Information Theory

Shi Yu
2023K8009926030

University of Chinese Academy of Sciences

Contents

§ 1 Entropy,Relative Entropy and Matual information

1.1 Entropy

Setting $X \sim P$ discrete random variable

"P" is the probability massfunction(PMF) of X.

$P_X(X = x) = P_r[X = x] \qquad P_X(x) \longleftrightarrow p(x)$

Definition 1.1.1. ***Entropy:***

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \qquad (log_2 : bit \quad log_e : nat)$$

Convention: $0log0 = 0$

Actually,$H(X) = H[P]$

Accordingly,$\bar{X} = \sum_{x \in X} p(x)x \sim \mathop{\mathbf{E}}_{X \sim p(x)} X$

$$H(X) \sim \mathop{\mathbf{E}}_{X \sim p(x)} log\frac{1}{p(x)}$$

Example 1.1.1. ***Binary Entropy Function:***

$$h(p) = \mathop{H}_{x \in \{0,1\}}(X) = -plog_2 p - (1-p)log_2(1-p)$$

Here,$p = P(X = 0)$

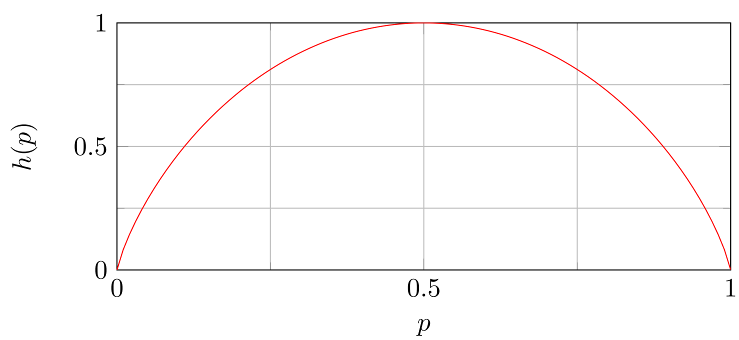


Figure 1: Binary Entropy Function

Each point on the curve represents one distribution of X.

H property:

1. $H[P] \geq 0$
2. $H_b(X) = log_b a H_a(X) \qquad$ nat,when a = e

1.2 Joint Entropy and Conditional Entropy

Definition 1.2.1. ***Joint Entropy:*** $\qquad (X, Y) \sim p_{X,Y}(X = x, Y = y)$

$$H[p_{X,Y}] = -\sum_{x \in \mathfrak{X}, y \in \mathfrak{Y}} p_{X,Y}(x, y) \log p_{X,Y}(x, y)$$

Definition 1.2.2. ***Conditional Entropy:*** $p_{X|Y}(X = x|Y = y)$

$$H[X|Y] = -\sum_{x\in\mathfrak{X},y\in\mathfrak{Y}} \underset{\substack{\uparrow \\ joint}}{p_{X,Y}(x,y)} \log \underset{\substack{\uparrow \\ conditional}}{p_{X|Y}(x|y)}$$

Chain Rule: $P(x, y) = P_Y(y)P_{X|Y}(x|y)$

Theorem 1.2.1. $H(X, Y) = H(Y) + H(X|Y)$

Proof.

$$\begin{aligned} H(X,Y) &= -\mathop{\mathbf{E}}_{X,Y\sim p_{X,Y}} \log p_{X,Y}(x,y) \\ &= -\mathop{\mathbf{E}}_{X,Y\sim p_{X,Y}} \log(p_{X|Y}(x|y)p_Y(y)) \\ &= -\mathop{\mathbf{E}}_{X,Y\sim p_{X,Y}} \log p_{X|Y}(x|y) - \mathop{\mathbf{E}}_{X,Y\sim p_{X,Y}} \log p_Y(y) \\ &= H(X|Y) + H(Y) \end{aligned}$$

□

1.3 Mutual Information

$$X, Y \sim p_{X,Y}(x, y)$$

Definition 1.3.1. ***Mutual Information:***

$$\begin{aligned} I(X;Y) &\overset{def}{=} \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)} \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

Property: if $X \perp Y \Leftrightarrow I(X,Y) = 0$
$I(X,Y) \geq 0$
$[I(X,Y)]_{max} = min\{H(X), H(Y)\}$

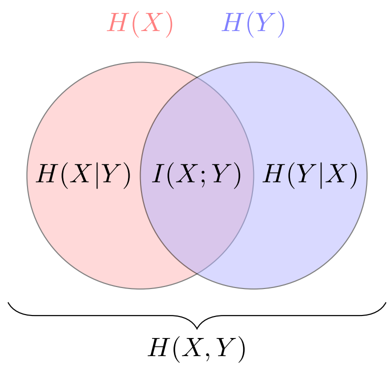


Figure 2: Mutual Information

1.4 KL-Divergence

$$X \sim p_X(x) \qquad X \sim q_X(x)$$

Definition 1.4.1. ***KL-Divergence(Relative Entropy):***
Kullback-Leiblar Divergence between two PMF $p(x)$ and $q(x)$ is defined as:

$$D[p \parallel q] \stackrel{def}{=} \sum_{x \in \mathfrak{X}_p} p(x) \log \frac{p(x)}{q(x)} \in [0, +\infty]$$

KL-Divergence is used to measure the difference between two PMF.

Convention:

1. $0 \log 0 = 0$
2. $0 \log \frac{0}{\tilde{q}} = 0$
3. $\tilde{p} \log \frac{\tilde{p}}{0} = +\infty$

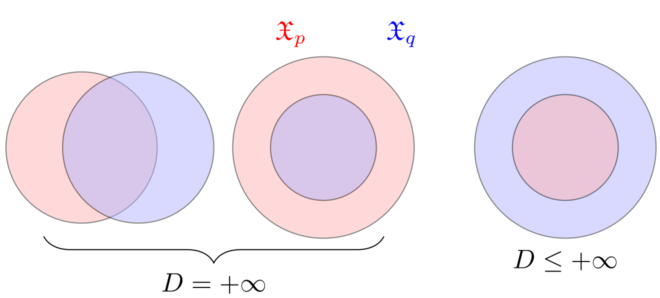


Figure 3: Value selection of KL-Divergence

Property: if $\exists x \in \mathfrak{X}$,st $p(x) > 0$ while $q(x) = 0$ then $D[p \parallel q] = +\infty$.

