# GML Fall 25, Homework 2: Generalization bounds

## 1 Gaussian and Rademacher complexities

**Read MW Chapter 5 as a reference.** The Gaussian complexity of a class of functions $\mathcal{H}$, for a fixed set of covariates $x_1, \ldots, x_n$, is defined as $\mathbb{E} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n w_i h(x_i)$ where $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. When deriving uniform bounds in regression problems with Gaussian noise, we will see that the Gaussian complexity arises naturally and is relatively easier to bound.

Let us now define, for an arbitrary set $\mathbb{T} \subset \mathbb{R}^n$, the Rademacher complexity as $\widetilde{\mathcal{R}}_n(\mathbb{T}) := \mathbb{E} \sup_{\theta \in \mathbb{T}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \theta_i$ with $\epsilon_i$ being i.i.d. Rademacher random variables and the Gaussian complexity as $\widetilde{\mathcal{G}}_n(\mathbb{T}) := \mathbb{E} \sup_{\theta \in \mathbb{T}} \frac{1}{n} \sum_{i=1}^n w_i \theta_i$ with $w_i$ i.i.d. standard normal random variables. In this question, we show how the Rademacher complexity is related to the Gaussian complexity. Formally, we prove that

$$\frac{\widetilde{\mathcal{G}}_n(\mathbb{T})}{2\sqrt{\log n}} \leq \widetilde{\mathcal{R}}_n(\mathbb{T}) \leq \sqrt{\frac{\pi}{2}} \widetilde{\mathcal{G}}_n(\mathbb{T}).$$

(a) **Show that** for any set $\mathbb{T}$ the Rademacher complexity satisfies the upper bound $\widetilde{\mathcal{R}}_n(\mathbb{T}) \leq \sqrt{\frac{\pi}{2}} \widetilde{\mathcal{G}}_n(\mathbb{T})$. Give an example of a set for which this bound is met with equality.

(b) **Show that** $\widetilde{\mathcal{G}}_n(\mathbb{T}) \leq 2\sqrt{\log n} \widetilde{\mathcal{R}}_n(\mathbb{T})$ for any set $\mathbb{T}$. Give an example for which this upper bound is tight up to a constant factor. You may use that $\widetilde{\mathcal{R}}_n(\phi(\mathbb{T})) \leq \widetilde{\mathcal{R}}_n(\mathbb{T})$ for any contraction (see Gaussian analog MW Proposition 5.28).

### Solution

(a) Using the fact that $\mathbb{E}|w_i| = \sqrt{\frac{2}{\pi}}$ if $w_i \sim N(0,1)$, we can write:

$$\begin{aligned}
\sqrt{\frac{2}{\pi}} \mathcal{R}_n(\mathbb{T}) &= \mathbb{E}_\epsilon \left[ \sup_\theta \frac{1}{n} \sum_{i=1}^n \epsilon_i \theta_i \sqrt{\frac{2}{\pi}} \right] = \mathbb{E}_\epsilon \left[ \sup_\theta \frac{1}{n} \sum_{i=1}^n \epsilon_i \theta_i \mathbb{E}_w[|w_i|] \right] \\
&\leq \mathbb{E}_{\epsilon, w} \left[ \sup_\theta \frac{1}{n} \sum_{i=1}^n \theta_i \epsilon_i |w_i| \right] \overset{(i)}{=} \mathbb{E}_{w'} \left[ \sup_\theta \frac{1}{n} \sum_{i=1}^n \theta_i w_i' \right] = \widetilde{\mathcal{G}}_n(\mathbb{T})
\end{aligned}$$

Step (i) comes from the fact that $w_i' := \epsilon_i |w_i|$ is distributed like a standard normal if $\epsilon_i$ is a Rademacher random variable and $w_i \sim N(0,1)$

Equality happens for instance when we take $\mathbb{T} = \mathbb{B}_\infty^d(1) = \{\theta \in \mathbb{T} \mid \|\theta\|_\infty \leq 1\}$

(b) We can write

$$\begin{aligned}
\widetilde{\mathcal{G}}_n(\mathbb{T}) &= \mathbb{E}_w\left[\frac{1}{n}\sup_\theta \sum_i w_i\theta_i\right] \\
&\overset{(i)}{=} \mathbb{E}_{w,\epsilon}\left[\frac{1}{n}\sup_\theta \sum_i \epsilon_i w_i\theta_i\right] = \mathbb{E}_{w,\epsilon}\left[\frac{1}{n}\sup_\theta \sum_i \epsilon_i |w_i|\theta_i\right] \\
&= \mathbb{E}_{w,\epsilon}\left[\max_j |w_j| \sup_\theta \frac{1}{n}\sum_i \epsilon_i \frac{|w_i|\theta_i}{\max_j |w_j|}\right] \\
&= \mathbb{E}_w\left[\max_j |w_j| \mathbb{E}_\epsilon\left[\sup_\theta \frac{1}{n}\sum_i \epsilon_i \frac{|w_i|\theta_i}{\max_j |w_j|}\right]\right] \\
&\overset{(ii)}{=} \mathbb{E}_w\left[\max_j |w_j| \widetilde{\mathcal{R}}_n(\phi_w(\mathbb{T}))\right] \overset{(iii)}{\leq} \mathbb{E}_w\left[\max_j |w_j| \widetilde{\mathcal{R}}_n(\mathbb{T})\right] \\
&= \widetilde{\mathcal{R}}_n(\mathbb{T})\mathbb{E}_w\left[\max_j |w_j|\right] \overset{(iv)}{\leq} 2\sqrt{\log n}\widetilde{\mathcal{R}}_n(\mathbb{T}).
\end{aligned}$$

For (i) we use symmetrization: By symmetry, the random variables $w_i$, $\epsilon_i w_i$ and $\epsilon_i|w_i|$ with independent (of $w_i$) Rademacher variables $\epsilon_i$ have the same distribution. In (ii) we define the function $\phi_{i,w}(\theta_i) = \frac{w_i\theta_i}{\max_j |w_j|}$ for arbitrarily fixed $w$ yields. It is then easy to verify that $\phi_{i,w}$ is a contraction and hence the contraction inequality can be used to obtain (iii). In (iv) we use the inequality proved in the first homework.

Equality (up to constant factors) happens for instance when we choose $\mathbb{T} = \mathbb{B}_1^d(1) = \{\theta \in \mathbb{T} \mid ||\theta||_1 \leq 1\}$.

# 2 Rates for smooth functions

Read MW Examples 5.10. through Example 5.12. (notice typos in Example 5.11. - it should be $\delta = \epsilon^{\alpha+\gamma}$ everywhere). The non-parametric least-squares estimate is defined as

$$\widehat{f} = \arg\min_{f\in\mathcal{F}} R_n(f) := \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2.$$

In this exercise we derive the prediction error bound for the examples of twice-differentiable functions $\mathcal{F}_{(2)}$ and $\alpha$-th order Sobolev spaces $\mathcal{W}_2^\alpha([0,1])$ on $[0,1]$.

$$\begin{aligned}
\mathcal{F}_{(2)} &:= \{f : [0,1] \to \mathbb{R} \mid \|f\|_\infty + \|f^{(1)}\|_\infty + \|f^{(2)}\|_\infty \leq C < \infty\} \\
\mathcal{W}_2^\alpha([0,1]) &:= \{f : [0,1] \to \mathbb{R} \mid f^{(i)} \in \mathcal{L}^2([0,1]) \text{ and } f^{(i)}(0) = 0 \,\forall i = 0,\ldots,\alpha-1\}
\end{aligned}$$

where $f^{(a)}$ stands for the $\alpha$-th (weak) derivative. Throughout the problem, we assume that $f^\star \in \mathcal{F}$.

