Finite-blocklength schemes in information theory

Li Cheuk Ting

Department of Information Engineering,
The Chinese University of Hong Kong

ctli@ie.cuhk.edu.hk

Part of this presentation is based on my lecture notes for Special Topics in Information Theory

Overview

- In this talk, we study an unconventional approach to code construction
  - An alternative to conventional random coding
  - Gives tight one-shot/finite-blocklength/asymptotic results
  - Very simple (proof of Marton's inner bound for broadcast channel can be written in one slide!)
- Apply to channel coding, channel with state, broadcast channel, multiple access channel, lossy source coding (with side information), etc

How to measure information?

How to measure information?

- How many bits are needed to store a piece of information?
  - E.g. We can use one bit to represent whether it will rain tomorrow
  - In general, to represent k possibilities, need $\lceil \log_2 k \rceil$ bits
- How much information does "it will rain tomorrow" really contain?
  - For a place that always rains, this contains no information
  - The less likely it will rain, the more information ("surprisal") it contains


It isn't going to rain tomorrow!
Huh.
A
B


It's going to rain tomorrow!
Hurray!
A
B

Self-information

- For probability mass function p_X of random variable X, the **self-information** of the value x is

$$\iota_X(x) = \log \frac{1}{p_X(x)}$$

- We use $\log$ to base 2 (unit is bit)
- For joint pmf $p_{X,Y}$ of a random variables X, Y,

$$\iota_{X,Y}(x, y) = \log \frac{1}{p_{X,Y}(x, y)}$$

Self-information

- E.g. in English text, the most frequent letter is "e" (13%), and the least frequent letter is "z" (0.074%)
  (according to https://en.wikipedia.org/wiki/Letter_frequency)
- Let $X \in \{\mathrm{a}, \ldots, \mathrm{z}\}$ be a random letter
- Have

$$\iota_X(\mathrm{e}) = \log \frac{1}{0.13} \approx 2.94 \text{ bits}$$

$$\iota_X(\mathrm{z}) = \log \frac{1}{0.00074} \approx 10.40 \text{ bits}$$

Self-information - Properties

- $\iota_X(x) \geq 0$
- If p_X is the uniform distribution over $[1..k]$, $\iota_X(x) = \log k$ for $x \in [1..k]$
- (**Invariant under relabeling**) If f is an injective function, then $\iota_{f(X)}(f(x)) = \iota_X(x)$
- (**Additive**) If X, Y are independent,

$$\iota_{X,Y}(x, y) = \iota_X(x) + \iota_Y(y)$$

Information spectrum

- If X is a random variable, $\iota_X(X)$ is random as well
  - Some values of X may contain more information than others
- The distribution of $\iota_X(X)$ (or its cumulative distribution function) is called the **information spectrum**
- $\iota_X(X)$ is a constant if and only if X follows a uniform distribution
- Information spectrum is a probability distribution, which can be unwieldy
  - We sometimes want a single number to summarize the amount of information of X

Entropy

- The **Shannon entropy**

$$H(X) = H(p_X) = \mathbf{E}[\iota_X(X)] = \sum_x p_X(x) \log \frac{1}{p_X(x)}$$

  is the average of the self-information
  - A number (not random) that roughly corresponds to the amount of information in X
  - Treat $0 \log(1/0) = 0$
- Similarly the **joint entropy** of X and Y is

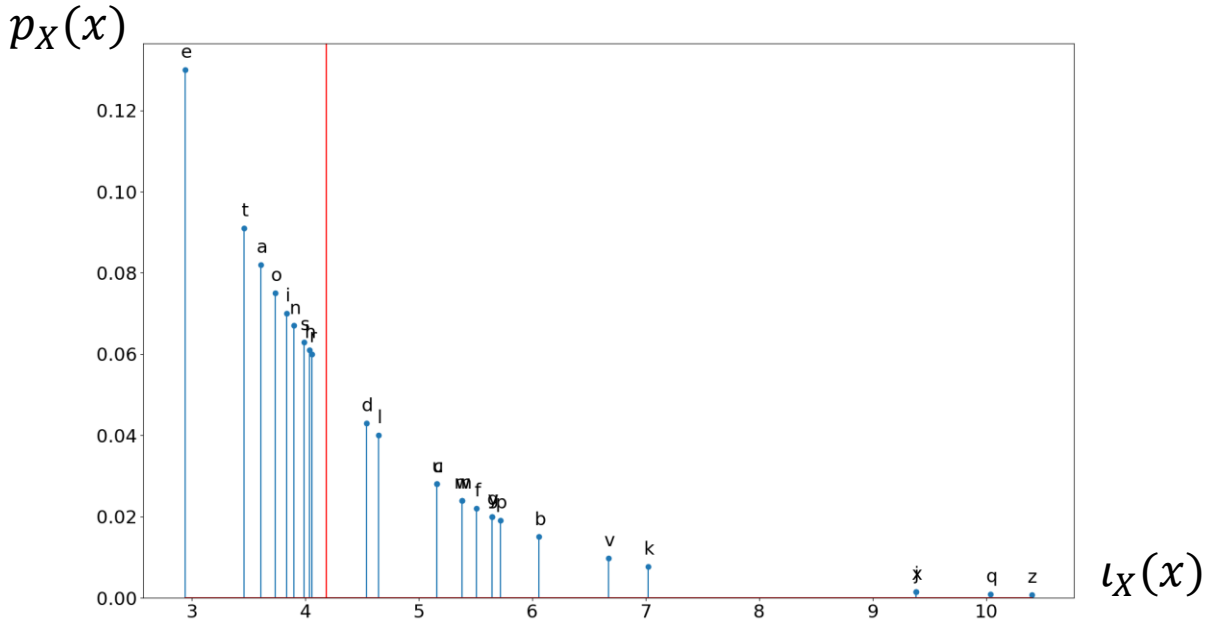
$$H(X, Y) = \mathbf{E}\left[\iota_{X,Y}(X, Y)\right]$$

