Problem Set 2

15-859 Information Theory and Applications in TCS

Name: Shashank Singh

Email: sss1@andrew.cmu.edu

Due: Thursday, February 28, 2013

Problem 1

(a) i. $p[\hat{X} \neq X]$ is minimized when $g(Y) = Y$, in which case $p[\hat{X} \neq X] = p[Y \neq X] = \boxed{1 - \frac{1}{2^{n/2}}.}$
$H(X \mid Y) = n/2$ and, for $\mathcal{X} = \operatorname{Dom} X$, $\log_2 |\mathcal{X}| = n$, so, the weak Fano's Inequality gives

$$p[\hat{X} \neq X] \geq \frac{H(X \mid Y) - 1}{\log_2(|\mathcal{X}|)} = \boxed{\frac{n-2}{2n}.}$$

ii. Note that $p[X = Y] = \alpha + (1-\alpha)\frac{1}{2^n}$, since $X = Y$ whenever $X = 0$. $p[\hat{X} \neq X]$ is minimized when $g(Y) = Y$, in which case $p[\hat{X} \neq X] = p[Y \neq X] = \boxed{(1-\alpha)\left(1 - \frac{1}{2^n}\right).}$
Let Z be an indicator random variable with $Z = 0$ if Y was sampled from the first distribution ($Y = X$) and $Z = 1$ otherwise ($Y = 0$). Then,

$$\begin{aligned} H(X \mid Y) &= H(X \mid Y, Z) + H(Z) \\ &= \alpha H(X \mid Y, Z = 0) + (1-\alpha) H(X \mid Y, Z = 0) + h(\alpha) \\ &= \alpha \cdot 0 + (1-\alpha)n + h(\alpha) = (1-\alpha)n + h(\alpha) \end{aligned}$$

(where h denotes the binary entropy function) and, for $\mathcal{X} = \operatorname{Dom}(X)$, $\log_2(|\mathcal{X}|) = n$, using the weakened form of Fano's Inequality,

$$p[\hat{X} \neq X] \geq \frac{H(X \mid Y) - 1}{\log_2(|\mathcal{X}|)} = \boxed{\frac{(1-\alpha)n + h(\alpha) - 1}{n}.}$$

(b) Define $E = \Delta(X, Y)$, the Hamming distance of X and Y. By definition of θ_i, $H(E) = \sum_{i=1}^n \theta_i \log_2\left(\frac{1}{\theta_i}\right)$. For $i \in \{1, \dots, n\}$, $H(X \mid Y, E = i)$ is maximized when X is distributed uniformly among the $\binom{n}{i}$ strings $Z \in \{0,1\}^n$ with $\Delta(Y, Z) = i$, so $H(X \mid Y, E = i) \leq \log_2 \binom{n}{i}$. Since E is a function of X and Y, $H(E \mid X, Y) = 0$. Thus, as in the proof of Fano's Inequality,

$$\begin{aligned} H(X \mid Y) \leq H(E) + H(X \mid E, Y) &= \sum_{i=1}^n \theta_i \log_2\left(\frac{1}{\theta_i}\right) + \sum_{i=1}^n \theta_i \cdot H(X \mid Y, E = i) \\ &\leq \sum_{i=1}^n \theta_i \log_2\left(\frac{1}{\theta_i}\right) + \sum_{i=1}^n \theta_i \log_2 \binom{n}{i} \\ &\leq \sum_{i=1}^n \theta_i \left(\log_2\left(\frac{1}{\theta_i}\right) + \log_2 \binom{n}{i}\right) \\ &\leq \sum_{i=1}^n \theta_i \log_2\left(\binom{n}{i} \frac{1}{\theta_i}\right). \quad \blacksquare \end{aligned}$$

