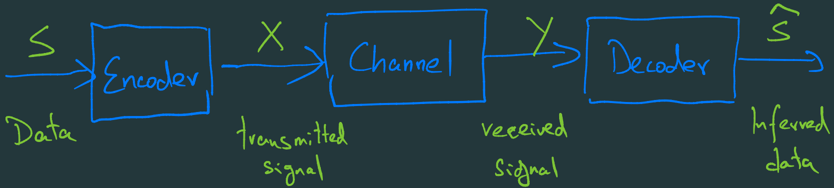


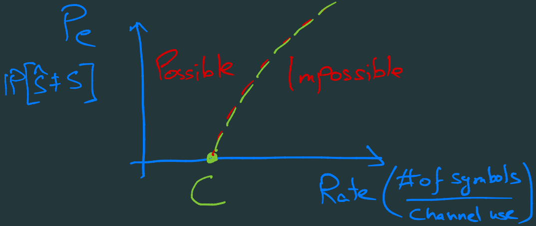
ORIE 4742 - Info Theory and Bayesian ML

Chapter 5: Channel Coding

Februaru 11, 2020

Sid Banerjee, ORIE, Cornell

Want - $\mathbb{P}[\hat{S} \neq S] \searrow 0$



dependent rv and information content

Chapter 8 from Mackay

entropy: basic properties

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

mutual information

mutual information

for any rvs X, Y:

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

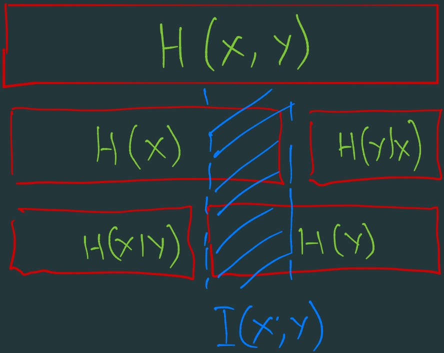
moreover, given any other conditioning rv Z

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

$$D_{KL}(P\|Q)$$

$$= \sum p(x) \log \frac{p(x)}{q(x)}$$

$$= 0 \text{ if } P = Q$$



$$I(x;y) =$$

$$D_{KL}\big(P(x)P(y) \,\|\, P(x,y)\big)$$

conditional entropy

conditional entropy

for any rvs X, Y: $H(X|Y) = \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y)$
$$= \sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log_2(1/p(x|y))$$

(Joint) $H(X,Y) = \sum_{(x,y)} p(x,y) \log_2\left(1/p(x,y)\right) = \sum_{(x,y)} p(x,y)\, h(x,y)$

(Marginals) $H(X) = \sum_{x} p(x)\, h(x)$, $H(Y) = \sum_{y} p(y)\, h(y)$

(conditional). $\left\{ \underbrace{p(x|y)}_{h(x|y) = \log_2(1/p(x|y))} = \mathbb{P}[X=x \mid Y=y] \right\}_{x \in \mathcal{X}} \quad \forall\, y \in \mathcal{Y}$

$$H(X|Y) = \sum_{y} p(y) \left(\sum_{x} p(x|y)\, h(x|y) \right)$$

the chain rule

the chain rule (information content)

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

$$h(x,y) = h(x) + h(y|x) = h(y) + h(x|y)$$

$h(x,y) = \log_2\left(1/p(x,y)\right)$, $h(x) = \log_2 1/p(x)$, $h(x|y) = \log_2 1/p(x|y)$

- $\log_2 \dfrac{1}{p(x,y)} = \log_2\left(\dfrac{1}{p(x)\,p(y|x)}\right) = h(x) + h(y|x)$

the chain rule

the chain rule (entropy)

for any rvs X, Y:

