THEORY OF MACHINE LEARNING

LECTURE 11

CONVEX OPTIMIZATION, GRADIENT DESCENT

[Wed AM — office hours — 9:30- 11AM.].

(same zoom link as class.).

RECAP: LOSS MINIMIZATION

feature vector $\in \mathbb{R}^d$

$(\vec{x}_1, y_1), (x_2, y_2) \ldots (x_N, y_N)$

label (± 1)

$$\min_{h} \sum_{i=1}^{N} \ell(h(x_i), y_i)$$

- ERM is hard, so we consider minimization of loss
- General problem

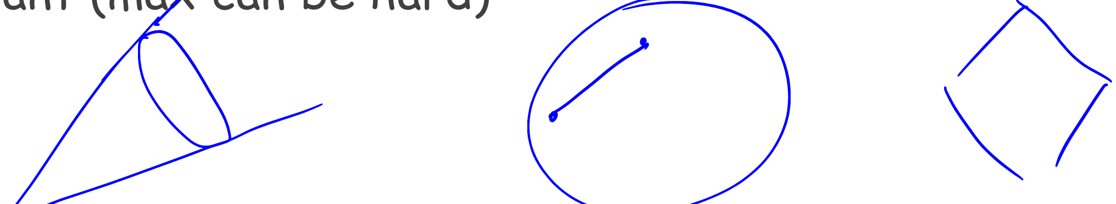
loss function ℓ

parametrize hypotheses using $\underline{\theta}$.

- Optimization can be hard in general, we study "easy" case of convex optimization

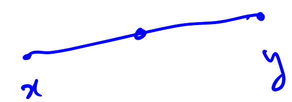
(hard unless ℓ is "nice" $\longrightarrow$ ℓ is convex in θ, where θ are parameters that define h.)

- Min f(x) over D, where f is convex, domain D is convex
- Minimization is important (max can be hard)



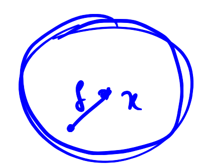
RECAP: CONVEX OPTIMIZATION

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



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

- Local opt = global opt (just due to convexity)
- <u>Question:</u> how to find a "locally better" point? (assume f is continuous, differentiable)
- Gradient descent – inspired by Taylor approximation



$f(x+\delta) \approx f(x) + \delta f'(x)$ (in 1D)

in general, $f(x+\delta) \approx f(x) + \langle \delta, \nabla f(x) \rangle$.

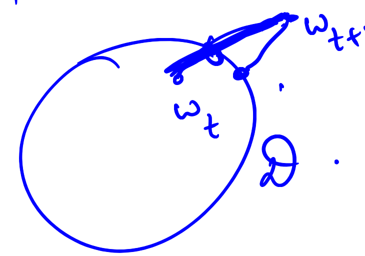
↳ vector

GRADIENT DESCENT ALGORITHM

- Start with w_0 = some feasible point; η is some fixed param.
- For $t = 0, 1, \ldots, T-1$: (T: # iterations)

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

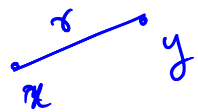
→ if w_{t+1} is outside $\mathcal{D}$, set w_{t+1} to be the projection.



- Generally applicable – even to non-convex functions (in which case you only find local opt)
- Choosing how much to move! (aka *learning rate*)
  - how should we set η?
  - should it depend on t?
- Staying in the domain

VANILLA ANALYSIS

(As simple as possible)



$|f(x) - f(y)| \leq L \|x - y\|$.

$\Updownarrow$

$\|\nabla f(x)\| \leq L$.

- Suppose f is L-Lipschitz, and domain $D = R^d$

  (no projection needed)

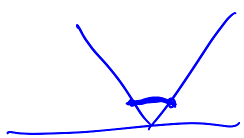
- Suppose OPT was distance B away from initial point

  $\rightarrow \operatorname{argmin}_x f(x)$

- **Theorem.** Consider running T steps of gradient descent with a fixed learning rate η. Then we have

$\|w_0 - w^*\| \leq B$

w_T



$$\frac{1}{T}\sum_{t=1}^{T}\left(f(w_t) - f(w^*)\right) \leq \frac{B^2}{2\eta T} + \frac{L^2\eta}{2} \rightarrow \eta \cdot \sqrt{\frac{B^2 L^2}{2T}}.$$

$w_t' = \frac{1}{T}(w_1 + \dots + w_T)$

(if η is very small, say $\frac{\varepsilon}{L^2}$ and T "large"

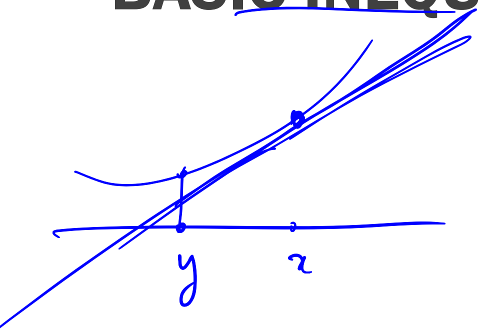
(ε is some error of interest).

