MIT 6.854/18.415: Advanced Algorithms Spring 2016

Lecture 15 – March 30, 2016

Prof. Ankur Moitra *Scribe: Amin Manna, Andrew Xia, Brandon Araki*

1 Last Time

Linear Programming Relaxations, Vertex Cover, Set Cover, Primal Dual method

2 Gradient Descent

There are many variations, we'll focus on **Unconstrained Minimization**. Given a convex, differentiable function, $f : \mathbb{R}^n \to \mathbb{R}$

A convex function :

$$f(\lambda x + (1+\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \quad \forall x, y \in \mathbb{R}^n, \lambda \in [0,1]$$

We want to minimize f(x).

2.1 Gradient Descent

Let's start with n = 1 case.

for $t = 1$ to T
set $X_{t+1} = X_t - \eta f'(X_t)$

where η is the learning rate.

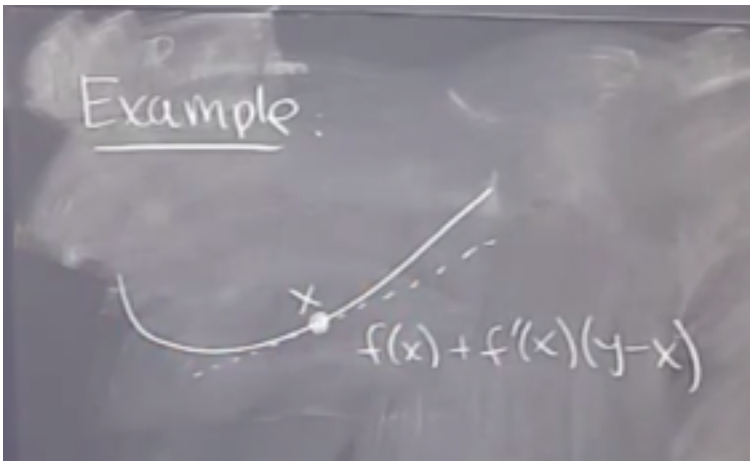
Recall, if f is twice differentiable: **Taylor's Thm:**

$$f(y) = f(x) + f'(x)(y-x) + f''(x)/2(y-x)^2 + o((y-x)^2)$$

Fact: If f is twice differentiable, then f is convex $\iff$ $f''(x) \geq 0 \forall x$.

Lagrange Remainder:
$f(y) = f(x) + f'(x)(y-x) + \frac{f''(x')}{2}(y-x)^2$ for some $x' \in [x, y]$

Example:



(see 14.50min in video)
Note that the tangent line is a lower bound for ALL $f(x)$.

Observation: $f(y) \geq f(x) + (y-x)f'(x)$
We see that strongly convex functions can fix a quadratic on the right hand side to give a lower bound of $f(y)$. The further away you move from, the further above the lower bound you are. *intuition: gradient descent is a simple greedy algorithm, using linear approximation*

2.2 Gradient Descent, n $\geq$ 2

Theorem 1. *(MultiVariate Taylor)*

$$f(y) = f(x) + \sum_{i=1}^{n} \frac{df(x)}{dx_i}(y_i - x_i) + 1/2 \sum_{i=1,j=1}^{n} \frac{d^2 f(x)}{dx_i dX_j}(y_i - x_i)(y_j - x_j)$$

gradient:

$$\nabla f = [\frac{df(x)}{dx_1}...]$$

$$\nabla^2 f = [\frac{df(x)}{dx_1^2}...]$$

[see 21min of lecture to see further notes]

There are many variants of the analysis of gradient descent, but we will study the strongest possible assumptions that give us the fastest (geometric) convergence.

1. β**-smooth** if $||\nabla f(y) - \nabla f(x)|| \leq \beta ||y - x||$
   equivalently, if f is twice differentiable:
   $||\nabla^2 f(x)||_{op} \leq \beta$
   Explaining the operator norm: $z^T \nabla^2 f(x) z \leq \beta ||z||^2$

2. α**-strongly convex** if:
   $(y-x)^T \nabla^2 f(x)(y-x) \geq \alpha ||y-x||^2$
   $f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\alpha}{2}||y-x||^2$

Intuition: Not only are you convex, but linear approximation grows quadratically. Gives lower bound as you are further away, intuition for strong convexity. When very far away, you can know how far off you are from original function.

Note: We want to show that the eigenvalues of the Hessian are sandwiched between β and α; in other words, that the Hessian has a good condition number.

Theorem 2. *(smooth, strongly convex)*

If $\eta \leq \frac{1}{2\beta}$ we have:

$$f(x_t) - f(x^*) \leq \beta(q - \frac{\eta\alpha}{2})^{t-1}||x_1 - x^*||^2$$

where x^* is the unique minimizer.

