A Lyapunov Analysis of Momentum Methods in Optimization

Ashia C. Wilson Benjamin Recht Michael I. Jordan

Abstract

Momentum methods play a central role in convex optimization. Several momentum methods are provably optimal and all use a technique called *estimate sequences* to analyze their behavior. The technique of estimate sequences has long been considered difficult to understand. 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 several existing algorithms, introduce a couple of new algorithms, and 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 [25], 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 [26].

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 [25]. Indeed, when applied to positive definite quadratic cost functions, Polyak's Heavy Ball method is equivalent to Chebyshev's Iterative Method [6]. Despite its intuitive elegance, Polyak's eigenvalue analysis does not apply globally for general convex functions. Indeed, Lessard *et al.* derived a simple one-dimensional counterexample where the standard Heavy Ball method does not converge [15].

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 [18]. However, to achieve such general applicability, Nesterov's proof techniques abandoned the physical intuition of Polyak [18]; 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 with understanding 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 sequences technique, several authors have recently proposed schemes to achieve acceleration without appealing to estimate sequences [10, 4, 15]. 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 [27, 14, 31]. 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 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 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 explain the need for the extragradient step inside Nesterov's method which does not appear in Polyak's method. Finally, we show there is an 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.

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:

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

corresponds to the equations of motion of an object with mass m in the presence of friction or viscosity as well as a potential f. Note that the fixed points of such dynamics are those where the gradient of f vanishes. The trajectories of this dynamical system will tend to have directions of motion that persist over time. In the limit as the mass goes to zero, and letting $\gamma \propto 1/m$, we recover a gradient-flow equation, which is the continuous-time limit of gradient descent [31]. 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). 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 [27], showed that the continuous-time limit of Nesterov's accelerated gradient descent method corresponds to the following second-order ODE,

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

This is similar to (2) but has a time-dependent damping term $\gamma_t = r\log t$. 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 [31, Sec 2], which does not correspond to a straightforward discretization technique; and they do not demonstrate what connection—if any—there is between the Lypaunov 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 [14] 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 [19]. In particular, their analysis entailed an additional smoothness assumption which is not necessary for the original algorithm [19], or for the original proof technique using estimate sequences.[1] See also Nemirovski and Yudin's original work [17] for a continuous-time perspective on mirror descent and a Lyapunov analysis of convergence.

Wibisono, Wilson, and Jordan [31] introduced a class of dynamical systems whose discretizations give rise to a general family of accelerated gradient algorithms, including gradient descent 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 [31, (78)], 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 [17]).

Wibisono *et al.* [31] imposed the following conditions on the time-scale functions α_t, β_t, and γ_t:

$$\dot{\beta}_t \;\le\; e^{\alpha_t} \tag{5a}$$

$$\dot{\gamma}_t \;=\; e^{\alpha_t}, \tag{5b}$$

refering 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. Wibosono *et al.* [31] 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.

3 Rate-Generating Lyapunov Functions

Lyapunov's method [16] 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

[1] In particular, they assume $\frac{\ell}{2}\|x-y\|^2 \le D_h(x,y) \le \frac{L}{2}\|x-y\|^2$. 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 $\frac{\ell}{2}\|x-y\|^2 \le D_h(x,y)$.

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 the optimization setting, 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. 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.* [31], 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 gradient descent and its non-Euclidean variants. 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 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 we 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 uses 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 $\dot\beta_t = e^{\alpha_t}$. Notice, now the entire 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 (4); this derivation is similar in spirit to the Lyapunov analysis of mirror descent by Nemirovski and Yudin.

$$\begin{aligned}
\frac{d}{dt}D_h\left(x, Z_t\right) &= \frac{d}{dt}\left(h(x) - h(Z_t) - \langle \nabla h(Z_t), x - Z_t\rangle\right)\\
&= -\langle \nabla h(Z_t), \dot Z_t\rangle - \left\langle \frac{d}{dt}\nabla h(Z_t), x - Z_t\right\rangle + \langle \nabla h(Z_t), \dot Z_t\rangle\\
&= -\left\langle \frac{d}{dt}\nabla h\left(Z_t\right), x - Z_t\right\rangle\\
&= \dot\beta_t e^{\beta_t}\left\langle \nabla f(X_t), x - X_t - \frac{1}{\dot\beta_t}\dot X_s\right\rangle & \text{(11a)}\\
&= \dot\beta_t e^{\beta_t}\langle \nabla f(X_t), x - X_t\rangle\; - e^{\beta_t}\langle \nabla f(X_t), \dot X_t\rangle dt\\
&= \dot\beta_t e^{\beta_t}\langle \nabla f(X_t), x - X_t\rangle\; - \frac{d}{dt}\left\{e^{\beta_t} f(X_t)\right\}\; + \dot\beta_t e^{\beta_t} f(X_t)\\
&= \dot\beta_t e^{\beta_t}[f(X_t) + \langle \nabla f(X_t), x - X_t\rangle] - \frac{d}{dt}\left\{e^{\beta_t} f(X_t)\right\}\\
&\le \dot\beta_t e^{\beta_t} f(x) - \frac{d}{dt}\left\{e^{\beta_t} f(X_t)\right\} & \text{(11b)}\\
&= -\frac{d}{dt}\left\{e^{\beta_t}\left(f(X_t) - f(x)\right)\right\}. & \text{(11c)}
\end{aligned}$$

Here (11a) uses the momentum dynamics (10b) and (10a). The inequality (11b) follows from the convexity of f. Rearranging terms and taking $x = x^*$, we have show 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 us to analyze several discretizations of momentum dynamics (4). In the following subsections, we demonstrate how to analyze the quasi-monotone subgradient method [23], the class of accelerated gradient methods [3, 31], and the general family of conditional gradient algorithms [22] 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,

$$Z_t = X_t + \frac{e^{\beta_t}}{\frac{d}{dt}e^{\beta_t}}\dot{X}_t, \tag{12a}$$

$$\frac{d}{dt}\nabla h(Z_t) = -\frac{d}{dt}\left(e^{\beta_t}\right)\nabla f(X_t). \tag{12b}$$

Taking $e^{\beta_t} = A_k$ and $\frac{d}{dt}e^{\beta_t} = \frac{A_{k+1}-A_k}{\delta}$, the implicit Euler method applied to (12a) results in the following sequence:

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

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

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

For equation (12b), the implicit Euler method results in the following sequence

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

whereas the explicit Euler method results in the sequence

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

In what follows, 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 straight-forward discretization technique, that can also be analyzed using the above Lyapunov analysis.

4.3 Momentum-Based Algorithms

We give a short analysis of the result of implicit Euler discretization of both (12a) and (12b). 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 $\tau_{k+1} = \frac{A_{k+1}-A_k}{A_k}$. Define recursively,

$$z_k = x_k + \frac{1}{\tau_k}(x_k - x_{k-1}), \tag{13a}$$

$$x_{k+1} = \underset{\substack{x\in\mathcal{X} \\ z=\frac{1+\tau_{k+1}}{\tau_{k+1}}x-\frac{1}{\tau_{k+1}}x_k}}{\arg\min} \left\{ f(x) + \frac{1}{A_{k+1}-A_k} D_h\left(z, z_k\right) \right\}, \tag{13b}$$

---

the optimality conditions for which constitute our discretization

$$z_{k+1} = x_{k+1} + \frac{A_k}{A_{k+1}-A_k}(x_{k+1} - x_k), \tag{14a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(A_{k+1} - A_k)\nabla f(x_{k+1}). \tag{14b}$$

To analyze this algorithm, we follow a similar argument to that in (11). First, we define our discrete-time energy function (6),

$$E_k = A_k(f(x_k) - f(x)) + D_h(x, z_k), \tag{15}$$

which is the result of making the same continuous to discrete time identifications. Note that,

$$\begin{aligned} E_{k+1} - E_k &= A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) - \langle \nabla h(z_{k+1}) - \nabla h(z_k), x - z_{k+1}\rangle - D_h(z_{k+1}, z_k) \\ &\overset{(14b)}{=} A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) + (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x - z_{k+1}\rangle - D_h(z_{k+1}, z_k) \\ &\overset{(14a)}{=} A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) + (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x - x_{k+1}\rangle \\ &\quad + A_k\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle - D_h(z_{k+1}, z_k) \\ &\le A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) + (A_{k+1} - A_k)(f(x) - f(x_{k+1})) \\ &\quad + A_k(f(x_k) - f(x_{k+1})) - D_h(z_{k+1}, z_k), \end{aligned} \tag{16a}$$

where the inequality follows from the convexity of f. From (16), we have shown that $E_{k+1} - E_k \le \varepsilon_k$, where the error $\varepsilon_k = -D_h(z_{k+1}, z_k) \le 0$ is negative. Taking $x = x^*$, writing $E_k \le E_0$ allows us to obtain the convergence rate guarantee,

$$f(x_k) - f(x^*) \le \frac{D_h(x^*, x_0) + A_0(f(x_0) - f(x^*))}{A_k}, \tag{17}$$

which is equivalent to (8), using the same discrete-time identifications.

