INFORMATION THEORY

信息论自学笔记

朧千夏

December 28, 2025

Contents

1 熵，相对熵与互信息 **1**
1.1 熵 (Entropy) . . . 1
1.2 联合熵与条件熵 . . . 3
1.3 相对熵与互信息 . . . 4
1.3.1 相对熵 (Kullback-Leibler Distance) . . . 4
1.3.2 互信息 (Mutual Information) . . . 5
1.4 熵与互信息的关系 . . . 5
1.5 链式法则扩展 . . . 6
1.6 Jensen 不等式与信息不等式 . . . 8
1.7 Log Sum 不等式与凸性 . . . 11
1.8 数据处理不等式 (Data-Processing Inequality) . . . 12
1.9 充分统计量 (Sufficient Statistics) . . . 13
1.10 Fano 不等式 (Fano's Inequality) . . . 13

2 渐进均分性 **16**
2.1 AEP 定理 . . . 16
2.1.1 AEP 定理陈述与证明 . . . 16
2.2 典型集 (The Typical Set) . . . 17
2.2.1 典型集的性质 . . . 17
2.3 AEP 的应用：数据压缩 . . . 18
2.3.1 编码方案 . . . 18
2.3.2 平均码长分析 . . . 18
2.4 高概率集与典型集 . . . 19

3 随机过程的熵率 **20**
3.1 马尔可夫链 (Markov Chains) . . . 20
3.1.1 基本定义 . . . 20
3.1.2 时不变性与平稳分布 . . . 20
3.2 熵率 (Entropy Rate) . . . 21
3.2.1 两种定义及其等价性 . . . 21
3.2.2 马尔可夫链的熵率 . . . 22
3.2.3 加权图上的随机游走 . . . 22
3.3 热力学第二定律 (Second Law of Thermodynamics) . . . 23
3.3.1 相对熵随时间减少 . . . 23
3.3.2 熵的增加与均匀分布 . . . 23
3.4 马尔可夫链的函数 (Functions of Markov Chains) . . . 24
3.4.1 上界和下界 . . . 24

4 数据压缩 **25**

4.1 码的例子 . . . 25
4.1.1 码的分类 . . . 25
4.2 Kraft 不等式 . . . 26
4.3 最优码 (Optimal Codes) . . . 27
4.3.1 直观推导 . . . 27
4.3.2 主要定理 . . . 27
4.4 最优码长的界 . . . 28
4.4.1 香农编码 (Shannon Coding) . . . 28
4.4.2 块编码与熵率 . . . 28
4.4.3 错误分布的代价 . . . 29
4.5 唯一可译码的 Kraft 不等式 . . . 30
4.6 霍夫曼编码 (Huffman Codes) . . . 31
4.6.1 算法描述与示例 . . . 31
4.7 霍夫曼编码的最优性 (Optimality) . . . 33
4.8 香农-法诺-伊莱亚斯编码 (SFE Code) . . . 35
4.8.1 算法定义 . . . 35
4.8.2 前缀性质的证明 . . . 36
4.8.3 平均码长界 . . . 36
4.9 香农码的竞争最优性 . . . 36
4.10 从公平硬币生成离散分布 . . . 37
4.10.1 引理与下界 . . . 38
4.10.2 一般分布的生成算法 . . . 39

1 熵，相对熵与互信息

本章奠定了信息论的基石。我们将引入熵（Entropy）、相对熵（Relative Entropy）和互信息（Mutual Information）这三个核心概念。这些量不仅是通信理论的基础，也广泛应用于统计学、计算机科学等领域。理解它们之间的关系（如链式法则、数据处理不等式）是关键。

1.1 熵 (Entropy)

熵是信息论中最基本的概念，它是对随机变量“不确定性”的度量。

Definition 1.1 (熵)**.** 设 X 是一个离散随机变量，其字母表为 $\mathcal{X}$，概率质量函数为 $p(x) = \Pr\{X = x\}, x \in \mathcal{X}$。$X$ 的熵 $H(X)$ 定义为：

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{1.1}$$

Remark 1.1. 关于定义的说明:

- **对数底数**：通常以 2 为底，此时熵的单位是比特（bits）。若以 e 为底，单位为奈特（nats）。除非特别说明，我们默认底数为 2。当使用底数 b 时，我们将熵记为 $H_b(X)$。
- **约定**：我们定义 $0\log 0 = 0$。这是合理的，因为 $\lim_{x\to 0} x\log x = 0$。这意味着零概率的事件不会改变熵的值。
- **函数属性**：熵是概率分布 p 的函数，记作 $H(p)$。它不依赖于 X 具体取什么值（是取 0, 1 还是 apple, banana），只依赖于这些值出现的概率。

Remark 1.2. 如何深入理解熵?

1. **惊奇度的期望**：我们可以将 $\log\frac{1}{p(x)}$ 看作是观察到事件 $X = x$ 时的“惊奇程度”。如果 $p(x) \approx 1$，事件发生理所应当，惊奇度接近 0；如果 $p(x) \approx 0$，事件发生非常罕见，惊奇度很高。惊奇度对应着该事件所蕴含信息量的大小（信息量越大越令人惊奇嘛），熵 $H(X) = \mathbb{E}[\log\frac{1}{p(X)}]$ 正是这种惊奇度的平均值。
2. **最短描述长度**：在编码理论中，熵代表了描述随机变量 X 平均所需的“最少比特数”。如果我们要通过提问“是/否”问题来确定 X 的值，熵给出了平均需要提问次数的下界。
3. **期望形式**：

$$H(X) = \mathbb{E}_p\left[\log\frac{1}{p(X)}\right] \tag{1.2}$$

Remark 1.3. 熵的公理化定义

你可能会问：公式 $H(X) = -\sum p(x)\log p(x)$ 看起来很精妙，但它是怎么想出来的？又怎么和我们对于信息量多少的直观对应上呢？事实上，Shannon 在 1948 年的论文中并没有直接给出这个定义，而是提出：如果我们想要定义一个度量 H 来描述“不确定性”或“信息量”，这个度量应该满足哪些**符合直觉的必要条件**？

Shannon 发现，只要我们坚持以下三条最基本的“常识”，数学上就能证明唯一满足这些条件的函数就是 $H(X) = -\sum p_i \log p_i$。

如果一个对称函数序列 $H_m(p_1, p_2, \ldots, p_m)$ 满足以下性质：

- **归一化 (Normalization)**：
$$H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1 \tag{1.3}$$
- **连续性 (Continuity)**：$H_2(p, 1-p)$ 是关于 p 的连续函数。
- **分组律 (Grouping)**：
$$H_m(p_1, p_2, \ldots, p_m) = H_{m-1}(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \tag{1.4}$$

则可以证明，H_m 必须具有如下形式：

$$H_m(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i, \quad m = 2, 3, \ldots \tag{1.5}$$

直观理解就是

- **归一化**规定了信息量的单位（比特）。
- **连续性**保证了概率微小的变化不会导致信息量的剧烈跳变。
- **分组律**体现了计算的一致性：通过多步决策（先确定属于哪个大组，再确定组内具体元素）得到的总不确定性，应该等于一步到位直接决策的不确定性。其中 (p_1+p_2) 是进入该分支的权重。

  假设我们要区分三个事件 A, B, C，概率分别为 p_1, p_2, p_3。我们可以直接问：“是 A、B 还是 C?”

  或者，我们可以分两步走：
  - 第一步：问“是 (A 或 B)，还是 C?”（此时 (A 或 B) 的概率是 $p_1 + p_2$）。
  - 第二步：如果第一步的答案是 (A 或 B)，再问“是 A 还是 B?”（此时是在 $p_1 + p_2$ 发生的前提下，条件概率分别为 $\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}$）。

  公理要求这两种方式计算出的总信息量必须相等：

$$H_3(p_1, p_2, p_3) = \underbrace{H_2(p_1 + p_2, p_3)}_{\text{第一步的不确定性}} + \underbrace{(p_1 + p_2)}_{\text{权重}} \cdot \underbrace{H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)}_{\text{第二步的不确定性}} \tag{1.6}$$

  注意第二项乘了一个权重 (p_1+p_2)，因为只有当第一步结果落在 (A, B) 组里时，我们才需要进行第二步辨别。

Lemma 1.1. $H(X) \geq 0$。

Proof. 由于 $0 \leq p(x) \leq 1$，故 $\log \frac{1}{p(x)} \geq 0$。非负数的加权和是非负的。 □

Lemma 1.2. $H_b(X) = (\log_b a) H_a(X)$。

Proof. 由对数换底公式 $\log_b p = \log_b a \cdot \log_a p$ 易证。这说明改变对数底数只是改变了度量单位。 □

Example 1.1 (二元熵函数)**.** 设 X 是伯努利分布，$X=1$ 概率为 p，$X=0$ 概率为 $1-p$。则

$$H(X) = -p\log p - (1-p)\log(1-p) \stackrel{\text{def}}{=} H(p)$$

1.2　联合熵与条件熵

Definition 1.2 (联合熵)**.** 一对离散随机变量 (X,Y) 的联合熵 $H(X,Y)$ 定义为：

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = -\mathbb{E}\log p(X,Y) \tag{1.7}$$

这实际上就是把 (X,Y) 看作一个单一的向量值随机变量时的熵。

Definition 1.3 (条件熵)**.** 若 $(X,Y)\sim p(x,y)$，条件熵 $H(Y|X)$ 定义为：

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) && (1.8)\\ &= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) && (1.9)\\ &= -\mathbb{E}\log p(Y|X) && (1.10) \end{aligned}$$

