# FDU 脑科学 6. 信息论

## 6.1 信号

### 6.1.1 信号 & 采样

信号的定义:

- 物理: 信号是携带信息的一种物理变化
- 信息论: 可用数学函数表示的一种信息流
- 数学: 信号是一个或多个变量的函数或序列
- 自变量: 时间、位移、周期、频率、幅度、相位

信号的分类:

- 确定信号: 信号可以用一个确定的时间函数 (或序列) 表示
- 随机信号: 信号在任意时刻由于某些 "不确定性" 的因素而造成信号无法用一个确定的时间函数 (或序列) 表示

---

确定信号 $x(t)$ 的表征:

- 峰值 (peak value) $P$ 信号的最大瞬时强度
- 双峰值 (double peak value) $P_{p\text{-}p}$
- 均值 (mean value) $\mu_x(t) := \mathrm{E}[x(t)]$
- 均方值 $\Phi_x^2(t) := \mathrm{E}[x^2(t)]$ (信号的平均功率)
  有效值 (RMS) $\Phi_x(t)$
- 方差 (variance) $\sigma_x^2(t) := \mathrm{Var}[x(t)] = \mathrm{E}[(x(t) - \mathrm{E}[x(t)])^2]$ (信号相对于均值的离散程度)
  (去除直流分量后，信号的平均功率)
  标准差 (standard deviation) $\sigma_x$
- 信号总功率 = 交流量功率 + 直流量功率

$$\Phi_x^2(t) = \sigma_x^2(t) + \mu_x^2(t)$$

---

**采样定理:**
如果信号含有的最高频率为 $f_{\max}$，那么采样频率 $f_s$ 应至少高于信号中最高频率 $f_{\max}$ 的 2 倍
这样原来的连续信号可以从采样数据中基本重建出来.
实际应用中采样频率往往应大于信号最高频率的 $2.56 \sim 4$ 倍，防止数字信号失真.
采样点越密集，那越接近信号波原始的样子，损失信息越少，越方便还原信号.

采样频率 $f_s < 2f_{\max}$ 时，采样后信号的频率就会重叠，
即高于采样频率一半的频率成分将被重建成低于采样频率一半的信号.
这种频谱的重叠导致的失真称为**混叠** (alias)

### 6.1.2 Fourier 变换

当函数受到平移不变的线性算子的作用时，Fourier 变换可以为它们提供了一种有用的表示.
函数 $f(t)$ 的 Fourier 变换是一个关于实数 Fourier 变换变量 $\omega$ (角频率) 或 $\mu$ (频率) 的复值函数:

$$\tilde{f}(\omega) := \int_{-\infty}^{\infty} f(t) \exp(-i\omega t) dt$$

$$\tilde{f}(\mu) := \int_{-\infty}^{\infty} f(t) \exp(-i(2\pi\mu) t) dt = \frac{1}{2\pi} \tilde{f}(\omega)$$

其逆变换提供了原函数 $f$ 的另一种表示:

$$
\begin{aligned}
&\frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{f}(\omega)\exp(i\omega t)d\omega \\
&=\frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty}f(s)\exp(-i\omega s)ds\right\}\exp(i\omega t)d\omega \\
&=\int_{-\infty}^{\infty}f(s)\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-iw(s-t))d\omega\right\}ds \\
&=\int_{-\infty}^{\infty}f(s)\delta(s-t)ds \\
&=f(t) \\
\hline
\Rightarrow f(t)&=\frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{f}(\omega)\exp(i\omega t)d\omega \\
&=\int_{-\infty}^{\infty}\tilde{f}(\mu)\exp(i(2\pi\mu)t)d\mu
\end{aligned}
$$

这表明函数 $f(t)$ 可由其 Fourier 变换 $\tilde{f}(\omega)$ 或 $\tilde{f}(\mu)$ 复原.

为保证 $f$ 的 Fourier 变换存在且可逆，它需要满足 Dirichlet 条件:

- ① 周期性: 函数 $f$ 是周期的，即存在 $T>0$ 使得 $f(t+T)=f(t)$ 恒成立.
- ② 有界性: 函数 $f$ 是有界的，即存在 $M>0$ 使得 $|f(t)|\leq M$ 恒成立.
- ③ 绝对可积性: 函数 $f$ 在周期长度的区间上是绝对可积的
- ④ 有限的不连续点: 函数 $f$ 只能有有限个不连续点，并且在这些点处的跳跃幅度必须是有限的
- ⑤ 有限的极值: 函数 $f$ 在任意区间内的极值点的个数必须是有限的

---

关于离散变量的两个函数的卷积是指将一个函数关于其原点翻转 (旋转 $180^\circ$)，并将其滑过另一个函数.
在滑动过程中，对每个位置执行乘积求和运算.
类似地，关于连续变量 $t$ 的两个连续函数 $f,g$ 的卷积 $f\star g$ 定义为:

$$
(f\star g)(t):=\int_{-\infty}^{\infty}f(\tau)g(t-\tau)d\tau
$$

其中 $\tau$ 是积分虚拟变量，负号代表翻转，$t$ 是函数 $g$ 滑过函数 $f$ 的位移.

