A Lyapunov Analysis of Momentum Methods in Optimization

Ashia C. Wilson Benjamin Recht Michael I. Jordan

University of California, Berkeley

March 13, 2018

Abstract

Momentum methods play a significant role in optimization. Examples include Nesterov's accelerated gradient method and the conditional gradient algorithm. Several momentum methods are provably optimal under standard oracle models, and all use a technique called *estimate sequences* to analyze their convergence properties. The technique of estimate sequences has long been considered difficult to understand, leading many researchers to generate alternative, "more intuitive" methods and analyses. 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 connection allows us to develop a simple and unified analysis of many existing momentum algorithms, introduce several 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 studied formally by Polyak [27], and have been employed in many practical optimization solvers. As an example, since the 1980s, momentum methods have been popular in neural networks as a way to accelerate the backpropagation algorithm. The conventional intuition is that momentum allows local search to avoid "long ravines" and "sharp curvatures" in the sublevel sets of cost functions [29].

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 difficult to make rigorous mathematically. 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 [27]. Despite its intuitive elegance, however, Polyak's eigenvalue analysis does not apply globally for general convex cost functions. In fact, Lessard *et al.* derived a simple one-dimensional counterexample where the standard Heavy Ball Method does not converge [15].

¹Indeed, when applied to positive-definite quadratic cost functions, Polyak's Heavy Ball Method is equivalent to Chebyshev's Iterative Method [7].

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]. To achieve such general applicability, Nesterov's proof techniques abandoned the physical intuition of Polyak [18]; in lieu of differential equations and Lyapunov functions, Nesterov devised the method of estimate sequences to verify the correctness of these momentum-based methods. Researchers have struggled to understand the foundations and scope 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 fundamental understanding of the estimate sequence technique, several authors have recently proposed schemes to achieve acceleration without appealing to it [9, 5, 15, 8]. One promising general approach to the analysis of acceleration has been to analyze the continuoustime limit of accelerated methods [30, 13], or to derive these limiting ODEs directly via an underlying Lagrangian [34], and to prove that the ODEs are stable via a Lyapunov function argument. 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. One recent line of attack on the discretization problem is via the use of a time-varying Hamiltonian and symplectic integrators [17]. In this paper, we present a different approach, one based on a fuller development of Lyapunov theory. In particular, we present Lyapunov functions for both the continuous and discrete settings, and we show how to move between these Lyapunov functions. Our Lyapunov functions are time-varying and they thus allow us to establish rates of convergence. They allow us to dispense with estimate sequences altogether, in favor of a dynamical-systems perspective that encompasses both continuous time and discrete time.

2 A Dynamical View of Momentum Methods

Problem setting. We are concerned with the following 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., $\dot{X}_t = \frac{d}{dt}X_t$.

2.1 The Bregman Lagrangian

Wibisono, Wilson and Jordan recently introduced the following function on curves,

$$\mathcal{L}(x,v,t) = e^{\alpha_t + \gamma_t} \left(D_h \left(x, x + e^{-\alpha_t} v \right) - e^{\beta_t} f(x) \right), \tag{2}$$

where $x \in \mathcal{X}$, $v \in \mathbb{R}^d$, and $t \in \mathbb{R}$ represent position, velocity and time, respectively [34]. They called (2) the *Bregman Lagrangian*. The functions $\alpha, \beta, \gamma : \mathbb{R} \to \mathbb{R}$ are arbitrary smooth increasing functions of time that determine the overall damping of the Lagrangian functional, as well as the weighting on the velocity and potential function. They also introduced the following "ideal scaling conditions," which are needed to obtain optimal rates of convergence:

$$\dot{\gamma}_t = e^{\alpha_t} \tag{3a}$$

$$\dot{\beta}_t \le e^{\alpha_t}.$$
 (3b)

Given $\mathcal{L}(x,v,t)$, we can define a functional on curves $\{X_t:t\in\mathbb{R}\}$ called the *action* via integration of the Lagrangian: $\mathcal{A}(X)=\int_{\mathbb{R}}\mathcal{L}(X_t,\dot{X}_t,t)dt$. Calculation of the Euler-Lagrange equation, $\frac{\partial \mathcal{L}}{\partial x}(X_t,\dot{X}_t,t)=\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial v}(X_t,\dot{X}_t,t)$, allows us to obtain a stationary point for the problem of finding the curve which minimizes the action. Wibisono, Wilson, and Jordan showed [34, (2.7)] that under the first scaling condition (3a), the Euler-Lagrange equation for the Bregman Lagrangian reduces to the following ODE:

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

Second Bregman Lagrangian. We introduce a second function on curves,

$$\mathcal{L}(x,v,t) = e^{\alpha_t + \gamma_t + \beta_t} \left(\mu D_h \left(x, x + e^{-\alpha_t} v \right) - f(x) \right), \tag{5}$$

using the same definitions and scaling conditions. The Lagrangian (5) places a different damping on the kinetic energy than in the original Bregman Lagrangian (2).

Proposition 1. Under the same scaling condition (3a), the Euler-Lagrange equation for the second Bregman Lagrangian (5) reduces to:

$$\frac{d}{dt}\nabla h(X_t + e^{-\alpha_t}\dot{X}_t) = \dot{\beta}_t\nabla h(X_t) - \dot{\beta}_t\nabla h(X_t + e^{-\alpha_t}\dot{X}_t) - \frac{e^{\alpha_t}}{\mu}\nabla f(X_t). \tag{6}$$

We provide a proof of Proposition 1 in Appendix A.1. In what follows, we pay close attention to the special case of the dynamics in (6) where h is Euclidean and the damping $\beta_t = \gamma t$ is linear:

$$\ddot{X}_t + 2\gamma \dot{X}_t + \frac{\gamma^2}{\mu} \nabla f(X_t) = 0.$$
 (7)

When $\gamma = \sqrt{\mu}$, we can discretize the dynamics in (7) to obtain accelerated gradient descent in the setting where f is μ -strongly convex.

2.2 Lyapunov function for the Euler-Lagrange equation

To establish a convergence rate associated with solutions to the Euler-Lagrange equation for both families of dynamics (4) and (6), under the ideal scaling conditions, we use Lyapunov's method [16]. Lyapunov's method 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 $\dot{X}_t = v(X_t)$:

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

The existence of such a Lyapunov function guarantees that the dynamical system converges: if the function is positive yet strictly decreasing along all trajectories, then the dynamical system must eventually approach a region where $\mathcal{E}(X)$ is minimal. If this region coincides with the stationary points of the dynamics, then all trajectories must converge to a stationary point. We now discuss the derivation of time-dependent Lyapunov functions for dynamical systems with bounded level sets. The Lyapunov functions will imply convergence rates for dynamics (2) and (6).

Proposition 2. Assume f is convex, h is strictly convex, and the second ideal scaling condition (3b) holds. The Euler-Lagrange equation (4) satisfies

$$\frac{d}{dt} \Big\{ D_h(x, X_t + e^{-\alpha_t} \dot{X}_t) \Big\} \le -\frac{d}{dt} \Big\{ e^{\beta_t} (f(X_t) - f(x)) \Big\},\tag{8}$$

when $x = x^*$. If the ideal scaling holds with equality, $\dot{\beta}_t = e^{\alpha_t}$, the solutions satisfy (8) for $\forall x \in \mathcal{X}$. Thus,

$$\mathcal{E}_t = D_h(x, X_t + e^{-\alpha_t} \dot{X}_t) + e^{\beta_t} (f(X_t) - f(x))$$
(9)

is a Lyapunov function for dynamics (4).

A similar proposition holds for the second family of dynamics (5) under the additional assumption that f is μ -uniformly convex with respect to h:

$$D_f(x,y) > \mu D_h(x,y). \tag{10}$$

When $h(x) = \frac{1}{2}||x||^2$ is the Euclidean distance, (10) is equivalent to the standard assumption that f is μ -strongly convex. Another special family is obtained when $h(x) = \frac{1}{p}||x||^p$, which, as pointed out by Nesterov [20, Lemma 4], yields a Bregman divergence that is σ -uniformly convex with respect to the p-th power of the norm:

$$D_h(x,y) \ge \frac{\sigma}{p} ||x - y||^p, \tag{11}$$

where $\sigma = 2^{-p+2}$. Therefore, if f is uniformly convex with respect to the Bregman divergence generated by the p-th power of the norm, it is also uniformly convex with respect to the p-th power of the norm itself. We are now ready to state the main proposition for the continuous-time dynamics.

Proposition 3. Assume f is μ -uniformly convex with respect to h (10), h is strictly convex, and the second ideal scaling condition (3b) holds. Using dynamics (6), we have the following inequality:

$$\frac{d}{dt} \left\{ e^{\beta_t} \mu D_h(x, X_t + e^{-\alpha_t} \dot{X}_t) \right\} \le -\frac{d}{dt} \left\{ e^{\beta_t} (f(X_t) - f(x)) \right\},$$

for $x = x^*$. If the ideal scaling holds with equality, $\dot{\beta}_t = e^{\alpha_t}$, the inequality holds for $\forall x \in \mathcal{X}$. In sum, we can conclude that

$$\mathcal{E}_t = e^{\beta_t} \left(\mu D_h(x, X_t + e^{-\alpha_t} \dot{X}_t) + f(X_t) - f(x) \right)$$
(12)

is a Lyapunov function for dynamics (6).

The proof of both results, which can be found in Appendix A.2, uses the fundamental theorem of calculus and basic properties of dynamics (6). Taking $x = x^*$ and writing the Lyapunov property $\mathcal{E}_t \leq \mathcal{E}_0$ explicitly,

$$f(X_t) - f(x^*) \le \frac{D_h(x^*, X_0 + e^{-\alpha_0} \dot{X}_0) + e^{\beta_0} (f(X_0) - f(x^*))}{e^{\beta_t}}$$
(13)

for (9), and

$$f(X_t) - f(x^*) \le \frac{e^{\beta_0} (\mu D_h(x^*, X_0 + e^{-\alpha_0} \dot{X}_0) + f(X_0) - f(x^*))}{e^{\beta_t}}, \tag{14}$$

for (12), allows us to infer a $O(e^{-\beta t})$ convergence rate for the function value for both families of dynamics (4) and (6).

So far, we have introduced two families of dynamics (4) and (6) and illustrated how to derive Lyapunov functions for these dynamics which certify a convergence rate to the minimum of an objective function f under suitable smoothness conditions on f and h. Next, we will discuss how various discretizations of dynamics (4) and (6) produce algorithms which are useful for convex optimization. A similar discretization of the Lyapunov functions (9) and (12) will provide us with tools we can use to analyze these algorithms. We defer discussion of additional mathematical properties of the dynamics that we introduce—such as existence and uniqueness—to Appendix C.4.

3 Discretization Analysis

In this section, we illustrate how to map from continuous-time dynamics to discrete-time sequences. We assume throughout this section that the second ideal scaling (3b) holds with equality, $\dot{\beta}_t = e^{\alpha_t}$.

Explicit and implicit methods. Consider a general vector field $\dot{X}_t = v(X_t)$, where $v : \mathbb{R}^n \to \mathbb{R}^n$ is smooth. The explicit Euler method evaluates the vector field at the current point to determine a discrete-time step

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

The implicit Euler method, on the other hand, evaluates the vector field at the future point

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

An advantage of the explicit Euler method is that it is easier to implement in practice. The implicit Euler method has greater stability and convergence properties but requires solving an expensive implicit equation. We evaluate what happens when we apply these discretization techniques to both families of dynamics (4) and (6). To do so, we write these dynamics as systems of first-order equations. The implicit and explicit Euler method can be combined in four separate ways to obtain algorithms we can analyze; for both families, we provide results on several combinations of the explicit and implicit methods, focusing on the family that gives rise to accelerated methods.

3.1 Methods arising from the first Euler-Lagrange equation

We apply the implicit and explicit Euler schemes to dynamics (4), written as the following system of first-order equations:

$$Z_t = X_t + \frac{e^{\beta_t}}{\frac{d}{dt}e^{\beta_t}}\dot{X}_t, \tag{15a}$$

$$\frac{d}{dt}\nabla h(Z_t) = -\left(\frac{d}{dt}e^{\beta_t}\right)\nabla f(X_t). \tag{15b}$$

Wibisono, Wilson and Jordan showed that the polynomial family $\beta_t = p \log t$ is the continuous-time limit of a family of accelerated disrete-time methods [34], Here, we consider any parameter β_t whose time derivative $\frac{d}{dt}e^{\beta_t} = (A_{k+1} - A_k)/\delta$ can be well-approximated by a discrete-time sequence $(A_i)_{i=1}^k$. The advantage of choosing an arbitrary time scaling δ is that it leads to a broad family of algorithms. To illustrate this, make the approximations $Z_t = z_k$, $X_t = x_k$, $\frac{d}{dt}\nabla h(Z_t) = \frac{\nabla h(z_{k+1}) - \nabla h(z_k)}{\delta}$, $\dot{X}_t = \frac{d}{dt}X_t = \frac{x_{k+1} - x_k}{\delta}$, and denote $\tau_k = \frac{A_{k+1} - A_k}{A_k} := \frac{\alpha_k}{A_k}$, so that $\frac{e^{\beta_t}}{\frac{d}{dt}e^{\beta_t}} = \delta/\tau_k$. With these approximations, we explore various combinations of the explicit and implicit discretizations.

Implicit-Implicit-Euler. Written as an algorithm, the implicit Euler method applied to (15a) and (15b) has the following update equations:

$$z_{k+1} = \underset{\substack{z \in \mathcal{X} \\ x = \frac{\tau_k}{1 + \tau_k} z + \frac{1}{1 + \tau_k} x_k}}{\arg \min} \left\{ A_k f(x) + \frac{1}{\tau_k} D_h(z, z_k) \right\},$$
(16a)

$$x_{k+1} = \frac{\tau_k}{1 + \tau_k} z_{k+1} + \frac{1}{1 + \tau_k} x_k. \tag{16b}$$

We now state our main proposition for the discrete-time dynamics.

Proposition 4. Using the discrete-time Lyapunov function,

$$E_k = D_h(x^*, z_k) + A_k(f(x_k) - f(x^*)), \tag{17}$$

the bound $\frac{E_{k+1}-E_k}{\delta} \leq 0$ holds for algorithm (16).

In particular, this allows us to conclude a general $O(1/A_k)$ convergence rate for the implicit method (16).

Proof. The implicit scheme (16), with the aforementioned discrete-time approximations, satisfies the following variational inequalities:

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -(A_{k+1} - A_k) \nabla f(x_{k+1})$$
(18a)

$$(A_{k+1} - A_k)z_{k+1} = (A_{k+1} - A_k)x_{k+1} + A_k(x_{k+1} - x_k).$$
(18b)

Using these identities, we have the following derivation:

$$\begin{split} E_{k+1} - E_k &= D_h(x, z_{k+1}) - D_h(x, z_k) + A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) \\ &= -\langle \nabla h(z_{k+1}) - \nabla h(z_k), x - z_{k+1} \rangle - D_h(z_{k+1}, z_k) \\ &+ A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) \\ &\stackrel{(18a)}{=} (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x - z_{k+1} \rangle - D_h(z_{k+1}, z_k) \\ &+ A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x)) \\ &\stackrel{(18b)}{=} (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), x - x_{k+1} \rangle + A_k\langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \\ &- D_h(z_{k+1}, z_k) + A_k(f(x_{k+1}) - f(x_k)) + (A_{k+1} - A_k)(f(x_{k+1}) - f(x)) \\ &\leq 0. \end{split}$$

The inequality on the last line follows from the convexity of f and the strict convexity of h.

Accelerated gradient family. We study families of algorithms which give rise to a family of accelerated methods. These methods can be thought of variations of the explicit Euler scheme applied to (15a) and the implicit Euler scheme applied to (15b).² The first family of methods can be written as the following general sequence:

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

$$\nabla h(z_{k+1}) = \nabla h(z_k) - \alpha_k \nabla f(x_{k+1}) \tag{19b}$$

$$y_{k+1} = \mathcal{G}(x), \tag{19c}$$

where $\mathcal{G}: \mathcal{X} \to \mathcal{X}$ is an arbitrary map whose domain is the previous state, $x = (x_{k+1}, z_{k+1}, y_k)$. The second family can be written:

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

$$y_{k+1} = \mathcal{G}(x) \tag{20b}$$

$$\nabla h(z_{k+1}) = \nabla h(z_k) - \alpha_k \nabla f(y_{k+1}), \tag{20c}$$

where $\mathcal{G}: \mathcal{X} \to \mathcal{X}$ is an arbitrary map whose domain is the previous state, $x = (x_{k+1}, z_k, y_k)$. When $\mathcal{G}(x) = x_{k+1}$ for either algorithm, we recover a classical explicit discretization applied to (15a) and implicit discretization applied to (15b). We will show that the additional sequence y_k allows us to obtain better error bounds in our Lyapunov analysis. Indeed, we will show that accelerated gradient descent [18, 19], accelerated higher-order methods [20, 3], accelerated universal methods [11], accelerated proximal methods [32, 4, 21] all involve particular choices for the map \mathcal{G} and for the smoothness assumptions on f and h. Furthermore, we demonstrate how the analyses contained in all of these papers implicitly show the following discrete-time Lyapunov function,

$$E_k = D_h(x^*, z_k) + A_k(f(y_k) - f(x^*)), \tag{21}$$

is decreasing for each iteration k. To show this, we begin with the following proposition.

²Here we make the identification $\tau_k = A_{k+1} - A_k/A_{k+1} := \alpha_k/A_{k+1}$.

