The Approximate Gap Technique: A Unified Approach to Optimal First-Order Methods

Jelena Diakonikolas and Lorenzo Orecchia
Computer Science Department, Boston University
{jelenad, orecchia}@bu.edu

Abstract

We provide a general technique for the analysis of first-order methods. The technique relies on constructions of upper and lower bounds on the optimal objective value, leading to approximate optimality gaps. We show that in continuous time enforcement of an invariant that the approximate gap decreases at a certain rate exactly recovers a wide range of first-order continuous-time methods. We then characterize the discretization errors incurred by different discretization methods, and derive optimal (in terms of iteration complexity) methods for various classes of problems, including: solving variational inequalities for smooth monotone operators, convex minimization for Lipschitz-continuous objectives, smooth convex minimization, composite minimization, and smooth and strongly convex minimization.

1 Introduction

First-order optimization methods have recently gained high popularity due to their applicability to large-scale problem instances arising from modern datasets, their relatively low computational complexity and potential for parallelizing computation [16]. Moreover, such methods have also been successfully applied in discrete optimization [7]. Most first-order optimization methods can be derived from the discretization of continuous-time dynamical systems that converge to optimal solutions. In the case of mirror descent, the continuous view was the original motivation for the algorithm [9], while more recent work has focused on deducing continuous-time interpretations of accelerated methods [6, 17, 19, 20].

Motivated by these works, we focus on providing a unified analysis of various first-order methods. We term this general framework of analyzing first-order methods the *Approximate Gap Technique (AGT)*. In addition to providing an intuitive and unified analysis of various first-order methods that often requires only a few lines to analyze the convergence bound, the general technique is valuable in developing new first-order methods with tight convergence bounds, as well as in analyzing properties such as noise robustness and model misspecification [3].

Unlike recent works that start from a certain continuous-time description of a method and then rely on the use of Lyapunov stability criteria to prove convergence bounds [6, 17, 19, 20], our approach relies on purely optimization-motivated arguments that naturally lead to continuous-time dynamics. In particular, we derive a general technique for the construction of an *approximate optimality gap*. Further, we show that various continuous-time methods can be directly obtained by enforcing an invariant that the gap decreases at a certain rate, i.e., that $\alpha^{(t)}G^{(t)}$ is a non-increasing function of time t, where $\alpha^{(t)}$ is a strictly increasing positive function and $G^{(t)}$ is the approximate gap. We then pinpoint the precise places where the discretization of the continuous-time dynamics (and the invariance condition on the gap reduction) incur the discretization error, and provide a characterization of the discretization error for various methods. Then, optimal first-order discrete-time methods for different classes of problems follow by cancelling out the discretization error (or keeping it "sufficiently small").

We provide example analysis for several well-known first order methods, such as mirror descent/dual averaging [9], mirror-prox/extra-gradient method [5, 12], accelerated methods [10], composite minimization methods [4, 15] and Frank-Wolfe methods [14], to illustrate the power and generality of the technique. The same ideas naturally extend to other classes of convex optimization problems and their corresponding optimal first-order methods. Here, "optimal" is in the sense that the methods yield worst-case iteration complexity bounds for which there is a matching lower bound (i.e., "optimal" is in terms of iteration complexity).

1.1 Related Work

There has been a very large body of research in optimization and first-order methods in particular, and, while we cannot provide a thorough literature review, we refer the reader to recent books [2,16]. There have been multiple efforts to unify the analysis of first-order methods, with a particular focus on accelerated methods. Tseng gives a formal framework that unifies all the different instantiations of accelerated gradient methods [18]. More recently, Allen-Zhu and Orecchia [1] interpret acceleration as coupling of mirror descent and gradient descent steps. Bubeck *et al.* provide an elegant geometric interpretation of the Euclidean instantiation of Nesterov's method [2]. Wibisono, *et al.* [19] and Wilson *et al.* [20] interpret accelerated methods using Lyapunov stability analysis and drawing ideas from Lagrangian mechanics.

1.2 Notation

We use non-boldface letters to denote scalars and boldface letters to denote vectors. Superscript index $(\cdot)^{(t)}$ will denote the value of $(\cdot)$ at time t. The "dot" notation will be used to denote the time derivative, i.e., $\dot{x} = \frac{dx}{dt}$. Given a measure $\alpha^{(\tau)}$ defined on $\tau \in [t_0, t]$, we will use the Lebesgue-Stieltjes notation for the integral. In particular, given $\phi^{(\tau)}$ defined on $\tau \in [t_0, t]$:

$$\int_{t_0}^{t} \phi^{(\tau)} \dot{\alpha}^{(\tau)} d\tau = \int_{t_0}^{t} \phi^{(\tau)} d\alpha^{(\tau)}.$$

When $\alpha^{(\tau)}$ is a discrete measure and both $\phi^{(\tau)}$ and $\alpha^{(\tau)}$ are defined on $\tau \in \{t_0, t_0+1, ..., t\}$, we have that $\int_{t_0}^{t} \phi^{(\tau)} d\alpha^{(\tau)} = \sum_{\tau=t_0}^{t} \phi^{(\tau)} d\alpha^{(\tau)}$. We denote $A^{(t)} = \int_{t_0}^{t} d\alpha^{(\tau)}$, so that $\frac{1}{A^{(t)}} \int_{t_0}^{t} d\alpha^{(\tau)} = 1$. We will assume throughout the paper that $\dot{\alpha}^{(t)} > 0, \forall t \geq t_0$, where t_0 is the initial point of the (continuous or discrete) dynamics, and use the following notation for the aggregated gradients:

$$\mathbf{z}^{(t)} \stackrel{\text{def}}{=} -\int_{t_0}^{t} \nabla f(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}. \tag{1.1}$$

For all considered problems, we will assume a linear, finite-dimensional (primal) vector space $X \subseteq \mathbb{R}^n$ on which the problems are defined. We will assume that there is a (fixed) norm $\|\cdot\|$ associated with space X and define its dual norm in a standard way: $\|\mathbf{z}\|_* = \max_{\mathbf{x} \in X} \{\langle \mathbf{z}, \mathbf{x} \rangle : \|\mathbf{x}\| \leq 1\}$, where $\langle \cdot, \cdot \rangle$ denotes the inner product.

1.3 Preliminaries

We focus on minimizing a continuous and differentiable convex function $f(\cdot)$ defined on a convex set $X \subseteq \mathbb{R}^n$, and we let $\mathbf{x}^* = \arg\min_{\mathbf{x} \in X} f(\mathbf{x})$ denote the minimizer of $f(\cdot)$ on X. The following definitions will be useful in our analysis, and thus we state them here for completeness.

Definition 1.1. 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 1.2. A function $f : X \to \mathbb{R}$ is smooth on X with smoothness parameter L and with respect to a norm $\|.\|$, if for all $\mathbf{x}, \hat{\mathbf{x}} \in X$: $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x} \rangle + \frac{L}{2} \|\hat{\mathbf{x}} - \mathbf{x}\|^2$. Equivalently: $\|\nabla f(\mathbf{x}) - \nabla f(\hat{\mathbf{x}})\|_* \leq L \|\mathbf{x} - \hat{\mathbf{x}}\|$.

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

Definition 1.4. (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 1.5. (Convex Conjugate) Function ψ^* is the convex conjugate of $\psi : X \to \mathbb{R}$, if $\psi^*(\mathbf{z}) = \max_{\mathbf{x} \in X} \{\langle \mathbf{z}, \mathbf{x} \rangle - \psi(\mathbf{x})\}$, $\forall \mathbf{z} \in \mathbb{R}$.

We will assume that there is a differentiable function $\phi : X \to \mathbb{R}$, possibly dependent on t (in which case we denote it as ϕ_t), such that $\max_{\mathbf{x} \in X} \{\langle \mathbf{z}, \mathbf{x} \rangle - \phi(\mathbf{x})\}$ is easily solvable, possibly in a closed form. Notice that this problem defines the convex conjugate of $\phi(\cdot)$, i.e., $\phi^*(\mathbf{z}) = \max_{\mathbf{x} \in X} \{\langle \mathbf{z}, \mathbf{x} \rangle - \phi(\mathbf{x})\}$. The following standard fact will be extremely useful in carrying out the analysis of the algorithms in this paper.

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

$$\nabla \phi^*(\mathbf{z}) = \arg\max_{\mathbf{x} \in X} \{\langle \mathbf{z}, \mathbf{x} \rangle - \phi(\mathbf{x})\} = \arg\min_{\mathbf{x} \in X} \{-\langle \mathbf{z}, \mathbf{x} \rangle + \phi(\mathbf{x})\}.$$

In particular, Fact 1.6 implies:

$$\nabla\phi^*(\mathbf{z}^{(t)}) = \arg\min_{\mathbf{x}\in X}\left\{\int_{t_0}^{t}\left\langle\nabla f(\mathbf{x}^{(\tau)}), \mathbf{x}-\mathbf{x}^{(\tau)}\right\rangle d\alpha^{(\tau)} + \phi(\mathbf{x})\right\} \tag{1.2}$$

Some other useful properties of Bregman divergence can be found in the appendix.

Overview of Continuous-Time Operations In continuous time, changes in the variables are described by differential equations. Of particular interest are (weighted) aggregation and averaging. Aggregation of a function $g(x)$ is given as $\dot{y}^{(t)} = \dot{\alpha}^{(t)} g(x^{(t)})$, where the "dot" notation is used to denote the time derivative. Observe that, by integrating both sides from t_0 to t, this is equivalent to: $y^{(t)} = y^{(t_0)} + \int_{t_0}^{t} g(x^{(\tau)}) d\alpha^{(\tau)}$. Averaging of a function $g(x)$ is given as $\dot{y}^{(t)} = \dot{\alpha}^{(t)} \frac{g(x^{(t)}) - y^{(t)}}{\alpha^{(t)}}$. This can be equivalently written as $\frac{d}{dt}(\alpha^{(t)} y^{(t)}) = \dot{\alpha}^{(t)} g(x^{(t)})$, implying $y^{(t)} = \frac{\alpha^{(t_0)}}{\alpha^{(t)}} y^{(t_0)} + \frac{1}{\alpha^{(t)}} \int_{t_0}^{t} g(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}$.

The following simple proposition (based on Danskin's theorem) will be very useful in our analysis.

Proposition 1.7. $\frac{d}{dt}\min_{\mathbf{x}\in X}\left\{-\left\langle\mathbf{z}^{(t)}, \mathbf{x}\right\rangle + \phi(\mathbf{x})\right\} = -\left\langle\dot{\mathbf{z}}^{(t)}, \nabla\phi^*(\mathbf{z}^{(t)})\right\rangle$.

Proof. Observe that $\phi^*(\mathbf{z}^{(t)}) = -\min_{\mathbf{x}\in X}\left\{-\left\langle\mathbf{z}^{(t)}, \mathbf{x}\right\rangle + \phi(\mathbf{x})\right\}$. By Fact 1.6, $\nabla\phi^*(\mathbf{z}^{(t)}) = \min_{\mathbf{x}\in X}\left\{-\left\langle\mathbf{z}^{(t)}, \mathbf{x}\right\rangle + \phi(\mathbf{x})\right\}$. Using Danskin's theorem (which allows us to differentiate inside the min), we can compute $\frac{d}{dt}\phi^*(\mathbf{z}^{(t)})$ by differentiating first w.r.t. $\mathbf{z}^{(t)}$ and then w.r.t. t, which yields the claimed identity. □

2 The Approximate Gap Technique

Underlying the analysis of all the first-order methods considered here is the notion of a lower bound $L^{(t)}$ and an upper bound $U^{(t)}$ of the optimal objective value $f(\mathbf{x}^*)$, together with the approximate optimality gap defined as $G^{(t)} = U^{(t)} - L^{(t)}$. The explicit construction of the upper and lower bounds allows us to take a unified primal-dual view of the methods and quantify the convergence rate as the rate at which the approximate gap $G^{(t)}$ decreases as a function of time t. We will refer to this general framework of constructing and quantifying the optimality gap $G^{(t)}$ as the Approximate Gap Technique (AGT).

2.1 Upper Bound

Since we are interested in minimizing a function $f(\cdot)$, any point $\mathbf{x} \in X$ leads to a valid upper bound, as $f(\mathbf{x}) \geq f(\mathbf{x}^*)$. Suppose that a minimization algorithm constructs a sequence of points $\mathbf{x}^{(\tau)}$. To keep the gap between upper and lower bounds as small as possible, a natural choice of $\mathbf{x} \in X$ at time t is the best seen point, that is, $\mathbf{x} = \arg\min_{\tau\in[t_0,t]} f(\mathbf{x}^{(\tau)})$. However, different choices of upper bounds will turn out to be more useful for the analysis, such as the function value at the last seen point or a convex combination of the function values $f(\mathbf{x}^{(\tau)})$ at all seen points $\mathbf{x}^{(\tau)}$ for $\tau \in [t_0, t]$, i.e., $U^{(t)} = \frac{1}{A^{(t)}} \int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}$.

2.2 Lower Bound

The construction of the lower bound is more interesting, since it allows us to take a dual viewpoint of the algorithm behavior. Suppose that we are given a sequence of points $\mathbf{x}^{(\tau)} \in X$, for $\tau \in [t_0, t]$. The convexity of $f(\cdot)$ leads to the lower-bounding hyperplanes, since $f(\mathbf{x}) \geq f(\mathbf{x}^{(\tau)}) + \left\langle\nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)}\right\rangle$, $\forall \mathbf{x} \in X$. Given a non-negative measure $\alpha^{(\tau)}$ and $A^{(t)} = \int_{t_0}^{t} d\alpha^{(\tau)}$, we have the following natural lower bound:

$$f(\mathbf{x}) \geq \frac{1}{A^{(t)}} \int_{t_0}^{t} \left(f(\mathbf{x}^{(\tau)}) + \left\langle\nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)}\right\rangle\right) d\alpha^{(\tau)}, \quad \forall \mathbf{x} \in X. \tag{2.1}$$

Taking a minimum over $\mathbf{x} \in X$ in the last equation would give a valid (and natural) lower bound that only depends on the points seen so far. However, such a lower bound would in general be non-differentiable and even non-continuous as a function of t (which is very important, since we will be interested in observing and controlling the changes in the lower bound as a function of t). To attain more stable behavior of the lower bound, we can

add a convex continuously-differentiable function $\phi(\mathbf{x})$ to both sides of (2.1) and then take the minimum over $\mathbf{x} \in X$ on the right-hand side of (2.1).[1] Setting $\mathbf{x} = \mathbf{x}^*$ on the left-hand side of (2.1):

$$f(\mathbf{x}^*) \geq L^{(t)} \stackrel{\text{def}}{=} \frac{\int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha + \min_{\mathbf{x} \in X} \left\{ \int_{t_0}^{t} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)} \right\rangle d\alpha + \phi(\mathbf{x}) \right\} - \phi(\mathbf{x}^*)}{A^{(t)}}. \tag{2.2}$$

The potential drawback of using (2.2) is that in continuous time it may not be defined at the initial point t_0. To circumvent this issue, we can mix in the trivial lower bound $f(\mathbf{x}^*) \geq f(\mathbf{x}^*)$, obtaining the following lower bound:

$$L^{(t)} \stackrel{\text{def}}{=} \frac{\int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha + \min_{\mathbf{x} \in X} \left\{ \int_{t_0}^{t} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)} \right\rangle d\alpha + \phi(\mathbf{x}) \right\} - \phi(\mathbf{x}^*)}{\alpha^{(t)}} + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}} f(\mathbf{x}^*). \tag{2.3}$$

Note that in the discrete time $A^{(t)} = \alpha^{(t)}$, and thus $L^{(t)}$ is equivalent to the right-hand side of (2.2).

