
On the Curved Geometry of Accelerated Optimization

Aaron Defazio
Facebook AI Research
New York

Abstract

In this work we propose a differential geometric motivation for Nesterov’s accelerated gradient method (AGM) for strongly-convex problems. By considering the optimization procedure as occurring on a Riemannian manifold with a natural structure, The AGM method can be seen as the proximal point method applied in this curved space. This viewpoint can also be extended to the continuous time case, where the accelerated gradient method arises from the natural block-implicit Euler discretization of an ODE on the manifold. We provide an analysis of the convergence rate of this ODE for quadratic objectives.

1 Introduction

The core algorithms of convex optimization are gradient descent (GD) and the accelerated gradient method (AGM). These methods are rarely used directly, more often they occur as the building blocks for distributed, composite, or non-convex optimization. In order to build upon these components, it is helpful to understand not just *how* they work, but *why*. The gradient method is well understood in this sense. It is commonly viewed as following a direction of steepest descent or as minimizing a quadratic upper bound. These interpretations provide a motivation for the method as well as suggesting a potential convergence proof strategy.

The accelerated gradient method in contrast has an identity crisis. Its equational form is remarkably malleable, allowing for many different ways of writing the same updates. We list a number of these forms in Table 1. Nesterov’s original motivation for the AGM method used the concept of estimate sequences. Unfortunately, estimate sequences do not necessarily yield the simplest accelerated methods when generalized, such as for the composite case (Beck and Teboulle 2009, Nesterov 2007), and they have not been successfully applied in the important finite-sum (variance reduced) optimization setting.

Because of the complexity of estimate sequences, the AGM method is commonly motivated as a form of momentum. This view is flawed as a way of introducing the AGM method from first principles, as the most natural way of using momentum yields the heavy ball method instead:

$$x^{k+1} = x^k - \gamma \nabla f(x^k) + \beta (x^k - x^{k-1}),$$

which arises from discretizing the physics of a particle in a potential well with additional friction. The heavy-ball method does not achieve an accelerated convergence rate on general convex problems, suggesting that momentum, *per se*, is not the reason for acceleration. Another contemporary view is the linear-coupling interpretation of Allen-Zhu and Orecchia [2017], which views the AGM method as an interpolation between gradient descent and mirror descent. We take a more geometric view in our interpretation.

In this work we motivate the AGM by introducing it as an application of the proximal-point method:

$$x^k = \arg \min_x \left\{ f(x) + \frac{\eta}{2} \|x - x^{k-1}\|^2 \right\}.$$

The proximal point (PP) method is perhaps as foundational as the gradient descent method, although it sees even less direct use as each step requires solving a regularized subproblem, in contrast to the closed form steps for GD and AGM. The PP method gains remarkable convergence rate properties in exchange for the computational difficulty, including convergence for any positive step-size.

We construct the AGM by applying a dual form of the proximal point method in a curved space. Each step follows a geodesic on a manifold in a sense we make precise in Section 4. We use the term curved with respect to a coordinate system, rather than a coordinate free notion of curvature such as the Riemannian curvature. We first give a brief introduction to the concepts from differential geometry necessary to understand our motivation. The equational form that our argument yields is much closer to those that have been successfully applied in practice, particularly for the minimization of finite sums [Lan and Zhou, 2017, Zhang and Xiao, 2017].

2 Connections

An (*affine*) *connection* is a type of structure on a manifold that can be used to define and compute geodesics. Geodesics in this sense represent curves of zero acceleration. These geodesics are more general concepts than Riemannian geodesics induced by the Riemannian connection, not necessarily representing the shortest path in any metric. Indeed, we will define multiple connections on the same manifold that lead to completely different geodesics.

Given a n dimensional coordinate system, a connection is defined by n^3 numbers at every point x on the manifold, called the connection coefficients (or Christoffel symbols) $\Gamma_{ij}^k(x)$. A geodesic is a path $\gamma : [0, 1] \rightarrow \mathcal{M}$ (in our case $\mathcal{M} = \mathbb{R}^n$) between two points x and y can then be computed as the unique solution $\gamma(t) = x(t)$ to the system of ordinary differential equations [Lee, 1997, Page 58, Eq 4.11]:

$$\frac{d^2 \gamma^i}{dt^2} = \frac{d^2 x^i}{dt^2} + \sum_{j,k} \Gamma_{jk}^i(x) \frac{dx^j}{dt} \frac{dx^k}{dt} = 0,$$

with boundary conditions $x(0) = x$ and $x(1) = y$. Here x^i denotes the i th component of x expressed in the same coordinate system as the connection.

3 Divergences induce Hessian manifold structure

Let ϕ be a smooth strongly convex function defined on \mathbb{R}^n . The Bregman divergence generated by ϕ :

$$B_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle,$$

and its derivatives can be used to define a remarkable amount of structure on the domain of ϕ . In particular, we can define a *Riemannian manifold, together with two dually flat connections with biorthogonal coordinate systems* [Amari and Nagaoka, 2000, Shima, 2007]. This structure is also known as a Hessian manifold. Topologically it is $\mathcal{M} = \mathbb{R}^n$ with the following additional geometric structures.

Riemannian structure

Riemannian manifolds have the additional structure of a *metric tensor* (a generalized dot-product), defined on their tangent spaces. We denote the vector space of tangent vectors at a point x as $T_x \mathcal{M}$. If we express the tangent vectors with respect to the Euclidean basis, the metric at a point x is a quadratic form with the Hessian matrix $H(x) = \nabla_x^2 B(x, y) = \nabla^2 \phi(x)$ of ϕ at x :

$$g_x(u, v) = u^T H(x) v.$$

Biorthogonal coordinate systems

Central to the notion of a manifold is the invariance to the choice of coordinate system. We can express a point on the manifold as well as a point in the tangent space using any coordinate system that is most convenient. Of course, when we wish to perform calculations on the manifold we must be careful to express all quantities in that coordinate system. Euclidean coordinates e_i are the most natural on our Hessian manifold, however there is another coordinate system which is naturally dual to e_i , and ties the manifold structure directly to duality theory in optimization.

Table 1: Equivalent forms of Nesterov’s method for μ -strongly convex, L -smooth f . Proofs of the stated relations are available in the appendix.

Form Name	Algorithm	Relations
Nesterov [2013] form I	$y^k = \frac{\alpha\gamma v^k + \gamma x^k}{\alpha\mu + \gamma}$ $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k),$ $v^{k+1} = (1 - \alpha)v^k + \frac{\alpha\mu}{\gamma}y^k - \frac{\alpha}{\gamma}\nabla f(y^k)$	$\alpha_{\text{Nes}} = \sqrt{\mu/L}$ $\gamma_{\text{Nes}} = \mu.$
Nesterov [2013] form II	$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k),$ $y^{k+1} = x^{k+1} + \beta(x^{k+1} - x^k)$	$\beta_{\text{Nes}} = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$
Sutskever et al. [2013]	$p^{k+1} = \beta p^k - \frac{1}{L}\nabla f(x^k + \beta p^k),$ $x^{k+1} = x^k + p^{k+1}$	$p_{\text{Sut}}^{k+1} = x_{\text{Nes}}^{k+1} - x_{\text{Nes}}^k,$ $y_{\text{Nes}}^k = x_{\text{Sut}}^k + \beta p_{\text{Sut}}^k.$
Modern Momentum ¹	$p^{k+1} = \beta p^k + \nabla f(x^k),$ $x^{k+1} = x^k - \frac{1}{L}(\nabla f(x^k) + \beta p^{k+1}).$	$x_{\text{mod}}^k = x_{\text{Sut}}^k + \beta p_{\text{Sut}}^k = y_{\text{Nes}}^k,$ $p_{\text{mod}}^k = -L p_{\text{Sut}}^k.$
Auslender and Teboulle [2006]	$y^k = (1 - \theta)\hat{x}^k + \theta z^k,$ $z^{k+1} = z^k - \frac{\gamma}{\theta}\nabla f(y^k),$ $\hat{x}^k = (1 - \theta)\hat{x}^k + \theta z^{k+1}.$	$\theta_{\text{AT}} = 1 - \beta_{\text{Nes}},$ $\hat{x}_{\text{AT}}^k = x_{\text{Nes}}^k,$ $y_{\text{AT}}^k = y_{\text{Nes}}^k = x_{\text{mod}}^k,$ $\gamma_{\text{AT}} = 1/L.$
Lan and Zhou [2017]	$\hat{x}^k = \alpha(x^{k-1} - x^{k-2}) + x^{k-1},$ $\underline{x}^k = \frac{\hat{x}^k + \tau \underline{x}^{k-1}}{1 + \tau},$ $g^k = \nabla f(\underline{x}^k),$ $x^k = x^{k-1} - \frac{1}{\eta}g^k.$	$x_{\text{Lan}}^k = z_{\text{AT}}^k,$ $\underline{x}_{\text{Lan}}^k = y_{\text{AT}}^k,$ $\eta_{\text{Lan}} = \frac{\gamma_{\text{AT}}}{\theta_{\text{AT}}},$ $\tau_{\text{Lan}} = \frac{1 - \theta_{\text{AT}}}{\theta_{\text{AT}}},$ $\alpha_{\text{Lan}} = 1 - \theta_{\text{AT}}.$

