SIAM J. OPTIM.
Vol. 29, No. 1, pp. 660–689

THE APPROXIMATE DUALITY GAP TECHNIQUE: A UNIFIED THEORY OF FIRST-ORDER METHODS*

JELENA DIAKONIKOLAS† AND LORENZO ORECCHIA‡

Abstract. We present a general technique for the analysis of first-order methods. The technique relies on the construction of a duality gap for an appropriate approximation of the objective function, where the function approximation improves as the algorithm converges. We show that in continuous time the enforcement of an invariant, which corresponds to the approximate duality gap decreasing at a certain rate, exactly recovers a wide range of first-order continuous-time methods. We characterize the discretization errors incurred by different discretization methods, and show how iteration-complexity-optimal methods for various classes of problems cancel out the discretization error. The techniques are illustrated on various classes of problems—including convex minimization for Lipschitz-continuous objectives, smooth convex minimization, composite minimization, smooth and strongly convex minimization, solving variational inequalities with monotone operators, and convex-concave saddle-point optimization—and naturally extend to other settings.

Key words. first-order methods, continuous-time optimization, discretization

AMS subject classifications. 90C06, 90C25, 49N15, 65K05

DOI. 10.1137/18M1172314

1. Introduction.

First-order optimization methods have recently gained in popularity due to their applicability to large-scale problems arising from modern data sets, their relatively low computational complexity, and their potential for parallelizing computation [37]. Moreover, such methods have also been successfully applied in discrete optimization, leading to faster numerical methods [36, 20], graph algorithms [19, 35, 23], and submodular optimization methods [17].

Most first-order optimization methods can be obtained from the discretization of continuous-time dynamical systems that converge to optimal solutions. In the case of mirror descent, the continuous-time view was the original motivation for the algorithm [26], while more recent work has focused on deducing continuous-time interpretations of accelerated methods [40, 41, 22, 38, 34].

Motivated by these works, we provide a unifying theory of first-order methods as discretizations of continuous-time dynamical systems. We term this general framework the *approximate duality gap technique (ADGT)*. In addition to providing an intuitive and unified convergence analysis of various first-order methods that is often only a few lines long, ADGT is also valuable in developing new first-order methods with tight convergence bounds [13, 9], in clarifying interactions between the acceleration and noise [9], and in obtaining width-independent[1] algorithms for problems with

*Received by the editors February 23, 2018; accepted for publication (in revised form) December 4, 2018; published electronically March 5, 2019.
http://www.siam.org/journals/siopt/29-1/M117231.html
Funding: Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing and while the first author was a postdoctoral researcher at Boston University. It was partially supported by NSF grants CCF-1718342 and CCF-1740855, by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF grant CCF-1740425, and by DHS-ALERT subaward 505035-78050.

†Department of Statistics, UC Berkeley, Berkeley, CA 94720 (jelena@jelena-diakonikolas.com).
‡Department of Computer Science, Boston University, Boston, MA 02215 (orecchia@bu.edu).

[1]Width-independent algorithms enjoy poly-logarithmic dependence of their convergence times on the constraints matrix width, i.e., the ratio between the constraint matrix maximum and minimum nonzero elements. By contrast, standard first-order methods incur (at best) linear dependence on the matrix width, which is not even considered to be polynomial-time convergence [28].

positive linear constraints [12, 10]. Further, we have extended ADGT to the setting of block coordinate descent methods [14].

Unlike traditional approaches that start from an algorithm description and then use arguments such as Lyapunov stability criteria to prove convergence bounds [26, 40, 41, 22, 38], our approach takes the opposite direction: *continuous-time methods are obtained from the analysis*, using purely optimization-motivated arguments.

In particular, ADGT can be summarized as follows. Given a convex optimization problem $\min_{\mathbf{x}\in X} f(\mathbf{x})$, to show that a method converges to a minimizer $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in X} f(\mathbf{x})$ at rate $1/\alpha^{(t)}$ (e.g., $\alpha^{(t)} = t$ or $\alpha^{(t)} = t^2$), we need to show that $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq Q/\alpha^{(t)}$, where $\mathbf{x}^{(t)} \in X$ is the solution produced by the method at time t and $Q \in \mathbb{R}_+$ is some bounded quantity that is independent of time t. In general, keeping track of the true optimality gap $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)$ is challenging, as the minimum function value $f(\mathbf{x}^*)$ is typically not known to the method. Instead, the main idea of ADGT is to create an estimate of the optimality gap $G^{(t)}$ that can be easily tracked and controlled and to ensure that $\alpha^{(t)}G^{(t)}$ is a nonincreasing function of time. The estimate corresponds to the difference between an upper bound on $f(\mathbf{x}^{(t)})$, $U^{(t)} \geq f(\mathbf{x}^{(t)})$, and a lower bound on $f(\mathbf{x}^*)$, $L^{(t)} \leq f(\mathbf{x}^*)$, so that $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq U^{(t)} - L^{(t)} = G^{(t)}$. Since ADGT ensures that $\alpha^{(t)}G^{(t)}$ is a nonincreasing function of time, it follows that $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq G^{(t)} \leq \alpha^{(0)}G^{(0)}/\alpha^{(t)}$, which is precisely what we want to show, as long as we ensure that $\alpha^{(0)}G^{(0)}$ is bounded.

To illustrate the power and generality of the technique, we show how to obtain and analyze several well-known first-order methods, such as gradient descent, dual averaging [30], the mirror-prox/extra-gradient method [21, 29, 25], accelerated methods [27, 28], composite minimization methods [16, 31], and Frank–Wolfe methods [32]. 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 exists a 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 the recent books [37, 8, 4, 33].

Multiple approaches to unifying analyses of first-order methods have been developed, with a particular focus on explaining the acceleration phenomenon. Tseng gives a formal framework that unifies all the different instantiations of accelerated gradient methods [39]. Allen-Zhu and Orecchia [1] interpret acceleration as a coupling of mirror-descent and gradient descent steps. Bubeck, Lee, and Singh provide an elegant geometric interpretation of the Euclidean instantiation of Nesterov's method [8]. Drusvyatskiy, Fazel, and Roy [15] interpret the geometric descent of Bubeck, Lee, and Singh [8] as a sequence minimizing quadratic lower models of the objective function and obtain limited-memory extensions with improved performance. Lin, Mairal, and Harchaoui [24] provide a universal scheme for accelerating nonaccelerated first-order methods.

Su, Boyd, and Candes [38] and Krichene, Bayen, and Bartlett [22] interpret Nesterov's accelerated method as a discretization of a certain continuous-time dynamics and analyze it using Lyapunov stability criteria. Scieur et al. [34] interpret acceleration as a multi-step integration method from numerical analysis applied to the gradient flow differential equation. Wibisono, Wilson, and Jordan [40] and Wilson, Recht, and Jordan [41] interpret accelerated methods by using Lyapunov stability analysis and drawing ideas from Lagrangian mechanics. A recent note of Bansal and

Gupta [2] provides an intuitive, potential-based interpretation of many of the common convergence proofs for first-order methods.

The two references [40, 41] are most closely related to our work, as they are motivated by the Lagrangian view from classical mechanics that describes system dynamics through the principle of stationary action. In a similar spirit, most dynamics described in our work maintain the invariant that $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)}) = 0$. However, unlike our work, which relies on enforcing an invariant that leads both to the algorithms and their convergence analysis, the results from [40, 41] rely on the use of separate Lyapunov functions to obtain convergence results. It is unclear how these Lyapunov functions relate to the problems' optimality gaps, which makes them harder to generalize, especially in nonstandard settings such as [12, 10, 14].

The approximate duality gap presented here is closely related to Nesterov's estimate sequence (see, e.g., [33]). In particular, up to the regularization term $\phi(\mathbf{x}^*)/\alpha^{(t)}$, our lower bound $L^{(t)}$ is equivalent to Nesterov's estimate sequence, providing a natural interpretation of this powerful and commonly used technique.

1.2. Notation. We use italic letters to denote scalars, and boldface letters to denote vectors. Superscript index $(\cdot)^{(t)}$ denotes the value of $(\cdot)$ at time t. The "dot" notation is used to denote the time derivative, i.e., $\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t}$. Given a measure $\alpha^{(\tau)}$ defined on $\tau \in [0, t]$, we use the Lebesgue–Stieltjes notation for the integral. In particular, given $\phi^{(\tau)}$ defined on $\tau \in [0, t]$, we have

$$\int_0^t \phi^{(\tau)} \dot{\alpha}^{(\tau)} \mathrm{d}\tau = \int_0^t \phi^{(\tau)} \mathrm{d}\alpha^{(\tau)}.$$

In the discrete-time setting, we will assume that $\alpha^{(\tau)}$ is an increasing piecewise constant function, with discontinuities occurring only at discrete time points $i \in Z_+$, and such that $\alpha^{(t)} = 0$ for $t < 0$. Hence, $\dot{\alpha}^{(\tau)}$ can be expressed as a train of Dirac Delta functions: $\dot{\alpha}^{(\tau)} = \sum_{i=0}^{\infty} a_i \delta(\tau - i)$, where $a_i = \alpha^{(i+\Delta)} - \alpha^{(i-\Delta)}$ for $\Delta \in (0, 1)$. This means that $\dot{\alpha}^{(\tau)}$ samples the function under the integral, so that $\int_0^t \phi^{(\tau)} \mathrm{d}\alpha^{(\tau)} = \sum_{i=0}^{\lfloor t \rfloor} a_i \phi^{(i)}$.

We define

$$A^{(t)} = \int_0^t \mathrm{d}\alpha^{(\tau)},$$

so that $\frac{1}{A^{(t)}} \int_0^t \mathrm{d}\alpha^{(\tau)} = 1$. In continuous time $A^{(t)} = \alpha^{(t)} - \alpha^{(0)}$, while in discrete time $A^{(t)} = \sum_{i=0}^{\lfloor t \rfloor} a_i = \alpha^{(t)}$. We assume throughout the paper that $\alpha^{(0)} > 0$ and $\dot{\alpha}^{(t)} > 0 \ \forall t \geq 0$, and use the following notation for the aggregated negative gradients:

$$\mathbf{z}^{(t)} \stackrel{\text{def}}{=} -\int_0^t \nabla f(\mathbf{x}^{(\tau)}) \mathrm{d}\alpha^{(\tau)}. \tag{1.1}$$

For all problems considered, we assume that the feasible region is a closed convex set $X \subseteq \mathbb{R}^n$, for a finite n. We assume that there is a (fixed) norm $\|\cdot\|$ associated with 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[2] convex function $f(\cdot)$ defined on a convex set $X \subseteq \mathbb{R}^n$, and we let $\mathbf{x}^* = \arg\min_{\mathbf{x} \in X} f(\mathbf{x})$

[2]The differentiability assumption is not always necessary and can be relaxed to subdifferentiability in the case of dual-averaging/mirror-descent methods. Nevertheless, we will assume differentiability throughout the paper for simplicity of exposition.

