10 Feb 2025 — Hoeffding's Inequality and Applications

Suppose $X_1, \ldots, X_n$ are independent RV's

with $X_i \in [a_i, b_i]$ for $1 \le i \le n$

Let $X = X_1 + \cdots + X_n$. Then

Hoeffding

$$\Pr(X \ge \mathbb{E}[X] + \lambda) \le \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

$$\Pr(X \le \mathbb{E}[X] - \lambda) \le \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Comparing the power of Chernoff and Hoeffding: suppose a fair coin is tossed n times. What is probab. of getting less than $\left(\frac{1-\varepsilon}{2}\right)\cdot n$ heads?

CHERNOFF: $X_i = \begin{cases} 1 & \text{if } i^{th} \text{ toss is heads} \\ \emptyset & \text{if } i^{th} \text{ toss is tails} \end{cases}$

$X = X_1 + \cdots + X_n$
$= \#$ heads

$\mathbb{E}X = \frac{n}{2}$.

$$\Pr\left(< \left(\tfrac{1-\varepsilon}{2}\right)\cdot n \text{ heads}\right) = \Pr\left(X < (1-\varepsilon)\cdot \mathbb{E}[X]\right)$$

$$< \exp\left(-\tfrac{1}{2}\varepsilon^2 \mathbb{E}[X]\right) = \exp\left(-\frac{\varepsilon^2 n}{4}\right)$$

HOEFFDING: $\Pr\left(< \left(\tfrac{1-\varepsilon}{2}\right)\cdot n \text{ heads}\right) = \Pr\left(X < \mathbb{E}[X] - \underbrace{\tfrac{\varepsilon n}{2}}_{\lambda}\right)$

$a_i = 0$
$b_i = 1$

$$< \exp\left(-\frac{2(\varepsilon n/2)^2}{\sum (b_i - a_i)^2}\right) = \exp\left(-\frac{2\varepsilon^2 n^2/4}{n}\right) = \exp\left(-\frac{\varepsilon^2 n}{2}\right)$$

Summary. Chernoff and Hoeffding deliver similar results but:

(a) Hoeffding usually has better constant factor

(b) Hoeffding can be applied when variables may be negative or positive

(c) Chernoff gives good bounds as long as $\mathbb{E}[X]$ is large, doesn't care how many variables there are.

Proof (sketch) of Hoeffding. Very similar to Chernoff.

If $X \geq \mathbb{E}[X] + \lambda$ then $\boxed{e^{tX} \geq e^{t\mathbb{E}[X]} \cdot e^{t\lambda}}$

Use Markov inequality on this event to show it is unlikely.

That means we need to bound $M_X(t) = \mathbb{E}[e^{tX}]$ or equivalently $K_X(t) = \ln M_X(t)$.

$$K_X(t) = K_{X_1}(t) + K_{X_2}(t) + \dots + K_{X_n}(t).$$

Hoeffding's Lemma: If X_i is supported on $[a_i, b_i]$ then $K_{X_i}(t) \leq \exp\left(-\frac{(b_i - a_i)^2 t^2}{8}\right)$.

Proof details in online lecture notes.

APPLICATION 1. Generalization error of ERM.

Given n random samples $z_1, z_2, \ldots, z_n$ drawn from distribution $\mathcal{D}$ on set $\mathcal{Z}$.

Given hypotheses $\mathcal{H} = \{h_1, \ldots, h_m\}$.

Given loss function $L : \mathcal{H} \times \mathcal{Z} \to [0,1]$.

$L(h, z) =$ "how badly does hypoth h fail to explain data point z?"

ERM algorithm: select $h^{emp} \in \mathcal{H}$ to minimize the empirical average loss

$$L^{emp}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, z_i).$$

What you really want to minimize is population loss $\overline{L}(h) = \mathop{E}_{z \sim \mathcal{D}} [L(h,z)]$.

Reducing error of randomized algorithms.

Suppose Π is a decision problem.

$\forall x \quad \Pi(x) \in \{0,1\}$.

Suppose A is a randomized algorithm running in time $poly(|x|)$ with random bits $r \in \{0,1\}^*$ s.t.

$$\forall x \quad \Pr_r (A(x,r) = \Pi(x)) \geq 2/3.$$

We say $\Pi \in BPP$ and A is a (poly-time) randomized decision procedure for Π.

Boosting confidence: run A on independent coin toss sequences $r_1, \ldots, r_m$ and take majority vote.

Claim: $\forall \delta > 0$ If $m > 18 \ln(1/\delta)$ and we take $MAJ(A(x,r_1), A(x,r_2), \ldots, A(x,r_m))$ then $\Pr(MAJ \neq \Pi(x)) < \delta$.

Proof. Let $X_i = \begin{cases} 1 & \text{if } A(x,r_i) = \Pi(x) \\ 0 & \text{otherwise.} \end{cases}$

$\forall i \quad \mathbb{E}[X_i] \geq 2/3.$

But if $MAJ \neq \Pi(x)$ it means

$$X_1 + \cdots + X_m \leq \frac{m}{2}.$$

$$\mathbb{E}(X_1 \cdots X_m) \geq 2m/3.$$

$$\Pr\left(X < \mathbb{E}X - \frac{m}{6}\right) < \exp\left(-\frac{2(m/6)^2}{m}\right)$$

$$= \exp\left(-\frac{m^2/18}{m}\right) = \exp\left(-\frac{m}{18}\right) < \delta$$