Observe that the minimum in $L^{(t)}$ is equal to $-\phi^*(\mathbf{z}^{(t)}) - \int_{t_0}^{t} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x}^{(\tau)} \right\rangle d\alpha^{(\tau)}$, where $\mathbf{z}^{(t)} = -\int_{t_0}^{t} \nabla f(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}$. It will be useful to think about $\mathbf{z}^{(t)}$ as the dual variable defined over the linear space of the gradients of $f(\cdot)$, and about maximizing $L^{(t)}$ as a dual problem to minimizing $f(\cdot)$. Then, $-\phi^*(\mathbf{z}^{(t)})$ can be interpreted as the "shape" of the dual function, while $\frac{1}{A^{(t)}}$ determines its scale and the remaining terms in $L^{(t)}$ determine its position.

Extension to Strongly Convex Objectives. When the objective is σ-strongly convex for some $\sigma > 0$, we can use σ-strong convexity (instead of just regular convexity) in the construction of the lower bound. This will generally give us a better lower bound which will in general lead to the better convergence guarantees in the discrete-time domain. It is not hard to verify (by repeating the same arguments as above) that in this case we have the following lower bound:

$$L^{(t)} = \frac{\int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha + \min_{\mathbf{x} \in X} \left\{ \int_{t_0}^{t} \left(\left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)} \right\rangle + \frac{\sigma}{2} \|\mathbf{x} - \mathbf{x}^{(\tau)}\|^2 \right) d\alpha + \phi(\mathbf{x}) \right\} - \phi(\mathbf{x}^*)}{\alpha^{(t)}} + \frac{A^{(t)} - \alpha^{(t)}}{\alpha^{(t)}} f(\mathbf{x}^*). \tag{2.4}$$

Extension to Composite Objectives. Suppose now that we have a composite objective $\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x})$. Then, we can choose to apply the convexity argument only to $f(\cdot)$ and use $\psi(\cdot)$ as a regularizer (this will be particular useful in the discrete-time domain in the settings where $f(\cdot)$ has some smoothness properties while $\psi(\cdot)$ is generally non-smooth). Therefore, we could start with $\bar{f}(\mathbf{x}) \geq f(\hat{\mathbf{x}}) + \langle \nabla f(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle + \psi(\mathbf{x})$. Repeating the same arguments as in the general construction of the lower bound presented earlier in this subsection:

$$L^{(t)} \stackrel{\text{def}}{=} \frac{\int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha + \min_{\mathbf{x} \in X} \left\{ \int_{t_0}^{t} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)} \right\rangle d\alpha + A^{(t)} \psi(\mathbf{x}) + \phi(\mathbf{x}) \right\} - \phi(\mathbf{x}^*)}{\alpha^{(t)}} + \frac{A^{(t)} - \alpha^{(t)}}{\alpha^{(t)}} \bar{f}(\mathbf{x}^*). \tag{2.5}$$

2.3 The Approximate Gap

The gap $G^{(t)}$ is simply defined as $G^{(t)} = U^{(t)} - L^{(t)}$. The crux of the analysis is to show that $G^{(t)}$ decays proportionally to $\frac{1}{A^{(t)}}$. That is, the "invariance" in the analysis of all considered first-order methods will be that $A^{(t+\Delta t)} G^{(t+\Delta t)} - A^{(t)} G^{(t)} \leq E_{t,\Delta t}$ for $E_{t,\Delta t} = 0$ or $E_{t,\Delta t} > 0$ but "sufficiently small".

Extension to Monotone Operators and Saddle-Point Formulations. The notion of the approximate gap can be defined for classes beyond convex functions. Examples are monotone operators and convex-concave saddle point problems. Given a monotone operator $F : X \to \mathbb{R}^n$, the goal is to find a point $\mathbf{x}^* \in X$ such that $\langle F(\mathbf{u}), \mathbf{x}^* - \mathbf{u} \rangle \leq 0$, $\forall \mathbf{u} \in X$. The approximate version of this problem is:

$$\text{Find } \mathbf{x}_\epsilon \in X \text{ such that } \langle F(\mathbf{u}), \mathbf{x}_\epsilon - \mathbf{u} \rangle \leq \epsilon, \quad \forall \mathbf{u} \in X, \tag{2.6}$$

and we can think of ϵ on the right-hand side of (2.6) as the optimality gap.

[1] This approach is similar to the Moreau-Yosida regularization.

The property of monotone operators $F(\cdot)$ useful for the approximate gap analysis (that also applies to the aforementioned saddle-point formulations) is $\forall \mathbf{x}, \mathbf{u} \in X$: $\langle F(\mathbf{u}), \mathbf{x} - \mathbf{u} \rangle \leq \langle F(\mathbf{x}), \mathbf{x} - \mathbf{u} \rangle$. The approximate gap can be constructed using the same ideas as in the case of a convex function, which, letting $\hat{\mathbf{x}}^{(t)} = \frac{1}{\alpha^{(t)}} \int_{t_0}^{t} \mathbf{x}^{(\tau)} d\alpha + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}} \mathbf{x}^{(0)}$, gives, $\forall \mathbf{u} \in X$:

$$\begin{aligned} \left\langle F(\mathbf{u}), \hat{\mathbf{x}}^{(t)} - \mathbf{u} \right\rangle \leq G^{(t)} \stackrel{\text{def}}{=} & \frac{\max_{\mathbf{x} \in X} \left\{ \int_{t_0}^{t} \left\langle F(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)} \right\rangle d\alpha + \phi(\mathbf{x}) \right\} - \max_{\mathbf{u} \in X} \phi(\mathbf{u})}{\alpha^{(t)}} \\ & + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}} \max_{\mathbf{v} \in X} \left\langle F(\mathbf{v}), \mathbf{x}^{(0)} - \mathbf{v} \right\rangle, \end{aligned} \tag{2.7}$$

Now assume that we want to find a saddle point of a function $\Phi(\mathbf{v}, \mathbf{w}) : V \times W \to \mathbb{R}$ that is convex in $\mathbf{v}$ and concave in $\mathbf{w}$. By convexity in $\mathbf{v}$ and concavity in $\mathbf{w}$, we have that, for all $\mathbf{v}, \mathbf{v}^{(\tau)} \in Y$ and all $\mathbf{w}, \mathbf{w}^{(\tau)} \in Z$:

$$\Phi(\mathbf{v}, \mathbf{w}^{(\tau)}) - \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}) \geq \left\langle \nabla_{\mathbf{v}} \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}), \mathbf{v} - \mathbf{v}^{(\tau)} \right\rangle, \tag{2.8}$$

$$\Phi(\mathbf{v}^{(\tau)}, \mathbf{w}) - \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}) \leq \left\langle \nabla_{\mathbf{w}} \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}), \mathbf{w} - \mathbf{w}^{(\tau)} \right\rangle, \tag{2.9}$$

where $\nabla_{\mathbf{v}}$ (resp. $\nabla_{\mathbf{w}}$) denotes the gradient w.r.t. $\mathbf{v}$ (resp. $\mathbf{w}$).

Combining (2.8) and (2.9), it follows that, $\forall \mathbf{v}, \mathbf{v}^{(\tau)} \in Y$, $\forall \mathbf{w}, \mathbf{w}^{(\tau)} \in Z$:

$$\Phi(\mathbf{v}^{(\tau)}, \mathbf{w}) - \Phi(\mathbf{v}, \mathbf{w}^{(\tau)}) \leq \left\langle \nabla_{\mathbf{v}} \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}), \mathbf{v}^{(\tau)} - \mathbf{v} \right\rangle - \left\langle \nabla_{\mathbf{w}} \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}), \mathbf{w} - \mathbf{w}^{(\tau)} \right\rangle.$$

Letting $\mathbf{x} = [\mathbf{v}, \mathbf{w}]^T$, $F(\mathbf{x}) = [\nabla_{\mathbf{v}} \Phi(\mathbf{v}, \mathbf{w}), -\nabla_{\mathbf{w}} \Phi(\mathbf{v}, \mathbf{w})]^T$, and denoting $\overline{\mathbf{v}} = \frac{1}{A^{(t)}} \int_{t_0}^{t} \mathbf{v}^{(\tau)} d\alpha^{(\tau)}$, $\overline{\mathbf{w}} = \frac{1}{A^{(t)}} \int_{t_0}^{t} \mathbf{w}^{(\tau)} d\alpha^{(\tau)}$, we have, $\forall \mathbf{x} = [\mathbf{v}, \mathbf{w}]^T \in V \times W$:

$$\Phi(\overline{\mathbf{v}}, \mathbf{w}) - \Phi(\mathbf{v}, \overline{\mathbf{w}}) \leq \frac{\int_{t_0}^{t} \left\langle F(\mathbf{x}), \mathbf{x}^{(\tau)} - \mathbf{x} \right\rangle d\alpha^{(\tau)}}{A^{(t)}},$$

and using the same arguments as before, we obtain the same bound for the gap as in (2.7). Therefore, we can focus only on analyzing the decrease of $G^{(t)}$ from (2.7) as a function of t and the same result will follow for the gap of convex-concave saddle-point problems.

3 First-Order Methods in Continuous Time

We now show how different choices of the upper bound and the lower bound (and, consequently, the gap) lead to different first-order methods. The beauty of the approach is that by choosing some obvious upper bounds (such as, e.g., convex combination of function values at all seen points or the function value at the last seen point), most well-known first-order methods will follow directly from the invariance condition that $A^{(t)} G^{(t)}$ be non-increasing with time t.

3.1 Mirror Descent Methods

Recall the expression for the general lower bound (2.3). Since $\frac{\int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}}{\alpha^{(t)}} + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}} f(\mathbf{x}^*)$ is a convex combination of the objective function values evaluated at feasible points, a natural choice of an upper bound would be $U^{(t)} = \frac{\int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}}{\alpha^{(t)}} + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}} f(\mathbf{x}^*)$. However, since we do not know $\mathbf{x}^*$ and we would like to have point $\hat{\mathbf{x}}^{(t)}$ such that $f(\hat{\mathbf{x}}^{(t)}) \leq U^{(t)}$ (this would be the point that the algorithm returns at time t), we can choose:

$$U^{(t)} = \frac{\int_{t_0}^{t} f(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}}{\alpha^{(t)}} + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}} f(\mathbf{x}^{(0)}),$$

where $\mathbf{x}^{(0)} = \mathbf{x}^{(t_0)} \in X$ is an arbitrary initial point. Then, the gap becomes:

$$G^{(t)} = \frac{-\min_{\mathbf{x} in X} \left\{ \int_{t_0}^{t} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)} \right\rangle d\alpha^{(\tau)} + \phi(\mathbf{x}) \right\} + \phi(\mathbf{x}^*)}{\alpha^{(t)}} + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}} (f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)). \tag{3.1}$$

Observe that $G^{(t_0)} = \frac{\phi(\mathbf{x}^*)}{\alpha^{(t_0)}} + f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)$. Thus, if we show that $\frac{d}{dt}(\alpha^{(t)}G^{(t)}) \leq 0$, this would immediately imply:

$$f(\hat{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*) \leq U^{(t)} - L^{(t)} \leq \frac{\alpha^{(t_0)}}{\alpha^{(t)}} G^{(t_0)} \leq \frac{\alpha^{(t_0)}(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)) + \phi(\mathbf{x}^*)}{\alpha^{(t)}}.$$

To ensure that $\phi(\mathbf{x}^*)$ is bounded, a common choice is $\phi(\mathbf{x}) = D_\psi(\mathbf{x}, \mathbf{x}^{(0)})$ for some convex function $\psi(\cdot)$.

Now we show that ensuring the invariance $\frac{d}{dt}(\alpha^{(t)}G^{(t)}) = 0$ exactly produces the continuous-time mirror descent dynamics. Recall that $\mathbf{z}^{(t)} = -\int_{t_0}^{t} \nabla f(\mathbf{x}^{(\tau)}) d\alpha^{(t)}$. Using (1.2) and Proposition (1.7):

$$\frac{d}{dt}(\alpha^{(t)}G^{(t)}) = -\left\langle \nabla f(\mathbf{x}^{(t)}), \nabla\phi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)} \right\rangle.$$

Thus, to have $\frac{d}{dt}(\alpha^{(t)}G^{(t)}) = 0$, we need $\mathbf{x}^{(t)} = \nabla\phi^*(\mathbf{z}^{(t)})$, which is precisely the mirror descent dynamics from [9]:

$$\begin{aligned} \mathbf{z}^{(t)} &= -\dot{\alpha}^{(t)} \nabla f(\mathbf{x}^{(t)}), \\ \mathbf{x}^{(t)} &= \nabla\phi^*(\mathbf{z}^{(t)}), \\ \hat{\mathbf{x}}^{(t)} &= \dot{\alpha}^{(t)} \frac{\mathbf{x}^{(t)} - \hat{\mathbf{x}}^{(t)}}{\alpha^{(t)}}, \\ \mathbf{z}^{(t_0)} = 0,\ \hat{\mathbf{x}}^{(t_0)} &= \mathbf{x}^{(0)},\ \text{for arbitrary initial point } \mathbf{x}^{(0)} \in X. \end{aligned} \tag{CT-MD}$$

Observe that, by replacing $\nabla f(\mathbf{x}^{(t)})$ by $F(\mathbf{x}^{(t)})$ and by the same arguments, (CT-MD) also ensures $\frac{d}{dt}(\alpha^{(t)}G^{(t)}) = 0$ for the gap (2.7) derived for monotone operators and saddle point problems.

The described results are summarized in the following lemma:

Lemma 3.1. Let $\mathbf{x}^{(t)}, \hat{\mathbf{x}}^{(t)}$ evolve according to (CT-MD) for some convex function $f : X \to \mathbb{R}$. Then, $\forall t > t_0$:

$$f(\hat{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*) \leq \frac{\alpha^{(t_0)}(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)) + \phi(\mathbf{x}^*)}{\alpha^{(t)}}.$$

If instead of convex minimization we are given a variational inequality problem with monotone operator $F : \mathbb{R}^n \to \mathbb{R}^n$, then a version of (CT-MD) that replaces $\nabla f(\mathbf{x}^{(t)})$ by $F(\mathbf{x}^{(t)})$ ensures that, $\forall t > t_0$, $\forall \mathbf{u} \in X$:

$$\left\langle F(\mathbf{u}), \hat{\mathbf{x}}^{(t)} - \mathbf{u} \right\rangle \leq \frac{\max_{\mathbf{x}' \in X} \phi(\mathbf{x}') + \alpha^{(t_0)} \max_{\mathbf{x}'' \in X} \left\langle F(\mathbf{x}''), \mathbf{x}^{(0)} - \mathbf{x}'' \right\rangle}{\alpha^{(t)}}.$$

Moreover, for a convex-concave saddle-point problem $\min_{\mathbf{v} \in V} \max_{\mathbf{w} \in W} \Phi(\mathbf{v}, \mathbf{w})$, taking $\mathbf{x} = [\mathbf{v}, \mathbf{w}]^T$, $F(\mathbf{x}) = [\nabla_{\mathbf{v}}\Phi(\mathbf{v}, \mathbf{w}), -\nabla_{\mathbf{w}}\Phi(\mathbf{v}, \mathbf{w})]^T$, then the version of (CT-MD) that uses the monotone operator $F(\mathbf{x})$ ensures that, $\forall t > t_0$, $\forall (\mathbf{v}, \mathbf{w}) \in V \times W$:

$$\Phi(\hat{\mathbf{v}}^{(t)}, \mathbf{w}) - \Phi(\mathbf{v}, \hat{\mathbf{w}}^{(t)}) \leq \frac{\max_{\mathbf{x}' \in X} \phi(\mathbf{x}') + \alpha^{(t_0)} \max_{\mathbf{v}'' \in V, \mathbf{w}'' \in W} \left\{ \Phi(\hat{\mathbf{v}}^{(t_0)}, \mathbf{w}'') - \Phi(\mathbf{v}'', \hat{\mathbf{w}}^{(t_0)}) \right\}}{\alpha^{(t)}}.$$

