A Lyapunov Analysis of Momentum Methods in Optimization

Ashia C. Wilson   Benjamin Recht   Michael I. Jordan

November 8, 2016

Abstract

Momentum methods play a central role in optimization. Several momentum methods are provably optimal, and all use a technique called *estimate sequences* to analyze their convergence properties. The technique of estimate sequences has long been considered difficult to understand, leading many researchers to generate alternative, "more intuitive" methods and analyses. In this paper we show there is an equivalence between the technique of estimate sequences and a family of Lyapunov functions in both continuous and discrete time. This framework allows us to develop a simple and unified analysis of many existing momentum algorithms, introduce several new algorithms, and most importantly, strengthen the connection between algorithms and continuous-time dynamical systems.

1   Introduction

Momentum is a powerful heuristic for accelerating the convergence of optimization methods. One can intuitively "add momentum" to a method by adding to the current step a weighted version of the previous step, encouraging the method to move along search directions that had been previously seen to be fruitful. Such methods were first popularized by Polyak [23], and have been employed in many practical optimization solvers. In particular, since the 1980s, momentum methods have been popular in neural networks as a way to accelerate the backpropagation algorithm. The intuition is that momentum allows local search to avoid "long ravines" and "sharp curvatures" in the sublevel sets of cost functions [24].

Polyak motivated momentum methods by an analogy to a "heavy ball" moving in a potential well defined by the cost function. However, Polyak's physical intuition was exceedingly difficult to make rigorous. For quadratic costs, Polyak was able to provide an eigenvalue argument that showed that his Heavy Ball method required no more iterations than the method of conjugate gradients [23]. Indeed, when applied to positive-definite quadratic cost functions, Polyak's Heavy Ball method is equivalent to Chebyshev's Iterative Method [5]. Despite its intuitive elegance, however, Polyak's eigenvalue analysis does not apply globally for general convex cost functions. Indeed, Lessard *et al.* derived a simple one-dimensional counterexample where the standard Heavy Ball method does not converge [13].

In order to make momentum methods rigorous, a different approach was required. In celebrated work, Nesterov devised a general scheme to accelerate convex optimization methods, achieving optimal running times under oracle models in convex programming [16]. However, to achieve such general applicability, Nesterov's proof techniques abandoned the physical intuition of Polyak [16];

in lieu of differential equations and potential functions, Nesterov devised the method of *estimate sequences* to verify the correctness of these momentum-based methods. Researchers have struggled to understand the intuition and underpinnings of the estimate sequence methodology since Nesterov's initial papers. The associated proof techniques are often viewed as an "algebraic trick."

To overcome the lack of intuition for the estimate sequence technique, several authors have recently proposed schemes to achieve acceleration without appealing to estimate sequences [8, 3, 13]. One promising general approach to the analysis of acceleration has been to analyze the continuous-time limit of accelerated methods and to show that the stability of the resulting ODEs can be verified via a Lyapunov function [25, 12, 28]. However, these methods stop short of providing principles for deriving a discrete-time optimization algorithm from a continuous-time ODE. There are many ways to discretize ODEs, but not all of them give rise to convergent methods or to acceleration. Indeed, for unconstrained optimization on Euclidean spaces in the setting where the objective is strongly convex, Polyak's Heavy Ball method and Nesterov's accelerated gradient descent have the same continuous-time limit.

In this paper, we propose a bridge between continuous-time limits and discrete-time algorithms. Our method takes, as its primary object of study, a general Lyapunov function for momentum methods in continuous time. Through a diverse set of examples, we demonstrate how the proofs of convergence of momentum methods can be understood as bounding discretization errors of the Lyapunov function when moving to discrete time. In particular, we show how the discretization of the associated continuous-time ODE needs to be performed in such a way such that the Lyapunov function remains valid when transitioning to discrete time. Using this technique, we provide a simple, direct proof of the convergence of Nesterov's method for strongly convex functions in Euclidean spaces. Our proof makes clear why only certain parameterizations of the continuous ODEs lead to valid discrete-time methods. Moreover, we provide new perspectives on the need for the extragradient step inside Nesterov's method which does not appear in Polyak's method. Finally, we show there is a striking equivalence between estimate sequences and Lyapunov functions. In continuous and discrete time, estimate sequences can be derived directly from the Lyapunov function and vice versa. We show that the associated continuous-time estimate sequence can also be directly discretized to give standard estimate-sequence based proofs of momentum methods. This clarifies how Nesterov's "algebraic trick" is closely related to invariant reduction proofs that are common in algorithm analysis.

The paper proceeds as follows. We first describe the related work in viewing optimization algorithms as discretizations of continuous-time dynamical systems. We then introduce the Lyapunov functions that we will use throughout the paper to demonstrate convergence of algorithms. In particular, we will introduce time-varying Lyapunov functions that not only establish algorithm convergence but additionally verify a *convergence rate*. We then turn to various momentum-based methods, providing a general Lyapunov analysis for Nesterov's accelerated method, the quasi-monotone subgradient method, the conditional gradient method, and a few novel algorithms. We pay particular attention to the strongly convex case in Euclidean space, highlighting some unique properties of this setup. Finally, we describe the connection between estimate sequences and Lyapunov functions and discuss directions for future investigation.

We try to include as many of the mathematical derivations as possible in the main body of the paper. Indeed, all proofs will follow a similar flow of analysis using a discretized Lyapunov function. However, as is often the case, the specifics of individual algorithms often require involved algebraic calculations to fully verify. In the spirit of reproducibility, we provide detailed proofs of

convergence rates of all of the algorithms we discuss in the Appendix.

2   The Dynamical View of Momentum Methods

Problem setting   This paper is concerned with the class of constrained optimization problems

$$\min_{x \in \mathcal{X}} f(x), \tag{1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed convex set and $f \colon \mathcal{X} \to \mathbb{R}$ is a continuously differentiable convex function. We use the standard Euclidean norm $\|x\| = \langle x, x \rangle^{1/2}$ throughout. We consider the general non-Euclidean setting in which the space $\mathcal{X}$ is endowed with a distance-generating function $h \colon \mathcal{X} \to \mathbb{R}$ that is convex and essentially smooth (i.e., h is continuously differentiable in $\mathcal{X}$, and $\|\nabla h(x)\|_* \to \infty$ as $\|x\| \to \infty$). The function h can be used to define a measure of distance in $\mathcal{X}$ via its Bregman divergence:

$$D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle,$$

which is nonnegative since h is convex. The *Euclidean setting* is obtained when $h(x) = \frac{1}{2}\|x\|^2$.

We denote a discrete-time sequence in lower case, e.g., x_k with $k \geq 0$ an integer. We denote a continuous-time curve in upper case, e.g., X_t with $t \in \mathbb{R}$. An over-dot means derivative with respect to time, i.e., $\dot{X}_t = \frac{d}{dt} X_t$.

Momentum methods   The intuition for solving problem (1) with momentum methods comes from viewing optimization algorithms as discrete-time approximations to ODEs. In particular, the second-order ODE:

$$m\frac{d^2 x}{dt^2} = -\nabla f(x) - m\dot{\gamma}_t \frac{dx}{dt} \tag{2}$$

corresponds to the equations of motion of an object with mass m in the presence of friction or viscosity, captured by the derivative of a function γ_t, and a potential f. The trajectories of this dynamical system will tend to have directions of motion that persist over time; however, for some values of γ_t, the dynamics will converge to the minimum of f. In the limit as the mass goes to zero, and letting $\gamma_t \propto t/m$, we recover the gradient-flow dynamics, which is the continuous-time limit of gradient descent [28]. Thus, the mass term here serves to *accelerate* the progress towards the bottom of the well.

In seminal work, Polyak defined and analyzed the Heavy Ball method as an application of Euler's method to the dynamical system (2), when γ_t is linear. He was able to show *local convergence* of this method, but was unable to prove global convergence. Moreover, his local convergence result required the assumption that f is strongly convex in the neighborhood of the associated stationary point.

Until recently, this ODE perspective on optimization algorithms was largely abandoned in favor of Nesterov's estimate sequence framework. However, there has been a recent resurgence in interest in ODE-based analyses. Su, Boyd and Candès [25], showed that the continuous-time limit of Nesterov's accelerated gradient descent method corresponds to the following second-order ODE,

$$\ddot{X}_t + \frac{r}{t}\dot{X}_t + \nabla f(X_t) = 0, \quad r \geq 3. \tag{3}$$

3

This is equivalent to (2), where the object being modeled has unit mass and the damping term is logarithmic, $\gamma_t = r \log t$. They also established a convergence rate of $O(1/t^2)$ for this continuous-time system via a Lyapunov function argument. This matches the $O(1/\epsilon k^2)$ convergence rate of accelerated gradient descent, given the discretization $t = \sqrt{\epsilon} k$. Thus, the convergence rate is maintained in passing from continuous time to discrete time. Two things remain unclear from their analysis however: they introduce an additional sequence $\{y_k\}_{k=1}^\infty$ while discretizing the ODE [28, Sec 2], which does not correspond to a straightforward discretization technique; and they do not demonstrate what connection—if any—there is between the Lyapunov argument used to analyze the continuous-time dynamics and the technique of estimate sequences used to analyze the discrete-time algorithm.

Krichene, Bartlett and Bayen [12] derived a modified accelerated mirror descent algorithm using a "discretized" Lyapunov function. However, their proof targeted a different algorithm than the algorithm in Nesterov's original work [17]. In particular, their analysis entailed an additional smoothness assumption which is not necessary for the original algorithm [17], or for the original proof technique using estimate sequences.[1] See also Nemirovski and Yudin's original work [15] for a continuous-time perspective on mirror descent and a Lyapunov analysis of convergence.

Wibisono, Wilson, and Jordan [28] introduced a class of dynamical systems whose discretizations give rise to a general family of accelerated gradient algorithms, including accelerated gradient and mirror descent. These algorithms are obtained as a discretization of the following second-order ODE:

$$\frac{d}{dt} \nabla h(X_t + e^{-\alpha_t} \dot{X}_t) = -e^{\alpha_t + \beta_t} \nabla f(X_t),\tag{4}$$

where h is a strongly convex function associated with the geometry of the space and where α_t and β_t are smooth functions of time. In the case where $\alpha_t = -\log m$, and $\beta_t = \log m$, the continuous-time limit $m \to 0$ of this ODE recovers the mirror descent dynamics [28], with h playing the role of the distance-generating function. Indeed, the original motivation for mirror descent by Nemirovski and Yudin was from this continuous perspective, and they demonstrated a Lyapunov argument for the convergence of the continuous-time dynamics (see Section 3.1 of Nemirovski and Yudin [15]).

Wibisono *et al.* [28] imposed the following conditions on the functions α_t, β_t, and γ_t:

$$\dot{\beta}_t \leq e^{\alpha_t}\tag{5a}$$
$$\dot{\gamma}_t = e^{\alpha_t},\tag{5b}$$

referring to these conditions as *ideal scaling*, and showing their necessity for achieving accelerated rates. These conditions will also play an important role in the present work. Wibisono *et al.* [28] also require that f and h be differentiable and convex, and that h be strictly convex. Under the ideal scaling, they demonstrated that the momentum ODE (4) is the Euler-Lagrange equation for a functional that they referred to as the *Bregman Lagrangian*. This means that the family of accelerated gradient dynamics—like the gradient flows which generate gradient methods—can be understood as being generated from a *variational principle*. They also presented a Lyapunov function to demonstrate a convergence rate of $O(e^{-\beta_t})$ for the momentum dynamics (4). However, they did not provide a Lyapunov analysis of the discrete-time algorithm, nor did they make explicit the connection between the Lyapunov function and the estimate sequence technique.

---

[1]In particular, they assume $\frac{\ell}{2}\|x-y\|^2 \leq D_h(x,y) \leq \frac{L}{2}\|x-y\|^2$. This additional assumption greatly simplifies the proof. However, only the assumption of an upper bound is needed for the most general form of accelerated methods is that $\frac{\ell}{2}\|x-y\|^2 \leq D_h(x,y)$.

3 Rate-Generating Lyapunov Functions

Lyapunov's method [14] is based on the idea of constructing a positive definite quantity $\mathcal{E} : \mathcal{X} \to \mathbb{R}$ which decreases along the trajectories of the dynamical system $\dot{X}_t = v(X_t)$:

$$\dot{\mathcal{E}}(X_t) = \langle \nabla \mathcal{E}(X_t), v(X_t) \rangle < 0.$$

The existence of such a *Lyapunov function* guarantees that the dynamical system converges: if the function is positive yet strictly decreasing along all trajectories, then the dynamical system must eventually approach a region where $\mathcal{E}(X)$ is minimal. If this region coincides with the stationary points of the dynamics, then all trajectories must converge to a stationary point.

In this paper, we are interested in obtaining Lyapunov functions for dynamical systems designed for optimization. In optimization, we are typically interested in the *rate of convergence* of a method, not mere convergence. To facilitate such rate analysis, we study time-dependent Lyapunov functions. Though it is unconventional to have Lyapunov functions depend explicitly on time, our Lyapunov function is parameterized explicitly by α_t, β_t, and γ_t, reflecting the dependence of the dynamics (4) on these functions.

The family of dynamics that obey the ideal scaling constraints (5) is invariant under time reparameterizations. Such time-invariance properties provide more flexibility in algorithm design and analysis, as we will see throughout this paper.

We begin by presenting time-varying Lyapunov functions that can verify the convergence rates of continuous-time momentum flows, and we then turn to the corresponding discrete-time analysis.

General accelerated mirror descent   In Wibisono *et al.* [28], the following Lyapunov function:

$$\mathcal{E}_t = e^{\beta_t}(f(X_t) - f(x^*)) + D_h\left(x^*, X_t + e^{-\alpha_t}\dot{X}_t\right), \tag{6}$$

was introduced to analyze the flow that forms the basis for accelerated mirror descent. In Section 4, we show how this Lyapunov function emerges from a general framework for analyzing momentum dynamics; furthermore, we show how to discretize the Lyapunov function and thereby obtain a tool which can be used to analyze a variety of algorithms that are discretizations of momentum dynamics.

Accelerated gradient descent (strong convexity)   In this paper, we show that the following function

$$\mathcal{E}_t = e^{\beta_t}\left(f(X_t) - f(x^*) + \mu D_h\left(x^*, X_t + e^{-\alpha_t}\dot{X}_t\right)\right), \tag{7}$$

is a Lyapunov function for a family of momentum dynamics when f is μ-strongly convex ($D_f(y,x) \geq \mu D_h(y,x)$) and $e^{\alpha_t} = \dot{\beta}_t$. In Section 5, we derive this Lyapunov function and show how it can be used to demonstrate a linear rate of convergence for the family of accelerated gradient descent dynamics. (For simplicity, we focus on the setting in which $h = \frac{1}{2}\|\cdot\|^2$ is Euclidean.) We also show that a particular discretization of the family of dynamics provides an algorithm for which a commensurate discretization of (7) is a Lyapunov function.

Note that in both the strongly convex and weakly convex setting, the structure of the Lyapunov function allows us to derive a rate of convergence for the optimality gap $f(X_t) - f(x^*)$. In particular,

the Lyapunov property $\mathcal{E}_t \leq \mathcal{E}_0$ implies

$$f(X_t) - f(x^*) \leq \frac{e^{\beta_0}(f(X_0) - f(x^*)) + D_h(x^*, X_0 + e^{-\alpha_0}\dot{X}_0)}{e^{\beta_t}}, \tag{8}$$

and

$$f(X_t) - f(x^*) \leq \frac{e^{\beta_0}(f(X_0) - f(x^*) + \mu D_h(x^*, X_0 - e^{-\alpha_0}\dot{X}_0))}{e^{\beta_t}}, \tag{9}$$

for (6) and (7) respectively. The typical convention is that the dynamics start from rest $\dot{X}_0 = 0$. Notice that we obtain the same rates of convergence in these two settings. However, the family of dynamics corresponding to the Lyapunov function in the strongly convex setting exploits the strong convexity of f, and this allows us to obtain a tighter bound in discrete time.

4   Momentum Methods

In this section, we introduce two families of momentum dynamics, both which are generated by functionals defined on continuous-time curves. We illustrate how to derive a Lyapunov function from these dynamics. Next, we show how these dynamics can be discretized to obtain optimization algorithms. The same discrete-time mappings can be applied to the Lyapunov function, resulting in a tool which can be used to analyze the discrete-time algorithms. We provide examples of this procedure for the quasi-monotone subgradient algorithm, accelerated gradient descent, the conditional gradient algorithm, and several other novel algorithms.