Entropy - Properties

- $H(X) \geq 0$, and $H(X) = 0$ iff X is (almost surely) a constant
- If $X \in [1..k]$, then $H(X) \leq \log k$
  - Equality iff X is uniform over $[1..k]$
  - Proof: Jensen's ineq. on concave function $z \mapsto z \log(1/z)$
- If f is a function, then $H(f(X)) \leq H(X)$
  - If f is injective, equality holds (invariant under relabeling)
  - Consequences: $H(X,Y) \geq H(X)$, $H\big(X, f(X)\big) = H(X)$
- (**Subadditive**) $H(X,Y) \leq H(X) + H(Y)$
  - Equality holds iff X, Y independent (**additive**)
- $H(X)$ is concave in p_X

A random English letter

(according to https://en.wikipedia.org/wiki/Letter_frequency)



- Self-information ranges from $\iota_X(\mathrm{e}) \approx 2.94$ to $\iota_X(\mathrm{z}) \approx 10.40$
- $H(X) \approx 4.18$

Why is entropy a reasonable measure of information?

- Axiomatic characterization:
  $H(X)$ is the only measure that satisfies
  - **Subadditivity.** $H(X,Y) \leq H(X) + H(Y)$
  - **Additivity**. $H(X,Y) = H(X) + H(Y)$ if X, Y independent
  - Invariant under relabeling and adding a zero mass
  - $H(X)$ is continuous in p_X
  - $H(X) = 1$ when $X \sim \text{Unif}\{0,1\}$

  [Aczél, J., Forte, B., & Ng, C. T. (1974). Why the Shannon and Hartley entropies are 'natural']
- Operational characterizations:
  - $H(X)$ is approximately the number of coin flips needed to generate X [D. E. Knuth & A. C. Yao. (1976). The complexity of nonuniform random number generation]
  - $H(X)$ is approximately the number of bits needed to compress X

Information density

- The **information density** between two random variables X, Y is

$$\iota_{X;Y}(x;y) = \iota_Y(y) - \iota_{Y|X}(y|x)$$
$$= \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} = \log \frac{p_{Y|X}(y|x)}{p_Y(y)}$$

  - $\iota_Y(y)$ is the info of $Y = y$ without knowing $X = x$
  - $\iota_{Y|X}(y|x)$ is the info of $Y = y$ after knowing $X = x$
  - $\iota_{X;Y}(x;y)$ measures how much knowing $X = x$ reduces the info of $Y = y$
  - Can be positive/negative/zero
  - Zero if X, Y independent

Information density

- $\iota_{X;Y}(x;y) = \iota_Y(y) - \iota_{Y|X}(y|x)$
  $= \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} = \log \frac{p_{Y|X}(y|x)}{p_Y(y)}$
- E.g. X, Y are the indicators of whether it rains today/tomorrow resp., with the following prob. matrix

	$Y = 0$	$Y = 1$
$X = 0$	0.6	0.1
$X = 1$	0.1	0.2

- $\iota_{X;Y}(1;1) = \log \frac{0.2}{0.3 \cdot 0.3} \approx 1.15$
  - Knowing it rains today decreases the info of "tomorrow will rain"
- $\iota_{X;Y}(1;0) = \log \frac{0.1}{0.3 \cdot 0.7} \approx -1.07$
  - Knowing it rains today increases the info of "tomorrow will not rain"

Mutual information

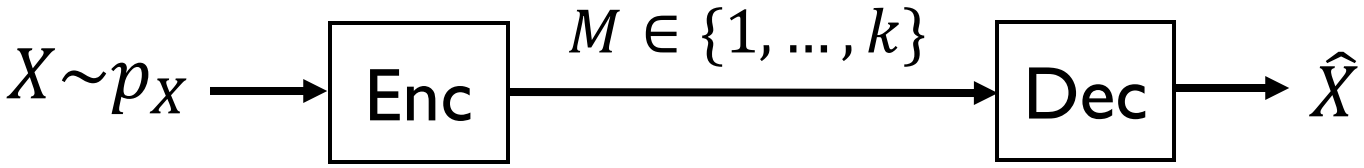
- The **mutual information** between two random variables X, Y is

$$
\begin{aligned}
I(X;Y) &= \mathbf{E}\left[\iota_{X;Y}(X;Y)\right] \\
&= \mathbf{E}\left[\log \frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right] \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned}
$$

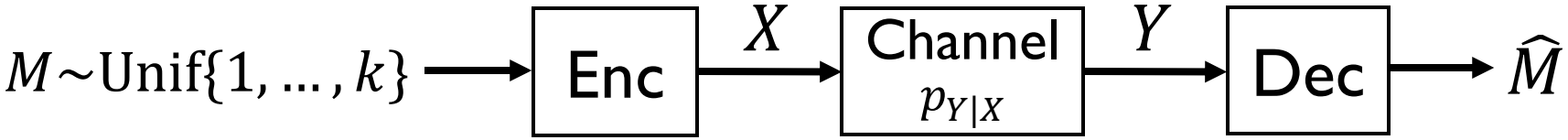
  - Always nonnegative since $H(Y) \geq H(Y|X)$
  - Measures the dependency between X, Y
    - Zero iff X, Y independent

