Accelerated Extra-Gradient Descent: A Novel Accelerated First-Order Method*

Jelena Diakonikolas[1] and Lorenzo Orecchia[2]

1 **Department of Computer Science, Boston University, Boston MA 02215, USA**
jelenad@bu.edu

2 **Department of Computer Science, Boston University, Boston MA 02215, USA**
orecchia@bu.edu

Abstract

We provide a novel accelerated first-order method that achieves the asymptotically optimal convergence rate for smooth functions in the first-order oracle model. To this day, Nesterov's Accelerated Gradient Descent (AGD) and variations thereof were the only methods achieving acceleration in this standard blackbox model. In contrast, our algorithm is significantly different from AGD, as it relies on a *predictor-corrector approach* similar to that used by Mirror-Prox [18] and Extra-Gradient Descent [14] in the solution of convex-concave saddle point problems. For this reason, we dub our algorithm Accelerated Extra-Gradient Descent (AXGD).

Its construction is motivated by the discretization of an accelerated continuous-time dynamics [15] using the classical method of implicit Euler discretization. Our analysis explicitly shows the effects of discretization through a conceptually novel primal-dual viewpoint. Moreover, we show that the method is quite general: it attains optimal convergence rates for other classes of objectives (e.g., those with generalized smoothness properties or that are non-smooth and Lipschitz-continuous) using the appropriate choices of step lengths. Finally, we present experiments showing that our algorithm matches the performance of Nesterov's method, while appearing more robust to noise in some cases.

1998 ACM Subject Classification F.2.1 Theory of Computation: Analysis of Algorithms and Problem Complexity, Numerical Algorithms and Problems, G.1.6 Mathematics of Computing: Numerical Analysi, Optimization, Convex programming, Gradient methods

Keywords and phrases Acceleration, dynamical systems, discretization, first-order methods

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.23

1 Introduction

First-order methods for convex optimization have come to play an important role in the design of algorithms and in Theoretical Computer Science in general, with applications including numerical methods [30, 13], graph algorithms [12, 29], submodular optimization [8] and complexity theory [11].

A classical setting for convex optimization is that of smooth optimization, i.e., minimizing a convex differentiable function f over a convex set $X \subseteq \mathbb{R}^n$, with the smoothness assumption

* Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing. It was partially supported by NSF grant #CCF-1718342 and by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF grant #CCF-1740425.

9th Innovations in Theoretical Computer Science Conference (ITCS 2018).
Editor: Anna R. Karlin; Article No. 23; pp. 23:1–23:19
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

that the gradient of f be L-Lipschitz continuous[1] for some positive real L, i.e.:

$$\forall\, x, y \in X,\ \|\nabla f(x) - \nabla f(y)\|_* \leq L \cdot \|x - y\|.$$

In this setting, it is also assumed that the algorithm can access the input function f only via queries to a first-order oracle, i.e., a *blackbox* that on input $x \in X$, returns the vector $\nabla f(x)$ in constant time.[2]

Smooth optimization is of particular interest because it is the simplest setting in which the phenomenon of *acceleration* arises, i.e., the optimal algorithms in the blackbox model achieve an error that scales as $O(1/t^2)$, where t is the number of queries [22]. This should be compared to the convergence of steepest-descent methods, which attempt to locally minimize the first-order approximation to the function and only yield $O(1/t)$-convergence [3, 28]. Acceleration has proved an active topic of algorithmic research, both for the promise of obtaining generic speed-ups for problems having some smoothness condition and for the unintuitive nature of the fact that faster algorithms can be obtained by not moving in the direction of steepest-descent.

Recently, a number of papers have helped demystify the concept behind accelerated algorithms by providing interpretations based on continuous dynamics and their discretization [15, 33, 31], geometric ideas [5], and on trading off the performances of two slower first-order methods [1]. Despite these efforts, to this day, Nesterov's Accelerated Gradient Descent (AGD) methods remain the only paradigm [22, 23] through which to obtain accelerated algorithms in the blackbox model and in related settings, where all existing accelerated algorithms are variations of Nesterov's general method [32].

Our Main Contributions

We present a novel accelerated first-order method that achieves the optimal convergence rate for smooth functions and is significantly different from Nesterov's method, as it relies on a predictor-corrector approach, similar to that of Mirror-Prox [18] and Extra-Gradient Descent [14]. For this reason, we name our method *Accelerated Extra-Gradient Descent* (AXGD). Our derivation of the AXGD algorithm is based on the discretization of a recently proposed continuous-time accelerated algorithm [15, 33]. The continuous-time view is particularly helpful in clarifying the relation between AGD, AXGD, and Mirror-Prox. Following [15], given a gradient field ∇f and a prox function ψ, it is possible to define two continuous-time evolutions: the mirror-descent dynamics and the accelerated-mirror-descent dynamics (see Section 2.2). With this setup, Nesterov's AGD can be seen as a variant of the classical forward-Euler discretization applied to the accelerated-mirror-descent dynamics. In contrast, Mirror-Prox and extra-gradient methods arise from an approximate backward-Euler discretization [9] on the mirror-descent dynamics. Finally, our algorithm AXGD is the result of an approximate backward-Euler discretization of the accelerated mirror-descent dynamics.

Another conceptual contribution of our paper is the application of a primal-dual viewpoint on the convergence of first-order methods, both in continuous and discrete time. At every time instant t, our algorithm explicitly maintains a current primal solution $\mathbf{x}^{(t)}$ and a current dual solution $\mathbf{z}^{(t)}$, the latter in the form of a convex combination of gradients of the convex

1 Lipschitz continuity is defined w.r.t to a pair of dual norms $\|\cdot\|, \|\cdot\|_*$. At a first reading, these can be taken as $\|\cdot\|_2$.

2 In general, we may assume that the blackbox also returns the function value $f(x)$. However, for the general class of problems we consider this information is not necessary and the gradient suffices [28]. For intuition about this, see the expression for the change in duality gap in Equation 4.

objective, i.e., a lower-bounding hyperplane. This primal-dual pair of solutions yields, for every t, both an upper bound U_t and a lower bound L_t on the optimum: $U_t \geq f(\mathbf{x}^*) \geq L_t$. In all cases, we obtain convergence bounds by explicitly quantifying the rate at which the duality gap $G_t = U_t - L_t$ goes to zero. We believe that this primal-dual viewpoint makes the analysis and design of first-order methods easier to carry out. We provide its application to proving other classical results in first-order methods, including Mirror Descent, Mirror-Prox, and Frank-Wolfe algorithms in the upcoming manuscript [7].

Other Technical Contributions

In Section 2.6, we provide a unified convergence proof for standard smooth functions (as defined above) and for functions with Hölder-continuous gradients, a more general notion of smoothness [20]. While this paper focuses on the standard smooth setup, the same techniques easily yield results matching those of AGD methods for the strongly-convex-and-smooth case. Indeed, it is possible to prove that our method is universal, in the sense of Nesterov [21], meaning that it can be composed with a line-search algorithm to yield near-optimal algorithms even when the smoothness parameters of the functions are unknown. We illustrate this phenomenon by showing that (AXGD) also achieves the optimal rate for the optimization of Lipschitz-continuous convex functions, a non-smooth problem.

Finally, we present a suite of experiments comparing AGD, AXGD, and standard gradient methods, showing that the performance of AXGD closely matches that of AGD methods. We also explore the empirical performance of AXGD in the practically and theoretically relevant case in which the queried gradients are corrupted by noise. We show that AXGD exhibits better stability properties than AGD in some cases, leading to a number of interesting theoretical questions on the convergence of AXGD.

1.1 Related Work

In his seminal work [22, 23], Nesterov gave a method for the minimization of convex functions that are smooth with respect to the Euclidean norm, where the function is accessed through a first-order oracle. Nesterov's method converges quadratically faster than gradient descent, at a rate of $O(\frac{1}{t^2})$, which has been shown to be asymptotically optimal [23] for smooth functions in this standard blackbox model [28]. More recently, Nesterov generalized this method to allow non-Euclidean norms in the definition of smoothness [25]. We refer to this generalization of Nesterov's method and to instantiations thereof as AGD methods. Accelerated gradient methods have been widely extended and modified for different settings, including composite optimization [27, 16], cubic regularization [26], and universal methods [21]. They have also found a number of fundamental applications in many algorithmic areas, including machine learning (see [4]) and discrete optimization [17].

An important application of AGD methods concerns the solution of various convex-concave saddle point problems. While these are examples of non-smooth problems, for which the optimal rate is known to be $\Omega(\frac{1}{\sqrt{k}})$ [20], Nesterov showed that the saddle-point structure can be exploited by smoothing the original problem and applying AGD methods on the resulting smooth function [25]. This approach [25, 24] yields an $O(\frac{1}{k})$-convergence for convex-concave saddle point problems with smooth gradients. Surprisingly, at around the same time, Nemirovski [18] gave a very different algorithm, known as Mirror-Prox, which achieves the same complexity for the saddle point problem. Mirror-Prox does not rely on the algorithm or analysis underlying AGD, but is based instead on the idea of an *extra-gradient* step, i.e., a correction step that is performed at every iteration to speed up convergence.

Mirror-Prox can be viewed as an approximate solution to the implicit Euler discretization of the standard mirror descent dynamics of Nemirovski and Yudin [20]. In this fashion, our AXGD algorithm resembles Mirror-Prox as it also makes use of an approximate implicit Euler step to discretize a *different*, accelerated dynamic.

A number of interpretations have been proposed to explain the phenomenon of acceleration in first-order methods. Tseng gives a formal framework that unifies all the different instantiations of AGD methods [32]. More recently, Allen-Zhu and Orecchia [1] cast AGD methods as the result of coupling mirror descent and gradient descent steps. Bubeck *et al.* give an elegant geometric interpretation of the Euclidean instantiation of Nesterov's method [5]. At the same time, Su *et al.* [31], Krichene *et al.* [15], and Wibisono *et al.* [33] have provided characterizations of accelerated methods as discretizations of certain families of ODEs related to the gradient flow of the objective f. Our algorithm is strongly influenced by these works: in particular, the starting point for the derivation of AXGD is the continuous-time accelerated-mirror-descent (AMD) dynamics [15].

