ECE 515
Information Theory

Joint Entropy, Equivocation and Mutual Information

Entropy

$$H(X) = -\sum_{i=1}^{N} p(x_i) \; \log_b p(x_i)$$

Joint Entropy

$$H(XY) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i, y_j)$$

Conditional Entropy

$$H(X|Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i|y_j)$$

Chain Rule

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

$$H(X_1, \ldots, X_N) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_N|X_1, \ldots, X_{N-1})$$

$$
\begin{aligned}
H(XYZ) &= H(X) + H(Y|X) + H(Z|XY) \\
&= H(X) + H(Z|X) + H(Y|XZ) \\
&= H(Y) + H(X|Y) + H(Z|XY) \\
&= H(Y) + H(Z|Y) + H(X|YZ) \\
&= H(Z) + H(X|Z) + H(Y|XZ) \\
&= H(Z) + H(Y|Z) + H(X|YZ).
\end{aligned}
$$

Example

- Random vector [X,Y,Z]
- Equally likely to take any of the 4 values

  [0,0,0]

  [0,1,0]

  [1,0,0]

  [1,0,1]

- Find H(XYZ) using

  H(XYZ) = H(X) + H(Y|X) + H(Z|XY)

Example

$$H(X) = -\sum_{i=1}^{2} p(x_i) log_2 \, p(x_i)$$

$$H(Y|X) = -\sum_{i=1}^{2}\sum_{j=1}^{2} p(x_i, y_j) log_2 \, p(y_j|x_i)$$

$$H(Z|XY)$$

$$= -\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} p(x_i, y_j, z_k) log_2 \, p(z_k|x_i, y_j)$$

Example

$$p(x_1) = p(x_2) = \tfrac{1}{2}$$
$$H(X) = -\tfrac{1}{2}\log_2\tfrac{1}{2} - \tfrac{1}{2}\log_2\tfrac{1}{2} = 1 \text{ bit}$$

$p(x_i,y_j)$	$p(y_j\mid x_j)$
$p(0,0) = \tfrac{1}{4}$	$p(0\mid0) = \tfrac{1}{2}$
$p(1,1) = 0$	$p(1\mid1) = 0$
$p(0,1) = \tfrac{1}{4}$	$p(1\mid0) = \tfrac{1}{2}$
$p(1,0) = \tfrac{1}{2}$	$p(0\mid1) = 1$

$$H(Y\mid X) = -\tfrac{1}{4}\log_2\tfrac{1}{2} - \tfrac{1}{4}\log_2\tfrac{1}{2} - \tfrac{1}{2}\log_2 1$$
$$= \tfrac{1}{2}\log_2 2 = \tfrac{1}{2} \text{ bit}$$

Example

$p(x_i, y_j, z_k)$ $p(z_k | x_j, y_j)$

$p(0,0,0) = \frac{1}{4}$ $p(0|0,0) = 1$

$p(0,1,0) = \frac{1}{4}$ $p(0|0,1) = 1$

$p(1,0,0) = \frac{1}{4}$ $p(0|1,0) = \frac{1}{2}$

$p(1,0,1) = \frac{1}{4}$ $p(1|1,0) = \frac{1}{2}$

$H(Z|XY) = -\frac{1}{4}\log_2 1 - \frac{1}{4}\log_2 1 - \frac{1}{4}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{2}$

$$= \frac{1}{2}\log_2 2 = \frac{1}{2} \text{ bit}$$
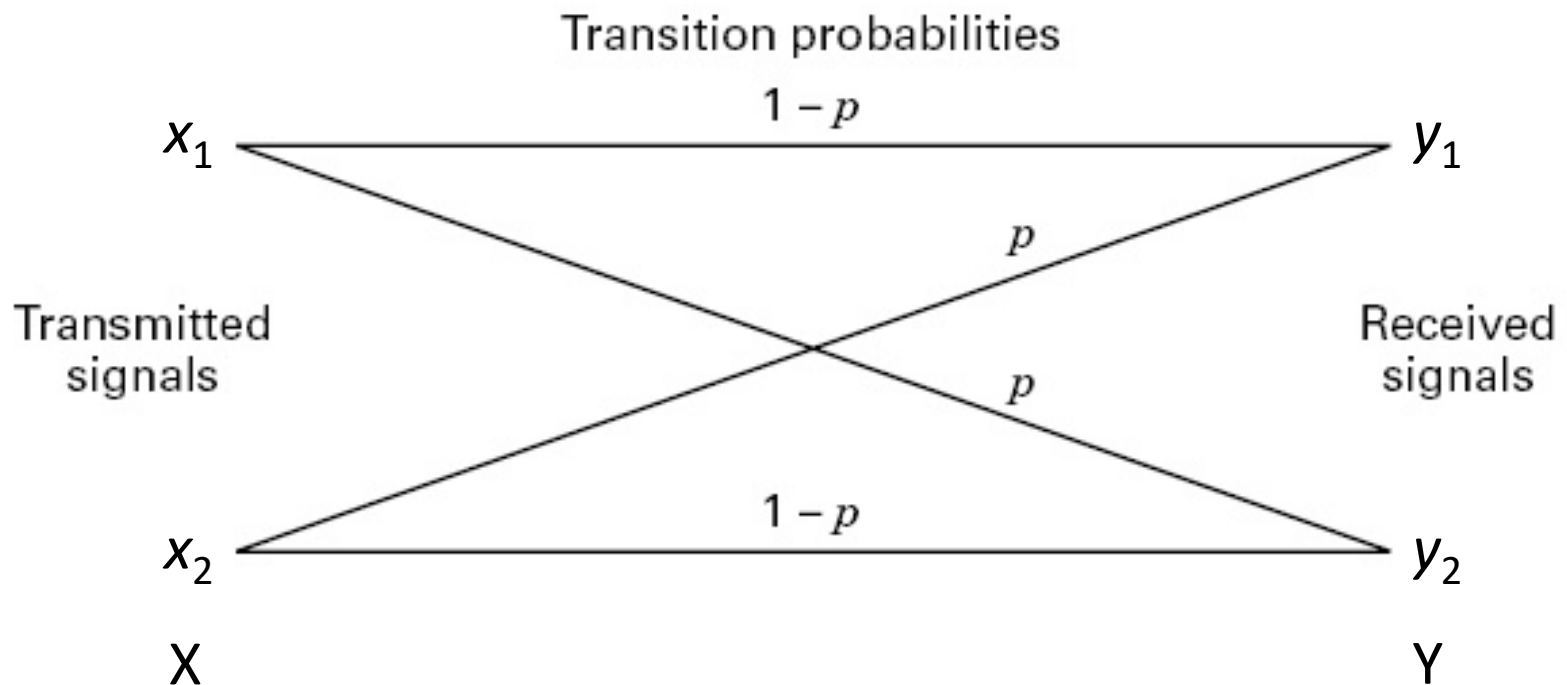
Example

- $H(XYZ) = H(X) + H(Y|X) + H(Z|XY)$
- $H(X) = 1$ bit
- $H(Y|X) = \frac{1}{2}$ bit
- $H(Z|XY) = \frac{1}{2}$ bit
- $H(XYZ) = 1 + \frac{1}{2} + \frac{1}{2} = 2$ bits

- $p(y_1) = \frac{3}{4} \quad p(y_2) = \frac{1}{4}$
- $H(Y) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = .811$ bit $> H(Y|X)$

Information Channels
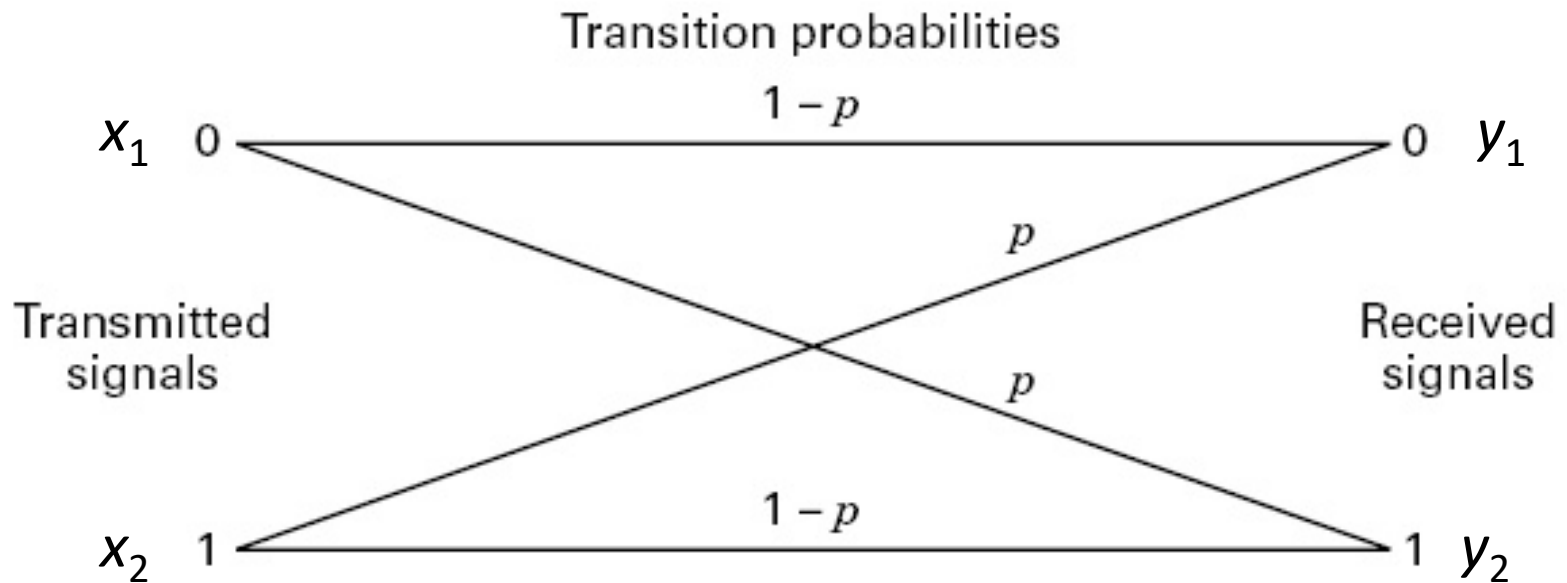
- An information channel is described by an
- Input random variable X
- Output random variable Y
- Set of conditional probabilities p(y_j|x_i)

X $\longrightarrow$ | Channel | $\longrightarrow$ Y

Binary Symmetric Channel

Transition probabilities

$$1 - p$$

x_1 y_1

p

Transmitted signals   Received signals

p

$$1 - p$$

x_2 y_2

X   Y

Binary Symmetric Channel

Transition probabilities

$1-p$

x_1 0 ————————————————— 0 y_1

p

Transmitted
signals

Received
signals

p

$1-p$

x_2 1 ————————————————— 1 y_2

$$p_{Y|X}(0|1) = p_{Y|X}(1|0) = p$$
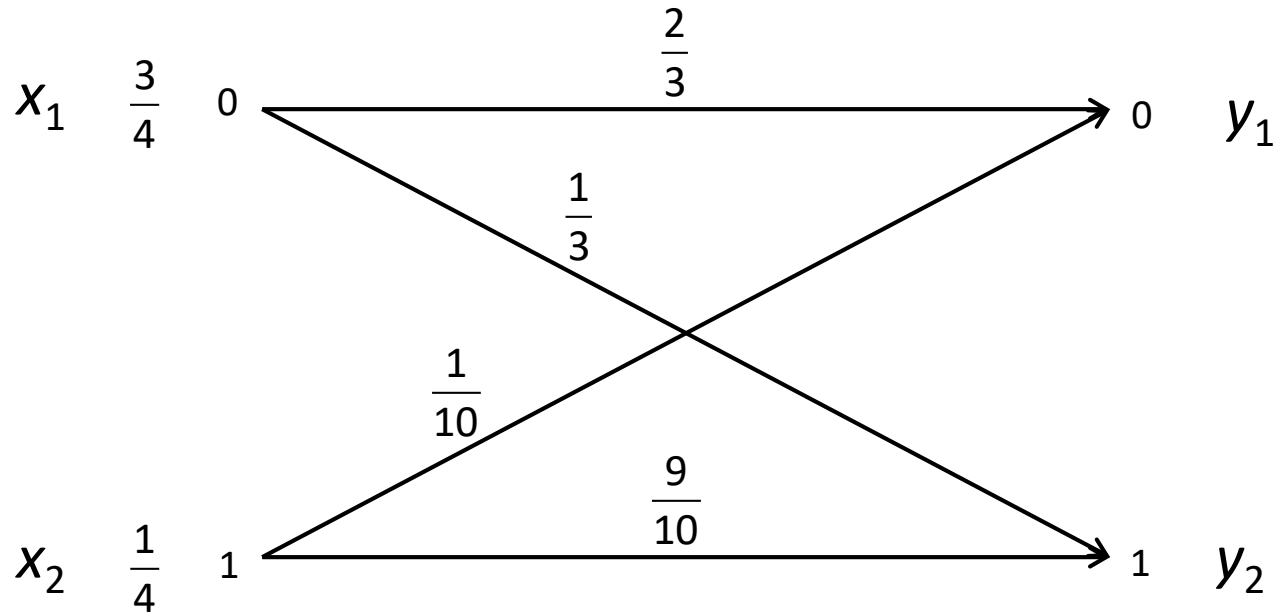$$p_{Y|X}(0|0) = p_{Y|X}(1|1) = 1\text{-}p$$

$$\begin{bmatrix} p(y_1 \,|\, x_1) & p(y_2 \,|\, x_1) \\ p(y_1 \,|\, x_2) & p(y_2 \,|\, x_2) \end{bmatrix} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

