Validation via information theory

Organization

What is PA?

Rationale of PA

Roadmap

Shannon's coding theorem

Formalization of PA

Posterior agreement

How to validate algorithms ?

$A : X \longmapsto$ distr. $p(\cdot|X)$ over solutions

clustering	dataset of points	cluster assignments
shortest path from A to B	graph	paths from A to B
linear regression	dataset of points	linear models

Posterior agreement

How to validate algorithms ?

$$A: X \longmapsto p(\cdot|x)$$

Expected log posterior agreement:

$$\frac{1}{\log|C|}\, \mathbb{E}_{x',x''}\left[\log\left(|C|\,\kappa(x',x'')\right)\right]$$

where

$$\kappa(x',x'') = \sum_c p(c|x')\,p(c|x'')$$

$$= \mathbb{E}_{p(\cdot|x')}\left[p(\cdot|x'')\right]$$

Posterior agreement

Expected log posterior agreement:

$$\frac{1}{\log|\mathcal{C}|}\mathbb{E}_{x',x''}\left[\log\left(|\mathcal{C}|\kappa(x',x'')\right)\right]$$

In practice, emp. log PA $= \frac{1}{\log|\mathcal{C}|}\log\left(|\mathcal{C}|\kappa(x',x'')\right)$.

When comparing A_1 and A_2, choose the one that maximizes $\kappa(x',x'')$.

Organization

What is PA?