1.2 Preliminaries

We focus on continuous and differentiable functions defined on a closed convex set $X \subseteq \mathbb{R}^n$. We assume that there is an arbitrary (but fixed) norm $\|\cdot\|$ associated with the space, and all the statements about function properties are stated with respect to that norm. We also define the dual norm $\|\cdot\|_*$ in the standard way: $\|\mathbf{z}\|_* = \sup\{\langle \mathbf{z}, \mathbf{x}\rangle : \|\mathbf{x}\| = 1\}$. The following definitions will be useful in our analysis, and thus we state them here for completeness.

▶ **Definition 1.** A function $f : X \to \mathbb{R}$ is convex on X, if for all $\mathbf{x}, \hat{\mathbf{x}} \in X$: $f(\hat{\mathbf{x}}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}\rangle$.

▶ **Definition 2.** A function $f : X \to \mathbb{R}$ is smooth on X with smoothness parameter L and with respect to a norm $\|\cdot\|$, if for all $\mathbf{x}, \hat{\mathbf{x}} \in X$: $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}\rangle + \frac{L}{2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2$.

Definition 2 can equivalently be stated as: $\|\nabla f(\mathbf{x}) - \nabla f(\hat{\mathbf{x}})\|_* \leq L\|\mathbf{x} - \hat{\mathbf{x}}\|$.

▶ **Definition 3.** A function $f : X \to \mathbb{R}$ is strongly convex on X with strong convexity parameter σ and with respect to a norm $\|\cdot\|$, if for all $\mathbf{x}, \hat{\mathbf{x}} \in X$: $f(\hat{\mathbf{x}}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}\rangle + \frac{\sigma}{2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2$.

▶ **Definition 4.** (Bregman Divergence) $D_\psi(\mathbf{x}, \hat{\mathbf{x}}) \stackrel{\text{def}}{=} \psi(\mathbf{x}) - \psi(\hat{\mathbf{x}}) - \langle \nabla\psi(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}}\rangle$.

▶ **Definition 5.** (Convex Conjugate) Function ψ^* is the convex conjugate of $\psi : X \to \mathbb{R}$, if $\psi^*(\mathbf{z}) = \max_{\mathbf{x} \in X}\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}$, $\forall \mathbf{z} \in \mathbb{R}$.

In the rest of the paper, we will assume that $\psi(\mathbf{x})$ is continuously differentiable, so that Fenchel-Moreau Theorem implies that $\psi^{**} = \psi$.[3] We are interested in minimizing a convex function f over $X \subseteq \mathbb{R}^n$. We let $\mathbf{x}^* = \arg\min_{\mathbf{x} \in X} f(\mathbf{x})$.

We will refer to any step that decreases the value of f as a gradient descent step. In the special case of a smooth function f the gradient descent step from a point $\mathbf{x} \in X$ will be given as $\mathrm{Grad}(\mathbf{x}) = \arg\min_{\hat{\mathbf{x}} \in X}\{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}\rangle + \frac{L}{2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2\}$.

We will assume that there is a strongly-convex differentiable function $\psi : X \to \mathbb{R}$ such that $\max_{\mathbf{x} \in X}\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}$ is easily solvable, possibly in a closed form. Notice that this

[3] Note that Fenchel-Moreau Theorem requires ψ to only be lower-semicontinuous for $\psi^{**} = \psi$ to hold, which is a weaker property than continuity or continuous differentiability.

problem defines the convex conjugate of $\psi(\cdot)$, i.e., $\psi^*(\mathbf{z}) = \max_{\mathbf{x}\in X}\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}$. The following standard fact will be extremely useful in carrying out the analysis of the algorithms in this paper.

▶ **Fact 6.** Let $\psi : X \to \mathbb{R}$ be a differentiable strongly-convex function. Then:

$$\nabla\psi^*(\mathbf{z}) = \arg\max_{\mathbf{x}\in X}\left\{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\right\}.$$

Additional useful properties of Bregman divergence are provided in Appendix A.

2 Accelerated Extra-Gradient Descent

In this section, we describe the AXGD method and analyze its convergence. For comparison, steps of AGD and AXGD are shown next to each other in the box below. In continuous time, both algorithms follow the same dynamics. However, due to the different discretization methods used in constructing AGD and AXGD, they follow different discrete-time updates. In particular, we show in [7] that AGD can be interpreted as performing explicit (forward) Euler discretization plus a gradient step to correct the discretization error. In contrast, AXGD uses an approximate implementation of implicit (backward) Euler discretization to directly control the discretization error.

Accelerated Gradient Descent (AGD)

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= \frac{A_k}{A_{k+1}}\hat{\mathbf{x}}^{(k)} + \frac{a_{k+1}}{A_{k+1}}\nabla\psi^*(\mathbf{z}^{(k)}),\\
\mathbf{z}^{(k+1)} &= \mathbf{z}^{(k)} - a_{k+1}\nabla f(\mathbf{x}^{(k+1)}),\\
\hat{\mathbf{x}}^{(k+1)} &= \mathrm{Grad}(\mathbf{x}^{(k+1)}).
\end{aligned} \tag{1}$$

Accelerated Extra-Gradient Descent (AXGD)

$$\begin{aligned}
\hat{\mathbf{x}}^{(k)} &= \frac{A_k}{A_{k+1}}\mathbf{x}^{(k)} + \frac{a_{k+1}}{A_{k+1}}\nabla\psi^*(\mathbf{z}^{(k)}),\\
\hat{\mathbf{z}}^{(k)} &= \mathbf{z}^{(k)} - a_{k+1}\nabla f(\hat{\mathbf{x}}^{(k)}),\\
\mathbf{x}^{(k+1)} &= \frac{A_k}{A_{k+1}}\mathbf{x}^{(k)} + \frac{a_{k+1}}{A_{k+1}}\nabla\psi^*(\hat{\mathbf{z}}^{(k)}),\\
\mathbf{z}^{(k+1)} &= \mathbf{z}^{(k)} - a_{k+1}\nabla f(\mathbf{x}^{(k+1)}).
\end{aligned} \tag{2}$$

The idea behind AXGD is similar to the dual-averaging version of Nemirovski's mirror prox algorithm [18, 7], with the main difference coming from the discretization of the accelerated dynamics in Equation (5) (as opposed to the standard mirror descent dynamics used in [18]). As we will show, an exact implicit Euler step would have $\nabla\psi^*(\mathbf{z}^{(k+1)})$ instead of $\nabla\psi^*(\hat{\mathbf{z}}^{(k)})$ in the third line of AXGD. However, obtaining $\mathbf{x}^{(k+1)}$ in a such a manner could be computationally prohibitive since $\mathbf{z}^{(k+1)}$ implicitly depends on $\mathbf{x}^{(k+1)}$ through its gradient. Instead, we opt for an extra prox-step $\nabla\psi^*(\hat{\mathbf{z}}^{(k)})$ that adds the gradient at an intermediate point $\hat{\mathbf{x}}^{(k)}$ constructed using $\mathbf{x}^{(k)}$ and $\mathbf{z}^{(k)}$ from the previous iteration. Thanks to this extra-gradient step, AXGD can correct the discretization error without using a gradient step.

Convergence proof for AXGD together with the sufficient conditions for obtaining optimal convergence bounds are provided in Section 2.4. For example, Theorem 11 shows that when the objective function is smooth, AXGD converges at the optimal rate of $1/k^2$. The analysis of AGD is provided in [7].

2.1 Approximate Optimality Gap

The analysis relies on the construction of an approximate optimality gap G_t, which is defined as the difference of an upper bound U_t and a lower bound L_t to the optimal function value $f(\mathbf{x}^*)$. In particular, for an increasing function of time t, $\alpha^{(t)}$, the convergence analysis will work on establishing the following:

Invariance condition: $\alpha^{(t)}G_t$ is non-increasing with time t.

Such a condition immediately implies: $G_t \leq \frac{\alpha^{(t_0)}}{\alpha^{(t)}} G_{t_0}$, leading to the $\frac{1}{\alpha^{(t)}}$ convergence rate. We sketch the main ideas that relate to the accelerated methods and AXGD in particular here for completeness, while the more general arguments that recover a number of known first-order methods are provided in [7].

We now describe the upper bound and the lower bound choices, which will take the same form in both continuous time and discrete time domains. To do so, we will rely on the Lebesgue-Stieltjes integration, which allows us to treat continuous and discrete choice of $\alpha^{(t)}$ in a unified manner. Observe that when $\alpha^{(t)}$ is a discrete measure, $\dot{\alpha}^{(t)}$ is a train of (scaled) Dirac Delta functions. Denote $A^{(t)} = \int_{t_0}^{t} d\alpha^{(\tau)} = \int_{t_0}^{t} \dot{\alpha}^{(\tau)} d\tau$.

Upper Bound

As $\mathbf{x}^* \in X$ is the minimizer of $f(\cdot)$, $f(\mathbf{x})$ for any $\mathbf{x} \in X$ constitutes a valid upper bound. In particular, our choice of the upper bound will be $U_t = f(\mathbf{x}^{(t)})$, where $\mathbf{x}^{(t)}$ is the solution maintained by the algorithm at time t.