Proposition 5. Assume that the distance-generating function h is σ -uniformly convex with respect to the p-th power of the norm $(p \ge 2)$ (11) and the objective function f is convex. Using only the updates (19a) and (19b), and using the Lyapunov function (21), we have the following bound:

$$\frac{E_{k+1} - E_k}{\delta} \le \varepsilon_{k+1},\tag{22}$$

where the error term scales as

$$\varepsilon_{k+1} = \frac{p-1}{p} \sigma^{-\frac{1}{p-1}} \frac{(A_{k+1} - A_k)^{\frac{p}{p-1}}}{\delta} \|\nabla f(x_{k+1})\|^{\frac{p}{p-1}} + \frac{A_{k+1}}{\delta} (f(y_{k+1}) - f(x_{k+1})). \tag{23a}$$

If we use the updates (20a) and (20c) instead, the error term scales as

$$\varepsilon_{k+1} = \frac{p-1}{p} \sigma^{-\frac{1}{p-1}} \frac{(A_{k+1} - A_k)^{\frac{p}{p-1}}}{\delta} \|\nabla f(y_{k+1})\|^{\frac{p}{p-1}} + \frac{A_{k+1}}{\delta} \langle \nabla f(y_{k+1}), y_{k+1} - x_{k+1} \rangle. \tag{23b}$$

The error bounds in (23) were obtained using no smoothness assumption on f and h; they also hold when full gradients of f are replaced with elements in the subgradient of f. The proof of this proposition can be found in Appendix B.1. The bounds in Proposition 5 were obtained without using the arbitrary update $y_{k+1} = \mathcal{G}(x)$. In particular, accelerated methods are obtained by picking a map \mathcal{G} that results in a better bound on the error than the straightforward discretization $y_{k+1} = x_{k+1}$. We immediately see that any algorithm for which the map \mathcal{G} satisfies the progress condition $f(y_{k+1}) - f(x_{k+1}) \propto -\|\nabla f(x_{k+1})\|^{\frac{p}{p-1}}$ or $\langle \nabla f(y_{k+1}), y_{k+1} - x_{k+1} \rangle \propto -\|\nabla f(y_{k+1})\|^{\frac{p}{p-1}}$ will have a $O(1/\epsilon \sigma k^p)$ convergence rate. We now show how this general analysis applied concretely to each of the aforementioned five methods.

Quasi-monotone method [24]. The quasi-monotone subgradient method, which uses the map

$$\mathcal{G}(x) = x_{k+1}$$

for both algorithms (19) and (20), was introduced by Nesterov in 2015. Under this map, assuming the strong convexity of h (which implies p = 2), we can write the error (23) as

$$\varepsilon_{k+1} = \frac{(A_{k+1} - A_k)^2}{2\sigma\delta} \|\nabla f(x_{k+1})\|^2.$$
 (24)

If we assume all the (sub)gradients of f are upper bounded in norm, then maximizing $\sum_{i=1}^k \varepsilon_i/A_i$ results in an $O(1/\sqrt{k})$ convergence rate. This matches the lower bound for (sub)gradient methods designed for Lipschitz-convex functions.³

Accelerated gradient/mirror descent [18, 19]. In 1983, Nesterov introduced accelerated gradient decent, which uses the following family of operators $\mathcal{G} \equiv \mathcal{G}_{\epsilon}$, parameterized by a scaling constant $\epsilon > 0$:

$$\mathcal{G}_{\epsilon}(x) = \arg\min_{y \in \mathcal{X}} \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\epsilon} ||y - x||^2 \right\}.$$
 (25)

Nesterov assumed the use of full gradients ∇f which are $(1/\epsilon)$ -smooth; thus, the gradient map is scaled according to the Lipschitz parameter.

³The same convergence bound can be shown to hold for the (sub)gradient method under this smoothness class, when one assesses convergence for the average/minimum iterate [18].

Lemma 6. Assume h is σ -strongly convex and f is $(1/\epsilon)$ -smooth. Using the gradient update, $y_{k+1} = \mathcal{G}_{\epsilon}(x_{k+1})$, for updates (19c) and (20b), where \mathcal{G}_{ϵ} is defined in (25), the error for algorithm (19) can be written as follows:

$$\varepsilon_{k+1} = \frac{(A_{k+1} - A_k)^2}{2\sigma\delta} \|\nabla f(x_{k+1})\|^2 - \frac{\epsilon A_{k+1}}{2\delta} \|\nabla f(x_{k+1})\|^2, \tag{26a}$$

and for algorithm (20), we have:

$$\varepsilon_{k+1} = \frac{(A_{k+1} - A_k)^2}{2\sigma\delta} \|\nabla f(y_{k+1})\|^2 - \frac{\epsilon A_{k+1}}{2\delta} \|\nabla f(y_{k+1})\|^2.$$
 (26b)

Proof. The optimality condition for the gradient update (25) is

$$\nabla f(x) = \frac{1}{\epsilon} (x - \mathcal{G}_{\epsilon}(x)). \tag{27}$$

The bound (26a) follows from smoothness of the objective function f,

$$f(\mathcal{G}_{\epsilon}(x)) \leq f(x) + \langle \nabla f(x), \mathcal{G}_{\epsilon}(x) - x \rangle + \frac{1}{2\epsilon} \|\mathcal{G}_{\epsilon}(x) - x\|^{2}$$

$$\stackrel{(27)}{=} f(x) - \frac{\epsilon}{2} \|\nabla f(x)\|^{2}.$$

For the second bound (26b), we use the $(1/\epsilon)$ -smoothness of the gradient,

$$\|\nabla f(\mathcal{G}_{\epsilon}(x)) - \nabla f(x)\| \le \frac{1}{\epsilon} \|\mathcal{G}_{\epsilon}(x) - x\|; \tag{28}$$

substituting (27) into (28), squaring both sides, and expanding the square on the left-hand side, yields the desired bound:

$$\langle \nabla f(\mathcal{G}_{\epsilon}(x)), x - \mathcal{G}_{\epsilon}(x) \rangle \leq -\frac{\epsilon}{2} \|\nabla f(\mathcal{G}_{\epsilon}(x))\|^{2}.$$

The error bounds we have just obtained depend explicitly on the scaling ϵ . This restricts our choice of sequences A_k ; they must satisfy the following inequality:

$$\frac{(A_{k+1} - A_k)^2}{A_{k+1}} \le \epsilon \sigma,\tag{29}$$

for the error to be bounded. Choosing A_k to be a polynomial in k of degree two, with leading coefficients $\epsilon \sigma$, optimizes the bound (29); from this we can conclude $f(y_k) - f(x^*) \leq O(1/\epsilon \sigma k^2)$, which matches the lower bound for algorithms which only use full gradients of the objective function. Furthermore, if we take the discretization step to scale according to the smoothness as $\delta = \sqrt{\epsilon}$, then both $||x_k - y_k|| = O(\sqrt{\epsilon})$ and $\varepsilon_k = O(\sqrt{\epsilon})$; therefore, as $\sqrt{\epsilon} \to 0$, we recover the dynamics (15) and the statement $\dot{\mathcal{E}}_t \leq 0$ for Lyapunov function (4) in the limit.

Accelerated universal methods [20, 3, 22, 11]. The term "universal methods" refers to the algorithms designed for the class of functions with (ϵ, ν) -Hölder-continuous higher-order gradients $(2 \le p \in \mathbb{N}, \nu \in (0, 1], \epsilon > 0)$,

$$\|\nabla^{p-1}f(x) - \nabla^{p-1}f(y)\| \le \frac{1}{\epsilon} \|x - y\|^{\nu}.$$
 (30)

Typically, practitioners care about the setting where we have Hölder-continuous gradients (p=2) or Hölder-continuous Hessians (p=3), since methods which use higher-order information are often too computationally expensive. In the case $p \geq 3$, the gradient update

$$\mathcal{G}_{\epsilon,p,\nu,N}(x) = \arg\min_{y \in \mathcal{X}} \left\{ f_{p-1}(x;y) + \frac{N}{\epsilon \tilde{p}} \|x - y\|^{\tilde{p}} \right\}, \quad \tilde{p} = p - 1 + \nu, N > 1$$
(31)

can be used to simplify the error (23b) obtained by algorithm (20). Notice, the gradient update is regularized by the smoothness parameter \tilde{p} . We summarize this result in the following proposition.

Lemma 7. Assume f has Hölder-continuous higher-order gradients. Using the map $y_{k+1} = \mathcal{G}_{\epsilon,p,\nu,N}(x_{k+1})$, defined by (31), in update (20b) yields the following progress condition:

$$\langle \nabla f(y_{k+1}), y_{k+1} - x_{k+1} \rangle \le -\frac{(N^2 - 1)^{\frac{\tilde{p} - 1}{2\tilde{p} - 2}}}{2N} \epsilon^{\frac{1}{\tilde{p} - 1}} \|\nabla f(y_{k+1})\|^{\frac{\tilde{p}}{\tilde{p} - 1}}, \tag{32}$$

where $\tilde{p} = p - 1 + \nu$ and $p \geq 3$.

Lemma 7 demonstrates that if the Taylor approximation is regularized according to the smoothness of the function, the progress condition scales as a function of the smoothness in a particularly nice way. Using this inequality, we can simplify the error (23b) in algorithm (20) to the following,

$$\varepsilon_{k+1} = \frac{\tilde{p} - 1}{\tilde{p}} \sigma^{-\frac{1}{\tilde{p} - 1}} \frac{(A_{k+1} - A_k)^{\frac{\tilde{p}}{\tilde{p} - 1}}}{\delta} \|\nabla f(y_{k+1})\|^{\frac{\tilde{p}}{\tilde{p} - 1}} - \frac{A_{k+1}}{\delta} \frac{(N^2 - 1)^{\frac{\tilde{p} - 1}{2\tilde{p} - 2}}}{2N} \epsilon^{\frac{1}{\tilde{p} - 1}} \|\nabla f(y_{k+1})\|^{\frac{\tilde{p}}{\tilde{p} - 1}},$$

where we have assumed that the geometry scales nicely with the smoothness condition: $D_h(x,y) \ge \frac{\sigma}{\tilde{p}} ||x-y||^{\tilde{p}}$. This requires the condition $p \ge 3$. To ensure a non-positive error we choose a sequence which satisfies the bound,

$$\frac{(A_{k+1} - A_k)^{\frac{\tilde{p}}{\tilde{p}-1}}}{A_{k+1}} \le (\epsilon \sigma)^{\frac{1}{\tilde{p}-1}} \frac{\tilde{p}}{\tilde{p}-1} \frac{(N^2 - 1)^{\frac{\tilde{p}-1}{2\tilde{p}-2}}}{2N} := C_{\epsilon,\sigma,\tilde{p},N}.$$

This bound is maximized by polynomials in k of degree \tilde{p} with leading coefficient proportional to $C_{\epsilon,\sigma,\tilde{p},N}^{\tilde{p}-1}$; this results in the convergence rate bound $f(y_k) - f(x^*) \leq O(1/\epsilon\sigma k^{\tilde{p}}) = O(1/\epsilon\sigma k^{p-1+\nu})$. We can compare this convergence rate to that obtained by using just the gradient map $y_{k+1} = \mathcal{G}_{\epsilon,p,\tilde{p},N}(y_k)$; this algorithm yields a slower $f(y_k) - f(x^*) \leq O(1/\epsilon\sigma k^{\tilde{p}-1}) = O(1/\epsilon\sigma k^{p-2+\nu})$ convergence rate under the same smoothness assumptions. Proofs of these statements can be found in Appendix B.2. This result unifies and extends the analyses of the accelerated (universal) cubic regularized Newton's method [20, 11] and accelerated higher-order methods [3]. Wibisono et al. [34]

show that $||x_k - y_k|| = O(\epsilon^{1/\tilde{p}})$ and $\varepsilon_k = O(\epsilon^{1/\tilde{p}})$ so that as $\epsilon^{1/\tilde{p}} \to 0$ we recover the dynamics (15) and the statement $\dot{\mathcal{E}}_t \leq 0$ for Lyapunov function (4).

We end by mentioning that in the special case p = 2, Nesterov [22] showed that a slightly modified gradient map,

$$\mathcal{G}_{\tilde{\epsilon}}(x) = x - \tilde{\epsilon} \,\nabla f(x),\tag{33}$$

has the following property when applied to functions with Hölder-continuous gradients.

Lemma 8. ([22, Lemma 1]) Assume f has (ϵ, ν) -Hölder-continuous gradients, where $\nu \in (0, 1]$. Then for $1/\tilde{\epsilon} \geq (1/2\tilde{\delta})^{\frac{1-\nu}{1+\nu}} (1/\epsilon)^{\frac{2}{1+\nu}}$ the following bound:

$$f(y_{k+1}) - f(x_{k+1}) \le -\frac{\tilde{\epsilon}}{2} \|\nabla f(x_{k+1})\|^2 + \tilde{\delta},$$

holds for $y_{k+1} = \mathcal{G}_{\tilde{\epsilon}}(x_{k+1})$ given by (33).

That is, if we take a gradient descent step with increased regularization and assume h is σ strongly convex, the error for algorithm (19) when f is (ϵ, ν) -Hölder-continuous can be written as,

$$\varepsilon_{k+1} = \frac{(A_{k+1} - A_k)^2}{2\sigma\delta} \|\nabla f(x_{k+1})\|^2 - \frac{\tilde{\epsilon}A_{k+1}}{2\delta} \|\nabla f(x_{k+1})\|^2 + \tilde{\delta}.$$
(34)

This allows us to infer a $O(1/\tilde{\epsilon}\sigma k^2)$ convergence rate of the function to within $\tilde{\delta}$, which is controlled by the amount of regularization $\tilde{\epsilon}$ we apply in the gradient update. Having discussed algorithms "derived" from dynamics (4), we next discuss algorithms arising from the second family of dynamics (6) and a proximal variant of it. The derivations and analyses will be remarkably similar to those presented in this section.

3.2 Methods arising from the second Euler-Lagrange equation

We apply the implicit and explicit Euler schemes to the dynamics (6) written as the following system of equations:

$$Z_t = X_t + \frac{e^{\beta_t}}{\frac{d}{dt}e^{\beta_t}}\dot{X}_t, \tag{35a}$$

$$\frac{d}{dt}\nabla h(Z_t) = \frac{\frac{d}{dt}e^{\beta_t}}{e^{\beta_t}} \left(\nabla h(X_t) - \nabla h(Z_t) - \frac{1}{\mu}\nabla f(X_t)\right),\tag{35b}$$

As in the previous setting, we consider any parameter β_t whose time derivative $\frac{d}{dt}e^{\beta_t} = (A_{k+1} - A_k)/\delta$ can be well-approximated by a discrete-time sequence $(A_i)_{i=1}^k$. In addition, we make the discrete-time approximations $\frac{d}{dt}\nabla h(Z_t) = \frac{\nabla h(z_{k+1}) - \nabla h(z_k)}{\delta}$ and $\frac{d}{dt}\dot{X}_t = \frac{x_{k+1} - x_k}{\delta}$, and denote $\tau_k = \frac{A_{k+1} - A_k}{A_k}$. We have the following proposition.

Proposition 9. Written as an algorithm, the implicit Euler scheme applied to (35a) and (35b) results in the following updates:

$$z_{k+1} = \underset{x = \frac{\tau_k}{1 + \tau_k} z + \frac{1}{1 + \tau_k} x_k}{arg \min} \left\{ f(x) + \mu D_h(z, x) + \frac{\mu}{\tau_k} D_h(z, z_k) \right\},$$
(36a)

$$x_{k+1} = \frac{\tau_k}{1 + \tau_k} z_{k+1} + \frac{1}{1 + \tau_k} x_k. \tag{36b}$$

Using the following discrete-time Lyapunov function:

$$E_k = A_k(\mu D_h(x^*, z_k) + f(x_k) - f(x^*)), \tag{37}$$

we obtain the bound $E_{k+1} - E_k \leq 0$ for algorithm (16). This allows us to conclude a general $O(1/A_k)$ convergence rate for the implicit scheme (16).

Proof. The algorithm that follows from the implicit discretization of the dynamics (36) satisfies the variational conditions

$$\nabla h(z_{k+1}) - \nabla h(z_k) = \tau_k \left(\nabla h(x_{k+1}) - \nabla h(z_{k+1}) - \frac{1}{\mu} \nabla f(x_{k+1}) \right)$$
(38a)

$$(x_{k+1} - x_k) = \tau_k(z_{k+1} - x_{k+1}), \tag{38b}$$

where $\tau_k = \frac{\alpha_k}{A_k}$. Using these variational inequalities, we have the following argument:

$$E_{k+1} - E_k = \alpha_k \mu D_h(x, z_{k+1}) + A_k \mu D_h(x, z_{k+1}) - A_k \mu D_h(x, z_k)$$

$$+ A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x))$$

$$= \alpha_k \mu D_h(x, z_{k+1}) - A_k \mu \langle \nabla h(z_{k+1}) - \nabla h(z_k), x - z_{k+1} \rangle - \mu A_k D_h(z_{k+1}, z_k)$$

$$+ A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x))$$

$$\stackrel{(38a)}{=} \alpha_k \mu D_h(x, z_{k+1}) + A_k \tau_k \langle \nabla f(x_{k+1}), x - z_{k+1} \rangle - A_k \mu D_h(z_{k+1}, z_k))$$

$$+ A_k \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + A_k \tau_k \mu \langle \nabla h(x_{k+1}) - \nabla h(z_{k+1}), x - z_{k+1} \rangle$$

$$+ A_{k+1}(f(x_{k+1}) - f(x)) - A_k(f(x_k) - f(x))$$

$$\stackrel{(38b)}{=} \alpha_k \mu D_h(x, z_{k+1}) + \alpha_k \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle - A_k \mu D_h(z_{k+1}, z_k))$$

$$+ A_k \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \alpha_k \mu \langle \nabla h(x_{k+1}) - \nabla h(z_{k+1}), x - z_{k+1} \rangle$$

$$+ A_k(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_{k+1}) - f(x))$$

$$\leq -\alpha_k \mu D_h(x_{k+1}, z_{k+1}) - A_k \mu D_h(z_{k+1}, z_k)$$

The inequality uses the Bregman three-point identity (60) and μ -uniform convexity of f with respect to h (10).

