ACCELERATED FIRST-ORDER METHODS: DIFFERENTIAL EQUATIONS AND LYAPUNOV FUNCTIONS

JONATHAN W. SIEGEL*

Abstract. We develop a theory of accelerated first-order optimization from the viewpoint of differential equations and Lyapunov functions. Building upon the work in [11] and [9], we consider differential equations which model the behavior of accelerated gradient descent. Our main contribution is to provide a general framework for discretizing the differential equations to produce accelerated methods. An important novelty in our approach is the treatment of stochastic discretizations, which introduce randomness at each iteration. This leads to a unified derivation of a wide variety of methods, which include Nesterov’s accelerated gradient descent, FISTA, and accelerated coordinate descent as special cases.

Key words. Nesterov’s accelerated scheme, convex optimization, differential equations, accelerated first-order optimization

AMS subject classifications. 90C25, 90C06, 65K10

1. Introduction. Minimizing convex and strongly convex functions is a fundamental problem which arises in many areas of science. We concern ourselves here with the problem

\[(1.1) \quad \arg\min_{x \in \mathbb{R}^n} f(x)\]

where \(f\) is a strongly convex function. Due to the importance of very large scale problems of the form 1.1 arising from machine learning and data science, first-order methods have gained popularity. In practice, methods which only utilize gradient information are often the only ones which can be applied to large scale problems of the form 1.1.

As a result, there has been a lot of research into developing optimal first-order methods for convex optimization. Beginning with the seminal discovery of Nesterov’s accelerated gradient descent [7], many different accelerated methods have been developed by many authors. For instance, accelerated methods for solving composite optimization problems are developed in [2] and [8] and an accelerated version of coordinate descent was developed in [6], to name only a few.

In spite of this progress, these methods have remained somewhat mysterious and difficult to understand. Consequently, there has been a lot of work in explaining these methods. For instance, in [3], a geometric explanation of acceleration is given, in [1] accelerated methods are derived as a coupling of gradient and mirror descent, and in [9] and [11] these methods are studied via the differential equations which they discretize.

However, we believe that a deeper understanding of accelerated first-order methods can still be reached. In this paper, we provide a step in this direction. We build upon the work in [11] and [9] and study accelerated first-order methods from the viewpoint of the underlying differential equations.

Our main contribution is the derivation of a general framework for discretizing the differential equations to produce accelerated methods. An important novelty in this regard is our treatment of stochastic discretizations, which introduce randomness.

*Department of Mathematics, Pennsylvania State University, State College, PA (jus1949@psu.edu, jwsiegel2510@gmail.com).
in every iteration. The question of whether such methods, for example accelerated coordinate descent, could systematically be connected to differential equations was posed in [11]. Our treatment leads to a unified derivation of a wide variety of methods in the literature, including deterministic methods such as Nesterov’s accelerated gradient descent and FISTA, and stochastic methods like accelerated coordinate descent.

The key to our theory is that many of these methods can be derived as special discretizations of a certain damped Hamiltonian dynamics. The discretizations we consider all look very similar; they consist of an explicit forward step in position, a semi-implicit step in velocity, and finally a small perturbation (second order in the step size) which ensures a sufficient decrease in the objective. This general framework, which we introduce in sections three through six, provides a unified derivation of a wide variety of accelerated first-order methods. We find it remarkable that so many methods can be obtained as discretizations of the same equations in such a simple way.

The paper is organized as follows. In the next section, we analyze the differential equations underlying accelerated methods. We introduce simple equations for the strongly convex case, which are similar to the approach in [11]. In the third section, we discuss how to discretize the differential equations to obtain accelerated methods when the objective is smooth, obtaining a version of Nesterov’s accelerated gradient descent. The method derived in this first step is then modified in the next section to incorporate non-smooth objectives. In the fifth section, we modify the theory to incorporate stochastic discretizations. In both of these later cases, an important condition is that we can obtain a sufficient decrease in the objective, a condition which we precisely formulate. In the fourth and fifth sections, we also show how accelerated methods for composite optimization and accelerated coordinate descent follow as special cases of our theory. Finally, in the last section, we provide some concluding remarks and further research directions.

2. The Differential Equations. In this section, we analyze the differential equations underlying accelerated first-order optimization methods. The theory in the convex case was already considered in [9] and [11]. We only consider the strongly convex case and our approach is simpler than what we have found in the literature, even though similar work was done in [11].

We first briefly recall the notion of strong convexity.

**Definition 2.1.** Let $0 < \alpha$. A convex function $f$ is $\alpha$-strongly convex if for all $x, y$ and $g \in \partial f(y)$, it holds that

\[
 f(x) \geq f(y) + \langle g, x - y \rangle + \frac{\alpha}{2} \|x - y\|^2
\]

**2.1. Strongly Convex Dynamics.** If $f$ is $\alpha$-strongly convex and differentiable, we consider the following damped Hamiltonian dynamics with potential energy $f$

\[
 \dot{x} = v, \quad \dot{v} = -2\sqrt{\alpha}v - \nabla f(x)
\]

For non-differentiable $f$, we replace the gradient by an element of the sub-differential to obtain the dynamics

\[
 \dot{x} = v, \quad -2\sqrt{\alpha}v - \dot{v} \in \partial f(x)
\]

We use a Lyapunov function to prove that the objective error decreases at a linear rate of $-\sqrt{\alpha}$ under this dynamics.
THEOREM 2.2. Let \( f \) be \( \alpha \)-strongly convex and differentiable. Assume that \( x(t) \) and \( v(t) \) obey the dynamics 2.3 (or equivalently 2.2) and \( v(0) = 0 \). Then we have

\[
(2.4) \quad f(x(t)) - f(x^*) \leq 2e^{-\sqrt{\alpha}t}(f(x(0)) - f(x^*))
\]

where \( x^* \) minimizes \( f \).

