# 信息论

## ——笔记整理

BarryMafu

2025 年 6 月 10 日

# 前言

临时抱佛脚，~ 复习 25Spr 信息论期末考试用的，比较草率请谅解！

第一版中，将直接采用按照课堂的分割方式。之后第二次整理时再重新划分。

BarryMafu

2025 年 6 月 10 日

# 目录

# 第一章　Lecture 2

讲师：王立威　　课程时间：25.Feb.25th　　笔记：25.June.7th

首先，我们不失一般性地考虑如下的通讯需求：

- 无限次的通信 (infinite communication)
- 信息服从某个固定分布 (A prob. distribution over the messages)

在这样的背景下，我们力求最小化平均编码长度.

注意到，我们为了接收方可以唯一解码，我们考虑一组无前缀码 (Prefix-Free Codes). 值得注意，这只是一个充分条件，而并不是必要条件. （下面的例子中就可以看出）

**例 1.0.1.** *考虑码字为 $\{0, 01\}$，这并不是无前缀的，但可以唯一解码.*

但事实上，考虑可唯一译码，可以归约到考虑无前缀码，这是由如下的定理保证的：

**定理 1.0.2.** *记 $A$ 表示所有无前缀码，$B$ 表示所有可唯一译码，显然有 $A \subsetneq B$. 用 $l(C, m)$ 表示使用码 $C$ 编译信息 $m$ 时的码长，那么可以证明*

$$\min_{C \in B} \mathbb{E}_{m \sim P}\Big[l(C, m)\Big] = \min_{C \in A} \mathbb{E}_{m \sim P}\Big[l(C, m)\Big]$$

*证明.* 此略去. 主要思想是证明对于任意 $C^* \in B$ 使得 $\mathbb{E}l(C^*)$ 达到最小值，都存在一个对应的 $\tilde{C}^* \in A$ 使得 $\mathbb{E}l(\tilde{C}^*) = \mathbb{E}l(C^*)$. □

有了上面的基础理论，我们来尝试求无前缀码的最短平均码长。首先，Kraft 不等式为我们估计了一个下界.

**定理 1.0.3** (无前缀码的 Kraft 不等式). *假设码 $C = (c_1, \ldots, c_n)$ 是无前缀的，记 $l_1, \ldots, l_n$ 分别是 $c_1, \ldots, c_n$ 的长度（此指比特数），则*

$$\sum_{i=1}^{n} 2^{-l_i} \leqslant 1$$

*证明.* 这个定理的证明是简单的，构造一棵二叉树，根据码的内容嵌入该二叉树。由于 $C$ 是无前缀码，所以不存在一个码对应的结点是另一个码对应的节点的祖先。故所有码 $c_i$ 都在叶结点上，再归纳地证明满二叉树叶节点的 $\sum 2^{-d_i} = 1$ 即可. □

接下来,我们给定信息为 $M = \{m_1, \ldots, m_n\}$,对应的分布列为 $P = (p_1, \ldots, p_n)$，其中 $p_i = \Pr[m_i]$。我们的目标是找到一个无前缀码 $C = (c_1, \ldots, c_n)$，其长度为 $l_1, \ldots, l_n$。使得平均码长最短，可见这事实上是一个优化问题：

$$\begin{aligned} \underset{l}{\text{minimize}} \ & \sum_{i=1}^{n} p_i l_i \\ \text{s.t.} \ & \sum_{i=1}^{n} 2^{-l_i} \leqslant 1 \\ & l_i \in \mathbb{N}, i = 1, 2, \ldots, n \end{aligned}$$

但这样的问题还是有些复杂，我们不妨去掉整数的约束，并且直接限定 Kraft 不等式中等号成立：

$$\begin{aligned} \underset{l}{\text{minimize}} \ & \sum_{i=1}^{n} p_i l_i \\ \text{s.t.} \ & \sum_{i=1}^{n} 2^{-l_i} = 1 \\ & l_i \geqslant 0, i = 1, 2, \ldots, n \end{aligned}$$

使用 Jensen 不等式或 Lagrange 乘子法（此略去）易得

$$\min_l \sum_{i=1}^{n} p_i l_i = \sum_{i=1}^{n} p_i \log_2 1/p_i$$

由此延伸出**熵**的定义：

**定义 1.0.4** (熵). *给定随机元 $X$，及其分布列 $(p_1, \ldots, p_n)$，定义 $X$ 的*信息熵 *(entropy) 为*

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

在不导致歧义的情况下，可以简称信息熵为熵。这个量有如下意义和特性：

- 最短编码长度，即描述该变量所需的长度下界
- 量化了“信息”这一概念
- 均匀分布时 $H(X)$ 最大（最不确定，无先验），而退化分布时 $H(X) = 0$
- 度量了 $X$ 的不确定性

$n = 2$ 时：

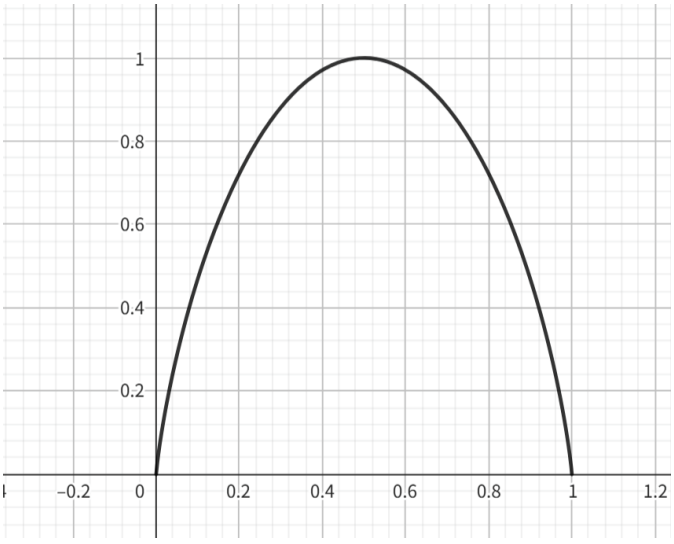



图 1.1: $H = x \log 1/x + (1-x) \log 1/(1-x)$ 的图像

# 第二章　Lecture 3

讲师：王立威　　课程时间：25.Mar.4th　　笔记：25.June.7th

考虑一般的 $p$ 进制编码（经典比特为 $p=2$ 进制，但也有 3 进制计算机）此时，随机元 $X$ 的熵为

$$H(X) := \sum_{i=1}^{n} p_i \log_p 1/p_i$$

相应地，这个量的单位也应该从 bit 变为其他的单位.

上节课说到均匀分布时熵最大，可以总结成如下命题

**命题 2.0.1.** *对于 $X=(m_1,\ldots,m_n)$，那么 $H(X)\leqslant \log_2 n$*

*证明.* 使用 Jensen 不等式即可，此略. □

事实上，熵的定义非常符合直觉，我们通过一个例子感受

**例 2.0.2.** *我们考虑随机变量 $X,Y,Z$. $X$ 有分布列 $(p_1,\ldots,p_n)$，其中 $p_n=q_1+q_2$；$Y$ 的分布列正比于 $(q_1,q_2)$；$Z$ 有分布列 $p_1,\ldots,p_{n-1},q_1,q_2$.*

*可以写成如下形式：*

$$\begin{array}{lccccc} X: & p_1 & p_2 & \cdots & p_n(= & q_1+q_2) \\ Y: & & & & \dfrac{q_1}{q_1+q_2} & \dfrac{q_2}{q_1+q_2} \\ Z: & p_1 & p_2 & \cdots & q_1 & q_2 \end{array}$$

*此时 $H(X),H(Y),H(Z)$ 满足什么关系？*

解答. 应该满足 $H(X)+p_n\cdot H(Y)=H(Z)$，下面通过计算来说明.

$$\begin{aligned}
&H(Z)-H(X)\\
=&\left[\sum_{i=1}^{n-1}p_i\log_2 1/p_i+\sum_{j=1}^{2}q_j\log_2 1/q_j\right]-\left[\sum_{i=1}^{n-1}p_i\log_2 1/p_i+p_n\log_2 1/p_n\right]\\
=&\sum_{j=1}^{2}q_j\left(\log 1/q_j-\log_2 1/p_n\right)\\
=&\sum_{j=1}^{2}q_j\log_2\frac{q_1+q_2}{q_j}\\
=&\,p_n\sum_{j=1}^{2}\frac{q_j}{q_1+q_2}\log_2\frac{q_1+q_2}{q_j}=p_n\cdot H(Y)
\end{aligned}$$

通过上面的例子，我们发现熵的定义十分合理.$Y$ 可以视作对于 $X$ 在 $m_n^{(X)}$ 的情况下的进一步阐释（细化）；而 $Z$ 直接融合了 $X$ 和 $Y$ 对其的阐释. 阐释发生的概率是 $p_n$，所以 $Z$ 的信息量就是 $X$ 的信息量加上 $p_n$ 倍 $Y$ 的信息量. ■

以上结论对于 $\boldsymbol{q}$ 是更高维的情况下也成立，称作**熵的加性**.

