Validation via information theory II

Statistical Learning Theory 2020

Organization

What is PA?

Roadmap

Formalization of PA

→ Algorithms as channels

→ Derivation of PA

Applications

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?

Roadmap

Formalization of PA

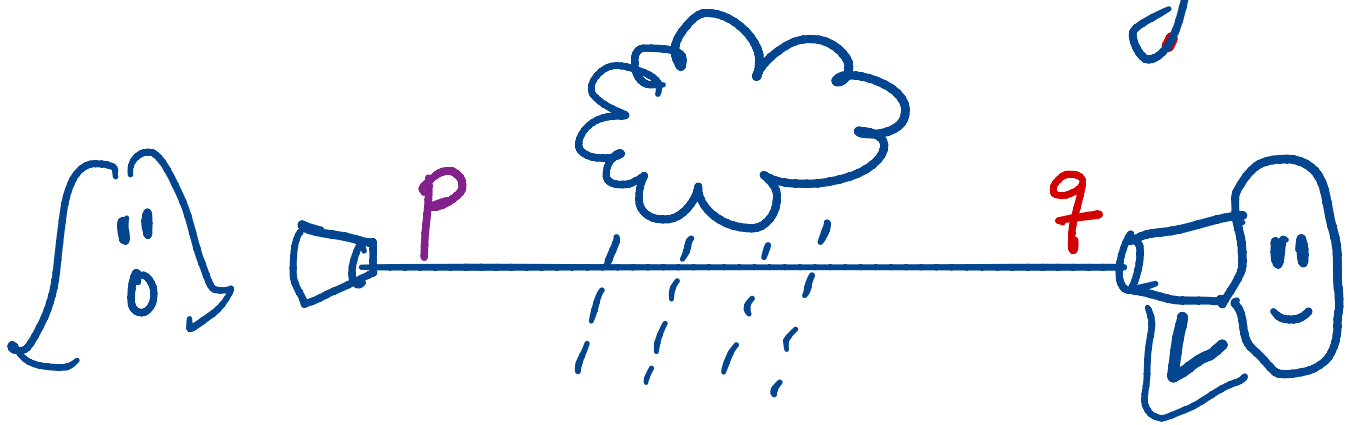
→ Algorithms as channels

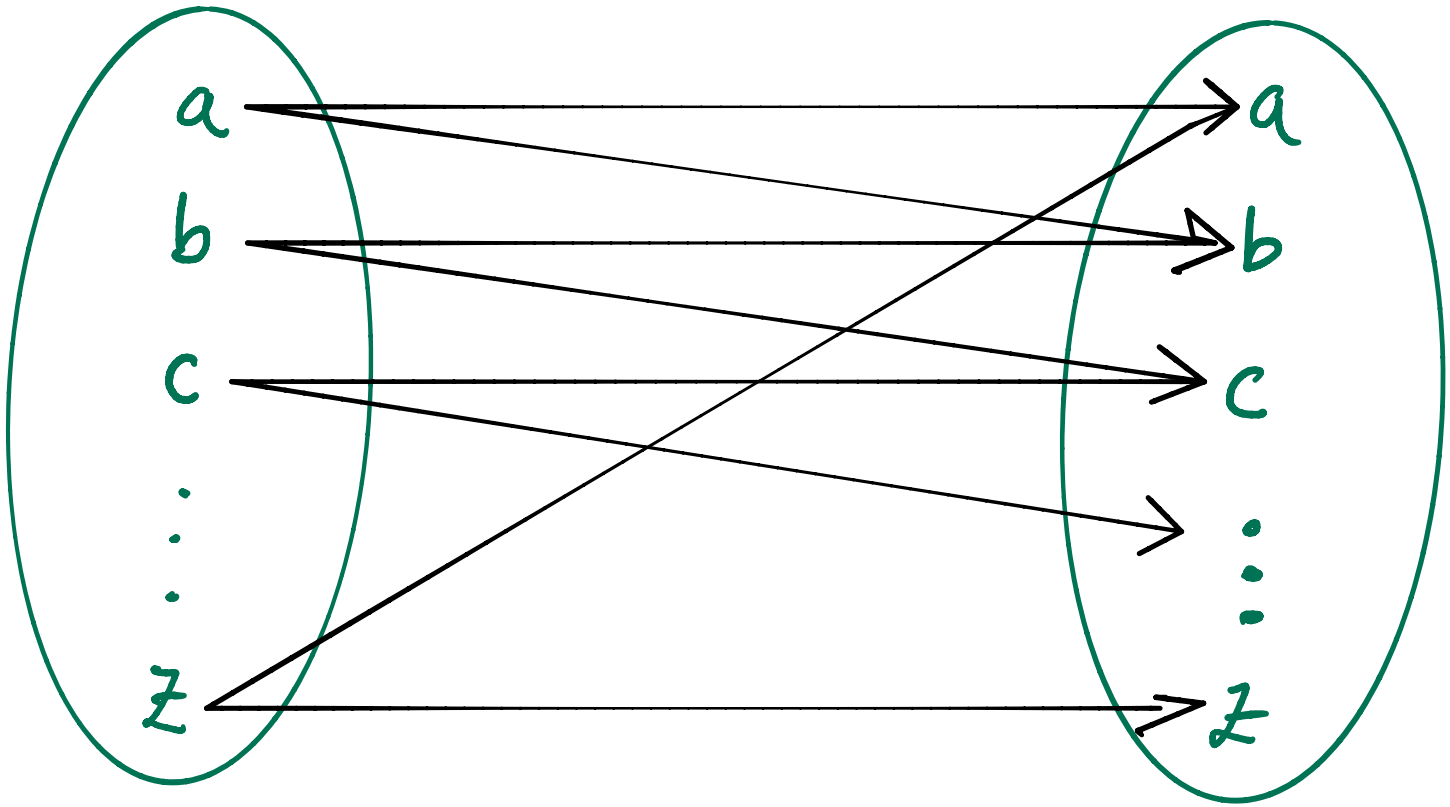
→ Derivation of PA

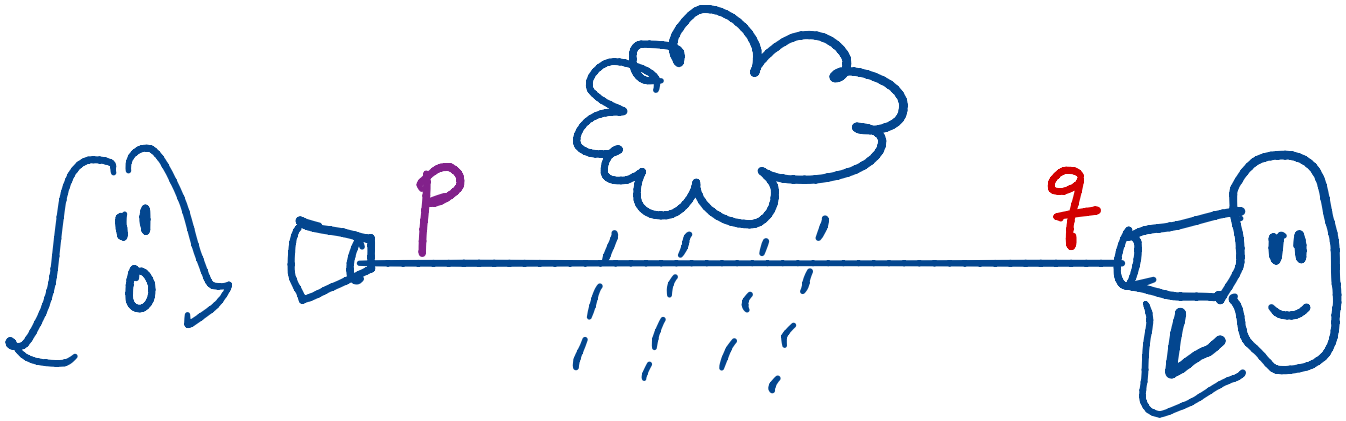
Applications

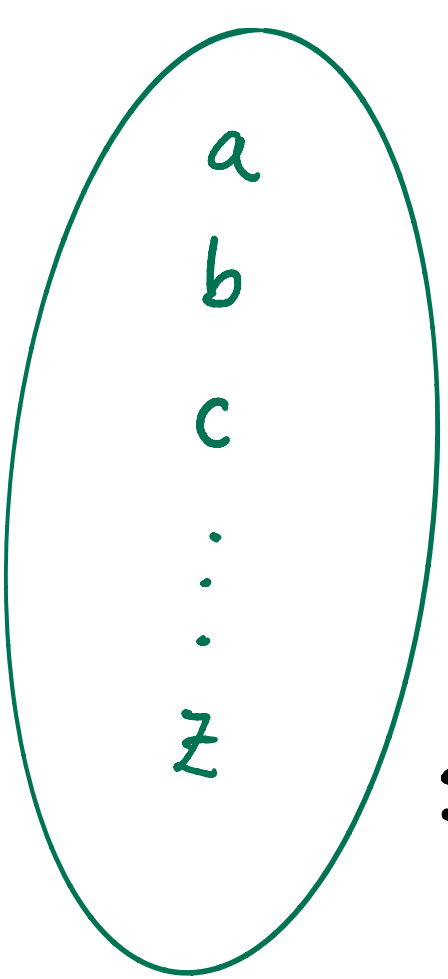
PA originates from modeling and evaluating algorithms as channels

Communication on a noisy channel

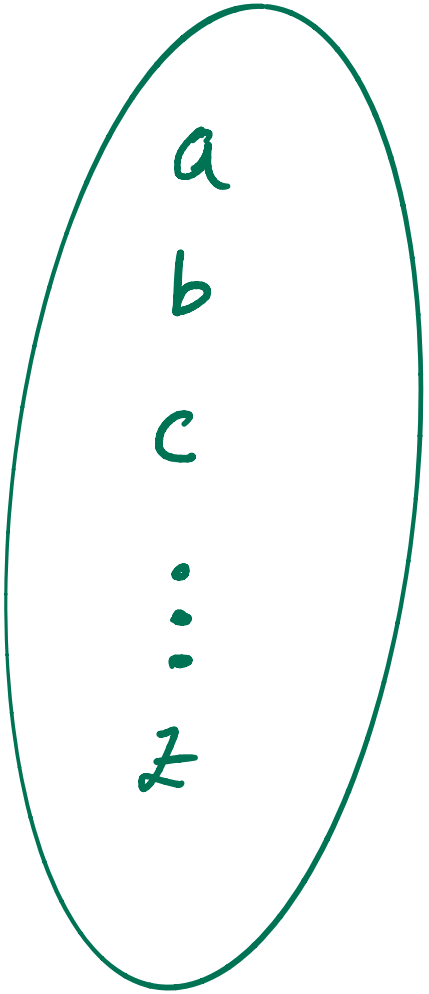


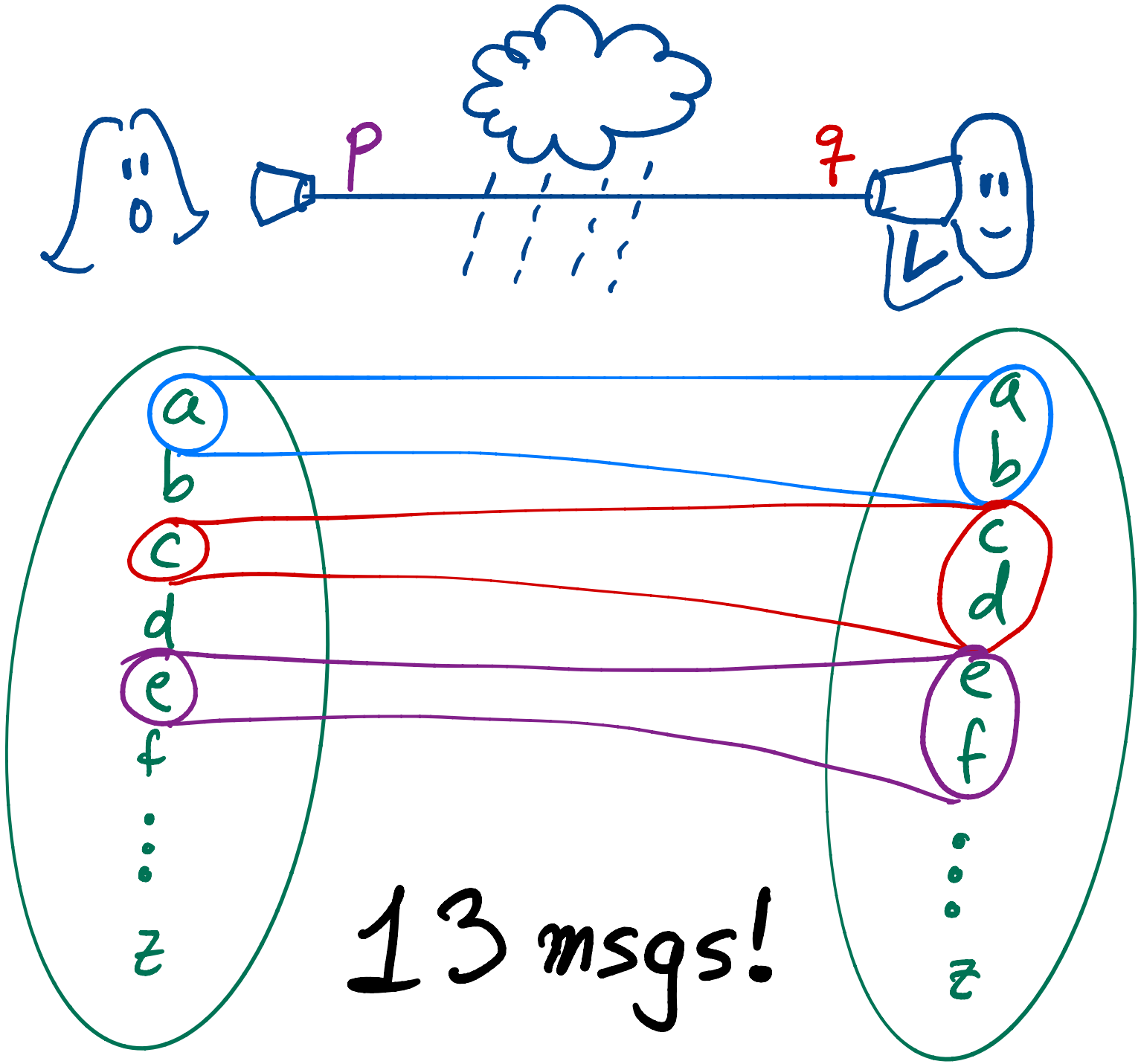


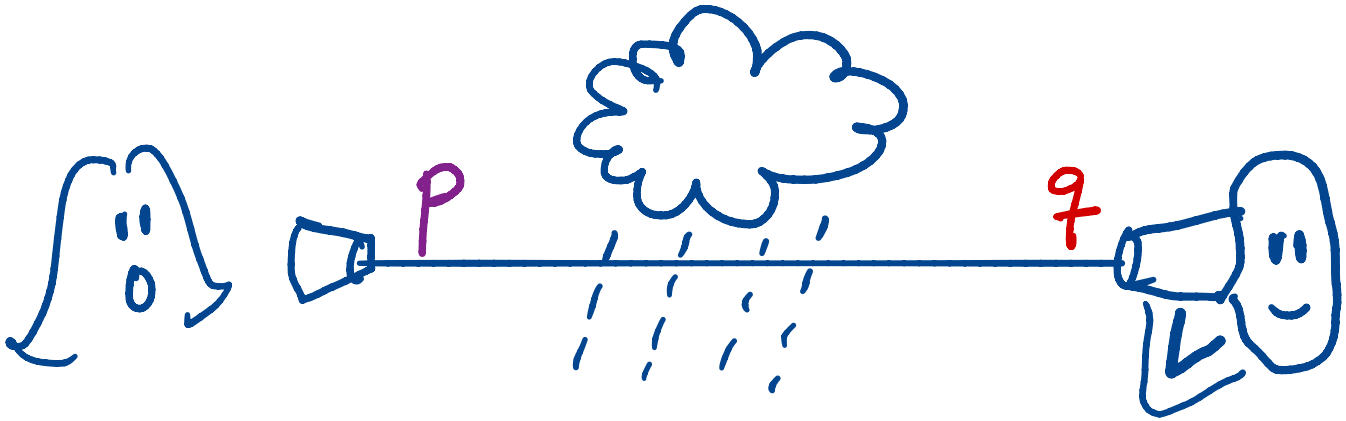


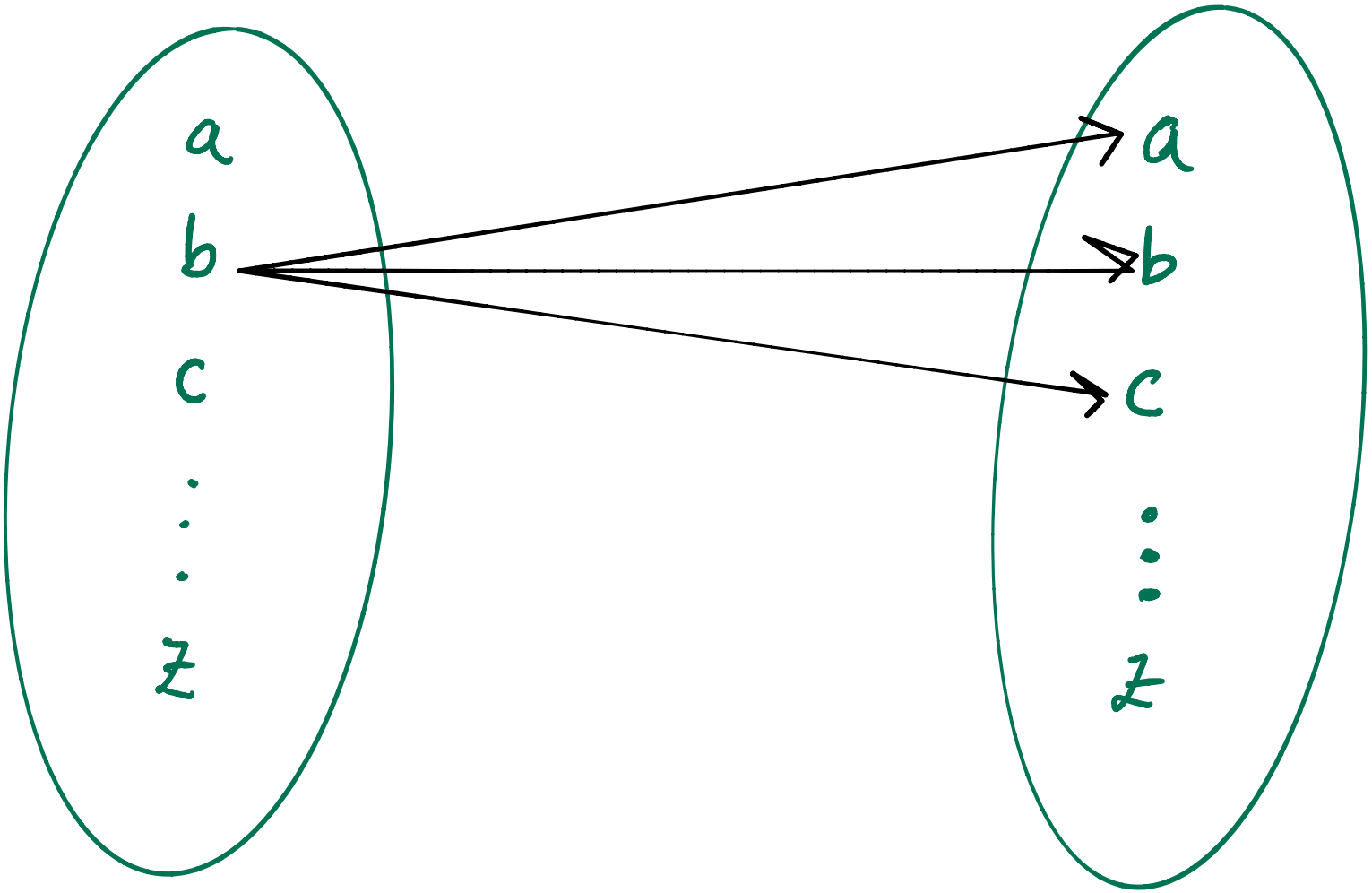


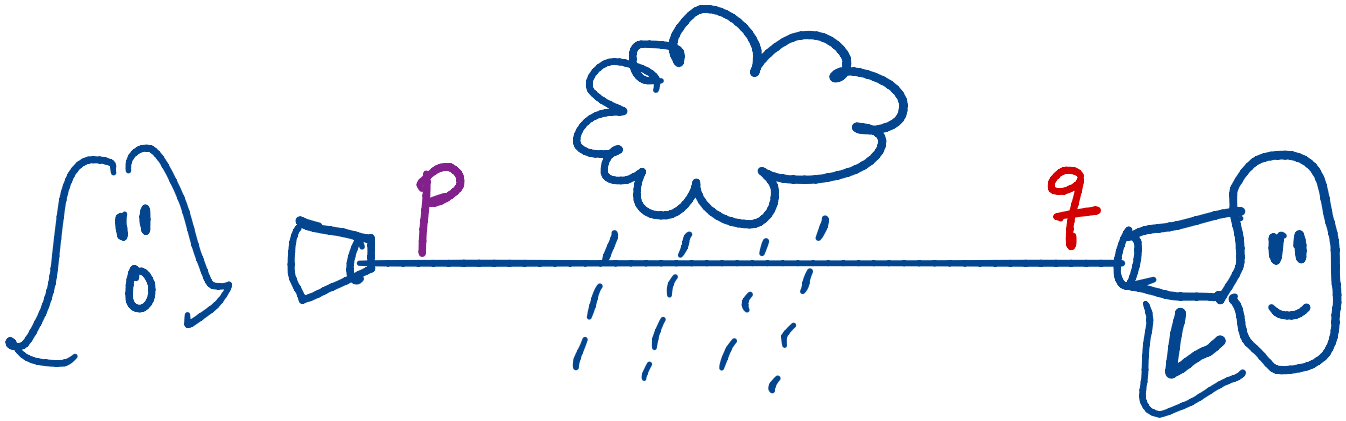
max # msgs
s.t. Pr. error=0?

