Collaborations: Matteo Palo

1 Optimality of polynomial Markov

1.a

To commence, let us begin by recalling a fundamental result from [6], which is presented as follows:

Proposition 1 (Markov's inequality)**.** *Given a non-negative random variable X with finite mean, we have*

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \textit{for all } t > 0.$$

Now, if we proceed to treat X as a Bernoulli-distributed random variable with $\mathbb{P}(X = 0) = 1 - p$ and $\mathbb{P}(X = 1) = p$, we can readily ascertain that $\mathbb{E}[X] = (1 - p) \cdot 0 + p \cdot 1 = p$. Furthermore, it holds that $\mathbb{P}(X \geq 1) = p$. Substituting these values into Markov's Inequality for $t = 1$, we arrive at the desired equality.

1.b

We compute

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &= \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{(\lambda X)^i}{i!}\right] \quad \text{(Taylor expansion of } e^{\lambda X}\text{)} \\
&= \sum_{i=0}^{\infty} \mathbb{E}\left[X^i\right] \frac{\lambda^i}{i!} \quad \text{(Linearity of expectation)} \\
&= \sum_{i=0}^{\infty} \mathbb{E}\left[|X|^i\right] \frac{\lambda^i}{i!} \cdot \frac{\delta^i}{\delta^i} \quad (|X| = X \text{ since } X \geq 0) \\
&\geq \sum_{i=0}^{\infty} \left(\inf_{k=0,1,\ldots} \frac{\mathbb{E}\left[|X|^k\right]}{\delta^k}\right) \frac{(\lambda\delta)^i}{i!} \\
&= \left(\inf_{k=0,1,\ldots} \frac{\mathbb{E}|X|^k}{\delta^k}\right) \underbrace{\sum_{i=0}^{\infty} \frac{(\lambda\delta)^i}{i!}}_{=e^{\delta\lambda}}
\end{aligned}$$

and so we have that

$$\inf_{k=0,1,\ldots} \frac{\mathbb{E}\left[|X|^k\right]}{\delta^k} \leq \inf_{\lambda>0} \frac{\mathbb{E}\left[\mathrm{e}^{\lambda X}\right]}{\mathrm{e}^{\lambda\delta}}$$

2 Concentration and kernel density estimation

First recall from lecture 2 the following definition and theorem:

Definition 1 (Bounded difference property). *Define for given $z, z' \in \mathcal{Z}^n$ a new vector $z^{\setminus k}$ with the k-th element from z' and all other from z : ${z_j}^k = \begin{cases} z_j & \text{if } j \neq k \\ z'_k & \text{if } j = k \end{cases}$. We say that $g_n : \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded difference inequality if for each $k = 1, \ldots, n$ it holds that*

$$\left| g_n(z) - g_n\left(z^{\setminus k}\right) \right| \leq \sigma_k \text{ for all } z, z' \in \mathcal{Z}^n$$

Theorem 1 (McDiarmid). *If $g_n : \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded difference condition with $\{\sigma_k\}_{k=1}^n$ and Z is a random vector with n independent entries, then*

$$\mathbb{P}\left(g_n(Z) - \mathbb{E}g_n(Z) \geq t\right) \leq e^{-\frac{2t^2}{\sum_{k=1}^n \sigma_k^2}}$$

Let $f_n(x, \mathbf{X})$ denote the kernel density estimate for the observations $\mathbf{X} = (X_1, \ldots, X_n)$, with f being the density on the real line. We also define $\mathbf{X}^{\setminus k}$ similarly to $\mathbf{X}$ according to Definition 1. Additionally, let $g_n(\mathbf{X}) = \|f_n - f\|_1$. We aim to prove that g_n satisfies the bounded difference property.
Consider X_k and $\tilde{X}_k$ as the k^{th} entries of $\mathbf{X}$ and $\mathbf{X}^{\setminus k}$, respectively. We have the following inequality:

$$\begin{aligned}
\left| g_n(\mathbf{X}) - g(\mathbf{X}^{\setminus k}) \right| &= \left| \|f_n(\cdot, \mathbf{X}) - f\|_1 - \|f_n(\cdot, \mathbf{X}^{\setminus k}) - f\|_1 \right| \\
&\leq \left\| f_n(\cdot, \mathbf{X}) - f_n(\cdot, \mathbf{X}^{\setminus k}) \right\|_1 && \text{(Reverse triangular inequality)} \\
&= \int_{\mathbb{R}} |f_n(t, \mathbf{X}), -f_n(t, \mathbf{X}^{\setminus k})| \, dt \\
&= \frac{1}{nh} \int_{\mathbb{R}} \left| K\left(\frac{t - X_k}{h}\right) - K\left(\frac{t - \tilde{X}_k}{h}\right) \right| dt && \text{(Only differ in one entry)} \\
&\leq \frac{1}{nh} \left[\int_{\mathbb{R}} K\left(\frac{t - X_k}{h}\right) dt + \int_{\mathbb{R}} K\left(\frac{t - \tilde{X}_k}{h}\right) dt \right] && \text{(Triangular inequality)} \\
&= \frac{2}{n} && \text{(Variable change and } \int_{\mathbb{R}} K(t)\, dt = 1)
\end{aligned}$$

Now, applying McDiarmid's inequality, we have:

$$\mathbb{P}\left(g_n(\mathbf{X}) - \mathbb{E}g_n(\mathbf{X}) \geq \delta\right) \leq e^{-\frac{2\delta^2}{\sum_{k=1}^n \sigma_k^2}} = e^{-\frac{n\delta^2}{2}} \leq e^{-\frac{n\delta^2}{18}}$$

This completes the proof.

3 Sub-Gaussian maxima

3.a

To address this problem, we employ two widely used tools: the property that $\{X_i\}_{i=1}^n$ is a sequence of zero-mean random variables, each of which is subgaussian, and Jensen's inequality applied to convex functions, such as the exponential function. Let $\lambda > 0$. The derivation proceeds as follows:

$$\begin{aligned}
e^{\lambda \cdot \mathbb{E}[\max_{i=1,\dots,n} X_i]} &\leq \mathbb{E}\left[e^{\lambda \cdot (\max_{i=1,\dots,n} X_i)}\right] && \text{(Jensen's inequality for convex functions)} \\
&= \mathbb{E}\left[\max_{i=1,\dots,n} e^{\lambda X_i}\right] && \text{(Monotonicity of the exponential function)} \\
&\leq \mathbb{E}\left[\sum_{i=1}^{n} e^{\lambda X_i}\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right] && \text{(Linearity of expectation)} \\
&\leq n e^{\frac{\lambda^2 \sigma^2}{2}} && (X_i \text{ are subgaussian, use Thm 2.6 [6]})
\end{aligned}$$

Taking the natural logarithm of both sides, we obtain:

$$e^{\lambda \cdot \mathbb{E}[\max_{i=1,\dots,n} X_i]} \leq n e^{\frac{\lambda^2 \sigma^2}{2}} \Leftrightarrow \mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] \leq \frac{\log(n)}{\lambda} + \frac{\lambda \sigma^2}{2}$$

By optimizing the right-hand side of the preceding inequality with respect to λ we get:

$$-\frac{\log(n)}{\lambda^2} + \frac{\sigma^2}{2} = 0 \Leftrightarrow \lambda = \frac{\sqrt{2\log(n)}}{\sigma} > 0$$

Substituting this optimal value back into the previous inequality, we obtain:

$$\mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] \leq \sqrt{2\sigma^2 \log(n)}$$

3.b

The second inequality follows directly from the following observation

$$\begin{aligned}
\sqrt{2\sigma^2 \log(2n)} &= \sqrt{2\sigma^2 \log(2) + 2\sigma^2 \log(n)} \\
&\leq \sqrt{2\sigma^2 \log(n) + 2\sigma^2 \log(n)} && \text{(since } n \geq 2\text{)} \\
&= 2\sqrt{\sigma^2 \log(n)}
\end{aligned}$$

Hence, it remains to prove the inequality $\mathbb{E}\left[\max_{i=1,\dots,n} |X_i|\right] \leq \sqrt{2\sigma^2 \log(2n)}$. Notably, we observe that:

$$\mathbb{E}\left[\max_{i=1,\dots,n} |X_i|\right] = \mathbb{E}\left[\underbrace{\max\{X_1, -X_1, X_2, -X_2, \dots, X_n, -X_n\}}_{2n \text{ random variables}}\right]$$

We can readily derive the desired bound by applying the procedure outlined in the previous subquestion (with a sample size of $2n$) since both X_i and $-X_i$ are σ-subgaussian.