我们至此讨论了熵的定义和性质，但回归到实际实践当中，我们当然需要 $l_i$ 全部为整数. 换言之，对于随机变量 $X$ 服从分布列 $P=(p_1,\ldots,p_n)$，我们想要构造一个具体的编码算法，得到平均码长最短的实际编码 $C=(c_1,\ldots,c_n)$，称**最优码**. 以下，我们记 $c_i$ 的长度为 $|c_i|$

可以得到最优码的一些性质：

- 如果 $p_1\geqslant p_2\geqslant\cdots\geqslant p_n$，那么对于最优码一定有 $|c_1|\leqslant|c_2|\leqslant\cdots\leqslant|c_n|$
- Kraft 不等式取等条件成立：$\sum_{i=1}^{n}2^{-|c_i|}=1$
- 如果 $p_1\geqslant p_2\geqslant\cdots\geqslant p_n$，那么会有 $|c_n|=|c_{n-1}|$. （直观而言，这是在说二叉树中 $c_n$ 一定会有兄弟；数学而言，这是为了满足上一条性质的等号要求）

- 若 $(c_1,\ldots,c_n)$ 是 $(p_1,\ldots,p_n)$ 的最优码，那么 $(c_1,\ldots,\tilde{c}_{n-1})$ 是 $(p_1,\ldots,p_{n-1}+p_n)$ 的最优码. 其中我们假定 $c_{n-1},c_n$ 在二叉树中是兄弟，仅有最后一位不同，而 $\tilde{c}_{n-1}$ 代表它们二者前面的公共前缀.（该性质可通过反证法得出）

结合这些性质便可以设计算法，逐步将问题的维度降低. 这个算法就是 Huffman 编码（自行查阅该算法，此不介绍）.

接下来拓展一下熵的定义，对于多个变量可定义联合熵.

**定义 2.0.3** (联合熵). *对于两个随机变量 $X,Y$，联合分布列为 $P_{X,Y}=(p_{i,j})$，定义***联合熵** *(joint entropy) 为*

$$H(X,Y):=\sum_{i,j}p_{i,j}\log_2 1/p_{i,j}$$

可见这样的定义和原本熵的定义很类似，意义是将 $X,Y$ 一起编码时的最短平均码长；$X,Y$ 一共的信息量等等.

**定理 2.0.4.** *对于两个随机变量 $X,Y$，有 $H(X,Y)\leqslant H(X)+H(Y)$，取等当且仅当 $X,Y$ 独立.*

**例 2.0.5.** *取 $X=Y$，可以发现 $H(X,Y)=H(X)\leqslant 2H(X)=H(X)+H(Y)$*

类似地，可以定义条件熵

**定义 2.0.6** (条件熵). *对于两个随机变量 $X,Y$，对于 $X$ 的可能取值 $x_i$（满足 $\Pr[X=x_i]>0$），定义*

$$H(Y|X=x_i):=\sum_j \Pr[Y=y_j|X=x_i]\log_2\frac{1}{\Pr[Y=y_j|X=x_i]}$$

*总体的***条件熵** *(conditional entropy) 就定义为*

$$H(Y|X):=\sum_i \Pr[X=x_i]H(Y|X=x_i)$$

以下定理说明条件熵 $H(Y|X)$ 的意义为“$Y$ 在 $X$ 的基础上引入的新信息”：

**定理 2.0.7.** *对于两个随机变量 $X,Y$，有*

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

# 第三章　Lecture 4

讲师：王立威　　课程时间：25.Mar.11th　　笔记：25.June.7th

接着上节课的内容，根据上节课的两个定理可以得到：

$$H(Y) \geqslant H(Y|X), \qquad H(X) \geqslant H(X|Y)$$

这说明原有的信息量永远不小于条件下的信息量. 这称作"Conditioning reduces entropy."，可以认为这是因为加上条件等价于在某种程度上引入了信息，从而降低了“新”的信息量.

事实上，我们能推出

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

据此引出新的定义

**定义 3.0.1** (互信息). *对于两个随机变量 $X, Y$，定义其*互信息 *(mutual information) 为*

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

互信息是可以用联合概率分布列表示的，表示法如下（请自行验证）：

$$I(X;Y) = \sum_{i,j} \Pr[X = x_i, Y = y_j] \log_2 \frac{\Pr[X = x_i, Y = y_j]}{\Pr[X = x_i] \cdot \Pr[Y = y_j]}$$

自然地，可类似定义任意多元随机变量 $\boldsymbol{X} = (X_i)_{i=1}^n$ 的联合熵 $H(\boldsymbol{X})$；也可定义两组随机变量 $\boldsymbol{X} = (X_i)_{i=1}^n, \boldsymbol{Y} = (Y_j)_{j=1}^m$ 的条件熵 $H(\boldsymbol{X}|\boldsymbol{Y})$ 和互信息 $I(\boldsymbol{X};\boldsymbol{Y})$.

**定理 3.0.2.** 对于多元随机变量 $\boldsymbol{X} = (X_i)_{i=1}^n$，有

$$H(\boldsymbol{X}) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_{1\sim(n-1)})$$

至此的讨论都基于随机变量 $X$ 的分布列 $P$ 是已知的，但在真实世界中我们往往无法精确观测或获取真实分布列. 但取而代之，我们往往有一个估计（近似）$Q = (q_1, \ldots, q_n)$，此时我们依照这个分布列编码，码长应为 $(\log 1/q_1, \ldots, \log 1/q_n)$，那么平均码长为 $\sum_{i=1}^n p_i \log 1/q_i$. 相较与最短码长 $H(P)$，这样的编码存在冗余 $\sum_{i=1}^n p_i \log p_i/q_i$，这就是著名的 K-L 散度.

**定义 3.0.3** (Kullback-Leibler 散度). 对于两个分布列 $P = (p_1, \ldots, p_n)$ 和 $Q = (q_1, \ldots, q_n)$，其 ***Kullback-Leibler* 散度** *(divergence)* 定义为

$$D(P\|Q) := \sum_{i=1}^n p_i \log_2 \frac{p_i}{q_i}$$

*Kullback-Leibler* 散度也称作 *K-L* 散度或相对熵 *(relative entropy).* 并且也可记作 $KL(P\|Q)$.

某种意义而言，K-L 散度是比信息熵更加基础和重要的概念. 在以后的例子中，我们会发现在有些情况下难以定义信息熵，但却可以定义 K-L 散度.

另外，请注意 K-L 散度不是对称的，故不是距离度量.

**例 3.0.4** (K-L 散度的凸性). 考虑以下情况：固定 $Q$，$D$ 对 $P$ 是凸的吗？固定 $P$，$D$ 对 $Q$ 是凸的吗？$D$ 对 $(P, Q)$ 是凸的吗？

解答. 事实上，K-L 散度是凸的. ■

除了 K-L 散度，我们还有别的定义分布列距离的方式，一个比较经典的是全变分距离（1-范数）

**定义 3.0.5** (全变分距离). *对于两个分布列 $P = (p_1, \ldots, p_n)$ 和 $Q = (q_1, \ldots, q_n)$，其*全变分距离 *(total variance) 定义为*

$$\|P - Q\|_1 := \sum_{i=1}^{n} |p_i - q_1|$$

KL-散度和全变分距离之间有如下关系：

**定理 3.0.6** (Pinsker 不等式). *对于两个分布列 $P, Q$，有*

$$\|P - Q\|_1 \leqslant \sqrt{2 \ln 2 \cdot D(P\|Q)}$$

*证明.* 首先我们对最简单的 Bernoulli 分布形式进行验证. 也就是设 $P = (p, 1-p), Q = (q, 1-q)$，此时通过暴力计算可以验证 Pinsker 不等式.

对于一般情况的 $P, Q$，我们定义：

$$\tilde{P} = \left( \sum_{i:p_i \geqslant q_i} p_i, \sum_{i:p_i < q_i} p_i \right), \quad \tilde{Q} = \left( \sum_{i:p_i \geqslant q_i} q_i, \sum_{i:p_i < q_i} q_i \right)$$

不难验证 $\|P - Q\|_1 = \left\| \tilde{P} - \tilde{Q} \right\|_1$. 而另一方面，根据 K-L 散度的凸性可以知道 $D(\tilde{P}\|\tilde{Q}) \leqslant D(P\|Q)$

而 $\tilde{P}, \tilde{Q}$ 都是 Bernoulli 分布，我们通过先前的计算已经获知其满足 Pinsker 不等式，于是

$$\|P - Q\|_1 = \left\| \tilde{P} - \tilde{Q} \right\|_1 \leqslant \sqrt{2 \ln 2 \cdot D(\tilde{P}\|\tilde{Q})} \leqslant \sqrt{2 \ln 2 \cdot D(P\|Q)}$$

至此证毕. □

# 第四章　Lecture 5

讲师：王立威　　课程时间：25.Mar.18th　　笔记：25.June.7th

首先讨论了 Pinsker 不等式的证明方法，我直接放在了上一章当中.

接下来看新的内容. 我们先来看一个例子来深入理解最短编码和信息熵之间的差距. 也就是要求整数和不要求整数之间的差距.