Definition 1.4.2. ***Conditional Relative Entropy:***

The Conditional Relative Entropy between $p(x, y)$ and $q(x, y)$ is defined as the average KL-Divergence between $p(y|x)$ and $q(y|x)$ by $p(x)$:

$$\begin{aligned} D[p(y|x) \parallel q(y|x)] &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} \end{aligned}$$

1.5 Chain Rule

- $p(x_1, x_2) = p(x_1)p(x_2|x_1)$

  $p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \ldots p(x_n|x_1, x_2, \ldots, x_{n-1})$
- $H(X_1, X_2, \ldots, X_n) = H(X_1)+H(X_2|X_1)+\ldots+H(X_n|X_1, X_2, \ldots, X_{n-1})$
- Conditional Mutual Information:

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

  means the information of X and Y given Z.
- $I(X_1, \ldots, X_n; Y) = I(X_1; Y)+I(X_2; Y|X_1)+\ldots+I(X_n; Y|X_1, \ldots, X_{n-1})$
- Chain Rule for KL-Divergence:

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

  Proof.

  $$\begin{aligned} D[p(x, y)||q(x, y)] &\stackrel{def}{=} \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \\ &= \sum_{x,y} p(x, y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\ &= \sum_{x,y} p(x, y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_{x,y} p(x) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} \\ &= D[p(x)||q(x)] + D[p(y|x)||q(y|x)] \end{aligned}$$

  □

1.6 Jensen Inequality

Definition 1.6.1. ***Convex Function:***
A function $f(x)$ is convex over (a,b),if $\forall x_1, x_2 \in (a,b), f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$, where $\lambda \in [0,1]$.

Example 1.6.1. ***Common Convex and Concave Functions:***
Convex Functions: $f(x) = x^2, e^x \quad \leftrightarrow \quad f^{(2)}(x) \geq 0$
Concave Functions: $f(x) = \log x \quad \leftrightarrow \quad f^{(2)}(x) \leq 0$

Theorem 1.6.1. ***Jensen Inequality:***
For a random variable$x \in \mathfrak{X}$,if $f(x)$ is convex, then:

$$f(\mathbf{E}X) \leq \mathbf{E}f(X) \sim \sum_x p(x)f(x) \geq f(\sum_x p(x)x) \tag{1}$$

Proof. Suppose (1) holds for $|X| \leq K-1$

$$\begin{aligned}
\sum_{i=1}^{K} p(x_i)f(x_i) &= p(x_K)f(x_K) + \sum_{i=1}^{K-1} p(x_i)f(x_i) \\
&= (1-p(x_K)) \sum_{i=1}^{K-1} \frac{p(x_i)}{1-p(x_K)} f(x_i) + p(x_K)f(x_K) \\
&\geq (1-p(x_K)) f(\sum_{i=1}^{K-1} \frac{p(x_i)}{1-p(x_K)} x_i) + p(x_K)f(x_K) \\
&= f(\sum_{i=1}^{K} p(x_i)x_i)
\end{aligned}$$

□

Theorem 1.6.2. ***Information Inequality:***

$$D[p \parallel q] \geq 0 \quad \textit{with equality iff} \quad p(x) = q(x)$$

Proof.

$$\begin{aligned}
D[p \parallel q] &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= -\sum_x p(x) \log \frac{q(x)}{p(x)} \\
\because -logx & \text{ is convex} \\
\therefore D[p \parallel q] &\geq -\log \sum_x p(x) \frac{q(x)}{p(x)} \\
&= -\log \sum_x q(x) \\
&=0
\end{aligned}$$

Here, the equality holds iff $-log\frac{q(x)}{p(x)} = const \Rightarrow q(x) = p(x)$ $\square$

Corollary 1.6.1.

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

Corollary 1.6.2.

$$D(p(X|Y) \parallel q(X|Y)) \geq 0$$

Corollary 1.6.3.

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

Theorem 1.6.3.

$$x \in X \qquad H(X) \leq \log|X|$$

Proof.

$$\begin{aligned}
u(x) =& \frac{1}{|X|} \\
D[p \parallel u] =& \sum_x p(x) \log \frac{p(x)}{u(x)} \geq 0 \\
=& \sum_x p(x) \log |X| + \sum_x p(x) log \frac{1}{u(x)} \\
=& \log |X| - H(X) \geq 0 \\
\Rightarrow H(X) \leq& \log |X|
\end{aligned}$$

$\square$

Theorem 1.6.4.

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

Proof.

$$\begin{aligned}
H(X) &= I(X;Y) + H(X|Y) \\
\because I(X;Y) &\geq 0 \\
\therefore H(X) &\geq H(X|Y)
\end{aligned}$$

$\square$

Example 1.6.2. *P(X,Y) is defined as follows:*

X \ Y	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

$$\begin{aligned} H(X) &= -\sum_x p(x)\log p(x) \\ &= -(\frac{1}{8}\log\frac{1}{8} + \frac{7}{8}\log\frac{7}{8}) \\ &\approx 0.544(bit) \end{aligned}$$

$$\begin{aligned} H(X|Y) &= -\sum_{x,y} p(x,y)\log p(x|y) \\ &= 0 - \frac{3}{4}\log 1 - \frac{1}{8}\log\frac{1}{2} - \frac{1}{8}\log\frac{1}{2} \\ &= 0.25(bit) \end{aligned}$$

1.7 log-sum Inequality and convexity of D, H, I

Theorem 1.7.1. ***log-sum Inequality:***

$$\sum_{i=1}^{n} a_i \log\frac{a_i}{b_i} \geq (\sum_{i=1}^{n} a_i)\log\frac{\sum\limits_{i=1}^{n} a_i}{\sum\limits_{i=1}^{n} b_i} \quad \textit{with equality iff} \quad \frac{a_i}{b_i} = const$$

Proof. Define $f(x) = x\log x$,then $f^{(2)}(x) = \frac{1}{x} > 0$,so $f(x)$ is convex.

$$\begin{aligned} \sum_i a_i \log\frac{a_i}{b_i} &= (\sum_j b_j)\sum_i \frac{b_i}{\sum\limits_j b_j}\frac{a_i}{b_i}\log\frac{a_i}{b_i} \\ &\geq (\sum_j b_j)(\sum_i \frac{b_i}{\sum\limits_j b_j}\frac{a_i}{b_i})\log(\sum_i \frac{b_i}{\sum\limits_j b_j}\frac{a_i}{b_i}) \\ &= \sum_i a_i \log\frac{\sum\limits_{i=1}^{n} a_i}{\sum\limits_{i=1}^{n} b_i} \end{aligned}$$

□

Theorem 1.7.2. ***KL-Divergence is a convex function.***

For two pair of PMF (p_1, q_1) *and* (p_2, q_2),*we have:*

$$D[\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2] \leq \lambda D[p_1 \parallel q_1] + (1-\lambda)D[p_2 \parallel q_2]$$

Also can be noted as:

$$(D(\lambda \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} + (1-\lambda)\begin{bmatrix} p_2 \\ q_2 \end{bmatrix}) \leq \lambda D(\begin{bmatrix} p_1 \\ q_1 \end{bmatrix}) + (1-\lambda)D(\begin{bmatrix} p_2 \\ q_2 \end{bmatrix}))$$