Problem 2

(a) Since $p(0) = p(1) = \frac{1}{2}$,

$$\begin{aligned}
I(B;Y) &= \sum_{b\in\{0,1\}} \sum_{y\in\mathcal{Y}} p(b,y) \log_2\left(\frac{p(b,y)}{p(b)p(y)}\right) \\
&= \sum_{b\in\{0,1\}} \sum_{y\in\mathcal{Y}} p(y\,|\,b)p(b) \log_2\left(\frac{p(y\,|\,b)}{p(y)}\right) \\
&= \sum_{b\in\{0,1\}} \sum_{y\in\mathcal{Y}} p(y\,|\,b)p(b) \log_2\left(\frac{p(y\,|\,b)}{p(y\,|\,1)p(1) + p(y\,|\,0)p(0)}\right) \\
&= \frac{1}{2} \sum_{b\in\{0,1\}} \sum_{y\in\mathcal{Y}} p(y\,|\,b) \log_2\left(\frac{2p(y\,|\,b)}{p(y\,|\,1) + p(y\,|\,0)}\right). \quad \blacksquare
\end{aligned}$$

(b) If, for some $y \in \mathcal{Y}$, $p(y\,|\,0) > p(y\,|\,1)$, then the probability of an error in decoding y is

$$p(\text{error}\,|\,y) = p(1\,|\,y) = \frac{p(y\,|\,1)p(1)}{p(y)} \leq \frac{p(y\,|\,1)}{p(y)} = \frac{\sqrt{p(y\,|\,1)^2}}{p(y)} \leq \frac{\sqrt{p(y\,|\,1)p(y\,|\,0)}}{p(y)}.$$

Similarly, if $p(y\,|\,0) \leq p(y\,|\,1)$, $p(\text{error}\,|\,y) \leq \frac{\sqrt{p(y\,|\,1)p(y\,|\,0)}}{p(y)}$. Thus,

$$p(\text{error}) = \sum_{y\in\mathcal{Y}} p(\text{error}\,|\,y)p(y) \leq \sum_{y\in\mathcal{Y}} \frac{\sqrt{p(y\,|\,1)p(y\,|\,0)}}{p(y)} p(y) = \sum_{y\in\mathcal{Y}} \sqrt{p(y\,|\,1)p(y\,|\,0)}. \quad \blacksquare$$

(c) Let $\rho : C \times C \to \mathbb{N}$ denote the Hamming metric. $\rho(\underline{c}, \underline{c}_0) > 0$ if and only if $\underline{c} \neq \underline{c}_0$, so that

$$p(\underline{c} \neq \underline{c}_0) = \sum_{\underline{d}\in C} p(\underline{c} = \underline{d}) = \sum_{j=1}^{n} d_j \cdot p(\rho(\underline{c}, \underline{c}_0) = j) = \sum_{j=1}^{n} d_j \cdot Z(W)^j,$$

since the bits are assumed to be independent. $\blacksquare$

Problem 3

For notational convenience, we define $\beta := 2^{-n(I(\underline{X};\underline{Y})+3\epsilon}$.

(a) Since we have already shown that $p\left[(\underline{X}, \underline{Y}) \in A_\epsilon^n\right] \to 1$ as $n \to \infty$ and the joint probability of two events whose probabilities approach 1 also approaches 1 as $n \to \infty$, it suffices to show that

$$p\left[\left|\Delta(\underline{X}, \underline{Y}) - \mathbb{E}\left[\Delta(\underline{X}, \underline{Y})\right]\right| < \epsilon\right] \to 1$$

as $n \to \infty$. Since $\Delta(\underline{X}, \underline{Y})$ is an average over n draws from the joint distribution of $(\underline{X}, \underline{Y})$, this is immediate from the Law of Large Numbers. $\blacksquare$

(b) By definition of $A^n_{\epsilon,\Delta}$ and the Triangle Inequality,

$$\begin{aligned}-\log_2\left(\frac{p(\underline{x})p(\underline{y})}{p(\underline{x},\underline{y})}\right) &= \log_2 p(\underline{x},\underline{y}) - \log_2 p(\underline{x}) - \log_2 p(\underline{y}) \\ &\leq n(H(\underline{X},\underline{Y}) - H(\underline{Y}) - H(\underline{Y}) + 3\epsilon) = n(I(\underline{X};\underline{Y}) + 3\epsilon),\end{aligned}$$

so that $\frac{p(\underline{x})p(\underline{y})}{p(\underline{x},\underline{y})} \geq \beta$. Then, by definition of Conditional Probability,

$$p(\underline{y}) = p(\underline{y}\,|\,\underline{x})\frac{p(\underline{x})p(\underline{y})}{p(\underline{x},\underline{y})} \geq p(\underline{y}\,|\,\underline{x})\beta. \quad \blacksquare$$