Remark 1.4. 深刻理解条件熵：不要把 $H(Y|X)$ 理解为“给定某个特定的 x 后 Y 的熵”。它是“给定 X 这一事件发生后，对 Y 的熵的平均值”。它是我们在观测到 X 之后，关于 Y 仍然保留的平均不确定性。

Theorem 1.1 (链式法则 Chain Rule)**.**

$$H(X,Y) = H(X) + H(Y|X) \tag{1.11}$$

Proof. 利用对数性质 $\log p(x,y) = \log(p(x)p(y|x)) = \log p(x) + \log p(y|x)$，再两边求期望即可得证。 □

Remark 1.5. 链式法则的物理意义：描述一对随机变量 (X,Y) 所需的信息量，等于先描述 X 所需的信息量 $H(X)$，加上在已知 X 的情况下描述 Y 所需的额外信息量 $H(Y|X)$。

Corollary 1.1. 条件形式的链式法则：$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$。

1.3 相对熵与互信息

1.3.1 相对熵 (Kullback-Leibler Distance)

Definition 1.4 (相对熵). 两个概率质量函数 $p(x)$ 和 $q(x)$ 之间的相对熵（或 KL 散度）定义为：

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = \mathbb{E}_p \log\frac{p(X)}{q(X)} \tag{1.12}$$

Remark 1.6. 相对熵的本质与解释：

1. 编码低效性：假设真实分布是 p，我们针对 p 的最优编码的平均长度是 $H(p)$。但如果我们误以为是 q 并据此设计了最优编码，那么对 q 的最优编码的平均长度是 $H(p)+D(p||q)$。$D(p||q)$ 度量了因为这种误判，平均每个符号“多浪费了多少比特”。
2. 距离度量：相对熵常被视为分布间的“距离”，但要注意它不是真正的距离度量：
   - 非对称性：$D(p||q) \neq D(q||p)$。
   - 不满足三角不等式。
3. 无穷大情况：如果存在 x 使得 $p(x) > 0$ 但 $q(x) = 0$，则 $D(p||q) = \infty$。这意味着如果我们将可能发生的事件误判为不可能事件，我们要付出的代价（编码长度）是无限大的。

Exercise 1.1. 分别给出相对熵不满足对称性和三角不等式的例子。

Solution. 我们需要构造二元随机变量的概率分布来验证。记分布 $p = (p_1, 1-p_1)$。

1. 对称性的反例

取 $p = (1, 0)$ 为确定性分布，$q = (1/2, 1/2)$ 为均匀分布。

- 计算 $D(p||q)$：

$$D(p||q) = 1\log\frac{1}{0.5} + 0\log\frac{0}{0.5} = \log 2 = 1 \text{ bit} \tag{1.13}$$

- 计算 $D(q||p)$：

$$D(q||p) = 0.5\log\frac{0.5}{1} + 0.5\log\frac{0.5}{0} = -0.5 + \infty = \infty \tag{1.14}$$

显然 $1 \neq \infty$，故 $D(p||q) \neq D(q||p)$。

2. 三角不等式的反例

取 $p = (1, 0)$，$q = (0.5, 0.5)$，$r = (0.25, 0.75)$

分别计算三段相对熵：

$$D(p||r) = 1\log\frac{1}{0.25} + 0 = \log 4 = 2 \text{ bits} \tag{1.15}$$

$$D(p||q) = 1\log\frac{1}{0.5} + 0 = \log 2 = 1 \text{ bit} \tag{1.16}$$

$$\begin{aligned} D(q||r) &= 0.5\log\frac{0.5}{0.25} + 0.5\log\frac{0.5}{0.75} \\ &= 0.5(1) + 0.5(\log 2 - \log 3) = 1 - 0.5\log 3 \approx 0.2075 \text{ bit} \end{aligned} \tag{1.17}$$

验证不等式关系：

$$D(p||r) = 2 \quad \text{vs} \quad D(p||q) + D(q||r) \approx 1.2075 \tag{1.18}$$

显然 $2 > 1.2075$，即 $D(p||r) > D(p||q) + D(q||r)$。三角不等式不成立。

□

1.3.2 互信息 (Mutual Information)

Definition 1.5 (互信息). 两个随机变量 X 和 Y 之间的互信息 $I(X;Y)$ 定义为联合分布与乘积分布之间的相对熵：

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) \tag{1.19}$$

Remark 1.7. 互信息的含义：

- 它是 X 和 Y 之间依赖程度的度量。
- 如果 X 和 Y 独立，则 $p(x,y) = p(x)p(y)$，此时 $\log 1 = 0$，互信息为 0。
- 它可以理解为：由于知道了 Y，我们对 X 的不确定性减少了多少？即 $I(X;Y) = H(X) - H(X|Y)$。
- 互信息是对称的：X 包含关于 Y 的信息量，等于 Y 包含关于 X 的信息量。

1.4 熵与互信息的关系

Theorem 1.2 (互信息与熵的关系).

$$I(X;Y) = H(X) - H(X|Y) \tag{1.20}$$
$$I(X;Y) = H(Y) - H(Y|X) \tag{1.21}$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1.22}$$
$$I(X;Y) = I(Y;X) \tag{1.23}$$
$$I(X;X) = H(X) \quad \text{(自信息)} \tag{1.24}$$

Proof. 根据互信息的定义：

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{1.25}$$
$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \tag{1.26}$$
$$= \sum_{x,y} p(x,y) \log p(x|y) - \sum_{x,y} p(x,y) \log p(x) \tag{1.27}$$

对于第二项，利用边缘分布性质 $\sum_y p(x,y) = p(x)$：

$$\sum_{x,y} p(x,y) \log p(x) = \sum_{x} p(x) \log p(x) = -H(X)$$

对于第一项，直接利用条件熵定义：

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

代回原式，得：

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

由对称性同理可证 $I(X;Y) = H(Y) - H(Y|X)$。

至于第三个等式，将 $\log \frac{p(x,y)}{p(x)p(y)}$ 展开为 $\log p(x,y) - \log p(x) - \log p(y)$ 并求期望即可得：

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

□

Remark 1.8. 韦恩图记忆法（Venn Diagram）： 想象两个集合（圆）代表 $H(X)$ 和 $H(Y)$。

- $H(X)$：左边的圆（X 的总不确定性）。
- $H(Y)$：右边的圆。
- $H(X,Y)$：两个圆的并集（系统的总不确定性）。
- $I(X;Y)$：两个圆的交集（共享的信息）。
- $H(X|Y)$：左边圆减去交集剩下的部分（已知 Y 后 X 剩余的不确定性）。

公式 $I(X;Y) = H(X) + H(Y) - H(X,Y)$ 对应集合论中的 $|A \cap B| = |A| + |B| - |A \cup B|$。

1.5 链式法则扩展

Theorem 1.3 (熵的链式法则)**.** 设 $X_1, X_2, \ldots, X_n$ 服从 $p(x_1, \ldots, x_n)$，则

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \tag{1.28}$$

Proof. 我们反复使用两个变量的熵的链式法则 $H(X,Y) = H(X) + H(Y|X)$。

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1) \tag{1.29}$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1) \tag{1.30}$$

$$= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \tag{1.31}$$

$$\vdots \tag{1.32}$$

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, \ldots, X_1) \tag{1.33}$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \tag{1.34}$$

□

Remark 1.9. 这解释了“总不确定性”是逐步累积的：$H(X_1)$（第一个变量的不确定性）$+H(X_2|X_1)$（已知第一个后，第二个的新增不确定性）$+\dots$

Theorem 1.4 (互信息的链式法则)**.**

$$I(X_1,\dots,X_n;Y)=\sum_{i=1}^{n}I(X_i;Y|X_{i-1},\dots,X_1) \tag{1.35}$$

其中条件互信息定义为 $I(X;Y|Z)=H(X|Z)-H(X|Y,Z)$。

Proof. 利用互信息与熵的关系 $I(A;B)=H(A)-H(A|B)$，我们将 $X_1,\dots,X_n$ 看作整体：

$$I(X_1,\dots,X_n;Y)=H(X_1,\dots,X_n)-H(X_1,\dots,X_n|Y)$$

对右边两项分别应用熵的链式法则（注意第二项有条件 Y）：

$$H(X_1,\dots,X_n)=\sum_{i=1}^{n}H(X_i|X_{i-1},\dots,X_1) \tag{1.36}$$

$$H(X_1,\dots,X_n|Y)=\sum_{i=1}^{n}H(X_i|X_{i-1},\dots,X_1,Y) \tag{1.37}$$

将两式相减：

$$I(X_1,\dots,X_n;Y)=\sum_{i=1}^{n}[H(X_i|X_{i-1},\dots,X_1)-H(X_i|X_{i-1},\dots,X_1,Y)] \tag{1.38}$$

$$=\sum_{i=1}^{n}I(X_i;Y|X_{i-1},\dots,X_1) \tag{1.39}$$

最后一步用到了条件互信息的定义：$I(A;B|C)=H(A|C)-H(A|B,C)$。 □

Definition 1.6 (条件相对熵)**.** 设 $p(x,y)$ 和 $q(x,y)$ 是两个联合概率质量函数。条件相对熵 $D(p(y|x)||q(y|x))$ 定义为条件分布之间的相对熵在边缘分布 $p(x)$ 下的平均值：

$$D(p(y|x)||q(y|x))=\sum_{x}p(x)\sum_{y}p(y|x)\log\frac{p(y|x)}{q(y|x)} \tag{1.40}$$

$$=\mathbb{E}_{p(x,y)}\left[\log\frac{p(Y|X)}{q(Y|X)}\right] \tag{1.41}$$

Remark 1.10. 符号说明：虽然记号写作 $D(p(y|x)||q(y|x))$，但这并不是关于某个特定 x 的函数，而是一个期望值。它衡量了在平均意义上，当我们已知 X 后，关于 Y 的真实条件分布 $p(y|x)$ 与假设条件分布 $q(y|x)$ 之间的差异。

