Lecture #15

Last Time: Linear Programming Relaxations

(approx. algorithms for vertex cover, set cover)

Today: Gradient Descent

Many variations, will focus on unconstrained minimization

Given a convex, differentiable function $f: \mathbb{R}^n \to \mathbb{R}$ i.e.

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

want to minimize $f(x)$.

Let's start with univariate case:

Gradient Descent ($n=1$)

For $t = 1$ to T

set $x_{t+1} = x_t - \eta f'(x_t)$ (learning rate; step size)

Recall Taylor's theorem says (if f is twice differentiable)

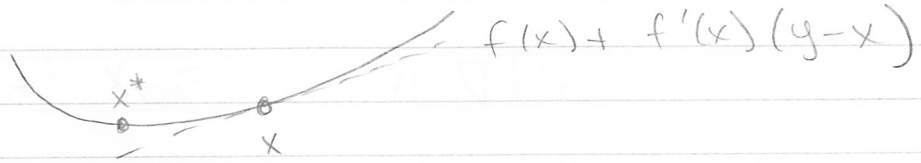
$$f(y) = f(x) + f'(x)(y-x) + \frac{f''(x)(y-x)^2}{2} + o((y-x)^2)$$

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

It is easy to show if f is twice differentiable

f is convex $\iff f''(x) \geq 0 \quad \forall x$

Taylor's Theorem gives us a linear approximation to f at x



It follows that $f(y) \geq f(x) + (y-x)f'(x)$

Gradient Descent ≡ Greedy, using Linear Approx

What should gradient descent do, in higher dimensions?

Multivariate Taylor's Theorem ($n > 1$)

$$f(y) = f(x) + \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}(y_i - x_i)$$

($f(y)$: length n vector)

$$+ \frac{1}{2} \sum_{i=1, j=1}^{n} \frac{\partial^2 f(x)}{\partial x_i \partial x_j}(y_i - x_i)(y_j - x_j) \ldots$$

more succinctly, $\nabla f = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2} \ldots\right]$ is the gradient,

$\nabla^2 f = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1^2} & \frac{\partial f(x)}{\partial x_1 \partial x_2} \\ & \ddots \end{bmatrix}$ is hessian

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

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

Gradient Descent ($n > 1$)

For $t = 1$ to T

Set $x_{t+1} = x_t - \eta \nabla f(x_t)$

There are many analyses of gradient descent, we'll just do one with assumptions

(1) B-smooth: $\|\nabla f(y) - \nabla f(x)\| \leq B \|y-x\|$
equivalently, if twice differentiable $\|\nabla^2 f(x)\|_{op} \leq B$

(2) α-strongly convex

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

equivalently: $f(y) \geq f(x) + \nabla f(x)^T (y-x)$

$+ \frac{\alpha}{2} \|y-x\|^2$

Theorem [smooth, strongly convex]

If $\eta \leq \frac{1}{B}$ we have

$f(x_t) - f(x^*) \leq B \left(1 - \frac{\eta\alpha}{2}\right)^{(t-1)} \|x_1 - x_*\|^2$

x^*: minimizer, must be unique under s.c.

We will prove:

Lemma: If f is β-smooth and α-strongly convex, then

$$\nabla f(x_t)^\top (x_t - x^*) \geq \frac{\alpha}{4} \|x_t - x^*\|^2 + \frac{1}{2\beta} \|\nabla f(x_t)\|^2$$

Proof of theorem: Let $\alpha' = \frac{\alpha}{4}$, $\beta' = \frac{1}{2\beta}$

$$\|x_{t+1} - x^*\|^2 = \|x_t - x^* - \eta \nabla f(x_t)\|^2$$

$$= \|x_t - x^*\|^2 - 2\eta \nabla f(x_t)^\top (x_t - x^*) + \eta^2 \|\nabla f(x_t)\|^2$$

$$\underbrace{\leq}_{\text{by lemma}} -2\eta \left(\alpha' \|x_t - x^*\|^2 + \beta' \|\nabla f(x_t)\|^2\right)$$

$$= (1 - 2\eta\alpha') \|x_t - x^*\|^2 + \underbrace{(\eta^2 - 2\eta\beta')}_{\leq 0} \|\nabla f(x_t)\|^2$$

To complete the proof note

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

Rearranging

$$\underbrace{\nabla f(x_t)^\top (x_t - x^*)}_{= (\nabla f(x_t) - \nabla f(x^*))^\top (x_t - x^*) \leq \beta \|x_t - x^*\|^2} \geq f(x_t) - f(x^*)$$

Thus putting it all together:

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

Now we prove the lemma; by establishing

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

(2) $\nabla f(x)^T (x - x^*) \geq \frac{1}{\beta} \|\nabla f(x)\|^2$

Let's prove (1). By strong convexity

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

But $f(x) \geq f(x^*)$, rearranging we get (1)

Let's prove (2). We need multivariate lagrange remainder

$$\nabla f(x) = \cancel{\nabla f(x^*)}^{\,0} + \nabla^2 f(x')(x - x^*)$$

for some x' between x and x^*. Then

$$\nabla f(x)^T \left(\nabla^2 f(x')\right)^{-1} \nabla f(x) = \nabla f(x)^T (x - x^*)$$

$$\geq \frac{1}{\beta} \|\nabla f(x)\|^2$$

Now $\frac{1}{2}(1) + \frac{1}{2}(2)$ finishes proof

An example

$$f(x) = \|Ax - b\|_2^2 \quad \text{least squares}$$

$$\nabla f(x) = 2A^T(\underbrace{Ax - b}_{r})$$

Gradient descent is $x_{t+1} = x_t - 2A^T r_t$

$r_{t+1} = Ax_{t+1} - b$

Can also do logistic regression

Further Discussion:

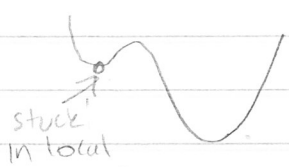
Stochastic gradient descent:

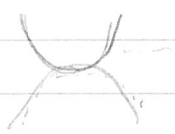
$$x_{t+1} = x_t - \eta \overset{\text{r.v.}}{g_t} \quad \text{where } \mathbb{E}[g_t] = \nabla f(x_t)$$

This is natural when $f(x) = \sum_{y \in \text{examples}} f_y(x)$ ← loss on ex. y

then $\nabla f(x) = \sum_y \nabla f_y(x)$, can choose $g_t = \text{random } \nabla f_y(x)$

What about gradient descent on nonconvex functions?





Saddle point, $\nabla f = 0$ but not even local min

Deep learning: $\min_\theta \sum_y \text{loss}(y;\theta)$ ← highly nonconvex, do gradient descent anyways!