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{align*}
H(\text{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{align*}
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

\[ p(x_1) = p(x_2) = \frac{1}{2} \]

\[ H(X) = - \frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \text{ bit} \]

\[
\begin{align*}
p(x_i,y_j) & \quad p(y_j|x_j) \\
p(0,0) = \frac{1}{4} & \quad p(0|0) = \frac{1}{2} \\
p(1,1) = 0 & \quad p(1|1) = 0 \quad \text{(empty)} \\
p(0,1) = \frac{1}{4} & \quad p(1|0) = \frac{1}{2} \\
p(1,0) = \frac{1}{2} & \quad p(0|1) = 1 \quad \text{(empty)}
\end{align*}
\]

\[ H(Y|X) = - \frac{1}{4}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{2} - \frac{1}{2}\log_21 \\
= \frac{1}{2}\log_22 = \frac{1}{2} \text{ bit} \]
Example

\[ p(x_i, y_j, z_k) \quad p(z_k | x_j, y_j) \]
\[ p(0,0,0) = \frac{1}{4} \quad p(0 | 0,0) = 1 \]
\[ p(0,1,0) = \frac{1}{4} \quad p(0 | 0,1) = 1 \]
\[ p(1,0,0) = \frac{1}{4} \quad p(0 | 1,0) = \frac{1}{2} \]
\[ p(1,0,1) = \frac{1}{4} \quad p(1 | 1,0) = \frac{1}{2} \]

\[ H(Z | XY) = - \frac{1}{4}\log_2 1 - \frac{1}{4}\log_2 1 - \frac{1}{4}\log_2 \frac{1}{2} - \frac{1}{4}\log_2 \frac{1}{2} \]
\[ = \frac{1}{2}\log_2 2 = \frac{1}{2} \text{ bit} \]
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 $p(y_j|x_i)$
Binary Symmetric Channel

Transition probabilities

Transmitted signals

Received signals

$X_1 \rightarrow Y_1$ with probability $1-p$

$X_2 \rightarrow Y_2$ with probability $1-p$

$X_1 \rightarrow Y_2$ with probability $p$

$X_2 \rightarrow Y_1$ with probability $p$
Binary Symmetric Channel

Transition probabilities

Transmitted signals

$x_1$ 0

Received signals

$y_1$ 0

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

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

$\begin{bmatrix}
  p(y_1 | x_1) & p(y_2 | x_1) \\
  p(y_1 | x_2) & p(y_2 | x_2)
\end{bmatrix} = \begin{bmatrix}
  1-p & p \\
  p & 1-p
\end{bmatrix}$
The probabilities \( p(y_j|x_i) \) are called the **forward transition probabilities**

Using Bayes’ Theorem

\[
p(x_i | y_j) = \frac{p(y_j | x_i)p(x_i)}{p(y_j)}
\]

The probabilities \( p(x_i|y_j) \) are called the **backward transition probabilities**
Non-symmetric Binary Channel

Channel matrix

\[
\begin{pmatrix}
\frac{2}{3} & \frac{1}{3} \\
\frac{1}{10} & \frac{9}{10}
\end{pmatrix}
= \begin{pmatrix}
\frac{2}{3} & \frac{1}{3} \\
\frac{1}{10} & \frac{9}{10}
\end{pmatrix}
\]

\[
\begin{pmatrix}
\frac{1}{4} & \frac{1}{4} \\
\frac{3}{4} & \frac{3}{4}
\end{pmatrix}
\]
### Backward Transition Probabilities

\[
\begin{bmatrix}
p(x_1 | y_1) & p(x_2 | y_1) \\
p(x_1 | y_2) & p(x_2 | y_2)
\end{bmatrix} =
\begin{bmatrix}
p_{x|y}(0|0) & p_{x|y}(1|0) \\
p_{x|y}(0|1) & p_{x|y}(1|1)
\end{bmatrix} =
\begin{bmatrix}
20 & 1 \\
21 & 21 \\
10 & 9 \\
19 & 19
\end{bmatrix}
\]

- \( H(X | y=0) = - p(1|0)\log_2 p(1|0) - p(0|0)\log_2 p(0|0) \)
  \[
  = - \left(\frac{1}{21}\right)\log_2\left(\frac{1}{21}\right) - \left(\frac{20}{21}\right)\log_2\left(\frac{20}{21}\right)
  \]
  \[
  = .209 + .067 = .276 \text{ bit}
  \]

