11.3 Generalization + online learning

$(\mathcal{F}, Z, \ell)$ $\quad\quad$ $\ell(f_t, z_t)$ = loss at time

Suppose $z_1, \dots z_n$ are independent, identically dist'd

Use projected gradient descent alg:

f_0 - fixed

$$f_{t+1} = \Pi\left(f_t + \alpha_t \nabla \ell(f_t, z_t)\right)$$

(as in Zinkevich)

(Use each data sample once.)

<u>Thm 11.2</u> <u>Generalization</u> ability of on-line algorithm. [$\ell \in [0,1]$] Then for any $T \geq 1$, with probability at least $1-\delta$:

(a) $$\frac{1}{T} \sum_{t=1}^{T} L(f_t) \leq \frac{1}{T} \sum_{t=1}^{T} \ell(f_t, z_t) + \sqrt{\frac{2 \log \frac{1}{\delta}}{T}}$$

generalization risk $\quad\quad$ empirical risk (observed)

(b) If $\ell(f, z)$ is convex in f and $\bar{f}_T = \frac{1}{T} \sum_{t=1}^{T} f_t$

then $$L(\bar{f}_T) \leq \frac{1}{T} \sum_{t=1}^{T} \ell(f_t, z_t) + \sqrt{\frac{2 \log \frac{1}{\delta}}{T}}$$

Proof Let $Y_0 = 0$, $Y_t = \sum_{s=1}^{t} L(f_s) - \ell(f_s, z_s)$

(conditional mean zero given f_t)

so $E[Y_{t+1} - Y_t \mid f_1 \cdots f_t, z_1, \cdots z_t, Y_1 \cdots Y_t] = 0$ i.e. Y is a martingale

($Y_{t+1} - Y_t$: martingale difference sequence with values in $[-1, 1]$)

(Hoeffding lemma: $E[e^{s(A - E[A])}] \le e^{\frac{2s^2}{(a-b)^2}}$ if $A \in [a,b]$)

$$E[e^{s(Y_{t+1} - Y_t)} \mid \text{past}] \le e^{\frac{2s^2}{4}} = e^{s^2/2}$$

$4 = (1 - (-1))^2$

So Azuma Hoeffding bound:

$$E[Y_T \ge tT] \le e^{-\frac{2t^2T^2}{T \cdot 4}} = e^{-\frac{t^2 T}{2}} = \delta$$

$$P\left[\frac{Y_T}{T} \ge \sqrt{\frac{2 \log \frac{1}{\delta}}{T}}\right]$$

$\sqrt{\frac{2 \log \frac{1}{\delta}}{T}} = t$

Proof of (b) $L(f) = E_z \ell(f, z)$ is also convex in f.

So (b) follows from (a) by Jensen's inequality.

Corollary 11.1

$$\bar{f}_n = \frac{1}{n}\sum_{t=1}^{n} f_t$$

Use projected gradient descent

$$\alpha_t = \frac{D}{L\sqrt{2t}}$$

D = diameter of $\mathcal{F}$

$= \max_{f, f' \in \mathcal{F}} \|f - f'\|$

Assume $\ell(\cdot, z)$ is L-Lipschitz and convex for each z.

Then with probability at least $1 - 2\delta$

$$L(\bar{f}_n) \leq L^* + DL\sqrt{\frac{2}{n}} + \sqrt{\frac{8\log 1/\delta}{n}}$$

Proof

(1) $$L(\bar{F}_n) \leq \frac{1}{n}\sum_{t=1}^{n} \ell(f_t, z_t) + \sqrt{\frac{2\log 1/\delta}{n}} \quad \text{wp } \geq 1-\delta$$

(generalization discussed above)

(2) $$\frac{1}{n}\sum_{t=1}^{n} \ell(f_t, z_t) \leq \frac{1}{n}\sum_{t=1}^{n} \ell(f^*, z_t) + DL\sqrt{\frac{2}{n}}$$

(by Zinkevich)

(3) $$\frac{1}{n}\sum_{t=1}^{n} \underbrace{\ell(f^*, z_t)}_{\text{independent, mean } L^*} \leq L^* + \sqrt{\frac{2\log 1/\delta}{n}} \quad \text{w/p } \geq 1-\delta$$

← Hoeffding inequality

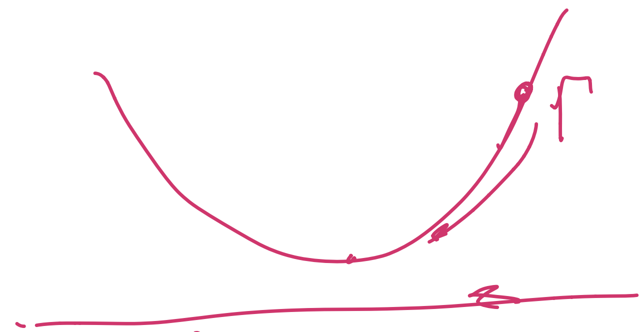
Combine (1)–(3) to complete proof.

More fresh samples:

$z_t, z_{t+1}, \cdots z_n$ are fresh samples

for f_t so $L(f_t) \approx \frac{1}{n-t+1} \sum_{s=t}^{n} \ell(f_t, z_s)$

#3 (a) $\dot{f} = -\nabla \Gamma(f)$



$f(z_t) - f^* = O(\frac{1}{t})$

$\dot{E}_t(f) \le 0$ $E_t(f_t) \le E_0(f_0)$

(b) - Accelerated gradient descent (convex optimization) - rate $O(\frac{1}{t^2})$ convergence.

4. Zinkevich with f^* - time varying (f^*_t)

constraint on expert — $\sum_{t=1}^{T-1} \|f^*_{t+1} - f^*_t\| \leq W$

5 $\sqrt{T}$ is best possible for Zinkevich

6 Python SGD

Chapter 12

12.1 Bound on average error probability for binary hypothesis testing

(Background)

Observe Y

H_0: Y has pmf p_0 ← Bayesian prior $\pi_0 = \pi_1 = 1/2$

H_1: Y has pmf p_1 ←

$f^*(y) \to \{0, 1\}$ decision rule (classifier)

$$f^*(y) = \begin{cases} 1 & \text{if } p_1(y) > p_0(y) \\ 0 & \text{if } p_1(y) < p_0(y) \end{cases}$$ (Bayesian rule)

$$\left(\bar{P}_e = \frac{1}{2} \sum_y p_0(y) f(y) + p_1(y)(1 - f(y)) \right)$$

$(\mathbb{P}\{f(Y) = H\}$ $H = H_0$ or H_1

Then $\bar{P}_e^* = \frac{1}{2} \sum_y p_0(y) \wedge p_1(y)$

$$\rho = \sum_y \sqrt{p_0(y)\,p_1(y)}$$ ← Bhattacharyya coefficient

Lemma 12.1 (c)

$$\frac{\rho^2}{4} \le P_e^* \le \frac{\rho}{2}$$

(b)

$Y_1, \dots, Y_n$

$H_0 \quad Y_i \sim p_{0,i} \quad 1 \le i \le n$

$H_1 \quad Y_i \sim p_{1,i} \quad 1 \le i \le n$

$$\rho\left(p_{0,1}(y_1)\cdots p_{0,n}(y_n),\ p_{1,1}(y_1)\cdots p_{1,n}(y_n)\right)$$
$$= \prod_{j=1}^{n} \rho(p_{0,j}, p_{1,j})$$