ENGG5301 Information Theory Tutorial

Tutorial 9: Final Review

Yanxiao Liu

Department of Information Engineering

Nov 22, 2022

Note

- I will review lectures 1-7 today.
- I will focus on the big picture and some important practices and proofs.
  The goal of this review is to let you know which part you are not familiar with.
  Details are in the lecture slides.
- **Important**: The course is for you to learn something, we are from different background, it is not a competition!

Lecture 1 Review: Definitions

Self-information

- Self-information: $\iota_X(x) = \log \frac{1}{p_X(x)}$
- Joint pmf $p_{X,Y}$: $\iota_{X,Y}(x, y) = \log \frac{1}{p_{X,Y}(x,y)}$

1. $\iota_X(x) \geq 0$
2. For a function f, $\iota_{f(X)}(f(x)) \leq \iota_X(x)$, equality iff f is injective.
3. (Additive) If X, Y are independent, $\iota_{X,Y}(x, y) = \iota_X(x) + \iota_Y(y)$.
4. $\iota_X(x)$ is constant iff X follows a uniform distribution
5. Weakness: information spectrum is a probability distribution, but we want a single number to summarize the amount of information.

Lecture 1 Review

Entropy

- Shannon entropy: $H(X) = \mathbf{E}[\iota_X(X)] = \sum_x p_X(x) \log \frac{1}{p_X(x)}$, which is the average of the self-information.
- Joint entropy: $H(X, Y) = \mathbf{E}[\iota_{X,Y}(X, Y)]$
  1. Positivity: $H(X) \geq 0$ with equality iff X is a constant.
  2. Uniform distribution maximizes entropy: For $|\mathcal{X}| < \infty$, $H(X) \leq \log |\mathcal{X}|$.
  3. Invariance under relabeling: $H(X) = H(f(X))$ for any bijective f.
  4. Conditioning reduces entropy: $H(X|Y) \leq H(X)$ with equality iff X, Y indpt.
  5. Full chain rule: $H(X_1, \ldots, X_n) = \sum_{i=1}^n H(X_i|X^{i-1}) \leq \sum_{i=1}^n H(X_i)$.
  6. $H(X)$ is concave in p_X.

Lecture 1 Review

Convexity

- $f : S \mapsto \mathbb{R}$ is convex if $f(\alpha x + \bar{\alpha} y) \leq \alpha f(x) + \bar{\alpha} f(y)$ for $\alpha \in [0, 1]$.
- Jensen's inequality: is f is convex, then for $X \in \mathbb{R}^n$, $f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)]$.
    1. If f strictly convex, then $f(\mathbf{E}[X]) = \mathbf{E}[f(X)]$ iff X is constant.

Log sum ineq

For $a_1, \ldots, a_n, b_1, \ldots, b_n \geq 0$, $a = \sum_i a_i, b = \sum_i b_i$,

$$\sum_i a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b}$$

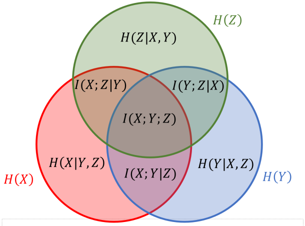
Lecture 2 Review

Overview

- Venn diagrams: Combinatorics VS information theory
- Conditional entropy
- Concavity of entropy
- Conditional Mutual Information

Lecture 2 Review

Information diagram for 3 RVs

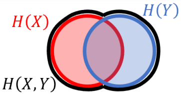


Lecture 2 Review

- Venn diagrams
  1. Shannon-type inequality: inequality implied by $I(X;Y|Z) \geq 0$
  2. $I(X;Y;Z)$ might be negative!
- Intuitively, information is similar to set.

