EE 231A
Information Theory
Instructor: Rick Wesel

Handout #20, Solution Set 7
Wednesday, May 20, 2020

80 pts
Reading: Chapter 13

Lecture 13A-B: Computing $R(D)$ for a discrete-alphabet source

1. (10 pts) *4-ary Hamming distortion.*
   A random variable X uniformly takes on values $\{0, 1, 2, 3\}$. The distortion function is the usual Hamming distortion.

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \end{cases} \tag{1}$$

   Compute the rate distortion function $R(D)$ by finding a lower bound on $I(x; \hat{x})$ and showing this lower bound to be achievable. *Hint:* Fano's inequality.

   Answer:

   Consider any joint distribution that satisfies the distortion constraint D. Since $D = P(X \neq \hat{X})$, we have by Fano's inequality that

$$H(X|\hat{X}) \leq H(D) + D \log(m-1), \tag{2}$$

   where m is the size of the alphabets $\hat{\mathcal{X}}, \mathcal{X}$. Now we apply this inequality and the fact that $H(X) = \log |\mathcal{X}| = 2$ for a uniform distribution to $I(X; \hat{X})$.

$$\begin{aligned} I(X; \hat{X}) &= H(X) - H(X|\hat{X} &\quad (3) \\ &\geq 2 - H(D) - D \log(3) &\quad (4) \end{aligned}$$

   We can achieve this lower bound by choosing $p(\hat{x}$ to be uniform. To do this we choose $p(x|\hat{x})$ to be

$$p(x|\hat{x}) = \begin{cases} 1 - D & \text{if } \hat{x} = x \\ \frac{D}{3} & \text{if } \hat{x} \neq x \end{cases} \tag{5}$$

   Thus

$$R(D) = \begin{cases} 2 - H(D) - D \log(3) & \text{for } 0 \leq D \leq \frac{3}{4} \\ 0 & \text{for } D \geq \frac{3}{4} \end{cases} \tag{6}$$

2. (10 pts) *Rate distortion for uniform source with Hamming distortion.* X is uniformly distributed on the set $\{1, 2, \ldots, m\}$. The distortion measure is

$$d(x, \hat{x}) = \left\{ \begin{array}{ll} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \end{array} \right.$$

Consider any joint distribution that satisfies the distortion constraint D. Since $D = \Pr(X \neq \hat{X})$, we have by Fano's inequality

$$H(X|\hat{X}) \leq H(D) + D \log(m-1),$$

and hence

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) & (7) \\ &\geq \log m - H(D) - D\log(m-1). & (8) \end{aligned}$$

We can achieve this lower bound by choosing $p(\hat{x})$ to be the uniform distribution, and the conditional distribution of $p(x|\hat{x})$ to be

$$p(x|\hat{x}) \begin{cases} = 1 - D & \text{if } \hat{x} = x \\ = D/(m-1) & \text{if } \hat{x} \neq x. \end{cases}$$

It is easy to verify that this gives the right distribution on X and satisfies the bound with equality for $D < 1 - \frac{1}{m}$. Hence

$$R(D) \begin{cases} = \log m - H(D) - D\log(m-1) & \text{if } 0 \leq D \leq 1 - \frac{1}{m} \\ 0 & \text{if } D > 1 - \frac{1}{m}. \end{cases}$$

3. (10 pts) *Scaled Hamming Distortion.*

   X is uniformly distributed on the set $\{1, 2, 3\}$. The distortion measure is

$$d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 2 & \text{if } x \neq \hat{x} \end{cases} \tag{9}$$

   Consider any joint distribution that satisfies the distortion constraint D. Since $D = 2\Pr(X \neq \hat{X})$, we have by Fano's inequality

$$H(X|\hat{X}) \leq H\left(\frac{D}{2}\right) + \frac{D}{2}\log(m-1),$$

   and hence

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) & (10) \\ &\geq \log m - H\left(\frac{D}{2}\right) - \frac{D}{2}\log(m-1) & (11) \\ &= \log 3 - H\left(\frac{D}{2}\right) - \frac{D}{2}. & (12) \end{aligned}$$

   We can achieve this lower bound by choosing $p(\hat{x})$ to be the uniform distribution, and the conditional distribution of $p(x|\hat{x})$ to be

$$p(x|\hat{x}) = \begin{cases} 1 - \frac{D}{2} & \text{if } \hat{x} = x \\ \frac{D}{4} & \text{if } \hat{x} \neq x \end{cases}. \tag{13}$$

   Thus $R(D) = \log 3 - H\left(\frac{D}{2}\right) - \frac{D}{2}$.

