ORIE 4742 - Info Theory and Bayesian ML

Chapter 5: Channel Coding

March 1, 2021

Sid Banerjee, ORIE, Cornell
dependent rv and info content
rv $X$ taking values $\mathcal{X} = \{a_1, a_2, \ldots, a_k\}$, with pmf $\mathbb{P}[X = a_i] = p_i$

**Shannon’s entropy function**

- outcome $X = a_i$ has *information content*: $h(a_i) = \log_2 \left( \frac{1}{p_i} \right)$
- random variable $X$ has *entropy*: $H(X) = \mathbb{E}[h(X)] = \sum_{i=1}^{k} p_i \log_2 \left( \frac{1}{p_i} \right)$

- only depends on distribution of $X$ (i.e., $H(X) = H(p_1, p_2, \ldots, p_k)$)
- $H(X) \geq 0$ for all $X$
- if $X \sim$ uniform on $\mathcal{X}$, then $H(X) = \log_2 |\mathcal{X}|$; else, $H(X) \leq \log_2 |\mathcal{X}|$
- if $X \perp \!\!\!\!\perp Y$, then $H(X, Y) = H(X) + H(Y)$

where joint entropy $H(X, Y) \triangleq \sum_{(x,y)} p(x, y) \log_2 1/p(x, y)$
conditional entropy

for any rvs $X$, $Y$:  
\[ H(X|Y) = \sum_{y \in Y} p(y) H(X|Y = y) \]
\[ = \sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 \left( \frac{1}{p(x|y)} \right) \]
the chain rule

the chain rule (information content)

for any rvs $X, Y$ and realizations $x, y$:

$$h(x, y) = h(x) + h(y|x) = h(y) + h(x|y)$$
the chain rule

the chain rule (entropy)
for any rvs $X, Y$:

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$
for any rvs $X, Y$:

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

moreover, given any other conditioning rv $Z$

\[ I(X; Y|Z) = H(X|Z) - H(X|Y, Z) = H(Y|Z) - H(Y|X, Z) \]
<table>
<thead>
<tr>
<th>( P(x, y) )</th>
<th>( x )</th>
<th>( P(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1  2  3  4</td>
<td>1/4</td>
</tr>
<tr>
<td>1</td>
<td>( \frac{1}{8} ) ( \frac{1}{16} ) ( \frac{1}{32} ) ( \frac{1}{32} )</td>
<td>1/4</td>
</tr>
<tr>
<td>2</td>
<td>( \frac{1}{16} ) ( \frac{1}{8} ) ( \frac{1}{32} ) ( \frac{1}{32} )</td>
<td>1/4</td>
</tr>
<tr>
<td>3</td>
<td>( \frac{1}{16} ) ( \frac{1}{16} ) ( \frac{1}{16} ) ( \frac{1}{16} )</td>
<td>1/4</td>
</tr>
<tr>
<td>4</td>
<td>( \frac{1}{4} ) 0 0 0</td>
<td>1/4</td>
</tr>
<tr>
<td>( P(x) )</td>
<td>( \frac{1}{2} ) ( \frac{1}{4} ) ( \frac{1}{8} ) ( \frac{1}{8} )</td>
<td></td>
</tr>
</tbody>
</table>
mutual information and KL divergence

**mutual information**

for any rvs $X$, $Y$:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$
visualizing mutual information

\[ H(X, Y) \]

\[ H(X) \]

\[ H(Y) \]

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

\[ H(X) \]

\[ H(Y) \]

\[ H(X,Y) \]

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

\[ H(X) \]

\[ H(Y) \]

\[ H(X,Y) \]
channel coding
mutual information for the BSC

Binary symmetric channel. $A_X = \{0, 1\}$. $A_Y = \{0, 1\}$.

\[
\begin{align*}
&x \quad 0 \quad 1 \\
y &\quad 0 \quad 1
\end{align*}
\]

$P(y = 0 \mid x = 0) = 1 - f$; $P(y = 0 \mid x = 1) = f$; $P(y = 1 \mid x = 0) = f$; $P(y = 1 \mid x = 1) = 1 - f$.

