Problem Set 4

Instructions:

- The topics for this problem set are:
  Unit 6 – Rademacher Complexity.
  Unit 7 – Covering Numbers and Chaining.
- Before you start, make sure you are familiar with the course's Homework Policy.

1. Let $\mathcal{X}$ be a nonempty set, and let $\mathcal{F}$ and $\mathcal{G}$ be a classes of functions $\mathcal{X} \to [-1, 1]$. Prove the following properties of the Rademacher complexity.

(a) Boundedness. $\mathsf{Rad}_m(\mathcal{F}) \le \sup_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}} |f(x)|$. [1 pt]
(b) Singleton. If $|\mathcal{F}| = 1$ then $\mathsf{Rad}_m(\mathcal{F}) = 0$. [1 pt]
(c) Monotonicity. If $\mathcal{F} \subseteq \mathcal{G}$ then $\mathsf{Rad}_m(\mathcal{F}) \le \mathsf{Rad}_m(\mathcal{G})$. [1 pt]
(d) Linear combination. $\mathsf{Rad}_m(\mathcal{F} + \mathcal{G}) = \mathsf{Rad}_m(\mathcal{F}) + \mathsf{Rad}_m(\mathcal{G})$, where

$$\mathcal{F} + \mathcal{G} = \{f + g : \ f \in \mathcal{F} \ \wedge \ g \in \mathcal{G}\}.$$ [2 pts]

(e) Scaling. $\forall c \in \mathbb{R} : \ \mathsf{Rad}_m(c\mathcal{F}) = |c|\mathsf{Rad}_m(\mathcal{F})$. [2 pts]
(f) Convex hull. Assume $\mathcal{F} = \{f_1, \ldots, f_n\}$. Then

$$\mathsf{ConvexHull}(\mathcal{F}) = \left\{ \sum_{i=1}^{n} \alpha_i f_i \;\middle|\; \forall i \in [m] : \ \alpha_i \in [0, 1] \ \wedge \ \sum_{i=1}^{n} \alpha_i \le 1 \right\}$$

satisfies $\mathsf{Rad}_m(\mathsf{ConvexHull}(\mathcal{F})) = \mathsf{Rad}_m(\mathcal{F})$. [2 pts]

2. Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \{0, 1\}$, and let $\mathcal{D}$ be a distribution over $\mathcal{X}$. Prove that

$$\mathsf{Rad}_{\mathcal{D},m}(\mathcal{F}) \le \sqrt{\frac{2\mathsf{VCEnt}_{\mathcal{D},\mathcal{F}}(m)}{m}}.$$ [10 pts]

(Roughly, this shows that learning bounds obtained using Rademacher complexity will be at least as good as bounds obtained using VC entropy.)

3. Let $m \in \mathbb{N}$, let $A \subseteq \{0, 1\}^m$ and let $\rho(x, y) = \|x - y\|_2$. Prove that for any $\varepsilon \in [0, 1]$,

$$\mathsf{Rad}(A) \le 4\varepsilon + \frac{12}{\sqrt{m}} \int_{\varepsilon}^{1} \sqrt{\ln\left(N(A, \varepsilon, \rho)\right)}\, \mathrm{d}\varepsilon.$$ [16 pts]

4. In this question we show that ε-covering numbers can be roughly understood as being the number of bits necessary to specify any given point in a metric space upto precision ε. Formally, consider the following definition.

▶ **Definition 1.** *Let (Ω, ρ) be a metric space, let $n \in \mathbb{N}$ and let $\varepsilon \ge 0$. An encoding of (Ω, ρ) with length n and precision ε is a function $f : \ \{0, 1\}^n \to \Omega$ such that*

$$\forall x \in \Omega \ \exists w \in \{0, 1\}^n : \ \rho(x, f(w)) \le \varepsilon.$$

▶ **Notation 2.** Let $\mathsf{EncodingLength}(\Omega, \rho, \varepsilon)$ denote the integer

$$\min \left\{ n \in \mathbb{N} \;\middle|\; \exists\, f\text{: } f \text{ is an encoding of } (\Omega, \rho) \text{ with length } n \text{ and precision } \varepsilon \right\}.$$

Prove that for any metric space (Ω, ρ) and any $\varepsilon \geq 0$,

$$\log_2 N(\Omega, \rho, \varepsilon) \leq \mathsf{EncodingLength}(\Omega, \rho, \varepsilon) \leq \lceil \log_2 M(\Omega, \rho, \varepsilon) \rceil. \qquad [10 \text{ pts}]$$

5. Let $\mathcal{F}$ be the set of monotone non-decreasing functions $\mathbb{R} \to [0, 1]$. Let $x_1, \ldots, x_m \in \mathbb{R}$. Consider the pseudo-metric space $(\mathcal{F}, \rho)$ where $\rho(f, g) = \max_{i \in [m]} |f(x_i) - g(x_i)|$. Let $\varepsilon > 0$ and $k = \lceil \frac{1}{\varepsilon} \rceil$.

