# Problem Set 8

For the Exercise Session on Dec 17

| Last name | First name | SCIPER Nr | Points |
| --- | --- | --- | --- |
| | | | |

**Problem 1: Prediction and coding**

After observing a binary sequence $u_1, \ldots, u_i$, that contains $n_0(u^i)$ zeros and $n_1(u^i)$ ones, we are asked to estimate the probability that the next observation, $u_{i+1}$ will be 0. One class of estimators are of the form

$$\hat{P}_{U_{i+1}|U^i}(0|u^i) = \frac{n_0(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha} \quad \hat{P}_{U_{i+1}|U^i}(1|u^i) = \frac{n_1(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha}.$$

We will consider the case $\alpha = 1/2$, this is known as the Krichevsky–Trofimov estimator. Note that for $i = 0$ we get $\hat{P}_{U_1}(0) = \hat{P}_{U_1}(1) = 1/2$.

Consider now the joint distribution $\hat{P}(u^n)$ on $\{0,1\}^n$ induced by this estimator,

$$\hat{P}(u^n) = \prod_{i=1}^{n} \hat{P}_{U_i|U^{i-1}}(u_i|u^{i-1}).$$

(a) Show, by induction on $n$ that, for any $n$ and any $u^n \in \{0,1\}^n$,

$$\hat{P}(u_1, \ldots, u_n) \geq \frac{1}{2\sqrt{n}} \left(\frac{n_0}{n}\right)^{n_0} \left(\frac{n_1}{n}\right)^{n_1},$$

where $n_0 = n_0(u^n)$ and $n_1 = n_1(u^n)$.
[Hint: if $0 \leq m \leq n$, then $(1 + 1/n)^{n+1/2} \geq \frac{m+1}{m+1/2}(1 + 1/m)^m$ ]

(b) Conclude that there is a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ such that

$$\text{length}\, \mathcal{C}(u_1, \ldots, u_n) \leq n h_2\left(\frac{n_0(u^n)}{n}\right) + \frac{1}{2}\log n + 2,$$

with $h_2(x) = -x \log x - (1-x)\log(1-x)$.

(c) Show that if $U_1, \ldots, U_n$ are i.i.d. Bernoulli, then

$$\frac{1}{n}\mathbb{E}[\text{length}\, \mathcal{C}(U_1, \ldots, U_n)] \leq H(U_1) + \frac{1}{2n}\log n + \frac{2}{n}$$

**Solution 1.** (a) For $n = 1$, we have $\hat{P}(u_1) = \hat{P}_{U_1}(u_i) = \frac{1}{2}$. If $u_1 = 0$, $n_0(u_1) = 1$ and $n_1(u_1) = 0$. Hence, $\hat{P}(u_1) = \frac{1}{2} = \frac{1}{2\sqrt{n}}(\frac{n_0}{n})^{n_0}(\frac{n_1}{n})^{n_1}$. It is easy to show that for $u_1 = 1$, the inequality still holds with equality.

For $n=k\geq 1$, let's assume that $\hat{P}(u_1,\ldots,u_k)\geq \frac{1}{2\sqrt{k}}\left(\frac{n_0}{k}\right)^{n_0}\left(\frac{n_1}{k}\right)^{n_1}$. For $n=k+1$, it is sufficient to check $u_{k+1}=0$, as the case $u_{i+1}=1$ is the same if we also exchange the roles of $n_0$ and $n_1$. In this case, $n_0(u^{k+1})=n_0(u^k)+1$ and $n_1(u^{k+1})=n_1(u^k)$.

