Efron-Stein and McDiarmid Inequalities

So far

- Markov, Chebyshev & Chernoff
- Subgaussian & Subexponential
- (Exponential) tail bounds on X & $\sum_{i=1}^{n} \alpha_i X_i$

Want: Tail bounds on functions of $(X_1 \ldots X_n)$ iid

① Largest eigenvalue of a random matrix

② Max degree of a random graph

Suppose $X_1 \cdots X_n$

$$Z = g(X_1 \cdots X_n)$$

$$Z_i = \mathbb{E}[Z \mid X_1 \cdots X_i]$$

$$= \int g(X_1 \cdots X_i, x_{i+1} \cdots x_n) \, f_{X_{i+1} \cdots X_n \mid X_1 \cdots X_i}(x_{i+1} \cdots x_n) \, dx_{i+1} \cdots dx_n$$

$$Z_0 = \mathbb{E} g(X_1 \cdots X_n) = \mathbb{E} Z$$

$$Z_n = Z = g(X_1 \cdots X_n)$$

$$\mathrm{Var}(g(X_1 \cdots X_n)) = \mathbb{E}[(Z_n - Z_0)^2]$$

$$Z_n - Z_0 = \sum_{i=1}^{n} Z_i - Z_{i-1}$$

$$Var(Z) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \underbrace{(Z_i - Z_{i-1})}_{\Delta_i} \right)^2 \right]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{n} \Delta_i \right)^2 \right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} \Delta_i^2 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \Delta_i \Delta_j \right] \quad \text{— ①}$$

$$\Delta_i = Z_i - Z_{i-1}$$

$$Z_i = \mathbb{E}[Z \mid X_1 \cdots X_i] = \mathbb{E}_{X_{i+1} \cdots X_n}[Z]$$

$$Z_{i-1} = \mathbb{E}[Z \mid X_1 \cdots X_{i-1}] = \mathbb{E}_{X_i \cdots X_n}[Z]$$

$$Z_{i-1} = \mathbb{E}_{X_i \cdots X_n}[Z] = \mathbb{E}_{X_i}\, \mathbb{E}_{X_{i+1} \cdots X_n}[Z]$$

$$= \mathbb{E}_{X_i} Z_i$$

$$\Delta_i = Z_i - \mathbb{E}_{X_i} Z_i = \mathbb{E}_{X_{i+1}^n} f(X^n) - \mathbb{E}_{X_i^n} f(X^n)$$

$$\mathbb{E}_{X^n} \Delta_i = 0$$

Consider $j > i$ $\quad \mathbb{E}_{X^n}(\Delta_i \Delta_j)$

$$\mathbb{E}_{X^n}\left[\left(\mathbb{E}_{X_{i+1}^n} Z - \mathbb{E}_{X_i^n} Z\right) \times \left(\mathbb{E}_{X_{j+1}^n} Z - \mathbb{E}_{X_j^n} Z\right)\right]$$

$$X_{i+1}^n = (X_{i+1} \cdots X_n)$$

$$\underbrace{\mathbb{E}_{X_i, X_{j-1}, X_{j+1}, X_j}}_{\mathbb{E}_{X^n}} \mathbb{E}_{X_j}\left[\underbrace{\left(\mathbb{E}_{X_{i+1}^n} Z - \mathbb{E}_{X_i^n} Z\right)}_{\substack{\text{not a function} \\ \text{of } X_j \\ \text{for } j>i}} \times \left(\mathbb{E}_{X_{j+1}^n} Z - \mathbb{E}_{X_j^n} Z\right)\right]$$

$$= \left(\mathbb{E}_{X_{i+1}^n} Z - \mathbb{E}_{X_i^n} Z\right) \underbrace{\mathbb{E}_{X_j}\left(\mathbb{E}_{X_{j+1}^n} Z - \mathbb{E}_{X_j^n} Z\right)}_{\mathbb{E}_{X_j}\Delta_j}$$

$$\Rightarrow \mathbb{E}_{X^n} \Delta_i \Delta_j = 0 \quad \text{for } j>i$$

from ①,

$$\mathrm{Var}(Z) = \mathbb{E}\left[\sum_{i=1}^{n} \Delta_i^2\right]$$

$$= \sum_{i=1}^{n} \mathrm{Var}(\Delta_i)$$ → true even for non iid $X_1 \cdots X_n$

$Z_1 \text{ — } Z_n \qquad (X_1 \cdots X_n)$

If $\mathbb{E}[Z_{i+1} / X_1 \cdots X_i] = Z_i \quad \forall i,$

we say that $Z_1 \cdots Z_n$ is a martingale wrt $X_1 \cdots X_n$

Theorum (Efron-Stein inequality)

