Journal of Machine Learning Research 22 (2021) 1-34 Submitted 2/20; Revised 10/20; Published 4/21

A Lyapunov Analysis of Accelerated Methods in Optimization

Ashia C. Wilson ASHIA07@MIT.EDU
Department of Electrical Engineering and Computer Sciences
Massachusetts Instituted of Technology
Cambridge, MA, 02139, USA

Ben Recht BRECHT@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720-1776, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Editor: Prateek Jain

Abstract

Accelerated optimization methods, such as Nesterov's accelerated gradient method, play a significant role in optimization. Several accelerated methods are provably optimal under standard oracle models. Such optimality results are obtained using a technique known as *estimate sequences* which yields upper bounds on convergence properties. The technique of estimate sequences has long been considered difficult to understand and deploy, 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 unified analysis of many existing accelerated algorithms, introduce new algorithms, and strengthen the connection between accelerated algorithms and continuous-time dynamical systems.

Keywords: gradient descent, Nesterov acceleration, dynamical systems, Lyapunov functions, estimate sequences

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 have been previously seen to be fruitful. Such methods were first studied formally by Polyak (1964), and have been employed in many practical optimization solvers. As an example, beginning in the 1980s, momentum methods have been used in neural network research 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 (Rumelhart et al., 1986).

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 (Polyak, 1964).[1] 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 (Lessard et al., 2016).

In order to make momentum methods rigorous, a different approach was required. In celebrated work, Nesterov relied on algebraic arguments (Nesterov, 1983), and later devised a general scheme to accelerate convex optimization methods, achieving optimal running times under oracle models in convex programming (Nesterov, 2004).[2] To achieve such general applicability, Nesterov's proof techniques abandoned the physical intuition of Polyak (Nesterov, 2004); indeed, in lieu of differential equations and Lyapunov functions, Nesterov devised the method of *estimate sequences* to verify the correctness of these momentum-based accelerated methods and used it extensively to offer a library of accelerated methods (e.g., Nesterov, 2005, 2008, 2013). Researchers have struggled to understand the foundations and scope of the estimate-sequence methodology since Nesterov's early papers.

To overcome the lack of fundamental understanding of the estimate-sequence technique, several authors have proposed schemes that achieve acceleration without appealing to it (Drusvyatskiy et al., 2016; Bubeck et al., 2015; Lessard et al., 2016; Drori and Teboulle, 2014; Beck and Teboulle, 2009; Tseng, 2008). One promising general approach to the analysis of acceleration has been to analyze the continuous-time limit of accelerated methods (Su et al., 2016; Krichene et al., 2015), or to derive these limiting ODEs directly via an underlying Lagrangian (Wibisono et al., 2016), and to prove that the ODEs are stable via a Lyapunov function argument. Another recent line of attack on the discretization problem is via the use of a time-varying Hamiltonian and symplectic integrators (Betancourt et al., 2018; Muehlebach and Jordan, 2021). 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 in Euclidean spaces in the setting where the objective is strongly convex, Polyak's Heavy Ball method and Nesterov's accelerated gradient descent have the same continuous-time limit.

In this paper, we 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-time 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. Most importantly, they allow us to dispense with estimate sequences altogether, in favor of a dynamical-systems perspective that encompasses both continuous time and discrete time.

1. Indeed, when applied to positive-definite quadratic cost functions, Polyak's Heavy Ball Method is equivalent to Chebyshev's Iterative Method (Chebyshev, 1854).
2. Notably, it is easier to extract a Lyapunov argument from Nesterov's original 1983 paper.

A Dynamical View of Accelerated Methods

We begin by presenting families of dynamical systems for optimization. To do so, we review the Lagrangian framework introduced by Wibisono et al. (2016) and introduce a second Bregman Lagrangian for the strongly convex setting.

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 $\|x\| = \langle x, x\rangle^{1/2}$. We consider the 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 Gâteaux differentiable on the interior of its domain. 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. An overdot means derivative with respect to time, i.e., $\dot{X}_t = \frac{d}{dt}X_t$. We denote $x^* \in \arg\min f(x)$.

The Bregman Lagrangian

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

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

where $x \in \mathcal{X}$, $v \in \mathbb{R}^d$, and $t \in \mathbb{R}$ represent position, velocity and time, respectively (Wibisono et al., 2016). 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:

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

$$\dot{\beta}_t \;\leq\; e^{\alpha_t}. \tag{3b}$$

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 et al. (2016) showed that under the first scaling condition (3a), the Euler-Lagrange equation for the Bregman Lagrangian reduces to the following ODE:

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

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

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

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); this change of scaling is important for obtaining dynamics with convergence rate guarantees when the objective function is *strongly* convex. We summarize in the following proposition.

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

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

The proof of Proposition 1 can be found in Appendix A.1. As discussed by Wibisono, Wilson and Jordan, when $h(x) = \frac{1}{2}\|x\|^2$, the Bregman Lagrangians (2) and (5) resemble the standard Lagrangian used in physics for dissipative dynamical systems, where the kinetic energy is given by $k(v) = \frac{1}{2}\|v\|^2$ and the potential energy is the objective function, both scaled by damping parameters. More generally, our Bregman Lagrangian uses the Bregman divergence $D_h(x, x+e^{-\alpha}v)$, which is closely related the Hessian metric $\|v\|_x^2 = \langle v, \nabla^2 h(x) v \rangle$, to measure kinetic energy. We refer the reader to Wibisono et al. (2016) for in-depth discussion on the structure of the Bregman Lagrangian and its relation to the Hessian Lagrangian and Hessian Riemannian gradient flows (Alvarez et al., 2004, 2002).

In what follows, we pay close attention to the special case of the dynamics in (6) where $h(x) = \frac{1}{2}\|x\|^2$, the ideal scaling (3b) holds with equality, and the damping $\beta_t = \sqrt{\mu} t$ is linear:

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

In this setting, we can discretize the dynamics in (7) to obtain accelerated gradient descent in the setting where f is μ-strongly convex.

Related work The connection between dynamical systems, particularly gradient flows, and optimization methods has a long history (Polyak, 1964; Attouch, 1996). The main motivation of our work comes from Su et al. (2016) and Wibisono et al. (2016); both works introduce families of dynamical systems modeling accelerated methods for weakly convex functions (the latter from a variational perspective) and suggest that Lyapunov functions can be used to analyze accelerated mirror descent, but stop short of describing how the Lyapunov perspective is useful for the analysis of accelerated algorithms more broadly (e.g. for analyzing higher order methods or for in obtaining linear rates for strongly convex functions). Our work is similar to other bodies of work (Krichene et al., 2015; Attouch and Peypouquet, 2015) that utilize Lyapunov functions to deduce convergence rates for accelerated methods; however, our framework is more general, encompassing the analysis of several additional methods, including accelerated gradient descent for strongly convex functions and composite optimization methods. It also makes the connection between estimate sequences and Lyapunov functions explicit. Note, moreover, that in subsequent work our Lyapunov framework has been used to generate novel methods (Tu et al., 2017; Betancourt et al., 2018) and analyses, including methods (18), (30) and (33) in the current paper.

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 (3), we use Lyapunov's method (Lyapunov, 1992). Lyapunov's method is based on the idea of constructing a positive definite quantity $\mathcal{E} : \mathcal{X} \to \mathbb{R}$ which does not increase along the trajectories of the dynamical system $\frac{d}{dt}X_t = v(X_t)$:

$$\tfrac{d}{dt}\mathcal{E}(X_t) = \left\langle \nabla\mathcal{E}(X_t), \tfrac{d}{dt}X_t \right\rangle = \langle \nabla\mathcal{E}(X_t), v(X_t)\rangle \leq 0. \tag{8}$$

The existence of a Lyapunov function often provides the dynamical system with a qualitative description. For example, if $\mathcal{E}(X_t) = d(x, X_t)$ where $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is a differentiable function and $\mathcal{E}(X_t) = 0$ iff $x = X_t$, then the implication of (8), which we write as $\mathcal{E}(X_t) \leq \mathcal{E}(X_0)$, is that the dynamical system does not leave a bounded region defined by $d(x, X_0)$. Since we are interested in quantifying the rate at which $\dot{X}_t = v(X_t)$ finds a solution to (1), we will use *time-dependent* Lyapunov functions of the form $\mathcal{E}_t = e^{\tilde{\beta}_t} d(x^*, X_t)$, $\mathcal{E}_t = e^{\hat{\beta}_t}(f(X_t) - f(x^*))$, or combinations thereof, where $\tilde{\beta}_t, \hat{\beta}_t : \mathbb{R}^+ \to \mathbb{R}^+$ are increasing functions of time. For example, if $\mathcal{E}_t = e^{\tilde{\beta}_t}(f(X_t) - f(x^*))$ satisfies (8), integrating both sides results in the upper bound $f(X_t) - f(x^*) \leq e^{-\tilde{\beta}_t}\mathcal{E}_0$. Next, we demonstrate how this works for the Euler-Lagrange equation when the second ideal scaling (3b) holds.

Remark 2 *Assuming f is convex, h is strictly convex, and the second ideal scaling condition* (3b) *holds, Wibisono et al. (2016) show that the Euler-Lagrange equation* (4) *satisfies*

$$\tfrac{d}{dt}\Big\{ D_h(x, X_t + e^{-\alpha_t}\dot{X}_t)\Big\} \leq -\tfrac{d}{dt}\Big\{ e^{\beta_t}(f(X_t) - f(x))\Big\}, \tag{9}$$

when $x = x^*$. If the ideal scaling holds with equality, $\dot{\beta}_t = e^{\alpha_t}$, the solutions satisfy (9) *for $\forall x \in \mathcal{X}$. Thus,*

$$\mathcal{E}_t = D_h(x, X_t + e^{-\alpha_t}\dot{X}_t) + e^{\beta_t}(f(X_t) - f(x)) \tag{10}$$

is a Lyapunov function for dynamics (4)*.*

A result similar to Remark 2 holds for the second family of dynamics (5) under the additional assumption that f is μ-uniformly convex with respect to h:

$$D_f(x, y) \geq \mu D_h(x, y). \tag{11}$$

When $h(x) = \frac{1}{2}\|x\|^2$, (11) 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 in Lemma 4 of Nesterov (2008), yields a Bregman divergence that is σ-uniformly convex with respect to the p-th power of the norm ($p \geq 2$):

$$D_h(x, y) \geq \frac{\sigma}{p}\|x - y\|^p, \tag{12}$$

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 for $p \geq 2$. We are now ready to state our main proposition for the continuous-time dynamics.

Proposition 3 *Assume f is μ-uniformly convex with respect to h* (11)*, h is strictly convex, and the second ideal scaling condition* (3b) *holds. Using dynamics* (6)*, we have the following inequality:*

$$\tfrac{d}{dt}\Big\{e^{\beta_t}\mu D_h(x, X_t + e^{-\alpha_t}\dot{X}_t)\Big\} \leq -\tfrac{d}{dt}\Big\{e^{\beta_t}(f(X_t) - f(x))\Big\},$$

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

$$\mathcal{E}_t = e^{\beta_t}\left(\mu D_h(x, X_t + e^{-\alpha_t}\dot{X}_t) + f(X_t) - f(x)\right) \tag{13}$$

is a Lyapunov function for dynamics (6)*.*

The proof of this result can be found in Appendix A.2. Taking $x = x^*$ and writing the Lyapunov property $\mathcal{E}_t \leq \mathcal{E}_0$ explicitly,

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

for (10), and

$$f(X_t) - f(x^*) \leq \frac{e^{\beta_0}(\mu D_h(x^*, X_0 + e^{-\alpha_0}\dot{X}_0) + f(X_0) - f(x^*))}{e^{\beta_t}}, \tag{15}$$

for (13), allows us to infer a $O(e^{-\beta_t})$ convergence rate for the function value for both families of dynamics (4) and (6).

Remark 4 (Ideal Scaling Conditions) *While the first ideal scaling condition simplified the Euler-Lagrange equation, the second ideal scaling established the validity of our Lyapunov functions. In particular, for a given α_t, the optimal convergence rate is achieved by setting $\dot{\beta}_t = e^{\alpha_t}$, resulting in convergence rate $O(e^{-\beta_t}) = O(e^{-\int_0^t \alpha_s ds})$.*

So far, we have discussed two families of dynamics (4) and (6) and shown 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 x dynamics (4) and (6) produce algorithms which are useful for convex optimization. A similar discretization of the Lyapunov functions (10) and (13) will provide us with a tool we can use to analyze these algorithms.

Discretization Analysis

We now show how accelerated methods can be viewed as mapping these continuous-time dynamics to discrete-time algorithms.

Explicit and implicit methods. Consider a 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 generally easy to implement in practice. The implicit Euler method, on the other hand, has greater stability and favorable convergence properties but requires the expensive solution of an implicit equation (Rapp, 2017). 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 a system of two first-order differential 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 of these combinations, focusing on the family that gives rise to accelerated methods. For the remainder of the paper we make the following assumption, which restricts our analysis to dynamical systems that are simpler and converge the fastest (see Remark 4).