Proof uses "basic inequality" of convexity

$f(w_T') \leq \frac{1}{T}\sum_t f(w_t)$ (by convexity)

$f(w_t') - f(w^*) \leq \dots$

BASIC INEQUALITY, POTENTIAL FUNCTION ANALYSIS



$$\forall y, \quad f(y) \geq f(x) + \langle y - x, \nabla f(x) \rangle .$$

[$\Leftrightarrow$ old definition of convexity.) .

$$f(w^*) \geq f(w_t) + \langle w^* - w_t, \nabla f(w_t) \rangle . \longrightarrow \bigstar$$

(intuition) Tells us that keeping track of $\| w^* - w_t \|$ may be useful.

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

(hope: this reduces with time..)

$\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b\rangle$.

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

$$= -2\eta \underbrace{\langle w^* - w_t, \nabla f(w_t) \rangle}_{(\text{use } \bigstar)} - \eta^2 \| \nabla f(w_t) \|^2$$

$\| w^* - w_{t+1} \|^2 \leq \| w^* - w_{t+\frac{1}{2}} \|^2$

f-divergences

($\bigstar$ HW)

$$\geq \underbrace{-2\eta\left[f(w^*) - f(w_t)\right]}_{= 2\eta\left[f(w_t) - f(w^*)\right]} - \eta^2 \|\nabla f(w_t)\|^2$$

("local" analysis)

$$\therefore \cancel{2}\eta\left[f(w_t) - f(w^*)\right] \le \frac{\Phi_t - \Phi_{t+1}}{2\eta} + \frac{\eta}{2}\|\nabla f(w_t)\|^2$$

(single t.)

$$\sum_{t=0}^{T-1}\left(f(w_t) - f(w^*)\right) \le \frac{\Phi_0 - \Phi_T}{2\eta} + \frac{\eta \cdot L^2 \cdot T}{2}$$

$$\sum_t A_t - A_{t+1}$$

$$A_1 - \cancel{A_2} + \cancel{A_2} - \cancel{A_3} + \cancel{A_3} - A_4 + \dots$$

$$= A_0 - A_T$$

$$\le \frac{B^2}{2\eta} + \frac{\eta \cdot L^2 T}{2}$$

What is the best choice of η for a given T? Must set η so that $\frac{B^2}{2\eta} = \frac{\eta \cdot L^2 T}{2}$, i.e., $\eta = \sqrt{\frac{B^2}{L^2 T}} := \frac{B}{L\sqrt{T}}$

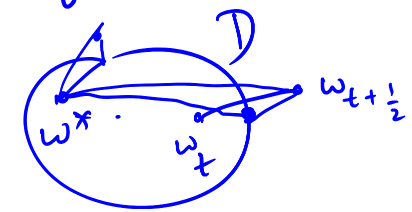
(in theorem)

RHS becomes $\frac{BL}{\sqrt{T}}$.

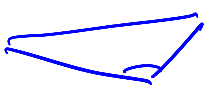
DEALING WITH THE DOMAIN – PROJECTED GD

→ Entire analysis goes through with projections!

→ Formally when D is bounded, GD is the foll:



Say we have $w_t \in D$



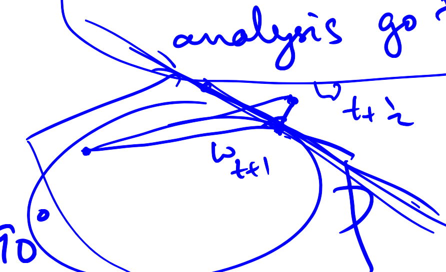
$$w_{t+\frac{1}{2}} = w_t - \eta \cdot \nabla f(w_t)$$

$$w_{t+1} = \Pi_D (w_{t+\frac{1}{2}}) = \underset{x \in D}{\operatorname{argmin}} \| w_{t+\frac{1}{2}} - x \|$$

Obsn: $\| w^* - w_{t+1} \|^2 \le \| w^* - w_{t+\frac{1}{2}} \|^2$ → lets the original analysis go through!

(angle with separating plane is $90°$)

$\therefore$ angle with any pt on the other side of the plane is $\geq 90°$



EXTENSIONS

- What if function is “smooth”? Get improved ‘rate’
- What if function is “strongly convex”?
- What if functions at different steps are different? (!)

What if at time t, $f_t(x) \rightarrow$ convex, L-Lipschitz, varies with time.

$\forall\ w^* \in D$, $\quad f_t(w^*) \geq f_t(w_t) + \langle w^* - w_t, \nabla f_t(w_t) \rangle$

$$\frac{1}{T} \sum_t \left(f_t(w_t) - f_t(w^*) \right) \leq \frac{B^2}{2\eta T} + \frac{L^2 \eta}{2}.$$

[Online Convex opt.]