ETH Zurich Guarantees for Machine Learning 23 **(Due: 9/11/23)**

Homework #2: Generalization bounds

Name: XXX, *Student ID:* XXX *Students discussed with:* None

Problem 1: Data-dependent generalization bound for hard-margin SVM

a) **Solution:** Given that:

$$R^0(f) = \mathbb{P}(Yf(X) \leq 0) \leq R_n^\gamma(f) + \frac{2DB}{\gamma\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)}{n}}$$

Take $B = \|w^*\|_2$ and $\gamma = 1$. Consider the functional class $\mathcal{F}_B = \{f(x) = \langle w, x \rangle : \|w^*\|_2 \leq B\}$. We also know that there exists w^* with the smallest l2-norm such that $\mathbb{P}(y\langle w, x \rangle \geq 1) = 1$, which means that there at least exists an $f \in \mathcal{F}_B$, such that:

$$R_n^\gamma(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i f(x_i) \leq 1} = 0$$

f made no mistake on the dataset (correctly classify dataset). This will futher leads to:

$$R^0(f) = \mathbb{P}(Yf(X) \leq 0) \leq \frac{2D\|w^*\|_2}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)}{n}}, \quad \text{where } f = f_{SVM}$$

b) **Solution:** It is equivalent to prove that:

$$\mathbb{P}\left(\bigcap_{k=k(f)}^{\infty}\bigcap_{f\in\mathcal{F}_k} E_{k,f}\right) \geq 1-\delta$$

We now show why it holds: Firstly, we note that if f is not in $\mathcal{F}_{k(f)}$, it is not possible to appear in $\mathcal{F}_{k(f)-1}$ and in lower k since $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ....$ Therefore the index should start with the smallest index $k(f)$ s.t. f is contained in $\mathcal{F}_k$. And we also infer from the nested sequence that $E_{k(f),f} \subset E_{k(f)+1,f} \subset E_{k(f)+2,f} \subset ...$, so:

$$\bigcap_{k=k(f)}^{\infty}\bigcap_{f\in\mathcal{F}_k} E_{k,f} = E_{k(f),f}, \quad f \in \mathcal{F}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{k=k(f)}^{\infty}\bigcap_{f\in\mathcal{F}_k} E_{k,f}\right) &= \mathbb{P}(E_{k(f),f}) \\
&= \mathbb{P}\left(R(f) - R_n(f) \leq c\sqrt{\frac{\log(1/\delta_{k(f)})}{n}} + 2\mathcal{R}_n(\mathcal{F}_{k(f)})\right) \\
&\geq 1-\delta \quad \text{need to be shown}
\end{aligned}$$

Since

$$\mathbb{P}\left(\bigcap_{f\in\mathcal{F}_k} E_{k,f}\right) \geq 1-\delta_k$$

which we could instead use the complement:

$$\mathbb{P}\left(\bigcup_{f\in\mathcal{F}_k} E_{k,f}^c\right) \leq \delta_k$$

Using union bound we get:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k=k(f)}^{\infty}\bigcup_{f\in\mathcal{F}_k} E_{k,f}^c\right) &\leq \sum_{k=k(f)}^{\infty}\mathbb{P}\left(\bigcup_{f\in\mathcal{F}_k} E_{k,f}^c\right) \\
&= \sum_{k=k(f)}^{\infty}\delta_k \\
&\leq \sum_{k=1}^{\infty}\delta_k \leq \delta
\end{aligned}$$

Again using the complement we got:

$$\mathbb{P}\left(\bigcap_{k=k(f)}^{\infty}\bigcap_{f\in\mathcal{F}_k} E_{k,f}\right) \geq 1-\delta$$

which means that

$$\mathbb{P}\left(R(f) - R_n(f) \leq c\sqrt{\frac{\log(1/\delta_{k(f)})}{n}} + 2\mathcal{R}_n(\mathcal{F}_{k(f)})\right) \geq 1-\delta$$

c) **Solution:** From the lecture we know that:

$$\mathcal{R}_n(\mathcal{F}_B) \le \sup_{x\in X} \hat{\mathcal{R}_n}(\mathcal{F}_B(x)) \le \frac{\max \|w\|_2 \max \|x\|_2}{\sqrt{n}}$$