We now focus on analyzing the accelerated gradient family, which can be viewed as a discretization that contains easier subproblems.

3.2.1 Accelerated gradient descent [18]

We study a family of algorithms which can be thought of as slight variations of the implicit Euler scheme applied to (35a) and the explicit Euler scheme applied to (35b)

$$x_k = \frac{\tau_k}{1 + \tau_k} z_k + \frac{1}{1 + \tau_k} y_k \tag{39a}$$

$$\nabla h(z_{k+1}) - \nabla h(z_k) = \tau_k \left(\nabla h(x_k) - \nabla h(z_k) - \frac{1}{\mu} \nabla f(x_k) \right)$$
(39b)

$$y_{k+1} = \mathcal{G}(x), \tag{39c}$$

where $x = (x_k, z_{k+1}, y_k)$ is the previous state and $\tau_k = \frac{A_{k+1} - A_k}{A_{k+1}}$. Note that when $\mathcal{G}(x) = x_k$, we recover classical discretizations. The additional sequence $y_{k+1} = \mathcal{G}(x)$, however, allows us to obtain better error bounds using the Lyapunov analysis. To analyze the general algorithm (39), we use the following Lyapunov function:

$$E_k = A_k(\mu D_h(x^*, z_k) + f(y_k) - f(x^*)). \tag{40}$$

We begin with the following proposition, which provides an initial error bound for algorithm (39) using the general update (39c).

Proposition 10. Assume the objective function f is μ -uniformly convex with respect to h (10) and h is σ -strongly convex. In addition, assume f is $(1/\epsilon)$ -smooth. Using the sequences (39a) and (39b), the following bound holds:

$$\frac{E_{k+1} - E_k}{\delta} \le \varepsilon_{k+1},\tag{41}$$

where the error term has the following form:

$$\varepsilon_{k+1} = \frac{A_{k+1}}{\delta} (f(y_{k+1}) - f(x_k)) + \frac{A_{k+1}}{\delta} \left(\frac{\tau_k}{2\epsilon} - \frac{\sigma\mu}{2\tau_k} \right) \|x_k - y_k\|^2 - \frac{A_{k+1}\mu\sigma}{2\delta} \|x_k - y_k\|^2 + \frac{\alpha_k}{\delta} \langle \nabla f(x_k), y_k - x_k \rangle + \frac{A_{k+1}\sigma\mu}{2\delta} \|\tau_k (\nabla h(x_k) - \nabla h(z_k) - \frac{1}{\mu} \nabla f(x_k))\|^2.$$

When h is Euclidean, the error simplifies to the following form

$$\varepsilon_{k+1} = \frac{A_{k+1}}{\delta} \left(f(y_{k+1}) - f(x_k) + \frac{\tau_k^2}{2\mu} \|\nabla f(x_k)\|^2 + \left(\frac{\tau_k}{2\epsilon} - \frac{\mu}{2\tau_k}\right) \|x_k - y_k\|^2 \right).$$

We present a proof of Proposition 10 in Appendix B.4. The result for accelerated gradient descent can be summed up in the following corollary, which is a consequence of Propositions 6 and 10.

Corollary 11. Using the gradient step.

$$\mathcal{G}(x) = x_k - \epsilon \nabla f(x_k),$$

for update (39c) results in an error which scales as

$$\varepsilon_{k+1} = \frac{A_{k+1}}{\delta} \left(\frac{\tau_k^2}{2\mu} - \frac{\epsilon}{2} \right) \|\nabla f(x_k)\|^2 + \frac{A_{k+1}}{\delta} \left(\frac{\tau_k}{2\epsilon} - \frac{\mu}{2\tau_k} \right) \|x_k - y_k\|^2,$$

when h is Euclidean.

The parameter choice $\tau_k \leq \sqrt{\mu\epsilon} = 1/\sqrt{\kappa}$ ensures the error is non-positive. With this choice, we obtain a linear $O(e^{-\sqrt{\mu\epsilon}k}) = O(e^{-k/\sqrt{\kappa}})$ convergence rate. Again, if we take the discretization step to scale according to the smoothness as $\delta = \sqrt{\epsilon}$, then both $||x_k - y_k|| = O(\sqrt{\epsilon})$ and $\varepsilon_k = O(\sqrt{\epsilon})$, so we recover the dynamics (7) and the continuous Lyapunov argument $\dot{\mathcal{E}}_t \leq 0$ in the limit $\sqrt{\epsilon} \to 0$.

3.2.2 Quasi-monotone method

We end this section by studying a family of algorithms which can be thought of as a variation of the implicit Euler scheme applied to (35b) and (35b),

$$x_{k+1} = \frac{\tau_k}{1 + \tau_k} z_k + \frac{1}{1 + \tau_k} x_k \tag{42a}$$

$$\nabla h(z_{k+1}) = \nabla h(z_k) + \tau_k \left(\nabla h(x_{k+1}) - \nabla h(z_{k+1}) - (1/\mu) \nabla f(x_{k+1}) \right), \tag{42b}$$

where $\tau_k = \frac{A_{k+1} - A_k}{A_k} := \frac{\alpha_k}{A_k}$. In discretization (42a), the state z_{k+1} has been replaced by the state z_k . When h is Euclidean, we can write (42b) as the following update:

$$z_{k+1} = \arg\min_{z \in \mathcal{X}} \left\{ \langle \nabla f(x_{k+1}), z \rangle + \frac{\mu}{2\tau_k} \|z - \tilde{z}_{k+1}\|^2 \right\}.$$

where $\tilde{z}_{k+1} = \frac{z_k + \tau_k x_{k+1}}{1 + \tau_k}$. The update (42b) involves optimizing a linear approximation to the function regularized by a weighted combination of Bregman divergences. This yields the result summarized in the following proposition.

Proposition 12. Assume f is μ -strongly convex with respect to h and h is σ -strongly convex. The following error bound:

$$\frac{E_{k+1} - E_k}{\delta} \le \varepsilon_{k+1},$$

can be shown for algorithm (42) using Lyapunov function (37), where the error scales as

$$\varepsilon_{k+1} = \frac{A_k \tau_k^2}{2\mu\sigma\delta} \|\nabla f(x_{k+1})\|^2. \tag{43}$$

No smoothness assumptions on f and h are needed to show this bound, and we can replace all the gradients with subgradients. If we assume that all the subgradients of f are upper bounded in norm, then optimizing this bound results in an $f(x_k) - f(x^*) \leq O(1/k)$ convergence rate for the function value, which is optimal for subgradient methods designed for strongly convex functions.⁴

3.3 Frank-Wolfe algorithms

In this section we describe how Frank-Wolfe algorithms can, in a sense, be considered as discretetime mappings of dynamics which satisfy the conditions,

$$Z_t = X_t + \dot{\beta}_t^{-1} \dot{X}_t, \tag{44a}$$

$$0 \le \langle \nabla f(X_t), x - Z_t \rangle, \quad \forall x \in \mathcal{X}. \tag{44b}$$

⁴In particular, this rate is achieved by taking $\tau_k = \frac{2}{k+2}$.

These dynamics are not guaranteed to exist; however, they are remarkably similar to the dynamics (4), where instead of using the Bregman divergence to ensure nonnegativity of the variational inequality $0 \leq \dot{\beta}_t e^{\beta_t} \langle \nabla f(X_t), x - Z_t \rangle$, we simply assume (44b) holds on the domain \mathcal{X} . We summarize the usefulness of dynamics (44) in the following proposition.

Proposition 13. Assume f is convex and the ideal scaling (3b) holds. The following function:

$$\mathcal{E}_t = e^{\beta_t} (f(X_t) - f(x)), \tag{45}$$

is a Lyapunov function for the dynamics which satisfies (44). We can therefore conclude an $O(e^{-\beta_t})$ convergence rate of dynamics (44) to the minimizer of the function.

The proof of this Proposition is in Appendix B.6. Here, we will analyze two Frank-Wolfe algorithms that arise from dynamics (44). Applying the backward-Euler scheme to (44a) and (44b), with the same approximations, $\frac{d}{dt}X_t = \frac{x_{k+1}-x_k}{\delta}$, $\frac{d}{dt}e^{\beta_t} = \frac{A_{k+1}-A_k}{\delta}$, and denoting $\tau_k = \frac{A_{k+1}-A_k}{A_{k+1}}$, we obtain the variational conditions for the following algorithm:

$$z_k = \arg\min_{z \in \mathcal{X}} \langle \nabla f(x_k), z \rangle,$$
 (46a)

$$x_{k+1} = \tau_k z_k + (1 - \tau_k) x_k. \tag{46b}$$

Update (46a) requires the assumptions that \mathcal{X} be convex and compact; under this assumption, (46a) satisfies

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

consistent with (44b). The following proposition describes how a discretization of (45) can be used to analyze the behavior of algorithm (46).

Proposition 14. Assume f is convex and \mathcal{X} is convex and compact. If f is $(1/\epsilon)$ -smooth, using the Lyapunov function,

$$E_k = A_k(f(x_k) - f(x)), \tag{47}$$

we obtain the error bound,

$$\frac{E_{k+1} - E_k}{\delta} \le \varepsilon_{k+1},$$

where the error for algorithm (46) scales as

$$\varepsilon_{k+1} = \frac{A_{k+1}\tau_k^2}{2\epsilon\delta} \|z_k - x_k\|^2. \tag{48}$$

If instead we assume f has (ϵ, ν) -Hölder-continuous gradients (30), the error in algorithm (46) now scales as

$$\varepsilon_{k+1} = \frac{A_{k+1}\tau_k^{1+\nu}}{(1+\nu)\epsilon\delta} \|z_k - x_k\|^{1+\nu}.$$
 (49)

Taking $x = x^*$ we infer the convergence rates $O(1/\epsilon k)$ and $O(1/\epsilon k^{\nu})$, respectively. We provide a proof of Proposition 14 in Appendix B.7.

4 Equivalence to Estimate Sequences

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

4.1 Estimate sequences

We provide a brief review of the technique of estimate sequences [18]. We begin with the following definition.

Definition 1. [18, 2.2.1] A pair of sequences $\{\phi_k(x)\}_{k=1}^{\infty}$ and $\{A_k\}_{k=0}^{\infty}$ $A_k \ge 1$ is called an estimate sequence of function f(x) if

$$A_k^{-1} \to 0$$
,

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

$$\phi_k(x) \le \left(1 - A_k^{-1}\right) f(x) + A_k^{-1} \phi_0(x). \tag{50}$$

The following lemma, due to Nesterov, explains why estimate sequences are useful.

Lemma 15. [18, 2.2.1] If for some sequence $\{x_k\}_{k>0}$ we have

$$f(x_k) \le \phi_k^* \equiv \min_{x \in \mathcal{X}} \phi_k(x), \tag{51}$$

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

Proof. The proof is straightforward:

$$f(x_k) \stackrel{\text{(51)}}{\leq} \phi_k^* \equiv \min_{x \in \mathcal{X}} \phi_k(x) \stackrel{\text{(50)}}{\leq} \min_{x \in \mathcal{X}} \left[\left(1 - A_k^{-1} \right) f(x) + A_k^{-1} \phi_0(x) \right] \leq \left(1 - A_k^{-1} \right) f(x^*) + A_k^{-1} \phi_0(x^*).$$

Rearranging gives the desired inequality.

Notice that this definition is not constructive. Finding sequences which satisfy these conditions is 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, 24, 23], and is, to the best of our knowledge, the only known formal way to construct an estimate sequence. We will see below that this particular class of estimate sequences can be turned into our Lyapunov functions with a few algebraic manipulations (and vice versa).

Proposition 16. [3, 2.2] Let $\phi_0 : \mathcal{X} \to \mathbb{R}$ be a convex function such that $\min_{x \in \mathcal{X}} \phi_0(x) \geq f^*$. Suppose also that we have a sequence $\{f_k\}_{k \geq 0}$ of functions from \mathcal{X} to \mathbb{R} that underestimates f:

$$f_k(x) \le f(x)$$
 for all $x \in \mathcal{X}$ and all $k \ge 0$. (52)

Define recursively $A_0 = 1$, $\tau_k = \frac{A_{k+1} - A_k}{A_{k+1}} := \frac{\alpha_k}{A_k}$, and

$$\phi_{k+1}(x) := (1 - \tau_k)\phi_k(x) + \tau_k f_k(x) = A_{k+1}^{-1} \left(A_0 \phi_0(x) + \sum_{i=0}^k a_i f_i(x) \right), \tag{53}$$

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

From (51) and (53), we observe that the following invariant:

$$A_{k+1}f(x_{k+1}) \le \min_{x} A_{k+1}\phi_{k+1}(x) = \min_{x} \sum_{i=0}^{k} \alpha_i f_i(x) + A_0\phi_0(x), \tag{54}$$

is maintained. In [24, 23], this technique was extended to incorporate an error term $\{\tilde{\varepsilon}_k\}_{k=1}^{\infty}$,

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

where $\varepsilon_k \geq 0, \forall k$. Rearranging, we have the following bound:

$$A_{k+1}f(x_{k+1}) \le \min_{x} A_{k+1}\phi_{k+1}(x) = \min_{x} \sum_{i=0}^{k} \alpha_{i}f_{i}(x) + A_{0}\left(\phi_{0}(x) - A_{0}^{-1}\tilde{\varepsilon}_{0}\right) + \tilde{\varepsilon}_{k+1}.$$

Notice that an argument analogous to that of Lemma 15 holds:

$$A_{k+1}f(x_{k+1}) \le \sum_{i=0}^{k} \alpha_i f_i(x^*) + A_0(\phi_0(x^*) - \tilde{\varepsilon}_0) + \tilde{\varepsilon}_{k+1} \le \sum_{i=0}^{k} \alpha_i f(x^*) + A_0\phi_0(x^*) + \tilde{\varepsilon}_{k+1}$$
$$= A_{k+1}f(x^*) + A_0\phi_0(x^*) + \tilde{\varepsilon}_{k+1}.$$

Rearranging, we obtain the desired bound,

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

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

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

Table 1: Choices of estimate sequences for various algorithms

In Table 1 "linear" is defined as $f_i(x) = f(x_i) + \langle \nabla f(x_i), x - x_i \rangle$, and "quadratic" is defined as $f_i(x) = f(x_i) + \langle \nabla f(x_i), x - x_i \rangle + \frac{\mu}{2} ||x - x_i||^2$. The estimate-sequence argument is inductive; one must know the three sequences $\{\varepsilon_k, A_k, \phi_k(x)\}$ a priori in order to check the invariants hold. This aspect of the estimate-sequence technique has made it hard to discern its structure and scope.

4.2 Equivalence to Lyapunov functions

We now demonstrate an equivalence between these two frameworks. The continuous-time view shows that the errors in both the Lyapunov function and estimate sequences are due to discretization errors. We demonstrate how this works for accelerated methods, and defer the proofs for the other algorithms discussed earlier in the paper to Appendix C.

Equivalence in discrete time. The discrete-time estimate sequence (53) for accelerated gradient descent can be written:

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

$$\stackrel{(53)}{=} (1 - \tau_k) \phi_k(x) + \tau_k f_k(x)$$

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

Multiplying through by A_{k+1} , we have the following argument, which follows directly from our definitions:

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

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

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

The last inequality follows from definition (52). Rearranging, we obtain the inequality $E_{k+1} \leq E_k$ for our Lyapunov function (21). Going the other direction, from our Lyapunov analysis we can derive the following bound:

$$E_{k} \leq E_{0}$$

$$A_{k}(f(x_{k}) - f(x)) + D_{h}(x, z_{k}) \leq A_{0}(f(x_{0}) - f(x)) + D_{h}(x, z_{0})$$

$$A_{k}(f(x_{k}) - A_{k}^{-1}D_{h}(x, z_{k})) \leq (A_{k} - A_{0})f(x) + A_{0}(f(x_{0}) + A_{0}^{-1}D_{h}(x^{*}, z_{0}))$$

$$A_{k}\phi_{k}(x) \leq (A_{k} - A_{0})f(x) + A_{0}\phi_{0}(x).$$
(55)

Rearranging, we obtain the estimate sequence (50), with $A_0 = 1$:

$$\phi_k(x) \le \left(1 - A_k^{-1} A_0\right) f(x) + A_k^{-1} A_0 \phi_0(x) = \left(1 - A_k^{-1}\right) f(x) + A_k^{-1} \phi_0(x).$$

Writing $\mathcal{E}_t \leq \mathcal{E}_0$, one can simply rearrange terms to extract an estimate sequence:

$$f(X_t) + e^{-\beta_t} D_h(x, Z_t) \le \left(1 - e^{-\beta_t} e^{\beta_0}\right) f(x^*) + e^{-\beta_t} e^{\beta_0} \left(f(X_0) + e^{-\beta_0} D_h(x, Z_0)\right).$$

Comparing this to (55), matching terms allows us to extract the continuous-time estimate sequence $\{\phi_t(x), e^{\beta_t}\}$, where $\phi_t(x) = f(X_t) + e^{-\beta_t}D_h(x, Z_t)$.