**例 4.0.1.** *随机变量 $X$，分布列 $P = (0.01, 0.99)$*

- *信息熵 $H(X) \approx 0.07 bit.$*
- *最短编码 1 bit.*

事实上，我们有如下推论

**推论 4.0.2.** *对于任意随机变量 $X$，最短编码和信息熵的差严格小于 1.*

因此，通过差值来衡量编码效率是不合适的，我们应该使用比值来衡量. 另一方面，我们设定要传输可列次信息，也就是说存在一个 i.i.d 的序列 $X_1, X_2, \ldots$ 和 $X$ 同分布. 我们传输这些信息时，不妨每 $T$ 个打包成一组进行传输，可见这样一次传输的平均最小码长 $l_{\min}$ 满足

$$l_{\min}(X_1, \ldots, X_T) \leqslant H(X_1, \ldots, H_T) + 1 = T \cdot H(X) + 1$$

于是比值

$$\text{Ratio} = \frac{l_{\min}(X_1, \ldots, X_T)}{H(X_1, \ldots, H_T)} \leqslant 1 + \frac{1}{T \cdot H(X)} = 1 + O\left(\frac{1}{T}\right)$$

平均到编码每一个信息所需的长度有上界 $\frac{TH(X)+1}{T} \to H(X)$. 因此，在 $T \to \infty$ 时，最终的平均码长应该趋于 $H(X)$，而非 $l_{\min}(X)$. 这也是为什么我们认为信息熵是一个本质的定义.

值得指出，这样的操作在实际实践中会带来更大的时延.

上面的模型其实对实际情况做了一个简化，也就是假设每个时刻的信息都是 i.i.d 的. 为了泛化我们的模型，我们应该将信源建模成随机过程 (stochastic process)$\mathcal{X} = (X_t)_{t\geqslant 1}$，下面讨论这个模型.

首先需要定义这个信源的平均码长.

**定义 4.0.3** (熵率). *对于随机信源 $\mathcal{X} = (X_t)_{t\geqslant 1}$，定义其*熵率 *(entropy rate) 为*

$$H(\mathcal{X}) := \lim_{T\to\infty} \frac{1}{T} \cdot H(X_1, X_2, \ldots, H_T)$$

*我们假定上述极限存在（这个技术细节过于琐碎，略去）*

另一种角度而言，我们也可以认为 $H(\mathcal{X})$ 是每个时刻增加的新的信息.

**定理 4.0.4.** *给定信源 $\mathcal{X} = (X_t)_{t\geqslant 1}$，假设以下极限存在，那么下式成立.*

$$H(\mathcal{X}) = \lim_{T\to\infty} H(X_T | X_1, X_2, \ldots, H_{T-1})$$

*证明.* 注意到

$$H(X_1, X_2, \ldots, H_T) = H(X_1) + \sum_{i=2}^{T} H(X_i | X_{1\sim(i-1)})$$

使用数学分析 (I) 中的结论便可完成证明. □

接下来考虑连续随机变量的编码，但显然它没有所谓“最短平均码长”这一概念，我们只能先形式上给出一个熵的定义：

**定义 4.0.5** (微分熵). *对于连续型随机变量 $X$，设其 p.d.f. 为 $f(x)$，则定义其*微分熵 *(differential entropy) 为*

$$h(X) := -\int_{\mathbb{R}} f(x) \log_2 f(x) \, \mathrm{d}x$$

这个形式上定义并非单纯的望文生义，而是真正的有本之木！我们考虑连续型随机变量 $X$ 对应的离散化版本 $X_\Delta$，其中 $\Delta > 0$ 是一个离散化的粒度. 具体而言，$X_\Delta$ 只取值于 $k\Delta$, $(k \in \mathbb{Z})$，并且

$$\Pr[X_\Delta = k\Delta] = \Pr[k\Delta \leqslant X < (k+1)\Delta] = \int_{k\Delta}^{(k+1)\Delta} f(x)\,\mathrm{d}x$$

可以得到如下的近似

**命题 4.0.6.** *对于连续型随机变量 $X$ 对应的离散化版本 $X_\Delta$，有*

$$h(X) + \log\frac{1}{\Delta} \approx H(X_\Delta)$$

另外，微分熵和香农熵在性质上也有一些不同的地方. 我们知道对于离散随机变量 $X$，

- 若 $Y = X + c$，则 $H(X) = H(Y)$
- 若 $Z = aX (a > 0)$，则 $H(X) = H(Z)$

但对于连续随机变量和微分熵，对应的结论会变为

**命题 4.0.7.** *对于连续型随机变量 $X$，若 $Y = X + c$（其中 $c$ 是常数），那么*

$$h(Y) = h(X)$$

**命题 4.0.8.** *对于连续型随机变量 $X$，若 $Z = aX$（其中 $a > 0$ 是常数），那么*

$$h(Z) = h(X) + \log a$$

下面举隅一例来计算其微分熵.

**例 4.0.9.** *已知随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$，求 $h(X)$.*

*解答.* 我们记 $X$ 的概率密度函数为 $f(x)$，有

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

因此，

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{\infty} f(x)\log_2 f(x)\,\mathrm{d}x \\
&= -\int_{-\infty}^{\infty} f(x)\cdot\left(-\log_2(\sigma\sqrt{2\pi})-\frac{(x-\mu)^2}{2\ln 2\cdot\sigma^2}\right)\mathrm{d}x \\
&= \log_2(\sigma\sqrt{2\pi})\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x+\frac{1}{2\ln 2\cdot\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,\mathrm{d}x \\
&= \log_2(\sigma\sqrt{2\pi})+\frac{1}{2\ln 2\cdot\sigma^2}\mathsf{Var}(X) \\
&= \log_2(\sigma\sqrt{2\pi})+\frac{1}{2\ln 2} \\
&= \frac{1+\log_2(\pi e\sigma^2)}{2}
\end{aligned}$$

至此便得到了答案. ■

# 第五章　Lecture 6

讲师：王立威　　课程时间：25.Mar.25th　　笔记：25.June.7th

接着将之前对于离散型随机变量定义的各种熵迁移到连续型随机变量上来

**定义 5.0.1** (联合微分熵). *对于两个连续型随机变量 $X,Y$，设其联合 p.d.f. 为 $f_{X,Y}(x,y)$，定义其***联合微分熵** *(joint differential entropy) 为*

$$h(X,Y) := -\int\int f_{X,Y}(x,y)\log f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

**定义 5.0.2** (条件微分熵). *对于两个连续型随机变量 $X,Y$，设其联合 p.d.f. 为 $f_{X,Y}(x,y)$，定义其***条件微分熵** *(conditional differential entropy) 为*

$$h(X|Y) := -\int\int f_{X,Y}(x,y)\log f_{X|Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

下面来讨论一个很重要的量：互信息。

**定义 5.0.3** (连续型随机变量的互信息). *对于两个连续型随机变量 $X,Y$，具有概率密度 $f_{X,Y}, f_{X|Y}, f_{Y|X}$ 等，定义它们的***互信息** *(mutual information) 为*

$$I(X;Y) := h(X) - h(X|Y)$$

这样从形式上定义的互信息竟然保留了离散版本的诸多性质！

**定理 5.0.4.** *对于两个连续型随机变量 $X,Y$，一定有*

$$I(X;Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y)$$

另外，还可以用联合概率密度和边缘概率密度来表达

$$I(X;Y)=\int\int f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\,\mathrm{d}x\,\mathrm{d}y$$

为什么会造成这样的性质呢？我们考虑它们二者的对应的离散化版本 $X_\Delta$ $Y_\Delta$，会有

$$I(X;Y)=\lim_{\Delta\to 0^+} I(X_\Delta,Y_\Delta)$$

这说明虽然在连续意义下，微分熵的物理意义不是非常明确，但互信息的物理意义是充分明晰的。

另外，也可以接着定义 K-L 散度（相对熵）

**定义 5.0.5.** 给定 $f,g$ 分别是连续型随机变量 $X,Y$ 的概率密度函数，定义 ***K-L* 散度**为

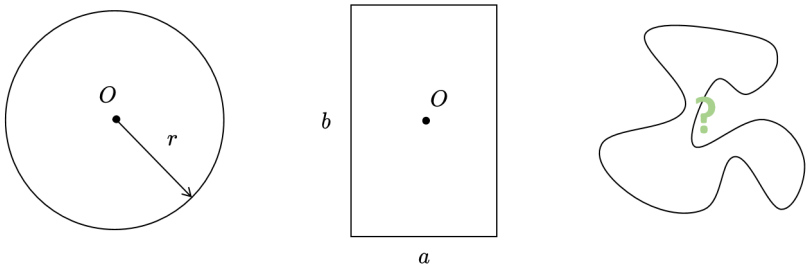
$$KL(f\|g):=\int f(x)\log\frac{f(x)}{g(x)}\,\mathrm{d}x$$

继续考虑离散化版本 $X_\Delta,Y_\Delta$ 以及他们对应的分布列 $P_\Delta,Q_\Delta$。可以发现

$$\lim_{\Delta\to 0^+}\mathrm{KL}(P_\Delta\|Q_\Delta)=\mathrm{KL}(f\|g)$$

至此，关于信源编码就告一段落了。

现在考虑编码的问题，对于一个随机对象，希望知道最小的描述长度。但这个问题是不平凡的，因为就算是退化到了确定对象也难以严格定义“描述长度”。好比对于圆，我们可以使用圆心和半径；对于矩形可以使用中心、长和宽；但对于一个比较混沌的图形就难以描述了。