Lower Bound

More interesting than the upper bound is the construction of a lower bound to $f(\mathbf{x}^*)$. From convexity of f, we have the standard lower-bounding hyperplanes $\forall \mathbf{x}, \hat{\mathbf{x}} \in X$: $f(\mathbf{x}) \geq f(\hat{\mathbf{x}}) + \langle \nabla f(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle$. A natural choice of a lower bound to the optimum at time $t \geq t_0$, is obtained by averaging such hyperplanes over $[t_0, t]$ according to the measure α:

$$f(\mathbf{u}) \geq \frac{\int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}}{A^{(t)}} + \frac{\int_{t_0}^{t} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{u} - \mathbf{x}^{(\tau)} \right\rangle d\alpha^{(\tau)}}{A^{(t)}}, \ \forall \mathbf{u} \in X.$$

While we could take the minimum over $\mathbf{u} \in X$ on the right-hand side of this equation as our notion of lower bound, this choice has two serious drawbacks. First, it is non-smooth, and in general not even differentiable, as a function of t. Second, in continuous-time, it is not defined for our initial time t_0, meaning that we do not have a natural concept of initial lower bound and initial duality gap. (In the discrete time, we can ensure that α contains a Dirac Delta function at t_0, which overcomes this issue.) We address the first problem by applying regularization, i.e., by adding to both sides of the inequality a regularizer term that is strongly-convex in $\mathbf{x}$ and then minimizing the right-hand side with respect to $\mathbf{u} \in X$.[4] Without loss of generality, the regularizer can be taken to be the Bregman divergence of a σ-strongly convex function ψ taken from an input point $\mathbf{x}^{(t_0)}$. This yields:

$$\begin{aligned} f(\mathbf{x}^*) &+ \frac{D_\psi(\mathbf{x}^*, \mathbf{x}^{(t_0)})}{A^{(t)}} \\ &\geq \frac{\int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}}{A^{(t)}} + \frac{\min_{\mathbf{u} \in X} \left\{ \int_{t_0}^{t} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{u} - \mathbf{x}^{(\tau)} \right\rangle d\alpha^{(\tau)} + D_\psi(\mathbf{u}, \mathbf{x}^{(t_0)}) \right\}}{A^{(t)}}. \end{aligned}$$

To address the second problem, we mix into the α-combination of hyperplanes the optimal lower bound $f(\mathbf{x}^*)$ with weight $\alpha^{(t)} - A^{(t)}$ (which is just zero in the discrete time, as in that case $A^{(t)} = \alpha^{(t)}$). Rescaling the normalization factor, we obtain our notion of *regularized*

[4] This is similar to the well-known Moreau-Yosida regularization.

lower bound:

$$
\begin{aligned}
L_t \stackrel{\text{def}}{=} & \frac{\int_{t_0}^t f(\mathbf{x}^{(\tau)})d\alpha^{(\tau)}}{\alpha^{(t)}} + \frac{\min_{\mathbf{u}\in X}\left\{\int_{t_0}^t \left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{u}-\mathbf{x}^{(\tau)}\right\rangle d\alpha^{(\tau)} + D_\psi(\mathbf{u}, \mathbf{x}^{(t_0)})\right\}}{\alpha^{(t)}} \\
& + \frac{(\alpha^{(t)} - A^{(t)})f(\mathbf{x}^*) - D_\psi(\mathbf{x}^*, \mathbf{x}^{(t_0)})}{\alpha^{(t)}}.
\end{aligned} \tag{3}
$$

2.2 Accelerated Mirror Descent in Continuous Time

We now show that the accelerated dynamics can be obtained by enforcing the invariance condition from previous subsection with $\alpha^{(t)}G_t$ being constant; i.e., we enforce that $\frac{d}{dt}(\alpha^{(t)}G_t) = 0$. Towards that goal, assume that $\alpha^{(t)}$ is continuously differentiable, and observe that $\alpha^{(t)} - A^{(t)} = \alpha^{(t_0)}$ is constant. To simplify the notation when taking the time derivative of $\alpha^{(t)}G^{(t)}$, we first show the following:

▶ **Proposition 7.** Let $\mathbf{z}^{(t)} = \nabla\psi(\mathbf{x}^{(t_0)}) - \int_{t_0}^t \nabla f(\mathbf{x}^{(\tau)})d\alpha^{(\tau)}$. Then:

$$
\nabla\psi^*(\mathbf{z}^{(t)}) = \arg\min_{\mathbf{u}\in X}\left\{\int_{t_0}^t \left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{u}-\mathbf{x}^{(\tau)}\right\rangle d\alpha^{(\tau)} + D_\psi(\mathbf{u}, \mathbf{x}^{(t_0)})\right\}.
$$

I.e., $\nabla\psi^*(\mathbf{z}^{(t)})$ is the argument of the minimum appearing in the definition of lower bound L_t. The proof is simple and is provided in the appendix.

Recalling that $U_t = f(\mathbf{x}^{(t)})$ and using (3) and Danskin's theorem (which allows us to differentiate inside the min):

$$
\begin{aligned}
\frac{d}{dt}(\alpha^{(t)}G_t) &= \frac{d}{dt}(\alpha^{(t)}f(\mathbf{x}^{(t)})) - \dot{\alpha}^{(t)}f(\mathbf{x}^{(t)}) - \dot{\alpha}^{(t)}\left\langle \nabla f(\mathbf{x}^{(t)}), \nabla\psi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}\right\rangle \\
&= \left\langle \nabla f(\mathbf{x}^{(t)}), \alpha^{(t)}\mathbf{x}^{(t)} - \dot{\alpha}^{(t)}\left(\nabla\psi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}\right)\right\rangle.
\end{aligned} \tag{4}
$$

Hence, to obtain $\frac{d}{dt}(\alpha^{(t)}G_t) = 0$, it suffices to set $\alpha^{(t)}\mathbf{x}^{(t)} = \dot{\alpha}^{(t)}(\nabla\psi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)})$, resulting in the accelerated dynamics from [15]:

$$
\begin{aligned}
\dot{\mathbf{z}}^{(t)} &= -\dot{\alpha}^{(t)}\nabla f(\mathbf{x}^{(t)}), \\
\dot{\mathbf{x}}^{(t)} &= \dot{\alpha}^{(t)}\frac{\nabla\psi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}, \\
\mathbf{z}^{(t_0)} &= \nabla\psi(\mathbf{x}^{(t_0)}),\ \mathbf{x}^{(t_0)} \in X \text{ is an arbitrary initial point.}
\end{aligned} \tag{5}
$$

It is not hard to see that (5) constructs a sequence of points $\mathbf{x}^{(t)}$ that are feasible, that is, $\mathbf{x}^{(t)} \in X$. This is because $\mathbf{x}^{(t)}$ can equivalently be written as $\frac{d}{dt}(\alpha^{(t)}\mathbf{x}^{(t)}) = \dot{\alpha}^{(t)}\nabla\psi^*(\mathbf{z}^{(t)})$, which, after integrating over $\tau \in [t_0, t]$, gives $\mathbf{x}^{(t)} = \frac{\alpha^{(t_0)}}{\alpha^{(t)}}\mathbf{x}^{(t_0)} + \frac{1}{\alpha^{(t)}}\int_{t_0}^t \nabla\psi^*(\mathbf{z}^{(\tau)})d\alpha^{(\tau)}$ – a convex combination of $\mathbf{x}^{(t_0)}$ and $\nabla\psi^*(\mathbf{z}^{(\tau)})$ for $\tau \in [t_0, t]$. By (5), $\mathbf{x}^{(t_0)} \in X$, while $\nabla\psi^*(\mathbf{z}^{(\tau)}) \in X$ by Proposition 7.

We immediately obtain the following continuous-time convergence guarantee:

▶ **Lemma 8.** *Let $\mathbf{x}^{(t)}$ evolve according to (5). Then, $\forall t \geq t_0$:*

$$
f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{\alpha^{(t_0)}(f(\mathbf{x}^{(t_0)}) - f(\mathbf{x}^*)) + D_\psi(\mathbf{x}^*, \mathbf{x}^{(t_0)})}{\alpha^{(t)}}.
$$

Proof. We have already established that $\frac{d}{dt}(\alpha^{(t)}G^{(t)}) = 0$, and, therefore, $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq G_t = \frac{\alpha^{(t_0)}}{\alpha^{(t)}}G_{t_0}$. Observing that $G_{t_0} = f(\mathbf{x}^{(t_0)}) - f(\mathbf{x}^*) + D_\psi(\mathbf{x}^*, \mathbf{x}^{(t_0)})/\alpha^{(t_0)}$, the proof follows. ◀

2.3 Discretization

As discussed in Section 2.1, our construction of the approximate optimality gap is valid both in the continuous time and in the discrete time domain. To understand where the discretization error occurs, we make the following observations. First, the upper bound does not involve any integration, and thus cannot incur a discretization error. In the lower bound (3), the role of the first integral is only to perform weighted averaging, which is the same in the continuous time and in the discrete time, and, therefore, does not incur a discretization error. The terms that are not integrated over look the same whether or not $\alpha^{(t)}$ is discrete. Therefore, the only term that can incur the discretization error is the integral under the min: $I^{(t_0,t)} = \int_{t_0}^{t} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \nabla\psi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(\tau)} \right\rangle d\alpha^{(\tau)}$.

As mentioned before, when α is a discrete measure, we can express it as $\alpha^{(t)} = \sum_{i=1}^{\infty} a_i\delta(t-(t_0+i-1))$, where $\delta(\cdot)$ denotes the Dirac Delta function and a_i's are positive. Then $A^{(t)} = \int_{t_0}^{t} d\alpha^{(\tau)} = \sum_{i:t_0+i-1\leq t} a_i$. To simplify the notation, we will use $i \in \mathbb{Z}_+$ to denote the discrete time points corresponding to $t_0 + i - 1$ on the continuous line. Therefore, the discretization error incurred in $A^{(t)}L_t$ between the discrete time points i and $i+1$ (understood as integrating from i^+ to $(i+1)^+$) is $I^{(i,i+1)} - I_c^{(i,i+1)}$, where $I_c^{(i,i+1)}$ is the continuous approximation of $I^{(i,i+1)}$ (i.e., we allow continuous integration rules in $I_c^{(i,i+1)}$). We can now establish the following bound on the discretization error.

