

Solutions for Final Examination of April 17, 2006

1. Short Snappers

- (a) **False.** For example, $C = \{0, 0\}$ satisfies the Kraft inequality, but is singular, and hence not uniquely decodable.
- (b) **True.** We have $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, so if $H(X) = H(Y)$, then $H(X|Y) = H(Y|X)$. (We can also see this from the equality $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.)
- (c) **True.** Let X be the channel input and let Y be the channel output. We have $I(X; Y) = H(Y) - H(r)$, where r is any row of the channel transition matrix. To maximize $I(X; Y)$ we must maximize $H(Y)$, i.e., make Y uniform. (This is achievable, e.g., by making X uniform.)
- (d) **False.** It is sufficient to make X uniform, but not always necessary. For example, the capacity of the noisy typewriter channel discussed in the text is achieved with a nonuniform input distribution.
- (e) **True.** The maximum entropy distribution with a fixed variance σ^2 is Gaussian, and the corresponding differential entropy $h = \frac{1}{2} \log(2\pi\sigma^2 e)$ is finite. (Here I have assumed a nonzero variance; as $\sigma^2 \rightarrow 0$, $h \rightarrow -\infty$.)
- (f) **True.** Let C_1 be a rate-distortion code of length n_1 that achieves (R_1, D_1) , and let C_2 be a rate-distortion code of length n_2 that achieves (R_2, D_2) . Let n be the lowest common multiple of n_1 and n_2 . Form a rate-distortion code C of length $2n$ by using C_1 a total of $a = n/n_1$ times, followed by using C_2 a total of $b = n/n_2$ times. (In other words C_1 and C_2 are time-shared, with each one active half of the time.) The rate of C is

$$R = \frac{1}{2n}(an_1R_1 + bn_2R_2) = \frac{1}{2n}(nR_1 + nR_2) = (R_1 + R_2)/2,$$

and the expected distortion is

$$D = \frac{1}{2n}(an_1D_1 + bn_2D_2) = \frac{1}{2n}(nD_1 + nD_2) = (D_1 + D_2)/2.$$

- (g) **True.** We know that for any $\epsilon > 0$ and any $R < C$ there exists (for some sufficiently large n) a binary $(2^{nR}, n)$ code C with maximal error probability at most ϵ . Let C' be the code obtained from C by appending an extra bit to each codeword of C , where the extra bit is chosen so that each codeword of C' has an even number of ones. Then C' has length $n + 1$ and 2^{nR} codewords, which corresponds to a rate of $R' = R(1 - \frac{1}{n+1})$. The rate R' can be made to approach arbitrarily closely to C by choosing R and n large enough. Furthermore, the maximal error probability for C' is at most ϵ (since the decoder can always ignore the extra bit, and simply decode using the decoding rule for C). Thus, constraining all codewords to have an even number of ones does *not* restrict the set of rates that are achievable on the binary symmetric channel.

- (h) Obtained by the Huffman procedure or otherwise, an optimum prefix code for this source is given in the following table:

x	a	b	c	d	e
$C(x)$	0	100	101	110	111

- (i) This channel is the sum of a channel with capacity $C_1 = 0$ and a symmetric channel with capacity

$$C_2 = \log_2 3 - \mathbf{H}\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) = \log_2 3 - \frac{3}{2}.$$

The capacity of the channel is therefore

$$C = \log_2(2^{C_1} + 2^{C_2}) = \log_2(1 + 2^{\log_2 3 - 1.5}) = \log_2(1 + 3/2\sqrt{2}).$$

- (j) Since the channel output is binary, $0 \leq R_1 \leq 1$, with $R_1 = 1$ achieved by setting $X_2 = 0$. Similarly $0 \leq R_2 \leq 1$, and $R_2 = 1$ is achieved by setting $X_1 = 0$. Again, since the channel output is binary, we find that $R_1 + R_2 \leq 1$. The capacity region is therefore the triangular region defined by

$$R_1 \geq 0, \quad R_2 \geq 0, \quad R_1 + R_2 \leq 1.$$

Any point on the $R_1 + R_2 = 1$ frontier can be obtained by time-sharing between $(R_1, R_2) = (1, 0)$ and $(R_1, R_2) = (0, 1)$.

2. Entropy of Functions of Random Variables

- (a) Since $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$, and since $H(Y|X) = 0$ as Y is a deterministic function of X , it follows that

$$H(X) = H(Y) + H(X|Y).$$

Thus $H(X) = H(Y)$ if $H(X|Y) = 0$, which means that X can be recovered deterministically from Y , i.e., $f(x)$ is one-to-one, and hence invertible.