$(O,r)$ $(O,a,b)$

这个所谓的“对象”空间太大了，需要限制一下范围，并挖掘本质。前苏联的数学家 Kolmogorov 给出了回答，将其限制在了字符串 $\{0,1\}^*$ 当中。

**例 5.0.6.** *直觉上，可以将 $000\cdots0$ 描述成 $n$ 个 0；将 $011011011\cdots$ 描述成 $m$ 个 011；但如果原字符串很混乱，那最好的描述不过是将其复述一遍。*

这样的直觉需要转为严谨的定义，思考一下就会发现我们需要发掘“计算”的本质。这个问题 Turing 已经回答了，是“Turing 机”（如今的 Boole 线路与这个模型也是等价的）。那么所有问题都一定有解吗？并不是，证明也很简单。

**定理 5.0.7.** *我们考虑所有的决定问题，即给定集合 $L \subseteq \{0,1\}^*$，需要对任意的 $x \in \{0,1\}^*$ 判断 $x$ 是否属于 $L$。一定存在某个 $L_0$ 对应的决定问题不存在 Turing 机上可以解决该问题的算法。*

*证明.* 记所有的问题构成的集合基数为 $\kappa$，记所有算法构成的集合基数为 $\lambda$。对于 Turing 机，算法本质上就是一个字符串，所以 $\lambda$ 是 $\{0,1\}^*$ 的基数。而上面的讨论说明了一个 $L$ 可以决定一个问题，故 $\kappa$ 是 $\mathcal{P}(\{0,1\}^*)$ 的基数。

根据 Cantor 定理，有

$$\lambda = \operatorname{card}\{0,1\}^* < \operatorname{card}\mathcal{P}\Big(\{0,1\}^*\Big) = \kappa$$

基数不同，所以不存在双射。因此一定存在一个不可解的问题。 □

我们可以讲出这样的不可计算问题，例如说“停机问题”。可以证明：给定算法 $A$ 和输入 $i$，不存在算法保证一定可以计算 $A$ 是否可以在有限步内停下。

Turing 的工作其实在一定程度上收到了 Godel 的启发，在 Turing 关于计算本质问题的研究工作发表前，Godel 证明了任意一阶逻辑的不完备性，也正是熟知的“Godel 不完备定理”.

# 第六章　Lecture 7

讲师：王立威　　课程时间：25.Apr.1st　　笔记：25.June.8th

接着上一课程未解决的问题，我们如何定义一个确定对象的“最短描述长度”。此前，我们已经将这个对象的范围缩小到 $x \in \{0,1\}^*$，加上 Turing 给出的计算模型，我们已经具备定义 Kolmogorov 复杂度的能力了！（关于下面用到的通用 Turing 机 (universal Turing machine) 这一概念，请自行浏览资料）

**定义 6.0.1** (Kolmogorov 复杂度). *对于一个字符串 $x \in \{0,1\}^*$，其对于一个通用 Turing 机 $U$ 的* ***Kolmogorov* 复杂度** *(complexity) 定义为*

$$K_U(x) := \min_{p,U(p)=x} |p|$$

*也就是 $U$ 要决定 $x$ 所需的最短“代码”长度，这也简称为 $K$-复杂度，并简写为 $K(x)$.*

若没有学过 Turing 机相关的知识，可以暂时理解成 $U$ 是某个固定的编程语言，例如 C++ 或 Python 等。以下命题说明了上面将 $U$ 简写是合理的：

**命题 6.0.2.** *设 $U, U'$ 是两个通用 Turing 机，那么存在一个常数 $c \in \mathbb{N}$ 使得对于任意 $x \in \{0,1\}^*$，如下不等式成立*

$$K_U(x) \leqslant K_{U'}(x) + c$$

*其中 $c$ 只依赖于 $U, U'$ 而不依赖于 $x$.*

但这个定义听起来很难实践。一种尝试计算 $x$ 的 K-复杂度的方式，是从根据长度小到大遍历所有的 $p$（由于 $K_U(x)$ 一定存在有界，所以需要遍历的 $p$ 是有限多的），但这会遇到一个难题：“停机问题”。我们无法判断对于特定的 $p$，$U$ 是否会停机。归根结底，这是因为 $K_U$ 本质上是不可以计算的！

**定理 6.0.3.** *$K_U(x)$ 是不可计算的.*

*证明.* 这里只给出大致的思想，UTM 的细节略去。假设存在一个算法 $P_0$，可以对于任意 $x$ 计算 $K_U(x)$.

记 $|P_0| = l$，下面给出一个新的算法 $P_0'$。首先我们将 $\{0,1\}^*$ 中的元素排序成 $s_1, s_2, \ldots, s_i, \ldots$。

$P_0'$ 如下：

- 令变量 $i$ 遍历 $1, 2, \ldots$
  - 使用 $P_0$ 计算 $K_U(s_i)$
  - 若 $K_U(s_i) \geqslant L$，则输出 $s_i$

那么可见 $|P_0'| = l + \alpha \log L + \beta$（$\alpha, \beta$ 常数）. 而 $P_0'$ 可以决定识别字符串 $s$，其中 $K_U(s) \geqslant L$。但根据 K-复杂度的定义，会有

$$L \leqslant K_U(s) \leqslant |P_0'| = l + \alpha \log L + \beta$$

所以取 $L$ 充分大便可得到矛盾. 因此 $K_U$ 不可计算. □

关于这部分告一段落，我们来看概率分布估计的问题，也就是说对于一些函数 $g_i$，已知 $\mathbb{E}_X \sim P[g_i(X)]$ 的值，求原始分布 $P$。这类问题往往具有无穷多个解，我们应当关心其中某个具有优良性质的。

**定义 6.0.4** (最大熵估计). *给定一族函数 $\{g_i(x)\}_{i\in I}$ 和 $\{r_i\}_{i\in I}$。**最大熵估计**指的是如下问题的解*

$$\begin{aligned} \underset{f}{\text{maximize}} \ & -\int f(x)\ln f(x)\,\mathrm{d}x \\ s.t. \ & \int f(x)\,\mathrm{d}x = 1 \\ & \int g_i(x)f(x)\,\mathrm{d}x = r_i \quad , i \in I \end{aligned}$$

这里将所谓“熵”从 $\log_2$ 替换成 ln 是因为他们只差一个倍数。我们试着求解一个最大熵问题：

**例 6.0.5.** *给定一个连续型随机变量 $X$ 及其 p.d.f. $f(x)$，已知*

$$\mathbb{E}[X] = 0, \qquad \mathsf{Var}(X) = \sigma^2$$

*求最大熵分布.*

*解答.* 一般而言，这类问题的通法时使用 Lagrange 乘子和变分。不过这里我们可以给出一个使用 K-L 散度的解法。

■

# 第七章　Lecture 8

讲师：王立威　　课程时间：25.Apr.8th　　笔记：25.June.8th

来讨论信道编码，动机在于真实世界的信道都是含噪的，噪声会干扰信息的传递致使接收到的信息和发送的信息不完全相同. 自然地，我们想要设计一套算法来纠正由噪声引起的错误，这套算法应该包含编码 (encoding) 和解码 (decoding) 两部分.

**例 7.0.1.** 一个最简单的想法就是编码时重复三次，解码时取众数.

上面的例子虽然简单，但却内蕴了深刻的思想——解码时将内容映射到最近邻的合理码字. 这里的最近邻使用的是 Hamming 距离（请自行查阅）.

我们设 $M_i$ 对应的编码是 $c_i$（对于所有 $1 \leqslant i \leqslant n$），那么有如下定理：

**定理 7.0.2.** 对于正整数 $t > 0$，若对于任意 $i \neq j$ 都有 $d_H(c_i, c_j) \geqslant 2t + 1$，那么这样的编码可以纠正 $t$ 比特的错误.

我现在希望将信息空间映射到编码空间

$$\text{Message Space} \quad \{0,1\}^m \qquad \longrightarrow \qquad \text{Coding Space} \quad \{0,1\}^n$$

现在给定 $t, m$，来估计一下 $n$ 的下界.

**命题 7.0.3.** 上述条件下，需要满足

$$2^n \geqslant 2^m \cdot \sum_{i=1}^{t} \binom{n}{i}$$

证明. 统计一下球形邻域 $B(x) := \{y : d_H(x,y) \leqslant t\}$ 的大小，注意到对于任意 $i \neq j$，要有 $B(c_i) \cap B(c_j) = \varnothing$ □

接着来计算其上界. 这里我们将目标重述为给定 $m$ 和 $\delta \in (0, 1/2)$，找一个尽量小的 $n$ 使得一定存在编码 $c_1, c_2, \ldots, c_{2^m} \in \{0,1\}^n$ 使得

$$d_H(c_i, c_j) \geqslant \delta n, \quad \forall i \neq j$$

采用概率方法（这是现代组合数学的常用技巧）可以证明如下结论.

