

COMS 4773 Fall 2026 Problem Set 3

Problem 1. The diameter of a metric space $(\mathcal{Z}, \mathbf{d})$ is $\sup_{z, z' \in \mathcal{Z}} \mathbf{d}(z, z')$. For any $L \geq 0$, we say a real-valued function $f: \mathcal{Z} \rightarrow \mathbb{R}$ defined on a metric space $(\mathcal{Z}, \mathbf{d})$ is L -Lipschitz (with respect to the metric $\mathbf{d}$) if, for all $z, z' \in \mathcal{Z}$, we have $f(z) - f(z') \leq L\mathbf{d}(z, z')$.

- (a) For any metric spaces $(\mathcal{Z}_1, \mathbf{d}_1)$ and $(\mathcal{Z}_2, \mathbf{d}_2)$, show that $(\mathcal{Z}, \mathbf{d})$ is a metric space, where $\mathcal{Z} := \mathcal{Z}_1 \times \mathcal{Z}_2$ and $\mathbf{d}: \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is defined by $\mathbf{d}((z_1, z_2), (z'_1, z'_2)) := \mathbf{d}_1(z_1, z'_1) + \mathbf{d}_2(z_2, z'_2)$.

- (b) Let $\mathbf{z}$ be a random variable taking values in $\mathcal{Z}$ for a metric space $(\mathcal{Z}, \mathbf{d})$. Define

$$\mathcal{F}_{\mathbf{z}} := \{f: \mathcal{Z} \rightarrow \mathbb{R} \mid f \text{ is 1-Lipschitz, } \mathbb{E}(f(\mathbf{z})) = 0\}$$

and

$$\mathbb{K}_{\mathbf{z}}(\lambda) := \sup_{f \in \mathcal{F}_{\mathbf{z}}} \ln \mathbb{E}[\exp(\lambda f(\mathbf{z}))], \quad \forall \lambda \in \mathbb{R}.$$

Show that if $(\mathcal{Z}, \mathbf{d})$ has diameter D , then

$$\mathbb{K}_{\mathbf{z}}(\lambda) \leq \frac{D^2 \lambda^2}{8}, \quad \forall \lambda \in \mathbb{R}.$$

- (c) Let $\mathbf{z}_1$ and $\mathbf{z}_2$ be independent random variables, with each $\mathbf{z}_i$ taking values in $\mathcal{Z}_i$ for a metric space $(\mathcal{Z}_i, \mathbf{d}_i)$. Show that

$$\mathbb{K}_{(\mathbf{z}_1, \mathbf{z}_2)}(\lambda) \leq \mathbb{K}_{\mathbf{z}_1}(\lambda) + \mathbb{K}_{\mathbf{z}_2}(\lambda), \quad \forall \lambda \in \mathbb{R},$$

where $\mathbb{K}_{\mathbf{z}_1}$, $\mathbb{K}_{\mathbf{z}_2}$, and $\mathbb{K}_{(\mathbf{z}_1, \mathbf{z}_2)}$ are as defined in Part (b).

Hint: For $f \in \mathcal{F}_{(\mathbf{z}_1, \mathbf{z}_2)}$, define $g: \mathcal{Z}_2 \rightarrow \mathbb{R}$ by $g(z_2) := \mathbb{E}[f(\mathbf{z}_1, \mathbf{z}_2) \mid \mathbf{z}_2 = z_2]$, and for each $z_2 \in \mathcal{Z}_2$, define $h_{z_2}: \mathcal{Z}_1 \rightarrow \mathbb{R}$ by $h_{z_2}(z_1) := f(z_1, z_2) - g(z_2)$. Show that each of g and h_{z_2} is 1-Lipschitz.

- (d) Let $\mathbf{z}_1, \dots, \mathbf{z}_n$ be independent random variables, with each $\mathbf{z}_i$ taking values in $\mathcal{Z}_i$ for a metric space $(\mathcal{Z}_i, \mathbf{d}_i)$ of diameter D_i . Show that for any real-valued 1-Lipschitz function $f: \mathcal{Z}_1 \times \dots \times \mathcal{Z}_n \rightarrow \mathbb{R}$ and any $t > 0$,

$$\Pr(f(\mathbf{z}_1, \dots, \mathbf{z}_n) - \mathbb{E}[f(\mathbf{z}_1, \dots, \mathbf{z}_n)] > t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n D_i^2}\right).$$

- (e) Use the result from Part (d) to prove McDiarmid's inequality (<https://www.cs.columbia.edu/~djhsu/coms4773-s24/lectures/mcdiarmid.pdf>).