Theorem 1.5 (相对熵的链式法则)**.** 两个联合分布之间的相对熵，可以分解为边缘分布的相对熵与条件相对熵之和：

$$D(p(x,y)||q(x,y))=D(p(x)||q(x))+D(p(y|x)||q(y|x)) \tag{1.42}$$

证明. 利用对数的性质 $\log(ab) = \log a + \log b$ 以及联合概率公式 $p(x,y) = p(x)p(y|x)$：

$$D(p(x,y)||q(x,y)) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \tag{1.43}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \tag{1.44}$$

$$= \sum_{x,y} p(x,y) \left(\log \frac{p(x)}{q(x)} + \log \frac{p(y|x)}{q(y|x)} \right) \tag{1.45}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \tag{1.46}$$

我们分别处理这两项：

- **第一项**：求和与 y 无关，利用 $\sum_y p(x,y) = p(x)$：

$$\sum_x \left(\sum_y p(x,y) \right) \log \frac{p(x)}{q(x)} = \sum_x p(x) \log \frac{p(x)}{q(x)} = D(p(x)||q(x))$$

- **第二项**：直接展开即为条件相对熵的定义：

$$\sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} = D(p(y|x)||q(y|x))$$

将两项相加，得证。 □

Remark 1.11. 直观理解： 这就好比比较两个二维形状（联合分布）的差异。总差异等于：

1. 在 X 轴投影上的差异（边缘相对熵 $D(p(x)||q(x))$）；
2. 加上在 X 轴位置对齐后，Y 轴剖面形状的平均差异（条件相对熵 $D(p(y|x)||q(y|x))$）。

这在热力学第二定律的推导中会有重要应用。

1.6 Jensen 不等式与信息不等式

这是本章证明不等式的核心数学工具。

Definition 1.7 (凸函数 Convex Function)**.** 函数 $f(x)$ 被称为在区间 (a,b) 上是**凸的**，如果对任意 $x_1, x_2 \in (a,b)$ 和 $0 \le \lambda \le 1$，满足：

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2) \tag{1.47}$$

若仅当 $\lambda = 0$ 或 $\lambda = 1$ 时等号成立，则称函数是**严格凸**的。

Definition 1.8 (凹函数 Concave Function)**.** 如果 $-f$ 是凸函数，则称 f 是凹函数。

Proposition 1.1 (凸性判定)**.** 如果函数 f 在某区间上的二阶导数非负（$f''(x) \ge 0$），则该函数是凸的。

Theorem 1.6 (Jensen 不等式)**.** 如果 f 是凸函数，且 X 是随机变量，则：

$$\mathbb{E}f(X) \geq f(\mathbb{E}X) \tag{1.48}$$

此外，如果 f 是严格凸的，当且仅当 X 是常数（即 $X = \mathbb{E}X$ 概率为 1）时等号成立。

Proof. 我们对概率质量点的个数 k 进行归纳。

1. **当 $k = 2$ 时：**

   $$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

   这直接由凸函数的定义得到。

2. **假设当 $k-1$ 时定理成立**。对于 k 个点分布，设 p_i 为概率。令 $p'_i = \frac{p_i}{1-p_k}$ $(i = 1, \ldots, k-1)$。注意到 $\sum_{i=1}^{k-1} p'_i = 1$。

$$\sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p'_i f(x_i) \tag{1.49}$$

$$\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i \right) \quad \text{(由归纳假设)} \tag{1.50}$$

$$\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p'_i x_i \right) \quad \text{(由 } k=2 \text{ 时的凸性定义)} \tag{1.51}$$

$$= f\left(\sum_{i=1}^{k} p_i x_i \right) = f(\mathbb{E}X) \tag{1.52}$$

对于连续分布，可以通过连续性论证扩展得到。 □

Theorem 1.7 (相对熵非负性)**.** 对于任意两个概率质量函数 $p(x), q(x)$：

$$D(p||q) \geq 0 \tag{1.53}$$

当且仅当 $p(x) = q(x)$ 对所有 x 成立时，等号成立。

Proof. 令 $A = \{x : p(x) > 0\}$ 为 $p(x)$ 的支撑集。

$$-D(p||q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \tag{1.54}$$

$$= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \tag{1.55}$$

$$\leq \log \left(\sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \tag{1.56}$$

$$= \log \left(\sum_{x \in A} q(x) \right) \tag{1.57}$$

$$\leq \log \left(\sum_{x \in \mathcal{X}} q(x) \right) = \log 1 = 0 \tag{1.58}$$

所以 $D(p||q) \geq 0$。由于 $\log t$ 是严格凹的，Jensen 不等式取等号当且仅当随机变量是常数，即 $q(x)/p(x) = c$（常数）。由概率归一化条件可知 $c = 1$，故 $p(x) = q(x)$。 □

这个定理也被称为吉布斯不等式（Gibbs' inequality）。它是信息论许多定理的源头。

Corollary 1.2 (互信息的非负性). 对于任意随机变量 X, Y，

$$I(X;Y) \geq 0 \tag{1.59}$$

当且仅当 X 和 Y 独立时等号成立。

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

Remark 1.12. 这意味着：知晓另一个随机变量 Y 不会增加关于 X 的不确定性。最坏的情况是 Y 与 X 无关，不确定性保持不变，互信息为 0。

Theorem 1.8 (最大熵定理). $H(X) \leq \log|\mathcal{X}|$，其中 $|\mathcal{X}|$ 是取值空间的大小。等号成立当且仅当 X 服从均匀分布。

Proof. 设 $u(x) = 1/|\mathcal{X}|$ 为均匀分布，$p(x)$ 为任意分布。

$$\begin{aligned} D(p||u) &= \sum p(x) \log \frac{p(x)}{u(x)} && (1.60)\\ &= \sum p(x) \log p(x) - \sum p(x) \log u(x) && (1.61)\\ &= -H(X) - \sum p(x) \log(1/|\mathcal{X}|) && (1.62)\\ &= -H(X) + \log|\mathcal{X}| && (1.63) \end{aligned}$$

由相对熵非负性 $D(p||u) \geq 0$，得 $\log|\mathcal{X}| - H(X) \geq 0$，即 $H(X) \leq \log|\mathcal{X}|$。 □

Theorem 1.9 (条件作用减少熵).

$$H(X|Y) \leq H(X) \tag{1.64}$$

等号成立当且仅当 X 和 Y 独立。

Proof. 由互信息的定义和非负性：

$$0 \leq I(X;Y) = H(X) - H(X|Y)$$

故 $H(X|Y) \leq H(X)$。 □

Remark 1.13. 该定理说的是“平均”而言，知道 Y 会降低 X 的不确定性。但是，对于特定的 $Y = y$，条件熵 $H(X|Y = y)$ 有可能大于 $H(X)$。例如：如果 X 大概率取 0，小概率取 1。此时熵很低。如果来了一个观测 Y，告诉我们 X 不可能取 0，那么 X 的分布可能变得更均匀，导致该特定情况下的熵增加。但如果把所有可能的 y 平均起来，熵一定是减少的。

1.7 Log Sum 不等式与凸性

Theorem 1.10 (Log Sum 不等式). 对于非负数 $a_1, \ldots, a_n$ 和 $b_1, \ldots, b_n$：

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^{n} a_i) \log \frac{\sum a_i}{\sum b_i} \tag{1.65}$$

这个纯数学不等式用于证明信息量的凸性/凹性。

约定：我们使用以下约定：$0 \log 0 = 0$，$a \log \frac{a}{0} = \infty$ (若 $a > 0$)，$0 \log \frac{0}{0} = 0$。

Proof. 不失一般性，假设 $a_i > 0$ 且 $b_i > 0$。函数 $f(t) = t \log t$ 是严格凸函数。

根据 Jensen 不等式，对于 $\alpha_i \geq 0$ 且 $\sum_i \alpha_i = 1$，有：

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right) \tag{1.66}$$

令 $\alpha_i = \frac{b_i}{\sum_{j=1}^{n} b_j}$ 且 $t_i = \frac{a_i}{b_i}$，代入上式：

$$\sum \frac{b_i}{\sum b_j} \left(\frac{a_i}{b_i} \log \frac{a_i}{b_i}\right) \geq \left(\sum \frac{b_i}{\sum b_j} \frac{a_i}{b_i}\right) \log \left(\sum \frac{b_i}{\sum b_j} \frac{a_i}{b_i}\right) \tag{1.67}$$

整理后即得到对数和不等式。 □

Theorem 1.11 (相对熵的凸性). $D(p||q)$ 是关于对 (p, q) 的凸函数。即，如果 (p_1, q_1) 和 (p_2, q_2) 是两对概率质量函数，则对于任意 $0 \leq \lambda \leq 1$：

$$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda) D(p_2||q_2) \tag{1.68}$$

Proof. 对不等式左边的每一项应用对数和不等式：

$$(\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \tag{1.69}$$

$$\leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda) p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \tag{1.70}$$

