EE5585 Data Compression February 28, 2013

Lecture 11

Instructor: Arya Mazumdar *Scribe: Nanwei Yao*

Rate Distortion Basics

When it comes to rate distortion about random variables, there are four important equations to keep in mind.

1. The entropy
$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log p(x)$$
2. Conditional entropy
$$H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y)$$
3. Joint entropy
$$H(X,Y) = \sum_{x \in \mathcal{X}} p(x) \log p(x)$$
4. Mutual Information
$$I(X;Y) = H(X) - H(X|Y)$$

We already know from previous lecture that

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

But previous proof is a little bit complex, thus we want to prove this equation again in a simpler way to make it clearer.

<u>Proof</u>:

$$\begin{aligned}
H(X,Y) &= -\sum p(x,y) \log p(x,y) \\
&= -\sum p(x,y) \log[p(x)p(y|x)] \\
&= -\sum p(x,y) \log p(x) - \sum p(x,y) \log p(y|x) \\
&= -\sum_{x} \log p(x) \sum_{y} p(x,y) + H(Y|X) \\
&= -\sum_{x} \log p(x) \sum_{x} p(x) + H(Y|X) \\
&= -\sum_{x} p(x) \log p(x) + H(Y|X) \\
&= H(X) + H(Y|X)
\end{aligned}$$

Definition:For random variables (X,Y) whose probabilities are given by p(x,y), the *conditional entropy* $H(Y|X)$ is defined by

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

$$
\begin{aligned}
&= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\log(p(y|x)
\end{aligned}
$$

One important thing to notice is that $H(X) \geq H(X|Y)$, this means that conditioning always reduces the entropy. The entropy of a pair of random variables is the summation of the entropy of one plus the conditional entropy of the other. This is based on *Chain rule*:

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

Thus, $I(X;Y) = H(X) + H(Y) - H(X,Y)$. We can understand the mutual information $I(X;Y)$ as the reduction in the uncertainty of X because of some of our knowledge of Y. By symmetry, we also have $I(X;Y) = I(Y;X)$. Thus we know that the information X provides us has the ”same amount” that Y provides us.

Rate Distortion New Materials

Recall that in *lossy coding*, we cannot compress a file without error, and we want the average distortion to be bounded above. For a binary file which is of our interest, we use *Hamming Distance* (probability of error distortion) for distortion function.

Another important case is Quantization. Suppose we are given a Gaussian random variable, we quantize it and represent it by bits. Thus we lose some information. What is the best Quantization level that we can achieve? Here is what we do.

Define a random variable $X \in \mathcal{X}$. Our source produces a n length vector and we denote it by $X^n = X_1, X_2, ..., X_n$, where the vector is i.i.d. and produced according to the distribution of random variable X and $p(x) = Pr(X = x)$. What we do is to encode the file. After the encoder, the function f_n gives us a compressed string. Then we map the point in the space to the nearest codeword and we obtain an index which is represented by $\log M_n$ bits. Finally, we use a decoder to map the index back which gives us one of the codeword $\hat{x}$ in the space.

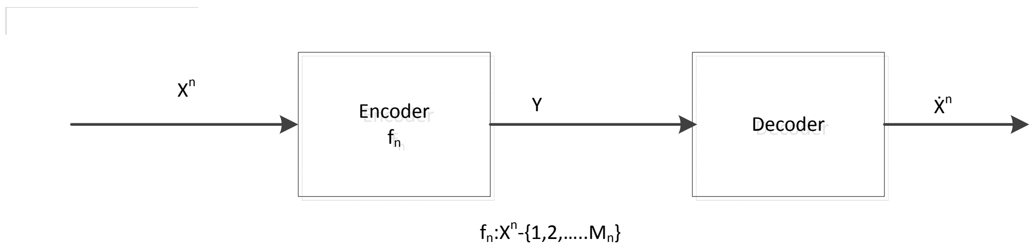
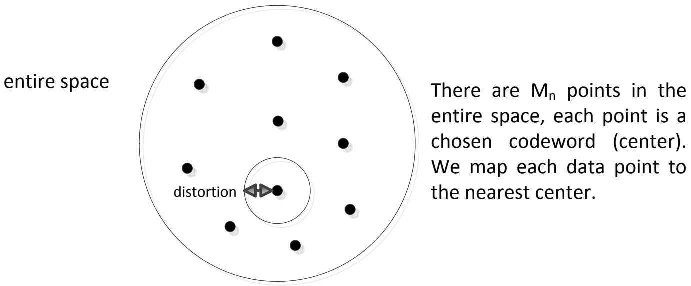


Figure 1: Rate Distortion Encoder and Decoder

Above is a figure of rate distortion encoder and decoder. Where $f_n = X^n \rightarrow \{1, 2, ...M_n\}$ is the encoder, and $\hat{X}^n$ is the actual code we choose.