The subequations corresponding to (13) may be difficult to solve. For reference, the proximal minimization method [7],

$$x_{k+1} = \arg\min_{x\in\mathcal{X}} \left\{ f(x) + \frac{1}{A_{k+1}-A_k} D_h\left(x, x_k\right) \right\}, \tag{18}$$

involves applying the implicit Euler method to the mirror descent dynamic, 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 [23] as an alternative to the dual averaging method. It also achieves a $O(1/\sqrt{k})$ 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.* 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 (12a) and the implicit Euler method applied to (12b) (where now, we take $e^{\beta_t} = A_{k+1}$):

Algorithm 2 The Quasi-Monotone Subgradient 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,

$$x_{k+1} = \tau_k z_k + (1-\tau_k)x_k, \tag{19a}$$

$$z_k = \arg\min_{z\in\mathcal{X}} \left\{ \sum_{i=1}^{k} \alpha_i \langle \nabla f(x_i), z\rangle + D_h(z, x_0) \right\}. \tag{19b}$$

This modified quasi-monotone subgradient method has optimality conditions

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

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(A_{k+1}-A_k)\nabla f(x_{k+1}). \tag{20b}$$

The following result illustrates how the Lyapunov function can be used to analyze (20):

Theorem 1. *Using the Lyapunov function (15), we can show the quasi-monotone method as defined above 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:

$$f(x_k) - f(x^*) \leq \frac{D_h(x^*, x_0) + A_0(f(x_0) - f(x^*)) + \frac{1}{2}\sum_{i=0}^{k}\varepsilon_i}{A_k}. \tag{21}$$

If we optimize the bound (21) 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 (21). 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 [23] 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 (12), and analyze them using the discretized Lyapunov function (15). The proofs of these results can be found in the Appendix B.3. The first method is the result of applying the explicit Euler method to (12b) and the implicit Euler method to (12a):

Algorithm 3 Method 1

Assumptions: f is smooth 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,

$$z_{k+1} = \arg\min_{z \in \mathcal{X}} \left\{ \langle \nabla f(x_k), z \rangle + \frac{1}{\alpha_k} D_h(z, z_k) \right\} \tag{22a}$$

$$x_{k+1} = \frac{\tau_k}{\tau_k - 1} z_{k+1} + \frac{1}{\tau_k - 1} x_k \tag{22b}$$

This algorithm has optimality conditions,

$$z_{k+1} = x_{k+1} - \frac{A_k}{A_{k+1} - A_k}(x_{k+1} - x_k), \tag{23a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(A_{k+1} - A_k) \nabla f(x_k). \tag{23b}$$

Choosing $A_k = k(k+1)/2$ results in an $O(1/k)$ convergence rate. Note that structurally this algorithm is 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 updates,

[2] See [23, 2] for more details on this history on subgradient methods.

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,

$$x_{k+1} = \frac{\tau_k}{\tau_k+1} z_k + \frac{1}{\tau_k+1} x_k, \tag{24a}$$

$$z_{k+1} = \arg\min_{z\in\mathcal{X}} \left\{ \langle \nabla f(x_{k+1}), z\rangle + \frac{1}{\alpha_k} D_h(z, z_k) \right\}. \tag{24b}$$

This algorithm has optimality conditions,

$$z_k = x_{k+1} + \frac{A_k}{A_{k+1} - A_k}(x_{k+1} - x_k), \tag{25a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(A_{k+1} - A_k)\nabla f(x_{k+1}). \tag{25b}$$

Notice, (25) is very similar to the quasi-monotone subgradient method (19). 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 (the accelerated gradient descent algorithm in the setting where $h = \frac{1}{2}\|\cdot\|^2$). Neither of these correspond to a straightforward discretization of (6)—they both entail introducing an additional sequence $\{y_k\}$ to obtain a better bound on the error. The first was the version was introduced by Michel Baes [3]:

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,

$$z_k = \arg\min_{z\in\mathcal{X}} \left\{ \sum_{i=1}^{k} \alpha_i \langle \nabla f(y_i), z\rangle + D_h(z, z_0) \right\} \tag{26a}$$

$$y_k = \arg\min_{x\in\mathcal{X}} \left\{ \langle \nabla f(x_k), x\rangle + \frac{L}{4M}\|y - x_k\|^2 \right\} \tag{26b}$$

$$x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k \tag{26c}$$

This algorithm satisfies the following optimality conditions:

$$z_k = y_k + \frac{\tilde{A}_{k+1}}{\tilde{A}_{k+1} - \tilde{A}_k}(x_{k+1} - y_k), \tag{27a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(\tilde{A}_{k+1} - \tilde{A}_k)\nabla f(y_{k+1}) \tag{27b}$$

$$M\|L^{-1}\nabla f(y_k)\|_*^2 \leq L^{-1}\langle \nabla f(y_k), x_k - y_k\rangle \tag{27c}$$

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 [19], entails computing the mirror descent update (26a) from the gradient at x:

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

which results in an algorithm with the following optimality conditions,

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

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

$$f(y_k) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|_*^2 \tag{28c}$$

As mentioned above, to obtain a convergence guarantee we have replaced the current state x_k by another sequence y_k which simply needs to satisfy (27c). 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 (15) which we sum up in the following theorem:

Theorem 2. *Using the following Lyapunov function*

$$E_k = \tilde{A}_k(f(y_k) - f(x)) + D_h(x, z_k), \tag{4}$$

we can show

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

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 (26).

Choosing $x = x^*$ and A_k such that $\tilde{A}_{k+1}^{-1}(\tilde{A}_{k+1} - \tilde{A}_k)^2 \leq 2M$ so that the error is non-positive, we obtain the following convergence rate,

$$f(y_k) - f(x^*) \leq \frac{\tilde{A}_0(f(x_0) - f(x^*)) + D_h(x^*, x_0)}{\tilde{A}_k}. \tag{5}$$

Note, this condition requires that A_k can grow at most quadratically which gives an $O(1/k^2)$ rate of convergence. 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 [11], has garnered renewed interest over the last several years. This interest has inspired several different analyses [12, 2] of the algorithm. Of particular note, is the analysis provided by Nesterov [22], 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

$$0 \le \langle \nabla f(X_t), x - Z_t \rangle, \qquad \forall x \in \mathcal{X} \tag{6}$$

where

$$Z_t = X_t + \frac{e^{\beta_t}}{\frac{d}{dt} e^{\beta_t}} \dot{X}_t, \tag{7}$$

then the Lyapunov analysis (11) follows without the use of the Bregman divergence:

$$\begin{aligned} 0 &\le \dot{\beta}_t e^{\beta_t} \langle \nabla f(X_t), x - Z_t \rangle \\ &\le \dot{\beta}_t e^{\beta_t} f(x) - \frac{d}{dt}\Big(e^{\beta_t} f(X_t) \Big) \\ &= -\frac{d}{dt}\Big(e^{\beta_t} \left(f(X_t) - f(x) \right) \Big). \end{aligned}$$

Integrating shows that the following function

$$\mathcal{E}_t = e^{\beta_t} (f(X_t) - f(x)). \tag{9}$$

is a Lyapunov function for any family of dynamics satisfying (6) and (7). We now turn our attention to the conditional gradient method, which maintains the following iterates:

Algorithm 6 Conditional Gradient Method

Assumptions: f is smooth and $\mathcal{X}$ is convex and compact.
Choose $A_0 = 1$, $x_0 = z_0$ $\tau_k = \frac{A_{k+1} - A_k}{A_{k+1}}$.

$$\begin{aligned} z_k &= \arg\min_{z \in \mathcal{X}} \langle \nabla f(x_k), z \rangle, \\ x_{k+1} &= \tau_k z_k + (1 - \tau_k) x_k. \end{aligned}$$

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

$$0 \le \langle \nabla f(x_k), x - z_k \rangle \quad \forall x \in \mathcal{X}, \tag{11a}$$

$$z_k = x_k + \frac{A_{k+1}}{A_{k+1} - A_k} (x_{k+1} - x_k). \tag{11b}$$

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

Theorem 3. *Using the Lyapunov function*

$$E_k = A_k(f(x_k) - f(x^*)), \tag{12}$$

we can show that the conditional gradient method satisfies

$$E_{k+1} - E_k \le \varepsilon_k, \tag{13}$$

where the error scales in the following way,

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

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 [12, 2]. In particular, choosing $A_k = \frac{k(k+1)}{2}$ *gives the standard* $O(1/k)$ *convergence rate (see Nesterov [22, 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 (6) and (7),

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

results in an algorithm for which we can show $E_{k+1} - E_k \le 0$ using (12). 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 [22, (5.1)]

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

$$x_{k+1} \in \arg\min_{y=(1-\tau_k)x_k+\tau_k x} \left\{\langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2}\langle \nabla f(x_k)(y - x_k), y - x_k\rangle : x \in dom(\mathcal{X})\right\} \tag{15a}$$

This algorithm has the following optimality conditions

$$z_k = x_k + \frac{A_{k+1}}{A_{k+1} - A_k}(x_{k+1} - x_k), \tag{16a}$$
$$\langle \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k), y - x_{k+1}\rangle \ge 0, \tag{16b}$$
$$\forall y = (1 - \tau_k)x_k + \tau_k x, \quad x \in dom(\mathcal{X}). \tag{16c}$$

In Appendix B.4 we show how to recover the bound shown by Nesterov using the Lyapunov function (12).