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

$$H(X,Y) = \sum_{(x,y)} p(x,y) \log_2\left(1/p(x,y)\right)$$

$$H(X) = \sum_x p(x) \log_2\left(1/p(x)\right) = \sum_{x,y} p(x,y) \log_2\left(1/p(x)\right)$$

$$H(X|Y) = \sum_y p(y)\left(\sum_x p(x|y) \log_2\left(1/p(x|y)\right)\right)$$

$$= \sum_{x,y} p(x,y) \log_2\left(1/p(x|y)\right)$$

mutual information

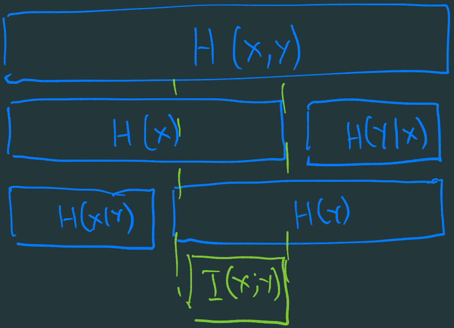
mutual information

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

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



$H(x,y) = H(x) + H(y) - I(x;y)$

example

$P(x,y)$		x = 1	2	3	4	$P(y)$
y	1	1/8 (3)	1/16 (4)	1/32 (5)	1/32 (5)	1/4
	2	1/16 (4)	1/8 (3)	1/32 (5)	1/32 (5)	1/4
	3	1/16 (4)	1/16 (4)	1/16 (4)	1/16 (4)	1/4
	4	1/4 (2)	0	0	0	1/4
$P(x)$		1/2	1/4	1/8	1/8	
$h(x)$		1	2	3	3	

$h(x,y)$

$$H(Y)=2,\quad H(X)=\frac{1}{2}+\frac{2}{4}+\frac{3}{8}+\frac{3}{8}=7/4$$

$$\Rightarrow H(X,Y)=\frac{1}{4}\cdot 2+\frac{2}{8}\cdot 3+\frac{6}{16}\cdot 4+\frac{4}{32}\cdot 5$$

$$=\frac{4+6+12+5}{8}=\frac{27}{8}$$

$$H(X)+H(Y)=\frac{30}{8}\geq H(X,Y)$$

$P(x|y)$

y \ x	1	2	3	4	$H(X \mid Y=y)$
1	1/2	1/4	1/8	1/8	
2	1/4	1/2	1/8	1/8	
3	1/4	1/4	1/4	1/4	
4	1	0	0	0	

$P(y|x)$

y \ x	1	2	3	4
1	1/4	1/4	1/4	1/4
2	1/8	1/2	1/4	1/4
3	1/8	1/4	1/2	1/2
4	1/2	0	0	0

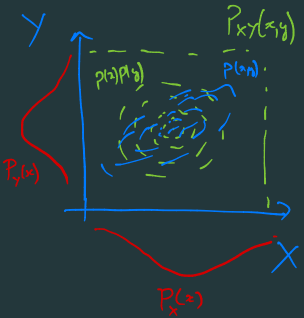
$$\Rightarrow \boxed{I(X;Y)=3/8}$$

mutual information and KL divergence

mutual information

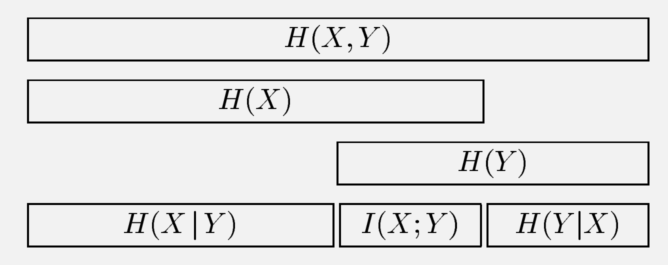
for any rvs X, Y: $\quad I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

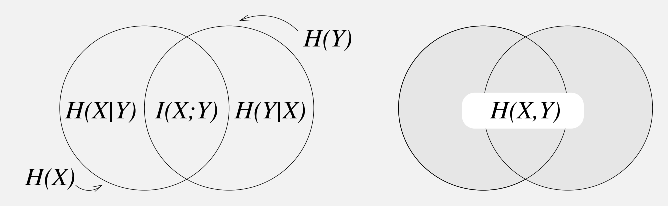
$$I(X;Y) = D_{KL}\Big(\underbrace{P(x,y)}_{\text{True dist of }(X,Y)} \,\Big\|\, \underbrace{P(x)P(y)}_{\text{Dist}^{\text{n}}\text{ of }(X,Y)\text{ if } X \perp\!\!\!\perp Y}\Big)$$



increase in codesize if you encode $P(x,y)$ using optimal code for $P(x)P(y)$

visualizing mutual information



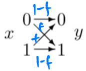


chapter 9 in Mackay

channel coding

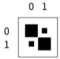
mutual information for the BSC

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



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

$f=0.1$



assume input distribution $\mathcal{P}_X = \{1-p, p\} = \{p_0, p_1\}$, $p_0 + p_1 = 1$

