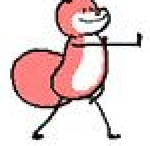
# 6 Information Theory



Authored by Yuhang Xie

# 6.1 Entropy

- Definition

The entropy of a **probability distribution** can be interpreted as a measure of **uncertainty.** To measure the amount of information.

- Examples

1. The sun rises in the east
2. Aliens came to Earth last night
3. No compilation principle exam on the 20th

# 6.1.1 Amount of Information

- Definition

$$h(x) = -\log_2(p(x))$$

- Why log?
  Assuming *x* and *y* are independent

$$h(x, y) = h(x) + h(y)$$

- Information Entropy

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k) = \boxed{\mathbb{E}_X[h(x)]} = -\mathbb{E}_X[\log p(X)]$$

# 6.1.2 Maximum Entropy

The discrete distribution with maximum entropy is the uniform distribution.

$$\mathbb{H}(X) = -\sum_{k=1}^{K} \frac{1}{K} \log(1/K) = -\log(1/K) = \log(K)$$

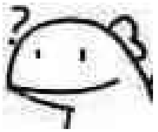
- Proof

$$f(p_1, p_2, \cdots, p_K) = -\sum_{k=1}^{K} p_k \log_2 p_k$$

$$g(p_1, p_2, \cdots, p_k) = \sum_{k=1}^{K} p(X = k) = 1$$

$$F(p_1, p_2, \cdots, p_k) = f + \lambda(1 - g)$$

$$\frac{\partial F}{\partial p_k} = -\left(\frac{p_k}{\ln 2 p_k} + \log_2 p_k\right) + \lambda = 0 \Rightarrow -\left(\frac{1}{\ln 2} + \log_2 p_k\right) + \lambda = 0$$



$$\therefore p_1 = p_2 = \cdots p_k = \frac{1}{K}$$

# 6.1.3 Cross Entropy

- Definition

Cross entropy is a measure of the difference between **two probability** distributions for **a given random variable** or set of events

$$\mathbb{H}(p,q) \triangleq -\sum_{k=1}^{K} p_k \log q_k$$

- Notation

$p$ is the true distribution and $q$ is the estimated distribution.

# 6.1.3 Cross Entropy

$P$: [0.5, 0.25, 0.25, 0]
$Q$: [0.25, 0.25, 0.25, 0.25]

$$\mathbb{H}(p) = 0.5 \times \log 2 + 2 \times 0.25 \times \log 4 = 1.5$$

$$\mathbb{H}(p,q) = (0.5 + 0.25 \times 2 + 0) \times \log 4 = 2$$

$$\mathbb{H}(p) < \mathbb{H}(p,q)$$

In fact, the following inequality between **positive quantities** holds:

$$\mathbb{H}(p) \leq \mathbb{H}(p,q)$$

## 6.1.3 Cross Entropy

$$\mathbb{H}(p) \leq \mathbb{H}(p,q)$$

- Proof

$$\frac{\partial^2 \ln x}{\partial x} = -\frac{1}{x^2} < 0$$

According to the Jenson's inequality,

$$f\left(\frac{\sum a_i x_i}{\sum a_i}\right) \geq \frac{\sum a_i f(x_i)}{\sum a_i}$$

Let $p_i$ equals $a_i$, $\frac{q_i}{p_i}$ equals $x_i$

$$0 = \ln\left(\sum p_i \frac{q_i}{p_i}\right) \geq \sum p_i \ln \frac{q_i}{p_i}$$

$$\sum p_i \ln \frac{q_i}{p_i} \leq 0 \Rightarrow -\sum p_i \ln q_i \geq -\sum p_i \ln p_i$$

we get equality when

$$\frac{q_1}{p_1} = \frac{q_2}{p_2} = \dots \frac{q_K}{p_K}$$

$$q_i = k p_i \Rightarrow \sum_i q_i = k \sum_i p_i \Rightarrow k = 1$$

$$p_i = q_i$$

Therefore, in some machine learning algorithms with cross-entropy loss function, we always minimize it to find the Q distribution that mostly approximates the true distribution P.

