Inequalities in Information Theory

A Brief Introduction

Xu Chen

Department of Information Science
School of Mathematics
Peking University

Mar.20, 2012

Part I

Basic Concepts and Inequalities

Outline

The Entropy

- Definition
  1. The Shannon information content of an outcome x is defined to be

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

  2. The entropy of an ensemble X is defined to be the average Shannon information content of an outcome:

$$H(X) = \sum_{x \in \mathcal{X}} P(X) \log_2 \frac{1}{P(X)} \quad (1)$$

  3. Conditional Entropy: the entropy of a r.v.,given another r.v.

$$H(X|Y) = -\sum_i \sum_j p(x_i, y_j) \log_2 p(x_i|y_j) \quad (2)$$

The Entropy

- Definition
  1. The Shannon information content of an outcome x is defined to be

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

  2. The entropy of an ensemble X is defined to be the average Shannon information content of an outcome:

$$H(X) = \sum_{x \in \mathcal{X}} P(X) \log_2 \frac{1}{P(X)} \tag{1}$$

  3. Conditional Entropy: the entropy of a r.v.,given another r.v.

$$H(X|Y) = -\sum_i \sum_j p(x_i, y_j) \log_2 p(x_i|y_j) \tag{2}$$

The Entropy

- Definition
  1. The Shannon information content of an outcome x is defined to be

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

  2. The entropy of an ensemble X is defined to be the average Shannon information content of an outcome:

$$H(X) = \sum_{x \in \mathcal{X}} P(X) \log_2 \frac{1}{P(X)} \quad (1)$$

  3. Conditional Entropy: the entropy of a r.v.,given another r.v.

$$H(X|Y) = -\sum_i \sum_j p(x_i, y_j) \log_2 p(x_i|y_j) \quad (2)$$

The Entropy

The Joint Entropy

The joint entropy of X; Y is:

$$H(X, Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log_2 \frac{1}{p(x, y)} \tag{3}$$

Remarks

- The entropy H answers the question that what is the ultimate data compression.
- The entropy is a measure of the average uncertainty in the random variable.It is the number of bits on the average required to describe the random variable.

- Reference for [[2]Thomas and [4]David]

The Entropy

The Joint Entropy

The joint entropy of X; Y is:

$$H(X,Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2\frac{1}{p(x,y)} \tag{3}$$

Remarks

1. The entropy H answers the question that what is the ultimate data compression.
2. The entropy is a measure of the average uncertainty in the random variable.It is the number of bits on the average required to describe the random variable.

- Reference for [[2]Thomas and [4]David]

The Entropy

The Joint Entropy

The joint entropy of X; Y is:

$$H(X, Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log_2 \frac{1}{p(x, y)} \tag{3}$$

Remarks

1. The entropy H answers the question that what is the ultimate data compression.
2. The entropy is a measure of the average uncertainty in the random variable.It is the number of bits on the average required to describe the random variable.

- Reference for [[2]Thomas and [4]David]

The Entropy

The Joint Entropy

The joint entropy of X; Y is:

$$H(X,Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2 \frac{1}{p(x,y)} \tag{3}$$

Remarks

1. The entropy H answers the question that what is the ultimate data compression.
2. The entropy is a measure of the average uncertainty in the random variable.It is the number of bits on the average required to describe the random variable.

- Reference for [[2]Thomas and [4]David]

The Entropy

The Joint Entropy

The joint entropy of X; Y is:

$$H(X,Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2 \frac{1}{p(x,y)} \tag{3}$$

Remarks

1. The entropy H answers the question that what is the ultimate data compression.
2. The entropy is a measure of the average uncertainty in the random variable.It is the number of bits on the average required to describe the random variable.

- Reference for [[2]Thomas and [4]David]

The Mutual Information

Definition

The mutual information is the reduction in uncertainty when given another r.v., for two r.v. X and Y this reduction is

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

- The capacity of channel is

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

The Mutual Information

Definition

The mutual information is the reduction in uncertainty when given another r.v., for two r.v. X and Y this reduction is

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

- The capacity of channel is

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

The relationships

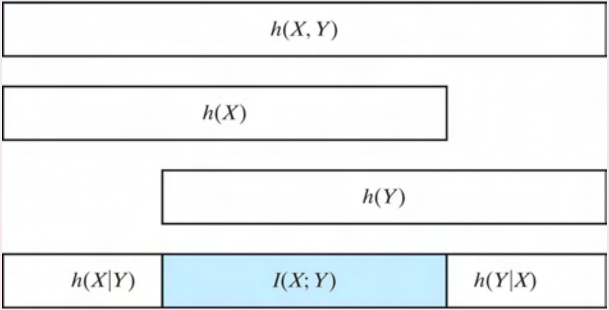


Figure: The relationships between Entropy and Mutual Information

- Graphic from [[3]Simon,2011].

The relative entropy

Definition

The relative entropy or Kullback Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}. \tag{5}$$

1. The relative entropy and mutual information

$$I(X;Y) = D(p(x,y) \parallel p(x)p(y)) \tag{6}$$

2. Pythagorean decomposition: let $X = AU$, then

$$D(p_x \parallel p_u) = D(p_x \parallel \tilde{p}_x) + D(\tilde{p}_x \parallel p_u). \tag{7}$$

The relative entropy

Definition

The relative entropy or Kullback Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}. \tag{5}$$

1. The relative entropy and mutual information

$$I(X; Y) = D(p(x, y) \parallel p(x)p(y)) \tag{6}$$

2. Pythagorean decomposition: let $X = AU$, then

$$D(p_x \parallel p_u) = D(p_x \parallel \tilde{p}_x) + D(\tilde{p}_x \parallel p_u). \tag{7}$$

The relative entropy

Definition

The relative entropy or Kullback Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}. \tag{5}$$

1. The relative entropy and mutual information

$$I(X;Y) = D(p(x,y) \parallel p(x)p(y)) \tag{6}$$

2. Pythagorean decomposition: let $X = AU$, then

$$D(p_x \parallel p_u) = D(p_x \parallel \tilde{p}_x) + D(\tilde{p}_x \parallel p_u). \tag{7}$$

Conditional definitions

Conditional mutual information

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{8}$$

$$= E_{p(x,y,z)} \log \frac{p(X,y|Z)}{p(X|Z)p(Y|Z)}. \tag{9}$$

Conditional relative entropy

$$D(p(y|x) \parallel q(y|x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \tag{10}$$

$$= E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}. \tag{11}$$

Conditional definitions

Conditional mutual information

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{8}$$

$$= E_{p(x,y,z)} \log \frac{p(X,y|Z)}{p(X|Z)p(Y|Z)}. \tag{9}$$

Conditional relative entropy

$$D(p(y|x) \parallel q(y|x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \tag{10}$$

$$= E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}. \tag{11}$$

Differential entropy

Definition 1

The differential entropy $h(X_1, X_2, \ldots, X_n)$, some times written $h(f)$, is defined by

$$h(X_1, X_2, \ldots, X_n) = -\int f(x) \log f(x) dx \tag{12}$$

Definition 2

The relative entropy between probability densities f and g is

$$D(f \parallel g) = -\int f(x) \log(f(x)/g(x)) dx \tag{13}$$

Differential entropy

Definition 1

The differential entropy $h(X_1, X_2, ..., X_n)$, some times written $h(f)$, is defined by

$$h(X_1, X_2, ..., X_n) = -\int f(x) \log f(x) dx \tag{12}$$

Definition 2

The relative entropy between probability densities f and g is

$$D(f \parallel g) = -\int f(x) \log(f(x)/g(x)) dx \tag{13}$$

Chain Rules

1. Chain rule for entropy

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \quad (14)$$

2. Chain rule for information

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1). \quad (15)$$

3. Chain rule for entropy

$$D(p(x, y) \parallel q(x, y)) = D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x)). \quad (16)$$

Chain Rules

1. Chain rule for entropy

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1). \tag{14}$$

2. Chain rule for information

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1). \tag{15}$$

3. Chain rule for entropy

$$D(p(x,y) \parallel q(x,y)) = D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x)). \tag{16}$$

Chain Rules

1. Chain rule for entropy

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1). \tag{14}$$

2. Chain rule for information

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \dots, X_1). \tag{15}$$

3. Chain rule for entropy

$$D(p(x,y) \parallel q(x,y)) = D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x)). \tag{16}$$

Outline

Jensen's inequality

Definition

A function f is said to be convex if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad (17)$$

for all $0 \leq \lambda \leq 1$ and all x_1 and x_2 in the convex domain of f.

