MTHE 474 Notes

Timothy Liu

Fall 2022

Contents

1 Chapter 2 Topics

1.1 Information Measures for Discrete Systems

1.1.1 Definitions

- **Definition 2.2:** Entropy of discrete random variable X with pmf $P_X(*)$ is defined as

$$H(X) := -\sum_{x\in X} P_X(x) * \log_2 P_X(x)$$

- **Definition 2.2:** Entropy of discrete random variable X with pmf $P_X(*)$ is defined as

$$H(X) := -\sum_{x\in X} P_X(x) * \log_2 P_X(x)$$

- **Definition 2.8 (Joint entropy):**

$$H(X,Y) := -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y) * \log_2 P_{(X,Y)}(x,y)$$

- **Definition 2.9 (Conditional entropy):**

$$H(Y|X) := \sum_{x\in\mathcal{X}} P_X(x)(-\sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) * \log_2 P_{Y|X}(y|x))$$

-

1.1.2 Lemmas/Theorems

- **Lemma 2.4 (Fundamental Inequality):** $\forall\ x > 0$ and $D > 1$ we have

$$\log_D(x) \leq \log_D e * (x-1)$$

- **Lemma 2.5 (Non-negativity):** $H(X) \geq 0$
- **Lemma 2.6 (Entropy Upper-Bound):** $H(X) \leq \log_2 |\mathcal{X}|$ where random variable X takes values from finite set $\mathcal{X}$
- **Lemma 2.7 (Log-Sum inequality):** For nonnegative numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$

$$\sum_{i=1}^{n}(a_i \log_D \frac{a_i}{b_i}) \leq (\sum_{i=1}^{n} a_i) \log_D \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

  with equality iff for all $i = 1, \ldots, n$

$$\frac{a_i}{b_i} = \frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j}$$

  is constand and does not depend on i

- **Theorem 2.10 (Chain rule for entropy):** $H(X,Y) = H(X) + H(Y|X)$
- **Theorem 2.12 (Conditioning never increases entropy):** $H(X|Y) \leq H(X)$ with equality holding iff X and Y are independent
- **Lemma 2.13 (Entropy is additive for independent RVs):** For independent X, Y

$$H(X,Y) = H(X) + H(Y)$$

- **Lemma 2.14 (Conditional entropy is lower additive):** $H(X_1, X_2|Y_1, Y_2) \leq H(X_1|Y_1) + H(X_2|Y_2)$ with equality holding iff

$$P_{X_1,X_2|Y_1,Y_2}(x_1, x_2|y_1, y_2) = P_{X_1|Y_1}(x_1|y_1)P_{X_2|Y_2}(x_2|y_2)$$

  for all x_1, x_2, y_1, y_2

1.2 Mutual Information

1.2.1 Definitions

- **Definition 2.2.1 (Mutual Information):**

$$I(X;Y) := H(X) - H(X|Y)$$

- **Definition 2.2.2 (Conditional Mutual Information):**

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

1.2.2 Lemmas

- **Lemma 2.15 (Properties of Mutual Information):**

$$1.\ I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log_2\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \tag{1}$$

$$2.\ I(X;Y) = I(Y;X) = H(Y) - H(Y|X) \tag{2}$$

$$3.\ I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3}$$

$$4.\ I(X;Y) \leq H(X)\ \ \text{equality iff X is a function of Y} \tag{4}$$

$$5.\ I(X;Y) \leq 0 \text{ with equality iff X and Y are independent} \tag{5}$$

$$6.\ I(X;Y) \leq \min\{\log_2|\mathcal{X}|, \log_2|\mathcal{Y}|\} \tag{6}$$

- **Lemma 2.16 (Chain Rule for Mutual Information):**

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

- **Theorem 2.17 (Chain Rule for entropy):** $X^n := (X_1,\ldots,X_n)$ and $x^n := (x_1,\ldots,x_n)$

$$H(X^n) = \sum_{i=1}^{n} H(X_i|X^{i-1})$$

- **Theorem 2.18 (Chain Rule for conditional entropy):**

$$H(X_1,X_2,\ldots,X_n|Y) = \sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1,Y)$$

- **Theorem 2.19 (Chain Rule for Mutual information):**

$$I(X_1,X_2,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_{i-1},\ldots,X_1)$$

Where $I(X_i;Y|X_{i-1},\ldots,X_1) := I(X_1,Y)$ for $i=1$

1.3 Conditional Divergence

1.3.1 Definitions

- **Definition 2.29 (Divergence):** Given 2 discrete random variables X and $\hat{x}$ defined over common alphabet $\mathcal{X}$ divergence is defined by,

$$D(X||\hat{X}) := E_x[\log_2\frac{P_X(X)}{P_{\hat{X}}(X)}] = \sum_{x\in\mathcal{X}} P_X(x)\log_2\frac{P_X(x)}{P_{\hat{X}}(x)}$$

