SIAM J. OPTIM.
Vol. 28, No. 3, pp. 2384–2405

OPTIMAL AFFINE-INVARIANT SMOOTH MINIMIZATION ALGORITHMS*

ALEXANDRE D'ASPREMONT†, CRISTÓBAL GUZMÁN‡, AND MARTIN JAGGI§

Abstract. We formulate an affine-invariant implementation of the accelerated first-order algorithm in [Y. Nesterov, *Dokl. Math.*, 27 (1983), pp. 372–376]. Its complexity bound is proportional to an affine-invariant regularity constant defined with respect to the Minkowski gauge of the feasible set. We extend these results to more general problems, optimizing Hölder smooth functions using p-uniformly convex prox terms, and derive an algorithm whose complexity better fits the geometry of the feasible set and adapts to both the best Hölder smoothness parameter and the best gradient Lipschitz constant. Finally, we detail matching complexity lower bounds when the feasible set is an ℓ_p ball. In this setting, our upper bounds on iteration complexity for the algorithm in [Y. Nesterov, *Dokl. Math.*, 27 (1983), pp. 372–376] are thus optimal in terms of target precision, smoothness, and problem dimension.

Key words. convex optimization, optimal methods, affine invariance, complexity theory

AMS subject classifications. 90C25, 90C60, 65K05

DOI. 10.1137/17M1116842

1. Introduction. Here, we show how to implement the smooth minimization algorithm described in Nesterov (1983, 2005) so that both its iterations and its complexity bound are invariant with respect to a change of coordinates in the problem. We focus on a generic convex minimization problem written as

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q, \end{array} \tag{1}$$

where f is a convex function with Lipschitz continuous gradient and Q is a compact convex set. Without too much loss of generality, we will assume that the interior of Q is nonempty and contains zero. When Q is sufficiently simple, in a sense that will be made precise later, Nesterov (1983) showed that this problem can be solved with a complexity $O(1/\sqrt{\varepsilon})$, where ε is the target precision. Furthermore, it can be shown that this complexity bound is optimal in ε for large dimensions, for the class of smooth problems (Nemirovskiĭ and Yudin, 1979; Nesterov, 2003).

While the dependence in $1/\sqrt{\varepsilon}$ of the complexity bound in Nesterov (1983) is optimal in ε, the various factors in front of that bound contain parameters which can heavily vary with implementation, i.e., the choice of norm and prox regularization function. In fact, the full upper bound on the iteration complexity of the optimal

*Received by the editors February 15, 2017; accepted for publication (in revised form) June 6, 2018; published electronically August 23, 2018.

http://www.siam.org/journals/siopt/28-3/M111684.html

Funding: The authors would like to acknowledge support from the chaire *Économie des nouvelles données*, the *data science* joint research initiative with the *fonds AXA pour la recherche* and a gift from Société Générale Cross Asset Quantitative Research. The first and third authors would like to acknowledge support from the European Research Council (Project SIPA). The second author would like to acknowledge support from FONDECYT Iniciación (Project 11160939). The third author also acknowledges support from the Swiss National Science Foundation (SNSF).

†CNRS & D.I., UMR 8548, École Normale Supérieure, Paris 75005, France (aspremon@ens.fr).

‡Institute for Mathematical & Computational Engineering, Facultad de Matemáticas & Escuela de Ingeniería, Pontificia Universidad Católica de Chile, Santiago RM 7820436, Chile (crguzmanp@uc.cl).

§EPFL, Lausanne CH-1015, Switzerland (martin.jaggi@epfl.ch).

algorithm in Nesterov (2003) is written

$$\sqrt{\frac{8L\Phi(x^*)}{\sigma\varepsilon}}, \tag{2}$$

where L is the Lipschitz constant of the gradient, $\Phi(x^*)$ the value of the prox at the optimum, and σ its strong convexity parameter, all varying with the choice of norm and prox. This means in particular that, everything else being equal, this bound is not invariant with respect to an affine change of coordinates.

Arguably then, the complexity bound varies while the intrinsic complexity of problem (1) remains unchanged. Optimality in ε is thus no guarantee of computational efficiency, and a poorly parameterized optimal method can exhibit far from optimal numerical performance. On the other hand, optimal choices of norm and prox, and hence of L and Φ, should produce affine-invariant bounds. Hence, affine invariance, besides its implications in terms of numerical stability, can also act as a guide to optimally choose norm and prox. In other words, affine invariance is a necessary but not sufficient condition for the optimality of the bound in (2) with respect to the choice of norm and prox.

Here, we show how to choose this underlying norm and prox term for the algorithm in Nesterov (1983, 2005) such that we make both the iterations and complexity bounds invariant by a change of coordinates. In section 4, we construct the norm as the Minkowski gauge of centrally symmetric sets Q, then derive the prox using a definition of the regularity of Banach spaces used by (Juditsky and Nemirovski, 2008) to derive concentration inequalities. These systematic choices allow us to derive an affine-invariant bound on the complexity of the algorithm in Nesterov (1983).

When Q is an ℓ_p ball, with $1 \leq p \leq 2$, we show that this complexity bound matches lower complexity bounds from Guzmán and Nemirovski (2015) for the large-scale regime (more precisely, $n = \Omega(1/\varepsilon^2)$).[1] In section 5, we extend our results to much more general problems, deriving a new algorithm to optimize Hölder smooth functions using p-uniformly convex prox functions. This extends the results of Nemirovskii and Nesterov (1985) by incorporating adaptivity to the Hölder continuity of the gradient, and those of (Nesterov, 2015) by allowing general uniformly convex prox functions, not just strongly convex ones.

These additional degrees of freedom allow us to match optimal complexity lower bounds from Guzmán and Nemirovski (2015) when optimizing on ℓ_p balls, with $2 < p < \infty$, in the large-scale regime (namely, $n = \Omega(1/\varepsilon^{1/\lfloor 1+2/p \rfloor})$),[2] with adaptivity in the Hölder smoothness parameter and Lipschitz constant as a bonus. This means that, on ℓ_p-balls at least, our complexity bounds are optimal not only in terms of target precision ε, but also in terms of smoothness and problem dimension. This shows that, in the ℓ_p setting at least, affine invariance does indeed lead to optimal complexity.

2. Notation and preliminaries. For $1 \leq p \leq \infty$ we define the conjugate exponent $1 \leq p_* \leq \infty$ as the one such that $\frac{1}{p} + \frac{1}{p_*} = 1$. It is well known that the space dual to $\ell_p = (\mathbb{R}^n, \|\cdot\|_p)$ with $1 \leq p \leq \infty$ can be isometrically identified with ℓ_{p_*}.

[1] In this bound we omit the dependence on the Lipschitz constant. For a more precise statement we refer the reader to section 4.

[2] In this bound we omit the dependence on the Lipschitz constant. For a more precise statement we refer the reader to section 5.

Convexity and duality. Given a convex proper lower semicontinuous (l.s.c.) function $f : Q \to \mathbb{R}$, its *Fenchel conjugate* is the convex proper l.s.c. function $f^* : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f^*(z) = \sup_{x \in Q} \langle z, x \rangle - f(x).$$

It is standard in convex analysis to consider globally defined functions $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, where $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$; we can naturally extend our original $f : Q \to \mathbb{R}$ as $+\infty$ outside of Q, which is consistent with the definition of the Fenchel conjugate.

Smoothness. We say that a function $f : Q \to \mathbb{R}$ is smooth if it has a Lipschitz continuous gradient w.r.t. a norm $\|\cdot\|$:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} L \|y - x\|^2, \quad x, y \in Q, \tag{3}$$

Since in finite dimensions all norms are equivalent, smoothness is a property which is invariant under the choice of norm; however, the constant L heavily depends on this choice.

Strong and uniform convexity. Regularization is a key tool in the analysis of optimal first-order methods. We say a function $\Phi : Q \to \mathbb{R}$ is a prox if it is strictly convex and subdifferentiable in Q. Notice the subdifferentiability property is mild as any convex l.s.c. function is locally Lipschitz, and thus subdifferentials are guaranteed to exist in the relative interior of Q. Given a prox, we will denote by $\nabla\Phi(x)$ any choice of subgradient of Φ at x. Given a prox function we define its prox-center as

$$x_0 := \operatorname*{argmin}_{x \in Q} \Phi(x) \tag{4}$$

and its constant of variation as

$$D_\Phi(Q) := \sup_{x \in Q} \Phi(x) - \inf_{x \in Q} \Phi(x). \tag{5}$$

In sections 3 and 4 we further require the prox function to be strongly convex, and in section 5 we relax this condition to allow for p-uniformly convex prox, where $2 \leq p < \infty$.

DEFINITION 2.1 (uniform convexity and strong convexity). *Let $2 \leq p < \infty$, $\mu > 0$, and $Q \subseteq \mathbb{R}^n$, a closed convex set. A function $\Phi : Q \to \mathbb{R}$ that is subdifferentiable on Q is p-uniformly convex with constant μ w.r.t. $\|\cdot\|$ iff, for all $x \in Q$, $y \in Q$,*

$$\Phi(y) \geq \Phi(x) + \langle \nabla\Phi(x), y - x \rangle + \frac{\mu}{p} \|y - x\|^p. \tag{6}$$

Finally, we say Φ is strongly convex if it is 2-uniformly convex.

From now on, whenever the constant μ of p-uniform convexity is not explicitly stated, $\mu = 1$.

Duality between strong convexity and smoothness. A classical result in convex analysis relates the properties of strong convexity and smoothness via the Fenchel conjugate.