Recall that for a convex function ϕ we may define the convex conjugate $\phi^*(y) = \max_x \{ \langle x, y \rangle - \phi(x) \}$. The dual coordinate system we define simply identifies each point x , when expressed in Euclidean (“primal”) coordinates, with the vector of “dual” coordinates:

$$y = \nabla \phi(x).$$

Our assumptions of smoothness and strong convexity imply this is a one-to-one mapping, with inverse given by $x = \nabla \phi^*(y)$. The remarkable fact that the gradient of the conjugate is the inverse of the gradient is a key building block of the theory in this paper.

The notion of *biorthogonality* refers to natural tangent space coordinates of these two systems. A tangent vector v at a point x can be converted to a vector u of dual (tangent space) coordinates using matrix multiplication with the Hessian [Shima, 2007]:

$$u = H(x)v, \tag{1}$$

Given the definition of the metric above, it is easy to see that if we have two vectors v_1 and v_2 , we may express v_2 in dual coordinates u_2 so that the metric tensor takes the simple form:

$$g_x(v_1, v_2) = v_1^T H(x) v_2 = v_1^T H(x) (H(x)^{-1} u_2) = v_1^T u_2,$$

which is the *biorthogonal* relation between the two tangent space coordinate systems.

Dual Flat Connections

There is an obvious connection $\Gamma^{(E)}$ we can apply to the Hessian manifold, the Euclidean connection that trivially identifies straight lines in \mathbb{R}^n as geodesics. Normally when we perform gradient descent

¹PyTorch & Tensorflow (for instance) implement this form. Evaluating the gradient and function at the current iterate x^k instead of a shifted point makes it more consistent with gradient descent when wrapped in a generic optimization layer.

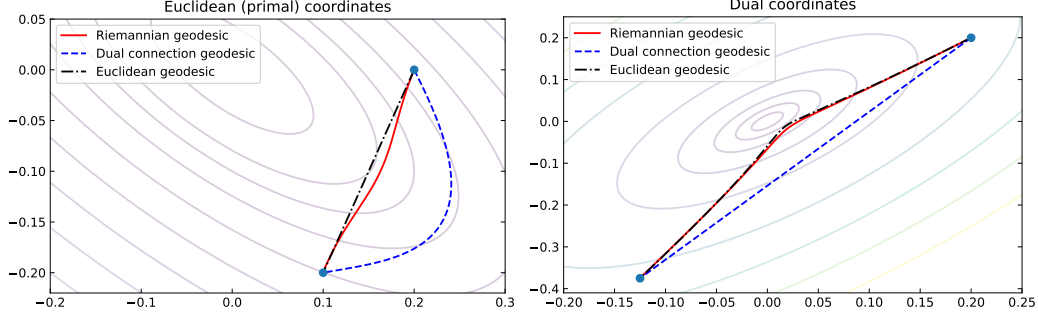


Figure 1: Illustrative geodesics for $f(x) = \frac{1}{4} \|Ax\|^4$, with $A = [2, 1; 1, 3]$. Viewing them from both coordinate systems highlights the duality. Contour lines are for f and f^* respectively.

in \mathbb{R}^n we are implicitly following a geodesic of this connection. The connection coefficients $\Gamma_{ij}^{(E)k}$ are all zero when this connection is expressed in Euclidean coordinates. A connection that has $\Gamma_{ij}^k = 0$ with respect to some coordinate system is a *flat* connection.

The Hessian manifold admits another flat connection, which we will call the dual connection, as it corresponds to straight lines in the dual coordinate system established above. In particular each dual geodesic can be expressed in primal coordinates $\gamma(t)$ as a solution to the equation:

$$\nabla\phi(\gamma(t)) = at + b,$$

for vectors a, b representing the initial velocity and point respectively (both represented in dual coordinates) that depend on the boundary conditions. This is quite easy to solve using the relation $\nabla\phi^{-1} = \nabla\phi^*$ discussed above. For instance, a geodesic $\gamma : [0, 1] \rightarrow \mathcal{M}$ between two arbitrary points x and y under the dual connection could be computed explicitly in Euclidean coordinates as:

$$\gamma(t) = \nabla\phi^*(t\nabla\phi(y) + (1-t)\nabla\phi(x)).$$

If we instead know the initial velocity we can find the endpoint with:

$$y = \nabla\phi^*(\nabla\phi(x) + H(x^k)v). \quad (2)$$

The flatness of the dual connection $\Gamma^{(D)}$ is crucial to its computability in practice. If we instead try to compute the geodesic in Euclidean coordinates using the geodesic ODE, we have to work with the connection coefficients which involve third derivatives of ϕ (taking the form of double those of the Riemannian connection $\Gamma^{(R)}$):

$$\Gamma_{ij}^{(D)k}(x) = 2\Gamma_{ij}^{(R)k} = [H(x)^{-1}(\nabla H(x))_i]_{kj},$$

The Riemannian connection's geodesics are similarly difficult to compute directly from the ODE (they also can't generally be expressed in a simpler form).

4 Bregman proximal operators follow geodesics

Bregman divergences arise in optimization primarily through their use in proximal steps. A Bregman proximal operation balances finding a minimizer of a given function f with maintaining proximity to a given point y , measured using a Bregman divergence instead of a distance metric:

$$x^k = \arg \min_x \{f(x) + \rho B_\phi(x, x^{k-1})\}. \quad (3)$$

A core application of this would be the mirror descent step [Nemirovski and Yudin, 1983, Beck and Teboulle, 2003], where the operation is applied to a linearized version of f instead of f directly:

$$x^k = \arg \min_x \{\langle x, \nabla f(x^{k-1}) \rangle + \rho B_\phi(x, x^{k-1})\}.$$

Bregman proximal operations can be interpreted as geodesic steps with respect to the dual connection. The key idea is that given an input point x^{k-1} , they output a point x such that the velocity of the connecting geodesic is equal to $-\nabla_{\frac{1}{\rho}} f(x)$ at x . This velocity is measured in the flat coordinate system of the connection, the dual coordinates. To see why, consider a geodesic $\gamma(t) = (1 -$

$t)\nabla\phi(x^{k-1}) + t\nabla\phi(x^k)$. Here x^{k-1} and x^k are in primal coordinates and $\gamma(t)$ is in dual coordinates. The velocity is $\frac{d}{dt}\gamma(t) = \nabla\phi(x^k) - \nabla\phi(x^{k-1})$. Contrast to the optimality condition of the Bregman prox (Equation 3):

$$-\frac{1}{\rho}\nabla f(x^k) = \nabla\phi(x^k) - \nabla\phi(x^{k-1}).$$

For instance, when using the Euclidean penalty the step is:

$$x^k = \arg \min_x \left\{ f(x) + \frac{\rho}{2} \|x - x^{k-1}\|^2 \right\}.$$

The final velocity is just $x^k - x^{k-1}$, and so $x^k - x^{k-1} = -\frac{1}{\rho}\nabla f(x^k)$, which is the solution of the proximal operation.