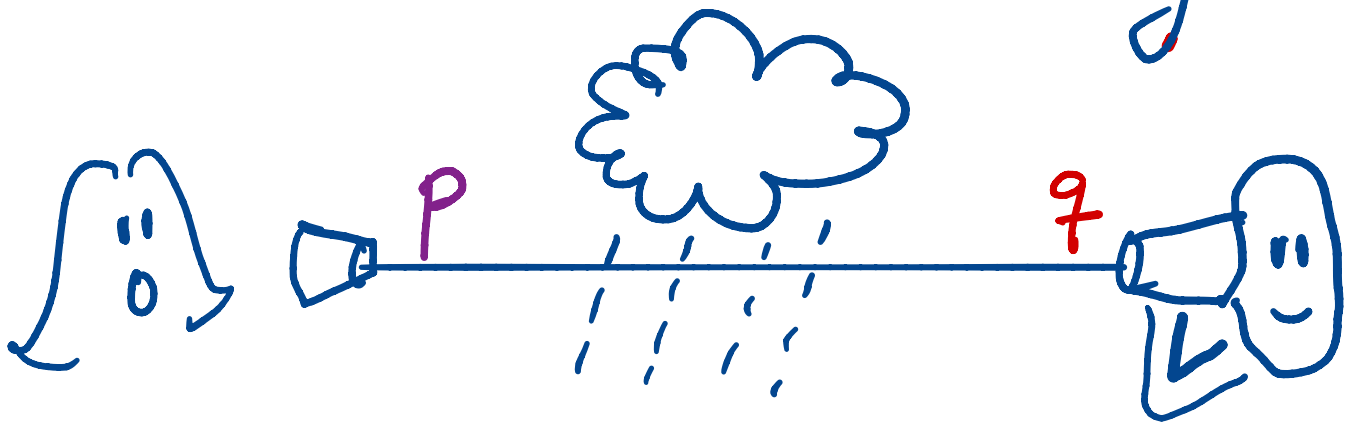
Rationale of PA

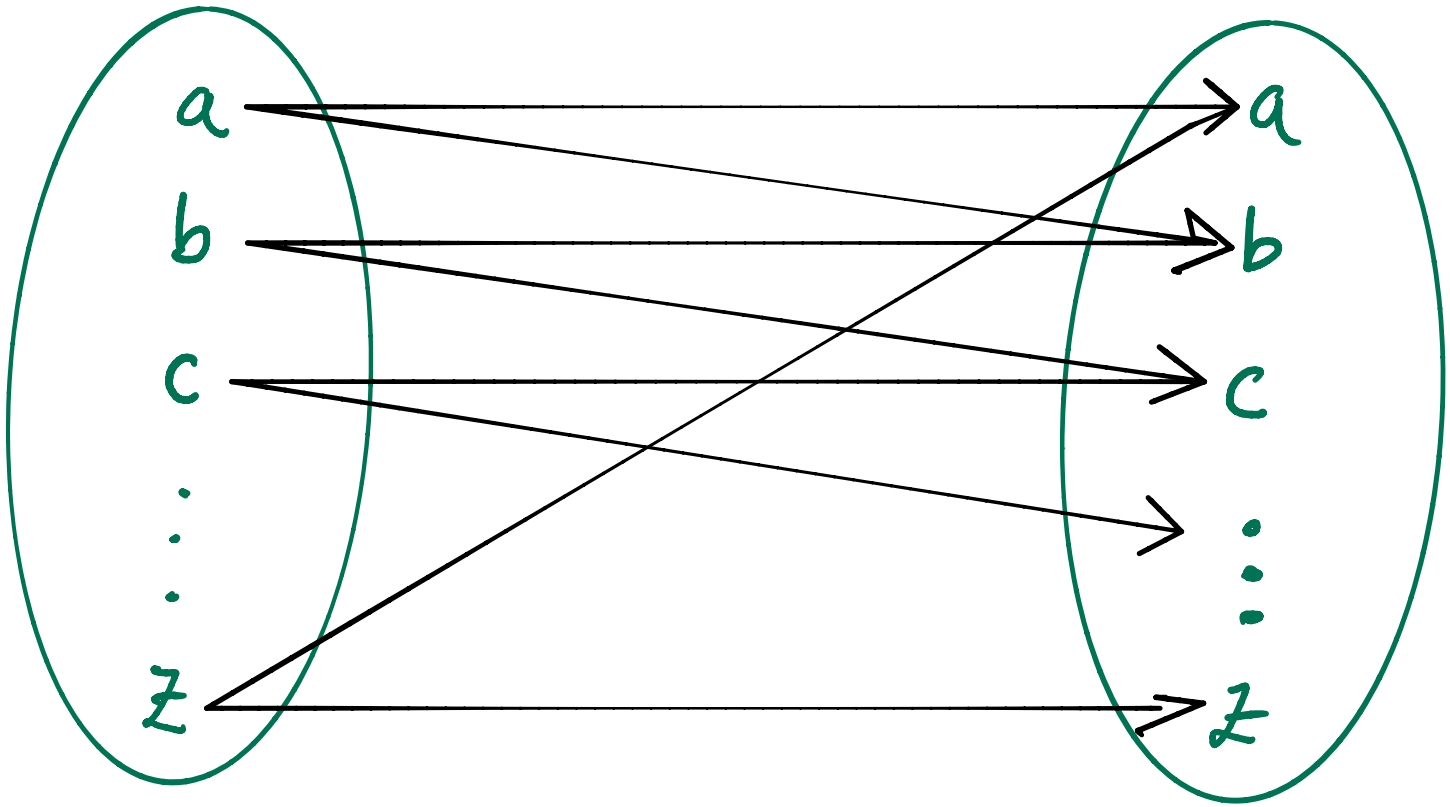
Roadmap

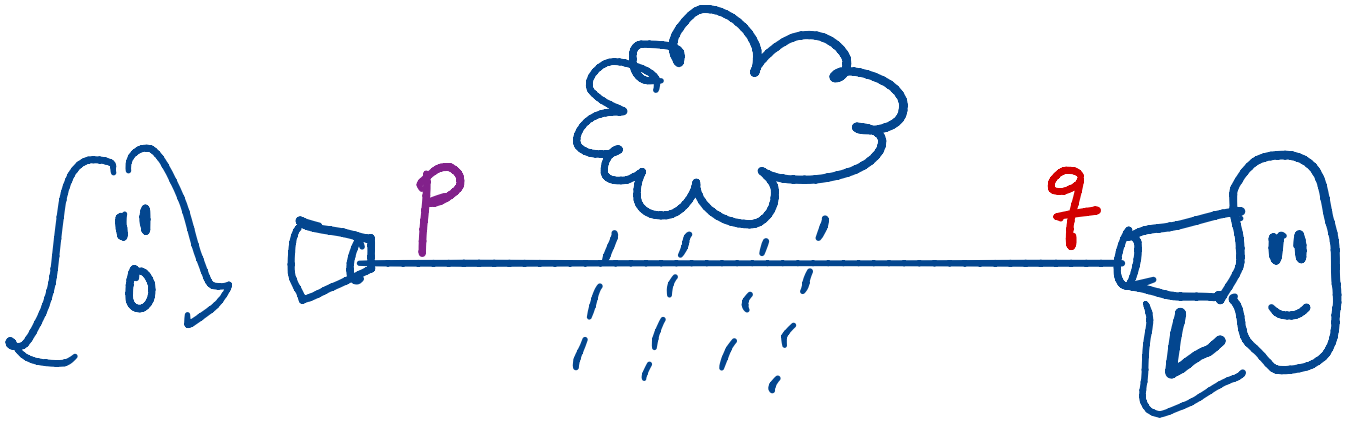
Shannon's coding theorem

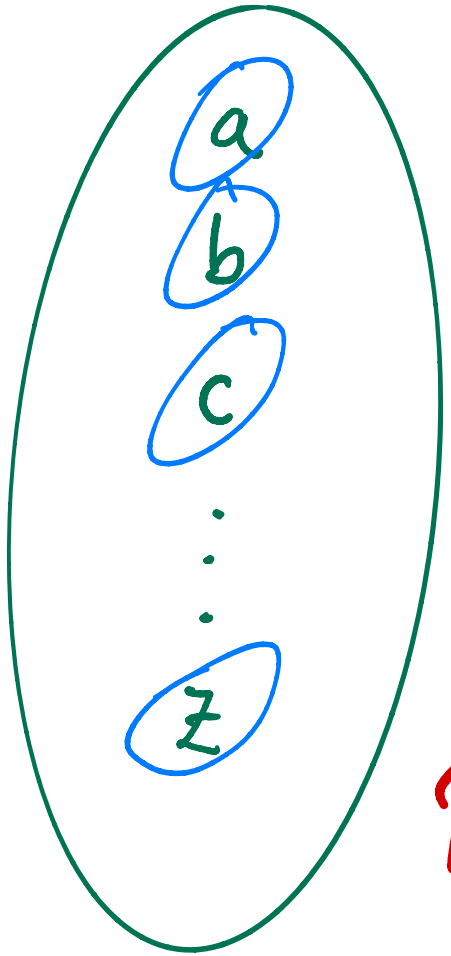
Formalization of PA

Communication on a noisy channel



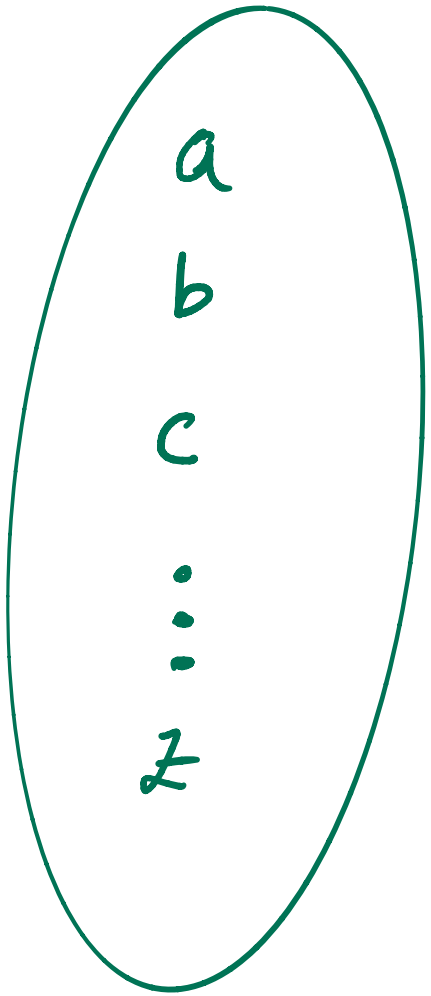


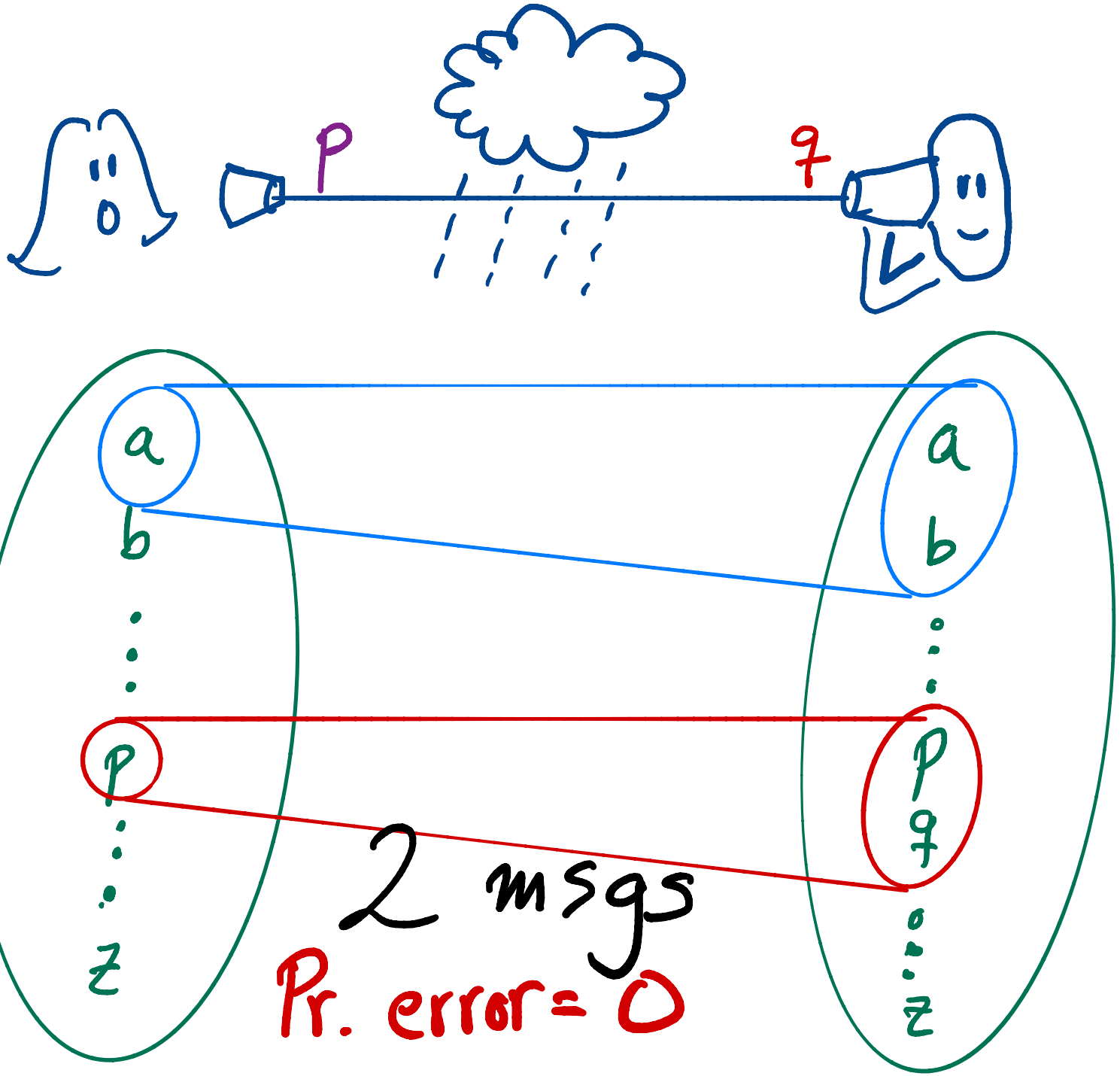


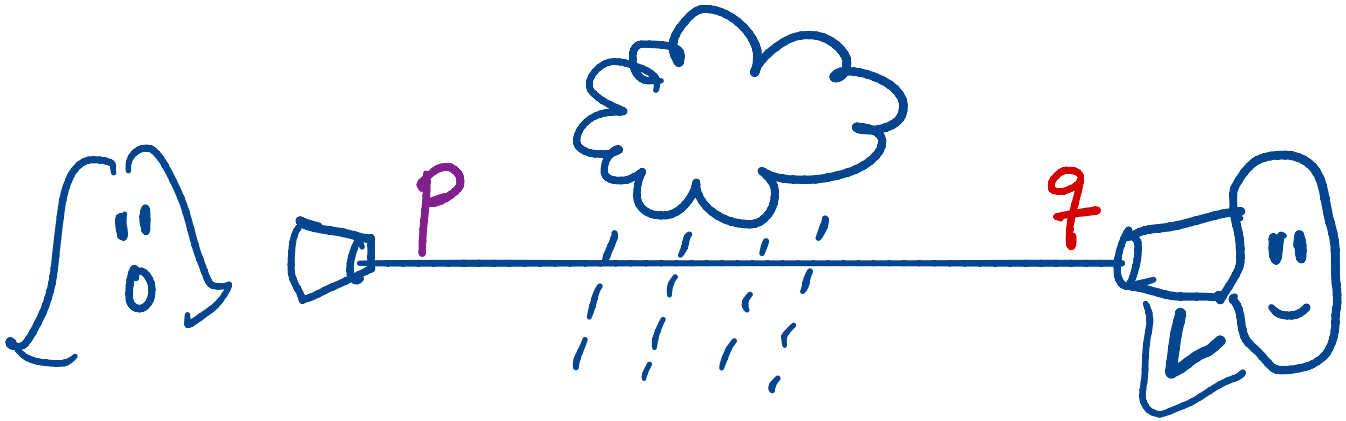


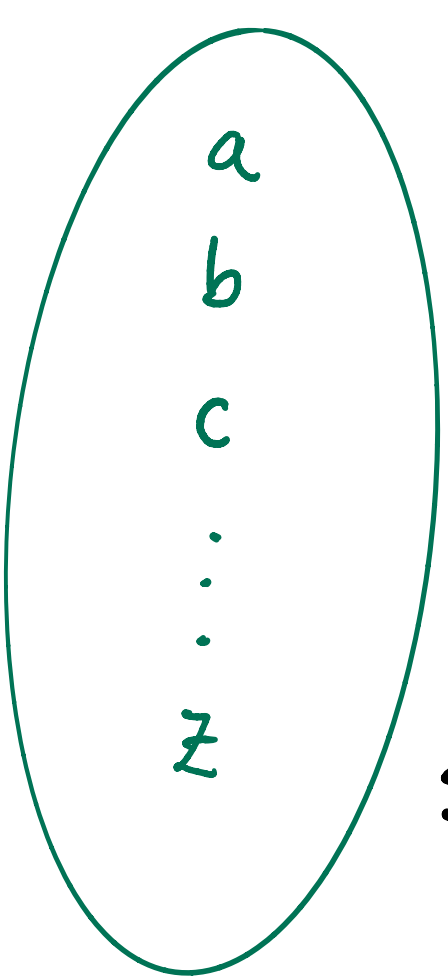
26 msgs

Pr. error > 0

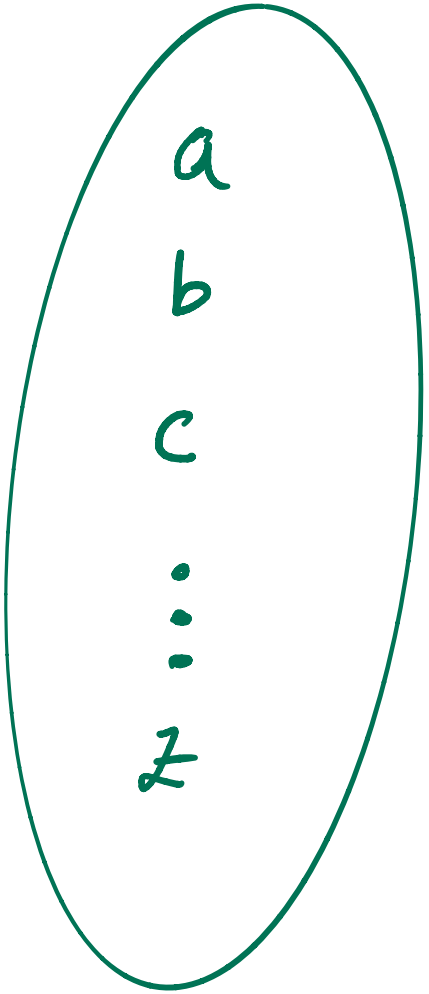


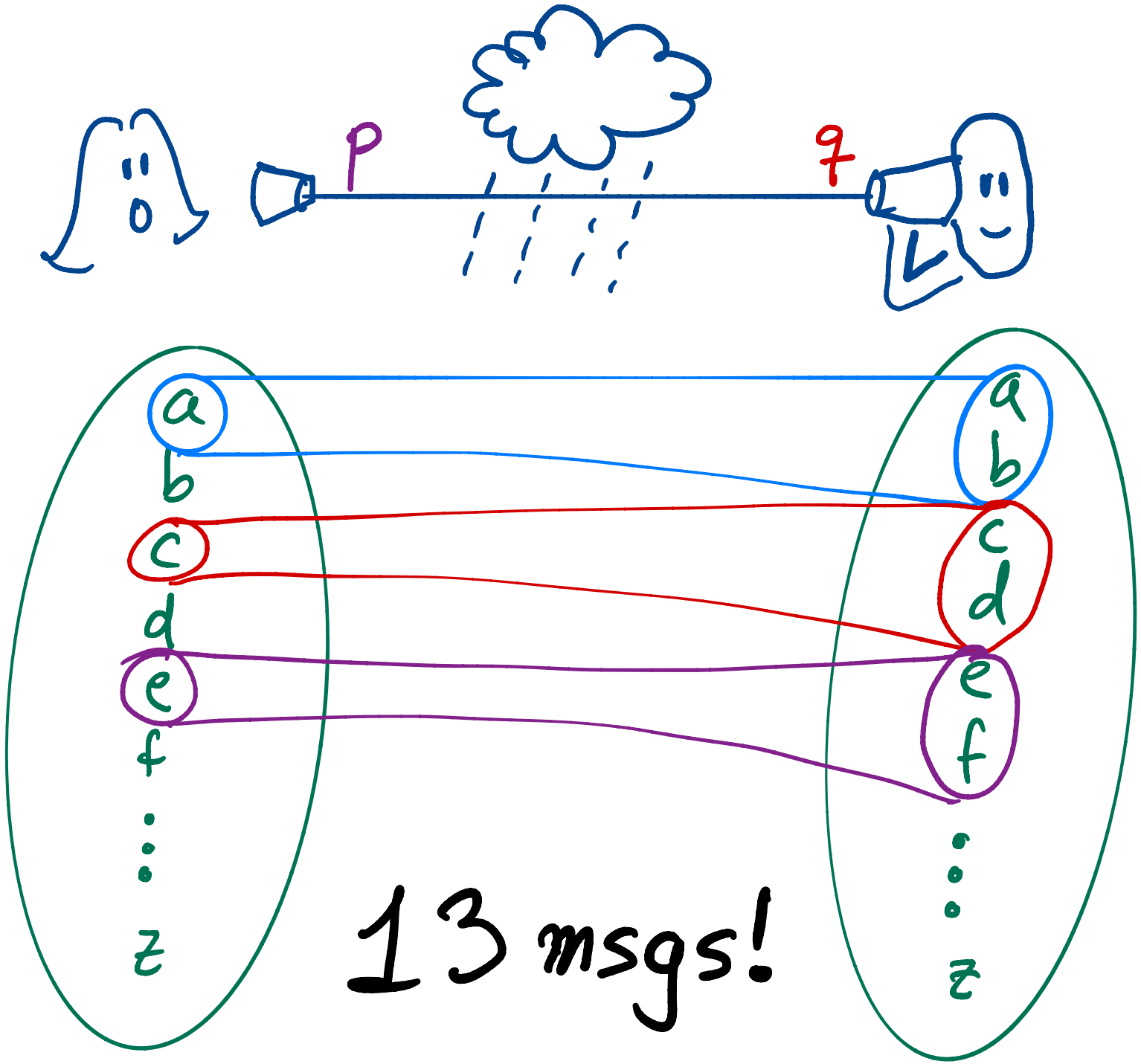

p
q
a
b
p
z
a
b
p
q
z
2 msgs
Pr. error = 0

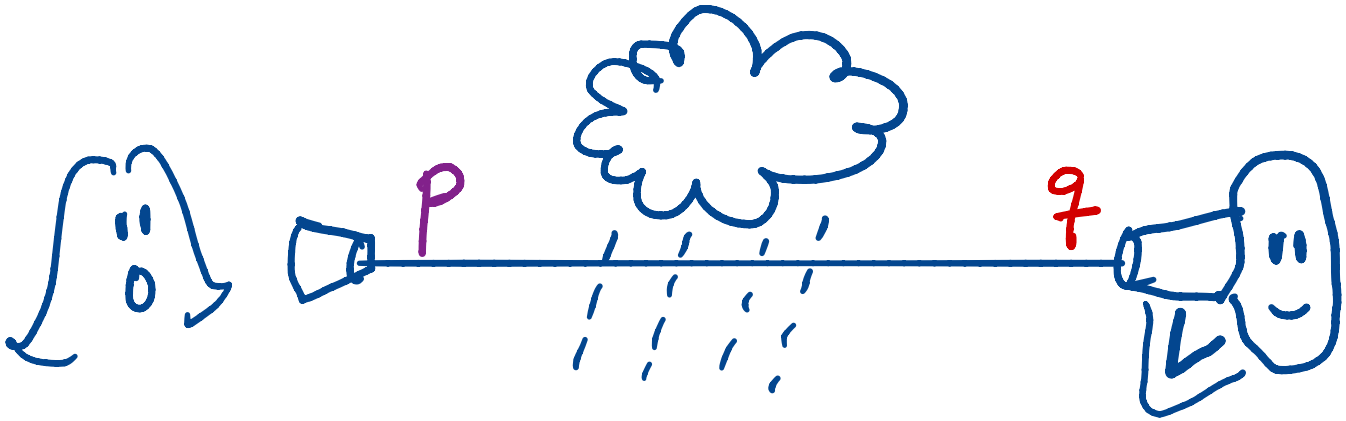


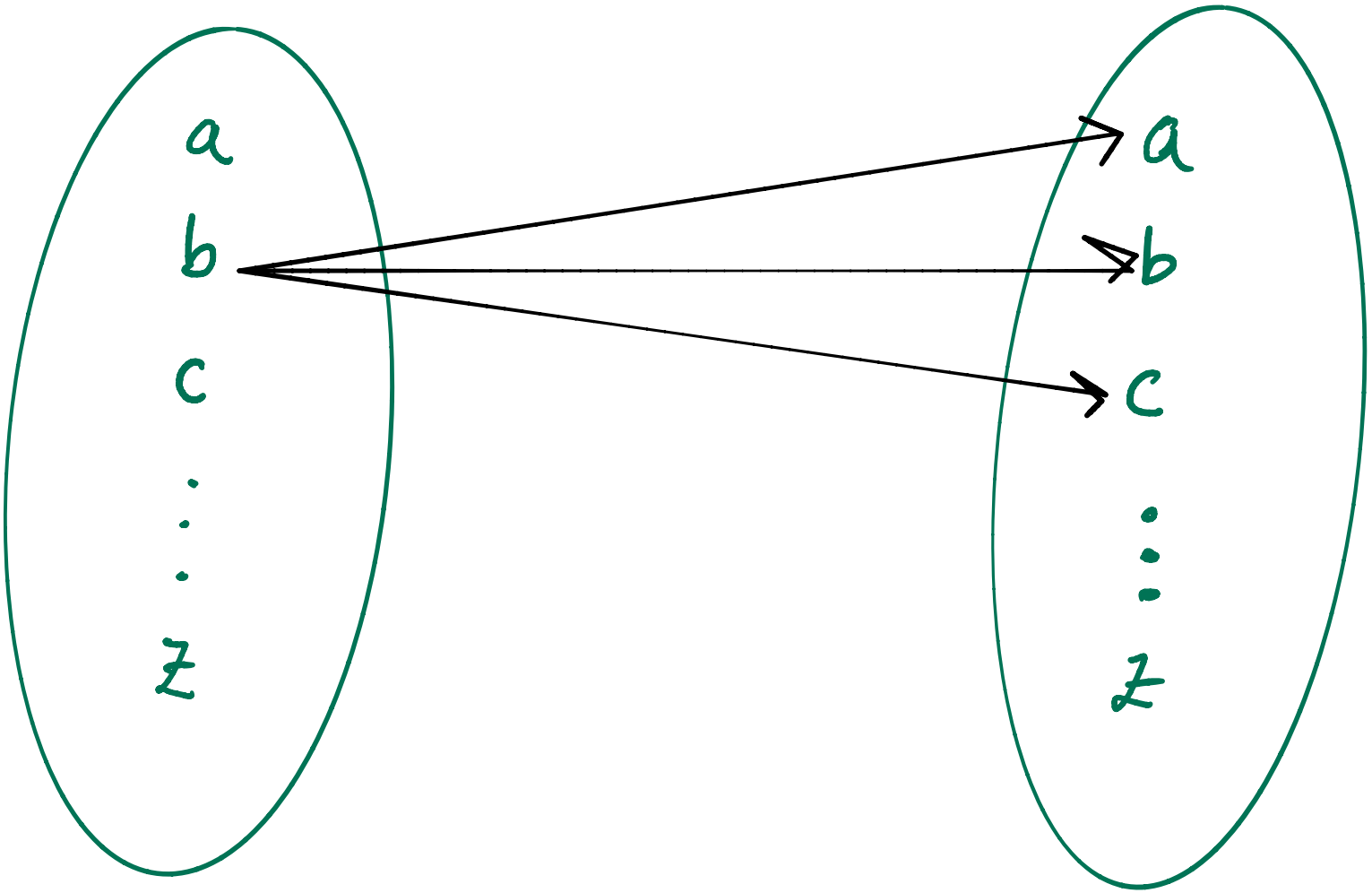


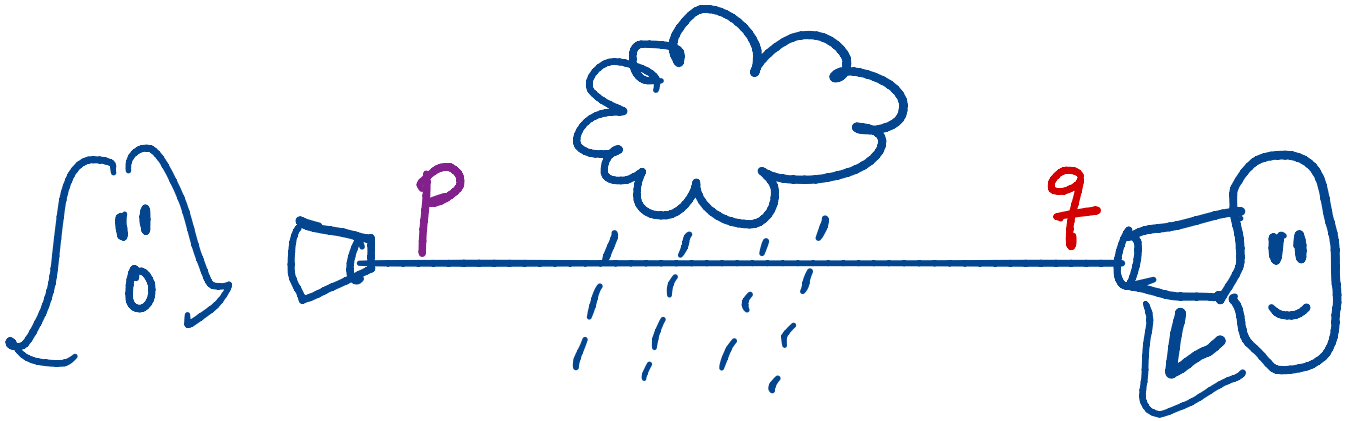
max # msgs
s.t. Pr. error=0?

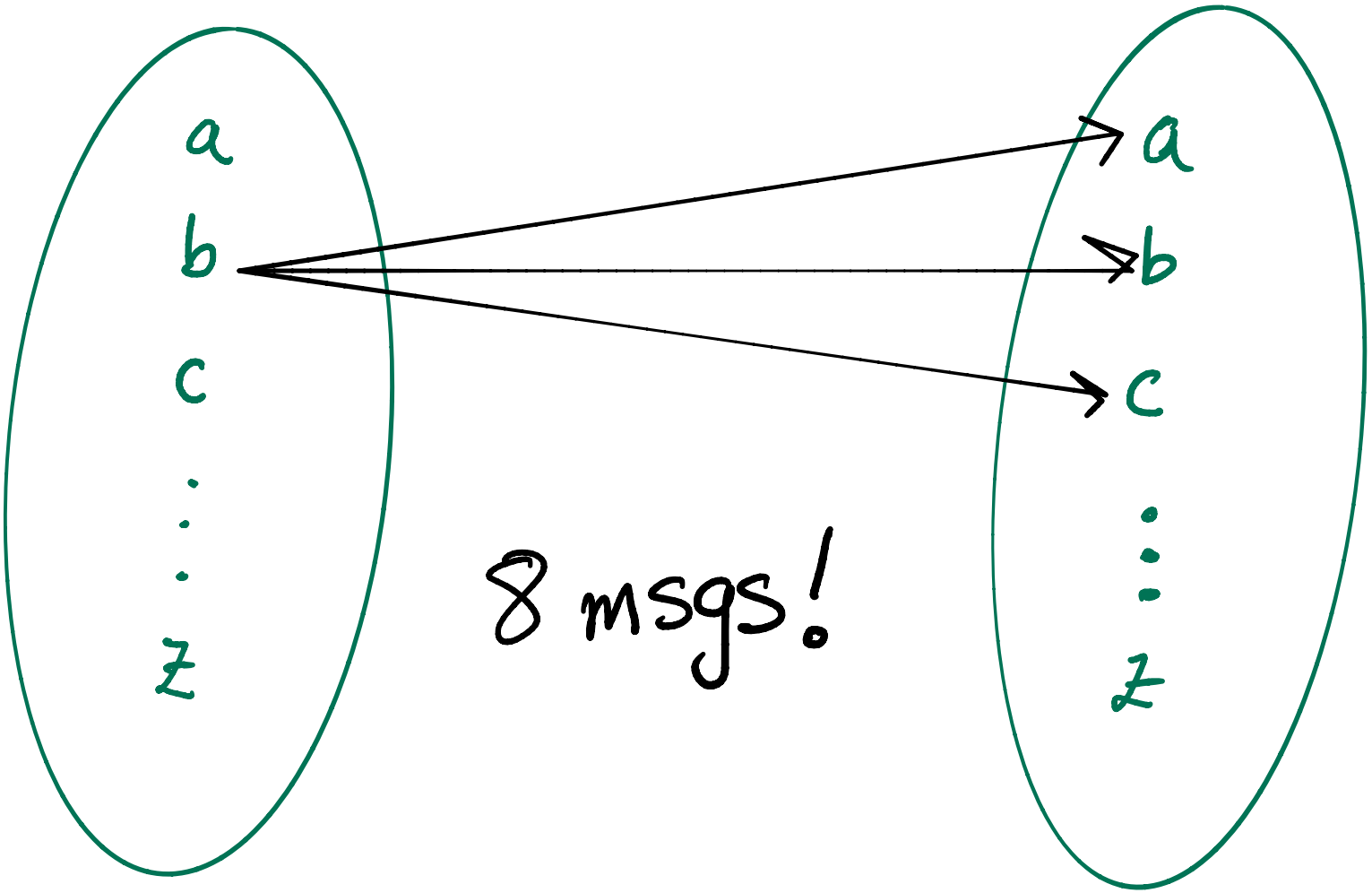


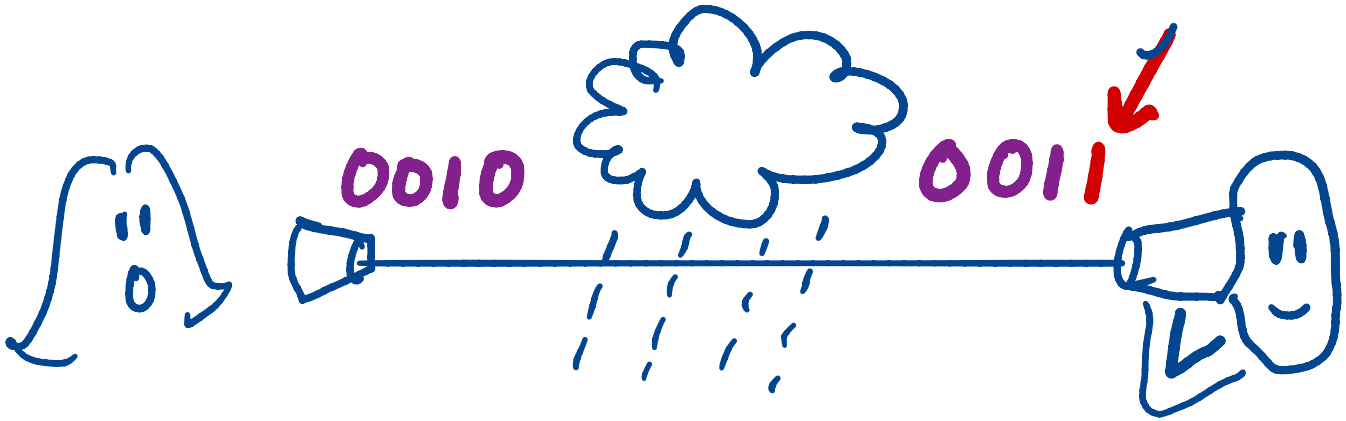

p
q
a
b
c
d
e
f
z
a
b
c
d
e
f
z
13 msgs!


p
q


a
b
c
z
a
b
c
z


p
q


a
b
c
z
a
b
c
z
8 msgs!



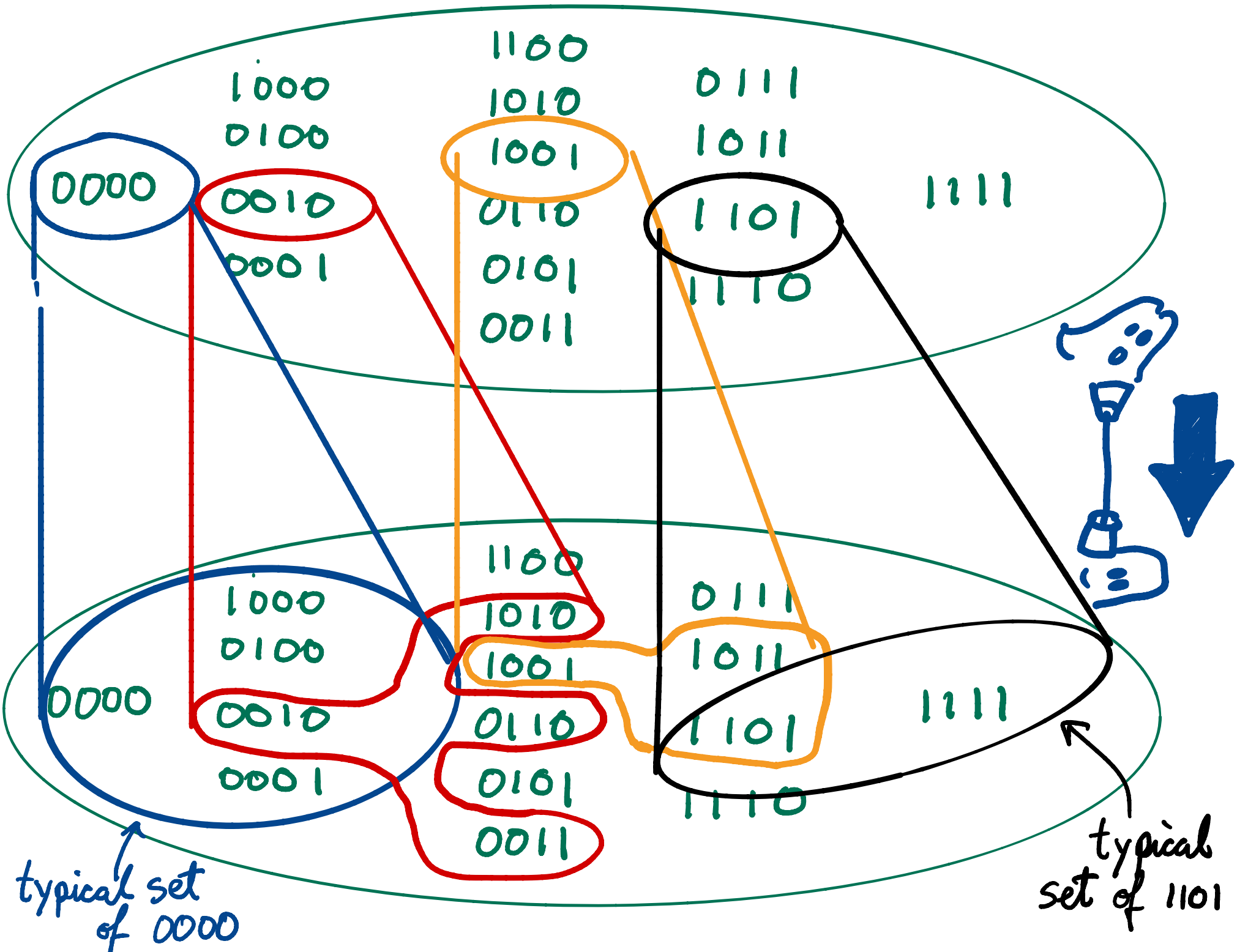
n=4

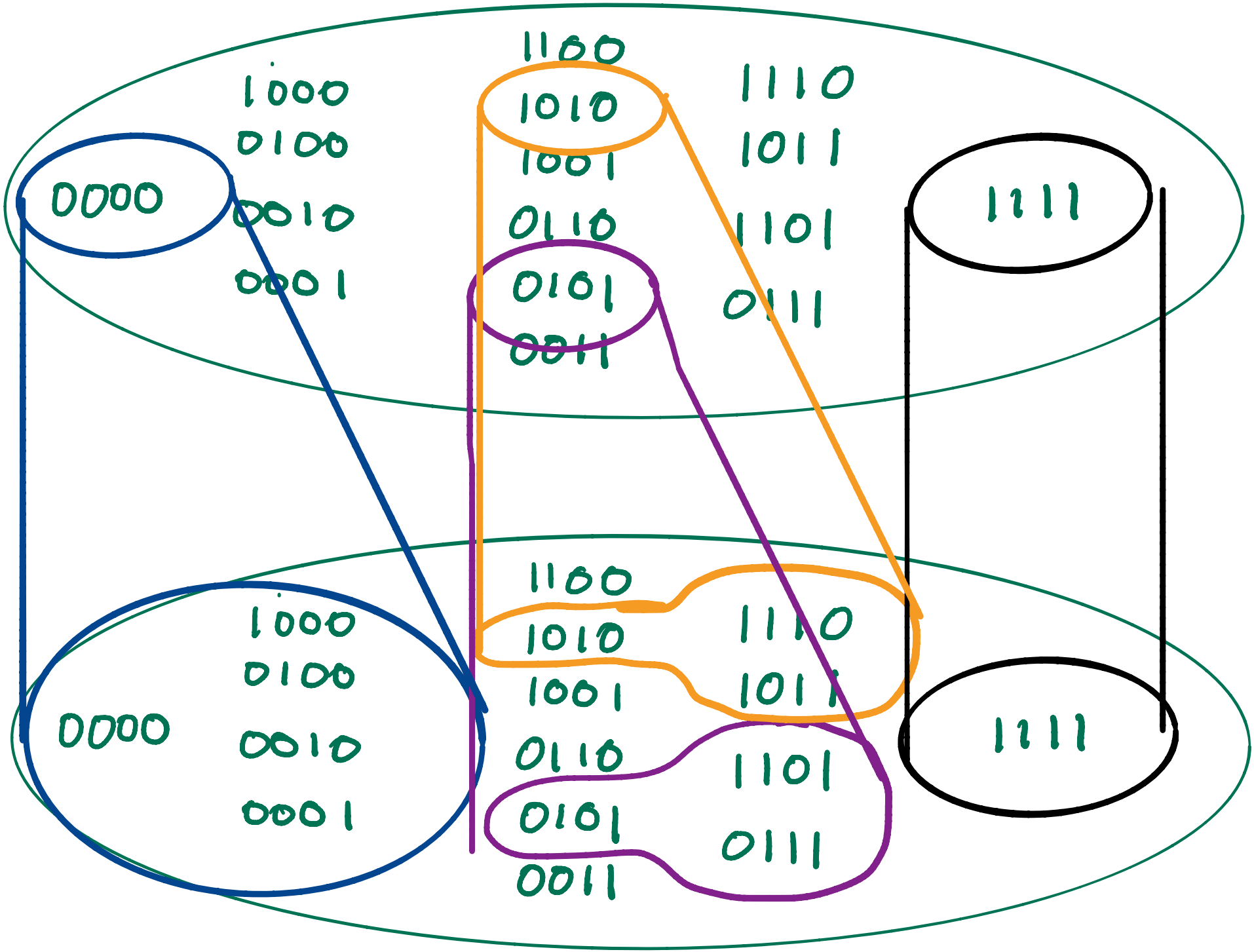
0000
1000
0100
0010
0001
1100
1010
1001
0110
0101
0011
0111
1011
1101
1110
1111

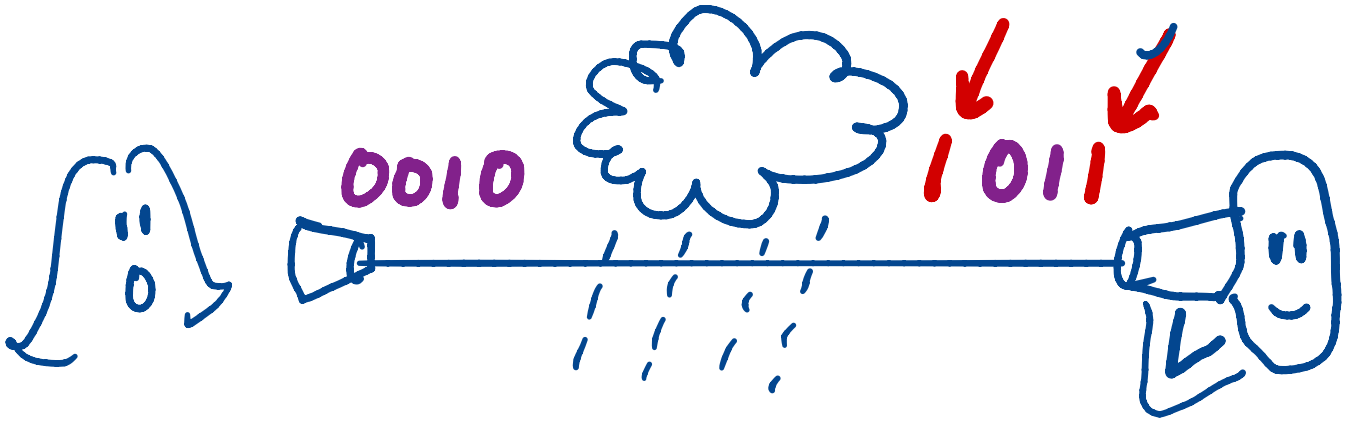
codeword

code space

max # msgs s.t. Pr. error = 0?


