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

Jelena Diakonikolas and Lorenzo Orecchia 
Computer Science Department, Boston University 
email: \{jelenad, 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 [11] and Extra-Gradient Descent [7] 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 [8] using the classical method of implicit Euler discretization. Our analysis explicitly shows the effects of discretization through a conceptually novel primal-dual viewpoint. Finally, we present experiments showing that our algorithm matches the performance of Nesterov’s method, while appearing more robust to noise in some cases. 

1 Introduction 

In his seminal work [13,14], 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 $O(1/k^2)$, which has been shown to be asymptotically optimal [14] for smooth functions in this standard blackbox model [19]. More recently, Nesterov generalized this method to allow non-Euclidean norms in the definition of smoothness [16]. We refer to this generalization of Nesterov’s method and to instantiations thereof as Accelerated Gradient Descent (AGD) methods. Accelerated gradient methods have been widely extended and modified for different settings, including composite optimization [9,18], cubic regularization [17] and universal methods [20]. They have also found a number of fundamental applications in many algorithmic areas, including machine learning (see [2]) and discrete optimization [10]. However, to this day, Nesterov’s AGD methods remain the only paradigm 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 [22]. In this paper, 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 [11] and Extra-Gradient Descent [7]. 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 [8,23]. The continuous-time view is particularly helpful in clarifying the relation between AGD, AXGD and Mirror-Prox. Following Krichene’s presentation [8], given a gradient
field $\nabla f$ and a prox function $\psi$, it is possible to define two continuous-time evolutions: the mirror-descent dynamics and the accelerated-mirror-descent dynamics (see Section 2.1). 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 similar extra-gradient methods arise from the application of an approximate backward-Euler discretization [6] to the mirror-descent dynamics. Finally, our algorithm AXGD is the result of an approximate backward-Euler discretization of the accelerated mirror-descent dynamics.

While this paper focuses on the smooth case, the same techniques easily yield results matching those of AGD methods both for the strongly-convex-and-smooth case and the case of Hölder-continuous gradients. A more thorough description of the techniques together with their applications to a range of convex optimization problems will be included in the upcoming full version of this paper.

Another conceptual contribution of this 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 algorithms explicitly maintain a current primal solution $x(t)$ and a current dual solution $z(t)$, the latter in the form of a convex combination of gradients of the convex 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: $U_t \geq f(x^*) \geq L_t$. 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 [5].

Finally, we present an analysis of AXGD for the case of a noisy first-order oracle and a number of experiments showing that the performance of AXGD closely matches that of AGD methods while exhibiting better stability properties in some cases.

### 1.1 Related Work

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 $\Omega(1/\sqrt{t})$ [12], Nesterov showed that the saddle-point structure can be exploited by smoothing the original problem and applying AGD methods on the resulting smooth function [16]. This approach [15,16] yields an $O(1/k)$-convergence for convex-concave saddle point problems with smooth gradients. Surprisingly, at around the same time, Nemirovski [11] 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 [12]. 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 [22]. 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 [3]. At the same time, Krichene et al. [8] and Wibisono et al. [23] 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 [8].
1.2 Preliminaries

We will focus on continuous and differentiable functions defined on a convex set $X \subseteq \mathbb{R}^n$. The following definitions will be useful in our analysis, and thus we state them here for completeness.

**Definition 1.1.** A function $f : X \rightarrow \mathbb{R}$ is convex on $X$, if for all $x, \hat{x} \in X$: $f(\hat{x}) \geq f(x) + \langle \nabla f(x), \hat{x} - x \rangle$.

**Definition 1.2.** A function $f : X \rightarrow \mathbb{R}$ is smooth on $X$ with smoothness parameter $L$ and with respect to a norm $\| \cdot \|$, if for all $x, \hat{x} \in X$: $f(\hat{x}) \leq f(x) + \langle \nabla f(x), \hat{x} - x \rangle + \frac{L}{2} \| \hat{x} - x \|^2$.

**Definition 1.3.** A function $f : X \rightarrow \mathbb{R}$ is strongly convex on $X$ with strong convexity parameter $\sigma$ and with respect to a norm $\| \cdot \|$, if for all $x, \hat{x} \in X$: $f(\hat{x}) \geq f(x) + \langle \nabla f(x), \hat{x} - x \rangle + \frac{\sigma}{2} \| \hat{x} - x \|^2$.

**Definition 1.4.** (Bregman Divergence) $D_\psi(x, \hat{x}) \overset{\text{def}}{=} \psi(x) - \psi(\hat{x}) - \langle \nabla \psi(\hat{x}), x - \hat{x} \rangle$.

**Definition 1.5.** (Convex Conjugate) Function $\psi^*$ is the convex conjugate of $\psi : X \rightarrow \mathbb{R}$, if $\psi^*(z) = \max_{x \in X} \{ \langle z, x \rangle - \psi(x) \}$, $\forall g \in \mathbb{R}$.

In the rest of the paper, we will assume that $\psi(x)$ is continuously differentiable, so that Fenchel-Moreau Theorem implies that $\psi^{**} = \psi^1$. We are interested in minimizing a convex function $f$ over $X \subseteq \mathbb{R}^n$. We let $x^* = \arg \min_{x \in X} f(x)$.

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 $x \in X$ will be given as $\text{Grad}(x) = \arg \min_{\hat{x} \in X} \{ f(x) + \langle \nabla f(x), \hat{x} - x \rangle + \frac{L}{2} \| \hat{x} - x \|^2 \}$.

We will assume that there is a strongly-convex differentiable function $\psi : X \rightarrow \mathbb{R}$ such that $\max_{x \in X} \{ \langle z, x \rangle - \psi(x) \}$ is easily solvable, possibly in a closed form. Notice that this problem defines the convex conjugate of $\psi(., \cdot)$, i.e., $\psi^*(z) = \max_{x \in X} \{ \langle z, x \rangle - \psi(x) \}$. The following standard fact will be extremely useful in carrying out the analysis of the algorithms in this paper.

**Fact 1.6.** Let $\psi : X \rightarrow \mathbb{R}$ be a differentiable strongly-convex function. Then:

$$\nabla \psi^*(z) = \arg \max_{x \in X} \{ \langle z, x \rangle - \psi(x) \}. \tag{1.1}$$

**Properties of the Bregman Divergence.** The following properties of Bregman divergence will be useful in our analysis.

**Proposition 1.7.** $D_\psi(\nabla \psi^*(z), x) = D_{\psi^*}(\nabla \psi(x), z)$, $\forall x, z$.

**Proof.** From the definition of $\psi^*$ and Fact 1.6

$$\psi^*(z) = \langle \nabla \psi^*(z), z \rangle - \psi(\nabla \psi^*), \forall z. \tag{1.1}$$

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

$$\psi(x) = \langle \nabla \psi(x), x \rangle - \psi^*(\nabla \psi(x)), \forall x. \tag{1.2}$$

Using the definition of $D_\psi(\nabla \psi^*(z), x)$ and Fact 1.6,

$$D_\psi(\nabla \psi^*(z), x) = \psi(\nabla \psi^*(z)) - \psi(x) - \langle \nabla \psi(x), \nabla \psi^*(z) \rangle$$

$$= \psi(\nabla \psi^*(z)) + \psi^*(\nabla \psi(x)) - \langle \nabla \psi(x), \nabla \psi^*(z) \rangle. \tag{1.3}$$

\(^1\text{Note that Fenchel-Moreau Theorem requires } \psi \text{ to only be lower-semicontinuous for } \psi^{**} = \psi \text{ to hold, which is a weaker property than continuity or continuous differentiability.}\)
Similarly, using the definition of $D_{\psi^*}(\nabla \psi(x), z)$ combined with (1.1):

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

(1.4)

Comparing (1.3) and (1.4), the proof follows.

**Proposition 1.8.** If $\psi(.)$ is $\sigma$-strongly convex, then $D_{\psi^*}(z, \hat{z}) \geq \frac{\sigma}{2} \| \nabla \psi^*(z) - \nabla \psi^*(\hat{z}) \|^2$.

