Faster Acceleration for Steepest Descent

Cedar Site Bai BAI123@PURDUE.EDU
Department of Computer Science
Purdue University

Brian Bullins BBULLINS@PURDUE.EDU
Department of Computer Science
Purdue University

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Recent advances (Sherman, 2017; Sidford and Tian, 2018; Cohen et al., 2021) have overcome the fundamental barrier of dimension dependence in the iteration complexity of solving ℓ_∞ regression with first-order methods. Yet it remains unclear to what extent such acceleration can be achieved for general ℓ_p smooth functions. In this paper, we propose a new accelerated first-order method for convex optimization under non-Euclidean smoothness assumptions. In contrast to standard acceleration techniques, our approach uses primal-dual iterate sequences taken with respect to *differing* norms, which are then coupled using an *implicitly* determined interpolation parameter. For ℓ_p norm smooth problems in d dimensions, our method provides an iteration complexity improvement of up to $O(d^{1-\frac{2}{p}})$ in terms of calls to a first-order oracle, thereby allowing us to circumvent long-standing barriers in accelerated non-Euclidean steepest descent.

Keywords: First-order acceleration, convex optimization, non-Euclidean smoothness, steepest descent

1. Introduction

Large-scale optimization tasks are a central part of modern machine learning, and many of the algorithms that find success in training these models, such as SGD (Robbins and Monro, 1951), AdaGrad (Duchi et al., 2011), and Adam (Kingma and Ba, 2014), among others, build on classic approaches in (convex) optimization. One prominent example is that of *momentum* (Polyak, 1964), and the related acceleration technique of Nesterov (1983), which use both current and previous (accumulated) gradient information to accelerate beyond the basic gradient descent method. For smooth, convex problems, this *accelerated gradient descent* (AGD) method converges (in terms of optimality gap) at a rate of $O(1/T^2)$, which improves upon the basic gradient descent rate of $O(1/T)$, and this rate is furthermore known to be tight due to matching lower bounds (Nesterov, 2018).

The general iterative scheme when moving from x_t to x_{t+1} for heavy ball momentum (with parameters $\alpha > 0$, $\beta \in [0, 1]$) is given as

$$x_{t+1} = x_t - \alpha \nabla f(x_t) - \beta(x_t - x_{t-1})$$

while the iterations of accelerated gradient descent can be expressed as

$$\begin{aligned} y_{t+1} &= x_t - \alpha \nabla f(x_t) \\ x_{t+1} &= y_{t+1} - \beta(y_{t+1} - y_t). \end{aligned}$$

Crucially, there is a natural (Euclidean) interpretation of the trajectory of these updates, whereby each gradient step is slightly "pushed" in the direction of $x_{t-1} - x_t$ (respectively, $y_t - y_{t+1}$), with the amount of "force" applied depending on the choice of β. Indeed, this structure leads to an analysis whose final rate of convergence has a Euclidean-based (ℓ_2 norm) dependence on the smoothness parameter, as well as the initial distance to the minimizer.

Nesterov (2005) later generalized these techniques to non-Euclidean settings, introducing an *estimate sequence* approach to acceleration, whose iterates are interpolations of two additional iterate sequences: one given by a steepest descent update (in the appropriate norm), and another given by the minimizer of a function that comprises a term linear in the accumulated gradients (which provide, in a sense, a certain *dual* characterization, as also found in the basic analysis for, e.g., mirror descent (Nemirovski and Yudin, 1983) and regret minimization (Hazan, 2016)) along with the Bregman divergence of a distance generating function. Other interpretations of acceleration have since been presented (e.g., (Bubeck et al., 2015)), including the linear coupling framework of Allen-Zhu and Orecchia (2017), which views acceleration as a certain *coupling* between gradient (steepest) descent and mirror descent updates, each step of which takes the form (for some $\alpha, \gamma > 0$, $\beta \in [0, 1]$):

$$
\begin{aligned}
x_{t+1} &= \beta z_t + (1-\beta) y_t \\
y_{t+1} &= \arg\min_{y \in \mathbb{R}^d} \left\{ \langle \nabla f(x_{t+1}),\, y - x_{t+1} \rangle + \frac{1}{2\alpha} \|y - x_{t+1}\|^2 \right\} \\
z_{t+1} &= \arg\min_{z \in \mathbb{R}^d} \left\{ \gamma \langle \nabla f(x_{t+1}),\, z - z_t \rangle + V_{z_t}(z) \right\},
\end{aligned}
$$

where $V_x(y)$ is the Bregman divergence with respect to the (distance generating) function $\phi(\cdot)$, i.e., $V_x(y) := \phi(y) - \phi(x) - \langle \nabla \phi(x),\, y - x \rangle$, and all of which further suggests a natural primal-dual interpretation of acceleration.

A notable use of these approaches occurs when optimizing over the simplex, in which case the fact that the Bregman divergence of the negative entropy function is strongly convex w.r.t. ℓ_1 norm (via Pinsker's inequality) suffices to yield the desired accelerated rate (see, e.g., Appendix A in (Allen-Zhu and Orecchia, 2017) for further details), and it even provides a natural means of deriving the (smooth) softmax approximation (Nesterov, 2005; Beck and Teboulle, 2012).

Turning to the problem of ℓ_∞ regression (Kelner et al., 2014) or its softmax approximation (Nesterov, 2005), which is smooth with respect to the ℓ_∞ norm, a key challenge arises. Specifically, while the algorithm requires $\phi(\cdot)$ to be strongly convex w.r.t. the ℓ_∞ norm (for the convergence guarantees to hold), *any such $\phi(\cdot)$ that satisfies this condition will have a range of at least $O(d)$* (Sidford and Tian, 2018, Appendix A.1). While various approaches have been proposed (Sherman, 2017; Sidford and Tian, 2018; Cohen et al., 2021) to overcome this fundamental barrier of dimension dependence in the iteration complexity in this special case of $p = \infty$, it remains unclear, in the more general ℓ_p-smooth setting for $p > 2$, to what extent any such acceleration is possible without breaking the known lower bound of $\Omega(L \|x^*\|_p^2 / T^{\frac{p+2}{p}})$ (Guzmán and Nemirovski, 2015).

Our contributions. In this work, we aim to circumvent this barrier in general by providing a faster accelerated non-Euclidean steepest descent method, called Hyper-Accelerated Steepest Descent (HASD) (Algorithm 1), that improves upon previous results in accelerated steepest descent (Nesterov, 2005; Allen-Zhu and Orecchia, 2017) by a factor of up to $O(d^{1-\frac{2}{p}})$, in terms of calls to a first-order oracle, where d is the problem dimension. (The full presentation of our convergence guarantees may be found in Theorem 4.) Our approach is based on a similar estimate sequence-type approaches as in (Nesterov, 2005), though with a key difference: rather than setting the interpolation parameter as a (fixed) function of the iteration index t, we instead choose the parameter implicitly, in a manner depending on *local* properties of the function (specifically, the gradient at the subsequent iterate, which itself depends on the choice of parameter).

In this way, our approach exhibits certain similarities with that of the (accelerated) HPE framework (Monteiro and Svaiter, 2010, 2013), though *we emphasize a crucial difference in the breaking of primal-dual symmetry and its eventual adjustment*. (We refer the reader to Section 5 for a detailed discussion of this matter.) We further believe our results complement the view of A-HPE as (approximate) proximal point acceleration (e.g., (Carmon et al., 2020)) by offering a more general perspective in terms of primal-dual asymmetry, the algorithmic “compensation” for which leads to improved convergence guarantees. In addition, our analysis offers a principled framework for analyzing the oracle complexity of minimizing smooth convex functions under *nonstandard geometries*—where the regularity of the objective is measured in norms *differing from* the feasible domain—a longstanding open problem posed in (Guzmán, 2015).

1.1. Related work

Steepest descent and acceleration. The (unnormalized) steepest descent direction of f at $x \in \mathbb{R}^d$ w.r.t. a general norm $\|\cdot\|$ is given by $\delta_{\text{sd}}(x) := \|\nabla f(x)\|_* \delta_{\text{nsd}}(x)$, where we define $\delta_{\text{nsd}}(x) := \arg\min_{v:\|v\|\leq 1} \nabla f(x)^\top v$ (Boyd and Vandenberghe, 2004), and so it follows that the gradient descent direction is a special case when taken w.r.t. the Euclidean norm, i.e., $\|\cdot\|_2$. Under appropriate L-smoothness assumptions w.r.t. $\|\cdot\|_2$, gradient descent (initialized at $x_0 \in \mathbb{R}^d$) may be further accelerated (Nesterov, 1983), improving the rate of convergence from $O(L\|x_0 - x^*\|_2^2/T)$ to $O(L\|x_0 - x^*\|_2^2/T^2)$, whereby the latter matches known lower bounds (Nesterov, 2018). Meanwhile, steepest descent w.r.t. $\|\cdot\|_p$ can be shown (under L-smoothness w.r.t. $\|\cdot\|_p$) to converge at a rate of $O(LR_p^2/T)$ (e.g., (Kelner et al., 2014)), where R_p represents a bound (in terms of $\|\cdot\|_p$) on the diameter of the problem, thereby functioning in a manner similar to the $\|x_0 - x^*\|_2^2$ term in the gradient descent rates.

Accelerated first-order methods have since been extended to non-Euclidean smoothness settings (Nesterov, 2005) (see also (Allen-Zhu and Orecchia, 2017) for further details). As discussed, however, the techincal requirements of these approaches lead to $\|\cdot\|_2$ dependence in the problem diameter for $p > 2$. On the other hand, previous work by Nemirovskii and Nesterov (1985) has shown how to achieve convergence rates of $O(LR_p^2/T^{\frac{p+2}{p}})$ for $p \geq 2$ (which trades off between the norm measuring the diameter and the power of T), and these are also known to be tight (Guzmán and Nemirovski, 2015; Diakonikolas and Guzmán, 2024).

Also of note is the appearance of (momentum-based) steepest descent methods in the context of deep learning and training Large Language Models (Bernstein et al., 2018; Balles et al., 2020; Chen et al., 2023), for which we believe our results may provide an alternative (practical) viewpoint to acceleration for steepest descent methods. (We discuss this topic further in Section 5.)

Optimal higher-order acceleration. Although in this work we are primarily interested in first-order methods, some aspects of our technical contributions share similarities with recent advances in acceleration for *higher-order* methods, i.e., methods which employ derivative information beyond first-order. We first recall that cubic regularization (Nesterov and Polyak, 2006) was shown to be amenable to (generalized) acceleration techniques (Nesterov, 2008), and such techniques extend to k^{th}-order methods (that is, methods which involve minimizing a regularized k^{th} order Taylor expansion), for $k > 2$ (Baes, 2009), achieving a rate of $O(1/T^{k+1})$. This has since been improved to $O(1/T^{\frac{3k+1}{2}})$ (Monteiro and Svaiter, 2013; Gasnikov et al., 2019) (the key idea behind which we discuss in Section 5), and furthermore these rates have been shown to be tight (Agarwal and Hazan, 2018; Arjevani et al., 2019).

1.2. Outline

We begin by establishing our setting and assumptions in Section 2, along with a general discussion of the particulars that occur when working with non-Euclidean norms. In Section 3, we present our main algorithm (Algorithm 1), as well as the key convergence guarantees (including Theorem 4) and their proofs. Section 4 presents natural extensions of our approach to additional settings, such as strongly convex and gradient norm minimization. Finally, we conclude with a discussion of our method as well as the opportunities presented for future work, in Sections 5 and 6, respectively.

2. Preliminaries

In this work, we consider the unconstrained convex minimization problem:

$$\min_{x\in\mathbb{R}^d} f(x), \tag{1}$$

where we let x^* denote the minimizer of f. Letting $\|\cdot\|_p$ and $\|\cdot\|_{p^*}$ denote the standard ℓ_p norm and its dual norm, respectively, we are interested in the case where f is *L-smooth* w.r.t. $\|\cdot\|_p$, i.e., $\forall\, x, y \in \mathbb{R}^d$,

$$\|\nabla f(y) - \nabla f(x)\|_{p^*} \le L\,\|y - x\|_p \quad . \tag{2}$$

Throughout, we specify $\|\cdot\|_2$ when referring to the Euclidean norm, and we may observe that (2) captures the standard (Euclidean) notion of smoothness for $p = 2$. We will consider only the case where $p \ge 2$, and we further use the notation $x[i]$ to refer to the i^{th} coordinate of x.

