THEORY OF MACHINE LEARNING

LECTURE 10

CONVEX OPTIMIZATION, GRADIENT DESCENT

OPTIMIZATION

SOLVING ERM EFFICIENTLY

RECAP: ERM IS OFTEN HARD WITH FINITE CLASSES

- Finding best linear classifier (fewest mistakes) is NP hard even to approximate!
- Common remedy: loss functions
- Many candidate loss functions

$(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$

x_1 → features, y_1 → label

ERM: find h s.t. $\sum_{i=1}^{N} \mathbb{1}[h(x_i) \neq y_i]$ is min

Loss min: $\sum_{i=1}^{N} \underbrace{loss(h(x_i), y_i)}$

$\sum_{i=1}^{n} (h(x_i) - y_i)^2$

if h has some parametric form,

e.g. $h(x) = w^T x$ (in lin. class.)

Then $\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2$ is an easy problem.

Moving to loss min converts ERM into (hopefully more tractable) optimization probs.

RECAP: CONVEX OPTIMIZATION

$x \in \mathbb{R}^d$

$\ell(x) \in \mathbb{R}$.

f: convex

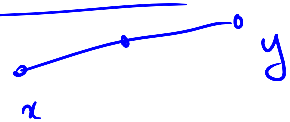
$D \subseteq \mathbb{R}^d$ is a convex set

$\arg\min_{x \in D} f(x)$

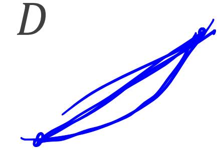
- **Problem.** Given a convex function defined over a convex domain, find the minimizer (or min value).

$f: \mathbb{R}^d \to \mathbb{R}$.

- $f(tx + (1-t)y) \le t\,f(x) + (1-t)f(y)$ for all $t \in (0,1)$ and $x, y \in D$

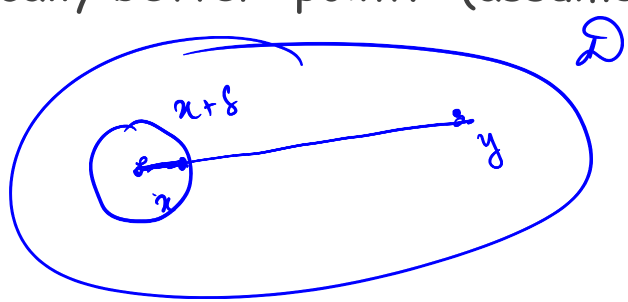


$f\left(\frac{x+y}{2}\right) \le \frac{f(x)+f(y)}{2}$.



- Local opt = global opt (just due to convexity) ✓

- Question: how to find a "locally better" point? (assume f is continuous, differentiable)



can you find some $x+\delta$ s.t. $f(x+\delta) < f(x)$?

TAYLOR APPROXIMATION

- Functions over R^d, gradients, Hessian
- First order approximation

(one dim): $f(x+\delta) = f(x) + \delta \cdot f'(x) + \frac{\delta^2}{2!} f''(x) + \ldots$

$\nabla^2 f$

$\downarrow$

Hessian

First order approx: $\boxed{f(x+\delta) \simeq f(x) + \delta f'(x)}$

$f(x+\delta) = f(x) + \delta \cdot f'(x')$

for some $x' \in (x, x+\delta)$ [Mean value theorem].

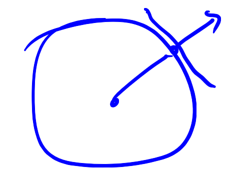
$f''(x) \leq C$

$f^{(r)}(x) \leq c^r \cdot r!$

Higher dimensions ($f : \mathbb{R}^d \to \mathbb{R}$)

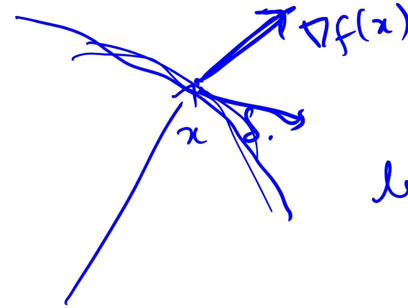
$\forall \; x, \delta \in \mathbb{R}^d$. ($\delta$ is "small")

$f(x) = \|x\|^2$



$$f(x+\delta) \approx f(x) + \boxed{\langle \delta, \nabla f(x) \rangle}.$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}$$



level set:
$\{ x' : f(x) = f(x') \}$

if we want to find δ s.t. $f(x+\delta) < f(x)$

idea: what if we set $\delta = -c \cdot \nabla f(x)$, for some $c > 0$?

$\langle y, y \rangle = \|y\|^2$.

$$\langle \delta, \nabla f(x) \rangle = -c \, \|\nabla f(x)\|^2$$

This ensures that either $\nabla f(x) = 0$, or $f(x+\delta) < f(x)$. (strictly)

(already at local opt $\longrightarrow$ global opt).

