Problem Set 8

For the Exercise Session on Dec 17

Last name	First name	SCIPER Nr	Points

Problem 1: Choose the Shortest Description

Suppose $\mathcal{C}_0 : \mathcal{U} \to \{0,1\}^*$ and $\mathcal{C}_1 : \mathcal{U} \to \{0,1\}^*$ are two prefix-free codes for the alphabet $\mathcal{U}$. Consider the code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ defined by

$$\mathcal{C}(u) = \begin{cases} [0, \mathcal{C}_0(u)] & \text{if length}\mathcal{C}_0(u) \leq \text{length}\mathcal{C}_1(u) \\ [1, \mathcal{C}_1(u)] & \text{else.} \end{cases}$$

Observe that $\text{length}(\mathcal{C}(u)) = 1 + \min\{\text{length}(\mathcal{C}_0(u)), \text{length}(\mathcal{C}_1(u))\}$.

(a) Is $\mathcal{C}$ a prefix-free code? Explain.

(b) Suppose $\mathcal{C}_0, \ldots, \mathcal{C}_{K-1}$ are K prefix-free codes for the alphabet $\mathcal{U}$. Show that there is a prefix-free code $\mathcal{C}$ with

$$\text{length}(\mathcal{C}(u)) = \lceil \log_2 K \rceil + \min_{0 \leq k < K-1} \text{length}(\mathcal{C}_k(u)).$$

(c) Suppose we are told that U is a random variable taking values in $\mathcal{U}$, and we are also told that the distribution p of U is one of K distributions $p_0, \ldots, p_{K-1}$, but we do not know which. Using (b) describe how to construct a prefix-free code $\mathcal{C}$ such that

$$\mathbb{E}[\text{length}(\mathcal{C}(U))] \leq \lceil \log_2 K \rceil + 1 + H(U).$$

[Hint: From class we know that for each k there is a prefix-free code $\mathcal{C}_k$ that descibes each letter u with at most $\lceil -\log_2 p_k(u) \rceil$ bits.]

Solution 1. (a) Yes, $\mathcal{C}$ is a prefix-free code. We can prove it by contradiction. Suppose there exist $u, v \in \mathcal{U}$ such that $\mathcal{C}(u)$ is a prefix of $\mathcal{C}(v)$. Then they must start with the same bit. Without loss of generality, let us assume they start with 0, then we have $\mathcal{C}(u) = 0\mathcal{C}_0(u)$ is a prefix of $\mathcal{C}(v) = 0\mathcal{C}_0(v)$. This requires $\mathcal{C}_0(u)$ is a prefix of $\mathcal{C}_0(v)$ which contradicts to $\mathcal{C}_0$ is prefix free code.

(b) Generalizing the given construction, we can construct the code $\mathcal{C}(u)$ for any $u \in \mathcal{U}$ as follows.

$$\mathcal{C}(u) = \text{Bin}(i^*)\mathcal{C}_{i^*}(u) \tag{1}$$

where $i^* = \arg\min_{0 \leq k \leq K-1\}} \text{length}\mathcal{C}_i(u)$ and $\text{Bin}(i^*)$ is the binary representation of number i^*. The length of such code is exactly the given expression and by the same reason in (a), we can show that it is prefix-free.

(c) As the hint suggests, we can use prefix free code $\mathcal{C}_k$ such that $\text{length}(\mathcal{C}_k) \leq \lceil -\log_2 p_k(u) \rceil$ and construct the prefix-free code $\mathcal{C}$ as in [b]. Then we have

$$\begin{aligned}
\text{length}(\mathcal{C}(u)) &= \lceil \log_2 K \rceil + \min_{0 \leq k < K-1} \text{length}(\mathcal{C}_k(u)) && (2)\\
&\leq \lceil \log_2 K \rceil + 1 - \min_{0 \leq k < K-1} \log_2 p_k(u) && (3)\\
&\leq \lceil \log_2 K \rceil + 1 - \log_2 p(u) && (4)
\end{aligned}$$

Taking expectation at both sides, we get that

$$\mathbb{E}[\text{length}(\mathcal{C}(U))] \leq \lceil \log_2 K \rceil + 1 + H(U). \quad (5)$$

Problem 2: 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) 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