In addition, we say f is *μ-strongly convex* w.r.t. $\|\cdot\|_p$ if, for all $x, y \in \mathbb{R}^d$,

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \ge \frac{\mu}{2}\|y - x\|_p^2 \quad . \tag{3}$$

2.1. Comparing smoothness parameters

It is useful to observe that L-smoothness w.r.t. $\|\cdot\|_p$ implies L-smoothness w.r.t. $\|\cdot\|_q$ for $2 \le q \le p$, since, by standard norm inequalities,

$$\|\nabla f(y) - \nabla f(x)\|_{q^*} \le \|\nabla f(y) - \nabla f(x)\|_{p^*} \le L\,\|y - x\|_p \le L\,\|y - x\|_q \quad . \tag{4}$$

On the other hand, L-smoothness w.r.t. $\|\cdot\|_q$ (for $2 \le q \le p$) implies $(d^{\frac{2}{q}-\frac{2}{p}}L)$-smoothness w.r.t. $\|\cdot\|_p$, i.e.,

$$\|\nabla f(y) - \nabla f(x)\|_{p^*} \le d^{\frac{2}{q}-\frac{2}{p}} L\,\|y - x\|_p\,,$$

since

$$\frac{1}{d^{\frac{1}{q}-\frac{1}{p}}}\|\nabla f(y) - \nabla f(x)\|_{p^*} \le \|\nabla f(y) - \nabla f(x)\|_{q^*} \le L\,\|y - x\|_q \le d^{\frac{1}{q}-\frac{1}{p}} L\,\|y - x\|_p\,.$$

3. Main results

In this section, we present the main results of our paper, starting with our algorithm.

3.1. Algorithm

Our algorithm, called Hyper-Accelerated Steepest Descent (HASD) (Algorithm 1), is inspired by the estimate sequence-based approaches to acceleration (e.g., (Nesterov, 2005, 2008)). The key difference to observe, however, is the addition of the (simultaneous) finding of ρ_t and x_{t+1} such that the conditions outlined in the algorithm hold. We would further note that a line search procedure similar to (Bubeck et al., 2019) can be used to find such a satisfying pair of ρ_t and x_{t+1}, although the implicit relationship between ρ_t and x_{t+1} is (crucially) different from that in, for example, (Monteiro and Svaiter, 2013; Bubeck et al., 2019), which requires some technical modifications we later elaborate in Section 3.3.

Algorithm 1 Hyper-Accelerated Steepest Descent (HASD)

```
Input: $x_0 \in \mathbb{R}^d$, $A_0 = 0$. Define $\psi_0(x) := \frac{1}{2}\|x - x_0\|_2^2$.
for $t = 0$ to $T - 1$ do
    $v_t = \arg\min_{x \in \mathbb{R}^d} \psi_t(x)$
    Determine $\rho_t > 0$, $x_{t+1} \in \mathbb{R}^d$ for which the following hold simultaneously:
        • $\frac{1}{2}\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2} \leq \rho_t \leq 2\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2}$
        • $a_{t+1} > 0$ s.t. $a_{t+1}^2 = \frac{(A_t + a_{t+1})}{18L\rho_t}$
        • Set $A_{t+1} = A_t + a_{t+1}$, $\tau_t = \frac{a_{t+1}}{A_{t+1}}$
        • $y_t = (1 - \tau_t)x_t + \tau_t v_t$
        • $x_{t+1} = \arg\min_{x \in \mathbb{R}^d}\{\langle \nabla f(y_t), x - y_t \rangle + L\|x - y_t\|_p^2\}$
    $\psi_{t+1}(x) = \psi_t(x) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle]$
end for
Output: $x_T$
```

3.2. Convergence results

We now establish the main theoretical guarantees of our work. To begin, we show the following lemma, which establishes our basic guarantee, analogous to the standard progress guarantee, as in the case of smooth optimization. Rather than measuring how much progress we make in a single step, however, we require a bound on an term relating to the gradient *at the subsequent point*.

Lemma 1 *Let f be L-smooth w.r.t. $\|\cdot\|_p$. Consider the update:*

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^d}\{\langle \nabla f(y_t), x - y_t \rangle + L\|x - y_t\|_p^2\}.$$

Then, x_{t+1} can be expressed in closed form as

$$x_{t+1} = y_t - \frac{1}{2L}\|\nabla f(y_t)\|_{p^*}^{\frac{p-2}{p-1}} g_t \quad \textit{where } g_t[i] := \frac{\nabla f(y_t)[i]}{|\nabla f(y_t)[i]|^{\frac{p-2}{p-1}}} \; \forall i \in \{1, \ldots, d\}. \tag{5}$$

In addition, we have that

$$\langle \nabla f(x_{t+1}),\, y_t - x_{t+1} \rangle \geq L \left\| x_{t+1} - y_t \right\|_p^2 \geq \frac{1}{9L} \left\| \nabla f(x_{t+1}) \right\|_{p^*}^2 \ . \tag{6}$$

Proof First, we note that, by first-order optimality conditions, for all $i \in \{1, \ldots, d\}$,

$$\nabla f(y_t)[i] = -2L \left\| x_{t+1} - y_t \right\|_p^{2-p} \left| x_{t+1}[i] - y_t[i] \right|^{p-2} (x_{t+1}[i] - y_t[i]). \tag{7}$$

It may be checked by verification that $x_{t+1} - y_t = -\frac{1}{2L} \left\| \nabla f(y_t) \right\|_{p^*}^{\frac{p-2}{p-1}} g_t$ from the update in Eq. (5) indeed satisfies the optimality condition. Furthermore,

$$\begin{aligned}
\langle \nabla f(x_{t+1}), &y_t - x_{t+1} \rangle = \langle \nabla f(x_{t+1}) - \nabla f(y_t) + \nabla f(y_t),\, y_t - x_{t+1} \rangle \\
&= \langle \nabla f(x_{t+1}) - \nabla f(y_t),\, y_t - x_{t+1} \rangle + \langle \nabla f(y_t),\, y_t - x_{t+1} \rangle \\
&= \langle \nabla f(x_{t+1}) - \nabla f(y_t),\, y_t - x_{t+1} \rangle \\
&\qquad + 2L \left\| x_{t+1} - y_t \right\|_p^{2-p} \sum_{i=1}^{d} \left| x_{t+1}[i] - y_t[i] \right|^{p-2} (x_{t+1}[i] - y_t[i])^2 \\
&= \langle \nabla f(x_{t+1}) - \nabla f(y_t),\, y_t - x_{t+1} \rangle + 2L \left\| x_{t+1} - y_t \right\|_p^{2-p} \left\| x_{t+1} - y_t \right\|_p^{p} \\
&= \langle \nabla f(x_{t+1}) - \nabla f(y_t),\, y_t - x_{t+1} \rangle + 2L \left\| x_{t+1} - y_t \right\|_p^{2} \ .
\end{aligned}$$

Next, we observe that

$$\begin{aligned}
\langle \nabla f(x_{t+1}) - \nabla f(y_t),\, x_{t+1} - y_t \rangle &\leq \left\| \nabla f(x_{t+1}) - \nabla f(y_t) \right\|_{p^*} \left\| x_{t+1} - y_t \right\|_p \\
&\leq L \left\| x_{t+1} - y_t \right\|_p^2 \ ,
\end{aligned}$$

which implies that $\langle \nabla f(x_{t+1}) - \nabla f(y_t),\, y_t - x_{t+1} \rangle \geq -L \left\| x_{t+1} - y_t \right\|_p^2$.

Combining this with the expression from before, it follows that

$$\langle \nabla f(x_{t+1}), y_t - x_{t+1} \rangle \geq -L \left\| x_{t+1} - y_t \right\|_p^2 + 2L \left\| x_{t+1} - y_t \right\|_p^2 = L \left\| x_{t+1} - y_t \right\|_p^2 \ .$$

Using again the fact that, by first-order optimality conditions, we have, for all $i \in \{1, \ldots, d\}$,

$$\nabla f(y_t)[i] + 2L \left\| x_{t+1} - y_t \right\|_p^{2-p} \left| x_{t+1}[i] - y_t[i] \right|^{p-2} (x_{t+1}[i] - y_t[i]) = 0, \tag{8}$$

it follows that, letting δ be such that $\delta[i] = \left| x_{t+1}[i] - y_t[i] \right|^{p-2} (x_{t+1}[i] - y_t[i])$,

$$\begin{aligned}
\left\| \nabla f(x_{t+1}) \right\|_{p^*}^2 &= \left\| \nabla f(x_{t+1}) - \nabla f(y_t) - 2L \left\| x_{t+1} - y_t \right\|_p^{2-p} \delta \right\|_{p^*}^2 \\
&\leq \left(\left\| \nabla f(x_{t+1}) - \nabla f(y_t) \right\|_{p^*} + \left\| 2L \left\| x_{t+1} - y_t \right\|_p^{2-p} \delta \right\|_{p^*} \right)^2 \\
&\leq \left\| \nabla f(x_{t+1}) - \nabla f(y_t) \right\|_{p^*}^2 + \left\| 2L \left\| x_{t+1} - y_t \right\|_p^{2-p} \delta \right\|_{p^*}^2 \\
&\quad + 2 \left\| \nabla f(x_{t+1}) - \nabla f(y_t) \right\|_{p^*} \left\| 2L \left\| x_{t+1} - y_t \right\|_p^{2-p} \delta \right\|_{p^*} \\
&\leq L^2 \left\| x_{t+1} - y_t \right\|_p^2 + 4L^2 \left\| x_{t+1} - y_t \right\|_p^{2(2-p)} \left\| \delta \right\|_{p^*}^2 + 4L^2 \left\| x_{t+1} - y_t \right\|_p^{3-p} \left\| \delta \right\|_{p^*} \\
&= 9L^2 \left\| x_{t+1} - y_t \right\|_p^2 \quad ,
\end{aligned}$$

where the final equality used the fact that

$$\|\delta\|_{p^*}^2 = \left(\sum_{i=1}^{d}\left(|x_{t+1}[i] - y_t[i]|^{p-1}\right)^{\frac{p}{p-1}}\right)^{\frac{2(p-1)}{p}} = \left(\|x_{t+1} - y_t\|_p^p\right)^{\frac{2(p-1)}{p}} = \|x_{t+1} - y_t\|_p^{2(p-1)}.$$

Combining these, it follows that $\langle \nabla f(x_{t+1}), y_t - x_{t+1}\rangle \geq L\,\|x_{t+1} - y_t\|_p^2 \geq \frac{1}{9L}\,\|\nabla f(x_{t+1})\|_{p^*}^2$. ■

Next, we proceed via the estimate sequence analysis (as in, e.g., (Nesterov, 2018), Section 4.3), adjusting for our per-step descent guarantee in terms of the ℓ_{p^*} norm.

Lemma 2 *Consider Algorithm 1. $\forall\ t \geq 0$, we have for $B_t = \frac{1}{18L}\sum_{i=0}^{t-1} A_{i+1}\,\|\nabla f(x_{i+1})\|_{p^*}^2$,*

$$A_t f(x_t) + B_t \leq \psi_t^* \coloneqq \min_{x\in\mathbb{R}^d} \psi_t(x). \tag{9}$$

Proof We proceed with a proof by induction. First we observe that for the base case $t = 0$, the inequality holds as both sides are 0. Next, suppose the inequality holds for some $t > 0$. Then, for any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}\psi_{t+1}(x) &\geq \psi_t^* + \frac{1}{2}\,\|x - v_t\|_2^2 + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}),\, x - x_{t+1}\rangle] \\ &\geq A_t f(x_t) + B_t + \frac{1}{2}\,\|x - v_t\|_2^2 + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}),\, x - x_{t+1}\rangle] \\ &\geq A_{t+1} f(x_{t+1}) + B_t + \frac{1}{2}\,\|x - v_t\|_2^2 + \langle \nabla f(x_{t+1}),\, A_t(x_t - x_{t+1}) + a_{t+1}(x - x_{t+1})\rangle \\ &= A_{t+1} f(x_{t+1}) + B_t + \frac{1}{2}\,\|x - v_t\|_2^2 + \langle \nabla f(x_{t+1}),\, a_{t+1}(x - v_t) + A_{t+1}(y_t - x_{t+1})\rangle\,.\end{aligned}$$