$x \backslash y$	0	1	$p(x)$
0	$p_0(1-f)$	$p_0 f$	p_0
1	$p_1 f$	$p_1(1-f)$	p_1
$p(y)$	$p_0(1-f) + p_1 f$	$p_0 f + p_1(1-f)$	

$\mathrm{Ber}(p)$

$$\sim \mathrm{Ber}\big(\underbrace{p_1 f + p_0(1-f)}_{q}\big) \sim \mathrm{Ber}\big(\underbrace{p_0 f + p_1(1-f)}_{1-q}\big)$$

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

$$= h_2(q) -$$

$$P(x=1)\,H(Y \mid x=1) + P[x=0]\,H[Y \mid x=0]$$

$$= h_2(q) - p_1 h_2(f) - p_0 h_2(f)$$

$$= \boxed{h_2(q) - h_2(f)}$$

$$\Rightarrow I(x;y) = \underbrace{h_2\left(p_1 f + p_0(1-f)\right)}_{H(y)} - \underbrace{h_2(f)}_{H(y|x)}$$

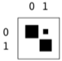
<u>Exercise</u> - $I(x;y) = H(x) - H(x|y)$

$$= h_2(p_1) - \quad ?$$

mutual information for the Z-channel

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

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



assume input distribution $\mathcal{P}_X = \{1-p, p\} = \{p_0, p_1\}$

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

$$\boxed{Y = \begin{cases} 0 & \text{wp } p_0 + p_1 f \\ 1 & \text{wp } p_1(1-f) \end{cases}}$$

$$= h_2(p_1(1-f)) - p_0 \overbrace{H(Y|X=0)}^{0} - p_1 \underbrace{H(Y|X=1)}_{h_2(f)}$$

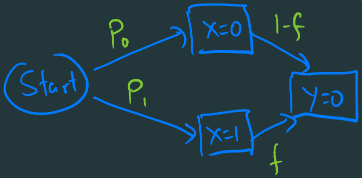
$$= \boxed{h_2(p_1(1-f)) - p_1 h_2(f)}$$

$\underline{\text{Alt}}$. $I(X;Y) = H(X) - H(X|Y)$

$= h_2(p_1) - \underbrace{H(X|Y=1)}_{=0}\,\overbrace{\mathbb{P}[Y=1]}^{p_1(1-f)}$

$- \underbrace{H(X|Y=0)}_{?}\,\underbrace{\mathbb{P}[Y=0]}_{p_0+p_1 f}$

Bayes Thm



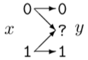
$$P_{X|Y=0} \sim \left\{ \frac{p_0(1-f)}{\underbrace{q_1}_{\mathbb{P}[Y=1]}}, \frac{p_1 f}{q_1} \right\}$$

$= h_2(p_1) - (p_0 + p_1 f)\, h_2(\theta)$

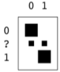
where $\theta = \dfrac{p_1 f}{p_1 f + p_0(1-f)}$

mutual information for the erasure channel

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 $\mathcal{P}_X = \{1-p, p\} = \{p_0, p_1\}$

