Distribution Approximation Techniques for Security, Differential Privacy, and Learning

Paul Cuff (Princeton University)

Information Theory

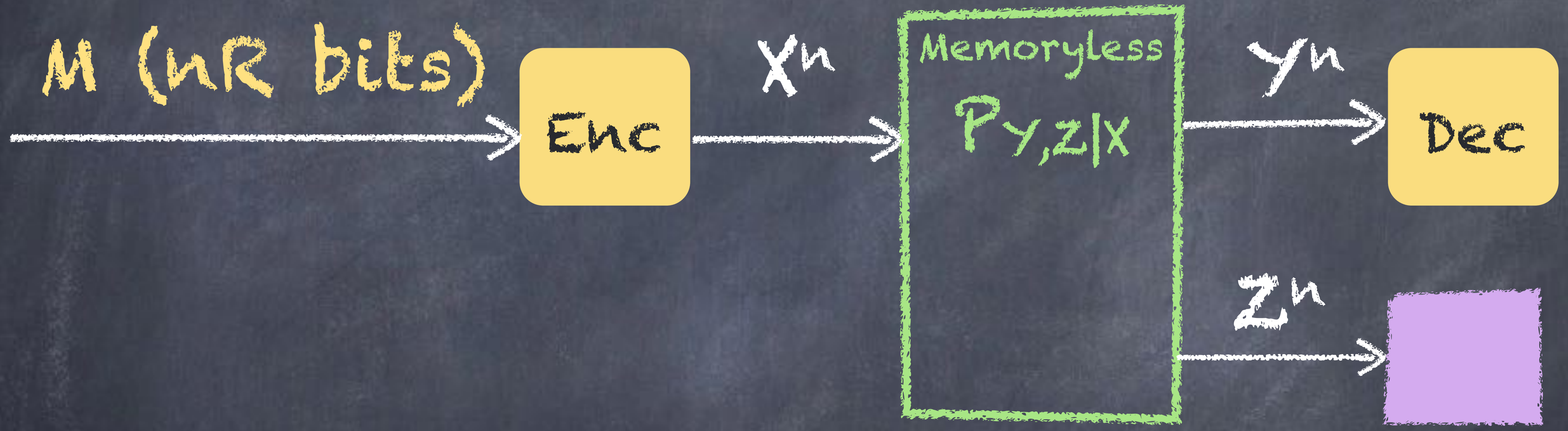
New results on secure communication

Wiretap Example

- Transmit n bits
- Eavesdropper sees all but one bit

0 1 1 0 1 0 0 $\blacksquare$ 1 0 1 1

Wiretap Channel



Secrecy Capacity:

- Reliable communication
- Z^n contains no information about M

Solutions

- 1975: Wyner introduced the problem and gave solution for degraded channels

- 1978: Csiszár and Körner gave solution for all channels

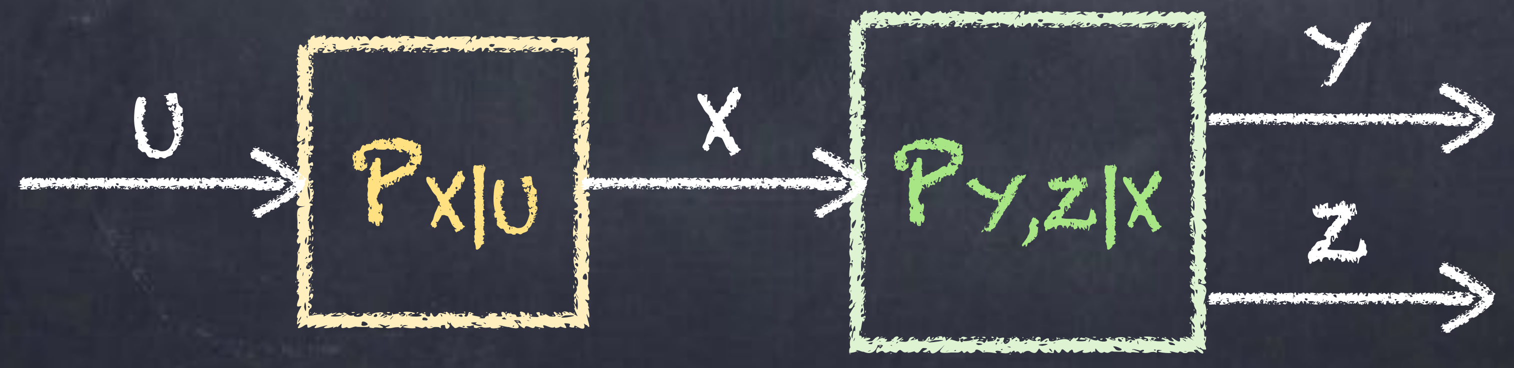
Solution

Degraded:

$$C_s = \max_{P_X} I(X;Y) - I(X;Z)$$

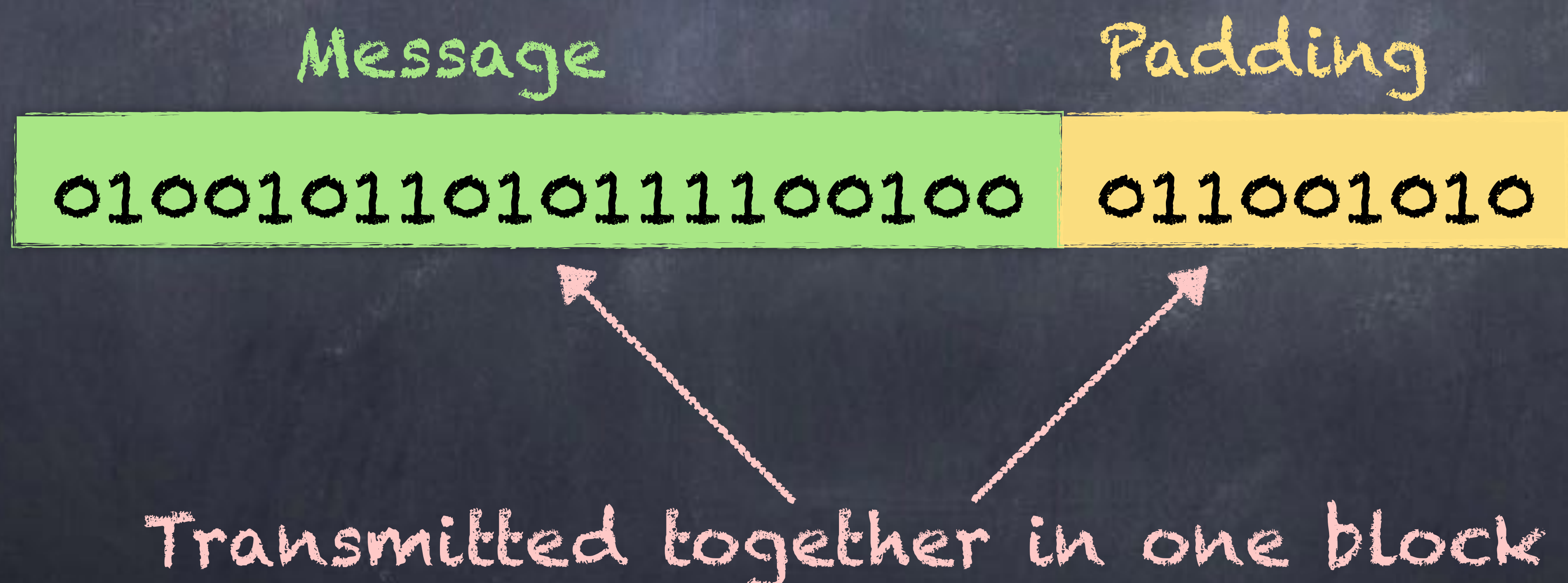
General:

$$C_s = \max_{P_{XU}} I(U;Y) - I(U;Z)$$

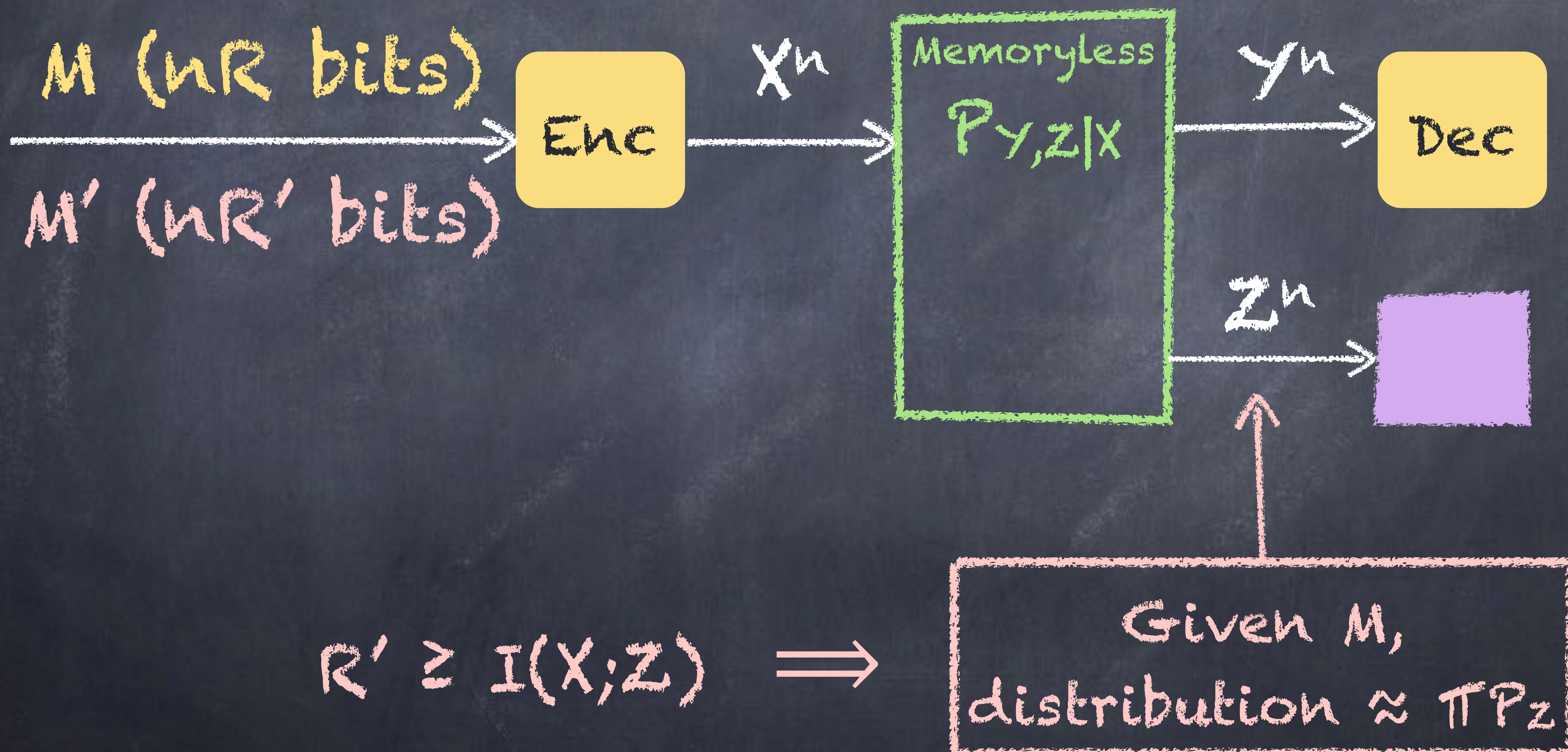


Encoding

- Random Codebook
- Pad with random garbage bits



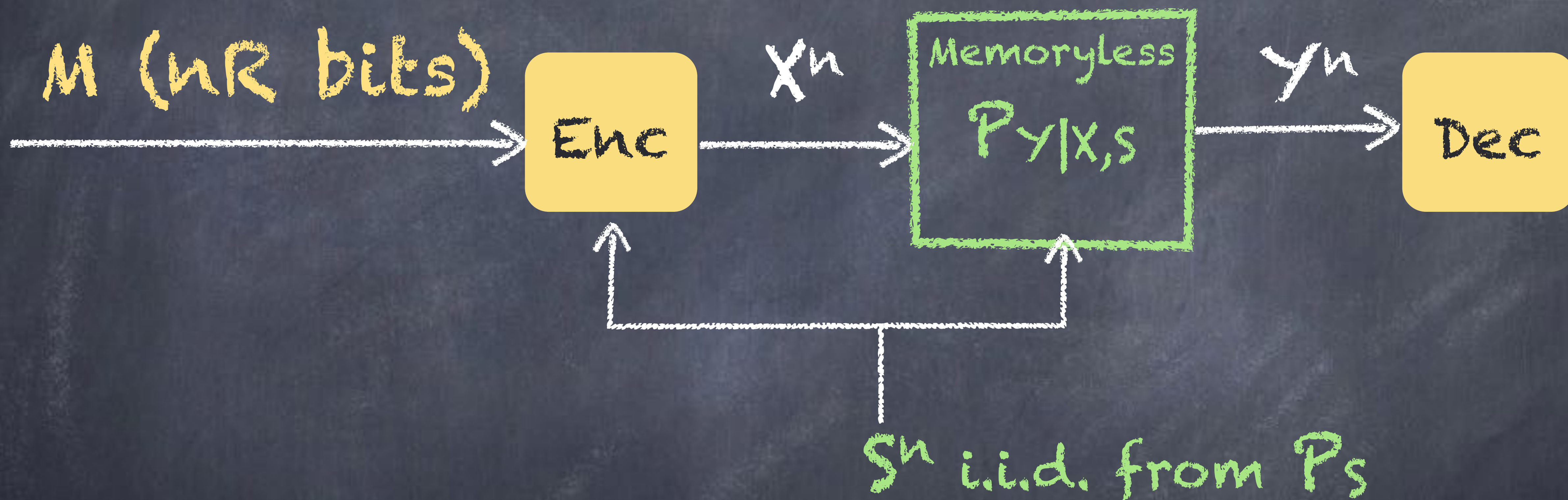
Encoding Concept



Channel Capacity with Random State

Puzzle

Gelfand-Pinsker
(state known to encoder)

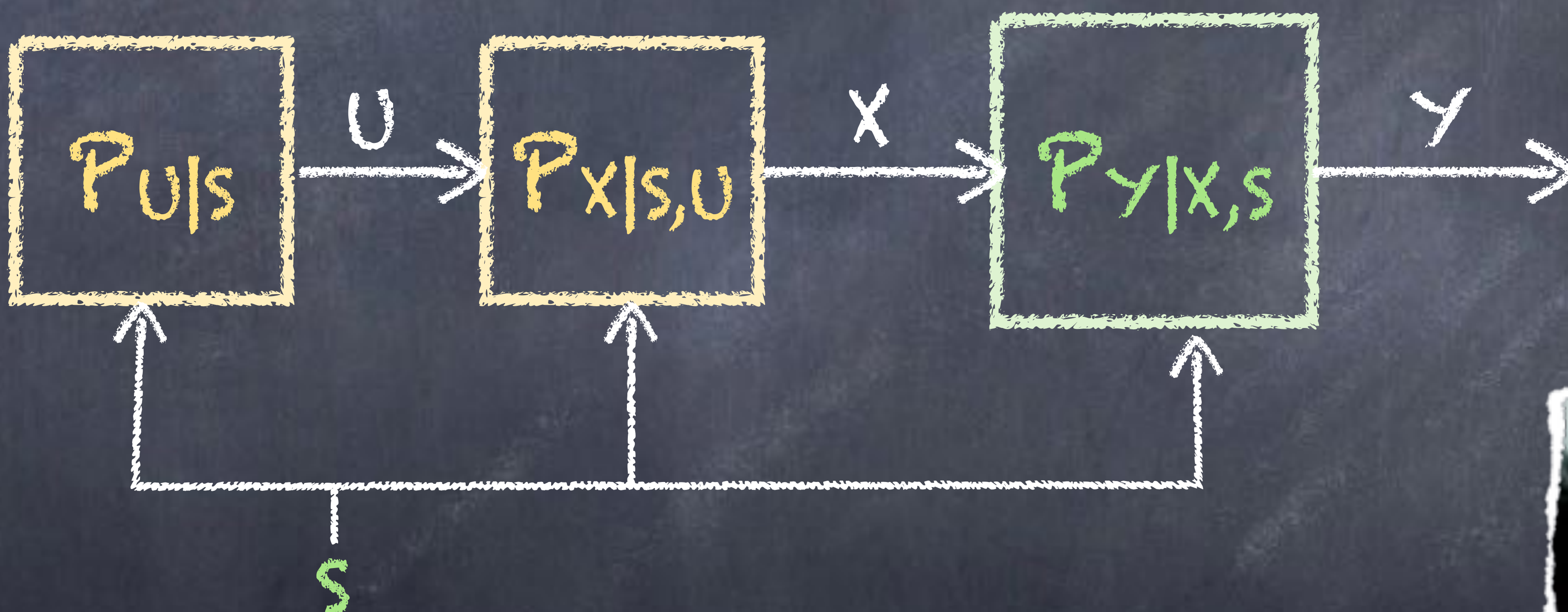


<u>Capacity:</u>

- Reliable communication

Solution (1980)
Gelfand-Pinsker

$$C = \max_{P_{X,U|S}} I(U;Y) - I(U;S)$$



Encoding

- Random Codebook
- Pad with skillfully chosen bits

Message

Padding

Message	Padding
0100101101011100100	011001010

Transmitted together in one block

Similarities

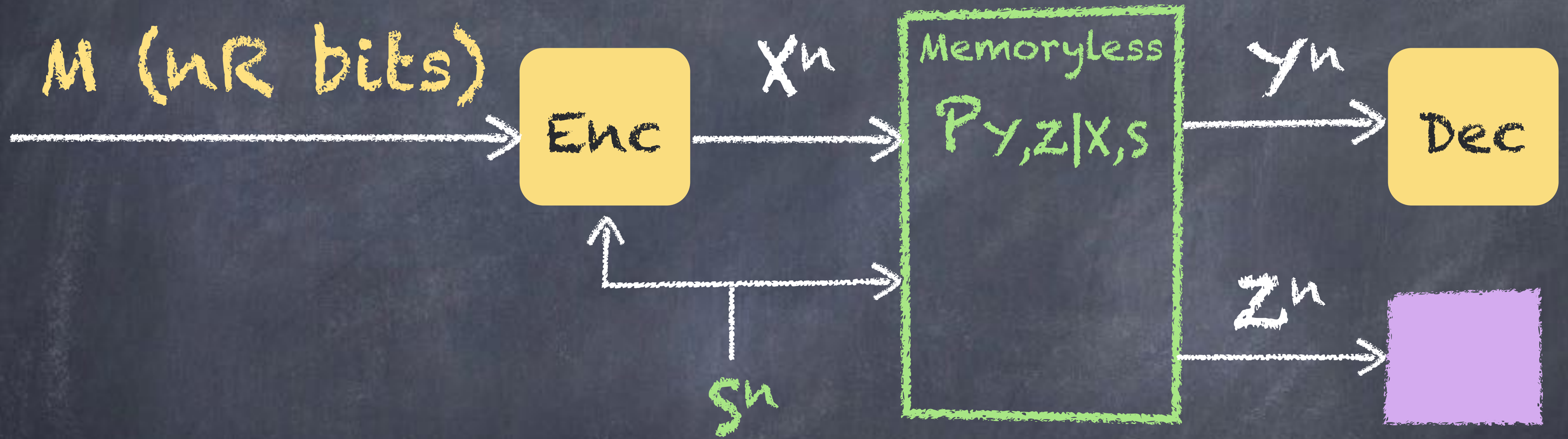
- Virtually the same
  - Same encoding
  - Same converse (except, iid S^n allows a skipped step)
  - Same problem statement:
    - Wiretap: M independent of Z^n
    - Gelfand-Pinsker: M independent of S^n

