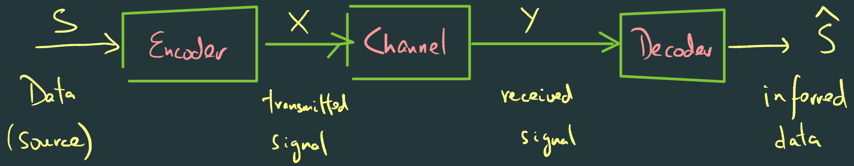
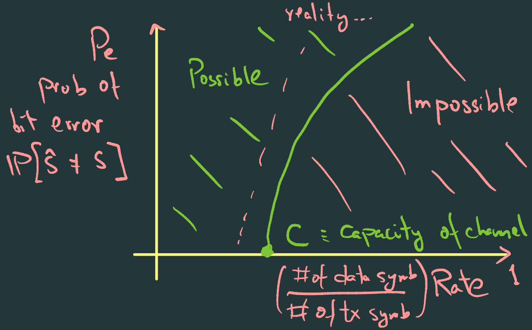


ORIE 4742 - Info Theory and Bayesian ML

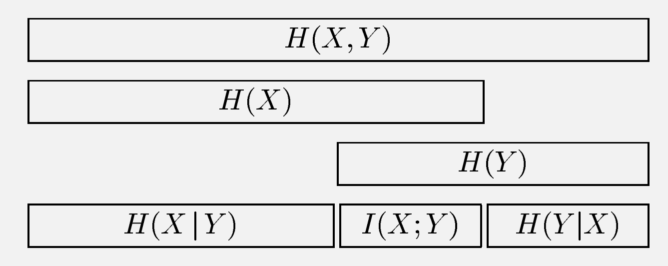
Chapter 5: Channel Coding

March 1, 2021

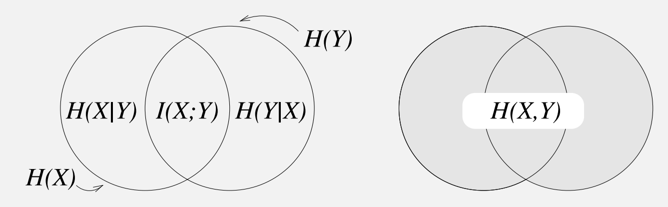
Sid Banerjee, ORIE, Cornell



visualizing mutual information



Correct picture

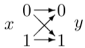


misleading

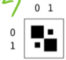
channel coding

mutual information for the BSC

Binary symmetric channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, 1\}$. wp f, bit flip (without loss of generality $f < \frac{1}{2}$)



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



assume input distribution $\mathcal{P}_X = \{1-p, p\}$, ie, Ber$(p)$

Ber(f)

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

$$= \boxed{H_2(q) - H_2(f)}$$

$p\,H_2(f) + (1-p)\,H_2(f)$

$\because$ distn of Y is $(f, 1-f)$ for $X = 0$ and 1

not under our control...

X \ Y	0	1
0	$(1-p)(1-f)$	$(1-p)f$
1	pf	$p(1-f)$

$q = p(1-f) + (1-p)f$

$Y \sim$ Ber(q)

Alt

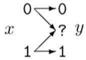
$$I(X;Y) = H(X) - H(X|Y)$$

$$= H_2(p) - q\,H(X|Y=1) - (1-q)\,H(X|Y=0)$$

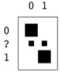
need Bayes thm

mutual information for the erasure channel (good model for compression)

Binary erasure channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, ?, 1\}$.



$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= 0; \\ P(y=? \mid x=0) &= f; & P(y=? \mid x=1) &= f; \\ P(y=1 \mid x=0) &= 0; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$



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

$$I(X;Y) = \underbrace{H(X)}_{H_2(p)} - \underbrace{H(X|Y)}_{\text{messy}\ldots}$$

$$H(Y) = \underbrace{H(\text{'erasure'})}_{H_2(f)} + \underbrace{H(Y|\text{'erasure'})}_{f\cdot 0 + (1-f)H_2(p)}$$