- The probabilities p(y_j|x_i) are called the **forward transition probabilities**

- Using Bayes' Theorem

$$p(x_i \mid y_j) = \frac{p(y_j \mid x_i)p(x_i)}{p(y_j)}$$

- The probabilities p(x_i|y_j) are called the **backward transition probabilities**

Non-symmetric Binary Channel

x_1 $\dfrac{3}{4}$ 0 $\xrightarrow{\quad\dfrac{2}{3}\quad}$ 0 y_1

$\dfrac{1}{3}$

$\dfrac{1}{10}$

$\dfrac{9}{10}$

x_2 $\dfrac{1}{4}$ 1 1 y_2

channel matrix $\begin{bmatrix} p(y_1 \mid x_1) & p(y_2 \mid x_1) \\ p(y_1 \mid x_2) & p(y_2 \mid x_2) \end{bmatrix} = \begin{bmatrix} \dfrac{2}{3} & \dfrac{1}{3} \\ \dfrac{1}{10} & \dfrac{9}{10} \end{bmatrix}$

Backward Transition Probabilities

$$\begin{bmatrix} p(x_1|y_1) & p(x_2|y_1) \\ p(x_1|y_2) & p(x_2|y_2) \end{bmatrix} = \begin{bmatrix} p_{X|Y}(0|0) & p_{X|Y}(1|0) \\ p_{X|Y}(0|1) & p_{X|Y}(1|1) \end{bmatrix} = \begin{bmatrix} \dfrac{20}{21} & \dfrac{1}{21} \\ \dfrac{10}{19} & \dfrac{9}{19} \end{bmatrix}$$

- H(X|y=0) = - p(1|0)log$_2$p(1|0) - p(0|0)log$_2$p(0|0)
    = - (1/21)log$_2$(1/21) - (20/21)log$_2$(20/21)
    = .209 + .067 = .276 bit

- H(X|y=1) = - p(1|1)log$_2$p(1|1) - p(0|1)log$_2$p(0|1)
    = - (9/19)log$_2$(9/19) - (10/19)log$_2$(10/19)
    = .511 + .487 = .998 bit

- H(X|Y) = p(y=0) H(X|y=0) + p(y=0) H(X|y=0)
    = (21/40)×(.276) + (19/40)×(.998) = .619 bit

Two Questions

- Given two random variables X and Y

    – How much information does Y give about X?

    – How much information does X give about Y?

Mutual Information

x_i ⟶ | Channel | ⟶ y_j

Mutual Information

$$I(x_i; y_j) = I(x_i) - I(x_i|y_j)$$
$$I(x_i; y_j) = -\log_b p(x_i) - [-\log_b p(x_i|y_j)]$$

$$I(x_i; y_j) = \log_b \frac{p(x_i|y_j)}{p(x_i)}$$

$$I(y_j; x_i) = I(y_j) - I(y_j|x_i)$$
$$I(y_j; x_i) = -\log_b p(y_j) - [-\log_b p(y_j|x_i)]$$

$$I(y_j; x_i) = \log_b \frac{p(y_j|x_i)}{p(y_j)}$$

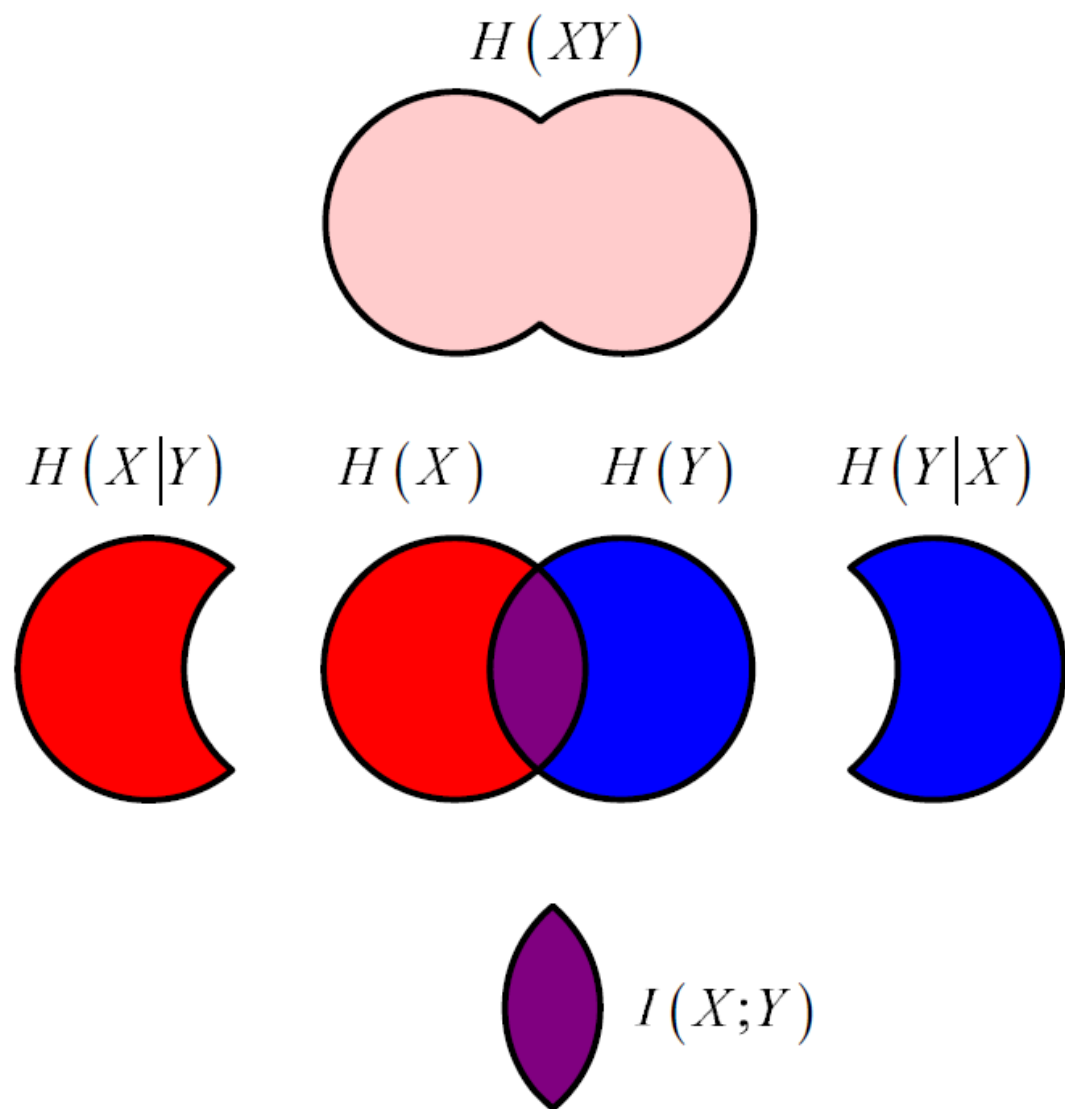$$\boxed{I(x_i; y_j) = I(y_j; x_i)}$$

Average Mutual Information

$$I(X;Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \, I(x_i; y_j)$$

$$I(X;Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \, \log_b \frac{p(x_i|y_j)}{p(x_i)}$$

Average Mutual Information

$$I(X;Y) \;=\; \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b \frac{p(x_i|y_j)}{p(x_i)}$$

$$=\; -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i)$$

$$+\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i|y_j)$$
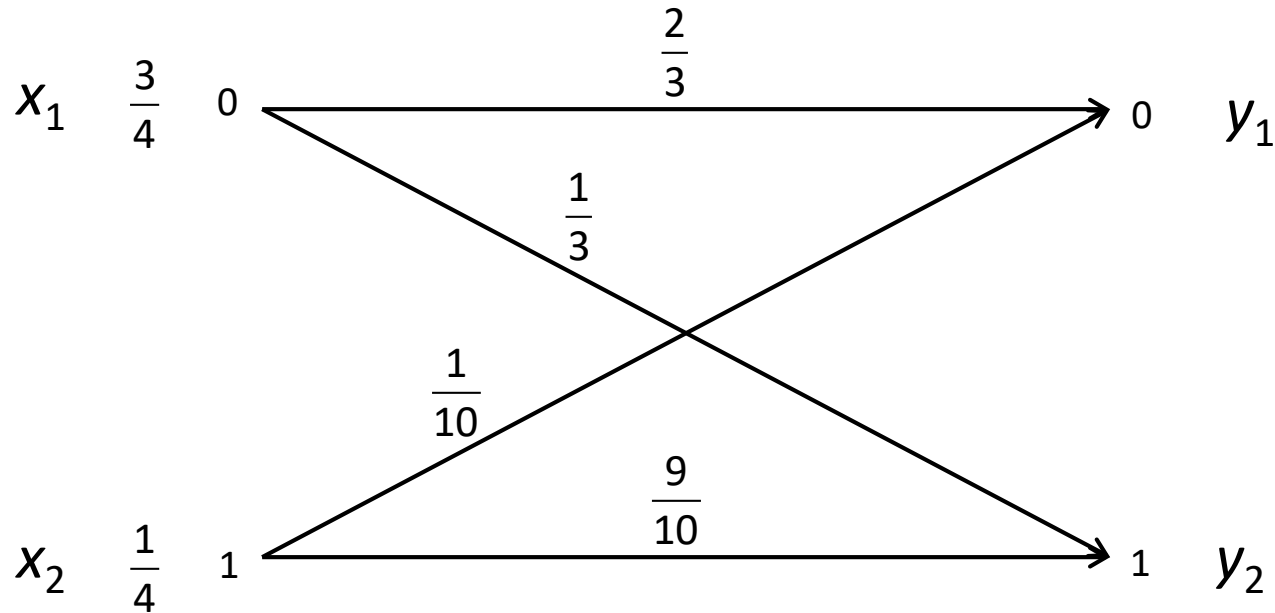
$$I(X;Y) \;=\; H(X) - H(X|Y)$$

$H(XY)$

$H(X|Y)$ $H(X)$ $H(Y)$ $H(Y|X)$

$I(X;Y)$

$H(X,Y)$

$H(X)$

$H(Y)$

$H(X|Y)$ $I(X;Y)$ $H(Y|X)$

Four Vectors Example

- [0,0,0], [0,1,0], [1,0,0], [1,0,1]
- p(y=1) = .250 → H(Y) = .811 bit
- I(X;Y) = H(Y) - H(Y|X)

  $\qquad$ = .811 - .500 = .311 bit
- p(x=1) = .50 → H(X) = 1 bit
- I(X;Y) = H(X) - H(X|Y)

  $\qquad$ = 1.0 - .689 = .311 bit
- H(X;Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) = 1.5 bits

Non-symmetric Binary Channel

x_1 $\frac{3}{4}$ $\quad$ 0 $\qquad\qquad \frac{2}{3} \qquad\qquad$ 0 $\quad y_1$

$\frac{1}{3}$

$\frac{1}{10}$

$\frac{9}{10}$

x_2 $\frac{1}{4}$ $\quad$ 1 $\qquad\qquad\qquad\qquad\qquad$ 1 $\quad y_2$

channel matrix $\begin{bmatrix} p(y_1 \mid x_1) & p(y_2 \mid x_1) \\ p(y_1 \mid x_2) & p(y_2 \mid x_2) \end{bmatrix} = \begin{bmatrix} \dfrac{2}{3} & \dfrac{1}{3} \\ \dfrac{1}{10} & \dfrac{9}{10} \end{bmatrix}$