## 6.1.4 Joint Entropy

- Definition

The joint entropy of two random variables $X$ and $Y$ is defined as

$$\mathbb{H}(X,Y) = -\sum_{x,y} p(x,y)\log_2 p(x,y)$$

- Example

For example, consider choosing an integer from 1 to 8, $n \in \{1, \ldots, 8\}$. Let $X(n) = 1$ if $n$ is even, and $Y(n) = 1$ if $n$ is prime:

| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $X$ | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| $Y$ | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |

The joint distribution is

| $p(X,Y)$ | $Y=0$ | $Y=1$ |
|---|---|---|
| $X=0$ | $\frac{1}{8}$ | $\frac{3}{8}$ |
| $X=1$ | $\frac{3}{8}$ | $\frac{1}{8}$ |

so the joint entropy is given by

$$\mathbb{H}(X,Y) = -\left[\frac{1}{8}\log_2\frac{1}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{1}{8}\log_2\frac{1}{8}\right] = 1.81 \text{ bits} \tag{6.9}$$

# 6.1.4 Joint Entropy

$$\mathbb{H}(X,Y) = -\sum_{x,y} p(x,y)\log_2 p(x,y)$$

$$= -\sum_{x,y} p(x,y)\log_2(p(x)\,p(y|x)) = -\sum_{x,y} p(x,y)\log_2 p(x) - \sum_{x,y} p(x,y)\log_2 p(y|x)$$

$$= -\sum_{x} p(x)\log_2 p(x) - \sum_{x,y} p(x,y)\log_2 p(y|x)$$

$$= \mathbb{H}(X) + \mathbb{H}(Y|X)$$

If $X$ and $Y$ are independent, then $p(y|x) = p(y)$. So we can get

$$\mathbb{H}(X,Y)$$

$$= -\sum_{x} p(x)\log_2 p(x) - \sum_{y} p(y)\log_2 p(y)$$

$$= \mathbb{H}(X) + \mathbb{H}(Y)$$

We can draw a more generalized conclusion, also known as **chain rule for entropy**

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

# 6.1.5 Conditional Entropy

- Definition

$$
\begin{aligned}
\mathbb{H}(Y|X) &\triangleq \mathbb{E}_{p(X)}[\mathbb{H}(p(Y|X))] \\
&= \sum_x p(x)\,\mathbb{H}(p(Y|X=x)) = -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\
&= -\sum_{x,y} p(x,y) \log p(x,y) + \sum_x p(x) \log p(x) \\
&= \mathbb{H}(X,Y) - \mathbb{H}(X)
\end{aligned}
$$

- Information Gain

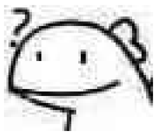
$$
\mathbb{I}(D;A) = \mathbb{H}(D) - \mathbb{H}(D|A)
$$

# 6.1.5.1 Information Gain

| Gender | Clever | Long hair |
|---|---|---|
| Man | 1 | 0 |
| Man | 0 | 0 |
| Woman | 1 | 1 |
| Woman | 1 | 1 |
| Woman | 0 | 1 |
| Man | 1 | 0 |
| Woman | 1 | 1 |
| Man | 1 | 0 |

| | |
|---|---|
| Man | 0 |

?



$$\mathbb{H}(D) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$\mathbb{H}(D|A) = p(A=M)\,\mathbb{H}(D|A=M) + p(A=W)\,\mathbb{H}(D|A=W)$$
$$= \frac{1}{2}\times(-1\times\log_2 1) + \frac{1}{2}\times(-1\times\log_2 1) = 0$$

$$\mathbb{H}(D|B) = p(B=1)\,\mathbb{H}(D|B=1) + p(B=0)\,\mathbb{H}(D|B=0)$$
$$= \frac{3}{4}\times\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{4}\times\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

$$\mathbb{I}(D;A) = \mathbb{H}(D) - \mathbb{H}(D|A) = 1$$

$$\mathbb{I}(D;B) = \mathbb{H}(D) - \mathbb{H}(D|B) = 0$$

# 6.1.6 Entropy for continuous random variables

- Definition

$$h(X) \triangleq -\int_{\mathcal{X}} p(x)\log p(x)\,dx$$

- Example

$X \sim U(0,a)$

$$h(X) = -\int_0^a \frac{1}{a}\log\frac{1}{a}\,dx = \log a$$

$X \sim \mathcal{N}(\mu,\sigma^2)$

$$h(X) = -\int_{\mathcal{X}} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

$$= \frac{1}{2}\ln[2\pi e\sigma^2]$$

$X \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ ?