Proof. Consider the Lyapunov function

\[
L(t) = f(x_t) - f(x^*) + \frac{1}{2}\|\sqrt{\alpha}(x_t - x^*) + v_t\|^2
\]

Here I have written \( x_t \) for \( x(t) \) and \( v_t \) for \( v(t) \) to simplify notation. We will show that \( L'(t) \leq -\sqrt{\alpha}L(t) \). This completes the proof since

\[
(2.5) \quad f(x_t) - f(x^*) \leq L(t) \leq e^{-\sqrt{\alpha}t}L(0) = e^{-\sqrt{\alpha}t}(f(x_0) - f(x^*) + \frac{\alpha}{2}\|x_0 - x^*\|^2)
\]

and \( \frac{\alpha}{2}\|x_0 - x^*\|^2 \leq f(x_0) - f(x^*) \) by the strong convexity of \( f \), so that

\[
(2.6) \quad f(x_t) - f(x^*) \leq 2e^{-\sqrt{\alpha}t}(f(x(0)) - f(x^*))
\]

We bound \( L'(t) \) as follows

\[
(2.7) \quad L'(t) = \langle \nabla f(x_t), \dot{x}_t \rangle + \langle \sqrt{\alpha}\dot{x}_t + \dot{v}_t, \sqrt{\alpha}(x_t - x^*) + v_t \rangle
\]

using the dynamics 2.2 to evaluate \( \dot{x}_t \) and \( \dot{v}_t \), we obtain

\[
(2.8) \quad L'(t) = \langle \nabla f(x_t), v_t \rangle + \langle -\sqrt{\alpha}v_t - \nabla f(x_t), \sqrt{\alpha}(x_t - x^*) + v \rangle
\]

where \( g_t \in \partial f(x_t) \). Simplifying this, we see that

\[
(2.9) \quad L'(t) = -\sqrt{\alpha}\langle \nabla f(x_t), (x_t - x^*) \rangle - \alpha\langle v_t, (x_t - x^*) \rangle - \sqrt{\alpha}\langle v_t, v_t \rangle
\]

It is here that we use strong convexity, namely

\[
\langle \nabla f(x_t), (x_t - x^*) \rangle \geq f(x_t) - f(x^*) + \frac{\alpha}{2}\|x_t - x^*\|^2
\]

Plugging this into 2.9 and simplifying the inner products we get

\[
(2.10) \quad L'(t) \leq -\sqrt{\alpha}\left(f(x_t) - f(x^*) + \frac{1}{2}\|\sqrt{\alpha}(x_t - x^*) + v_t\|^2\right) - \frac{\sqrt{\alpha}}{2}\|v_t\|^2
\]

which implies

\[
(2.11) \quad L'(t) \leq -\sqrt{\alpha}L(t)
\]

as desired.

3. Discrete Dynamics for Smooth Objectives. In this section, we consider smooth objectives and derive a version of Nesterov’s accelerated gradient descent. We derive this method as a particular discretization of 2.2. The method is made up of a forward step in position \( x \), a semi-implicit step in velocity \( v \), and finally a small perturbation (second order in the step size) which ensures a sufficient decrease in the Lyapunov function. This perturbation moves \( x \) to decrease the objective and moves \( v \) to compensate, so that the second part of the Lyapunov function (the squared norm) doesn’t increase. These steps are all explained in detail in this section.

We begin by briefly recalling the definition of smoothness.
**Definition 3.1.** Let $L > 0$. A differentiable function $f$ is $L$-smooth if

\[(3.1) \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|\]

It is an easy consequence of the above definition that

\[(3.2) \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2\]

Our goal will now be to discretize 2.2 so that a discrete version of the Lyapunov function in the proof of theorem 2.2, for instance,

\[L_n = f(x_n) - f(x^*) + \frac{1}{2}\|\sqrt{\alpha}(x_n - x^*) + v_n\|^2\]

will be decreased by a constant factor in each timestep. It actually turns out that the correct discrete Lyapunov function to consider is

\[(3.3) \quad L_n = f(x_n) - f(x^*) + \frac{1}{2}\|\sqrt{\alpha}(x_n - x^*) + (1 + s\sqrt{\alpha})v_n\|^2\]

where $s$ is the step size of the discretization.

We first note that the $L$-smoothness of $f$ implies that

\[f\left(x_n - \frac{1}{L}\nabla f(x_n)\right) - f(x^*) \leq f(x_n) - f(x^*) - \frac{1}{2L}\|\nabla f(x_n)\|^2\]

i.e. that taking a small gradient step ensures a decrease of the objective. Observe also that the second part of the Lyapunov function,

\[\frac{1}{2}\|\sqrt{\alpha}(x_n - x^*) + (1 + s\sqrt{\alpha})v_n\|^2\]

can be kept constant by adjusting $v_n$ appropriately. Putting this together, we see that the following update

\[(3.4) \quad x_n \leftarrow x_n - \frac{1}{L}\nabla f(x_n), \quad v_n \leftarrow v_n + (1 + s\sqrt{\alpha})^{-1}\sqrt{\alpha}\nabla f(x_n)\]

decreases the Lyapunov function by $\frac{1}{2L}\|\nabla f(x_n)\|^2$, so that

\[L_n \leftarrow L_n - \frac{1}{2L}\|\nabla f(x_n)\|^2\]

We call this a sufficient decrease update and it plays an important role in stabilizing our discretization, which we now discuss.

Consider the following discretization of 2.2, which consists of a forward Euler step in $x$ and a semi-implicit step in $v$, followed by a sufficient decrease update.

\[x' = x + sv_n\]
\[v' = v_n - (1 + s\sqrt{\alpha})^{-1}\left(s\sqrt{\alpha}v_n + s\nabla f(x')\right) - s\sqrt{\alpha}v_n\]
\[x_{n+1} = x' - \frac{1}{L}\nabla f(x')\]
\[v_{n+1} = v' + (1 + s\sqrt{\alpha})^{-1}\frac{\sqrt{\alpha}}{L}\nabla f(x')\]
The most complicated part of this discretization is the semi-implicit update for $v$. Unfortunately, we haven’t found a simpler update for which the discretization remains stable.

We now show that under a mild step size restriction, this discretization produces an accelerated method.

**Theorem 3.2.** If $s \leq \frac{1}{\sqrt{L}}$ and $f$ is $\alpha$-strongly convex and $L$-smooth, then the iteration 3.5 satisfies

$$L_{n+1} \leq (1 + s\sqrt{\alpha})^{-1}L_n$$

Before proving this, we give a simple corollary showing that this implies an accelerated convergence rate.