$x \rightsquigarrow x - c \cdot \nabla f(x)$

GRADIENT DESCENT ALGORITHM

- Generally applicable – even to non-convex functions
  (in which case you only find ~~local opt~~) some critical pt. (derivative $=0$)

$\rightarrow$ Start with some pt. $w^{(0)} \in \Theta$; pick η as some param.

$\rightarrow$ For T steps, do: $t=0,1,\ldots,T-1$

$$\boxed{w^{(t+1)} = w^{(t)} - \eta_t \nabla f(w^{(t)})}$$

η_t

$\eta_t \cdot \nabla f(w^{(t)})$

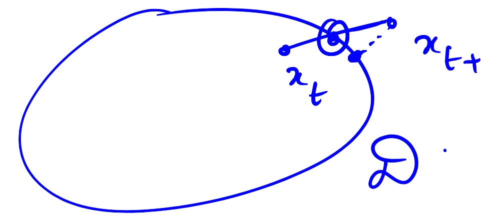
Choose T s.t. $\|\nabla f(w^{(T)})\|$ is "small enough".

NATURAL ISSUES

$\eta_t = \frac{1}{t}$.

$\eta_t = \eta$.

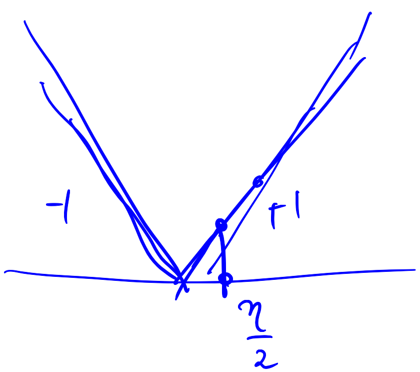
$\boxed{\eta_t = \frac{1}{\sqrt{t}}}$.

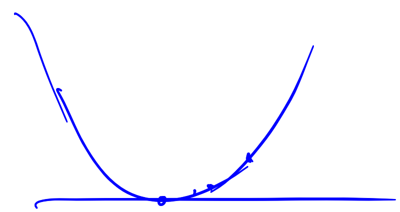


- Choosing how much to move! (aka *learning rate*)
- Staying in the domain

project $x^{(t+1)}$ to $\mathcal{D}$.

projection can be tricky! (if $\mathcal{D}$ is not "simple")





GRADIENT DESCENT – VANILLA ANALYSIS

$|f(x+\delta) - f(x)| \le L\|\delta\|$

$\approx |\langle \delta, \nabla f(x) \rangle|$

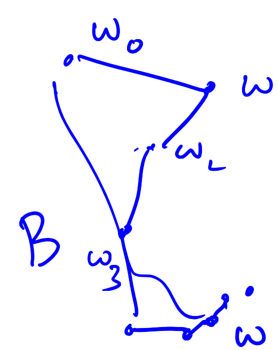
L-Lipschitz: $|f(x) - f(y)| \le L\|x-y\|_2$

convex,

$\Longrightarrow \|\nabla f(x)\| \le L.$

- Suppose f is L-Lipschitz, and domain $D = R^d$
- Suppose OPT was distance B away from initial point
- **Theorem.** Consider running T steps of gradient descent with a fixed learning rate η. Then we have

$$\frac{1}{T}\sum_{t=1}^{T} \left(f(w_t) - f(w) \right) \le \frac{B^2}{2\eta T} + \frac{L^2 \eta}{2}$$



w: true minimizer of f, i.e., $w = \arg\min_{x \in \mathbb{R}^d} f(x)$

$\frac{L^2 \cdot \eta}{2} \approx \eta$

Theorem tells us that if we perform "sufficiently many" steps of GD with fixed l. rate η, $\frac{1}{T}\sum_{t=1}^{T}\left(f(w_t)-f(w)\right) \le O(\eta)$.

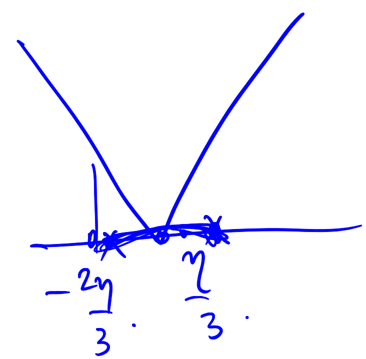
Can you find one pt $\bar{w}$ s.t. $f(\bar{w}) - f(w) \le O(\eta)$?

