10-725: Optimization **Fall 2012**

Lecture 5: Gradient Desent Revisited

Lecturer: Geoff Gordon/Ryan Tibshirani *Scribes: Cong Lu/Yu Zhao*

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.*

5.1 Choose step size

Recall that we have $f : \mathbb{R}^n \to \mathbb{R}$, convex and differentiable. We want to solve

$$\min_{x \in \mathbb{R}^n} f(x)$$

i.e, to find $x^\star$ such that $f(x^\star)$=min $f(x)$.

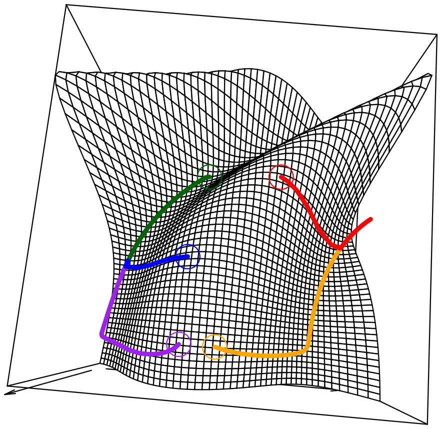
Gradient descent: choose initial $x^{(0)} \in \mathbb{R}^n$, repeat :

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), k = 1, 2, 3, ...$$

Stop at some point(When to stop is quite dependent on what problems you are looking at).

Figure 5.1 is shows a example that we cannot always continue and it depends where we start. i.e. If we start at a spot somewhere between the purple and orange, it would stay there and go nowhere.

Figure 5.1:



At each iteration, consider the expansion

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|^2$$

We can use quadratic approximation, replacing usual $\nabla^2\ f(x)$ by $\frac{1}{t}I$, then we have

$$f(x) + \nabla f(x)^T(y-x),$$

which is a linear combination to f, and

$$\frac{1}{2t}\|y-x\|^2,$$

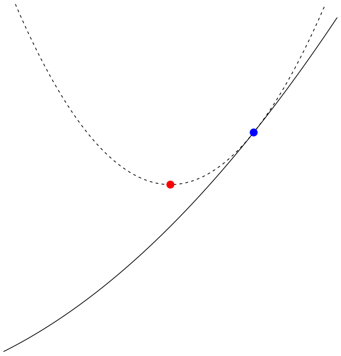
which is a proximity term to x, with weight $\frac{1}{2t}$.

Then, choose next point $y = x^+$ to minimize quadratic approximation

$$x^+ = x - t\nabla f(x)$$

as shown in Figure 5.2.

Figure 5.2:



blue point is x, red point is x^+

5.1.1 Fixed step size

Simply take $t_k = t$ for all $k = 1, 2, 3, ,$ can diverge if t is too big. Consider $f(x) = (10{x_1}^2 + {x_2}^2/2)$, Figure 5.3 shows the gradient descent after 8 steps. It can be slow if t is too small . As for the same example, gradient descent after 100 steps in Figure 5.4, and gradient descent after 40 appropriately sized steps in Figure 5.5.

Convergence analysis will give us a better idea which one is just right.

5.1.2 Backtracking line search

Adaptively choose the step size:

First, fix a parameter $0 < \beta < 1$, then at each iteration, start with $t = 1$, and while

$$f(x - \nabla f(x)) > f(x) - \frac{t}{2}\|\nabla f(x)\|^2,$$

update $t = \beta t$, as shown in Figure 5.6 (from B & V page 465), for us $\triangle x = -\nabla f(x)$, $\alpha = 1/2$.

Backtracking line search is simple and work pretty well in practice. Figure 5.7 shows that backpacking picks up roughly the right step size(13 steps) for the same example, with $\beta = 0.8$ (B & V recommend $\beta \in (0.1, 0.8)$).

Figure 5.3:

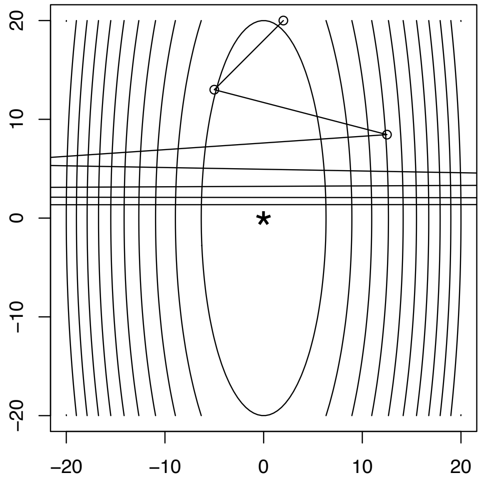
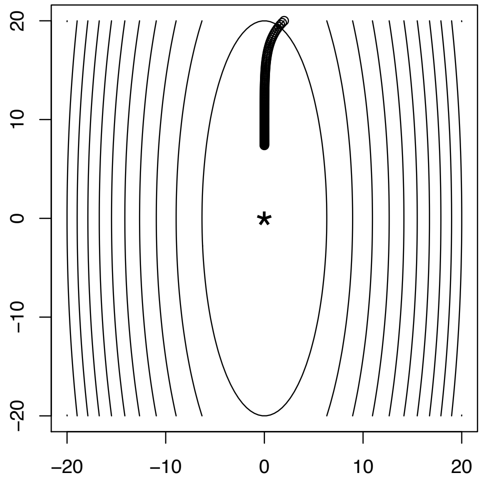


