

# ECE253/CSE208 Introduction to Information Theory

## Lecture 10: Channel Capacity

Dr. Yu Zhang

ECE Department

University of California, Santa Cruz

- Chap 7 of *Elements of Information Theory (2nd Edition)* by Thomas Cover & Joy Thomas

## Discrete Memoryless Channels (DMC)

We are seeking the answers to the following two core questions in information theory:

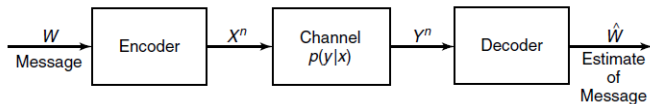
- The fundamental limit of data compression is  $H(\mathcal{X})$ : Chap 5
- The fundamental limit of data transmission: Chap 7

There is a duality between data compression and transmission. That is, we want to reduce redundancy as much as possible for data compression while adding controlled redundancy to combat errors for data transmission.

### Definition

A **discrete channel** to be a system of an input alphabet  $\mathcal{X}$  and output alphabet  $\mathcal{Y}$  and a probability transition matrix  $p(y|x)$  that expresses the probability of observing the output symbol  $y$  given that we send the symbol  $x$ . The channel is **memoryless** if the probability distribution of the output depends only on the input at the time and is conditionally independent of previous channel inputs or outputs.

## Definitions of Channel Capacity



### Definition

"Information" channel capacity of a DMC is defined as  $C = \max_{p(x)} I(X; Y)$ .

### Definition

"Operational" channel capacity is the number of bits we can transmit reliably (with an arbitrarily low probability of error) per channel usage.

$C = \log(\text{number of identifiable inputs by passing through the channel with low error})$ .

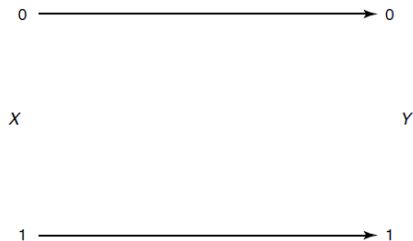
### Shannon's second theorem (Noisy-channel coding theorem):

"Information" channel capacity = "operational" channel capacity

# Noiseless Binary Channel

## Example (Noiseless binary channel (NBC))

Operational-wise, we can transmit one bit without any error per channel use. Hence  $C = 1$  bit. Meanwhile, we can calculate the *information* capacity as

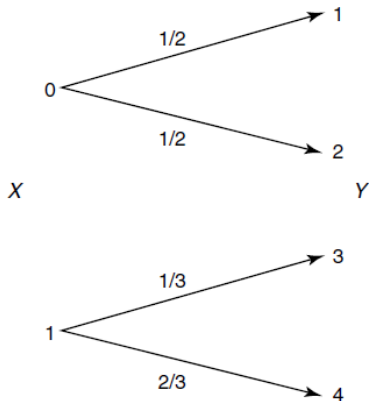


$$\begin{aligned} C &= \max_{p(x)} I(X; Y) \\ &= \max_{p(x)} [H(X) - H(X|Y)] \\ &= \max_{p(x)} H(X) \\ &= 1 \text{ bit} \quad \text{if } p(x) = (1/2, 1/2) \end{aligned}$$

## Noisy Channel with Non-overlapping Outputs

### Example (Noisy channel with non-overlapping outputs)

Similar to the previous example, we can transmit one bit without any error per channel use. Hence  $C = 1$  bit.



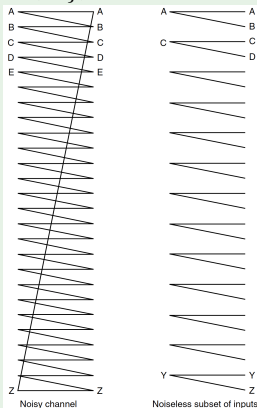
$$\begin{aligned} C &= \max_{p(x)} I(X; Y) \\ &= \max_{p(x)} [H(X) - H(X|Y)] \\ &= \max_{p(x)} H(X) \\ &= 1 \text{ bit} \quad \text{if } p(x) = (1/2, 1/2) \end{aligned}$$

[Exercise: find the solution by an alternative way  $C = \max_{p(x)} (H(Y) - H(Y|X)).$ ]

# Noisy Typewriter

## Example (Noisy Typewriter)

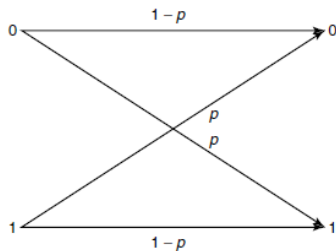
The channel input is received correctly with probability 0.5, or is transformed into the next letter with probability 0.5. Note that if we transmit every other symbols  $\{A, C, E, \dots, Z\}$ , there will be no error in the communication. Therefore,  $C = \log(13)$  bits.