**Proof.** Using the definition $D_{\psi^*}(z, \hat{z})$ and (1.1), we can write $D_{\psi^*}(z, \hat{z})$ as:

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

Since $\psi(.)$ is $\sigma$-strongly convex, it follows that:

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

As, from Fact 1.6 $\nabla \psi^*(z) = \arg \max_{x \in X} \{ \langle x, z \rangle - \psi(x) \}$, by the first-order optimality condition

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

completing the proof.

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

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

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

**Proposition 1.9.** For all $x, y, z \in X$

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

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. Steps in each iteration of AGD come from explicit (forward) Euler discretization, followed by the gradient step that is used to correct the discretization error. The points $x^{(k)}$ are used in the construction of the lower bound, while the points $\hat{x}^{(k)} = \text{Grad}(x^{(k)})$ are used in the upper bound.

| Accelerated Gradient Descent (AGD) | Accelerated Extra-Gradient Descent (AXGD) |
|-----------------------------------|------------------------------------------|
| $x^{(k+1)} = \frac{A_k}{A_{k+1}} \hat{x}^{(k)} + \frac{a_{k+1}}{A_{k+1}} \nabla \psi^*(z^{(k)})$, | $\hat{x}^{(k)} = \frac{A_k}{A_{k+1}} x^{(k)} + \frac{a_{k+1}}{A_{k+1}} \nabla \psi^*(z^{(k)})$, |
| $z^{(k+1)} = z^{(k)} - a_{k+1} \nabla f(x^{(k+1)})$, | $\hat{z}^{(k)} = z^{(k)} - a_{k+1} \nabla f(\hat{x}^{(k)})$, |
| $\hat{x}^{(k+1)} = \text{Grad}(x^{(k+1)})$. | $x^{(k+1)} = \frac{A_k}{A_{k+1}} x^{(k)} + \frac{a_{k+1}}{A_{k+1}} \nabla \psi^*(\hat{z}^{(k)})$, |
| $z^{(k+1)} = z^{(k)} - a_{k+1} \nabla f(x^{(k+1)})$. | $z^{(k+1)} = z^{(k)} - a_{k+1} \nabla f(x^{(k+1)})$. |

The iterations of AXGD come from an approximate implementation of implicit (backward) Euler discretization. The idea is similar to Nemirovski’s mirror prox algorithm [11], with the main difference
We now discuss the continuous-time analogues of first-order methods, upon which our new algorithm is based. While the analogy between discrete optimization algorithms and continuous-time dynamics was known from the outset [12], recent works by Krichene et al. [8] and Wibisono et al. [23] have shed more light on this connection, reaffirming the importance of the choice of discretization, which was also studied by Scieur et al. [21].

Our description of the dynamics closely follows that of Krichene et al. [8]. We use “dot” notation to denote the time derivative. That is, \( \dot{x} = \frac{dx}{dt} \). We allow for a slightly more general scaling of time, by introducing a continuously-differentiable monotonically increasing function \( \alpha(t) \) be of time \( t \) with \( \alpha(1) = 1 \). Mirror Descent then has following continuous-time analogue [12]:

\[
\begin{align*}
\dot{z}(t) &= -\dot{\alpha}(t) \cdot \nabla f(\psi^*(z(t))), \\
\dot{x}(t) &= \dot{\alpha}(t) \cdot \frac{\nabla^2 \psi^*(z(t))-x(t)}{\alpha(t)}. 
\end{align*}
\] (2.3)

Recently, Krichene et al. [8] introduced a different dynamics, the accelerated-mirror-descent (AMD) dynamics, as a continuous-time analogue of accelerated first-order methods. The continuous-time AMD dynamics can be described as:

\[
\begin{align*}
\dot{z}(t) &= -\dot{\alpha}(t) \cdot \nabla f(x(t)), \\
\dot{x}(t) &= \dot{\alpha}(t) \cdot \frac{\nabla^2 \psi^*(x(t))-x(t)}{\alpha(t)}. 
\end{align*}
\] (2.4)

The following proposition gives an integral expression for \( z(t) \) and \( x(t) \) in the AMD dynamics. Its simple proof appears in the appendix. To avoid singularities at the initial point, we will start the evolution of (2.4) at \( t_0 = 1 \) from an arbitrary point \( x(1) \in X \) and \( z(1) = \nabla \psi(x(1)) \). In this proposition and for the rest of the paper, we will make use of the Lesbegue-Stieljes notation for the integral with respect to the density \( \dot{\alpha}(t) \).

**Proposition 2.1.** Let \( x(t) \) and \( z(t) \) evolve according to (2.4), starting at \( t_0 = 1 \) from an arbitrary point \( x(1) \in X \) and \( z(1) = \nabla \psi(x(1)) \). Then, \( \forall t \geq 1 \):

1. \( z(t) = z(1) - \int_1^t \dot{\alpha}(\tau) \cdot \nabla f(x(\tau))d\tau = z(1) - \int_1^t \nabla f(x)d\alpha \)

2. \( x(t) = \frac{x(1) + \int_1^t \dot{\alpha}(\tau) \nabla^2 \psi^*(x(\tau))d\tau}{\alpha(t)} = \frac{x(1) + \int_1^t \nabla^2 \psi^*(x)d\alpha}{\alpha(t)} \) and \( x(t) \in X \).

At this point, we briefly remark on the difference between the mirror descent dynamic and its accelerated counterpart, which lies in the form of \( \dot{z}(t) \). In the basic mirror-descent dynamic, this derivative equals the negative gradient of the function at the point \( \nabla \psi^*(z(t)) \), which can be thought of as the primal image of the current dual solution \( z(t) \) under the mirror map \( \nabla \psi^* \). In contrast, for the accelerated
dynamics, the gradient is actually taken at $x^{(t)}$, which, by the previous proposition, is the average of the mirror maps of the dual solutions observed so far. Ultimately it is this additional averaging that enables acceleration.

We now describe an analysis of the convergence of $f(x^{(t)})$ to $f(x^*)$ in the AMD dynamics. Our argument differs from that of Krichene et al. \cite{Krichene2015} by taking a primal-dual viewpoint. This analysis will provide a blueprint for the convergence arguments of all accelerated algorithms in this paper.

**Constructing Upper and Lower Bounds to the Optimum.** We interpret the primal-dual pair $(x^{(t)}, z^{(t)})$ maintained by the evolution of (2.4) as constructing both an upper and a lower bound to the optimal value $f(x^*)$. The upper bound $U_t$ is naturally given by the function value at time $t$, i.e., $U_t \overset{\text{def}}{=} f(x^{(t)})$. In continuous time, this has derivative

$$
\dot{U}_t = \left\langle \nabla f(x^{(t)}), \dot{x}^{(t)} \right\rangle. 
$$

(2.5)

The construction of the lower bound is more interesting. At each time $t$, the convexity of the objective $f(.)$ yields the lower-bounding hyperplane: $\forall x \in X, f(x) \geq f(x^{(t)}) + \left\langle \nabla f(x^{(t)}), x - x^{(t)} \right\rangle$. A natural choice of lower bound to the optimum at time $t \geq 1$, is given by averaging such hyperplanes over $[1, t]$ according to the measure $\alpha$ and by minimizing over $x \in X$

$$
f(x^*) \geq \frac{\int_1^t f(x) d\alpha}{\alpha(t) - 1} + \min_{u \in X} \left\{ \int_1^t \left( \langle \nabla f(x), u - x \rangle d\alpha \right) \right\}.
$$

While we could use 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, it is not defined for our initial time $t_0 = 1$, meaning that we do not have a natural concept of initial lower bound and initial duality gap. We overcome the first obstacle by applying Moreau-Yosida regularization, i.e., by adding to the minimization a regularizer term that is strongly-convex in $x$. Without loss of generality, this term can be taken to be the Bregman divergence of a $\sigma$-strongly-convex function $\psi$ taken from an input point $x^{(1)}$. This yields:

$$
f(x^*) + \frac{D_\psi(x^*, x^{(1)})}{\alpha(t) - 1} \geq \frac{\int_1^t f(x) d\alpha}{\alpha(t) - 1} + \min_{u \in X} \left\{ \int_1^t \left( \langle \nabla f(x), u - x \rangle + D_\psi(u, x^{(1)}) \rangle d\alpha \right) \right\}.
$$

To address the second problem, we mix into the $\alpha$-combination of hyperplanes the optimal lower bound $f(x^*)$ with weight 1. Rescaling the normalization factor, we obtain our notion of regularized lower bound:

$$
L_t \overset{\text{def}}{=} \frac{\int_1^t f(x) d\alpha}{\alpha(t)} + \min_{u \in X} \left\{ \int_1^t \left( \langle \nabla f(x), u - x \rangle + D_\psi(u, x^{(1)}) \rangle d\alpha \right) \right\} + \left( f(x^*) - D_\psi(x^*, x^{(1)}) \right) \frac{1}{\alpha(t)}.
$$

With this definition, a simple computation yields the derivative of the lower bound $L_t$.