Non-symmetric Channel Example

$I(X;Y)$ = .192 bit

$H(X)$ = .811 bit

- $H(X|Y) = H(X) - I(X;Y) = .619$ bit

$H(Y)$ = .998 bit

- $H(Y|X) = H(Y) - I(X;Y) = .806$ bit

- $H(XY) = H(X) + H(Y|X)$
  $= H(Y) + H(X|Y) = 1.617$ bits

Mutual Information for the BSC

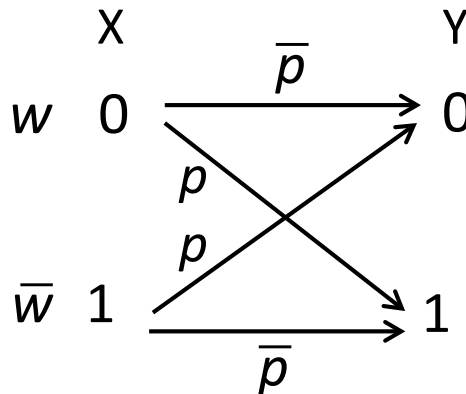X $\longrightarrow$ [BSC] $\longrightarrow$ Y

crossover probability p

$$\overline{p} = 1 - p$$

channel matrix

$$\begin{bmatrix} \overline{p} & p \\ p & \overline{p} \end{bmatrix}$$

X $\qquad$ Y

w 0 $\xrightarrow{\ \overline{p}\ }$ 0

p

p

$\overline{w}$ 1 $\xrightarrow{\ \overline{p}\ }$ 1

$\mathrm{p}(x=0)=w$

$\mathrm{p}(x=1)=1-w=\overline{w}$

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

Mutual Information for the BSC

$$
\begin{aligned}
H(Y|X) &= -\sum_i \sum_j p(x_i, y_j) \log p(y_j|x_i) \\
&= -\sum_i \sum_j p(x_i) p(y_j|x_i) \log p(y_j|x_i) \\
&= -\sum_i p(x_i) \sum_j p(y_j|x_i) \log p(y_j|x_i) \\
&= -\sum_i p(x_i) [\, p \log p + \overline{p} \log \overline{p} \,] \\
&= -[\, p \log p + \overline{p} \log \overline{p} \,]
\end{aligned}
$$

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) + [\, p \log p + \overline{p} \log \overline{p} \,] \\
&= H(Y) - h(p)
\end{aligned}
$$

Binary Entropy Function

$$h(p) = -p\log_2 p - (1-p)\log_2(1-p) \qquad 0 \le p \le 1$$

h(p)

p
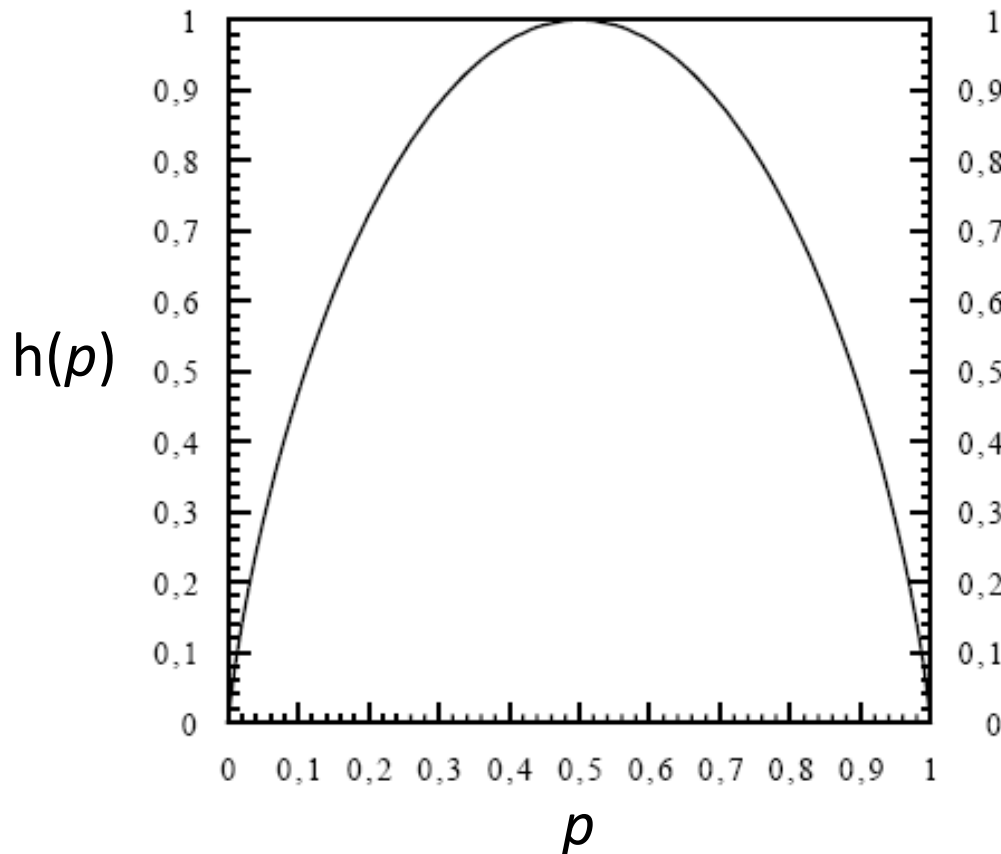
$$p(y = 0) \;\; = \;\; w\bar{p} + \bar{w}p$$
$$p(y = 1) \;\; = \;\; wp + \overline{wp}$$

$$
\begin{aligned}
H(Y) \;\; &= \;\; -[\,(w\bar{p} + \bar{w}p)\log(w\bar{p} + \bar{w}p) + (wp + \overline{wp})\log(wp + \overline{wp})\,] \\
&= \;\; h(wp + \overline{wp}) \\
I(X;Y) \;\; &= \;\; h(wp + \overline{wp}) - h(p)
\end{aligned}
$$

For $w = 1/2$

$$
\begin{aligned}
I(X;Y) \;\; &= \;\; h[1/2(p) + 1/2(1 - p)] - h(p) \\
&= \;\; h(1/2) - h(p) \\
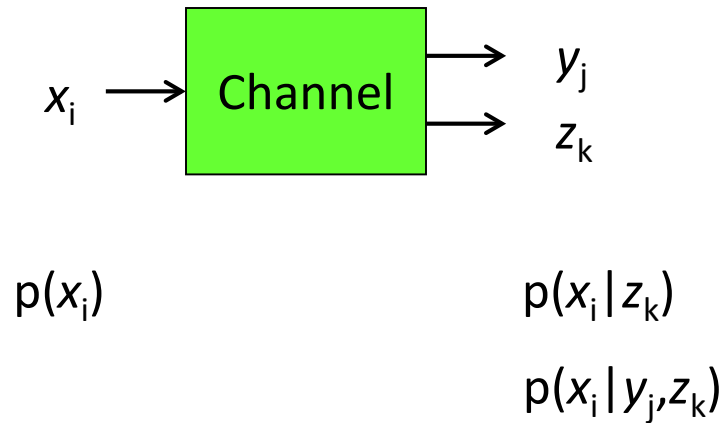&= \;\; 1 - h(p)
\end{aligned}
$$

For $w = 0$ or 1

$$
\begin{aligned}
I(X;Y) \;\; &= \;\; h(p) - h(p) \\
&= \;\; 0
\end{aligned}
$$

Entropy function

$$H(Y) = h(wp + \overline{wp})$$

$$H(Y|X) = h(p)$$

$$I(X;Y)$$

$$\omega p + \overline{\omega}\overline{p}$$

0 p $\frac{1}{2}$ $\overline{p}$ 1.0

34

Conditional Mutual Information

$x_i \rightarrow$ Channel $\rightarrow y_j$, z_k

$p(x_i)$ $\qquad\qquad p(x_i|z_k)$
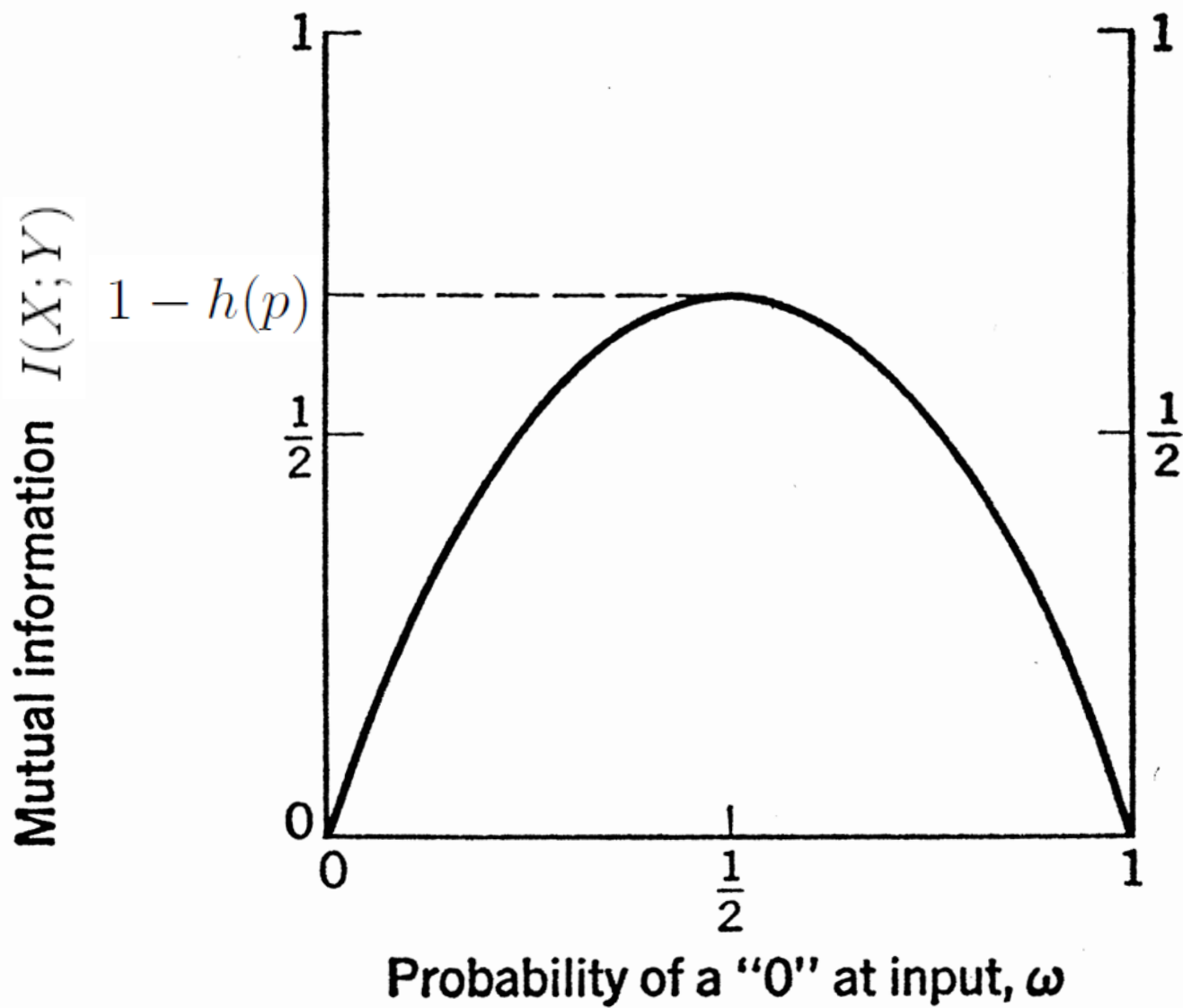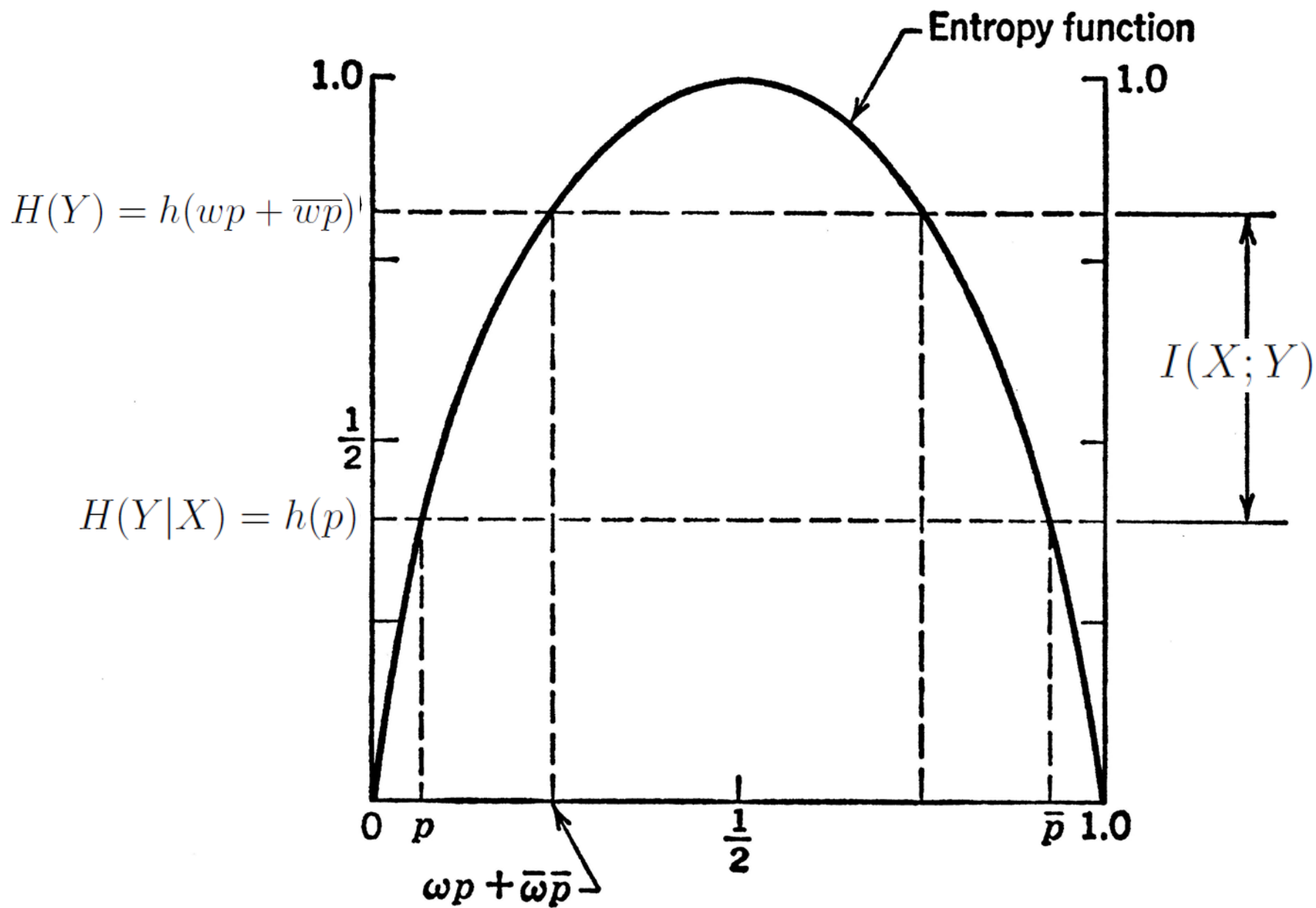
$\qquad\qquad\qquad\qquad p(x_i|y_j,z_k)$

