Lecture 13 Online learning

Recap: Convex optimization

	Convex	+ Smooth	+ Strong CVX	+ Hessian Lipschitz (M)	+ Constraints
GD	$\frac{1}{\varepsilon^2}$	$\frac{1}{\varepsilon} \to \sqrt{\frac{1}{\varepsilon}}$	$\frac{L}{m}\log\frac{1}{\varepsilon} \to \sqrt{\frac{L}{m}}\log\frac{1}{\varepsilon}$		
Prox. Gradient	$\frac{1}{\varepsilon^2}$	$\frac{1}{\varepsilon}$	same as GD	not used	~~not used~~ n.a.
(projected) SGD	$\frac{1}{\varepsilon^2}$	$\frac{1}{\varepsilon^2}$	$\frac{1}{m\varepsilon}$ even if ~~smoothness~~		
Newton	n.a.	n.a.	n.a.	$K_0(m,L,M,\alpha,\beta) + C(m,L,M,\alpha,\beta)\log\log\left(\frac{f}{\varepsilon}\right)$ $\Downarrow$ $K_0(\alpha,\beta) + C\cdot\log\log\left(\frac{f}{\varepsilon}\right)$	Affine
IPM (Barrier Method) PD·IPM	Reformulate	Reformulate	Reformulate	$O\left(\log\left(\frac{c}{\varepsilon}\right)\right)$ # of Newton steps $O\left(\log\frac{\#\text{of constraints}}{\varepsilon}\right)$	same as left.

Online Learning

Setting of Statistical Learning (Batch learning, PAC learning) (Valiant, Vapnik)

$(x_1, y_1) \cdots (x_N, y_N) \overset{iid}{\sim} D$ $\quad x \in \mathcal{X} = \mathbb{R}^d$ or $\{0,1\}^d$

$y \in \mathcal{Y} = \{0,1\}$

Hypothesis Class $\mathcal{H}$. Realizable: $\exists h^* \in \mathcal{H}$, $\mathbb{P}_{(x,y) \sim D}(h^*(x) = y) = 1$

Goal of learning: find $h \in \mathcal{H}$, s.t. $err_D(h) = \mathbb{E}_{x \sim D}\left[\mathbb{1}(h^*(x) = h(x))\right] \xrightarrow[N \to \infty]{} 0$

Online Learning: You observe one data pair at a time

$h_1 \leftarrow x_1$, predict $\hat{y}_1 = h_1(x_1)$, y_1 is revealed

$h_2 \leftarrow (x_1, y_1)$; x_2

$\vdots$

$h_t \leftarrow \underbrace{(x_1, y_1) \cdots (x_{t-1}, y_{t-1})}$, x_t predict $h_t(x_t)$, reveal y_t

Data can be adversarially chosen
Not i.i.d.

Mistake bound M

If Alg. A satisfy that A makes $\le M$ mistakes for all sequences of $(x_1, h^*(x_1)) \cdots (x_t, h^*(x_t)) \cdots (x_T, y_T)$ (T)

then Alg A has a mistake bound of M.

Alg A: "Consistency" ~~Follow-the-Leader~~ or ERM

1. $V_1 = \mathcal{H}$
2. for $t = 1, 2, 3, \ldots$

   Receive x_t, pick any $h \in V_t$

   predict $\hat{y}_t = h(x_t)$

   Receive $y_t = h^*(x_t)$

   Update $V_{t+1} = \{ h \in V_t \mid h(x_t) = y_t \}$

   any $h \in V_{t+1}$, $h^*(x_i) = h(x_i)$ $\forall i = 1, 2, \ldots, t$

Each mistake, we can eliminate at least 1 hypothesis

Trivial Upper bound

$$M[\text{consistency}] \leq |\mathcal{H}| - 1$$

Example: $\mathcal{X} = \{1, 2, \ldots, |\mathcal{H}|\}$

$\mathcal{H} = \{h_1, h_2, \ldots, h_{|\mathcal{H}|}\}$

$h_i(x) = \begin{cases} 0 & \text{if } x < i \\ 1 & \text{otherwise} \end{cases}$