Figure 5.4:



5.1.3 Exact line search

At each iteration, do the best e can along the direction of the gradient,

$$t = \underset{s\geq 0}{\text{argmin}} f(x - s\nabla f(x)).$$

Usually, it is not possible to do this minimization exactly.

Approximations to exact line search are often not much more efficient than backtracking, and it's not worth it.

Figure 5.5:

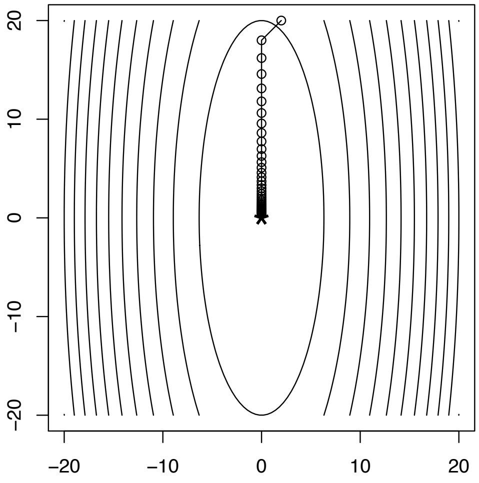
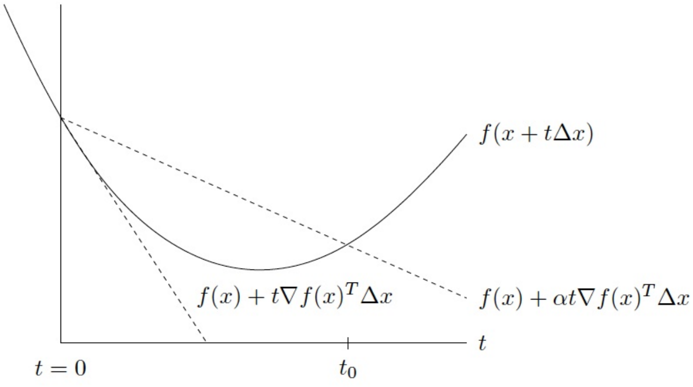


Figure 5.6:


$f(x + t\Delta x)$
$f(x) + t\nabla f(x)^T \Delta x$
$f(x) + \alpha t\nabla f(x)^T \Delta x$
t
$t = 0$
t_0

5.2 Convergence analysis

5.2.1 Convergence analysis for fixed step size

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and additionally

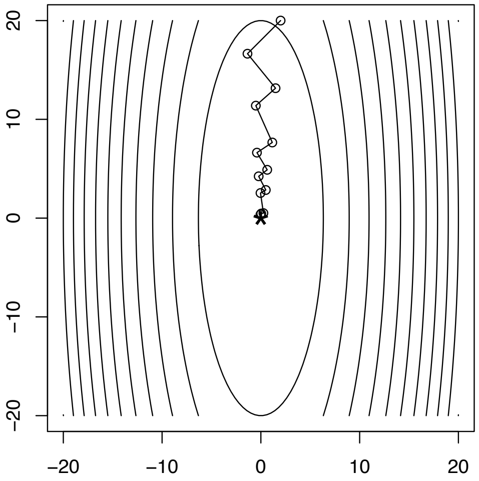
$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| for\ any\ x, y$$

i.e. , ∇f is Lipschitz continuous with constant $L > 0$

Theorem 5.1 *Gradient descent with fixed step size $t \leq 1/L$ satisfies*

$$f(x^{(k)} - f(x^\star) \leq \frac{\|x^{(0)} - x^\star\|^2}{2tk}$$

Figure 5.7:



i.e. gradient descent has convergence rate $O(1/k)$
i.e. to get $f(x^{(k)}) - f(x^\star) \leq \epsilon$, we need $O(1/\epsilon)$ iterations

