Information and Coding Theory **Autumn 2022**

Homework 3

Due: November 17, 2022

Note: *You may discuss these problems in groups. However, you must write up your own solutions and mention the names of the people in your group. Also, please do mention any books, papers or other sources you refer to. It is recommended that you typeset your solutions in LaTeX.*

1. **Generalization and Mutual Information** **[2 + 3 = 5 points]**

   A central problem in machine learning is to use (training) data samples to learn predictors, which we hope will perform well on *future* examples which we may have not seen during training, which is known as generalization.

   Consider a randomized algorithm, which receives a sequence i.i.d. training examples, $\overline{z} = (z_1, \ldots, z_n) \sim D^n$ drawn from some unknown distribution D. The algorithm outputs a (possibly random) hypothesis $h \in \mathcal{H}$. We can represent the data and the prediction by a pair of random variables $(\overline{Z}, H)$, which has some joint distribution based on the learning algorithm. Let the error of the predictor h on a data point z be measured by a loss function $\ell(z, h)$. The *training error* is defined as the average loss of the predictor on the training data set, which equals

   $$L(\overline{z}, h) \ = \ \frac{1}{n} \cdot \sum_{i=1}^{n} \ell(z_i, h) \, .$$

   On the other hand, the *test error* is the expected error on a future unknown sample, which equals $\mathbb{E}_{W \sim D}\left[\ell(W, h)\right]$. We are interested in bounding the *generalization error* which is the difference between the training and test error of the (possibly random) predictor given by our learning algorithm. This is given by the expression

   $$\varepsilon_G \ = \ \left| \underset{\overline{Z},H}{\mathbb{E}} \left[\frac{1}{n} \cdot \sum_{i=1}^{n} \ell(Z_i, H) \right] - \underset{W.H}{\mathbb{E}} \left[\ell(W, H) \right] \right| \ = \ \left| \underset{\overline{Z},H}{\mathbb{E}} \left[L(\overline{Z}, H) \right] - \underset{W,H}{\mathbb{E}} \left[\ell(W, H) \right] \right| \, .$$

   Note that while $\overline{Z}$ and H are correlated, the future sample W is independent of H.

   (a) Let $\overline{W} \sim D^n$ be a sequence of i.i.d. samples from D, which is independent of both $\overline{Z}$ and H. Show that the generalization error can be written as

   $$\varepsilon_G \ = \ \left| \underset{\overline{Z},H}{\mathbb{E}} \left[L(\overline{Z}, H) \right] - \underset{\overline{W},H}{\mathbb{E}} \left[L(\overline{W}, H) \right] \right| \, .$$

(b) Suppose that the random variable $L(\overline{W}, H)$ is σ-subgaussian. Show that this implies that

$$\varepsilon_G = \left| \mathop{\mathbb{E}}_{\overline{Z},H} \left[L(\overline{Z}, H) \right] - \mathop{\mathbb{E}}_{\overline{W},H} \left[L(\overline{W}, H) \right] \right| \leq \sqrt{2 \ln 2 \cdot \sigma^2 \cdot I(\overline{Z}; H)}$$

(**Hint:** Use Homework 2.)

Remark: To apply the above result, one can use the fact that if $\ell(W, H)$ is α-subgaussian (which is true, for example, if the loss is a bounded function), one can show that $L(\overline{W}, H)$ is $\frac{\alpha}{\sqrt{n}}$-subgaussian. This gives us a generalization bound of

$$\varepsilon_G \leq \sqrt{\frac{2 \ln 2 \cdot \alpha^2 \cdot I(\overline{Z}; H)}{n}} .$$

which decreases with increasing number of samples n. This can interpreted as saying that as long as the learning algorithm does not "memorize" an amount of information which grows linearly with the amount of training data, the generalization error decreases with the number of training examples.

2. **Minimax rates for denoising.** **[3 × 5 = 15 points]**

We consider the problem of learning a function $f : [0, 1] \to \mathbb{R}$, given noisy samples. For this problem, we will also assume that the function is L-Lipschitz i.e., for any $x_1, x_2 \in [0, 1]$, we have

$$|f(x_1) - f(x_2)| \leq L \cdot |x_1 - x_2| .$$

Note that without any such assumptions, it hard to learn f in a meaningful way even if there is no noise: given the value of f at a few sample points, we have no information about the value of f at other points in the interval.

(a) Let a sample Y be of the form

$$Y = f(X) + G ,$$

where $X \in [0, 1]$ is chosen uniformly at random, and $G \sim N(0, \sigma^2)$ is a one-dimensional Gaussian random variable (independent of X) with mean 0 and variance σ^2. Note that given a value x for the random variable X, Y is simply a Gaussian with mean $f(x)$ and variance σ^2.

Also, note that the distribution of (X, Y) depends on the function f. We denote this distribution as by P_f. Show that for two functions f and g,

$$D(P_f \| P_g) = \frac{\|f - g\|_2^2}{2 \ln 2 \cdot \sigma^2} \qquad \text{where} \quad \|f - g\|_2^2 = \int_0^1 |f(x) - g(x)|^2 \, dx .$$

(**Hint**: Consider the density for Y.)

