Q1 : Sparse principal component analysis

(a) Under H_1, $A = \lambda u v^T + B$,

$$\mathbb{P}\left[\mathrm{val}(A) > \lambda - \frac{t}{\sqrt{d}} \right]$$

$$= \mathbb{P}\left[\sup_{\substack{x,y \in B_2^d \\ \|x\|_0 = \|y\|_0 = k}} x^T A y > \lambda - \frac{t}{\sqrt{d}} \right]$$

$$\leq \mathbb{P}\left[\sup u^T u v^T v \cdot \lambda + u^T B v > \lambda - \frac{t}{\sqrt{d}} \right]$$

$$\leq \mathbb{P}\left[\lambda + u^T B v > \lambda - \frac{t}{\sqrt{d}} \right]$$

$$= \mathbb{P}\left[u^T (B\sqrt{d}) v > -t \right]$$

$$= 1 - \mathbb{P}\left[u^T (B\sqrt{d}) v \leq -t \right]$$

But $u^T \sqrt{d} B v = \sum_{i,j=1}^{d} u_i v_j \sqrt{d} B_{ij}$ and

$u_i v_j \sqrt{d} B_{ij}$ is $|u_i v_j|$ - subgaussian by HW 6, 5a)

So $u^T \sqrt{d} B v$ is the sum of iid $|u_i v_j|$ - subgaussian RVs and hence it is a $\sum_{i,j=1}^{d} u_i^2 v_j^2 = 1$ - subgaussian RV, with mean zero.

Hence $\mathbb{P}\left[u^T B v \leq -t\right] \leq e^{-t^2/2}$ and hence

$$\mathbb{P}\left[u^T B v > -t\right] \geqslant 1 - e^{-t^2/2}.$$

(b) WTS $\exists\ c < 10$ s.t. if $A = B$,

$$\mathbb{P}\left[\ \mathrm{val}(A) > c\sqrt{\frac{k}{d}\left(1 + \log\frac{d}{k}\right) + \frac{1}{d}\log\left(\frac{1}{\delta}\right)}\ \right] \leq d$$

Let $\tilde{B} = B_0^d(k) \cap B_1^d(1)$. We have $\mathrm{Val}(A) = \sup_{(x,y) \in \tilde{B}^2} x^T A y$.

Consider N an ε-net of $\tilde{B}$ for the 2-norm, s.t.
$\forall x \in \tilde{B}$, $\exists\ y \in N$ s.t. $\|x - y\| \leq \varepsilon$ and $\mathrm{supp}(x) = \mathrm{supp}(y)$, and
s.t. N is minimal. Then

$$|N| \leq \binom{d}{k}\left(1 + \frac{2}{\varepsilon}\right)^k$$

since there are maximum $\binom{d}{k}$ combinations of k points from d-dimensional vects and using the upper bound from lectures on the size of an ε-net (minimal).

$$|N| \leq \left(\frac{ed}{k}\right)^k \left(1 + \frac{2}{\varepsilon}\right)^k \qquad \text{(Stirling's)}$$

Then if we take $(x,y) \in \tilde{B}_1^2$ such that $x^T A y = \mathrm{val}(A)$, and $(\tilde{x}, \tilde{y}) \in N^2$ the corresponding vectors in N, then

$$x^T A y - \tilde{x}^T A \tilde{y} = (x-\tilde{x}) A y + \tilde{x}^T A (y - \tilde{y})$$

$$= \|x-\tilde{x}\|_2 \frac{(x-\tilde{x})^T}{\|x-\tilde{x}\|_2} A y + \tilde{x}^T A \frac{(y-\tilde{y})}{\|y-\tilde{y}\|_2} \|y-\tilde{y}\|_2$$

$$\le \|x-\tilde{x}\|_2 \mathrm{val}(A) + \mathrm{Val}(A) \|y-\tilde{y}\|_2$$

$$\le 2 \varepsilon x^T A y$$

Then $\mathbb{P}[\mathrm{val}(A) > t] \le \mathbb{P}\left[\sup_{\tilde{x},\tilde{y} \in N} \tilde{x}^T A \tilde{y} \geq t(1-2\varepsilon)\right]$

$$\le |N|^2 \sup_{\tilde{x},\tilde{y} \in N} \mathbb{P}\Big[\underbrace{\tilde{x}^T B \tilde{y}}_{\sqrt{d}\text{- subgaussian by (a)}} \geq t(1-2\varepsilon)\Big] \quad \text{(union bound)}$$

$$\le |N|^2 \sup_{\tilde{x},\tilde{y} \in N} \exp\{-t^2(1-2\varepsilon)^2 d/2\}$$

$$\le \left(\frac{ed}{k}\right)^{2k} \left(1+\frac{2}{\varepsilon}\right)^{2k} \exp\{-t^2(1-2\varepsilon)^2 \tfrac{d}{2}\}$$

Take $\varepsilon = \frac{1}{4}$ Then

$$\mathbb{P}_{H_0}[\mathrm{val}(A) > t] \le \left[\left(\frac{ed}{k}\right)^{2k} \cdot 9^{2k} \cdot \exp\left\{\frac{-t^2}{8} d\right\}\right]$$

$$\le \left[\left(\frac{9ed}{k}\right) \exp\left\{\frac{-t^2}{16} \frac{d}{k}\right\}\right]^{2k}$$

To have $P_{H_0}[\mathrm{val}(A) > t] \le \delta$, we need $\delta \geq \left(\left(\frac{9ed}{k}\right) \cdot \exp\left\{\frac{-t^2 d}{16 k}\right\}\right)^{2k}$ i.e.

It is sufficient to choose

$$t^2 \geq 16\left(\frac{k}{d}\right)\left(\log 9e + \log\frac{d}{k} + \frac{1}{2k}\log\frac{1}{\delta}\right)$$

So taking $t(k,d) = 4\sqrt{\log 9e}\sqrt{\frac{k}{d}\left(1+\log\frac{d}{k}\right) + \frac{1}{d}\log\frac{1}{\delta}}$

is enough, hence we get the result with

$$c = 4\sqrt{\log 2e} \leq 7.2 < 10$$

(c) For a fixed threshold $\tau = O(1)$, we want to have non-negligible power and level, power > level.

i.e. $$\begin{cases} \tau = 1 - \frac{t}{\sqrt{d}} = c\sqrt{\frac{k}{d}\left(1+\log\frac{k}{d}\right) + \frac{1}{d}\log\left(\frac{1}{\delta}\right)} \\ 1-e^{-t^2/2} > \delta \end{cases}$$

$1-e^{-t^2/2} > \delta$ gives us $\log(1-\delta) > -\frac{t^2}{2}$ but $\log(1-\delta) \simeq -\delta$ for small δ so $t > \sqrt{2\delta}$. Plugging into the first equation gives

$$1 - \sqrt{\frac{2\delta}{d}} > c\sqrt{\frac{k}{d}\left(1+\log\left(\frac{k}{d}\right)\right) + \frac{1}{d}\log\frac{1}{\delta}}$$

Q2 Rademacher Complexity

(a) Let $\bar{\mathcal{F}} := \{ f - Pf \mid f \in \mathcal{F} \}$.

