ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Handout 31 Information Theory and Coding
Solutions to graded homework Dec. 17, 2024

Notation:

$\exp_2(a) := 2^a$.

for a positive integer m, $[m] := \{1, \dots, m\}$.

PROBLEM 1. Suppose $\mathrm{scr}\colon \mathcal{A} \times \mathcal{B} \to \mathbb{R}$. We will call scr to be a 'score function'. Suppose $a(1), \dots, a(m)$ are elements of $\mathcal{A}$. For this collection, the *maximum score decoder* is the function $\mathrm{dec}_{\mathrm{scr}}\colon \mathcal{B} \to \{0, 1, \dots, m\}$ given by

$$\mathrm{dec}_{\mathrm{scr}}(b) = \arg\max_{i \in [m]} \mathrm{scr}(a(i), b),$$

if there is a unique maximizer; if not we set $\mathrm{dec}_{\mathrm{scr}}(b) = 0$ (i.e., in the case of a tie the decoder says "I can't decide").

(a) Suppose two score functions scr and scr′ are such that $\mathrm{scr}(a, b) - \mathrm{scr}'(a, b)$ is only a function of b. What can you say about $\mathrm{dec}_{\mathrm{scr}}$ and $\mathrm{dec}_{\mathrm{scr}'}$?

Solution: Assume $\mathrm{scr}(a, b) = \mathrm{scr}'(a, b) + f(b)$ for some function f. Since

$$\max_{i \in [m]} \mathrm{scr}'(a(i), b) = \max_{i \in [m]} \{\mathrm{scr}(a(i), b) + f(b)\} = f(b) + \max_{i \in [m]} \mathrm{scr}(a(i), b),$$

the set of elements achieving the maximum is the same whether we use scr or scr′, and hence $\mathrm{dec}_{\mathrm{scr}'}(b) = \mathrm{dec}_{\mathrm{scr}}(b)$.

(b) Suppose two score functions scr and scr′ are such that $\mathrm{scr}(a, b) = \lambda\,\mathrm{scr}'(a, b)$ for a given constant $\lambda > 0$. What can you say about $\mathrm{dec}_{\mathrm{scr}}$ and $\mathrm{dec}_{\mathrm{scr}'}$?

Solution: Similar to part (a), since

$$\max_{i \in [m]} \mathrm{scr}'(a(i), b) = \max_{i \in [m]} \lambda\,\mathrm{scr}(a(i), b) = \lambda \max_{i \in [m]} \mathrm{scr}(a(i), b),$$

the set of elements achieving the maximum is the same whether we use scr or scr′, and hence $\mathrm{dec}_{\mathrm{scr}'}(b) = \mathrm{dec}_{\mathrm{scr}}(b)$.

For $t \in \mathbb{R}$, consider the *threshold decoder* $\mathrm{dec}_{t,\mathrm{scr}}$ that operates as follows: given b, form the list $L_t = \{i : \mathrm{scr}(a(i), b) \geq t\}$. If L_t consists of a single element i_0, we set $\mathrm{dec}_{t,\mathrm{scr}}(b) = i_0$. Otherwise $\mathrm{dec}_{t,\mathrm{scr}}(b) = 0$.

(c) Show that if $\mathrm{dec}_{t,\mathrm{scr}}(b) \neq 0$, then $\mathrm{dec}_{\mathrm{scr}}(b) = \mathrm{dec}_{t,\mathrm{scr}}(b)$.
Moral: If the threshold decoder makes the correct decision, so does the maximum score decoder.

Solution: Since $\mathrm{dec}_{t,\mathrm{scr}}(b) \neq 0$, L_t consists of one element i_0. By definition of L_t, we thus have $\mathrm{scr}(a(i_0), b) \geq t$, while $\mathrm{scr}(a(j), b) < t$ for all $j \in [m] \setminus \{i_0\}$. In particular this means $\arg\max_{i \in [m]} \mathrm{scr}(a(i), b) = \{i_0\}$, so that the maximizer is unique and hence $\mathrm{dec}_{\mathrm{scr}}(b) = \mathrm{dec}_{t,\mathrm{scr}}(b)$.

