Problem Set 8 **(not graded)**

For the Exercise Session on Dec 23

Last name	First name	SCIPER Nr	Points

Problem 1: Code Extension

Suppose $|\mathcal{U}| \geq 2$. For $n \geq 1$ and a code $c : \mathcal{U} \to \{0,1\}^*$ we define its *n-extension* $c^n : \mathcal{U}^n \to \{0,1\}*$ via $c^n(u^n) = c(u_1) \dots c(u_n)$. In other words $c^n(u^n)$ is the concatenation of the binary strings $c(u_1), \dots, c(u_n)$. A code c is said to be *uniquely decodeable* if for any u^k and $\tilde{u}^m$ with $u^k \neq \tilde{u}^m$, $c^k(u^k) \neq c^m(\tilde{u}^m)$.

(a) Show that if c is uniquely decodable, then for all $n \geq 1$, c^n is injective.

(b) Show that if c is not uniquely decodable, there are u^k and $\tilde{u}^m$ with $u_1 \neq \tilde{u}_1$ and $c^k(u^k) = c^m(\tilde{u}^m)$.

(c) Show that if c is not uniquely decodable, then there is an n for which c^n is not injective. [Hint: try $n = k + m$.]

Problem 2: Elias coding

Let 0^n denote a sequence of n zeros. Consider the code (the subscript U a mnemonic for 'Unary'), $\mathcal{C}_U : \{1, 2, \dots\} \to \{0,1\}^*$ for the positive integers defined as $\mathcal{C}_U(n) = 0^{n-1}$.

(a) Is $\mathcal{C}_U$ injective? Is it prefix-free?

Consider the code (the subscript B a mnenonic for 'Binary'), $\mathcal{C}_B : \{1, 2, \dots\} \to \{0,1\}^*$ where $\mathcal{C}_B(n)$ is the binary expansion of n. I.e., $\mathcal{C}_B(1) = 1$, $\mathcal{C}_B(2) = 10$, $\mathcal{C}_B(3) = 11$, $\mathcal{C}_B(4) = 100$, ... Note that

$$\operatorname{length} \mathcal{C}_B(n) = \lceil \log_2(n+1) \rceil = 1 + \lfloor \log_2 n \rfloor.$$

(b) Is $\mathcal{C}_B$ injective? Is it prefix-free?

With $k(n) = \operatorname{length} \mathcal{C}_B(n)$, define $\mathcal{C}_0(n) = \mathcal{C}_U(k(n))\mathcal{C}_B(n)$.

(c) Show that $\mathcal{C}_0$ is a prefix-free code for the positive integers. To do so, you may find it easier to describe how you would recover $n_1, n_2, \dots$ from the concatenation of their codewords $\mathcal{C}_0(n_1)\mathcal{C}_0(n_2) \dots$.

(d) What is $\operatorname{length}(\mathcal{C}_0(n))$?

Now consider $\mathcal{C}_1(n) = \mathcal{C}_0(k(n))\mathcal{C}_B(n)$.

(e) Show that $\mathcal{C}_1$ is a prefix-free code for the positive integers, and show that $\operatorname{length}(\mathcal{C}_1(n)) = 2 + 2\lfloor \log(1 + \lfloor \log n \rfloor) \rfloor + \lfloor \log n \rfloor \leq 2 + 2 \log(1 + \log n) + \log n$.

Suppose U is a random variable taking values in the positive integers with $\Pr(U=1) \geq \Pr(U=2) \geq \dots$.

(f) Show that $\mathbb{E}[\log U] \leq H(U)$, [Hint: first show $i \Pr(U=i) \leq 1$], and conclude that

$$E[\text{length}\, \mathcal{C}_1(U)] \leq H(U) + 2\log(1+H(U)) + 2.$$

Problem 3: Lower bound on Expected Length

Suppose U is a random variable taking values in $\{1,2,\dots\}$. Set $L = \lfloor \log_2 U \rfloor$. (I.e., $L=j$ if and only if $2^j \leq U < 2^{j+1}$; $j=0,1,2,\dots$.

(a) Show that $H(U|L=j) \leq j$, $j=0,1,\dots$.

(b) Show that $H(U|L) \leq \mathbb{E}[L]$.

(c) Show that $H(U) \leq \mathbb{E}[L] + H(L)$.

(d) Suppose that $\Pr(U=1) \geq \Pr(U=2) \geq \dots$. Show that $1 \geq i \Pr(U=i)$.

(e) With U as in (d), and using the result of (d), show that $\mathbb{E}[\log_2 U] \leq H(U)$ and conclude that $\mathbb{E}[L] \leq H(U)$.