THEOREM 2.2 (see, e.g., Hiriart-Urruty and Lemaréchal (1993, Chapter X, Theorem 4.2.1)). *Let $f : Q \to \mathbb{R}$ be a convex l.s.c. function. Then f is strongly convex w.r.t. the norm $\|\cdot\|$ with constant $\mu > 0$ if and only f^* has Lipschitz continuous gradient w.r.t. the norm $\|\cdot\|_*$ with constant $L = 1/\mu$.*

3. Smooth optimization algorithm. We first recall the basic structure of the algorithm in Nesterov (1983). While many variants of this method have been derived, we use the formulation in Nesterov (2005). We *choose a norm* $\|\cdot\|$ and assume that the function f in problem (1) is convex with Lipschitz continuous gradient, so inequality (3) holds for some $L > 0$. We also *choose a prox function* $\Phi(\cdot)$ for the set Q, in this case a strongly convex function on Q with constant σ, which is subdifferentiable in Q (see Nesterov (2003) or Hiriart-Urruty and Lemaréchal (1993) for a discussion of regularization techniques using strongly convex functions). We let x_0 be the prox-center for Φ so that

$$x_0 := \operatorname*{argmin}_{x \in Q} \Phi(x),$$

assuming without loss of generality (w.l.o.g.) that $\Phi(x_0) = 0$, we then get in particular

$$\Phi(x) \geq \frac{1}{2}\sigma\|x - x_0\|^2. \tag{7}$$

We write $T_Q(x)$ as a solution to the following subproblem:

$$T_Q(x) := \operatorname*{argmin}_{y \in Q} \left\{ \langle \nabla f(x), y - x\rangle + \frac{1}{2}L\|y - x\|^2 \right\}. \tag{8}$$

We let $y_0 := T_Q(x_0)$, where x_0 is defined above. We recursively define three sequences of points: the current iterate x_t, the corresponding $y_t = T_Q(x_t)$, and the points

$$z_t := \operatorname*{argmin}_{x \in Q} \left\{ \frac{L}{\sigma}\Phi(x) + \sum_{i=0}^{t} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i\rangle] \right\} \tag{9}$$

given a step-size sequence $\alpha_k \geq 0$ with $\alpha_0 \in (0, 1]$ so that

$$\begin{aligned} x_{t+1} &= \tau_t z_t + (1 - \tau_t) y_t, \\ y_{t+1} &= T_Q(x_{t+1}), \end{aligned} \tag{10}$$

where $\tau_t = \alpha_{t+1}/A_{t+1}$ with $A_t = \sum_{i=0}^{t} \alpha_i$. We implicitly assume here that Q is simple enough so that the two subproblems defining y_t and z_t can be solved very efficiently. We have the following convergence result.

THEOREM 3.1 (Nesterov (2005)). *Suppose* $\alpha_t = (t + 1)/2$ *with the iterates* x_t, y_t, *and* z_t *defined in* (9) *and* (10). *Then for any* $t \geq 0$ *we have*

$$f(y_t) - f(x^*) \leq \frac{4L\Phi(x^*)}{\sigma\,(t + 1)^2},$$

where x^* *is an optimal solution to problem* (1).

If $\varepsilon > 0$ is the target precision, Theorem 3.1 ensures that Algorithm 1 will converge to an ε-accurate solution in no more than

$$\sqrt{\frac{8L\Phi(x^*)}{\sigma\varepsilon}} \tag{11}$$

iterations. In practice of course, $\Phi(x^*)$ needs to be bounded a priori and L and σ are often hard to evaluate.

While most of the parameters in Algorithm 1 are set explicitly, the norm $\|\cdot\|$ and the prox function $\Phi(\cdot)$ are chosen arbitrarily. In what follows, we will see that a natural choice for both makes the algorithm affine invariant.

Algorithm 1 Smooth minimization.

Require: x_0, the prox center of the set Q.

1: **for** $t = 0, \dots, T$ **do**
2: Compute $\nabla f(x_t)$.
3: Compute $y_t = T_Q(x_t)$.
4: Compute $z_t = \operatorname{argmin}_{x \in Q} \left\{ \frac{L}{\sigma}\Phi(x) + \sum_{i=0}^{t} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \right\}$.
5: Set $x_{t+1} = \tau_t z_t + (1 - \tau_t) y_t$.
6: **end for**

Ensure: $x_T, y_T \in Q$.

4. Affine-invariant implementation. We can define an affine change of coordinates $x = Aw$, where $A \in \mathbb{R}^{n \times n}$ is a nonsingular matrix, for which the original optimization problem in (1) is transformed so that

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & x \in Q & \end{array} \quad \text{becomes} \quad \begin{array}{lll} \text{minimize} & \hat{f}(w) & \\ \text{subject to} & w \in \hat{Q}, & \end{array} \tag{12}$$

in the variable $w \in \mathbb{R}^n$, where

$$\hat{f}(w) := f(Aw) \quad \text{and} \quad \hat{Q} := A^{-1}Q. \tag{13}$$

Unless A is pathologically ill-conditioned, both problems are equivalent and should have identical complexity bounds and iterations. In fact, the complexity analysis of Newton's method based on the self-concordance argument developed in Nesterov and Nemirovskii (1994) produces affine-invariant complexity bounds and the iterates themselves are invariant. Here we will show how to choose the norm $\|\cdot\|$ and the prox function $\Phi(\cdot)$ to get a similar behavior for Algorithm 1.

4.1. Choosing the norm. We start with a few classical results and definitions. Recall that the *Minkowski gauge* of a set Q is defined as follows.

DEFINITION 4.1. *Given $Q \subset \mathbb{R}^n$ containing zero, we define the Minkowski gauge of Q as*

$$\gamma_Q(x) := \inf\{\lambda \geq 0 : x \in \lambda Q\}$$

with $\gamma_Q(x) = 0$ when Q is unbounded in the direction x.

When Q is a compact convex, centrally symmetric set with respect to the origin and has nonempty interior, the Minkowski gauge defines a *norm*. We write this norm $\|\cdot\|_Q := \gamma_Q(\cdot)$. From now on, we will assume that the set Q is centrally symmetric or use, for example, $\bar{Q} = Q - Q$ (in the Minkowski sense) for the gauge when it is not (this can be improved and extending these results to the nonsymmetric case is a classical topic in functional analysis). Note that any linear transform of a centrally symmetric convex set remains centrally symmetric. The following simple result shows why $\|\cdot\|_Q$ is potentially a good choice of norm for Algorithm 1.

LEMMA 4.2. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$, Q is a centrally symmetric convex set with nonempty interior, and let $A \in \mathbb{R}^{n \times n}$ be a nonsingular matrix. Then f has Lipschitz continuous gradient with respect to the norm $\|\cdot\|_Q$ with constant $L > 0$, i.e.,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} L \|y - x\|_Q^2, \quad x, y \in Q,$$

if and only if the function $\hat{f}(w) := f(Aw)$ *has Lipschitz continuous gradient with respect to the norm* $\|\cdot\|_{A^{-1}Q}$ *with the same constant* L.

Proof. Let $w, y \in Q$, with $y = Az$ and $x = Aw$. Then

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} L \|y - x\|_Q^2, \quad x, y \in Q,$$

is equivalent to

$$f(Az) \leq f(Aw) + \left\langle A^{-T} \nabla_w f(Aw), Az - Aw \right\rangle + \frac{1}{2} L \|Az - Aw\|_Q^2, \quad z, w \in A^{-1}Q,$$

and, using the fact that $\|Aw\|_Q = \|w\|_{A^{-1}Q}$, this is also

$$f(Az) \leq f(Aw) + \left\langle \nabla_w f(Aw), A^{-1}(Az - Aw) \right\rangle + \frac{1}{2} L \|z - w\|_{A^{-1}Q}^2, \quad z, w \in A^{-1}Q,$$

hence the desired result. ∎

An almost identical argument shows the following analogous result for the property of *strong convexity* with respect to the norm $\|\cdot\|_Q$ and affine changes of coordinates. However, when starting from Lemma 4.2, this can also be seen as a consequence of the well-known duality between smoothness and strong convexity (see Theorem 2.2).

Combining these two results, we immediately have the following lemma.

LEMMA 4.3. *Let* $f : \mathbb{R}^n \to \mathbb{R}$, *let* Q *be a centrally symmetric convex set with nonempty interior, and let* $A \in \mathbb{R}^{n \times n}$ *be a nonsingular matrix. Suppose* f *is strongly convex with respect to the norm* $\|\cdot\|_Q$ *with parameter* $\sigma > 0$, *i.e.,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \sigma \|y - x\|_Q^2, \quad x, y \in Q,$$

if and only if the function $\hat{f}(w) := f(Aw)$ *is strongly convex with respect to the norm* $\|\cdot\|_{A^{-1}Q}$ *with the same parameter* σ.

We now turn our attention to the choice of prox function in Algorithm 1.

4.2. Choosing the prox. Choosing the norm as $\|\cdot\|_Q$ allows us to define a norm without introducing an arbitrary geometry in the algorithm, since the norm is extracted directly from the problem definition. Notice furthermore that by Theorem 2.2 when $(\|\cdot\|_Q^2)^*$ is smooth, we can set $\Phi(x) = \|x\|_Q^2$. The immediate impact of this choice is that the term $\Phi(x^*)$ in (11) is bounded by one, by construction. This choice has other natural benefits which are highlighted below. We first recall a result showing that the conjugate of a squared norm is the squared dual norm.