# 6.2 KL Divergence(Relative Entropy)

- Definition

$$D_{\mathbb{KL}}(p||q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

$$D_{\mathbb{KL}}(p||q) \triangleq \int p(x) \log \frac{p(x)}{q(x)} dx$$

- Interpretation

$$D_{\mathbb{KL}}(p||q) \triangleq \sum_{k=1}^{K} p_k \log p_k - \sum_{k=1}^{K} p_k \log q_k$$

$$\boxed{-\mathbb{H}(p)} + \mathbb{H}(p,q)$$

**Non-negative** 

# 6.2.1 KL Divergence

- Example

$$D_{\mathbb{KL}}\left(\mathcal{N}(x|\mu_1,\sigma_1)||\mathcal{N}(x|\mu_2,\sigma_2)\right) = \int_{\mathcal{X}} \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \log \frac{\frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}} dx$$

$$= \int_{\mathcal{X}} \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \left[\log\frac{\sigma_2}{\sigma_1} - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right] dx$$

$$\int_{\mathcal{X}} \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \log\frac{\sigma_2}{\sigma_1}\, dx = \log\frac{\sigma_2}{\sigma_1}$$

$$\mathbb{E}\left[(X-\mathbb{E}(X))^2\right] = \int f(x)\,(x-\mathbb{E}(x))^2\, dx$$

$$\int_{\mathcal{X}} \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \cdot -\frac{(x-\mu_1)^2}{2\sigma_1^2}\, dx = -\frac{\sigma_1^2}{2\sigma_1^2} = -\frac{1}{2}$$

$$\int_{\mathcal{X}} \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{(x-\mu_1)^2 + 2(\mu_1-\mu_2)(x-\mu_1) + (\mu_2-\mu_1)^2}{2\sigma_2^2}\, dx = \frac{\sigma_1^2 + (\mu_2-\mu_1)^2}{2\sigma_2^2}$$

odd function

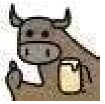
$$D_{\mathbb{KL}} = \boxed{\log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}}$$

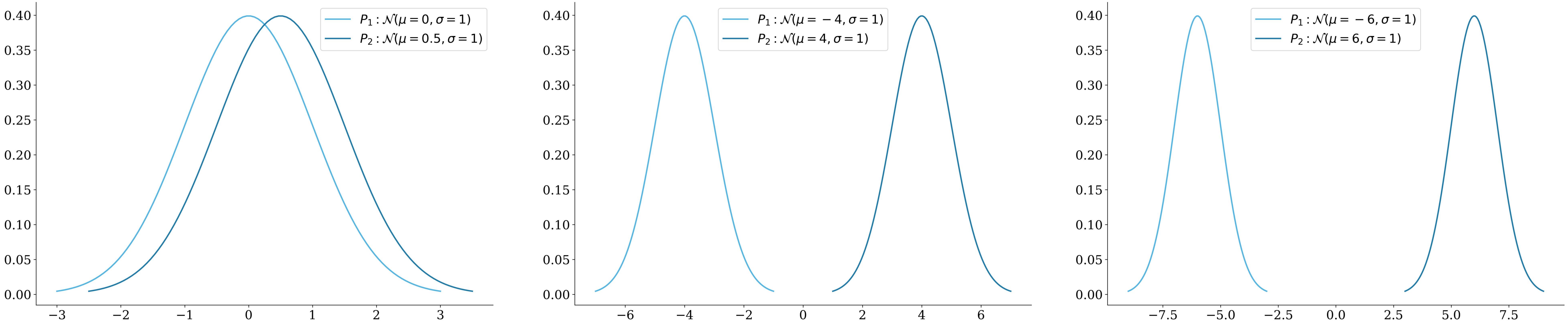
?

# 6.2.2 JS Divergence

- Definition

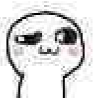
$$D_{\mathbb{JS}}(P_1||P_2) = \frac{1}{2} D_{\mathbb{KL}}\left(P_1 || \frac{P_1+P_2}{2}\right) + \frac{1}{2} D_{\mathbb{KL}}\left(P_2 || \frac{P_1+P_2}{2}\right)$$






$$D_{\mathbb{KL}}(p||q) \triangleq \boxed{\sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}}$$