(f) Suppose that N is a random variable taking values in $\{0,1,\dots\}$ with distribution p_N and $\mathbb{E}[N] = \mu$. Let G be a geometric random variable with mean μ, i.e., $p_G(n) = \mu^n/(1+\mu)^{1+n}$, $n \geq 0$.

Show that $H(G) - H(N) = D(p_N \| p_G)$, and conclude that $H(N) \leq g(\mu)$ with $g(x) = (1+x)\log_2(1+x) - x\log_2 x$.

[Hint: Let $f(n,\mu) = -\log_2 p_G(n) = (n+1)\log_2(1+\mu) - n\log_2(\mu)$. First show that $\mathbb{E}[f(G,\mu)] = \mathbb{E}[f(N,\mu)]$, and consequently $H(G) = \sum_n p_N(n)\log_2(1/p_G(n))$.]

(g) Show that for U as in (d) and $g(x)$ as in (f),

$$E[L] \geq H(U) - g(H(U)).$$

[Hint: combine (f), (e), (c).]

(h) Now suppose U is a random variable taking values on an alphabet $\mathcal{U}$, and $c : \mathcal{U} \to \{0,1\}^*$ is an injective code. Show that

$$E[\text{length}\, c(U)] \geq H(U) - g(H(U)).$$

[Hint: the best injective code will label $\mathcal{U} = \{a_1, a_2, a_3, \dots\}$ so that $\Pr(U=a_1) \geq \Pr(U=a_2) \geq \dots$, and assign the binary sequences $\lambda, 0, 1, 00, 01, 10, 11, \dots$ to the letters $a_1, a_2, \dots$ in that order. Now observe that the i'th binary sequence in the list $\lambda, 0, 1, 00, 01, \dots$ is of length $\lfloor \log_2 i \rfloor$.]

Problem 4: Dependence and large error events

In the lecture notes we have seen how to bound the expected generalization error using information measures. With this exercise we will work on large error events and provide bounds on the probabilities of such events. The setting is the same: we observe n iid samples $D = (X_1, \dots, X_n)$ (according to some unknown distribution P) and based on this observation we will choose a hypothesis $w \in W$. We also consider the usual definition of empirical and population risk, *i.e.* given a loss function ℓ, some hypothesis w, $L_D(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, X_i)$, and $L_P(w) = \mathbb{E}_P[\ell(w, X)]$. We are interested in controlling the following quantity:

$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right). \tag{1}$$

(a) Suppose that the loss is such that $\ell(w, x) \in \{0, 1\}$ for every $w \in W$ and $x \in \mathcal{X}$. Suppose also that $|\mathcal{W}| < \infty$, *i.e.*, the number of hypotheses is finite.

1. Show that for every **fixed** $w \in W$ $\Pr\left(|L_P(w) - L_D(w)| > \epsilon\right) \leq 2\exp(-2n\epsilon^2)$;
2. Show that
$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq |W| \cdot 2\exp(-2n\epsilon^2); \tag{2}$$
Hint: denote with $\mathbb{E} = \{(d, w) : |L_P(w) - L_d(w)| > \epsilon\}$.
You have that $\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) = \Pr(E) = \sum_{(w,d)\in E} P(w, d)$.
(be careful: $\Pr\left(|L_P(W) - L_D(W)| > \epsilon | W = w\right)$ is not necessarily $\leq 2\exp(-2n\epsilon^2)$. *Why?*)

(b) Now consider the following information measure, given two discrete random variables X, Y:

$$\mathcal{L}(X \to Y) = \log \sum_{y} \max_{x: P_X(x) > 0} P_{Y|X}(y|x). \tag{3}$$

This quantity is known in the literature as Maximal Leakage and quantifies the leakage of information between X and Y.

1. Show that if the alphabet of Y (denoted with $\mathcal{Y}$) is finite then
$$\mathcal{L}(X \to Y) \leq \log |\mathcal{Y}|,$$
which distributions achieve the bound with equality?
2. It is possible to show that
$$\mathcal{L}(X \to Y) \geq 0,$$
which distributions achieve the bound with equality?
3. Let X be a binary random variable and let Y be an observation of X after passing through a Binary Symmetric Channel with parameter δ. More precisely we have $P_{Y|X=x}(x) = 1 - \delta$, for $x \in \{0, 1\}$.
What is the maximal leakage $\mathcal{L}(X \to Y)$?
Which values of δ allow you to achieve the bounds in (1), (2) with equality?
4. Suppose further that the space of samples $\mathcal{D}$ is finite. Denote with $E_w = \{d : (d, w) \in E\}$, for $w \in \mathcal{W}$; Show that:
$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq \exp(\mathcal{L}(D \to W)) \max_{w \in \mathcal{W}} \Pr(E_w);$$
5. Conclude that
$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq 2\exp(\mathcal{L}(D \to W) - 2n\epsilon^2);$$
6. Compare the two bound retrieved in *(a2)* and *(b4)*, what do you notice? Is one of the two better than the other? When are they equal? What conclusions can you draw?