5 Further Observations

The dynamical perspective can be extended to the derivation and analysis of a range of other methods. In this section, we provide sketches of some of these analyses, providing a detailed treatment in Appendix D.

Proximal methods. Methods for minimizing the composite of two convex functions, $\varphi(x) = f(x) + \psi(x)$, were introduced by Nesterov [21] and studied by Beck and Teboulle [4], Tseng [32] and several others. In Appendix D.1, we present a dynamical perspective on these methods and show how to recover their convergence theory via the Lyapunov functions presented in this paper.

Stochastic methods. We sketch a high-level view of algorithms which use stochastic estimates of gradients, and provide a more detailed analysis in Appendix D.2. Our scope is a Lyapunov-based analysis of four algorithms—stochastic mirror descent with momentum, accelerated (proximal) coordinate descent [2, 25, 33, 10, 28], accelerated stochastic variance reduction (SVRG) [1], and accelerated stochastic composite methods [14]. We study these methods under two smoothness settings and present proofs for several explicit methods. Broadly, we consider algorithms (19), (42) and (39), where stochastic gradients are used instead of full gradients. For these methods, we show the bound $\mathbb{E}[E_{k+1}] - E_k \leq \mathbb{E}[\varepsilon_{k+1}]$ for Lyapunov function (17) and $\mathbb{E}[E_{k+1}] - E_k \leq -\tau_k E_k + \mathbb{E}[\varepsilon_{k+1}]$ for Lyapunov function (40), where the expectation is taken conditioned on the previous state. By summing, we obtain convergence rates for the aforementioned algorithms, provided the sequence $(A_i)_{i=1}^{\infty}$ is chosen so that $\mathbb{E}[\sum_{i=1}^{\infty} \varepsilon_i] < \infty$.

6 Discussion

The main contributions in this paper are twofold: We have presented a unified analysis of a wide variety of algorithms using three Lyapunov functions—(21), (40) and (47), and we have demonstrated the equivalence between Lyapunov functions and estimate sequences, under the formalization of the latter due to Baes [3]. More generally, we have provided a dynamical-systems perspective that builds on Polyak's early intuitions, and elucidates connections between discrete-time algorithms and continuous-time, dissipative second-order dynamics. We believe that the dynamical perspective renders the design and analysis of accelerated algorithms for optimization particularly transparent, and we also note in passing that Lyapunov analyses for non-accelerated gradient-based methods, such as mirror descent and natural gradient descent, can be readily derived from analyses of gradient-flow dynamics.

We close with a brief discussion of some possible directions for future work. First, we remark that requiring a continuous-time Lyapunov function to remain a Lyapunov function in discrete time places significant constraints on which ODE solvers can be used. In this paper, we show that we can derive new algorithms using a restricted set of ODE techniques (several of which are nonstandard) but it remains to be seen if other methods can be applied in this setting. Techniques such as the midpoint method and Runge Kutta provide more accurate solutions of ODEs than Euler methods [6]. 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. In a similar vein, it would be of interest to analyze the symplectic integrators studied by [17] within our Lyapunov framework.

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 [30]. Therefore, many authors [26, 30, 13] 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 [30] and Krichene, Bayen and Bartlett [13] 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 (6) might lead to better results.

Earlier work by Drori and Teboulle [8], Kim and Fessler [12], Taylor *et al* [31], and Lessard *et al* [15] have shown that optimization algorithms can 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 as well as Orianna Demassi and Stephen Tu for the many helpful discussions involving this paper. ACW was supported by an NSF Graduate Research Fellowship. This work was supported in part by the Army Research Office under grant number W911NF-17-1-0304 and by the Mathematical Data Science program of the Office of Naval Research.

References

- [1] Z. Allen-Zhu, Katyusha: The first direct acceleration of stochastic gradient methods, in STOC, 2017.
- [2] Z. Allen-Zhu, P. Richtárik, Z. Qu, and Y. Yuan, Even faster accelerated coordinate descent using non-uniform sampling, in Proceedings of the 33rd International Conference on Machine Learning, ICML '16, 2016.
- [3] M. BAES, Estimate sequence methods: Extensions and approximations. Manuscript, available at http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf, August 2009.
- [4] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.
- [5] S. Bubeck, Y. T. Lee, and M. Singh, A geometric alternative to Nesterov's accelerated gradient descent, ArXiv preprint arXiv:1506.08187, (2015).
- [6] J. Butcher, Numerical methods for ordinary differential equations in the 20th century, Journal of Computational and Applied Mathematics, 125 (2000), pp. 1–29.
- [7] P. L. Chebyshev, Théorie des mécanismes connus sous le nom de parallélogrammes, Mémoires Présentés à l'Académie Impériale des Sciences de St-Pétersbourg, VII (1854).
- [8] Y. Drori and M. Teboulle, Performance of first-order methods for smooth convex minimization: a novel approach, Math. Program., 145 (2014), pp. 451–482.
- [9] D. Drusvyatskiy, M. Fazel, and S. Roy, An optimal first order method based on optimal quadratic averaging, ArXiv preprint arXiv:1604.06543, (2016).
- [10] O. Fercoq and P. Richtárik, Accelerated, parallel, and proximal coordinate descent, SIAM Journal on Optimization, 25 (2015), pp. 1997–2023.

- [11] G. N. Grapiglia and Y. Nesterov, Regularized Newton methods for minimizing functions with Hölder continuous Hessians, SIAM Journal on Optimization, 27 (2017), pp. 478–506.
- [12] D. Kim and J. A. Fessler, Optimized first-order methods for smooth convex minimization, Mathematical Programming, 159 (2016), pp. 81–107.
- [13] W. KRICHENE, A. BAYEN, AND P. BARTLETT, Accelerated mirror descent in continuous and discrete time, in Advances in Neural Information Processing Systems (NIPS) 29, 2015.
- [14] G. Lan, An optimal method for stochastic composite optimization, Mathematical Programming, 133 (2012), pp. 365–397.
- [15] L. LESSARD, B. RECHT, AND A. PACKARD, Analysis and design of optimization algorithms via integral quadratic constraints, SIAM Journal on Optimization, 26 (2016), pp. 57–95.
- [16] A. M. LYAPUNOV AND A. T. FULLER, General problem of the stability of motion, 1992.
- [17] B. MICHAEL, J. MICHEAL, AND W. ASHIA, On symplectic optimization. Arxiv preprint arXiv1802.03653, March 2018.
- [18] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Applied Optimization, Kluwer, Boston, 2004.
- [19] Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical Programming, 103 (2005), pp. 127–152.
- [20] Y. Nesterov, Accelerating the cubic regularization of Newton's method on convex problems, Mathematical Programming, 112 (2008), pp. 159–181.
- [21] Y. Nesterov, Gradient methods for minimizing composite functions, Mathematical Programming, 140 (2013), pp. 125–161.
- [22] Y. Nesterov, Universal gradient methods for convex optimization problems, Mathematical Programming, (2014), pp. 1–24.
- [23] Y. Nesterov, Complexity bounds for primal-dual methods minimizing the model of objective function, tech. report, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2015.
- [24] Y. Nesterov and V. Shikhman, Quasi-monotone subgradient methods for nonsmooth convex minimization, Journal of Optimization Theory and Applications, 165 (2015), pp. 917–940.
- [25] Y. Nesterov and S. U. Stich, Efficiency of the accelerated coordinate descent method on structured optimization problems, SIAM Journal on Optimization, 27 (2017), pp. 110–123.
- [26] B. O'DONOGHUE AND E. CANDÈS, Adaptive restart for accelerated gradient schemes, Foundations of Computational Mathematics, 15 (2015), pp. 715–732.
- [27] B. T. Polyak, Some methods of speeding up the convergence of iteration methods, USSR Computational Mathematics and Mathematical Physics, 4 (1964), pp. 1–17.

- [28] L. X. Qihang Lin, Zhaosong Lu, An accelerated proximal coordinate gradient method, in Advances in Neural Information Processing Systems (NIPS) 27, 2014.
- [29] D. E. RUMELHART, G. E. HINTON, AND R. J. WILLIAMS, Learning representations by back-propagating errors, Nature, 323 (1986), pp. 533–536.
- [30] W. Su, S. Boyd, and E. Candes, A differential equation for modeling nesterov's accelerated gradient method: Theory and insights, in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., Curran Associates, Inc., 2014, pp. 2510–2518.
- [31] A. B. Taylor, J. M. Hendrickx, and F. Glineur, Smooth strongly convex interpolation and exact worst-case performance of first-order methods, Mathematical Programming, (2016), pp. 1–39.
- [32] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, SIAM Journal on Optimization, (2008).
- [33] S. Tu, S. Venkataraman, A. C. Wilson, A. Gittens, M. I. Jordan, and B. Recht, *Breaking locality accelerates block gauss-seidel*, in Proceedings of the 34th International Conference on Machine Learning, ICML 2017, 2017, pp. 1549–1557.
- [34] A. Wibisono, A. C. Wilson, and M. I. Jordan, A variational perspective on accelerated methods in optimization, Proceedings of the National Academy of Sciences, 133 (2016), pp. E7351–E7358.

A Dynamics

A.1 Proof of Proposition 1

We compute the Euler-Lagrange equation for the second Bregman Lagrangian (5). Denote $z = x + e^{-\alpha t}\dot{x}$. The partial derivatives of the Bregman Lagrangian can be written,

$$\frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t) = \mu e^{\beta_t + \gamma_t} \left(\nabla h(Z_t) - \nabla h(X_t) \right)
\frac{\partial \mathcal{L}}{\partial x}(X_t, \dot{X}_t, t) = \mu e^{\alpha_t} \frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t) - \mu e^{\beta_t + \gamma_t} \frac{d}{dt} \nabla h(X_t) - e^{\alpha_t + \beta_t + \gamma_t} \nabla f(X_t).$$

We also compute the time derivative of the momentum $p = \frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t)$,

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t) = (\dot{\beta}_t + \dot{\gamma}_t)\frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t) + \mu e^{\beta_t + \gamma_t}\frac{d}{dt}\nabla h(Z_t) - \mu e^{\beta_t + \gamma_t}\frac{d}{dt}\nabla h(X_t).$$

The terms involving $\frac{d}{dt}\nabla h(X)$ cancel and the terms involving the momentum will simplify under the scaling condition (3a) when computing the Euler-Lagrange equation $\frac{\partial \mathcal{L}}{\partial x}(X_t, \dot{X}_t, t) = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t)$. Compactly, the Euler-Lagrange equation can be written

$$\frac{d}{dt}\mu\nabla h(Z_t) = -\dot{\beta}_t\mu\left(\nabla h(Z_t) - \nabla h(X_t)\right) - e^{\alpha_t}\nabla f(x).$$

Remark. It is interesting to compare with the partial derivatives of the first Bregman Lagrangian (2),

$$\frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t) = e^{\gamma_t} \left(\nabla h(Z_t) - \nabla h(X_t) \right)
\frac{\partial \mathcal{L}}{\partial x}(X_t, \dot{X}_t, t) = e^{\alpha_t} \frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t) - e^{\gamma_t} \frac{d}{dt} \nabla h(X_t) - e^{\alpha_t + \beta_t + \gamma_t} \nabla f(X_t),$$

as well as the derivative of the momentum,

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t) = \dot{\gamma}_t \frac{\partial \mathcal{L}}{\partial v}(X_t, \dot{X}_t, t) + e^{\gamma_t} \frac{d}{dt} \nabla h(Z_t) - e^{\gamma_t} \frac{d}{dt} \nabla h(X_t).$$

For Lagrangian (2), not only do the terms involving $\frac{d}{dt}\nabla h(X)$ cancel when computing the Euler-Lagrange equation, but the ideal scaling will also force the terms involving the momentum to cancel as well.

A.2 Deriving the Lyapunov functions

A.2.1 Proof of Proposition 2

We demonstrate how to derive the Lyapunov function (21) for the momentum dynamics (4); this derivation is similar in spirit to the Lyapunov analysis of mirror descent by Nemirovski and Yudin.

Denote $Z_t = X_t + e^{-\alpha_t} \dot{X}_t$. We have:

$$\begin{split} \frac{d}{dt}D_{h}\left(x,Z_{t}\right) &= \frac{d}{dt}\left(h(x) - h(Z_{t}) - \left\langle\nabla h(Z_{t}), x - Z_{t}\right\rangle\right) \\ &= -\left\langle\nabla h(Z_{t}), \dot{Z}_{t}\right\rangle - \left\langle\frac{d}{dt}\nabla h(Z_{t}), x - Z_{t}\right\rangle + \left\langle\nabla h(Z_{t}), \dot{Z}_{t}\right\rangle \\ &= -\left\langle\frac{d}{dt}\nabla h\left(Z_{t}\right), x - Z_{t}\right\rangle. \end{split}$$

Using this identity, we obtain the following argument:

$$\frac{d}{dt}D_{h}(x,Z_{t}) = -\left\langle \frac{d}{dt}\nabla h\left(Z_{t}\right), x - Z_{t}\right\rangle
= e^{\alpha_{t}+\beta_{t}}\left\langle \nabla f(X_{t}), x - X_{t} - e^{-\alpha_{t}}\dot{X}_{t}\right\rangle
= e^{\alpha_{t}+\beta_{t}}\langle \nabla f(X_{t}), x - X_{t}\rangle - e^{\beta_{t}}\langle \nabla f(X_{t}), \dot{X}_{t}\rangle
= e^{\alpha_{t}+\beta_{t}}\langle \nabla f(X_{t}), x - X_{t}\rangle - \frac{d}{dt}\left(e^{\beta_{t}}f(X_{t})\right) + \dot{\beta}_{t}e^{\beta_{t}}f(X_{t})
= \dot{\beta}_{t}e^{\beta_{t}}[f(X_{t}) + \langle \nabla f(X_{t}), x - X_{t}\rangle] - \frac{d}{dt}\left(e^{\beta_{t}}f(X_{t})\right) + e^{\beta_{t}}\left(e^{\alpha_{t}} - \dot{\beta}_{t}\right)\langle \nabla f(X_{t}), x - X_{t}\rangle
\leq \dot{\beta}_{t}e^{\beta_{t}}f(x) - \frac{d}{dt}\left(e^{\beta_{t}}f(X_{t})\right)
= -\frac{d}{dt}\left\{e^{\beta_{t}}\left(f(X_{t}) - f(x)\right)\right\}.$$
(58a)

(58a)

Here (58a) uses the momentum dynamics (15b) and (15a). The inequality (58b) follows from the convexity of f. If $\dot{\beta}_t = e^{\alpha_t}$, simply by rearranging terms and taking $x = x^*$, we have shown that the function (9) has nonpositive derivative for all t and is hence a Lyapunov function for the family of momentum dynamics (4). If $\dot{\beta}_t \leq e^{\alpha_t}$, the Lyapunov function is only decreasing for $x = x^*$.

A.2.2 Proof of Proposition 3

We demonstrate how to derive the Lyapunov function (12) for the momentum dynamics (6). Using the same identity (57), we have the following initial,

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

The Bregman three-point identity,

$$\langle \nabla h(Z_t) - \nabla h(X_t), x - Z_t \rangle + D_h(x, Z_t) = D_h(x, X_t) - D_h(Z_t, X_t), \tag{60}$$

will now be useful. Proceeding from the last line, we have

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

The first inequality follows from the μ -uniform convexity of f with respect to h. The second inequality follows from nonnegativity of the Bregman divergence, and the ideal scaling condition (3b), where we must take $x = x^*$ if $\dot{\beta}_t \leq e^{\alpha_t}$.

B Algorithms derived from dynamics (4)

B.1 Proof of Proposition 5

We show the initial bounds (23a) and (23b). We begin with algorithm (19):

$$\begin{split} E_{k+1} - E_k &= D_h(x, z_{k+1}) - D_h(x, z_k) + A_{k+1}(f(y_{k+1}) - f(x)) - A_k(f(y_k) - f(x)) \\ &= -\langle \nabla h(z_{k+1}) - \nabla h(z_k), x - z_{k+1} \rangle - D_h(z_{k+1}, z_k) + A_{k+1}(f(y_{k+1}) - f(x)) - A_k(f(y_k) - f(x)) \\ &\stackrel{\text{(19b)}}{=} \alpha_k \langle \nabla f(x_{k+1}), x - z_{k+1} \rangle - D_h(z_{k+1}, z_k) + \alpha_k(f(x_{k+1}) - f(x)) + A_k(f(x_{k+1}) - f(y_k)) \\ &+ A_{k+1}(f(y_{k+1}) - f(x_{k+1})) \\ &\leq \alpha_k \langle \nabla f(x_{k+1}), x - z_k \rangle + \alpha_k \langle \nabla f(x_{k+1}), z_k - z_{k+1} \rangle - \frac{\sigma}{p} \|z_{k+1} - z_k\|^p + \alpha_k(f(x_{k+1}) - f(x)) \\ &+ A_k(f(x_{k+1}) - f(y_k)) + A_{k+1}(f(y_{k+1}) - f(x_{k+1})) \\ &\leq \alpha_k \langle \nabla f(x_{k+1}), x - z_k \rangle + A_k(f(x_{k+1}) - f(y_k)) + \alpha_k(f(x_{k+1}) - f(x)) \\ &+ \frac{p-1}{p} \sigma^{-\frac{1}{p-1}} (A_{k+1} - A_k)^{\frac{p}{p-1}} \|\nabla f(x_{k+1})\|^{\frac{p}{p-1}} + A_{k+1}(f(y_{k+1}) - f(x_{k+1})). \end{split}$$

The first inequality follows from the σ -uniform convexity of h with respect to the p-th power of the norm and the last inequality follows from the Fenchel Young inequality. If we continue with our argument, and plug in the identity (23a), it simply remains to use our second update (19a):

$$E_{k+1} - E_{k} \leq \alpha_{k} \langle \nabla f(x_{k+1}), x - z_{k} \rangle + A_{k} (f(x_{k+1}) - f(y_{k})) + \alpha_{k} (f(x_{k+1}) - f(x))$$

$$+ \frac{p-1}{p} \sigma^{-\frac{1}{p-1}} (A_{k+1} - A_{k})^{\frac{p}{p-1}} \| \nabla f(x_{k+1}) \|^{\frac{p}{p-1}} + A_{k+1} (f(y_{k+1}) - f(x_{k+1}))$$

$$\leq \alpha_{k} \langle \nabla f(x_{k+1}), x - y_{k} \rangle + A_{k+1} \langle \nabla f(x_{k+1}), y_{k} - x_{k+1} \rangle + A_{k} (f(x_{k+1}) - f(y_{k}))$$

$$+ \alpha_{k} (f(x_{k+1}) - f(x)) + \varepsilon_{k+1}$$

$$= \alpha_{k} (f(x_{k+1}) - f(x)) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle) + A_{k} (f(x_{k+1}) - f(y_{k})) + \langle \nabla f(x_{k+1}), y_{k} - x_{k+1} \rangle$$

$$+ \varepsilon_{k+1}.$$

From here, we can conclude $E_{k+1} - E_k \leq \varepsilon_k$ using the convexity of f.

