GML Fall 23 - HW 2: Generalization bounds

1 Data-dependent generalization bound for hard-margin SVM

In this exercise, we will derive a refined upper bound on the population risk of the hard-margin SVM (support vector machine) solution.

Recall the setting of the first in-lecture exercise, where we analyzed max-margin linear classifiers. The function class of bounded linear functions is given by $\mathcal{F}_B = \{f(x) = \langle w, x\rangle : w \in \mathbb{R}^d, \|w\|_2 \leq B\}$, and we assume that (x, y) come from a joint distribution $\mathbb{P}$ and we are given n training datapoints $\{(x_i, y_i), i \in [n]\}$. We made the following assumption:

Assumption A: Covariates x are bounded, $\mathbb{P}\left(\|x\|_2 \leq D\right) = 1$.

Given $\gamma \geq 0$, we define the margin risk $R^\gamma(f) = \mathbb{P}\left(Yf(X) \leq \gamma\right)$ and its empirical version $R_n^\gamma(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{y_i f(x_i) \leq \gamma}$. In the in-lecture exercise, we proved that with probability at least $1-\delta$, it holds that for all $f \in \mathcal{F}_B$

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

for some constant $c > 0$. This statement holds for all fixed B and γ. However, the bound eq. 1 becomes less and less useful as the margin of the data distribution gets smaller, as in this case $R_n^\gamma(f)$ remains large for any f. If we make an additional margin assumption on the *distribution*, and instead of any f, consider the specific hard-margin SVM solution, we can obtain a more useful bound:

Assumption B: Data is linearly separable, i.e. there exists w^* with the smallest ℓ_2-norm such that $\mathbb{P}\left(y\langle w, x\rangle \geq 1\right) = 1$.

Definition 1. *The hard-margin SVM solution is*

$$f_{SVM} = \langle w_{SVM}, \cdot\rangle \quad \textit{where } w_{SVM} = \textit{arg}\min_w \|w\|_2 \textit{ s.t. } y_i\langle w, x_i\rangle \geq 1$$

In particular, for the hard-margin SVM solution the following holds:

Theorem 1 (Distribution-dependent margin bound)**.** *Under Assumption A and B, with probability at least* $1-\delta$ *it holds that*

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

where $c > 0$ *is some constant.*

a) **Prove** Theorem 1.

The bound of the preceding theorem depends on $\|w^*\|_2$, which is unknown. In the following, we will derive a bound which depends on the norm of the output of SVM; hence it can be calculated from the training set itself. For some training data data, the margin could be larger, and thus we could instantiate eq. 1 with a larger γ to get a tighter bound:

Theorem 2 (Data-dependent margin bound)**.** *Under Assumptions A and B, with probability at least* $1-\delta$ *it holds that*

$$\mathbb{P}\left(Yf_{SVM}(X)<0\right) \leq \frac{2eD\|w_{SVM}\|_2}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)+\log(4\log\|w_{SVM}\|_2)}{n}}.$$

Notice that this bound could be tighter than Theorem 1 since $\|w_{SVM}\|_2 \leq \|w^*\|_2$.

The proof of Theorem 2 is based on the principle called *Structural Risk Minimization* (SRM). SRM aims to alleviate the problem of overfitting which arises when minimizing the empirical risk within a large *preselected* function class $\mathcal{F}$. Instead, we could rewrite a (too) complex function class $\mathcal{F}$ as a nested sequence of function classes with increasing complexity: $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ...$, $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$. One then minimizes the empirical risk penalized by some complexity measure of the function class, so that we can find an optimally complex predictor $f \in \mathcal{F}_k$ for some k (e.g. a polynomial of degree 5 when $\mathcal{F}$ is the space of all polynomials). For more context read Shalev-Schwartz, Ben-David Chapter 7.