1100
1000
1010
0111
0100
1001
1011
0000
0010
0110
1101
1111
0001
0101
1110
0011
1100
1000
1010
0111
0100
1001
1011
0000
0010
0110
1101
1111
0001
0101
1110
0011
typical set of 0000
typical set of 1101

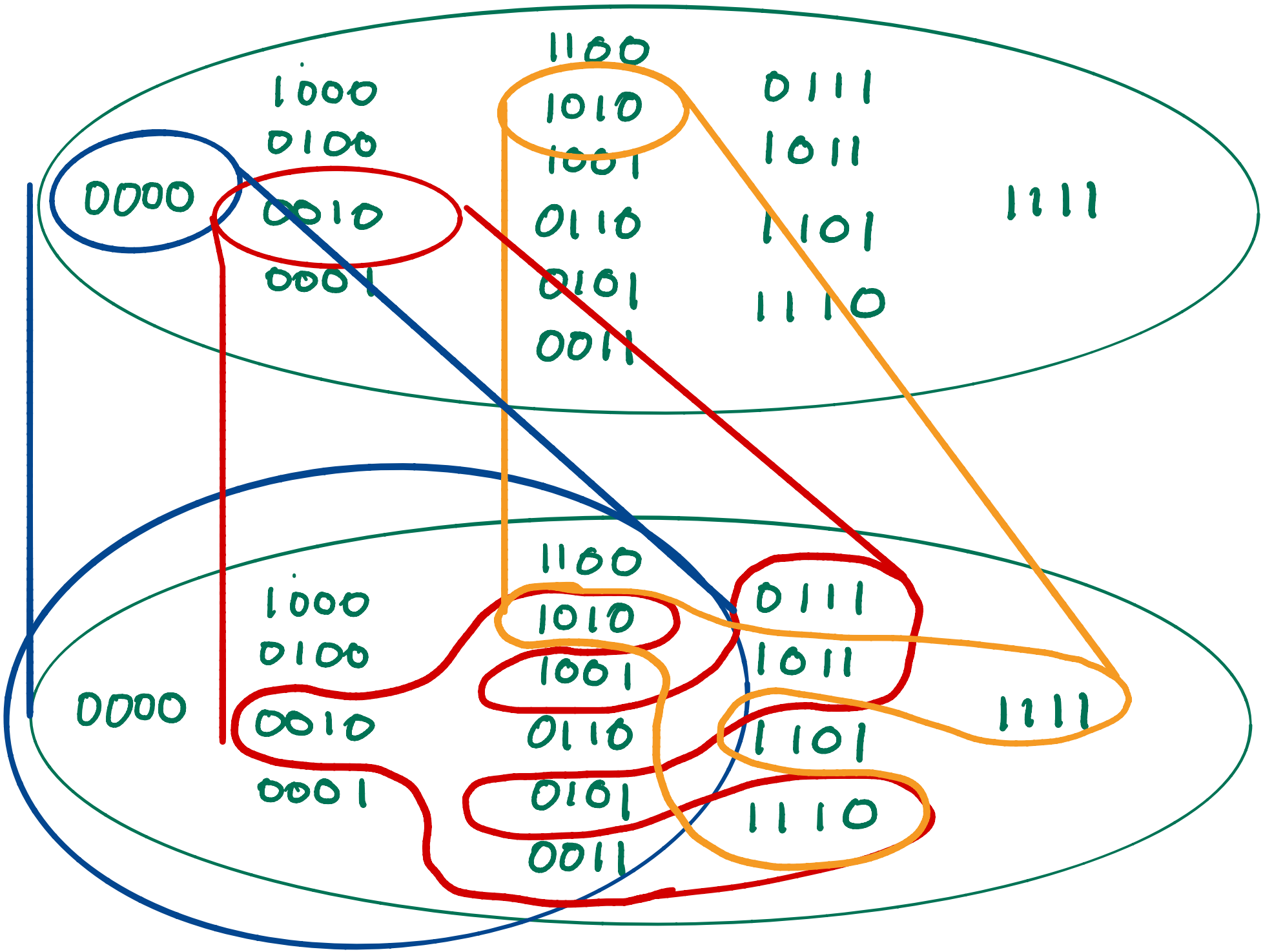

1000
0100
0010
0001
0000
1100
1010
1001
0110
0101
0011
1110
1011
1101
0111
1111
1000
0100
0010
0001
0000
1100
1010
1001
0110
0101
0011
1110
1011
1101
0111
1111

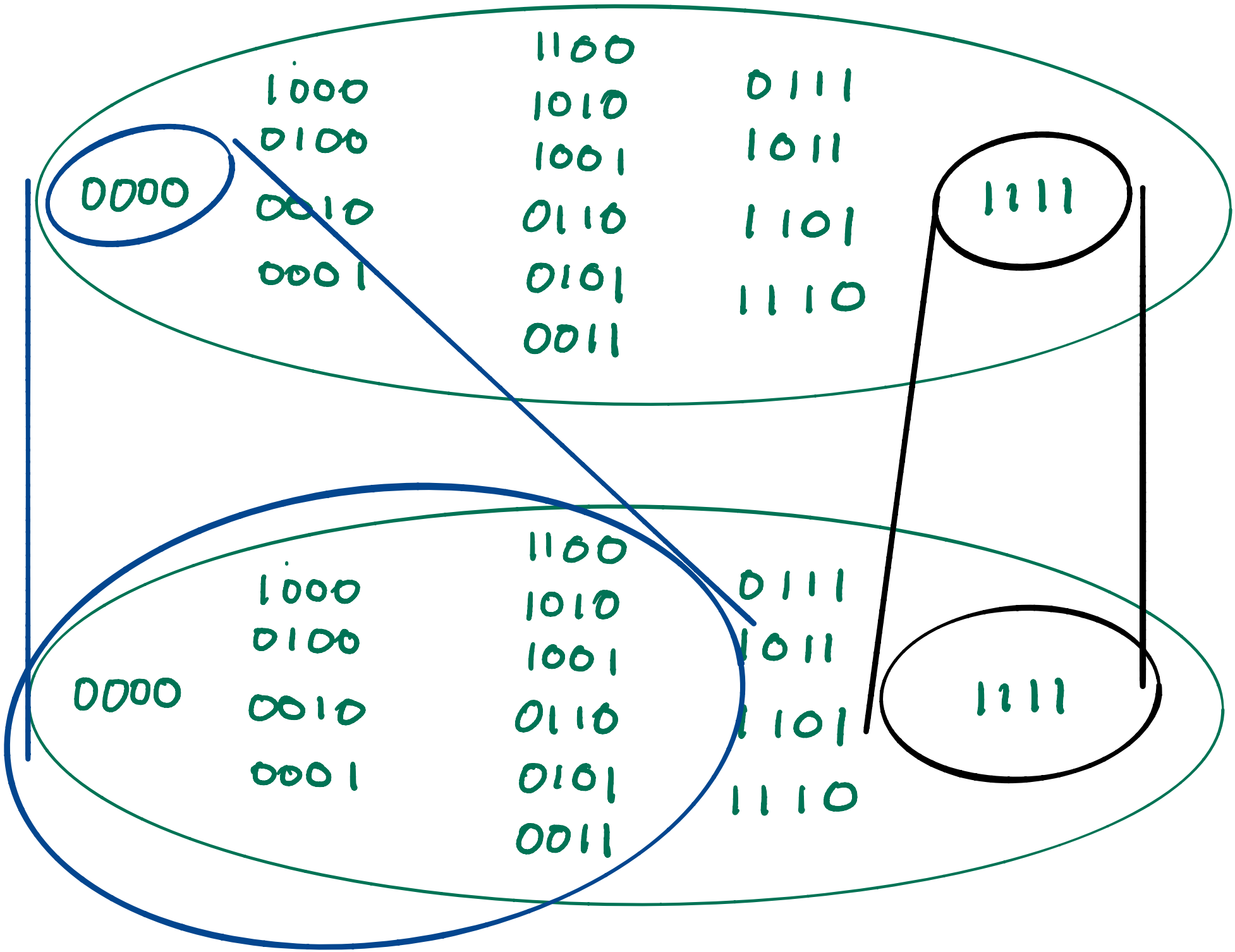


n=4

0000
1000
0100
0010
0001
1100
1010
1001
0110
0101
0011
0111
1011
1101
1110
1111

max # msgs s.t. Pr. error = 0?

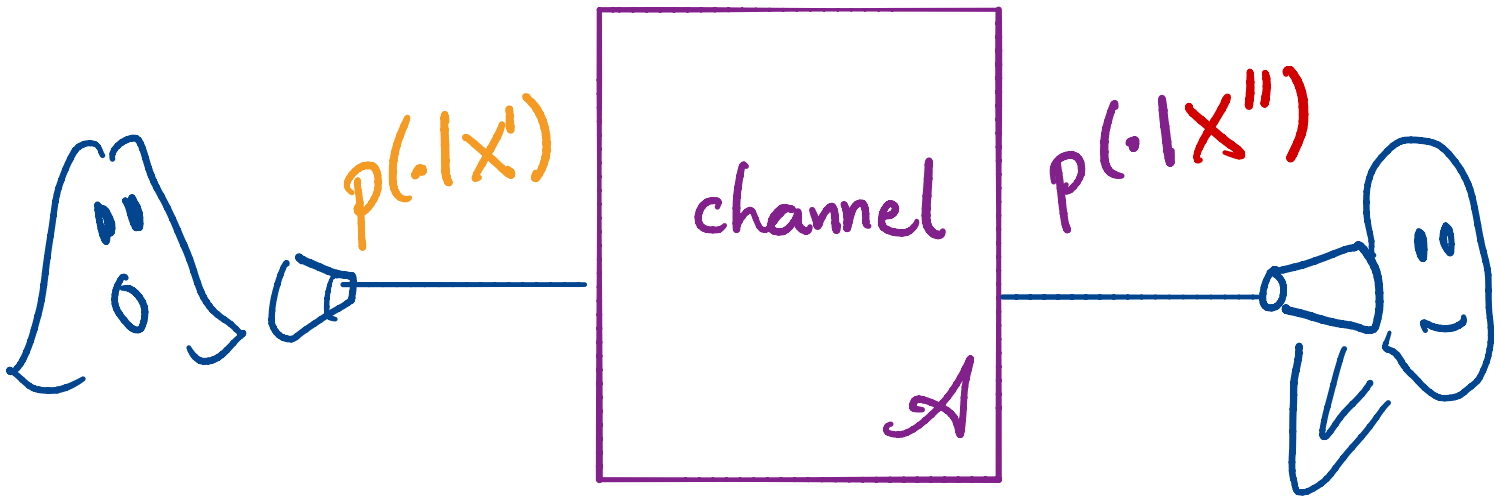

1000
0100
0000
0010
0001
1100
1010
1001
0110
0101
0011
0111
1011
1101
1110
1111
1000
0100
0000
0010
0001
1100
1010
1001
0110
0101
0011
0111
1011
1101
1110
1111


1000
0100
0000
0010
0001
1100
1010
1001
0110
0101
0011
0111
1011
1101
1110
1111
1000
0100
0000
0010
0001
1100
1010
1001
0110
0101
0011
0111
1011
1101
1110
1111

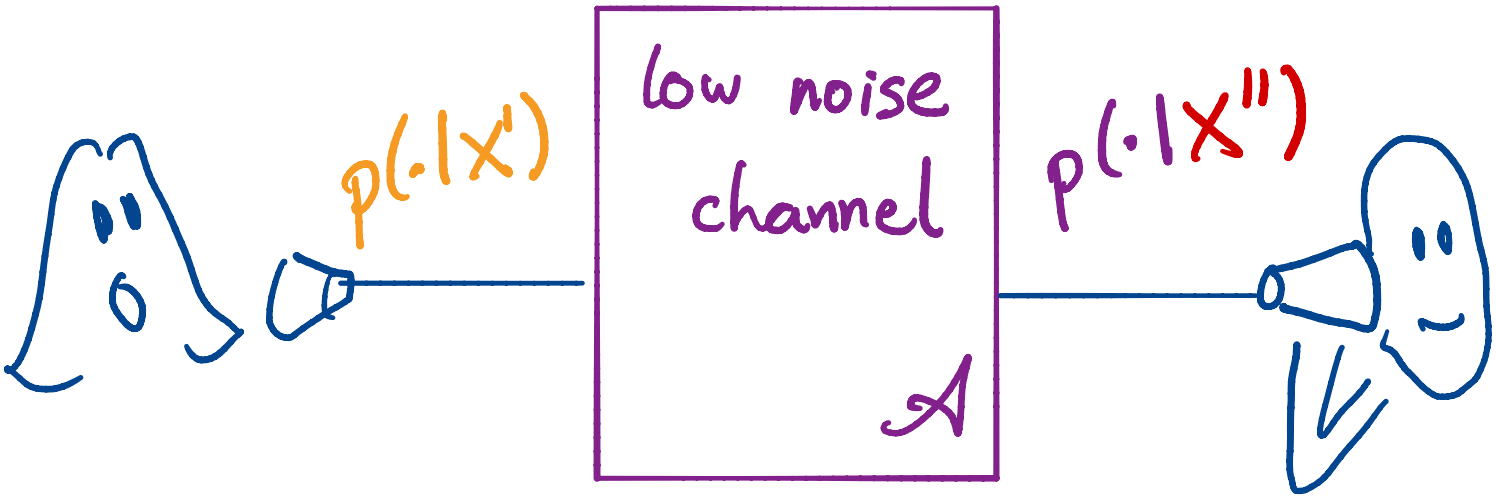
For a noisy channel,
the max # of msgs
is defined by the
capacity $\max_{p} I(S;\hat{S})$

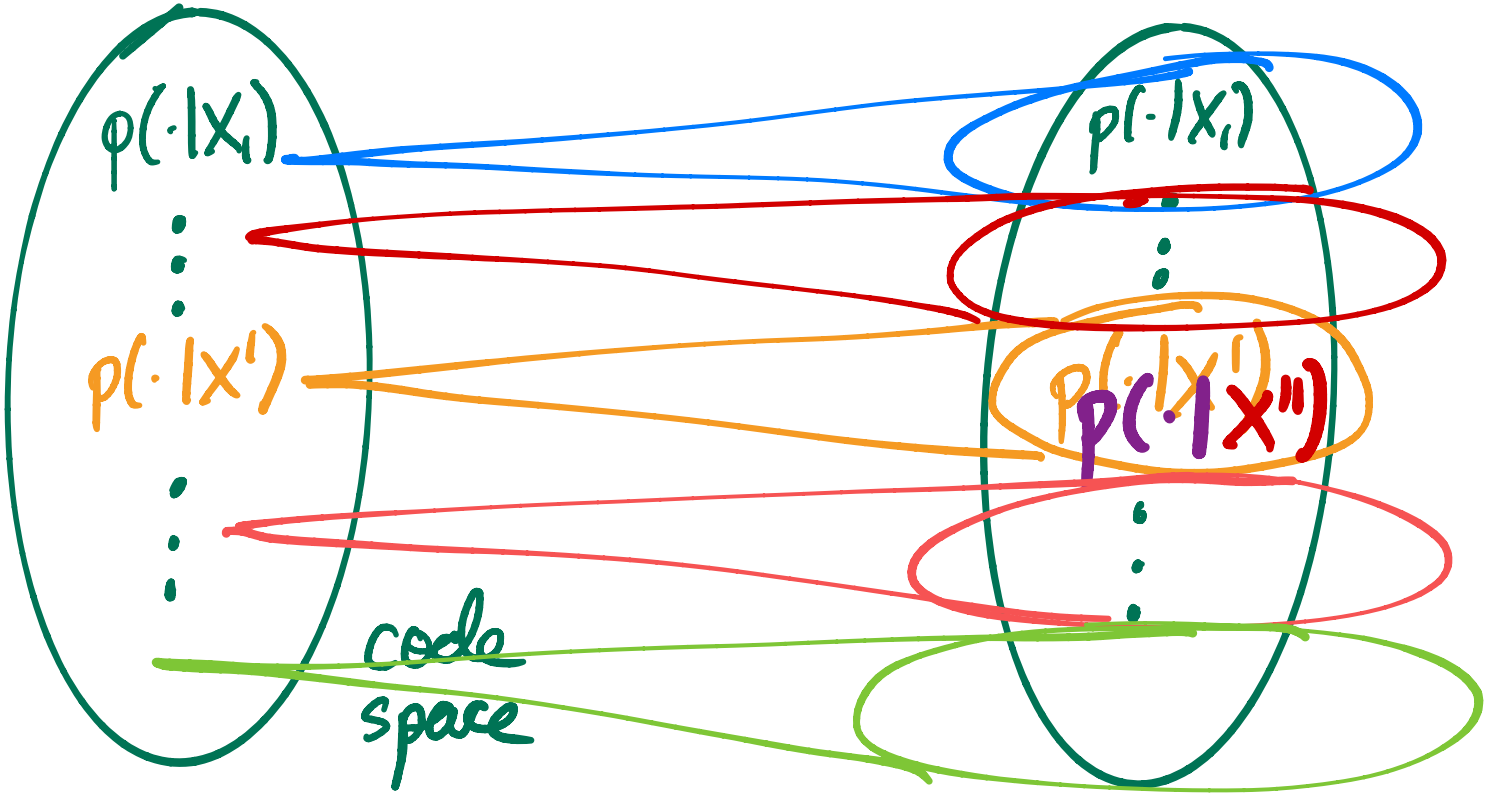
The max # of msgs is 2^{nc}

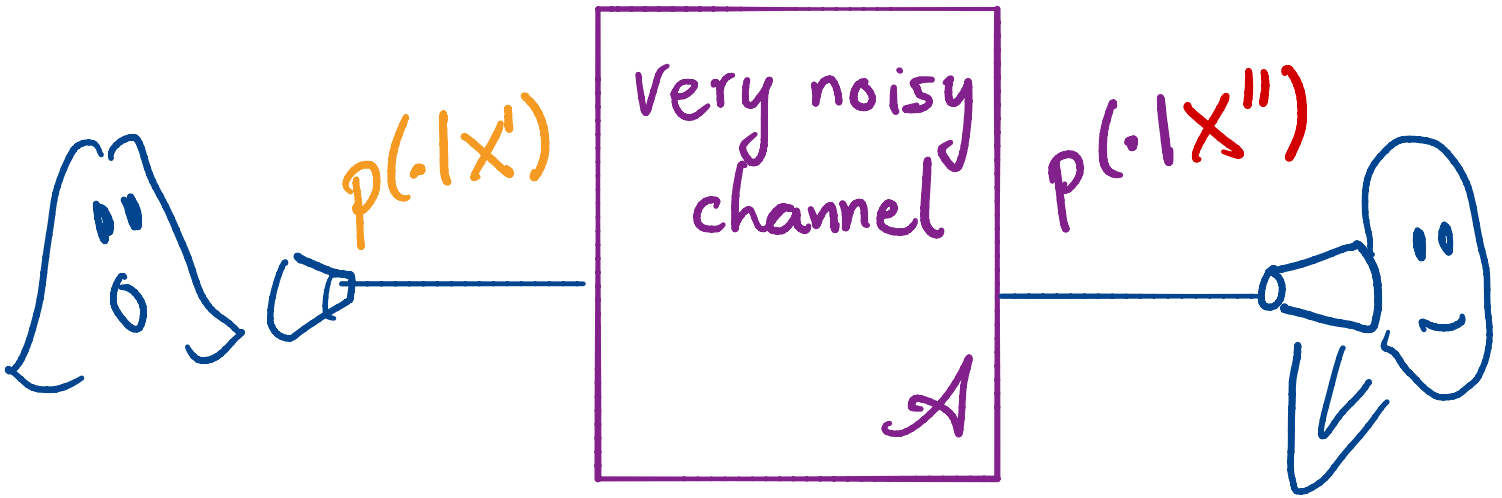
A channel can be evaluated by its channel capacity

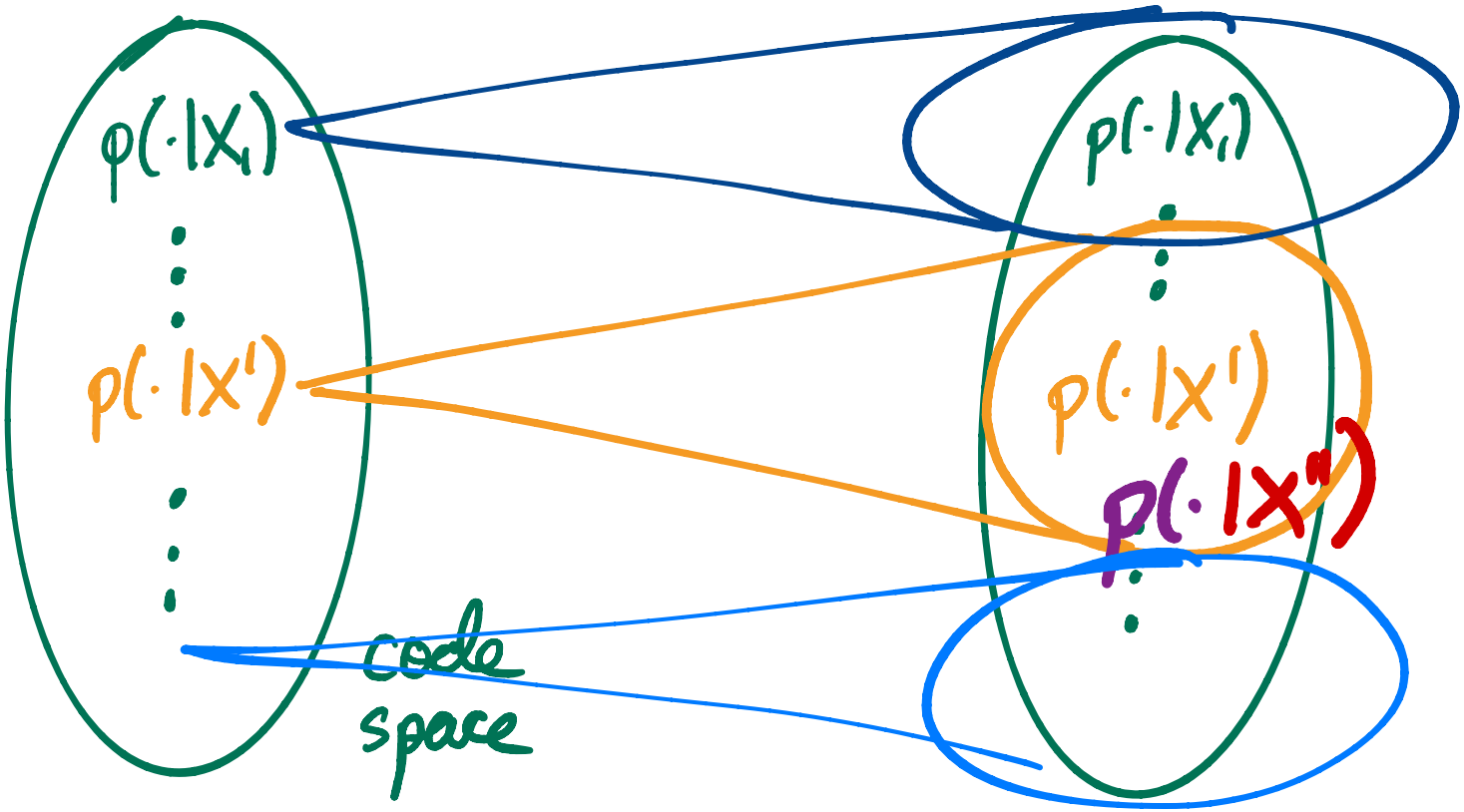


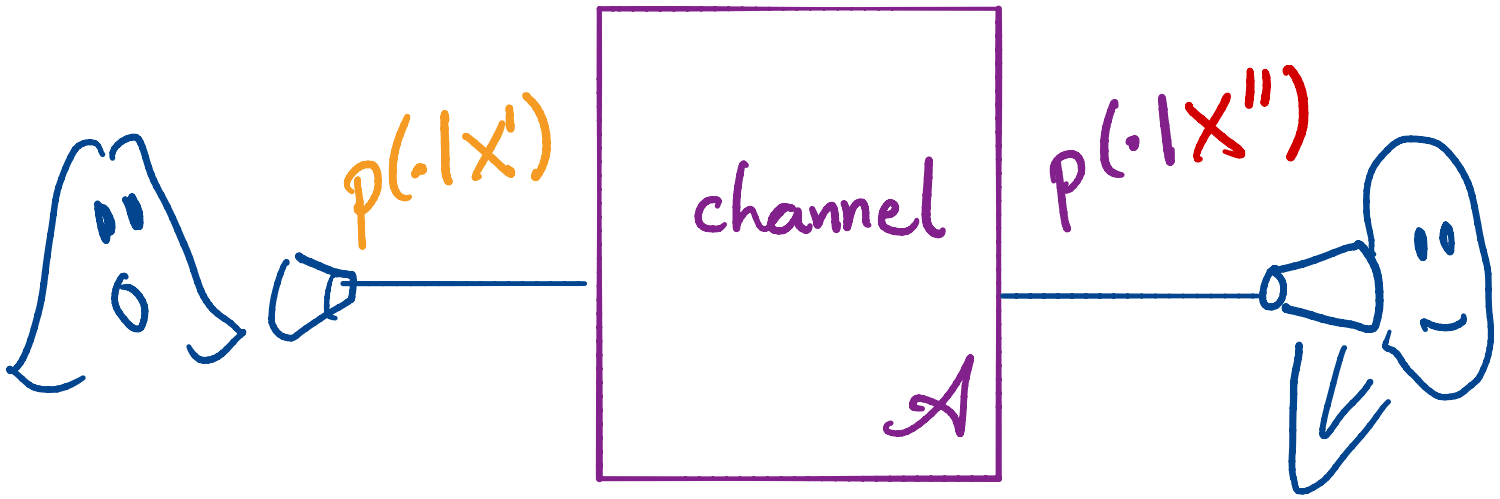
Algorithms are also channels!


low noise
channel
$p(\cdot|x')$
$p(\cdot|x'')$
$\mathcal{A}$


$p(\cdot|x_1)$
$p(\cdot|x')$
$p(\cdot|x_1)$
$p(\cdot|x')$
$p(\cdot|x'')$
code
space


p(·|x')
very noisy channel
A
p(·|x'')


p(·|x1)
p(·|x')
code space
p(·|x1)
p(·|x')
p(·|x'')



An algo is evaluated by its "channel's capacity"

Shannon's coding thm

Aim for **channels** with high **capacity**

More capacity
$\Rightarrow$ more messages

Posterior agreement

Aim for **algorithms** with high **exp. log. post. agr.**

Higher exp. log. PA
$\Rightarrow$ more messages

Organization

What is PA?