$x_1 = 1$, $x_2 = 2$, ..., $x_{|\mathcal{H}|} = |\mathcal{H}|$

$y_1 = 0$ $y_2 = 0$ $y_i = 0$ $y_{|\mathcal{H}|} = 1$

pick $h = h_1$ h_2 h_3

mistake $m+1$ $m+1$ ..., $h^* = h_{|\mathcal{H}|}$

Alg B. "Halving"

1. $V_1 = \mathcal{H}$
2. for $t = 1, 2, 3, \ldots$

   Receive x_t

   different > predict $\hat{y}_t = \text{Vote}_{h \in V_t}(h(x_t)) = \arg\max_{r \in \{0,1\}} |\{ h \in V_t : h(x_t) = r \}|$

   Receive $y_t = h^*(x_t)$

   Update $V_{t+1} = \{ h \in V_t \mid h(x_t) = y_t \}$

Thm. (Halving) $M \leq \log_2(|\mathcal{H}|)$

Proof: For each mistake, at least $\frac{|V_t|}{2}$ hypothesis are removed.

$$1 \leq |V_{t+1}| \leq |\mathcal{H}| 2^{-M}$$

$$2^M \leq |\mathcal{H}| \qquad M \leq \log_2(|\mathcal{H}|) \quad \square$$

"~~Realizability~~" . agnostic learning

Compete v.s. the best $h \in \mathcal{H}$ in the hindsight

Regret $= \sum_{t=1}^{T} \mathbb{1}(h_t(x_t) \neq y_t) - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}(h(x_t) \neq y_t)$

(x_t, y_t) chosen by an adversary. as $T \to \infty$

ff Regret $= o(T)$

Example: Stock Forecasting

	Exp 1. (Alex)	Exp 2 (Esther)	Exp 3 (Boyuan)	Exp 4 (Raffles the Cat)	Outcome
Day 1:	Down	UP	UP	Down	Down
Day 2:	UP	UP	Down	Down	Down
Day 3:	UP	Down	UP	UP	UP

Weighted Majority

	1	1	1	1
Day 1	1	$\frac{1}{2}$	$\frac{1}{2}$	1
Day 2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	1
Day 3	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{2}$	1

Alg C (Weighted Majority)

predict $\hat{y}_t = \mathbb{1}\left(\underbrace{\sum_{h\in H} w_h^t \cdot h(x_t)}_{\text{total weight on } \{UP/1\}} > \frac{1}{2}\underbrace{\sum_{h\in H} w_h^t}_{W}\right)$

receive y_t, discount $w_h^{t+1} \leftarrow w_h^t \cdot \frac{1}{2}$ for all h that made a mistake.

Thm. (WM) $M \le 2.4(m + \log n)$

M: total # of mistakes WM makes

m: # of mistakes of the best expert in the hindsight

n: # of hypotheses / experts

Proof: claim: Each mistake $\Rightarrow$ W drops at least by 25% (why?)

$$\left(\frac{1}{2}\right)^m \le W \le n\left(\frac{3}{4}\right)^M$$

best expert is part of the pool

$$m \log\frac{1}{2} \le \log n + M \log\frac{3}{4}$$

$$M \log\frac{4}{3} \le \log n + m \cdot \log 2$$

not 1

$$M \le \frac{\log 2 \cdot m + \log n}{\log\frac{4}{3}} \le 2.41(m + \log n) \quad \square$$

Idea 1: discount by $1-\varepsilon$, choose ε appropriately.

$$(1-\varepsilon)^m \leq W \leq n\cdot\left(1-\frac{1}{2}\cdot(1-\varepsilon)\right)^M = n\left(\frac{1}{2}+\frac{\varepsilon}{2}\right)^M$$

$$m\log(1-\varepsilon) \leq \log n + M\log\left(\frac{1}{2}+\frac{\varepsilon}{2}\right)$$

$$M \leq \frac{\log n - m\log(1-\varepsilon)}{-\log\left(\frac{1}{2}+\frac{\varepsilon}{2}\right)} = \frac{-\log(1-\varepsilon)}{-\log\left(\frac{1}{2}(1+\varepsilon)\right)}m - \frac{\log n}{\log\left(\frac{1}{2}+\frac{\varepsilon}{2}\right)}$$

$$\leq \frac{\varepsilon+\varepsilon^2}{\frac{1}{2}(1-\varepsilon)}m - O(\log n)$$

$$= 2(1+o(\varepsilon))m - O(\log n)$$

Lem: $0 \leq x \leq \frac{1}{2}$

$$-x-x^2 \leq \log(1-x) \leq -x$$

Idea 2: Randomize. Weighted Majority (RWM)