Next, letting $m(x) \coloneqq \frac{1}{2}\,\|x - v_t\|_2^2 + a_{t+1}\,\langle \nabla f(x_{t+1}),\, x - v_t\rangle$, it follows that, for all $x \in \mathbb{R}^d$, $m(x) \geq -\frac{1}{2}a_{t+1}^2\,\|\nabla f(x_{t+1})\|_2^2$. Therefore, we may observe that

$$\begin{aligned}\psi_{t+1}^* &\geq A_{t+1} f(x_{t+1}) + B_t - \frac{1}{2}a_{t+1}^2\,\|\nabla f(x_{t+1})\|^2 + A_{t+1}\,\langle \nabla f(x_{t+1}),\, y_t - x_{t+1}\rangle \\ &= A_{t+1} f(x_{t+1}) + B_t - \frac{A_{t+1}}{36L\rho_t}\,\|\nabla f(x_{t+1})\|_2^2 + A_{t+1}\,\langle \nabla f(x_{t+1}),\, y_t - x_{t+1}\rangle \\ &\geq A_{t+1} f(x_{t+1}) + B_t - \frac{A_{t+1}}{36L\rho_t}\,\|\nabla f(x_{t+1})\|_2^2 + \frac{A_{t+1}}{9L}\,\|\nabla f(x_{t+1})\|_{p^*}^2 \\ &\geq A_{t+1} f(x_{t+1}) + B_t - \frac{A_{t+1}}{18L}\,\|\nabla f(x_{t+1})\|_{p^*}^2 + \frac{A_{t+1}}{9L}\,\|\nabla f(x_{t+1})\|_{p^*}^2 \\ &= A_{t+1} f(x_{t+1}) + B_t + \frac{A_{t+1}}{18L}\,\|\nabla f(x_{t+1})\|_{p^*}^2 \\ &= A_{t+1} f(x_{t+1}) + B_{t+1},\end{aligned}$$

where the first equality follows from the algorithm that $a_{t+1}^2 = \frac{A_t + A_{t+1}}{18L\rho_t}$, the second inequality follows from Lemma 1, and the last inequality follows from the fact that $\rho_t \geq \frac{1}{2}\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2}$. ■

Lemma 3 *Let $\mathcal{G}_0 = 0$ and $\mathcal{G}_t \coloneqq \frac{1}{t}\sum\limits_{i=0}^{t-1} \frac{\|\nabla f(x_{t+1})\|_{p^*}}{\|\nabla f(x_{t+1})\|_2}$ for $t > 0$. Then, for all $t \geq 0$, we have*

$$A_t^{1/2} \geq \frac{1}{18L^{1/2}} \sum_{i=0}^{t-1} \frac{\|\nabla f(x_{i+1})\|_{p^*}}{\|\nabla f(x_{i+1})\|_2} = \frac{t}{18L^{1/2}} \left(\frac{1}{t} \sum_{i=0}^{t-1} \frac{\|\nabla f(x_{i+1})\|_{p^*}}{\|\nabla f(x_{i+1})\|_2} \right) = \frac{\mathcal{G}_t t}{18L^{1/2}}.$$

Proof We proceed with a proof by induction. For $t = 0$, $A_0 = 0$, and so the inequality holds. Suppose for $t > 0$, $A_t^{1/2} \geq \frac{1}{18L^{1/2}} \sum\limits_{i=0}^{t-1} \frac{\|\nabla f(x_{i+1})\|_{p^*}}{\|\nabla f(x_{i+1})\|_2}$. Observe that

$$A_{t+1}^{1/2} - A_t^{1/2} = \frac{a_{t+1}}{A_{t+1}^{1/2} + A_t^{1/2}} = \frac{1}{A_{t+1}^{1/2} + A_t^{1/2}} \left(\frac{A_{t+1}}{18L\rho_t} \right)^{1/2} \geq \frac{1}{9L^{1/2}\rho_t^{1/2}}. \tag{10}$$

Thus, we have that

$$\begin{aligned} A_{t+1}^{1/2} &\geq A_t^{1/2} + \frac{1}{9L^{1/2}\rho_t^{1/2}} \geq \frac{1}{18L^{1/2}} \sum_{i=0}^{t-1} \frac{\|\nabla f(x_{t+1})\|_{p^*}}{\|\nabla f(x_{t+1})\|_2} + \frac{1}{9L^{1/2}\rho_t^{1/2}} \\ &\geq \frac{1}{18L^{1/2}} \sum_{i=0}^{t-1} \frac{\|\nabla f(x_{i+1})\|_{p^*}}{\|\nabla f(x_{i+1})\|_2} + \frac{1}{18L^{1/2}} \frac{\|\nabla f(x_{t+1})\|_{p^*}}{\|\nabla f(x_{t+1})\|_2} = \frac{1}{18L^{1/2}} \sum_{i=0}^{t} \frac{\|\nabla f(x_{i+1})\|_{p^*}}{\|\nabla f(x_{i+1})\|_2}, \end{aligned}$$

where the second inequality follows from our inductive hypothesis, and the final inequality follows from the fact that $\rho_t \leq 2\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2}$, which yields the desired result. ∎

We now have the requisite tools to prove the main theorem of our work.

Theorem 4 (Main theorem) *Let f be convex and L-smooth w.r.t. $\|\cdot\|_p$. Then, after $T > 0$ iterations, and letting $\mathcal{G} \coloneqq \frac{1}{T}\sum\limits_{t=0}^{T-1} \frac{\|\nabla f(x_{t+1})\|_{p^*}}{\|\nabla f(x_{t+1})\|_2}$, it holds that* **HASD** *(Algorithm 1) outputs x_T such that*

$$f(x_T) - f(x^*) \leq \frac{324L\|x_0 - x^*\|_2^2}{\mathcal{G}^2 T^2}. \tag{11}$$

Proof By convexity, $\forall\, x \in \mathbb{R}^d$, $\forall\, t \in [T]$,

$$\psi_t(x) \leq \psi_{t-1}(x) + a_t f(x) \leq \psi_0(x) + \sum_{i=1}^{t} a_i f(x) = \frac{1}{2}\|x - x_0\|_2^2 + A_t f(x).$$

Further by Lemma 2,

$$A_T f(x_T) \leq \psi_T^* \leq \psi_T(x^*) \leq \frac{1}{2}\|x^* - x_0\|_2^2 + A_T f(x^*),$$

which yields $f(x_T) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2A_T}$. Applying Lemma 3 completes the proof. ∎

We would note that the key difference between the rate of Theorem 4 and that as presented in, e.g., (Allen-Zhu and Orecchia, 2017), is precisely the addition of the $\mathcal{G}^2$ term, *which may be as large as*

$d^{1-\frac{2}{p}}$ (and which is furthermore always ≥ 1). We would also note that there exists an instance of minimizing the softmax function f^s, as shown in Appendix B, for which $\mathcal{G}^2 = d^{1-\frac{2}{p}}$, thus yielding the rate, when $p=\infty$, $f^s(x_T) - f^s(x^*) \leq \frac{324 L_s \|x_0 - x^*\|_2^2}{dT^2} \leq \frac{324 L_s \|x_0 - x^*\|_\infty^2}{T^2}$ where $L_s \in \mathcal{O}\left(\varepsilon^{-1}\right)$ when approximating ℓ_∞ regression, thereby matching, for this particular instance,[1]the rate given by (Sherman, 2017; Sidford and Tian, 2018; Cohen et al., 2021). We cannot, however, achieve rates of this sort for general ℓ_p-smooth functions, due to the lower bound (Guzmán and Nemirovski, 2015).

3.3. Complexity of Binary Search

In this section, we characterize the complexity of binary search to simultaneously find a pair of ρ_t and x_{t+1} that satisfies the implicit relation of $\frac{1}{2}\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2} \leq \rho_t \leq 2\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2}$ at each iteration. We show in the following theorem that such binary search takes at most $\mathcal{O}\left(\log(d) + \log\left(\frac{1}{\varepsilon}\right)\right)$ calls to the first-order steepest descent oracle. The proof follows the general framework of Theorem 18 in (Bubeck et al., 2019), with several non-trivial technical modifications to accommodate the ℓ_p norm and the specific formulation of the condition, which involves the ratio of the ℓ_2 and ℓ_{p^*} norms of the gradient. We provide a proof sketch and defer the complete proof to Appendix D.

Theorem 5 *For any iteration t in Algorithm 1, with at most $9 + \frac{5(p-2)}{2p}\log_2(d) + \log_2\left(\frac{LD_R}{\varepsilon}\right)$ calls to the first-order ℓ_p steepest descent oracle, we can find either a point x_{t+1} such that $f(x_{t+1}) - f(x^*) \leq \varepsilon$ or a pair of ρ_t, x_{t+1} that satisfies the condition $\frac{1}{2}\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2} \leq \rho_t \leq 2\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2}$ for $0 < \varepsilon \leq \frac{LD_R}{6}$ where $D_R = \left(R + 1458R^2\right)\left(20R + 4374R^2\right)$ for $R = \|x_0 - x^*\|_2$.*

Proof sketch Given the implicit dependence of x_{t+1} on ρ_t and vice versa, we make such dependence explicit by letting $\theta := \frac{A_t}{A_{t+1}}$, so that other variables can all be expressed as a function of θ, denoted as $x_\theta := x_{t+1}$, $y_\theta := y_t = \theta x_t + (1-\theta)v_t$, and $\rho_\theta := \rho_t = \frac{\theta}{18L(1-\theta)^2 A_t}$. As a result, the search for ρ_t, x_{t+1} that satisfy $\frac{1}{2}\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2} \leq \rho_t \leq 2\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2}$ is equivalent to finding θ such that

$$\frac{1}{2} \leq \zeta(\theta) := \frac{18L(1-\theta)^2 A_t}{\theta}\frac{\|\nabla f(x_\theta)\|_2^2}{\|\nabla f(x_\theta)\|_{p^*}^2} \leq 2.$$

Noting that $\zeta(0) = \infty$, $\zeta(1) = 0$, $\exists\, \theta^*$ such that $\zeta(\theta^*) = \frac{5}{4}$. Then one can use $\log_2\left(\frac{1}{\delta}\right)$ binary search steps to find θ such that $|\theta - \theta^*| \leq \delta$. It remains to verify that with certain choice of δ, we indeed have $\zeta(\theta) \in \left[\frac{1}{2}, 2\right]$. How the function value changes with respect to the input within some δ-neighborhood is characterized by the Lipschitz constant of $\zeta(\theta)$, which we analyze by showing

$$\left|\frac{d}{d\theta}\log\left(\zeta(\theta)\right)\right| \leq \frac{2}{1-\theta} + \frac{1}{\theta} + \frac{4d^{\frac{p-2}{2p}}}{\|\nabla f(x_\theta)\|_{p^*}}\left\|\nabla^2 f(x_\theta)\right\|_p \left\|\frac{d}{d\theta}x_\theta\right\|_p.$$

We bound it by bounding each of the relevant terms, $\|\nabla f(x_\theta)\|_{p^*}$ by Lemma 17, $\left\|\nabla^2 f(x_\theta)\right\|_p$ by Lemma 13, and $\left\|\frac{d}{d\theta}x_\theta\right\|_p$ by Lemma 15, which involves nontrivially showing for $x \in \mathbb{R}^d$,

[1] While we include this softmax example to provide an instance where $O(d^{1-\frac{2}{p}})$ acceleration is achieved, it remains to be investigated how to further incorporate the affine transformation for more general ℓ_∞ regression.

$\nabla^2 \left\|x\right\|_p^2$ is the sum of two positive semidefinite matrices, after which it can be further simplified to $\left|\frac{d}{d\theta}\log\left(\zeta\left(\theta\right)\right)\right| \leq \omega\left(\theta\right)\left(1+\frac{1}{\zeta(\theta)}+\zeta\left(\theta\right)\right)$ for $\omega\left(\theta\right) = 4d^{\frac{5(p-2)}{2p}}\left(6+9LA_t+\frac{LD_R}{f(x_\theta)-f(x^*)}\right)$ where $D_R = \left(R+1458R^2\right)\left(20R+4374R^2\right)$. Finally, we show in Lemma 18 that if the Lipschitz constant (as a function of θ) is bounded as such, by properly choosing the neighborhood $\delta = \frac{\varepsilon}{320d^{\frac{5(p-2)}{2p}}LD_R} \leq \frac{1}{10\omega(\theta)}$, $\zeta\left(\theta\right)$ falls within the range $\left[\frac{1}{2},2\right]$ when $\left|\theta-\theta^*\right| \leq \delta$. ■

4. Extensions

In this section, we consider natural extensions of our method to additional related problem settings, including minimizing strongly convex objectives, as well as minimizing the ℓ_{p^*} norm of the gradient. The latter—explored, for example, in (Gratton and Toint, 2023; Diakonikolas and Guzmán, 2024)—may be of independent interest, as it allows us to deviate from the typical goal of minimizing in terms of the ℓ_2 norm.

4.1. Strongly convex setting

We begin by considering the case in which f is additionally μ-strongly convex, whereby combining our method with the usual restarting scheme lets us straightforwardly improve from a sublinear to a linear rate. Furthermore, it is important to note that the improvements our method offers in the smooth and (weakly) convex setting appear, in the strongly convex setting, outside of the log factor, along with the condition number of the problem. Due to space constraints, we provide the full details of our algorithm for the strongly convex setting (**HASD + Restarting**) in Appendix A.

Using our results from Section 3, we now arrive at the following corollary.

