facebook Artificial Intelligence Research

# ON THE CURVED GEOMETRY OF ACCELERATED OPTIMIZATION

Aaron Defazio


## 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.


## Bregman proximal operators and 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:

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

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:

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

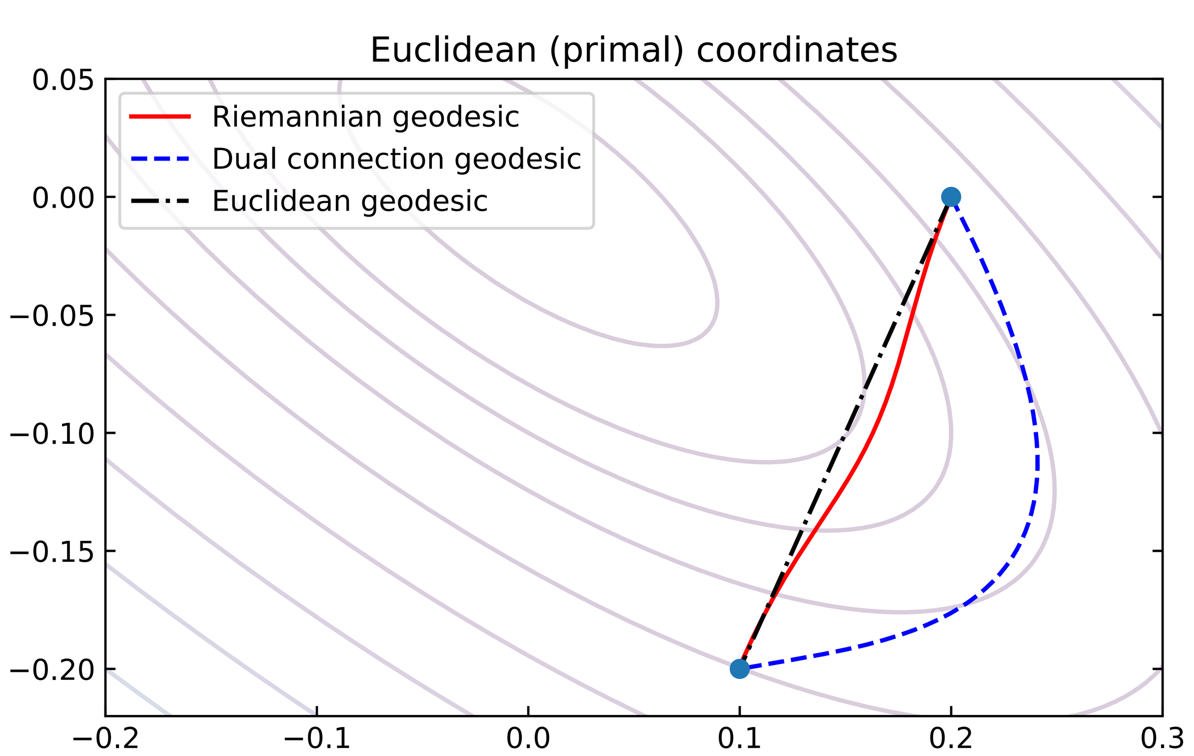
Bregman proximal operations can be interpreted as geodesic steps with respect to the dual connection. The key idea is that given an input point $x^{k-1}$, they output a point $x$ such that the velocity of the connecting geodesic is equal to $-\nabla \frac{1}{\rho} f(x)$ at $x$. This velocity is measured 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):