Lecture 2 Review

Sets vs RVs

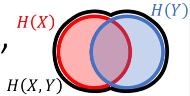


Sets	RVs
Cardinality $\|\cdots\|$	$H(\cdots)$ or $I(\cdots)$
$A \cup B$	(X,Y) (is an RV)
$A \cap B$	"$X;Y$" (not an RV!)
$A \backslash B$	"$X\|Y$" (not an RV!)
$A \cap B = \emptyset$	$X \perp Y$
$B \subseteq A$	Y is a function of X

- $|A| = |A \cap B| + |A \backslash B|$ becomes $H(X) = I(X;Y) + H(X|Y)$
- $|A \cap (B \cup C)| = |A \cap B| + |(A \cap C)\backslash B|$ becomes $I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$
  - Operator precedence: "," then ";" then "|"

Lecture 2 Review

Random variables as "sets"



Union $A \cup B$ of sets A, B	Joint RV (X, Y) of X, Y
$A, B \subseteq A \cup B$	Both X and Y can be determined (i.e., are functions of) (X, Y)
For any C satisfying $A, B \subseteq C$, we have $A \cup B \subseteq C$	For any Z satisfying that both X, Y are functions of Z, then (X, Y) is a function of Z
$\max\{\lvert A\rvert, \lvert B\rvert\}$ $\leq \lvert A \cup B\rvert \leq \lvert A\rvert + \lvert B\rvert$	$\max\{H(X), H(Y)\}$ $\leq H(X,Y) \leq H(X) + H(Y)$
If A, B disjoint, then $\lvert A \cup B\rvert = \lvert A\rvert + \lvert B\rvert$	If X, Y indep. ,then $H(X,Y) = H(X) + H(Y)$

Lecture 2 Review

Lecture 2 Review

- Conditional entropy of Y given X:

$$\begin{aligned} H(Y|X) &= \sum_{x} P_X(x) H(Y|X=x) \\ &= \mathbf{E}\left[\log \frac{1}{p_{Y|X}(Y|X)}\right] \\ &= \sum_{x,y} p_{X,Y}(x,y) \log \frac{1}{p_{Y|X}(y|x)} \end{aligned}$$

  1. Average amount of new info in Y if we already know X.

- Conditional entropy vs set difference:

$$H(Y|X) = H(X,Y) - H(X)$$

- Concavity of entropy

Lecture 2 Review

Lecture 2 Review

- Mutual information: $I(X;Y) = \mathbf{E}\left[\log \frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right]$.
  1. Measures how much information do X, Y share
  2. $I(X;Y) \geq 0$
  3. $I(X;Y) \leq \min\{H(X), H(Y)\}$
  4. $I(X;Y) = H(Y)$ iff Y is a function of X, by $I(X;Y) = H(Y) - H(Y|X)$.
- $I(X;Y)$ is convex in $p_{Y|X}$ and concave in p_X

Lecture 2 Review

Lecture 2 Review

- Conditional mutual information:

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

  $I(X;Y|Z) \geq 0$ with equality iff $X \perp\!\!\!\perp Y|Z$.

  1. Condition may increase/decrease mutual information

  $$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) = H(Y|Z) - H(Y|X,Z)$$

- Information diagram for 4 and 5 RVs

Lecture 2 Review

Lecture 2 Review

- Chain rule: $H(X, Y, Z) = H(X) + H(Y|X) + H(Z|X, Y)$
- Generally,

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1})$$

- For mutual information,

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, \ldots, X_{i-1})$$

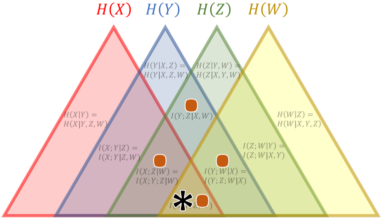
Lecture 2 Review

Markov chain

$$P(X_{i+1} = x_{i+1} | X_1 = x_1, \ldots, X_i = x_i) = P(X_{i+1} = x_{i+1} | X_i = x_i)$$

- Data processing inequality: If $X \to Y \to Z \to W$, then $I(X; W) \leq I(Y; Z)$.

- $I(Y;Z) = I(Y;W) + I(Y;Z|W)$
  $= I(X;W) + I(Y;W|X) + I(Y;Z|W)$



Lecture 2 Review

Lecture 2 Review

- Kullback-Leibler divergence: $D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$.
  1. $D(p\|q) \geq 0$ with equality iff $p = q$.
  2. $I(X;Y) = D(p_{X,Y}\|p_X(x)p_Y(y))$: Mutual information is the divergence between the true joint distribution and the hypothetical joint distribution if X, Y were independent.
  3. It is not a distance measure! (not symmetric)
- Total variation distance: $\delta_{TV}(p,q) = \sup_{A\subseteq\mathcal{X}} |p(A) - q(A)|.$[a].
- Pinsker's inequality: $\delta_{TV}(p,q) \leq \sqrt{\frac{1}{2\log e} D(p\|q)}$.