Source coding & channel coding

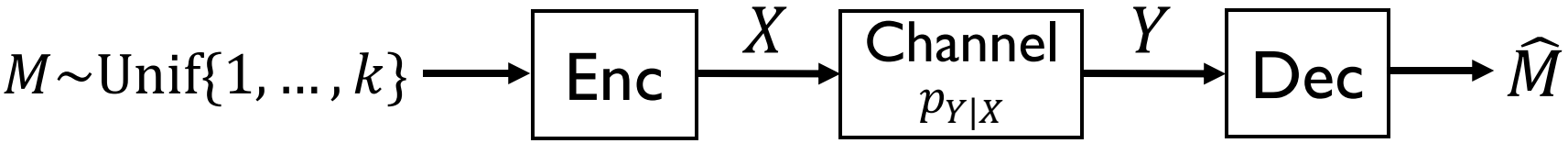
- Source coding: compressing a source $X \sim p_X$



- Channel coding: transmitting a message M through a noisy channel

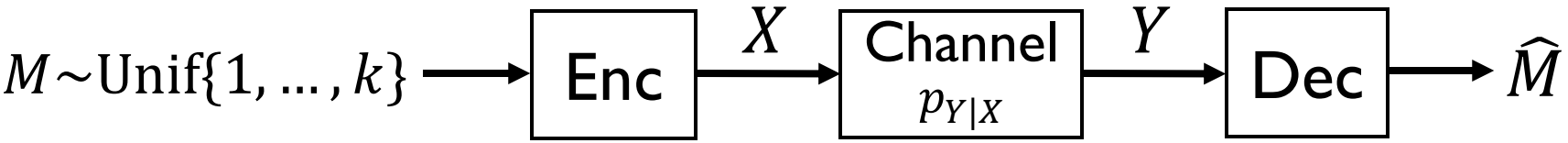


One-shot channel coding



- Message $M \sim \text{Unif}\{1, \ldots, k\}$
- Encoder maps message to channel input $X = f(M)$
  - The set $\mathcal{C} = \{f(m) : m \in \{1, \ldots, k\}\}$ is the **codebook**
  - Its elements $f(m)$ are called **codewords**
- Channel output Y follows conditional distribution $p_{Y|X}$
- Decoder maps Y to decoded message $\widehat{M} = g(Y)$
- Goal: error prob $\mathbf{P}(\widehat{M} \neq M)$ is small

One-shot channel coding



- Want $\mathbf{P}\left(\widehat{M} \neq M\right) \leq \epsilon$

Thm [Yassaee et al. 2013]. Fix any p_X. There exists code with

$$\mathbf{P}\left(\widehat{M} \neq M\right) \leq 1 - \mathbf{E}\left[\frac{1}{1 + k2^{-\imath_{X;Y}(X;Y)}}\right]$$

$$\leq \mathbf{E}\left[\min\{k2^{-\imath_{X;Y}(X;Y)}, 1\}\right]$$

where $(X, Y) \sim p_X p_{Y|X}$

[Yassaee, Aref, and Gohari, "A technique for deriving one-shot achievability results in network information theory," ISIT 2013.]

One-shot channel coding

- Random codebook generation: generate $f(m) \sim p_X$ i.i.d. for $m \in \{1, \ldots, k\}$

Given Y, the decoder:

- (Maximum likelihood decoder) Find $\hat{m}$ that maximizes $p_{Y|X}(Y|f(\hat{m}))$
  - Optimal – attains the lowest error prob. for a fixed f
- (Stochastic likelihood decoder) Chooses $\hat{m}$ with prob.

$$\mathbf{P}(\hat{m}|Y) = \frac{p_{Y|X}(Y|f(\hat{m}))}{\sum_{m'} p_{Y|X}(Y|f(m'))} = \frac{2^{\imath_{X;Y}(f(\hat{m});Y)}}{\sum_{m'} 2^{\imath_{X;Y}(f(m');Y)}}$$

[Yassaee-Aref-Gohari 2013]

- $\mathbf{P}(\widehat{m}|Y) = \frac{2^{\iota X;Y(f(\widehat{m});Y)}}{\sum_{m'} 2^{\iota X;Y(f(m');Y)}}$ [Yassaee-Aref-Gohari 2013]

$$\mathbf{P}\left(M = \widehat{M}\right)$$

$$= \mathbf{E}_{\mathcal{C}}\left[\frac{1}{k}\sum_{m,y} p_{Y|X}(y|f(m))\frac{2^{\iota X;Y(f(m);y)}}{\sum_{m'} 2^{\iota X;Y(f(m');y)}}\right]$$

$$= \mathbf{E}_{\mathcal{C}}\left[\sum_{y} p_{Y|X}(y|f(1))\frac{2^{\iota X;Y(f(1);y)}}{\sum_{m'} 2^{\iota X;Y(f(m');y)}}\right] \quad \text{(Symmetry)}$$

$$= \sum_{y} \mathbf{E}_{f(1)}\mathbf{E}_{f(2),\ldots,f(k)}\left[p_{Y|X}(y|f(1))\frac{2^{\iota X;Y(f(1);y)}}{2^{\iota X;Y(f(1);y)}+\sum_{m'\neq 1} 2^{\iota X;Y(f(m');y)}}\right]$$