(c) By definition of $A^n_{\epsilon,\Delta}$ and the fact that $\mathbb{E}\left[\Delta(\underline{X},\underline{Y})\right] \leq D$,

$$\mathbb{E}\left[\Delta(\underline{X},g(f(\underline{X})))\,|\,(\underline{X},g(f(\underline{X}))) \in A^n_{\epsilon,\Delta}\right] \leq \mathbb{E}\left[\Delta(\underline{X},\underline{Y})\right] + \epsilon \leq D + \epsilon.$$

Then, conditioning on whether $g(f(\underline{X})) \in A^n_{\epsilon,\Delta}$ gives

$$\begin{aligned}\mathbb{E}\left[\Delta(\underline{X},g(f(\underline{X})))\right] &= \mathbb{E}\left[\Delta(\underline{X},g(f(\underline{X})))\,|\,(\underline{X},g(f(\underline{X}))) \in A^n_{\epsilon,\Delta}\right](1-p_0) \\ &\quad + \mathbb{E}\left[\Delta(\underline{X},g(f(\underline{X})))\,|\,(\underline{X},g(f(\underline{X}))) \notin A^n_{\epsilon,\Delta}\right]p_0 \\ &\leq \mathbb{E}\left[\Delta(\underline{X},g(f(\underline{X})))\,|\,(\underline{X},g(f(\underline{X}))) \in A^n_{\epsilon,\Delta}\right] + d_{max}p_0 \\ &\leq D + \epsilon + d_{max}p_0. \quad \blacksquare\end{aligned}$$

(d) $\forall \underline{x} \in \mathcal{X}$, $(\underline{x}, f(\underline{x})) \notin A^n_{\epsilon,\Delta}$ if and only if none of the 2^{nR} values y in the code book $\mathcal{C}$ satisfies $(\underline{x},\underline{y}) \in A^n_{\epsilon,\Delta}$, which occurs with probability

$$p\left[(\underline{X},f(\underline{X})) \in A^n_{\epsilon,\Delta}\,|\,\underline{X}=\underline{x}\right] = \left[1 - p\left[A(\underline{x},\underline{y})=1\right]\right]^{2^{nR}} = \left[1 - \sum_{\underline{y}} p(\underline{y})A(\underline{x},\underline{y})\right]^{2^{nR}}.$$

Thus, conditioning on the value of $\underline{X}$,

$$p_0 = \sum_{\underline{x}} p(\underline{x})p\left[(\underline{X},\underline{Y}) \in A^n_{\epsilon,\Delta}\,|\,\underline{X}=\underline{x}\right] = \sum_{\underline{x}} p(\underline{x})\left[1 - \sum_{\underline{y}} p(\underline{y})A(\underline{x},\underline{y})\right]^{2^{nR}}. \quad \blacksquare$$

(e) By definition of A and the inequality derived in part (b), $\forall \underline{x} \in \mathcal{X}$,

$$\sum_{\underline{y}} p(\underline{y})A(\underline{x},\underline{y}) = \sum_{\underline{y}:(\underline{x},\underline{y})\in A^n_{\epsilon,\Delta}} p(\underline{y}) \geq \sum_{\underline{y}:(\underline{x},\underline{y})\in A^n_{\epsilon,\Delta}} \beta p(\underline{y}\,|\,\underline{x}) = \beta\sum_{\underline{y}} p(\underline{y}\,|\,\underline{x}) \cdot A(\underline{x},\underline{y}).$$

