10-725: Optimization **Fall 2013**

Lecture 6: September 12

Lecturer: Ryan Tibshirani *Scribes: Micol Marchetti-Bowick*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

6.1 Gradient Descent: Convergence Analysis

Last class, we introduced the gradient descent algorithm and described two different approaches for selecting the step size t. The first method was to use a fixed value for t, and the second was to adaptively adjust the step size on each iteration by performing a backtracking line search to choose t. Next, we will discuss the convergence properties of gradient descent in each of these scenarios.

6.1.1 Convergence of gradient descent with fixed step size

Theorem 6.1 *Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and that its gradient is Lipschitz continuous with constant $L > 0$, i.e. we have that $\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2$ for any x, y. Then if we run gradient descent for k iterations with a fixed step size $t \le 1/L$, it will yield a solution $f^{(k)}$ which satisfies*

$$f(x^{(k)}) - f(x^*) \le \frac{\|x^{(0)} - x^*\|_2^2}{2tk}, \tag{6.1}$$

where $f(x^*)$ is the optimal value. Intuitively, this means that gradient descent is guaranteed to converge and that it converges with rate $O(1/k)$.

Proof: Our assumption that ∇f is Lipschitz continuous with constant L implies that $\nabla^2 f(x) \preceq LI$, or equivalently that $\nabla^2 f(x) - LI$ is a negative semidefinite matrix. Using this fact, we can perform a quadratic expansion of f around $f(x)$ and obtain the following inequality:

$$\begin{aligned} f(y) &\le f(x) + \nabla f(x)^T(y - x) + \tfrac{1}{2}\nabla^2 f(x)\|y - x\|_2^2 \\ &\le f(x) + \nabla f(x)^T(y - x) + \tfrac{1}{2}L\|y - x\|_2^2 \end{aligned}$$

Now let's plug in the gradient descent update by letting $y = x^+ = x - t\nabla f(x)$. We then get:

$$\begin{aligned} f(x^+) &\le f(x) + \nabla f(x)^T(x^+ - x) + \tfrac{1}{2}L\|x^+ - x\|_2^2 \\ &= f(x) + \nabla f(x)^T(x - t\nabla f(x) - x) + \tfrac{1}{2}L\|x - t\nabla f(x) - x\|_2^2 \\ &= f(x) - \nabla f(x)^T t\nabla f(x) + \tfrac{1}{2}L\|t\nabla f(x)\|_2^2 \\ &= f(x) - t\|\nabla f(x)\|_2^2 + \tfrac{1}{2}Lt^2\|\nabla f(x)\|_2^2 \\ &= f(x) - \left(1 - \tfrac{1}{2}Lt\right) t\|\nabla f(x)\|_2^2 \end{aligned} \tag{6.2}$$

Using $t \leq 1/L$, we know that $-(1-\frac{1}{2}Lt) = \frac{1}{2}Lt - 1 \leq \frac{1}{2}L(1/L) - 1 = \frac{1}{2} - 1 = -\frac{1}{2}$. Plugging this in to (**??**), we can conclude the following:

$$f(x^+) \leq f(x) - \tfrac{1}{2}t\|\nabla f(x)\|_2^2 \tag{6.3}$$

Since $\frac{1}{2}t\|\nabla f(x)\|_2^2$ will always be positive unless $\nabla f(x) = 0$, this inequality implies that the objective function value strictly decreases with each iteration of gradient descent until it reaches the optimal value $f(x) = f(x^*)$. Note that this convergence result only holds when we choose t to be small enough, i.e. $t \leq 1/L$. This explains why we observe in practice that gradient descent diverges when the step size is too large.

Next, we can bound $f(x^+)$, the objective value at the next iteration, in terms of $f(x^*)$, the optimal objective value. Since f is convex, we can write

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) \\ f(x) &\leq f(x^*) + \nabla f(x)^T(x - x^*) \end{aligned}$$

where the first inequality yields the second through simple rearrangement of terms. Plugging this in to (**??**), we obtain:

$$\begin{aligned} f(x^+) &\leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ f(x^+) - f(x^*) &\leq \frac{1}{2t}\Big(2t\,\nabla f(x)^T(x - x^*) - t^2\|\nabla f(x)\|_2^2\Big) \\ f(x^+) - f(x^*) &\leq \frac{1}{2t}\Big(2t\,\nabla f(x)^T(x - x^*) - t^2\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\Big) \\ f(x^+) - f(x^*) &\leq \frac{1}{2t}\Big(\|x - x^*\|_2^2 - \|x - t\nabla f(x) - x^*\|_2^2\Big) \end{aligned} \tag{6.4}$$

where the final inequality is obtained by observing that expanding the square of $\|x - t\nabla f(x) - x^*\|_2^2$ yields $\|x - x^*\|_2^2 - 2t\,\nabla f(x)^T(x - x^*) + t^2\|\nabla f(x)\|_2^2$. Notice that by definition we have $x^+ = x - t\nabla f(x)$. Plugging this in to (**??**) yields:

$$f(x^+) - f(x^*) \leq \frac{1}{2t}\Big(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\Big) \tag{6.5}$$