**Proposition 2.2.** Recalling that $G_t = U_t - L_t$, we have:

$$
\dot{L}_t = \frac{\dot{\alpha}(t)}{\alpha(t)} \left( \langle \nabla f(x^{(t)}), \nabla \psi^*(z^{(t)}) - x^{(t)} \rangle + G_t \right).
$$

\footnotetext{This construction does not allow the algorithm to explicitly maintain $U_t$ and $L_t$, as the algorithm has no initial knowledge of $f(x^*)$. In specific cases, such as for smooth objectives, the functions $U_t$ and $L_t$ can be slightly modified to be explicitly computable by the algorithm \cite{Krichene2015}.}
We now show how to discretize the continuous time algorithms from the previous subsection, while using the AMD dynamics from Equation (2.4). The continuous-time dynamics described by Eq. (2.4) and the initial condition\( \mathbf{z}^{(1)} = \nabla \psi(\mathbf{x}^{(1)}) \) converges as follows: for any \( t > 1 \),

\[
    f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*) + D_\psi(\mathbf{x}^*, \mathbf{x}^{(1)})}{\alpha^{(t)}}.
\]

**Proof.** Consider the evolution of the duality gap \( G_t = U_t - L_t \). By Equation (2.5) and Proposition 2.2, we have:

\[
    \dot{G}_t = \dot{U}_t - \dot{L}_t = \left\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{x}^{(t)} - \mathbf{x}^* \right\rangle \cdot \frac{\nabla \psi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}} - \frac{\dot{\alpha}^{(t)} G_t}{\alpha^{(t)}}.
\]

By the definition of the dynamics, this yields \( \dot{G}_t = -\frac{\dot{\alpha}^{(t)} G_t}{\alpha^{(t)}} \). Because this is equivalent to \( \frac{d}{dt} \left( \alpha^{(t)} G_t \right) = 0 \), we obtain that \( G_t = G_1/\alpha^{(t)} \) for \( t \geq 1 \).\( \square \)

### 2.2 Discretization

We now show how to discretize the continuous time algorithms from the previous subsection, while using similar arguments in the convergence analysis. We will consider unit-length steps (i.e., the algorithm will be making updates at times \( t_0, t_0 + 1, t_0 + 2, \ldots \)) and replace integrals from the convergence analysis with Riemann sums. The measure \( \alpha \) will be replaced in the Riemann sums by non-negative coefficients \( a_i \), with \( A_k = \sum_{i=1}^{k} a_i \). As in the continuous-time case, the growth of \( A_k \) as a function of \( k \) will determine the convergence rate. From this point on, to distinguish between the continuous-time and discrete-time cases, we will use \( t, \tau \in \mathbb{R}_+ \) to denote time in the continuous-time domain, and \( k, i \in \mathbb{Z}_+ \) to denote discrete time points in the discrete-time domain. In particular, we will construct a sequence of primal-dual pairs \( (\mathbf{x}^{(k)}, \mathbf{z}^{(k)}) \) that act as a discretization of the continuous curves \( (\mathbf{x}^{(t)}, \mathbf{z}^{(t)}) \). Following the argument in the previous section, we can define a notion of upper bound \( U_k \) and lower bound \( L_k \) in the discrete case. We have \( U_k \overset{\text{def}}{=} f(\mathbf{x}^{(k)}) \) and:

\[
    L_k = \frac{\sum_{i=1}^{k} a_i f(\mathbf{x}^{(i)}) + \min_{\mathbf{u} \in \mathcal{X}} \left\{ \sum_{i=1}^{k} a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{u} - \mathbf{x}^{(i)} \right\rangle + D_\psi(\mathbf{u}, \mathbf{x}^{(1)}) \right\} - D_\psi(\mathbf{x}^*, \mathbf{x}^{(1)})}{A_k}.
\]

The following theorem provides an expression for the discretization error accrued by the duality gap \( G_k \overset{\text{def}}{=} U_k - L_k \). Its proof uses the same arguments as in the continuous case and is presented in the appendix for completeness.
In particular, we have

While this discretization is not computationally feasible in practice, as it requires solving for the implicitly defined next iterate \( x^{(k+1)} \), in the case of the \( \text{AMD} \) dynamics, implicit Euler discretization yields the following algorithm: let \( x^{(1)} \in X \) be an arbitrary initial point that satisfies \( x^{(1)} = \nabla \psi^*(z^{(1)}) \), where \( z^{(1)} = \nabla \psi(x^{(1)}) - \nabla f(x^{(1)}) \); for all \( k \geq 1 \)

\[
\begin{align*}
  x^{(k+1)} &= \frac{A_k}{A_{k+1}} x^{(k)} + a_{k+1} \nabla \psi^*(z^{(k)}) - a_{k+1} \nabla f(x^{(k)}) \\
  z^{(k+1)} &= z^{(k)} - a_{k+1} \nabla f(x^{(k+1)})
\end{align*}
\] 

(2.7)

While this discretization is not computationally feasible in practice, as it requires solving for the implicitly defined \( x^{(k+1)} \), it also boasts a negative discretization error, i.e., it converges faster than the continuous-time \( \text{AMD} \), as shown in the next theorem. Ultimately, we will use this extra slack to trade-off the error arising from an approximate implicit discretization.

**Theorem 2.5.** For the evolution of Equation (2.7), we have

\[
G_k \leq G_1 - \sum_{i=2}^{k} D_{\psi^*}(z^{(k)}, z^{(k+1)}) / A_k
\]

(2.8)

**Proof.** We bound the discretization error \( E_{k+1} \) in Theorem 2.4 for all \( k \geq 1 \). By convexity, \( f(x^{(k+1)}) - f(x^{(k)}) = \langle \nabla f(x^{(k+1)}), x^{(k+1)} - x^{(k)} \rangle \), and therefore:

\[
E_{k+1} \leq \left\langle \nabla f(x^{(k+1)}), A_{k+1} x^{(k+1)} - A_{k} x^{(k)} - a_{k+1} \nabla \psi^*(z^{(k+1)}) \right\rangle - D_{\psi^*}(z^{(k)}, z^{(k+1)})
\]

(from (2.7)) \( = - D_{\psi^*}(z^{(k)}, z^{(k+1)}) \).

The result follows by Theorem 2.4. \( \square \)

### 2.3 Convergence of AXGD

A standard way to implement implicit Euler discretization in the solution of ODEs 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:

\[
x^{(k+1)} = A_k / A_{k+1} x^{(k)} + a_{k+1} / A_{k+1} \nabla \psi^*(z^{(k)}) - a_{k+1} \nabla f(x^{(k+1)}).
\]

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

\[
\begin{align*}
\hat{x}^{(k)} &= A_k / A_{k+1} x^{(k)} + a_{k+1} / A_{k+1} \nabla \psi^*(z^{(k)}) \\
\hat{x}^{(k+1)} &= A_k / A_{k+1} x^{(k)} + a_{k+1} / A_{k+1} \nabla \psi^*(z^{(k)}) - a_{k+1} \nabla f(\hat{x}^{(k)})
\end{align*}
\]
We can now analyze AXGD as producing an approximate solution to the implicit Euler discretization problem. The following theorem gives a general bound on the convergence of AXGD for a convex and differentiable \( f(.) \) without additional assumptions. The only (mild) difference is replacing \( D_\psi(x, x^{(1)}) \) and \( D_\psi(x^*, x^{(1)}) \) by \( D_\psi(x, \tilde{x}^{(0)}) \) and \( D_\psi(x^*, \tilde{x}^{(0)}) \), since we start from the “intermediate” point \( \tilde{x}^{(0)} \). This change is only important for bounding the initial gap \( G_1 \); everything else remains the same.