Conditional Mutual Information

$$I(x_i; y_j | z_k) = I(x_i | z_k) - I(x_i | y_j, z_k)$$

$$I(x_i; y_j | z_k) \equiv \log_b \frac{p(x_i | y_j, z_k)}{p(x_i | z_k)}$$

Conditional Mutual Information

$$I(X;Y|Z) \;=\; \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} p(x_i, y_j, z_k) \log_b \left[\frac{p(x_i|y_j, z_k)}{p(x_i|z_k)} \right]$$

$$I(X;Y|Z) \;=\; \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} p(x_i, y_j, z_k) \log_b p(x_i|y_j, z_k)$$

$$-\; \sum_{i=1}^{N}\sum_{k=1}^{L} p(x_i, z_k) \log_b p(x_i|z_k)$$

$$I(X;Y|Z) = H(X|Z) - H(X|YZ)$$

Conditional Mutual Information

$$I(X;Y|Z) = H(X|Z) - H(X|YZ)$$

$H(X|Z)$: average uncertainty remaining in X after the observation in Z

$H(X|YZ)$: average uncertainty remaining in X after the observation in both Z and Y

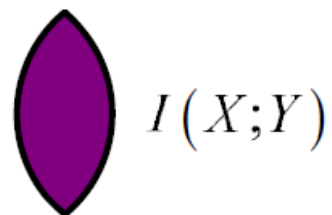$I(X;Y|Z)$: average amount of uncertainty in X *resolved* by the observation in Y

$$H(XY)$$

$$H(X|Y) \qquad H(X) \qquad H(Y) \qquad H(Y|X)$$
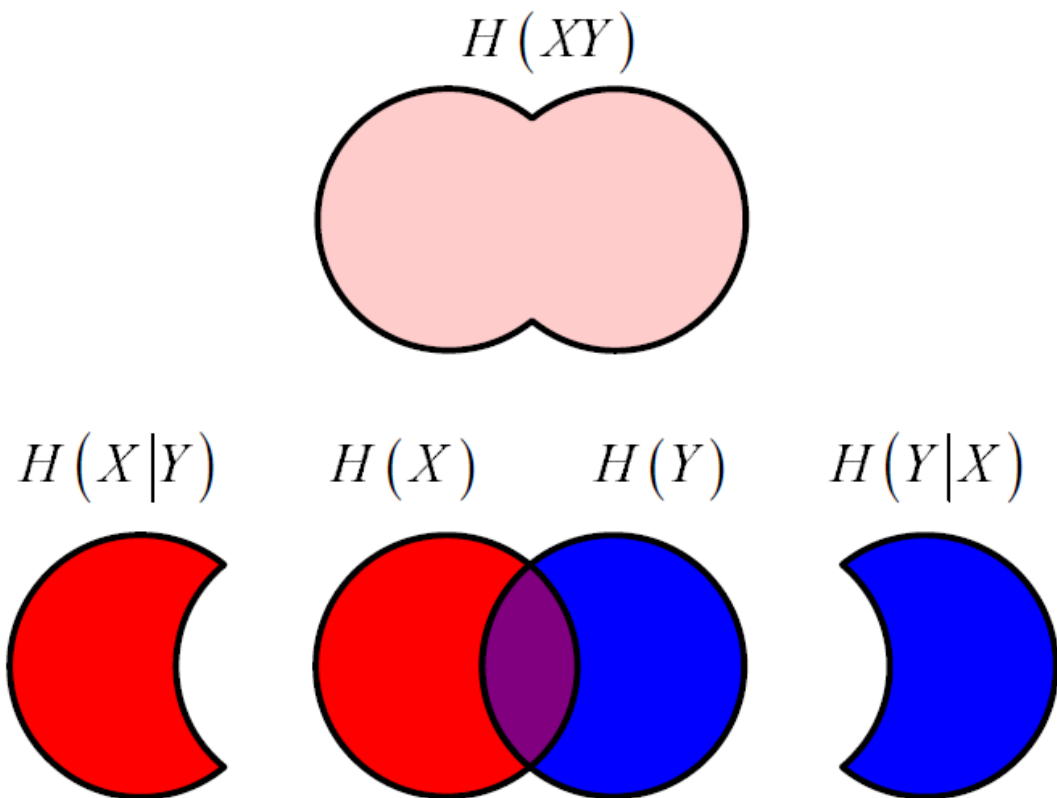
I(X;Y) = I(Y;X)
I(X;Y) ≥ 0
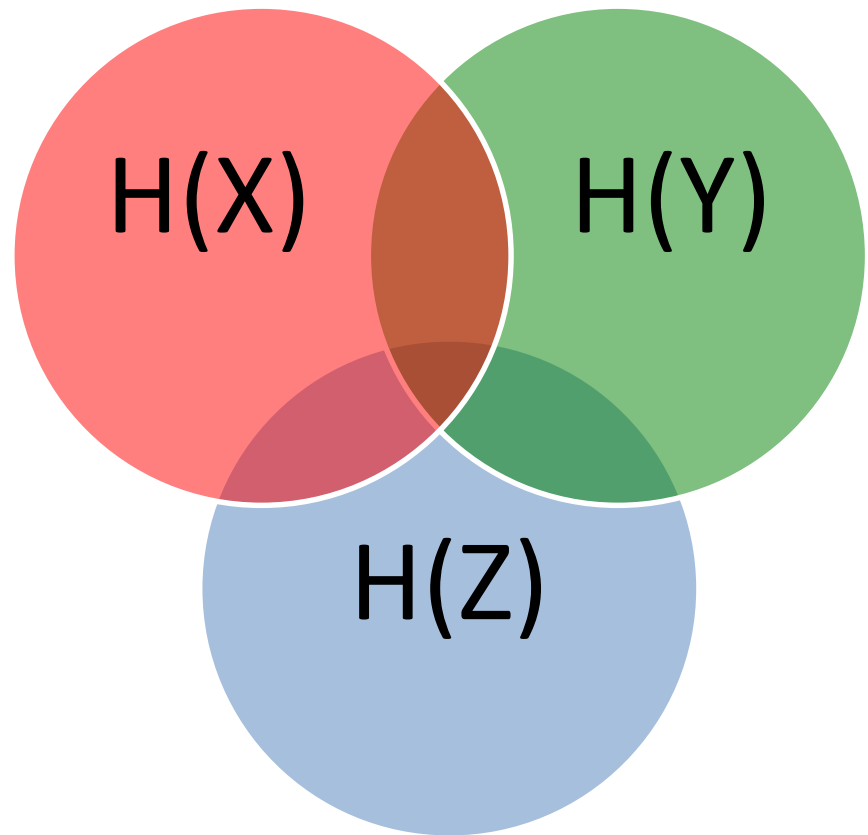I(X;Y) ≤ min[H(X),H(Y)]
I(X;Y) = H(X) − H(X|Y)

$$I(X;Y)$$

$I(X;Y|Z) = I(Y;X|Z)$

$I(X;Y|Z) \geq 0$

$I(X;Y|Z) \leq \min[H(X|Z),H(Y|Z)]$

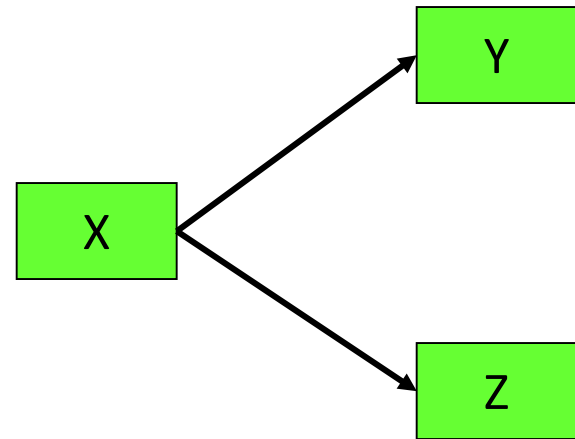$I(X;Y|Z) = H(X|Z) - H(X|YZ)$

Joint Mutual Information

$$I(X;YZ) = I(X;Y) + I(X;Z|Y)$$
$$I(X;YZ) = I(X;Z) + I(X;Y|Z)$$

Joint Mutual Information

Example: Broadcast Network
- Source X
- Receivers Y, Z
- Transmissions can be encrypted or unencrypted
- Separate encryption for Y and Z

I(X;Y) information received at Y (encrypted and unencrypted)
I(X;Z|Y) information received at Z (encrypted)
I(X;Z) information received at Z (encrypted and unencrypted)
I(X;Y|Z) information received at Y (encrypted)
I(X;YZ) = I(X;Y) + I(X;Z|Y) = I(X;Z) + I(X;Y|Z)

Mutual Information
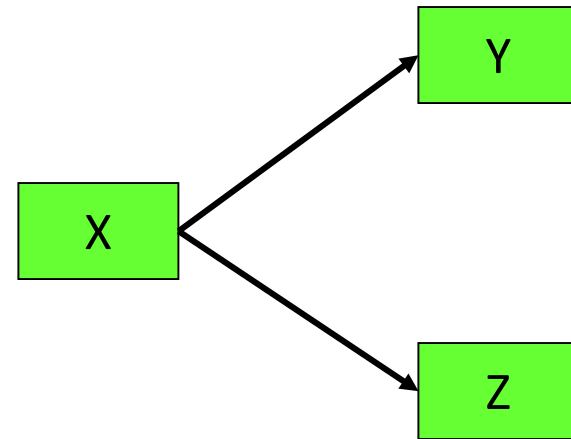
- For two random variables

  I(X;Y) = H(X) − H(X|Y)

- For three random variables

  I(X;Y;Z) = I(X;Y) − I(X;Y|Z)