(a) **Prove that** the set $\{f_\beta, \beta \in \{-1,+1\}^M\}$ in Example 5.10. forms a $2\epsilon L$-covering in the sup-norm.

(b) For the function class

$$\mathcal{F}_{\alpha,\gamma} = \{f : [0,1] \to \mathbb{R} \mid \|f^{(j)}\|_\infty \leq C_j \,\forall j = 0,\ldots,\alpha, |f^{(\alpha)}(x) - f^{(\alpha)}(x')| \leq L|x-x'|^\gamma \,\forall x, x' \in [0,1]\}$$

we have $\log \mathcal{N}(\epsilon; \mathcal{F}_{\alpha,\gamma}, \|\cdot\|_\infty) = O((\frac{1}{\epsilon})^{\frac{1}{\alpha+\gamma}})$. Use this fact to **establish the following prediction error bound** for the non-parametric least-squares estimate $\widehat{f}$ with $\mathcal{F} = \mathcal{F}_{(2)}$ for positive constants $c_0, c_1, c_2$ which may depend on $C$ but not on $n, \sigma^2$

$$\mathbb{P}(\|\widehat{f} - f^\star\|_n^2 \geq c_0(\frac{\sigma^2}{n})^{\frac{4}{5}}) \leq c_1 \mathrm{e}^{-c_2(n/\sigma^2)^{1/5}}$$

(c) For $\alpha$-th order Sobolev kernels, assume that the empirical eigenvalues decay with rate $\hat{\mu}_j = j^{-2\alpha}$ and we minimize the square loss in the constrained function class $\mathcal{F} = \{f \in \mathcal{W}_2^\alpha([0,1]) : \|f\|_\mathcal{F} \leq 1\}$. **Show that** the prediction error of the non-parametric least-squares estimate reads

$$\mathbb{P}\big[\|\widehat{f} - f^\star\|_n^2 \geq c_0(\frac{\sigma^2}{n})^{\frac{2\alpha}{2\alpha+1}}\big] \leq c_1 \mathrm{e}^{-c_2(\frac{n}{\sigma^2})^{\frac{1}{2\alpha+1}}}.$$

## Solution

(a) We prove that the set $\{f_\beta, \beta \in \{-1,+1\}^M\}$ is a $2\epsilon L$-cover of $\mathcal{F}_L$ by showing that for any $f \in \mathcal{F}_L$ it is possible to construct a sequence $\beta$ such that $\|f - f_\beta\|_\infty \leq 2\epsilon L$. For an arbitrary $f \in \mathcal{F}_L$, let us construct $\beta = \{\beta_1, ..., \beta_M\}$ in the following way:

$$\beta_1 = \text{sgn}(f(\epsilon));\ \beta_{k+1} = \text{sgn}\left(f((k+1)\epsilon) - l_k \epsilon L)\right), \forall k \geq 1$$

where $l_k \in \mathbb{Z}$ is the level in the grid on the vertical axis that approximates $f(k\epsilon)$ according to the previous choices of $\{\beta_1, ..., \beta_k\}$. Assuming the whole $\beta$ is known and the function $f$ is completely determined, we can write $f_\beta(k\epsilon) = l_k \epsilon L$. As shown in Exercise 5.10 from MW, $f_\beta \in \mathcal{F}_L, \forall \beta \in \{-1,+1\}^M$. So what remains to be proved is that an arbitrary $f \in \mathcal{F}_L$ is $2\epsilon L$-covered by $f_\beta$, with $\beta$ defined as above. More formally, we have to show that $\|f - f_\beta\|_\infty \leq 2\epsilon L$.

We propose a proof by induction over the M intervals that $|f(k\epsilon) - f_\beta(k\epsilon)| \leq \epsilon L, \forall k \in [M]$. An essential premise for several steps in the proof is that $f$ is L-Lipschitz. For the first interval we have for any $x \in [0, \epsilon]$ that:

$$\begin{aligned}
\sup_{x\in[0,\epsilon]} \big|f(x) - f_\beta(x)\big| &= \sup_{x\in[0,\epsilon]} \left|f(x) - \epsilon L \cdot \text{sgn}(f(x))\frac{x}{\epsilon}\right| \\
&\leq \sup_{x\in[0,\epsilon]} \big|f(x)\big| + \left|\epsilon L \cdot \text{sgn}(f(x))\frac{x}{\epsilon}\right| \\
&\leq 2\epsilon L
\end{aligned}$$

For the inductive step, we assume that $\sup_{x\in[0,k\epsilon]} |f(x) - f_\beta(x)| \leq 2\epsilon L$ and want to show that $\sup_{x\in(k\epsilon,(k+1)\epsilon]} |f(x) - f_\beta(x)| \leq 2\epsilon L$.

$$\begin{aligned}
\sup_{x\in(k\epsilon,(k+1)\epsilon]} |f(x) - f_\beta(x)| &= \sup_{x\in(k\epsilon,(k+1)\epsilon]} \left|f(x) - \left(f_\beta(k\epsilon) + \epsilon L \cdot \text{sgn}(f(x) - f_\beta(k\epsilon))\frac{x - k\epsilon}{\epsilon}\right)\right| \\
&= \sup_{x\in(k\epsilon,(k+1)\epsilon]} \left|f(x) - f_\beta(k\epsilon) + f_\beta(k\epsilon) - \left(f_\beta(k\epsilon) + \epsilon L \cdot \text{sgn}(f(x) - f_\beta(k\epsilon))\frac{x - k\epsilon}{\epsilon}\right)\right| \\
&\leq \sup_{x\in(k\epsilon,(k+1)\epsilon]} \big|\text{sgn}(f(x) - f_\beta(k\epsilon))\big| \cdot \left|\big|f(x) - f_\beta(k\epsilon)\big| - \epsilon L\frac{x - k\epsilon}{\epsilon}\right| \\
&\leq \sup_{x\in(k\epsilon,(k+1)\epsilon]} \left|\big|f(x) - f_\beta(k\epsilon)\big| - \epsilon L\frac{x - k\epsilon}{\epsilon}\right| \\
&\leq 2\epsilon L
\end{aligned}$$

The last inequality holds because on the one hand we have that $0 \leq \epsilon L\frac{x-k\epsilon}{\epsilon} \leq \epsilon L$ and on the other hand $0 \leq \big|f(x) - f_\beta(k\epsilon)\big| \leq \big|f(x) - f(k\epsilon) + f(k\epsilon) - f_\beta(k\epsilon)\big| \leq 2\epsilon L$.