1.3.2 Theorems

- **Lemma 2.30 (Nonnegativity of Divergence):** $D(X||\hat{X}) \geq 0$, with equality iff $P_X(x) = P_{\hat{X}}(x)$ for all $x \in \mathcal{X}$

1.4 Fano's Inequality

1.4.1 Definitions

1.4.2 Theorems/Lemmas

- **Lemma 2.6 (Fano's inequality):** Let X and Y be two random variables with alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively ($\mathcal{X}$ is finite but $\mathcal{Y}$ can be countably infinite). Let $\hat{X} := g(Y)$ represent the estimate of X by observing Y and $P_e := \Pr[\hat{X} \neq X]$ represent the probability of error of this observation
  Then the following holds
  $$H(X|Y) \leq h_b(P_e) + P_e * \log_2(\mid \mathcal{X} \mid -1)$$
  Where $h_b(P_e)$ is the binary entropy with probability P_e

1.5 Data Processing Inequality

1.5.1 Definitions

- **Lecture 7 Definition (Markov Chain):** Three jointly distributed random variables X, Y, Z are said to form a Markov Chain (in that order), denoted by $X \to Y \to Z$
  if:
  $$P_{XZ|Y}(x, y|z) = P_{X|Y}(x|y)P_{Z|Y}(z, y) \iff P_{Z|XY}(z|x, y) = P_{Z|Y}(z|y)$$

  $\forall x \in X, y \in Y, z \in Z$

  - The probability of each event ONLY depends on the state attained on the previous event

1.5.2 Theorems

- **Lecture 7 Theorem (Data Processing Inequality):** If $X \to Y \to Z$, then
  $$I(X;Y) \leq I(X;Z)$$
  - Another way to think of this is that the futher the RVs are along the markov chain, the less relevant the RVs are with eachother and the less information we get
- **Lecture 8 Theorem (DPI for Divergence):** Given fixed conditional PMF $P_{Y|X}$ on $y \times x$, which describes a channel with input x and output y, let P_x and q_x be 2 possible PMFs for input x with corresponding output PMFs P_y and q_y respectively, then
  $$D(P_x||q_x) \leq D(P_y||q_y)$$

1.6 Convex/Concavity of Information Measures

1.6.1 Definitions

- **Lecture 6 Definition (Convex Set):**

  A subset K of $\mathbb{R}$ is called convex if the line segment joining any two points in K also lies in K

- **Lecture 6 Definition (Convex Function):** The function $f : k \to \mathbb{R}$ where k is a convex subset of $\mathbb{R}^n$, is called convex on k if $\forall x_1, x_2 \in k$ and $\lambda \in [0, 1]$,
  $$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$
  Strict equality holds whenever $x_1 \neq x_2$ and $0 < \lambda < 1$ then f is called strictly convex
- **Lecture 6 Definition (Concave Function):** $f : k \to \mathbb{R}$ is concave on k (where $k \subseteq \mathbb{R}^n$ is a concave subset) if $-f$ is convex. In other words: if $\forall x_1, x_2 \in k$ and $\lambda \in [0, 1]$,
  $$f(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

1.6.2 Theorems

- **Lecture 6 Theorem (Jensen's Inequality):** Let $K \subseteq \mathbb{R}$ (where K is a convex set?) and let $f : k \to \mathbb{R}$ be a convex function. Also let x be a RV with alphabet $\mathcal{X} \subseteq k$ and finite mean, then

$$E[f(x)] \leq f(E[x])$$

Also if f is strictly convex, then the inequality is strict unless x is deterministic

- **Lecture 7 Theorem (Convexity/Concavity of Information Measures):**
i. $D(p||q)$ is convex in the pair (p, q) (ie: if p_1, q_1 and p_2, q_2 are two pairs of PMFs defined on $\mathcal{X}$) then:

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

$\forall \lambda \in [0, 1]$

ii. if $x \sim P_x$, then

$$H(x) = H(p_x) \text{ is concave in } P_x$$

iii. If (x, y) $P_X P_{Y|X}$, then $I(X;Y) = I(P_X, P_{Y|X})$ is concave in P_X for fixed $P_{Y|X}$ and convex in $P_{Y|X}$ for fixed P_X

2 Chapter 3 Topics

2.1 Principles of Data Compression (Week 4)

2.1.1 Definitions

- **Lecture 9 Definition (Discrete Memoryless Source):** A DMS is an infinite sequence of i.d.d random variables $\{x_i\}_{i=1}^{\infty} = \{x_1, x_2, \ldots\}$, such that all the random variables have a common PMF P_x defined on the alphabet/finite set $\mathcal{X}$
i.d.d property: $P(X_1 = a_1, \ldots, X_n = a_n) = \prod_{i=1}^{n} P(X_i = a_i)$