We first prove a result on SRM in b) and then use it to prove 2.

b) (Structural Risk Minimization) As above, assume we are given a function class $\mathcal{F}$ which is a union of a nested sequence of function spaces, i.e. $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$ and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ....$ Given a positive sequence $(\delta_1, \delta_2, ...)$ which satisfies $\sum_{i=1}^{\infty} \delta_i \leq \delta$, we define for each k and each function $f \in \mathcal{F}_k$ the event

$$E_{k,f} = \{R^0(f) - R_n^0(f) \leq c\sqrt{\frac{\log 1/\delta_k}{n}} + 2\mathcal{R}_n(\mathcal{F}_k)\}.$$

Assume that for each k, the intersection of these events holds with probability at least $1-\delta_k$, i.e.

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

Prove that with probability at least $1-\delta$ it holds for all $f \in \mathcal{F}$ that

$$R(f) - R_n(f) \leq c\sqrt{\frac{\log(1/\delta_{k(f)})}{n}} + 2\mathcal{R}_n(\mathcal{F}_{k(f)}),$$

where $k(f)$ is the smallest index k s.t. f is contained in $\mathcal{F}_k$.

c) (Data-dependent generalization bound) **Prove** Theorem 2. *Hint*: for the proof, you might want to utilize b) with an appropriate choice of $\mathcal{F}_k$ and δ_k.

1.1 Solution

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

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

We now minimize the right-hand side of the above inequality with respect to γ. For $\gamma \geq 1$, 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,

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

..... **5 points**

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:

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

. . . . . **5 points**

c) Suppose $\|w_{SVM}\| > 1$. Consider the function class $\mathcal{F} = \cup_{k=1}^{\infty}\mathcal{F}_{B_k}$, with $\mathcal{F}_{B_k} = \left\{f(x) = \langle w, x\rangle : w \in \mathbb{R}^d, 1 < \|w\|_2 \leq B_k\right\}$ and $B_k = e^k$. Let $\delta_k = \frac{\delta}{4k^2}$ (which satisfies $\sum_{k=1}^{\infty}\delta_k \leq \delta$). 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:

$$\mathbb{P}(Yf(X) < 0) \leq \frac{2DB_k}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta_k)}{n}}$$

If we now let $k = \lceil \log(\|w\|_2)\rceil$ we have that:

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

. . . . . **5 points**

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 α-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 α-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

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

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, σ^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 α-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

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

2.1 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 β 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:

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

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 β 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_β, with β defined as above. More formally, we have to show that $\|f - f_\beta\|_\infty \leq 2\epsilon L$.

..... **1 pt**

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:

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

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

$\sup_{x\in(k\epsilon,(k+1)\epsilon]} |f(x) - f_\beta(x)| \leq 2\epsilon L$.

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

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 $0 \leq \left| f(x) - f_\beta(k\epsilon) \right| \leq \left| f(x) - f(k\epsilon) + f(k\epsilon) - f_\beta(k\epsilon) \right| \leq 2\epsilon L$.

. . . . . **4 pt**

Remark: A similar argument can be used to show that the same set is a ϵL-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 δ_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 δ_n.

Concretely, for the function class $\mathcal{F}_{\alpha,\gamma}$, we can start by rewriting the integral as follows:

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

. . . . . **3 pt**

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

. . . . . **1 pt**

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).

. . . . . **1 pt**

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 δ_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 δ.

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

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, $k\delta^2 \geq (k+1)^{1-2\alpha}$.

. . . . . **3 pt**

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:

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

. . . . . **1 pt**

Using Corollary 13.7 from MW it follows that in order to satisfy the critical inequality, it suffices to choose a δ 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.