Theorem

If f is convex,then

$$f(EX) \leq Ef(x) \quad (18)$$

Proof

We consider discrete distributions only. The proof is given by induction. For a two mass point distribution, by definition. for k mass points, let $p'_i = p_i/(1-p_k)$ for $i \leq k-1$, the result can be derived easily.

Jensen's inequality

Definition

A function f is said to be convex if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \quad (17)$$

for all $0 \leq \lambda \leq 1$ and all x_1 and x_2 in the convex domain of f.

Theorem

If f is convex,then

$$f(EX) \leq Ef(x) \quad (18)$$

Proof

We consider discrete distributions only. The proof is given by induction. For a two mass point distribution, by definition. for k mass points, let $p'_i = p_i/(1-p_k)$ for $i \leq k-1$, the result can be derived easily.

Jensen's inequality

Definition

A function f is said to be convex if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{17}$$

for all $0 \leq \lambda \leq 1$ and all x_1 and x_2 in the convex domain of f.

Theorem

If f is convex,then

$$f(EX) \leq Ef(x) \tag{18}$$

Proof

We consider discrete distributions only. The proof is given by induction. For a two mass point distribution, by definition. for k mass points, let $p_i' = p_i/(1-p_k)$ for $i \leq k-1$, the result can be derived easily.

Log sum inequality

Theorem

For positive numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^{n} a_i) \log(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}) \tag{19}$$

with equality iff $\frac{a_i}{b_i}$ = constant.

Proof

We substitute discrete distribution parameters in Jensen's Inequality by $\alpha_i = b_i / \sum_{j=1}^{n} b_j$ and the variables by $t_i = a_i / b_i$, we obtain the inequality.

Log sum inequality

Theorem

For positive numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right) \tag{19}$$

with equality iff $\frac{a_i}{b_i} =$ constant.

Proof

We substitute discrete distribution parameters in Jensen's Inequality by $\alpha_i = b_i / \sum_{j=1}^{n} b_j$ and the variables by $t_i = a_i / b_i$, we obtain the inequality.

Inequalities in Entropy Theory

- By Jensen's inequality and Log Sum inequality, we can easily prove following basic conclusions:

$$0 \leq H(X) \leq \log \mid \mathcal{X} \mid \tag{20}$$

$$D(p \parallel q) \geq 0 \tag{21}$$

Further more,

$$I(X;Y) \geq 0 \tag{22}$$

- Note:the conditions when the equalities holds.

Inequalities in Entropy Theory

- By Jensen's inequality and Log Sum inequality, we can easily prove following basic conclusions:

$$0 \leq H(X) \leq \log \mid \mathcal{X} \mid \tag{20}$$

$$D(p \parallel q) \geq 0 \tag{21}$$

Further more,

$$I(X;Y) \geq 0 \tag{22}$$

- Note:the conditions when the equalities holds.

Inequalities in Entropy Theory(cont.)

- Conditioning reduces entropy:

$$H(X|Y) \leq H(X)$$

- The chain rule and independence bound on entropy:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \leq \sum_{i=1}^{n} H(X_i) \qquad (23)$$

- Note: the conclusions continue to hold for differential entropy.
- If X and Y are independent, then

$$h(X+Y) \geq h(Y)$$

Inequalities in Entropy Theory(cont.)

- Conditioning reduces entropy:

$$H(X|Y) \leq H(X)$$

- The chain rule and independence bound on entropy:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \leq \sum_{i=1}^{n} H(X_i) \qquad (23)$$

- Note: the conclusions continue to hold for differential entropy.
- If X and Y are independent, then

$$h(X+Y) \geq h(Y)$$

Inequalities in Entropy Theory(cont.)

- Conditioning reduces entropy:

$$H(X|Y) \leq H(X)$$

- The chain rule and independence bound on entropy:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \leq \sum_{i=1}^{n} H(X_i) \tag{23}$$

- Note: the conclusions continue to hold for differential entropy.
- If X and Y are independent, then

$$h(X+Y) \geq h(Y)$$

Inequalities in Entropy Theory(cont.)

- Conditioning reduces entropy:

$$H(X|Y) \leq H(X)$$

- The chain rule and independence bound on entropy:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \leq \sum_{i=1}^{n} H(X_i) \tag{23}$$

- Note: the conclusions continue to hold for differential entropy.
- If X and Y are independent, then

$$h(X+Y) \geq h(Y)$$

Convexity & concavity entropy theory

Theorem

$D(p \parallel q)$ is convex in the pair (p, q),i.e., if (p_1, q_1) and $(p_2, q_2)$are two pairs of probability mass functions, then

$$D(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1 \parallel q_1) + (1-\lambda)D(p_2 \parallel q_2) \tag{24}$$

for all $0 \leq \lambda \leq 1$.

- Apply the log sum inequality to the term on the left hand side of (24).

Convexity & concavity entropy theory

Theorem

$D(p \parallel q)$ is convex in the pair (p, q),i.e., if (p_1, q_1) and $(p_2, q_2)$are two pairs of probability mass functions, then

$$D(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1 \parallel q_1) + (1-\lambda)D(p_2 \parallel q_2) \tag{24}$$

for all $0 \leq \lambda \leq 1$.

- Apply the log sum inequality to the term on the left hand side of (24).

Convexity & concavity in entropy theory(cont.)

Theorem

$H(p)$ is a concave function of p.

- Let u be the uniform distribution on $|\mathcal{X}|$ outcomes, then the concavity of H then follows directly from then convexity of D, since the following equality holds.

$$H(p) = \log|\mathcal{X}| - D(p \parallel u) \tag{25}$$

Convexity & concavity in entropy theory(cont.)

Theorem

$H(p)$ is a concave function of p.

- Let u be the uniform distribution on $|\mathcal{X}|$ outcomes, then the concavity of H then follows directly from then convexity of D, since the following equality holds.

$$H(p) = \log|\mathcal{X}| - D(p \parallel u) \tag{25}$$

Convexity & concavity in entropy theory(cont.)

Theorem

Let$(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(X)$.

- The detailed proof can be found in $[[2] Thomas, section 2.7]$. An alternative proof is given in [1],P51-52.

Convexity & concavity in entropy theory(cont.)

Theorem

Let$(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(X)$.

- The detailed proof can be found in [[2]$Thomas, section2.7$]. An alternative proof is given in [1],P51-52.

Outline

$\mathcal{L}_1$ bound on entropy

Theorem

Let p and q be two probability mass functions on $\mathcal{X}$ such that

$$\| p - q \|_1 = \sum_{x \in \mathcal{X}} | p(x) - q(x) | \leq \frac{1}{2}.$$

Then

$$| H(p) - H(q) | \leq - \| p - q \|_1 \log \frac{\| p - q \|_1}{| \mathcal{X} |}. \tag{26}$$

Proof of $\mathcal{L}_1$ bound on entropy

Proof

Consider the function $f(t) = -t\log t$,it is concave and positive on $[0, 1]$, since $f(0) = f(1) = 0$.