Fourier 变换在处理卷积时非常有用，因为**卷积的 Fourier 变换是被卷积的两个函数的 Fourier 变换的乘积**:
考虑卷积 $h=f\star g$

$$
\begin{aligned}
\tilde{h}(\omega)&=\int_{-\infty}^{\infty}h(t)\exp(-i\omega t)dt \\
&=\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty}f(s)g(t-s)ds\right\}\exp(-i\omega t)dt \\
&=\int_{-\infty}^{\infty}f(s)\exp(-i\omega s)\left\{\int_{-\infty}^{\infty}g(t-s)\exp(-i\omega(t-s))dt\right\}ds \\
&=\int_{-\infty}^{\infty}f(s)\exp(-i\omega s)\left\{\int_{-\infty}^{\infty}g(\tau)\exp(-i\omega\tau)d\tau\right\}ds\quad(\text{denote }\tau:=t-s) \\
&=\int_{-\infty}^{\infty}f(s)\exp(-i\omega s)ds\cdot\int_{-\infty}^{\infty}g(\tau)\exp(-i\omega\tau)d\tau \\
&=\tilde{f}(\omega)\cdot\tilde{g}(\omega) \\
\hline
\tilde{h}(\mu)&=\tilde{f}(\mu)\cdot\tilde{g}(\mu)
\end{aligned}
$$

因此空间域的卷积对应于频率域的乘积.
从中我们还可以推出 **Parseval 定理**:

$$
\int_{-\infty}^{\infty}|f(t)|^2dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}|\tilde{f}(\omega)|^2d\omega
$$

反过来，空间域的乘积对应于频率域的卷积.

---

实函数 $f(t)$ 的 Fourier 变换通常是复函数，我们记:

$$
\begin{aligned}
\tilde{f}(\mu) &= \mathrm{Re}(\tilde{f}(\mu)) + i \cdot \mathrm{Im}(\tilde{f}(\mu)) \\
&= |\tilde{f}(\mu)| \exp\{i \cdot \phi(\mu)\} \\
&\text{where} \begin{cases} |\tilde{f}(\mu)| := \sqrt{[\mathrm{Re}(\tilde{f}(\mu))]^2 + [\mathrm{Im}(\tilde{f}(\mu))]^2} \\ \phi(\mu) := \arctan\left(\frac{\mathrm{Im}(\tilde{f}(\mu))}{\mathrm{Re}(\tilde{f}(\mu))}\right) \end{cases}
\end{aligned}
$$

其中我们称 $|\tilde{f}(\mu)|$ 为 **Fourier 频谱**，称 $\phi(\mu)$ 为**相位谱**
能量谱的定义为 $P(\mu) := |\tilde{f}(\mu)|^2 = [\mathrm{Re}(\tilde{f}(\mu))]^2 + [\mathrm{Im}(\tilde{f}(\mu))]^2$
若 $f(t)$ 为周期信号，则功率谱可以定义为:

$$
P(\mu) := \lim_{T\to\infty} \frac{1}{T} \left| \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \exp\{-i(2\pi\mu)t\} dt \right|^2
$$

## 6.1.3 Fourier 级数

**Fourier 级数:**
正如多项式是由单项式 $x^n$ 构成的那样，三角多项式由特征函数 $e^{2\pi inx}$ 构成.
**(特征函数)**
对于任意整数 $n$，我们定义频率为 $n$ 的特征函数 $e_n(x) := e^{2\pi inx}$.
它属于连续 $\mathbb{Z}$ 周期复值函数空间 $C(\mathbb{R}/\mathbb{Z};\mathbb{C})$