- (b) Let $\mathcal{S}_X = \{1, 2, \dots, m\}$, with $P[X = i] = p_i$ and let $Y = f(X)$, where

$$f(X) = \begin{cases} 0 & 1 \leq X \leq r \\ 1 & r + 1 \leq X \leq m \end{cases}$$

Then $Y = 0$ with probability $z_1 = \sum_{i=1}^r p_i$, and $Y = 1$ with probability $z_2 = \sum_{i=r+1}^m p_i$. Furthermore, the conditional probability mass function for X , given that $Y = 0$, is

$$P[X = i|Y = 0] = \begin{cases} p_i/z_1 & 1 \leq i \leq r \\ 0 & r + 1 \leq i \leq m \end{cases}$$

and the conditional probability mass function for X , given that $Y = 1$, is

$$P[X = i|Y = 1] = \begin{cases} 0 & 1 \leq i \leq r \\ p_i/z_2 & r + 1 \leq i \leq m \end{cases}.$$

From the result of part (a), we have

$$\begin{aligned} H(X) &= H(Y) + H(X|Y) \\ &= H(Y) + P[Y = 0]H(X|Y = 0) + P[Y = 1]H(X|Y = 1), \end{aligned}$$

so that

$$\begin{aligned} \mathbf{H}(p_1, p_2, \dots, p_m) &= \mathbf{H}(z_1, z_2) + z_1 \mathbf{H}(p_1/z_1, p_2/z_1, \dots, p_r/z_1) \\ &\quad + z_2 \mathbf{H}(p_{r+1}/z_2, p_{r+2}/z_2, \dots, p_m/z_2) \end{aligned}$$

- (c) Let $Y = X_1 + X_2$, where X_1 and X_2 are discrete random variables. Note that Y is a deterministic function of X_1 and X_2 . Thus from (a) it follows that

$$H(X_1, X_2) = H(Y) + H(X_1, X_2|Y)$$

so

$$\begin{aligned} H(Y) &= H(X_1, X_2) - H(X_1, X_2|Y) \\ &\leq H(X_1, X_2) \quad (\text{equality iff } H(X_1, X_2|Y) = 0) \\ &\leq H(X_1) + H(X_2) \quad (\text{equality iff } X_1 \text{ and } X_2 \text{ are independent}) \end{aligned}$$

Thus $H(Y) = H(X_1) + H(X_2)$ if and only if (i) X_1 and X_2 are independent and (ii) X_1 and X_2 can be determined uniquely from their sum Y .

An example where this holds is obtained as follows. Take $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 2\}$, independently and with uniform probability. Then $Y = X_1 + X_2$ is uniform in $\{0, 1, 2, 3\}$, and $H(Y) = 2 = H(X_1) + H(X_2)$. (Note that X_1 and X_2 are both uniquely determined from the value of Y .)

3. A Correlated Noise Channel

- (a) If the channels are to behave independently, then the probability that an error occurs simultaneously in the two channels (given by parameter a) must be p^2 , i.e., $a = p^2$. The capacity of the resulting channel is the sum of the individual channel capacities, namely $C = 2(1 - \mathcal{H}(p))$.
- (b) The possible inputs and outputs of the overall channel are $\{00, 01, 10, 11\}$. The channel transition matrix is given as

$$M = \begin{bmatrix} 1 + a - 2p & p - a & p - a & a \\ p - a & 1 + a - 2p & a & p - a \\ p - a & a & 1 + a - 2p & p - a \\ a & p - a & p - a & 1 + a - 2p \end{bmatrix}.$$

We see that channel is symmetric, and hence the capacity is

$$C = 2 - \mathcal{H}(1 + a - 2p, p - a, p - a, a).$$

- (c) When $a = p = 1/2$ we have $C = 2 - \mathcal{H}(1/2, 0, 0, 1/2) = 1$ bit per channel use. In this case, the sub channels are individually useless, but the noise in the two channels is identical, i.e., $Z_1 = Z_2$. If we transmit a single bit in channel 1, and the symbol 0 in channel 2, the observation of Y_2 will completely determine the value of the noise $Z_2 = Z_1$, which can then be subtracted from Y_1 . Thus, given a message bit b , the scheme is to transmit $(X_1, X_2) = (b, 0)$, receive $(Y_1, Y_2) = (b \oplus Z_1, Z_1)$ and decode $Y_1 \oplus Y_2 = b$, thereby achieving capacity.
- (d) The channel capacity can be written as

$$\begin{aligned} C &= 2 - H(Z_1, Z_2) \\ &= 2 - H(Z_1) - H(Z_2|Z_1) \\ &= 2 - \mathcal{H}(p) - H(Z_2|Z_1). \end{aligned}$$

To minimize C , we should maximize $H(Z_2|Z_1)$ by making Z_2 independent of Z_1 , i.e., select $a = p^2$.