1. Let $0 \le \nu \le \frac{1}{2}$, for any $0 \le t \le 1-\nu$,we have

$$| f(t) - f(t+\nu) | \le \max\{f(\nu), f(1-\nu)\} = -\nu\log\nu. \quad (27)$$

2. Let $r(x) = | p(x) - q(x) |$. Then

$$| H(p) - H(q) | = | \sum_{x\in\mathcal{X}} (-p(x)\log p(x) + q(x)\log q(x) | \quad (28)$$

$$\le \sum_{x\in\mathcal{X}} | (-p(x)\log p(x) + q(x)\log q(x) | \quad (29)$$

Proof of $\mathcal{L}_1$ bound on entropy

Proof

Consider the function $f(t) = -t \log t$,it is concave and positive on $[0, 1]$, since $f(0) = f(1) = 0$.

1. Let $0 \leq \nu \leq \frac{1}{2}$, for any $0 \leq t \leq 1 - \nu$,we have

$$\mid f(t) - f(t+\nu) \mid \leq \max\{f(\nu), f(1-\nu)\} = -\nu \log \nu. \tag{27}$$

2. Let $r(x) = \mid p(x) - q(x) \mid$. Then

$$\mid H(p) - H(q) \mid = \mid \sum_{x \in \mathcal{X}} (-p(x) \log p(x) + q(x) \log q(x) \mid \tag{28}$$

$$\leq \sum_{x \in \mathcal{X}} \mid (-p(x) \log p(x) + q(x) \log q(x) \mid \tag{29}$$

Proof of $\mathcal{L}_1$ bound on entropy

Proof

Consider the function $f(t) = -t \log t$,it is concave and positive on $[0, 1]$, since $f(0) = f(1) = 0$.

1. Let $0 \leq \nu \leq \frac{1}{2}$, for any $0 \leq t \leq 1 - \nu$,we have

$$| f(t) - f(t+\nu) | \leq \max\{f(\nu), f(1-\nu)\} = -\nu \log \nu. \tag{27}$$

2. Let $r(x) = | p(x) - q(x) |$. Then

$$| H(p) - H(q) | = | \sum_{x \in \mathcal{X}} (-p(x) \log p(x) + q(x) \log q(x) | \tag{28}$$

$$\leq \sum_{x \in \mathcal{X}} | (-p(x) \log p(x) + q(x) \log q(x) | \tag{29}$$

Proof of $\mathcal{L}_1$ bound on entropy

Proof(cont.)

By using (27), we have

$$Left \le \sum_{x\in\mathcal{X}} -r(x)\log r(x) \tag{30}$$

$$= \| p-q \|_1 \sum_{x\in\mathcal{X}} -\frac{r(x)}{\| p-q \|_1} \log \frac{r(x)}{\| p-q \|_1} \| p-q \|_1 \tag{31}$$

$$= - \| p-q \|_1 \log \| p-q \|_1 + \| p-q \|_1 H\left(\frac{r(x)}{\| p-q \|_1}\right) \tag{32}$$

$$\le - \| p-q \|_1 \log \| p-q \|_1 + \| p-q \|_1 \log | \mathcal{X} | . \tag{33}$$

The lower bound of relative entropy

Theorem

$$D(P_1 \parallel P_2) \geq \frac{1}{2\ln 2} \parallel P_1 - P_2 \parallel_1^2 . \tag{34}$$

Proof

(1)Binary case. Consider two binary distribution with parameter p and q with $p \leq q$. We will show that

$$p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \geq \frac{4}{2\ln 2}(p-q)^2.$$

Let

$$g(p,q) = p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q} - \frac{4}{2\ln 2}(p-q)^2.$$

The lower bound of relative entropy

Proof(cont.)

Then

$$\frac{\partial g(p,q)}{\partial q} \leq 0$$

since $q(1-q) \leq \frac{1}{4}$ and $q \leq p$. For $q = p$, $g(p,q) = 0$, and hence $g(p,q) \geq 0$ for $q \leq p$, which proves the binary case.

The lower bound of relative entropy

Proof(cont.)

(2)For the general case, for any two distribution P_1 and P_2,let $A=\{x:P_1(x)>P_2(x)\}$. Define $Y=\phi(X)$, the indicator of the set A,and let $\hat{P}_1$ and $\hat{P}_2$ be the distribution of Y. By the data processing inequality([2]Thomas,section 2.8) applied to relative entropy, we have

$$D(P_1 \parallel P_2) \geq D(\hat{P}_1 \parallel \hat{P}_2) \geq \frac{4}{2\ln 2}(P_1(A)-P_2(A))^2 = \frac{1}{2\ln 2}\parallel P_1 - P_2 \parallel_1^2 .$$

Part II

Entropy in Statistics

Outline

Data processing inequality and its corollaries

Data processing inequality

If $X \to Y \to Z$, then

$$I(X;Y) \geq I(X;Z). \quad (35)$$

Corollary

In particular, if $Z = g(Y)$,we have

$$I(X;Y) \geq I(X;g(Y)). \quad (36)$$

Corollary

If $X \to Y \to Z$, then

$$I(X;Y|Z) \geq I(X;Y). \quad (37)$$

Data processing inequality and its corollaries

Data processing inequality

If $X \to Y \to Z$, then

$$I(X;Y) \geq I(X;Z). \tag{35}$$

Corollary

In particular, if $Z = g(Y)$,we have

$$I(X;Y) \geq I(X;g(Y)). \tag{36}$$

Corollary

If $X \to Y \to Z$, then

$$I(X;Y|Z) \geq I(X;Y). \tag{37}$$

Data processing inequality and its corollaries

Data processing inequality

If $X \to Y \to Z$, then

$$I(X;Y) \geq I(X;Z). \tag{35}$$

Corollary

In particular, if $Z = g(Y)$,we have

$$I(X;Y) \geq I(X;g(Y)). \tag{36}$$

Corollary

If $X \to Y \to Z$, then

$$I(X;Y|Z) \geq I(X;Y). \tag{37}$$

Entropy in Markov chain

Theorem

For a *Markov Chain*:

1. Relative entropy $D(\mu_n \parallel \mu'_n)$ decreases with time.
2. Relative entropy $D(\mu_n \parallel \mu)$ between a distribution and the stationary distribution decreases with time.
3. Entropy $H(X_n)$ increases if the stationary distribution is uniform.
4. The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.
5. Shuffles increase entropy.

Entropy in Markov chain

Theorem

For a *Markov Chain*:

1. Relative entropy $D(\mu_n \parallel \mu'_n)$ decreases with time.
2. Relative entropy $D(\mu_n \parallel \mu)$ between a distribution and the stationary distribution decreases with time.
3. Entropy $H(X_n)$ increases if the stationary distribution is uniform.
4. The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.
5. Shuffles increase entropy.

Entropy in Markov chain

Theorem

For a *Markov Chain*:

1. Relative entropy $D(\mu_n \parallel \mu'_n)$ decreases with time.
2. Relative entropy $D(\mu_n \parallel \mu)$ between a distribution and the stationary distribution decreases with time.
3. Entropy $H(X_n)$ increases if the stationary distribution is uniform.
4. The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.
5. Shuffles increase entropy.

