EE 381V: Large Scale Optimization **Fall 2012**

Lecture 4 — September 11

Lecturer: Caramanis & Sanghavi *Scribe: Gezheng Wen, Li Fan*

4.1 Gradient Descent

The idea relies on the fact that $-\nabla f(x^{(k)})$ is a descent direction.

Algorithm description

$$x^{(k+1)} = x^{(k)} - \eta^{(k)} \nabla f(x^{(k)}) \tag{4.1}$$

One important parameter to control is the step sizes $\eta^{(k)} > 0$. Too small values of $\eta^{(k)}$ will cause our algorithm to converge very slowly. On the other hand, too large η could cause our algorithm to overshoot the minima and diverge.
Intuitively, at each iterate, we would like to ensure that the next step taken by this algorithm results in a smaller function value at the next iterate.

Definition: *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called L-Lipschitz if and only if*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L||x - y||_2, \qquad \forall x, y \in \mathbb{R}^n \tag{4.2}$$

We denote this condition by $f \in C_L$, where C_L is the class of *L-Lipschitz* functions.

Lemma 4.1. *If $f \in C_L$, then $|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}||y - x||^2$*

Proof: Refer text. □

Theorem 4.2. *If $f \in C_L$ and $f^* = \min_x f(x) > -\infty$, then the gradient descent algorithm with fixed step size satisfying $\eta < \frac{2}{L}$ will converge to a stationary point.*

Proof: Let $x^+ = x - \eta \nabla f(x)$. Using lemma (4.1), we can write

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

This leads to:

$$\|\nabla f(x)\|^2 \leq \frac{1}{\eta(1-\frac{\eta L}{2})}(f(x) - f(x^+))$$

$\Rightarrow$

$$\begin{aligned}\sum_{k=1}^{N} \|\nabla f(x^{(k)})\|^2 &\leq \frac{1}{\eta(1-\frac{\eta}{2}L)}(f(x^{(0)}) - f(x^{(N)}) \\ &\leq \frac{1}{\eta(1-\frac{\eta}{2}L)}(f(x^{(0)}) - f^*)\end{aligned}$$

This implies that $\lim_{k\to\infty} \nabla f(x^{(k)}) = 0$. [1] $\square$

Suppose that the function $f(x)$ is convex. The above automatically implies that $x^{(k)}$ converges to an optimum x^*. Now, the natural next question is how fast this happens. For this, two natural metrics are: how fast do the following decrease to 0:

- $f(x^{(k)}) - f^*$
- $\|x^{(k)} - x^*\|$

Note that if we only impose that f be convex, we cannot ensure an upper bound on these two metrics. Examples are functions which are really flat at their minima (Figure 4.1).

Figure 4.1. Visualize convex function: (left)descent direction (right)a 'flat' convex function

[1] One can actually show that $\nabla f(x^{(k)})$ goes to 0 as $\frac{c}{\sqrt{k}}$, where the constant c depends on $x^{(0)}$

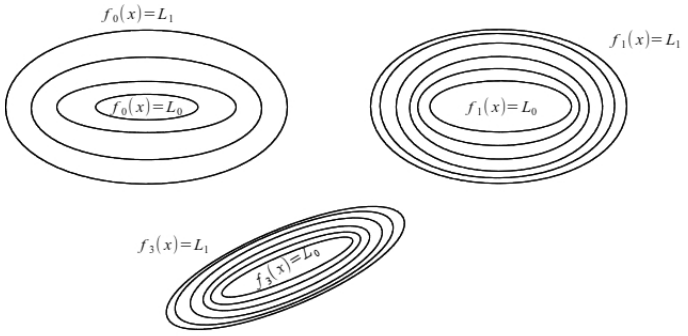

Figure 4.2. Sublevel sets for functions f_0, f_1, and f_3

Sublevel Set

A sublevel set S_L is defined as: $S_L = \{x : f(x) \leq L\}$. We note that this set is convex. Sublevel sets show if the gradient of the function $f(x)$ is steep or flat in a given direction of descent and also if the gradients are asymmetric in different directions. Sublevel sets are convenient for visualizing convex functions.

For example, in Figure (4.2), $f_0(x)$ has flat gradients while $f_1(x)$ has steeper gradients. Again, $f_3(x)$ has different gradients in different directions as shown in the figure.

Fixed Step Size:

Some gradient descent methods tend to use fixed step size for simplicity but the choice of appropriate step sizes is not easy. As shown in Figure (4.3), a too small η will cause the algorithm to converge very slowly. On the other hand, a too large η could cause the algorithm to overshoot the minima and diverge.