To maximize C , we should minimize $H(Z_2|Z_1)$. This minimum is zero, obtained when $a = p$. (If $p = 1/2$, a maximum of C occurs also at $a = 0$.)

The largest capacity occurs when $Z_1 = Z_2$, i.e., when the noise in the two channels is identical. Indeed, the capacity is one bit more than the capacity of one of the channels used in isolation. To exploit this strong correlation, we would use an ordinary (close to) capacity-achieving code in one of the channels. After decoding this code (which we can do correctly with probability arbitrarily close to one), we can subtract the transmitted codeword from the received word, to get a (near) perfect estimate of the noise in the first channel. But this noise sequence is identical to that in the second channel, so it can be subtracted from the output of the second channel, allowing for decoding of any information transmitted in that channel.

This problem tells us that having parallel channels with correlated noise is a good thing, as the correlation can be exploited in decoding (estimating the noise in one channel helps us to better decode the information transmitted in the other channel).

4. Rate-Distortion

At $D = 0$, we require $R \geq H(X) = \log_2 m$. At $R = 0$, if we guess a symbol at random, the probability of error, which is the expected Hamming distortion, is $D_{\max} = (m - 1)/m$.

Let us assume that $D \leq D_{\max}$. We have

$$\begin{aligned} I(X; \hat{X}) &= H(X) - H(X|\hat{X}) \\ &= H(X) - H(X - \hat{X}|\hat{X}) \\ &\geq H(X) - H(X - \hat{X}) \\ &= \log_2 m - H(X - \hat{X}) \end{aligned}$$

where $X - \hat{X}$ denotes the modulo- m difference between X and \hat{X} . The distortion is restricted to D or less, so we must have $P[X - \hat{X} \neq 0] \leq D$. The maximum entropy distribution under this constraint is $(1 - D, D/(m - 1), D/(m - 1), \dots, D/(m - 1))$, where the first component corresponds to $X - \hat{X} = 0$. Thus we find that

$$\begin{aligned} I(X; \hat{X}) &\geq \log_2 m - H\left(1 - D, \frac{D}{m - 1}, \dots, \frac{D}{m - 1}\right) \\ &= \log_2 m - \mathcal{H}(D) - D \log(m - 1). \end{aligned} \tag{1}$$

(This second equality follows from the grouping property of the entropy function, or by expansion and grouping of like terms.) To show that this lower bound is achievable (and hence that this lower bound is in fact the rate-distortion function), we observe that $I(X; \hat{X})$ represents the mutual information between the output X and the input \hat{X} of an m -ary symmetric channel, with (for $x, \hat{x} \in \{1, 2, \dots, m\}$)

$$P[X = x|\hat{X} = \hat{x}] = \begin{cases} 1 - D & \text{if } x = \hat{x}, \\ D/(m - 1) & \text{if } x \neq \hat{x}. \end{cases}$$

Since a uniform input distribution over \hat{X} induces the required uniform distribution over X , the lower bound given is in fact achievable. Thus

$$R(D) = \begin{cases} \log_2 m - \mathcal{H}(D) - D \log_2(m - 1) & \text{if } 0 \leq D \leq (m - 1)/m, \\ 0 & \text{if } D > (m - 1)/m. \end{cases}$$

Remark: the inequality (1) also follows from Fano's inequality, which states (with $P_e = D$) that

$$H(X|\hat{X}) \leq \mathcal{H}(D) + D \log_2(m-1).$$

From this it follows that

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) \geq H(X) - \mathcal{H}(D) - D \log_2(m-1).$$

Equivalently, we could have obtained Fano's inequality from (1). Thus Fano's inequality may be regarded as a result of rate-distortion theory.

5. **A Periodic Jammer**—This problem is about parallel Gaussian channels: the OFF periods correspond to channels with noise variance σ_0^2 ; the ON periods correspond to channels with noise variance σ_1^2 . The optimum transmission strategy in this case will be the one that corresponds to water-pouring.