Alg:

1. Set $W_h^1 = 1$ for all $h \in \mathcal{H}$
2. at time t, we choose $\begin{cases} 1 & \text{w.p. } \dfrac{\sum_{h\in\mathcal{H}} W_h^t \mathbb{1}(\hat{y}_h = 1)}{\sum_{h\in\mathcal{H}} W_h^t} \\ 0 & \text{otherwise} \end{cases}$
3. $W_h^{(t+1)} \leftarrow W_h^{(t)}(1-\varepsilon)$ for all $h \in \mathcal{H}$ that has made a mistake.

Analysis of RWM: At time t, F_t the fraction of the weights of the experts that make mistakes,

$$W = n \cdot (1-\varepsilon F_1)(1-\varepsilon F_2)\cdots(1-\varepsilon F_t)\cdots(1-\varepsilon F_T)$$

$$W_{t+1} = (1-F_t)\cdot W_t + F_t W_t (1-\varepsilon)$$
$$= W_t - F_t W_t \cdot \varepsilon = (1-F_t\varepsilon)W_t$$

$\boxed{\text{Lem } \log(1-x) \le -x}$

$$\log W = \log n + \sum_{t=1}^{T} \log(1-\varepsilon F_t)$$
$$\le \log n + (-\varepsilon \sum_t F_t)$$
$$= \log n - \varepsilon M$$

$$\sum_t F_t = E[\#\text{ of mistakes}] = M$$

best expert made m mistakes

weight of the best expert $\le W$ total weights

$$(1-\varepsilon)^m \le W$$

$(1-\varepsilon)^m \le (1-\varepsilon F_{i_1})\cdots(1-\varepsilon F_{i_m}) = W_{h^*}$

$\boxed{\text{Lem } -x-x^2 \le \log(1-x) \text{ if } x \in [0,\tfrac{1}{2}]}$

$$\log((1-\varepsilon)^m) \le \log n - \varepsilon M$$

$$\text{Regret} \le \varepsilon m + \frac{\log n}{\varepsilon}$$

$$m\log(1-\varepsilon) \le \log n - \varepsilon M$$

$$M \le m\cdot\left(\frac{-\log(1-\varepsilon)}{\varepsilon}\right) + \frac{\log n}{\varepsilon} \le m\cdot\left(\frac{\varepsilon+\varepsilon^2}{\varepsilon}\right) + \frac{\log n}{\varepsilon} = (1+\varepsilon)m + \frac{\log n}{\varepsilon}$$

each expert h, incur $\ell_t(h)$.

$$\text{Regret} = \sum_t \ell_t(h_t) - \min_{h \in \mathcal{H}} \sum_t \ell_t(h)$$

$\boxed{\text{Hedge alg}}$ ① Pick expert h_t w.p. $\propto \frac{w_t(h)}{\sum_{h'} w_t(h')}$.

② incur loss $\ell_t(h_t)$

③ $w_h^{t+1} = w_h^t e^{-\varepsilon \ell_t(h)}$ for all $h \in \mathcal{H}$

analysis: exercise ..

the same as RWM.

Optimization View:

$$\min_{\theta \in \mathcal{H}} \sum_i f_i(\theta) \quad \text{online} \qquad f_i(\theta) = x_i^T \theta = \left\langle \begin{pmatrix} \ell_1 \\ \ell_2 \\ \vdots \\ \ell_{|\mathcal{H}|} \end{pmatrix}, P(h \mid \theta_i) \right\rangle$$

probability simplex

generalizations ① f_i is convex. + smooth + strongly convex

② $\mathcal{H}$ affine constraints, general convex set, $\mathcal{H}$ norm, second order cone

Online Convex Optimization. Marrying Online Learning with Convex optimization

Robustness to adversarial nature

flexible modeling language