Corollary 6 *Let $\varepsilon > 0$, let $x_{outer,0} \in \mathbb{R}^d$, and let $K = O(\log(1/\varepsilon))$. Consider f that is L-smooth w.r.t. $\left\|\cdot\right\|_p$ and μ-strongly convex w.r.t. $\left\|\cdot\right\|_2$. Assume that, for all $i \in \{1,\ldots,K\}$, the respective average term $\mathcal{G}_i \geq \hat{\mathcal{G}} \geq 1$. Then, the method* **HASD + Restarting** *(Algorithm 2) outputs $x_{outer,K}$ such that*

$$f(x_{outer,K}) - f(x^*) \leq \varepsilon. \tag{12}$$

Proof Note that by Theorem 4, for all i, it holds that

$$f(x_{\text{outer},i+1}) - f(x^*) \leq \frac{324L\left\|x_{\text{outer},i} - x^*\right\|_2^2}{\mathcal{G}_i^2 T^2}$$

where $\mathcal{G}_i := \frac{1}{T}\sum\limits_{t=0}^{T-1}\frac{\left\|\nabla f(x_{t+1})\right\|_{p^*}}{\left\|\nabla f(x_{t+1})\right\|_2}$ (where x_t are w.r.t. the i^{th} outer iteration). By μ-strong convexity, we have that $\left\|x_{\text{outer},i} - x^*\right\|_2^2 \leq \frac{2}{\mu}(f(x_{\text{outer},i}) - f(x^*))$, and so it follows that

$$f(x_{\text{outer},i+1}) - f(x^*) \leq \frac{648L(f(x_{\text{outer},i}) - f(x^*))}{\mu\mathcal{G}_i^2 T^2}.$$

Thus, by setting $T = \frac{36}{\hat{\mathcal{G}}}\sqrt{\frac{L}{\mu}}$, we have that $f(x_{\text{outer},i+1}) - f(x^*) \leq \frac{f(x_{\text{outer},i})-f(x^*)}{2}$, and so, since we halve the optimality gap each time, the desired result follows from the recurrence after $K = O(\log(1/\varepsilon))$ (outer) iterations. ■

Interestingly, due to the diameter term being $\|x_0 - x^*\|_2^2$ (i.e., in terms of the ℓ_2 norm) in the final convergence expression for the smooth and (weakly) convex case, we similarly need the strong convexity assumption to be w.r.t. ℓ_2 for the analysis to hold.

4.2. Gradient norm minimization

As a natural consequence of the results in both the (weakly) convex and strongly convex setting, we may additionally derive guarantees in terms of minimizing the ℓ_{p^*} norm of the gradient, as has been considered (in the case of the ℓ_2 norm) in both convex (e.g., (Nesterov, 2012; Allen-Zhu, 2018)) and non-convex (e.g., (Agarwal et al., 2017; Carmon et al., 2018)) settings.

4.2.1. First attempt: directly relating to optimality gap

Aiming to minimize the gradient norm, we begin with the following lemma, the proof of which follows the standard transition from optimality gap to gradient norm by convexity and smoothness and may be found in Appendix C.

Lemma 7 *Let f be convex and L-smooth w.r.t. $\|\cdot\|_p$, and let $x \in \mathbb{R}^d$ be such that $f(x) - f(x^*) \leq \frac{\varepsilon^2}{2L}$. Then, $\|\nabla f(x)\|_{p^*} \leq \varepsilon$.*

Combining this with our main convergence guarantee (Theorem 4) leads to the following corollary.

Corollary 8 *Let $R > 0$ be such that $\|x_0 - x^*\|_2 \leq R$, and let x_T be the output of* **HASD** *(Algorithm 1) after $T = \left\lceil \frac{18\sqrt{2}LR}{\hat{\mathcal{G}}\varepsilon} \right\rceil$ iterations, where $\hat{\mathcal{G}}$ is such that $\mathcal{G} \geq \hat{\mathcal{G}} \geq 1$. Then,*

$$\|\nabla f(x_T)\|_{p^*} \leq \varepsilon.$$

4.2.2. Improved rate: using additional B_t term

By observing more closely the recurrence relation established in Lemma 2, we may further improve the rate for minimizing $\|\nabla f(x)\|_{p^*}$.

Corollary 9 *Let $R > 0$ be such that $\|x_0 - x^*\|_2 \leq R$, and let x_t ($t \in \{1, \dots T\}$) be the iterates generated by* **HASD** *(Algorithm 1) after $T = \left\lceil \frac{21 L^{\frac{2}{3}} R^{\frac{2}{3}}}{\hat{\mathcal{G}}^{\frac{2}{3}} \varepsilon^{\frac{2}{3}}} \right\rceil$ iterations, where $\hat{\mathcal{G}}$ is such that, for all $t \in \{1, \dots T\}$, $\mathcal{G}_t \geq \hat{\mathcal{G}} \geq 1$. Then,*

$$\min_{t \in \{1, \dots T\}} \|\nabla f(x_t)\|_{p^*} \leq \varepsilon.$$

Proof The corollary follows from the fact that, by Lemma 2, we know

$$\left(\frac{1}{18L} \sum_{t=0}^{T-1} A_{t+1}\right) \left(\min_{t \in \{1, \dots T\}} \|\nabla f(x_t)\|_{p^*}^2\right) \leq B_T = \frac{1}{18L} \sum_{t=0}^{T-1} A_{t+1} \|\nabla f(x_{t+1})\|_{p^*}^2 \leq \frac{1}{2} \|x_0 - x^*\|_2^2.$$

Therefore, given $A_t^{1/2} \geq \frac{\mathcal{G}_t t}{18 L^{1/2}}$ from Lemma 3,

$$\min_{t \in \{1, \dots T\}} \|\nabla f(x_t)\|_{p^*}^2 \leq \frac{9LR^2}{\sum_{t=0}^{T-1} A_{t+1}} \leq \frac{2916 L^2 R^2}{\sum_{t=1}^{T} \mathcal{G}_t^2 t} \leq \frac{2916 L^2 R^2}{\hat{\mathcal{G}}^2 \frac{T(T+1)(2T+1)}{6}} \leq \frac{8748 L^2 R^2}{\hat{\mathcal{G}}^2 T^3}$$

From $\frac{8748L^2R^2}{\hat{\mathcal{G}}^2T^3} \leq \epsilon^2$ we solve for $T \geq \frac{21L^{\frac{2}{3}}R^{\frac{2}{3}}}{\hat{\mathcal{G}}^{\frac{2}{3}}\varepsilon^{\frac{2}{3}}}$. ■

5. Discussion

We believe the relative simplicity of our algorithm affords it the opportunity to be expanded and simplified (from both theoretical *and* practical perspectives), as it is a readily implementable and efficient first-order algorithm. Recent developments in sign (stochastic) gradient methods (which are a special case of steepest descent w.r.t. $\|\cdot\|_\infty$) (Bernstein et al., 2018; Balles et al., 2020; Chen et al., 2023) highlight the critical importance of better understanding the interplay between acceleration and steepest descent, and we feel that our work brings additional insights and perspectives to this context.

Comparison with (Monteiro and Svaiter, 2013). In contrast to work by Nesterov (2008), which achieves a rate of $O(1/T^3)$ for accelerated cubic regularization, Monteiro and Svaiter (2013) establish an improved $\tilde{O}(1/T^{7/2})$ rate for the same (convex, second-order smooth) setting, by using the Accelerated Hybrid Proximal Extragradient (A-HPE) method. A key algorithmic difference between the two approaches lies in certain choice of regularization function. Namely, whereas Nesterov (2008) uses $\frac{1}{3}\|x - x_0\|_2^3$—here, we note that the exponent is 3, matching that of the cubic regularization term and thus *maintaining a certain symmetry*—Monteiro and Svaiter (2013) instead use a *quadratic* term, thereby *breaking the symmetry*. Much of the subsequent analysis (which gives the near-optimal rate) is based around *adjusting for this broken symmetry*.

We wish to emphasize this (high-level) observation, as our approach follows a similar means of "symmetry-breaking," though instead of doing so w.r.t. the exponent of the regularizer, we do so *w.r.t. the norm itself*. Specifically, we combine ℓ_p norm-based ($p \geq 2$) steepest descent steps with ℓ_2 norm-based mirror descent steps, *and algorithmically adjust for this discrepancy*. It follows that, in appropriately accounting for this adjustment, our algorithm HASD yields additional convergence gains (as made explicit in Theorem 4) *beyond* what has been previously shown.

Practical considerations. We provide in Appendix E preliminary experimental evidence to validate (and complement) the main guarantees of our algorithm. We would also note that, while a line search procedure is needed in theory, in practice this may likely be relaxed, by using, e.g., a heuristic procedure. Furthermore, it may be the case that the line search could be (effectively) removed altogether, as has been shown in the case of high-order optimization (Carmon et al., 2022; Kovalev and Gasnikov, 2022), and doing so may provide an interesting future research direction.

6. Conclusion and future work

We have presented a new method for accelerating non-Euclidean steepest descent, based on an implicit interpolation of steepest and mirror descent updates in differing norms, which offers up to $O(d^{1-\frac{2}{p}})$ improvement, in terms of iteration complexity, when considering smoothness w.r.t. $\|\cdot\|_p$. We believe our results suggest there are many more interesting directions yet to be explored, even in the case of *first-order* acceleration. Due to the role optimization landscapes and geometry play in training deep learning models, we are also optimistic that our approach might lend itself to more practical considerations—with promising evidence of such possibilites recently presented by Luo et al. (2025) for stochastic settings—though we leave a full exploration of this to future work.

Another possibility would be to consider whether we might achieve a more fine-grained analysis, as has been shown by Sidford and Tian (2018) for the problem of ℓ_∞ regression. Determining (matching) lower bounds would also help to clarify the placement of our results. In addition, given the fact that standard (Euclidean) acceleration has been extended to non-convex (Agarwal et al., 2017; Carmon et al., 2018) and stochastic (Ghadimi and Lan, 2013) settings, we believe it may be possible to extend our results in a similar manner. As our analysis provides a principled framework for characterizing the oracle complexity of minimizing smooth convex functions under *nonstandard geometries* (Guzmán, 2015), we suspect our techniques could also apply to more general domains whose diameters are measured in alternative q-norms.

References

Naman Agarwal and Elad Hazan. Lower bounds for higher-order convex optimization. In *Conference On Learning Theory*, pages 774–792. PMLR, 2018. (Cited on page 4.)

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199, 2017. (Cited on pages 11 and 13.)

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on page 11.)

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017. (Cited on pages 2, 3, and 8.)

Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1):327–360, 2019. (Cited on page 4.)

Michel Baes. Estimate sequence methods: extensions and approximations. *Institute for Operations Research, ETH, Zürich, Switzerland*, 2(1), 2009. (Cited on page 4.)

Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *arXiv preprint arXiv:2002.08056*, 2020. (Cited on pages 3, 12, and 24.)

Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012. (Cited on page 2.)

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018. (Cited on pages 3 and 12.)

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. (Cited on page 3.)

Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015. (Cited on page 2.)

Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near-optimal method for highly smooth convex optimization. In *Conference on Learning Theory*, pages 492–507. PMLR, 2019. (Cited on pages 5 and 9.)

Brian Bullins. Highly smooth minimization of non-smooth problems. In *Conference on Learning Theory*, pages 988–1030. PMLR, 2020. (Cited on page 29.)

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018. (Cited on pages 11 and 13.)

Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. *Advances in Neural Information Processing Systems*, 33:19052–19063, 2020. (Cited on page 3.)

Yair Carmon, Danielle Hausler, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Optimal and adaptive Monteiro-Svaiter acceleration. *Advances in Neural Information Processing Systems*, 35: 20338–20350, 2022. (Cited on page 12.)

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 36, 2023. (Cited on pages 3 and 12.)

Michael B Cohen, Aaron Sidford, and Kevin Tian. Relative lipschitzness in extragradient methods and a direct recipe for acceleration. In *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021. (Cited on pages 1, 2, and 9.)

Jelena Diakonikolas and Cristóbal Guzmán. Complementary composite minimization, small gradients in general norms, and applications. *Mathematical Programming*, pages 1–45, 2024. (Cited on pages 3 and 10.)

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011. (Cited on page 1.)

Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, César A Uribe, Bo Jiang, Haoyue Wang, Shuzhong Zhang, Sébastien Bubeck, et al. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives. In *Conference on Learning Theory*, pages 1392–1393. PMLR, 2019. (Cited on page 4.)

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. (Cited on page 13.)

S Gratton and Ph L Toint. Adaptive regularization minimization algorithms with nonsmooth norms. *IMA Journal of Numerical Analysis*, 43(2):920–949, 2023. (Cited on page 10.)

Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015. (Cited on pages 2, 3, and 9.)