Joint Mutual Information
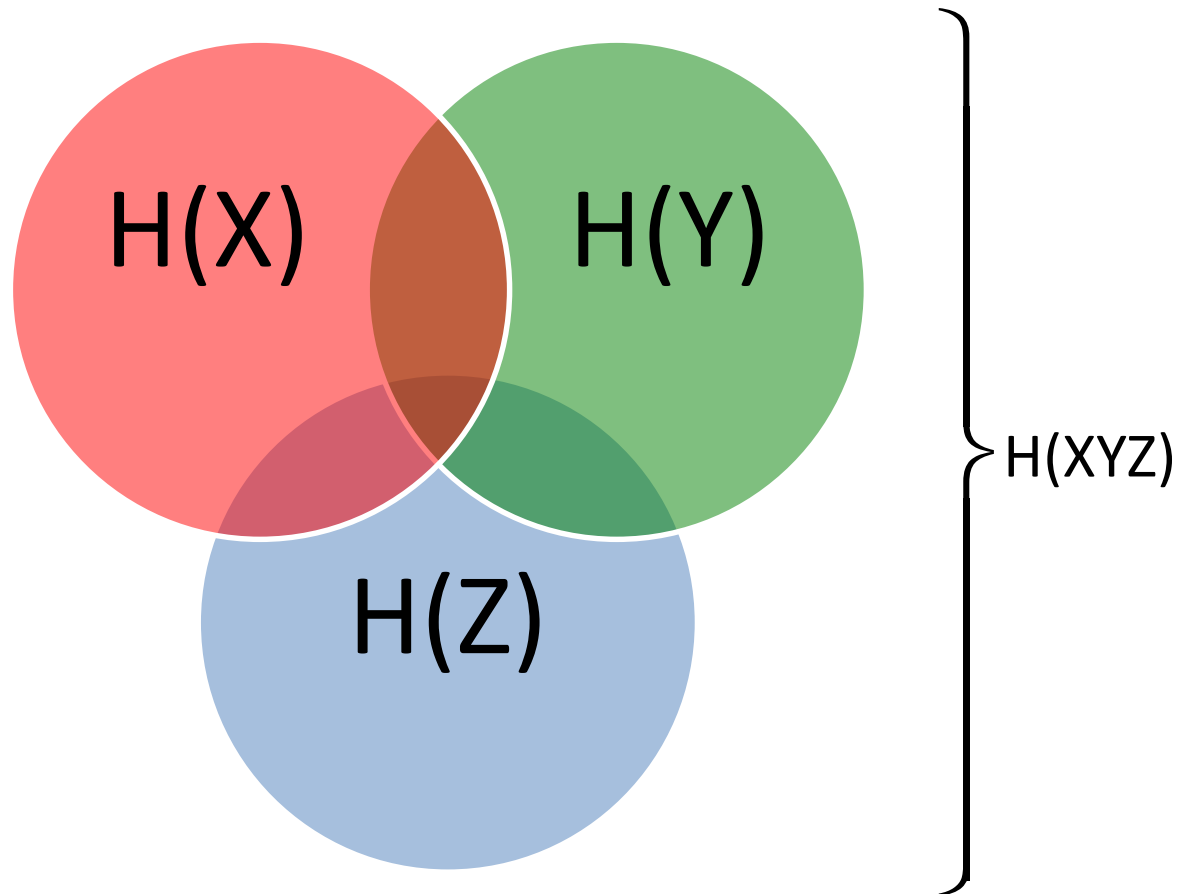
Example: Broadcast Network
- Source X
- Receivers Y, Z
- Transmissions can be encrypted or unencrypted

$I(X;YZ)$
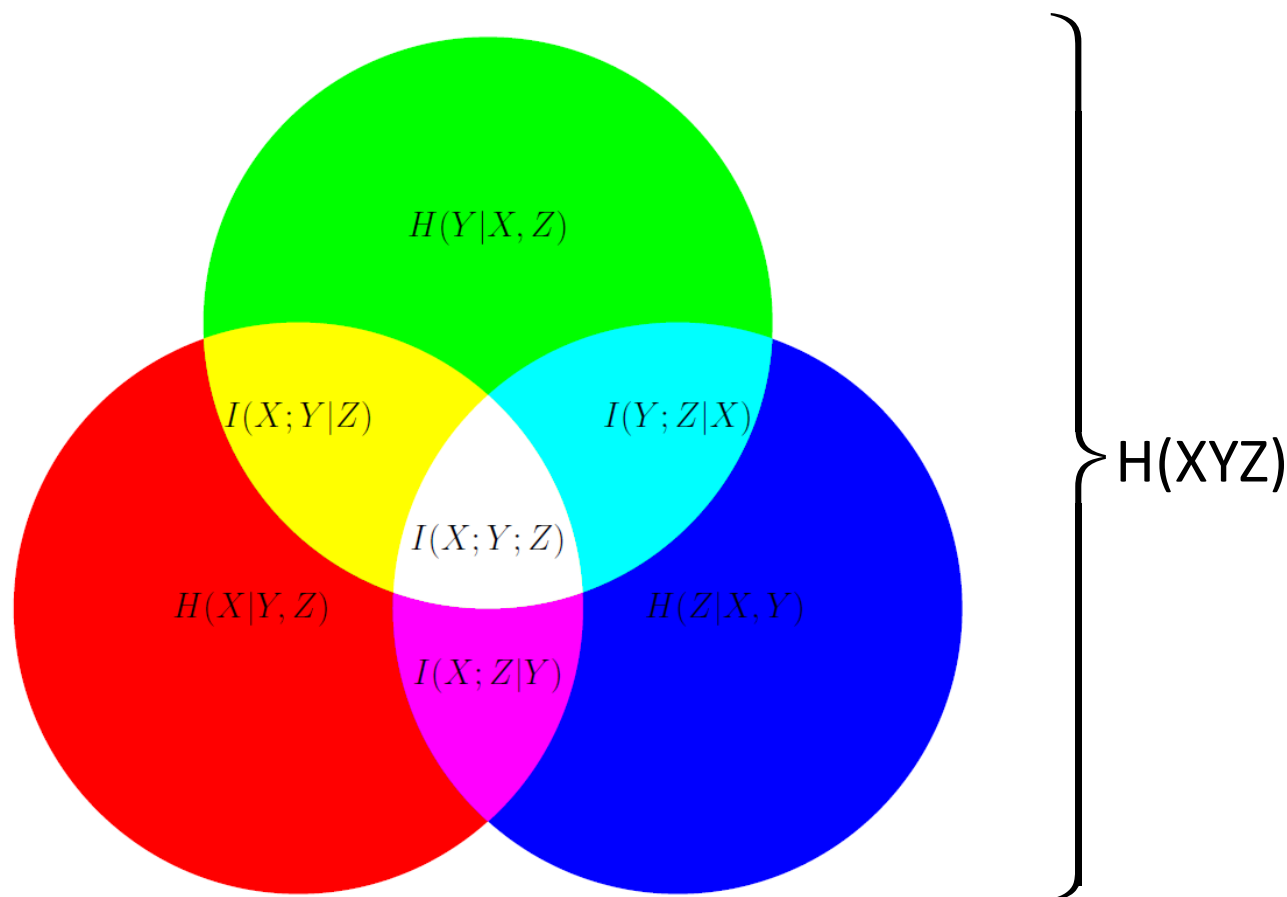$= I(X;Z|Y)$ information received at Z (encrypted)
$+ I(X;Y|Z)$ information received at Y (encrypted)
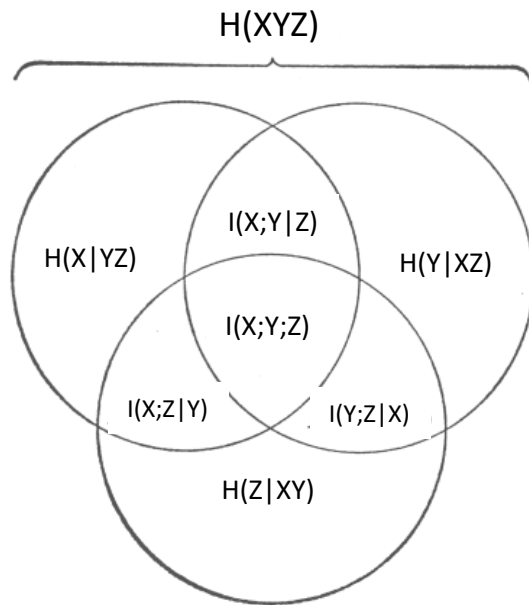$+ I(X;Y;Z)$ information received at Y and Z (unencrypted)

Three Random Variables X, Y and Z

$$I(X;Y;Z) = H(X) + H(Y) + H(Z) - H(XY) - H(XZ) - H(YZ) + H(XYZ)$$

Three Random Variables X, Y and Z

H(XYZ)

H(XYZ)

I(X;Y|Z)

H(X|YZ)   H(Y|XZ)

I(X;Y;Z)

I(X;Z|Y)   I(Y;Z|X)

H(Z|XY)

(a)

H(X)

H(YZ|X)

(b)

H(X|YZ)   H(Y|X)

I(X;YZ)

H(Z|XY)

(c)

H(X|YZ)

H(YZ)

(d)

47

XOR Gate

X ——⊐
Y ——⊐⟩—— Z
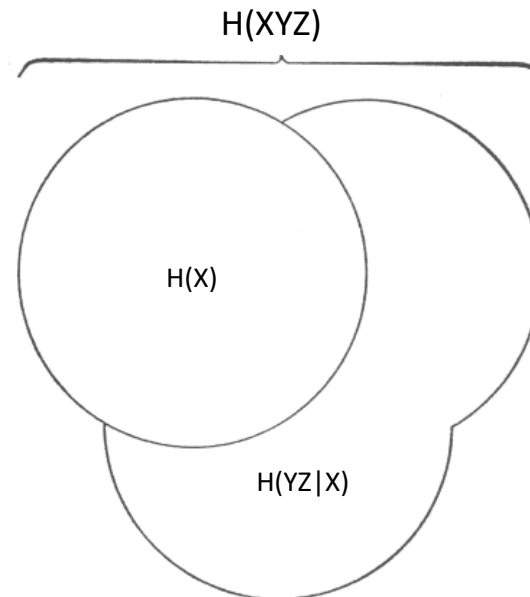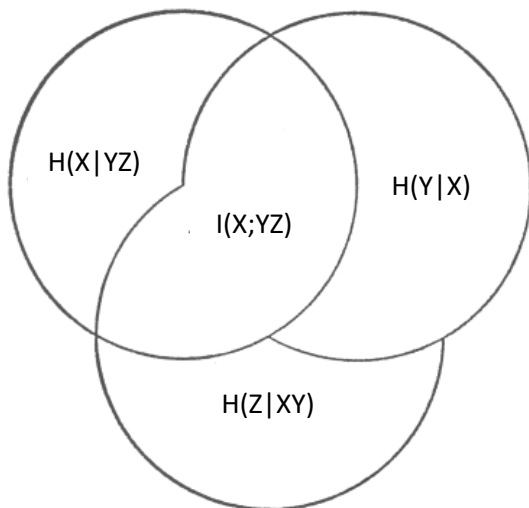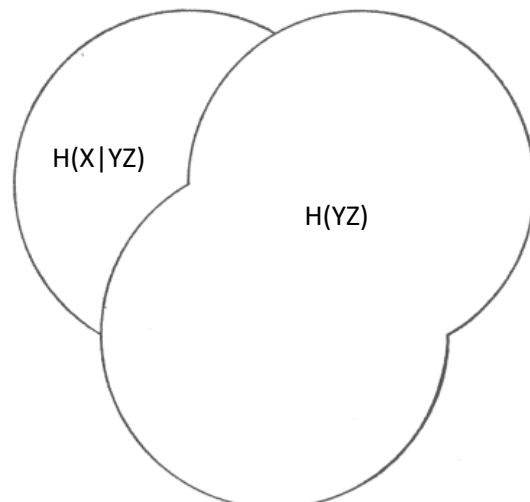
$$z_k = x_i \oplus y_j$$

$$z_k = \begin{cases} 0 & x_i = y_j \\ 1 & x_i \neq y_j \end{cases}$$

x	y	z
0	0	0
0	1	1
1	0	1
1	1	0

$p(x_i) = p(y_j) = 0.5$
X and Y are SI

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z)$$

Probabilities for Three RVs

x_i	y_j	z_k	$p(x_i, y_j, z_k)$	$p(x_i \vert y_j, z_k)$	$p(x_i \vert z_k)$	$p(x_i, y_j)$	$p(x_i)$
0	0	0	1/4	1	1/2	1/4	1/2
0	0	1	0	0	1/2	1/4	1/2
0	1	0	0	0	1/2	1/4	1/2
0	1	1	1/4	1	1/2	1/4	1/2
1	0	0	0	0	1/2	1/4	1/2
1	0	1	1/4	1	1/2	1/4	1/2
1	1	0	1/4	1	1/2	1/4	1/2
1	1	1	0	0	1/2	1/4	1/2