Proof: Since ∇f Lipschitz with constant L, which means $\nabla^2 f \preceq LI$, we have $\forall x, y, z$

$$(x-y)^T(\nabla^2 f(z) - LI)(x-y) \leq 0$$

Which means

$$L\|x-y\|^2 \geq (x-y)^T \nabla^2 f(z)(x-y)$$

Based on Taylor's Remainder Theorem, we have $\forall x, y$, $\exists z \in [x, y]$

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(x-y)^T\nabla^2 f(z)(x-y) \\ &\leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2 \end{aligned} \tag{5.1}$$

Plugging in $x^+ = x - t\nabla f(x)$,

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)^T(x - t\nabla x - x) + \frac{L}{2}\|x - t\nabla x - x\|^2 \\ &= f(x) - (1 - \frac{Lt}{2})t\|\nabla f(x)\|^2 \end{aligned} \tag{5.2}$$

Taking $0 < t \leq 1/L$, $1 - Lt/2 \geq 1/2$, we have

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

Since f is convex, $f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$ we have

$$
\begin{aligned}
f(x^+) &\leq f(x) - \frac{t}{2}\|\nabla f(x)\|^2 \\
&\leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{t}{2}\|\nabla f(x)\|^2 \\
&= f(x^*) + \frac{1}{2t}(\|x - x^*\|^2 - \|x - x^* - t\nabla f(x)\|^2) \\
&= f(x^*) + \frac{1}{2t}(\|x - x^*\|^2 - \|x^+ - x^*\|^2)
\end{aligned}
\tag{5.3}
$$

Summing over iterations, we have

$$
\begin{aligned}
\sum_{i=1}^{k}(f(x^{(i)} - f(x^*)) &\leq \frac{1}{2t}(\|x^{(0)} - x^*\|^2 - \|x^{(k)} - x^*\|^2) \\
&\leq \frac{1}{2t}\|x^{(0)} - x^*\|^2
\end{aligned}
\tag{5.4}
$$

From (**??**), we can see that $f(x^{(k)})$ is nonincreasing. Then we have

$$
f(x^{(k)}) - f(x^*) \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)} - f(x^*)) \leq \frac{\|x^{(0)} - x^*\|^2}{2tk}
$$

■

5.2.2 Convergence analysis for backtracking

For backtracking, it's the same assumptions, $f : \mathbb{R}_n \to \mathbb{R}$ is convex and differentiable, and ∇f is Lipschitz continuous with constant $L > 0$.

But we don't have to choose a step size that is small or equal to $1/L$ to begin with. We just get the same rate assuming that the function is Lipschitz.

Theorem 5.2 *Gradient descent with backtracking line search satisfies*

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

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

So the gradient descent has convergence rate $O(1/k)$. The constants are the same as there before, but since t is adapted in each iteration, we replace t by t_{min}, where $t_{min} = \min\{1, \beta/L\}$.

If β is not very tiny, then we don't lose much compared to fixed step size (β/L vs $1/L$).

The proof is very similar to the proof of fixed step theorem.

5.2.3 Convergence analysis for strong convexity

There is also a statement of convergence on strong convexity. Strong convexity is a condition that the smallest eigenvalue of the Hessian matrix of function f is uniformly bounded for any x, which means for some $d > 0$,

$$\nabla f(x) \succeq dI, \forall x$$

Then the function has a better lower bound than that from usual convexity:

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

The strong convexity adds a quadratic term and still has a lower bound. If a function has both strong convexity and Lipschitz assumption, it has both lower and upper bound by quadratics. We will have some strong things about it since the function is well behaved.

Theorem 5.3 *Gradient descent with fixed step size $t \leq 2/(d + L)$ or with backtracking line search satisfies*

$$f(x^{(k)}) - f(x^*) \leq c^k \frac{L}{2}\|x^{(0)} - x^*\|^2$$

where $0 < c < 1$.

The proof is on the textbook.

Under strong convextiy and Lipschitz assumption, we have a theorem that it goes better than $1/k$ and the rate is $O(c^k)$, which is exponentially fast. It is called *linear convergence*, because if we plot iterations on the x-axis, and we plot difference in the function values on the y-axis on a log scale, it looks like a linear straight line. If we want $f(x^{(k)} - f(x^*) \leq \epsilon$, we need $O(\log(1/\epsilon))$ iterations.

The constant c depends adversely on condition number L/d. If the condtion number is very high, that is a slower rate.

5.2.4 How realistic are these conditions?

How realistic is Lipschitz continuity of ∇f? This means $\nabla^2 f(x) \preceq LI$.