"Last iterate convergence"

$\underset{t}{\operatorname{argmin}}\ f(w_t)$

$\bar{w} = \frac{1}{T}\sum_t w_t$

$f(\bar{w}) \le \frac{1}{t}\sum_t f(w_t)$

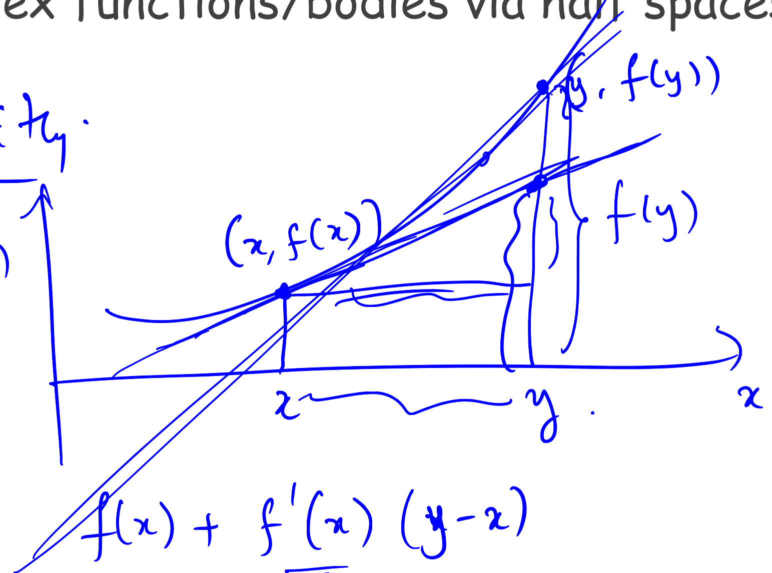


ALTERNATE DEFINITION OF CONVEXITY

- Function lies "above" the tangent plane *at any point* !
- Related to the definition of convex functions/bodies via half spaces

"Basic inequality" in convexity.

$$\underbrace{f(y) - f(x)}_{} \geq \langle \nabla f(x), \underbrace{y-x}_{} \rangle$$



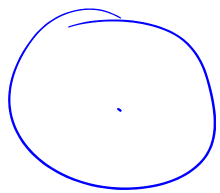
$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2}.$$

(Homework 2).

$f(x) + f'(x)(y-x)$

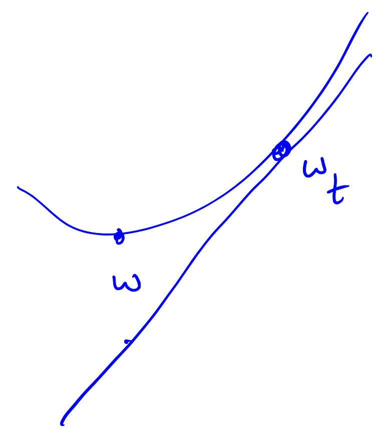
$\langle \nabla f(x), y-x \rangle$

GRADIENT DESCENT – VANILLA ANALYSIS



- Suppose f is L-Lipschitz, and domain $D = R^d$
- Suppose OPT was distance B away from initial point
- **Theorem.** Consider running T steps of gradient descent with a fixed learning rate η. Then we have

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w) \leq \frac{B^2}{2\eta T} + \frac{L\cancel{\rho}^2\eta}{2}$$



$f(w) - f(w_t) \geq \ldots$

$f(w_t) - f(w) \leq \ldots$

Basic ineq of convexity:

$$f(w) - f(w_t) \geq \langle \nabla f(w_t), w - w_t \rangle$$

$$\Leftrightarrow \boxed{f(w_t) - f(w)} \leq \boxed{\langle \nabla f(w_t), w_t - w \rangle}$$

How to connect to grad. descent?

$$w_{t+1} = w_t - \eta \cdot \nabla f(w_t)$$

(Algebraic magic)

$$\Phi_t = \| w_t - w \|^2$$

$$\| z - \eta \nabla f(w_t) \|^2 - \|z\|^2$$

$$\|z\|^2 + \eta^2 \|\nabla f(w_t)\|^2 - 2 \langle z, \eta \nabla f .. \rangle$$

$$\boxed{\Phi_{t+1} - \Phi_t} = \| w_t - \eta \nabla f(w_t) - w \|^2 - \| \overbrace{w_t - w}^{z} \|^2$$

$$= -2\eta \boxed{\langle \nabla f(w_t), w_t - w \rangle} + \eta^2 \| \nabla f(w_t) \|^2$$

$$\Phi_{t+1} - \cancel{\Phi_t} + \cancel{\Phi_t} - \cancel{\Phi_{t-1}} + \ldots + \cancel{\Phi_1} - \Phi_0$$

ANALYSIS VIA POTENTIAL FUNCTIONS

→ drop in potential is "related" to ~~cha~~ change in function value...

DEALING WITH THE DOMAIN – PROJECTED GD