$$\begin{aligned}
\hat{P}(u_1,\ldots,u_k,0) &= \hat{P}_{U_{k+1}|U^k}(0|u^k)\hat{P}_{U^k}(u^k)\\
&\geq \frac{n_0(u^k)+\frac{1}{2}}{n_0(u^k)+n_1(u^k)+1}\frac{1}{2\sqrt{k}}\Big(\frac{n_0(u^k)}{k}\Big)^{n_0(u^k)}\Big(\frac{n_1(u^k)}{k}\Big)^{n_1(u^k)}\\
&= \underbrace{\frac{(k+1)^{k+1/2}}{k^{k+1/2}}\frac{(n_0(u^k)+\frac{1}{2})n_0(u^k)^{n_0(u^k)}}{(n_0(u^k)+1)^{n_0(u^k)+1}}}_{f(u^k)}\frac{1}{2\sqrt{k+1}}\left(\frac{n_0(u^{k+1})}{k+1}\right)^{n_0(u^{k+1})}\left(\frac{n_1(u^{k+1})}{k+1}\right)^{n_1(u^{k+1})}
\end{aligned}$$

We need to show that $f(u^k)\geq 1$ for any $u^k\in\{0,1\}^k$, but this follows from the hint. Therefore, we proved that our induction hypothesis is true for any $n=k+1$, given the condition that $n=k$ cases is satisfied. By induction, we have for any integer $n\geq 1$

$$\hat{P}(u_1,\ldots,u_n)\geq \frac{1}{2\sqrt{n}}\Big(\frac{n_0}{n}\Big)^{n_0}\Big(\frac{n_1}{n}\Big)^{n_1},$$

**Proof the hint**: We need to show that:

$$\left(1+\frac{1}{k}\right)^{k+1/2}\geq \underbrace{\frac{n_0(u^k)+1}{n_0(u^k)+\frac{1}{2}}\left(1+\frac{1}{n_0(u^k)}\right)^{n_0(u^k)}}_{g(n_0(u^k))=g(n_0)}.$$

Now, consider the function $g(x)=\frac{x+1}{x+\frac{1}{2}}(1+\frac{1}{x})^x$ for $x\geq 1$. Since we have that $n_0(u^k)\leq k$, if $g(x)$ is an increasing function then we would have:

$$\begin{aligned}
g(n_0(u^k))\leq g(k)=\frac{k+1}{k+\frac{1}{2}}(1+\frac{1}{k})^k &= \frac{k+1}{(k+\frac{1}{2})\sqrt{1+\frac{1}{k}}}(1+\frac{1}{k})^{k+1/2}\\
&= \frac{\sqrt{k(k+1)}}{k+\frac{1}{2}}(1+\frac{1}{k})^{k+1/2}\\
&< \left(1+\frac{1}{k}\right)^{k+1/2},
\end{aligned}$$

and the result would follow (the last inequality is due to $\sqrt{k(k+1)}<\sqrt{k(k+1)+1/4}=k+1/2$). Hence, we just need to show that $g(x)$ is an increasing function, *i.e.* that $\frac{d}{dx}g(x)\geq 0$. A simple way of doing this is by showing that $\ln g(x)$ is an increasing function, which would then imply the result for $g(x)$. If we compute the differentiation of $\ln g(x)$, we get

$$\frac{d}{dx}\ln g(x)=\frac{1}{x+1}-\frac{1}{x+\frac{1}{2}}+\ln\left(1+\frac{1}{x}\right)-\frac{1}{x+1}=\ln(x+1)-\ln x-\frac{1}{x+\frac{1}{2}}$$

Now observe:

$$\ln(x+1)-\ln x=\int_x^{x+1}\frac{1}{u}du=\mathbb{E}\left[\frac{1}{U}\right],$$

where $U$ is a unifom random variable between $x$ and $x+1$. Also,

$$\frac{1}{x+1/2}=\frac{1}{\mathbb{E}[U]}.$$

Thus:

$$\frac{d}{dx}\ln g(x)=\mathbb{E}\left[\frac{1}{U}\right]-\frac{1}{\mathbb{E}[U]}$$

and the positivity of $\frac{d}{dx}\ln g(x)$ follows from the convexity of the function $u \to 1/u$ (and Jensen's inequality).