Entropy in Markov chain

Theorem

For a *Markov Chain*:

1. Relative entropy $D(\mu_n \parallel \mu'_n)$ decreases with time.
2. Relative entropy $D(\mu_n \parallel \mu)$ between a distribution and the stationary distribution decreases with time.
3. Entropy $H(X_n)$ increases if the stationary distribution is uniform.
4. The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.
5. Shuffles increase entropy.

Entropy in Markov chain

Theorem

For a *Markov Chain*:

1. Relative entropy $D(\mu_n \parallel \mu'_n)$ decreases with time.
2. Relative entropy $D(\mu_n \parallel \mu)$ between a distribution and the stationary distribution decreases with time.
3. Entropy $H(X_n)$ increases if the stationary distribution is uniform.
4. The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.
5. Shuffles increase entropy.

Entropy in Markov chain

Theorem

For a *Markov Chain*:

1. Relative entropy $D(\mu_n \parallel \mu'_n)$ decreases with time.
2. Relative entropy $D(\mu_n \parallel \mu)$ between a distribution and the stationary distribution decreases with time.
3. Entropy $H(X_n)$ increases if the stationary distribution is uniform.
4. The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.
5. Shuffles increase entropy.

Proof for item 1

Let μ_n and μ'_n be two probability distributions on the state space of a Markov chain at time n, corresponding to p and q as joint mass functions. By the chain rule:

$$
\begin{aligned}
&D(p(x_n, x_{n+1}) \parallel q(x_n, x_{n+1})) \\
&\quad = D(p(x_n) \parallel q(x_n)) + D(p(x_{n+1}|x_n) \parallel q(x_{n+1}|x_n)) \\
&\quad = D(p(x_{n+1}) \parallel q(x_{n+1})) + D(p(x_n|x_{n+1}) \parallel q(x_n|x_{n+1}))
\end{aligned}
$$

Proof for item 1(cont.)

Since the probability transition function $p(x_{n+1}|x_n) = q(x_{n+1}|x_n)$ from the Markov chain, hence $D(p(x_{n+1}|x_n) \parallel q(x_{n+1}|x_n)) = 0$, and also$D(p(x_n|x_{n+1}) \parallel q(x_n|x_{n+1})) \geq 0$, we have

$$D(p(x_n) \parallel q(x_n)) \geq D(p(x_{n+1}) \parallel q(x_{n+1}))$$

or

$$D(\mu_n \parallel \mu_n') \geq D(\mu_{n+1} \parallel \mu_{n+1}').$$

Proof for item 2

Let $\mu'_n = \mu$, and $\mu'_{n+1} = \mu$, μ can be any stationary distribution. By item 1, the inequality holds.

Remarks

The monotonically non-increasing non-negative sequence $D(\mu_n \parallel \mu)$ has 0 as its limit if the stationary distribution is unique.

Remark on item 3

Let the stationary distribution μ be uniform, then by

$$D(\mu_n \parallel \mu) = \log|\mathcal{X}| - H(\mu_n) = \log|\mathcal{X}| - H(X_n)$$

we know the conclusion holds.

Proof for item 2

Let $\mu'_n = \mu$, and $\mu'_{n+1} = \mu$, μ can be any stationary distribution. By item 1, the inequality holds.

Remarks

The monotonically non-increasing non-negative sequence $D(\mu_n \parallel \mu)$ has 0 as its limit if the stationary distribution is unique.

Remark on item 3

Let the stationary distribution μ be uniform, then by

$$D(\mu_n \parallel \mu) = \log|\mathcal{X}| - H(\mu_n) = \log|\mathcal{X}| - H(X_n)$$

we know the conclusion holds.

Proof for item 2

Let $\mu'_n = \mu$, and $\mu'_{n+1} = \mu$, μ can be any stationary distribution. By item 1, the inequality holds.

Remarks

The monotonically non-increasing non-negative sequence $D(\mu_n \parallel \mu)$ has 0 as its limit if the stationary distribution is unique.

Remark on item 3

Let the stationary distribution μ be uniform, then by

$$D(\mu_n \parallel \mu) = \log |\mathcal{X}| - H(\mu_n) = \log |\mathcal{X}| - H(X_n)$$

we know the conclusion holds.

Proof for item 4

$$H(X_n|X_1) \geq H(X_n|X_1, X_2) = H(X_n|X_2) = H(X_{n-1}|X_1)$$

Remarks on item 5

If T is a shuffle *permutation*of cards and X is the initial *random* position, and if T is independent of X, then

$$H(TX) \geq H(X)$$

where TX is the permutation by the shuffle T on X.

- Proof

$$H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)$$

- Reference for [[2]Thomas, section 4.4.]

Proof for item 4

$$H(X_n|X_1) \geq H(X_n|X_1, X_2) = H(X_n|X_2) = H(X_{n-1}|X_1)$$

Remarks on item 5

If T is a shuffle *permutation*of cards and X is the initial *random* position, and if T is independent of X, then

$$H(TX) \geq H(X)$$

where TX is the permutation by the shuffle T on X.

- Proof

$$H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)$$

- Reference for [[2]Thomas, section 4.4.]

Proof for item 4

$$H(X_n|X_1) \geq H(X_n|X_1, X_2) = H(X_n|X_2) = H(X_{n-1}|X_1)$$

Remarks on item 5

If T is a shuffle *permutation*of cards and X is the initial *random* position, and if T is independent of X, then

$$H(TX) \geq H(X)$$

where TX is the permutation by the shuffle T on X.

- Proof

$$H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)$$

- Reference for [[2]Thomas, section 4.4.]

Proof for item 4

$$H(X_n|X_1) \geq H(X_n|X_1, X_2) = H(X_n|X_2) = H(X_{n-1}|X_1)$$

Remarks on item 5

If T is a shuffle *permutation*of cards and X is the initial *random* position, and if T is independent of X, then

$$H(TX) \geq H(X)$$

where TX is the permutation by the shuffle T on X.

- Proof

$$H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)$$

- Reference for [[2]Thomas, section 4.4.]

Entropy in Markov chain

Theorem(Fano's inequality)

For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$,with $P_e = Pr(X \neq \hat{X})$, we have

$$H(P_e) + P_e \log(|\mathcal{X}|) \geq H(X|\hat{X}) \geq H(X|Y) \tag{38}$$

this inequality can be weakened to

$$1 + P_e \log \mid \mathcal{X} \mid \geq H(X|Y) \tag{39}$$

or

$$P_e \geq \frac{H(X|Y) - 1}{\log \mid \mathcal{X} \mid}. \tag{40}$$

Proof of Fano's inequality

Proof

Define an error random varible,

$$E = \begin{cases} 1, & \text{if } \hat{X} \neq X \\ 0, & \text{if } \hat{X} = X \end{cases}$$

Then,

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} = \underbrace{H(E|\hat{X})}_{\leq H(E) = H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log(|\mathcal{X}|)}.$$

since

$$\begin{aligned} H(X|E, \hat{X}) &= Pr(E = 0)H(X|\hat{X}, E = 0) + Pr(E = 1)H(X|\hat{X}, E = 1) \\ &\leq (1 - P_e)0 + P_e \log | \mathcal{X} | . \end{aligned}$$

Proof of Fano's inequality

Proof(cont.)

By the data-processing inequality, we have $I(X;\hat{X}) \geq I(X;Y)$ since $X \to Y \to \hat{X}$ is a Markov chain, and therefore $H(X|\hat{X}) \geq H(X|Y)$. Thus we have (38) holds.

- For any two random variables X and Y, if the estimator $g(Y)$ takes values in the set X, we can strengthen the inequality slightly by replacing $\log|\mathcal{X}|$ with $\log(|\mathcal{X}|-1)$.

Proof of Fano's inequality

Proof(cont.)