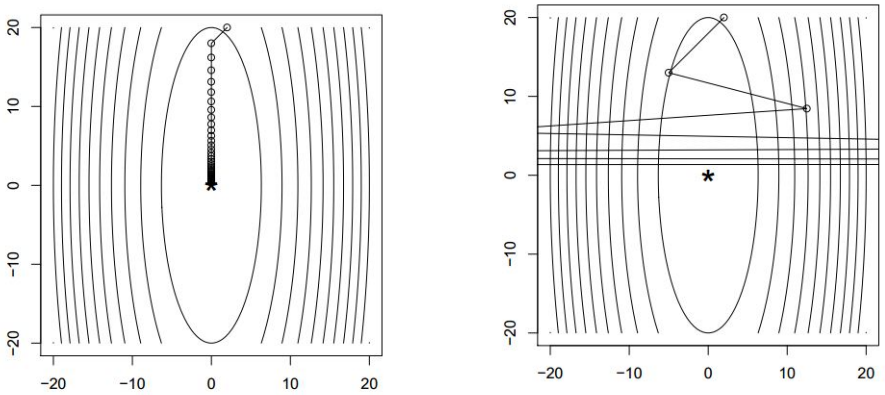

Figure 4.3. (left)a too small step size that leads to slow convergence; (right)a too large step size that leads to divergence

4.1.1 Strong Convexity and implications

Definition: If there exist a constant $m > 0$ such that $\nabla^2 f \succeq mI$ for all $x \in S$, then the function $f(x)$ is a *strongly convex function* on S.

When $m = 0$, we recover the basic inequality characterizing convexity; for $m > 0$, we obtain a better lower bound on $f(y)$ than that from convexity alone. The value of m reflects the shape of convex functions. Typically as shown in Figure (4.4) , a small m corresponds to a 'flat' convex function while a large m corresponds to a 'steep' convex function.

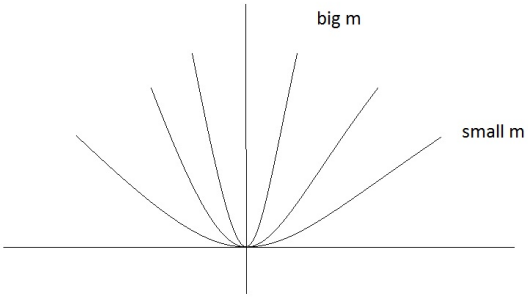

Figure 4.4. A strongly convex function with different parameter m. The larger m is, the steeper the function looks like.

Lemma 4.3. *If f is strongly convex on S, we have the following inequality:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|^2 \tag{4.3}$$

for all x and y in S.

Proof: For $x, y \in S$, we have

$$f(y) \;=\; f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x)$$

for some z on the line segment $[x, y]$. By the strong convexity assumption, the last term on the righthand side is at least $\frac{m}{2}\|y - x\|_2^2$ □

Strong convexity has several interesting consequences. We will first show that the inequality can be used to bound $f(x) - f^*$, which is the suboptimality of the point x, in terms of $\|\nabla f(x)\|_2$. The righthand size is a convex quadratic function of y (for fixed x). Setting the

gradient with respect to y equal to zero, we find that $\tilde{y} = x - (1/m)\nabla f(x)$ minimizes the righthand side. Therefore we have

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|^2 \\ &\geq f(x) + \langle \nabla f(x), \tilde{y} - x \rangle + \frac{m}{2}\|\tilde{y} - x\|^2 \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \end{aligned}$$

Since this holds for any $y \in S$. we have

$$f^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \tag{4.4}$$

This allows us to realize how fast you get to a minimum as a function of gradient. If the gradient is small at a point, then the point is nearly optimal.

Similarly, we can also derive a bound on $\|x - x^*\|_2$, the distance between x and any optimal point x^*, in terms of $\|\nabla f(x)\|_2$:

$$\|x - x^*\|_2 \leq \frac{2}{m}\|\nabla f(x)\|_2^2 \tag{4.5}$$

where $x^* = \arg\min_x f(x)$.

To see this, we apply (4.3) with $y = x^*$ to obtain:

$$\begin{aligned} f^* = f(x^*) &\geq f(x) + < \nabla f(x), x^* - x > + \frac{m}{2}\|x^* - x\|_2^2 \\ &\geq f(x) - \|\nabla f(x)\|_2\|x^* - x\|_2 + \frac{m}{2}\|x^* - x\|_2^2, \end{aligned}$$

where we use the Cauchy-Schwarz inequality in the second inequality. Since $f^* \leq f(x)$, we must have

$$-\|\nabla f(x)\|_2 - \|x^* - x\|_2 + \frac{m}{2}\|x^* - x\|_2^2 \leq 0,$$

from which (4.5) follows. One consequence of (4.5) is x^* is unique and the solution locates within a ball of radius of $\|\nabla f(x)\|^2$ around the optimal solution.