WTS $\frac{1}{2} R_n(\bar{\mathcal{F}}) \le \mathbb{E} \| P_n - P \|_{\mathcal{F}}$

$$\frac{1}{2} \mathbb{E}_{X,\varepsilon} \sup_{f \in \bar{\mathcal{F}}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|$$

$$= \frac{1}{2} \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - \mathbb{E}_{Y \sim P} f(Y) \right) \right|$$

$$\le \frac{1}{2} \mathbb{E}_{X,Y,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - f(Y_i) \right) \right| \quad \text{Jensen's inequality}$$

$$\le \frac{1}{2} \mathbb{E}_{X,Y} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(Y_i) \right| \quad \text{symmetrization}$$

$$= \frac{1}{2} \mathbb{E}_{X,Y} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f + \mathbb{E} f - f(Y_i) \right|$$

$$\le \frac{1}{2} \mathbb{E}_{X,Y} \sup_{f \in \mathcal{F}} \left(\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f \right| + \left| \frac{1}{n} \sum_{i=1}^{n} f(Y_i) - \mathbb{E} f \right| \right)$$

$$\le \mathbb{E}_X \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f \right| \quad \text{due to } X, Y \text{ iid}$$

$$= \mathbb{E}_X \| P_n - P \|_{\mathcal{F}}$$

(b) WTS: $R_n(\bar{\mathcal{F}}) \ge R_n(\mathcal{F}) - \frac{1}{\sqrt{n}} \|P\|_{\mathcal{F}}$

$$\mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - \mathbb{E} f(X_i) \right) \right| \ge \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| -$$

$$- \frac{1}{\sqrt{n}} \mathbb{E}_X \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f \right|$$

By expanding $R_n(\mathcal{F})$,

$$R_n(\mathcal{F}) = \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|$$

$$= \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - \mathbb{E} f + \mathbb{E} f \right) \right|$$

$$\leq \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left(\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - \mathbb{E} f \right) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mathbb{E} f \right| \right)$$

$$= \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - \mathbb{E}(f) \right) \right| + \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mathbb{E} f \right|$$

$$= R_n(\bar{\mathcal{F}}) + \mathbb{E}_{\varepsilon} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \right| \cdot \sup_{f \in \mathcal{F}} |\mathbb{E} f|$$

Since $\varepsilon_i \sim \mathrm{unif}(\{\pm 1\})$,

$$\mathbb{E}_{\varepsilon} \left| \sum_{i=1}^{n} \varepsilon_i \right| = \mathbb{E}_{\varepsilon} \sqrt{\left(\sum_{i=1}^{n} \varepsilon_i \right)^2} = \mathbb{E}_{\varepsilon} \sqrt{n + 2 \sum_{1 \leq i < j \leq n} \varepsilon_i \varepsilon_j} \overset{(*)}{\leq} \sqrt{n + \mathbb{E}_{\varepsilon} \sum_{1 \leq i < j \leq n} \varepsilon_i \varepsilon_j} = \sqrt{n}$$

$(*)$ follows from Jensen's inequality

Hence $\quad R_n(\mathcal{F}) \leq R_n(\bar{\mathcal{F}}) + \dfrac{\sup_{f \in \mathcal{F}} |\mathbb{E} f|}{\sqrt{n}} = R_n(\bar{\mathcal{F}}) + \dfrac{1}{\sqrt{n}} \|P\|_{\mathcal{F}}$

(c) From (a) and (b)

$$\mathbb{E}_X \|P_n - P\|_{\mathcal{F}} \geq \frac{1}{2} R_n(\mathcal{F}) - \frac{1}{2\sqrt{n}} \|P\|_{\mathcal{F}}$$

WTS (sufficient to show)

$$\|P_n - P\|_{\mathcal{F}} \geq \mathbb{E}_X \|P_n - P\|_{\mathcal{F}} - \delta \text{ wp at least } 1 - e^{-n\delta^2/2b^2}$$

Now, if we write $g = \|P_n - P\|_{\mathcal{F}}$, we have

$$\mathbb{P}[g - \mathbb{E}g > -\delta] = \mathbb{P}[\mathbb{E}g - g \leq \delta] = 1 - \mathbb{P}[\mathbb{E}g - g > \delta]$$

Note g has bounded difference property with parameter $L = \frac{2b}{n}$ (shown in class)

Hence $\mathbb{P}[\mathbb{E}g - g > \delta] \leq \exp\left(\frac{-n\delta^2}{2b^2}\right)$

This is what we wanted to show.

Q3.

First we show $\mathbb{E}_X \sup_{f \in \mathcal{F}} \pm (P_n f - Pf) \leq 2\tilde{R}_n(\mathcal{F})$

$$\mathbb{E}_X \sup_{f\in\mathcal{F}} \pm (P_n f - Pf) = \mathbb{E}_X \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^n \pm \left(f(X_i) - \mathbb{E}_X f(X_i) \right) \right]$$

$$= \mathbb{E}_X \sup_{f\in\mathcal{F}} \mathbb{E}_Y \left[\frac{1}{n}\sum_{i=1}^n \pm \left(f(X_i) - f(Y_i) \right) \right]$$

because $\mathbb{E}_\theta \sup_\theta h(\theta) \geq \sup_\theta \mathbb{E}_\theta h(\theta)$

$$\leq \mathbb{E}_{X,Y} \left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \pm \left(f(X_i) - f(Y_i) \right) \right]$$

symmetrization, $\varepsilon_i, -\varepsilon_i$ same distribution

and $\varepsilon_i (f(X_i) - f(Y_i)) \overset{d}{=} (f(X_i) - f(Y_i))$

$$= \mathbb{E}_{X,Y,\varepsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(f(X_i) - f(Y_i) \right)$$

$$\leq \mathbb{E}_{X,Y,\varepsilon} \left(\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \right) + \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_i) \right) \right)$$

$$= \mathbb{E}_{X,\varepsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) + \mathbb{E}_{Y,\varepsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_i)$$

$$= 2\tilde{R}_n(\mathcal{F}).$$

So $\mathbb{E}_X \sup_{f\in\mathcal{F}} P_n f - Pf \leq 2\tilde{R}_n(\mathcal{F})$ and $\mathbb{E}_X \sup_{f\in\mathcal{F}} -P_n f + Pf \leq 2\hat{R}_n(\mathcal{F})$

Now we show the second part

Consider $\tilde{g}(X_1, \dots X_n) = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \bar{f}(X_i) = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i) - E[f(X_i)]$

Define $X^{(i)}$ as the same vector as X, but index i is changed.

