ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Handout 26 Information Theory and Coding

Solutions Graded homework Dec. 4, 2017

Problem 1.

(a) Note that

$$\sum_{x,y} p_{XY}(x,y)\log\frac{q_{X|Y}(x|y)}{p_X(x)} = \sum_{x,y} p_{XY}(x,y)\log\frac{p_{X|Y}(x|y)}{p_X(x)} - \sum_{x,y} p_{XY}(x,y)\log\frac{p_{X|Y}(x|y)}{q_{X|Y}(x|y)}$$

The first term on the right is $I(X;Y)$. Writing $p_{XY}(x,y) = p_Y(y)p_{X|Y}(x|y)$, we see that the second term equals $\sum_y p_Y(y)D(p_{X|Y=y}\|q_{X|Y=y})$. The desired inequality now follows from the positivity of $D(p\|q)$. For equality to hold $D(p_{X|Y=y}\|q_{X|Y=y})$ needs to equal 0 for each y for which $p_Y(y) > 0$, that is, $p_{X|Y}(x|y)p_Y(y) = q_{X|Y}(x|y)p_Y(y)$ for all x, y.

(b) Noting that $D(p_Y\|q_Y) = \sum_y p_Y(y)\log\frac{p_Y(y)}{q_Y(y)} = \sum_{x,y} p_{XY}(x,y)\log\frac{p_Y(y)}{q_Y(y)}$, the inequality follows.

(c) From (a) we see that $I(X;Y) = \max_{q_{X|Y}} \sum_{x,y} p_X(x)W(y|x)\log\frac{q_{X|Y}(x|y)}{p_X(x)}$, so $C(W) = \max_{p_X} I(X;Y)$ can be written as the double maximization

$$C(W) = \max_{p_X, q_{X|Y}} \sum_{x,y} p_X(x)W(y|x)\log\frac{q_{X|Y}(x|y)}{p_X(x)}.$$

Most numerical methods that compute $C(W)$ use this form and alternating between maximizing over p_X and $q_{X|Y}$. Write the objective function in the maximization as $f(p_X, q_{X|Y})$. Start with some initial guess $p_X^{(0)}$ (e.g., the uniform distribution on $\mathcal{X}$), then iterate

$$q_{X|Y}^{(k)} = \arg\max_{q_{X|Y}} f(p_X^{(k)}, q_{X|Y}), \qquad p_X^{(k+1)} = \arg\max_{p_X} f(p_X, q_{X|Y}^{(k)}).$$

Each of these maximizations turns out to be easy and can be found in closed form (e.g., we know from (a) that $q_{X|Y}^{(k)} = p_{X|Y}^{(k)}$), and we repeat until some termination condition (e.g., the approximate satisfaction of the KKT conditions derived in class) is reached. (The method just outlined above is the Arimoto–Blahut algorithm.)

(d) By the KKT conditions derived in class $\sum_y W(y|x)\log\frac{W(y|x)}{p_Y^*(y)} \le C(W)$ for every x. Multiplying both sides by $p_X(x)$ and summing over x gives

$$\sum_{x,y} p_{XY}(x,y)\log\frac{W(y|x)}{p_Y^*(y)} \le C(W)$$

which is the left hand inequality to be shown. The right hand inequality follows from (b) with p_X^* and p_Y playing the roles of p_X, and q_Y.

(e) From (b), $D(p_Y\|p_Y^*) = \sum_{x,y} p(x,y)\frac{W(y|x)}{p_Y^*(y)} - I(X;Y)$. By (d) this is upper bounded by $C(W) - I(X;Y)$. But $I(X;Y) = C(W)$ as p_X maximizes $I(X;Y)$. We thus see that $D(p_Y\|p_Y^*) = 0$. So we have shown that while the input distribution that maximizes $I(X;Y)$ may not be unique the corresponding output distribution is unique.

Problem 2.

(a) The transmitted message does not appear in the list only when the list is empty, which happens when the channel erases k or more times, which has probability $\sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$. Note that this is the probability that $\sum_{i=1}^{n} Z_i \geq k$ where $Z_1, \dots, Z_n$ are $\{0,1\}$-valued, i.i.d., with $\Pr(Z_i = 1) = p$.

(b) Given that Y^n contains j erasures, a particular incorrect message m' will appear on the decoders list if $\mathrm{Enc}(m')_i = Y_i$ for every i for which Y_i is not an erasure. As there are $n-j$ such i's and $\mathrm{Enc}(m')_i$ is chosen independently and is equally likely to be 0 or 1, a particular incorrect message m' will appear in the decoder's list with probability $1/2^{n-j}$. Since there are $M-1$ such messages, the expected number of incorrect messages on the list is $(M-1)/2^{n-j}$.