XOR Gate
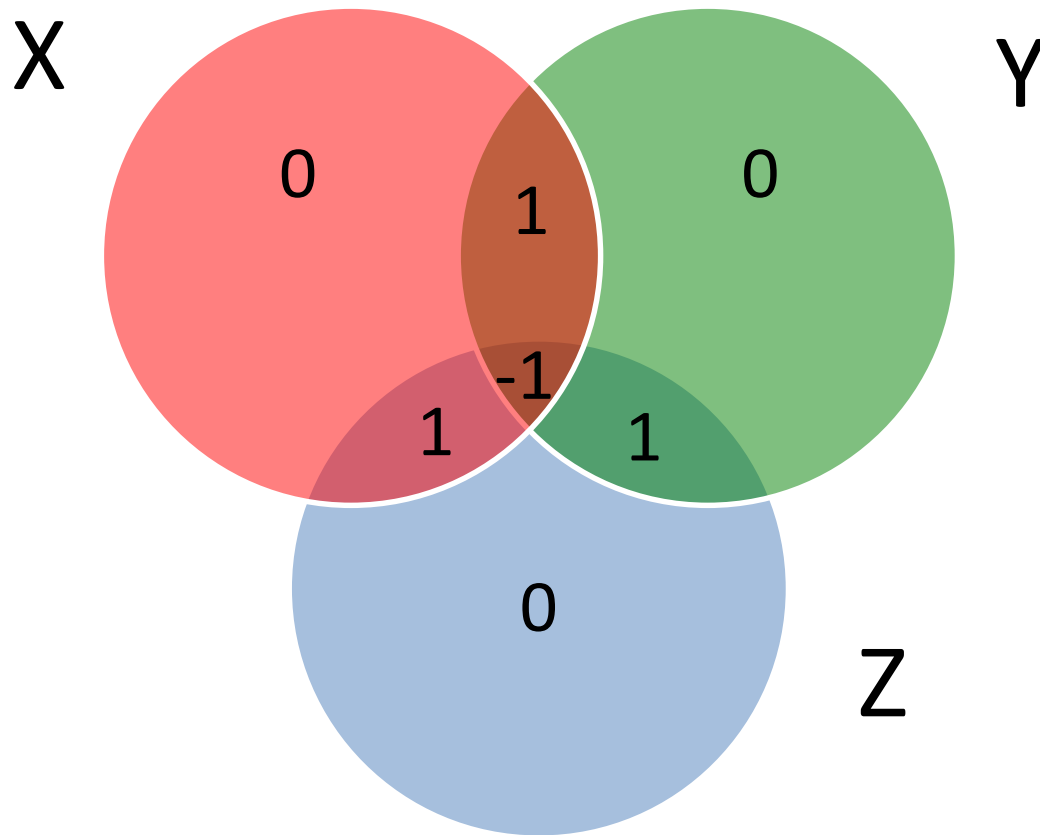
XOR Gate

- $I(X;Y;Z) = I(X;Y) - I(X;Y|Z)$
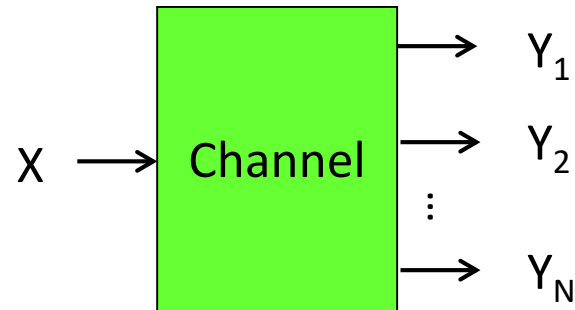- X and Y are independent so $I(X;Y) = 0$
- $I(X;Y|Z) = 1$ bit
- $I(X;Y;Z) = 0 - 1 = -1$ bit

- $I(X;Y;Z)$ is called the Interaction Information

I(X;Y;Z)

- X – rain
- Y – dark
- Z – cloudy

- Which is larger?

  I(rain;dark) or I(rain;dark|cloudy)

Additivity of Mutual Information

Additivity of Mutual Information

$I(X;Y_1Y_2...Y_N) = I(X;Y_1) + I(X;Y_2|Y_1) + I(X;Y_3|Y_1Y_2)$
$+ ... + I(X;Y_N|Y_1Y_2...Y_{N-1})$

$I(X;Y_1Y_2...Y_N) \leq H(X)$

All terms on the RHS ≥ 0

Binary Symmetric Channel

X → [BSC] → Y

crossover probability p

$$\bar{p} = 1 - p$$

X   Transition probabilities   Y

$1 - p$

0 ────────────── 0

p

Transmitted signals

p

Received signals

$1 - p$

1 ────────────── 1

channel matrix

$$\begin{bmatrix} \bar{p} & p \\ p & \bar{p} \end{bmatrix}$$

Figure showing mutual information $I(X;Y)$ versus probability of a "0" at input, ω. The curve peaks at $\omega = \frac{1}{2}$ with value $1 - h(p)$.

Vertical axis (left): Mutual information $I(X;Y)$, marked at 0, $\frac{1}{2}$, $1 - h(p)$, and 1.

Vertical axis (right): marked at $\frac{1}{2}$ and 1.

Horizontal axis: Probability of a "0" at input, ω, marked at 0, $\frac{1}{2}$, and 1.
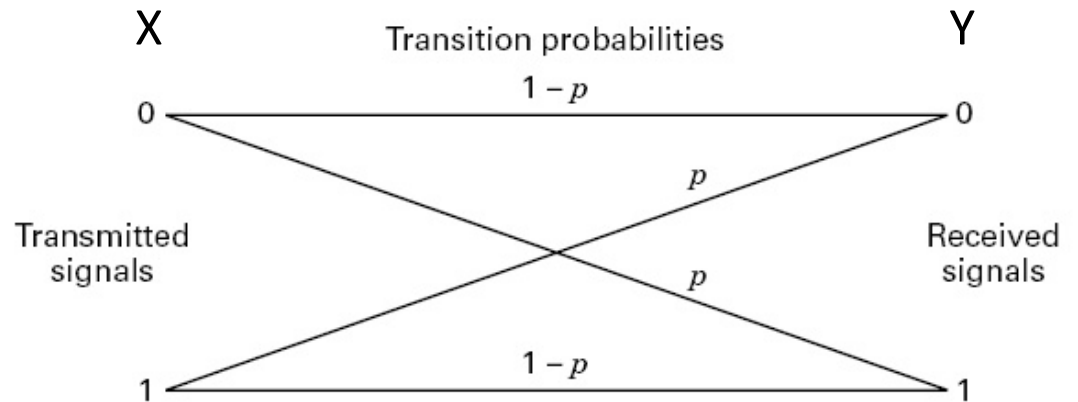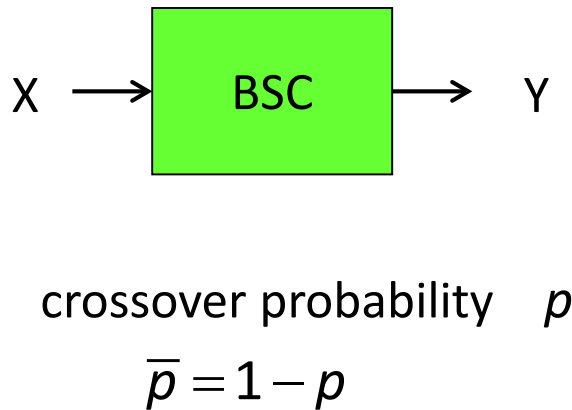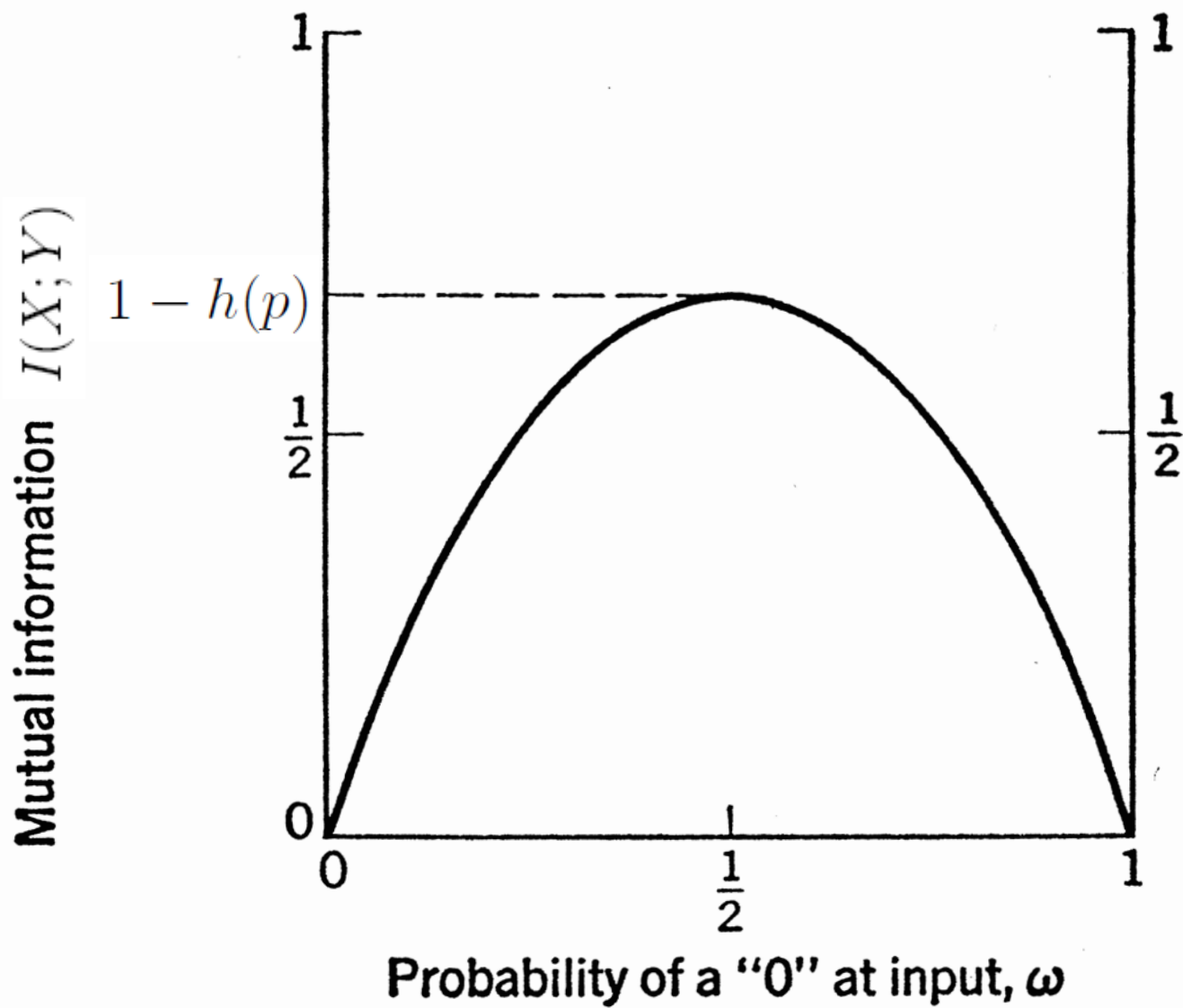
Additivity of Mutual Information

Probabilities for a Repetitive BSC

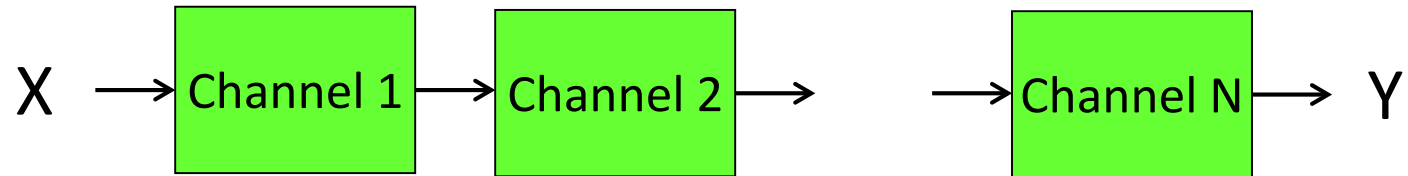| x_i	y_j	z_k	$p(x_i)$	$p(x_i, y_j, z_k)$	$p(y_j, z_k)$
0	0	0	1/2	$1/2(\overline{p}^2)$	$1/2(p^2 + \overline{p}^2)$
0	0	1	1/2	$1/2(p\overline{p})$	$p\overline{p}$
0	1	0	1/2	$1/2(p\overline{p})$	$p\overline{p}$
0	1	1	1/2	$1/2(p^2)$	$1/2(p^2 + \overline{p}^2)$
1	0	0	1/2	$1/2(p^2)$	$1/2(p^2 + \overline{p}^2)$
1	0	1	1/2	$1/2(p\overline{p})$	$p\overline{p}$
1	1	0	1/2	$1/2(p\overline{p})$	$p\overline{p}$
1	1	1	1/2	$1/2(\overline{p}^2)$	$1/2(p^2 + \overline{p}^2)$

Additivity of Mutual Information

$$I(X;Y) \;=\; 1 - h(p)$$

$$I(X;YZ) \;=\; (p^2 + \overline{p}^2)\left[1 - h\left(\frac{p^2}{p^2 + \overline{p}^2}\right)\right]$$