4 Sharper tail bounds for bounded variables: Bennett's inequality

4.a

Since X_i has zero-mean, we have that $\sigma_i^2 = \mathrm{Var}(X_i) = \mathbb{E}\left[X_i^2\right]$. Now we compute:

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda X_i}\right] &= \mathbb{E}\left[\sum_{j=0}^{\infty} \frac{(\lambda X_i)^j}{j!}\right] \\
&= 1 + \sum_{j=2}^{\infty} \frac{\lambda^j}{j!} \mathbb{E}\left[X_i^j\right] \\
&\leq 1 + \sum_{j=2}^{\infty} \frac{\lambda^j}{j!} \underbrace{\mathbb{E}\left[X_i^2\right]}_{=\sigma_i^2} b^{j-2} && (|X_i| < b \text{ is bounded}) \\
&= 1 + \frac{\sigma_i^2}{b^2} \sum_{j=2}^{\infty} \frac{(\lambda b)^j}{j!} \\
&= 1 + \frac{\sigma_i^2}{b^2} \left(-1 - \lambda b + \underbrace{\sum_{j=0}^{\infty} \frac{(\lambda b)^j}{j!}}_{=e^{\lambda b}} \right) \cdot \frac{\lambda^2}{\lambda^2} \\
&= 1 + \sigma_i^2 \lambda^2 f(\lambda b) && \left(f(x) = \frac{e^x - 1 - x}{x^2}\right)
\end{aligned}$$

If we now take the natural logarithm on both sides and use the fact that $\log(1+x) \leq x$ for $x \geq 0$ we obtain

$$\log \mathbb{E} e^{\lambda X_i} \leq \sigma_i^2 \lambda^2 f(\lambda b) = \sigma_i^2 \frac{\mathrm{e}^{\lambda b} - 1 - \lambda b}{b^2}$$

4.b

To commence, let us begin by recalling a fundamental result from [6], which is presented as follows:

Proposition 2 (Markov's inequality)**.** *Given a non-negative random variable X with finite mean, we have*

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \textit{for all } t > 0.$$

We can no employ the Chernoff bound method, which was introduced during the initial lecture. For $\lambda \geq 0$

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \delta\right) &= \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq n\delta\right) \\
&= \mathbb{P}\left(e^{\lambda \sum_{i=1}^n X_i} \geq e^{\lambda n \delta}\right) \\
&\leq \frac{\mathbb{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right]}{e^{\lambda n \delta}} && \text{(Markov's inequality, see Prop. 2)} \\
&= \frac{\mathbb{E}\left[\prod_{i=1}^n e^{\lambda X_i}\right]}{e^{\lambda n \delta}} \\
&= \frac{\prod_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right]}{e^{\lambda n \delta}} && (X_i \text{ are i.i.d.}) \\
&\leq \frac{\prod_{i=1}^n e^{\sigma^2\left(\frac{\mathrm{e}^{\lambda b} - 1 - b\lambda}{b^2}\right)}}{e^{\lambda n \delta}} && \text{(Result from previous subquestion)} \\
&= e^{n\sigma^2\left(\frac{\mathrm{e}^{\lambda b} - 1 - b\lambda}{b^2}\right) - \lambda n \delta} && (*)
\end{aligned}$$

Now, we optimize the bound by finding the value of λ that minimizes the right-hand side of $(*)$. We do this by setting the derivative of the expression with respect to λ equal to zero. Solving for λ yield

$$e^{n\sigma^2\left(\frac{\mathrm{e}^{\lambda b} - 1 - b\lambda}{b^2}\right) - \lambda n \delta} \cdot \left(\frac{n\sigma^2}{b^2}\left(be^{\lambda b} - b\right) - n\delta\right) = 0 \Leftrightarrow \lambda = \frac{1}{b}\log\left(\frac{b\delta}{\sigma^2} + 1\right)$$

Inserting into $(*)$ gives

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \delta\right) \leq e^{\frac{n\sigma^2}{b^2}\left(\frac{b\delta}{\sigma^2}-\log\left(\frac{b\delta}{\sigma^2}+1\right)-\frac{b\delta}{\sigma^2}\log\left(\frac{b\delta}{\sigma^2}\right)\right)} = e^{-\frac{n\sigma^2}{b^2}h\left(\frac{b\delta}{\sigma^2}\right)} \quad \text{with } h(t) := (1+t)\log(1+t) - t, t \geq 0$$

4.c

Definition 2 (Bernstein's condition, [6]). *Given a random variable X with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}\left[X^2\right] - \mu^2$, we say that Bernstein's condition with parameter b holds if*

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \leq \frac{1}{2}k!\sigma^2 b^{k-2} \quad \textit{for } k = 2, 3, 4, \ldots.$$

Proposition 3 (Bernstein's inequality, [6]). *For any random variable with zero mean satisfying the Bernstein condition*

$$\mathbb{P}[X \geq t] \leq e^{-\frac{t^2}{2(\sigma^2+bt)}} \quad \textit{for all } t \geq 0.$$

From the previous subquestion, we have established that for a single random variable satisfying the conditions in the problem statement, Bennett's inequality states:

$$\mathbb{P}[X \geq t] \leq e^{-\frac{\sigma^2}{b^2}h\left(\frac{bt}{\sigma^2}\right)} \quad \text{for all } t \geq 0.$$

Now, it remains to prove that:

$$\frac{\sigma^2}{b^2}h\left(\frac{bt}{\sigma^2}\right) \geq \frac{t^2}{2(\sigma^2+bt)} \tag{1}$$

We start by manipulating the right-hand side of Equation 1 to bring it into a form more similar to the left-hand side of the equation:

$$\frac{t^2}{2\left(\sigma^2+bt\right)} = \frac{\sigma^2}{b^2}\frac{\left(\frac{bt}{\sigma^2}\right)^2}{2\left(1+\frac{bt}{\sigma^2}\right)} = \frac{\sigma^2}{b^2}f\left(\frac{bt}{\sigma^2}\right), \quad \text{with } f(x) = \frac{x^2}{2(1+x)}.$$

Hence, it is necessary for us to establish that the function $h(x)$ is greater than or equal to the function $f(x)$ for all $x \geq 0$. One way to do so is as follows: since both $h(x)$ and $f(x)$ are differentiable for $x \geq 0$, we can compute the difference between them and demonstrate that the derivative of this difference is always non-negative for $x \geq 0$. Therefore we have that

$$\frac{d}{dx}(h(x)-f(x)) = \underbrace{\log(1+x)}_{\frac{d}{dx}h(x)} - \underbrace{\frac{x(x+2)}{2(x+1)^2}}_{\frac{d}{dx}f(x)}$$

To establish if this quantity is always bigger than zero we again differentiate and see if the derivative is always non-negative. The calculations are:

$$\frac{d^2}{dx^2}(h(x)-f(x)) = \underbrace{\frac{1}{1+x}}_{\frac{d^2}{dx^2}h(x)} - \underbrace{\frac{1}{(1+x)^3}}_{\frac{d^2}{dx^2}f(x)} \geq 0, \forall x \geq 0$$

Since $h(0) = f(0) = 0, h'(0) = f'(0) = 0$ and $(h'' - f'')(x) \geq 0$ it holds that $(h' - f')(x) \geq 0, \forall x \geq 0$ and so $(h-f)(x) \geq 0, \forall x \geq 0$. We have successfully proven that the Bennett's inequality is at least as tight as Bernstein's inequality.

5 Sharp upper bounds on binomial tails

5.a

Let $\lambda > 0$. To address this problem, we can employ again the Chernoff bound method. First, we consider the probability that the random variable Z_n is less than or equal to δn:

$$\begin{aligned}\mathbb{P}\left(Z_n \leq \delta n\right) &\leq \mathbb{P}\left(e^{-\lambda Z_n} \geq e^{-\lambda \delta n}\right) \\ &\leq \mathbb{E}\left[e^{-\lambda Z_n}\right] \cdot e^{\lambda \delta n} \qquad \text{(Markov's inequality, see 2)}\end{aligned}$$

Then, we calculate the expectation of the random variable $e^{-\lambda Z_n}$, which requires us to use the probability generating function (PGF) of a Bernoulli distribution:

- The PGF of a Bernoulli variable with parameter α is defined as $P(t) = \alpha t + (1-\alpha)$,. To find the PGF of the sum of n i.i.d. Bernoulli variables (i.e., Z_n), we raise the PGF of a single Bernoulli variable to the power of n: $P_{Z_n}(t) = [P(t)]^n = (\alpha t + (1-\alpha))^n$.
- We substitute t with $e^{-\lambda}$ in the PGF and compute the expectation: $\mathbb{E}\left[e^{-\lambda Z_n}\right] = \left(e^{-\lambda}\alpha + (1-\alpha)\right)^n$.