Proof.

$$
\begin{aligned}
left =& \sum_x (\lambda p_1 + (1-\lambda)p_2) \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2} \\
=& \sum_x (\sum_{l=1}^{2} \lambda_l p_l) \log \frac{\sum\limits_{l=1}^{2} \lambda_l p_l}{\sum\limits_{l=1}^{2} \lambda_l q_l} \quad (\lambda_1 = \lambda, \lambda_2 = 1-\lambda) \\
\leq& \sum_x \sum_{l=1}^{2} \lambda_l p_l \log \frac{\lambda_l p_l}{\lambda_l q_l} \\
=& \lambda \sum_x p_1 \log \frac{p_1}{q_1} + (1-\lambda) \sum_x p_2 \log \frac{p_2}{q_2} \\
=& \lambda D[p_1 \parallel q_1] + (1-\lambda) D[p_2 \parallel q_2] = right
\end{aligned}
$$

$$\therefore D[\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2] \leq \lambda D[p_1 \parallel q_1] + (1-\lambda) D[p_2 \parallel q_2]$$

□

Theorem 1.7.3. ***Concavity of Entropy:***

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda) H(p_2)$$

Proof.

$$
\begin{aligned}
H[p] =& -\sum_x p(x) \log p(x) \qquad u(x) = \frac{1}{M} \quad M = |\mathfrak{X}| \\
D[p \parallel u] =& \sum_x p(x) \log \frac{p(x)}{u(x)} = \sum_x p(x) \log p(x) - \sum_x p(x) \log u(x) \\
=& -H[p] - \log M = -H[p] - \log |\mathfrak{X}| \\
& \because D \text{ is a convex function} \\
& \therefore H \text{ is a concave function} \\
& \therefore H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda) H(p_2)
\end{aligned}
$$

□

Alternative proof:

Proof.

$$1.\text{Generate an R.V: } \theta = \begin{cases} 1 & \text{with probability: } \lambda \\ 2 & \text{with probability: } 1-\lambda \end{cases}$$

$$2.\text{Generate an R.V: } X \sim \begin{cases} p_1 & \text{if } \theta = 1 \\ p_2 & \text{if } \theta = 2 \end{cases}$$

$$\begin{aligned}
&\Rightarrow p(x) = \sum_{\theta} p(x,\theta) = \sum_{\theta=1}^{2} p(x|\theta)p(\theta) \\
&\qquad\quad = \lambda p_1(x) + (1-\lambda)p_2(x) \\
&\Rightarrow H[\lambda p_1 + (1-\lambda)p_2] \\
&= H(X) \geq H(X|\theta) = -\sum_{x,\theta} p(x,\theta)\log p(x|\theta) \\
&\qquad\qquad\qquad\qquad = -\sum_{x}\sum_{\theta=1}^{2} p(x|\theta)p(\theta)\log p(x|\theta) \\
&\qquad\qquad\qquad\qquad = -\lambda\sum_{x} p_1 \log p_1 - (1-\lambda)\sum_{x} p_2 \log p_2 \\
&\qquad\qquad\qquad\qquad = \lambda H(p_1) + (1-\lambda)H(p_2) \\
&\therefore H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)
\end{aligned}$$

□

Theorem 1.7.4. ***Convexity of Mutual Information:***
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 $p(x)$.

$$I(X;Y) \begin{cases} \textit{concave of } \ p(x), \textit{for fixed } \ p(y|x) \\ \textit{convex of } \ p(y|x), \textit{for fixed } \ p(x) \end{cases}$$

Proof.

$$\begin{aligned}
I(X;Y) =& \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= H(Y) - H(Y|X) = H(Y) - \sum_{x} p(x)H(Y|X=x)
\end{aligned}$$

If $p(y|x)$ is fixed, then $p(y) = \int p(x)p(y|x)dx$ is a linear function of $p(x)$. Because $H(Y)$ is a concave function of $p(y)$, so H(Y) is a concave function of $p(x)$. The latter term $p(x)H(Y|X=x)$ is a linear function of $p(x)$, so $I(X;Y)$ is a concave function of $p(x)$.

Fix $p(x)$, set two CPMF $p_1(y|x), p_2(y|x)$

$$\begin{aligned}
p_\lambda(y|x) =& \lambda p_1(y|x) + (1-\lambda)p_2(y|x)\\
p_\lambda(x,y) =& p(x)p_\lambda(y|x) = \lambda p_1(x,y) + (1-\lambda)p_2(x,y)\\
p_\lambda(y) =& \int p_\lambda(x,y)dx = \lambda p_1(y) + (1-\lambda)p_2(y)\\
&\text{Set } \ q_\lambda(x,y) = p(x)p_\lambda(y)\\
q_\lambda(x,y) =& \lambda q_1(x,y) + (1-\lambda)q_2(x,y)\\
I(X;Y) =& \sum_{x,y} p_\lambda(x,y)\log\frac{p_\lambda(x,y)}{p(x)p_\lambda(y)} = D[p_\lambda \parallel q_\lambda]\\
\because D[p_\lambda \parallel q_\lambda]& \ \text{ is a convex function of } p_\lambda\\
&p_\lambda(x,y) = p(x)p_\lambda(y|x) \text{ is a linear function of } p_\lambda(y|x)\\
\therefore I(X;Y)& \ \text{ is a convex function of } \ p(y|x)
\end{aligned}$$

□

1.8 Data Processing Inequality

Definition 1.8.1. ***Markov Chain:***
R.V X,Y,Z form a MC: $X \to Y \to Z$ *if* $p(x,y,z) = p(x)p(y|x)p(z|y)$*, which also means* $p(x,z|y) = p(x|y)p(z|y)$*.*

If any part of a process only depends on the previous part,then any three continuous parts of the process form a Markov Chain.

Example 1.8.1. *If a Checker is placed on a chessboard,and the probability of next move is:*

$$P(X) = \begin{cases} p_1, X = up\\ p_2, X = down\\ p_3, X = left\\ p_4, X = right \end{cases}$$

any three continuous moves form a Markov Chain.

Theorem 1.8.1. ***Data Processing Inequality:***
If $X \to Y \to Z$ *form a Markov Chain,then* $I(X;Y) \geq I(X;Z)$*.*

Proof.

$$\begin{aligned}
I(X;Y,Z) =& I(X;Z) + I(X;Y|Z)\\
=& I(X;Y) + I(X;Z|Y)\\
\because X|Y \perp Z|Y \Rightarrow& X \perp Z|Y \Rightarrow I(X;Z|Y) = 0\\
\therefore I(X;Y,Z) =& I(X;Y) = I(X;Z) + I(X;Y|Z)\\
\because I(X;Y|Z) \geq& 0\\
\therefore I(X;Y) \geq& I(X;Z)
\end{aligned}$$

□

Corollary 1.8.1. *If* $Z = f(Y) \Rightarrow I(X;Y) \geq I(X;f(Y))$

1.9 Fano's Inequality

We want to estimate an unknown R.V X with a distribution $p(x)$. We observe an R.V Y that is related to X by the conditional distribution $p(y|x)$. From Y,we caculate a function $f(Y) = \hat{X}$. $X, Y, \hat{X}$ form a MC $X \to Y \to \hat{X}$.