5 Primal-Dual form of the proximal point method

The proximal point method is the building block from which we will construct the accelerated gradient method. Consider the basic form of the proximal point method applied to a strongly convex function f . At each step, the iterate x^k is constructed from x^{k-1} by solving the proximal operation subproblem given an inverse step size parameter η :

$$x^k = \arg \min_x \left\{ f(x) + \frac{\eta}{2} \|x - x^{k-1}\|^2 \right\}. \quad (4)$$

This step can be considered an implicit form of the gradient step, where the gradient is evaluated at the end-point of the step instead of the beginning:

$$x^k = x^{k-1} - \frac{1}{\eta}\nabla f(x^k),$$

which is just the optimality condition of the subproblem in Equation 4, found by taking the derivative $\nabla f(x) + \eta x - \eta x^{k-1}$ to be zero. A remarkable property of the proximal operation becomes apparent when we rearrange this formula, namely that the solution to the operation is not a single point but a *primal-dual pair*, whose weighted sum is equal to the input point:

$$x^k + \frac{1}{\eta}\nabla f(x^k) = x^{k-1}.$$

If we define $g^k = \nabla f(x^k)$, the primal-dual pair obeys a duality relation: $g^k = \nabla f(x^k)$ and $x^k = \nabla f^*(g^k)$, which allows us to interchange primal and dual quantities freely. Indeed we may write the condition in a dual form as:

$$\nabla f^*(g^k) + \frac{1}{\eta}g^k = x^{k-1}, \quad (5)$$

which is the optimality condition for the proximal operation:

$$g^k = \arg \min_g \left\{ f^*(g) + \frac{1}{2\eta} \|g - \eta x^{k-1}\|^2 \right\}.$$

Our goal in this section is to express the proximal point method in terms of a dual step, and while this equation involves the dual function f^* , it is not a *step* in the sense that g^k is formed by a proximal operation from g^{k-1} .

We can manipulate this formula further to get an update of the form we want, by simply adding and subtracting g^{k-1} from 5:

$$\nabla f^*(g^k) + \frac{1}{\eta}g^k = \frac{1}{\eta}g^{k-1} + \left(x^{k-1} - \frac{1}{\eta}g^{k-1} \right),$$

Which gives the updates:

$$\begin{aligned} g^k &= \arg \min_g \left\{ f^*(g) - \left\langle g, x^{k-1} - \frac{1}{\eta}g^{k-1} \right\rangle + \frac{1}{2\eta} \|g - g^{k-1}\|^2 \right\}, \\ x^k &= x^{k-1} - \frac{1}{\eta}g^k. \end{aligned}$$

We call this the primal-dual form of the proximal point method.

6 Acceleration as a change of geometry

The proximal point method is rarely used in practice due to the difficulty of computing the solution to the proximal subproblem. It is natural then to consider modifications of the subproblem to make it more tractable. The subproblem becomes particularly simple if we replace the proximal operation with a Bregman proximal operation with respect to f^* ,

$$g^k = \arg \min_g \left\{ f^*(g) - \left\langle g, x^{k-1} - \frac{1}{\eta} g^{k-1} \right\rangle + \tau B_{f^*}(g, g^{k-1}) \right\}.$$

We have additionally changed the penalty parameter to a new constant τ , which is necessary as the change to the Bregman divergence changes the scaling of distances. We discuss this further below.

Recall from Section 4 that Bregman proximal operations follow geodesics. The key idea is that we are now following a geodesic in the dual connection of $\phi = f^*$, using the notation of Section 3, which is a *straight-line in the primal coordinates* of f due to the flatness of the connection (Section 3). Due to the flatness property, a simple closed-form solution can be derived by equating the derivative to 0:

$$\begin{aligned} \nabla f^*(g^k) - \left[x^{k-1} - \frac{1}{\eta} g^{k-1} \right] + \tau \nabla f^*(g^k) - \tau \nabla f^*(g^{k-1}) &= 0, \\ \text{therefore } g^k &= \nabla f \left((1 + \tau)^{-1} \left[x^{k-1} - \frac{1}{\eta} g^{k-1} + \tau \nabla f^*(g^{k-1}) \right] \right). \end{aligned}$$

This formula gives g^k in terms of the derivative of known quantities, as $\nabla f^*(g^{k-1})$ is known from the previous step as the point at which we evaluated the derivative at. We will denote this argument to the derivative operation y , so that $g^k = \nabla f(y^k)$. It no longer holds that $g^k = \nabla f(x^k)$ after the change of divergence. Using this relation, y can be computed each step via the update:

$$y^k = \frac{x^{k-1} - \frac{1}{\eta} g^{k-1} + \tau y^{k-1}}{1 + \tau}.$$

In order to match the accelerated gradient method exactly we need some additional flexibility in the step size used in the y^k update. To this end we introduce an additional constant α in front of g^{k-1} , which is 1 for the proximal point variant. The full method is as follows:

Bregman form of the accelerated gradient method

$$\begin{aligned} y^k &= \frac{x^{k-1} - \frac{\alpha}{\eta} g^{k-1} + \tau y^{k-1}}{1 + \tau}, \\ g^k &= \nabla f(y^k), \\ x^k &= x^{k-1} - \frac{1}{\eta} g^k. \end{aligned} \tag{6}$$

This is very close to the equational form of Nesterov's method explored by Lan and Zhou [2017], with the change that they assume an explicit regularizer is used, whereas we assume strong convexity of f . Indeed we have chosen our notation so that the constants match. This form is algebraically equivalent to other known forms of the accelerated gradient method for appropriate choice of constants. Table 1 shows the direct relation between the many known ways of writing the accelerated gradient method in the strongly-convex case (Proofs of these relations are in the Appendix). When f is μ -strongly convex and L -smooth, existing theory implies an accelerated geometric convergence rate of at least $1 - \sqrt{\frac{\mu}{L}}$ for the parameter settings [Nesterov, 2013]:

$$\eta = \sqrt{\mu L}, \quad \tau = \frac{L}{\eta}, \quad \alpha = \frac{\tau}{1 + \tau}.$$

In contrast, the primal-dual form of the proximal point method achieves at least that convergence rate for parameters:

$$\eta = \sqrt{\mu L}, \quad \tau = \frac{1}{\eta}, \quad \alpha = 1.$$

The difference in τ arises from the difference in the scaling of the Bregman penalty compared to the Euclidean penalty. The Bregman generator f^* is strongly convex with constant $1/L$ whereas the Euclidean generator $\frac{1}{2} \|\cdot\|^2$ is strongly convex with constant 1, so the change in scale requires rescaling by L .

6.1 Interpretations

After the change in geometry, the g update no longer gives a dual point that is directly the gradient of the primal iterate. However, notice that the term we are attempting to minimize in the g step:

$$f^*(g) - \langle g, x^{k-1} - \frac{\alpha}{\eta} g^{k-1} \rangle,$$

has a fixed point of $\nabla f^*(g^k) = x^{k-1} - \frac{\alpha}{\eta} g^k$, which is precisely an α -weight version of the proximal point's key property from Equation 5. Essentially we have relaxed the proximal-point method. Instead of this relation holding precisely at every step, we are instead constantly taking steps which pull g closer to satisfying it.