- **(全体特征构成了一个标准正交系, 实分析, 引理 16.3.5)**
  对于任意整数 $n$ 和 $m$，我们有 $\langle e_n, e_m\rangle = \int_{[0,1]} e_n \bar{e}_m dx = \begin{cases} 1 & \text{if } n = m \\ 0 & \text{if } n \neq m \end{cases}$ 成立.
  并且 $\|e_n\|_2 = (\int_{[0,1]} e_n \bar{e}_n dx)^{\frac{1}{2}} = (\int_{[0,1]} |e_n|^2 dx)^{\frac{1}{2}} = 1$

若 $f$ 的周期为 $T$，则我们可以将其表示为 Fourier 级数:

$$
f(t) = \sum_{n=-\infty}^{\infty} c_n \exp\{i\frac{2\pi n}{T}t\}
$$

$$
c_n = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\exp\{-i\frac{2\pi n}{T}t\}dt \ (n = 0, \pm1, \pm2, \ldots)
$$

它同样满足卷积定理和 Parseval 定理.

根据 Euler 公式 $e^{i\theta} = \cos(\theta) + i\sin(\theta)\ (\theta \in \mathbb{R})$ 可将其写为等价形式:

$$
f(t) = c_0 + \sum_{n=1}^{\infty}\{a_n \cos(\frac{2\pi n}{T}t) + b_n \sin(\frac{2\pi n}{T}t)\}
$$

$$
c_0 = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)dt
$$

$$
\begin{cases} a_n = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\cos(\frac{2\pi n}{T}t)dt \\ b_n = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\sin(\frac{2\pi n}{T}t)dt \end{cases} (n \geq 1)
$$

$$
c_n = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\exp\{-i\frac{2\pi n}{T}t\}dt = \begin{cases} \frac{1}{2}(a_n - ib_n) & n \geq 1 \\ \frac{1}{2}(a_{-n} + ib_{-n}) & n \leq -1 \end{cases}
$$

数值计算时，我们可以截断至某个最大频率 $n_{\max}$，即有**离散 Fourier 变换**:

$$
\begin{aligned}
f_{n_{\max}}(t) &:= \sum_{k=-n_{\max}}^{n_{\max}} c_k \exp\{-i\frac{2\pi k}{T}t\} \\
&= c_0 + \sum_{k=1}^{n_{\max}} \left\{ a_k \cos\left(\frac{2\pi k}{T}t\right) + b_k \sin\left(\frac{2\pi k}{T}t\right) \right\}
\end{aligned}
\quad \text{where} \begin{cases} c_0 = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)dt \\ c_k = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\exp\{-i\frac{2\pi k}{T}t\}\ (k = 0, \pm1, \ldots) \\ a_k = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\cos(\frac{2\pi k}{T}t)dt \quad (k \geq 1) \\ b_k = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\sin(\frac{2\pi k}{T}t)dt \quad (k \geq 1) \end{cases}
$$

我们称 $c_0$ 为直流系数，$a_k$ 为余弦分量系数，$b_k$ 为正弦分量系数.
振幅频谱

### 6.1.4 Gauss 白噪声

**Gauss 白噪声:**
$\mathbf{X}=\{X(t):t\geq 0\}$是一个随机过程，满足:

- **① Gauss 分布:** 对于任意给定的 $t\geq 0$ 都有 $X(t)\sim N(\mu,\sigma^2)$
- **② 白噪声 (white noise):**
  **零均值:** 对于任意 $t\geq 0$ 都有 $\mu_{\mathbf{X}}(t)=\mathrm{E}[X(t)]=0$
  **不相关性:** 对于任意 $s\neq t\geq 0$ 都有自相关函数 $r_{\mathbf{X}}(s,t)=\mathrm{E}[X(s)X(t)]=0$
  在时间上是独立的，意味着不同时刻的噪声值之间没有相关性.
  实际上其自协方差函数 $\mathrm{Cov}_{\mathbf{X}}(s,t)=\mathrm{Cov}(X(s),X(t))=r_{\mathbf{X}}(s,t)-\mu_{\mathbf{X}}(s)\mu_{\mathbf{X}}(t)=r_{\mathbf{X}}(s,t)=\sigma^2\delta(s-t)$
  其中 $\delta(\cdot)$ 为 Dirac $\delta$ 函数.