**divide by zero ?**



# 6.2.2 JS Divergence

$$D_{JS}(P_1||P_2) = \frac{1}{2} D_{KL}\left(P_1 || \frac{P_1+P_2}{2}\right) + \frac{1}{2} D_{KL}\left(P_2 || \frac{P_1+P_2}{2}\right)$$

$$= \frac{1}{2}\sum p(x)\log\left(\frac{p(x)}{\frac{p(x)+q(x)}{2}}\right) + \frac{1}{2}\sum q(x)\log\left(\frac{q(x)}{\frac{p(x)+q(x)}{2}}\right) = \frac{1}{2}\sum p(x)\log\left(\frac{2p(x)}{p(x)+q(x)}\right) + \frac{1}{2}\sum q(x)\log\left(\frac{2q(x)}{p(x)+q(x)}\right)$$

$$= \log 2 \times \frac{1}{2}\left(\sum p(x)+q(x)\right) + \frac{1}{2}\left[\sum p(x)\log\left(\frac{p(x)}{p(x)+q(x)}\right) + \sum q(x)\log\left(\frac{q(x)}{p(x)+q(x)}\right)\right]$$

$$= \log 2 + \frac{1}{2}\left[\sum p(x)\log\left(\frac{p(x)}{p(x)+q(x)}\right) + \sum q(x)\log\left(\frac{q(x)}{p(x)+q(x)}\right)\right]$$

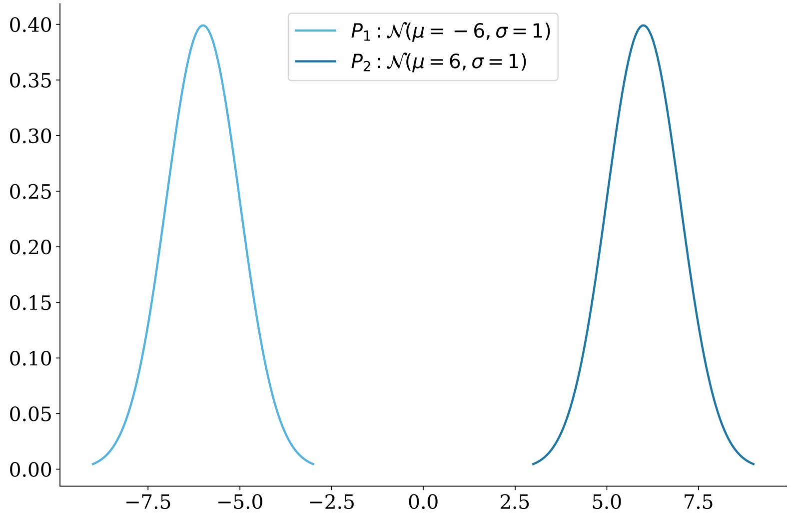
**scale**

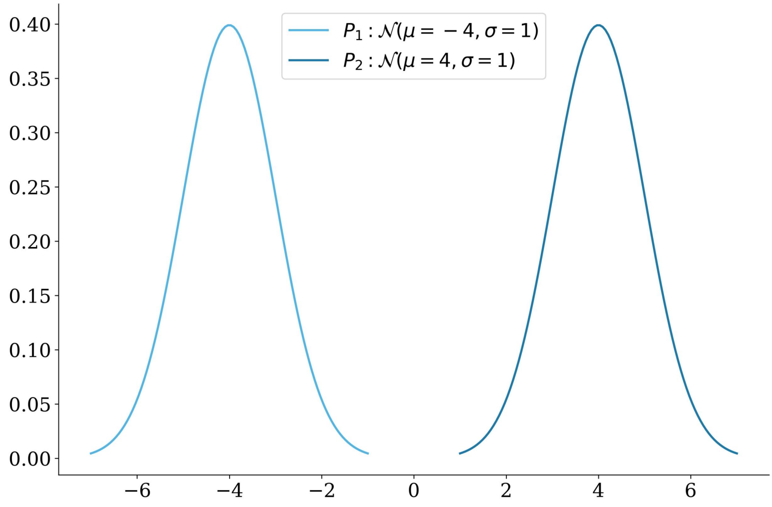
$$\approx \boxed{\log 2}$$

**Gradient Vanish**

?