Take $f \in \mathcal{F}$ s.t. $\frac{1}{n} \sum_{i=1}^{n} \bar{f}(X_i) \geq \tilde{g}(X) - \varepsilon$, $\varepsilon > 0$

$$\tilde{g}(X) - \tilde{g}(X^{(i)}) \leq \frac{1}{n} \sum_{i=1}^{n} \bar{f}(X_i) + \varepsilon - \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^{n} \bar{f}(X_i^{(i)}) + \frac{1}{n} \bar{f}(X_1^{(i)}) \right)$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \bar{f}(X_i) + \varepsilon - \left(\frac{1}{n} \sum_{i=2}^{n} \bar{f}(X_i^{(i)}) + \frac{1}{n} \bar{f}(X_1^{(i)}) \right)$$

$$= \frac{1}{n} \left(f(X_1) - f(X_1^{(i)}) \right) + \varepsilon \leq \frac{2b}{n} + \varepsilon$$

Similarly, taking $f \in \mathcal{F}$ s.t. $\tilde{h}(X) - \varepsilon \leq \frac{1}{n} \sum_{i=1}^{n} -\bar{f}(X_i)$, we have

$$\tilde{h}(X) - \tilde{h}(X^{(i)}) \leq \frac{1}{n} \sum_{i=1}^{n} -\bar{f}(X_i) + \varepsilon - \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} -\bar{f}(X_i^{(i)})$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} -\bar{f}(X_i) + \varepsilon - \frac{1}{n} \sum_{i=1}^{n} -\bar{f}(X_i^{(i)})$$

$$= \frac{1}{n} \sum_{i=1}^{n} -\bar{f}(X_i) + \varepsilon - \frac{1}{n} \left(\sum_{i=2}^{n} -\bar{f}(X_i^{(i)}) + \bar{f}(X_1^{(i)}) \right)$$

$$= \frac{1}{n} \left(\bar{f}(X_1^{(i)}) - \bar{f}(X_1) \right) + \varepsilon \leq \frac{2b}{n} + \varepsilon$$

Taking $\varepsilon \downarrow 0$, $\tilde{g}$ and $\tilde{h}$ both satisfy the bounded differences property with parameter $\frac{2b}{n}$.

By the bounded differences inequality, with $\mathbb{E}[\tilde{g}(X)] = \mathbb{E}[\tilde{h}(X)] = 0$,

$$\mathbb{P}\left[\tilde{g}(X) > \mathbb{E}[\tilde{g}(X)] + \delta\right] = \mathbb{P}\left[\tilde{g}(X) > \delta\right] \leq \exp\left(-\frac{\delta^2 n}{2b^2}\right)$$

$$\mathbb{P}\left[\tilde{h}(X) > \mathbb{E}[\tilde{h}(X)] + \delta\right] = \mathbb{P}\left[\tilde{h}(X) \geq \delta\right] \leq \exp\left(-\frac{\delta^2 n}{2b^2}\right)$$

Hence $\sup_{f\in\mathcal{F}} (P_n f - Pf) \leq \delta$ with probability $\geq 1 - e^{-\delta^2 n/2b^2}$

and $\sup_{f\in\mathcal{F}} (Pf - P_n f) \leq \delta$ with probability $\geq 1 - e^{-\delta^2 n/2b^2}$

We write: let $m(f) := \sup_{f\in\mathcal{F}} P_n f - Pf$, $\quad m^-(f) = \sup_{f\in\mathcal{F}} Pf - P_n f$.

$$m(f) = \underbrace{m(f) - \mathbb{E}_X m(f)}_{\leq \delta \text{ wp at least } 1-\exp\left(\frac{-\delta^2 n}{2b^2}\right)} + \underbrace{\mathbb{E}_X m(f)}_{\leq 2\hat{R}_n(\mathcal{F})}$$

Similarly,

$$m^-(f) = \underbrace{m^-(f) - \mathbb{E}_X m^-(f)}_{\leq \delta \text{ wp at least } 1-\exp\left(\frac{-\delta^2 n}{2b^2}\right)} + \underbrace{\mathbb{E}_X m^-(f)}_{\leq 2\hat{R}_n(\mathcal{F})}$$

Hence $m(f) \leq 2\hat{R}_n(\mathcal{F}) + \delta$ w prob at least $1 - e^{-\delta^2 n/2b^2}$
$m^-(f) \leq 2\hat{R}_n(\mathcal{F}) + \delta$ w prob at least $1 - e^{-\delta^2 n/2b^2}$

Now $\sup_{f \in \mathcal{F}} |Pf - P_n f| = \max\left(m(f), m^-(f) \right)$

$$\Rightarrow \mathbb{P}\left(\sup_{f \in \mathcal{F}} |P_n f - Pf| \geq 2\hat{R}_n(\mathcal{F}) + \delta \right)$$

$$= \mathbb{P}\left(\{ m(f) > 2\hat{R}_n(\mathcal{F}) + \delta \} \cup \{ m^-(f) \quad 2\hat{R}_n(\mathcal{F}) + \delta \} \right)$$

$$\leq \mathbb{P}\left(m(f) > 2\hat{R}_n(\mathcal{F}) + \delta \right) + \mathbb{P}\left(m^-(f) > 2\hat{R}_n(\mathcal{F}) + \delta \right)$$

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

Hence $\|P_n f - Pf\|_{\mathcal{F}} \leq 2\hat{R}_n(\mathcal{F}) + \delta$ w.p.

at least $1 - 2\exp\left(\frac{-n\delta^2}{2b^2} \right)$.