对所有 x 求和即得证。 □

Theorem 1.12 (熵的凹性). $H(p)$ 是关于 p 的凹函数（Concave function）。

Proof. 我们可以将熵写为：

$$H(p) = \log |\mathcal{X}| - D(p||u) \tag{1.71}$$

其中 u 是 $|\mathcal{X}|$ 个结果上的均匀分布。由于 $D(p||u)$ 是 p 的凸函数，且前面有负号，因此 $H(p)$ 是凹函数。 □

Theorem 1.13 (互信息的凹凸性). 设 $(X, Y) \sim p(x, y) = p(x)p(y|x)$。互信息 $I(X; Y)$ 是：

1. 固定 $p(y|x)$ 时，关于 $p(x)$ 的**凹函数**。
2. 固定 $p(x)$ 时，关于 $p(y|x)$ 的**凸函数**。

Proof. **第一部分**：展开互信息：

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X=x) \tag{1.72}$$

如果 $p(y|x)$ 固定，则 $p(y)$ 是 $p(x)$ 的线性函数。因为 $H(Y)$ 是 $p(y)$ 的凹函数，所以也是 $p(x)$ 的凹函数。第二项是 $p(x)$ 的线性函数。两者的差（凹函数 - 线性函数）仍是凹函数。

第二部分：固定 $p(x)$。考虑两个条件分布 $p_1(y|x)$ 和 $p_2(y|x)$ 及其混合 $p_\lambda(y|x)$。互信息可以写为联合分布与边缘分布乘积的相对熵：

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \tag{1.73}$$

由于 $D(p||q)$ 是 (p,q) 的凸函数，互信息也是条件分布的凸函数。 □

Remark 1.14. 1. 相对熵的凸性：$D(p||q)$ 关于 (p,q) 是凸函数。这在证明信道容量定理等最优化问题时非常重要。
2. 熵的凹性：$H(p)$ 关于分布 p 是凹函数。这意味着混合两个分布（例如混合两瓶气体），混合后的熵不小于各自分布熵的加权和（混合增加混乱度）。

1.8 数据处理不等式 (Data-Processing Inequality)

Definition 1.9 (马尔可夫链)**.** 随机变量 X, Y, Z 构成马尔可夫链（记为 $X \to Y \to Z$），如果 Z 的条件分布仅依赖于 Y，而与 X 条件独立。即：

$$p(x,y,z) = p(x)p(y|x)p(z|y) \tag{1.74}$$

推论：$X \to Y \to Z \iff X, Z$ 给定 Y 条件独立 (即 $p(x,z|y) = p(x|y)p(z|y)$)。直观上，这意味着 Z 仅依赖于 Y，而不直接依赖于 X。

Theorem 1.14 (数据处理不等式 DPI)**.** 若 $X \to Y \to Z$，则

$$I(X;Y) \geq I(X;Z) \tag{1.75}$$

Proof. 利用链式法则展开 $I(X;Y,Z)$ 两种方式：

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \tag{1.76}$$

$$= I(X;Y) + I(X;Z|Y) \tag{1.77}$$

由于 $X \to Y \to Z$ 是马尔可夫链，给定 Y 时 X 和 Z 独立，故 $I(X;Z|Y) = 0$。又因为 $I(X;Y|Z) \geq 0$（互信息的非负性），所以：

$$I(X;Y) \geq I(X;Z) \tag{1.78}$$

等号成立当且仅当 $I(X;Y|Z) = 0$，即 $X \to Z \to Y$ 也是马尔可夫链。 □

Remark 1.15. DPI 的深刻内涵：
1. 无法通过处理增加信息：Z 是基于 Y 处理得到的（可能是计算、加噪声等）。定理告诉我们，没有任何针对 Y 的数据处理能够增加 Y 中原本包含的关于 X 的信息。

2. 信息流失：在传输链 $X \to Y \to Z$ 中，信息只会丢失或保持不变，绝不会凭空产生。

Corollary 1.3. 若 $Z = g(Y)$，则 $I(X;Y) \geq I(X;g(Y))$。即对数据做函数变换不会增加信息。

1.9 充分统计量 (Sufficient Statistics)

这是数据处理不等式在统计学中的一个重要应用。

Definition 1.10 (充分统计量)**.** 统计量 $T(X)$ 对于参数 θ 是充分的，如果给定 $T(X)$ 后，X 的分布与 θ 独立（即 $\theta \to T(X) \to X$ 构成马尔可夫链）。

这等价于数据处理不等式取等号的条件：

$$I(\theta;X) = I(\theta;T(X)) \tag{1.79}$$

即充分统计量保留了样本中关于参数的所有互信息。

例子：

1. **伯努利序列：** $X_i \in \{0,1\}$ i.i.d.，$\theta = P(X_i = 1)$。$T(X) = \sum X_i$ 是充分统计量。
2. **正态分布：** $X_i \sim \mathcal{N}(\theta, 1)$。样本均值 $\bar{X}_n$ 是 θ 的充分统计量。
3. **均匀分布：** $X_i \sim \text{Uniform}(\theta, \theta+1)$。$T(X) = (\max X_i, \min X_i)$ 是充分统计量。

Definition 1.11 (最小充分统计量)**.** 如果是任何其他充分统计量的函数，则称该统计量为最小充分统计量。它对数据进行了最大程度的压缩。

1.10 Fano 不等式 (Fano's Inequality)

这是从信息论联系到估计理论的桥梁。它给出了估计误差的下界。

假设我们想从 Y 估计 X，估计量为 $\hat{X} = g(Y)$。定义误差概率 $P_e = \Pr(X \neq \hat{X})$。注意到 $X \to Y \to \hat{X}$ 构成马尔可夫链。我们没有要求 $\hat{X}$ 的取值也在 $\mathcal{X}$ 上。

Theorem 1.15 (Fano 不等式)**.** 对于任何满足 $X \to Y \to \hat{X}$ 的估计量 $\hat{X}$，令 $P_e = \Pr(X \neq \hat{X})$，则：

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y) \tag{1.80}$$

此不等式可弱化为：

$$1 + P_e \log |\mathcal{X}| \geq H(X|Y) \tag{1.81}$$

或

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|} \tag{1.82}$$

Proof. 定义误差随机变量 E：

$$E = \begin{cases} 1 & \text{若} \hat{X} \neq X \\ 0 & \text{若} \hat{X} = X \end{cases} \tag{1.83}$$

利用链式法则展开 $H(E, X|\hat{X})$：

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) && (1.84) \\ &= H(X|\hat{X}) + 0 \quad (\text{给定} X, \hat{X}, E \text{ 是确定的}) && (1.85) \\ &= H(X|\hat{X}) && (1.86) \end{aligned}$$

另一种展开方式：

$$\begin{aligned} H(E, X|\hat{X}) &= H(E|\hat{X}) + H(X|E, \hat{X}) && (1.87) \\ &\leq H(E) + H(X|E, \hat{X}) \quad (\text{条件减少熵}) && (1.88) \\ &= H(P_e) + P(E=0)H(X|\hat{X}, E=0) + P(E=1)H(X|\hat{X}, E=1) && (1.89) \end{aligned}$$

分析各项：

- 若 $E = 0$，则 $X = \hat{X}$，没有不确定性，熵为 0。
- 若 $E = 1$，估计值在集合 $\mathcal{X}$ 中取值，则 $X \neq \hat{X}$，X 可以是剩下的 $|\mathcal{X}| - 1$ 个值中的任意一个，熵上界为 $\log(|\mathcal{X}| - 1) \leq \log|\mathcal{X}|$。若估计值不在集合 $\mathcal{X}$ 中取值，则我们只能用最大熵定理以 $\log|\mathcal{X}|$ 控制上界。

因此：

$$H(X|\hat{X}) \leq H(P_e) + P_e \log|\mathcal{X}| \tag{1.90}$$

再由数据处理不等式 $I(X; \hat{X}) \leq I(X; Y)$ 可得 $H(X|\hat{X}) \geq H(X|Y)$，证毕。 □

Corollary 1.4. 如果估计值必须在集合 $\mathcal{X}$ 中取值，可以将不等式加强为：

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y) \tag{1.91}$$

Remark 1.16. Fano 不等式的意义：

- 如果条件熵 $H(X|Y)$ 很大（说明即使知道 Y，对 X 的不确定性依然很高），那么任何估计器的错误概率 P_e 都不可能很小。
- 它是香农信道编码定理逆定理证明的关键：如果我们要无误差传输（$P_e \to 0$），必须使得条件熵 $H(X|Y) \to 0$。

Lemma 1.3. 如果 X 和 X' 是独立同分布 (i.i.d.) 的，具有熵 $H(X)$，则：

$$\Pr(X = X') \geq 2^{-H(X)} \tag{1.92}$$

当且仅当 X 为均匀分布时等号成立。

Proof. 利用 Jensen 不等式：$2^{\mathbb{E}\log p(X)} \leq \mathbb{E}[2^{\log p(X)}]$。

$$2^{-H(X)} = 2^{\sum p(x)\log p(x)} \leq \sum p(x) 2^{\log p(x)} = \sum p^2(x) = \Pr(X = X') \tag{1.93}$$