We now show the bound (23b) for algorithm (20) using a similar argument.

$$\begin{split} E_{k+1} - E_k &= D_h(x, z_{k+1}) - D_h(x, z_k) + A_{k+1}(f(y_{k+1}) - f(x)) - A_k(f(y_k) - f(x)) \\ &\stackrel{(19b)}{=} \alpha_k \langle \nabla f(y_{k+1}), x - z_{k+1} \rangle - D_h(z_{k+1}, z_k) + \alpha_k(f(y_{k+1}) - f(x)) + A_k(f(y_{k+1}) - f(y_k)) \\ &\leq \alpha_k \langle \nabla f(y_{k+1}), x - z_k \rangle + \alpha_k \langle \nabla f(y_{k+1}), z_k - z_{k+1} \rangle - \frac{\sigma}{p} \|z_{k+1} - z_k\|^p \\ &+ \alpha_k(f(y_{k+1}) - f(x)) + A_k(f(y_{k+1}) - f(y_k)) \\ &\leq \alpha_k \langle \nabla f(y_{k+1}), x - z_k \rangle + A_k(f(y_{k+1}) - f(y_k)) + \alpha_k(f(y_{k+1}) - f(x)) \\ &- A_{k+1} \langle \nabla f(y_{k+1}), y_{k+1} - x_{k+1} \rangle + \varepsilon_{k+1}. \end{split}$$

The first inequality follows from the uniform convexity of h and the second uses the Fenchel Young inequality and definition (23b). Using the second update (20a), we obtain our initial error bound:

$$\begin{split} E_{k+1} - E_k &\leq \alpha_k \langle \nabla f(y_{k+1}), x - y_k \rangle + A_k (f(y_{k+1}) - f(y_k)) + \alpha_k (f(y_{k+1}) - f(x)) \\ &\quad + A_{k+1} \langle \nabla f(y_{k+1}), y_k - x_{k+1} \rangle - A_{k+1} \langle \nabla f(y_{k+1}), y_{k+1} - x_{k+1} \rangle + \varepsilon_{k+1} \\ &= \alpha_k (f(y_{k+1}) - f(x) + \langle \nabla f(y_{k+1}), x - y_{k+1} \rangle) \\ &\quad + A_k (f(y_{k+1}) - f(y_k) + \langle \nabla f(y_{k+1}), y_k - y_{k+1} \rangle) + \varepsilon_{k+1}. \end{split}$$

The last line can be upper bounded by the error ε_{k+1} using convexity of f.

B.2 Proof of Proposition 7

A similar progress bound was proved in Wibisono, Wilson and Jordan [34, Lem 3.2]. Note that $y = \mathcal{G}(x)$ satisfies the optimality condition

$$\sum_{i=1}^{p-1} \frac{1}{(i-1)!} \nabla^{i} f(x) (y-x)^{i-1} + \frac{N}{\epsilon} ||y-x||^{\tilde{p}-2} (y-x) = 0.$$
 (62)

Furthermore, since $\nabla^{p-1}f$ is Hölder-continuous (30), we have the following error bound on the (p-2)-nd order Taylor expansion of ∇f ,

$$\left\| \nabla f(y) - \sum_{i=0}^{p-1} \frac{1}{(i-1)!} \nabla^i f(x) (y-x)^{i-1} \right\| = \left\| \int_0^1 \left[\nabla^{p-1} f(ty + (1-t)x) - \nabla^{p-1} f(x) \right] (y-x)^{p-2} dt \right\|$$

$$\leq \frac{1}{\epsilon} \|y - x\|^{p-2+\nu} \int_0^1 t^{\nu} = \frac{1}{\epsilon} \|y - x\|^{\tilde{p}-1}. \quad (63)$$

Substituting (62) to (63) and writing r = ||y - x||, we obtain

$$\left\|\nabla f(y) + \frac{Nr^{\tilde{p}-2}}{\epsilon} (y-x)\right\|_{*} \le \frac{r^{\tilde{p}-1}}{\epsilon}.$$
 (64)

Now the argument proceeds as in [34]. Squaring both sides, expanding, and rearranging the terms, we get the inequality

$$\langle \nabla f(y), x - y \rangle \ge \frac{\epsilon}{2Nr^{\tilde{p}-2}} \|\nabla f(y)\|_*^2 + \frac{(N^2 - 1)r^{\tilde{p}}}{2N\epsilon}.$$
 (65)

Note that if $\tilde{p}=2$, then the first term in (65) already implies the desired bound (32). Now assume $\tilde{p} \geq 3$. The right-hand side of (65) is of the form $A/r^{\tilde{p}-2}+Br^{\tilde{p}}$, which is a convex function of r>0 and minimized by $r^*=\left\{\frac{(\tilde{p}-2)}{\tilde{p}}\frac{A}{B}\right\}^{\frac{1}{2\tilde{p}-2}}$, yielding a minimum value of

$$\frac{A}{(r^*)^{\tilde{p}-2}} + B(r^*)^p = A^{\frac{\tilde{p}}{2\tilde{p}-2}} B^{\frac{\tilde{p}-2}{2\tilde{p}-2}} \left[\left(\frac{\tilde{p}}{\tilde{p}-2} \right)^{\frac{\tilde{p}-2}{2\tilde{p}-2}} + \left(\frac{\tilde{p}-2}{\tilde{p}} \right)^{\frac{\tilde{p}}{\tilde{p}-2}} \right] \geq A^{\frac{\tilde{p}}{2\tilde{p}-2}} B^{\frac{\tilde{p}-2}{2\tilde{p}-2}}.$$

Substituting the values $A = \frac{\epsilon}{2N} \|\nabla f(y)\|_*^2$ and $B = \frac{1}{2N\epsilon} (N^2 - 1)$ from (65), we obtain

$$\langle \nabla f(y), x - y \rangle \geq \left(\frac{\epsilon}{2N} \|\nabla f(y)\|_*^2 \right)^{\frac{\tilde{p}}{2\tilde{p}-2}} \left(\frac{1}{2N\epsilon} (N^2 - 1) \right)^{\frac{\tilde{p}-2}{2\tilde{p}-2}} = \frac{(N^2 - 1)^{\frac{\tilde{p}-2}{2\tilde{p}-2}}}{2N} \epsilon^{\frac{1}{\tilde{p}-1}} \|\nabla f(y)\|_*^{\frac{\tilde{p}}{\tilde{p}-1}},$$

which proves the progress bound (32).

B.3 Proof of Universal Gradient Method

We present a convergence rate for higher-order gradient method $y_{k+1} = \mathcal{G}_{\epsilon,p,\nu,N}(x_{k+1})$ where \mathcal{G} is given by (31) and f has (ϵ,ν) -Hölder-continuous gradients (30). The proof is inspired by the proof of the rescaled gradient flow $\dot{X}_t = -\nabla f(X_t)/\|\nabla f(X_t)\|_*^{\frac{p-2}{p-1}}$, outlined in [34, Appendix G], which is the continuous-time limit of the algorithm. Using the Lyapunov function

$$\mathcal{E}_t = t^p(f(X_t) - f(x)),$$

the following argument can be made using the convexity of f and the dynamics:

$$\dot{\mathcal{E}}_{t} = t^{p} \langle \nabla f(X_{t}), \dot{X}_{t} \rangle + pt^{p-1} (f(X_{t}) - f(x^{*}))
\leq t^{p} \langle \nabla f(X_{t}), \dot{X}_{t} \rangle + pt^{p-1} \langle \nabla f(X_{t}), X_{t} - x^{*} \rangle
= -t^{p} \|\nabla f(X_{t})\|_{*}^{\frac{p}{p-1}} + pt^{p-1} \langle \nabla f(X_{t}), X_{t} - x^{*} \rangle
\leq \frac{1}{p-1} \|(p-1)(X_{t} - x^{*})\|^{p}
\leq (p-1)^{p-1} R^{p}.$$

The last two inequalities use the Fenchel-Young inequality and the fact that $||X_t - x^*|| \le R$ since rescaled gradient flow is a descent method. We can conclude $O(t^{p-1})$ convergence rate by integrating. We now proceed with the discrete-time argument by using the Lyapunov function (47) $(\tilde{p} \ge 2)$:

$$E_k = A_k(f(x_k) - f(x))$$

We argue as follows:

$$E_{k+1} - E_k = A_k(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_{k+1}) - f(x^*))$$

$$\leq A_k \langle \nabla f(x_{k+1}), x_{k+1} - x_k \rangle + \alpha_k \langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle$$

$$\leq -A_k \frac{(N^2 - 1)^{\frac{\tilde{p} - 2}{2\tilde{p} - 2}}}{2N} \epsilon^{\frac{1}{\tilde{p} - 1}} \|\nabla f(x_{k+1})\|_{*}^{\frac{\tilde{p}}{\tilde{p} - 1}} + \alpha_k \langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle$$

$$\leq \frac{1}{\epsilon} \alpha_k^{\tilde{p}} A_k^{1 - \tilde{p}} \frac{(\tilde{p} - 1)^{\tilde{p} - 1}}{\tilde{p}^{\tilde{p}}} \left(\frac{(N^2 - 1)^{\frac{\tilde{p} - 2}{2\tilde{p} - 2}}}{2N} \right)^{-\frac{\tilde{p} - 1}{\tilde{p}}} \|x_{k+1} - x^*\|_{*}^{\tilde{p}},$$

where the first inequality follows from convexity, the second inequality uses 7, and the third line uses Young's inequality, $\langle s,u\rangle+\frac{1}{p}\|u\|^p\leq -\frac{p-1}{p}\|s\|_*^{\frac{p-1}{p}}$, $2\geq p\in\mathbb{R}$ with the identifications

$$s = \epsilon^{1/\tilde{p}} \nabla f(x_{k+1}) \left(A_k \frac{(N^2 - 1)^{\frac{\tilde{p} - 2}{2\tilde{p} - 2}}}{2N} \right)^{\frac{\tilde{p} - 1}{\tilde{p}}}$$
$$u = (x_{k+1} - x^*) \epsilon^{-\frac{1}{\tilde{p}}} \left(A_k \frac{(N^2 - 1)^{\frac{\tilde{p} - 2}{2\tilde{p} - 2}}}{2N} \right)^{-\frac{\tilde{p} - 1}{\tilde{p}}} \alpha_k \frac{\tilde{p} - 1}{\tilde{p}}.$$

From Lemma 7, it follows that this method is a descent method. Furthermore, we can choose $\alpha_k^{\tilde{p}} A_k^{1-\tilde{p}} \leq C$, for some constant C, by choosing A_k to be a polynomial of degree \tilde{p} . By summing we obtain the desired $O(1/k^{\tilde{p}-1})$ convergence rate.

B.4 Proof of Proposition 10

We show the initial error bound (41). To do so, we define the Lyapunov function,

$$\tilde{E}_k = f(y_k) - f(x^*) + \mu D_h(x, z_k).$$
(66)

Note that we simply need to show $\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k \tilde{E}_k + \varepsilon_{k+1}/A_{k+1}$ where $\tau_k = \frac{A_{k+1} - A_k}{A_{k+1}}$. Thus, we begin with the following bound:

$$\begin{split} \tilde{E}_{k+1} - \tilde{E}_k &= f(y_{k+1}) - f(y_k) - \mu \langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_{k+1} \rangle - \mu D_h(z_{k+1}, z_k) \\ &\leq f(y_{k+1}) - f(x_k) + f(x_k) - f(y_k) - \mu \langle \nabla h(z_{k+1}) - \nabla h(z_k), x^* - z_k \rangle + \frac{\sigma \mu}{2} \|\nabla h(z_{k+1}) - \nabla h(z_k)\|^2 \\ &\leq f(y_{k+1}) - f(x_k) + \langle \nabla f(x_k), x_k - y_k \rangle - \mu D_h(x_k, y_k) - \mu \langle \nabla h(z_{k+1}) - \nabla h(z_k), x - z_k \rangle \\ &+ \frac{\sigma \mu}{2} \|\nabla h(z_{k+1}) - \nabla h(z_k)\|^2 \\ \stackrel{\text{(39b)}}{=} f(y_{k+1}) - f(x_k) + \langle \nabla f(x_k), x_k - y_k \rangle - \mu D_h(x_k, y_k) + \tau_k \langle \nabla f(x_k), x - z_k \rangle \\ &- \mu \tau_k \langle \nabla h(x_k) - \nabla h(z_k), x^* - z_k \rangle + \frac{\sigma \mu}{2} \|\nabla h(z_{k+1}) - \nabla h(z_k)\|^2 \\ \stackrel{\text{(39a)}}{=} f(y_{k+1}) - f(x_k) + \langle \nabla f(x_k), x_k - y_k \rangle - \mu D_h(x_k, y_k) + \tau_k \langle \nabla f(x_k), x - x_k \rangle \\ &- \tau_k \mu \langle \nabla h(x_k) - \nabla h(z_k), x - z_k \rangle + \langle \nabla f(x_k), y_k - x_k \rangle + \frac{\sigma \mu}{2} \|\nabla h(z_{k+1}) - \nabla h(z_k)\|^2 \\ \leq -\tau_k (f(x_k) - f(x^*) + \mu D_h(x, x_k)) + f(y_{k+1}) - f(x_k) - \frac{\sigma \mu}{2} \|x_k - y_k\|^2 \\ &- \tau_k \mu \langle \nabla h(x_k) - \nabla h(z_k), x^* - z_k \rangle + \frac{\sigma \mu}{2} \|\nabla h(z_{k+1}) - \nabla h(z_k)\|^2 \\ = -\tau_k (f(y_k) - f(x^*) + \mu D_h(x, z_k)) + f(y_{k+1}) - f(x_k) - \frac{\mu \sigma}{2} \|x_k - y_k\|^2 \\ &+ \frac{\sigma \mu}{2} \|\nabla h(z_{k+1}) - \nabla h(z_k)\|^2 - \tau_k \mu D_h(x_k, z_k) + \tau_k (f(y_k)) - f(x_k)) \\ \leq -\tau_k \tilde{E}_k + f(y_{k+1}) - f(x_k) - \frac{\mu \sigma}{2} \|x_k - y_k\|^2 + \tau_k \langle \nabla f(x_k), y_k - x_k \rangle \\ &+ \frac{\sigma \mu}{2} \|\tau_k (\nabla h(x_k) - \nabla h(z_k) - \frac{1}{\mu} \nabla f(x_k))\|^2 - \left(\frac{\sigma \mu}{2\tau_k} - \frac{\tau_k}{2\epsilon}\right) \|x_k - y_k\|^2 \\ \leq -\tau_k \tilde{E}_k + \varepsilon_{k+1}/A_{k+1}. \end{split}$$

The first inequality uses the σ -strong convexity of h and the Fenchel-Young inequality. The second inequality uses the μ -strong convexity of f with respect to h. The third inequality uses the strong convexity of f and σ -strong convexity of h. The following line uses the Bregman three point identity (60) and the subsequent inequality uses the strong convexity of f. The last line follows from the smoothness of f. Now we turn to the case where h is Euclidean (so $\sigma = 1$):

$$\begin{split} \tilde{E}_{k+1} - \tilde{E}_{k} &\leq -\tau_{k} \tilde{E}_{k} + f(y_{k+1}) - f(x_{k}) - \frac{\mu}{2} \|x_{k} - y_{k}\|^{2} + \tau_{k} \langle \nabla f(x_{k}), y_{k} - x_{k} \rangle \\ &+ \frac{\mu}{2} \|\tau_{k}(x_{k} - z_{k}) - \frac{\tau_{k}}{\mu} \nabla f(x_{k}))\|^{2} - \left(\frac{\mu}{2\tau_{k}} - \frac{\tau_{k}}{2\epsilon}\right) \|x_{k} - y_{k}\|^{2} \\ &= -\tau_{k} \tilde{E}_{k} + f(y_{k+1}) - f(x_{k}) - \frac{\mu}{2} \|x_{k} - y_{k}\|^{2} + \tau_{k} \langle \nabla f(x_{k}), y_{k} - x_{k} \rangle + \frac{\mu}{2} \|\tau_{k}(x_{k} - z_{k})\| \\ &- \tau_{k} \langle \nabla f(x_{k}), \tau_{k}(x_{k} - z_{k}) \rangle + \frac{\tau_{k}^{2}}{2\mu} \|\nabla f(x_{k})\|^{2} - \left(\frac{\mu}{2\tau_{k}} - \frac{\tau_{k}}{2\epsilon}\right) \|x_{k} - y_{k}\|^{2} \\ &= -\tau_{k} \tilde{E}_{k} + f(y_{k+1}) - f(x_{k}) + \frac{\tau_{k}^{2}}{2\mu} \|\nabla f(x_{k})\|^{2} - \left(\frac{\mu}{2\tau_{k}} - \frac{\tau_{k}}{2\epsilon}\right) \|x_{k} - y_{k}\|^{2}. \end{split}$$

In the second line we have expanded the square. The last line uses the update (39a).