**Theorem 2.6.** Consider the AXGD algorithm as described in Equation (2.2), starting from an arbitrary point \( \tilde{x}^{(0)} \) with \( z^{(0)} = \nabla \psi(\tilde{x}^{(0)}) \) and \( A_0 = 0 \). Then the error from Theorem 2.4 is bounded by:

\[
E_{k+1} \leq \left\langle \nabla f(x^{(k+1)}), A_{k+1}x^{(k+1)} - A_kx^{(k)} - a_{k+1}\nabla \psi^*(z^{(k+1)}) \right\rangle - D_\psi(z^{(k)}, z^{(k+1)})
\]

**Proof.** As in the proof of Theorem 2.5, we use convexity to bound the error given by Theorem 2.4 followed by the definition of AXGD:

\[
E_{k+1} \leq \left\langle \nabla f(x^{(k+1)}), A_{k+1}x^{(k+1)} - A_kx^{(k)} - a_{k+1}\nabla \psi^*(z^{(k+1)}) \right\rangle - D_\psi(z^{(k)}, z^{(k+1)})
\]

We now use the fact that \( a_{k+1}f(\tilde{x}^{(k)}) = z^{(k)} - \tilde{z}^{(k)} \) together with the standard triangle-inequality for Bregman divergences (see Proposition 1.9) to show that:

\[
E_{k+1} \leq \left\langle \nabla f(\tilde{x}^{(k)}), \nabla \psi^*(\tilde{z}^{(k)}) - \nabla \psi^*(z^{(k+1)}) \right\rangle - D_\psi(z^{(k)}, z^{(k+1)})
\]

Combining the results of the two previous equations, we have the final bound on the error:

\[
E_{k+1} \leq \left\langle \nabla f(x^{(k+1)}), \nabla \psi^*(\tilde{z}^{(k)}) - \nabla \psi^*(z^{(k+1)}) \right\rangle - D_\psi(\tilde{z}^{(k)}, z^{(k+1)}) - D_\psi(z^{(k)}, \tilde{z}^{(k)}).
\]

\[\square\]

### 2.4 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 \( f(.) \) by applying Theorem 2.6. The crux of the proof is that we can take sufficiently large steps while maintaining the error from Theorem 2.6 non-positive. In other words, we are able to move quickly through the continuous evolution of AMD by taking large discrete steps.

**Theorem 2.7.** Let \( f : X \to \mathbb{R} \) be an \( L \)-smooth convex function and let \( x^{(k)}, z^{(k)}, \tilde{x}^{(k)}, \tilde{z}^{(k)} \) be updated according to the AXGD algorithm in Equation (2.2), starting from an arbitrary initial point \( \tilde{x}^{(0)} \) with the following initial conditions: \( z^{(0)} = \nabla \psi(\tilde{x}^{(0)}) \) and \( A_0 = 0 \). Let \( \psi : X \to \mathbb{R} \) be \( \sigma \)-strongly convex. If the steps \( a_k \) satisfy the inequality \( a_k^2 \leq \frac{\sigma^2}{L} \), then for all \( k \geq 1 \),

\[
f(x^{(k)}) - f(x^*) \leq \frac{D_\psi(x^*, \tilde{x}^{(0)})}{A_k}.
\]

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

\[
f(x^{(k)}) - f(x^*) \leq \frac{L}{(k+1)^2} \|x^* - \tilde{x}^{(0)}\|^2.
\]
We now assume that we do not have the exact information about the gradients at queried points. Instead, with an inexact gradient, we will assume that the gradients are corrupted by additive noise. This model is fundamentally different than the model considered in e.g., \[4\]. In \[4\], it is assumed that a function \(f\) is associated with a \((\delta, L)\) oracle, such that

\[
\langle \nabla f(\mathbf{x}), \hat{x} - x \rangle + \frac{L}{2} \| \hat{x} - x \|^2 \geq \Delta_k, \forall x, \hat{x} \in X
\]

(as opposed to \(f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{x} - x \rangle + \frac{L}{2} \| \hat{x} - x \|^2\) that we normally get from \(f\)’s smoothness). Such a model seems more suitable for incorrectly specified functions (e.g., non-smooth functions treated as being smooth) rather than for inexact gradients, which is our focus.

### 3.1 Constructing a Lower Bound from Noise-Corrupted Gradients

Suppose that the sequence of queried points is \(x^{(1)}, x^{(2)}, \ldots, x^{(k)}, \ldots\), where each point \(x^{(k)}\) is associated with an inexact gradient \(\tilde{\nabla} f(x^{(k)}) = \nabla f(x^{(k)}) + \eta_k\), where \(\nabla f(x^{(k)})\) is the true gradient and \(\eta_k\) is the additive noise. Using convexity of \(f(.)\), \(\forall x \in X\) and any positive \(a_1, \ldots, a_k\), where \(A_k = \sum_{i=1}^{k} a_i\):

\[
f(x) \geq \frac{1}{A_k} \sum_{i=1}^{k} a_i f(x^{(i)}) + \frac{1}{A_k} \sum_{i=1}^{k} a_i \langle \tilde{\nabla} f(x^{(i)}), x - x^{(i)} \rangle - \frac{1}{A_k} \sum_{i=1}^{k} a_i \langle \eta_i, x - x^{(i)} \rangle.
\]

Therefore, using the same construction of the lower bound as in the case of exact gradients, we have:

\[
f(x^*) + \frac{1}{A_k} \sum_{i=1}^{k} a_i \langle \eta_i, x^* - x^{(i)} \rangle \geq \Lambda_k,
\]

where

\[
\Lambda_k = \frac{\sum_{i=1}^{k} a_i f(x^{(i)}) + \min_{x \in X} \{ \sum_{i=1}^{k} a_i \langle \tilde{\nabla} f(x^{(i)}), x - x^{(i)} \rangle + D_\psi(x, x^{(i)}) \} - D_\psi(x^*, x^{(1)})}{A_k}.
\]
We can observe that $\Lambda_k$ is equivalent to $L_k$ with the only difference of the exact gradients being replaced by the noisy gradients. In particular, if for some suitably chosen upper-bound $U_k$, $U_k - \Lambda_k \leq \epsilon$, then the duality gap is bounded by $\epsilon + \frac{1}{A_k} \sum_{i=1}^{k} a_i \langle \eta_i, x^* - x^{(i)} \rangle$. This is good, because when we use noisy gradients, the lower bound is only corrupted by the average of the noise-distance to the optimum inner-product. Moreover, since $\forall i$, $x^{(i)}$ is independent of the gradient at time $i$ (and, therefore, independent of $\eta_i$), when $\eta$’s elements have mean zero, $\frac{1}{A_k} \sum_{i=1}^{k} a_i \langle \eta_i, x^* - x^{(i)} \rangle$ also has mean equal to zero.

### 3.2 Upper Bound and the AGD Gap

Suppose that AGD is run using noisy gradient (the algorithm may be unaware of the noise). Observe that in this case $z^{(k+1)} = z^{(k)} - a_{k+1} \nabla f(x^{(k+1)})$. As before, let the upper bound be $U_k = f(\hat{x}^{(k)})$, where

$$\hat{x}^{(k)} = \text{Grad}(x^{(k)}) = \arg\min_{x \in X} \left\{ \left\langle \nabla f(x^{(k)}), x - x^{(k)} \right\rangle + \Delta f(x, x^{(k)}) \right\}.$$

Using that, by convexity of $f(.)$, $f(\hat{x}^{(k)}) - f(x^{(k)}) \geq \left\langle \nabla f(x^{(k)}), \hat{x}^{(k)} - x^{(k)} \right\rangle - \langle \eta_{k+1}, \hat{x}^{(k)}, x^{(k+1)} \rangle$ and repeating the same arguments as for the exact gradients (Appendix B, with the same assumptions as in Theorem B.1), it follows that:

$$\Lambda_{k+1} \geq \frac{A_k}{A_{k+1}} (\Lambda_k - U_k) + \frac{A_k}{A_{k+1}} \left\langle \eta_{k+1}, \hat{x}^{(k)} - x^{(k)} \right\rangle + \min_{y \in X} \left\{ \left\langle \nabla f(x^{(k+1)}), y - x^{(k+1)} \right\rangle + \Delta f(y, x^{(k+1)}) \right\}. \tag{3.1}$$