**Corollary 3.1.** If $s \leq \frac{1}{\sqrt{L}}$ and $f$ is $\alpha$-strongly convex and $L$-smooth, then the iteration 3.5 with $v_0 = 0$ satisfies

$$f(x_n) - f(x^*) \leq 2 \left(1 + s\sqrt{\alpha}\right)^{-n}(f(x_0) - f(x^*))$$

**Proof.** By theorem 3.2, we have

$$f(x_n) - f(x^*) \leq L_n \leq (1 + s\sqrt{\alpha})^{-n}L_0$$

Since $v_0 = 0$ and $f$ is $\alpha$-strongly convex, we get

$$L_0 = f(x_0) - f(x^*) + \frac{\alpha}{2}\|x_0 - x^*\|^2 \leq 2(f(x_0) - f(x^*))$$

**Proof of Theorem 3.2.** Throughout the proof, we will use the following elementary fact. Let $t_n$ be some quantity which changes throughout our iteration. Then

$$\frac{1}{2}\|t_{n+1}\|^2 - \frac{1}{2}\|t_n\|^2 = \langle t_{n+1} - t_n, t_n \rangle + \frac{1}{2}\|t_{n+1} - t_n\|^2$$

and

$$\frac{1}{2}\|t_{n+1}\|^2 - \frac{1}{2}\|t_n\|^2 = \langle t_{n+1} - t_n, t_{n+1} \rangle - \frac{1}{2}\|t_{n+1} - t_n\|^2$$

Applying each of these identities once, we see that if $t_{n+1} - t_n = a + b$, then

$$\frac{1}{2}\|t_{n+1}\|^2 - \frac{1}{2}\|t_n\|^2 = \langle a, t_n \rangle + \langle b, t_{n+1} \rangle + \frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2$$

We will use these identities without explicit mention in what follows. We now prove that

$$L_{n+1} \leq L_n - s\sqrt{\alpha}L_{n+1}$$

To do so, we calculate the change in $L$ due to the forward step in $x$,

$$L(x'_n, v_n) - L_n = f(x'_n) - f(x_n) + s\langle \sqrt{\alpha}v_n, \sqrt{\alpha}(x'_n - x^*) \rangle + (1 + s\sqrt{\alpha})v_n \tag{3.6}$$

and the change due to the semi-implicit step for $v$, noting that the change in $v$ can be broken up as

$$(1 + s\sqrt{\alpha})(v'_n - v_n) = -(s\sqrt{\alpha}v_n + s\nabla f(x'_n)) - s\sqrt{\alpha}(1 + s\sqrt{\alpha})v'_n$$
to get

\[
L(x'_n, v'_n) - L(x'_n, v_n) = -s(\sqrt{\alpha}v_n + \nabla f(x'_n), \sqrt{\alpha}(x'_n - x*) + (1 + s\sqrt{\alpha})v_n) \\
- s(\sqrt{\alpha}(1 + s\sqrt{\alpha})v'_n, \sqrt{\alpha}(x'_n - x*) + (1 + s\sqrt{\alpha})v'_n)
\]

(3.7)

\[
+ \frac{s^2}{2} \|\sqrt{\alpha}v_n + \nabla f(x'_n)\|^2 \\
- \frac{s^2\alpha}{2} \|v'_n\|^2
\]

Adding these, collecting terms, and recalling that \(x'_n - x_n = s v_n\) we obtain

\[
L(x'_n, v'_n) - L_n = f(x'_n) - f(x_n) - (\nabla f(x'_n), x'_n - x_n) \\
- s\sqrt{\alpha}(\nabla f(x'_n), x'_n - x*) \\
- s\sqrt{\alpha}((1 + s\sqrt{\alpha})v'_n, \sqrt{\alpha}(x'_n - x*) + (1 + s\sqrt{\alpha})v'_n)
\]

(3.8)

\[
+ \frac{s^2}{2} \|\nabla f(x'_n)\|^2 - \frac{s^2\alpha}{2} \|v'_n\|^2
\]

The terms on the first line are \(\leq 0\), by the convexity of \(f\). The inner product on the second line can be bounded using the strong convexity of \(f\), as

\[
\langle \nabla f(x'_n), x'_n - x* \rangle \geq f(x'_n) - f(x*) + \frac{\alpha}{2} \|x'_n - x*\|^2
\]

Plugging this bound into the above equation and completing the square with line three, yields

\[
L(x'_n, v'_n) - L_n \leq -s\sqrt{\alpha}L(x'_n, v'_n) + \frac{s^2}{2} \|\nabla f(x'_n)\|^2 \\
- \frac{s^2\alpha}{2} \|v'_n\|^2 - \frac{s\sqrt{\alpha}}{2}(1 + s\sqrt{\alpha})^2 \|v'_n\|^2
\]

(3.9)

The sufficient decrease update now decreases the Lyapunov function by at least \(\frac{s^2}{2} \|\nabla f(x'_n)\|^2\) since \(s \leq \frac{1}{2\sqrt{\alpha}}\), and we obtain

\[
L_{n+1} - L_n \leq -s\sqrt{\alpha}L(x'_n, v'_n)
\]

(3.10)

To complete the proof, we merely note that \(L_{n+1} \leq L(x'_n, v'_n)\) (the sufficient decrease update decreased the Lyapunov function). This implies

\[
L_{n+1} - L_n \leq -s\sqrt{\alpha}L_{n+1}
\]

(3.11)

as desired.

4. Discrete Dynamics for Non-smooth Objectives. In this section we consider the situation where the objective \(f\) is not smooth, but we still assume that \(f\) is \(\alpha\)-strongly convex. We first consider simply replacing \(\nabla f(x'_n)\) by some element in the sub-differential of \(f\), say \(g_n \in \partial f(x'_n)\). We observe that the argument for smooth functions continues to apply as long as we can find an analog of the sufficient decrease update, i.e. if we have an update

\[
x_n \leftarrow x_n - \delta_n, \ v_n \leftarrow v_n + (1 + s\sqrt{\alpha})^{-1}\sqrt{\alpha}\delta_n
\]

(4.1)
such that
\[ f(x_n - \delta_n) \leq f(x_n) - \frac{s^2}{2} \|g_n\|^2 \]

Unfortunately, in many cases of interest this is not possible. To get around this, we allow \( g_n / \in \partial f(x'_n) \). Then, we observe that our Lyapunov argument still works as long as the following decrease condition can be guaranteed.

\[
\forall z, \quad f(x_{n+1}) - f(z) \leq (g_n, x'_n - z) - \frac{\alpha}{2} \|x'_n - z\|^2 - \frac{s^2}{2} \|g_n\|^2
\]

Let us examine this condition for a moment. Note that if \( g_n / \in \partial f(x'_n) \), then by the strong convexity, we have

\[
\forall z, \quad f(x'_n) - f(z) \leq (g_n, x'_n - z) - \frac{\alpha}{2} \|x'_n - z\|^2
\]

So, in this case, the above condition is equivalent to a decrease in the objective.

\[
f(x_{n+1}) = f(x'_n - \delta_n) \leq f(x'_n) - \frac{s^2}{2} \|g_n\|^2
\]

What the new condition 4.2 does is simply to allow \( g_n / \in \partial f(x'_n) \), but still to enforce a combined decrease and strong convexity condition. We will see that for many problems of interest, in particular composite optimization, \( g_n \) and \( \delta_n \) can be chosen to satisfy this condition. This will lead to an accelerated version of forward-backward iteration, of which accelerated projected gradient descent and FISTA [2] are special cases.