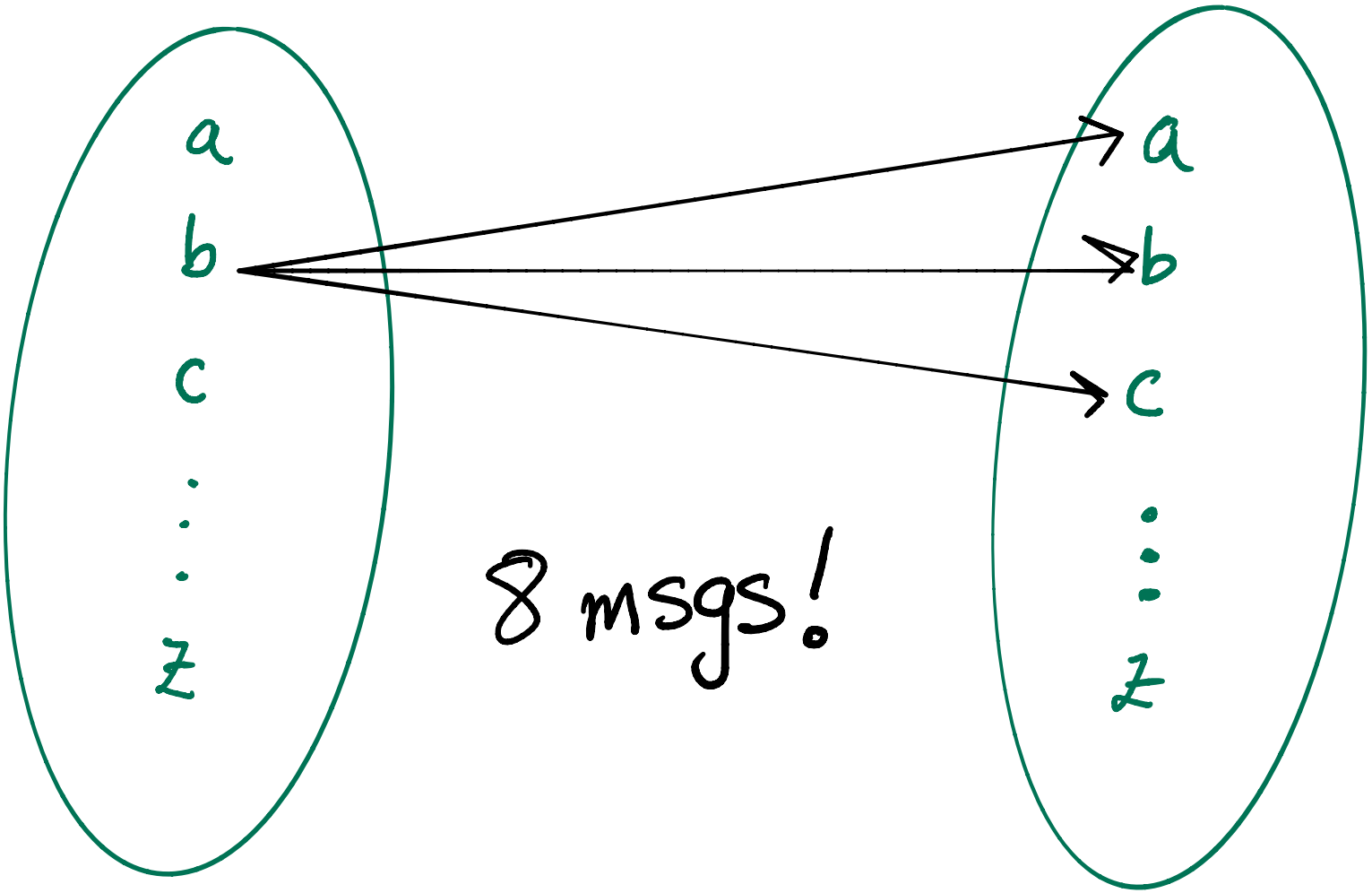


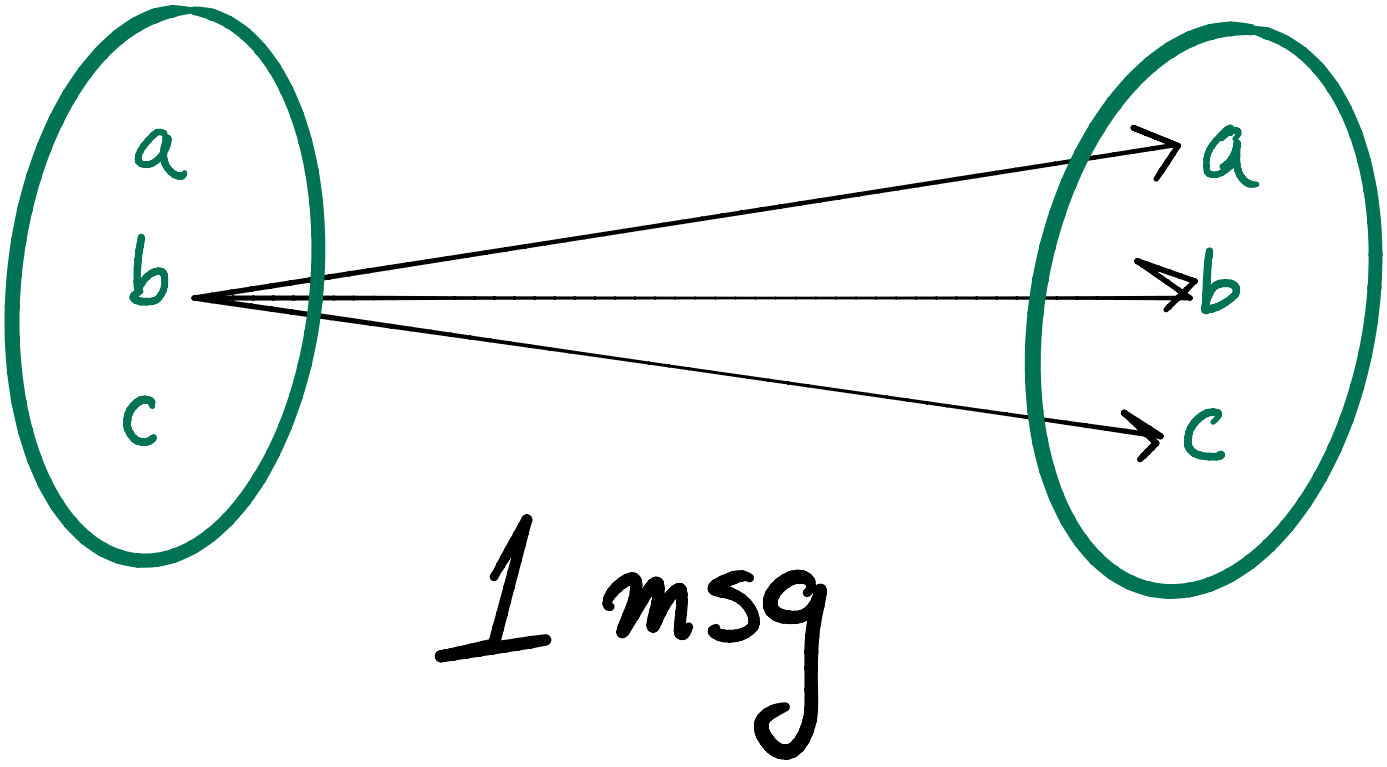

p
q
a
b
c
d
e
f
z
a
b
c
d
e
f
z
13 msgs!


p
q


a
b
c
z
a
b
c
z


p
q


a
b
c
z
a
b
c
z
8 msgs!



1 msg

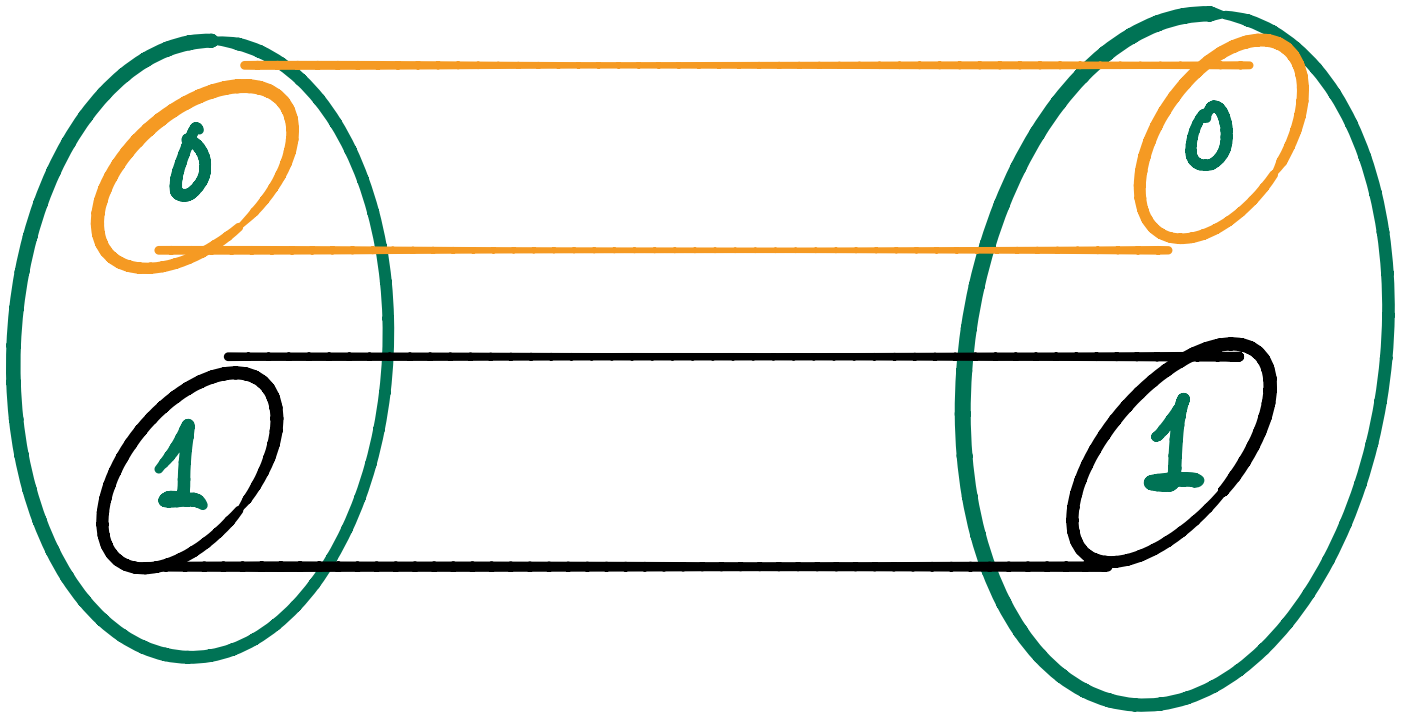
Good channels

<=>

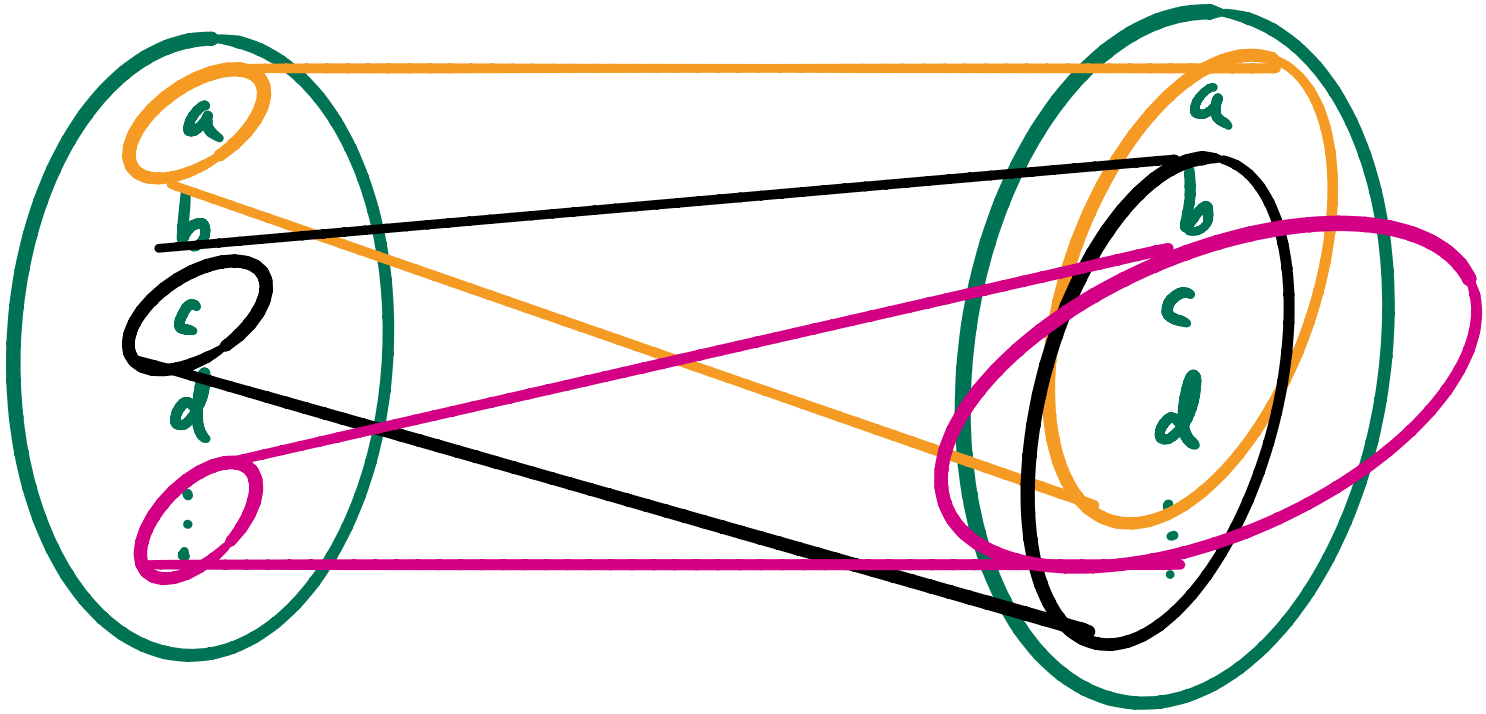
high info transfer

= codewords have content + channel robust against noise

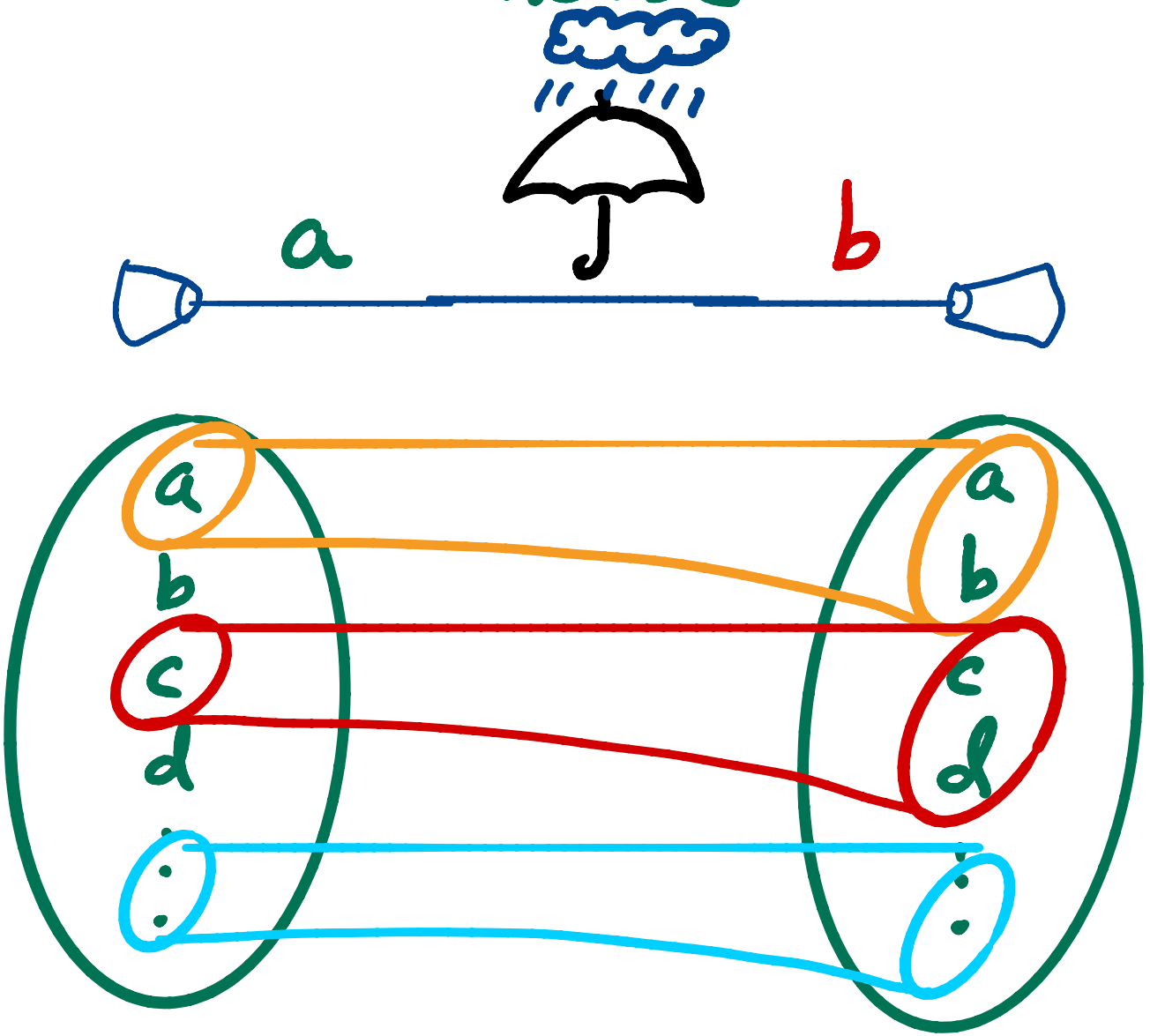
poor content + robust against noise = low information transfer



rich content + sensitive to noise = low information transfer



rich content + robust against noise = high information transfer



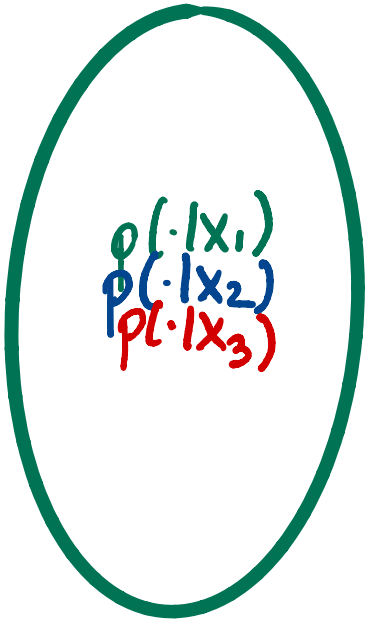
Good algorithms

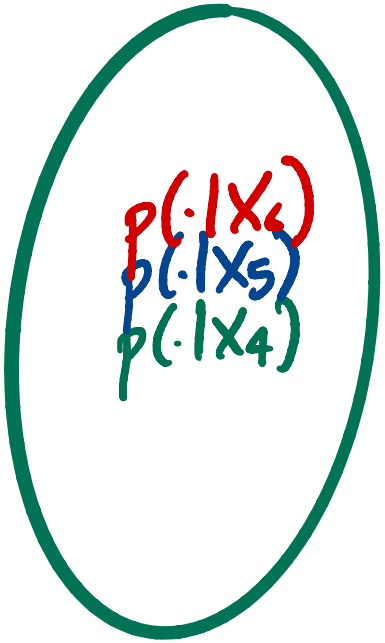
$\Longleftrightarrow$

high info extraction

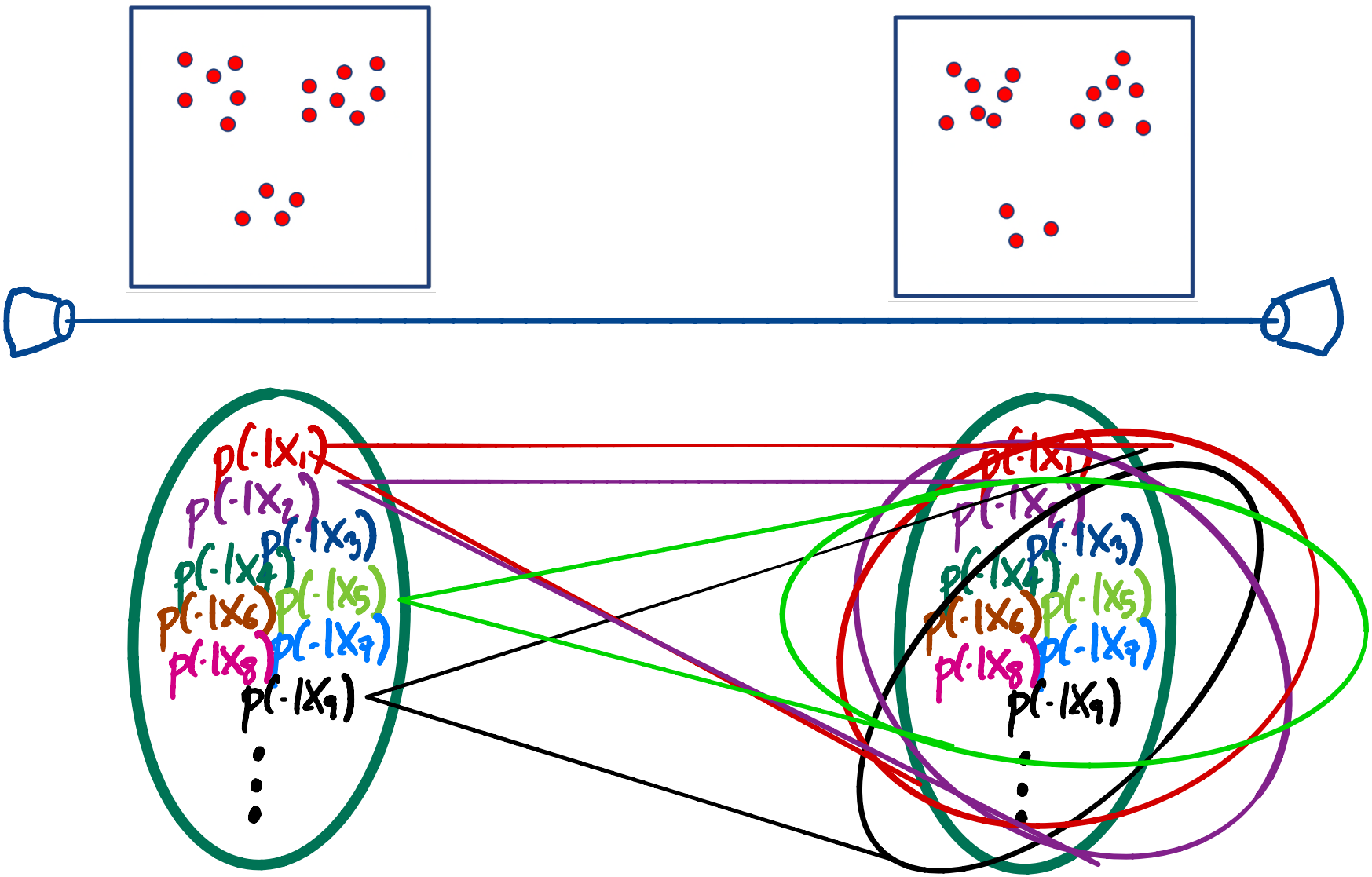
= outputs have content + algorithm robust against noise

poor content + robust = low info extraction

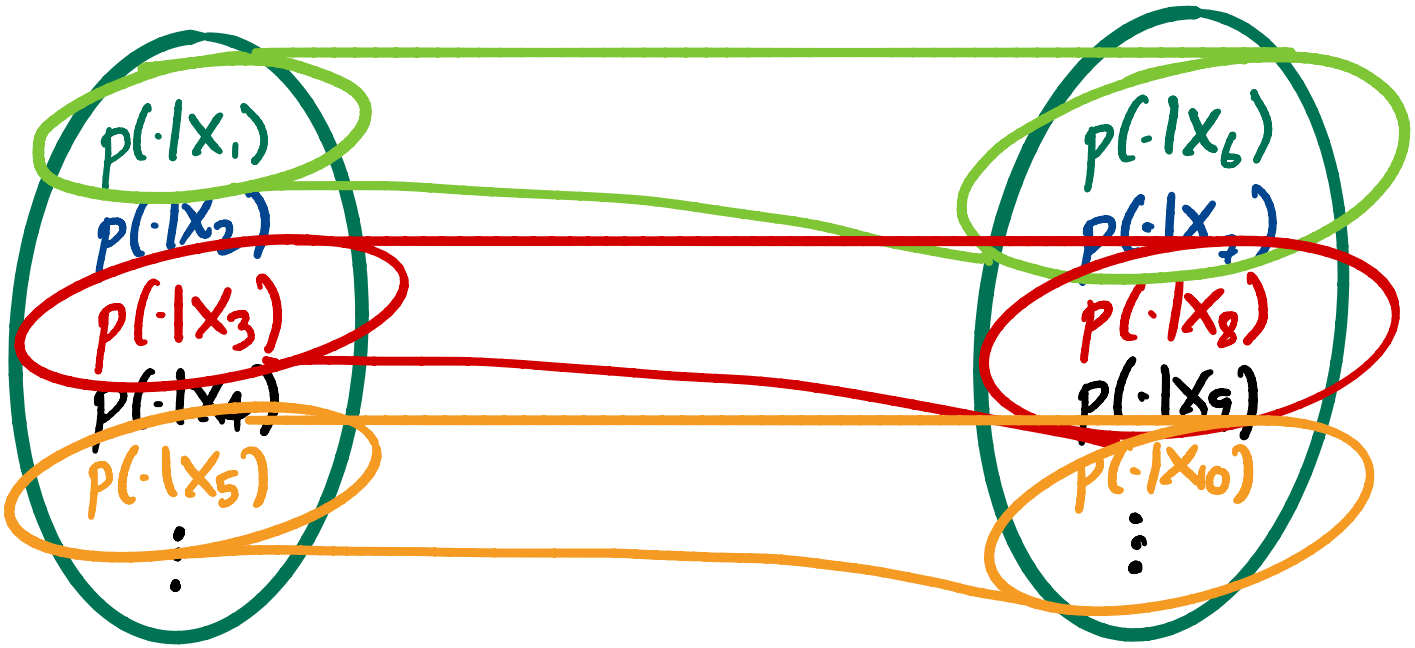




rich content + no robustness = low info extraction



rich content + robust = high info extraction



Information extraction	Information transmission
* Outputs have content * Algorithm is robust	* Codewords have content * Channel is robust
Adjust your algos to increase exp. log. PA	Adjust your channels to increase 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?