(b) Consider the code with length function $L(u^n) = \lceil -\log \hat{P}(u^n) \rceil$. We can check that such code satisfies the Kraft Inequity.

$$\sum_{u^n} 2^{-L(u^n)} = \sum_{u^n} 2^{-\lceil -\log \hat{P}(u^n) \rceil} \le \sum_{u^n} \hat{P}(u^n) = 1$$

Hence, there exists a prefix-free code with length function $L(u^n)$.

$$\begin{aligned}
\text{length}\, \mathcal{C}(u_1, \dots, u_n) = \lceil -\log \hat{P}(u^n) \rceil &\le -\log \hat{P}(u^n) + 1 \\
&\le -\log \left( \frac{1}{2\sqrt{n}} \left(\frac{n_0}{n}\right)^{n_0} \left(\frac{n_1}{n}\right)^{n_1} \right) + 1 \\
&= 2 + \frac{1}{2}\log n + n \left[ -\frac{n_0}{n} \log(\frac{n_0}{n}) - \frac{n_1}{n} \log \frac{n_1}{n} \right] \\
&= 2 + \frac{1}{2}\log n + n h_2(\frac{n_0}{n})
\end{aligned}$$

(c) Let $\Pr(U_i = 0) = \theta$, $\forall i \in \{1, \dots, n\}$. Since $U_1, \dots, U_n$ are i.i.d, we have $\mathbb{E}[n_0(u^n)] = \sum_{i=1}^n \mathbb{E}[n_0(u_i)] = n\theta$ and $H(U_i) = h_2(\theta)$ for all $i$.

$$\begin{aligned}
\mathbb{E}[\text{length}\, \mathcal{C}(U_1, \dots, U_n)] &\le \mathbb{E}[n h_2(\frac{n_0(u^n)}{n}) + \frac{1}{2}\log n + 2] \\
&= n\mathbb{E}[h_2(\frac{n_0(u^n)}{n})] + \frac{1}{2}\log n + 2 \\
&\le n h_2(\frac{\mathbb{E}[n_0(u^n)]}{n}) + \frac{1}{2}\log n + 2 \\
&= n h_2(\theta) + \frac{1}{2}\log n + 2 \\
&= n H(U_1) + \frac{1}{2}\log n + 2
\end{aligned}$$

Therefore,

$$\frac{1}{n}\mathbb{E}[\text{length}\, \mathcal{C}(U_1, \dots, U_n)] \le H(U_1) + \frac{1}{2n}\log n + \frac{2}{n}$$

**Problem 2: 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 $\mu$, 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 $H(N) \leq g(\mu)$ 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 $\lambda, 0, 1, 00, 01, 10, 11, ...$ 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 $2^j \leq U < 2^{j+1}$, 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,

$$H(U|L = j) \leq \log_2(2^j) = j. \tag{1}$$

*(b)* We have that:

$$\begin{aligned} H(U|L) &= \sum_j p_L(j) H(U|L=j) && (2)\\ &\leq \sum_j p_L(j) j && (3)\\ &= \mathbb{E}[L]. && (4) \end{aligned}$$

*(c)* We have that:

$$\begin{aligned} H(U) &\leq H(UL) && (5)\\ &= H(L) + H(U|L) && (6)\\ &\leq H(L) + \mathbb{E}[L]. && (7) \end{aligned}$$

Where (7) follows from *(b)*. Notice that Ineq. (5) 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

$$1 = \sum_j \Pr(U = j) \geq \sum_{j=1}^{i} \Pr(U = j) \geq i \Pr(U = i). \tag{8}$$

