Mid Review

Hongjie Chen

June 4, 2018

Review Outline

- 1 Introduction
  - Stirling approximation
- 2 Entropy and Mutual Information
  -
- 3 Error Correcting
  -
- 4 Lossy Source Coding
  - can explain typical set
  - Chebyshev's inequality 1, 2 and week law of large numbers
- 5 Symbol Codes
  - kraft inequality
  - Source coding theorem for symbol codes
  - Huffman coding
- 6 Noisy Channel Coding
  - def of channel capacity
  - computation of typical channel capacity (such as BSC, BEC, Z).
  - def of DMC.
  - def of Gaussian Channel, derivation of capacity of Gaussian Channel

Homework

HW 1

1. The normal distribution maximizes the entropy for a given variance a

   Proof. $D(f \parallel g) \geq 0,\ g \sim \mathcal{N}(0, a)$ □

2. **entropy of a sum**: Let X and Y be two discrete r.v., and $Z = X + Y$.

   (a) $H(Z|X) = H(Y|X)$ and $H(Z|Y) = H(X|Y)$.

   (b) If $X \perp Y$, then $H(Z) \geq H(X)$ and $H(Z) \geq H(Y)$.

   (c) $H(X+Y) = H(X)+H(Y)$ if and only if $Z = X+Y$ is an one-to-one function of (X, Y) and $X \perp Y$.

HW 2

1. **weighing problem**

HW 3

1. For a symmetric channel with any number of inputs, the uniform distribution over the inputs is an optimal input distribution.

2. Optimal distribution of Z channel.

3. All optimal input distributions of a channel have the same output probability distribution $P(y) = \sum_x P(x)Q(y|x)$

Lecture 1 - introduction

Stirling approximation

$$
\begin{aligned}
n! &\approx \sqrt{2\pi n}(\frac{n}{e})^n \\
\log_2 \binom{N}{r} &\approx NH_2(\frac{r}{N}) + \frac{1}{2}\log_2 \frac{N}{2\pi(N-r)r} \\
\binom{N}{r} &\approx 2^{NH_2(r/N)}
\end{aligned}
$$

Lecture 2 - Entropy and Mutual Information

Support Set

$$supp(X) := \{x | p(x) > 0, x \in \mathfrak{X}\}$$

Entropy

$$H(X) := -\sum_{x\in supp(X)} p(x)\log x$$

Joint Entropy

$$H(X,Y) := -\sum_{x,y} p(x,y)\log p(x,y)$$

Conditional Entropy

$$H(Y|X) := -\sum_{x,y} p(x,y)\log p(y|x)$$

Mutual Information

$$I(X;Y) := D\big(p(x,y) \parallel p(x)p(y)\big) = \sum_{x,y} p(x,y)\frac{p(x,y)}{p(x)p(y)}$$

Conditional Mutual Information

$$I(X;Y|Z) := H(X|Z) - H(X|Y,Z) = -\sum_{x,y,z} p(x,y,z)\log\frac{p(x,y,z)}{p(x|z)p(y|z)}$$

KL Distance or **Relative Entropy**

$$D(p \parallel q) := \sum_{x} p(x)\log\frac{p(x)}{q(x)}$$

- $D(p \parallel q) \geq 0$

  Proof. $-\log(x)$ is convex, by Jensen's inequality:

  $$D(p \parallel q) = \sum_{x} p(x)(-\log\frac{q(x)}{p(x)}) \geq -\log\sum_{x} p(x)\frac{q(x)}{p(x)} = 0$$

  □

- $D(p \parallel q)$ is convex w.r.t. (p,q)
- $H(X)$ is a concave function of $p(x)$

  Proof. $H(X) = \log|\mathfrak{X}| - D(p(x) \parallel u(x))$, where $u(x) = \frac{1}{|\mathfrak{X}|}$ is uniform distribution. □

