CIS3333: Mathematics of Machine Learning Fall 2025

Homework 4

Due: October 8, 2025, 11:59 PM ET

Submission Instructions: Submit a single PDF (clear scans or photos compiled) to Gradescope and assign pages for each problem. Show key steps and justify answers.

Collaboration & AI Policy: You may discuss approaches with classmates, but write up your own solutions and list collaborators. If you use computational tools (including LLMs) for checking, cite them and ensure the reasoning is your own.

Note: This homework walks you through the generalization bounds we discussed in class in more technical detail and helps you build intuition for the proofs. **The proofs of some parts can be found in the lecture notes, but the goal here is to derive the proofs yourself and give detailed explanations for each step.**

Problem 1: From Excess Risk to Uniform Convergence (6 points)

Our goal in machine learning is to show that empirical risk minimization works: the function $\hat{f}$ we learn from data should have true risk $R(\hat{f})$ close to the best achievable, $R(f^*)$. This problem shows how bounding the *excess risk*, $R(\hat{f}) - R(f^*)$, reduces to bounding a different quantity: the *uniform deviation*, $\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)|$.

a. (2 points) Let $\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{R}(f)$ and $f^* = \arg\min_{f\in\mathcal{F}} R(f)$. Can $\hat{R}(\hat{f}) - \hat{R}(f^*)$ be positive? Explain clearly.

b. (2 points) Starting from the decomposition

$$R(\hat{f}) - R(f^*) = \left(R(\hat{f}) - \hat{R}(\hat{f})\right) + \left(\hat{R}(\hat{f}) - \hat{R}(f^*)\right) + \left(\hat{R}(f^*) - R(f^*)\right),$$

use part (a) to show that

$$R(\hat{f}) - R(f^*) \le \left|R(\hat{f}) - \hat{R}(\hat{f})\right| + \left|\hat{R}(f^*) - R(f^*)\right|.$$

c. (2 points) Prove that

$$\left|R(\hat{f}) - \hat{R}(\hat{f})\right| + \left|\hat{R}(f^*) - R(f^*)\right| \le 2 \cdot \sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)|.$$

Problem 2: Uniform Convergence for Finite Classes (14 points)

Problem 1 showed that bounding excess risk requires bounding $\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)|$. This problem shows how to actually prove such a uniform bound. In lecture, we used Hoeffding's inequality and the union bound. Here, you'll derive the same result using McDiarmid's inequality instead, which is more general and reveals the key proof structure more clearly.

Lemma 1 (McDiarmid's Inequality)**.** [1] *Let $Z_1, \ldots, Z_N$ be independent, not nececessarily identically distributed, random variables taking values in a set $\mathcal{Z}$, and let $g : \mathcal{Z}^N \to \mathbb{R}$. If for all $i \in [N]$ and $z_1, \ldots, z_N \in \mathcal{Z}$,*

$$\sup_{z_i' \in \mathcal{Z}} \left| g(z_1, \ldots, z_i, \ldots, z_N) - g(z_1, \ldots, z_i', \ldots, z_N) \right| \leq c_i,$$

then for any $t > 0$,

$$\mathbb{P}\left(|g(Z_1, \ldots, Z_N) - \mathbb{E}[g(Z_1, \ldots, Z_N)]| \geq t \right) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^N c_i^2} \right).$$

Let $\mathcal{F}$ be a finite hypothesis class and assume $0 \leq \ell(f(x), y) \leq B$ for all $f \in \mathcal{F}$.

a. (4 points) Let our random variables be the training samples $Z_i = (x_i, y_i)$, so our training set is $S = (Z_1, \ldots, Z_N)$. Fix $f \in \mathcal{F}$ and define $g(S) = \hat{R}(f; S) = \frac{1}{N} \sum_{i=1}^N \ell(f(x_i), y_i)$. Show that g satisfies McDiarmid's condition with constants $c_i = B/N$.

b. (2 points) Use the above to bound $\mathbb{P}(B_f)$, where $B_f = \{|\hat{R}(f) - R(f)| > \epsilon\}$ is the event that f has a large gap between its empirical and true risk.

c. (6 points) Now we extend from a bound for one function to a bound for all functions simultaneously. We want to bound the probability that *at least one* function has a large gap between its empirical and true risk.