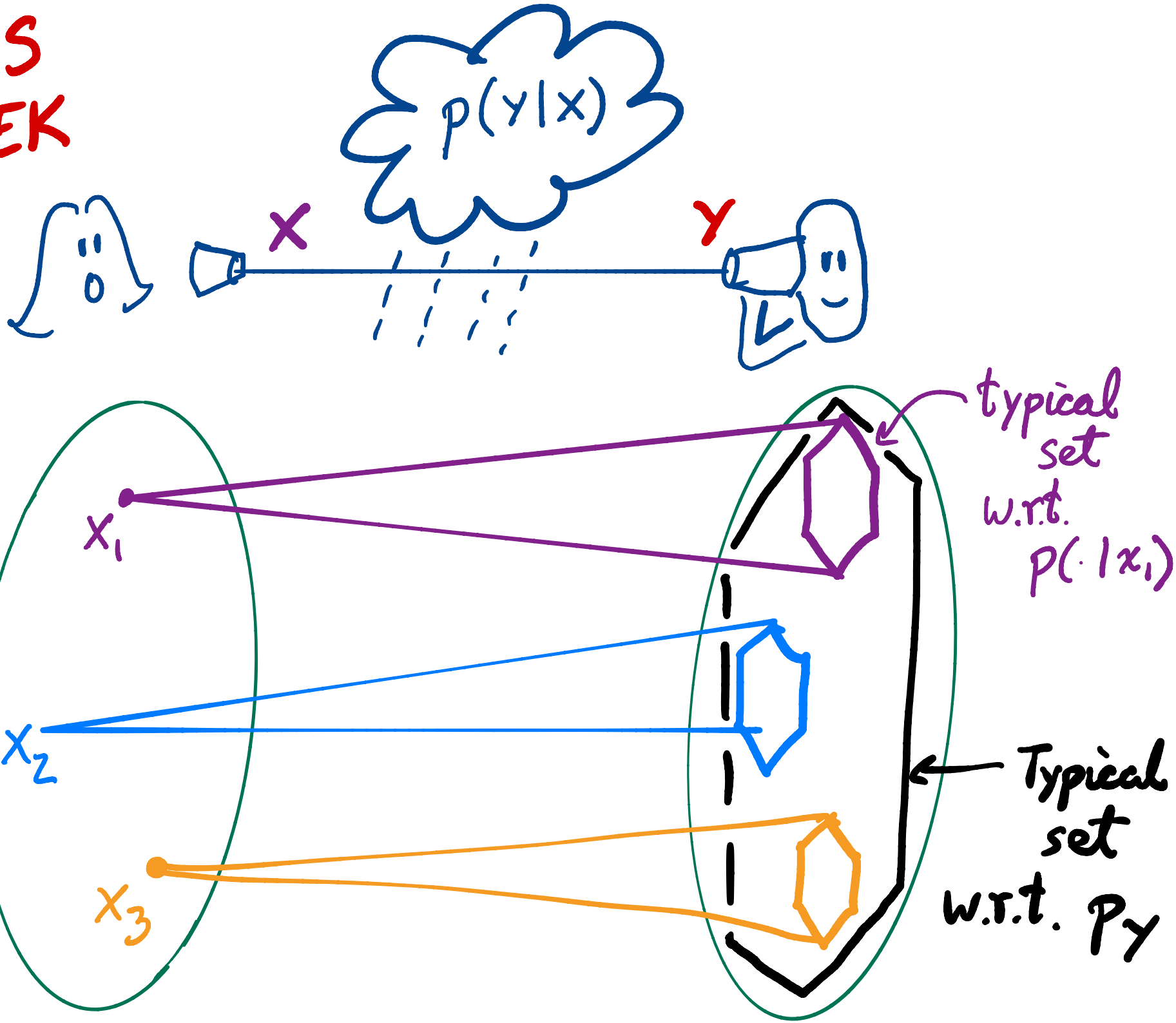
Rationale of PA

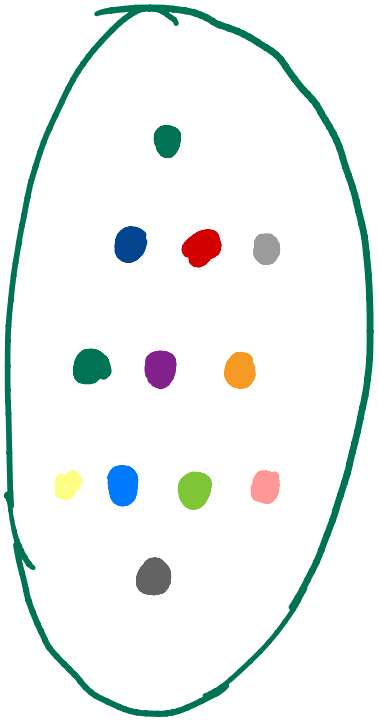
Roadmap

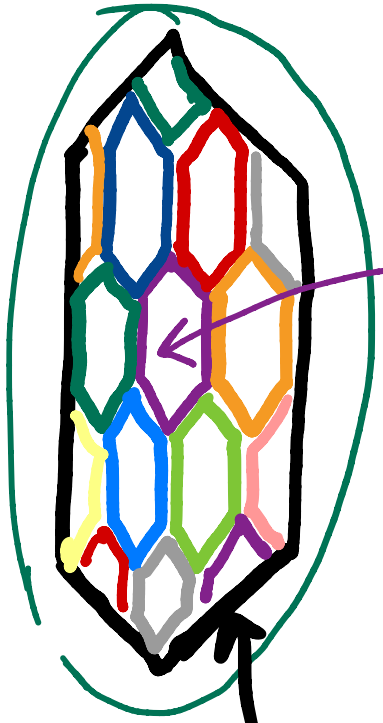
Shannon's coding theorem

Formalization of PA

THIS WEEK







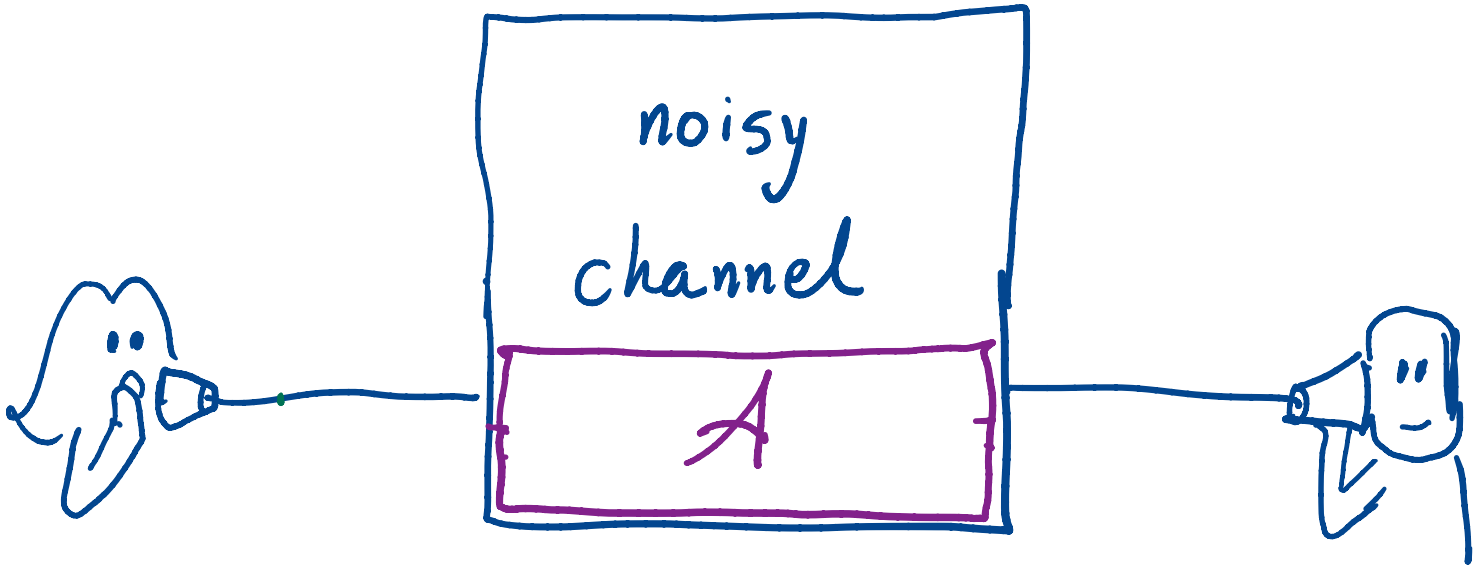
Typical set

$\approx 2^{nH(Y|X)}$

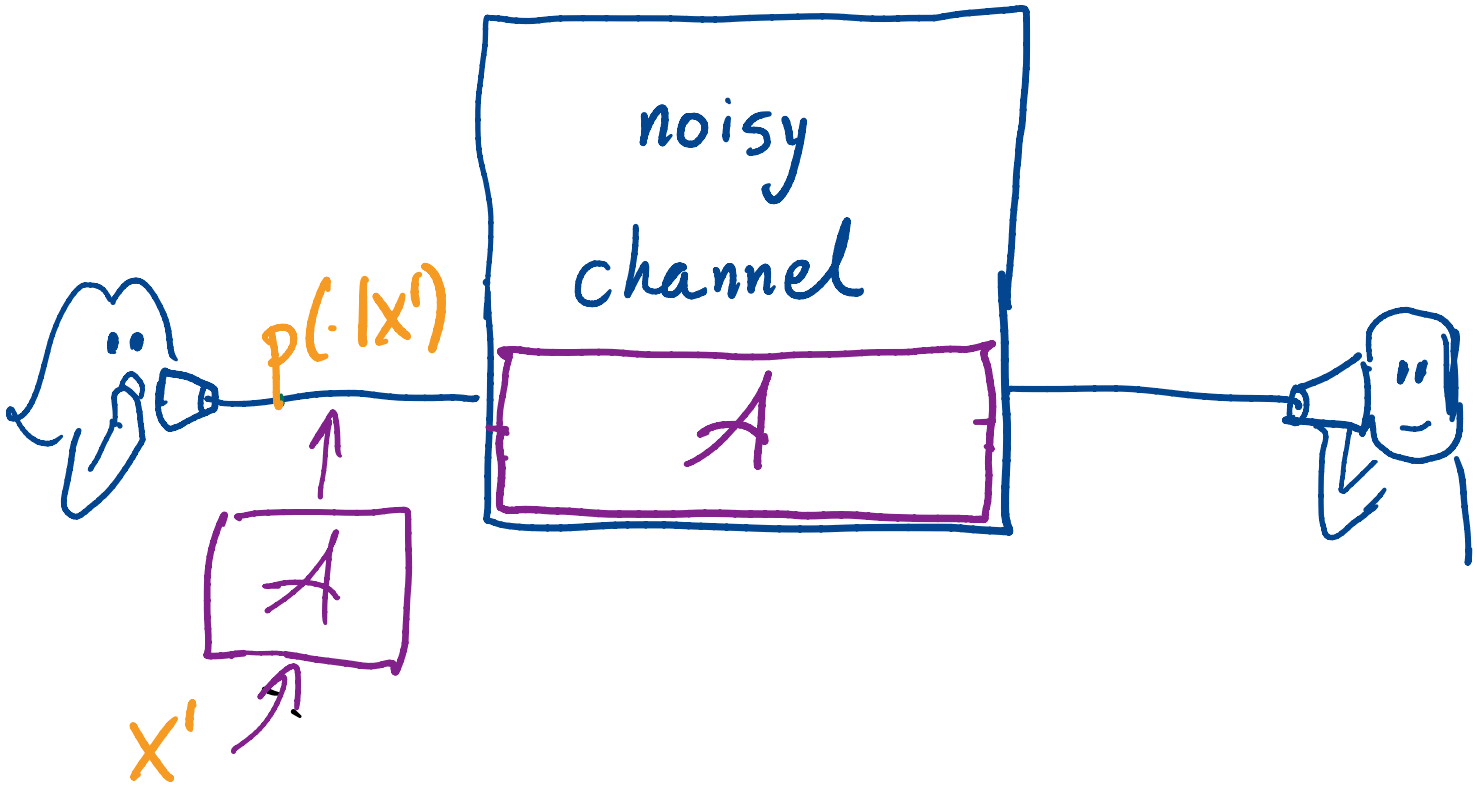
Typical set $\approx 2^{nH(Y)}$

$$\text{Optimal \# of msgs} \leq \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X;Y)}$$

$$= 2^{nC}$$

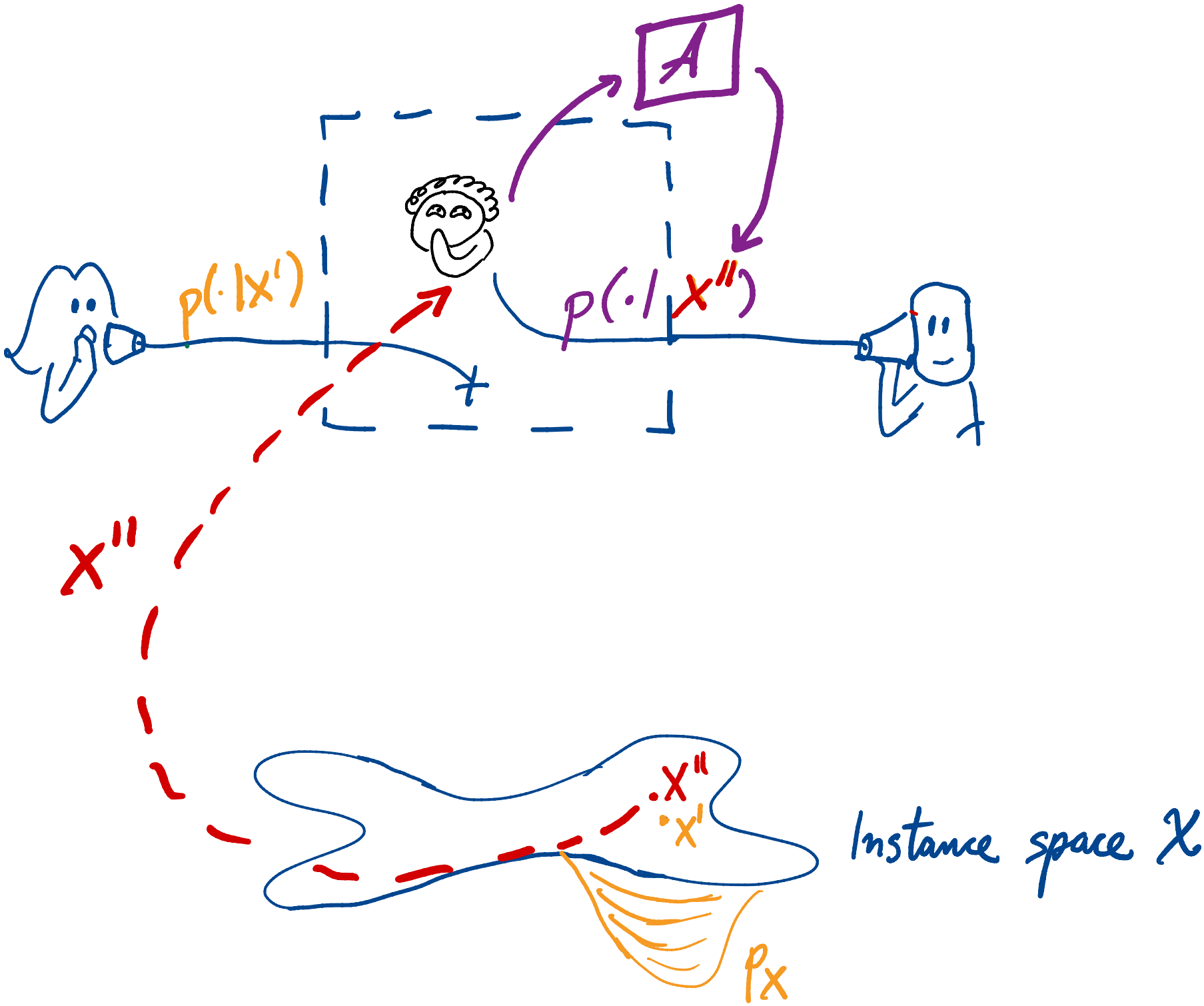


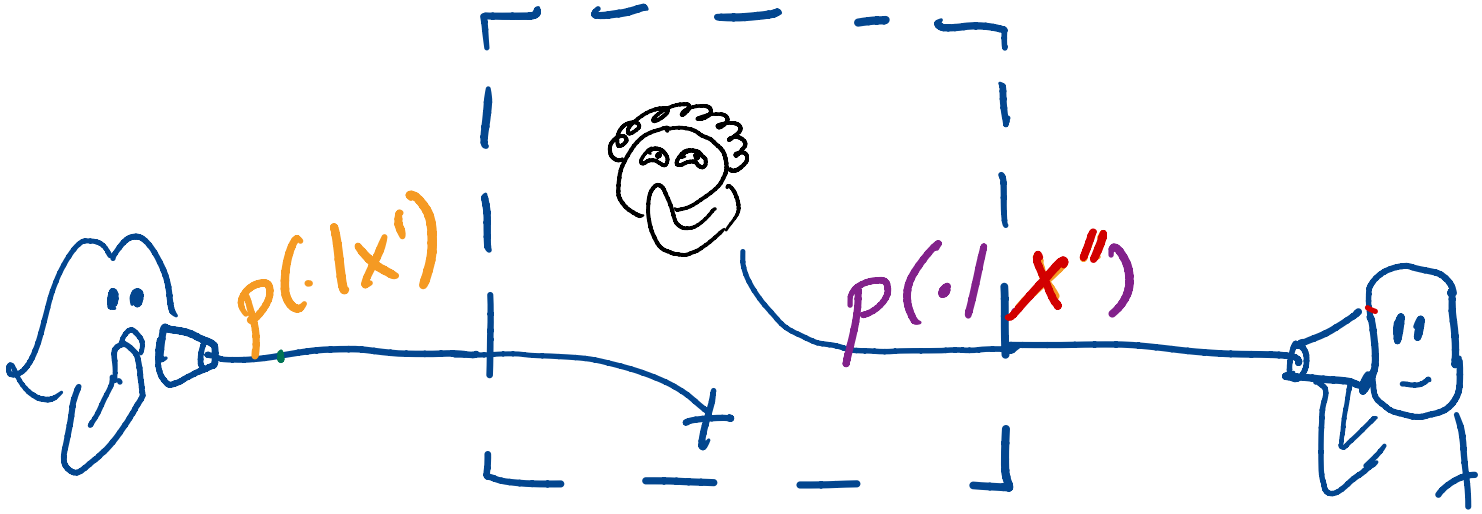
NEXT WEEK

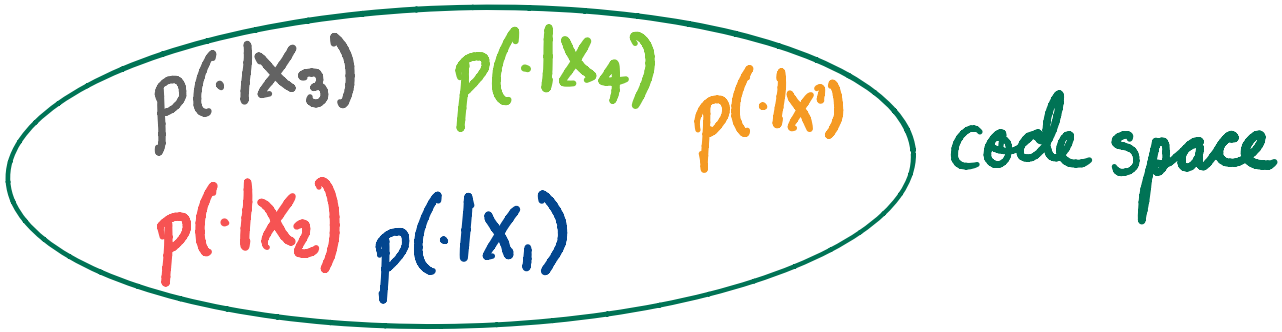


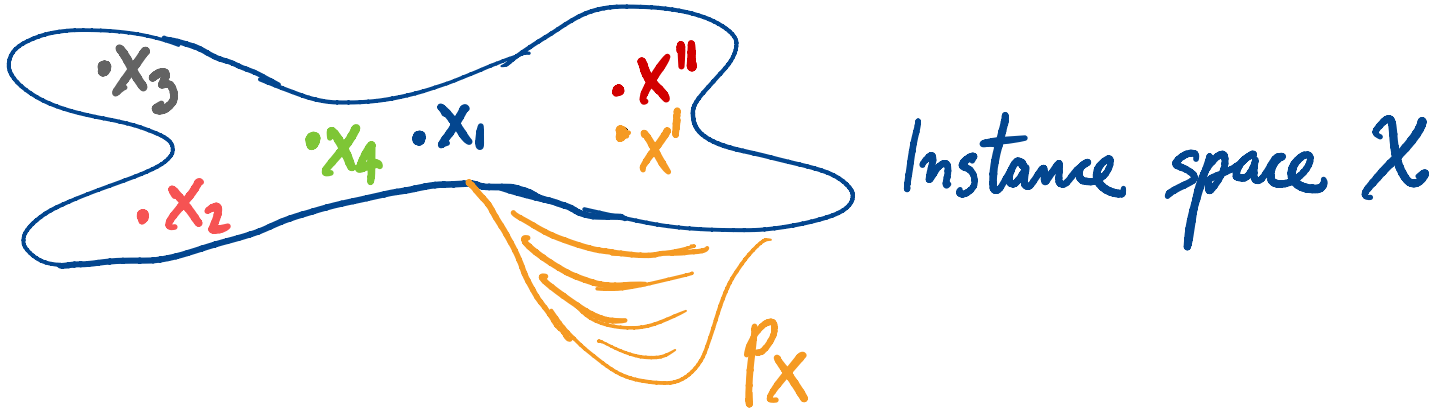
Instance space $\mathcal{X}$
(all instances of "size" n)