Roadmap

Formalization of PA

→ Algorithms as channels

→ Derivation of PA

Applications

Algorithms as communication channels

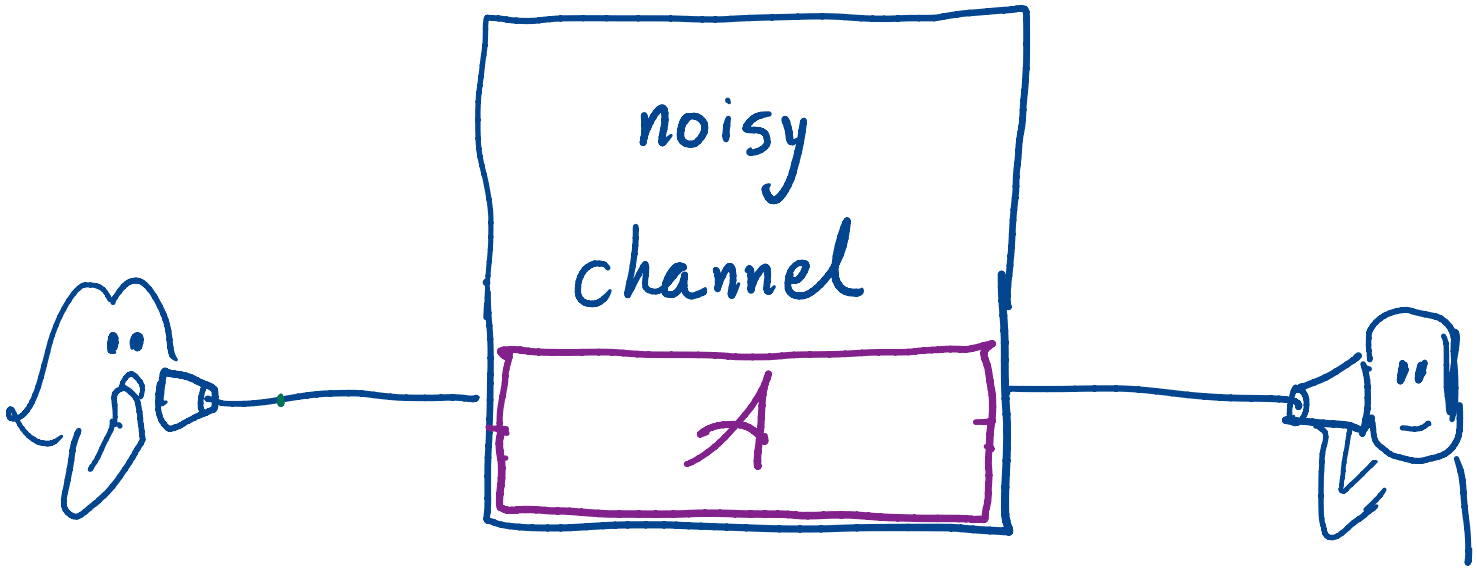
Assumptions

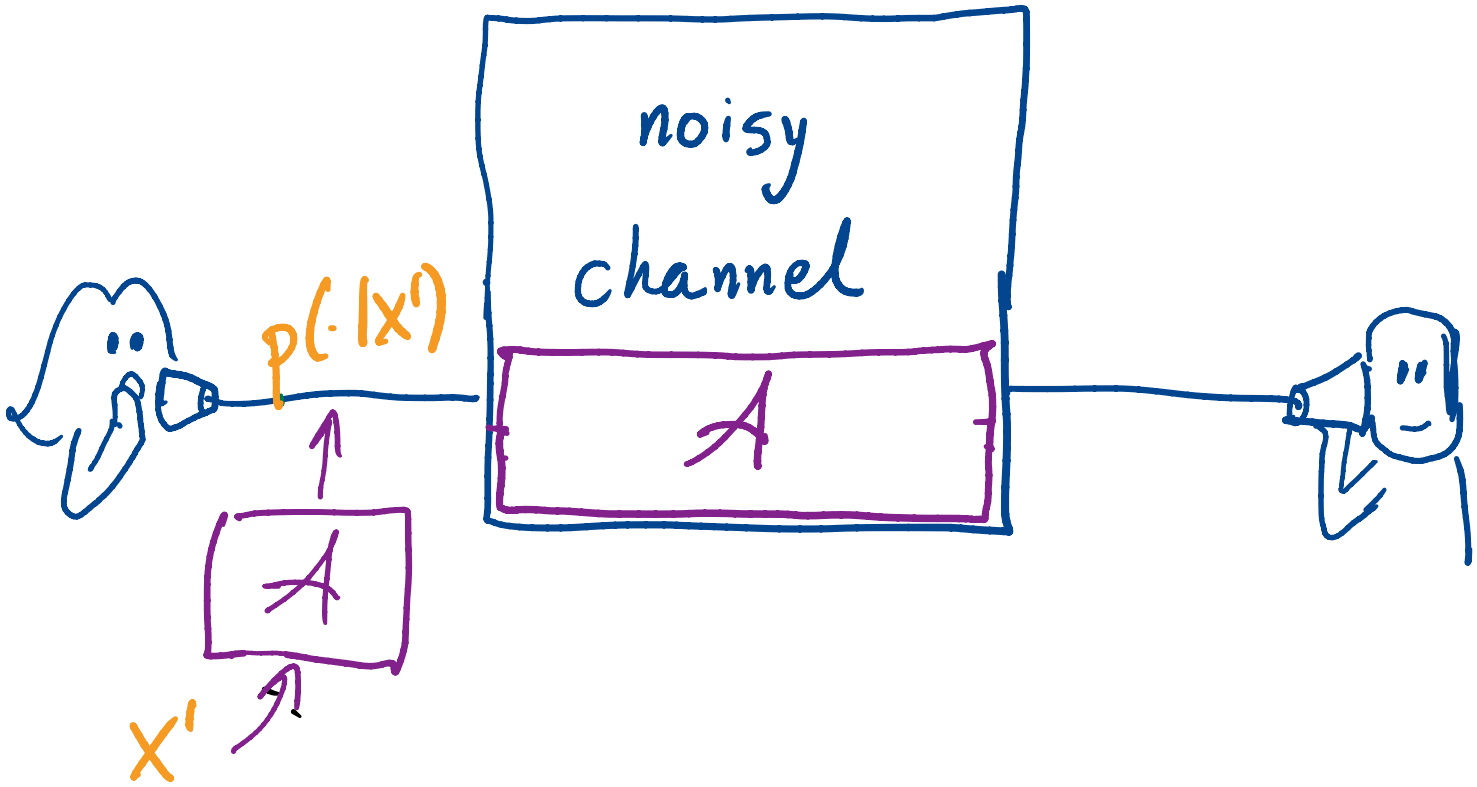
* Exponential sol space: $\log |\mathcal{C}| = O(n)$.

* Prob outputs: $x' \xrightarrow{\mathcal{A}} p(\cdot|x')$

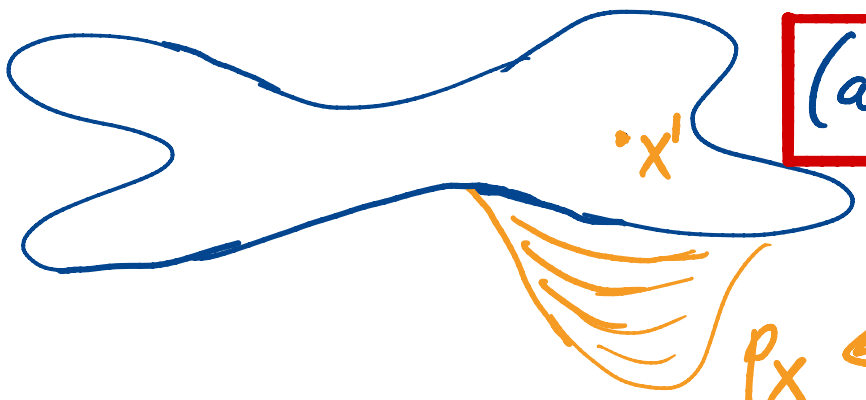
* Asymptotic equip. property:

$$\underbrace{\text{exp.}\log.\text{PA}}_{\approx \frac{1}{\log|\mathcal{C}|}\sum_i \log Y_i} \xrightarrow[n\to\infty]{\text{in prob}} \underbrace{\text{exp.}\log.\text{PA}}_{\approx \mathbb{E}_Y[\log Y]}$$

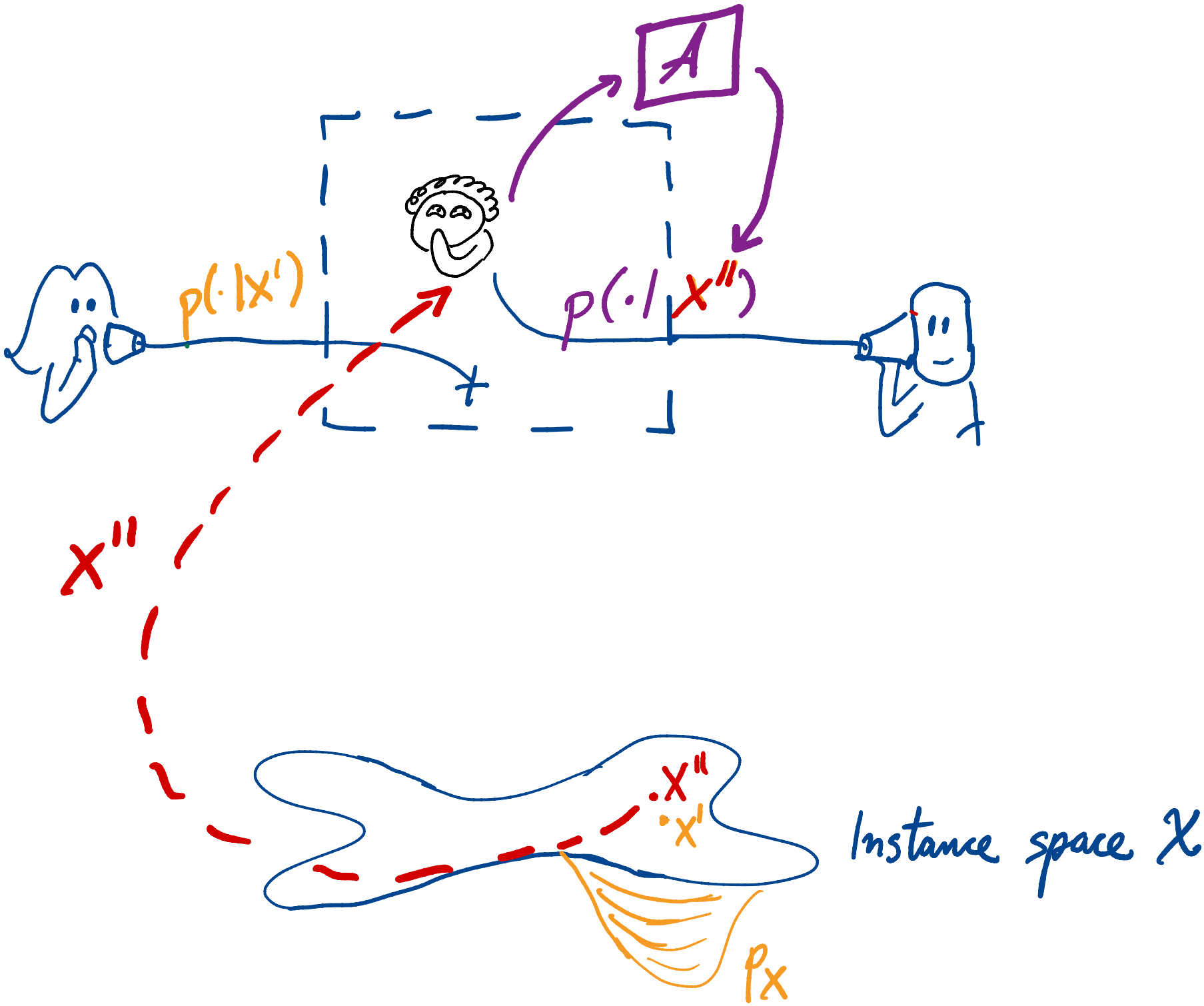

noisy
channel
A

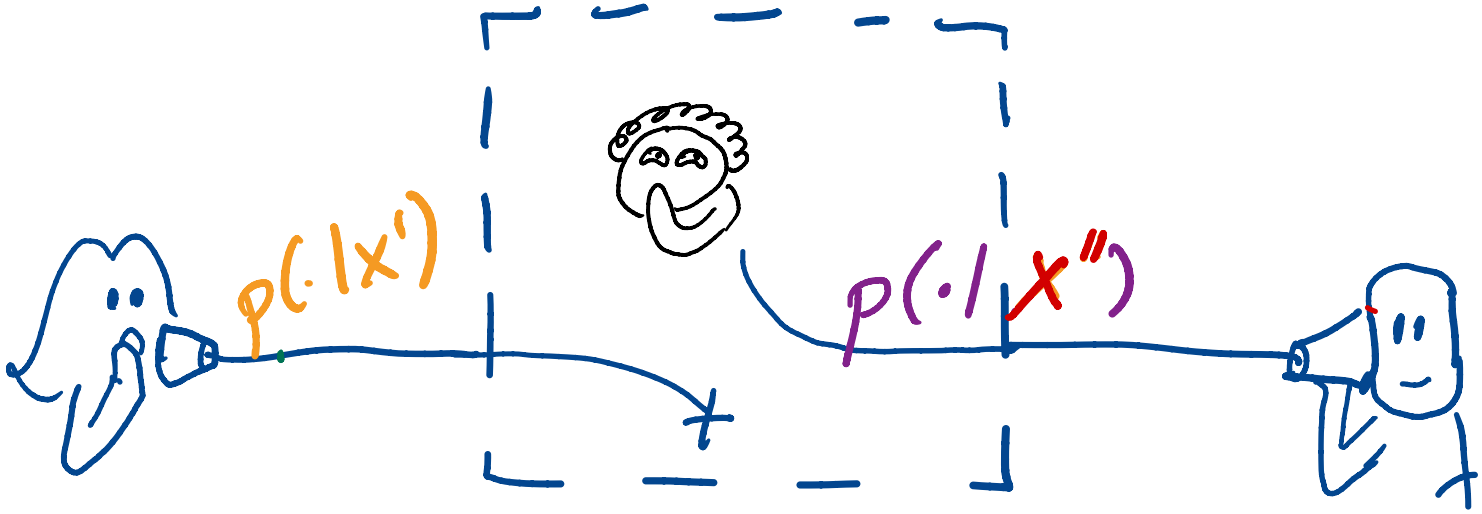


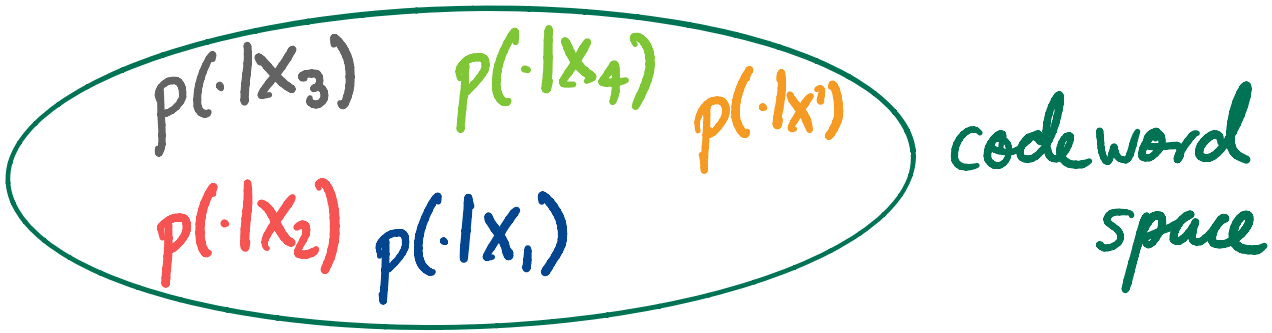
Instance space $\mathcal{X}$
(all instances of "size" n)

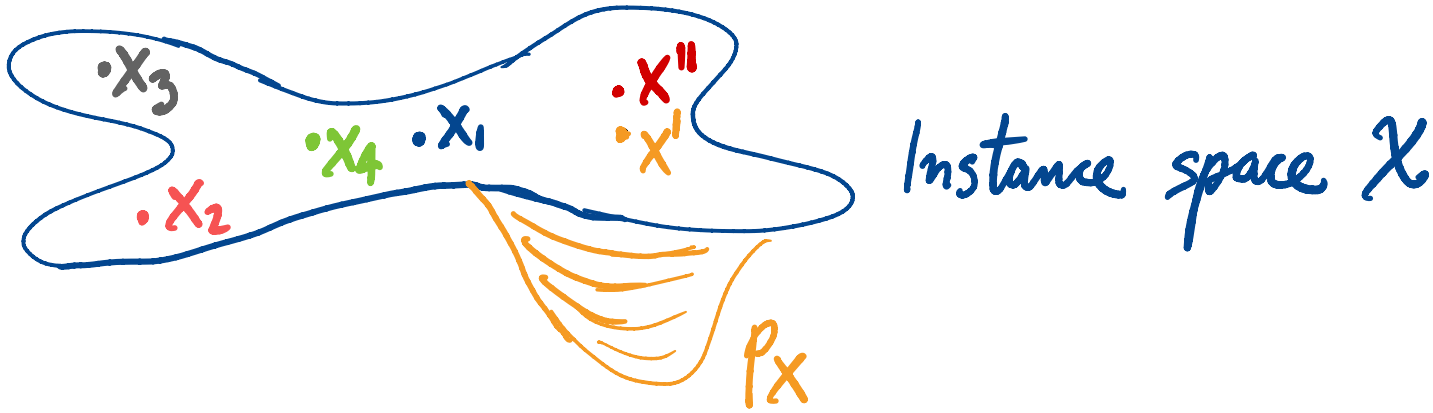


$p_X \leftarrow$ experiment

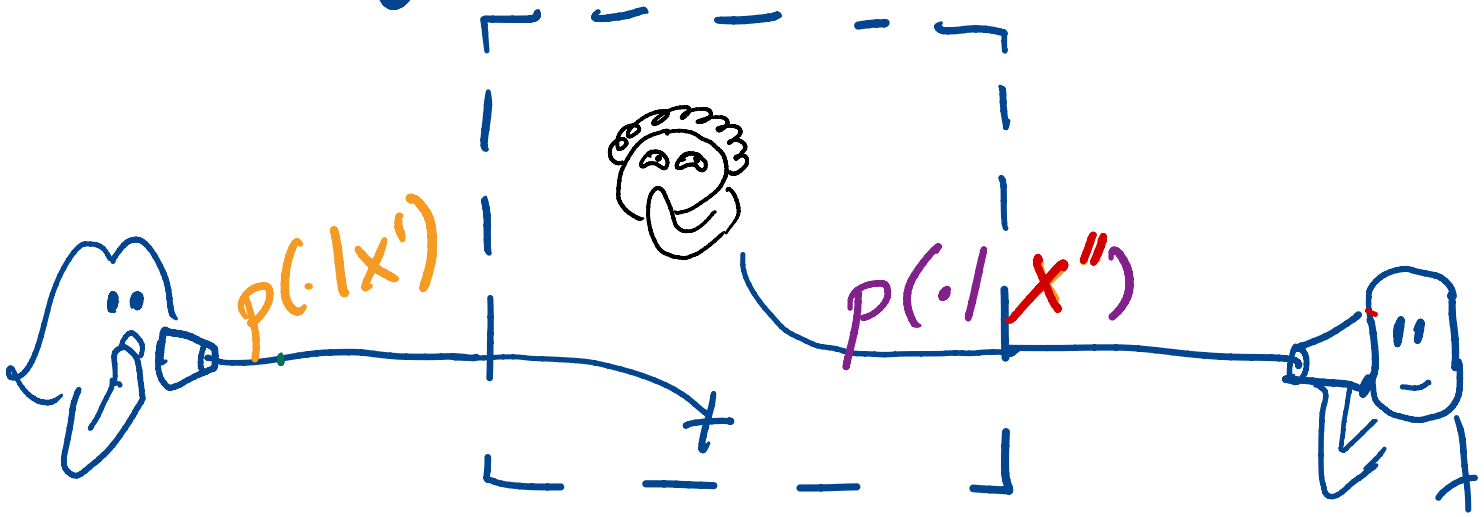

A
p(·|x')
p(·|x'')
x''
x''
x'
Instance space X
p_X

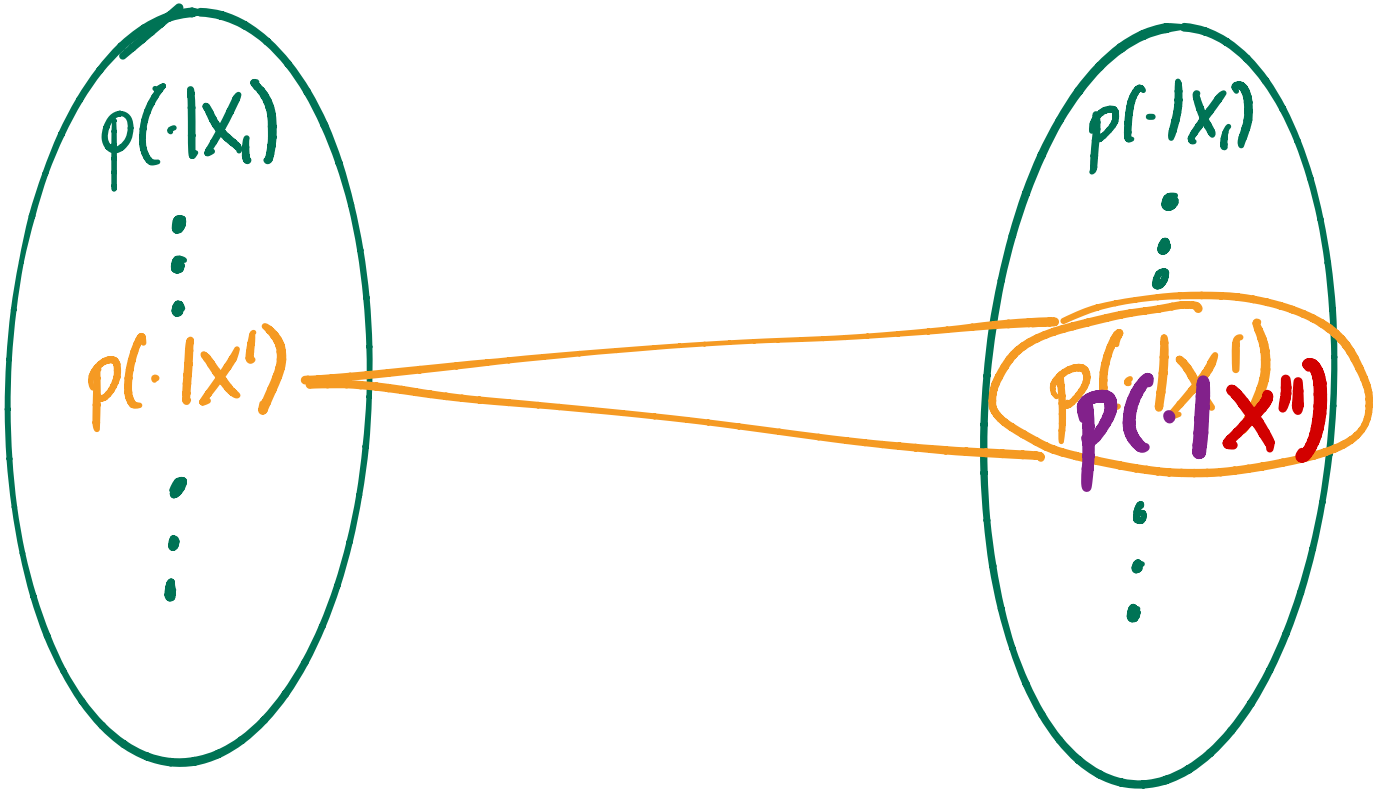

p(·|x')
p(·|x'')


