[1]

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 [Nesterov, 1983]. 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 [Nesterov, 1983] are thus optimal in terms of target precision, smoothness and problem dimension.

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

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

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 the class of smooth problems [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 algorithm in [Nesterov, 2003] is written

$$\sqrt{\frac{8Ld(x^\star)}{\sigma\varepsilon}}$$

where L is the Lipschitz constant of the gradient, $d(x^\star)$ 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, hence of L and d 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.

Here, we show how to choose an underlying norm and a prox term for the algorithm in [Nesterov, 1983, 2005] which make its iterations and complexity invariant by a change of coordinates. In Section 3, 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

[1]Part of this work was done during a postdoctoral position at Centrum Wiskunde & Informatica
Date: November 8, 2016.

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, we show that this complexity bound is optimal for most, but not all, high dimensional regimes in (n, ε). In Section 4, we thus 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 derived in Section 5 from [Guzmán and Nemirovski, 2015] when optimizing on ℓ_p balls, 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. 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

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

for some $L > 0$. We also *choose a prox function* $d(x)$ for the set Q, i.e. a continuous, strongly convex function on Q with parameter σ (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 center of Q for the prox-function $d(x)$ so that

$$x_0 \triangleq \operatorname*{argmin}_{x \in Q} d(x),$$

assuming w.l.o.g. that $d(x_0) = 0$, we then get in particular

$$d(x) \geq \frac{1}{2} \sigma \|x - x_0\|^2. \tag{3}$$

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

$$T_Q(x) \triangleq \operatorname*{argmin}_{y \in Q} \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2} L \|y - x\|^2 \right\}. \tag{4}$$

We let $y_0 \triangleq T_Q(x_0)$ 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

$$z_t \triangleq \operatorname*{argmin}_{x \in Q} \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^{t} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \right\} \tag{5}$$

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

$$\begin{aligned} x_{t+1} &= \tau_t z_t + (1 - \tau_t) y_t \\ y_{t+1} &= T_Q(x_{t+1}) \end{aligned} \tag{6}$$

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 2.1. **Nesterov [2005].** *Suppose $\alpha_t = (t+1)/2$ with the iterates x_t, y_t and z_t defined in (5) and (6), then for any $t \geq 0$ we have*

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

where $x^\star$ is an optimal solution to problem (1).

If $\varepsilon > 0$ is the target precision, Theorem 2.1 ensures that Algorithm 1 will converge to an ε-accurate solution in no more than

$$\sqrt{\frac{8Ld(x^\star)}{\sigma\varepsilon}} \tag{7}$$

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

Algorithm 1 Smooth minimization.

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

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

Output: $x_T, y_T \in Q$.

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

3. Affine Invariant Implementation

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

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q, \end{array} \quad \text{becomes} \quad \begin{array}{ll} \text{minimize} & \hat{f}(y) \\ \text{subject to} & y \in \hat{Q}, \end{array} \tag{8}$$

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

$$\hat{f}(y) \triangleq f(Ay) \quad \text{and} \quad \hat{Q} \triangleq A^{-1}Q. \tag{9}$$

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 $d(x)$ to get a similar behavior for Algorithm 1.

3.1. Choosing the Norm.

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

Definition 3.1. *Given $Q \subset \mathbb{R}^n$ containing zero, we define the Minkowski gauge of Q as*

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

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 $\|\cdot\|_Q \triangleq \gamma_Q(\cdot)$. 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 3.2. *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 $\hat{f}(w) \triangleq f(Aw)$ 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 the above Lemma 3.2, this can also be seen as a consequence of the well-known duality between smoothness and strong convexity (see e.g. [Hiriart-Urruty and Lemaréchal, 1993, Chap. X, Theorem 4.2.1]).

Theorem 3.3. *Let $f : Q \to \mathbb{R}$ be a convex l.s.c. function. Then f is strongly convex w.r.t. norm $\|\cdot\|$ with constant $\mu > 0$ if and only f^* has Lipschitz continuous gradient w.r.t. norm $\|\cdot\|_*$ with constant $L = 1/\mu$.*

From the previous two results, we immediately have the following lemma.