[a] Rudin, Walter. Principles of mathematical analysis. Vol. 3. New York: McGraw-hill, 1976.

Lecture 3 Review

- If X is uniformly distributed, you need $n \approx H(X) = \log_2 k$ bits to compress X.
- You can do better if you allow n to change according to value of X: **Variable-length compression**
- You should be able to **uniquely decode**: let the decoder know the boundaries of the codewords $m = f(X_1) \| \cdots \| f(X_n)$
- Prefix-free code: can be represented as a binary tree

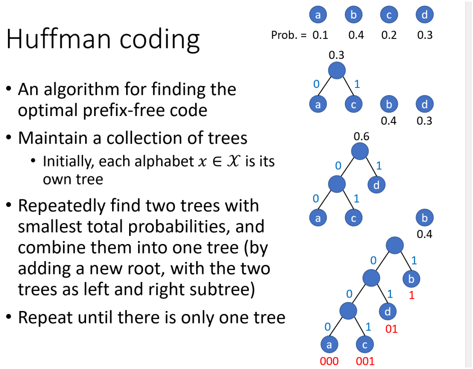
Kraft's inequality

There exists a prefix-free code with $L(f(x)) = \ell_x$ for $x \in \mathcal{X}$ if and only if $\sum_{x \in \mathcal{X}} 2^{-\ell_x} \leq 1$

- Expected length: $\mathbb{E}[L(f(x))] = \mathbb{E}[\ell_x] = \sum_x p_X(x)\ell_x$
- Expected length must be at least $H(X)$(proved in lec3)

Huffman coding

- An algorithm for finding the optimal prefix-free code
- Optimality: attains the smallest possible $\mathbb{E}[\ell_X]$(proved in lec3)



Huffman coding

- An algorithm for finding the optimal prefix-free code
- Maintain a collection of trees
  - Initially, each alphabet $x \in \mathcal{X}$ is its own tree
- Repeatedly find two trees with smallest total probabilities, and combine them into one tree (by adding a new root, with the two trees as left and right subtree)
- Repeat until there is only one tree

Lecture 3: Lossless Compression

Fano's inequality

$X, \hat{X}$ are r.v. over $\mathcal{X}$, $P_e = \mathbb{P}(X \neq \hat{X})$, then

$$H(X|\hat{X}) \leq H_b(P_e) + P_e \log(|\mathcal{X}| - 1) \leq 1 + P_e \log|\mathcal{X}|$$

where $H_b(a) = H(Bern(a))$) is the binary entropy function.

Compree $X_1, \ldots, X_n$ i.i.d. following p_X into fixed length codeword $M = \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ with error probability ϵ_n.

Shannon's source coding theorem

If $R > H(X)$, then there is a code with $\epsilon_n \to 0$. If $R < H(X)$, then there does not exist code with $\epsilon_n \to 0$.

Lecture 3: Lossless Compression

Compree $X_1, \ldots, X_n$ i.i.d. following p_X into fixed length codeword $M = \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ with error probability ϵ_n.

Shannon's source coding theorem

If $R > H(X)$, then there is a code with $\epsilon_n \to 0$. If $R < H(X)$, then there does not exist code with $\epsilon_n \to 0$.

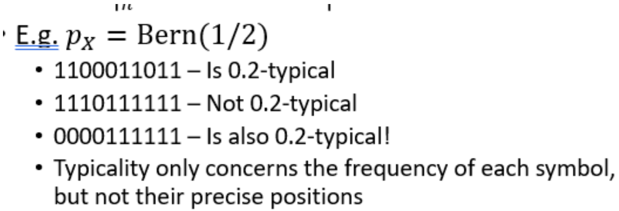
- Achievability follows from Huffman coding
- For converse, assume $\epsilon_n \to 0$. By Fano's inequality,
  $H\left(X^n \middle| \hat{X}^n\right) \leq 1 + \epsilon_n \log(|\mathcal{X}|^n) = 1 + n\epsilon_n \log|\mathcal{X}|$
- $H(X) = \frac{1}{n} H(X^n) \leq \frac{1}{n}\left(H\left(\hat{X}^n\right) + H\left(X^n \middle| \hat{X}^n\right)\right)$
  $\leq \frac{1}{n}(H(M) + 1 + n\epsilon_n \log|\mathcal{X}|)$
  $\leq R + \frac{1}{n} + \epsilon_n \log|\mathcal{X}| \to R$ as $n \to \infty$