Pinkser's Inequality

$$d(p,q) := \sum_{x\in\mathfrak{X}} |p(x) - q(x)|$$

$$D(p \parallel q) \geq \frac{1}{2\ln 2}d^2(p,q)$$

Chain Rule

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

  Proof. $I(X_1, ..., X_n; Y) = H(X_1, ..., X_n) - H(X_1, ..., X_n|Y)$ □

- $I(X_1, ..., X_n; Y|Z) = \sum_{i=1}^{n} I(X_i; Y|X_1, ..., X_{i-1}, Z)$
- $D(p(x, y) \parallel q(x, y)) = D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x))$

  Proof. $D(p(x, y) \parallel q(x, y)) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{q(x,y)} = \sum_{x,y} p(x, y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)}$ □

Data Process Inequality

If $X \to Y \to Z$, then $I(X; Y) \geq I(X; Z)$ and similarly $I(Y; Z) \geq I(X; Z)$.

Proof.

$I(X; Y) - I(X; Z) = H(X|Z) - H(X|Y) = I(X; Y|Z) - I(X; Z|Y) = I(X; Y|Z) \geq 0$

□

- For any function $g, I(X; Y) \geq I(X; g(Y))$.
- If $X \to Y \to Z$, then $I(X; Y|Z) \leq I(X; Y)$.

Some Information Inequalities

- $I(X; Y|Z) \geq 0$ with equality if and only if X and Y are conditionally independent given Z.
- $I(X; Y) = 0 \iff$ X and Y are independent.
- $H(Y|X) = 0 \iff$ Y is a function of X.
- (Conditioning reduces entropy) $H(X|Y) \leq H(X)$
- (Independence bound on entropy) $H(X_1, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$
- (Caveat) conditioning may increase or decrease mutual information

Differential Entropy

$$h(x) := -\int_x f(x) \log f(x) dx$$

- $X \sim U[0, a] : h(X) = \log a$
- $X \sim N(0, \sigma^2) : h(X) = \frac{1}{2} \ln^{2\pi e \sigma^2}$

Lecture 3 - Error Correcting

$$s \to t \to r \to \hat{s}$$

Repetition Code

Optimal Decoding

The optimal decoding decision is to find which value of s is most probable, given r, i.e. $\max_s P(s|r)$.
If $f < 0.5$, majority vote is optimal.

Proof.

$$P(r|s) = P(r|t(s)) = \prod_{n=1}^{N} P(r_n|t_n(s))$$

$$P(r_n|t_n) = \begin{cases} 1-f & r_n = t_n \\ f & r_n \neq t_n \end{cases}$$

$$\frac{P(r|s=1)}{P(r|s=0)} = \prod_{n=1}^{N} \frac{P(r_n|t_n(1))}{P(r_n|t_n(0))}$$

□

Hamming Code

Block Code

A block code is a rule for converting a sequence of source bits s, of length K, into a transmitted sequence of length N bits. The extra $N-K$ bits are linear functions of the original K bits, called parity-check bits.

(7,4) Hamming code

- 1 bit is flipped ⇒ Syndrome decoding: find a unique bit that lies inside all the unhappy circles and outside all the happy circles. Flip this bit for correction.
- more than 1 bits are flipped

Decomposition of Entropy

$$\begin{aligned} H(P) =& H(p_1, ..., p_n) \\ & \text{define } p := \sum_{i=1}^{m} p_i \\ =& H_2(p) + pH(\frac{p_1}{p}, ..., \frac{p_m}{p}) + (1-p)H(\frac{p_{m+1}}{1-p}, ..., \frac{p_n}{1-p}) \end{aligned}$$

Lecture 4 - Lossy Source Coding

Ensemble and Shannon Information Content

ensemble

$$(x; A_X; P_X)$$

x is the value of a r.v. $A_X = \{a_1, ..., a_N\}, P_X = \{p_1, ..., p_N\}$, with $P(x = a_i) = p_i$ and $\sum_{i=1}^{N} p_i = 1$.

