Concentration Inequalities (Ch. 2 of course notes)

Recap: $z_1, \dots, z_n \overset{iid}{\sim} P$ (data) $\quad f: \mathcal{Z} \to \mathbb{R}_+$

Learning algo: data $\longrightarrow \hat{f}_n \in \mathcal{F}$

$\mathbb{E}_P[\hat{f}_n(z)]$ (expectation on a fresh sample)

$$= \int_{\mathcal{Z}} \hat{f}_n(z) P(dz)$$

random! ($\hat{f}_n$ depends on z^n)

Goal: construct LA s.t.

$$\mathbb{E}_P[\hat{f}_n(z)] \approx \min_{f \in \mathcal{F}} \mathbb{E}_P[f(z)]$$

with high probability (w.r.t. draw of z^n)

Ex.: $Z = (X, Y) \qquad Y \in \{0,1\}$

$f(z) = 1_{\{y \neq \tilde{f}(x)\}} \qquad \tilde{f} : \mathcal{X} \to \{0,1\}$

Main concerns:

1) When is this possible? (restrictions on $\mathcal{F}$)
2) how much data do we need?

Example (linear classification)

$(X, Y) \in \mathbb{R}^d \times \{0,1\}$

Classifiers $\quad \tilde{f}_w(x) = 1_{\{w^T x \geq 0\}}, \quad \|w\| = 1$

Loss: $f(z) = 1_{\{\tilde{f}_w(x) \neq y\}}$

$$L_P(w) := \mathbb{P}[Y \neq \tilde{f}_w(X)]$$

want to choose $\hat{w}_n \in \mathbb{R}^d$ based on data, s.t.

$$L_P(\hat{w}_n) \approx \min_{\|w\|=1} L_P(w)$$

with high prob.

Revisit coin tossing: $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$

$\theta \in (0,1)$ unknown

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

$$\mathbb{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2} \quad \text{(Lec. 1)}$$

— this is a concentration inequality

$$\mathbb{E}\,\hat{\theta}_n = \theta$$

$$\mathbb{P}(|\hat{\theta}_n - \mathbb{E}\,\hat{\theta}_n| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$$

— $\hat{\theta}_n$ concentrates around its expectation

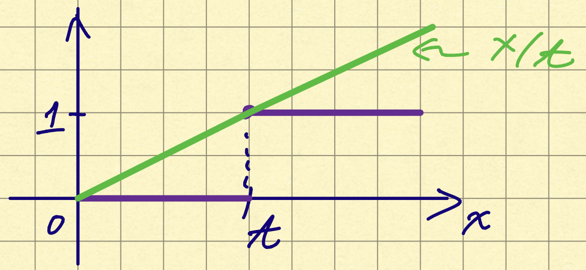
Build up:

1) Markov's inequality

X is a r.v. that takes ≥ 0 values

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}X}{t} \quad (\text{any } t > 0)$$

Proof $\mathbb{P}[X \geq t] = \mathbb{E}[1_{\{X \geq t\}}]$



$$\mathbb{E}[1_{\{X \geq t\}}]$$
$$\leq \mathbb{E}\left[\frac{X}{t} 1_{\{X \geq t\}}\right]$$
$$\leq \mathbb{E}\left[\frac{X}{t}\right] = \frac{\mathbb{E}[X]}{t} \quad \blacksquare$$

$$\mathbb{P}[X \geq t] \leq \min\left\{\frac{\mathbb{E}[X]}{t}, 1\right\}$$

2) Chebyshev's inequality

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\mathrm{Var}[X]}{t^2}$$

Proof $U := |X - \mathbb{E}X|^2$

$\mathbb{E}U = \mathrm{var}[X]$

Apply Markov: $\mathbb{P}\{|X - \mathbb{E}X| \geq t\}$

$$= \mathbb{P}\{|X - \mathbb{E}X|^2 \geq t^2\}$$
$$= \mathbb{P}\{U \geq t^2\}$$
$$\leq \frac{\mathbb{E}[U]}{t^2} \qquad \text{(Markov)}$$
$$= \frac{\mathrm{var}[X]}{t^2}. \quad \blacksquare$$

Back to coins: $\mathbb{P}\{|\hat{\theta}_n - \theta| \geq \varepsilon\}$

$$= \mathbb{P}\{|\hat{\theta}_n - \mathbb{E}\hat{\theta}_n| \geq \varepsilon\}$$
$$\leq \frac{\mathrm{Var}\{\hat{\theta}_n\}}{\varepsilon^2}$$

$$\mathrm{Var}\{\hat\theta_n\} = \mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\}$$

$$= \frac{\theta(1-\theta)}{n}$$

$X_i \overset{iid}{\sim} \mathrm{Bern}(\theta)$

$\mathrm{Var}(X_i) = \theta(1-\theta)$

Chebyshev: $\mathbb{P}\{|\hat\theta_n - \theta| \geq \varepsilon\} \leq \frac{\theta(1-\theta)}{n}$

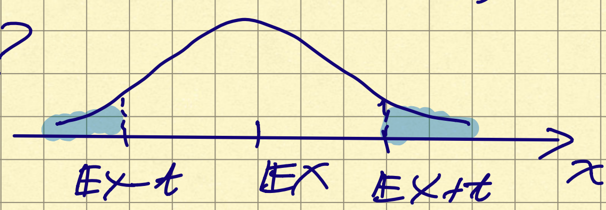
– loose!