Concluding Remarks

- Def of (joint) entropy, **properties**
- **Venn diagrams**
- **Conditional entropy**
- (Conditional) Mutual Information
- Karnaugh map
- Total variation distance
- **Variable-length compression**: uniquely decodability, Prefix-free code, Kraft's inequality, Optimality
- Fano's inequality
- Shannon's source coding theorem

Review of Lecture 4

- Strong typical sequences
- For $x^n = (x_1, \dots, x_n)$, $N(a; x^n) = |\{i : x_i = a\}|$
- δ-strongly typical set $\mathcal{T}_\delta^n(X)$ w.r.t. p_X includes x^n s.t.:
  1. $N(a; x^n) = 0$ for a not in $\mathrm{supp}(p_X)$
  2. $\sum_a \left| \frac{1}{n} N(a; x^n) - p_X(a) \right| \le \delta$
- Eg:



E.g. $p_X = \mathrm{Bern}(1/2)$

- 1100011011 – Is 0.2-typical
- 1110111111 – Not 0.2-typical
- 0000111111 – Is also 0.2-typical!
- Typicality only concerns the frequency of each symbol, but not their precise positions

Review of Lecture 4

Asymptotic equipartition property

There exists $\eta = \eta(p_X, \delta)$ with $\eta(p_X, \delta) \to 0$ as $\delta \to 0$ s.t. $\forall x^n \in \mathcal{T}_\delta^n(X)$,

$$2^{-n(H(X)+\eta)} \leq p_X^n(x^n) \leq 2^{-n(H(X)-\eta)} \tag{1}$$

- All typical sequences have similar probabilities

Cor

$$|\mathcal{T}_\delta^n(X)| \leq 2^{n(H(X)+\eta)} \tag{2}$$

Review of Lecture 4

Most sequences are typical

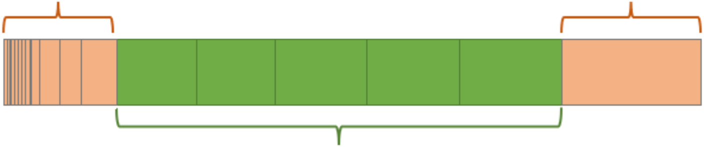
For $\delta > 0$ and i.i.d. $X_1, \ldots .X_n \sim p_X$,

$$\lim_{n\to\infty} \mathbf{P}(X^n \in \mathcal{T}_\delta^n(X)_{=}1 \qquad (3)$$

- Each $x^n \in \mathcal{T}_\delta^n(X)$ has probability approximately $2^{-nH(X)}$.
- A random sequence is probably typical.

- Illust. of the pmf $p_X^n(x^n)$ sorted in ascending order:

There might be more atypical sequences, but they make up a small portion of the space



Typical sequences $\mathcal{T}_\delta^n(X)$ have similar probabilities, contribute to most of the space, and approximately "equipartition" the space

Review of Lecture 4

Shannon's source coding theorem

Compress $X_1, \ldots, X_n$ i.i.d. following p_X into codeword $M = \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ with error prob ϵ_n.
If $R > H(X)$, there is a code with $\epsilon_n \to 0$.

- Proof: using typical set.

Review of Lecture 4

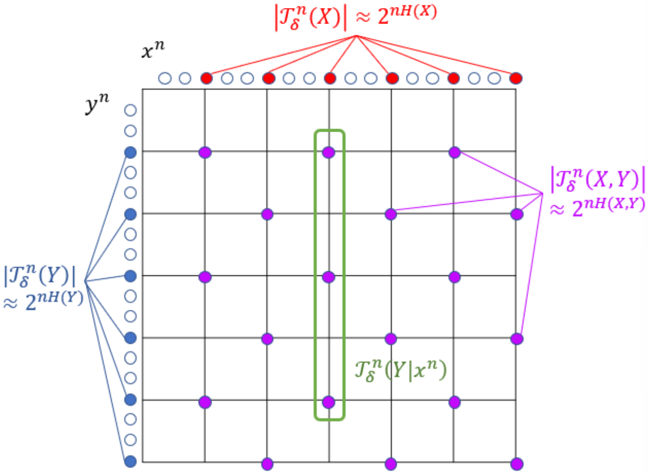
Jointly typical sequence

$$\mathcal{T}_\delta^n(X, Y) =$$
$$\{((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n : \hat{p}_{x^n, y^n} << p_{X,Y}, \delta_{TV}(\hat{p}_{x^n, y^n}, p_{X,Y}) \leq \delta/2\}$$

