Lecture 13: Channel coding theorem, joint typicality

- Preview and set-up
- Channel-coding theorem
- Joint typical sequences

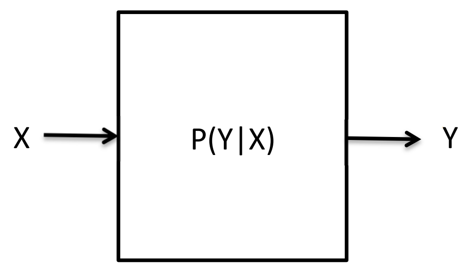
Dr. Yao Xie, ECE587, Information Theory, Duke University

Information channel capacity

For discrete memoryless channel (DMC)

$$C = \max_{p(x)} I(X;Y)$$

- $C \geq 0$ since $I(X;Y) \geq 0$, $C \leq \log |\mathcal{X}|$, $C \leq \log |\mathcal{Y}|$



Discrete: $\mathcal{X}$, $\mathcal{Y}$ discrete
Memoryless: $p(Y^n|X^n) = \prod_{i=1}^{n} p(y_i|x_i)$

Communication system model

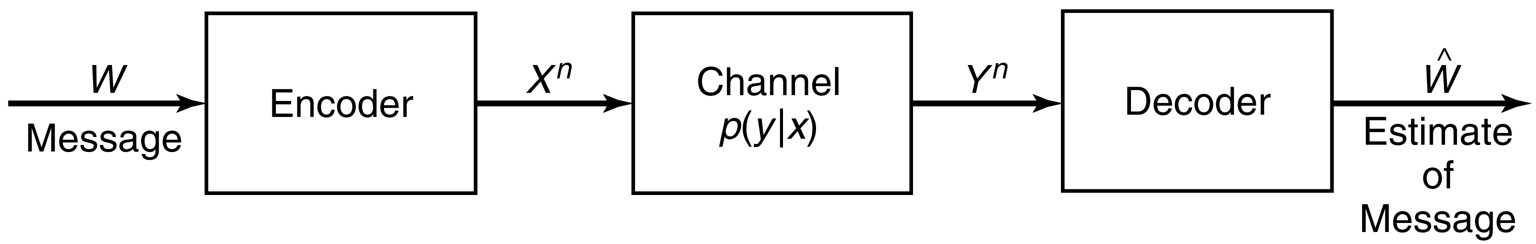


FIGURE 7.1. Communication system.

- $W \in \{1, 2, \ldots, M\}$: source message
- X^n: sequence of channel symbols
- Y^n: output sequence, $Y^n \sim p(y^n|x^n)$
- $\hat{W}$: recovered message, according to decoding function $\hat{W} = g(Y^n)$

Fundamental question

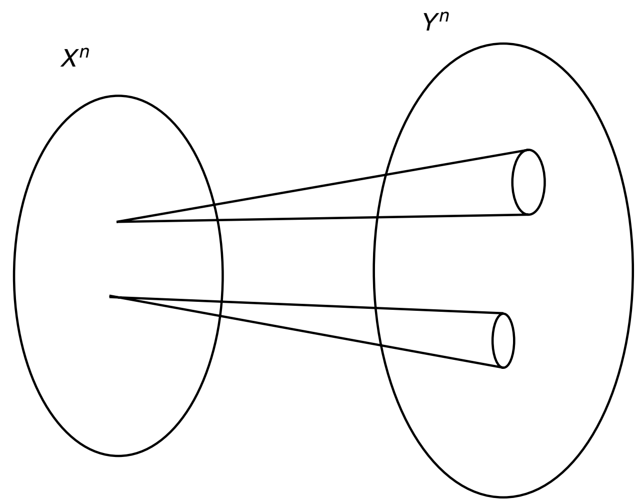
- How fast can we transmit information over a communication channel?
- suppose a source sends r messages per second, and the entropy of a message is H bits per message, information rate is $R = rH$ bits/second
- intuition: as R increases, error will increase
- surprisingly, error can be nearly zero, as long as

$$R < \underbrace{R_{\max}}_{\text{“operational channel capacity”}}$$

- Shannon showed $R_{\max} = C$

Basic idea

- For large block length, every channel looks like the noisy type writer channel
- Channel has a subset of inputs that produce "disjoint" sequences at the output



Code rate

- Rate of an (M, n) code is

$$R = \frac{\log M}{n} \text{ bit per transmission}$$

- On the other hand, we usually write

$$M = \lceil 2^{nR} \rceil$$

Assumption about channel

- Transmit large block length: n over n transmissions
- DMC

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$$

- channel without feedback:

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k), k = 1, \ldots, n$$

Model for encode and decode

- (M, n) code
- An encoder function:

$$f : \{1, \ldots, M\} \to \mathcal{X}^n$$

- Codebook: $[x^n(1), \ldots, x^n(M)]$, each is a codeword
- A decoding function

$$g : \mathcal{Y}^n \to \{1, \ldots, M\}$$

Performance metric

- Conditional probability of error

$$\lambda_i = P\{g(Y^n) \neq i | X^n = x^n(i)\}$$

- Maximal probability of error

$$\lambda^{(n)} = \max_{i=1}^{m} \lambda_i$$

- Average probability of error

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

- $P_e^{(n)} \leq \lambda^{(n)}$
- If W uniform distributed,

$$P_e^{(n)} = P\{W \neq g(Y^n))\}$$

Achievable rate

A rate R is achievable:

if exists a sequence of $\lceil 2^{nR}, n \rceil$ codes such that $\lambda^{(n)} \to 0$ as $n \to 0$.