▶ **Lemma 9.** *Let $A_{i+1}G_{i+1} - A_iG_i \equiv E_{i+1}$ be the discretization error. Then*

$$G_k = \frac{A_1}{A_k}G_1 + \frac{\sum_{i=1}^{k} E_i}{A_k}$$

and

$$E_{i+1} \leq \left\langle \nabla f(\mathbf{x}^{(i+1)}), A^{(i+1)}\mathbf{x}^{(i+1)} - A^{(i)}\mathbf{x}^{(i)} - a_{i+1}\nabla\psi^*(\mathbf{z}^{(i+1)}) \right\rangle - D_{\psi^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}).$$

Proof. The first part of the lemma follows by summing over $1 \leq i \leq k$. For the second part, we have already argued that $E_{i+1} = I_c^{(i,i+1)} - I^{(i,i+1)}$. For the discrete integral $I^{(i,i+1)}$, as $\dot{\alpha}^{(t)}$ just samples the function under the integral at point $i+1$, we have:

$$I^{(i,i+1)} = a_{i+1} \left\langle \nabla f(\mathbf{x}^{(i+1)}), \nabla\psi^*(\mathbf{z}^{(i+1)}) - \mathbf{x}^{(i+1)} \right\rangle. \tag{6}$$

For the continuous integral, using (5) and integration by parts:

$$\begin{aligned} I_c^{(i,i+1)} &= \int_i^{i+1} \alpha^{(\tau)} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \dot{\mathbf{x}}^{(\tau)} \right\rangle d\tau \\ &\quad + \int_i^{i+1} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \nabla\psi^*(\mathbf{z}^{(i+1)}) - \nabla\psi^*(\mathbf{z}^{(\tau)}) \right\rangle d\alpha^{(\tau)} \\ &= A^{(i)}(f(\mathbf{x}^{(i+1)}) - f(\mathbf{x}^{(i)})) - \int_i^{i+1} \left\langle \dot{\mathbf{z}}^{(\tau)}, \nabla\psi^*(\mathbf{z}^{(i+1)}) - \nabla\psi^*(\mathbf{z}^{(\tau)}) \right\rangle d\tau \\ &= A^{(i)}(f(\mathbf{x}^{(i+1)}) - f(\mathbf{x}^{(i)})) - D_{\psi^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}), \end{aligned} \tag{7}$$

where we have used $\dot{\mathbf{z}}^{(\tau)} = -\dot{\alpha}^{(\tau)}\nabla f(\mathbf{x}^{(\tau)})$, $\nabla_{\mathbf{z}^{(\tau)}} D_{\psi^*}(\mathbf{z}^{(\tau)}, \mathbf{z}^{(i+1)}) = \nabla\psi^*(\mathbf{z}^{(\tau)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})$, and $D_{\psi^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i)}) = 0$.

By convexity of f, $f(\mathbf{x}^{(i+1)}) - f(\mathbf{x}^{(i)}) \leq \left\langle \nabla f(\mathbf{x}^{(i+1)}), \mathbf{x}^{(i+1)} - \mathbf{x}^{(i)} \right\rangle$. Combining with (6) and (7):

$$E_{i+1} \leq \left\langle \nabla f(\mathbf{x}^{(i+1)}), A^{(i+1)}\mathbf{x}^{(i+1)} - A^{(i)}\mathbf{x}^{(i)} - a_{i+1}\nabla\psi^*(\mathbf{z}^{(i+1)}) \right\rangle - D_{\psi^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}),$$

as claimed. ◀

We remark that the same result for the discretization error can be obtained by directly computing $A_{i+1}G_{i+1} - A_iG_i$ under a discrete measure α (where all the integrals in the definition of the duality gap are replaced by summations). We have chosen to work with the integration error described above to demonstrate the cause of the discretization error.

We now describe how AXGD cancels out the discretization error by (approximately) implementing implicit Euler discretization of $\dot{\mathbf{x}}^{(t)}$.

Implicit Euler Discretization

Implicit Euler discretization is an abstract discretization method which defines the next iterate $\mathbf{x}^{(k+1)}$ implicitly as a function of the gradient at $\mathbf{x}^{(k+1)}$. In the case of the AMD dynamics, implicit Euler discretization yields the following algorithm: let $\mathbf{x}^{(1)} \in X$ be an arbitrary initial point that satisfies $\mathbf{x}^{(1)} = \nabla\psi^*(\mathbf{z}^{(1)})$, where $\mathbf{z}^{(1)} = \nabla\psi(\mathbf{x}^{(1)}) - \nabla f(\mathbf{x}^{(1)})$; for all $k \geq 1$

$$\begin{cases} \mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} - a_{k+1}\nabla f(\mathbf{x}^{(k+1)}), \\ \mathbf{x}^{(k+1)} = \frac{A_k}{A_{k+1}}\mathbf{x}^{(k)} + \frac{a_{k+1}}{A_{k+1}}\nabla\psi^*(\mathbf{z}^{(k+1)}) \end{cases} \tag{8}$$

Observe that $\mathbf{x}^{(k+1)}$ in (8) exactly sets the inner product in E_{i+1} (Lemma 9) to zero, leaving only the negative term $-D_{\psi^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i+1)})$. While this discretization is not computationally feasible in practice, as it requires solving for the implicitly defined $\mathbf{x}^{(k+1)}$, it also boasts a negative discretization error, i.e., it converges faster than the continuous-time AMD. Ultimately, we will use this extra slack to trade-off the error arising from an *approximate* implicit discretization.

2.4 Convergence of AXGD

A standard way to implement implicit Euler discretization in the solution of ODEs [9] is to replace the exact solution of the implicit equation with a small number of fixed point iterations of the same equation. In our case, the implicit equation can be written as:

$$\mathbf{x}^{(k+1)} = \frac{A_k}{A_{k+1}}\mathbf{x}^{(k)} + \frac{a_{k+1}}{A_{k+1}}\nabla\psi^*(\mathbf{z}^{(k)} - a_{k+1}\nabla f(\mathbf{x}^{(k+1)})).$$

Two steps of the fixed-point iteration yield the following updates, which are exactly those performed by AXGD:

$$\begin{cases} \hat{\mathbf{x}}^{(k)} = \frac{A_k}{A_{k+1}}\mathbf{x}^{(k)} + \frac{a_{k+1}}{A_{k+1}}\nabla\psi^*(\mathbf{z}^{(k)}). \\ \mathbf{x}^{(k+1)} = \frac{A_k}{A_{k+1}}\mathbf{x}^{(k)} + \frac{a_{k+1}}{A_{k+1}}\nabla\psi^*(\mathbf{z}^{(k)} - a_{k+1}\nabla f(\hat{\mathbf{x}}^{(k)})) \end{cases}$$

We can now analyze AXGD as producing an approximate solution to the implicit Euler discretization problem. The following lemma gives a general bound on the convergence of AXGD for a convex and differentiable $f(\cdot)$ without additional assumptions. The only (mild) difference is replacing $D_\psi(\mathbf{x}, \mathbf{x}^{(1)})$ and $D_\psi(\mathbf{x}^*, \mathbf{x}^{(1)})$ by $D_\psi(\mathbf{x}, \hat{\mathbf{x}}^{(0)})$ and $D_\psi(\mathbf{x}^*, \hat{\mathbf{x}}^{(0)})$, since we start from the "intermediate" point $\hat{\mathbf{x}}^{(0)}$. This change is only important for bounding the initial gap G_1; everything else is the same as before.

▶ **Lemma 10.** *Consider the AXGD algorithm as described in Equation (2), starting from an arbitrary point $\hat{\mathbf{x}}^{(0)}$ with $\mathbf{z}^{(0)} = \nabla\psi(\hat{\mathbf{x}}^{(0)})$ and $A_0 = 0$. Then the error from Lemma 9 is bounded by:*

$$\begin{aligned} E_{i+1} \leq & a_{i+1}\left\langle \nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})\right\rangle \\ & - D_{\psi^*}(\hat{\mathbf{z}}^{(i)}, \mathbf{z}^{(i+1)}) - D_{\psi^*}(\mathbf{z}^{(i)}, \hat{\mathbf{z}}^{(i)}). \end{aligned}$$

Proof. From Lemma 9:

$$\begin{aligned}E_{i+1} \leq & a_{i+1}\left\langle \nabla f(\mathbf{x}^{(i+1)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})\right\rangle - D_{\psi^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}) \\ = & a_{i+1}\left\langle \nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)}) + \nabla f(\hat{\mathbf{x}}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})\right\rangle \\ & - D_{\psi^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}).\end{aligned}$$

We now use the fact that $a_{i+1} f(\hat{\mathbf{x}}^{(i)}) = \mathbf{z}^{(i)} - \hat{\mathbf{z}}^{(i)}$ together with the standard triangle-inequality for Bregman divergences (see Proposition 17) to show that:

$$\begin{aligned}a_{i+1}\left\langle \nabla f(\hat{\mathbf{x}}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})\right\rangle &= \left\langle \mathbf{z}^{(i)} - \hat{\mathbf{z}}^{(i)}, \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})\right\rangle \\ &= D_{\psi^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i+1)}) - D_{\psi^*}(\hat{\mathbf{z}}^{(i)}, \mathbf{z}^{(i+1)}) - D_{\psi^*}(\mathbf{z}^{(i)}, \hat{\mathbf{z}}^{(i)}),\end{aligned}$$

Combining the results of the last two equations, we get the claimed bound on the error. ◀

2.5 Smooth Minimization with AXGD

We show that AXGD achieves the asymptotically optimal convergence rate of $1/k^2$ for the minimization of an L-smooth convex objective $f(\cdot)$ by applying Lemma 10. The crux of the proof is that we can take sufficiently large steps while keeping the error from Lemma 10 non-positive. In other words, we are able to move quickly through the continuous evolution of AMD by taking large discrete steps.

