Foundations of Data Science
Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2022
December 12, 2022

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 $n = k + m$.]

Solution 1. *(a)* Suppose that c^n is not injective, then there exists $u^n \neq \tilde{u}^n$ such that $c^n(u^n) = c^n(\tilde{u}^n)$, hence c is not uniquely decodable, which is a contradiction.

(b) If c is not uniquely decodable, then there exists u^k and $\tilde{u}^m$ such that $c^k(u^k) = c^m(\tilde{u}^m)$. First suppose that u^k is a prefix of $\tilde{u}^m$, then $c(\tilde{u}_{k+1}) = \lambda$ which means that for any $a \in \mathcal{U} \setminus \{\tilde{u}_{k+1}\}$ we have that $c^2(\tilde{u}_{k+1}a) = c^2(a\tilde{u}_{k+1})$ which proves the statement. If $\tilde{u}^m$ is a prefix of u^k a similar reasoning can be applied. Otherwise let p be the first index where $u_p \neq \tilde{u}_p$, then if $u_1^{p-1} = u_1u_2\dots u_{p-1}$, $u_p^k = u_pu_{p+1}\dots u_k$ and $\tilde{u}_p^m = \tilde{u}_p\tilde{u}_{p+1}\dots\tilde{u}_m$ we have that

$$c^{p-1}(u_1^{p-1})c^{k-p+1}(u_p^k) = c^k(u^k) = c^m(\tilde{u}^m) = c^{p-1}(u_1^{p-1})c^{m-p+1}(\tilde{u}_p^m)$$

Hence $c^{k-p+1}(u_p^k) = c^{m-p+1}(\tilde{u}_p^m)$ and $u_p \neq \tilde{u}_p$ which proves the statement.

(c) As shown in subquestion b, if c is not uniquely decodable then there exists u^k and $\tilde{u}^m$ such that $u_1 \neq \tilde{u}_1$ and $c^k(u^k) = c^m(\tilde{u}^m)$, now if $n = m + k$, we have that $c^n(u^k\tilde{u}^m) = c^k(u^k)c^m(\tilde{u}^m) = c^m(\tilde{u}^m)c^k(u^k) = c^n(\tilde{u}^m u^k)$ and since $u_1 \neq \tilde{u}_1$, $u^k\tilde{u}^m \neq \tilde{u}^m u^k$ so c^n is not injective.

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., $\mathcal{C}_B(1) = 1$, $\mathcal{C}_B(2) = 10$, $\mathcal{C}_B(3) = 11$, $\mathcal{C}_B(4) = 100$, $\dots$. Note that

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

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

With $k(n) = \text{length}\, \mathcal{C}_B(n)$, 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 $n_1, n_2, \ldots$ from the concatenation of their codewords $\mathcal{C}_0(n_1)\mathcal{C}_0(n_2)\ldots$.

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

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 $\text{length}(\mathcal{C}_1(n)) = 2 + 2\lfloor \log(1 + \lfloor \log n \rfloor)\rfloor + \lfloor \log n \rfloor \leq 2 + 2\log(1 + \log n) + \log n$.

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

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

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

Solution 2. (a) As $\mathcal{C}_U(n)$ and $\mathcal{C}_U(m)$ are of different lengths when $n \neq m$, the code is injective. It is not prefix free, in particular $\mathcal{C}_U(1) =$ empty-string is a prefix of all other codewords.

(b) As different integers have different binary expansions, $\mathcal{C}_B$ is injective. It is not prefix free, e.g., $\mathcal{C}_B(1) = 1$ is a prefix of all other codewords.

