University of California, Berkeley Lecturer: Benjamin Recht

CS227C/STAT260 Convex Optimization and Approximation: Optimization for Modern Data Analysis

January 22, 2015 Notes: Ashia Wilson and Benjamin Recht

Lecture 3: the gradient method (loose ends)

Let's wrap up a few loose ends about the gradient method. First, we will discuss convex and strongly convex functions and estimate the convergence of the gradient method when applied to these functions. Then, we'll turn to how to estimate step-sizes without knowledge of the Lipschitz constants or strong convexity parameters.

1 Strong Convexity

A function $f : \mathbb{R}^d \to \mathbb{R}$ is *strongly convex* if there is a scalar $m > 0$ such that

$$f(z) \geq f(x) + \nabla f(x)^T(z - x) + \tfrac{m}{2}\|z - x\|^2 \tag{1}$$

Strong convexity asserts that f can be lower bounded by quadratic functions. These functions change from point to point, but only in the linear term. It also tells us that the curvature of the function is bounded away from zero. Note that if a function is strongly convex *and* has L-Lipschitz gradients, then f is bounded above and below by simple quadratics. This sandwiching will enable us to prove the linear convergence of the gradient method.

The simplest strongly convex function is the squared Euclidean norm $\|x\|^2$. Any convex function can be perturbed to form a strongly convex function by adding any small multiple of the squared Euclidean norm. In fact, if f is any differentiable function with L-Lipschitz gradients, then

$$f_\mu(x) = f(x) + \mu\|x\|^2$$

is strongly convex for μ large enough. Verifying this fact is a fun exercise.

As another canonical example, note that a quadratic function $f(x) = \frac{1}{2}x^TQx$ is strongly convex if and only if the smallest eigenvalue of Q is strictly positive.

Strongly convex functions are in essence the "easiest" functions to optimize by first-order methods. First, the norm of the gradient provides useful information about how far away we are from optimality. Note that if we minimize the right hand side of our strong convexity condition with respect to x, we find that the minimizer is $x - \frac{1}{m}\nabla f(x)$. Plugging that into (1), we find

$$\begin{aligned} f(z) &\geq \min_x f(x) + \nabla f(x)^T(z - x) + \frac{m}{2}||z - x||^2 \\ &\geq f(x) - \nabla f(x)^T\frac{1}{m}\nabla f(x) + \frac{m}{2}\left\|\frac{1}{m}\nabla f(x)\right\|^2 \\ &\geq f(x) - \frac{1}{2m}||\nabla f(x)||^2 \end{aligned} \tag{2}$$

Now if $||\nabla f(x)|| < \delta$ then

$$f(x) - f(x_\star) \leq \frac{||\nabla f(x)||^2}{2m} \leq \frac{\delta^2}{2m}$$

Thus, when the gradient is small, we are close to having found a point with minimal function value. We can even derive a stronger result about the distance of x to the optimal point $x_\star$. Using (1) and Cauchy-Schwartz, we have

$$\begin{aligned} f(x_{opt}) &\geq f(x) + \nabla f(x)^T (x_\star - x) + \frac{m}{2}||x - x_\star||^2 \\ &\geq f(x) - ||\nabla f(x)||\, ||x_\star - x|| + \frac{m}{2}||x - x_\star||^2 \end{aligned}$$

Rearranging terms proves that

$$||x - x_\star|| \leq \frac{2}{m}||\nabla f(x)||\,. \tag{3}$$

This says we can estimate the distance to the optimal value purely in terms of the norm of the gradient.

An immediate consequence of (3) is the following

Corollary 1 *If f is strongly convex then f has a unique optimal solution.*

Essentially, strongly convex functions are nice, wide bowls, and we just need to roll downhill to the bottom.

We close this discussion of strong convexity by proving that for differentiable functions, strong convexity is equivalent to the Hessian having positive eigenvalues. We encountered this fact when we were discussing contraction mappings in the previous lecture.

Proposition 2 *If f is strongly convex and two-times differentiable, then $\nabla^2 f(x) \succeq mI$*

Proof Using the fact that

$$f(x) \leq f(y) + \nabla f(y)^T (x - y) + \frac{m}{2}||x - y||^2$$

and

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}||y - x||^2$$

we can add these two inequalities together and get

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq m||x - y||^2\,.$$

Setting $x = u + \alpha v$ and $y = u$, for u and v in $\mathbb{R}^d$ yields

$$\langle \nabla f(u + \alpha d) - \nabla f(v), \alpha v \rangle \geq m\alpha^2 ||d||^2$$

Dividing through by α^2 and taking the limit as α goes to zero proves

$$d^T \nabla^2 f(x) d \geq m||x - y||^2$$

as desired. ■

2 Three proofs of convergence of the gradient method on strongly convex functions

Strongly convex functions allow us to show off three popular search strategies for proving convergence of optimization algorithms.

