ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Handout 16 — Information Theory and Coding

Midterm Solutions — Oct. 29, 2013

PROBLEM 1.

(a) Since $\ell(u) := \text{length}(\mathcal{C}(u)) \geq \log \frac{Q}{q(u)}$, we see that

$$\sum_u 2^{-\ell(u)} \leq \sum_u q(u)/Q = 1.$$

Thus, the prescribed lengths satisfy Kraft's inequality and we conclude that a prefix code with these lengths exist.

(b) Suppose p_α is the true distrubution. Since $q(u) \geq p_\alpha(u)$, the codeword lengths satisfy

$$\ell(u) = \left\lceil \log \frac{Q}{p_\alpha(u)} \right\rceil \leq 1 + \log Q + \log \frac{1}{p_\alpha(u)}$$

Multiplying both sides by $p_\alpha(u)$ and summer over u gives the inequality $E[\text{length}(\mathcal{C}(U))] \leq 1 + \log Q + H(U)$.

(c) Observe that $q(u) = \max_{\alpha \in A} p_\alpha(u) \leq \sum_{\alpha \in A} p_\alpha(u)$. Thus

$$Q = \sum_u q(u) \leq \sum_{\alpha \in A} \sum_u p_\alpha(u) = \sum_{\alpha \in A} 1 = |A|.$$

(d) By the hypothesis of the problem $q(u) = \max_{\alpha \in B} p_\alpha(u)$. Repeating the computation in (c) gives $Q \leq |B|$.

(e) We claim that when we maximize $f(\alpha) = \alpha^k (1-\alpha)^{n-k}$ over the choice of $\alpha \in [0,1]$, the maximum occurs at $\alpha = k/n$ which is an element of B: to see this, note that we may equivalently maximize $\ln f(\alpha) = k \ln \alpha + (n-k)\ln(1-\alpha)$, by setting $\frac{d}{d\alpha} \log f(\alpha)$ to zero. This yields

$$\frac{k}{\alpha} = \frac{n-k}{1-\alpha}$$

from which we find $\alpha = k/n$ as the maximizer.

Thus, for any $(u_1, \dots, u_n)$, $\max_{\alpha \in A} p_\alpha(u_1, \dots, u_n)$ equals $\max_{\alpha \in B} p_\alpha(u_1, \dots, u_n)$.

(f) With $\alpha = \Pr(U_1 = 1)$, p_α in (e) is the distribution of i.i.d. binary random variables $U_1, \dots, U_n$. Using (b), we see that there is a code $\mathcal{C}$ for $(U_1, \dots, U_n)$ for which

$$E[\text{length}(\mathcal{C}(U_1, \dots, U_n)] - H(U_1, \dots, U_n) \leq 1 + \log Q. \qquad (*)$$

By (d), $Q \leq |B| = (n+1)$. Also $H(U_1, \dots, U_n) = nH(U_1)$. Dividing both sides of (**??**) by n yields the desired conclusion.

Problem 2.

(a) Let $\ell_{\max} = \max_u \operatorname{length}(\mathcal{C}(u))$ be the length of the longest codeword, and $\ell_{\min} = \min_u \operatorname{length}(\mathcal{C}(u))$ be the length of the shortest codeword. In a Huffman code there are (at least) two sibling codewords of longest length, let u_1 and u_2 be corresponding letters and $w0$ and $w1$ the corresponding codewords; let u_3 be a letter assigned the shortest codeword, let v be the corresponding codeword. We can now construct a new prefix-free code that assigns to u_3 the codeword w and assigns to u_1 and u_2 the codewords $v0$ and $v1$.

Set $d := \ell_{\max} - \ell_{\min}$. We will show that $d \leq 1$ by contradiction. Accordingly, suppose $d > 1$. Then, in the new code, the codewords of u_1 and u_2 have become shorter by $d-1$ bits and codeword of u_3 will have become longer by $d-1$ bits. The expected length has thus change by $(d-1)[p(u_3) - p(u_1) - p(u_2)]$. As

$$p(u_3) \leq \max_u p(u) < 2 \min_u p(u) \leq p(u_1) + p(u_2),$$

the new code has a strictly smaller expected length, contradicting the optimality of the Huffman code.

(b) If the inequality in $(*)$ is not strict, then the argument in (a) shows that if $d > 1$, then, the new code has smaller or equal expected length (and thus is also optimal), but at the same time, has fewer (perhaps zero) codewords of lengths $\ell_{\max}$ or $\ell_{\min}$. Repeating the reduction in (a) until no such codewords remain shows that there exists an optimal (and thus Huffman) code with the desired property.

(c) By part (a) we know that the Huffman code will only have codewords of lengths k and $k+1$ for some k. Let M_k and M_{k+1} be the number of such codewords. Since the Huffman code tree is complete, we have $2M_k + M_{k+1} = 2^{k+1}$. At the same time, $M_k + M_{k+1} = |\mathcal{U}| = 2^j + r$. These two equations yield

$$M_k = 2^{k+1} - 2^j - r \quad \text{and} \quad M_{k+1} = 2^{j+1} + 2r - 2^{k+1}.$$

From these we find that $k = j$, $M_j = 2^j - r$, $M_{j+1} = 2r$.