Assumption 1 *The second ideal scaling* (3b) *holds with equality.*

Proposition 5 (Three-point identity) *For all* $x \in \operatorname{dom} h$ *and* $y, z \in \operatorname{int}(\operatorname{dom} h)$

$$D_h(x,y) - D_h(x,z) = -\langle \nabla h(y) - \nabla h(z), x - y\rangle - D_h(y,z). \tag{16}$$

The Bregman three-point identity plays a key role in the analysis of all accelerated methods. For a fixed $x \in \mathcal{X}$, (16) can be viewed as an approximation of the identity

$$\tfrac{d}{dt} D_h(x, X_t) = -\langle \tfrac{d}{dt}\nabla h(X_t), x - X_t\rangle.$$

Methods arising from the first Euler-Lagrange equation

We begin by writing the dynamics (4) as the following system of first-order equations:

$$Z_t = X_t + \frac{e^{\beta_t}}{\frac{d}{dt}e^{\beta_t}}\dot{X}_t, \tag{17a}$$

$$\tfrac{d}{dt}\nabla h(Z_t) = -\left(\tfrac{d}{dt}e^{\beta_t}\right)\nabla f(X_t). \tag{17b}$$

As in Wibisono et al. (2016), we focus on the family of dynamical systems with the scaling $\beta_t = p\log t + \log C$, where $p > 1$ is an integer. Using the identification $t = \delta k$, we approximate $e^{\beta_t} = Ct^p$ with the discrete sequence $A_k = C\delta^p k^{(p)}$, where instead of k^p we use the rising factorial $k^{(p)} = k(k+1)\cdots(k+p-1) = \Theta(k^p)$. We also approximate the time derivative $\frac{d}{dt}e^{\beta_t} = Cpt^{p-1}$ with the difference sequence $\alpha_k := \frac{A_{k+1}-A_k}{\delta} = Cp\delta^{p-1}k^{(p-1)}$. Finally, we make the approximations $Z_t = z_k$, $X_t = x_k$, $\frac{d}{dt}\nabla h(Z_t) = \frac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}$, $\frac{d}{dt}X_t = \frac{x_{k+1}-x_k}{\delta}$, and denote $\tau_k := \frac{\alpha_k}{A_k} = \frac{p}{\delta(k+p-1)} = \Theta(\frac{p}{\delta k})$ which approximates $\frac{d}{dt}e^{\beta_t}/e^{\beta_t} = \frac{p}{t}$. With these identifications, we explore various combinations of the explicit and implicit discretization methods.

Implicit-Euler. Written as an algorithm, the implicit Euler method applied to (17b) and (17a) has the following update equations:

$$z_{k+1} = \arg\min_{z\in\mathcal{X}}\left\{A_k f(x) + \tfrac{1}{\delta\tau_k}D_h\left(z, z_k\right)\right\}, \tag{18a}$$

$$x_{k+1} = \tfrac{\delta\tau_k}{1+\delta\tau_k}z_{k+1} + \tfrac{1}{1+\delta\tau_k}x_k, \tag{18b}$$

where $x = \frac{\delta\tau_k}{1+\delta\tau_k}z + \frac{1}{1+\delta\tau_k}x_k$. We now state a convergence rate for these dynamics.

Proposition 6 *Using the discrete-time Lyapunov function,*

$$E_k = D_h(x^*, z_k) + A_k(f(x_k) - f(x^*)), \tag{19}$$

the bound $\frac{E_{k+1}-E_k}{\delta} \leq 0$ *holds for algorithm* (18) *when* f *is convex and* h *is strictly convex.*

In particular, this allows us to conclude a general $O(1/A_k)$ convergence rate for the implicit method (18). While this illustrates our methodology, we note that the update (18a) is typically as hard to solve as the original optimization problem.

Proof The proof for the implicit scheme, with the aforementioned discrete-time approximations, satisfies the variational equality,

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

$$\frac{A_{k+1}-A_k}{\delta} z_{k+1} = \frac{A_{k+1}-A_k}{\delta} x_{k+1} + A_k \frac{x_{k+1}-x_k}{\delta}. \tag{20b}$$

Using these identities, we have the following derivation:

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &\overset{(16)}{=} -\left\langle \tfrac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}, x^* - z_{k+1} \right\rangle - \tfrac{1}{\delta} D_h(z_{k+1}, z_k) \\
&\quad + \tfrac{A_{k+1}}{\delta}(f(x_{k+1}) - f(x^*)) - \tfrac{A_k}{\delta}(f(x_k) - f(x^*)) \\
&\overset{(20a)}{=} \tfrac{A_{k+1}-A_k}{\delta}\langle \nabla f(x_{k+1}), x^* - z_{k+1}\rangle - \tfrac{1}{\delta} D_h(z_{k+1}, z_k) \\
&\quad + \tfrac{A_{k+1}}{\delta}(f(x_{k+1}) - f(x^*)) - \tfrac{A_k}{\delta}(f(x_k) - f(x^*)) \\
&\overset{(20b)}{=} \tfrac{A_{k+1}-A_k}{\delta}\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + A_k \left\langle \nabla f(x_{k+1}), \tfrac{x_k - x_{k+1}}{\delta} \right\rangle \\
&\quad - \tfrac{1}{\delta} D_h(z_{k+1}, z_k) + A_k \tfrac{f(x_{k+1})-f(x_k)}{\delta} + \tfrac{A_{k+1}-A_k}{\delta}(f(x_{k+1}) - f(x^*)) \leq 0.
\end{aligned}$$

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

Family of Accelerated algorithms

Given algorithm (18) is expensive to implement, it is natural to consider whether fast and computationally efficient algorithms can be obtain using an explicit-Euler discretization of one of the sequences. In this section, we illustrate that such techniques yield fast quasi-monotone methods, and that with an additional trick, we obtain the famed family of accelerated gradient methods. In particular, we study families of algorithms which can be thought of variations of the explicit Euler scheme applied to (17a) and the implicit Euler scheme applied to (17b).[3] The first family of methods can be written as the updates,

$$x_{k+1} = \delta\tau_k z_k + (1 - \delta\tau_k) y_k \tag{21a}$$

$$\nabla h(z_{k+1}) = \nabla h(z_k) - \delta\alpha_k \nabla f(x_{k+1}), \tag{21b}$$

and the second family can be written as the updates,

$$x_{k+1} = \delta\tau_k z_k + (1 - \delta\tau_k) y_k \tag{22a}$$

$$\nabla h(z_{k+1}) = \nabla h(z_k) - \delta\alpha_k \nabla f(y_{k+1}). \tag{22b}$$

3. Here we make the identification $\tau_k = \frac{A_{k+1}-A_k}{\delta A_{k+1}} := \frac{\alpha_k}{A_{k+1}} = \frac{p}{\delta(k+p)}$.

In both algorithms, we have replaced x_k with a sequences y_k whose update we leave unspecified for now. Without this replacement, the sequences (21) and (22) are equivalent, and both algorithms are optimal for non-smooth optimization. However, the addition of the sequence y_k results in optimal convergence for smooth optimization. The update (21b) is the variational condition for the mirror descent update

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

The same is true of update (22b) where the gradient of the function is evaluated at y_{k+1}. We show that accelerated gradient descent (Nesterov, 2004, 2005), accelerated higher-order methods (Nesterov, 2008; Baes, 2009) and accelerated universal gradient methods (Nesterov, 2014) all entail choosing y_{k+1} so thatthe following discrete-time Lyapunov function,

$$E_k = D_h(x^*, z_k) + A_k(f(y_k) - f(x^*)), \tag{23}$$

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

Proposition 7 *Assume that the distance-generating function h is σ-uniformly convex with respect to the p-th power of the norm ($p \geq 2$) (12) and the objective function f is convex. Using only the updates* (21a) *and* (21b)*, and using the Lyapunov function* (23)*, we have the following bound:*

$$\tfrac{E_{k+1}-E_k}{\delta} \leq \varepsilon_{k+1}, \tag{24}$$

where the error term scales as

$$\varepsilon_{k+1} = \tfrac{p-1}{p}(\sigma/\delta)^{-\frac{1}{p-1}} \alpha_k^{\frac{p}{p-1}} \|\nabla f(x_{k+1})\|^{\frac{p}{p-1}} + \tfrac{A_{k+1}}{\delta}(f(y_{k+1}) - f(x_{k+1})). \tag{25a}$$

If we use the updates (22a) *and* (22b) *instead, the error term scales as*

$$\varepsilon_{k+1} = \tfrac{p-1}{p}(\sigma/\delta)^{-\frac{1}{p-1}} \alpha_k^{\frac{p}{p-1}} \|\nabla f(y_{k+1})\|^{\frac{p}{p-1}} + \tfrac{A_{k+1}}{\delta}\langle \nabla f(y_{k+1}), y_{k+1} - x_{k+1}\rangle. \tag{25b}$$

The error bounds in (25) are obtained using the σ-uniform convexity with respect to the p-th power of the norm (11), and no smoothness assumption on f; they also hold when full gradients of f are replaced with an element in the subgradient of f. The proof of this proposition can be found in Appendix B.1.

With the choices $A_k = C\delta^p k^{(p)}$ and $0 < C \leq 1/\sigma p^p$, it is possible to ensure $\varepsilon_{k+1} \leq 0$ simply by choosing an update y_{k+1} which satisfies

$$f(y_{k+1}) - f(x_{k+1}) \leq -\delta^{\frac{p}{p-1}} \|\nabla f(x_{k+1})\|^{\frac{p}{p-1}}, \tag{26a}$$

or

$$\langle \nabla f(y_{k+1}), y_{k+1} - x_{k+1}\rangle \leq -\delta^{\frac{p}{p-1}} \|\nabla f(y_{k+1})\|^{\frac{p}{p-1}}, \tag{26b}$$

An algorithm with these choices satisfies the convergence rate guarantee $f(y_k) - f(x^*) \leq 1/A_k = O(1/(\delta k)^p)$.

Remark 8 (Quasi-monotone method (Nesterov and Shikhman, 2015)) *The quasi-monotone subgradient method (QMS), introduced by Nesterov in 2015, is algorithm* (21) *where $y_{k+1} = x_{k+1}$, $p = 2$, and when subgradients of f are used instead of full gradients. This results in error* (25) *given by $\varepsilon_{k+1} = \delta\frac{\alpha_k^2}{2\sigma}\|\nabla f(x_{k+1})\|^2$ where $\nabla f(x) \in \partial f(x)$. Combined with the assumptions $\sup_{x\in\mathcal{X}} \|\nabla f(x)\|_2 \leq G$ and $\sup_{x,y\in\mathcal{X}} D_h(x,y) \leq R$, choosing $\alpha_k = R/\delta G\sqrt{k+1}$ results in the upper bound $f(x_k) - f(x^*) = O(1/\sqrt{k})$.*

Acceleration of Descent Methods

In convex optimization, the term "acceleration" in the phrase "accelerated methods" stems from the observation that any sequence satisfying (26a) and (26b), already yields a convergence rate $f(y_k) - f(x^*) = O(1/\delta^p k^{p-1})$, provided f has bounded level sets (i.e., $R := \sup_{x:f(x)\leq f(x_0)} \|x - x^*\| < \infty$). Adding the additional updates contained in (21) and (22) requires at most one additional gradient step, no additional assumptions on f, and results in a superior convergence rate bound of $f(y_k) - f(x^*) = O(1/(\delta k)^p)$. Thus, we can interpret algorithms that satisfy the descent conditions (26a) and (26b) (which refer to as "descent methods") as being "accelerated."

A simple demonstration of this claim follows from introducing the following function $E_k = \delta^p k^{(p)}(f(y_k) - f(x^*))$ where $k^{(p)}$ is the rising factorial $k^{(p)} = k(k+1)\cdots(k+p-1) = \Theta(k^p)$ and showing that the difference $\frac{E_{k+1}-E_k}{\delta}$ is upper bounded by a constant. Summing gives the result. Details of this argument is in Appendix B.2.

Acceleration of gradient descent (Nesterov, 2004, 2005) Accelerating gradient decent entails chooses y_{k+1} as the gradient update:

$$y_{k+1} = \arg\min_{y\in\mathcal{X}} \left\{ f(x_{k+1}) + \langle \nabla f(x_{k+1}), y - x_{k+1}\rangle + \frac{1}{2\eta}\|y - x_{k+1}\|^2 \right\}. \tag{27}$$

When ∇f is L-Lipschitz, and $0 < \eta \leq 1/L$, the gradient update satisfies conditions (26a) and (26b) with $p = 2$ and $\delta = \sqrt{1/2L}$. Indeed, plugging in the update (27) into the smoothness condition $f(y_{k+1}) \leq f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle + \frac{L}{2}\|y_{k+1} - x_{k+1}\|^2$ results in the first bound (26a). Substituting (27) into the smoothness condition $\|\nabla f(y_{k+1}) - \nabla f(x_{k+1})\| \leq L\|y_{k+1} - x_{k+1}\|$, squaring both sides, and expanding the square on the left-hand side, yields the desired second bound (26b).

Acceleration of tensor methods (Nesterov, 2008; Baes, 2009) Higher-order gradient methods choose y_{k+1} as the tensor update

$$y_{k+1} = \arg\min_{x\in\mathcal{X}} \left\{ f_{p-1}(x; y) + \frac{1}{p\eta}\|x - y\|^p \right\}, \tag{28}$$

where $f_{p-1}(x; y) = \sum_{i=0}^{p-1} \frac{1}{i!}\nabla^i f(x)(y - x)^i, p \geq 1$ is the $(p-1)$-st order Taylor expansion of f centered at $x \in \mathcal{X}$. When the p-th order gradient $\nabla^p f$ is L-Lipschitz, the gradient update (28) with step size $0 < \eta \leq \frac{\sqrt{3}(p-1)!}{2L}$ satisfies (26b) with $\delta^{\frac{p}{p-1}} = \eta^{\frac{p}{p-1}}/2^{\frac{2p-3}{p-1}}$. Details are presented in Appendix B.3.