The discretization we arrive at is the following.

\[
\begin{align*}
x'_n &= x_n + sv_n \\
v'_n &= v_n - (1 + s\sqrt{\alpha})^{-1} (s\sqrt{\alpha}v_n + sg_n) - s\sqrt{\alpha}v'_n \\
x_{n+1} &= x'_n - \delta_n \\
v_{n+1} &= v'_n + (1 + s\sqrt{\alpha})^{-1} \sqrt{\alpha}\delta_n
\end{align*}
\]

where \( g_n \) and \( \delta_n \) are chosen so that 4.2 holds.

This is the same as in the smooth case, except that the gradients have been replaced by \( g_n \) and the sufficient decrease update has been changed. We now prove that this scheme leads to an accelerated method. The proof is very similar to the smooth case. In this case, we don’t even need to assume the strong convexity of \( f \). This assumption is subsumed by 4.2.

**Theorem 4.1.** Assume that \( g_n \) and \( \delta_n \) are chosen so that the condition 4.2 holds at every iteration of the scheme 4.4. Then we will have

\[
L_{n+1} \leq (1 + s\sqrt{\alpha})^{-1} L_n
\]

where \( L_n \) is the same Lyapunov function as in the smooth case (equation 3.3).

As in the smooth case, we have a corollary which shows that this method achieves an accelerated convergence rate. The proof is exactly the same as in the smooth case so we omit it.

**Corollary 4.1.** Assume that \( g_n \) and \( \delta_n \) are chosen so that the condition 4.2 holds at every iteration of the scheme 4.4. Then if \( v_0 = 0 \), we have

\[
f(x_n) - f(x^*) \leq 2 (1 + s\sqrt{\alpha})^{-n} (f(x_0) - f(x^*))
\]
Proof of Theorem 4.1. We proceed exactly as in the proof of theorem 3.2, replacing \( \nabla f(x'_n) \) by \( g_n \) to obtain, in place of equation 3.8,

\[
\begin{align*}
L(x'_n, v'_n) - L_n &= f(x'_n) - f(x_n) - (g_n, x'_n - x_n) \\
&\quad - s\sqrt{\alpha}(g_n, x'_n - x^*) \\
&\quad - s\sqrt{\alpha}((1 + s\sqrt{\alpha})v'_n, \sqrt{\alpha}(x'_n - x^*) + (1 + s\sqrt{\alpha})v'_n) \\
&\quad + \frac{s^2}{2}\|g_n\|^2 - \frac{s^2\alpha}{2}\|v'_n\|^2
\end{align*}
\]

(4.5)

Noting that by the construction of the sufficient decrease update (the last two lines of 4.4),

\[ L_{n+1} - L(x'_n, v'_n) = f(x_{n+1}) - f(x'_n) \]

we get

\[
\begin{align*}
L_{n+1} - L_n &= f(x_{n+1}) - f(x_n) - (g_n, x'_n - x_n) \\
&\quad - s\sqrt{\alpha}(g_n, x'_n - x^*) \\
&\quad - s\sqrt{\alpha}((1 + s\sqrt{\alpha})v'_n, \sqrt{\alpha}(x'_n - x^*) + (1 + s\sqrt{\alpha})v'_n) \\
&\quad + \frac{s^2}{2}\|g_n\|^2 - \frac{s^2\alpha}{2}\|v'_n\|^2
\end{align*}
\]

(4.6)

We now apply the decrease condition 4.2 with \( z = x_n \) and \( z = x^* \) to the first two lines of this equation. This gives

\[
\begin{align*}
L_{n+1} - L_n &\leq -\frac{\alpha}{2}\|x'_n - x_n\|^2 - \frac{s^2}{2}\|g_n\|^2 \\
&\quad - s\sqrt{\alpha} \left( f(x_{n+1}) - f(x^*) + \frac{\alpha}{2}\|x'_n - x^*\|^2 + \frac{s^2}{2}\|g_n\|^2 \right) \\
&\quad - s\sqrt{\alpha}((1 + s\sqrt{\alpha})v'_n, \sqrt{\alpha}(x'_n - x^*) + (1 + s\sqrt{\alpha})v'_n) \\
&\quad + \frac{s^2}{2}\|g_n\|^2 - \frac{s^2\alpha}{2}\|v'_n\|^2
\end{align*}
\]

(4.7)

Completing the square and noting that the sufficient decrease update (the last two lines of 4.4) is designed so that

\[
\frac{1}{2}\|\sqrt{\alpha}(x'_n - x^*) + (1 + s\sqrt{\alpha})v'_n\|^2 = \frac{1}{2}\|\sqrt{\alpha}(x_{n+1} - x^*) + (1 + s\sqrt{\alpha})v_{n+1}\|^2
\]

we see that

\[
L_{n+1} - L_n \leq -s\sqrt{\alpha}L_{n+1}
\]

(4.8)

as desired. \( \square \)

4.1. Accelerated Forward-Backward Splitting. In this subsection, we apply theorem 4.1 to strongly convex composite objectives, i.e. objectives of the form

\[
f(x) = g(x) + h(x)
\]

(4.9)

where \( g \) is \( \alpha \)-strongly convex and \( L \)-smooth, and \( h \) is an arbitrary convex function. We also assume that we are able to compute a proximal update for \( h \), i.e. solve

\[
y^* = \text{prox}_{s,h}(x) = \arg\min_y h(y) + \frac{1}{2s}\|y - x\|^2
\]

(4.10)
The proximal update is essentially a step of backward Euler with step size $s$, hence the name accelerated forward-backward method. For many convex functions of interest, the proximal update can be efficiently computed.

For example, if $h$ is the characteristic function of a convex set $S$, then 4.10 is just a projection onto $S$. In this case, Theorem 4.1 recovers accelerated projected gradient descent.

Another example of interest is $h(x) = \|x\|_1$, in which case 4.10 is just soft-thresholding with parameter $s$. In this instance Theorem 4.1 recovers a version of FISTA [2].

Our goal is to show how $g_n$ and $\delta_n$ can be chosen to satisfy the decrease condition 4.2. The next lemma answers this question for us.

**Lemma 4.2.** Assume that $f(x) = g(x) + h(x)$, with $g$ an $L$-smooth, $\alpha$-strongly convex function and $h$ a convex function. Then setting

$$g_n = \frac{1}{s^2} \left( x_n' - \text{prox}_{s^2 h} \left( x_n' - s^2 \nabla g(x_n') \right) \right)$$

and $\delta_n = s^2 g_n$ will satisfy the condition 4.2 as long as $s \leq \frac{1}{\sqrt{\alpha}}$.