Assume input distribution $\mathcal{P}_X = \{1 - p, p\}$.
mutual information for the Z-channel

Z channel. \( A_X = \{0, 1\} \). \( A_Y = \{0, 1\} \).

\[
\begin{align*}
0 & \rightarrow 0 \\
1 & \rightarrow 1
\end{align*}
\]

\[
P(y = 0 \mid x = 0) = 1; \quad P(y = 0 \mid x = 1) = f; \\
P(y = 1 \mid x = 0) = 0; \quad P(y = 1 \mid x = 1) = 1 - f.
\]

assume input distribution \( \mathcal{P}_X = \{1 - p, p\} \)
mutual information for the erasure channel

Binary erasure channel. $A_X = \{0, 1\}$. $A_Y = \{0, ?, 1\}$.

\[
\begin{align*}
0 & \quad \Rightarrow \quad 0 \quad \Rightarrow \quad 0 \\
1 & \quad \Rightarrow \quad ? \quad \Rightarrow \quad y \\
1 & \quad \Rightarrow \quad 1
\end{align*}
\]

\[
\begin{align*}
P(y = 0 \mid x = 0) & = 1 - f; & P(y = 0 \mid x = 1) & = 0; \\
P(y = ? \mid x = 0) & = f; & P(y = ? \mid x = 1) & = f; \\
P(y = 1 \mid x = 0) & = 0; & P(y = 1 \mid x = 1) & = 1 - f.
\end{align*}
\]

assume input distribution $\mathcal{P}_X = \{1 - p, p\}$
channel capacity

the capacity of a channel $Q$, with input $A_X$ and output $A_Y$, is

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

any arg max $P_X^*$ is called the optimal input distribution

Shannon's channel coding theorem

can communicate $\leq C$ bits of information per channel use without error!
capacity of the BSC

Binary symmetric channel. $A_X = \{0, 1\}$. $A_Y = \{0, 1\}$.

\[
\begin{align*}
P(y=0 \mid x=0) &= 1 - f; & P(y=0 \mid x=1) &= f; \\
P(y=1 \mid x=0) &= f; & P(y=1 \mid x=1) &= 1 - f.
\end{align*}
\]

assume input distribution $\mathcal{P}_X = \{1 - p, p\}$
capacity of the Z-channel

Z channel. \( A_X = \{0, 1\} \). \( A_Y = \{0, 1\} \).

\[
\begin{array}{c}
0 \leftrightarrow 0 \\
1 \rightarrow 1
\end{array}
\]

\[
P(y=0 \mid x=0) = 1; \quad P(y=0 \mid x=1) = f; \\
P(y=1 \mid x=0) = 0; \quad P(y=1 \mid x=1) = 1 - f.
\]

assume input distribution \( \mathcal{P}_X = \{1 - p, p\} \)
the noisy typewriter

**Noisy typewriter.** $A_X = A_Y = \{A, B, \ldots, Z, \}$: The letters are arranged in a circle, and when the typist attempts to type B, what comes out is either A, B or C, with probability $1/3$ each; when the input is C, the output is B, C or D; and so forth, with the final letter ‘-’ adjacent to the first letter A.

\[
\begin{align*}
P(y = F \mid x = G) &= 1/3; \\
P(y = G \mid x = G) &= 1/3; \\
P(y = H \mid x = G) &= 1/3; \\
&\vdots
\end{align*}
\]
capacity of noisy typewriter
coding with noisy typewriter
another view of the noisy typewriter
expanded channel for the BSC

Binary symmetric channel. $A_X = \{0, 1\}$. $A_Y = \{0, 1\}$.

$$
x \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ \end{pmatrix} y$$

$$
P(y = 0 \mid x = 0) = 1 - f; \quad P(y = 0 \mid x = 1) = f; 
\quad P(y = 1 \mid x = 0) = f; \quad P(y = 1 \mid x = 1) = 1 - f.
$$
expanded channel for the Z-channel