LEMMA 4.4. *Let* $\|\cdot\|$ *be a norm and* $\|\cdot\|^*$ *its dual norm. Then*

$$\frac{1}{2} \left(\|y\|^*\right)^2 = \sup_x \; y^T x - \frac{1}{2} \|x\|^2.$$

Proof. We recall the proof in (Boyd and Vandenberghe, 2004, Example 3.27) as it will prove useful in what follows. By definition, $x^T y \leq \|y\|^* \, \|x\|$, and hence

$$y^T x - \frac{1}{2} \|x\|^2 \leq \|y\|^* \, \|x\| - \frac{1}{2} \|x\|^2 \leq \frac{1}{2} \left(\|y\|^*\right)^2$$

because the second term is a quadratic function of $\|x\|^2$, with maximum $(\|y\|^*)^2/2$. This maximum is attained by any x such that $x^T y = \|y\|^* \, \|x\|$ (there must be one by construction of the dual norm), normalized so $\|x\| = \|y\|^*$, which yields the desired result. ∎

Computing the prox-mapping in (8) amounts to taking the conjugate of $\|\cdot\|^2$. We now recall another simple result showing that the dual of the norm $\|\cdot\|_Q$ is given by $\|\cdot\|_{Q^\circ}$, where

$$Q^\circ := \{x \in \mathbb{R}^n : \ x^T y \leq 1 \text{ for all } y \in Q\},$$

is the polar of the set Q.

LEMMA 4.5. *Let Q be a centrally symmetric convex set with nonempty interior. Then $\|\cdot\|_Q^* = \|\cdot\|_{Q^\circ}$.*

Proof. We write

$$\begin{aligned}\|x\|_{Q^\circ} &= \inf\{\lambda \geq 0 : x \in \lambda Q^\circ\} = \inf\{\lambda \geq 0 : x^T y \leq \lambda \text{ for all } y \in Q\} \\ &= \inf\left\{\lambda \geq 0 : \sup_{y \in Q} x^T y \leq \lambda\right\} = \sup_{y \in Q} x^T y = \|x\|_Q^*,\end{aligned}$$

which is the desired result. ∎

In light of the results above, we conclude that whenever Q° is smooth we obtain a natural prox function $\Phi(x) = \|x\|_Q^2$, whose strong convexity parameter is controlled by the Lipschitz constant of the gradient of $\|\cdot\|_{Q^\circ}^2$. However, this does not cover the case in which the squared norm $\|\cdot\|_Q$ is not strongly convex. In that scenario, we need to pick the norm based on Q but find a strongly convex prox function not too different from $\|\cdot\|_Q^2$. This is exactly the dual of the problem studied by Juditsky and Nemirovski (2008), who worked on concentration inequalities for vector-valued martingales and defined the regularity of a Banach space $(\mathbb{E}, \|\cdot\|_{\mathbb{E}})$ in terms of the smoothness of the best smooth approximation of the norm $\|\cdot\|_{\mathbb{E}}$.

We first recall a few more definitions, and we will then show that the regularity constant defined in Juditsky and Nemirovski (2008) produces an affine-invariant bound on the term $\Phi(x^*)/\sigma$ in the complexity of the smooth algorithm in Nesterov (1983).

DEFINITION 4.6. *Suppose $\|\cdot\|_X$ and $\|\cdot\|_Y$ are two norms on a space $\mathbb{E}$. The* distortion *$d(\|\cdot\|_X, \|\cdot\|_Y)$ between these two norms is equal to the smallest product $ab > 0$ such that*

$$\frac{1}{b}\|x\|_Y \leq \|x\|_X \leq a\|x\|_Y$$

over all $x \in \mathbb{E}$.

Note that $\log d(\|\cdot\|_X, \|\cdot\|_Y)$ defines a metric on the set of all symmetric convex bodies in $\mathbb{R}^n$, called the *Banach–Mazur distance*. We then recall the regularity definition in Juditsky and Nemirovski (2008).

DEFINITION 4.7. *The* regularity constant *of a Banach space $(\mathbb{E}, \|\cdot\|)$ is the smallest constant $\Delta > 0$ for which there exists a smooth norm $P(x)$ such that*

(i) *$P(x)^2/2$ has a Lipschitz continuous gradient with constant μ w.r.t. the norm $P(x)$, with $1 \leq \mu \leq \Delta$;*
(ii) *the norm $P(x)$ satisfies*

$$\|x\|^2 \leq P(x)^2 \leq \frac{\Delta}{\mu}\|x\|^2 \quad \text{for all } x \in \mathbb{E}, \tag{14}$$

and hence $d(P(\cdot), \|\cdot\|) \leq \sqrt{\Delta/\mu}$.

Note that in finite dimension, since all norms are equivalent to the Euclidean norm with distortion at most $\sqrt{\dim \mathbb{E}}$, we know that all finite-dimensional Banach spaces are at least $(\dim \mathbb{E})$-regular. Furthermore, the regularity constant is invariant with respect to an affine change of coordinates since both the distortion and the smoothness bounds are.

PROPOSITION 4.8. *Let $\varepsilon > 0$ be the target precision, suppose that the function f has a Lipschitz continuous gradient with constant L_Q with respect to the norm $\|\cdot\|_Q$, and that the space $(\mathbb{R}^n, \|\cdot\|_Q^*)$ is Δ_Q-regular. Then there exists a prox function for which Algorithm* 1 *will produce an ε-solution to problem* (1) *in at most*

$$\sqrt{\frac{4L_Q\Delta_Q}{\varepsilon}} \tag{15}$$

iterations. The constants L_Q and Δ_Q are affine invariant.

Proof. If $(\mathbb{R}^n, \|\cdot\|_Q^*)$ is Δ_Q-regular, then by Definition 4.7, there exists a norm $P^*(x)$ such that $P^*(x)^2/2$ has a Lipschitz continuous gradient with constant μ with respect to the norm $P^*(x)$, and Proposition 3.2 in Juditsky and Nemirovski (2008) shows by conjugacy that the prox function $\Phi(x) := (P^*(x)^2/2)^* = P(x)^2/2$ (where P is the norm conjugate to P^*) is strongly convex with respect to the norm $P(x)$ with constant $1/\mu$. Now (14) means that

$$\sqrt{\frac{\mu}{\Delta_Q}}\,\|x\|_Q \le P(x) \le \|x\|_Q \quad \text{for all } x \in Q$$

since $\|\cdot\|^{**} = \|\cdot\|$, and hence

$$\begin{aligned}\Phi(x+y) &\ge \Phi(x) + \langle \partial\Phi(x), y\rangle + \frac{1}{2\mu}P(y)^2\\ &\ge \Phi(x) + \langle \partial\Phi(x), y\rangle + \frac{1}{2\Delta_Q}\|y\|_Q^2,\end{aligned}$$

so $\Phi(x)$ is strongly convex with respect to $\|\cdot\|_Q$ with constant $\sigma = 1/\Delta_Q$, and, using (14) as above,

$$\frac{\Phi(x^*)}{\sigma} = \frac{P(x^*)^2\Delta_Q}{2} \le \frac{\|x^*\|_Q^2\Delta_Q}{2} \le \frac{\Delta_Q}{2},$$

by definition of $\|\cdot\|_Q$, if x^* is an optimal (hence feasible) solution of problem (1). The bound in (15) then follows from (11) and its affine invariance follows directly from affine invariance of the distortion and Lemmata 4.2 and 4.3. ∎

One important observation about the result above is that the constant Δ_Q is not necessarily the regularity constant (it should always be an upper bound, nevertheless), yet the resulting method still enjoys affine invariance. This may be important in settings where exactly computing the regularity constant or the associated prox function are difficult. Of course, it is always desirable to make the constant Δ_Q as small as possible, but one can trade-off this improvement on iteration complexity with tractability of the prox.

4.3. Example. ℓ_p balls. To illustrate our results, first consider the problem

$$\begin{array}{ll}\text{minimize} & f(x)\\ \text{subject to} & \|x\|_p \le 1\end{array} \tag{16}$$

in the variable $x \in \mathbb{R}^n$, where $\mathcal{B}_p := \{x \in \mathbb{R}^n : \sum_{i=1}^n |x_i|^p \le 1\}$ is the unit ℓ_p ball.

When $p \in [1,2]$, Example 3.2 in Nemirovski et al. (2009) shows that the dual norm $\|\cdot\|_{\frac{p}{p-1}}$ is Δ_p regular, with

$$\Delta_p = \inf_{2 \leq \rho < \frac{p}{p-1}} (\rho - 1) n^{\frac{2}{\rho} - \frac{2(p-1)}{p}} \leq \min\left\{ \frac{p}{p-1}, C \log n \right\} \quad \text{when } p \in [1,2].$$

When $p \in [2, \infty]$, the regularity is only controlled by the distortion $d(\|\cdot\|_{\frac{p}{p-1}}, \|\cdot\|_2)$, since $\|\cdot\|_\alpha$ is only smooth when $\alpha \geq 2$. This means that $\|\cdot\|_{\frac{p}{p-1}}$ is Δ_p regular, with

$$\Delta_p = n^{\frac{p-2}{p}} \quad \text{when } p \in [2, \infty].$$

This means that the complexity of solving (16) using Algorithm 1 is bounded by

$$\sqrt{\frac{4 L_p \Delta_p}{\varepsilon}}, \tag{17}$$

where L_p is the Lipschitz constant of ∇f with respect to the ℓ_p norm. We will see below that this bound is nearly optimal when $1 \leq p \leq 2$. However, for $p > 2$ the complexity bound contains dimension-dependent factors which are undesirable for practical purposes. In fact, these bounds are essentially suboptimal, and in order to obtain optimal methods in this range we will need to extend our theory to p-uniformly convex prox functions.