This inequality holds for x^+ on every iteration of gradient descent. Summing over iterations, we get:

$$\begin{aligned} \sum_{i=1}^{k} f(x^{(i)} - f(x^*) &\leq \sum_{i=1}^{k} \frac{1}{2t}\Big(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2\Big) \\ &= \frac{1}{2t}\Big(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2\Big) \\ &\leq \frac{1}{2t}\Big(\|x^{(0)} - x^*\|_2^2\Big) \end{aligned} \tag{6.6}$$

where the summation on the right-hand side disappears because it is a telescoping sum. Finally, using the fact that f decreasing on every iteration, we can conclude that

$$\begin{aligned} f(x^{(k)}) - f(x^*) &\leq \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) \\ &\leq \frac{\|x^{(0)} - x^*\|_2^2}{2tk} \end{aligned} \tag{6.7}$$

where in the final step, we plug in (**??**) to get the inequality from (**??**) that we were trying to prove. ■

6.1.2 Convergence of gradient descent with adaptive step size

We will not prove the analogous result for gradient descent with backtracking to adaptively select the step size. Instead, we just present the result with a few comments.

Theorem 6.2 *Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and that its gradient is Lipschitz continuous with constant $L > 0$, i.e. we have that $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ for any x, y. Then if we run gradient descent for k iterations with step size t_i chosen using backtracking line search on each iteration i, it will yield a solution $f^{(k)}$ which satisfies*

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|_2^2}{2t_{\min}k}, \tag{6.8}$$

where $t_{\min} = \min\{1, \beta/L\}$

Notice that the only difference between Theorems **??** and **??** is that the fixed step size t is replaced by $t_{\min}$. Notice that if we choose β to be large enough, the rate of convergence is very similar to what we got for gradient descent with fixed step size.

6.1.3 Convergence rates for gradient descent

Convex f. From Theorem **??**, we know that the convergence rate of gradient descent with convex f is $O(1/k)$, where k is the number of iterations. This implies that in order to achieve a bound of $f(x^{(k)}) - f(x^*) \leq \epsilon$, we must run $O(1/\epsilon)$ iterations of gradient descent. This rate is referred to as "sub-linear convergence."

Strongly convex f. In contrast, if we assume that f is strongly convex, we can show that gradient descent converges with rate $O(c^k)$ for $0 < c < 1$. This means that a bound of $f(x^{(k)}) - f(x^*) \leq \epsilon$ can be achieved using only $O(\log(1/\epsilon))$ iterations. This rate is typically called "linear convergence."

6.1.4 Pros and cons of gradient descent

The principal advantages and disadvantages of gradient descent are:

- Simple algorithm that is easy to implement and each iteration is cheap; just need to compute a gradient
- Can be very fast for smooth objective functions, i.e. well-conditioned and strongly convex
- However, it's often slow because many interesting problems are not strongly convex
- Cannot handle non-differentiable functions (biggest downside)

6.2 Subgradients

Subgradients are the analog to gradients for non-differentiable functions. They are one of the fundamental mathematical concepts underlying convexity.

Definition 6.3 *A subgradient of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at some point x is any vector $g \in \mathbb{R}^n$ that achieves the same lower bound as the tangent line to f at x, i.e. we have*

$$f(y) \geq f(x) + g^T(y - x) \qquad \forall\ x, y$$

The subgradient g always exists for convex functions on the relative interior of their domain. Furthermore, if f is differentiable at x, then there is a unique subgradient $g = \nabla f(x)$. Note that subgradients need not exist for nonconvex functions (for example, cubic functions do not have subgradients at their inflection points).

6.2.1 Examples of subgradients

absolute value. $f(x) = |x|$. Where f is differentiable, the subgradient is identical to the gradient, sign(x). At the point $x = 0$, the subgradient is any point in the range $[-1, 1]$ because any line passing through $x = 0$ with a slope in this range will lower bound the function.

ℓ_2 norm. $f(x) = \|x\|_2$. For $x \neq 0$, f is differentiable and the unique subgradient is given by $g = x/\|x\|_2$. For $x = 0$, the subgradient is any vector whose ℓ_2 norm is at most 1. This holds because, by definition, in order for g to be a subgradient of f we must have that

$$f(y) = \|y\|_2 \geq f(x) + g^T(y - x) = g^T y \qquad \forall\ y.$$

In order for $\|y\|_2 \geq g^T y$ to hold, g must have $\|g\|_2 \leq 1$.

