Lecture 19

* HW2 due today on Gradescope (mark pages when submitting)

RECAP

follow the regularized leader (FTRL)

- FTRL algorithm: At every time t, choose

$$w_t \in \operatorname*{argmin}_{w \in S} \left(\psi(w) + \sum_{i=1}^{t-1} f_i(w) \right)$$

(FTL is FTRL with $\psi = 0$).

linear f_t, quadratic ψ

Thm For any $\eta > 0$, FTRL with $S \subseteq \mathbb{R}^d$ a convex set & $\psi(w) = \frac{\|w\|_2^2}{2\eta}$, $f_t(w) = \langle w, v_t \rangle$ satisfies,

$$\text{Regret}(u, T) \leq \frac{1}{2\eta} \|u\|^2 + \eta \sum_{t=1}^{T} \|v_t\|_2^2 .$$

If $\|u\|_2 \leq B$ & $\|v_t\| \leq L$, then choosing $\eta = \frac{B}{L\sqrt{T}}$ gives $\text{Regret}(T) = O(BL\sqrt{T})$

Beyond linear : online convex optimization.

Consider convex functions to avoid intractability.

for convex functions, a linear approximation suffices.

Algorithm : online gradient descent (OGD).

① Let $w_1 = 0$, $\theta_1 = 0$

② for $t = 1, \ldots, T$

③ Predict w_t, receive f_t

④ Find gradient $v_t = \nabla f_t(w_t)$

⑤ If $S = \mathbb{R}^d$

⑥ $w_{t+1} = w_t - \eta v_t$

⑦ Else

⑧ $w_{t+1} = \Pi_S(\eta \theta_{t+1})$, $\theta_{t+1} = \theta_t - v_t$

Note : If f is not differentiable, use subgradients.

Thm OGD enjoys the following regret bound for every $w^* \in S$,

$$\text{Regret}(w^*, T) \leq \frac{\|w^*\|^2}{2\eta} + \eta \sum_{t=1}^{T} \|v_t\|_2^2$$

If $\|v_t\| \leq \rho \ \forall t$ (f_t is ρ-Lipschitz $\forall t$) & $\|w^*\|_2 \leq B$, then setting $\eta = \frac{B}{\rho\sqrt{T}}$ yields,

$$\text{Regret}(T) \leq B\rho\sqrt{T}.$$

TODAY:

Proof: from our earlier analysis of FTRL, we have a regret bound for linearized losses:

$$\sum_{t=1}^{T} \left[\langle w_t, r_t \rangle - \langle w^*, r_t \rangle \right]$$

$$\leq \frac{\|w^*\|^2}{2\eta} + \eta \sum_{t=1}^{T} \|r_t\|_2^2$$

Actual regret: $\sum_{t=1}^{T} \left[f_t(w_t) - f_t(w^*) \right]$

Using convexity, since r_t is gradient at w_t:

$$f_t(w^*) \geq f_t(w_t) + \langle r_t, w^* - w_t \rangle$$

$$\therefore \quad f_t(w_t) - f_t(w^*) \leq \langle w_t, r_t \rangle - \langle w^*, r_t \rangle.$$

and the result follows. ∎

Example 1: Learning with expert advice

S = Simplex in $\mathbb{R}^d$ ($\Delta_d = \{ w : w \in \mathbb{R}^d, \; w_i \geq 0 \; \forall i \in [d], \; \sum w_i = 1 \}$)

$f_t(w) = \langle w, r_t \rangle$

$r_t = \left(l(h_1(x_t), y_t), \; l(h_2(x_t), y_t), \; \ldots, \; l(h_d(x_t), y_t) \right)$

where $\ell(h_i(x_t), y_t) \in [0,1] \quad \forall i \in [d]$.

Corollary: FTRL with quadratic regularizer (or OGD) for the experts setting gets $\text{Regret}(T) \leq \sqrt{dT}$.

Proof

Bound on set of experts (B)

experts are in Δ_d

$\|u\|_1 = 1 \quad \forall \quad u \in \Delta_d$

Since $\|u\|_2 \leq \|u\|_1 \Rightarrow \|u\|_2 \leq 1 \quad \forall u \in \Delta_d$

$\therefore B \leq 1$

Gradient: $\nu \in$

since $(\nu_t)_i \in [0,1] \quad \therefore \|\nu_t\|_2 \leq \sqrt{d}$

$\therefore L \leq \sqrt{d}$

$\therefore \text{Regret}(T) \leq \sqrt{dT}$.

Note: This depends on $\sqrt{d}$ instead of $\sqrt{\log(d)}$.