3.2 Accelerated Convex Minimization

The previous subsection considered the choice of the upper bound that takes all seen points into account. We may hope, however, that there is an algorithm whose objective function value at the last seen point is "good enough". Interestingly, we will see that choosing $U^{(t)} = f(\mathbf{x}^{(t)})$ and $L^{(t)}$ from (2.3) results in the accelerated dynamics. To do so, as before, we need to show that the invariance condition $\frac{d}{dt}(\alpha^{(t)}G^{(t)}) \leq 0$ is satisfied for some choice of $\mathbf{x}^{(t)}$. We have:

$$\begin{aligned} \frac{d}{dt}(\alpha^{(t)}G^{(t)}) =& \frac{d}{dt}(\alpha^{(t)} f(\mathbf{x}^{(t)})) - \dot{\alpha}^{(t)} \left(f(\mathbf{x}^{(t)}) + \left\langle \nabla f(\mathbf{x}^{(t)}), \nabla\phi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)} \right\rangle \right) \\ =& \alpha^{(t)} \frac{d}{dt}(f(\mathbf{x}^{(t)})) - \dot{\alpha}^{(t)} \left\langle \nabla f(\mathbf{x}^{(t)}), \nabla\phi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)} \right\rangle \\ =& \left\langle \nabla f(\mathbf{x}^{(t)}), \alpha^{(t)}\dot{\mathbf{x}}^{(t)} - \dot{\alpha}^{(t)}(\nabla\phi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}) \right\rangle, \end{aligned}$$

where we have used $\frac{d}{dt}(f(\mathbf{x}^{(t)})) = \left\langle \nabla f(\mathbf{x}^{(t)}), \dot{\mathbf{x}}^{(t)} \right\rangle$. Therefore, choosing $\dot{\mathbf{x}}^{(t)} = \dot{\alpha}^{(t)} \frac{\nabla \phi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}$, we get $\frac{d}{dt}(\alpha^{(t)} G^{(t)}) = 0$. This is precisely the accelerated mirror descent dynamics [6,19], and we immediately get the convergence result stated as Lemma 3.2 below.

$$\begin{aligned} \dot{\mathbf{z}}^{(t)} &= -\dot{\alpha}^{(t)} \nabla f(\mathbf{x}^{(t)}), \\ \dot{\mathbf{x}}^{(t)} &= \dot{\alpha}^{(t)} \frac{\nabla \phi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}, \\ \mathbf{z}^{(t_0)} = 0,\ \mathbf{x}^{(t_0)} &= \mathbf{x}^{(0)},\ \text{for arbitrary initial point } \mathbf{x}^{(0)} \in X. \end{aligned} \tag{CT-AMD}$$

Lemma 3.2. Let $\mathbf{x}^{(t)}, \mathbf{z}^{(t)}$ evolve according to (CT-AMD), for some convex function $f : X \to \mathbb{R}$. Then, $\forall t > t_0$:

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{\alpha^{(t_0)}(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)) + \phi(\mathbf{x}^*)}{\alpha^{(t)}}.$$

3.3 Accelerated Strongly Convex Minimization

We can also get a similar accelerated dynamics when the function is, in addition, strongly convex. In that case, we use $U^{(t)} = f(\mathbf{x}^{(t)})$ and the lower bound from (2.4). Let $\phi_t(\mathbf{x}) = \int_{t_0}^{t} \frac{\sigma}{2} \|\mathbf{x} - \mathbf{x}^{(\tau)}\|^2 + \phi(\mathbf{x})$. Observe that $\frac{d}{dt}\phi^{(t)}(\mathbf{x}) \geq 0$, $\forall \mathbf{x} \in X$. Then, we have the following result for the change in the gap:

$$\begin{aligned} \frac{d}{dt}(\alpha^{(t)} G^{(t)}) =& \frac{d}{dt}(\alpha^{(t)} f(\mathbf{x}^{(t)})) - \dot{\alpha}^{(t)} \left(f(\mathbf{x}^{(t)}) + \left\langle \nabla f(\mathbf{x}^{(t)}), \nabla \phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)} \right\rangle \right) - \frac{d}{d\tau} \left. (\phi_\tau(\nabla \phi_t^*(\mathbf{z}^{(t)})))\right|_{\tau=t} \\ \leq& \left\langle \nabla f(\mathbf{x}^{(t)}), \alpha^{(t)} \dot{\mathbf{x}}^{(t)} - \dot{\alpha}^{(t)} (\nabla \phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}) \right\rangle. \end{aligned}$$

Therefore, choosing $\dot{\mathbf{x}}^{(t)} = \dot{\alpha}^{(t)} \frac{\nabla \phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}$ gives $\frac{d}{dt}(\alpha^{(t)} G^{(t)}) \leq 0$, and the convergence result stated as Lemma 3.3 below follows.

$$\begin{aligned} \dot{\mathbf{z}}^{(t)} &= -\dot{\alpha}^{(t)} \nabla f(\mathbf{x}^{(t)}), \\ \dot{\mathbf{x}}^{(t)} &= \dot{\alpha}^{(t)} \frac{\nabla \phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}, \\ \mathbf{z}^{(t_0)} = 0,\ \mathbf{x}^{(t_0)} &= \mathbf{x}^{(0)} \text{ for arbitrary initial point } \mathbf{x}^{(0)} \in X. \end{aligned} \tag{CT-ASC}$$

Lemma 3.3. Let $\mathbf{x}^{(t)}$ evolve according to (CT-ASC), for an arbitrary initial point $\mathbf{x}^{(0)} \in X$ and $\phi_t(\mathbf{x}) = \int_{t_0}^{t} \frac{\sigma}{2} \|\mathbf{x} - \mathbf{x}^{(\tau)}\|^2 + \phi(\mathbf{x})$. Then, $\forall t > t_0$:

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{\alpha^{(t_0)}(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)) + \phi(\mathbf{x}^*)}{\alpha^{(t)}}.$$

We note that, while there is no difference in the algorithm or in the convergence bound for (CT-AMD) and (CT-ASC) in the continuous-time domain, in the discrete time these two algorithms will lead to very different convergence bounds, due to the different discretization errors they incur.

3.4 Composite Mirror Descent

Now assume that the objective is composite: $\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x})$, where $f(\mathbf{x})$ is convex and $\nabla \phi_t^*(\cdot)$ is easily computable, for $\phi_t(\mathbf{x}) = A^{(t)}\psi(\mathbf{x}) + \phi(\mathbf{x})$. Then, we can use the lower bound for composite functions (2.5). Let upper bound be:

$$U^{(t)} = \frac{1}{\alpha^{(t)}} \int_{t_0}^{t} \bar{f}(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)} + \frac{\alpha^{(t_0)}}{\alpha^{(t)}} \bar{f}(\mathbf{x}^{(0)}) = \frac{1}{\alpha^{(t)}} \int_{t_0}^{t} (f(\mathbf{x}^{(\tau)}) + \psi(\mathbf{x}^{(\tau)})) d\alpha^{(\tau)} + \frac{\alpha^{(t_0)}}{\alpha^{(t)}} (f(\mathbf{x}^{(0)}) + \psi(\mathbf{x}^{(0)})).$$

Then, the change in the gap is:

$$\frac{d}{dt}(\alpha^{(t)} G^{(t)}) = \dot{\alpha}^{(t)} \psi(\mathbf{x}^{(t)}) - \dot{\alpha}^{(t)} \left\langle \nabla f(\mathbf{x}^{(t)}), \nabla \phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)} \right\rangle - \dot{\alpha}^{(t)} \psi(\nabla \phi_t^*(\mathbf{z}^{(t)})). \tag{3.2}$$

Thus, when $\dot{\mathbf{x}}^{(t)} = \nabla\phi_t^*(\mathbf{z}^{(t)})$, where $\phi_t(\cdot) = \alpha^{(t)}\psi(\cdot)$, we have $\frac{d}{dt}(\alpha^{(t)}G^{(t)}) = 0$, and Lemma 3.4 follows immediately. The algorithm can be thought of as mirror descent for composite minimization.

$$\begin{aligned}\dot{\mathbf{z}}^{(t)} &= -\dot{\alpha}^{(t)}\nabla f(\mathbf{x}^{(t)}),\\ \mathbf{x}^{(t)} &= \nabla\phi_t^*(\mathbf{z}^{(t)}),\\ \hat{\mathbf{x}}^{(t)} &= \dot{\alpha}^{(t)}\frac{\mathbf{x}^{(t)} - \hat{\mathbf{x}}^{(t)}}{\alpha^{(t)}},\\ \mathbf{z}^{(t_0)} = 0,\ \hat{\mathbf{x}}^{(t_0)} &= \mathbf{x}^{(0)},\ \text{for arbitrary initial point } \mathbf{x}^{(0)} \in X.\end{aligned} \tag{CT-CMD}$$

Lemma 3.4. Let $\mathbf{x}^{(t)}, \hat{\mathbf{x}}^{(t)}$ evolve according to (CT-CMD), for some convex composite function $\bar{f} = f + \psi : X \to \mathbb{R}$. Then, $\forall t > t_0$:

$$\bar{f}(\hat{\mathbf{x}}^{(t)}) - \bar{f}(\mathbf{x}^*) \leq \frac{\alpha^{(t_0)}(\bar{f}(\mathbf{x}^{(0)}) - \bar{f}(\mathbf{x}^*)) + \phi(\mathbf{x}^*)}{\alpha^{(t)}}.$$

3.5 Accelerated Composite Minimization

Similar as before, we now consider an upper bound $U^{(t)} = \bar{f}(\mathbf{x}^{(t)})$, while the lower bound is given by (2.5). Let $\phi_t(\cdot) = A^{(t)}\psi(\cdot) + \phi(\cdot)$. Then, the change in the gap is given as:

$$\frac{d}{dt}(\alpha^{(t)}G^{(t)}) = \frac{d}{dt}(\alpha^{(t)}(f(\mathbf{x}^{(t)}) + \psi(\mathbf{x}^{(t)}))) - \dot{\alpha}^{(t)}\left(f(\mathbf{x}^{(t)}) + \left\langle \nabla f(\mathbf{x}^{(t)}), \nabla\phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}\right\rangle\right) - \dot{\alpha}^{(t)}\psi(\nabla\phi_t^*(\mathbf{z}^{(t)})).$$

When $\dot{\mathbf{x}}^{(t)} = \dot{\alpha}^{(t)}\frac{\nabla\phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}$, then, by the same arguments as in Section 3.2,

$$\frac{d}{dt}(\alpha^{(t)}f(\mathbf{x}^{(t)})) - \dot{\alpha}^{(t)}\left(f(\mathbf{x}^{(t)}) + \left\langle \nabla f(\mathbf{x}^{(t)}), \nabla\phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}\right\rangle\right) = 0.$$

To conclude that $\frac{d}{dt}(\alpha^{(t)}G^{(t)}) \leq 0$, it remains to show that for this choice of $\mathbf{x}^{(t)}$ also $\frac{d}{dt}(\alpha^{(t)}\psi(\mathbf{x}^{(t)})) - \dot{\alpha}^{(t)}\psi(\nabla\phi_t^*(\mathbf{z}^{(t)})) \leq 0$. But this follows from Jensen's inequality, as $\mathbf{x}^{(t)} = \frac{1}{\alpha^{(t)}}\int_{t_0}^t \nabla\phi_\tau^*(\mathbf{z}^{(\tau)})d\alpha^{(\tau)} + \frac{\alpha^{(t_0)}}{\alpha^{(t)}}\mathbf{x}^{(0)}$, and we immediately obtain the convergence result stated as Lemma 3.5, and recover a continuous-time version of the Fast Gradient Method from [15].

$$\begin{aligned}\dot{\mathbf{z}}^{(t)} &= -\dot{\alpha}^{(t)}\nabla f(\mathbf{x}^{(t)}),\\ \dot{\mathbf{x}}^{(t)} &= \dot{\alpha}^{(t)}\frac{\nabla\phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}},\\ \mathbf{z}^{(t_0)} = 0,\ \mathbf{x}^{(t_0)} &= \mathbf{x}^{(0)},\ \text{for arbitrary initial point } \mathbf{x}^{(0)} \in X.\end{aligned} \tag{CT-ACMD}$$

Lemma 3.5. Let $\mathbf{x}^{(t)}, \mathbf{z}^{(t)}$ evolve according to (CT-ACMD), for some convex composite function $\bar{f} = f + \psi : X \to \mathbb{R}$. Then, $\forall t > t_0$:

$$\bar{f}(\hat{\mathbf{x}}^{(t)}) - \bar{f}(\mathbf{x}^*) \leq \frac{\alpha^{(t_0)}(\bar{f}(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)) + \phi(\mathbf{x}^*)}{\alpha^{(t)}}.$$

3.6 Frank-Wolfe Method

Let $\phi_t(\mathbf{x}) = A^{(t)}\psi(\mathbf{x}) + \phi(\mathbf{x})$, where $\phi(\mathbf{x}) = 0$. We now show that it is possible to replace the map $\nabla\phi_t^*(\mathbf{z}^{(t)})$ (the mirror map of the aggregated gradient) by $\nabla\psi^*(\hat{\mathbf{z}}^{(t)})$, where $\hat{\mathbf{z}}^{(t)} = -\nabla f(\mathbf{x}^{(t)})$ (the mirror map at the last seen gradient) in the lower bound for composite functions. When $\psi(\cdot)$ is the indicator function of the feasible set, the method reduces to the classical Frank-Wolfe method. Observe that, as $\phi_t(\cdot) = A^{(t)}\psi(\cdot)$, we have that $\nabla\phi_t^*(\mathbf{z}^{(t)}) = \nabla\psi^*(\mathbf{z}^{(t)}/A^{(t)})$. By optimality of $\nabla\psi^*(\hat{\mathbf{z}}^{(t)})$ (Fact 1.6):

$$\begin{aligned}\left\langle \nabla f(\mathbf{x}^{(t)}), \nabla\phi_t^*(\mathbf{z}^{(t)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(t)})\right\rangle - \psi(\nabla\phi_t^*(\mathbf{z}^{(t)})) &= -\left\langle \hat{\mathbf{z}}^{(t)}, \nabla\psi^*(\mathbf{z}^{(t)}/A^{(t)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(t)})\right\rangle - \psi(\nabla\psi^*(\mathbf{z}^{(t)}/A^{(t)}))\\ &\leq -\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(t)})).\end{aligned} \tag{3.3}$$

Combining (3.2) and (3.3) and choosing $U^{(t)} = f(\mathbf{x}^{(t)}) + \psi(\mathbf{x}^{(t)})$ as the upper bound:

$$\frac{d}{dt}(\alpha^{(t)}G^{(t)}) \leq \frac{d}{dt}(\alpha^{(t)}(f(\mathbf{x}^{(t)}) + \psi(\mathbf{x}^{(t)}))) - \dot{\alpha}^{(t)}\left(f(\mathbf{x}^{(t)})\left\langle \nabla f(\mathbf{x}^{(t)}), \nabla\psi^*(\hat{\mathbf{z}}^{(t)}) - \mathbf{x}^{(t)}\right\rangle\right) - \dot{\alpha}^{(t)}\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(t)})).$$

Thus, when $\dot{\mathbf{x}}^{(t)} = \dot{\alpha}^{(t)} \frac{\nabla\psi^*(\hat{\mathbf{z}}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}$, we have that

$$\frac{d}{dt}(\alpha^{(t)} G^{(t)}) \leq \frac{d}{dt}(\alpha^{(t)}\psi(\mathbf{x}^{(t)})) - \dot{\alpha}^{(t)}\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(t)})),$$

which is non-positive by Jensen's inequality, as $\mathbf{x}^{(t)} = \frac{\alpha^{(t_0)}}{\alpha^{(t)}}\mathbf{x}^{(0)} + \frac{1}{\alpha^{(t)}}\int_{t_0}^{t} \nabla\psi^*(\hat{\mathbf{z}}^{(\tau)}) d\alpha^{(\tau)}$. Thus, we recover the continuous-time version of the Frank-Wolfe algorithm [14], and the convergence result stated as Lemma 3.6 follows.

