Digital Communication Systems
ECS 452

Asst. Prof. Dr. Prapun Suksompong

prapun@siit.tu.ac.th

4. Mutual Information and Channel Capacity

Office Hours:

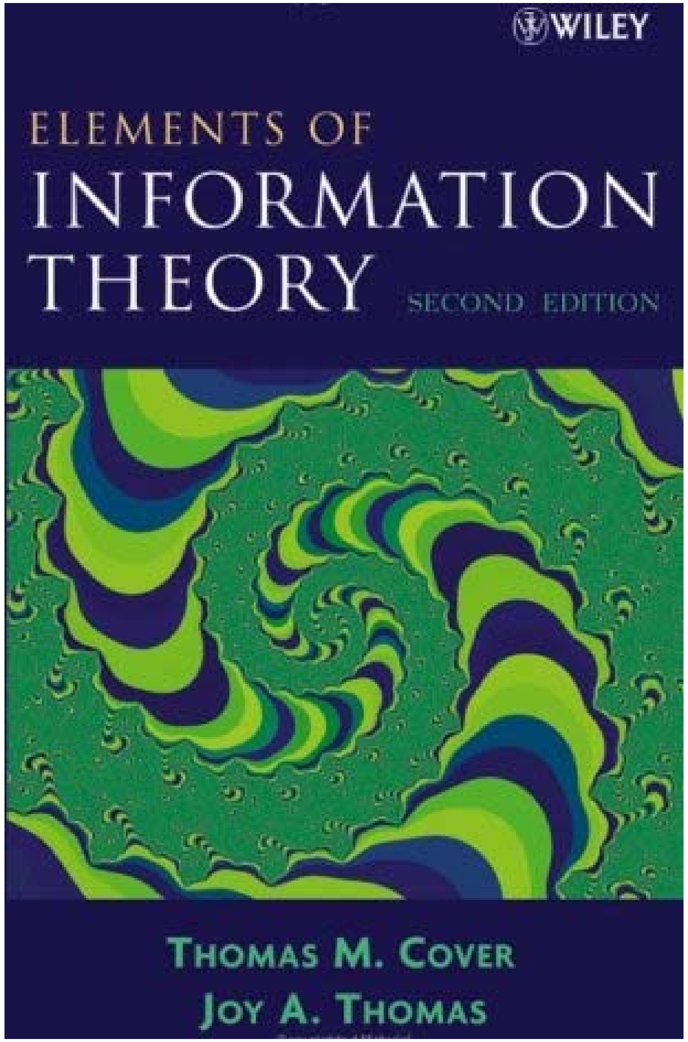
Check Google Calendar on the course website.

Dr.Prapun's Office:

6th floor of Sirindhralai building, BKD

Reference for this chapter

- Elements of Information Theory
- By Thomas M. **Cover** and Joy A. **Thomas**
- 2nd Edition (Wiley)
- Chapters 2, 7, and 8
- 1st Edition available at SIIT library: Q360 C68 1991



Recall: Entropy

4.29. Reminder:

(a) Some definitions involving entropy

(i) Binary entropy function: $h(p) = -p \log_2 p - (1-p) \log_2 (1-p)$

(ii) $H(X) = -\sum_x p(x) \log_2 p(x)$

(iii) $H(\underline{\mathbf{p}}) = -\sum_i p_i \log_2 (p_i)$

(b) A key entropy property that will be used frequently in this section is that for any random variable X,

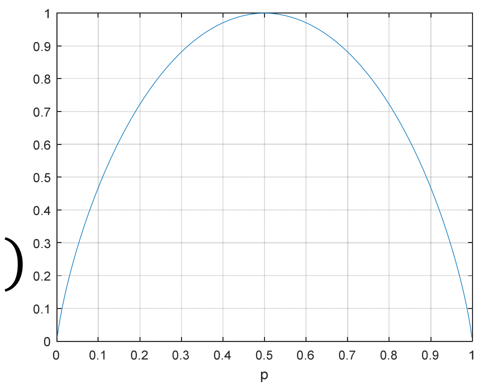
$$H(X) \le \log_2 |\mathcal{X}| \text{ with equality iff } X \text{ is uniform.}$$

[Page 72]

Recall: Entropy

- **Entropy** measures the amount of uncertainty (randomness) in a RV.
- Three formulas for calculating entropy:
  - [Defn 2.41] Given a pmf $p_X(x)$ of a RV X,
    - $\boxed{\boldsymbol{H(X)} \equiv -\sum_x p_X(x) \log_2 p_X(x).}$ Set $0\log_2 0 = 0$.
  - [2.44] Given a probability vector $\underline{\mathbf{p}}$,
    - $\boldsymbol{H(\underline{\mathbf{p}})} \equiv -\sum_i p_i \log_2 p_i.$
  - [Defn 2.47] Given a number $p \in [0,1]$,
    - $\boldsymbol{H(p)} \equiv \boldsymbol{h_b(p)} = -p\log_2 p - (1-p)\log_2(1-p)$

binary entropy function



- [2.56] Operational meaning: Entropy of a random variable is the average length of its shortest description.

Recall: Entropy

- Important Bounds

$$\underset{\text{deterministic}}{0} \leq H(X) \leq \underset{\text{uniform}}{\log_2 |S_X|}$$

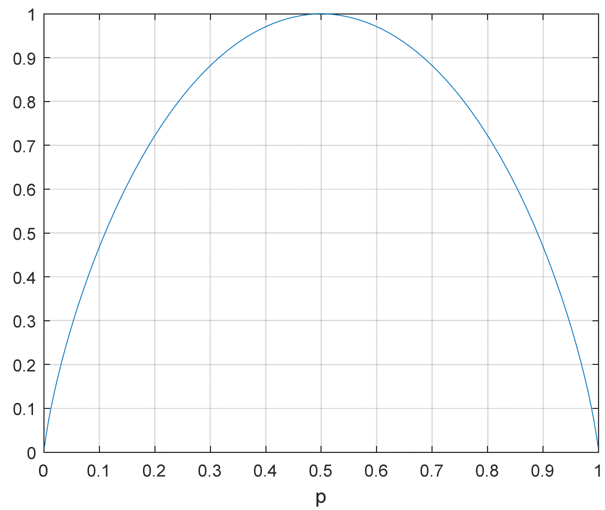
- The entropy of a uniform (discrete) random variable:

$$H(X) = \log_2 |S_X|$$

- The entropy of a Bernoulli random variable:

$$\boldsymbol{H(p) \equiv h_b(p)} = -p\log_2 p - (1-p)\log_2(1-p)$$

  - **binary entropy function**



Digital Communication Systems ECS 452

Asst. Prof. Dr. Prapun Suksompong

prapun@siit.tu.ac.th

Information-Theoretic Quantities

ECS315 vs. ECS452

ECS315	**ECS452**
We talked about **randomness** but we did not have a quantity that formally measures the amount of randomness. Back then, we studied **variance** and **standard deviation**.	We study **entropy**.
We talked about **independence** but we did not have a quantity that completely measures the amount of **dependency**. Back then, we studied **correlation**, **covariance**, and **uncorrelated** random variables.	We study **mutual information**.

Recall: ECS315 2019/1

11.46. To quantify the amount of ***dependence*** between two random variables, we may calculate their ***mutual information***. This quantity is crucial in the study of digital communications and information theory. However, in introductory probability class (and introductory communication class), it is traditionally omitted.

11.4 Linear Dependence

Definition 11.47. Given two random variables X and Y, we may calculate the following quantities:

(a) **Correlation**: $\mathbb{E}[XY]$.