Definition: The *distortion between sequences* x^n and $\hat{x^n}$ is defined by

$$d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i) \tag{1}$$



There are $\mathrm{M_n}$ points in the entire space, each point is a chosen codeword (center). We map each data point to the nearest center.

Figure 2: Codewords and Distortion

Thus, we know that the distortion of a sequence is the average distortion of the symbol-to-symbol distortion. We would like $d(x^n, \hat{x}^n) \leq D$. The compression rate we could achieve is $R = \lim_{x \to +\infty} \frac{\log M_n}{n}$ bits/symbol.The *Rate distortion* function $R(D)$ is the minimization of $\frac{\log M_n}{n}$ such that $Ed(x^n, \hat{x}^n) \leq D$ at the limit of $n \to \infty$.

Theorem(Fundamental theory of source coding): The *information rate distortion function* $R(D)$ for a source X with distortion measure $d(x, \hat{x})$ is defined as

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

From the above equation, we could see that no n involved. This is called a single letter characterization. For the subscript of the summation, we know that $\sum p(x, \hat{x})d(x, \hat{x}) = \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D$.
To prove this theorem, we want to show $R(D) \geq \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X})$ first, and the $R(D) \leq \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X})$ will be shown in the next lecture.

First of all, let's see what is *Chain Rule*. It is defined as below:

$$H(X_1, X_2, X_3, ..., X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + ... + H(X_n|X_1, X_2, ..., X_{n-1})$$

Chain rule can be easily proved by *induction*.
Besides chain rule, we also so need the fact that

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

is convex.

Two Properties of R(D)

We now show two properties of R(D) that are useful in proving the converse to the rate-distortion theorem.

1. R(D) is a decreasing function of D.
2. R(D) is a convex function in D.

For the first property, we can prove its correctness intuitively: If R(D) is an increasing function, this means that the more the distortion, the worse the compression. This is definitely what we don't want.

Now, let's prove that second property.

Proof: Choose two points $(R1, D1), (R2, D2)$ on the boundary of R(D) with distributions $P_{\hat{X}_1|X}$ and $P_{\hat{X}_2|X}$. Then, we can construct another distribution $P_{\hat{X}_\lambda|X}$ such that

$$P_{\hat{X}_\lambda|X} = \lambda P_{\hat{X}_1|X} + (1-\lambda) P_{\hat{X}_2|X}$$

where $0 \leq \lambda \leq 1$.The average distortion D_λ can be given as

$$\begin{aligned} EP_{X,\hat{X}_\lambda}[d(X,\hat{X}_\lambda)] &= \lambda EP_{X,\hat{X}_1}[d(X,\hat{X}_1)] + (1-\lambda) EP_{X,\hat{X}_2}[d(X,\hat{X}_2)] \\ &= \lambda D_1 + (1-\lambda) D_2 \end{aligned}$$

We know that $I(X;Y)$ is a convex function of $p_{\hat{X},X}()$ for a given $p_X()$. Therefore,

$$I(\hat{X}_\lambda; X) \leq \lambda I(\hat{X}_1; X) + (1-\lambda) I(\hat{X}_2; X)$$

Thus,

$$\begin{aligned} R(D_\lambda) = I(\hat{X}_\lambda) &\leq \lambda I(\hat{X}_1; X) + (1-\lambda) I(\hat{X}_2; X) \\ &= \lambda R(D_1) + (1-\lambda) R(D_2) \end{aligned}$$

Therefore, R(D) is a convex function of D.

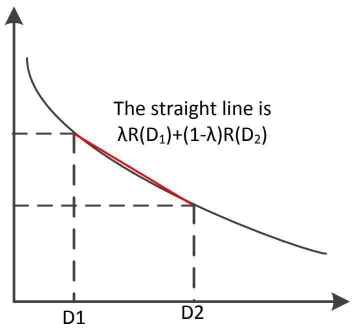


Figure 3: Function of $R(D)$ and $R(\hat{D})$

For the above proof, we need to make the argument that distortion D is a linear function of $p(\hat{x}|x)$. We know that the D is the expected distortion and it is given as $D = \sum p(x,\hat{x}) d(\hat{x}, x) = \sum p(x) p(\hat{x}|x) d(x,\hat{x})$. If we treat $p(\hat{x}|x)$ as a variable and both $p(x)$ and $d(x,\hat{x})$ as known quantities, we know that D is a linear function of $p(\hat{x}|x)$. Therefore, $R(D)$ is a convex function of $p(\hat{x}|x)$. The proof that $I(X;\hat{X})$ is a convex function of $p(\hat{x}|x)$ will not be shown here.