4.1   Lyapunov Analysis

We study dynamics that are based on the Euler-Lagrange equation (4) in the setting where the ideal scaling (5a) holds with equality $\dot{\beta}_t = e^{\alpha_t}$. This implies that the Bregman Lagrangian is parameterized by β_t, which determines the convergence rate:

$$Z_t = X_t + \frac{1}{\dot{\beta}_t}\dot{X}_t, \tag{10a}$$

$$\frac{d}{dt}\nabla h(Z_t) = -\dot{\beta}_t e^{\beta_t}\nabla f(X_t). \tag{10b}$$

We begin by demonstrating how to derive the Lyapunov function (6) for the momentum dynamics (10); this derivation is similar in spirit to the Lyapunov analysis of mirror descent by Nemirovski and Yudin. We have:

$$\begin{aligned}
\frac{d}{dt}D_h(x, Z_t) &= \frac{d}{dt}\left(h(x) - h(Z_t) - \langle\nabla h(Z_t), x - Z_t\rangle\right) \\
&= -\langle\nabla h(Z_t), \dot{Z}_t\rangle - \left\langle\frac{d}{dt}\nabla h(Z_t), x - Z_t\right\rangle + \langle\nabla h(Z_t), \dot{Z}_t\rangle \\
&= -\left\langle\frac{d}{dt}\nabla h(Z_t), x - Z_t\right\rangle.
\end{aligned}$$

Using this identity, we obtain the following argument:

$$\frac{d}{dt} D_h(x, Z_t) = -\left\langle \frac{d}{dt} \nabla h(Z_t), x - Z_t \right\rangle$$

$$= \dot{\beta}_t e^{\beta_t} \left\langle \nabla f(X_t), x - X_t - \frac{1}{\dot{\beta}_t} \dot{X}_s \right\rangle \tag{12a}$$

$$= \dot{\beta}_t e^{\beta_t} \langle \nabla f(X_t), x - X_t \rangle - e^{\beta_t} \langle \nabla f(X_t), \dot{X}_t \rangle dt$$

$$= \dot{\beta}_t e^{\beta_t} \langle \nabla f(X_t), x - X_t \rangle - \frac{d}{dt} \left(e^{\beta_t} f(X_t) \right) + \dot{\beta}_t e^{\beta_t} f(X_t)$$

$$= \dot{\beta}_t e^{\beta_t} [f(X_t) + \langle \nabla f(X_t), x - X_t \rangle] - \frac{d}{dt} \left(e^{\beta_t} f(X_t) \right)$$

$$\leq \dot{\beta}_t e^{\beta_t} f(x) - \frac{d}{dt} \left(e^{\beta_t} f(X_t) \right) \tag{12b}$$

$$= -\frac{d}{dt} \left(e^{\beta_t} (f(X_t) - f(x)) \right). \tag{12c}$$

Here (12a) uses the momentum dynamics (10b) and (10a). The inequality (12b) follows from the convexity of f. Rearranging terms and taking $x = x^*$, we have shown that the function (6) has nonpositive derivative for all t and is hence a Lyapunov function for the family of momentum dynamics (10).

The structure of this derivation provides a tool for the analysis of several discretizations of momentum dynamics (4). In the following subsections, we demonstrate how to analyze the quasi-monotone subgradient method [21], the class of accelerated gradient methods [2, 28], and the general family of conditional gradient algorithms [20] using the Lyapunov argument above.

4.2 Implicit and explicit Euler methods

The methods that we study are based on compositions of Euler discretizations of the ODE $\dot{X}_t = v(X_t)$. Recall that the *explicit Euler method* evaluates the vector field at the current point to determine a discrete-time step:

$$\frac{x_{k+1} - x_k}{\delta} = \frac{X_{t+\delta} - X_t}{\delta} = v(X_t) = v(x_k),$$

whereas the *implicit Euler method* evaluates the vector field at the future point:

$$\frac{x_{k+1} - x_k}{\delta} = \frac{X_{t+\delta} - X_t}{\delta} = v(X_{t+\delta}) = v(x_{k+1}).$$

In our case, the momentum dynamics (10) are a system of two first-order equations, and one can compose the implicit and explicit Euler methods in four different ways, leading to four separate algorithms. To illustrate how this works, we write (10) as the following system of first-order ODEs,

$$Z_t = X_t + \frac{e^{\beta_t}}{\frac{d}{dt} e^{\beta_t}} \dot{X}_t, \tag{13a}$$

$$\frac{d}{dt} \nabla h(Z_t) = -\left(\frac{d}{dt} e^{\beta_t} \right) \nabla f(X_t). \tag{13b}$$

7

Taking $e^{\beta_t} = A_k$, $\dot{X}_t = \frac{x_{k+1} - x_k}{\delta}$, and $\frac{d}{dt} e^{\beta_t} = \frac{A_{k+1} - A_k}{\delta}$, the implicit Euler method applied to (13a) results in the following sequence:

$$z_{k+1} = x_{k+1} + \frac{A_k}{A_{k+1} - A_k}(x_{k+1} - x_k).$$

The explicit Euler method, on the other hand, results in the sequence

$$z_k = x_k + \frac{A_k}{A_{k+1} - A_k}(x_{k+1} - x_k).$$

For equation (13b), using the identification $\frac{d}{dt} \nabla h(Z_t) = \frac{1}{\delta}(\nabla h(z_{k+1}) - \nabla h(z_k))$, the implicit Euler method results in the following sequence

$$\nabla h(z_{k+1}) - \nabla h(z_k) = (A_{k+1} - A_k)\nabla f(x_{k+1}),$$

whereas the explicit Euler method results in the sequence

$$\nabla h(z_{k+1}) - \nabla h(z_k) = (A_{k+1} - A_k)\nabla f(x_k).$$

In the following, we analyze three combinations of the implicit and explicit Euler methods using our Lyapunov framework. Furthermore, we show there are at least two slightly different methods, which do not comport to a straightforward discretization technique, that can also be analyzed using our Lyapunov analysis.

4.3 Momentum-Based Algorithms

We give a short analysis of the result of implicit Euler discretization of both (13a) and (13b). We can write the algorithm as follows:

Algorithm 1 Implicit Euler Based Method

Assumptions: f, h are convex and differentiable.
Choose $A_0 = 1$, $x_{-1} = x_0 = z_0$ and $\tau_k = \frac{A_{k+1} - A_k}{A_k}$. Define recursively,

$$z_k = x_k + \frac{1}{\tau_{k-1}}(x_k - x_{k-1}), \tag{14a}$$

$$x_{k+1} = \underset{\substack{x \in \mathcal{X} \\ z = \frac{1+\tau_k}{\tau_k} x - \frac{1}{\tau_k} x_k}}{\arg\min} \left\{ f(x) + \frac{1}{A_{k+1} - A_k} D_h(z, z_k) \right\}, \tag{14b}$$

---

the optimality conditions for which constitute our discretization

$$z_{k+1} = x_{k+1} + \frac{A_k}{A_{k+1} - A_k}(x_{k+1} - x_k), \tag{15a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(A_{k+1} - A_k)\nabla f(x_{k+1}). \tag{15b}$$

To analyze this algorithm, we make the same continuous to discrete time identifications as in (6), which results in the following discrete-time energy function:

$$E_k = A_k(f(x_k) - f(x)) + D_h(x, z_k). \tag{16}$$

8

Theorem 1. *Consider the Euler discretization* (15). *Using the discrete Lyapunov function* (16), *the following bound holds:*

$$E_{k+1} - E_k \leq \varepsilon_k,$$

where the error is negative,

$$\varepsilon_k = -D_h(z_{k+1}, z_k).$$

Proof. In view of definition (16), the following holds,

$$
\begin{aligned}
E_{k+1} - E_k &= A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) - \langle \nabla h(z_{k+1}) - \nabla h(z_k), x - z_{k+1} \rangle - D_h(z_{k+1}, z_k) \\
&\overset{(15b)}{=} A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) + (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x - z_{k+1} \rangle - D_h(z_{k+1}, z_k) \\
&\overset{(15a)}{=} A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) + (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x - x_{k+1} \rangle \\
&\quad + A_k \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle - D_h(z_{k+1}, z_k) \\
&\leq A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) + (A_{k+1} - A_k)(f(x) - f(x_{k+1})) \\
&\quad + A_k(f(x_k) - f(x_{k+1})) - D_h(z_{k+1}, z_k) \\
&= -D_h(z_{k+1}, z_k),
\end{aligned}
$$

where the inequality follows from the convexity of f. $\qquad\square$

Taking $x = x^*$, and writing $E_k \leq E_0$ allows us to obtain a convergence rate guarantee analogous to (8):

$$f(x_k) - f(x^*) \leq \frac{D_h(x^*, x_0) + A_0(f(x_0) - f(x^*))}{A_k}. \tag{18}$$

We remark that the subequations corresponding to (14) are difficult to solve. For reference, the proximal minimization method [6],

$$x_{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f(x) + \frac{1}{A_{k+1} - A_k} D_h(x, x_k) \right\}, \tag{19}$$

involves applying the implicit Euler method to the mirror descent dynamics, and both yield a $O(1/A_k)$ convergence rate.

4.4 The Quasi-Monotone Subgradient Method

The quasi-monotone subgradient method was developed by Nesterov [21] as an alternative to the dual averaging method. It also achieves the $O(1/\sqrt{k})$ lower bound for the class of methods with bounded subgradients. The advantage of the quasi-monotone method however is that a convergence-rate guarantee can be shown for the entire sequence of iterates instead of the average (or minimum) iterate, as is the case with the dual averaging method in the non-Euclidean setting. *We begin by studying the quasi-monotone method in the situation where we have full gradients, which we assume to be bounded* and show later how the analysis extends to the setting where we only have access to bounded subgradients. We also set the dual averaging term $\gamma_k \equiv 1$ for convenience and give an analysis of the full algorithm in Appendix B.2. With these minor modifications, the algorithm can be understood as the explicit Euler method applied to (13a) and the implicit Euler method applied to (13b) (where now we take $e^{\beta_t} = A_{k+1}$, while leaving our identification for $\frac{d}{dt}e^{\beta_t}$ unchanged):

Algorithm 2 The Quasi-Monotone (Sub)gradient Method $\gamma_k \equiv 1$

Assumptions: f, h are convex and differentiable. h is strongly convex and f has bounded gradients, $\sup_{x \in \mathcal{X}} \|\nabla f(x)\| = G < \infty$.

Let $A_0 = 1$, $x_0 = z_0$, $\tau_k = \frac{A_{k+1} - A_k}{A_{k+1}}$, $\alpha_k = A_k - A_{k-1}$. Define recursively,

$$x_{k+1} = \tau_k z_k + (1 - \tau_k) x_k, \tag{20a}$$

$$z_k = \arg\min_{z \in \mathcal{X}} \left\{ \sum_{i=1}^{k} \alpha_i \langle \nabla f(x_i), z \rangle + D_h(z, x_0) \right\}. \tag{20b}$$

This modified quasi-monotone subgradient method has optimality conditions

$$z_k = x_k + \frac{A_{k+1}}{A_{k+1} - A_k}(x_{k+1} - x_k), \tag{21a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(A_{k+1} - A_k)\nabla f(x_{k+1}). \tag{21b}$$

The following result illustrates how the discrete Lyapunov function (16) can be used to analyze (21):

Theorem 2. *Using the Lyapunov function* (16)*, we can show the quasi-monotone method* (21) *satisfies*

$$E_{k+1} - E_k \leq \varepsilon_k,$$

where the error scales in the following way,

$$\varepsilon_k = \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2 \leq \frac{(A_{k+1} - A_k)^2}{2}G^2.$$

Taking $x = x^*$ and summing over k results in the following convergence rate:

$$f(x_k) - f(x^*) \leq \frac{D_h(x^*, x_0) + A_0(f(x_0) - f(x^*)) + \frac{1}{2}\sum_{i=0}^{k} \varepsilon_i}{A_k}. \tag{22}$$

If we optimize the bound (22) over A_k, we can obtain an $O(1/\sqrt{k})$ convergence rate guarantee by setting a time-horizon for the algorithm to be run, and choosing $A_{k+1} - A_k = \frac{D_h(x,x_0)}{G\sqrt{k+1}}$. Without this step, we suffer an additional $\log k$ factor in the the numerator.[2]

In the proof given in Appendix B.1, convexity is the only property of f that is necessary to show (22). Thus, the proof can be extended to include subgradient steps instead of full gradient steps, where now the condition on f is that its subgradients are bounded. This recovers the result of Nesterov [21] using the technique of estimate sequences.

4.5 Other Discretizations

We give examples of two other algorithms that are the result of discretizing the momentum dynamics (13), and analyze them using the discretized Lyapunov function (16). The proofs of these

---

[2]See [21, 2] for more details on this history on subgradient methods as well as step-size analysis.

results can be found in the Appendix B.3. The first method is the result of applying the explicit Euler method to (13b) and the implicit Euler method to (13a):

Algorithm 3 Method 1
Assumptions: f is continuously differentiable, $\|\nabla^2 f\| \leq L$, and $\mathcal{X}$ is convex and compact.
Let $A_0 = 1$, $x_0 = z_0$, $\alpha_k = A_{k+1} - A_k$ and $\tau_k = \frac{A_{k+1} - A_k}{A_k}$. Define recursively,

$$z_{k+1} = \arg\min_{z \in \mathcal{X}} \left\{ \langle \nabla f(x_k), z \rangle + \frac{1}{\alpha_k} D_h(z, z_k) \right\} \tag{23a}$$

$$x_{k+1} = \frac{\tau_k}{\tau_k - 1} z_{k+1} + \frac{1}{\tau_k - 1} x_k. \tag{23b}$$

This algorithm has optimality conditions,

$$z_{k+1} = x_{k+1} - \frac{A_k}{A_{k+1} - A_k}(x_{k+1} - x_k), \tag{24a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(A_{k+1} - A_k)\nabla f(x_k). \tag{24b}$$

Choosing $A_k = k(k+1)/2$ results in an $O(1/k)$ convergence rate. We remark that this algorithm is structurally very similar to the conditional gradient method (see B.3 for details). The second method does not correspond to a straightforward discretization technique and is given by the following sequences:

Algorithm 4 Method 2
Assumptions: f, h are convex and differentiable. h is strongly-convex and f has bounded gradients, $\sup_{x \in \mathcal{X}} \|\nabla f(x)\| = G < \infty$. Let $A_0 = 1$, $x_0 = z_0$ and $\tau_k = \frac{A_{k+1} - A_k}{A_k}$. Define recursively,

$$x_{k+1} = \frac{\tau_k}{\tau_k + 1} z_k + \frac{1}{\tau_k + 1} x_k, \tag{25a}$$

$$z_{k+1} = \arg\min_{z \in \mathcal{X}} \left\{ \langle \nabla f(x_{k+1}), z \rangle + \frac{1}{\alpha_k} D_h(z, z_k) \right\}. \tag{25b}$$

This algorithm has optimality conditions,

$$z_k = x_{k+1} + \frac{A_k}{A_{k+1} - A_k}(x_{k+1} - x_k), \tag{26a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(A_{k+1} - A_k)\nabla f(x_{k+1}). \tag{26b}$$

Notice that (26) is very similar to the quasi-monotone subgradient method (20). In particular, under the same assumptions, it achieves a $O(1/\sqrt{k})$ convergence rate. Furthermore, like the quasi-monotone subgradient method, it can be extended to functions which have bounded subgradients instead of full gradients and an additional weighting term can be added.