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

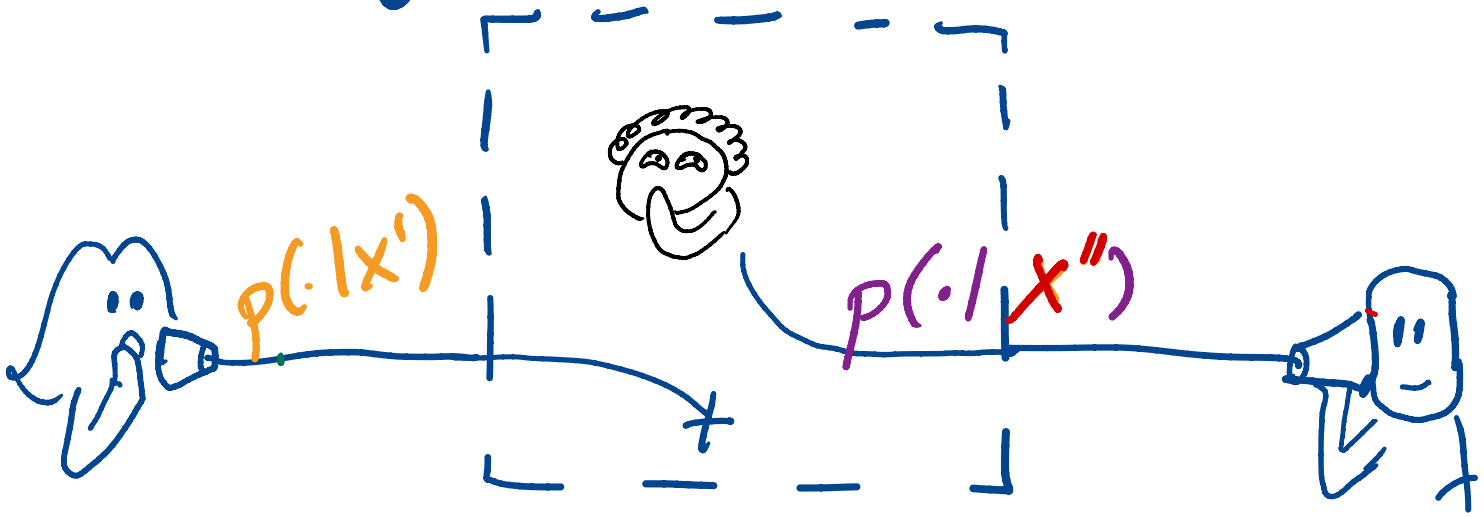

x3
x4
x1
x''
x'
x2
px
Instance space X

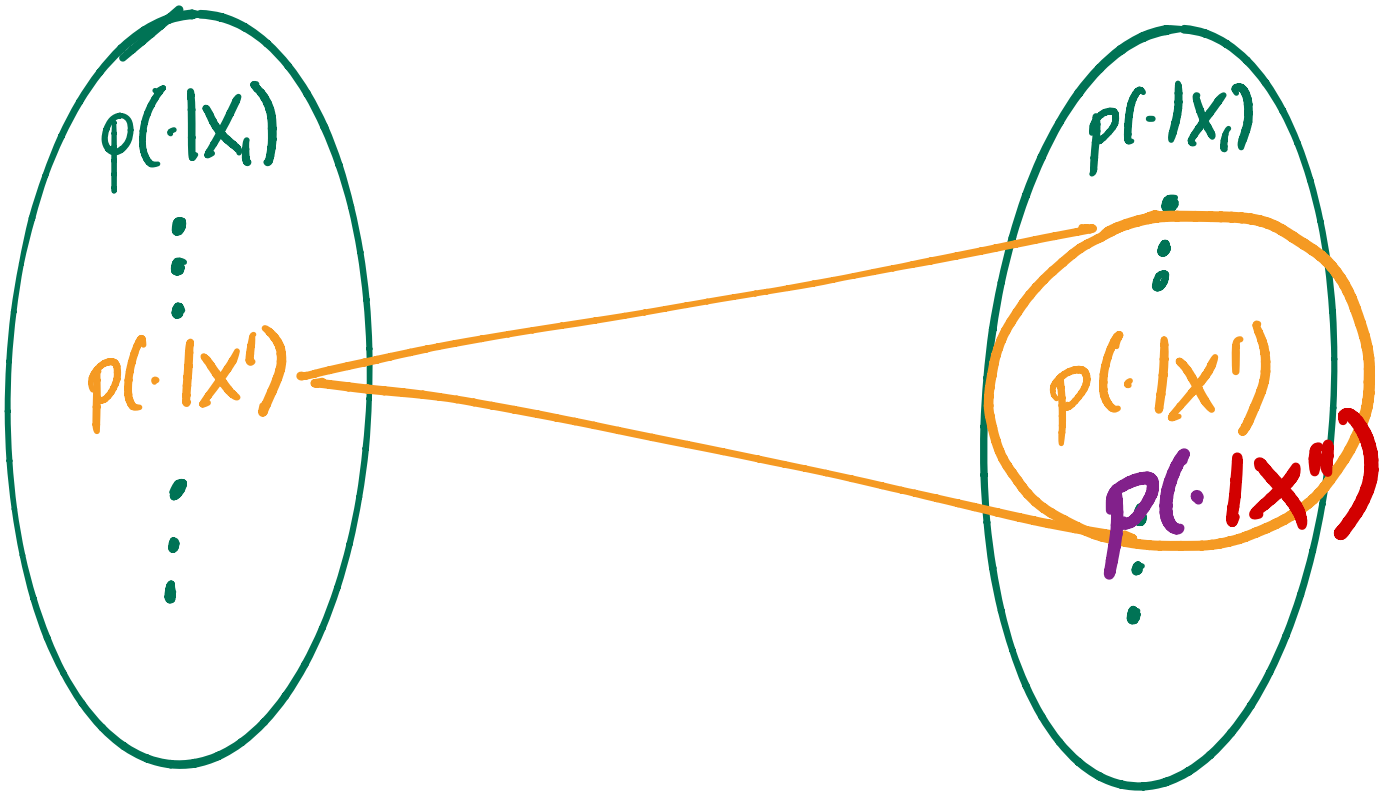
If the algo is robust...

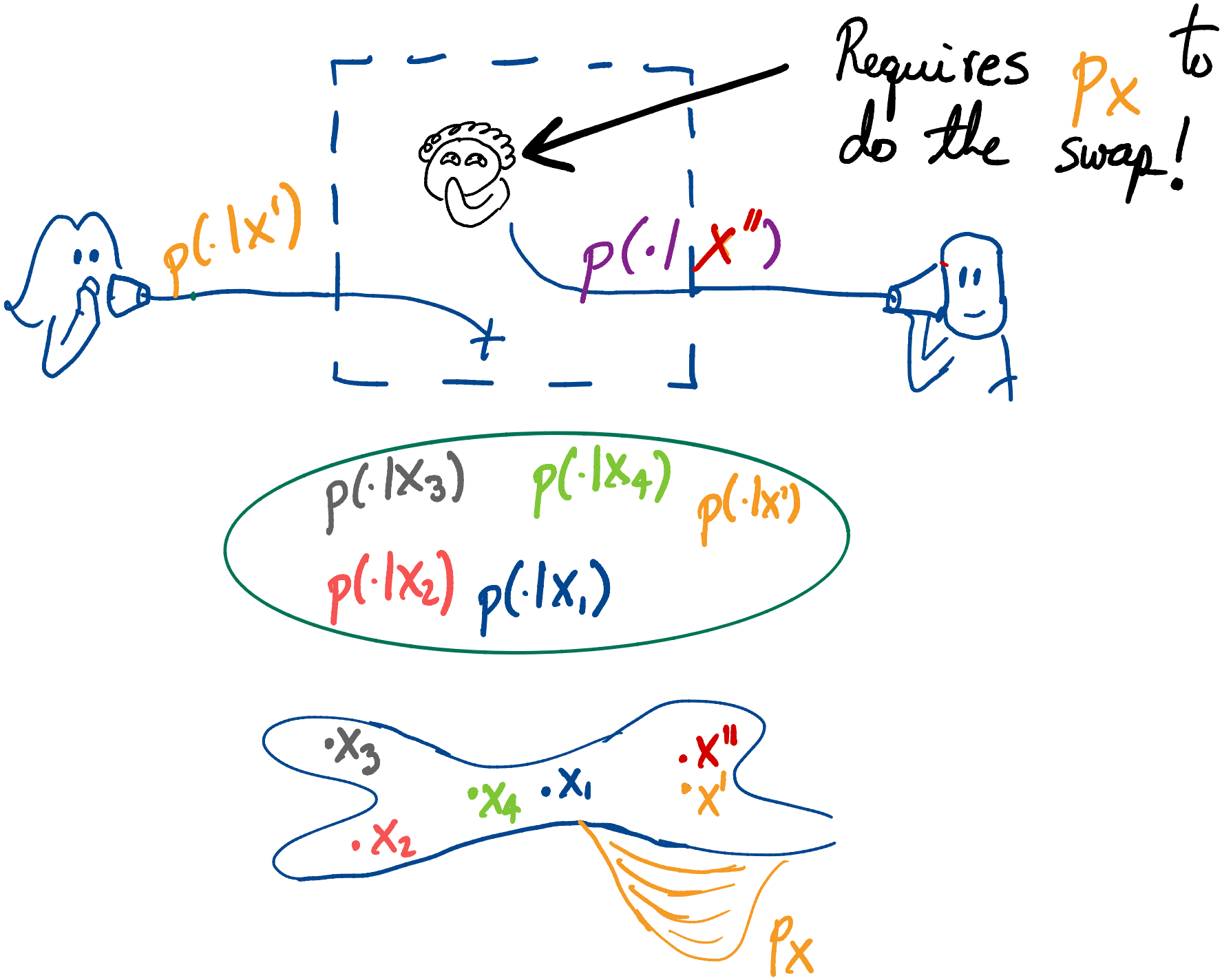




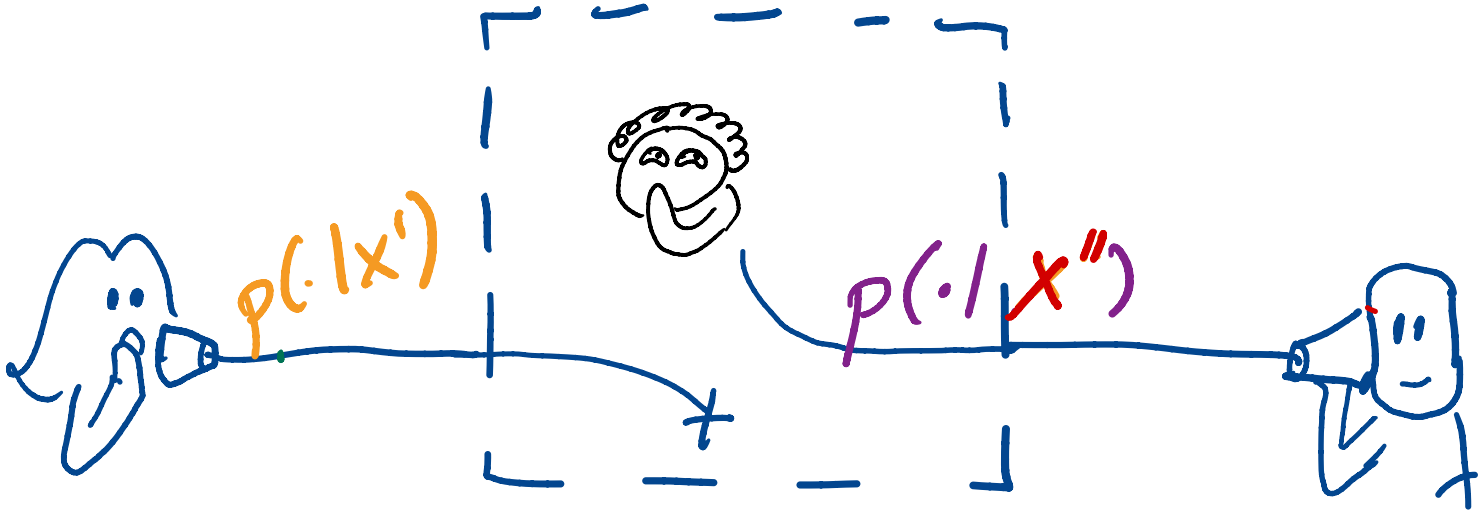
If the algo is not robust...

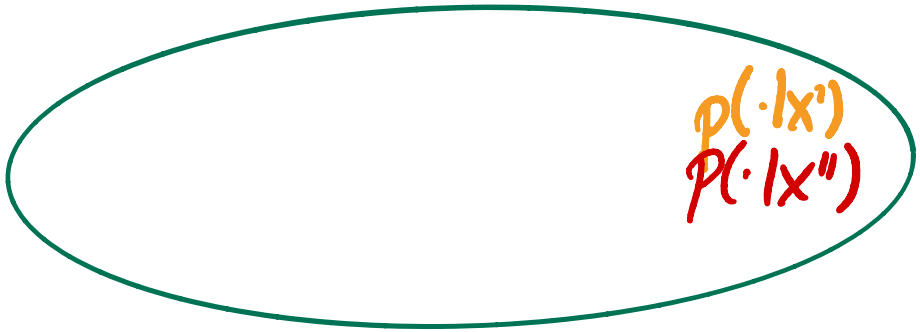


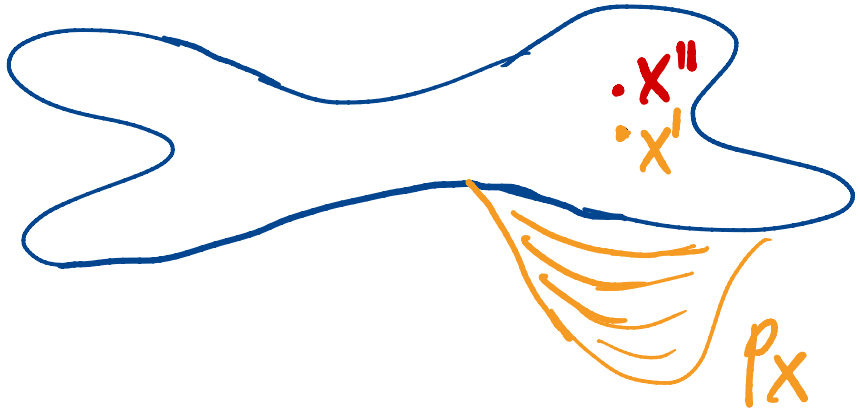



Requires p_X to do the swap!
$p(\cdot|x')$
$p(\cdot|x'')$
$p(\cdot|x_3)$
$p(\cdot|x_4)$
$p(\cdot|x')$
$p(\cdot|x_2)$
$p(\cdot|x_1)$
x_3
x_4
x_1
x''
x'
x_2
p_X

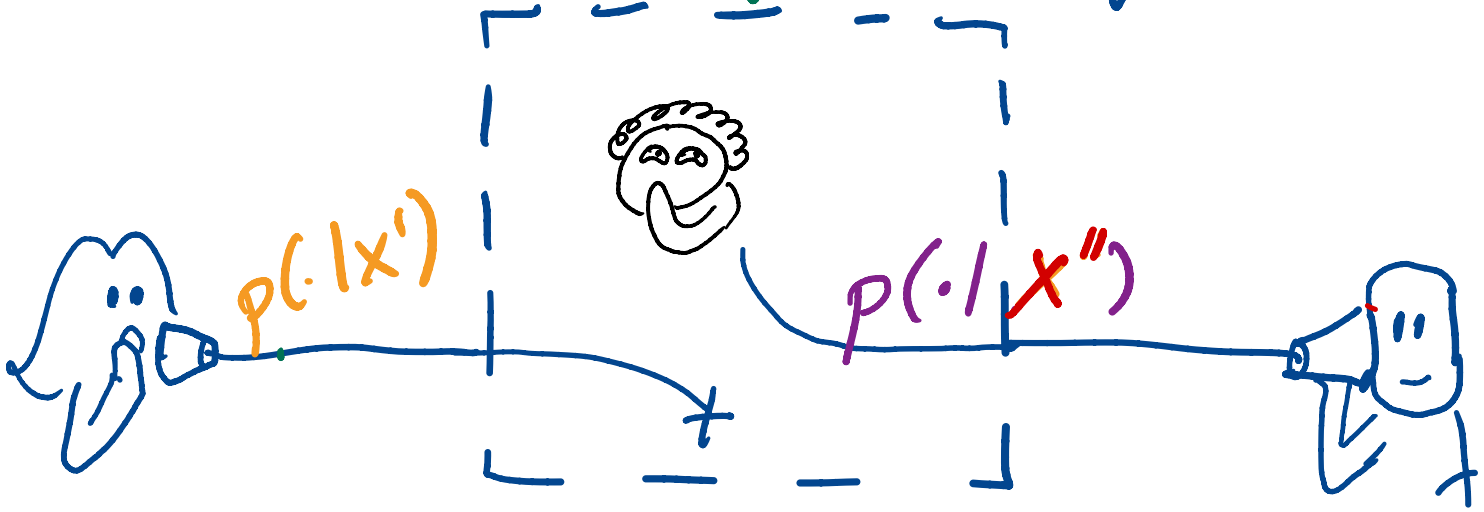
Problem: We only have one experiment p_X

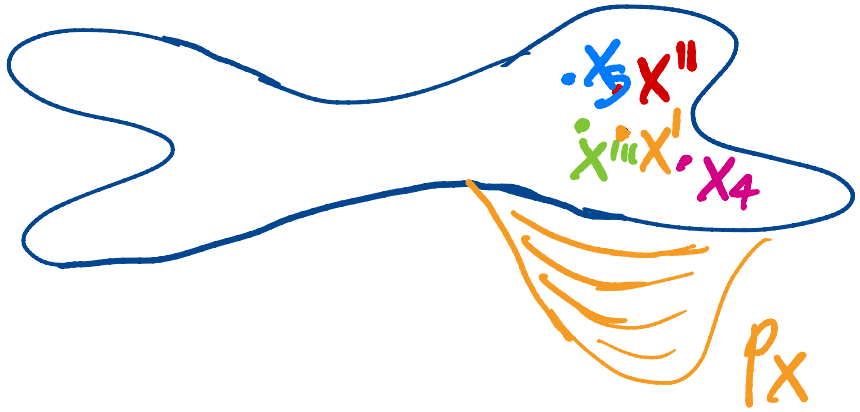




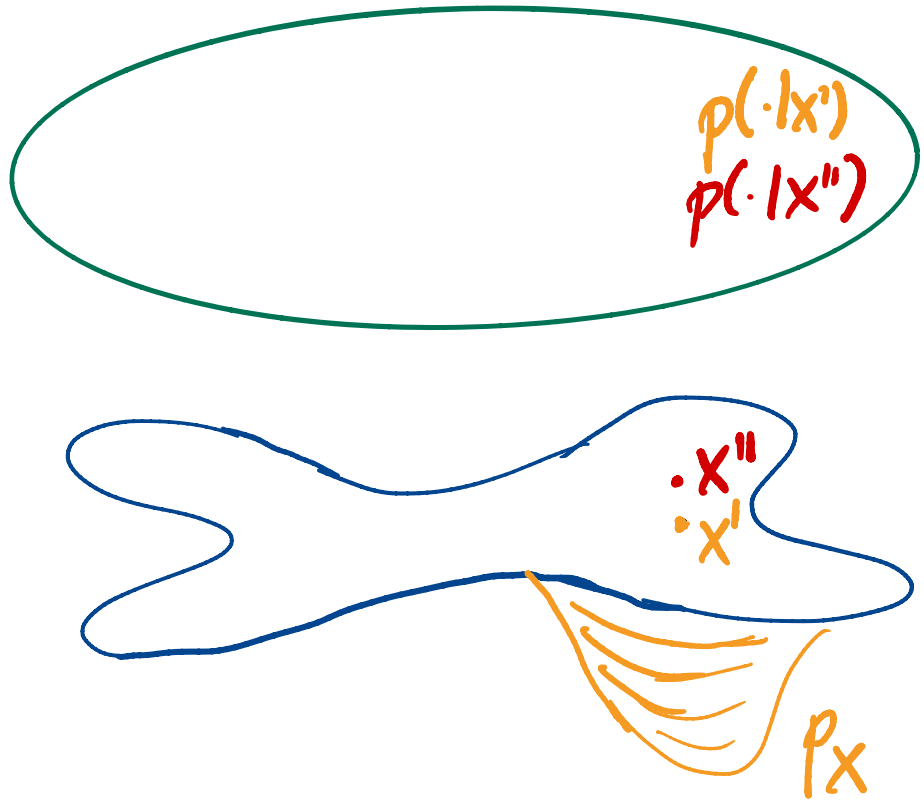


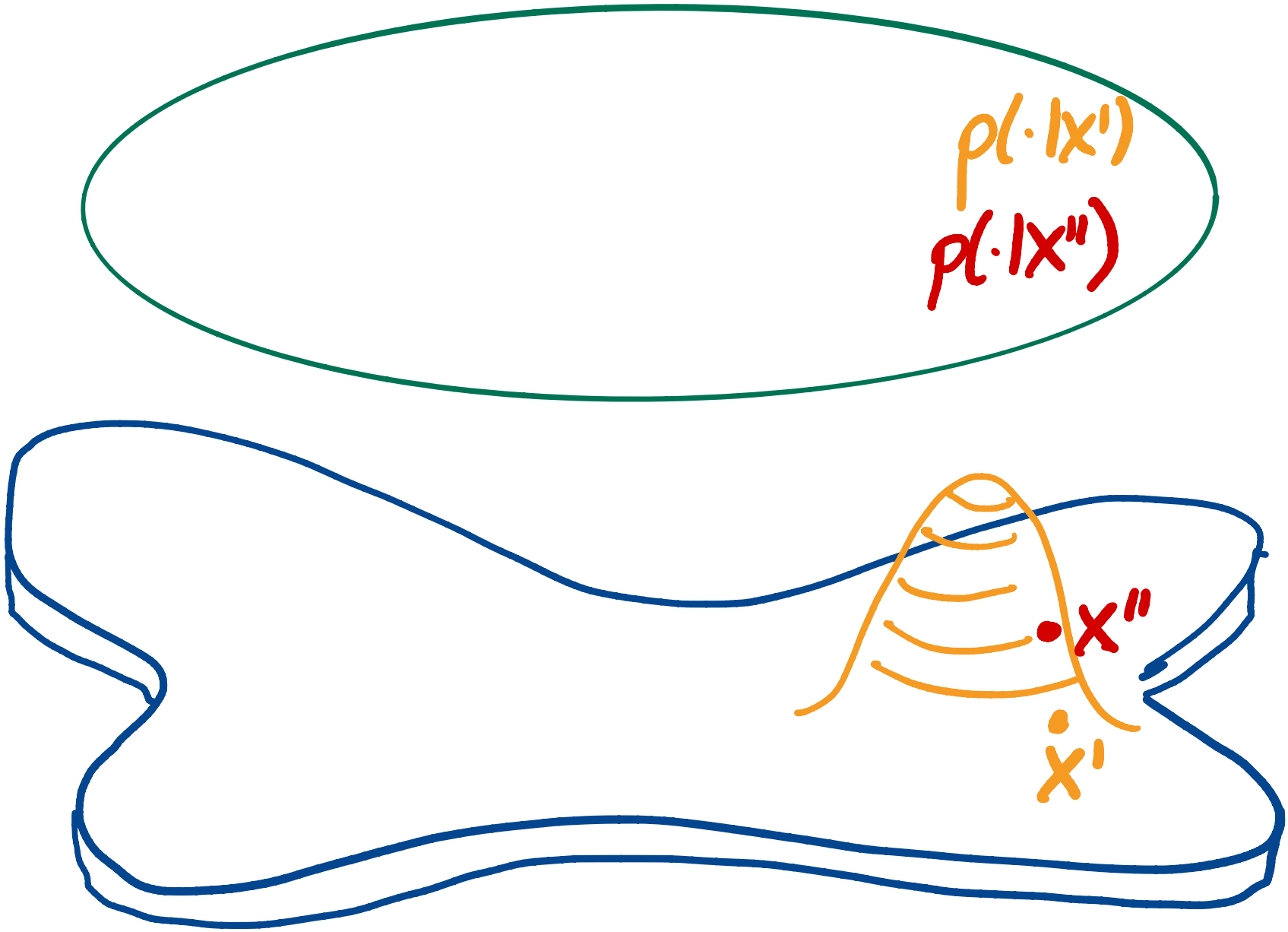
Problem: Robust and bad-content algorithm / Robust and good-content algorithm } yield similar channels

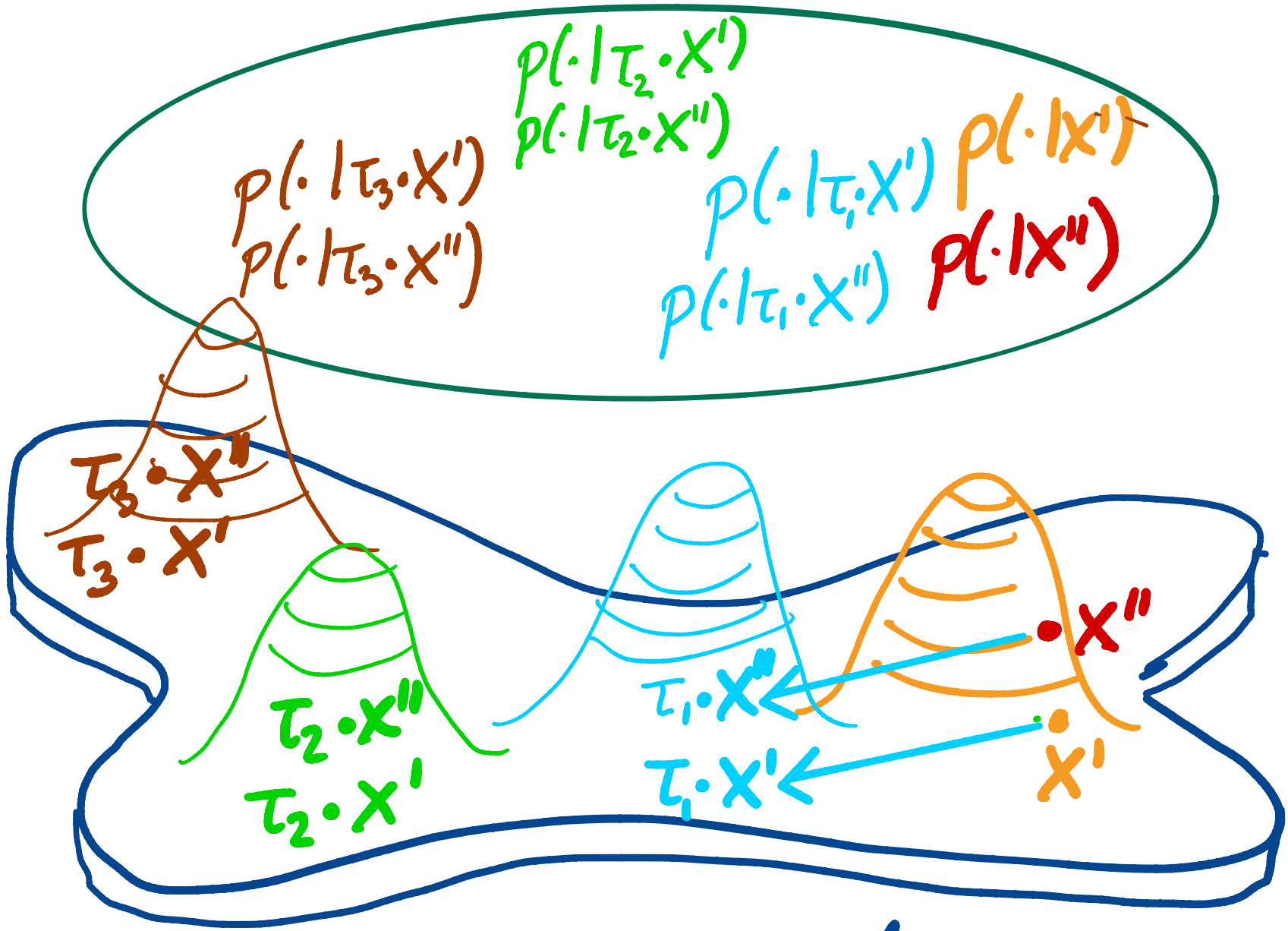




How to get more experiments ?