B.5 Proof of Proposition 12

We show the convergence bound for the quasi-monotone method (42). We have,

$$\begin{split} \tilde{E}_{k+1} - \tilde{E}_k &= -\langle \nabla h(z_{k+1}) - \nabla h(z_k), x - z_{k+1} \rangle - D_h(z_{k+1}, z_k) + f(x_{k+1}) - f(x_k) \\ &\stackrel{(42b)}{=} \tau_k \langle \nabla f(x_{k+1}), x - z_k \rangle + \tau_k \langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x - z_{k+1} \rangle \\ &+ \tau_k \langle \nabla f(x_{k+1}), z_k - z_{k+1} \rangle - D_h(z_{k+1}, z_k) + f(x_{k+1}) - f(x_k) \\ &\leq \tau_k \langle \nabla f(x_{k+1}), x - z_k \rangle + \tau_k \langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x - z_{k+1} \rangle \\ &+ \frac{\tau_k^2 \sigma}{2} \|\nabla f(x_{k+1})\|^2 + f(x_{k+1}) - f(x_k) \\ &= \tau_k \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle + \tau_k \langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x - z_{k+1} \rangle \\ &+ \frac{\tau_k^2 \sigma}{2} \|\nabla f(x_{k+1})\|^2 + f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \\ &\leq -\tau_k (f(x_{k+1}) - f(x^*) + D_h(x, x_{k+1})) + \tau_k \langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x - z_{k+1} \rangle \\ &+ \frac{\tau_k^2 \sigma}{2} \|\nabla f(x_{k+1})\|^2 \\ &\leq -\tau_k (f(x_{k+1}) - f(x^*) + D_h(x, z_{k+1})) + \frac{\tau_k^2 \sigma}{2} \|\nabla f(x_{k+1})\|^2 \end{split}$$

The first inequality from the strong convexity of h as well as Hölder's inequality. The second inequality from the uniform convexity of f with respect to h and convexity of f. The last line follows from the Bregman three-point identity (60) and non-negativity of the Bregman divergence. Taking $\tau_k = \frac{A_{k+1} - A_k}{A_k}$ gives the desired error bound.

B.6 Proof of Proposition 13

We show that (47) is a Lyapunov function for dynamics (44). The argument is simple:

$$0 \leq \dot{\beta}_{t}e^{\beta_{t}}\langle \nabla f(X_{t}), x - Z_{t}\rangle = \dot{\beta}_{t}e^{\beta_{t}}\langle \nabla f(X_{t}), x - X_{t}\rangle - e^{\beta_{t}}\langle \nabla f(X_{t}), \dot{X}_{t}\rangle$$

$$= \dot{\beta}_{t}e^{\beta_{t}}\langle \nabla f(X_{t}), x - X_{t}\rangle - \frac{d}{dt}\left\{e^{\beta_{t}}f(X_{t})\right\} + \dot{\beta}_{t}e^{\beta_{t}}f(X_{t})$$

$$\leq -\frac{d}{dt}\left\{e^{\beta_{t}}(f(X_{t}) - f(x))\right\}.$$

B.7 Proof of Proposition 14

If we take $\nu = 1$ bound (49) implies (48); therefore we simply show the bound (49). To that end,

$$\begin{split} E_{k+1} - E_k &= A_{k+1}(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_k) - f(x)) \\ &\leq A_{k+1} \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{A_{k+1}}{(1+\nu)\epsilon} \|x_{k+1} - x_k\|^{1+\nu} + \alpha_k \langle \nabla f(x_k), x - x_k \rangle \\ &\stackrel{\text{(46b)}}{=} \alpha_k \langle \nabla f(x_k), z_k - x_k \rangle + \frac{A_{k+1} \alpha_k^{1+\nu}}{(1+\nu)\epsilon} \|z_k - x_k\|^{1+\nu} + \alpha_k \langle \nabla f(x_k), x - x_k \rangle \\ &\stackrel{\text{(46a)}}{\leq} \frac{A_{k+1} \alpha_k^{1+\nu}}{(1+\nu)\epsilon} \|z_k - x_k\|^{1+\nu}. \end{split}$$

The first inequality follows from the Hölder continuity and convexity of f. The rest simply follows from plugging in our identities.

C Estimate Sequences

C.1 The quasi-monotone subgradient method

The discrete-time estimate sequence (53) for quasi-monotone subgradient method can be written:

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

$$\stackrel{(53)}{=} (1 - \tau_k) \left(\phi_k(x) - A_k^{-1} \tilde{\varepsilon}_k \right) + \tau_k f_k(x)$$

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

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

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

$$- (A_{k+1} - \alpha_k)A_k^{-1}\tilde{\varepsilon}_k + \alpha_k f_k(x)$$

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

$$\stackrel{(52)}{\leq} A_k f(x_k) + D_h(x, z_k) - \tilde{\varepsilon}_k + \alpha_k f(x).$$

Rearranging, we obtain our Lyapunov argument $E_{k+1} \leq E_k + \varepsilon_{k+1}$ for (21):

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

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

$$E_{k} \leq E_{0} + \tilde{\varepsilon}_{k}$$

$$A_{k}(f(x_{k}) - f(x)) + D_{h}(x, z_{k}) \leq A_{0}(f(x_{0}) - f(x)) + D_{h}(x, z_{0}) + \tilde{\varepsilon}_{k}$$

$$A_{k}\left(f(x_{k}) - \frac{1}{A_{k}}D_{h}(x, z_{k})\right) \leq (A_{k} - A_{0})f(x) + A_{0}\left(f(x_{0}) + \frac{1}{A_{0}}D_{h}(x^{*}, z_{0})\right) + \tilde{\varepsilon}_{k}$$

$$A_{k}\phi_{k}(x) \leq (A_{k} - A_{0})f(x) + A_{0}\phi_{0}(x) + \tilde{\varepsilon}_{k}.$$

$$(68)$$

Rearranging, we obtain our estimate sequence (50) $(A_0 = 1)$ with an additional error term:

$$\phi_k(x) \le \left(1 - \frac{A_0}{A_k}\right) f(x) + \frac{A_0}{A_k} \phi_0(x) + \frac{\tilde{\varepsilon}_k}{A_k} = \left(1 - \frac{1}{A_k}\right) f(x) + \frac{1}{A_k} \phi_0(x) + \frac{\tilde{\varepsilon}_k}{A_k}. \tag{69a}$$

C.2 Frank-Wolfe

The discrete-time estimate sequence (53) for conditional gradient method can be written:

$$\phi_{k+1}(x) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} := f(x_{k+1}) - \frac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} \stackrel{(53)}{=} (1 - \tau_k) \left(\phi_k(x) - \frac{\tilde{\varepsilon}_k}{A_k} \right) + \tau_k f_k(x)$$

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

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

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

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

$$\stackrel{(52)}{\leq} A_k f(x_k) - \tilde{\varepsilon}_k + (A_{k+1} - A_k) f(x).$$

Rearranging, we obtain our Lyapunov argument $E_{k+1} - E_k \le \varepsilon_{k+1}$ for (47):

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

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

$$E_k \le E_0 + \tilde{\varepsilon}_k$$

$$A_k f(x_k) \le (A_k - A_0) f(x) + A_0 f(x_0) + \tilde{\varepsilon}_k$$

$$A_k \phi_k(x) \le (A_k - A_0) f(x) + A_0 \phi_0(x) + \tilde{\varepsilon}_k$$

Rearranging, we obtain our estimate sequence (50) $(A_0 = 1)$ with an additional error term:

$$\phi_k(x) \le \left(1 - \frac{A_0}{A_k}\right) f(x) + \frac{A_0}{A_k} \phi_0(x) + \frac{\tilde{\varepsilon}_k}{A_k} = \left(1 - \frac{1}{A_k}\right) f(x) + \frac{1}{A_k} \phi_0(x) + \frac{\tilde{\varepsilon}_k}{A_k}.$$

Given that the Lyapunov function property allows us to write

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

we can extract $\{f(X_t), e^{\beta_t}\}$ as the continuous-time estimate sequence for Frank-Wolfe.

C.3 Accelerated gradient descent (strong convexity)

The discrete-time estimate sequence (53) for accelerated gradient descent can be written:

$$\phi_{k+1}(x) := f(x_{k+1}) + \frac{\mu}{2} \|x - z_{k+1}\|^2 \stackrel{(53)}{=} (1 - \tau_k) \phi_k(x) + \tau_k f_k(x) \stackrel{(52)}{\leq} (1 - \tau_k) \phi_k(x) + \tau_k f(x)$$

Therefore, we obtain the inequality $\tilde{E}_{k+1} - \tilde{E}_k \leq -\tau_k \tilde{E}_k$ for our Lyapunov function (66) by simply writing $\phi_{k+1}(x) - f(x) + f(x) - \phi_k(x) \leq -\tau_k (\phi_k(x) - f(x))$:

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

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

Going the other direction, we have,

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

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

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

Summing over the right-hand side, we obtain the estimate sequence (50):

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

Since the Lyapunov function property allows us to write

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

we can extract $\{f(X_t) + \frac{\mu}{2} ||x - Z_t||^2, e^{\beta_t}\}$ as the continuous-time estimate sequence for accelerated gradient descent in the strongly convex setting.

C.4 Existence and uniqueness

In this section, we show existence and uniqueness of solutions for the differential equations (6), when h is Euclidean. To do so, we write the dynamics as the following system of equations

$$\dot{X}_t = \sqrt{\mu}(W_t - 2X_t) \tag{72a}$$

$$\dot{W}_t = -\frac{1}{\sqrt{\mu}} \nabla f(X_t), \tag{72b}$$

where we have taken $W_t = Z_t + X_t$ and $\beta_t = \sqrt{\mu}t$. Now if we assume ∇f is Lipschitz continuous, then over any bounded interval $[t_0, t_1]$ with $0 \le t_0 < t_1$, the right-hand side of (72) is a Lipschitz-continuous vector field. Therefore, by the Cauchy-Lipschitz theorem, for any initial conditions $(X_{t_0}, W_{t_0}) = (x_0, w_0)$ at time $t = t_0$, the system of differential equations has a unique solution over the time interval $[t_0, t_1]$. Since t_1 is arbitrary and the energy is decreasing in t_1 , this shows that there is a unique maximal solution for any $t_1 \to \infty$. To show a unique solution exists for an arbitrary β_t , we show the family of dynamics (6) is closed under time-dilation (similar to dynamics (4)). Thus, if a unique solution exists for any setting β_t , we can conclude it exists for all β_t . To demonstrate the time-dilation property, we calculate the velocity and acceleration of the reparameterized curve $Y_t = X_{\tau_t}$, where $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function of time:

$$\begin{split} \dot{Y}_t &= \dot{\tau}_t \dot{X}_{\tau_t} \\ \ddot{Y}_t &= \ddot{\tau}_t \dot{X}_{\tau_t} + \dot{\tau}_t^2 \ddot{X}_{\tau_t} \\ \dot{\tilde{\beta}}_t &= \frac{d}{dt} \beta_{\tau_t} = \dot{\tau}_t \dot{\beta}_{\tau_t} \\ \ddot{\tilde{\beta}}_t &= \frac{d}{dt} \dot{\tau}_t \dot{\beta}_{\tau_t} = \ddot{\tau}_t \dot{\beta}_{\tau_t} + \dot{\tau}_t^2 \ddot{\beta}_{\tau_t}. \end{split}$$

Inverting the first of these relations, we get

$$\begin{split} \dot{X}_{\tau_t} &= \frac{1}{\dot{\tau}_t} \dot{Y}_t \\ \ddot{X}_{\tau_t} &= \frac{1}{\dot{\tau}_t^2} \ddot{Y}_t - \frac{\ddot{\tau}_t}{\dot{\tau}_t^3} \dot{Y}_t. \\ \dot{\beta}_{\tau_t} &= \frac{1}{\dot{\tau}_t} \dot{\tilde{\beta}}_t \\ \ddot{\beta}_{\tau_t} &= \frac{1}{\dot{\tau}_t^2} \ddot{\tilde{\beta}}_t - \frac{\ddot{\tau}_t}{\dot{\tau}_t^2} \dot{\beta}_{\tau_t}. \end{split}$$

Computing the time-dilated Euler-Lagrange equation, we get

$$Z_{\tau_t} = X_{\tau_t} + \frac{1}{\dot{\beta}_{\tau_t}} \dot{X}_{\tau_t} = Y_t + \frac{1}{\dot{\beta}_{\tau_t} \dot{\tau}_t} \dot{Y}_t = Y_t + \frac{1}{\dot{\beta}_t} \dot{Y}_t$$

for the first equation, as well as the identity

$$\begin{split} \dot{Z}_{\tau_t} &= \dot{X}_{\tau_t} + \frac{1}{\dot{\beta}_{\tau_t}} \ddot{X}_{\tau} - \frac{\ddot{\beta}_{\tau_t}}{\dot{\beta}_{\tau_t}^2} \dot{X}_{\tau_t} \\ &= \frac{1}{\dot{\tau}_t} \dot{Y}_t + \frac{1}{\dot{\tau}_t \dot{\tilde{\beta}}_{\tau_t}} \ddot{Y}_t - \frac{\ddot{\tau}_t}{\dot{\tilde{\beta}}_t \dot{\tau}_t^2} \dot{Y}_t - \frac{\ddot{\beta}_{\tau_t}}{\dot{\tau}_t \dot{\beta}_{\tau_t}^2} \dot{Y}_t \\ &= \frac{1}{\dot{\tau}_t} \dot{Y}_t + \frac{1}{\dot{\tau}_t \dot{\tilde{\beta}}_{\tau_t}} \ddot{Y}_t - \frac{\ddot{\tau}_t}{\dot{\tilde{\beta}}_t \dot{\tau}_t^2} \dot{Y}_t - \frac{\ddot{\tilde{\beta}}_t}{\dot{\tau}_t \dot{\tilde{\beta}}_t^2} \dot{Y}_t + \frac{\ddot{\tau}_t}{\dot{\tau}_t^2 \dot{\tilde{\beta}}_{\tau_t}} \dot{Y}_t \\ &= \frac{1}{\dot{\tau}_t} \left(\dot{Y}_t + \frac{1}{\dot{\tilde{\beta}}_{\tau_t}} \ddot{Y}_t - \frac{\ddot{\tilde{\beta}}_t}{\dot{\tilde{\beta}}_t^2} \dot{Y}_t \right). \end{split}$$

Therefore the second equation

$$\nabla^2 h(Z_{\tau_t}) \dot{Z}_{\tau_t} = -\dot{\beta}_{\tau_t} (\nabla h(X_{\tau_t}) - \nabla h(Z_{\tau_t}) - \frac{1}{\mu} \nabla f(X_{\tau_t}))$$

can be written,

$$\frac{1}{\dot{\tau}_t} \left(\nabla^2 h \left(Y_t + \frac{1}{\dot{\tilde{\beta}}_t} \dot{Y}_t \right) \left(\dot{Y}_t + \frac{1}{\dot{\tilde{\beta}}_{\tau_t}} \ddot{Y}_t - \frac{\ddot{\tilde{\beta}}_t}{\dot{\tilde{\beta}}_t^2} \dot{Y}_t \right) \right) = -\frac{\dot{\tilde{\beta}}_t}{\dot{\tau}_t} \left(\nabla h(Y_t) - \nabla h \left(Y_t + \frac{1}{\dot{\tilde{\beta}}_t} \dot{Y}_t \right) - \frac{1}{\mu} \nabla f(Y_t) \right),$$

which is the Euler-Lagrange equation for the sped-up curve, where the ideal scaling holds with equality. Finally, we mention that we can deduce the existence/uniqueness of solution for the proximal dynamics (74) and (80) from the existence/uniqueness of solution for dynamics (4) and (6), given the difference between these dynamics is that (74) (80) have an extra Lipschitz-continuous vector field. Thus, the Cauchy-Lipschitz theorem can be readily applied to the proximal dynamics and the same arguments can be made regarding time-dilation.

D Additional Observations

D.1 Proximal algorithms

D.1.1 Convex functions [32, 4, 21]

In 2009, Beck and Teboulle introduced FISTA, which is a method for minimizing the composite of two convex functions

$$f(x) = \varphi(x) + \psi(x) \tag{73}$$

where φ is $(1/\epsilon)$ -smooth and ψ is simple. The canonical example of this is $\psi(x) = ||x||_1$, which defines the ℓ_1 -ball. The following proposition provides dynamical intuition for momentum algorithms derived for this setting.

