Recall :

GD: $x^{(t+1)} = x^{(t)} - \mu^{(t)} \nabla f(x^{(t)})$

We developed necessary & sufficient conditions for optimality;

- Today, we'd like to build towards making the choice of $\mu^{(t)}$ rigorous.
- Along the way, we will develop an understanding of the rate of convergence of GD.

(we'd like to know how $f(x^{(t)}) - f(x^{*})$ behaves as a function of t.)

Defn: A function $f : \Omega \to \mathbb{R}$
$\hookrightarrow \subset \mathbb{R}^n$

is L-Lipschitz if $\forall x, y \in \Omega$

$$\underbrace{|f(x) - f(y)|}_{\text{"distance" between } f(x) \,\&\, f(y)} \leq L \underbrace{\|x - y\|}_{\text{distance between } x \,\&\, y}$$

So how much f changes between x & y is determined by how far x & y are from each other.

Example: $f : \mathbb{R} \to \mathbb{R}$

$f(x) = |x|$ is Lipschitz

because $|f(x) - f(y)| = \big|\,|x| - |y|\,\big|$

$$\leq |x - y|.$$

Moreover the Lipshitz constant is 1.

$f : \Omega \to \mathbb{R}$

Lemma: If f is L-Lipschitz, convex, and differentiable then $\|\nabla f(x)\| \leq L \;\; \forall x \in \Omega$

Proof: For any $x, y \in \Omega$

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

convexity

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

Lipschitz.

Pick $x = y + \nabla f(y)$

So:

$$L\|\nabla f(y)\| \geq \|\nabla f(y)\|^2$$

$$\Rightarrow \quad \|\nabla f(y)\| \leq L.$$

$\blacksquare$

$\boxed{\begin{array}{l}\text{Theorem: } \|\nabla f(x)\| \leq L \ \forall x \in \Omega \\ \qquad \text{then } |f(x)-f(y)| \leq L\|x-y\|\end{array}}$

Proof: By Taylor's theorem:

$$f(x) - f(y) = \nabla f(tx+(1-t)y)^T(x-y)$$

for some $t \in (0,1)$

So $$|f(x)-f(y)| = |\nabla f(tx+(1-t)y)^T(x-y)|$$
$$\leq \|\nabla f(\quad)\| \, \|x-y\|$$
$$\leq L \, \|x-y\|$$

bec $w^T z = \|w\| \, \|z\| \cos\theta$

& $|w^T z| \leq \|w\| \, \|z\|$

$\blacksquare$

We are now ready to present our result on the choice of μ & the convergence of GD.

Theorem: Let f be convex, differentiable, L-Lipschitz

& let $\begin{cases} \|x^{(0)} - x^*\| \leq R \\ \|\nabla f(x)\| \leq L \end{cases}$

Choose $\mu = \dfrac{R}{L\sqrt{t}}$

Conv. rate theorem

Then

$$f\left(\frac{1}{t}\sum_{s=0}^{t-1} x^{(s)}\right) - \underbrace{f(x^*)}_{\text{optimal value}} \leq \frac{RL}{\sqrt{t}}$$

gap between $\frac{1}{t}\sum_{s=0}^{t-1} x^{(s)}$ & opt. value

Example: Consider

$$f(x_1, x_2) = \sin(x_1) + x_2$$

and notice that $\nabla f(x_1, x_2) = \begin{pmatrix} \cos x_1 \\ 1 \end{pmatrix}$

$\Rightarrow \|\nabla f\| \leq \sqrt{\cos^2(x_1) + 1} \leq \sqrt{2}$

$\Rightarrow f$ is Lipschitz with $L = \sqrt{2}$

So if we run GD for t-steps

with $\mu^{(t)} = \dfrac{R}{\sqrt{2}\sqrt{t}}$ and we'll get

within $\dfrac{R\sqrt{2}}{\sqrt{t}}$ of a local min.