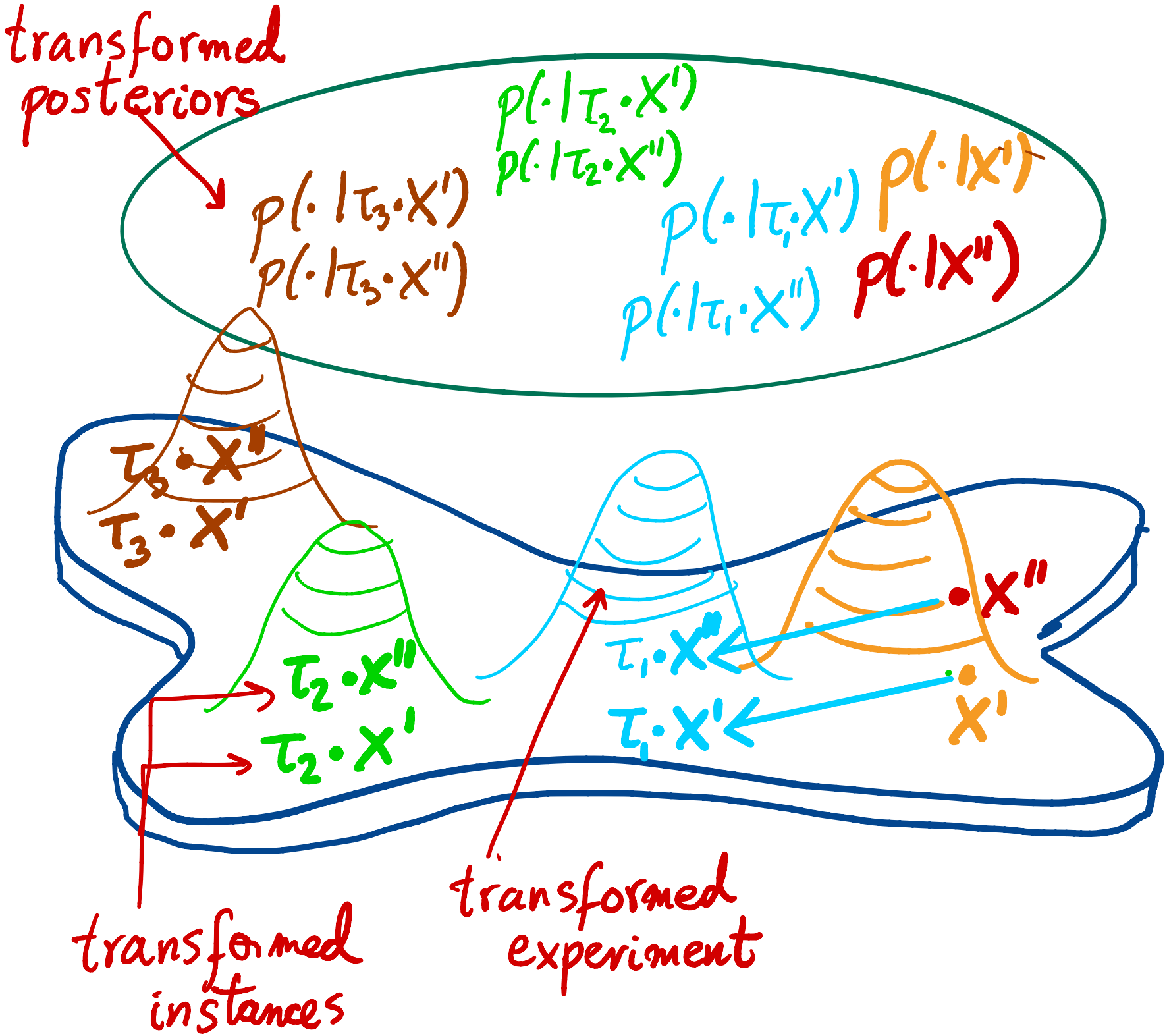


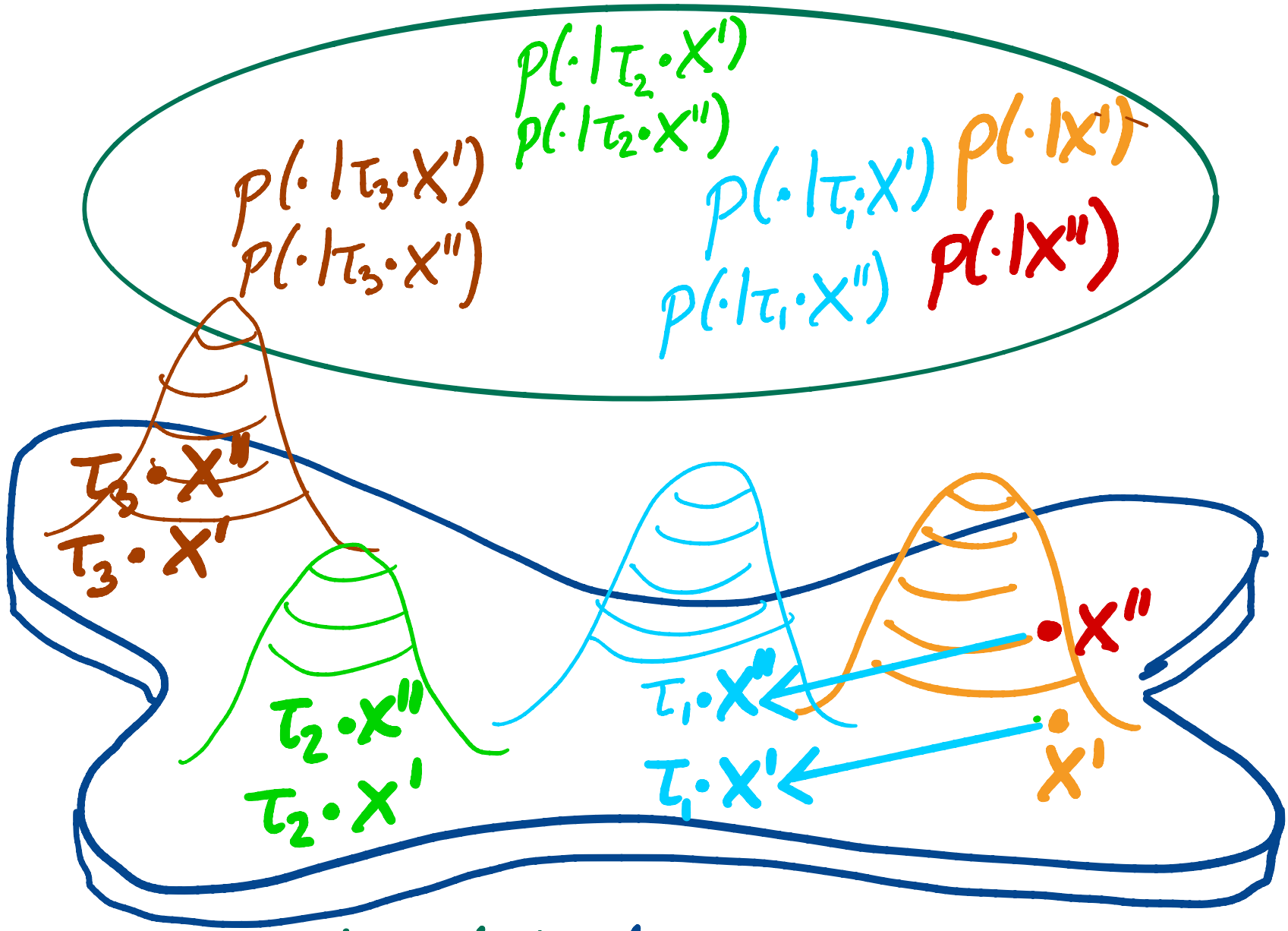

p(·|x')
p(·|x'')
x''
x'



We define a series of transformations

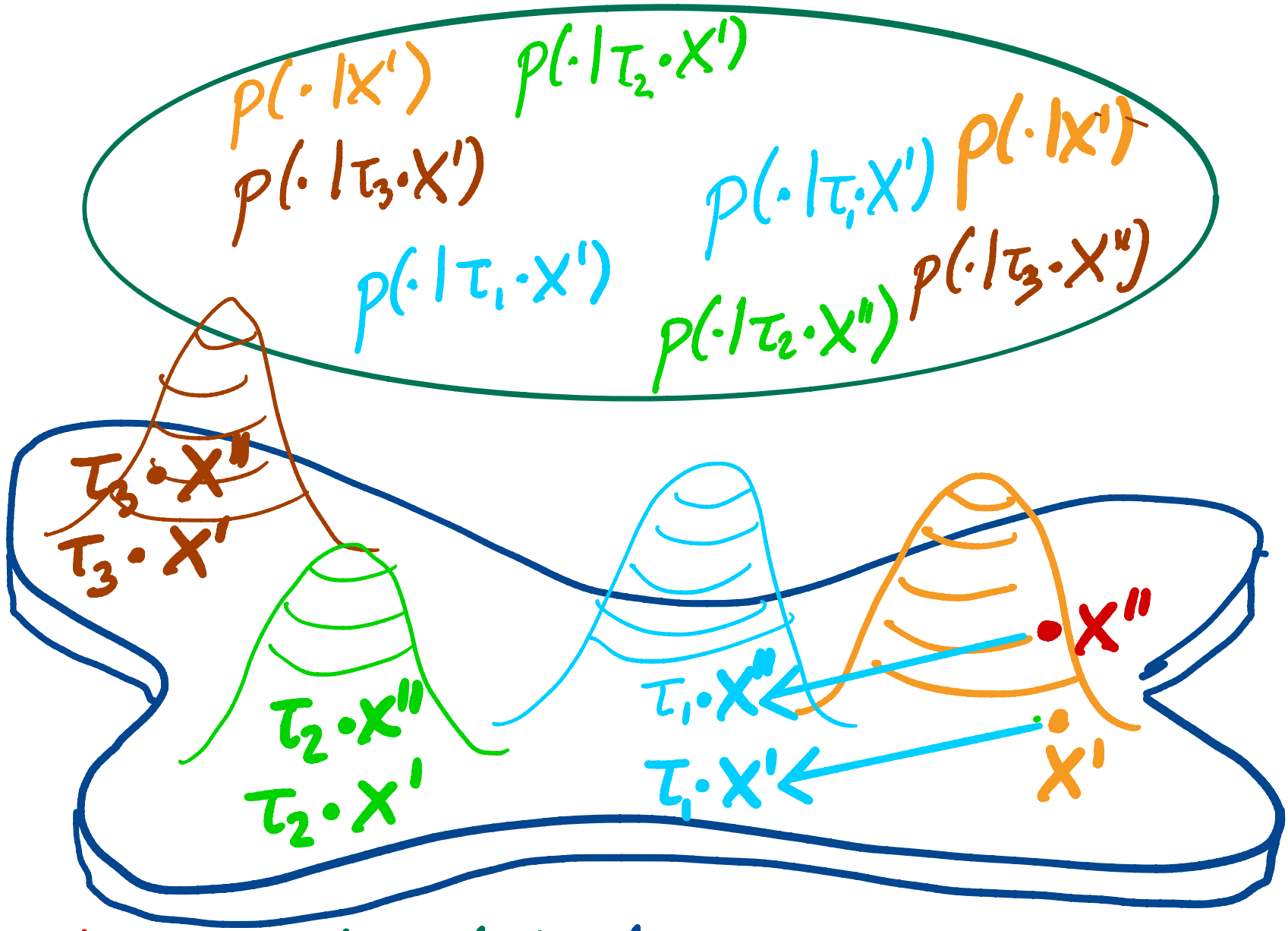
$$\mathcal{T} := \{\tau_1, \tau_2, \tau_3, \ldots\}$$


transformed posteriors
$p(\cdot|\tau_2 \cdot x')$
$p(\cdot|\tau_2 \cdot x'')$
$p(\cdot|\tau_3 \cdot x')$
$p(\cdot|\tau_3 \cdot x'')$
$p(\cdot|\tau_1 \cdot x')$
$p(\cdot|\tau_1 \cdot x'')$
$p(\cdot|x')$
$p(\cdot|x'')$
$\tau_3 \cdot x''$
$\tau_3 \cdot x'$
$\tau_2 \cdot x''$
$\tau_2 \cdot x'$
$\tau_1 \cdot x''$
$\tau_1 \cdot x'$
x''
x'
transformed instances
transformed experiment



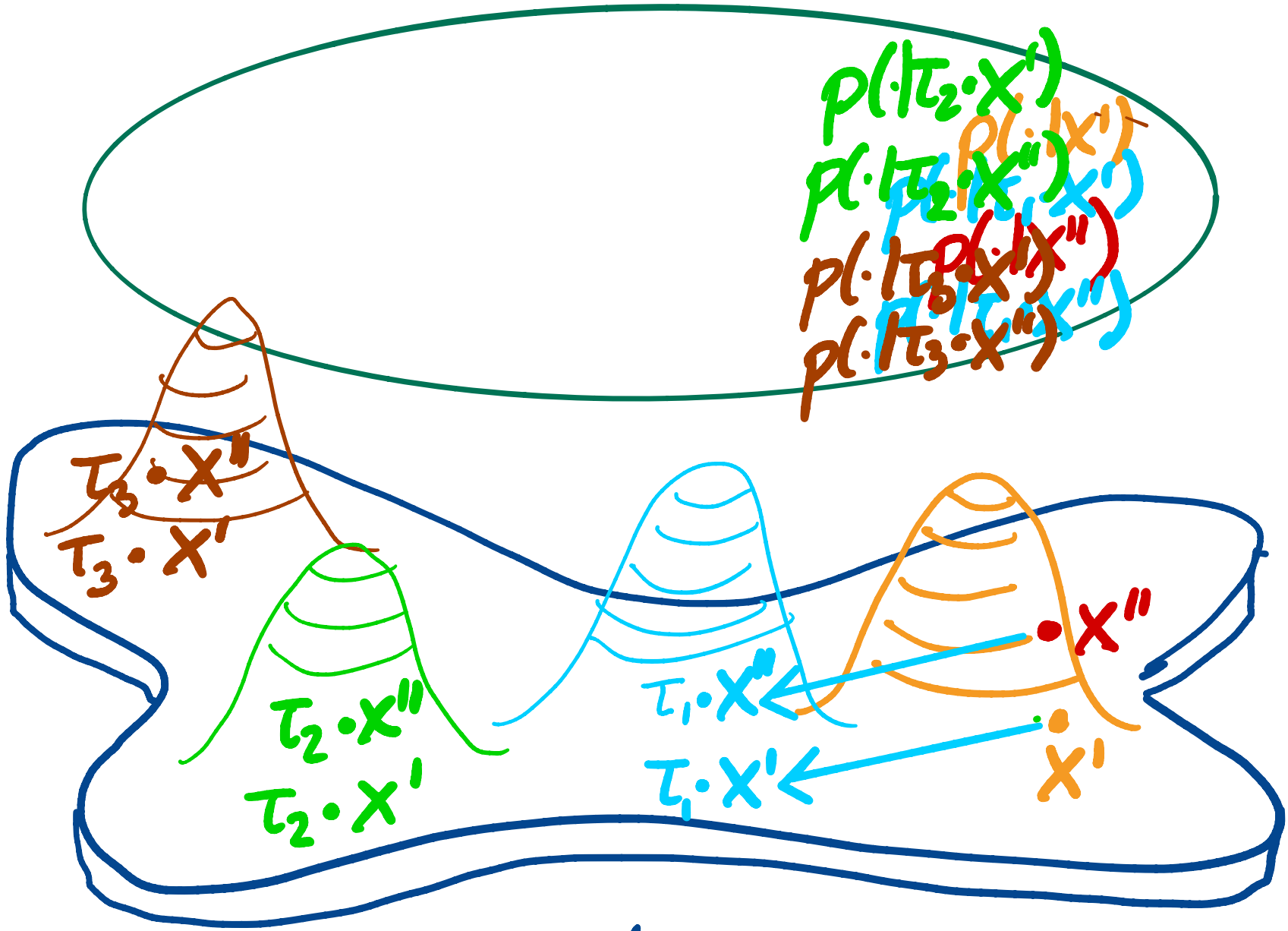
Robust and good-content algos:

* different experiments $\Rightarrow$ distant posteriors
* same experiment $\Rightarrow$ close posteriors



Unstable and good-content algos:

* different experiments $\Rightarrow$ distant posteriors
* same experiment $\Rightarrow$ distant posteriors



Robust and poor-content algos:

* different experiments $\Rightarrow$ close posteriors
* same experiment $\Rightarrow$ close posteriors

Requirements for transformations

The total mass of the transformed posteriors should "uniformly" cover G.

That is $\sum_{\tau} p(c \mid \tau \cdot x^l) \tilde{\approx} \frac{|\Pi|}{|G|}$ ← Total mass of $p(\cdot \mid \tau_1 \cdot x^l), p(\cdot \mid \tau_2 \cdot x^l) \ldots$

$$\sum_{\tau} \underbrace{\sum_{c} p(c \mid \tau \cdot x^l)}_{1} = |\Pi|$$

$$* \sum_{\tau} p(c \mid \tau \cdot x^l) \in \left[\frac{|\Pi|}{|G|}(1-\rho), \frac{|\Pi|}{|G|}(1+\rho) \right],$$

for any $x^l \in X$

Requirements for transformations

* $\| p(\cdot \mid \tau_i \circ X^l) - p(\cdot \mid \tau_j \circ X^l) \|_1 > 0$, for $i \neq j$

That is, different transformations yield different posteriors

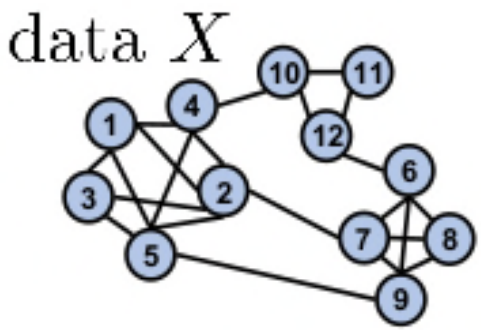
* The experimental measurements of X^l are not changed after a transformation.

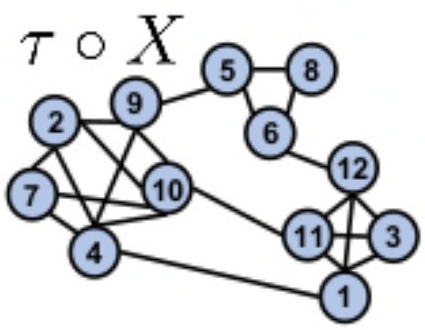
That is, the noise in X^l is not changed after a transformation.