$$\begin{aligned} \hat{\mathbf{z}}^{(t)} &= -\nabla f(\mathbf{x}^{(t)}), \\ \dot{\mathbf{x}}^{(t)} &= \dot{\alpha}^{(t)} \frac{\nabla\psi^*(\hat{\mathbf{z}}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}, \\ \mathbf{x}^{(t_0)} &= \mathbf{x}^{(0)}, \text{ for arbitrary } \mathbf{x}^{(0)} \in X. \end{aligned} \tag{CT-FW}$$

Lemma 3.6. Let $\mathbf{x}^{(t)}, \hat{\mathbf{x}}^{(t)}$ evolve according to (CT-FW), for some convex composite function $\bar{f} = f + \psi : X \to \mathbb{R}$. Then, $\forall t > t_0$:

$$\bar{f}(\hat{\mathbf{x}}^{(t)}) - \bar{f}(\mathbf{x}^*) \leq \frac{\alpha^{(t_0)}(\bar{f}(\mathbf{x}^{(0)}) - \bar{f}(\mathbf{x}^*))}{\alpha^{(t)}}.$$

4 Discretization and Incurred Errors

Suppose now that $\alpha^{(t)}$ is a discrete measure. In particular, let $\alpha^{(t)}$ be a piecewise constant function, with $\alpha^{(t)} = 0$ for $t < t_0$, $\alpha^{(t)}$ constant in intervals $(t_0 + i, t_0 + i + 1)$ for $i \in \mathbb{Z}_+$, and $\alpha^{((t_0+i)+)} - \alpha^{((t_0+i)-)} = a_i$ for some $a_i > 0$ and $i \in \mathbb{Z}_+$. Then $\dot{\alpha}^{(t)}$ is equal to $a_i\delta$ for $t = t_0 + i$, $i \in \mathbb{Z}_+$, where δ is the Dirac Delta function, and is equal to zero elsewhere. In other words, $\dot{\alpha} = \sum_{i=0}^{\infty} \delta(t - (t_0 + i))$ meaning that $\dot{\alpha}$ samples the function under the integral at discrete time points $t_0 + i$ for $i \in \mathbb{Z}_+$. If we interpret all the integrals as starting from t_0- and going to $t+$, then the same upper and lower bounds as presented before are still valid with $A^{(t)} = \alpha^{(t)}$.

For the continuous-time algorithms (and their analysis) presented in Section 3, there are generally two causes of the discretization error: (i) different integration rules applying to continuous and discrete measures, and (ii) discontinuities in the algorithm updates. We discuss these two causes in more details below.

Integration errors. To understand where the integration errors occur, we first note that such errors cannot occur in integrals whose sole purpose is weighted averaging, since for these integrals there is no functional difference in the continuous- and discrete-time domains. Thus, the only place where the integration errors can occur is in the integral appearing under the minimum in the lower bound. In $\alpha^{(t)}G^{(t)} = A^{(t)}G^{(t)}$, the integral appears as:

$$I^{(t_0,t)} = -\int_{t_0}^{t} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \nabla\phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(\tau)} \right\rangle d\alpha^{(\tau)},$$

where $\phi_t(\cdot) = \phi(\cdot)$ in the case of mirror descent and accelerated convex minimization. Let $I_c^{(t_0,t)}$ denote the value of $I^{(t_0,t)}$ when α is a continuous measure. To simplify the notation, let (i) in the superscript denote the value of the corresponding entity at time $(t_0 + i)$, for $i \in \mathbb{Z}_+$. Then, for $i \geq 1$, the integration error is $I^{(i-1,i)} - I_c^{(i-1,i)}$.

Observe that, as between times $i - 1$ and i $\dot{\alpha}^{(\tau)}$ samples the function under the integral at time i, we have:

$$I^{(i-1,i)} = -\left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi_i(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle. \tag{4.1}$$

Discontinuities in the algorithm updates. In all the described algorithms, the updates for $\mathbf{x}^{(t)}$ (and possibly $\hat{\mathbf{x}}^{(t)}$) depend on $\nabla\phi_t^*(\mathbf{z}^{(t)})$. Recall that $\mathbf{z}^{(t)}$ aggregates gradients up to time t and thus also depends on $\nabla f(\mathbf{x}^{(t)})$. In the continuous time, this is not a problem, since the updates in $\mathbf{x}^{(t)}$ can follow updates in $\mathbf{z}^{(t)}$ with an arbitrarily small delay, meaning that in the limit we can take that $\mathbf{x}^{(t)}$ changes simultaneously with $\mathbf{z}^{(t)}$. In the discrete time, though, the delay between the two updates cannot be neglected, and solving $\mathbf{x}^{(t)} = g(\nabla\phi_t^*(\mathbf{z}^{(t)}))$ for some function $g(\cdot)$ is in general either not possible or requires many fixed-point iterations.

Apart from affecting the value of $I^{(i-1,i)}$ described above, the discontinuities will also contribute additional discretization error in the case of composite minimization. The reason for the additional discretization error is that the analysis of the gap reduction relies on bounding the change in $\psi(\mathbf{x}^{(t)})$ (or the $\dot{\alpha}^{(\tau)}$-weighted average

of $\psi(\mathbf{x}^{(\tau)})$'s for $\tau \in [t_0, t]$) from the upper bound by $\dot{\alpha}^{(t)}\psi(\nabla\phi_t^*(\mathbf{z}^{(t)}))$ from the lower bound. For composite mirror descent, this discretization error at time i will amount to $a_i\left(\psi(\mathbf{x}^{(i)}) - \psi(\nabla\phi_i^*(\mathbf{z}^{(i)}))\right)$, while for accelerated composite mirror descent, the error will be $A^{(i)}\psi(\mathbf{x}^{(i)}) - A^{(i-1)}\psi(\mathbf{x}^{(i-1)}) - a_i\psi(\nabla\phi_t^*(\mathbf{z}^{(i)}))$. Similar to composite mirror descent, Frank-Wolfe will accrue discretization error $a_i\left(\psi(\mathbf{x}^{(i)}) - \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)}))\right)$.

Effect of discretization errors on the gap. Since in continuous time we had that $\frac{d}{dt}(\alpha^{(t)}G^{(t)}) \leq 0$, if the discretization error between discrete time points $i-1$ and i is $E_d^{(i)}$, then $A^{(i)}G^{(i)} - A^{(i-1)}G^{(i-1)} \leq E_d^{(i)}$, and we can conclude that:

$$G^{(k)} \leq \frac{a_0 G^{(0)}}{A^{(k)}} + \frac{\sum_{i=1}^k E_d^{(i)}}{A^{(k)}}. \tag{4.2}$$

We are now ready to bound the discretization errors of the algorithms from Section 3. We note that the versions of mirror descent and mirror prox presented here are in fact "lazy" versions of these methods more similar to Nesterov's dual averaging [12]. Nevertheless, the standard versions of the methods can be obtained without much additional effort, and we choose to present the "lazy" versions because they follow more directly from the discretization.

4.1 Mirror Descent

Recall that in mirror descent, $\phi_i(\cdot) = \phi(\cdot)$. The discretization error can be bounded as follows.

Proposition 4.1. The discretization error for (CT-MD) is:

$$E_d^{(i)} = -a_i\left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}).$$

Proof. In the continuous-time version of the algorithm (CT-MD), $\mathbf{x}^{(\tau)} = \nabla\phi^*(\mathbf{z}^{(\tau)})$, and thus:

$$\begin{aligned} I_c^{(i-1,i)} &= \int_{i-1}^{i}\left\langle \dot{\mathbf{z}}^{(\tau)}, \nabla\phi^*(\mathbf{z}^{(i)}) - \nabla\phi^*(\mathbf{z}^{(\tau)})\right\rangle d\tau \\ &= -\int_{i-1}^{i}\frac{dD_{\phi^*}(\mathbf{z}^{(\tau)}, \mathbf{z}^{(i)})}{d\tau}d\tau = D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \end{aligned} \tag{4.3}$$

Since for non-composite functions $E_d^{(i)} = I^{(i-1,i)} - I_c^{(i-1,i)}$, combining (4.1) and (4.3) completes the proof. □

We now consider two different discretization error methods that lead to discrete-time algorithms known as mirror descent and mirror prox (or extra-gradient method).

Forward Euler Discretization: Mirror Descent. Forward Euler discretization leads to the following algorithm updates:

$$\begin{aligned} \mathbf{z}^{(i)} &= \mathbf{z}^{(i-1)} - a_i\nabla f(\mathbf{x}^{(i)}), \\ \mathbf{x}^{(i)} &= \nabla\phi^*(\mathbf{z}^{(i-1)}), \\ \hat{\mathbf{x}}^{(i)} &= \frac{A^{(i-1)}}{A^{(i)}}\hat{\mathbf{x}}^{(i-1)} + \frac{a_i}{A^{(i)}}\mathbf{x}^{(i)}, \\ \mathbf{z}^{(0)} &= -a_0 f(\mathbf{x}^{(0)}),\ \hat{\mathbf{x}}^{(0)} = \mathbf{x}^{(0)},\ \text{for arbitrary } \mathbf{x}^{(0)} \in X. \end{aligned} \tag{MD}$$

It follows (from Proposition 4.1) that in this case the discretization error is given as:

$$E_d^{(i)} = -a_i\left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \tag{4.4}$$

When $f(\cdot)$ is Lipschitz-continuous, we recover the classical mirror descent convergence result [9]:

Theorem 4.2. *Let $f : X \to \mathbb{R}$ be an L-Lipschitz-continuous convex function, and let $\psi : X \to \mathbb{R}$ be σ-strongly convex for some $\sigma > 0$. Let $\mathbf{x}^{(i)}, \hat{\mathbf{x}}^{(i)}$ evolve according to (MD) for $i \leq k$ and $k \geq 1$. Then, if $a_i = \frac{1}{L}\sqrt{\frac{2\sigma D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)})}{k+1}}$ and $\phi(\cdot) = D_\psi(\cdot, \mathbf{x}^{(0)})$:*

$$f(\hat{\mathbf{x}}^{(k)}) - f(\mathbf{x}^*) \leq \sqrt{\frac{2D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)})}{\sigma}} \cdot \frac{L}{\sqrt{k+1}}.$$

Proof. By Proposition A.2, $D_{\psi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \geq \frac{\sigma}{2}\|\nabla\psi^*(\mathbf{z}^{(i-1)}) - \nabla\psi^*(\mathbf{z}^{(i)})\|^2 = \frac{\sigma}{2}\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\|^2$. As $D_\phi(\cdot,\cdot) = D_\psi(\cdot,\cdot)$ and $f(\cdot)$ is Lipschitz continuous with parameter L, using Cauchy-Schwartz Inequality:

$$E_d^{(i)} \leq a_i L\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\| - \frac{\sigma}{2}\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\|^2 \leq \frac{{a_i}^2 L^2}{2\sigma},$$

where the second inequality follows from $2ab - b^2 \leq -a^2$, $\forall a, b$. Therefore, from (4.2):

$$G^{(k)} \leq \frac{a_0 G^{(0)}}{A^{(k)}} + \frac{L^2}{2\sigma} \cdot \frac{\sum_{i=1}^k {a_i}^2}{A^{(k)}}. \tag{4.5}$$

Similarly, we can bound the initial gap as:

$$\begin{aligned} a_0 G^{(0)} &= -a_0 \left\langle \nabla f(\mathbf{x}^{(0)}), \nabla\phi^*(\mathbf{z}^{(0)}) - \mathbf{x}^{(0)} \right\rangle - \phi(\nabla\phi^*(\mathbf{z}^{(0)})) + \phi(\mathbf{x}^*) \\ &= -a_0 \left\langle \nabla f(\mathbf{x}^{(0)}), \mathbf{x}^{(1)} - \mathbf{x}^{(0)} \right\rangle - D_\psi(\mathbf{x}^{(1)}, \mathbf{x}^{(0)}) + D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)}) \\ &\leq \frac{{a_0}^2 L^2}{2\sigma} + D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)}). \end{aligned} \tag{4.6}$$

Finally, combining (4.5), (4.6), the choice of a_i's, and $f(\hat{\mathbf{x}}^{(k)}) - f(\mathbf{x}^*) \leq G^{(k)}$, the result follows. □

Approximate Backward Euler Discretization: Mirror Prox/Extra-gradient. Observe that if we could set $\mathbf{x}^{(i)} = \nabla\phi^*(\mathbf{z}^{(i)})$ (i.e., if we were using backward Euler discretization for $\mathbf{x}$), then the discretization error would be negative: $E_d^{(i)} = -D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})$. However, backward Euler is only an implicit discretization method, as it involves solving $\mathbf{x}^{(i)} = \nabla\phi^*(\mathbf{z}^{(i-1)} - a_i\nabla f(\mathbf{x}^{(i)}))$. Fortunately, the fact that the discretization error is negative enables an approximate implementation of backward Euler, where only two fixed-point iteration steps are performed. The resulting discrete-time method is known as mirror prox [8] or extra-gradient descent [5].

$$\begin{aligned} \tilde{\mathbf{x}}^{(i-1)} &= \nabla\phi^*(\mathbf{z}^{(i-1)}), \\ \tilde{\mathbf{z}}^{(i-1)} &= \mathbf{z}^{(i-1)} - a_i\nabla f(\tilde{\mathbf{x}}^{(i-1)}), \\ \mathbf{x}^{(i)} &= \nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)}), \\ \mathbf{z}^{(i)} &= \mathbf{z}^{(i-1)} - a_i\nabla f(\mathbf{x}^{(i)}), \\ \hat{\mathbf{x}}^{(i)} &= \frac{A^{(i-1)}}{A^{(i)}}\hat{\mathbf{x}}^{(i)} + \frac{a_i}{A^{(i)}}\mathbf{x}^{(i)}, \\ \mathbf{z}^{(0)} = -a_0\nabla f(\mathbf{x}^{(0)}),&\ \text{and } \mathbf{x}^{(0)} \in X \text{ is an arbitrary initial point.} \end{aligned} \tag{MP}$$

We are now ready to show the following well-known result [8]:

Theorem 4.3. *Let $F : X \to \mathbb{R}^n$ be an L-smooth monotone operator and let $\psi(\cdot)$ be a σ-strongly convex function. Let $\mathbf{x}^{(i)}, \hat{\mathbf{x}}^{(i)}$ evolve according to (MP), where $\nabla f(\cdot)$ is replaced by $F(\cdot)$. If $a_i = \sigma/L$ and $\phi(\cdot) = D_\psi(\cdot, \tilde{\mathbf{x}}^{(0)})$, then $\forall k \geq 1$ and $\forall \mathbf{u} \in X$:*

$$\left\langle F(\mathbf{u}), \hat{\mathbf{x}}^{(k)} - \mathbf{u} \right\rangle \leq \frac{L}{\sigma} \cdot \frac{\max_{\mathbf{x}\in X} D_\psi(\mathbf{x}, \tilde{\mathbf{x}}^{(0)})}{k}.$$

Proof. From (4.1) and similarities between mirror-descent gaps for convex functions and monotone operators, we have that the discretization error is:

$$\begin{aligned} E_d^{(i)} &= -a_i \left\langle F(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \\ &= a_i \left\langle F(\tilde{\mathbf{x}}^{(i-1)}) - F(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle - a_i \left\langle F(\tilde{\mathbf{x}}^{(i-1)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \end{aligned} \tag{4.7}$$

As $a_i F(\tilde{\mathbf{x}}^{(i-1)}) = \mathbf{z}^{(i-1)} - \tilde{\mathbf{z}}^{(i-1)}$ and $\mathbf{x}^{(i)} = \nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)})$, Proposition A.3 implies:

$$\begin{aligned} -a_i \left\langle F(\tilde{\mathbf{x}}^{(i-1)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) &= -D_{\phi^*}(\mathbf{z}^{(i-1)}, \tilde{\mathbf{z}}^{(i-1)}) - D_{\phi^*}(\tilde{\mathbf{z}}^{(i-1)}, \mathbf{z}^{(i)}) \\ \leq -\frac{\sigma}{2}\left(\|\nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)})\|^2 + \|\nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\|^2\right), \end{aligned} \tag{4.8}$$

where the inequality is by Proposition A.2.