Using the entropic regularizer

$$\psi(w) = \sum_{i=1}^{d} w_i \log(w_i)$$

(negative entropy of w)

gives the $O\left(\sqrt{T \log(d)}\right)$ regret bound (and recovers the Weighted-Majority algorithm).

Example 2: Online Perceptron.

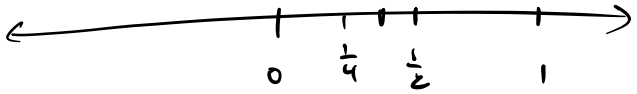
$X = \mathbb{R}^d$

$Y = \{-1, 1\}$

At every time t, learner receives $x_t \in \mathbb{R}^d$.

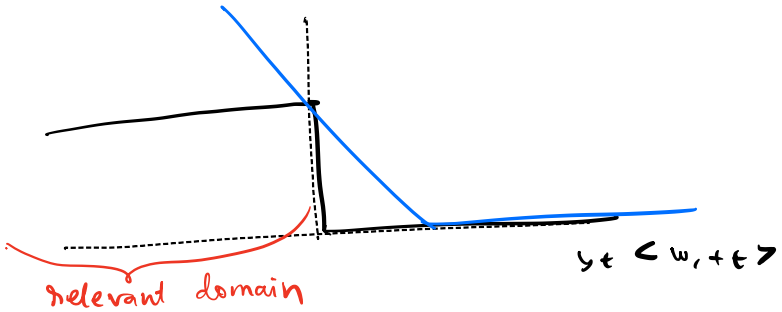
Maintain $w_t \in \mathbb{R}^d$ and predict $p_t = \text{sign}(\langle w_t, x_t \rangle)$.

We showed earlier that thresholds have $Ldim = \infty$, therefore no hope of getting small mistake bound without further assumptions.



Convex surrogate loss to avoid this

$$\ell_{0,1}(w, (x,y)) = \mathbb{1}(y \langle w, x \rangle \leq 0)$$



* Whenever the algorithm makes a mistake, we use the hinge loss:

$$f_t(w) = \max\{0, 1 - y_t \langle w, x_t \rangle\} = [1 - y_t \langle w, x_t \rangle]_+$$

* when the algorithm is correct, define

$$f_t(w) = 0 .$$

Note :

- $f_t(w)$ is convex
- for all $l_{0,1}(w, (x_t, y_t)) \leq f_t(w)$

Use OGD to learn this,

$$\nabla f_t(w_t) = \begin{cases} 0 , & \text{if } y_t \langle w_t, x_t \rangle > 0 \\ & (\text{since } f_t(w) = 0) \\ -y_t x_t , & \text{if } y_t \langle w_t, x_t \rangle \leq 0 \end{cases}$$

since $f_t(w) = [1 - y_t \langle w, x_t \rangle]_+$

$= 1 - y_t \langle w, x_t \rangle$

$\therefore$ OGD updates :

- $w_1 = 0$
- $w_{t+1} = \begin{cases} w_t & \text{if } y_t \langle w_t, x_t \rangle > 0 \\ w_t + \eta y_t x_t , & \text{otw} \end{cases}$

Alg : Perceptron

- $w_1 = 0$
- for $t = 1, \dots, T$

  receive x_t

  predict $p_t = \text{sign} \langle w_t, x_t \rangle$

  if $y_t \langle w_t, x_t \rangle \leq 0$

    $w_{t+1} = w_t + y_t x_t$ (we can drop η since we just use the sign)

  else

    $w_{t+1} = w_t$

Thm Suppose we run Perceptron on a sequence $(x_1, y_1), \dots, (x_T, y_T)$ & let $R = \max_t \|x_t\|_2$. Let $\mathcal{M}$ be the rounds where Perceptron makes a mistake and let $f_t(w) = \mathbb{1}(t \in \mathcal{M}) [1 - y_t \langle w, x_t \rangle]_+$.

Then for every w^*,

$$|\mathcal{M}| \leq \sum_t f_t(w^*) + R \|w^*\| \sqrt{\sum_t f_t(w^*)} + R^2 \|w^*\|^2 \quad - (1)$$

If $\exists\, w^*$ s.t. $\|w^*\| = 1$ & $y_t \langle w^*, x_t \rangle \geq \gamma \ \forall t$, then

$$|\mathcal{M}| \leq R^2 / \gamma^2 \quad - (2)$$

Proof By OGD guarantee,

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^*) \le \frac{1}{2\eta}\|w^*\|_2^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|r_t\|_2^2$$

where r_t is the gradient

$$\|r_t\| = \begin{cases} 0 & \text{if } t \notin M \\ \|\ell_t\| & \text{if } t \in M \end{cases}$$

$$\therefore \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^*) \le \frac{\|w^*\|_2^2}{2\eta} + \frac{\eta}{2}|M|R^2$$