We show now that the complexity bounds derived from affine invariance are optimal for the regime $1 \leq p \leq 2$. Before doing this, it is worth mentioning that this optimality only holds for large-scale problems, namely where dimension n is larger than the number of iterations T: if one can afford a superlinear (in dimension) number of iterations, methods such as the center of gravity or ellipsoid can achieve better complexity estimates (Nemirovskiĭ and Yudin, 1979). It was proved in Guzmán and Nemirovski (2015) that the class of problems (16), where f is convex and has L_p-Lipschitz continuous gradient w.r.t. $\|\cdot\|_p$, satisfies the following lower bound on minimax risk:

$$\Omega\left(\frac{L_p}{T^2 \log(T+1)} \right),$$

where T is the number of iterations. This translates into the following lower bound on iteration complexity:

$$\Omega\left(\sqrt{\frac{L_p}{\varepsilon \log n}} \right)$$

as a function of the target precision $\varepsilon > 0$; observe that due to the large-scale condition, $\varepsilon \geq \Omega(L_p / [n^2 \log n])$. Therefore, for large-scale problems the affine-invariant algorithm is optimal, up to poly-logarithmic factors, in this range.

Our results on affine invariance can also be exploited to guide the choices of prox function for nonsymmetric feasible sets. Consider an optimization problem over the unit simplex, written

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \mathbf{1}^T x \leq 1, \ x \geq 0 \end{array} \tag{18}$$

in the variable $x \in \mathbb{R}^n$. As discussed in section 3.3 in Nemirovski et al. (2009), choosing $\|\cdot\|_1$ as the norm and $\Phi(x) = \log n + \sum_{i=1}^n x_i \log x_i$ as the prox function,

we have $\kappa = 1$ and $\Phi(x^*) \leq \log n$, which means the complexity of solving (18) using Algorithm 1 is bounded by

$$\sqrt{8\frac{L_1 \log n}{\varepsilon}}, \tag{19}$$

where L_1 is the Lipschitz constant of ∇f with respect to the ℓ_1 norm. This choice of norm and prox has a double advantage here. First, the prox term $\Phi(x^*)$ grows only as $\log n$ with the dimension. Second, the ℓ_∞ norm being the smallest among all ℓ_p norms, the smoothness bound L_1 is also minimal among all choices of ℓ_p norms.

Let us now follow our affine-invariant construction. The simplex $C = \{x \in \mathbb{R}^n : \mathbf{1}^T x \leq 1,\ x \geq 0\}$ is not centrally symmetric, but we can use the ℓ_1 ball as a symmetric proxy to choose a prox function. We do not modify the original problem, but simply choose the prox function as if we were optimizing over the ℓ_1-ball instead of the simplex. The Minkowski norm associated with the ℓ_1-ball is then of course equal to the ℓ_1-norm, so $\|\cdot\|_Q = \|\cdot\|_1$ here. The space $(\mathbb{R}^n, \|\cdot\|_\infty)$ is $2\log n$ regular (see Juditsky and Nemirovski (2008, Example 3.2)) with the prox function chosen here as $\|\cdot\|_\alpha^2/2$, with $\alpha = 2\log n/(2\log n - 1)$. Proposition 4.8 then shows that the complexity bound we obtain using this procedure is identical to that in (19). A similar result holds in the matrix case.

5. Hölder smooth functions and uniformly convex prox. We now extend the results of section 4 to problems in which the objective $f(x)$ is Hölder smooth and the prox function is p-uniformly convex, with arbitrary p. This generalization is necessary to derive optimal complexity bounds for smooth convex optimization over ℓ_p-balls when $p > 2$, and will require some extensions of the ideas we presented for the standard analysis, which was based on a strongly convex prox. We will consider a slightly different accelerated method that can be seen as a combination of mirror and gradient descent steps (Allen-Zhu and Orecchia, 2014). This variant of the accelerated gradient method is not substantially different, however, from the one used in the previous section, and its purpose is to make the step-size analysis more transparent. It is worth emphasizing that an interesting byproduct of our method is the analysis of an adaptive step-size policy, which can exploit weaker levels of Hölder continuity for the gradient.

In order to show the motivation behind our choice of p-uniformly convex prox, we begin with an example highlighting how the difficult geometries of ℓ_p-spaces when $p > 2$ necessarily lead to weak (dimension-dependent) complexity bounds for any strongly convex prox.

Example 5.1. Let $2 < p \leq \infty$ and let $\mathcal{B}_p$ be the unit p-ball on $\mathbb{R}^n$. Let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be any strongly convex (prox) function w.r.t. $\|\cdot\|_p$ with constant 1, and suppose w.l.o.g. that $\Phi(0) = 0$. We will prove that

$$\sup_{x \in \mathcal{B}_p} \Phi(x) \geq n^{1-2/p}/2.$$

We start from point $x_0 = 0$ and choose a direction $e_{1+} \in \{e_1, -e_1\}$ in order that $\langle \nabla\Phi(0), e_{1+}\rangle \geq 0$. By strong convexity, we have, for $x_1 := x_0 + \frac{e_1}{n^{1/p}}$,

$$\Phi(x_1) \geq \Phi(x_0) + \langle \nabla\Phi(0), e_{1+}\rangle + \frac{1}{2}\|x_1 - x_0\|^2 \geq \frac{1}{2n^{2/p}}.$$

Inductively, we can proceed by adding coordinate vectors one by one, $x_i := x_{i-1} + \frac{e_{i+}}{n^{1/p}}$ for $i = 1, \ldots, n$, where $e_{i+} \in \{e_i, -e_i\}$ is chosen so that $\langle \nabla\Phi(x_{i-1}), e_{i+}\rangle \geq 0$. For this

choice we can guarantee

$$\Phi(x_i) \geq \Phi(x_{i-1}) + \langle \nabla\Phi(x_{i-1}), e_{i+}\rangle + \frac{1}{2n^{2/p}} \geq \frac{i}{2n^{2/p}}.$$

At the end, the vector $x_n \in \mathcal{B}_p$ and $\Phi(x_n) \geq n^{1-2/p}/2$.

5.1. Uniform convexity. The previous example shows that strong convexity of the prox function is too restrictive when dealing with certain domain geometries, such as $Q = \mathcal{B}_p$ when $p > 2$ (see also Example 4.3 for a related discussion). In order to obtain dimension-independent bounds for these cases, we will have to consider relaxed notions of regularity for the prox, namely p-uniform convexity (see section 2 for the formal definition).

We turn our attention to the question of how to obtain an affine-invariant prox in the uniformly convex setup. In the previous section it was observed that the regularity constant of the dual space provided such tuning among strongly convex prox functions, however we are not aware of extensions of this notion to the uniformly smooth setup. Nevertheless, the same purpose can be achieved by directly minimizing the growth factor among the class of uniformly convex functions, which leads to the following notion.

DEFINITION 5.2 (minimal growth factor). *Given $Q \subseteq \mathbb{R}^n$, a convex and compact set on $(\mathbb{R}^n, \|\cdot\|)$, we define*

$$D_{p,Q} := \inf_{\Phi} \left\{ D_\Phi(Q) \;\middle|\; \Phi : Q \to \mathbb{R} \text{ is } p\text{-uniformly convex w.r.t. } \|\cdot\| \right\}, \tag{20}$$

where $D_\Phi(Q) = \sup_{x\in Q} \Phi(x) - \inf_{x\in Q} \Phi(x)$ is the constant of variation of Φ.

Some comments are in order. First, for fixed p, the constant $D_{p,Q}$ provides the optimal constant of variation among p-uniformly convex functions over Q, which means that, by construction, $D_{p,Q}$ is affine-invariant. Second, Example 5.1 showed that when $2 < p < \infty$, we have $D_{2,\mathcal{B}_p} \geq n^{1-2/p}/2$, and the function $\Phi(x) = \|x\|_2^2/2$ shows this bound is tight. We will later see that $D_{p,\mathcal{B}_p} = 1$, which is a major improvement for large dimensions. If we denote by Δ_Q the regularity constant defined in (4.7), then in Proposition 3.3 in Juditsky and Nemirovski (2008) it is shown that $\Delta_Q \geq D_{2,Q} \geq c\,\Delta_Q$, where $c > 0$ is an absolute constant, since $\Phi(\cdot)$ is not required to be a norm here.

When Q is the unit ball of a norm, a classical result by Pisier (1975) links the constant of variation in (20) above with the notion of martingale cotype. A Banach space $(\mathbb{E}, \|\cdot\|)$ has M-cotype q iff there is some constant $C > 0$ such that for any $T \geq 1$ and martingale difference $d_1, \ldots, d_T \in \mathbb{E}$ we have

$$\left(\sum_{t=1}^{T} \mathbf{E}\left[\|d_t\|^q\right] \right)^{1/q} \leq C\, \mathbf{E}\left[\left\| \textstyle\sum_{t=1}^{T} d_t \right\| \right].$$

Pisier (1975) then shows the following result.

THEOREM 5.3 (Pisier (1975)). *A Banach space $(\mathbb{E}, \|\cdot\|)$ has M-cotype q iff there exists a q-uniformly convex norm equivalent to $\|\cdot\|$.*

In the same spirit, there exists a concrete characterization of a function achieving the optimal constant of variation (see, e.g., Srebro, Sridharan, and Tewari (2011)). Unfortunately, this characterization does not lead to an efficiently computable prox.