On the other hand, by L-smoothness of $F(\cdot)$, using Cauchy-Schwartz Inequality:

$$\begin{aligned} a_i \left\langle F(\tilde{\mathbf{x}}^{(i-1)}) - F(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle &\leq a_i L \|\tilde{\mathbf{x}}^{(i-1)} - \mathbf{x}^{(i)}\| \cdot \|\nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\| \\ &= a_i L \|\nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)})\| \cdot \|\nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\|. \end{aligned} \tag{4.9}$$

As $a_i = \sigma/L$, combining (4.7)-(4.9) with the inequality $2ab - a^2 - b^2 \leq 0$, $\forall a, b$, we get that $E_d^{(i)} \leq 0$, $\forall i$. By (4.2), it follows that $G^{(k)} \leq \frac{a_0 G^{(0)}}{A^{(k)}}$.

To bound the initial gap, we use a slight modification of the gap that starts from $i = 1$ instead of $i = 0$. Observe that we still have $G^{(k)} \leq \frac{a_1 G^{(1)}}{A^{(k)}}$, but now $a_0 = 0$ and, therefore, $A^{(k)} = \frac{\sigma}{L}k$. As $D_\phi(\cdot,\cdot) = D_\psi(\cdot,\cdot)$:

$$\begin{aligned} a_1 G^{(1)} &= -a_1 \left\langle F(\mathbf{x}^{(1)}), \nabla\phi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)} \right\rangle - D_\psi(\nabla\phi^*(\mathbf{z}^{(1)}), \tilde{\mathbf{x}}^{(0)}) + D_\psi(\mathbf{x}^*, \tilde{\mathbf{x}}^{(0)}) \\ &= -a_1 \left\langle F(\mathbf{x}^{(1)}), \nabla\phi^*(\mathbf{z}^{(1)}) - \mathbf{x}^{(1)} \right\rangle - D_\phi(\nabla\phi^*(\mathbf{z}^{(1)}), \tilde{\mathbf{x}}^{(0)}) + D_\psi(\mathbf{x}^*, \tilde{\mathbf{x}}^{(0)}). \end{aligned}$$

Observing that $a_1 F(\tilde{\mathbf{x}}^{(0)}) = \mathbf{z}^{(0)} - \tilde{\mathbf{z}}^{(0)}$, $\mathbf{x}^{(1)} = \nabla\phi^*(\tilde{\mathbf{z}}^{(0)})$, and applying the same arguments as in bounding $E_d^{(i)}$ above, it follows that $a_1 G^{(1)} \leq D_\psi(\mathbf{x}^*, \tilde{\mathbf{x}}^{(0)})$, completing the proof. □

Similarly as before, as convex-concave saddle point problems have the same gap as variational inequalities, it is straightforward to extend Theorem 4.3 to this setting (see Section 3.1 and [8]).

4.2 Accelerated Smooth Minimization

In this and in the following subsection, we will only consider forward Euler discretization of the accelerated dynamics, which corresponds to the Nesterov's accelerated algorithm. Approximate backward Euler discretization using similar ideas as in the proof of convergence of mirror prox from the previous subsection is also possible and leads to the recent accelerated extra-gradient descent (AXGD) algorithm that we presented in [3].

As before, we can bound the discretization error by computing $I^{(i-1,i)}$ to obtain the following result.

Proposition 4.4. The discretization for (CT-AMD) is:

$$\begin{aligned} E_d^{(i)} &= -a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle + A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \\ &\leq \left\langle \nabla f(\mathbf{x}^{(i)}), A^{(i)}\mathbf{x}^{(i)} - A^{(i-1)}\mathbf{x}^{(i-1)} - a_i \nabla\phi^*(\mathbf{z}^{(i)}) \right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \end{aligned}$$

Proof. Recall continuous-time accelerated dynamics (CT-AMD), where $\dot{\mathbf{x}}^{(t)} = \dot{\alpha}^{(t)} \frac{\nabla\phi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}$ and $\phi_i(\cdot) = \phi(\cdot)$. We have:

$$\begin{aligned} I_c^{(i-1,i)} &= -\int_{i-1}^{i} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(\tau)} \right\rangle d\alpha^{(\tau)} \\ &= -\int_{i-1}^{i} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \alpha^{(\tau)}\dot{\mathbf{x}}^{(\tau)} \right\rangle d\tau + \int_{i-1}^{i} \left\langle \dot{\mathbf{z}}^{(\tau)}, \nabla\phi^*(\mathbf{z}^{(i)}) - \nabla\phi^*(\mathbf{z}^{(\tau)}) \right\rangle d\tau. \end{aligned}$$

Integrating by parts, the first integral is $-A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)}))$, while the second integral is (as we have seen in the previous subsection) $D_{\psi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})$. Thus, using (4.1), the discretization error is:

$$\begin{aligned} E_d^{(i)} &= -a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle + A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \\ &\leq \left\langle \nabla f(\mathbf{x}^{(i)}), A^{(i)}\mathbf{x}^{(i)} - A^{(i-1)}\mathbf{x}^{(i-1)} - a_i \nabla\phi^*(\mathbf{z}^{(i)}) \right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}), \end{aligned} \tag{4.10}$$

where we have used $f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})$, by $f(\cdot)$'s convexity. □

Standard forward Euler discretization sets $\mathbf{x}^{(i)} = \frac{A^{(i-1)}}{A^{(i)}}\mathbf{x}^{(i-1)} + \frac{a_i}{A^{(i)}}\nabla\phi^*(\mathbf{z}^{(i-1)})$, which results in the discretization error equal to $D_{\phi^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i-1)})$. We cannot bound such a discretization error, since we are not assuming that

$f(\cdot)$ is Lipschitz-continuous. However, since $f(\cdot)$ is L-smooth, we can introduce an additional gradient step whose role is to cancel out the discretization error by reducing the upper bound. The algorithm then becomes:

$$\begin{aligned}\mathbf{z}^{(i)} &= \mathbf{z}^{(i-1)} - a_i \nabla f(\mathbf{x}^{(i)}),\\ \mathbf{x}^{(i)} &= \frac{A^{(i-1)}}{A^{(i)}}\hat{\mathbf{x}}^{(i-1)} + \frac{a_i}{A^{(i)}}\nabla\phi^*(\mathbf{z}^{(i-1)}),\\ \hat{\mathbf{x}}^{(i)} &= \mathrm{Grad}(\mathbf{x}^{(i)}),\\ \mathbf{z}^{(0)} = -a_0 f(\mathbf{x}^{(0)}),\ \hat{\mathbf{x}}^{(0)} &= \mathrm{Grad}(\mathbf{x}^{(0)}),\ \text{for arbitrary } \mathbf{x}^{(0)} \in X,\end{aligned} \tag{AMD}$$

where

$$\mathrm{Grad}(\mathbf{x}^{(i)}) = \arg\min_{\mathbf{x}\in X}\left\{\left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}-\mathbf{x}^{(i)}\right\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{x}^{(i)}\|^2\right\}. \tag{4.11}$$

The introduced gradient steps only affect the upper bound, changing it from $U^{(i)} = f(\mathbf{x}^{(i)})$ to $U^{(i)} = f(\hat{\mathbf{x}}^{(i)})$. Thus, correcting (4.10) for the change in the upper bound, we get:

$$\begin{aligned}E_d^{(i)} =& -a_i\left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\right\rangle + A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})\\ &+ A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)})) - A^{(i-1)}(f(\hat{\mathbf{x}}^{(i-1)}) - f(\mathbf{x}^{(i-1)}))\\ \le& A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)})) + \left\langle \nabla f(\mathbf{x}^{(i)}), A^{(i)}\mathbf{x}^{(i)} - A^{(i-1)}\hat{\mathbf{x}}^{(i-1)} - a_i\nabla\phi^*(\mathbf{z}^{(i)})\right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})\\ =& A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)})) + a_i\left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}).\end{aligned} \tag{4.12}$$

We are now ready to prove the convergence of Nesterov's algorithm for smooth functions [10]:

Theorem 4.5. *Let $f : X \to \mathbb{R}$ be an L-smooth function, $\psi : X \to \mathbb{R}$ be a σ-strongly convex function, and let $\phi(\cdot) = D_\psi(\cdot, \mathbf{x}^{(0)})$. If $\mathbf{x}^{(t)}, \hat{\mathbf{x}}^{(t)}$ evolve according to (AMD) for $a_i = \frac{\sigma}{L}\frac{i+1}{2}$, then $\forall k \ge 1$:*

$$f(\hat{\mathbf{x}}^{(k)}) - f(\mathbf{x}^*) \le \frac{4L}{\sigma}\cdot\frac{D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)})}{(k+1)(k+2)}.$$

Proof. As $f(\cdot)$ is L-smooth, by the definition of $\hat{\mathbf{x}}^{(i)}$:

$$f(\hat{\mathbf{x}}^{(i)}) \le f(\mathbf{x}^{(i)}) + \min_{\mathbf{x}\in X}\left\{\left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}-\mathbf{x}^{(i)}\right\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{x}^{(i)}\|^2\right\}. \tag{4.13}$$

Since the gradient step was introduced to cancel out the discretization error, intuitively, it is natural to try to cancel out the second two terms from (4.12) (that correspond to the original discretization error (4.10)) by $A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)}))$, the decrease due to the gradient step. A point $\mathbf{x} \in X$ that would charge the gradient term from (4.12) to the gradient term in (4.13) is $\mathbf{x} = \mathbf{x}^{(i)} - \frac{a_i}{A^{(i)}}(\nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})) = \frac{A^{(i-1)}}{A^{(i)}}\hat{\mathbf{x}}^{(i-1)} + \frac{a_i}{A^{(i)}}\nabla\phi^*(\mathbf{z}^{(i)}) \in X$. It follows from (4.13) that:

$$A^{(i)}f(\hat{\mathbf{x}}^{(i)}) \le A^{(i)}f(\mathbf{x}^{(i)}) - a_i\left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\right\rangle + \frac{La_i{}^2}{2A^{(i)}}\|\nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\|^2. \tag{4.14}$$

Recall that, by Proposition A.2, $D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \ge \frac{\sigma}{2}\|\nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\|^2$. Therefore, for the quadratic term in (4.14) to cancel the remaining term in (4.12), $D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})$, it suffices to have $\frac{a_i{}^2}{A^{(i)}} \le \frac{\sigma}{L}$. It is easy to verify that $a_i = \frac{\sigma}{L}\frac{i+1}{2}$ from the theorem statement satisfies $a_i = \frac{\sigma}{L}\frac{i+1}{2}$, and thus it follows that $G^{(k)} \le \frac{a_0 G^{(0)}}{A^{(k)}}$.

It remains to bound the initial gap, while the final bound will follow by simple computation of $A^{(k)}$. We have:

$$\begin{aligned}a_0 G^{(0)} &= a_0(f(\hat{\mathbf{x}}^{(0)}) - f(\mathbf{x}^{(0)})) - a_0\left\langle \nabla f(\mathbf{x}^{(0)}), \nabla\phi^*(\mathbf{z}^{(0)}) - \mathbf{x}^{(0)}\right\rangle - D_\psi(\nabla\phi^*(\mathbf{z}^{(0)}), \mathbf{x}^{(0)}) + D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)})\\ &\le D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)}),\end{aligned}$$

by the same arguments as in bounding the discretization error above. □

4.3 Accelerated Smooth and Strongly Convex Minimization

Recall the accelerated dynamics for σ-strongly convex objectives (CT-ASC). The dynamics is almost the same as (CT-AMD), except that instead of a fixed $\phi_i(\cdot)$, we now have: $\phi_i(\mathbf{x}) = \int_0^i \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^{(\tau)}\|^2 d\alpha^{(\tau)} + \phi(\mathbf{x})$. Observe that for $i \geq j$, $\phi_i(\mathbf{x}) \geq \phi_j(\mathbf{x})$, $\forall \mathbf{x} \in X$. We can compute the discretization error for (CT-ASC) as follows.

Proposition 4.6. The discretization error for (CT-ASC) is:

$$E_d^{(i)} \leq \left\langle \nabla f(\mathbf{x}^{(i)}), A^{(i)}\mathbf{x}^{(i)} - A^{(i-1)} - a_i \nabla \phi_{i-1}^*(\mathbf{z}^{(i)}) \right\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}).$$

Proof. To compute the discretization error, we first need to compute $I_c^{(i-1,i)}$:

$$\begin{aligned} I_c^{(i-1,i)} &= -\int_{i-1}^{i} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \nabla \phi_i^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(\tau)} \right\rangle d\alpha^{(\tau)} \\ &= -\int_{i-1}^{i} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \alpha^{(\tau)} \dot{\mathbf{x}}^{(\tau)} \right\rangle d\tau + \int_{i-1}^{i} \left\langle \dot{\mathbf{z}}^{(\tau)}, \nabla \phi_i^*(\mathbf{z}^{(i)}) \right\rangle d\tau - \int_{i-1}^{i} \left\langle \dot{\mathbf{z}}^{(\tau)}, \nabla \phi_\tau^*(\mathbf{z}^{(\tau)}) \right\rangle d\tau \\ &= -A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) + \left\langle \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \nabla \phi_i^*(\mathbf{z}^{(i)}) \right\rangle - \phi_i^*(\mathbf{z}^{(i)}) + \phi_{i-1}^*(\mathbf{z}^{(i-1)}). \end{aligned} \tag{4.15}$$

Combining (4.1), (4.15), and the fact that $-a_i \nabla f(\mathbf{x}^{(i)}) = \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}$:

$$E_d^{(i)} = A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - \left\langle \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \mathbf{x}^{(i)} \right\rangle + \phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}). \tag{4.16}$$

As $\phi_i(\mathbf{x}) \geq \phi_{i-1}(\mathbf{x})$, $\forall \mathbf{x} \in X$, it follows that also $\phi_i^*(\mathbf{z}) \leq \phi_{i-1}^*(\mathbf{z})$, $\forall \mathbf{z}$. Using the definition of Bregman divergence:

$$\begin{aligned} E_d^{(i)} &\leq A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - \left\langle \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \mathbf{x}^{(i)} - \nabla \phi_{i-1}^*(\mathbf{z}^{(i)}) \right\rangle - D_{\phi_{i-1}^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \\ &\leq \left\langle \nabla f(\mathbf{x}^{(i)}), A^{(i)}\mathbf{x}^{(i)} - A^{(i-1)} - a_i \nabla \phi_{i-1}^*(\mathbf{z}^{(i)}) \right\rangle - D_{\phi_{i-1}^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \end{aligned}$$

□

Comparing the discretization error from Proposition 4.6 with the discretization error (4.10) from previous subsection, we can observe that they take the same form, with the only difference of ϕ^* being replaced by ϕ_{i-1}^*. Thus, introducing a gradient descent step into the discrete algorithm leads to the same changes in the discretization error, and we can use the same arguments to analyze the convergence. The algorithm is given as:

$$\begin{aligned} \mathbf{z}^{(i)} &= \mathbf{z}^{(i-1)} - a_i \nabla f(\mathbf{x}^{(i)}), \\ \mathbf{x}^{(i)} &= \frac{A^{(i-1)}}{A^{(i)}} \hat{\mathbf{x}}^{(i-1)} + \frac{a_i}{A^{(i)}} \nabla \phi_{i-1}^*(\mathbf{z}^{(i-1)}), \\ \hat{\mathbf{x}}^{(i)} &= \text{Grad}(\mathbf{x}^{(i)}), \\ \mathbf{z}^{(0)} = -a_0 f(\mathbf{x}^{(0)}), \hat{\mathbf{x}}^{(0)} &= \text{Grad}(\mathbf{x}^{(0)}), \text{ for arbitrary } \mathbf{x}^{(0)} \in X, \end{aligned} \tag{ASC}$$

while the discretization error for (ASC) becomes:

$$E_d^{(i)} \leq A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)})) + a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla \phi_{i-1}^*(\mathbf{z}^{(i-1)}) - \nabla \phi_{i-1}^*(\mathbf{z}^{(i)}) \right\rangle - D_{\phi_{i-1}^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \tag{4.17}$$