Wiretap Channels with Random States

with

Ziv Golfeld and Haim Permuter

Wiretap Channel with State



Secrecy Capacity:

- Reliable communication
- Z^n contains no information about M

Same Encoding

$$C_s \geq \max_{P_{X,U|S}} I(U;Y) - \max \left\{ \begin{matrix} I(U;Z), \\ I(U;S) \end{matrix} \right\}$$

Message	Padding
01001011010111100100	011001010

Transmitted together in one block

Chen and Han Vinck, 2006

Extract Key

Assume S is known to the intended receiver as well:

$$C_s \geq \max_{P_{X,U|S}} \min \left\{ \begin{array}{c} I(U;Y|S), \\ H(S|Z,U) \end{array} \right\}$$

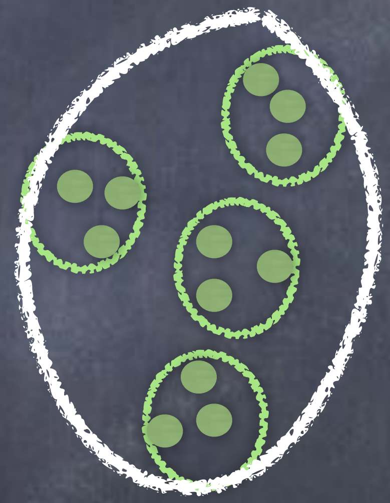
Chia and El Gamal, 2012

Note: They consider causal state information.
This region is adapted to take advantage of non-causal state information.

Our Scheme

Superposition code

Codebook



Two auxiliary variables
U for the clusters
V for the codewords in each cluster

U^n index is padding only
V^n index is message and padding

All secrecy comes from V
U^n is decoded by the eavesdropper

Our Scheme

$$C_s \geq \max_{P_{X,U,V|S} \,:\, I(U;Y) \geq I(U;S)} \min \left\{ \begin{array}{c} I(U,V;Y) - I(U,V;S), \\ I(V;Y|U) - I(V;Z|U) \end{array} \right\}$$

Can mimic Chia and El Gamal's key extraction by setting V=S

Beats previous regions

Other Related Work

- Prabhakaran, Eswaran, and Ramchandran, 2012:
  - Same superposition code but require $U-V-(S,X)$ and $U \perp S$.
- Bassi, Bunin, Piantanida, and Shamai, 2016 (several papers):
  - Key generation and secure communication
  - Sources independent of channel
  - Generalized feedback

Simple Special case

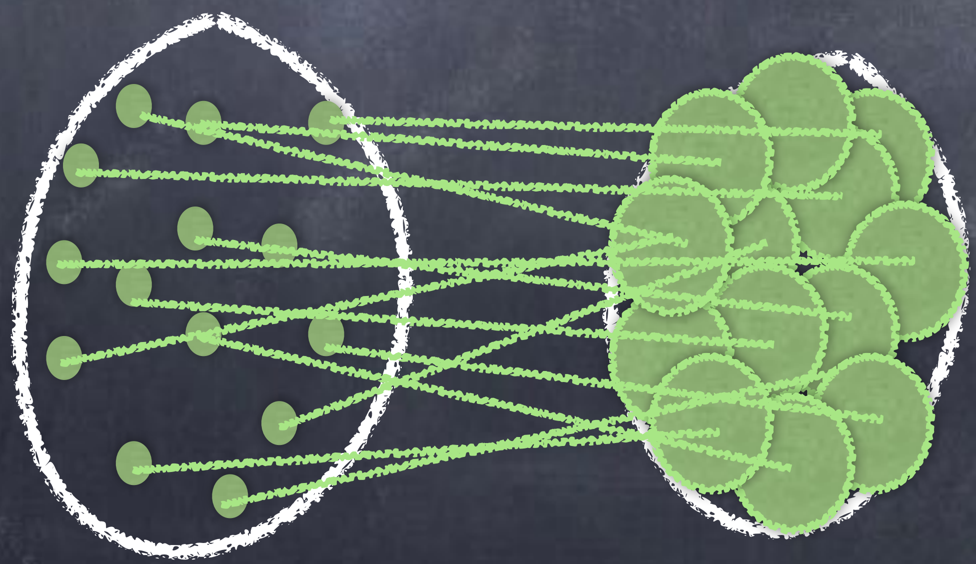
- Unlimited public noise-free channel
- "Key Capacity with one-way communication"

$$C_s = \max_{P_{U,V|S_x}} I(V;S_y|U) - I(V;S_z|U)$$

Achieved by our scheme

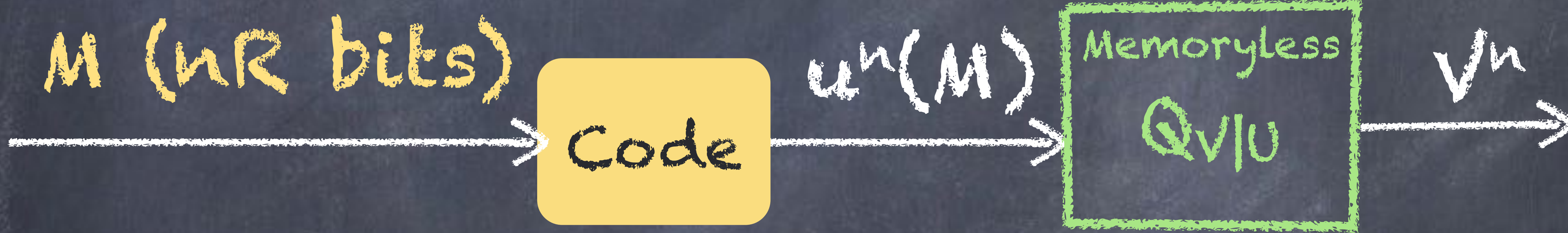
Distribution Approximation Tool

Soft Covering



Soft Covering

- Theorem 6.3 of Wyner's C.I. paper:



Randomly select a codeword
Pass through a memoryless channel
Does induced output distribution match desired?

Output Distribution

Desired output distribution:

$$Q_V(v) = \sum_u Q_{V|U}(v|u) Q_U(u)$$

$$Q_{V^n} = \prod Q_V$$

$$Q_{U^n} = \prod Q_U$$

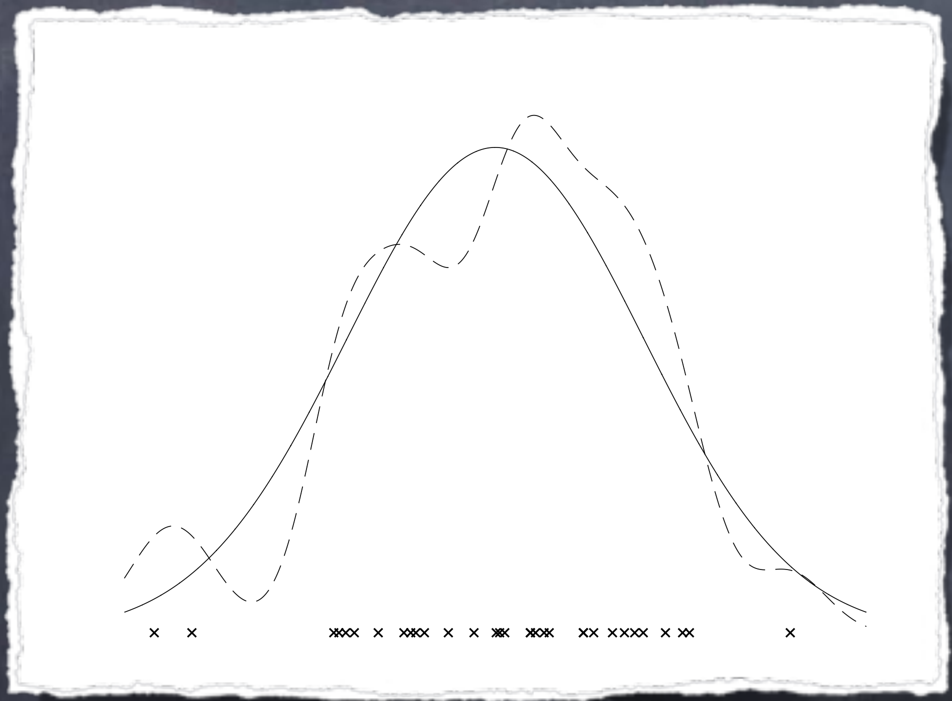
$$Q_{V^n|U^n} = \prod Q_{V|U}$$

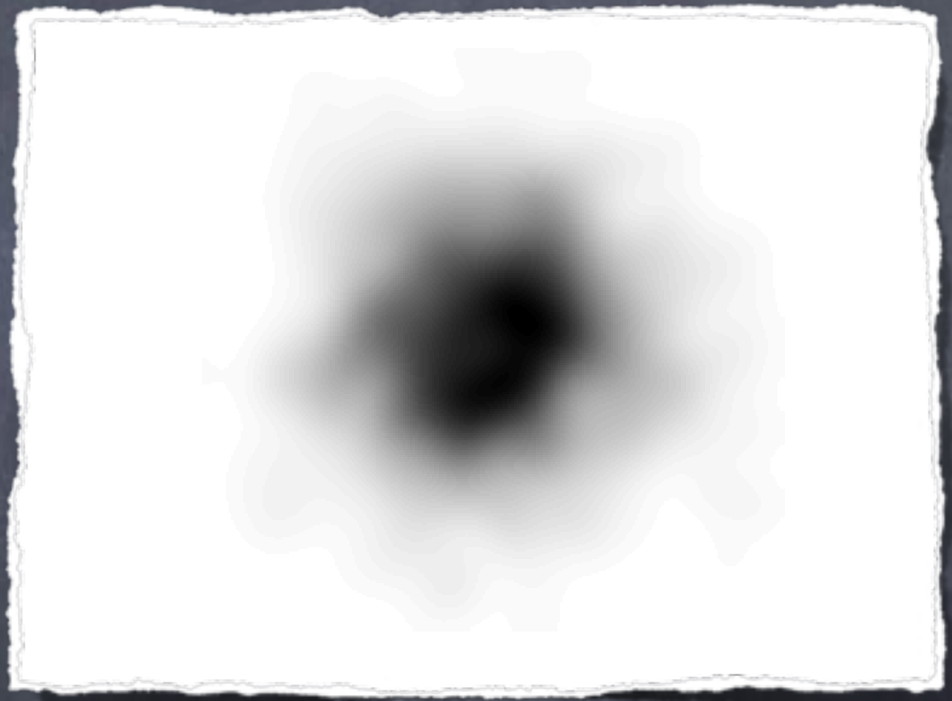
Induced output distribution:

$$P_{V^n|\mathcal{C}} = 2^{-nR} \sum_{u^n(m) \in \mathcal{C}} Q_{V^n|U^n = u^n(m)}$$

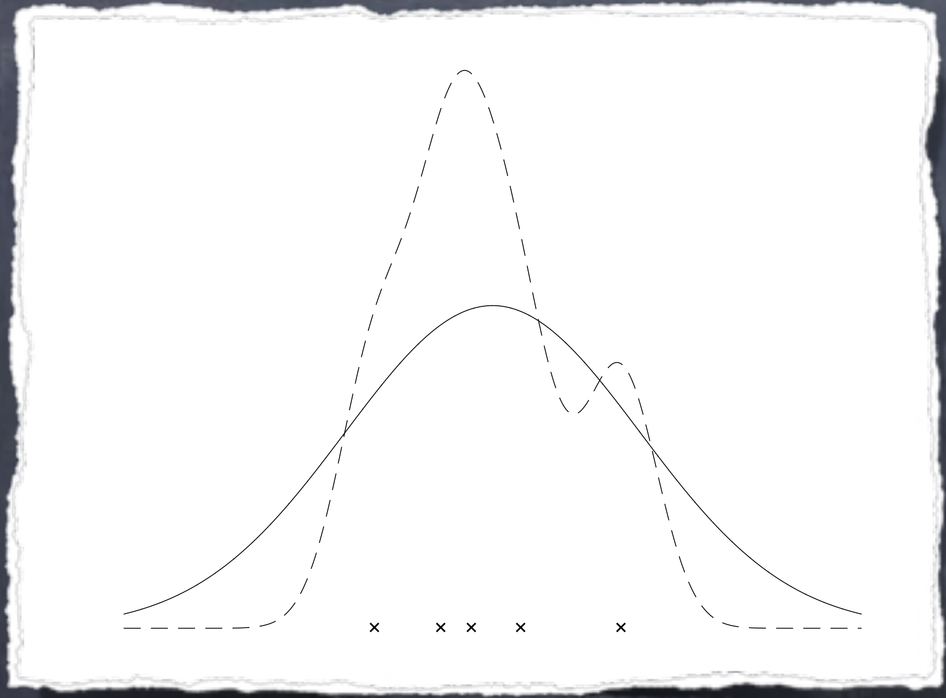
Output Distribution

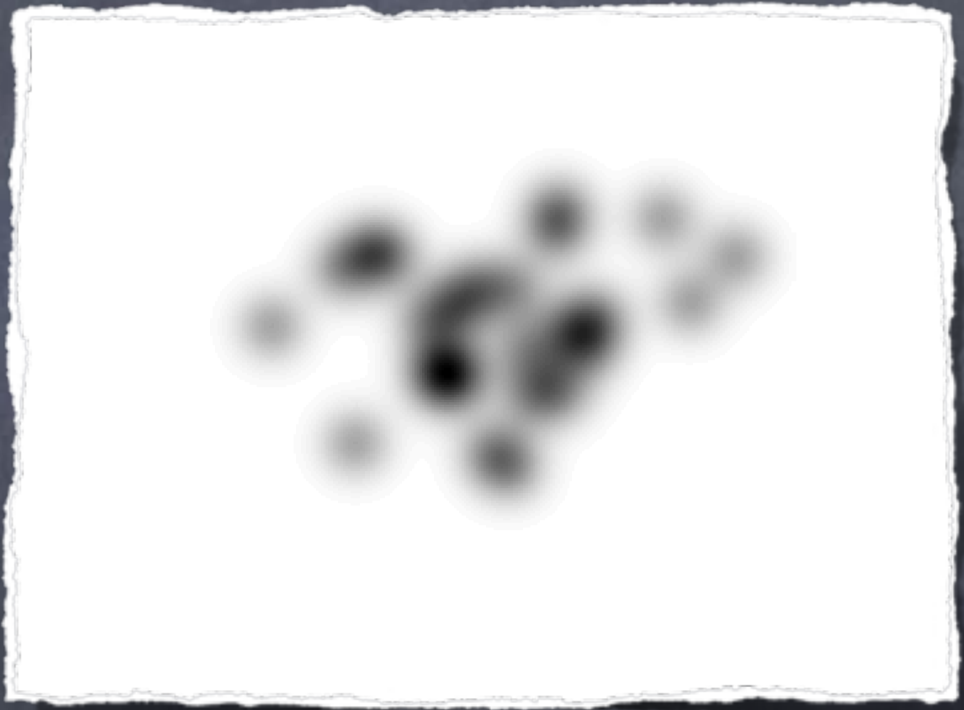
Gaussian Example





Gaussian Example

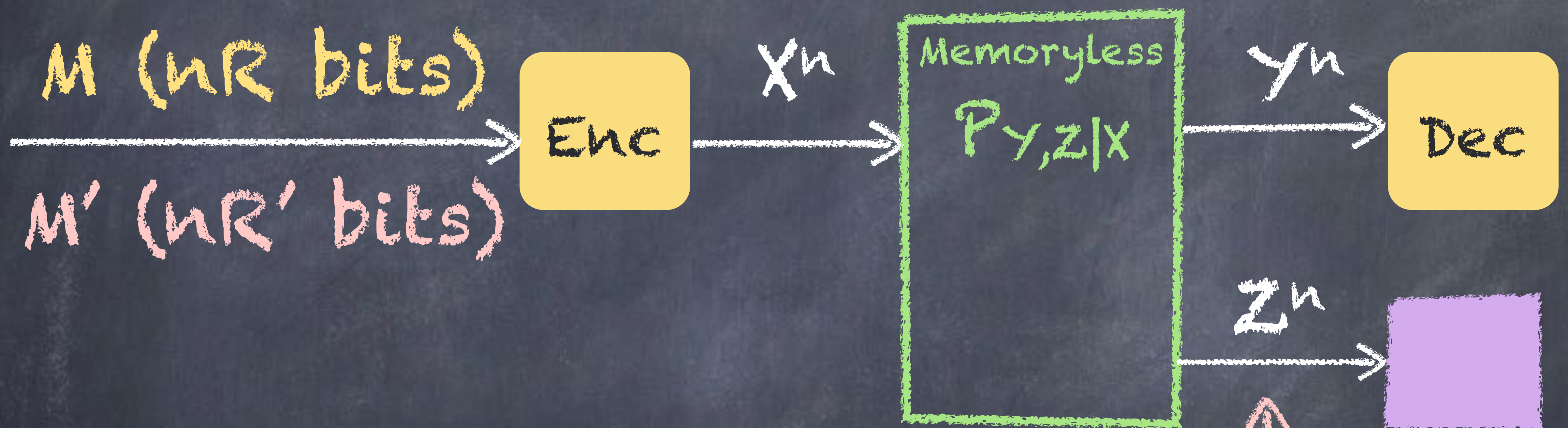




Soft Covering Lemma

- Codebook size: $R > I(U;V)$
- Codebook generation: $U^n(m) \sim Q_U$ i.i.d.
- Success: $P_{V^n|\mathcal{C}} \approx Q_{V^n}$