Remark 9 (Hölder-continuous gradients Nesterov (2014)) *Suppose f has (L, ν)-Hölder-continuous gradients, where $\nu \in (0, 1)$ and $p = 2$. For $1/\tilde{L} \geq (1/2\tilde{\delta})^{\frac{1-\nu}{1+\nu}}(1/L)^{\frac{2}{1+\nu}}$, Nesterov showed that the gradient update (27) with $\eta = \tilde{L}$ satisfies $f(y_{k+1}) - f(x_{k+1}) \leq -\frac{1}{2\tilde{L}}\|\nabla f(x_{k+1})\|^2 + \frac{\tilde{\delta}}{2}$. The resulting error bound, $\varepsilon_{k+1} = \delta\frac{\alpha_k^2}{2\sigma}\|\nabla f(x_{k+1})\|^2 - \frac{A_{k+1}}{2\delta\tilde{L}}\|\nabla f(x_{k+1})\|^2 + A_{k+1}\frac{\tilde{\delta}}{2\delta}$, allows us to conclude a $O(1/k^2)$ convergence rate of the function to within $\tilde{\delta}$ using the parameter choices $A_k = \delta^2 k^{(2)}/4$ where $\delta = \sqrt{\sigma/\tilde{L}}$.*

Remark 10 (Acceleration of proximal algorithms) *Proximal algorithms, such as FISTA (Beck and Teboulle, 2009), also fit readily within our Lyapunov framework. We refer the reader to Appendix B.4 for details.*

Methods arising from the second Euler-Lagrange equation

We write the dynamics (6) as the following system of equations:

$$Z_t = X_t + \tfrac{e^{\beta_t}}{\frac{d}{dt}e^{\beta_t}}\dot{X}_t, \tag{29a}$$

$$\tfrac{d}{dt}\nabla h(Z_t) = \tfrac{\frac{d}{dt}e^{\beta_t}}{e^{\beta_t}}\Big(\nabla h(X_t) - \nabla h(Z_t) - \tfrac{1}{\mu}\nabla f(X_t)\Big). \tag{29b}$$

We focus on the family obtained when $\beta_t = \sqrt{\mu}t$. Using the identification $t = \delta k$, we approximate $e^{\beta_t} = e^{\sqrt{\mu}t}$ with the discrete sequence $A_k = (1+\sqrt{\mu}\delta)^k$. We also approximate the time derivatives $\frac{d}{dt}e^{\beta_t} = \sqrt{\mu}e^{\sqrt{\mu}t}$, $\frac{d}{dt}\nabla h(Z_t)$, $\frac{d}{dt}X_t$ and $\frac{d}{dt}e^{\beta_t}/e^{\beta_t} = \sqrt{\mu}$, with the discrete sequences $\frac{A_{k+1}-A_k}{\delta} = \sqrt{\mu}(1+\sqrt{\mu}\delta)^k$, $\frac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}$, $\frac{x_{k+1}-x_k}{\delta}$ and $\tau_k := \frac{\alpha_k}{A_k} = \sqrt{\mu}$, respectively. We begin with the following proposition.

Proposition 11 *Assume h is strictly convex. Written as an algorithm, the implicit Euler scheme applied to* (29a) *and* (29b) *results in the following updates:*

$$z_{k+1} = \arg\min_{z\in\mathcal{X}}\Big\{f(x) + \mu D_h(z,x) + \tfrac{\mu}{\delta\tau_k}D_h(z,z_k)\Big\}, \tag{30a}$$

$$x_{k+1} = \tfrac{\delta\tau_k}{1+\delta\tau_k}z_{k+1} + \tfrac{1}{1+\delta\tau_k}x_k, \tag{30b}$$

where $x = \frac{\delta\tau_k}{1+\delta\tau_k}z + \frac{1}{1+\delta\tau_k}x_k$. Using the following discrete-time Lyapunov function:

$$E_k = A_k(\mu D_h(x^*, z_k) + f(x_k) - f(x^*)), \tag{31}$$

we obtain the bound $\frac{E_{k+1}-E_k}{\delta} \le 0$ for algorithm (30) *under assumption* (11)*. This allows us to conclude a general $O(1/A_k)$ convergence rate for the implicit scheme* (30)*.*

Proof The algorithm that is obtained from the implicit discretization of the dynamics (30) satisfies the variational equalities

$$\tfrac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta} = \tau_k\Big(\nabla h(x_{k+1}) - \nabla h(z_{k+1}) - \tfrac{1}{\mu}\nabla f(x_{k+1})\Big) \tag{32a}$$

$$\tfrac{x_{k+1}-x_k}{\delta} = \tau_k(z_{k+1} - x_{k+1}), \tag{32b}$$

Using these variational equalities, we have the following argument:

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &\overset{(16)}{=} \alpha_k\mu D_h(x^*, z_{k+1}) - A_k\mu\left\langle\tfrac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}, x^* - z_{k+1}\right\rangle - \mu\tfrac{A_k}{\delta}D_h(z_{k+1}, z_k)\\
&\quad + \tfrac{A_{k+1}}{\delta}(f(x_{k+1}) - f(x^*)) - \tfrac{A_k}{\delta}(f(x_k) - f(x^*))\\
&\overset{(32a)}{=} \alpha_k\mu D_h(x^*, z_{k+1}) + A_k\tau_k\langle\nabla f(x_{k+1}), x^* - x_{k+1}\rangle + A_k\left\langle\nabla f(x_{k+1}), \tfrac{x_k - x_{k+1}}{\delta}\right\rangle\\
&\quad + A_k\tau_k\mu\left\langle\nabla h(x_{k+1}) - \nabla h(z_{k+1}), x^* - z_{k+1}\right\rangle - \mu\tfrac{A_k}{\delta}D_h(z_{k+1}, z_k)\\
&\quad + \tfrac{A_{k+1}}{\delta}(f(x_{k+1}) - f(x^*)) - \tfrac{A_k}{\delta}(f(x_k) - f(x^*))\\
&\overset{(32b)}{=} \alpha_k\mu D_h(x^*, z_{k+1}) + \alpha_k\langle\nabla f(x_{k+1}), x^* - x_{k+1}\rangle - \mu\tfrac{A_k}{\delta}D_h(z_{k+1}, z_k))\\
&\quad + A_k\left\langle\nabla f(x_{k+1}), \tfrac{x_k - x_{k+1}}{\delta}\right\rangle + \alpha_k\mu\langle\nabla h(x_{k+1}) - \nabla h(z_{k+1}), x^* - z_{k+1}\rangle\\
&\quad + \tfrac{A_k}{\delta}(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_{k+1}) - f(x^*))\\
&\le -\alpha_k\mu D_h(x_{k+1}, z_{k+1}) - \mu\tfrac{A_k}{\delta}D_h(z_{k+1}, z_k) \le 0.
\end{aligned}$$

The inequality uses the Bregman three-point identity (16) and μ-uniform convexity of f with respect to h (11). ■

Remark 12 (Quasi-monotone method) *A variation of the implicit Euler scheme applied to* (29b) *and* (29b),

$$x_{k+1} = \tfrac{\delta\tau_k}{1+\delta\tau_k} z_k + \tfrac{1}{1+\delta\tau_k} x_k \tag{33a}$$

$$\tfrac{\nabla h(z_{k+1}) - \nabla h(z_k)}{\delta} = \tau_k \left(\nabla h(x_{k+1}) - \nabla h(z_{k+1}) - \tfrac{1}{\mu} \nabla f(x_{k+1}) \right), \tag{33b}$$

results in what can be regarded as the quasi-monotone method for strongly convex functions. When $h(x) = \frac{1}{2}\|x\|^2$, we can write (33b) *as a mirror-descent update:*

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

where $\tilde{z}_{k+1} = \frac{z_k + \delta\tau_k x_{k+1}}{1+\delta\tau_k}$. More generally, the update (33b) *involves optimizing a linear approximation to the function regularized by a weighted combination of Bregman divergences. Assuming f is differentiable and μ-strongly convex with respect to h and that h is σ-strongly convex we obtain the error bound*

$$\tfrac{E_{k+1} - E_k}{\delta} \leq \delta \tfrac{A_k \tau_k^2}{2\mu\sigma} \|\nabla f(x_{k+1})\|^2, \tag{34}$$

for algorithm (33) *using Lyapunov function* (31). *The choice $A_k = \frac{\delta^2 k^{(2)}}{2}$ so that $\tau_k := \frac{A_{k+1} - A_k}{\delta A_k} = \frac{\alpha_k}{A_k} = \frac{2}{\delta k}$ results in the upper bound $f(x_k) - f(x^*) = O(1/k)$. This bound matches the subgradient oracle lower bound for strongly convex Lipschitz functions. Details of this result are in Appendix C.3.*

Accelerated gradient descent (Nesterov, 2004)

We study a family of algorithms which can be thought of as variations of the implicit Euler scheme applied to (29a) and the explicit Euler scheme applied to (29b):

$$x_k = \tfrac{\delta\tau_k}{1+\delta\tau_k} z_k + \tfrac{1}{1+\delta\tau_k} y_k \tag{35a}$$

$$\tfrac{\nabla h(z_{k+1}) - \nabla h(z_k)}{\delta} = \tau_k \left(\nabla h(x_k) - \nabla h(z_k) - \tfrac{1}{\mu} \nabla f(x_k) \right), \tag{35b}$$

where y_{k+1} satisfies (26a) with $p = 2$. We make the identification $A_k = (1 - \sqrt{\mu}\delta)^{-k}$ which is a first-order Taylor approximation of $e^{\beta_t} = e^{\sqrt{\mu}t}$ using the identification $t = \delta k$. Denote $\alpha_k := \frac{A_{k+1} - A_k}{\delta} = \sqrt{\mu}(1 - \sqrt{\mu}\delta)^{-(k+1)}$ and $\tau_k := \frac{\alpha_k}{A_{k+1}} = \sqrt{\mu}$ which approximates $\frac{d}{dt} e^{\beta_t} / e^{\beta_t} = \sqrt{\mu}$ exactly. To analyze the general algorithm (35), we use the following Lyapunov function:

$$E_k = A_k(\mu D_h(x^*, z_k) + f(y_k) - f(x^*)). \tag{36}$$

We begin with the following proposition, which provides an error bound for algorithm (35).

Proposition 13 *Assume the objective function f is μ-uniformly convex with respect to h (11) and h is σ-strongly convex. In addition, assume f is L-smooth. Using the sequences* (35a) *and* (35b)*, we obtain the bound $\frac{E_{k+1}-E_k}{\delta} \leq \varepsilon_{k+1}$, where the error term has the following form:*

$$\varepsilon_{k+1} = \tfrac{A_{k+1}}{\delta}(f(y_{k+1}) - f(x_k)) + \tfrac{A_{k+1}}{\delta}\left(\tfrac{\delta\tau_k L}{2} - \tfrac{\sigma\mu}{2\delta\tau_k}\right)\|x_k - y_k\|^2 - \tfrac{A_{k+1}\mu\sigma}{2\delta}\|x_k - y_k\|^2 + \tfrac{\alpha_k}{\delta}\langle\nabla f(x_k), y_k - x_k\rangle + \tfrac{A_{k+1}\mu}{2\sigma\delta}\|\delta\tau_k(\nabla h(x_k) - \nabla h(z_k) - \tfrac{1}{\mu}\nabla f(x_k))\|^2. \quad (37a)$$

When $h(x) = \frac{1}{2}\|x\|^2$, the error simplifies to the following form

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

Given the original update (26a) has a $O(e^{-\mu k})$ convergence rate, we consider (35) an "accelerated" algorithm. We present a proof of Proposition 13 in Appendix C.1. The result for accelerated gradient descent, which satisfies (26a) with $p = 2$, can be summed up in the following corollary:

Corollary 14 *Using the gradient step* (27) *for the sequence y_{k+1} results in an error which scales as*

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

when $h(x) = \frac{1}{2}\|x\|^2$.

Given $\tau_k = \sqrt{\mu}$, the parameter choice $\delta \leq \sqrt{1/L}$ so that $\delta\tau_k \leq 1/\sqrt{\kappa}$, where $\kappa = \mu/L$ is the condition number, ensures the error is nonpositive. With this choice, we obtain a linear $O(e^{-\sqrt{\mu}\delta k}) = O(e^{-k/\sqrt{\kappa}})$ upper bound. In particular, when $h(x) = \frac{1}{2}\|x\|^2$ and $\delta\tau_k = 1/\sqrt{\kappa}$, the algorithm (35) can be reduced to the familiar two-sequence accelerated gradient descent algorithm of Nesterov, where we set $\gamma_0 = \mu$ (Nesterov, 2004, (p. 78-79)). The upper bound for (35) matches the oracle lower-bound for gradient-based methods designed for smooth strongly convex functions.

Remark 15 (Hölder-continuous gradients) *Assume f is μ-strongly convex and has (L, ν)-Hölder-continuous gradients, where $\nu \in (0, 1]$. For $1/\tilde{L} \geq (1/2\tilde{\delta})^{\frac{1-\nu}{1+\nu}}(1/L)^{\frac{2}{1+\nu}}$, the gradient update $y_{k+1} = x_k - \frac{1}{\tilde{L}}\nabla f(x_k)$ results in an error for algorithm* (35) *that scales as*

$$\varepsilon_{k+1} = \tfrac{A_{k+1}}{\delta}\left(\tfrac{(\delta\tau_k)^2}{2\mu} - \tfrac{1}{2\tilde{L}}\right)\|\nabla f(x_k)\|^2 + \tfrac{A_{k+1}}{\delta}\left(\tfrac{\delta\tau_k\tilde{L}}{2} - \tfrac{\mu}{2\delta\tau_k}\right)\|x_k - y_k\|^2 + \left(\tfrac{\alpha_k}{2} + \tfrac{A_{k+1}}{2\delta}\right)\tilde{\delta}.$$

With the parameter choices $A_k = (1 - \sqrt{\mu}\delta)^{-k}$, $\alpha_k = \sqrt{\mu}(1 - \mu\delta)^{-(k+1)}$, $\tau_k = \sqrt{\mu}$ and $\delta = (1/\tilde{L})^{1/2}$, we obtain the upper bound $f(y_k) - f(x^*) \leq \tilde{\varepsilon}_{k+1} := A_k^{-1}E_0 + \tilde{\delta}_1$, where $\tilde{\delta}_1 = \frac{\tilde{\delta}}{2}((\mu/\tilde{L})^{1/2} + 1)$ determines the threshold of convergence. In particular, choosing a sufficiently small δ_1 requires $L \ll \tilde{L}$ which negatively affects the linear convergence rate.