Problem 2. Let p_{XY} be a probability distribution on $\mathcal{X}\times\mathcal{Y}$. Suppose the triple $(\tilde{X}, X, Y)$ has a joint distribution given by $p_{\tilde{X}XY}(\tilde{x}, x, y) = p_X(\tilde{x})p_{XY}(x, y)$, i.e., (X, Y) is drawn according to p_{XY} and $\tilde{X}$ is independent of the pair (X, Y) but with the same marginal distribution as X. Suppose $\big((\tilde{X}_i, X_i, Y_i) : i = 1, 2, \dots\big)$ is a collection of i.i.d. random triples with distribution $p_{\tilde{X}XY}$. Given $s\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$, let $\mathrm{scr}(x^n, y^n) = \sum_{i=1}^n s(x_i, y_i)$.

(a) Let $t_0 = \mathbb{E}\left[s(X, Y)\right] = \sum_{x,y} p_{XY}(x, y)s(x, y)$. Show that for any $t < t_0$,

$$\lim_{n\to\infty} \Pr(\mathrm{scr}(X^n, Y^n) < nt) = 0.$$

Solution: By the Weak Law of Large Numbers, we have that for all $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^n s(X^n, Y^n) - t_0\right| > \varepsilon\right) = 0.$$

In particular this means that for any $t < t_0$, which we can write as $t = t_0 - \varepsilon$ for some $\varepsilon > 0$,

$$\begin{aligned}
\lim_{n\to\infty} \Pr(s(X^n, Y^n) < nt) &= \lim_{n\to\infty} \Pr\left(\frac{1}{n}s(X^n, Y^n) < t_0 - \varepsilon\right)\\
&= \lim_{n\to\infty} \Pr\left(\varepsilon < t_0 - \frac{1}{n}s(X^n, Y^n)\right)\\
&\le \lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^n s(X^n, Y^n) - t_0\right| > \varepsilon\right)\\
&= 0,
\end{aligned}$$

where the inequality follows from the fact that for any random variable X and constant $c \in \mathbb{R}$, $\Pr(X > c) \le \Pr(|X| > c)$.

(b) Show that

$$\Pr(\mathrm{scr}(\tilde{X}^n, Y^n) \ge nt) \le \exp_2[-n(t - \alpha)],$$

where $\alpha = \log_2 \mathbb{E}\left[\exp_2\left(s(\tilde{X}, Y)\right)\right] = \log_2 \sum_{x,y} p_X(x)p_Y(y)\exp_2(s(x, y))$.
Hint: $\mathbb{1}\{Z \ge z\} \le \exp_2(Z)\exp_2(-z)$.

Solution: We can use Chebyshev's inequality to get

$$\begin{aligned}
\Pr(\mathrm{scr}(\tilde{X}^n, Y^n) \ge nt) &= \Pr\left(\exp_2(\mathrm{scr}(\tilde{X}^n, Y^n)) \ge \exp_2(nt)\right)\\
&\le \frac{\mathbb{E}[\exp_2(\mathrm{scr}(\tilde{X}^n, Y^n)]}{2^{nt}}\\
&= \exp_2(\log_2 \mathbb{E}[\exp_2(\mathrm{scr}(\tilde{X}^n, Y^n)] - nt)\\
&= \exp_2\left(\log_2 \mathbb{E}\left[\prod_{i=1}^n \exp_2(s(\tilde{X}_i, Y_i))\right] - nt\right)\\
&= \exp_2\left(\log_2 \prod_{i=1}^n \mathbb{E}\left[\exp_2(s(\tilde{X}_i, Y_i))\right] - nt\right)\\
&= \exp_2\left(\log_2 \mathbb{E}\left[\exp_2(s(\tilde{X}, Y))\right]^n - nt\right)\\
&= \exp_2(n\log_2 \mathbb{E}[\exp_2(s(\tilde{X}, Y)] - nt)\\
&= \exp_2(-n(t - \alpha)).
\end{aligned}$$