Q4. Binary Regression

(a) $\mathcal{F}_{sgn} = \{ f_\theta(x) = \text{sgn}(\langle x, \theta \rangle) \mid \|\theta\| = 1, \theta \in \mathbb{R}^d \}$

Suppose $X = X_1, \dots X_n \in \mathbb{R}^d$ linearly independent

WTS: $R(\mathcal{F}_{sgn}, X) = 1$

$$R(\mathcal{F}_{sgn}, X) = \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}_{sgn}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right|$$

Note $\frac{1}{n} \sum_{i=1}^n \varepsilon_i \text{sgn}(\langle X_i, \theta \rangle) \le \frac{1}{n} \sum_{i=1}^n |\varepsilon_i| = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 = 1$

$$\Rightarrow \sup_{\substack{\theta \in \mathbb{R}^d \\ \|\theta\| = 1}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \text{sgn}(\langle X_i, \theta \rangle) \le 1$$

Since $X_i \perp\!\!\!\perp X_j \ \forall i \neq j$, and $\text{span}\{X_i\}_{i=1}^n = \mathbb{R}^n$, $\exists \theta \in \mathbb{R}^d$ s.t.

$\text{sgn}(\langle X_i, \theta \rangle) = \varepsilon_i \ \forall i \in [n]$. (take $\theta = \frac{\sum_{i=1}^n X_i \varepsilon_i}{\| \sum_{i=1}^n X_i \varepsilon_i \|_2} \in \mathbb{R}^d$, then $\|\theta\|_2 = 1$ and

$$\langle X_i, \theta \rangle = \left\langle X_i, \frac{\sum_{i=1}^n X_i \varepsilon_i}{\| \sum_{i=1}^n X_i \varepsilon_i \|_2} \right\rangle = \frac{\|X_i\|_2^2 \varepsilon_i}{\| \sum_{i=1}^n X_i \varepsilon_i \|_2} \Rightarrow \text{sgn}(\langle X_i, \theta \rangle) = \text{sgn}(\varepsilon_i)$$

$$\Rightarrow \sup_{\substack{\theta \in \mathbb{R}^d \\ \|\theta\| = 1}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \text{sgn}(\langle X_i, \theta \rangle) = 1, \text{ i.e.}$$

$$\sup_{f \in \mathcal{F}_{sgn}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| = \sup_{\theta \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| = \frac{1}{n} \cdot n = 1$$

and hence $R(\mathcal{F}_{sgn}, X) = 1$.

This is a problem because it implies overfitting.
For $d \geq n$, and x_i's linearly independent, we showed $R_n(\mathcal{F}) = 1$.
So for the boolean loss function, we can chose the joint distribution of (X, Y) and $\mathcal{F}$ such that $R_n(l \circ \mathcal{F}) = O(1)$ which means $\mathcal{F}$ a fixed proportion of the points and $\|P_n - P\|_{\mathcal{F}}$ won't decrease as $n \to \infty$ (still having $d \geq n$).

This is problematic in the context of binary regression, eg with $X_i \sim N(0, \Sigma)$ and $l(z, y) = \frac{1}{2}(1 - zy) = \mathbb{1}\{y \neq z\}$
because we will likely encounter overfitting as the function will perfectly fit random noise in a high-dimensional setting with independent covariates.

(b) WTS if $h : \mathbb{R} \to \mathbb{R}$ is γ-Lipschitz, $\mathcal{F}$ any function class, $h \circ \mathcal{F} = \{ h \circ f \mid f \in \mathcal{F} \}$, then $R_n(h \circ \mathcal{F}) \leq \gamma R_n(\mathcal{F})$

Consider ϕ a 1-Lipschitz function, $T \subseteq \mathbb{R}^2$ bounded.
Take $t^1, t^2 \in T$.

$$\begin{aligned} t_1^1 + \phi(t_2^1) + t_1^2 + \phi(t_2^2) &\leq t_1^1 + t_1^2 + |t_2^1 - t_2^2| \\ &\leq \max\{ t_1^1 + t_1^2 + t_1^2 - t_2^2, t_1^1 - t_1^2 + t_2^1 + t_2^2 \} \\ &\leq \sup_{t \in T} (t_1 + t_2) + \sup_{t \in T} (t_1 + t_2) \end{aligned}$$

Hence $\sup_{t^1, t^2 \in T} t_1^1 + \phi(t_2^1) + t_1^2 - \phi(t_2^2) \leq \sup_{t \in T} (t_1 + t_2) + \sup_{t \in T} (t_1 - t_2)$

i.e. $\sup_{t \in T} (t_1 + \phi(t_2)) + \sup_{t \in T} (t_1 - \phi(t_2)) \leq \sup_{t \in T} (t_1 + t_2) + \sup_{t \in T} (t_1 - t_2)$

Now let us consider $\gamma = 1$ and show by induction on $0 \leq j \leq n$,

$$\hat{R}_n(h \circ f) \leq \mathbb{E}_{X, Y, \varepsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n-j} \varepsilon_i h(f(X_i), Y_i) + \sum_{k=n-j+1}^{n} \varepsilon_k \cdot f(X_k)$$

Base case $n = 0$: Nothing to prove

$1 \leq j \leq n$: By induction hypothesis,

$$\hat{R}_n(h \circ \mathcal{F}) \leq \mathbb{E}_{X, Y, \varepsilon} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left(\sum_{i=1}^{n-j+1} \varepsilon_i h(f(X_i), Y_i) + \sum_{k=n-j+2}^{n} \varepsilon_k f(X_k) \right) \right]$$

let us define $t^d_{X,\varepsilon} = \sum_{i=1}^{n-j} \varepsilon_i h(f(X_i), Y_i) + \sum_{k=n-j+1}^{n} \varepsilon_k f(X_k)$.