**定理 7.0.4** (Gilbert-Vashamov 界). *如果对于某个常数 $c$（$c$ 依赖于 $\delta$），$n \geqslant cm$，那么存在一组编码符合条件.*

证明. 我们在 $\{0,1\}^n$ 上独立且均匀地采样 $c_1, c_2, \ldots, c_n$，相当于每个比特都服从均匀两点分布. 故 $d_H(c_i, c_j) \sim B(n, 1/2)$. 根据 Chernoff Bound，可以得到

$$\Pr\Big[d_H(c_i, c_j) < \delta n\Big] < e^{-\ln 2 D(\delta \| 1/2) \cdot n} = e^{-O(n)}$$

根据 Union Bound，又有

$$\Pr\left[\bigcup_{i,j \in [2^m] : i \neq j} d_H(c_i, c_j) < \delta n\right] < \binom{2^m}{2} e^{-O(n)}$$

所以

$$\Pr\Big[\forall i \neq j \in [2^m],\ d_H(c_i, c_j) \geqslant \delta n\Big] \geqslant 1 - \binom{2^m}{2} e^{-O(n)}$$

我们只需要上式右侧大于 0 即可，化简可以得到这需要

$$n \geqslant \Omega(m)$$

也就证明了结论. 事实上，这个 $c$ 可以取成 $\dfrac{2}{D(\delta \| \frac{1}{2})}$. □

总结一下设计一套纠错码的流程：

1. 首先要设计一组码字 $c_1, c_2, \ldots, c_N$，使得它们彼此 Hamming 距离充分大（达到预期纠错的能力）.

2. 然后给出一个编码算法，将消息映射到码字.

3. 最后给出解码算法，找到接受消息最近的码字.

这里的第 3 步比较特殊，一般 $m$ 可以来到上百的量级，码字计算会来到 2 的上百次方量级，朴素寻找最近邻码字是不现实的. 我们必须要考虑编码和解码的计算复杂度. 这部分也是该领域着重努力的方向.

一个最早期的做法是 Hamming 码，在介绍 Hamming 码之前，先通过一个例子感受一下. 在陈述例子之前，我们先定义零空间.

**定义 7.0.5.** *对于域 $F$ 上的线性映射 $T : U \to V$，定义其*零空间 *(null space) 为*

$$\mathrm{Null}(T) := \{x \in U : Tx = 0\}$$

*如果 $T$ 有矩阵表示 $A$，那么也称为矩阵的零空间 $\mathrm{Null}(A)$.*

**例 7.0.6.** *考虑 $GF(2)$ 上的矩阵*

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}_{3\times 7}$$

*试计算* $\min\limits_{x,y\in\mathrm{Null}(H):x\neq y} d_H(x,y)$.

我们用列向量表示矩阵 $H = \begin{bmatrix} v_1, & v_2, & \cdots & , v_7 \end{bmatrix}$. 可以看出 $v_i$ 事实上就是 $i$ 的二进制表示. 可以看出这个 7 个向量张成的空间是 3 维的.

不难验证 $\mathrm{Null}(H)$ 事实上是 $[2]^7$ 的一个线性子空间. 而 $d_H(x,y)$ 其实是 $x+y$ 为 1 的比特数，也就是 1-范数. （这里 $+$ 是 $GF(2)$ 中的运算符）这里我们称 $\|x\|_1$ 为 $x$ 的**权重**. 而若

$$Hx = \sum_{i=1}^{7} v_i x_i = 0$$

结合之前关于维度的讨论可知 $\|x\|_1 \geqslant 3$，所以 $d_H(x,y) \geqslant 3$. 另一方面取

$$x = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^\top, \quad y = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}^\top$$

可以达到下界，故答案就是 3.

# 第八章　Lecture 9

讲师：王立威　　课程时间：25.Apr.15th　　笔记：25.June.9th

首先回答了上一节的最后一个例子，我已经放在了前面的讨论中.

这个例子其实给了我们一个设计上的启发，Null($H$) 中一共有 $2^{7-3}=16$ 个元素，将这些元素作为码字可以保证其两两距离均大于等于 3，因此足够纠正 1 比特的错误！由此定义 Hamming 码：

**定义 8.0.1** (Hamming $(n,m)$ 码). *对于某个正整数 $t$，我们记 $1,2,\ldots,2^t-1$ 的二进制列向量表达为 $v_1,v_2,\ldots,v_{2^t-1}$，得到矩阵*

$$H=\begin{bmatrix} v_1, & v_2, & \cdots & , v_{2^t-1} \end{bmatrix}_{t\times 2^t-1}$$

*该矩阵的零空间 Null($H$) 中的所有元素作为码字称作 Hamming$(n,r)$ 码，其中 $n=2^t-1, m=2^t-t-1$.*

上一个例子中的给出的就是 Hamming $(7,4)$ 码，是极其常用的. 并且它完美契合了之前得到的下界.

**定义 8.0.2** (完美码). *对于一个码 $C$，如果达到了球形邻域给出的下界，则称其为*完美码 *(perfect code). 即要满足*

$$2^n=2^m\cdot\sum_{i=1}^{t}\binom{n}{i}$$

**定理 8.0.3.** *Hamming 码是完美码.*

解答. 请读者对 Hamming $(7,4)$ 码自行验证. ■

现在来考虑如何解码，事实上这需要将不在码字中的元素映射到最近邻的码字，如下图所示：

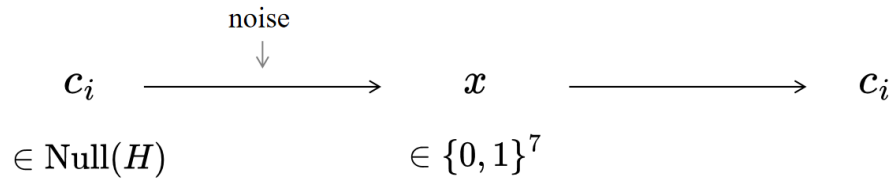



图 8.1: 解码图示

如果 $d_H(x,c_i) \leqslant 1\text{bit}$，那么只有两种情况，如果 $x = c_i$，那么 $Hx = 0$；如果 $x = c_i + \boldsymbol{e}_j$ 其中 $\boldsymbol{e}_j$ 是仅在第 $j$ 位上是 1，其余全是 0 的向量，那么 $Hx = Hc_i + H\boldsymbol{e}_j = v_j$. 对于后者的情况，我们可以轻易从 $v_j$ 中推断出 $j$，只需要取 $c_i = x - v_j$ 即可. 这也说明了我们按二进制排序 $H$ 的列向量的合理性，于是我们称这个 $H$ 为**校验矩阵** (check matrix).

下面考虑编码的方法，一个十分理想的编码是将消息 $m_i$ 映射到码字 $c_i$，其中保证 $m_i$ 是 $c_i$ 的部分前缀（换言之 $c_i$ 的前 $m$ 位就是 $m_i$）我们现在取 $\mathrm{Null}(H)$ 的一组基 $\{g_i\}_i$，这组基的大小应该是 $m$，记

$$G = \Big[g_1,\ \ g_2,\ \ \cdots, g_m\Big]_{n\times m}$$

我们规定编码的映射 $f$ 如下：

$$f(a) = \sum_{i=1}^{4} a_i g_i = Ga \in \mathrm{Null}(H)$$

我们称这样的 $G$ 为**生成矩阵** (generator matrix).

**定理 8.0.4.** *生成矩阵 $G$ 和校验矩阵 $H$ 满足 $HG = O$*

*证明.* 这个定理十分显然，因为 $Hg_i = 0$ 对于任意 $i$. □

但至此我们还没有保证 $c_i$ 的前 $m$ 位就是 $m_i$，但思考一下，我们只需要 $G$ 具有如下形式即可：

$$G = \left[\begin{array}{c} I_{m\times m} \\ \hline \tilde{G}_{(n-m)\times m} \end{array}\right] \quad \Rightarrow \quad Ga = \left[\begin{array}{c} a \\ \hline \tilde{G}a \end{array}\right]$$

当然只要 $G$ 是满秩的，我们就可以将其上部的方阵调整为单位矩阵.

下面接着介绍另一种码，线性码. 它是一种更广泛的定义，记作 $[n,k,d]$ 码，$n$ 是长度，$k$ 是维度，$d$ 是 Hamming 距离. 具体而言：

**定义 8.0.5** (线性码). *一个**线性码** (linear code) 是 $\{0,1\}^n$ 的一个 $k$ 维子空间，满足这个空间中任意两个不同元素之间的 Hamming 距离都大于等于 $d$. 这样的码记作 $[n,k,d]$ 码.*

**例 8.0.6.** *之前提到的 Hamming $(7,4)$ 码，其实是一个 $[7,4,3]$ 码.*

在线性码中，编码的步骤也是这样的，有一个生成矩阵 $G_{n\times k}$，其列向量就是子空间的一组基.

回到原来的 Hamming 码，我们有一组标准的方式构造 $H$ 和 $G$，这样的得到的码称作系统码或标准码：

**定义 8.0.7** (系统码). *一组**系统码** (systematic code) 是由如下形式的 $G,H$ 生成的：*

$$G = \left[\begin{array}{c} I_{m\times m} \\ \hline P_{(n-m)\times m} \end{array}\right], \quad H = \left[\begin{array}{c|c} P_{(n-m)\times m} & I_{(n-m)\times(n-m)} \end{array}\right]$$

当然并不是所有码都构成一个线性空间，也有大量的**非线性码**，一般用 $(n,M,d)$ 表记，其中 $M$ 表示码字的个数（对应线性码中的 $2^m$）. 对于这样的码，方法生成新的码：

**定义 8.0.8** (缩短). *对于一组 $(n,M,d)$ 码 $C$，将其拆分为两个集合*

$$C_0 := \{c : c_n = 0\}, \quad C_1 := \{c : c_n = 1\}$$

则其中必然有一个集合的元素个数大于等于 $M/2$，取该集合中每个元素的前 $n-1$ 位得到新的码 $\tilde{C}$，是 $(n-1,\tilde{M},\tilde{d})$ 的. 一定有 $\tilde{M}\geqslant M/2,\tilde{d}\geqslant d$. 这种方法称作**缩短** *(shorten).*

现在考虑与缩短相反的操作，对于一个 $(n,M,d)$ 码 $C$，满足 $d$ 是奇数. 我们现在对于 $C$ 中的每一个码字 $c_i$，在后面添加一个比特 $b$ 得到 $\tilde{c}_i$，使得 $\|\tilde{c}_i\|_1$ 是偶数，这样一来码字之间的 Hamming 距离也必然是偶数，且不小于 $d$，故不小于 $d+1$. 因此如此得到的码 $\tilde{C}$ 是 $(n+1,M,\tilde{d})$ 的，满足 $\tilde{d}\geqslant d+1$.