Key Lemma:

Lemma 3. *If f is β-smooth, α-strongly convex, then*
$\nabla f(x_t)^T(x_t - x^*) \geq \alpha/4||x_t - x^*||^2 + \frac{1}{2\beta}||\nabla f(x_t)||^2$

Proof. Proof of Lemma 3:

1. $\nabla f(x)^T(x - x^*) \geq \alpha/2||x - x^*||^2$
2. $\nabla f(x)^T(x - x^*) \geq 1/\beta||\nabla f(X)||^2$.

Let's prove 1:

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

Observe $f(x) \geq f(x^*) \Rightarrow$
$\nabla f(x)^T(x - x^*) \geq \alpha/2||x - x^*||^2$

Let's prove 2: We will use miltivariate lagrange remainder:
$\nabla f(x) = \nabla f(x^*) + \nabla^2 f(x')(x - x^*) \Rightarrow$
$\nabla f(x)^T(\nabla^2 f(x'))^{-1}\nabla f(x) = \nabla f(x)^T(x - x*)$

Now $1/2(1) + 1/2(2)$ □

2.3 Using Lemma 3 to prove Theorem 2

Proof. Let $\alpha' = \alpha/4$, $\beta' = \frac{1}{2\beta}$

$$\begin{aligned} ||X_{t+1} - x^*||^2 &= ||x_t - x^* - \eta\nabla f(x_t)||^2 \\ &= ||x_t - x^*||^2 - 2\eta\nabla(f(x_t)^T(x_t - x^*) - \eta^2||\nabla f(x_t)||^2) \\ &\leq (1 - 2\eta\alpha')||x_t - x^*||^2 + (\eta^2 - 2\eta\beta')||\nabla f(x_t)||^2 \\ &\leq (1 - 2\eta\alpha')||x_t - x^*||^2 \end{aligned}$$

Last step:

$$f(x^*) \geq f(x_t) + \nabla f(x_t)^T(x^* - x_t)$$

Rearrange

$$\nabla f(x_t)^T(x_t - x*) \geq f(x_t) - f(x^*)$$
$$(\nabla f(x_t) - \nabla f(x^*))^T(x_t 0 x^*) \leq \beta||x_t - x^*||^2$$

Putting it together:

$$\beta(||x_t - x^*||^2 \leq (1 - \eta\alpha/2)^{t-1}||x_q - x^*||^2)$$

□

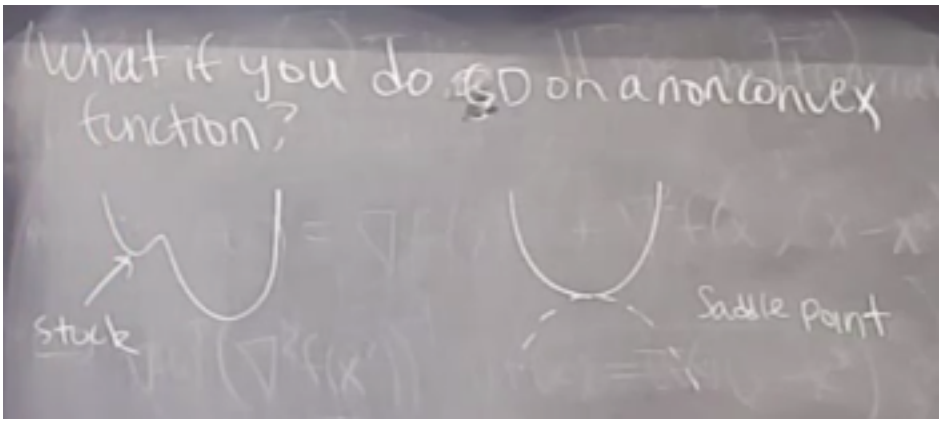
2.4 An Example:

$f(x) = ||Ax - b||_2^2$ least squares
$\nabla f(x) = 2A^T(Ax - b)$. $Ax - b$ is the remainder of the error.
Gradient Descent is:
$x_{t+1} = x_t - 2\eta A^T f_t$
$r_t + 1 = AX_t + 1 - b$

3 Stochastic GD

Useful in Machine Learning optimizations.
Take $x_{t+1} = x_t - \eta g_t$, where g_t is a random variable. We have $\mathbb{E}[g_t] = \nabla f(x_t)$.
This is natural when $f(x) = \sum_{y \in \text{examples in training set}} f_y(x)$. y are some parameters for training set x.
Then, $\nabla f(x) = \sum_y f_y(x)$, choose $g_t = \text{random} \nabla f_y(x)$.

4 What if you do GD on a non-convex function?



- Stuck at a local minimum picture.
  Oops, there's not much you can do.

- Stuck at a saddle point:
  Saddle points have Gradient to 0, can add second order derivative or stochastics to get out.