$t^d_{X,\varepsilon} \subseteq T$, T bounded. then

$$\tilde{R}_n(h \circ \mathcal{F}) \leq \frac{1}{2n} \mathbb{E}_{X,Y} \mathbb{E}_{\varepsilon / \{\varepsilon_{n-j+1}\}} \Big[\sup_{f \in \mathcal{F}} t^d_{X,\varepsilon} + h(f(X_{n-j+1}), Y_{n-j+1})$$

all ε's except $n-j+1$

$$+ \sup_{f \in \mathcal{F}} (t^d_{X,\varepsilon} - h(f(X_{n-j+1}), Y_{n-j+1})) \Big)$$

By the hint, with $T = \{ t^d_{X,\varepsilon} \mid f \in \mathcal{F}\} \times \{ f(X_{n-j+1})\} \mid f \in \mathcal{F}\} \subseteq \mathbb{R}^2$ bounded.

$$\leq \frac{1}{2n} \mathbb{E}_{X,Y} \mathbb{E}_{\varepsilon \setminus \{\varepsilon_{n-j+1}\}} \Big[\sup_{f \in \mathcal{F}} (t^d_{X,\varepsilon} + f(X_{n-j+1})) + \sup_{f \in \mathcal{F}} (t^d_{X,\varepsilon} - f(X_{n-j+1})) \Big]$$

$$= \frac{1}{n} \mathbb{E}_{X,Y} \mathbb{E}_{\varepsilon} \sup_{f \in \mathcal{F}} \Big[t^d_{X,\varepsilon} + \varepsilon_j f(X_{n-j+1}) \Big]$$

$$= \mathbb{E}_{X,Y,\varepsilon} \sup_{f \in \mathcal{F}} \Big[\frac{1}{n} \sum_{i=1}^{n-j} \varepsilon_i h(f(X_i), Y_i) + \sum_{k=n-j+1}^{n} \varepsilon_k f(X_k) \Big]$$

Which is what we wanted to prove. Now with $j = n$,

$$\tilde{R}_n(h \circ \mathcal{F}) \leq \mathbb{E}_{X,Y,\varepsilon} \sup_{f \in \mathcal{F}} \Big[\frac{1}{n} \sum_{k=1}^{n} \varepsilon_k f(X_k) \Big] = \tilde{R}_n(\mathcal{F}).$$

This shows for h 1-Lipschitz in its first argument, $\hat{R}_n(h \circ F) \leq \tilde{R}_n(\mathcal{F})$.
Now let h be γ-Lipschitz, $\gamma \geq 1$ in its first argument.

Take $\tilde{\phi}_h(f(X)) = \frac{1}{\gamma}\phi(f(X))$

where $\phi_h(f(X)) = h(f(X), Y)$ with h γ-Lipschitz in first argument

Note $\tilde{\phi}_h(f(X))$ is 1-Lipschitz, and applying previous argument, $\forall n \in \mathbb{N}$

$$\mathbb{E}_{X,Y,\varepsilon} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \phi(f(X_i)) = \gamma\, \mathbb{E}_{X,Y,\varepsilon} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \tilde{\phi}(f(X_i))$$

$$\leq \gamma\, \mathbb{E}_{X,Y,\varepsilon} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)$$

Hence $\hat{R}_n(h \circ F) \leq \gamma \tilde{R}_n(\mathcal{F})$.

This is what we wanted to show

(c) Suppose $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{Y}$, $\ell((x,y)) \in [-b,b]$, ℓ is γ-Lipschitz in z defined by $\ell_\theta((x,y)) = \ell(f_\theta(x), y)$

Consider $\mathcal{L} := \{ \ell(f_\theta(x), y) \mid f_\theta \in \mathcal{F} \}$.

WTS: $\| L_n - L \|_{\mathcal{L}} \leq 2\gamma R_n(\mathcal{F})$, with

$L_n = \mathbb{P}_n \ell_\theta(X,Y)$, $L = \mathbb{E} L_\theta(X,Y)$

From theorem proved in lecture, since $\mathcal{L}$ is class of b-bounded functions, $\forall \delta \geq 0$, $n \geq 1$, we have $\| L_n - L \|_{\mathcal{F}} \leq 2 R_n(\mathcal{L}) + \delta$.

with probability at least $1 - \exp\left(\frac{-\delta^2 n}{2b^2} \right)$

Note by Question 3, we know $\| L_n - L \|_{\mathcal{F}} \leq 2 \tilde{R}_n(\mathcal{L}) + \delta$.

From above, with $\ell \circ \mathcal{F} := \{ \ell(f_\theta(X), Y) : f_\theta \in \mathcal{F} \} = \mathcal{L}$,

we have $\tilde{R}_n(\mathcal{L}) \leq \gamma \cdot \tilde{R}_n(\mathcal{F})$.

Hence $\| L_n - L \|_{\mathcal{L}} \leq 2 \cdot \gamma \cdot \tilde{R}_n(\mathcal{F})$ with probability $\geq 1 - \exp\left(\frac{-\partial^2 n}{2\delta^2} \right)$

(d) $\ell_c(z,y) = \max\left(0, \min\left(1, \frac{1}{2} - \frac{1}{2}cyz\right)\right)$. Let $|y| \leq 1$

- If $y=0$, then $\ell_c(z,y) = \frac{1}{2}$, ℓ_c is trivially $\frac{c}{2}$-Lipschitz ($c \geq 0$ is assumed) since it is 0-Lipschitz.

- If $y \neq 0$, we consider

Let $z_1, z_2 \in \mathbb{R}$.

$$|\ell_c(z_1,y) - \ell_c(z_2,y)| = \begin{cases} \left|\min\left(1, \frac{1}{2} - \frac{1}{2}cyz_2\right)\right| & \text{if } (1) \\ \left|\min\left(1, \frac{1}{2} - \frac{1}{2}cyz_1\right) - \min\left(1, \frac{1}{2} - \frac{1}{2}cyz_2\right)\right| & \text{if } (2) \\ \left|\min\left(1, \frac{1}{2} - \frac{1}{2}cyz_1\right)\right| & \text{if } (3) \\ 0 & \text{if } (4) \end{cases}$$

(1) $0 \geq \min\{1, \frac{1}{2} - \frac{1}{2}cyz_1\}$ and $0 \leq \min(1, \frac{1}{2} - \frac{1}{2}cyz_2)$
(2) $0 \geq \min\{1, \frac{1}{2} - \frac{1}{2}cyz_1\}$ and $0 \geq \min(1, \frac{1}{2} - \frac{1}{2}cyz_2)$
(3) $0 \leq \min\{1, \frac{1}{2} - \frac{1}{2}cyz_1\}$ and $0 \geq \min\{1, \frac{1}{2} - \frac{1}{2}cyz_2)$
(4) $0 \leq \min\{1, \frac{1}{2} - \frac{1}{2}cyz_1\}$ and $0 \leq \min\{1, \frac{1}{2} - \frac{1}{2}cyz_2\}$