Hint: Choose the metric spaces to use with the result from Part (d) in terms of the parameters of the bounded differences property of the given function.

Problem 2. Suppose $\mathbf{x}_1, \dots, \mathbf{x}_n$ are 1-subgaussian mean-zero random variables (not necessarily independent nor identically distributed), and let

$$\mathbf{z} := \max_{i \in [n]} \mathbf{x}_i.$$

In this problem, you will prove a bound on $E[\mathbf{z}]$ two (or three) ways. (Throughout this problem, assume $n > 1$ so $\ln(n) > 0$. This will simplify the form of the bounds.)

- (a) Prove the following: for any $\lambda > 0$,

$$E[\mathbf{z}] \leq \frac{1}{\lambda} K_{\mathbf{z}}(\lambda).$$

Hint: Use Jensen's inequality. Not a trick question.

- (b) Use the result from Part (a) to prove the following: for some absolute constant $C > 0$,

$$E[\mathbf{z}] \leq C \sqrt{\ln(n)}.$$

Hint: For any real numbers $a_1, \dots, a_n$, we have $e^{\max_{i \in [n]} a_i} = \max_{i \in [n]} e^{a_i} \leq \sum_{i=1}^n e^{a_i}$.

Now we start the second way to prove the same bound on $E[\mathbf{z}]$.

- (c) Prove the following: for any $t > 0$,

$$\Pr(\mathbf{z} \geq t) \leq n \cdot e^{-t^2/2}.$$

Hint: Not a trick question.

- (d) Use the result from Part (c) and the fact that $E[\mathbf{z}] \leq E[\max\{0, \mathbf{z}\}] = \int_0^\infty \Pr(\mathbf{z} \geq t) dt$ to prove the following: for some absolute constant $C > 0$,

$$E[\mathbf{z}] \leq C \sqrt{\ln(n)}.$$

Hint: Break the integral into two parts, $[0, w)$ and $[w, \infty)$, for some judicious choice of $w > 0$; use a “trivial” bound for the first part, and use a bound that takes advantage of the lower integral limit for the second part.

Here is a generalization.

- (e) (Optional.) Suppose $\mathbf{x}_1, \mathbf{x}_2, \dots$ is an infinite sequence of 1-subgaussian mean-zero random variables (not necessarily independent nor identically distributed), and let

$$\mathbf{y} := \max_{i \in \mathbb{N}} \frac{\mathbf{x}_i}{\sqrt{1 + \ln(i)}}.$$

Prove the following: for some absolute constant $C > 0$,

$$E[\mathbf{y}] \leq C.$$

Problem 3. Let P be a probability distribution over $\mathcal{X} \times \mathcal{Y}$, $T \in \mathbb{N}$, $\mathbf{S} \sim P^T$, and $\delta \in (0, 1)$.

- (a) Prove the following. There is an absolute constant $C > 0$ such that the following holds. If ERM (with hypothesis class $\mathcal{H}$) is run with training data $\mathbf{S}$, then with probability at least $1 - \delta$, ERM returns a hypothesis $\mathbf{h}$ from $\mathcal{H}$ satisfying

$$\text{err}(\mathbf{h}; P) \leq \text{err}(h^*; P) + C \left(\sqrt{\frac{\text{err}(h^*; P) \log(|\mathcal{H}|/\delta)}{T}} + \frac{\log(|\mathcal{H}|/\delta)}{T} \right)$$

for all $h^* \in \mathcal{H}$.

Hint: The following algebraic fact may be useful: for any non-negative numbers x, y, z ,

$$x \leq y + z\sqrt{x} \implies x \leq y + z^2 + z\sqrt{y}.$$

Two more useful facts are $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and $\sqrt{xy} \leq (x+y)/2$. You may use all of these facts without proof.