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

$\because H(Y|X) = \sum p(x) H(Y|X=x)$

$= p_0 H(Y|X=0) + p_1 H(Y|X=1)$

However, both $H(Y|X=1)$ & $H(Y|X=0)$ are $h_2(f)$

$$= h_2(f) + (1-f)\, h_2(p_1)$$

$$\Big[\ p_0 h_2(f) - p_1 h_2(f)$$

$$= \boxed{(1-f)\, h_2(p_1)}$$

$$Y = \begin{cases} 0 & \text{wp } p_0(1-f) \\ ? & \text{wp } f \\ 1 & \text{wp } p_1(1-f) \end{cases}$$

$$H(Y) = p_0(1-f)\log_2\left(\frac{1}{p_0(1-f)}\right) + p_1(1-f)\log_2\left(\frac{1}{p_1(1-f)}\right) + f\log_2\frac{1}{f}$$

$$= h_2(f) + (1-f)\, h_2(p_1)$$

Let $Z = \mathbb{1}\{Y = ?\}$, then

$H(Y) = H(Z) + H(Y|Z)$

(Since Z is a fn of Y)

Note - $I(X;Y)$ separates into terms depending on f ad on p_1

capacity of a channel

channel capacity

the capacity of a channel $\mathcal{Q}$, with input $\mathcal{A}_{\mathcal{X}}$ and output $\mathcal{A}_{\mathcal{Y}}$, is

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

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!

Mackay. Ch 9 (Defines capacity, channel coding)

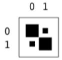
Ch 10 (Pf of Shannon's coding thm)

capacity of the BSC

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

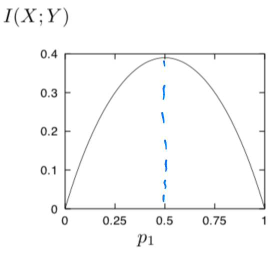
x 0 → 0 y, 1 → 1

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



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

$f = 0.1$



$$I(X;Y) = h_2(q) - h_2(f)$$

$$q = p_1 f + p_0(1-f)$$

$$C(f) = \max_{(p_0,p_1)\ \text{s.t.}\ p_0+p_1=1} I(X;Y)$$

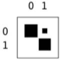
$$\Rightarrow P^* \equiv \text{set } q = 1/2 \Rightarrow p_1 f + (1-p_1)(1-f) = 1/2 \Rightarrow p_1^* = p_0^* = 1/2$$

$$\Rightarrow \boxed{C(f) = 1 - h_2(f)}$$

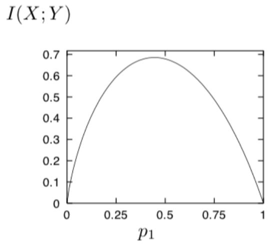
capacity of the Z-channel

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

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



assume input distribution $\mathcal{P}_X = \{1-p, p\} = \{P_0, P_1\}$

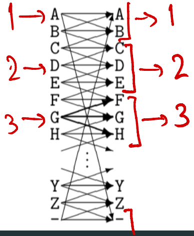


$$I(X;Y) = h_2(P_1(1-f)) - P_1 h_2(f)$$

This is not symmetric in P_1 - Somewhat complicated to maximize (see fig)

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.



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

$$I(X;Y) = \underbrace{H(Y)}_{\le \log|\mathcal{A}_Y|} - \underbrace{H(Y|X)}_{=\sum_x p(x)\,H(Y|X=x) \,=\, \log_2 3}$$

$$\le \underbrace{\log_2 27} - \log_2 3 = \log_2 9$$

Can be achieved, Eg, set $P_X = (1/27, 1/27, \ldots\ldots 1/27)$

Capacity | Coding for noisy typewriter

$$C(Q) = \max_{P_X} I(X;Y) = \log_2 9 \text{ bits}$$

(P_X^* can be $\{1/27, 1/27, \ldots, 1/27\}$)

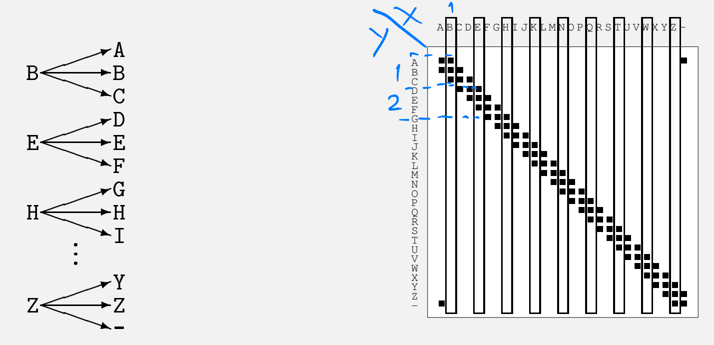
- Code for noisy typewriter: $\phi : \{1,2,\ldots,9\} \to \{A, B, \ldots, Z, -\}$

Encoder $\phi(1) = A$, $\phi(2) = D$, $\phi(3) = G$ ...