By the data-processing inequality, we have $I(X;\hat{X}) \geq I(X;Y)$ since $X \to Y \to \hat{X}$ is a Markov chain, and therefore $H(X|\hat{X}) \geq H(X|Y)$. Thus we have (38) holds.

- For any two random variables X and Y, if the estimator $g(Y)$ takes values in the set X, we can strengthen the inequality slightly by replacing $\log|\mathcal{X}|$ with $\log(|\mathcal{X}|-1)$.

Empirical probability mass function

Theorem

Let $X_1, X_2, \ldots, X_n$ be i.i.d $\sim p(x)$. Let $\tilde{p}_n$ be the empirical probability mass function of $X_1, X_2, \ldots, X_n$. Then

$$ED(\hat{p}_n \parallel p) \leq ED(\hat{p}_{n-1} \parallel p) \tag{41}$$

Proof

Use $D(\hat{p}_n \parallel p) = E_{\hat{p}_n} \log \frac{\hat{p}_n}{p(x)} = E_{\hat{p}_n} \log \hat{p}_n - \log p(x)$, we have$E_p D(\hat{p}_n \parallel p) = H(p) - H(\hat{p}_n)$,then by item 3 in Markov Chain.

Outline

4 Entropy in Markov chain

5 Bounds on entropy on distributions

Entropy of a multivariate normal distribution

Lemma

Let$X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with mean μ and covariance matrix $\mathbf{K}$. Then

$$h(X_1, X_2, \ldots, X_n) = h(\mathcal{N}(\mu, \mathbf{K})) = \frac{1}{2}\log(2\pi e)^n \mid \mathbf{K} \mid \text{ bits}, \tag{42}$$

where $\mid \mathbf{K} \mid$ denotes the determinant of K.

Bounds on differential entropies

Theorem

Let the random vector $\mathbf{X} \in \mathbf{R}^n$ have zero mean and covariance $\mathbf{K} = E\mathbf{X}\mathbf{X}^t$, i.e., $K_{ij} = EX_iX_j, 1 \leq j, j \leq n$. Then

$$h(\mathbf{X}) \leq \frac{1}{2} \log (2\pi e)^n |\mathbf{K}|, \tag{43}$$

with equality *iff* $\mathbf{X} \sim \mathcal{N}(0, \mathbf{K})$.

Bounds on differential entropies

Proof

Let $g(\mathbf{x})$ be any density satisfying $\int g(\mathbf{x}) x_i x_j d\mathbf{x} = K_{ij}$ for all i, j. Let $\phi_K \sim \mathcal{N}(0, K)$. Note that $\log \phi_K(x)$ is a quadratic form and $\int x_i x_j \phi_K(\mathbf{x}) d\mathbf{x} = K_{ij}$. Then

$$\begin{aligned}
0 &\leq D(g \parallel \phi_K) \\
&= \int g \log(g/\phi_K) \\
&= -h(g) - \int g \log \phi_K \\
&= -h(g) - \int \phi_K \log \phi_K \\
&= -h(g) + h(\phi_K)
\end{aligned}$$

since $h(\phi_K) = \frac{1}{2} \log (2\pi e)^n |\mathbf{K}|$, the conclusion holds.

Bounds on discrete entropies

Theorem

$$H(p_1, p_2, \ldots) \le \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i\right)^2 + \frac{1}{12}\right) \qquad (44)$$

Proof

Define new r.v. X, with the distribution $Pr(X = i) = p_i$, $U \sim \mathcal{U}(0,1)$, define $\tilde{X}$ by $\tilde{X} = X + U$. Then

$$\begin{aligned} H(X) &= -\sum_{i=1}^{\infty} p_i \log p_i \\ &= -\sum_{i=1}^{\infty}\left(\int_i^{i+1} f_{\tilde{X}}(x)dx\right)\log\left(\int_i^{i+1} f_{\tilde{X}}(x)dx\right) \end{aligned}$$

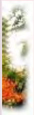
Bounds on discrete entropies

Proof(cont.)

$$H(X) = -\sum_{i=1}^{\infty} \int_{i}^{i+1} f_{\tilde{X}}(x) \log f_{\tilde{X}}(x) dx$$
$$= -\int_{1}^{\infty} f_{\tilde{X}}(x) \log f_{\tilde{X}}(x) dx$$
$$= h(\tilde{X})$$

since $f_{\tilde{X}}(x) = p_i$ for $i \leq x < i+1$. Hence

$$h(\tilde{X}) \leq \frac{1}{2} \log(2\pi e) Var(\tilde{X}) = \frac{1}{2} \log(2\pi e)(Var(X) + Var(U))$$
$$= \frac{1}{2} \log(2\pi e) \left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i \right)^2 + \frac{1}{12} \right).$$



Entropy and fisher information

- The Fisher information matrix is a measure of the minimum error in estimating a parameter vector of a distribution.
- The Fisher information matrix of the distribution of X with a parameter vector θ is defined as

$$J(\theta) = E\{\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^T\} \tag{45}$$

for any $\theta \in \Theta$.

- If f_θ is twice differentiable in θ, and alternative expression is

$$J(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(X)\right]. \tag{46}$$

- Reference in [5].

Entropy and fisher information

- The Fisher information matrix is a measure of the minimum error in estimating a parameter vector of a distribution.
- The Fisher information matrix of the distribution of X with a parameter vector θ is defined as

$$J(\theta) = E\{\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^T\} \tag{45}$$

  for any $\theta \in \Theta$.
- If f_θ is twice differentiable in θ, and alternative expression is

$$J(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(X)\right]. \tag{46}$$

- Reference in [5].

Entropy and fisher information

- The Fisher information matrix is a measure of the minimum error in estimating a parameter vector of a distribution.
- The Fisher information matrix of the distribution of X with a parameter vector θ is defined as

$$J(\theta) = E\{\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^T\} \tag{45}$$

  for any $\theta \in \Theta$.
- If f_θ is twice differentiable in θ, and alternative expression is

$$J(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(X)\right]. \tag{46}$$

- Reference in [5].

Entropy and fisher information

- The Fisher information matrix is a measure of the minimum error in estimating a parameter vector of a distribution.
- The Fisher information matrix of the distribution of X with a parameter vector θ is defined as

$$J(\theta) = E\left\{ \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right] \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right]^T \right\} \tag{45}$$

  for any $\theta \in \Theta$.
- If f_θ is twice differentiable in θ, and alternative expression is

$$J(\theta) = -E\left[\frac{\partial^2}{\partial \theta \partial \theta^T} \log f_\theta(X) \right]. \tag{46}$$

- Reference in [5].