Therefore, we obtain the expression:

$$\mathbb{P}\left(Z_n \leq \delta n\right) \leq e^{\lambda \delta n} \cdot \left(\alpha e^{-\lambda} + (1-\alpha)\right)^n \Leftrightarrow \log \mathbb{P}\left(Z_n \leq \delta n\right) \leq \lambda \delta n + n \log\left(\alpha e^{-\lambda} + (1-\alpha)\right) \tag{2}$$

Now, we optimize the bound by finding the value of λ that minimizes the right-hand side of (2). We do this by setting the derivative of the right-hand side with respect to λ equal to zero. Solving for λ yield we have

$$\delta n - n\frac{\alpha e^{-\lambda}}{\alpha e^{-\lambda} + (1-\alpha)} = 0 \Leftrightarrow \lambda = \log\left(\frac{(1-\delta)\alpha}{\delta(1-\alpha)}\right) > 0$$

By plugging this solution back into (2) we obtain

$$\begin{aligned}\log \mathbb{P}\left(Z_n \leq \delta n\right) &\leq \delta n \cdot \left(-\log\frac{\delta}{\alpha} - \log\frac{1-\alpha}{1-\delta}\right) + n\log\left(-\alpha \cdot \left(\frac{(1-\delta)\alpha}{\delta(1-\alpha)}\right) + (1-\alpha)\right) \\ &= -n\left(\delta \log\frac{\delta}{\alpha} + (1-\delta)\log\frac{1-\delta}{1-\alpha}\right) \\ &= -nD(\delta\|\alpha)\end{aligned}$$

And so $\mathbb{P}\left(Z_n \leq \delta n\right) \leq e^{-nD(\delta\|\alpha)}$.

5.b

We start by introducing the concept of total variation distance, accompanied by two important inequalities.

Definition 3 (Total variation distance, [5])**.** *Consider a measurable space $(\Omega, \mathcal{F})$ and probability measures P and Q defined on $(\Omega, \mathcal{F})$. The total variation distance between P and Q is defined as:*

$$TV(P,Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

Proposition 4 ((first) Pinsker's inequality, [5])**.** *If P and Q are two probability distributions on a measurable space (X, Σ), then*

$$TV(P,Q) \leq \sqrt{\frac{1}{2}D(P\|Q)},$$

where $TV(P,Q)$ and $D(P\|Q)$ are respectively the total variation distance and the Kullback–Leibler divergence between the distributions P and Q.

Proposition 5 (Hoeffding inequality, [1])**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely. Consider the sum of these random variables, $S_n = X_1 + \cdots + X_n$. Then, for all $t > 0$*

$$\mathrm{P}\left(S_n - \mathrm{E}\left[S_n\right] \leq -t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n \left(b_i - a_i\right)^2}\right)$$

Proposition 6. *Let $\alpha \in (0, \frac{1}{2}], \delta \in (0, \alpha)$. Then it holds that $D(\delta\|\alpha) > 2(\alpha - \delta)^2$, where $D(\delta\|\alpha)$ is defined as in the problem assignment.*

Proof. The total variation distance between two Bernoulli distributions with parameters δ and α is given by:

$$TV(\delta, \alpha) = \frac{1}{2}(|\alpha - \delta| + |1 - \alpha - (1 - \delta)|) = |\alpha - \delta|$$

By Proposition 4 we have that

$$TV(\delta, \alpha) \leq \sqrt{D(\delta\|\alpha)/2} \Leftrightarrow D(\delta\|\alpha) \geq 2(\alpha - \delta)^2$$

However, equality holds only if $\delta = \alpha$ which can't happen in our case since $\alpha \in (0, \frac{1}{2}], \delta \in (0, \alpha)$. Therefore $D(\delta\|\alpha) > 2(\alpha - \delta)^2$. □

We first rewrite the tail probability as follows:

$$\mathbb{P}(Z_n \leq \delta n) = \mathbb{P}(Z_n - \alpha n \leq -(\alpha - \delta)n)$$

Since $\{X_i\}_{i=1}^n$ are i.i.d. sequence of Bernoulli variables with parameter $\alpha \in \left(0, \frac{1}{2}\right]$ and $Z_n = \sum_{i=1}^n X_i$ is a binomial random variable with expectation αn, we can apply Hoeffding inequality (5) with $t = (\alpha - \delta)n$, $a_i = 0, b_i = 1$:

$$\mathbb{P}(Z_n \leq \delta n) \leq e^{-2(\alpha-\delta)^2 n}$$

Using Proposition 6 we therefore have that the bound derived from previous subquestion is strictly better than the Hoeffding bound.

6 Robust estimation of the mean

Following the hint, we first perform an equi-partition of our data set into $\mathcal{O}(\log(\delta^{-1}))$ (for the sake of notation, suppose that we partition the data set into $q \cdot \log(\delta^{-1})$ subsets with $q \geq 1$ and assume that $q \cdot \log(\delta^{-1})$ is an integer; otherwise, we can round it up), each with the same sample size N_i (so that $N_1 = N_2 = \cdots = N_{q\cdot\log(\delta^{-1})}$). A weak estimate, denoted as $\hat{\mu}_i$, is calculated for the true mean μ of the data-set within each of these partitions. In our pursuit of achieving an estimate of the mean for a 1-dimensional random variable with variance σ^2 that possesses an ϵ-level of accuracy, drawn independently from a sample denoted as $X_1, X_2, \ldots, X_n$, we can employ Chebyshev's inequality (see [4]) to establish the following:

$$\mathbb{P}\left(|\hat{\mu}_i - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{N_i \epsilon^2} = \frac{1}{k}$$

Here, we have defined $N_i = k \cdot \frac{\sigma^2}{\epsilon^2} = \mathcal{O}\left(\frac{\sigma^2}{\epsilon^2}\right)$, where $k > 1$ is chosen to be sufficiently large, in line with our objective of minimizing the above probability.
Therefore we have a sample size of $N = q \cdot \log(\delta^{-1}) \cdot N_i = \mathcal{O}\left(\log\left(\delta^{-1}\right) \frac{\sigma^2}{\epsilon^2}\right)$. It remains to prove that the median of the previous constructed weak learners falls indeed with probability $\geq 1 - \delta$ in the interval $[\mu - \epsilon, \mu + \epsilon]$.To achieve this, we can construct a variable γ_i that indicates whether the weak estimate $\hat{\mu}_i$ falls within the specified interval of $[\mu - \epsilon, \mu + \epsilon]$. We can then utilize a concentration inequality to constrain $\mathbb{P}(|\operatorname{median}(\hat{\mu}_1, \ldots, \hat{\mu}_{q\cdot\log(\delta^{-1})}) - \mu| > \epsilon)$. Therefore, we define γ_i as follows:

$$\gamma_i = \begin{cases} 1 & \text{if } \hat{\mu}_i \in [\mu - \epsilon, \mu + \epsilon] \\ 0 & \text{otherwise} \end{cases}$$