Cristóbal Guzmán. Open problem: The oracle complexity of smooth convex optimization in nonstandard settings. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1761–1763, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Guzman15.html. (Cited on pages 3 and 13.)

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. (Cited on page 2.)

Jonathan A Kelner, Yin Tat Lee, Lorenzo Orecchia, and Aaron Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 217–226. SIAM, 2014. (Cited on pages 2, 3, and 29.)

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 1.)

Dmitry Kovalev and Alexander Gasnikov. The first optimal acceleration of high-order methods in smooth convex optimization. *Advances in Neural Information Processing Systems*, 35:35339–35351, 2022. (Cited on page 12.)

Xinyu Luo, Cedar Site Bai, Bolian Li, Petros Drineas, Ruqi Zhang, and Brian Bullins. Stacey: Promoting stochastic steepest descent via accelerated ℓ_p-smooth nonconvex optimization. In *International Conference on Machine Learning*. PMLR, 2025. (Cited on page 12.)

Renato DC Monteiro and Benar Fux Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010. (Cited on page 3.)

Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013. (Cited on pages 3, 4, 5, and 12.)

Arkadi Semenovič Nemirovski and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983. (Cited on page 2.)

Nemirovskii and Y. Nesterov. Optimal methods of smooth convex optimization (in russian). *Zh. Vychisl. Mat. i Mat. Fiz.*, 25(3):356—-369, 1985. (Cited on page 3.)

Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k** 2). *Doklady Akademii Nauk SSSR*, 269(3):543, 1983. (Cited on pages 1 and 3.)

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103: 127–152, 2005. (Cited on pages 2, 3, 5, and 17.)

Yurii Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008. (Cited on pages 4, 5, and 12.)

Yurii Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012. (Cited on page 11.)

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018. (Cited on pages 1, 3, and 7.)

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical programming*, 108(1):177–205, 2006. (Cited on page 4.)

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964. (Cited on page 1.)

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. (Cited on page 1.)

Jonah Sherman. Area-convexity, ℓ_∞ regularization, and undirected multicommodity flow. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 452–460, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345286. doi: 10.1145/3055399.3055501. URL https://doi.org/10.1145/3055399.3055501. (Cited on pages 1, 2, and 9.)

Aaron Sidford and Kevin Tian. Coordinate methods for accelerating l-infty regression and faster approximate maximum flow. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 922–933. IEEE, 2018. (Cited on pages 1, 2, 9, and 13.)

Appendix A. Algorithm for Strongly Convex Setting

We include here the algorithm for the strongly convex setting.

Algorithm 2 HASD + Restarting

```
Input: $x_{\text{outer},0} \in \mathbb{R}^d$, $\varepsilon > 0$, $\hat{\mathcal{G}} > 0$, $K = O(\log(1/\varepsilon))$
  for $i = 0$ to $K - 1$ do
    $A_0 = 0$
    Define $\psi_0(x) := \frac{1}{2}\|x - x_{i,0}\|_2^2$.
    $x_0 = x_{outer,i}$
    $T = \frac{36}{\hat{\mathcal{G}}}\sqrt{\frac{L}{\mu}}$
    for $t = 0$ to $T - 1$ do
      $v_t = \arg\min_{x \in \mathbb{R}^d} \psi_t(x)$
      Determine $\rho_t > 0$, $x_{t+1} \in \mathbb{R}^d$ for which the following hold simultaneously:
        •  $\frac{1}{2}\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2} \leq \rho_t \leq 2\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2}$
        •  $a_{t+1} > 0$ s.t. $a_{t+1}^2 = \frac{(A_t + a_{t+1})}{18L\rho_t}$
        •  Set $A_{t+1} = A_t + a_{t+1}$, $\tau_t = \frac{a_{t+1}}{A_{t+1}}$
        •  $y_t = (1 - \tau_t)x_t + \tau_t v_t$
        •  $x_{t+1} = \arg\min_{x \in \mathbb{R}^d}\{\langle \nabla f(y_t), x - y_t \rangle + L\|x - y_t\|_p^2\}$
      $\psi_{t+1}(x) = \psi_t(x) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle]$
    end for
    $x_{outer,i+1} = x_T$
  end for
Output: $x_{outer,K}$
```

Appendix B. Example: The Softmax Function

Consider, as an illustration, the (symmetric) softmax objective:

$$\min_{x \in \mathbb{R}^d} f^s(x) := \alpha \log\left(\sum_{i=1}^{d} e^{\frac{x[i]}{\alpha}} + e^{\frac{-x[i]}{\alpha}}\right)$$

For appropriate choice of α, the softmax function closely approximates ℓ_∞ regression (Nesterov, 2005). Further, suppose we initialize at $x_0 = [1, \ldots, 1]^\top \in \mathbb{R}^d$. Then we may observe that, by symmetry, the iterates of the algorithm $x_0, \ldots, x_T$ satisfy $\forall\ i, j$, $\frac{\partial}{\partial x[i]} f^s(x_t) = \frac{\partial}{\partial x[j]} f^s(x_t)$ for $t = 0, \ldots, T$. That is to say, $\forall\ t \in [T]$, $\frac{\|\nabla f^s(x_t)\|_{p^*}}{\|\nabla f^s(x_t)\|_2} = d^{\frac{1}{2} - \frac{1}{p}}$, and so $\mathcal{G} = d^{\frac{1}{2} - \frac{1}{p}}$.

Appendix C. Naive Analysis for Gradient Norm Minimization

Lemma 7 *Let f be convex and L-smooth w.r.t. $\|\cdot\|_p$, and let $x \in \mathbb{R}^d$ be such that $f(x) - f(x^*) \leq \frac{\varepsilon^2}{2L}$. Then, $\|\nabla f(x)\|_{p^*} \leq \varepsilon$.*

Proof Let $z = x - \frac{1}{L} \left\|\nabla f(x)\right\|_{p^*}^{\frac{p-2}{p-1}} g$, where g is such that $g[i] := \frac{\nabla f(x)[i]}{|\nabla f(x)[i]|^{\frac{p-2}{p-1}}}$ for all $i \in \{1, \ldots, d\}$. Using our smoothness assumption, along with the fact that $\nabla f(x^*) = 0$, it follows that

$$\begin{aligned} f(x^*) - f(x) &= f(x^*) - f(z) + f(z) - f(x) \\ &\leq \langle \nabla f(x^*),\, x^* - z\rangle + \langle \nabla f(x),\, z - x\rangle + \frac{L}{2} \left\|z - x\right\|_p^2 \\ &= -\frac{1}{L} \left\|\nabla f(x)\right\|_{p^*}^{\frac{p-2}{p-1}} \langle \nabla f(x),\, g\rangle + \frac{1}{2L} \left\|\nabla f(x)\right\|_{p^*}^{\frac{2(p-2)}{p-1}} \left\|g\right\|_p^2 \\ &= -\frac{1}{2L} \left\|\nabla f(x)\right\|_{p^*}^2, \end{aligned}$$

where the inequality follows from convexity and L-smoothness of f. Rearranging and multiplying both sides by $2L$ gives us $\left\|\nabla f(x)\right\|_{p^*}^2 \leq 2L(f(x) - f(x^*))$, and so, using

$$f(x) - f(x^*) \leq \frac{\varepsilon^2}{2L},$$

it follows that

$$\left\|\nabla f(x)\right\|_{p^*} \leq \varepsilon,$$

as desired. ■

Appendix D. Complexity of Binary Search

D.1. Proof of Theorem 5

Theorem 5 *For any iteration t in Algorithm 1, with at most* $9 + \frac{5(p-2)}{2p} \log_2(d) + \log_2\left(\frac{LD_R}{\varepsilon}\right)$ *calls to the first-order* ℓ_p *steepest descent oracle, we can find either a point* x_{t+1} *such that* $f(x_{t+1}) - f(x^*) \leq \varepsilon$ *or a pair of* ρ_t, x_{t+1} *that satisfies the condition* $\frac{1}{2}\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2} \leq \rho_t \leq 2\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2}$ *for* $0 < \varepsilon \leq \frac{LD_R}{6}$ *where* $D_R = \left(R + 1458R^2\right)\left(20R + 4374R^2\right)$ *for* $R = \left\|x_0 - x^*\right\|_2$.

Proof For every iteration of the algorithm, we seek for $\rho_t > 0$, $x_{t+1} \in \mathbb{R}^d$ for which the following hold simultaneously:

- $\frac{1}{2}\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2} \leq \rho_t \leq 2\frac{\|\nabla f(x_{t+1})\|_2^2}{\|\nabla f(x_{t+1})\|_{p^*}^2}$
- $a_{t+1} > 0$ s.t. $a_{t+1}^2 = \frac{(A_t + a_{t+1})}{18L\rho_t}$
- Set $A_{t+1} = A_t + a_{t+1}$, $\tau_t = \frac{a_{t+1}}{A_{t+1}}$
- $y_t = (1 - \tau_t)x_t + \tau_t v_t$
- $x_{t+1} = \arg\min_{x \in \mathbb{R}^d} \{\langle \nabla f(y_t),\, x - y_t\rangle + L\left\|x - y_t\right\|_p^2\}$

One can see the circular dependence of x_{t+1} on y_t, which depends on a_{t+1}, which depends on ρ_t, which then depends on x_{t+1}. We break such circular dependence by letting $\theta := \frac{A_t}{A_{t+1}}$, and express other variables as a function of θ, denoted as $x_\theta := x_{t+1}$, $y_\theta := y_t$, and $\rho_\theta := \rho_t$. We further denote $z_\theta := x_\theta - y_\theta$. By definition, $\frac{a_{t+1}}{A_{t+1}} = 1 - \theta$ and we have $y_\theta = \theta x_t + (1 - \theta)v_t$. As a result,

$\rho_\theta = \frac{A_t + a_{t+1}}{18La_{t+1}^2} = \frac{\theta}{18L(1-\theta)^2 A_t}$. All conditions are now satisfied except for the first condition, which is equivalent to

$$\frac{1}{2} \leq \frac{\|\nabla f(x_\theta)\|_2^2}{\rho_\theta \|\nabla f(x_\theta)\|_{p^*}^2} \leq 2$$

Defining

$$\zeta(\theta) := \frac{\|\nabla f(x_\theta)\|_2^2}{\rho_\theta \|\nabla f(x_\theta)\|_{p^*}^2} = \frac{18L(1-\theta)^2 A_t}{\theta} \frac{\|\nabla f(x_\theta)\|_2^2}{\|\nabla f(x_\theta)\|_{p^*}^2},$$

we can search for θ such that $\frac{1}{2} \leq \zeta(\theta) \leq 2$, which then yields all conditions satisfied simultaneously.

Given that $\zeta(0) = \infty$, $\zeta(1) = 0$, $\exists\ \theta^* \in [0,1]$ such that $\zeta(\theta^*) = \frac{5}{4}$. Then one can use $\log_2\left(\frac{1}{\delta}\right)$ bineary search step to find θ such that $|\theta - \theta^*| \leq \delta$. Now we verify that with certain choice of δ, $\zeta(\theta) \in \left[\frac{1}{2}, 2\right]$. How the function value changes with respect to the input within some δ-neighborhood is characterized by the Lipschitz constant of $\zeta(\theta)$, i.e., the upper bound on $\left|\frac{d}{d\theta}\zeta(\theta)\right|$. For simplicity, we start by analyzing $\left|\frac{d}{d\theta}\log(\zeta(\theta))\right|$. It's trivial that

$$\log(\zeta(\theta)) = 2\log(1-\theta) - \log(\theta) + \log(18LA_t) + 2\log(\|\nabla f(x_\theta)\|_2) - 2\log\left(\|\nabla f(x_\theta)\|_{p^*}\right).$$

Taking the derivative, we have

$$\begin{aligned}\frac{d}{d\theta}\log(\zeta(\theta)) = &-\frac{2}{1-\theta} - \frac{1}{\theta} + \frac{2}{\|\nabla f(x_\theta)\|_2^2}\nabla^2 f(x_\theta)\nabla f(x_\theta)\frac{d}{d\theta}x_\theta \\ &- \frac{2}{\|\nabla f(x_\theta)\|_{p^*}^{p^*}}\nabla^2 f(x_\theta)\begin{bmatrix}|\nabla f(x_\theta)[1]|^{p^*-2}\nabla f(x_\theta)[1] \\ \vdots \\ |\nabla f(x_\theta)[d]|^{p^*-2}\nabla f(x_\theta)[d]\end{bmatrix}\frac{d}{d\theta}x_\theta.\end{aligned}$$

