ECE4304 C-term 2007: Lecture 26 Supplemental Slides

D. Richard Brown III
Worcester Polytechnic Institute, Department of Electrical and Computer Engineering

February 20, 2007

Mutual Information: Part I

X is selected from alphabet $\mathcal{A}_X = \{\theta_1, \ldots, \theta_M\}$ with probabilities $P(X = \theta_m) = p_m$. X has entropy

$$H(X) := -\sum_{m=1}^{M} p_m \log_2(p_m).$$

Y is selected from alphabet $\mathcal{A}_Y = \{\phi_1, \ldots, \phi_K\}$ with probabilities $P(Y = \phi_k) = q_k$. Y has entropy

$$H(Y) := -\sum_{k=1}^{K} q_k \log_2(q_k).$$

The DMC is completely specified by the transition probabilities

$$p_{k|m} = P(Y = \phi_k \,|\, X = \theta_m)$$

Mutual Information: Part II

We want to know how much information knowing Y reveals about X (on average). This is the intuition behind the concept of mutual information $I(X;Y)$.

- If $I(X;Y)$ is high, then Y reveals lots of information about X, on average.
- If $I(X;Y)$ is low, then Y doesn't reveal much information about X, on average.

It is a bit like correlation.

Knowing something about Y tells us something about X and vice-versa. Their mutual information is defined as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Note that the units of "information" here are bits per symbol (as usual).

What is $H(X|Y)$? $H(X|Y)$ is the "conditional entropy" of X given that you know Y. Intuitively, this is the amount of average information (or uncertainty) in X left when you observe Y. Note that $H(X|Y) \leq H(X)$, hence $I(X;Y) \geq 0$.

Conditional Entropy

Suppose you observe $Y = \phi_k$. The amount of average information left in X when you observe $Y = \phi_k$ can be written as

$$H(X|Y = \phi_k) := -\sum_{m=1}^{M} p_{m|k} \log_2(p_{m|k}).$$

where $p_{m|k} = P(X = \theta_m|Y = \phi_k)$. But we want to know the amount of average information left in X, averaged over all possible observations (not just for this observation). Averaging this last expression over all of the possible observations, we can write

$$\begin{aligned} H(X|Y) &= \mathsf{E}[H(X|Y = \phi_k)] \\ &= \sum_{k=1}^{K} q_k H(X|Y = \phi_k) \\ &= -\sum_{k=1}^{K} \sum_{m=1}^{M} q_k p_{m|k} \log_2(p_{m|k}) \\ &= -\sum_{k=1}^{K} \sum_{m=1}^{M} p_{m,k} \log_2(p_{m|k}) \end{aligned}$$

where $p_{m,k} = P(X = \theta_m, Y = \phi_k)$.

Intuition and Summary

- If $Y = X$ always (the channel doesn't make errors), then $H(X|Y) = 0$ and $I(X;Y) = H(X)$.
- If Y is completely independent of X (symbols coming out of the channel are totally corrupted), then $H(X|Y) = H(X)$ and $I(X;Y) = 0$.

Put it all together...

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= -\sum_{m=1}^{M} p_m \log_2(p_m) + \sum_{k=1}^{K}\sum_{m=1}^{M} p_{m,k} \log_2(p_{m|k}) \\ &= H(Y) - H(Y|X) \\ &= -\sum_{k=1}^{K} q_k \log_2(q_k) + \sum_{m=1}^{M}\sum_{k=1}^{K} p_{m,k} \log_2(p_{k|m}) \end{aligned}$$

where $p_{m|k} = P(X = \theta_m | Y = \phi_k)$; $p_{m,k} = P(X = \theta_m, Y = \phi_k)$; $p_{k|m} = P(Y = \phi_k | X = \theta_m)$.

Note: Both expressions give the same answer but sometimes one is easier to compute than the other.