(c) The codeword of $\mathcal{C}_0(n) = \mathcal{C}_U(k(n))\mathcal{C}_B(n)$ is concatenated by two parts. The first part, $\mathcal{C}_U(k(n))$, is the sequence of zeros with length of $k(n) - 1$. And the second part, $\mathcal{C}_B(n)$ is a binary representation for n. For any two different positive integers n_1 and n_2, let's assume that $n_1 < n_2$, which implies that $\text{length}(\mathcal{C}_0(n_1)) \leq \text{length}(\mathcal{C}_0(n_2))$ and $k(n_1) \leq k(n_2)$. We show that $\mathcal{C}_0(n_1)$ is not a prefix of $\mathcal{C}_0(n_2)$.

If $k(n_1) < k(n_2)$, the first $k(n_1)$ bits of $\mathcal{C}_0(n_1)$ are $0 \ldots 01$[1], while the first $k(n_1)$ bits of $\mathcal{C}_0(n_2)$ are all zeros. So in such cases, $\mathcal{C}_0(n_1)$ cannot be a prefix of $\mathcal{C}_0(n_2)$. If $k(n_1) = k(n_2)$, we have $\text{length}(\mathcal{C}_0(n_1)) = \text{length}(\mathcal{C}_0(n_2))$. Although the first $k(n_1)$ bits of $\mathcal{C}_0(n_1)$ and $\mathcal{C}_0(n_2)$ are the same, the second parts, $\mathcal{C}_B(n_1)$ and $\mathcal{C}_B(n_2)$ are different. So $\mathcal{C}_0(n_1)$ cannot be a prefix of $\mathcal{C}_0(n_2)$. Therefore, $\mathcal{C}_0(n_1)$ cannot be a prefix of $\mathcal{C}_0(n_2)$ for any positive integers $n_1 < n_2$. In other words, $\mathcal{C}_0$ is a prefix-free code for the positive integers.

(d)Since $k(n) = \text{length}(\mathcal{C}_B(n)) = 1 + \lfloor \log_2 n \rfloor$,

$$\begin{aligned} \text{length}(\mathcal{C}_0(n)) &= \text{length}(\mathcal{C}_U(k(n))) + \text{length}(\mathcal{C}_B(n)) \\ &= k(n) - 1 + 1 + \lfloor \log_2 n \rfloor \\ &= 1 + 2\lfloor \log_2 n \rfloor \end{aligned}$$

(e) Similarly, as we did in (c), we can show that for any positive integers $n_1 < n_2$, $\mathcal{C}_1(n_1)$ cannot be a prefix of $\mathcal{C}_1(n_2)$. If $k(n_1) < k(n_2)$, $\mathcal{C}_0(k(n_1))$ is not a prefix of $\mathcal{C}_0(k(n_2))$, since $\mathcal{C}_0$ is prefix-free for positive integers. Hence, in such cases, $\mathcal{C}_1(n_1)$ cannot be a prefix of $\mathcal{C}_1(n_2)$. If $k(n_1) = k(n_2)$, we have $\text{length}(\mathcal{C}_1(n_1)) = \text{length}(\mathcal{C}_1(n_2))$. Although the first $\text{length}(\mathcal{C}_0(k(n_1)))$ bits of $\mathcal{C}_1(n_1)$ and $\mathcal{C}_1(n_2)$ are

[1] If $k(n_1) = 1$, then there is no zeros and sequence starts with 1.

the same, the second parts, $\mathcal{C}_B(n_1)$ and $\mathcal{C}_B(n_2)$ are different. So $\mathcal{C}_1(n_1)$ cannot be a prefix of $\mathcal{C}_1(n_2)$. Therefore, $\mathcal{C}_1(n_1)$ cannot be a prefix of $\mathcal{C}_1(n_2)$ for any positive integers $n_1 < n_2$. In other words, $\mathcal{C}_1$ is a prefix-free code for the positive integers.