Here we did substitution since we've been given $\|w\|_2 = \|w_{SVM}\|_2$ and $\|x\|_2 \le D$ (Assumption A). So

$$\mathcal{R}_n(\mathcal{F}_B) \le \frac{\|w_{SVM}\|_2 D}{\sqrt{n}}$$

Let $B_k = 2^k$, $\mathcal{F}_k = \{f(x) = \langle w, x\rangle : \|w_{SVM}\|_2 \le B_k\}$ and $\delta_k = \frac{\delta}{2k^2}$. We can observe that it satisfies the condition that $\sum_{k=1}^{\infty} \delta_k = \frac{\pi^2}{12}\delta \approx 0.82\delta \le \delta$, also $k \ge \lceil \log \|w_{SVM}\|_2 \rceil$. Just take $k = \lceil \log \|w_{SVM}\|_2 \rceil$, then

$$\begin{aligned}\frac{1}{\delta_{k(f)}} &= \frac{1}{2}\frac{(2k(f))^2}{\delta} \\ &\le \frac{1}{8}\frac{(4\log \|w_{SVM}\|_2)^2}{\delta}\end{aligned}$$

We already knew that:

$$\mathbb{P}\left(R(f) - R_n(f) \le c\sqrt{\frac{\log(1/\delta_{k(f)})}{n}} + 2\mathcal{R}_n(\mathcal{F}_{k(f)})\right) \ge 1-\delta$$

Substitute $1/\delta_{k(f)}$ and $\mathcal{F}_{k(f)}$:

$$\begin{aligned}&\mathbb{P}\left(R(f) - R_n(f) \le c\sqrt{\frac{\log(1/\delta_{k(f)})}{n}} + 2\mathcal{R}_n(\mathcal{F}_{k(f)})\right) \ge 1-\delta \\ \to\ &\mathbb{P}\left(R(f) - R_n(f) \le c\sqrt{\frac{\log(\frac{1}{8}\frac{(4\log\|w_{SVM}\|_2)^2}{\delta})}{n}} + 2\frac{\|w_{SVM}\|_2 D}{\sqrt{n}}\right) \ge 1-\delta \\ \to\ &\mathbb{P}\left(R(f) - R_n(f) \le c\sqrt{\frac{\log(\frac{4\log\|w_{SVM}\|_2}{\delta})}{n}} + \frac{2eD\|w_{SVM}\|_2}{\sqrt{n}}\right) \ge 1-\delta\end{aligned}$$

Again for any $f \in \mathcal{F}_k$, we have empirical loss $R_n(f) = 0$ given Assumption B, so that

$$\mathbb{P}(Yf_{SVM}(X) \le 0) = R^0(f) \le c\sqrt{\frac{\log(\frac{4\log\|w_{SVM}\|_2}{\delta})}{n}} + \frac{2eD\|w_{SVM}\|_2}{\sqrt{n}}$$

with probability at least $1-\delta$

Problem 2: Rates for smooth functions

a) **Solution:**

b) **Solution:** Remember the prediction error bound from the lecture 8 (MW Theorem 13.5):

$$\mathbb{P}(\|\hat{f} - f^*\|_n^2 \geq 16t\delta_n^2) \leq e^{-\frac{nt\delta_n^2}{2\sigma^2}}$$

So we need to show that $\delta_n^2 = c\left(\frac{\sigma^2}{n}\right)^{\frac{4}{5}}$ which will lead to the conclusion:

$$\mathbb{P}\left(\|\hat{f} - f^*\|_n^2 \geq c_0\left(\frac{\sigma^2}{n}\right)^{\frac{4}{5}}\right) \leq c_1 e^{-\frac{nt\left(\frac{n}{\sigma^2}\right)^{-\frac{4}{5}}}{2\sigma^2}}$$

$$\leftrightarrow \mathbb{P}\left(\|\hat{f} - f^*\|_n^2 \geq c_0\left(\frac{\sigma^2}{n}\right)^{\frac{4}{5}}\right) \leq c_1 e^{-c_2\left(\frac{n}{\sigma^2}\right)^{\frac{1}{5}}}$$