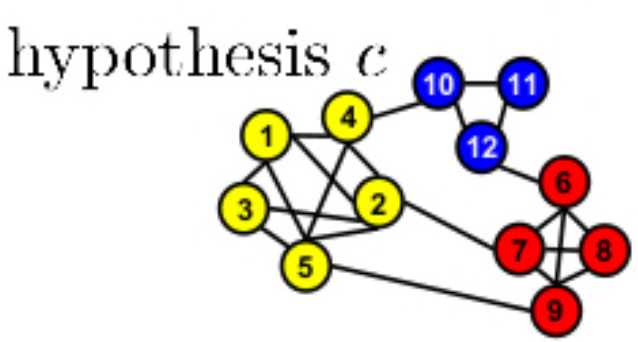
$\underbrace{X', X'', X'''}_{\text{instances of same exp.}}$ $\longleftrightarrow$ diff. exps. $\longrightarrow$ $\underbrace{\tau \circ X', \tau \cdot X'', \tau \cdot X'''}_{\text{instances of same exp.}}$

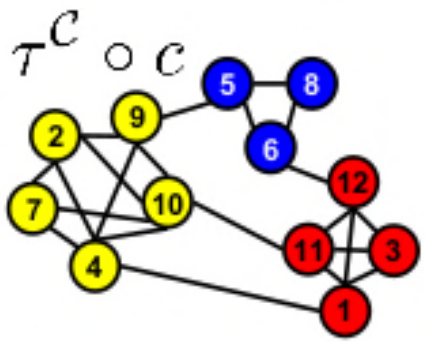
Examples of transformations

mean estimation	Translation δ	$X \overset{\tau}{\mapsto} X + \delta$
linear regression	Scaling s	$(X, Y) \overset{\tau}{\mapsto} (X, sY)$
Graph-based clustering	Vertex permutation π	$(V, E) \overset{\tau}{\mapsto} (V^{\pi}, E)$ $V_i^{\pi} = V_{\pi(i)}$


data X
10
11
4
1
12
6
3
2
7
8
5
9

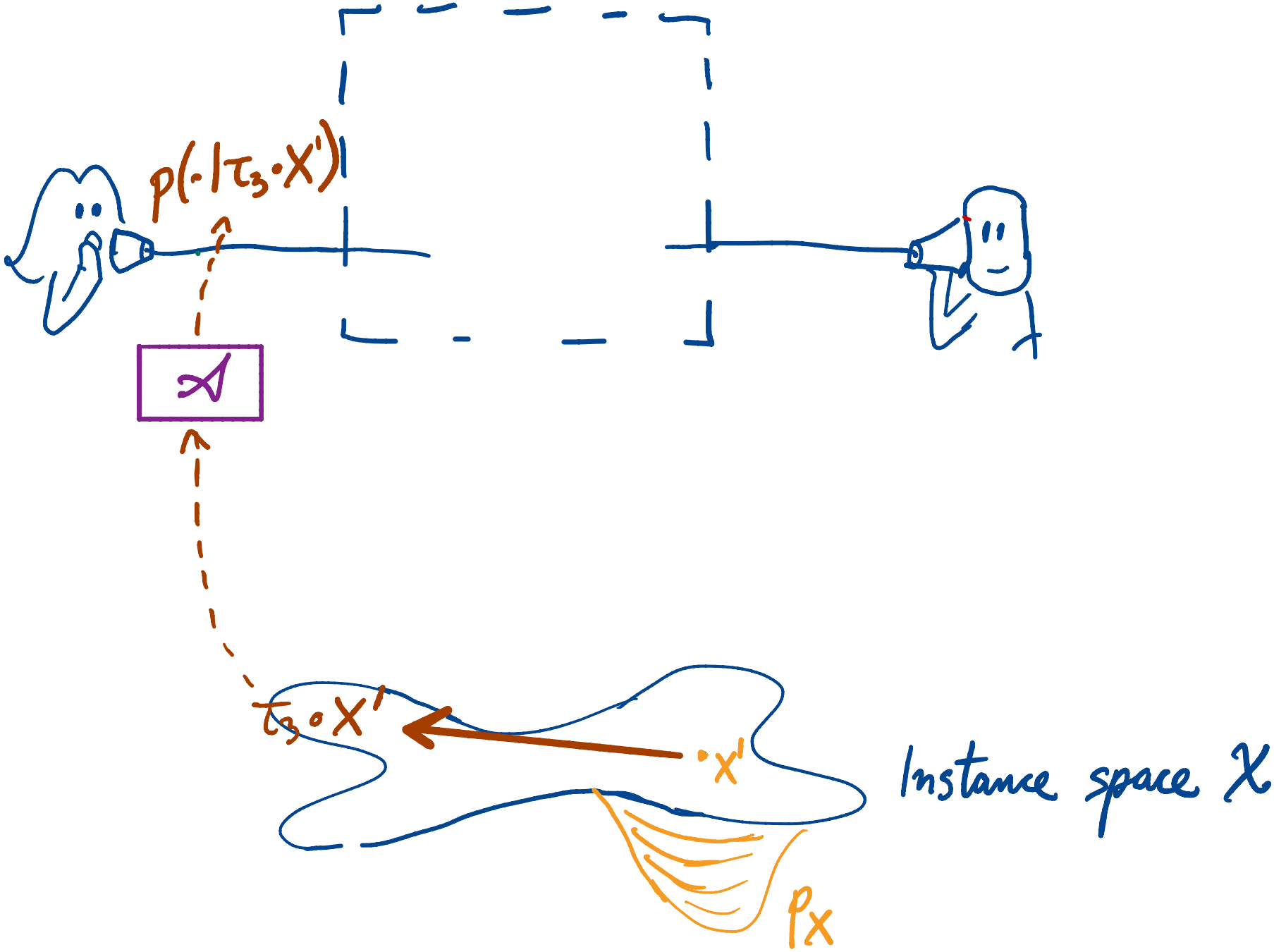

$\tau \circ X$
5
8
9
2
6
12
7
10
11
3
4
1

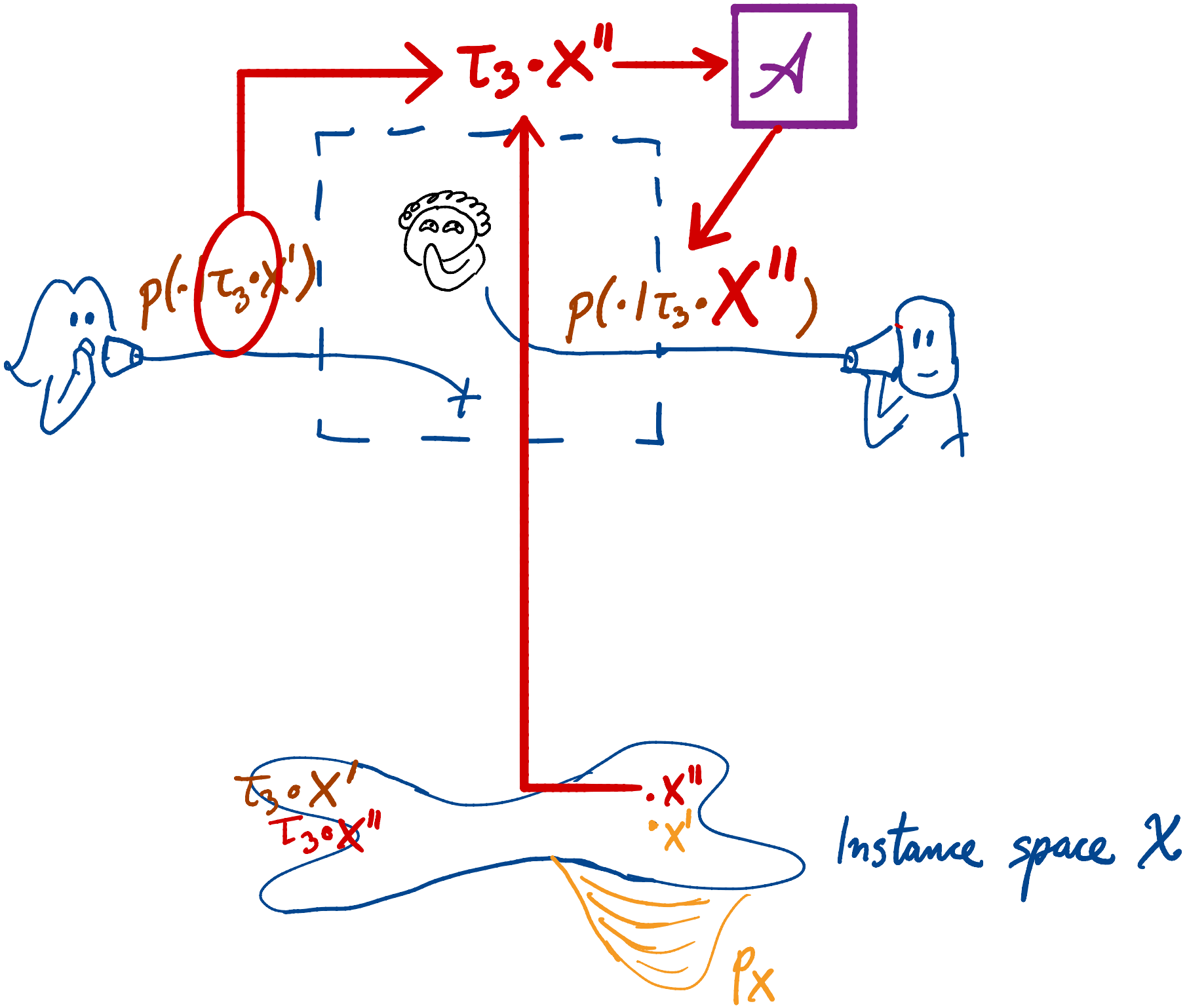

hypothesis c
10
11
4
1
12
6
3
2
7
8
5
9


$\tau^{\mathcal{C}} \circ c$
5
8
9
2
6
12
7
10
11
3
4
1

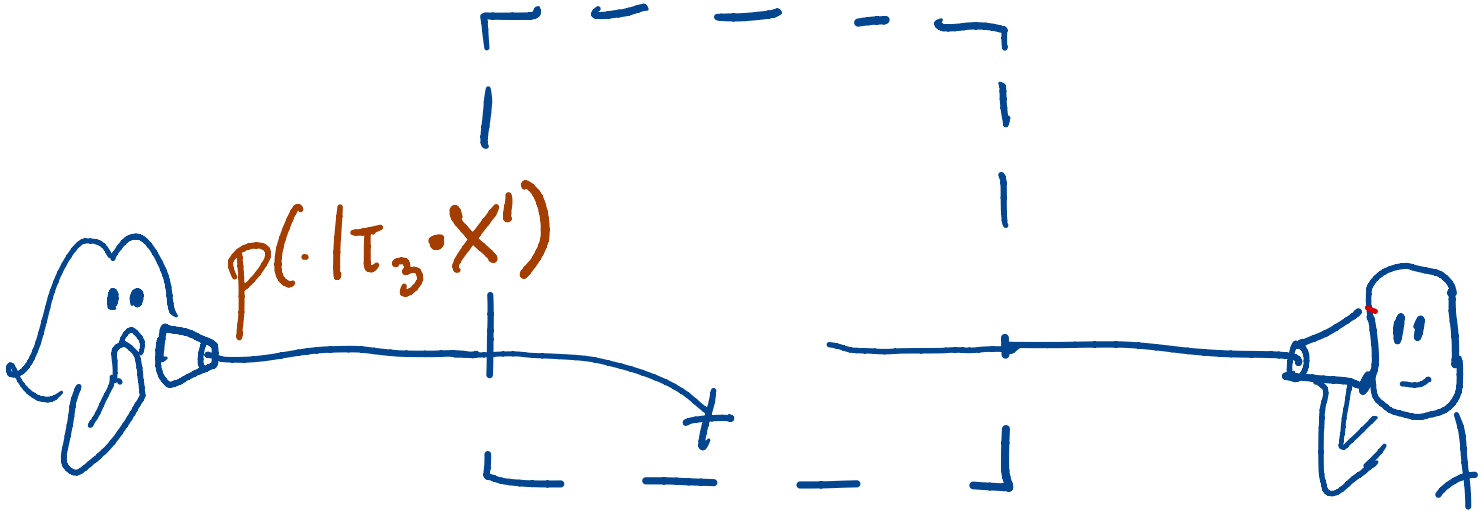

1 2 3 4 5 6 7 8 9 10 11 12


1 2 3 4 5 6 7 8 9 10 11 12


$p(\cdot|\tau_3 \cdot x')$
$\mathcal{A}$
$\tau_3 \circ x'$
x'
Instance space $\mathcal{X}$
$p_{\mathcal{X}}$


$\tau_3 \cdot x''$
$\mathcal{A}$
$p(\cdot|\tau_3 \cdot x')$
$p(\cdot|\tau_3 \cdot x'')$
$\tau_3 \cdot x'$
$\tau_3 \cdot x''$
x''
x'
p_X
Instance space $\mathcal{X}$

Input codeword space



$p(\cdot|\tau_3 \cdot X')$ $p(\cdot|\tau_6 \cdot X')$ $p(\cdot|\tau_4 \cdot X')$

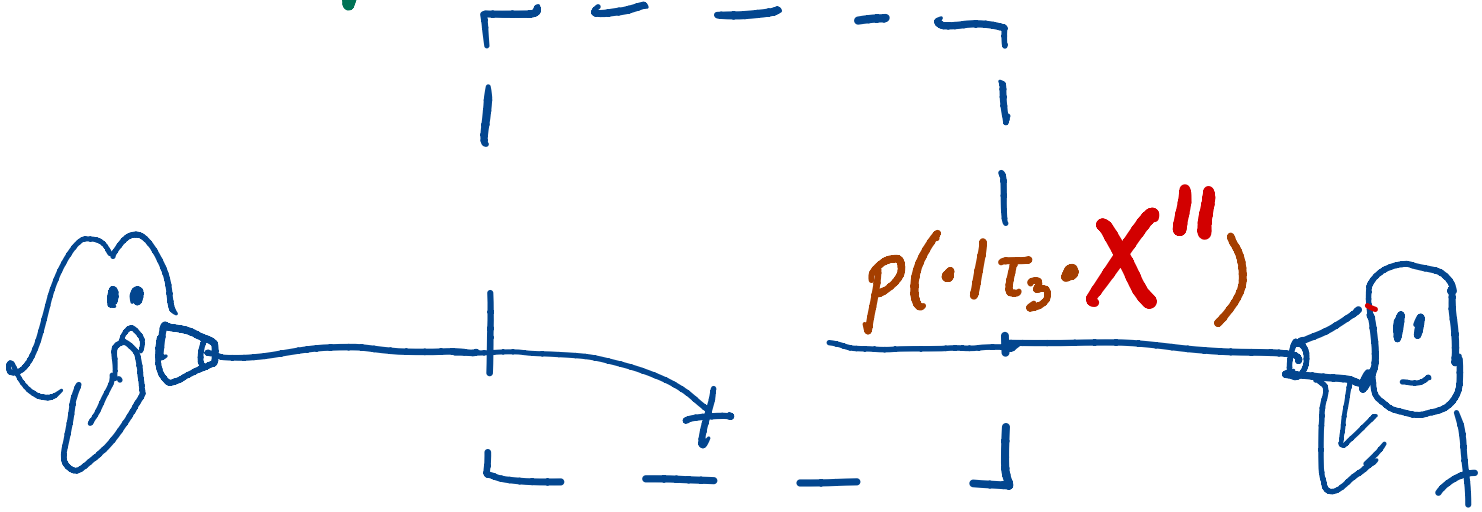
$p(\cdot|\tau_8 \cdot X')$ $p(\cdot|\tau_1 \cdot X')$ $p(\cdot|\tau_0 \cdot X')$

$p(\cdot|\tau_7 \cdot X')$

$p(\cdot|\tau_2 \cdot X')$

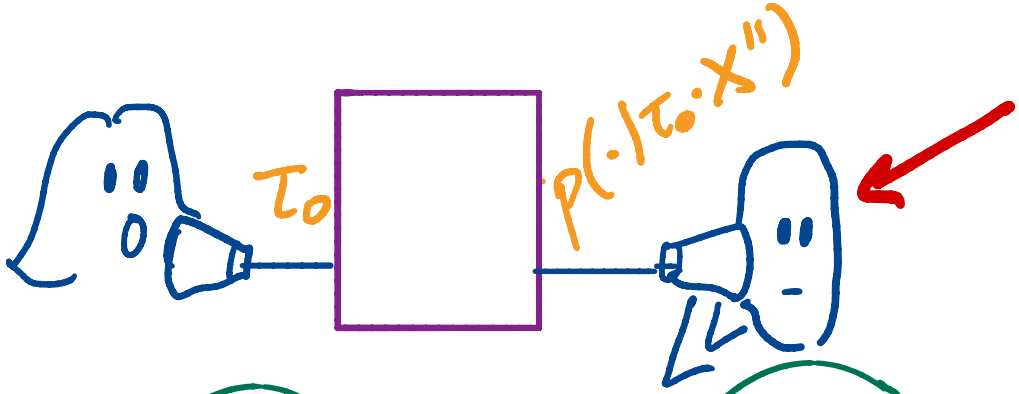
$p(\cdot|\tau_5 \cdot X')$ $p(\cdot|\tau_9 \cdot X')$

Output codeword space



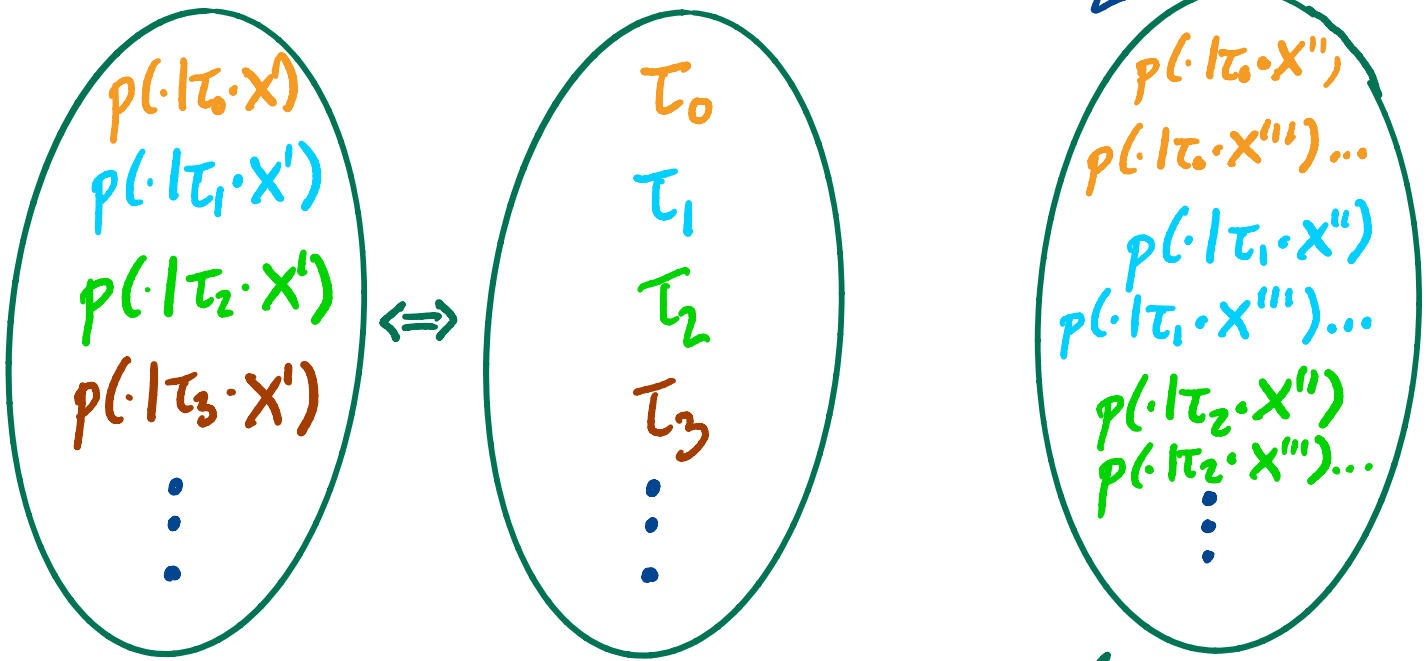
$p(\cdot|\tau_3 \cdot X'')$ $p(\cdot|\tau_6 \cdot X'')$ $p(\cdot|\tau_4 \cdot X'')$ $p(\cdot|\tau_8 \cdot X'')$ $p(\cdot|\tau_1 \cdot X'')$ $p(\cdot|\tau_0 \cdot X'')$ $p(\cdot|\tau_7 \cdot X'')$ $p(\cdot|\tau_2 \cdot X'')$ $p(\cdot|\tau_5 \cdot X'')$ $p(\cdot|\tau_9 \cdot X'')$

The algorithm as a channel



Remember:

Receiver knows input and output codeword spaces

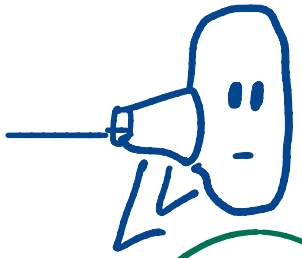


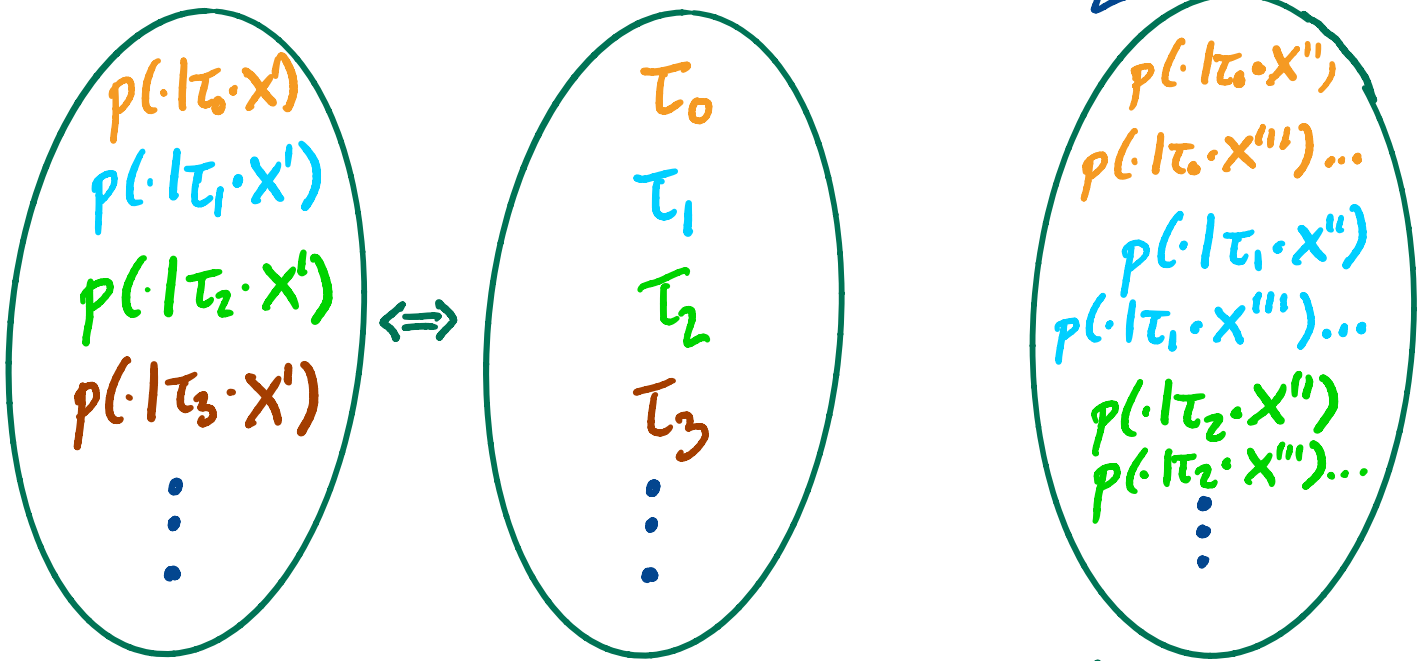
input codeword space $= \Pi$ $P(I, X)$ $=$ output codeword space

$|\Pi| \ll |P(I, X)|$

Receiver's knowledge

* Receiver knows X', Π, A. That is, he knows input and output codeword spaces and the channel's conditional distribution.





input codeword space $= \Pi$ $P(\Pi, \mathcal{X})$ $=$ output codeword space

$|\Pi| \ll |P(\Pi, \mathcal{X})|$

The algorithm as a channel

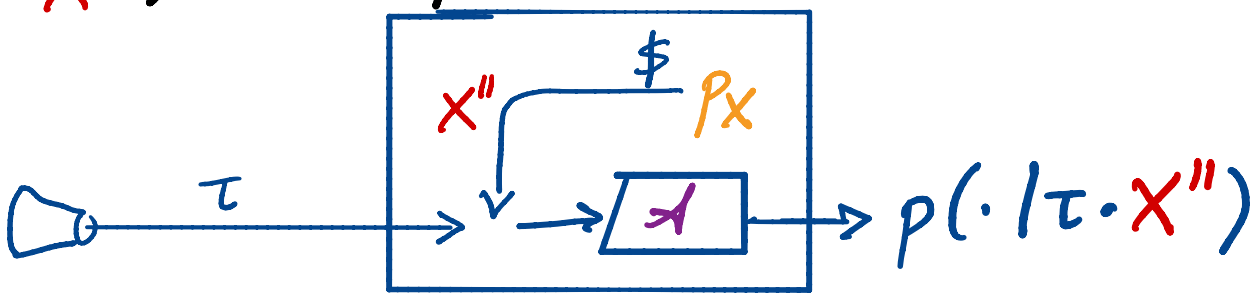
* Input alphabet:

$$\mathbb{I} = \{\tau_0, \tau_1, \ldots, \tau_S\}.$$

* Output alphabet:

$$\mathcal{P}(\mathbb{I}, \mathcal{X}) := \{p(\cdot|\tau \circ x'') : \tau \in \mathbb{I},\ x'' \in \mathcal{X}\}$$

* When τ is input, a new instance x'' is drawn $\sim P_X$ and $p(\cdot|\tau \circ x'')$ is output.



Organization

What is PA?

Roadmap

Formalization of PA

→ Algorithms as channels

→ Derivation of PA

Applications

Derivation of PA

* We follow Shannon's strategy.
  Define a code for M messages by choosing M codewords at random.

* We demonstrate that the prob. of a comm. error

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

* We compare this with Shannon's code's prob. of comm. error

$$\leq 2^{-n\left(\text{capacity} - \frac{\log M}{n} - 3\varepsilon\right)}$$

* Therefore, we propose the exp. log. PA as a capacity measure.