For the analysis of our accelerated method with uniformly convex prox, we will also need the notion of *Bregman divergence*.

DEFINITION 5.4 (Bregman divergence). *Let* $(\mathbb{R}^n, \|\cdot\|)$ *be a normed space and* $\Phi : Q \to \mathbb{R}$ *be a p-uniformly convex function w.r.t.* $\|\cdot\|$. *We define the Bregman divergence as*

$$V_x(y) := \Phi(y) - \langle \nabla\Phi(x), y - x\rangle - \Phi(x) \quad \forall x \in \mathrm{int}Q,\ \forall y \in Q.$$

Observe that $V_x(x) = 0$ *and* $V_x(y) \geq \frac{1}{p}\|y - x\|^p$.

For starters, let us prove a simple fact that will be useful in the complexity bounds.

LEMMA 5.5 (three-points identity (Chen and Teboulle, 1993)). *Let* $\Phi : Q \to \mathbb{R}$ *be a p-uniformly convex function and* $V_x(\cdot)$ *the corresponding Bregman divergence. Then, for all* x, x', *and* u *in* Q,

$$V_x(u) - V_{x'}(u) - V_x(x') = \langle \nabla V_x(x'), u - x'\rangle.$$

Proof. From simple algebra

$$\begin{aligned}
&V_x(u) - V_{x'}(u) - V_x(x')\\
&= \Phi(u) - \langle \nabla\Phi(x), u - x\rangle - \Phi(x) - [\Phi(u) - \langle \nabla\Phi(x'), u - x'\rangle - \Phi(x')] - V_x(x')\\
&= \langle \nabla\Phi(x') - \nabla\Phi(x), u - x'\rangle + \underbrace{\Phi(x') - \langle \nabla\Phi(x), x' - x\rangle - \Phi(x)}_{=V_x(x')} - V_x(x')\\
&= \langle \nabla\Phi(x') - \nabla\Phi(x), u - x'\rangle = \langle \nabla V_x(x'), u - x'\rangle,
\end{aligned}$$

which is the desired result. ∎

5.2. An accelerated method for minimizing Hölder smooth functions. We consider classes of weakly smooth convex functions. For a Hölder exponent $\kappa \in (1, 2]$ we denote the class $\mathcal{F}^{\kappa}_{\|\cdot\|}(Q, L_\kappa)$ as the set of convex functions $f : Q \to \mathbb{R}$ such that, for all $x, y \in Q$,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L_\kappa \|x - y\|^{\kappa - 1}.$$

Before describing the method, we first define a step sequence that is useful in the algorithm. For a given p such that $2 \leq p < \infty$, consider the sequence $(\gamma_t)_{t \geq 0}$ defined by $\gamma_1 = 1$ and, for any $t > 1$, γ_{t+1} is the major root of

$$\gamma^p_{t+1} = \gamma^{p-1}_{t+1} + \gamma^p_t.$$

This sequence has the following properties.

PROPOSITION 5.6. *The following properties hold for the auxiliary sequence* $(\gamma_t)_{t \geq 0}$*:*

(i) *the sequence is increasing;*
(ii) $\gamma^p_t = \sum_{s=1}^t \gamma^{p-1}_s$*;*
(iii) $\frac{t}{p} \leq \gamma_t \leq t$*; and*
(iv) $\sum_{s=1}^t \gamma^p_s \leq t\gamma^p_t$.

Proof.

(i) By definition, $\gamma^p_{t+1} = \gamma^{p-1}_{t+1} + \gamma^p_t \geq \gamma^p_t$, and thus $\gamma_{t+1} \geq \gamma_t$.
(ii) By telescoping the recursion, $\gamma^p_t = \sum_{s=1}^t \gamma^{p-1}_s$.

(iii) For the lower bound, a Fenchel-type inequality yields

$$\gamma_t = \gamma_{t+1}^{1/p_*}[\gamma_{t+1}-1]^{1/p} \leq \frac{\gamma_{t+1}}{p_*} + \frac{\gamma_{t+1}-1}{p} = \gamma_{t+1} - \frac{1}{p}.$$

The upper bound is proved by induction as follows:

$$\begin{aligned}(t+1)^p &= (t+1)^{p-1} + t(t+1)^{p-1} \;>\; (t+1)^{p-1} + t[t^{p-1} + (p-1)t^{p-2}] \\ &\geq (t+1)^{p-1} + \gamma_t^p,\end{aligned}$$

where the last inequality holds by induction hypothesis, $\gamma_t \leq t$. As a conclusion, the major root defining γ_{t+1} has to be at most $t+1$.

(iv) By (ii), we have

$$\sum_{s=1}^{t}\gamma_s^p = \sum_{s=1}^{t}\sum_{r=1}^{s}\gamma_s^{p-1} = \sum_{r=1}^{t}(t-r)\gamma_r^{p-1} \leq t\sum_{r=1}^{t}\gamma_r^{p-1} = t\gamma_t^p,$$

which concludes the proof. ∎

We now prove a simple lemma controlling the smoothness of f in terms of $\|\cdot\|^p$. This idea is a minor extension of the "inexact gradient trick" proposed in Devolder, Glineur, and Nesterov (2011) and further studied in Nesterov (2015), which corresponds to the special case in which $p = 2$ for the results described here. As in Devolder, Glineur, and Nesterov (2011), this trick will allow us to minimize Hölder smooth functions by treating their gradient as an inexact oracle on the gradient of a smooth function.

LEMMA 5.7. *Let $f \in \mathcal{F}^{\kappa}_{\|\cdot\|}(Q, L_\kappa)$. Then for any $\delta > 0$ and*

$$M \geq \left[\frac{2}{p}\left(\frac{p-\kappa}{\kappa}\right)\frac{1}{\delta}\right]^{\frac{p-\kappa}{\kappa}} L_\kappa^{\frac{p}{\kappa}} \tag{21}$$

we have that, for all $x, y \in Q$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{p}M\|y-x\|^p + \frac{\delta}{2}.$$

Proof. By assumption on f, the following bound holds for any $x, y \in Q$:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_\kappa}{\kappa}\|y-x\|^\kappa.$$

Notice first that it suffices to show that, for all $t \geq 0$,

$$\frac{L_\kappa}{\kappa}t^\kappa \leq \frac{M}{p}t^p + \frac{\delta}{2}. \tag{22}$$

This can be seen by letting $t = \|y - x\|$ and using (22) in the preceding inequality. Let us prove (22). First recall the following Fenchel-type inequality: if $r, s \geq 1$ and $1/r + 1/s = 1$, then for all x and y we have that $xy \leq \frac{1}{r}x^r + \frac{1}{s}y^s$. For $r = p/\kappa$, $s = p/(p-\kappa)$, and $x = t^\kappa$, we obtain

$$\frac{L_\kappa}{\kappa}t^\kappa \leq \frac{1}{p}\frac{L_\kappa}{y}t^p + \frac{L_\kappa(p-\kappa)}{p\kappa}y^{\frac{\kappa}{p-\kappa}}.$$

Now we choose y so that $\frac{\delta}{2} = \frac{L_\kappa(p-\kappa)}{p\kappa} y^{\frac{\kappa}{p-\kappa}}$, which leads to the inequality

$$\frac{L_\kappa}{\kappa} t^\kappa \leq \frac{1}{p}\frac{L_\kappa}{y} t^p + \frac{\delta}{2}.$$

Finally, by our choice of M we have that $M \geq L_\kappa / y$, proving (22) and therefore the result. ∎

Algorithm 2 Accelerated method with uniformly convex prox.

Require: $x_0 \in Q$, $2 \leq p < \infty$, $\delta > 0$, $T \in \mathbb{N} \setminus \{0\}$, and $M > 0$ as in (21) achieving equality.

1: $y_0 = x_0$, $z_0 = x_0$, and $A_0 = 0$.
2: **for** $t = 0, \ldots, T-1$ **do**
3: $\alpha_{t+1} = \gamma_{t+1}^{p-1}/M$,
4: $A_{t+1} = A_t + \alpha_{t+1}$,
5: $\tau_t = \alpha_{t+1}/A_{t+1}$,
6: $x_{t+1} = \tau_t z_t + (1-\tau_t) y_t$.
7: Obtain from oracle $\nabla f(x_{t+1})$, and update

$$y_{t+1} = \arg\min_{y \in Q} \left\{ \frac{M}{p}\|y - x_{t+1}\|^p + \langle \nabla f(x_{t+1}), y - x_{t+1}\rangle \right\}, \tag{23}$$

$$z_{t+1} = \arg\min_{z \in Q} \left\{ V_{z_t}(z) + \alpha_{t+1}\langle \nabla f(x_{t+1}), z - z_t\rangle \right\}. \tag{24}$$

8: **end for**
9: **return** y^T.

As we will show below, the accelerated method described in Algorithm 2 extends the ℓ_p-setting for acceleration first proposed by Nemirovskii and Nesterov (1985) to nonsmooth spaces, using Bregman divergences. This gives us more flexibility in the choice of prox function and allows us in particular to better fit the geometry of the feasible set.

PROPOSITION 5.8. *Let $f \in \mathcal{F}^\kappa_{\|\cdot\|}(Q, L_\kappa)$ and let $\Phi : Q \to \mathbb{R}$ be p-uniformly convex w.r.t. $\|\cdot\|$. Then for any $\varepsilon > 0$, setting $\delta := \varepsilon/T$, and for any M satisfying* (21)*, the accelerated method in Algorithm* 2 *guarantees an accuracy*

$$f(y_T) - f(y^*) \leq \frac{D_\Phi(Q)}{A_T} + \frac{\varepsilon}{2}$$

after T iterations.

