Contents

1 Introduction **3**
1.1 Entropy 3
1.2 Mutual information 4
1.3 Channel Capacity 5
1.4 Relative entropy 5
1.5 Convex function and inequalities 5
1.6 Sufficient statistics 8

Chapter 1

Introduction

1.1 Entropy

Let X be a random variable with probability mass function $p(x)$, then the **entropy** of X is defined as

$$H(X) = \mathbb{E}[-\log(p(X))] = -\sum_{x \in \mathcal{X}} p(x)\log(p(x))$$

which intuitively measures the uncertainty of a single variable. Depending one the base of the logarithm, the entropy is measured in bits, for base 2, nats, for base e. Entropy can also be viewed as the average amount information revealed after sampling X. We can define conditional entropy of X given that $Y = y$ to be

$$H(X|Y = y) = -\sum_{x \in \mathcal{X}} p_{X|Y}(x|y) \lg\left(\frac{p_{XY}(x, y)}{p_Y(y)}\right)$$

and conditional entropy of X given Y is

$$\begin{aligned} H(X|Y) &= \sum_{y \in \mathcal{Y}} p_Y(y) H(X|Y = y) \\ &= -\sum_y \sum_x p_{XY}(x, y) \lg\left(\frac{p_{XY}(x, y)}{p_Y(y)}\right) \end{aligned}$$

Lastly, the joint entropy to variables is defineds

$$H(X, Y) = \mathbb{E}_{X,Y}[-\log(p_{XY}(X, Y))] = -\sum_{x,y} p_{XY}(x, y) \lg(p_{XY}(x, y))$$

From now on we omit the subscript for the PMFs unless it can not be inferred from the context.

Proposition 1.1 (Chain rule for entropy). *For any two random variables X and Y*

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

furthermore if Z is another random variable then

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

which then can be used to generalize the chain rule

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

Proof. For the conditional case

$$
\begin{aligned}
H(X|Z) &= -\sum_{x,z} p(x,z)\lg\left(\frac{p(x,z)}{p(z)}\right)\\
H(Y|X,Z) &= -\sum_{x,y,z} p(x,y,z)\lg\left(\frac{p(x,y,z)}{p(x,z)}\right)\\
\implies H(X|Z)+H(Y|X,Z) &= -\sum_{x,y,z} p(x,y,z)\lg\left(\frac{p(x,y,z)}{p(z)}\right)\\
&= H(X,Y|Z)
\end{aligned}
$$

1.2 Mutual information

Mutual information is the reduction in entropy due to another random variable.

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y)\\
&= \mathbb{E}_{x,y}\left[\lg\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right]\\
&= \sum_x\sum_y p(x,y)\lg\left(\frac{p(x,y)}{p(x)p(y)}\right)\\
&= H(Y) - H(Y|X) = I(Y;X)
\end{aligned}
$$

Proposition 1.2. *$I(X;Y)$ is zero if and only if X and Y are independent.*

For conditional mutual information we have

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

Proposition 1.3 (Chain rule for mutual information). *For a random variable Y and random variables $X_1,\dots,X_n$ we have*

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

Proof. We have

$$
\begin{aligned}
I(X_1,\dots,X_n;Y) &= H(X_1,\dots,X_n) - H(X_1,\dots,X_n|Y)\\
&= \sum_{i=1}^{n} H(X_i|X_{i-1},\dots,X_1) - H(X_i|X_{i-1},\dots,X_1,Y)\\
&= \sum_{i=1}^{n} I(X_i;Y|X_{i-1},\dots,X_1)
\end{aligned}
$$

1.3 Channel Capacity

A *communication channel* is a system in which output depends probabilistically on its input. It is characterized by a probability transition matrix $p(y|x)$. **Capacity** of a communication channel with input X and output Y is defined as

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

1.4 Relative entropy

Relative entropy or *Kullback–Leibler divergence* measures how one probability distribution differs from another.

$$D(p||q) = \mathbb{E}_{p(x)}\left[\lg\left(\frac{p(X)}{q(X)}\right)\right] = \sum_x p(x)\lg\left(\frac{p(x)}{q(x)}\right)$$

Even though it is not a metric, if $D(p||q) = 0 \implies p = q$.