We have $\log\mathcal{N}(\epsilon; \mathcal{F}_{\alpha,\gamma}, \|\cdot\|_\infty) = O(\frac{1}{\epsilon}^{\frac{1}{\alpha+\gamma}})$, which is related to Dudley's integral:

$$\begin{aligned}
\frac{16}{\sqrt{n}}\int_{\frac{\delta^2}{4\sigma}}^{\delta}\sqrt{\log\mathcal{N}(\epsilon; \mathcal{F}_{\alpha,\gamma}, \|\cdot\|_\infty)}d\epsilon &\leq \frac{16}{\sqrt{n}}\int_0^{\delta}\sqrt{\log\mathcal{N}(\epsilon; \mathcal{F}_{\alpha,\gamma}, \|\cdot\|_\infty)}d\epsilon \\
&\leq \frac{C}{\sqrt{n}}\int_0^{\delta}\sqrt{\frac{1}{\epsilon}^{\frac{1}{\alpha+\gamma}}}d\epsilon \\
&\leq \frac{C}{\sqrt{n}}\int_0^{\delta}\epsilon^{-\frac{1}{2(\alpha+\gamma)}}d\epsilon \\
&\leq \frac{C}{\sqrt{n}}\delta^{1-\frac{1}{2(\alpha+\gamma)}}
\end{aligned}$$

We need to let the right hand side less equal than $\frac{\delta^2}{4\sigma}$ to satisfy critical inequality:

$$\begin{aligned}
&\frac{C}{\sqrt{n}}\delta^{1-\frac{1}{2(\alpha+\gamma)}} \leq \frac{\delta^2}{4\sigma} \\
\rightarrow &\frac{C\sigma}{\sqrt{n}} \leq \delta^{1+\frac{1}{2(\alpha+\gamma)}} \\
\rightarrow &C(\frac{\sigma}{\sqrt{n}}^{\frac{2(\alpha+\gamma)}{1+2(\alpha+\gamma)}}) \leq \delta \\
\rightarrow &C(\frac{\sigma^2}{n}^{\frac{2(\alpha+\gamma)}{1+2(\alpha+\gamma)}}) \leq \delta^2
\end{aligned}$$

If we let $\alpha = 1, \gamma = 1$, a possible value for δ_n^2 is equal to $\delta_n^2 = c\left(\frac{\sigma^2}{n}\right)^{\frac{4}{5}}$, which ends the proof. We need to show that $\mathcal{F}_2 \in \mathcal{F}_{1,1}$. Since the second derivative $\|f^{(2)}\|$ is bounded for each entry ($\|\cdot\|_\infty$ indicates a max), then its first derivative is a Lipschitz function. proof is easy since we could just use the definition of second derivative.

c) **Solution:** Similarly, we want to show that $\delta_n^2 = c\left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$. We have been given: $\hat{\mu_j} = j^{-2\alpha}$. We could use corollary 13.18 from MW:

$$\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n} \min\{\delta^2, \hat{\mu_j}\}} \leq \frac{R}{4\sigma}\delta^2$$

Assume after k rounds, $\hat{\mu_{k+1}}$ starts to smaller than δ^2. Lets expand the LHS of the inequality by inserting $\hat{\mu_j}$:

$$\begin{aligned}\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n} \min\{\delta^2, \hat{\mu_j}\}} &= \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + \sum_{j=k+1}^{n} j^{-2\alpha}} \\ &\leq \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + \int_{k+1}^{\infty} j^{-2\alpha} dj} \\ &\leq \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + O((k+1)^{1-2\alpha})}\end{aligned}$$

we know that $\delta^2 \geq (k+1)^{-2\alpha}$, so the term inside the square root is dominated by $k\delta^2$:

$$\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n} \min\{\delta^2, \hat{\mu_j}\}} \leq \sqrt{\frac{2}{n}}\sqrt{O(k\delta^2)}$$

On the other side,

$$\begin{aligned} k^{-2\alpha} &\geq \delta^2 \\ \to k^{2\alpha} &\leq \delta^{-2} \\ \to k &\leq \delta^{-\frac{1}{\alpha}} \\ \to k\delta^2 &\leq \delta^{2-\frac{1}{\alpha}}\end{aligned}$$

Therefore,

$$\begin{aligned}\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n} \min\{\delta^2, \hat{\mu_j}\}} &\leq \sqrt{\frac{2}{n}}\sqrt{O(\delta^{2-\frac{1}{\alpha}})} \\ &\leq O\left(\sqrt{\frac{\delta^{2-\frac{1}{\alpha}}}{n}}\right) \\ &\leq \frac{R}{4\sigma}\delta^2 \quad \text{mentioned before}\end{aligned}$$