*Remark*: A similar argument can be used to show that the same set is a $\epsilon L$-cover of $\mathcal{F}_l$, but in this case one would have to be more careful to keep into account the smoothness of a function $f \in \mathcal{F}_L$ *inside* the quadrants as well.

(b) The main idea is to bound the error of the non-parametric least-square estimate using the prediction error bound in Lecture 4/5 (MW Theorem 13.5). We set out to find a $\delta_n$ that satisfies the critical inequality and thus makes the bound in the theorem hold. We can use Dudley's integral to bound the localized Gaussian complexity in the critical inequality. One such result is given by Theorem on slide 7 Lecture 5 (MW Corollary 3.17). We use this to select the $\delta_n$. Concretely, for the function class $\mathcal{F}_{\alpha,\gamma}$, we can start by rewriting the integral as follows:

$$\begin{aligned}
l\frac{1}{\sqrt{n}}\int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}_{\alpha,\gamma}, \|\cdot\|_\infty)}dt &\leq \frac{1}{\sqrt{n}}\int_0^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}_{\alpha,\gamma}, \|\cdot\|_\infty)}dt \\
&= \frac{1}{\sqrt{n}}\int_0^{\delta} \left(\frac{1}{t}\right)^{\frac{1}{2(\alpha+\gamma)}} dt \\
&= \mathcal{O}\left(\frac{1}{\sqrt{n}}\delta^{1-\frac{1}{2(\alpha+\gamma)}}\right)
\end{aligned}$$

Using Corollary 13.7 from MW we can conclude that in order to choose a $\delta_n$ that satisfies the critical inequality it is sufficient to select a value that satisfies $\frac{1}{\sqrt{n}}\delta^{1-\frac{1}{2(\alpha+\gamma)}} \leq \mathcal{O}\left(\frac{\delta^2}{4\sigma}\right)$.

By rearranging the terms we obtain $\delta_n^2 \approx \frac{\sigma^2}{n}^{\frac{2(\alpha+\gamma)}{2(\alpha+\gamma)+1}}$.

By the definition of $\mathcal{F}_{(2)}$, we see that $\mathcal{F}_{(2)} \subset \mathcal{F}_{1,1}$ by the fundamental theorem of calculus. The final result follows now by plugging the value of $\delta_n^2 = c\frac{\sigma^2}{n}^{\frac{4}{5}}$ into the prediction error bound (note that we choose $t = 1$ in the bound by notation in lecture, which differs from the notation in the book).

(c) The solution follows the derivation in Example 13.20 in MW. We use the bound on the localized Gaussian complexity of a norm-bounded RKHS introduced in lecture 6 (see Lemma on slide 8). We then plug this into the critical inequality to choose a $\delta_n$ that satisfies it, thus bounding the prediction error with high probability. We start from the aforementioned lemma in the lecture. Let us choose $k \in \mathbb{N}$ such that $\hat{\mu}_k = k^{-2\alpha} \geq \delta^2 \geq (k+1)^{-2\alpha} = \hat{\mu}_{k+1}$ i.e. the index $k$ of the smallest eigenvalue larger than $\delta$.

$$\begin{aligned}
\widetilde{\mathcal{G}}_n(\mathcal{W}_2^\alpha([0,1]);\delta) &\leq \sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n}\min\{\delta^2, \hat{\mu}_j\}} \\
&= \sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n}\min\{\delta^2, \frac{1}{j^{2\alpha}}\}} \\
&= \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + \sum_{j=k+1}^{n}\frac{1}{j^{2\alpha}}} \\
&\overset{(i)}{\leq} \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + \int_{k+1}^{\infty}\frac{1}{t^{2\alpha}}dt} \\
&= \sqrt{\frac{2}{n}}\sqrt{k\delta^2 + \mathcal{O}\left((k+1)^{1-2\alpha}\right)} \\
&\overset{(ii)}{=} \sqrt{\frac{2}{n}}\sqrt{\mathcal{O}\left(k\delta^2\right)}
\end{aligned}$$

The resulting second term can be then upper bounded by an integral as we did in (i). In (ii) we use the fact that, by the definition of $k$, $k\delta^2 \geq (k+1)^{1-2\alpha}$.

In order to get rid of the dependence on $k$, we can further upper bound $k\delta^2$ like $k\delta^2 \leq \delta^{2-\frac{1}{\alpha}}$ by using the left-hand side inequality in the definition of $k$. We obtain that:

$$\widetilde{\mathcal{G}}_n(\mathcal{W}_2^\alpha([0,1]);\delta) \leq \sqrt{\frac{2}{n}}\sqrt{\mathcal{O}\left(k\delta^2\right)} \leq \mathcal{O}\left(\sqrt{\frac{\delta^{2-\frac{1}{\alpha}}}{n}}\right)$$

Using Corollary 13.7 from MW it follows that in order to satisfy the critical inequality, it suffices to choose a $\delta$ such that $\sqrt{\frac{\delta^{2-\frac{1}{\alpha}}}{n}} \leq \mathcal{O}\left(\frac{\delta^2}{\sigma}\right)$. After conveniently rearranging the terms we arrive at $\delta_n^2 \approx \left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$. Plugging everything into the statement of Theorem 13.5 from MW, like we did for part a), concludes the proof.

# 3 Sparse linear functions

We already looked at the complexity of linear function classes with a margin $\gamma$ and $\ell_2$ norm constraint in previous homeworks and lectures. In this exercise, we bound the Gaussian complexity of a smaller subset of $\ell_2$ constrained balls, i.e.,

$$\mathcal{F}_{B,s} = \{f(\cdot) = \langle\theta, x\rangle : \|\theta\|_0 \leq s, \|\theta\|_2 \leq B\}$$

This is a useful quantity as it gives intuition for why constraining the function class to sparse linear models can help to decrease the sample complexity below dimension $d$.

(a) Define $X \in \mathbb{R}^{n\times d}$ as consisting of rows $x_1, \ldots, x_n$ the sample covariate vectors. Let the matrix $X_S \in \mathbb{R}^{n\times|S|}$ be the submatrix of $X$ consisting of columns of $X$ that are indexed by $S$. First **show that** the Gaussian complexity $\mathcal{G}_n(\mathcal{F}_{B,s}(x_1^n))$ can be rewritten as $\frac{1}{\sqrt{n}}\mathbb{E}\sup_\theta\langle\theta, \frac{X^T w}{\sqrt{n}}\rangle$ where $w \sim \mathcal{N}(0, I_n)$. Use this fact to **establish** $\mathcal{G}_n(\mathcal{F}_{B,s}(x_1^n)) \leq B\mathbb{E}_w \max_{|S|=s}\|\frac{X_S^\top w}{n}\|_2$.

(b) Define $w_S = \frac{1}{\sqrt{n}} X_S^\top w$. Assuming that for all subsets $S$ of cardinality $s$ we have $\lambda_{\max}\big(\frac{X_S^\top X_S}{n}\big) \leq C^2$, **prove that**

$$\mathbb{P}(\|w_S\|_2 \geq \sqrt{s}C + \delta) \leq \mathrm{e}^{-\frac{\delta^2}{2C^2}}.$$

*Hint*: The Euclidean norm is a Lipschitz function.