Let $M_{k+1} = \min_{y \in X} \left\{ \left\langle \nabla f(x^{(k+1)}), y - x^{(k+1)} \right\rangle + \Delta f(y, x^{(k+1)}) \right\}$. From the definition of $\hat{x}^{(k+1)}$:

$$M_{k+1} = \left\langle \nabla f(x^{(k+1)}), \hat{x}^{(k+1)} - x^{(k+1)} \right\rangle + \Delta f(\hat{x}^{(k+1)}, x^{(k+1)}) \tag{3.2}$$

$$\geq \left\langle \eta_{k+1}, \hat{x}^{(k+1)} - x^{(k+1)} \right\rangle + f(\hat{x}^{(k+1)}) - f(x^{(k+1)}),$$

where the inequality comes from $\nabla f(x^{(k+1)}) = \nabla f(x^{(k+1)}) + \eta_{k+1}$ and the assumption that $\forall x, \hat{x}$: $f(\hat{x}) \leq f(x) + \langle \nabla f(x), \hat{x} - x \rangle + \Delta f(x, \hat{x})$. Combining with (3.1):

$$U_{k+1} - \Lambda_{k+1} \leq \frac{A_k}{A_{k+1}} (U_k - \Lambda_k) + \frac{A_k}{A_{k+1}} \left\langle \eta_{k+1}, \hat{x}^{(k)} - x^{(k+1)} \right\rangle - \left\langle \eta_{k+1}, \hat{x}^{(k+1)} - x^{(k+1)} \right\rangle$$

$$= \frac{A_k}{A_{k+1}} (U_k - \Lambda_k) + \left\langle \eta_{k+1}, \frac{A_k}{A_{k+1}} \hat{x}^{(k)} + \frac{a_{k+1}}{A_{k+1}} x^{(k+1)} - \hat{x}^{(k+1)} \right\rangle. \tag{3.3}$$

The bound (3.3) we get for the decrease in the duality gap is not encouraging for our techniques: each step contributes to additional noise, which even if it had a low mean, could still have a very high variance. Moreover, $\hat{x}^{(k+1)}$ is not independent of $\eta_{k+1}$, since $\hat{x}^{(k+1)}$ depends on the noisy gradient $\nabla f(x^{(k+1)})$.

Note that the noise-accumulating component does not seem avoidable, since it comes from the inexact gradient within the minimization problem in $M_{k+1}$ when we take a gradient step from $x^{(k+1)}$. The only inequalities we have used in deriving the gap come from the convexity and second-order approximation of $f(.)$, which are deterministic properties. Thus, it seems unlikely that the noise component from (3.3) can be removed without any changes to the algorithm (or techniques).

### 3.3 Upper Bound and the AXGD Gap

Consider now AXGD (Eq. (2.2)) with noisy gradients. Let $U_k = f(\hat{x}^{(k)})$ be the upper bound, same as in the case of exact gradients. Using the same arguments as in Section 2.3

$$\Lambda_{k+1} \geq \frac{A_k}{A_{k+1}} (U_k - \Lambda_k) + f(x^{(k+1)}) - \frac{E_{k+1}}{A_{k+1}},$$

11
We illustrate the performance of \( E_{k+1} = A_k (f(x^{(k+1)}) - f(x^{(k)})) - a_{k+1} \left( \nabla f(x^{(k+1)}), \nabla \psi^* (z^{(k+1)}) - x^{(k+1)} \right) - D_{\psi^*} (z^{(k)}, z^{(k+1)}) \).

By convexity of \( f(.) \) and the definition of \( \nabla f(x^{(k+1)}) \):

\[
E_{k+1} \leq A_k \left( \eta_{k+1}, x^{(k)} - x^{(k+1)} \right) - \left( \nabla f(x^{(k+1)}), a_{k+1} \nabla \psi^* (z^{(k+1)}) + A_k x^{(k)} - A_{k+1} x^{(k+1)} \right) - D_{\psi^*} (z^{(k)}, z^{(k+1)})
\]

\[
= A_k \left( \eta_{k+1}, x^{(k)} - x^{(k+1)} \right) - D_{\psi^*} (z^{(k)}, z^{(k+1)})
\]

\[
- a_{k+1} \left( \nabla f(x^{(k+1)}), \nabla \psi^* (z^{(k+1)}) - \nabla \psi^* (\hat{z}^{(k)}) \right) .
\]

Observing that \( a_{k+1} \nabla f(\hat{z}^{(k)}) = z^{(k)} - \hat{z}^{(k)} \), and using Prop. 1.9, we obtain a similar bound as in the case of exact gradients:

\[
E_{k+1} \leq A_k \left( \eta_{k+1}, x^{(k)} - x^{(k+1)} \right) - D_{\psi^*} (z^{(k)}, \hat{z}^{(k)}) - D_{\psi^*} (\hat{z}^{(k)}, z^{(k+1)})
\]

In particular, if we assume that the discretization error

\[
E_{k+1} = a_{k+1} \left( \nabla f(x^{(k+1)}), \nabla \psi^* (z^{(k)}) - \nabla \psi^* (z^{(k+1)}) \right) - D_{\psi^*} (z^{(k)}, \hat{z}^{(k)}) - D_{\psi^*} (\hat{z}^{(k)}, z^{(k+1)})
\]

is non-positive (see Theorems 2.6 and 2.7), we have:

\[
E_{k+1} \leq A_k \left( \eta_{k+1}, x^{(k)} - x^{(k+1)} \right) - a_{k+1} \left( \eta_{k+1} - \hat{\eta}_k, \nabla \psi^* (z^{(k+1)}) - \nabla \psi^* (\hat{z}^{(k)}) \right)
\]

\[
= a_{k+1} \left( \eta_{k+1}, x^{(k+1)} - \nabla \psi^* (z^{(k+1)}) \right) + a_{k+1} \left( \hat{\eta}_k, \nabla \psi^* (z^{(k+1)}) - \nabla \psi^* (\hat{z}^{(k)}) \right)
\]

Therefore, we have the following bound on the duality gap:

\[
f(x^{(k)}) - f(x^*) \leq \frac{D_{\psi^*}(x^*, x^{(0)}) + \sum_{i=1}^{k} a_i \left( \langle \eta_i, x^* - \nabla \psi^* (z^{(i)}) \rangle + \langle \hat{\eta}_{i-1}, \nabla \psi^* (z^{(i)}) - \nabla \psi^* (\hat{z}^{(i-1)}) \rangle \right)}{A_k}.
\]

Suppose that \( D = \max_{x, \hat{x} \in X} ||x - \hat{x}|| \) is bounded. If the elements of \( \eta_k \)'s are chosen adversarially from the interval \([-\epsilon, \epsilon]\), then the error due to noise on the right-hand side of (3.4) is a random variable with mean at most \( 2\epsilon D \) and standard deviation at most \( 2\epsilon D \).

Now suppose that \( \eta_k \)'s are vectors whose elements are independent Gaussians with zero mean and standard deviation \( \epsilon \). Unfortunately, because \( \nabla \psi^* (z^{(i)}) \) depends on both \( \nabla f(\hat{x}^{(i)}) \) and \( \nabla f(x^{(i)}) \), it is correlated with both \( \eta_i \) and \( \hat{\eta}_i \), and thus we cannot obtain zero mean and \( O(\epsilon D) \) standard deviation on the right-hand side. However, using Cauchy-Schwartz Inequality, we can bound the mean and the standard deviation of the error on the right-hand side of (3.4) with \( 2\epsilon \sqrt{n} D \), which means that in the case of AXGD when the diameter of the region is bounded, the noise cannot accumulate (i.e., it remains bounded as a function of \( \epsilon, n \), and \( D \)).

## 4 Experiments

We now illustrate the performance of AGD and AXGD for (i) an unconstrained problem over \( \mathbb{R}^n \) with the objective function \( f(x) = \frac{1}{2} (A x, x) - (b, x) \), and (ii) for the problem with the same objective and unit simplex as the feasible region, where \( A \) is the Laplacian of a cycle graph\(^3\) and \( b \) is a vector whose
Figure 1: (a)(c) Exact and (b)(d) approximate duality gaps for AGD and AXGD with exact gradients.