So we need to find such δ_n s.t.

$$\begin{aligned} C\left(\sqrt{\frac{\delta_n^{2-\frac{1}{\alpha}}}{n}}\right) &\leq \frac{R}{4\sigma}\delta_n^2 \\ \to C\frac{\sigma^2}{n} &\leq \delta_n^{2+\frac{1}{\alpha}} \\ \to C\left(\frac{\sigma^2}{n}\right)^{\frac{\alpha}{2\alpha+1}} &\leq \delta_n \\ \to C\left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}} &\leq \delta_n^2\end{aligned}$$

Then we could take $\delta_n^2 = c\left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$. Proof finish.

Problem 3: Sparse linear functions

a) **Solution:** From the lecture, we know that the localized Gaussian complexity can be written as:

$$\tilde{\mathcal{G}_n}(\mathcal{F}_{B,s}(x_1^n)) = \frac{1}{n}\mathbb{E}\sup_{\hat{\Delta}\in\mathcal{F}^*}\sum_{i=1}^n w_i\hat{\Delta}$$

, where $\hat{\Delta} \in \mathcal{F}^* := \{f(\cdot) = \langle\theta, x_i\rangle\}$. Therefore:

$$\begin{aligned}
\tilde{\mathcal{G}_n}(\mathcal{F}_{B,s}(x_1^n)) &= \frac{1}{n}\mathbb{E}\sup_{\theta}\sum_{i=1}^n w_i\langle\theta, x_i\rangle \\
&= \frac{1}{n}\mathbb{E}\sup_{\theta}\sum_{i=1}^n\langle\theta, w_i x_i\rangle \\
&= \frac{1}{n}\mathbb{E}\sup_{\theta}\langle\theta, X^T w\rangle \quad x_i \text{ is the row of } X \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta}\langle\theta, \frac{X^T w}{\sqrt{n}}\rangle
\end{aligned}$$

We know that $\|\theta\|_0 \leq s$. We could rewrite θ with elementwise multiplication $\theta = \theta \odot \mathbb{1}_S$, where we know that the maximum number of non-zero elements in θ is less or equal than s, which means that $|S| \leq s$. $\mathbb{1}_S$ has the same size of θ. Therefore:

$$\begin{aligned}
\tilde{\mathcal{G}_n}(\mathcal{F}_{B,s}(x_1^n)) &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta}\langle\theta, \frac{X^T w}{\sqrt{n}}\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\leq s}\langle\theta\odot\mathbb{1}_S, \frac{X^T w}{\sqrt{n}}\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\leq s}\langle\theta, \mathbb{1}_S^T\odot\frac{X^T w}{\sqrt{n}}\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\leq s}\langle\theta, \frac{(X\odot\mathbb{1}_S)^T w}{\sqrt{n}}\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\leq s}\langle\theta, \frac{X_S^T w}{\sqrt{n}}\rangle
\end{aligned}$$

Recall Cauchy–Schwarz inequality that $\langle a, b\rangle \leq \|a\|\|b\|$, therefore:

$$\begin{aligned}
\tilde{\mathcal{G}_n}(\mathcal{F}_{B,s}(x_1^n)) &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\leq s}\langle\theta, \frac{X_S^T w}{\sqrt{n}}\rangle \\
&\leq \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\leq s}\|\theta\|_2\|\frac{X_S^T w}{\sqrt{n}}\|_2 \\
&\leq \frac{B}{\sqrt{n}}\mathbb{E}\sup_{|S|\leq s}\|\frac{X_S^T w}{\sqrt{n}}\|_2 \\
&= B\mathbb{E}_w\max_{|S|=s}\|\frac{X_S^T w}{n}\|_2
\end{aligned}$$

b) **Solution:** We first recall the Theorem 2.26 from MW: let $X_1, ..., X_n$ be some random gaussian variables. If $f(X_i)$ is L-Lipschitz w.r.t Euclidean norm, then $f(X_i) - \mathbb{E}(f(X_i))$ is sub-gaussian with parameter at most L, and further

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq t] \leq e^{-\frac{t^2}{2L^2}}$$

We observe that w_S belongs to Gaussian family since we define such a w_i so that $w_i \sim \mathcal{N}(0,1)$, therefore, we can regard w_S as a linear combination of w_i. Therefore what we left in this question is to solve two sub-questions:

1. Euclidean norm $\|\cdot\|_2$ is C-Lipschitz
2. $\mathbb{E}[\|w_S\|_2] \leq C\sqrt{s}$, so that $\mathbb{P}[\|w_S\|_2 - C\sqrt{s} \geq t] \leq \mathbb{P}[\|w_S\|_2 - \mathbb{E}[\|w_S\|_2] \geq t]$

w_S is defined as: $w_S = \frac{1}{\sqrt{n}} X_S^T w$, where the largest eigenvalue of $\frac{X_S^T X_S}{n}$ is bounded by C^2.

$$\begin{aligned}
\|w_S\|_2^2 = w_S^T w_S &= \frac{1}{n} w^T X_S X_S^T w \\
&= w^T \frac{X_S X_S^T}{n} w \\
&\leq w^T C^2 w \\
&= C^2 \|w\|_2^2 \\
\rightarrow \|w_S\|_2 &\leq C\|w\|_2
\end{aligned}$$

Therefore we have proved that $\|w_S\|_2$ is C-Lipschitz. Then we bound $\mathbb{E}[\|w_S\|_2]$ to let it less or equal than $C\sqrt{s}$:

$$\begin{aligned}
\mathbb{E}[\|w_S\|_2] &\leq \mathbb{E}[\sqrt{\frac{1}{n} w^T X_S X_S^T w}] \\
&= \sqrt{\mathbb{E}[\frac{1}{n} w^T X_S X_S^T w]} \quad \text{Jensen's inequality} \\
&= \sqrt{\mathbb{E}[\text{tr}\{\frac{1}{n} w^T X_S X_S^T w\}]} \quad w^T X_S X_S^T w \text{ is a scalar} \\
&= \sqrt{\mathbb{E}[\text{tr}\{\frac{1}{n} X_S X_S^T w w^T\}]} \\
&= \sqrt{\text{tr}\{\frac{1}{n} X_S X_S^T \mathbb{E}[w w^T]\}}
\end{aligned}$$

Since the diagonal of $\mathbb{E}[ww^T]$ is $\mathbb{E}[w_i^2] = \mathbb{E}[w_i]^2 + \mathbb{V}[w_i] = 1$ where other entries are zero since w_i is somehow independent. Therefore, multiplying a $\mathbb{E}[ww^T]$ is equivalent to multiplying an identity matrix therefore we could remove that. So

$$\begin{aligned}
\mathbb{E}[\|w_S\|_2] &= \sqrt{\text{tr}\{\frac{1}{n} X_S X_S^T \mathbb{E}[w w^T]\}} \\
&= \sqrt{\text{tr}\{\frac{1}{n} X_S X_S^T\}} \\
&= \sqrt{\sum_{i=1}^{s} \lambda_i \left(\frac{1}{n} X_S X_S^T\right)} \quad \text{sum of eigenvalues equals to the trace} \\
&\leq \sqrt{s \cdot \lambda_{max} \left(\frac{1}{n} X_S X_S^T\right)} = \sqrt{s \cdot C^2} = C\sqrt{s}
\end{aligned}$$

Hence, following the MW 2.26 we achieved the conclusion that:

$$\mathbb{P}[f(\|w_S\|_2) - C\sqrt{s} \geq \delta] \leq e^{-\frac{\delta^2}{2C^2}}$$

c) **Solution:** From a) we know that

$$\tilde{\mathcal{G}_n}(\mathcal{F}_{B,s}(x_1^n)) \leq \frac{B}{\sqrt{n}}\mathbb{E}_w \max_{|S|=s} \|\frac{X_S^T w}{\sqrt{n}}\|_2$$