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 $\|\cdot\|$ if $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x} \rangle + \frac{L}{2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2$ for all $\mathbf{x}, \hat{\mathbf{x}} \in X$. 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 $\|\cdot\|$ if, for all $\mathbf{x}, \hat{\mathbf{x}} \in X$, $f(\hat{\mathbf{x}}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \hat{\mathbf{x}} - \mathbf{x} \rangle + \frac{\sigma}{2}\|\hat{\mathbf{x}} - \mathbf{x}\|^2$. Equivalently, $\|\nabla f(\mathbf{x}) - \nabla f(\hat{\mathbf{x}})\|_* \geq \sigma\|\mathbf{x} - \hat{\mathbf{x}}\|$.*

DEFINITION 1.4 (Bregman divergence). *The Bregman divergence of a function ψ is defined as $D_\psi(\mathbf{x}, \hat{\mathbf{x}}) \stackrel{\text{def}}{=} \psi(\mathbf{x}) - \psi(\hat{\mathbf{x}}) - \langle \nabla\psi(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle$.*

DEFINITION 1.5 (convex conjugate). *Function $\psi^*(\cdot)$ is the convex conjugate of $\psi : X \to \mathbb{R}$ if $\psi^*(\mathbf{z}) = \sup_{\mathbf{x} \in X}\{\langle \mathbf{z}, \mathbf{x} \rangle - \psi(\mathbf{x})\}$ $\forall \mathbf{z} \in \mathbb{R}$.*

As X is assumed to be closed, sup in Definition 1.5 can be replaced by max.

We assume there is a differentiable strictly convex function $\phi : X \to \mathbb{R}$, possibly dependent on t (in which case we denote it by 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., $\phi^*(\mathbf{z}) = \max_{\mathbf{x} \in X}\{\langle \mathbf{z}, \mathbf{x} \rangle - \phi(\mathbf{x})\}$. We will further assume without loss of generality that $\min_{\mathbf{x} \in X} \phi(\mathbf{x}) \geq 0$.[3] The role of function ϕ will be to regularize the lower bound in the construction of the approximate duality gap.

The following standard fact, based on Danskin's theorem (see, e.g., [5, 6]), will be extremely useful in analyzing 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

$$\nabla\phi^*(\mathbf{z}^{(t)}) = \arg\min_{\mathbf{x} \in X}\left\{\int_0^t \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)} \rangle \mathrm{d}\alpha^{(\tau)} + \phi(\mathbf{x})\right\}. \tag{1.2}$$

Some other useful properties of Bregman divergence can be found in Appendix A.

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 $\dot{y}^{(t)} = \dot{\alpha}^{(t)} g(x^{(t)})$. Observe that, by integrating both sides from 0 to t, this is equivalent to $y^{(t)} = y^{(0)} + \int_0^t g(x^{(\tau)}) \mathrm{d}\alpha^{(\tau)}$. Averaging of a function $g(x)$ is $\dot{y}^{(t)} = \dot{\alpha}^{(t)} \frac{g(x^{(t)}) - y^{(t)}}{\alpha^{(t)}}$. This can equivalently be written as $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)} y^{(t)}) = \dot{\alpha}^{(t)} g(x^{(t)})$, implying

$$y^{(t)} = \frac{\alpha^{(0)}}{\alpha^{(t)}} y^{(0)} + \frac{1}{\alpha^{(t)}} \int_0^t g(\mathbf{x}^{(\tau)}) \mathrm{d}\alpha^{(\tau)}.$$

The following simple proposition will be useful in our analysis.

PROPOSITION 1.7. *We have $\frac{\mathrm{d}}{\mathrm{d}t} \min_{\mathbf{x} \in X}\{-\langle \mathbf{z}^{(t)}, \mathbf{x} \rangle + \phi(\mathbf{x})\} = -\langle \dot{\mathbf{z}}^{(t)}, \nabla\phi^*(\mathbf{z}^{(t)}) \rangle$.*

Proof. The proof follows by observing that $\phi^*(\mathbf{z}^{(t)}) = -\min_{\mathbf{x} \in X}\{-\langle \mathbf{z}^{(t)}, \mathbf{x} \rangle + \phi(\mathbf{x})\}$ and applying the chain rule. ∎

[3] This assumption can be easily satisfied by taking $\phi(\cdot)$ to be, for example, a Bregman divergence: $\phi(\mathbf{x}) = D_\psi(\mathbf{x}, \mathbf{x}^{(0)})$ for some strictly convex and differentiable ψ and fixed $\mathbf{x}^{(0)} \in X$.

2. The approximate duality gap technique. As mentioned in the introduction, to unify the analysis of a large class of first-order methods, we will show how to construct an upper estimate $G^{(t)}$ of the optimality gap $f(\hat{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*)$, where $\hat{\mathbf{x}}^{(t)}$ is the output of a first-order method at time t. This upper estimate is defined as $G^{(t)} = U^{(t)} - L^{(t)}$, where $U^{(t)} \geq f(\hat{\mathbf{x}}^{(t)})$ is an upper bound on $f(\hat{\mathbf{x}}^{(t)})$ and $L^{(t)} \leq f(\mathbf{x}^*)$ is a lower bound on $f(\mathbf{x}^*)$. We refer to $G^{(t)}$ as the approximate duality gap, due to the connections between the lower bound $L^{(t)}$ and the Fenchel dual of a certain approximation of the objective function $f(\mathbf{x}^{(t)})$, further discussed in section 2.2. To show that the method converges at some rate $\alpha^{(t)}$ (e.g., $\alpha^{(t)} = t$), we will show that $\alpha^{(t)}G^{(t)}$ is a nonincreasing function of time, so that $\alpha^{(t)}G^{(t)} \leq \alpha^{(0)}G^{(0)}$, and, consequently, $f(\hat{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*) \leq G^{(t)} \leq {}^{\alpha^{(0)}G^{(0)}}/_{\alpha^{(t)}}$.

2.1. Upper bound. The simplest upper bound on $f(\hat{\mathbf{x}}^{(t)})$ is $f(\hat{\mathbf{x}}^{(t)})$ itself, i.e., $U^{(t)} = f(\hat{\mathbf{x}}^{(t)})$. In this case, $\hat{\mathbf{x}}^{(t)}$ will be the last point constructed by the algorithm, i.e., $\hat{\mathbf{x}}^{(t)} = \mathbf{x}^{(t)}$. We will make this choice of the upper bound whenever we can assume that $f(\cdot)$ is differentiable (e.g., in the setting of accelerated and Frank–Wolfe methods), so that in the continuous-time setting we can differentiate $\alpha^{(t)}U^{(t)}$ with respect to t and write $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}U^{(t)}) = \dot{\alpha}^{(t)}f(\hat{\mathbf{x}}^{(t)}) + \alpha^{(t)}\langle \nabla f(\hat{\mathbf{x}}^{(t)}), \frac{\mathrm{d}}{\mathrm{d}t}\hat{\mathbf{x}}^{(t)}\rangle$. In the settings where $f(\cdot)$ is typically not assumed to be differentiable but only subdifferentiable (e.g., in the setting of dual-averaging/mirror-descent methods), $\hat{\mathbf{x}}^{(t)}$ will be a weighted average of the points $\mathbf{x}^{(\tau)}$ constructed by the method up to time t: $\hat{\mathbf{x}}^{(t)} = \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}}\mathbf{x}^{(0)} + \frac{1}{\alpha^{(t)}}\int_0^t \mathbf{x}^{(\tau)}\mathrm{d}\alpha^{(\tau)}$, and we will choose $U^{(t)} = \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}}f(\mathbf{x}^{(0)}) + \frac{1}{\alpha^{(t)}}\int_0^t f(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)}$. Due to Jensen's inequality, $f(\hat{\mathbf{x}}^{(t)}) \leq U^{(t)}$, i.e., $U^{(t)}$ is a valid upper bound on $f(\hat{\mathbf{x}}^{(t)})$. Observe that this choice of $U^{(t)}$ allows us to differentiate $\alpha^{(t)}U^{(t)}$ with respect to t in the continuous-time setting, and, thus, we can write $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}U^{(t)}) = \dot{\alpha}^{(t)}f(\mathbf{x}^{(t)})$, as $\alpha^{(t)} - A^{(t)} = \alpha^{(0)}$ is a constant. These choices of upper bounds easily extend to the setting of composite objectives $\bar{f}(\cdot) = f(\cdot) + \psi(\cdot)$ (see section 3 for more details).

2.2. Lower bound. The simplest lower bound on $f(\mathbf{x}^*)$ is $f(\mathbf{x}^*)$ itself. However, it is not clear how to use $L^{(t)} = f(\mathbf{x}^*)$ and guarantee $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)}) \leq 0$, as in that case in the continuous-time domain $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}L^{(t)}) = \dot{\alpha}^{(t)}f(\mathbf{x}^*)$ is not possible to evaluate, as we do not know $f(\mathbf{x}^*)$ (recall that, by assumption, $\dot{\alpha}^{(t)} > 0$). Observe that if, instead, $f(\mathbf{x}^*)$ appeared in the lower bound as $\frac{c}{\alpha^{(t)}}f(\mathbf{x}^*)$ for some constant c, we would not have this problem anymore, as $f(\mathbf{x}^*)$ would not appear in $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}L^{(t)})$. This is true because $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}\frac{c}{\alpha^{(t)}}f(\mathbf{x}^*)) = \frac{\mathrm{d}}{\mathrm{d}t}(cf(\mathbf{x}^*)) = 0$. On the other hand, convexity of $f(\cdot)$ leads to the following lower-bounding hyperplanes for all $\mathbf{x} \in X$: $f(\mathbf{x}^*) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x}\rangle$.[4] In particular, taking a convex combination of the trivial lower bound $f(\mathbf{x}^*)$ and the lower-bounding hyperplanes defined by points $\mathbf{x}^{(\tau)}$ constructed by the method up to time t, we have

$$f(\mathbf{x}^*) \geq \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}}f(\mathbf{x}^*) + \frac{1}{\alpha^{(t)}}\int_0^t (f(\mathbf{x}^{(\tau)}) + \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x}^* - \mathbf{x}^{(\tau)}\rangle)\mathrm{d}\alpha^{(\tau)}. \tag{2.1}$$

As in the continuous-time domain $\alpha^{(t)} - A^{(t)} = \alpha^{(0)}$ is a positive constant, the lower bound equal to the right-hand side of (2.1) is well defined at $t = 0$ and $f(\mathbf{x}^*)$ does not appear in $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}L^{(t)})$. However, it would still not be possible to evaluate $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}L^{(t)})$,

[4] Observe here that if $f(\cdot)$ is not differentiable but only subdifferentiable, we can still obtain lower-bounding hyperplanes by using subgradients in place of the gradients.

as $\mathbf{x}^*$ is not known. One way of addressing this issue is to use that

$$\int_0^t (f(\mathbf{x}^{(\tau)}) + \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x}^* - \mathbf{x}^{(\tau)} \rangle) \mathrm{d}\alpha^{(\tau)} \geq \int_0^t \min_{\mathbf{u} \in X} (f(\mathbf{x}^{(\tau)}) + \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{u} - \mathbf{x}^{(\tau)} \rangle) \mathrm{d}\alpha^{(\tau)}, \tag{2.2}$$

or

$$\int_0^t (f(\mathbf{x}^{(\tau)}) + \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x}^* - \mathbf{x}^{(\tau)} \rangle) \mathrm{d}\alpha^{(\tau)} \geq \min_{\mathbf{u} \in X} \int_0^t (f(\mathbf{x}^{(\tau)}) + \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{u} - \mathbf{x}^{(\tau)} \rangle) \mathrm{d}\alpha^{(\tau)}. \tag{2.3}$$

While the inequality (2.3) is tighter than (2.2), as a minimum of affine functions it is not differentiable w.r.t. $\mathbf{u}$ (and consequently not differentiable w.r.t. t). The use of (2.2) leads to the continuous-time version of the standard Frank–Wolfe method [18].