Let $T = \sum_{i=1}^{q\cdot\log(\delta^{-1})} \gamma_i$. The probability that the median of the week estimates falls outside the required interval is then bounded by the probability that T is less than half of $q \cdot \log(\delta^{-1})$. It is evident that the equality $\mu_T = \sum_{i=1}^{q\cdot\log\left(\delta^{-1}\right)} p_i = q \cdot \log\left(\delta^{-1}\right) \cdot p$, holds, where $p_i = p = 1 - \frac{1}{k}$ represents the lower bound probability that each weak estimate falls within the interval $[\mu - \epsilon, \mu + \epsilon]$. Consequently, since each $\gamma_i \in [0, 1]$, the variable T is confined within the range $[0, q \cdot \log(\delta^{-1})]$, contingent upon the number of γ_i values equaling 1, we can now use Hoeffding's inequality. Considering the fact that from the previous observations we have that the γ_i are sub-gaussian with $\sigma = \frac{1-0}{2} = \frac{1}{2}$, we compute

$$\begin{aligned}
\mathbb{P}(|\operatorname{median}(\hat{\mu}_1, \ldots, \hat{\mu}_{q\cdot\log(\delta^{-1})}) - \mu| > \epsilon) &\leq \mathbb{P}\left(T < \frac{q \cdot \log(\delta^{-1})}{2}\right) \\
&= \mathbb{P}\left(T - \mu_T < \underbrace{\frac{q \cdot \log(\delta^{-1})}{2} - \mu_T}_{=t}\right) \\
&\leq \exp\left\{-\frac{t^2}{2 \cdot \sum_{i=1}^{q\cdot\log(\delta^{-1})} \frac{1}{2^2}}\right\} \qquad \text{(Hoeffding's inequality[1])} \\
&= \exp\left\{-\frac{2 \cdot t^2}{q \cdot \log(\delta^{-1})}\right\}
\end{aligned}$$

Therefore we are left to prove the following equality for some value of k, q:

$$\delta = \exp\left\{-2 \cdot \frac{\left(\frac{q\cdot\log(\delta^{-1})}{2} - q \cdot \log(\delta^{-1}) \cdot p\right)^2}{q \cdot \log(\delta^{-1})}\right\}$$

$$\frac{1}{2} = q \cdot \left(\frac{1}{2} - (1 - \frac{1}{k})\right)^2$$

which is true for all values of $k = 3, 4, 5, \ldots$ and $q = \frac{1}{2\cdot\left(\frac{1}{k} - \frac{1}{2}\right)^2}$.

[1]Note that the condition $t < 0 \Leftrightarrow k > 2$ must also be satisfied.

7 Best-arm identification

7.a

From the problem assignment sheet we know that at any-time confidence interval, such that for any arm k,

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty}\{|\hat{\mu}_{k,t}-\mu_k|\geq U(t,\delta)\}\right)\leq\delta. \tag{3}$$

Therefore we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &= \mathbb{P}\left(\bigcup_{k=1}^{K}\bigcup_{t=1}^{\infty}\{|\hat{\mu}_{k,t}-\mu_k|>U(t,\delta/K)\}\right)\\
&\leq \sum_{k=1}^{K}\mathbb{P}\left(\bigcup_{t=1}^{\infty}\{|\hat{\mu}_{k,t}-\mu_k|>U(t,\delta/K)\}\right) && \text{(Union bound)}\\
&\leq \sum_{k=1}^{K}\frac{\delta}{K}=\delta && \text{(Equation 3 with } \delta'=\frac{\delta}{K}\text{)}
\end{aligned}$$

7.b

Any arm k with $k\in\{1,\dots,K\}$ will only be removed from S_{t-1} under the condition that there exists at a time-step $t\geq 1$ an arm $i\in S_{t-1}$ such that

$$\hat{\mu}_{i,t}-U(t,\delta/K)>\hat{\mu}_{k,t}+U(t,\delta/K). \tag{4}$$

If we assume that $\mathcal{E}^c=\bigcap_{k=1}^{K}\bigcap_{t=1}^{\infty}\{|\hat{\mu}_{k,t}-\mu_k|\leq U(t,\delta/K)\}$ is satisfied (event that occurs with $\mathbb{P}(\mathcal{E}^c)\geq 1-\delta$ due to the previous subquestion), we know that for any arm $i\in\{1,\dots,K\}$ at any step $t\geq 1$

$$\mu_i+U(t,\delta/K)\geq\hat{\mu}_{i,t}\geq\mu_i-U(t,\delta/K)$$

For the best arm $k^\star$ it holds that $\mu_{k^\star}\geq\mu_i,\forall i\in\{1,\dots,K\}$. Therefore we have

$$\hat{\mu}_{k^\star,t}+U(t,\delta/K)\geq\mu_{k^\star}\geq\mu_i\geq\hat{\mu}_{i,t}-U(t,\delta/K)$$

So, the inequality in Equation 4 is false for all $t\geq 1$, and thus, for any $t\geq 1$, the best arm $k^\star$ is included in the set S_t.

7.c

For the random variable Z_t we have that $\hat{\mu}_{k,t}=\frac{1}{t}\sum_{i=1}^{t}Z_i$. We have to prove that

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty}\left\{\left|\frac{1}{t}\sum_{i=1}^{t}Z_i-\mathbb{E}\left[Z_i\right]\right|\geq\sqrt{\frac{(b-a)^2\log\left(4t^2/\delta\right)}{2t}}\right\}\right)\leq\delta$$

Since the Z_i's are i.i.d bounded random variables with $Z_i\in[a,b]$, they are also subgaussian with parameter $\sigma=\frac{b-a}{2}$. We now compute

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{t=1}^{\infty}\left\{\left|\frac{1}{t}\sum_{i=1}^{t}Z_i-\mathbb{E}\left[Z_i\right]\right|\geq\sqrt{\frac{(b-a)^2\log\left(4t^2/\delta\right)}{2t}}\right\}\right) &\overset{\text{U.B}}{\leq} \sum_{t=1}^{\infty}\mathbb{P}\left(\left|\frac{1}{t}\sum_{i=1}^{t}Z_i-\mathbb{E}\left[Z_i\right]\right|\geq\sqrt{\frac{(b-a)^2\log\left(4t^2/\delta\right)}{2t}}\right)\\
&= \sum_{t=1}^{\infty}\mathbb{P}\left(\left|\sum_{i=1}^{t}Z_i-\mathbb{E}\left[Z_i\right]\right|\geq\underbrace{t\cdot\sqrt{\frac{(b-a)^2\log\left(4t^2/\delta\right)}{2t}}}_{\geq 0}\right)\\
&\leq \sum_{t=1}^{\infty}2e^{-\frac{t^2\cdot(b-a)^2\log\left(4t^2/\delta\right)}{2t\cdot 2\cdot t\frac{(b-a)^2}{4}}} && \text{(Hoeffding's inequality)}\\
&= \sum_{t=1}^{\infty}2e^{-\log\left(4t^2/\delta\right)}\\
&= \frac{\delta}{2}\sum_{t=1}^{\infty}\frac{1}{t^2}=\frac{\delta\cdot\pi^2}{12}\leq\delta
\end{aligned}$$

7.d

To prove the second statement in Theorem 1 we will construct a reasoning process similar to the one outlined in [2]. Similar to sub-question 7.b if we assume that $\mathcal{E}^c$ holds (is satisfied) we have that at any step $t \geq 1$ and for all $k \in \{1, \ldots, K\}$:

$$\mu_k + U(t, \delta/K) \geq \hat{\mu}_{k,t} \geq \mu_k - U(t, \delta/K) \tag{5}$$

By employing the update rules outlined in Algorithm 1, it is evident that for any arm denoted as $k \neq k^\star$, the algorithm eliminates it when the subsequent condition is satisfied:

$$\hat{\mu}_{k^\star,t} - U(t, \delta/K) > \hat{\mu}_{k,t} + U(t, \delta/K)$$

Due to Equation 5, the above inequality is satisfied as long as

$$\mu^\star - 2U(t, \delta/K) > \mu_k + 2U(t, \delta/K) \Leftrightarrow \Delta_k > 4U(t, \delta/K) \quad (\text{with } 0 < \Delta_k \leq 1)$$

In our setup, since the random variables $X_{k,t}$ are i.i.d bounded random variables, we can establish a valid time confidence interval as described in the previous subsection, given by:

$$U = \sqrt{\frac{\log\left(4t^2/\delta\right)}{2t}}$$

It is evident that U exhibits a monotonically decreasing behavior with respect to time t. Thus, our objective is to determine the minimum time T_k such that $\Delta_k > 4U\left(T_k, \delta/K\right)$. This condition ensures that each non-optimal arm k will be removed from the set S_t after T_k steps (assuming T_k is an integer; otherwise, we round it up). Using the confidence bound from [3], presented in Equation $(\star)$ of [2], it could be shown that

$$T_k \geq c\Delta_k^{-2} \log\left(\frac{K\Delta_k^{-1}}{\delta}\right)$$

where $c > 0$ is a constant (since T_k must be bigger than 0). If we sum up the bound for each non-optimal arm, we find that the total number of samples needed to remove all non-optimal arms is given by:

$$\mathcal{O}\left(\sum_{k \neq k^\star}^{K} T_k\right) = \mathcal{O}\left(\sum_{k \neq k^\star}^{K} \Delta_k^{-2} \log\left(K \log\left(\Delta_k^{-1}\right)\right)\right)$$

References

[1] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[2] Kevin Jamieson. Lecture 4: Stochastic multi-armed bandits, pure exploration. PDF Document, Winter 2018. CSE599i: Online and Adaptive Machine Learning.