We have the following convergence result:

Theorem 4.7. *Let $f : X \to \mathbb{R}$ be an L-smooth and σ-strongly convex function, $\psi : X \to \mathbb{R}$ be a σ_0-strongly convex function, for $\sigma_0 = L - \sigma$, $\phi(\cdot) = D_\psi(\cdot, \mathbf{x}^{(0)})$, and let $\mathbf{x}^{(i)}, \hat{\mathbf{x}}^{(i)}, \mathbf{z}^{(i)}$ evolve according to (ASC), where $\phi_i(\mathbf{x}) = \sum_{j=0}^{i} a_j \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^{(j)}\|^2 + \phi(\mathbf{x})$, $\forall \mathbf{x} \in X$. If $a_0 = 1$ and $\frac{a_i}{A^{(i)}} = \frac{\sqrt{4\kappa+1}-1}{2\kappa}$, where $\kappa = L/\sigma$ is $f(\cdot)$'s condition number, then:*

$$f(\hat{\mathbf{x}}^{(k)}) - f(\mathbf{x}^*) \leq \left(1 - \frac{\sqrt{4\kappa+1}-1}{2\kappa}\right)^k D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)}).$$

Proof. The proof follows by applying the same arguments as in the proof of Theorem 4.5. To obtain the convergence bound, we observe that $\phi_i(\cdot)$ is σ_i-strongly convex for $\sigma_i = \sigma \sum_{j=0}^{i} a_j + \sigma_0 = A^{(i)}\sigma + \sigma_0$. Thus, we only need to show that $\frac{{a_i}^2}{A^{(i)}} \leq \frac{\sigma_i}{L}$. A sufficient condition is that $\frac{{a_i}^2}{A^{(i)}A^{(i-1)}} \leq \frac{\sigma}{L} = \frac{1}{\kappa}$, which is equivalent to $\frac{{a_i}^2}{(A^{(i)})^2} \leq \frac{1}{\kappa}(1 - \frac{a_i}{A^{(i)}})$. Solving $\frac{{a_i}^2}{(A^{(i)})^2} = \frac{1}{\kappa}(1 - \frac{a_i}{A^{(i)}})$ gives the a_i's from the theorem statement for $i \geq 1$. The choice of $a_0 = 1$, $\sigma_0 = L - \sigma$ ensures $a_0 G^{(0)} \leq \phi(\mathbf{x}^*) = D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)})$. □

Remark 4.8. When $X = \mathbb{R}^n$, it is possible to obtain a tighter convergence bound. Namely, we can recover the standard guarantee $f(\hat{\mathbf{x}}^{(k)}) - f(\mathbf{x}^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)})$ [13]. More details are provided in Appendix B.1.

4.4 Composite Mirror Descent

Consider the forward Euler discretization of (CT-CMD), recovering updates similar to [4][2]:

$$\begin{aligned} \mathbf{z}^{(i)} &= \mathbf{z}^{(i-1)} - a_i \nabla f(\mathbf{x}^{(i)}), \\ \mathbf{x}^{(i)} &= \nabla \phi_i^*(\mathbf{z}^{(i-1)}), \\ \hat{\mathbf{x}}^{(i)} &= \frac{A^{(i-1)}}{A^{(i)}} + \frac{a_i}{A^{(i)}}\mathbf{x}^{(i)}, \\ \mathbf{z}^{(0)} = -a_0 f(\mathbf{x}^{(0)}), \hat{\mathbf{x}}^{(0)} &= \mathbf{x}^{(0)}, \text{ for arbitrary } \mathbf{x}^{(0)} \in X. \end{aligned} \tag{CMD}$$

Unlike in the standard (non-composite) convex minimization, as discussed at the beginning of the section, in the composite case the discretization error needs to take into account an extra term. The additional term appears due to the discontinuous solution updates and $\psi(\cdot)$ in the objective; in the continuous-time case $\mathbf{x}^{(t)} = \nabla \phi_t^*(\mathbf{z}^{(t)})$ and the change in the upper bound term $\int_{t_0}^{t} \psi(\mathbf{x}^{(\tau)}) d\alpha^{(\tau)}$ matches the change in $\psi(\nabla \phi_t^*(\mathbf{z}^{(t)}))$. In the discrete time, however, $\mathbf{x}^{(i)} = \nabla \phi_i^*(\mathbf{z}^{(i-1)})$, and thus the error also includes $a_i(\psi(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) - \psi(\nabla \phi_i^*(\mathbf{z}^{(i)})))$, leading to the following bound on the discretization error.

Proposition 4.9. The discretization error for forward Euler discretization of (CT-CMD) is:

$$E_d^{(i)} \leq D_{\phi_i^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i-1)}).$$

Proof. In continuous-time regime, $\mathbf{x}^{(\tau)} = \nabla \phi_\tau^*(\mathbf{z}^{(\tau)})$, and thus:

$$\begin{aligned} I_c^{(i-1,i)} &= \int_{i-1}^{i} \left\langle \dot{\mathbf{z}}^{(\tau)}, \nabla \phi_i^*(\mathbf{z}^{(i)}) - \nabla \phi_\tau^*(\mathbf{z}^{(\tau)}) \right\rangle d\tau \\ &= \left\langle \nabla \phi_i^*(\mathbf{z}^{(i)}), \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)} \right\rangle - \int_{i-1}^{i} \left\langle \nabla \phi_\tau^*(\mathbf{z}^{(\tau)}), \dot{\mathbf{z}}^{(\tau)} \right\rangle d\tau \\ &= \left\langle \nabla \phi_i^*(\mathbf{z}^{(i)}), \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)} \right\rangle - \int_{i-1}^{i} \left(\frac{d}{d\tau} \phi_\tau^*(\mathbf{z}^{(\tau)}) - \frac{d}{ds}\, \phi_s^*(\mathbf{z}^{(\tau)})\Big|_{s=\tau} \right) d\tau \\ &= \left\langle \nabla \phi_i^*(\mathbf{z}^{(i)}), \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)} \right\rangle - \phi_i^*(\mathbf{z}^{(i)}) + \phi_{i-1}^*(\mathbf{z}^{(i-1)}) + \int_{i-1}^{i} \frac{d}{ds}\, \phi_s^*(\mathbf{z}^{(\tau)})\Big|_{s=\tau} d\tau. \end{aligned}$$

Recalling that $\phi_t(\mathbf{x}) = \alpha^{(t)}\psi(\mathbf{x}) + \phi(\mathbf{x})$ and using Danskin's theorem, we have:

$$\begin{aligned} \int_{i-1}^{i} \frac{d}{ds}\, \phi_s^*(\mathbf{z}^{(\tau)})\Big|_{s=\tau} d\tau &= \int_{i-1}^{i} \frac{d}{ds} \max_{\mathbf{x} \in X} \left\{ \left\langle \mathbf{z}^{(\tau)}, \mathbf{x} \right\rangle - \alpha^{(s)}\psi(\mathbf{x}) - \phi(\mathbf{x}) \right\}\Big|_{s=\tau} d\tau \\ &= -\int_{i-1}^{i} \dot{\alpha}^{(\tau)} \psi(\nabla \phi_\tau^*(\mathbf{z}^{(\tau)})) d\tau = -a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i)})). \end{aligned}$$

Therefore:

$$I_c^{(i-1,i)} = \left\langle \nabla \phi_i^*(\mathbf{z}^{(i)}), \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)} \right\rangle - \phi_i^*(\mathbf{z}^{(i)}) + \phi_{i-1}^*(\mathbf{z}^{(i-1)}) - a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i)})).$$

[2]The updates from (CMD) are equivalent to the updates from COMID algorithm in [4] when $\nabla \phi_i^*(\nabla \phi_i(\mathbf{x})) = \mathbf{x}$, which is true in e.g., the unconstrained case – $X = \mathbb{R}^n$.

On the other hand, as:

$$I^{(i-1,i)} = -a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla \phi_i^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle = \left\langle \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \nabla \phi_i^*(\mathbf{z}^{(i)}) - \nabla \phi_i^*(\mathbf{z}^{(i-1)}) \right\rangle,$$

the discretization error is:

$$\begin{aligned} E_d^{(i)} &= \phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) - \left\langle \nabla \phi_i^*(\mathbf{z}^{(i-1)}), \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)} \right\rangle + a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) \\ &= D_{\phi_i^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i-1)}) + \phi_i^*(\mathbf{z}^{(i-1)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) + a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i-1)})). \end{aligned}$$

It remains to show that $\phi_i^*(\mathbf{z}^{(i-1)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) + a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) \leq 0$. Observing that $a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) = \phi_i(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) - \phi_{i-1}(\nabla \phi_i^*(\mathbf{z}^{(i-1)}))$ and using the definition of a convex conjugate together with Fact 1.6:

$$\begin{aligned} &\phi_i^*(\mathbf{z}^{(i-1)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) + a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) \\ &= \phi_i^*(\mathbf{z}^{(i-1)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) + \phi_i(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) - \phi_{i-1}(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) \\ &= \left\langle \mathbf{z}^{(i-1)}, \nabla \phi_i^*(\mathbf{z}^{(i-1)}) - \nabla \phi_{i-1}^*(\mathbf{z}^{(i-1)}) \right\rangle + \phi_{i-1}(\nabla \phi_{i-1}^*(\mathbf{z}^{(i-1)})) - \phi_{i-1}(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) \\ &\leq 0, \end{aligned}$$

where the inequality follows from $\nabla \phi_{i-1}(\mathbf{z}^{(i-1)}) = \arg\min_{\mathbf{x} \in X} \{ - \left\langle \mathbf{z}^{(i-1)}, \mathbf{x} \right\rangle + \phi_{i-1}(\mathbf{x}) \}$, by Fact 1.6. □

Finally, we can obtain the following convergence result for the composite functions, similar to the classical case of mirror descent.

Theorem 4.10. *Let $\bar{f} = f + \psi : X \to \mathbb{R}$ be a composite function, such that $f(\cdot)$ is L-Lipschitz-continuous and convex, and $\psi(\cdot)$ is "simple" and convex. Here, "simple" means that $\nabla \phi_i^*(\mathbf{z})$ is easily computable for $\phi_i(\cdot) = A^{(i)} \psi(\cdot) + D_\phi(\cdot, \mathbf{x}^{(0)})$ and some σ-strongly convex $\phi(\cdot)$ where $\sigma > 0$. Fix any $k \geq 1$ and let $\mathbf{x}^{(i)}, \hat{\mathbf{x}}^{(i)}$ evolve according to (CMD) for $a_i = \frac{1}{L} \sqrt{\frac{2\sigma \phi(\mathbf{x}^*)}{k+1}}$. Then:*

$$\bar{f}(\hat{\mathbf{x}}^{(k)}) - \bar{f}(\mathbf{x}^*) \leq \sqrt{\frac{2 D_\phi(\mathbf{x}^*, \mathbf{x}^{(0)})}{\sigma}} \frac{L}{\sqrt{k+1}}.$$

Proof. Observe that since $\phi(\cdot)$ is σ-strongly convex, $\phi_i(\cdot)$ is also σ-strongly convex. The rest of the proof follows the same argument as the proof of Theorem 4.2 (mirror descent convergence), as $D_{\phi_i^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i-1)}) = -a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i+1)} - \mathbf{x}^{(i)} \right\rangle - D_{\phi_i^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})$, and is omitted. □

Remark 4.11. Observe that if $\psi(\cdot)$ was σ-strongly convex for some $\sigma > 0$, we could have obtained a stronger convergence result, as in that case we would have $D_{\phi_i^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \geq A^{(i)} \frac{\sigma}{2} \|\mathbf{x}^{(i)} - \mathbf{x}^{(i+1)}\|^2$, which would allow choosing larger steps a_i.

4.5 Accelerated Composite Minimization

Consider the following forward Euler discretization of (CT-ACMD):

$$\begin{aligned} \mathbf{z}^{(i)} &= \mathbf{z}^{(i-1)} - a_i \nabla f(\mathbf{x}^{(i)}), \\ \mathbf{x}^{(i)} &= \frac{A^{(i-1)}}{A^{(i)}} \mathbf{x}^{(i-1)} + \frac{a_i}{A^{(i)}} \nabla \psi_i^*(\mathbf{z}^{(i-1)}), \\ \mathbf{z}^{(0)} &= -a_0 f(\mathbf{x}^{(0)}), \mathbf{x}^{(0)} \in X \text{ is an arbitrary point.} \end{aligned} \tag{ACMD-FE}$$

The discretization for (ACMD-FE) can be bounded as in the following proposition, whose proof can be found in Appendix B.2.

Proposition 4.12. Forward Euler discretization (ACMD-FE) of (CT-ACMD) incurs discretization error:

$$E_d^{(i)} \leq a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla \phi_i^*(\mathbf{z}^{(i-1)}) - \nabla \phi_i^*(\mathbf{z}^{(i)}) \right\rangle - D_{\phi_i^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}).$$

Similar as in the case of smooth accelerated (non-composite) minimization, we need to correct the discretization error. Unfortunately, correcting the error is not as simple as in the non-composite case. One can verify that in the presence of a non-smooth component $\psi(\cdot)$ taking a gradient step defined as, e.g.,

$$\mathrm{Grad}(\hat{\mathbf{x}}) = \arg\min_{\mathbf{u}\in X}\left\{f(\hat{\mathbf{x}}) + \langle\nabla f(\hat{\mathbf{x}}), \mathbf{u}-\hat{\mathbf{x}}\rangle + \frac{L}{2}\|\mathbf{u}-\hat{\mathbf{x}}\|^2 + \psi(\mathbf{u})\right\}$$

does not suffice. Instead, we need to resort to a two-step error correction. An extrapolation step (known as the dual extrapolation [11]) that can correct the discretization error is motivated by the following proposition.

Proposition 4.13. $\phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) \leq -\min_{\mathbf{x}\in X}\left\{a_i\left\langle\nabla f(\mathbf{x}^{(i)}), \mathbf{x}\right\rangle + a_i\psi(\mathbf{x}) + D_{\phi_{i-1}}(\mathbf{x}, \nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)}))\right\}$.

