Convergence Rate of Proximal-Gradient with a General Step-Size

Mark Schmidt
University of British Columbia

September 12, 2014

Abstract

We extend the previous analysis of Schmidt et al. [2011] to derive the linear convergence rate obtained by the proximal-gradient method under a general step-size scheme, for the problem of optimizing the sum of a smooth strongly-convex function and a simple (but potentially non-smooth) convex function.

1 Overview and Assumptions

We consider minimization problems of the form

$$\min_{x \in \mathbb{R}^d} f(x) := g(x) + h(x), \tag{1.1}$$

where g a is strongly-convex function with parameter μ, g' is Lipschitz-continuous with parameter L, and h is only required to be a lower semi-continuous proper convex function. This class includes the elastic-net regularized least-squares problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}\|Ax - b\|^2 + \frac{\lambda_2}{2}\|x\|^2 + \lambda_1 \|x\|_1,$$

with $g(x) = \frac{1}{2}\|Ax - b\|^2 + \frac{\lambda_2}{2}\|x\|^2$ and $h(x) = \lambda_1\|x\|_1$. In this case, $L = \sigma_{\max}(A^T A) + \lambda_2$ and $\mu = \sigma_{\min}(A^T A) + \lambda_2$. In this work we'll analyze the proximal-gradient algorithm, which uses iterations of the form

$$x^{k+1} = \text{prox}[x^k - \alpha g'(x^k)], \tag{1.2}$$

where $\alpha > 0$ is the step-size and the proximal operator is

$$\text{prox}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^d} \frac{1}{2}\|x - y\|^2 + \alpha h(y). \tag{1.3}$$

Our prior results in Schmidt et al. [2011, Proposition 3] show that with a step-size of $\alpha = 1/L$ that the iterates of this algorithm have a linear convergence rate,

$$\left\|x^k - x^*\right\| \leq \left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^*\|,$$

where x^* is the optimal solution. In this note show that for a general step-size α we have

$$\left\|x^k - x^*\right\| \leq Q(\alpha)^k \|x_0 - x^*\|,$$

where $Q(\alpha) = \max\{|1 - \alpha L|, |1 - \alpha\mu|\}$. This matches the known rate of the gradient method with a constant step-size for solving strictly-convex quadratic problems [Bertsekas, 1999, Section 1.3], and the rate of the projected-gradient algorithm with a constant step-size for minimizing strictly-convex quadratic functions over convex sets [Bertsekas, 1999, Section 2.3]. This result includes the previous result as a speical case since $Q(\frac{1}{L}) = 1 - \frac{\mu}{L}$, and also gives a faster rate if we miniminze Q in terms of α to give $\alpha = \frac{2}{L+\mu}$ which yields $Q\left(\frac{2}{L+\mu}\right) = 1 - \frac{2\mu}{L+\mu} = \frac{L-\mu}{L+\mu}$.

2 Useful inequalitites

We note that x^* is a fixed-point of the iterations,

$$x^* = \text{prox}[x^* - \alpha g'(x^*)]. \tag{2.1}$$

This follows because by the definition of x^* is satifies the optimality condition for (1.1),

$$0 \in g'(x^*) + \partial h(x^*). \tag{2.2}$$

The optimality conditions that define the solution to the proximal problem (1.3) are

$$0 \in -(x - y) + \alpha \partial h(y),$$

and plugging in $x = x^* - \alpha g'(x^*)$ we have

$$0 \in (y - x^*) + \alpha g'(x^*) + \alpha \partial h(y),$$

which in light of (2.2) is solved by setting $y = x^*$.

We'll also use that the proximal operator is non-expansive [Combettes and Wajs, 2005],

$$\|\text{prox}[x] - \text{prox}[y]\|^2 \leq \langle \text{prox}[x] - \text{prox}[y], x - y \rangle,$$

which implies by Cauchy-Schwartz that

$$\|\text{prox}[x] - \text{prox}[y]\| \leq \|x - y\|, \tag{2.3}$$

Because g' is L-Lipschitz continuous we have

$$\left\|g'(x) - g'(y)\right\| \leq L\|x - y\|,$$

and because g is μ-strongly convex we have

$$\left\|g'(x) - g'(y)\right\| \geq \mu\|x - y\|,$$

so putting these together (noting that $L \geq \mu$) we have for any β (positive or negative) that

$$\beta\left\|g'(x) - g'(y)\right\|^2 \leq \max\{\beta L^2, \beta\mu^2\}\|x - y\|^2. \tag{2.4}$$

Finally, because g' is L-Lipschitz and μ-strongly convex we have [Nesterov, 2004, Theorem 2.1.12]

$$\langle g'(x) - g'(y), x - y \rangle \geq \frac{1}{L + \mu}\left\|f'(x) - f'(y)\right\|^2 + \frac{L\mu}{L + \mu}\|x - y\|^2. \tag{2.5}$$

3 Derivation

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &= \left\|\text{prox}[x^k - \alpha g'(x^k)] - \text{prox}[x^* - \alpha g'(x^*)]\right\|^2 && (1.2), (2.1) \\
&\leq \left\|(x^k - \alpha g'(x^k) - (x^* - \alpha g'(x^*))\right\|^2 && (2.3) \\
&= \left\|(x^k - x^*) - \alpha(g'(x^k) - g'(x^*))\right\|^2 \\
&= \left\|(x^k - x^*)\right\|^2 - 2\alpha\langle g'(x^k) - g'(x^*), x^k - x^* \rangle + \alpha^2\left\|g'(x^k) - g'(x^*)\right\|^2 \\
&\leq \left\|(x^k - x^*)\right\|^2 - 2\alpha\left(\frac{1}{L + \mu}\left\|g'(x^k) + g'(x^*)\right\|^2 + \frac{L\mu}{L + \mu}\left\|x^k - x^*\right\|^2\right) + \alpha^2\left\|g'(x^k) - g'(x^*)\right\|^2 && (2.5) \\
&= \left(1 - \frac{2\alpha L\mu}{L + \mu}\right)\left\|(x^k - x^*)\right\|^2 + \alpha\left(\alpha - \frac{2}{L + \mu}\right)\left\|g'(x^k) - g'(x^*)\right\|^2 \\
&\leq \left(1 - \frac{2\alpha L\mu}{L + \mu}\right)\left\|(x^k - x^*)\right\|^2 + \alpha\max\left\{L^2\left(\alpha - \frac{2}{L + \mu}\right), \mu^2\left(\alpha - \frac{2}{L + \mu}\right)\right\}\left\|x^k - x^*\right\|^2 && (2.4) \\
&= \max\left\{\left(1 - \frac{2\alpha L\mu}{L + \mu}\right) + \alpha L^2\left(\alpha - \frac{2}{L + \mu}\right), \left(1 - \frac{2\alpha L\mu}{L + \mu}\right) + \alpha\mu^2\left(\alpha - \frac{2}{L + \mu}\right)\right\}\left\|x^k - x^*\right\|^2 \\
&= \max\left\{1 - \frac{2\alpha L(L + \mu)}{L + \mu} + \alpha^2 L^2, 1 - \frac{2\alpha\mu(L + \mu)}{L + \mu} + \alpha^2\mu^2\right\}\left\|x^k - x^*\right\|^2 \\
&= \max\left\{(1 - \alpha L)^2, (1 - \alpha\mu)^2\right\}\left\|x^k - x^*\right\|^2 \\
&= Q^2\left\|x^k - x^*\right\|^2.
\end{aligned}$$

Taking the square root and applying it repeatedly gives the result.

References

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 2nd edition, 1999.

P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Netherlands, 2004.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *Neural Information Processing Systems*, 2011.