Converse Argument of R(D)

The converse argument of R(D) tells us that for any coding scheme whose expected distortion is at most to be D, there doesn't exist a code such that its rate is less than R(D). Now, let's prove it.
Proof

$$\begin{aligned}
\log M_n &\geq H(\hat{X}^n)\\
&\geq H(\hat{X}^n) - H(\hat{X}^n|X^n)\\
&= I(\hat{X}^n; X^n)\\
&= H(X^n) - H(X^n|\hat{X}^n)\\
&= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_1, X_2, ...X_n)\\
&\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i)\\
&= \sum_{i=1}^{n} I(X_i; \hat{X}_i)
\end{aligned}$$

Recall that $R(D) = \min\limits_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X})$, thus

$$\begin{aligned}
\log M_n &\geq \sum_{i=1}^{n} I(X_i; \hat{X}_i)\\
&\geq \sum_{i=1}^{n} R(Ed(X_i; \hat{X}_i)\\
&= n\sum_{i=1}^{n} \frac{1}{n} R(Ed(X_i; \hat{X}_i))\\
&\geq nR(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i; \hat{X}_i))\\
&= nR(\frac{1}{n}E[\sum_{i=1}^{n} d(X_i, \hat{X})])\\
&\geq nR(D)
\end{aligned}$$

We see from the proof that $R = \lim_{x\to+\infty} \frac{\log M_n}{n} \geq R(D)$, thus, our proof is finished.

Example of Rate Distortion Theorem

An interesting example to look at is a binary file. What is $R(D)$ for a given binary file?
We already know from previous lectures that $R(D) = 1 - H(D)$. But this equation is too general to use for our given file.
So given a source $\mathcal{X} = \{0, 1\}$, suppose X has a Bernoulli(p) distribution, i.e. $Pr(X = 1) = p$ and $Pr(X = 0) = 1 - p$.Then

$$\begin{aligned}
I(x; \hat{x}) &= H(X) - H(X|\hat{X})\\
&= h(p) - H(X|\hat{X})
\end{aligned}$$

$$\begin{aligned} &= h(p) - H(X \oplus \hat{X}|\hat{X}) \\ &\geq h(p) - H(X \oplus \hat{X}) \; (conditionality\, reduces\, entropy) \\ &= h(p) - H(Y) \end{aligned}$$

where $Y = X \oplus \hat{X}$. It is clear that $Pr(Y=1) = Pr(X \neq \hat{X})$, and

$$\begin{aligned} \sum p(X,\hat{X})d(X,\hat{X}) &= p(0,1) + p(1,0) \\ &= Pr(X \neq \hat{X}) \\ &= Pr(Y=1) \leq D \end{aligned}$$

Recall from previous lectures, for any $D \leq \frac{1}{2}$, the binary entropy function $h(D)$ is increasing. Thus $h(p) - H(Y) \geq h(p) - h(D)$ for $D \leq \frac{1}{2}$. We have showed that $R(D) \geq h(p) - h(D)$, and we will show $R(D) = h(p) - h(D)$. When $p = \frac{1}{2}$, $R(D) = 1 - h(D)$.

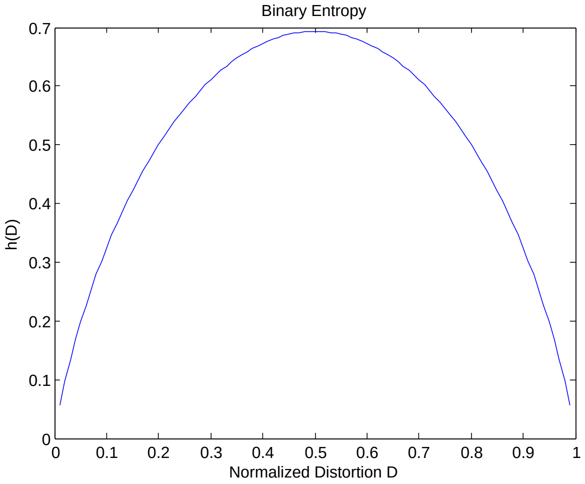


Figure 4: Binary Entropy

Up to this point, we want to show $H(X \oplus \hat{X}|X)$ has the same value as $h(D)$.

