GML Fall 23 - HW 2: Generalization bounds

Instructions:

- If you notice any unclear points, please start a thread on Moodle.
- Please upload to gradescope **by Friday, 9.11.23 by 23:59**.
- Typeset (Latex, Markdown etc.), do not add your name on homeworks and midterms to avoid biases when grading, start the answer to each question on a new page.
- See website for details regarding collaboration and honor code.
- All questions will be picked and graded by the TAs.
- MW refers to Martin Wainwright's book, SSBD to Shalev-Schwartz, Ben-David's book.
- You will receive partial points if you attempt a question and zero if you don't.
- In order to get full points for your own question, please see detailed instructions. Failure to hand in will result in zero points.

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*

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

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*

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

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*

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

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.

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 will 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 $[0,1]$.

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

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

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

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, 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

Read MW Sections 7.1 - 7.4. In previous exercises and lectures, we have looked at the complexity of linear function classes with a margin γ and an ℓ_2 norm constraint. In the first part of this exercise, we will bound the Gaussian complexity of the function class induced by the set of s-sparse, ℓ_2-bounded vectors:

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

Gaussian complexity of this function class 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 $\widetilde{\mathcal{G}}_n(\mathcal{F}_{B,s}(x_1^n))$ can be rewritten as $\frac{1}{\sqrt{n}} \mathbb{E} \sup_{\theta \in \mathcal{F}_{B,s}(x_1^n)} \langle \theta, \frac{X^T w}{\sqrt{n}} \rangle$. Use this fact to **establish**

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

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** that

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

d) In Lecture 8, we have discussed how localized Gaussian complexity enters the prediction error bound in the regression setting via the critical radius. In this exercise, we will prove an upper bound for the localized Gaussian complexity which we have used in the lecture to obtain the prediction error bound for ℓ_0-constrained sparse linear regression:

$$\|\hat{f} - f^*\|_n \leq \mathcal{O}\left(\frac{s\log(ed/s)\log(1/\delta)}{n}\right).$$

Define

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

Prove that

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

4 Bonus: Classification error bounds for hard margin 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 $1 - 2e^{-t^2/2}$.

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

Writing your own problems is a very important way to really learn the material. The famous *Bloom's Taxonomy* that lists the levels of learning is: Remember, Understand, Apply, Analyze, Evaluate, and Create. Using what you know to create is the top-level. As a practical matter, having some practice at trying to create problems also helps you to study for the exam. This is because thinking about how to create an interesting problem forces you to really look at the material from the perspective of those who are going to create the exams. In this exercise you are asked to come up with your own question related to the material learned in the first 9 lectures. The open-ended question should offer a bit of a glimpse into how theory research could feel like in real life. In both cases your answers will help the learning experience of both your peers and yourself (see details below).

5.1 Instructions for writing your own question

In terms of style, your own question can be a technical problem, one including examples, proof derivations, extensions or a conceptual question like the one above etc. Anything that deepens the understanding of the material.

Please include the question and your answer in your homework submission. In addition, also post your question on **Moodle** (see instructions below) so that others can use them to study for the midterm and engage in discussions. You only receive full points if you post your question on Moodle. In the forum, anybody can upvote a question if they find it helpful and engage in discussions. Note that peers will thank you for a good question that helps them to prepare for the exam as well!

Instructions for posting problems: We will use a Moodle forum as the platform to post your proposed problems. Please post each problem in a separate post, in the category **HW2-Own Question**. You are encouraged to engage in discussions in the comments section of each thread and upvote the questions that you liked.