- (a) Over three time units, the total available energy budget is $3P$. During the OFF periods, the transmitter sees two parallel Gaussian noise channels of equal noise variance, and the optimum strategy is to assign equal energy (of $3P/2$) to each channel. These channels each permit reliable transmission at rates up to

$$\frac{1}{2} \log_2 \left(1 + \frac{3P/2}{\sigma_0^2} \right) \text{ bits per channel use.}$$

Since 2 channel uses are available every 3 units of time, the overall achievable rate R must satisfy

$$R \leq \frac{1}{3} \log_2 \left(1 + \frac{3P}{2\sigma_0^2} \right) \text{ bits per unit time.}$$

- (b) In this strategy, the transmitter sees two channel uses that permit transmission at a rate of $\frac{1}{2} \log_2(1 + P/\sigma_0^2)$ bits per channel use and one channel use that permits transmission at a rate of $\frac{1}{2} \log_2(1 + P/\sigma_1^2)$. The overall achievable rate R must satisfy

$$R \leq \frac{1}{3} \log_2 \left(1 + \frac{P}{\sigma_0^2} \right) + \frac{1}{6} \log_2 \left(1 + \frac{P}{\sigma_1^2} \right).$$

- (c) The optimum transmission strategy is to perform water-pouring, as shown in Fig. 1. Over three units of time, the available energy budget is $3P$. If P is small enough, the strategy of part (a) is in fact optimal. When P becomes larger, it pays to transmit even when the jammer is ON. The break-point between these two strategies occurs when $3P/2 = \sigma_1^2 - \sigma_0^2$, or $P = 2(\sigma_1^2 - \sigma_0^2)/3$. Power in excess of this value is shared equally among all three channels. Thus, if P_0 denotes the power allocated when the jammer is OFF and P_1 denotes the power allocated when the jammer is ON, we have

$$P_0 = \begin{cases} \frac{3}{2}P & \text{if } P \leq \frac{2}{3}(\sigma_1^2 - \sigma_0^2), \\ P + \frac{1}{3}(\sigma_1^2 - \sigma_0^2) & \text{if } P > \frac{2}{3}(\sigma_1^2 - \sigma_0^2); \end{cases}$$

and

$$P_1 = \begin{cases} 0 & \text{if } P \leq \frac{2}{3}(\sigma_1^2 - \sigma_0^2), \\ P - \frac{2}{3}(\sigma_1^2 - \sigma_0^2) & \text{if } P > \frac{2}{3}(\sigma_1^2 - \sigma_0^2). \end{cases}$$

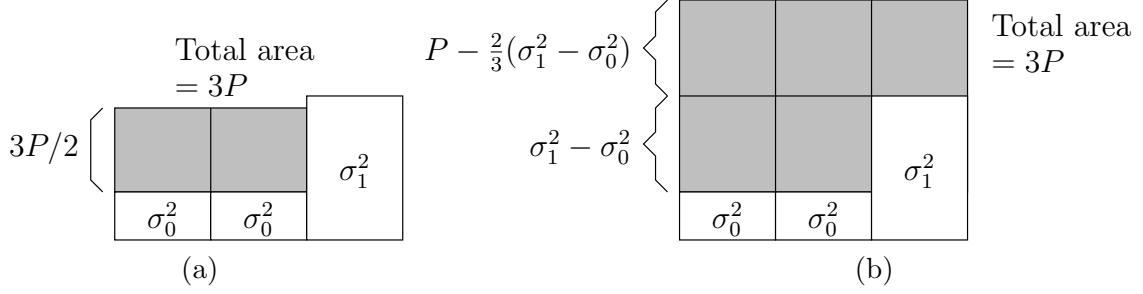


Figure 1: Waterpouring solution (a) when $3P/2 < \sigma_1^2 - \sigma_0^2$ and (b) when $3P/2 > \sigma_1^2 - \sigma_0^2$.

The overall achievable rate R must satisfy

$$R \leq \frac{1}{3} \log_2 \left(1 + \frac{P_0}{\sigma_0^2} \right) + \frac{1}{6} \log_2 \left(1 + \frac{P_1}{\sigma_1^2} \right).$$

The maximum achievable rate R is plotted as a function of P in Fig. 2, for the case where $\sigma_1^2 = 3$ and $\sigma_0^2 = 1$. Also plotted in Fig. 2 is the maximum achievable rate for the transmission strategy of part (a).