The length of $\mathcal{C}_1(n)$ can be computed as

$$\begin{aligned}\text{length}(\mathcal{C}_1(n)) &= \text{length}(\mathcal{C}_0(k(n))) + \text{length}(\mathcal{C}_B(n))\\ &= 1 + 2\lfloor \log_2 k(n) \rfloor + k(n)\\ &= 2 + 2\lfloor \log_2(1 + \lfloor \log_2 n \rfloor) \rfloor + \lfloor \log_2 n \rfloor\\ &\leq 2 + 2\log_2(1 + \log_2 n) + \log_2 n\end{aligned}$$

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

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

Taking log at both sides, we get $-\log \Pr(U = i) \geq \log i, \forall i$.

$$\mathbb{E}[\log U] = \sum_i \Pr(U = i) \log i \leq -\sum_i \Pr(U = i) \log \Pr(U = i) = H(U)$$

Using the results from (e) we have

$$\begin{aligned}\mathbb{E}[\text{length}(\mathcal{C}_1(U))] &\leq \mathbb{E}[2 + 2\log(1 + \log U) + \log U]\\ &= 2 + 2\mathbb{E}[\log(1 + \log U)] + \mathbb{E}[\log U]\\ &\leq 2 + 2\log(1 + H(U)) + H(U)\end{aligned}$$

where we used $\mathbb{E}[\log(x)] \leq \log(\mathbb{E}[x])$ for the second term because $\log(x)$ is a concave and monotonically increasing function.

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 $\lambda, 0, 1, 00, 01, 10, 11, ...$ 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 3. *(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:

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

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

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

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:

$$H(N) = H(G) - D(p_N\|p_G) \leq H(G) = (1+\mu)\log(1+\mu) - \mu\log\mu = g(\mu). \quad (19)$$

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

$$H(L) \leq g(\mu). \quad (20)$$

From *(e)* we have that

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

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

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

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

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

(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 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:

$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right). \tag{29}$$

(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
$$\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq |W| \cdot 2\exp(-2n\epsilon^2); \tag{30}$$
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:

$$\mathcal{L}(X \to Y) = \log \sum_y \max_{x: P_X(x) > 0} P_{Y|X}(y|x). \tag{31}$$

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?

Solution 4. *(a1)* For a given $w \in W$, we have, by assumption, that $\ell(w, X_i) - \mathbb{E}[\ell(w, X)]$ is a 0-mean Bernoulli random variable. Hence, $L_D(w) - L_P(w)$ is a $\frac{1}{4}$-sub-Gaussian random variable. By Hoeffding's inequality for σ^2-sub-Gaussian (Lemma 5.5 in the Lecture Notes):

$$\Pr\left(|L_P(w) - L_D(w)| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) = 2\exp(-2n\epsilon^2) \tag{32}$$

(a2) Using the hint, we have that:

$$\begin{aligned}
\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) &= \Pr(E) &(33)\\
&= \sum_{(w,d)\in E} P_{WD}(w,d) &(34)\\
&= \sum_{w\in W}\sum_{d\in E_w} P_{W|D=d}(w)P_D(d) &(35)\\
&\leq \sum_{w\in W}\sum_{d\in E_w} P_D(d) &(36)\\
&= \sum_{w\in W} P_D(E_w) &(37)\\
&\leq \sum_{w\in W} 2\exp(-2n\epsilon^2) &(38)\\
&= 2|W|\exp(-2n\epsilon^2). &(39)
\end{aligned}$$

Where we denoted with $E_w = \{d : (d,w) \in E)\}$. The important thing to notice here is that splitting the summation in this way, and considering E_w for a given $w \in W$ is, in a way, equivalent to fixing the hypothesis w, just like we assumed in *(a1)*. This, along with upper-bounding $P_{W|D=d}(w)$ by 1, allows us to 'ignore' the dependence between W and D. Better bounds can be obtained, as we will soon see, by actually expoliting the dependence and not trivially upper-bounding the conditional probabilities.