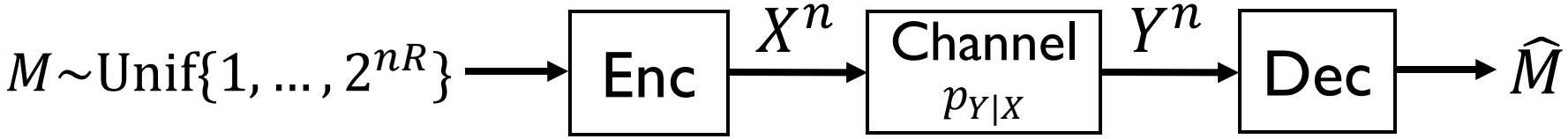
$$\geq \sum_{y} \mathbf{E}_{f(1)}\left[p_{Y|X}(y|f(1))\frac{2^{\iota X;Y(f(1);y)}}{2^{\iota X;Y(f(1);y)}+k-1}\right] \quad \text{(Jensen)}$$

$$\geq \sum_{y} \mathbf{E}_{f(1)}\left[p_{Y|X}(y|f(1))\frac{1}{1+k2^{-\iota X;Y(f(1);y)}}\right]$$

$$= \sum_{y}\sum_{x} p_X(x)\, p_{Y|X}(y|x)\frac{1}{1+k2^{-\iota X;Y(x;y)}}$$

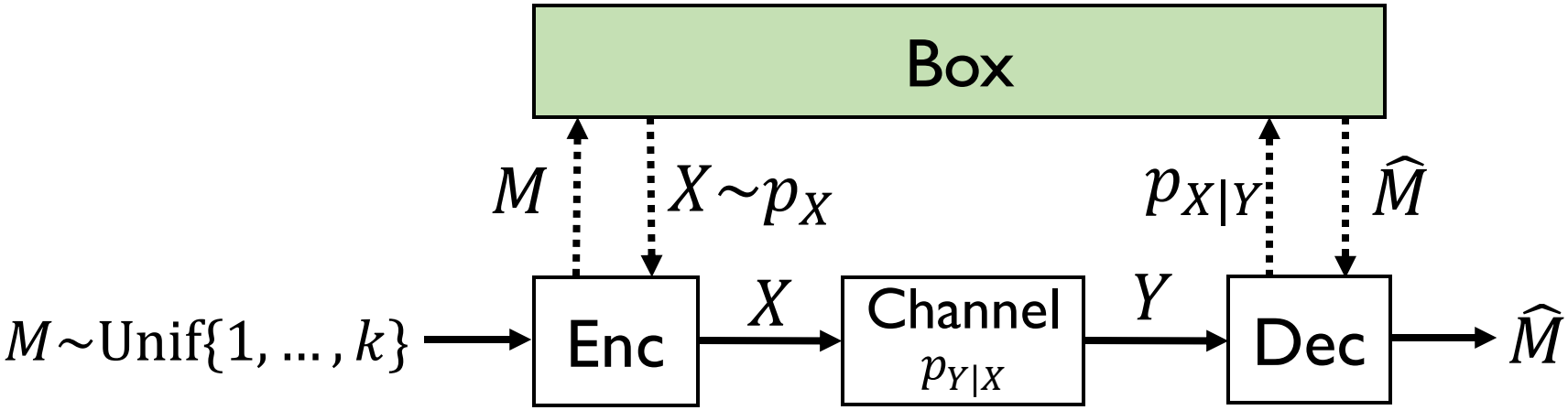
$$= \mathbf{E}\left[\frac{1}{1+k2^{-\iota X;Y(X;Y)}}\right]$$

Asymptotic channel coding



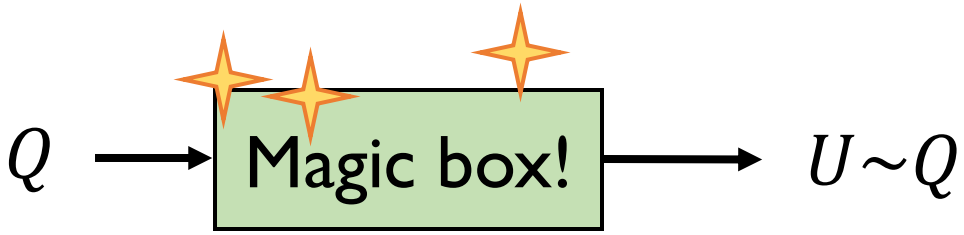
- Memoryless: $p_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^n p_{Y|X}(y_i|x_i)$
- Applying one-shot:
  $P_e = \mathbf{P}\left(\widehat{M} \neq M\right) \leq \mathbf{E}\left[\min\{2^{nR - \sum_{i=1}^n \imath_{X;Y}(X_i;Y_i)}, 1\}\right]$,
  where $(X_i, Y_i) \sim p_X p_{Y|X}$ i.i.d. for $i = 1, \ldots, n$
- Asymptotic ($n \to \infty$): have
  $\sum_{i=1}^n \imath_{X;Y}(X_i; Y_i) \approx nI(X;Y)$ by law of large numbers,
  so $P_e \to 0$ if $R < I(X;Y)$
- Recovers (achievability part of) Shannon's channel coding theorem: Channel capacity is
  $$C = \max_{p_X} I(X;Y)$$