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

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

(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

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., [19]). 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 $\ell_2$ 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 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 $\epsilon \eta I$. When the region is
unconstrained (top row in Fig. 2), the algorithms may be taking large steps, which leads to a high error (see Section 3 for more details). GD overall exhibits higher tolerance to noise, since it takes steps based only on the last point, and thus it cannot accumulate noise from previous steps.

In the case of the unit simplex region (bottom row in Fig. 2), the diameter is one and thus the norm of the step size taken by any of the algorithms is bounded by one, which makes all the algorithms more tolerant to noise than in the unconstrained case. Since the points come from the unit simplex, the steps correspond to the vector rotation. In this case, averaging over steps helps, since it, intuitively, reduces the chance of high vector rotation due to noise. This explains the more stable behavior of AXGD compared to AGD and GD, which both take greedy gradient steps in each iteration.

5 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) non-smooth objectives with Hölder-continuous gradients, (iii) and non-smooth Lipschitz-continuous objectives [5]. In the experiments from Section 4 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 particular interest, since the gradients can often come from noise-corrupted measurements. Further, our experiments from Fig. 2 suggest that there is 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 accelerated methods in practice.

References

[1] Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent, 2014. arXiv preprint, arXiv:1407.1537.

[2] S. Bubeck. Theory of Convex Optimization for Machine Learning. 2014. arXiv preprint, arXiv:1405.4980v1.

[3] S. Bubeck, Y. T. Lee, and M. Singh. A geometric alternative to Nesterov’s accelerated gradient descent. 2015.

[4] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. Math. Prog., 146(1-2):37–75, 2014.

[5] J. Diakonikolas and L. Orecchia. The approximate gap technique: A unified approach to optimal first-order methods, 2017. Manuscript.

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

[7] 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.

[8] W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In Proc. NIPS’15, 2015.

[9] G. Lan. An optimal method for stochastic composite optimization. Math. Prog., 133(1-2):365–397, Jan. 2011.

[10] Y. T. Lee, S. Rao, and N. Srivastava. A new approach to computing maximum flows using electrical flows. In Proc. ACM STOC ’13, 2013.
[11] A. 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.

[12] A. Nemirovskii and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[13] Y. 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.

[14] Y. Nesterov. *Introductory Lectures on Convex Programming Volume: A Basic course*, volume I. Kluwer Academic Publishers, 2004.

[15] Y. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optimiz.*, 16(1):235–249, Jan. 2005.

[16] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Prog.*, 103(1):127–152, Dec. 2005.

[17] Y. Nesterov. Accelerating the cubic regularization of newton’s method on convex problems. *Math. Prog.*, 112(1):159–181, 2008.

[18] Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Prog.*, 140(1):125–161, 2013.

[19] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

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

[21] D. Scieur, V. Roulet, F. Bach, and A. D’Aspremont. Integration methods and accelerated optimization algorithms, Feb. 2017. arXiv preprint, arXiv:1702.06751.

[22] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008.

[23] A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. In *Proc. Natl. Acad. Sci. U.S.A.*, 2016.
A Proofs for Section 2 (Accelerated Extra-Gradient Descent)

Proposition 2.1 Let \( x^{(t)} \) and \( z^{(t)} \) evolve according to (2.4), starting at \( t_0 = 1 \) from an arbitrary point \( x^{(1)} \in X \) and \( z^{(1)} = \nabla \psi(x^{(1)}) \). Then, \( \forall t \geq 1 \):

1. \( z^{(t)} = z^{(1)} - \int_1^t \dot{\alpha}(\tau) \cdot \nabla f(x(\tau)) d\tau = z^{(1)} - \int_1^t \nabla f(x) d\alpha \)

2. \( x^{(t)} = \frac{x^{(1)} + \int_1^t \dot{\alpha}(\tau) \nabla \psi^*(z(\tau)) d\tau}{\alpha(t)} = \frac{x^{(1)} + \int_1^t \nabla \psi^*(z) d\alpha}{\alpha(t)} \) and \( x^{(t)} \in X \).

Proof. The first part follows by integrating both sides of the first equation in (2.4) from 1 to \( t \). For the second part, we can write the second equation in (2.4) equivalently as \( \alpha(t)x^{(t)} + \dot{\alpha}(t)x^{(t)} = \alpha(t)\nabla \psi^*(z^{(t)}) \).

Observing that \( \alpha(t)x^{(t)} + \dot{\alpha}(t)x^{(t)} = \frac{d(\alpha(t)x^{(t)})}{dt} \) and integrating both sides from 1 to \( t \), we get the required statement:

\[ \alpha(t)x^{(t)} - x^{(1)} = \int_1^t \dot{\alpha}(t)\nabla \psi^*(z(\tau)) d\tau. \]

Finally, from Fact 1.6, it must be that \( \nabla \psi(z(\tau)) \in X \) for all \( \tau \in [1, t] \). As \( x^{(t)} \) is a convex combination of \( x^{(1)} \in X \) and \( \frac{1}{\alpha(t) - 1} \int_1^t \dot{\alpha}(t)\nabla \psi(z(\tau)) d\tau \in X \), the convexity of \( X \) implies that \( x^{(t)} \in X \). \( \square \)

Theorem 2.4 Let \( E_{k+1}/A_{k+1} = G_{k+1} - \frac{A_k}{A_{k+1}} G_k \) be the discretization error in Equation (2.6). Then:

\[ E_{k+1} = A_k \left( f(x^{(k+1)}) - f(x^{(k)}) \right) - a_{k+1} \left( \nabla f(x^{(k+1)}), \nabla \psi^*(z^{(k+1)}) - x^{(k+1)} \right) - D_{\psi^*}(z^{(k)}, z^{(k+1)}). \]

In particular, we have \( G_k = G_1/A_k + \sum_{i=2}^k E_i/A_k \).

Proof. Using the same arguments as in the construction of the lower bound in the continuous-time domain (beginning of the proof of Theorem 2.3), the lower bound in the discrete-time domain can be written as:

\[ L_{k+1} = \frac{1}{A_{k+1}} \sum_{i=1}^{k+1} a_i f(x^{(i)}) + \min_{x \in X} \left\{ \frac{1}{A_{k+1}} \sum_{i=1}^{k+1} a_i \left( \nabla f(x^{(i)}), x - x^{(i)} \right) + \frac{1}{A_{k+1}} D_{\psi}(x, x^{(1)}) \right\} - \frac{1}{A_{k+1}} D_{\psi}(x^*, x^{(1)}). \]

Letting \( z^{(k+1)} = \nabla \psi(x^{(1)}) - \sum_{i=1}^{k+1} \nabla f(x^{(i)}) \), we can further write:

\[ L_{k+1} = \frac{1}{A_{k+1}} \left( \sum_{i=1}^{k+1} a_i f(x^{(i)}) + \sum_{i=1}^{k+1} a_i \left( \nabla f(x^{(i)}), \nabla \psi^*(z^{(k+1)}) - x^{(i)} \right) \right) + D_{\psi}(\nabla \psi^*(z^{(k+1)}), x^{(1)}) - D_{\psi}(x^*, x^{(1)}) \]

\[ = \frac{A_k}{A_{k+1}} L_k + \frac{a_{k+1}}{A_{k+1}} \left( f(x^{(k+1)}) + \left( \nabla f(x^{(k+1)}), \nabla \psi^*(z^{(k+1)}) - x^{(k+1)} \right) \right) + \frac{1}{A_{k+1}} \left( \sum_{i=1}^{k} a_i \left( \nabla f(x^{(i)}), \nabla \psi^*(z^{(k+1)}) - \nabla \psi^*(z^{(k)}) \right) + D_{\psi}(\nabla \psi^*(z^{(k+1)}), x^{(1)}) - D_{\psi}(\nabla \psi^*(z^{(k)}), x^{(1)}) \right). \]