□

Corollary 1.5. 设 X 和 X' 是独立的随机变量，其中 $X \sim p(x)$，$X' \sim r(x)$，$x, x' \in \mathcal{X}$。那么：

$$\Pr(X = X') \geq 2^{-H(p)-D(p||r)} \tag{1.94}$$

$$\Pr(X = X') \geq 2^{-H(r)-D(r||p)} \tag{1.95}$$

Proof. 我们证明第一个不等式：

$$\begin{aligned}
2^{-H(p)-D(p||r)} &= 2^{\sum p(x)\log p(x)+\sum p(x)\log\frac{r(x)}{p(x)}} && (1.96)\\
&= 2^{\sum p(x)(\log p(x)+\log r(x)-\log p(x))} && (1.97)\\
&= 2^{\sum p(x)\log r(x)} && (1.98)\\
&\leq \sum p(x)2^{\log r(x)} \quad \text{(由 Jensen 不等式及} y = 2^x \text{ 的凸性)} && (1.99)\\
&= \sum p(x)r(x) && (1.100)\\
&= \Pr(X = X') && (1.101)
\end{aligned}$$

其中，不等式步骤使用了 Jensen 不等式：对于凸函数 $f(y) = 2^y$，有 $2^{\mathbb{E}[Y]} \leq \mathbb{E}[2^Y]$。这里期望是针对分布 $p(x)$ 取的，随机变量是 $\log r(X)$。 □

Remark 1.17. 解释：之前引理说的是两个 X 分布相同时，碰撞概率（$X = X'$）至少是 $2^{-H(X)}$。这个推论说的是如果两个分布 p 和 r 不同，碰撞概率会更低。指数部分不仅仅是熵 $H(p)$，还要减去相对熵 $D(p||r)$。由于 $D(p||r) \geq 0$，这说明分布差异越大（D 越大），2^{-H-D} 这个下界就越小，符合直觉。

2 渐进均分性

在信息论中，渐进均分性质 (Asymptotic Equipartition Property, AEP) 是大数定律 (Law of Large Numbers, LLN) 的对应物。它是弱大数定律的直接推论。

- **大数定律回顾**：对于独立同分布 (i.i.d.) 的随机变量 $X_1, X_2, \ldots, X_n$，当 n 很大时，样本均值 $\frac{1}{n}\sum_{i=1}^{n} X_i$ 接近于其期望值 $\mathbb{E}X$。
- **AEP 的核心思想**：对于 i.i.d. 随机变量，$\frac{1}{n}\log\frac{1}{p(X_1,\ldots,X_n)}$ 接近于熵 $H(X)$。这意味着，观测序列 $X_1, \ldots, X_n$ 的概率 $p(X_1, \ldots, X_n)$ 接近于 2^{-nH}。

直观解释： 这使我们能够将所有可能的序列分为两类：

1. 典型集 (Typical Set)：样本熵接近真实熵的集合。我们要重点关注这一类。
2. 非典型集 (Nontypical Set)：包含其他所有序列的集合。

任何在典型序列上被证明的性质，都将以极高的概率成立，从而决定了大样本的平均行为。

2.1 AEP 定理

在正式陈述定理之前，我们需要回顾随机变量收敛的概念。

Definition 2.1 (随机变量的收敛)**.** 给定随机变量序列 $X_1, X_2, \ldots$，我们说该序列收敛于随机变量 X：

1. **依概率收敛 (In probability)**：如果对于任意 $\epsilon > 0$，
$$\lim_{n\to\infty} \Pr\{|X_n - X| > \epsilon\} = 0$$
2. **均方收敛 (In mean square)**：如果 $\mathbb{E}(X_n - X)^2 \to 0$。
3. **以概率 1 收敛 (With probability 1 / almost surely)**：如果 $\Pr\{\lim_{n\to\infty} X_n = X\} = 1$。

2.1.1 AEP 定理陈述与证明

Theorem 2.1 (AEP)**.** 如果 $X_1, X_2, \ldots$ 是 i.i.d. $\sim p(x)$，则
$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \quad \text{依概率收敛.} \tag{2.1}$$

Proof. 独立随机变量的函数也是独立的随机变量。因此，由于 X_i 是 i.i.d. 的，那么 $\log p(X_i)$ 也是 i.i.d. 的。根据弱大数定律 (WLLN)：

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) \tag{2.2}$$
$$\to -\mathbb{E}[\log p(X)] \quad \text{依概率收敛} \tag{2.3}$$
$$= H(X). \tag{2.4}$$

证毕。 □

2.2 典型集 (The Typical Set)

AEP 告诉我们观测序列的概率大致是 2^{-nH}。我们将其形式化为典型集的定义。

Definition 2.2 (典型集). 关于 $p(x)$ 的典型集 $A_\epsilon^{(n)}$ 是序列 $(x_1,\dots,x_n)\in\mathcal{X}^n$ 的集合，满足以下性质：

$$2^{-n(H(X)+\epsilon)}\le p(x_1,\dots,x_n)\le 2^{-n(H(X)-\epsilon)}. \tag{2.5}$$

这等价于：

$$H(X)-\epsilon\le -\frac{1}{n}\log p(x^n)\le H(X)+\epsilon.$$

2.2.1 典型集的性质

作为 AEP 的推论，典型集具有以下四个重要性质：

Theorem 2.2 (典型集的性质).

1. 如果 $(x_1,\dots,x_n)\in A_\epsilon^{(n)}$，则 $p(x^n)\approx 2^{-nH(X)}$。具体来说：
$$H(X)-\epsilon\le -\frac{1}{n}\log p(x^n)\le H(X)+\epsilon.$$
2. 对于足够大的 n，$\Pr\{A_\epsilon^{(n)}\}>1-\epsilon$。
3. $|A_\epsilon^{(n)}|\le 2^{n(H(X)+\epsilon)}$，其中 $|A|$ 表示集合 A 的元素个数。
4. 对于足够大的 n，$|A_\epsilon^{(n)}|\ge(1-\epsilon)2^{n(H(X)-\epsilon)}$。

Proof. **性质 1**：直接由典型集的定义可得。

性质 2：由定理 2.1 (AEP) 可知，事件 $\{|-\frac{1}{n}\log p(X^n)-H(X)|<\epsilon\}$ 的概率随着 $n\to\infty$ 趋向于 1。因此，对于任意 $\delta>0$，存在 n_0，使得对于所有 $n\ge n_0$，概率大于 $1-\delta$。令 $\delta=\epsilon$ 即得证。

性质 3：利用全概率和为 1 的性质：

$$\begin{aligned}
1&=\sum_{x^n\in\mathcal{X}^n}p(x^n) && (2.6)\\
&\ge\sum_{x^n\in A_\epsilon^{(n)}}p(x^n) && (2.7)\\
&\ge\sum_{x^n\in A_\epsilon^{(n)}}2^{-n(H(X)+\epsilon)}\quad\text{(根据定义中的概率下界)} && (2.8)\\
&=|A_\epsilon^{(n)}|2^{-n(H(X)+\epsilon)}. && (2.9)
\end{aligned}$$

移项得 $|A_\epsilon^{(n)}|\le 2^{n(H(X)+\epsilon)}$。

性质 4：对于足够大的 n，我们要利用 $\Pr\{A_\epsilon^{(n)}\}>1-\epsilon$：

$$\begin{aligned}
1-\epsilon&<\Pr\{A_\epsilon^{(n)}\} && (2.10)\\
&\le\sum_{x^n\in A_\epsilon^{(n)}}2^{-n(H(X)-\epsilon)}\quad\text{(根据定义中的概率上界)} && (2.11)\\
&=|A_\epsilon^{(n)}|2^{-n(H(X)-\epsilon)}. && (2.12)
\end{aligned}$$

移项得 $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$。 □

总结与深入理解： 这四个性质告诉我们关于典型集的三个核心事实：

- 典型集的概率接近 1。
- 典型集中的所有元素几乎是等概率的。
- 典型集中的元素个数接近 2^{nH}。

注意：总序列数是 $|\mathcal{X}|^n = 2^{n\log|\mathcal{X}|}$。因为 $H(X) \leq \log|\mathcal{X}|$，所以典型集通常只是所有可能序列集的一个极小的子集，但它承载了绝大多数的概率质量。

2.3 AEP 的应用：数据压缩

既然大部分概率都集中在典型集 $A_\epsilon^{(n)}$ 中，且其中仅有约 2^{nH} 个序列，我们可以利用这一事实进行高效编码。

2.3.1 编码方案

我们将所有序列 $\mathcal{X}^n$ 分为两个集合：典型集 $A_\epsilon^{(n)}$ 和非典型集 $(A_\epsilon^{(n)})^c$。
编码策略：

- 对典型集中的元素进行排序。因为元素个数 $\leq 2^{n(H+\epsilon)}$，我们需要的索引长度不超过 $n(H+\epsilon)+1$ 位。为了区分，在这些码字前加一个 0。
- 对非典型集中的元素进行排序。元素个数最多为 $|\mathcal{X}|^n$，需要的索引长度不超过 $n\log|\mathcal{X}|+1$ 位。在这些码字前加一个 1。

这个前缀位充当了标志位，使得码字长度可变但易于解码。

2.3.2 平均码长分析

设 $l(x^n)$ 为序列 x^n 的码字长度。期望长度为：

$$\mathbb{E}[l(X^n)] = \sum_{x^n} p(x^n)l(x^n) \tag{2.13}$$

我们可以将其分为典型集和非典型集两部分：

$$\mathbb{E}[l(X^n)] = \sum_{x^n\in A_\epsilon^{(n)}} p(x^n)l(x^n) + \sum_{x^n\in (A_\epsilon^{(n)})^c} p(x^n)l(x^n) \tag{2.14}$$

$$\leq \Pr(A_\epsilon^{(n)})[n(H+\epsilon)+2] + \Pr((A_\epsilon^{(n)})^c)[n\log|\mathcal{X}|+2] \tag{2.15}$$