Equivalence to Estimate Sequences of f

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

Estimate sequences of a function $f(x)$

We provide a brief review of the technique of estimate sequences (Nesterov, 2004). We begin with the following definition.

Definition 16 *(Nesterov, 2004, 2.2.1) A pair of sequences, $\{\phi_k(x)\}_{k=1}^{\infty}$ and $\{A_k\}_{k=0}^{\infty}$, for $A_k \geq 1$, is called an* estimate sequence *of a 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

$$\phi_k(x) \leq \left(1 - A_k^{-1}\right) f(x) + A_k^{-1}\phi_0(x). \tag{38}$$

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

Lemma 17 *(Nesterov, 2004, 2.2.1) If for some sequence $\{x_k\}_{k\geq 0}$ we have*

$$f(x_k) \leq \phi_k^* \equiv \min_{x\in\mathcal{X}} \phi_k(x), \tag{39}$$

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

Proof Observe that

$$f(x_k) \overset{(39)}{\leq} \min_{x\in\mathcal{X}} \phi_k(x) \overset{(38)}{\leq} \min_{x\in\mathcal{X}} \left(\left(1 - A_k^{-1}\right) f(x) + A_k^{-1}\phi_0(x)\right) \leq \left(1 - A_k^{-1}\right) f(x^*) + A_k^{-1}\phi_0(x^*).$$

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 (2009) as an extension of Nesterov's Lemma 2.2.2 (Nesterov, 2004), provides guidance for constructing estimate sequences. This construction is used in Nesterov (2004, 2005, 2008); Baes (2009); Nesterov and Shikhman (2015); Nesterov (2015), and is, to the best of our knowledge, the only existing formal way to construct an estimate sequence. We will see below that this particular class of estimate sequences can be transformed into our Lyapunov arguments with a few algebraic manipulations (and vice versa).

Proposition 18 *(Baes, 2009, 2.2) Let $\phi_0 : \mathcal{X} \to \mathbb{R}$ be a convex function such that $\min_{x\in\mathcal{X}} \phi_0(x) \geq f^*$. Suppose also that we have a sequence $\{f_k\}_{k\geq 0}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ that underestimates f:*

$$f_k(x) \leq f(x) \quad \text{for all } x \in \mathcal{X} \text{ and all } k \geq 0. \tag{40}$$

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

$$\phi_{k+1}(x) := (1 - \delta\tau_k)\phi_k(x) + \delta\tau_k f_k(x) = A_{k+1}^{-1}\left(A_0\phi_0(x) + \textstyle\sum_{i=0}^{k} \delta a_i f_i(x)\right), \tag{41}$$

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

From (39) and (41), we observe that the following invariant:

$$A_{k+1}f(x_{k+1}) \leq \min_x A_{k+1}\phi_{k+1}(x) = \min_x \sum_{i=0}^{k} \delta\alpha_i f_i(x) + A_0\phi_0(x), \tag{42}$$

is maintained. In Nesterov and Shikhman (2015) and Nesterov (2015), this technique was extended to incorporate an error term $\{\tilde{\varepsilon}_k\}_{k=1}^{\infty}$,

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

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

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

An argument analogous to that of Lemma 17 holds:

$$\begin{aligned}A_{k+1}f(x_{k+1}) &\leq \sum_{i=0}^{k}\delta\alpha_i f_i(x^*) + A_0(\phi_0(x^*) - \tilde{\varepsilon}_0) + \tilde{\varepsilon}_{k+1} \\ &\overset{(40)}{\leq} \sum_{i=0}^{k}\delta\alpha_i f(x^*) + A_0\phi_0(x^*) + \tilde{\varepsilon}_{k+1} = A_{k+1}f(x^*) + A_0\phi_0(x^*) + \tilde{\varepsilon}_{k+1}.\end{aligned}$$

Rearranging, we obtain the desired bound,

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

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	$f_i(x)$	$\phi_k(x)$	$\tilde{\varepsilon}_{k+1}$
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

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 $\{\varepsilon_k, A_k, \phi_k(x)\}$ in order to check a priori that the invariants hold. This aspect of the estimate-sequence technique has made it hard to discern its structure and scope.

Equivalence to Lyapunov arguments

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 D. The discrete-time estimate sequence (41) for accelerated gradient descent can be written:

$$\begin{aligned}\phi_{k+1}(x) &:= f(x_{k+1}) + A_{k+1}^{-1} D_h(x, z_{k+1}) \\ &\overset{(41)}{=} (1 - \delta\tau_k)\phi_k(x) + \delta\tau_k f_k(x) \\ &\overset{\text{Table 1}}{=} \Big(1 - A_{k+1}^{-1}\delta\alpha_k\Big)\Big(f(x_k) + A_k^{-1} D_h(x, z_k)\Big) + A_{k+1}^{-1}\delta\alpha_k f_k(x).\end{aligned}$$

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

$$\begin{aligned}A_{k+1}f(x_{k+1}) + D_h(x, z_{k+1}) &= (A_{k+1} - \delta\alpha_k)\Big(f(x_k) + A_k^{-1} D_h(x, z_k)\Big) + \delta\alpha_k f_k(x) \\ &= A_k\Big(f(x_k) + A_k^{-1} D_h(x, z_k)\Big) + (A_{k+1} - A_k) f_k(x) \\ &\leq A_k f(x_k) + D_h(x, z_k) + (A_{k+1} - A_k) f(x).\end{aligned}$$

The last inequality follows from definition (40). Rearranging, we obtain the inequality $E_{k+1} \leq E_k$ for our Lyapunov function (23) with $x = x^*$. Going the other direction, from our Lyapunov analysis we can derive the following bound:

$$\begin{aligned}E_k &\leq E_0 \\ A_k(f(x_k) - f(x)) + D_h(x, z_k) &\leq A_0(f(x_0) - f(x)) + D_h(x, z_0) \\ A_k\Big(f(x_k) - A_k^{-1} D_h(x, z_k)\Big) &\leq (A_k - A_0) f(x) + A_0\Big(f(x_0) + A_0^{-1} D_h(x^*, z_0)\Big) \\ A_k\phi_k(x) &\leq (A_k - A_0) f(x) + A_0\phi_0(x). \qquad (43)\end{aligned}$$

Rearranging, with $x = x^*$ we obtain the estimate sequence (38), with $A_0 = 1$:

$$\phi_k(x) \leq \Big(1 - A_k^{-1}A_0\Big) f(x) + A_k^{-1}A_0\phi_0(x) = \Big(1 - A_k^{-1}\Big) f(x) + A_k^{-1}\phi_0(x).$$

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

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

Comparing this to (43), matching terms allows us to extract the continuous-time estimate sequence $\{\phi_t(x), e^{\beta_t}\}$, where $\phi_t(x) = f(X_t) + e^{-\beta_t} D_h(x, Z_t)$.

Discussion

The main contributions in this paper are twofold: We have presented a unified analysis of a wide variety of algorithms using Lyapunov functions—equations (23) and (36)—and we have demonstrated the equivalence between Lyapunov arguments and estimate sequences

of f, under the formalization of the latter due to Baes (2009). 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 (Butcher, 2000). 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 Betancourt et al. (2018) 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 (Su et al., 2016). Therefore, many authors (O'Donoghue and Candès, 2015; Su et al., 2016; Krichene et al., 2015) 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 et al. (2016) and Krichene et al. (2015) 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 (2014), Kim and Fessler (2016), Taylor et al. (2016), and Lessard et al. (2016) 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

Felipe Alvarez, Hédy Attouch, Jérôme Bolte, and Patrick Redont. A second-order gradient-like dissipative dynamical system with Hessian-driven damping: Application to optimization and mechanics. *Journal de Mathématiques Pures et Appliquées*, 81(8):747–779, 2002.

Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. Hessian Riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, 43(2):477–501, 2004.

Hédy Attouch. Viscosity solutions of minimization problems. *SIAM Journal on Optimization*, 6(3):769–806, 1996.

Hédy Attouch and Juan Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually $o(k^{-2})$. *ArXiv e-prints arXiv:1510.08740v2*, November 2015.

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

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, March 2009. ISSN 1936-4954.

Michael Betancourt, Michael I. Jordan, and Ashia Wilson. On symplectic optimization. Arxiv preprint arXiv1802.03653, March 2018.

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

John C. Butcher. Numerical methods for ordinary differential equations in the 20th century. *Journal of Computational and Applied Mathematics*, 125(1–2):1–29, 2000.

Pafnuty L. Chebyshev. Théorie des mécanismes connus sous le nom de parallélogrammes. *Mémoires Présentés à l'Académie Impériale des Sciences de St-Pétersbourg*, VII(539-568), 1854.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

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

Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1):81–107, 2016. ISSN 1436-4646.

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

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

Alexander M. Lyapunov. General problem of the stability of motion. *International Journal of Control*, 55:531–773, 1992.

Michael Muehlebach and Michael I. Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *Journal of Machine Learning Research*, 22:1–50, 2021.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

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

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

Yurii Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008. ISSN 0025-5610.

Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, Aug 2013.

Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, pages 1–24, 2014. ISSN 0025-5610.

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

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

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

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

Bastian E. Rapp. Numerical methods for solving differential equations. In Bastian E. Rapp, editor, *Microfluidics: Modelling, Mechanics and Mathematics*, Micro and Nano Technologies, pages 549 – 593. Elsevier, Oxford, 2017.

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

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17:1–43, 2016.

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

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *SIAM Journal on Optimization*, 12:724–739, 2008.

Stephen Tu, Shivaram Venkataraman, Ashia C. Wilson, Alex Gittens, Michael I. Jordan, and Benjamin Recht. Breaking locality accelerates block Gauss-Seidel. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 1549–1557, 2017.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 133:E7351–E7358, 2016.

Appendix A. Dynamics

A.1 Proof of Proposition 1: Computing the Euler-Lagrange equation

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:

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

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

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

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

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

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

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

as well as the derivative of the momentum,

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

For Lagrangian (2), not only do the terms involving $\tfrac{d}{dt}\nabla h(X)$ 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 Proof of Proposition 3: Deriving the Lyapunov function

We compute the time derivative of the Lyapunov function (13):

$$\begin{aligned}
\tfrac{d}{dt}\mathcal{E}_t &= e^{\beta_t}\Big(\dot{\beta}_t(f(X_t) - f(x^*)) + \langle \nabla f(X_t), \dot{X}_t\rangle - \mu\left\langle \nabla h\left(Z_t\right), x^* - Z_t\right\rangle + \mu\dot{\beta}_t D_h\left(x^*, Z_t\right)\Big) \\
&\overset{(29b)}{=} e^{\beta_t}\Big(\dot{\beta}_t(f(X_t) - f(x^*)) + \langle \nabla f(X_t), \dot{X}_t\rangle + \dot{\beta}_t\mu\left\langle \nabla h(Z_t) - \nabla h(X_t), x^* - Z_t\right\rangle \\
&\quad + \dot{\beta}_t\langle \nabla f(X_t), x^* - Z_t\rangle + \mu\dot{\beta}_t D_h\left(x^*, Z_t\right)\Big) \\
&\overset{(16)}{=} e^{\beta}\Big(\dot{\beta}_t\left(f(X_t) - f(x^*) + \langle \nabla f(X_t), x - X_t\rangle + \mu D_h(x^*, X_t)\right) - \mu\dot{\beta}_t D_h\left(Z_t, X_t\right)\Big) \leq 0
\end{aligned}$$

The second equality uses (29b) and third equality uses the Bregman three-point identity with $x = x$, $y = X_t$ and $z = Z_t$ as well as (29a). We conclude the desired result from the μ-uniform convexity of f with respect to h and the nonnegativity of the Bregman divergence.

Appendix B. Algorithms derived from dynamics (4)

We show the initial error bound has an appealing form.

