# Information Theory I

Fall Semester 2017

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

## Model Answers to Exercise 6 of October 25, 2017

http://www.isi.ee.ethz.ch/teaching/courses/it1.html

---

### Problem 1 *Strong Versus Weak Typicality*

By the definition of the strongly typical set (note that we do not follow the definition from the book *Elements of Information Theory*)

$$\mathcal{T}_\epsilon^{(n)}(P) \triangleq \left\{\mathbf{x} \in \mathcal{X}^n : \left|\frac{1}{n}N(\xi|\mathbf{x}) - P(\xi)\right| \leq \epsilon \cdot P(\xi) \quad \forall \xi \in \mathcal{X}\right\},$$

a sequence $\mathbf{x} \in \mathcal{X}^{100}$ is strongly 0.01-typical if and only if the number of occurrences of "True" and "False" is between 49.5 and 50.5. The set of strongly typical sequences therefore consists of all sequences where "True" and "False" occur exactly 50 times. (This means that only a fraction of $\binom{100}{50}/2^{100} \approx 0.08$ of the sequences are strongly 0.01-typical.)
By the definition of the weakly typical set, a sequence $\mathbf{x} \in \mathcal{X}^{100}$ is weakly $\epsilon$-typical if and only if its probability is not smaller than $2^{-100(H(X)+\epsilon)}$ and not larger than $2^{-100(H(X)-\epsilon)}$. Because "True" and "False" are equiprobable, $H(X) = 1$. Since the $X_i$ are IID, every sequence $\mathbf{x} \in \mathcal{X}^{100}$ occurs with probability $2^{-100}$. Consequently, every sequence is weakly typical, i.e., the set of weakly typical sequences is equal to $\mathcal{X}^{100}$ (irrespective of $\epsilon > 0$).

### Problem 2 *Random Box Size*

a) We have

$$\mathsf{E}[V_n] = \mathsf{E}\left[\prod_{i=1}^n X_i\right] \overset{\text{(i)}}{=} \prod_{i=1}^n \mathsf{E}[X_i] = \prod_{i=1}^n \frac{1}{2} = \left(\frac{1}{2}\right)^n,$$

where (i) holds because $X_1, \ldots, X_n$ are independent. Consequently,

$$\lim_{n\to\infty} (\mathsf{E}[V_n])^{1/n} = \frac{1}{2}.$$

b) We have

$$\mathsf{E}[\ln X_1] = \int_0^1 \ln x \, \mathrm{d}x = (x \ln x - x)\Big|_{x=0}^{x=1} = -1,$$

and because the $X_i$ (and therefore also $\ln X_i$) are IID, the claim follows from the weak law of large numbers.

(For a proof of the weak law of large numbers, see Problem 4 of Exercise 1, which assumes that the variance is finite; more generally, the existence of the expectation is sufficient for the weak law of large numbers to hold, so we do not need to check whether the variance is finite or not.)

c) Intuitively, the claim follows from Part b) because

$$L_n = \left(\prod_{i=1}^{n} X_i\right)^{1/n} = e^{\frac{1}{n}\ln\left(\prod_{i=1}^{n} X_i\right)} = e^{\frac{1}{n}\sum_{i=1}^{n}\ln X_i},$$

because $\frac{1}{n}\sum_{i=1}^{n}\ln X_i$ converges to $-1$ in probability, and because the exponential function is continuous.

(For the interested reader, we provide a mathematically rigorous argument. Let $\epsilon > 0$ be fixed and let $Z_n = \frac{1}{n}\sum_{i=1}^{n}\ln X_i$. By the continuity of the exponential function, there exists a $\delta > 0$ such that for all $z \in \mathbb{R}$, $|e^z - e^{-1}| < \epsilon$ whenever $|z-(-1)| < \delta$. Therefore,

$$\Pr\left[\left|e^{Z_n} - e^{-1}\right| \geq \epsilon\right] \leq \Pr[|Z_n - (-1)| \geq \delta]. \tag{1}$$

From Part b) we know that the RHS of (1) tends to zero as $n$ tends to infinity. Thus, the LHS of (1) must also tend to zero as $n$ tends to infinity, which proves the claim.)