$$\begin{aligned} C &= \max_{p(x)} I(X; Y) \\ &= \max_{p(x)} [H(Y) - H(Y|X)] \\ &= \max_{p(x)} \left[ H(Y) - \sum p(x) H(Y|X = x) \right] \\ &= \max_{p(x)} [H(Y) - 1] \\ &= \log(26) - 1 \\ &= \log(13) \text{ bits} \end{aligned}$$

# Binary Symmetric Channel

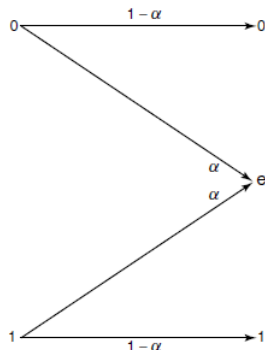
## Example (Binary Symmetric Channel (BSC))



$$\begin{aligned} C &= \max_{p(x)} I(X; Y) \\ &= \max_{p(x)} [H(Y) - H(Y|X)] \\ &= \max_{p(x)} [H(Y) - H(p)] \quad H(Y|X = i) = H(p), i = 1, 2 \\ &= 1 - H(p) \text{ bits} \quad \text{if } p(x) = (1/2, 1/2) \Rightarrow p(y) = (1/2, 1/2) \end{aligned}$$

# Binary Erasure Channel

## Example (Binary Erasure Channel (BEC))



$$\begin{aligned} C &= \max_{p(x)} I(X; Y) \\ &= \max_{p(x)} [H(Y) - H(Y|X)] \\ &= \max_{p(x)} [H(Y) - H(\alpha)] \\ &= \max_{\pi} (1 - \alpha)H(\pi) \\ &= (1 - \alpha) \text{ bits} \quad \text{when } \pi = 1/2 \end{aligned}$$

Define  $p(X = 1) := \pi$ , we have

$$p(Y = 0) = (1 - \pi)(1 - \alpha), p(Y = 1) = \pi(1 - \alpha), p(Y = e) = (1 - \pi)\alpha + \pi\alpha = \alpha \implies$$

$$H(Y) = H((1 - \pi)(1 - \alpha), \pi(1 - \alpha), \alpha) = H(\alpha) + (1 - \alpha)H(\pi).$$



## Symmetric Channel

### Example (Symmetric Channel (SC))

Given the channel, i.e., the doubly stochastic matrix collects all conditional probabilities:

$$p(y|x) = \begin{pmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{pmatrix}.$$

All rows and columns are permutations of each other, the channel  $p(y|x)$  is symmetric.

$$\begin{aligned} C &= \max_{p(x)} H(Y) - H(Y|X) \\ &= \max_{p(x)} H(Y) - H(\mathbf{r}) \\ &= \log |\mathcal{Y}| - H(\mathbf{r}) \quad [\mathbf{r} \text{ is any row of the matrix}] \end{aligned}$$

$$\begin{aligned} p(y) &= \sum_x p(y|x)p(x) = \frac{1}{|\mathcal{X}|} \sum_x p(y|x) \\ &= \frac{c}{|\mathcal{X}|} = \frac{1}{|\mathcal{Y}|} \quad [\text{uniform input} \rightarrow \text{uniform output}] \end{aligned}$$

## Weakly Symmetric Channel

### Example (Weakly Symmetric Channel (WSC))

If all rows are permutations of each other, while the sums of all columns  $\sum_x p(y|x)$  are equal (not necessarily equal to 1), then the channel is called *weakly symmetric*; e.g.,

$$p(y|x) = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{pmatrix}.$$

### Theorem

*For a weakly symmetric channel, we have  $C = \log |\mathcal{Y}| - H(\mathbf{r})$ . This is achieved by a uniform distribution over  $\mathcal{X}$ .*

# Computing Capacity of General Channels

Channel capacity has the following properties:

- Naive lower and upper bounds:  $0 \leq C \leq \min(\log |\mathcal{X}|, \log |\mathcal{Y}|)$
- Given the channel ( $p(y|x)$  is fixed),  $I(X; Y)$  is a continuous concave function in  $p(x)$ ; see Theorem 2.7.4 in the book.

$$C = \max_{p(x) \in \mathcal{P}} I(X; Y),$$

where  $\mathcal{P} = \{p(x) \mid 0 \leq p(x) \leq 1, \sum_x p(x) = 1\}$  denotes the probabilistic simplex.

Generally, no closed-form solution for the above nonlinear optimization problem. We can leverage various algorithms to numerically evaluate  $C$ , e.g.,

1. Gradient search based algorithms
2. KKT conditions
3. Other iterative algorithms

# Projected Gradient Ascent

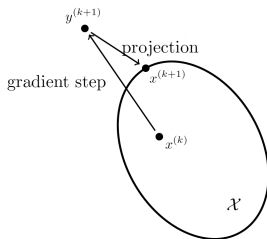
$$C = \max_{p(x) \in \mathcal{P}} I(X; Y) = \max_{p(x) \in \mathcal{P}} \sum_{x, y} p(x) p(y|x) \log \frac{p(x|y)}{p(x)} \quad (1)$$

$$= \max_{p(x) \in \mathcal{P}} \sum_{x, y} p(x) p(y|x) \log \frac{\frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}}{p(x)} \quad (2)$$

Projected gradient ascent:

$$p^{\ell+1}(x) = \text{Proj}_{\mathcal{P}} \left[ p^{\ell}(x) + \alpha^{\ell} \nabla f(p^{\ell}(x)) \right],$$

where  $\ell$  is the iteration index;  $\alpha^{\ell}$  is the step size; and  $\nabla f(\cdot)$  is the gradient of the objective function.