6.2 Inertial form

The primal-dual view of the proximal point method can also be written in terms of the quantity $z^{k-1} = x^{k-1} - \frac{\alpha}{\eta} g^{k-1}$ instead of x^{k-1} . This form is useful for the construction of ODEs that model the discrete dynamics. Under this change of variables the updates are:

$$\begin{aligned} g^k &= \arg \min_g \left\{ f^*(g) - \langle g, z^{k-1} \rangle + \frac{1}{2\eta} \|g - g^{k-1}\|^2 \right\}, \\ z^k &= z^{k-1} - \frac{1}{\eta} g^k - \frac{\alpha}{\eta} (g^k - g^{k-1}). \end{aligned} \quad (7)$$

6.3 Relation to the heavy ball method

Consider Equation 6 with $\alpha = 0$, which removes the over-extrapolation before the proximal operation. If we define $\beta = \frac{\tau}{1+\tau}$ we may write the method as:

$$x^k = x^{k-1} - \frac{1}{\eta} f'(y^{k-1}), \quad y^k = \beta y^{k-1} + (1 - \beta) x^k.$$

We can eliminate x^k from the y^k update above by plugging in the x^k step equation, then using the y^k update from the previous step in the form $(1 - \beta) x^{k-1} = y^{k-1} - \beta y^{k-2}$:

$$\begin{aligned} y^k &= \beta y^{k-1} + (1 - \beta) \left(x^{k-1} - \frac{1}{\eta} f'(y^{k-1}) \right) \\ &= \beta y^{k-1} - (1 - \beta) \frac{1}{\eta} f'(y^{k-1}) + [y^{k-1} - \beta y^{k-2}] \\ &= y^{k-1} - (1 - \beta) \frac{1}{\eta} f'(y^{k-1}) + \beta [y^{k-1} - y^{k-2}]. \end{aligned}$$

This has the exact form of the heavy ball method with step size $(1 - \beta) / \eta$ and momentum β . We can also derive the heavy ball method by starting from the saddle-point expression for f :

$$\min_x f(x) = \min_x \max_g \{ \langle x, g \rangle - f^*(g) \}.$$

The alternating-block gradient descent/ascent method on the objective $\langle x, g \rangle - f^*(g)$ with step-size γ is simply:

$$g^k = g^{k-1} + \frac{1}{\gamma} [x^{k-1} - \nabla f^*(g^{k-1})], \quad x^k = x^{k-1} - \gamma g^k.$$

If we instead perform a Bregman proximal update in the dual geometry for the g part, we arrive at the same equations as we had for the primal-dual proximal point method but with $\alpha = 0$, yielding the heavy ball method. In order to get the accelerated gradient method instead of the heavy ball method, the extra inertia that arises from starting from the proximal point method instead of the saddle point formulation appears to be crucial.

7 Dual geometry in continuous time

The inertial form (Equation 7) of the proximal point method can be formulated as an ODE in a very natural way, by mapping $z^k - z^{k-1} \rightarrow \dot{z}$ and $g^k - g^{k-1} \rightarrow \dot{g}$, and taking x and g to be at time t .

This is the inverse of the Euler class of discretizations applied separately to the two terms, which is the most natural way to discretize an ODE. The resulting proximal point ODE is:

$$\begin{aligned}\dot{g} &= f_g(z, g, t) \doteq -\frac{1}{\tau} \nabla f^*(g) + \frac{1}{\tau} z, \\ \dot{z} &= f_z(z, g, t) \doteq -\frac{1}{\eta} g - \frac{\alpha}{\eta} \dot{g}.\end{aligned}$$

We have suppressed the dependence on t of each quantity for notational simplicity. We can treat g more formally as a point $g \in \mathcal{M}$ on a Hessian manifold \mathcal{M} . Then the solution for the g variable of the ODE is a curve $\gamma(t) : I \rightarrow \mathcal{TM}$ from an interval I to the tangent bundle on the manifold so the velocity $\dot{\gamma}(t) \in T_g \mathcal{M}$ obeys the ODE: $\dot{\gamma}(t) = f_g(z, g, t)$. The right hand side of the ODE is a point in the tangent space of the manifold at $\gamma(t)$, expressed in Euclidean coordinates.

We can now apply the same partial change of geometry that we used in the discrete case. We will consider the quantity $f_g(z, g, t)$ to be a tangent vector in dual tangent space coordinates. For the $\phi = f^*$ Hessian manifold, instead of its primal tangent space coordinates (which would leave the ODE unchanged). The variable g remains in primal coordinates with respect to ϕ , so we must add to the ODE a change of coordinates for the tangent vector, yielding:

$$\dot{g} = (\nabla^2 f^*(g))^{-1} f_g(z, g, t),$$

where we have used the inverse of Equation 1, with $\phi = f^*$. We can rewrite this as:

$$f_g(z, g, t) = \nabla^2 f^*(g) \dot{g} = \frac{d}{dt} \nabla f^*(g),$$

giving the AGM ODE system:

$$\frac{d}{dt} \nabla f^*(g) = -\frac{1}{\tau} \nabla f^*(g) + \frac{1}{\tau} z, \quad \dot{z} = -\frac{1}{\eta} g - \frac{\alpha}{\eta} \dot{g}.$$

It is now easily seen that the implicit Euler update for the g variable with z fixed now corresponds to the solution of the Bregman proximal operation considered in the discrete case. So this ODE is a natural continuous time analogue to the accelerated gradient method.

Convergence in continuous time

The natural analogy to convergence in continuous time is known as the decay rate of the ODE. A sufficient condition for an ODE with parameters $u = [z; g]$ to decay with constant ρ is:

$$\|u(t) - u^*\| \leq \exp(-t\rho) \|u(0) - u^*\|,$$

where u^* is a fixed point. We can relate this to the discrete case by noting that $\exp(-t\rho) = \lim_{k \rightarrow \infty} (1 - \frac{t}{k}\rho)^k$, so given our discrete-time convergence rate is proportional to $(1 - \sqrt{\mu/L})^k$, we would expect values of ρ proportional to $\sqrt{\mu/L}$ if the ODE behaves similarly to the discrete process. We have been able to establish this result for both the proximal and AGM ODEs for quadratic objectives (proof in the Appendix in the supplementary material).

Theorem 1. *The proximal and AGM ODEs decay with at least the following rates for μ -strongly convex and L -smooth quadratic objective functions when using the same hyper-parameters as in the discrete case:*

$$\rho_{\text{prox}} \geq \frac{\sqrt{\mu}}{\sqrt{\mu} + \sqrt{L}}, \quad \rho_{\text{AGM}} \geq \frac{1}{2} \sqrt{\frac{\mu}{L}}.$$

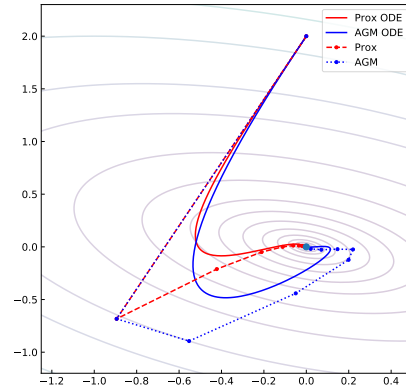


Figure 2: Paths for the quadratic problem $f(x) = \frac{1}{2} x^T A x$ with $A = [2, 1; 1, 3]$.

Figure 2 contrasts the convergence of the discrete and continuous variants. The two methods have quite distinct paths whose shape is shared by their ODE counterparts.

8 Related Work

The application of Bregman divergence to the analysis of continuous time views of the accelerated gradient method has recently been explored by Wibisono et al. [2016] and Wilson et al. [2018]. Their approaches do not use the Bregman divergence of f^* , a key factor of our approach. The Bregman divergence of a function ϕ occurs explicitly as a term in a Hamiltonian, in contrast to our view of ϕ as curving space. The accelerated gradient method has been shown to be modeled by a momentum of the form ODE $\ddot{X} + c(t)\dot{X} + \nabla f(x) = 0$ by Su et al. [2014]. Natural discretizations of their ODE result in the heavy-ball method instead of the accelerated gradient method, unlike our form which can produce both based on the choice of α . The fine-grained properties of momentum ODEs have also been studied in the quadratic case by Scieur et al. [2017].

A primal-dual form of the regularized accelerated gradient method appears in Lan and Zhou [2017]. Our form can be seen as a special case of their form when the regularizer is zero. Our work extends theirs, providing an understanding of the role that geometry plays in unifying acceleration and implicit steps.

The Riemannian connection induced by a function has been heavily explored in the optimization literature as part of the natural gradient method [Amari, 1998], although other connections on this manifold are less explored. The dual-flat connections have primarily seen use in the information-geometry setting for optimization over distributions [Amari and Nagaoka, 2000].