For example, consider linear regression

$$f(x) = \frac{1}{2}\|y - Ax\|^2$$

Then

$$\nabla f(x) = -A^T(y - Ax)$$

$$\nabla^2 f(x) = A^T A$$

Take $L = \sigma^2_{max}(A) = \|A\|^2$, then $\nabla^2 f(x) = A^T A \preceq LI$. ∇f Lipschitz with L. Then we can choose a fixed step size that is smaller than $1/L$ or use backtracking search to get a converge rate of $O(1/k)$.

How realistic is strong convexity of f? Recall this is $\nabla^2 f(x) \succeq dI$.

That is not easily realistic. Again consider

$$f(x) = \frac{1}{2}\|y - Ax\|^2$$

Now we need $d = \sigma_{min}^2(A)$.

If A is wide, then $\sigma_{min}(A) = 0$, and f can't be strongly convex.

Even if $\sigma_{min}(A) > 0$, we can still have a very large condition number $L/d = \sigma_{max}(A)/\sigma_{min}(A)$.

5.3 Pracalities

5.3.1 Stopping rule

We can basicly stop when the gradient $\|\nabla f(x)\|$ is small. It is reasonable because $\nabla f(x^*) = 0$. If $\|\nabla f(x)\|$ is small, we think that $f(x)$ is close to the minimum $f(x^*)$.

If f is strongly convex with parameter d, then

$$\|\nabla f(x)\| \leq \sqrt{2d\epsilon} \Rightarrow f(x) - f(x^*) \leq \epsilon$$

5.3.2 Pros and cons

Pros:

- It is a simple idea, and each iteration is cheap.
- It is very fast for well-conditioned, strongly convex problems.

Cons:

- It is often slow, because interesting problems aren't strongly convex or well-conditioned.
- It can't handle nondifferentiable functions.

5.4 Forward stagewise regression

5.4.1 Forward stagewise regression

Let's go back to the linear regression function

$$f(x) = \frac{1}{2}\|y - Ax\|^2$$

A is $n \times p$, its columns $A_1, \ldots, A_p$ are predictor variables.

Forward stagewise regression is the algorithm below:

```
Start with $x^{(0)} = 0$, **repeat**
    Find variable $i$ such that $|A_i^T r|$ is largest, for $r = y - Ax^{(k-1)}$(largest absolute correlation with residual)
    Update $x_i^{(k)} = x_i^{(k-1)} + \gamma \cdot \text{sign}(A_i^T r)$
```

Here $\gamma > 0$ is small and fixed, called learning rate.

In each iteration, forward stagewise regression just updates one of the variables in x with a small rate γ.

5.4.2 Steepest descent

It is a close cousin to gradient descent and just change the choice of norm.

Let's suppose q, r are complementary: $1/q + 1/r = 1$.

Steepest descent just update $x^+ = x + t \cdot \Delta x$, where

$$\Delta x = \|u\|_r \cdot u$$

$$u = \underset{\|v\|_q \leq 1}{\operatorname{argmin}} \nabla f(x)^T v$$

If $q = 2$, then $\Delta x = -\nabla f(x)$, which is exactly gradient descent.

If $q = 1$, then $\Delta x = -\partial f(x)/\partial x_i \cdot e_i$, where

$$\left| \frac{\partial f}{\partial x_i}(x) \right| = \max_{j=1,\dots,n} \left| \frac{\partial f}{\partial x_j}(x) \right| = \|\nabla f(x)\|_\infty$$

The normalized steepest descent just takes $\Delta x = u$ (unit q-norm).

5.4.3 Equivalence

Normalized steepest descent with 1-norm: updates are

$$x_i^+ = x_i - t \cdot \operatorname{sign} \left\{ \frac{\partial f}{\partial x_i}(x) \right\}$$

where i is the largest component of $\nabla f(x)$ in absolute value.

Compare forward stagewise: updates are

$$x_i^+ = x_i + \gamma \cdot \operatorname{sign}(A_i^T r), r = y - Ax$$

Recall here $\Delta f(x) = -A_i^T(y - Ax)$, so

$$\frac{\partial f}{\partial x_i}(x) = -A_i^T(y - Ax)$$

Forward stagewise regression is exactly normalized steepest descent under 1-norm.

5.4.4 Early stopping and regularization

Forward stagewise is like a slower version of forward stepwise. If we stop early, i.e.m don't continue all the way to the least squares solution, then we get a sparese approximation. Can this be used as a form of regularization?

Recall lasso problem:

$$\min_{\|x\|_1 \leq t} \frac{1}{2}\|y - Ax\|^2$$

Solution $x^*(t)$, as function of t, also exhibits varying amounts of regularization.

For some problems (some y, A), with a small enough step size, forward stagewise iterates trace out lasso solution path.