where $\hat{p}_{x^n, y^n}(a, b) = \frac{1}{n}|\{i : (x_i, y_i) = (a, b)\}|$

- Each $(x^n, y^n) \in \mathcal{T}_\delta^n(X, Y)$ has prob approximately $2^{-nH(X,Y)}$
- $\lim\limits_{n\to\infty} \mathbf{P}((X^n, Y^n) \in \mathcal{T}_\delta^n(X, Y) = 1$
- $|\mathcal{T}_\delta^n(X, Y)| \approx 2^{-nH(X,Y)}$
- Preservation: $Y = f(X)$, $x^n \in \mathcal{T}_\delta^n(X) \Rightarrow y^n \in \mathcal{T}_\delta^n(Y)$ s.t. $y_i = f(x_i)$
- Consistency: $(x^n, y^n) \in \mathcal{T}_\delta^n(X, Y) \Rightarrow x^n \in \mathcal{T}_\delta^n(X), y^n \in \mathcal{T}_\delta^n(Y)$

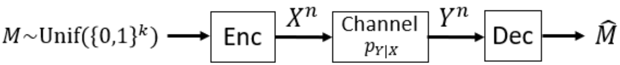
Review of Lecture 4


$|\mathcal{T}_\delta^n(X)| \approx 2^{nH(X)}$
x^n
y^n
$|\mathcal{T}_\delta^n(X,Y)|$
$\approx 2^{nH(X,Y)}$
$|\mathcal{T}_\delta^n(Y)|$
$\approx 2^{nH(Y)}$
$\mathcal{T}_\delta^n(Y|x^n)$

Review of Lecture 5

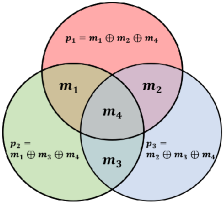
- Channel: A channel is a conditional distribution $p_{Y|X}$: with input X, you have an output Y.
- Memoryless channel: The channel is memoryless, the different channel uses are independent of each other
- Discrete memoryless channel:
  1. Binary symmetric channel (BSC)
  2. Binary erasure channel (BEC)

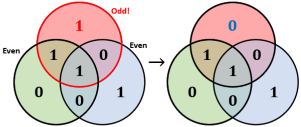
Review of Lecture 5



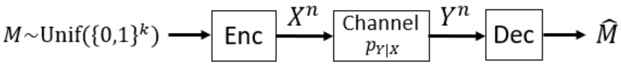
- Similar def of Encoder, Decoder, Block error probability
- Rate of the code (the number of message bits sent per channel use) is k/n.
  1. Repetition code: Rate $1/t$.
  2. Hamming (7,4) code for BSC:

$$p_1 = m_1 \oplus m_2 \oplus m_4$$
$$p_2 = m_1 \oplus m_3 \oplus m_4$$
$$p_3 = m_2 \oplus m_3 \oplus m_4$$


$p_1 = m_1 \oplus m_2 \oplus m_4$
m_1
m_2
m_4
$p_2 = m_1 \oplus m_3 \oplus m_4$
m_3
$p_3 = m_2 \oplus m_3 \oplus m_4$

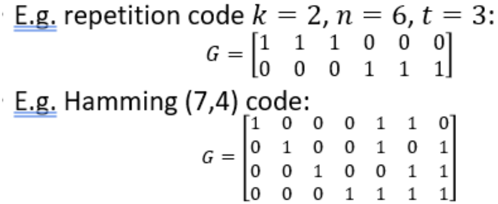


Review of Lecture 5



Linear Codes

Encode $m \in \mathbb{F}_2^k$ into $f(m) = mG$ s.t. G is called the generator matrix.