2.1 Descent analysis

Recall from last lecture that if we run the gradient method with stepsize $1/L$ (or use exact line search), we have the inequality

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}||\nabla f(x_k)||^2 .$$

Subtracting $f(x_\star)$ from both sides and using (2) gives

$$f(x_{k+1}) - f(x_{opt}) \leq \left[1 - \frac{m}{L}\right] (f(x_k) - f(x_{opt}))$$

Here $\frac{m}{L}$ is an estimate of the worst-case "condition number" of the Hessian. When m is significantly smaller than L, f will have very eccentric level sets. Since the gradient is orthogonal to the contours of f, this will cause the gradient method to oscillate rapidly, and convergence will be quite slow.

2.2 Contraction Mapping

If f is twice continuously differentiable, note that the map

$$\Phi(x) = x - \alpha \nabla f(x)$$

is a contraction mapping for any $0 < \alpha < \frac{2}{L+m}$. To see this, observe

$$\begin{aligned}
\|\Phi(x) - \Phi(y)\| &= \|x - \alpha\nabla f(x) - (y - \alpha \nabla f(y))\| \\
&\leq \left\| \int_0^\infty (I - \alpha \nabla^2 f(x + t(y-x))(y-x) dt \right\| \\
&\leq \sup_z \|I - \alpha \nabla^2 f(z)\| \|y - z\| .
\end{aligned}$$

Note that the minimum eigenvalue of $\nabla^2 f(z)$ is at least m and the maximum eigenvalue is at least L. Therefore the eigenvalues of $I - \alpha \nabla^2 f(z)$ are at most $\max(1 - \alpha L, 1 - \alpha m)$ and at least $\min(1 - \alpha L, 1 - \alpha m)$. Therefore,

$$\|I - \alpha \nabla^2 f(z)\| \leq \max(|1 - \alpha L|, |1 - \alpha \ell|) .$$

The right-hand side is minimized when $\alpha = \frac{2}{L+m}$. Moreover, the quantity is less than 1 and Φ is contractive whenever $0 < \alpha < 2/L$.

Note that with the stepsize $\alpha = \frac{2}{L+m}$, we see that the iterates converge linearly with the rate

$$\frac{L - m}{L + m}$$

which is slightly faster than the rate $(1 - L/m)$ derived above.

2.3 Lyapunov Analysis

A function $V : \mathbb{R}^d \to \mathbb{R}$ is a Lypaunov function for an algorithm if

1. $V(x) \geq 0$
2. $V(x_\star) = 0$
3. $V(x_{k+1}) < V(x_k)$ for all iterates of the algorithm

The existence of a Lyapunov question is clearly sufficient to guarantee asymptotic convergence of a method.

Note that we showed in Section 2.1 that $f(x) - f_\star$ was a Lyapunov function. Here, we can show that $\|x_k - x_\star\|$ is also a Lyapunov function. This follows from the contraction argument, but it's worth mentioning this connection to Lyapunov analysis, as this will be another tool in our belt moving forward.

2.4 From rates to iterations

When we have linear convergence, we are often left with expressions of the form

$$f(x_k) - f(x_\star) \leq (1 - \beta)^k D_0$$

where D_0 defines some initial distance or deviation in the function value. We'd like to turn these expressions into bounds on how many iterations it takes to get $f(x_k) - f(x_\star) \leq \epsilon$.

Let's say we have run for N iterations. Taking logarithms, we have the inequality

$$\log(\epsilon) \geq N \log [1 - \beta] + \log D_0$$

Rearranging terms, we find that

$$N \geq \frac{-\log(D_0/\epsilon)}{\log(1 - \beta)}$$

Using the inequality $1 - x \leq \exp(-x)$ or equivalently $\log(1 - x) \leq -x$ we get the sufficient condition that

$$N \geq \beta^{-1} \log(D_0/\epsilon)\,.$$

are required to get $f(x_k) - f(x_\star) \leq \epsilon$.

For the gradient method, this means

$$N \geq \frac{L}{m} \log \left(\frac{f(x_0) - f(x_\star)}{\epsilon} \right)$$

iterations suffice to guarantee convergence to tolerance ϵ. This ratio L/m governs how regular the curvature of f is, and it is akin to the condition number of a matrix. Note that this is not the condition number of the Hessian, per se. We can define 1-dimensional functions where L/m is arbitrarily large, but the Hessian has condition number 1 everywhere.