If $X_1 \cdots X_n$ independent & have bdd variance

$g : \mathcal{X}^n \to \mathbb{R}$

$Z = g(X_1 \cdots X_n)$

$$Var(Z) \leq \sum_{i=1}^{n} \mathbb{E}(\underbrace{Z - \mathbb{E}_{X_i} Z}_{})^2 = \sum_{i=1}^{n} \mathbb{E}_{X^n}\left(f(X^n) - \mathbb{E}_{X_i} f(X^n)\right)^2$$

$Y_i = Z - \mathbb{E}_{X_i} Z$

$$= \sum_{i=1}^{n} Var(Y_i)$$

Define $Z_i' = g(X_1 \cdots X_{i-1}, X_i', X_{i+1} \cdots X_n)$

replace X_i with an iid copy

$$\mathrm{Var}(Z) \le \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}(Z - Z_i')^2$$

$$\le \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}([Z - Z_i']_+)^2$$

$$= \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}([Z - Z_i']_-)^2$$

$[\alpha]_+ = \max\{\alpha, 0\}$

$[\alpha]_- = -[-\alpha]_+$

$$\le \sum_{i=1}^{n} \inf_{U_i = h(X_1 \cdots X_{i-1}, X_{i+1} \cdots X_n)} \mathbb{E}[(Z - U_i)^2]$$

$\mathrm{Var}(U_i) < \infty$

Suppose

$$|g(x_1 \cdots x_n) - g(x_1 \cdots x_{i-1}, x_i', x_{i+1} \cdots x_n)| \leq c_i$$

$$Var(Z) = Var(g(X^n)) \leq \frac{1}{2} \sum_{i=1}^{n} c_i^2$$

Proof of Efron-Stein inequality

$$\mathrm{Var}(Z) = \sum_{i=1}^{n} \mathbb{E}\Delta_i^2 \le \sum_{i=1}^{n} \mathbb{E}_{X^n}\left(g(X^n) - \mathbb{E}_{X_i} g(X^n)\right)^2$$

$$\Delta_i = \mathbb{E}_{X_{i+1}^n} g(X^n) - \mathbb{E}_{X_i^n} g(X^n)$$

$$= \mathbb{E}_{X_{i+1}^n}\left[g(X^n) - \mathbb{E}_{X_i} g(X^n)\right]$$

$$\Delta_i^2 = \left[\mathbb{E}_{X_{i+1}^n}\left(g(X^n) - \mathbb{E}_{X_i} g(X^n)\right)\right]^2$$

$$\le \mathbb{E}_{X_{i+1}^n}\Big[\underbrace{\left(g(X^n) - \mathbb{E}_{X_i} g(X^n)\right)^2}_{Y_i}\Big]$$

$$\mathbb{E}_{X^n}\Delta_i^2 \le \mathbb{E}_{X^n}\left[\left(g(X^n) - \mathbb{E}_{X_i} g(X^n)\right)^2\right]$$

$\mathbb{E}Y_i^2 \ge$

$(\mathbb{E}Y_i)^2$

$\Updownarrow$

$\mathrm{Var}(Y_i) \ge 0$

$$\mathrm{Var}(g(X^n)) \le \sum_{i=1}^{n} \mathbb{E}_{X^n} \underbrace{\left(g(X^n) - \mathbb{E}_{X_i} g(X^n)\right)^2}$$

$$Z_i' = g(X_1 \cdots X_{i-1}, X_i', X_{i+1} \cdots X_n)$$

X_i' ↓ iid copy

CI: Conditional on $X_1 \cdots X_{i-1}, X_{i+1} \cdots X_n$, Z, Z_i' are iid

$$\mathbb{E}_{X_i X_i'} (Z_i - Z_i')^2 = 2\, \mathbb{E}_{X_i} \left(Z_i - \mathbb{E}_{X_i} Z \right)^2$$

Claim: X & Y iid

$$\mathrm{Var}(X) = \tfrac{1}{2} \mathbb{E}(X - Y)^2$$

If C_1 is true,

$$\mathrm{Var}(Z) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}_{X^n, X_i'} (Z - Z_i')^2$$