其中 $+2$ 来自于向上取整所需的 $+1$ 和前缀标志位 $+1$。

利用 $\Pr(A_\epsilon^{(n)}) \leq 1$ 和 $\Pr((A_\epsilon^{(n)})^c) < \epsilon$ (当 n 足够大时)：

$$\mathbb{E}[l(X^n)] \leq 1\cdot[n(H+\epsilon)+2] + \epsilon\cdot[n\log|\mathcal{X}|+2] \tag{2.16}$$

$$= n(H+\epsilon+\epsilon\log|\mathcal{X}|) + 2(1+\epsilon) \tag{2.17}$$

$$= n(H+\epsilon') \tag{2.18}$$

其中 ϵ' 可以通过选择足够小的 ϵ 和足够大的 n 变得任意小。

由此我们得到定理：

Theorem 2.3 (信源编码定理). 设 X^n 是 i.i.d. $\sim p(x)$。对于任意 $\epsilon > 0$，存在一个将长度为 n 的序列 x^n 映射为二进制串的一对一编码，使得当 n 足够大时：

$$\mathbb{E}\left[\frac{1}{n}l(X^n)\right] \leq H(X) + \epsilon. \tag{2.19}$$

这意味着我们可以平均用 $nH(X)$ 个比特来表示序列 X^n。

2.4 高概率集与典型集

一个自然的问题是：典型集 $A_\epsilon^{(n)}$ 是否是包含大部分概率的最小集合？

Definition 2.3 (最小大概率集). 对于 $\delta < 1/2$，令 $B_\delta^{(n)} \subset \mathcal{X}^n$ 是满足 $\Pr\{B_\delta^{(n)}\} \geq 1-\delta$ 的最小集合。

我们需要比较 $|A_\epsilon^{(n)}|$ 和 $|B_\delta^{(n)}|$。
首先定义一个新的符号来表示“指数阶相等”。

Definition 2.4 (指数阶相等). 符号 $a_n \doteq b_n$ 表示

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{a_n}{b_n} = 0.$$

这意味着 a_n 和 b_n 在指数的一阶上是相等的。

Theorem 2.4 (最小集与典型集的关系). 令 $X_1, \ldots, X_n$ i.i.d. $\sim p(x)$。对于 $\delta < 1/2$ 和任意 $\delta' > 0$，如果 $\Pr\{B_\delta^{(n)}\} > 1-\delta$，则对于足够大的 n：

$$\frac{1}{n}\log|B_\delta^{(n)}| > H - \delta'. \tag{2.20}$$

结论：

$$|B_\delta^{(n)}| \doteq |A_\epsilon^{(n)}| \doteq 2^{nH}. \tag{2.21}$$

深入理解：典型集 vs. 最可能序列 为了理解这一节的精髓，我们来看一个重要的例子（Bernoulli 序列）：假设 $X \in \{0,1\}$，且 $p(1) = 0.9, p(0) = 0.1$。

- 最可能的单个序列：全 1 序列 $(1,1,\ldots,1)$。它的概率是 0.9^n。这个序列**不属于**典型集，因为它的经验熵是 0，而真实熵 $H(X) > 0$。
- 典型序列：那些包含大约 $0.9n$ 个 1 和 $0.1n$ 个 0 的序列。单个典型序列的概率约为 2^{-nH}，远小于全 1 序列的概率。
- 为什么关注典型集？虽然全 1 序列概率最大，但这类极端序列的数量极少。典型序列虽然单个概率小，但它们的数量巨大。这大量的典型序列加起来占据了几乎所有的概率空间。
- $B_\delta^{(n)}$ 会首先包含全 1 序列（因为它是最可能的），但为了凑够 $1-\delta$ 的总概率，它必须包含那些典型的序列。因此，$B_\delta^{(n)}$ 的大小在指数级上与 $A_\epsilon^{(n)}$ 相同。

3 随机过程的熵率

渐进均分性（AEP）确立了 $nH(X)$ 比特平均足以描述 n 个独立同分布 (i.i.d.) 的随机变量。如果随机变量是相关的，特别是形成平稳过程时，情况会如何？

我们将证明，对于平稳过程，熵 $H(X_1,\dots,X_n)$ 随 n 渐进线性增长，其增长率为 $H(\mathcal{X})$，称为该过程的**熵率**。

3.1 马尔可夫链 (Markov Chains)

3.1.1 基本定义

Definition 3.1 (随机过程)**.** 随机过程 $\{X_i\}$ 是一个随机变量的索引序列。该过程由其联合概率质量函数表征：

$$\Pr\{(X_1,\dots,X_n)=(x_1,\dots,x_n)\}=p(x_1,\dots,x_n),\quad n=1,2,\dots \tag{3.1}$$

Definition 3.2 (平稳性)**.** 如果一个随机过程的任何子集的联合分布对时间下标的平移是不变的，则称该随机过程是平稳的。即对于任意 n 和位移 l：

$$\Pr\{X_1=x_1,\dots,X_n=x_n\}=\Pr\{X_{1+l}=x_1,\dots,X_{n+l}=x_n\} \tag{3.2}$$

Remark 3.1. 平稳性意味着过程的统计特性不随时间推移而改变。例如，$p(X_1=x)$ 必须等于 $p(X_n=x)$。这并不意味着 X_1 和 X_2 独立，只是说它们的边际分布和相关结构在时间轴上是“平移不变”的。

Definition 3.3 (马尔可夫链)**.** 如果一个离散随机过程 $\{X_i\}$ 满足以下条件，则称为马尔可夫链：

$$\Pr(X_{n+1}=x_{n+1}|X_n=x_n,X_{n-1}=x_{n-1},\dots,X_1=x_1)=\Pr(X_{n+1}=x_{n+1}|X_n=x_n) \tag{3.3}$$

即：给定现在，未来与过去独立。

对于马尔可夫链，联合分布可以简化为：

$$p(x_1,\dots,x_n)=p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}) \tag{3.4}$$

3.1.2 时不变性与平稳分布

对于一个马尔科夫链 $\{X_i\}$，我们称 X_n 为时间 n 时的状态。如果条件概率 $p(x_{n+1}|x_n)$ 不依赖于 n，则称马尔可夫链是**时不变的 (Time Invariant)**。如未特别声明，我们总假设马尔可夫链是时不变的。此时马尔科夫链由初始状态和概率转移矩阵 $P=[P_{ij}]$ 刻画，其中 $P_{ij}=\Pr(X_{n+1}=j|X_n=i)$。下面再给出一些定义：

- **不可约 (Irreducible)**: 从任意状态出发，能在有限步内以正概率到达任意其他状态。
- **非周期 (Aperiodic)**: 状态回路长度的最大公约数为 1。

如果概率质量函数为 $p(x_n)$，则下一时刻的分布为 $p(x_{n+1}) = \sum_{x_n} p(x_n)P_{x_n x_{n+1}}$。用向量表示即 $\mu_{n+1} = \mu_n P$。

Definition 3.4 (平稳分布). 如果分布 μ 满足 $\mu P = \mu$，则称 μ 为平稳分布。

Remark 3.2. 如果初始状态 X_1 服从平稳分布 μ，那么整个过程 $X_1, X_2, \ldots$ 就是平稳过程。对于不可约非周期的有限状态马尔可夫链，平稳分布存在且唯一。

3.2 熵率 (Entropy Rate)

对于一个随机变量序列，我们关心其熵随 n 的增长速度。

3.2.1 两种定义及其等价性

Definition 3.5 (熵率定义 1). 随机过程 $\mathcal{X} = \{X_i\}$ 的熵率定义为：

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{3.5}$$

如果该极限存在。这代表**平均每个符号所需的比特数**。

Definition 3.6 (熵率定义 2).

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1) \tag{3.6}$$

如果该极限存在。这代表**给定过去的历史，当前符号带来的不确定性（新信息量）**。

对于平稳过程，这两个定义是等价的。这是一个非常重要的结论。

Theorem 3.1 (平稳过程的熵率). 对于平稳随机过程，定义 1 和定义 2 的极限均存在且相等：

$$H(\mathcal{X}) = H'(\mathcal{X}) \tag{3.7}$$

为了证明这一点，我们需要先引入 Cesáro 均值定理。

Theorem 3.2 (Cesáro Mean). 如果序列 $a_n \to a$，且 $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$，那么 $b_n \to a$。

Remark 3.3. 直观理解：如果一组成绩逐渐稳定在 80 分附近，那么平均成绩最终也会稳定在 80 分。

定理 3.1 的证明思路：

1. 首先证明 $H(X_n | X_{n-1}, \ldots, X_1)$ 随 n 非递增（因为条件越多，熵越小），且非负，因此极限 $H'(\mathcal{X})$ 存在。

$$H(X_{n+1} | X_1, \ldots, X_n) \le H(X_{n+1} | X_2, \ldots, X_n) = H(X_n | X_1, \ldots, X_{n-1}) \tag{3.8}$$

(不等式由“条件降低熵”得，等式由平稳性得)。

2. 利用链式法则：
$$\frac{H(X_1,\dots,X_n)}{n}=\frac{1}{n}\sum_{i=1}^{n}H(X_i|X_{i-1},\dots,X_1) \tag{3.9}$$
这是条件熵序列的“算术平均值”。
3. 根据 Cesáro 均值定理，既然项 $H(X_i|X_{i-1},\dots)$ 收敛到 $H'(\mathcal{X})$，那么它们的平均值也收敛到 $H'(\mathcal{X})$。