CLT : $\mathbb{P}\{|\hat\theta_n - \theta| \geq \varepsilon\} \lessapprox 2\exp\left(-\frac{n\varepsilon^2}{2\theta(1-\theta)}\right)$

Exponential Chebyshev trick

$\mathbb{P}\{X - \mathbb{E}X \geq t\} \leq ?$ $(t \geq 0)$

$\mathbb{P}\{X - \mathbb{E}X \leq -t\} \leq ?$



$$\mathbb{P}\{X - \mathbb{E}X \geq t\}$$

$$\leq \mathbb{P}\{\underbrace{e^{s(X-\mathbb{E}X)}}_{\geq 0} \geq e^{st}\}$$

$X - \mathbb{E}X \geq t$

$\Rightarrow \underbrace{\exp(s(X-\mathbb{E}X))}_{>0} \geq e^{st}$

$$\leq \frac{\mathbb{E}[e^{s(X-\mathbb{E}X)}]}{e^{st}}$$

$$= e^{-st}\,\mathbb{E}[e^{s(X-\mathbb{E}X)}]$$

(holds for every $s \geq 0$)

Chernoff bound:

$$\mathbb{P}[X - \mathbb{E}X \geq t] \leq \inf_{s\geq 0} \{e^{-st}\mathbb{E}[e^{s(X-\mathbb{E}X)}]$$

Good (tight) bounds on $\mathbb{E}[e^{s(X-\mathbb{E}X)}]$ needed!

Assume w.l.o.g. $\mathbb{E}X = 0$, analyze

$$\psi(s) := \log \mathbb{E}[e^{sX}] \qquad (s \geq 0)$$

Chernoff: $$\mathbb{P}[X \geq t] \leq \inf_{s \geq 0} \{e^{-st + \psi(s)}\}$$
$$= \exp\left(-\sup_{s\geq 0}[st - \psi(s)]\right).$$

<u>Lemma</u> (Hoeffding) Suppose X takes values between a and b w.p. 1 (a, b finite). Then

$$\mathbb{E}[e^{s(X-\mathbb{E}X)}] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right), \quad \forall s \geq 0.$$

Implications:

1) If $a \leq X \leq b$ a.s., then

$$\psi(s) \leq \frac{s^2(b-a)^2}{8} \qquad (\text{assume } \mathbb{E}X = 0)$$

$$\Rightarrow \mathbb{P}[X \geq t] \leq \exp\left(-\sup_{s \geq 0}\left(st - \frac{s^2(b-a)^2}{8}\right)\right)$$

$$g(s) = -\frac{(b-a)^2}{8}s^2 + st$$

$$g'(s) = -\frac{(b-a)^2}{4}s + t = 0$$

max. uniquely attained at $\bar{s} = \frac{4t}{(b-a)^2}$

$$\bar{s}t - \frac{1}{8}(b-a)^2\bar{s}^2 = \frac{4t^2}{(b-a)^2} - \frac{1}{8}(b-a)^2 \cdot \frac{16t^2}{(b-a)^4}$$

$$= \frac{4t^2}{(b-a)^2} - \frac{2t^2}{(b-a)^2} = \frac{2t^2}{(b-a)^2}$$

$$\Rightarrow \mathbb{P}\{X \geq t\} \leq \exp\left(-\frac{2t^2}{(b-a)^2}\right)$$

2) $X_1, \dots, X_n \overset{iid}{\sim} P$ $\quad a \leq X_i \leq b$ a.s.

$\mathbb{E}X_i = 0$

$$\mathbb{P}\left\{\sum_{i=1}^n X_i \geq t\right\}$$

$$\leq \inf_{s \geq 0} e^{-st}\,\mathbb{E}\left[e^{s\sum_{i=1}^n X_i}\right]$$

$$\mathbb{E}\left[e^{s\sum_{i=1}^n X_i}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{sX_i}\right]$$

$$= \prod_{i=1}^n \mathbb{E}\left[e^{sX_i}\right] \quad \text{(by indep.)}$$

$$\leq \left(e^{\frac{s^2(b-a)^2}{8}}\right)^n$$

$$= \exp\left(\frac{(b-a)^2}{8} n s^2\right)$$

$$\Rightarrow \mathbb{P}\left\{\sum_{i=1}^n X_i \geq t\right\} \leq \min_{s \geq 0} e^{-st + \frac{(b-a)^2}{8} n s^2}$$

$$= \exp\left(-\frac{2t^2}{n(b-a)^2}\right)$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \varepsilon\right\} = \mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq n\varepsilon\right\}$$

$$\leq \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right).$$

3) Coin tossing: $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{Bern}(\theta)$

$$0 \leq X_i \leq 1$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \theta + \varepsilon\right\} \leq \exp(-2n\varepsilon^2)$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \leq \theta - \varepsilon\right\} \leq \exp(-2n\varepsilon^2).$$

Proof of Hoeffding's lemma: next lecture.

\+ McDiarmid's inequality.