Note that

$$I(X;Y) = \sum_{x,y} p(x,y)\lg\left(\frac{p(x,y)}{p(x)p(y)}\right) = D(p(x,y)||p(x)p(y))$$

Conditional relative entropy is defined as

$$\begin{aligned} D(p(y|x)||q(y|x)) &= \mathbb{E}_{p(x,y)}\left[\lg\left(\frac{p(Y|X)}{q(Y|X)}\right)\right] \\ &= \sum_x p(x)\sum_y p(y|x)\lg\left(\frac{p(y|x)}{q(y|x)}\right) \end{aligned}$$

Similarly we define the following chain rule

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

1.5 Convex function and inequalities

A function f is said to be convex over an interval $[a,b]$ if for every $x_1, x_2 \in\ ]a,b[$ and $0 \le \lambda \le 1$

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2)$$

f is said to be strictly convex if equality holds only if $\lambda = 0, 1$.

Theorem 1.4. *If f is twice differentiable and has non-negative (positive) second derivative over an interval, then f is convex (strictly convex) over that interval.*

Theorem 1.5 (Jensen's inequality). *If f is a convex function and X is a random variable*

$$f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$$

Moreover, if f is strictly convex, the equality implies that $X = \mathbb{E}[X]$ with probability 1.

Corollary 1.6. *The followings can be shown using the Jensen's inequality*

1. *For non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$*

$$\sum a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum a_i\right) \log\left(\frac{\sum b_i}{\sum a_i}\right)$$

equality holds if and only $\frac{a_i}{b_i} = c, \quad \forall i$. This is called log sum inequality.

2. *$D(p||q) \geq 0$ and equality holds when $p = q$.*

3. *$I(X;Y) \geq 0$ and equality holds when X and Y are independent.*

4. *$D(p(y|x)||q(y|x)) \geq 0$ and equality holds when $p(y|x) = q(y|x)$ for all x and y such that $p(x) > 0$.*

5. *$I(X;Y|Z) \geq 0$ and equality holds when X and Y are conditionally independent given Z.*

Proof. 1. Suppose $\lambda_i = b_i$, $x_i = \frac{a_i}{b_i}$, and $f(x) = x \log x$ then

$$\begin{aligned} \frac{\sum \lambda_i f(x_i)}{\sum \lambda_i} &= \frac{\sum a_i \log \frac{a_i}{b_i}}{\sum b_i} \\ &\geq \frac{\sum a_i}{\sum b_i} \log\left(\frac{\sum a_i}{\sum b_i}\right) \\ \implies \sum a_i \log\left(\frac{a_i}{b_i}\right) &\geq \left(\sum a_i\right) \log\left(\frac{\sum b_i}{\sum a_i}\right) \end{aligned}$$

For the Jensen inequality, equality holds when $x_1 = \cdots = x_n$ and thus $\frac{a_i}{b_i} = c, \forall i$.

2. Using the log sum inequality for $a_i = p(x_i)$ and $b_i = q(x_i)$

$$\begin{aligned} \sum_i p(x_i) \log\left(\frac{p(x_i)}{q(x_i)}\right) &\geq \left(\sum p(x_i)\right) \log\left(\frac{\sum p(x_i)}{\sum q(x_i)}\right) \\ &= 0 \end{aligned}$$

equality holds when $p(x) = cq(x)$, and since both are PMFs $c = 1$.

3. we know that

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$$

and equality holds when $p(x,y) = p(x)p(y)$ which means X and Y are independent.