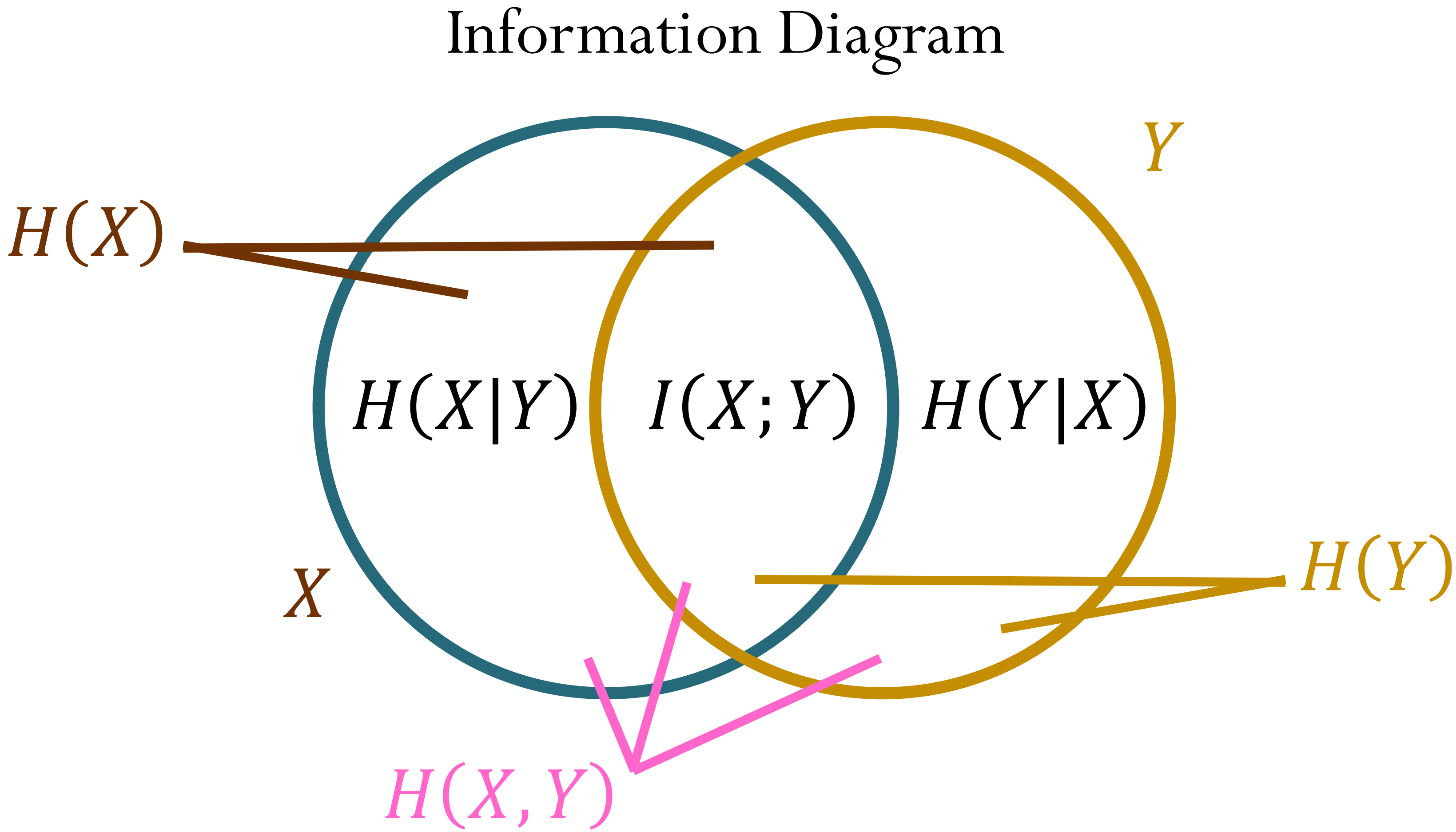
(b) **Covariance**: $\text{Cov}[X,Y] = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$.

Definition 11.51. X and Y are said to be ***uncorrelated*** if and only if $\text{Cov}[X,Y] = 0$.

11.53. Independence implies uncorrelatedness; that is if $X \perp\!\!\!\perp Y$, then $\text{Cov}[X,Y] = 0$.

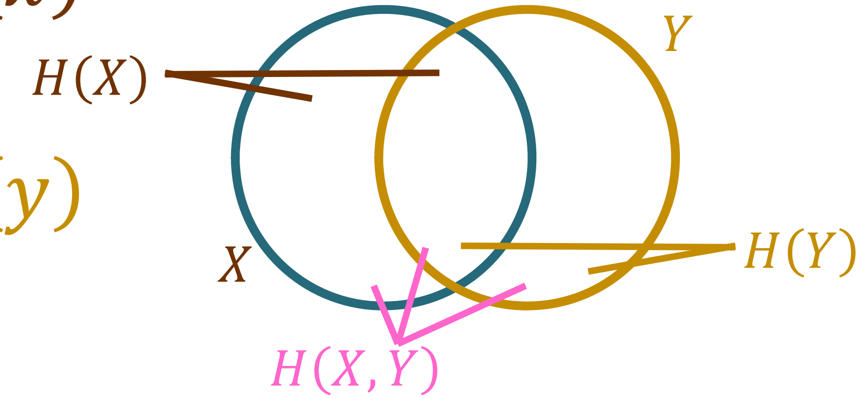
The converse is not true. Uncorrelatedness does not imply independence. See Example 11.54.

Information-Theoretic Quantities



Entropy and Joint Entropy

- **Entropy**
  - $H(X) = -\sum_x p(x) \log_2 p(x)$
    - Amount of randomness in X
  - $H(Y) = -\sum_y q(y) \log_2 q(y)$
    - Amount of randomness in Y



- **Joint Entropy**
  - $H(X,Y) = -\sum_{(x,y)} p(x,y) \log_2 p(x,y)$
    - Amount of randomness in (X, Y) pair
  - In general, $H(X,Y) \neq H(X) + H(Y)$
    - There might be some shared randomness between X and Y.

Conditional Entropies

Amount of randomness in Y

$$H(Y) \equiv -\sum_{y \in \mathcal{Y}} q(y) \log_2 \overbrace{q(y)}^{P[Y=y]} \equiv H(\underline{\mathbf{q}})$$

Amount of randomness still remained in Y when we know that $X = x$.

$$H(Y|\overbrace{X=x}^{\text{given a particular value } x}) \equiv H(Y|x) \equiv -\sum_{y \in \mathcal{Y}} Q(y|x) \log_2 \overbrace{Q(y|x)}^{P[Y=y|X=x]}$$