Lemma 3.4. *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. 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 $\hat{f}(w) \triangleq f(Aw)$ 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.

3.2. **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 3.3 when $(\|\cdot\|_Q^2)^*$ is smooth, we can set $d(x) = \|x\|_Q^2$. The immediate impact of this choice is that the term $d(x^\star)$ in (7) 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 3.5. *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, $x^T y \leq \|y\|^* \|x\|$, 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 $\left(\|y\|^*\right)^2/2$. 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 (4) amounts to taking the conjugate of $\|\cdot\|^2$, so this last result (and its proof) shows that, in the unconstrained case, solving the prox mapping is equivalent to finding a vector aligned with the gradient, with respect to the Minkowski norm $\|\cdot\|_Q$. We now recall another simple result showing that the dual of the norm $\|\cdot\|_Q$ is given by $\|\cdot\|_{Q^\circ}$ where Q° is the polar of the set Q.

Lemma 3.6. *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 the light of the results above, we conclude that whenever Q° is smooth we obtain a natural prox function $d(x) = \|x\|_Q^2$, 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 where 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 $\|.\|_{\mathbb{E}}$.

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

Definition 3.7. *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 3.8. *The* regularity constant *of a Banach space $(\mathbb{E}, \|.\|)$ 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*

$$\|x\|^2 \leq p(x)^2 \leq \frac{\Delta}{\mu}\|x\|^2, \quad \textit{for all } x \in \mathbb{E} \tag{10}$$

hence $d(p(x), \|.\|) \leq \sqrt{\Delta/\mu}$.

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. We are now ready to prove the main result of this section.

Proposition 3.9. *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 Algorithm 1 will produce an ε-solution to problem (1) in at most*

$$\sqrt{\frac{4L_Q\Delta_Q}{\varepsilon}} \tag{11}$$

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 3.8, 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 [Juditsky and Nemirovski, 2008, Prop. 3.2] shows by conjugacy that the prox function $d(x) \triangleq p(x)^2/2$ is strongly convex with respect to the norm $p(x)$ with constant $1/\mu$. Now (10) means that

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

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

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

so $d(x)$ is strongly convex with respect to $\|\cdot\|_Q$ with constant $\sigma = 1/\Delta_Q$, and using (10) as above

$$\frac{d(x^\star)}{\sigma} = \frac{p(x^\star)^2\Delta_Q}{2} \leq \frac{\|x^\star\|_Q^2\Delta_Q}{2} \leq \frac{\Delta_Q}{2}$$

by definition of $\|\cdot\|_Q$, if $x^\star$ is an optimal (hence feasible) solution of problem (1). The bound in (11) then follows from (7) and its affine invariance follows directly from affine invariance of the distortion and Lemmas 3.2 and 3.4. ■

In Section 5, we will see that, when Q is an ℓ_p ball, the complexity bound in (11) is optimal for most, but not all, high dimensional regimes in (n, ε) where n is greater than $\varepsilon^{-1/2}$. In the section that follows, we thus extend Algorithm 1 to much more general problems, optimizing Hölder smooth functions using p-uniformly convex prox functions (not just strongly convex ones). These additional degrees of freedom will allow us to match optimal complexity lower bounds, with adaptivity in the Hölder smoothness parameter as a bonus.

4. Hölder Smooth Functions & Uniformly Convex Prox

We now extend the results of Section 3 to problems where 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 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 motivate 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 prox.

Example 4.1. *Let $2 < p \leq \infty$ and let $\mathcal{B}_p$ be the unit p-ball on $\mathbb{R}^n$. Let $\Psi : \mathbb{R}^n \to \mathbb{R}$ be any strongly convex (prox) function w.r.t. $\|\cdot\|_p$ with constant 1, and suppose w.l.o.g. that $\Psi(0) = 0$. We will prove that*

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

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

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

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

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

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

4.1. Uniform Convexity and Smoothness.

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 $p > 2$. 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 and its dual notion of q-uniform smoothness. For simplicity, we will only consider the case of subdifferentiable convex functions, which suffices for our purposes.