Define the probability of error:

$$P_e = P(\hat{X} \neq X) \qquad Y \sim P(Y|X)$$

We can set the R.V E:

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X \end{cases} \qquad P_e = P(E = 1)$$

Theorem 1.9.1. ***Fano's Inequality:***

$$\begin{aligned} &H(P_e) + P_e \log|\mathfrak{X}| \geq H(X|\hat{X}) \geq H(X|Y) \\ \Rightarrow &1 + P_e \log|\mathfrak{X}| \geq H(X|Y) \\ \Rightarrow &P_e \geq \frac{H(X|Y) - 1}{\log|\mathfrak{X}|} \end{aligned}$$

Proof.

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

If $X, \hat{X}$ is fixed,then E is also fixed, so

$$H(E|X, \hat{X}) = 0 \Rightarrow H(E, X|\hat{X}) = H(X|\hat{X})$$

Since conditioning reduces entropy, we have:

$$H(E|\hat{X}) \leq H(E) = H(P_e)$$

It is easy to see that E is a binary-valued R.V, so $H(X|E, \hat{X})$ can be bounded as:

$$H(X|E, \hat{X}) = P_r(E = 0)H(X|\hat{X}, E = 0) + P_r(E = 1)H(X|\hat{X}, E = 1)$$

Since $E = 0$ means $\hat{X} = X$, so $H(X|\hat{X}, E = 0) = 0$.
Since the upper bound of H is $\log|\mathfrak{X}|$, so: $H(X|E, \hat{X}) \leq P_e \log|\mathfrak{X}|$
Combine the above results, we have:

$$H(E|\hat{X}) + H(X|E, \hat{X}) \leq H(E) + (1 - P_e)0 + P_e \log|\mathfrak{X}|$$

$$\Rightarrow H(X|\hat{X}) \leq H(P_e) + P_e \log|\mathfrak{X}|$$

□

§ 2 AEP(Asymptotic Equipartition Property)

2.1 AEP

Review:Law of large numbers:

For $X_1, X_2, \ldots, X_n$ i.i.d $\sim P \rightarrow$ note as $\underline{X}_n$:

$$\frac{1}{n}\sum_i X_i \overset{n\to\infty}{\underset{p.}{\rightarrow}} \mathbf{E}X = \sum_x xp(x)$$

Theorem 2.1.1. ***AEP:***

$$\frac{1}{n}\log\frac{1}{p(\underline{X}_n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{p(X_i)} \overset{n\to\infty}{\underset{p.}{\rightarrow}} \mathbf{E}\log\frac{1}{p(x)} = H(X)$$

Definition 2.1.1. ***Typical Set:***

$A_\epsilon^{(n)}$ is a set of sequences $(x_1, x_2, \ldots, x_n) \in X^{(n)}$ with the property that:

$$2^{-n(H(X)+\epsilon)} \le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(X)-\epsilon)}$$

1. $\left|A_\epsilon^{(n)}\right| = 2^{nH(X)}$

2. $(1-\epsilon)2^{nH(X)} \le P(A_\epsilon^{(n)}) \to 1$

Example 2.1.1.

$$X_i \sim Bernollip = \begin{cases} 1 & \textit{with probability: } p = 0.9 \\ 0 & \textit{with probability: } p = 0.1 \end{cases}$$

Now we sample $n = 100$ times,then we have:

$$(1, 1, 1, \ldots, 1, 1)\textit{with 100 "1"}$$

$$(1, 0, 1, 1, 0, \ldots, 0, 1)\textit{with 90 "1" and 10 "0"}$$

Obviously, the second one is more likely to happen.

Theorem 2.1.2. ***Property of the typical set:***

(1) If $(x_1, \ldots, x_n) \in A_\epsilon^{(n)}$ then

$$H(X) - \varepsilon \le -\frac{1}{n}\log p(x_1, \ldots, x_n) \le H(X) + \varepsilon$$

(2) $P_r(X_1, \ldots, X_n) \in A_\epsilon^{(n)} \ge 1 - \varepsilon$ for n sufficiently large.

(3) $\left|A_\epsilon^{(n)}\right| \le 2^{n(H(X)+\varepsilon)}$

(4) $\left|A_\epsilon^{(n)}\right| \geq 2^{n(H(X)-\varepsilon)}$

From (3) and (4),we have:

$$\left|A_\epsilon^{(n)}\right| \approx 2^{nH(X)}$$

Proof. (2) From AEP, $-\frac{1}{n}\log p(x_1,\ldots,x_n) \xrightarrow{p.} H(X)$

$\therefore$ For any $\delta > 0, \varepsilon > 0, \exists n_0, \forall n \geq n_0$:

$$P_r\{\left|-\frac{1}{n}\log p(x_1,\ldots,x_n) - H(X)\right| < \varepsilon\} \geq 1-\delta$$

set $\delta = \varepsilon$,then: $P_r\{A_\varepsilon^{(n)}\} \geq 1-\varepsilon$

(3)

$$1 = \sum_{\underline{x}_n \in X^{(n)}} P(x) \geq P_r\{A_\varepsilon^{(n)}\} = \sum_{\underline{x}_n \in A_\varepsilon^{(n)}} P(x) \geq \sum_{\underline{x}_n \in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)}$$

$$= \left|A_\epsilon^{(n)}\right| 2^{-n(H(X)+\varepsilon)}$$

$\Rightarrow \left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\varepsilon)}$

(4)

$$1-\varepsilon \leq P_r\{\left|A_\epsilon^{(n)}\right|\} \leq \sum_{\underline{x}_n \in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = \left|A_\epsilon^{(n)}\right| 2^{-n(H(X)-\varepsilon)}$$

$\Rightarrow \left|A_\epsilon^{(n)}\right| \geq (1-\varepsilon)2^{n(H(X)-\varepsilon)}$

$\square$

2.2 Consequences of AEP:Data Compression

Compression Scheme:

1. If $\underline{x}_n = (x_1,\ldots,x_n) \in A_\varepsilon^{(n)}$,we use $\lceil n(H(X)+\varepsilon)\rceil$ to encode $\underline{x}_n$;
2. If $\underline{x}_n \notin A_\varepsilon^{(n)}$,we use $\lceil n\log|X|\rceil$ to encode $\underline{x}_n$;
3. Use extra 1 bit to identify whether $\underline{x}_n \in A_\varepsilon^{(n)}$ or not.