▶ **Theorem 11.** *Let $f : X \to \mathbb{R}$ be an L-smooth convex function and let $\mathbf{x}^{(k)}, \mathbf{z}^{(k)}, \hat{\mathbf{x}}^{(k)}, \hat{\mathbf{z}}^{(k)}$ be updated according to the* AXGD *algorithm in Equation (2), starting from an arbitrary initial point $\hat{\mathbf{x}}^{(0)} \in X$ with the following initial conditions: $\mathbf{z}^{(0)} = \nabla\psi(\hat{\mathbf{x}}^{(0)})$ and $A_0 = 0$. Let $\psi : X \to \mathbb{R}$ be σ-strongly convex. If $\frac{{a_k}^2}{A_k} \leq \frac{\sigma}{L}$, then for all $k \geq 1$,*

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \frac{D_\psi(\mathbf{x}^*, \hat{\mathbf{x}}^{(0)})}{A_k}.$$

In particular, if $a_k = \frac{k+1}{2} \cdot \frac{\sigma}{L}$ and $\psi(\mathbf{x}) = \frac{\sigma}{2}\|\mathbf{x}\|^2$, then:

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \frac{2L}{(k+1)^2}\|\mathbf{x}^* - \hat{\mathbf{x}}^{(0)}\|^2.$$

Proof. The proof follows directly by applying Lemma 10 and using L-smoothness of f and σ-strong convexity of ψ. In particular, by Cauchy-Schwartz inequality and smoothness:

$$\begin{aligned}&\left\langle \nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\hat{\mathbf{x}}^{(k)}), \nabla\psi^*(\hat{\mathbf{z}}^{(k)}) - \nabla\psi^*(\mathbf{z}^{(k+1)})\right\rangle \\ &\qquad\leq L\|\mathbf{x}^{(k+1)} - \hat{\mathbf{x}}^{(k)}\| \cdot \|\nabla\psi^*(\mathbf{z}^{(k+1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(k)})\|,\end{aligned}$$

and, by Proposition 16

$$\begin{aligned}&D_{\psi^*}(\hat{\mathbf{z}}^{(k)}, \mathbf{z}^{(k+1)}) + D_{\psi^*}(\mathbf{z}^{(k)}, \hat{\mathbf{z}}^{(k)}) \\ &\qquad\geq \frac{\sigma}{2}\left(\|\nabla\psi^*(\hat{\mathbf{z}}^{(k)}) - \nabla\psi^*(\mathbf{z}^{(k+1)})\|^2 + \|\nabla\psi^*(\mathbf{z}^{(k)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(k)})\|^2\right).\end{aligned} \tag{9}$$

From the definition of the steps, $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = \frac{a_{k+1}}{A_{k+1}}(\nabla\psi^*(\hat{\mathbf{z}}^{(k)}) - \nabla\psi^*(\mathbf{z}^{(k)}))$, and, therefore:

$$E_{k+1} \leq \frac{{a_{k+1}}^2}{A_{k+1}} L \cdot pq - \frac{\sigma}{2}(p^2 + q^2),$$

where $p = \|\nabla\psi^*(\hat{\mathbf{z}}^{(k)}) - \nabla\psi^*(\mathbf{z}^{(k+1)})\|$ and $q = \|\nabla\psi^*(\mathbf{z}^{(k)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(k)})\|$. Since, for any p, q, $p^2 + q^2 - 2\alpha pq \geq 0$ whenever $\alpha \leq 1$, it follows that $E_{k+1} \leq 0$ whenever $\frac{{a_{k+1}}^2}{A_{k+1}}\frac{L}{\sigma} \leq 1$, which is true by the theorem assumptions. In particular, for $a_k = \frac{k+1}{2} \cdot \frac{\sigma}{L}$, $A_k = \frac{\sigma}{L}(\frac{(k+1)(k+2)}{4}) \geq \frac{\sigma}{L}\frac{(k+1)^2}{4}$. This proves that $f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \frac{G_1}{A_k}$. It remains to bound G_1. This a simple computation, shown in the appendix, which yields: $G_1 \leq \frac{1}{A_1} D_\psi(\mathbf{x}^*, \hat{\mathbf{x}}^{(0)})$. ◀

2.6 Generalized Smoothness: Hölder-Continuous Gradients

Suppose that $f(\cdot)$ has Hölder-continuous gradients, namely, $f(\cdot)$ then satisfies:

$$\|\nabla f(\hat{\mathbf{x}}) - \nabla f(\mathbf{x})\| \leq L_\nu \|\hat{\mathbf{x}} - \mathbf{x}\|^\nu, \tag{10}$$

which also implies:

$$\forall \mathbf{x}, \hat{\mathbf{x}} \in X: \; f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}\rangle + \frac{L_\nu}{1+\nu}\|\hat{\mathbf{x}} - \mathbf{x}\|^{1+\nu}, \tag{11}$$

where $\nu \in (0, 1]$, $L_\nu \in \mathbb{R}_{++}$. In particular, if $\nu = 1$, then $f(\cdot)$ is L_ν-smooth. Thus, the functions with Hölder-continuous gradients represent a class of functions with generalized/relaxed smoothness properties.

The lower iteration complexity bound for (unconstrained) minimization of convex functions with Hölder-continuous gradients was established in [20] and equals $O\left(L_\nu D_1^{1+\nu}\epsilon^{-\frac{2}{1+3\nu}}\right)$, where D_1 is the distance from the initial point to the optimal solution. A matching upper bound was obtained in [19].

To recover the optimal convergence rate in the minimization of convex functions with Hölder-continuous gradients, as before, we bound the discretization error from Lemma 10. Before doing so, we will need the following technical proposition (which appears in a similar form as Lemma 3.1 a) in [18]).

▶ **Proposition 12.**

$$\begin{aligned} a_{i+1}\left\langle \nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)}), \nabla\psi^*(\mathbf{z}^{(i+1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\right\rangle \\ \leq \sigma^{-1}{a_{i+1}}^2\|\nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)})\|^2. \end{aligned}$$

The proof is provided in the appendix.

▶ **Theorem 13.** *Let $f(\cdot)$ be a convex function that satisfies (10), and let $\psi(\cdot)$ be σ-strongly convex. Let $\mathbf{x}^{(k)}, \mathbf{z}^{(k)}, \hat{\mathbf{x}}^{(k)}, \hat{\mathbf{z}}^{(k)}$ be updated according to the* AXGD *algorithm in Equation (2), starting from an arbitrary initial point $\hat{\mathbf{x}}^{(0)} \in X$ with the following initial conditions: $\mathbf{z}^{(0)} = \nabla\psi(\hat{\mathbf{x}}^{(0)})$ and $A_0 = 0$. Let $a_k = c\frac{\sigma}{L_\nu}D^{1-\nu}k^{\frac{-1+3\nu}{2}}$, where $D = \max_{\mathbf{x}, \hat{\mathbf{x}} \in X}\|\mathbf{x} - \hat{\mathbf{x}}\|$ and $c = 2^{\frac{3\nu(\nu+1)-1}{2}}$. Then, $\forall k \geq 1$:*

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq 2^{\frac{1-3\nu(\nu+1)}{2}}\frac{L_\nu}{\sigma}\frac{D^{\nu-1}D_\psi(\mathbf{x}^*, \hat{\mathbf{x}}^{(0)})}{k^{\frac{1+3\nu}{2}}}.$$

In particular, if $\psi(\mathbf{x}) = \frac{\sigma}{2}\|\mathbf{x}\|^2$, then:

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq 2^{\frac{1-3\nu(\nu+1)}{2}}L_\nu D^{1+\nu}k^{-\frac{1+3\nu}{2}}.$$

Proof. We prove the theorem by bounding the discretization error E_{i+1} from Lemma 10. Applying Propositions 16 and 12:

$$\begin{aligned}
E_{i+1} =& a_{i+1} \left\langle \nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)}) \right\rangle \\
& - D_{\psi^*}(\hat{\mathbf{z}}^{(i)}, \mathbf{z}^{(i+1)}) - D_{\psi^*}(\mathbf{z}^{(i)}, \hat{\mathbf{z}}^{(i)}) \\
\le& \sigma^{-1} {a_{i+1}}^2 \|\nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)})\|^2 \\
& - \frac{\sigma}{2}\left(\|\nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})\|^2 + \|\nabla\psi^*(\mathbf{z}^{(i)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\|^2 \right) \\
\le& \sigma^{-1} {a_{i+1}}^2 {L_\nu}^2 \|\mathbf{x}^{(i+1)} - \hat{\mathbf{x}}^{(i)}\|^{2\nu} - \frac{\sigma}{2}\|\nabla\psi^*(\mathbf{z}^{(i)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\|^2 \\
\le& \sigma^{-1} {L_\nu}^2 \frac{{a_{i+1}}^{2+2\nu}}{{A_{i+1}}^{2\nu}} \|\nabla\psi^*(\mathbf{z}^{(i)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\|^{2\nu} - \frac{\sigma}{2}\|\nabla\psi^*(\mathbf{z}^{(i)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\|^2, \quad (12)
\end{aligned}$$

where the second inequality is by (10) and the third inequality is by the step definition (2).

Taking $a_k = c\frac{\sigma}{L_\nu} D^{1-\nu} k^{\frac{-1+3\nu}{2}}$, where $c = 2^{\frac{3\nu(\nu+1)-1}{2}}$, it follows that $A_k = \sum_{i=1}^k a_i \ge \sum_{i=\lceil k/2 \rceil}^k a_i \ge \frac{c}{2} D^{1-\nu} \frac{\sigma}{L_\nu} \left(\frac{k}{2}\right)^{\frac{1+3\nu}{2}}$. Therefore, the expression in (12) is at least:

$$\Big(- c^2 2^{3\nu(\nu+1)}(k+1)^{\nu-1} + \frac{1}{2}\Big)\sigma\|\nabla\psi^*(\mathbf{z}^{(k)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(k)})\|^2 \ge 0,$$

as $(k+1)^{\nu-1} \le 1$. Therefore, we have that $G_k \le \frac{A^{(1)}}{A^{(k)}} G_1$, and using similar arguments to bound the initial gap G_1, the proof follows. ◀

2.7 Non-Smooth Minimization: Lipschitz-Continuous Objective

We now show that we can recover the well-known $\frac{1}{\sqrt{k}}$ convergence rate for the class of non-smooth L-Lipschitz objectives by using AXGD. This is summarized in the following theorem. We note that, as in the analysis of classical mirror descent (see, e.g., [3]), the factor $\log(k)$ can be removed if we fix the approximation error (and, consequently, the number of steps k) in advance.