4.6 Accelerated Mirror Descent

We present two variants of the accelerated mirror descent algorithm (which is the accelerated gradient descent algorithm in the setting where $h = \frac{1}{2}\|\cdot\|^2$). Neither of these correspond to a straightforward discretization of (10)—they both entail introducing an additional sequence $\{y_k\}$ to obtain a better bound on the error. The first is the version introduced by Baes [2]:

Algorithm 5 Accelerated Mirror Descent (Weakly Convex Setting)

Assumptions: f, h are convex and differentiable. h is 1-strongly convex and f has smooth gradients $\|\nabla^2 f\| \leq L$

Choose $A_0 = 1$, $M > 0$, $\tilde{A}_{k+1} = L^{-1}A_{k+1}$, $\tau_k = \frac{\tilde{A}_{k+1} - \tilde{A}_k}{\tilde{A}_{k+1}} := \frac{\alpha_k}{\tilde{A}_{k+1}}$ and $x_0 = z_0 = y_0$. Define recursively,

$$z_k = \arg\min_{z \in \mathcal{X}} \left\{ \sum_{i=1}^{k} \alpha_i \langle \nabla f(y_i), z \rangle + D_h(z, z_0) \right\} \tag{27a}$$

$$y_k = \arg\min_{x \in \mathcal{X}} \left\{ \langle \nabla f(x_k), x \rangle + \frac{L}{4M}\|y - x_k\|^2 \right\} \tag{27b}$$

$$x_{k+1} = \tau_k z_k + (1 - \tau_k)y_k \tag{27c}$$

This algorithm satisfies the following optimality conditions:

$$z_k = y_k + \frac{\tilde{A}_{k+1}}{\tilde{A}_{k+1} - \tilde{A}_k}(x_{k+1} - y_k), \tag{28a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(\tilde{A}_{k+1} - \tilde{A}_k)\nabla f(y_{k+1}) \tag{28b}$$

$$M\|L^{-1}\nabla f(y_k)\|_*^2 \leq L^{-1}\langle \nabla f(y_k), x_k - y_k \rangle. \tag{28c}$$

The advantage of this version of the accelerated mirror descent algorithm is that it can be extended to accelerate higher-order gradient methods (the details of which we give in Appendix B.5). The second version, originally introduced by Nesterov [17], entails setting $M = \frac{1}{2}$ and computing the mirror descent update (27a) from the gradient at x:

$$z_k = \arg\min_{z \in \mathcal{X}} \left\{ \sum_{i=1}^{k} \alpha_i \langle \nabla f(x_i), z \rangle + D_h(z, z_0) \right\}.$$

The resulting algorithm has the following optimality conditions:

$$z_k = y_k + \frac{\tilde{A}_{k+1}}{\tilde{A}_{k+1} - \tilde{A}_k}(x_{k+1} - y_k), \tag{29a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(\tilde{A}_{k+1} - \tilde{A}_k)\nabla f(x_{k+1}) \tag{29b}$$

$$f(y_k) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|_*^2. \tag{29c}$$

As mentioned above, to obtain a convergence guarantee for both algorithms we have replaced the current state x_k by another sequence y_k which needs to satisfy (28c). Furthermore, we have

replaced our sequence A_{k+1} by $\tilde{A}_{k+1}$ which depends on the Lipschitz parameter. We make the same adjustments to the Lyapunov function (16) which we summarize in the following theorem:

Theorem 3. *Using the following Lyapunov function*

$$E_k = \tilde{A}_k(f(y_k) - f(x)) + D_h(x, z_k), \tag{30}$$

we can show

$$E_{k+1} - E_k \le \varepsilon_k,$$

where the error scales in the following way

$$\varepsilon_k = \left(\frac{1}{2}(\tilde{A}_{k+1} - \tilde{A}_k)^2 - \tilde{A}_{k+1}M \right) \|\nabla f(y_{k+1})\|^2$$

for the accelerated gradient method (27).

We choose $x = x^*$ and A_k such that $\tilde{A}_{k+1}^{-1}(\tilde{A}_{k+1} - \tilde{A}_k)^2 \le 2M$ to ensure the error is nonpositive. With these choices, we obtain the following convergence rate,

$$f(y_k) - f(x^*) \le \frac{\tilde{A}_0(f(x_0) - f(x^*)) + D_h(x^*, x_0)}{\tilde{A}_k}. \tag{31}$$

The condition to ensure the error is nonpositive restricts the choice of sequence $\{A_k\}_{k=1}^{\infty}$ so that it can grow at most quadratically. This results in a $O(1/k^2)$ convergence rate. We give a proof of both of these methods in Appendix B.5.

4.7 The Conditional Gradient Algorithm

The conditional gradient method, also called the *Frank-Wolfe* algorithm [9], has garnered renewed interest over the last several years. This interest has inspired several different analyses [10, 1] of the algorithm. Of particular note is the analysis provided by Nesterov [20], who used to technique of estimate sequences to evaluate the algorithm under varying smoothness assumptions and step-size conditions. Nesterov also extended the algorithm to provide an analysis of a trust-region Newton-type method. Here, we explore the conditional gradient method from a dynamical view, and give an alternate analysis of both methods. To do so, we first revisit the derivation of the Lyapunov function. Notice, that if $\dot{\beta}_t > 0$ and we can ensure the following inequality for our dynamical system

$$0 \le \langle \nabla f(X_t), x - Z_t \rangle, \tag{32}$$

$\forall x \in \mathcal{X}$, where

$$Z_t = X_t + \frac{1}{\dot{\beta}_t} \dot{X}_t, \tag{33}$$

then the Lyapunov analysis (12) follows without the use of the Bregman divergence:

$$0 \le \dot{\beta}_t e^{\beta_t} \langle \nabla f(X_t), x - Z_t \rangle$$

$$\overset{(12)}{\le} \dot{\beta}_t e^{\beta_t} f(x) - \frac{d}{dt} \left(e^{\beta_t} f(X_t) \right)$$

$$= -\frac{d}{dt} \left(e^{\beta_t} \left(f(X_t) - f(x) \right) \right).$$

13

Integrating shows that the following function

$$\mathcal{E}_t = e^{\beta_t}(f(X_t) - f(x)). \tag{35}$$

is a Lyapunov function for any family of dynamics satisfying (32) and (33). We now turn our attention to the conditional gradient method, which maintains the following iterates:

Algorithm 6 Conditional Gradient Method

Assumptions: f is continuously differentiable, $\|\nabla^2 f\| \leq L$, and $\mathcal{X}$ is convex and compact. Choose $A_0 = 1$, $x_0 = z_0$ $\tau_k = \frac{A_{k+1} - A_k}{A_{k+1}} := \frac{\alpha_{k+1}}{A_{k+1}}$.

$$z_k = \arg\min_{z \in \mathcal{X}} \langle \nabla f(x_k), z \rangle, \tag{36a}$$

$$x_{k+1} = \tau_k z_k + (1 - \tau_k) x_k. \tag{36b}$$

In particular, the conditional gradient method satisfies the following variational inequalities:

$$0 \leq \langle \nabla f(x_k), x - z_k \rangle \quad \forall x \in \mathcal{X}, \tag{37a}$$

$$z_k = x_k + \frac{A_{k+1}}{A_{k+1} - A_k}(x_{k+1} - x_k). \tag{37b}$$

This suggests the following Lyapunov analysis of the conditional gradient method:

Theorem 4. *Using the Lyapunov function*

$$E_k = A_k(f(x_k) - f(x^*)), \tag{38}$$

we can show that the conditional gradient method (37) *satisfies*

$$E_{k+1} - E_k \leq \varepsilon_k, \tag{39}$$

where the error scales in the following way,

$$\varepsilon_k = \frac{L}{2}\frac{(A_{k+1} - A_k)^2}{A_{k+1}}diam(\mathcal{X})^2.$$

This matches the analysis of the conditional gradient method by Nesterov using the technique of estimate sequences as well as the analysis provided by others [10, 1]. In particular, choosing $A_k = \frac{k(k+1)}{2}$ gives the standard $O(1/k)$ convergence rate (see Nesterov [20, Eq (2.16)] for details).

The proof of this theorem can be found in Appendix B.4.

4.7.1 Extensions

We briefly point out that the algorithm resulting from applying an implicit Euler method to both (32) and (33),

$$z_{k+1} = \arg\min_{z \in \mathcal{X}} \langle \nabla f(x_{k+1}), z \rangle$$

$$z_{k+1} = x_{k+1} + \frac{A_k}{A_{k+1} - A_k}(x_{k+1} - x_k),$$

results in an algorithm for which we can show $E_{k+1} - E_k \leq 0$ using (38). However, like most implicit discretization techniques applied to the momentum dynamics (4), this does not lead to an algorithm with tractable subproblems. We now turn our attention to analyzing the trust region Newton-like method introduced by Nesterov. This method is given by the following update,

---

Algorithm 7 Conditional Gradient Method [20, (5.1)]

Assumptions: f is twice continuously differentiable, $\|\nabla^3 f\| \leq L$, and $\mathcal{X}$ is convex and compact. Choose $A_0 = 1$, $x_0 = z_0$ $\tau_k = \frac{A_{k+1} - A_k}{A_{k+1}}$.

$$x_{k+1} \in \arg\min_{y=(1-\tau_k)x_k+\tau_k x} \left\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2} \langle \nabla f(x_k)(y - x_k), y - x_k \rangle : x \in dom(\mathcal{X}) \right\} \quad (41a)$$

This algorithm has the following optimality conditions:

$$z_k = x_k + \frac{A_{k+1}}{A_{k+1} - A_k}(x_{k+1} - x_k), \quad (42a)$$

$$\langle \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k), y - x_{k+1} \rangle \geq 0, \quad (42b)$$

$$\forall y = (1 - \tau_k)x_k + \tau_k x, \quad x \in dom(\mathcal{X}). \quad (42c)$$

In Appendix B.4 we show how to recover the bound shown by Nesterov using the Lyapunov function (38).

5 Strong Convexity

In this section, we introduce a new family of momentum dynamics which can be discretized to obtain the accelerated gradient descent algorithm in the setting where the objective function f is strongly convex. In Appendix C, we show how to derive a Lyapunov function using the momentum dynamics. Furthermore, we show how to discretize the dynamics to obtain a discrete-time algorithm which can be analyzed by performing an analogous discretization of the Lyapunov function.

5.1 Lyapunov Analysis

In this section, we study the Lyapunov function (7) in the Euclidean setting,

$$\mathcal{E}_t = e^{\beta_t} \left(f(X_t) - f(x^*) + \frac{\mu}{2} \|x^* - Z_t\|^2 \right), \quad (43)$$

and use it to analyze the following dynamics:

$$Z_t = X_t + \frac{1}{\beta_t} \dot{X}_t \quad (44a)$$

$$\mu \dot{Z}_t = -\mu \dot{X}_t - \dot{\beta}_t \nabla f(X_t). \quad (44b)$$

Note that (44) is the Euler-Lagrange equation for the following Lagrangian:

$$\mathcal{L}(x, \dot{x}, t) = e^{\gamma_t + \beta_t + \alpha_t} \left(\frac{\mu}{2} e^{-2\alpha_t} \|\dot{x}\|^2 - f(x) \right), \quad (45)$$

15

under the same ideal scaling conditions (5), where we take $e^{\alpha_t} = \dot{\beta}_t$. Furthermore, this Lagrangian can be generalized to include non-Euclidean geometries; however, further exploration of this Lagrangian is outside the scope of this work. Here, we demonstrate that (43) can be used to establish an $O(e^{-\beta_t})$ rate of convergence for (44) with a strong convexity assumption. To do so, note that if we can ensure $\dot{\mathcal{E}}_t = e^{\beta_t} \dot{\beta}_t \tilde{\mathcal{E}}_t + e^{\beta_t} \dot{\tilde{\mathcal{E}}}_t \leq 0$, which amounts to ensuring $\dot{\tilde{\mathcal{E}}}_t \leq -\dot{\beta}_t \tilde{\mathcal{E}}_t$ for

$$\tilde{\mathcal{E}}_t = f(X_t) - f(x^*) + \frac{\mu}{2}\|x^* - Z_t\|^2, \tag{46}$$

then (43) is a Lyapunov function. To that end, we have the following argument:

$$
\begin{aligned}
\dot{\tilde{\mathcal{E}}}_t &= \langle \nabla f(X_t), \dot{X}_t \rangle - \mu \left\langle \dot{Z}_t, x^* - X_t - \frac{1}{\dot{\beta}_t} \dot{X}_t \right\rangle \\
&= \langle \nabla f(X_t), \dot{X}_t \rangle + \left\langle \mu \dot{X}_t + \dot{\beta}_t \nabla f(X_t), x^* - X_t - \frac{1}{\dot{\beta}_t} \dot{X}_t \right\rangle \\
&= \dot{\beta}_t \langle \nabla f(X_t), x^* - X_t \rangle + \mu \left\langle \dot{X}_t, x^* - X_t - \frac{1}{\dot{\beta}_t} \dot{X}_t \right\rangle \\
&\leq -\dot{\beta}_t \left(f(X_t) - f(x^*) + \frac{\mu}{2}\|x^* - X_t\|^2 \right) + \mu \left\langle \dot{X}_t, x^* - X_t - \frac{1}{\dot{\beta}_t} \dot{X}_t \right\rangle \qquad\qquad (47a) \\
&= -\dot{\beta}_t \left(f(X_t) - f(x^*) + \frac{\mu}{2}\left\| x^* - X_t - \frac{1}{\dot{\beta}_t} \dot{X}_t \right\|^2 \right) - \mu \left\langle \dot{X}_t, x^* - X_t - \frac{1}{\dot{\beta}_t} \dot{X}_t \right\rangle - \frac{\mu}{2\dot{\beta}_t}\|\dot{X}_t\|^2 \\
&\quad + \mu \left\langle \dot{X}_t, x^* - X_t - \frac{1}{\dot{\beta}_t} \dot{X}_t \right\rangle \\
&\leq -\dot{\beta}_t \left(f(X_t) - f(x^*) + \frac{\mu}{2}\|x^* - Z_t\|^2 \right),
\end{aligned}
$$

where (47a) uses our strong convexity assumption. In the following subsections, we demonstrate how to discretize the dynamic (44), and how this discretization can be analyzed using a Lyapunov argument analogous to (47).

5.2 Discretizing the Dynamics

As a proof of concept, we begin by analyzing an implicit discretization of the dynamics (44). Denoting $\tau_k = \frac{A_{k+1} - A_k}{A_k}$, $\dot{Z}_t = \frac{z_{k+1} - z_k}{\delta}$, $\dot{X}_t = \frac{x_{k+1} - x_k}{\delta}$, $\frac{1}{\dot{\beta}_t} = \frac{e^{\beta_t}}{\frac{d}{dt} e^{\beta_t}} = \frac{A_k}{\frac{A_{k+1} - A_k}{\delta}}$, we obtain the following algorithm:

Algorithm 8 Implicit Euler Based Method (Strong Convexity)

Assumptions: f, h are convex and differentiable.