Codebook as a black box



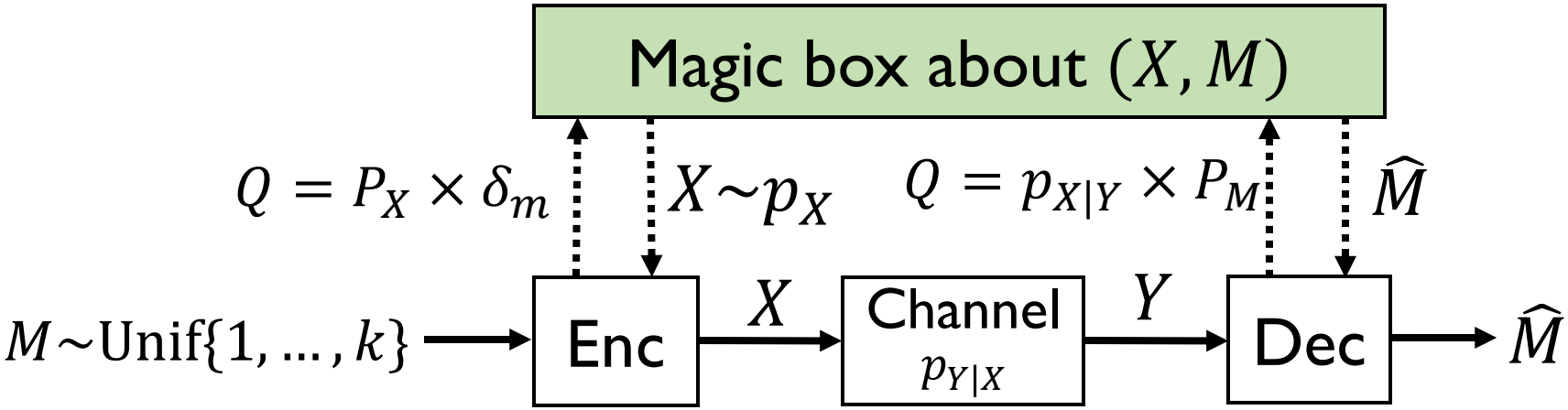
- Random codebook:
  $\mathcal{C} = \{f(m)\} \sim p_X$ i.i.d. for $m \in \{1, \ldots, k\}$
- Decoder: Find $\hat{m} = \operatorname{argmax} p_{X|Y}(f(\hat{m})|Y)/p_X(f(\hat{m}))$
- Treat codebook $\mathcal{C}$ as a box:
  - **Operation 1:** Query M, get $X \sim p_X$
  - **Operation 2:** Query posterior distribution $p_{X|Y}$, get $\widehat{M}$

A general black box



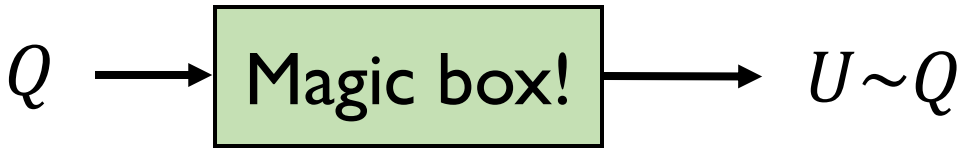
- Consider random variable U
- Only one operation: Query distribution Q, get $U \sim Q$
- Want box to have "memory"
  - If we query the same Q twice, should get the same U
  - If we query similar Q_1, Q_2, then U_1, U_2 are equal with high probability

Using the general black box



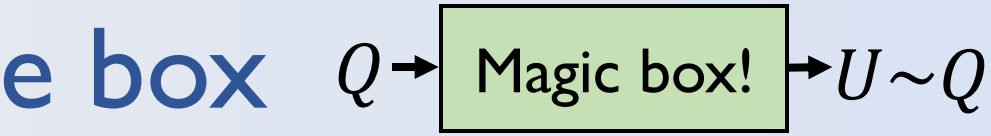
- Let $U = (X, M)$
- Encoding: Query $Q = P_X \times \delta_m$ (δ_m is degenerate distribution $\mathbf{P}(M = m) = 1$), get (X, m)
- Decoding: Query $Q = P_{X|Y} \times P_M$ (P_M is $\mathrm{Unif}\{1, \ldots, k\}$), get $(\hat{X}, \hat{m})$
- Input partial knowledge into box, get full knowledge

How to build the box



- **Operation:** Query distribution Q, get $U \sim Q$
- **Memory:** If we query similar Q_1, Q_2, then U_1, U_2 are equal with high probability
- Attempt 1: Generate $U \sim Q$ afresh for each query?
  - Does not have memory!
- Attempt 2: Generate random seed Z at the beginning, then use the same seed to generate all $U \sim Q$?
  - Only guarantees to give the same U for the same Q
  - No guarantee for similar but different Q_1, Q_2
- Need a way to generate U that is not sensitive to small changes to Q

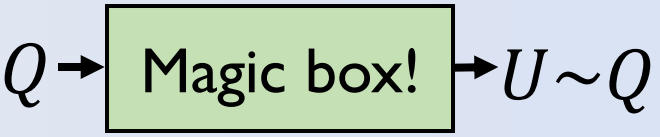
How to build the box



- Generate random seed Z at the beginning, then use the same seed to generate all $U \sim Q$?
- **Exponential distribution with rate λ**
  $\mathrm{Exp}(\lambda)$ has prob. density function
  $$f(z;\lambda) = \lambda e^{-\lambda z} \quad \text{for } z \geq 0$$
  - If $Z \sim \mathrm{Exp}(\lambda)$, then $aZ \sim \mathrm{Exp}(\lambda/a)$
  - For $Z_i \sim \mathrm{Exp}(\lambda_i)$ indep. for $i = 1, \ldots, l$, have
  $$\mathbf{P}(\mathrm{argmin}_i Z_i = j) = \frac{\lambda_j}{\lambda_1 + \cdots + \lambda_l}$$
- Let $Z = (Z_1, \ldots, Z_l)$ be the seed, $Z_u \sim \mathrm{Exp}(1)$ i.i.d.
- Query Q, output $U = \mathrm{argmin}_u \frac{Z_u}{Q(u)}$

C.T. Li and A. El Gamal. Strong functional representation lemma and applications to coding theorems. IEEE Trans. Inf. Theory, 64(11):6967–6978, 2018
C.T. Li and V. Anantharam, "A Unified Framework for One-Shot Achievability via the Poisson Matching Lemma," IEEE Trans. Inf. Theory, vol. 67, no. 5, pp. 2624-2651, 2021.