NEXT WEEK

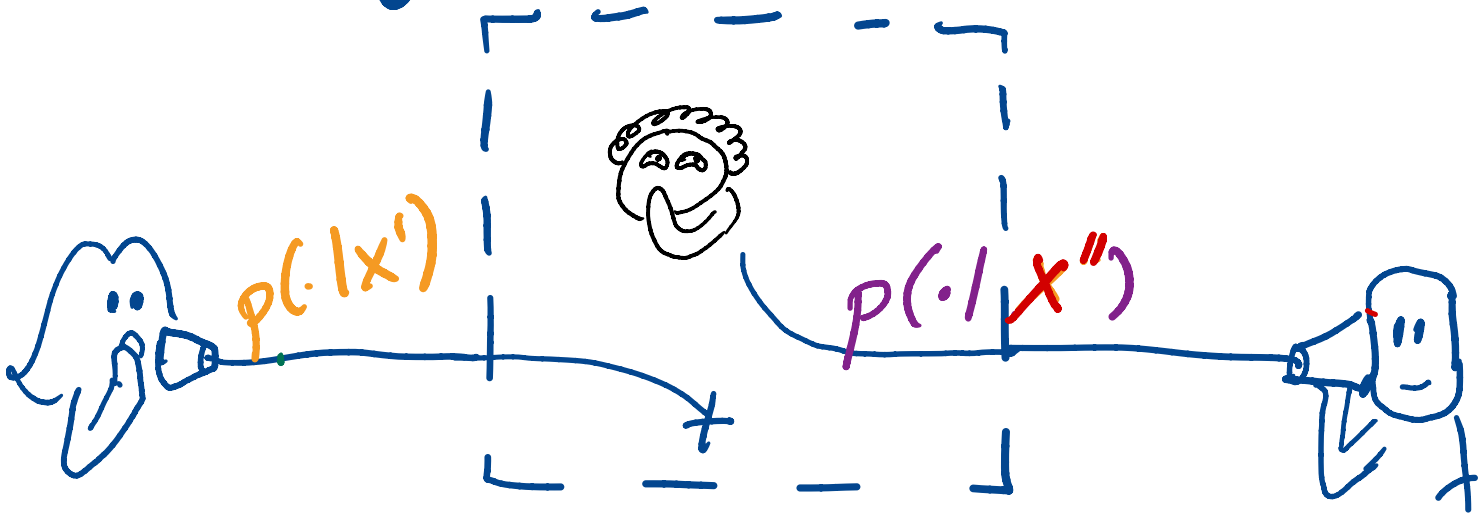

A
p(·|x')
p(·|x'')
x''
x''
x'
Instance space X
p_X

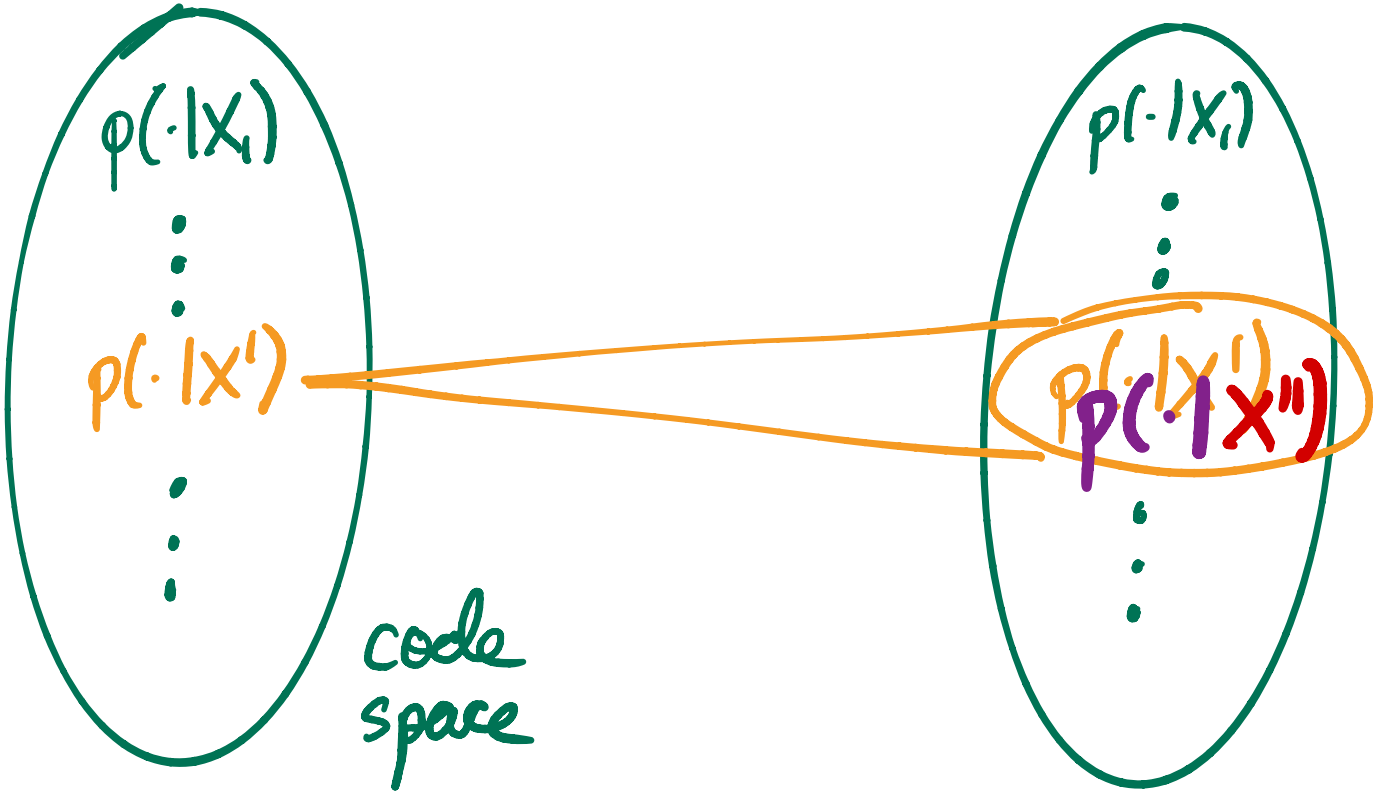

p(·|x')
p(·|x'')


p(·|x3)
p(·|x4)
p(·|x')
p(·|x2)
p(·|x1)
code space

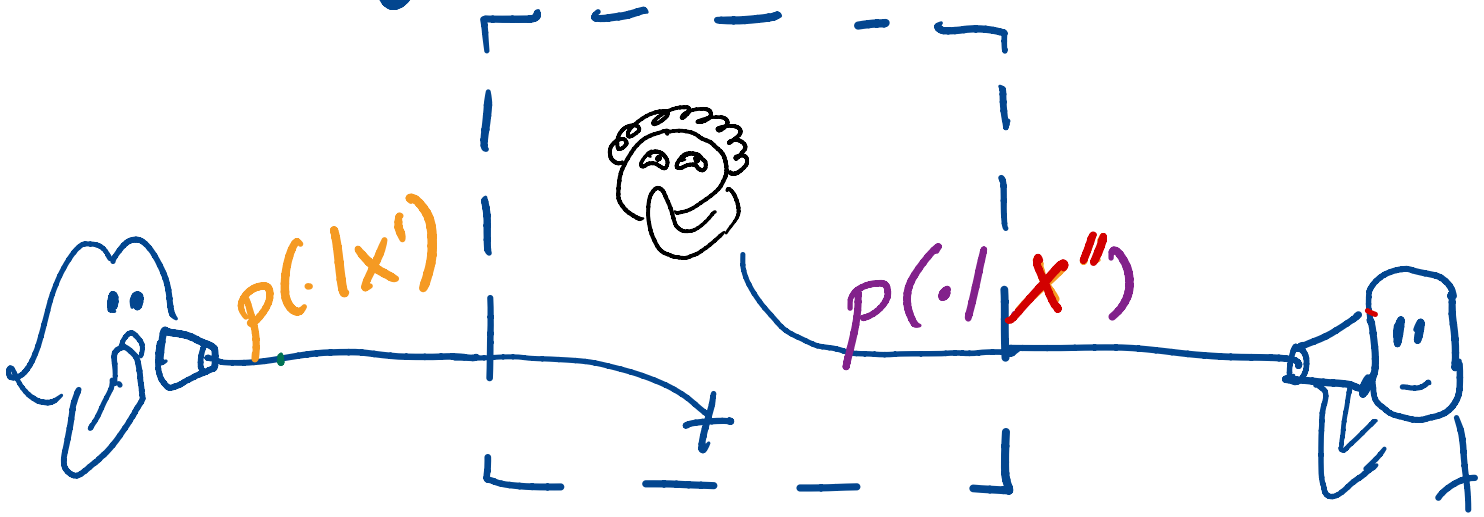

x3
x4
x1
x''
x'
x2
px
Instance space X

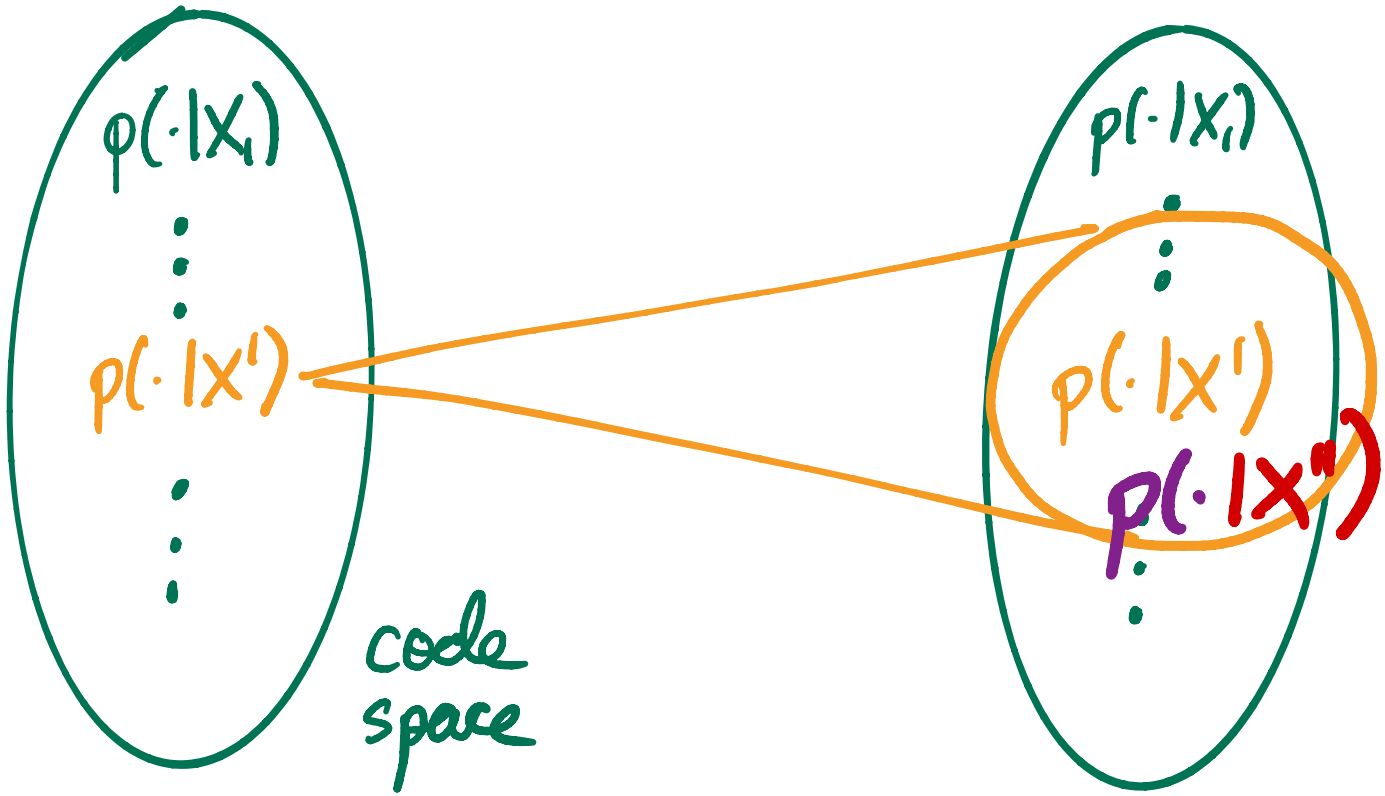
If the algo is robust...

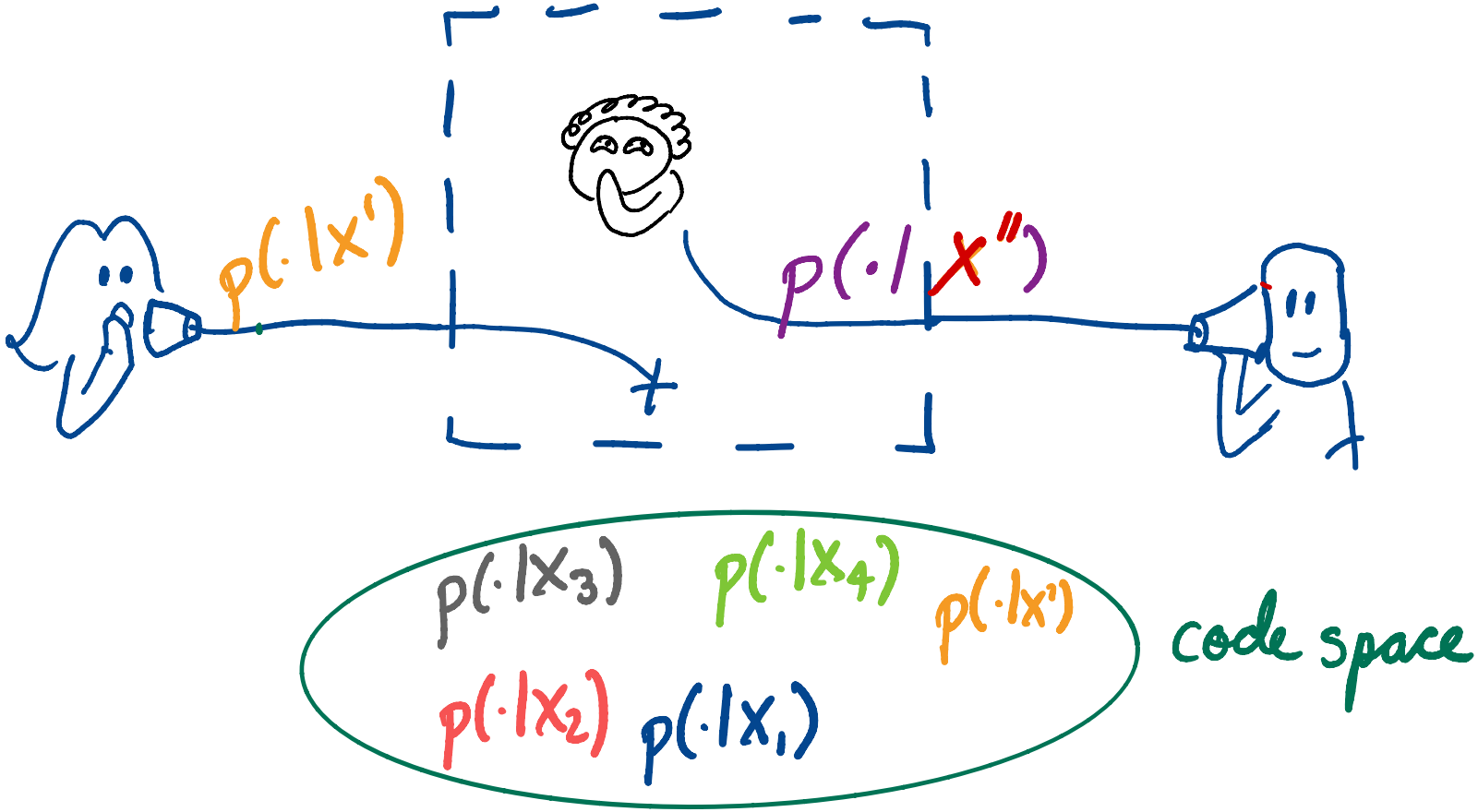




If the algo is not robust...







Channel capacity = exp log. post. agr.

We prove later that, with high probability,

$$\text{prob fail} \leq \text{const} \cdot \exp\left(-\log|C|\left(\exp.\log PA - \frac{\log m_n}{\log|C|} - \varepsilon\right)\right)$$

(m_n: # of msgs)

So, if $\exp.\log PA > \frac{\log m_n}{\log|C|}$ then

$$\text{prob. fail} \xrightarrow[n\to\infty]{\text{in prob.}} 0$$

$$\text{const} \cdot \exp\left(-\log|\mathcal{C}|\left(\exp.\log PA - \frac{\log m_n}{\log|\mathcal{C}|} - \varepsilon\right)\right)$$

An algorithm can reduce prob of fail by:

maximizing exp. log. post. agreement.

This allows a large number m_n of msgs that sender can communicate to the receiver.

Organization

What is PA?

Rationale of PA

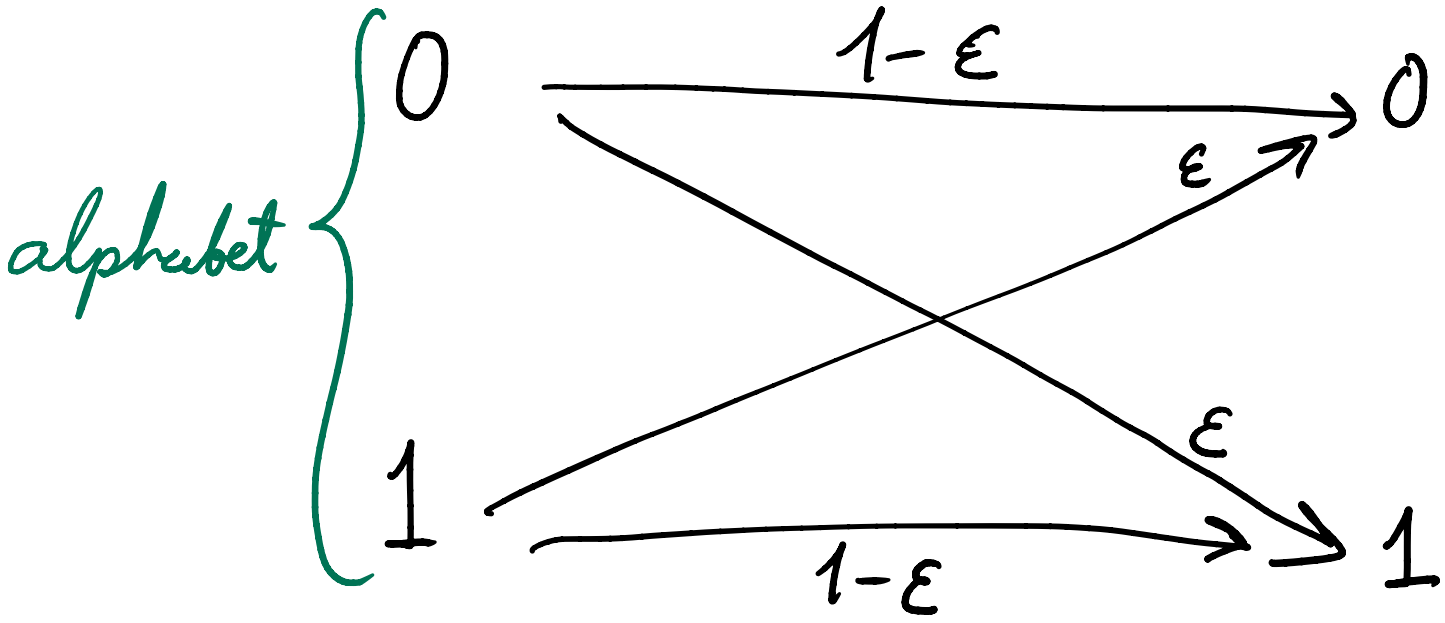
Roadmap

Shannon's coding theorem

Formalization of PA

Shannon's channel coding theorem

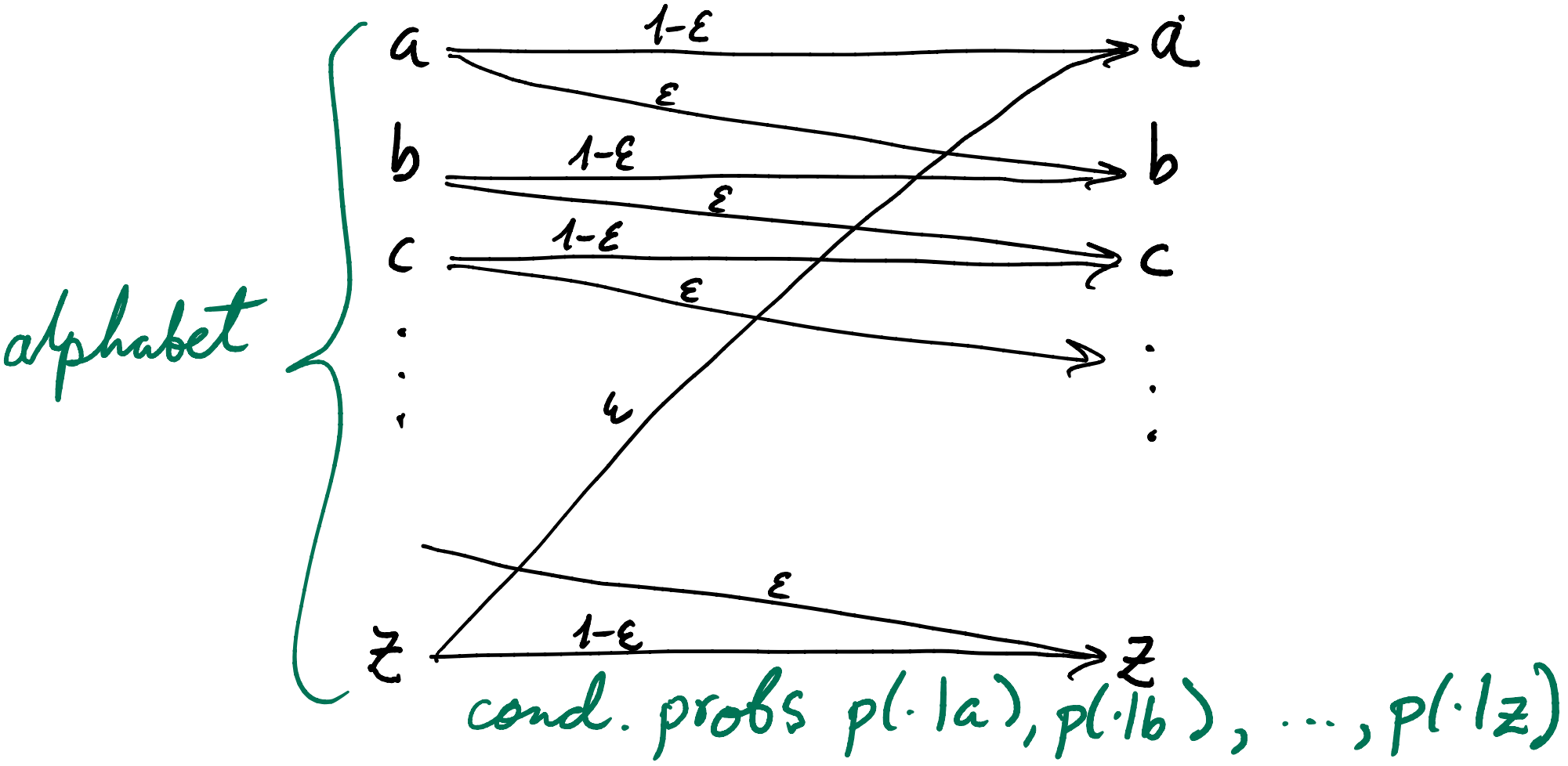
Channels



condi. probs $p(\cdot|0), p(\cdot|1)$

Channels

noisy typewriter



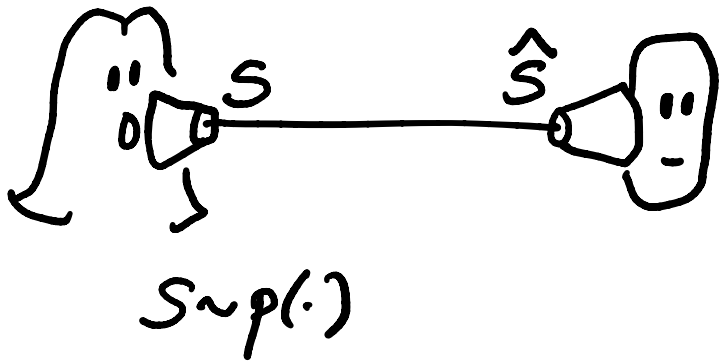
Channel capacity

$$\max_{p(\cdot)} I(S;\hat{S})$$

output

$p(\cdot)$: distr over alphabet

S: random input



Observe: Typewriter's capacity > binary channel's capacity

Channel capacity

* The max rate in bits / transmission with arbitrarily low prob. of error.