$f H_2(Y|\text{erasure}) + (1-f) H_2(Y|\text{non erasure})$

$$I(X;Y) = \underbrace{H(Y)}_{H_2(f) + \boxed{(1-f)H_2(p)}} - \underbrace{H(Y|X)}_{H_2(f)}$$

$$Y = \begin{cases} 0 : & p(1-f) \\ 1 : & (1-p)(1-f) \\ ? : & f \end{cases}$$

$$\Rightarrow H(Y) = (1-f)\left(p \log_2\left(\frac{1}{p(1-f)}\right) + (1-p)\log_2\left(\frac{1}{(1-p)(1-f)}\right)\right) + f \log_2 \frac{1}{f}$$

$$= H_2(f) + (1-f)H_2(p)$$

capacity of a channel

channel capacity

the capacity of a channel $\mathcal{Q}$, with input $\mathcal{A}_X$ and output $\mathcal{A}_Y$, is

received signal

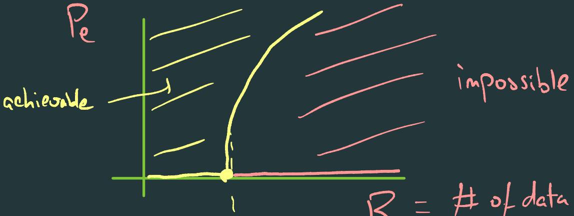
$$C(\mathcal{Q}) = \underbrace{\max_{\mathcal{P}_X}}_{\text{encoding}} I(X;Y)$$

tx signal

any $\arg\max \mathcal{P}_X^\star$ is called the optimal input distribution

Shannon's channel coding theorem

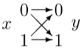
can communicate $\leq C$ bits of information per channel use without error!



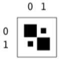
$$R = \frac{\#\text{ of data symbols}}{\#\text{ of transmitted symbols}}$$

capacity of the BSC

Binary symmetric channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, 1\}$.

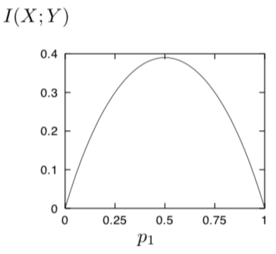


$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= f; \\ P(y=1 \mid x=0) &= f; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$



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

$I(X;Y) = H_2(q) - H_2(f), \quad q = p(1-f) + (1-p)f$



max $I(X;Y)$ over p $\iff$ max $H_2(q)$ over p

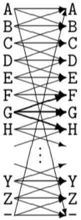
- $H_2(q) \leq 1, \; = 1$ if $q = \frac{1}{2}$

$\Rightarrow$ we need $p(1-f) + (1-p)f = \frac{1}{2} \Rightarrow p = \frac{1}{2}$

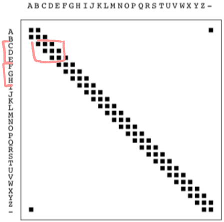
$$\Rightarrow \quad C_{BSC} = \underbrace{1 - H_2(f)}_{\substack{\geq 0 \\ = 0 \text{ if } f = 1/2}} \quad \text{for } \mathcal{P}_X^* = \{\tfrac{1}{2}, \tfrac{1}{2}\}$$

the noisy typewriter

Noisy typewriter. $\mathcal{A}_X = \mathcal{A}_Y =$ the 27 letters $\{$`A`, `B`, …, `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`.



$$\begin{aligned} &\vdots \\ P(y = \texttt{F} \mid x = \texttt{G}) &= 1/3; \\ P(y = \texttt{G} \mid x = \texttt{G}) &= 1/3; \\ P(y = \texttt{H} \mid x = \texttt{G}) &= 1/3; \\ &\vdots \end{aligned}$$



$$I(X;Y) = H(X) - H(X|Y) = \underbrace{H(Y)}_{\leq \log_2 27} - \underbrace{H(Y|X)}_{\log_2 3}$$

$Y \in \{a, b, \ldots, z, \_\} \ni A_Y = 27$

$$\leq \log_2 9$$

(and also, $I(X;Y) = \log_2 9$ if X is uniform over A_X)

capacity of noisy typewriter

$$\Rightarrow \quad C_{NT} = \log_2 9$$

(i.e, can send 9 symbols without error per channel use)

coding with noisy typewriter

Idea - 'Use every 3rd letter'

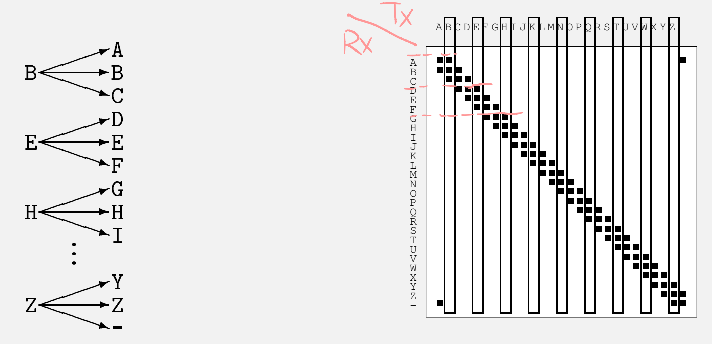
encoder $1 \to A, 2 \to D, 3 \to G, \dots, 9 \to Y$

decoder
$$\left.\underbrace{\begin{matrix} (\_, A, B) \to 1 \\ (C, D, E) \to 2 \\ \vdots \\ (X, Y, Z) \to 9 \end{matrix}}_{\text{Syndrome decoding}}\right]$$

no error!