$$\underline{x}_n \to b_1 b_2 \cdots b_{l(\underline{x}_n)}$$

Here, b_1=1 or 0, $l(\underline{x}_n) \leq \begin{cases} n(H(X)+\varepsilon)+2 & (1) \\ n\log|X|+2 & (2) \end{cases}$

Theorem 2.2.1.

$$\mathbf{E}[\frac{1}{n}l(\underline{x}_n)] \leq H(X)+\varepsilon$$

Proof.

$$
\begin{aligned}
\mathbf{E}[\frac{1}{n}l(\underline{x}_n)] &= \sum_{\underline{x}_n} p(\underline{x}_n)\cdot\frac{1}{n}l(\underline{x}_n) \\
&= \sum_{\underline{x}_n\in A_\varepsilon^{(n)}} p(\underline{x}_n)\cdot\frac{1}{n}l(\underline{x}_n) + \sum_{\underline{x}_n\notin A_\varepsilon^{(n)}} p(\underline{x}_n)\cdot\frac{1}{n}l(\underline{x}_n) \\
&\leq \sum_{\underline{x}_n\in A_\varepsilon^{(n)}} p(\underline{x}_n)\cdot\frac{1}{n}(n(H(X)+\varepsilon)+2) + \sum_{\underline{x}_n\notin A_\varepsilon^{(n)}} p(\underline{x}_n)\cdot\frac{1}{n}(n\log|X|+2) \\
&= P_r\{A_\varepsilon^{(n)}\}\cdot\frac{1}{n}(n(H(X)+\varepsilon)+2) + (1-P_r\{A_\varepsilon^{(n)}\})\cdot\frac{1}{n}(n\log|X|+2) \\
&\leq \frac{1}{n}(n(H(X)+\varepsilon)+2) + \varepsilon\frac{1}{n}(n\log|X|+2) \\
&= H(X)+\varepsilon+\frac{2}{n}+\frac{\varepsilon}{n}\log|X|+\frac{2\varepsilon}{n} \\
&= H(X)+\varepsilon' \qquad \text{Here,we set } \varepsilon' = \varepsilon+\frac{2}{n}+\frac{\varepsilon}{n}\log|X|+\frac{2\varepsilon}{n}
\end{aligned}
$$

□

§ 3 Data Compreession

3.1 Code

Definition 3.1.1. ***Source Code:***

(1) For a R.V. X is a map

$$C : X \to D^* \qquad x \mapsto d_1 d_2 \cdots d_{l(x)} = c(x)$$

(2) $c(x)$ is called ***codeword*** *of x.*

(3) $l(x)$ is called ***length*** *of the codeword,$l(x) \leq \infty$.*

Example 3.1.1. $\mathcal{X} = \{1, 2, 3, 4\}$

x	$p(x)$	Codeword(*)	Codeword(Native)
1	$\frac{1}{2}$	0	00
2	$\frac{1}{4}$	10	01
3	$\frac{1}{8}$	110	10
4	$\frac{1}{8}$	111	11

$$\bar{l}(x) = H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = 1.75 \ bit$$

Definition 3.1.2. ***Nonsigular Code:***
A code C is nonsigualr, if $\forall x \neq x'$ then $C(x) \neq C(x')$.

Definition 3.1.3. ***Extension of Code:***
For a code:

$$C : x \longmapsto C(x)$$

The extension of code is defined as:

$$C^* : x_1 x_2 \cdots x_n \longmapsto C(x_1)C(x_2)\cdots C(x_n)$$

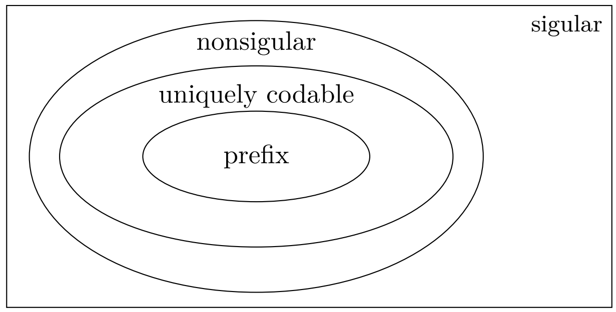
Definition 3.1.4. ***Uniquely Codable:***
A code is uniquely codable if C^* is nonsingular.

Example 3.1.2. *Here are some examples of codes:*

x	singular	nonsigualr	uniquely codable but no prefix	prefix
1	0	0	10	0
2	0	010	00	10
3	1	01	11	110
4	1	10	110	111
$C^*(324)$	101	0101010	1100110	11010111

Obviously, the nonsigular code 0101010 *can be decoded as* 324 *or* 3331.

Definition 3.1.5. ***Prefix Code/Instantaneous Code:***
A code is a prefix code if no codeword is a prefix of any other code.



3.2 Kraft Inequality

Theorem 3.2.1. ***Kraft Inequality:***

1) For any prefix code over an alphabet of size D. The code length $l_1, l_2, \ldots, l_m$ must satisfy:

$$\sum_{i=1}^{m} D^{-l_i} \leq 1$$

2) Conversely, given a set of word length $\{l_1, \ldots, l_m\}$ then there exists a prefix code with those lengths, if the set satisfies the Kraft inequality.

3.3 Optimal Codes

Here we will show some inferences:

$$x \in \{x_1, \dots, x_m\} \quad p(x) = p_1, \dots, p_m$$

$$C(x_1), \dots, C(x_m) \quad l(x_1), \dots, l(x_m) \quad \bar{l} = \sum_{i=1}^{m} p_i l(x_i)$$

$$\text{We want to find:} \min_{l_i} \bar{l} = \sum_{i=1}^{m} p_i l(x_i) \quad s.t. \sum_i D^{-l_i} \leq 1$$

$$l_i \in \mathbf{Z}^* \Rightarrow l_i \in \mathbf{R}^*$$

$$J = \sum_i p_i l_i + \lambda(\sum_i D^{-l_i} - 1)$$

$$\frac{\partial J}{\partial l_j} = p_j - \lambda D^{-l_j} \ln D = 0 \Rightarrow D^{-l_j} = \frac{p_j}{\lambda \ln D}$$

$$\frac{\partial J}{\partial lambda} = \sum_i D^{-l_i} - 1 = 0 \Rightarrow \frac{\sum_i p_i}{\lambda \ln D} = 1 \Rightarrow \lambda = \frac{1}{\ln D}$$

$$D^{-l_j} = \frac{p_j}{\lambda \ln D} = p_j \Rightarrow l_j^* = -\log_D p_j$$

$$\bar{l}^* = \sum_i p_i l_i^* = \sum_i p_i(-\log_D p_i) = -\sum_i p_i \log_D p_i = H_D[X]$$