Shannon Information Content

$$h(x = a_i) = \log_2 \frac{1}{p_i}$$

Lossy Compression

raw bit content

$$H_0(X) = \log_2 |A_X|$$

smallest δ-sufficient subset S_δ

$$P(x \in S_\delta) \geq 1 - \delta \iff P(x \notin S_\delta) \leq \delta$$

essential bit content

$$H_\delta(X) = \log_2 |S_\delta|$$

Shannon's source coding theorem

Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 s.t. for $N > N_0$,

$$|\frac{1}{N} H_\delta(X^N) - H| < \epsilon$$

Typicality

By law of large numbers, the probability of a typical string $x \in A_X^N$ is

$$P(x) = \prod_{i=1}^{N} P(x_i) \approx p_1^{Np_1} p_2^{Np_2} \cdots p_I^{Np_I}$$

Information content of a typical string

$$\log_2 \frac{1}{P(x)} \approx NH$$

Typical Set

$$T_{N\beta} := \left\{ x \in A_X^N \;\middle|\; |\frac{1}{N} \log_2 \frac{1}{P(x)} - H| < \beta \right\}$$

$$T_{N\beta} := \left\{ x \in A_X^N \;\middle|\; 2^{-N(H+\beta)} < P(x) < 2^{-N(H-\beta)} \right\}$$

Asymptotic equipartition principle (AEP)
For an ensemble of N i.i.d. random variables $X^N \equiv (X_1, ..., X_N)$, with N sufficiently large, the outcome $x = (x_1, ..., x_N)$ is almost certain to belong to a subset of A_X^N having only $2^{NH(X)}$ members, each having probability close to $2^{-NH(X)}$.

Proof. For $x = (x_1, ..., x_N)$ drawn from A_X^N, $\frac{1}{N}\log_2 \frac{1}{P(x)} = \frac{1}{N}\sum_{n=1}^{N} \log_2 \frac{1}{P(x_n)}$. $\{\log_2 \frac{1}{P(x_n)}\}$ can be viewed as a set of IID r.v. where x_n is drawn from A_X, and $\mathbb{E}[\log_2 \frac{1}{P(x_n)}] = H(X), \sigma^2 = var[\log_2 \frac{1}{P(x_n)}]$. Then $\frac{1}{N}\log_2 \frac{1}{P(x)}$ can be regared as their mean. By weak law of large numbers, for $\forall \beta > 0$

$$\begin{aligned} & P\Big((\frac{1}{N}\log_2 \frac{1}{P(x)} - H)^2 > \beta^2 \Big) \leq \frac{\sigma^2}{N\beta^2} \\ \iff & P(x \in T_{N\beta}) \geq 1 - \frac{\sigma^2}{N\beta^2} \\ \Rightarrow & \lim_{N\to\infty} P(x \in T_{N\beta}) = 1 \end{aligned}$$

□

Markov Inequality
Let X be a non-negative r.v. and $\mathbb{E}[X]$ exists. For $\forall t > 0$,

$$P(X > t) \leq \frac{\mathbb{E}[X]}{t}$$

Proof.

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty xf(x)dx = \int_0^t + \int_t^\infty \\ &\geq \int_t^\infty xf(x)dx \geq t\int_t^\infty f(x)dx = tP(X > t) \end{aligned}$$

□

Chebyshev's Inequality Let $\mu = \mathbb{E}$ and $\sigma^2 = var[X]$. Then $\forall t > 0$,

$$P(|X-\mu| > t) \leq \frac{\sigma^2}{t^2} \text{ or } P((X-\mu)^2 > t^2) \leq \frac{\sigma^2}{t^2}$$

Proof.

$$\begin{aligned} P(|X-\mu| > t) &= P((X-\mu)^2 > t^2) \\ &\text{By Markov Inequlity} \\ &\leq \frac{\mathbb{E}[(X-\mu)^2]}{t^2} \\ &= \frac{\sigma^2}{t^2} \end{aligned}$$