**Wasserstein Distance**



# 6.2.3 KL Divergence and MLE

- Review

Model is given, but parameters are unknown

$$\hat{\boldsymbol{\theta}}_{\mathrm{mle}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n | \boldsymbol{\theta})$$

negative log likelihood (**NLL**)

$$\mathrm{NLL}(\boldsymbol{\theta}) \triangleq - \sum_{n=1}^{N_{\mathcal{D}}} \log p(\boldsymbol{x}_n | \boldsymbol{\theta})$$

Dirac delta function

$$\delta(x) = 0, (x \neq 0)$$

$$\int \delta(x) dx = 1$$

# 6.2.3 KL Divergence and MLE

Suppose $p$ is the empirical distribution, every $x$ is a probability atom.

$$p_{\mathcal{D}}(x) = \frac{1}{N_{\mathcal{D}}} \sum_{n=1}^{N_{\mathcal{D}}} \delta(x - x_n)$$


$$N_{\mathcal{D}} \to \infty, p_{\mathcal{D}} \to \frac{1}{N_{\mathcal{D}}}$$

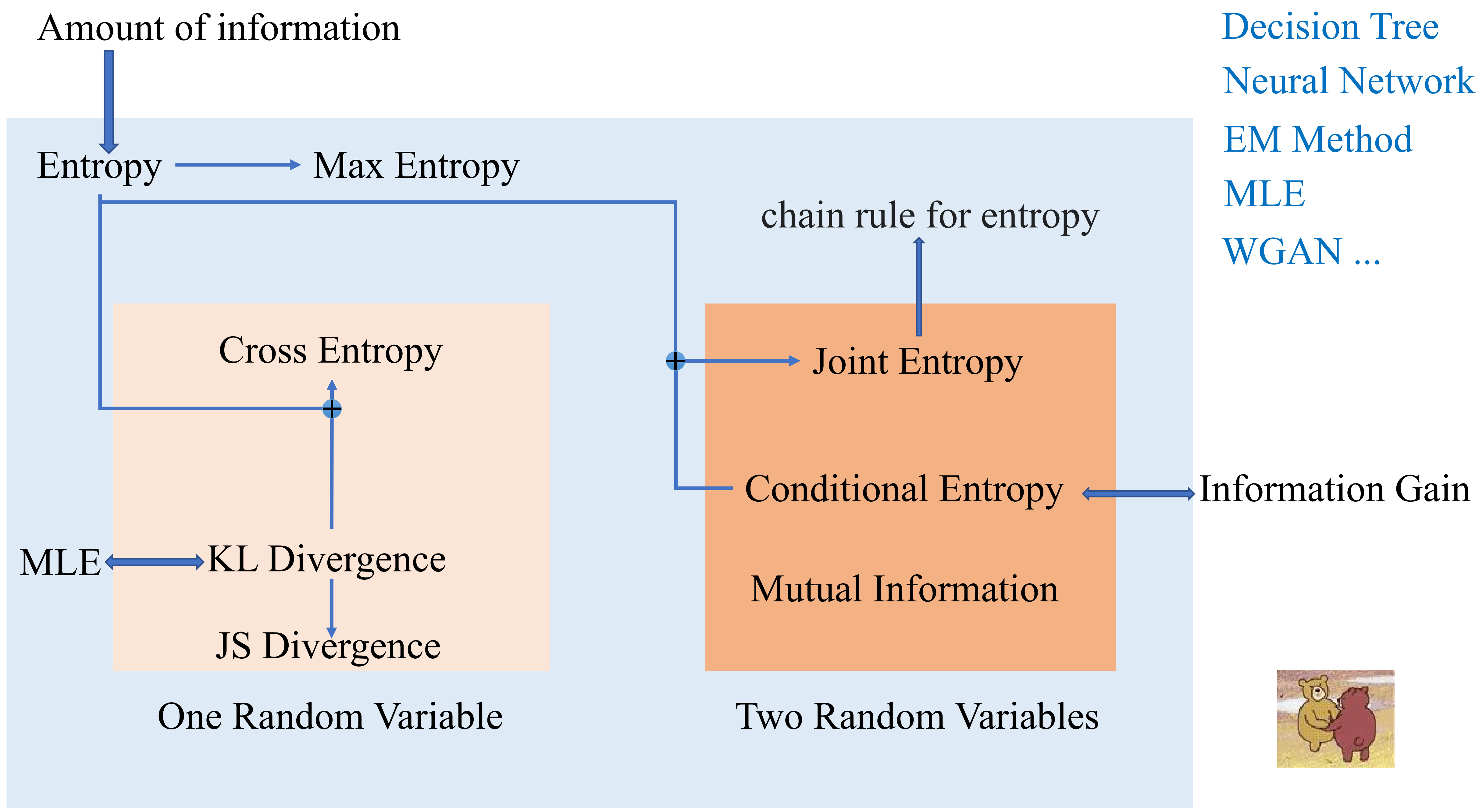
Set $p(x_n|\theta) = q(x)$

$$D_{\mathbb{KL}}(p||q) \triangleq \int p(x) \log \frac{p(x)}{q(x)} dx$$

$$= \text{const} - \int p_{\mathcal{D}}(x) \log q(x) dx = -\int \left[ \frac{1}{N_{\mathcal{D}}} \sum_n \delta(x - x_n) \right] \log q(x) dx + C$$