Since $l_{0,1}((x_t, y_t), w) \le f_t(w)$

$$\sum_{t=1}^{T} f_t(w_t) \ge |M|$$

$$|M| - \sum_{t=1}^{T} f_t(w^*) \le \frac{1}{2\eta}\|w^*\|_2^2 + \frac{\eta}{2}|M|R^2$$

This is true $\forall\, \eta > 0$

Setting $\eta = \dfrac{\|w^*\|}{R\sqrt{|M|}}$

$$|M| \le \sum_{t=1}^{T} f_t(w^*) + R\|w^*\|\sqrt{|M|}$$

By solving quadratic, we get ①.

$$|M| \leq \sum_{t=1}^{T} f_t(w^*) + \frac{1}{2\eta} + \frac{\eta}{2} |M| R^2$$

Claim: $f_t(w^*) \leq |M|(1-\gamma)$

Proof for all $t \notin M$ $f_t(w^*) = 0$

for all $t \in M$, $f_t(w^*) \leq 1-\gamma$.

$$|M| \leq |M|(1-\gamma) + \frac{1}{2\eta} + \frac{\eta}{2} |M| R^2$$

Setting $\eta = \frac{1}{R\sqrt{|M|}}$

$$\gamma |M| \leq R\sqrt{|M|}$$

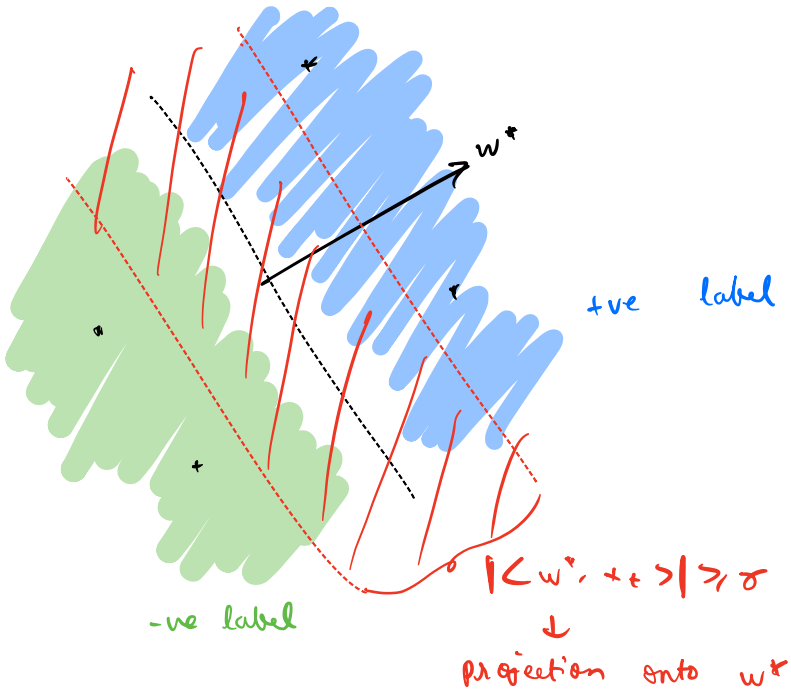
$$\Rightarrow |M| \leq R^2/\gamma^2.$$

The assumption that $y_t \langle w, x_t \rangle \geq \gamma$ is called separability with a margin.

$\|w^*\| = 1$

we require $y_t \langle w^*, x_t \rangle \geq \gamma \quad \forall t$

say for $y = 1$, $\langle w^*, x_t \rangle \geq \gamma$.



Computational - Statistical tradeoffs

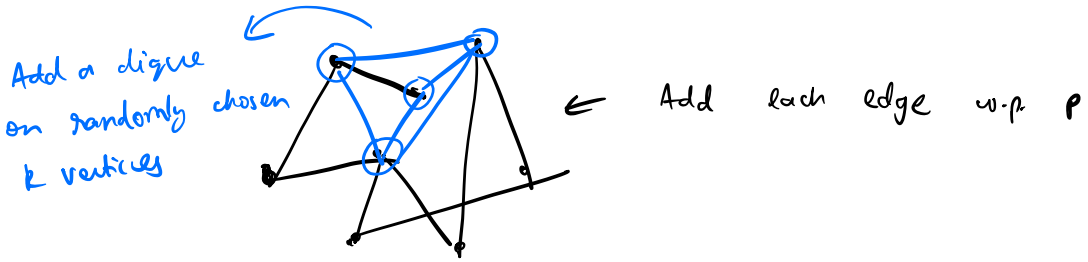
Statistical power often comes at the cost of computational efficiency.