The accelerated gradient method is not the only way to achieve accelerated rates among first order methods. Other techniques include the Geometric descent method of Bubeck et al. [2015], where a bounding ball is updated at each step that encloses two other balls, a very different approach. The method described by Nemirovski and Yudin [1983] is also notable as being closer to the conjugate gradient method than other accelerated approaches, but at the expense of requiring a 2D search instead of a 1D line search at each step.

9 Conclusion

We believe the tools of differential geometry may provide a new and insightful avenue for the analysis of accelerated optimization methods. The analysis we provide in this work is a first step in this direction. The advantage of the differential geometric approach is that it provides high level tools that make the derivation of acceleration easier to state. This derivation, from the proximal point method to the accelerated gradient method, is in our opinion not nearly as mysterious as the other known approaches to understanding acceleration.

References

- Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science*, ITCS '17, 2017. Full version available at <http://arxiv.org/abs/1407.1537>.
- Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.
- Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.
- Alfred Auslender and Marc Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.
- Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 2003.
- Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. IMAGING SCIENCES*, 2009.
- Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*. Society for Industrial and Applied Mathematics, 1994.

- Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov’s accelerated gradient descent, 2015.
- Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, pages 1–49, 2017.
- John M. Lee. *Riemannian Manifolds : An introduction to curvature*. Springer-Verlag, 1997.
- Arkadi Nemirovski and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
- Yu. Nesterov. Accelerating the cubic regularization of newton’s method on convex problems. *Mathematical Programming*, 2008.
- Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.
- Yurii Nesterov. Gradient methods for minimizing composite objective function,. Technical report, CORE, 2007.
- Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- Ernest K Ryu and Stephen Boyd. Primer on monotone operator methods. *Appl. Comput. Math*, 15(1): 3–43, 2016.
- Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d’Aspremont. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems 30*, 2017.
- Hirohiko Shima. *The Geometry of Hessian Structures*. World Scientific, 2007.
- Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights. *Advances in Neural Information Processing Systems*, 2014.
- Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
- Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.
- Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2018.
- Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

Appendix

A Reformulations of the accelerated gradient method

Form II

This simplification is described in Nesterov [2013] which we reproduce for completeness. Recall that Form I is given by the updates:

$$\begin{aligned} y^k &= \frac{\alpha\gamma v^k + \gamma x^k}{\alpha\mu + \gamma}, \\ x^{k+1} &= y^k - \frac{1}{L} \nabla f(y^k), \\ v^{k+1} &= (1 - \alpha) v^k + \frac{\alpha\mu}{\gamma} y^k - \frac{\alpha}{\gamma} \nabla f(y^k). \end{aligned}$$

Nesterov specifies the requirement that $\gamma = (1 - \alpha)\gamma + \alpha\mu$. When $\alpha = \sqrt{\mu/L}$ then γ must then satisfy:

$$\gamma = (1 - \sqrt{\mu/L})\gamma + \mu\sqrt{\mu/L},$$

$$\therefore 1 = (1 - \sqrt{\mu/L}) + \frac{\mu}{\gamma} \sqrt{\mu/L},$$

$$\therefore \sqrt{\mu/L} = \frac{\mu}{\gamma} \sqrt{\mu/L},$$

$$\therefore \mu = \gamma$$

We may rewrite the y^k definition as:

$$(\alpha\mu + \gamma) y^k = \alpha\gamma v^k + \gamma x^k,$$

$$\therefore v^k = \frac{1}{\alpha\gamma} [(\alpha\mu + \gamma) y^k - \gamma x^k].$$

Plugging this into the v step:

$$\begin{aligned} v_{k+1} &= (1 - \alpha) v^k + \frac{\alpha\mu}{\gamma} y^k - \frac{\alpha}{\gamma} \nabla f(y^k) \\ &= \frac{1 - \alpha}{\alpha\gamma} [(\alpha\mu + \gamma) y^k - \gamma x^k] + \frac{\alpha\mu}{\gamma} y^k - \frac{\alpha}{\gamma} \nabla f(y^k) \\ &= \frac{1}{\gamma} \left[(1 - \alpha) \mu y^k + \left(\frac{1 - \alpha}{\alpha} \right) \gamma y^k + \alpha \mu y^k \right] - \frac{1 - \alpha}{\alpha} x^k - \frac{\alpha}{\gamma} \nabla f(y^k) \\ &= \frac{1}{\alpha\gamma} [\alpha\mu y^k + (1 - \alpha) \gamma y^k] - \frac{1 - \alpha}{\alpha} x^k - \frac{\alpha}{\gamma} \nabla f(y^k) \\ &= \frac{1}{\alpha} [y^k] - \frac{1 - \alpha}{\alpha} x^k - \frac{\alpha}{\gamma} \nabla f(y^k) \\ &= x^k + \frac{1}{\alpha} (y^k - x^k) - \frac{\alpha}{\gamma} \nabla f(y^k) \\ &= x^k + \frac{1}{\alpha} (x^{k+1} - x^k). \end{aligned}$$

For y^k , we start with using $\gamma = \mu$, then apply the v simplification;

$$\begin{aligned}
y^k &= \frac{\alpha\gamma v^k + \gamma x^k}{\alpha\gamma + \gamma} \\
&= \frac{(\alpha\gamma x^{k-1} + \gamma(x^k - x^{k-1})) + \gamma x^k}{\alpha\gamma + \gamma} \\
&= x^k + \frac{(\alpha\gamma x^{k-1} + \gamma(x^k - x^{k-1})) - \alpha\gamma x^k}{\alpha\gamma + \gamma} \\
&= x^k + \frac{\alpha\gamma(x^{k-1} - x^k) + \gamma(x^k - x^{k-1})}{\alpha\gamma + \gamma} \\
&= x^k + \frac{\gamma - \alpha\gamma}{\alpha\gamma + \gamma} (x^k - x^{k-1}).
\end{aligned}$$

Note that by multiplying by \sqrt{L}/μ we get:

$$\frac{\mu - \mu\sqrt{\mu/L}}{\mu\sqrt{\mu/L} + \mu} = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{\mu} + \sqrt{L}} = \beta.$$

Sutskever's form

Recall Sutskever's form:

$$\begin{aligned}
p_{\text{Sut}}^{k+1} &= \beta p_{\text{Sut}}^k - \frac{1}{L} \nabla f(x_{\text{Sut}}^k + \beta p_{\text{Sut}}^k), \\
x_{\text{Sut}}^{k+1} &= x_{\text{Sut}}^k + p_{\text{Sut}}^{k+1}.
\end{aligned}$$

and Nesterov's form:

$$\begin{aligned}
x_{\text{Nes}}^{k+1} &= y_{\text{Nes}}^k - \frac{1}{L} \nabla f(y_{\text{Nes}}^k), \\
y_{\text{Nes}}^{k+1} &= x_{\text{Nes}}^{k+1} + \beta (x_{\text{Nes}}^{k+1} - x_{\text{Nes}}^k).
\end{aligned}$$

We will show that using the substitutions:

$$\begin{aligned}
p_{\text{Sut}}^{k+1} &= x_{\text{Nes}}^{k+1} - x_{\text{Nes}}^k, \\
y_{\text{Nes}}^k &= x_{\text{Sut}}^k + \beta p_{\text{Sut}}^k,
\end{aligned}$$

applied to Sutskever's form gives Nesterov's form. We start with the momentum term:

$$\begin{aligned}
p_{\text{Sut}}^{k+1} &= \beta p_{\text{Sut}}^k - \frac{1}{L} \nabla f(y_{\text{Nes}}^k), \\
\therefore x_{\text{Nes}}^{k+1} - x_{\text{Nes}}^k &= \beta (x_{\text{Nes}}^k - x_{\text{Nes}}^{k-1}) - \frac{1}{L} \nabla f(y_{\text{Nes}}^k), \\
\therefore x_{\text{Nes}}^{k+1} + \frac{1}{L} \nabla f(y_{\text{Nes}}^k) &= x_{\text{Nes}}^k + \beta (x_{\text{Nes}}^k - x_{\text{Nes}}^{k-1}).
\end{aligned}$$