(b1) Starting from the expression:

$$\begin{aligned}
\mathcal{L}(X \to Y) &= \log\sum_y \max_{x:P_X(x)>0} P_{Y|X}(y|x) &(40)\\
&\leq \log\sum_y 1 = \log|\mathcal{Y}|. &(41)
\end{aligned}$$

Fixed a distribution P_X a set of distributions $P_{Y|X}$ that achieves the bound with equality is the one induced by a deterministic mapping (*i.e.*, $Y = f(X)$, and f is deterministic).

(b2) Fixed a distribution P_X, if Y is independent from X we have that $P_{Y|X} = P_Y$ and:

$$\begin{aligned}
\mathcal{L}(X \to Y) &= \log\sum_y \max_{x:P_X(x)>0} P_{Y|X}(y|x) &(42)\\
&= \log\sum_y P_Y(y) = \log 1 = 0. &(43)
\end{aligned}$$

(b3) The Maximal Leakage in this case is $\mathcal{L}(X \to Y) = \log(2(1-\delta))$.

With $\delta = 0$, we have a determinstic channel, as $P_{Y|X=0}(0) = 1$ and $P_{Y|X=0}(1) = 0$ and we have that $\mathcal{L}(X \to Y) = \log(2) = \log(|\mathcal{Y}|)$, recovering *(b1)* with equality.

With $\delta = 1/2$, we have that Y is independent of X as $P_{Y|X=0}(0) = P_{Y|X=1}(0) = 1/2$ and we have that $\mathcal{L}(X \to Y) = \log(1) = 0$, recovering *(b2)* with equality.

(b4)

$$
\begin{aligned}
\Pr\left(|L_P(W) - L_D(W)| > \epsilon\right) &= \Pr(E) && (44)\\
&= \sum_{(w,d)\in E} P_{WD}(w,d) && (45)\\
&= \sum_{w\in W}\sum_{d\in E_w} P_{W|D=d}(w)P_D(d) && (46)\\
&\leq \sum_{w\in W}\max_{d:P_D(d)>0} P_{W|D=d}(w)\sum_{d\in E_w} P_D(d) && (47)\\
&= \sum_{w\in W}\max_{d:P_D(d)>0} P_{W|D=d}(w)P_D(E_w) && (48)\\
&\leq \max_{w\in W} P_D(E_w)\sum_{w\in W}\max_{d:P_D(d)>0} P_{W|D=d}(w) && (49)\\
&= \max_{w\in W} P_D(E_w)\exp(\mathcal{L}(D\to W)). && (50)
\end{aligned}
$$

(b5) As noticed in *(a1)*, for every $w \in W$ we have that $\Pr\left(|L_P(w) - L_D(w)| > \epsilon\right) \leq 2\exp(-2n\epsilon^2)$. The bound follows from this and *(b2)*.

(b6) Under the assumption that $|W| < \infty$ as seen in *(b1)*, the Maximal Leakage $\mathcal{L}(D \to W) \leq \log|W|$. Thus, the bound in *(b4)* can be tighter that the one in *(a2)*. More precisely: when $W = f(D)$ where f is deterministic, *i.e.* maximum dependence of W on D, we recover the bound in *(a2)*. When W and D are independent, we have that $\mathcal{L}(D \to W) = 0$ and we recover the Hoeffding's bound. Measuring the dependence (or, in other words, the amount of adaptivity) via Maximal Leakage, allows us to go from the classical Hoeffding's bound to the 'worst-case' scenario provided by the union bound in *(a2)*. The first case represents **no** dependence, the second case represents **maximum** dependence, with a whole spectrum of behaviours in the middle. Another interpretation of the bound (specific to the measure) is the following: if the learning algorithm leaks too much information about the training set, then it will overfit. If you limit the amount of leakage, then you generalize (and potentially, with an exponentially decaying bound).

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.