(c) Use the preceding parts to **show**

$$\mathcal{G}_n(\mathcal{F}_{B,s}(x_1^n)) \leq \mathcal{O}\left(BC\sqrt{\frac{s\log(\frac{ed}{s})}{n}}\right)$$

(d) We use the set

$$\widetilde{\mathcal{F}}_{B,s} = \left\{ f(\cdot) = \langle \theta, x\rangle : \|\theta\|_0 \leq s, \frac{\|X\theta\|_2}{\sqrt{n}} \leq B \right\}$$

for bounding the prediction error of the best linear sparse approximation. **Prove that**

$$\mathcal{G}_n(\widetilde{\mathcal{F}}_{B,s}(x_1^n)) \leq \mathcal{O}\left(B\sqrt{\frac{s\log(\frac{ed}{s})}{n}}\right).$$

(e) Consider the following model: We observe $y_i = f^\star(x_i) + w_i$ with i.i.d. noise $w_i \sim \mathcal{N}(0,1)$ and $f^\star = \langle \theta^\star, \cdot\rangle \in \mathcal{F}_{\infty,s}$. Consider the (computationally infeasible) estimator $\hat{\theta} \in \arg\min_{\theta:\|\theta\|_0 \leq s} \|y - X\theta\|_2^2$. Use Theorem 13.13 in MW and the previous bound to **prove** that with high probability,

$$\frac{1}{n}\left\|X\theta^\star - X\hat{\theta}\right\|_2^2 \lesssim \frac{s\log(ed/s)}{n}.$$

## Solution

(a) To rewrite the Gaussian complexity we simply rearrange some terms and use the matrix notation for the points $x_1^n$. We then use Cauchy-Schwarz inequality to pull out the supremum of $\|\theta\|_2$ and arrive at the final result. In what follows, we denote with $\odot$ the elementwise product and for a set $S \subseteq [d]$ and the vector $\mathbb{1}_S \in \mathbb{R}^n$ is defined as $(\mathbb{1}_S)_i = 1$, for $i \in S$ and 0 otherwise. It is important to observe that any sparse $\theta$ with $\|\theta\|_0 \leq s$ can be written as $\theta = \theta \odot \mathbb{1}_{S_\theta}$, where $S_\theta \subset [d]$ is the set of indices of the non-zero values of $\theta$ and thus $|S_\theta| \leq s$.

$$\begin{aligned}
\mathcal{G}_n(\mathcal{F}_{B,s}(x_1^n)) &= \frac{1}{n}\mathbb{E}\sup_\theta \sum_{i=1}^n w_i\langle\theta, x_i\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_\theta \sum_{i=1}^n \langle\theta, \frac{w_i x_i}{\sqrt{n}}\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_\theta \langle\theta, \frac{X^T w}{\sqrt{n}}\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S_\theta|=s} \langle\theta\odot\mathbb{1}_{S_\theta}, \frac{X^T w}{\sqrt{n}}\rangle \\
&\overset{CS}{\leq} \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S_\theta|=s} \|\theta\|_2 \frac{\|\mathbb{1}_{S_\theta}^T \odot X^T w\|_2}{\sqrt{n}} \\
&\leq B\mathbb{E}\max_{|S|=s} \frac{\|X_S^T w\|_2}{n}
\end{aligned}$$

(b) A key insight for solving this is to notice that for any $i \in [s]$, $(w_S)_i$ is a linear combination of iid standard Gaussians. This means that it is itself distributed according to a Gaussian $\mathcal{N}(0, \sum_{j=0}^n (X_S)_{ij}^2)$. Moreover it is important to point out that the norm of $w_S$ is C-Lipschitz wrt $w$ because $\|w_S\| = \|\frac{1}{\sqrt{n}} X_S^T w\| \leq C\|w\|$.

This allows us to use Theorem 2.26 from MW from which it follows that $\|w_S\| - \mathbb{E}\|w_S\|$ is sub-Gaussian with parameter $C$.

$$\begin{aligned}
\mathbb{E}[\|w_S\|_2] = \mathbb{E}\left[\|\frac{X_S^T w}{\sqrt{n}}\|_2\right] &= \mathbb{E}\left[\sqrt{\frac{w^T X_S X_S^T w}{n}}\right] \\
&\overset{(i)}{\leq} \sqrt{\mathbb{E}\left[\frac{w^T X_S X_S^T w}{n}\right]} = \sqrt{\mathbb{E}\left[\frac{\text{tr}\left(w^T X_S X_S^T w\right)}{n}\right]} \\
&\overset{(ii)}{=} \sqrt{\mathbb{E}\left[\frac{\text{tr}\left(X_S X_S^T w w^T\right)}{n}\right]} = \sqrt{\frac{\text{tr}\left(X_S X_S^T \mathbb{E}\left[w w^T\right]\right)}{n}} = \sqrt{\frac{\text{tr}\left(X_S X_S^T\right)}{n}} \\
&\overset{(iii)}{=} \sqrt{\sum_{i=0}^{s} \lambda_i\left(\frac{X_S X_S^T}{n}\right)} \leq \sqrt{s\lambda_{max}\left(\frac{X_S X_S^T}{n}\right)} \leq C\sqrt{s}
\end{aligned}$$

This yields the following:

$$\mathbb{P}\left[\|w_S\| \geq C\sqrt{s} + \delta\right] \leq \mathbb{P}\left[\|w_S\| \geq \mathbb{E}[\|w_S\|] + \delta\right] \leq e^{\frac{-\delta^2}{2C^2}}$$

Inequality $(i)$ follows from Jensen, in $(ii)$ we have used the cyclic property of the trace. The identity (iii) uses the fact that the trace of the matrix $A$ is equal to the sum of its eigenvalues, denoted by $\lambda_i(A)$.

(c) For point a) we have proved that the Gaussian complexity is bounded by the expectation of the maximum of a finite collection of random variables. As we stated in part b), the random variable $\|w_S\| - \mathbb{E}\|w_S\|$ is zero-mean and sub-Gaussian with parameter $C$ for any $S$. Notice that there are $\binom{d}{s}$ ways to select the set $S \subset [d]$. We can use the inequality for the expectation of the maximum of sub-Gaussian random variables that we derived in the previous homework, because it applies for random variables that are not independent as well (as is the case here). Thus we arrive at the following:

$$\begin{aligned}
\mathcal{G}_n(\mathcal{F}_{B,s}(x_1^n)) &\leq B\mathbb{E}\max_{|S|=s} \frac{\|X_S^T w\|_2}{n} \\
&= B\frac{C\sqrt{s}}{\sqrt{n}} + B\mathbb{E}\max_{|S|=s} \frac{\|w_S\|_2 - C\sqrt{s}}{\sqrt{n}} \\
&\leq B\frac{C\sqrt{s}}{\sqrt{n}} + B\mathbb{E}\max_{|S|=s} \frac{\|w_S\|_2 - \mathbb{E}\|w_S\|_2}{\sqrt{n}} \\
&\leq B\frac{C\sqrt{s}}{\sqrt{n}} + B\mathcal{O}\left(C\sqrt{\frac{\log\binom{d}{s}}{n}}\right) \\
&\overset{(i)}{\leq} B\frac{C\sqrt{s}}{\sqrt{n}} + B\mathcal{O}\left(C\sqrt{\frac{s\log\left(\frac{ed}{s}\right)}{n}}\right) \\
&\overset{(ii)}{\leq} BC\mathcal{O}\left(\sqrt{\frac{s\log(\frac{ed}{s})}{n}}\right)
\end{aligned}$$

Inequality (i) employs the fact that $\binom{d}{s} \leq \left(\frac{ed}{s}\right)^s$ and inequality (ii) follows from the fact that we ignore constants (hiding them inside the big-O notation) and $\sqrt{s} \leq \sqrt{s\log\left(\frac{ed}{s}\right)}$.