Example 2.1 (standard continuous-time Frank–Wolfe method). Using (2.2), we have the following lower bound:

$$f(\mathbf{x}^*) \geq L^{(t)} \stackrel{\text{def}}{=} \frac{\int_0^t \min_{\mathbf{u} \in X} (f(\mathbf{x}^{(\tau)}) + \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{u} - \mathbf{x}^{(\tau)} \rangle) \mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}} + \frac{(\alpha^{(t)} - A^{(t)}) f(\mathbf{x}^*)}{\alpha^{(t)}}. \tag{2.4}$$

Since the standard assumption in this setting is that $f(\cdot)$ is smooth (or, at the very least, continuously differentiable), we take $U^{(t)} = f(\mathbf{x}^{(t)})$ and $\hat{\mathbf{x}}^{(t)} = \mathbf{x}^{(t)}$. Setting $\mathbf{v}^{(t)} \in \operatorname{argmin}_{\mathbf{u} \in X} (f(\mathbf{x}^{(\tau)}) + \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{u} - \mathbf{x}^{(\tau)} \rangle)$ and computing $\alpha^{(t)} G^{(t)}$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t} (\alpha^{(t)} G^{(t)}) = \langle \nabla f(\mathbf{x}^{(t)}), \alpha^{(t)} \dot{\mathbf{x}}^{(t)} - \dot{\alpha}^{(t)} (\mathbf{v}^{(t)} - \mathbf{x}^{(t)}) \rangle.$$

Setting $\alpha^{(t)} \dot{\mathbf{x}}^{(t)} - \dot{\alpha}^{(t)} (\mathbf{v}^{(t)} - \mathbf{x}^{(t)}) = 0$ gives $\frac{\mathrm{d}}{\mathrm{d}t} (\alpha^{(t)} G^{(t)}) = 0$ and precisely recovers the continuous-time version of the Frank–Wolfe algorithm, as in this case $\dot{\mathbf{x}}^{(t)} = \frac{\dot{\alpha}^{(t)} (\mathbf{v}^{(t)} - \mathbf{x}^{(t)})}{\alpha^{(t)}}$, i.e., $\mathbf{x}^{(t)}$ is a weighted average of $\mathbf{v}^{(t)}$'s (as explained in section 1.3).

Notice that the use of (2.2) in the construction of the lower bound makes sense only when linear minimization over X is possible. However, there are insights we can take from the construction of the lower bound (2.4). In particular, we can alternatively view (2.4) as being constructed as a lower bound on $f(\mathbf{x}^*) + \psi(\mathbf{x}^*)$, where $\psi(\mathbf{x}^*)$ is the indicator of X. Hence, we can view $L^{(t)}$ from (2.4) as being constructed as follows:

$$\begin{aligned} f(\mathbf{x}^*) + \psi(\mathbf{x}^*) &\geq \frac{\int_0^t \left(f(\mathbf{x}^{(\tau)}) + \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x}^* - \mathbf{x}^{(\tau)} \rangle + \psi(\mathbf{x}^*) \right) \mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}} \\ &\quad + \frac{(\alpha^{(t)} - A^{(t)}) f(\mathbf{x}^*)}{\alpha^{(t)}} \\ &\geq L^{(t)}. \end{aligned}$$

We will see later, in section 3, how this leads to a more general version of the Frank–Wolfe method for composite functions, along the lines of the method from [32].

Constructing a lower bound on $f(\mathbf{x}^*)+\psi(\mathbf{x}^*)$ when $\psi(\mathbf{x}^*)$ was an indicator function had no effect, as in that case $f(\mathbf{x}^*)+\psi(\mathbf{x}^*)=f(\mathbf{x}^*)$. To generalize this idea, we can construct a lower bound on a function that closely approximates $f(\cdot)$ around $\mathbf{x}^*$. Since we want to obtain convergent methods, any error we introduce into this approximation should vanish at rate $1/\alpha^{(t)}$. Hence, a natural choice is to create a lower bound on $f(\mathbf{x}^*)+\frac{1}{\alpha^{(t)}}\phi(\mathbf{x}^*)$, where $\phi(\mathbf{x}^*)$ is bounded:[5]

$$f(\mathbf{x}^*)+\frac{1}{\alpha^{(t)}}\phi(\mathbf{x}^*)\geq\frac{\int_0^t f(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)}+\int_0^t\langle\nabla f(\mathbf{x}^{(\tau)}),\mathbf{x}^*-\mathbf{x}^{(\tau)}\rangle\mathrm{d}\alpha^{(\tau)}+\phi(\mathbf{x}^*)}{\alpha^{(t)}}+\frac{(\alpha^{(t)}-A^{(t)})f(\mathbf{x}^*)}{\alpha^{(t)}}.$$

Now, if $\phi(\cdot)$ is strictly convex, $\min_{\mathbf{u}\in X}\{\int_0^t\langle\nabla f(\mathbf{x}^{(\tau)}),\mathbf{u}-\mathbf{x}^{(\tau)}\rangle\mathrm{d}\alpha^{(\tau)}+\phi(\mathbf{u})\}$ is differentiable (i.e., we can generalize the stronger inequality from (2.3)). We can view this as regularization of the minimum from (2.3), leading to the following lower bound:

$$\begin{aligned}(2.5)\qquad f(\mathbf{x}^*)+\frac{1}{\alpha^{(t)}}\phi(\mathbf{x}^*)&\geq L^{(t)}+\frac{\phi(\mathbf{x}^*)}{\alpha^{(t)}}\\&=\frac{\int_0^t f(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)}+\min_{\mathbf{u}\in X}\{\int_0^t\langle\nabla f(\mathbf{x}^{(\tau)}),\mathbf{u}-\mathbf{x}^{(\tau)}\rangle\mathrm{d}\alpha^{(\tau)}+\phi(\mathbf{u})\}}{\alpha^{(t)}}\\&\quad+\frac{(\alpha^{(t)}-A^{(t)})f(\mathbf{x}^*)}{\alpha^{(t)}}.\end{aligned}$$

Another advantage of the lower bound from (2.5) over the previous one is that for many feasible sets X there are natural choices of ϕ for which the minimization inside the lower bound from (2.5) is easily solvable, often in a closed form (see, e.g., [4]).

Dual view of the lower bound. An alternative view of the lower bound from (2.5) is through the concept of Fenchel duality, which is defined for the sum of two convex functions (or the difference of a convex and a concave function). In particular, the Fenchel dual of $f(\mathbf{x})+\phi_t(\mathbf{x})$ is defined as $-f^*(-\mathbf{u})-\phi_t^*(\mathbf{u})$ (see, e.g., [3, Chapter 15.2]). Let $\phi_t(\cdot)=\frac{1}{A^{(t)}}\phi(\cdot)$ and

$$\mathbf{u}^{(t)}=-\frac{\int_0^t\nabla f(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)}}{A^{(t)}}.$$

Observe that the minimization problem from (2.5) defines a convex conjugate of ϕ_t. Thus, we can equivalently write (2.5) as

$$\begin{aligned}f(\mathbf{x}^*)+\frac{A^{(t)}}{\alpha^{(t)}}\phi_t(\mathbf{x}^*)&\geq L^{(t)}+\frac{A^{(t)}}{\alpha^{(t)}}\phi_t(\mathbf{x}^*)\\&=\frac{\int_0^t(f(\mathbf{x}^{(\tau)})-\langle\nabla f(\mathbf{x}^{(\tau)}),\mathbf{x}^{(\tau)}\rangle)\mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}}-\frac{A^{(t)}}{\alpha^{(t)}}\phi_t^*(\mathbf{u}^{(t)})+\frac{\alpha^{(t)}-A^{(t)}}{\alpha^{(t)}}f(\mathbf{x}^*)\\&=\frac{-\int_0^t f^*(\nabla f(\mathbf{x}^{(\tau)}))\mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}}-\frac{A^{(t)}}{\alpha^{(t)}}\phi_t^*(\mathbf{u}^{(t)})+\frac{\alpha^{(t)}-A^{(t)}}{\alpha^{(t)}}f(\mathbf{x}^*).\end{aligned}$$

[5] A common choice of $\phi(\cdot)$ that ensures boundedness of $\phi(\mathbf{x}^*)$ and nonnegativity of $\phi(\cdot)$ is the Bregman divergence of some function ψ; namely, $\phi(\cdot)=D_\psi(\cdot,\mathbf{x}^{(0)})$. Hence, in this case we can interpret $\phi(\mathbf{x}^*)$ as a generalized notion of the initial distance to the optimal solution.

Rearranging the terms in the last equality, we can equivalently write

$$\begin{aligned} L^{(t)} &+ \phi_t(\mathbf{x}^*) \\ &= -\frac{A^{(t)}}{\alpha^{(t)}}\left(\frac{\int_0^t f^*(\nabla f(\mathbf{x}^{(\tau)}))\mathrm{d}\alpha^{(\tau)}}{A^{(t)}} + \phi_t^*(\mathbf{u}^{(t)})\right) + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}}(f(\mathbf{x}^*) + \phi_t(\mathbf{x}^*)) \\ &\geq \frac{A^{(t)}}{\alpha^{(t)}}(-f^*(-\mathbf{u}^{(t)}) - \phi_t^*(\mathbf{u}^{(t)})) + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}}(f(\mathbf{x}^*) + \phi_t(\mathbf{x}^*)), \end{aligned}$$

where the last line is obtained by Jensen's inequality. Hence, we can view the general lower bound from (2.5) as being slightly stronger than the weighted average of $f(\mathbf{x}^*)+\phi_t(\mathbf{x}^*)$ and its Fenchel dual $-f^*(-\mathbf{u}^{(t)}) - \phi_t^*(\mathbf{u}^{(t)})$ evaluated at the average negative gradient $\mathbf{u}^{(t)}$, and corrected by the introduced approximation error $\phi_t(\mathbf{x}^*)$. This means that we can think about the lower bound $L^{(t)}$ as encoding the Fenchel dual of $f(\mathbf{x}^*) + \phi_t(\mathbf{x}^*)$—an approximation to $f(\mathbf{x}^*)$ that converges to $f(\mathbf{x}^*)$ at rate $1/\alpha^{(t)}$—and constructing dual solutions from the history of the gradients of f.