ℓ_1 norm. $f(x) = \|x\|_1$. Since $\|x\|_1 = \sum_{i=1}^n |x_i|$, we can consider each element g_i of the subgradient separately. The result is very analogous to the subgradient of the absolute value function. For $x_i \neq 0$, $g_i = \text{sign}(g_i)$. For $x_i = 0$, $g_i isanypointin[-1, 1]$.

maximum of two functions. $f(x) = \max\{f_1(x), f_2(x)\}$, where f_1 and f_2 are convex and differentiable. Here we must consider three cases. First, if $f_1(x) > f_2(x)$, then $f(x) = f_1(x)$ and therefore there is a unique subgradient $g = \nabla f_1(x)$. Likewise, if $f_2(x) > f_1(x)$, then $f(x) = f_2(x)$ and $g = \nabla f_2(x)$. Finally, if $f_1(x) = f_2(x)$, then f may not be differentiable at x and the subgradient will be any point on the line segment that joints $\nabla f_1(x)$ and $\nabla f_2(x)$.

6.2.2 Subdifferential

Definition 6.4 *The subdifferential of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at some point x is the set of all subgradients of f at x, i.e. we say*

$$\partial f(x) = \{g \in \mathbb{R}^n : g \textit{ is a subgradient of } f \textit{ at } x\}$$

An important property of the subdifferential $\partial f(x)$ is that it is a closed and convex set, which holds even for nonconvex f. To verify this, suppose we have two subgradients $g_1, g_2 \in \partial f(x)$. We need to show that $g_0 = \alpha g_1 + (1 - \alpha) g_2$ is also in $\partial f(x)$ for arbitrary α. If we write the following inequalities,

$$\alpha\Big(f(y) \geq f(x) + g_1^T(y - x)\Big)\alpha$$
$$(1 - \alpha)\Big(f(y) \geq f(x) + g_2^T(y - x)\Big)(1 - \alpha),$$

which follow from the definition of subgradient applied to g_1 and g_2, we can add them together to yield $f(y) \geq f(x) + \alpha\, g_1^T(y - x) + (1 - \alpha)\, g_2^T(y - x) = g_0^T(y - x)$.

6.2.3 Connection between sugradients and convex geometry

Suppose we have a convex set $C \subseteq \mathbb{R}^n$ and consider the indicator function $\mathbb{I}_C : \mathbb{R}^n \to \mathbb{R}$, defined by

$$\mathbb{I}_C(x) = \mathbb{I}\{x \in C\} = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

We would like to determine the subgradients of $\mathbb{I}_C$. If $x \in C$, $\partial \mathbb{I}_C = N_C(x)$, where $N_C(x)$ is the normal cone of C at x and is defined as

$$N_C(x) = \{g \in \mathbb{R}^n : g^T x \geq g^T y \text{ for any } y \in C\}$$

This result comes directly out of the definition of subgradient $\mathbb{I}_C(y) \geq \mathbb{I}_C(x) + g^T(y - x)$.

The subgradients of indicator functions are interesting for the following reason. Eventually, we will prove that a point x^* minimizes a particular function f if and only if $0 \in \partial f(x^*)$. Now suppose we want to know whether x^* minimized $f(x)$ subject to the constraint that x be a member of the set C. We can rewrite this optimization problem as

$$\min_x f(x) + \mathbb{I}_C(x)$$

Thus we can determine whether x^* is a solution to this problem by checking whether $0 \in \partial\,(f(x^*) + \mathbb{I}_C(x^*))$.

6.2.4 Subgradient calculus

Subgradients can be computed by knowing the subgradients for a basic set of functions and then applying the rules of subgradient calculus. Here are the set of rules.

Scaling. $\partial(af) = a \cdot \partial f$ provided that $a > 0$

Addition. $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$

Affine composition. If $g(x) = f(Ax) + b$ then

$$\partial g(x) = A^T \partial f(Ax + b)$$

Finite pointwise maximum. If $f(x) = \max_{i=1,\ldots,m} f_i(x)$ then

$$\partial f(x) = \operatorname{conv}\left(\bigcup_{i : f_i(x) = f(x)} \partial f_i(x)\right)$$

General pointwise maximum. If $f(x) = \max_{s \in \mathcal{S}} f_s(x)$ then

$$\partial f(x) \supseteq \operatorname{cl}\left\{\operatorname{conv}\left(\bigcup_{s : f_s(x) = f(x)} \partial f_s(x)\right)\right\}$$