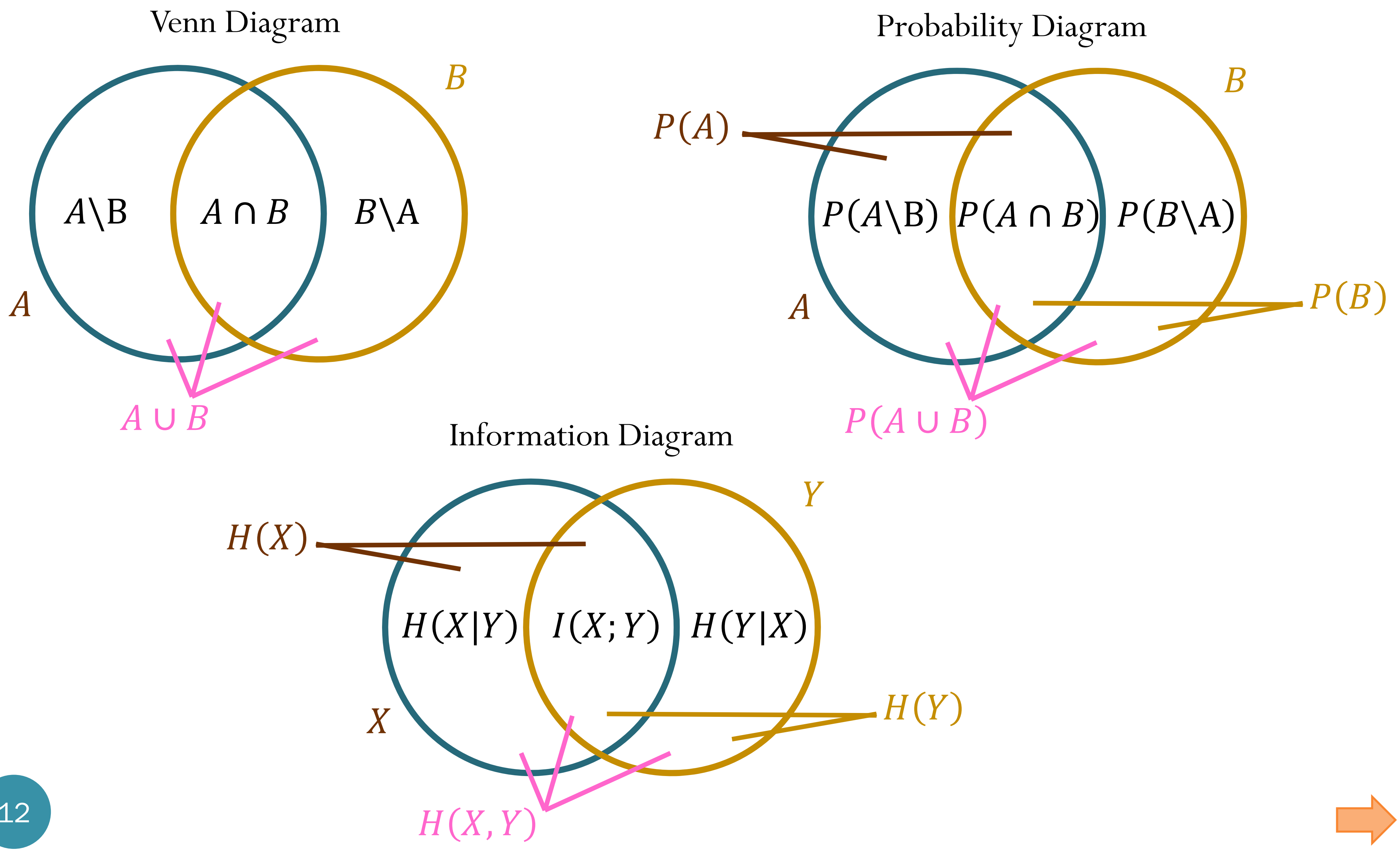
Apply the entropy calculation to a row from the $\mathbf{Q}$ matrix

The **average** amount of randomness still remained in Y when we know X

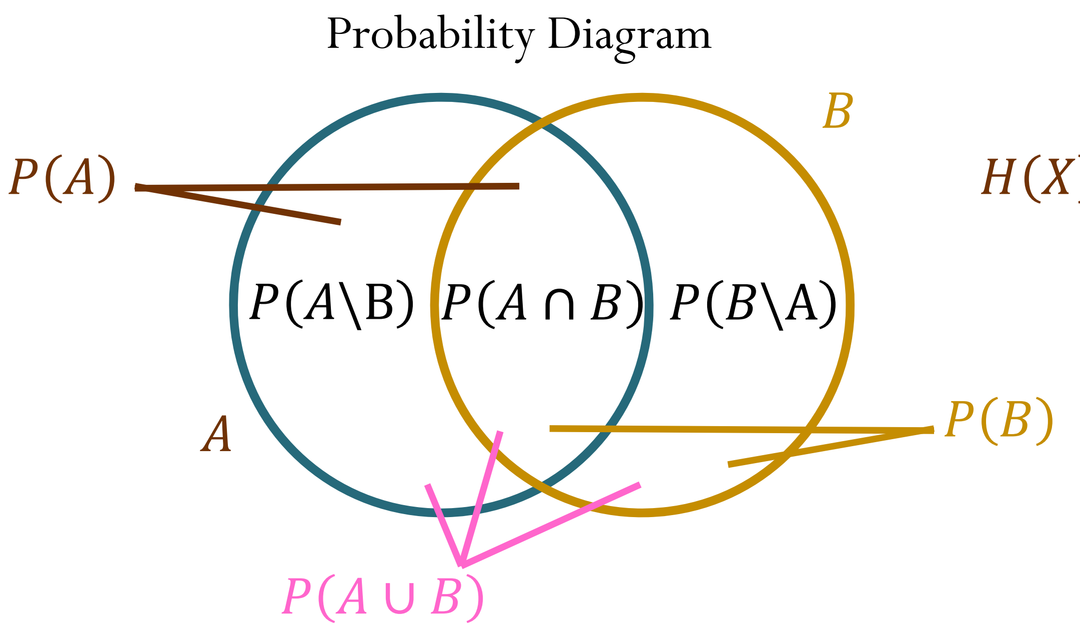
$$H(Y|X) \equiv \overbrace{\sum_{x \in \mathcal{X}} p(x) H(Y|x)}^{\textbf{average} \text{ of } H(Y|x)}$$

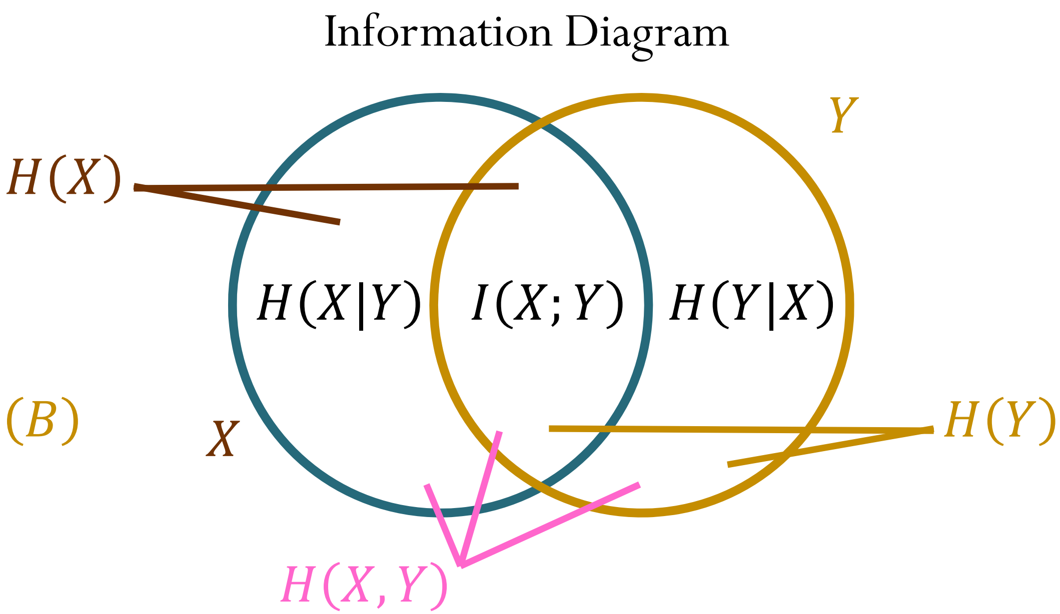
$$= H(X,Y) - H(X)$$

Diagrams [Figure 16]