(b) Consider the problem of finding an "estimator" for the function f given n samples (of the form (X, Y)) from the distribution P_f i.e., we consider the family

$$\Pi \;=\; \left\{ P_f \;\mid\; f : [0,1] \to \mathbb{R} \text{ is } L\text{-Lipschitz} \right\} ,$$

and the property $\theta(P_f) = f$. We consider the loss function

$$\ell(f, g) \;:=\; \|f - g\|_2^2 \;=\; \int_0^1 |f(x) - g(x)|^2 \, dx \,.$$

Let $\{f_a\}_{a \in S}$ be a collection of L-Lipschitz functions such that for any two $a, b \in S$, we have

$$2\delta \;\le\; \|f_a - f_b\|_2 \;\le\; 8\delta \,.$$

Show that the minimax loss for n samples is lower bounded as

$$\mathcal{M}_n(\Pi, \ell) \;\ge\; \delta^2 \cdot \left(1 - \frac{(32\delta^2 n)/(\sigma^2 \cdot \ln 2) + 1}{\log |S|} \right)$$

(c) We will now construct such a family of functions using the "bump" functions $B_\varepsilon : [-1, 1] \to \mathbb{R}$ defined as

$$B_\varepsilon(x) \;=\; \begin{cases} L \cdot (\varepsilon - |x|) & |x| \le \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

Note that this function is bump around the origin of width 2ε. Show that $B(x)$ is L-Lipschitz and (assuming $\varepsilon < 1$)

$$\int_{-1}^{1} (B_\varepsilon(x))^2 dx \;=\; \frac{2\varepsilon^3 L^2}{3} \,.$$

(d) Let $z_1, \ldots, z_m \in (\varepsilon, 1 - \varepsilon)$ be a set of points which are at least 2ε apart. For a set $S \subseteq \{0,1\}^m$, define the function f_a for each $a \in S$ as

$$f_a \;=\; \sum_{i=1}^{m} a_i \cdot B_\varepsilon(x - z_i) \,,$$

f_a is a collection of (non-intersecting) bumps around points z_i depending on which positions i have $a_i = 1$. Show that if $d_H(a, b)$ denotes the Hamming distance between a and b, then

$$\|f_a - f_b\|_2^2 \;=\; \frac{2\varepsilon^3 L^2}{3} \cdot d_H(a, b) \,.$$

(e) Assume that there exists a set $S \subseteq \{0,1\}^m$ such that $|S| \geq 2^{m/8}$ and $d_H(a,b) \geq m/8$ for all $a, b \in S$ (note that this is just a good code). Use this to show that there exists a constant c_0 such that

$$\mathcal{M}_n(\Pi, \ell) \;\geq\; c_0 \cdot \left(\frac{\sigma^2 \cdot L}{n}\right)^{2/3}$$

This bound is known to be tight for Lipschitz functions.

3. **Loaded dice.** **[3 + 4 = 7 points]**

Consider the following game played using a dice: a single dice is rolled and we gain a dollar if the outcome is 2, 3, 4 or 5, and lose a dollar if it's 1 or 6.

(a) What is our expected gain assuming all outcomes in $\{1,2,3,4,5,6\}$ are equally likely.

(b) Find the maximum entropy distribution over the universe $\mathcal{X} = \{1,2,3,4,5,6\}$ such that the expected gain is at least α (say α is greater than the expected gain for the uniform distribution).

4. **Exponential families and maximum entropy.** **[3 + 3 + 2 = 8 points]**

In the class, we proved that for a linear family defined as

$$\mathcal{L} = \left\{ P \;\middle|\; \sum_{x \in \mathcal{X}} P(x) \cdot f_i(x) \;=\; \mathop{\mathbb{E}}_{x \sim P}\left[f_i(x)\right] \;=\; \alpha_i, \; \forall i \in [k] \right\},$$

the maximum entropy distribution P^* is of the form

$$P^*(x) \;=\; \exp\left(\lambda_0 + \sum_{i \in [k]} \lambda_i \cdot f_i(x)\right),$$

where $\lambda_0, \ldots, \lambda_k$ are chosen so that

$$\sum_{x \in \mathcal{X}} P^*(x) \;=\; 1 \qquad \text{and} \qquad \sum_{x \in \mathcal{X}} P^*(x) \cdot f_i(x) \;=\; \alpha_i \;\; \forall i \in [k].$$

In this exercise, we consider the converse. Let $f_1, \ldots, f_k : \mathcal{X} \to \mathbb{R}$ be any functions and Q be *any* a distribution of the form

$$Q(x) \;=\; \exp\left(\lambda_0 + \sum_{i \in [k]} \lambda_i \cdot f_i(x)\right).$$

and let $\alpha_1, \ldots, \alpha_k$ be *defined* as

$$\alpha_i \;:=\; \sum_{x \in \mathcal{X}} Q(x) \cdot f_i(x) \;=\; \mathop{\mathbb{E}}_{x \sim Q}\left[f_i(x)\right].$$

We now consider the linear family defined by $f_1, \ldots, f_k$ and $\alpha_1, \ldots, \alpha_k$.

$$\mathcal{L} = \left\{ P \mid \sum_{x \in \mathcal{X}} P(x) \cdot f_i(x) \; = \; \mathop{\mathbb{E}}_{x \sim P} \left[f_i(x) \right] \; = \; \alpha_i, \; \forall i \in [k] \right\} .$$

Thus, $\mathcal{L}$ is the family of distributions which have the same expected value for the "statistics" $f_1, \ldots, f_k$, as the distribution Q. We will show that Q is indeed the maximum entropy distribution in the family $\mathcal{L}$ (this is a generalization of the often stated fact that the Gaussian distribution has the highest entropy among all distributions with the same covariance).

(a) Show that

$$H(Q) \; = \; -\frac{1}{\ln 2} \cdot \left(\lambda_0 + \sum_{i \in [k]} \lambda_i \cdot \alpha_i \right) .$$

(b) Show that for any distribution $P \in \mathcal{L}$, we have

$$D(P \| Q) \; = \; H(Q) - H(P) .$$

(c) Deduce that Q is the maximum entropy distribution in the family $\mathcal{L}$.