Fisher information of a distribution

- Let X be any r.v. with density $f(x)$, for a location parameter θ, the fisher information w.r.t. θ is given by

$$J(\theta) = \int_{-\infty}^{\infty} f(x-\theta)\left[\frac{\partial}{\partial\theta}\ln f(x-\theta)\right]^2 dx.$$

- As the differentiation w.r.t. x is equivalent to θ, so we can rewrite the Fisher information as

$$\begin{aligned} J(X) = J(\theta) &= \int_{-\infty}^{\infty} f(x)\left[\frac{\partial}{\partial x}\ln f(x)\right]^2 dx \\ &= \int_{-\infty}^{\infty} f(x)\left[\frac{\frac{\partial}{\partial x} f(x)}{f(x)}\right]^2 dx. \end{aligned}$$

Cramér-Rao inequality

Theorem

The mean-squared error of any unbiased estimator $T(X)$ of the parameter θ is lower bounded by the reciprocal of the Fisher information:

$$Var[T(X)] \geq [J(\theta)]^{-1}. \tag{47}$$

Proof

By Cauchy-Schwarz inequality,

$$Var[T(X)]Var\left(\frac{\partial \log f}{\partial \theta}\right) \geq Cov^2\left(T(X), \frac{\partial \log f}{\partial \theta}\right)$$

Then

$$Cov^2\left(T(X), \frac{\partial \log f}{\partial \theta}\right) = E\left(T(X)\frac{\partial \log f}{\partial \theta}\right) = \frac{\partial}{\partial \theta}E_\theta(T(X)) = 1.$$

Cramér-Rao inequality

Theorem

The mean-squared error of any unbiased estimator $T(X)$ of the parameter θ is lower bounded by the reciprocal of the Fisher information:

$$Var[T(X)] \geq [J(\theta)]^{-1}. \tag{47}$$

Proof

By Cauchy-Schwarz inequality,

$$Var[T(X)]Var\left(\frac{\partial \log f}{\partial \theta}\right) \geq Cov^2\left(T(X), \frac{\partial \log f}{\partial \theta}\right)$$

Then

$$Cov^2\left(T(X), \frac{\partial \log f}{\partial \theta}\right) = E\left(T(X)\frac{\partial \log f}{\partial \theta}\right) = \frac{\partial}{\partial \theta}E_\theta(T(X)) = 1.$$

Entropy and Fisher information

Theorem

Let X be any random variable with a finite variance with a density $f(x)$. Let Z be an independent normally distributed random variable with zero mean and unit variance. Then

$$\frac{\partial}{\partial t} h_e(X + \sqrt{t}Z) = \frac{1}{2} J(X + \sqrt{t}Z), \tag{48}$$

where h_e is the differential entropy to base e. In particular, if the limit exists as $t \to 0$,

$$\frac{\partial}{\partial t} h_e(X + \sqrt{t}Z) \mid_{t=0} = \frac{1}{2} J(X). \tag{49}$$

Proof

- Let $Y_t = X + \sqrt{t}Z$. Then the density of Y_t is

$$g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} dx.$$

- It's easy to verify that

$$\frac{\partial}{\partial t} g_t(y) = \frac{1}{2} \frac{\partial^2}{\partial y^2} g_t(y). \tag{50}$$

Proof

- Let $Y_t = X + \sqrt{t}Z$. Then the density of Y_t is

$$g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} dx.$$

- It's easy to verify that

$$\frac{\partial}{\partial t} g_t(y) = \frac{1}{2} \frac{\partial^2}{\partial y^2} g_t(y). \tag{50}$$

Proof

- Since $h_e(Y_t) = -\int_{-\infty}^{\infty} g_t(y) \ln g_t(y) dy$ Differentiating, by $\int g_t(y)dy = 1$ and (50), then integrate by parts, we obtain

$$\frac{\partial}{\partial t} h_e(Y_t) = -\frac{1}{2}\left[\frac{\partial g_t(y)}{\partial y} \ln g_t(y)\right]_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty}\left[\frac{\partial}{\partial y} g_t(y)\right]^2 \frac{1}{g_t(y)} dy.$$

- The first term above goes to 0 at both limit, and by definition, the first term is $\frac{1}{2}J(Y_t)$. Thus the theorem is prove.

Proof

- Since $h_e(Y_t) = -\int_{-\infty}^{\infty} g_t(y) \ln g_t(y) dy$ Differentiating, by $\int g_t(y)dy = 1$ and (50), then integrate by parts, we obtain

$$\frac{\partial}{\partial t} h_e(Y_t) = -\frac{1}{2}\left[\frac{\partial g_t(y)}{\partial y} \ln g_t(y)\right]_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty}\left[\frac{\partial}{\partial y} g_t(y)\right]^2 \frac{1}{g_t(y)} dy.$$

- The first term above goes to 0 at both limit, and by definition, the first term is $\frac{1}{2}J(Y_t)$. Thus the theorem is prove.

Part III

Some important theories deduced from entropy

Outline

6. Entropy rates of subsets

7. The Entropy power inequality

Entropy on subsets

Definition: Average Entropy Rate

Let $(X_1, X_2, \ldots, X_n)$ have a density, and for every $S \subseteq \{1, 2, \ldots, n\}$,denote by $X(S)$ the subset $\{X_i : i \in S\}$. Let

$$h_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} \frac{h(X(S))}{k}. \tag{51}$$

Here $h_k^{(n)}$ is the average entropy in bits per symbol of a randomly drawn k-element subset of $(X_1, X_2, \ldots, X_n)$.

- The average conditional entropy rate and average mutual information rate can be defined similarly on $h(X(S)|X(S^c))$ and$I(X(S); X(S^c))$.

Entropy on subsets

Definition: Average Entropy Rate

Let $(X_1, X_2, \ldots, X_n)$ have a density, and for every $S \subseteq \{1, 2, \ldots, n\}$,denote by $X(S)$ the subset $\{X_i : i \in S\}$. Let

$$h_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} \frac{h(X(S))}{k}. \tag{51}$$

Here $h_k^{(n)}$ is the average entropy in bits per symbol of a randomly drawn k-element subset of $(X_1, X_2, \ldots, X_n)$.

- The average conditional entropy rate and average mutual information rate can be defined similarly on $h(X(S)|X(S^c))$ and$I(X(S); X(S^c))$.

Entropy on subsets

Theorem

1. For average entropy rate,

$$h_1^{(n)} \geq h_2^{(n)} \geq \ldots \geq h_n^{(n)}. \tag{52}$$

2. For average conditional entropy rate,

$$g_1^{(n)} \leq g_2^{(n)} \leq \ldots \leq g_n^{(n)}. \tag{53}$$

3. For average mutual information,

$$f_1^{(n)} \geq f_2^{(n)} \geq \ldots \geq f_n^{(n)}. \tag{54}$$

Proof for Theorem, item 1

- We first proof $h_n^{(n)} \le h_{n-1}^{(n)}$. Since for $i = 1, 2, \ldots, n$,

$$\begin{aligned} h(X_1, X_2, \ldots, X_n) &= h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ &\quad + h(X_i | X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ &\le h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ &\quad + h(X_i | X_1, X_2, \ldots, X_{i-1}) \end{aligned}$$

- Adding these n inequalities and using the chain rule, we obtain

$$\frac{1}{n} h(X_1, X_2, \ldots, X_n) \le \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}{n-1}$$

Thus $h_n^{(n)} \le h_{n-1}^{(n)}$ holds.

Proof for Theorem, item 1

- We first proof $h_n^{(n)} \leq h_{n-1}^{(n)}$. Since for $i = 1, 2, \ldots, n$,

$$\begin{aligned} h(X_1, X_2, \ldots, X_n) &= h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ &\quad + h(X_i | X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ &\leq h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ &\quad + h(X_i | X_1, X_2, \ldots, X_{i-1}) \end{aligned}$$

- Adding these n inequalities and using the chain rule, we obtain

$$\frac{1}{n} h(X_1, X_2, \ldots, X_n) \leq \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}{n-1}$$

Thus $h_n^{(n)} \leq h_{n-1}^{(n)}$ holds.

Proof for Theorem, item 1(cont.)

- For each k-element subset, $h_k^{(k)} \leq h_{k-1}^{(k)}$,
- and hence the inequality remains true after taking the expectation over all k-element subsets chosen uniformly from the n elements.

Proof for Theorem, item 1(cont.)

- For each k-element subset, $h_k^{(k)} \leq h_{k-1}^{(k)}$,
- and hence the inequality remains true after taking the expectation over all k-element subsets chosen uniformly from the n elements.

