Inference & Information Measures for NOISY CHANNELS

- The Capacity
Information theory, pattern recognition, and neural networks

- 1 Noisy-channel coding
- Source coding (Data compression)
  - 2 Information content, entropy
  - 3 Typicality and the source coding theorem
  - 4 Symbol codes
  - 5 Symbol codes and Arithmetic coding
- Noisy-channel coding
  - 6 Inference and Information measures for noisy channels
  - 7 Capacity of a noisy channel
  - 8 The noisy-channel coding theorem
Last time

- inference
- information measures for dependent variables
THREE DOORS

- Gameshow host hides prize behind one door
The rules

- Gameshow host hides prize behind one door
- Player chooses a door e.g. door 1
The rules

- Host opens another door, promising the prize will not be revealed.
The rules

- Host opens another door, promising the prize will not be revealed.
- Player chooses to stick or switch i.e., receive what's behind his original door, or the other closed door.
Player chooses door 1
Host opens door 3, revealing nothing, as promised

A should stick
B should switch
C makes no difference
Z don't know
### Solving Three Doors by Probability Theory

**Bayes's theorem**

\[
P(\mathcal{H} | \text{Data}) = \frac{P(\text{Data} | \mathcal{H}) \ P(\mathcal{H})}{P(\text{Data})}
\]

| HYPOTHESIS | PRIOR \( P(\mathcal{H}) \) | LIKELIHOOD \( P(D=3 | \mathcal{H}) \) | POSTERIOR \( P(\mathcal{H} | D=3) \) |
|------------|-----------------------------|-----------------------------|-----------------------------|
| \( \mathcal{H} = 1 \) | \( \frac{1}{3} \) | \( \frac{1}{2} \) | |
| \( \mathcal{H} = 2 \) | \( \frac{1}{3} \) | \( 1 \) | |
| \( \mathcal{H} = 3 \) | \( \frac{1}{3} \) | \( 0 \) | |
Solving Three Doors by Probability Theory

Bayes's theorem

\[
P(\mathcal{H} | \text{Data}) = \frac{P(\text{Data} | \mathcal{H}) P(\mathcal{H})}{P(\text{Data})}
\]

| HYPOTHESIS | PRIOR \( P(\mathcal{H}) \) | LIKELIHOOD \( P(D=3 | \mathcal{H}) \) | POSTERIOR \( P(\mathcal{H} | D=3) \) |
|------------|-----------------------------|-----------------------------|-----------------------------|
| \( \mathcal{H} = 1 \) | \( \frac{1}{3} \)               | \( \frac{1}{2} \)                      | \( \frac{1}{3} \times \frac{1}{2} / \frac{1}{2} = \frac{1}{3} \) |
| \( \mathcal{H} = 2 \) | \( \frac{1}{3} \)               | \( 1 \)                          | \( \frac{1}{3} \times 1 / \frac{1}{2} = \frac{2}{3} \) |
| \( \mathcal{H} = 3 \) | \( \frac{1}{3} \)               | \( 0 \)                          | \( 0 \)                      |

\[
P(D=3) = \sum_{\mathcal{H}} P(D=3 | \mathcal{H}) P(\mathcal{H})
\]

\[
= \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 1 = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}
\]
Information measures

<table>
<thead>
<tr>
<th>$H(X,Y)$</th>
<th>$H(X)$</th>
<th>$H(Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H(X</td>
<td>Y)$</td>
<td>$I(X;Y)$</td>
</tr>
</tbody>
</table>
Communication over noisy channels

- inference
- information measures
Inference for channels

Channel Q defines conditional probabilities

\[ P(y \mid x) \]

If we choose an input distribution \( P(x) \), we have a joint distribution

\[ P(x, y) = P(x) \, P(y \mid x) \]

with which we can do inference.
Channel $Q$ defines conditional probabilities

$$P(y \mid x)$$

If we choose an input distribution $P(x)$, we have a joint distribution

$$P(x, y) = P(x) P(y \mid x)$$

for which we can compute the mutual information

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

$$= H(Y) - H(Y \mid X)$$
Some channels

BSC

$0 \rightarrow 0$

$\times$

$f$

$1 \rightarrow 1$
\[ Q = y_0 \begin{bmatrix} 1 & 1-f & f \\ 1-f & f & 1 \end{bmatrix} \times 1 \]
Binary erasure channel