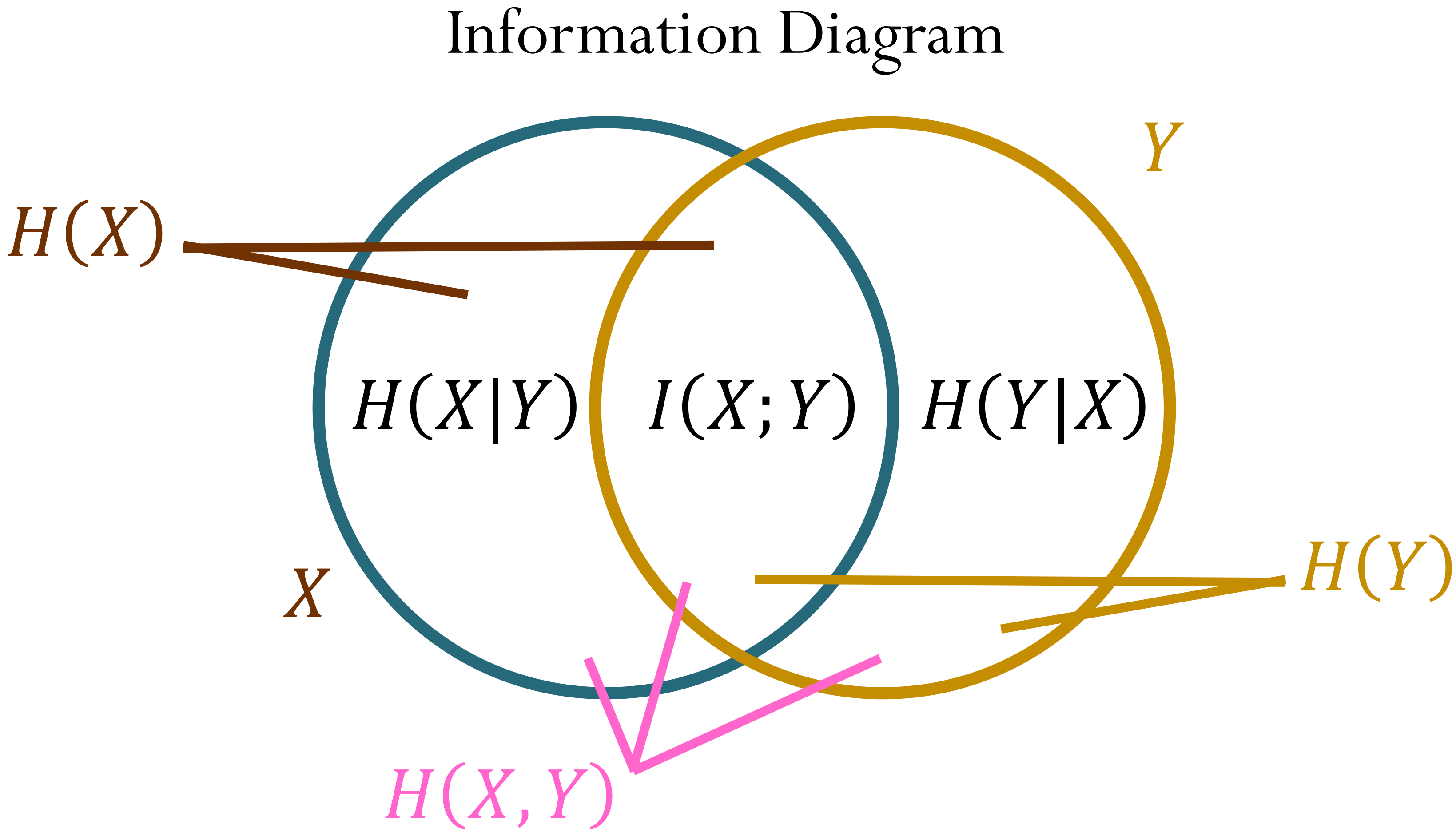


Diagrams [Figure 16]





Information-Theoretic Quantities

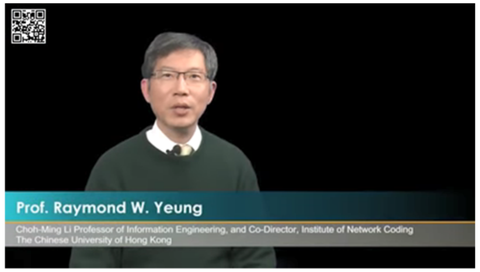


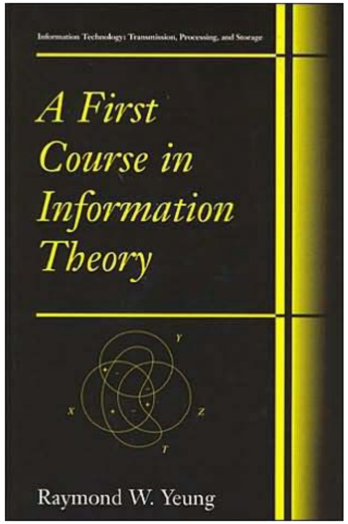
Toby Berger with Berger plaque

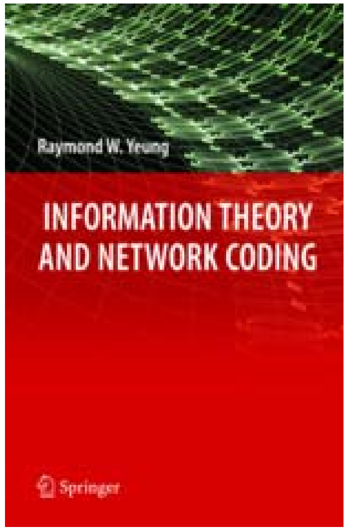
เรย์มอนด์ ยึง

Raymond Yeung

- BS, MEng and PhD degrees in electrical engineering from **Cornell** University in 1984, 1985, and 1988, respectively.





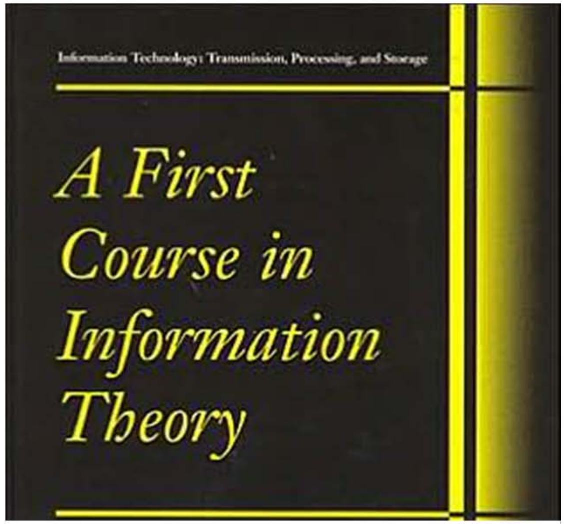


Prof. Yeung Wai-Ho, Raymond 楊偉豪教授

Choh-Ming Li Professor of Information Engineering (FIEEE, FHKIE)
卓敏信息工程學講座教授
Co-Director, Institute of Network Coding
Education: BS, MEng, PhD (Cornell)
Research Area: Communications and Information Theory
Contact
Tel: (852) 3943-8375
Fax: (852) 2603-5032
Address: Rm 733, Ho Sin Hang Engineering Building, CUHK
Email: whyeung [@] ie.cuhk.edu.hk

Website

Foreword



The first course usually is an appetizer. In the case of Raymond Yeung's *A First Course in Information Theory*, however, another delectable dish gets served up in each of the sixteen chapters. Chapters 1 through 7 deal with the basic concepts of entropy and information with applications to lossless source coding. This is the traditional early fare of an information theory text, but Yeung flavors it uniquely. No one since Shannon has had a better appreciation for the mathematical structure of information quantities than Prof. Yeung. In the early chapters this manifests itself in a careful treatment of information measures via both Yeung's analytical theory of I-Measure and his geometrically intuitive information diagrams. (This material, never before presented in a textbook, is rooted in works by G. D. Hu, by H. Dyckman
Fundamental interrelations among information meas
are developed with precision and unity. New slants are
the divergence inequality, the data processing theorer
There is also a clever, Kraft-inequality-free way of
length of the words in a lossless prefix source code
entropy. An easily digestible treatment of the redur
source codes also is served up, an important topic i
slighted in textbooks.