Defining: $y_{\text{Nes}}^k := x_{\text{Nes}}^{k+1} + \frac{1}{L} \nabla f(y_{\text{Nes}}^k)$ and applying on the right gives Nesterov's y update:

$$y_{\text{Nes}}^k = x_{\text{Nes}}^k + \beta (x_{\text{Nes}}^k - x_{\text{Nes}}^{k-1}).$$

Modern form

We want:

$$\begin{aligned}
p_{\text{Mod}}^{k+1} &= \beta p_{\text{Mod}}^k + \nabla f(x_{\text{Mod}}^k), \\
x_{\text{Mod}}^{k+1} &= x_{\text{Mod}}^k - \frac{1}{L} (\nabla f(x_{\text{Mod}}^k) + \beta p_{\text{Mod}}^{k+1}).
\end{aligned}$$

Starting from Sutskever's form,

$$\begin{aligned}
p_{\text{Sut}}^{k+1} &= \beta p_{\text{Sut}}^k - \frac{1}{L} \nabla f(x_{\text{Sut}}^k + \beta p_{\text{Sut}}^k), \\
x_{\text{Sut}}^{k+1} &= x_{\text{Sut}}^k + p_{\text{Sut}}^{k+1}.
\end{aligned}$$

Define $x_{\text{Mod}}^k = x_{\text{Sut}}^k + \beta p_{\text{Sut}}^k$. Note that this is equal to y_{Nes}^k by definition. So we have:

$$x_{\text{Sut}}^k = x_{\text{Mod}}^k - \beta p_{\text{Sut}}^k.$$

Plugging that into the Sutskever step

$$\begin{aligned} x_{\text{Mod}}^{k+1} - \beta p_{\text{Sut}}^{k+1} &= x_{\text{Mod}}^k - \beta p_{\text{Sut}}^k + p_{\text{Sut}}^{k+1}, \\ \therefore x_{\text{Mod}}^{k+1} &= x_{\text{Mod}}^k + (p_{\text{Sut}}^{k+1} - \beta p_{\text{Sut}}^k) + \beta p_{\text{Sut}}^{k+1} \\ &= x_{\text{Mod}}^k - \frac{1}{L} f'(x_{\text{Mod}}^k) + \beta p_{\text{Sut}}^{k+1}. \end{aligned}$$

Then define $p_{\text{mod}}^k = -L p_{\text{Sut}}^k$, so the update becomes:

$$x_{\text{Mod}}^{k+1} = x_{\text{Mod}}^k - \frac{1}{L} (\nabla f(x_{\text{Mod}}^k) + \beta p_{\text{Sut}}^{k+1}).$$

The momentum update changes from:

$$p_{\text{Sut}}^{k+1} = \beta p_{\text{Sut}}^k - \frac{1}{L} \nabla f(x_{\text{Sut}}^k + \beta p_{\text{Sut}}^k),$$

to:

$$\begin{aligned} -\frac{1}{L} p_{\text{mod}}^{k+1} &= -\beta \frac{1}{L} p_{\text{mod}}^k - \frac{1}{L} \nabla f(x_{\text{Mod}}^k), \\ \therefore p_{\text{mod}}^{k+1} &= \beta p_{\text{mod}}^k + \nabla f(x_{\text{Mod}}^k). \end{aligned}$$

Auslender & Teboulle form

$$\begin{aligned} y_{\text{AT}}^k &= (1 - \theta_{\text{AT}}) \hat{x}_{\text{AT}}^k + \theta_{\text{AT}} z_{\text{AT}}^k, \\ z_{\text{AT}}^{k+1} &= z_{\text{AT}}^k - \frac{\gamma_{\text{AT}}}{\theta_{\text{AT}}} \nabla f(y_{\text{AT}}^k), \\ \hat{x}_{\text{AT}}^{k+1} &= (1 - \theta_{\text{AT}}) \hat{x}_{\text{AT}}^k + \theta_{\text{AT}} z_{\text{AT}}^{k+1}. \end{aligned}$$

We first eliminate y_{AT}^k from the $\hat{x}_{\text{AT}}^{k+1}$ update:

$$\begin{aligned} \hat{x}_{\text{AT}}^{k+1} &= (1 - \theta_{\text{AT}}) \hat{x}_{\text{AT}}^k + \theta_{\text{AT}} z_{\text{AT}}^{k+1} \\ &= (1 - \theta_{\text{AT}}) \hat{x}_{\text{AT}}^k + \theta_{\text{AT}} \left(z_{\text{AT}}^k - \frac{\gamma_{\text{AT}}}{\theta_{\text{AT}}} \nabla f(y_{\text{AT}}^k) \right) \\ &= \hat{x}_k + \theta_{\text{AT}} (z_{\text{AT}}^k - \hat{x}_{\text{AT}}^k) - \gamma_{\text{AT}} \nabla f(\hat{x}_k + \theta_{\text{AT}} (z_{\text{AT}}^k - \hat{x}_{\text{AT}}^k)). \end{aligned}$$

This notational similarity with Nesterov's $x_{\text{Nes}}^{k+1} = y_{\text{Nes}}^k - \frac{1}{L} \nabla f(y_{\text{Nes}}^k)$ suggests matching

$$y_{\text{Nes}}^k = x_{\text{mod}}^k = y_{\text{AT}}^k = \hat{x}_k + \theta_{\text{AT}} (z_{\text{AT}}^k - \hat{x}_{\text{AT}}^k),$$

as well as $\gamma_{\text{AT}} = \frac{1}{L}$ and

$$x_{\text{Nes}}^k = \hat{x}_{\text{AT}}^k.$$

Note that using the step for \hat{x} we can rearrange to get:

$$z_{\text{AT}}^{k+1} = \frac{1}{\theta_{\text{AT}}} \hat{x}_{\text{AT}}^{k+1} - \frac{(1 - \theta_{\text{AT}})}{\theta_{\text{AT}}} \hat{x}_{\text{AT}}^k.$$

Now to determine θ we simplify using this substitution:

$$\begin{aligned} y_{\text{AT}}^k &= \hat{x}_{\text{AT}}^k + \theta_{\text{AT}} (z_{\text{AT}}^k - \hat{x}_{\text{AT}}^k) \\ &= \hat{x}_{\text{AT}}^k + \theta_{\text{AT}} \left(\frac{1}{\theta_{\text{AT}}} \hat{x}_{\text{AT}}^k - \frac{(1 - \theta_{\text{AT}})}{\theta_{\text{AT}}} \hat{x}_{\text{AT}}^{k-1} \right) \\ &= \hat{x}_{\text{AT}}^k + \theta_{\text{AT}} \left(\frac{1 - \theta_{\text{AT}}}{\theta_{\text{AT}}} \hat{x}_{\text{AT}}^k - \frac{(1 - \theta_{\text{AT}})}{\theta_{\text{AT}}} \hat{x}_{\text{AT}}^{k-1} \right) \\ &= \hat{x}_{\text{AT}}^k + (1 - \theta_{\text{AT}}) (\hat{x}_{\text{AT}}^k - \hat{x}_{\text{AT}}^{k-1}). \end{aligned}$$

So my matching constants against Nesterov's method we have $\beta_{\text{Nes}} = 1 - \theta_{\text{AT}}$.