Given a channel $p_{Y|X}$, a probability distribution p_X, a blocklength n, and a rate R, set $m = \lceil 2^{nR} \rceil$, and construct a random encoder $\mathrm{enc}\colon [m] \to \mathcal{X}^n$ by setting $\mathrm{enc}(i) = (E_{i1}, \ldots, E_{in})$, where $(E_{ij} : i \in [m]\, j \in [n])$ is a collection of i.i.d. random variables with distribution p_X. Let the decoder be the threshold decoder $\mathrm{dec}_{nt,\mathrm{scr}}$ using the score function scr as defined above and threshold nt.

Let W be the transmitted message (uniformly chosen in $[m]$), and let $\hat{W}$ be the threshold decoder's decision.

(c) With $\tilde{X}^n$, X^n and Y^n as in the first paragraph, show that

$$\begin{aligned}\Pr(\hat{W} \neq W) &\leq \Pr(\mathrm{scr}(X^n, Y^n) < nt) + (m-1)\Pr(\mathrm{scr}(\tilde{X}^n, Y^n) \geq nt)\\ &\leq \Pr(\mathrm{scr}(X^n, Y^n) < nt) + 2^{nR}\Pr(\mathrm{scr}(\tilde{X}^n, Y^n) \geq nt).\end{aligned}$$

Solution: In the following, for a message $i \in [m]$, we denote its random encoding by $(E_{i1}, \ldots, E_{in})$ and the corresponding output through the channel $p_{Y|X}$ as $(F_{i1}, \ldots, F_{in})$.

There are two kinds of errors to consider:

1. *The threshold decoder returns a set L_{nt} which is empty (type 1 error).*
   If message $i \in [m]$ is sent, the error occurs in particular because $\mathrm{scr}((E_{i1}, \ldots, E_{in}), (F_{i1}, \ldots, F_{in})) < nt$. Since $(E_{i1}, \ldots, E_{in})$ are i.i.d. according to p_X, and $(F_{i1}, \ldots, F_{in})$ is the corresponding output according to the channel $p_{X|Y}$, we can write

$$\begin{aligned}\Pr(\text{Type 1 error}|W = i) &\leq \Pr(\mathrm{scr}((E_{i1}, \ldots, E_{in}), (F_{i1}, \ldots, F_{in})) < nt)\\ &= \Pr(\mathrm{scr}(X^n, Y^n) < nt)\end{aligned}$$

2. *The threshold decoder returns a set L_{nt} with more than one element (type 2 error).*
   If message $i \in [m]$ is sent, this means there is at least one $j \neq i$ such that $\mathrm{scr}((E_{j1}, \ldots, E_{jn}), (F_{i1}, \ldots, F_{in})) \geq nt$. That is, although $\mathrm{enc}(j)$ was not sent through the channel, it scores high when compared to the actual output. Since all the $(E_{j1}, \ldots, E_{jn})$ are i.i.d. according to p_X, and independent of the actual output of the channel $(F_{i1}, \ldots, F_{in})$, we can write

$$\begin{aligned}\Pr(\text{Type 2 error}|W = i) &= \Pr(\exists j \neq i \text{ s.t. } \mathrm{scr}((E_{j1}, \ldots, E_{jn}), (F_{i1}, \ldots, F_{in})) \geq nt)\\ &\leq \sum_{j \in [m] : j \neq i} \Pr(\mathrm{scr}((E_{j1}, \ldots, E_{jn}), (F_{i1}, \ldots, F_{in})) \geq nt)\\ &\leq \sum_{j \in [m] : j \neq i} \Pr(\mathrm{scr}(\tilde{X}^n, Y^n) \geq nt)\\ &= (m-1)\Pr(\mathrm{scr}(\tilde{X}^n, Y^n) \geq nt)\\ &= (\lceil 2^{nR} \rceil - 1)\Pr(\mathrm{scr}(\tilde{X}^n, Y^n) \geq nt)\\ &\leq 2^{nR}\Pr(\mathrm{scr}(\tilde{X}^n, Y^n) \geq nt)\end{aligned}$$