The concept of weakly typical sequences is introdu
chor Yeung's proof of the lossless block source codi
of strongly typical sequences is introduced next. La
icality, this provides a foundation for proving the ch
Chapter 8, the lossy source coding (rate-distortion) t
selected multi-source network coding theorems in C
proof of the channel coding theorem follows standa
development of the interplay between information q
ness readies one's palate for a rigorous proof that fe
memoryless channel does not increase its capacity.

vii

The first course usually is an appetizer. In the case of Raymond Yeung's *A First Course in Information Theory*, however, another delectable dish gets served up in each of the sixteen chapters. Chapters 1 through 7 deal with the basic concepts of entropy and information with applications to lossless source coding. This is the traditional early fare of an information theory text, but Yeung flavors it uniquely. No one since Shannon has had a better appreciation for the mathematical structure of information quantities than Prof. Yeung. In the early chapters this manifests itself in a careful treatment of information measures via both Yeung's analytical theory of I-Measure and his geometrically intuitive information diagrams. (This material, never before presented in a textbook, is rooted in works by G. D. Hu, by H. Dyckman, and by R. Yeung *et al.*)

FOREWORD

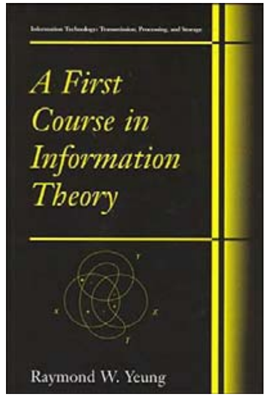
hensive theory remains elusive. Non
information inequalities developed i
tool for attacking this class of problems. The closing chapter linking entropy to the theory of groups is mouthwateringly provocative, having the potential to become a major contribution of information theory to this renowned branch of mathematics and mathematical physics.

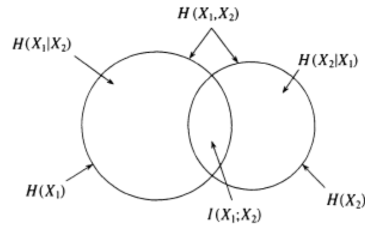
Savor this book; I think you will agree the proof is in the pudding.

Toby Berger
Irwin and Joan Jacobs Professor of Engineering
Cornell University, Ithaca, New York

Raymond Yeung

- Introduce, for the first time in a textbook,
  - analytical theory of I-Measure and
  - geometrically intuitive information diagrams
  - Establish a one-to-one correspondence between Shannon’s information measures and set theory.
- Rooted in works by G. D. Hu, by H. Dyckman, and by R. Yeung et al.





Chapter 6

THE *I*-MEASURE

In Chapter 2, we have shown the relationship between Shannon’s information measures for two random variables by the diagram in Figure 2.2. For convenience, Figure 2.2 is reproduced in Figure 6.1 with the random variables X and Y replaced by X_1 and X_2, respectively. This diagram suggests that Shannon’s information measures for any $n \geq 2$ random variables may have a set-theoretic structure.

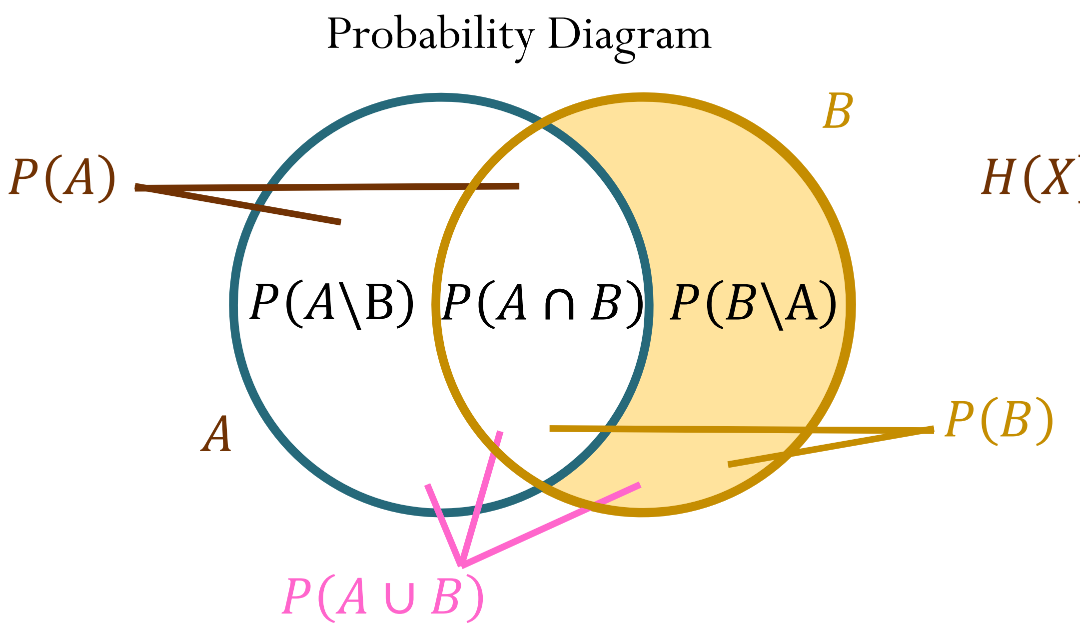
In this chapter, we develop a theory which establishes a one-to-one correspondence between Shannon’s information measures and set theory in full generality. With this correspondence, manipulations of Shannon’s information measures can be viewed as set operations, thus allowing the rich suite of tools in set theory to be used in information theory. Moreover, the structure of Shannon’s information measures can easily be visualized by means of an

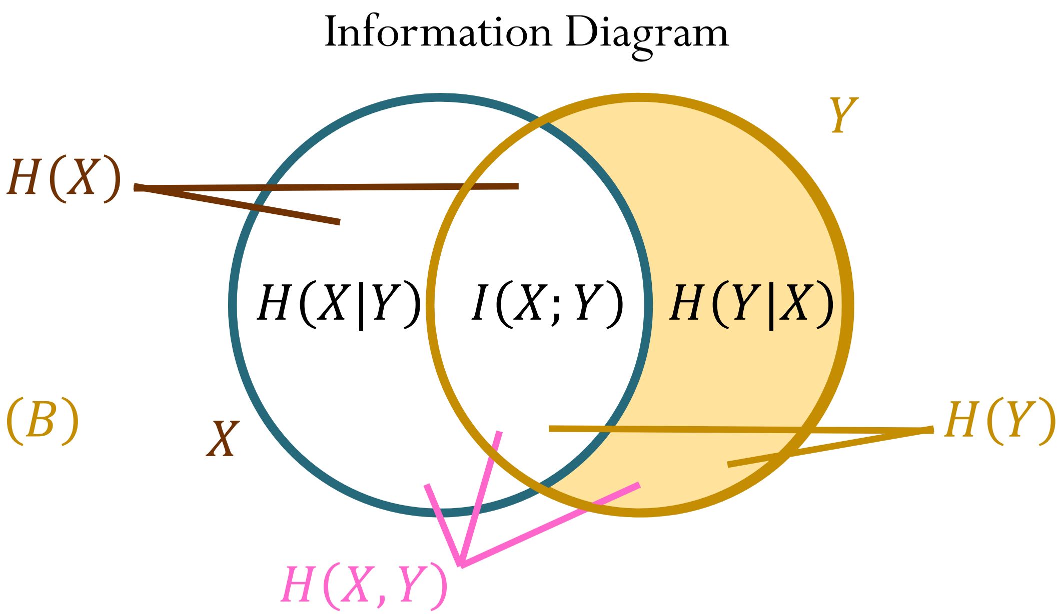
Figure 6.1. Relationship between entropies and mutual information for two random variables.