Entropy on subsets

Proof for Theorem,item 2 and 3

(1) We prove $g_n^{(n)} \leq g_{n-1}^{(n)}$ first.By

$$h(X_1, X_2, \ldots, X_n) \leq \sum_{i-1}^{n} h(X_i)$$

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

Similar as the proof of item 1, we have $g_k^{(k)} \leq g_{k-1}^{(k)}$.
(2) Since $I(X(S); X(S^c) = h(X(S)) - h(X(S)|X(S^c))$, item 3 holds.

Outline

The Entropy power inequality

Theorem

If $\mathbf{X}$ and $\mathbf{Y}$ are independent random n-vectors with densities, then

$$2^{\frac{2}{n}h(\mathbf{X}+\mathbf{Y})} \geq 2^{\frac{2}{n}h(\mathbf{X})} + 2^{\frac{2}{n}h(\mathbf{Y})}. \tag{55}$$

Remarks

For normal distributions, since $2^{2h(X)} = (2\pi e)\sigma_X^2$, we have a new statement of the entropy power inequality.

The Entropy power inequality

Theorem

If $\mathbf{X}$ and $\mathbf{Y}$ are independent random n-vectors with densities, then

$$2^{\frac{2}{n}h(\mathbf{X}+\mathbf{Y})} \geq 2^{\frac{2}{n}h(\mathbf{X})} + 2^{\frac{2}{n}h(\mathbf{Y})}. \tag{55}$$

Remarks

For normal distributions, since $2^{2h(X)} = (2\pi e)\sigma_X^2$, we have a new statement of the entropy power inequality.

The entropy power inequality

Theorem: the entropy power inequality

For two independent random variables X and Y,

$$h(X+Y) \geq h(X'+Y')$$

where X' and Y' are independent normal random variables with $h(X') = h(X)$ and $h(Y') = h(Y)$.

Definitions

- The set sum $A + B$ of two sets $A, B \subset \mathcal{R}^n$ is defined as the set $\{x + y : x \in A, y \in B\}$.
- Example: The set sum of two spheres of radius 1 at the origins is a sphere of radius 2 at the origin.
- Let the $\mathcal{L}_r$ norm of the density be defined by $\| f \|_r = \left(\int f^r(x)dx\right)^{\frac{1}{r}}$.
- The Rényi entropy $h_r(X)$ of order r is defined as

$$h_r(X) = \frac{1}{1-r} \log \left[\int f^r(x)dx \right] \tag{56}$$

for $0 < r < \infty, r \neq 1$.

Remarks on definition

Remarks

- If we take the limit as $r \to 1$, we obtain the Shannon entropy function

$$h(X) = h_1(x) = -\int f(x) \log f(x) dx.$$

- If we take the limit as $r \to 0$, we obtain the logarithm of the support set,

$$h_0 = \log(\mu\{x : f(x) > 0\}).$$

- Thus the zeroth order Rényi entropy gives the measure of the support set of the density of f.

The Brunn-Minkowski inequality

Theorem: Brunn-Minkowski inequality

The volume of the set sum of two sets A and B is greater than the volume of the set sum of two spheres A' and B' with the same volume as A and B, respectively, i.e.,

$$V(A+B) \geq V(A'+B')$$

where A' and B' are spheres with $V(A') = V(A)$ and $V(B') = V(B)$.

The Rényi Entropy Power

Definition

The Rényi entropy power $V_r(X)$ of order r is defined as

$$V_r(X) = \begin{cases} \left[\int f^r(x)dx\right]^{\frac{2}{n}\frac{r'}{r}}, & 0 < r \leq \infty, r \neq 1, \frac{1}{r} + \frac{1}{r'} = 1 \\ \exp[\frac{2}{n}h(X)], & r = 1 \\ \mu(\{x : f(x) > 0\})^{\frac{2}{n}}, & r = 0 \end{cases}$$

Theorem

For two independent random variables X and Y and any $0 \leq r < \infty$ and any $0 \leq \lambda \leq 1$, let $p = \frac{r}{r+\lambda(1-r)}$, $q = \frac{r}{r+(1-\lambda)(1-r)}$, we have

$$\log V_r(X+Y) \geq \lambda \log V_p(X) + (1-\lambda) \log V_q(Y) + H(\lambda) \quad (57)$$

$$+ \left(\frac{1+r}{1-r}\right)\left[H\left(\frac{r+\lambda(1-r)}{1+r}\right) - H\left(\frac{r}{1+r}\right)\right]. \quad (58)$$

Remarks on the Rényi Entropy Power

- The Entropy power inequality. Taking the limit of (58) as $r \to 1$ and setting $\lambda = \frac{V_1(X)}{V_1(X)+V_1(Y)}$, we obtain

$$V_1(X+Y) \geq V_1(X) + V_1(Y).$$

- The Brunn-Minkowski inequality. Similarly letting $r \to 0$ and choosing $\lambda = \frac{\sqrt{V_0(X)}}{\sqrt{V_0(X)}+\sqrt{V_0(Y)}}$, we obtain

$$\sqrt{V_0(X+Y)} \geq \sqrt{V_0(X)} + \sqrt{V_0(Y)}$$

Now let A and B be the support set of X and Y. Then $A+B$ is the support set of $X+Y$, and the equation above reduces to

$$[\mu(A+B)]^{1/n} \geq [\mu(A)]^{1/n} + [\mu(B)]^{1/n},$$

which is the Brunn-Minkowski inequality.

Part IV

Important applications

Outline

Basic concepts

Definition

1. The type P_x of a sequence $x_1, x_2, \ldots, x_n$ is the relative proportion of occurrences in $\mathcal{X}$,i.e., $P_{\mathbf{x}}(a) = N(a|\mathbf{x})/n$ for all $a \in \mathcal{X}$.
2. Let $\mathcal{P}_n$ denote the set of types with a sequence of n symbols.
3. If $P \in \mathcal{P}_n$, then the type class of P, denoted $T(P)$ is defined as:

$$T(P) = \{\mathbf{x} \in \mathcal{X}^n : P_{\mathbf{x}} = P\}$$

Bound on number of types

Theorem: the probability of x

If $X_1, X_2, \ldots, X_n$ are drawn i.i.d.$\sim Q(x)$, then the probability of $\mathbf{x}$ depends only on its type and is given by

$$Q^{(n)}(\mathbf{x}) = 2^{-n(H(P_{\mathbf{x}})+D(P_{\mathbf{x}}\|Q))} \tag{59}$$

Proof

$$\begin{aligned} Q^{(n)}(\mathbf{x}) &= \prod_{i=1}^{n} Q(X_i) = \prod_{a\in\mathcal{X}} Q(a)^{N(a|\mathbf{x})} \\ &= \prod_{a\in\mathcal{X}} Q(a)^{nP_{\mathbf{x}}(a)} = \prod_{a\in\mathcal{X}} 2^{nP_{\mathbf{x}}\log Q(a)} \\ &= 2^{n\sum_{a\in\mathcal{X}}(-P_{\mathbf{x}}(a)\log\frac{P_{\mathbf{x}}(a)}{Q(a)}+P_{\mathbf{x}}(a)\log P_{\mathbf{x}}(a))}. \end{aligned}$$

Bound on number of types

Theorem: the probability of x

If $X_1, X_2, \ldots, X_n$ are drawn i.i.d. $\sim Q(x)$, then the probability of $\mathbf{x}$ depends only on its type and is given by

$$Q^{(n)}(\mathbf{x}) = 2^{-n(H(P_{\mathbf{x}})+D(P_{\mathbf{x}}\|Q))} \tag{59}$$