- \( H(X | y=1) = - p(1|1)\log_2 p(1|1) - p(0|1)\log_2 p(0|1) \)
  \[
  = - \left(\frac{9}{19}\right)\log_2\left(\frac{9}{19}\right) - \left(\frac{10}{19}\right)\log_2\left(\frac{10}{19}\right)
  \]
  \[
  = .511 + .487 = .998 \text{ bit}
  \]

- \( H(X | Y) = p(y=0) H(X | y=0) + p(y=0) H(X | y=0) \)
  \[
  = \left(\frac{21}{40}\right)\times(.276) + \left(\frac{19}{40}\right)\times(.998) = .619 \text{ bit}
  \]
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

\[ x_i \rightarrow \text{Channel} \rightarrow y_j \]
Mutual Information

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

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

\[ I(x_i; y_j) = \log_b \frac{p(x_i | y_j)}{p(x_i)} \]
\[ I(y_j; x_i) = I(y_j) - I(y_j|x_i) \]
\[ I(y_j; x_i) = - \log_b p(y_j) - [ - \log_b p(y_j|x_i) ] \]

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

\[ I(x_i; y_j) = I(y_j; x_i) \]
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

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

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

\[ + \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i|y_j) \]

\[ I(X; Y) = H(X) - H(X|Y) \]
\[ H(XY) \]

\[ H(X|Y) \quad H(X) \quad H(Y) \quad H(Y|X) \]

\[ I(X;Y) \]
Four Vectors Example

- \([0,0,0], [0,1,0], [1,0,0], [1,0,1]\)
- \(p(y=1) = .250 \rightarrow H(Y) = .811 \text{ bit}\)
- \(I(X;Y) = H(Y) - H(Y|X) = .811 - .500 = .311 \text{ bit}\)
- \(p(x=1) = .50 \rightarrow H(X) = 1 \text{ bit}\)
- \(I(X;Y) = H(X) - H(X|Y) = 1.0 - .689 = .311 \text{ bit}\)
- \(H(X;Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) = 1.5 \text{ bits}\)
Non-symmetric Binary Channel

Channel matrix

\[
\begin{bmatrix}
    p(y_1 | x_1) & p(y_2 | x_1) \\
    p(y_1 | x_2) & p(y_2 | x_2)
\end{bmatrix} =
\begin{bmatrix}
    \frac{2}{3} & \frac{1}{3} \\
    \frac{3}{10} & \frac{3}{10} \\
    \frac{1}{10} & \frac{9}{10}
\end{bmatrix}
\]
Non-symmetric Channel Example

\[ I(X;Y) = .192 \text{ bit} \]
\[ H(X) = .811 \text{ bit} \]

- \[ H(X|Y) = H(X) - I(X;Y) = .619 \text{ bit} \]

\[ H(Y) = .998 \text{ bit} \]

- \[ H(Y|X) = H(Y) - I(X;Y) = .806 \text{ bit} \]

- \[ H(XY) = H(X) + H(Y|X) \]
\[ = H(Y) + H(X|Y) = 1.617 \text{ bits} \]
Mutual Information for the BSC

channel matrix

\[
\begin{bmatrix}
\overline{p} & p \\
p & \overline{p}
\end{bmatrix}
\]

crossover probability \( p \)

\( \overline{p} = 1 - p \)

\[\begin{aligned}
\mathbb{P}(x = 0) &= w \\
\mathbb{P}(x = 1) &= 1 - w = \overline{w}
\end{aligned}\]

\[
I(X; Y) = H(Y) - H(Y | X)
\]

\[
I(X; Y) = H(X) - H(X | Y)
\]
Mutual Information for the BSC

\[ H(Y|X) = - \sum_{i} \sum_{j} p(x_i, y_j) \log p(y_j|x_i) \]

\[ = - \sum_{i} \sum_{j} p(x_i)p(y_j|x_i) \log p(y_j|x_i) \]

\[ = - \sum_{i} p(x_i) \sum_{j} p(y_j|x_i) \log p(y_j|x_i) \]

\[ = - \sum_{i} p(x_i) [p \log p + \bar{p} \log \bar{p}] \]

\[ = - [p \log p + \bar{p} \log \bar{p}] \]

\[ I(X;Y) = H(Y) - H(Y|X) \]

\[ = H(Y) + [p \log p + \bar{p} \log \bar{p}] \]

\[ = H(Y) - h(p) \]
Binary Entropy Function

\[ h(p) = -p \log_2 p - (1-p) \log_2 (1-p) \quad 0 \leq p \leq 1 \]


\[ p(y = 0) = w\bar{p} + \bar{w}p \]
\[ p(y = 1) = wp + \bar{wp} \]

\[ H(Y) = -[(w\bar{p} + \bar{w}p) \log(w\bar{p} + \bar{w}p) + (wp + \bar{wp}) \log(wp + \bar{wp})] \]
\[ = h(wp + \bar{wp}) \]

\[ I(X; Y) = h(wp + \bar{wp}) - h(p) \]