## 6.2 信息论

### 6.2.1 信息熵

**信息熵:**
概率 $p$ 的不确定性可由 $h:=-\log_2 p$ 表征.
设某一离散随机变量 $X$ 有 $x_1,\dots,x_N$ 取值，概率质量分别为 $p_1,\dots,p_N$
则我们定义其信息熵为:

$$H:=-\sum_{i=1}^{N}p_i\log_2(p_i)$$

**信息量:** 观测前的信息熵减去观测后的信息熵.

$$\Delta I:=H_{\text{before}}-H_{\text{after}}\geq 0$$

信息论中常用的对数底数为 $2$，其单位为 $\mathrm{bit}$
若取自然对数，则信息量的单位为奈特 $(\mathrm{nat})$
若以 $10$ 为对数底数，则信息量的单位为笛特 $(\mathrm{det})$

$$1\ \mathrm{nat}=\log_2 e\approx 1.433\ \mathrm{bit}$$
$$1\ \mathrm{det}=\log_2 10\approx 3.322\ \mathrm{bit}$$

一个系统越是有序，信息熵就越低;
一个系统越是混乱，信息熵就越高.
所以信息熵是系统通讯过程有序化程度的一个度量.

**Shannon: 信息 = 不确定性减少**
信息是用来消除随机不定性的东西

### 6.2.2 条件熵

以下论述中 $X$ 相当于输入，$Y$ 相当于响应.

给定条件 $Y=y$ 下 $X$ 的信息熵称为**条件熵**:
设离散随机变量 $X$ 有 $N$ 个取值 $x_1,\dots,x_N$

$$H(X|Y=y):=-\sum_{i=1}^{N}\mathrm{P}\{X=x_i|Y=y\}\log_2\mathrm{P}\{X=x_i|Y=y\}$$

设离散随机变量 $Y$ 有 $M$ 个取值 $y_1,\dots,y_M$
我们进而定义:

$$
\begin{aligned}
H(X|Y) &:= \sum_{j=1}^{M} \mathrm{P}\{Y=y_j\} H(X|Y=y_j) \\
&= \sum_{j=1}^{M} \mathrm{P}\{Y=y_j\} \cdot -\sum_{i=1}^{N} \mathrm{P}\{X=x_i|Y=y_j\} \log_2 \mathrm{P}\{X=x_i|Y=y_j\} \\
&= -\sum_{j=1}^{M} \mathrm{P}\{Y=y_j\} \sum_{i=1}^{N} \mathrm{P}\{X=x_i|Y=y_j\} \log_2 \mathrm{P}\{X=x_i|Y=y_j\} \\
&= -\sum_{j=1}^{M} \sum_{i=1}^{N} \mathrm{P}\{X=x_i, Y=y_j\} \log_2 \mathrm{P}\{X=x_i|Y=y_j\}
\end{aligned}
$$