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

Solution:

$$\begin{aligned}
||\mathbb{E}_{P_{DH}} [L_{P_X}(H) - L_D(H)]|| &\leq \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{P_{X_i H}} [\ell(H, X_i)] - \mathbb{E}_{P_{X_i} P_H} [\ell(H, X_i)] \right|\\
&\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{P_H} \left[\left| \mathbb{E}_{P_{X_i|H}} [\ell(H, X_i)] - \mathbb{E}_{P_{X_i}} [\ell(H, X_i)] \right| \right] && (11.14)\\
&\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{P_H} \left[\sqrt{2\sigma^2 D(P_{X_i|H} || P_{X_i})} \right] && (11.12)\\
&\leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 \mathbb{E}_{P_H} \left[D(P_{X_i|H} || P_{X_i}) \right]} && (11.15)\\
&= \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 I(X_i;H)} && (11.15)\\
&\leq \sqrt{\frac{2\sigma^2 \sum_{i=1}^n I(X_i;H)}{n}}
\end{aligned}$$

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

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

Solution:

$$I(D;H) = I(X_1,...,X_n;H) = \sum_{i=1}^{n} I(X_i;H|X^{i-1}) \qquad \text{(chain rule for MI)}$$
$$= \sum_{i=1}^{n} I(X_i;HX^{i-1}) \qquad \text{(independence of } X_i\text{'s)}$$
$$\geq \sum_{i=1}^{n} I(X_i;H) \qquad \text{(chain rule and non-negativity of MI)}$$

Therefore the new upper bound is never larger than the previous upper bound.

(iii) 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

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

How does the old bound perform in this example?
Hint: Adding independent gaussian random variables, you get a gaussian random variable.
Solution: Note that the learning algorithm is a deterministic one, that is given a training set D, the learning algorithm outputs a deterministic number. Note also that by property of Gaussian, $H \sim \mathcal{N}(\theta, 1/n)$. Therefore,

$$I(D;H) = h(H) - h(H|D) = \frac{1}{2}\log(2\pi e\frac{1}{n}) - \frac{1}{2}\log(2\pi e 0) = \infty \tag{6}$$

which gives a vacuous bound. Let us compute $I(X_1;H) = h(H) - h(H|X_1)$. Fix x_1, Then,

$$H = \frac{1}{n}x_1 + \frac{1}{n}\sum_{i=2}^{n} X_i \tag{7}$$

which is Gaussian around some mean (which we do not care about) and with variance $(n-1)/n^2$, and note that the variance does not depend on x_1. Therefore the mutual information can be computed as,

$$I(X_1;H) = h(H) - h(H|X_1) = \frac{1}{2}\log(2\pi e\frac{1}{n}) - \frac{1}{2}\log(2\pi e\frac{n-1}{n^2}) = \frac{1}{2}\log(\frac{n}{n-1}) \tag{8}$$

This is true for all $I(X_i;H)$. Also, this loss function is bounded between $0-1$ therefore it is $1/4-$subgaussian. We get the bound,

$$|\mathbb{E}_{P_{DH}}\left[L_{P_X}(H) - L_D(H)\right]| \leq \sqrt{\frac{2\sigma^2\sum_{i=1}^{n} I(X_i;H)}{n}} = \sqrt{\frac{2\sigma^2 n\frac{1}{2}\log(\frac{n}{n-1})}{n}} \tag{9}$$
$$= \sqrt{\frac{1}{4}\log(\frac{n}{n-1})} \tag{10}$$
$$= \sqrt{\frac{1}{4}\log(1+\frac{1}{n-1})} \tag{11}$$
$$\leq \sqrt{\frac{1}{4}\frac{1}{n-1}} \tag{12}$$

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 $2^j \le U < 2^{j+1}$; $j = 0, 1, 2, \dots$.

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

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

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

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

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

(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., $p_G(n) = \mu^n/(1+\mu)^{1+n}$, $n \ge 0$.

Show that $H(G) - H(N) = D(p_N \| p_G)$, and conclude that $H(N) \le g(\mu)$ 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),

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

[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

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

[Hint: the best injective code will label $\mathcal{U} = \{a_1, a_2, a_3, \dots\}$ so that $\Pr(U = a_1) \ge \Pr(U = a_2) \ge \dots$, and assign the binary sequences $\lambda, 0, 1, 00, 01, 10, 11, ...$ 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 $\lfloor \log_2 i \rfloor$.]

Solution 2. *(a)* We know that if $L = j$ then $2^j \le U < 2^{j+1}$, meaning that if $L = j$ then U can take at most $2^{j+1} - 2^j = 2^j$ values. We also know that the entropy of a discrete random variable is at most the logarithm of the number of possible values it assumes. Thus,

$$H(U|L = j) \le \log_2(2^j) = j. \tag{13}$$

(b) We have that:

$$\begin{aligned} H(U|L) &= \sum_j p_L(j) H(U|L = j) && (14)\\ &\le \sum_j p_L(j) j && (15)\\ &= \mathbb{E}[L]. && (16) \end{aligned}$$

(c) We have that:

$$\begin{aligned} H(U) &\le H(UL) && (17)\\ &= H(L) + H(U|L) && (18)\\ &\le H(L) + \mathbb{E}[L]. && (19) \end{aligned}$$

Where (19) follows from *(b)*. Notice that Ineq. (17) is actually an equality, since L is a function of U (and thus, $H(L|U) = 0$).