* max # of different messages that can be communicated $= 2^{n \text{Capacity}}$

Quick recap

$$H(X) := -\sum_x p(x) \log p(x)$$ "Info in X"

$$H(X|Y) := -\sum_y p(y) \sum_x p(x|y) \log p(x|y)$$ "residual info left in X after learning Y"

$$I(X;Y) := \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$ "Info that Y reveals about X"

$H(X) = ?$

Quick recap

$$H(X) := -\sum_x p(x) \log p(x)$$ "Info in X"

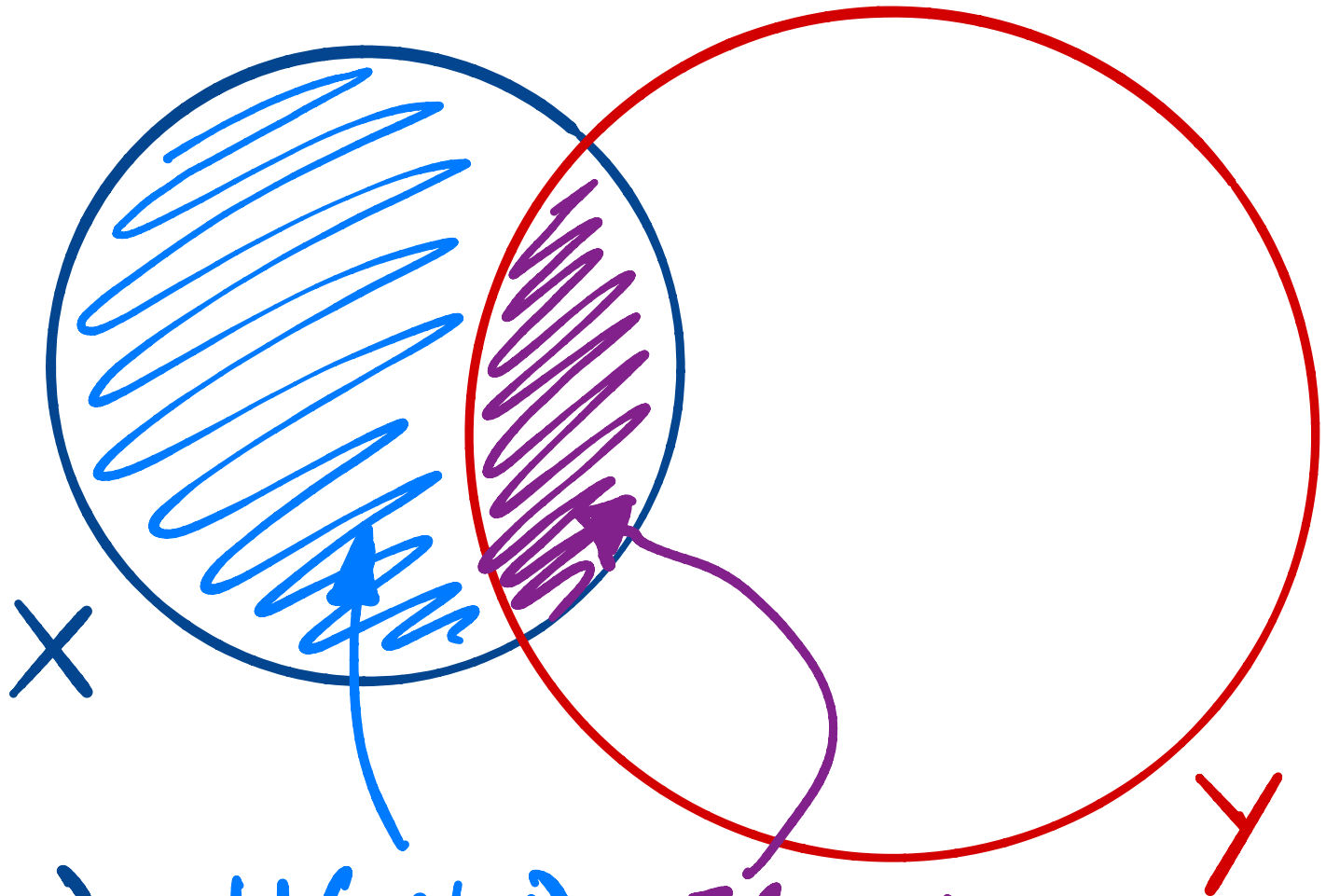
$$H(X|Y) := -\sum_y p(y) \sum_x p(x|y) \log p(x|y)$$ "residual info left in X after learning Y"

$$I(X;Y) := \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$ "Info that Y reveals about X"

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

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

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



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

Proof:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

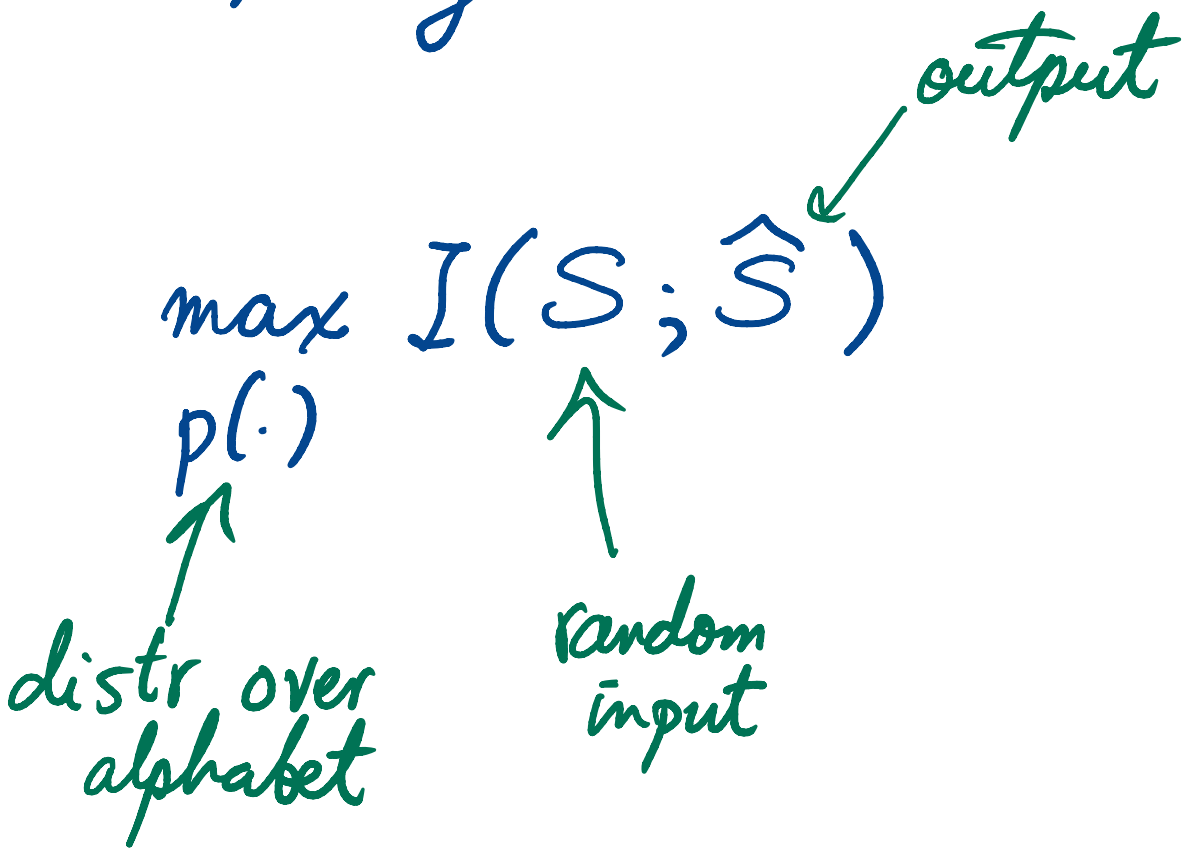
$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)p(y)}{p(x)p(y)}$$

$$= \sum_{x,y} p(x,y) \log p(x|y) - \sum_{x,y} p(x,y) \log p(x)$$

$$= \underbrace{\sum_{y} p(y) \sum_{x} p(x|y) \log p(x|y)}_{} - \underbrace{\sum_{x} p(x) \log p(x) \sum_{y} p(y|x)}_{1}$$

$$= -H(X|Y) + H(X).$$

Channel capacity



Observe: Typewriter's capacity > binary channel's capacity

Examples:

Zero-noise bin channel

$$\max_{p} I(S;\hat{S}) = ?$$

Examples:

Zero-noise bin channel

$$\max_{p} I(S;\hat{S}) = \max_{p} H(S) - H(S|\hat{S}) \to 0$$

$$= \max_{p} H(S) = 1$$

1 bit per transmission

Zero-noise typewriter

$$\max_{p} I(S;\hat{S}) = ?$$

Zero-noise typewriter

$$\max_{P} I(S;\hat{S}) = \max_{P} H(S) - H(S|\hat{S})$$

$$= \max_{P} H(S) = \log 26 \simeq 4.7$$

4.7 bits per transmission

Noisy typewriter with $\varepsilon = 0.5$

$$\max_{p} I(S;\hat{S}) = ?$$

Noisy typewriter with $\varepsilon = 0.5$

$$\max_p I(S;\hat{S}) = \ldots = \log(26/2)$$

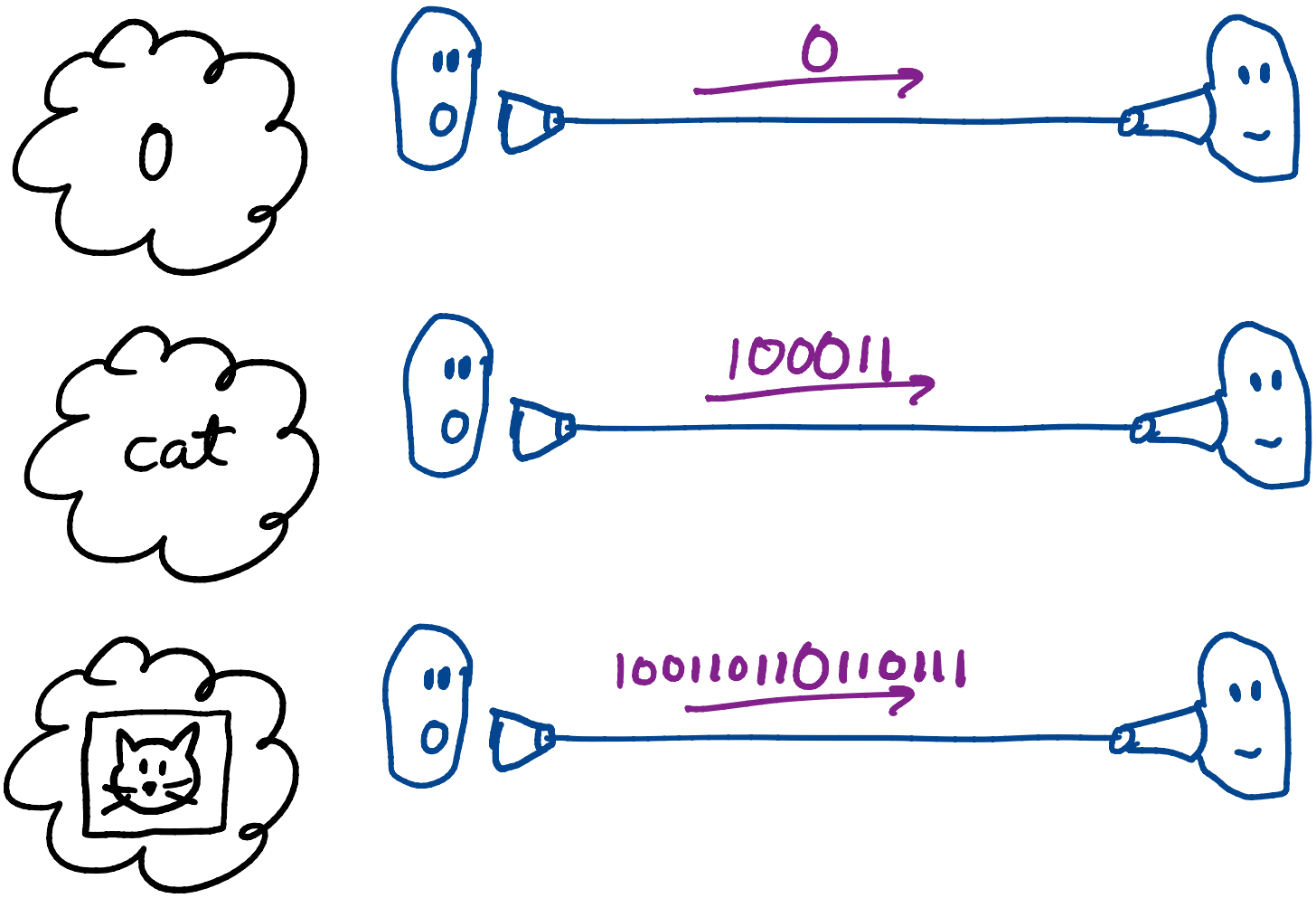
$$= \log 26 - 1$$

$$\approx 4.7 - 1$$

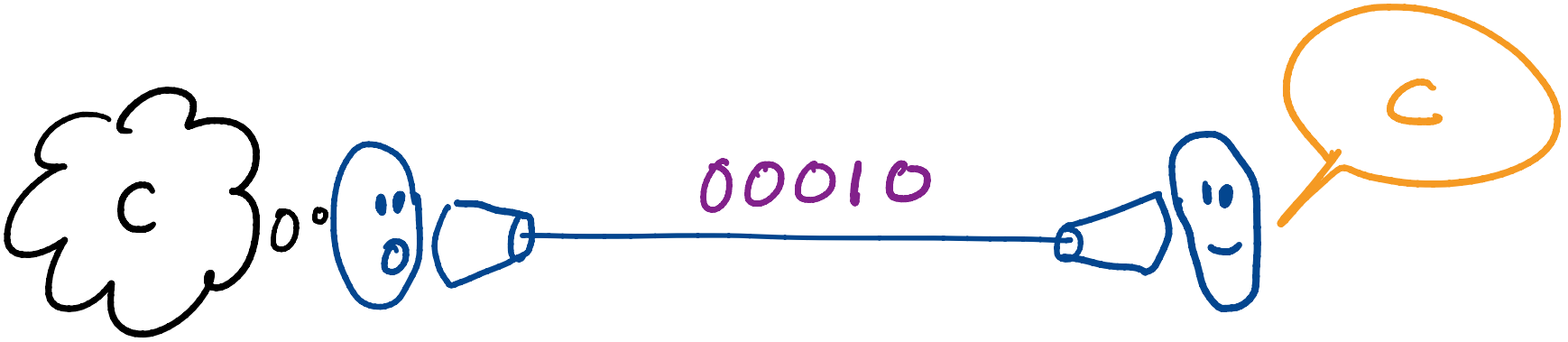
$$= 3.7$$

3.7 bits per transmission

Codes:

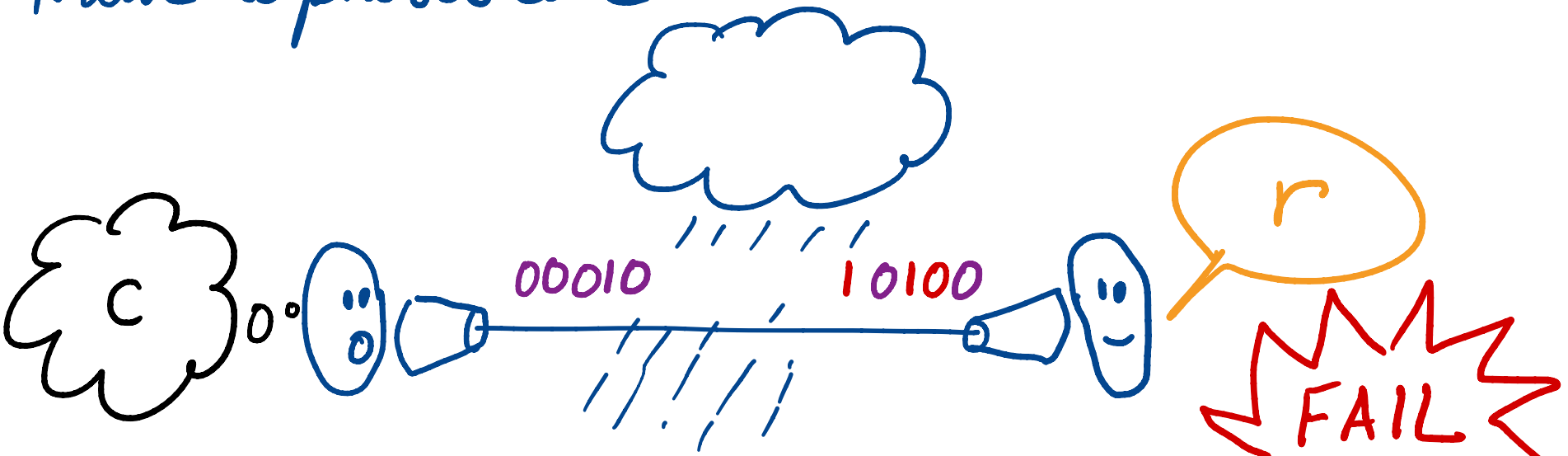


Naïve alphabet code



a	00000 ← codewords
b	00001
c	00010
⋮	
z	11001

Naïve alphabet code

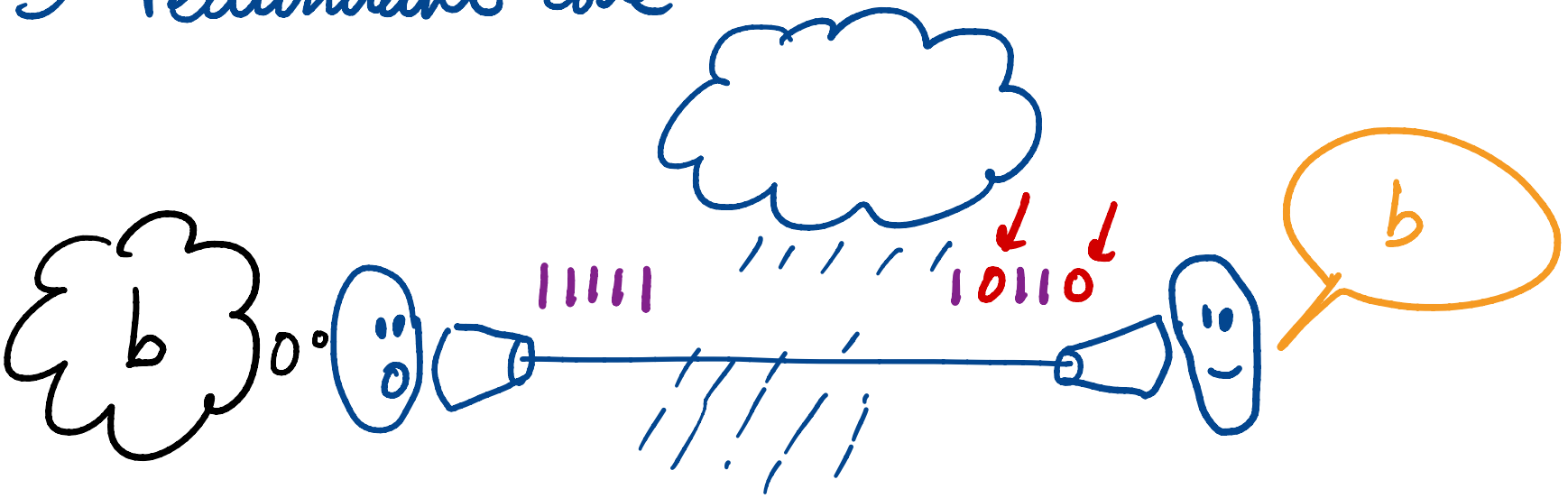


a	00000
b	00001
c	00010
⋮	
z	11010

Info theory has come up with robust codes...