Proof of claim:

$$\frac{1}{2} \mathbb{E}(X-Y)^2$$

$$= \frac{1}{2} \mathbb{E}\left[(X - \mathbb{E}X) - (Y - \mathbb{E}Y)\right]^2$$

$$= \frac{1}{2} \Big[\mathbb{E}(X - \mathbb{E}X)^2 + \mathbb{E}(Y - \mathbb{E}Y)^2 - 2 \underbrace{\mathbb{E}\left[(X-\mathbb{E}X)(Y-\mathbb{E}Y)\right]}_{0} \Big]$$

$$= \mathrm{Var}(X)$$

HW: Show that if X & Y are iid,

$$\frac{1}{2}\mathbb{E}(X-Y)^2 = \mathbb{E}\left((X-Y)_+\right)^2$$

$$= \mathbb{E}\left((X-Y)_-\right)^2$$

$$(X-Y)_+ = \max\{(X-Y), 0\}$$

$$(X-Y)_- = \min\{(X-Y), 0\}$$

Example: Bin packing problem

$$0.4, \quad 0.3, \quad 0.6, \quad 0.5$$

$X_1 \dots X_n$ iid $\qquad X_i \in [0, 1]$

Goal: pack $X_1 \dots X_n$ into min # bins.

(each bin has size $= 1$)

$g(X_1 \dots X_n) =$ min # bins required

Changing x_i can change $g(x_1 \dots x_n)$ by at most 1

$$\mathrm{Var}(g(X_1 \dots X_n)) \le \frac{n}{4}$$

Example 2: Longest common subsequence

X^n = r e c e n t

Y^n = excellent

$g(X^n, Y^n)$ = length of longest common subsequence

X^n, Y^n iid

$\frac{\mathbb{E}\, g(X^n, Y^n)}{n}$ is conjectured $\frac{2}{1+\sqrt{2}}$ for Ber$(\frac{1}{2})$

$\mathrm{Var}(g(X^n, Y^n)) \leq \frac{n}{2}$

$$P_n\left[\, |g(x^n, y^n) - \mathbb{E}\, g(x^n, y^n)| > \delta\, \mathbb{E}\, g(x^n, y^n) \right]$$

$$\leq \frac{\mathrm{Var}(g)}{\delta^2 (\mathbb{E} g)^2} \leq \frac{\frac{n}{2}}{\delta^2 (C n)^2}$$

$$= \frac{1}{C\delta^2} \frac{1}{n}$$

McDiarmid's inequality — If $X_1 \dots X_n$ are independent

$$g: \mathcal{X}^n \to \mathbb{R}$$

$$|g(x_1 \dots x_n) - g(x_1 \dots x_{i-1}, x_i', x_{i+1} \dots x_n)| \leq c_i$$

$$\forall x^n, x_i'$$

then,

$$\Pr\left[\,|g(X^n) - \mathbb{E}g(X^n)| > t\right] \leq e^{-2t^2/\sum_{i=1}^n c_i^2}$$

In fact, suppose

$$\sum_{i=1}^n (z - z_i')^2 \leq \nu^2 \quad \text{with prob } 1,$$

$$\Pr\left[\,|g(X^n) - \mathbb{E}g(X^n)| > t\right] \leq e^{-t^2/\nu^2}$$

Kernel density estimation

$X_1 \dots X_n \sim \text{iid } f_X$

$\hookrightarrow$ unknown

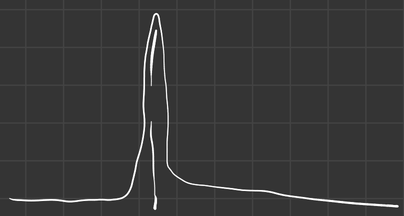
$K : \mathbb{R} \to \mathbb{R}$ smooth (Kernel)

$$\int_{-\infty}^{\infty} K(x)\, dx = 1$$

$K(x) \geq 0$

$\forall x$

$$\phi_n(x) = \frac{1}{n h_n} \sum_{i=1}^{n} \underbrace{K\left(\frac{x - X_i}{h_n}\right)}$$



Mean X_i

$\text{var} = h_n^2$

$$\text{Error} = \int_{-\infty}^{\infty} |f_X(x) - \phi_n(x)|\, dx$$

$$\Pr\left[\text{Error} > \mathbb{E}(\text{error}) + \delta\right]$$

$$g(X_1 - X_n) = \int_{-\infty}^{\infty} |f_X(x) - \phi_n(x)|\, dx$$

$$|g(x_1 - x_n) - g(x_1 \cdots x_{i-1}, x_i', x_{i+1} \cdots x_n)|$$

$$= \left| \int_{-\infty}^{\infty} |f_X(x) - \phi_n(x)| - |f_X(x) - \phi_n'(x)|\, dx \right|$$