4. (10 pts) *Rate Distortion function with infinite distortion.*
We wish to evaluate the rate distortion function

$$R(D) = \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X;\hat{X}).$$

Since $d(0,1) = \infty$, we must have $p(0,1) = 0$ for a finite distortion. Thus, the distortion $D = p(1,0)$, and hence we have the following joint distribution for $(X, \hat{X})$ (assuming $D \leq \frac{1}{2}$).

$$p(x,\hat{x}) = \begin{bmatrix} \frac{1}{2} & 0 \\ D & \frac{1}{2} - D \end{bmatrix}$$

The mutual information for this joint distribution is

$$\begin{aligned} R(D) = I(X;\hat{X}) &= H(X) - H(X|\hat{X}) && (14) \\ &= H(\frac{1}{2},\frac{1}{2}) - (\frac{1}{2} + D)H\left(\frac{\frac{1}{2}}{\frac{1}{2}+D}, \frac{D}{\frac{1}{2}+D}\right) && (15) \\ &= 1 + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}+D} + D\log\frac{D}{\frac{1}{2}+D}, && (16) \end{aligned}$$

which is the rate distortion function for this binary source if $0 \leq D \leq \frac{1}{2}$. Since we can achieve $D = \frac{1}{2}$ with zero rate (use $p(\hat{x} = 0) = 1$), we have $R(D) = 0$ for $D \geq \frac{1}{2}$.

5. (10 pts) *Erasure distortion.* Consider $X \sim \text{Bernoulli}(\frac{1}{2})$, and the distortion measure

$$d(x,\hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}.$$

The infinite distortion constrains $p(0,1) = p(1,0) = 0$. Hence by symmetry the joint distribution of $(X, \hat{X})$ is of the form shown in Figure 1.

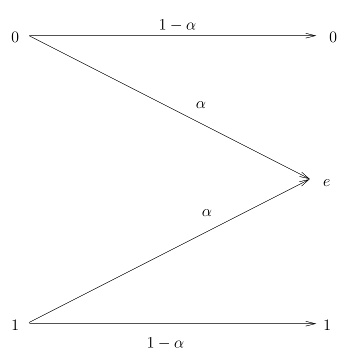

Figure 1: Joint distribution for erasure rate distortion of a binary source

For this joint distribution, it is easy to calculate the distortion $D = \alpha$ and that $I(X;\hat{X}) = H(X) - H(X|\hat{X}) = 1 - \alpha$. Hence we have $R(D) = 1 - D$ for $0 \leq D \leq 1$. For $D > 1$, $R(D) = 0$.

Now lets see how we could achieve this rate distortion function. If D is rational, say k/n, then we send only the first $n-k$ of any block of n bits. We reproduce these bits exactly and reproduce the remaining bits as erasures. Hence we can send information at rate $1-D$ and achieve a distortion D. If D is irrational, we can get arbitrarily close to D by using longer and longer block lengths.