Definition 4.2 (Uniform convexity and uniform smoothness)**.** *Let $2 \leq p < \infty$, $\mu > 0$ and $Q \subseteq \mathbb{R}^n$, a closed convex set. A subdifferentiable function $\Psi : Q \to \mathbb{R}$ is p-uniformly convex with constant μ w.r.t. $\|\cdot\|$ iff for all $x \in \mathring{Q}$, $y \in Q$,*

$$\Psi(y) \geq \Psi(x) + \langle \nabla\Psi(x), y - x\rangle + \frac{\mu}{p}\|y - x\|^p. \tag{12}$$

Now let $1 < q \leq 2$ and $L > 0$. A subdifferentiable function $\Phi : Q \to \mathbb{R}$ is q-uniformly smooth with constant L w.r.t. $\|\cdot\|$ iff for all $x \in \mathring{Q}$, $y \in Q$,

$$\Phi(y) \leq \Phi(x) + \langle \nabla\Phi(x), y - x\rangle + \frac{L}{q}\|y - x\|^q. \tag{13}$$

From now on, whenever the constant μ of p-uniform convexity is not explicitly stated, $\mu = 1$. 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 4.3 (Constant of variation)**.** *Given a p-uniformly convex function $\Psi : \mathbb{R}^n \to \mathbb{R}$, we define its constant of variation on Q as $D_\Psi(Q) \triangleq \sup_{x\in Q} \Psi(x) - \inf_{x\in Q} \Psi(x)$. Furthermore, we define*

$$D_{p,Q} \triangleq \inf_{\Psi} \left\{ \sup_{x\in Q} \Psi(x) \;\middle|\; \Psi : Q \to \mathbb{R}_+ \text{ is } p\text{-uniformly convex w.r.t. } \|\cdot\|,\ \Psi(0) = 0 \right\}. \tag{14}$$

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 4.1 showed that when $2 < p < \infty$, we have $D_{2,\mathcal{B}_p} \geq n^{1-2/p}/2$, and the function $\Psi(x) = \|x\|_2^2/2$ shows this bound is tight. We will later see that $D_{p,\mathcal{B}_p} = 1$, which is a major improvement for large dimensions. [Juditsky and Nemirovski, 2008, Prop. 3.3] also shows that $\Delta_Q \geq D_{2,Q} \geq c\,\Delta_Q$, where Δ_Q is the regularity constant defined in (3.8) and $c > 0$ is an absolute constant, since $\Psi(x)$ 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 (14) above with the notion of martingale cotype. A Banach space $(\mathbb{E}, \|.\|)$ 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 4.4. *[Pisier, 1975] A Banach space $(\mathbb{E}, \|.\|)$ 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 et al., 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 4.5 (Bregman divergence)**.** *Let $(\mathbb{R}^n, \|\cdot\|)$ be a normed space, and $\Psi : Q \to \mathbb{R}$ be a p-uniformly convex function w.r.t. $\|\cdot\|$. We define the Bregman divergence as*

$$V_x(y) \triangleq \Psi(y) - \langle \nabla\Psi(x), y - x\rangle - \Psi(x) \quad \forall x \in \mathring{Q},\ \forall y \in Q.$$

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 4.6. *Let $\Psi : Q \to \mathbb{R}$ 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

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

which is the desired result. ∎

4.2. **An Accelerated Method for Minimizing Hölder Smooth Functions.** We consider classes of weakly-smooth convex functions. For a Hölder exponent $\sigma \in (1, 2]$ we denote the class $\mathcal{F}^{\sigma}_{\|\cdot\|}(Q, L_\sigma)$ as the set of convex functions $f : Q \to \mathbb{R}$ such that for all $x, y \in Q$

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

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

$$\gamma_{t+1}^p = \gamma_{t+1}^{p-1} + \gamma_t^p.$$

This sequence has the following properties.

Proposition 4.7. *The following properties hold for the auxiliary sequence* $(\gamma_t)_{t\geq 0}$

(i) *The sequence is increasing.*
(ii) $\gamma_t^p = \sum_{s=1}^t \gamma_s^{p-1}$.
(iii) $\frac{t}{p} \leq \gamma_t \leq t$.
(iv) $\sum_{s=1}^t \gamma_s^p \leq t\gamma_t^p$.