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 lead to 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:

$$\begin{aligned} L^{(t)} = &\frac{\int_0^t f(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}} + \frac{(\alpha^{(t)} - A^{(t)})f(\mathbf{x}^*) - \phi(\mathbf{x}^*)}{\alpha^{(t)}} \\ &+ \frac{\min_{\mathbf{x}\in X}\{\int_0^t(\langle\nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)}\rangle + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^{(\tau)}\|^2)\mathrm{d}\alpha + \phi(\mathbf{x})\}}{\alpha^{(t)}}. \end{aligned} \tag{2.6}$$

Remark 2.2. Note that, due to the strong convexity of f, we do not need additional regularization in the lower bound to ensure that the minimum inside it is differentiable, i.e., we could use $\phi(\cdot) = 0$. This choice of ϕ will have no effect on the continuous-time convergence. In discrete time, however, if we choose $\phi(\cdot) = 0$, the initial gap (and, consequently, the convergence bound) would scale with $\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|^2$. Adding a little bit of regularization (i.e., choosing a nonzero ϕ) will allow us to replace $\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|^2$ with $\|\mathbf{x}^* - \mathbf{x}^{(0)}\|^2$.

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 particularly useful in the discrete-time domain in the settings where $f(\cdot)$ has some smoothness properties while $\psi(\cdot)$ is generally nonsmooth). 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, we get

$$\begin{aligned} L^{(t)} \stackrel{\text{def}}{=} &\frac{\int_0^t f(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}} + \frac{(\alpha^{(t)} - A^{(t)})\bar{f}(\mathbf{x}^*) - \phi(\mathbf{x}^*)}{\alpha^{(t)}} \\ &+ \frac{\min_{\mathbf{x}\in X}\{\int_0^t\langle\nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)}\rangle\mathrm{d}\alpha^{(\tau)} + A^{(t)}\psi(\mathbf{x}) + \phi(\mathbf{x})\}}{\alpha^{(t)}}. \end{aligned} \tag{2.7}$$

2.3. Extension to monotone operators and saddle-point formulations. The notion of the approximate gap can be defined for problem classes beyond convex minimization. Examples are monotone operators and convex-concave saddle-point problems. More details are provided in Appendix B.

3. First-order methods in continuous time. We now show how different assumptions about the problem, leading to different choices of the upper and lower bounds (and, consequently, the gap), yield different first-order methods.

3.1. Mirror-descent/dual-averaging methods. Let us start by making minimal assumptions about the objective function f: we will assume that f is convex and subdifferentiable (with abuse of notation; in this case $\nabla f(\mathbf{x}^{(t)})$ denotes an arbitrary but fixed subgradient of f at $\mathbf{x}^{(t)}$). As discussed in the previous section, since we are not assuming that f is differentiable, we will take

$$U^{(t)} = \frac{\alpha^{(0)}}{\alpha^{(t)}} f(\mathbf{x}^{(0)}) + \frac{\int_0^t f(\mathbf{x}^{(\tau)}) \mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}},$$

so that $\alpha^{(t)} U^{(t)}$ is differentiable w.r.t. the time and well defined at the initial time point. As there are no additional assumptions about the objective (such as composite structure and strong convexity), we will use the generic lower bound from (2.5). Hence, we have the following expression for the gap:

$$G^{(t)} = \frac{-\min_{\mathbf{x}\in X}\{\int_0^t \langle \nabla f(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)}\rangle \mathrm{d}\alpha^{(\tau)} + \phi(\mathbf{x})\}}{\alpha^{(t)}} + \frac{\alpha^{(0)}(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)) + \phi(\mathbf{x}^*)}{\alpha^{(t)}}. \tag{3.1}$$

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

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

Now we show that ensuring the invariance $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)} G^{(t)}) = 0$ produces exactly the continuous-time mirror-descent dynamics. Using (1.2) and Proposition 1.7 with $\mathbf{z}^{(t)} = -\int_0^t \nabla f(\mathbf{x}^{(\tau)}) \mathrm{d}\alpha^{(t)}$ yields

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

Thus, to have $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)} G^{(t)}) = 0$, we can set $\mathbf{x}^{(t)} = \nabla \phi^*(\mathbf{z}^{(t)})$, which is precisely the mirror-descent dynamics from [26]:

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

We immediately obtain 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,* $\forall t > 0$,

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

3.2. Accelerated convex minimization. Let us now assume more about the objective function: we will assume that f is continuously differentiable, which means that our choice of the upper bound will be $U^{(t)} = f(\hat{\mathbf{x}}^{(t)}) = f(\mathbf{x}^{(t)})$. Since there are no additional assumptions about f, the lower bound we use is the generic one from (2.5). Differentiating $\alpha^{(t)}G^{(t)}$, we have

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

where we have used $\frac{\mathrm{d}}{\mathrm{d}t}(f(\mathbf{x}^{(t)})) = \langle \nabla f(\mathbf{x}^{(t)}), \dot{\mathbf{x}}^{(t)}\rangle$. Choosing $\dot{\mathbf{x}}^{(t)} = \dot{\alpha}^{(t)}\frac{\nabla\phi^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}$, we get $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)}) = 0$. This is precisely the accelerated mirror-descent dynamics [22, 40],

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

and we immediately get the convergence result stated as in Lemma 3.2 below.

LEMMA 3.2. *Let* $\mathbf{x}^{(t)}$, $\mathbf{z}^{(t)}$ *evolve according to* (CT-AMD), *for some continuously differentiable convex function* $f : X \to \mathbb{R}$. *Then,* $\forall t > 0$,

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

3.3. Gradient descent. Using the same approximate gap as in the previous subsection, now consider the special case when $\phi(\mathbf{x}) = \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^{(0)}\|_2^2$ for an arbitrary initial point $\mathbf{x}^{(0)} \in X$, $X = \mathbb{R}^n$, and some $\sigma > 0$. Then, $\nabla\phi^*(\mathbf{z}^{(t)}) = \mathbf{x}^{(0)} + \mathbf{z}^{(t)}/\sigma$. Using the result for $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)})$ from the previous subsection and setting $\mathbf{x}^{(t)} = \nabla\phi^*(\mathbf{z}^{(t)})$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)}) = \alpha^{(t)}\langle \nabla f(\mathbf{x}^{(t)}), \dot{\mathbf{x}}^{(t)}\rangle = -\frac{\alpha^{(t)}\dot{\alpha}^{(t)}}{\sigma}\|\nabla f(\mathbf{x}^{(t)})\|_2^2 \le 0.$$

The choice $\mathbf{x}^{(t)} = \nabla\phi^*(\mathbf{z}^{(t)}) = \mathbf{x}^{(0)} + \mathbf{z}^{(t)}/\sigma$ precisely defines the gradient descent algorithm:

$$\text{(CT-GD)}\qquad \begin{aligned}\dot{\mathbf{z}}^{(t)} &= -\dot{\alpha}^{(t)}\nabla f(\mathbf{x}^{(t)}),\\ \mathbf{x}^{(t)} &= \nabla\phi^*(\mathbf{z}^{(t)}) = \mathbf{x}^{(0)} + \mathbf{z}^{(t)}/\sigma,\\ \mathbf{z}^{(0)} &= 0,\ \mathbf{x}^{(0)} \in \mathbb{R}^n \text{ is an arbitrary initial point},\end{aligned}$$

and the convergence result, stated as in Lemma 3.3, immediately follows.

LEMMA 3.3. *Let* $\mathbf{x}^{(t)}$, $\mathbf{z}^{(t)}$ *evolve according to* (CT-GD), *for some continuously differentiable convex function* $f : \mathbb{R}^n \to \mathbb{R}$. *Then,* $\forall t > 0$,

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \le \frac{\alpha^{(0)}(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)) + \frac{\sigma}{2}\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2^2}{\alpha^{(t)}}.$$

3.4. Accelerated strongly convex minimization. Let us now assume that, in addition to being continuously differentiable, f is strongly convex. In that case, we use $U^{(t)} = f(\mathbf{x}^{(t)})$ and the lower bound from (2.6). Let

$$\phi_t(\mathbf{x}) = \int_0^t \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^{(\tau)}\|^2 \mathrm{d}\alpha^{(\tau)} + \phi(\mathbf{x}).$$

Observe that $\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(\mathbf{x}) \geq 0 \ \forall \mathbf{x} \in X$. Then, we have the following result for the change in the gap:

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

Therefore, choosing $\dot{\mathbf{x}}^{(t)} = \dot{\alpha}^{(t)}\frac{\nabla\phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}$, namely

$$\text{(CT-ASC)} \qquad \begin{aligned}\dot{\mathbf{z}}^{(t)} &= -\dot{\alpha}^{(t)}\nabla f(\mathbf{x}^{(t)}), \\ \dot{\mathbf{x}}^{(t)} &= \dot{\alpha}^{(t)}\frac{\nabla\phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}, \\ \mathbf{z}^{(0)} = 0,\ &\mathbf{x}^{(0)} \in \mathbb{R}^n \text{ is an arbitrary initial point,}\end{aligned}$$

gives $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)}) \leq 0$, and the convergence result, stated as in Lemma 3.4 below, follows.

LEMMA 3.4. *Let* $\mathbf{x}^{(t)}$ *evolve according to* (CT-ASC) *for some continuously differentiable and* σ*-strongly convex function* F*, where* $\phi_t(\mathbf{x}) = \int_0^t \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^{(\tau)}\|^2 + \phi(\mathbf{x})$*. Then,* $\forall t > 0$*,*

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

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

3.5. Composite dual averaging. 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 (2.7). Since we are not assuming that either f or ψ is continuously differentiable, the upper bound of choice is

$$\begin{aligned}U^{(t)} &= \frac{1}{\alpha^{(t)}}\int_0^t \bar{f}(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)} + \frac{\alpha^{(0)}}{\alpha^{(t)}}\bar{f}(\mathbf{x}^{(0)}) \\ &= \frac{1}{\alpha^{(t)}}\int_0^t (f(\mathbf{x}^{(\tau)}) + \psi(\mathbf{x}^{(\tau)}))\mathrm{d}\alpha^{(\tau)} + \frac{\alpha^{(0)}}{\alpha^{(t)}}(f(\mathbf{x}^{(0)}) + \psi(\mathbf{x}^{(0)})).\end{aligned}$$

Then, the change in the gap is

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

Thus, when $\dot{\mathbf{x}}^{(t)} = \nabla\phi_t^*(\mathbf{z}^{(t)})$, where $\phi_t(\cdot) = \alpha^{(t)}\psi(\cdot)$, we have $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)}) = 0$, and Lemma 3.5 follows immediately. The following algorithm can be thought of as mirror descent (or dual averaging) for composite minimization:

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

LEMMA 3.5. *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,* $\forall t > 0$,