Decoder $\phi^{-1}(\{-, A, B\}) = 1$, $\phi^{-1}(\{C, D, E\}) = 2$, ...

Can send $\log_2 9$ bits per channel use without error

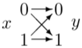
another view of the noisy typewriter



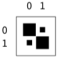
Syndrome decoding - map set of outputs to same input

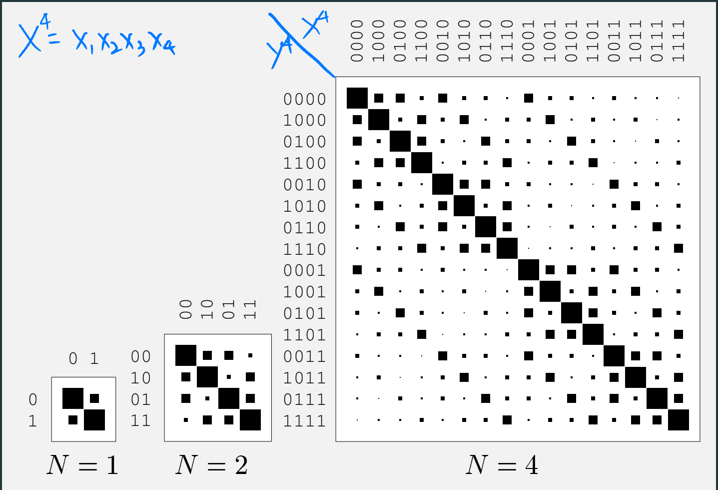
expanded channel for the BSC

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



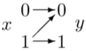
$$
\begin{aligned}
P(y=0 \,|\, x=0) &= 1-f; & P(y=0 \,|\, x=1) &= f;\\
P(y=1 \,|\, x=0) &= f; & P(y=1 \,|\, x=1) &= 1-f.
\end{aligned}
$$



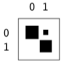


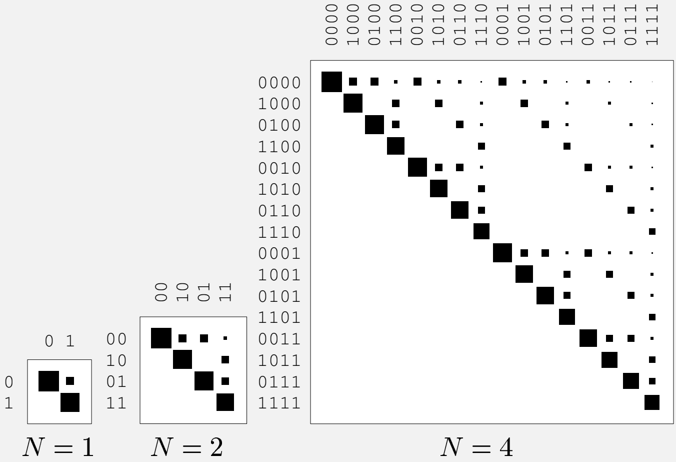
expanded channel for the Z-channel

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



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





lossless compression via typical set encoding

typical set

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

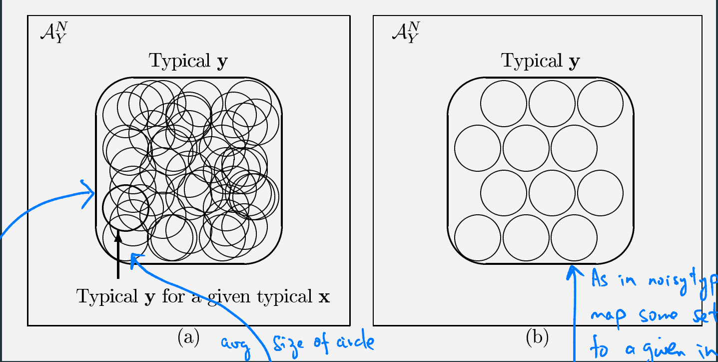
then X^N is very likely to be one of $\approx 2^{NH(X)}$ typical strings,