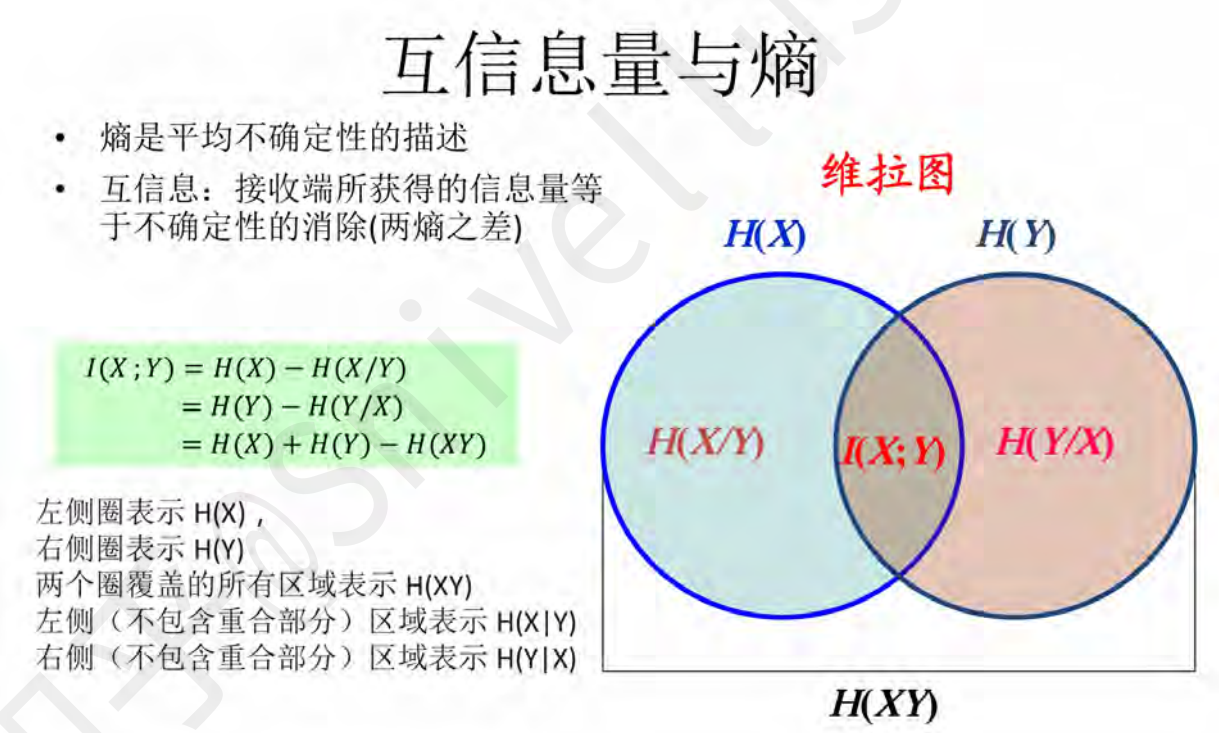
### 6.2.3 联合熵

$X$ 和 $Y$ 联合分布的的信息熵称为**联合熵**:

$$
H(X,Y) := -\sum_{j=1}^{M} \sum_{i=1}^{N} \mathrm{P}\{X=x_i, Y=y_j\} \log_2 \mathrm{P}\{X=x_i, Y=y_j\}
$$

联合熵 $H(X,Y)$ 表示两个随机变量 $X$ 和 $Y$ 的联合不确定性
或两个随机变量 $X$ 和 $y$ 的联合概率分布的自信息的期望值 (平均值)
对于联合分布来说，我们关注的是这两个变量同时取特定值时所提供的信息量.

### 6.2.4 互信息




**互信息** (Mutual Information) 用于量化两个随机变量之间的信息共享程度或相关性.
它可以被理解为接收端从一个随机变量中获得关于另一个随机变量的信息量.

$$
\begin{aligned}
I(X,Y) &:= H(X) + H(Y) - H(X,Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) - H(X|Y) \\
&= -\sum_{i=1}^{N} \mathrm{P}\{Y=y_i\} \log_2 \mathrm{P}\{X=x_i\}) + (-\sum_{j=1}^{M} \sum_{i=1}^{N} \mathrm{P}\{X=x_i, Y=y_j\} \log_2 \mathrm{P}\{X=x_i|Y=y_j\}) \\
&= -\sum_{i=1}^{N}
\end{aligned}
$$

它表示的是当我们联合考虑 $X$ 和 $Y$ 时，相较于独立情况，整体不确定性减少的量.

**不确定性消除:**
当我们知道一个随机变量 (例如 $Y$) 的值时，关于另一个随机变量 (例如 $X$) 的不确定性会减少.
这个减少的程度可以用互信息来量化.
换句话说，互信息量化了知道 $Y$ 的情况下，关于 $X$ 的不确定性减少了多少.

---

下面我们证明 $H(X,Y)=H(Y)+H(X|Y)$:

$$
\begin{aligned}
H(Y)+H(X|Y) &= (-\sum_{j=1}^{M}\mathrm{P}\{Y=y_j\}\log_2\mathrm{P}\{Y=y_j\})+(-\sum_{j=1}^{M}\sum_{i=1}^{N}\mathrm{P}\{X=x_i,Y=y_j\}\log_2\mathrm{P}\{X=x_i|Y=y_j\}) \\
&= -\sum_{j=1}^{M}[\mathrm{P}\{Y=y_j\}\log_2\mathrm{P}\{Y=y_j\}+\sum_{i=1}^{N}\mathrm{P}\{X=x_i,Y=y_j\}\log_2\mathrm{P}\{X=x_i|Y=y_j\}] \\
&= -\sum_{j=1}^{M}[\sum_{i=1}^{N}\mathrm{P}\{X=x_i,Y=y_j\}\log_2\mathrm{P}\{Y=y_j\}+\sum_{i=1}^{N}\mathrm{P}\{X=x_i,Y=y_j\}\log_2\mathrm{P}\{X=x_i|Y=y_j\}] \\
&= -\sum_{j=1}^{M}\sum_{i=1}^{N}\mathrm{P}\{X=x_i,Y=y_j\}[\log_2\mathrm{P}\{Y=y_j\}+\log_2\mathrm{P}\{X=x_i|Y=y_j\}] \\
&= -\sum_{j=1}^{M}\sum_{i=1}^{N}\mathrm{P}\{X=x_i,Y=y_j\}\log_2[\mathrm{P}\{Y=y_j\}\mathrm{P}\{X=x_i|Y=y_j\}] \\
&= -\sum_{j=1}^{M}\sum_{i=1}^{N}\mathrm{P}\{X=x_i,Y=y_j\}\log_2\mathrm{P}\{X=x_i,Y=y_j\} \\
&= H(X,Y)
\end{aligned}
$$