\[
\begin{align*}
0 & \overset{1-f}{\rightarrow} 0 \\
0 & \overset{f}{\rightarrow} ? \\
1 & \overset{f}{\rightarrow} ? \\
1 & \overset{1-f}{\rightarrow} 1
\end{align*}
\]
\[ Q = \begin{bmatrix} 1 & -f & 0 \\ f & f & f \\ 0 & 0 & 1 \end{bmatrix} \]
eraswe channel

\[ Q = \begin{bmatrix}
1-f & 0 & 0 \\
0 & f & f \\
0 & 0 & 1-f
\end{bmatrix} \]
Z channel

\[
\begin{align*}
0 & \xrightarrow{1} 0 \\
1 & \xrightarrow{1-f} 1 \\
& \frac{f}{1-f}
\end{align*}
\]

\[
Q = \begin{bmatrix}
1 & f \\
0 & 1-f
\end{bmatrix}
\]
Q_{j|i} = P(y = b_j | x = a_i)
Nasy typewriter
Information measures for noisy channels

Channel \( Q \) defines conditional probabilities

\[ P(y \mid x) \]

If we choose an input distribution \( P(x) \), we have a joint distribution

\[ P(x, y) = P(x) P(y \mid x) \]

for which we can compute the mutual information

\[ I(X; Y) = H(X) - H(X \mid Y) \]
\[ = H(Y) - H(Y \mid X) \]
$f = 0.1$

Pick this input distribution:

$P(x=0) = 0.9$

$P(x=1) = 0.1$
Same channel's

BSC

$\begin{array}{c}
0 \\
\times \\
1
\end{array}$
observe \( y = 1 \)

Given this output, what's the probability that \( x = 1 \)?
\[ \begin{align*}
A & : 0.1 \\
B & : 0.5 \\
C & : 0.9 \\
D & : \text{Something else} \\
Z & : \text{Don't know}
\end{align*} \]

\[ P(x=1 | y) \]
\[ P(x=1 \mid y=1) = \frac{P(y=1 \mid x=1) P(x=1)}{ \sum_{x'} P(y=1 \mid x') P(x')} \]

\[ = \frac{0.9 \times 0.1}{0.9 \times 0.1 + 0.1 \times 0.9} \]

\[ P(y=1) = 0.18 \]
\[ P(x=1 \mid y=0) = \frac{P(y=0 \mid x=1) \cdot P(x=1)}{P(y=0)} \]

\[ = \frac{0.1 \times 0.1}{0.01 + 0.81} = \frac{1}{82} \]

\[ P(y=0) = 0.82 \]
\[
P(x=1 \mid y=1) = \frac{P(y=1 \mid x=1) P(x=1)}{P(y=1)} \sum_{x'} P(y=1 \mid x') P(x')
\]

\[
= \frac{0.9 \times 0.1}{0.9 \times 0.1 + 0.1 \times 0.9} = \frac{1}{2}
\]
$I(X; Y)$

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

\[ = H(Y) - H(Y|X) \]
Input by the output

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

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

\[ H_2(0.18) - H_2(f) \]
\[ H_2(0.18) - H_2(f) \]

\[ = 0.68 - 0.47 \]

\[ = 0.21 \text{ bits} \]
\[ H(x) - H(x|y) \frac{1}{\sum_y P(y) H(x|y)} \]

\[
0.47 - \begin{bmatrix}
0.18 \times H_2\left(\frac{1}{2}\right) + 0.82 \times H_2\left(\frac{1}{82}\right)
\end{bmatrix}
\]

\[
(y=1)
\]

\[
(y=0)
\]

\[
= 0.47 - 0.26
\]

\[
= 0.21 \text{ bits}
\]
Mutual information for the BSC

\[
Q = \begin{bmatrix}
1-f & f \\
 f & 1-f \\
\end{bmatrix}
\]

Consider \( f = 0.1 \).

Assume input distribution \( \mathcal{P}_X = \{p_0, p_1\} = \{0.9, 0.1\} \).