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

3.6. Generalized Frank–Wolfe method. We have already discussed the standard version of Frank–Wolfe method in Example 2.1. As discussed there, we can view standard Frank–Wolfe method as minimizing a composite objective $\bar{f} = f + \psi$, where ψ is the indicator function of set X, and using the assumption that problems of the form $\min_{\mathbf{u}\in X}\{\langle \mathbf{z}, \mathbf{u}\rangle + \psi(\mathbf{u})\}$ are easily solvable for any fixed $\mathbf{z}$. We will now show how the method generalizes for any (possibly nondifferentiable) ψ. The lower bound from (2.4) derived for the standard Frank–Wolfe method immediately generalizes to

$$\text{(3.3)}\qquad \begin{aligned} L^{(t)} = &\frac{\int_0^t f(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)} + \int_0^t \min_{\mathbf{u}\in X}\{\langle \nabla f(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)}, \mathbf{u} - \mathbf{x}^{(\tau)}\rangle + \psi(\mathbf{u})\}}{\alpha^{(t)}}\\ &+ \frac{(\alpha^{(t)} - A^{(t)})\bar{f}(\mathbf{x}^*)}{\alpha^{(t)}}. \end{aligned}$$

By the (generalized) Danskin theorem [5], we have that

$$\nabla\psi^*(-\nabla f(\mathbf{x}^{(\tau)})) \in \operatorname*{argmin}_{\mathbf{u}\in X}\{\langle \nabla f(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)}, \mathbf{u} - \mathbf{x}^{(\tau)}\rangle + \psi(\mathbf{u})\}$$

(note that the minimizer may not be unique as ψ is not necessarily strictly convex; with abuse of notation, $\nabla\psi^*(-\nabla f(\mathbf{x}^{(\tau)}))$ may be a subgradient of ψ^*). Since f is assumed to be continuously differentiable, we would like to use $f(\mathbf{x}^{(t)})$ in the upper bound. On the other hand, ψ is not necessarily differentiable, and hence we cannot use $\psi(\mathbf{x}^{(t)})$ in the upper bound. Instead, we need to have

$$\frac{\alpha^{(0)}}{\alpha^{(t)}}\psi(\mathbf{x}^{(0)}) + \frac{\int_0^t \psi(\mathbf{v}^{(\tau)})\mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}}$$

for some points $\mathbf{v}^{(\tau)} \in X$ to ensure that $\alpha^{(t)}U^{(t)}$ is differentiable. Hence, based on the rules for choosing the upper bound discussed in section 2, we would like the upper bound to be of the form

$$\text{(3.4)}\qquad U^{(t)} = f(\mathbf{x}^{(t)}) + \frac{\alpha^{(0)}}{\alpha^{(t)}}\psi(\mathbf{x}^{(0)}) + \frac{\int_0^t \psi(\mathbf{v}^{(\tau)})\mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}}.$$

Using Jensen's inequality, this is a valid upper bound on $\bar{f}(\mathbf{x}^{(t)})$ if $\mathbf{x}^{(t)} = \frac{\alpha^{(0)}}{\alpha^{(t)}}\mathbf{x}^{(0)} + \frac{\int_0^t \mathbf{v}^{(\tau)}\mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}}$. Taking a leap of faith, let us consider the gap constructed based on the

lower and upper bounds from (3.3), (3.4). Differentiating $\alpha^{(t)}G^{(t)}$ yields

$$\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)}) = \langle \nabla f(\mathbf{x}^{(t)}), \alpha^{(t)}\dot{\mathbf{x}}^{(t)} - \dot{\alpha}^{(t)}(\nabla\psi^*(-\nabla f(\mathbf{x}^{(t)})) - \mathbf{x}^{(t)})\rangle + \dot{\alpha}^{(t)}\psi(\mathbf{v}^{(t)}) - \dot{\alpha}^{(t)}\psi(\nabla\psi^*(-\nabla f(\mathbf{x}^{(t)}))).$$

For the terms from the second line to cancel out, we need $\mathbf{v}^{(t)} = \nabla\psi^*(-\nabla f(\mathbf{x}^{(t)}))$. Then, if we set $\alpha^{(t)}\dot{\mathbf{x}}^{(t)} - \dot{\alpha}^{(t)}(\mathbf{v}^{(t)} - \mathbf{x}^{(t)}) = 0$, we get $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)}) = 0$. But this precisely defines $\mathbf{x}^{(t)}$ as

$$\mathbf{x}^{(t)} = \frac{\alpha^{(0)}}{\alpha^{(t)}}\mathbf{x}^{(0)} + \frac{\int_0^t \mathbf{v}^{(\tau)}\mathrm{d}\alpha^{(\tau)}}{\alpha^{(t)}},$$

which is what we needed for the upper bound to be valid. Hence, we precisely recover the continuous-time counterpart of the generalized Frank–Wolfe method from [32], i.e.,

$$\text{(CT-FW)} \qquad \begin{aligned} \hat{\mathbf{z}}^{(t)} &= -\nabla f(\mathbf{x}^{(t)}), \\ \dot{\mathbf{x}}^{(t)} &= \dot{\alpha}^{(t)}\frac{\nabla\psi^*(\hat{\mathbf{z}}^{(t)}) - \mathbf{x}^{(t)}}{\alpha^{(t)}}, \\ \mathbf{x}^{(0)} &\in X \text{ is an arbitrary initial point,} \end{aligned}$$

and Lemma 3.6 follows.

LEMMA 3.6. *Let $\mathbf{x}^{(t)}$, $\hat{\mathbf{z}}^{(t)}$ evolve according to* (CT-FW) *for some convex composite function $\bar{f} = f + \psi : X \to \mathbb{R}$, where f is continuously differentiable. Then*

$$\bar{f}(\mathbf{x}^{(t)}) - \bar{f}(\mathbf{x}^*) \leq \frac{\alpha^{(0)}(\bar{f}(\mathbf{x}^{(0)}) - \bar{f}(\mathbf{x}^*))}{\alpha^{(t)}} \quad \forall t \geq 0.$$

4. Discretization and incurred errors. Suppose now that $\alpha^{(t)}$ is a discrete measure. In particular, let $\alpha^{(t)}$ be an increasing piecewise constant function, with $\alpha^{(t)} = 0$ for $t < 0$, $\alpha^{(t)}$ constant in intervals $(0 + i, 0 + i + 1)$ for $i \in \mathbb{Z}_+$, and $\alpha^{((0+i)+)} - \alpha^{((0+i)-)} = a_i$ for some $a_i > 0$ and $i \in \mathbb{Z}_+$, as discussed in section 1.2.

For the continuous-time algorithms (and their analyses) presented in section 3, the discretization error generally has two causes: (i) different integration rules applying to continuous and discrete measures, and (ii) discontinuities in the algorithm updates. We discuss these two causes in more detail 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

$$I^{(0,t)} = -\int_0^t \langle \nabla f(\mathbf{x}^{(\tau)}), \nabla\phi_t^*(\mathbf{z}^{(t)}) - \mathbf{x}^{(\tau)}\rangle \mathrm{d}\alpha^{(\tau)},$$

where $\phi_t(\cdot) = \phi(\cdot)$ in the case of mirror descent and accelerated convex minimization. Let $I_c^{(0,t)}$ denote the value that $I^{(0,t)}$ would take if α were a continuous measure, i.e., if the rules of continuous integration applied.

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

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

Discontinuities in the algorithm updates. In all of the algorithms described, 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 negative gradients up to time t and thus also depends on $\nabla f(\mathbf{x}^{(t)})$. In the continuous-time domain, 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, $\mathbf{x}^{(t)}$ changes simultaneously with $\mathbf{z}^{(t)}$. In the discrete-time domain, however, the delay between the two updates cannot be neglected, and using implicit updates of the form $\mathbf{x}^{(t)} = g(\nabla\phi_t^*(\mathbf{z}^{(t)}))$ for some function $g(\cdot)$ either is not possible in general 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 $\tau \in [0,t]$) from the upper bound by $\dot{\alpha}^{(t)}\psi(\nabla\phi_t^*(\mathbf{z}^{(t)}))$ from the lower bound. For composite dual averaging, this discretization error at time i will amount to $a_i(\psi(\mathbf{x}^{(i)}) - \psi(\nabla\phi_i^*(\mathbf{z}^{(i)})))$. Similar to composite dual averaging, Frank–Wolfe will accrue the discretization error $a_i(\psi(\mathbf{x}^{(i)}) - \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)})))$.

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

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

Note that the $E_d^{(i)}$'s contain both the integration error and the error due to the discontinuities in the algorithm updates discussed above. These discretization errors will ultimately determine the algorithms' convergence rates: while in the continuous-time domain we could choose $\alpha^{(t)}$ (and $A^{(t)}$) to grow arbitrarily fast as a function of time, in the discrete-time domain the discretization errors $E_d^{(k)}$ will depend on the choice of $A^{(k)}$. In particular, this co-dependence between $A^{(k)}$ and $E_d^{(k)}$ will determine the choice of $A^{(k)}$ leading to the greatest decrease in the bound on $G^{(k)}$ from (4.2) as a function of k.

We are now ready to bound the discretization errors of the algorithms from section 3. Before doing so, we make the following two remarks.

Remark 4.1. The versions of mirror descent and mirror prox presented here are in fact the "lazy" versions of these methods that are known as Nesterov's dual averaging [30]. The "lazy" and standard versions of the methods are equivalent whenever $X = \mathbb{R}^n$. The reason we choose to work with the "lazy" versions of the methods is that they follow more directly from the discretization of the continuous-time dynamics presented in the previous section.

Remark 4.2. It is possible to obtain the discretization error $E_d^{(i)}$ by directly computing $A^{(k)}G^{(k)} - A^{(k-1)}G^{(k-1)}$ for the discrete version of the gap. We have chosen the approach presented here to illustrate the effects of discretization.

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

PROPOSITION 4.3. *The discretization error for* (CT-MD) *is*

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

Proof. In (CT-MD), $\mathbf{x}^{(\tau)} = \nabla\phi^*(\mathbf{z}^{(\tau)})$, and thus

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

Since, for noncomposite 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 methods that lead to discrete-time algorithms known as (lazy) mirror descent and mirror prox.

Forward Euler discretization: Lazy 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 4.3) that in this case the discretization error is given as

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

When $f(\cdot)$ is Lipschitz-continuous, we recover the classical mirror-descent/dual-averaging convergence result [26].