We want to use the fact that:

$$\mathbb{E}\sup_{\theta} X_s \leq \sqrt{2\sigma^2 \log N}$$

if X_s is σ-subgaussian, where from b) we know that $\|w_S\|_2 - \mathbb{E}\|w_S\|_2$ is C-subgaussain. Therefore:

$$\begin{aligned}
\tilde{\mathcal{G}_n}(\mathcal{F}_{B,s}(x_1^n)) &\leq \frac{B}{\sqrt{n}}\mathbb{E}_w \max_{|S|=s} \|\frac{X_S^T w}{\sqrt{n}}\|_2 \\
&= \frac{B}{\sqrt{n}}\mathbb{E}_w \max_{|S|=s}\{\|w_S\|_2 - \mathbb{E}[\|w_S\|_2] + \mathbb{E}[\|w_S\|_2]\} \\
&\leq \frac{BC\sqrt{s}}{\sqrt{n}} + \frac{B}{\sqrt{n}}\mathbb{E}_w \max_{|S|=s}\{\|w_S\|_2 - \mathbb{E}[\|w_S\|_2]\} \\
&\leq \frac{BC\sqrt{s}}{\sqrt{n}} + \frac{B}{\sqrt{n}}\sqrt{2C^2 \log N}
\end{aligned}$$

where here N is equal to the number of possible choices for $\|w_S\|_2 - \mathbb{E}[\|w_S\|_2]$, which is equal to the number of possible non-zero indexes hat $\mathbb{1}_S$ could provide. This number N is then equal to $\binom{d}{s}$. We then applied the following bounds for binomial coefficients:

$$\binom{d}{s} \leq \frac{d^s}{s!} \leq \left(\frac{e \cdot d}{s}\right)^s$$

Simple proof here:

$$\begin{aligned}
\binom{d}{s} &= \frac{d \times (d-1) \times ... \times (d-s+1)}{s!} \\
&\leq \frac{d \times d \times ... \times d}{s!} \\
&= \frac{d^s}{s!}
\end{aligned}$$

We remain to prove that: $e^s \geq \frac{s^s}{s!}$. We perform the Taylor expansion to e^s: $e^s = \sum_{i=0}^{\infty} \frac{s^i}{i!}$, then we could easily find that $\frac{s^s}{s!}$ is one of the term when i = s, therefore, $e^s \geq \frac{s^s}{s!}$. Then,

$$\begin{aligned}
\tilde{\mathcal{G}_n}(\mathcal{F}_{B,s}(x_1^n)) &\leq \frac{BC\sqrt{s}}{\sqrt{n}} + \frac{B}{\sqrt{n}}\sqrt{2C^2 \log N} \\
&= \frac{BC\sqrt{s}}{\sqrt{n}} + \frac{B}{\sqrt{n}}\sqrt{2C^2 \log \binom{d}{s}} \\
&\leq \frac{BC\sqrt{s}}{\sqrt{n}} + \frac{B}{\sqrt{n}}\sqrt{2C^2 \log \left(\frac{e \cdot d}{s}\right)^s} \\
&\leq \frac{BC\sqrt{s}}{\sqrt{n}} + BC\sqrt{\frac{2}{n}}\sqrt{s \log \left(\frac{e \cdot d}{s}\right)}
\end{aligned}$$

We use the fact that $\log\left(\frac{e \cdot d}{s}\right) > 1$, so that the Gaussian complexity is mainly dominated by the second term. Using big-O notation, we get:

$$\begin{aligned}
\tilde{\mathcal{G}_n}(\mathcal{F}_{B,s}(x_1^n)) &\leq \frac{BC\sqrt{s}}{\sqrt{n}} + BC\sqrt{\frac{2}{n}}\sqrt{s \log \left(\frac{e \cdot d}{s}\right)} \\
&\leq O\left(BC\sqrt{\frac{2}{n}}\sqrt{s \log \left(\frac{e \cdot d}{s}\right)}\right) \\
&\leq O\left(BC\sqrt{\frac{1}{n}}\sqrt{s \log \left(\frac{e \cdot d}{s}\right)}\right) \quad \text{remove constant achieved the answer}
\end{aligned}$$

d) **Solution:** Again by a) we know that

$$\begin{aligned}\tilde{\mathcal{G}}_n(\mathcal{F}_{B,s}(x_1^n)) &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\le s}\langle\theta,\frac{X_S^T w}{\sqrt{n}}\rangle\\ &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\le s}\langle\frac{X_S\theta}{\sqrt{n}},w\rangle\end{aligned}$$