Proof. We get

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

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

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 terms of $\|\cdot\|^p$. This idea is a minor extension of the "inexact gradient trick" proposed in [Devolder et al., 2011] and further studied in [Nesterov, 2015], which corresponds to the special case where $p = 2$ in the results described here. As in [Devolder et al., 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 4.8. *Let* $f \in \mathcal{F}^{\sigma}_{\|\cdot\|}(Q, L_\sigma)$*, then for any* $\delta > 0$ *and*

$$M \geq \left[\frac{2}{p}\left(\frac{p-\sigma}{\sigma}\right)\frac{1}{\delta}\right]^{\frac{p-\sigma}{\sigma}} L_\sigma^{\frac{p}{\sigma}} \tag{15}$$

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 $x, y \in Q$

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

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

$$\frac{L_\sigma}{\sigma} t^\sigma \leq \frac{M}{p} t^p + \frac{\delta}{2}. \tag{16}$$

This can be seen by letting $t = \|y - x\|$ and using (16) in the preceding inequality. Let us prove (16). 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 $r = p/\sigma$, $s = p/(p - \sigma)$ and $x = t^\sigma$, we obtain

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

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

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

Finally, by our choice of M we have that $M \geq L_\sigma / y$, proving (16) and therefore the result. ■

Algorithm 2 Accelerated Method with Bregman Prox

Input: $x_0 \in Q$

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

$$y_{t+1} = \arg\min_{y \in Q} \left\{ \frac{M}{p} \|y - x_{t+1}\|^p + \langle \nabla f(x_{t+1}), y - x_{t+1} \rangle \right\} \tag{17}$$

$$z_{t+1} = \arg\min_{z \in Q} \{ V_{z_t}(z) + \alpha_{t+1} \langle \nabla f(x_{t+1}), z - z_t \rangle \} \tag{18}$$

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 4.9. *Let $f \in \mathcal{F}^\sigma_{\|\cdot\|}(Q, L_\sigma)$ and $\Psi : Q \to \mathbb{R}$ be p-uniformly convex w.r.t. $\|\cdot\|$. Then for any $\varepsilon > 0$, setting $\delta \triangleq \varepsilon / T$, and M satisfying (15), the accelerated method in Algorithm 2 guarantees an accuracy*

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

after T iterations.

Proof. Let $u \in Q$ be an arbitrary vector. Using the optimality conditions for subproblem (18), and Lemma 4.6, we get

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

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 4.7 (ii),

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

From this we obtain

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

where the first inequality holds by the definition of y_{t+1}, and the last inequality holds by Lemma 4.8 and the choice of M. This means that

$$\alpha_{t+1}\langle \nabla f(x_{t+1}), z_t - u\rangle \leq A_{t+1}\left[f(x_{t+1}) - f(y_{t+1}) + \delta/2\right] + V_{z_t}(u) - V_{z_{t+1}}(u). \tag{19}$$

From (19) and other simple estimates

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

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 inequlities, 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 4.7, we have $\frac{1}{A_T}\sum_{t=1}^{T} A_t \leq \frac{1}{\gamma_T^p} T\gamma_T^p \leq T$, thus by the choice $\delta = \varepsilon/T$, we obtain

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

Definition 4.5 together with the fact that $\langle \nabla\Psi(x), y - x\rangle \geq 0$ when x minimizes $\Psi(x)$ over Q then yields the desired result. ■

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 (15) is satisfied with equality. Since $A_T = \gamma_T^p/M$ we can use Proposition 4.7 (iii), so that

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

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

Corollary 4.10. *Let $f \in \mathcal{F}^\sigma_{\|\cdot\|}(Q, L_\sigma)$ and $\Psi : X \to \mathbb{R}$ be p-uniformly convex w.r.t. $\|\cdot\|$. Setting $\delta \triangleq \varepsilon/T$, and M satisfying (15), the accelerated method in Algorithm 2 requires*

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

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 σ (the smoothness) and L_σ which sets the step-size. We now focus on designing an adaptive step-size policy, that does not require L_σ as input, and adapts itself to the best weak smoothness parameter $\sigma \in (1, 2]$.