**例 8.0.9.** *试将 Hamming* $[7,4,3]$ *码改造为* $[8,4,4]$ *码.*

最后，看一个例子（这里老师有一些混用 $G$ 和 $G^\top$ 的表记，大部分教材上使用的是将 $g_i$ 作为行向量的版本，不过我相信只要提前声明这就是等价的）

**例 8.0.10.** *考察这样的线性码生成矩阵，它和 Hamming* $[7,4,3]$ *码等价：*

$$G=\begin{bmatrix}1&1&0&1&0&0&0\\0&1&1&0&1&0&0\\0&0&1&1&0&1&0\\0&0&0&1&1&0&1\end{bmatrix}$$

这样满足每一行都是上一行的位移的码，称作**循环码** (cyclic code). 有关于循环码的更多内容，若有兴趣请自行查阅.

# 第九章　Lecture 10

讲师：王立威　　课程时间：25.Apr.22nd　　笔记：25.June.9th

关于编码的部分就到此结束，我们转而关注信道传输中的根本问题. 这章的核心是信道容量 (channel capacity). 先总结一下信道通信的基本框架，图示如下：

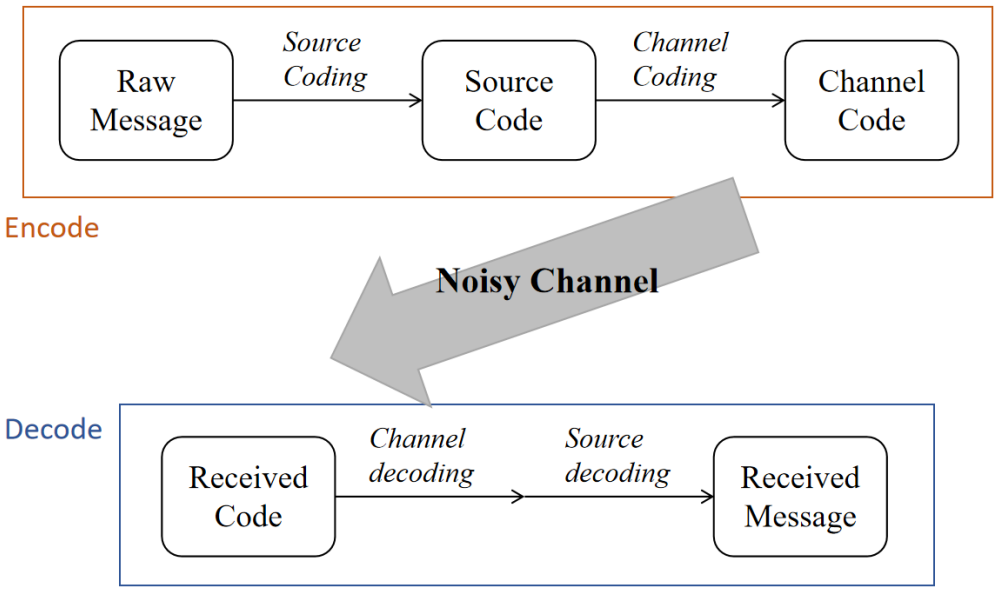



图 9.1: 信道通信的图示

如何刻画一个含噪的信道呢？我们将信源的消息看作随机变量 $X$，通过信道后的信息为 $Y$，那么含噪信道在意义上完全可以由似然（条件概率）$\Pr[Y|X]$ 决定（一下可能使用 $P(Y|X)$ 代替 $\Pr[Y|X]$）.

在这个模型的基础上，当给定一个含噪信道 $P(Y|X)$ 后，接收方能够接受到有多少有关 $X$ 的“信息”. 如果我们单纯用信息熵来刻画是不够合理的：

**命题 9.0.1.** *存在含噪信道 $P(Y|X)$ 使得 $H(Y) > H(X)$.*

这是因为接收方可能收到了很大一部分来自噪声的信息，而非来自 $X$ 的信息，故我们再次强调该信息一定是关于 $X$ 的. 自然我们应该使用互信息而非信息熵来刻画，即使用 $I(X;Y) = D(P_{X,Y}\|P_XP_Y)$.

因此延伸出信道容量的定义：

**定义 9.0.2** (信道容量)**.** *对于一个含噪信道 $P(Y|X)$，其***信道容量*** (channel capacity) 定义为*

$$C := \max_{P_X} I(X;Y)$$

可以看出信道容量的物理意义是有效信息最高传输速率. 为加深理解，来看几则例子.

**例 9.0.3.** *$X, Y$ 均取值于 $\{0,1\}$，若信道是无噪声的，即似然如下，求容量.*

| $P(Y\|X)$ | $X=0$ | $X=1$ |
|---|---|---|
| $Y=0$ | 1 | 0 |
| $Y=1$ | 0 | 1 |

解答. 根据定义有

$$C = \max_{P_X} I(X;Y) = \max_{P_X} H(X) = 1 \text{ bit}$$

■

**例 9.0.4.** *$X, Y$ 均取值于 $\{0,1\}$，若信道是完全翻转的，即似然如下，求容量.*

| $P(Y\|X)$ | $X=0$ | $X=1$ |
|---|---|---|
| $Y=0$ | 0 | 1 |
| $Y=1$ | 1 | 0 |

解答. 这和上面的情况本质上完全一样，所以容量 $C = 1$ bit. ■

**例 9.0.5.** $X$ 取值于 $\{A_1, A_2, B_1, B_2\}$，而 $Y$ 取值于 $\{A, B\}$. 我们考虑信道合并了同类的信息，即似然如下，求容量.

| $P(Y\|X)$ | $X = A_1$ | $X = A_2$ | $X = B_1$ | $X = B_2$ |
|---|---|---|---|---|
| $Y = A$ | 1 | 1 | 0 | 0 |
| $Y = B$ | 0 | 0 | 1 | 1 |

解答. 我们知道 $I(X;Y) \leqslant H(X), H(Y)$，故

$$C = \max_{P_X} I(X;Y) \leqslant \max_{P_X} H(Y) = 1 \text{ bit}$$

取 $P_X(A_1) + P_X(A_2) = 1/2$ 即可. ■

**例 9.0.6.** $X, Y$ 均取值于 $\{A, B, C, D, E\}$，似然如下，求容量.

| $P(Y\|X)$ | $X = A$ | $X = B$ | $X = C$ | $X = D$ | $X = E$ |
|---|---|---|---|---|---|
| $Y = A$ | $1/2$ | 0 | 0 | 0 | $1/2$ |
| $Y = B$ | $1/2$ | $1/2$ | 0 | 0 | 0 |
| $Y = C$ | 0 | $1/2$ | $1/2$ | 0 | 0 |
| $Y = D$ | 0 | 0 | $1/2$ | $1/2$ | 0 |
| $Y = E$ | 0 | 0 | 0 | $1/2$ | $1/2$ |

解答. 注意到 $H(Y|X)$ 是固定的

$$C = \max_{P_X} I(X;Y) = \max_{P_X} H(Y) - H(Y|X) = \max_{P_X} H(Y) - 1 = \log_2 5 - 1 \text{ bit}$$

■

**例 9.0.7.** $X, Y$ 均取值于 $\{0, 1\}$，可虑一个较现实的信道，即似然如下 $(\varepsilon > 0)$，求容量.

| $P(Y\|X)$ | $X = 0$ | $X = 1$ |
|---|---|---|
| $Y = 0$ | $1 - \varepsilon$ | $\varepsilon$ |
| $Y = 1$ | $\varepsilon$ | $1 - \varepsilon$ |

解答. 和上一个例子很像，注意到 $H(Y|X)$ 是固定的

$$C = \max_{P_X} I(X;Y) = \max_{P_X} H(Y) - H(Y|X) = \max_{P_X} H(Y) - H_2(\varepsilon) = 1 - H_2(\varepsilon) \text{ bit}$$

其中 $H_2(\varepsilon) = \varepsilon \log_2 \varepsilon + (1-\varepsilon)\log_2(1-\varepsilon)$. ■

**例 9.0.8.** 试设计一个含噪信道，使得 $H(Y)$ 远大于 $I(X;Y)$.

我们此前讨论了纠错码的应用，纠错码实际上添加了一部分冗余，但增加了纠错的能力. 冗余越多，纠错能力越强，可恢复概率越大，但相应地传输效率越低. 但我们希望让正确率无限接近 1（也就是任给 $\varepsilon > 0$，正确率都可以大于 $1-\varepsilon$），而此时“效率”时受限的，在下一章我们量化地讨论，这里先不严谨地理解一下.