(d) For random variable U with $\Pr(U = 1) \geq \Pr(U = 2) \geq \dots$, we have

$$1 = \sum_j \Pr(U = j) \geq \sum_{j=1}^{i} \Pr(U = j) \geq i \Pr(U = i). \tag{20}$$

(e) From *(d)* we get that for a given i, $\log_2 i \leq -\log_2 \Pr(U = i)$. Thus:

$$\begin{aligned}
\mathbb{E}[\lfloor \log_2 U \rfloor] &= \sum_i \Pr(U = i) \lfloor \log_2 i \rfloor && (21)\\
&\leq \sum_i \Pr(U = i) \log_2 i && (22)\\
&\leq -\sum_i \Pr(U = i) \log_2 \Pr(U = i) && (23)\\
&= H(U) && (24)
\end{aligned}$$

(f) It is easy to see that, for any integer valued random variable Q:

$$\begin{aligned}
\mathbb{E}[f(Q, \mu)] &= \sum_n ((n+1)\log(1+\mu) - n \log \mu) p_Q(n) && (25)\\
&= \log(1+\mu) \sum_n (n+1) p_Q(n) - \log \mu \sum_n n p_Q(n) && (26)\\
&= \log(1+\mu)(\mathbb{E}[Q] + 1) - \log \mu \mathbb{E}[Q] && (27)
\end{aligned}$$

Thus, since $\mathbb{E}[N] = \mathbb{E}[G]$, we have that $\mathbb{E}[f(N, \mu)] = \mathbb{E}[f(G, \mu)]$.

This implies that $H(G) = \sum_n p_N(n) \log(1/p_G(n))$ as $H(G) = \mathbb{E}_G[-\log(p_G)] = \mathbb{E}_N[-\log(p_G)]$. Computing the difference:

$$\begin{aligned}
H(G) - H(N) &= \sum_n p_N(n) \left(\log \frac{1}{p_G(n)} - \log \frac{1}{p_N(n)} \right) && (28)\\
&= \sum_n p_N(n) \log \left(\frac{p_N(n)}{p_G(n)} \right) && (29)\\
&= D(p_N \| p_G). && (30)
\end{aligned}$$

To conclude:

$$H(N) = H(G) - D(p_N \| p_G) \leq H(G) = (1+\mu)\log(1+\mu) - \mu \log \mu = g(\mu). \tag{31}$$

(g) Let us denote with $\mu = \mathbb{E}[L]$. L takes values in $\{0, 1, \dots\}$ and from *(f)* we know that

$$H(L) \leq g(\mu). \tag{32}$$

From *(e)* we have that

$$\mu = \mathbb{E}[L] \leq H(U). \tag{33}$$

As $g(x)$ a non-decreasing function for $x > 0$ (the derivative is $\log_2(1+x) - \log_2(x) > 0$ for $x > 0$), we can see that

$$g(\mu) = g(\mathbb{E}[L]) \leq g(H(U)). \tag{34}$$

To conclude, from *(c)* we have that:

$$\mathbb{E}[L] \geq H(U) - H(L) \tag{35}$$
$$\geq H(U) - g(\mu) \tag{36}$$
$$\geq H(U) - g(H(U)). \tag{37}$$

(h) Consider the following random variable V taking values in the alphabet $\mathcal{V} = \{1, 2, \ldots\}$ and such that $\Pr(V = i) = \Pr(U = a_i)$ for every $i = 1, 2 \ldots$, *i.e.* a bijective mapping from U to V . We have that $\mathbb{E}[\text{length } c(U)] = \mathbb{E}[\lfloor \log_2 V \rfloor]$. Let us denote with $\hat{L} = \lfloor \log_2 V \rfloor$: this random variable will play the same role played by L until now. We can say that:

$$\mathbb{E}[\text{length } c(U)] = \mathbb{E}[\hat{L}] \tag{38}$$
$$\geq H(V) - g(H(V)) \tag{39}$$
$$= H(U) - g(H(U)). \tag{40}$$

Where (39) follows from *(g)* and (40) is true since V is a bijective function of U and entropy is preserved under bijective mappings.