Quick Bayes Rule Refresher

$$\begin{aligned} P(A,B) &= P(B)P(A|B) = P(A)P(B|A) \\ &= P(A)P(B) \text{ only if events A and B are independent} \end{aligned}$$

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \text{ as long as } P(B) > 0$$

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} \text{ as long as } P(A) > 0$$

Mutual Information Example: Part I

Suppose we have a Binary Symmetric Channel (BSC) with crossover probability ε. The input/output alphabets are $\mathcal{A}_X = \mathcal{A}_Y = \{0, 1\}$. The input probabilities are $P(X = 0) = p$ and $P(X = 1) = 1 - p$.

Let's compute the mutual information this way...

$$I(X;Y) = H(Y) - H(Y|X) = -\sum_{k=1}^{K} q_k \log_2(q_k) + \sum_{m=1}^{M} \sum_{k=1}^{K} p_{m,k} \log_2(p_{k|m})$$

The first thing we need is the output probabilities $P(Y = 0)$ and $P(Y = 1)$.

$$q_1 = P(Y = 0) \quad =$$

$$q_2 = P(Y = 1) \quad =$$

Mutual Information Example: Part II

The second thing we need is the transition probabilities $P(Y=0|X=0)$, $P(Y=1|X=0)$, $P(Y=1|X=1)$, $P(Y=0|X=1)$

$$
\begin{aligned}
p_{1|1} = P(Y=0|X=0) &= \\
p_{2|1} = P(Y=1|X=0) &= \\
p_{1|2} = P(Y=0|X=1) &= \\
p_{2|2} = P(Y=1|X=1) &=
\end{aligned}
$$

The last thing we need is the joint probabilities $P(Y=0, X=0)$, $P(Y=1, X=0)$, $P(Y=1, X=1)$, $P(Y=0, X=1)$

$$
\begin{aligned}
p_{1,1} = P(Y=0, X=0) &= \\
p_{2,1} = P(Y=1, X=0) &= \\
p_{1,2} = P(Y=0, X=1) &= \\
p_{2,2} = P(Y=1, X=1) &=
\end{aligned}
$$