(d) The main idea is to use the same arguments as before in parts a), b) and c) but applied for a different Lipschitz function. From part a) we have that:

$$\mathcal{G}_n(\widetilde{\mathcal{F}}_{B,s}(x_1^n)) = \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta}\langle\frac{X^T w}{\sqrt{n}}, \theta\rangle = \frac{1}{\sqrt{n}}\mathbb{E}\sup_{|S|=s}\sup_{\theta_S}\langle\frac{X_S^T w}{\sqrt{n}}, \theta_S\rangle$$

We can rewrite the inner product to take advantage of the upper bound on $\|\frac{X\theta}{\sqrt{n}}\|_2$.

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

We denoted by $w_S$ the orthogonal projection of $w$ onto $\text{span}(X_S)$ and by $P[X_S] \in \mathbb{R}^{s\times n}$ the projection operator. By the orthogonality of the projection, the norm of $w_S$ is 1-Lipschitz wrt $w$. So given parts b) and c), the conclusion follows by taking $C = \frac{1}{\sqrt{n}}$.

(e) Denote $\partial\widetilde{\mathcal{F}}_{B,s} := \widetilde{\mathcal{F}}_{B,s} - \widetilde{\mathcal{F}}_{B,s}$. By Theorem 13.13, any positive solution $\delta_n > 0$ to

$$\frac{\mathcal{G}_n(\partial\widetilde{\mathcal{F}}_{\delta,s}(x_1^n))}{\delta} \leq \frac{\delta}{2}$$

will be a bound on the prediction error. By the inclusion $\partial\widetilde{\mathcal{F}}_{B,s} \subset \widetilde{\mathcal{F}}_{2B,2s}$ and the previous derivation, we know that

$$\mathcal{G}_n(\partial\widetilde{\mathcal{F}}_{\delta,s}(x_1^n)) \lesssim \delta\sqrt{\frac{s\log(ed/s)}{n}}.$$

Solving $\delta\sqrt{\frac{s\log(ed/s)}{n}}/\delta = \sqrt{\frac{s\log(ed/s)}{n}} \leq \frac{\delta}{2}$ yields that, up to constants,

$$\delta_n^2 \asymp \frac{s\log(ed/s)}{n}$$

bounds the prediction error by Theorem 13.13 in MW.

# 4 A minimax lower bound for dictionary learning

Consider the problem of *dictionary learning*: Suppose you want to solve a regression problem, and are given a finite set of candidate models $\mathcal{F} = \{f_1, \ldots, f_m\}$, called a dictionary. The elements in $\mathcal{F}$ may have been trained before using an independent dataset, or may simply be good candidates for the learning task at hand (think, for instance, of using "foundation models"). For this exercise, we treat them as fixed, deterministic functions, but do not assume anything else about them (beyond boundedness). A central question then becomes: How can we find a predictor that is as good as the best one from the dictionary? And what is the minimal amount of data necessary to achieve this (say, up to $\epsilon$-error)? It turns out that, despite the simple setup, answering these questions yields a rich theory.

In this exercise, we prove a minimax lower bound on the problem from above in the following setting. Let $\mu$ be the Lebesgue measure on $[0,1]$, let $X_1, \ldots, X_n$ be i.i.d. samples from $\mu$, and let

$$Y_i = f^\star(X_i) + \xi_i$$

for some unknown function $f^\star : [0,1] \to \mathbb{R}$ and i.i.d. Gaussian variables $\xi_1, \ldots, \xi_n \sim \mathcal{N}(0,1)$. For any $f : [0,1] \to \mathbb{R}$, denote by $P_f$ the distribution of $(X_1, Y_1)$ if $f^\star = f$, so that $D := ((X_1,Y_1),\ldots,(X_n,Y_n)) \sim P_f^{\otimes n}$ is an i.i.d. dataset from this distribution. Let $\mathcal{F}_0 = \{f : [0,1] \to \mathbb{R} : \|f\|_\infty \leq 1\}$ and $\mathcal{F} = \{f_1, \ldots, f_m\} \subset \mathcal{F}_0$ be an arbitrary set of measurable functions $f : [0,1] \to \mathbb{R}$ where $3 \leq m \leq \exp n$. We further denote $P_j = P_{f_j}$. Under these assumptions, it turns out that the following lower bound is true, and in fact tight up to constant factors.

**Theorem 1** *There exists a constant $c > 0$ such that for all regressors $\widehat{f} \equiv \widehat{f}(D,\mathcal{F}) : [0,1] \to \mathbb{R}$ it holds*

$$\sup_{\substack{f^\star\in\mathcal{F}_0 \\ \mathcal{F}\subset\mathcal{F}_0, |\mathcal{F}|=m}} \left[\mathbb{E}_{D\sim P_{f^\star}^{\otimes n}}\left\|\widehat{f} - f^\star\right\|_{L^2(\mu)}^2 - \min_{f\in\mathcal{F}}\|f - f^\star\|_{L^2(\mu)}^2\right] \geq c\frac{\log m}{n}.$$

In this exercise, we will prove Theorem 1, by first proving a lower bound on general metric spaces, and then applying this bound to the setting from Theorem 1.

(a) **Describe in words** what Theorem 1 tells us about the dictionary learning problem.

(b) Let $m \geq 2$, and let $(\Theta, d)$ be some metric space that contains the elements $\theta_0, \theta_1, \ldots, \theta_m$. Let $P_\theta, \theta \in \Theta$ be a family of probability measures on $\mathcal{X}$, denote $P_j = P_{\theta_j}$, and assume that $P_j \ll P_0$ and $P_0 \ll P_j$ (they are absolutely continuous with respect to each other) for all $j = 1, \ldots, m$.