- Case 1: $y > 0$ $\quad \frac{1}{2} - \frac{1}{2}cyz \leq 1 \Leftrightarrow z \geq \frac{-1}{cy}$

and $\frac{1}{2} - \frac{1}{2}cyz \geq 0 \Leftrightarrow z \leq \frac{1}{cy}$

Then

$$\ell_c(z,y) = \begin{cases} 1 & \text{if } z \leq -\frac{1}{cy} \\ \frac{1}{2} - \frac{1}{2}cyz & \text{if } z \in \left]\frac{-1}{cy}, \frac{1}{cy}\right[\\ 0 & \text{if } z \geq \frac{1}{cy} \end{cases}$$

if $z \notin \left]\frac{-1}{cy}, \frac{1}{cy}\right]$, $\ell_c(z,y)$ is constant and hence trivially Lipschitz $\forall c > 0$. If $z \in \left]\frac{-1}{cy}, \frac{1}{cy}\right[$, then
$\left|\left(\frac{1}{2} - \frac{1}{2}cyz_1\right) - \left(\frac{1}{2} - \frac{1}{2}cyz_2\right)\right| = \left|\frac{1}{2}cy(z_2 - z_1)\right| \leq \frac{c}{2}|z_2 - z_1|$.

Case 2: If $y < 0$, then $\frac{1}{2} - \frac{1}{2} czy \le 1 \Leftrightarrow z \le \frac{-1}{cy}$

$\frac{1}{2} - \frac{1}{2} cyz \geq 0 \Leftrightarrow z \geq \frac{1}{cy}$.

Here we have $\frac{1}{cy} < \frac{-1}{cy}$. and

$$\ell_c(z,y) = \begin{cases} 1 & \text{if } z > \frac{-1}{cy} \\ \frac{1}{2} - \frac{1}{2} cyz & \text{if } z \in [\frac{1}{cy}, \frac{-1}{cy}] \\ 0 & \text{if } z < \frac{1}{cy}. \end{cases}$$

Similarly as above, $\ell_c(z,y)$ is constant and hence 0-Lipschitz if $z \notin [\frac{1}{cy}, \frac{-1}{cy}]$, and otherwise

$$\left|\left(\frac{1}{2} - \frac{1}{2} cyz_1\right) - \left(\frac{1}{2} - \frac{1}{2} cyz_2\right)\right| \le \left|\frac{1}{2} cy (z_2 - z_1)\right| \le \frac{c}{2} |z_1 - z_2|.$$

So for $|y| \le 1$, $\ell_c(z,y)$ is $\frac{c}{2}$-Lipschitz.

(e) $\mathcal{F}_{lin} \{ f_\theta(x) = \langle x, \theta \rangle \mid \theta \in \mathbb{R}^d, \|\theta\| = 1 \}$

WTS: $\hat{R}_n(\mathcal{F}_{lin}) \underset{(1)}{=} \mathbb{E}_{X,\varepsilon} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\| \underset{(2)}{\le} \sqrt{\frac{\mathbb{E}[\|X_i\|^2]}{n}}$

$$\hat{R}_n(\mathcal{F}_{lin}) = \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}_{lin}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i)$$

$$= \mathbb{E}_{X,\varepsilon} \sup_{\substack{\theta \in \mathbb{R}^d \\ \|\theta\|=1}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle X_i, \theta \rangle$$

$$= \frac{1}{n} \mathbb{E}_{X,\varepsilon} \sup_{\|\theta\| \leq 1} \langle X^T \varepsilon, \theta \rangle$$

$$\leq \frac{1}{n} \left(\mathbb{E}_{X,\varepsilon} [\|X^T \varepsilon\|_2^2] \right)^{1/2} \qquad \text{Cauchy-Schwartz, Jensen's}$$

$$= \frac{1}{n} \sqrt{\sum_{i,j=1}^{n} \mathbb{E}[\langle X_i, X_j \rangle \varepsilon_i \varepsilon_j]}$$

$$= \frac{1}{n} \sqrt{\sum_{i=1}^{n} \mathbb{E}[\|X_i\|^2 \varepsilon_i^2] + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathbb{E}[\langle X_i, X_j \rangle \varepsilon_j \mathbb{E}[\varepsilon_i]]}$$

$$= \sqrt{\frac{\mathbb{E}[\|X_i\|^2]}{n}}$$

(f). Using part (c), $\forall n \geq 1$, $\delta \geq 0$, (L_c is bounded by 1) if $\|X_i\| \leq 1$, then we have for $|Y_i| \leq 1 \ \forall Y_i$

$$\|L_c - L\|_{\mathcal{F}_{lin}} \leq c\tilde{R}(\mathcal{F}_{lin}) + \delta \quad \text{w.p. at least } 1 - \exp\left\{-\frac{\delta^2 n}{2}\right\}$$

$$\leq \frac{c}{\sqrt{n}} + \delta \leq c + \delta.$$

This is due to $\|X_i\| \leq 1$ a.s. $\Rightarrow \sqrt{\frac{1}{n}\mathbb{E}\|X_i\|^2} \leq \frac{1}{\sqrt{n}}$

We want $N \in \mathbb{N}$ s.t. $\forall n \geq N$, $\|L_n - L\|_{\mathcal{F}_{lin}} \leq \varepsilon$ with probability at least $1 - e^{-\delta^2 n/2} \geq 1 - \delta'$, for some $\delta' > 0$

From the above, it suffices for $\delta^2 n \geq 2\log(1/\delta')$ and $\frac{c}{\sqrt{n}} + \delta \leq \varepsilon$.

i.e. we select $\delta(n)$ s.t. $\delta(n) \to 0$ and $\delta^2 n \to \infty$.
Take $\delta = n^{-1/4}$. This is enough to get uniform convergence in probability $\|L - L_n\|_{\mathcal{F}_{lin}} \xrightarrow{P} 0$.

Now, we can bound excess risk for fixed X, Y by
$L(\hat{\theta}_n) - L(\theta^*) = |L(\hat{\theta}_n) - L_n(\hat{\theta}_n)| + |L_n(\hat{\theta}_n) - L_n(\theta^*)| + |L_n(\theta^*) - L(\theta^*)|$
$\leq 2\|L - L_n\|_{\mathcal{F}_{lin}}$.