Proof. Let $u \in Q$ be an arbitrary vector. Using the optimality conditions for subproblem (24), and Lemma 5.5, we get

$$\begin{aligned}
&\alpha_{t+1}\langle \nabla f(x_{t+1}), z_t - u\rangle \\
&\quad = \alpha_{t+1}\langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle + \alpha_{t+1}\langle \nabla f(x_{t+1}), z_{t+1} - u\rangle \\
&\quad \leq \alpha_{t+1}\langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle - \langle \nabla V_{z_t}(z_{t+1}), z_{t+1} - u\rangle \\
&\quad = \alpha_{t+1}\langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle + V_{z_t}(u) - V_{z_{t+1}}(u) - V_{z_t}(z_{t+1}) \\
&\quad \leq \left[\alpha_{t+1}\langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle - \frac{1}{p}\|z_t - z_{t+1}\|^p\right] + V_{z_t}(u) - V_{z_{t+1}}(u).
\end{aligned}$$

Let us examine the latter term in brackets closely. For this, let $v = \tau_t z_{t+1} + (1-\tau_t) y_t$ and note that $x_{t+1} - v = \tau_t(z_t - z_{t+1})$. With $\tau_t = \alpha_{t+1}/A_{t+1}$ we also have, using Proposition 5.6(ii),

$$\frac{1}{\tau_t^p} = \left(\frac{L \sum_{s=1}^{t+1} \gamma_s^{p-1}}{L \gamma_{t+1}^{p-1}} \right)^p = \gamma_{t+1}^p = M A_{t+1}.$$

From this we obtain

$$\begin{aligned}
&\alpha_{t+1} \langle \nabla f(x_{t+1}), z_t - z_{t+1} \rangle - \frac{1}{p} \|z_t - z_{t+1}\|^p \\
&\quad = \left\langle \frac{\alpha_{t+1}}{\tau_t} \nabla f(x_{t+1}), x_{t+1} - v \right\rangle - \frac{1}{p \tau_t^p} \|x_{t+1} - v\|^p \\
&\quad = A_{t+1} \left[\langle \nabla f(x_{t+1}), x_{t+1} - v \rangle - \frac{M}{p} \|x_{t+1} - v\|^p \right] \\
&\quad \leq A_{t+1} \left[\langle \nabla f(x_{t+1}), x_{t+1} - y_{t+1} \rangle - \frac{M}{p} \|x_{t+1} - y_{t+1}\|^p \right] \\
&\quad \leq A_{t+1} \left[f(x_{t+1}) - f(y_{t+1}) + \frac{\delta}{2} \right],
\end{aligned}$$

where the first inequality holds by the definition of y_{t+1}, and the last inequality holds by Lemma 5.7 and the choice of M. This means that

$$\alpha_{t+1} \langle \nabla f(x_{t+1}), z_t - u \rangle \leq A_{t+1} \left[f(x_{t+1}) - f(y_{t+1}) + \delta/2 \right] + V_{z_t}(u) - V_{z_{t+1}}(u). \tag{25}$$

From (25) and other simple estimates,

$$\begin{aligned}
&\alpha_{t+1}[f(x_{t+1}) - f(u)] \\
&\quad \leq \alpha_{t+1} \langle \nabla f(x_{t+1}), x_{t+1} - u \rangle \\
&\quad = \alpha_{t+1} \langle \nabla f(x_{t+1}), x_{t+1} - z_t \rangle + \alpha_{t+1} \langle \nabla f(x_{t+1}), z_t - u \rangle \\
&\quad = \frac{(1-\tau_t)\alpha_{t+1}}{\tau_t} \langle \nabla f(x_{t+1}), y_t - x_{t+1} \rangle + \alpha_{t+1} \langle \nabla f(x_{t+1}), z_t - u \rangle \\
&\quad \leq \frac{(1-\tau_t)\alpha_{t+1}}{\tau_t} [f(y_t) - f(x_{t+1})] + \alpha_{t+1} \langle \nabla f(x_{t+1}), z_t - u \rangle \\
&\quad \leq \frac{(1-\tau_t)\alpha_{t+1}}{\tau_t} [f(y_t) - f(x_{t+1})] + A_{t+1}[f(x_{t+1}) - f(y_{t+1}) + \delta/2] \\
&\qquad + V_{z_t}(u) - V_{z_{t+1}}(u) \\
&\quad = (A_{t+1} - \alpha_{t+1})[f(y_t) - f(x_{t+1})] + A_{t+1}[f(x_{t+1}) - f(y_{t+1}) + \delta/2] \\
&\qquad + V_{z_t}(u) - V_{z_{t+1}}(u).
\end{aligned}$$

Therefore

$$A_{t+1} f(y_{t+1}) - A_t f(y_t) + V_{z_{t+1}}(u) - V_{z_t}(u) \leq \alpha_{t+1} f(u) + A_{t+1} \frac{\delta}{2}.$$

Summing these inequalities, we obtain

$$A_T f(y_T) + [V_{z_{t+1}}(u) - V_{z_0}(u)] \leq A_T f(u) + \sum_{t=1}^{T} A_t \frac{\delta}{2}.$$

Now, by Proposition 5.6, we have $\frac{1}{A_T}\sum_{t=1}^{T} A_t \le \frac{1}{\gamma_T^p} T\gamma_T^p \le T$, and thus, by the choice $\delta = \varepsilon/T$, we obtain

$$f(y_T) - f(u) \le \frac{V_{z_0}(u)}{A_T} + \frac{\varepsilon}{2}.$$

Definition 5.4 together with the fact that $\langle \nabla\Phi(x), y - x\rangle \ge 0$ when x minimizes $\Phi(x)$ over Q then yields the desired result. ∎

Algorithm 3 Accelerated method with uniformly convex prox and adaptive step size.

Require: $x^0 \in Q$, $2 \le p < \infty$, and $T \in \mathbb{N} \setminus \{0\}$.

Set $y_0 = x_0$, $z_0 = x_0$, $M_0 = 1$ and $A_0 = 0$.
for $t = 0, \dots, T-1$ **do**
  $M = M_t/2$.
  repeat
    Set

$$\begin{aligned} M &= 2M, \\ \alpha &= \max\left\{a : M^{\frac{1}{p-1}} a^{p_*} - a = A_t\right\}, \\ A &= A_t + \alpha, \\ \tau &= \alpha/A, \\ x_{t+1} &= \tau z_t + (1-\tau) y_t. \end{aligned}$$

    Obtain $\nabla f(x_{t+1})$ and compute

$$y_{t+1} = \arg\min_{y\in Q}\left\{\frac{M}{p}\|y - x_{t+1}\|^p + \langle \nabla f(x_{t+1}), y - x_{t+1}\rangle\right\}$$

  until

$$f(y_{t+1}) \le f(x_{t+1}) + \langle \nabla f(x_{t+1}), y_{t+1} - x_{t+1}\rangle + \frac{M}{p}\|y_{t+1} - x_{t+1}\|^p + \frac{\tau\varepsilon}{2}. \tag{26}$$

  Set $M_{t+1} = M/2$, $\alpha_{t+1} = \alpha$, $A_{t+1} = A$, $\tau_t = \tau$.
  Compute

$$z_{t+1} = \arg\min_{z\in Q}\left\{V_{z_t}(z) + \alpha_{t+1}\langle \nabla f(x_{t+1}), z - z_t\rangle\right\}.$$

end for
return y.

In order to obtain the convergence rate of the method, we need to estimate the value of A_T given the choice of M. For this we assume the bound in (21) is satisfied with equality. Since $A_T = \gamma_T^p/M$ we can use Proposition 5.6(iii), so that

$$\begin{aligned} A_T &= \gamma_T^p \left[\frac{p}{2}\left(\frac{\kappa}{p-\kappa}\right)\frac{\varepsilon}{T}\right]^{\frac{p-\kappa}{\kappa}} L_\kappa^{-\frac{p}{\kappa}} \\ &\ge p^{-p} T^{p+1-\frac{p}{\kappa}} \varepsilon^{\frac{p}{\kappa}-1}\left[\frac{p}{2}\left(\frac{\kappa}{p-\kappa}\right)\right]^{\frac{p-\kappa}{\kappa}} L_\kappa^{-\frac{p}{\kappa}}. \end{aligned}$$

Notice that to obtain an ε-solution it suffices to have $A_T \geq 2D_\Phi(Q)/\varepsilon$. By imposing this lower bound on the lower bound obtained for A_T we get the following complexity estimate.

COROLLARY 5.9. *Let $f \in \mathcal{F}^\kappa_{\|\cdot\|}(Q, L_\kappa)$ and let $\Phi : Q \to \mathbb{R}$ be p-uniformly convex w.r.t. $\|\cdot\|$. Setting $\delta := \varepsilon/T$, and with M satisfying (21), the accelerated method in Algorithm 2 requires*

$$T < p\left[2^p\left(\frac{p-\kappa}{\kappa}\right)^{p-\kappa}\frac{D_\Phi(Q)^\kappa L_\kappa^p}{\varepsilon^p}\right]^{\frac{1}{(p+1)\kappa-p}} + 1$$

iterations to reach an accuracy ε.