[3] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models, 2016.

[4] P. Tchébychef. Des valeurs moyennes (traduction du russe, n. de khanikof. *Journal de Mathématiques Pures et Appliquées*, pages 177–184, 1867.

[5] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[6] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Collaborations: Matteo Palo

1 Data-dependent generalization bound for hard-margin SVM

1.a

Let $B = \|w^\star\|_2$, and consider the set $\mathcal{F}_B = \{f(x) = \langle w, x\rangle : w \in \mathbb{R}^d, \|w\|_2 \le B\}$. By the definition of the hard-SVM and since $\|w_{\text{SVM}}\|_2 \le \|w^\star\|_2$, it holds that $f_{\text{SVM}} \in \mathcal{F}_B$. Given $\gamma \ge 0$, using Equation (1) from the assignment sheet, we can see that it holds with probability at least $1-\delta$ for f_{SVM}:

$$\mathbb{P}\left(Y f_{SVM}(X) < 0\right) \le \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{y_i f_{SVM}(x_i)<\gamma} + \frac{2D\,\|w^*\|_2}{\gamma\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)}{n}}$$

We now minimize the right-hand side of the above inequality with respect to γ. For $\gamma \le 1$, we find that $R^{\gamma}_{\text{SVM},n}$ is 0. Considering the fact that $\frac{1}{\gamma}$ (from the Rademacher term in the above bound) is decreasing in the interval $\gamma \in [0,1]$, we conclude that the minimum of the RHS is obtained with $\gamma = 1$. Therefore,

$$\mathbb{P}\left(Y f_{SVM}(X) < 0\right) \le \frac{2D\,\|w^*\|_2}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)}{n}}\,.$$

1.b

From the definition of $\mathcal{F}$ we observe that for each f there exists a $k(f)$ that is the smallest index k s.t. f is contained in $\mathcal{F}_k$. We now compute

$$\begin{aligned}
1-\delta &\le 1-\delta_{k(f)} \\
&\le \mathbb{P}\left(\bigcap_{f'\in\mathcal{F}_{k(f)}} E_{k(f),f'}\right) \\
&\le \mathbb{P}\left(E_{k(f),f}\right) \\
&= \mathbb{P}\left(R(f) - R_n(f) \le c\sqrt{\frac{\log\left(1/\delta_{k(f)}\right)}{n}} + 2\mathcal{R}_n\left(\mathcal{F}_{k(f)}\right)\right)
\end{aligned}$$

1.c

Suppose $\|w_{SVM}\| > 1$[1].
Consider the function class $\mathcal{F} = \cup_{k=1}^{\infty}\mathcal{F}_{B_k}$, with $\mathcal{F}_{B_k} = \{f(x) = \langle w, x\rangle : w \in \mathbb{R}^d, 1 < \|w\|_2 \le B_k\}$ and $B_k = e^k$. Let $\delta_k = \frac{\delta}{4k^2}$ (which satisfies $\sum_{k=1}^{\infty}\delta_k \le \delta$). Using Equation (1) from the assignment sheet and the previous sub-question, we have that for all $f \in \mathcal{F}$ with probability at least $1-\delta$, it holds that (use notation from now $k := k(f)$):

$$\mathbb{P}(Y f(X) < 0) \le \frac{2DB_k}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta_k)}{n}} \tag{1}$$

If we now let $k = \lceil \log(\|w\|_2)\rceil$ we have that $B_k \le e\|w\|_2$ and $\frac{1}{\delta_k} = \frac{4\lceil\log(\|w\|_2)\rceil^2}{\delta} \le \frac{(4\log(\|w\|_2))^2}{\delta}$. By inserting these values in Equation (1) and choosing $f = f_{SVM}$ we get the desired result

$$\begin{aligned}
\mathbb{P}(Y f_{SVM}(X) < 0) &\le \frac{2eD\|w_{SVM}\|_2}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta) + 2\log(4\log(\|w_{SVM}\|_2))}{n}} \\
&\le \frac{2eD\|w_{SVM}\|_2}{\sqrt{n}} + c\sqrt{\frac{2\log(1/\delta) + 2\log(4\log(\|w_{SVM}\|_2))}{n}} \\
&\le \frac{2eD\|w_{SVM}\|_2}{\sqrt{n}} + c'\sqrt{\frac{\log(1/\delta) + \log(4\log(\|w_{SVM}\|_2))}{n}}
\end{aligned}$$

for some constant c'.

[1] We need that $\|w_{SVM}\| > 1$ since in the result we have to prove the argument of the nested $\log$ must be strictly larger than 1 so that this last is positive.

2 Rates for smooth functions

2.a

We have to prove the following:

$$\|f - f_\beta\|_\infty \leq 2\epsilon L$$

I couldn't menage to figure out a proof.

2.b

First recall from Lecture 8 (Theorem 13.5 of [2]):

Theorem 1. *If $\mathcal{F}^\star$ is star-shaped, we have for the square loss minimizer $\hat{f}$ for any $t \geq 1$:*

$$\mathbb{P}\left[\left\|\hat{f} - f^*\right\|_n^2 \geq 16t\delta_n^2\right] \leq e^{-\frac{nt\delta_n^2}{2\sigma^2}}.$$

The problem therefore reduces to finding a δ_n that satisfies the critical inequality (13.17) of [2]. To do so, we can apply the Corollary of Dudley's integral from Lecture 8 (also 13.7 of [2]) that states:

Corollary 1. *If $\mathcal{F}$ is star-shaped, any $\delta \in [0, \sigma]$ such that*

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}\left(t; \mathcal{F}^\star\left(x_1^n\right) \cap \mathbb{B}_n(\delta), \|\cdot\|_n\right)} dt \leq \frac{\delta^2}{4\sigma}$$

satisfies the critical inequality.

Additionally recall that:

Proposition 1. *For $\mathcal{F}_{(2)}$ and $\mathcal{F}_{1,1}$ defined as in the assignment sheet, it holds that $\mathcal{F}_{(2)} \subset \mathcal{F}_{1,1}$.*

Proof. Consider a function denoted as f belonging to the class $\mathcal{F}_{(2)}$. The first property of $\mathcal{F}_{(2)}$ is straightforwardly met. When it comes to the second property, which involves the Lipschitz continuity of $f^{(1)}$, it follows from the fact that a function is Lipschitz if and only if its first derivative is bounded (apply this reasoning to $f^{(1)}$, which has a bounded derivative because of its definition). □

In the special case of $\mathcal{F} = \mathcal{F}_{(2)}$, Proposition 1 says that $\mathcal{F}_{(2)} \subset \mathcal{F}_{1,1}$. Therefore we now compute the LHS of the inequality in Corollary 1 above for $\mathcal{F}_{1,1}$ (since it follows trivially from the Definition of $\mathcal{F}_{1,1}$ that it is star shaped):

$$\begin{aligned}
\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}\left(t; \mathcal{F}_{1,1}, \|\cdot\|_\infty\right)} dt &= \frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \left(\frac{1}{t}\right)^{\frac{1}{4}} dt \\
&\leq \frac{16}{\sqrt{n}} \int_{0}^{\delta} \left(\frac{1}{t}\right)^{\frac{1}{4}} dt \qquad \text{("trick" from Lecture 7)} \\
&= \frac{64}{3\sqrt{n}} \delta^{3/4}
\end{aligned}$$

So, the LHS of the inequality of Corollary 1 is of the order $\mathcal{O}\left(\frac{\delta^{3/4}}{\sqrt{n}}\right)$. On the other hand we have that the RHS of the inequality of Corollary 1 is of the order $\mathcal{O}\left(\frac{\delta^2}{\sigma}\right)$. By comparing the two terms and solving for $\delta_n = \delta$ we obtain a value for δ_n that satisfies the critical inequality: $\delta_n = \mathcal{O}\left(\left(\sigma/\sqrt{n}\right)^{4/5}\right) = c_0 \cdot \left(\sigma/\sqrt{n}\right)^{4/5}$ for some constant c_0. By plugging in δ_n in the inequality from Theorem 1, grouping some constants, and choosing $t = 1$, the desired result is obtained.

