t(x)$ in the $\|\cdot\|$-ball $B$ of radius $\delta/2$ centered at $x_{t-1}$, whence, by (20),
$$x \in B \quad\Rightarrow\quad g^k(x) = g^{t-1}(x).$$
Since $\chi R = \delta/2$, it follows that $g^{t-1} \in \mathcal{C}_{\|\cdot\|}$ and $g^k \in \mathcal{C}_{\|\cdot\|}$ coincide on the set $x_{t-1} + \chi G$, whence, as we know from item S.3 in section 3.2, $f^{t-1}(\cdot) = \beta\mathcal{S}_\chi[g^{t-1}](\cdot)$ and $f^k(\cdot) = \beta\mathcal{S}_\chi[g^k](\cdot)$ coincide in a neighbourhood of $x_{t-1}$, as claimed.

$5^0$. We have $g^k(x_k) \geq \xi_k\langle\omega_{\sigma(k)}, x_k\rangle - (k-1)\delta = |\langle\omega_{\sigma(k)}, x_k\rangle| - (k-1)\delta \geq -(k-1)\delta$, whence, by item S.2 in section 3.2, $\mathcal{S}_\chi[g^k](x_k) \geq -(k-1)\delta - \chi R \geq -k\delta = -\Delta/2$, implying that
$$f^k(x_k) \geq -\beta\Delta/2.$$
On the other hand, by (15) there exists $x_* \in X$ such that $g^k(x_*) \leq \max_{1\leq i\leq k}\xi_i\langle\omega_{\sigma(i)}, x_*\rangle \leq -\Delta$, whence $\mathcal{S}_\chi[g^k](x_*) \leq g^k(x_*) \leq -\Delta$ and thus $\mathrm{Opt}(f^k) \leq f^k(x_*) \leq -\beta\Delta$. Since, as we have seen, $x_1, ..., x_k$ is the trajectory of $\mathcal{M}$ as applied to $f^k$, $x_k$ is the approximate solution generated by $\mathcal{M}$ as applied to $f^k \in \mathcal{F}(L)$, and we see that the inaccuracy of this solution, in terms of the objective, is at least $\frac{\beta\Delta}{2} = \frac{\Delta^\kappa}{2^{\kappa+1}(RM)^{\kappa-1}} \cdot \frac{L}{k^{\kappa-1}}$, as required. Besides this, $f^k$ is of the form $f^{\sigma(\cdot),\xi^k}$, and we have seen that all these functions belong to $\mathcal{F}_{\|\cdot\|}(\kappa, L)$. □

### 4.1 Illustration: Smooth Convex Minimization over $\|\cdot\|_p$-Balls

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

$$\min_{\|x\|_p \leq 1} \max_{1 \leq i \leq k} \xi_i \langle \omega_i, x \rangle \leq -k^{-1/p}. \tag{21}$$

Invoking the results from section 3.3 (cf. Proposition 3.1), we see that *whenever $X \subset \mathbf{R}^n$ is a convex set containing the unit ball of the norm $\|\cdot\|_p$, assumptions II and III in Proposition 4.1 are satisfied with $M = O(1)\min[p, \ln n]$, $R = 1$ and $\Delta = k^{-1/p}$.* Applying Proposition 4.1, we arrive at

**Corollary 4.1.** *Let $p \in [2, \infty]$, $\kappa \in (1, 2]$, let $L > 0$, and let $X \subset \mathbf{R}^n$ be a convex set containing the unit ball of $\|\cdot\|_p$. Then, for every $k \leq n$, every local oracle $\mathcal{O}$, and $k$-step method $\mathcal{M}$ associated with the oracle, there exists an objective $f \in \mathcal{F}_{\|\cdot\|_p}(\kappa, L)$ such that for the approximate solution $x_k$ generated by $\mathcal{M}$ as applied to the problem*