We will later see that the algorithm above leads to optimal complexity bounds (that is, unimprovable up to constant factors) for ℓ_p-setups. However, our algorithm is highly sensitive to several parameters, the most important being κ (the smoothness) and L_κ, which sets the step size. We now focus on designing an adaptive step-size policy that does not require L_κ as input and adapts itself to the best weak smoothness parameter $\kappa \in (1, 2]$.

5.3. An adaptive gradient method. We will now extend the adaptive algorithm in Theorem 3 of Nesterov (2015) to handle p-uniformly convex prox functions using Bregman divergences. This new method with adaptive step-size policy is described as Algorithm 3. From line 5 in Algorithm 3 we get the following identities:

$$A^{p-1} = \alpha^p M, \tag{27}$$

$$\frac{1}{\tau^p} = MA. \tag{28}$$

These identities are analogous to the ones derived for the nonadaptive variant. For this reason, the analysis of the adaptive variant is almost identical. There are a few extra details to address, which is what we do now. First, we need to show that the line-search procedure is feasible. That is, it always terminates in finite time. This is intuitively true from Lemma 5.7, but let us make this intuition precise. From (27) and (28) we have

$$M\tau^{\frac{p-\kappa}{\kappa}} = \frac{A^{p-1}}{\alpha^p}\left(\frac{\alpha}{A}\right)^{\frac{p}{\kappa}-1} = \frac{1}{\alpha}\left(\frac{A}{\alpha}\right)^{p-\frac{p}{\kappa}} \geq \frac{1}{\alpha}.$$

Notice that whenever the condition (26) of Algorithm 3 is not satisfied, M is doubled. Suppose the line-search does not terminate; then $\alpha \to 0$. However, by Lemma 5.7, the termination condition (26) is guaranteed to be satisfied as soon as

$$M \geq \left[\frac{2}{p}\left(\frac{p-\kappa}{\kappa}\right)\frac{1}{\varepsilon\tau}\right]^{\frac{p-\kappa}{\kappa}} L_\kappa^{\frac{p}{\kappa}},$$

which is a contradiction with $\alpha \to 0$.

To produce convergence rates, we need a lower bound on the sequence A_t. Unfortunately, the analysis in Nesterov (2015) only works when $p = 2$, and we will thus use a different argument. First, notice that, by the line-search rule,

$$\frac{M_{t+1}}{2} \leq \left[\frac{2}{p}\left(\frac{p-\kappa}{\kappa}\right)\frac{1}{\varepsilon\tau_t}\right]^{\frac{p-\kappa}{\kappa}} L_\kappa^{\frac{p}{\kappa}},$$

from which we obtain

$$\begin{aligned}\alpha_{t+1}^p = \tau_t^p A_{t+1}^p &= \frac{A_{t+1}^{p-1}}{M_{t+1}} \\ &\geq A_{t+1}^{p-1}\frac{1}{2}\left[\frac{p}{2}\left(\frac{\kappa}{p-\kappa}\right)\varepsilon\tau_t\right]^{\frac{p}{\kappa}-1} L_\kappa^{-\frac{p}{\kappa}} \\ &\geq \frac{1}{2}\left[\frac{\varepsilon p}{2}\left(\frac{\kappa}{p-\kappa}\right)\right]^{\frac{p-\kappa}{\kappa}} L_\kappa^{-\frac{p}{\kappa}} A_{t+1}^{p-\frac{p}{\kappa}} \alpha_{t+1}^{\frac{p}{\kappa}-1}.\end{aligned}$$

This allows us to conclude

$$\alpha_{t+1}^{\frac{(p+1)\kappa-p}{\kappa}} \geq \frac{1}{2}\left[\frac{\varepsilon p}{2}\left(\frac{\kappa}{p-\kappa}\right)\right]^{\frac{p-\kappa}{\kappa}} L_\kappa^{-\frac{p}{\kappa}} A_{t+1}^{\frac{p\kappa-p}{\kappa}},$$

which gives an inequality involving α_{t+1} and A_{t+1}:

$$\alpha_{t+1} \geq \left(2^{-\frac{\kappa}{(p+1)\kappa-p}}\left[\frac{\varepsilon p}{2}\left(\frac{\kappa}{p-\kappa}\right)\right]^{\frac{p-\kappa}{(p+1)\kappa-p}} L_\kappa^{-\frac{p}{(p+1)\kappa-p}}\right) A_{t+1}^{\frac{p\kappa-p}{(p+1)\kappa-p}}.$$

Here is where we need to depart from Nesterov's analysis, as the condition $\gamma \geq 1/2$ in that proof does not hold. Instead, we show the following bound.

LEMMA 5.10. *Suppose* $\alpha_t \geq 0$, $\alpha_0 = 0$, *and* $A_t = \sum_{j=0}^t \alpha_j$ *satisfy*

$$\alpha_t \geq \beta A_t^s$$

for some $s \in [0, 1[$ *and* $\beta \geq 0$. *Then,*

$$A_t \geq ((1-s)\beta t)^{\frac{1}{1-s}}$$

for any $t \geq 0$.

Proof. The sequence A_t follows the recursion $A_t - A_{t-1} \geq \beta A_t^s$. The function $h(x) := x - \beta x^s$ satisfies $h(0) = 0$, $h'(0^+) < 0$ and $h'(x)$ only has a single positive root. Hence, when $A_{t-1} > 0$, the equation

$$A_t - \beta A_t^s = A_{t-1}$$

in the variable A_t only has a single positive root, after which $h(A_t)$ is increasing. This means that to get a lower bound on A_t it suffices to consider the extreme case of the sequence satisfying

$$A_t - A_{t-1} = \beta A_t^s.$$

Because A_t is increasing, the sequence $A_t - A_{t-1}$ is increasing, and hence there exists an increasing, convex, piecewise affine function $A(t)$ that interpolates A_t, whose breakpoints are located at integer values of t. By construction, this function $A(t)$ satisfies

$$A'(t) = A_{\lfloor t+1 \rfloor} - A_{\lfloor t \rfloor} = \alpha_{\lfloor t+1 \rfloor} \geq \beta A_{\lfloor t+1 \rfloor}^s \geq \beta A(t)^s$$

for any $t \notin \mathbb{N}$. In particular, the interpolant satisfies

$$A'(t) \geq \beta A^s(t) \tag{29}$$

for any $t \geq 0$. Note that $1/A^s(t)$ is a convergent integral around 0, as $A(t)$ is linear around 0, and $A'(\cdot)$ can be defined as a right continuous nondecreasing function, which is furthermore constant around 0; therefore the involved functions are integrable, and the theorem of change of variables holds. Integrating the differential inequality we get

$$\beta t \leq \int_0^t \frac{A'(t)}{A^s(t)} dt = \int_0^{A(t)} \frac{du}{u^s} = \frac{A(t)^{1-s}}{1-s},$$

yielding the desired result. ∎

Using Lemma 5.10 with $s = (p\kappa - p)/((p+1)\kappa - p)$ produces the following bound on A_T:

$$A_T \geq \frac{1}{2} \left(\frac{\kappa}{(p+1)\kappa - p} \right)^{\frac{(p+1)\kappa - p}{\kappa}} \left(\frac{\varepsilon p}{2} \frac{\kappa}{p - \kappa} \right)^{\frac{p-\kappa}{\kappa}} L_\kappa^{-\frac{p}{\kappa}} T^{\frac{(p+1)\kappa - p}{\kappa}}.$$

To guarantee that $A_T \geq 2D_\Phi(Q)/\varepsilon$, it suffices to impose

$$T \geq C(p,\kappa) \left(\frac{D_\Phi^\kappa(Q) L_\kappa^p}{\varepsilon^p} \right)^{\frac{1}{(p+1)\kappa - p}},$$

where

$$C(p,\kappa) := \left(\frac{(p+1)\kappa - p}{\kappa} \right) \left(\frac{2(p-\kappa)}{p\kappa} \right)^{\frac{p-\kappa}{(p+1)\kappa - p}} 2^{\frac{2\kappa}{(p+1)\kappa - p}}.$$

COROLLARY 5.11. *Let $f \in \mathcal{F}^\kappa_{\|\cdot\|}(Q, L_\kappa)$ and let $\Phi : X \to \mathbb{R}$ be p-uniformly convex w.r.t. $\|\cdot\|$. Then the number of iterations required by Algorithm 3 to produce a solution with accuracy ε is bounded by*

$$T \leq \inf_{1 < \kappa \leq 2} \left[C(p,\kappa) \left(\frac{D_\Phi^\kappa(Q) L_\kappa^p}{\varepsilon^p} \right)^{\frac{1}{(p+1)\kappa - p}} \right].$$

From Corollary 5.11 we obtain the affine-invariant bound on iteration complexity. Given a centrally symmetric convex body $Q \subseteq \mathbb{R}^n$, we choose the norm as its Minkowski gauge $\|\cdot\| = \|\cdot\|_Q$, and p-uniformly convex prox as the minimizer defining the optimal p-variation constant, $\sup_{x \in Q} \Phi(x) - \inf_{x \in Q} \Phi(x) = D_{p,Q}$. With these choices, the iteration complexity is

$$T \leq \inf_{1 < \kappa \leq 2} \left[C(p,\kappa) \left(\frac{D_{p,Q}^\kappa L_{\kappa,Q}^p}{\varepsilon^p} \right)^{\frac{1}{(p+1)\kappa - p}} \right],$$

where $L_{\kappa,Q}$ is the Hölder constant of f quantified in the Minkowski gauge norm $\|\cdot\|_Q$. As a consequence, the bound above is affine-invariant, since $D_{p,Q}$ is affine-invariant by construction. Observe our iteration bound automatically adapts to the best possible weak smoothness parameter $\kappa \in (1, 2]$; note however that an implementable algorithm requires an accuracy certificate in order to stop with this adaptive bound. These details are beyond the scope of this paper, but we refer the reader to (Nesterov, 2015) for details.