Proof. From the definition of $\phi_i^*(\mathbf{z}^{(i)})$:

$$\begin{aligned}
\phi_i^*(\mathbf{z}^{(i)}) &= \max_{\mathbf{x}\in X}\left\{\left\langle\mathbf{z}^{(i)}, \mathbf{x}\right\rangle - \phi_i(\mathbf{x})\right\} \\
&= \max_{\mathbf{x}\in X}\left\{\left\langle\mathbf{z}^{(i)}, \mathbf{x}\right\rangle - A^{(i)}\psi(\mathbf{x}) - \phi(\mathbf{x})\right\} \\
&= \max_{\mathbf{x}\in X}\left\{\left\langle\mathbf{z}^{(i-1)}, \mathbf{x}\right\rangle - A^{(i-1)}\psi(\mathbf{x}) - \phi(\mathbf{x}) + \left\langle\mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \mathbf{x}\right\rangle + a_i\psi(\mathbf{x})\right\}. \qquad (4.18)
\end{aligned}$$

Let $m(\mathbf{x}) = \left\langle\mathbf{z}^{(i-1)}, \mathbf{x}\right\rangle - A^{(i-1)}\psi(\mathbf{x}) - \phi(\mathbf{x})$. By Fact 1.6, $\phi_{i-1}^*(\mathbf{z}^{(i-1)}) = m(\nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)}))$. Moreover, observing that $D_m(\mathbf{x}, \mathbf{y}) = -D_{\phi_i}(\mathbf{x}, \mathbf{y})$, $\forall\mathbf{x}, \mathbf{y}$, we further have:

$$\begin{aligned}
m(\mathbf{x}) &= m(\nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})) + \left\langle\nabla m(\nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})), \mathbf{x} - \nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})\right\rangle + D_m(\mathbf{x}, \nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})) \\
&= \phi_{i-1}^*(\mathbf{z}^{(i-1)}) + \left\langle\mathbf{z}^{(i-1)} - \nabla\phi_{i-1}(\nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})), \mathbf{x} - \nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})\right\rangle - D_{\phi_{i-1}}(\mathbf{x}, \nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})). \qquad (4.19)
\end{aligned}$$

By first-order optimality from Fact 1.6, $\left\langle\mathbf{z}^{(i-1)} - \nabla\phi_{i-1}(\nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})), \mathbf{x} - \nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})\right\rangle \leq 0$. Thus, combining (4.18) and (4.19) and recalling that $\mathbf{z}^{(i-1)} - \mathbf{z}^{(i)} = a_i\nabla f(\mathbf{x}^{(i)})$:

$$\phi_i^*(\mathbf{z}^{(i)}) \leq \phi_{i-1}^*(\mathbf{z}^{(i-1)}) + \max_{\mathbf{x}\in X}\left\{-a_i\left\langle\nabla f(\mathbf{x}^{(i)}), \mathbf{x}\right\rangle - a_i\psi(\mathbf{x}) - D_{\phi_{i-1}}(\mathbf{x}, \nabla\psi_{i-1}^*(\mathbf{z}^{(i-1)}))\right\},$$

implying the proposition. □

The extrapolation step is simply defined as the argument of the minimum from Proposition 4.13. Denoting $\hat{\mathbf{z}}^{(i-1)} = \nabla\phi_{i-1}(\nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})) - a_i\nabla f(\mathbf{x}^{(i)})$, it is not hard to see that the extrapolation step is equal to $\nabla\phi_{i-1}^*(\hat{\mathbf{z}}^{(i-1)})$. The algorithm involving the extrapolation step is given below and is equivalent to the fast gradient method from [15]:

$$\begin{aligned}
\mathbf{z}^{(i)} &= \mathbf{z}^{(i-1)} - a_i\nabla f(\mathbf{x}^{(i)}), \\
\mathbf{x}^{(i)} &= \frac{A^{(i-1)}}{A^{(i)}}\mathbf{y}^{(i-1)} + \frac{a_i}{A^{(i)}}\nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)}), \\
\hat{\mathbf{z}}^{(i-1)} &= \nabla\phi_{i-1}(\nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})) - a_i\nabla f(\mathbf{x}^{(i)}), \qquad \text{(ACMD)} \\
\mathbf{y}^{(i)} &= \frac{A^{(i-1)}}{A^{(i)}}\mathbf{y}^{(i-1)} + \frac{a_i}{A^{(i)}}\nabla\phi_{i-1}^*(\hat{\mathbf{z}}^{(i-1)}), \\
\mathbf{z}^{(0)} &= -a_0 f(\mathbf{x}^{(0)}),\ \mathbf{y}^{(0)} = \nabla\phi_0^*(\mathbf{z}^{(0)}) \text{ for arbitrary } \mathbf{x}^{(0)} \in X.
\end{aligned}$$

Using similar arguments as in Section 4.2, we can obtain the discretization error for (ACMD) as follows.

Lemma 4.14. Let $f(\cdot)$ be L-smooth and $\phi(\cdot)$ be σ-strongly convex. Then, the discretization error for (ACMD) discretization of (CT-ACMD) is:

$$E_d^{(i)} \leq \left(A^{(i)}\frac{L}{2} - \frac{(A^{(i)})^2}{{a_i}^2}\frac{\sigma}{2}\right)\|\mathbf{y}^{(i)} - \mathbf{x}^{(i)}\|^2.$$

The proof can be found in Appendix B.2.

Theorem 4.15. *Let $\bar{f} = f + \psi : X \to \mathbb{R}$ be such that $f(\cdot)$ is L-smooth and $\nabla\phi_i^*(\mathbf{z}^{(i)})$ is easily computable for $\phi_i(\cdot) = A^{(i)}\psi(\cdot) + D_\phi(\cdot, \mathbf{x}^{(0)})$ and some σ-strongly convex $\phi(\cdot)$. Let $\mathbf{x}^{(i)}, \mathbf{y}^{(i)}$ evolve according to (ACMD) for $a_i = \frac{\sigma}{L}\frac{i+1}{2}$. Then, $\forall k \geq 1$:*

$$\bar{f}(\mathbf{y}^{(i)}) - f(\mathbf{x}^*) \leq \frac{4L}{\sigma}\frac{D_\phi(\mathbf{x}^*, \mathbf{x}^{(0)})}{(k+1)(k+2)}.$$

Proof. By the choice of a_i's it is not hard to verify that $\frac{{a_i}^2}{A^{(i)}} \leq \frac{\sigma}{L}$, and thus, by Lemma 4.14, $E_d^{(i)} \leq 0$. Therefore, $G^{(k)} \leq \frac{a_0}{A^{(k)}}G^{(0)}$. It remains to bound $G^{(0)}$:

$$\begin{aligned} a^{(0)}G^{(0)} &= a_0\left(f(\mathbf{y}^{(0)}) - f(\mathbf{x}^{(0)}) - \left\langle \nabla f(\mathbf{x}^{(0)}), \nabla\phi_0^*(\mathbf{z}^{(0)}) - \mathbf{x}^{(0)}\right\rangle\right) - D_{\phi_0}(\nabla\phi_0^*(\mathbf{z}^{(0)}), \mathbf{x}^{(0)}) + D_\phi(\mathbf{x}^*, \mathbf{x}^{(0)}) \\ &\leq D_\phi(\mathbf{x}^*, \mathbf{x}^{(0)}), \end{aligned}$$

as $\mathbf{y}^{(0)} = \nabla\phi_0^*(\mathbf{z}^{(0)})$ and using L-smoothness of f, σ-strong convexity of ϕ_0, and the choice of a_0. □

4.6 Frank-Wolfe Method

For the discretization of continuous-time Frank-Wolfe method (CT-FW), we need to take into account the different gap we obtained by using inequality (3.3). In particular, the integral that accrues a discretization error is $-\int_{i-1}^{i}\left\langle \nabla f(\mathbf{x}^{(\tau)}), \nabla\psi^*(\hat{\mathbf{z}}^{(\tau)}) - \mathbf{x}^{(\tau)}\right\rangle d\alpha^{(\tau)}$, where $\hat{\mathbf{z}}^{(\tau)} = -\nabla f(\mathbf{x}^{(\tau)})$. The forward Euler discretization gives the following algorithm:

$$\begin{aligned} \hat{\mathbf{z}}^{(i)} &= -\nabla f(\mathbf{x}^{(i)}), \\ \mathbf{x}^{(i)} &= \frac{A^{(i-1)}}{A^{(i)}}\mathbf{x}^{(i-1)} + \frac{a_i}{A^{(i)}}\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)}), \\ \mathbf{x}^{(0)} &\in X \text{ is an arbitrary initial point.} \end{aligned} \tag{FW}$$

As discussed before, the discretization error needs to include $a_i(\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})) - \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)})))$ in addition to $I^{(i-1,i)} - I_c^{(i-1,i)}$, and is bounded as follows.

Proposition 4.16. The discretization error for forward Euler discretization of (CT-FW) is:

$$E_d^{(i)} \leq a_i\left\langle \nabla f(\mathbf{x}^{(i)}) - \nabla f(\mathbf{x}^{(i-1)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i-1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\right\rangle.$$

Proof. In the discrete-time case, $I^{(i-1,i)} = -a_i\left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \mathbf{x}^{(i)}\right\rangle$, while in continuous-time case, as $\dot{\mathbf{x}}^{(t)} = \dot{\alpha}^{(t)}\frac{\nabla\psi^*(\hat{\mathbf{z}}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}$ and integrating by parts:

$$I_c^{(i-1,i)} = -\int_{i-1}^{i}\left\langle \nabla f(\mathbf{x}^{(\tau)}), \alpha^{(\tau)}\dot{\mathbf{x}}^{(\tau)}\right\rangle d\tau = -A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})).$$

Therefore:

$$\begin{aligned} E_d^{(i)} &= A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - a_i\left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \mathbf{x}^{(i)}\right\rangle + a_i(\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})) - \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)}))) \\ &\leq a_i\left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i-1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\right\rangle + a_i(\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})) - \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)}))), \end{aligned} \tag{4.20}$$

where we have used convexity of $f(\cdot)$ and $\mathbf{x}^{(i)} = \frac{A^{(i-1)}}{A^{(i)}} + \frac{a_i}{A^{(i)}}\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})$. Further, by Fact 1.6,

$$\begin{aligned} &-\left\langle \hat{\mathbf{z}}^{(i-1)}, \nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})\right\rangle + \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})) \leq -\left\langle \hat{\mathbf{z}}^{(i-1)}, \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\right\rangle + \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)})) \\ \Leftrightarrow\quad &\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})) - \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)})) \leq \left\langle \nabla f(\mathbf{x}^{(i-1)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})\right\rangle, \end{aligned} \tag{4.21}$$

where we have used $\hat{\mathbf{z}}^{(i-1)} = -\nabla f(\mathbf{x}^{(i-1)})$.

Combining (4.20) and (4.21), the claimed bound on discretization error follows. □

We can now recover the convergence result from [14].

Theorem 4.17. *Let $\bar{f} = f + \psi : X \to \mathbb{R}$ be a composite function, where $\psi(\cdot)$ is convex and $f(\cdot)$ is convex with Hölder-continuous gradients, i.e., for some fixed $L_\nu < \infty$, $\nu \in (0,1]$[3]:*

$$\|\nabla f(\mathbf{x}) - \nabla f(\hat{\mathbf{x}})\|_* \leq L_\nu \|\mathbf{x} - \hat{\mathbf{x}}\|^\nu, \ \forall \mathbf{x}, \hat{\mathbf{x}} \in X. \tag{4.22}$$

Let $D \stackrel{\text{def}}{=} \max_{\mathbf{x},\hat{\mathbf{x}} \in X} \|\mathbf{x} - \hat{\mathbf{x}}\|$ denote the diameter of X. If $\mathbf{x}^{(i)}$ evolves according to (FW), then, $\forall k \geq 1$:

$$\bar{f}(\mathbf{x}^{(k)}) - \bar{f}(\mathbf{x}^*) \leq L_\nu D^{1+\nu} \frac{\sum_{i=0}^{k} \frac{{a_i}^{1+\nu}}{(A^{(i)})^\nu}}{A^{(k)}}.$$

In particular, if $a_i = i + 1$, then

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

Proof. Applying Cauchy-Schwartz Inequality to the discretization error given by Proposition 4.16, we have:

$$\begin{aligned} E_d^{(i)} &\leq a_i \|\nabla f(\mathbf{x}^{(i)}) - \nabla f(\mathbf{x}^{(i-1)})\|_* \cdot \|\nabla \psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla \psi^*(\hat{\mathbf{z}}^{(i-1)})\| \\ &\leq \frac{{a_i}^{1+\nu}}{(A^{(i)})^\nu} L_\nu \|\mathbf{x}^{(i-1)} - \nabla \psi^*(\hat{\mathbf{z}}^{(i-1)})\|^\nu \cdot \|\nabla \psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla \psi^*(\hat{\mathbf{z}}^{(i-1)})\| \\ &\leq \frac{{a_i}^{1+\nu}}{(A^{(i)})^\nu} L_\nu D^{1+\nu}, \end{aligned}$$

where the second inequality follows from (4.22) and $\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)} = \frac{a_i}{A^{(i)}}(\mathbf{x}^{(i-1)} - \nabla \psi^*(\hat{\mathbf{z}}^{(i-1)}))$ (by (FW)). Therefore, it follows that $G^{(k)} \leq \frac{a_0 G^{(0)}}{A^{(k)}} + L_\nu D^{1+\nu} \frac{\sum_{i=1}^{k} \frac{{a_i}^{1+\nu}}{(A^{(i)})^\nu}}{A^{(k)}}$. We can use the same arguments as for bounding the discretization error to bound $G^{(0)}$. As $\mathbf{x}^{(0)}$ can be mapped to $\nabla \psi^*(\hat{\mathbf{z}}^{(-1)})$, for some $\hat{\mathbf{z}}^{(-1)}$, we have:

$$G^{(0)} = -\left\langle \nabla f(\mathbf{x}^{(0)}), \nabla \psi^*(\hat{\mathbf{z}}^{(0)}) - \mathbf{x}^{(0)} \right\rangle - \psi(\nabla \psi^*(\hat{\mathbf{z}}^{(0)})) + \psi(\mathbf{x}^{(0)}) \leq L_\nu D^{1+\nu}.$$

Therefore, $G^{(k)} \leq L_\nu D^{1+\nu} \frac{\sum_{i=0}^{k} \frac{{a_i}^{1+\nu}}{(A^{(i)})^\nu}}{A^{(k)}}$. In particular, when $a_i = i + 1$, then $A^{(i)} = \frac{(i+1)(i+2)}{2}$. Therefore, $\sum_{i=0}^{k} \frac{{a_i}^{1+\nu}}{(A^{(i)})^\nu} < 2^\nu \sum_{i=0}^{k} (i+1)^{1-\nu} < 2^\nu (k+1)^{2-\nu}$, and the convergence bound follows. □

5 Conclusion

We presented a general technique for the analysis of first-order methods. The technique is intuitive and follows the argument of reducing approximate optimality gap at a certain rate. Besides the unified interpretation of many first-order methods, the technique is generally useful for obtaining new optimization methods and analyzing properties such as noise robustness and model misspecification. More details about the technique's applicability to the analysis of robustness will be provided in a subsequent version of this paper.

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, Y. T. Lee, and M. Singh. A geometric alternative to Nesterov's accelerated gradient descent. 2015.

[3] J. Diakonikolas and L. Orecchia. Accelerated extra-gradient descent: A novel, accelerated first-order method, June 2017. arXiv preprint, arXiv:1706.04680.

[4] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT'10*, 2010.

[5] 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.

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

[3] Observe that when $\nu = 1$, $f(\cdot)$ is L_ν-smooth.

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

[8] 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.

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

[10] 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.

[11] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Prog.*, 109(2):319–344, 2007.

[12] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Math. Prog.*, 120(1):221–259, 2009.

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

[14] Y. Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *CORE Discussion Paper*, 2015.

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

[16] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning*. MIT Press, 2012.

[17] W. Su, S. Boyd, and E. J. Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.*, 17(153):1–43, 2016.

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

[19] 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.

[20] A. C. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

A Properties of the Bregman Divergence

The following properties of Bregman divergence are useful in our analysis.

Proposition A.1. Let $\psi(\cdot)$ be a continuously-differentiable function. Then, $D_{\psi}(\nabla\psi^*(\mathbf{z}), \mathbf{x}) = D_{\psi^*}(\nabla\psi(\mathbf{x}), \mathbf{z})$, $\forall \mathbf{x}, \mathbf{z}$.