▶ **Theorem 14.** *Let $f(\cdot)$ be a Lipschitz-continuous function with parameter L. Let $\mathbf{x}^{(k)}$, $\mathbf{z}^{(k)}$, $\hat{\mathbf{x}}^{(k)}$, $\hat{\mathbf{z}}^{(k)}$ be updated according to the* AXGD *algorithm in Equation (2), starting from an arbitrary initial point $\hat{\mathbf{x}}^{(0)} \in X$ with the following initial conditions: $\mathbf{z}^{(0)} = \nabla\psi(\hat{\mathbf{x}}^{(0)})$ and $A_0 = 0$. If $a_k = \frac{\sqrt{\sigma}}{2\sqrt{2}L}\sqrt{\frac{D_\psi(\mathbf{x}^*, \hat{\mathbf{x}}^{(0)})}{k}}$, then, $\forall k \ge 1$:*

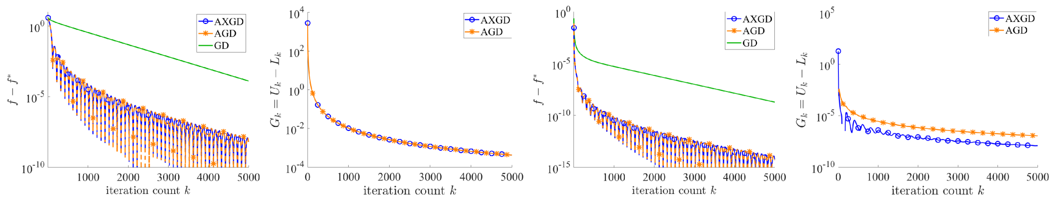
$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \le 8(2 + \log(k)) \frac{L \cdot \sqrt{D_\psi(\mathbf{x}^*, \hat{\mathbf{x}}^{(0)})}}{\sqrt{\sigma}\sqrt{k}}.$$

In particular, for $\psi(\mathbf{x}) = \frac{\sigma}{2}\|\mathbf{x}\|^2$:

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \le 4\sqrt{2}(2 + \log(k)) \frac{L \cdot \|\mathbf{x}^* - \hat{\mathbf{x}}^{(0)}\|}{\sqrt{k}}.$$

Proof. As before, we bound the discretization error from Lemma 10. As $f(\cdot)$ is L-Lipschitz, using Proposition 16:

$$\begin{aligned}
E_{i+1} \le& a_{i+1} \left\langle \nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)}) \right\rangle \\
& - D_{\psi^*}(\hat{\mathbf{z}}^{(i)}, \mathbf{z}^{(i+1)}) - D_{\psi^*}(\mathbf{z}^{(i)}, \hat{\mathbf{z}}^{(i)}) \\
\le& 2a_{i+1} L \|\nabla\psi^*(\mathbf{z}^{(i+1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\| - \frac{\sigma}{2}\|\nabla\psi^*(\mathbf{z}^{(i+1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\|^2 \\
\le& \frac{8(a^{(i+1)}L)^2}{\sigma},
\end{aligned}$$



(a) unconstrained **(b)** unconstrained **(c)** simplex **(d)** simplex

Figure 1 (a),(c) Exact and (b),(d) approximate duality gaps for AGD and AXGD with exact gradients.

where we have used the inequality $2xy - x^2 \le y^2$, $\forall x, y$. As $\sigma \ge L$ and

$$A_k \cdot \frac{2\sqrt{2}L}{\sqrt{\sigma D_\psi(\mathbf{x}^*, \hat{\mathbf{x}}^{(0)})}} = \sum_{i=1}^{k} \frac{1}{\sqrt{k}} \ge \sum_{i=\lceil k/2 \rceil}^{k} \frac{1}{\sqrt{k}} \ge \frac{1}{2} \cdot \sqrt{\frac{k}{2}},$$

we have that

$$\sum_{i=1}^{k} \frac{E_i}{A_k} \le 8 \cdot \frac{L \cdot \sqrt{D_\psi(\mathbf{x}^*, \hat{\mathbf{x}}^{(0)})}}{\sqrt{\sigma}\sqrt{k}} (\log(k) + 1),$$

which, after bounding the initial gap by similar arguments, completes the proof. ◀

3 Experiments

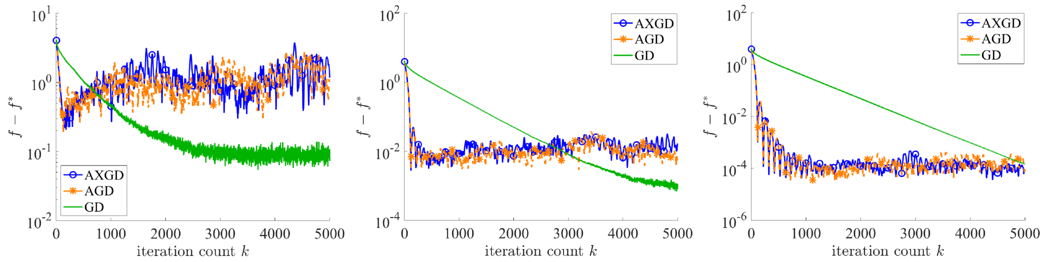
We now illustrate the performance of AGD and AXGD for (i) an unconstrained problem over $\mathbb{R}^n$ with the objective function $f(\mathbf{x}) = \frac{1}{2}\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle$, and (ii) for the problem with the same objective and unit simplex as the feasible region, where $\mathbf{A}$ is the Laplacian of a cycle graph[5] and $\mathbf{b}$ is a vector whose first element is one and the remaining elements are zero. This example is known as a "hard" instance for smooth minimization – it is typically used in proving the lower iteration complexity bound for first-order methods (see, e.g., [28]). We also include Gradient Descent (GD) in the exact gap graphs for comparison. In the experiments, we take $n = 100$ and $\sigma = L$ $(= 4)$. We use the ℓ_2 norm in the gradient steps.

In the figures, f denotes the objective value at the upper-bound point and f^* denotes the optimal objective value, so that $f - f^*$ is the true distance to the optimum (the exact gap). Fig. 1 shows the distance to the optimum and the approximate duality gap $G_k = U_k - L_k$ obtained using our analysis. We can observe that AGD and AXGD exhibit similar performance in these examples. The approximate gap overestimates the actual duality gap, however, the difference between the two decreases with the number of iterations.

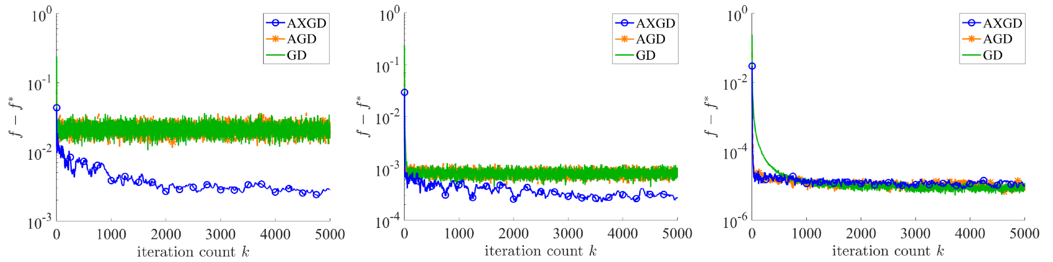
Acceleration and Noise

We now consider the setting in which the gradients output by our oracle are corrupted by additive noise, which has significant applications in practice [10] and theory [2]. We note

[5] Namely, the sum of a tridiagonal matrix $\mathbf{B}$ with 2's on its main diagonal and -1's on its remaining two diagonals and a matrix $\mathbf{C}$ whose all elements are zero except for the $\mathbf{C}_{1,n} = \mathbf{C}_{n,1} = -1$.



(a) $\epsilon_\eta = 10^{-1}$, unconstrained **(b)** $\epsilon_\eta = 10^{-2}$, unconstrained **(c)** $\epsilon_\eta = 10^{-3}$, unconstrained



(d) $\epsilon_\eta = 10^{-1}$, simplex **(e)** $\epsilon_\eta = 10^{-2}$, simplex **(f)** $\epsilon_\eta = 10^{-3}$, simplex

Figure 2 Exact gap for additive Gaussian noise in the gradients with zero mean and covariance $\epsilon_\eta I$ (a)-(c) in the unconstrained-region case and (d)-(f) in unit simplex.

that this model is fundamentally different from the inexact model considered by Devolder *et al.* [6], for which tight lower bounds preventing acceleration exist.[6]

Specifically, we experimentally evaluate the performance of AGD and AXGD under additive Gaussian noise. Fig. 2 illustrates the performance of AGD and AXGD when the gradients are corrupted by zero-mean additive Gaussian noise with covariance matrix $\epsilon_\eta I$, where I is the identity matrix. When the region is unconstrained (top row in Fig. 2), both AGD and AXGD exhibit high sensitivity to noise. The GD method overall exhibits higher tolerance to noise (at the expense of slower convergence). In the case of the unit simplex region (bottom row in Fig. 2), all the algorithms appear more tolerant to noise than in the unconstrained case. Interestingly, on this example AXGD exhibits higher tolerance to noise than GD and AGD, both in terms of mean and in terms of variance. Explaining this phenomenon analytically is an interesting question that merits further investigation.

4 Conclusion

We have presented a novel accelerated method – AXGD– that combines ideas from the Nesterov's AGD and Nemirovski's mirror prox. AXGD achieves optimal convergence rates for a range of convex optimization problems, such as the problems with the (i) smooth objectives, (ii) objectives with Hölder-continuous gradients, (iii) and non-smooth Lipschitz-continuous objectives. In the constrained-regime experiments from Section 3, the method demonstrates favorable performance compared to AGD when subjected to zero-mean Gaussian noise.