It then follows immediately from the result of part (d) that

$$p_0 \leq \sum_{\underline{x}} p(\underline{x})\left(1 - \beta\sum_{\underline{y}} p(\underline{y}\,|\,\underline{x}) \cdot A(\underline{x},\underline{y})\right)^{2^{nR}}. \quad \blacksquare$$

(f) By part (e) and the given inequality,

$$\begin{aligned} p_0 &\le \sum_{\underline{x}} p(\underline{x}) \left(1 - \sum_{\underline{y}} p(\underline{y} \,|\, \underline{x}) \cdot A(\underline{x}, \underline{y}) + e^{-\beta \cdot 2^{nR}} \right) \\ &\le \sum_{\underline{x}} p(\underline{x}) \left(1 - \sum_{\underline{y}} p(\underline{y} \,|\, \underline{x}) \cdot A(\underline{x}, \underline{y}) + e^{-2^{n(I(X;Y)+3\epsilon-R)}} \right). \end{aligned}$$

Since $R > I(X;Y) + 3\epsilon$, $e^{-2^{n(R-I(X;Y)+3\epsilon)}} \to 0$ as $n \to \infty$. $\sum_{\underline{y}} p(\underline{y} \,|\, \underline{x}) \cdot A(\underline{x}, \underline{y})$ is the probabilty that $\exists y \in \mathcal{C}$ with $(\underline{x}, \underline{y}) \in A^n_{\epsilon,\Delta}$, which approaches 1 for all $x \in \mathcal{X}$. Thus, the upper bound on p_0 approaches 0 as $n \to \infty$, so that, for sufficiently large n, $p_0 \le \epsilon$. ■

(g) By part (f), $\forall \epsilon > 0$, by choosing an appropriate conditional distribution for y given x, we can have $R \in (R^*, R^* + 3\epsilon)$, and expected distortion at most $D + \epsilon + d_{max}\epsilon$, so that, for sufficiently long messages, $R \to R^*$ and distortion approaches D. ■

(h) For a single bit, $p(X \neq Y) = D$, so that

$$I(X;Y) = H(X) - H(X \,|\, Y) = H(X) - H(X \neq Y) = h(p) - h(D).$$

Thus, the uniform distribution achieves rate $h(p) - h(D)$, so that the optimal rate $R^* \le h(p) - h(D)$. ■

Problem 4

(a) I didn't quite understand this question. Doesn't the existence of linear, capacity-achieving codes (such as Arikan's construction) immediately prove the result for all channels?

(b) i. This can be shown by induction on m, by repeatedly choosing random binary vectors of dimension k and computing the probability distribution of the dimension of their span (since $\text{rank}(M) = k$ if and only if M has k linearly independent rows). ■

ii. Erasing $|J|$ bits from Gx is equivalent to removing the corresponding $|J|$ rows from G (creating a matrix $H \in \{0,1\}^{(n-|J|)\times k}$). Then, a decoding error occurs precisely when $\text{rank}(H) < k$, which, by part i., occurs with probability $2^{k-(n-|J|)} = 2^{k-n+|J|}$. ■

iii. Since $k = Rn$ and $|J| = \alpha n$, by the result of part ii.,

$$\mathop{\mathbb{E}}_{G \in \{0,1\}^{n\times k}} [P_{\text{err}}(G)] \le 2^{k-n+|J|} = 2^{Rn-n+\alpha n} = 2^{(R-(1-\alpha))n} \to 0$$

exponentially as $n \to \infty$, since $R - (1 - \alpha) < 0$. ■

iv. By part iii., $\forall \varepsilon > 0$, for sufficiently large k, length-k messages can be sent at rate $R < 1-\alpha$ according to a random linear code with expected probability of a decoding failure

$$\mathop{\mathbb{E}}_{G \in \{0,1\}^{n\times k}} [P_{\text{err}}(G)] < \varepsilon$$

(where $n = k/R$). It follows that $\exists G \in \{0,1\}^{n\times k}$ with $P_{\text{err}}(G) < \varepsilon$, so that a message encoded by G can be decoded correctly with probability $1 - \varepsilon$. ■