Choose $A_0 = 1$, $x_{-1} = x_0 = z_0$ and $\tau_k = \frac{A_{k+1} - A_k}{A_{k+1}}$. Define recursively,

$$z_k = x_k + \frac{1}{\tau_{k-1}}(x_k - x_{k-1}) \tag{48a}$$

$$x_{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f(x) - \frac{\mu}{2\tau_k}\|\tau_k(x_k - z_k) - (x - x_k)\|^2 \right\}, \tag{48b}$$

16

The optimality conditions of this algorithm corresponds to our discretization:

$$z_{k+1} - z_k = \tau_k \left(x_{k+1} - z_{k+1} - \frac{1}{\mu} \nabla f(x_{k+1}) \right) \tag{49a}$$

$$z_{k+1} = x_{k+1} + \frac{1}{\tau_k}(x_{k+1} - x_k). \tag{49b}$$

Theorem 5. *The following bound*

$$\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k \tilde{E}_{k+1} + \varepsilon_k,$$

can be shown for (49b) *using the discrete-time Lyapunov function*

$$\tilde{E}_k = f(x_k) - f^* + \frac{\mu}{2} \|x^* - z_k\|^2, \tag{50}$$

where the error is negative,

$$\varepsilon_k = -\frac{\mu}{2} \|z_{k+1} - z_k\|^2.$$

Proof. Notice that a similar argument to (47) holds:

$$\tilde{E}_{k+1} - \tilde{E}_k = f(x_{k+1}) - f(x_k) - \mu\langle z_{k+1} - z_k, x^* - z_{k+1}\rangle - \frac{\mu}{2}\|z_{k+1} - z_k\|^2$$

$$\overset{(49a)}{=} f(x_{k+1}) - f(x_k) + \tau_k\langle \nabla f(x_{k+1}) - \mu(x_{k+1} - z_{k+1}), x^* - z_{k+1}\rangle - \frac{\mu}{2}\|z_{k+1} - z_k\|^2$$

$$= \tau_k\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + f(x_{k+1}) - f(x_k) + \tau_k\langle \nabla f(x_{k+1}), x_{k+1} - z_{k+1}\rangle - \frac{\mu}{2}\|z_{k+1} - z_k\|^2$$

$$+ \mu\tau_k\langle z_{k+1} - x_{k+1}, x^* - z_{k+1}\rangle$$

$$\leq -\tau_k\left(f(x_k) - f(x^*) + \frac{\mu}{2}\|x^* - x_{k+1}\|^2\right) + \tau_k\langle \nabla f(x_{k+1}), x_{k+1} - z_{k+1}\rangle$$

$$- \frac{\mu}{2}\|z_{k+1} - z_k\|^2 + \mu\tau_k\langle z_{k+1} - x_{k+1}, x^* - z_{k+1}\rangle$$

$$\leq -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_{k+1}\|^2\right) + f(x_{k+1}) - f(x_k) + \tau_k\langle \nabla f(x_{k+1}), x_{k+1} - z_{k+1}\rangle$$

$$- \frac{\mu}{2}\|z_{k+1} - z_k\|^2.$$

$$\overset{(49b)}{=} -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_{k+1}\|^2\right) + f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle$$

$$- \frac{\mu}{2}\|z_{k+1} - z_k\|^2.$$

$$\leq -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_{k+1}\|^2\right) - \frac{\mu}{2}\|z_{k+1} - z_k\|^2.$$

Therefore,

$$\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_{k+1}\|^2\right) = -\tau_k\tilde{E}_{k+1}. \tag{52}$$

$\square$

Choosing $\tau_k = \sqrt{\kappa}$ gives the $O(e^{-\sqrt{\kappa}k})$ convergence rate. However, there is no restriction on τ_k; we are free to scale it arbitrarily. Given the subproblems for this algorithm are difficult to solve, we consider other discretizations based on the explicit Euler method. In particular, we utilize the

same trick of introducing a new sequence $\{y_k\}_{k=1}^{\infty}$, as in accelerated gradient descent (28), in order to bound the error. We find that we can use a Lyapunov argument to analyze the following two sequences:

$$z_{k+1} - z_k = \tau_k \left(x_{k+1} - z_k - \frac{1}{\mu} \nabla f(x_{k+1}) \right), \tag{53a}$$

$$\tau_k (x_{k+1} - z_k) = y_k - x_{k+1} \tag{53b}$$

$$y_{k+1} = x_{k+1} - \frac{1}{L} \nabla f(x_{k+1}), \tag{53c}$$

and

$$z_{k+1} - z_k = \tau_k \left(x_k - z_k - \frac{1}{\mu} \nabla f(x_k) \right), \tag{54a}$$

$$\tau_k (x_k - z_k) = y_k - x_k \tag{54b}$$

$$y_{k+1} = x_k - \frac{1}{L} \nabla f(x_k). \tag{54c}$$

The second sequence (54) is the accelerated gradient scheme introduced by Nesterov [16, (2.2.8)], which can be further simplified into two sequences. We summarize our results in the following theorem:

Theorem 6. *Using the following Lyapunov function,*

$$\tilde{E}_k = f(y_k) - f(x^*) + \frac{\mu}{2} \|z_k - x^*\|^2, \tag{55}$$

we can show

$$\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k \tilde{E}_k + \varepsilon_k \tag{56}$$

for both methods (53) *and* (54)*, where*

$$\varepsilon_k = \left(\frac{\tau_k^2}{2\mu} - \frac{1}{2L} \right) \|\nabla f(x_{k+1})\|^2 + \left(\frac{\tau_k L}{2} - \frac{\mu}{2\tau_k} \right) \|x_{k+1} - y_k\|^2 \tag{57}$$

and

$$\varepsilon_k = \left(\frac{\tau_k^2}{2\mu} - \frac{1}{2L} \right) \|\nabla f(x_k)\|^2 + \left(\frac{\tau_k L}{2} - \frac{\mu}{2\tau_k} \right) \|x_k - y_k\|^2, \tag{58}$$

for (53) *and* (54) *respectively.*

The proof of this result is in Appendix C. In both cases, to ensure the error is nonpositive we must set $\tau_k \leq 1/\sqrt{\kappa}$, where $\kappa = L/\mu$ is the condition number. Indeed, setting $\tau_k = 1/\sqrt{\kappa}$ results in the recurrence

$$\tilde{E}_{k+1} \leq \left(1 - \tfrac{1}{\sqrt{\kappa}} \right) \tilde{E}_k,$$

which ensures the standard convergence rate for Nesterov's method applied to strongly convex functions [16].

6 Estimate Sequences

In this section, we connect our Lyapunov framework directly to the technique of estimate sequences. We derive continuous-time estimate sequences directly from our Lyapunov function and demonstrate how these two techniques are equivalent in continuous and discrete time.

We begin with a brief review of the technique of estimate sequences. In [16], Nesterov introduced estimate sequences by giving the following definition:

Definition 1. *[16, 2.2.1] A pair of sequences $\{\phi_k(x)\}_{k=1}^\infty$ and $\{A_k\}_{k=0}^\infty$ $A_k \geq 1$ is called an* estimate sequence *of function $f(x)$ if*

$$\frac{1}{A_k} \to 0$$

and for any $x \in \mathbb{R}^n$ and all $k \geq 0$, we have

$$\phi_k(x) \leq \left(1 - \frac{1}{A_k}\right)f(x) + \frac{1}{A_k}\phi_0(x). \tag{59}$$

The following lemma, given by Nesterov, explains why estimate sequences are useful.

Lemma 7. *[16, 2.2.1] If for some sequence $\{x_k\}_{k\geq 0}$ we have*

$$f(x_k) \leq \phi_k^* \equiv \min_{x \in \mathcal{X}} \phi_k(x), \tag{60}$$

then $f(x_k) - f(x^*) \leq \frac{1}{A_k}[\phi_0(x^*) - f(x^*)]$.

Proof. The proof is straightforward:

$$f(x_k) \overset{(60)}{\leq} \phi_k^* \equiv \min_{x \in \mathcal{X}} \phi_k(x) \overset{(59)}{\leq} \min_{x \in \mathcal{X}} \left[\left(1 - \frac{1}{A_k}\right)f(x) + \frac{1}{A_k}\phi_0(x)\right]$$
$$\leq \left(1 - \frac{1}{A_k}\right)f(x^*) + \frac{1}{A_k}\phi_0(x^*).$$

Rearranging gives the desired inequality. $\square$

Notice that this definition is not constructive and the definition of an estimation sequence is rather general. Finding sequences which satisfy these conditions is a non-trivial task. The next proposition, formalized by Baes in [2] as an extension of Nesterov's Lemma 2.2.2 [16], provides guidance for constructing estimate sequences. This construction is used in [16, 17, 18, 2, 21, 20], and is the only known way so far to construct an estimate sequence. We will see below that this particular class of estimate sequences can be turned into our Lyapunov functions with a few algebraic manipulations (and vice-versa).

Proposition 8. *[2, 2.2] Let $\phi_0 : \mathcal{X} \to \mathbb{R}$ be a convex function such that $\min_{x \in \mathcal{X}} \phi_0(x) \geq f^*$. Suppose also that we have a sequence $\{f_k\}_{k\geq 0}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ that underestimates f:*

$$f_k(x) \leq f(x) \quad \text{for all } x \in \mathcal{X} \text{ and all } k \geq 0. \tag{61}$$

Define recursively $A_0 = 1$

$$\alpha_k := A_{k+1} - A_k \tag{62}$$

$$\tau_k := \frac{a_k}{A_{k+1}}, \tag{63}$$

and

$$\phi_{k+1}(x) := (1 - \tau_k)\phi_k(x) + \tau_k f_k(x) = \frac{1}{A_{k+1}}\left(A_0\phi_0(x) + \sum_{i=0}^{k} a_i f_i(x)\right), \tag{64}$$

for all $k \geq 0$. Then $(\{\phi_k\}_{k\geq 0}, \{A_k\}_{k\geq 0})$ is an estimate sequence.

From (60) and (64), we observe that the following invariant:

$$A_{k+1}f(x_{k+1}) \leq \min_x A_{k+1}\phi_{k+1}(x) = \min_x \sum_{i=0}^{k} \alpha_i f_i(x) + A_0\phi_0(x), \tag{65}$$

is maintained. In [21, 20], this technique was extended to incorporate a error term $\{\tilde{\varepsilon}_k\}_{k=1}^{\infty}$,

$$\phi_{k+1}(x) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} := (1 - \tau_k)\left(\phi_k(x) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \tau_k f_k(x) = \frac{1}{A_{k+1}}\left(A_0(\phi_0(x) - \tilde{\varepsilon}_0) + \sum_{i=0}^{k} a_i f_i(x)\right) \tag{66}$$

Rearranging, we have

$$A_{k+1}f(x_{k+1}) \leq \min_x A_{k+1}\phi_{k+1}(x) = \min_x \sum_{i=0}^{k} \alpha_i f_i(x) + A_0\left(\phi_0(x) - \frac{\tilde{\varepsilon}_0}{A_0}\right) + \tilde{\varepsilon}_{k+1}. \tag{67}$$

Notice that the analogous argument to that of Lemma 7 holds:

$$\begin{aligned}
A_{k+1}f(x_{k+1}) &\leq \sum_{i=0}^{k} \alpha_i f_i(x^*) + A_0(\phi_0(x^*) - \tilde{\varepsilon}_0) + \tilde{\varepsilon}_{k+1} \\
&\overset{(61)}{\leq} \sum_{i=0}^{k} \alpha_i f(x^*) + A_0\phi_0(x^*) + \tilde{\varepsilon}_{k+1} \\
&\overset{(62)}{=} A_{k+1}f(x^*) + A_0\phi_0(x^*) + \tilde{\varepsilon}_{k+1},
\end{aligned} \tag{68a}$$

where in the second inequality we use the fact that $\varepsilon_k \geq 0$, $\forall k$. Rearranging,

$$f(x_{k+1}) - f(x^*) \leq \frac{1}{A_{k+1}}\left(A_0\phi_0(x^*) + \tilde{\varepsilon}_{k+1}\right).$$

Thus, we simply need to choose our sequence $\{A_k\}_{k=1}^{\infty}$ to ensure $\tilde{\varepsilon}_{k+1}/A_{k+1} \to 0$. The following table illustrates the choices of $\phi_k(x)$ and $\tilde{\varepsilon}_k$ for the four methods discussed earlier:

Here "linear" is defined as

$$f_i(x) = f(x_i) + \langle \nabla f(x_i), x - x_i \rangle \tag{69}$$

and "quadratic" is defined as

$$f_i(x) = f(x_i) + \langle \nabla f(x_i), x - x_i \rangle + \frac{\mu}{2}\|x - x_i\|^2. \tag{70}$$

Algorithm	$f_i(x)$	$\phi_k(x)$	$\tilde{\varepsilon}_k$
Quasi-Monotone Subgradient Method	linear	$\frac{1}{A_k}D_h(x, z_k) + f(x_k)$	$\frac{1}{2}\sum_{i=0}^{k}\frac{(A_i - A_{i-1})^2}{2}\|\nabla f(x_i)\|^2$
Accelerated Gradient Method (Weakly Convex)	linear	$\frac{1}{A_k}D_h(x, z_k) + f(x_k)$	0
Accelerated Gradient Method (Strongly Convex)	quadratic	$f(x_k) + \frac{\mu}{2}\|x - z_k\|^2$	0
Conditional Gradient Method	linear	$f(x_k)$	$\frac{1}{2}\sum_{i=0}^{k}\frac{(A_{i+1} - A_i)^2}{A_{i+1}}diam(\mathcal{X})^2$

Table 1: Choices of estimate sequences for various algorithms

6.1 Equivalence between Estimate Sequences and Lyapunov Functions

We now demonstrate an equivalence between these two frameworks. The continuous-time view shows that the errors in both the Lyapunov function and estimate sequences are due to discretization errors. We provide a sketch for how this works for most of these methods, leaving some of the details for Appendix D.

6.1.1 Accelerated Gradient Descent

Equivalence in discrete time The discrete-time estimate sequence (64) for accelerated gradient descent can be written:

$$\phi_{k+1}(x) := f(x_{k+1}) + \frac{1}{A_{k+1}}D_h(x, z_{k+1})$$

$$\stackrel{(64)}{=} (1 - \tau_k)\phi_k(x) + \tau_k f_k(x)$$

$$\stackrel{\text{Table } 1}{=} \left(1 - \frac{\alpha_k}{A_{k+1}}\right)\left(f(x_k) + \frac{1}{A_k}D_h(x, z_k)\right) + \frac{\alpha_k}{A_{k+1}}f_k(x).$$

Multiplying through by A_{k+1}, we have

$$A_{k+1}\left(f(x_{k+1}) + \frac{1}{A_{k+1}}D_h(x, z_{k+1})\right) = (A_{k+1} - \alpha_k)\left(f(x_k) + \frac{1}{A_k}D_h(x, z_k)\right) + \alpha_k f_k(x)$$

$$= (A_{k+1} - (A_{k+1} - A_k))\left(f(x_k) + \frac{1}{A_k}D_h(x, z_k)\right) + (A_{k+1} - A_k)f_k(x)$$

$$= A_k\left(f(x_k) + \frac{1}{A_k}D_h(x, z_k)\right) + (A_{k+1} - A_k)f_k(x)$$

$$\stackrel{(61)}{\leq} A_k f(x_k) + D_h(x, z_k) + (A_{k+1} - A_k)f(x).$$

Rearranging, we obtain the inequality $E_{k+1} \leq E_k$ for our Lyapunov function (16). Going the other direction, from our Lyapunov analysis we can derive the following bound:

$$E_k \leq E_0 \tag{71a}$$

$$A_k(f(x_k) - f(x)) + D_h(x, z_k) \leq A_0(f(x_0) - f(x)) + D_h(x, z_0)$$

$$A_k\left(f(x_k) - \frac{1}{A_k}D_h(x, z_k)\right) \leq (A_k - A_0)f(x) + A_0\left(f(x_0) + \frac{1}{A_0}D_h(x^*, z_0)\right)$$

$$A_k\phi_k(x) \leq (A_k - A_0)f(x) + A_0\phi_0(x). \tag{71b}$$

Rearranging, we obtain the estimate sequence (59), with $A_0 = 1$:

$$\phi_k(x) \leq \left(1 - \frac{A_0}{A_k}\right) f(x) + \frac{A_0}{A_k} \phi_0(x) \tag{72a}$$

$$\leq \left(1 - \frac{1}{A_k}\right) f(x) + \frac{1}{A_k} \phi_0(x). \tag{72b}$$

Equivalence in continuous time From the derivation of the Lyapunov function (12), we have the following equality:

$$\frac{d}{ds} D_h(x, Z_s) = \dot{\beta}_t e^{\beta_t} [f(X_s) + \langle \nabla f(X_s), x - X_s \rangle] - \frac{d}{ds} \left(e^{\beta_s} f(X_s)\right).$$

If we integrate and assume we start from rest $\dot{X}_0 = 0$, we can write this as,

$$D_h(x, Z_t) \leq \int_0^t \dot{\beta}_t e^{\beta_t} [f(X_s) + \langle \nabla f(X_s), x - X_s \rangle] ds - e^{\beta_t} f(X_t) + e^{\beta_0} f(X_0) + D_h(x, X_0). \tag{73}$$

Now, by pattern matching, we can use this inequality to extract a continuous-time estimate sequence. From (73) we have the inequality,

$$e^{\beta_t} f(X_t) + D_h(x, Z_t) \leq \int_0^t \dot{\beta}_t e^{\beta_t} [f(X_s) + \langle \nabla f(X_s), x - X_s \rangle] ds + e^{\beta_0} f(X_0) + D_h(x, Z_0)$$

$$\leq \int_0^t \dot{\beta}_t e^{\beta_t} f(x) + e^{\beta_0} f(X_0) + D_h(x, Z_0)$$

$$= (e^{\beta_t} - e^{\beta_0}) f(x) + e^{\beta_0} f(X_0) + D_h(x, Z_0).$$

Comparing this to (110a), if we define

$$\phi_t(x) = f(X_t) + e^{-\beta_t} D_h(x, Z_t)$$

then the above discussion shows that $\{\phi_t(x), e^{\beta_t}\}$ is a continuous-time estimate sequence. In Appendix D we extend this argument to the quasi-monotone subgradient method by adding an error term ε_k.