There are several directions that merit further investigation. A more thorough analytical and experimental study of acceleration when the gradients are corrupted by noise is of

[6] In [6], it is assumed that a function $f(\cdot)$ is associated with a (δ, L) oracle, such that $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x}\rangle + \frac{L}{2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2 + \frac{\delta}{2}$, $\forall \mathbf{x}, \hat{\mathbf{x}} \in X$. Such a model seems more suitable for incorrectly specified functions (e.g., non-smooth functions treated as being smooth) and adversarially perturbed functions.

particular interest, since the gradients can often come from noise-corrupted measurements. Further, our experiments from Fig. 2 suggest that there are cases that incur a trade-off between noise tolerance and acceleration. A systematic study of this trade-off is thus another important direction, since it would guide the choice of accelerated/non-accelerated algorithms in practice depending on the application. Finally, it is interesting to investigate whether restart schemes can improve the algorithms' noise tolerance, since in the noiseless setting several restart schemes are known to improve the convergence of AGD in practice.

References

1 Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proc. ITCS'17*, 2017.

2 Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proc. IEEE FOCS'14*, 2014.

3 Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: Analysis, algorithms, and engineering applications*. MPS-SIAM Series on Optimization. SIAM, 2001.

4 Sébastien Bubeck. Theory of convex optimization for machine learning. *CoRR*, abs/1405.4980, 2014. `arXiv:1405.4980v1`.

5 Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. *CoRR*, abs/1506.08187, 2015. `arXiv:1506.08187`.

6 Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Prog.*, 146(1-2):37–75, 2014.

7 Jelena Diakonikolas and Lorenzo Orecchia. The approximate gap technique: A unified approach to optimal first-order methods, 2017. Manuscript.

8 A. Ene and H. L. Nguyen. Constrained submodular maximization: Beyond 1/e. In *Proc. IEEE FOCS'16*, 2016.

9 E Hairer, SP Nørsett, and G Wanner. *Solving Ordinary Differential Equations I (2nd Revised. Ed.): Nonstiff Problems*. Springer Ser. Comput. Math. Springer-Verlag New York, Inc., 1993.

10 Moritz Hardt. Robustness vs acceleration, 2014. URL: `http://blog.mrtz.org/2014/08/18/robustness-versus-acceleration.html`.

11 Rahul Jain, Zhengfeng Ji, Sarvagya Upadhyay, and John Watrous. QIP = PSPACE. *Journal of the ACM (JACM)*, 58(6):30, 2011.

12 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 *Proc. ACM-SIAM SODA'14*, 2014.

13 Jonathan A. Kelner, Lorenzo Orecchia, Aaron Sidford, and Zeyuan Allen Zhu. A simple, combinatorial algorithm for solving SDD systems in nearly-Linear time. In *Proc. ACM STOC'13*, 2013.

14 G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matekon : translations of Russian & East European mathematical economics*, 13(4):35–49, 1977.

15 Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Proc. NIPS'15*, 2015.

16 Guanghui Lan. An optimal method for stochastic composite optimization. *Math. Prog.*, 133(1-2):365–397, January 2011.

17 Yin Tat Lee, Satish Rao, and Nikhil Srivastava. A new approach to computing maximum flows using electrical flows. In *Proc. ACM STOC '13*, 2013.

18 Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optimiz.*, 15(1):229–251, 2004.

19 Arkadi S Nemirovski and Yurii Evgen'evich Nesterov. Optimal methods of smooth convex minimization. *Zh. Vychisl. Mat. i Mat. Fiz.*, 25(3):356–369, 1985.

20 Arkadii Nemirovskii and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

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

22 Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Doklady AN SSSR (translated as Soviet Mathematics Doklady)*, volume 269, pages 543–547, 1983.

23 Yurii Nesterov. *Introductory Lectures on Convex Programming Volume: A Basic course*, volume I. Kluwer Academic Publishers, 2004.

24 Yurii Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optimiz.*, 16(1):235–249, January 2005.

25 Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Prog.*, 103(1):127–152, December 2005.

26 Yurii Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Math. Prog.*, 112(1):159–181, 2008.

27 Yurii Nesterov. Gradient methods for minimizing composite functions. *Math. Prog.*, 140(1):125–161, 2013.

28 Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

29 Jonah Sherman. Nearly maximum flows in nearly linear time. In *Proc. IEEE FOCS'13*, 2013.

30 Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proc. ACM STOC '04*, 2004.

31 Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Proc. NIPS'14*, 2014.

32 Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008.

33 Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. In *Proc. Natl. Acad. Sci. U.S.A.*, 2016.

A Properties of the Bregman Divergence

The following properties of Bregman divergence will be useful in our analysis.

▶ **Proposition 15.** $D_\psi(\nabla\psi^*(\mathbf{z}), \mathbf{x}) = D_{\psi^*}(\nabla\psi(\mathbf{x}), \mathbf{z})$, $\forall \mathbf{x}, \mathbf{z}$.

Proof. From the definition of ψ^* and Fact 6,

$$\psi^*(\mathbf{z}) = \langle \nabla\psi^*(\mathbf{z}), \mathbf{z}\rangle - \psi(\nabla\psi^*),\ \forall \mathbf{z}. \tag{13}$$

Similarly, as in the light of Fenchel-Moreau Theorem $\psi^{**} = \psi$,

$$\psi(\mathbf{x}) = \langle \nabla\psi(\mathbf{x}), \mathbf{x}\rangle - \psi^*(\nabla\psi(\mathbf{x})),\ \forall \mathbf{x}. \tag{14}$$

Using the definition of $D_\psi(\nabla\psi^*(\mathbf{z}), \mathbf{x})$ and Fact 6:

$$\begin{aligned} D_\psi(\nabla\psi^*(\mathbf{z}), \mathbf{x}) &= \psi(\nabla\psi^*(\mathbf{z})) - \psi(\mathbf{x}) - \langle \nabla\psi(\mathbf{x}), \nabla\psi^*(\mathbf{z}) - \mathbf{x}\rangle \\ &= \psi(\nabla\psi^*(\mathbf{z})) + \psi^*(\nabla\psi(\mathbf{x})) - \langle \nabla\psi(\mathbf{x}), \nabla\psi^*(\mathbf{z})\rangle\,. \end{aligned} \tag{15}$$

Similarly, using the definition of $D_{\psi^*}(\nabla\psi(\mathbf{x}), \mathbf{z})$ combined with (13):

$$\begin{aligned} D_{\psi^*}(\nabla\psi(\mathbf{x}), \mathbf{z}) &= \psi^*(\nabla\psi(\mathbf{x})) - \psi^*(\mathbf{z}) - \langle\nabla\psi^*(\mathbf{z}), \nabla\psi(\mathbf{x}) - \mathbf{z}\rangle \\ &= \psi^*(\nabla\psi(\mathbf{x})) + \psi(\nabla\psi^*(\mathbf{z})) - \langle\nabla\psi^*(\mathbf{z}), \nabla\psi(\mathbf{x})\rangle. \end{aligned} \tag{16}$$

Comparing (15) and (16), the proof follows. ◀

▶ **Proposition 16.** If $\psi(\cdot)$ is σ-strongly convex, then $D_{\psi^*}(\mathbf{z}, \hat{\mathbf{z}}) \geq \frac{\sigma}{2}\|\nabla\psi^*(\mathbf{z}) - \nabla\psi^*(\hat{\mathbf{z}})\|^2$.

Proof. Using the definition of $D_{\psi^*}(\mathbf{z}, \hat{\mathbf{z}})$ and (13), we can write $D_{\psi^*}(\mathbf{z}, \hat{\mathbf{z}})$ as:

$$D_{\psi^*}(\mathbf{z}, \hat{\mathbf{z}}) = \psi(\nabla\psi^*(\hat{\mathbf{z}})) - \psi(\nabla\psi^*(\mathbf{z})) - \langle\mathbf{z}, \nabla\psi^*(\hat{\mathbf{z}}) - \nabla\psi^*(\mathbf{z})\rangle.$$

Since $\psi(\cdot)$ is σ-strongly convex, it follows that:

$$D_{\psi^*}(\mathbf{z}, \hat{\mathbf{z}}) \geq \frac{\sigma}{2}\|\nabla\psi^*(\hat{\mathbf{z}}) - \nabla\psi^*(\mathbf{z})\|^2 + \langle\nabla\psi(\nabla\psi^*(\mathbf{z})) - \mathbf{z}, \nabla\psi^*(\hat{\mathbf{z}}) - \nabla\psi^*(\mathbf{z})\rangle.$$

As, from Fact 6, $\nabla\psi^*(\mathbf{z}) = \arg\max_{\mathbf{x}\in X}\{\langle\mathbf{x}, \mathbf{z}\rangle - \psi(\mathbf{x})\}$, by the first-order optimality condition

$$\langle\nabla\psi(\nabla\psi^*(\mathbf{z})) - \mathbf{z}, \nabla\psi^*(\hat{\mathbf{z}}) - \nabla\psi^*(\mathbf{z})\rangle \geq 0,$$

completing the proof. ◀

The Bregman divergence $D_{\psi^*}(\mathbf{x}, \mathbf{y})$ captures the difference between $\psi^*(\mathbf{x})$ and its first order approximation at $\mathbf{y}$. Notice that, for a differentiable ψ^*, we have:

$$\nabla_{\mathbf{x}} D_{\psi^*}(\mathbf{x}, \mathbf{y}) = \nabla\psi^*(\mathbf{x}) - \nabla\psi^*(\mathbf{y}).$$

The Bregman divergence $D_{\psi^*}(\mathbf{x}, \mathbf{y})$ is a convex function of $\mathbf{x}$. Its Bregman divergence is itself.