Theorem 3.3.1. *The expected length L of any prefix D-adic code satisfies:*

$$L \geq H_D[X]$$

3.4 Upper bound on the optimal code length

Theorem 3.4.1.

$$H_D[X] \leq \bar{l}^* \leq H_D[X] + 1$$

Proof.

$$l_i^* \in \mathbf{R}^* \Rightarrow l_i = \lceil l_i^* \rceil \in \mathbf{Z}^*$$

$$\sum_i D^{-l_i} \leq \sum_i D^{-l_i^*} = 1 \quad \text{Kraft Inequality holds}$$

$$\sum_i p_i \lceil l_i^* \rceil \leq \sum_i p_i(l_i^* + 1) = \sum_i p_i l_i^* + \sum_i p_i = H_D[X] + 1$$

□

***Wrong Code:**If we use anothor distribution $q(x)$ instead of the true $p(x)$, then we will get:

Theorem 3.4.2. ***Wrong Code:***

$$l(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil$$

$$\bar{l} = \mathbf{E}l(x) = \sum_i p(x_i) l(x_i) = \sum_i p(x_i) \left\lceil \log \frac{1}{q(x)} \right\rceil$$

$$\begin{aligned}
\sum_i p(x_i) \left\lceil \log \frac{1}{q(x)} \right\rceil &< \sum_i p(x_i)(\log \frac{1}{q(x)} + 1) \\
&= \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)} - \sum_i p(x_i) \log p(x_i) + \sum_i p(x_i) \\
&= D(p \parallel q) + H[p] + 1 \\
\sum_i p(x_i) \left\lceil \log \frac{1}{q(x)} \right\rceil &> D(p \parallel q) + H[p]
\end{aligned}$$

We call $D(p \parallel q)$ the puhishment of wrong code.

3.5 Huffman Code

Observation:

1. Smaller probability $\Rightarrow$ longer codeword.
2. The two longest codewords must have the same length.
3. Two longest codewords merges to one single source symbol, with the probability being the sum of the replaced two symbols.

Here we have the Huffman algorithm:
Input:$\{(x_i, p_i) | i = 1, 2, \cdots, n\}$
Output: $\{C(x_i)\}$ A tree representing Huffman code.
Algorithm:

```
Initialize Q as the PriorityQueue ({p_i,x_i,N_i})//N_i is the tree node
While Q.size()>1:
    p_1,x_1,N_1 = Q.pop()
    p_2,x_2,N_2 = Q.pop()
    N_3 = NewTreeNode(N_1,N_2)
    Q.push(p_1+p_2,Null,N_3)
return Q.pop()
```

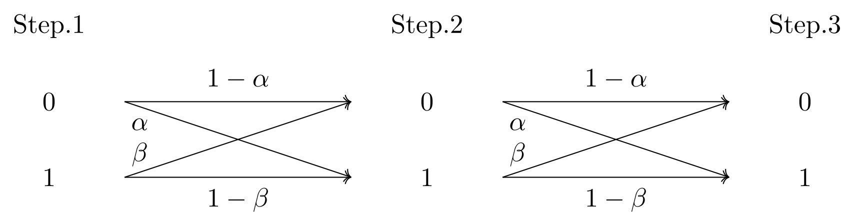
§ 4 Entropy Rate of a stochastic process

$X \leftarrow H[X]$
$X_1, X_2, \ldots, X_n i.i.d \sim p(x) \leftarrow H[X_1, X_2, \ldots, X_n] = nH[X]$
Normally, $X_1, \ldots, X_n, X_i \not\perp X_j, H[X_1, \ldots, X_n] =? \propto n \cdot h$
Here, h is called the entropy rate of the process.

4.1 Markove Chain

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

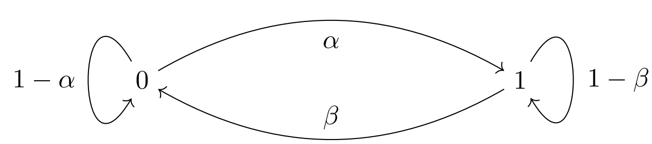
Example 4.1.1. $x \in \{0, 1\}$



This is a Markov Chain with 3 steps.
To describe a Markov Chain, we need:

1. $P_0(X_0)$
2. $P(X_{n+1}|X_n) = \begin{bmatrix} P(0|0) = 1-\alpha & P(1|0) = \alpha \\ P(0|1) = \beta & P(1|1) = 1-\beta \end{bmatrix}$

We can also use a map to describe a Markov Chain:



Definition 4.1.1. ***Time invariant Markov Chain:***
A Markov Chain is time invariant if the transition probability does not depend on time:

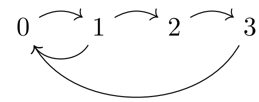
$$P_{k+1}(X_{k+1}|X_k) = P_k(X_k|X_{k-1})$$

Notions of Markov Chain:

1. **State:** X_i is a state of the Markov Chain, X_0 is the initial state.

2. **Irreducable:**$\forall i, j, \exists n, s.t. P(X_n = j | X_i = i) > 0$
3. **Aperiodic:**The largest common factoer of the length of paths from a state to itself is 1.

   Example 4.1.2. *Here is a Markov Chain with 4 states:*

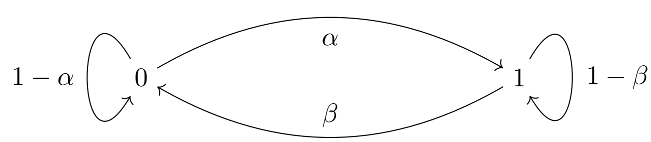
   Length of path 1 is 2, length of path 2 is 4, the largest common factor is 2.

   This Markov Chain is not aperiodic.

4. **Probability trainsition matrix:**$p_{ij} = P(X_{n+1} = X_j | X_n = X_i)$

   $P = [p_{ij}] \qquad P(X_{n+1}) = \sum_{X_n} P(X_{n+1}, X_n) = [P(X_1) \dots P(X_n)]P$