Wiretap Application



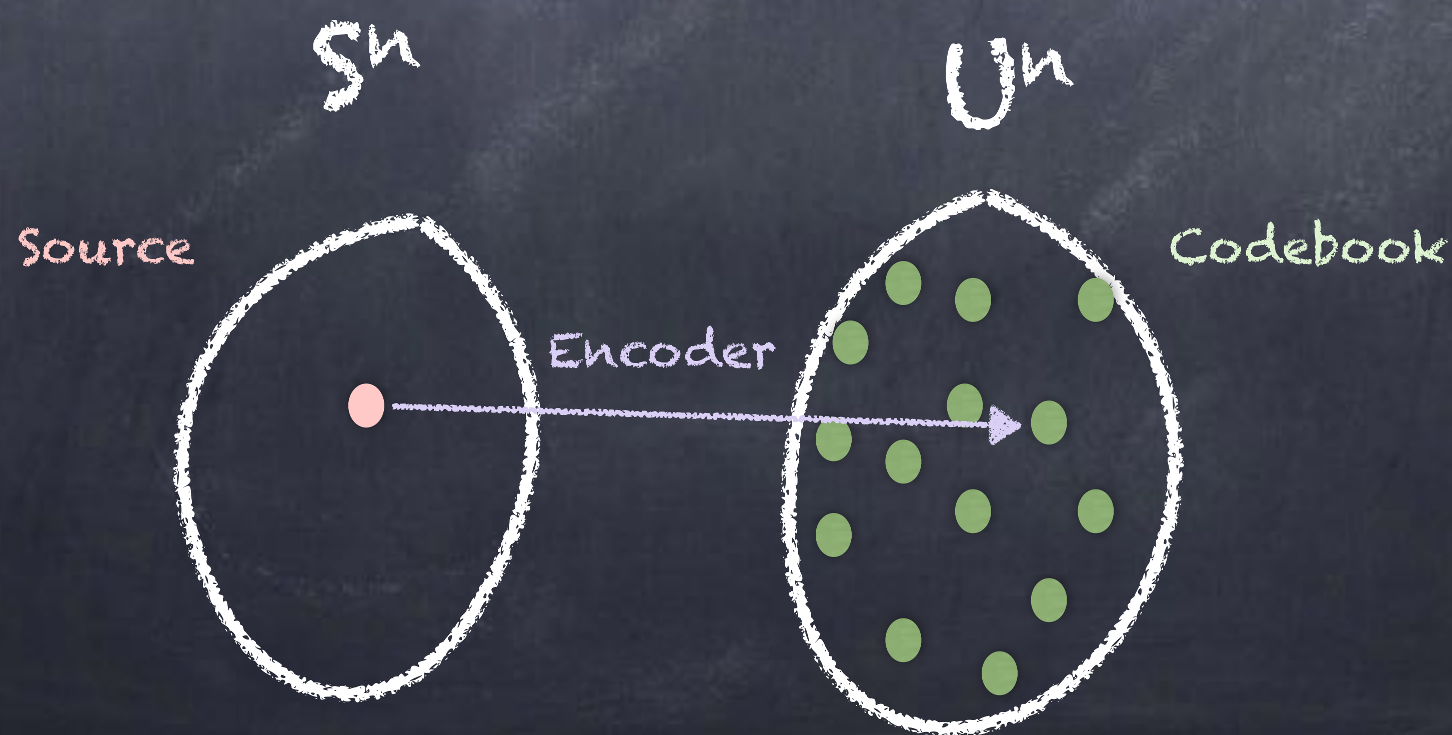
$R' \geq I(X;Z) \implies$ Given M, distribution $\approx \prod P_Z$

Distribution Approximation Trick

"likelihood Encoder" + Soft Covering

Source Coding

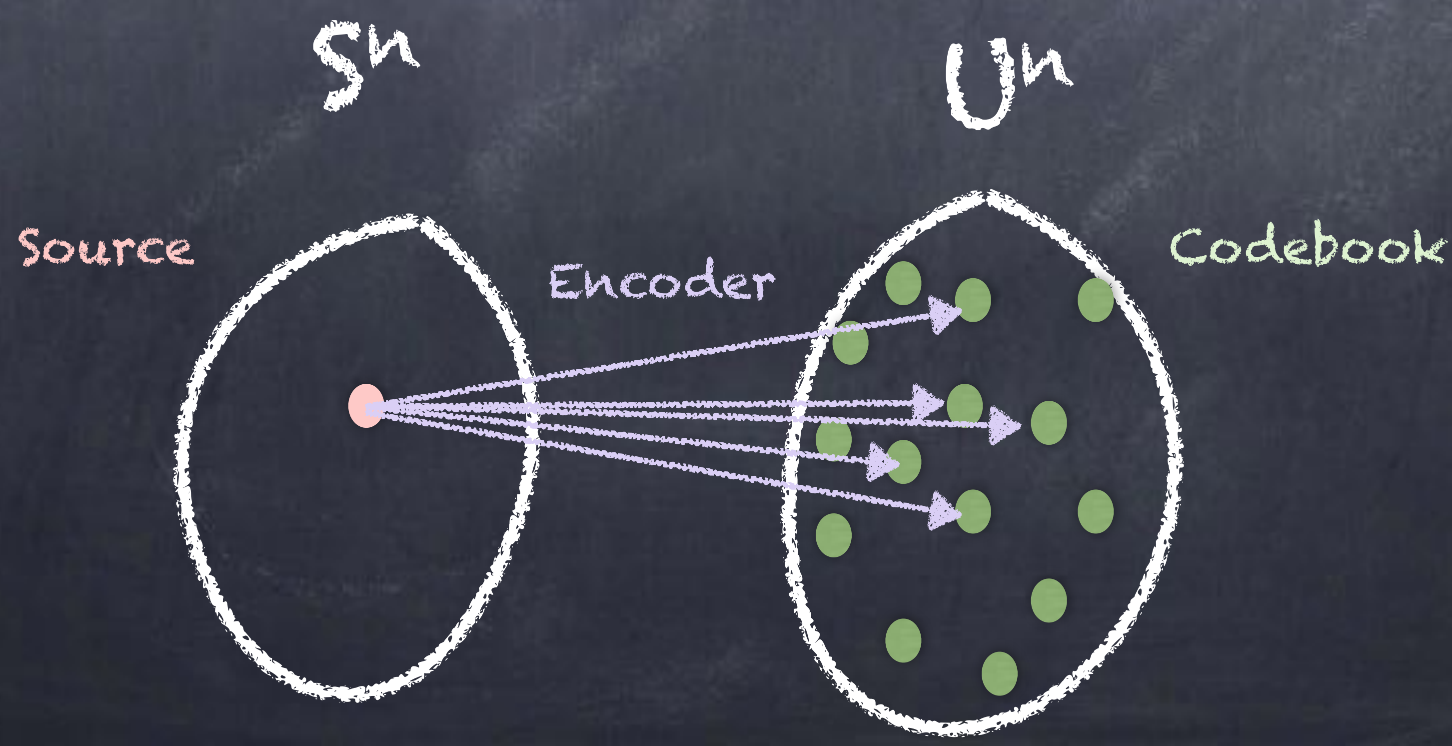
- Source (random process) with known distribution P_S (i.i.d.)
- Desired correlation $P_{U|S}$
- Codebook of U^n sequences
- Encoder selects codeword to empirically match the desired distribution $P_{U,S}$



Lossy Compression

Likelihood Encoder

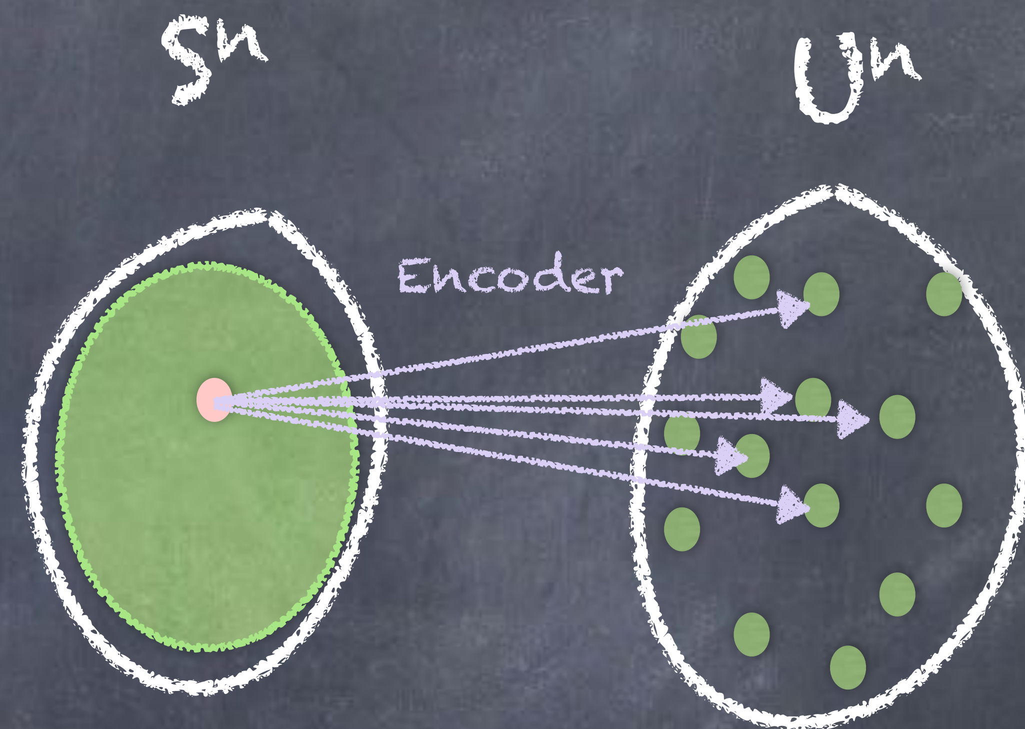
- Source (random process) with known distribution P_S (i.i.d.)
- Desired correlation $P_{U|S}$
- Codebook of U^n sequences
- Encoder stochastically selects codeword proportional to likelihood under $P_{S|U}$