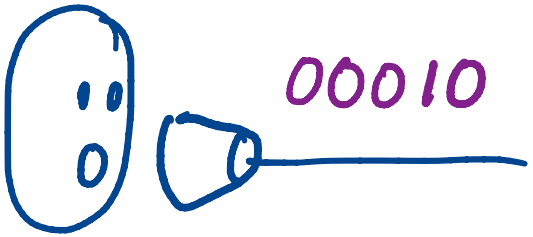
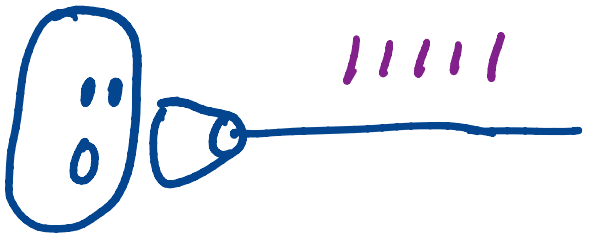
... but at the cost of less bits per transmission

5-redundant code



a 00000

b 11111

	Naïve	n-redundant
		
codeword length	n	n
# messages	2^n ☺	2 ☹
prob. fail $n \to \infty$	1 ☹	0 ☺
bits/trans $n \to \infty$	1 ☺	0 ☹

Goal: Max. rate as prob. fail $\to 0$ as $n \to \infty$

Shannon's coding thm

If # of msgs $= 2^{nr}$ (with $r < C$)

$\Rightarrow$ there is a code

prob. error $\xrightarrow[n \to \infty]{} 0$

(The converse is true)

This means $\Rightarrow$ max # msgs $= 2^{nC}$

$\Rightarrow$ capacity as a channel measure

Shannon's random code

$$
\begin{aligned}
&m_1 && s_1^n := s_{11}^n s_{12}^n \cdots s_{1n}^n \\
&m_2 && s_2^n := s_{21}^n s_{22}^n \cdots s_{2n}^n \\
&\vdots \\
&m_{2^{nr}} && s_{2^{nr}}^n := \cdots
\end{aligned}
$$

Shannon's random code

$m_1 \qquad S_1^n := S_{11}^n S_{12}^n \cdots S_{1n}^n$

$m_2 \qquad S_2^n := S_{21}^n S_{22}^n \cdots S_{2n}^n$

$\vdots$

$m_{2^{nr}} \qquad S_{2^{nr}}^n := \cdots$

$S_{ij}^n \xleftarrow{\$} P_S^* = \arg\max_P I(S;\hat{S})$

S_{ij} symbol

Every symbol of every codeword is chosen indep. at random from $P_S^* = \arg\max_P I(S;\hat{S})$

Shannon's random code

$$m_1 \qquad S_1^n := S_{11}^n S_{12}^n \cdots S_{1n}^n$$

$$m_2 \qquad S_2^n := S_{21}^n S_{22}^n \cdots S_{2n}^n$$

$$\vdots$$

$$m_{2^{nr}} \qquad S_{2^{nr}}^n := \cdots$$

$S_{ij}^n \xleftarrow{\$} P_S^* = \arg\max_P I(S;\hat{S})$

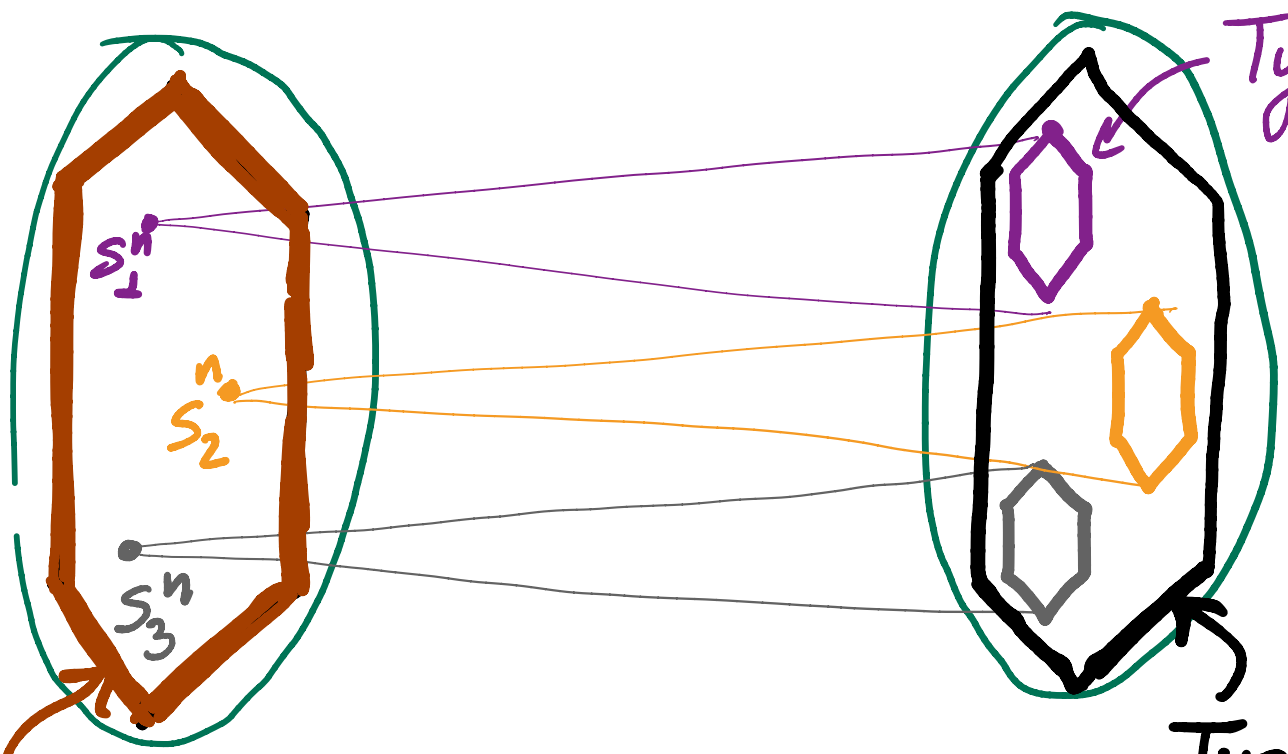
symbol

$S_i^n \xleftarrow{\$} P_{S^n}^* := \prod_{j \le n} P_S^*$

codeword

Every symbol of every codeword is chosen indep. at random from $P_S^* = \arg\max_P I(S;\hat{S})$

What's the max # of msgs we can communicate?

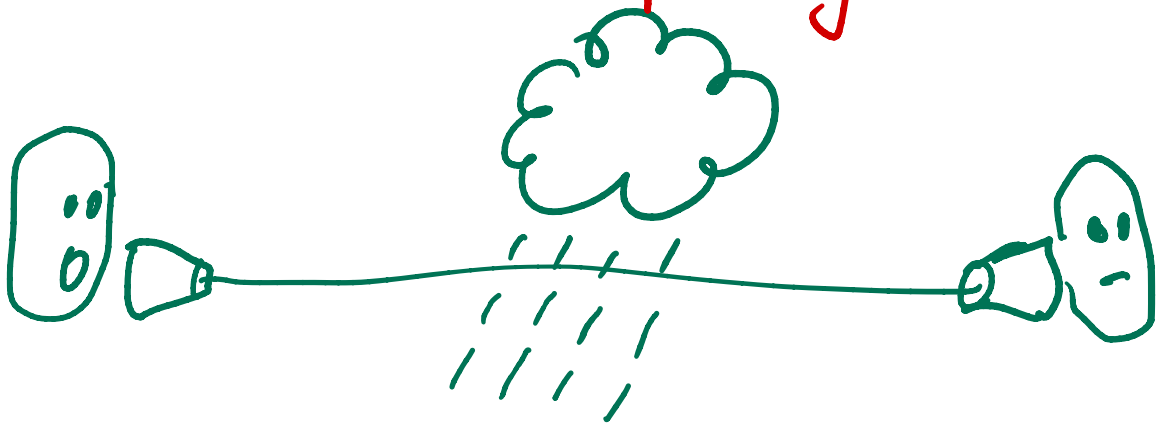


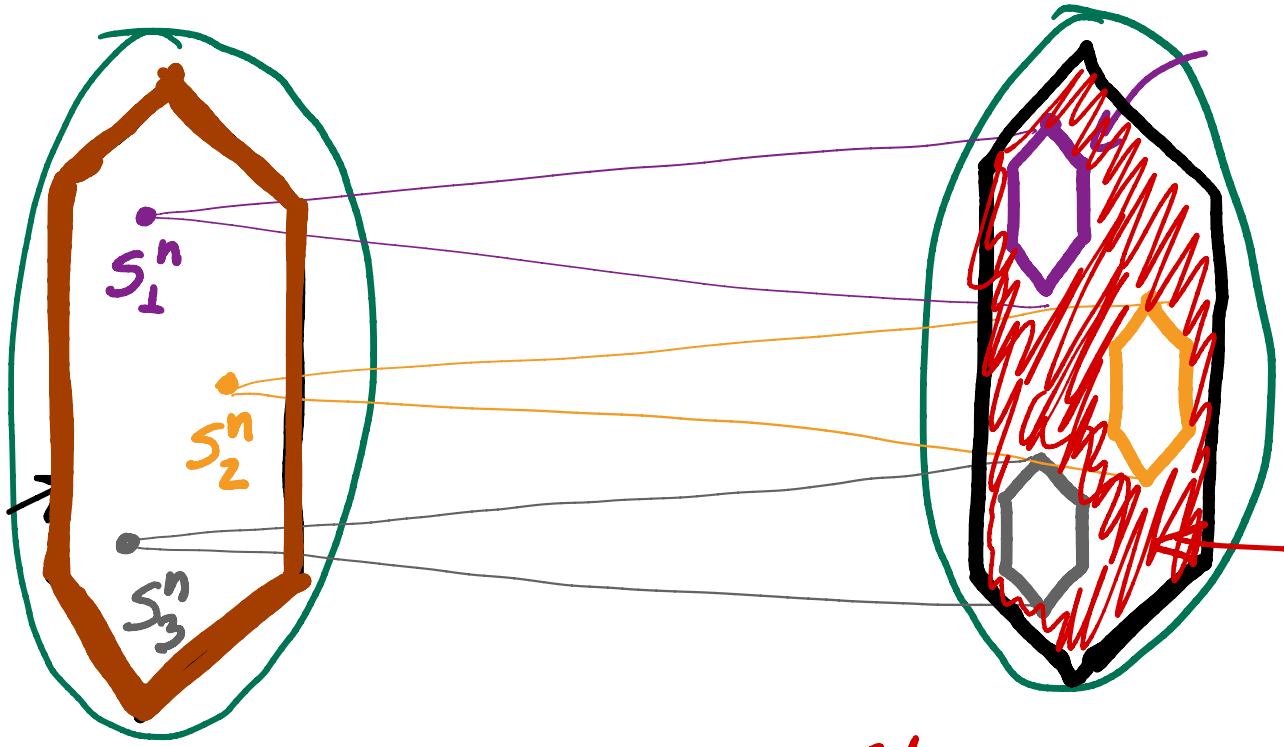
Typical set w.r.t. $P_{S|s}$

Typical set w.r.t. P_S^*

Typical set w.r.t. P_S

What's the max # of msgs we can communicate?

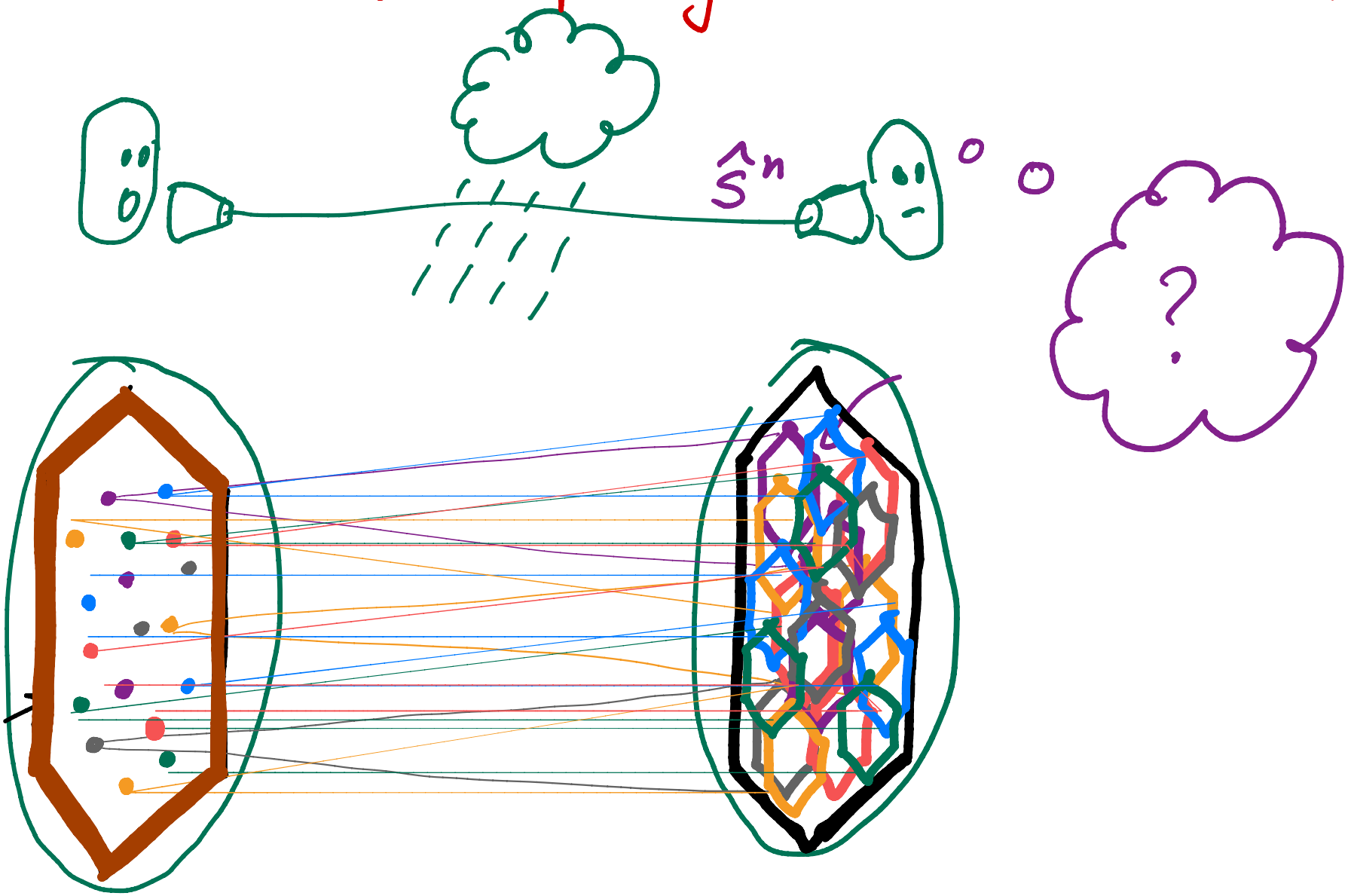




WASTED

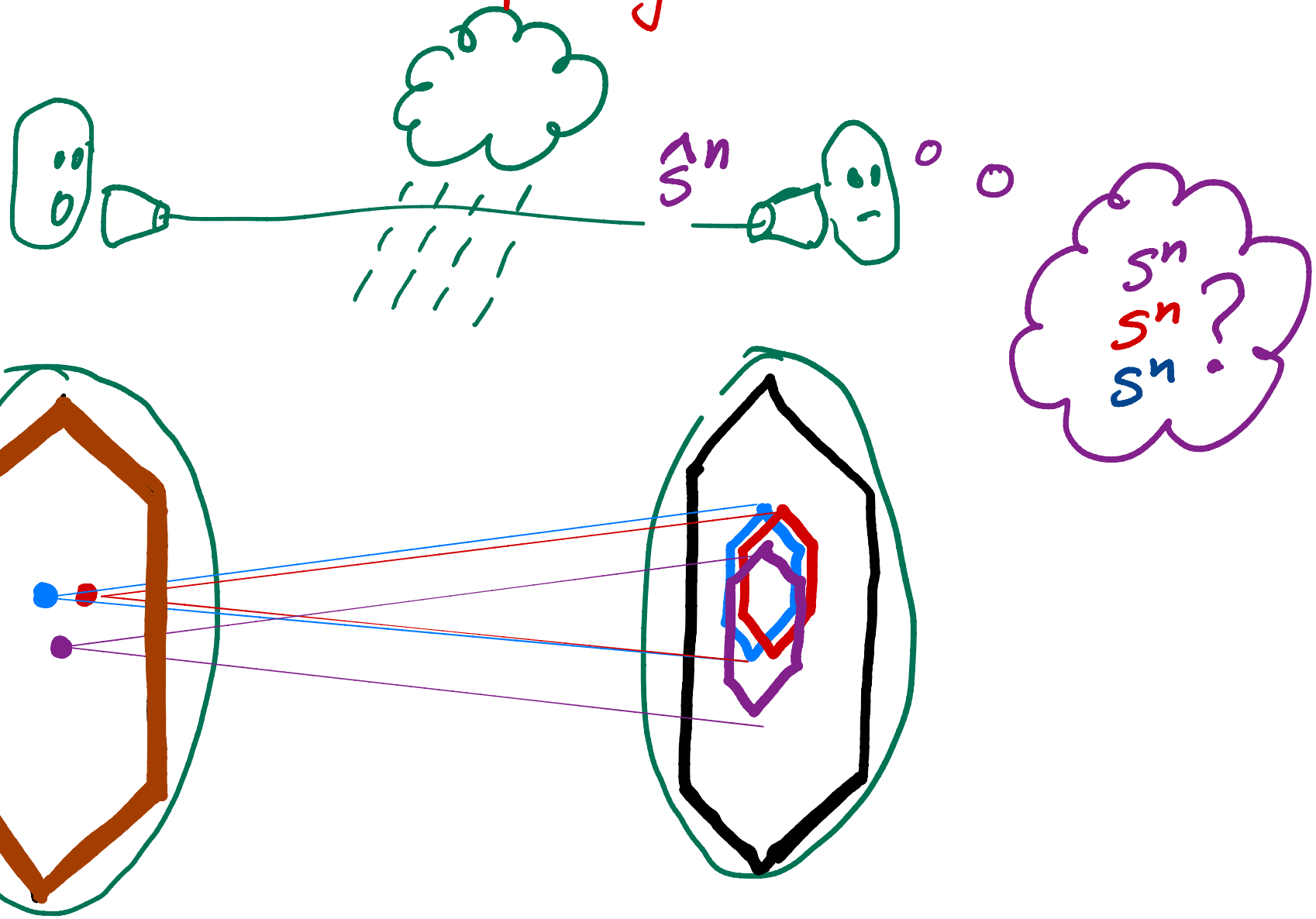
You can send more msgs!

What's the max # of msgs we can communicate?

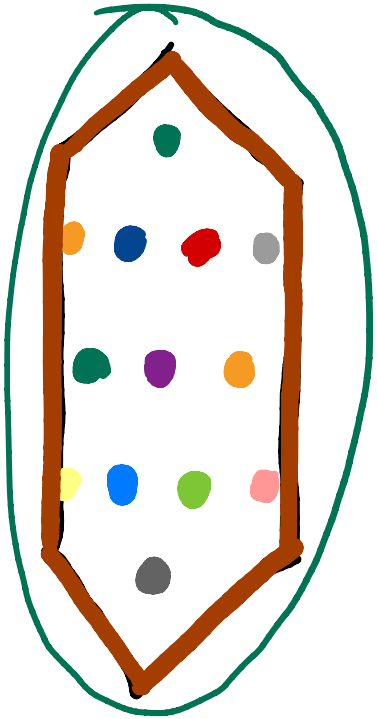


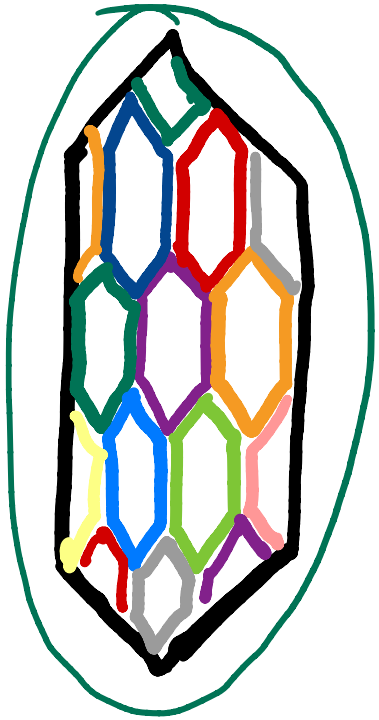
Too many messages!

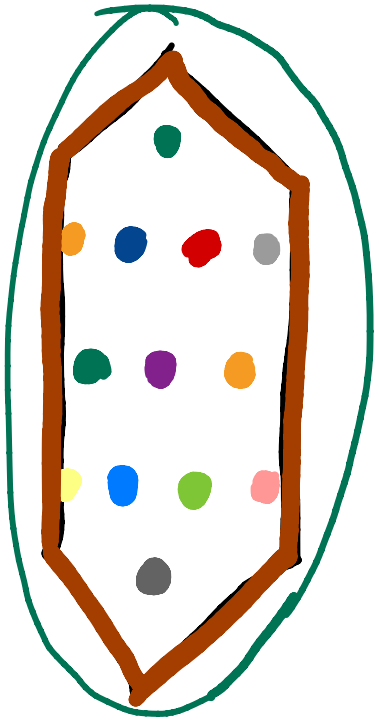
What's the max # of msgs we can communicate?

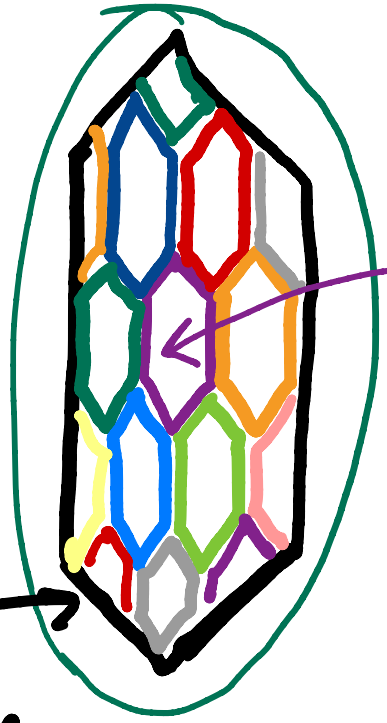


What's the max # of msgs we can communicate?