$$\leq \int_{-\infty}^{\infty} |\cancel{f_X(x)} - \phi_n(x) - (\cancel{f_X(x)} - \phi_n'(x))|\, dx$$

$$= \int_{-\infty}^{\infty} | \phi_n'(x) - \phi_n(x) | \, dx$$

$$= \int_{-\infty}^{\infty} \left| \frac{1}{nh_n} \left(K\left(\frac{x - x_i}{h_n}\right) - K\left(\frac{x - x_i'}{h_n}\right) \right) \right| dx$$

$$\leq \int_{-\infty}^{\infty} \frac{1}{nh_n} \left(K\left(\frac{x - x_i}{h_n}\right) + K\left(\frac{x - x_i'}{h_n}\right) \right) dx$$

$$y = \frac{x - x_i}{h_n} \qquad dy = \frac{dx}{h_n}$$

$$\leq \int_{-\infty}^{\infty} \frac{1}{n} \left(K(y) + K(y') \right) dy = \frac{2}{n}$$

$$\text{Error} = \int_{-\infty}^{\infty} |f_X(x) - \phi_n(x)|\, dx \le \int_{-\infty}^{\infty} (f_X(x) + \phi_n(x))\, dx$$

$$\le 2$$

$$\mathbb{E}\,\text{error} \le 2$$

$$\text{Var}(\text{error}) \le \frac{1}{4} \times \sum_{i=1}^{n} c_i^2 \le \frac{1}{n}$$

$$\frac{1}{\sqrt{n}}$$

$$\Pr\left[\text{Error} \geq \mathbb{E}\,\text{error}\,(1+\delta)\right] \le \frac{1}{\delta^2 n}$$

$$e^{-\delta^2 n}$$

Empirical Risk Minimization

Classification

X → image

Y → is object present

$X \in \mathcal{X}$ $\quad$ $Y \in \{1, -1\}$

$$(X, Y) \sim P_{XY}$$

Goal: Design $g : \mathcal{X} \to \{1, -1\}$

g → classifier

$$R_g = \Pr[g(X) \neq Y]$$ → risk for classifier g

$R_g = \mathbb{E}\, \ell(g(X), Y)$

$\downarrow$

$\mathbb{E}(\text{risk}) = \mathbb{E}\, \mathbb{1}_{\{g(X) \neq Y\}}$

Suppose that we knew p_{XY}. What g minimizes R_g?

$$g^*(x) = \underset{y \in \{1, \dots\}}{\operatorname{argmax}}\; p_{Y|X}(y \mid x) \quad \text{(MAP estimate)}$$

$R(g^*) = $ Minimum Bayes risk

But we do not have p_{XY}

Dataset: $(X_1, Y_1)\, (X_2, Y_2) \cdots (X_n, Y_n) \sim \text{iid}\, (p_{XY})$

$$R_n(g) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{g(x_i)\neq y_i\}}}_{\text{fraction of time prediction is wrong}} \rightarrow \text{Empirical risk}$$

$R_n(g)$ ↓ empirical risk

Empirical risk minimization

$$g_n = \operatorname*{argmin}_{g\in\mathcal{G}} R_n(g)$$

What can we say about

$$|R(g) - R_n(g)| = \left|\mathbb{E}\,\mathbb{1}_{\{g(x)\neq y\}} - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{g(x_i)\neq y_i\}}\right.$$

$$\mathbb{E}\,R_n(g) = R(g)$$

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(x_i) \neq y_i\}$$

$$R(g) = \mathbb{E}_{X,Y}\, \mathbb{1}\{g(X) \neq Y\} = \Pr[g(X) \neq Y]$$

$$\Pr\left[\, |R_n(g) - R(g)| \geq \varepsilon R(g) \right] \leq 2 e^{-c \varepsilon^2 n R(g)}$$

Want :

$$R_n(g_n) - \underbrace{R(g^*)}_{\text{optimum}}$$

$$\downarrow$$

$$\operatorname*{argmin}_{g \in \mathcal{G}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(x_i) \neq y_i\}}$$