$$= -\frac{1}{N_{\mathcal{D}}} \sum_n \log p(x_n|\theta) + C$$

$$\mathrm{NLL}(\boldsymbol{\theta}) \triangleq -\sum_{n=1}^{N_{\mathcal{D}}} \log p(\boldsymbol{x}_n|\boldsymbol{\theta})$$


Amount of information
Entropy
Max Entropy
chain rule for entropy
Cross Entropy
Joint Entropy
Conditional Entropy
Information Gain
MLE
KL Divergence
Mutual Information
JS Divergence
One Random Variable
Two Random Variables
Decision Tree
Neural Network
EM Method
MLE
WGAN ...

# 6.3 Mutual Information

- Definition

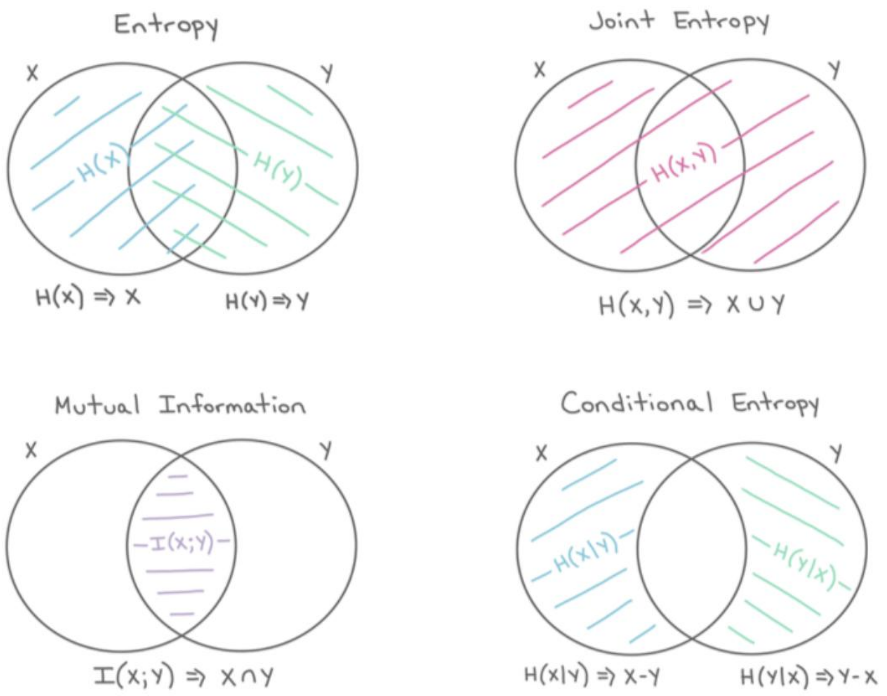

The mutual information between $X$ and $Y$ is defined as follows:

$$\mathbb{I}(X;Y) \triangleq D_{\mathbb{KL}}(p(x,y) \mid\mid p(x)p(y)) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$\mathbb{I}(X;Y) \triangleq \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X)$$

or $$\mathbb{I}(X;Y) \triangleq \mathbb{H}(X,Y) - \mathbb{H}(X|Y) - \mathbb{H}(Y|X)$$

or $$\mathbb{I}(X;Y) \triangleq \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y)$$



## 6.4 References


[1] https://www.zhihu.com/question/65288314/answer/244557337
[2] https://www.zhihu.com/question/310100965
[3] https://blog.csdn.net/luixiao1220/article/details/107530514
[4] https://en.wikipedia.org/wiki/Gibbs%27_inequality
[5] https://en.wikipedia.org/wiki/Information_gain_(decision_tree)
[6] https://github.com/probml/pmlbook/releases/latest/download/book1.pdf
[7] https://www.mckinleylu.com/2021/09/03/san-du/#!

# Thanks for your watching

2022.11.18