3.2.2 马尔可夫链的熵率

对于平稳马尔可夫链，计算熵率非常简单，因为其“记忆”有限。

$$H(\mathcal{X})=\lim_{n\to\infty}H(X_n|X_{n-1},\dots,X_1)=\lim_{n\to\infty}H(X_n|X_{n-1})=H(X_2|X_1) \tag{3.10}$$
其中最后一个等式是在平稳分布假设下成立的。

Theorem 3.3 (马尔可夫链熵率公式)**.** 设 $\{X_i\}$ 为平稳马尔可夫链，平稳分布为 μ，转移矩阵为 P。则熵率为：
$$H(\mathcal{X})=-\sum_i \mu_i\sum_j P_{ij}\log P_{ij} \tag{3.11}$$

Proof.
$$H(\mathcal{X})=H(X_2|X_1)=\sum_i p(x_1=i)H(X_2|X_1=i) \tag{3.12}$$
$$=\sum_i \mu_i\left(-\sum_j P_{ij}\log P_{ij}\right) \tag{3.13}$$
□

Example 3.1 (两状态马尔可夫链)**.** 转移矩阵 $P=\begin{bmatrix}1-\alpha & \alpha\\ \beta & 1-\beta\end{bmatrix}$。平稳分布为 $\mu_1=\frac{\beta}{\alpha+\beta},\mu_2=\frac{\alpha}{\alpha+\beta}$。熵率为：
$$H(\mathcal{X})=\frac{\beta}{\alpha+\beta}H(\alpha)+\frac{\alpha}{\alpha+\beta}H(\beta) \tag{3.14}$$
注意这里的 $H(\alpha)$ 指的是二元熵函数 $-\alpha\log\alpha-(1-\alpha)\log(1-\alpha)$。

3.2.3 加权图上的随机游走

考虑一个无向连通图，边上有权重 $W_{ij}\geq 0$ ($W_{ij}=W_{ji}$)。粒子从节点 i 转移到 j 的概率与权重成正比：
$$P_{ij}=\frac{W_{ij}}{\sum_k W_{ik}}=\frac{W_{ij}}{W_i} \tag{3.15}$$
其中 $W_i=\sum_j W_{ij}$ 是从节点 i 出发的总权重。

平稳分布：我们可以猜测并验证平稳分布非常简单：

$$\mu_i = \frac{W_i}{2W} \tag{3.16}$$

其中 $W = \sum_{i,j:j>i} W_{ij}$ 是所有边的总权重之和。

Remark 3.4. 直观理解：连接边权重越大的节点，被访问的概率越高。

随机游走的熵率：

$$H(\mathcal{X}) = H\left(\ldots, \frac{W_{ij}}{2W}, \ldots\right) - H\left(\ldots, \frac{W_i}{2W}, \ldots\right) \tag{3.17}$$

特例：如果所有边的权重相等（即简单随机游走），设 E 为总边数，E_i 为节点 i 的度数，则：

$$H(\mathcal{X}) = \log(2E) - H\left(\frac{E_1}{2E}, \frac{E_2}{2E}, \ldots, \frac{E_m}{2E}\right) \tag{3.18}$$

这个公式说明熵率只取决于图的几何结构（边数和度数分布）。

3.3 热力学第二定律 (Second Law of Thermodynamics)

虽然孤立系统的物理熵被认为是不减的，但在随机过程模型中，我们发现 ** 相对熵（Relative Entropy）总是减少的 **。

3.3.1 相对熵随时间减少

设 μ_n 和 μ'_n 是马尔可夫链在时刻 n 的两个不同的概率分布。它们随时间演化为 μ_{n+1} 和 μ'_{n+1}。

Theorem 3.4.

$$D(\mu_n||\mu'_n) \geq D(\mu_{n+1}||\mu'_{n+1}) \tag{3.19}$$

Proof. 利用相对熵的链式法则和数据处理不等式的思想。因为 $p(x_{n+1}|x_n) = q(x_{n+1}|x_n)$（同一个马尔可夫链），条件相对熵项消失。 □

推论：如果 μ' 是平稳分布 μ，则：

$$D(\mu_n||\mu) \geq D(\mu_{n+1}||\mu) \tag{3.20}$$

这意味着：**任何初始分布都会随时间推移越来越接近平稳分布（在相对熵意义下）。**

3.3.2 熵的增加与均匀分布

通常情况下，分布趋向于平稳分布意味着 $H(X_n)$ 不一定是单调的。但是，**如果平稳分布是均匀分布**，那么：

$$D(\mu_n||\mu) = \log|\mathcal{X}| - H(X_n) \tag{3.21}$$

因为 $D(\mu_n||\mu)$ 随 n 减小，这意味着 $\log|\mathcal{X}| - H(X_n)$ 减小，即 $**H(X_n)$ 随 n 单调增加 **。

什么情况下平稳分布是均匀的？当且仅当转移矩阵 P 是**双随机矩阵 (Doubly Stochastic)**，即行和为 1 且列和为 1 ($\sum_i P_{ij} = 1$ 且 $\sum_j P_{ij} = 1$)。

3.4 马尔可夫链的函数 (Functions of Markov Chains)

设 $X_1, X_2, \ldots$ 是平稳马尔可夫链，$Y_i = \phi(X_i)$ 是状态的函数。注意：Y 序列本身通常**不是**马尔可夫链。这在隐马尔可夫模型 (HMM) 中很常见。

计算 $H(\mathcal{Y})$ 的问题在于直接计算极限收敛很慢。我们可以使用上界和下界逼近。

3.4.1 上界和下界

上界： $H(Y_n|Y_{n-1}, \ldots, Y_1)$。由于“条件降低熵”，这个序列单调递减趋向于 $H(\mathcal{Y})$。

下界： $H(Y_n|Y_{n-1}, \ldots, Y_1, X_1)$。这是一个技巧，利用了 X_1 包含了关于 Y 历史的所有信息（甚至更多）。

Theorem 3.5 (函数的熵率收敛)**.** 如果 $Y_i = \phi(X_i)$，则：

$$H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) \leq H(\mathcal{Y}) \leq H(Y_n|Y_{n-1}, \ldots, Y_1) \tag{3.22}$$

并且随着 $n \to \infty$，上下界收敛到同一个值 $H(\mathcal{Y})$。

4 数据压缩

本章确立了熵（Entropy）作为数据压缩的基本极限。数据压缩的核心思想是：**给频繁出现的信源符号分配短的描述，给不常出现的符号分配长的描述**。

本章的主要目标是找到随机变量的最短平均描述长度（Shortest Average Description Length）。我们将证明：

1. 期望描述长度必须大于或等于熵 $H(X)$。
2. 我们可以构建编码使期望长度渐进地达到熵。
3. 霍夫曼编码（Huffman Coding）是最优的。

4.1 码的例子

Definition 4.1 (信源编码)**.** 随机变量 X 的信源编码（source code）C 是从 X 的取值范围 $\mathcal{X}$ 到 $\mathcal{D}^*$ 的映射。其中 $\mathcal{D}^*$ 是 D 元字母表 $\mathcal{D}$ 上的有限长符号串集合。$C(x)$ 表示对应于 x 的码字（codeword），$l(x)$ 表示 $C(x)$ 的长度。不失一般性的，我们可以假定 $\mathcal{D} = \{0, 1, \ldots, D-1\}$.

Definition 4.2 (期望长度)**.** 设随机变量 X 的概率质量函数为 $p(x)$，信源编码 $C(x)$ 的期望长度 $L(C)$ 定义为：

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x) \tag{4.1}$$

4.1.1 码的分类

为了保证能够无歧义地解码，我们需要对码加一些限制条件。

Definition 4.3 (非奇异码 Nonsingular)**.** 如果 $\mathcal{X}$ 中的每个元素都映射到 $\mathcal{D}^*$ 中不同的字符串，即：

$$x \neq x' \Rightarrow C(x) \neq C(x') \tag{4.2}$$

则称该码为非奇异的。

非奇异性只能保证单个符号能被区分。但通常我们要发送符号序列。如果仅仅是非奇异，发送序列时可能产生歧义。

Definition 4.4 (码的扩展 Extension)**.** 码 C 的扩展 C^* 是从 $\mathcal{X}$ 的有限长字符串到 $\mathcal{D}$ 的有限长字符串的映射，定义为：

$$C(x_1 x_2 \cdots x_n) = C(x_1) C(x_2) \cdots C(x_n) \tag{4.3}$$

即码字的直接拼接。

Definition 4.5 (唯一可译码 Uniquely Decodable)**.** 如果一个码的扩展是非奇异的，则称该码是唯一可译的。

唯一可译码意味着任何编码后的字符串只能对应一种唯一的信源符号序列。但是，唯一可译码可能需要等到接收完整个字符串才能开始解码（需要“前瞻”）。

Definition 4.6 (前缀码/即时码 Prefix Code / Instantaneous Code)**.** 如果没有任何一个码字是另一个码字的前缀，则称该码为前缀码或即时码。

- **即时性**：一收到某个码字的最后一个符号，立刻就能识别出该信源符号，无需等待后续符号。就像在一段话中看到了逗号，立刻知道句子结束了（Self-punctuating）。
- **包含关系**：所有前缀码都是唯一可译码；所有唯一可译码都是非奇异码。即：前缀码 $\subset$ 唯一可译码 $\subset$ 非奇异码。