Lan form

$$\begin{aligned}\tilde{x}_{\text{Lan}}^k &= \alpha_{\text{Lan}}(x_{\text{Lan}}^{k-1} - x_{\text{Lan}}^{k-2}) + x_{\text{Lan}}^{k-1}, \\ \underline{x}_{\text{Lan}}^k &= \frac{\tilde{x}_{\text{Lan}}^k + \tau_{\text{Lan}} \underline{x}_{\text{Lan}}^{k-1}}{1 + \tau_{\text{Lan}}}, \\ g_{\text{Lan}}^k &= \nabla f(\underline{x}_{\text{Lan}}^k), \\ x_{\text{Lan}}^k &= x_{\text{Lan}}^{k-1} - \frac{1}{\eta_{\text{Lan}}} g_{\text{Lan}}^k.\end{aligned}$$

Now we can eliminate \tilde{x} from the updates to give

$$\underline{x}_{\text{Lan}}^k = \frac{\tau_{\text{Lan}} \underline{x}_{\text{Lan}}^{k-1} + x_{\text{Lan}}^{k-1} + \alpha_{\text{Lan}}(x_{\text{Lan}}^{k-1} - x_{\text{Lan}}^{k-2})}{1 + \tau_{\text{Lan}}},$$

Also consider the y^k update for the AT method:

$$y_{\text{AT}}^k = (1 - \theta_{\text{AT}}) \hat{x}_{\text{AT}}^k + \theta_{\text{AT}} z_{\text{AT}}^k,$$

We can write this to not involve \hat{x} , giving:

$$y_{\text{AT}}^k = (1 - \theta_{\text{AT}}) y_{\text{AT}}^{k-1} + \theta_{\text{AT}} z_{\text{AT}}^k + (1 - \theta_{\text{AT}}) \theta_{\text{AT}} (z_{\text{AT}}^k - z_{\text{AT}}^{k-1}).$$

This form suggests matching the iterates with:

$$\underline{x}_{\text{Lan}}^k = y_{\text{AT}}^k, \quad x_{\text{Lan}}^k = z_{\text{AT}}^k.$$

Under this matching, the constants need to satisfy the following relations:

$$\begin{aligned}(1 - \theta_{\text{AT}}) &= \frac{\tau_{\text{Lan}}}{1 + \tau_{\text{Lan}}}, \\ (\theta_{\text{AT}} + \theta_{\text{AT}} - \theta_{\text{AT}}^2) &= \frac{1 + \alpha_{\text{Lan}}}{1 + \tau_{\text{Lan}}}, \\ (1 - \theta_{\text{AT}}) \theta_{\text{AT}} &= \frac{\alpha_{\text{Lan}}}{1 + \tau_{\text{Lan}}}.\end{aligned}$$

The settings:

$$\tau_{\text{Lan}} = \frac{1 - \theta_{\text{AT}}}{\theta_{\text{AT}}}, \quad \alpha_{\text{Lan}} = 1 - \theta_{\text{AT}},$$

result in these three constraints being satisfied:

$$\begin{aligned}\frac{\tau}{1 + \tau} &= \frac{\frac{1 - \theta}{\theta}}{1 + \frac{1 - \theta}{\theta}} = \frac{1 - \theta}{\theta + 1 - \theta} = 1 - \theta \checkmark, \\ \frac{1 + \alpha}{1 + \tau} &= \frac{1 + 1 - \theta}{1 + \frac{1 - \theta}{\theta}} = \frac{2\theta - \theta^2}{\theta + 1 - \theta} = 2\theta - \theta^2 \checkmark, \\ \frac{\alpha}{1 + \tau} &= \frac{1 - \theta}{1 + \frac{1 - \theta}{\theta}} = \frac{\theta - \theta^2}{1} \checkmark.\end{aligned}$$

Matching the x_{Lan}^k step $x_{\text{Lan}}^k = x_{\text{Lan}}^{k-1} - \frac{1}{\eta_{\text{Lan}}} g_{\text{Lan}}^k$ requires $\frac{1}{\eta_{\text{Lan}}} = \frac{\gamma_{\text{AT}}}{\theta_{\text{AT}}}$ also.

B Continuous time theory

Lemma 2. Consider a linear ODE of the form:

$$\dot{u} = A(u - u^*),$$

Where A is real and diagonalizable but not necessarily normal or symmetric, and u^* is the fixed point. Then

$$\|u(t) - u^*\| \leq \max_j \exp(t \operatorname{Re}[\lambda_j]) \|u(0) - u^*\|,$$

where λ_j are eigenvalues of A with real part $\operatorname{Re}[\lambda_j]$.

Proof. Because the inhomogeneous term Au^* is constant w.r.t time, without loss of generality we can reduce our problem to a homogenous ODE by shifting the origin:

$$\dot{u} = Au.$$

The solution of a linear ODE of this form is given in closed form using the matrix exponential:

$$u(t) = \exp(tA) u(0),$$

we can use this formula to bound the norm of $u(t)$:

$$\therefore \|u(t)\| \leq \|\exp(tA)\| \|u(0)\|,$$

where $\|\cdot\|$ is the spectral norm. If A is diagonalizable with $A = U \text{diag}(\lambda_1, \dots, \lambda_d) U^{-1}$ then the matrix exponential can be expressed as:

$$\exp(tA) = U \begin{bmatrix} \exp(t\lambda_1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \exp(t\lambda_d) \end{bmatrix} U^{-1}.$$

The spectral norm is given by the largest absolute value of the eigenvalues, so we must consider the interaction of the real and complex parts. For an eigenvalue $\lambda = a + bi$ of A , the norm takes the simple form:

$$\begin{aligned} |\exp(t\lambda)| &= \exp(ta + tbi) \\ &= |\exp(ta)| |\exp(tbi)| \\ &= |\exp(ta)|. \end{aligned}$$

So the spectral norm is given by the maximum over the eigenvalues λ_j of $\exp(t\text{Re}[\lambda_j])$. \square

Theorem 3. Consider the following linear ODE:

$$\begin{aligned} \dot{u} &= A(u - u^*), \\ A : 2n \times 2n &= \begin{bmatrix} -I & -\frac{1}{\eta}I + H^{-1} \\ \eta I & -\eta H^{-1} \end{bmatrix}, \end{aligned}$$

where $H : n \times n$ is a real, positive definite and symmetric matrix with minimum eigenvalue μ and maximum eigenvalue L . This corresponds to the proximal ODE for a quadratic function $f(x) = \frac{1}{2} (x - x^*)^T H (x - x^*)$, with $u = [x; g]$ and $u^* = [x^*; 0]$. Then the decay rate of the ODE towards the origin $u^* = 0$ can be bounded as follows for $\eta = \sqrt{\mu L}$:

$$\begin{aligned} \|u(t) - u^*\| &\leq \exp(-t\rho) \|u(0) - u^*\|, \\ \rho &= \frac{\sqrt{\mu}}{\sqrt{\mu} + \sqrt{L}}. \end{aligned}$$

Proof. We will take the approach of bounding the real parts of the eigenvalues of A , so that we can directly apply Lemma 2.

Let $U\Lambda U^T = H$ be the eigen-decomposition of H . Note that the operation of conjugation by U leaves the identity matrix unchanged ($UIU^T = I$). Each block of A is just a weighted combination of the identity matrix and H^{-1} , so this implies that conjugation of a block by its self gives a diagonal matrix. We can use this idea to define a similarity transform that converts A into a matrix where each of the four blocks are diagonal. In particular we have:

$$\begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}^T \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} = \begin{bmatrix} UA_{11}U^T & UA_{12}U^T \\ UA_{21}U^T & UA_{22}U^T \end{bmatrix},$$

where we have written A in terms of its 4 constituent $n \times n$ blocks. Each block is a weighted sum of the identity matrix and the diagonal matrix of inverse eigenvalues of H , for instance:

$$UA_{12}U^T = -\frac{1}{\eta}I + \Lambda^{-1}.$$

Next we construct a permutation matrix $\Pi : 2n \times 2n$ with the goal of converting D into a block diagonal matrix with 2×2 blocks along the diagonal, where each block has the structure of A as if it was applied to a 1D optimization problem, with H being replaced by one of the d eigenvalues of H . This is achieved with the permutation matrix Π that is zero except for:

$$\Pi_{2i,i} = 1, \Pi_{2i+1,d+i} = 1, i = 1 \dots n.$$

For instance, in the $n = 2$ case the matrix is:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

This matrix has the effect of interleaving the primal dual pairs (per coordinate) instead of having all the primal coordinates together followed by all the dual coordinates. So when we conjugate using Π we get:

$$\Pi^T \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}^T A \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} \Pi = \begin{bmatrix} T_1 & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & T_d \end{bmatrix},$$

where each T is a 2×2 matrix of the described form:

$$T_i = \begin{bmatrix} -1 & -\frac{1}{\eta} + \lambda_i^{-1} \\ \eta & -\eta\lambda_i^{-1} \end{bmatrix}.$$

The eigenvalues of a block diagonal matrix are just the eigenvalues of the blocks concatenated, and since there is a similarity transform between A and this block diagonal matrix, we have effectively reduced our problem to considering 1D quadratics, with curvature between μ and L , for fixed η .