Asymptote

N=3
N=2
N=1

Mutual information

Probability of error *p*

60

Cascaded Channels

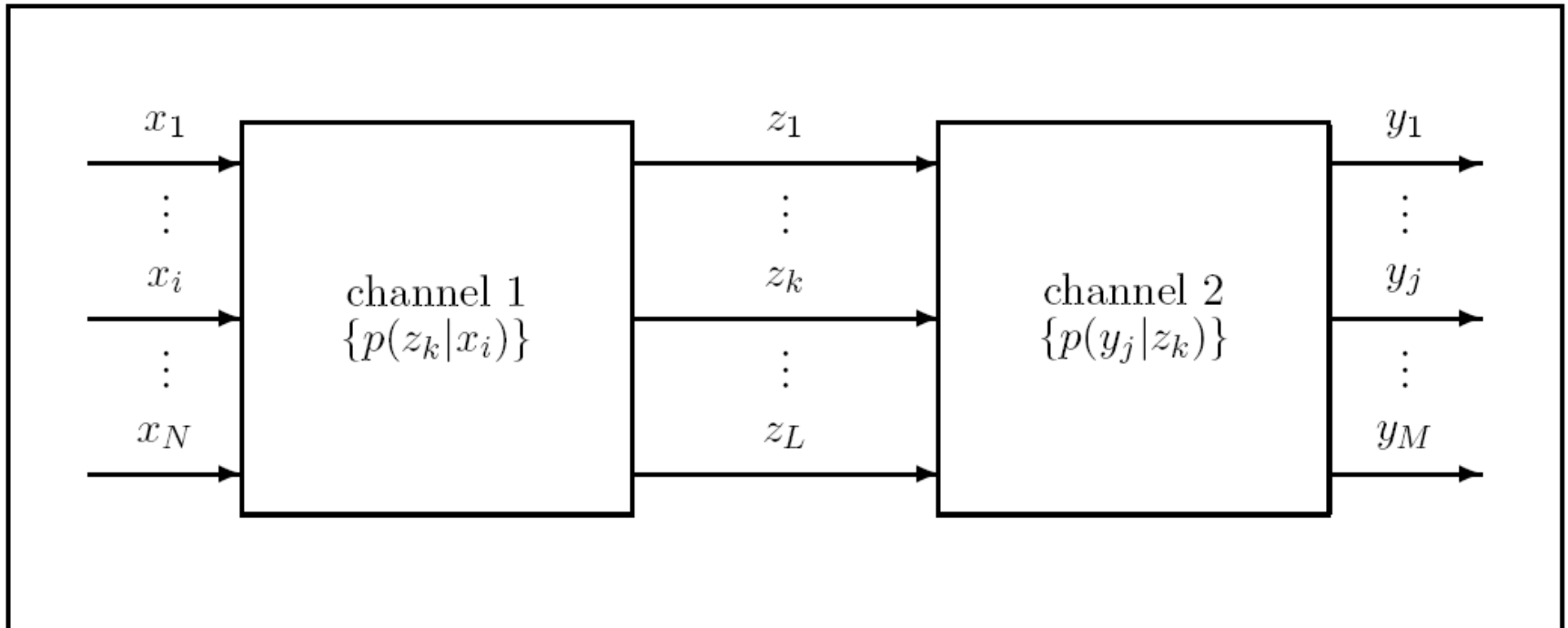X → Channel 1 → Channel 2 → → Channel N → Y
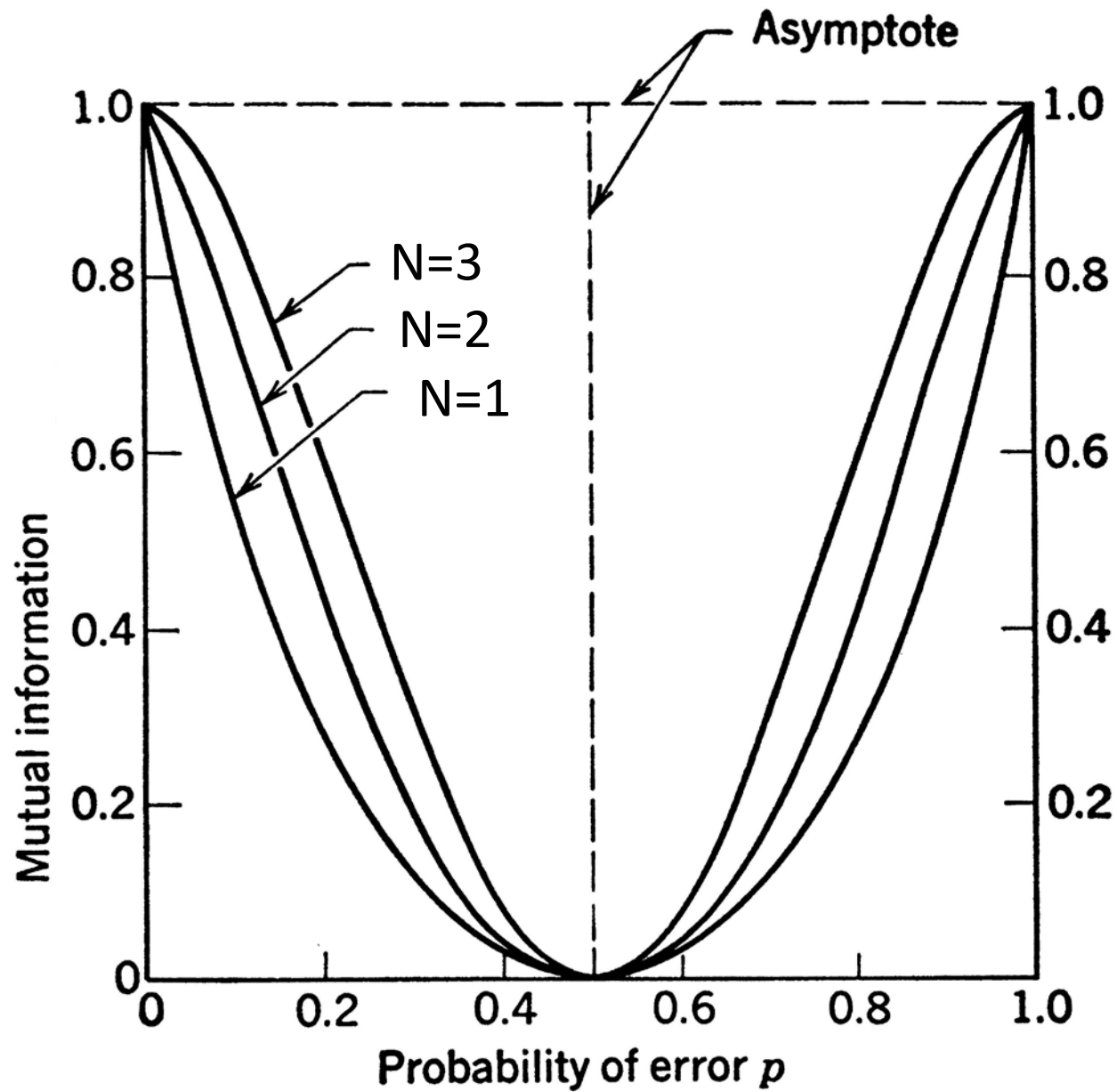
Cascaded Channels

The RVs from X to Y form a

Markov chain

f the conditional distributions of the channel outputs depend only on the immediate inputs and are conditionally independent of the previous RVs

Two Cascaded Channels

$$p(y_j | x_i, z_k) = p(y_j | z_k)$$
$$p(x_i | z_k, y_j) = p(x_i | z_k)$$

Two Cascaded Channels

[Cover and Thomas p. 34]

Three random variables X,Y,Z form a Markov chain, denoted by X → Z → Y, if their joint probability

$$p(x_i,y_j,z_k) = p(x_i)p(z_k|x_i)p(y_j|x_i,z_k)$$

can be factored as

$$p(x_i,y_j,z_k) = p(x_i)p(z_k|x_i)p(y_j|z_k)$$

Two Cascaded Channels

- If X,Y,Z form a Markov chain, then

  $I(X;Y) \leq I(X;Z)$

- To prove this, note that $I(X;Y|Z) = 0$ and

  $I(X;YZ) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y)$

  so that

  $I(X;Z) = I(X;Y) + I(X;Z|Y)$

  or

  $I(X;Y) = I(X;Z) - I(X;Z|Y)$

Two Cascaded Channels

- $I(X;Y|Z) = 0$
- $I(X;Y) \leq I(X;Z)$

$$H(X) - H(X|Y) \leq H(X) - H(X|Z)$$

$$H(X|Y) \geq H(X|Z)$$

- $I(Y;X) \leq I(Y;Z)$

$$H(Y) - H(Y|X) \leq H(Y) - H(Y|Z)$$

$$H(Y|X) \geq H(Y|Z)$$

Data Processing Inequality

- The mutual information between the input and output can never exceed the mutual information between the input and an intermediate point
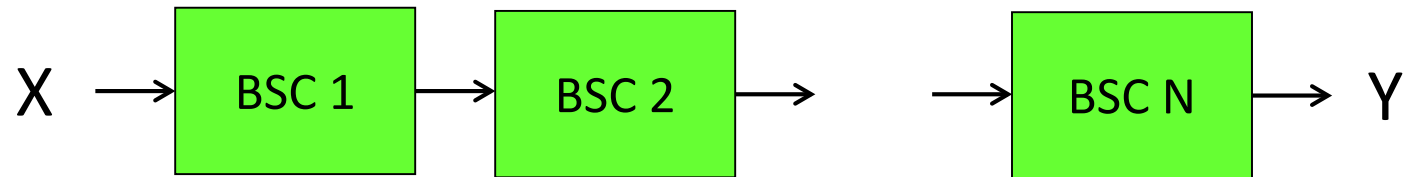
$$I(X;Y) \leq I(X;Z)$$

- The mutual information between the output and input can never exceed the mutual information between the output and an intermediate point

$$I(Y;X) \leq I(Y;Z)$$

- **Data processing cannot increase the amount of information**

Cascaded BSCs

BSC channel matrix

$$\begin{bmatrix} \bar{p} & p \\ p & \bar{p} \end{bmatrix}$$

Cascaded BSCs p=.01

Number of Channels N	Equivalent Crossover Probability	I(X;Y)
1	.01	.919
2	.0198	.860
3	.0294	.809
4	.0388	.763
5	.0480	.722
10	.0915	.559
20	.166	.352
30	.227	.227
40	.277	.149
50	.318	.0978
64	.363	.0549
256	.497	.0000260

Mutual information $I(X;Y)$ versus single channel error probability p, for $N=1$, $N=2$, and $N=3$.
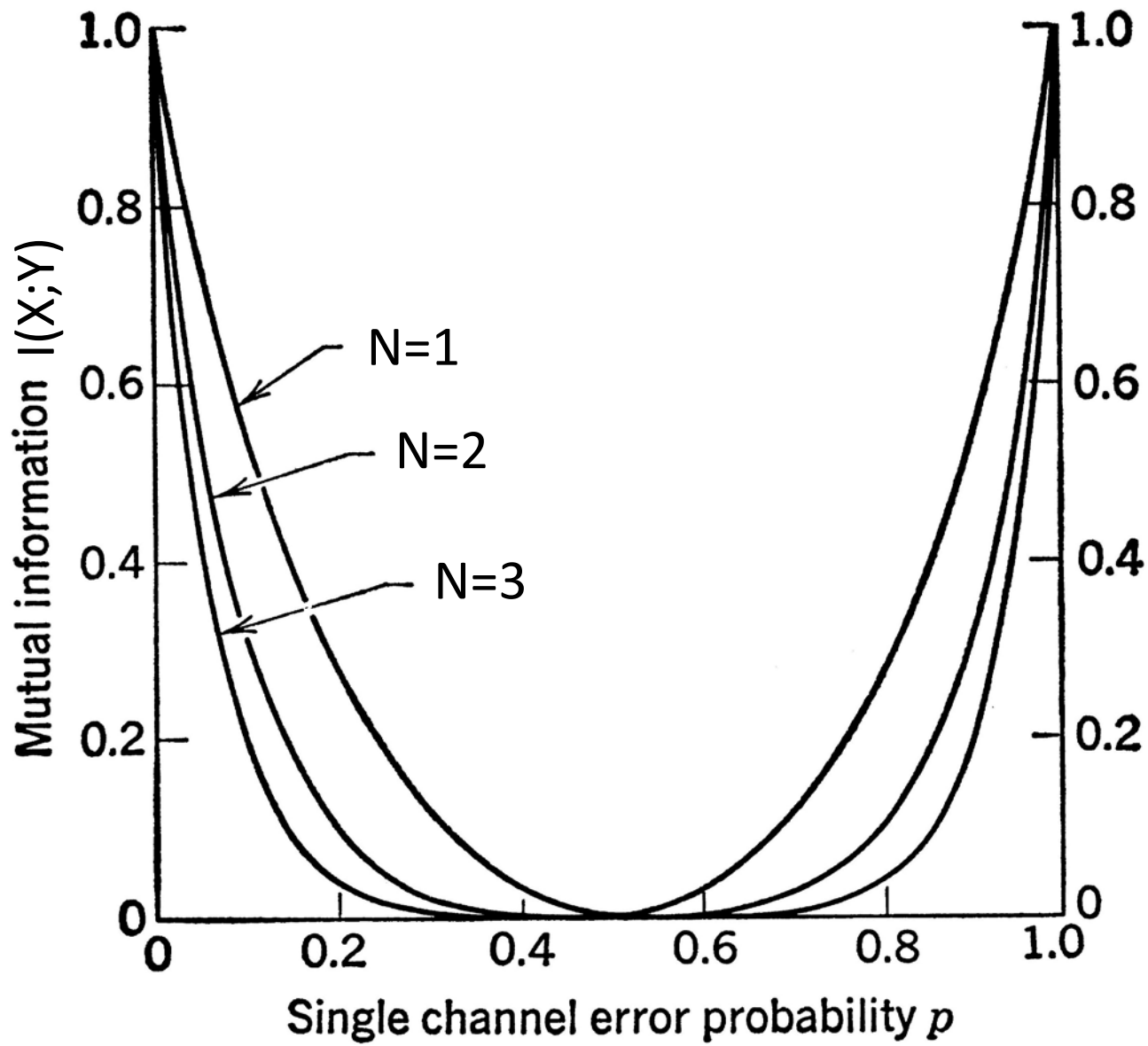
A Mathematical Theory of Communications, BSTJ July, 1948

``The fundamental problem of communication is that of reproducing at one point exactly or approximately a message selected at another point. … If the channel is noisy it is not in general possible to reconstruct the original message or the transmitted signal with certainty by any operation on the received signal.''