(a) Show that $N_{\text{in}}(\mathcal{F}, \rho, \varepsilon) \leq k^m$. [5 pts]

(b) Show that $N_{\text{in}}(\mathcal{F}, \rho, \varepsilon) \leq (m + 1)^k$. [10 pts]

6. In this question we will prove the following theorem from Unit 6.

▶ **Theorem 3** (McDiarmid's Inequality)**.** *Let Ω be a set and let $f : \Omega^m \to \mathbb{R}$ be a function. Assume there exist $c_1, \ldots, c_m \in \mathbb{R}$ such that f satisfies the following* bounded differences property*:*

$$\forall x_1, \ldots, x_m, x'_1, \ldots, x'_m \in \Omega \ \forall i \in [m] : \\ |f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x'_i, \ldots, x_m)| \leq c_i.$$

Let $X_1, \ldots, X_m$ be independent random variables taking values in Ω. Assume that $\mathbb{E}[|f(X_1, \ldots, X_m)|] < \infty$. Then for any $\varepsilon > 0$,

$$\mathbb{P}\Big[f(X_1, \ldots, X_m) - \mathbb{E}[f(X_1, \ldots, X_m)] \geq \varepsilon\Big] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

and

$$\mathbb{P}\Big[\mathbb{E}[f(X_1, \ldots, X_m)] - f(X_1, \ldots, X_m) \geq \varepsilon\Big] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

We will use the following definitions.

▶ **Notation 4.** Let $v_1, v_2, \ldots$ be a sequence. For any two indices i, j, we write $v_{i:j}$ to denote the sub-sequence $v_i, v_{i+1}, \ldots, v_j$. If $i > j$ then $v_{i:j}$ is an empty sequence.

▶ **Definition 5.** *Let $X_1, \ldots, X_m$ and $Z_0, Z_1, \ldots, Z_m$ be random variables. We say that $Z_{0:m}$ is a* <u>*martingale with respect to $X_{1:m}$*</u> *if the following conditions hold:*

(i) $\forall i \in \{0, \ldots, m\} : \ \mathbb{E}[|Z_i|] < \infty$.

(ii) $\forall i \in \{0, \ldots, m\} :$ *Z_i is a deterministic function of $X_{1:i}$. (In particular, Z_0 is constant.)*

(iii) $\forall i \in \{1, \ldots, m\} : \ \mathbb{E}[Z_i \mid X_{1:i-1}] = Z_{i-1}$.

In other words, a martingale is a sequence of random variables where the differences between consecutive variables are independent and each difference has expectation 0. An example of a martingale is a random walk $Z_{0:m}$ such that $Z_0 = 0$ and for each $i \in [m]$, $Z_i = Z_{i-1} + X_i$, where $X_{1:m}$ is a sequence of random variables chosen independently and uniformly from $\{-1, 1\}$.

(a) Prove the following lemma.

▶ **Lemma 6.** *Let $m \in \mathbb{N}$. Let $Z_{0:m}$ be a martingale with respect to $X_{1:m}$. Suppose there exist real numbers $\sigma_{1:m}$ such that for each $i \in [m]$, the difference $D_i = Z_i - Z_{i-1}$ is conditionally sub-Gaussian with variance factor σ_i^2, namely*

$$\forall \lambda \in \mathbb{R} : \ \ln \mathbb{E}\left[e^{\lambda D_i} \mid X_{1:i-1}\right] \leq \frac{\lambda^2 \sigma_i^2}{2}.$$

Then $Z_m - Z_0$ is sub-Gaussian with variance factor $\sigma^2 = \sum_{i=1}^m \sigma_i^2$.

[10 pts]

Hint: Proceed by induction on m.

(b) In the context of Theorem 3, denote $Z_i = \mathbb{E}[f(X_1, \ldots, X_m) \mid X_{1:i}]$ for all $i \in \{0, \ldots, m\}$. Prove that $Z_{0:m}$ is a martingale with respect to $X_{1:m}$. [10 pts]

Hint: You may use without proof the following version of the law of total expectation: for any real valued random variable Q and random variables A, B,

$$\mathbb{E}\Big[\mathbb{E}[Q \mid A, B] \mid A\Big] = \mathbb{E}[Q \mid A] \quad \text{(a.s.)}.$$

(c) Let $Z_{0:m}$ be as in (6b). Prove that for each $i \in [m]$, the difference $D_i = Z_i - Z_{i-1}$ is conditionally sub-Gaussian with variance factor $c_i^2/4$. [10 pts]

(d) Prove Theorem 3. [5 pts]

(e) Consider the special case where: (1) $f(X_1, \ldots, X_m) = \frac{1}{m}\sum_{i=1}^m X_i$; and (2) $X_{1:m}$ are i.i.d. with $a, b, \mu \in \mathbb{R}$ such that for all $i \in [m]$, $\mathbb{E}[X_i] = \mu$ and $\mathbb{P}[X_i \in [a, b]] = 1$.

Use Theorem 3 to derive an upper bound on $\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^m X_i - \mu\right| \geq \varepsilon\right]$. How does this bound compare with Hoeffding's inequality? [5 pts]