3 Gradient method with "weak" convexity

The gradient method for functions that are convex but not strongly convex also converges faster than in the nonconvex case we previously studied. We'll refer to such functions as "weakly convex" to contrast them to strongly convex functions. Note that all convex functions are weakly convex, but not all convex functions are strongly convex.

The analysis of this case uses an important property of convex functions. Indeed, the following inequality completely characterizes the set of convex functions with L-Lipschitz gradients.

Lemma 3 *f is convex with L-Lipschitz gradients if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{L}||\nabla f(x) - \nabla f(y)||^2$$

Proof Define $\varphi_{x_0}(y) = f(y) - \langle \nabla f(x_0), y \rangle$ (i.e. φ_{x_0} is just f perturbed by a linear function). φ_{x_0} has lipschitz gradients. Furthermore

$$\nabla_y \varphi_{x_0}(y) = \nabla f(y) - \nabla f(x_0)$$

implying that $x_0 \in \arg\min \varphi_{x_0}$. Therefore we have

$$\begin{aligned} \varphi(x_0) &= \varphi(y - \frac{1}{L}(\nabla f(y) + \nabla f(x_0)) \\ &\leq \varphi(y) - \frac{1}{2L}||\nabla f(y) - \nabla f(x_0)||^2 \end{aligned}$$

where the last step follows from a standard Lipschitz upper bound $f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}||x - y||^2$

For the converse, note that convexity follows because the quadratic term is nonnegative. The Lipschitz gradients follows because if we swap the role of x and y and add the inequalities, we have

$$|\langle \nabla f(x) - \nabla f(y), x - y \rangle| \geq \frac{1}{2L}||\nabla f(x) - \nabla f(y)||^2$$

Cauchy-Schwartz completes the proof. ■

It is interesting about this lemma is that the set of convex functions with L-Lipschitz gradients is characterized by *one* inequality. Rather than two. This property is called *co-coercivity* and will be used many times in the course.

With the co-coercivity lemma in hand, the following theorem gives our convergence rate for general convex functions.

Lemma 4 *If f is convex with L-Lipschitz gradients, the gradient method with $t = \frac{1}{L}$ satisfies*

$$f(x_k) - f(x^*) \leq \frac{2L||x_0 - x^*||^2}{k + 1}$$

Proof Recall again that we have

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}||\nabla f(x_k)||^2$$

We also have (by first order conditions),

$$f(x_k) - f(x_\star) \leq \langle \nabla f(x_k), x_k - x* \rangle \leq ||\nabla f(x_k)||\, ||x_k - x^*|| \tag{4}$$

Now by the lemma

$$\begin{aligned} ||x_{k+1} - x^*||^2 &= ||x_k - \frac{1}{L}\nabla f(x_k) - x^*||^2 \\ &= ||x_k - x^*||^2 - \frac{2}{L}\langle \nabla f(x_k), x_k - x^* \rangle + \frac{1}{L^2}||\nabla f(x_k)||^2 \end{aligned}$$

By first order conditions

$$\begin{aligned} -\frac{2}{L}\langle \nabla f(x_k), x_k - x^* \rangle &\leq \frac{2}{L}(f(x^*) - f(x_k)) \\ &\leq \frac{2}{L}(f(x_{k+1}) - f(x_k)) \\ &\leq \frac{2}{L}\left(-\frac{1}{2L}||\nabla f(x_k)||^2\right) \\ &\leq -\frac{1}{L^2}||\nabla f(x_k)||^2 \end{aligned}$$

Therefore we get

$$||x_{k+1} - x^*||^2 \leq ||x_k - x^*||^2 \leq ||x_0 - x^*||^2$$

This means that the iterates live in a bounded set. Using (4),

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq f(x_k) - f(x^*) - \frac{1}{2L\|x_k - x_\star\|^2}(f(x_k) - f(x^*))^2 \\ &\leq f(x_k) - f(x^*) - \frac{1}{2L\|x_0 - x_\star\|^2}(f(x_k) - f(x^*))^2 \end{aligned}$$

Defining $D_0 \equiv ||x_0 - x^*||$ we have

$$f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \frac{1}{2LD_0^2}(f(x_k) - f(x^*))^2 \tag{5}$$

This recursion is similar to the bound we had for linear convergence, but now we have an additional quadratic term. The rest of the proof is just algebraic manipulation to show that the convergence rate is achieved.