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 dynamic. 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),

$$\mathcal{E}_t = e^{\beta_t}\left(f(X_t) - f(x^*) + \frac{\mu}{2}\|x^* - Z_t\|^2\right), \tag{17}$$

and use it to analyze the following dynamic:

$$Z_t = X_t + \frac{1}{\dot{\beta}_t}\dot{X}_t \tag{18a}$$

$$\mu\dot{Z}_t = -\mu\dot{X}_t - \dot{\beta}_t\nabla f(X_t). \tag{18b}$$

We briefly remark that (18) is the Euler-Lagrange equation for the following Lagrangian,

$$\mathcal{L}(x,\dot{x},t) = \dot{\beta}_t e^{2\beta_t}\left(\frac{\mu}{2}\left\|\frac{\dot{x}}{\dot{\beta}_t}\right\|^2 - f(x)\right), \tag{19}$$

and that this Lagrangian can generalized to include non-Euclidean geometries; however, further exploration of this Lagrangian is outside the scope of this work. Here, we demonstrate that (17) can be used to show an $O(e^{-\beta_t})$ rate of convergence for (18) 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

$$\tilde{\mathcal{E}}_t = f(X_t) - f(x^*) + \frac{\mu}{2}\|x^* - Z_t\|^2, \tag{20}$$

then (17) is a Lyapunov function. To that end, we have the following argument:

$$
\begin{aligned}
\dot{\mathcal{E}}_t &= \langle \nabla f(X_t), \dot{X}_t \rangle - \mu \left\langle \dot{Z}_t, x^* - X_t - \frac{1}{\dot{\beta}_t}\dot{X}_t \right\rangle \\
&= \langle \nabla f(X_t), \dot{X}_t \rangle + \left\langle \mu \dot{X}_t + \dot{\beta}_t \nabla f(X_t), x^* - X_t - \frac{1}{\dot{\beta}_t}\dot{X}_t \right\rangle \\
&= \dot{\beta}_t \langle \nabla f(X_t), x^* - X_t \rangle + \mu \left\langle \dot{X}_t, x^* - X_t - \frac{1}{\dot{\beta}_t}\dot{X}_t \right\rangle \\
&\leq -\dot{\beta}_t \left(f(X_t) - f(x^*) + \frac{\mu}{2}\|x^* - X_t\|^2 \right) + \mu \left\langle \dot{X}_t, x^* - X_t - \frac{1}{\dot{\beta}_t}\dot{X}_t \right\rangle \qquad (21a) \\
&= -\dot{\beta}_t \left(f(X_t) - f(x^*) + \frac{\mu}{2}\left\| x^* - X_t - \frac{1}{\dot{\beta}_t}\dot{X}_t \right\|^2 \right) - \mu \left\langle \dot{X}_t, x^* - X_t - \frac{1}{\dot{\beta}_t}\dot{X}_t \right\rangle - \frac{\mu}{2\dot{\beta}_t}\|\dot{X}_t\|^2 \\
&\quad + \mu \left\langle \dot{X}_t, x^* - X_t - \frac{1}{\dot{\beta}_t}\dot{X}_t \right\rangle \\
&\leq -\dot{\beta}_t \left(f(X_t) - f(x^*) + \frac{\mu}{2}\|x^* - Z_t\|^2 \right),
\end{aligned}
$$

where (21a) uses our strong convexity assumption. In the following subsections, we demonstrate how to discretize the dynamic (18), and how this discretization can be analyzed using a Lyapunov argument analogous to (21).

5.2 Discretizing the Dynamics

As a proof of concept, we begin by analyzing an implicit discretization of the dynamics (18). 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 $\tau_{k+1} = \frac{A_{k+1}-A_k}{A_{k+1}}$. Define recursively,

$$
z_k = x_k + \frac{1}{\tau_{k-1}}(x_k - x_{k-1}) \qquad (22a)
$$

$$
x_{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f(x) - \frac{\mu}{2\tau_k}\|\tau_k(x_k - z_k) - (x - x_k)\|^2 \right\}, \qquad (22b)
$$

The optimality conditions of this algorithm corresponds to our discretization:

$$
z_{k+1} - z_k = \tau_k \left(x_{k+1} - z_{k+1} - \frac{1}{\mu}\nabla f(x_{k+1}) \right) \qquad (23a)
$$

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

Using the discrete-time Lyapunov function

$$\tilde{E}_k = f(x_k) - f^* + \frac{\mu}{2}\|x^* - z_k\|^2, \tag{24}$$

notice that a similar argument to (21) holds:

$$\begin{aligned}
\tilde{E}_{k+1} - \tilde{E}_k &= f(x_{k+1}) - f(x_k) - \mu\langle z_{k+1} - z_k, x^* - z_{k+1}\rangle - \frac{\mu}{2}\|z_{k+1} - z_k\|^2 \\
&\overset{(23a)}{=} f(x_{k+1}) - f(x_k) + \tau_k\langle \nabla f(x_{k+1}) - \mu(x_{k+1} - z_{k+1}), x^* - z_{k+1}\rangle - \frac{\mu}{2}\|z_{k+1} - z_k\|^2 \\
&= \tau_k\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + f(x_{k+1}) - f(x_k) + \tau_k\langle \nabla f(x_{k+1}), x_{k+1} - z_{k+1}\rangle - \frac{\mu}{2}\|z_{k+1} - z_k\|^2 \\
&\quad + \mu\tau_k\langle z_{k+1} - x_{k+1}, x^* - z_{k+1}\rangle \\
&\leq -\tau_k\left(f(x_k) - f(x^*) + \frac{\mu}{2}\|x^* - x_{k+1}\|^2\right) + \tau_k\langle \nabla f(x_{k+1}), x_{k+1} - z_{k+1}\rangle \\
&\quad - \frac{\mu}{2}\|z_{k+1} - z_k\|^2 + \mu\tau_k\langle z_{k+1} - x_{k+1}, x^* - z_{k+1}\rangle \\
&\leq -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_{k+1}\|^2\right) + f(x_{k+1}) - f(x_k) + \tau_k\langle \nabla f(x_{k+1}), x_{k+1} - z_{k+1}\rangle \\
&\quad - \frac{\mu}{2}\|z_{k+1} - z_k\|^2. \\
&\overset{(23b)}{=} -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_{k+1}\|^2\right) + f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle \\
&\quad - \frac{\mu}{2}\|z_{k+1} - z_k\|^2. \\
&\leq -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_{k+1}\|^2\right).
\end{aligned}$$

Therefore,

$$\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k\left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x^* - z_{k+1}\|^2\right) = -\tau_k\tilde{E}_{k+1}. \tag{26}$$

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 $\{y_k\}_{k=1}^\infty$ in order to bound the error. We find that we can use a Lyapunov argument to analyze the following two sequences:

$$z_{k+1} - z_k = \tau_k\left(x_{k+1} - z_k - \frac{1}{\mu}\nabla f(x_{k+1})\right), \tag{27a}$$

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

$$y_{k+1} = x_{k+1} - \frac{1}{L}\nabla f(x_{k+1}), \tag{27c}$$

and

$$z_{k+1} - z_k = \tau_k\left(x_k - z_k - \frac{1}{\mu}\nabla f(x_k)\right), \tag{28a}$$

$$\tau_k(x_k - z_k) = y_k - x_k \tag{28b}$$

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k). \tag{28c}$$

The second sequence is the accelerated gradient scheme introduced by Nesterov [18, (2.2.8)], which can be further simplified into two sequences. We summarize our results in the following theorem:

Theorem 4. *Using the following Lyapunov function,*

$$\tilde{E}_k = f(y_k) - f(x^*) + \frac{\mu}{2}\|z_k - x^*\|^2, \tag{29}$$

we can show

$$\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k \tilde{E}_k + \varepsilon_k \tag{30}$$

for both methods (27) *and* (28)*, where*

$$\varepsilon_k = \left(\frac{\tau_k^2}{2\mu} - \frac{1}{2L}\right)\|\nabla f(x_{k+1})\|^2 + \left(\frac{\tau_k L}{2} - \frac{\mu}{2\tau_k}\right)\|x_{k+1} - y_k\|^2 \tag{31}$$

and

$$\varepsilon_k = \left(\frac{\tau_k^2}{2\mu} - \frac{1}{2L}\right)\|\nabla f(x_k)\|^2 + \left(\frac{\tau_k L}{2} - \frac{\mu}{2\tau_k}\right)\|x_k - y_k\|^2 \tag{32}$$

for (27) *and* (28) *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.