Find
\[
I(X; Y) = H(Y) - H(Y | X) = H_2(0.18) - H_2(0.1) = 0.68 - 0.47 = 0.21 \text{ bits.}
\]

Complete the curve
\[ I(x, y) \]
\[ (x', y') \]
\[ 0.2 \]
\[ 0.4 \]
\[ 1.2 \]
\[ P(x) \]
\[ P(x') \]
\[ (x, y) \]
\[ (x', y') \]
Definition

Capacity of a channel

\[ C(\mathcal{Q}) = \max_P P_x I(X;Y) \]
The **Capacity** of a channel

is the maximum, over all input distributions $P(x)$, of the mutual information:

$$C \equiv \max_{P_x} I(X;Y)$$

The distribution $P_x^*$ that achieves the maximum is called the **optimal input distribution**.
Definition

Capacity of a channel

\[ C(Q) = \max_{P_X} I(X; Y) \]

eg. for BSC with \( p = 0.1 \), the opt. input dist. is \( (\frac{1}{2}, \frac{1}{2}) \)
\[ I(X;Y) = H(Y) - H(Y|X) \]
\[ I(X;Y) = H_2 \left( p_i (1-f) + (1-p_i) \times f \right) \]
\[ C = 1 - H_2(f) \]
The Capacity of a channel

is the maximum, over all input distributions $P(x)$, of the mutual information:

$$C \equiv \max_{P_X} I(X;Y)$$

The distribution $P_X^*$ that achieves the maximum is called the optimal input distribution.

Shannon's noisy-channel coding theorem:

Reliable (virtually error-free) communication is possible at rates up to $C$
\( H_2(x) \) \( ^{1/2} \)

\( 1 \)

\( 0 \)

\( \frac{1}{2} \)

\( \sigma \)

\( C_{BSC} \geq 0.53 \)

\( f = 0.1 \)
What is the optimal input distribution?

\[ \hat{p}_x = \frac{1}{3}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{5} \]

\[ C = 2 \text{ bits} \]

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

\[ = \log_2 4 - 0 \text{ bits} \]

\[ = 2 \text{ bits} \]
Ternary confusion channel

Assume input distribution $P_X = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$.

- If $y = M$, what’s $x$?

- What is the mutual information $I(X; Y)$?
Assume $P_x = \left\{ \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\}$

If $y = M$, what's $x$?

$P(x \mid y = M) = \left\{ \frac{2}{3}, \frac{1}{3}, 0 \right\}$

What's $I(X; Y)$?
$I(X;Y) = H(X) - H(X|Y)$

- $w$: Something else

- A: $\log 3 - H_2(\frac{1}{3})$

- B: $1 - \frac{1}{3} = \frac{2}{3}$

- C: $\log 3 - \frac{1}{3}$

- D: $\log 3 - 1$

- E: $\frac{1}{2} \log 3$

- Z: Don't know
$w$

Something else

A

\[ \log_2 3 - H_2 \left( \frac{1}{3} \right) \]

B

1 - \frac{1}{3} = \frac{2}{3}

C

\[ \log 3 - \frac{1}{3} \]

D

\[ \log 3 - 1 \]

E

\[ \frac{1}{2} \log 3 \]

F

Don't know

G

4

H

0

I

1

J

0

K

6
\( I(Y) = H(Y) - H(Y|X) \)

\[
\frac{1}{3} \times 0 + \frac{1}{3} \times H_2(\frac{1}{2}) = \frac{1}{3} \times 0 + \frac{1}{3} \times 1 = \frac{1}{3}.
\]
Ternary confusion channel

$P^*_x = \begin{pmatrix} 2 & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

& Capacity = 1 bit
Noisy typewriter

\[ P(y = F \mid x = G) = \frac{1}{3}; \]
\[ P(y = G \mid x = G) = \frac{1}{3}; \]
\[ P(y = H \mid x = G) = \frac{1}{3}; \]
\[ \vdots \]

What is the optimal input distribution?

What is the capacity?
Recommended homework

- Reading: Chapters 1-6; Advance reading: Chapters 8, 9, 10
- Arithmetic coding: 6.17 (p125)  6.3 (p118)  6.7 (p123)
- Data compression recap:
  - 5.26 (p103)
  - Invent a question about data compression.
- See handout 2 on website for more recommended exercises
  - Huffman programs huffman.p, huffman.py are on website
  - also a 'bent coin' file 0010000.... as a compression benchmark