GML Fall 25, Homework 2: Generalization bounds

- Please send via email to Tobias Wegel (`twegel@ethz.ch`) by Tuesday, 11.11.25, at 23:59.
- Typeset (Latex or Markdown) and start the answer to each question on a new page.
- Name the file `firstname_lastname.pdf` (e.g., `max_mustermann.pdf`).
- See website for details regarding collaboration and honor code.
- MW refers to Martin Wainwright's book.
- The homeworks are pass/fail: You pass if you properly attempted all questions (except the bonus ones). A genuine attempt means showing your reasoning, intermediate steps, or an explanation of why you are stuck (in case that you are).

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 $x_1, \dots, x_n$, 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 ϵ_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

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

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

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

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 $[0,1]$.

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

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

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

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

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

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 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,

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

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 $\mathcal{F} = \{f_1, \dots, f_m\}$, 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 ϵ-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 μ be the Lebesgue measure on $[0,1]$, let $X_1, \dots, X_n$ be i.i.d. samples from μ, 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 $\xi_1, \dots, \xi_n \sim \mathcal{N}(0,1)$. 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 $D := ((X_1, Y_1), \dots, (X_n, Y_n)) \sim P_f^{\otimes n}$ is an i.i.d. dataset from this distribution. Let $\mathcal{F}_0 = \{f : [0,1] \to \mathbb{R} : \|f\|_\infty \le 1\}$ and $\mathcal{F} = \{f_1, \dots, f_m\} \subset \mathcal{F}_0$ be an arbitrary set of measurable functions $f : [0,1] \to \mathbb{R}$ where $3 \le m \le \exp n$. 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*

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

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 $m \ge 2$, and let (Θ, d) be some metric space that contains the elements $\theta_0, \theta_1, \dots, \theta_m$. 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 $j = 1, \dots, m$.

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

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

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

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

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

Again, $P_\theta(d(\hat{\theta}, \theta) \ge s)$ is short for $P_\theta(\{x \in \mathcal{X} : d(\hat{\theta}(x), \theta) \ge s\})$.
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

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

(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).

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-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 as a 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 $\tilde{X}$.

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