4.1.2 Upper Bound on $\nabla^2 f(x)$

The inequality (4.3) implies that the sublevel sets contained in S are bounded, so in particular, S is bounded. Therefore the maximum eigenvalue of $\nabla^2 f(x)$, which is a continuous function of x on S, is bounded above on S. And there exists a constant M such that $\nabla^2 f(x) \preceq MI$ for all $x \in S$.

Lemma 4.4. *For any $x, y \in S$, if $\nabla^2 f(x) \preceq MI$ for all $x \in S$ then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2 \tag{4.6}$$

Proof: The proof is analogous to the proof of (4.3). □

4.1.3 Condition Number

From the strong convexity inequality (4.3)and the inequality (4.6), we have:

$$mI \preceq \nabla^2 f(x) \preceq MI \tag{4.7}$$

for all $x \in S$. The ratio $k = M/m$ is thus an upper bound on the condition number of the matrix $\nabla^2 f(\mathbf{x})$, i.e., the ratio of its largest eigenvalue to its smallest eigenvalue.
When the ratio is close to 1, we call it *well-conditioned*. When the ratio is much larger than 1, we call it *ill-conditioned*. When the ratio is exactly 1, it is the best case that only one step will lead to the optimal solution (i.e., there is no wrong direction).

It must be kept in mind that constants m and M are known only in the rare cases, so the the inequality cannot be used as a practical stopping criterion. It can be considered as *conceptual* stopping criterion; it shows that if the gradient of f at x is small enough, then the difference between $f(x)$ and f^* is small. If we terminate an algorithm when $\|\nabla f(x^k)\|_2 \leq \eta$, where η is chosen small enough to be smaller than $(m\epsilon)^{\frac{1}{2}}$, then we have $f(x^k) - f^* \leq \epsilon$, where ϵ is some positive tolerance.

Though these bounds involve the (usually) unknown constants m and M, they establish that the algorithm converges, even if the bound on the number of iterations required to reach a given accuracy depends on constants that are unknown.

See in text for the discussion about condition number of shape of level sets.

Theorem 4.5. *Gradient descent for a strongly convex function f and step size $\eta = 1/M$ will converge as*

$$f(x^*) - f^* \leq c^k (f(x^0) - f^*), \tag{4.8}$$

where $c \leq 1 - \frac{m}{M}$.

Since we usually do not know the value of M, we do line search. The following sections introduce the two line search methods: exact line search and backtracking line search. The proof of (4.5) is also provided.

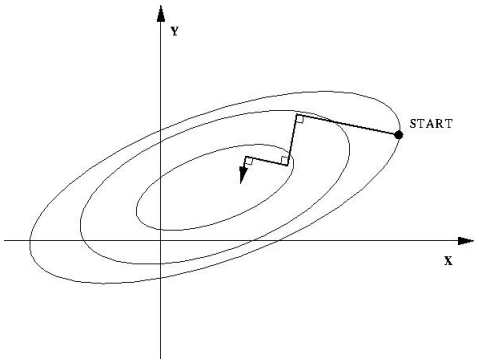

Figure 4.5. Exact Line Search

4.1.4 Exact Line Search

The optimal line search method is *exact line search*, in which η is chosen to minimize f along the ray $\{x - \eta\nabla f(x)\}$, as shown in Figure (4.5)

Algorithm (Gradient descent with exact line search)

1. Set iteration counter $k = 0$, and make an initial guess x_0 for the minimum
2. Compute $\nabla f(x^{(k)})$
3. Choose $\eta^{(k)} = \arg\min_{\eta}\{f\left(x^{(k)} - \eta\nabla f(x^{(k)})\right)\}$
4. Update $x^{(k+1)} = x^{(k)} - \eta^{(k)}\nabla f(x^{(k)})$ and $k = k + 1$.
5. Go to 2 until $\|\nabla f(x^{(k)})\| < \epsilon$

An exact line search is used when the cost of the minimization problem with one variable is low compared to the cost of computing the search direction itself. However, the algorithm is not very practical.

Convergence Analysis

$$\begin{aligned} f(x^+) &\leq f(x - \frac{1}{M}\nabla f(x)) \\ &\leq f(x) - \frac{1}{M}\|\nabla f(x)\|_2^2 + \frac{M}{2}(\frac{1}{M})^2\|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2 \end{aligned}$$

$\Rightarrow$

$$f(x^+) - f^* \leq f(x) - f^* - \frac{1}{2M}\|\nabla f(x)\|_2^2 \tag{4.9}$$

Recall the analysis for strong convexity:

$$\|\nabla f(x)\|_2^2 \;\geq\; 2m(f(x) - f(y))$$

Thus, the following inequality holds:

$$f(x^+) - f^* \;\leq\; \left(1 - \frac{m}{M}\right)(f(x) - f^*) \tag{4.10}$$

Thus we see that $|f(x^{(k)}) - f^*|$ decreases by at least a constant factor in every iteration, converging to 0 geometrically fast. This is commonly called *linear convergence* (as the log-log plot is linear).