6. (10 pts) *Bounds on the rate distortion function for squared error distortion.*

   We assume that X has zero mean and variance σ^2. To prove the lower bound, we use the same techniques as used for the Guassian rate distortion function. Let $(X, \hat{X})$ be random variables such that $E(X-\hat{X})^2 \leq D$. Then

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) && (17)\\ &= h(X) - h(X-\hat{X}|\hat{X}) && (18)\\ &\geq h(X) - h(X-\hat{X}) && (19)\\ &\geq h(X) - h(\mathcal{N}(0, E(X-\hat{X})^2)) && (20)\\ &= h(X) - \frac{1}{2}\log(2\pi e)E(X-\hat{X})^2 && (21)\\ &\geq h(X) - \frac{1}{2}\log(2\pi e)D. && (22) \end{aligned}$$

   To prove the upper bound, we consider the joint distribution as shown in the figure and calculate the distortion and the mutual information between X and $\hat{X}$. Since

$$\hat{X} = \frac{\sigma^2 - D}{\sigma^2}(X+Z), \tag{23}$$

   we have

$$\begin{aligned} E(X-\hat{X})^2 &= E\left(\frac{D}{\sigma^2}X - \frac{\sigma^2-D}{\sigma^2}Z\right)^2 && (24)\\ &= \left(\frac{D}{\sigma^2}\right)^2 EX^2 + \left(\frac{\sigma^2-D}{\sigma^2}\right)^2 EZ^2 && (25)\\ &= \left(\frac{D}{\sigma^2}\right)^2 \sigma^2 + \left(\frac{\sigma^2-D}{\sigma^2}\right)^2 \frac{D\sigma^2}{\sigma^2-D} && (26)\\ &= D, && (27) \end{aligned}$$

   since X and Z are independent and zero mean. Also the mutual information is

$$\begin{aligned} I(X;\hat{X}) &= h(\hat{X}) - h(\hat{X}|X) && (28)\\ &= h(\hat{X}) - h(\frac{\sigma^2-D}{\sigma^2}Z). && (29) \end{aligned}$$

Now

$$E\hat{X}^2 = \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 E(X+Z)^2 \tag{30}$$

$$= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 (EX^2 + EZ^2) \tag{31}$$

$$= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \left(\sigma^2 + \frac{D\sigma^2}{\sigma^2 - D}\right) \tag{32}$$

$$= \sigma^2 - D. \tag{33}$$

Hence, we have

$$I(X;\hat{X}) = h(\hat{X}) - h(\frac{\sigma^2 - D}{\sigma^2} Z) \tag{34}$$

$$= h(\hat{X}) - h(Z) - \log\frac{\sigma^2 - D}{\sigma^2} \tag{35}$$

$$\leq h(\mathcal{N}(0, \sigma^2 - D)) - \frac{1}{2}\log(2\pi e)\frac{D\sigma^2}{\sigma^2 - D} - \log\frac{\sigma^2 - D}{\sigma^2} \tag{36}$$

$$= \frac{1}{2}\log(2\pi e)(\sigma^2 - D) - \frac{1}{2}\log(2\pi e)\frac{D\sigma^2}{\sigma^2 - D} - \frac{1}{2}\log\left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \tag{37}$$

$$= \frac{1}{2}\log\frac{\sigma^2}{D}, \tag{38}$$

which combined with the definition of the rate distortion function gives us the required upper bound.

For a Gaussian random variable, $h(X) = \frac{1}{2}\log(2\pi e)\sigma^2$ and the lower bound is equal to the upper bound. For any other random variable, the lower bound is strictly less than the upper bound and hence non-Gaussian random variables cannot require more bits to describe to the same accuracy than the corresponding Gaussian random variables. This is not surprising, since the Gaussian random variable has the maximum entropy and we would expect that it would be the most difficult to describe.

Synthesis of ideas about Gaussian channel capacity and lossy compression of a Gaussian source

7. (10 pts) *Simplicity is best.* *Fall 99 oral*
This problem considers transmission of a Gaussian source W with mean zero and variance P over a Gaussian channel $Y = X + Z$ with power constraint P where $Z \sim \mathcal{N}(0, N)$.

   (a) Combine the known results of the capacity of the Gaussian channel and $R(D)$ for a Gaussian source and squared error distortion to derive the smallest distortion possible in this scenario.