Proof. From the definition of ψ^* and Fact 1.6,

$$\psi^*(\mathbf{z}) = \langle \nabla\psi^*(\mathbf{z}), \mathbf{z}\rangle - \psi(\nabla\psi^*),\ \forall \mathbf{z}. \tag{A.1}$$

Similarly, as in the light of Fenchel-Moreau Theorem $\psi^{**} = \psi$[4],

$$\psi(\mathbf{x}) = \langle \nabla\psi(\mathbf{x}), \mathbf{x}\rangle - \psi^*(\nabla\psi(\mathbf{x})),\ \forall \mathbf{x}. \tag{A.2}$$

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

$$\begin{aligned} D_{\psi}(\nabla\psi^*(\mathbf{z}), \mathbf{x}) &= \psi(\nabla\psi^*(\mathbf{z})) - \psi(\mathbf{x}) - \langle \nabla\psi(\mathbf{x}), \nabla\psi^*(\mathbf{z}) - \mathbf{x}\rangle \\ &= \psi(\nabla\psi^*(\mathbf{z})) + \psi^*(\nabla\psi(\mathbf{x})) - \langle \nabla\psi(\mathbf{x}), \nabla\psi^*(\mathbf{z})\rangle\,. \end{aligned} \tag{A.3}$$

Similarly, using the definition of $D_{\psi^*}(\nabla\psi(\mathbf{x}), \mathbf{z})$ combined with (A.1):

$$\begin{aligned} D_{\psi^*}(\nabla\psi(\mathbf{x}), \mathbf{z}) &= \psi^*(\nabla\psi(\mathbf{x})) - \psi^*(\mathbf{z}) - \langle \nabla\psi^*(\mathbf{z}), \nabla\psi(\mathbf{x}) - \mathbf{z}\rangle \\ &= \psi^*(\nabla\psi(\mathbf{x})) + \psi(\nabla\psi^*(\mathbf{z})) - \langle \nabla\psi^*(\mathbf{z}), \nabla\psi(\mathbf{x})\rangle\,. \end{aligned} \tag{A.4}$$

Comparing (A.3) and (A.4), the proof follows. □

Proposition A.2. 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 (A.1), 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 1.6, $\nabla\psi^*(\mathbf{z}) = \arg\max_{\mathbf{x}\in X}\{\langle \mathbf{x}, \mathbf{z}\rangle - \psi(\mathbf{x})\}$, 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 A.3. For all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in X$

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

[4]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.

B Proofs for Section 4 (Discretization and Incurred Errors)

B.1 Smooth and Strongly Convex Unconstrained Minimization

From (4.16), regardless of the particular discretization of (CT-ASC), the discretization error is:

$$E_d^{(i)} = A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) + a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i)} \right\rangle + \phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}).$$

As discussed in Section 4.2, incorporating a gradient step at the end of each iteration effectively adds $A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)})) - A^{(i-1)}(f(\hat{\mathbf{x}}^{(i-1)}) - f(\mathbf{x}^{(i-1)}))$ to $E_d^{(i)}$, where $\hat{\mathbf{x}}^{(i)} = \mathrm{Grad}(\mathbf{x}^{(i)})$, $\forall i$, since adding a gradient step only affects the upper bound. Therefore, the discretization error for a discrete-time method obtained from on (CT-ASC) that adds a gradient step at the end of each iteration is:

$$\begin{aligned} E_d^{(i)} &= A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)})) + A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\hat{\mathbf{x}}^{(i-1)})) + a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i)} \right\rangle + \phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) \\ &\leq A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)})) + \left\langle \nabla f(\mathbf{x}^{(i)}), A^{(i)}\mathbf{x}^{(i)} - A^{(i-1)}\hat{\mathbf{x}}^{(i-1)} \right\rangle + \phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}), \end{aligned}$$

where the inequality follows by convexity of $f(\cdot)$.

What makes the unconstrained case special is that we can set

$$\mathbf{x}^{(i)} = \frac{A^{(i-1)}}{A^{(i)}}\hat{\mathbf{x}}^{(i)} + \frac{a_i}{A^{(i)}}\nabla\phi_i^*(\mathbf{z}^{(i-1)}) \tag{B.1}$$

(compare with $\mathbf{x}^{(i)} = \frac{A^{(i-1)}}{A^{(i)}}\hat{\mathbf{x}}^{(i)} + \frac{a_i}{A^{(i)}}\nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})$ in the constrained case). This is because $\nabla\phi_i^*(\mathbf{z}^{(i-1)}) = \frac{\mathbf{z}^{(i-1)} + \sigma\sum_{j=0}^{i} a_j\mathbf{x}^{(j)} + \sigma_0\mathbf{x}^{(0)}}{\sigma_i}$, where $\sigma_i = A^{(i)}\sigma + \sigma_0$. Therefore, $\mathbf{x}^{(i)}$ can be determined from (B.1) in a closed form. This seemingly minor difference in the discretization (everything else is the same as in the case of (ASC)) allows us to get an improved convergence bound. In particular, for $\phi(\mathbf{x}) = \frac{\sigma_0}{2}\|\mathbf{x}-\mathbf{x}^{(0)}\|^2 = \frac{L-\sigma}{2}\|\mathbf{x}-\mathbf{x}^{(0)}\|^2$, we have that $\phi_i(\cdot)$ is both σ_i-smooth and σ_i-strongly convex. On the other hand, by smoothness of $f(\cdot)$, in the unconstrained case we have $f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)}) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}^{(i)})\|_*^2$. Therefore, we can bound the discretization error as:

$$\begin{aligned} E_d^{(i)} &\leq A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)})) + a_i \left\langle \nabla f(\mathbf{x}^{(i)}), a_i\nabla\phi_i^*(\mathbf{z}^{(i-1)}) \right\rangle + \phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) \\ &\leq -\frac{A^{(i)}}{2L}\|\nabla f(\mathbf{x}^{(i)})\|_*^2 + \phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) - \left\langle \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \nabla\phi_i^*(\mathbf{z}^{(i-1)}) \right\rangle \\ &\leq -\frac{A^{(i)}}{2L}\|\nabla f(\mathbf{x}^{(i)})\|_*^2 + D_{\phi_i^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i-1)}) \\ &= -\frac{A^{(i)}}{2L}\|\nabla f(\mathbf{x}^{(i)})\|_*^2 + D_{\phi_i}(\nabla\phi_i^*(\mathbf{z}^{(i)}), \nabla\phi_i^*(\mathbf{z}^{(i-1)})) \\ &\leq -\frac{A^{(i)}}{2L}\|\nabla f(\mathbf{x}^{(i)})\|_*^2 + \frac{A^{(i)}\sigma}{2}\left\|\frac{a_i\nabla f(\mathbf{x}^{(i)})}{A^{(i)}\sigma}\right\|_*^2, \end{aligned}$$

where we have used $\phi_{i-1}^*(\mathbf{z}^{(i-1)}) \geq \phi_i^*(\mathbf{z}^{(i-1)})$ and $D_{\phi_i^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i-1)}) = D_{\phi_i}(\nabla\phi_i^*(\mathbf{z}^{(i)}), \nabla\phi_i^*(\mathbf{z}^{(i-1)}))$ (which holds in the unconstrained case). Therefore, whenever $\frac{a_i}{A^{(i)}} \leq \sqrt{\frac{\sigma}{L}}$, $E_d^{(i)} \leq 0$. Setting $\frac{a_i}{A^{(i)}} = \sqrt{\frac{\sigma}{L}} = \frac{1}{\sqrt{\kappa}}$, the improved convergence bound $f(\hat{\mathbf{x}}^{(k)}) - f(\mathbf{x}^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \frac{L-\sigma}{2}\|\mathbf{x}^* - \mathbf{x}^{(0)}\|^2$ follows.

B.2 Proofs from Section 4.5 (Accelerated Composite Minimization)

Proposition 4.12. Forward Euler discretization (ACMD-FE) of (CT-ACMD) incurs discretization error:

$$E_d^{(i)} \leq a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi_i^*(\mathbf{z}^{(i-1)}) - \nabla\phi_i^*(\mathbf{z}^{(i)}) \right\rangle - D_{\phi_i^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}).$$

Proof. Recall that in the continuous-time algorithm $\dot{\mathbf{x}}^{(t)} = \dot{\alpha}^{(t)} \frac{\nabla \phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}$. Therefore:

$$\begin{aligned} I_c^{(i-1,i)} &= -\int_{i-1}^{i} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \alpha^{(\tau)} \dot{\mathbf{x}}^{(\tau)} \right\rangle d\tau - \int_{i-1}^{i} \left\langle \nabla f(\mathbf{x}^{(\tau)}), \nabla \phi_i^*(\mathbf{z}^{(i)}) - \nabla \phi_\tau^*(\mathbf{z}^{(\tau)}) \right\rangle d\tau \\ &= -A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) + \int_{i-1}^{i} \left\langle \dot{\mathbf{z}}^{(\tau)}, \nabla \phi_i^*(\mathbf{z}^{(i)}) - \nabla \phi_\tau^*(\mathbf{z}^{(\tau)}) \right\rangle d\tau. \end{aligned}$$

By the same arguments as in the proof of Proposition 4.9,

$$\int_{i-1}^{i} \left\langle \dot{\mathbf{z}}^{(\tau)}, \nabla \phi_i^*(\mathbf{z}^{(i)}) - \nabla \phi_\tau^*(\mathbf{z}^{(\tau)}) \right\rangle d\tau = \left\langle \nabla \phi_i^*(\mathbf{z}^{(i)}), \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)} \right\rangle - \phi_i^*(\mathbf{z}^{(i)}) + \phi_{i-1}^*(\mathbf{z}^{(i-1)}) - a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i)})),$$

and, therefore:

$$I_c^{(i-1,1)} = -A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) + \left\langle \nabla \phi_i^*(\mathbf{z}^{(i)}), \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)} \right\rangle - \phi_i^*(\mathbf{z}^{(i)}) + \phi_{i-1}^*(\mathbf{z}^{(i-1)}) - a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i)})).$$

As $I^{(i-1,i)} = -a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla \phi_i^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle$, we can bound the discretization error as:

$$\begin{aligned} E_d^{(i)} =& A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla \phi_i^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle \\ & - \left\langle \nabla \phi_i^*(\mathbf{z}^{(i)}), \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)} \right\rangle + \phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) + a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i-1)})). \end{aligned} \tag{B.2}$$

By convexity of $f(\cdot)$ and (ACMD-FE):

$$A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla \phi_i^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)} \right\rangle \leq a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla \phi_i^*(\mathbf{z}^{(i-1)}) - \nabla \phi_i^*(\mathbf{z}^{(i)}) \right\rangle. \tag{B.3}$$

On the other hand, by similar arguments as in the proof of Proposition 4.9:

$$\phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) - \left\langle \nabla \phi_i^*(\mathbf{z}^{(i)}), \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)} \right\rangle + a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i-1)})) \leq -D_{\phi_i^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \tag{B.4}$$

Combining (B.2)-(B.4), the proof follows. □

Lemma 4.14. Let $f(\cdot)$ be L-smooth and $\phi(\cdot)$ be σ-strongly convex. Then, the discretization error for (ACMD) discretization of (CT-ACMD) is:

$$E_d^{(i)} \leq \left(A^{(i)} \frac{L}{2} - \frac{(A^{(i)})^2}{{a_i}^2} \frac{\sigma}{2} \right) \|\mathbf{y}^{(i)} - \mathbf{x}^{(i)}\|^2.$$

Proof. The added steps define a new upper bound $U^{(t)} = f(\mathbf{y}^{(t)})$. Accounting for the new upper bound, the discretization error is:

$$\begin{aligned} E_d^{(i)} =& I^{(i-1,i)} - I_c^{(i-1,i)} + A^{(i)} \psi(\mathbf{x}^{(i)})) - A^{(i-1)} \psi(\mathbf{x}^{(i-1)}) - a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i)})) \\ & + A^{(i)}(\bar{f}(\mathbf{y}^{(i)}) - \bar{f}(\mathbf{x}^{(i)})) - A^{(i-1)}(\bar{f}(\mathbf{y}^{(i-1)}) - \bar{f}(\mathbf{x}^{(i-1)})). \end{aligned} \tag{B.5}$$

Combining (B.2) and $I^{(i-1,i)} - I_c^{(i-1,i)}$ from the proof of Proposition 4.13:

$$\begin{aligned} I^{(i-1,i)} - I_c^{(i-1,i)} =& A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) + a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i)} \right\rangle + a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i)})) \\ & - \min_{\mathbf{x} \in X} \left\{ a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x} \right\rangle + a_i \psi(\mathbf{x}) + D_{\phi_{i-1}}(\mathbf{x}, \nabla \phi_{i-1}^*(\mathbf{z}^{(i-1)})) \right\} \\ =& A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla \phi_{i-1}^*(\hat{\mathbf{z}}^{(i-1)}) - \mathbf{x}^{(i)} \right\rangle + a_i \psi(\nabla \phi_i^*(\mathbf{z}^{(i)})) \\ & - a_i \psi(\nabla \phi_i^*(\hat{\mathbf{z}}^{(i-1)})) - D_{\phi_{i-1}}(\nabla \phi_i^*(\hat{\mathbf{z}}^{(i-1)}), \nabla \phi_i^*(\mathbf{z}^{(i-1)})). \end{aligned} \tag{B.6}$$

Now we bound the error correction due to the changes in the upper bound. First, for $f(\cdot)$:

$$\begin{aligned} & A^{(i)}(f(\mathbf{y}^{(i)}) - f(\mathbf{x}^{(i)})) - A^{(i-1)}(f(\mathbf{y}^{(i-1)}) - f(\mathbf{x}^{(i-1)})) \\ &= A^{(i)}(f(\mathbf{y}^{(i)}) - f(\mathbf{x}^{(i)})) - A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) + A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{y}^{(i-1)}) \\ &\leq A^{(i)}(f(\mathbf{y}^{(i)}) - f(\mathbf{x}^{(i)})) - A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) + a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i)} - \nabla \phi_{i-1}^*(\mathbf{z}^{(i-1)}) \right\rangle, \end{aligned} \tag{B.7}$$

where the inequality is by convexity of $f(\cdot)$ and (ACMD). The error correction for $\psi(\cdot)$ is:

$$\begin{aligned}
&A^{(i)}(\psi(\mathbf{y}^{(i)}) - \psi(\mathbf{x}^{(i)})) - A^{(i-1)}(\psi(\mathbf{y}^{(i-1)}) - \psi(\mathbf{x}^{(i-1)})) \\
&= A^{(i)}\psi(\mathbf{y}^{(i)}) - A^{(i-1)}\psi(\mathbf{y}^{(i-1)}) - (A^{(i)}\psi(\mathbf{x}^{(i)}) - A^{(i-1)}\psi(\mathbf{x}^{(i-1)})) \\
&\leq \psi(\nabla\phi_{i-1}(\hat{\mathbf{z}}^{(i-1)})) - (A^{(i)}\psi(\mathbf{x}^{(i)}) - A^{(i-1)}\psi(\mathbf{x}^{(i-1)})),
\end{aligned} \tag{B.8}$$

where we have used Jensen's Inequality and (ACMD). Combining (B.5)-(B.8):

$$E_d^{(i)} \leq A^{(i)}(f(\mathbf{y}^{(i)}) - f(\mathbf{x}^{(i)})) - a_i \left\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi_{i-1}^*(\hat{\mathbf{z}}^{(i-1)}) - \nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)}) \right\rangle - D_{\phi_{i-1}}(\nabla\phi_i^*(\hat{\mathbf{z}}^{(i-1)}), \nabla\phi_i^*(\mathbf{z}^{(i-1)})).$$

From (ACMD), $a_i(\nabla\phi_{i-1}^*(\hat{\mathbf{z}}^{(i-1)}) - \nabla\phi_{i-1}^*(\mathbf{z}^{(i-1)})) = A^{(i)}(\mathbf{y}^{(i)} - \mathbf{x}^{(i)})$. As $f(\cdot)$ is L-smooth and $\phi(\cdot)$ is σ-strongly convex:

$$E_d^{(i)} \leq A^{(i)}\frac{L}{2}\|\mathbf{y}^{(i)} - \mathbf{x}^{(i)}\|^2 - \frac{\sigma}{2}\frac{(A^{(i)})^2}{{a_i}^2}\|\mathbf{y}^{(i)} - \mathbf{x}^{(i)}\|^2,$$

as claimed. □