2.c

First recall from Lecture 9 the following:

Lemma 1 (Local Gaussian Complexity for norm-bounded RKHS (Corollary 13.18 of [2]))**.** *Defining $\hat{\mu}_j$ as eigenvalues of the kernel matrix K, we have*

$$\widetilde{\mathcal{G}}_n\left(\mathcal{F}_1; \delta\right) \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^{n} \min\left\{\delta^2, \hat{\mu}_j\right\}}$$

Definition 1 (R-modified critical quantity $\delta_{n;R}$)**.** *We define $\delta_{n;R}$ to be the smallest $\delta > 0$ satisfying*

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

Theorem 2 (Prediction error of norm-bounded RKHS)**.** *Assume $f^\star \in \mathcal{F}_R$. Then we have for least-squares estimate $\widehat{f}_R \in \mathcal{F}_R$*

$$\left\|\widehat{f}_R - f^\star\right\|_n^2 \leq c_0 R^2 \delta_{n;R}^2$$

with probability $\geq 1 - c_1 e^{-c' \frac{n R^2 \delta_{n;R}^2}{\sigma^2}}$.

We will now follow the same procedure that the professor used to solve example 1 in slide 10 of Lecture 9 on the blackboard. Hence:

$$\begin{aligned}
\widetilde{\mathcal{G}}_n\left(W_2^\alpha([0,1]);\delta\right) &\leq \sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n}\min\left\{\delta^2, \hat{\mu}_j\right\}} \\
&= \sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n}\min\left\{\delta^2, j^{-2\alpha}\right\}} \\
&= \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + \sum_{j=k+1}^{n} j^{-2\alpha}} && \text{(for smallest } k \text{ s.t. } (k+1)^{-2\alpha} \leq \delta^2\text{)} \\
&\leq \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + \int_{k+1}^{\infty}\frac{1}{j^{2\alpha}}dj} \\
&= \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + \frac{1}{2\alpha-1}(k+1)^{1-2\alpha}} \\
&= \mathcal{O}\left(\sqrt{\frac{k\delta^2}{n}}\right) && \text{($k\delta^2 \geq (k+1)^{1-2\alpha}$ by definition of } k\text{)} \\
&\leq \mathcal{O}\left(\sqrt{\frac{\delta^{2-1/\alpha}}{n}}\right) && \text{(by def. of } k\text{, } k^{-2\alpha} \geq \delta^2 \Leftrightarrow k \leq \delta^{-1/\alpha}\text{)}
\end{aligned}$$

From now on, we proceed similarly to the previous subquestion by first finding a $\delta_{n;R}$ that satisfies Definition 1 with $R = 1$. We have that the right-hand side (RHS) of the inequality in Definition 1 is of the order $\mathcal{O}\left(\frac{\delta^2}{\sigma}\right)$. By comparing the term obtained in the above align of equations (same order as LHS of Definition 1) with the RHS of Definition 1 and solving for $\delta_{n;R} = \delta$, we obtain a value for $\delta_{n;R}$ that satisfies the R-modified critical inequality: $\delta_{n;R} = \mathcal{O}\left((\sigma/\sqrt{n})^{\frac{2\alpha}{2\alpha+1}}\right) = c \cdot (\sigma/\sqrt{n})^{\frac{2\alpha}{2\alpha+1}}$ for some constant c. By plugging in $\delta_{n;R}$ into the inequality from Theorem 2, grouping some constants, and choosing $R = 1$, as mentioned before, the desired result is obtained.

3 Sparse linear functions

3.a

From the definition of Gaussian complexity it follows immediately that

$$\begin{aligned}\widetilde{\mathcal{G}}_n\left(\mathcal{F}_{B,s}\left(x_1^n\right)\right) &= \frac{1}{n}\mathbb{E}\sup_{\theta\in\mathcal{F}_{B,s}\left(x_1^n\right)}\sum_{i=1}^{n} w_i\left\langle\theta, x_i\right\rangle \\ &= \frac{1}{n}\mathbb{E}\sup_{\theta\in\mathcal{F}_{B,s}\left(x_1^n\right)}\sum_{i=1}^{n}\left\langle\theta, w_i\cdot x_i\right\rangle \quad \text{(linearity of } \langle\cdot,\cdot\rangle) \\ &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta\in\mathcal{F}_{B,s}\left(x_1^n\right)}\left\langle\theta, \frac{X^\top w}{\sqrt{n}}\right\rangle .\end{aligned}$$

Remark 1. *Without loss of generality, we can observe that if the j-th entry in θ (with $j\in[d]$) is zero, then we could remove the j-th column from $X^\top$ since that column would be multiplied by 0. In the following, we denote θ_S as the vector obtained from θ by retaining only the non-zero entries indexed by S. We can then apply the Cauchy-Schwarz inequality to obtain the desired result.*

Following the previous remark we now compute:

$$\begin{aligned}\widetilde{\mathcal{G}}_n\left(\mathcal{F}_{B,s}\left(x_1^n\right)\right) &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta\in\mathcal{F}_{B,s}\left(x_1^n\right)}\left\langle\theta, \frac{X^\top w}{\sqrt{n}}\right\rangle \\ &= \mathbb{E}\sup_{\theta_S\in\mathcal{F}_{B,s}\left(x_1^n\right),|S|=s}\left\langle\theta_S, \frac{X_S^\top w}{n}\right\rangle \\ &\le \mathbb{E}\sup_{\theta_S\in\mathcal{F}_{B,s}\left(x_1^n\right),|S|=s}\|\theta_S\|_2\cdot\left\|\frac{X_S^\top w}{n}\right\|_2 \quad \text{(Cauchy-Schwartz inequality)} \\ &\le B\cdot\mathbb{E}\max_{|S|=s}\left\|\frac{X_S^\top w}{n}\right\|_2 \quad (\|\theta_S\|_2\le B)\end{aligned}$$

3.b

We first recall the following Theorem:

Theorem 3 (2.26 of [2])**.** *Let $(X_1,\dots,X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f:\mathbb{R}^n\to\mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then the variable $f(X)-\mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence*

$$\mathbb{P}[|f(X)-\mathbb{E}[f(X)]|\ge t]\le 2e^{-\frac{t^2}{2L^2}} \quad \textit{for all } t\ge 0.$$

First, note that $\lambda_{\max}\left(\frac{X_S^\top X_S}{n}\right)\le C^2$ implies that $\frac{X_S X_S^\top}{n}\preceq C^2 I$, where $\preceq$ denotes the matrix inequality, and I is the identity matrix. We will now proceed following the given hint.
We now prove that w_S is a C-Lipschitz function, and so is its l_2 norm. Let $f(w)=w_S$. For $w,w'\in\mathbb{R}^n$ we therefore have

$$\|f(w)-f(w')\|_2 = \left\|\frac{1}{\sqrt{n}}X_S^\top(w-w')\right\|_2 \le \left\|\frac{X_S}{\sqrt{n}}\right\|_2\|w-w'\|_2 \le C\|w-w'\|_2$$

where above we used that $\left\|\frac{X_S}{\sqrt{n}}\right\|_2 = s_{\max}\left(\frac{X_S}{\sqrt{n}}\right) = \sqrt{\lambda_{\max}\left(\frac{X_S^\top X_S}{n}\right)}\le C$ due to Proposition 2.
Since for any $i\in[s]$, $(w_S)_i$ is a linear combination of i.i.d standard Gaussian variables and $\|w_S\|_2$ is C-Lipschitz, if we can prove that $\mathbb{E}[\|w_S\|_2]\le C\sqrt{s}$, then by Theorem 3, it follows that:

$$\mathbb{P}\left(\|w_S\|_2\ge\sqrt{s}C+\delta\right)\le\mathbb{P}\left[\|w_S\|_2\ge\mathbb{E}\left[\|w_S\|_2\right]+\delta\right]\le\mathrm{e}^{-\frac{\delta^2}{2C^2}}$$