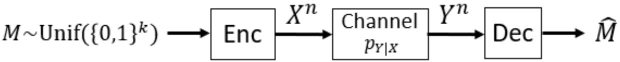
- E.g. repetition code $k = 2, n = 6, t = 3$:

$$G = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

- E.g. Hamming (7,4) code:

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

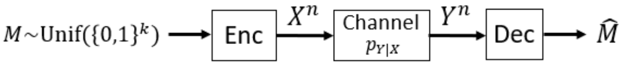
Review of Lecture 5



Asymptotic channel coding

- Send message of $\approx nR$ bits using n channel uses
- R is achievable if there is a sequence of codes f_n, g_n s.t. $\epsilon_n \to 0$ as $n \to \infty$.

Review of Lecture 5



Channel coding theorem

The information capacity of a discrete memoryless hannel $p_{Y|X}$ is

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

s.t. $p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y|x)$

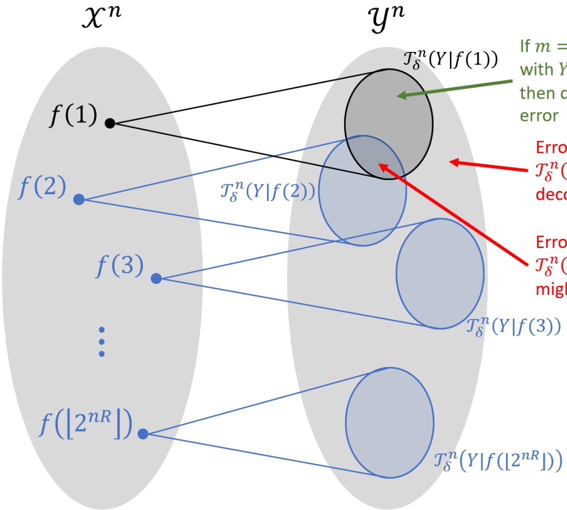
- Achievability
- Converse

Review of Lecture 6

Achievability: Random coding

- If we construct the code randomly, then it is good with high probability
- Random codebook: Generate $f(1), \dots, f(\lfloor 2^{nR} \rfloor)$ i.i.d. following p_X^n.
- Joint typicality decoder: For $y^n \in \mathcal{Y}^n$, if there is a unique m s.t. $(f(m), y^n) \in \mathcal{T}_\delta^n(X, Y)$, take $g(y^n) = m$; o.w. set $g(y^n)$ to be an arbitrary value.
- Assume $m = 1$ is sent, error event:
  1. E_1: $(f(1), y^n) \notin \mathcal{T}_\delta^n(X, Y)$
  2. E_2: There is a wrong $m \neq 1$ with $(f(1), y^n) \in \mathcal{T}_\delta^n(X, Y)$

Review of Lecture 6



$\mathcal{X}^n$
$\mathcal{Y}^n$
$f(1)$
$f(2)$
$f(3)$
$f(\lfloor 2^{nR} \rfloor)$
$\mathcal{T}_\delta^n(Y|f(1))$
$\mathcal{T}_\delta^n(Y|f(2))$
$\mathcal{T}_\delta^n(Y|f(3))$
$\mathcal{T}_\delta^n(Y|f(\lfloor 2^{nR} \rfloor))$

If $m = 1$ is the unique m with $Y^n \in \mathcal{T}_\delta^n(Y|f(m))$, then decode $\widehat{M} = 1$, no error

Error event E_1: If $Y^n \notin \mathcal{T}_\delta^n(Y|f(1))$, will not decode $\widehat{M} = 1$

Error event E_2: If $Y^n \in \mathcal{T}_\delta^n(Y|f(m))$, $m \neq 1$, might decode wrong $\widehat{M}$

Review of Lecture 6

Avg error prob. vs Max error prob

- If M_n follows another distribution, the error prob. may no longer be small! $\bar{\epsilon}_n = \max_n \mathbf{P}(\hat{M}_n \neq m | M_n = m)$
- Given a sequence of codes with $\epsilon_n \to 0$, convert it to a sequence of codes with $\bar{\epsilon}_n \to 0$.
- Use Markov's inequality and show $\mathbf{P}(\hat{M}_n \neq m | M_n = m) < 2\epsilon_n$.