How to build the box

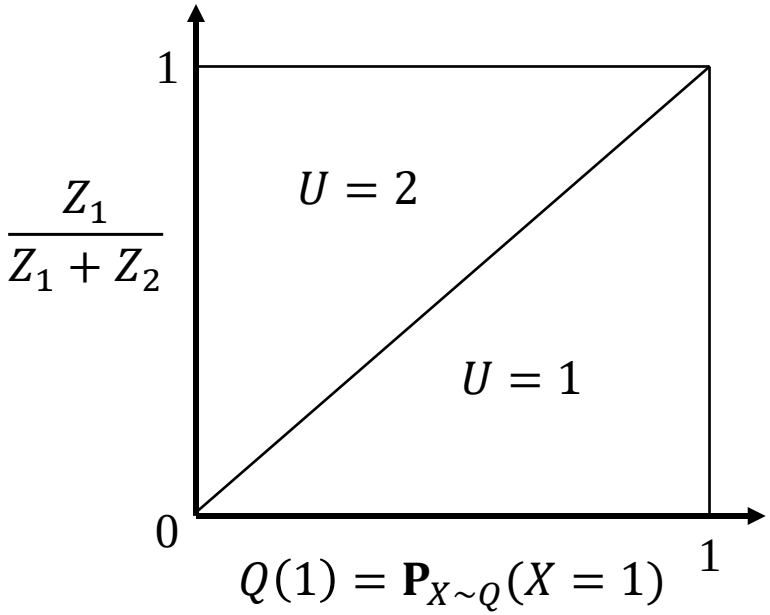


- Let $Z = (Z_1, \ldots, Z_l)$ be the seed, $Z_i \sim \mathrm{Exp}(1)$ i.i.d.
- Query Q, output $U = \mathrm{argmin}_u \frac{Z_u}{Q(u)}$
- $\mathbf{P}(U = u) = \frac{Q(u)}{Q(1)+\cdots+Q(l)} = Q(u)$ OK!
- Give same U for same Q since U is a function of Q and Z (fixed at the beginning) OK!
- Small changes to Q is unlikely to affect $\mathrm{argmin}_u \frac{Z_u}{Q(u)}$ OK!

How to build the box

$Q \rightarrow$ Magic box! $\rightarrow U \sim Q$

- Let $Z = (Z_1, \ldots, Z_l)$ be the seed, $Z_i \sim \mathrm{Exp}(1)$ i.i.d.
- Query Q, output $U = \mathrm{argmin}_u \frac{Z_u}{Q(u)}$
- If $l = 2$, then $U = 1$ iff $\frac{Z_1}{Q(1)} < \frac{Z_2}{Q(2)} \Leftrightarrow \frac{Z_1}{Z_1 + Z_2} < Q(1)$



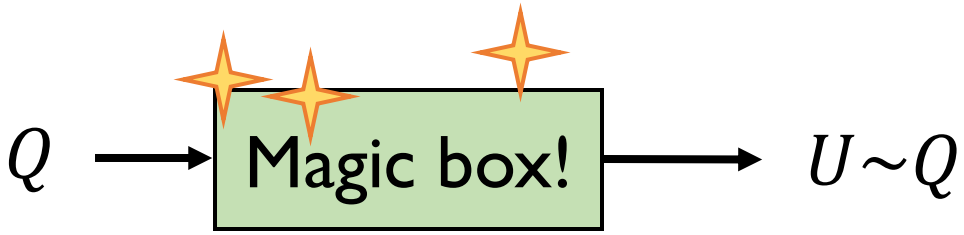
Poisson matching lemma

- Let $Z = (Z_1, \ldots, Z_l)$ be the seed, $Z_i \sim \mathrm{Exp}(1)$ i.i.d.
- Query Q, output $U_Q = \mathrm{argmin}_u \frac{Z_u}{Q(u)}$
- **Poisson matching lemma** [Li-Anantharam 2018]: If we query P, Q to get U_P, U_Q respectively, then

$$\mathbf{P}\left(U_Q \neq U_P \middle| U_P\right) \leq \frac{P(U_P)}{Q(U_P)}$$

C.T. Li and A. El Gamal. Strong functional representation lemma and applications to coding theorems. IEEE Trans. Inf. Theory, 64(11):6967–6978, 2018.
C.T. Li and V. Anantharam, "A Unified Framework for One-Shot Achievability via the Poisson Matching Lemma," IEEE Trans. Inf. Theory, vol. 67, no. 5, pp. 2624-2651, 2021.