Therefore, we will now prove that $\mathbb{E}[\|w_S\|_2] \leq C\sqrt{s}$:

$$\begin{aligned}
\mathbb{E}[\|w_S\|_2] &= \mathbb{E}\left[\left\|\frac{X_S^\top w}{\sqrt{n}}\right\|_2\right] \\
&\leq \sqrt{\mathbb{E}\left[\frac{w^T X_S X_S^\top w}{n}\right]} && \text{(Jensen's inequality)} \\
&= \sqrt{\mathbb{E}\left[\frac{\operatorname{Tr}\left(w^T X_S X_S^\top w\right)}{n}\right]} && \text{(Eq. (17) of [1])} \\
&= \sqrt{\frac{\operatorname{Tr}\left(X_S^\top X_S \mathbb{E}\left[ww^\top\right]\right)}{n}} && \text{(Eq. (14) of [1] and linearity of expectation)} \\
&= \sqrt{\frac{\operatorname{Tr}\left(X_S^\top X_S\right)}{n}} && \left(\mathbb{E}\left[ww^\top\right] = 1\right) \\
&= \sqrt{\sum_i \lambda_i\left(\frac{X_S^\top X_S}{n}\right)} && \text{(Eq. (11) of [1])} \\
&\leq \sqrt{s\lambda_{\max}\left(\frac{X_S^\top X_S}{n}\right)} \leq C\sqrt{s} && \left(\lambda_{\max}\left(\frac{X_S^\top X_S}{n}\right) \leq C^2\right)
\end{aligned}$$

3.c

First recall from HW1 exercise 3.a that for a sequence of subgaussian (dependent or not) random variable $X_1, \ldots, X_n$, for all $n \geq 1$ we have

$$\mathbb{E} \max_{i=1,\ldots,n} X_i \leq \sqrt{2\sigma^2 \log n}\,.$$

We will begin with the result obtained in the first subquestion of this exercise. Also, recall from the previous subquestion that $\|w_S\|_2 - \mathbb{E}[\|w_S\|_2]$ is subgaussian distributed with a parameter of at most C. We proceed to compute:

$$\begin{aligned}
\widetilde{\mathcal{G}}_n\left(\mathcal{F}_{B,s}\left(x_1^n\right)\right) &\leq \frac{B}{\sqrt{n}}\mathbb{E}\max_{|S|=s}\|w_S\|_2 \\
&= \frac{B}{\sqrt{n}}\mathbb{E}\max_{|S|=s}\|w_S\|_2 - \frac{BC\sqrt{s}}{\sqrt{n}} + \frac{BC\sqrt{s}}{\sqrt{n}} \\
&\leq \frac{B}{\sqrt{n}}\mathbb{E}\max_{|S|=s}\|w_S\|_2 - \frac{B\mathbb{E}\left[\|w_S\|_2\right]}{\sqrt{n}} + \frac{BC\sqrt{s}}{\sqrt{n}} \\
&\leq \frac{2B}{\sqrt{n}}\sqrt{2C^2\log\binom{d}{s}} + \frac{BC\sqrt{s}}{\sqrt{n}} && \left(\text{3.a HW1, } \binom{d}{s} = \#\text{ ways select set } S \subset [d]\right) \\
&\leq \frac{2B}{\sqrt{n}}\sqrt{2C^2 s\log\left(\frac{e\cdot d}{s}\right)} + \frac{BC\sqrt{s}}{\sqrt{n}} && \left(\text{using } \binom{n}{k} \leq (ne/k)^k\right) \\
&\leq \mathcal{O}\left(BC\sqrt{\frac{s\log\left(\frac{ed}{s}\right)}{n}}\right) && (\text{since } d \geq s)
\end{aligned}$$

3.d

Starting from the intermediate result from the first subquestion of this problem we compute:

$$\begin{aligned}
\widetilde{\mathcal{G}}_n\left(\mathcal{F}_{B,s}\left(x_1^n\right)\right) &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta\in\mathcal{F}_{B,s}\left(x_1^n\right)}\left\langle\theta, \frac{X^\top w}{\sqrt{n}}\right\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta\in\mathcal{F}_{B,s}\left(x_1^n\right)}\left\langle\frac{X\theta}{\sqrt{n}}, w\right\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta_S\in\mathcal{F}_{B,s}\left(x_1^n\right),|S|=s}\left\langle\frac{X_S\theta_S}{\sqrt{n}}, w\right\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta_S\in\mathcal{F}_{B,s}\left(x_1^n\right),|S|=s}\left\langle\frac{X_S\theta_S}{\sqrt{n}}, P_S w\right\rangle \\
&\leq \frac{B}{\sqrt{n}}\mathbb{E}\sup_{|S|=s}\|P_S w\|_2
\end{aligned}$$

where $P_S = X_S(X_S^\top X_S)^{-1}X_S^\top$ is the orthogonal projection matrix onto the $\mathrm{span}(X_S)$. We now prove that $w_{S'}$ is a 1-Lipschitz function, which implies that the l_2 norm of $w_{S'}$ is also a 1-Lipschitz function. Let $w, w' \in \mathbb{R}^n$:

$$\|P_S w - P_S w'\|_2 \leq \|P_S\|_2\|w - w'\|_2 = 1 \cdot \|w - w'\|_2$$

where we used the Cauchy-Schwarz inequality. Similarly to the second subquestion, we can therefore also apply here Theorem 3. The proof now follows directly by the previous 2 subquestions with $C = 1$.

4 Classification error bounds for hard margin SVM

4.a

Let $\hat{\boldsymbol{\theta}} = [r, \gamma\tilde{\theta}]$, and let $\mathbf{x} = [yr, \tilde{x}]$ with $\tilde{x} \sim \mathcal{N}(0, I_{d-1})$. We compute:

$$\begin{aligned}
\mathbb{P}\left[y\hat{\boldsymbol{\theta}}^\top \mathbf{x} < 0\right] &= \mathbb{P}\left[y \cdot \left([r, \gamma\tilde{\theta}]^\top \cdot [yr, \tilde{x}]\right) < 0\right] \\
&= \mathbb{P}\left[r^2 + \gamma \sum_{i=1}^{d-1} (y\tilde{x}_i)\tilde{\theta}_i < 0\right] \quad (y^2 = 1) \\
&= \mathbb{P}\left[r^2 + \gamma \sum_{i=1}^{d-1} \tilde{x}_i\tilde{\theta}_i < 0\right],
\end{aligned}$$

In the third equality, we used the fact that since $\tilde{x}_i$ is Gaussian distributed, $-\tilde{x}_i$ is also Gaussian distributed with the same distribution. Now, let $Z = \sum_{i=1}^{d-1} \tilde{x}_i\tilde{\theta}_i$. Since Z is composed of a sum of independent standard Gaussian random variables multiplied with some $\tilde{\theta}$ with $\|\tilde{\theta}\|_2 = 1$, then Z itself is standard Gaussian distributed[2]. Therefore, we have:

$$\mathbb{P}\left[r^2 + \gamma \cdot Z < 0\right] = \mathbb{P}\left[Z < -\frac{r^2}{\gamma}\right] = \Phi\left(-\frac{r^2}{\gamma}\right), \tag{2}$$

The dependence on r of the quantity in Equation (2) is as follows: it monotonically decreases with respect to r.

4.b

First recall the following property of the l_2 norm:

Proposition 2. *Let $\mathbf{A} \in \mathbb{R}^{N\times N}$ be a real matrix. The operator norm of $\mathbf{A}$ is defined as*

$$\|\mathbf{A}\|_2 \overset{def}{=} \sup_{\boldsymbol{x}\neq 0} \frac{\|\mathbf{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$$

It holds that $\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A})$, i.e., the largest singular value of $\mathbf{A}$.