(i) **Prove that** for any test (that is, measurable function) $\psi : \mathcal{X} \to \{0, \ldots, m\}$ and any $\tau > 0$,

$$\max_{j \in \{0,\ldots,m\}} P_j(\psi \neq j) \geq \frac{\tau m}{1 + \tau m} \left( \frac{1}{m} \sum_{j=1}^{m} P_j \left( \frac{dP_0}{dP_j} \geq \tau \right) \right).$$

Here $P_j(\psi \neq j)$ is short for $P_j(\{x \in \mathcal{X} : \psi(x) \neq j\})$.
*Hint:* To prove this bound, you may find it useful to proceed as follows: First show a lower bound on $P_0(\psi \neq 0)$ using the events $A_j = \{x \in \mathcal{X} : (dP_0/dP_j)(x) \geq \tau\}$. Then combine it with a lower bound on all other $P_j(\psi \neq j)$ using $\max_{j \in \{0,\ldots,m\}} P_j(\psi \neq j) \geq \lambda P_0(\psi \neq 0) + (1-\lambda) \max_{j \in \{1,\ldots,m\}} P_j(\psi \neq j)$ for a well-chosen $\lambda \in [0, 1]$.

(ii) **Conclude** from the previous step that, if for some $s > 0$ it holds $d(\theta_j, \theta_k) \geq 2s > 0$ for all $j \neq k$ and $\frac{1}{m} \sum_{j=1}^{m} \mathrm{KL}(P_j, P_0) \leq \alpha \log m$ for some $0 < \alpha < 1/8$, any estimator $\hat{\theta} : \mathcal{X} \to \Theta$ (measurable function) satisfies

$$\sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{m}}{1 + \sqrt{m}} \left( 1 - 2\alpha - \sqrt{\frac{2\alpha}{\log m}} \right) > 0. \tag{1}$$

Again, $P_\theta(d(\hat{\theta}, \theta) \geq s)$ is short for $P_\theta(\{x \in \mathcal{X} : d(\hat{\theta}(x), \theta) \geq s\})$.
*Hint:* For each estimator, construct a corresponding test and use the previous bound. You may then use the following version of Pinsker's inequality *without proof*: If $P \ll Q$, then

$$\int \max \left\{ \log \frac{dP}{dQ}, 0 \right\} dP \leq \mathrm{KL}(P, Q) + \sqrt{\mathrm{KL}(P, Q)/2}.$$

(c) **Prove** Theorem 1.
*Hint*: Prove the statement for a particular choice of $\mathcal{F}$ consisting orthogonal system in $L^2(\mu)$ where each function has $L^2(\mu)$-norm $\asymp \sqrt{\log(m)/n}$, and use the lower bound from Eq. (1).

## Solution

(a) This question is open ended, but this is one of the take-aways: In the worst case, we can "only" handle an exponential number of models $m = o(\exp(n))$ (in particular, no infinite function class). In the learning settings we have considered in the class so far, we learned how to deal with function classes that are infinite through covering numbers, Rademacher complexity, VC dimension, etc. But for those to be meaningful, we need stronger assumptions on the function class, which are not satisfies in the counterexample of this lower bound.

Arguably, though, the logarithmic dependence on $m$ is more of a blessing than a curse, and it stems from the fact that we only want to do as well as the best in the dictionary. In contrast, suppose you would want to match the performance of the best linear combination of the dictionary functions $\sum_i w_i f_i$ with $w \in \mathbb{R}^m$. Then, using similar arguments, one can show a lower bound of order $m/n$.

(b) (i) We begin by defining the events $A_j = \left\{ x \in \mathcal{X} : \frac{dP_0}{dP_j}(x) \geq \tau \right\} \subset \mathcal{X}$. A calculation shows that

$$\begin{aligned}
P_0(\psi \neq 0) &= \sum_{j=1}^{m} P_0(\psi = j) \\
&\geq \sum_{j=1}^{m} \tau P_j(\{\psi = j\} \cap A_j) \\
&\geq \tau m \left( \frac{1}{m} \sum_{j=1}^{m} P_j(\psi = j) \right) - \tau \sum_{j=1}^{m} P_j(A_j^c) \\
&= \tau m (p_0 - t)
\end{aligned}$$

where we defined the two helper quantities

$$p_0 = \frac{1}{m}\sum_{j=1}^{m} P_j(\psi = j) \qquad \text{and} \qquad t = \frac{1}{m}\sum_{j=1}^{m} P_j\left(\frac{dP_0}{dP_j} < \tau\right).$$

Therefore, we get that for $\lambda = 1/(1+\tau m) \in [0,1]$

$$\begin{aligned}
\max_{j\in\{0,\dots,m\}} P_j(\psi \neq j) &= \max\left\{P_0(\psi \neq 0), \max_{j\in\{1,\dots,m\}} P_j(\psi \neq j)\right\} \\
&\geq \max\left\{\tau m(p_0 - t), 1 - p_0\right\} \\
&\geq \lambda\tau m(p_0 - t) + (1-\lambda)(1-p_0) \\
&= \frac{\tau m}{1+\tau m}(1-t)
\end{aligned}$$

This yields the claim by plugging in the definition of $t$.

(ii) Let $\beta = \alpha \log m$, $\frac{1}{m}\sum_{j=1}^{m} \mathrm{KL}(P_j, P_0) \leq \beta$, and choose $\tau \in (0,1)$. For each estimator $\hat{\theta}$, we can construct the following test:

$$\psi(x) = \operatorname*{arg\,min}_{j\in\{0,\dots,m\}} d(\hat{\theta}(x), \theta_j).$$

Because $d(\theta_j, \theta_k) \geq 2s > 0$, we get that

$$\sup_{\theta\in\Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \max_{j\in\{0,\dots,m\}} P_j(d(\hat{\theta}, \theta_j) \geq s) \geq \max_{j\in\{0,\dots,m\}} P_j(\psi \neq j).$$

We can now lower bound this quantity using the previous bound. Specifically, we have to lower bound the term $\frac{1}{m}\sum_{j=1}^{m} P_j\left(\frac{dP_0}{dP_j} \geq \tau\right)$. To that end, we obtain from Markov's and Pinsker's inequalities that

$$\begin{aligned}
P_j\left(\frac{dP_0}{dP_j} \geq \tau\right) &= 1 - P_j\left(\log\frac{dP_j}{dP_0} > \log\frac{1}{\tau}\right) \\
&\geq 1 - \frac{1}{\log(1/\tau)}\int \max\left\{\log\frac{dP_j}{dP_0}, 0\right\} dP_j \\
&\geq 1 - \frac{1}{\log(1/\tau)}\left(\mathrm{KL}(P_j, P_0) + \sqrt{\mathrm{KL}(P_j, P_0)/2}\right).
\end{aligned}$$

From Jensen's inequality and the assumption, we know that $\frac{1}{m}\sum_{j=1}^{m}\sqrt{\mathrm{KL}(P_j, P_0)} \leq \sqrt{\beta}$, and so it holds that

$$\frac{1}{m}\sum_{j=1}^{m} P_j\left(\frac{dP_0}{dP_j} \geq \tau\right) \geq 1 - \frac{\beta + \sqrt{\beta/2}}{\log(1/\tau)}.$$

The result then follows from the previous derivation with $\tau = 1/\sqrt{m}$. A calculation shows that this is positive for any $0 < \alpha < 1/8$.