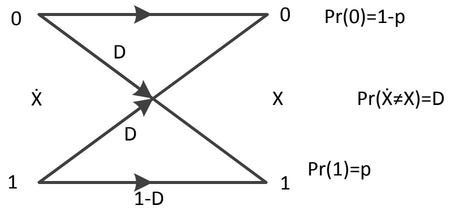


Figure 5: Binary Encoding Demonstration

We know that $Pr(\hat{X} \neq X) = D$. Assume $Pr(\hat{X} = 0) = 1 - r$ and $Pr(\hat{X} = 1) = r$, so $Pr(X = 0) = (1-r)(1-D) + rD = 1 - p$. Solve this equation for r, we obtain $r = \frac{p-D}{1-2D}$ for $D \leq \frac{1}{2}$.

In this case, $I(X;\hat{X}) = H(X) - H(Y) = h(p) - h(D)$, thus we proved both sides of the main theory for binary source.

Midterm Solutions

1. $1 - h(D)$ is the optimal rate of compression that is achievable.$1 - h(D) = 1 - h(1/20)$, where $h(x) = -x \log x - (1-x)\log(1-x)$.

2. $\frac{\pi(\frac{1}{4})^2}{(\frac{1}{2})^2} = \frac{1}{4}\pi$

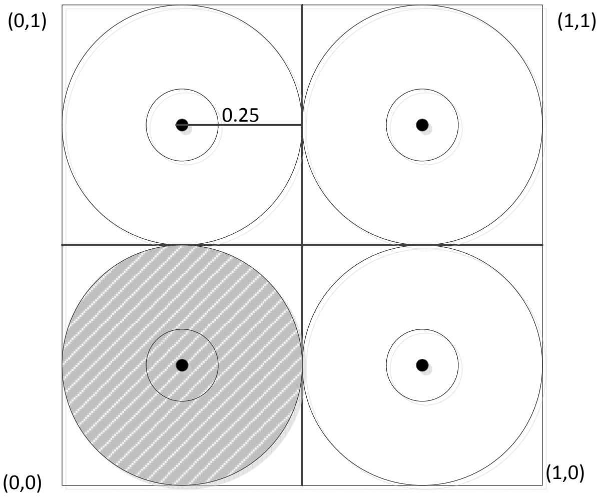


Figure 6: Distortion Demonstration

3. $p_1, p_2, p_3, ...$

$$\begin{aligned}\sum_{i=n+1}^{\infty} p_i &= \sum_{i=n+1}^{\infty} \frac{9}{10}(\frac{1}{10})^{i-1} \\ &= \frac{9}{10}(\frac{1}{10})^n + \frac{9}{10}(\frac{1}{10})^{n+1} + ... \\ &= \frac{9}{10}(\frac{1}{10})^n[1 + \frac{1}{10} + (\frac{1}{10})^2 + ...] \\ &= \frac{9}{10}\frac{1}{\frac{9}{10}} \\ &= \frac{1}{10}(\frac{1}{10})^{n-1}\end{aligned}$$

$p_n = \frac{9}{10}(\frac{1}{10})^{n-1}$

Thus $p_n > \sum_{i=n+1}^{\infty} p_i$.

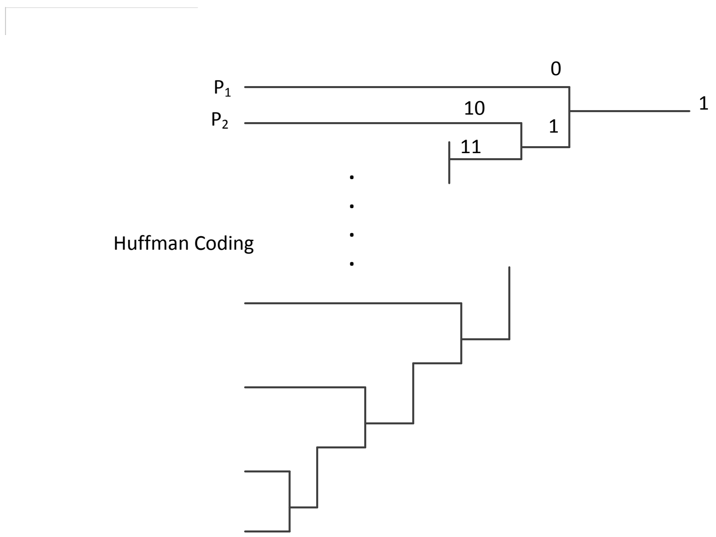


Figure 7: Huffman Coding

So length of the series is 1,2,3,4,.......

4. $\frac{n(n+1)}{2} = 5050 \rightarrow n = 100$
   Length is $100X \lceil \log(100) + 1 \rceil = 800$
   $R = \frac{m(\log(m)+1)}{m(m+1)/2} = \frac{2(\log(m)+1)}{(m+1)}$

5. Covered in the previous course.