Proof. Let $\boldsymbol{A} = U\Sigma V^\top$ be the SVD decomposition of $\boldsymbol{A} \in \mathbb{R}^{N\times N}$. Recall that U and V are unitary matrices. We have that

$$\begin{aligned}
\|\mathbf{A}\|_2 &= \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \\
&= \sup_{\mathbf{u}\in\mathbb{R}^n : \|\mathbf{u}\|_2 = 1} \|\mathbf{A}\mathbf{u}\|_2 \\
&= \sup_{\mathbf{u}\in\mathbb{R}^n : \|\mathbf{u}\|_2 = 1} \left\|\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top\mathbf{u}\right\|_2 \\
&= \sup_{\mathbf{u}\in\mathbb{R}^n : \|\mathbf{u}\|_2 = 1} \left\|\boldsymbol{\Sigma}\mathbf{V}^\top\mathbf{u}\right\|_2 \\
&= \sup_{\mathbf{y}\in\mathbb{R}^n : \|\mathbf{y}\|_2 = 1} \|\boldsymbol{\Sigma}\mathbf{y}\|_2
\end{aligned}$$

Since $\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_n)$, where σ_1 is the largest singular value. The max for the above, σ_1, is attained when $y = (1, \cdots, 0)^T$. □

From the problem assignment sheet we have that γ is defined as follows

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

Similar to how we did in the introduction to machine learning lecture, we can now rewrite the max-l_2-margin problem as follows (note that we also use the fact that if a random variable X is Gaussian distributed, so is its negative $-X$):

$$\gamma = \max_{\xi,\theta\in\mathbb{R}^{d-1}} \xi \qquad \text{s.t. } \|\theta\|_2 = 1,\ \theta^\top \tilde{X} \overset{\circ}{\geq} \boldsymbol{\xi},$$

[2]The zero mean follows trivially from linearity of expectation, while for the variance we note that $\operatorname{var}\left(\sum_{i=1}^{d-1} \tilde{x}_i\tilde{\theta}_i\right) = \sum_{i=1}^{d-1} \tilde{\theta}_i^2 = 1$ using $\|\tilde{\theta}\|_2 = 1$

where $\boldsymbol{\xi} = (\xi, \xi, \ldots, \xi)^\top \in \mathbb{R}^n$. By considering the two constraints and taking the norm on both sides of the second one, we obtain

$$\|\theta^\top \tilde{X}\|_2 \geq \xi\sqrt{n}$$

Since $\gamma \leq \xi$ we have that

$$\begin{aligned}
\gamma &\leq \frac{\|\theta^\top \tilde{X}\|_2}{\sqrt{n}} \\
&\leq \frac{\|\theta\|_2 \|\tilde{X}\|_2}{\sqrt{n}} && \text{(Cauchy-Schwarz Inequality)} \\
&= \frac{s_{\max}(\tilde{X})}{\sqrt{n}} && (\|\theta\|_2 = 1 \text{ and Proposition 2})
\end{aligned}$$

4.c

i) From the assignment sheet we have that $X_{u,v} = \langle Xu, v\rangle$. Using the definition of X it follows that $X_{u,v} = \sum_{i=1}^{d}\sum_{j=1}^{n} x_{ij}u_i v_j$ is Gaussian distributed with mean 0 and variance $\sum_{i=1}^{d}\sum_{j=1}^{n}(u_i v_j)^2 = \sum_{i=1}^{d} u_i^2 \sum_{j=1}^{n} v_j^2 = \|u\|_2\|v\|_2$. Similarly we obtain that $Y_{u,v}$ is Gaussian distributed with mean 0 and variance $\sum_{i=1}^{d}\sum_{j=1}^{n} u_i^2 + v_j^2 = \|u\|_2 + \|v\|_2$.
To prove the result we therefore need to prove that:

$$\operatorname{var}\left(X_{u,v} - X_{u',v'}\right) \leq \operatorname{var}\left(Y_{u,v} - Y_{u',v'}\right)$$

Since for a Gaussian random variable X, its negative $-X$ has the same distribution, and for independent and identically distributed (i.i.d.) random variables, the variance of the sum is equal to the sum of the variances, the problem reduces to proving that:

$$\|u\|_2\|v\|_2 + \|u'\|_2\|v'\|_2 \leq \|u\|_2 + \|v\|_2 + \|u'\|_2 + \|v'\|_2$$

Since u, u' and v, v' are unitary vectors, the above inequality is always satisfied.

ii) Let $\mathcal{S}^k = \left\{y \in \mathbb{R}^{k+1} \mid \|y\|_2 = 1\right\}$. We compute:

$$\begin{aligned}
\mathbb{E}\left[s_{\max}(X)\right] &= \mathbb{E}\left[\max_{u\in\mathcal{S}^{d-1}, v\in\mathcal{S}^{n-1}} \langle Xu, v\rangle\right] \\
&= \mathbb{E}\left[\max_{u\in\mathcal{S}^{d-1}, v\in\mathcal{S}^{n-1}} X_{u,v}\right] \\
&\leq \mathbb{E}\left[\max_{u\in\mathcal{S}^{d-1}, v\in\mathcal{S}^{n-1}} Y_{u,v}\right] && \text{(Lemma 2 on assignment sheet)} \\
&= \mathbb{E}\left[\max_{u\in\mathcal{S}^{d-1}, v\in\mathcal{S}^{n-1}} \langle g, u\rangle + \langle h, v\rangle\right] \\
&= \mathbb{E}\left[\|g\|_2 + \|h\|_2\right] && (\max \text{ obtained with } u = g/\|g\|_2 \text{ and } v = h/\|h\|_2) \\
&= \sqrt{d} + \sqrt{n}
\end{aligned}$$

where in the last equality we used the fact that $\|g\|_2 \sim \sqrt{\chi_d^2}, \|h\|_2 \sim \sqrt{\chi_n^2}$ since $g \in \mathbb{R}^d$ and $h \in \mathbb{R}^n$ are independent standard normal distributed variables.

4.d

In order to apply Theorem 3 and prove that $s_{\max}(\tilde{X}) \leq \sqrt{d} + \sqrt{n} + t$ with a probability of at least $1 - 2e^{-t^2/2}$, we need a vector of Gaussian random variables ($\operatorname{vec}(\tilde{X})$), and it is required that the function $f(\tilde{X}) = s_{\max}(\tilde{X})$ is 1-Lipschitz. This requirement arises from the following: consider two matrices X and Y in $\mathbb{R}^{n\times m}$

$$\|s_{\max}(X) - s_{\max}(Y)\|_2 = \|\|X\|_2 - \|Y\|_2\|_2 \leq \|X - Y\|_2 .$$

where for the last inequality we used the reverse triangular inequality.

5 Collective learning: crowdsourcing an answer and collecting good practice questions from the group

5.a Question

Discuss the relationship between the VC dimension and the number of parameters in a function class. Prove that the VC dimension is not always equivalent to the number of parameters, and provide an example of a function class where the VC dimension is equal to the number of parameters.

5.b Answer

As seen in the lecture, the Vapnik-Chervonenkis (VC) dimension and the number of parameters in a function class are both essential concepts in machine learning, particularly in the context of statistical learning theory. The VC dimension characterizes the capacity of a hypothesis class to shatter a set of data points. On the other hand, the number of parameters in a function class represents the flexibility or complexity of the class in terms of the functions it can represent.

Proving That VC Dimension is Not Always Equivalent to the Number of Parameters:
The VC dimension and the number of parameters do not always align. To illustrate this, we can present a counter-example:

Consider a scalar value, represented as $t \in \mathbb{R}$, and the function $f_t(x) = \text{sign}(\sin(t \cdot x))$. We define a function class, denoted as $\mathcal{F}$, as $\{f_t : [-1, 1] \to \mathbb{R} \mid t \in \mathbb{R}\}$. Notably, this class demonstrates an infinite VC dimension, as demonstrated by the following intuition:
For any natural number d and any set of samples $x_1, x_2, \ldots, x_d \in [-1, 1]$, the function class $\mathcal{F}$ can achieve any possible labeling for the given sample. This versatility arises from the fact that, for sufficiently large values of t, the function $\sin(t \cdot x)$ oscillates at an extremely high frequency. Given the freedom to select this frequency, any labeling for the sample $x_1, x_2, \ldots, x_d$ can be attained.
Consequently, this characteristic endows the function class with the ability to shatter any set of d points, regardless of the natural number d. This result leads to an infinite VC dimension, despite the class having only a single parameter.

Example of a Function Class Where VC Dimension Equals the Number of Parameters:
A simple linear regression model is an example where the VC dimension equals the number of parameters. The hypothesis class consists of lines defined by two parameters: the slope and the intercept. You can shatter any set of two data points using this class, and it also has two parameters. Thus, the VC dimension is equal to the number of parameters in this case.

In summary, the VC dimension and the number of parameters in a function class are related but not always equivalent. The VC dimension characterizes the capacity of a hypothesis class to shatter data points, while the number of parameters represents the complexity and flexibility of the class in terms of the functions it can represent.

References

[1] K. B. Petersen and M. S. Pedersen. The matrix cookbook, October 2008. Version 20081110.

[2] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.