Taking the ℓ_p norm on both sides, we have

$$\begin{aligned}\left|\frac{d}{d\theta}\log(\zeta(\theta))\right| &\leq \frac{2}{1-\theta} + \frac{1}{\theta} + \frac{2}{\|\nabla f(x_\theta)\|_2^2}\left\|\nabla^2 f(x_\theta)\right\|_p \|\nabla f(x_\theta)\|_p \left\|\frac{d}{d\theta}x_\theta\right\|_p \\ &\quad + \frac{2}{\|\nabla f(x_\theta)\|_{p^*}}\left\|\nabla^2 f(x_\theta)\right\|_p \left\|\frac{d}{d\theta}x_\theta\right\|_p \\ &\leq \frac{2}{1-\theta} + \frac{1}{\theta} + \frac{4}{\|\nabla f(x_\theta)\|_2}\left\|\nabla^2 f(x_\theta)\right\|_p \left\|\frac{d}{d\theta}x_\theta\right\|_p \\ &\leq \frac{2}{1-\theta} + \frac{1}{\theta} + \frac{4d^{\frac{p-2}{2p}}}{\|\nabla f(x_\theta)\|_{p^*}}\left\|\nabla^2 f(x_\theta)\right\|_p \left\|\frac{d}{d\theta}x_\theta\right\|_p.\end{aligned}$$

For the first two terms, we have

$$\frac{1}{1-\theta} \leq 1 + \frac{\theta}{(1-\theta)^2} = 1 + \frac{18LA_t}{\zeta(\theta)}\frac{\|\nabla f(x_\theta)\|_2^2}{\|\nabla f(x_\theta)\|_{p^*}^2} \leq 1 + \frac{18LA_t}{\zeta(\theta)},$$

and

$$\frac{1}{\theta} \leq 2 + \frac{(1-\theta)^2}{\theta} = 2 + \frac{\zeta(\theta)}{18LA_t} \frac{\|\nabla f(x_\theta)\|_{p^*}^2}{\|\nabla f(x_\theta)\|_2^2} \leq 2 + \frac{d^{\frac{p-2}{p}}\zeta(\theta)}{18LA_t}.$$

For each of the relevant components in the third term, we lower bound $\|\nabla f(x_\theta)\|_{p^*}$ in Lemma 17, upper bound $\left\|\nabla^2 f(x_\theta)\right\|_p$ and $\left\|\frac{d}{d\theta}x_\theta\right\|_p$ by Lemma 13, and Lemma 15. With these results, we have

$$\begin{aligned}
\left|\frac{d}{d\theta}\log(\zeta(\theta))\right| &\leq \frac{2}{1-\theta} + \frac{1}{\theta} + \frac{4d^{\frac{p-2}{2p}}\left\|\nabla^2 f(x_\theta)\right\|_p}{\|\nabla f(x_\theta)\|_{p^*}}\left\|\frac{d}{d\theta}x_\theta\right\|_p \\
&\leq 4 + \frac{36LA_t}{\zeta(\theta)} + \frac{d^{\frac{p-2}{p}}\zeta(\theta)}{18LA_t} + \frac{4d^{\frac{p-2}{2p}}\left\|\nabla^2 f(x_\theta)\right\|_p}{\|\nabla f(x_\theta)\|_{p^*}}\left\|\frac{d}{d\theta}x_\theta\right\|_p \\
&\leq 4 + \frac{36LA_t}{\zeta(\theta)} + \frac{d^{\frac{p-2}{p}}\zeta(\theta)}{18LA_t} + \frac{4L\left(R+1458R^2\right)d^{\frac{3(p-2)}{2p}}}{\frac{f(x_\theta)-f(x^*)}{\left(19+d^{\frac{p-2}{p}}\right)R+\left(2916+1458d^{\frac{p-2}{p}}\right)R^2}} && \text{(Lemma 15, 17, 13)} \\
&\leq 4 + \frac{36LA_t}{\zeta(\theta)} + \frac{d^{\frac{p-2}{p}}\zeta(\theta)}{18LA_t} + \frac{4LD_R d^{\frac{5(p-2)}{2p}}}{f(x_\theta)-f(x^*)} \\
&\leq 4 + \frac{36LA_t}{\zeta(\theta)} + 18d^{\frac{p-2}{p}}\zeta(\theta) + \frac{4LD_R d^{\frac{5(p-2)}{3p}}}{f(x_\theta)-f(x^*)} && \text{(Lemma 3)} \\
&\leq \omega(\theta)\left(1 + \frac{1}{\zeta(\theta)} + \zeta(\theta)\right)
\end{aligned}$$

for $\omega(\theta) = 4d^{\frac{5(p-2)}{2p}}\left(6 + 9LA_t + \frac{LD_R}{f(x_\theta)-f(x^*)}\right)$ in which $D_R = \left(R + 1458R^2\right)\left(20R + 4374R^2\right)$.

Now we choose the proper δ. First, we claim that $A_t \leq \frac{R^2}{2\varepsilon}$ and $f(x_\theta) - f(x^*) \geq \varepsilon$ or otherwise we have $f(x_t) - f(x^*) \leq \varepsilon$ by Lemma 11 (1) or $f(x_\theta) - f(x^*) \leq \varepsilon$ by direct negation of $f(x_\theta) - f(x^*) \geq \varepsilon$, in either case we have found a desired solution. Now for $\omega(\theta) = 4d^{\frac{5(p-2)}{2p}}\left(6 + 9LA_t + \frac{LD_R}{f(x_\theta)-f(x^*)}\right)$, we have

$$\begin{aligned}
10\omega(\theta) &= 40d^{\frac{5(p-2)}{2p}}\left(6 + 9LA_t + \frac{LD_R}{f(x_\theta)-f(x^*)}\right) \\
&\leq 40d^{\frac{5(p-2)}{2p}}\left(6 + \frac{9LR^2}{2\varepsilon} + \frac{2LD_R}{\varepsilon}\right) \\
&\leq 320d^{\frac{5(p-2)}{2p}}\left(\frac{LD_R}{\varepsilon}\right)
\end{aligned}$$

for $\varepsilon \leq \frac{LD_R}{6}$ in which $D_R = \left(R + 1458R^2\right)\left(20R + 4374R^2\right)$. Therefore, by choosing $\delta = \frac{\varepsilon}{320d^{\frac{5(p-2)}{2p}}LD_R}$, we have $|\theta - \theta^*| \leq \frac{1}{10\omega(\theta)}$, and by Lemma 18 we confirm that $\zeta(\theta) \in \left[\frac{1}{2}, 2\right]$. And the complexity of binary search to find such θ is $\log_2\left(\frac{1}{\delta}\right) \leq 9 + \frac{5(p-2)}{2p}\log_2(d) + \log_2\left(\frac{LD_R}{\varepsilon}\right)$.

■

D.2. Proof of Supporting Lemmas

Lemma 10 $\forall\, t \in [T]$, $\psi_t\left(x\right) \leq A_t f\left(x\right) + \frac{1}{2}\left\|x - x_0\right\|_2^2$.

Proof By definition,

$$\begin{aligned}\psi_t(x) &= \psi_{t-1}(x) + a_t[f(x_t) + \langle \nabla f(x_t),\, x - x_t\rangle] \\ &\leq \psi_{t-1}(x) + a_t f(x) \\ &\leq \psi_0(x) + \sum_{i=1}^{t} a_i f(x) \\ &= \frac{1}{2}\left\|x - x_0\right\|_2^2 + A_t f\left(x\right),\end{aligned}$$

where the first inequality holds by convexity and the second by applying the first recursively. ■

Lemma 11 $\forall\, t \in [T]$,

(1) $f\left(x_t\right) - f\left(x^*\right) \leq \frac{1}{2A_t}\left\|x_0 - x^*\right\|_2^2$,

(2) $\left\|v_t - x^*\right\|_p \leq \left\|v_t - x^*\right\|_2 \leq \left\|x_0 - x^*\right\|_2 = R$,

(3) $B_t \leq \frac{1}{2}\left\|x_0 - x^*\right\|_2^2$.

Proof Given the definition that $\forall\ t \in [T]$, $\psi_t(x) = \psi_{t-1}(x) + a_t[f(x_t) + \langle \nabla f(x_t),\, x - x_t\rangle]$, we have $\nabla\psi_t(x) = \nabla\psi_{t-1}(x) + a_t\nabla f(x_t)$. Applying this equality recursively, $\nabla\psi_t(x) = \nabla\psi_0(x) + \sum_{i=1}^{t} a_i \nabla f(x_i)$. Given that $\psi_0(x) = \frac{1}{2}\left\|x - x_0\right\|_2^2$, we have $\forall\ t \in [T]$, $\nabla^2\psi_t(x) = \nabla^2\psi_0(x) = \mathbf{I}$ and third-order derivative of $\psi_t(x)$ being zero. As a result, we have for the second-order Taylor expansion of $\psi_t(x)$ at v_t that $\forall\ x$,

$$\begin{aligned}\psi_t(x) &= \psi_t(v_t) + \langle \nabla\psi_t(v_t),\, x - v_t\rangle + \frac{1}{2}\nabla^2\psi_t(v_t)\left\|x - v_t\right\|_2^2 \\ &= \psi_t(v_t) + \frac{1}{2}\left\|x - v_t\right\|_2^2 \qquad (13)\end{aligned}$$

where the first-order term vanishes by the definition $v_t = \underset{x\in\mathbb{R}^d}{\arg\min}\,\psi_t(x)$ which indicates that $\nabla\psi_t(v_t) = 0$. Plugging in x^*, we have by Lemma 2,

$$\begin{aligned}A_t f(x_t) + B_t &\leq \min_{x\in\mathbb{R}^d}\psi_t(x) \\ &\leq \psi_t(v_t) \\ &= \psi_t(x^*) - \frac{1}{2}\left\|x^* - v_t\right\|_2^2 \\ &\leq A_t f(x^*) + \frac{1}{2}\left\|x^* - x_0\right\|_2^2 - \frac{1}{2}\left\|x^* - v_t\right\|_2^2\end{aligned}$$

where the equality follows from Eq. (13) and the last inequality from Lemma 10. Rearranging the terms, we have

$$A_t\left[f(x_t) - f(x^*)\right] + B_t + \frac{1}{2}\left\|x^* - v_t\right\|_2^2 \leq \frac{1}{2}\left\|x^* - x_0\right\|_2^2.$$

Given that $A_t \geq 0$, $f(x_t) - f(x^*) \geq 0$, $B_t \geq 0$, and $\frac{1}{2}\left\|x^* - v_t\right\|_2^2 \geq 0$, we have $A_t\left[f(x_t) - f(x^*)\right] \leq \frac{1}{2}\left\|x^* - x_0\right\|_2^2$ which yields (1), $\frac{1}{2}\left\|x^* - v_t\right\|_2^2 \leq \frac{1}{2}\left\|x^* - x_0\right\|_2^2$ which yields (2), and $B_t \leq \frac{1}{2}\left\|x^* - x_0\right\|_2^2$ which completes the proof. ■

Lemma 12 *$\forall\, t \in [T]$, for $\left\|x_0 - x^*\right\|_2 = R$,*

(1) $\left\|x_t - x^*\right\|_p \leq \left\|x_0 - x^*\right\|_2 + \frac{2916}{t}\left\|x_0 - x^*\right\|_2^2 = R + 2916R^2$,

(2) $\left\|x_t - v_t\right\|_p \leq 2R + 2916R^2$.