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

(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

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

Hint: Recall from the lecture notes that

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

Solution:

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

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

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

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) [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

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

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,

$$I(D;H)=h(H)-h(H|D)=\frac{1}{2}\log(2\pi e\frac{1}{n})-\frac{1}{2}\log(2\pi e0)=\infty \tag{51}$$

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

$$H = \frac{1}{n}x_1 + \frac{1}{n}\sum_{i=2}^{n} X_i \tag{52}$$

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,

$$I(X_1; H) = h(H) - h(H|X_1) = \frac{1}{2}\log(2\pi e\frac{1}{n}) - \frac{1}{2}\log(2\pi e\frac{n-1}{n^2}) = \frac{1}{2}\log(\frac{n}{n-1}) \tag{53}$$

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,

$$\begin{aligned}
|\mathbb{E}_{P_{DH}}[L_{P_X}(H) - L_D(H)]| \le \sqrt{\frac{2\sigma^2\sum_{i=1}^{n} I(X_i; H)}{n}} &= \sqrt{\frac{2\sigma^2 n\frac{1}{2}\log(\frac{n}{n-1})}{n}} && (54)\\
&= \sqrt{\frac{1}{4}\log(\frac{n}{n-1})} && (55)\\
&= \sqrt{\frac{1}{4}\log(1+\frac{1}{n-1})} && (56)\\
&\le \sqrt{\frac{1}{4}\frac{1}{n-1}} && (57)
\end{aligned}$$

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 $\mathcal{Y}$, $D(P_{XY}||P_X P_Y) \le D(P_{XY}||P_X Q)$.

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 $\ell \in [0,1]$.

Solution 5. *(a)* For any marginal distribution Q on $\mathcal{Y}$,

$$D(P_{XY}||P_XP_Y) - D(P_{XY}||P_XQ) = \sum_{x,y} P_{XY}(x,y)\left(\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} - \log\frac{P_{XY}(x,y)}{P_X(x)Q(y)}\right) \tag{58}$$

$$= \sum_{x,y} P_{XY}(x,y)\log\frac{Q(y)}{P_Y(y)} \tag{59}$$

$$= \sum_{y} P_Y(y)\log\frac{Q(y)}{P_Y(y)} \tag{60}$$

$$\overset{(*)}{\leq} \log\sum_{y} P_Y(y)\frac{Q(y)}{P_Y(y)} \tag{61}$$

$$= \log\sum_{y} Q(y) = 0 \tag{62}$$

where $(*)$ is because $\log(x)$ is a concave function of x.

(b1)

$$\min_{P_{W|D}}\left(\mathbb{E}[L_D(W)] + \frac{1}{\beta}D(P_{DW}||P_DQ)\right) \tag{63}$$

$$= \min_{P_{W|D}}\left(\mathbb{E}_D[\mathbb{E}[L_D(W)|D=d]] + \frac{1}{\beta}\sum_{w,d} P_{W|D}(w|d)P_D(d)\log\frac{P_{W|D}(w|d)P_D(d)}{P_D(d)Q}\right) \tag{64}$$

$$= \min_{P_{W|D}}\left(\mathbb{E}_D[\mathbb{E}[L_D(W)|D=d]] + \frac{1}{\beta}\sum_{w,d} P_{W|D}(w|d)P_D(d)\log\frac{P_{W|D}(w|d)}{Q}\right) \tag{65}$$

$$= \min_{P_{W|D}}\left(\mathbb{E}_D[\mathbb{E}[L_D(W)|D=d]] + \mathbb{E}_D[\frac{1}{\beta}D(P_{W|D}||Q)|D=d]\right) \tag{66}$$

$$= \mathbb{E}_D\left[\min_{P_{W|D=d}}\left(\mathbb{E}[L_d(W)] + \frac{1}{\beta}D(P_{W|D=d}||Q)\right)\right] \tag{67}$$

(b2) Given $D=d$, we know that $P_W(w)=\sum_{d'} P_{W|D}(w|d')P_D(d')=P_{W|D}(w|d)$.