**Proof.** Note that the choice of $g_n$ and $\delta_n$ implies that

$$x_{n+1} = \arg \min_x h(x) + \frac{1}{2s^2} \|x - \left( x_n' - s^2 \nabla g(x_n') \right) \|^2$$

We denote by $y_n$ the intermediate point $x_n' - s^2 \nabla g(x_n')$ so that the above becomes

$$x_{n+1} = \arg \min_x h(x) + \frac{1}{2s^2} \|x - y_n\|^2$$

Now let $z$ be arbitrary and note that our goal is to bound

$$f(x_{n+1}) - f(z) = (g(x_{n+1}) - g(z)) + (h(x_{n+1}) - h(z))$$

We consider the second of these terms first. Note that equation 4.12 implies that $L(y_n - x_{n+1}) \in \partial h(x_{n+1})$. This means that (since $h$ is convex)

$$h(x_{n+1}) - h(z) \leq \frac{1}{s^2} \langle (y_n - x_{n+1}), x_{n+1} - z \rangle$$

We proceed to bound $g(x_{n+1}) - g(z)$.

$$g(x_{n+1}) - g(z) = (g(x_{n+1}) - g(x_n')) + (g(x_n') - g(z))$$

The first term above is bounded due to the $L$-smoothness of $g$ and the assumption that $s^2 \leq \frac{1}{L}$:

$$g(x_{n+1}) - g(x_n') \leq \langle \nabla g(x_n'), x_{n+1} - x_n' \rangle + \frac{1}{2s^2} \|x_{n+1} - x_n'\|^2$$

and the second is bounded due to the strong convexity of $g$

$$g(x_n') - g(z) \leq \langle \nabla g(x_n'), x_n' - z \rangle - \frac{\alpha}{2} \|x_n' - z\|^2$$

Combining these two bounds with equation 4.13 and noting that

$$g_n = \frac{1}{s^2} (x_n' - x_{n+1}) = \frac{1}{s^2} (y_n - x_{n+1}) + \nabla g(x_n')$$
we obtain

\begin{equation}
(4.15) \quad f(x_{n+1}) - f(z) \leq \langle g_n, x_{n+1} - z \rangle - \frac{\alpha}{2} \|x_n' - z\|^2 + \frac{s^2}{2}\|g_n\|^2
\end{equation}

Now we write

\begin{equation}
x_{n+1} - z = x_{n+1} - x_n' + x_n' - z = -s^2 g_n + x_n' - z
\end{equation}

to obtain

\begin{equation}
(4.16) \quad f(x_{n+1}) - f(z) \leq \langle g_n, x_n' - z \rangle - \frac{\alpha}{2} \|x_n' - z\|^2 - \frac{s^2}{2}\|g_n\|^2
\end{equation}

which is exactly 4.2.

5. Discrete Stochastic Dynamics. In this section, we extend our theory to stochastic discretizations. By this we simply mean schemes which introduce randomness in each iteration. The important new step is to modify the sufficient decrease update appropriately when the gradient is sampled randomly. We will first present the ideas for smooth functions and then show how to modify it (analogous to the previous section) for non-smooth objectives.

Recall that for smooth objectives the sufficient decrease update was critical for obtaining a stable accelerated method. What we needed was a way of decreasing the objective sufficiently by perturbing \(x\). Then we could perturb \(v\) appropriately to keep the second term in our Lyapunov function constant. We obtained an update of the form 3.4, which reduced the Lyapunov function by at least \(s^2\|g_n\|^2\).

When deriving stochastic accelerated methods, we want to replace \(\nabla f(x_n')\) by some sample \(g_n\), where \(E_n(g_n) = \nabla f(x_n')\) (here \(E_n\) denotes the expectation taken with respect to randomness introduced in iteration \(n\), essentially a conditional expectation). It turns out that this will work as long as we can guarantee an objective decrease of at least \(s^2\|g_n\|^2\). Note here that there is no expectation inside of the norm, i.e. this is the norm of the actual gradient sample encountered at iteration \(n\).

We obtain the following discretization.

\begin{equation}
(5.1) \quad \begin{aligned}
x_n' &= x_n + sv_n \\
v_n' &= v_n - (1 + s\sqrt{\alpha})^{-1} (s\sqrt{\alpha} v_n + sg_n) - s\sqrt{\alpha} v_n' \\
x_{n+1} &= x_n' - \delta_n \\
v_n+1 &= v_n' + (1 + s\sqrt{\alpha})^{-1} \sqrt{\alpha} \delta_n
\end{aligned}
\end{equation}

where \(E_n(g_n) = \nabla f(x_n')\) and \(\delta_n\) is chosen (dependent on \(g_n\)) so that

\begin{equation}
(5.2) \quad f(x_{n+1}) \leq f(x_n') - \frac{s^2}{2}\|g_n\|^2
\end{equation}

This method will achieve an accelerated convergence rate under these conditions.

**Theorem 5.1.** Assume that \(f\) is \(\alpha\)-strongly convex and differentiable. Then as long as condition 5.2 holds, the iterates of 5.1 will satisfy

\[E_n(L_{n+1}) \leq (1 + s\sqrt{\alpha})^{-1} L_n\]

where \(L_n\) is the same Lyapunov function introduced in 3.3.
This theorem will follow as a special case of the theorem for non-smooth schemes, so we omit a proof. Before continuing on to stochastic schemes for non-smooth functions, we state a corollary which gives us the accelerated convergence rate.

**Corollary 5.1.** Assume that $f$ is $\alpha$-strongly convex and differentiable. Then as long as condition 5.2 holds, the iterates of 5.1 with $v_0 = 0$ will satisfy

$$\mathbb{E}(f(x_n) - f(x^*)) \leq 2(1 + s\sqrt{\alpha})^{-n}(f(x_0) - f(x^*))$$

**Proof.** Note that Theorem 5.1 implies that

$$\mathbb{E}(L_{n+1}) \leq (1 + s\sqrt{\alpha})^{-1} L_n$$

Taking the expectation with respect to the randomness introduced in previous iterations, we obtain

$$\mathbb{E}(L_{n+1}) \leq (1 + s\sqrt{\alpha})^{-1} \mathbb{E}(L_n)$$

so that

$$\mathbb{E}(f(x_n) - f(x^*)) \leq \mathbb{E}(L_{n+1}) \leq (1 + s\sqrt{\alpha})^{-n} L_0$$