Proposition 17. Define $f = \varphi + \psi$ and assume φ and ψ are convex. Under the ideal scaling condition (3b), Lyapunov function (9) can be used to show that solutions to dynamics

$$Z_t = X_t + e^{-\alpha_t} \dot{X}_t \tag{74a}$$

$$\frac{d}{dt}\nabla h(Z_t) = -e^{\alpha_t + \beta_t}(\nabla \varphi(X_t) + \nabla \psi(Z_t))$$
(74b)

satisfy $f(X_t) - f(x^*) \le O(e^{-\beta_t})$.

The same Lyapunov argument can be made for the dynamics (74) if we replace $\nabla \psi(Z_t)$ with a directional subgradient at the position Z_t , provided $\beta_t = p \log t$ for $p \in \mathbb{R}$.

Proof.

$$\frac{d}{dt}D_{h}(x,Z_{t}) = -\left\langle \frac{d}{dt}\nabla h(Z_{t}), x - Z_{t} \right\rangle
= e^{\alpha_{t}+\beta_{t}} \left\langle \nabla \varphi(X_{t}), x - X_{t} - e^{-\alpha_{t}}\dot{X}_{t} \right\rangle + e^{\alpha_{t}+\beta_{t}} \left\langle \nabla \psi(Z_{t}), x - Z_{t} \right\rangle
\leq -\frac{d}{dt} \left\{ e^{\beta_{t}} \left(\varphi(X_{t}) - \varphi(x) \right) \right\} + e^{\alpha_{t}+\beta_{t}} \left\langle \nabla \psi(Z_{t}), x - Z_{t} \right\rangle
\leq -\frac{d}{dt} \left\{ e^{\beta_{t}} \left(\varphi(X_{t}) - \varphi(x) \right) \right\} + \dot{\beta}_{t} e^{\beta_{t}} (\psi(x) - \psi(Z_{t}))
\leq -\frac{d}{dt} \left\{ e^{\beta_{t}} \left(\varphi(X_{t}) - f(x) \right) \right\} - \dot{\beta}_{t} e^{\beta_{t}} (\psi(X_{t}) + \langle \nabla \psi(X_{t}), Z_{t} - X_{t} \rangle)
= -\frac{d}{dt} \left\{ e^{\beta_{t}} \left(\varphi(X_{t}) - f(x) \right) \right\} - \dot{\beta}_{t} e^{\beta_{t}} \psi(X_{t}) - e^{\beta_{t}} \langle \nabla \psi(X_{t}), \dot{X}_{t} \rangle)
= -\frac{d}{dt} \left\{ e^{\beta_{t}} \left(f(X_{t}) - f(x) \right) \right\}.$$

The first line follows from the Bregman identity (57). The second line plugs in the dynamics (80a) and (80b). The third lines follows from (58). The fourth and fifth lines follow from convexity. The sixth line plugs in the dynamics (80b) and the last line follows from application of the chain rule.

Next, to show results for dynamics when subgradients of the function are used, we adopt the setting of Su, Boyd and Candes [30, p.35]. First, we define the subgradient through the following lemma.

Lemma 18 (Rockafellar, 1997). For any convex function f and any $x, v \in \mathbb{R}^n$, the directional derivative $\lim_{\delta \to 0+} (f(x+\delta v) - f(x))/\delta$ exists, and can be evaluated as

$$\lim_{\delta \to 0+} (f(x+\delta v) - f(x))/\delta = \sup_{w \in \partial f(x)} \langle w, v \rangle.$$

Definition 2. A Borel-measurable function $G_f(x,v)$ defined on $\mathbb{R}^n \times \mathbb{R}^n$ is said to be a directional subgradient of f if

$$G_f(x, v) \in \partial f(X)$$

 $\langle G_f(x, v), v \rangle = \sup_{w \in \partial f(x)} \langle w, v \rangle,$

for all x, v.

This guarantees the existence of a directional derivative. Now we establish the following theorem (similar to [30, Thm 24]):

Theorem 19. Given the sum of two convex functions $f(x) = \varphi(x) + \psi(x)$ with directional subgradient $G_{\psi}(x, v)$, assume that the second-order ODE

$$Z_t = X_t + \frac{t}{p}\dot{X}_t$$

$$\frac{d}{dt}\nabla h(Z_t) = -pt^{p-1}(G_{\varphi}(X_t, \dot{X}_t) + G_{\psi}(Z_t, \dot{Z}_t))$$

admits a solution X_t on $[0, \alpha)$ for some $\alpha > 0$. Then for any $0 < t < \alpha$, we have $f(X_t) - f(x) \le O(1/t^p)$.

Proof. We follow the framework of Su, Boyd and Candes [30, pg. 36]. It suffices to establish that our Lyapunov function is monotonically decreasing. Although \mathcal{E}_t may not be differentiable, we can study $\mathcal{E}(t + \Delta t) - \mathcal{E}(t)/\Delta t$ for small $\Delta t > 0$. For the first term, note that

$$(t + \Delta t)^{p}(f(X_{t+\Delta t}) - f(x)) - t^{p}(f(X_{t}) - f(x)) = t^{p}(f(X_{t+\Delta t}) - f(X_{t})) + pt^{p-1}(f(X_{t+\Delta t}) - f(x))\Delta t + o(\Delta t) = t^{p}\langle G_{f}(X_{t}, \dot{X}_{t}), \dot{X}_{t}\rangle \Delta t + pt^{p-1}(f(X_{t+\Delta t}) - f(x))\Delta t + o(\Delta t),$$

where the second line follows since we assume f is locally Lipschitz. The $o(\Delta t)$ does not affect the function in the limit:

$$f(X_{t+\Delta t}) = f(X + \Delta t \dot{X}_t + o(\Delta t)) = f(X + \Delta t \dot{X}_t) + o(\Delta t)$$

= $f(X_t) + \langle G_f(X_t, \dot{X}_t), \dot{X}_t \rangle \Delta t + o(\Delta t).$ (75)

The second term, $D_h(x, X_t + \frac{t}{p}\dot{X}_t)$, is differentiable, with derivative $-\langle \frac{d}{dt}\nabla h(Z_t), x - Z_t \rangle$. Hence,

$$D_{h}\left(x, X_{t+\Delta t} + \frac{t+\Delta t}{p}\dot{X}_{t+\Delta t}\right) - D_{h}\left(x, X_{t} + \frac{t}{p}\dot{X}_{t}\right)$$

$$= -\left\langle \frac{d}{dt}\nabla h(Z_{t}), x - Z_{t}\right\rangle \Delta t + o(\Delta t)$$

$$= pt^{p-1}\langle G_{\varphi}(X_{t}, \dot{X}_{t}) + G_{\psi}(Z_{t}, \dot{Z}_{t})), x - Z_{t}\rangle \Delta t + o(\Delta t)$$

$$= pt^{p-1}\langle G_{\varphi}(X_{t}, \dot{X}_{t}), x - X_{t}\rangle \Delta t + t^{p}\langle G_{\varphi}(X_{t}, \dot{X}_{t}), \dot{X}_{t}\rangle \Delta t + pt^{p-1}\langle G_{\psi}(Z_{t}, \dot{Z}_{t})), x - Z_{t}\rangle \Delta t + o(\Delta t)$$

$$\leq -pt^{p-1}(\varphi(X_{t}) - \varphi(x))\Delta t + t^{p}\langle G_{\varphi}(X_{t}, \dot{X}_{t}), \dot{X}_{t}\rangle \Delta t - pt^{p-1}(\psi(Z_{t}) - \psi(x))\Delta t + o(\Delta t)$$

$$\leq -pt^{p-1}(\varphi(X_{t}) - \varphi(x))\Delta t + t^{p}\langle G_{\varphi}(X_{t}, \dot{X}_{t}), \dot{X}_{t}\rangle \Delta t - pt^{p-1}(\psi(X_{t}) - \psi(x))\Delta t$$

$$+ t^{p}\langle G_{\psi}(X_{t}, \dot{X}_{t}), \dot{X}_{t}\rangle \Delta t + o(\Delta t)$$

$$= -pt^{p-1}(f(X_{t}) - f(x))\Delta t + t^{p}\langle G_{f}(X_{t}, \dot{X}_{t}), \dot{X}_{t}\rangle \Delta t.$$

The last two inequalities follows from the convexity of $f = \varphi + \psi$. In the last inequality, we have used the identity $Z_t - X_t = \frac{t}{p} \dot{X}_t$ in the term $pt^{p-1} \langle G_{\psi}(X_t, Z_t - X_t), Z_t - X_t \rangle$. Combining everything we have shown

$$\lim \sup_{\Delta t \to 0^+} \frac{\mathcal{E}_{t+\Delta t} - \mathcal{E}_t}{\Delta t} \le 0,$$

which along with the continuity of \mathcal{E}_t , ensures \mathcal{E}_t is a non-increasing of time.

Algorithm. Now we will discretize the dynamics (74). We assume the ideal scaling (3b) holds with equality. Using the same identifications $\dot{X}_t = \frac{x_{k+1} - x_k}{\delta}$, $\frac{d}{dt} \nabla h(Z_t) = \frac{\nabla h(z_{k+1}) - \nabla h(z_k)}{\delta}$ and $\frac{d}{dt} e^{\beta_t} = \frac{A_{k+1} - A_k}{\delta}$, we apply the implicit-Euler scheme to (74b) and the explicit-Euler scheme to (74a). Doing so, we obtain a proximal mirror descent update,

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

and the sequence (19a), respectively. We write the algorithm as

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

$$\nabla h(z_{k+1}) - \nabla h(z_k) = -\alpha_k \nabla \varphi(x_{k+1}) - \alpha_k \nabla \psi(z_{k+1})$$
(76b)

$$y_{k+1} = \mathcal{G}(x), \tag{76c}$$

where we have similarly substituted the state x_k with a sequence y_k , and added the update $y_{k+1} = \mathcal{G}(x)$. We summarize how the initial bound scales for algorithm (76) in the following proposition.

Proposition 20. Assume h is strongly convex, φ is $(1/\epsilon)$ -smooth and ψ is simple but not necessarily smooth. Using the Lyapunov function (21), the following initial bound

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

can be shown for algorithm (76), where the error scales as

$$\varepsilon_{k+1} = -\frac{\sigma}{2} \|z_{k+1} - z_k\|^2 + \frac{A_{k+1}}{2\epsilon} \|\tau_k z_k + (1 - \tau_k) y_k - y_{k+1}\|^2 + A_{k+1} \psi(y_{k+1}) - A_k \psi(y_k) - \alpha_k \psi(z_{k+1}).$$

Tseng [32, Algorithm 1] showed that the map

$$\mathcal{G}(x) = \tau_k z_{k+1} + (1 - \tau_k) y_k \tag{77}$$

can be used to simplify the error to the following,

$$\varepsilon_{k+1} = -\frac{\sigma}{2} \|z_{k+1} - z_k\|^2 + \frac{A_{k+1}\tau_k^2}{2\epsilon} \|z_{k+1} - z_k\|^2.$$
 (78)

Notice that the condition necessary for the error to be non-positive is the same as the condition for accelerated gradient descent (29). Using the same polynomial, we can conclude an $O(1/\epsilon\sigma k^2)$ convergence rate.

Proof. We begin with the observation that the update (77) and the convexity of ψ allow us to show the inequality

$$A_{k+1}\psi((1-\tau_k)y_k + \tau_k z_{k+1}) \le A_{k+1}(1-\tau_k)\psi(y_k) + A_{k+1}\tau_k\psi(z_{k+1}) \tag{79}$$

Thus we can conclude $A_{k+1}\psi(y_{k+1}) - A_k\psi(y_k) \leq \alpha_k\psi(z_{k+1})$. With this, the standard Lyapunov analysis follows:

$$\begin{split} E_{k+1} - E_k &= D_h(x, z_{k+1}) - D_h(x, z_k) + A_{k+1}(f(y_{k+1}) - f(x)) - A_k(f(y_k) - f(x)) \\ &\leq D_h(x, z_{k+1}) - D_h(x, z_k) + A_{k+1}(\varphi(y_{k+1}) - \varphi(x)) - A_k(\varphi(y_k) - \varphi(x)) + \alpha_k(\psi(z_{k+1}) - \psi(x)) \\ &= \alpha_k \langle \nabla \varphi(x_{k+1}) + \nabla \psi(z_{k+1}), x - z_{k+1} \rangle - D_h(z_{k+1}, z_k) \\ &+ A_{k+1}(\varphi(y_{k+1}) - \varphi(x)) - A_k(\varphi(y_k) - \varphi(x)) + \alpha_k(\psi(z_{k+1}) - \psi(x)) \\ &\leq \alpha_k \langle \nabla \varphi(x_{k+1}), x - z_k \rangle + \alpha_k \langle \nabla \varphi(x_{k+1}), z_k - z_{k+1} \rangle - D_h(z_{k+1}, z_k) \\ &+ A_{k+1}(\varphi(y_{k+1}) - \varphi(x)) - A_k(\varphi(y_k) - \varphi(x)) \\ &= \alpha_k \langle \nabla \varphi(x_{k+1}), x - z_k \rangle + A_{k+1} \langle \nabla \varphi(x_{k+1}), x_{k+1} - y_{k+1} \rangle - D_h(z_{k+1}, z_k) \\ &+ A_{k+1}(\varphi(y_{k+1}) - \varphi(x)) - A_k(\varphi(y_k) - \varphi(x)) \\ &\leq \alpha_k \langle \nabla \varphi(x_{k+1}), x - z_k \rangle + \frac{A_{k+1}}{2\epsilon} \|x_{k+1} - y_{k+1}\|^2 - D_h(z_{k+1}, z_k) \\ &+ \alpha_k (\varphi(x_{k+1}) - \varphi(x)) + A_k (\varphi(x_{k+1}) - \varphi(y_k)). \end{split}$$

The first inequality uses the identity (79). The second inequality follows from the convexity of ψ . The last line uses the $\frac{1}{\epsilon}$ -smoothness of φ . It simply remains to use the σ -strong convexity of h and the identities (19a) and $x_{k+1} - y_{k+1} = \tau_k(z_{k+1} - z_k)$. Continuing from the last line, and using these properties, we have

$$\begin{split} E_{k+1} - E_k &\leq \alpha_k \langle \nabla \varphi(x_{k+1}), x - x_{k+1} \rangle + \frac{A_{k+1} \tau_k^2}{2\epsilon} \|z_{k+1} - z_k\|^2 - \frac{\sigma}{2} \|z_{k+1} - z_k\|^2 \\ &+ \alpha_k (\varphi(x_{k+1}) - \varphi(x)) + A_k (\varphi(x_{k+1}) - \varphi(y_k) + \langle \nabla \varphi(x_{k+1}), y_k - x_{k+1} \rangle) \\ &\leq \frac{A_{k+1} \tau_k^2}{2\epsilon} \|z_{k+1} - z_k\|^2 - \frac{\sigma}{2} \|z_{k+1} - z_k\|^2. \end{split}$$

The last line follows from the convexity of φ .

D.1.2 Strongly convex functions

We study the problem of minimizing the composite objective $f = \varphi + \psi$ in the setting where φ is $(1/\epsilon)$ -smooth and μ -strongly convex and ψ is simple but not smooth. Like the setting where f is weakly convex, we begin with the following proposition concerning dynamics that are relevant for this setting.

Proposition 21. Define $f = \varphi + \psi$ and assume φ is μ -strongly convex with respect to h and ψ is convex. Under the ideal scaling condition (3b), Lyapunov function (12) can be used to show that solutions to dynamics,

$$Z_t = X_t + e^{-\alpha_t} \dot{X}_t \tag{80a}$$

$$\frac{d}{dt}\nabla h(Z_t) = \dot{\beta}_t \nabla h(X_t) - \dot{\beta}_t \nabla h(Z_t) - \frac{e^{\alpha_t}}{\mu} (\nabla \varphi(X_t) + \nabla \psi(Z_t)), \tag{80b}$$

satisfy $f(X_t) - f(x) < O(e^{-\beta_t})$.