THEOREM 4.4. *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 $i \le k$ and $k \ge 1$. Then, if*

$$a_i = \frac{1}{L}\sqrt{\frac{2\sigma D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)})}{k+1}}$$

and $\phi(\cdot) = D_\psi(\cdot, \mathbf{x}^{(0)})$, we have

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

Proof. By Proposition A.1, $D_{\psi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \ge \frac{\sigma}{2}\|\nabla\psi^*(\mathbf{z}^{(i-1)}) - \nabla\psi^*(\mathbf{z}^{(i)})\|^2 = \frac{\sigma}{2}\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\|^2$. As $D_\phi(\cdot,\cdot) = D_\psi(\cdot,\cdot)$ and $f(\cdot)$ is Lipschitz continuous with parameter L, using the Cauchy–Schwarz inequality, we obtain

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

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

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

Similarly, we can bound the initial gap as

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

Finally, by combining (4.5), (4.6), the choice of a_i's, and $f(\hat{\mathbf{x}}^{(k)}) - f(\mathbf{x}^*) \le G^{(k)}$, 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 the method, where only two fixed-point iteration steps are performed.[6] The resulting discrete-time method is known as mirror prox [25] or extra-gradient descent [21]:

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

This method is typically used for solving variational inequalities with monotone operators [25]. Its convergence bound is provided in Theorem B.2.

4.2. Accelerated smooth minimization. In this subsection and the following one, we will consider only forward Euler discretization of the accelerated dynamics, which corresponds to the Nesterov's accelerated algorithm. Approximate backward Euler discretization using similar ideas to the proof of convergence of the mirror prox from the previous subsection is also possible and leads to the recent accelerated extra-gradient descent (AXGD) algorithm that we presented in [13].

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

PROPOSITION 4.5. *The discretization error for* (CT-AMD) *is*

$$\begin{aligned} E_d^{(i)} &= -a_i\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\rangle + A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) \\ &\quad - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \\ &\le \langle \nabla f(\mathbf{x}^{(i)}), A^{(i)}\mathbf{x}^{(i)} - A^{(i-1)}\mathbf{x}^{(i-1)} - a_i\nabla\phi^*(\mathbf{z}^{(i)})\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \end{aligned} \tag{4.7}$$

[6] This discretization can also be viewed as the predictor-corrector method.

Proof. Recall the 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

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

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

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

where we used $f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)}) \leq \langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i)} - \mathbf{x}^{(i-1)} \rangle$, 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 a 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 the familiar Nesterov accelerated method [27]:

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

where

$$\mathrm{Grad}(\mathbf{x}^{(i)}) = \arg\min_{\mathbf{x} \in X} \left\{ \langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x} - \mathbf{x}^{(i)} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{x}^{(i)}\|^2 \right\}. \tag{4.8}$$

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 (4.7) for the change in the upper bound yields

(4.9)

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

We are now ready to prove the convergence of Nesterov's algorithm for smooth functions [27].

THEOREM 4.6. *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, $\forall k \geq 1$,*

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

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

$$f(\hat{\mathbf{x}}^{(i)}) \leq f(\mathbf{x}^{(i)}) + \min_{\mathbf{x}\in X}\left\{\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x} - \mathbf{x}^{(i)}\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^{(i)}\|^2\right\}. \tag{4.10}$$

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 (4.2) (which correspond to the original discretization error (4.7)) 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 (4.2) to the gradient term in (4.10) 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 (4.10) that

$$\begin{aligned} A^{(i)} f(\hat{\mathbf{x}}^{(i)}) \leq A^{(i)} f(\mathbf{x}^{(i)}) &- a_i\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\rangle \\ &+ \frac{L{a_i}^2}{2A^{(i)}}\|\nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\|^2. \end{aligned} \tag{4.11}$$

By Proposition A.1, $D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}) \geq \frac{\sigma}{2}\|\nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\|^2$. Therefore, for the quadratic term in (4.11) to cancel the remaining term in (4.2), $D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})$, it suffices to have $\frac{{a_i}^2}{A^{(i)}} \leq \frac{\sigma}{L}$. It is easy to verify that $a_i = \frac{\sigma}{L}\frac{i+1}{2}$ in the theorem statement satisfies $\frac{{a_i}^2}{A^{(i)}} \leq \frac{\sigma}{L}$, and thus it follows that $G^{(k)} \leq \frac{a_0 G^{(0)}}{A^{(k)}}$.

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

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

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

4.3. Gradient descent. The discretization error of gradient descent is the same as the discretization error of the accelerated method from the previous subsection (see (4.7)), since the two methods use the same approximate duality gap. Classical gradient descent uses forward Euler discretization, which sets $\mathbf{x}^{(i)} = \mathbf{x}^{(0)} - \mathbf{z}^{(i-1)}/\sigma$. Thus, the algorithm can be stated as

$$\begin{aligned} \mathbf{z}^{(i)} &= \mathbf{z}^{(i-1)} - a_i \nabla f(\mathbf{x}^{(i)}), \\ \mathbf{x}^{(i)} &= \nabla\phi^*(\mathbf{z}^{(i-1)}) = \mathbf{x}^{(0)} + \frac{\mathbf{z}^{(i-1)}}{\sigma}, \\ \mathbf{z}^{(0)} &= 0,\ \mathbf{x}^{(0)} \in \mathbb{R}^n \text{ is an arbitrary initial point.} \end{aligned} \tag{GD}$$

To cancel out the discretization error, we only need to use the fact that gradient steps (corresponding to the steps of the algorithm) reduce the function value, assuming that the function is L-smooth for some $L \in \mathbb{R}_{++}$. This is achieved by setting $U^{(i)} = f(\mathbf{x}^{(i+1)})$. Correcting the discretization error by the change in the upper bound yields the following.

PROPOSITION 4.7. *The discretization error of* (GD) *is*

$$E_d^{(i)} = A^{(i)}(f(\mathbf{x}^{(i+1)}) - f(\mathbf{x}^{(i)})) - a_i\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\rangle - \frac{{a_i}^2}{2\sigma}\|\nabla f(\mathbf{x}^{(i)})\|_2^2.$$

Proof. The proof follows directly by combining (4.7), $U^{(i)} = f(\mathbf{x}^{(i+1)})$, and (GD). ∎

We can now recover the classical bound for gradient descent (see, e.g., [7]).

THEOREM 4.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function for $L \in \mathbb{R}_{++}$, $\phi(\mathbf{x}) = \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^{(0)}\|^2$, and let $\mathbf{x}^{(i)}$ evolve according to* (GD). *If $a_i = \frac{\sigma}{L}$ $\forall i \geq 0$, then, $\forall k \geq 0$,*

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

Proof. By the smoothness of $f(\cdot)$ and (GD),

$$\begin{aligned} E_d^{(i)} &\leq A^{(i-1)}\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\rangle + A^{(i)}\frac{L}{2}\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\|_2^2 - \frac{{a_i}^2}{2\sigma}\|\nabla f(\mathbf{x}^{(i)})\|_2^2 \\ &= \left(-\frac{a_i A^{(i-1)}}{\sigma} + \frac{A^{(i)}{a_i}^2 L}{2\sigma^2} - \frac{{a_i}^2}{2\sigma}\right)\|\nabla f(\mathbf{x}^{(i)})\|_2^2. \end{aligned}$$

By type-checking the last expression, it follows that a_i needs to be proportional to $\frac{\sigma}{L}$ to obtain $E_d^{(i)} \leq 0$. Choose $a_i = \frac{\sigma}{L}$. It remains to bound the initial gap. Observing that, for $a_i = \frac{\sigma}{L}$, $U^{(0)} = f(\mathbf{x}^{(0)}) - \frac{1}{2L}\|\nabla f(\mathbf{x}^{(0)})\|_2^2$ and $L^{(0)} = f(\mathbf{x}^{(0)}) - \frac{1}{2L}\|\nabla f(\mathbf{x}^{(0)})\|_2^2 + \frac{L}{2a_0}\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2^2$, the claimed bound on the gap follows. ∎

4.4. 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

$$\phi_i(\mathbf{x}) = \int_0^i \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^{(\tau)}\|^2 \mathrm{d}\alpha^{(\tau)} + \phi(\mathbf{x}).$$

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 4.9. *The discretization error for* (CT-ASC) *is*

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

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

$$\begin{aligned} I_c^{(i-1,i)} &= -\int_{i-1}^{i} \langle \nabla f(\mathbf{x}^{(\tau)}), \nabla \phi_i^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(\tau)} \rangle \mathrm{d}\alpha^{(\tau)} \\ &= -\int_{i-1}^{i} \langle \nabla f(\mathbf{x}^{(\tau)}), \alpha^{(\tau)} \dot{\mathbf{x}}^{(\tau)} \rangle \mathrm{d}\tau + \int_{i-1}^{i} \langle \dot{\mathbf{z}}^{(\tau)}, \nabla \phi_i^*(\mathbf{z}^{(i)}) \rangle \mathrm{d}\tau \\ &\quad - \int_{i-1}^{i} \langle \dot{\mathbf{z}}^{(\tau)}, \nabla \phi_\tau^*(\mathbf{z}^{(\tau)}) \rangle \mathrm{d}\tau \\ &= -A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) + \langle \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \nabla \phi_i^*(\mathbf{z}^{(i)}) \rangle \\ &\quad - \phi_i^*(\mathbf{z}^{(i)}) + \phi_{i-1}^*(\mathbf{z}^{(i-1)}). \end{aligned} \tag{4.12}$$

Combining (4.1), (4.12), and the fact that $-a_i \nabla f(\mathbf{x}^{(i)}) = \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}$, we have

$$E_d^{(i)} = A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - \langle \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \mathbf{x}^{(i)} \rangle + \phi_i^*(\mathbf{z}^{(i)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}). \tag{4.13}$$

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

$$\begin{aligned} E_d^{(i)} &\leq A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - \langle \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \mathbf{x}^{(i)} - \nabla \phi_{i-1}^*(\mathbf{z}^{(i)}) \rangle \\ &\quad - D_{\phi_{i-1}^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \end{aligned} \tag{4.14}$$

∎

Comparing the discretization error from Proposition 4.9 with the discretization error (4.7) from the previous subsection, we can observe that they take the same form, with the only difference being that ϕ^* is 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

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

while the discretization error for (ASC) becomes

$$\begin{aligned} E_d^{(i)} &\leq A^{(i)}(f(\hat{\mathbf{x}}^{(i)}) - f(\mathbf{x}^{(i)})) + a_i \langle \nabla f(\mathbf{x}^{(i)}), \nabla \phi_{i-1}^*(\mathbf{z}^{(i-1)}) - \nabla \phi_{i-1}^*(\mathbf{z}^{(i)}) \rangle \\ &\quad - D_{\phi_{i-1}^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}). \end{aligned} \tag{4.15}$$

We have the following convergence result.

THEOREM 4.10. *Let $f : X \to \mathbb{R}$ be an L-smooth and σ-strongly convex function, let $\psi : X \to \mathbb{R}$ 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 4.6. 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 $\frac{{a_i}^2}{A^{(i)}} \leq \frac{\sigma_{i-1}}{L}$. 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

$$\frac{{a_i}^2}{(A^{(i)})^2} = \frac{1}{\kappa}\left(1 - \frac{a_i}{A^{(i)}}\right)$$

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 4.11. When $X = \mathbb{R}^n$, we obtain a tighter convergence bound. Namely, assuming that $\|\cdot\| = \|\cdot\|_2$, we can recover the standard guarantee $f(\hat{\mathbf{x}}^{(k)}) - f(\mathbf{x}^*) \leq (1 - \frac{1}{\sqrt{\kappa}})^k D_\psi(\mathbf{x}^*, \mathbf{x}^{(0)})$ [33]. More details can be found in [11, Appendix B].