$$\operatorname{argmin} Z(g, X^n, Y^n)$$

Toy example : $(X_1, Y_1) \cdots (X_n, Y_n)$

$$\mathcal{G} = \{ g_1 , g_{-1} \}$$

$$g_1(x) = 1 \quad \forall x$$

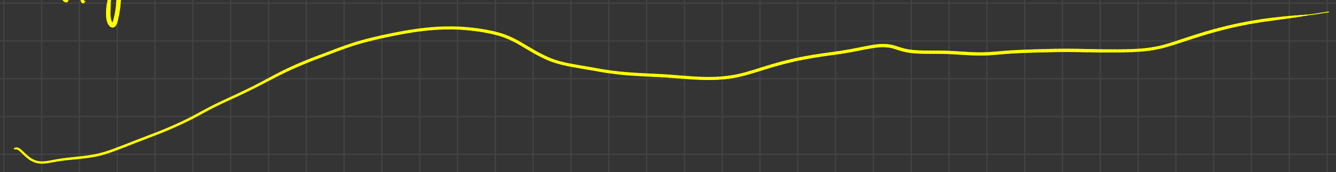
$$g_{-1}(x) = -1 \quad \forall x$$

$$P(Y=1) = \alpha \qquad P(Y=-1) = 1-\alpha$$

$$\sum_{i=1}^{n} 1_{\{g(X_i) \neq Y_i\}} \begin{cases} \text{Bin}(n, 1-\alpha) & \text{if } g = g_1 \\ \text{Bin}(n,\alpha) & \text{if } g = g_{-1} \end{cases}$$

$$\min_{g \in \{g_1, g_{-1}\}} \sum_{i=1}^{n} 1_{\{g(X_i) \neq Y_i\}} = \min\{ \#\text{ of } 1\text{'s}, \#(-1)\text{s} \}$$

$$\inf_{g \in \mathcal{G}} R_n(g)$$



$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(X_i) \neq Y_i\}$$

$$g_n = \operatorname*{argmin}_{g \in \mathcal{G}} R_n(g)$$

$$g^* = \operatorname*{argmin}_{g \in \mathcal{G}} P_n[g(X) \neq Y]$$

$R(g^*)$

$R_n(g^*)$

$$R^* = \operatorname*{argmin}_{g} P_n[g(X) \neq Y]$$

Observables: X^n, Y^n $\quad g_n \quad R_n(g_n)$

Performance of g_n on a (new) test sample:

$$P_n[g_n(x) \neq y] = R(g_n)$$

$$|R_n(g_n) - R(g_n)|$$

$$R(g_n) - R(g^*)$$

min (Test) risk for g from ERM

min (test) risk over all classifiers in $\mathcal{G}$

Goal: ST $R(g_n) - R(g^*)$ is small

OR: What is min n st $R(g_n) \approx R(g^*)$ whp

We know: for a given $g \in \mathcal{G}$

$$\Pr\left[\,|R_n(g) - R(g)| > \varepsilon R(g)\right] \leq e^{-cn\varepsilon^2}$$

For any g

$$R(g) = R_n(g) + R(g) - R_n(g)$$

$$\leq R_n(g) + \sup_{g \in \mathcal{G}} \left(R(g) - R_n(g)\right)$$

Want:

$$\Pr\left[\sup_{g \in \mathcal{G}} \left(R(g) - R_n(g)\right) > \varepsilon\right]$$

$$\leq \sum_{g \in \mathcal{G}} \Pr\left[R(g) - R_n(g) > \varepsilon\right] \leq |\mathcal{G}|\, e^{-nc\varepsilon^2}$$

<u>Theorum</u>: If $\mathcal{G}$ is finite

$$\Pr\left[R(g) \geq R_n(g) + 2\sqrt{\frac{\log|\mathcal{G}| + \log 2/\delta}{2n}}\,\right] \leq \delta$$

$$\Pr\left[R_n(g) \geq R(g^*) + 2\sqrt{\frac{\log|\mathcal{G}| + \log 2/\delta}{2n}} \right] \leq \delta$$

What if $\mathcal{G}$ is infinite

$$\eta_g(X_i, Y_i) = \mathbb{1}_{\{g(X_i) \neq Y_i\}}$$

$$\mathcal{J}_{X^n, Y^n} = \{ (\eta_g(X_1, Y_1) \cdots \eta_g(X_n, Y_n)) : g \in \mathcal{G} \}$$

$$|\mathcal{J}_{X^n, Y^n}| = 2^n$$

For given $\mathcal{G}$

Growth function: $S_{\mathcal{G}} = \sup_{(x^n, y^n)} \mathcal{J}_{x^n, y^n}$ $\longrightarrow$ Measures how diverse $\mathcal{G}$ is

Theorem (Vapnik - Chervonenkis)

$$\Pr\left[R(g) > R_n(g) + 2\sqrt{\frac{2\log S_{\mathcal{G}}(2n) + \log\frac{2}{\delta}}{n}} \quad \text{for any } g \in \mathcal{G}\right]$$

$$\leq \delta$$

" Introduction to Statistical Learning Theory "

Vapnik, SLT