4. Using the log sum inequality for $a_i = p(y_i|x)$ and $b_i = q(y_i|x)$

$$\begin{aligned} \sum_x p(x) \sum_{y_i} p(y_i|x) \log\left(\frac{p(y_i|x)}{q(y_i|x)}\right) &\geq \sum_x p(x) \left(\sum p(y|x_i)\right) \log\left(\frac{\sum p(y|x_i)}{\sum q(y_i|x)}\right) \\ &= 0 \end{aligned}$$

equality holds when $p(y|x) = q(y|x)$ for all y and x with $p(x) > 0$.

5. Since
$$I(X;Y|Z) = D(p(x,y|z)||p(x|z)p(y|z)) \geq 0$$
and equality holds when X and Y are independent given Z. ■

Theorem 1.7. *For any random variable X*
$$H(X) \leq \log|X|$$
with equality if and only if X has a uniform distribution.

Proof. Let $u(X)$ be the uniform distribution on X. Then
$$\begin{aligned} D(p||u) &= \sum p(x)\log\left(\frac{p(x)}{u(x)}\right) \\ &= \sum p(x)\log(p(x)) + \log(|X|) \\ &= -H(X) + \log|X| \geq 0 \\ \implies \log|X| &\geq H(X) \end{aligned}$$

Theorem 1.8 (Conditioning reduces entropy).
$$H(X|Y) \geq H(X)$$
However this is on average. That is, $H(X|Y=y)$ might be greater than $H(X)$.

Proof. Mutual information $I(X;Y)$ is greater than zero. ■

Corollary 1.9 (Independence bound on entropy).
$$H(X_1,\ldots,X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Theorem 1.10 (Convexity of relative entroy). *For any two pairs probability mass functions (p_1,q_1) and (p_2,q_2)*
$$D(\lambda p_1 + (1-\lambda)p_1||\lambda q_1 + (1-\lambda)q_1) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$
for all $0 \leq \lambda \leq 1$.

Proof. Note that using the log sum inequality on each term
$$(\lambda p_1 + (1-\lambda)p_2)\log\left(\frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2}\right) \leq \lambda p_1 \log\frac{p_1}{q_1} + (1-\lambda)\log\frac{p_2}{q_2}$$

Theorem 1.11 (Concavity of entropy). *$H(X)$ is a concave function of its distribution, $p(x)$.*

Proof.
$$H(X) = \log|X| - D(p||u)$$

Theorem 1.12. *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 $p(x)$*

Definition (Markov chain): Let X, Y, Z be random variables are said to form a Markov chain in that order $X \to Y \to Z$ if the conditional distribution of Z depends only Y and is conditionally independent of X. Specifically

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

For example $Z = f(Y)$ then $X \to Y \to Z$ is a Markov chain. Note that

$$X \to Y \to Z \implies Z \to Y \to Z$$

and hence we can write $X \leftrightarrow Y \leftrightarrow Z$.

Theorem 1.13 (Data processing inequality). *If $X \to Y \to Z$ is a Markov chain, then*

$$I(X;Y) \geq I(X;Z)$$

equality happens if $I(X;Y|Z) = 0$ which implies $X \to Z \to Y$.

1.6 Sufficient statistics

Let $\{f_\theta(x)\}_\theta$ be a family of PMSs and let X be a sample from a distribution in this family. Let $T(X)$ be any statistics. Then, $\theta \to X \to T(X)$ is Markov chain and hence

$$I(\theta; X) \geq I(\theta; T(X))$$

$T(X)$ is sufficient statistics for parameter θ if the conditional distribution of X given $T(X)$ does not depend on θ. Therefore, for a sufficient statistics $\theta \to T(X) \to X$ and thus the data processing inequality becomes an equality

$$I(\theta; X) = I(\theta; T(X))$$

A statistics $T(X)$ is a minimal sufficient statistics relative to $\{f_\theta(x)\}$ if it is a function of every other sufficient statistics U. Equivalently

$$\theta \to T(X) \to U(X) \to X$$

We observe a random variable Y and we guess the correlated variable X using $\hat{X} = f(Y)$ for some function f. Then we wish to know the probability of error

$$P_e = \mathbb{P}\left(X \neq \hat{X}\right)$$

Fano's inequality gives bound on P_e.