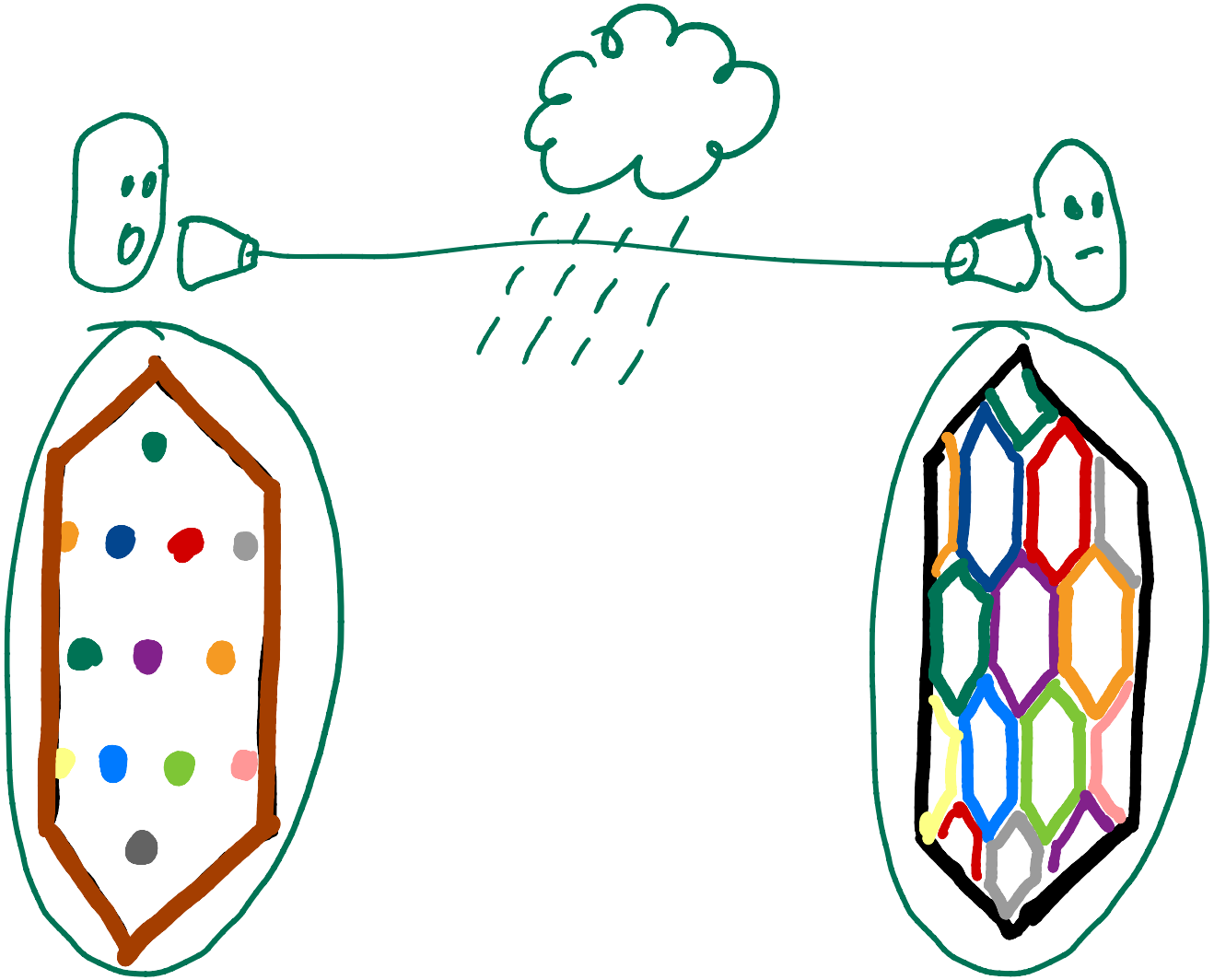
Typical set w.r.t. $P_{\hat{S}|S}$ $\approx 2^{nH(\hat{S}|S)}$

Typical set w.r.t. $P_{\hat{S}}$ $\approx 2^{nH(\hat{S})}$

$$\text{max \# of msgs} \leq \frac{2^{nH(\hat{S})}}{2^{nH(\hat{S}|S)}} = 2^{n(H(\hat{S})-H(\hat{S}|S))}$$

$$= 2^{nI(S;\hat{S})} = 2^{nC}$$

$$\| \; 2^{nr}$$



If you choose 2^{nr} codewords at random, then w.h.p., the typical sets don't overlap!

Definition

* Channel: $(\Sigma, \{p(\cdot \mid a)\}_{a \in \Sigma})$

* An (M,n)-code is a pair (Enc, Dec)

$$Enc : \{1, \ldots, M\} \longrightarrow \Sigma^n$$

$$Dec : \Sigma^n \longrightarrow \{1, \ldots, M\}$$

* The rate is $\frac{\log M}{n}$ bits/transmission

* Prob. comm error $= \frac{1}{M} \sum_{i \in M} \Pr(Dec(\hat{S}^n) \neq i \mid S^n = Enc(i))$

Typicality:

Thm: (Asymptotic equipartition) If $S_1, S_2, \ldots, S_n$ are i.i.d. $\sim p(\cdot)$ then

$$-\frac{1}{n} \log p(\underbrace{S_1, \ldots, S_n}_{\text{codeword}}) \xrightarrow[n\to\infty]{} H(p) \quad \text{in prob.}$$

As $n \to \infty$, if you choose a codeword s^n at random, by choosing each of its symbols at random using $p(\cdot)$, then $p(s^n) \approx 2^{-nH(p)}$ w.h.p.

Proof:

$$-\frac{1}{n}\log p(S_1 \dots S_n)$$

$$= -\frac{1}{n}\sum_{i\le n}\log p(S_i) = \frac{1}{n}\sum_{i\le n} -\log p(S_i)$$

$$\xrightarrow[n\to\infty]{} \mathbb{E}_S[-\log p(S)] \qquad \text{in prob. with } S\sim p(\cdot)$$

$$= H(p) = H(S).$$

Def: The typical set $A_\varepsilon^{(n)}$ w.r.t. $p(\cdot)$ is the set of sequences such that

$(s_1, \ldots, s_n) \in A_\varepsilon^{(n)}$ if

$$\left| -\frac{1}{n} \log p(s_1, \ldots, s_n) - H(p) \right| \leq \varepsilon.$$

A codeword is typical if

$$p(\text{codeword}) \approx 2^{-nH(p)}$$

Thm:

1) If $(s_1 \dots s_n) \in A_\varepsilon^{(n)}$ then

$$2^{-n(H(p)+\varepsilon)} \leq p(s_1 \dots s_n) \leq 2^{-n(H(p)-\varepsilon)}$$

2) $\Pr[A_\varepsilon^{(n)}] \geq 1-\varepsilon$,
for sufficiently large n

3) $(1-\varepsilon)2^{n(H(p)-\varepsilon)} \leq |A_\varepsilon^{(n)}| \leq 2^{n(H(p)+\varepsilon)}$,
for large n

$p(\text{typical codeword}) \approx 2^{-nH(p)}$

If you choose a codeword at random then w.h.p. that codeword is typical.

$|A_\varepsilon^{(n)}| \approx 2^{nH(p)}$.

Proof:

1) If $(s_1, \dots, s_n) \in A_\varepsilon^{(n)}$,

$$\left| -\frac{1}{n} \log p(s_1 \dots s_n) - H(p) \right| \leq \varepsilon$$

$$H(p) - \varepsilon \leq -\frac{1}{n} \log p(s_1 \dots s_n) \leq H(p) + \varepsilon$$

$$2^{-n(H(p)+\varepsilon)} \leq p(s_1 \dots s_n) \leq 2^{-n(H(p)-\varepsilon)}$$

$$p(\text{typical codeword}) \approx 2^{-nH(p)}$$

Proof:

2) $\Pr[A_\varepsilon^{(n)}] \geq 1-\varepsilon$, for sufficiently large n.

$$-\frac{1}{n}\log p(s_1,\dots s_n) \xrightarrow[n\to\infty]{\text{in. prob}} H[p]$$

for $\varepsilon>0$, there is N, for $n \geq N$

$$\Pr\Big(\underbrace{\Big|-\frac{1}{n}\log p(s_1,\dots s_n) - H[p]\Big| < \varepsilon}_{(s_1,\dots s_n)\in A_\varepsilon^{(n)}}\Big) \geq 1-\varepsilon$$

$$\Pr\big((s_1,\dots s_n] \in A_\varepsilon^{(n)}\big) \geq 1-\varepsilon$$

$$\Pr\big(A_\varepsilon^{(n)}\big) \geq 1-\varepsilon.$$

If you choose a codeword s^n at random, then w.h.p. s^n is typical w.r.t. $p(\cdot)$.

Proof

3. $(1-\varepsilon)2^{n(H(p)-\varepsilon)} \leq |A_\varepsilon^{(n)}| \leq 2^{n(H(p)+\varepsilon)}$ for large n

$$1 = \sum_{(s_1 \dots s_n)} p(s_1 \dots s_n) \geq \sum_{(s_1 \dots s_n) \in A_\varepsilon^{(n)}} p(s_1 \dots s_n)$$

$$\geq \sum_{(s_1 \dots s_n) \in A_\varepsilon^{(n)}} 2^{-n(H(p)+\varepsilon)}$$

$$= 2^{-n(H(p)+\varepsilon)} |A_\varepsilon^{(n)}|$$

$$|A_\varepsilon^{(n)}| \leq 2^{-n(H(p)+\varepsilon)}$$

Proof

3. $(1-\varepsilon)2^{n(H(p)-\varepsilon)} \leq |A_\varepsilon^{(n)}| \leq 2^{n(H(p)+\varepsilon)}$ for large n

$\Pr[A_\varepsilon^{(n)}] \geq 1-\varepsilon$, for suff. large n

$$1-\varepsilon \leq \Pr[A_\varepsilon^{(n)}] = \sum_{(s_1 \ldots s_n) \in A_\varepsilon^{(n)}} p(s_1 \ldots s_n) \leq \sum_{(s_1 \ldots s_n) \in A_\varepsilon^{(n)}} 2^{-n(H(p)-\varepsilon)}$$

$$\Rightarrow (1-\varepsilon)2^{n(H(p)-\varepsilon)} \leq |A_\varepsilon^{(n)}| = 2^{-n(H(p)-\varepsilon)} |A_\varepsilon^{(n)}|$$

$$|A_\varepsilon^{(n)}| \approx 2^{nH(p)}$$

Def: The set $A_\varepsilon^{(n)}$ of jointly typical pairs of seqs w.r.t. $P_{S,\hat{S}}(\cdot,\cdot)$ is the set of pairs of sequences such that

$$(s^n, \hat{s}^n) \in A_\varepsilon^{(n)} \quad \text{iff}$$

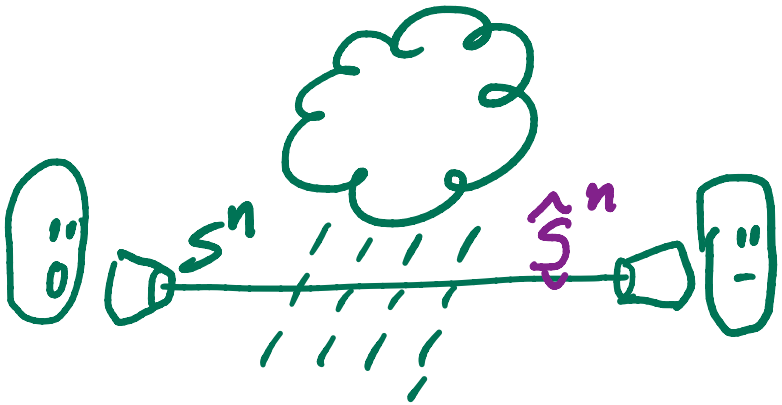
1. $\left| -\frac{1}{n} \log P_{S^n}(s^n) - H(S) \right| < \varepsilon$ 2. $\left| -\frac{1}{n} \log P_{\hat{S}^n}(\hat{s}^n) - H(\hat{S}) \right| < \varepsilon$

3. $\left| -\frac{1}{n} \log P_{S^n,\hat{S}^n}(s^n,\hat{s}^n) - H(S,\hat{S}) \right| < \varepsilon$

In our context : S^n input codeword, $\hat{S}^n$ output codeword

Thm:

1. $\Pr\left((S^n, \hat{S}^n) \in A_\varepsilon^{(n)}\right) \to 1$ as $n \to \infty$



If s^n drawn from $p^*_{S^n}$ then $(s^n, \hat{s}^n) \in A_\varepsilon^{(n)}$ w.h.p.

If

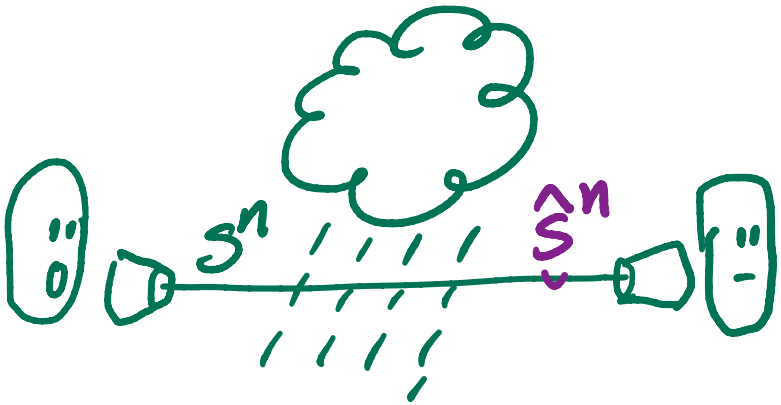
1. $s^n \overset{\$}{\leftarrow} p^*_{S^n}(\cdot)$
2. $\hat{s}^n \overset{\$}{\leftarrow} p_{\hat{S}^n | S^n = s^n}(\cdot)$

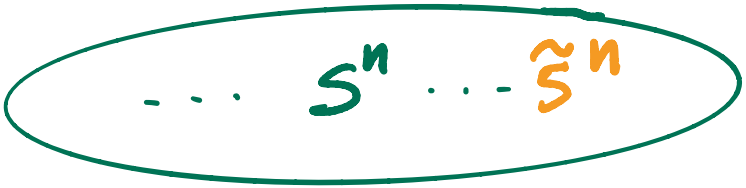
then $(s^n, \hat{s}^n) \in A_\varepsilon^{(n)}$ w.h.p.

Thm:

2. If $\tilde{S}^n \sim p_{S^n}(\cdot)$ and $\hat{S}^n \sim p_{\hat{S}^n}(\cdot)$. then

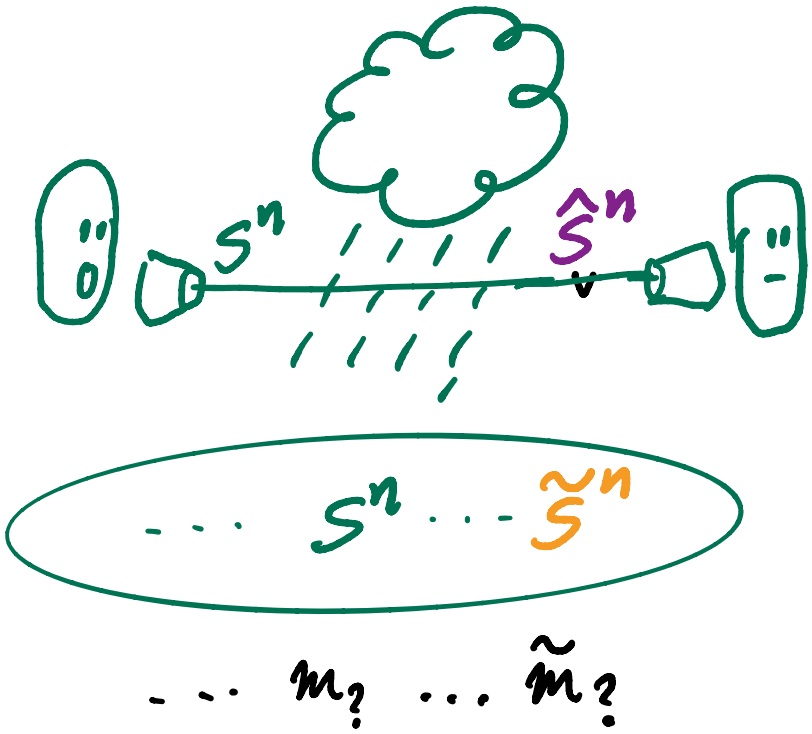
$$\Pr\left[\tilde{S}^n, \hat{S}^n \in A_\varepsilon^{(n)}\right] \leq 2^{-n(I(S;\hat{S})-3\varepsilon)}$$





$\ldots \; m_? \; \ldots \; \tilde{m}_?$

$(\tilde{S}^n, \hat{S}^n) \notin A_\varepsilon^{(n)}$ w.h.p.



$(s^n, \hat{s}^n) \in A_\varepsilon^{(n)}$ and $(\tilde{s}^n, \hat{s}^n) \notin A_\varepsilon^{(n)}$ w.h.p.

Shannon's coding thm

If $r <$ channel's capacity then there is a code s.t.

* rate $= r$ (i.e., $\#\text{msgs} = 2^{nr}$)
* prob. error $\xrightarrow[n\to\infty]{} 0$

(The converse is also true)

If $\#\text{msgs} = 2^{nr}$ with $r < C$, then there is a $(\lfloor 2^{nr} \rfloor, n)$-code with pr. error $\xrightarrow[n\to\infty]{} 0$.

Remember:

A $(\lfloor 2^{nr} \rfloor, n)$-code is a pair (Enc, Dec)

* $Enc: \{1 \dots \lfloor 2^{nr} \rfloor\} \longrightarrow \mathcal{Z}^n$

* $Dec: \mathcal{Z}^n \longrightarrow \{1 \dots \lfloor 2^{nr} \rfloor\}$

$$\text{Prob. comm. error} = \frac{1}{\lfloor 2^{nr} \rfloor} \sum_{m < \lfloor 2^{nr} \rfloor} \Pr\left(Dec(\hat{S}^n) \neq m \mid S^n = Enc(m)\right)$$

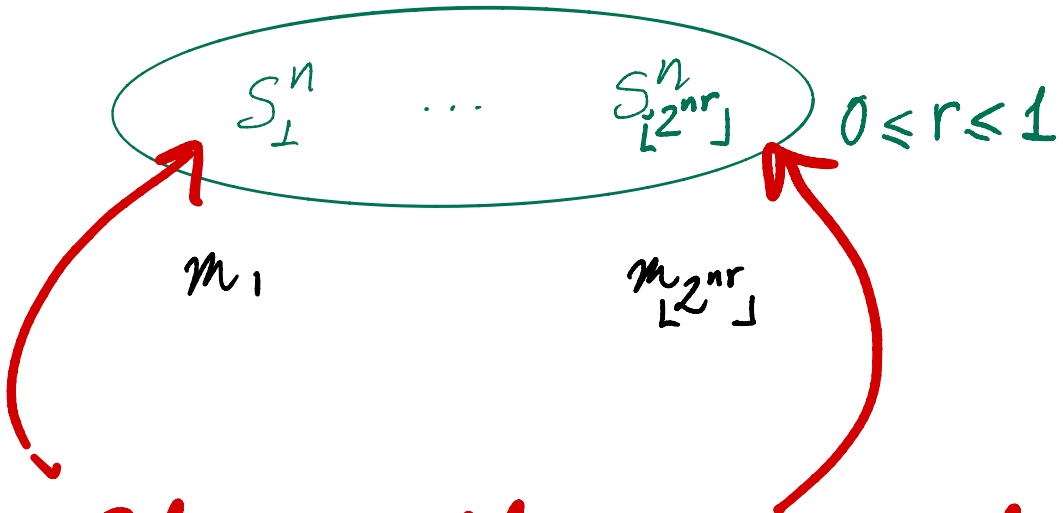
Shannon's random code

Let $r <$ channel cap.

m_1 $\quad\quad$ $m_{\lfloor 2^{nr} \rfloor}$

Pick $\lfloor 2^{nr} \rfloor$ msgs.

Shannon's random code



Encoder

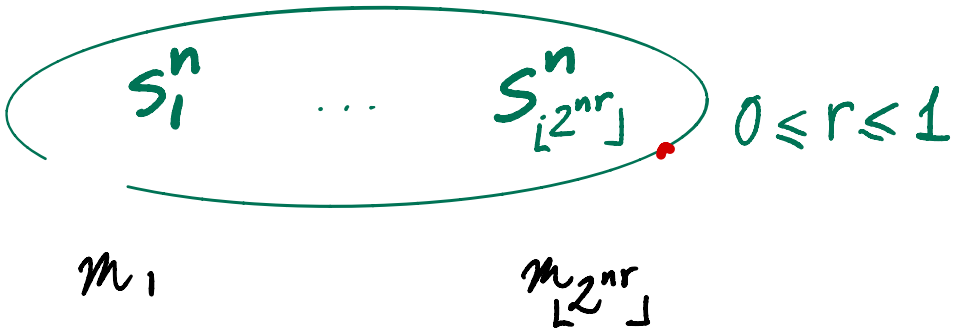
Choose the codewords at random from $p_S^* = \arg\max_p I(S;\hat{S})$

The codewords are typical w.h.p.

Shannon's random code

For which s_i^n

$(s_i^n, \hat{s}^n) \in A_\varepsilon^{(n)}$?

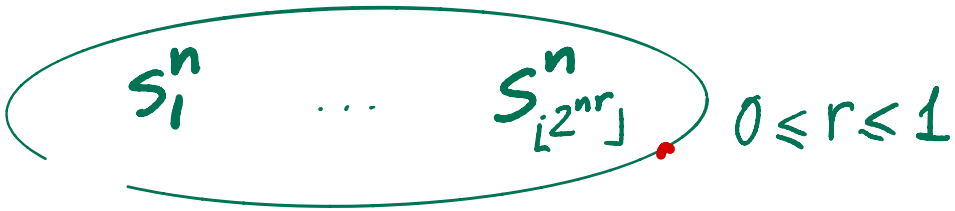


Decoder

Shannon's random code

For which s_i^n

$(s_i^n, \hat{s}^n) \in A_\varepsilon^{(n)}$?



m_1 $\quad$ $m_{\lfloor 2^{nr} \rfloor}$

We now prove that as $n \longrightarrow \infty$

* rate $= r$

* prob. error $\longrightarrow 0$

$$\text{rate} = \frac{\log \lfloor 2^{nr} \rfloor}{n} \xrightarrow[n \to \infty]{} r$$

Proof :

$$P_r(\mathcal{E}) = P_r(\mathcal{E} \mid \#1) \qquad S^n = Enc(1)$$

$$= P_r\left[\begin{array}{l} Enc(1), \hat{S}^n \notin A_\varepsilon^{(n)} \text{ or} \\ Enc(2), \hat{S}^n \in A_\varepsilon^{(n)} \text{ or} \\ \quad \vdots \\ Enc(\lfloor 2^{nr} \rfloor), \hat{S}^n \in A_\varepsilon^{(n)} \end{array} \middle| \; \#1 \right]$$

$$\leq P_r\left[Enc(1), \hat{S}^n \notin A_\varepsilon^{(n)} \mid \#1\right] + \sum_{2 \leq w \leq \lfloor 2^{nr} \rfloor} P_r\left[Enc(w), \hat{S}^n \in A_\varepsilon^{(n)} \mid \#1\right]$$

$$\leq \underbrace{\Pr\left[\mathrm{Enc}(1), \hat{S}^n \notin A_\varepsilon^{(n)} \mid \#1\right]}_{\eta_n \xrightarrow[n\to\infty]{} 0} + \sum_{2 \leq w \leq \lfloor 2^{nr} \rfloor} \underbrace{\Pr\left[\mathrm{Enc}(w), \hat{S}^n \in A_\varepsilon^{(n)} \mid \#1\right]}_{\leq 2^{-n(I(S;\hat{S})-3\varepsilon)}}$$

$$\leq \eta_n + \sum_{2 \leq w \leq \lfloor 2^{nr} \rfloor} 2^{-n(I(S;\hat{S})-3\varepsilon)}$$

$$= \eta_n + \left(\lfloor 2^{nr} \rfloor - 1\right) 2^{-n(I(S;\hat{S})-3\varepsilon)}$$

$$\leq \eta_n + 2^{nr}\, 2^{-n(I(S;\hat{S})-3\varepsilon)}$$

$$= \eta_n + 2^{-n\left(I(S;\hat{S}) - r - 3\varepsilon\right)}$$

if $r < I(S;\hat{S})$ $= C$

$$\xrightarrow[n\to\infty]{} 0$$

■

Shannon's coding thm

If # of msgs $= 2^{nr}$ (with $r < C$)

$\Rightarrow$ there is a code

prob. error $\xrightarrow[n \to \infty]{} 0$

(The converse is true)

This means $\Rightarrow$ max # msgs $= 2^{nC}$

$\Rightarrow$ capacity as a channel measure