Recalling that \( z^{(k)} = \nabla \psi(x^{(1)}) - \sum_{i=1}^{k} a_i \nabla f(x^{(i)}) \), we can bound \( . \) as:

\[ (.)_2 = \psi(\nabla \psi^*(z^{(k+1)})) - \psi(\nabla \psi^*(z^{(k)})) - \left( z^{(k)}, \nabla \psi^*(z^{(k+1)}) - \nabla \psi^*(z^{(k)}) \right) = D_{\psi^*}(z^{(k)}, z^{(k+1)}), \]
and thus we arrive at:

\[ L_{k+1} = \frac{A_k}{A_{k+1}} L_k + \frac{1}{A_{k+1}} \left( a_{k+1} f(x^{(k+1)}) + a_{k+1} \left\langle \nabla f(x^{(k+1)}), \nabla \psi^*(z^{(k+1)}) - x^{(k+1)} \right\rangle \right) \]

which, adding and subtracting \( U_k = f(x^{(k)}) \), gives

\[
L_{k+1} = \frac{A_k}{A_{k+1}} (L_k - U_k) + f(x^{(k+1)}) \\
+ \left( \frac{A_k}{A_{k+1}} (f(x^{(k)}) - f(x^{(k+1)})) \right) \\
+ \frac{a_{k+1}}{A_{k+1}} \left\langle \nabla f(x^{(k+1)}), \nabla \psi^*(z^{(k+1)}) - x^{(k+1)} \right\rangle \\
+ \frac{1}{A_{k+1}} D_{\psi^*}(z^{(k)}, z^{(k+1)})_E,}

Remaining Proof of Theorem 2.7 (The Bound on \( G_1 \)). To bound \( G_1 \), we recall the definition of \( L_1 \):

\[
L_1 = f(x^{(1)}) + \min_{x \in A} \left\{ \left\langle \nabla f(x^{(1)}), x - x^{(1)} \right\rangle + \frac{1}{A_1} D_\psi(x, x^{(0)}) \right\} - \frac{1}{A_1} D_\psi(x^*, x^{(0)})
\]

\[
= f(x^{(1)}) + \left\langle \nabla f(x^{(1)}), \nabla \psi^*(z^{(1)}) - x^{(1)} \right\rangle + \frac{1}{A_1} D_\psi(\nabla \psi^*(z^{(1)}), x^{(0)}) - \frac{1}{A_1} D_\psi(x^*, x^{(0)}).
\]

As \( a_1 = A_1 \), \( x^{(1)} = \nabla \psi^*(\hat{z}^{(0)}) \), and \( a_1 \nabla f(\hat{x}^{(0)}) = z^{(0)} - \hat{z}^{(0)} \), using Proposition 1.9 we have that:

\[
\left\langle \nabla f(\hat{x}^{(0)}), \nabla \psi^*(z^{(1)}) - x^{(1)} \right\rangle = \frac{1}{A_1} \left\langle z^{(0)} - \hat{z}^{(0)}, \nabla \psi^*(z^{(1)}) - \nabla \psi^*(\hat{z}^{(0)}) \right\rangle
\]

\[
= \frac{1}{A_1} \left( D_\psi(z^{(0)}, \hat{z}^{(0)}) - D_\psi(z^{(0)}, z^{(1)}) + D_\psi(\hat{z}^{(0)}, z^{(1)}) \right). \tag{A.1}
\]

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

\[
\left\langle \nabla f(x^{(1)}), \nabla \psi^*(z^{(1)}) - x^{(1)} \right\rangle \geq -L \| \nabla \psi^*(\hat{z}^{(0)}) - \hat{x}^{(0)} \| \| \nabla \psi^*(z^{(1)}) - x^{(1)} \| . \tag{A.2}
\]

Finally, by Proposition 1.7 and the initial condition \( z^{(0)} = \nabla \psi(\hat{x}^{(0)}) \), we have that \( D_\psi(z^{(0)}, z^{(1)}) = D_\psi(\nabla \psi^*(z^{(1)}), \hat{x}^{(0)}) \). Combining with (A.1), (A.2), and \( G_1 = U_1 - L_1 = f(x^{(1)}) - L_1 \):

\[
G_1 \leq L \| \nabla \psi^*(\hat{z}^{(0)}) - \hat{x}^{(0)} \| \| \nabla \psi^*(z^{(1)}) - x^{(1)} \| - \frac{1}{A_1} \left( D_\psi(z^{(0)}, \hat{z}^{(0)}) + D_\psi(\hat{z}^{(0)}, z^{(1)}) \right) + \frac{1}{A_1} D_\psi(x^*, \hat{x}^{(0)})
\]

\[
= L \| \nabla \psi^*(\hat{z}^{(0)}) - \hat{x}^{(0)} \| \| \nabla \psi^*(z^{(1)}) - x^{(1)} \| - \frac{1}{A_1} \left( D_\psi(\nabla \psi^*(\hat{z}^{(0)}), \hat{x}^{(0)}) + D_\psi(\hat{z}^{(0)}, z^{(1)}) \right) + \frac{1}{A_1} D_\psi(x^*, \hat{x}^{(0)})
\]

\[
\leq L \| \nabla \psi^*(\hat{z}^{(0)}) - \hat{x}^{(0)} \| \| \nabla \psi^*(z^{(1)}) - x^{(1)} \| \leq \frac{\sigma}{2A_1} \left( \| \nabla \psi^*(\hat{z}^{(0)}) - \hat{x}^{(0)} \|^2 + \| \nabla \psi^*(z^{(1)}) - x^{(1)} \|^2 \right) + \frac{1}{A_1} D_\psi(x^*, \hat{x}^{(0)})
\]

\[
\leq \frac{1}{A_1} D_\psi(x^*, \hat{x}^{(0)}),
\]

where we have used Proposition 1.7 \( x^{(1)} = \nabla \psi^*(\hat{z}^{(0)}) \), and \( \frac{\sigma^2}{A_1} = A_1 \leq \frac{\sigma}{L} \).

\[ \Box \]

**B Discretization for Nesterov’s Algorithm**

We now consider a different discretization method: namely, the explicit Euler discretization which leads to AGD. This method allows us to perform one step per iteration (compared to AXGD that performs two
steps per iteration), at the cost of performing an additional gradient step to improve the upper bound. Observe that, by the definition of Bregman divergence:

$$
\forall x, \tilde{x} \in X : f(x) = f(\tilde{x}) + (\nabla f(\tilde{x}), x - \tilde{x}) + D_f(x, \tilde{x}), \text{ and}
$$

(B.1)

$$
\forall x, \tilde{x} \in X : \psi(x) = \psi(\tilde{x}) + (\psi(\tilde{x}), x - \tilde{x}) + D_\psi(x, \tilde{x}).
$$

(B.2)

Therefore:

We will further assume that we can use (B.1) to perform a gradient step, defined as:

$$
\text{Grad}(\tilde{x}) = \min_{x \in X} \{ (\nabla f(\tilde{x}), x - \tilde{x}) + D_f(x, \tilde{x}) \},
$$

(B.3)

where $\Delta_f(x, \tilde{x})$ is any function that bounds $D_f(x, \tilde{x})$ from above, i.e., $\Delta_f(x, \tilde{x}) \geq D_f(x, \tilde{x}), \forall x, \tilde{x} \in X$.

With these assumptions on hand, we can show the following result:

**Theorem B.1.** Let $x^{(1)} \in X$ be an arbitrary initial point and let $z^{(1)} = \nabla \psi(x^{(1)}) - f(x^{(1)})$, and $\hat{x}^{(1)} = \text{Grad}(x^{(1)})$. For $k \geq 1$, let $x^{(k)}, z^{(k)}, \hat{x}^{(k)}$ evolve according to (B.1). If, $\forall x \in X$, $\frac{1}{A_k} D_\psi(x, x^{(1)}) \geq D_f(x, x^{(1)})$ and $\forall k \geq 1 : \frac{1}{\lambda_{k+1}} D_\psi(x, \nabla \psi^*(z^{(k)})) \geq D_f(\frac{A_{k+1}}{\lambda_{k+1}} x + \frac{A_k}{\lambda_{k+1}} \hat{x}^{(k)}, x^{(k+1)})$, then

$$
f(x^{(k)}) - f(x^*) \leq \frac{1}{A_k} D_\psi(x^*, x^{(1)}).
$$

Proof. The sequence of points $\hat{x}^{(1)}, \hat{x}^{(2)}, \ldots$ is used in the construction of the upper bounds. In particular, $\forall k \geq 1 : U_k = f(\hat{x}^{(k)})$. The construction of the lower bound is as before, based on the points $x^{(1)}, x^{(2)}, \ldots$ In particular:

$$
L_{k+1} = \frac{A_k}{A_{k+1}} L_k + f(x^{(k+1)}) - \frac{A_k}{A_{k+1}} f(\hat{x}^{(k)}) - \frac{E_{k+1}}{A_{k+1}}
$$

where

$$
E_{k+1} = A_k \left( f(x^{(k+1)}) - f(\hat{x}^{(k)}) \right) \quad \text{satisfies, } \forall x \in X.
$$