   Example 4.1.3. $P(X_{n+1}|X_n) = \begin{bmatrix} P(0|0) = 1-\alpha & P(1|0) = \alpha \\ P(0|1) = \beta & P(1|1) = 1-\beta \end{bmatrix}$



$$V_n = \begin{matrix} P(X_n = 0) \\ P(X_n = 1) \end{matrix} = \begin{matrix} 0.1 \\ 0.9 \end{matrix} \qquad V_{n+1} = V_n^T P \qquad V_\infty = ?$$

$$V_\infty = V_\infty P \Rightarrow \begin{cases} (1-\alpha)V_1 + \beta V_2 = V_1 \\ \alpha V_1 + (1-\beta)V_2 = V_2 \\ V_1 + V_2 = 1 \end{cases} \Rightarrow \begin{cases} (1-\alpha)V_1 + \beta V_2 = V_1 \\ V_1 + V_2 = 1 \end{cases}$$

$$\Rightarrow \frac{V_1}{V_2} = \frac{\beta}{\alpha}$$

4.2 Entropy Rate

Definition 4.2.1. ***Entropy Rate:***

1) The entropy rate of a stochastic process $\{X_i\}$ is defined as:

$$H[\mathcal{X}] = \lim_{n\to\infty} \frac{1}{n} H[X_1, \dots, X_n]$$

2) Conditional entropy rate:

$$H'[\mathcal{X}] = \lim_{n\to\infty} H[X_n|X_1, \dots, X_{n-1}]$$

Definition 4.2.2. ***Stationary stochastic process:***
A stochastic process $\{X_i\}$ is stationary if the joint distribution of $X_1, \dots, X_n$ does not depend on n:

$$P(X_{l+1}, \dots, X_{l+n}) = P(X_{l+2}, \dots, X_{l+n+1})$$

Theorem 4.2.1. *For a stationary stochastic process:*

$$\begin{aligned} &H[X_n|X_1, \dots, X_{n-1}] \geq H[X_{n+1}|X_1, \dots, X_n] \\ \Rightarrow &H'[\mathcal{X}] \text{ exists a limit} \end{aligned}$$

Proof.

$$H[X_n|X_1, \dots, X_{n-1}] = H[X_{n+1}|X_1, \dots, X_n] \geq H[X_{n+1}|X_1, \dots, X_n]$$

□

Theorem 4.2.2. ***Cesáro mean:***
If $a \to a_n, b_n = \frac{1}{n}\sum\limits_{i=1}^{n} a_i$, then $b_n \to a$

Based on the above theorem, we can get:

$$\begin{aligned} H[\mathcal{X}] &= \lim_{n\to\infty} \frac{1}{n} H[X_1, \dots, X_n] \\ &= \lim_{n\to\infty} (H[X_1] + H[X_2|X_1] + \dots + H[X_n|X_1, \dots, X_{n-1}]) \\ H[X_n|X_1, \dots, X_{n-1}] &\overset{def}{=} b_n \\ H[\mathcal{X}] &= \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} b_m = \lim_{n\to\infty} b_n = H'[\mathcal{X}] \end{aligned}$$

Theorem 4.2.3. ***Entropy rate of a Markov Chain:***
Obviously, the entropy rate only depends on the transition probability matrix:

$$H[\mathcal{X}] = F(P)$$

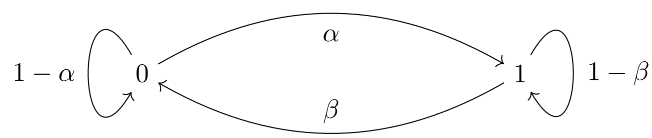
$$\begin{aligned} H[\mathcal{X}] = H'[\mathcal{X}] &= \lim_{n\to\infty} H[X_n|X_1, \dots, X_{n-1}] = \lim_{n\to\infty} H[X_n|X_{n-1}] \\ &= H[X_2|X_1] \qquad X \sim V_\infty \qquad V_\infty = V_\infty P \\ &= -\sum_{X_1} V_\infty P(X_2|X_1) \log P(X_2|X_1) \end{aligned}$$

Theorem 4.2.4. *Let u and P be the stationary distribution and transition probability matrix respectively, then the entropy rate:*

$$H[\mathcal{X}] = -\sum_{i,j} u_i P_{ij} \log P_{ij}$$

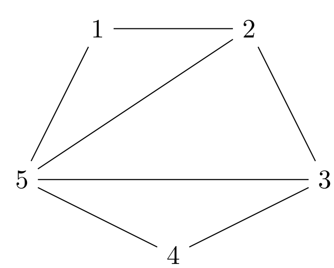
where $u_j = \sum_i u_i p_{ij}$

Example 4.2.1. *For the Markov Chain:*



$$\begin{aligned}
H[\mathcal{X}] =& H[X_2|X_1] \\
=& -\frac{\beta}{\alpha+\beta}((1-\alpha)\log(1-\alpha) + \alpha\log\alpha) - \frac{\alpha}{\alpha+\beta}((1-\beta)\log(1-\beta) + \beta\log\beta) \\
=& \frac{\beta}{\alpha+\beta} H[X_1] + \frac{\alpha}{\alpha+\beta} H[X_2]
\end{aligned}$$

Example 4.2.2. ***Random walk on a graph:***



$$X_k \in \{1,2,3,4,5\} \quad k = 0,1,2 \quad X_0 = l$$

$$P(X_{k+1} = j | X_k = i) = \frac{A_{ij}}{d_i} \quad A_{ij} = \begin{cases} 1 & \textit{if } (i,j) \in E \\ 0 & \textit{otherwise} \end{cases}$$

$$P_0(X_0) = \begin{cases} 1 & \textit{if } X_0 = l \\ 0 & \textit{otherwise} \end{cases} \quad v^* = v^* P \quad v^* = [v_1, v_2, \ldots, v_5]$$

$$v^* P = [v_1 \frac{A_{11}}{d_1} + v_2 \frac{A_{21}}{d_2} + v_3 \frac{A_{31}}{d_3} + v_4 \frac{A_{41}}{d_4} + v_5 \frac{A_{51}}{d_5}, \ldots]$$

$$\Rightarrow v_i^* = \frac{d_i}{2D} \quad D = |E|$$

$$\begin{aligned} H[\mathcal{X}] &= -\sum_{i,j} v_i^* p_{ij} \log p_{ij} = -\sum_{i,j} \frac{d_i}{2D} \frac{A_{ij}}{d_i} \log \frac{A_{ij}}{d_i} = -\sum_{i,j} \frac{A_{ij}}{2D} \log(\frac{A_{ij}}{2D} \frac{2D}{d_i}) \\ &= -\sum_{i,j} \frac{A_{ij}}{2D} \log \frac{A_{ij}}{2D} - \sum_{i,j} \frac{A_{ij}}{2D} \log \frac{2D}{d_i} \\ &= -\sum_{i,j} \frac{A_{ij}}{2D} \log \frac{A_{ij}}{2D} + \sum_{i} \frac{d_i}{2D} \log \frac{d_i}{2D} \\ &= log(2D) - H[v^*] \end{aligned}$$

§ 5 Mutual Information Estimation

5.1 Fenchel-Legendre Transform

Definition 5.1.1. ***F-L transform***

For a given $f(u)$, Fenchel-Legredre transform of f is defined by:

$$f^*(t) = \sup_u \{ut - f(u)\}$$

Corollary 5.1.1. *If f is convex, the $ut - f(u)$ is concave.*

$$u^* : \frac{d(ut - f(u))}{du} = 0 \Rightarrow t = f'(u^*) \Rightarrow u^* = f'^{-1}(t)$$