## 6.2.5 补充习题

Monty Hall 问题，它是一个概率论中的经典问题，通常用来说明直觉在概率判断中的不足.
问题背景如下:
有三个房间，其中两个是陷阱，一个是大奖. 你不知道每个房间后面是什么.
你首先选择一个房间 (例如，房间 A)
主持人知道每个房间后面的内容，故意打开一个没有大奖的房间 (例如，房间 B)，并告诉你这是一个陷阱.
现在你有两个选择: 坚持你原来的选择 (房间 A)，或者换到未选择的房间 (房间 C)

在做出选择后，许多人可能会认为坚持或换的机会是相等的，即每个房间都有 $\frac{1}{2}$ 的概率
实际上这个问题的概率并不对称.

- 选择房间 $A$ 后，大奖有 $\frac{1}{3}$ 的概率在房间 $A$ 中.
- 不论你选择哪个房间，主持人总是会打开一个没有大奖的房间.
  若你的初始选择是正确的 ($\frac{1}{3}$ 的概率)，换房间将失去大奖.
  若你的初始选择是错误的 ($\frac{2}{3}$ 的概率)，换房间会得到大奖.

因此从概率的角度来看，换房间是更有利的选择，成功获得大奖的概率为 $\frac{2}{3}$，而坚持原选择的成功概率只有 $\frac{1}{3}$

**计算信息量:**

- ① 主持人提供信息前:

$$
\begin{aligned}
H_{\text{before}} &= -(\frac{1}{3}\log_2(\frac{1}{3})+\frac{1}{3}\log_2(\frac{1}{3})+\frac{1}{3}\log_2(\frac{1}{3})) \\
&= -\log_2(\frac{1}{3}) \\
&= \log_2 3 \\
&\approx 1.585\ \text{bit}
\end{aligned}
$$

- ② 主持人提供信息后:

$$
\begin{aligned}
H_{\text{after}} &= -(\frac{1}{3}\log_2(\frac{1}{3}) + 0\log_2(0) + \frac{2}{3}\log_2(\frac{2}{3})) \\
&= \frac{1}{3}\log_2 3 + \frac{2}{3}\log_2(\frac{3}{2}) \\
&= \log_2 3 - \frac{2}{3} \\
&\approx 0.918 \text{ bit}
\end{aligned}
$$

- ③ 获得的信息量:

$$
\Delta I := H_{\text{before}} - H_{\text{after}} = \frac{2}{3} = 0.667 \text{ bit}
$$

---

人工智能发展的，下一个10年如何借鉴生物脑来突破 (alpha-fold)
以后工作、职业怎么办? 人工智能的能力在不断增长.

首先，生物脑的复杂性和高效性值得我们深入研究.
生物脑处理信息的方式比现有的AI系统更为复杂和灵活.
未来的AI系统可以从这一点出发，借鉴生物脑的学习和适应机制.

- 像 AlphaGo 这样的人工智能系统的训练依赖于大量的GPU，训练阶段的功耗可以达到数十千瓦时.
  生物脑以非常低的能耗运行，典型的人类大脑在静息状态下的能耗约为20瓦特，约占人体总能耗的20%.
  即使在进行复杂思维和决策时，生物脑的能耗也相对较低，对于大多数认知任务，生物脑的能耗远低于现有的人工智能系统
- 生物脑能够通过少量样本进行快速学习，这一能力是当前许多机器学习模型所欠缺的.
  因此，研究如何设计更高效的学习算法，模仿大脑的神经可塑性和联想学习，将是未来AI发展的关键。

在职业和工作方面，AI 会对很多职业构成威胁:

- 白领工作和知识型工作更容易被大语言模型替代，例如财务/审计/税务、翻译、银行、销售业务等.
  其共同特点在于工作任务包含较多的文本处理、资料收集整理等内容，
  而这些知识型的工作任务正是大语言模型人工智能的长项.
- 新增岗位逐渐集中于那些不容易被人工智能技术所替代的岗位
  换句话说，那些比较容易被人工智能所替代的岗位正在逐渐消失。
- 这不是数字技术第一次改变我们的工作，但不同于以往数字技术创新替代较低层次、低薪的劳动者，大型语言模型类AI工具的横空出世，对白领类型工作或相对高报酬的知识型高收入工作具有替代效应