University of California, Berkeley | Lecturer: Benjamin Recht

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

February 4, 2014 | Scribe: Ashia Wilson

Lecture 5: Gradient Descent for (Strong) Convex Functions

1 Properties

Recall: A function f is strongly convex if there exists an $\ell > 0$ such that

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

Strong convexity gives us a nice quadratic lower bound at any point x of our function f. A nice way to form a strongly convex function is to take a convex function $f_0(x)$ and add $\frac{\mu}{2}||x||^2$ to the function.

We have the following implications:

- If we minimize the right hand side of our strong convexity condition over x, we see that setting the gradient equal to 0 give $x = y - \frac{1}{\ell}\nabla f(y)$. Plugging that in we get

$$\begin{aligned} f(x) &\leq \min_x f(y) + \nabla f(y)^T(x-y) + \frac{\ell}{2}||x-y||^2 \\ &\leq f(y) - \nabla f(y)^T \frac{1}{\ell}\nabla f(y) + \frac{\ell}{2}||\frac{1}{\ell}\nabla f(y)||^2 \\ &\leq f(y) - \frac{1}{2\ell}||\nabla f(y)||^2 \end{aligned}$$

- If $||\nabla f(x)|| < \delta$ then

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

  Thus, the gradient tells you how far you are from being optimal.

- We also have

$$\begin{aligned} f(x_{opt}) &\geq f(x) + \nabla f(x)^T(x_{opt} - x) + \frac{\ell}{2}||x - x_{opt}||^2 \\ &\geq f(x) - ||\nabla f(x)||\ ||x_{opt} - x|| + \frac{\ell}{2}||x - x_{opt}||^2 \qquad \text{(Cauchy Schwartz)} \\ &\implies ||x - x_{opt}|| \leq \frac{2}{\ell}||\nabla f(x)|| \end{aligned}$$

  This says we have control of the distance to the optimal value totally in terms of the gradient

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

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

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

Proof.

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

and

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

Therefore adding these two together, we get

$$\langle \nabla f(x) - \nabla f(y), x-y \rangle \geq \ell ||x-y||^2$$
$$\Big\langle \frac{\nabla f(x+\alpha d) - \nabla f(y)}{\alpha}, \alpha d \Big\rangle \geq \alpha \ell ||x-y||^2$$
$$\Big\langle \frac{\nabla f(x+\alpha d) - \nabla f(y)}{\alpha}, d \Big\rangle \geq \ell ||x-y||^2$$

Taking the limit as $\alpha \to 0$

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

□

2 Gradient Descent with Strong Convexity

If we do exact line search (back tracking is similar), for f strongly convex, we get convergence at a linear rate

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

for exact line search (or $t = \frac{1}{L}$), we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - t||\nabla f(x_k)||^2 + \frac{Lt^2}{2}||\nabla f(x_k)||^2 \\ &\leq f(x_k) - \frac{1}{2L}||\nabla f(x_k)||^2 \end{aligned}$$

Therefore

$$f(x_{k+1}) - f(x_{opt}) \leq \underbrace{(f(x_k) - f(x_{opt}))\left[1 - \frac{\ell}{L}\right]}_{\text{linear convergence}}$$

Where $\frac{\ell}{L}$ the worst case condition number of the Hessian. When ℓ 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.

When f is not strongly convex, our convergence rate is even slower as it is difficult to get a quadratic lower bound on the function in terms of its gradients. The best we can provide is the following

Lemma 3. *If f has L-Lipschitz gradients and is convex then*

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

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

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

$\implies x_0 \in \arg\min \varphi_{x_0}$. Therefore

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

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

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.

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

We also have (by first order conditions),

$$\begin{aligned}
f(x_k) - f(x^*) &\leq \langle \nabla f(x_k), x_k - x* \rangle \\
&\leq ||\nabla f(x_k)||\ ||x_k - x^*||
\end{aligned}$$

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

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

Defining $D_0 \equiv ||x_0 - x^*||$ and $\Delta_k = f(x_k) - f(x^*)$, we have

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq (f(x_k) - f(x^*)) - \frac{1}{2L}\frac{(f(x_k) - f(x^*))^2}{D_0^2} \\ \Delta_{k+1} &\leq \Delta_k - \frac{1}{2LD_0^2}\Delta_k^2 \end{aligned}$$

$\implies$

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

Thus,

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

□

2.1 Lower Bound for Strongly Convex Functions

Can we do better than the gradient method? And is the bound

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

optimal for strongly convex functions. In order to check this, we are going to contract hard functions to help understand the worst case runtime.

Worst Case Instances for Strongly Convex Functions are Quadratics

$$f(x) = (1/2)x^T Ax - b^T x \qquad A \succeq 0 \quad \begin{aligned} A &\preceq LI \\ A &\succeq \ell I \end{aligned}$$

What does gradient iteration do? The optimal solution is given by $\nabla f(x) = 0 \Rightarrow \nabla f(x) = Ax - b$

$$
\begin{aligned}
x_0 &= 0 \qquad \leftarrow \text{worst case bounds are deterministic in terms of starting values} \\
x_1 &= t_1 b \\
x_2 &= t_{21} Ab + t_{20} b \qquad \leftarrow \text{gradient is in the span of } \{b, Ab\} \\
&\vdots \\
x_k &= \sum_{i=0}^{k-1} t_{k-i} A^i b \qquad \leftarrow \text{i.e. it is in the span}\{b, Ab, A^2 b, \ldots, A^{k-1} b\}
\end{aligned}
$$

which is a Krylov subspace generated by A and b. Since $x_{opt} = A^{-1}b$, we want to generate a subspace that is far from this. We will explore this more next lecture.