6.1.2 Conditional Gradient Method

Continuous-time estimate sequence For the conditional gradient method, the following condition needs to hold:

$$0 \leq \int_0^t \dot{\beta}_t e^{\beta_t} [f(X_s) + \langle \nabla f(X_s), x - X_s \rangle] - \frac{d}{ds} \left(e^{\beta_s} f(X_s)\right) ds. \tag{74}$$

We can write this as

$$e^{\beta_t} f(X_t) \leq \int_0^t \dot{\beta}_t e^{\beta_t} [f(X_s) + \langle \nabla f(X_s), x - X_s \rangle] ds + e^{\beta_0} f(X_0),$$

and, from the derivation of the Lypaynov function, we conclude

$$e^{\beta_t} f(X_t) \leq (e^{\beta_t} - e^{\beta_0}) f(x) + e^{\beta_0} f(X_0).$$

Thus, $\{f(X_t), e^{\beta_t}\}$ is a continuous-time estimate sequence for this method. In Appendix D, we show the corresponding equivalence in discrete time.

6.2 Accelerated Gradient Strong Convexity

Notice that for the dynamics in (44), the following holds:

$$
\begin{aligned}
\frac{d}{dt}\left(e^{\beta_t}\frac{\mu}{2}\|x - Z_t\|^2\right) &= \dot{\beta}_t e^{\beta_t}\frac{\mu}{2}\|x - Z_t\|^2 - e^{\beta_t}\mu\langle\dot{Z}_t, x - Z_t\rangle \\
&= \dot{\beta}_t e^{\beta_t}\frac{\mu}{2}\|x - Z_t\|^2 + e^{\beta_t}\mu\langle\dot{X}_t, x - Z_t\rangle + \dot{\beta}_t e^{\beta_t}\left\langle\nabla f(X_t), x - X_t - \frac{1}{\dot{\beta}_t}\dot{X}_t\right\rangle \\
&= \dot{\beta}_t e^{\beta_t}\frac{\mu}{2}\|x - Z_t\|^2 + \dot{\beta}_t e^{\beta_t}\mu\langle Z_t - X_t, x - X_t\rangle - \dot{\beta}_t e^{\beta_t}\frac{\mu}{2}\|Z_t - X_t\|^2 \\
&\quad + \dot{\beta}_t e^{\beta_t}\langle\nabla f(X_t), x - X_t\rangle - e^{\beta_t}\langle\nabla f(X_t), \dot{X}_t\rangle \\
&= \dot{\beta}_t e^{\beta_t}\frac{\mu}{2}\|x - X_t\|^2 + \dot{\beta}_t e^{\beta_t}\langle\nabla f(X_t), x - X_t\rangle - e^{\beta_t}\langle\nabla f(X_t), \dot{X}_t\rangle \\
&= \dot{\beta}_t e^{\beta_t}\left(f(X_t) + \langle\nabla f(X_t), x - X_t\rangle + \frac{\mu}{2}\|x - X_t\|^2\right) - \frac{d}{dt}\left(e^{\beta_t}f(X_t)\right), \quad (75)
\end{aligned}
$$

where the the second to last line follows from completing the square.[3] Integrating both sides results in the following inequality,

$$
\begin{aligned}
e^{\beta_t}f(X_t) + e^{\beta_t}\frac{\mu}{2}\|x - Z_t\|^2 &\le \int_0^t \dot{\beta}_t e^{\beta_t}[f(X_s) + \langle\nabla f(X_s), x - X_s\rangle + \frac{\mu}{2}\|x - X_s\|^2]\,ds \\
&\quad + e^{\beta_0}f(X_0) + e^{\beta_0}\frac{\mu}{2}\|x - X_0\|^2. \\
&\le \int_0^t \dot{\beta}_t e^{\beta_t}f(x)\,ds + e^{\beta_0}f(X_0) + e^{\beta_0}\frac{\mu}{2}\|x - X_0\|^2.
\end{aligned}
$$

Therefore,

$$
e^{\beta_t}f(X_t) + e^{\beta_t}\frac{\mu}{2}\|x - Z_t\|^2 \le (e^{\beta_t} - e^{\beta_0})f(x) + e^{\beta_0}f(X_0) + e^{\beta_0}\frac{\mu}{2}\|x - Z_0\|^2,
$$

and hence

$$
\{f(X_t) + \frac{\mu}{2}\|x - Z_t\|^2, e^{\beta_t}\}
$$

is a continuous-time estimate sequence. In Appendix D, we show the equivalence in discrete time between estimate sequences and our Lyapunov framework in this setting.

7 Discussion

In this paper, we have presented a Lyapunov framework for analyzing a variety of gradient-based methods for optimization. We showed that a single family of Lyapunov functions can be used to verify the convergence rates of a large collection of momentum methods, making Polyak's original physical intuition rigorous. We demonstrated that convergence rates could be understood as the consequence of discretization errors incurred when passing from continuous to discrete time. Consistently, implicit discretization schemes result in harder subproblems for algorithms, but provide

---

[3]We remark also that this derivation also shows how to derive the Lyapunov function (43) from the dynamics in (44).

an almost exact approximation of the continuous-time dynamics. On the other hand, when explicit discretization schemes are used, the algorithms incur discretization error, and require assumptions about the problem instance to guarantee good convergence properties.

We close with a discussion of some possible directions for future work. First, note that requiring that the continuous-time Lyapunov function remain a Lyapunov function in discrete time places significant constraints on which ODE solvers can be used. In this paper, we show that we can derive new algorithms using a restricted set of ODE techniques (several of which are nonstandard) but it remains to be seen if other methods can be applied in this setting. Techniques such as the midpoint method and Runge Kutta provide more accurate solutions of ODEs than Euler methods [4]. Is it possible to analyze such techniques as optimization methods? We expect that these methods do not achieve better asymptotic convergence rates, but may inherit additional favorable properties. Determining the advantages of such schemes could provide more robust optimization techniques in certain scenarios.

Several restart schemes have been suggested for the strongly convex setting based on the momentum dynamics (4). In many settings, while the Lipschitz parameter can be estimated using backtracking line-search, the strong convexity parameter is often hard—if not impossible—to estimate [26]. Therefore, many authors [22, 26, 12] have developed heuristics to empirically speed up the convergence rate of the ODE (or discrete-time algorithm), based on model misspecification. In particular, both Su, Boyd, and Candes [26] and Krichene, Bayen and Bartlett [12] develop restart schemes designed for the strongly convex setting based on the momentum dynamics (4). Our analysis suggests that restart schemes based on the dynamics (44) might lead to better results.

As mentioned in Section 5, the dynamics (44) developed for the setting when f is strongly convex can be viewed variationally, as the Euler-Lagrange equation for a different Lagrangian functional. While outside the scope of this paper, the Lagrangian (45) can be generalized to a non-Euclidean setting. In follow-up work, we plan to study this second Lagrangian family and its properties further. Other extensions of the Lyapunov analysis left for followup work are to nondeterministic settings such as accelerated coordinate ascent and stochastic gradient descent. A natural question we hope to answer is whether a dynamical perspective can be fruitfully applied in these settings.

Earlier work by Drori and Teboulle [7], Kim and Fessler [11], Taylor *et al* [27], and Lessard *et al* [13] have shown that optimization algorithms could be analyzed by solving convex programming problems. In particular, Lessard *et al* show that Lyapunov-like potential functions called *integral quadratic constraints* can be found by solving a constant-sized semidefinite programming problem. It would be interesting to see if these results can be adapted to directly search for Lyapunov functions like those studied in this paper. This would provide a method to automate the analysis of new techniques, possibly moving beyond momentum methods to novel families of optimization techniques.

Acknowledgements

References

[1] Francis R. Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 1(25):115–129, 2015.

[2] Michel Baes. Estimate sequence methods: Extensions and approximations. Manuscript, available at http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf, August 2009.

[3] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. *ArXiv preprint arXiv:1506.08187*, 2015.

[4] J.C. Butcher. Numerical methods for ordinary differential equations in the 20th century. *Journal of Computational and Applied Mathematics*, 125(1–2):1 – 29, 2000. Numerical Analysis 2000. Vol. VI: Ordinary Differential Equations and Integral Equations.

[5] P. L Chebyshev. Théorie des mécanismes connus sous le nom de parallélogrammes. *Mémoires présentés á l'Académie Impériale des Sciences de St-Pétersbourg*, VII(539-568), 1854.

[6] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[7] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, pages 1–32, 2013.

[8] Dmitry Drusvyatskiy, Maryam Fazel, and Scott Roy. An optimal first order method based on optimal quadratic averaging an optimal first order method based on optimal quadratic averaging. *ArXiv preprint arXiv:1604.06543*, 2016.

[9] M Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Res. Logis. Quart*, 3:95–110, 1956.

[10] Robert M. Freund and Paul Grigas. New analysis and results for the franke-wolfe method. Arxiv preprint arXiv, 2014.

[11] Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1):81–107, 2016.

[12] Walid Krichene, Alexandre Bayen, and Peter Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems (NIPS) 29*, 2015.

[13] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[14] A. M. Lyapunov and A. T. Fuller. General problem of the stability of motion, 1992.

[15] Arkadi Nemirovskii and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

[16] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Applied Optimization. Kluwer, Boston, 2004.

[17] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[18] Yurii Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[19] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[20] Yurii Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2015.

[21] Yurii Nesterov and Vladimir Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *J. Optimization Theory and Applications*, 165(3):917–940, 2015.

[22] Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

[23] Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[24] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 10 1986.

[25] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems (NIPS) 27*, 2014.

[26] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. *ArXiv e-prints arXiv:1503.01243*, 2015.

[27] Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, pages 1–39, 2016.

[28] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *ArXiv e-prints arXiv:1509.03616*, September 2015.

Appendix

We present the analyses for several results presented in the main text. The Appendix is organized as follows: Appendix A presents a review of mirror descent from a dynamical perspective. In particular, a Lyapunov analysis is derived similar to (12). This part of the presentation is based on [15]. Appendix B presents the proofs for all the algorithms in Section 4 using our Lyapunov framework and Appendix C presents results for the strong convexity setting. Finally, Appendix D presents results showing the equivalence between estimate sequences, as formulated by Baes [2] and our Lyapunov framework.

A   Mirror Descent Dynamics

We begin with a smooth manifold $\mathcal{X}$ with a local metric $g(x)$ and a function $f : \mathcal{X} \to \mathbb{R}$ we would like to minimize. The gradient flow associated with f is the flow induced by the differential equation,

$$\dot{X}_t = v(X_t), \tag{76}$$

where the vector field $v(X_t)$ is the "steepest descent" direction (the direction that makes f decrease the fastest):

$$v(x) = \arg\min_v \left\{ \langle \nabla f(X_t), v \rangle + \frac{1}{2}\|v\|^2_{g(X_t)} \right\}. \tag{77}$$

We can write the gradient flow equation explicitly as,

$$\dot{X}_t = -g(X_t)^{-1}\nabla f(X_t). \tag{78}$$

The dynamics which forms the basis for the mirror descent algorithm arises from a special structure that appears when the metric is chosen to be the Hessian of a strictly convex function h:

$$\dot{X}_t = -[\nabla^2 h(X_t)]^{-1}\nabla f(X_t). \tag{79}$$

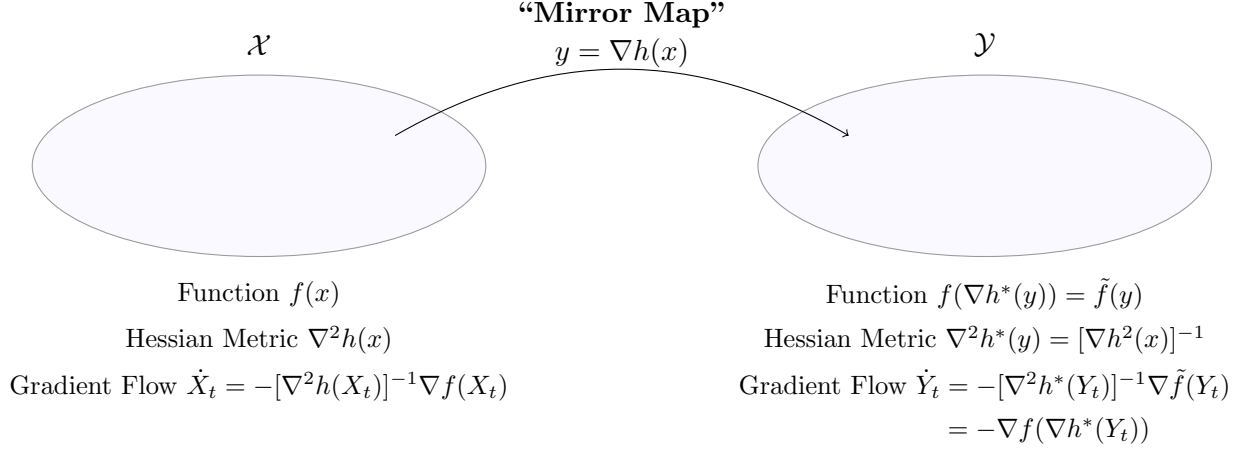which we can also write as,

$$\frac{d}{dt}\nabla h(X_t) = -\nabla f(X_t). \tag{80}$$

The ability to rewrite (79) as (80) is a manifestation of the following fact:

> When $h : \mathcal{X} \to \mathbb{R}$ is strictly convex and $g(x) = \nabla^2 h(x)$, there is a special map $\phi : \mathcal{X} \to \mathcal{Y}$ called the mirror map, given by $\phi = \nabla h$, whose push-forward maps a gradient flow on $\mathcal{X}$ to a gradient flow on $\mathcal{Y}$. [28]

It can be checked that the following holds under the mirror map:

"Mirror Map"
$$y = \nabla h(x)$$

$\mathcal{X}$ $\mathcal{Y}$

Function $f(x)$ Function $f(\nabla h^*(y)) = \tilde{f}(y)$

Hessian Metric $\nabla^2 h(x)$ Hessian Metric $\nabla^2 h^*(y) = [\nabla h^2(x)]^{-1}$

Gradient Flow $\dot{X}_t = -[\nabla^2 h(X_t)]^{-1} \nabla f(X_t)$ Gradient Flow $\dot{Y}_t = -[\nabla^2 h^*(Y_t)]^{-1} \nabla \tilde{f}(Y_t)$
$$= -\nabla f(\nabla h^*(Y_t))$$

Time-dilation The gradient flow equation does not explicitly depend on time. As in [28], let $\tau : \mathbb{T} \to \mathbb{T}'$ be a smooth twice-continuously differential function, where $\mathbb{T}' = \tau(\mathbb{T}) \subseteq \mathbb{R}$ is the image of $\mathbb{T}$. Given a curve $X : \mathbb{T}' \to \mathcal{X}$, we consider the reparameterized curve $Y : \mathbb{T} \to \mathcal{X}$ given by

$$Y_t = X_{\tau(t)}.$$

That is, the new curve is obtained by traversing the old curve at a new speed. If we consider the arbitrary time-dilation function $\tau(t) = e^{\beta(t)} := e^{\beta_t}$, where $\dot{\beta}_t > 0$, applied to the mirror descent dynamics (80), we obtain the following equation,

$$\frac{d}{dt}\nabla h(X_t) = -\dot{\tau}(t)\nabla f(X_t) = -\dot{\beta}_t e^{\beta_t}\nabla f(X_t). \tag{81}$$

Since the vector field explicitly depends on time, (81) is no longer a gradient flow. Nevertheless, a we can generate a Lyapunov function from the general time-dilated dynamics.