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

*one has*

$$f(x_k) - \mathrm{Opt}(f) \geq \frac{O(1)}{\min^{\kappa-1}[p, \ln n]} \frac{L}{k^{\kappa + \frac{\kappa}{p} - 1}}. \tag{22}$$

*In other words, in the situation under consideration the minimax risk of the family of problems $(P_{f,X})$ with $f \in \mathcal{F}_{\|\cdot\|_p}(\kappa, L)$ admits the lower bound*

$$\mathrm{Risk}(k) := \mathrm{Risk}_{\mathcal{F}_{\|\cdot\|_p}(\kappa,L),\mathcal{O}}(k) \geq \frac{O(1)}{[\min[p, \ln n]]^{\kappa-1}} \frac{L}{k^{\kappa + \frac{\kappa}{p} - 1}}, \tag{23}$$

*whatever local oracle $\mathcal{O}$ be used.*

**Discussion.** Some comments are in order.

**A.** Corollary 4.1 implies that when $X$ is the unit $\|\cdot\|_\infty$ ball in $\mathbf{R}^n$, the $k$-step minimax risk $\mathrm{Risk}_{\mathcal{F},X,\mathcal{O}}(k)$ of minimizing over $X$ of objectives from the family $\mathcal{F} = \mathcal{F}_{\|\cdot\|_\infty}(\kappa, L)$ in the range $k \leq n$ is lower-bounded by $O(1/\ln n)Lk^{1-\kappa}$. Comparing this lower risk bound with the efficiency estimate (2) we conclude that *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 $O(1)\ln n$ factor, whatever be a local oracle in use.* In fact, comparing (22) and (2), we see that the same conclusion remains true when $\|\cdot\|_\infty$ and $X$ are replaced with $\|\cdot\|_p$ and the unit ball of $\|\cdot\|_p$, respectively, and *$p$ is large*, specifically, $p \geq O(1)\ln n$.

**B.** In fact, in the case of $2 \leq p < \infty$ the lower complexity bounds for smooth convex minimization over $\|\cdot\|_p$-balls established in Corollary 4.1, are tight: it is shown in [5], see also [2, Chapter 2] 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 unit $\|\cdot\|_p$-ball $X$ in $\mathbf{R}^n$, for *every* number $k = 1, 2, ...$ of steps ensures that

$$f(x_k(\mathcal{N}, f)) - \min_{x \in X} f(x) \leq C(p) \frac{L}{k^{\kappa + \frac{\kappa}{p} - 1}},$$

with $C(p)$ depending solely on $p$, which is in full accordance with (23).

**C.** Corollary 4.1 remains true when replacing in it the embedding space $E = \mathbf{R}^n$ of $X$ with the space $E = \mathbf{R}^{n \times n}$ of $n \times n$ matrices, the norm $\|\cdot\|_p$ on $\mathbf{R}^n$ with the Shatten norm $\|\cdot\|_{\mathrm{Sh},p}$ [3)], 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 $\|\cdot\|_{\mathrm{Sh},p}$."

The latter 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 $\|\cdot\|_{\mathrm{Sh},p}$ 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. $\|\cdot\|_{\mathrm{Sh},p}$, over the unit $\|\cdot\|_{\mathrm{Sh},p}$ 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 $\mathcal{P}(\mathcal{F}_{\|\cdot\|_{\mathrm{Sh},p}}(\kappa, L), X^+)$.

[3)] the norm $\|x\|_{\mathrm{Sh},p}$ of an $n \times n$ matrix $x$ is, by definition the $\|\cdot\|_p$-norm of the vector of singular values of $x$.

# References

[1] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[2] K.-H. Elster. *Modern Mathematical Methods in Optimization.* Academie Verlag, Berlin, 1993.

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

[4] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization.* Wiley -Interscience, 1 edition, 1983.

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

[6] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Math. Dokl.*, 27:2:372–376, 1983.

[7] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming Series A*, 103:127–152, 2005.

[8] M. Raginsky and A. Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10):7036–7056, 2011.