Theorem 1.14 (Fano's inequality). *For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$ with $P_e = \mathbb{P}\left(X \neq \hat{X}\right)$ we have*

$$H(P_e) + P_e \lg|X| \geq H\left(X|\hat{X}\right) \geq H(X|Y)$$

and thus

$$\begin{aligned} 1 + P_e \lg|X| &\geq H(X|Y) \\ \implies P_e &\geq \frac{H(X|Y) - 1}{\lg|X|} \end{aligned}$$

Intuitively, this inequality says that if Y does not give much information about X then P_e is greater than when Y has a lot information about X.

Proof. Let E be the random variable with

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

then

$$\begin{aligned} H\left(E, X|\hat{X}\right) &= H\left(E|\hat{X}\right) + H\left(X|E, \hat{X}\right) \\ &= H\left(X|\hat{X}\right) + H\left(E|X, \hat{X}\right) = H\left(X|\hat{X}\right) \end{aligned}$$

therefore

$$\begin{aligned} H\left(X|\hat{X}\right) &= H\left(E|\hat{X}\right) + H\left(X|E, \hat{X}\right) \\ &\leq H(E) + H\left(X|E=0, \hat{X}\right)\mathbb{P}(E=0) + H\left(X|E=1, \hat{X}\right)\mathbb{P}(E=1) \\ &\leq H(P_e) + P_e H(X) \\ &\leq H(P_e) + P_e \lg|X| \end{aligned}$$

and by data processing inequality

$$H\left(X|\hat{X}\right) \geq H(X|Y)$$

Corollary 1.15. *For any two random variables X, Y let $p = \mathbb{P}(X \neq Y)$ then*

$$H(p) + p\lg|X| \geq H(X|Y)$$

Proof. Let $\hat{X} = Y$ in Fano's inequality. ■

Corollary 1.16. *Let $P_e = \mathbb{P}\left(X \neq \hat{X}\right)$ where $\hat{X} : \mathcal{Y} \to \mathcal{X}$ then*

$$H(P_e) + P_e \lg(|X| - 1) \geq H(X|Y)$$

Lemma 1.17. *If X, X' are i.i.d with entropy $H(X)$ then*

$$\mathbb{P}(X = X') \geq 2^{-H(X)}$$

Proof.

$$\mathbb{P}(X = X') = \sum_x p^2(x) = \sum_x p(x) 2^{\lg p(x)} \geq 2^{\sum p(x)\lg p(x)} = 2^{-H(X)}$$

Corollary 1.18. *Let X, X' be independent variables with $X \sim p(x)$ and $X' \sim r(x)$, $x, x' \in \mathcal{X}$ then*

$$\begin{aligned} \mathbb{P}(X = X') &\geq 2^{-H(p) - D(p||r)} \\ &\geq 2^{-H(r) - D(r||p)} \end{aligned}$$

Example 1.1. We will prove the fact there are infinitely many primes. Let

$$\pi(x) = |\{p \leq x \,|\, p \text{ is a prime}\}|$$

Let $N \sim \mathrm{Unif}\{1, \dots, n\}$ then by the prime factorization theorem

$$N = \prod_{i=1}^{\pi(n)} p_i^{X_i}$$

where X_i are random variables.

$$2^{X_i} \leq p_i^{X_i} \leq N \leq n \implies X_i \leq \lg n$$

Furthemore

$$H(N) = H\big(X_1, \dots, X_{\pi(n)}\big) \leq \sum_{i=1}^{\pi(n)} H(X_i)$$

therefore

$$\lg n \leq \sum_{i=1}^{\pi(n)} H(X_i) \leq \pi(n) \lg(\lg n + 1)$$

implying that

$$\pi(n) \geq \frac{\lg n}{\lg(\lg n + 1)}$$

hence $\pi(n) \to \infty$ as $n \to \infty$.