(d) Since j and $j+1$ are the two possible codeword lengths, the expected codeword length equals j plus the total probability of the letters that get assigned codewords of length $j+1$. By (c) we know there are $2r$ such letters. In an optimal code, the less probable letters must receive codewords of longer length. Consequently, the expected codeword length exceeds j by exactly the sum of the probabilities of $2r$ least likely codewords.

PROBLEM 3.

(a) Since the Huffman code $\mathcal{C}_y$ is designed for the distribution p_y, where $p_y(x) = p(x|y)$, its expected length satisfies

$$\sum_x p_y(x) \log \frac{1}{p_y(x)} \le \sum_x p_y(x) \operatorname{length}(\mathcal{C}_y(x)) \le \sum_x p_y(x) \log \frac{1}{p_y(x)} + 1.$$

Multipying all sides by $p(y)$ and summing over y we get $H(Y|X) \le E[\operatorname{length}(\mathcal{C}_Y(X))] \le H(X|Y) + 1$.

(b) From the first $\lceil m \log |\mathcal{U}| \rceil$ bits of the description we learn U_1^m, and thus Y_1. The rest of the description starts with a codeword of $\mathcal{C}_{Y_1}$. This code being prefix free, we can decode $X_1 = U_{m+1}^{m+k}$. From Y_1 and X_1 we know U_1^{m+k}, in particular Y_2. Knowing Y_2 we know that the rest of the description starts with a codeword of $\mathcal{C}_{Y_2}$. This code being prefix free, we can decode $X_2 = U_{m+k+1}^{m+2k}$. Since we already knew U_1^{m+k} we now know U_1^{m+2k}, and thus learn Y_3. Continuing in this manner, after n decoding operations we know U_1^{m+nk}.

(c) Note that $L_n = \lceil m \log |\mathcal{U}| \rceil + \sum_{i=1}^n \operatorname{length}(\mathcal{C}_{Y_i}(X_i))$. By stationarity, (X_i, Y_i) has the same distribution as (X_1, Y_1), and thus

$$E[L_n] = \lceil m \log |\mathcal{U}| \rceil + nE[\operatorname{length}(\mathcal{C}_{Y_1}(X_1))].$$

Dividing both sides of this equality by $m + nk$ and taking the limit as n gets large, we find that $\rho = 0 + \frac{1}{k} E[\operatorname{length}(\mathcal{C}_{Y_1}(X_1))]$. By (a), $E[\operatorname{length}(\mathcal{C}_{Y_1}(X_1))]$ is between $H(X_1|Y_1)$ and $H(X_1|Y_1) + 1$ and thus

$$\frac{1}{k} H(X_1|Y_1) \le \rho \le \frac{1}{k}[H(X_1|Y_1) + 1].$$

Noting $X_1 = U_{m+1}^{m+k}$ and $Y_1 = U_1^m$ concludes the proof.

(d) Let $b_{k,m} = \frac{1}{k} H(U_{m+1}^{m+k}|U_1^m)$. We have

$$b_{k,m+1} = \frac{1}{k} H(U_{m+2}^{m+2+k}|U_1^{m+1}) \le \frac{1}{k} H(U_{m+2}^{m+2+k}|U_2^{m+1}) = \frac{1}{k} H(U_{m+1}^{m+1+k}|U_1^m) = b_{k,m}.$$

The inequality is due to “conditioning reduces entropy” and the following equality is due to stationarity.

(e) Define $a_m = H(U_{m+1}|U_1^m) = b_{1,m}$. By (d) we see that a_m is a non-increasing sequence, in particular, any term is smaller than the average of any terms that preceede it,

$$a_{m+k} \le \frac{1}{k}[a_m + \cdots + a_{m+k-1}].$$

Expressing $b_{k+1,m}$ by the chain rule and using the inequality just shown,

$$\begin{aligned}
b_{k+1,m} &= \frac{1}{k+1}[a_m + a_{m+1} + \cdots + a_{m+k-1} + a_{m+k}] \\
&\le \frac{1}{k+1}[a_m + a_{m+1} + \cdots + a_{m+k-1}] + \frac{1}{k+1}\frac{1}{k}[a_m + a_{m+1} + \cdots + a_{m+k-1}] \\
&= \frac{1}{k}[a_m + a_{m+1} + \cdots + a_{m+k-1}] = b_{k,m}.
\end{aligned}$$

(f) Let $H_U = \lim_{m\to\infty} \frac{1}{m} H(U^m)$ denote the entropy rate of the process. By the chain rule $H(U_{m+1}^{2m}|H_1^m) = H(U_1^{2m}) - H(U_1^m)$. Thus

$$\lim_{m\to\infty} \frac{1}{m} H(U_{m+1}^{2m}|U_1^m) = \lim_{m\to\infty} \frac{2}{2m} H(U_1^{2m}) - \lim_{m\to\infty} \frac{1}{m} H(U_1^m) = 2H_U - H_U = H_U.$$