B.1 Proof of Proposition 7: Initial bounds (25a) and (25b)

We begin with algorithm (21) using Lyapunov function (23):

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &\overset{(16)}{=} -\left\langle \tfrac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}, x^* - z_{k+1}\right\rangle - \tfrac{1}{\delta} D_h(z_{k+1}, z_k) + \tfrac{A_{k+1}}{\delta}(f(y_{k+1}) - f(x^*)) \\
&\quad - \tfrac{A_k}{\delta}(f(y_k) - f(x^*)) \\
&\overset{(21b)}{=} \alpha_k \langle \nabla f(x_{k+1}), x^* - z_{k+1}\rangle - \tfrac{1}{\delta} D_h(z_{k+1}, z_k) + \alpha_k(f(x_{k+1}) - f(x^*)) \\
&\quad + A_k \tfrac{f(x_{k+1})-f(y_k)}{\delta} + A_{k+1}\tfrac{f(y_{k+1})-f(x_{k+1})}{\delta} \\
&\le \alpha_k \langle \nabla f(x_{k+1}), x^* - z_k\rangle + \alpha_k \langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \tfrac{\sigma}{\delta p}\|z_{k+1} - z_k\|^p \\
&\quad + \alpha_k(f(x_{k+1}) - f(x^*)) + A_k \tfrac{f(x_{k+1})-f(y_k)}{\delta} + A_{k+1}\tfrac{f(y_{k+1})-f(x_{k+1})}{\delta} \\
&\le \alpha_k \langle \nabla f(x_{k+1}), x^* - z_k\rangle + A_k \tfrac{f(x_{k+1})-f(y_k)}{\delta} + \alpha_k(f(x_{k+1}) - f(x^*)) \\
&\quad + \tfrac{p-1}{p}(\sigma/\delta)^{-\frac{1}{p-1}} \alpha_k^{\frac{p}{p-1}} \|\nabla f(x_{k+1})\|^{\frac{p}{p-1}} + A_{k+1}\tfrac{f(y_{k+1})-f(x_{k+1})}{\delta}.
\end{aligned}$$

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 (25a), it simply remains to use our second update (21a):

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &\le \alpha_k \langle \nabla f(x_{k+1}), x^* - z_k\rangle + A_k \tfrac{f(x_{k+1})-f(y_k)}{\delta} + \alpha_k(f(x_{k+1}) - f(x^*)) \\
&\quad + \tfrac{p-1}{p}(\sigma/\delta)^{-\frac{1}{p-1}} \alpha_k^{\frac{p}{p-1}} \|\nabla f(x_{k+1})\|^{\frac{p}{p-1}} + A_{k+1}\tfrac{f(y_{k+1})-f(x_{k+1})}{\delta} \\
&\le \alpha_k \langle \nabla f(x_{k+1}), x^* - y_k\rangle + \tfrac{A_{k+1}}{\delta}\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle + A_k \tfrac{f(x_{k+1})-f(y_k)}{\delta} \\
&\quad + \alpha_k(f(x_{k+1}) - f(x^*)) + \varepsilon_{k+1} \\
&= \alpha_k(f(x_{k+1}) - f(x^*) + \langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle) \\
&\quad + \tfrac{A_k}{\delta}(f(x_{k+1}) - f(y_k) + \langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle) + \varepsilon_{k+1}.
\end{aligned}$$

From here, we can conclude $\frac{E_{k+1}-E_k}{\delta} \le \varepsilon_{k+1}$ using the convexity of f. Using update (26a), we have $\frac{E_{k+1}-E_k}{\delta} \le \left((\delta/\sigma)^{\frac{1}{p-1}}(Cp\delta^{p-1}(k+1)^{(p-1)})^{\frac{p}{p-1}} - C\delta^{\frac{1}{p-1}}\delta^p(k+1)^{(p)}\right)\|\nabla f(x_k)\|_*^{\frac{p}{p-1}}$. Given $((k+1)^{(p-1)})^{\frac{p}{p-1}}/(k+1)^{(p)} \le 1$, it suffices that $C \le 1/\sigma p^p$ to ensure $\frac{E_{k+1}-E_k}{\delta} \le 0$. Summing the Lyapunov function gives the convergence rate $f(y_k) - f(x^*) = O(1/A_k) = O(1/(\delta k)^p)$.

We now show the bound (25b) for algorithm (22) using a similar argument:

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &\overset{(16)}{=} -\left\langle \tfrac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}, x^* - z_{k+1}\right\rangle - \tfrac{1}{\delta} D_h(z_{k+1}, z_k) + \tfrac{A_{k+1}}{\delta}(f(y_{k+1}) - f(x^*)) \\
&\quad - \tfrac{A_k}{\delta}(f(y_k) - f(x^*)) \\
&\overset{(21b)}{=} \alpha_k \langle \nabla f(y_{k+1}), x^* - z_{k+1}\rangle - \tfrac{1}{\delta} D_h(z_{k+1}, z_k) + \alpha_k(f(y_{k+1}) - f(x^*)) + A_k \tfrac{f(y_{k+1})-f(y_k)}{\delta} \\
&\le \alpha_k \langle \nabla f(y_{k+1}), x^* - z_k\rangle + \alpha_k \langle \nabla f(y_{k+1}), z_k - z_{k+1}\rangle - \tfrac{\sigma}{\delta p}\|z_{k+1} - z_k\|^p \\
&\quad + \alpha_k(f(y_{k+1}) - f(x^*)) + A_k \tfrac{f(y_{k+1})-f(y_k)}{\delta} \\
&\le \alpha_k \langle \nabla f(y_{k+1}), x^* - z_k\rangle + A_k \tfrac{f(y_{k+1})-f(y_k)}{\delta} + \alpha_k(f(y_{k+1}) - f(x^*)) \\
&\quad - \tfrac{A_{k+1}}{\delta}\langle \nabla f(y_{k+1}), y_{k+1} - x_{k+1}\rangle + \varepsilon_{k+1}.
\end{aligned}$$

The first inequality follows from the uniform convexity of h and the second uses the Fenchel-Young inequality and definition (25b). Using the second update (22a), we obtain our initial error bound:

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &\le \alpha_k \langle \nabla f(y_{k+1}), x^* - y_k\rangle + A_k \tfrac{f(y_{k+1})-f(y_k)}{\delta} + \alpha_k(f(y_{k+1}) - f(x^*)) \\
&\quad + \tfrac{A_{k+1}}{\delta}\langle \nabla f(y_{k+1}), y_k - x_{k+1}\rangle - \tfrac{A_{k+1}}{\delta}\langle \nabla f(y_{k+1}), y_{k+1} - x_{k+1}\rangle + \varepsilon_{k+1} \\
&= \alpha_k(f(y_{k+1}) - f(x^*) + \langle \nabla f(y_{k+1}), x^* - y_{k+1}\rangle) \\
&\quad + \tfrac{A_k}{\delta}(f(y_{k+1}) - f(y_k) + \langle \nabla f(y_{k+1}), y_k - y_{k+1}\rangle) + \varepsilon_{k+1}.
\end{aligned}$$

From here, we can conclude $\frac{E_{k+1}-E_k}{\delta} \le \varepsilon_{k+1}$ using the convexity of f. Using (26b), we have $\frac{E_{k+1}-E_k}{\delta} \le -\delta^{\frac{1}{p-1}} C (k+1)^{(p)} \|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}} + (\delta/\sigma)^{\frac{1}{p-1}} (Cp(k+1)^{(p-1)})^{\frac{p}{p-1}} \|\nabla f(y_{k+1})\|_*^{\frac{p}{p-1}}$. For $\frac{E_{k+1}-E_k}{\delta} \le 0$ it suffices that $C \le 1/\sigma p^p$. Summing the Lyapunov function gives the convergence rate $f(y_k) - f(x^*) = O(1/A_k) = O(1/(\delta k)^p)$.

B.2 Descent Methods: Convergence of algorithms satisfying (26a) and (26b)

We show that any algorithm that satisfies (26b) obtains a $O(1/\delta^p k^{p-1})$ convergence upper bound using the function $E_k = \delta^p k^{(p)}(f(x_k) - f(x^*))$. To do so, we compute,

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &= p\delta^{p-1} k^{(p-1)}(f(x_k) - f(x^*)) + \delta^p (k+1)^{(p)} \tfrac{f(x_{k+1})-f(x_k)}{\delta} \\
&\le p\delta^{p-1} k^{(p-1)} \langle \nabla f(x_k), x_k - x^*\rangle + \delta^p (k+1)^{(p)} \tfrac{f(x_{k+1})-f(x_k)}{\delta} \\
&\le p\delta^{p-1} k^{(p-1)} \langle \nabla f(x_k), x_k - x^*\rangle - \delta^p (k+1)^{(p)} \|\nabla f(x_k)\|^{\frac{p}{p-1}} \\
&\le (p-1)^p \|x_k - x^*\|^p \le (p-1)^p R^p.
\end{aligned}$$

The first inequality follows from convexity and the second from (26a). The last inequality follows from Young's inequality, $\langle s, u\rangle + \frac{1}{p}\|u\|^p \ge -\frac{p-1}{p}\|s\|_*^{\frac{p}{p-1}}$, with $s = \delta^{p-1}\nabla f(x_k)[(k+1)^{(p)}]^{\frac{p-1}{p}}$ and $u = (p-1)\{k^{(p-1)}/[(k+1)^{(p)}]^{\frac{p-1}{p}}\}(x_k - x^*)$. The descent condition implies $\|x_k - x^*\| \le R$. Summing over k shows $f(x_k) - f(x^*) = O(1/\delta^p k^{p-1})$. For (26b), we

similarly compute

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} &= p\delta^{p-1}k^{(p-1)}(f(x_{k+1})-f(x^*)) + \delta^p k^{(p)}\tfrac{f(x_{k+1})-f(x_k)}{\delta}\\ &\le p\delta^{p-1}k^{(p-1)}\langle\nabla f(x_{k+1}), x_{k+1}-x^*\rangle + \delta^p k^{(p)}\tfrac{f(x_{k+1})-f(x_k)}{\delta}\\ &\le p\delta^{p-1}k^{(p-1)}\langle\nabla f(x_{k+1}), x_{k+1}-x^*\rangle - \delta^p k^{(p)}\|\nabla f(x_{k+1})\|^{\frac{p}{p-1}}\\ &\le 2(p-1)^p\|x_{k+1}-x^*\|^p \le 2(p-1)^p R^p.\end{aligned}$$

The first inequality follows from convexity and the second from (26a). The last inequality follows from Young's inequality, $\langle s,u\rangle + \frac{1}{p}\|u\|^p \ge -\frac{p-1}{p}\|s\|_*^{\frac{p}{p-1}}$, with $s = \delta^{p-1}\nabla f(x_k)[(k+1)^{(p)}]^{\frac{p-1}{p}}$ and $u = (p-1)\{k^{(p-1)}/[(k+1)^{(p)}]^{\frac{p-1}{p}}\}(x_k - x^*)$. The descent condition implies $\|x_k - x^*\| \le R$. Summing over k gives the bound. An analogous argument holds for (26b).

B.3 Higher-order Tensor Method (28) satisfies (26b)

Let $\tilde{p} = p - 1 + \nu$. The optimality condition for (28) is

$$\textstyle\sum_{i=1}^{p-1}\frac{1}{(i-1)!}\nabla^i f(x_k)\,(x_{k+1}-x_k)^{i-1} + \frac{1}{\eta}\|x_{k+1}-x_k\|^{\tilde{p}-2}\,(x_{k+1}-x_k) = 0. \tag{46}$$

Since $\nabla^{p-1}f$ is L-Lipschitz, we have the following error bound on the $(p-2)$th order Taylor expansion of ∇f:

$$\left\|\nabla f(x_{k+1}) - \textstyle\sum_{i=1}^{p-1}\frac{1}{(i-1)!}\nabla^i f(x_k)\,(x_{k+1}-x_k)^{i-1}\right\|_* \le \tfrac{L}{(p-2)!}\|x_{k+1}-x_k\|^{p-2+\nu}. \tag{47}$$

Substituting (46) into (47) and writing $r_k = \|x_{k+1}-x_k\|$, we obtain

$$\left\|\nabla f(x_{k+1}) + \tfrac{r_k^{\tilde{p}-2}}{\eta}\,(x_{k+1}-x_k)\right\|_* \le \tfrac{L}{(p-2)!}r_k^{\tilde{p}-1}. \tag{48}$$

Squaring both sides, expanding, and rearranging the terms, we get the inequality

$$\langle\nabla f(x_{k+1}), x_k - x_{k+1}\rangle \ge \tfrac{\eta}{2r_k^{\tilde{p}-2}}\|\nabla f(x_{k+1})\|_*^2 + \tfrac{\eta r_k^{\tilde{p}}}{2}\left(\tfrac{1}{\eta^2} - \tfrac{L^2}{(p-2)!^2}\right). \tag{49}$$

If $p = 2$, then the first term in (49) already implies the desired bound. Now assume $p \ge 3$. The right-hand side of (49) is of the form $A/r^{\tilde{p}-2} + Br^{\tilde{p}}$, which is a convex function of $r > 0$ and is 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

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

Substituting the values $A = \frac{\eta}{2}\|\nabla f(x_{k+1})\|_*^2$ and $B = \frac{\eta}{2}(\frac{1}{\eta^2} - \frac{L^2}{(p-2)!^2})$ from (49), we obtain

$$\langle\nabla f(x_{k+1}), x_k - x_{k+1}\rangle \ge \tfrac{\eta}{2}\left(\tfrac{1}{\eta^2} - \tfrac{L^2}{(p-2)!^2}\right)^{\frac{\tilde{p}-2}{2\tilde{p}-2}}\|\nabla f(x_{k+1})\|_*^{\frac{\tilde{p}}{\tilde{p}-1}}.$$

Finally, using the inequality $f(x_k) - f(x_{k+1}) \ge \langle\nabla f(x_{k+1}), x_k - x_{k+1}\rangle$ by the convexity of f yields the progress bound

$$f(x_{k+1}) - f(x_k) \le -\tfrac{\eta^{\frac{1}{\tilde{p}-1}}}{2}\left(1 - \tfrac{(L\eta)^2}{(p-2)!^2}\right)^{\frac{\tilde{p}-2}{2\tilde{p}-2}}\|\nabla f(x_{k+1})\|_*^{\frac{\tilde{p}}{\tilde{p}-1}} \le -\tfrac{\eta^{\frac{1}{\tilde{p}-1}}}{2^{\frac{2\tilde{p}-3}{\tilde{p}-1}}}\|\nabla f(x_{k+1})\|_*^{\frac{\tilde{p}}{\tilde{p}-1}},$$

where the least inequality uses the fact that $\eta \le \frac{\sqrt{3}(p-2)!}{2L}$.

B.4 Details of Remark 10: Lyapunov analysis of FISTA (convex case)

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

$$f(x) = \varphi(x) + \psi(x), \tag{50}$$

where φ is L-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 19 *Define $f = \varphi + \psi$ and assume φ and ψ are convex. Under the ideal scaling condition* (3b)*, Lyapunov function* (10) *can be used to show that solutions to the system*

$$Z_t = X_t + e^{-\alpha_t}\dot{X}_t \tag{51a}$$
$$\tfrac{d}{dt}\nabla h(Z_t) = -e^{\alpha_t+\beta_t}(\nabla\varphi(X_t) + \nabla\psi(Z_t)) \tag{51b}$$

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

Proof We begin by plugging in the dynamics (51a) and (51b) into Lyapunov function (10).