Approximate Distribution

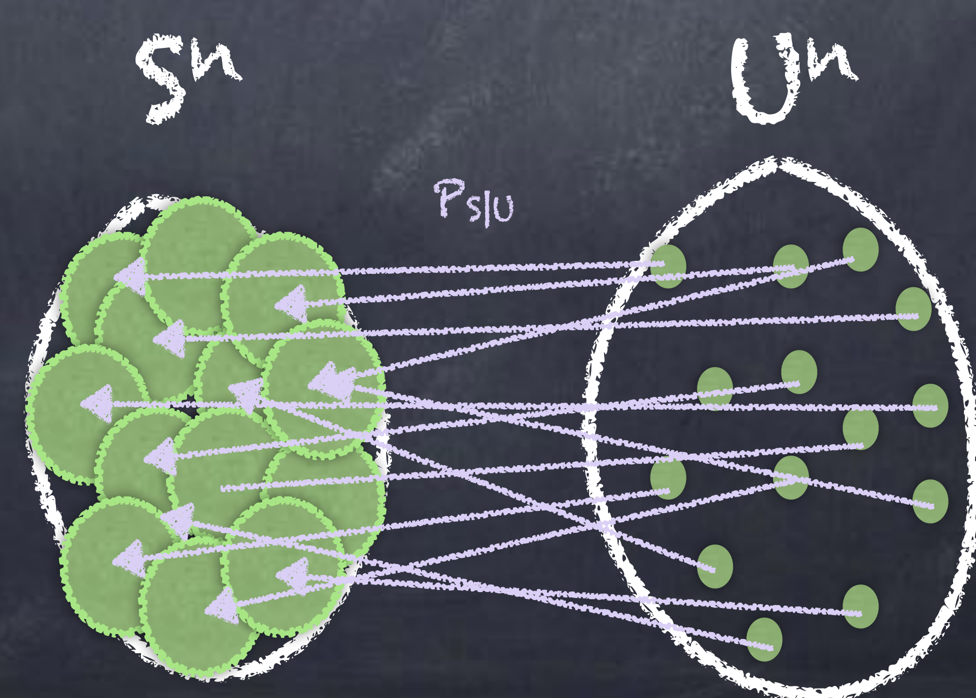
Distribution 1 (induced by encoding):

- S^n is i.i.d. $\sim P_S$
- Likelihood encoder produces U^n

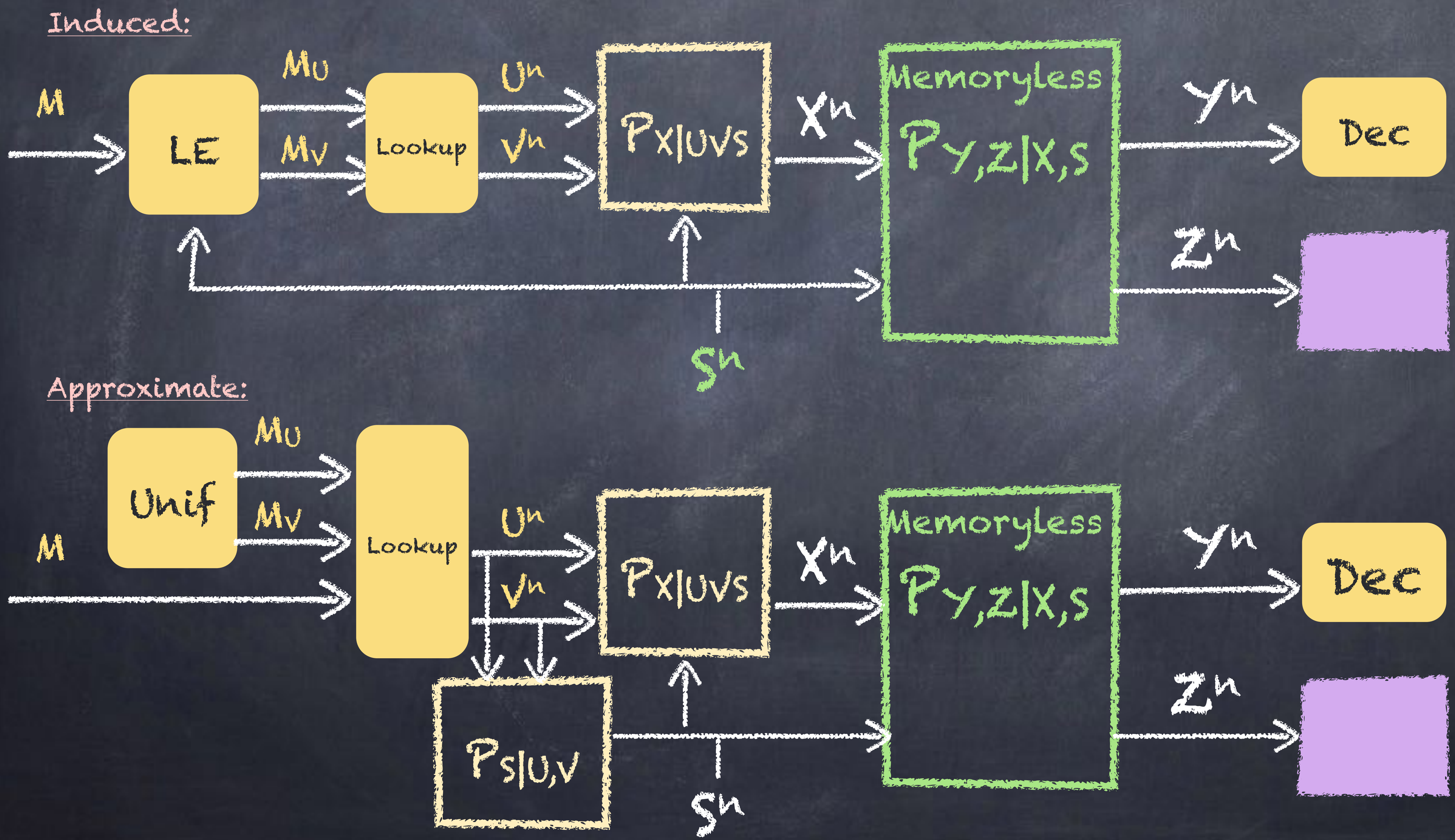


Distribution 2:

- Choose U^n codeword uniformly at random
- Generate S^n memorylessly from $U^n \sim P_{S|U}$



Wiretap with Random States



Differential Privacy as a Mutual Information Constraint

Paul Cuff and Lanqing Yu

Database Privacy

- Let $X_1, X_2, \ldots, X_n$ be entries in a database
    - E.g. X_i is personal information about person i
- Let Y be the response to a query
- The job of the information provider is to answer queries and protect individual privacy

Design $P(y|x)$

Differential Privacy

- ε-DP:
  - Let x and x′ differ in only one entry (i.e. $x_i = x'_i$ for all but one i)
  - $p(y|x) \leq e^{\varepsilon} p(y|x')$
- Why x and x′ differ in only one spot?
  - Convince someone to put their data in your database
- Why multiplicative constraint?
  - Posterior update is small

Dwork, 2006

A Common Technique

- Add Laplacean noise

Weaker DP

- (ε,δ)-DP:
    - Let x and x′ differ in only one entry
    - $P(Y \in A|x) \leq e^{\varepsilon} P(Y \in A|x') + \delta$
- Additive Gaussian noise often provides privacy

Dwork, 2006

Mutual Information Differential Privacy

ε-MI-DP:

$$\max_{i, P_{X^n}} I(X_i; Y | X^{i-1}, X_{i+1}^n) < \epsilon$$

Claim

ε-DP > MI-DP > (ε,δ)-DP

Furthermore, if input or output alphabet is finite,

MI-DP = (ε,δ)-DP

similar to semantic security proof, Bellare, Tessaro, Vardy, 2012

Privacy Ordering

- α-DP > β-DP if for all $\beta>0$ there exists α such that α-DP $\Rightarrow$ β-DP.

Subadditivity of DP

- Multiple queries:
  - If k queries $Y_{q1}, Y_{q2}, \ldots, Y_{qk}$ each have differential privacy ε and are conditionally independent, the combined they have kε privacy.

Simple MI-DP Proof:

$$
\begin{aligned}
I(X;Y_1,Y_2) &= I(X;Y_1) + I(X;Y_2|Y_1) \\
&\leq I(X;Y_1) + I(X;Y_2)
\end{aligned}
$$

For clarity, conditioned database variables are omitted.

Common complaint

- Differentially privacy doesnt not mean that you cant learn about X_i.
  - Consider a database with correlated entries.