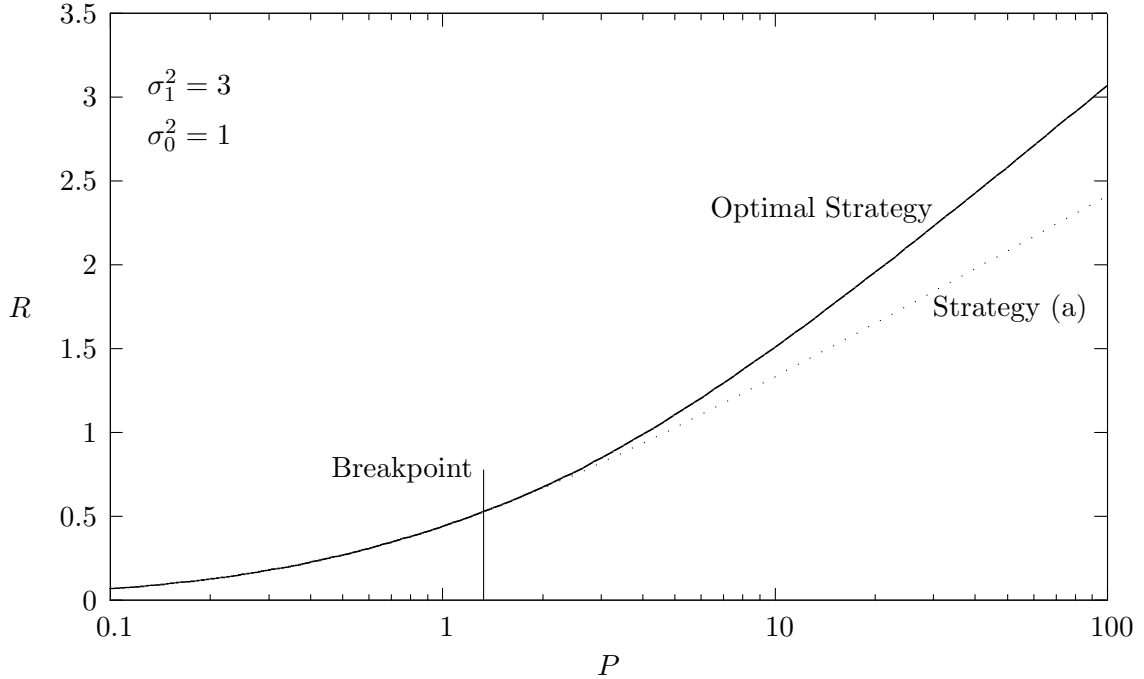


Figure 2: Maximum achievable rate R for the optimum transmission strategy and for the transmission strategy of part (a).

In case the duty cycle of the jammer is a positive rational number $r = a/b$, we effectively have b parallel Gaussian channels, a of them with noise variance σ_1^2 and $b-a$ of them with noise variance σ_0^2 . Over b time units, the total energy budget is bP . If P is small enough, the optimal strategy is to transmit only during the $b-a$ OFF periods, transmitting at a power of $bP/(b-a)$ for each such channel use. If $P > \frac{b-a}{b}(\sigma_1^2 - \sigma_0^2)$ then it pays to transmit even during the ON periods.

Again denoting the power transmitted while the jammer is OFF as P_0 and the power transmitted while the transmitter is ON as P_1 , the optimum power allocation is

$$P_0 = \begin{cases} \frac{bP}{b-a} & \text{if } P \leq \frac{b-a}{b}(\sigma_1^2 - \sigma_0^2) \\ P + \frac{a}{b}(\sigma_1^2 - \sigma_0^2) & \text{if } P > \frac{b-a}{b}(\sigma_1^2 - \sigma_0^2) \end{cases}$$

and

$$P_1 = \begin{cases} 0 & \text{if } P \leq \frac{b-a}{b}(\sigma_1^2 - \sigma_0^2) \\ P - \frac{b-a}{b}(\sigma_1^2 - \sigma_0^2) & \text{if } P > \frac{b-a}{b}(\sigma_1^2 - \sigma_0^2) \end{cases}$$

Expressed in terms of the duty cycle r , we have

$$P_0 = \begin{cases} \frac{P}{1-r} & \text{if } P \leq (1-r)(\sigma_1^2 - \sigma_0^2) \\ P + r(\sigma_1^2 - \sigma_0^2) & \text{if } P > (1-r)(\sigma_1^2 - \sigma_0^2) \end{cases}$$

and

$$P_1 = \begin{cases} 0 & \text{if } P \leq (1-r)(\sigma_1^2 - \sigma_0^2) \\ P - (1-r)(\sigma_1^2 - \sigma_0^2) & \text{if } P > (1-r)(\sigma_1^2 - \sigma_0^2) \end{cases}$$

The overall achievable rate R must satisfy

$$\begin{aligned} R &\leq \frac{b-a}{2b} \log_2 \left(1 + \frac{P_0}{\sigma_0^2} \right) + \frac{a}{2b} \log_2 \left(1 + \frac{P_1}{\sigma_1^2} \right) \\ &= \frac{1-r}{2} \log_2 \left(1 + \frac{P_0}{\sigma_0^2} \right) + \frac{r}{2} \log_2 \left(1 + \frac{P_1}{\sigma_1^2} \right). \end{aligned}$$