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

The second line plugs in the dynamics (51a) and (51b). The third line follows from choosing $e^{\alpha_t} = \dot{\beta}_t$. The fourth and fifth lines follow from convexity. The sixth line plugs in the dynamics (51b) and the last line follows from application of the chain rule. ■

Algorithm. We now discretize the dynamics (51) when the ideal scaling (3b) holds with equality. We use 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 identify $e^{\beta_t} = (1/4)t^2$ with the discrete sequence $A_k = \frac{\delta^2 k^{(2)}}{4}$. We also approximate $\frac{d}{dt}e^{\beta_t} = t/2$ and $\frac{d}{dt}e^{\beta_t}/e^{\beta_t} = \frac{2}{t}$ with the discrete sequences $\alpha_k := \frac{A_{k+1}-A_k}{\delta} = \frac{\delta(k+1)}{2}$, and $\tau_k := \frac{A_{k+1}-A_k}{\delta A_{k+1}} = \frac{2}{\delta(k+2)}$, respectively. We apply the implicit-Euler scheme to (51b) and the explicit-Euler scheme for (51a). Doing so, we obtain a proximal mirror descent update,

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

and the sequence (21a), respectively. We write the variational equality as

$$x_{k+1} = \delta\tau_k z_k + (1-\delta\tau_k)y_k \tag{52a}$$
$$\nabla h(z_{k+1}) - \nabla h(z_k) = -\delta\alpha_k \nabla\varphi(x_{k+1}) - \delta\alpha_k \nabla\psi(z_{k+1}), \tag{52b}$$

where y_{k+1} is chosen to simplify the error bound. We summarize how the initial bound scales for algorithm (52) in the following proposition.

Proposition 20 *Assume h is strongly convex, φ is L-smooth and ψ is simple but not necessarily smooth. Using the Lyapunov function* (23)*, the following initial bound:*

$$\tfrac{E_{k+1}-E_k}{\delta} \leq \varepsilon_{k+1},$$

can be shown for algorithm (52)*, where the error scales as*

$$\begin{aligned}\varepsilon_{k+1} = &\tfrac{A_{k+1}L}{2\delta}\|\delta\tau_k z_k + (1-\delta\tau_k)y_k - y_{k+1}\|^2 - \tfrac{\sigma}{2\delta}\|z_{k+1}-z_k\|^2 \\ &+ \langle \nabla\varphi(x_{k+1}), \tfrac{A_{k+1}}{\delta}y_{k+1} - \tfrac{A_k}{\delta}y_k - \alpha_k z_{k+1}\rangle + \tfrac{A_{k+1}}{\delta}\psi(y_{k+1}) - \tfrac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}).\end{aligned}$$

The update

$$y_{k+1} = \delta\tau_k z_{k+1} + (1-\delta\tau_k)y_k \tag{52c}$$

provides the upper bound $\frac{A_{k+1}}{\delta}\psi(y_{k+1}) - \frac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}) \leq 0$ using the convexity of ψ, and eliminates the inner product. Furthermore, combined with update (52a), the norm in the error is simplified, so that the error scales as $\varepsilon_k = \left(\frac{A_{k+1}\delta^2\tau_k^2 L}{2\delta} - \frac{\sigma}{2\delta}\right)\|z_{k+1}-z_k\|^2$. Using the same choice $A_k = \delta^2 k^{(2)}/4$, $\delta = \sqrt{\sigma/L}$ results in an $O(1/(\delta k)^2)$ convergence rate.

Proof The proof of Proposition 20 begins with our Lyapunov bound

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} &\overset{(16)}{=} -\left\langle \tfrac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}, x^* - z_{k+1}\right\rangle - \tfrac{1}{\delta}D_h(z_{k+1},z_k) + \tfrac{A_{k+1}}{\delta}f(y_{k+1}) - f(x^*)) \\ &\quad - \tfrac{A_k}{\delta}(f(y_k) - f(x^*)) \\ &\overset{(52b)}{=} \alpha_k\langle\nabla\varphi(x_{k+1}), x^* - z_{k+1}\rangle - \tfrac{1}{\delta}D_h(z_{k+1},z_k) + \tfrac{A_{k+1}}{\delta}(f(y_{k+1}) - f(x^*)) \\ &\quad - \tfrac{A_k}{\delta}(f(y_k) - f(x^*)) + \alpha_k\langle\nabla\psi(z_{k+1}), x^* - z_{k+1}\rangle.\end{aligned}$$

Using the convexity of ψ we obtain the upper bound

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} \leq &\alpha_k\langle\nabla\varphi(x_{k+1}), x^* - z_{k+1}\rangle - \tfrac{1}{\delta}D_h(z_{k+1},z_k) + \tfrac{A_{k+1}}{\delta}(\varphi(y_{k+1}) - \varphi(x^*)) \\ &- \tfrac{A_k}{\delta}(\varphi(y_k) - \varphi(x^*)) + \tfrac{A_{k+1}}{\delta}\psi(y_{k+1}) - \tfrac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}).\end{aligned}$$

It remains to use the smoothness and convexity of φ as well as the σ-strong convexity of h:

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} \leq &\alpha_k\langle\nabla\varphi(x_{k+1}), x^* - x_{k+1}\rangle - \tfrac{\sigma}{2\delta}\|z_{k+1}-z_k\|^2 + \alpha_k(\varphi(x_{k+1}) - \varphi(x^*)) \\ &+ \langle\nabla\varphi(x_{k+1}), \tfrac{A_{k+1}}{\delta}y_{k+1} - \tfrac{A_k}{\delta}y_k - \alpha_k z_{k+1}\rangle + \tfrac{A_{k+1}L}{2\delta}\|x_{k+1}-y_{k+1}\|^2.\end{aligned}$$

Using the convexity of φ and update (52a) we obtain the desired bound on the error.

■

Appendix C. Algorithms derived from dynamics (6)

C.1 Proof of Proposition 13: Initial bound (37)

We begin by expanding the Lyapunov bound, followed by using the strong convexity of h:

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} &= \tfrac{A_{k+1}}{\delta}(f(y_{k+1})-f(x^*)) - \tfrac{A_k}{\delta}(f(y_k)-f(x^*)) \\ &\quad - \mu A_{k+1}\left\langle \tfrac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}, x^*-z_{k+1}\right\rangle - \tfrac{A_{k+1}}{\delta}\mu D_h(z_{k+1},z_k) + \alpha_k \mu D_h(x^*,z_k) \\ &\leq \tfrac{A_{k+1}}{\delta}(f(y_{k+1})-f(x_k)) + \boldsymbol{\tfrac{A_{k+1}}{\delta}(f(x_k)-f(y_k))} + \alpha_k(f(y_k)-f(x^*)+\mu D_h(x^*,z_k)) \\ &\quad - \mu A_{k+1}\left\langle \tfrac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}, x^*-z_k\right\rangle + \tfrac{A_{k+1}\mu}{2\sigma\delta}\|\nabla h(z_{k+1})-\nabla h(z_k)\|^2.\end{aligned}$$

Using the $\mu\sigma$-strong convexity of f w.r.t the norm on the bolded term, we obtain:

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} &\leq \tfrac{A_{k+1}}{\delta}(f(y_{k+1})-f(x_k)+\langle\nabla f(x_k),x_k-y_k\rangle - \tfrac{\sigma\mu}{2}\|x_k-y_k\|^2 + \tfrac{\mu}{2\sigma}\|\nabla h(z_{k+1})-\nabla h(z_k)\|^2) \\ &\quad + \alpha_k(f(y_k)-f(x^*)+\mu D_h(x^*,z_k)) - \tfrac{\mu A_{k+1}}{\delta}\langle \nabla h(z_{k+1})-\nabla h(z_k), x^*-z_k\rangle \\ &\overset{(35b)}{=} \tfrac{A_{k+1}}{\delta}(f(y_{k+1})-f(x_k)+\langle\nabla f(x_k),x_k-y_k\rangle - \tfrac{\sigma\mu}{2}\|x_k-y_k\|^2 + \tfrac{\mu}{2\sigma}\|\nabla h(z_{k+1})-\nabla h(z_k)\|^2) \\ &\quad + \alpha_k(f(y_k)-f(x^*)+\mu D_h(x^*,z_k)+\langle\nabla f(x_k),x^*-z_k\rangle - \mu\langle\nabla h(x_k)-\nabla h(z_k),x^*-z_k\rangle) \\ &\overset{(35a)}{=} \tfrac{A_{k+1}}{\delta}(f(y_{k+1})-f(x_k)-\tfrac{\sigma\mu}{2}\|x_k-y_k\|^2 + \tfrac{\mu}{2\sigma}\|\nabla h(z_{k+1})-\nabla h(z_k)\|^2) + \alpha_k(f(y_k)-f(x^*)) \\ &\quad + \alpha_k(\boldsymbol{\langle\nabla f(x_k), x^*-x_k\rangle} - \mu\langle\nabla h(x_k)-\nabla h(z_k), x^*-z_k\rangle + \mu D_h(x^*,z_k)).\end{aligned}$$

Using the μ-strong convexity of f with respect to h (11) on the bolded term in the last line, we have:

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} &\leq \tfrac{A_{k+1}}{\delta}(f(y_{k+1})-f(x_k)) + \alpha_k(f(y_k)-f(x_k)-\boldsymbol{\mu D_h(x^*,x_k)}) - \tfrac{A_{k+1}\sigma\mu}{2\delta}\|x_k-y_k\|^2 \\ &\quad - \boldsymbol{\mu\alpha_k}(\langle\boldsymbol{\nabla h(x_k)-\nabla h(z_k), x^*-z_k}\rangle + \boldsymbol{D_h(x^*,z_k)}) + \tfrac{A_{k+1}\mu}{2\sigma\delta}\|\nabla h(z_{k+1})-\nabla h(z_k)\|^2 \\ &\overset{(16)}{=} \tfrac{A_{k+1}}{\delta}(f(y_{k+1})-f(x_k)) + \boldsymbol{\alpha_k(f(y_k)-f(x_k))} - \tfrac{A_{k+1}\sigma\mu}{2\delta}\|x_k-y_k\|^2 \\ &\quad + \tfrac{A_{k+1}\sigma\mu}{2\delta}\|\nabla h(z_{k+1})-\nabla h(z_k)\|^2 - \alpha_k\mu D_h(x_k,z_k) \qquad (53) \\ &\leq \tfrac{A_{k+1}}{\delta}(f(y_{k+1})-f(x_k)) + \alpha_k\langle\nabla f(x_k).y_k-x_k\rangle - \tfrac{A_{k+1}\sigma\mu}{2\delta}\|x_k-y_k\|^2 \\ &\quad + \tfrac{A_{k+1}\mu}{2\sigma\delta}\|\delta\tau_k(\nabla h(x_k)-\nabla h(z_k)-\tfrac{1}{\mu}\nabla f(x_k))\|^2 - \tfrac{A_{k+1}}{\delta}\left(\tfrac{\sigma\mu}{2\delta\tau_k}-\tfrac{\delta\tau_k L}{2}\right)\|x_k-y_k\|^2.\end{aligned}$$

The second line applies the Bregman three-point identity to the bolded terms in the line before. The last line, our final error bound, is obtained from applying the μ-strong convexity of f on term in bold on the previous line.

C.2 Details of Remark 15: Hölder-continuous gradients bound

To analyze the setting where f has Hölder continuous gradients and $h(x) = \frac{1}{2}\|x\|^2$ we proceed from (53) using the following bound:

$$f(y)-f(x) \leq \langle\nabla f(x), y-x\rangle + \tfrac{\tilde{L}}{2}\|x-y\|^2 + \tfrac{\tilde{\delta}}{2}, \qquad (54)$$

for $x, y \in \mathcal{X}$ where $1/\tilde{L} \geq (1/2\tilde{\delta})^{\frac{1-\nu}{1+\nu}}(1/L)^{\frac{2}{1+\nu}}$ (Nesterov, 2014, Lemma 1). We have

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} &\leq \tfrac{A_{k+1}}{\delta}(f(y_{k+1}) - f(x_k)) + \alpha_k\langle \nabla f(x_k).y_k - x_k\rangle - \tfrac{A_{k+1}\mu}{2\delta}\|x_k - y_k\|^2 \\ &\quad + \tfrac{A_{k+1}\mu}{2\delta}\|\delta\tau_k(x_k - z_k - \tfrac{1}{\mu}\nabla f(x_k))\|^2 - \tfrac{A_{k+1}}{\delta}\left(\tfrac{\mu}{2\delta\tau_k} - \tfrac{\delta\tau_k\tilde{L}}{2}\right)\|x_k - y_k\|^2 + \alpha_k\tfrac{\tilde{\delta}}{2} \\ &= \tfrac{A_{k+1}}{\delta}(f(y_{k+1}) - f(x_k) + \tfrac{(\delta\tau_k)^2}{2\mu}\|\nabla f(x_k)\|^2) - \tfrac{A_{k+1}}{\delta}\left(\tfrac{\mu}{2\delta\tau_k} - \tfrac{\delta\tau_k\tilde{L}}{2}\right)\|x_k - y_k\|^2 + \alpha_k\tfrac{\tilde{\delta}}{2}.\end{aligned}$$

The last line follows from expanding the square and plugging in update (35a).