□

Weak Law of Large Numbers
If $X_1, ..., X_N$ are IID, then $\bar{X} := \frac{1}{N}\sum_{i=1}^{N} X_i \xrightarrow{P} \mu$

Proof. Suppose $\mathbb{E}[X_i] = \mu, var[X_i] = \sigma^2$. Then,

$$\mathbb{E}[\bar{X}] = \mu, var[\bar{X}] = \frac{1}{N}\sigma^2$$

By Chebyshev's inequality, $\forall \epsilon > 0, P(|\bar{X} - \mu| > \epsilon) \leq \frac{\sigma^2}{N\epsilon^2}$. Therefore,

$$\forall \epsilon > 0, \lim_{N \to \infty} P(|\bar{X} - \mu| > \epsilon) = 0$$

□

Lecture 5 - Symbol Codes

Symbol Code

(binary) symbol code
A (binary) symbol code C for an ensemble X is a mapping from the range of x, $A_X = \{a_1, ..., a_I\}$ to $\{0,1\}^+$. $c(x)$ will denote the codeword corresponding to x, and $l(x)$ will denote its length, with $l_i = l(a_i)$.
Unique decoding
A code $C(X)$ is uniquely decodeable if, under the extended code C^+ , no two distinct strings have the same encoding, i.e., $\forall x, y \in A_X^+,\ x \neq y \Rightarrow c^+(x) \neq c^+(y)$
Prefix code
A symbol code is called a prefix code if no codeword is a prefix of any other codeword.
Expected length $L(C, X)$
The expected length $L(C, X)$ of a symbol code C for ensemble X is

$$L(C, X) = \sum_{x \in A_X} P(x) l(x) = \sum_{i=1}^{I} p_i l_i$$

Kraft Inequality

For any uniquely decodeable code $C(X)$ over the binary alphabet $\{0,1\}$, the codeword lengths must satisfy:

$$\sum_{i=1}^{I} 2^{-l_i} \leq 1$$

Conversely, given a set of codeword lengths $\{l_1, ..., l_I\}$ satisfying Kraft inequality, we can always construct a prefix code.

Proof. Define $S = \sum_{i=1}^{I} 2^{-l_i}$. Then

$$S^N = [\sum_{i=1}^{I} 2^{-l_i}]^N = \sum_{i_1=1}^{I} \cdots \sum_{i_N=1}^{I} 2^{-(l_{i_1}+\cdots+l_{i_N})}$$

Note $(l_{i_1} + \cdots + l_{i_N})$ is the length of the encoding of a string $x = a_{i_1} \cdots a_{i_N}$. Define A_l as the number of strings x with $l(x) = l$, $l_{min} = \min_i l_i$, $l_{max} = \max_i l_i$.Then,

$$S^N = \sum_{Nl_{min}}^{Nl_{max}} 2^{-l} A_l$$

There are 2^l distinct bit strings of length l. Since C is uniquely decodeable, we have $A_l \leq 2^l$. Then,

$$S^N = \sum_{Nl_{min}}^{Nl_{max}} 2^{-l} A_l \leq \sum_{Nl_{min}}^{Nl_{max}} 1 \leq Nl_{max}$$
$$S \leq (Nl_{max})^{1/N}$$

Since the above inequality holds true for any positive integer N, it is true as $N \to \infty$. Since $\lim_{N\to\infty} (Nl_{max})^{1/N} = 1$, thus

$$\sum_{i=1}^{I} 2^{-l_i} \leq 1$$

Label the first node (lexicographically) of depth l_1 as codeword 1, and remove its descendants from the tree. Then label the first remaining node of depth l_2 as codeword 2, and so on. Proceeding this way, we construct a prefix code with the specified $l_1, l_2, ..., l_m$. $\square$

Source Coding Theorem for Symbol Codes

For an ensemble X, there always exists a prefix code C with the expected length satisfying

$$H(X) \leq L(C, X) < H(X) + 1$$

Proof.