Problem 5: Tighter Generalization Bound

[10pts] Let $D = X_1, ..., X_n$ iid from an unknown distribution P_X, let $\mathcal{H}$ be a hypothesis space, and $\ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$ be a σ^2-subgaussian loss function for every h. In the lecture we have seen that the generalization error can be upper bounded using the mutual information.

$$|\mathbb{E}_{P_{DH}} [L_{P_X}(H) - L_D(H)]| \leq \sqrt{\frac{2\sigma^2 I(D;H)}{n}}$$

(i) [4 pts] Modify the proof of the *Mutual Information Bound (11.2.2)* to show that if for all $h \in \mathcal{H}$, $\ell(h, X)$ is σ^2-subgaussian in X, then

$$|\mathbb{E}_{P_{DH}} [L_{P_X}(H) - L_D(H)]| \leq \sqrt{\frac{2\sigma^2 \sum_{i=1}^n I(X_i;H)}{n}}.$$

Hint: Recall from the lecture notes that

$$|\mathbb{E}_{P_{DH}} [L_{P_X}(H) - L_D(H)]| \leq \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{P_{X_i H}} [\ell(H, X_i)] - \mathbb{E}_{P_{X_i} P_H} [\ell(H, X_i)] \right|.$$

(ii) [3 pts] Show that, this new bound is never worse than the previous bound by showing that,

$$I(D;H) \geq \sum_{i=1}^n I(X_i;H).$$

(iii) [3 pts] Let us consider an example. Assume that $D = X_1, .., X_n$, $n > 1$, are i.i.d. from $\mathcal{N}(\theta, 1)$, and that we do not know θ. We want to learn θ assuming the loss $\ell(h, x) = \min(1, (h - x)^2)$ (which is bounded) and $\mathcal{H} = \mathbb{R}$. Our learning algorithm outputs $H = \frac{1}{n} \sum_{i=1}^n X_i$. Use the new bound to show that

$$|\mathbb{E}_{P_{DH}} [L_{P_X}(H) - L_D(H)]| \leq \sqrt{\frac{1}{4(n-1)}}.$$

How does the old bound perform in this example?
Hint: Adding independent gaussian random variables, you get a gaussian random variable.

Problem 6: Gibbs Algorithm

Let $\mathcal{X}$ be the sample space, $\mathcal{W}$ the hypothesis space, and let $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_+$ be a corresponding loss function. On a dataset $D = (X_1, X_2, \ldots, X_n)$, the empirical risk for a hypothesis w is given by $L_D(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, X_i)$. We saw in class that $I(D;W)$ can be used to bound the generalization error. Hence, we can use it as a *regularizer* in empirical risk minimization.

(a) First, show that given any joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$ and marginal distribution Q on $\mathcal{Y}$, $D(P_{XY}||P_X P_Y) \leq D(P_{XY}||P_X Q)$.

Since we cannot directly compute $D(P_{DW}||P_D P_W)$, we will use $D(P_{DW}||P_D Q)$ as a proxy, where Q is a distribution on $\mathcal{W}$.

(b) Let

$$P^{\star}_{W|D} = \arg\min_{P_{W|D}} \left(\mathbb{E}[L_D(W)] + \frac{1}{\beta} D(P_{DW}||P_D Q) \right).$$

1. Show that

$$\min_{P_{W|D}} \left(\mathbb{E}[L_D(W)] + \frac{1}{\beta} D(P_{DW}||P_D Q) \right) = \mathbb{E}_D \left[\min_{P_{W|D=d}} \left(\mathbb{E}[L_d(W)] + \frac{1}{\beta} D(P_{W|D=d}||Q) \right) \right].$$

2. Show that the minimizer on the right-hand side $P^{\star}_{W|D=d}$ is given by

$$P^{\star}_{W|D=d} = \frac{e^{-\beta L_d(w)} Q(w)}{\mathbb{E}_Q\left[e^{-\beta L_d(W)}\right]}.$$

This is known in the literature as the Gibbs algorithm. (Hint: Write $\mathbb{E}[\beta L_d(W)] = \mathbb{E}[\log e^{\beta L_d(W)}]$, combine with the KL divergence term and use non-negativity of KL divergence.)

3. Show that $P^{\star}_{W|D=d}$ is $2\beta/n$-differential private if $\ell \in [0,1]$.