C.3 Details of Remark 12: Quasi-monotone gradient method

We show the convergence bound for the quasi-monotone method (33). We have:

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} &= A_k\tfrac{f(x_{k+1})-f(x_k)}{\delta} + \alpha_k(f(x_{k+1}) - f(x^*)) + \alpha_k\mu D_h(x^*, z_{k+1}) \\ &\quad - A_k\mu\left\langle \tfrac{\nabla h(z_{k+1})-\nabla h(z_k)}{\delta}, x^* - z_{k+1}\right\rangle - \tfrac{A_k}{\delta}\mu D_h(z_{k+1}, z_k) \\ &\overset{(33b)}{=} A_k\tfrac{f(x_{k+1})-f(x_k)}{\delta} + \alpha_k(f(x_{k+1}) - f(x^*) + \mu D_h(x^*, z_{k+1}) + \langle \nabla f(x_{k+1}), x^* - z_k\rangle) \\ &\quad + \alpha_k(\mu\langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle + \langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \tfrac{A_k\mu}{\alpha_k\delta}D_h(z_{k+1}, z_k)) \\ &\leq A_k\tfrac{f(x_{k+1})-f(x_k)}{\delta} + \alpha_k(f(x_{k+1}) - f(x^*) + \mu D_h(x^*, z_{k+1}) + \langle \nabla f(x_{k+1}), x^* - z_k\rangle) \\ &\quad + \mu\langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle) + \tfrac{\alpha_k^2\delta}{2\mu\sigma A_k}\|\nabla f(x_{k+1})\|^2 \\ &\overset{(33a)}{=} A_k\tfrac{f(x_{k+1})-f(x_k)}{\delta} + \alpha_k(f(x_{k+1}) - f(x^*) + \langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + \mu D_h(x^*, z_{k+1})) \\ &\quad + A_k\left\langle \nabla f(x_{k+1}), \tfrac{x_k - x_{k+1}}{\delta}\right\rangle + \alpha_k\mu\langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle \\ &\quad + \tfrac{\alpha_k^2\delta}{2\mu\sigma A_k}\|\nabla f(x_{k+1})\|^2 \\ &\leq \alpha_k\mu(\langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle + D_h(x^*, z_{k+1}) - D_h(x^*, x_{k+1})) \\ &\quad + \tfrac{\alpha_k^2\delta}{2\mu\sigma A_k}\|\nabla f(x_{k+1})\|^2.\end{aligned}$$

The first inequality comes from the strong convexity of h and Hölder's inequality. The second inequality follows from the uniform convexity of f with respect to h and convexity of f. The final error bound follows from using the Bregman three-point identity (16) and nonnegativity of the Bregman divergence on the last line.

C.4 Details of Remark 10: Lyapunov analysis of FISTA (strongly convex case)

We study the problem of minimizing the composite objective $f = \varphi + \psi$ in the setting where φ is L-smooth and μ-strongly convex and ψ is simple but not smooth:

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* (13) *can be used to show that solutions to the system*

$$Z_t = X_t + e^{-\alpha_t}\dot{X}_t \tag{55a}$$

$$\frac{d}{dt}\nabla h(Z_t) = \dot{\beta}_t\nabla h(X_t) - \dot{\beta}_t\nabla h(Z_t) - \frac{e^{\alpha_t}}{\mu}(\nabla\varphi(X_t) + \nabla\psi(Z_t)), \tag{55b}$$

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

Proof Let $e^{\alpha_t} = \dot{\beta}_t$. Using (13) we compute

$$\begin{aligned}
\tfrac{d}{dt}\mathcal{E}_t &= e^{\beta_t}\Big(\dot{\beta}_t(f(X_t) - f(x^*)) + \langle \nabla f(X_t), \dot{X}_t\rangle + \mu\dot{\beta}_t D_h(x^*, Z_t) - \mu\left\langle \tfrac{d}{dt}\nabla h(Z_t), x^* - Z_t\right\rangle\Big) \\
&\overset{(55)}{=} e^{\beta_t}\Big(\dot{\beta}_t(f(X_t) - f(x^*)) + \dot{\beta}_t\langle \nabla f(X_t), Z_t - X_t\rangle + \mu\dot{\beta}_t D_h(x^*, Z_t)\Big) \\
&\quad + \dot{\beta}_t e^{\beta_t}\Big(\langle \nabla\varphi(X_t), x^* - Z_t\rangle - \mu\langle \nabla h(Z_t) - \nabla h(X_t), x^* - Z_t\rangle + \langle \nabla\psi(Z_t), x^* - Z_t\rangle\Big) \\
&\le \dot{\beta}_t e^{\beta_t}\Big((\psi(Z_t) - \psi(x^*)) + (\varphi(X_t) - \varphi(x^*)) + \langle \nabla\varphi(X_t), Z_t - X_t\rangle + \mu D_h(x^*, Z_t) \\
&\quad - \mu\langle \nabla h(Z_t) - \nabla h(X_t), x^* - Z_t\rangle + \langle \nabla\varphi(X_t), x^* - Z_t\rangle + \langle \nabla\psi(Z_t), x^* - Z_t\rangle\Big) \\
&\le \dot{\beta}_t e^{\beta_t}\Big(\varphi(X_t) - \varphi(x^*) + \langle \nabla\varphi(X_t), x^* - X_t\rangle + \mu(D_h(x^*, Z_t) - \langle \nabla h(Z_t) - \nabla h(X_t), x^* - Z_t\rangle)\Big) \\
&\le -\mu\dot{\beta}_t e^{\beta_t} D_h(Z_t, X_t).
\end{aligned}$$

The second line comes from plugging in the dynamics (55b) and (55a). The third and fourth lines use the convexity of ψ and the fifth line uses the strong convexity of φ and the Bregman three-point identity with $x = x$, $y = X_t$ and $z = Z_t$. ■

Algorithm Assume $h(x) = \frac{1}{2}\|x\|^2$ and the ideal scaling (3b) holds with equality $\dot{\beta}_t = e^{\alpha_t}$. To discretize the dynamics (55b), we split the vector field 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 $v_1(x, z, t)$. We also approximate $e^{\beta_t} = e^{-\mu t}$ with a first-order Taylor approximation $A_k = (1 - \sqrt{\mu}\delta)^{-k}$ so that $\tau_k := \frac{A_{k+1} - A_k}{\delta A_{k+1}} = \sqrt{\mu}$ yields $\frac{d}{dt}e^{\beta_t}/e^{\beta_t} = \sqrt{\mu}$. This results in the proximal update

$$z_{k+1} = \arg\min_z \left\{\psi(z) + \langle \nabla\varphi(x_k), z\rangle + \tfrac{\mu}{2\delta\tau_k}\|z - (1 - \delta\tau_k)z_k - \delta\tau_k x_k\|^2\right\}. \tag{56}$$

In full, we can write the algorithm as

$$x_k = \tfrac{\delta\tau_k}{1+\delta\tau_k} z_k + \tfrac{1}{1+\delta\tau_k} y_k \tag{57a}$$

$$z_{k+1} - z_k = \delta\tau_k\left(x_k - z_k - \tfrac{1}{\mu}\nabla\varphi(x_k) - \tfrac{1}{\mu}\nabla\psi(z_{k+1})\right), \tag{57b}$$

where y_{k+1} is chosen to simplify the error bound. We summarize how the initial bound changes with this modified update in the following proposition.

Proposition 22 *Assume $h(x) = \frac{1}{2}\|x\|^2$, φ is strongly convex, φ is L-smooth, and ψ is convex and simple. Using the Lyapunov function* (36)*, we have following bound $\frac{E_{k+1} - E_k}{\delta} \le \varepsilon_{k+1}$, for algorithm* (57)*, where the error scales as*

$$\begin{aligned}
\varepsilon_{k+1} = \tfrac{A_{k+1}L}{2\delta}\|y_{k+1} - x_k\|^2 - \tfrac{A_{k+1}\mu}{2\delta}\|z_{k+1} - z_k - \delta\tau_k(x_k - z_k)\|^2 + \Big(\tfrac{\alpha_k L}{2} - \tfrac{\alpha_k\mu}{2(\delta^2\tau_k^2)}\Big)\|y_k - x_k\|^2 \\
+ \langle \nabla\varphi(x_k), \tfrac{A_{k+1}}{\delta}y_{k+1} - \tfrac{A_k}{\delta}y_k - \alpha_k z_{k+1}\rangle + \tfrac{A_{k+1}}{\delta}\psi(y_{k+1}) - \tfrac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}).
\end{aligned}$$

The same update,

$$y_{k+1} = \delta\tau_k z_{k+1} + (1 - \delta\tau_k)y_k, \tag{57c}$$

provides the upper bound $\frac{A_{k+1}}{\delta}\psi(y_{k+1}) - \frac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}) \leq 0$ using the convexity of ψ, and eliminates the inner product. Furthermore, the identity $x_k - y_{k+1} \overset{(57c)}{=} x_k - y_k - \delta\tau_k(z_{k+1} - y_k) \overset{(57a)}{=} \delta\tau_k(z_k - z_{k+1} + y_k - x_k) \overset{(57a)}{=} \delta\tau_k(\delta\tau_k(x_k - z_k) - (z_k - z_{k+1}))$. allows us to simplify the norm in the error so that we conclude a new error that scales as

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

Given $\tau_k = \sqrt{\mu}$, choosing $\delta = \sqrt{1/L}$ results in an results in a $O(e^{-\sqrt{\mu}\delta k}) = O(e^{-k/\sqrt{\kappa}})$ convergence rate which matches the lower bound for the class of L-smooth and μ-strongly convex functions.

Proof We begin with the Bregman three point identity:

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &\overset{(16)}{=} -A_{k+1}\mu\left\langle \tfrac{z_{k+1}-z_k}{\delta}, x^* - z_{k+1}\right\rangle - \tfrac{A_{k+1}\mu}{2\delta}\|z_{k+1} - z_k\|^2 + \tfrac{\alpha_k\mu}{2}\|x^* - z_k\|^2 \\
&\quad + A_{k+1}\tfrac{f(y_{k+1})-f(y_k)}{\delta} + \alpha_k(f(y_k) - f(x^*)) \\
&\overset{(57b)}{=} -\alpha_k\mu\langle x_k - z_k, x^* - z_{k+1}\rangle - \tfrac{A_{k+1}\mu}{2\delta}\|z_{k+1} - z_k\|^2 + \tfrac{\alpha_k\mu}{2}\|x^* - z_k\|^2 \\
&\quad + A_{k+1}\tfrac{\varphi(y_{k+1})-\varphi(y_k)}{\delta} + \boldsymbol{\alpha_k\langle\nabla\varphi(x_k), x^* - x_k\rangle} + \alpha_k\langle\nabla\varphi(x_k), x_k - z_{k+1}\rangle \\
&\quad + \alpha_k(\varphi(y_k) - f(x^*)) + \tfrac{A_{k+1}}{\delta}\psi(y_{k+1}) - \tfrac{A_k}{\delta}\psi(y_k) + \boldsymbol{\alpha_k\langle\nabla\psi(z_{k+1}), x^* - z_{k+1}\rangle} \\
&\leq -\alpha_k\mu\langle x_k - z_k, x^* - z_{k+1}\rangle - \tfrac{A_{k+1}\mu}{2\delta}\|z_{k+1} - z_k\|^2 + \tfrac{\alpha_k\mu}{2}\|x^* - z_k\|^2 \\
&\quad + A_{k+1}\tfrac{\varphi(y_{k+1})-\varphi(y_k)}{\delta} + \alpha_k(\varphi(x_k) - \varphi(x^*) - \tfrac{\mu}{2}\|x^* - x_k\|^2 + \langle\nabla\varphi(x_k), x_k - z_{k+1}\rangle) \\
&\quad + \alpha_k(\varphi(y_k) - \varphi(x^*)) + \tfrac{A_{k+1}}{\delta}\psi(y_{k+1}) - \tfrac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}).
\end{aligned}$$

The inequality follows from strong convexity of φ and the convexity of ψ which was used to upper bound the bolded inner products on the second line. Using the L-smoothness and μ-strong convexity of φ, we obtain the upper bound:

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &\leq -\alpha_k(\boldsymbol{\mu\langle x_k - z_k, x^* - z_{k+1}\rangle + \tfrac{\mu}{2}\|x^* - z_k\|^2} + \tfrac{L}{2}\|y_k - x_k\|^2) - \tfrac{A_{k+1}\mu}{2\delta}\|z_{k+1} - z_k\|^2 \\
&\quad + \tfrac{A_{k+1}L}{2\delta}\|y_{k+1} - x_k\|^2 - \tfrac{A_{k+1}\mu}{2\delta}\|y_k - x_k\|^2 + \langle\nabla\varphi(x_k), \tfrac{A_{k+1}}{\delta}y_{k+1} - \tfrac{A_k}{\delta}y_k - \alpha_k z_{k+1}\rangle \\
&\quad - \boldsymbol{\tfrac{\alpha_k\mu}{2}\|x^* - x_k\|^2} + \tfrac{A_{k+1}}{\delta}\psi(y_{k+1}) - \tfrac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}) \\
&\overset{(16)}{=} \alpha_k\left(\tfrac{L}{2}\|y_k - x_k\|^2 - \mu\langle x_k - z_k, z_k - z_{k+1}\rangle\right) + \tfrac{A_{k+1}}{\delta}\left(\tfrac{L}{2}\|y_{k+1} - x_k\|^2 - \tfrac{\mu}{2}\|z_{k+1} - z_k\|^2\right) \\
&\quad - \boldsymbol{\tfrac{A_{k+1}\mu}{2\delta}\|y_k - x_k\|^2} + \langle\nabla\varphi(x_k), \tfrac{A_{k+1}}{\delta}y_{k+1} - \tfrac{A_k}{\delta}y_k - \alpha_k z_{k+1}\rangle - \boldsymbol{\tfrac{\alpha_k\mu}{2}\|x_k - z_k\|^2} \\
&\quad + \tfrac{A_{k+1}}{\delta}\psi(y_{k+1}) - \tfrac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}).
\end{aligned}$$