Recall that for a matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ the eigenvalues are given by the two roots of a quadratic, namely:

$$\nu_{\pm} = \frac{a + d \pm \sqrt{(a + d)^2 - 4(ad - bc)}}{2}.$$

We use the notation ν to avoid confusion between the eigenvalues of the T blocks and those of H . For a block T_i , this expression is

$$\nu_{\pm}^{(i)} = -\frac{1}{2} - \frac{\eta\lambda_i^{-1}}{2} \pm \frac{1}{2} \sqrt{(1 + \eta\lambda_i^{-1})^2 - 4(\eta\lambda_i^{-1} + 1 - \eta\lambda_i^{-1})}$$

Suppose that the discriminate (the quantity under the square root) is negative, then

$$\text{Re} [\nu_{\pm}^{(i)}] = -\frac{1}{2} - \frac{\eta\lambda_i^{-1}}{2},$$

this is obviously at least as small as $-\rho = -\frac{\sqrt{\mu}}{\sqrt{\mu} + \sqrt{L}}$, since the largest value of ρ possible is when $\mu = L$, in which case $\rho = \frac{1}{2}$. So consider instead the case where the discriminate is positive. We need only consider the ν_+ root as it is strictly larger. Then we will use the concavity of the square root function to bound ν_+ ,

$$h(x) \leq h(y) + \langle \nabla h(y), x - y \rangle$$

For $h = \sqrt{\cdot}$, with $y = (-1 - \eta\lambda_i^{-1})^2$ and $x = (-1 - \eta\lambda_i^{-1})^2 - 4$. We get:

$$\begin{aligned} \nu_+^{(i)} &\leq -\frac{1}{2} - \frac{\eta\lambda_i^{-1}}{2} + \frac{1}{2} \sqrt{(1 + \eta\lambda_i^{-1})^2} - \frac{4}{4(1 + \eta\lambda_i^{-1})} \\ &= -\frac{1}{(1 + \eta\lambda_i^{-1})}. \end{aligned}$$

Therefore:

$$\nu_+^{(i)} \leq -\frac{1}{(1 + \eta\lambda_i^{-1})} \leq -\frac{1}{(1 + \eta/\mu)} = -\frac{\sqrt{\mu}}{\sqrt{\mu} + \sqrt{L}} = -\rho.$$

Thus we have shown that the real parts of all eigenvalues of A are less than $-\rho$. \square

Theorem 4. Consider the following linear ODE:

$$\dot{u} = A(u - u^*),$$

$$A : 2d \times 2d = \begin{bmatrix} -\frac{\alpha}{\eta\tau}H & -\frac{1}{\eta}I + \frac{\alpha}{\eta\tau}I \\ \frac{1}{\tau}H & -\frac{1}{\tau}I \end{bmatrix},$$

where $H : d \times d$ is a real, positive definite and symmetric matrix with minimum eigenvalue μ and maximum eigenvalue L . This corresponds to the AGM ODE for a quadratic objective $f(x) = \frac{1}{2}(x - x^*)^T H (x - x^*)$, with $u = [x; g]$ and $u^* = [x^*; 0]$. The decay rate of this ODE towards the origin $u^* = 0$ can be bounded as follows for $\eta = \sqrt{\mu L}$, $\tau = L/\eta$, and $\alpha \in [0, 1]$:

$$\|u(t)\| \leq \exp(-t\rho) \|u(0)\|,$$

$$\rho = \frac{1}{2}\sqrt{\frac{\mu}{L}}.$$

Proof. We can apply the same proof technique as for Theorem 3, we omit the details. The 2×2 block diagonal matrices are:

$$T_i = \begin{bmatrix} -\frac{\alpha}{\eta\tau}\lambda_i & -\frac{1}{\eta} + \frac{\alpha}{\eta\tau} \\ \frac{1}{\tau}\lambda_i & -\frac{1}{\tau} \end{bmatrix}.$$

For eigenvalues λ_i of H . The two eigenvalues of T_i are:

$$\nu_{\pm}^{(i)} = -\frac{\alpha}{2\eta\tau}\lambda_i - \frac{1}{2\tau} \pm \frac{1}{2}\sqrt{\left(\frac{\alpha}{\eta\tau}\lambda_i + \frac{1}{\tau}\right)^2 - 4\frac{\lambda_i}{\eta\tau}}$$

The discriminate here is always non-positive. To see why, we plug in the constants η and τ :

$$\begin{aligned} \left(\frac{\alpha}{L}\lambda_i + \sqrt{\frac{\mu}{L}}\right)^2 - 4\frac{\lambda_i}{L} &\leq 2\frac{\alpha^2}{L^2}\lambda_i^2 + 2\frac{\mu}{L} - 4\frac{\lambda_i}{L} \\ &\leq 2\frac{\lambda_i}{L} + 2\frac{\mu}{L} - 4\frac{\lambda_i}{L} \\ &= 2\frac{\mu}{L} - 2\frac{\lambda_i}{L} \\ &\leq 0. \end{aligned}$$

So the real part is given by the quantity outside the square root, namely:

$$\operatorname{Re} [\nu_{\pm}^{(i)}] = -\frac{\alpha}{2\eta\tau}\lambda_i - \frac{1}{2\tau} = -\frac{\alpha\lambda_i}{2L} - \frac{1}{2}\sqrt{\frac{\mu}{L}} \leq -\frac{1}{2}\sqrt{\frac{\mu}{L}}.$$

Using Lemma 2 gives the result. □

C The standard heavy ball ODE

The standard heavy-ball ODE for a quadratic $f(x) = \frac{1}{2}x^T H x$ is

$$\ddot{x} + (1 - \beta)\dot{x} + \gamma H x = 0.$$

Which can be written in first-order form in terms of a momentum parameter p as:

$$\begin{aligned} \dot{x} &= p, \\ \dot{p} &= -(1 - \beta)p - \gamma H x. \end{aligned} \tag{8}$$

The constants that result in optimal convergence rate for the discretize heavy ball method:

$$\begin{aligned} x^{k+1} &= x^k + p^k \\ p^k &= \beta p^{k-1} - \gamma \nabla f(x^k) \end{aligned}$$

which can also be written as:

$$x^{k+1} = x^k - \gamma H x^k + \beta (x^k - x^{k-1}),$$

are:

$$\beta = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \quad \gamma = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}. \quad (9)$$

Theorem 5. Consider the following ODE:

$$\begin{aligned} \dot{u} &= A(u - u^*), \\ A &= \begin{bmatrix} 0 & I \\ -\gamma H & -(1 - \beta)I \end{bmatrix}. \end{aligned}$$

This is the ODE in Equation 8 written in matrix form for a combined iterate u . For the parameters given in Equation 9, this ODE has decay rate at least:

$$\|u(t) - u^*\| \leq \exp(-t\rho) \|u(0) - u^*\|,$$

$$\text{where } \rho = \frac{\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}.$$

Proof. We can reduce the problem to considering the eigenvalues of 2×2 matrices as we did for the proximal and AGM ODEs. We have matrices of the form:

$$T_i = \begin{bmatrix} 0 & 1 \\ -\gamma\lambda_i & -(1 - \beta) \end{bmatrix},$$

whose eigenvalues are given by the general formula

$$\nu_{\pm} = \frac{a + d \pm \sqrt{(a + d)^2 - 4(ad - bc)}}{2}.$$

Simplifying:

$$\nu_{\pm} = -\frac{1}{2}(1 - \beta) \pm \frac{1}{2}\sqrt{(1 - \beta)^2 - 4\gamma\lambda_i}.$$

Note that:

$$1 - \beta = 2 \frac{\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}.$$

The choice of step size ensures that the discriminant is always non-positive:

$$\left(\frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2 - 4 \frac{4\lambda_i}{(\sqrt{L} + \sqrt{\mu})^2} \leq \left(\frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2 - 4 \left(\frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2 \leq 0$$

Therefore the decay rate is bounded by the real part of the eigenvalues, which is $-\frac{1}{2}(1 - \beta)$. \square