95

R. W. Yeung, *A First Course in Information Theory*

Diagrams





$$P(B\backslash A) = P(A \cup B) - P(A) \qquad H(Y|X) = H(X,Y) - H(X)$$

Conditional Entropies

Amount of randomness in Y

$$H(Y) \equiv -\sum_{y \in \mathcal{Y}} q(y) \log_2 \overbrace{q(y)}^{P[Y=y]} \equiv H(\underline{\mathbf{q}})$$

Amount of randomness still remained in Y when we know that $X = x$.

$$H(\overbrace{Y|X=x}^{\text{given a particular value } x}) \equiv H(Y|x) \equiv -\sum_{y \in \mathcal{Y}} Q(y|x) \log_2 \overbrace{Q(y|x)}^{P[Y=y|X=x]}$$

Apply the entropy calculation to a row from the $\mathbf{Q}$ matrix

The **average** amount of randomness still remained in Y when we know X

$$H(Y|X) \equiv \overbrace{\sum_{x \in \mathcal{X}} p(x) H(Y|x)}^{\textbf{average} \text{ of } H(Y|x)}$$

$$= H(X,Y) - H(X)$$

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

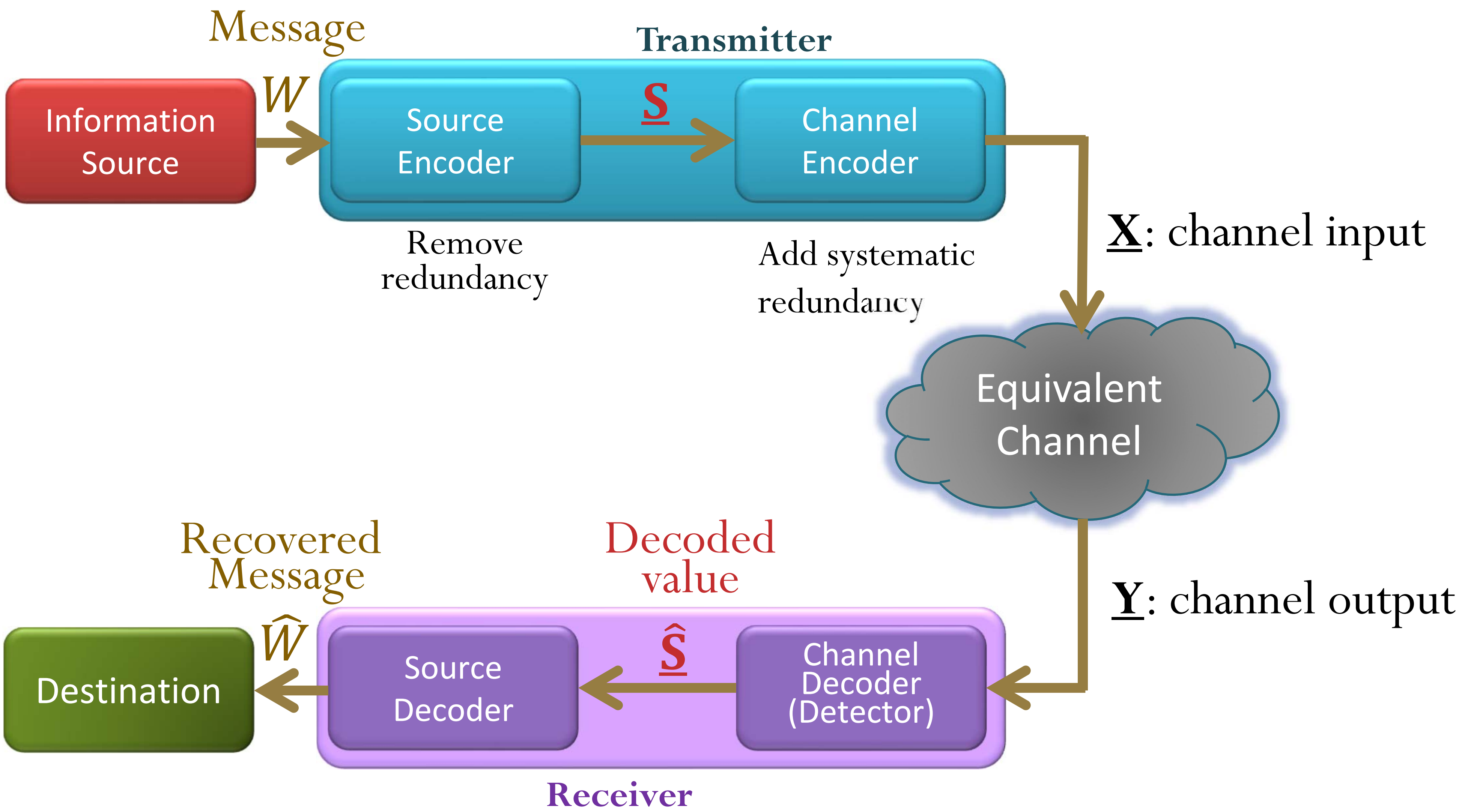
Digital Communication Systems ECS 452

Asst. Prof. Dr. Prapun Suksompong

prapun@siit.tu.ac.th

Information Channel Capacity

System Model for Section 3.5



Channel Capacity

[Section 4.2]

Channel Capacity

"**Operational**": max rate at which **reliable** communication is possible

Arbitrarily small error probability can be achieved.

"**Information**": $\boxed{\max_{\underline{\mathbf{p}}} I(X;Y)}$ [bpcu]

[Section 4.3]

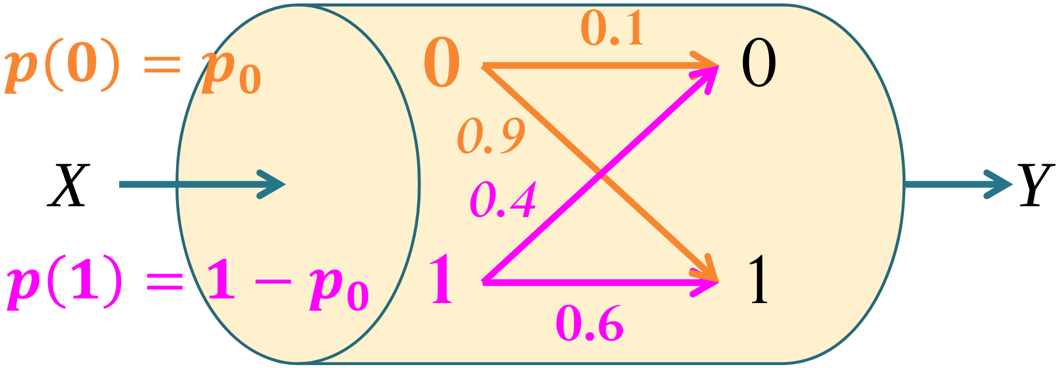
Shannon [1948] shows that these two quantities are actually the same.

MATLAB

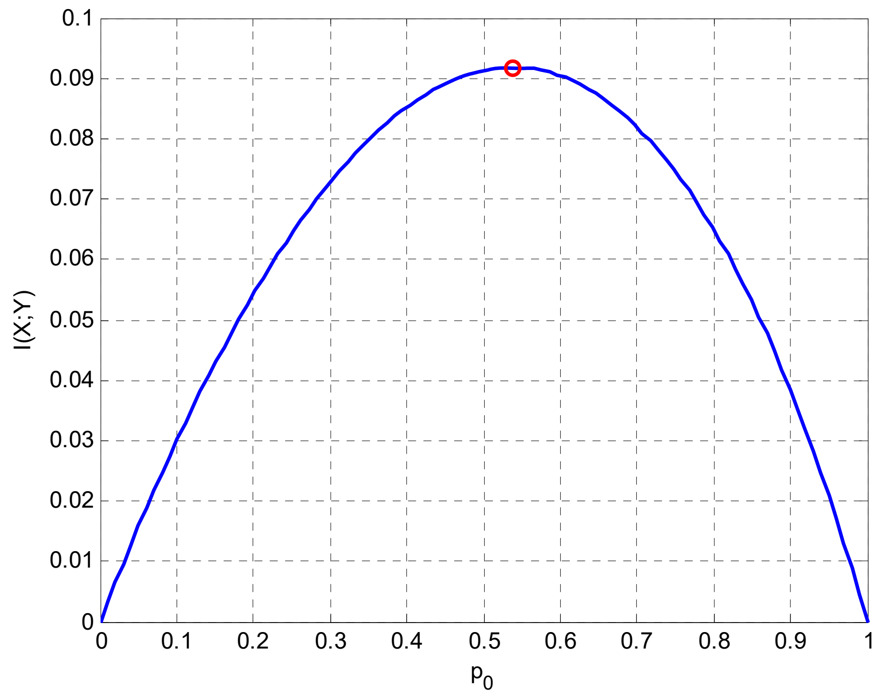
```
function H = entropy2s(p)
% ENTROPY2S accepts probability mass function
% as a row vector, calculate the corresponding
% entropy in bits.
p=p(find(abs(sort(p)-1)>1e-8)); % Eliminate 1
p=p(find(abs(p)>1e-8)); % Eliminate 0
if length(p)==0
    H = 0;
else
    H = simplify(-sum(p.*log(p))/log(sym(2)));
end
```

```
function I = informations(p,Q)
X = length(p);
q = p*Q;
HY = entropy2s(q);
temp = [];
for i = 1:X
    temp = [temp entropy2s(Q(i,:))];
end
HYgX = sum(p.*temp);
I = HY-HYgX;
```