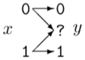
per channel use sends 9 symbols

another view of the noisy typewriter

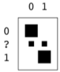


example: the erasure channel

Binary erasure channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, ?, 1\}$. , $P_X = \{1-p, p\}$



$$
\begin{aligned}
P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= 0; \\
P(y=? \mid x=0) &= f; & P(y=? \mid x=1) &= f; \\
P(y=1 \mid x=0) &= 0; & P(y=1 \mid x=1) &= 1-f.
\end{aligned}
$$



$$I(X;Y) = (1-f)H_2(p)$$

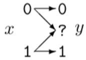
$$\Rightarrow \quad C_{BEC} = \max_p I(X;Y) = 1-f \quad \text{for} \quad P_X^* = \{1/2, 1/2\}$$

$\Rightarrow$ should try and encode data such that each encoded bit is 0 or 1 with prob 1/2 (i.e., optimal code!)

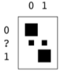
- This channel has a simple 'achievable' scheme (assuming we have perfect feedback) ← not necessary

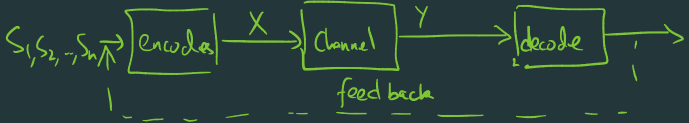
erasure channel capacity with perfect feed back

Binary erasure channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, ?, 1\}$.



$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= 0; \\ P(y=? \mid x=0) &= f; & P(y=? \mid x=1) &= f; \\ P(y=1 \mid x=0) &= 0; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$





S_1 0 ⟶ 0

S_2 1 ⟶ 1

S_3 $\cancel{\longrightarrow}$?

⟵ NACK = send it again

S_3 $\cancel{\longrightarrow}$?

S_3 1 ✓⟶ 1

Q: how many times do I send each symbol?

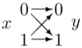
- A: Geom(1-f)

$\Rightarrow \mathbb{E}[\#\text{ of channel uses}] = \frac{1}{1-f}$

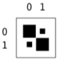
$\Rightarrow$ Rate $= 1-f$

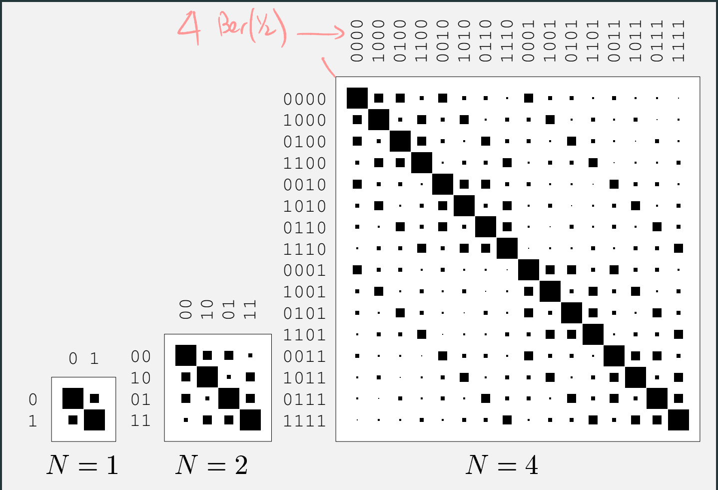
expanded channel for the BSC

Binary symmetric channel. $\mathcal{A}_X = \{0,1\}$. $\mathcal{A}_Y = \{0,1\}$.



$$
\begin{aligned}
P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= f; \\
P(y=1 \mid x=0) &= f; & P(y=1 \mid x=1) &= 1-f.
\end{aligned}
$$