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

$$
\begin{aligned}
\mathbb{E}[\lfloor \log_2 U \rfloor] &= \sum_i \Pr(U=i) \lfloor \log_2 i \rfloor && (9)\\
&\le \sum_i \Pr(U=i) \log_2 i && (10)\\
&\le -\sum_i \Pr(U=i) \log_2 \Pr(U=i) && (11)\\
&= H(U) && (12)
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}[f(Q,\mu)] &= \sum_n ((n+1)\log(1+\mu) - n\log\mu) p_Q(n) && (13)\\
&= \log(1+\mu)\sum_n (n+1)p_Q(n) - \log\mu \sum_n n p_Q(n) && (14)\\
&= \log(1+\mu)(\mathbb{E}[Q]+1) - \log\mu\mathbb{E}[Q] && (15)
\end{aligned}
$$

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:

$$
\begin{aligned}
H(G) - H(N) &= \sum_n p_N(n)\left(\log\frac{1}{p_G(n)} - \log\frac{1}{p_N(n)}\right) && (16)\\
&= \sum_n p_N(n)\log\left(\frac{p_N(n)}{p_G(n)}\right) && (17)\\
&= D(p_N \| p_G). && (18)
\end{aligned}
$$

To conclude:

$$
H(N) = H(G) - D(p_N\|p_G) \le H(G) = (1+\mu)\log(1+\mu) - \mu\log\mu = g(\mu). \tag{19}
$$

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

$$
H(L) \le g(\mu). \tag{20}
$$

From *(e)* we have that

$$
\mu = \mathbb{E}[L] \le H(U). \tag{21}
$$

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

$$
g(\mu) = g(\mathbb{E}[L]) \le g(H(U)). \tag{22}
$$

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

$$
\begin{aligned}
\mathbb{E}[L] &\ge H(U) - H(L) && (23)\\
&\ge H(U) - g(\mu) && (24)\\
&\ge H(U) - g(H(U)). && (25)
\end{aligned}
$$

*(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:

$$\begin{aligned}
\mathbb{E}[\text{length } c(U)] &= \mathbb{E}[\hat{L}] && (26)\\
&\geq H(V) - g(H(V)) && (27)\\
&= H(U) - g(H(U)). && (28)
\end{aligned}$$

Where (27) follows from *(g)* and (28) is true since $V$ is a bijective function of $U$ and entropy is preserved under bijective mappings.

**Problem 3: 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 $\sigma^2-$subgaussian loss function for every $h$. In the lecture we have seen that the generalization error can be upper bounded using the mutual information.

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

(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 $\sigma^2-$subgaussian in $X$, then

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

*Hint:* Recall from the lecture notes that

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

**Solution:**

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

(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:**

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

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 $\theta$. We want to learn $\theta$ 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,

$$I(D;H) = h(H) - h(H|D) = \frac{1}{2}\log(2\pi e\frac{1}{n}) - \frac{1}{2}\log(2\pi e 0) = \infty \tag{29}$$

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

$$H = \frac{1}{n}x_1 + \frac{1}{n}\sum_{i=2}^{n} X_i \tag{30}$$

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,

$$I(X_1;H) = h(H) - h(H|X_1) = \frac{1}{2}\log(2\pi e\frac{1}{n}) - \frac{1}{2}\log(2\pi e\frac{n-1}{n^2}) = \frac{1}{2}\log(\frac{n}{n-1}) \tag{31}$$

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,

$$\begin{aligned} |\mathbb{E}_{P_{DH}}\left[L_{P_X}(H) - L_D(H)\right]| \leq \sqrt{\frac{2\sigma^2\sum_{i=1}^{n} I(X_i;H)}{n}} &= \sqrt{\frac{2\sigma^2 n\frac{1}{2}\log(\frac{n}{n-1})}{n}} && (32) \\ &= \sqrt{\frac{1}{4}\log(\frac{n}{n-1})} && (33) \\ &= \sqrt{\frac{1}{4}\log(1+\frac{1}{n-1})} && (34) \\ &\leq \sqrt{\frac{1}{4}\frac{1}{n-1}} && (35) \end{aligned}$$