6 Estimate Sequences

In this section, we connect our Lyapunov framework 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 discrete time.

We begin with a brief reivew of the technique of estimate sequences. In [18], Nesterov introduced estimate sequences by giving the following definition

Definition 1. *[18, 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*

$$\phi_k(x) \leq \left(1 - \frac{1}{A_k}\right)f(x) + \frac{1}{A_k}\phi_0(x). \tag{33}$$

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

Lemma 5. *[18, 2.2.1] If for some sequence* $\{x_k\}_{k\geq 0}$ *we have*

$$f(x_k) \leq \phi_k^* \equiv \min_{x\in\mathcal{X}} \phi_k(x), \tag{34}$$

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

Proof. The proof is straightforward:

$$f(x_k) \le \phi_k^* \equiv \min_{x\in\mathcal{X}} \phi_k(x) \overset{(33)}{\le} \min_{x\in\mathcal{X}} \left[\left(1 - \frac{1}{A_k}\right) f(x) + \frac{1}{A_k}\phi_0(x)\right]$$
$$\overset{(34)}{\le} \left(1 - \frac{1}{A_k}\right) f(x^*) + \frac{1}{A_k}\phi_0(x^*).$$

Rearranging gives the desired inequality. □

Notice, this definition is not constructive. Finding sequences which satisfy these conditions is generally a non-trivial task. The next proposition, formalized by Baes in [3] as an extension of Nesterov's Lemma 2.2.2 [18], provides guidance for constructing estimate sequences. This construction is used in [18, 19, 20, 3, 23, 22], and is the only known way so far to construct an estimate sequence.

Proposition 6. *[3, 2.2] Let $\phi_0 : \mathcal{X} \to \mathbb{R}$ be a convex function such that $\min_{x\in\mathcal{X}} \phi_0(x) \ge f^*$. Suppose also that we have a sequence $\{f_k\}_{k\ge 0}$ of functions form $\mathcal{X}$ to $\mathbb{R}$ that underestimates f:*

$$f_k(x) \le f(x) \quad \text{for all } x \in \mathcal{X} \text{ and all } k \ge 0 \tag{35}$$

Define recursively $A_0 = 1$

$$\alpha_k := A_{k+1} - A_k \tag{36}$$
$$\tau_k := \frac{a_k}{A_{k+1}}, \tag{37}$$

and

$$\phi_{k+1}(x) := (1-\tau_k)\phi_k(x) + \tau_k f_k(x) = \frac{1}{A_{k+1}}\left(A_0\phi_0(x) + \sum_{i=0}^{k} a_i f_i(x)\right) \tag{38}$$

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

From (34) and (38), we observe that the following invariant,

$$A_{k+1} f(x_{k+1}) \le \min_x A_{k+1}\phi_{k+1}(x) = \min_x \sum_{i=0}^{k} \alpha_i f_i(x) + A_0\phi_0(x) \tag{39}$$

is maintained. In [23, 22], this technique was extended to incorporate a error term $\{\tilde{\varepsilon}_k\}_{k=1}^{\infty}$,

$$\phi_{k+1}(x) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} := (1-\tau_k)\left(\phi_k(x) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \tau_k f_k(x) = \frac{1}{A_{k+1}}\left(A_0(\phi_0(x) - \tilde{\varepsilon}_0) + \sum_{i=0}^{k} a_i f_i(x)\right) \tag{40}$$

Rearranging, we have

$$A_{k+1} f(x_{k+1}) \le \min_x A_{k+1}\phi_{k+1}(x) = \min_x \sum_{i=0}^{k} \alpha_i f_i(x) + A_0\left(\phi_0(x) - \frac{\tilde{\varepsilon}_0}{A_0}\right) + \tilde{\varepsilon}_{k+1}. \tag{41}$$

Notice the similar argument to Lemma 5 holds,

$$
\begin{aligned}
A_{k+1}f(x_{k+1}) &\leq \sum_{i=0}^{k}\alpha_i f_i(x^*) + A_0(\phi_0(x^*) - \tilde{\varepsilon}_0) + \tilde{\varepsilon}_{k+1} \\
&\overset{(35)}{\leq} \sum_{i=0}^{k}\alpha_i f(x^*) + A_0\phi_0(x^*) + \tilde{\varepsilon}_{k+1} \\
&\overset{(36)}{=} A_{k+1}f(x^*) + A_0\phi_0(x^*) + \tilde{\varepsilon}_{k+1}. \qquad (42a)
\end{aligned}
$$

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

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

we see that 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 analyzed using estimate sequences:

Algorithm	$f_i(x)$	$\phi_k(x)$	$\tilde{\varepsilon}_k$
Quasi-Monotone Subgradient Method	linear	$\frac{1}{A_k}D_h(x, z_k) + f(x_k)$	$\frac{1}{2}\sum_{i=0}^{k}\frac{(A_i - A_{i-1})^2}{2}\Vert\nabla f(x_i)\Vert^2$
Accelerated Gradient Method (Weakly Convex)	linear	$\frac{1}{A_k}D_h(x, z_k) + f(x_k)$	0
Accelerated Method Gradient (Strongly Convex)	quadratic	$f(x_k) + \frac{\mu}{2}\Vert x - z_k\Vert^2$	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: How the Estimate Sequence is defined for the various algorithms

where linear is defined as

$$
f_i(x) = f(x_i) + \langle\nabla f(x_i), x - x_i\rangle \qquad (43)
$$

and quadratic is defined as,

$$
f_i(x) = f(x_i) + \langle\nabla f(x_i), x - x_i\rangle + \frac{\mu}{2}\Vert x - x_i\Vert^2. \qquad (44)
$$

6.1 Equivalence between Estimate Sequences and Lyapunov Functions

Now we demonstrate how these two frameworks are equivalent. 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 (38) for accelerated gradient descent can be written:

$$\begin{aligned}
\phi_{k+1}(x) &:= f(x_{k+1}) + \frac{1}{A_{k+1}} D_h(x, z_{k+1}) \\
&\overset{(38)}{=} (1-\tau_k)\phi_k(x) + \tau_k f_k(x) \\
&\overset{\text{Tab } 1}{=} \left(1 - \frac{\alpha_k}{A_{k+1}}\right)\left(f(x_k) + \frac{1}{A_k} D_h(x, z_k)\right) + \frac{\alpha_k}{A_{k+1}} f_k(x)
\end{aligned}$$

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

$$\begin{aligned}
A_{k+1}\left(f(x_{k+1}) + \frac{1}{A_{k+1}} D_h(x, z_{k+1})\right) &= (A_{k+1} - (A_{k+1} - A_k))\left(f(x_k) + \frac{1}{A_k} D_h(x, z_k)\right) + (A_{k+1} - A_k) f_k(x) \\
&= A_k\left(f(x_k) + \frac{1}{A_k} D_h(x, z_k)\right) + (A_{k+1} - A_k) f_k(x) \\
&\overset{(35)}{\leq} A_k f(x_k) + D_h(x, z_k) + (A_{k+1} - A_k) f(x)
\end{aligned}$$

Rearranging, we obtain the inequality $E_{k+1} \leq E_k$ for our Lyapunov function (15). Going the other direction, from our Lyapunov analysis we can derive the following bound:

$$E_k \leq E_0 \tag{45a}$$

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

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

$$A_k \phi_k(x) \leq (A_k - A_0) f(x) + A_0 \phi_0(x) \tag{45b}$$

Rearranging, we obtain our estimate sequence (33) ($A_0 = 1$):

$$\phi_k(x) \leq \left(1 - \frac{A_0}{A_k}\right) f(x) + \frac{A_0}{A_k}\phi_0(x) \tag{46a}$$

$$\leq \left(1 - \frac{1}{A_k}\right) f(x) + \frac{1}{A_k}\phi_0(x). \tag{46b}$$

Equivalence in Continuous time From the derivation of the Lyapunov function (11), we have the following equality:

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

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

$$D_h(x, Z_t) \leq \int_0^t \frac{d}{ds}\left\{e^{\beta_s}\right\}[f(X_s) + \langle \nabla f(X_s), x - X_s\rangle] ds - e^{\beta_t} f(X_t) + e^{\beta_0} f(X_0) + D_h(x, X_0). \tag{47}$$

Now, by pattern matching, we can use this inequality to extract a continuous-time estimate sequence. From (47) we have the inequality,

$$\begin{aligned} e^{\beta_t} f(X_t) + D_h(x, Z_t) &\leq \int_0^t \frac{d}{ds}\left\{e^{\beta_s}\right\}[f(X_s) + \langle \nabla f(X_s), x - X_s\rangle]ds + e^{\beta_0} f(X_0) + D_h(x, Z_0) \\ &\leq \int_0^t \frac{d}{ds}\{e^{\beta_t}\} f(x) + e^{\beta_0} f(X_0) + D_h(x, Z_0) \\ &= (e^{\beta_t} - e^{\beta_0}) f(x) + e^{\beta_0} f(X_0) + D_h(x, Z_0). \end{aligned}$$

Comparing this to (85a), and ignoring the discretization error $\tilde{\varepsilon}_{k+1}$, 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 algorithm simply needed to ensure,

$$0 \leq \int_0^t \frac{d}{ds}\left\{e^{\beta_s}\right\}[f(X_s) + \langle \nabla f(X_s), x - X_s\rangle] - \frac{d}{ds}\left\{e^{\beta_s} f(X_s)\right\} ds \tag{48}$$

for our Lyapunov analysis to go through. Note, we can write (48) as,

$$e^{\beta_t} f(X_t) \leq \int_0^t \frac{d}{ds}\left\{e^{\beta_s}\right\}[f(X_s) + \langle \nabla f(X_s), x - X_s\rangle]ds + e^{\beta_0} f(X_0).$$

From the derivation of the Lypaynov function, we can conclude

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

which shows that $\{f(X_t), e^{\beta_t}\}$ is a continuous-time estimate sequence for this method. In Appendix D, we show the equivalence in discrete time between estimate sequences and our Lyapunov framework in this setting.