- (b) Let $(\mathcal{H}_k)_{k \in \mathbb{N}}$ be an infinite sequence of hypothesis classes of increasing sizes:

$$|\mathcal{H}_1| < |\mathcal{H}_2| < \dots$$

(For example, the classes could be nested $\mathcal{H}_1 \subsetneq \mathcal{H}_2 \subsetneq \dots$.)

Consider the following learning strategy. For each $k \in \mathbb{N}$, let $\hat{\mathbf{h}}_k := \arg \min_{h \in \mathcal{H}_k} \text{err}(h; \mathbf{S})$, and let $\hat{\mathbf{k}} := \arg \min_{k \in \mathbb{N}} \text{err}(\hat{\mathbf{h}}_k; \mathbf{S}) + 2\sqrt{\varepsilon_k}$, where

$$\varepsilon_k := \frac{\ln(2|\mathcal{H}_k|k(k+1)/\delta)}{2T}, \quad \forall k \in \mathbb{N}.$$

Return $\hat{\mathbf{h}} := \hat{\mathbf{h}}_{\hat{\mathbf{k}}}$. (Assume any “ties” in $\arg \min$ are broken arbitrarily.)

- (i) The above learning strategy is a special case of Structural Risk Minimization. However, it may not be obvious how to implement it, especially the “ $\arg \min_{k \in \mathbb{N}}$ ” part. Briefly explain how the strategy can be carried out in finite time. (Assume, for each $k \in \mathbb{N}$, that you have an algorithm for computing both $\hat{\mathbf{h}}_k$ and ε_k .)
- (ii) Show that, with probability at least $1 - \delta$,

$$\text{err}(\hat{\mathbf{h}}; P) \leq \min_{k \in \mathbb{N}} \min_{h \in \mathcal{H}_k} \text{err}(h; P) + 4\sqrt{\varepsilon_k}.$$

- (c) (Optional.) Describe a variant of the learning strategy from Part (b) that returns $\hat{\mathbf{h}}$ with the following guarantee (for some absolute constant $C > 0$): with probability at least $1 - \delta$,

$$\text{err}(\hat{\mathbf{h}}; P) \leq \min_{k \in \mathbb{N}} \min_{h \in \mathcal{H}_k} \text{err}(h; P) + C \left(\sqrt{\text{err}(h; P) \varepsilon_k} + \varepsilon_k \right)$$

where ε_k are as defined in Part (b), and prove that it has this guarantee. The learning strategy should NOT require knowledge of the values $\min_{h \in \mathcal{H}_k} \text{err}(h; P)$ for $k \in \mathbb{N}$.

Problem 4. Let $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ denote a hypothesis class on the domain $\mathcal{X}$.

(a) Let $\mathcal{X} := \{0, 1, 2, \dots, N-1\}$ for $N \in \mathbb{N}$, and

$$\mathcal{H} := \{h: \mathcal{X} \rightarrow \{0, 1\} \mid \exists \theta \in \mathbb{R} \text{ s.t. } h(x) = \mathbb{1}\{x > \theta\} \forall x \in \mathcal{X}\}.$$

What are $\text{Ldim}(\mathcal{H})$ and $\text{VCdim}(\mathcal{H})$? Give your answers in terms of N ; it is fine to use asymptotic notation.

(b) Let $\mathcal{H}^* := \{\mathbf{e}_x \mid x \in \mathcal{X}\}$ be the dual class where

$$\mathbf{e}_x(h) := h(x) \quad \forall h \in \mathcal{H}.$$

Show that

$$\text{VCdim}(\mathcal{H}^*) \leq O(2^{\text{VCdim}(\mathcal{H})}).$$

(c) Let $\mathcal{X} := \mathbb{R}^n$ and $\mathcal{H} := \{h: \mathbb{R}^n \rightarrow \{0, 1\} \mid \exists w \in \mathbb{R}^n \text{ s.t. } h(x) = \mathbb{1}\{\langle w, x \rangle > 0\} \forall x \in \mathbb{R}^n\}$. Show that $\text{VCdim}(\mathcal{H}^*) = \text{VCdim}(\mathcal{H})$.