We can now use the above computations to determine the probability of error:

$$\begin{aligned}
\Pr(\hat{W} \neq W) &= \sum_{i\in[m]} \Pr(\hat{W} \neq W | W = i) \Pr(W = i) \\
&= \sum_{i\in[m]} \Pr\left(\text{Type 1 error} \cup \text{Type 2 error} | W = i\right) \Pr(W = i) \\
&= \sum_{i\in[m]} \left[\Pr\left(\text{Type 1 error}|W = i\right) + \Pr\left(\text{Type 2 error}|W = i\right)\right] \Pr(W = i) \\
&\leq \sum_{i\in[m]} \left[\Pr(\text{scr}(X^n, Y^n) < nt) + 2^{nR} \Pr(\text{scr}(\tilde{X}^n, Y^n) \geq nt)\right] \Pr(W = i) \\
&= \Pr(\text{scr}(X^n, Y^n) < nt) + 2^{nR} \Pr(\text{scr}(\tilde{X}^n, Y^n) \geq nt)
\end{aligned}$$

(d) With t_0 and α as in (a) and (b), show that for $R < t_0 - \alpha$, and $\epsilon > 0$, there is a choice of t such that for n large enough, $\Pr(\hat{W} \neq W) < \epsilon$.

Hint: Since $R < t_0 - \alpha$, there is a t strictly in between $R + \alpha$ and t_0. Now use (a), (b), and (c).

Solution: Since $R < t_0 - \alpha$, we can find a t such that $R < t - \alpha < t_0 - \alpha$. From part 2(c), we know that using the threshold decoder with threshold nt, the probability of error can be bounded by

$$\Pr(\text{scr}(X^n, Y^n) < nt) + 2^{nR} \Pr(\text{scr}(\tilde{X}^n, Y^n) \geq nt).$$

The first term goes to 0 as n gets large due to part 2(a) since $t < t_0$. The second term is bounded by $2^{-n(t-\alpha-R)}$ by part 2(b), and can be made arbitrarily small since $R < t - \alpha$. Hence, for any choice of ϵ, we can always find n large enough such that the probability of error is at most ϵ.

(e) Again, with t_0 and α as in (a) and (b), show that for $R < t_0 - \alpha$ and $\epsilon > 0$, for large enough n, there is an encoder enc with $m = \lceil 2^{nR} \rceil$ codewords, and a maximum score decoder dec_{scr} using the score function scr with average error probability at most ϵ.

Solution: From part 2(d), we know that there exists a threshold decoder such that the probability of error can be made arbitrarily small when using a random encoder (note that the error probability is also with respect to the randomness of the encoder!). But this means that there must exists an encoder achieving this probability of error. Moreover, we have seen from Problem 1 c) that the maximum score decoder can only do better than the threshold decoder, so that overall we can guarantee the existence of an encoder and a maximum score decoder with the desired properties.

(f) Evaluate t_0 and α for the choice $s(x, y) := \log_2 \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}$. Use (e) together with 1(a) to conclude that for all rates $R < I(X; Y)$, there is an encoder of rate at least R with the decoder using the maximum likelihood rule $\text{dec}(y^n) = \arg\max_{i\in[m]} p_{Y^n|X^n}\big(y^n \mid \text{enc}(i)\big)$, which achieves an average error probability that can be made arbitrarily small.