6.2 Accelerated Gradient Strong Convexity

Notice that for the dynamic (18) we can show the following,

$$
\begin{aligned}
\frac{d}{dt}\left\{e^{\beta_t}\frac{\mu}{2}\|x-Z_t\|^2\right\} &= \dot\beta_t e^{\beta_t}\frac{\mu}{2}\|x-Z_t\|^2 - e^{\beta_t}\mu\langle \dot Z_t, x-Z_t\rangle\\
&= \dot\beta_t e^{\beta_t}\frac{\mu}{2}\|x-Z_t\|^2 + e^{\beta_t}\mu\langle \dot X_t, x-Z_t\rangle + \dot\beta_t e^{\beta_t}\left\langle \nabla f(X_t), x - X_t - \frac{1}{\dot\beta_t}\dot X_t\right\rangle\\
&= \dot\beta_t e^{\beta_t}\frac{\mu}{2}\|x-Z_t\|^2 + \dot\beta_t e^{\beta_t}\mu\langle Z_t - X_t, x-X_t\rangle - \dot\beta_t e^{\beta_t}\frac{\mu}{2}\|Z_t-X_t\|^2\\
&\quad + \dot\beta_t e^{\beta_t}\langle \nabla f(X_t), x-X_t\rangle - e^{\beta_t}\langle \nabla f(X_t), \dot X_t\rangle\\
&= \dot\beta_t e^{\beta_t}\frac{\mu}{2}\|x-X_t\|^2 + \dot\beta_t e^{\beta_t}\langle \nabla f(X_t), x-X_t\rangle - e^{\beta_t}\langle \nabla f(X_t), \dot X_t\rangle\\
&= \frac{d}{dt}\left\{e^{\beta_t}\right\}\left(\langle \nabla f(X_t), x-X_t\rangle + \frac{\mu}{2}\|x-X_t\|^2\right)\\
&\quad - \left(\frac{d}{dt}\left\{e^{\beta_t}f(X_t)\right\} - \frac{d}{ds}\left\{e^{\beta_t}\right\}f(X_t)\right), \qquad (49)
\end{aligned}
$$

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

$$e^{\beta_t}f(X_t) \le \int_0^t \frac{d}{ds}\left\{e^{\beta_s}\right\}[f(X_s) + \langle \nabla f(X_s), x - X_s\rangle + \frac{\mu}{2}\|x-X_s\|^2]\,ds + e^{\beta_0}f(X_0) + e^{\beta_0}\frac{\mu}{2}\|x-X_0\|^2.$$

From (49), we have

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

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 several method used in optimization. We showed that a single family of Lyapunov functions can be used to verify the convergence rates of a variety 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 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 properities.

[3] We remark also that this derivation also shows how to derive the Lyapunov function (17) from the dynamic (18).

We close with a discussion of some possible directions for future work.

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, but it remains to be seen if other methods can be analyzed. Techniques such as the midpoint method and Runge Kutta provide more accurate solutions of ODEs than Euler methods [5]. 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 [28]. Therefore, many [24, 28, 14] 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 [28] and Krinchene, Bayen and Bartlett [14] 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 (18) might lead to better results.

As mentioned in Section 5, the dynamics (18) 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 (19) can be generalized to a non-Euclidean setting. In follow-up work, we plan to study this second Lagrangian family and its properties further. A natural question is whether the the Lyapunov analysis introduced in this paper can be extended to other settings, such as accelerated coordinate ascent [1] and stochastic gradient descent, and whether a dynamical perspective can be developed in these settings.

Earlier work by Drori and Teboulle [9], Kim and Fessler [13], Taylor *et al* [29], and Lessard *et al* [15] 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

We would like to give special thanks to Andre Wibisono for the many helpful discussion involving this paper. We would also like to thank Orianna Demassi for her very useful feedback and Stephen Tu who caught a small error in a previous version of this paper.

References

[1] Zeyuan Allen-Zhu, Peter Richtárik, Zheng Qu, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML '16, 2016.

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001.

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

[10] 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.

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

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

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

[14] 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.

[15] 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.

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

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

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

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

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

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

[22] 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.

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

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

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

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

[27] 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.

[28] 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.

[29] 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.

[30] Andre Wibisono and Ashia C. Wilson. On Accelerated Methods in Optimization. *ArXiv e-prints arXiv:1509.03616*, September 2015.

[31] 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.

A Mirror Descent Dynamic

We begin with a smooth manifold $\mathcal{X}$ with a local metric $\mathrm{g}(x)$ 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,

$$\dot{X}_t = v(X_t), \tag{50}$$

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

$$v(x) = \arg\min_v \left\{ \langle \nabla f(X_t), v \rangle + \frac{1}{2}\|v\|^2_{\mathrm{g}(X_t)} \right\}. \tag{51}$$

We can write the gradient flow equation explicitly as,

$$\dot{X}_t = -\mathrm{g}(X_t)^{-1}\nabla f(X_t). \tag{52}$$

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:

$$\dot{X}_t = -[\nabla^2 h(X_t)]^{-1}\nabla f(X_t). \tag{53}$$

which we can also write as,

$$\frac{d}{dt}\nabla h(X_t) = -\nabla f(X_t). \tag{54}$$

While on the surface, the ability to rewrite (53) as (54) appears to be a nice algebraic trick, it is actually a manifestation of the following fact:

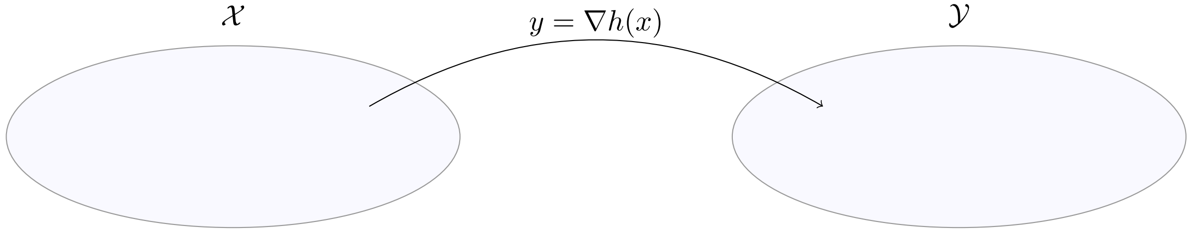
When $h : \mathcal{X} \to \mathbb{R}$ is strictly convex and $\mathrm{g}(x) = \nabla^2 h(x)$, there is a special map $\phi : \mathcal{X} \to \mathcal{Y}$ called the mirror map*, given by $\phi = \nabla h$, whose push-forward maps a gradient flow on $\mathcal{X}$ to a gradient flow on $\mathcal{Y}$. [31]*

As a quick exercise, we can check that the following holds under the mirror map:

"Mirror Map"

$y = \nabla h(x)$

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



Function $f(x)$

Hessian Metric $\nabla^2 h(x)$

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

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

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

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 [31], 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 of time. 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 (54), we obtain the following equation,

$$\frac{d}{dt}\nabla h(X_t) = -\dot{\tau}(t)\nabla f(X_t) = -\frac{d}{dt}\left(e^{\beta_t}\right)\nabla f(X_t). \tag{55}$$

Since the vector field explicitly depends on time, (55) is no longer a gradient flow. Nevertheless, a we can generate a Lyapunov function from the general time-dilated dynamic.