Finally, as in the previous proofs of this corollary, the strong convexity along with $v_0 = 0$, implies that

$$L_0 \leq 2(f(x_0) - f(x^*))$$

We now turn to stochastic schemes for non-smooth functions. The idea is very similar to the deterministic scheme for non-smooth functions. We want to choose $g_n$ as a random sample of an element in the subgradient $\partial f(x_n')$, however, we don’t restrict $\mathbb{E}(g_n) \in \partial f(x_n')$. Instead, we enforce a stochastic version of the constraint 4.2. The scheme we consider is same as 4.4

$$x_n' = x_n + sv_n$$

$$v_n' = v_n - (1 + s\sqrt{\alpha})^{-1} (s\sqrt{\alpha}v_n + sg_n) - s\sqrt{\alpha}v_n'$$

$$x_{n+1} = x_n' - \delta_n$$

$$v_{n+1} = v_n' + (1 + s\sqrt{\alpha})^{-1} \sqrt{\alpha}\delta_n$$

except that we allow $g_n$ and $\delta_n$ to be (dependent) random variables and enforce the following condition

$$\forall z, f(x_{n+1}) - f(z) \leq \langle \mathbb{E}(g_n), x_n' - z \rangle - \frac{\alpha}{2} \|x_n' - z\|^2 - \frac{s^2}{2} \|g_n\|^2$$

which only differs from 4.2 in the expectation taken in the first inner product.

Note first that the schemes 5.1 and 5.3 are exactly the same, except that the condition 5.4 is weaker than the conditions enforced in the smooth case. This follows since if $f$ is differentiable and $\alpha$-strongly convex, then

$$\forall z, f(x_{n+1}') - f(z) \leq \langle \nabla f(x_n'), x_n' - z \rangle - \frac{\alpha}{2} \|x_n' - z\|^2$$

This, combined with the requirement in 5.1 that $\mathbb{E}(g_n) = \nabla f(x_n')$ and that $\delta_n$ is chosen so that 5.2 holds, implies the condition 5.4. This means that the proof below will imply Theorem 5.1. In fact, the scheme 5.3 is most often applied in this way to smooth, strongly convex functions. In this case it leads to different versions of accelerated coordinate descent, as we will show later.

We now prove that the scheme 5.3 achieves the desired accelerated convergence rate. As in the non-smooth deterministic case, we don’t need to assume that our objective is strongly convex. This assumption is superseded by condition 5.4.
Theorem 5.2. Assume that $g_n$ and $\delta_n$ are chosen so that the condition 5.4 holds at every iteration of the scheme 5.3. Then we will have

$$E_n(L_{n+1}) \leq (1 + s\sqrt{\alpha})^{-1} L_n$$

where $L_n$ is the same Lyapunov function as in the smooth case (equation 3.3).

As in the smooth case, we have the corollary that this implies the accelerated convergence rate. Since the proof is exactly the same as in the smooth case, we omit it.

Corollary 5.2. Assume that $g_n$ and $\delta_n$ are chosen so that the condition 5.4 holds at every iteration of the scheme 5.3. Then, if $v_0 = 0$, we will have

$$E(f(x_n) - f(x^*)) \leq 2(1 + s\sqrt{\alpha})^{-n}(f(x_0) - f(x^*))$$

Proof of Theorem 5.2. As in the proof of Theorem 4.1, we obtain (equation 4.5)

$$L(x'_n, v'_n) = f(x'_n) - f(x_n) - (g_n, x'_n - x_n)$$

$$-s\sqrt{\alpha}(g_n, x'_n - x^*)$$

$$-s\sqrt{\alpha}((1 + s\sqrt{\alpha})v'_n, \sqrt{\alpha}(x'_n - x^*) + (1 + s\sqrt{\alpha})v'_n)$$

$$+ \frac{s^2}{2} \|g_n\|^2 - \frac{s^2\alpha}{2} \|v'_n\|^2$$

(5.5)

We note that the sufficient decrease update (the last two lines of 5.2) implies that $L_{n+1} - L(x'_n, v'_n) = f(x_{n+1}) - f(x'_n)$, so that

$$L_{n+1} - L_n = f(x_{n+1}) - f(x_n) - (g_n, x'_n - x_n)$$

$$-s\sqrt{\alpha}(g_n, x'_n - x^*)$$

$$-s\sqrt{\alpha}((1 + s\sqrt{\alpha})v'_n, \sqrt{\alpha}(x'_n - x^*) + (1 + s\sqrt{\alpha})v'_n)$$

$$+ \frac{s^2}{2} \|g_n\|^2 - \frac{s^2\alpha}{2} \|v'_n\|^2$$

(5.6)

We now apply the decrease condition 5.4 with $z = x_n$ and $z = x^*$ to the first two lines of this equation. Because we have an expectation $E_n(g_n)$ in the condition 5.4, this results in extra terms which are an inner produce with the difference $g_n - E_n(g_n)$. We get

$$L_{n+1} - L_n \leq -\frac{s\alpha}{2} \|x'_n - x_n\|^2 - \frac{s^2}{2} \|g_n\|^2 - (g_n - E_n(g_n), x'_n - x_n)$$

$$- s\sqrt{\alpha}\left(f(x_{n+1}) - f(x^*) + \frac{\alpha}{2} \|x'_n - x^*\|^2 + \frac{s^2}{2} \|g_n\|^2\right)$$

$$- s\sqrt{\alpha}(g_n - E_n(g_n), x'_n - x^*)$$

$$- s\sqrt{\alpha}((1 + s\sqrt{\alpha})v'_n, \sqrt{\alpha}(x'_n - x^*) + (1 + s\sqrt{\alpha})v'_n)$$

$$+ \frac{s^2}{2} \|g_n\|^2 - \frac{s^2\alpha}{2} \|v'_n\|^2$$

(5.7)

Collecting terms and completing the square as in the previous proofs, we get (throwing out unnecessary negative terms)

$$L_{n+1} - L_n \leq -s\sqrt{\alpha}\left(f(x_{n+1}) - f(x^*) + \frac{1}{2} \sqrt{\alpha}(x'_n - x^*) + (1 + s\sqrt{\alpha})v'_n\|^2\right)$$

$$- (g_n - E_n(g_n), x'_n - x_n)$$

$$- s\sqrt{\alpha}(g_n - E_n(g_n), x'_n - x^*)$$

(5.8)
Recalling that the sufficient decrease update (the last two lines of 5.2) was designed so that
\[
\frac{1}{2}\|\sqrt{\alpha}(x_n' - x^*) + (1 + s\sqrt{\alpha})v_n'\|^2 = \frac{1}{2}\|\sqrt{\alpha}(x_{n+1} - x^*) + (1 + s\sqrt{\alpha})v_{n+1}\|^2
\]
we see that
\[
L_{n+1} - L_n \leq -s\sqrt{\alpha}L_{n+1}
\]
(5.9)
\[
-\langle g_n - E_n(g_n), x_n' - x_n \rangle - s\sqrt{\alpha}(g_n - E_n(g_n), x_n' - x^*)
\]
Finally, we take the expectation \( E_n \) on both sides to obtain
\[
E_n(L_{n+1}) - L_n \leq -s\sqrt{\alpha}E_n(L_{n+1})
\]
which implies
\[
E_n(L_{n+1}) \leq (1 + s\sqrt{\alpha})^{-1}L_n
\]
as desired.