- **Lecture 9 Definition (Convergence in Probability):** Given sequence $\{x_i\}_{i=1}^{\infty}$ of RVs and RV Z,

$$X_n \xrightarrow{n\to\infty} \text{ in probability} \iff \forall \epsilon > 0, \lim_{n\to\infty} P(|X_n - Z| > \epsilon) = 0$$

- **Lecture 9 Definition (Typical Set):** For a DMS $\{x_i\}_{i=1}^{\infty}$ with PMF P_x and entropy $H(X)$, given integer $n \geq 1$ and $\epsilon > 0$, the typical set $A_\epsilon^{(n)}$ with respect to the source is

$$A_\epsilon^{(n)} = \{a^n \in \mathcal{X} : | -\frac{1}{n}\log_2 P_{X^n}(a^n) - H(X)| \leq \epsilon\}$$

- **Lecture 9 Definition (Code block):** Given integers $D \geq 2$, $n \geq 1$ and $k = k(n)$ a (k, n) D-ary Fixed length code φ for a DMS $\{x_i\}_{i=1}^{\infty}$ with alphabet $\mathcal{X}$ consists of the following pair of encoding and decoding functions

$$\text{Encoding: } f : \mathcal{X}^n \to \{0, 1, \ldots, D-1\}^k$$

$$\text{Decoding: } g : \{0, 1, \ldots, D-1\}^k \to \mathcal{X}$$

The range of f is called the *codebook*
The code (or Compression) rate is defined as $R = \frac{k}{n}$ in D-ary code symbols / Source symbols

- **Lecture 9 Definition (Probability of Decoding Error):** Measures the code's reliability and defined as

$$P_e := P(g(f(x^n))) \neq x^n$$

Predicament is that we want code to be efficient and reliable (ie code rate as small as possible and probability of error is also as small as possible)

- **Lecture 9 Definition (Lossless):** A (k, n) D-ary code for the source is called uniquely decodable or lessless if

$$f : \mathcal{X}^n \to \{0, 1, \ldots, D-1\}$$

is an invertable map and $g = f^{-1}$

2.1.2 Theorems/Lemmas

- **Lecture 9 Theorem (Weak Law of Large Numbers):** if $\{x_i\}_{i=i}^{\infty}$ is a DMS then

$$\frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{n\to\infty} E[X]$$

  in probability

- **Lecture 9 Theorem (Asymptotic Equipartition Property):** (also known as "entropy stability property") For a DMS $\{x_i\}_{i=1}^{\infty}$ with PMF P_x and alphabet $\mathcal{X}$,

$$-\frac{1}{n}\log_2 P_{X^n}(x^n) \xrightarrow{n\to\infty} H(X) \text{ in probability}$$

- **Lecture 9 Theorem (Consequence of AEP):** For a DMS $\{x_i\}_{i=1}^{\infty}$ with PMF P_x and entropy $H(X)$ the typical set satisfies
  - $\lim_{n\to\infty} P(A_\epsilon^{(n)}) = 1$
  - $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$ Where $|A|$ is the size of set A
  - $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ for n sufficiently large

- **Lecture 10 Theorem (Shannon's Fixed-length lossless source coding theorem for DMS):** For integer $D \leq 2$, consider a DMS $\{x_i\}_{i=i}^{\infty}$ with alphabet $\mathcal{X}$, PMF P_x and source entropy $H_D(X) = -\sum_{a\in\mathcal{X}} P_X(a)\log_D P_X(a)$ then the following hold:
  - (i. forward part) $\forall\epsilon \in (0,1)$ and $0 < \delta < \epsilon$, $\exists$ a sequences of D-ary (k,n) fixed length codes φ_n such that,

$$\limsup_{n\to\infty} \frac{k}{n} \leq H_d(x) + \delta$$

and

$$P_e(\varphi_n) < \epsilon \text{ for n sufficiently large}$$

  - (ii. strong converse part) $\forall\epsilon \in (0,1)$ and any sequence of D-ary (k,n) fixed-length codes φ_n for the source
    with $\limsup_{n\to\infty} \frac{k}{n} < H_D(X)$, we have

$$P_e(\varphi_n) > 1-\epsilon \text{ for n sufficiently large}$$

  Consequence from this theorem:

$$H_D(x) = \inf\{R : R achieveable\}$$

  where

$$R\ achieveable \iff \forall\epsilon > 0, \exists \text{ D-ary (k,n) fixed length codes } \varphi_n \text{ such that } \limsup_{n\to\infty}\frac{k}{n} \leq R$$

3 Tutorial Proofs

3.1 Week 2 Tutorial

- Given 2 discrete RVs, X, Y we have that

$$H(Y|X) = 0 \iff \text{Y is a function of X}$$

- Given RV X with alphabet $\mathcal{X}$ and function $f : x \to \mathbb{R}$

$$H(X) \leq H(f(X))$$

3.2 Week 3 Tutorial