A.1 Lyapunov Analysis

Using the Bregman divergence of h,

$$D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x\rangle, \tag{56}$$

we illustrate how the special property of the Hessian metric provides a Lyapunov analysis of the mirror descent dynamic (54). First, notice that

$$\begin{aligned}
\frac{d}{dt}D_h(x^*, X_t) &= \frac{d}{dt}\left\{h(x^*) - h(X_t) - \langle \nabla h(X_t), x^* - X_t\rangle\right\} ds \\
&= -\langle \nabla h(X_t), \dot{X}_t\rangle - \left\langle \frac{d}{dt}\nabla h(X_t), x^* - X_t\right\rangle + \langle \nabla h(X_t), \dot{X}_t\rangle \\
&= -\left\langle \frac{d}{dt}\nabla h(X_t), x^* - X_t\right\rangle \\
&\overset{(55)}{=} \frac{d}{dt}\left(e^{\beta_t}\right)\langle \nabla f(X_t), x^* - X_t\rangle \\
&\leq \frac{d}{dt}\left(e^{\beta_t}\right)(f(X_t) - f(x^*))dt, \qquad (57a)
\end{aligned}$$

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,

$$\begin{aligned}
f\left(\frac{\int_0^t \dot{\beta}_s e^{\beta_s} X_s ds}{e^{\beta_t} - e^{\beta_0}}\right) - f(x^*) &\leq \frac{-\int_0^t \frac{d}{ds}D_h(x^*, X_s)ds}{e^{\beta_t} - e^{\beta_0}} \\
&\leq \frac{D_h(x^*, X_0) - D_h(x^*, X_t)}{e^{\beta_t} - e^{\beta_0}} \\
&\leq \frac{D_h(x^*, X_0)}{e^{\beta_t} - e^{\beta_0}}.
\end{aligned}$$

However, notice that the primal form (53) of the mirror descent dynamics allows us to obtain a stronger guarantee,

$$\begin{aligned}\frac{d}{dt}\left(e^{\beta_t}f(X_t)\right) &= \frac{d}{dt}\left(e^{\beta_t}\right)f(X_t) + e^{\beta_t}\langle \nabla f(X_t), \dot{X}_t\rangle \\ &= \frac{d}{dt}\left(e^{\beta_t}\right)f(X_t) - \dot{\beta}_t e^{2\beta_t}\langle \nabla f(X_t), \nabla^2 h(X_t)^{-1}\nabla f(X_t)\rangle \\ &= \frac{d}{dt}\left(e^{\beta_t}\right)f(X_t) - \dot{\beta}_t e^{2\beta_t}\frac{1}{2}\|\nabla f(X_t)\|^2_{*,X_t} \\ &\leq \frac{d}{dt}\left(e^{\beta_t}\right)f(X_t). \end{aligned} \tag{58a}$$

If we plug this into (57a), we obtain the following inequality,

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

Integrating gives a Lyapunov function for the mirror descent dynamic (54),

B Momentum Algorithms

B.1 Proof of Theorem 1

We show how to used the Lyapunov function

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

to obtain a convergence guarantee for the quasi-monotone subgradient method (20). Denoting $D^*_{k+1,k} = D_h(x^*, z_{k+1}) - D_h(x^*, z_k)$, observe that

$$\begin{aligned} D^*_{k+1,k} &= -D_h(z_{k+1}, z_k) + \langle \nabla h(z_{k+1}) - \nabla h(z_k), z_{k+1} - x^*\rangle && (59a)\\ &\overset{(20b)}{\leq} -\frac{1}{2}\|z_k - z_{k+1}\|^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_{k+1} - x^*\rangle && (59b)\\ &= -\frac{1}{2}\|z_k - z_{k+1}\|^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_{k+1} - z_k\rangle - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - x^*\rangle \\ &\leq \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2_* - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - x^*\rangle && (59c)\\ &\overset{(20a)}{=} \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2_* - (A_{k+1} - A_k)\left\langle \nabla f(x_{k+1}), \frac{A_{k+1}}{(A_{k+1} - A_k)}(x_{k+1} - x_k) + x_k - x^*\right\rangle \\ &= \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2_* - A_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle \\ &\quad + (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle \\ &= \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2_* - A_k\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle. \\ &\leq -A_k(f(x_{k+1}) - f(x_k)) - (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + \frac{\alpha^2_{k+1}}{2}\|\nabla f(x_{k+1})\|^2_* \\ &= A_k(f(x_k) - f(x^*)) - A_{k+1}(f(x_{k+1}) - f(x^*)) + \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2_* \end{aligned}$$

where (59b) follows from the strong convexity assumption on h, and (59c) uses the Fenchel-Young inequality. Thus, we obtain the following bound,

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

Choosing $\sum_{i=0}^{k}(A_{i+1} - A_i)^2 < \infty$ gives the convergence rate

$$f(x_k) - f(x^*) \leq \frac{D_h(x^*, x_0) + A_0(f(x_0) - f(x^*)) + \frac{1}{2}\sum_{i=0}^{k}(A_{i+1} - A_i)^2\|\nabla f(x_i)\|_*^2}{A_k}. \tag{61}$$

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 (19b) 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,

$$\gamma_{k+1}\nabla h(x_{k+1}) - \gamma_k \nabla h(x_k) = -(A_{k+1} - A_k)\nabla f(x_{k+1}). \tag{62}$$

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

$$E_k = A_k(f(x_k) - f(x^*)) + \gamma_k(D_h(x^*, z_k) - D_h(x^*, x_0)), \tag{63}$$

however first, we analyze the dynamics to understand how one could obtain the Lyapunov function (63) for the dual-averaged algorithm.

The Dynamic We analyze the momentum dynamics (4) with an additional increasing weighting term $\gamma_0 \geq 0$, $\forall t \in \mathbb{R}$ (where $\dot{\gamma}_t \geq 0$):

$$\frac{d}{dt}\Big(\gamma_t \nabla h(Z_t)\Big) = \frac{d}{dt}\Big(e^{\beta_t}\Big)\nabla f(X_t) \tag{64a}$$

$$Z_t = X_t + \frac{1}{\dot{\beta}_t}\dot{X}_t \tag{64b}$$

With this additional weighting term we obtain the following bound,

$$\begin{aligned}
\frac{d}{dt}\left(\gamma_t D_h(x^*, Z_t)\right) dt &= \dot{\gamma}_t[h(x^*) - h(Z_t)] - \frac{d}{dt}\left(\gamma_t\right)\langle \nabla h(Z_t), x - Z_t\rangle - \gamma_t\left\langle \frac{d}{dt}\nabla h(Z_t), x - Z_t\right\rangle \\
&= \dot{\gamma}_t[h(x^*) - h(Z_t)] - \left\langle \frac{d}{dt}\left(\gamma_t \nabla h(Z_t)\right), x^* - Z_t\right\rangle \\
&\overset{(64)}{=} \gamma_t[h(x^*) - h(Z_t)] + \frac{d}{dt}\left(e^{\beta_t}\right)\left\langle \nabla f(X_t), x^* - X_t - \frac{e^{\beta_t}}{\frac{d}{dt}e^{\beta_t}}\dot{X}_t\right\rangle \\
&\overset{(11)}{\leq} \dot{\gamma}_t[h(x^*) - h(Z_t)] - \frac{d}{dt}\left(e^{\beta_t}\left(f(X_t) - f(x^*)\right)\right) \qquad (65a)
\end{aligned}$$

Prox-Center Here, we relax the assumption $\mathcal{X} = \mathbb{R}^d$ and now take it to be some bounded and compact set. We define X_0 to be the "prox-center," of $\mathcal{X}$:

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

and 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, from (65) we obtain the following bound:

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

from which we can conclude that the following function,

$$\mathcal{E}_t = e^{\beta_t}(f(X_t) - f(x^*)) + \gamma_t(D_h(x^*, Z_t) - D_h(x^*, Z_0)), \tag{66}$$

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

$$f(X_t) - f(X_0) \leq \frac{\gamma_t D_h(x^*, Z_0)}{e^{\beta_t}}. \tag{67}$$

The Algorithm Using the following equality,

$$\begin{aligned}\gamma_{k+1}D_h(x^*, z_{k+1}) - \gamma_k D_h(x^*, z_k) = \gamma_{k+1}D_h(z_{k+1}, z_k) + \langle \gamma_{k+1}\nabla h(z_{k+1}) - \gamma_k \nabla h(z_k), z_{k+1} - x^*\rangle \\ + (\gamma_{k+1} - \gamma_k)(h(x^*) - h(z_{k+1})),\end{aligned} \tag{68}$$

we can analyze what happens when we add an additional weighting term γ_t. We define $x_0 \in \mathcal{X}$ to be the *prox-center* over the set that is being optimized over. This amounts to the condition that $h(x) \geq D_h(x, x_0)^2 \geq \frac{\sigma}{2}\|x_0 - x\|^2, \forall x \in \mathcal{X}$, where we used the σ-strong convexity assumption on h in the last inequality. For simplicity, we rescale h so that $\sigma = 1$. That is, we choose $h(x_0) = 0$ and x_0 to be the minimizer of h ($\nabla h(x_0) = 0$). 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 (68), we obtain the bound

$$\begin{aligned}\gamma_{k+1}D_h(x^*, z_{k+1}) - \gamma_k D_h(x^*, z_k) \leq \gamma_k D_h(z_{k+1}, z_k) + \langle \gamma_{k+1}\nabla h(z_{k+1}) - \gamma_k \nabla h(z_k), z_{k+1} - x^*\rangle \\ + (\gamma_{k+1} - \gamma_k)h(x^*).\end{aligned} \tag{69}$$