Capacity calculation for BAC

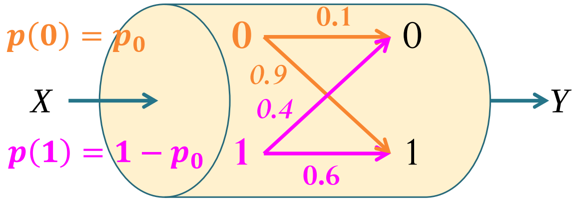


$$Q = \begin{bmatrix} 0.1 & 0.9 \\ 0.4 & 0.6 \end{bmatrix}$$



Capacity of 0.0918 bits is achieved by $\underline{p} = [0.5380,\ 0.4620]$

Capacity calculation for BAC



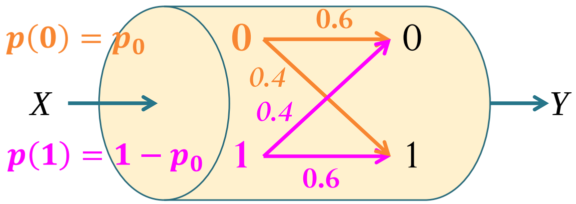
$$\mathbf{Q} = \begin{bmatrix} 0.1 & 0.9 \\ 0.4 & 0.6 \end{bmatrix}$$

```
close all; clear all;
syms p0
p = [p0 1-p0];
Q = [1 9; 4 6]/sym(10);

I = simplify(informations(p,Q))

p0o = simplify(solve(diff(I)==0))

po = eval([p0o 1-p0o])

C = simplify(subs(I,p0,p0o))

eval(C)
```

```
>> Capacity_Ex_BAC
I =
(log(2/5 - (3*p0)/10)*((3*p0)/10 - 2/5) - log((3*p0)/10 + 3/5)*((3*p0)/10 + 3/5))/log(2) + (log((5*2^(3/5)*3^(2/5))/6)*(p0 - 1))/log(2) + (p0*log((3*3^(4/5))/10))/log(2)
p0o =
(27648*2^(1/3))/109565 - (69984*2^(2/3))/109565 + 135164/109565
po =
    0.5376    0.4624
C =
(log((3*3^(4/5))/10)*((27648*2^(1/3))/109565 - (69984*2^(2/3))/109565 + 135164/109565))/log(2) - (log((104976*2^(2/3))/547825 - (41472*2^(1/3))/547825 + 16384/547825)*((104976*2^(2/3))/547825 - (41472*2^(1/3))/547825 + 16384/547825) + log((41472*2^(1/3))/547825 - (104976*2^(2/3))/547825 + 531441/547825)*((41472*2^(1/3))/547825 - (104976*2^(2/3))/547825 + 531441/547825))/log(2) + (log((5*2^(3/5)*3^(2/5))/6)*((27648*2^(1/3))/109565 - (69984*2^(2/3))/109565 + 25599/109565))/log(2)
ans =
    0.0918
```

[Ex. 4.25, Fig. 19]

Same procedure applied to BSC



$$\mathbf{Q} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

```
close all; clear all;
syms p0
p = [p0 1-p0];
Q = [6 4; 4 6]/sym(10);

I = simplify(informations(p,Q))

p0o = simplify(solve(diff(I)==0))

po = eval([p0o 1-p0o])

C = simplify(subs(I,p0,p0o))

eval(C)
```

```
>> Capacity_Ex_BSC
I =
(log((5*2^(3/5)*3^(2/5))/6)*(p0 - 1))/log(2) -
(p0*log((5*2^(3/5)*3^(2/5))/6))/log(2) - (log(p0/5 +
2/5)*(p0/5 + 2/5) - log(3/5 - p0/5)*(p0/5 -
3/5))/log(2)
p0o =
1/2
po =
   0.5000   0.5000
C =
log((2*2^(2/5)*3^(3/5))/5)/log(2)
ans =
   0.0290
```

Blahut, Richard (1972), "Computation of channel capacity and rate-distortion functions", IEEE Transactions on Information Theory, 18 (4): 460–473

Computation of Channel Capacity and Rate-Distortion Functions

RICHARD E. BLAHUT, MEMBER, IEEE

***Abstract*—By defining mutual information as a maximum over an appropriate space, channel capacities can be defined as double maxima and rate-distortion functions as double minima. This approach yields valuable new insights regarding the computation of channel capacities and rate-distortion functions. In particular, it suggests a simple algorithm for computing channel capacity that consists of a mapping from the set of channel input probability vectors into itself such that the sequence of probability vectors generated by successive applications of the mapping converges to the vector that achieves the capacity of the**

Arimoto [13] used the first of the preceding expressions in an investigation of C, thereby obtaining Theorems 1 and 3 as well as Corollary 2 of this paper.[1]

This approach places the existing theory of C and $R(D)$ in a more transparent setting and suggests several new results. In particular, the approach in question results in algorithms for determining C and $R(D)$ by means of map-

Arimoto, Suguru (1972), "An algorithm for computing the capacity of arbitrary discrete memoryless channels", IEEE Transactions on Information Theory, 18 (1): 14–20

An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels

SUGURU ARIMOTO

***Abstract*—A systematic and iterative method of computing the capacity of arbitrary discrete memoryless channels is presented. The algorithm is very simple and involves only logarithms and exponentials in addition to elementary arithmetical operations. It has also the property of monotonic convergence to the capacity. In general, the approximation error is at least inversely proportional to the number of iterations; in certain circumstances, it is exponentially decreasing. Finally, a few inequalities that give upper and lower bounds on the capacity are derived.**

Manuscript received September 9, 1970.

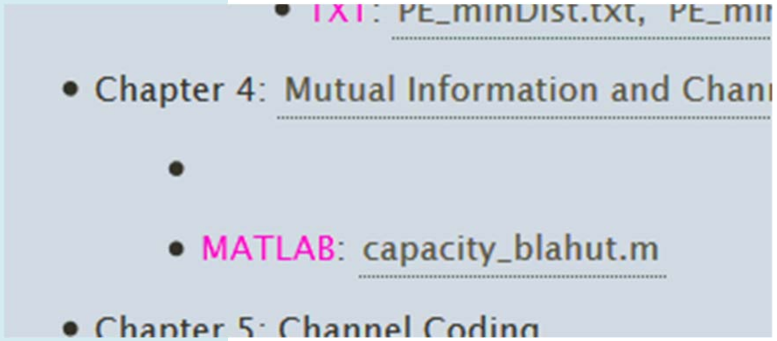
The author is with the Faculty of Engineering Science, Osaka University, Osaka, Japan.