To bound this by ε, we pick $\|L - L_n\|_{\mathcal{F}_{lin}} \leq \varepsilon/2$. To optimize the rate of convergence for the given probability δ', take

$$\delta = \sqrt{\frac{2\log(1/\delta')}{\sqrt{n}}}, \text{ and observe}$$

$$\|L - L_n\|_{\mathcal{F}_{lin}} \leq \frac{c + \sqrt{2\log(1/\delta)}}{\sqrt{n}} \underset{(*)}{\leq} \varepsilon/2$$

when $(*)$ is true for $n \geq 4\left(\frac{c + \sqrt{2\log(1/\delta')}}{\varepsilon}\right)^2$

When δ' is the δ used in the probability of deviation from mean of sample loss, and δ is in the bound on deviation from sample loss.

g) WTS $\underset{c}{\text{minimize}}\ \mathbb{P}_n \ell_c(z,y)$

$\left(\Leftrightarrow \underset{c}{\text{minimize}}\ \underset{c}{\min}\ \max\left(0, \min\left(1, \frac{1}{2}-\frac{1}{2}cyz\right)\right)\right)$

We have $s(z) = \text{sgn}(z)\mathbb{1}[|z| \geq 1] + z\mathbb{1}[|z| \leq 1]$ and let $\text{sgn}_c(z) = s(cy)$ be a "smooth classifier".

Note $\ell(\text{sgn}_c(z), y) = \frac{1}{2}\left(1-\text{sgn}_c(z)y\right)$

$$= \frac{1}{2}\left(1-s(cyz)\right)$$

$$= \max\left(0, \min\left(\frac{1}{2}(1-cyz), 1\right)\right)$$

$$= \ell_c(z,y)$$

This gives us the wanted mapping. Since $X_i \sim \mathcal{N}(0, \frac{I_d}{d})$, and $\|\theta\|=1$, we have (cf d)

$\langle X_i, \theta\rangle \sim \mathcal{N}(0, \frac{1}{d})$. To determine the size of c, we

we would like that most $\langle X_i, \theta\rangle$ are s.t. $|\langle X_i, \theta\rangle| \geq \frac{1}{c}$. The mean of $|\langle X_i, \theta\rangle|$ is $\sqrt{\frac{2}{\pi d}}$

Hence to classify most points (proba $\geq \frac{1}{c}$), we need $c \geq \sqrt{\frac{\pi d}{2}}$, because, as $\left(\mathbb{1}\{|\langle x_i, \theta\rangle| \geq 1/c\}\right)_{n \geq 1}$ a series of bounded variables, the concentration inequalities will give us that at least $\frac{1}{2}$ of the points will be classified whp.

Q5. Chaining Apetizer

(a) $\exp(s A_n(T)) = \exp(s \mathbb{E}_a \sup_{t \in T} \langle a, t \rangle)$

$$\le \mathbb{E}_a \exp(s \cdot \sup_{t \in T} \langle a, t \rangle) \quad \text{convexity}$$

$$\le |T| \sup_{t \in T} \mathbb{E}_a \exp(s \langle a, t \rangle) \quad |T| < \infty, \text{ union bound}$$

$$\le |T| \sup_{t \in T} \prod_{i=1}^{n} \mathbb{E}_{a_i} \exp(s a_i t_i) \quad a_i \text{ iid}$$

$$\le |T| \sup_{t \in T} \prod_{i=1}^{n} \exp\left(\frac{\sigma^2 t_i^2 s^2}{2}\right) \quad a_i \text{ subgaussian \& mean-zero}$$

$$= |T| \sup_{t \in T} \exp\left(\frac{\sigma^2 s^2}{2} \|t\|_2^2\right)$$

$$= |T| \exp\left(\frac{\sigma^2 s^2}{2} \sup_{t \in T} \|t\|_2^2\right)$$

This holds $\forall s \in \mathbb{R}_+$, so for $s \in \mathbb{R}_+$

$$\exp(sA_n(T)) \le |T| \exp\left(\frac{\sigma^2 s^2}{2} \sup_{t \in T} \|t\|_2^2\right)$$

$$\Leftrightarrow \quad A_n(T) \le \frac{1}{s}\left(\log |T| + \frac{s^2 \sigma^2}{2} \sup_{t \in T} \|t\|_2^2\right)$$

choose s such that $\quad c\sigma \sup_{t \in T} \|t\| \sqrt{\log |T|} \ge \frac{1}{s}\left(\log |T| + \frac{\sigma^2 s^2}{2} \sup_{t \in T} \|t\|_2^2\right)$

$$\Leftrightarrow \quad c\sigma \sup_{t \in T} \|t\|_2 \sqrt{\log |T|} \ge \frac{1}{s} \log |T| + \frac{\sigma^2 s}{2} \sup_{t \in T} \|t\|_2^2$$

Take $s = \frac{2}{\sigma \sup_{t\in T} \|t\|_2} \sqrt{\log |T|}$, then for $c \geq \frac{3}{2}$, we obtain the result.

(b) $\forall\, t \in T$, $\exists\, t' \in N_\varepsilon$ s.t. $\|t - t'\| \leq \tilde{\varepsilon}$. (we use $\hat{\varepsilon}$ for notation)

$$R_n(T) = \mathbb{E}_\varepsilon \sup_{t\in T} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i t_i$$

$$\leq \mathbb{E}_\varepsilon \sup_{t'\in N} \sup_{\|\delta\| \leq \tilde{\varepsilon}} \frac{1}{n} \langle \underline{\varepsilon}, t' + \delta \rangle,$$

$$\leq \frac{1}{n} \mathbb{E}_\varepsilon \left[\sup_{t'\in N} \langle \varepsilon, t' \rangle + \sup_{\|\delta\| \leq \tilde{\varepsilon}} \langle \varepsilon, \delta \rangle \right]$$

$$\leq R_n(N) + \frac{1}{n} \mathbb{E}_\varepsilon \sup_{\|\delta\| \leq \tilde{\varepsilon}} \langle \varepsilon, \delta \rangle$$

$$\mathbb{E}_\varepsilon \sup_{\|\delta\| \leq \tilde{\varepsilon}} \frac{1}{n} \langle \varepsilon, \delta \rangle = \mathbb{E}_\varepsilon\, \tilde{\varepsilon} \sup_{\|\delta\|_2 \leq 1} \frac{1}{n} \langle \varepsilon, \delta \rangle$$

$$\leq \frac{1}{n} \mathbb{E}_\varepsilon \sup_{\|\delta\|_2 \leq 1} \tilde{\varepsilon}\, \|\varepsilon\|_2 \cdot \|\delta\|_2 \leq \frac{1}{n} \mathbb{E}_\varepsilon\, \tilde{\varepsilon}\, \|\varepsilon\|_2 = \hat{\varepsilon} \sqrt{n}$$

using the fact that $\mathbb{E}\, \hat{\varepsilon}\, \|\varepsilon\|_2 = \mathbb{E}\, \tilde{\varepsilon} \sqrt{n}$