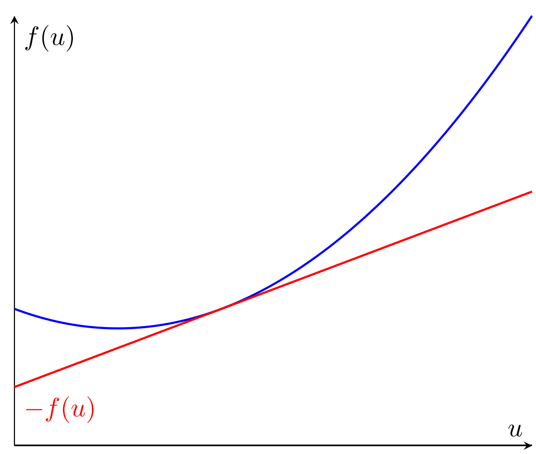
therefore, $f^*(t) = u^*t - f(u^*), u^* = f'^{-1}(t)$

Definition 5.1.2. ***Inverse FL transform***

$$f^{**}(u) = (f^*)^* = \sup_t \{ut - f^*(t)\}$$

Example 5.1.1. *Obviously, the tangent line of $f(u)$ at u^* is:*

$$g(u) = ut^* - f(u^*)$$



Each t corresponds to a tangent line of $f(u)$.

Theorem 5.1.1. ***F-L transform for a convex***

If $f(u)$ is strictly convex, then $f^{**} = f$.

Proof.

$$\begin{aligned}
f^*(t) &= u^*t - f(u^*) \quad \text{where } t = f'(u^*) \\
f^{**}(u) &= (f^*(t))^* = \sup_t \{ut - u^*t + f(u^*)\} \\
&= \sup_{u^*} \{uf'(u^*) - u^*f'(u^*) + f(u^*)\} \\
\frac{d[f'(u^*)(u - u^*) + f(u^*)]}{du^*} &= f''(u^*)(u - u^*) - f'(u^*) + f'(u^*) \\
&= f''(u^*)(u - u^*) = 0 \\
\because f \text{ is strictly convex} &\Rightarrow f''(u^*) > 0 \Rightarrow u = u^* \\
\therefore f^{**}(u) &= \sup_t \{ut - u^*t + f(u^*)\} = f(u)
\end{aligned}$$

□

5.2 Estimate Mutual Information/K-L Divergence via maximizing lower bound

- **Setting:**Suppose we have a set of observed data:

$$\{(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_m, Z_m)\} = D \quad (Y_i, Z_i) \sim P(Y, Z) \rightarrow \text{Unkown}$$

- **Task:**The objective is to estimate:

$$I[Y; Z] = \sum_{Y,Z} p(Y, Z) \log \frac{p(Y, Z)}{p(Y)p(Z)} = D[p(Y, Z) \parallel (p(Y)p(Z))]$$

Example 5.2.1.

$$Y \in \{0, 1\} \qquad Z \in \{0, 1\}$$

i	Y	Z
1	0	1
2	0	0
3	1	1
4	0	0
5	0	1
6	1	1
7	0	0
8	1	0

$$P(Y, Z) = \frac{\#(Y, Z)}{\#total}$$

When the dimension of observed data is too large, it is hard to estimate the distribution by the frequency.

Theorem 5.2.1. ***Nguyen 2010:***

$$D[P(X) \parallel Q(X)] = \mathop{\mathbf{E}}_{X\sim P} \log \frac{P(X)}{Q(X)} \geq \sup_{T\in\mathcal{T}}\{\mathop{\mathbf{E}}_{X\sim P} T(X) - \mathop{\mathbf{E}}_{X\sim Q} e^{T(X)-1}\}$$

Through this theorem, we can estimate the mutual information by machine learning.

Proof.

$$D[P(X) \parallel Q(X)] = \sum_X P(X)\log\frac{P(X)}{Q(X)} = \sum_X Q(X)\frac{P(X)}{Q(X)}\log\frac{P(X)}{Q(X)}$$

$$= \sum_X Q(X)f(u) \quad \begin{cases} u = \frac{P(X)}{Q(X)} \\ f(u) = u\log u \end{cases}$$

$$f'(u) = \log u + 1 \quad f'(u^*) = t = \log u^* + 1 \Rightarrow u^* = e^{t-1}$$

$$f^*(t) = u^*t - f(u^*) = te^{t-1} - f(e^{t-1}) = e^{t-1}$$

$$\therefore \sum_X Q(X)f(u) = \sum_X Q(X)(f^*)^* = \sum_X Q(X)\sup_t\{ut - f^*(t)\}$$

$$= \sum_X Q(X)\sup_t\{\frac{P(X)}{Q(X)}t - f^*(t)\}$$

$$\because f = u\log u \text{ is convex} \Rightarrow f^* \text{ is concave} \Rightarrow f^{**} \text{ is convex}$$

$$\therefore \sum_X Q(X)f(u) \geq \sup_t\{\sum_X Q(X)[\frac{P(X)}{Q(X)}t - f^*(t)]\}$$

$$= \sup_t\{\sum_X P(X)t - \sum_X Q(X)f^*(t)\}$$

$$= \sup_X\{\mathop{\mathbf{E}}_{X\sim P} t_X - \mathop{\mathbf{E}}_{X\sim Q} f^*(t)\}$$

$$= \sup_X\{\mathop{\mathbf{E}}_{X\sim P} t_X - \mathop{\mathbf{E}}_{X\sim Q} e^{t_X-1}\}$$

□

5.3 Implement the estimation of I using lower bound

Let $X = (Y, Z) \quad P(X) = P(Y, Z) \quad Q(X) = P(Y)P(Z)$

Critic function $T_\theta(X)$**:**Define a neural network $T_\theta(X)$ with parameter $\theta = \{\omega_1, \omega_2\}$.$T_\theta = f(\omega_2 f(\omega_1 X))$ while f is a non-linear function.

$$\max_\theta\{\sum_{Y,Z} P(Y,Z)T_\theta(Y,Z) - \sum_{Y,Z} P(Y)P(Z)e^{T_\theta(Y,Z)-1}\}$$

$$\approx \max_\theta\{\frac{1}{N}\sum_{i=1}^N T_\theta(Y_i, Z_i) - \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N e^{T_\theta(Y_i,Z_j)-1}\}$$

$$\approx \max_\theta\{\frac{1}{N}\sum_{i=1}^N T_\theta(Y_i, Z_i) - \frac{1}{M}\sum_{k=1}^M e^{T_\theta(Y_{i_k}Z_{j_k})-1}\} \quad i_k, j_k \quad i.i.d. \sim (1, \ldots, N)$$