4.5. Composite dual averaging. Consider the forward Euler discretization of (CT-CMD), recovering updates similar to [16]:[7]

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

Unlike in the standard (noncomposite) 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 $\int_0^t \psi(\mathbf{x}^{(\tau)})\mathrm{d}\alpha^{(\tau)}$ matches the change in $\psi(\nabla\phi_t^*(\mathbf{z}^{(t)}))$. In the discrete-time case, 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 4.12. *The discretization error for forward Euler discretization of* (CT-CMD) *is*

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

[7] Equation (CMD) is the "lazy" (dual-averaging) version of the COMID algorithm from [16].

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

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

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

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

Therefore,

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

On the other hand, as

$$\begin{aligned}
I^{(i-1,i)} &= -a_i\langle \nabla f(\mathbf{x}^{(i)}), \nabla\phi_i^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\rangle \\
&= \langle \mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}, \nabla\phi_i^*(\mathbf{z}^{(i)}) - \nabla\phi_i^*(\mathbf{z}^{(i-1)})\rangle,
\end{aligned}$$

the discretization error is

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

It remains to show that $\phi_i^*(\mathbf{z}^{(i-1)}) - \phi_{i-1}^*(\mathbf{z}^{(i-1)}) + a_i\psi(\nabla\phi_i^*(\mathbf{z}^{(i-1)})) \le 0$. 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, we get

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

where the inequality is obtained by Fact 1.6, as

$$\nabla\phi_{i-1}(\mathbf{z}^{(i-1)}) = \arg\min_{\mathbf{x}\in X}\{-\langle \mathbf{z}^{(i-1)}, \mathbf{x}\rangle + \phi_{i-1}(\mathbf{x})\}. \qquad \square$$

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

THEOREM 4.13. *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*

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

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 4.4 (dual-averaging convergence), as $D_{\phi_i^*}(\mathbf{z}^{(i)}, \mathbf{z}^{(i-1)}) = -a_i\langle \nabla f(\mathbf{x}^{(i)}), \mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\rangle - D_{\phi_i^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})$, and is omitted. ∎

Remark 4.14. 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 larger steps a_i to be chosen.

4.6. Frank–Wolfe method. For the discretization of (CT-FW), we need to take into account the different lower bound we obtained in (3.3). In particular, the integral that accrues a discretization error is

$$-\int_{i-1}^{i} \langle \nabla f(\mathbf{x}^{(\tau)}), \nabla\psi^*(\hat{\mathbf{z}}^{(\tau)}) - \mathbf{x}^{(\tau)}\rangle \mathrm{d}\alpha^{(\tau)},$$

where $\hat{\mathbf{z}}^{(\tau)} = -\nabla f(\mathbf{x}^{(\tau)})$. The forward Euler discretization gives the following algorithm:

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

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 4.15. *The discretization error for forward Euler discretization of* (CT-FW) *is*

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

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

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

Therefore,

$$\begin{aligned} E_d^{(i)} &= A^{(i-1)}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})) - a_i\langle \nabla f(\mathbf{x}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \mathbf{x}^{(i)}\rangle \\ &\quad + a_i(\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})) - \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)}))) \\ &\le a_i\langle \nabla f(\mathbf{x}^{(i)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i-1)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\rangle \\ &\quad + a_i(\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})) - \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)}))), \end{aligned} \tag{4.16}$$

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

$$-\langle \hat{\mathbf{z}}^{(i-1)}, \nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})\rangle + \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})) \le -\langle \hat{\mathbf{z}}^{(i-1)}, \nabla\psi^*(\hat{\mathbf{z}}^{(i)})\rangle + \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)})),$$

and, therefore, as $\hat{\mathbf{z}}^{(i-1)} = -\nabla f(\mathbf{x}^{(i-1)})$,

$$\psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})) - \psi(\nabla\psi^*(\hat{\mathbf{z}}^{(i)})) \le \langle \nabla f(\mathbf{x}^{(i-1)}), \nabla\psi^*(\hat{\mathbf{z}}^{(i)}) - \nabla\psi^*(\hat{\mathbf{z}}^{(i-1)})\rangle. \tag{4.17}$$

Combining (4.16) and (4.17), the claimed bound on discretization error follows. ∎

We can now recover the convergence result from [32].

THEOREM 4.16. *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 $L_\nu < \infty$, $\nu \in (0, 1]$,*[8]

$$\|\nabla f(\mathbf{x}) - \nabla f(\hat{\mathbf{x}})\|_* \le L_\nu \|\mathbf{x} - \hat{\mathbf{x}}\|^\nu \quad \forall \mathbf{x}, \hat{\mathbf{x}} \in X. \tag{4.18}$$

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, $\forall k \ge 1$,*

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

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

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

Proof. Applying the Cauchy–Schwarz inequality to the discretization error given by Proposition 4.15, we have

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

where the second inequality follows from (4.18) 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, $G^{(k)} \le \frac{a_0 G^{(0)}}{A^{(k)}} + L_\nu D^{1+\nu} \frac{1}{A^{(k)}} \sum_{i=1}^{k} \frac{{a_i}^{1+\nu}}{(A^{(i)})^\nu}$.

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

We now use the same arguments 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

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

Therefore, $G^{(k)} \leq L_\nu D^{1+\nu} \frac{1}{A^{(k)}} \sum_{i=0}^{k} \frac{{a_i}^{1+\nu}}{(A^{(i)})^\nu}$. In particular, if $a_i = i+1$, then $A^{(i)} = \frac{(i+1)(i+2)}{2}$. Finally,

$$\sum_{i=0}^{k} \frac{{a_i}^{1+\nu}}{(A^{(i)})^\nu} < 2^\nu \sum_{i=0}^{k} (i+1)^{1-\nu} < 2^\nu (k+1)^{2-\nu},$$

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 the approximate duality gap at a rate equal to the convergence rate. Besides the unified interpretation of many first-order methods, the technique is generally useful for obtaining new optimization methods [13, 12, 10, 9, 14]. An interesting direction for the future is extending this technique to other settings, such as geodesically convex optimization.

Appendix A. Properties of the Bregman divergence. The following properties of Bregman divergence are useful in our analysis.

PROPOSITION A.1. *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. From the definition of ψ^* and Fact 1.6,

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

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

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

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.2. *For all* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in X$,

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

Appendix B. Extension of ADGT to 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 the following:

$$\text{find } \mathbf{x}_\epsilon \in X \text{ such that } \langle F(\mathbf{u}), \mathbf{x}_\epsilon - \mathbf{u}\rangle \leq \epsilon \quad \forall \mathbf{u} \in X, \tag{B.1}$$

and we can think of ϵ on the right-hand side of (B.1) as the optimality gap.

The property of monotone operators $F(\cdot)$ useful for the approximate gap analysis 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

$$\hat{\mathbf{x}}^{(t)} = \frac{1}{\alpha^{(t)}} \int_0^t \mathbf{x}^{(\tau)} \mathrm{d}\alpha + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}} \mathbf{x}^{(0)},$$

gives, $\forall \mathbf{u} \in X$,

(B.2)

$$\begin{aligned}\langle F(\mathbf{u}), \hat{\mathbf{x}}^{(t)} - \mathbf{u}\rangle &\leq G^{(t)} \\ &\stackrel{\text{def}}{=} \frac{\max_{\mathbf{x}\in X}\{\int_0^t \langle F(\mathbf{x}^{(\tau)}), \mathbf{x} - \mathbf{x}^{(\tau)}\rangle \mathrm{d}\alpha + \phi(\mathbf{x})\} - \max_{\mathbf{u}\in X} \phi(\mathbf{u})}{\alpha^{(t)}} \\ &\quad + \frac{\alpha^{(t)} - A^{(t)}}{\alpha^{(t)}} \max_{\mathbf{v}\in X} \langle F(\mathbf{v}), \mathbf{x}^{(0)} - \mathbf{v}\rangle.\end{aligned}$$

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 $\mathbf{w}, \mathbf{w}^{(\tau)} \in Z$,

$$\Phi(\mathbf{v}, \mathbf{w}^{(\tau)}) - \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}) \geq \langle \nabla_{\mathbf{v}} \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}), \mathbf{v} - \mathbf{v}^{(\tau)}\rangle, \tag{B.3}$$

$$\Phi(\mathbf{v}^{(\tau)}, \mathbf{w}) - \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}) \leq \langle \nabla_{\mathbf{w}} \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}), \mathbf{w} - \mathbf{w}^{(\tau)}\rangle, \tag{B.4}$$

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

Combining (B.3) and (B.4), it follows that, $\forall \mathbf{v}, \mathbf{v}^{(\tau)} \in Y$, $\forall \mathbf{w}, \mathbf{w}^{(\tau)} \in Z$,

$$\begin{aligned}&\Phi(\mathbf{v}^{(\tau)}, \mathbf{w}) - \Phi(\mathbf{v}, \mathbf{w}^{(\tau)}) \\ &\quad \leq \langle \nabla_{\mathbf{v}} \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}), \mathbf{v}^{(\tau)} - \mathbf{v}\rangle - \langle \nabla_{\mathbf{w}} \Phi(\mathbf{v}^{(\tau)}, \mathbf{w}^{(\tau)}), \mathbf{w} - \mathbf{w}^{(\tau)}\rangle.\end{aligned}$$

Let $\mathbf{x} = [\mathbf{v}, \mathbf{w}]^T$, $F(\mathbf{x}) = [\nabla_{\mathbf{v}} \Phi(\mathbf{v}, \mathbf{w}), -\nabla_{\mathbf{w}} \Phi(\mathbf{v}, \mathbf{w})]^T$, and $\overline{\mathbf{v}} = \frac{1}{A^{(t)}} \int_0^t \mathbf{v}^{(\tau)} \mathrm{d}\alpha^{(\tau)}$, $\overline{\mathbf{w}} = \frac{1}{A^{(t)}} \int_0^t \mathbf{w}^{(\tau)} \mathrm{d}\alpha^{(\tau)}$. Then, we have, $\forall \mathbf{x} = [\mathbf{v}, \mathbf{w}]^T \in V \times W$,

$$\Phi(\overline{\mathbf{v}}, \mathbf{w}) - \Phi(\mathbf{v}, \overline{\mathbf{w}}) \leq \frac{\int_0^t \langle F(\mathbf{x}), \mathbf{x}^{(\tau)} - \mathbf{x}\rangle \mathrm{d}\alpha^{(\tau)}}{A^{(t)}},$$

and, using the same arguments as before, we obtain the same bound for the gap as (B.2). Therefore, we can focus on analyzing the decrease of $G^{(t)}$ from (B.2) as a function of t and the same result will follow for the gap of convex-concave saddle-point problems.