用 $R$ 表示每次传输时真正信息的比特数，那么在容量为 $C$ 的信道上传输时. 若 $R < C$，存在纠错码可让错误率无限接近 0；若 $R > C$，没有纠错码可让错误率无限接近 0.

# 第十章　Lecture 11

讲师：王立威　　课程时间：25.Apr.29th　　笔记：25.June.9th

在介绍信道编码定理 (Channel Coding Theorem, CCT) 之前，先了解一下渐进等分性 (Asymptotic Equipartition Property, AEP). 我们不加证明地使用最简单的弱大数律 (WLLN)

**定理 10.0.1** (弱大数律，WLLN)**.** *若随机变量 $X, X_1, X_2, \ldots, X_n, \ldots$ 是 i.i.d. 的，且 $\mathbb{E}X$ 存在（有界），那么对于任意 $\varepsilon > 0$，有*

$$\lim_{n\to\infty} \Pr\left[\left|\frac{1}{n}\sum_i X_i - \mathbb{E}X\right| \geqslant \varepsilon\right] = 0$$

**推论 10.0.2.** *若随机变量 $X, X_1, X_2, \ldots, X_n, \ldots$ 是 i.i.d. 的，且 $\mathbb{E}X$ 存在（有界），对于任意函数 $g$ 和 $\varepsilon > 0$，有*

$$\lim_{n\to\infty} \Pr\left[\left|\frac{1}{n}\sum_i g(X_i) - \mathbb{E}g(X)\right| \geqslant \varepsilon\right] = 0$$

**推论 10.0.3.** *若离散随机变量 $X, X_1, X_2, \ldots, X_n, \ldots$ 是 i.i.d. 的，且 $\mathbb{E}X$ 存在（有界），记其分布列为 $p(x)$，则对任意 $\varepsilon > 0$，有*

$$\lim_{n\to\infty} \Pr\left[\left|\frac{1}{n}\sum_i \log\frac{1}{p(X_i)} - \mathbb{E}\log\frac{1}{p(X)}\right| \geqslant \varepsilon\right] = 0$$

我们现在考虑 $X$ 服从 Bernoulli 分布，且 $\Pr[X = 1] = p$，仍然 i.i.d. 地采样出 $X_1, X_2, \ldots$，那么上面的推论就说明以 $\approx 1$ 的概率，会有

$$2^{-n(H(X)+\varepsilon)} \leqslant \Pr[X_1, \ldots, X_n] \leqslant 2^{-n(H(X)-\varepsilon)}$$

从概率的角度，我们可以忽视余下的情况. 此后开始把上式记作如下记号

$$\Pr[X_1, \ldots, X_n] \doteq 2^{-nH(X)}$$

我们记集合

$$A := \{(x_1, x_2, \ldots, x_n) : P(x_1, x_2, \ldots, x_n \doteq 2^{-nH(X)})\}$$

称 $A$ 为**典型集** (typical set)，$A$ 中的元素称为**典型序列** (typical sequence). 由于每个序列取到的概率是一样的，所以上面的分析导出

**定理 10.0.4** (AEP)**.** $|A| \approx 2^{nH(X)}$

当然，我们对于 $(X, Y), (X_1, Y_1), \ldots$ i.i.d. $\sim p(x, y)$，也可以定义其联合典型集和联合典型序列。$|A| \approx 2^{nH(X,Y)}$

另外，联合采样也可以看作依据似然（条件概率）分布采样，并保留性质 $2^{nH(Y|X)} = 2^{nH(X,Y)}/2^{nH(X)}$.

这样来说，可以考虑：如果我们分别独立地按照边缘概率分布在典型集中采样（其实不强调典型集也可以，以为几乎以概率 1 落在典型集内）$x_1, x_2, \ldots, x_n$ 和 $y_1, y_2, \ldots, y_n$. 试问将其融合得到的 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 也在典型集中的概率是多少？很显然是

$$\frac{2^{nH(X,Y)}}{2^{nH(X)} \cdot 2^{nH(Y)}} = \frac{2^{nH(Y|X)}}{2^{nH(Y)}} = 2^{-nI(X;Y)}$$

请注意，上面的两种计算方法对应了两种理解方式!

另外，本节内容均是为后面铺垫。

# 第十一章　Lecture 12

讲师：王立威　　课程时间：25.May.13rd　　笔记：25.June.9th

首先引入概念

**定义 11.0.1** (码率). 一个信道通信的**码率** *(rate)* 是指平均每次传输有效信息量.

**例 11.0.2.** 如果我传输一个比特时，将其重复 $n$ 次再传输，那么码率为 $1/n$.

然后叙述一直以来想要证明的重要定理：

**定理 11.0.3** (Shannon 信道编码定理). 对于一个含噪信道，记其信道容量为 $C$，希望码率为 $R$. 那么

- 若 $R < C$，那么对于任意 $\varepsilon > 0$ 存在码率至少为 $R$ 的纠错码，使得错误概率小于 $\varepsilon$.
- 若 $R > C$，那么存在 $\varepsilon_0 > 0$，对于任意码率至少是 $R$ 的纠错码其错误概率都至少是 $\varepsilon_0$.

证明. 我们这里主要阐述 Shannon 精妙的思想，后续严格化工作请自行补充.

(1) $R < C$ 时. 不失一般性（这里需要思考），假设信息集为 $\{m_1, m_2, \ldots, m_{nR}\}$ 并且它们均匀分布. 下面来设计码字集合 $(c_1, c_2, \ldots, c_{nR})$

使用随机编码（考虑所有可能的码字集合），设每个码字长度都为 $n$，记

$$c_i = (c_{i1}, c_{i2}, \ldots, c_{in}), \quad i = 1, 2, \ldots 2^{nR}$$

由于信息的熵为 $nR$，码字长度为 $n$，所以码率为 $R$。

根据信道容量的定义，取使得达到信道容量的先验分布：

$$P(X) = \arg\max_{P_X} I(X;Y)$$

我们现在令 $c_{ij}$ 独立同分布地服从分布 $P(X)$（对于所有 $i,j$）这一步是整个证明中很关键的一步！我们考察其平均错误率，如果平均错误率 $< \varepsilon$，那么就一定有一组码字的错误率 $< \varepsilon$. 不过我们仍需给出解码的方法。

解码时，当我们接收到 $(y_1,\ldots,y_n)$ 时。我们将其对应到 $c_i = (x_1,\ldots,x_n)$ 使得 $(x_1,\ldots,x_n;y_1,\ldots,y_n)$ 是关于联合分布 $P(X)P(Y|X)$ 的典型序列；如果没有这样的或者有多于一个的 $(x_1,\ldots,x_n)$，那么返回错误；如果唯一存在 $c_i$，则返回 $c_i$。

下面计算错误概率，不妨假设我们原定传输的信息是 $m_1$

$$\Pr\left[\text{找到多个解}\right] \leqslant \sum_{i=2}^{2^{nR}} \Pr\left[\boldsymbol{y}\text{和}c_i\text{匹配为典范序列}\right] \approx 2^{nR}\cdot 2^{-nI(X;Y)}$$

而根据我们的 $P(X)$ 的选择，会有 $I(X;Y) = C$，上式右侧变为 $2^{n(R-C)}$，所以当 $R < C$ 时取 $n$ 充分大即可. 而

$$\Pr\left[\text{找不到解}\right] \leqslant \varepsilon$$

所以加在一起的错误概率就是趋于零的，因此可以任意小.

(2) 当 $R > C$ 时，往证明存在 $\varepsilon_0$ 使得任意纠错码的错误率都 $\geqslant \varepsilon_0$. 我们需要使用 Fano 不等式（写在后面）

设我们发出的消息是 $M$，收到比特流 $Y_1,\ldots,Y_n$，则

$$H(M|Y_1,\ldots Y_n) = H(X) - I(M;Y_1,\ldots,Y_n) \geqslant nR - nC = n(R-C)$$

那么错误概率

$$P_{\text{error}} \geqslant \frac{H(M|Y_{1\sim n}) - 1}{\log|\{0,1\}^{nR}|} \geqslant \frac{n(R-C)-1}{nR} = \frac{R-C}{R} - O(n^{-1})$$

可以看出错误率存在正下界.

综合两部分便完成了定理的证明. □

**定理 11.0.4** (Fano 不等式). 令 $X, Y$ 是随机变量，$X \in \mathcal{H}$，其取值范围有限 $|\mathcal{H}| < \infty$。现在我们使用 $Y$ 来估计 $X$，记估计值 $\hat{X} = g(Y)$，那么错误概率满足

$$P_{error} = \Pr[\hat{X} \neq X] \geqslant \frac{H(X|Y) - 1}{\log|\mathcal{H}|}$$

# 第十二章　Lecture 13

讲师：王立威　　课程时间：25.May.20th　　笔记：25.June.9th

首先完成了信道编码定理的证明，放在了前一节. 下面介绍 Fisher 信息和 Cramér-Rao 不等式.

动机和来源是无偏估计. （若学过统计学可跳过）

**定义 12.0.1** (无偏估计). 我们（往往独立同分布地）采样了一组样本点 $X = (X_1, X_2, \ldots, X_n)$，设其密度函数为 $f(\cdot;\theta)$（其中 $\theta$ 是我们要估计的参数），那么有

$$f(x;\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

现在使用这些样本点得到 $\theta$ 的一个估计 $\hat{\theta} = \phi(X_1, \ldots, X_n)$. 如果满足 $\mathbb{E}[\hat{\theta}] = \theta$ 则称 $\hat{\theta}$ 是 $\theta$ 的**无偏估计** *(unbiased estimation).*

在正式定义 Fisher 信息之前先要引入得分函数的定义.