We want to relate w to w_S such that we could use the property in (b) and (c) to bound Gaussian complexity if $\|w_S\|$ is Lipschitz. The simplest idea is to let w_S be the orthogonal projection of w onto the subspace of X_s, which means:

$$\langle w - w_S, X_S\rangle = 0$$

therefore,

$$\begin{aligned}\tilde{\mathcal{G}}_n(\tilde{\mathcal{F}}_{B,s}(x_1^n)) &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\le s}\langle\frac{X_S\theta}{\sqrt{n}},w\rangle\\ &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\le s}\langle\frac{X_S\theta}{\sqrt{n}},w_S\rangle\\ &\le \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\le s}\|\frac{X_S\theta}{\sqrt{n}}\|_2\|w_S\|_2 \quad \text{cauchy schwarz}\\ &\le \frac{B}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S|\le s}\|w_S\|_2\end{aligned}$$

It is obvious that $\|w_s\|_2 \le \|w\|_2$ according to the definition of orthogonal projection, or equivalently $\|w_S\|_2 = \|w\|\cos\theta$. Therefore, $\|w_S\|_2$ is then 1-Lipschitz, then

$$\mathbb{E}\sup_{\theta,|S|\le s}\|w_S\|_2 \le \sqrt{2s\log\left(\frac{e\cdot d}{s}\right)}$$

if using the conclusion in c) when C = 1. Therefore,

$$\begin{aligned}\tilde{\mathcal{G}}_n(\tilde{\mathcal{F}}_{B,s}(x_1^n)) &\le \frac{B}{\sqrt{n}}\sqrt{2s\log\left(\frac{e\cdot d}{s}\right)}\\ &\le O\left(B\sqrt{\frac{s}{n}\log\left(\frac{e\cdot d}{s}\right)}\right)\end{aligned}$$

Problem 4: Bonus: Classification error bounds for hard margin SVM

a) **Solution:** we know that $\hat{\theta} = [r, \gamma\tilde{\theta}]$, and $x = [yr, \tilde{x}]$ so

$$\begin{aligned}\mathbb{P}[y\hat{\theta}^T x] &= \mathbb{P}[yrx_1 + y\gamma\tilde{\theta}^T x_{2:d}] \\ &= \mathbb{P}[y^2r^2 + y\gamma\tilde{\theta}^T x_{2:d}]\end{aligned}$$

Firstly, we note that y can only take two values $\{-1, +1\}$. And $x_{2:d}$ serves standard normal distribution, therefore $y\gamma\tilde{\theta}^T x_{2:d}$ could be regarded as sum of standard normal distributed random variables, which also serves Gaussian distribution, which has a mean of 0 and a variance of $y^2\gamma^2\frac{1*(d-1)}{d-1} = \gamma^2$ (standard deviation σ $= \gamma$). Therefore, $y^2r^2 + y\gamma\tilde{\theta}^T x_{2:d}$ serves $\mathcal{N}(r^2, \gamma^2)$.

We know that $\mathbb{P}(X < a) = \Phi(\frac{a-\mu}{\sigma})$, if a serves $\mathcal{N}(\mu, \sigma)$, where $\Phi(\cdot)$ is the error function. In this case,

$$\mathbb{P}[y\hat{\theta}^T x < 0] = \Phi(\frac{0 - r^2}{\gamma}) = \Phi(-\frac{r^2}{\gamma})$$

Review the error function:

$$\Phi(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$$

where erf function is monotonically increasing in $\mathbb{R}$. When r increases, $\Phi(-\frac{r^2}{\gamma})$ decreases, since $-\frac{r^2}{\gamma}$ decreases.

b) **Solution:** We have defined γ as:

$$\gamma = \max_{\theta \in \mathbb{R}^{d-1}} \min_{(x,y)\in D} y \frac{\langle \hat{\theta}, x_{2:d} \rangle}{\|\hat{\theta}\|_2}$$

, we can use the simpler argument that mentioned: $\|\theta\|_2 = 1$ to rewrite γ:

$$\gamma = \max_{\theta \in \mathbb{R}^{d-1}, \|\theta\|_2 = 1} \min_{(x,y)\in D} y \langle \hat{\theta}, x_{2:d} \rangle$$

We know $\tilde{X} \in \mathbb{R}^{n*(d-1)}$ and $\theta \in \mathbb{R}^{d-1}$, therefore, $\tilde{X}\theta \in \mathbb{R}^n$ Then,

$$\min_{(x,y)\in D} y \langle \hat{\theta}, x_{2:d} \rangle \leq \frac{\|\tilde{X}\theta\|_2}{\sqrt{n}}$$

since $\tilde{X}\theta$ contains n entries of $\langle \hat{\theta}, x_{2:d} \rangle$. It is not hard to observe that: n * min less than at least n non-min (or min) numbers, which leads to the above inequality. Then,

$$\gamma \leq \max_{\theta \in \mathbb{R}^{d-1}, \|\theta\|_2 = 1} \frac{\|\tilde{X}\theta\|_2}{\sqrt{n}}$$