For \( w = 1/2 \)

\[ I(X; Y) = h[1/2(p) + 1/2(1 - p)] - h(p) \]
\[ = h(1/2) - h(p) \]
\[ = 1 - h(p) \]

For \( w = 0 \) or \( 1 \)

\[ I(X; Y) = h(p) - h(p) \]
\[ = 0 \]
Mutual information $I(X;Y)$

$1 - h(p)$

Probability of a "0" at input, $\omega$
$H(Y) = h(\omega p + \bar{\omega} \bar{p})$

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

$I(X;Y)$
Conditional Mutual Information

\[ p(x_i) \quad \rightarrow \quad \text{Channel} \quad \rightarrow \quad y_j \]

\[ p(x_i | z_k) \quad \rightarrow \quad z_k \]

\[ p(x_i | y_j, z_k) \]
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 \(Y\)
\[ I(X;Y) = I(Y;X) \]
\[ I(X;Y) \geq 0 \]
\[ I(X;Y) \leq \min[H(X), H(Y)] \]
\[ I(X;Y) = H(X) - H(X|Y) \]
\[ I(X;Y|Z) = I(Y;X|Z) \]
\[ I(X;Y|Z) \geq 0 \]
\[ I(X;Y|Z) \leq \min[H(X|Z),H(Y|Z)] \]
\[ I(X;Y|Z) = H(X|Z) - H(X|YZ) \]
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

\[ I(X;YZ) = I(X;Z|Y) \text{ information received at Z (encrypted)} + I(X;Y|Z) \text{ information received at Y (encrypted)} + I(X;Y;Z) \text{ information received at Y and Z (unencrypted)} \]
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) = I(X; Y|Z) + I(Y; Z|X) + I(X; Y|Z) + H(X|Y, Z) + H(Z|X, Y) + H(Y|X, Z)

H(XYZ)
XOR Gate

\[ z_k = x_i \oplus y_j \]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ p(x_i) = p(y_j) = 0.5 \]

X and Y are SI

I(X;Y;Z) = I(X;Y) - I(X;Y|Z)
### Probabilities for Three RVs

| $x_i$ | $y_j$ | $z_k$ | $p(x_i, y_j, z_k)$ | $p(x_i | y_j, z_k)$ | $p(x_i | z_k)$ | $p(x_i, y_j)$ | $p(x_i)$ |
|-------|-------|-------|-------------------|-------------------|----------------|----------------|---------|
| 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     |
| 1     | 1     | 1     | 0                 | 0                 | 1/2            | 1/4            | 1/2     |
XOR Gate

X Y Z
0 1 1
1 0 1
-1 1 0
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

\[ I(X; Y_1 Y_2 \ldots Y_N) = I(X; Y_1) + I(X; Y_2 | Y_1) + I(X; Y_3 | Y_1 Y_2) + \ldots + I(X; Y_N | Y_1 Y_2 \ldots Y_{N-1}) \]

\[ I(X; Y_1 Y_2 \ldots Y_N) \leq H(X) \]

All terms on the RHS \( \geq 0 \)
Binary Symmetric Channel

crossover probability $p$

$\bar{p} = 1 - p$

channel matrix

$$\begin{pmatrix} \bar{p} & p \\ p & \bar{p} \end{pmatrix}$$
Mutual information $I(X; Y)$ vs. Probability of a “0” at input, $\omega$
Additivity of Mutual Information