Simple MI-DP Explanation:

$$I(X_i;Y) \quad \nleq \quad I(X_i;Y|X^{i-1},X_{i+1}^n)$$

Precise Bounds

(ε,δ)-Closeness

$$P \overset{(\epsilon,\delta)}{\approx} Q$$

if

$$P(A) \leq e^{\epsilon} Q(A) + \delta, \quad \forall A \in \mathcal{F},$$
$$Q(A) \leq e^{\epsilon} P(A) + \delta, \quad \forall A \in \mathcal{F}.$$

Special Cases

$$P \overset{(\epsilon,0)}{\approx} Q \quad \iff \quad \left|\ln \frac{dP}{dQ}(a)\right| \leq \epsilon \quad \forall a \in \Omega.$$

$$P \overset{(0,\delta)}{\approx} Q \quad \iff \quad \|P - Q\|_{TV} \leq \delta.$$

Simple Claim

$$P \overset{(\epsilon,0)}{\approx} Q \quad \Longrightarrow \quad \begin{aligned} D(P\|Q) &\leq \epsilon \text{ nats}, \\ D(Q\|P) &\leq \epsilon \text{ nats}. \end{aligned}$$

Tight Bound

$$P \overset{(\epsilon,0)}{\approx} Q \quad \Longrightarrow \quad \begin{aligned} D(P\|Q) &\leq \epsilon \frac{(e^{\epsilon}-1)\left(1-e^{-\epsilon}\right)}{(e^{\epsilon}-1)+\left(1-e^{-\epsilon}\right)} \text{ nats}, \\ D(Q\|P) &\leq \epsilon \frac{(e^{\epsilon}-1)\left(1-e^{-\epsilon}\right)}{(e^{\epsilon}-1)+\left(1-e^{-\epsilon}\right)} \text{ nats}. \end{aligned}$$

Relative entropy to Mutual Information

If

$$D\left(P_{Y|X=x_1} \| P_{Y|X=x_2}\right) \leq \epsilon \quad \forall x_1, x_2 \in \mathcal{X}$$

then

$$I(X;Y) \leq \epsilon$$

Hint: Radius of information ball

Mutual Information to Total Variation

$$\max_{P_X} I(X;Y) \leq \epsilon \quad \Longrightarrow \quad \begin{array}{c} \left\| P_{Y|X=x_1} - P_{Y|X=x_2} \right\|_{TV} \leq \delta' \\ \forall x_1, x_2 \in \mathcal{X} \end{array}$$

$$\begin{aligned} \delta' &= 1 - 2h^{-1}(\ln 2 - \epsilon) \\ &\leq \sqrt{2\epsilon} \end{aligned}$$

Tightest bound, achieved with binary channel

Finite Alphabet

$$\|P_{Y|X=x_1} - P_{Y|X=x_2}\|_{TV} \leq \delta \quad \forall x_1, x_2 \in \mathcal{X} \implies I(X;Y) \leq \epsilon'$$

$$\epsilon' = 2h(\delta) + 2\delta \ln\left(\min\left\{|\mathcal{Y}|, \max_i |\mathcal{X}_i| + 1\right\}\right)$$

harder step

Continuity of entropy

Continuity of conditional entropy

inspired by Alicki and Fannes, 2004

Estimation of Smoothed Entropy

Paul Cuff, Peter Park, Yucel Altug, Lanqing Yu
(Princeton University)

Estimation of Smoothed Support

Paul Cuff, Peter Park, Yucel Altug, Lanqing Yu
(Princeton University)

Problem

- Take n samples from an unknown distribution (i.id.)
- Estimate the entropy
- Estimate the support

Many Incarnations

- Shakespeare's vocabulary

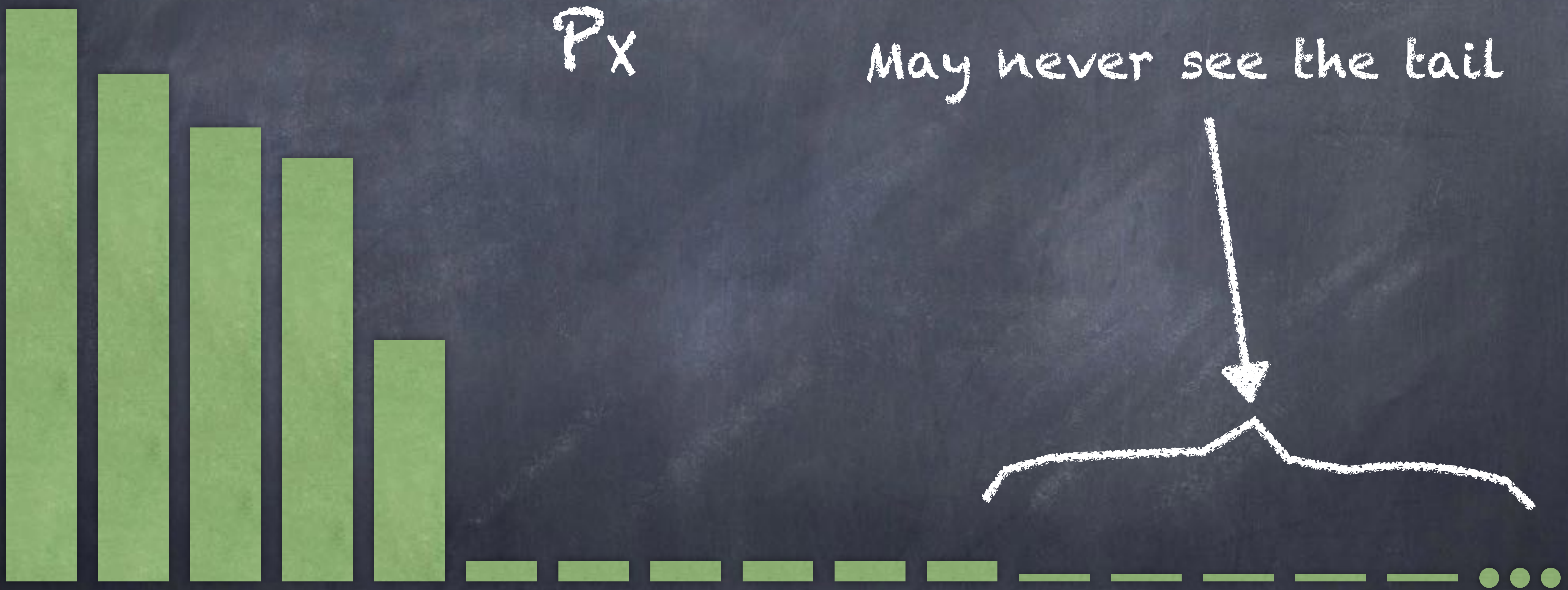
- How many species?

- Good-Turing estimator

Long History

- Recent:
  - [Valiant-Valiant 10]
  - [Acharya-Jafarpour-Orlitsky-Suresh-Wu 13, 15]
  - [Jiao-Venkat-Han-Weissman 15]

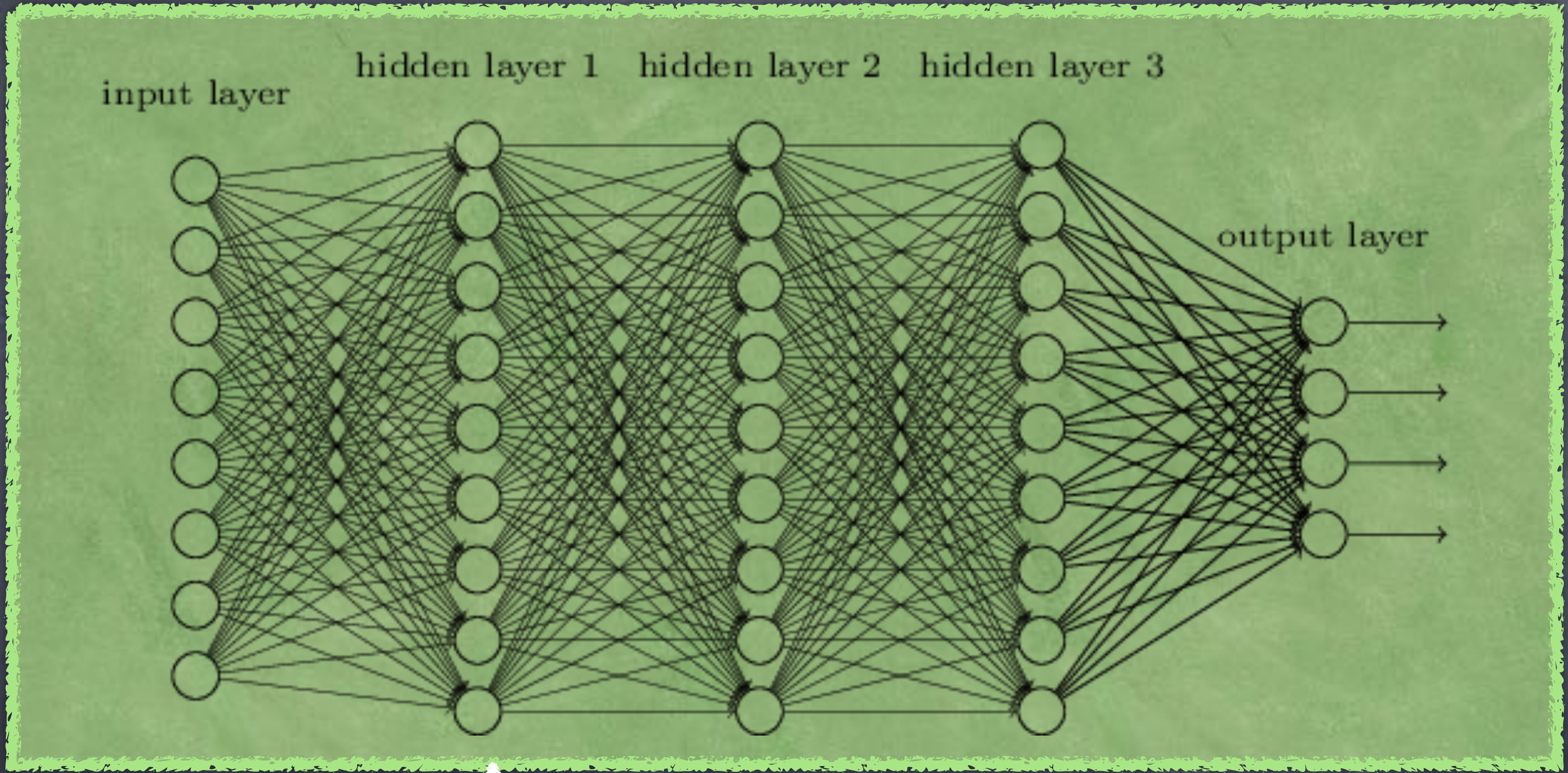
The problem



The usual assumption

- Recent work
  - Entropy: assume a bound on the support size (S)
  - Support: assume a minimum probability mass (1/S)
  - Sample complexity: $n \sim \frac{S}{\log S}$

Death by S



$S = 2^{2000}$

What can we do
with no assumption?