4.3. **An Adaptive Gradient Method.** We will now extend the adaptive algorithm in [Nesterov, 2015, Th.3] 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

$$A^{p-1} = \alpha^p M \tag{21}$$

$$\frac{1}{\tau^p} = MA. \tag{22}$$

These identities are analogous to the ones derived for the non-adaptive 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 4.8, but let us make this intuition precise. From (21) and (22) we have

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

Notice that whenever the condition (20) of Algorithm 3 is not satisfied, M is increased by a factor two. Suppose the line-search does not terminate, then $\alpha \to 0$. However, by Lemma 4.8, the termination condition (20) is guaranteed to be satisfied as soon as

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

Algorithm 3 Accelerated Method with Bregman Prox and Adaptive Stepsize

Input: $x^0 \in Q$

Set $y_0 = x_0$, $z_0 = x_0$, $M_0 = 1$ and $A_0 = 0$.
for $t = 0, \ldots, T-1$ **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

$$f(y_{t+1}) \leq f(x_{t+1}) + \langle \nabla f(x_{t+1}), y_{t+1} - x_{t+1} \rangle + \frac{M}{p} \|y_{t+1} - x_{t+1}\|^p + \frac{\tau\varepsilon}{2} \tag{20}$$

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

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

end for
return y

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 $p = 2$, we will thus use a different argument. First, notice that by the line-search rule

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

from which we obtain

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

This allows us to conclude

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

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

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

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 4.11. *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 $h(x) \triangleq x - \beta x^s$ 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, 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

$$A'(t) \geq \beta A^s(t) \tag{23}$$

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 4.11 with $s = (p\sigma - p)/((p+1)\sigma - p)$ produces the following bound on A_T

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

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

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

where

$$C(p,\sigma) \triangleq \left(\frac{(p+1)\sigma - p}{\sigma} \right) \left(\frac{2(p-\sigma)}{p\sigma} \right)^{\frac{p-\sigma}{(p+1)\sigma-p}} 2^{\frac{2\sigma}{(p+1)\sigma-p}}.$$

Corollary 4.12. *Let $f \in \mathcal{F}^{\sigma}_{\|\cdot\|}(Q, L_\sigma)$ and $\Psi : X \to \mathbb{R}$ is 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*

$$T \leq \inf_{1<\sigma\leq 2} \left[C(p,\sigma) \left(\frac{D^{\sigma}_{\Psi}(Q) L^{p}_{\sigma}}{\varepsilon^{p}} \right)^{\frac{1}{(p+1)\sigma - p}} \right].$$

From Corollary 4.12 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, $\sup_{x\in Q} \Psi(x) = D_{p,Q}$. With these choices, the iteration complexity is

$$T \leq \inf_{1<\sigma\leq 2} \left[C(p,\sigma) \left(\frac{D^{\sigma}_{p,Q} L^{p}_{\sigma,Q}}{\varepsilon^{p}} \right)^{\frac{1}{(p+1)\sigma - p}} \right],$$

where $L_{\sigma,Q}$ is the Hölder constant of f quantified in the Minkowski gauge norm $\|\cdot\|_Q$. As a consequence, the bound above is affine-invariant, since also $D_{p,Q}$ is affine-invariant by construction. Observe our iteration bound automatically adapts to the best possible weak smoothness parameter $\sigma \in (1, 2]$; 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 to [Nesterov, 2015] for details. Finally, we will see in what follows that this affine invariant bound also matches corresponding lower bounds when Q is an ℓ_p ball.

5. Explicit Bounds on Problems over ℓ_p Balls

5.1. **Upper Bounds.** To illustrate our results, first consider the problem of minimizing a smooth convex function over the unit simplex, written

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \mathbf{1}^T x \leq 1,\ x \geq 0, \end{array} \tag{24}$$

in the variable $x \in \mathbb{R}^n$.