- $H(X) \leq L(C, X)$

Define $z = \sum_i 2^{-l_i},\ q_i := \frac{2^{-l_i}}{z}$

$$\begin{aligned}
L(C,X) &= \sum_i p_i l_i \\
&= \sum_i p_i \log \frac{1}{q_i} - \log z \\
&= D_{KL}(p \parallel q) + \sum_i p_i \log \frac{1}{p_i} - \log z \\
&= H(X) + D_{KL}(p \parallel q) - \log z \\
&\text{Kraft inequality: } z \le 1 \text{ and } D_{KL}(p \parallel q) \ge 0 \\
&\ge H(X)
\end{aligned}$$

Optimal source code-lengths:

$$l_i = \log_2 \frac{1}{p_i}$$

- $L(C,X) < H(X) + 1$
  We set l_i a little bit larger than optimal length, i.e. $l_i = \lceil \log_2 \frac{1}{p_i} \rceil$. Then

$$L(C,X) = \sum_i p_i \lceil \log_2 \frac{1}{p_i} \rceil < \sum_i p_i (\log_2 \frac{1}{p_i} + 1) = H(X) + 1$$

□

Huffman Coding: optimal source coding with symbol codes

Lecture 6 - Noisy Channel Coding

DMC

A **discrete memoryless channel** Q is characterized by an input alphabet A_X, an output alphabet A_Y , and a set of conditional probability distributions $P(y|x)$, one for each $x \in A_X$. These transition probabilities may be written in a matrix

$$Q_{j|i} = P(y = b_j | x = a_i)$$

Definition 1 (symmetric DMC)**.** *A DMC is defined to be **symmetric**, if the set of outputs can be partitioned into subsets in such a way that for each subset the matrix of transition probability has the property that each row is a permutation of each other row and each column is a permutation of each other column.*

Useful model channels

- BSC

$$C(BSC) = 1 - H_2(f), \quad P_X^* = \{0.5, 0.5\}$$

Proof.

$$
\begin{aligned}
C(BSC) &= \max_{P_X} H(Y) - H(Y|X) \\
&= \max_{P_X} H(X+Z) - H(X+Z|X) \\
&= \max_{P_X} H(X+Z) - H(Z) \\
&= 1 - H_2(f)
\end{aligned}
$$

□

- BEC

$$C(BEC) = 1 - f, \quad P_X^* = \{0.5, 0.5\}$$

Proof.

$$C(BEC) = \max_{P_X} H(X) - H(X|Y) = \max_{P_X} H(X) - fH(X) = 1 - f$$

□

- Z Channel

$$p_1^* = \frac{1/(1-f)}{1 + 2^{H_2(f)/(1-f)}}$$

- Noisy typewriter

Lemma 1. *For a symmetric channel with any number of inputs, the uniform distribution over the inputs is an optimal input distribution.*

Channel coding theorem

The capacity of a channel Q is

$$C(Q) := \max_{P_X} I(X;Y)$$

The distribution P_X that achieves the maximum is called the optimal input distribution, denoted by P_X^* . [There may be multiple optimal input distributions.]

Gaussian channel

The Gaussian channel has a real input x and a real output y. The conditional distribution of y given x is a Gaussian distribution:

$$P(y|x) \sim N((y-x)|0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y-x)^2}{2\sigma^2}\}$$

For a Gaussian channel with power constraint v and variance of noise σ^2, its channel capacity is

$$C = \max_{var(X) \leq \sigma^2} I(X; X+Z) = \frac{1}{2}\log(1 + \frac{v}{\sigma^2})$$

Proof.

$$\begin{aligned}
\max_{var(X)\leq\sigma^2} I(X;X+Z) &= \max_{var(X)\leq\sigma^2} h(X+Z) - h(X+Z|X) \\
&= \max_{var(X)\leq\sigma^2} h(X+Z) - h(Z) \\
&\text{the above is maximized when } X+Z \sim \mathcal{N}(*, v+\sigma^2) \\
&= \frac{1}{2}\log(1+\frac{v}{\sigma^2})
\end{aligned}$$

□