- Show that the event $\left\{ \sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| > \epsilon \right\}$ is the same as $\bigcup_{f \in \mathcal{F}} B_f$.
- Prove the union bound: show that $\mathbb{P}\left(\bigcup_{f \in \mathcal{F}} B_f \right) \leq \sum_{f \in \mathcal{F}} \mathbb{P}(B_f)$. Then use this and part (a) to conclude that

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| > \epsilon \right) \leq 2|\mathcal{F}| \exp\left(-\frac{2N\epsilon^2}{B^2} \right).$$

- Set this probability equal to δ and solve for ϵ to conclude that with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| \leq B \sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2N}}.$$

d. (2 points) Combine this with Problem 1 to bound the excess risk $R(\hat{f}) - R(f^*)$.

Problem 3: Why We Need the Supremum (10 points)

Problem 1 showed we need to bound $\sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)|$. This may seem excessive if we only care about $\hat{f}$ and f^*. This problem gives a rigorous counterexample showing why a bound for a *fixed* function fails for $\hat{f}$ unless we take a supremum (or otherwise control class complexity).

[1] If you are curious about the proof, you can find it here.

Setup. Let the input domain be $\mathcal{X} = \{0,1\}^d$ with a fixed dimension $d = 50$ (so $|\mathcal{X}| = 2^{50}$), and let the label distribution be independent of the input: $y \sim \text{Bernoulli}(1/2)$, independent of input x. Consider the hypothesis class of *all* binary functions on $\mathcal{X}$:

$$\mathcal{F} = \{f : \{0,1\}^d \to \{0,1\}\},$$

and let the loss function be the 0-1 loss, $\ell(f(x), y) = \mathbb{1}(f(x) \neq y)$. Suppose we draw a training sample $S = ((x_1, y_1), \ldots, (x_N, y_N))$ of size $N = 200$.

a. (2 points) Show that for any $f \in \mathcal{F}$, $R(f) = \mathbb{P}(f(x) \neq y) = 1/2$ under this setup.

b. (3 points) Show that for any realized sample S with no two inputs (x_is) being the same[2], there exists a function $\hat{f} \in \mathcal{F}$ such that $\hat{f}(x_i) = y_i$ for all $i = 1, \ldots, N$. Conclude that $\hat{R}(\hat{f}; S) = 0$.

c. (3 points) Using Problem 2 (a), derive a high-probability bound for a *single, fixed* function f: with probability at least $1 - \delta$, $|\hat{R}(f) - R(f)| \leq \epsilon(N, \delta)$. Compute $\epsilon(N, \delta)$ for $\delta = 0.05$ and $N = 200$.

d. (4 points) From parts (a) and (b), show that $|\hat{R}(\hat{f}; S) - R(\hat{f})| = 0.5$. Why does this not contradict the bound from part (c)? How does the union bound we used in Problem 2 fix this?

(Optional) Problem 4: Beyond Finite Classes (0 points)

The uniform convergence bound fundamentally relies on the hypothesis class being finite. This problem explores what happens when that assumption fails, and hints at how more sophisticated theory can handle infinite classes.

a. (0 points) What happens to the bound $R(\hat{f}) - R(f^*)$ from Problem 2 (c) when $|\mathcal{F}| = \infty$?

b. (0 points) Consider 1D threshold classifiers $\mathcal{F} = \{f_c : x \mapsto \mathbb{1}(x \geq c) \mid c \in \mathbb{R}\}$. This class is infinite, but on N distinct points, it can only produce finitely many labelings $m_{\mathcal{F}}(N)$. Show that $m_{\mathcal{F}}(N) \leq N + 1$.

   Hint: Consider the threshold c between any two points say x_i and x_{i+1}. Does the value of c affect the labeling of x_i and x_{i+1}?

To handle infinite classes, we can replace (up to constants) the size of the class with $m_{\mathcal{F}}(N)$, the number of distinct labelings it can produce on any N-point sample. This allows us to apply the uniform convergence bound to infinite classes[3].

[2] This is true w.p. very close to 1 given our choice of $d = 50$ and $N = 200$. You can try to compute this probability and see how it's close to 1.

[3] If you're interested in learning more about this, I recommend taking CIS 6250.