A.1 Lyapunov Analysis

Using the Bregman divergence of h,

$$D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle, \tag{82}$$

we illustrate how the special property of the Hessian metric provides a Lyapunov analysis of the mirror descent dynamic (80). First, notice that

$$\frac{d}{dt}D_h(x^*, X_t) = \frac{d}{dt}\left\{h(x^*) - h(X_t) - \langle \nabla h(X_t), x^* - X_t \rangle\right\} ds$$

$$= -\langle \nabla h(X_t), \dot{X}_t \rangle - \left\langle \frac{d}{dt}\nabla h(X_t), x^* - X_t \right\rangle + \langle \nabla h(X_t), \dot{X}_t \rangle$$

$$= -\left\langle \frac{d}{dt}\nabla h(X_t), x^* - X_t \right\rangle$$

$$\overset{(81)}{=} \dot{\beta}_t e^{\beta_t} \langle \nabla f(X_t), x^* - X_t \rangle$$

$$\leq \dot{\beta}_t e^{\beta_t}(f(X_t) - f(x^*))dt, \tag{83a}$$

28

where in the last step we have used the convexity of f. Jensen's inequality ensures a $O(e^{-\beta_t})$ convergence rate on the average iterate,

$$f\left(\frac{\int_0^t \dot{\beta}_s e^{\beta_s} X_s ds}{e^{\beta_t} - e^{\beta_0}}\right) - f(x^*) \leq \frac{-\int_0^t \frac{d}{ds} D_h(x^*, X_s) ds}{e^{\beta_t} - e^{\beta_0}}$$

$$\leq \frac{D_h(x^*, X_0) - D_h(x^*, X_t)}{e^{\beta_t} - e^{\beta_0}}$$

$$\leq \frac{D_h(x^*, X_0)}{e^{\beta_t} - e^{\beta_0}}.$$

However, notice that the primal form (79) of the mirror descent dynamics allows us to obtain a stronger guarantee,

$$\frac{d}{dt}\left(e^{\beta_t} f(X_t)\right) = \dot{\beta}_t e^{\beta_t} f(X_t) + e^{\beta_t} \langle \nabla f(X_t), \dot{X}_t \rangle$$

$$= \dot{\beta}_t e^{\beta_t} f(X_t) - \dot{\beta}_t e^{2\beta_t} \langle \nabla f(X_t), \nabla^2 h(X_t)^{-1} \nabla f(X_t) \rangle$$

$$= \dot{\beta}_t e^{\beta_t} f(X_t) - \dot{\beta}_t e^{2\beta_t} \frac{1}{2} \|\nabla f(X_t)\|_{*, X_t}^2$$

$$\leq \dot{\beta}_t e^{\beta_t} f(X_t). \tag{84a}$$

If we plug this into (83a), we obtain the following inequality,

$$\frac{d}{dt}\left(e^{\beta_t}(f(X_t) - f(x^*))\right) \leq \dot{\beta}_t e^{\beta_t}\left(f(X_t) - f(x^*)\right) \leq -\frac{d}{dt} D_h(x^*, X_t).$$

Integrating gives a Lyapunov function for the mirror descent dynamics (80).

B   Momentum Algorithms

B.1   Proof of Theorem 2

We use the Lyapunov function

$$E_k = A_k(f(x_k) - f(x^*)) + D_h(x^*, z_k)$$

29

to obtain a convergence guarantee for the quasi-monotone subgradient method (21). Denoting $D_{k+1,k}^* = D_h(x^*, z_{k+1}) - D_h(x^*, z_k)$, observe that

$$D_{k+1,k}^* = -D_h(z_{k+1}, z_k) + \langle \nabla h(z_{k+1}) - \nabla h(z_k), z_{k+1} - x^* \rangle \tag{85a}$$

$$\overset{(21b)}{\leq} -\frac{1}{2}\|z_k - z_{k+1}\|^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_{k+1} - x^* \rangle \tag{85b}$$

$$= -\frac{1}{2}\|z_k - z_{k+1}\|^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_{k+1} - z_k \rangle - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - x^* \rangle$$

$$\leq \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - x^* \rangle \tag{85c}$$

$$\overset{(21a)}{=} \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\left\langle \nabla f(x_{k+1}), \frac{A_{k+1}}{(A_{k+1} - A_k)}(x_{k+1} - x_k) + x_k - x^* \right\rangle$$

$$= \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|_*^2 - A_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle$$

$$+ (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle$$

$$= \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|_*^2 - A_k\langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle.$$

$$\leq A_k(f(x_k) - f(x_{k+1})) - (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|_*^2$$

$$= A_k(f(x_k) - f(x^*)) - A_{k+1}(f(x_{k+1}) - f(x^*)) + \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|_*^2,$$

where (85b) follows from the strong convexity assumption on h, and (85c) uses the Fenchel-Young inequality. Thus, we obtain the following bound,

$$E_{k+1} - E_k \leq \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|_*^2.$$

B.2 Proof of General Quasi-Monotone Method

In B.1 we have set the dual averaging term $\gamma_k \equiv 1$ for simplicity. In its more general form, we add an increasing sequence $\{\gamma_k\}_{k=1}^\infty$ to the algorithm, and (20b) can be written as,

$$z_k = \arg\min_{z \in \mathcal{X}} \left\{ \sum_{i=1}^k \alpha_i \langle \nabla f(x_i), z \rangle + \gamma_k D_h(z, z_0) \right\},$$

which satisfies the optimality condition,

$$\gamma_{k+1}\nabla h(x_{k+1}) - \gamma_k\nabla h(x_k) = -(A_{k+1} - A_k)\nabla f(x_{k+1}). \tag{87}$$

For this method, the notion of a "prox-center" is essential. We provide an analysis of this algorithm using the discrete Lyapunov function

$$E_k = A_k(f(x_k) - f(x^*)) + \gamma_k(D_h(x^*, z_k) - D_h(x^*, x_0)). \tag{88}$$

First, however, we analyze the dynamics and show how to obtain the discrete Lyapunov function (88) for the dual-averaged algorithm.

The dynamics   We analyze the momentum dynamics (4) with an additional increasing weighting term ($\dot{\gamma}_t \geq 0$), which we require to be positive, $\gamma_t \geq 0$, $\forall t \in \mathbb{R}$:

$$\frac{d}{dt}\left(\gamma_t \nabla h(Z_t)\right) = -\dot{\beta}_t e^{\beta_t} \nabla f(X_t) \tag{89a}$$

$$Z_t = X_t + \frac{1}{\beta_t}\dot{X}_t. \tag{89b}$$

With this additional weighting term we obtain the following bound,

$$
\begin{aligned}
\frac{d}{dt}\left(\gamma_t D_h(x^*, Z_t)\right) dt &= \dot{\gamma}_t[h(x^*) - h(Z_t)] - \dot{\gamma}_t\langle \nabla h(Z_t), x - Z_t\rangle - \gamma_t\left\langle \frac{d}{dt}\nabla h(Z_t), x - Z_t\right\rangle \\
&= \dot{\gamma}_t[h(x^*) - h(Z_t)] - \left\langle \frac{d}{dt}\left(\gamma_t \nabla h(Z_t)\right), x^* - Z_t\right\rangle \\
&\overset{(89)}{=} \gamma_t[h(x^*) - h(Z_t)] + \dot{\beta}_t e^{\beta_t}\left\langle \nabla f(X_t), x^* - X_t - \frac{1}{\dot{\beta}_t}\dot{X}_t\right\rangle \\
&\overset{(12)}{\leq} \dot{\gamma}_t[h(x^*) - h(Z_t)] - \frac{d}{dt}\left(e^{\beta_t}\left(f(X_t) - f(x^*)\right)\right).
\end{aligned}
\tag{90a}
$$

Prox-center   We now require $\mathcal{X}$ to be a compact set, and define X_0 to be the "prox-center" of $\mathcal{X}$:

$$X_0 = \arg\min_{x \in \mathcal{X}} h(x).$$

Without loss of generality, we choose $h(X_0) = 0$, so that $h(x) = D_h(x, X_0) \geq 0$, $\forall x \in \mathcal{X}$. Using this definition and rearranging (90), we obtain the following bound:

$$\frac{d}{dt}\left(e^{\beta_t}\left(f(X_t) - f(x^*)\right)\right) \leq \frac{d}{dt}\left(\gamma_t D_h(x^*, Z_0) - \gamma_t D_h(x^*, Z_t)\right),$$

from which we can conclude that the following function,

$$\mathcal{E}_t = e^{\beta_t}(f(X_t) - f(x^*)) + \gamma_t(D_h(x^*, Z_t) - D_h(x^*, Z_0)), \tag{91}$$

is a Lyapunov function. We also obtain the following convergence rate guarantee,

$$f(X_t) - f(X_0) \leq \frac{\gamma_t D_h(x^*, Z_0)}{e^{\beta_t}}. \tag{92}$$

The algorithm   Using the following equality,

$$
\begin{aligned}
\gamma_{k+1} D_h(x^*, z_{k+1}) - \gamma_k D_h(x^*, z_k) = \gamma_{k+1} D_h(z_{k+1}, z_k) &+ \langle \gamma_{k+1}\nabla h(z_{k+1}) - \gamma_k\nabla h(z_k), z_{k+1} - x^*\rangle \\
&+ (\gamma_{k+1} - \gamma_k)(h(x^*) - h(z_{k+1})),
\end{aligned}
\tag{93}
$$

we can analyze what happens when use the additional weighting term (87) in the quasi-monotone method, instead of (21b). We define $x_0 \in \mathcal{X}$ to be the *prox-center* over the set that is being optimized over. In discrete time, this amounts to the analogous condition that $h(x) \geq D_h(x, x_0) \geq \frac{\sigma}{2}\|x_0 - x\|^2$, $\forall x \in \mathcal{X}$, where we used the σ-strong convexity assumption on h in the last inequality.

31

For simplicity, we rescale h so that $\sigma = 1$.[4] By defining the prox-center of the set, notice that $h(z_{k+1}) \geq \frac{1}{2}\|x_0 - z_{k+1}\|^2$. Therefore, from (93), we obtain the bound

$$\gamma_{k+1}D_h(x^*, z_{k+1}) - \gamma_k D_h(x^*, z_k) \leq \gamma_k D_h(z_{k+1}, z_k) + \langle \gamma_{k+1}\nabla h(z_{k+1}) - \gamma_k \nabla h(z_k), z_{k+1} - x^* \rangle$$
$$+ (\gamma_{k+1} - \gamma_k)h(x^*). \tag{94}$$

With this inequality in hand, the proof of convergence is essentially equivalent to the proof of Theorem 2. In particular, denoting $D^*_{\gamma_{k+1}, \gamma_k} = \gamma_{k+1}D_h(x^*, z_{k+1}) - \gamma_k D_h(x^*, z_k)$, we obtain the following upper bound,

$$D^*_{\gamma_{k+1}, \gamma_k} \stackrel{(94)}{\leq} \gamma_k D_h(z_{k+1}, z_k) + \langle \gamma_{k+1}\nabla h(z_{k+1}) - \gamma_k \nabla h(z_k), z_{k+1} - x^* \rangle + (\gamma_{k+1} - \gamma_k)h(x^*)$$

$$\stackrel{(87)}{\leq} -\frac{\gamma_k}{2}\|z_k - z_{k+1}\|^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_{k+1} - x^* \rangle + (\gamma_{k+1} - \gamma_k)h(x^*)$$

$$= -\frac{\gamma_k}{2}\|z_k - z_{k+1}\|^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_{k+1} - z_k \rangle - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - x^* \rangle$$
$$+ (\gamma_{k+1} - \gamma_k)h(x^*)$$

$$\leq \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - x^* \rangle + (\gamma_{k+1} - \gamma_k)h(x^*)$$

$$\stackrel{(21a)}{=} \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), \frac{A_{k+1}}{(A_{k+1} - A_k)}(x_{k+1} - x_k) + x_k - x^* \rangle$$
$$+ (\gamma_{k+1} - \gamma_k)h(x^*)$$

$$= \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle$$
$$+ (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle - A_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle + (\gamma_{k+1} - \gamma_k)h(x^*)$$

$$= \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle - A_k\langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle$$
$$+ (\gamma_{k+1} - \gamma_k)h(x^*)$$

$$= -A_k(f(x_{k+1}) - f(x_k)) - (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2$$
$$+ (\gamma_{k+1} - \gamma_k)h(x^*)$$

$$= A_k(f(x_k) - f(x^*)) - A_{k+1}(f(x_{k+1}) - f(x^*)) + \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2$$
$$+ (\gamma_{k+1} - \gamma_k)D_h(x^*, x_0),$$

where in the last inequality, we use Fenchel-Young. Therefore, for (88) we can ensure the following

$$E_{k+1} - E_k \leq \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2 \leq \frac{(A_{k+1} - A_k)^2}{2\gamma_k}G^2,$$

---

[4]Note, this is equivalent to choosing $h(x_0) = 0$ and x_0 to be the minimizer of h ($\nabla h(x_0) = 0$).

where $h(x^*) = D_h(x^*, x_0)$ follows from the definition of the prox-center. Taking $\gamma_{-1} = \gamma_0 = 1$ and choosing $\sum_{i=1}^{k} \frac{(A_{i+1} - A_i)^2}{\gamma_{i-1}} < \infty$, we obtain the convergence rate

$$f(x_k) - f(x^*) \leq \frac{A_0(f(x_0) - f(x^*)) + \gamma_k D_h(x^*, x_0) + \frac{1}{2}\sum_{i=0}^{k} \frac{(A_{i+1} - A_i)^2}{\gamma_{i-1}} G^2}{A_k}.$$

This matches the bound Nesterov obtained in [21]. Note that a very similar analysis can be presented for Nesterov's dual averaging algorithm [19].

B.3 Proof of Discretizations

B.3.1 Proof of Method 1

We analyze the algorithm (23) using the Lyapunov function (16):

$$
\begin{aligned}
E_{k+1} - E_k &= A_{k+1}(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_k) - f(x^*)) - \langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1} \rangle \\
&\quad - D_h(z_{k+1}, z_k) \\
&= A_{k+1}(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_k) - f(x^*)) + (A_{k+1} - A_k)\langle \nabla f(x_k), x^* - x_{k+1} \rangle \\
&\quad - A_k\langle \nabla f(x_k), x_{k+1} - x_k \rangle - D_h(z_{k+1}, z_k) \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad (96a) \\
&= A_{k+1}(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_k) - f(x^*)) + (A_{k+1} - A_k)\langle \nabla f(x_k), x^* - x_k \rangle \\
&\quad + (A_{k+1} - A_k)\langle \nabla f(x_k), x_k - x_{k+1} \rangle - A_k\langle \nabla f(x_k), x_{k+1} - x_k \rangle - D_h(z_{k+1}, z_k) \\
&= A_{k+1}(f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle) + \alpha_k(f(x_k) - f(x^*) + \langle \nabla f(x_k), x^* - x_k \rangle) \\
&\quad - D_h(z_{k+1}, z_k) \\
&\leq A_{k+1}\frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&\overset{(24b)}{\leq} \frac{L}{2}\frac{A_{k+1}\alpha_k^2}{A_k^2} diam(\mathcal{X})^2, \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad (96b)
\end{aligned}
$$

where in (96a) we used (24b) and (24a). Therefore, we obtain the following upper bound,

$$f(x_k) - f(x^*) \leq \frac{D_h(x^*, x_0) + A_0(f(x_0) - f(x^*)) + \frac{L}{2}\sum_{i=1}^{k}\frac{A_{i+1}(A_{i+1} - A_i)^2}{A_i^2}diam(\mathcal{X})^2}{A_k}.$$

Choosing $A_k = \frac{k(k+1)}{2}$ gives a $O(1/k)$ convergence rate.

33

B.3.2 Proof of Method 2

We analyze (26) using the Lyapunov function (16). Denoting $D^*_{k+1,k} = D_h(x^*, z_{k+1}) - D_h(x^*, z_k)$, observe that

$$
\begin{aligned}
D^*_{k+1,k} &= -\langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1} \rangle - D_h(z_{k+1}, z_k) \\
&\overset{(26b)}{=} (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x^* - z_k \rangle + (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - z_{k+1} \rangle - D_h(z_{k+1}, z_k) \\
&\overset{(26a)}{=} (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x^* - x_{k+1} \rangle + A_k \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle \\
&\quad - D_h(z_{k+1}, z_k) \\
&\leq (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x^* - x_{k+1} \rangle + A_k \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle + \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2 \\
&\leq (A_{k+1} - A_k)(f(x^*) - f(x_{k+1})) + A_k(f(x_k) - f(x_{k+1})) + \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2 \\
&= A_k(f(x_k) - f(x^*)) - A_{k+1}(f(x_{k+1}) - f(x^*)) + \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2.
\end{aligned}
$$

Therefore,

$$
E_{k+1} - E_k \leq \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2 \leq \frac{(A_{k+1} - A_k)^2}{2}G^2
$$

and we have the following convergence rate guarantee

$$
f(x_k) - f(x^*) \leq \frac{D_h(x^*, x_0) + A_0(f(x_0) - f(x^*)) + \sum_{i=1}^{k} \frac{(A_{i+1} - A_i)^2}{2}G^2}{A_k},
$$

which is the same convergence rate as the quasi-monotone subgradient method (21).