Hence $R_n(T) \leq R_n(N) + \tilde{\varepsilon}\sqrt{n}$

$$\leq \frac{3}{2} \sup_{t\in N} \|t\| \cdot \sqrt{\log |N|} + \varepsilon\sqrt{n}$$

$$\leq \frac{3}{2} \sup_{t\in T} \|t\| \cdot \sqrt{\log |N|} + \varepsilon\sqrt{n}$$

(c) We have $\sigma_{max}(B) = \|B\|_{op} = \sup_{\substack{\|x\|=1 \\ \|y\|=1}} y^T B x = \sup_{\substack{\|x\|\le 1 \\ \|y\|\le 1}} y^T B x$

We write $\|B\|_{op} = \sup_{x,y \in S^{d-1}} \sum_{i,j=1}^{d} B_{ij} y_i x_i$

Now consider $t_{ij} = x_i y_i$, and consider $\tilde{B}, \tilde{t} \in \mathbb{R}^{d^2}$ s.t.

$$\begin{cases} B_{ij} = \tilde{B}_{i+(d-1)j} \\ t_{ij} = \tilde{t}_{i+(d-1)j} \end{cases} \quad \forall i,j \in [d].$$

Then $\|x\| \le 1$, $\|y\| \le 1$

$$\Rightarrow \|\tilde{t}\|_2^2 = \sum_{i=1}^{d^2} |t_{ij}|^2$$

$$= \sum_{i,j=1}^{d^2} |x_i y_i|^2$$

$$= \sum_{i=1}^{d} |x_i|^2 \sum_{j=1}^{d} |y_j|^2$$

$$= \|x\|_2^2 \cdot \|y\|_2^2 \le 1$$

Hence $\|B\|_{op} \le \sup_{\tilde{t}} \{\tilde{B}, \tilde{t}\}$

with $T = \{ t \in B_2^{d^2}, \text{ s.t. } \exists\, x, y \in S^{d-1} \text{ s.t. } t_{i+(d-1)j} = y_i x_i \}$

each B_{ij} is $\text{unif}(\{\pm 1\})$ hence a Rademacher RV. To build an ε-net for T, it suffices to build separate $\varepsilon/2$ nets

N_1, N_2 of B_2^d and look at the vectors with entries mapped to the entries of xy^T $\forall x \in N_1, y \in N_2$.

Indeed, if $t \in T$, $\exists x \in N_1, y \in N_2$, $\|\delta_1\|_2 \leq \varepsilon/4$, $\|\delta_2\| \leq \varepsilon/4$ with $\delta_1, \delta_2 \in \mathbb{R}^d$ s.t.

$$t_{i+(d-1)j} = (x_i + \delta_{1j})(y_i + \delta_{2j}),$$

and if we look at the vector $z \in B_2^{d^2}$, which entries are mapped to the ones in xy^T

$$\|t - z\|_2^2 = \sum_{i,j=1}^{d} \left((x_i + \delta_{1i})(y_j + \delta_{2j}) - x_i x_j\right)^2$$

$$= \sum_{i,j=1}^{d} \left(x_i \delta_{2j} + y_j \delta_{1i} + \delta_{1i}\delta_{2j}\right)^2$$

$$= \|x\|_2^2 \|\delta_2\|_2^2 + \|y\|_2^2 \|\delta_1\|_2^2 + 2\langle x, \delta_1\rangle \langle y, \delta_2\rangle + 2\langle y, \delta_2\rangle \|\delta_1\|_2^2 + 2\langle x, \delta_1\rangle \|\delta_2\|_2^2 + \|\delta_1\|^2 \|\delta_2\|^2$$

$$\leq \frac{\varepsilon^2}{16} + \frac{\varepsilon^2}{16} + \frac{2\varepsilon^2}{16} + \frac{2\varepsilon^3}{64} + \frac{2\varepsilon^3}{64} + \frac{\varepsilon^4}{128}$$

$$\leq \frac{\varepsilon^2}{4} + \frac{\varepsilon^3}{2} + \frac{\varepsilon^4}{16} \leq \varepsilon^2$$

We can build $\frac{\varepsilon}{4}$ nets of B_2^d of size $\left(1 + \frac{8}{\varepsilon}\right)^d$ & $\exists$ an ε-net of T of size $\left(1 + \frac{8}{\varepsilon}\right)^{2d}$. Now applying (8), we get

$$\mathbb{E}_{B} \|B\|_{op} = \mathbb{E}_{\tilde{B}} \sup_{\tilde{t}\in T} \langle \tilde{B}, \tilde{t} \rangle$$

$$\leq c \cdot \underbrace{\sup_{\tilde{t}\in T} \|\tilde{t}\|}_{\leq 1 \text{ shown above}} \sqrt{\log\left(1+\frac{2}{\varepsilon}\right)^{2d}} + \varepsilon\sqrt{d^2}$$

$$= c\sqrt{d\log\left(1+{}^{8}/_{\varepsilon}\right)^2} + \varepsilon d$$

By taking $\varepsilon = \frac{1}{\sqrt{d}}$, $\mathbb{E}_{B} \|B\|_{op} \leq c\sqrt{2d\log\left(1+64d+16\sqrt{d}\right)} + \sqrt{d}$

$$= O(\sqrt{d\log d})$$

(d) The bound from (b) is not tight. The lossy term is $\varepsilon\sqrt{n}$. We obtained it by applying Cauchy-Schwartz to $\sum_{i=1}^{n} \varepsilon_i \delta_i$ to bound it by $\|\varepsilon\|_2 \|\delta\|_2$. But the bound is achieved only when δ is colinear to ε. However, taking the sup over all $\delta \in B_2^d(\varepsilon)$ was already lossy, as our covering might include include points out of our set T (in (b)). Then we might not have δ colinear to ε in the set which creates looseness in the Cauchy-Schwartz inequality.