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \tag{39}$$

$$R(D) = \frac{1}{2}\log\left(\frac{P}{D}\right) \tag{40}$$

Setting $R = C$ we find that we can at best achieve:

$$\frac{P}{D} = 1 + \frac{P}{N} \tag{41}$$

or

$$D = \frac{N}{\frac{N}{P} + 1} = \frac{NP}{N + P}. \tag{42}$$

(b) Show that directly transmitting the source $X = W$ and employing the unbiased receiver $\hat{W} = Y$ approaches the theoretical performance limit as $P/N \to \infty$.

With unbiased direct transmission, $\frac{P}{D} = \frac{P}{N}$ so we have $D = N$ which is the limit of $\frac{N}{\frac{N}{P}+1}$ as $P/N \to \infty$ (or $N/P \to 0$.)

(c) Now show that the proper choice of a produces a *biased* receiver $\hat{W} = aY$ that achieves the performance limit of part a for every value of P/σ^2 assuming direct transmission $X = W$ as in part b.

For biased direct transmission

$$D = E(W - aY)^2 \tag{43}$$

$$= E(W^2) - 2aE(WY) + a^2E(Y^2) \tag{44}$$

$$= P - 2aP + a^2(P + N) \tag{45}$$

Now we can simply set a derivative to zero to select the a that minimizes D.

$$\frac{dD}{da} = -2P + 2a(P + N) \tag{46}$$

and so we select $a = \frac{P}{P+N}$ which yields

$$D = P - 2\frac{P^2}{P + N} + \frac{P^2}{P + N} \tag{47}$$

$$= P - \frac{P^2}{P + N} \tag{48}$$

$$= \frac{P^2 + NP - P^2}{P + N} \tag{49}$$

$$= \frac{NP}{P + N} \tag{50}$$

Lecture 14D: Proof of Converse for $R(D)$.

8. (10 pts) *Properties of optimal rate distortion code.* The converse of the rate distortion theorem relies on the following chain of inequalities

$$\begin{aligned}
nR &\overset{(a)}{\geq} H(f_n(X^n)) && (51)\\
&\overset{(b)}{\geq} H(f_n(X^n)) - H(f_n(X^n)|X^n) && (52)\\
&= I(X^n; f_n(X^n)) && (53)\\
&\overset{(c)}{\geq} I(X^n; \hat{X}^n) && (54)\\
&= H(X^n) - H(X^n|\hat{X}^n) && (55)\\
&\overset{(d)}{=} \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) && (56)\\
&\overset{(e)}{=} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{i-1}, \ldots, X_1) && (57)\\
&\overset{(f)}{\geq} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) && (58)\\
&= \sum_{i=1}^{n} I(X_i; \hat{X}_i) && (59)\\
&\overset{(g)}{\geq} \sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i)) && (60)\\
&= n \sum_{i=1}^{n} \frac{1}{n} R(Ed(X_i, \hat{X}_i)) && (61)\\
&\overset{(h)}{\geq} nR\left(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i; \hat{X}_i)\right) && (62)\\
&\overset{(i)}{=} nR(Ed(X^n, \hat{X}^n)) && (63)\\
&\overset{(j)}{=} nR(D). && (64)
\end{aligned}$$

We will have equality in (a) if the 2^{nR} possible values of $f_n(X^n)$ are equally likely -i.e., if all the codewords were equally likely. For equality in (b), $f_n(X^n)$ must be a deterministic function of X^n. For equality in (c), each $f_n(X^n)$ needs to deterministically map to a unique $\hat{X}^n$. Equality in (d) follows from the chain rule and the fact that the X_i's are i.i.d. Equality in (e) follows from the chain rule. For equality in (f), each X_i depends on the corresponding $\hat{X}_i$ but is conditionally independent of every other $\hat{X}_j$. For equality in (g), the joint distribution of X_i and $\hat{X}_i$ is the one achieving the minimum in the definition of the rate distortion function. For equality in (h) either the rate distortion curve is a straight line or all the distortions (at each i) are equal. Equality in (j) follows from the definition of D.

Thus the optimal rate distortion code would be deterministic, and the joint distribution between the source symbol and the codeword at each instant of time would be independent and equal to the joint distribution that achieves the minimum of the rate distortion function. The distortion would be the same for each time instant.