B.4 Proof of Conditional Gradient Method

We use the Lyapunov function (38) to analyze conditional gradient method (37):

$$
\begin{aligned}
E_{k+1} - E_k &= A_{k+1}(f(x_{k+1}) - f(x^*)) - A_k(f(x_k) - f(x^*)) \\
&= A_{k+1}(f(x_{k+1}) - f(x_k)) - \alpha_{k+1}(f(x_k) - f(x^*)) \\
&\leq A_{k+1}\left(\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2\right) - \alpha_{k+1}(f(x_k) - f(x^*)) \\
&= A_{k+1}\left(\tau_k \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L\tau_k^2}{2}\|z_k - x_k\|^2\right) - \alpha_{k+1}(f(x_k) - f(x^*)) \\
&= \alpha_{k+1}\langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\frac{\alpha_{k+1}^2}{A_{k+1}}\|z_k - x_k\|^2 - \alpha_k(f(x_k) - f(x^*)) \\
&\leq \alpha_{k+1}(f(x^*) - f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle) + \frac{L}{2}\frac{\alpha_{k+1}^2}{A_{k+1}}\|z_k - x_k\|^2 \\
&\leq \frac{L}{2}\frac{\alpha_{k+1}^2}{A_{k+1}}diam(\mathcal{X})^2.
\end{aligned}
$$

Therefore,

$$E_k \le E_0 + \frac{L}{2} diam(\mathcal{X})^2 \sum_{i=0}^{k} \frac{\alpha_i^2}{A_k}.$$

Choosing $A_{k+1} = \frac{k(k+1)}{2}$ gives the coupling term $\tau_k = \frac{2}{k+2}$ and $O(1/k)$ convergence rate. This recovers the bounds shown by Nesterov [20].

Proof of extension Notice that our variational condition (42b) (taking $y = (1 - \tau_k)x_k + \tau_k x^*$, where as usual, $\tau_k = \frac{\alpha_k}{A_{k+1}}$, and $\alpha_k = A_{k+1} - A_k$) gives the inequality

$$-\alpha_k \langle \nabla f(x_k), x^* - x_k \rangle \le -A_{k+1} \langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k \rangle + \alpha_k \langle \nabla^2 f(x_k)(x_{k+1} - x_k), x^* - x_k \rangle$$
$$+ A_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle. \tag{99}$$

Noting this, we have

$$
\begin{aligned}
E_{k+1} - E_k &= A_{k+1}(f(x_{k+1}) - f(x_k)) - \alpha_k(f(x_k) - f(x^*)) \\
&\le A_{k+1}(f(x_{k+1}) - f(x_k)) - \alpha_k \langle \nabla f(x_k), x^* - x_k \rangle \\
&\overset{(99)}{\le} A_{k+1}(f(x_{k+1}) - f(x_k)) - A_{k+1} \langle \nabla f(x_k), x_{k+1} - x_k \rangle \\
&\quad - A_{k+1} \langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k \rangle + \alpha_k \langle \nabla^2 f(x_k)(x_{k+1} - x_k), x^* - x_k \rangle \\
&\le A_{k+1}(f(x_{k+1}) - f(x_k)) - A_{k+1} \langle \nabla f(x_k), x_{k+1} - x_k \rangle \\
&\quad - A_{k+1} \langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k \rangle + \frac{\alpha_k^2}{A_k} \langle \nabla^2 f(x_k)(z_k - x_k), z_k - x_k \rangle \\
&\le A_{k+1}(f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle - \frac{1}{2} \langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k \rangle) \\
&\quad + \frac{\alpha_k^2}{A_k} L D^2 \\
&\le A_{k+1} \frac{L}{6} \|x_{k+1} - x_k\|^3 + \frac{\alpha_k^2}{A_k} L D^2 \le \frac{\alpha_k^3}{A_k^2} \frac{L}{6} D^3 + \frac{\alpha_k^2}{A_k} L D^2,
\end{aligned}
$$

recovering the bound shown by Nesterov [20, (5.4)]. We remark that the following proof can easily be extended to incorporate the setting where f has Holder-continuous Hessians as in [20].

B.5   Proof of Accelerated Gradient Methods

We present the generalized accelerated gradient descent algorithm [2, 28] first outlined by Michel Baes (for the weakly convex setting). The algorithm maintains the following sequences:

Algorithm 9 Accelerated Gradient Descent (Weakly Convex Setting)

Assumptions: f, h are convex and differentiable. h satisfies smoothness condition $\frac{1}{p}\|x - y\|^p \leq D_h(x, y)$ and f satisfies smoothness condition $\|\nabla^p f\| \leq L$.

Choose $A_0 = 1$, $M > 0$, $\tilde{A}_{k+1} = L^{-1}A_{k+1}$, $\tau_k = \frac{\tilde{A}_{k+1} - \tilde{A}_k}{\tilde{A}_{k+1}} := \frac{\alpha_k}{\tilde{A}_{k+1}}$ and $x_0 = z_0 = y_0$. Define recursively,

$$x_{k+1} = \tau_k z_k + (1 - \tau_k)y_k \tag{101a}$$

$$z_k = \arg\min_{z \in \mathcal{X}} \left\{ \sum_{i=1}^{k} \alpha_i \langle \nabla f(y_i), z \rangle + D_h(z, z_0) \right\} \tag{101b}$$

y_k is a gradient update

$$G_{p,\epsilon,N}(x_k) = \arg\min_y \left\{ f_{p-1}(y; x_k) + \frac{LN}{p}\|y - x_k\|^p \right\}, \tag{101c}$$

where

$$f_{p-1}(y; x) = \sum_{i=0}^{p-1} \frac{1}{i!}\nabla^i f(x)(y - x)^i = f(x) + \langle \nabla f(x), y - x \rangle + \cdots + \frac{1}{(p-1)!}\nabla^{p-1} f(x)(y - x)^{p-1}$$

This algorithm satisfies the following optimality conditions,

$$z_k = y_k + \frac{\tilde{A}_{k+1}}{\tilde{A}_{k+1} - \tilde{A}_k}(x_{k+1} - y_k), \tag{102a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(\tilde{A}_{k+1} - \tilde{A}_k)\nabla f(y_{k+1}) \tag{102b}$$

$$M\|L^{-1}\nabla f(y_k)\|_*^{\frac{p}{p-1}} \leq L^{-1}\langle \nabla f(y_k), x_k - y_k \rangle \tag{102c}$$

where $M = \frac{(N^2-1)^{\frac{p-2}{2p-2}}}{2N}$. Denoting $D^*_{k+1,k} = D_h(x^*, z_{k+1}) - D_h(x^*, z_k)$, we obtain the following

bound:

$$D^*_{k+1,k} = -D_h(z_{k+1}, z_k) + \langle \nabla h(z_{k+1}) - \nabla h(z_k), z_{k+1} - x^* \rangle$$

$$\overset{(102b)}{\leq} -\frac{1}{p}\|z_k - z_{k+1}\|^p - (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), z_{k+1} - x^* \rangle \tag{103a}$$

$$= -\frac{1}{p}\|z_k - z_{k+1}\|^p - (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), z_{k+1} - z_k \rangle - (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), z_k - x^* \rangle$$

$$\leq \frac{p-1}{p}(\tilde{A}_{k+1} - \tilde{A}_k)^{\frac{p}{p-1}}\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}} - (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), z_k - x^* \rangle \tag{103b}$$

$$\overset{(102a)}{=} \frac{p-1}{p}(\tilde{A}_{k+1} - \tilde{A}_k)^{\frac{p}{p-1}}\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}} - (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), \frac{\tilde{A}_{k+1}}{(\tilde{A}_{k+1} - \tilde{A}_k)}x_{k+1} - y_k \rangle$$
$$- (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), y_k - x^* \rangle$$

$$= \frac{p-1}{p}(\tilde{A}_{k+1} - \tilde{A}_k)^{\frac{p}{p-1}}\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}} - \tilde{A}_{k+1}\langle \nabla f(y_{k+1}), x_{k+1} - y_k \rangle$$
$$- (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), y_{k+1} - x^* \rangle + (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), y_{k+1} - y_k \rangle$$

$$= \frac{p-1}{p}(\tilde{A}_{k+1} - \tilde{A}_k)^{\frac{p}{p-1}}\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}} - \tilde{A}_{k+1}\langle \nabla f(y_{k+1}), x_{k+1} - y_{k+1} \rangle$$
$$- (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), y_{k+1} - x^* \rangle + (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), y_{k+1} - y_k \rangle \tag{103c}$$
$$- \tilde{A}_{k+1}\langle \nabla f(y_{k+1}), y_{k+1} - y_k \rangle$$

$$= \frac{p-1}{p}(\tilde{A}_{k+1} - \tilde{A}_k)^{\frac{p}{p-1}}\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}} - \tilde{A}_{k+1}\langle \nabla f(y_{k+1}), x_{k+1} - y_{k+1} \rangle$$
$$+ (A_{k+1} - A_k)\langle \nabla \tilde{f}(y_{k+1}), x^* - y_{k+1} \rangle - \tilde{A}_k\langle \nabla f(y_{k+1}), y_{k+1} - y_k \rangle$$

$$\overset{(102c)}{\leq} \frac{p-1}{p}(\tilde{A}_{k+1} - \tilde{A}_k)^{\frac{p}{p-1}}\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}} - \tilde{A}_{k+1}L^{-\frac{1}{p-1}}M\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}}$$
$$+ (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(y_{k+1}), x^* - y_{k+1} \rangle - \tilde{A}_k\langle \nabla f(y_{k+1}), y_{k+1} - y_k \rangle$$

$$\leq -\tilde{A}_k(f(y_{k+1}) - f(y_k)) - (\tilde{A}_{k+1} - \tilde{A}_k)(f(y_{k+1}) - f(x^*))$$
$$+ \frac{p-1}{p}(\tilde{A}_{k+1} - \tilde{A}_k)^{\frac{p}{p-1}}\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}} - \tilde{A}_{k+1}L^{-\frac{1}{p-1}}M\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}}$$

$$= \tilde{A}_k(f(y_k) - f(x)) - \tilde{A}_{k+1}(f(y_{k+1}) - f(x^*))$$
$$+ \left(\frac{p-1}{p}(A_{k+1} - A_k)^{\frac{p}{p-1}} - A_{k+1}M\right)L^{-\frac{p}{p-1}}\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}},$$

where in (103a) we have used the uniform convexity assumption and in (103b) we have used the Fenchel-Young inequality. Therefore, for the general accelerated gradient methods we have the following bound

$$E_{k+1} - E_k \leq \varepsilon_k,$$

where the error scales in the following way

$$\varepsilon_k = \left(\frac{p-1}{p}(A_{k+1} - A_k)^{\frac{p}{p-1}} - A_{k+1}M\right)L^{-\frac{p}{p-1}}\|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}}.$$

Choosing the error to be nonpositive means that A_k can be at most a polynomial of degree p.

B.6 Accelerated Gradient Descent: Version 2

We now analyze the second version of the accelerated gradient descent algorithm using the same energy functional (30):

$$
\begin{aligned}
E_{k+1} - E_k &= \tilde{A}_{k+1}(f(y_{k+1}) - f(x^*)) - \tilde{A}_k(f(y_k) - f(x^*)) + D_h(x^*, z_{k+1}) - D_h(x^*, z_k) \\
&= \tilde{A}_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \tilde{A}_k(f(x_{k+1}) - f(y_k)) + (\tilde{A}_{k+1} - \tilde{A}_k)(f(x_{k+1}) - f(x^*)) \\
&\quad - \langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1} \rangle - D_h(z_{k+1}, z_k) \\
&\overset{(29b)}{\leq} \tilde{A}_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \tilde{A}_k(f(x_{k+1}) - f(y_k)) + (\tilde{A}_{k+1} - \tilde{A}_k)(f(x_{k+1}) - f(x^*)) \\
&\quad + (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(x_{k+1}), x^* - z_{k+1} \rangle - \frac{1}{2}\|z_{k+1} - z_k\|^2 \\
&= \tilde{A}_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \tilde{A}_k(f(x_{k+1}) - f(y_k)) + (\tilde{A}_{k+1} - \tilde{A}_k)(f(x_{k+1}) - f(x^*)) \\
&\quad + (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(x_{k+1}), x^* - z_k \rangle + (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(x_{k+1}), z_k - z_{k+1} \rangle - \frac{1}{2}\|z_{k+1} - z_k\|^2 \\
&\leq \tilde{A}_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \tilde{A}_k(f(x_{k+1}) - f(y_k)) + (\tilde{A}_{k+1} - \tilde{A}_k)(f(x_{k+1}) - f(x^*)) \\
&\quad + (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(x_{k+1}), x^* - z_k \rangle + \frac{(\tilde{A}_{k+1} - \tilde{A}_k)^2}{2}\|f(x_{k+1})\|^2 \\
&\overset{(29a)}{=} \tilde{A}_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \tilde{A}_k(f(x_{k+1}) - f(y_k)) + (\tilde{A}_{k+1} - \tilde{A}_k)(f(x_{k+1}) - f(x^*)) \\
&\quad + \tilde{A}_{k+1}\langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle + (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(x_{k+1}), x^* - y_k \rangle + \frac{(\tilde{A}_{k+1} - \tilde{A}_k)^2}{2}\|f(x_{k+1})\|^2 \\
&= \tilde{A}_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \tilde{A}_k(f(x_{k+1}) - f(y_k)) + (\tilde{A}_{k+1} - \tilde{A}_k)(f(x_{k+1}) - f(x^*)) \\
&\quad + \tilde{A}_k\langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle + (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(x_{k+1}), x^* - x_{k+1} \rangle + \frac{(\tilde{A}_{k+1} - \tilde{A}_k)^2}{2}\|f(x_{k+1})\|^2 \\
&\overset{(29c)}{\leq} -\frac{\tilde{A}_{k+1}}{2L}\|\nabla f(x_{k+1})\|^2 + \tilde{A}_k(f(x_{k+1}) - f(y_k)) + (\tilde{A}_{k+1} - \tilde{A}_k)(f(x_{k+1}) - f(x^*)) \\
&\quad + \tilde{A}_k\langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle + (\tilde{A}_{k+1} - \tilde{A}_k)\langle \nabla f(x_{k+1}), x^* - x_{k+1} \rangle + \frac{(\tilde{A}_{k+1} - \tilde{A}_k)^2}{2}\|f(x_{k+1})\|^2 \\
&\leq \frac{1}{2L^2}\left((A_{k+1} - A_k)^2 - A_{k+1}\right)\|\nabla f(x_{k+1})\|^2.
\end{aligned}
$$

Therefore we have shown

$$
E_{k+1} - E_k \leq \varepsilon_k,
$$

where

$$
\varepsilon_k = \frac{1}{2L^2}\left[(A_{k+1} - A_k)^2 - A_{k+1}\right]\|\nabla f(x_{k+1})\|^2.
$$

To ensure our error ε_k is nonpositive, A_k can be at most a polynomial of degree two.