The second line follows from applying the Bregman three point identity to the terms on the first line in bold. Next, we apply the coupling identity (57a) to the terms on the last line

in bold:

$$\begin{aligned}
\tfrac{E_{k+1}-E_k}{\delta} &\overset{(57a)}{\leq} \boldsymbol{\alpha_k \mu \langle z_k - x_k, z_k - z_{k+1}\rangle - \frac{A_{k+1}\mu}{2\delta}\|z_{k+1}-z_k\|^2} + \tfrac{A_{k+1}L}{2\delta}\|y_{k+1}-x_k\|^2 \\
&\quad + \tfrac{\alpha_k L}{2}\|y_k - x_k\|^2 - \boldsymbol{\frac{A_{k+1}\mu}{2\delta}\|\delta\tau_k(x_k - z_k)\|^2} + \langle \nabla\varphi(x_k), \tfrac{A_{k+1}}{\delta}y_{k+1} - \tfrac{A_k}{\delta}y_k - \alpha_k z_{k+1}\rangle \\
&\quad - \tfrac{\alpha_k \mu}{2(\delta^2\tau_k^2)}\|x_k - y_k\|^2 + \tfrac{A_{k+1}}{\delta}\psi(y_{k+1}) - \tfrac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}) \\
&\overset{(16)}{=} \tfrac{A_{k+1}L}{2\delta}\|y_{k+1}-x_k\|^2 - \tfrac{A_{k+1}\mu}{2\delta}\|z_{k+1}-z_k-\delta\tau_k(x_k-z_k)\|^2 + \left(\tfrac{\alpha_k L}{2} - \tfrac{\alpha_k\mu}{2(\delta^2\tau_k^2)}\right)\|y_k - x_k\|^2 \\
&\quad + \langle \nabla\varphi(x_k), \tfrac{A_{k+1}}{\delta}y_{k+1} - \tfrac{A_k}{\delta}y_k - \alpha_k z_{k+1}\rangle + \tfrac{A_{k+1}}{\delta}\psi(y_{k+1}) - \tfrac{A_k}{\delta}\psi(y_k) - \alpha_k\psi(z_{k+1}).
\end{aligned}$$

The final line follows from applying the Bregman three point identity (16) to the terms on the first line in bold.

Appendix D. Estimate Sequences

D.1 Lyapunov and estimate sequence frameworks for quasi-monotone method

The discrete-time estimate sequence (41) for the quasi-monotone subgradient method can be written:

$$\begin{aligned}
\phi_{k+1}(x) - A_{k+1}^{-1}\tilde{\varepsilon}_{k+1} &:= f(x_{k+1}) + A_{k+1}^{-1}D_h(x, z_{k+1}) - A_{k+1}^{-1}\tilde{\varepsilon}_{k+1} \\
&\overset{(41)}{=} (1-\delta\tau_k)\left(\phi_k(x) - A_k^{-1}\tilde{\varepsilon}_k\right) + \delta\tau_k f_k(x) \\
&= \left(1 - \frac{\delta\alpha_k}{A_{k+1}}\right)\left(f(x_k) + \frac{1}{A_k}D_h(x, z_k) - \frac{\tilde{\varepsilon}_k}{A_k}\right) + \frac{\delta\alpha_k}{A_{k+1}}f_k(x).
\end{aligned}$$

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

$$\begin{aligned}
A_{k+1}f(x_{k+1}) + D_h(x, z_{k+1}) - \tilde{\varepsilon}_{k+1} &= (A_{k+1} - \delta\alpha_k)(f(x_k) + A_k^{-1}D_h(x, z_k) - A_k^{-1}\tilde{\varepsilon}_k) \\
&\quad - (A_{k+1} - \delta\alpha_k)A_k^{-1}\tilde{\varepsilon}_k + \delta\alpha_k f_k(x) \\
&= A_k\left(f(x_k) + A_k^{-1}D_h(x, z_k) - A_k^{-1}\tilde{\varepsilon}_k\right) + \delta\alpha_k f_k(x) \\
&\overset{(40)}{\leq} A_k f(x_k) + D_h(x, z_k) - \tilde{\varepsilon}_k + \delta\alpha_k f(x).
\end{aligned}$$

Rearranging, we obtain our Lyapunov argument $E_{k+1} \leq E_k + \varepsilon_{k+1}$ for (23):

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

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

$$E_k \leq E_0 + \tilde{\varepsilon}_k \tag{58}$$

$$\begin{aligned}
A_k(f(x_k) - f(x)) + D_h(x, z_k) &\leq A_0(f(x_0) - f(x)) + D_h(x, z_0) + \tilde{\varepsilon}_k \\
A_k\left(f(x_k) - \frac{1}{A_k}D_h(x, z_k)\right) &\leq (A_k - A_0)f(x) + A_0\left(f(x_0) + \frac{1}{A_0}D_h(x^*, z_0)\right) + \tilde{\varepsilon}_k \\
A_k\phi_k(x) &\leq (A_k - A_0)f(x) + A_0\phi_0(x) + \tilde{\varepsilon}_k.
\end{aligned} \tag{59}$$

Rearranging, we obtain our estimate sequence (38) ($A_0 = 1$) with an additional error term:

$$\phi_k(x) \leq \left(1 - \frac{A_0}{A_k}\right)f(x) + \frac{A_0}{A_k}\phi_0(x) + \frac{\tilde{\varepsilon}_k}{A_k} = \left(1 - \frac{1}{A_k}\right)f(x) + \frac{1}{A_k}\phi_0(x) + \frac{\tilde{\varepsilon}_k}{A_k}. \tag{60a}$$

D.2 Lyapunov and estimate sequence frameworks for accelerated gradient descent

The discrete-time estimate sequence (41) for accelerated gradient descent can be written:

$$\phi_{k+1}(x) := f(x_{k+1}) + \frac{\mu}{2}\|x - z_{k+1}\|^2 \overset{(41)}{=} (1-\delta\tau_k)\phi_k(x) + \delta\tau_k f_k(x) \overset{(40)}{\leq} (1-\delta\tau_k)\phi_k(x) + \delta\tau_k f(x).$$

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

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

Going the other direction, we have,

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

Summing over the right-hand side, we obtain the estimate sequence (38):

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

Since the Lyapunov function property allows us to write

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

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.

Appendix E. Additional Methods

E.1 Frank-Wolfe algorithms

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

$$Z_t = X_t + \dot{\beta}_t^{-1}\dot{X}_t, \tag{61a}$$

$$0 \leq \langle \nabla f(X_t), x - Z_t\rangle, \quad \forall x \in \mathcal{X}. \tag{61b}$$

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 (61b) holds on the domain $\mathcal{X}$. We summarize the usefulness of dynamics (61) in the following proposition.

Proposition 23 *Assume f is convex and the ideal scaling* (3b) *holds. The following function:*

$$\mathcal{E}_t = e^{\beta_t}(f(X_t) - f(x^*)), \tag{62}$$

is a Lyapunov function for the dynamics which satisfies (61)*. We can therefore conclude an $O(e^{-\beta_t})$ convergence rate of dynamics* (61) *to the minimizer of the function.*

Before proving Proposition 23, we first analyze Frank-Wolfe algorithms, which are discretizations of the dynamics (61). Applying the backward-Euler scheme to (61a) and (61b), we use the same approximation $\frac{d}{dt}X_t = \frac{x_{k+1}-x_k}{\delta}$, and identify $e^{\beta_t} = Ct^2$ where $C = \frac{1}{2}$ with the discrete sequence $A_k = \frac{\delta^2 k^{(2)}}{2}$ so that $\alpha_k := \frac{A_{k+1}-A_k}{\delta} = \delta(k+1)$ and $\tau_k = \frac{A_{k+1}-A_k}{\delta A_{k+1}} = \frac{2}{\delta(k+2)}$ roughly approximates $\frac{d}{dt}e^{\beta_t} = t$ and $\frac{d}{dt}e^{\beta_t}/e^{\beta_t} = \dot{\beta}_t = \frac{2}{t}$, respectively. We obtain the following algorithm:

$$z_k = \arg\min_{z\in\mathcal{X}} \langle \nabla f(x_k), z\rangle, \tag{63a}$$

$$x_{k+1} = \delta\tau_k z_k + (1-\delta\tau_k)x_k. \tag{63b}$$

Update (63a) requires the assumptions that $\mathcal{X}$ be convex and compact; under this assumption, (63a) satisfies $0 \le \langle \nabla f(x_k), x - z_k\rangle, \forall x \in \mathcal{X}$, consistent with (61b). The following proposition describes how a discretization of (62) can be used to analyze the behavior of algorithm (63).

Proposition 24 *Assume f is convex and $\mathcal{X}$ is convex and compact, and f has (L,ν)-Hölder-continuous gradients $\nu \in (0,1]$. Using the Lyapunov function*

$$E_k = A_k(f(x_k) - f(x^*)), \tag{64}$$

we obtain the error bound, $\frac{E_{k+1}-E_k}{\delta} \le \varepsilon_{k+1}$, where the error for algorithm (63) *scales as*

$$\varepsilon_{k+1} = \delta^\nu \tfrac{A_{k+1}\tau_k^{1+\nu}L}{(1+\nu)}\|z_k - x_k\|^{1+\nu}. \tag{65}$$

The choice $A_k = \frac{\delta^2 k^{(2)}}{2}$ results in a convergence rate bound of $O(1/k^\nu)$.

Proof of Proposition 23 We show that (62) is a Lyapunov function for dynamics (61).

$$\begin{aligned}\tfrac{d}{dt}\mathcal{E}_t &= e^{\beta_t}\tfrac{d}{dt}\{f(X_t)\} + \dot{\beta}_t e^{\beta_t}(f(X_t) - f(x^*)) \\ &\le e^{\beta_t}\langle \nabla f(X_t), \dot{X}_t\rangle - \dot{\beta}_t e^{\beta_t}\langle \nabla f(X_t), x^* - X_t\rangle = -\dot{\beta}_t e^{\beta_t}\langle \nabla f(X_t), x^* - Z_t\rangle \le 0.\end{aligned}$$

Proof of Proposition 24 To show bound (65) we have

$$\begin{aligned}\tfrac{E_{k+1}-E_k}{\delta} &= \tfrac{A_{k+1}}{\delta}(f(x_{k+1}) - f(x_k)) + \alpha_k(f(x_k) - f(x^*)) \\ &\le \tfrac{A_{k+1}}{\delta}\langle \nabla f(x_k), x_{k+1} - x_k\rangle + \tfrac{A_{k+1}L}{\delta(1+\nu)}\|x_{k+1} - x_k\|^{1+\nu} + \alpha_k\langle \nabla f(x_k), x_k - x^*\rangle \\ &\overset{(63b)}{=} \alpha_k\langle \nabla f(x_k), z_k - x_k\rangle + \tfrac{\delta^\nu A_{k+1}\tau_k^{1+\nu}L}{(1+\nu)}\|z_k - x_k\|^{1+\nu} + \alpha_k\langle \nabla f(x_k), x_k - x^*\rangle \\ &\overset{(63a)}{\le} \tfrac{\delta^\nu A_{k+1}\tau_k^{1+\nu}L}{(1+\nu)}\|z_k - x_k\|^{1+\nu}.\end{aligned}$$

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

E.1.1 Lyapunov and estimate sequence frameworks for Frank-Wolfe

The discrete-time estimate sequence (41) for conditional gradient method can be written:

$$\begin{aligned}\phi_{k+1}(x) - \tfrac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} := f(x_{k+1}) - \tfrac{\tilde{\varepsilon}_{k+1}}{A_{k+1}} &\overset{(41)}{=} (1-\delta\tau_k)\left(\phi_k(x) - \tfrac{\tilde{\varepsilon}_k}{A_k}\right) + \delta\tau_k f_k(x) \\ &\overset{\text{Table 1}}{=} \left(1 - \tfrac{\delta\alpha_k}{A_{k+1}}\right)\left(f(x_k) - \tfrac{\tilde{\varepsilon}_k}{A_k}\right) + \tfrac{\delta\alpha_k}{A_{k+1}} f_k(x).\end{aligned}$$

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

$$\begin{aligned}A_{k+1}\left(f(x_{k+1}) - \tfrac{\tilde{\varepsilon}_{k+1}}{A_{k+1}}\right) &= (A_{k+1} - (A_{k+1} - A_k))\left(f(x_k) - \tfrac{\tilde{\varepsilon}_k}{A_k}\right) + \alpha_k f_k(x) \\ &= A_k\left(f(x_k) - A_k^{-1}\tilde{\varepsilon}_k\right) + (A_{k+1} - A_k) f_k(x) \\ &\overset{(40)}{\leq} A_k f(x_k) - \tilde{\varepsilon}_k + (A_{k+1} - A_k) f(x).\end{aligned}$$

Rearranging, we obtain our Lyapunov argument $E_{k+1} - E_k \leq \varepsilon_{k+1}$ for (64):

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

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

$$\begin{aligned}E_k &\leq E_0 + \tilde{\varepsilon}_k \\ A_k f(x_k) &\leq (A_k - A_0) f(x) + A_0 f(x_0) + \tilde{\varepsilon}_k \\ A_k \phi_k(x) &\leq (A_k - A_0) f(x) + A_0 \phi_0(x) + \tilde{\varepsilon}_k\end{aligned}$$

Rearranging, we obtain our estimate sequence (38) ($A_0 = 1$) with an additional error term:

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

Given that the Lyapunov function property allows us to write

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

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

■