4.1.5 Backtracking Line Search

In (unconstrained) optimization, the *backtracking line search* strategy is used as part of a line search method, to compute how far one should move along a given search direction. Usually it is undesirable to exactly minimize the function in the generic line search algorithm. One way to inexactly minimize is by finding an that gives a sufficient decrease in the objective function $f : \mathbb{R}^n \to \mathbb{R}$.

Backtracking line search is very simple and quite effective, and it depends on two constants α, β with $0 < \alpha < 0.5, 0 < \beta < 1$. It starts with unit step size and then reduces it by the factor β until the stopping condition $f(x - \eta\nabla f(x)) \leq f(x) - \alpha\eta\|\nabla f(x)\|^2$. Since $-\nabla f(x)$ is a descent direction and $-\|\nabla f(x)\|^2 < 0$, so for small enough step size η, we have:

$$f(x - \eta\nabla f(x)) \approx f(x) - \eta\|\nabla f(x)\|^2 < f(x) - \alpha\eta\|\nabla f(x)\|^2, \tag{4.11}$$

which shows that the backtracking line search eventually terminates. The constant α can be interpreted as the fraction of the decrease in f predicted by linear extrapolation that we we will accept.

The backtracking condition is illustrated in figure, which shows that the exit inequality holds for $\eta \geq 0$ in an interval $[0, \eta_0]$. It follows that the backtracking line search stops with a step size η that satisfies: $\eta = 1$, or $\eta \in (\beta\eta_0, \eta_0]$.

Algorithm

1. Set iteration counter $k = 0$. Make an initial guess x^0 and choose initial $\eta = 1$.
2. Update $\eta^k = \beta\eta^k$
3. Go to 2 until $f(x^k - \eta^k\nabla f(x^k)) \leq f(x^k) - \alpha\eta^k\|\nabla f(x^k)\|^2$.
4. Calculate $x^{k+1} = x^k - \eta^k\nabla f(x^k)$ and update $k = k + 1$.

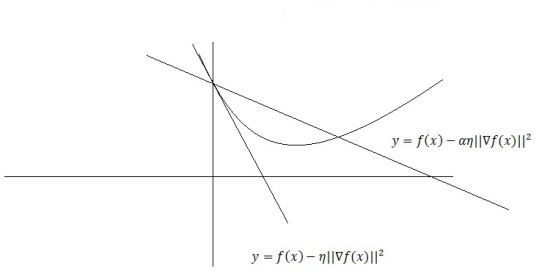

Figure 4.6. Backtracking line search

5. Go to 1 until $\|\nabla f(x^{(k)})\| < \epsilon$

The parameter α is typically chosen between 0.01 and 0.3, meaning that we accept a decrease in f between 1% and 30% of the prediction based on the extrapolation. The parameter β is often chosen to be between 0.1 (which corresponds to a very crude search) and 0.8(which corresponds to a less crude search).

Convergence Analysis

Claim: $\eta \leq \frac{1}{M}$ always satisfies the stopping condition.

Proof: Recall:

$$f(x^+) \quad \leq \quad f(x) - \eta\|\nabla f(x)\|^2 + \frac{\eta^2 M}{2}\|\nabla f(x)\|^2$$

With the assumption that $\eta \leq \frac{1}{M}$, the inequality implies that:

$$f(x^+) \quad \leq \quad f(x) - \frac{\eta}{2}\|\nabla f(x)\|^2$$

$\Rightarrow$

$$\eta \geq \frac{\beta}{M}$$

So overall,

$$\eta \quad \geq \quad \min(1, \frac{\beta}{M}) \tag{4.12}$$

$$f(x^+) \quad \leq \quad f(x) - \alpha \min(1, \frac{\beta}{M})|\nabla f(x)\|^2 \tag{4.13}$$

Now, we subtract $f*$ from both sides to get:

$$f(x^+) - f^* \;\leq\; f(x) - f^* - \alpha \min(1, \frac{\beta}{M})\|\nabla f(x)\|_2^2,$$

and combines with $\|\nabla f(x)\|_2^2 \geq 2m(f(x) - f^*)$ to obtain:

$$f(x^+) - f^* \;\leq\; (1 - \alpha \min(1, \frac{\beta}{M})(f(x) - f^*),$$

where

$$c \;=\; 1 - 2m\alpha \min\{1, \frac{\beta}{M}\} < 1 \tag{4.14}$$

□

In particular, $f(x^k)$ converges to f^* at least as fast as a geometric series with an exponent that depends (at least in part) on the condition number bound $\frac{M}{m}$. As before with exact line search, the convergence is at least linear (but with a different factor).