X → BSC → Y, Z
### Probabilities for a Repetitive BSC

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$y_j$</th>
<th>$z_k$</th>
<th>$p(x_i)$</th>
<th>$p(x_i, y_j, z_k)$</th>
<th>$p(y_j, z_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/2</td>
<td>1/2($\overline{p}^2$)</td>
<td>1/2($p^2 + \overline{p}^2$)</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1/2</td>
<td>1/2($p\overline{p}$)</td>
<td>$p\overline{p}$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1/2</td>
<td>1/2($p\overline{p}$)</td>
<td>$p\overline{p}$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1/2</td>
<td>1/2($p^2$)</td>
<td>1/2($p^2 + \overline{p}^2$)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1/2</td>
<td>1/2($p^2$)</td>
<td>1/2($p^2 + \overline{p}^2$)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1/2</td>
<td>1/2($p\overline{p}$)</td>
<td>$p\overline{p}$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1/2</td>
<td>1/2($p\overline{p}$)</td>
<td>$p\overline{p}$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1/2</td>
<td>1/2($\overline{p}^2$)</td>
<td>1/2($p^2 + \overline{p}^2$)</td>
</tr>
</tbody>
</table>
Additivity of Mutual Information

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

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

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

The RVs from X to Y form a Markov chain if 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 $X \rightarrow Z \rightarrow Y$, 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 $p=0.01$

<table>
<thead>
<tr>
<th>Number of Channels $N$</th>
<th>Equivalent Crossover Probability</th>
<th>$I(X;Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01</td>
<td>0.919</td>
</tr>
<tr>
<td>2</td>
<td>0.0198</td>
<td>0.860</td>
</tr>
<tr>
<td>3</td>
<td>0.0294</td>
<td>0.809</td>
</tr>
<tr>
<td>4</td>
<td>0.0388</td>
<td>0.763</td>
</tr>
<tr>
<td>5</td>
<td>0.0480</td>
<td>0.722</td>
</tr>
<tr>
<td>10</td>
<td>0.0915</td>
<td>0.559</td>
</tr>
<tr>
<td>20</td>
<td>0.166</td>
<td>0.352</td>
</tr>
<tr>
<td>30</td>
<td>0.227</td>
<td>0.227</td>
</tr>
<tr>
<td>40</td>
<td>0.277</td>
<td>0.149</td>
</tr>
<tr>
<td>50</td>
<td>0.318</td>
<td>0.0978</td>
</tr>
<tr>
<td>64</td>
<td>0.363</td>
<td>0.0549</td>
</tr>
<tr>
<td>256</td>
<td>0.497</td>
<td>0.0000260</td>
</tr>
</tbody>
</table>
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

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.
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.
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.
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

\[
D [p(X) \| q(X)] = \sum_{i=1}^{N} p(x_i) \log_b \left[ \frac{p(x_i)}{q(x_i)} \right]
\]
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
  \[ q(x_i) = \frac{1}{N} \]
- In this case
  \[ D[p(X) \parallel q(X)] = \log_2 N - H(p(X)) \]
Example 1: Four Symbol 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
- \( D[p(X) \mid q(X)] = \log_2 N - H(p(X)) = 0.25 \) bit
Example 2: Two Symbol Source

- \( p(x_1) = p(x_2) = \frac{1}{2} \)
- \( q(x_1) = \frac{1}{4} \quad q(x_2) = \frac{3}{4} \)

- \( D[p(X) \mid \mid q(X)] = .208 \text{ bits} \)
- \( D[q(X) \mid \mid p(X)] = .188 \text{ bits} \)
D[p(X) || q(X)] versus D[q(X) || p(X)]

• \( p(x_i) = \frac{1}{N} \)
• \( q(x_1) = \varepsilon \quad q(x_i) = \frac{1-\varepsilon}{(N-1)} \quad i \neq 1 \)
• as \( \varepsilon \to 0 \)
  
  \[ \text{D}[p(X) \| q(X)] \to \infty \]
  \[ \text{D}[q(X) \| p(X)] \to \log(N/(N-1)) \]
Cross Entropy

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

$$H(p,q) = H(p(X)) + D(p(X) \parallel q(X))$$

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

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

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

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

For joint probability density functions $p(XY)$ and $q(XY)$

the conditional relative entropy is

$$D[p(Y|X) \mid \mid q(Y|X)]$$
Chain Rule for Relative Entropy

$$D[p(XY) \mid q(XY)] = D[p(X) \mid q(X)] + D[p(Y \mid X) \mid q(Y \mid X)]$$
Three Random Variables X, Y and Z

\[ I(X;Y|Z) \]

\[ I(X;Y;Z) \]

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

\[ I(X;Z) = I(X;Y;Z) + I(X;Z|Y) \]