Perhaps nothing

- Cannot reliably decide that entropy or support is finite.
- Reason: Every distribution has an $H=\infty$ neighbor (in total variation)

Yikes

- After one million samples of seeing only one outcome, can we not say anything?

Two Changes

1. Estimate smoothed entropy/support

$$S_\delta(P_X) = \min_{Q : \|P_X - Q\|_{TV} \leq \delta} |\mathrm{Support}(Q)|$$

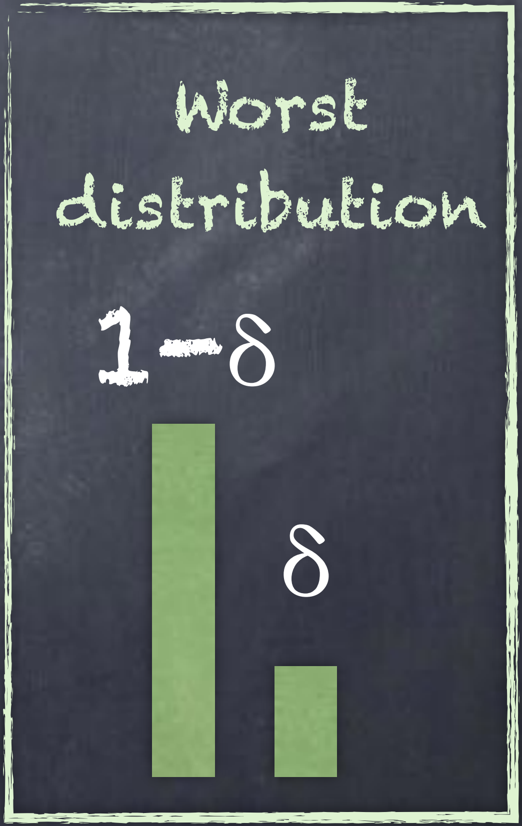
2. Confidence bounds: Estimator can fail as long as it knows when it fails

$$\Big(\quad S_\delta \quad \Big)$$

$\underline{S}_\delta(X^n)$ $\overline{S}_\delta(X^n)$

All Samples the Same

- Conclude: H=0, Support=1
- Error prob. < ε if $n \geq \frac{\log \frac{2}{\epsilon}}{\log \frac{1}{1-\delta}}$
- 459 samples (for δ=ε=0.01)



All Samples Different

- No upper bound possible
- Lower bound: Support = $\Omega(n^2)$

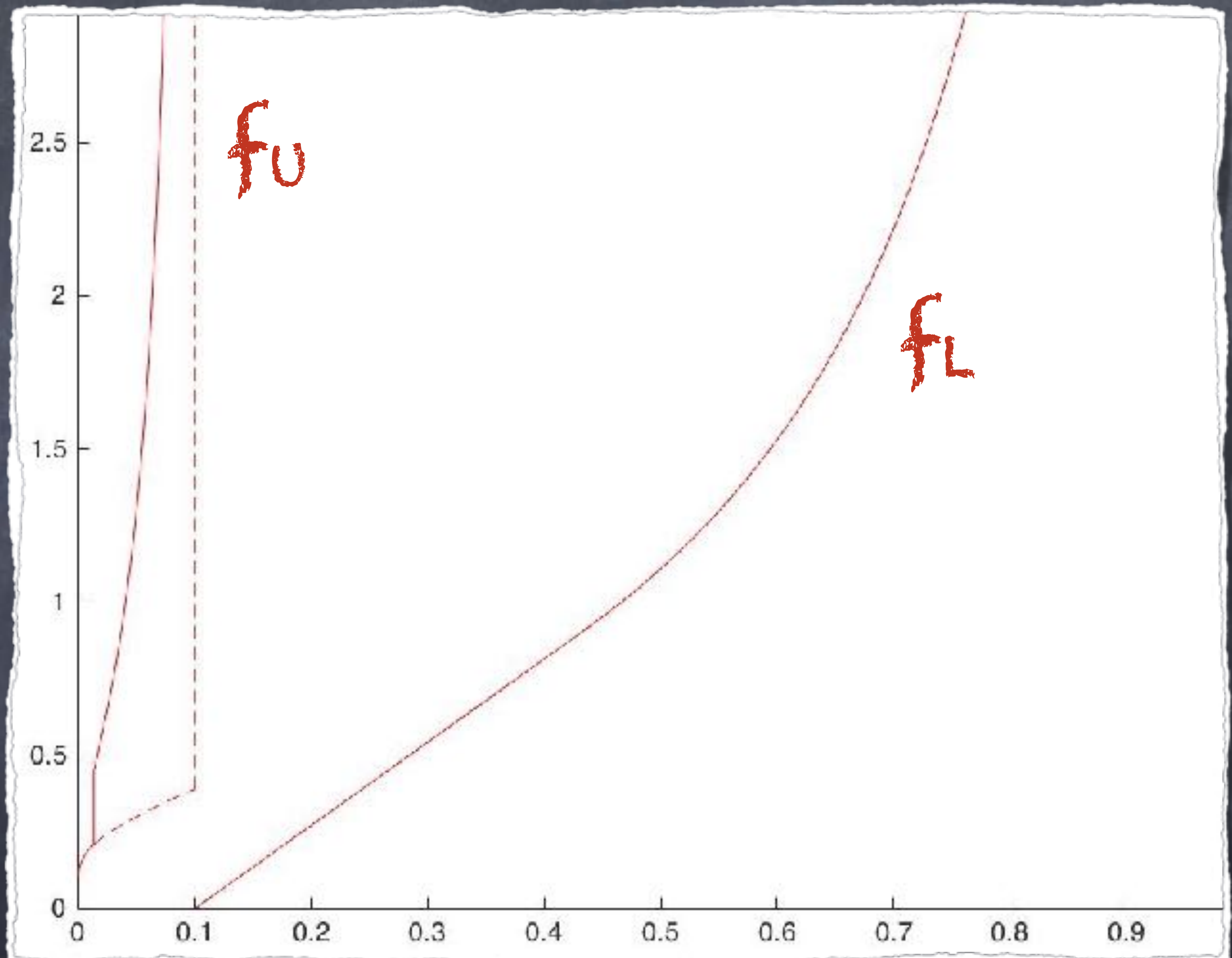
ε-Achieving

$$\sup_P \mathbb{P}\left(S_\delta(P) \notin \left[\underline{S}_\delta(X^n), \overline{S}_\delta(X^n)\right]\right) \leq \epsilon$$

Simple estimator

- Build estimator based on a simple statistic:
  - R = fraction of unique samples

$\delta = 0.1$



$\frac{S_\delta}{n}$

R

Claim

- Choose c>3:
- ε-achieving (for large enough n):

$$\underline{S}_\delta(R) = nf_L\left(R + c\sqrt{\frac{\log n}{n}}\right) \qquad \overline{S}_\delta(R) = nf_U\left(R - c\sqrt{\frac{\log n}{n}}\right)$$

$$f_L(r) = \begin{cases} 0 & r \le \delta \\ e(r-\delta) & \delta < r < \delta + e^{-1}(1-\delta) \\ \frac{1-\delta}{\log \frac{1-\delta}{r-\delta}} & r \ge \delta + e^{-1}(1-\delta) \end{cases} \qquad f_U(r) = \begin{cases} \frac{1-\delta}{\log \frac{\delta}{r}} & r < \delta \\ \infty & r \ge \delta \end{cases}$$

Summary

- Estimator works with no assumptions about the distribution
- Key step was to allow a total variation approximation

Proof – 2 Steps

1. Connect to Poisson Approximation
2. Analyze Poisson Approximation

Step 1

Poisson approximation

Non-discrete part

Bernstein

Discrete Tail

$$\mathbb{P}\left(|R - \mathbb{E}_{X^N} R| > 3\Delta\right) < e\sqrt{n}\left(\exp\left(-\frac{n\Delta^2}{2(1+\Delta)}\right) + \exp\left(-\frac{n\Delta^2}{2}\right) + \exp\left(-\frac{n\Delta^2}{2(1+\Delta/3)}\right) + \frac{1}{n}\right)$$

Plug in $\Delta = \dfrac{c}{3}\sqrt{\dfrac{\log n}{n}}$

Step 2

- Define fingerprint: $$X \sim P_X$$ $$Y = P_X(X) = e^{-\imath_X(X)}$$

- P_Y is fingerprint of P_X

$$S_\delta(P_X) = \mathbb{E}\,\frac{1}{Y} 1\{Y > \mathbb{F}_Y(\delta)\}$$

$$\mathbb{E}_{X^N} R = \mathbb{E}\, e^{-nY}$$