Review of Lecture 6

Converse

If there is a sequence of codes with $\epsilon_n \to 0$, then $R \leq C = \max_{p_X} I(X;Y)$

Proof

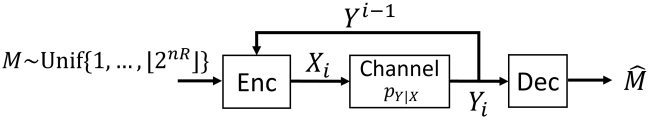
1. Lemma: $I(X^n; Y^n) \leq \sum_{i=1}^n I(X_i; Y_i)$
2. Fano's inequality: $H(M|\hat{M}) \leq 1 + \epsilon_n \log\lfloor 2^{nR} \rfloor$

$$\begin{aligned} \log\lceil 2^{nR} \rceil &= H(M) = I(M;\hat{M}) + H(M|\hat{M}) \\ &\leq I(X^n;Y^n) + o(n) \\ &\leq \sum_{i=1}^n I(X_i;Y_i) + o(n) \\ &\leq nC + o(n) \end{aligned}$$

3. Take $n \to \infty$

- There are alternative proofs.

Review of Lecture 6



Channel with feedback

The (operational) capacity of DMC with perfect feedback is the same as the capacity without feedback

- Note $p_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$ may fail for memoryless channels with feedback

Review of Lecture 6

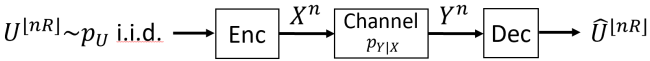
Converse

- $(M, Y^{i-1} \to X_i \to Y_i)$ forms a Markov chain
- Fano's ineq $H(M|\hat{M}) = o(n)$

$$\begin{aligned}
\log\lceil 2^{nR}\rceil &= H(M) = I(M;\hat{M}) + H(M|\hat{M}) \\
&\leq I(M;Y_i) + o(n) \\
&= \sum_{i=1}^{n} I(M;Y_i|Y^{i-1}) + o(n) \\
&\leq \sum_{i=1}^{n} I(M,Y^{i-1};Y_i) + o(n) \\
&\leq \sum_{i=1}^{n} I(X_i;Y_i) + o(n) \\
&\leq nC + o(n)
\end{aligned}$$

- Take $n \to \infty$

Review of Lecture 6



$U^{\lfloor nR \rfloor} \sim p_U$ i.i.d.

Enc

X^n

Channel

$p_{Y|X}$

Y^n

Dec

$\hat{U}^{\lfloor nR \rfloor}$

Joint source-channel coding

The supremum of achievable R is $C/H(U)$ where C is the capacity of $p_{Y|X}$.

Review of Lecture 7

Lossy Compression

Compress $X \in \mathcal{X}$ into $M = f(X)$ and decompress $\hat{X} = g(M)$.
Instead of lossless compression, lossy compression only requires $\hat{X}$ to be close to X

Distortion measure

Suppose $\hat{\mathcal{X}}$ is the reconstruction alphabet, a **distortion measure** is a function $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$:

1. $d(x, \hat{x})$ measures the **distance** between x and $\hat{x}$.
2. It is not required that $d(x, \hat{x}) = 0$ or $d(x, \hat{x}) = d(\hat{x}, x)$
3. Eg: $d(x, \hat{x}) = \log \frac{1}{\hat{x}(x)}$

Review of Lecture 7

One-shot lossy compression

- Compress $X \sim p_X$ into M and decompress $\hat{X} = g(M) \in \hat{\mathcal{X}}$.
- Minimize the **expected distortion** subject to:
  1. Cardinality constraint: $M \in \{1, \ldots, k\}$

$$\min_{p_{\hat{X}|X}: H_0(\hat{X}) \leq \log k} \mathbf{E}[d(X, \hat{X})]$$

  2. Entropy constraint: $H(M) \leq R$

$$\min_{p_{\hat{X}|X}: H_0(\hat{X}) \leq R} \mathbf{E}[d(X, \hat{X})]$$

Review of Lecture 7

Asymptotic lossy compression

Compress $X_1, \ldots, X_n \sim p_X$ into $M = f_n(X^n) \in \{1, \ldots, \lceil 2^{nR} \rceil\}$ and decompress into $\hat{X}^n = g_n(M) \in \hat{\mathcal{X}}^n$.