**定义 12.0.2** (得分函数). **得分函数** *(score function)* 是指

$$S(X;\theta) := \frac{\partial}{\partial\theta} \ln f(X;\theta)$$

**命题 12.0.3.** 得分函数满足 $\mathbb{E}[S(X;\theta)] = 0$

证明. 根据定义

$$\begin{aligned}\mathbb{E}[S(X;\theta)] &= \int S(x;\theta)f(x;\theta)\,\mathrm{d}x \\ &= \int \frac{\partial}{\partial\theta}\ln f(x;\theta)f(x;\theta) \\ &= \int \frac{\partial f(x;\theta)}{\partial\theta}\frac{1}{f(x;\theta)}f(x;\theta)\,\mathrm{d}x \\ &= \int \frac{\partial f(x;\theta)}{\partial\theta}\,\mathrm{d}x \\ &= \frac{\partial}{\partial\theta}\int f(x;\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta}\ 1 = 0\end{aligned}$$

□

现在可以定义 Fisher 信息了

**定义 12.0.4** (Fisher 信息)**.** *对于 $X,\theta$，定义其 Fisher 信息 (information) 为*

$$I(\theta) := \mathsf{Var}(S(X;\theta)) = \mathbb{E}[S^2(X;\theta)]$$

**命题 12.0.5.** $I(\theta) = -\mathbb{E}\left[\dfrac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right] = -\displaystyle\int \frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)f(x;\theta)\,\mathrm{d}x$

证明.

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right] &= \int \frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)f(x;\theta)\,\mathrm{d}x \\ &= \int \frac{\partial}{\partial\theta}\left(\frac{\nabla_\theta f(x;\theta)}{f(x;\theta)}\right)f(x;\theta)\,\mathrm{d}x \\ &= \int \frac{\nabla_\theta^2 f(x;\theta)\cdot f(x;\theta) - (\nabla_\theta f(x;\theta))^2}{f^2(x;\theta)}f(x;\theta)\,\mathrm{d}x \\ &= \int \nabla_\theta^2 f(x;\theta)\,\mathrm{d}x - \int\left(\frac{\nabla_\theta f(x;\theta)}{f(x;\theta)}\right)^2 f(x;\theta)\,\mathrm{d}x \\ &= \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)\,\mathrm{d}x - \int \frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)f(x;\theta)\,\mathrm{d}x \\ &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right]\end{aligned}$$

□

Fisher 信息的意义在衡量无偏估计最好能好到什么程度，这里的好是由方差来度量的. 也就是这里要介绍的不等式：

**定理 12.0.6** (Cramér-Rao 不等式)**.** *对于任意 $\theta$ 的无偏估计 $\phi : X \to \mathbb{R}$，我们有*

$$\mathsf{Var}(\phi(X)) \geqslant \frac{1}{I(\theta)}$$

*证明.* 证明此略. □

我们可以将得分函数，Fisher 信息，Cramér-Rao 不等式拓展到 $d$ 维的情况：

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ln f(X;\theta) \quad &\longrightarrow \quad \nabla_\theta \ln f(x;\theta) \\
-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right] \quad &\longrightarrow \quad -\mathbb{E}\left[\nabla_\theta^2 \ln f(X;\theta)\right] \\
\mathsf{Var}(\phi(X)) \geqslant \frac{1}{I(\theta)} \quad &\longrightarrow \quad \mathsf{Cov}(\phi(X)) \geqslant \frac{1}{I(\theta)}
\end{aligned}$$

# 第十三章　Lecture 14

讲师：王立威　　课程时间：25.May.27th　　笔记：25.June.9th

当我们在进行多方参与的计算时，每方只拿到了输入的一部分，我们希望讨论为了得到计算的结果，通信的代价几何？

**定义 13.0.1** (通信代价). *我们考虑 Boole 函数 $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$. 现在要计算 $f(x,y)$，且有两方参与者 Alice 和 Bob，但 Alice 只能看到 $x$ 而 Bob 只能看到 $y$.* **通信代价** *(communication complexity) 是指最差情况下所需要的通信总比特数，记作 $\mathrm{CC}(f)$.*

这里举了一个例子（太过复杂了，之后再相信陈述，请先看当日的录像 42min 处），最后可以发现，我们需要将 $(x,y)$ 所有可能的取值划分成若干等值矩形，设一共划分了 $R$ 个，那么通信代价有一个下界是 $\log_2 R$.

总结成定理就是

**定理 13.0.2.** *$\mathrm{CC}(f) \geqslant \log_2 \chi(f)$，其中 $\chi(f)$ 指最优等值矩形划分的矩形个数.*

**例 13.0.3.** *对于等值函数 $f_{Eq}(x,y) = \mathbb{I}_{x=y}$，试求 $\chi(f_{Eq})$.*

解答. 也就是要划分单位矩阵 $I_{2^n \times 2^n}$. 注意到对角线的元素两两必然不在同一个等值矩形中，所以个数满足 $\chi(f_{\mathrm{Eq}}) \geqslant 2^n$，推出 $\mathrm{CC}(f) \geqslant n$. ■

**例 13.0.4.** *对于函数*

$$f(x,y) = \begin{cases} 1 & , \textit{if } x \wedge y = 0 \\ 0 & , \textit{otherwise} \end{cases}$$

其中 $\wedge$ 表示按位与，试求 $\chi(f)$ 的一个下界.

解答. 关注副对角线上的所有元素，会发现对于 $x \neq \tilde{x}$，$f(x,\tilde{x})$ 和 $f(x^c,\tilde{x}^c)$ 不可能同时为 1，所以 $\chi(f) \geqslant 2^n$. ■

不难察觉，要精确计算 $\chi(f)$ 是很难的，因此往后我们都将重点放在下界的估计上. 下面将 $f$ 的取值矩阵记作 $M(f)_{2^n\times 2^n}$

**定理 13.0.5.** 有 $\operatorname{rank}(M(f)) \leqslant \chi_1(f) \leqslant \chi(f)$，其中 $\chi_1(f)$ 表示值为 *1* 的矩阵块个数.

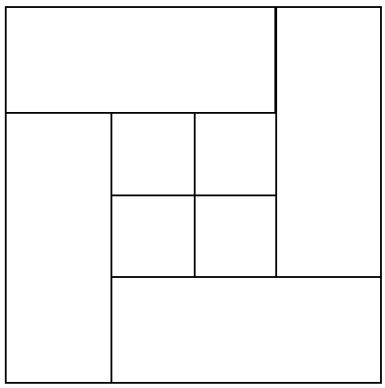
证明. 注意到 $\operatorname{rank}(A+B) \leqslant \operatorname{rank}(A) + \operatorname{rank}(B)$ 即可证明. □

**推论 13.0.6.** $\mathrm{CC}(f) \geqslant \log_2 \operatorname{rank}(M(f))$

事实上，有理论工作已经证明

$$\mathrm{CC}(f) \leqslant \operatorname{poly}\log \operatorname{rank}(M(f))$$

但使用 $\chi(f)$ 去估计得到的是一个下界，该下界不一定可达（例如下图），下一章会介绍一个上界.



$$\log_2 \chi = \log_2 8 = 3$$

图 13.1: 无法达到下界的例子

# 第十四章　Lecture 15

讲师：王立威　　课程时间：25.June.3rd　　笔记：25.June.9th

（前一部分内容不在考试范围之内，所以等待考试结束后再更新）

现在考虑一个比之前更泛化的最大熵问题

**例 14.0.1.** *设 $X$ 是一个 $d$ 维的随机变量，满足*

$$\mathbb{E}[X] = \mathbf{0}, \quad \mathsf{Cov}(X) = \mathbb{E}[XX^\top] = \mathbf{\Sigma}$$

*求最大熵的分布.*

解答. 我们直接证明最大熵分布是 $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$，取 $Y$ 服从该正态分布，往证明对于任意合意的 $X$，都有 $h(Y) \geqslant h(X)$.

注意到 $D(X\|Y) \geqslant 0$，而

$$\begin{aligned} D(X\|Y) &= -h(X) - \int f_X(t) \ln f_Y(t)\,\mathrm{d}t \\ &= -h(X) + \frac{1}{2}\ln\left((2\pi)^d \det \mathbf{\Sigma}\right) + \mathbb{E}[X^\top \mathbf{\Sigma}^{-1} X] \\ &= -h(X) + \frac{1}{2}\ln\left((2\pi)^d \det \mathbf{\Sigma}\right) + \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}) \\ &= -h(X) + \frac{1}{2}\ln\left((2\pi)^d \det \mathbf{\Sigma}\right) + d \\ &= -h(X) + h(Y) \geqslant 0 \end{aligned}$$

所以推出 $h(Y) \geqslant h(X)$. 至此证毕. ■

**例 14.0.2.** 当 $X$ 是离散型随机变量时,规定取值是 $\{0,1,2,\dots\}$,并限定 $\mathbb{E}[X]=\mu$. 试求最大熵分布.

解答. 可以算得，最大熵分布是几何分布. ■