As discussed in [Juditsky et al., 2009, §3.3], choosing $\|\cdot\|_1$ as the norm and $d(x) = \log n + \sum_{i=1}^{n} x_i \log x_i$ as the prox function, we have $\sigma = 1$ and $d(x^\star) \leq \log n$, which means the complexity of solving (24) using Algorithm 1 is bounded by

$$\sqrt{8 \frac{L_1 \log n}{\varepsilon}} \tag{25}$$

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 $d(x^\star)$ 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 the construction of Section 3. The simplex $C = \{x \in \mathbb{R}^n : \mathbf{1}^T x \leq 1, x \geq 0\}$ is not centrally symmetric, but we can symmetrize it as the ℓ_1 ball. The Minkowski norm associated with that set is then equal to the ℓ_1-norm, so $\|\cdot\|_Q = \|\cdot\|_1$ here. The space $(\mathbb{R}^n, \|\cdot\|_\infty)$ is $2\log n$ regular [Juditsky and Nemirovski, 2008, Example 3.2] with the prox function chosen here as $\|\cdot\|^2_\alpha/2$, with $\alpha = 2\log n/(2\log n - 1)$. Proposition 3.9 then shows that the complexity bound we obtain using this procedure is identical to that in (25). A similar result holds in the matrix case.

5.1.1. *Strongly Convex Prox.* We can generalize this result to all cases where Q is an l_p ball. When $p \in [1, 2]$, [Juditsky et al., 2009, Ex. 3.2] 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

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{B}_p \end{array} \tag{26}$$

in the variable $x \in \mathbb{R}^n$, where $\mathcal{B}_p$ is the ℓ_p ball, using Algorithm 1, is bounded by

$$\sqrt{\frac{4L_p\Delta_p}{\varepsilon}} \tag{27}$$

where L_p is the Lipschitz constant of ∇f with respect to the ℓ_p norm. We will later see that this bound is nearly optimal when $1 \leq p \leq 2$; however, the dimension dependence on the bounds when $p > 2$ is essentially suboptimal. In order to obtain the optimal methods in this range we will need our p-uniformly convex extensions.

5.1.2. *Uniformly Convex Bregman Prox.* In the case $2 \leq p < \infty$, the function $\Psi_p(w) = \frac{1}{p}\|w\|_p^p$ is p-uniformly convex w.r.t. $\|\cdot\|_p$ (see, e.g. [Ball et al., 1994]), and thus

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

As a consequence, Algorithm 2 with Ψ_p as p-uniformly requires

$$T \geq C(p)\left(\frac{L_p}{\varepsilon}\right)^{\frac{p}{p+2}} \tag{28}$$

iterations to reach a target precision ε, where $C(p)$ is a constant only depending on p (which nevertheless diverges as $p \to \infty$). This complexity guarantee admits passage to the limit $p \to \infty$ with a poly-logarithmic extra factor. Note however that in this case we can avoid any dimension dependence by the much simpler Frank-Wolfe method.

5.2. **Lower Bounds.** We show that in the case of ℓ_p balls estimates from the proposed methods are nearly optimal in terms of information-based complexity. We consider the class of problems given by the minimization of smooth convex objectives with a bound L_p on the Lipschitz constant of their gradients w.r.t. norm $\|\cdot\|_p$, and the feasible domain given the radius $R_p > 0$ ball $\mathcal{B}_p(R)$. We emphasize that the lower bounds we present only hold for the large-scale regime, where the number of iterations T is upper bounded by the dimension of the space, n. It is well-known that when one can afford a super-linear (in dimension) number of iterations, methods such as the center of gravity or ellipsoid can achieve better complexity estimates [Nemirovskii and Yudin, 1979].

First, in the range $1 \leq p \leq 2$ we can immediately use the lower bound on risk from [Guzmán and Nemirovski, 2015],

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

where T is the number of iterations, which translates into the following lower bound on iteration complexity

$$\Omega\left(\sqrt{\frac{L_p R_p^2}{\varepsilon \log n}}\right)$$

as a function of the target precision $\varepsilon > 0$. Therefore, the affine invariant algorithm is optimal, up to poly-logarithmic factors, in this range.