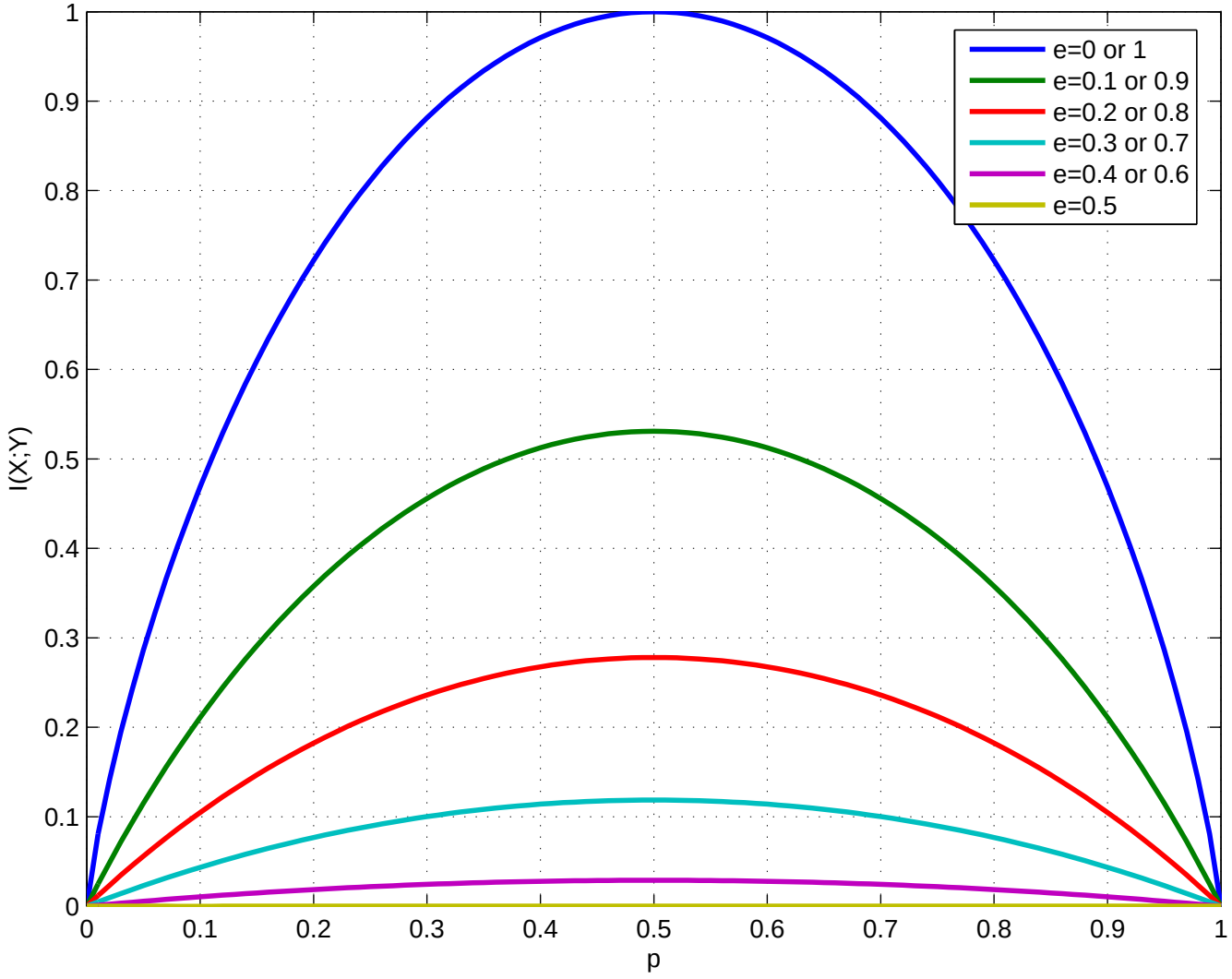
Mutual Information Example: Part III

Put it all together...

$$
\begin{aligned}
I(X;Y) &= -\sum_{k=1}^{K} q_k \log_2(q_k) + \sum_{m=1}^{M} \sum_{k=1}^{K} p_{m,k} \log_2(p_{k|m}) \\
I(X;Y) &=
\end{aligned}
$$

Note that $I(X;Y)$ is a function of both the crossover probability ε and the input probability p.

Mutual Information Example: Part IV



Channel Capacity and a Big Theorem

Definition: The "capacity" of a DMC is

$$C = \max_{\{p_m\}} I(X;Y)$$

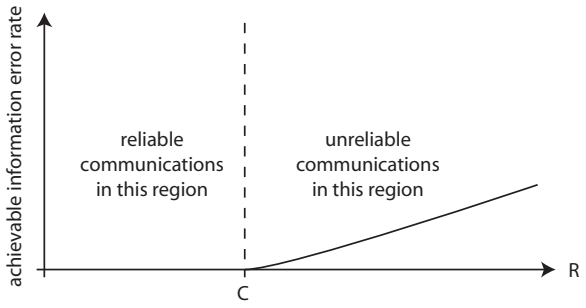
where $\{p_m\}$ are the input symbol probabilities.

Note that computing C requires the maximization of $I(X;Y)$ over all possible input probabilities. In our BSC example, $I(X;Y)$ is always maximized when $p = 0$, irrespective of the crossover probability ε.

Theorem 1 (Shannon's Noisy Channel Coding Theorem)**.** *Let R be the information rate (bits of information per symbol, on average) of the source X and let C be the capacity of a DMC. If $R \leq C$ then it is possible to achieve an arbitrarily small "information error rate" through this channel. If $R > C$, then there is a non-zero lower bound to the achievable "information error rate" through this channel.*

Remarks

1. You can think of C as the "speed limit" of the channel. If you keep the information rate below C bits/symbol, it is possible (if you are clever) to communicate with arbitrarily small loss of information.



2. Note that information error rate (IER) is not equal to SER or BER. We can't make SER$\rightarrow 0$ without making $\mathcal{E}_{b-av} \rightarrow \infty$. But Shannon's Noisy Channel Coding Theorem says that we can make IER$\rightarrow 0$ if we stay below the speed limit C, even if SER> 0 .

3. How do we do this? By cleverly encoding the information prior to sending it through the channel. The channel coder adds smart redundancy to the information to make the communication of information reliable, even when the communication of symbols isn't.