5.1. Accelerated Coordinate Descent. We now show how two versions of accelerated coordinate descent follow as special cases of Theorem 5.1.

The premise we consider is that the objective \( f \) is \( \alpha \)-strongly convex and the gradient is coordinate-wise \( L_i \)-smooth, i.e. if we denote by \( \nabla f(x) \), the \( i \)-th coordinate of the gradient of \( f \), then
\[
|\nabla f(x)_i - \nabla f(x + ce_i)_i| \leq L_i|c|
\]
Intuitively, this means that the \( i \)-th diagonal entry of the Hessian of \( f \) is bounded by \( L_i \) at each point.

Note also that the \( L_i \)-smoothness implies that we obtain a sufficient decrease when moving in the direction \( i \), i.e.
\[
f\left(x - \frac{1}{L_i}\nabla f(x)_i\right) \leq f(x) - \frac{1}{2L_i}\|\nabla f(x)_i\|^2
\]
This will be exactly what we need when choosing \( g_n \) and \( \delta_n \) in the scheme 5.1.

To obtain accelerated coordinate descent, we choose a coordinate \( i \) at random, where the probability of choosing coordinate \( i \) proportional to \( \sqrt{L_i} \). We then set
\[
g_n = \frac{1}{s\sqrt{L_i}}\nabla f(x'_n)_i
\]
and choose \( s \) so that \( E_n(g_n) = \nabla f(x'_n) \). This gives a step size of
\[
s = \left( \sum_{i=1}^{n} \sqrt{L_i} \right)^{-1}
\]
We now note that the required decrease in 5.2 is
\[
f(x_{n+1}) \leq f(x'_n) - \frac{s^2}{2}\|g_n\|^2 = f(x'_n) - \|sg_n\|^2 = f(x'_n) - \frac{1}{2L_i}\|\nabla f(x)_i\|^2
\]
So we must simply choose \( \delta_n \) so that

\[
(5.11) \quad f(x_n' - \delta_n) \leq f(x_n') - \frac{1}{2L_i} \| \nabla f(x)_i \|^2
\]

By the \( L_i \)-smoothness, the choice \( \delta_n = \frac{1}{L_i} \nabla f(x)_i \) will work. This recovers a version of the original accelerated coordinate descent in [6].

Another possible choice is to choose \( \delta_n = \frac{1}{L_j} \nabla f(x)_j \) where \( j \) is the coordinate which maximizes the decrease \( \frac{1}{2L_j} \| \nabla f(x)_j \|^2 \). This recovers the accelerated semi-greedy scheme presented in [5].

Finally, we would like to conclude by explaining why accelerated coordinate descent is very efficient if \( \nabla f(x)_i \) can be calculated using only a small number (say \( k \ll n \)) of entries \( x_{i_1}, ..., x_{i_k} \).

The issue is that each step of accelerated coordinate descent requires updating the full vectors \( x \) and \( v \). Namely in the first and second steps in 5.1 we update the whole vectors, while in the last two steps we only update the selected coordinate \( i \) (\( g_n \) and \( \delta_n \) are only non-zero in one coordinate).

The key observation is that we actually don’t need to update all of the \( x \) and \( v \) coordinates in each step. To see why, imagine for a moment that \( g_n = \delta_n = 0 \) in each step. Then the iteration becomes

\[
(5.12) \quad x_{n+1} = x_n + s v_n \\
v_{n+1} = (1 + s \sqrt{\alpha})^{-2} v_n
\]

which can be solved in closed form any number of iterations in the future. The key is to treat all of the coordinates in this manner between their selection times. When coordinate \( i \) is selected, this simple iteration must be modified to include \( \nabla f(x)_i \). However, until it is selected again, we can evaluate \( x_i \) and \( v_i \) at any iteration in closed form, as long as we know how many iterations have passed and what the values were the last time \( i \) was selected. This observation permits an efficient implementation of

5.1 as long as \( \nabla f(x)_j \) doesn’t depend on too many indices \( i_1, ..., i_k \). This idea can be used to construct fast solvers for sparse, symmetric, diagonally dominant linear systems; compare with the work in [4], for instance.

6. Conclusion. In was very remarkable to us that so many accelerated methods in the literature were discretizing the same underlying differential equations. What was even more interesting was that these differential equations had already been considered by the physics community in the context of electronic structure calculations, in [10]. In Appendix A we reproduce their argument, providing a detailed description of why this is the correct dynamics to consider, based on physics intuition and an analysis of quadratic objectives.

We believe that the ideas we have developed will help lead to the discovery of novel accelerated methods. In the future, we hope to use our general framework to derive and numerically test specialized accelerated first-order algorithms. Of particular interest are accelerated methods on manifolds. We believe the differential equation approach will prove important in understanding whether acceleration is possible in the presence of curvature.

Finally, we believe that our approach simplifies and clarifies the connections between the vast number of accelerated optimization methods in the literature and will help other researchers gain intuition about how they work and how to derive new ones.
7. Acknowledgements. We would like to thank Professors Russel Caflisch, Stanley Osher, and Jinchao Xu for their helpful suggestions and comments. This work was partially supported by AFOSR grant FA9550-15-1-0073.