Proof

(1) By definition, $y_t = \frac{A_t}{A_{t+1}}x_t + \frac{a_{t+1}}{A_{t+1}}v_t$. Thus,

$$\begin{aligned}
\left\|y_t - x^*\right\|_p &= \left\|\frac{A_t}{A_{t+1}}x_t + \frac{a_{t+1}}{A_{t+1}}v_t - \frac{A_t}{A_{t+1}}x^* - \frac{a_{t+1}}{A_{t+1}}x^*\right\|_p \\
&\leq \frac{A_t}{A_{t+1}}\left\|x_t - x^*\right\|_p + \frac{a_{t+1}}{A_{t+1}}\left\|v_t - x^*\right\|_p
\end{aligned}$$

Also, by Lemma 1 Eq. (6) we have

$$\begin{aligned}
L\left\|x_{t+1} - y_t\right\|_p^2 &\leq \langle \nabla f(x_{t+1}),\, y_t - x_{t+1}\rangle \\
&\leq \left\|\nabla f(x_{t+1})\right\|_{p^*}\left\|y_t - x_{t+1}\right\|_p,
\end{aligned}$$

which indicates that

$$\left\|x_{t+1} - y_t\right\|_p \leq \frac{1}{L}\left\|\nabla f(x_{t+1})\right\|_{p^*}. \tag{14}$$

Then we have

$$\begin{aligned}
\left\|x_{t+1} - x^*\right\|_p &= \left\|x_{t+1} - y_t + y_t - x^*\right\|_p \\
&\leq \left\|x_{t+1} - y_t\right\|_p + \left\|y_t - x^*\right\|_p \\
&\leq \left\|x_{t+1} - y_t\right\|_p + \frac{A_t}{A_{t+1}}\left\|x_t - x^*\right\|_p + \frac{a_{t+1}}{A_{t+1}}\left\|v_t - x^*\right\|_p \\
&\leq \frac{1}{L}\left\|\nabla f(x_{t+1})\right\|_{p^*} + \frac{A_t}{A_{t+1}}\left\|x_t - x^*\right\|_p + \frac{a_{t+1}}{A_{t+1}}\left\|x_0 - x^*\right\|_p
\end{aligned}$$

where the last inequality follows from Eq. (14) and Lemma 12 (1). Applying this inequality recursively,

$$\begin{aligned}
\|x_{t+1}-x^*\|_p &\le \frac{A_0}{A_{t+1}}\|x_0-x^*\|_p + \frac{\sum_{i=1}^{t+1} a_i}{A_{t+1}}\|x_0-x^*\|_p + \frac{1}{L}\sum_{i=0}^{t}\frac{A_{i+1}}{A_{t+1}}\|\nabla f(x_{i+1})\|_{p^*} \\
&= \|x_0-x^*\|_p + \frac{1}{L}\sum_{i=0}^{t}\frac{A_{i+1}^{\frac{1}{2}}}{A_{t+1}}A_{i+1}^{\frac{1}{2}}\|\nabla f(x_{i+1})\|_{p^*} \\
&\le \|x_0-x^*\|_p + \frac{1}{L}\left(\sum_{i=0}^{t}\frac{A_{i+1}}{A_{t+1}^2}\right)\left(\sum_{i=0}^{t}A_{i+1}\|\nabla f(x_{i+1})\|_{p^*}^2\right) \\
&= \|x_0-x^*\|_p + \left(18\sum_{i=0}^{t}\frac{A_{i+1}}{A_{t+1}^2}\right)B_{t+1} \\
&\le \|x_0-x^*\|_p + \left(9\sum_{i=0}^{t}\frac{A_{i+1}}{A_{t+1}^2}\right)\|x_0-x^*\|_2^2
\end{aligned}$$

where the first equality follows from $A_0 = 0$ and $A_{t+1} = \sum_{i=1}^{t+1} a_i$, for the second inequality we applied Cauchy-Schwarz inequality, and for the last inequality we applied Lemma 11 (3). Furthermore,

$$\begin{aligned}
\sum_{i=0}^{t}\frac{A_{i+1}}{A_{t+1}^2} &= \frac{1}{A_{t+1}}\sum_{i=0}^{t}\frac{A_{i+1}}{A_{t+1}} \\
&\le \frac{1}{A_{t+1}}\sum_{i=0}^{t}\frac{A_{i+1}}{A_{i+1}} \\
&= \frac{t+1}{A_{t+1}} \\
&\le \frac{t+1}{\left(\frac{\mathcal{G}_{t+1}(t+1)}{18L^{1/2}}\right)^2} \\
&\le \frac{324}{t+1}
\end{aligned}$$

where the first inequality follows from $A_{t+1} \ge A_{i+1}$ for $i \le t$ since $A_{t+1} = \sum_{i=1}^{t+1} a_i$, and the second inequality from Lemma 3. Therefore,

$$\|x_{t+1}-x^*\|_p \le \|x_0-x^*\|_p + \frac{2916}{t+1}\|x_0-x^*\|_2^2\,.$$

(2) $\|x_t - v_t\|_p = \|x_t - x^* + x^* - v_t\|_p \le \|x_t - x^*\|_p + \|v_t - x^*\|_p \le 2R + 2916R^2$ applying Lemma 11 (2) and Lemma 12 (1).

■

Lemma 13 (Proposition 3 in (Balles et al., 2020)) *$f\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth in the norm $\|\cdot\|_p$ if and only if $\forall\, x \in \mathbb{R}^d$, $\left\|\nabla^2 f(x)\right\|_p \leq L$.*

Lemma 14 *For $z \in \mathbb{R}^d$ and $s\left(z\right) = \left[|z[1]|^{p-2} z[1], \cdots, |z[d]|^{p-2} z[d]\right]^\top$,*

(1) $\nabla_z^2 \left\|z\right\|_p^2 = 2(p-1) \left\|z\right\|_p^{2-p} \operatorname{Diag}\left(|z[1]|^{p-2}, \cdots, |z[d]|^{p-2}\right) + 2(2-p) \left\|z\right\|_p^{2(1-p)} s\left(z\right) s\left(z\right)^\top$,

(2) $\nabla_z^2 \left\|z\right\|_p^2 \succeq 2 \left\|z\right\|_p^{2(1-p)} s\left(z\right) s\left(z\right)^\top$,

(3) $\left\|\nabla_z^2 \left\|z\right\|_p^2\right\|_p \geq \frac{2}{d^{\frac{p-2}{2}}}$.

Proof

(1) Given $\left\|z\right\|_p = \left(\sum_{i=1}^d |z[i]|^p\right)^{\frac{1}{p}}$, we have

$$\nabla_z \left\|z\right\|_p^2 = 2 \left\|z\right\|_p \begin{bmatrix} \frac{1}{p} \left\|z\right\|_p^{1-p} \cdot p \left|z[1]\right|^{p-2} z[1] \\ \vdots \\ \frac{1}{p} \left\|z\right\|_p^{1-p} \cdot p \left|z[d]\right|^{p-2} z[d] \end{bmatrix} = 2 \left\|z\right\|_p^{2-p} \begin{bmatrix} \left|z[1]\right|^{p-2} z[1] \\ \vdots \\ \left|z[d]\right|^{p-2} z[d] \end{bmatrix}$$

Therefore, $\forall\, i, j \in [d]$,

$$\frac{\partial^2}{\partial z[i] \partial z[j]} = \begin{cases} 2(2-p) \left\|z\right\|_p^{2(1-p)} \left|z[i]\right|^{2(p-1)} + 2(p-1) \left\|z\right\|_p^{2-p} \left|z[i]\right|^{p-2} & \text{if } i = j \\ 2(2-p) \left\|z\right\|_p^{2(1-p)} \left|z[i]\right|^{p-2} z[i] \left|z[j]\right|^{p-2} z[j] & \text{if } i \neq j \end{cases},$$

which yields

$$\nabla_z^2 \left\|z\right\|_p^2 = 2(2-p) \left\|z\right\|_p^{2(1-p)} s\left(z\right) s\left(z\right)^\top + 2(p-1) \left\|z\right\|_p^{2-p} \operatorname{Diag}\left(|z[1]|^{p-2}, \cdots, |z[d]|^{p-2}\right).$$

(2) By the result of (1),

$$\begin{aligned} \nabla_z^2 \left\|z\right\|_p^2 &= 2(p-1) \left\|z\right\|_p^{2-p} \operatorname{Diag}\left(|z[1]|^{p-2}, \cdots, |z[d]|^{p-2}\right) \\ &\quad - 2(p-1) \left\|z\right\|_p^{2(1-p)} s\left(z\right) s\left(z\right)^\top + 2 \left\|z\right\|_p^{2(1-p)} s\left(z\right) s\left(z\right)^\top \\ &= 2(p-1) \left\|z\right\|_p^{2(1-p)} \left(\underbrace{\left\|z\right\|_p^p \operatorname{Diag}\left(|z[1]|^{p-2}, \cdots, |z[d]|^{p-2}\right) - s\left(z\right) s\left(z\right)^\top}_{M}\right) \\ &\quad + 2 \left\|z\right\|_p^{2(1-p)} s\left(z\right) s\left(z\right)^\top. \end{aligned}$$

We now show that matrix M is positive semidefinite. $\forall\, v \in \mathbb{R}^d$,

$$\begin{aligned}
v^\top M v &= \|z\|_p^p \, v^\top \mathrm{Diag}\left(|z[1]|^{p-2}, \cdots, |z[d]|^{p-2}\right) v - v^\top s(z)\, s(z)^\top v \\
&= \|z\|_p^p \sum_{i=1}^d v_i^2 \, |z[i]|^{p-2} - \left(v^\top s(z)\right)^2 \\
&= \left(\sum_{i=1}^d |z[i]|^p\right)\left(\sum_{i=1}^d v_i^2 \, |z[i]|^{p-2}\right) - \left(\sum_{i=1}^d v_i \, |z[i]|^{\frac{p-2}{2}} \cdot |z[i]|^{\frac{p-2}{2}} z[i]\right)^2 \\
&\geq \left(\sum_{i=1}^d |z[i]|^p\right)\left(\sum_{i=1}^d v_i^2 \, |z[i]|^{p-2}\right) - \left(\sum_{i=1}^d v_i^2 \, |z[i]|^{p-2}\right)\left(\sum_{i=1}^d |z[i]|^{p-2} z[i]^2\right) \\
&= 0
\end{aligned}$$

where we applied the Cauchy-Schwarz inequality. Therefore, we have

$$\nabla_z^2 \|z\|_p^2 - 2\, \|z\|_p^{2(1-p)} s(z)\, s(z)^\top = 2(p-1)\, \|z\|_p^{2(1-p)} M \succeq \mathbf{0},$$

which completes the proof.

(3) By the result of (2),

$$\left\|\nabla_z^2 \|z\|_p^2\right\|_p \geq 2\, \|z\|_p^{2(1-p)} \left\|s(z)\, s(z)^\top\right\|_p.$$

By definition of induced matrix ℓ_p-norm,

$$\begin{aligned}
\left\|s(z)\, s(z)^\top\right\|_p &= \sup_{v:\, \|v\|_p = 1} \left\|s(z)\, s(z)^\top v\right\|_p \\
&= \|s(z)\|_p \sup_{v:\, \|v\|_p = 1} \left|s(z)^\top v\right| \\
&= \|s(z)\|_p \, \|s(z)\|_{p^*} \\
&\geq \|s(z)\|_2^2 \\
&= \sum_{i=1}^d |z[i]|^{2(p-1)} \\
&= \|z\|_{2(p-1)}^{2(p-1)}
\end{aligned}$$

where the second equality uses the fact that $s(z)^\top v$ is a scalar, the third equality follows from the definition of dual vector norm, and the inequality follows the Cauchy-Schwarz inequality. Finally,

$$\left\|\nabla_z^2 \|z\|_p^2\right\|_p \geq \frac{2\, \|z\|_{2(p-1)}^{2(p-1)}}{\|z\|_p^{2(p-1)}} \geq 2 \left(\frac{\|z\|_{2(p-1)}}{d^{\frac{1}{p} - \frac{1}{2(p-1)}} \|z\|_{2(p-1)}}\right)^{2(p-1)} = \frac{2}{d^{\frac{p-2}{2}}}.$$

■

Lemma 15 $\left\|\frac{d}{d\theta}x_\theta\right\|_p = \left\|\frac{d}{d\theta}z_\theta\right\|_p \leq d^{\frac{p-2}{p}}\left(R + 1458R^2\right)$ *for* $z_\theta = x_\theta - y_\theta$.

Proof Denote $F(x_\theta, y_\theta) = f(y_\theta) + \langle \nabla f(y_\theta), x_\theta - y_\theta\rangle + L\left\|x_\theta - y_\theta\right\|_p^2$. By the definition that $x_\theta = \underset{x\in\mathbb{R}^d}{\arg\min}\, F(x, y_\theta)$ and first optimality condition, we know that

$$\nabla_x F(x_\theta, y_\theta) = \nabla f(y_\theta) + \nabla_x\left(L\left\|x_\theta - y_\theta\right\|_p^2\right) = 0.$$

Taking the gradient with respect to θ on both sides yields

$$\nabla_x^2 F(x_\theta, y_\theta)\frac{d}{d\theta}x_\theta + \nabla_y\nabla_x F(x_\theta, y_\theta)\frac{d}{d\theta}y_\theta = 0, \tag{15}$$

in which $\nabla_x^2 F(x_\theta, y_\theta) = \nabla_x^2\left(L\left\|x_\theta - y_\theta\right\|_p^2\right)$, $\frac{d}{d\theta}y_\theta = x_t - v_t$, and

$$\begin{aligned}\nabla_y\nabla_x F(x_\theta, y_\theta) &= \nabla^2 f(y_\theta) + \nabla_y\nabla_x\left(L\left\|x_\theta - y_\theta\right\|_p^2\right)\\ &= \nabla^2 f(y_\theta)\end{aligned}$$

since $\nabla_y\nabla_x\left(L\left\|x_\theta - y_\theta\right\|_p^2\right) = 0$ by calculation or by symmetry. Therefore, from Eq. (15),

$$\frac{d}{d\theta}x_\theta = \left(\nabla_x^2\left(L\left\|x_\theta - y_\theta\right\|_p^2\right)\right)^{-1}\nabla^2 f(y_\theta)\left(v_t - x_t\right).$$

Taking the ℓ_p-norm on both sides

$$\begin{aligned}\left\|\frac{d}{d\theta}x_\theta\right\|_p &\leq \frac{\left\|\nabla^2 f(y_\theta)\right\|\left\|v_t - x_t\right\|_p}{L\left\|\nabla_x^2\left(\left\|x_\theta - y_\theta\right\|_p^2\right)\right\|_p}\\ &\leq \frac{d^{\frac{p-2}{p}}}{2}\left\|x_t - v_t\right\|_2\\ &\leq d^{\frac{p-2}{p}}\left(R + 1458R^2\right)\end{aligned}$$

where the second inequality follows from Lemma 13 and 14, and the last inequality follows from Lemma 12. By letting $F(z_\theta, y_\theta) = f(y_\theta) + \langle \nabla f(y_\theta), z_\theta\rangle + L\left\|z_\theta\right\|_p^2$ and following the same argument, we can show the result also holds for $\left\|\frac{d}{d\theta}z_\theta\right\|_p$. ■

Lemma 16 $\left\|x_\theta - y_\theta\right\|_p \leq 18R + d^{\frac{p-2}{p}}\left(R + 1458R^2\right)$.