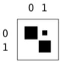


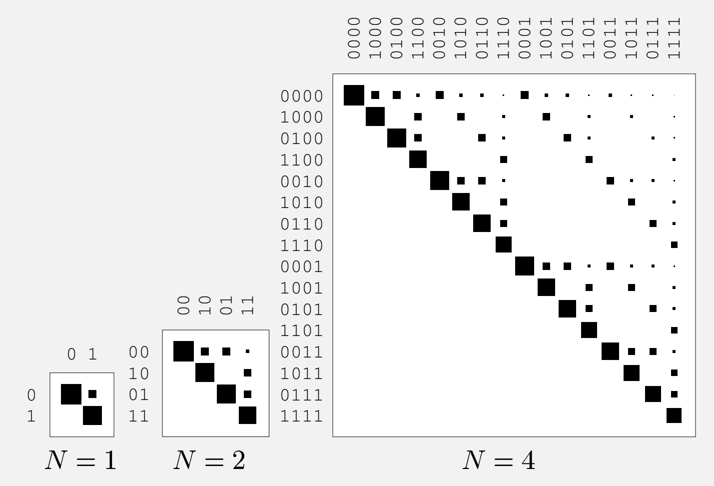


expanded channel for the Z-channel

Z channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, 1\}$.

$$x \begin{array}{c} 0 \longrightarrow 0 \\ \nearrow \\ 1 \longrightarrow 1 \end{array} y \qquad \begin{array}{rclrcl} P(y=0 \mid x=0) & = & 1; & P(y=0 \mid x=1) & = & f; \\ P(y=1 \mid x=0) & = & 0; & P(y=1 \mid x=1) & = & 1-f. \end{array}$$





lossless compression via typical set encoding

typical set

iid source produces $X^N = (X_1 X_2 \dots X_N)$; each $X_i \in \mathcal{X}$ has entropy $H(X)$

then X^N is very likely to be one of $\approx 2^{NH(X)}$ typical strings,

all of which have probability $\approx 2^{NH(X)}$

Eg - If $X_i \sim \text{Ber}(1/2)$

$np \mp 2\sqrt{n}\, p(1-p)$

$(X_1 X_2 \dots X_{100}) \in$ (All sequences $\{0,1\}^{100}$ with between 40 and 60 '1's)

with prob $\geq 1-\delta$

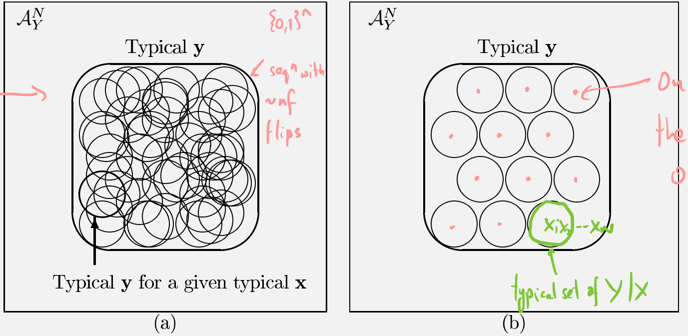
much smaller than $\{0,1\}^{100}$

typical set and non-confusable subset

(see ch 9 of Mackay)

used a channel n times

Eg - In BSC, this could be the sequence of flips



(a) (b)

Only send these values of X

'Volume of typical set of Y' $\approx 2^{nH(Y)}$ 'Volume of typical set of $Y|X$' $\approx 2^{nH(Y|X)}$

by volume cons, can 'pack' $2^{nH(Y)} / 2^{nH(Y|X)}$ 'spheres' in the typical set of Y $\Rightarrow$ can send $n(H(Y) - H(Y|X))$ symbols in n channel uses

typical set and non-confusable subset

This is the first example of 'the probabilistic method'

block codes, encoding, decoding

block code

for channel $\mathcal{Q}$ with input $\mathcal{A}_X$, an (N, K)-block code is a list of $\mathcal{S} = 2^K$ codewords $\{x^{(1)}, x^{(2)}, \ldots, x^{(2^K)}\}$ with $x^{(i)} \in \mathcal{A}_X^N$ (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 $y \in \mathcal{A}_Y^N$ of channel outputs to a codeword label $\widehat{s} \in \{\varphi, 1, 2, 3, \ldots, 2^K\}$ as $x(s)$
– φ indicates failure

block codes and capacity

block code

for channel $\mathcal{Q}$ with input $\mathcal{A}_X$, an (N, K)-block code is a list of $\mathcal{S} = 2^K$ codewords $\{x^{(1)}, x^{(2)}, \ldots, x^{(2^K)}\}$ with $x^{(i)} \in \mathcal{A}_X^N$ (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