Proof.

$$\frac{d}{dt} \left\{ \mu e^{\beta_t} D_h(x, Z_t) \right\} = \mu \dot{\beta}_t e^{\beta_t} D_h(x, Z_t) - \mu e^{\beta_t} \left\langle \frac{d}{dt} \nabla h(Z_t), x - Z_t \right\rangle
= \mu \dot{\beta}_t e^{\beta_t} \left(\left\langle \nabla h(Z_t) - \nabla h(X_t), x - Z_t \right\rangle + D_h(x, Z_t) \right)
+ e^{\alpha_t + \beta_t} \left\langle \nabla \varphi(X_t) + \nabla \psi(Z_t), x - Z_t \right\rangle
= \mu \dot{\beta}_t e^{\beta_t} \left(D_h(x, X_t) - D_h(Z_t, X_t) \right) + \dot{\beta}_t e^{\beta_t} \left\langle \nabla \varphi(X_t), x - X_t \right\rangle + e^{\beta_t} \left\langle \nabla \varphi(X_t), \dot{X}_t \right\rangle
+ e^{\beta_t} \left(e^{\alpha_t} - \dot{\beta}_t \right) \left\langle \nabla \varphi(X_t), x - X_t \right\rangle + e^{\alpha_t + \beta_t} \left\langle \nabla \psi(Z_t), x - Z_t \right\rangle.$$

The second line comes from plugging in dynamics (80b). The third line uses the Bregman three-point identity (60). We continue by using the strong convexity assumption:

$$\frac{d}{dt} \left\{ \mu e^{\beta_t} D_h(x, Z_t) \right\} = \leq -\dot{\beta}_t e^{\beta_t} (\varphi(X_t) - \varphi(x)) - e^{\beta_t} \langle \nabla \varphi(X_t), \dot{X}_t \rangle + e^{\beta_t} \left(e^{\alpha_t} - \dot{\beta}_t \right) \langle \nabla \varphi(X_t), x - X_t \rangle \\
- e^{\alpha_t + \beta_t} (\psi(Z_t) - \psi(x)) \\
\leq -\dot{\beta}_t e^{\beta_t} (f(X_t) - f(x)) - e^{\beta_t} \langle \nabla \varphi(X_t), \dot{X}_t \rangle - e^{\beta_t + \alpha_t} \langle \nabla \psi(X_t), Z_t - X_t \rangle \\
+ e^{\beta_t} \left(e^{\alpha_t} - \dot{\beta}_t \right) (\langle \nabla \varphi(X_t), x - X_t \rangle - (\psi(X_t) - \psi(x))) \\
\leq -\dot{\beta}_t e^{\beta_t} (f(X_t) - f(x)) - e^{\beta_t} \langle \nabla f(X_t), \dot{X}_t \rangle \\
+ e^{\beta_t} \left(e^{\alpha_t} - \dot{\beta}_t \right) (\langle \nabla \varphi(X_t), x - X_t \rangle - (\psi(X_t) - \psi(x))) \\
\leq -\frac{d}{dt} \left\{ e^{\beta_t} (f(X_t) - f(x)) \right\}.$$

The fourth line follows the strong convexity of φ and convexity of ψ . The fifth line (second inequality) uses the convexity of ψ once again. The third inequality plugs in the definition of $Z_t - X_t$ and the second-last inequality follows from the chain rule and the ideal scaling condition (3b). \square

Assume h is Euclidean and the ideal scaling (3b) holds with equality $\dot{\beta}_t = e^{\alpha_t}$. To discretize the dynamics (80b), we split the vector field (80b) into two components, $v_1(x,z,t) = \dot{\beta}_t(X_t - Z_t - (1/\mu)\nabla\varphi(X_t))$, and $v_2(x,z,t) = -\dot{\beta}_t/\mu\nabla\psi(Z_t)$ and apply the explicit Euler scheme to $v_2(x,z,t)$ and the implicit Euler scheme to $v_1(x,z,t)$, with the same identification $\dot{\beta}_t = \tau_k/\delta$ for both vector fields.⁵ This results in the proximal update

$$z_{k+1} = \arg\min_{z} \left\{ \psi(z) + \langle \nabla \varphi(x_k), z \rangle + \frac{\mu}{2\tau_k} \|z - (1 - \tau_k)z_k - \tau_k x_k\|^2 \right\}.$$
 (81)

In full, we can write the algorithm as

$$x_k = \frac{\tau_k}{1 + \tau_k} z_k + \frac{1}{1 + \tau_k} y_k \tag{82a}$$

$$z_{k+1} - z_k = \tau_k \left(x_k - z_k - \frac{1}{\mu} \nabla \varphi(x_k) - \frac{1}{\mu} \nabla \psi(z_{k+1}) \right)$$
 (82b)

$$y_{k+1} = \mathcal{G}(x). \tag{82c}$$

⁵While using the same identification of $\dot{\beta}_t$ for both vector fields is problematic—since one is being evaluated forward in time and the other backward in time—the error bounds only scale sensibly in the setting where $\dot{\beta}_t = \gamma \leq \sqrt{\mu}$ is a constant.

We summarize how the initial bound changes with this modified update in the following proposition.

Proposition 22. Assume h is Euclidean, φ is strongly convex, φ is $(1/\epsilon)$ -smooth, and ψ is convex and simple. Using the Lyapunov function (40), we have

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

for algorithm (82), where

$$\varepsilon_{k+1} = -A_{k+1} \frac{\mu}{2} \| (z_k - z_{k+1}) - \tau_k (z_k - x_k) \|^2 + \frac{A_{k+1}}{2\epsilon} \| x_k - y_{k+1} \|^2$$

$$+ A_{k+1} \left(\frac{\tau_k}{2\epsilon} - \frac{\mu}{2\tau_k} \right) \| x_k - y_k \|^2 + A_{k+1} \psi(y_{k+1}) - A_k \psi(y_k) - \alpha_k \psi(z_{k+1}).$$

Using the same update (77),

$$\mathcal{G}(x) = \tau_k z_{k+1} + (1 - \tau_k) y_k,$$

the bound simplifies nicely,

$$\varepsilon_{k+1}/A_{k+1} = \left(\frac{\tau_k^2}{2\epsilon} - \frac{\mu}{2}\right) \frac{1}{\mu} \|\nabla \varphi(x_k) + \nabla \psi(z_{k+1})\|^2 + \left(\frac{\tau_k}{2\epsilon} - \frac{\mu}{2\tau_k}\right) \|x_k - y_k\|^2.$$

The condition necessary for the error to be non-positive, $\tau_k \leq \sqrt{\epsilon \mu} = 1/\sqrt{\kappa}$, results in a $O(e^{-k/\sqrt{\kappa}})$ convergence rate. This matches the lower bound for the class of $(1/\epsilon)$ -smooth and μ -strongly convex functions. As in continuous time, this analysis also allows for the use of subgradients of ψ .

Proof.

$$\begin{split} \tilde{E}_{k+1} - \tilde{E}_k &= \frac{\mu}{2} \|x^* - z_{k+1}\|^2 - \frac{\mu}{2} \|x^* - z_k\|^2 + 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(y_k) \\ &\leq -\mu \langle z_{k+1} - z_k, x^* - z_{k+1} \rangle - \frac{\mu}{2} \|z_{k+1} - z_k\|^2 + \langle \nabla \varphi(x_k), y_{k+1} - y_k \rangle + \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 \\ &- \frac{\mu}{2} \|x_k - y_k\|^2 - \tau_k \psi(y_k) - \tau_k \psi(z_{k+1}) \\ &\stackrel{(77)}{=} -\mu \langle z_{k+1} - z_k, x^* - z_{k+1} \rangle - \frac{\mu}{2} \|z_{k+1} - z_k\|^2 + \tau_k \langle \nabla \varphi(x_k), z_{k+1} - y_k \rangle - \frac{\mu}{2} \|x_k - y_k\|^2 \\ &+ \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 - \tau_k \psi(y_k) - \tau_k \psi(z_{k+1}) \\ &\stackrel{(76b)}{=} \tau_k \langle \nabla \varphi(x_k), x^* - z_{k+1} \rangle - \frac{\mu}{2} \|z_{k+1} - z_k\|^2 + \tau_k \langle \nabla \varphi(x_k), z_{k+1} - y_k \rangle + \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 \\ &- \frac{\mu}{2} \|x_k - y_k\|^2 + \mu \tau_k \langle x_k - z_k, x^* - z_{k+1} \rangle + \tau_k \langle \nabla \psi(z_{k+1}), x^* - z_{k+1} \rangle - \tau_k \psi(y_k) - \tau_k \psi(z_{k+1}) \\ &\leq \tau_k \langle \nabla \varphi(x_k), x^* - x_k \rangle - \frac{\mu}{2} \|z_{k+1} - z_k\|^2 + \tau_k \langle \nabla \varphi(x_k), x_k - y_k \rangle + \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 \\ &- \frac{\mu}{2} \|x_k - y_k\|^2 + \mu \tau_k \langle x_k - z_k, x^* - z_{k+1} \rangle - \tau_k (\psi(y_k) - \psi(x^*)) \\ &\leq -\tau_k \left(\varphi(x_k) - \varphi(x^*) + \frac{\mu}{2} \|x^* - x_k\|^2 \right) + \mu \tau_k \langle x_k - z_k, x^* - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - y_k \rangle \\ &+ \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 - \frac{\mu}{2} \|x_k - y_k\|^2 + \mu \tau_k \langle x_k - z_k, x^* - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - y_k \rangle \\ &+ \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 - \frac{\mu}{2} \|x_k - y_k\|^2 + \mu \tau_k \langle x_k - z_k, x^* - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - y_k \rangle \\ &+ \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 - \frac{\mu}{2} \|x_k - y_k\|^2 + \mu \tau_k \langle x_k - z_k, x^* - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - y_k \rangle \\ &+ \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 - \frac{\mu}{2} \|x_k - y_k\|^2 + \mu \tau_k \langle x_k - z_k, x_k - z_k, x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - y_k \rangle \\ &+ \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 - \frac{\mu}{2} \|x_k - y_k\|^2 + \mu \tau_k \langle x_k - z_k, x_k - z_k, x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x_k - z_k \rangle + \tau_k \langle \nabla \varphi(x_k), x$$

The first inequality follows from the strong convexity and $(1/\epsilon)$ -smoothness of φ and (79), from which we can conclude $\varphi(y_{k+1}) - \varphi(y_k) \le -\tau_k \varphi(y_k) - \tau_k \varphi(z_{k+1})$. The second inequality follows from the convexity of ψ . The third inequality uses the strong convexity of f. Next, we use identity (60) and the smoothness of φ to simplify the bound as follows:

$$\begin{split} \tilde{E}_{k+1} - \tilde{E}_k &\overset{(39a)}{\leq} -\tau_k \left(\varphi(x_k) - f(x^*) + \frac{\mu}{2} \|x^* - z_k\|^2 \right) - \frac{\mu}{2\tau_k} \|x_k - y_k\|^2 + \tau_k \langle \nabla \varphi(x_k), x_k - y_k \rangle \\ &+ \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 - \frac{\mu}{2} \|x_k - y_k\|^2 + \mu \tau_k \langle x_k - z_k, z_k - z_{k+1} \rangle - \frac{\mu}{2} \|z_{k+1} - z_k\|^2 - \tau_k \psi(y_k) \\ &\leq -\tau_k \left(f(y_k) - f(x^*) + \frac{\mu}{2} \|x^* - z_k\|^2 \right) - \frac{\mu}{2\tau_k} \|x_k - y_k\|^2 + \frac{\tau_k}{2\epsilon} \|x_k - y_k\|^2 \\ &+ \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 - \frac{\mu}{2} \|x_k - y_k\|^2 + \mu \tau_k \langle x_k - z_k, z_k - z_{k+1} \rangle - \frac{\mu}{2} \|z_{k+1} - z_k\|^2 \\ &\overset{(39a)}{=} -\tau_k E_k - \frac{\tau_k^2 \mu}{2} \|x_k - z_k\|^2 + \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 + \mu \tau_k \langle x_k - z_k, z_k - z_{k+1} \rangle \\ &- \frac{\mu}{2} \|z_{k+1} - z_k\|^2 + \left(\frac{\tau_k}{2\epsilon} - \frac{\mu}{2\tau_k} \right) \|x_k - y_k\|^2 \\ &= -\tau_k E_k - \frac{\mu}{2} \|\tau_k (x_k - z_k) - (z_k - z_{k+1})\|^2 + \frac{1}{2\epsilon} \|x_k - y_{k+1}\|^2 + \left(\frac{\tau_k}{2\epsilon} - \frac{\mu}{2\tau_k} \right) \|x_k - y_k\|^2. \end{split}$$

It remains to check

$$x_k - y_{k+1} \stackrel{(77)}{=} x_k - y_k - \tau_k (z_{k+1} - y_k) \stackrel{(39a)}{=} \tau_k (z_k - x_k - z_{k+1} + y_k) \stackrel{(39a)}{=} \tau_k (\tau_k (x_k - z_k) - (z_k - z_{k+1})).$$

D.2 Stochastic methods

We begin with the following proposition.

Claim 23. Assume h is σ -strongly convex and f is convex. For algorithm (19), where stochastic gradients are used instead of full gradients and $\mathcal{G}(x) = x_{k+1}$, we can show the following error bound:

$$\frac{\mathbb{E}[E_{k+1}] - E_k}{\delta} \le \mathbb{E}[\varepsilon_{k+1}],\tag{83}$$

for Lyapunov function (17), where the error scales as

$$\mathbb{E}[\varepsilon_{k+1}] = \frac{(A_{k+1} - A_k)^2}{2\sigma\delta} \mathbb{E}[\|G(x_{k+1})\|^2]. \tag{84}$$

For algorithm (42), where stochastic gradients are used instead of full gradients, we can show the following error bound:

$$\frac{\mathbb{E}[E_{k+1}] - E_k}{\delta} \le -\frac{\tau_k}{\delta} E_k + \mathbb{E}[\varepsilon_{k+1}]$$

for Lyapunov function (40), where the error scales as

$$\mathbb{E}[\varepsilon_{k+1}] = \frac{A_k \tau_k^2}{2\mu\sigma\delta} \mathbb{E}[\|G(x_{k+1})\|^2]. \tag{85}$$

The proof of this claim follows from the proof of Proposition 5 and 12, where we simply take ∇f to be stochastic. Maximizing over this sequence gives a $O(1/\sqrt{k})$ for the first algorithm and O(1/k) for the second. This convergence rate is optimal and matches the rate of SGD. Notice, however, that the convergence rate is for the entire sequence of iterates, unlike SGD.

Stochastic dynamics. Having introduced the dynamics (6), it is clear that the following stochastic dynamics

$$dZ_t = \dot{\beta}_t (X_t dt - Z_t dt - (1/\mu)(\nabla f(X_t) dt + \sigma(X_t, t) dB_t))$$

$$dX_t = \dot{\beta}_t (Z_t - X_t) dt$$

is a natural candidate for approximating the stochastic variants of algorithms (42) and (39) in the setting where $h(x) = \frac{1}{2}||x||^2$ is Euclidean and f is μ -strongly convex.⁶ Here $B_t \in \mathbb{R}^d$ is a standard Brownian motion, and $\sigma(X_t, t) \in \mathbb{R}^{d \times d}$ is the diffusion coefficient. We assume $\mathbb{E}||\sigma(X_t, t)^{\top}\sigma(X_t, t)||^2 \le M$ for some positive constant $M \in \mathbb{R}^+$. We can use Itô's formula to calculate $d\mathcal{E}_t$ where \mathcal{E}_t is given by (5) as follows,

$$d\mathcal{E}_{t} = \frac{\partial \mathcal{E}_{t}}{\partial t}dt + \left\langle \frac{\partial \mathcal{E}_{t}}{\partial X_{t}}, dX_{t} \right\rangle + \left\langle \frac{\partial \mathcal{E}_{t}}{\partial Z_{t}}, dZ_{t} \right\rangle + \frac{\dot{\beta}_{t}^{2} e^{\beta_{t}}}{2\mu} \operatorname{tr}\left(\sigma(X_{t}, t)^{\top} \sigma(X_{t}, t)\right) dt.$$
 (86)

We compute:

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

$$\frac{\partial \mathcal{E}_t}{\partial X_t} = e^{\beta_t} \nabla f(X_t),$$

$$\frac{\partial \mathcal{E}_t}{\partial Z_t} = \mu(x^* - Z_t).$$

Plugging this into (86), we have

$$d\mathcal{E}_{t} = \dot{\beta}_{t}e^{\beta_{t}}\left(f(X_{t}) - f(x^{*}) + \frac{\mu}{2}\|x^{*} - Z_{t}\|^{2}\right)dt + \dot{\beta}_{t}e^{\beta_{t}}\left\langle\nabla f(X_{t}), Z_{t} - X_{t}\right\rangle dt + \dot{\beta}_{t}e^{\beta_{t}}\mu\left\langle x^{*} - Z_{t}, X_{t} - Z_{t}\right\rangle dt + \dot{\beta}_{t}e^{\beta_{t}}\left\langle x^{*} - Z_{t}, \nabla f(X_{t})\right\rangle dt + \dot{\beta}_{t}e^{\beta_{t}}\left\langle x^{*} - Z_{t}, \sigma(X_{t}, t)\right\rangle dt + \frac{\dot{\beta}_{t}^{2}e^{\beta_{t}}}{2\mu}\mathrm{tr}\left(\sigma(X_{t}, t)^{\top}\sigma(X_{t}, t)\right) dt + \dot{\beta}_{t}^{2}e^{\beta_{t}}\mathrm{tr}\left(\sigma(X_{t}, t)^{\top}\sigma(X_{t}, t)\right) dt,$$

where the inequality follows from the proof of proposition 3 which can be found in Appendix A.2.2. That is, we can conclude

$$\mathcal{E}_t \leq \mathcal{E}_0 - \int_0^t \dot{\beta}_s e^{\beta_s} \langle x^* - Z_s, \sigma(X_s, s) \rangle \mathrm{d}s + \int_0^t \frac{\dot{\beta}_s^2 e^{\beta_s}}{2\mu} \mathrm{tr} \left(\sigma(X_s, s)^\top \sigma(X_s, s) \right) \mathrm{d}s.$$

 $^{^6}$ Some of the following statements can be made more rigorous and motivates further study. The dynamics can also be generalized to the more general setting in the natural way, but for simplicity we take h to be Euclidean.

If we take the expectation of both sides, the middle term, $\int_0^t \dot{\beta}_s e^{\beta_s} \langle x^* - Z_s, \sigma(X_s, s) \rangle ds$, will vanish by the martingale property of the Itô integral. This allows us to conclude that

$$\mathbb{E}[f(X_t) - f(x^*)] \le \frac{\mathcal{E}_0 + \mathbb{E}\left[\int_0^t \frac{\dot{\beta}_s^2 e^{\beta_s}}{2\mu} \operatorname{tr}\left(\sigma(X_s, s)^\top \sigma(X_s, s)\right) ds\right]}{e^{\beta_t}}.$$

In particular, choosing $\beta_t = 2 \log t + \log(1/2)$, we obtain a $O(1/t^2) + O(1/t)$ convergence rate. We can compare this upper bound to the bound (85) with the identifications $\dot{\beta}_t = \tau_k$ and $e^{\beta_t} = A_k$.