ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Handout 17 Information Theory and Coding

Solutions to Midterm November 9, 2010, SG1 – 13:15-15:00

You have 2 hours. It is not necessarily expected that you finish all problems. Do not lose too much time on each problem but try to collect as many points as possible.

Closed-book, no calculators, cell-phones; only once piece of A4 paper is allowed. Write only what is relevant to the question!

Good Luck!!

Name: I. M. Perfect

Prob I	40 / 40
Prob II	40 / 40
Prob III	40 / 40
Prob IV	40/ 40
Total	160 / 160

Problem 1 (Walk the Line)**.** [40 pts] We consider a symmetric random walk on the integers starting with $X_0 = 0$. At each step with equal probability we either move one to the left or one to the right.

(i) [10 pts] Note that X_2 takes on the values $\{-2, 0, 2\}$ with probability $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, respectively. This implies that $H(X_2) = \frac{3}{2}$.

(ii) [10 pts] No.

(iii) [10 pts] Yes. To show this note that $H(X_i \mid X_0, \ldots, X_{i-1}) = H(X_i \mid X_{i-1})$ since, given X_{i-1} the rv X_i takes on the values $X_{i-1} \pm 1$. Since the respective probabilities are $\frac{1}{2}$ it follows that the entropy rate of this process is 1.

(iv) [10 pts] Note that from $X_0, \ldots, X_n$ one can compute the vector $Y_1, \ldots, Y_n$. It follows that $H(X_0, \ldots, X_n)/n \geq H(Y_1, \ldots, Y_n)/n$ for any n. But the map is invertible, since $X_i = Y_i - X_{i-1}$ and $X_0 = 0$ is known. So the inverse inequality is true as well.

Problem 2 (Compress That)**.** [40 pts]

(i) [10 pts] We have $\rho(X_1^\infty) = 0$. We show this by showing that $\rho(X_1^\infty) \leq \delta$ for any $\delta > 0$. To see the last statement, build an invertible FSM which "recognizes" a string of type "ab...ab" for a particular length, call it L, and outputs lets say "0" at the end of this string and returns to the starting state. Hence this machine will output an infinite string of "0" when the input is X_1^∞. Further, there is a loop in the starting state when the input is b. In this case output 11, where the first 1 indicates that it is not this special path and the second one that it is a b. Finally, from each state of the chain which recognizes the special string make an edge back to the starting state in the case the next input is not the correct one. The output for each such edge is an encoding for the sequence where all odd bits are 1 and even bits are 0 to denote an "a" and 1 to denote a "b". This machine is clearly lossless and has a compressibility of $1/L$ for the desired sequence.

(ii) [10 pts] A machine as described above will have $\rho_M(X_1^\infty) = 1/4$.

(iii) [10 pts] We have $\rho_{\text{LZ}} = 0$ since compressibility is non-negative and we know that the compressibility of LZ is at least as good as that of any FSM, i.e., we know that $\rho_{\text{LZ}}(X_1^\infty) \leq \rho(X_1^\infty)$.

(iv) [10 pts] The dictionary increases by 1 every time and has size 2 in the beginning. Hence, if we look at lets say c steps of the algorithm then we need in total

$$\sum_{i=1}^{c} \lceil \log(1+i) \rceil \leq c \log(2(c+1))$$

bits to describe the output.

What are the words which we are using. Note that the parsing is a, b, ab, aba, ba, bab,... Note that in average at most every second step the length of the used dictionary word increases by 1, i.e., we have a linear increase in the used dictionary words. Therefore, if we compute the total length which we have parsed after c steps, this length increases like the squre of c.

It follows that the ratio of the total number of bits used divided by the total length described behaves like $1/c$, i.e., it tends to 0.

Problem 3 (Smallest Most Probable Set). (i) $Pr(A_\epsilon^{(n)} \cap B) = Pr(A_\epsilon^{(n)}) + Pr(B) - Pr(A_\epsilon^{(n)} \cup B) \geq 2(1-\epsilon) - 1 = 1 - 2\epsilon$

(ii) Each $x^n \in A_\epsilon^{(n)} \cap B$ is in particular in $A_\epsilon^{(n)}$ and therefore by definition of $A_\epsilon^{(n)}$, we have

$$Pr(A_\epsilon^{(n)} \cap B) = \sum_{x^n \in A_\epsilon^{(n)} \cap B} p(x^n) \leq \sum_{x^n \in A_\epsilon^{(n)} \cap B} 2^{-n(H-\epsilon)}$$

.

(iii)

$$1 - 2\epsilon \leq \sum_{x^n \in A_\epsilon^{(n)} \cap B} 2^{-n(H-\epsilon)} \leq \sum_{x^n \in B} 2^{-n(H-\epsilon)} = |B| 2^{-n(H-\epsilon)}.$$

Therefore $|B| \geq (1-2\epsilon)2^{n(H-\epsilon)}$.

(iv) In the previous part, we set $B = C_\epsilon^{(n)}$. Then we use the fact that $|A_\epsilon^{(n)}| \geq |C_\epsilon^{(n)}| \geq (1-2\epsilon)2^{n(H-\epsilon)}$. Now, we take the logarithm of all three parts and we let n goes to ∞. The assertion follows due to the Sandwich property.

Problem 4 (Data Processing Inequality)**.**

(i)

$$\begin{aligned} I(X;Y,Z) &= H(Y,Z) - H(Y,Z|X) \\ &= (H(Y) + H(Z|Y)) - (H(Y|X) + H(Z|X,Y)) \\ &= (H(X) - H(Y|X)) + (H(Z|Y) - H(Z|X,Y)) \\ &= I(X;Z) + I(X;Y|Z). \end{aligned}$$

Interchanging Z and Y in the above, we get $I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$.

(ii) The mutual information $I(X;Z|Y)$ can also be expressed as the Kullback-Leibler divergence between the conditional joint distribution $p_{X,Z|Y}(x,z|y)$ and the product of the conditional joint distributions $p_{X|Y}(x|y)$ and $p_{Z|Y}(z|y)$, that is,

$$I(X;Z|Y) = D\left(p_{X,Z|Y}(x,z|y)||p_{X|Y}(x|y)p_{Z|Y}(z|y)\right).$$

However, the Kullback-Leibler divergence is zero as $p_{X,Z|Y}(x,z|y) = p_{X|Y}(x|y)p_{Z|Y}(z|y)$.

(iii) Using the above two results and the fact that $I(X;Y|Z) \geq 0$, we get $I(X;Z) \leq I(X;Y)$.

(iv) We have equality if $I(X;Y \mid Z) = 0$ which is equivalent to saying that $X \multimap Z \multimap Y$ forms a Markov chain.

The data processing inequality is an important result in information theory that shows that one cannot get back information that has been degraded. In this case, if X is the information of interest, and Y is a degraded version of X (for instance $Y = X + N$, where N is some noise), then one cannot get any more information about X by processing Y and obtaining Z (the mutual information $I(X;Z)$) than from Y itself (the mutual information $I(X;Y)$).