# Blahut-Arimoto Algorithm<sup>1</sup>

Blahut-Arimoto algorithm: Treat  $p(x)$  and  $p(x|y)$  as independent variables (blocks).

$$C = \max_{p(x)} I(X; Y) = \max_{p(x)} \sum_{x,y} p(x)p(y|x) \log \frac{p(x|y)}{p(x)} \quad (3)$$

$$= \max_{p(x)} \max_{q(x|y)} \sum_{x,y} p(x)p(y|x) \log \frac{q(x|y)}{p(x)} \quad (4)$$

For fixed  $p(x)$ , update  $q(x|y)$  as

$$q(x|y) = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}.$$

For fixed  $q(x|y)$ , update  $p(x)$  as

$$p(x) = \frac{\prod_y q(x|y)^{p(y|x)}}{\sum_{x'} \prod_y q(x'|y)^{p(y|x')}}.$$

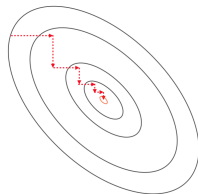


Figure: For block coordinate ascent, we only update one block at each iteration:  
 $p^1(x) \rightarrow q^1(x|y) \rightarrow p^2(x) \rightarrow q^2(x|y) \rightarrow \dots$

<sup>1</sup>S. Arimoto (1972), "An algorithm for computing the capacity of arbitrary discrete memoryless channels".  
R. Blahut (1972), "Computation of channel capacity and rate-distortion functions".

## Two Lemmas<sup>2</sup>

We will show two lemmas to validate the Blahut-Arimoto algorithm. Let

$$J(p) \triangleq \sum_{x,y} p(x)p(y|x) \log \frac{p(x|y)}{p(x)} \quad (5)$$

$$J(p, q) \triangleq \sum_{x,y} p(x)p(y|x) \log \frac{q(x|y)}{p(x)} \quad (6)$$

### Lemma

For any fixed  $p$ ,  $J(p, q) \leq J(p)$ , with equality iff  $q(x|y) = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')} = p(x|y)$ .

### Lemma

For any fixed  $q(x|y)$ ,  $J(p, q) \leq \log(\sum_x r(x))$ , where  $r(x) = \exp\left(\sum_y p(y|x) \log q(x|y)\right)$  with equality iff  $p(x) = \frac{r(x)}{\sum_{x'} r(x')}$ .

<sup>2</sup>[http://ecse.rpi.edu/~pearlman/lec\\_notes/arimoto\\_2.pdf](http://ecse.rpi.edu/~pearlman/lec_notes/arimoto_2.pdf)

## Proof of Two Lemmas

Recall that

$$J(p) \triangleq \sum_{x,y} p(x)p(y|x) \log \frac{p(x|y)}{p(x)} \quad \text{and} \quad J(p, q) \triangleq \sum_{x,y} p(x)p(y|x) \log \frac{q(x|y)}{p(x)}.$$

Proof of Lemma 1:

$$J(p) - J(p, q) = \sum_{x,y} p(x)p(y|x) \log \left[ \frac{p(x|y)}{q(x|y)} \right] = D(p(x|y) || q(x|y)) \geq 0$$

Proof of Lemma 2:

$$\begin{aligned} J(p, q) &= \sum_{x,y} p(x)p(y|x) \log \frac{q(x|y)}{p(x)} = \sum_{x,y} p(x)p(y|x) \log \frac{1}{p(x)} + \sum_{x,y} p(x)p(y|x) \log q(x|y) \\ &= \sum_x p(x) \log \left( \frac{\exp \sum_y p(y|x) \log q(x|y)}{p(x)} \right) = \sum_x p(x) \log \left( \frac{r(x)}{p(x)} \right) \\ &= \sum_x p(x) \log \left( \frac{\tilde{r}(x)}{p(x)} \times \sum_x r(x) \right), \quad / \tilde{r}(x) \triangleq \frac{r(x)}{\sum_x r(x)} / \\ &= \sum_x p(x) \log \left( \sum_x r(x) \right) - D(p || \tilde{r}) \leq \log \left( \sum_x r(x) \right) \end{aligned}$$

with equality iff  $p(x) = \tilde{r}(x)$ .

*Thank You!*

Email: <zhangy@ucsc.edu>

Homepage: <https://people.ucsc.edu/~yzhan419/>