## Problem 3 *From AEP to Kraft's Inequality*

Let $\ell_1, \ldots, \ell_d$ be the codeword lengths of a uniquely decodable one-to-variable code $\mathcal{C}$ and assume that

$$\alpha = \sum_{i=1}^{d} 2^{-\ell_i} > 1. \tag{2}$$

We construct a memoryless source, i.e., a source that emits IID messages, with message probabilities

$$p_i = \frac{2^{-\ell_i}}{\alpha} \quad \forall i \in \{1, \ldots, d\}. \tag{3}$$

Using $\mathcal{C}$ to encode this source yields the expected codeword length

$$\mathsf{E}[\ell(X)] = \sum_{i=1}^{d} p_i \ell_i \overset{(3)}{=} \sum_{i=1}^{d} p_i \log \frac{1}{\alpha p_i} = H(X) - \log \alpha, \tag{4}$$

where $\log \alpha > 0$ because of (2).
(You do not need to show this, but setting $p_i \propto 2^{-\ell_i}$ in fact minimizes $\mathsf{E}[\ell(X)] - H(X)$ because

$$\begin{aligned}
\mathsf{E}[\ell(X)] - H(X) &= \sum_{i=1}^{d} p_i \ell_i - \sum_{i=1}^{d} p_i \log \frac{1}{p_i} \\
&= \sum_{i=1}^{d} p_i \log \frac{p_i}{2^{-\ell_i}} \\
&= \underbrace{\sum_{i=1}^{d} p_i \log \frac{p_i}{2^{-\ell_i}/\alpha}}_{= D(p||q) \geq 0} + \underbrace{\sum_{i=1}^{d} p_i \log(1/\alpha)}_{= -\log \alpha},
\end{aligned}$$

so the choice (3) corresponds to $D(p||q) = 0$, which minimizes the expression.)
For every $n \in \mathbb{Z}^+$, we construct a code that maps $n$ source symbols to $\left\lceil n(H(X) - \frac{1}{2}\log\alpha)\right\rceil + 1$ bits as follows:

**Step 1:** Describe a source sequence $\mathbf{x} \in \mathcal{X}^n$ using the extension of $\mathcal{C}$.

 

**Step 2:** If the length of the description of $\mathbf{x}$ is less than or equal to $\lceil n(H(X) - \frac{1}{2}\log\alpha)\rceil$, append a single one to its end and then add zeros to its end until the length of the description is $\lceil n(H(X) - \frac{1}{2}\log\alpha)\rceil + 1$; use the resulting bitstring as the codeword for $\mathbf{x}$.

If the length of the description of $\mathbf{x}$ is greater than $\lceil n(H(X) - \frac{1}{2}\log\alpha)\rceil$, use the all-zero bitstring as the codeword for $\mathbf{x}$.

Observe that the rate of the described code satisfies

$$\lim_{n\to\infty} \rho_n = \lim_{n\to\infty} \frac{\lceil n(H(X) - \frac{1}{2}\log\alpha)\rceil + 1}{n} = H(X) - \frac{1}{2}\log\alpha < H(X),$$

where the last inequality follows from (2). Consequently, this code satisfies the condition for the converse of the general source coding theorem. We will show next that the probability of successful decoding of this code tends to one as $n$ tends to infinity, which contradicts the theorem and establishes that no uniquely decodable one-to-variable code $\mathcal{C}$ satisfying (2) can exist.
To analyze the error probability of the code, observe that the source message can be uniquely determined if the length of the description in Step 1 is less than or equal to $\lceil n(H(X) - \frac{1}{2}\log\alpha)\rceil$: removing the trailing zeros and the single one from the codeword recovers the original description (in other words, the padding scheme is injective). We bound the probability that the description length exceeds $\lceil n(H(X) - \frac{1}{2}\log\alpha)\rceil$ as follows:

$$\begin{aligned}
&\Pr\left[\sum_{k=1}^{n} \ell(X_k) > \left\lceil n\left(H(X) - \frac{1}{2}\log\alpha\right)\right\rceil\right] \\
&\quad\le \Pr\left[\sum_{k=1}^{n} \ell(X_k) > n\left(H(X) - \frac{1}{2}\log\alpha\right)\right] \\
&\quad= \Pr\left[\frac{1}{n}\sum_{k=1}^{n} \ell(X_k) > H(X) - \frac{1}{2}\log\alpha\right] \\
&\quad= \Pr\left[\frac{1}{n}\sum_{k=1}^{n} \ell(X_k) - (H(X) - \log\alpha) > \frac{1}{2}\log\alpha\right] \\
&\quad\le \Pr\left[\left|\frac{1}{n}\sum_{k=1}^{n} \ell(X_k) - (H(X) - \log\alpha)\right| \ge \frac{1}{2}\log\alpha\right]. \qquad (5)
\end{aligned}$$

By the weak law of large numbers, the right-hand side of (5) tends to zero as $n$ tends to infinity (we have an IID source; we computed the expected codeword length in (4); and $\log\alpha$ is positive). Hence, the probability that the description length is less than or equal to $\lceil n(H(X) - \frac{1}{2}\log\alpha)\rceil$, and also the probability of successful decoding, tend to one as $n$ tends to infinity.

## Problem 4 *Fano's Inequality*

a) For any guess $\hat{x} \in \mathcal{X}$, the probability of error is

$$P_{\mathrm{e}} = \Pr[X \neq \hat{x}] = \sum_{x\in\mathcal{X}\setminus\{\hat{x}\}} P_X(x) = 1 - P_X(\hat{x}),$$

so the smallest probability of error is achieved by guessing the most probable value of $X$ (or any of the most probable values if the most probable value is not unique). The associated probability of error is

$$P_{\mathrm{e}}^* = 1 - \max_{x\in\mathcal{X}} P_X(x).$$

b) From Part a) we know that for any probability distribution $(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m)$, the probability of error associated with the guess $\hat{x}$ is equal to $P_{\mathrm{e}}$ if and only if $\tilde{p}_{\hat{x}} = 1 - P_{\mathrm{e}}$. Without loss of generality assume from now on that $\hat{x} = 1$, which implies $p_1 = 1 - P_{\mathrm{e}}$.

Denote by $\tilde{\mathcal{P}}$ the set of all probability distributions $(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m)$ with $\tilde{p}_1 = 1 - P_{\mathrm{e}}$. Denote the maximum of $H(\tilde{X})$ over $\tilde{\mathcal{P}}$ by $H^*$, and let $Q_1 \triangleq (1 - P_{\mathrm{e}}, q_2, \ldots, q_m) \in \tilde{\mathcal{P}}$ be a probability distribution that achieves the maximum, so

$$H^* \triangleq \max_{(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m) \in \tilde{\mathcal{P}}} H(\tilde{X}) = H(Q_1).$$

(The existence of $Q_1$ follows from the extreme value theorem.) Because relabeling does not change the entropy, the following probability distributions all have entropy $H^*$:

$$\begin{aligned}
Q_1 &= (1 - P_{\mathrm{e}},\, q_2,\; q_3,\; \ldots,\; q_{m-1},\, q_m),\\
Q_2 &= (1 - P_{\mathrm{e}},\, q_3,\; q_4,\; \ldots,\; q_m,\quad q_2),\\
&\vdots\\
Q_{m-1} &= (1 - P_{\mathrm{e}},\, q_m,\, q_2,\; \ldots,\; q_{m-2},\, q_{m-1}).
\end{aligned}$$

Define the mixture

$$Q_{\mathrm{mix}} \triangleq \frac{1}{m-1} Q_1 + \frac{1}{m-1} Q_2 + \ldots + \frac{1}{m-1} Q_{m-1}.$$