(c) If the number of erasures $j \geq k$ then there are no messages (and thus no incorrect) messages in the decoder's list. So by (b), $E[\ell]$ (the expectation being over the random choice of Enc) is upper bounded by $(M-1)/2^{n-k} \leq 2^{n(R-1+q)}$. As $q < 1-R$, this quantity decays to zero as n gets large. At the same time, by (a) p_0 is the probability that $\frac{1}{n}\sum_{i=1}^{n} Z_i \geq q$. Since $E[Z_i] = p < q$, by the law of large number p_0 approaches 0 as n gets large. Thus, for large enough n we will have both $p_0 < \epsilon$ and $E[\ell] < \epsilon$, and consequently there is an encoder with $p_0 < \epsilon$ and $\ell < \epsilon$.

(d) Take an encoder as in (c) with $\epsilon/2$ in the role of ϵ. Modify the decoder to declare '?' whenever the list contains two or more messages (i.e., when $\ell \geq 1$) or is empty. Note that this decoder declares m only if it is the only message compatible with y^n. Now, the probability that the list contains two or more messages is upper bounded by $E[\ell] < \epsilon/2$, and the probability that the list is empty is upper bounded by $\epsilon/2$. By the union bound, we see that the decoder declares '?' with probability at most ϵ.

Problem 3.

(a) Note that $1 - \delta = \Pr\bigl(\lambda(X,Y) \geq t\bigr) = \Pr\bigl(Y \in T(X)\bigr)$. Use Bayes' rule to write the second as

$$\sum_x p_X(x) \Pr\bigl(Y \in T(X) \bigm| X = x\bigr) = \sum_x p_X(x) \Pr\bigl(Y \in T(x) \bigm| X = x\bigr).$$

If $\Pr(Y \in T(x) | X = x) < 1 - \epsilon$ for all x, the right hand side above would be strictly less than $1 - \epsilon$. But this is a contradiction, as the right hand side equals $1 - \delta$ and $\epsilon \geq \delta$.

(b) For $y \in T(x)$ we have $p_Y(y) \leq 2^{-t} p_{Y|X}(y|x)$. Thus

$$\Pr\bigl(Y \in T(x)\bigr) = \sum_{y \in T(x)} p_Y(y) \leq 2^{-t} \sum_{y \in T(x)} p_{Y|X}(y|x) = 2^{-t} \Pr\bigl(Y \in T(x) \bigm| X = x\bigr).$$

(c) As $\{\lambda(X,Y) \geq t\}$ is the same as $\{Y \in T(X)\}$, the event we are interested in is $\{Y \in \mathcal{S} \cap T(X)\}$. By $(*)$ and the Bayes rule, we have $\Pr\bigl(Y \in \mathcal{S} \cap T(X)\bigr) < 1 - \epsilon$ as to be shown.

(d) Note that $D_i \subset T(x(i))$. By the union bound $\Pr(Y \in \cup_{i=1}^{m} D_i) \leq \sum_{i=1}^{M} \Pr(Y \in D_i) \leq \sum_{i=1}^{M} \Pr(Y \in T(x(i))$. By (b) each term in this sum is upper bounded by 2^{-t} and the conclusion follows.

(e) The event $\{\lambda(X,Y) \ge t\}$ is included in the union of $\{\lambda(X,Y) \ge t\} \cap \{Y \in \mathcal{S}\}$ and $\{Y \notin \mathcal{S}\}$. Using (c) and (d) to upper bound the probabilities of these two events we find
$$1 - \delta = \Pr(\lambda(X,Y) \ge t) < 1 - \epsilon + 2^{-t}.$$

(f) Observe that $\lambda(X^n, Y^n) = \sum_{i=1}^n \lambda(X_i, Y_i)$, a sum of i.i.d. random variables each with expected value $I(X;Y)$. Thus, by the law of large numbers $\frac{1}{n}\lambda(X^n, Y^n)$ converges to $I(X;Y)$. As $t_n/n \to R$ and $R < I(X;Y)$, we see that $\delta_n \to 0$.

(g) By (e) the algorithm will terminate with $M \ge (\epsilon - \delta_n)2^{t_n} \ge (\epsilon/2)2^{nR+\log(2/\epsilon)} = 2^{nR}$, so the code constructed by the algorithm has rate at least R. Consider now the decoder that declares m when $y^n \in D_m$, and 0 if there is no such m. (As D_m's as disjoint y^n cannot belong to two or more D_m's.) By construction $\Pr(Y^n \in D_m | X^n = x^n(m)) \ge 1 - \epsilon$, so the error probability of this encoder and decoder is at most ϵ.

The 'greedy algorithm' outlined above to pick the codewords $x(1), \dots, x(M)$ and decoding sets $D_1, \dots, D_M$ is due to Feinstein (in late 1950's), and gives an alternative way to prove the coding theorem without using the random coding argument. Note however, that just like random coding, while proving the existence of a good code, the construction does not lead to a practical coding technique.