Proof By ℓ_p-smoothness we have $f(x_\theta) \leq f(y_\theta) + \langle \nabla f(y_\theta), x_\theta - y_\theta\rangle + \frac{L}{2}\left\|x_\theta - y_\theta\right\|_p$. Then for

$$\begin{aligned}F(x_\theta, y_\theta) &= f(y_\theta) + \langle \nabla f(y_\theta), x_\theta - y_\theta\rangle + L\left\|x_\theta - y_\theta\right\|_p^2\\ &\geq f(x_\theta) - \frac{L}{2}\left\|x_\theta - y_\theta\right\|_p^2 + L\left\|x_\theta - y_\theta\right\|_p^2\\ &= f(x_\theta) + \frac{L}{2}\left\|x_\theta - y_\theta\right\|_p^2.\end{aligned}$$

As a result, rearranging the terms we get

$$\begin{aligned}
\|x_\theta - y_\theta\|_p^2 &\leq \frac{2}{L}\left(F(x_\theta, y_\theta) - f(x_\theta)\right) \\
&\leq \frac{2}{L}\left(F(x_\theta, y_\theta) - f(x^*)\right) \\
&\leq \frac{2}{L}\left(F(y_\theta, y_\theta) - f(x^*)\right) \\
&= \frac{2}{L}\left(f(y_\theta) - f(x^*)\right)
\end{aligned}$$

where the last inequality follows from that $x_\theta = \underset{x \in \mathbb{R}^d}{\arg\min}\, F(x, y_\theta)$. When $\theta = 1$, we have $y_{\theta=1} = x_t$. Then we have

$$\begin{aligned}
\|x_{\theta=1} - y_{\theta=1}\|_p^2 &\leq \frac{2}{L}\left(f(x_t) - f(x^*)\right) \\
&\leq \frac{1}{LA_t}\|x_0 - x^*\|_2^2
\end{aligned}$$

where the second inequality follows from Lemma 11. Finally, for $\left\|\frac{d}{d\theta}(x_\theta - y_\theta)\right\|_p = \left\|\frac{d}{d\theta} z_\theta\right\|_p \leq d^{\frac{p-2}{p}}\left(R + 1458R^2\right)$, by Taylor's inequality,

$$\begin{aligned}
\|x_\theta - y_\theta\|_2 &\leq \|x_{\theta=1} - y_{\theta=1}\|_p + d^{\frac{p-2}{p}}\left(R + 1458R^2\right)|\theta - 1| \\
&\leq \frac{\|x_0 - x^*\|_2}{\sqrt{L}\frac{\mathcal{G}_t t}{18L^{1/2}}} + d^{\frac{p-2}{p}}\left(R + 1458R^2\right) \\
&\leq 18R + d^{\frac{p-2}{p}}\left(R + 1458R^2\right)
\end{aligned}$$

where we applied Lemma 12, 15, and 3. ■

Lemma 17 $\|\nabla f(x_\theta)\|_{p^*} \geq \frac{f(x_\theta) - f(x^*)}{\left(19 + d^{\frac{p-2}{p}}\right)R + \left(2916 + 1458 d^{\frac{p-2}{p}}\right)R^2}$.

Proof Starting with the definition of convexity and applying Cauchy-Schwarz inequality,

$$\begin{aligned}
f(x_\theta) &\leq f(x^*) + \langle \nabla f(x_\theta), x_\theta - x^* \rangle \\
&\leq f(x^*) + \|\nabla f(x_\theta)\|_{p^*}\|x_\theta - x^*\|_p \\
&\leq f(x^*) + \|\nabla f(x_\theta)\|_{p^*}\|x_\theta - y_\theta + y_\theta - x^*\|_p \\
&\leq f(x^*) + \|\nabla f(x_\theta)\|_{p^*}\left(\|x_\theta - y_\theta\|_p + \|y_\theta - x^*\|_p\right) \\
&\leq f(x^*) + \|\nabla f(x_\theta)\|_{p^*}\left(\|x_\theta - y_\theta\|_p + \|\theta x_t + (1-\theta)v_t - x^*\|_p\right) \\
&\leq f(x^*) + \|\nabla f(x_\theta)\|_{p^*}\left(\|x_\theta - y_\theta\|_p + \theta\|x_t - x^*\|_p + (1-\theta)\|v_t - x^*\|_p\right)
\end{aligned}$$

We can bound the three terms in the parenthesis by Lemma 16, 11, and 12 as follows:

$$\begin{aligned} f(x_\theta) &\le f(x^*) + \|\nabla f(x_\theta)\|_{p^*} \left(18R + d^{\frac{p-2}{p}}\left(R + 1458R^2\right) + \theta\left(R + 2916R^2\right) + (1-\theta)R\right) \\ &\le f(x^*) + \|\nabla f(x_\theta)\|_{p^*} \left(\left(19 + d^{\frac{p-2}{p}}\right)R + \left(2916 + 1458d^{\frac{p-2}{p}}\right)R^2\right) \end{aligned}$$

Rearranging the terms we have

$$\|\nabla f(x_\theta)\|_{p^*} \ge \frac{f(x_\theta) - f(x^*)}{\left(19 + d^{\frac{p-2}{p}}\right)R + \left(2916 + 1458d^{\frac{p-2}{p}}\right)R^2}.$$

■

Lemma 18 *For $\zeta\colon [0,1] \to \mathbb{R}_+$ that satisfies $\forall\, \theta \in [0,1]$, $\left|\frac{d}{d\theta}\log(\zeta(\theta))\right| \le \omega\left(1 + \frac{1}{\zeta(\theta)} + \zeta(\theta)\right)$, and $\exists\, \theta^* \in [0,1]$ such that $\zeta(\theta^*) = \frac{5}{4}$, if it holds for $\tilde{\theta}$ such that $\left|\tilde{\theta} - \theta^*\right| \le \frac{1}{10\omega}$ for some $\omega \ge 0$, then one has for such $\tilde{\theta}$ that*

$$\zeta\left(\tilde{\theta}\right) \in \left[\frac{1}{2}, 2\right].$$

Proof $\forall\, \theta$ such that $|\theta - \theta^*| \le \gamma$, we denote the range of $\zeta(\theta) \in [\alpha, \beta]$. Then we have

$$\begin{aligned} \left|\frac{d}{d\theta}\zeta(\theta)\right| &= \left|\zeta(\theta)\frac{1}{\zeta(\theta)}\frac{d}{d\theta}\zeta(\theta)\right| \\ &= \left|\zeta(\theta)\frac{d}{d\theta}\log(\zeta(\theta))\right| \\ &\le \omega\left(\zeta(\theta) + 1 + \zeta(\theta)^2\right) \\ &\le \left(1 + \beta + \beta^2\right)\omega \end{aligned}$$

By the mean value inequality,

$$|\zeta(\theta) - \zeta(\theta^*)| \le \left(1 + \beta + \beta^2\right)\omega\,|\theta - \theta^*| \le \left(1 + \beta + \beta^2\right)\omega\gamma.$$

Let $\gamma = \frac{3}{4(1+\beta+\beta^2)\omega}$, we have $|\zeta(\theta) - \zeta(\theta^*)| \le \frac{3}{4}$, which implies that if

$$|\theta - \theta^*| \le \frac{3}{4(1 + \beta + \beta^2)\,\omega}$$

holds, then one has $\frac{1}{2} = \zeta(\theta^*) - \frac{3}{4} \le \zeta(\theta) \le \zeta(\theta^*) + \frac{3}{4} = 2$, i.e., $\zeta(\theta) \in \left[\frac{1}{2}, 2\right]$. Therefore, we must have $[\alpha, \beta] \subset \left[\frac{1}{2}, 2\right]$, i.e., $\beta \le 2$. As a result, we verify for $\tilde{\theta}$,

$$\left|\tilde{\theta} - \theta^*\right| \le \frac{1}{10\omega} \le \frac{3}{4(1 + 2 + 2^2)\,\omega} \le \frac{3}{4(1 + \beta + \beta^2)\,\omega},$$

and therefore we have $\zeta\left(\tilde{\theta}\right) \in \left[\frac{1}{2}, 2\right]$. ■

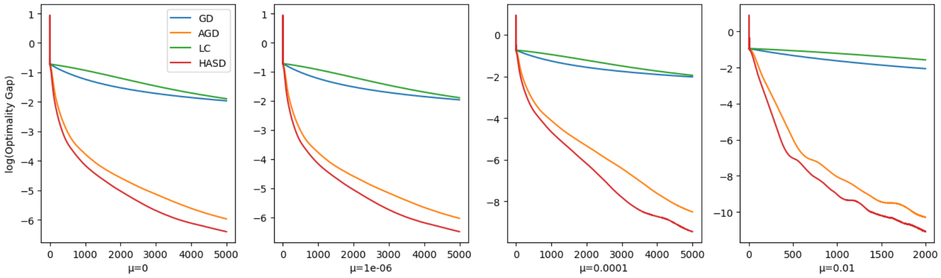

Figure 1: Comparison of HASD (our method) with Gradient Descent (GD), Accelerated Gradient Descent (AGD), and LC w.r.t. $\|\cdot\|_\infty$ (LC), across different choices of $\mu \in \{0, 1e-6, 1e-4, 1e-2\}$. All plots have been parameter tuned (in terms of stepsize) based on the set of possible stepsizes provided. The x-axis of each plot represents the number of iterations, and the y-axis represents $\log(\text{Optimality Gap})$. Notably, the addition of the implicit coupling term leads to significant improvements over the (non-implicitly-coupled) LC method, and our algorithm performs in a manner comparable to (or slightly better than) AGD.

Appendix E. Empirical Validation

We provide in this section experimental evidence to validate (and complement) our theoretical contributions. In order to do so, we consider the function $\text{LogSumExp}(v)$ (related to the softmax function, e.g., (Kelner et al., 2014; Bullins, 2020)), which is defined as

$$\text{LogSumExp}(v) := \log\left(\sum_{i=1}^{n} e^{v[i]}\right) \quad .$$

Next, we consider $A \in \mathbb{R}^{n\times d}$ such that each $A_{i,j}$ is drawn from a Bernouilli with $p = 0.8$, and $b \in \mathbb{R}^n$ such that each b_i is drawn from a normal distribution with mean 0 and variance 1. Following from this, the functions we aim to minimize are

$$f(x) = \text{LogSumExp}(Ax - b) + \frac{\mu}{2}\|x\|^2 ,$$

for $\mu \in \{0, 1e-6, 1e-4, 1e-2\}$. The experiments were run on a MacBook Pro laptop computer.

We compare the methods **Gradient Descent** (GD), **Accelerated Gradient Descent** (AGD), **Linear Coupling** w.r.t. $\|\cdot\|_\infty$ (LC), and our methods **Hyper-Accelerated Steepest Descent** (HASD) w.r.t. $\|\cdot\|_\infty$. We tune the stepsize parameter over the set $\{1e-10, 2e-10, 5e-10, 1e-9, 2e-9, 5e-9, 1e-8, 2e-8, 5e-8, 1e-7, 2e-7, 5e-7, 1e-6, 2e-6, 5e-6, 1e-5, 2e-5, 5e-5, 1e-4, 2e-4, 5e-4, 1e-3, 2e-3, 5e-3, 1e-2, 2e-2, 5e-2, 1e-1, 2e-1, 5e-1, 1\}$, and the results may be found in Figure 1. Notably, our algorithm performs slightly better than AGD, and significantly better than its (non-implicitly-coupled) counterpart LC.