Since the probabilities of $Q_1$ have to sum up to one, we have $q_2 + q_3 + \ldots + q_m = P_{\mathrm{e}}$, and

$$Q_{\mathrm{mix}} = \left(1 - P_{\mathrm{e}}, \frac{P_{\mathrm{e}}}{m-1}, \frac{P_{\mathrm{e}}}{m-1}, \ldots, \frac{P_{\mathrm{e}}}{m-1}, \frac{P_{\mathrm{e}}}{m-1}\right).$$

Now,

$$\begin{aligned}
H^* &\overset{\text{(i)}}{=} \frac{1}{m-1} H(Q_1) + \frac{1}{m-1} H(Q_2) + \ldots + \frac{1}{m-1} H(Q_{m-1})\\
&\overset{\text{(ii)}}{\leq} H(Q_{\mathrm{mix}})\\
&\overset{\text{(iii)}}{\leq} H^*,
\end{aligned}$$

where (i) holds because $H(Q_1) = \ldots = H(Q_{m-1}) = H^*$; (ii) holds because the entropy is concave; and (iii) holds because $Q_{\mathrm{mix}} \in \tilde{\mathcal{P}}$ and because $H^*$ is the maximum of $H(\tilde{X})$ over $\tilde{\mathcal{P}}$. Thus, $H^* = H(Q_{\mathrm{mix}})$ must hold, and we have

$$\begin{aligned}
H^* &= H(Q_{\mathrm{mix}})\\
&= (1 - P_{\mathrm{e}}) \log \frac{1}{1 - P_{\mathrm{e}}} + \frac{P_{\mathrm{e}}}{m-1} \log \frac{m-1}{P_{\mathrm{e}}} + \ldots + \frac{P_{\mathrm{e}}}{m-1} \log \frac{m-1}{P_{\mathrm{e}}}\\
&= (1 - P_{\mathrm{e}}) \log \frac{1}{1 - P_{\mathrm{e}}} + (m-1) \cdot \frac{P_{\mathrm{e}}}{m-1} \log \frac{m-1}{P_{\mathrm{e}}}\\
&= (1 - P_{\mathrm{e}}) \log \frac{1}{1 - P_{\mathrm{e}}} + P_{\mathrm{e}} \log \frac{1}{P_{\mathrm{e}}} + P_{\mathrm{e}} \log (m-1)\\
&= H_b(P_{\mathrm{e}}) + P_{\mathrm{e}} \log (m-1).
\end{aligned}$$

Finally,

$$\begin{aligned}
H(X) &\overset{\text{(i)}}{\leq} H^*\\
&= H_b(P_{\mathrm{e}}) + P_{\mathrm{e}} \log (m-1),
\end{aligned}$$

where (i) holds because $(p_1, p_2, \ldots, p_m) \in \tilde{\mathcal{P}}$ and because $H^*$ maximizes $H(\tilde{X})$ over $\tilde{\mathcal{P}}$.

c) i) This holds because $E$ is a binary random variable and $\Pr[E = 1] = P_{\mathrm{e}}$.

ii) This holds because conditioning does not increase entropy.

iii) This follows from the definition of the conditional entropy and from the law of total probability.

iv) Conditional on $Y = y$, $X$ has the distribution $P_{X|Y=y}(x)$. Let $P_{\mathrm{e},y}$ denote the probability of error conditional on $Y = y$. We infer from Part b) that

$$H(X|Y = y) \leq H_b(P_{\mathrm{e},y}) + P_{\mathrm{e},y} \log (m - 1)$$

must hold. Rewriting this inequality in terms of the error indicator variable $E$, we obtain

$$H(X|Y = y) \leq H(E|Y = y) + \Pr[E = 1|Y = y] \log (m - 1)$$

for every $y \in \mathcal{Y}$.

v) This follows from the definition of the conditional entropy.