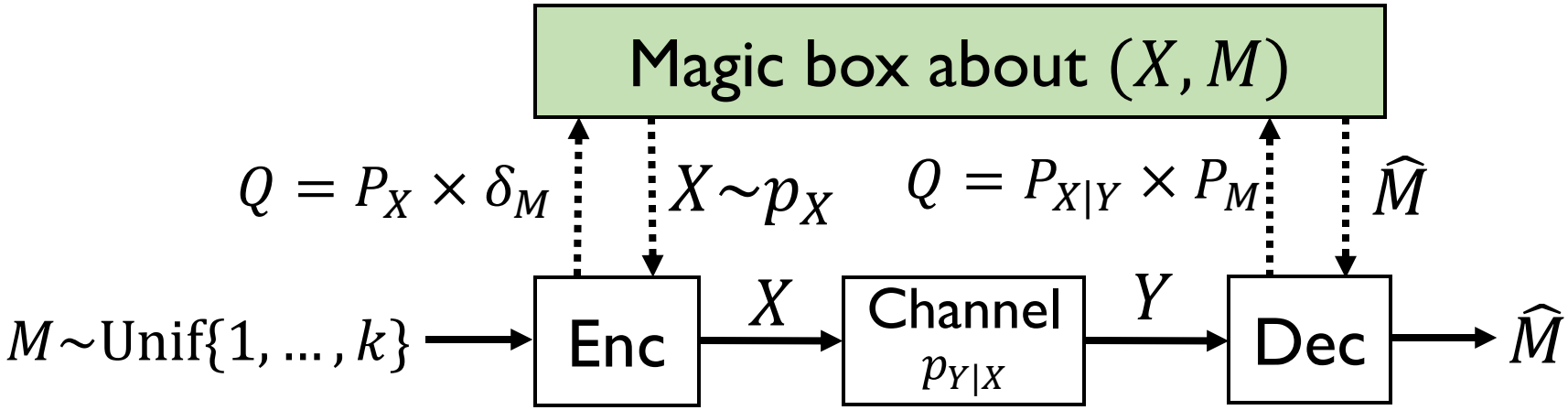
A general black box



- **Operation:** Query distribution Q, get $U \sim Q$
- **Guarantee:** If we query P, Q to get U_P, U_Q respectively, then

$$\mathbf{P}\left(U_Q \neq U_P \middle| U_P\right) \leq \frac{P(U_P)}{Q(U_P)}$$

- We can use this box alone to prove many tight one-shot/finite-blocklength/asymptotic coding results

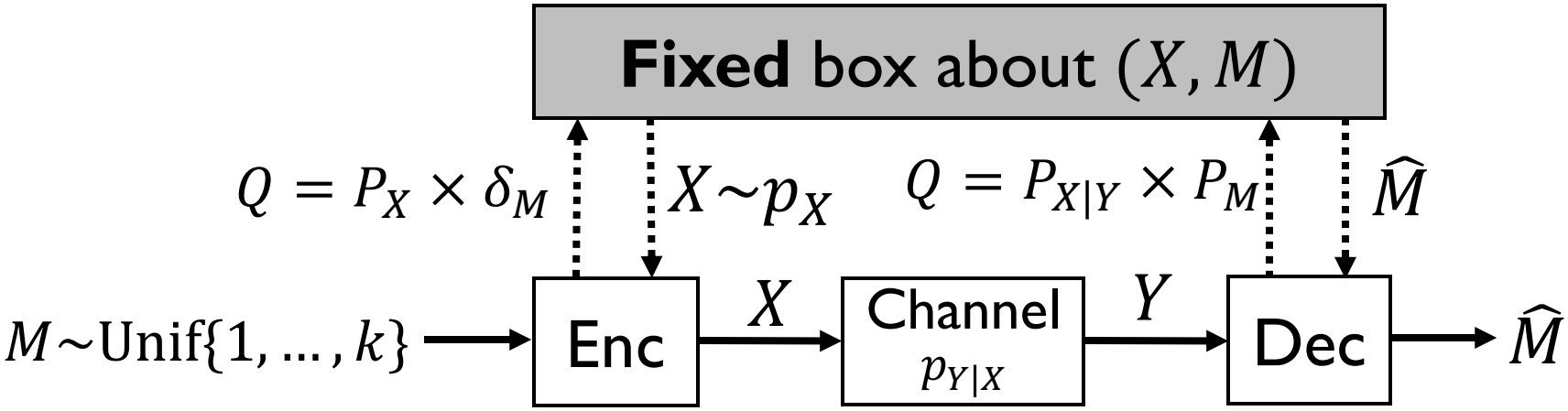


- Let $U = (X, M)$
- Encoding: Query $Q = P_X \times \delta_M$, get (X, M)
- Decoding: Query $Q = P_{X|Y} \times P_M$, get $(\hat{X}, \widehat{M})$
- Poisson matching lemma:

$$\mathbf{P}\left(M \neq \widehat{M}\right) \leq \mathbf{E}\left[\mathbf{P}\left(M \neq \widehat{M} \middle| M, X, Y\right)\right]$$
$$\leq \mathbf{E}\left[\min\left\{\frac{(P_X \times \delta_M)(X,M)}{(P_{X|Y} \times P_M)(X,M)}, 1\right\}\right]$$
$$= \mathbf{E}\left[\min\left\{\frac{P_X(X)}{P_{X|Y}(X|Y)/k}, 1\right\}\right]$$
$$= \mathbf{E}\left[\min\left\{k 2^{-\iota_{X;Y}(X;Y)}, 1\right\}\right]$$

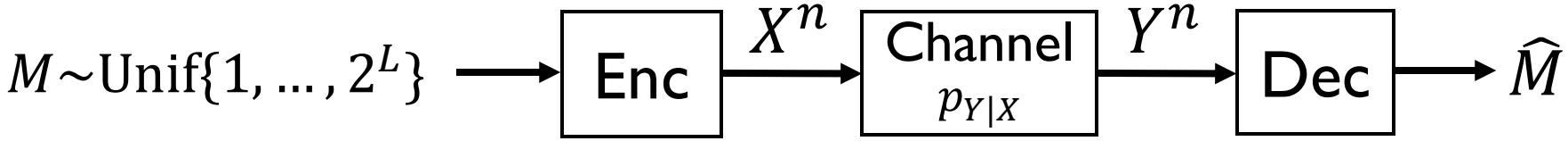
C. T. Li and V. Anantharam, "A Unified Framework for One-Shot Achievability via the Poisson Matching Lemma," IEEE Trans. Inf. Theory, vol. 67, no. 5, pp. 2624-2651, 2021.