According to the largest singular value definition: $s_{max}(\tilde{X}) = \max_{\|\theta\|_2 = 1} \|\tilde{X}\theta\|_2$, we then achieve:

$$\gamma \leq \frac{s_{max}(\tilde{X})}{\sqrt{n}}$$

c) **Solution: i)** We know that each entry of A is a random standard normal distributed variable, therefore,

$$\begin{aligned}
\mathbb{E}[X_{u,v} - X_{u',v'}] &= 0 \\
\rightarrow \mathbb{E}[|X_{u,v} - X_{u',v'}|^2] &= \mathbb{V}[|X_{u,v} - X_{u',v'}|] \\
&= |\langle u, v\rangle - \langle u', v'\rangle|^2 \\
&= |\langle u - u', v - v'\rangle|^2 \\
&\leq \|u - u'\|_2^2 + \|v - v'\|_2^2
\end{aligned}$$

Likewise, since g, h serves standard normal distribution,

$$\begin{aligned}
\mathbb{E}[|Y_{u,v} - Y_{u',v'}|] &= 0 \\
\rightarrow \mathbb{E}[|Y_{u,v} - Y_{u',v'}|^2] &= \mathbb{V}[|Y_{u,v} - Y_{u',v'}|] \\
&= \mathbb{V}[\langle g, u - u'\rangle + \langle h, v - v'\rangle] \\
&= \mathbb{V}[\langle g, u - u'\rangle] + \mathbb{V}[\langle h, v - v'\rangle] \\
&= \|u - u'\|_2^2 + \|v - v'\|_2^2 \\
&\geq \mathbb{E}[|X_{u,v} - X_{u',v'}|^2]
\end{aligned}$$

c) **Solution: ii)**

$$\begin{aligned}\mathbb{E}[s_{max}(X)] &= \mathbb{E}[\sup_{u\in\mathbb{S}^{d-1},v\in\mathbb{S}^{n-1}} X_{u,v}]\\ &\le \mathbb{E}[\sup_{u\in\mathbb{S}^{d-1},v\in\mathbb{S}^{n-1}} Y_{u,v}]\\ &\le \mathbb{E}[\sup_{u\in\mathbb{S}^{d-1},v\in\mathbb{S}^{n-1}} \langle g,u\rangle + \langle h,v\rangle]\end{aligned}$$

The supremum is achieved when $u = \frac{g}{\|g\|_2}$, where its the unit vector on the direction of g. Therefore,

$$\begin{aligned}\mathbb{E}[s_{max}(X)] &\le \mathbb{E}[\|g\|_2 + \|h\|_2]\\ &= \|g\|_2 + \|h\|_2\\ &\le \sqrt{d} + \sqrt{n}\end{aligned}$$

d) **Solution:**
We need to show that $s_{max}(\cdot)$ is 1-Lipschitz w.r.t the definition of largest singular value:

$$\begin{aligned} |s_{max}(X_1) - s_{max}(X_2)| &= |\max_{\|\theta\|_2=1} \|X_1\theta\|_2 - \max_{\|\theta\|_2=1} \|X_2\theta\|_2| \\ &\leq |\max_{\|\theta\|_2=1} (\|X_1\theta\|_2 - \|X_2\theta\|_2)| \\ &\leq \max_{\|\theta\|_2=1} \|(X_1 - X_2)\theta\|_2 \\ &\leq \|X_1 - X_2\|_{\mathcal{F}}, \quad \text{Cauchy-Swartz} \end{aligned}$$

It is element-wise 1-Lipschitz, therefore, using the theorem MW 2.26, we got:

$$\mathbb{P}[|s_{max}(X) - \mathbb{E}[s_{max}(X)]| \leq t] \geq 1 - 2e^{-t^2/2}$$

Or equivalently,

$$\mathbb{P}[s_{max}(X) - \mathbb{E}[s_{max}(X)] \leq t|] \geq 1 - e^{-t^2/2}$$

for one side.