Solution: We find that $t_0 = \mathbb{E}[s(X, Y)] = \mathbb{E}\left[\log_2 \frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right] = I(X; Y)$ and $\alpha = \log_2 \mathbb{E}\left[\exp_2(s(\tilde{X}, Y))\right] = \log_2 \mathbb{E}\left[\frac{p_{XY}(\tilde{X},Y)}{p_X(\tilde{X})p_Y(Y)}\right] = \log_2 \sum_{x,y} p_{XY}(x, y) = \log_2(1) = 0$. Hence, we have from part (e) that there exists an encoder with rate R and maximum decoder using the score $\text{scr}(x^n, y^n) = \sum_{i=1}^{n} s(x_i, y_i)$. In particular, this decoder computes

$$\begin{aligned}
\text{dec}_{\text{scr}}(y^n) &= \arg\max_{i\in[m]} \sum_{j=1}^{n} s(\text{enc}(i)_j, y_j)\\
&= \arg\max_{i\in[m]} \sum_{j=1}^{n} \log_2 \frac{p_{XY}(\text{enc}(i)_j, y_j)}{p_X(\text{enc}(i)_j)p_Y(y_j)}\\
&= \arg\max_{i\in[m]} \log_2 \prod_{j=1}^{n} \frac{p_{XY}(\text{enc}(i)_j, y_j)}{p_X(\text{enc}(i)_j)p_Y(y_j)}\\
\text{(from 1(a))} &= \arg\max_{i\in[m]} \log_2 \prod_{j=1}^{n} p_{Y|X}(y_j|\,\text{enc}(i)_j)\\
&= \arg\max_{i\in[m]} \log_2 p_{Y^n|X^n}(y^n|\,\text{enc}(i))\\
&= \arg\max_{i\in[m]} p_{Y^n|X^n}(y^n|\,\text{enc}(i)),
\end{aligned}$$

which is precisely the maximum likelihood rule.

(g) Suppose now the decoder uses the maximum likelihood rule adapted to the channel q instead of the true channel $p_{Y|X}$. Show that all rates up to

$$\max_{r_Y,\lambda}\left[\sum_{x,y} p_X(x)p_{Y|X}(y|x)\log_2\frac{q(y|x)^\lambda}{r_Y(y)} - \log_2\sum_{x,y} p_X(x)p_Y(y)\frac{q(y|x)^\lambda}{r_Y(y)}\right]$$

can be achieved with this (mismatched Maximum Likelihood) decoder. (The maximization is over all probability distributions r_Y and $\lambda > 0$.)

Hint: Use 1(b) and 1(c).

Solution: Note that using the maximum likelihood rule adapted to channel q is equivalent to using the score with $s(x,y) = \log_2\frac{p_X(x)q(y|x)}{p_X(x)q_Y(y)}$ with q_Y being the output distribution. However, as established in Problem 1, replacing $q_Y(y)$ with any function $f(y)$ (non-negative since it is inside a $\log_2$) gives an identical decoder. Moreover, multiplying the score by a constant does not change the decoder. Hence, we can consider $s(x,y) = \lambda\log_2\frac{q(y|x)}{f(y)} = \log_2\frac{q(y|x)^\lambda}{f(y)^\lambda}$. For fixed λ and f, writing $f(y)^\lambda = r_Y(y)$, the maximum rate achievable is given by $t_0 - \alpha$ from part (e), which in this case is

$$\sum_{x,y} p_{XY}(x,y)\log_2\frac{q(y|x)^\lambda}{r_Y(y)} - \log_2\sum_{x,y} p_x(x)p_Y(y)\frac{q(y|x)^\lambda}{r_Y(y)}.$$

Optimizing over λ and r_Y, the maximum rate achievable is thus

$$\max_{\lambda,r_Y}\left\{\sum_{x,y} p_{XY}(x,y)\log_2\frac{q(y|x)^\lambda}{r_Y(y)} - \log_2\sum_{x,y} p_x(x)p_Y(y)\frac{q(y|x)^\lambda}{r_Y(y)}\right\}.$$

(h) Suppose our channel $p_{Y|X}$ is the a binary input binary output channel $p(1|0) = \delta_0$ and $p(0|1) = \delta_1$, with $\delta_i \le 1/2$. Suppose the decoder is using the ML rule adapted to the BSC(ϵ) with $\epsilon < 1/2$. Show that all rates up to $1 - h_2\left(\frac{\delta_0+\delta_1}{2}\right)$ can be achieved by a suitable choice of p_X at the encoder.

Hint: Choose p_X, r_Y to be uniform distributions in (g) and find the maximizing λ.