Channel coding – removing the box



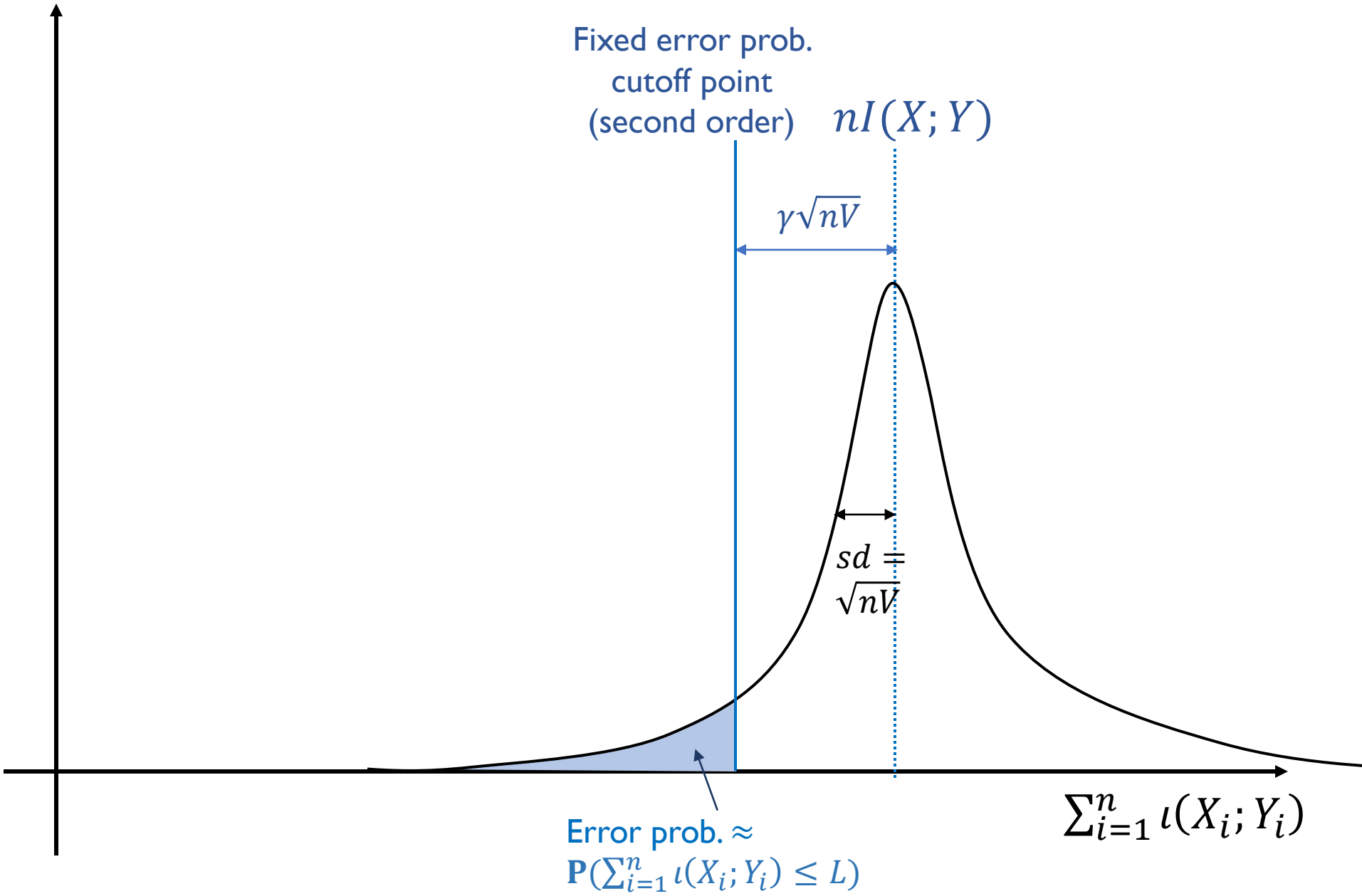
- The box contains a random seed in it
- In reality, encoder and decoder cannot share common randomness
- $P_e \leq \mathbf{E}\left[\min\left\{k2^{-\iota_{X;Y}(X;Y)}, 1\right\}\right]$ averaged over choices of seed
- There exists fixed seed s.t. $P_e \leq \mathbf{E}\left[\min\left\{k2^{-\iota_{X;Y}(X;Y)}, 1\right\}\right]$

Second-order asymptotics



- $P_e \leq \mathbf{E}\left[\min\left\{2^{L-\sum_{i=1}^n \iota_{X;Y}(X_i;Y_i)}, 1\right\}\right], (X_i, Y_i) \sim p_X p_{Y|X}$ i.i.d.
- $P_e \approx 0$ if $L \ll \sum_{i=1}^n \iota(X_i; Y_i)$, $P_e \approx 1$ if $L \gg \sum_{i=1}^n \iota(X_i; Y_i)$
- First-order: optimal $L \approx nI(X;Y)$
- Central limit theorem:
  $\sum_{i=1}^n \iota(X_i; Y_i)$ approximately follows $N(nI(X;Y), nV)$, where
  $$V = \text{Var}[\iota(X;Y)]$$
- For a fixed $P_e = \epsilon$, optimal $L \approx nI(X;Y) - \sqrt{nV}Q^{-1}(\epsilon)$
  where $Q^{-1}(\epsilon)$ is the inverse of the Q-function
  ($Q(\gamma) = 1 - \Phi(\gamma)$, Φ is the cdf of $N(0,1)$)
- The V when p_X is the capacity-achieving distribution (that maximizes $I(X;Y)$) is called the **channel dispersion**

Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 2307–2359, 2010.


Fixed error prob.
cutoff point
(second order)
$nI(X;Y)$
$\gamma\sqrt{nV}$
$sd = \sqrt{nV}$
$\sum_{i=1}^{n} \iota(X_i;Y_i)$
Error prob. ≈
$\mathbf{P}(\sum_{i=1}^{n} \iota(X_i;Y_i) \leq L)$