Therefore, to obtain the claimed result, we need to show that $f(x^{(k+1)}) - E_{k+1}/A_{k+1} \geq f(\hat{x}^{(k+1)}) = U_{k+1}$.

$$
\psi^*(z^{(k)}) - \psi^*(z^{(k+1)}) = \min_{x \in X} \left\{ - \langle z^{(k+1)}, x \rangle + \psi(x) \right\} - \psi(\nabla \psi^*(z^{(k)})) + \langle z^{(k)}, \nabla \psi^*(z^{(k)}) \rangle.
$$

From (B.2), $\psi(x) - \langle z^{(k)}, x \rangle$ satisfies, $\forall \tilde{x} \in X$:

$$
\psi(x) - \langle z^{(k)}, x \rangle = \psi(\tilde{x}) - \langle z^{(k)}, \tilde{x} \rangle + \langle \nabla \psi(\tilde{x}) - z^{(k)}, x - \tilde{x} \rangle + D_\psi(x, \tilde{x}),
$$

since $\psi(x)$ and $\psi(x) - \langle z, x \rangle$ have the same Bregman divergence for a fixed $z$. In particular, taking $\tilde{x} = \nabla \psi^*(z^{(k)})$:

$$
\psi(x) - \langle z^{(k)}, x \rangle = \psi(\nabla \psi^*(z^{(k)})) - z^{(k)} - \langle \nabla \psi^*(z^{(k)}), x - z^{(k)} \rangle
$$

$$
\geq \psi(\nabla \psi^*(z^{(k)})) - z^{(k)} + D_\psi(x, \nabla \psi^*(z^{(k)})).
$$
where the inequality follows by the first-order optimality condition implied by Fact 1.6. It follows that:

$$\psi^*(z^{(k)}) - \psi^*(z^{(k+1)}) \geq \min_{x \in X} \left\{ \left\langle z^{(k)} - z^{(k+1)}, x \right\rangle + D_\psi(x, \nabla \psi^*(z^{(k)})) \right\}$$

$$= \min_{x \in X} \left\{ a_{k+1} \left\langle \nabla f(x^{(k+1)}), x \right\rangle + D_\psi(x, \nabla \psi^*(z^{(k)})) \right\}.$$  \hspace{1cm} (B.5)

Combining (B.5) and (B.4), we have:

$$E_{k+1} \leq - \min_{x \in X} \left\{ \left\langle \nabla f(x^{(k+1)}), a_{k+1} x + A_k \hat{x}^{(k)} - A_{k+1} x^{(k+1)} \right\rangle + D_\psi(x, \nabla \psi^*(z^{(k)})) \right\}.$$

By the theorem’s assumption, \( \frac{1}{A_k+1} D_\psi(x, \nabla \psi^*(z^{(k)})) \geq D_f(A_k, x, x^{(k+1)}) (\geq A_k+1 f(x^{(k+1)})) \). Setting \( y = \frac{a_{k+1}}{A_k+1} x + \frac{A_k}{A_k+1} \hat{x}^{(k)} \in X \):

$$\frac{E_{k+1}}{A_{k+1}} \leq - \min_{y \in X} \left\{ \left\langle \nabla f(x^{(k+1)}), y - x^{(k+1)} \right\rangle + \Delta_f(y, x^{(k+1)}) \right\} = f(\hat{x}^{(k+1)}) + f(x^{(k+1)}),$$

as \( \hat{x}^{(k+1)} = \text{Grad}(x^{(k+1)}) \) and by (B.1). Therefore:

$$L_{k+1} \geq \frac{A_k}{A_k+1} L_k + f(\hat{x}^{(k+1)}) - \frac{A_k}{A_k+1} f(\hat{x}^{(k+1)}),$$

and it follows that:

$$U_{k+1} - L_{k+1} \leq \frac{A_k}{A_k+1} (U_k - L_k).$$

To complete the proof, it remains to bound the initial gap \( U_1 - L_1 \). To do so, it suffices to show that:

$$L_1 \geq f(\hat{x}^{(1)}) - \frac{1}{A_1} D_\psi(x^*, x^{(1)}).$$

From the definition of the lower bound and the initial conditions:

$$L_1 = f(x^{(1)}) + \min_{x \in X} \left\{ \frac{a_1}{A_1} \left\langle \nabla f(x^{(1)}), x - x^{(1)} \right\rangle + \frac{1}{A_1} D_\psi(x, x^{(1)}) \right\} - \frac{1}{A_1} D_\psi(x^*, x^{(1)})$$

$$\geq f(x^{(1)}) + \min_{x \in X} \left\{ \left\langle \nabla f(x^{(1)}), x - x^{(1)} \right\rangle + \frac{1}{A_1} \Delta_f(x, x^{(1)}) \right\} - D_\psi(x^*, x^{(1)})$$

$$\geq f(\hat{x}^{(1)}) - D_\psi(x^*, x^{(1)}),$$

where the second inequality follows by the theorem assumptions.

\( \square \)

**Theorem B.2.** Let \( f : X \to \mathbb{R} \) be an \( L \)-smooth convex function and let \( x^{(k)}, z^{(k)}, \hat{x}^{(k)} \) evolve as in Theorem B.1. Let \( \psi : X \to \mathbb{R} \) be \( \sigma \)-strongly convex. If the steps \( a_k \) satisfy \( \frac{a_k^2}{A_k} \leq \frac{\sigma}{2} \), then, \( \forall k \geq 1 \):

$$f(\hat{x}^{(k)}) - f(x^*) \leq \frac{D_\psi(x^*, x^{(1)})}{A_k}. \quad \text{In particular, if } a_k = \frac{k+1}{2} \cdot \frac{\sigma}{L} \text{ and } \psi(x) = \frac{\sigma}{2} \|x\|^2, \text{ then:}$$

$$f(\hat{x}^{(k)}) - f(x^*) \leq \frac{L}{(k+1)^2} \|x^* - x^{(1)}\|^2.$$  \hspace{1cm} \text{(B.6)}

**Proof.** The proof follows by applying Theorem B.1 and by using the smoothness of \( f(.) \) and the strong convexity of \( \psi(.) \). In particular, the smoothness of \( f(.) \) gives \( D_f(x, \hat{x}) \leq \frac{L}{2} \|x - \hat{x}\|^2, \forall x, \hat{x} \in X \), while the strong convexity of \( \psi(.) \) gives \( D_\psi(x, \hat{x}) \geq \frac{\sigma}{2} \|x - \hat{x}\|^2 \geq \frac{\sigma}{2} D_f(x, \hat{x}), \forall x, \hat{x} \in X \). Therefore:

$$\frac{1}{A_k} D_\psi(x, x^{(1)}) = \frac{1}{\sigma} D_\psi(x, x^{(1)}) \geq D_f(x, x^{(1)}).$$

For \( k \geq 1 \):

$$\frac{1}{A_{k+1}} D_\psi(x, \nabla \psi^*(z^{(k)})) \geq \frac{\sigma}{2A_{k+1}} \|x - \nabla \psi^*(z^{(k)})\|^2 \geq D_f(\frac{a_{k+1}}{A_{k+1}} x + \frac{A_k}{A_{k+1}} \hat{x}^{(k)}, x^{(k+1)}),$$

where the last inequality follows by \( \nabla \psi^*(z^{(k)}) = \frac{A_{k+1}}{a_{k+1}} (x^{(k+1)} - \frac{A_k}{A_{k+1}} \hat{x}^{(k)}) \) and from the choice of \( a_k \)’s.

Finally, choosing \( a_k = \frac{k+1}{2} \cdot \frac{\sigma}{L} \) and \( \psi(x) = \frac{\sigma}{2} \|x\|^2 \) gives \( f(\hat{x}^{(k)}) - f(x^*) \leq \frac{L}{(k+1)^2} \|x^* - x^{(1)}\|^2. \) \( \square \)