It is worth noting at this point that affine invariance of our algorithm comes from the choice of the prox, which minimizes the constant of variation. Unfortunately, we do not have a robust result, as opposed to the strongly convex case, where any bound on the regularity constant provides an affine-invariant method. We consider improving the affine invariance of this algorithm an interesting open problem.

5.4. Example. ℓ_p-balls with $2 \le p < \infty$. In the case in which $2 \le p < \infty$, the function $\Phi_p(w) = \frac{1}{p}\|w\|_p^p$ is p-uniformly convex w.r.t. $\|\cdot\|_p$ (see, e.g., Ball, Carlen, and Lieb (1994)), and thus

$$D_{p,\mathcal{B}_p} = 1 \quad \text{when } p \in [2, \infty].$$

As a consequence, Algorithm 2 with Φ_p guarantees $\varepsilon > 0$ accuracy within

$$T \le \left\lceil C(p,2) \left(\frac{L_p}{\varepsilon}\right)^{\frac{p}{p+2}} \right\rceil \tag{30}$$

iterations, where $C(p,2)$ is a constant only depending on p (which nevertheless diverges as $p \to \infty$). More precisely,

$$C(p,2) = \left(\frac{p+2}{2}\right)\left(\frac{(p-2)}{p}\right)^{\frac{p-2}{p+2}} 2^{\frac{4}{p+2}}.$$

This way, the complexity guarantee admits passage to the limit $p \to \infty$ with a logarithmic extra factor, since

$$C([\log n] - 2, 2) \le \frac{\log n}{2} 2^{\frac{4}{\ln n}},$$

and thus, for $p = \infty$, using the prox $\Phi_{\log n - 2}$ we get an iteration complexity bound

$$T = O\left(\frac{L_p \log n}{\varepsilon}\right).$$

Note however that in this case we can avoid any dimensional dependence by using the much simpler Frank–Wolfe method (Frank and Wolfe, 1956), and this is optimal up to an absolute constant factor, as we will see next.

5.4.1. Lower bounds. We show that the proposed methods are nearly optimal in the large-scale regime. For the range $2 < p \le \infty$, Guzmán and Nemirovski (2015) proved that the class of problems (16), where f is convex and has L_p-Lipschitz continuous gradient w.r.t. $\|\cdot\|_p$, satisfies the following lower bound on accuracy when the number of steps $T \le n$:

$$\Omega\left(\frac{L_p}{\min[p, \log n]\, T^{1+2/p}}\right),$$

which translates into the iteration complexity lower bound

$$\Omega\left(\left(\frac{L_p}{\min[p, \log n]\varepsilon}\right)^{\frac{p}{p+2}}\right)$$

when the accuracy $\varepsilon \ge \Omega(L_p/[\min\{p, \log n\} n^{1+2/p}])$. For fixed $2 \le p < \infty$, this lower bound matches—up to constant factors—our iteration complexity obtained for these setups. For the $p = \infty$ case, our algorithm also turns out to be optimal up to polylogarithmic in dimension factors.

6. Numerical results. We now briefly illustrate the numerical performance of our methods on a simple problem taken from Nesterov (2015). To test the adaptivity of Algorithm 3, we focus on solving the following *continuous Steiner problem*:

$$\min_{\|x\|_2 \leq 1} \sum_{i=1}^{m} \|x - x_i\|_q \tag{31}$$

in the variable $x \in \mathbb{R}^n$, with parameters $x_i \in \mathbb{R}^n$ for $i = 1, \ldots, m$. The parameter $q \in [1, 2]$ controls the Hölder continuity of the objective. We sample the points x uniformly at random in the cube $[0, 1]^n$. We set $n = 50$, $m = 10$, and the target precision $\varepsilon = 10^{-12}$. We compare iterates with the optimum obtained using CVX (Grant, Boyd, and Ye, 2001). The reader can find the results in Figure 1 (see online for color version). We observe that while it is the same algorithm solving the three cases $q = 1, 1.5, 2$, it is significantly faster on smoother problems, as forecast by the adaptive bound in Corollary 5.11.

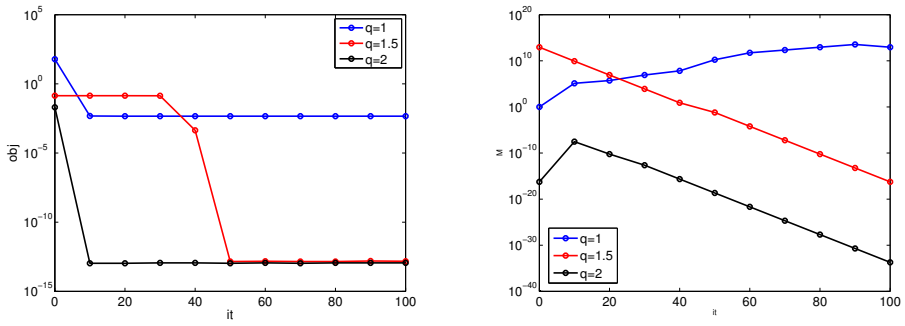

FIG. 1. *We test the adaptivity of Algorithm* 3. *Left: convergence plot of Algorithm* 3 *applied to the continuous Steiner problem* (31) *for* $q = 1, 1.5, 2$. *Right: value of the local smoothness parameter* M *across iterations.*

7. Conclusion. From a practical point of view, the results above offer guidance in the choice of a prox function depending on the geometry of the feasible set Q. On the theoretical side, these results provide affine-invariant descriptions of the complexity of an optimization problem based on both the geometry of the feasible set and of the smoothness of the objective function. In our first algorithm, this complexity bound is written in terms of the regularity constant of the polar of the feasible set and the Lipschitz constant of ∇f with respect to the Minkowski norm. In our last two methods, the regularity constant is replaced by a Bregman diameter constructed from an optimal choice of prox.

When Q is an ℓ_p ball, matching lower bounds on iteration complexity for the algorithm in Nesterov (1983) show that these bounds are optimal in terms of target precision, smoothness, and problem dimension, up to a polylogarithmic term.

However, while we show that it is possible to formulate an affine-invariant implementation of the optimal algorithm in Nesterov (1983), we do not yet show that this is always a good idea outside of the ℓ_p case... In particular, given our choice of norm the constants L_Q and Δ_Q are both affine invariant, with L_Q optimal by our choice of prox function minimizing Δ_Q over all smooth square norms. However, outside of the cases in which Q is an ℓ_p ball, this does not mean that our choice of norm (Minkowski gauge of a centrally symmetric feasible set) minimizes the product $L_Q \min\{\Delta_Q/2, n\}$,

and hence that we achieve the best possible bound for the complexity of the smooth algorithm in Nesterov (1983) and its derivatives. Furthermore, while our bounds give clear indications of what an optimal choice of prox should look like, given a choice of norm, this characterization is not constructive outside of special cases like ℓ_p-balls.

Acknowledgments. Part of this work was done while CG was a postdoc at Centrum Wiskunde & Informatica. We would like to thank Arkadi Nemirovski for valuable discussions, as well as suggesting the reference Nemirovskii and Nesterov (1985) for the uniformly convex setup.

REFERENCES

Z. Allen-Zhu and L. Orecchia, *Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent*, preprint, https://arxiv.org/abs/1407.1537, 2014.

K. Ball, E. Carlen, and E. Lieb, *Sharp uniform convexity and smoothness inequalities for trace norms*, Invent. Math., 115 (1994), pp. 463–482, https://doi.org/10.1007/BF01231769.

S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.

O. Devolder, F. Glineur, and Y. Nesterov, *First-Order Methods of Smooth Convex Optimization with Inexact Oracle*, CORE discussion paper 2011/2, ECORE, Brussels, 2011.

M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist., 3 (1956), pp. 95–110.

M. Grant, S. Boyd, and Y. Ye, *CVX: Matlab Software for Disciplined Convex Programming*, http://cvxr.com/cvx, 2001.

C. Guzmán and A. Nemirovski, *On lower complexity bounds for large-scale smooth convex optimization*, J. Complexity, 31 (2015), pp. 1–14.

J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Grundlehren Math. Wiss. 305, Springer, Berlin, 1993.

A. Juditsky and A. Nemirovski, *Large Deviations of Vector-Valued Martingales in 2-Smooth Normed Spaces*, preprint, https://arxiv.org/abs/0809.0813, 2008.

A. Nemirovskii and Y. E. Nesterov, *Optimal methods of smooth convex minimization*, U.S.S.R. Comput. Math. Math. Phys., 25 (1985), pp. 21–30.

A. S. Nemirovskiĭ and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Nauka, Moscow, 1979 (in Russian).

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Dokl. Math., 27 (1983), pp. 372–376.

Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Springer, Boston, MA, 2003.

Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

Y. Nesterov, *Universal gradient methods for convex optimization problems*, Math. Program., 152 (2015), pp. 381–404.

Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, Stud. Appl. Numer. Math. 13, SIAM, Philadelphia, PA, 1994.

G. Pisier, *Martingales with values in uniformly convex spaces*, Israel J. Math., 20 (1975), pp. 326–350.

N. Srebro, K. Sridharan, and A. Tewari, *On the universality of online mirror descent*, in Adv. Neural Inf. Process. Syst. 24, Curran Associates, Red Hook, NY, 2011, pp. 2645–2653.