C Strong Convexity

C.1 Proof of Accelerated Gradient (53)

$$\tilde{E}_{k+1} - \tilde{E}_k = f(y_{k+1}) - f(y_k) - \mu\langle z_{k+1} - z_k, x^* - z_{k+1}\rangle - \frac{\mu}{2}\|z_{k+1} - z_k\|^2$$

$$= f(y_{k+1}) - f(x_{k+1}) + f(x_{k+1}) - f(y_k) - \mu\langle z_{k+1} - z_k, x^* - z_k\rangle + \frac{\mu}{2}\|z_{k+1} - z_k\|^2$$

$$= f(y_{k+1}) - f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - y_k\rangle - \frac{\mu}{2}\|x_{k+1} - y_k\|^2$$

$$+ \tau_k\langle \nabla f(x_{k+1}) - \mu(x_{k+1} - z_k), x^* - z_k\rangle + \frac{\mu}{2}\|z_{k+1} - z_k\|^2$$

$$\overset{(53a)}{=} f(y_{k+1}) - f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_{k+1} - y_k\rangle - \frac{\mu}{2}\|x_{k+1} - y_k\|^2 + \tau_k\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle$$

$$- \tau_k\mu\langle x_{k+1} - z_k, x^* - z_k\rangle + \tau_k\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle + \frac{\mu}{2}\|z_{k+1} - z_k\|^2$$

$$\leq -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - x_{k+1}\|^2\right) + f(y_{k+1}) - f(x_{k+1}) - \tau_k\mu\langle x_{k+1} - z_k, x^* - z_k\rangle$$

$$+ \tau_k\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 + \langle \nabla f(x_{k+1}), x_{k+1} - y_k\rangle - \frac{\mu}{2}\|x_{k+1} - y_k\|^2$$

$$= -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + f(y_{k+1}) - f(x_{k+1}) + \tau_k\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle$$

$$+ \frac{\mu}{2}\|z_{k+1} - z_k\|^2 - \frac{\mu\tau_k}{2}\|z_k - x_{k+1}\|^2 + \langle \nabla f(x_{k+1}), x_{k+1} - y_k\rangle - \frac{\mu}{2}\|x_{k+1} - y_k\|^2$$

$$\overset{(53b)}{=} -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + f(y_{k+1}) - f(x_{k+1}) - \frac{\mu}{2\tau_k}\|y_k - x_{k+1}\|^2$$

$$+ \frac{\mu}{2}\|z_{k+1} - z_k\|^2 - \frac{\mu}{2}\|x_{k+1} - y_k\|^2$$

$$\overset{(53a)}{=} -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + f(y_{k+1}) - f(x_{k+1}) - \frac{\mu}{2\tau_k}\|y_k - x_{k+1}\|^2$$

$$+ \frac{\mu}{2}\|\tau_k(x_{k+1} - z_k)\|^2 + \tau_k\langle \nabla f(x_{k+1}), \tau_k(x_{k+1} - z_k)\rangle + \frac{\tau_k^2}{2\mu}\|\nabla f(x_{k+1})\|^2 - \frac{\mu}{2}\|x_{k+1} - y_k\|^2$$

$$\leq -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + f(y_{k+1}) - f(x_{k+1}) - \frac{\mu}{2\tau_k}\|y_k - x_{k+1}\|^2$$

$$+ \tau_k\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle + \frac{\tau_k^2}{2\mu}\|\nabla f(x_{k+1})\|^2$$

$$\leq -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) - \tau_k(f(y_k) - f(x_{k+1})) + \left(\frac{L\tau_k}{2} - \frac{\mu}{2\tau_k}\right)\|x_{k+1} - y_k\|^2$$

$$+ \left(\frac{\tau_k^2}{2\mu} - \frac{1}{2L}\right)\|\nabla f(x_{k+1})\|^2$$

$$= -\tau_k\left(f(y_k) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + \left(\frac{L\tau_k}{2} - \frac{\mu}{2\tau_k}\right)\|x_{k+1} - y_k\|^2 + \left(\frac{\tau_k^2}{2\mu} - \frac{1}{2L}\right)\|\nabla f(x_{k+1})\|^2.$$

Thus, we have shown

$$\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k\tilde{E}_k + \varepsilon_k, \tag{104}$$

where

$$\varepsilon_k = \left(\frac{L\tau_k}{2} - \frac{\mu}{2\tau_k}\right)\|x_{k+1} - y_k\|^2 + \left(\frac{\tau_k^2}{2\mu} - \frac{1}{2L}\right)\|\nabla f(x_{k+1})\|^2. \tag{105}$$

Choosing $\tau_k \leq 1/\sqrt{\kappa}$ ensures that $\varepsilon_k \leq 0$.

C.2   Proof of Accelerated Gradient (54)

$$
\begin{aligned}
\tilde{E}_{k+1} - \tilde{E}_k &= f(y_{k+1}) - f(y_k) - \mu\langle z_{k+1} - z_k, x^* - z_{k+1}\rangle - \frac{\mu}{2}\|z_{k+1} - z_k\|^2 \\
&= f(y_{k+1}) - f(x_k) + f(x_k) - f(y_k) - \mu\langle z_{k+1} - z_k, x^* - z_k\rangle + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 \\
&= f(y_{k+1}) - f(x_k) + \langle\nabla f(x_k), x_k - y_k\rangle - \frac{\mu}{2}\|x_k - y_k\|^2 - \mu\langle z_{k+1} - z_k, x^* - z_k\rangle + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 \\
&= f(y_{k+1}) - f(x_k) + \langle\nabla f(x_k), x_k - y_k\rangle - \frac{\mu}{2}\|x_k - y_k\|^2 + \tau_k\langle\nabla f(x_k) - \mu(x_k - z_k), x^* - z_k\rangle \\
&\quad + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 \\
&\overset{(54a)(54b)}{=} f(y_{k+1}) - f(x_k) + \langle\nabla f(x_k), x_k - y_k\rangle - \frac{\mu}{2}\|x_k - y_k\|^2 + \tau_k\langle\nabla f(x_k), x^* - x_k\rangle \\
&\quad - \tau_k\mu\langle x_k - z_k, x^* - z_k\rangle + \langle\nabla f(x_k), y_k - x_k\rangle + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 \\
&\leq -\tau_k\left(f(x_k) - f(x^*) + \frac{\mu}{2}\|x^* - x_k\|^2\right) + f(y_{k+1}) - f(x_k) - \frac{\mu}{2}\|x_k - y_k\|^2 \\
&\quad - \tau_k\mu\langle x_k - z_k, x^* - z_k\rangle + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 \\
&= -\tau_k\left(f(x_k) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + f(y_{k+1}) - f(x_k) - \frac{\mu}{2}\|x_k - y_k\|^2 \\
&\quad + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 - \frac{\tau_k\mu}{2}\|x_k - z_k\|^2 \\
&\overset{(54a)}{=} -\tau_k\left(f(x_k) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + f(y_{k+1}) - f(x_k) - \frac{\mu}{2}\|x_k - y_k\|^2 \\
&\quad + \frac{\mu}{2}\|\tau_k(x_k - z_k)\|^2 - \tau_k\langle\nabla f(x_k), \tau_k(x_k - z_k)\rangle + \frac{\tau_k^2}{2\mu}\|\nabla f(x_k)\|^2 - \frac{\mu}{2\tau_k}\|x_k - y_k\|^2 \\
&\overset{(54b)}{=} -\tau_k\left(f(x_k) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + f(y_{k+1}) - f(x_k) \\
&\quad - \tau_k\langle\nabla f(x_k), y_k - x_k\rangle + \frac{\tau_k^2}{2\mu}\|\nabla f(x_k)\|^2 - \frac{\mu}{2\tau_k}\|x_k - y_k\|^2 \\
&\leq -\tau_k\left(f(x_k) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + f(y_{k+1}) - f(x_k) - \tau_k(f(y_k) - f(x_k)) + \frac{\tau_k^2}{2\mu}\|\nabla f(x_k)\|^2 \\
&\quad + \left(\frac{\tau_k L}{2} - \frac{\mu}{2\tau_k}\right)\|x_k - y_k\|^2 \\
&= -\tau_k\left(f(y_k) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + f(y_{k+1}) - f(x_k) + \frac{\tau_k^2}{2\mu}\|\nabla f(x_k)\|^2 \\
&\quad + \left(\frac{\tau_k L}{2} - \frac{\mu}{2\tau_k}\right)\|x_k - y_k\|^2 \\
&\leq -\tau_k\left(f(y_k) - f(x^*) + \frac{\mu}{2}\|x^* - z_k\|^2\right) + \left(\frac{\tau_k^2}{2\mu} - \frac{1}{2L}\right)\|\nabla f(x_k)\|^2 + \left(\frac{\tau_k L}{2} - \frac{\mu}{2\tau_k}\right)\|x_k - y_k\|^2.
\end{aligned}
$$

If we choose $\tau_k \leq 1/\sqrt{\kappa}$ then we have shown

$$\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k\tilde{E}_k. \tag{106}$$

C.3 Derivation of Lyapunov Function

For completeness, we derive the Lyapunov function for the dynamics (44). From (75), the following inequality holds for (44):

$$\frac{d}{dt}\left(e^{\beta_t}\frac{\mu}{2}\|x - Z_t\|^2\right) = \dot{\beta}_t e^{\beta_t}\left(f(X_t) + \langle \nabla f(X_t), x - X_t\rangle + \frac{\mu}{2}\|x - X_t\|^2\right) - \frac{d}{dt}\left(e^{\beta_t}f(X_t)\right),$$

$$\leq \dot{\beta}_t e^{\beta_t}f(x) - \frac{d}{dt}\left(e^{\beta_t}f(X_t)\right),$$

$$= -\frac{d}{dt}\left(e^{\beta_t}(f(X_t) - f(x))\right).$$

Integrating and setting $x = x^*$ shows that (7) is a Lyapunov function for the dynamics (44).

D Estimate Sequences

D.1 The Quasi-Montone Subgradient Method

Equivalence in discrete time The discrete-time estimate sequence (64) for quasi-monotone subgradient method can be written:

$$\phi_{k+1}(x) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} := f(x_{k+1}) + \frac{1}{A_{k+1}}D_h(x, z_{k+1}) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}}$$

$$\overset{(64)}{=} (1 - \tau_k)\left(\phi_k(x) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \tau_k f_k(x)$$

$$= \left(1 - \frac{\alpha_k}{A_{k+1}}\right)\left(f(x_k) + \frac{1}{A_k}D_h(x, z_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \frac{\alpha_k}{A_{k+1}}f_k(x).$$

Multiplying through by A_{k+1}, we have

$$A_{k+1}\left(f(x_{k+1}) + \frac{1}{A_{k+1}}D_h(x, z_{k+1}) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}}\right) = (A_{k+1} - (A_{k+1} - A_k))\left(f(x_k) + \frac{1}{A_k}D_h(x, z_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right)$$

$$+ (A_{k+1} - A_k)f_k(x)$$

$$= A_k\left(f(x_k) + \frac{1}{A_k}D_h(x, z_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + (A_{k+1} - A_k)f_k(x)$$

$$\overset{(61)}{\leq} A_k f(x_k) + D_h(x, z_k) - \tilde{\varepsilon}_k + (A_{k+1} - A_k)f(x).$$

Rearranging, we obtain our Lyapunov argument $E_{k+1} \leq E_k + \varepsilon_{k+1}$ for (16):

$$A_{k+1}(f(x_{k+1}) - f(x)) + D_h(x, z_{k+1}) \leq A_k(f(x_k) - f(x)) + D_h(x, z_k) + \varepsilon_{k+1},$$

where $\varepsilon_{k+1} = \tilde{\varepsilon}_{k+1} - \tilde{\varepsilon}_k = \frac{A_{k+1} - A_k}{2}\|\nabla f(x_{k+1})\|^2$.

Going the other direction, from our Lyapunov analysis we can derive the following bound:

$$E_k \leq E_0 + \tilde{\varepsilon}_{k+1} \tag{107a}$$

$$A_k(f(x_k) - f(x)) + D_h(x, z_k) \leq A_0(f(x_0) - f(x)) + D_h(x, z_0) + \tilde{\varepsilon}_{k+1}$$

$$A_k\left(f(x_k) - \frac{1}{A_k}D_h(x, z_k)\right) \leq (A_k - A_0)f(x) + A_0\left(f(x_0) + \frac{1}{A_0}D_h(x^*, z_0)\right) + \tilde{\varepsilon}_{k+1}$$

$$A_k\phi_k(x) \leq (A_k - A_0)f(x) + A_0\phi_0(x) + \tilde{\varepsilon}_{k+1}. \tag{107b}$$

Rearranging, we obtain our estimate sequence (59) ($A_0 = 1$) with an additional error term:

$$\phi_k(x) \leq \left(1 - \frac{A_0}{A_k}\right) f(x) + \frac{A_0}{A_k} \phi_0(x) + \tilde{\varepsilon}_{k+1} \tag{108a}$$

$$\leq \left(1 - \frac{1}{A_k}\right) f(x) + \frac{1}{A_k} \phi_0(x) + \tilde{\varepsilon}_{k+1}. \tag{108b}$$

Note, given accelerated gradient method and quasi-monotone subgradient method have the same continuous-time limit, the equivalence between the Lyapunov argument and method estimate sequences holds for this setting.

D.2 The Conditional Gradient Method

Equivalence in discrete time The discrete-time estimate sequence (64) for conditional gradient method can be written:

$$\phi_{k+1}(x) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} := f(x_{k+1}) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} \stackrel{(64)}{=} (1 - \tau_k) \left(\phi_k(x) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \tau_k f_k(x)$$

$$\stackrel{\text{Table 1}}{=} \left(1 - \frac{\alpha_k}{A_{k+1}}\right) \left(f(x_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \frac{\alpha_k}{A_{k+1}} f_k(x).$$

Multiplying through by A_{k+1}, we have

$$A_{k+1} \left(f(x_{k+1}) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}}\right) = (A_{k+1} - (A_{k+1} - A_k)) \left(f(x_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + (A_{k+1} - A_k) f_k(x)$$

$$= A_k \left(f(x_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + (A_{k+1} - A_k) f_k(x)$$

$$\stackrel{(61)}{\leq} A_k f(x_k) - \tilde{\varepsilon}_k + (A_{k+1} - A_k) f(x).$$

Rearranging, we obtain our Lyapunov argument $E_{k+1} \leq E_k + \varepsilon_{k+1}$ for (38):

$$A_{k+1}(f(x_{k+1}) - f(x)) \leq A_k(f(x_k) - f(x)) + \varepsilon_{k+1},$$

where $\varepsilon_{k+1} = \tilde{\varepsilon}_{k+1} - \tilde{\varepsilon}_k = \frac{1}{2} \frac{(A_{k+1} - A_k)^2}{A_{k+1}} diam(\mathcal{X})^2$. Going the other direction, from our Lyapunov analysis we can derive the following bound:

$$E_k \leq E_0 + \tilde{\varepsilon}_{k+1}$$
$$A_k(f(x_k) - f(x)) \leq A_0(f(x_0) - f(x)) + \tilde{\varepsilon}_{k+1}$$
$$A_k f(x_k) \leq (A_k - A_0) f(x) + A_0 f(x_0) + \tilde{\varepsilon}_{k+1}$$
$$A_k \phi_k(x) \leq (A_k - A_0) f(x) + A_0 \phi_0(x) + \tilde{\varepsilon}_{k+1}.$$

Rearranging, we obtain our estimate sequence (59) ($A_0 = 1$) with an additional error term:

$$\phi_k(x) \leq \left(1 - \frac{A_0}{A_k}\right) f(x) + \frac{A_0}{A_k} \phi_0(x) + \tilde{\varepsilon}_{k+1} \tag{110a}$$

$$\leq \left(1 - \frac{1}{A_k}\right) f(x) + \frac{1}{A_k} \phi_0(x) + \tilde{\varepsilon}_{k+1}. \tag{110b}$$

D.3 Accelerated Gradient Descent (Strong Convexity)

The discrete-time estimate sequence (64) for accelerated gradient descent can be written:

$$\phi_{k+1}(x) := f(x_{k+1}) + \frac{\mu}{2}\|x - z_{k+1}\|^2$$

$$\overset{(64)}{=} (1 - \tau_k)\phi_k(x) + \tau_k f_k(x)$$

$$\overset{(61)}{\leq} (1 - \tau_k)\phi_k(x) + \tau_k f(x).$$

Therefore, we obtain the inequality $E_{k+1} - E_k \leq -\tau_k E_k$ for our Lyapunov function (55):

$$\phi_{k+1}(x) - \phi_k(x) \leq -\tau_k(\phi_k(x) - f(x))$$

$$f(x_{k+1}) + \frac{\mu}{2}\|x - z_{k+1}\|^2 - \left(f(x_k) + \frac{\mu}{2}\|x - z_{k+1}\|^2\right) \overset{\text{Table } 1}{\leq} -\tau_k\left(f(x_k) - f(x) + \frac{\mu}{2}\|x - z_{k+1}\|^2\right)$$

$$f(x_{k+1}) - f(x) + \frac{\mu}{2}\|x - z_{k+1}\|^2 - \left(f(x_k) - f(x) + \frac{\mu}{2}\|x - z_{k+1}\|^2\right) \leq -\tau_k\left(f(x_k) - f(x) + \frac{\mu}{2}\|x - z_{k+1}\|^2\right).$$

Going the other direction, we have,

$$E_{k+1} - E_k \leq -\tau_k E_k$$

$$\phi_{k+1} \leq (1 - \tau_k)\phi_k(x) + \tau_k f(x)$$

$$A_{k+1}\phi_{k+1} \leq A_k\phi_k + (A_{k+1} - A_k)f(x).$$

Summing over the right-hand side, we obtain the estimate sequence (59):

$$\phi_{k+1} \leq \left(1 - \frac{A_0}{A_{k+1}}\right)f(x) + \frac{A_0}{A_{k+1}}\phi_0(x)$$

$$= \left(1 - \frac{1}{A_{k+1}}\right)f(x) + \frac{1}{A_{k+1}}\phi_0(x).$$