Proof

$$\begin{aligned} Q^{(n)}(\mathbf{x}) &= \prod_{i=1}^{n} Q(X_i) = \prod_{a\in\mathcal{X}} Q(a)^{N(a|\mathbf{x})} \\ &= \prod_{a\in\mathcal{X}} Q(a)^{nP_{\mathbf{x}}(a)} = \prod_{a\in\mathcal{X}} 2^{nP_{\mathbf{x}}\log Q(a)} \\ &= 2^{n\sum_{a\in\mathcal{X}}(-P_{\mathbf{x}}(a)\log\frac{P_{\mathbf{x}}(a)}{Q(a)}+P_{\mathbf{x}}(a)\log P_{\mathbf{x}}(a))}. \end{aligned}$$

Size of type class $T(P)$

Theorem

$$| \mathcal{P}_n | \leq (n+1)^{|\mathcal{X}|}. \tag{60}$$

Theorem

For any type of $P \in \mathcal{P}_n$,

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq | T(P) | \leq 2^{nH(P)}. \tag{61}$$

Size of type class $T(P)$

Theorem

$$| \mathcal{P}_n | \leq (n+1)^{|\mathcal{X}|}. \tag{60}$$

Theorem

For any type of $P \in \mathcal{P}_n$,

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq | T(P) | \leq 2^{nH(P)}. \tag{61}$$

Size of type class $T(P)$

Proof

By (59), if $\mathbf{x} \in T(P)$, then $P^{(n)}(\mathbf{x}) = 2^{-nH(P)}$, we have

$$1 \geq P^{(n)}(T(P)) = \sum_{\mathbf{x} \in T(P)} P^{(n)}(\mathbf{x}) = \sum_{\mathbf{x} \in T(P)} 2^{-nH(P)} = \mid T(P) \mid 2^{-nH(P)}.$$

For the lower bound, we use the fact $P^{(n)}(T(P)) \geq P^{(n)}(T(\hat{P}))$, for all $\hat{P} \in \mathcal{P}_n$ without proof.

$$\begin{aligned} 1 = \sum_{Q \in \mathcal{P}_n} P^{(n)}(T(Q)) &\leq \sum_{Q \in \mathcal{P}_n} P^{(n)}(T(P)) \\ &\leq (n+1)^{|\mathcal{X}|} P^{(n)}(T(P)) = (n+1)^{|\mathcal{X}|} \mid T(P) \mid 2^{-nH(P)}. \end{aligned}$$

Probability of type class

Theorem

for any $P \in P_n$ and any distribution Q, the probability of the type class $T(P)$ under $Q^{(n)}$ is

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)} \leq |\ Q^{(n)}(T(P))\ | \leq 2^{-nD(P\|Q)}. \tag{62}$$

Proof

$$Q^{(n)}(T(P)) = \sum_{\mathbf{x} \in T(P)} Q^{(n)}(\mathbf{x}) = \sum_{\mathbf{x} \in T(P)} 2^{-n(D(P\|Q)+H(P))}$$
$$= |\ T(P)\ |\ 2^{-n(D(P\|Q)+H(P))}$$

Then use the bounds on $|\ T(P)\ |$ derived in last theorem.

Probability of type class

Theorem

for any $P \in P_n$ and any distribution Q, the probability of the type class $T(P)$ under $Q^{(n)}$ is

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)} \leq \mid Q^{(n)}(T(P)) \mid \leq 2^{-nD(P\|Q)}. \quad (62)$$

Proof

$$Q^{(n)}(T(P)) = \sum_{\mathbf{x} \in T(P)} Q^{(n)}(\mathbf{x}) = \sum_{\mathbf{x} \in T(P)} 2^{-n(D(P\|Q)+H(P))}$$
$$= \mid T(P) \mid 2^{-n(D(P\|Q)+H(P))}$$

Then use the bounds on $\mid T(P) \mid$ derived in last theorem.

Summarize

- We can summarize the basic theorems concerning types in four equations:

$$| \mathcal{P}_n | \leq (n+1)^{|\mathcal{X}|}, \tag{63}$$

$$Q^{(n)}(\mathbf{x}) = 2^{-n(H(P_{\mathbf{x}})+D(P_{\mathbf{x}}\|Q))}, \tag{64}$$

$$| T(P) | \doteq 2^{nH(P)}, \tag{65}$$

$$Q^{(n)}(T(P)) \doteq 2^{-nD(P\|Q)}. \tag{66}$$

- There are only a polynomial number of types and an exponential number of sequences of each type.
- We can calculate the behavior of long sequences based on the properties of the type of the sequence.

Outline

Tight bounds on the size of $\binom{n}{k}$

Lemma

For $0 < p < 1$, $q = 1 - p$, such that np is an integer,

$$\frac{1}{\sqrt{8npq}} \le \binom{n}{np} 2^{-nH(p)} \le \frac{1}{\sqrt{\pi npq}}. \tag{67}$$

Tight bounds on the size of $\binom{n}{k}$

Proof of Lemma

Applying a strong form of Stirling's approximation, which states that

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \le n! \le \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}. \tag{68}$$

we obtain

$$\begin{aligned}\binom{n}{np} &\le \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}}{\sqrt{2\pi np}\left(\frac{np}{e}\right)^{np}\sqrt{2\pi nq}\left(\frac{nq}{e}\right)^{nq}} \\ &= \frac{1}{\sqrt{2\pi npq}}\frac{1}{p^{np}q^{nq}}e^{\frac{1}{12n}} \\ &< \frac{1}{\sqrt{\pi npq}}2^{nH(p)}\end{aligned}$$

Since $e^{\frac{1}{12n}} < e^{\frac{1}{12}} < \sqrt{2}$. The lower bound is obtained similarly.

Tight bounds on the size of $\binom{n}{k}$

Proof of Lemma(cont.)

$$\begin{aligned}\binom{n}{np} &\geq \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{-\left(\frac{1}{12np}+\frac{1}{12nq}\right)}}{\sqrt{2\pi np}\left(\frac{np}{e}\right)^{np}\sqrt{2\pi nq}\left(\frac{nq}{e}\right)^{nq}} \\ &= \frac{1}{\sqrt{2\pi npq}}\frac{1}{p^{np}q^{nq}}e^{-\left(\frac{1}{12np}+\frac{1}{12nq}\right)} \\ &< \frac{1}{\sqrt{2\pi npq}}2^{nH(p)}e^{-\left(\frac{1}{12np}+\frac{1}{12nq}\right)}\end{aligned}$$

If $np \geq 1$, and $nq \geq 3$,then $e^{-\left(\frac{1}{12np}+\frac{1}{12nq}\right)} \geq e^{-\frac{1}{9}} = 0.8948 > \frac{\sqrt{\pi}}{2} = 0.8862$. For $np = 1$, $nq = 1$or 2, and $np = 2$, $nq = 2$ can easily be verified that the inequality still holds. Thus we proved the Lemma.

Reference I

石峰and 莫忠息.
信息论基础.
武汉大学出版社, 2rd edition, 2006.

Thomas M. Cover and Joy A. Thomas.
Elements of Information Theory.
John Wiley & Sons, Inc., 2rd edition, 2006.

Simon Haykin.
Neural Networks and Learning Machines.
China Machine Press, 3rd edition, 2011.

David J.C. MacKay.
Information Theory, Inference, and Learning Algorithms.
Cambridge University Press, 2003.

Reference II

Jun Shao.
Mathematical Statistics.
Springer, 2rd edition, 2003.

Thank You!!!

Thank You!!!

Thank You!!!

Thank You!!!

Thank You!!!

Thank You!!!

Thank You!!!

Thank You!!!

Thank You!!!

Thank You!!!