With this inequality in hand, the proof of convergence is essentially equivalent to the proof of Theorem 1. 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,

$$
\begin{aligned}
D^*_{\gamma_{k+1},\gamma_k} &\overset{(69)}{\leq} \gamma_k D_h(z_{k+1}, z_k) + \langle \gamma_{k+1}\nabla h(z_{k+1}) - \gamma_k \nabla h(z_k), z_{k+1} - x^*\rangle + (\gamma_{k+1} - \gamma_k)h(x^*)\\
&\overset{(62)}{\leq} -\frac{\gamma_k}{2}\|z_k - z_{k+1}\|^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_{k+1} - x^*\rangle + (\gamma_{k+1} - \gamma_k)h(x^*)\\
&= -\frac{\gamma_k}{2}\|z_k - z_{k+1}\|^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_{k+1} - z_k\rangle - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - x^*\rangle\\
&\quad + (\gamma_{k+1} - \gamma_k)h(x^*)\\
&\leq \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - x^*\rangle + (\gamma_{k+1} - \gamma_k)h(x^*)\\
&\overset{(20a)}{=} \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), \frac{A_{k+1}}{(A_{k+1} - A_k)}(x_{k+1} - x_k) + x_k - x^*\rangle\\
&\quad + (\gamma_{k+1} - \gamma_k)h(x^*)\\
&= \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle\\
&\quad + (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle - A_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle + (\gamma_{k+1} - \gamma_k)h(x^*)\\
&= \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2 - (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle - A_k\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle\\
&\quad + (\gamma_{k+1} - \gamma_k)h(x^*)\\
&= -A_k(f(x_{k+1}) - f(x_k)) - (A_{k+1} - A_k)(f(x_{k+1}) - f(x^*)) + \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2\\
&\quad + (\gamma_{k+1} - \gamma_k)h(x^*)\\
&= A_k(f(x_k) - f(x^*)) - A_{k+1}(f(x_{k+1}) - f(x^*)) + \frac{(A_{k+1} - A_k)^2}{2\gamma_k}\|\nabla f(x_{k+1})\|_*^2\\
&\quad + (\gamma_{k+1} - \gamma_k)D_h(x^*, x_0).
\end{aligned}
$$

Therefore, for (63) we can ensure the following

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

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 [23]. Note, that a very similar analysis can be presented for Nesterov's dual averaging algorithm [21].

B.3 Proof of Discretizations

Method 1 We analyze (22) using the Lyapunov function (15).

$$
\begin{aligned}
E_{k+1} - E_k &= A_{k+1}(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_k) - f(x^*)) - \langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1}\rangle \\
&\quad - D_h(z_{k+1}, z_k) \\
&= A_{k+1}(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_k) - f(x^*)) + (A_{k+1} - A_k)\langle \nabla f(x_k), x^* - x_{k+1}\rangle \\
&\quad - A_k\langle \nabla f(x_k), x_{k+1} - x_k\rangle - D_h(z_{k+1}, z_k) \qquad (71a) \\
&= A_{k+1}(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_k) - f(x^*)) + (A_{k+1} - A_k)\langle \nabla f(x_k), x^* - x_k\rangle \\
&\quad + (A_{k+1} - A_k)\langle \nabla f(x_k), x_k - x_{k+1}\rangle - A_k\langle \nabla f(x_k), x_{k+1} - x_k\rangle - D_h(z_{k+1}, z_k) \\
&= A_{k+1}(f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k\rangle) + \alpha_k(f(x_k) - f(x^*) + \langle \nabla f(x_k), x^* - x_k\rangle) \\
&\quad - D_h(z_{k+1}, z_k) \\
&\le A_{k+1}\frac{1}{2}\|x_{k+1} - x_k\|^2 \\
&\overset{(23b)}{\le} \frac{1}{2}\frac{A_{k+1}\alpha_k^2}{A_k^2} diam(\mathcal{X})^2 \qquad (71b)
\end{aligned}
$$

where in (71a) we used (23b) and (23a). Therefore, we obtain the following upper bound,

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

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

Method 2 We analyze (25) using the Lyapunov function (15). Denoting $D^*_{k+1,k} = D_h(x^*, z_{k+1}) - D_h(x^*, z_k)$, observe that

$$
\begin{aligned}
D^*_{k+1,k} &= -\langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1}\rangle - D_h(z_{k+1}, z_k) \\
&\overset{(25b)}{=} (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x^* - z_k\rangle + (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle - D_h(z_{k+1}, z_k) \\
&\overset{(25a)}{=} (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + A_k\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle \\
&\quad - D_h(z_{k+1}, z_k) \\
&\le (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + A_k\langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle + \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2 \\
&\le (A_{k+1} - A_k)(f(x^*) - f(x_{k+1})) + A_k(f(x_k) - f(x_{k+1})) + \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2 \\
&= A_k(f(x_k) - f(x^*)) - A_{k+1}(f(x_{k+1}) - f(x^*)) + \frac{(A_{k+1} - A_k)^2}{2}\|\nabla f(x_{k+1})\|^2.
\end{aligned}
$$

Therefore,

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

and we have the following convergence rate guarantee

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

which is the same convergence rate as the quasi-monotone subgradient method (61). A dual averaging term can also be added to this algorithm as well.

B.4 Proof of Conditional Gradient Method

We use the Lyapunov function (12) to analyze conditional gradient method (15)

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

This recovers the bounds shown by Freund and Grigas [12] , Bach [2] and Nesterov [22].

$$E_k \le E_0 + \frac{1}{2} diam(\mathcal{X})^2 \sum_{i=1}^k \frac{\alpha_i^2}{A_k},$$

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.

Proof of Extension Notice that our variational condition (16b) (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

$$\begin{aligned}
-\alpha_k\langle \nabla f(x_k), x^* - x_k\rangle &\le -A_{k+1}\langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k\rangle + \alpha_k\langle \nabla^2 f(x_k)(x_{k+1} - x_k), x^* - x_k\rangle \\
&\quad + A_k\langle \nabla f(x_k), x_k - x_{k+1}\rangle
\end{aligned} \tag{74}$$

Noting this, we have

$$
\begin{aligned}
E_{k+1} - E_k &= A_{k+1}(f(x_{k+1}) - f(x_k)) - \alpha_k(f(x_k) - f(x^*)) \\
&\le A_{k+1}(f(x_{k+1}) - f(x_k)) - \alpha_k\langle \nabla f(x_k), x^* - x_k\rangle \\
&\overset{(74)}{\le} A_{k+1}(f(x_{k+1}) - f(x_k)) - A_{k+1}\langle \nabla f(x_k), x_{k+1} - x_k\rangle \\
&\quad - A_{k+1}\langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k\rangle + \alpha_k\langle \nabla^2 f(x_k)(x_{k+1} - x_k), x^* - x_k\rangle \\
&\le A_{k+1}(f(x_{k+1}) - f(x_k)) - A_{k+1}\langle \nabla f(x_k), x_{k+1} - x_k\rangle \\
&\quad - A_{k+1}\langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k\rangle + \frac{\alpha_k^2}{A_k}\langle \nabla^2 f(x_k)(z_k - x_k), z_k - x_k\rangle \\
&\le A_{k+1}(f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k\rangle - \frac{1}{2}\langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k\rangle) \\
&\quad + \frac{\alpha_k^2}{A_k} L D^2 \\
&\le A_{k+1}\frac{L}{6}\|x_{k+1} - x_k\|^3 + \frac{\alpha_k^2}{A_k} L D^2 \le \frac{\alpha_k^3}{A_k^2}\frac{L}{6}D^3 + \frac{\alpha_k^2}{A_k} L D^2
\end{aligned}
$$

Recovering the bound shown by Nesterov [22, (5.4)]. The following proof can be extended to incorporate the setting where f has Holder-continuous Hessians as in [22].