Define $\Delta_k := f(x_k) - f(x_\star)$. Then the recursion (5) becomes

$$\Delta_{k+1} \leq \Delta_k - \frac{1}{2LD_0^2}\Delta_k^2\,.$$

Dividing both sides by $\frac{1}{\Delta_k \Delta_{k+1}}$ we have

$$\frac{1}{\Delta_k} \leq \frac{1}{\Delta_{k+1}} - \frac{1}{2LD_0^2}\frac{\Delta_k}{\Delta_{k+1}}$$

Rearranging the inequality gives

$$\begin{aligned}\frac{1}{\Delta_{k+1}} &\geq \frac{1}{\Delta_k} + \frac{1}{2LD_0^2}\frac{\Delta_k}{\Delta_{k+1}} \\ &\geq \frac{1}{\Delta_k} + \frac{1}{2LD_0^2} \\ &\geq \frac{1}{\Delta_0} + \frac{k+1}{2LD_0^2} \\ &\geq \left(\frac{1}{2} + \frac{k+1}{2}\right)\left(\frac{1}{LD_0^2}\right) \\ &= \frac{k+2}{2LD_0^2}\end{aligned}$$

Where we used $\Delta_0 \leq 2LD_0^2$. Thus,

$$f(x_{k+1}) - f(x^*) \leq \frac{2L||x_0 - x^*||^2}{k+2}$$

■

4 Backtracking Line Search

Finally, let's figure out how to automatically tune the step-size without knowledge of the Lipschitz constant. One of the most popular schemes is called *backtracking line seach*. The idea is simple: we choose a starting step-size, and if the function is not sufficiently decreased, we choose a smaller step. This is summarized as:

1. First we pick a direction $v = -\nabla f(x_k)$.
2. Start with $t = 1$
3. Test the Armijo Condition:

$$f(x_k - t\nabla f(x_k)) \leq f(x_k) - \gamma t||\nabla f(x_k)||^2$$

   (a) If fails $t = \beta t$
   (b) If succeeds, terminate

A decent rule of thumb for this procedure is to choose $\beta = .8$ and $\gamma = \frac{1}{2}$. But your mileage may vary.

Analyzing backtracking line search is not much more difficult than analyzing constant stepsize. Assume $\gamma \leq \frac{1}{2}$. Then at termination,

$$f(x_{k+1}) \leq f(x_k) - \gamma t||\nabla f(x_k)||^2$$

by assumption. Note that we *always* have

$$f(x_{k+1}) \leq f(x_k) - t\left(1 - \frac{tL}{2}\right)||\nabla f(x_k)||^2,$$

for our step t.

There are now two cases to analyze:

1. Case 1: $t_0 = 1$ satisfies the Armijo condition. In this case,
$$\Rightarrow ||\nabla f(x_k)||^2 \leq \gamma^{-1}(f(x_k) - f(x_{k+1}))\,.$$

2. Case 2: We succeed at some t after backtracking. But this means that $\beta^{-1}t$ failed to satisfy the Armijo condition. Writing this out explicitly:
$$\begin{aligned} f(x_k) - \gamma\beta^{-1}t||\nabla f(x_k)||^2 &\leq f(x_k - \beta^{-1}t\nabla f(x_k)) \\ &\leq f(x_k) - \beta^{-1}t\left(1 - \frac{tL}{2}\right)||\nabla f(x_k)||^2 \end{aligned}$$
If we match terms here, we find that
$$-\gamma\beta^{-1}t \leq \beta^{-1}t\left(1 - \frac{\beta^{-1}tL}{2}\right)$$
Or, again rearranging things,
$$t \geq \frac{2(1-\gamma)\beta}{L} \geq \frac{\beta}{L}$$

Therefore we can combine the cases and find
$$\begin{aligned} ||\nabla f(x_k)||^2 &\leq \frac{L}{\beta\gamma}\left[f(x_k) - f(x_{k+1})\right] \\ &\leq \frac{L}{\gamma\min(L,\beta)}\left[f(x_k) - f(x_{k+1})\right] \end{aligned}$$

Now we can apply the previous analyses.

5 Complexity

How many iterations will this take?

$$\begin{aligned} F &: \text{Time to compute the function} && \sim O(d) \\ G &: \text{Time to compute the gradient} && \sim O(d) \\ H &: \text{Time to compute the Hessian} && \sim O(d^2) \end{aligned}$$

Definition 1 (Gaxpy) *A gaxpy is a scale and add of a vector (i.e. our gradient method update* $x_k \leftarrow x_{k+1} + \alpha g$*)*

For Gradient method on convex function, we have
$$(\underbrace{\text{time for gaxpy}}_{O(d)} + \underbrace{\text{Gradient call}}_{O(d)} + \underbrace{\text{backtracking}}_{O(\log(L))F}) \cdot \frac{L||x_0 - x^*||}{N}$$