Z channel. $A_X = \{0, 1\}$. $A_Y = \{0, 1\}$.

$$x \begin{array}{c} \rightarrow 0 \\ 1 \rightarrow 1 \end{array} y \quad P(y=0 \mid x=0) = 1; \quad P(y=0 \mid x=1) = f; \quad P(y=1 \mid x=0) = 0; \quad P(y=1 \mid x=1) = 1 - f.$$
lossless compression via \textbf{typical set} encoding

\begin{tcolorbox}
\textbf{typical set}

iid source produces $X^N = (X_1 X_2 \ldots X_N)$; each $X_i \in \mathcal{X}$ has entropy $H(X)$; then $X^N$ is \underline{very likely} to be one of $\approx 2^{NH(X)}$ typical strings, all of which have probability $\approx 2^{NH(X)}$
\end{tcolorbox}
Typical set and non-confusable subset

\[ A_Y^N \]

(a) Typical \( y \) for a given typical \( x \)

(b) Typical \( y \)
typical set and non-confusable subset
### block code

For channel $Q$ with input $A_X$, an $(N, K)$-block code is a list of $S = 2^K$ codewords $\{x^{(1)}, x^{(2)}, \ldots, x^{(2^K)}\}$ with $x^{(i)} \in A_X^N$ (i.e., of length $N$).

### encoder

- Using $(N, K)$-block code, can encode signal $s \in \{1, 2, 3, \ldots, 2^K\}$ as $x(s)$.
- The rate of the code is $R = N/K$ bits per channel use.

### decoder

- Mapping from each length-$N$ string $y \in A_Y^N$ of channel outputs to a codeword label $\hat{s} \in \{\varnothing, 1, 2, 3, \ldots, 2^K\}$ as $x(s)$.
- $\varnothing$ indicates failure.
block codes and capacity

**block code**

for channel $Q$ with input $A_X$, an $(N, K)$-block code is a list of $S = 2^K$ codewords $\{x^{(1)}, x^{(2)}, \ldots, x^{(2^K)}\}$ with $x^{(i)} \in A_X^N$ (i.e., of length $N$)
– the rate of the code is $R = N/K$ bits per channel use

**Shannon's channel coding theorem**

For any $\epsilon > 0$ and $R < C$, for large enough $N$, there exists a block code of length $N$ and rate $\geq R$ such that probability of block error is $< \epsilon$. 

intuition behind proof
example: the erasure channel

Binary erasure channel. $A_X = \{0, 1\}$. $A_y = \{0, ?, 1\}$.

\[
\begin{align*}
&0 \quad \rightarrow \quad 0 \\
&x \quad \rightarrow \quad ? \quad y \\
&1 \quad \rightarrow \quad 1
\end{align*}
\]

\[
\begin{align*}
P(y = 0 \mid x = 0) &= 1 - f; & P(y = 0 \mid x = 1) &= 0; \\
P(y = ? \mid x = 0) &= f; & P(y = ? \mid x = 1) &= f; \\
P(y = 1 \mid x = 0) &= 0; & P(y = 1 \mid x = 1) &= 1 - f.
\end{align*}
\]
erasure channel capacity

Binary erasure channel. $A_X = \{0, 1\}$. $A_Y = \{0, ?, 1\}$.

\[
\begin{align*}
0 & \quad 0 \\
\rightarrow & \quad \searrow \\
x & \quad \uparrow \quad ? \quad y \\
1 & \quad 1
\end{align*}
\]

\[
\begin{align*}
P(y = 0 \mid x = 0) & = 1 - f; \\
P(y = 0 \mid x = 1) & = 0; \\
P(y = ? \mid x = 0) & = f; \\
P(y = ? \mid x = 1) & = f; \\
P(y = 1 \mid x = 0) & = 0; \\
P(y = 1 \mid x = 1) & = 1 - f.
\end{align*}
\]
feedback coding