Channel coding theorem

Theorem. *(Shannon, 1948)*
For a DMC

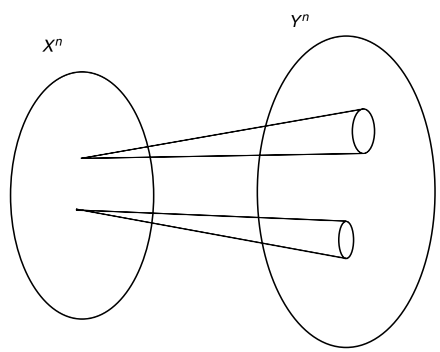
1. *all rates below capacity* $R < C$ *are achievable.*

2. *Converse: any sequence of* $(2^{nR}, n)$ *codes with* $\lambda^{(n)} \to 0$ *must have* $R \leq C$.

Reliable communication over noisy channel is possible!

Proof idea

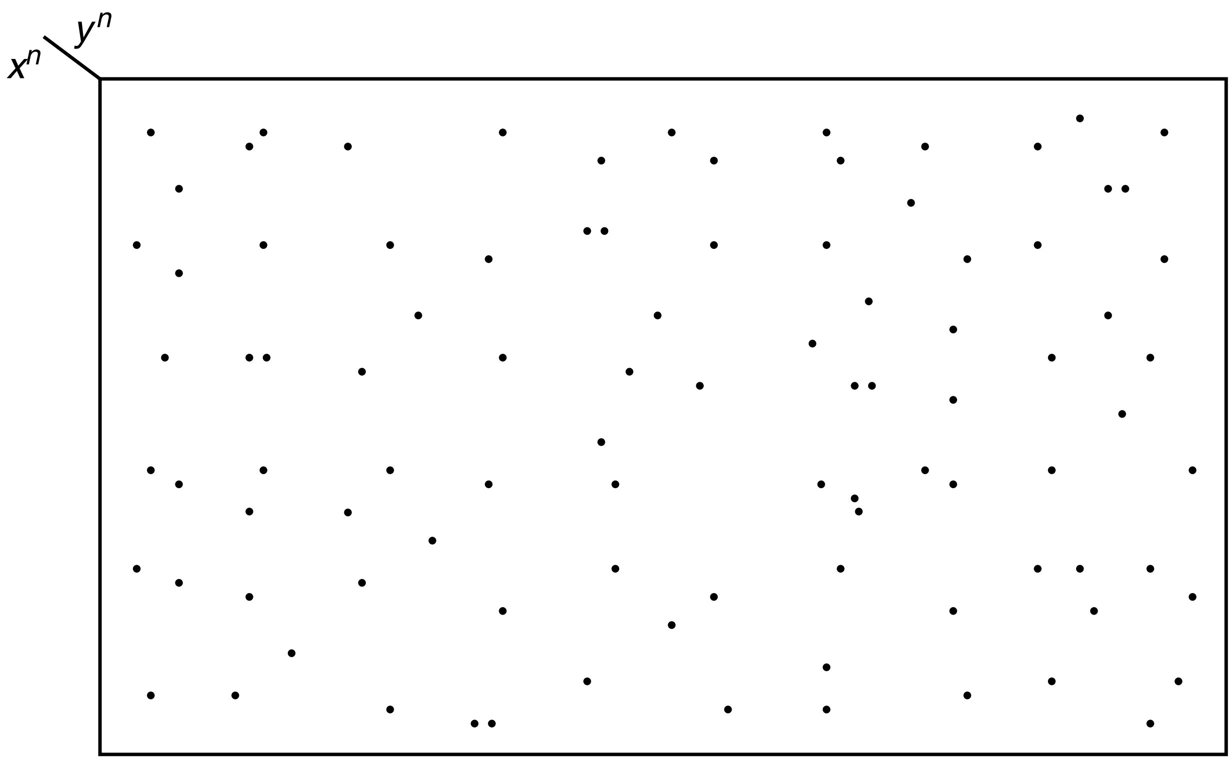
- for each (typical) X^n, there are $\approx 2^{nH(Y|X)}$ possible Y^n
- Total number of (typical) Y^n is $2^{nH(Y)}$
- Total number of disjoint inputs should be $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$
- To formalize these ideas, we need "joint typical sequences"



Joint typical sequences

- Associate a "fan" with each codeword X^n
- We decode Y^n as the ith index if the codeword $X^n(i)$ is "joint typical" with Y^n
- Set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$ is

$$\begin{aligned}A_\epsilon^{(n)} =&\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n :\\&|-\frac{1}{n}\log p(x^n) - H(X)| < \epsilon\\&|-\frac{1}{n}\log p(y^n) - H(Y)| < \epsilon\\&|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)| < \epsilon\}\end{aligned}$$



$2^{nH(X)}$ typical X^n, $2^{nH(Y)}$ typical Y, not all pairs of typical X^n and Y^n are also jointly typical. Any randomly chosen pair is jointly typical is $2^{-nI(X;Y)}$.

Joint AEP

- Let (X^n, Y^n) be sequences of length n drawn i.i.d. according to $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$. Then

  1. $P((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$
  2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$
  3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then

  $$P\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} \leq 2^{-n(I(X;Y)-3\epsilon)}$$

  For sufficient large n,

  $$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |A_\epsilon^{(n)}|$$

  $$P\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

Equipped with definitions and joint typicality, next time we will proof Shannon's channel coding theorem.