For the second range, $2 < p \leq \infty$, the lower bound states the accuracy after T steps is no better than

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

which translates into the iteration complexity lower bound

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

For fixed $2 \leq p < \infty$, this lower bound matches –up to constant factors– our iteration complexity obtained for these setups. For the case $p = \infty$, our algorithm also turns out to be optimal, up to polylogarithmic in the 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*

$$\min_{\|x\|_2 \leq 1} \sum_{i=1}^{m} \|x - x_i\|_q \tag{29}$$

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 et al., 2001]. We observe that while the algorithm solving the three cases $q = 1, 1.5, 2$ is identical, it is significantly faster on smoother problems, as forecast by the adaptive bound in Corollary 4.12.

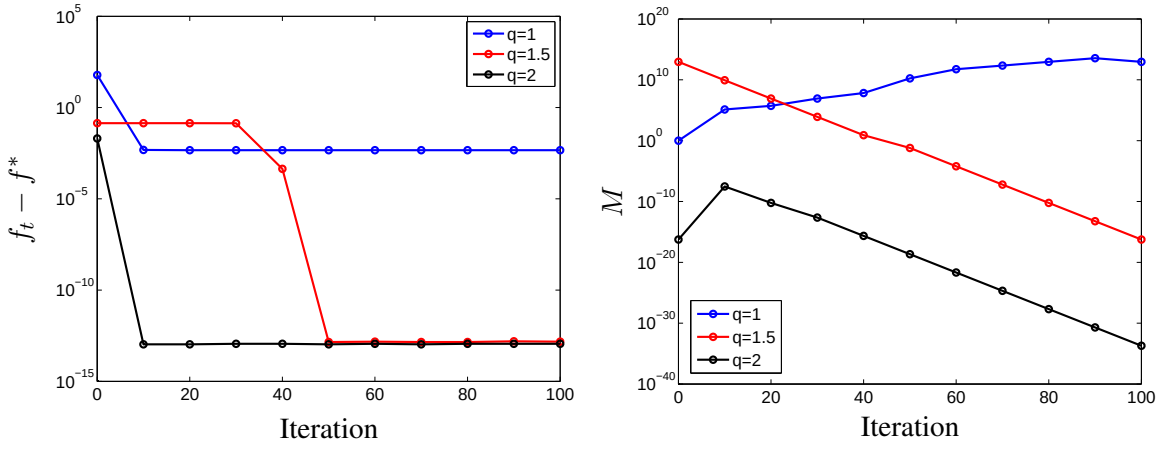

FIGURE 1. We test the adaptivity of of Algorithm 3. *Left:* Convergence plot of Algorithm 3 applied to the continuous Steiner problem (29) for $q = 1, 1.5, 2$. *Right:* Value of the local smoothness parameter M across iterations.

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 where 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 $L_Q \min\{\Delta_Q/2, n\}$, 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

AA and MJ would like to acknowledge support from the European Research Council (project SIPA). MJ also acknowledges support from the Swiss National Science Foundation (SNSF). The authors would also 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.

References

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

K. Ball, E.A. Carlen, and E.H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994. doi: 10.1007/BF01231769.

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

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *CORE Discussion Papers,(2011/02)*, 2011.

M. Grant, S. Boyd, and Y. Ye. CVX: Matlab software for disciplined convex programming. 2001.

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

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer, 1993.

A. Juditsky and A.S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro. Stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

A. Nemirovskii and D. Yudin. Problem complexity and method efficiency in optimization. *Nauka (published in English by John Wiley, Chichester, 1983)*, 1979.

AS Nemirovskii and Yu E Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985.

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

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer, 2003.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.

Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2): 381–404, 2015.

Gilles Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20(3-4):326–350, 1975.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011.*, pages 2645–2653, 2011.

CNRS & D.I., UMR 8548,
ÉCOLE NORMALE SUPÉRIEURE, PARIS, FRANCE.
E-mail address: aspremon@ens.fr

FACULTAD DE MATEMÁTICAS & ESCUELA DE INGENIERÍA, PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
E-mail address: crguzmanp@uc.cl

ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, SWITZERLAND.
E-mail address: m.jaggi@gmail.com