~~Proof of theorem:~~

First, note that

$$f(x^*) \geq f(x^{(s)}) + \nabla f(x^{(s)})^T (x^* - x^{(s)})$$

$$\Longleftrightarrow \quad \boxed{f(x^*) - f(x^{(s)}) \geq \nabla f(x^{(s)})^T (x^* - x^{(s)})}$$

(1)

Next, note that

$$x^{(s+1)} = x^{(s)} - \mu \nabla f(x^{(s)})$$

$$\Longrightarrow \quad \boxed{\nabla f(x^{(s)}) = \frac{x^{(s)} - x^{(s+1)}}{\mu}}$$

(2)

(1) & (2) $\Longrightarrow$

$$\boxed{f(x^{(s)}) - f(x^*) \leq \frac{1}{\mu} \langle x^{(s)} - x^{(s+1)}, x^{(s)} - x^* \rangle}$$

(3)

Fact: $a^T b$ or $(\langle a, b \rangle) = \dfrac{\|a\|^2 + \|b\|^2 - \|a-b\|^2}{2}$

Using this fact in 3 :

$$f(x^{(s)}) - f(x^*) \leq \frac{1}{2\mu}\Big(\overbrace{\|x^{(s)} - x^*\|^2}^{a}$$

$$+ \underbrace{\|x^{(s)} - x^{(s+1)}\|^2}_{b}$$

$$- \underbrace{\|x^{(s+1)} - x^*\|^2}_{a-b} \Big)$$

But $x^{(s)} - x^{(s+1)} = \mu \nabla f(x^{(s)})$

So

$$\boxed{f(x^{(s)}) - f(x^*) \leq \frac{1}{2\mu}\left(\|x^{(s)} - x^*\|^2 - \|x^{(s+1)} - x^*\|^2\right) + \frac{\mu}{2}\|\nabla f(x^{(s)})\|^2} \quad (4)$$

Summing from $s=0$ to $t-1$:

$$\sum_{s=0}^{t-1}\left(f(x^{(s)}) - f(x^*)\right)$$

$$\leq \frac{1}{2\mu}\left(\underbrace{\|x^{(0)}-x^*\|^2}_{\leq R^2} - \underbrace{\|x^{(t)}-x^*\|^2}_{\geq 0}\right)$$

$$+ \mu/2 \sum_{s=0}^{t-1} \underbrace{\|\nabla f(x^{(s)})\|^2}_{\leq L^2}$$

$$\leq \frac{1}{2\mu}\left(R^2 + \frac{\mu}{2}L^2 t\right)$$

dividing by t:

$$\frac{1}{t}\sum_{s=0}^{t-1} f(x^{(s)}) - f(x^*)$$

$$\leq \frac{1}{2\mu t}R^2 + \frac{\mu}{2}L^2$$

But convexity of f then gives

$$f\left(\frac{1}{t}\sum_{s=0}^{t-1} x^{(s)}\right) - f(x^*) \leq \frac{1}{2\mu t}R^2 + \frac{\mu}{2}L^2$$

by plugging in $\mu = \frac{R}{L\sqrt{t}}$ $\leftarrow$ $\leq \frac{RL}{\sqrt{t}}$ ∎

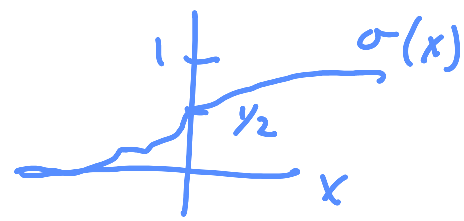
Note about HW2 (don't need to solve)

Problem $\{(x_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, 1\}$

Want to build classifier

Model probability of label w/ sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



$w^T x$ selects & weights features

$$\boxed{P(y_i \mid w^T x_i) = \sigma(y_i w^T x_i)}$$

Cross entropy loss to learn w looks to

minimize $-\mathbb{E}_{(x,y)} \log(p(x,y))$

$$L = -\frac{1}{n}\sum_{i=1}^{n} \log(\sigma(y_i w^T x_i)) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i w^T x_i}\right)$$