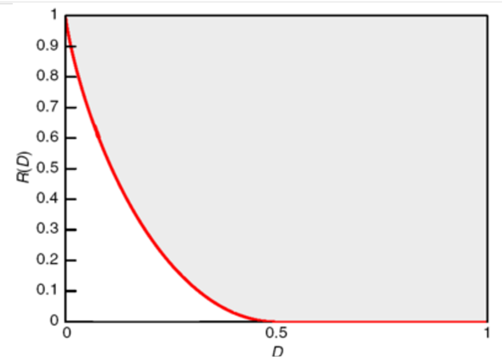
- Average distortion $d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$
- rate-distortion pair (R, D) is achievable if there is a sequence of codes f_n, g_n s.t.

$$\lim_{n \to \infty} \mathbf{E}\left[d(X^n, g_n(f_n(X^n)))\right] \leq D$$

- The rate-distortion region is the closure of the set of achievable rate-distortion pairs
- The (operational) rate-distortion function $R(D)$ is the infimum of rates R s.t. (R, D) is in rate-distortion region.

Review of Lecture 7

- $R(D) = 0$ if $D \geq D_{max} = \min_{\hat{x}} \mathbf{E}[d(X, \hat{x})]$
- $R(D) \leq H(X)$ if $D \geq D_{min} = \mathbf{E}[\min_{\hat{x}} d(X, \hat{x})]$
- The rate-distortion region is convex
-



Review of Lecture 7

Shannon's lossy source coding theorem

- The information rate-distortion function is

$$R_I(D) = \min_{p_{\hat{X}|X}:\mathbf{E}[d(X,\hat{X})]\le D} I(X;\hat{X})$$

- Theorem: $R(D) = R_I(D)$ for $D \geq D_{\min}$
  1. Achievability: $\forall \epsilon > 0$, and $p_{\hat{X}|X}$ with $\mathbf{E}[d(X,\hat{X})] \leq D$ and $R > I(X;\hat{X}) + \epsilon$, we can construct a scheme with
  $$\lim_n \mathbf{E}[d(X^n,\hat{X}^n)] \leq D + \epsilon \tag{4}$$
  2. Converse: Any scheme with $\lim_n \mathbf{E}[d(X^n,\hat{X}^n)] \leq D$ satisfies $R \geq R_I(D)$
- Better than one-shot scheme since $I(X;\hat{X}) \leq H(\hat{X})$

Review of Lecture 7

Computing $R(D)$

- $R(D) = \min_{p_{\hat{X}|X}:\mathbf{E}[d(X,\hat{X})]\leq D} I(X;\hat{X})$
- Since $I(X;\hat{X})$ is convex in $p_{\hat{X}|X}$ for fixed p_X and $\mathbf{E}[d(X,\hat{X})]$ is an affine function of $p_{\hat{X}|X}$, this is a convex optimization problem

Example

- $X \sim$Bern($\frac{1}{2}$) and Hamming distance $d(x,\hat{x}) = \mathbf{1}\{x \neq \hat{x}\}$
- Assume $\mathbf{E}[d(X,\hat{X})] = \mathbf{P}(X \neq \hat{X}) = \epsilon \leq D$
- Fano's ineq
- $I(X;\hat{X}) = H(X) - H(X|\hat{X}) \geq 1 - H_b(\epsilon)$
- If $D \leq 1/2$, we have $I(X;\hat{X}) \geq 1 - H_b(D)$ attained when $p_{\hat{X}|X}$ is BSC(D).
- If $D > 1/2$, this distortion is attained by any $\hat{X}$ indpt of X
- Hence $R(D) = 1 - H_b(D)$ if $D \leq 1/2$, $R(D) = 0$ if $D > 1/2$.

Review of Lecture 8

Lecture 8

- Conditionally typical sequence
- Conditional typicality lemma
- Covering lemma
- Lossy source coding
    1. Achievability
    2. Converse

Concluding Remarks

Why we study information theory?

- It provides theoretic guarantees of many practical problems.
  1. Beyond the digital communication, information theory finds its way to biology, computation and complexity, and machine learning.
- It is beautiful: intersection of math, engineering and science.
- Information Theory is the art of telling you how much can you possibly do.
- We are asking how information can be reinforced in a complex setting, ultimately giving us principles for better technology and greater understanding.

Remarks

- Review all the homeworks carefully!
- No cheating, and good luck!