A Mathematical Theory of Communications, BSTJ July, 1948

通信的基本问题是，在一个点上的再现准确或约在另一点选择的消息。如果通道是噪声是不一般未能重建原始消息，或确定所传输的信号，由接收到的信号上的任何操作。

Communication is a basic problem in accurate reproduction of a point or another point about the selected message. If the channel is the noise is generally possible to reconstruct the original message, or to determine the transmitted signal from the received signal to any operation.

A Mathematical Theory of Communications, BSTJ July, 1948

التواصل هو المشكلة الأساسية في الاستنساخ الدقيق للنقطة أو نقطة أخرى حول الرسالة المحددة. إذا القناة هو الضجيج عموما ممكن لإعادة بناء الرسالة الأصلية، أو لتحديد الإشارة المرسلة من إشارة وردت إلى أي عملية.

Networking is a fundamental problem in the exact reproduction of one point or another about the selected message. If the channel noise is generally possible to reconstruct the original message, or to determine the transmitted signal from the received signal to any process.

A Mathematical Theory of Communications, BSTJ July, 1948

Networking ist ein grundsätzliches Problem in der exakten Wiedergabe der einen oder anderen Punkt über die ausgewählte Nachricht. Wenn der Kanal Rauschen ist in der Regel möglich, die ursprüngliche Nachricht zu rekonstruieren, um die übertragenen Signals aus dem empfangenen Signal für jeden Prozess zu bestimmen.

Networking is a fundamental problem in the exact reproduction of one point or another over the selected message. If the channel noise is normally possible to reconstruct the original message in order to determine the transmitted signal from the received signal for each process.

A Mathematical Theory of Communications, BSTJ July, 1948

नेटवर्किंग चुने गए संदेश पर एक बिंदु या किसी अन्य की सटीक प्रजनन में एक मूलभूत समस्या है. चैनल शोर प्रत्येक प्रक्रिया के लिए प्राप्त संकेत से संकेत संचारित निर्धारित करने के लिए मूल संदेश को फिर से संगठित करने के लिए सामान्य रूप से संभव है.

Networking at one point or another of the selected message is a fundamental problem in accurate reproduction. Channel noise for each process receives the signal from the transmit signal to determine the message again to organize normally possible.

Entropy

- Let X be a random variable with probability distribution

    $p(X) = \{p(x_i)\}$

- $H(X) = E_p[-\log(p(X)]$

$$H(X) = -\sum_{i=1}^{N} p(x_i) \ \log_b p(x_i)$$

Relative Entropy

- Let X be a random variable with two different probability distributions

    $p(X) = \{p(x_i)\}$

    $q(X) = \{q(x_i)\}$

Relative Entropy

- The **relative entropy** between two probability distributions p(X) and q(X) is defined as the expectation of the logarithm of the ratio of the distributions

$$D[p(X)||q(X)] = E_p[\log(p(X)/q(X))]$$

Relative Entropy

$$D\left[p(X)\|q(X)\right] = \sum_{i=1}^{N} p(x_i)\ \log_b \left[\frac{p(x_i)}{q(x_i)}\right]$$

Relative Entropy

- The relative entropy is a measure of how different two probability distributions p and q (over the same sample space X) are.
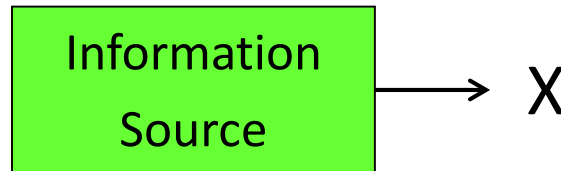
Relative Entropy

- Divergence Inequality

$$D[p(X)||q(X)] \geq 0$$

with equality iff p(X)=q(X)

Relative Entropy
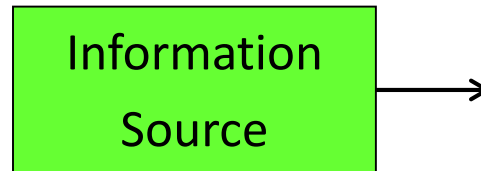

- If nothing is known about the source, the best approximation is a uniform distribution

$$q(x_i) = 1/N$$

- In this case

$$D[p(X)||q(X)] = \log_2 N - H(p(X))$$

Example 1: Four Symbol Source

Information Source

- $p(x_1) = 1/2$ $p(x_2) = 1/4$ $p(x_3) = p(x_4) = 1/8$
- $q(x_1) = q(x_2) = q(x_3) = q(x_4) = 1/4$ (equiprobable)

- $H(p(X)) = 1.75$ bits
- $H(q(X)) = \log_2 N = 2.00$ bits
- $D[p(X)||q(X)] = \log_2 N - H(p(X)) = 0.25$ bit

Example 2: Two Symbol Source

- $p(x_1) = p(x_2) = 1/2$
- $q(x_1) = 1/4$ $q(x_2) = 3/4$

- $D[p(X)||q(X)] = .208$ bits
- $D[q(X)||p(X)] = .188$ bits

D[p(X)||q(X)] versus D[q(X)||p(X)]

- $p(x_i) = 1/N$
- $q(x_1) = \varepsilon$ $q(x_i) = (1-\varepsilon)/(N-1)$ $i \neq 1$
- as $\varepsilon \rightarrow 0$

    $D[p(X)||q(X)] \rightarrow \infty$
    $D[q(X)||p(X)] \rightarrow \log(N/(N-1))$

Cross Entropy

- The **cross entropy** between the probability distributions p(X) and q(X) is defined as

$$H(p,q) = H(p(X))+D(p(X)||q(X))$$

$$H(p,q) = E_p[-\log(q(X)]$$

$$H(p,q) = -\sum_{i=1}^{N} p(x_i) \log q(x_i)$$

Example 3: Four Symbol Source

- $p(x_1) = 1/2$ $p(x_2) = 1/4$ $p(x_3) = p(x_4) = 1/8$
- $q(x_1) = 1/2$ $q(x_2) = q(x_3) = q(x_4) = 1/6$

- $H(p(X)) = 1.75$ bits
- $H(p,q) = 1.7925$ bits
- $D[p(X)||q(X)] = 0.0425$ bits

Cross-Entropy and Iterative Decoding

Michael Moher, *Member, IEEE*, and
T. Aaron Gulliver, *Senior Member, IEEE*

Abstract—In this correspondence, the relationship between iterative decoding and techniques for minimizing cross-entropy is explained. It is shown that minimum cross-entropy (MCE) decoding is an optimal lossless decoding algorithm but its complexity limits its practical implementation. Use of a maximum *a posteriori* (MAP) symbol estimation algorithm instead of the true MCE algorithm provides practical algorithms that are identical to those proposed in the literature. In particular, turbo decoding is shown to be equivalent to an optimal algorithm for iteratively minimizing cross-entropy under an implicit independence assumption.

Cross Entropy Minimization for Efficient Estimation of SRAM Failure Rate

Mohammed Abdul Shahid
Electrical Engineering Department,
University of California, Los Angeles, CA 90095, USA
Email: amohammed@ucla.edu

Abstract—As the semiconductor technology scales down to 45nm and below, process variations have a profound effect on SRAM cells and an urgent need is to develop fast statistical tools which can accurately estimate the extremely small failure probability of SRAM cells. In this paper, we adopt the Importance Sampling (IS) based information theory inspired *Minimum Cross Entropy* method, to propose a general technique to quickly evaluate the failure probability of SRAM cells. In particular, we first mathematically formulate the failure of SRAM cells such that the concept of 'Cross Entropy Distance' can be leveraged, and the distance between the *ideal distribution* for IS and the *practical distribution* for IS (which is used for generating samples), is well-defined. This cross entropy distance is now minimized resulting in a simple analytical solution to obtain the *optimal practical distribution* for IS, thereby expediting the convergence of estimation. The experimental results of a commercial 45nm SRAM cell demonstrate that for the same accuracy, the proposed method yields computational savings on the order of 17~50X over the existing state-of-the-art techniques.

tation of MC requires a prohibitive amount of time (100's of millions, or billions of samples in order to produce a handful of failures) to obtain accurate information. The accuracy of the analytical models and their ability to exactly capture the circuit behaviour is also a matter of concern. Thus, most of the traditional approaches fail to quickly estimate the extreme statistics of rare events of SRAM circuits.

The more recent approaches to improve the sampling efficiency of MC have been based on Importance Sampling (IS) [3]-[5], [7]-[9]. In IS, the original distribution (PDF) is *shifted* towards the rare infeasible failure region, and using this new shifted distribution called *practical distribution* for IS, the failure region is now directly sampled. The non-triviality lies in determining the *optimal shift*, which results in the *optimal practical distribution* for IS, in order to predict quickly and accurately. In [3]-[5], the *optimal practical distribution* for IS,
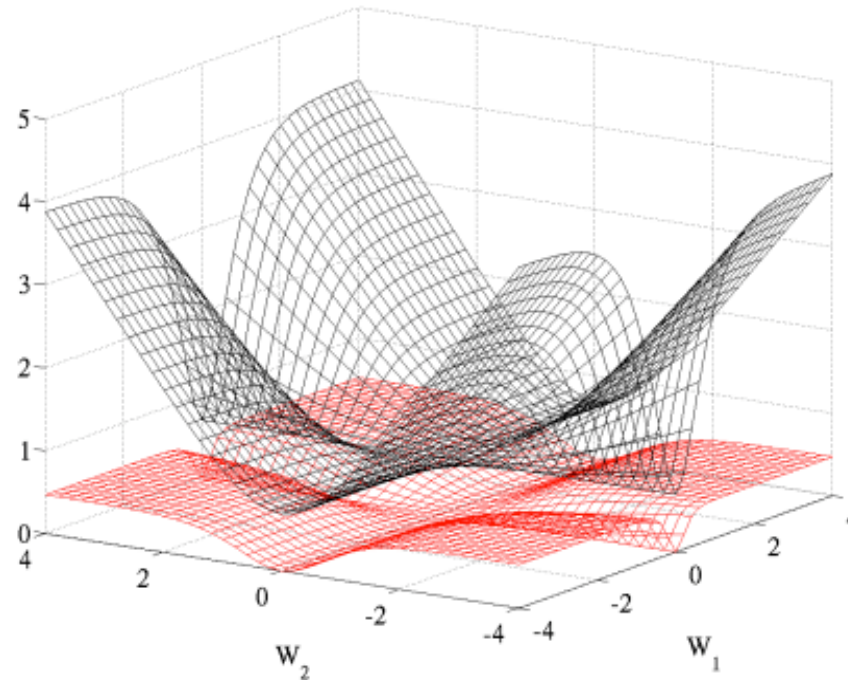
Cross Entropy in ANNs

Figure 5: Cross entropy (black, surface on top) and quadratic (red, bottom surface) cost as a function of two weights (one at each layer) of a network with two layers, W_1 respectively on the first layer and W_2 on the second, output layer.

Mutual Information

$$D\left[p(XY)\|p(X)p(Y)\right] = \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i, y_j) \, \log_b\left[\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right] = I(X;Y)$$

Conditional Relative Entropy

- For joint probability density functions

  $p(XY)$ and $q(XY)$

  the **conditional relative entropy** is

  $$D[p(Y|X)||q(Y|X)]$$

Chain Rule for Relative Entropy

$$D[p(XY)||q(XY)] = D[p(X)||q(X)] + D[p(Y|X)||q(Y|X)]$$

Three Random Variables X, Y and Z

I(X;Y;Z)

I(X;Y|Z)

Y

H(Y|X) = H(Y|XZ)+I(Y;Z|X)

H(X)

X

Z

H(X|YZ)

I(X;Z) = I(X;Y;Z) + I(X;Z|Y)