$$\arg\min_{P_{W|D=d}} \left(\mathbb{E}[L_d(W)]+\frac{1}{\beta}D(P_{W|D=d}||Q)\right) \tag{68}$$

$$= \arg\min_{P_{W|D=d}} \left(\mathbb{E}[\beta L_d(W)]+D(P_{W|D=d}||Q)\right) \tag{69}$$

$$= \arg\min_{P_{W|D=d}} \left(\mathbb{E}[\log e^{\beta L_d(W)}]+D(P_{W|D=d}||Q)\right) \tag{70}$$

$$= \arg\min_{P_{W|D=d}} \left(\sum_w \log e^{\beta L_d(w)} P_W(w) + \sum_w P_{W|D}(w|d)\log\frac{P_{W|D}(w|d)}{Q(w)}\right) \tag{71}$$

$$= \arg\min_{P_{W|D=d}} \left(\sum_w \log e^{\beta L_d(w)} P_{W|D}(w|d) + \sum_w P_{W|D}(w|d)\log\frac{P_{W|D}(w|d)}{Q(w)}\right) \tag{72}$$

$$= \arg\min_{P_{W|D=d}} \left(\sum_w P_{W|D}(w|d)(\log e^{\beta L_d(w)} + \log\frac{P_{W|D}(w|d)}{Q(w)})\right) \tag{73}$$

$$= \arg\min_{P_{W|D=d}} \left(\sum_w P_{W|D}(w|d)\log\frac{P_{W|D}(w|d)}{Q(w)e^{-\beta L_d(w)}}\right) \tag{74}$$

$$= \arg\min_{P_{W|D=d}} \left(\sum_w P_{W|D}(w|d)\log\frac{P_{W|D}(w|d)}{Q(w)e^{-\beta L_d(w)}}\frac{\mathbb{E}_Q[e^{-\beta L_d(W)}]}{\mathbb{E}_Q[e^{-\beta L_d(W)}]}\right) \tag{75}$$

$$= \arg\min_{P_{W|D=d}} D\left(P_{W|D}||\frac{Q(w)e^{-\beta L_d(w)}}{\mathbb{E}_Q[e^{-\beta L_d(W)}]}\right) - \log\mathbb{E}_Q[e^{-\beta L_d(W)}] \tag{76}$$

$$= \arg\min_{P_{W|D=d}} D\left(P_{W|D}||\frac{Q(w)e^{-\beta L_d(w)}}{\mathbb{E}_Q[e^{-\beta L_d(W)}]}\right) \tag{77}$$

$$= \frac{Q(w)e^{-\beta L_d(w)}}{\mathbb{E}_Q[e^{-\beta L_d(W)}]} \tag{78}$$

The reason why we added $\mathbb{E}_Q[e^{-\beta L_d(W)}]$ as a normalization term is $P_{W|D}$ has to be a valid pmf, i.e. $\sum_w P_{W|D}(w|d)=1$. However, the scaled version of Q, $Q(w)e^{-\beta L_d(w)}$, may not be a valid pmf.

(b3) Suppose d and d' differ at j-th entry only. Hence,

$$e^{-\beta L_d(w)}e^{\beta L_{d'}(w)} = e^{-\frac{\beta}{n}(e(w_j,X_j)-e(w_j,X'_j))} \leq e^{\beta/n} \tag{79}$$

Similarly,

$$\frac{\mathbb{E}_Q[e^{-\beta L_d(w)}]}{\mathbb{E}_Q[e^{-\beta L_{d'}(w)}]} \leq \frac{\mathbb{E}_Q[e^{\frac{\beta}{n}}e^{-\beta L_{d'}(w)}]}{\mathbb{E}_Q[e^{-\beta L_{d'}(w)}]} \leq e^{\frac{\beta}{n}} \tag{80}$$

Thus, we have $\frac{P^*_{W|D=b}}{P^*_{W|D=d'}} \leq e^{2\beta/n}$ and $P^*_{W|D=d}$ is $2\beta/n-$ differential private if $l\in[0,1]$.