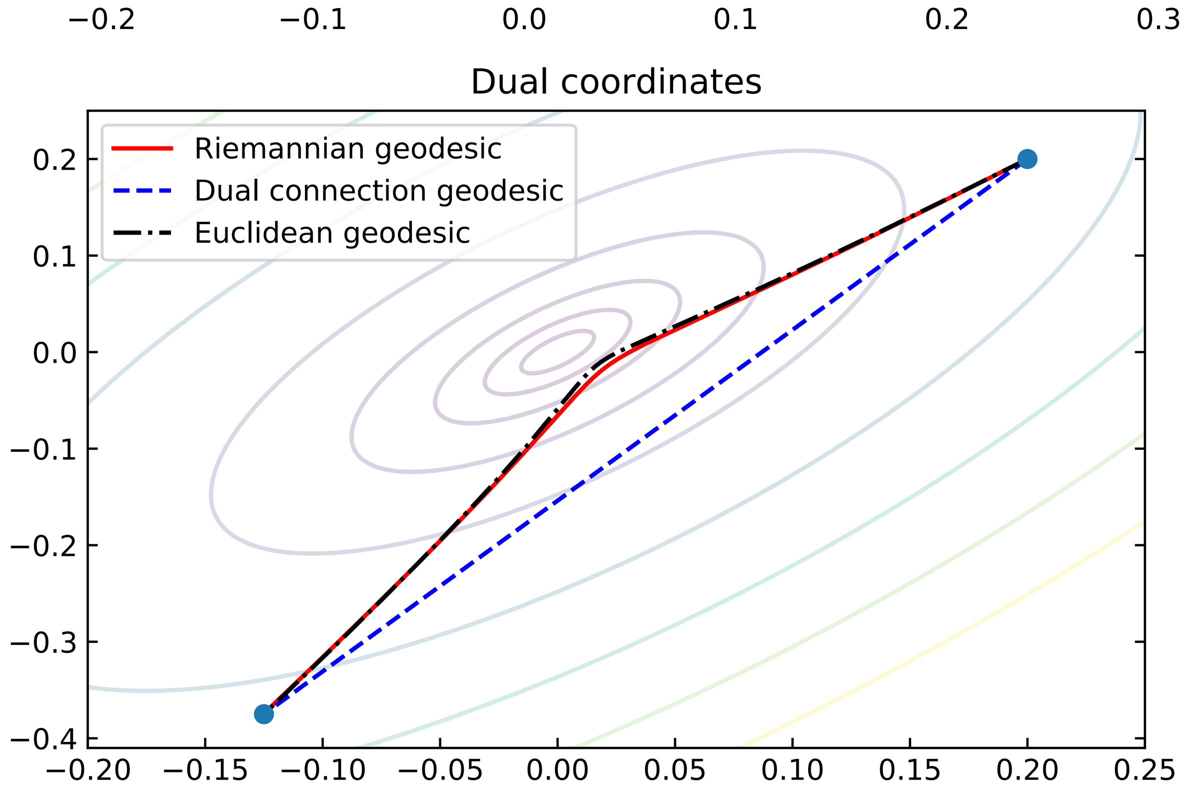
$$-\frac{1}{\rho}\nabla f(x^k) = \nabla\phi(x^k) - \nabla\phi(x^{k-1}).$$

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

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

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.




## 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 $\eta$:

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

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:

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

which is the optimality condition for the proximal operation:

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

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:

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

Which gives the updates:

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

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

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

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 $\tau$, 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:

$$\nabla f^*(g^k) - \left[x^{k-1} - \frac{1}{\eta}g^{k-1}\right] + \tau\nabla f^*(g^k) - \tau\nabla f^*(g^{k-1}) = 0,$$

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

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 $\alpha$ 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**

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

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 $\mu$-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]:

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

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

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

The difference in $\tau$ 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 $\frac{1}{2}\left\|\cdot\right\|^2$ is strongly convex with constant 1, so the change in scale requires rescaling by $L$.

### 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 $\rho$ is:

$$\left\| u(t) - u^* \right\| \le \exp(-t\rho)\left\| u(0) - u^* \right\|,$$

where $u^*$ is a fixed point. We can relate this to the discrete case by noting that $\exp(-t\rho) = \lim_{k\to\infty}(1 - \frac{t}{k}\rho)^k$, so given our discrete-time convergence rate is proportional to $(1 - \sqrt{\mu/L})^k$, we would expect values of $\rho$ 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).

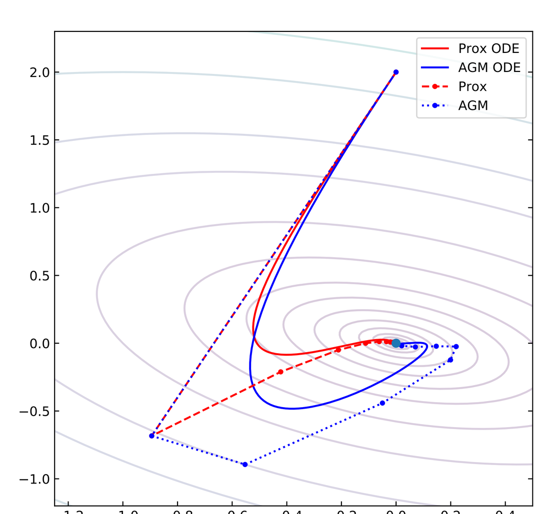



Figure 2: Paths for the quadratic problem $f(x) = \frac{1}{2}x^T Ax$ with $A = [2,\ 1;\ 1,\ 3]$.

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

$$\rho_{prox} \ge \frac{\sqrt{\mu}}{\sqrt{\mu}+\sqrt{L}}, \qquad \rho_{AGM} \ge \frac{1}{2}\sqrt{\frac{\mu}{L}}.$$

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.