(c) Let $\{\phi_j\}_{j=1}^{m}$ be an orthogonal set of functions in $L^2(\mu)$, such that

$$\forall j \neq k: \quad \int \phi_j(x)\phi_k(x)d\mu(x) = 0, \qquad \|\phi_j\|_{L^2(\mu)} = 1 \qquad \text{and} \qquad \|\phi_j\|_\infty \leq 1.$$

For instance, this could be the Rademacher functions defined as $\phi_1 = 1$ and for $j > 1$, $\phi_j(x) = \mathrm{sgn}(\sin(2^{j-1}\pi x))$.

Define for some $0 < \gamma \leq 1$ to be chosen later

$$f_j(x) = \gamma\sqrt{\frac{\log m}{n}}\phi_j(x).$$

Then $\mathcal{F} \subset \mathcal{F}_0$. If $f^\star \in \mathcal{F}$, then $\min_{f\in\mathcal{F}} \|f - f^\star\|_{L^2(\mu)} = 0$, and so

$$\inf_{\widehat{f}} \sup_{\substack{f^\star\in\mathcal{F}_0 \\ \mathcal{F}\subset\mathcal{F}_0, |\mathcal{F}|=m}} \mathbb{E}_{D\sim P_{f^\star}^{\otimes n}} \left\|\widehat{f} - f^\star\right\|_{L^2(\mu)}^2 - \min_{f\in\mathcal{F}} \|f - f^\star\|_{L^2(\mu)}^2 \geq \inf_{\widehat{f}} \sup_{f^\star\in\mathcal{F}} \mathbb{E}_{D\sim P_{f^\star}^{\otimes n}} \left\|\widehat{f} - f^\star\right\|_{L^2(\mu)}^2$$

Notice that by construction, we have that

$$\left\|f_j - f_k\right\|^2_{L^2(\mu)} = \int (f_j - f_k)^2 d\mu = \int f_j^2 d\mu + \int f_k^2 d\mu = 2\gamma^2 \frac{\log m}{n},$$

and moreover,

$$\mathrm{KL}(P_j^{\otimes n}, P_k^{\otimes n}) = n\,\mathrm{KL}(P_j, P_k) = \frac{n}{2}\left\|f_j - f_k\right\|^2_{L^2(\mu)} = \gamma^2 \log m.$$

Hence, we can now employ the lower bound from the previous section, with the metric space $(\Theta, d) = (\mathcal{F}_0, \|\cdot\|_{L^2(\mu)})$, $\theta_j = f_j$, and $s^2 = \gamma^2(\log m)/(2n)$ and $\alpha = \gamma^2$, which we choose to be $\gamma^2 = 1/16$. By Markov's inequality, we get that

$$\begin{aligned}
\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}} \mathbb{E}_{D \sim P_{f^\star}^{\otimes n}} \left\|\widehat{f} - f^\star\right\|^2_{L^2(\mu)} &\geq s^2 \inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}} P_{f^\star}\left(\left\|\widehat{f} - f^\star\right\|^2_{L^2(\mu)} \geq s^2\right) \\
&\geq \frac{\log m}{32n}\left[\frac{\sqrt{m}}{1+\sqrt{m}}\left(1 - \frac{2}{16} - \sqrt{\frac{2/16}{\log m}}\right)\right] \\
&\geq 0.01 \frac{\log m}{n}
\end{aligned}$$

where the last inequality holds because the second factor is minimized at $m = 3$. Since this holds for a particular choice of a subset of $m$ measurable functions, it also holds for the supremum. That concludes the proof with $c = 0.01$.

# 5 (BONUS) Classification error bounds for hard margin support vector machines (SVM)

Recal the material on max margin and SVMs from the lecture. In this exercise, we derive upper bounds for the $0-1$ classification error of hard margin SVMs, also called max-$\ell_2$-margin classifiers, and defined by:

$$\hat{\theta} = \arg\max_{\theta \in \mathbb{R}^d} \min_{(x,y) \in D} y \frac{\langle \theta, x \rangle}{\|\theta\|_2} \tag{2}$$

where $D = \{(x_i, y_i)\}_{i=1}^n$ is the dataset consisting of $n$ input features/label pairs. We remark that the hard-margin SVM is obtained when running logistic regression until convergence on separable data.

For this exercise, we assume that the dataset $D$ is generated by drawing iid samples form the following generative data distribution $(x, y) \sim \mathbb{P}$ where the labels $y$ are uniformly distributed on $\{-1, +1\}$ and the input features are in the form of $x = [yr, \tilde{x}]$ with $\tilde{x} \sim \mathcal{N}(0, I_{d-1})$. Furthermore, let $\gamma$ be the max-$\ell_2$-margin of $D$ in its last $d-1$ coordinates, defined by

$$\gamma = \max_{\theta \in \mathbb{R}^{d-1}} \min_{(x,y) \in D} y \frac{\langle \hat{\theta}, x_{2:d} \rangle}{\|\hat{\theta}\|_2} \tag{3}$$

A simple geometric argument shows that the max-$\ell_2$-margin classifier (up to rescalings) points in the same direction as

$$\hat{\theta} = [r, \gamma\tilde{\theta}] \tag{4}$$

where $\|\tilde{\theta}\|_2 = 1$.

(a) **Compute** the test error of the max-$\ell_2$-margin classifier as a function of $\gamma$ and $r$, i.e. for $(x, y) \sim \mathbb{P}$, what is $P[y\hat{\theta}^\top x < 0]$? What is the dependence on $r$?

(b) Note that $\gamma$ is a random variable dependent on $n$ and $d$. We aim to understand the dependence of the accuracy on $n$ and $d$. Hence, we want to derive non-asymptotic high probability bounds on $\gamma$. Let $\tilde{X} \in \mathbb{R}^{n \times (d-1)}$ be the datamatrix in the last $d-1$ dimensions, i.e. row $i$ in $\tilde{X}$ equals $x_{i,[2:d]}$. **Show that**

$$\gamma \leq \frac{s_{max}(\tilde{X})}{\sqrt{n}} \tag{5}$$

where $s_{max}(\tilde{X})$ is the largest singular value of the datamatrix $\tilde{X}$.

(c) Recall that each entry of $\tilde{X}$ is i.i.d. standard normal Gaussian distributed. To achieve non-asymptotic bounds on $s_{max}(\tilde{X})$, we first prove the following Lemma in two steps.