B.1. Continuous-time mirror descent. Replacing $\nabla f(\mathbf{x}^{(t)})$ by $F(\mathbf{x}^{(t)})$ in the gap for mirror descent for convex minimization from section 3.1, it follows that (CT-MD) also ensures $\frac{\mathrm{d}}{\mathrm{d}t}(\alpha^{(t)}G^{(t)}) = 0$ for the gap (B.2) derived for monotone operators and saddle-point problems. Hence, we have the following lemma.

LEMMA B.1. *Suppose 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, $\forall t > 0$, $\forall \mathbf{u} \in X$,*

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

Moreover, for a convex-concave saddle-point problem $\min_{\mathbf{v}\in V} \max_{\mathbf{w}\in W} \Phi(\mathbf{v}, \mathbf{w})$, taking $\mathbf{x} = [\mathbf{v}, \mathbf{w}]^T$, $F(\mathbf{x}) = [\nabla_{\mathbf{v}}\Phi(\mathbf{v}, \mathbf{w}), -\nabla_{\mathbf{w}}\Phi(\mathbf{v}, \mathbf{w})]^T$, the version of (CT-MD) *that uses the monotone operator $F(\mathbf{x})$ ensures that, $\forall t > 0$, $\forall (\mathbf{v}, \mathbf{w}) \in V \times W$,*

$$\begin{aligned}&\Phi(\hat{\mathbf{v}}^{(t)}, \mathbf{w}) - \Phi(\mathbf{v}, \hat{\mathbf{w}}^{(t)})\\ &\qquad \leq \frac{\max_{\mathbf{x}'\in X} \phi(\mathbf{x}') + \alpha^{(0)} \max_{\mathbf{v}''\in V, \mathbf{w}''\in W} \{\Phi(\hat{\mathbf{v}}^{(0)}, \mathbf{w}'') - \Phi(\mathbf{v}'', \hat{\mathbf{w}}^{(0)})\}}{\alpha^{(t)}}.\end{aligned}$$

B.2. Lazy mirror prox. As discussed in section 4, a lazy version of the mirror-prox method [25] follows as an approximate backward Euler (or predictor-corrector) discretization of the mirror-descent dynamics (CT-MD) stated in (MP).

THEOREM B.2. *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 $\forall \mathbf{u} \in X$,*

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

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

$$\begin{aligned}E_d^{(i)} &= -a_i\langle F(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})\\ &= a_i\langle F(\tilde{\mathbf{x}}^{(i-1)}) - F(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\rangle\\ &\quad - a_i\langle F(\tilde{\mathbf{x}}^{(i-1)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)}).\end{aligned} \tag{B.5}$$

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.2 implies

$$\begin{aligned}&-a_i\langle F(\tilde{\mathbf{x}}^{(i-1)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\rangle - D_{\phi^*}(\mathbf{z}^{(i-1)}, \mathbf{z}^{(i)})\\ &\quad = -D_{\phi^*}(\mathbf{z}^{(i-1)}, \tilde{\mathbf{z}}^{(i-1)}) - D_{\phi^*}(\tilde{\mathbf{z}}^{(i-1)}, \mathbf{z}^{(i)})\\ &\quad \leq -\frac{\sigma}{2}(\|\nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)})\|^2 + \|\nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\|^2),\end{aligned} \tag{B.6}$$

where the inequality is given by Proposition A.1.

On the other hand, by the L-smoothness of $F(\cdot)$, using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}&a_i\langle F(\tilde{\mathbf{x}}^{(i-1)}) - F(\mathbf{x}^{(i)}), \nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\rangle\\ &\quad \leq a_i L\|\tilde{\mathbf{x}}^{(i-1)} - \mathbf{x}^{(i)}\| \cdot \|\nabla\phi^*(\mathbf{z}^{(i)}) - \mathbf{x}^{(i)}\|\\ &\quad = a_i L\|\nabla\phi^*(\mathbf{z}^{(i-1)}) - \nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)})\| \cdot \|\nabla\phi^*(\tilde{\mathbf{z}}^{(i-1)}) - \nabla\phi^*(\mathbf{z}^{(i)})\|.\end{aligned} \tag{B.7}$$

As $a_i = \sigma/L$, combining (B.5)–(B.7) 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, starting 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 $D_\phi(\cdot, \cdot) = D_\psi(\cdot, \cdot)$, we obtain

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

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 $a_1 G^{(1)} \leq D_\psi(\mathbf{x}^*, \tilde{\mathbf{x}}^{(0)})$. ∎

As before, as convex-concave saddle-point problems have the same gap as variational inequalities, it is straightforward to extend Theorem B.2 to this setting (see [25]).

Acknowledgments. We thank the anonymous reviewers for their thoughtful comments and suggestions, which greatly improved the presentation of this paper. We also thank Ziye Tang for pointing out several typos in the earlier version of the paper, and providing useful suggestions for improving its presentation.

REFERENCES

[1] Z. Allen-Zhu and L. Orecchia, *Linear coupling: An ultimate unification of gradient and mirror descent*, in Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), LIPIcs. Leibniz Int. Proc. Inform. 67, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Wadern, 2017.

[2] N. Bansal and A. Gupta, *Potential-function Proofs for First-order Methods*, 2017, preprint, https://arxiv.org/abs/1712.04581.

[3] H. H. Bauschke and P. L. Combettes, eds., *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books Math. 408, Springer, Berlin, 2011.

[4] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS-SIAM Ser. Optim, SIAM, Philadelphia, PA, 2001.

[5] D. P. Bertsekas, *Control of Uncertain Systems with a Set-membership Description of the Uncertainty*, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1971.

[6] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Nashua, NH, 2003.

[7] S. Bubeck, *Theory of Convex Optimization for Machine Learning*, preprint, https://arxiv.org/abs/1405.4980v1, 2014.

[8] S. Bubeck, Y. T. Lee, and M. Singh, *A geometric alternative to Nesterov's accelerated gradient descent*, preprint, https://arxiv.org/abs/1506.08187, 2015.

[9] M. B. Cohen, J. Diakonikolas, and L. Orecchia, *On acceleration with noise-corrupted gradients*, in Proceedings of the 35th International Conference on Machine Learning, Proc. Mach. Learn. Res. 80, 2018, pp. 1019–1028; available at http://proceedings.mlr.press/v80/.

[10] J. Diakonikolas, M. Fazel, and L. Orecchia, *Width Independence beyond Linear Objectives: Distributed Fair Packing and Covering Algorithms*, preprint, https://arxiv.org/abs/1808.02517, 2018.

[11] J. Diakonikolas and L. Orecchia, *The Approximate Duality Gap Technique: A Unified Theory of First-Order Methods*, preprint, https://arxiv.org/abs/1712.02485, 2017.

[12] J. Diakonikolas and L. Orecchia, *Solving Packing and Covering Linear Programs in $\tilde{O}(\epsilon^{-2})$ Distributed Iterations with a Single Algorithm and Simpler Analysis*, preprint, https://arxiv.org/abs/1710.09002, 2017.

[13] J. Diakonikolas and L. Orecchia, *Accelerated extra-gradient descent: A novel, accelerated first-order method*, in Proceedings of the 9th Innovations in Theoretical Computer Science Conference (ITCS 2018), LIPIcs. Leibniz Int. Proc. Inform. 94, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Wadern, 2018.

[14] J. Diakonikolas and L. Orecchia, *Alternating randomized block coordinate descent*, in Proceedings of the 35th International Conference on Machine Learning, Proc. Mach. Learn. Res. 80, 2018, pp. 1224–1232; available at http://proceedings.mlr.press/v80/.

[15] D. Drusvyatskiy, M. Fazel, and S. Roy, *An optimal first order method based on optimal quadratic averaging*, SIAM J. Optim., 28 (2018), pp. 251–271.

[16] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, *Composite objective mirror descent*, in Proceedings of 23rd Annual Conference on Learning Theory, Omnipress, Madison, WI, 2010; available at http://www.learningtheory.org/colt2010/papers.html.

[17] A. Ene and H. L. Nguyen, *Constrained submodular maximization: Beyond* $1/e$, in 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), IEEE Press, Piscataway, NJ, 2016, pp. 248–257.

[18] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist., 3 (1956), pp. 95–110.

[19] J. A. Kelner, Y. T. Lee, L. Orecchia, and A. Sidford, *An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations*, in Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, PA, 2014, pp. 217–226.

[20] J. A. Kelner, L. Orecchia, A. Sidford, and Z. A. Zhu, *A simple, combinatorial algorithm for solving SDD systems in nearly-linear time*, in Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, New York, 2013, pp. 911–920.

[21] G. M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matekon: Transl. Russ. East Eur. Math. Econ., 13 (1977), pp. 35–49.

[22] W. Krichene, A. Bayen, and P. L. Bartlett, *Accelerated mirror descent in continuous and discrete time*, in Adv. Neural Inf. Process. Syst. 28, Curran Associates, Red Hook, NY, 2015, pp. 2845–2853.

[23] Y. T. Lee, S. Rao, and N. Srivastava, *A new approach to computing maximum flows using electrical flows*, in Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, New York, 2013, pp. 755–764.

[24] H. Lin, J. Mairal, and Z. Harchaoui, *A universal catalyst for first-order optimization*, in Adv. Neural Inf. Process. Syst. 28, Curran Associates, Red Hook, NY, 2015, pp. 3384–3392.

[25] 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. Optim., 15 (2004), pp. 229–251.

[26] A. Nemirovskii and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley, New York, 1983.

[27] Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Dokl. Akad. Nauk, 269 (1983), pp. 543–547.

[28] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

[29] Y. Nesterov, *Dual extrapolation and its applications to solving variational inequalities and related problems*, Math. Program., 109 (2007), pp. 319–344.

[30] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., 120 (2009), pp. 221–259.

[31] Y. Nesterov, *Universal gradient methods for convex optimization problems*, Math. Program., 152 (2015), pp. 381–404.

[32] Y. Nesterov, *Complexity bounds for primal-dual methods minimizing the model of objective function*, Math. Program., 171 (2018), pp. 311–330.

[33] Y. Nesterov, *Lectures on Convex Optimization*, Springer, Berlin, 2018.

[34] D. Scieur, V. Roulet, F. Bach, and A. D'Aspremont, *Integration methods and accelerated optimization algorithms*, in Adv. Neural Inf. Process. Syst. 30, Curran Associates, Red Hook, NY, 2017, pp. 1109–1118.

[35] J. Sherman, *Nearly maximum flows in nearly linear time*, in Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, IEEE Press, Piscataway, NJ, 2013, pp. 263–269.

[36] D. A. Spielman and S.-H. Teng, *Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems*, in Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, New York, 2004, pp. 81–90.

[37] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for Machine Learning*, MIT Press, Cambridge, MA, 2012.

[38] 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 (2016), pp. 1–43.
[39] P. Tseng, *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, preprint, https://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf, 2008.
[40] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. E7351–E7358.
[41] A. C. Wilson, B. Recht, and M. I. Jordan, *A Lyapunov Analysis of Momentum Methods in Optimization*, preprint, https://arxiv.org/abs/1611.02635, 2016.