. . . . . **1 pt**

3 Sparse linear functions

We already looked at the complexity of linear function classes with a margin γ and ℓ_2 norm constraint in previous homeworks and lectures. In this exercise we bound the Gaussian complexity of a smaller subset of

ℓ_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 classifiers (as in the computationally infeasible case of sparse linear regression) 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 $\frac{1}{\sqrt{n}} \mathbb{E} \sup_\theta \langle \theta, \frac{X^T w}{\sqrt{n}} \rangle$. 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**

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

d) We use the set

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

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

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

3.1 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 θ 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 θ and thus $|S_\theta| \leq s$.

$$\begin{aligned}
\mathcal{G}_n(\mathcal{F}_{B,s}(x_1^n)) &= \frac{1}{n}\mathbb{E}\sup_\theta \sum_{i=1}^n w_i\langle\theta, x_i\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_\theta \sum_{i=1}^n \langle\theta, \frac{w_i x_i}{\sqrt{n}}\rangle \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_\theta \langle\theta, \frac{X^T w}{\sqrt{n}}\rangle \ldots\ldots \textbf{2 pt} \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S_\theta|=s} \langle\theta \odot \mathbb{1}_{S_\theta}, \frac{X^T w}{\sqrt{n}}\rangle \\
&\overset{CS}{\leq} \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\theta,|S_\theta|=s} \|\theta\|_2 \frac{\|\mathbb{1}_{S_\theta}^T \odot X^T w\|_2}{\sqrt{n}} \\
&\leq B\mathbb{E}\max_{|S|=s} \frac{\|X_S^T w\|_2}{n} \ldots\ldots \textbf{3 pt}
\end{aligned}$$

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 $\mathcal{N}(0, \sum_{j=0}^n (X_S)_{ij}^2)$.
. . . . . **1 pt**

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.
. . . . . **1 pt**

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

. . . . . **2 pt**
This yields the following:

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

. . . . . **1 pt**

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:

$$\begin{aligned}
\mathcal{G}_n(\mathcal{F}_{B,s}(x_1^n)) &\leq B\mathbb{E}\max_{|S|=s}\frac{\|X_S^T w\|_2}{n} \\
&= B\frac{C\sqrt{s}}{\sqrt{n}} + B\mathbb{E}\max_{|S|=s}\frac{\|w_S\|_2 - C\sqrt{s}}{\sqrt{n}} \\
&\leq B\frac{C\sqrt{s}}{\sqrt{n}} + B\mathbb{E}\max_{|S|=s}\frac{\|w_S\|_2 - \mathbb{E}\|w_S\|_2}{\sqrt{n}} \\
&\leq B\frac{C\sqrt{s}}{\sqrt{n}} + B\mathcal{O}\left(C\sqrt{\frac{\log\binom{d}{s}}{n}}\right) \ldots\ldots \textbf{4 pt} \\
&\overset{(i)}{\leq} B\frac{C\sqrt{s}}{\sqrt{n}} + B\mathcal{O}\left(C\sqrt{\frac{s\log\left(\frac{ed}{s}\right)}{n}}\right) \\
&\overset{(ii)}{\leq} BC\mathcal{O}\left(\sqrt{\frac{s\log(\frac{ed}{s})}{n}}\right) \ldots\ldots\ldots\ldots \textbf{1 pt}
\end{aligned}$$

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:

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

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

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

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 $C = \frac{1}{\sqrt{n}}$.

..... **5 pt**

4 Bonus: Classification error bounds for hard margin support vector machines (SVM)

In this exercise, we derive upper bounds for the $0-1$ classification error of hard margin SVMs, also called max-ℓ_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 γ be the max-ℓ_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-ℓ_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-ℓ_2-margin classifier in function of γ 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 γ 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 γ. 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 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 $X \in \mathbb{R}^{(n\times d)}$ be such that all entries are i.i.d. normal distributed. 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:

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 $1 - 2e^{-t^2/2}$.

4.1 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 Φ denotes the cumulative density function of a normal distributed RV. Clearly, the test error is monotonically decreasing in r. ..... **5 pt**

b) We can rewrite the definition of the max-ℓ_2-margin γ 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 θ 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 $b < s_{max}/\sqrt{n}$. ..... **5 pt**

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

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

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

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

..... **3 pt**

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

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

..... **2 pt**

d) We can write the matrix X as a vector of size $\mathbb{R}^{d\times n}$. 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. ..... **5 pt**