**Lemma 1** *Let $X \in \mathbb{R}^{n\times d}$ have i.i.d. normally distributed entries. Then, $\mathbb{E}\left[s_{max}(X)\right] < \sqrt{d} + \sqrt{n}$*

(i) Recall that $s_{max}(X) = \max_{u\in\mathcal{S}^{d-1}, v\in\mathcal{S}^{n-1}} \langle Au, v\rangle$ equals the supremum of the Gaussian process $X_{u,v} = \langle Au, v\rangle$. Define $Y_{u,v} = \langle g, u\rangle + \langle h, v\rangle$ where $g \in \mathbb{R}^d$ and $h \in \mathbb{R}^n$ are independent standard normal distributed variables. **Show that**

$$\mathbb{E}\left|X_{u,v} - X_{u',v'}\right|^2 \leq \mathbb{E}\left|Y_{u,v} - Y_{u',v'}\right|^2 \tag{6}$$

(ii) To finish the proof of Lemma 1, we use the following important result: ¡br¿ ¡br¿

**Lemma 2 (Slepian's inequality)** *Consider two Gaussian processes $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ whose increments satisfy Equation (4) for all $((u, v), (u', v')) \in T$. Then $\mathbb{E}[sup_{t\in T} X_t] \leq \mathbb{E}[sup_{t\in T} Y_t]$.*

**Prove** Lemma 1 using Lemma 2.

(d) Use Theorem 2.26 in MW and Lemma 1 to prove that $s_{max}(\tilde{X}) \leq \sqrt{d} + \sqrt{n} + t$ with a probability of at least $1 - 2e^{-t^2/2}$.

## Solution

(a) Using that $\hat{\theta} = [r, \gamma\tilde{\theta}]$, we find that

$$P\left[y\hat{\theta}^\top x < 0\right] = P\left[yrx_1 + \gamma\sum_{i=2}^{d} x_i\tilde{\theta}_{i-1} < 0\right] = P\left[r^2 + \gamma\sum_{i=2}^{d} x_i\tilde{\theta}_{i-1} < 0\right], \tag{7}$$

where we used that $x_1 = yr$. Note that $\sum_{i=2}^{d} x_i\tilde{\theta}_{i-1}$ is a sum of independent Gaussian distributed random variables (RVs). Recall that the sum of two Gaussian distributed RVs is again a Gaussian distributed RV with a variance equaling the square sum of the variances and the mean the sum of the means. Using this fact, we find that $\sum_{i=2}^{d} x_i\tilde{\theta}_{i-1}$ is standard normal distributed since $\sum_{i=1}^{d-1} \tilde{\theta}^2 = 1$ and

$$P\left[y\hat{\theta}^\top x < 0\right] = \Phi\left(-\frac{r^2}{\gamma}\right), \tag{8}$$

where $\Phi$ denotes the cumulative density function of a normal distributed RV. Clearly, the test error is monotonically decreasing in $r$.

(b) We can rewrite the definition of the max-$\ell_2$-margin $\gamma$ as follows:

$$\gamma = \max_{\theta\in\mathbb{R}^{d-1}, \|\theta\|_2=1} \min_{(x,y)\in D} y\langle\tilde{\theta}, x_{2:d}\rangle. \tag{9}$$

Let $1_n$ denote the all ones vector of size $n$ and recall that the labels $y$ are independent of the last $d-1$ coordinates of the input features $x$. Using the definition of $\tilde{X}$ and the fact that a standard normal distributed RV times an independet RV which take the values $+1$ or $-1$ remains a standard normal distributed RV, we can write

$$\begin{aligned} \gamma = &\max_{\theta\in\mathbb{R}^{d-1}, \|\theta\|_2=1} b \\ &\text{subject to } \theta^\top\tilde{X} > b1_n, \end{aligned} \tag{10}$$

where the greater than sign is elementwise. Recall the following important property of the maximal singular value: for any vector $\theta$ with $\|\theta\|_2 = 1$, we have that $\|\theta^\top\tilde{X}\|_2 < s_{max}(\tilde{X})$. Hence, taking the norms on both sides yields $s_{max} > b\|1_n\|_2$ such that $b < s_{max}/\sqrt{n}$.

(c) (i) Using the definition of $X_{u,v}$, we find that

$$\mathbb{E}[|X_{u,v} - X_{u',v'}|^2] = \mathbb{E}\left[\left|\langle Au, v\rangle - \langle Au', v'\rangle\right|^2\right] = \mathbb{E}\left[\left|\sum_{i=1}^{d}\sum_{j=1}^{n} a_{i,j}(u_iv'_j - u'_iv_j)\right|^2\right], \tag{11}$$

where $a_{i,j}$ is the $(i,j)th$ entry of $A$ and normal distributed. Since all entries of $A$ are i.i.d. standard normal distributed the cross terms of the expectation are 0, i.e. $\mathbb{E}[a_{i,j}a_{i',j'}] = 0$ if $i \neq i'$ or $j \neq j'$ and the non-cross terms satisfy $\mathbb{E}[a_{i,j}^2] = 1$. We find that

$$\mathbb{E}\left[\left|X_{u,v} - X_{u',v'}\right|^2\right] = \left|\langle u, v\rangle - \langle u', v'\rangle\right|^2 = \left|\langle u - u', v - v'\rangle\right|^2 \leq \|u - u'\|_2^2 + \|v - v'\|_2^2. \tag{12}$$

Similarly, from the right hand side, where in this case $h, g$ are vectors as entries i.i.d. normal distributed RVs, we find that

$$\mathbb{E}\left[\left|Y_{u,v} - Y_{u',v'}\right|^2\right] = \|u - u'\|_2 + \|v - v'\|_2. \tag{13}$$

(ii) Using Slepian's Lemma, we find that

$$\mathbb{E}\left[s_{max}(X)\right] = \mathbb{E}\left[\max_{(u,v)} X_{u,v}\right] \leq \mathbb{E}\left[\max_{(u,v)} Y_{u,v}\right] = \mathbb{E}\left[\max_{(u,v)} \langle g, u\rangle + \langle h, v\rangle\right]. \tag{14}$$

Clearly, $\max_u \langle g, u\rangle$ is achieved by setting $u = \frac{g}{\|g\|_2}$. Hence, we find that

$$\mathbb{E}[s_{max}(X)] \leq \|g\|_2 + \|h\|_2 = \sqrt{d} + \sqrt{n}. \tag{15}$$

(d) We can write the matrix $X$ as a vector of size $\mathbb{R}^{d\cdot n}$. If the maximum singular value functional is a 1-Lipschitz function, then Theorem 2.26 yields the result directly. Note that for any matrices $A_1, A_2$ of size $\mathbb{R}^{n\times d}$ it holds that

$$\left|s_{max}(A_1) - s_{max}(A_2)\right| = \left|\max_{\theta\in\mathbb{R}^d, \|\theta\|_2=1} \|A_1\theta\|_2 - \max_{\theta'\in\mathbb{R}^d, \|\theta\|_2=1} \|A_2\theta'\|_2\right|. \tag{16}$$

Without loss of generality, we assume that $s_{max}(A_1) > s_{max}(A_2)$. We find

$$\left|\max_{\theta\in\mathbb{R}^d, \|\theta\|_2=1} \|A_1\theta\|_2 - \max_{\theta'\in\mathbb{R}^d, \|\theta\|_2=1} \|A_2\theta'\|_2\right| \leq \max_{\theta\in\mathbb{R}^d, \|\theta\|_2=1} \|A_1\theta\|_2 - \|A_2\theta\|_2 \leq \|A_1 - A_2\|_{\mathcal{F}}, \tag{17}$$

where $\|A_1 - A_2\|_{\mathcal{F}}$ is the Frobenius norm of $A_1 - A_2$. Hence, the maximum singular value functional is a 1-Lipschitz function, which concludes the proof.