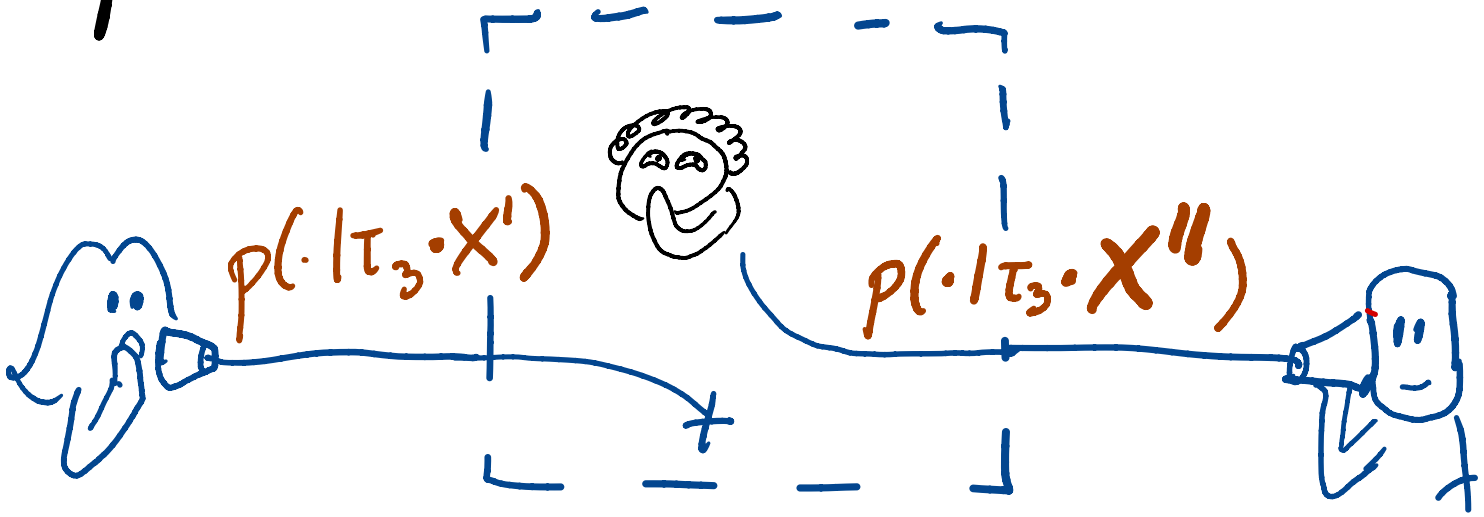
Remember:

A code is a pair (Enc, Dec),

$Enc: \{1 \dots M\} \longrightarrow$ input codeword space

$Dec:$ output codeword space $\longrightarrow \{1, \dots, M\}$

We follow Shannon's random code

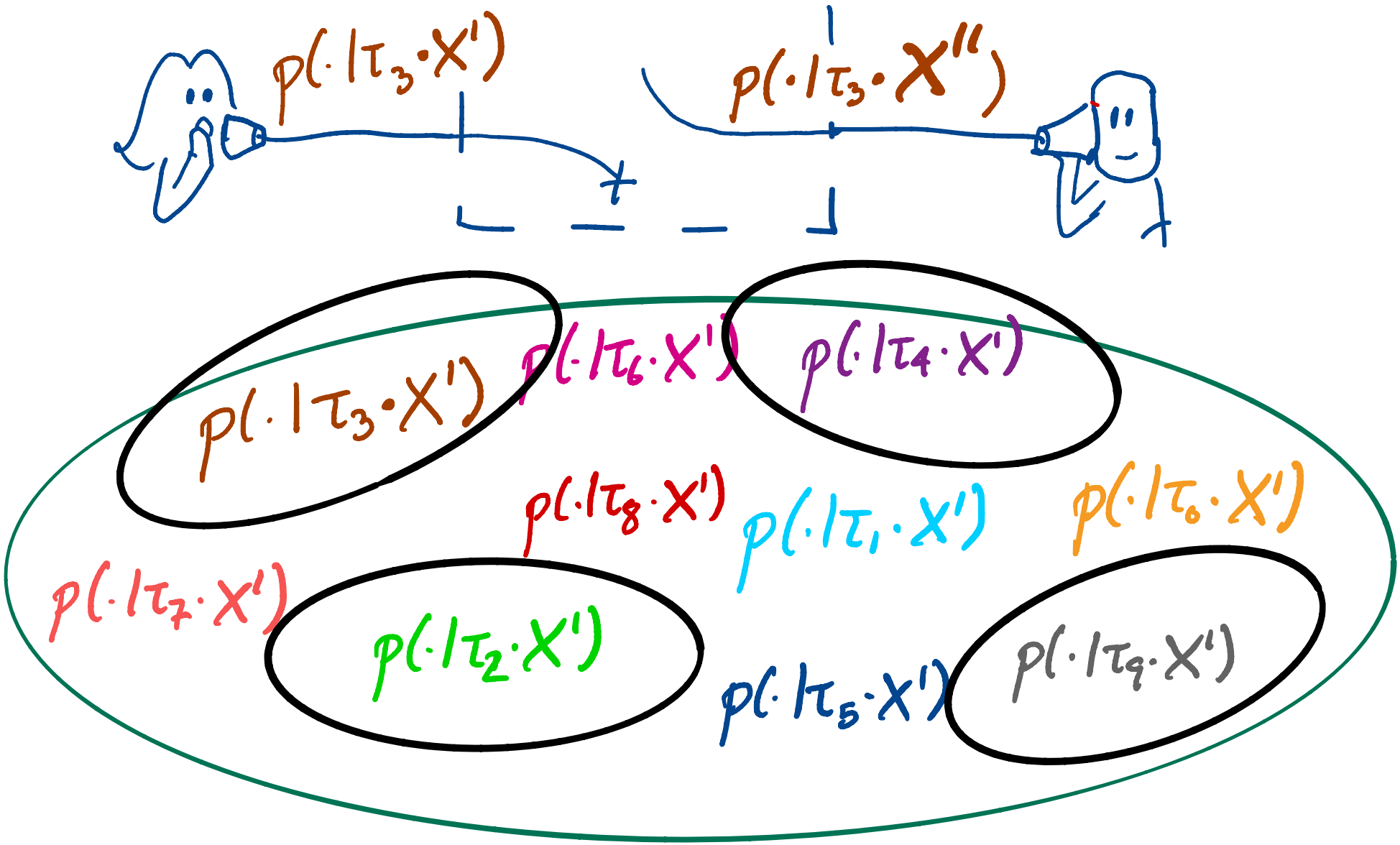


$p(\cdot|\tau_3 \cdot X')$ $p(\cdot|\tau_6 \cdot X')$ $p(\cdot|\tau_4 \cdot X')$

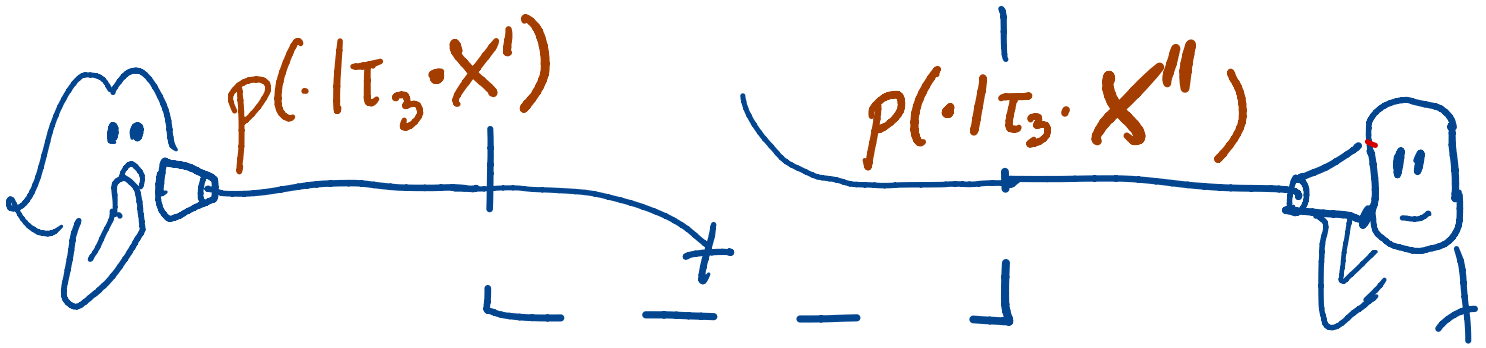
$p(\cdot|\tau_8 \cdot X')$ $p(\cdot|\tau_1 \cdot X')$ $p(\cdot|\tau_0 \cdot X')$

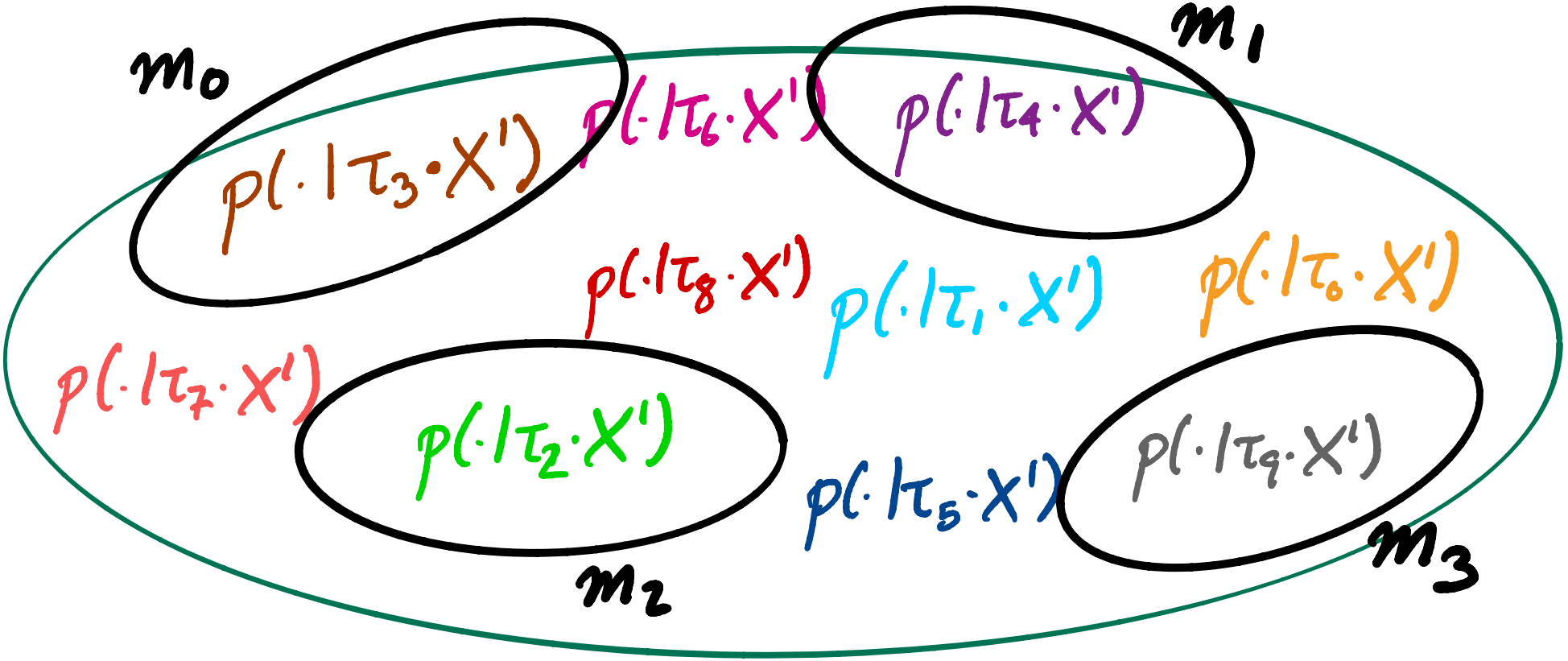
$p(\cdot|\tau_7 \cdot X')$ $p(\cdot|\tau_2 \cdot X')$ $p(\cdot|\tau_5 \cdot X')$ $p(\cdot|\tau_9 \cdot X')$

Let M be # msgs.
Pick M codewords at random.

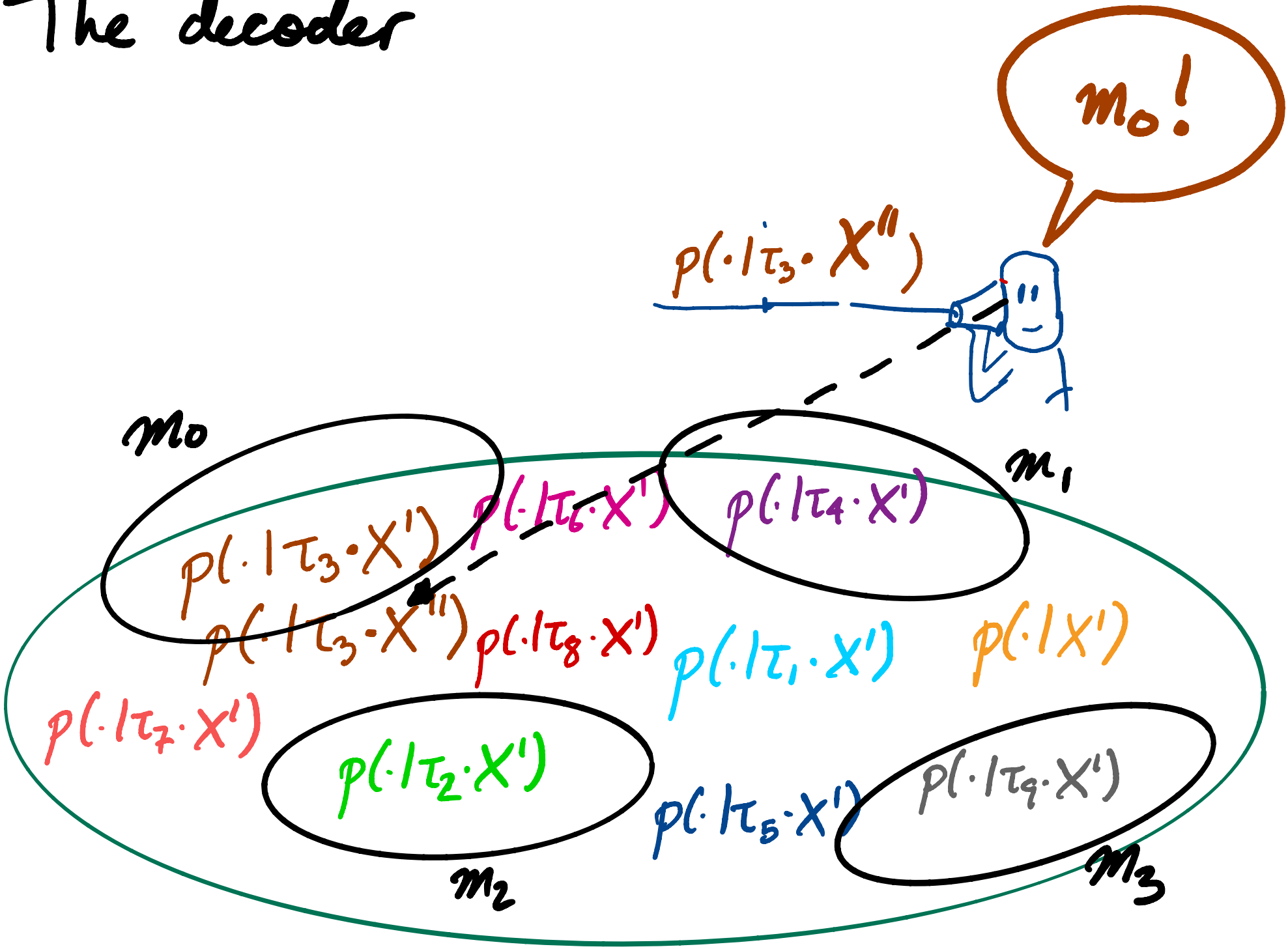


The encoder function maps each message to each of these codewords.





The decoder

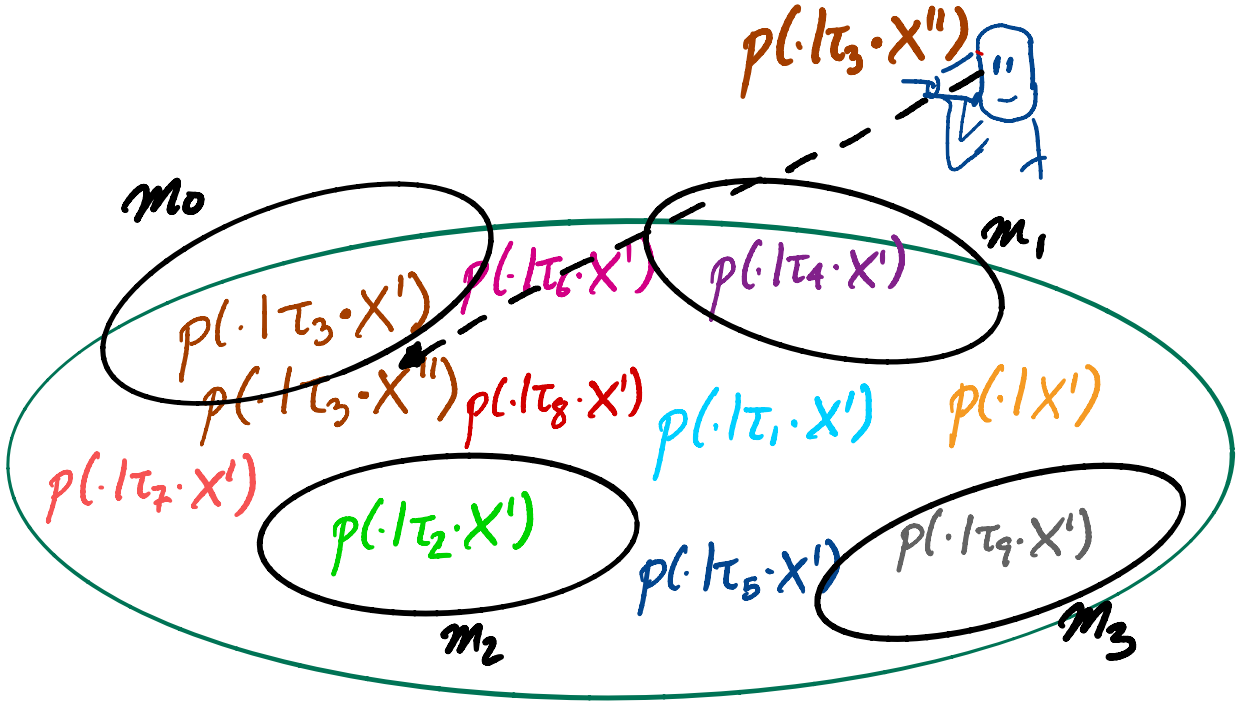

m_0!
$p(\cdot|\tau_3 \cdot X'')$
m_0
m_1
m_2
m_3
$p(\cdot|\tau_3 \cdot X')$
$p(\cdot|\tau_6 \cdot X')$
$p(\cdot|\tau_4 \cdot X')$
$p(\cdot|\tau_3 \cdot X'')$
$p(\cdot|\tau_8 \cdot X')$
$p(\cdot|\tau_1 \cdot X')$
$p(\cdot|X')$
$p(\cdot|\tau_7 \cdot X')$
$p(\cdot|\tau_2 \cdot X')$
$p(\cdot|\tau_5 \cdot X')$
$p(\cdot|\tau_9 \cdot X')$

The decoder outputs

$$\arg\max_{m \in M} \langle \mathrm{Enc}(m), p(\cdot \mid \tau_3 \cdot X'') \rangle$$

$$= \arg\max_{\tau} \langle p(\cdot \mid \tau \cdot X'), p(\cdot \mid \tau_3 \cdot X'') \rangle$$

$$= \arg\max_{\tau} K(\tau \circ X', \tau_3 \circ X'')$$



Reference: Shannon's random code

* Let $M \in \mathbb{N}$ and $X' \xleftarrow{\$} P_X$
* Choose $\tau_1, \ldots, \tau_M$ from Π uniformly at random.
* Sender and receiver learn $\{\tau_1, \ldots, \tau_M\}$, X', and $\mathcal{A}$.
* Encoder $\quad Enc : \{1 \ldots M\} \longrightarrow \{\tau_1 \ldots \tau_m\}$.

$$i \longmapsto \tau_i$$

* Decoder $\quad Dec_{X'} : P(\Pi, \mathcal{X}) \longrightarrow \{1, \ldots, M\}$

$$p(\cdot \mid \tau \cdot X'') \longmapsto \hat{\tau}$$

where $\hat{\tau} = \underset{\tau' \in \{\tau_1 \ldots \tau_M\}}{\arg\max} K(\tau' \cdot X', \tau \cdot X'')$

The receiver needs X', $\{\tau_1 \ldots \tau_M\}$, and $\mathcal{A}$ to do this!

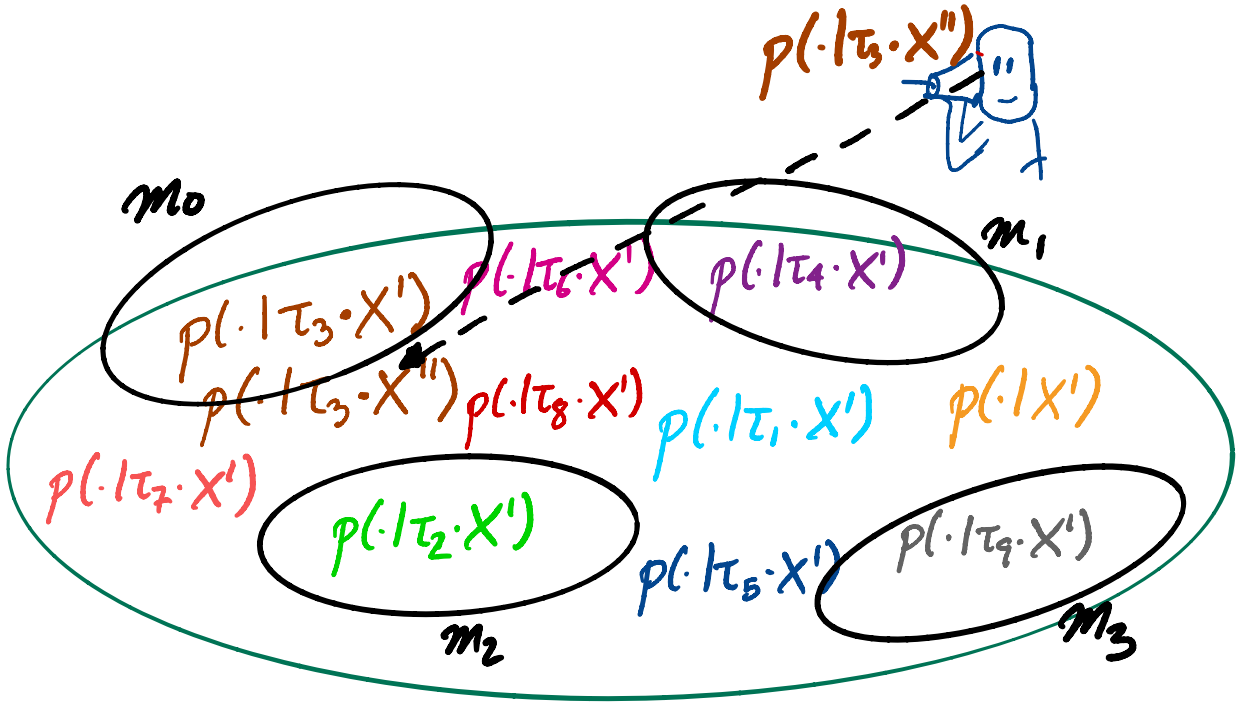
Communication error

$$\arg\max_{\tau} K(\tau \circ X', \tau_3 \circ X'') \neq \tau_3$$

i.e.,

$$K(\tau \cdot X', \tau_3 \cdot X'') \geq K(\tau_3 \cdot X', \tau_3 \cdot X'')$$

for some $\tau \neq \tau_3$



Probability of a communication error

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E} \mid \tau_1)$$

sender sent $\tau_1 = id$

$$= \Pr\left(K(\hat{\tau}\cdot X', \tau_1 \cdot X'') \geq K(\tau_1 \cdot X', \tau_1 \cdot X''),\ \hat{\tau} \neq \tau_1 \mid \tau_1\right)$$

$$= \Pr\left[\begin{array}{l} K(\tau_2 \cdot X', \tau_1 \cdot X'') \geq K(\tau_1 \cdot X', \tau_1 \cdot X'') \\ \vdots \\ K(\tau_M \cdot X', \tau_1 \cdot X'') \geq K(\tau_1 \cdot X', \tau_1 \cdot X'') \end{array} \,\middle|\, \tau_1\right]$$

Apply the union bound

$$= \Pr\Big[\, K(\tau_2 \circ X', \tau_1 \circ X'') \geq K(\tau_1 \circ X', \tau_1 \circ X'') \text{ or}$$
$$K(\tau_3 \circ X', \tau_1 \circ X'') \geq K(\tau_1 \circ X', \tau_1 \circ X'') \text{ or}$$
$$\vdots$$
$$K(\tau_M \circ X', \tau_1 \circ X'') \geq K(\tau_1 \circ X', \tau_1 \circ X'') \,\Big| \, \tau_1 \Big]$$

$$= \Pr\Big[\, K(\tau_2 \circ X', X'') \geq K(X', X'') \text{ or}$$
$$K(\tau_3 \circ X', X'') \geq K(X', X'') \text{ or}$$
$$\vdots$$
$$K(\tau_M \circ X', X'') \geq K(X', X'') \,\Big| \, \tau_1 \Big]$$

$$\leq \sum_{1<j\leq M} \Pr\big(K(\tau_j \circ X', X'') \geq K(X', X'') \mid \tau_1\big)$$