▶ **Proposition 17.** For all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in X$

$$D_{\psi^*}(\mathbf{x}, \mathbf{y}) = D_{\psi^*}(\mathbf{z}, \mathbf{y}) + \langle\nabla\psi^*(\mathbf{z}) - \nabla\psi^*(\mathbf{y}), \mathbf{x} - \mathbf{z}\rangle + D_{\psi^*}(\mathbf{x}, \mathbf{z}).$$

B Omitted Proofs from Section 2

▶ **Proposition 7** (restated). *Let* $\mathbf{z}^{(t)} = \nabla\psi(\mathbf{x}^{(t_0)}) - \int_{t_0}^{t} \nabla f(\mathbf{x}^{(\tau)})d\alpha^{(\tau)}$. *Then:*

$$\nabla\psi^*(\mathbf{z}^{(t)}) = \arg\min_{\mathbf{u}\in X}\left\{\int_{t_0}^{t}\left\langle\nabla f(\mathbf{x}^{(\tau)}), \mathbf{u} - \mathbf{x}^{(\tau)}\right\rangle d\alpha^{(\tau)} + D_\psi(\mathbf{u}, \mathbf{x}^{(t_0)})\right\}.$$

Proof. From the definition of Bregman divergence:

$$\begin{aligned} &\arg\min_{\mathbf{u}\in X}\left\{\int_{t_0}^{t}\left\langle\nabla f(\mathbf{x}^{(\tau)}), \mathbf{u} - \mathbf{x}^{(\tau)}\right\rangle d\alpha^{(\tau)} + D_\psi(\mathbf{u}, \mathbf{x}^{(t_0)})\right\} \\ &= \arg\min_{\mathbf{u}\in X}\left\{\int_{t_0}^{t}\left\langle\nabla f(\mathbf{x}^{(\tau)}), \mathbf{u} - \mathbf{x}^{(\tau)}\right\rangle d\alpha^{(\tau)} + \psi(\mathbf{u}) - \psi(\mathbf{x}^{(t_0)}) - \left\langle\nabla\psi(\mathbf{x}^{(t_0)}), \mathbf{u} - \mathbf{x}^{(t_0)}\right\rangle\right\} \\ &= \arg\min_{\mathbf{u}\in X}\left\{\left\langle\int_{t_0}^{t}\nabla f(\mathbf{x}^{(\tau)})d\alpha^{(\tau)} - \nabla\psi(\mathbf{x}^{(t_0)}), \mathbf{u}\right\rangle + \psi(\mathbf{u})\right\}. \end{aligned}$$

Using the definition of $\mathbf{z}^{(t)}$ and Fact 6, the proof follows. ◀

Remaining Proof of Theorem 11 (The Bound on G_1). To bound G_1, we recall the definition of L_1:

$$\begin{aligned}
L_1 &= f(\mathbf{x}^{(1)}) + \min_{\mathbf{x}\in X}\left\{\left\langle \nabla f(\mathbf{x}^{(1)}), \mathbf{x}-\mathbf{x}^{(1)}\right\rangle + \frac{1}{A_1}D_\psi(\mathbf{x},\hat{\mathbf{x}}^{(0)})\right\} - \frac{1}{A_1}D_\psi(\mathbf{x}^*,\hat{\mathbf{x}}^{(0)}) \\
&= f(\mathbf{x}^{(1)}) + \left\langle \nabla f(\mathbf{x}^{(1)}), \nabla\psi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)}\right\rangle \\
&\quad + \frac{1}{A_1}D_\psi(\nabla\psi^*(\mathbf{z}^{(1)}),\hat{\mathbf{x}}^{(0)}) - \frac{1}{A_1}D_\psi(\mathbf{x}^*,\hat{\mathbf{x}}^{(0)}).
\end{aligned}$$

As $a_1 = A_1$, $\mathbf{x}^{(1)} = \nabla\psi^*(\hat{\mathbf{z}}^{(0)})$, and $a_1\nabla f(\hat{\mathbf{x}}^{(0)}) = \mathbf{z}^{(0)} - \hat{\mathbf{z}}^{(0)}$, using Proposition 17, we have that:

$$\begin{aligned}
&\left\langle \nabla f(\hat{\mathbf{x}}^{(0)}), \nabla\psi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)}\right\rangle \\
&\quad= \frac{1}{A_1}\left\langle \mathbf{z}^{(0)} - \hat{\mathbf{z}}^{(0)}, \nabla\psi^*(\mathbf{z}^{(1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(0)})\right\rangle \\
&\quad= \frac{1}{A_1}\left(D_{\psi^*}(\mathbf{z}^{(0)},\hat{\mathbf{z}}^{(0)}) - D_{\psi^*}(\mathbf{z}^{(0)},\mathbf{z}^{(1)}) + D_{\psi^*}(\hat{\mathbf{z}}^{(0)},\mathbf{z}^{(1)})\right). \qquad (17)
\end{aligned}$$

On the other hand, by smoothness of $f(\cdot)$ and the initial condition:

$$\begin{aligned}
&\left\langle \nabla f(\mathbf{x}^{(1)}) - \nabla f(\hat{\mathbf{x}}^{(0)}), \nabla\psi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)}\right\rangle \\
&\quad\geq -L\|\nabla\psi^*(\hat{\mathbf{z}}^{(0)}) - \hat{\mathbf{x}}^{(0)}\|\|\nabla\psi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)}\|. \qquad (18)
\end{aligned}$$

Finally, by Proposition 15 and the initial condition $\mathbf{z}^{(0)} = \nabla\psi(\hat{\mathbf{x}}^{(0)})$, we have that $D_{\psi^*}(\mathbf{z}^{(0)},\mathbf{z}^{(1)}) = D_\psi(\nabla\psi^*(\mathbf{z}^{(1)}),\hat{\mathbf{x}}^{(0)})$. Combining with (17), (18), and $G_1 = U_1 - L_1 = f(\mathbf{x}^{(1)}) - L_1$:

$$\begin{aligned}
G_1 \leq& L\|\nabla\psi^*(\hat{\mathbf{z}}^{(0)}) - \hat{\mathbf{x}}^{(0)}\|\cdot\|\nabla\psi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)}\| \\
&- \frac{1}{A_1}\left(D_{\psi^*}(\mathbf{z}^{(0)},\hat{\mathbf{z}}^{(0)}) + D_{\psi^*}(\hat{\mathbf{z}}^{(0)},\mathbf{z}^{(1)})\right) + \frac{1}{A_1}D_\psi(\mathbf{x}^*,\hat{\mathbf{x}}^{(0)}) \\
=& L\|\nabla\psi^*(\hat{\mathbf{z}}^{(0)}) - \hat{\mathbf{x}}^{(0)}\|\cdot\|\nabla\psi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)}\| \\
&- \frac{1}{A_1}\left(D_\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(0)}),\hat{\mathbf{x}}^{(0)}) + D_{\psi^*}(\hat{\mathbf{z}}^{(0)},\mathbf{z}^{(1)})\right) + \frac{1}{A_1}D_\psi(\mathbf{x}^*,\hat{\mathbf{x}}^{(0)}) \\
\leq& L\|\nabla\psi^*(\hat{\mathbf{z}}^{(0)}) - \hat{\mathbf{x}}^{(0)}\|\cdot\|\nabla\psi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)}\| \\
&- \frac{\sigma}{2A_1}\left(\|\nabla\psi^*(\hat{\mathbf{z}}^{(0)}) - \hat{\mathbf{x}}^{(0)}\|^2 + \|\nabla\psi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)}\|^2\right) + \frac{1}{A_1}D_\psi(\mathbf{x}^*,\hat{\mathbf{x}}^{(0)}) \\
\leq& \frac{1}{A_1}D_\psi(\mathbf{x}^*,\hat{\mathbf{x}}^{(0)}),
\end{aligned}$$

where we have used Proposition 15, $\mathbf{x}^{(1)} = \nabla\psi^*(\hat{\mathbf{z}}^{(0)})$, and $\frac{{a_1}^2}{A_1} = A_1 \leq \frac{\sigma}{L}$. ◀

▶ **Proposition 12** (restated).

$$\begin{aligned}
a_{i+1}&\left\langle \nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)}), \nabla\psi^*(\mathbf{z}^{(i+1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\right\rangle \\
&\qquad\leq \sigma^{-1}{a_{i+1}}^2\|\nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)})\|^2.
\end{aligned}$$

Proof. From the first order optimality condition in Fact 6, $\forall \mathbf{x},\mathbf{y}\in X$:

$$\left\langle \nabla\psi(\nabla\psi^*(\mathbf{z}^{(i+1)})) - \mathbf{z}^{(i+1)}, \mathbf{x} - \nabla\psi^*(\mathbf{z}^{(i+1)})\right\rangle \geq 0, \text{ and} \qquad (19)$$

$$\left\langle \nabla\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)})) - \hat{\mathbf{z}}^{(i)}, \mathbf{y} - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\right\rangle \geq 0. \qquad (20)$$

Letting $\mathbf{x} = \nabla\psi^*(\hat{\mathbf{z}}^{(i)})$, $\mathbf{y} = \nabla\psi^*(\mathbf{z}^{(i+1)})$, and summing (19) and (20):

$$\begin{aligned}
&\left\langle \hat{\mathbf{z}}^{(i)} - \mathbf{z}^{(i+1)}, \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)}) \right\rangle \\
&\qquad\qquad \geq \left\langle \nabla\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)})) - \nabla\psi(\nabla\psi^*(\mathbf{z}^{(i+1)})), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)}) \right\rangle \\
&\qquad\qquad \geq \sigma\|\nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})\|^2, \qquad (21)
\end{aligned}$$

where (21) follows by the σ-strong convexity of $\psi(\cdot)$. Using the Cauchy-Schwartz inequality and dividing both sides by $\|\nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})\|$ gives $\|\hat{\mathbf{z}}^{(i)} - \mathbf{z}^{(i+1)}\| \geq \sigma\|\nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\mathbf{z}^{(i+1)})\|$.

Since, by the step definition (2), $\hat{\mathbf{z}}^{(i)} - \mathbf{z}^{(i+1)} = a_{i+1}(\nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)}))$, applying Cauchy-Schwartz Inequality to $a_{i+1}\left\langle \nabla f(\mathbf{x}^{(i+1)}) - \nabla f(\hat{\mathbf{x}}^{(i)}), \nabla\psi^*(\mathbf{z}^{(i+1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) \right\rangle$ completes the proof. ◀