B.5 Proof of Accelerated Gradient Methods

The generalized accelerated gradient descent algorithm [3, 31] first outlined by Michel Baes (for the weakly convex setting), can be written as the following iterations:

Algorithm 9 Accelerated Gradient Descent (Weakly Convex Setting)

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

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,

$$
x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k \tag{76a}
$$

$$
z_k = \arg\min_{z \in \mathcal{X}} \left\{ \sum_{i=1}^{k} \alpha_i \langle \nabla f(y_i), z\rangle + D_h(z, z_0) \right\} \tag{76b}
$$

y_k is a gradient update

$$
G_{p,\epsilon,N}(x_k) = \arg\min_y \left\{ f_{p-1}(y; x_k) + \frac{LN}{p}\|y - x_k\|^p \right\} \tag{76c}
$$

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,

$$z_k = y_k + \frac{\tilde{A}_{k+1}}{\tilde{A}_{k+1} - \tilde{A}_k}(x_{k+1} - y_k), \tag{77a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(\tilde{A}_{k+1} - \tilde{A}_k)\nabla f(y_{k+1}) \tag{77b}$$

$$M\|L^{-1}\nabla f(y_k)\|_*^{\frac{p}{p-1}} \le L^{-1}\langle \nabla f(y_k), x_k - y_k\rangle \tag{77c}$$

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

$$\begin{aligned}
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{(77b)}{\le} -\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 && (78a)\\
&= -\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 \\
&\le \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 && (78b)\\
&\overset{(77a)}{=} \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 \\
&\quad - (\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 \\
&\quad - (\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 \\
&\quad - (\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 && (78c)\\
&\quad - \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 \\
&\quad + (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{(77c)}{\le} \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}} \\
&\quad + (\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 \\
&\le -\tilde{A}_k(f(y_{k+1}) - f(y_k)) - (\tilde{A}_{k+1} - \tilde{A}_k)(f(y_{k+1}) - f(x^*)) \\
&\quad + \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^*)) \\
&\quad + \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}},
\end{aligned}$$

where in (78a) we have used the uniform convexity assumption and in (78b) we have used Holder's 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 non-positive mean 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 (4):

$$\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{(28b)}{\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{(28a)}{=} \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{(28c)}{\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 non-positive, A_k can be at most a polynomial of degree 2.

C Strong Convexity

C.1 Proof of Accelerated Gradient (27)

$$
\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+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 \\
&\quad + \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{(27a)}{=} 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 \\
&\quad - \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 \\
&\quad + \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 \\
&\quad + \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{(27b)}{=} -\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 \\
&\quad + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 - \frac{\mu}{2}\|x_{k+1} - y_k\|^2 \\
&\overset{(27a)}{=} -\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 \\
&\quad + \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 \\
&\quad + \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 \\
&\quad + \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.
\end{aligned}
$$

Thus, we have shown

$$
\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k\tilde{E}_k + \varepsilon_k \tag{79}
$$

where

$$
\varepsilon_k = \left(\frac{L\tau_k}{2} - \frac{\mu}{2\tau_k}\right)\|x_{k+1} - y_k\|^2 + \left(\frac{\tau_k^2}{2\mu} - \frac{1}{2L}\right)\|\nabla f(x_{k+1})\|^2 \tag{80}
$$

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

C.2 Proof of Accelerated Gradient (28)

$$
\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{(28a)(28b)}{=} 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{(28a)}{=} -\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{(28b)}{=} -\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

$$\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k \tilde{E}_k \tag{81}$$

D Estimate Sequences

D.1 The Quasi-Montone Subgradient Method

Equivalence in Discrete time The discrete-time estimate sequence (38) for quasi-monotone subgradient method can be written:

$$\begin{aligned}
\phi_{k+1}(x) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} &:= f(x_{k+1}) + \frac{1}{A_{k+1}} D_h(x, z_{k+1}) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} \\
&\overset{(38)}{=} (1-\tau_k)\left(\phi_k(x) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \tau_k f_k(x) \\
&= \left(1 - \frac{\alpha_k}{A_{k+1}}\right)\left(f(x_k) + \frac{1}{A_k} D_h(x, z_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \frac{\alpha_k}{A_{k+1}} f_k(x)
\end{aligned}$$

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

$$\begin{aligned}
A_{k+1}\left(f(x_{k+1}) + \frac{1}{A_{k+1}} D_h(x, z_{k+1}) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}}\right) &= (A_{k+1} - (A_{k+1} - A_k))\left(f(x_k) + \frac{1}{A_k} D_h(x, z_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) \\
&\quad + (A_{k+1} - A_k) f_k(x) \\
&= A_k\left(f(x_k) + \frac{1}{A_k} D_h(x, z_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + (A_{k+1} - A_k) f_k(x) \\
&\overset{(35)}{\leq} A_k f(x_k) + D_h(x, z_k) - \tilde{\varepsilon}_k + (A_{k+1} - A_k) f(x)
\end{aligned}$$

Rearranging, we obtain our Lyapunov argument $E_{k+1} \leq E_k + \varepsilon_{k+1}$ for (15):

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

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:

$$\begin{aligned}
E_k &\leq E_0 + \tilde{\varepsilon}_{k+1} && \text{(82a)} \\
A_k(f(x_k) - f(x)) + D_h(x, z_k) &\leq A_0(f(x_0) - f(x)) + D_h(x, z_0) + \tilde{\varepsilon}_{k+1} \\
A_k\left(f(x_k) - \frac{1}{A_k} D_h(x, z_k)\right) &\leq (A_k - A_0) f(x) + \left(f(x_0) + \frac{1}{A_0} D_h(x^*, z_0)\right) + \tilde{\varepsilon}_{k+1} \\
A_k \phi_k(x) &\leq (A_k - A_0) f(x) + A_0 \phi_0(x) + \tilde{\varepsilon}_{k+1} && \text{(82b)}
\end{aligned}$$

Rearranging, we obtain our estimate sequence (33) ($A_0 = 1$) with an additional error term:

$$\begin{aligned}
\phi_k(x) &\leq \left(1 - \frac{A_0}{A_k}\right) f(x) + \frac{A_0}{A_k} \phi_0(x) + \tilde{\varepsilon}_{k+1} && \text{(83a)} \\
&\leq \left(1 - \frac{1}{A_k}\right) f(x) + \frac{1}{A_k} \phi_0(x) + \tilde{\varepsilon}_{k+1}. && \text{(83b)}
\end{aligned}$$

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 (38) for conditional gradient method can be written:

$$\begin{aligned}\phi_{k+1}(x) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} := f(x_{k+1}) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} &\overset{(38)}{=} (1-\tau_k)\left(\phi_k(x) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \tau_k f_k(x)\\ &\overset{\text{Tab } 1}{=} \left(1 - \frac{\alpha_k}{A_{k+1}}\right)\left(f(x_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \frac{\alpha_k}{A_{k+1}} f_k(x).\end{aligned}$$

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

$$\begin{aligned}A_{k+1}\left(f(x_{k+1}) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}}\right) &= (A_{k+1} - (A_{k+1} - A_k))\left(f(x_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + (A_{k+1} - A_k) f_k(x)\\ &= A_k\left(f(x_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + (A_{k+1} - A_k) f_k(x)\\ &\overset{(35)}{\le} A_k f(x_k) - \tilde{\varepsilon}_k + (A_{k+1} - A_k) f(x).\end{aligned}$$

Rearranging, we obtain our Lyapunov argument $E_{k+1} \le E_k + \varepsilon_{k+1}$ for (12):

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

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:

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

Rearranging, we obtain our estimate sequence (33) ($A_0 = 1$) with an additional error term:

$$\phi_k(x) \le \left(1 - \frac{A_0}{A_k}\right) f(x) + \frac{A_0}{A_k}\phi_0(x) + \tilde{\varepsilon}_{k+1} \tag{85a}$$

$$\le \left(1 - \frac{1}{A_k}\right) f(x) + \frac{1}{A_k}\phi_0(x) + \tilde{\varepsilon}_{k+1}. \tag{85b}$$

D.3 Accelerated Gradient Descent (Strong Convexity)

The discrete-time estimate sequence (38) for accelerated gradient descent can be written:

$$\begin{aligned}\phi_{k+1}(x) &:= f(x_{k+1}) + \frac{\mu}{2}\|x - z_{k+1}\|^2\\ &\overset{(38)}{=} (1-\tau_k)\phi_k(x) + \tau_k f_k(x)\\ &\overset{(35)}{\le} (1-\tau_k)\phi_k(x) + \tau_k f(x).\end{aligned}$$

Therefore, we obtain the inequality $E_{k+1} - E_k \le -\tau_k E_k$ for our Lyapunov function (29):

$$\phi_{k+1}(x) - \phi_k(x) \le -\tau_k(\phi_k(x) - \tau_k f(x))$$
$$f(x_{k+1}) - f(x) + \frac{\mu}{2}\|x - z_{k+1}\|^2 - \left(f(x_k) - f(x) + \frac{\mu}{2}\|x - z_{k+1}\|^2\right) \overset{\text{Tab 1}}{\le} -\tau_k \left(f(x_k) - f(x) + \frac{\mu}{2}\|x - z_{k+1}\|^2\right).$$

Going the other direction, we have,

$$\begin{aligned} E_{k+1} - E_k &\le -\tau_k E_k \\ \phi_{k+1} &\le (1 - \tau_k)\phi_k(x) + \tau_k f(x) \\ A_{k+1}\phi_{k+1} &\le A_k\phi_k + (A_{k+1} - A_k)f(x). \end{aligned}$$

Summing over the righthand side, we obtain the estimate sequence (33):

$$\begin{aligned} \phi_{k+1} &\le \left(1 - \frac{A_0}{A_{k+1}}\right) f(x) + \frac{A_0}{A_{k+1}}\phi_0(x) \\ &= \left(1 - \frac{1}{A_{k+1}}\right) f(x) + \frac{1}{A_{k+1}}\phi_0(x). \end{aligned}$$