Apply Markov's inequality

$$\leq \sum_{1<j\leq M} \Pr\left[K(\tau_j \circ X', X'') \geq K(X', X'') \mid \tau_1\right]$$

randomness: $X', X'', \tau_1, \ldots, \tau_m$

$$= \sum_{1<j\leq M} \sum_{X', X''} p(X', X'') \Pr\left[K(\tau_j \circ X', X'') \geq K(X', X'') \mid \tau_1, X', X''\right]$$

$$\leq$$

$$\leq \sum_{1<j\leq M} \sum_{X', X''} p(X', X'') \underbrace{\mathbb{E}_{\tau_j}\left[K(\tau_j \circ X', X'') \mid \tau_1, X', X''\right]}_{K(X', X'')}$$

$$= \sum_{1<j\leq M} \mathbb{E}_{X', X''}\left[\frac{\mathbb{E}_{\tau_j}\left[K(\tau_j \circ X', X'') \mid \tau_1, X', X''\right]}{K(X', X'')}\right]$$

Bound the numerator

$$\leq \sum_{1<j\leq M} \mathbb{E}_{X',X''}\left[\frac{\mathbb{E}_{\tau_j}[\kappa(\tau_j\circ X', X'') \mid \tau_1, X', X'']}{\kappa(X', X'')}\right]$$

$$= \mathbb{E}_{\tau_j}\left[\sum_c p(c \mid \tau_j \circ X')\, p(c \mid X'') \mid \tau_1, X', X''\right]$$

$$= \sum_c p(c \mid X'')\, \mathbb{E}_{\tau_j}[p(c \mid \tau_j \circ X') \mid X']$$

$$= \sum_c p(c \mid X'') \sum_\tau \frac{1}{|\Pi|}\, p(c \mid \tau \circ X')$$

$$\sum_\tau p(c \mid \tau \circ X') \leq \frac{|\Pi|}{|C|}(1+\rho)$$

$$= \sum_c p(c \mid X'')\, \frac{1}{|\Pi|}\, \frac{|\Pi|}{|C|}(1+\rho) = \frac{(1+\rho)}{|C|} \sum_c p(c \mid X'') = \frac{(1+\rho)}{|C|}$$

$\sum_c p(c \mid X'') = 1$

Do the sum and get $\log \kappa(x', x'')$

$$\leq \sum_{1<j\leq M} \mathbb{E}_{x',x''}\left[\frac{\mathbb{E}_{\tau_j}[\kappa(\tau_j \circ x', x'') \mid \tau_1, x', x'']}{\kappa(x', x'')} \right]$$

$$\leq \sum_{1<j\leq M} \mathbb{E}_{x',x''}\left[\frac{1+\rho}{|C|\,\kappa(x', x'')} \right]$$

$$= (1+\rho)(M-1)\,\mathbb{E}_{x',x''}\left[\frac{1}{|C|\,\kappa(x', x'')} \right]$$

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

$$= (1+\rho)\,\mathbb{E}_{x',x''}\left[\exp\left(-\log(|C|\,\kappa(x', x'')) + \log M \right) \right]$$

Apply AEP

$$= (1+\rho)\, \mathbb{E}_{x',x''}\left[\exp\left(-\log\left(|\mathcal{C}|K(x',x'')\right) + \log M\right)\right]$$

$$= (1+\rho)\, \mathbb{E}_{x',x''}\left[\exp\left(-\log|\mathcal{C}|\left(\underbrace{\frac{1}{\log|\mathcal{C}|}\log\left(|\mathcal{C}|K(x',x'')\right)}_{\text{emp. log. PA}} - \frac{\log M}{\log|\mathcal{C}|}\right)\right)\right]$$

$$= (1+\rho)\, \mathbb{E}_{x',x''}\left[\exp\left(-\log|\mathcal{C}|\left[\begin{array}{l}\text{emp. log. PA} - \text{exp. log. PA} \\ + \text{exp. log. PA} - \dfrac{\log M}{\log|\mathcal{C}|}\end{array}\right]\right)\right]$$

$$= (1+\rho)\,\mathbb{E}_{X',X''}\left[\exp\left[-\log|\mathcal{C}|\left[\begin{array}{l}\text{emp. log. PA} - \text{exp. log. PA} \\ + \text{exp. log. PA} - \dfrac{\log M}{\log|\mathcal{C}|}\end{array}\right]\right]\right]$$

Asymptotic equipartition property:

$$\text{emp. log. PA} \xrightarrow[n\to\infty]{\text{in prob.}} \text{exp. log. PA}$$

$$-\log|\mathcal{C}|\,\text{emp. log PA} + \log|\mathcal{C}|\,\text{exp. log. PA}$$
$$\leq \log|\mathcal{C}|\,|\text{emp log PA} - \text{exp. log. PA}| \leq \log|\mathcal{C}|\,\varepsilon$$

with prob. $1-\varepsilon$ for large n.

$$\leq (1+\rho)\,\mathbb{E}_{X',X''}\left[\exp\left(-\log|\mathcal{C}|\left(-\varepsilon + \text{exp. log. PA} - \frac{\log M}{\log|\mathcal{C}|}\right)\right)\right]$$

$$= (1+\rho)\exp\left(-\log|\mathcal{C}|\left(\text{exp. log. PA} - \frac{\log M}{\log|\mathcal{C}|} - \varepsilon\right)\right).$$

with prob. $1-\varepsilon$ for large n.

$$\Pr(\mathcal{E}) \leq \text{const} \cdot \exp\left(-\log|\mathcal{C}|\left(\exp.\log.\mathrm{PA} - \frac{\log M}{\log|\mathcal{C}|} - \varepsilon\right)\right)$$

if $\exp.\log.\mathrm{PA} > \frac{\log M}{\log|\mathcal{C}|}$, then

$\Pr(\mathcal{E}) \xrightarrow[n\to\infty]{\text{in prob.}} 0$

codeword length — channel capacity — code rate

$$\Pr(\mathcal{E}) \leq 2^{-n\left(\text{capacity} - \frac{\log \#\text{msgs}}{n} - 3\varepsilon\right)}$$

if $\text{cap} > \frac{\log 2^{nr}}{n} = r$, then

$\Pr(\mathcal{E}) \xrightarrow[n\to\infty]{} 0.$

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

if $\text{exp.log.PA} > \frac{\log M}{\log|\mathcal{C}|}$, then

$$\Pr(\mathcal{E}) \xrightarrow[n\to\infty]{\text{in prob.}} 0$$

$\Rightarrow$ Large exp.log.PA

$\Rightarrow$ Large M

$\Rightarrow$ more messages can be communicated

$\Rightarrow$ the algorithm is more informative

Asymptotic equipartition property

emp. log. PA $\xrightarrow[n \to \infty]{\text{in prob.}}$ exp. log. PA. ?

→ Solutions can be decomposed into "components"

* cluster assig fun → assig for each point.
* graph algos → edge
  e.g. MST
  shortest paths
  clique

→ Posteriors can be factorized; one factor per component

possibly via MFA

Asymptotic equipartition property

→ Posteriors can be factorized; one factor per component

$O(\log |C|)$ components

→ $$\kappa(x', x'') = \sum_c p(c|x') p(c|x'') \approx \sum_c \prod_i p_i(c_i|x') p_i(c_i|x'')$$

see lecture 4 (DA) →

$$= \prod_i \sum_g p_i(g|x') p_i(g|x'')$$

$$= \prod_i \kappa_i(x', x'')$$

Asymptotic equipartition property

$$\text{emp. log. PA} = \frac{1}{\log|C|} \log(|C| K(x', x''))$$

$$= 1 + \frac{1}{\log|C|} \log K(x', x'')$$

$$\approx 1 + \frac{1}{\log|C|} \log \prod_i K_i(x', x'')$$

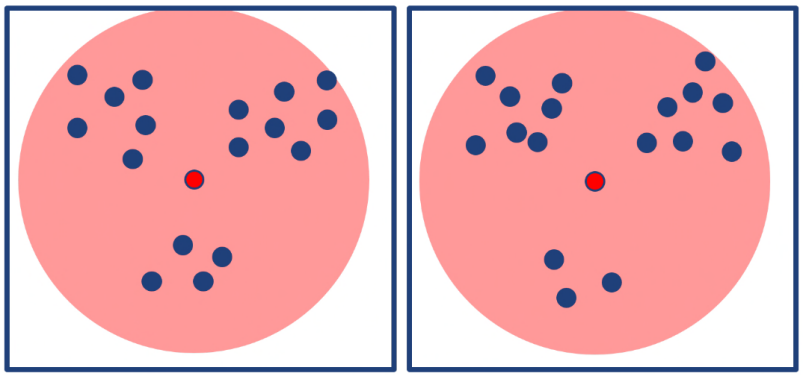
$$= 1 + \frac{1}{\log|C|} \sum_i \log K_i(x', x'')$$

components $= O(\log|C|)$

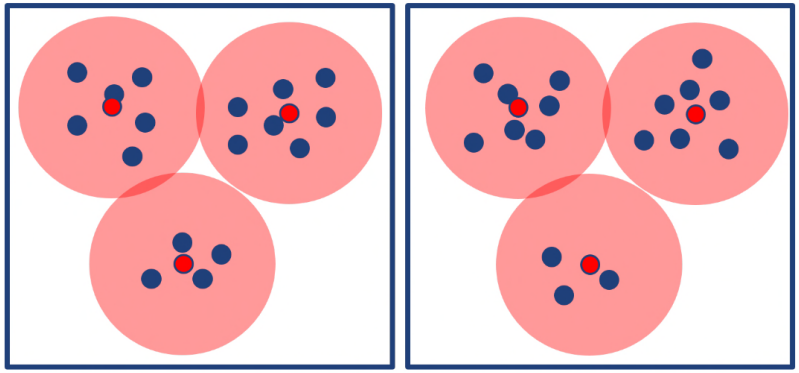
$$\xrightarrow[n \to \infty]{\text{in prob.}} \text{exp. log. PA}.$$

Summary

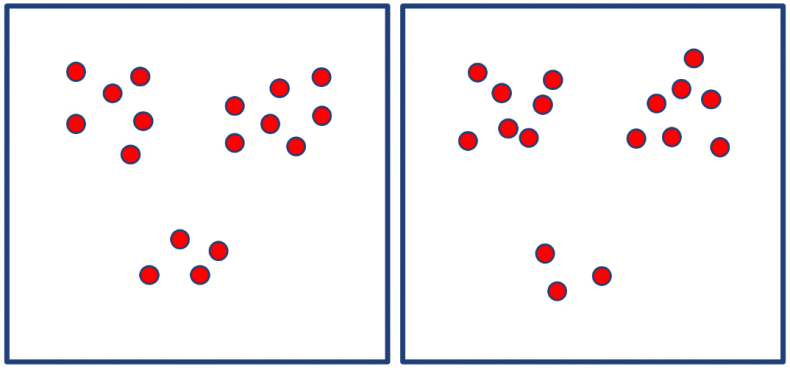
Algorithms must be robust and retrieve signals



robust — no signal

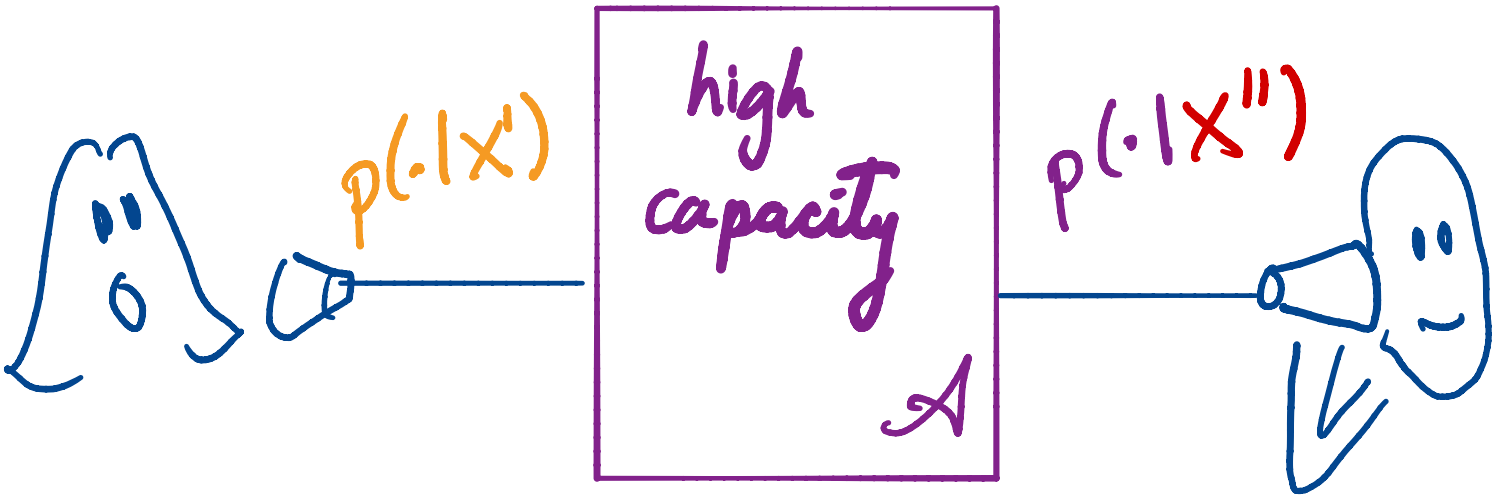


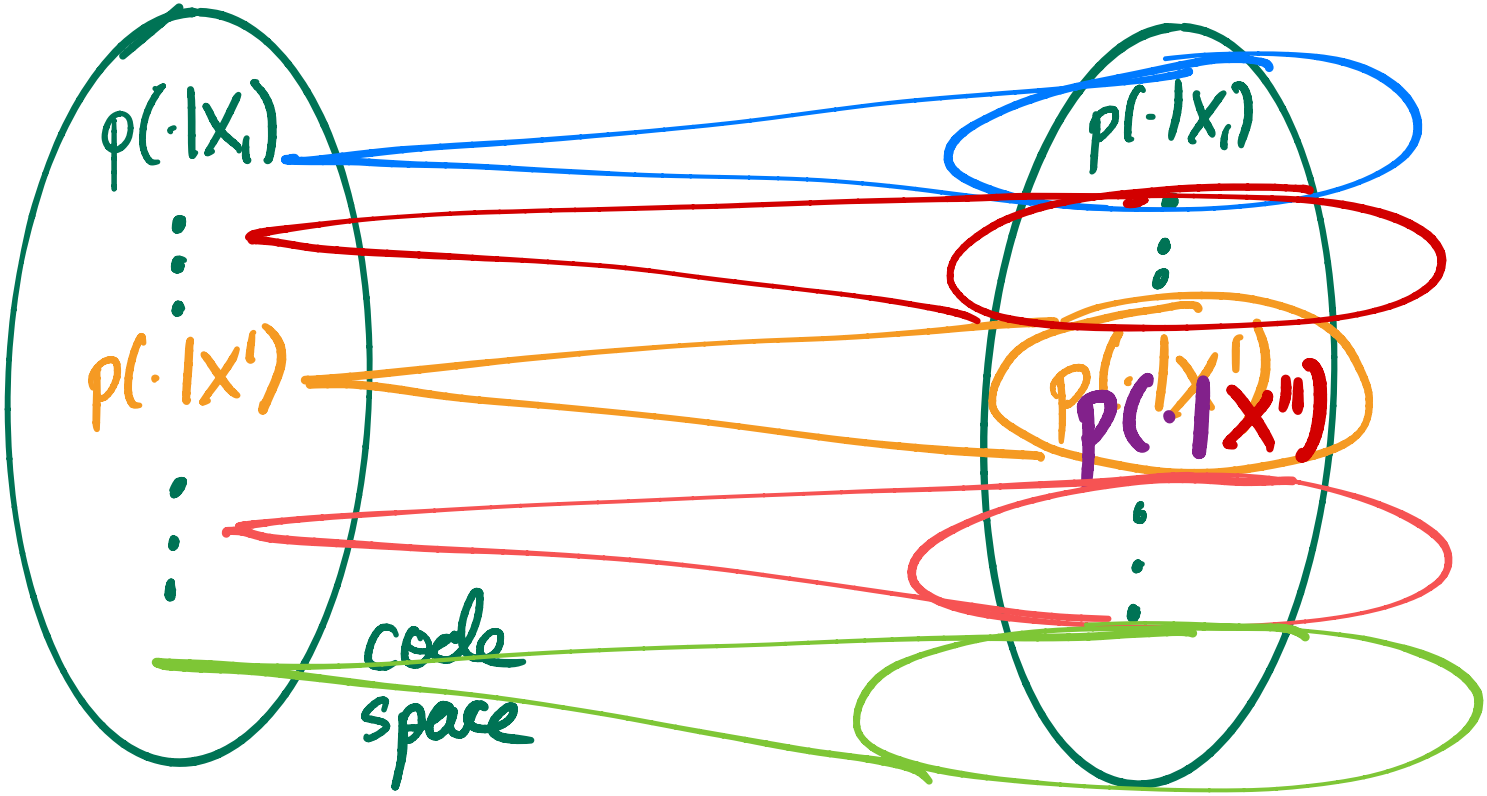
robust — signal

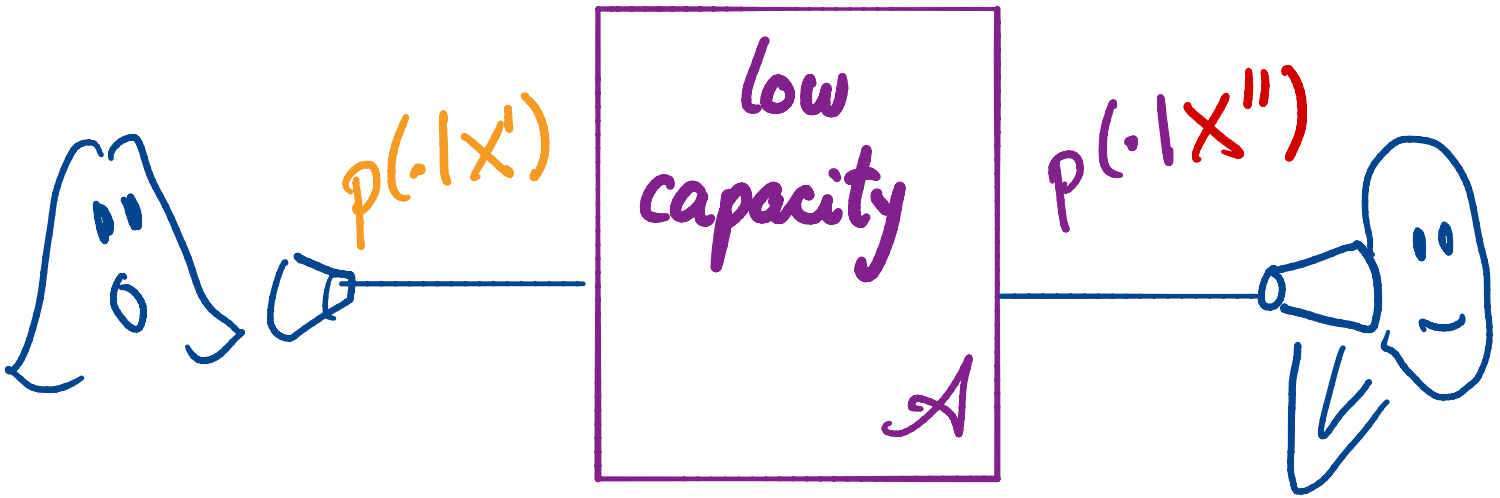


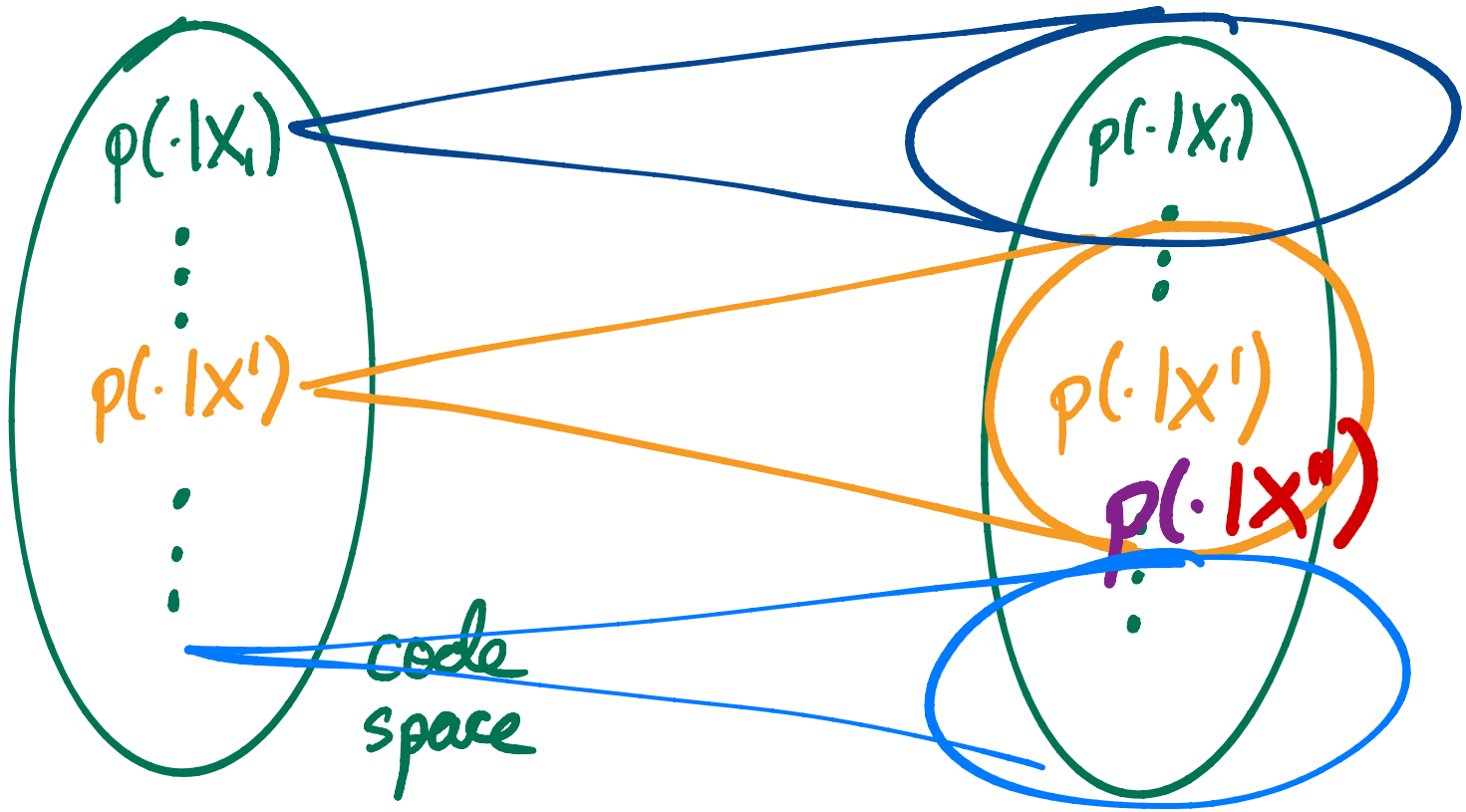
not robust — signal

Information extraction	Information transmission
* Outputs have content * Algorithm is robust	* Codewords have content * Channel is robust
Adjust your algos to increase exp. log. PA	Adjust your channels to increase capacity


p(·|x')
high capacity
A
p(·|x'')


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


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


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

$$\Pr(\mathcal{E}) \leq \text{const} \cdot \exp\left(-\log|\mathcal{C}|\left(\exp.\log.\mathrm{PA} - \frac{\log m}{\log|\mathcal{C}|} - \varepsilon\right)\right)$$

if $\exp.\log.\mathrm{PA} > \frac{\log m}{\log|\mathcal{C}|}$, then

$\Pr(\mathcal{E}) \xrightarrow[n\to\infty]{\text{in prob.}} 0$

codeword length — channel capacity — code rate

$$\Pr(\mathcal{E}) \leq 2^{-n\left(\text{capacity} - \frac{\log \#\text{msgs}}{n} - 3\varepsilon\right)}$$

if $\text{cap} > \frac{\log 2^{nr}}{n} = r$, then

$\Pr(\mathcal{E}) \xrightarrow[n\to\infty]{} 0.$

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 $\kappa(x',x'')$.