I. Introduction

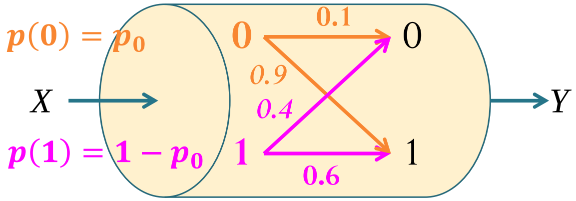
IT IS well known that the capacity of discrete memoryless channels that are symmetric from the input can easily be evaluated. Muroga [1] developed a method for straightforward evaluation of capacity, but unfortunately its usefulness is restricted to the case where 1) the channel

Blahut–Arimoto algorithm [4.26]

```
function [ps C] = capacity_blahut(Q)
% Input:     Q  = channel transition probability matrix
% Output:    C  = channel capacity
%            ps = row vector containing pmf that achieves capacity

tl = 1e-8; % tolerance (for the stopping condition)
n = 1000; % max number of iterations (in case the stopping condition
          % is "never" reached")
nx = size(Q,1); pT = ones(1,nx)/nx; % First, guess uniform X.
for k = 1:n
    qT = pT*Q;
    % Eliminate the case with 0
    % Column-division by qT
    temp = Q.*(ones(nx,1)*(1./qT));
    %Eliminate the case of 0/0
    l2 = log2(temp);
    l2(find(isnan(l2) | (l2==-inf) | (l2==inf)))=0;
    logc = (sum(Q.*(l2),2))';
    CT = 2.^(logc);
    A = log2(sum(pT.*CT)); B = log2(max(CT));
    if((B-A)<tl)
        break
    end
    % For the next loop
    pT = pT.*CT;     % un-normalized
    pT = pT/sum(pT); % normalized
    if(k == n)
        fprintf('\nNot converge within n loops\n')
    end
end
ps = pT;
C = (A+B)/2;
```



Capacity calculation for BAC: a revisit



$$\mathbf{Q} = \begin{bmatrix} 0.1 & 0.9 \\ 0.4 & 0.6 \end{bmatrix}$$

```
close all; clear all;

Q = [1 9; 4 6]/10;

[ps C] = capacity_blahut(Q)
```

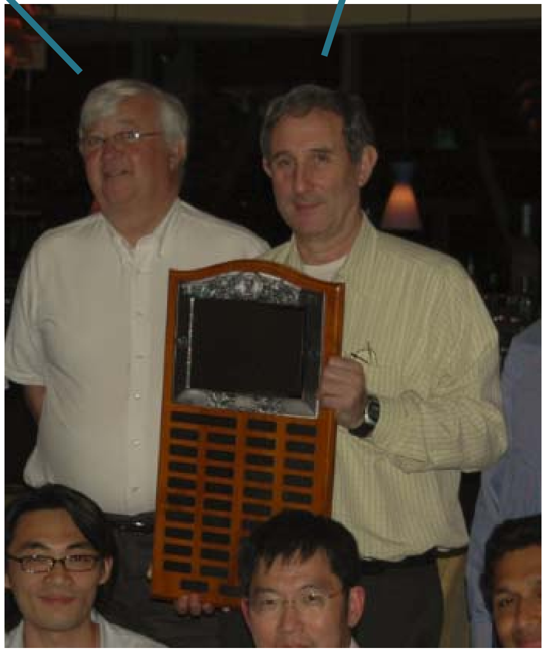
```
>> Capacity_Ex_BAC_blahut
ps =
    0.5376    0.4624
C =
    0.0918
```

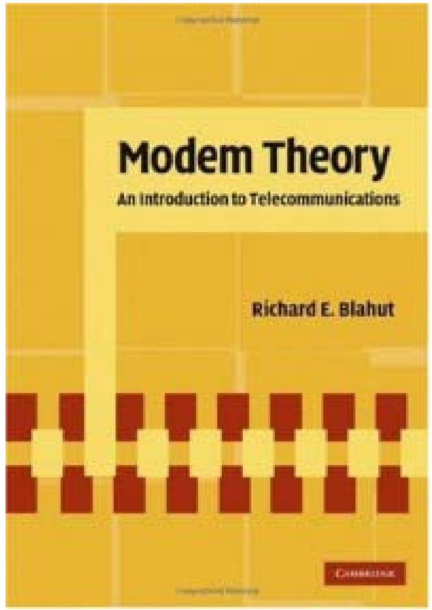
Toby Berger with Berger plaque

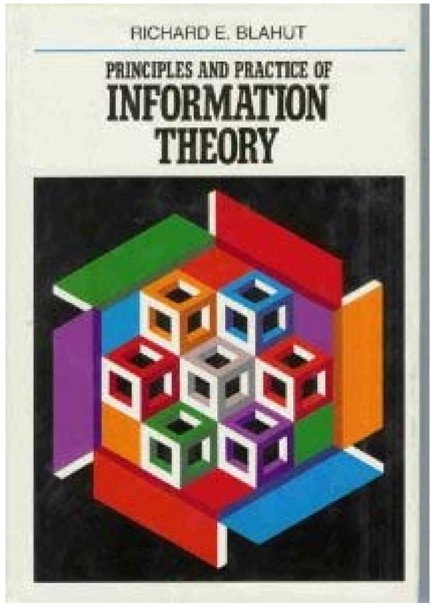
Richard Blahut

- Former chair of the Electrical and Computer Engineering Department at the University of Illinois at Urbana-Champaign
- Best known for **Blahut–Arimoto algorithm (Iterative Calculation of C)**

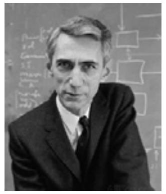
Toby Berger







Claude E. Shannon Award



Claude E. **Shannon** (1972)
David S. Slepian (1974)
Robert M. Fano (1976)
Peter Elias (1977)
Mark S. Pinsker (1978)
Jacob Wolfowitz (1979)
W. Wesley Peterson (1981)
Irving S. Reed (1982)
Robert G. Gallager (1983)
Solomon W. Golomb (1985)
William L. Root (1986)
James L. Massey (1988)
Thomas M. Cover (1990)
Andrew J. Viterbi (1991)

Elwyn R. Berlekamp (1993)
Aaron D. Wyner (1994)
G. David Forney, Jr. (1995)
Imre Csiszár (1996)
Jacob Ziv (1997)
Neil J. A. Sloane (1998)
Tadao Kasami (1999)
Thomas Kailath (2000)
Jack Keil Wolf (2001)
Toby **Berger** (2002)
Lloyd R. Welch (2003)
Robert J. McEliece (2004)
Richard Blahut (2005)
Rudolf Ahlswede (2006)

Sergio Verdu (2007)
Robert M. Gray (2008)
Jorma Rissanen (2009)
Te Sun Han (2010)
Shlomo Shamai (Shitz) (2011)
Abbas El Gamal (2012)
Katalin Marton (2013)
János Körner (2014)
Arthur Robert Calderbank (2015)
Alexander S. Holevo (2016)
David Tse (2017)
Gottfried Ungerboeck (2018)
Erdal Arıkan (2019)
Charles Bennett (2020)

[http://www.itsoc.org/honors/claude-e-shannon-award]
[https://en.wikipedia.org/wiki/Claude_E._Shannon_Award]

Digital Communication Systems ECS 452

Asst. Prof. Dr. Prapun Suksompong

prapun@siit.tu.ac.th

Special Cases for Calculation of Channel Capacity

Calculating channel capacity

1. Use (multi-variable) calculus
   - standard nonlinear optimization techniques
2. Use Blahut-Arimoto algorithm (MATLAB)
3. Check whether we can match the **Q** matrix with any known special cases.

Remark: Do not assume that the input probabilities will have to be uniform to obtain C.

- See BAC in Ex. 4.25.

Channel Capacity: Special Cases

- **Channel with Nonoverlapping Outputs (NO^2)**
  - There is only one non-zero element in each column of its **Q** matrix.
  - $\boxed{C = \log_2 |\mathcal{X}|}$ [4.30]
    is achieved by uniform input probabilities.
  - Ex. Noiseless Binary Channel: $C = 1$ [bpcu] [Ex. 4.27]
- **Weakly Symmetric Channel**
  - (1) all the rows of **Q** are permutations of each other and (2) all the column sums are equal. [Defn 4.36]
  - $\boxed{C = \log_2 |\mathcal{Y}| - H(\underline{\mathbf{r}})}$ where $\underline{\mathbf{r}}$ is any row from the **Q** matrix. [4.37]
    is achieved by uniform input probabilities.
  - Ex. Binary Symmetric Channel: $C = 1 - H(p)$ [bpcu]