Planted clique problem

Erdös - Renyi random graph $G(n,p)$

n : # vertices

There is an edge b/w vertices $i < j$ w.p p



for our lecture, take $p = \frac{1}{2}$.

Erdös - Renyi graph with "planted clique":

Choose a random subset of k vertices, add a clique on these vertices.

Def (Planted clique problem).

Given a graph generated from one of the following 2 distributions, decide which distribution generated the graph

① $G(n, \frac{1}{2})$

② Generate an instance of $G(n, \frac{1}{2})$ & plant a clique on k randomly chosen vertices of the graph.

What is the smallest k at which these two distributions are information-theoretically distinguishable?

Lemma An Erdös-Renyi random graph $G(n, \frac{1}{2})$ does not have a clique of $k \geq 3\log n$, w.h.p.

Proof Consider any subset of $k = 3\log n$ vertices.

$$\text{Prob}(\exists \text{ a clique on these } k \text{ vertices}) = \left(\frac{1}{2}\right)^{\frac{k(k-1)}{2}}$$

By union bound,

$$\text{Prob}(\exists \text{ a clique on } \underline{\text{any}} \text{ subset of } k \text{ vertices}) \leq \binom{n}{k} \cdot \left(\frac{1}{2}\right)^{\frac{k(k-1)}{2}}$$

Since $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$

$$\therefore \text{Prob(clique of size } k) \leq \left(\frac{ne}{k}\right)^k \cdot \left(\left(\frac{1}{2}\right)^k\right)^{\frac{k-1}{2}}$$

$$= \left(\frac{ne}{3\log n}\right)^{3\log n} \cdot \left(\frac{1}{n^3}\right)^{\frac{3\log n - 1}{2}}$$

$$= \left(\frac{ne}{3\log n}\right)^{3\log n} \left(\frac{1}{n^{1.5}}\right)^{3\log n - 1}$$

$$\leq \frac{1}{p(n)}$$

for any polynomial $p(\cdot)$. ∎

for $k \geq 3\log n$:

Algorithm (Brute-force search)

→ Search over every subset of k vertices

→ If ∃ a clique on any subset

return (graph comes from a planted clique model)

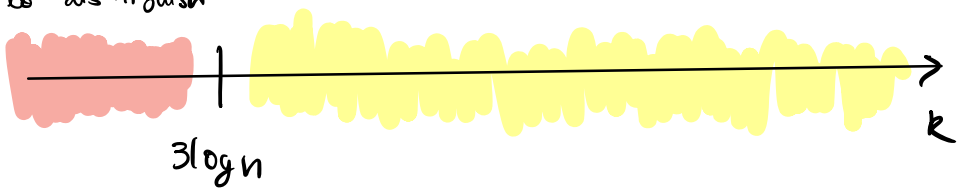
→ Else

return (graph comes from $G(n, \frac{1}{2})$

Running time : $k^2\binom{n}{k} = \Omega\left(n^{\log n}\right)$.

for $k \ll \log n$ information-theoretically impossible to distinguish

information theoretically possible to detect if $k \geq 3\log n$



When can we do this efficiently?

Simple alg. when $k \gg \sqrt{n \log n}$

Lemma The # edges in an Erdős-Renyi graph $G(n, \frac{1}{2})$ lies in $\left[\frac{n \cdot n-1}{4} - 100n\sqrt{\log n}, \frac{n \cdot n-1}{4} + 100n\sqrt{\log n} \right]$, whp.

Proof

X_{ij} = indicator r.v. denoting whether or not $\exists$ edge b/w i & j

$X_{ij} \sim \text{Ber}(\frac{1}{2})$

Let $m = \binom{n}{2}$.

Let $Z = \sum_{i<j} X_{ij}$

Then $\mathbb{E}[Z] = \frac{m}{2}$

By Chernoff bound / Hoeffding bound,

$$\Pr\left[|Z - \mathbb{E}[Z]| \geq \sqrt{m} \cdot t \right] \leq 2\exp(-2t^2).$$

$\therefore \quad t = \sqrt{10 \log n}$

$$\Pr\left[\left| z - \frac{m}{2} \right| \geq \sqrt{10 m \log n} \right] \leq \frac{2}{n^{20}}.$$

$\therefore$ whp, $z \in \left[\frac{n \cdot n-1}{4} - 100 n \sqrt{\log n}, \right.$

$\left. \frac{n \cdot n-1}{4} + 100 n \sqrt{\log n} \right]$.