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

Recall- bent coin lottery

- $\underbrace{X_1 X_2 \ldots\ldots X_{1000}} \sim \mathrm{Bin}(1000, f)$
- Most of the time, # of 1s $= 1000f \mp \sqrt{1000f}$

typical set and non-confusable subset



Union of all circles

avg size of circle

As in noisy typewriter, we map some set of outputs to a given input.

# of elements in the typical set of outputs $\approx 2^{NH(Y)}$

# of typical outputs for typical input X $\approx 2^{NH(Y|X)}$

# of non-overlapping circles $\approx 2^{NH(Y)} / 2^{NH(Y|X)}$

$= 2^{NI(X;Y)} \leq 2^{NC}$

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

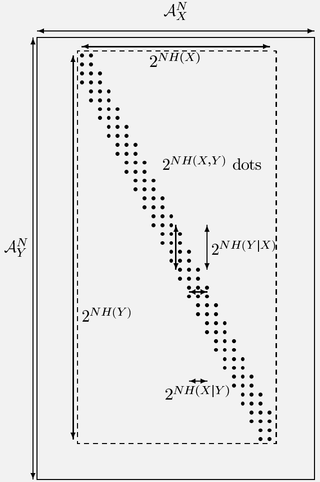
for channel $\mathcal{Q}$ with input $\mathcal{A}_X$, an (N, K)-block code is a list of $\mathcal{S} = 2^K$ codewords $\{x^{(1)}, x^{(2)}, \ldots, x^{(2^K)}\}$ with $x^{(i)} \in \mathcal{A}_X^N$ (i.e., of length N)
– the rate of the code is $R = N/K$ bits per channel use

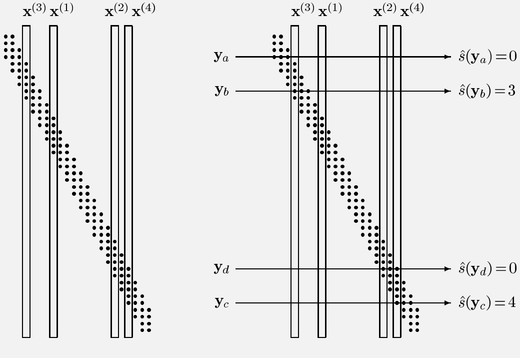
Shannon's channel coding theorem

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 $< \epsilon$.

- Only transmit 2^K out of $|A_X|^N$ symbols

intuition behind proof

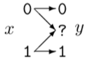




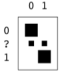
(a) (b)

erasure channel capacity

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



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



$P_X = \{p_0, p_1\}$

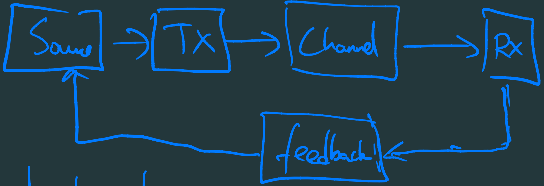
$$I(X;Y) = (1-f)\, h_2(p_1)$$

$$\Rightarrow \quad C = \max_{\{p_0, p_1 \mid p_0 + p_1 = 1\}} I(X;Y) = 1-f \quad \text{for } P_X = \{1/2, 1/2\}$$

How can we design a scheme to achieve this?

feedback coding

<u>Idea</u> - Suppose we have feedback from the receiver



- <u>Code</u> - If Rx gets ? , asks for a repeat character (retransmit / ACK protocol)

- $\mathbb{P}[\text{bit received correctly}] = 1-f$

$\Rightarrow$ # of retransmissions $\sim \text{Geom}(1-f)$, $\mathbb{E}[\#\text{ of retx}] = \frac{1}{1-f}$

(can do without feedback - fountain codes)