Solution: We choose p_X and r_Y to be uniform distributions. Denoting the maximum value computed in part 2(g) by $R^\star$, our choice yields

$$\begin{aligned} R^\star &\geq \max_{\lambda>0} \left\{ \frac{1}{2} \sum_{x,y} p_{Y|X}(y|x) \log_2(2q(y|x)^\lambda) - \log_2 \sum_{x,y} p_Y(y) q(y|x)^\lambda \right\} \\ &= \max_{\lambda>0} \left\{ 1 + \frac{1}{2} \sum_{x,y} p_{Y|X}(y|x) \log_2(q(y|x)^\lambda) - \log_2 \sum_y p_Y(y) \sum_x q(y|x)^\lambda \right\}. \end{aligned}$$

Moreover, plugging

$$p_Y(y) = \begin{cases} \frac{1}{2}(1-\delta_0+\delta_1), & \text{if } y=0 \\ \frac{1}{2}(1-\delta_1+\delta_0)), & \text{if } y=1, \end{cases} \tag{1}$$

in the above and defining $\delta := (\delta_0+\delta_1)/2$, we find

$$\begin{aligned} R^\star &\geq \max_{\lambda>0} \left\{ 1 + \frac{1}{2} \sum_{x,y} p_{Y|X}(y|x) \log_2(q(y|x)^\lambda) - \log_2 \sum_y p_Y(y) \sum_x q(y|x)^\lambda \right\} \\ &= \max_{\lambda>0} \left\{ 1 + \frac{1}{2} \left[(1-\delta_0)\log_2((1-\epsilon)^\lambda) + \delta_0 \log_2(\epsilon^\lambda) + \delta_1 \log_2(\epsilon^\lambda) + (1-\delta_1)\log_2((1-\epsilon)^\lambda)\right] \right. \\ &\qquad \left. - \log_2 \left[\frac{1}{2} \left(\epsilon^\lambda + (1-\epsilon)^\lambda\right)(2-\delta_0-\delta_1+\delta_0+\delta_1)) \right] \right\} \\ &= \max_{\lambda>0} \left\{ 1 + (1-\delta)\log_2((1-\epsilon)^\lambda) + \delta \log_2(\epsilon^\lambda) - \log_2\left(\epsilon^\lambda + (1-\epsilon)^\lambda\right) \right\} \\ &= \max_{\lambda>0} \left\{ 1 + (1-\delta)\log_2 \frac{(1-\epsilon)^\lambda}{\epsilon^\lambda + (1-\epsilon)^\lambda} + \delta \log_2 \frac{\epsilon^\lambda}{\epsilon^\lambda + (1-\epsilon)^\lambda} \right\}. \end{aligned} \tag{2}$$

To find the value of λ which achieves the maximum, we leave it as an exercise to show that for a given parameter $\delta \in [0,1]$, the function $f(x) = (1-\delta)\log_2(1-x) + \delta \log_2(x)$ is maximized at $x = \delta$. From this observation, we conclude that the optimal λ in Eq. (2) is such that

$$\begin{aligned} \delta = \frac{\epsilon^\lambda}{\epsilon^\lambda + (1-\epsilon)^\lambda} &\iff \delta(1-\epsilon)^\lambda = (1-\delta)\epsilon^\lambda \\ &\iff \left(\frac{1-\epsilon}{\epsilon}\right)^\lambda = \frac{1-\delta}{\delta} \\ &\iff \lambda = \frac{\log_2 \frac{1-\delta}{\delta}}{\log_2 \frac{1-\epsilon}{\epsilon}}. \end{aligned}$$

As mentioned above, this choice of λ gives

$$\begin{aligned} R^\star &\geq 1 + (1-\delta)\log_2(1-\delta) + \delta \log_2 \delta \\ &= 1 - h_2(\delta). \end{aligned}$$

Since all rates up to $R^\star$ can be achieved and $1 - h_2(\delta) \leq R^\star$, we see that in particular rates below $1 - h_2(\delta)$ can be achieved.