References.
[1] Zeyuan Allen-Zhu and Lorenzo Orecchia. “Linear coupling: An ultimate unification of gradient and mirror descent”. In: arXiv preprint arXiv:1407.1537 (2014).
[2] Amir Beck and Marc Teboulle. “A fast iterative shrinkage-thresholding algorithm for linear inverse problems”. In: SIAM journal on imaging sciences 2.1 (2009), pp. 183–202.
[3] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. “A geometric alternative to Nesterov’s accelerated gradient descent”. In: arXiv preprint arXiv:1506.08187 (2015).
[4] Yin Tat Lee and Aaron Sidford. “Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems”. In: arXiv preprint arXiv:1305.1924 (2013).
[5] Haihao Lu, Robert M Freund, and Vahab Mirrokni. “Accelerating Greedy Coordinate Descent Methods”. In: arXiv preprint arXiv:1806.02476 (2018).
[6] Yu Nesterov. “Efficiency of coordinate descent methods on huge-scale optimization problems”. In: SIAM Journal on Optimization 22.2 (2012), pp. 341–362.
[7] Yurii Nesterov. “A method of solving a convex programming problem with convergence rate O (1/k^2)”. In: Soviet Mathematics Doklady. Vol. 27. 2. 1983, pp. 372–376.
[8] Yuerg Nesterov et al. Gradient methods for minimizing composite objective function. 2007.
[9] Weijie Su, Stephen Boyd, and Emmanuel Candes. “A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights”. In: Advances in Neural Information Processing Systems. 2014, pp. 2510–2518.
[10] Francesco Tassone, Francesco Mauri, and Roberto Car. “Acceleration schemes for ab initio molecular-dynamics simulations and electronic-structure calculations”. In: Physical Review B 50.15 (1994), p. 10561.
[11] Andre Wibisono, Aishia C Wilson, and Michael I Jordan. “A variational perspective on accelerated methods in optimization”. In: Proceedings of the National Academy of Sciences 113.47 (2016), E7351–E7358.

Appendix A. Intuition Behind the Dynamics. The intuition behind the dynamics in 2.2 comes from imagining a particle in a potential defined by the objective f. Without any damping (friction), the particle will oscillate freely in this potential and the total energy f(x) + 1/2||v||^2 is conserved. The damping term −γv in 2.2 causes the particle to lose energy so that it will eventually settle at the minimum of f.

We want to choose the damping rate γ so that the particle will settle in the minimum energy configuration as fast as possible. The optimal damping rate can be determined by analyzing the dynamics when the objective f is quadratic.

So let f be a quadratic objective. We diagonalize f and write
\[ x(t) = \sum_k x_k(t)w_k, \quad v(t) = \sum_k v_k(t)w_k \]
where Aw_k = λ_k w_k. The dynamics decouples across each of the eigendirections and we get
\[ \dot{x}_k(t) = v_k(t), \quad \dot{v}_k(t) = -\gamma v_k(t) - \lambda_k x_k(t) \]
a damped harmonic oscillator for each eigendirection.

Analyzing a damped harmonic oscillator is an undergraduate physics exercise. The characteristic polynomial of $A_2$ is

(A.3) $p(z) = z^2 + \gamma z + \lambda_k$

The roots of this polynomial, $z = \frac{1}{2}(-\gamma \pm \sqrt{\gamma^2 - 4\lambda_k})$, determine the behavior of the harmonic oscillator. The qualitative behavior of the system depends upon whether the characteristic polynomial has two real roots, a repeated real root, or two imaginary roots.

If $\gamma^2 - 4\lambda_k > 0$, the characteristic polynomial has two real roots. This is called the over damped regime, and the oscillator stays on the same side of equilibrium throughout the dynamics. The decay rate of the harmonic oscillator is dominated by the largest root, $z = \frac{1}{2}(-\gamma \pm \sqrt{\gamma^2 - 4\lambda_k})$.

If $\gamma^2 - 4\lambda_k < 0$, the characteristic polynomial has two imaginary roots. This is called the under damped regime, because the oscillator swings back and forth, losing energy in each oscillation. The damping is not strong enough to keep the oscillator on the same side of equilibrium. The decay rate in this regime is equal to the real part of the roots, $\text{Re}(z) = -\frac{1}{2}\gamma$.

If $\gamma^2 - 4\lambda_k = 0$, the characteristic polynomial has repeated real roots. This is called a critically damped harmonic oscillator. The oscillator stays on the same side of equilibrium but decays toward equilibrium as fast as possible. The decay rate is equal to the root $z = -\frac{1}{2}\gamma$.

Fixing $\lambda_k$, we see that the fastest decay rate possible is $-\sqrt{\lambda_k}$, which occurs for a critically damped harmonic oscillator. This follows since in the under damped regime the decay rate is $-\frac{1}{2}\gamma$, with $\gamma < 2\sqrt{\lambda_k}$, and in the over damped regime we have

(A.4) $\gamma^2 - 4\lambda_k = (\gamma + 2\sqrt{\lambda_k})(\gamma - 2\sqrt{\lambda_k}) > (\gamma - 2\sqrt{\lambda_k})^2$

The last inequality occurs because $\gamma > 2\sqrt{\lambda_k}$. This means that

$$\sqrt{\gamma^2 - 4\lambda_k} > \gamma - 2\sqrt{\lambda_k}$$

and so for the larger real root

$$z = \frac{1}{2}(-\gamma + \sqrt{\gamma^2 - 4\lambda_k}) > -\sqrt{\lambda_k}$$

Now, in the dynamics $2.2$, we must choose a single damping rate $\gamma$ for all eigenvalues $\lambda_k$. In order to obtain the fastest possible decay rate, we want to maximize the slowest decay rate among the $\lambda_k$.

Notice that the fastest this decay rate can be is $-\sqrt{\lambda_1}$, since the harmonic oscillator corresponding to the smallest eigenvalue cannot decay any faster. Moreover, this decay rate is achieved when $\gamma$ is chosen so that the mode corresponding to $\lambda_1$ is critically damped, because then all other modes will be under damped and will also decay at the rate $-\frac{1}{2}\gamma$.

What about the step size required for a stable discretization of the dynamics? Just as we chose a single damping rate for all eigenmodes, we must choose a single step size for all of the eigenvalues $\lambda_k$. If we use an integrator whose region of stability contains the negative unit semicircle (such as Runge-Kutta 4), then the discretization of the harmonic oscillator corresponding to $\lambda_k$ will be stable if

$$\frac{1}{\Delta t} \geq \max\{|z_1|, |z_2|\}$$
where \( z_1 \) and \( z_2 \) are the (possibly complex) roots of the characteristic equation. Since the optimal damping parameter is \( 2\sqrt{\alpha} \), all of the modes are either critically damped or under damped. So the roots \( z_i \) are complex and we calculate

\[
|z_1| = |z_2| = \frac{1}{2} \sqrt{\gamma^2 + 4\lambda_k - \gamma^2} = \sqrt{\lambda_k} \leq \sqrt{\lambda_n}
\]

So we see that the dynamics decays with exponential rate \( -\sqrt{\lambda_1} \) and the step size required for a stable discretization is \( \Delta t < 1/\sqrt{\lambda_n} \). This explains why discretizing 2.2 produces a method which requires \( O(\sqrt{k}) \) iterations to converge.

Indeed, for the quadratic problem we can use any integrator whose region of stability contains the negative unit semicircle to obtain an accelerated method. The resulting methods are related to the Chebyshev method for solving a positive definite linear system.