4.2 Kraft 不等式

我们在寻找最短的即时码。我们不能让所有码字都很短，否则会违反前缀条件。Kraft 不等式限制了即时码可能的码长集合。

Theorem 4.1 (Kraft 不等式)**.** 对于 D 元字母表上的任意即时码，其码字长度 $l_1, l_2, \ldots, l_m$ 必须满足不等式：

$$\sum_i D^{-l_i} \leq 1 \tag{4.4}$$

反之，给定一组满足该不等式的码长，必存在一个具有这些码长的即时码。

证明. 考虑一棵 D 叉树。

1. 树的每个节点有 D 个子节点。从根节点出发的分支代表码字的符号。
2. 每个码字由树上的一个叶子节点表示。
3. **前缀条件**意味着：如果一个节点被选为码字，它的任何后代节点都不能再作为码字。因此，每个码字“消除”了树上其下方的所有潜在码字。
4. 令 $l_{\max}$ 为最大码长。考虑树在 $l_{\max}$ 层的节点总数，为 $D^{l_{\max}}$。
5. 一个长度为 l_i 的码字，在 $l_{\max}$ 层支配（遮挡）了 $D^{l_{\max}-l_i}$ 个后代节点。
6. 由于前缀性质，这些被支配的集合是不相交的。所有这些集合的节点总数不能超过 $l_{\max}$ 层的总节点数：

$$\sum_i D^{l_{\max}-l_i} \leq D^{l_{\max}} \tag{4.5}$$

7. 两边同时除以 $D^{l_{\max}}$，得 $\sum D^{-l_i} \leq 1$。

反之：如果满足不等式，我们可以利用同样的树结构，按长度从小到大排序，依次分配第一个可用的节点，并在分配后移除其子树，即可构造出前缀码。 □

Theorem 4.2 (扩展 Kraft 不等式)**.** 对于任何可数无限的码字集合形成的前缀码，码长满足 $\sum_{i=1}^{\infty} D^{-l_i} \leq 1$。反之亦然。

Proof. 将码字 $y_1 \cdots y_{l_i}$ 映射为实数区间 $[(0.y_1 \cdots y_{l_i})_D, (0.y_1 \cdots y_{l_i})_D + D^{-l_i})$，其中 $(0.y_1 \cdots y_{l_i})_D$ 是 D 进数表示。**不同码字对应的区间不然是包含关系不然是不相交的**。前缀条件保证了这些区间互不相交且都在 $[0,1]$ 内，因此总长度 ≤ 1。反之，若要构建前缀码，只需要对区间 $[0,1]$ 进行分割并将子区间左端点作为码字。 □

扩展 Kraft 不等式的证明是简洁且有趣的。类比扩展 Kraft 不等式的证明我们也自然能够给出 Kraft 不等式的另一个更加简单的证明。

4.3 最优码 (Optimal Codes)

我们的目标是：在满足 Kraft 不等式 $\sum D^{-l_i} \leq 1$ 的约束下，最小化期望长度 $L = \sum p_i l_i$。

4.3.1 直观推导

为了最小化期望长度，我们需要码字尽可能短。由 Kraft 不等式，如果 $\sum D^{-l_i} < 1$，直观上我们还能把码字变得更短使求和更接近 1，故我们假定约束为 $\sum D^{-l_i} = 1$。忽略 l_i 必须是整数的约束，使用拉格朗日乘数法：

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

对 l_i 求导并令为 0，可得 $p_i = \lambda (\ln D) D^{-l_i}$。代入约束条件解得 $\lambda = 1/\ln D$，最终得到最优码长：

$$l_i^* = -\log_D p_i \tag{4.6}$$

此时期望长度 $L = \sum p_i (-\log_D p_i) = H_D(X)$。这是理想情况。

4.3.2 主要定理

Theorem 4.3 (期望长度下界)**.** 任何随机变量 X 的即时码的期望长度 L 大于或等于熵 $H_D(X)$：

$$L \geq H_D(X) \tag{4.7}$$

当且仅当 $D^{-l_i} = p_i$ 时取等号。

证明. 计算 $L - H_D(X)$：

$$\begin{aligned} L - H_D(X) &= \sum p_i l_i - \sum p_i \log_D \frac{1}{p_i} && (4.8)\\ &= \sum p_i \log_D D^{l_i} + \sum p_i \log_D p_i && (4.9)\\ &= \sum p_i \log_D (p_i D^{l_i}) = \sum p_i \log_D \frac{p_i}{D^{-l_i}} && (4.10) \end{aligned}$$

令 $r_i = \frac{D^{-l_i}}{\sum_j D^{-l_j}}$，这是一个归一化的概率分布。令 $C = \sum D^{-l_i} \leq 1$。上式变形为：

$$\begin{aligned} \sum p_i \log_D \frac{p_i}{r_i C} &= \sum p_i \log_D \frac{p_i}{r_i} - \log_D C && (4.11)\\ &= D(\mathbf{p}||\mathbf{r}) + \log_D \frac{1}{C} && (4.12) \end{aligned}$$

由于相对熵 $D(\mathbf{p}||\mathbf{r}) \geq 0$ 且 $C \leq 1 \Rightarrow \log(1/C) \geq 0$，因此 $L - H_D(X) \geq 0$。 □

Remark 4.1. 提示:

- $l_i = -\log_D p_i$ 通常不是整数，所以实际编码长度通常略大于熵。
- 如果一个分布满足 $p_i = D^{-n_i}$（即所有概率都是 $1/D$ 的幂），称为 **D-adic 分布**，此时可以完美达到熵。

4.4 最优码长的界

既然 $l_i = -\log_D p_i$ 是理想长度，我们可以向上取整来构造整数码长。

4.4.1 香农编码 (Shannon Coding)

取码长 $l_i = \lceil \log_D \frac{1}{p_i} \rceil$。

- **满足 Kraft 不等式：** $\sum D^{-\lceil \log_D \frac{1}{p_i} \rceil} \leq \sum D^{-\log_D \frac{1}{p_i}} = \sum p_i = 1$。
- **长度界限：** 由于 $\log_D \frac{1}{p_i} \leq l_i < \log_D \frac{1}{p_i} + 1$，对 p_i 加权求和可得：

$$H_D(X) \leq L_{\text{shannon}} < H_D(X) + 1$$

香农码是一个很好的码，能保证 $L_{\text{shannon}} < H_D(X) + 1$，但它往往不是最优的。设 L^* 为最优码的期望长度，由期望长度下界可知其必须有 $H_D(X) \leq L^*$。而香农码已满足 $L_{\text{shannon}} < H_D(X) + 1$，$L^*$ 必然不大于 L_{shannon}，故有

Theorem 4.4. 设 L^* 为最优码的期望长度，则：

$$H_D(X) \leq L^* < H_D(X) + 1 \tag{4.13}$$

这说明我们可以找到一种编码，其平均长度与熵的差距在 1 bit (或 1 个 D 元符号) 以内。

4.4.2 块编码与熵率

为了进一步降低每个符号的平均编码长度，我们可以对长度为 n 的符号序列进行联合编码。

我们在上一小节看见了对于单个随机变量，存在一种编码（香农码）使得其期望长度 L 满足 $H(X) \leq L < H(X) + 1$。

现在我们考虑发送长度为 n 的符号序列 $(X_1, X_2, \ldots, X_n)$（为简单起见，本小节假设 $D = 2$）。我们将这个序列视为一个来自于乘积空间 $\mathcal{X}^n$ 的单一超符号。

定义 L_n 为每输入符号的期望码长。

$$L_n = \frac{1}{n}\mathbb{E}l(X_1, \ldots, X_n) \tag{4.14}$$

这里的 $l(X_1, \ldots, X_n)$ 指的是针对该序列构造的香农码（即取码长 $l = \lceil -\log p(x^n) \rceil$）。因为如果香农码能满足某个上界，那么最优码一定也能满足。我们将之前针对香农码推导出的界限应用到这个超符号序列上：

$$H(X_1, \ldots, X_n) \leq \mathbb{E}l(X_1, \ldots, X_n) < H(X_1, \ldots, X_n) + 1 \tag{4.15}$$

情况 1：独立同分布 (i.i.d.)

如果 $X_1, \ldots, X_n$ 是 i.i.d. 的，则联合熵等于边缘熵之和：$H(X_1, \ldots, X_n) = \sum H(X_i) = nH(X)$。

将式 (4.15) 两边同时除以 n，我们得到每符号平均长度的界：

$$H(X) \le L_n < H(X) + \frac{1}{n} \tag{4.16}$$

这意味着，虽然单个符号的编码可能有 1 bit 的冗余，但这 1 bit 的冗余被分摊到了 n 个符号上。通过使用大的块长，我们可以实现每符号期望码长任意接近熵。

情况 2：一般随机过程

对于不一定是 i.i.d. 的随机过程，我们使用相同的论证。依然利用香农码的存在性，我们有界限：

$$H(X_1, \dots, X_n) \le \mathbb{E}l(X_1, \dots, X_n) < H(X_1, \dots, X_n) + 1 \tag{4.17}$$

再次除以 n 并定义 L_n 为每符号期望描述长度，我们得到：

$$\frac{H(X_1, \dots, X_n)}{n} \le L_n < \frac{H(X_1, \dots, X_n)}{n} + \frac{1}{n} \tag{4.18}$$

如果该随机过程是平稳的，那么根据熵率的定义，$H(X_1, \dots, X_n)/n \to H(\mathcal{X})$。当 $n \to \infty$ 时，期望描述长度趋于熵率。

上述推导是基于香农码的。由于最优码（L_n^*）的长度必然小于等于香农码，因此最优码也必然满足上述不等式。我们得到以下定理：

Theorem 4.5. 每符号的**最小**期望码长 L_n^* 满足：

$$\frac{H(X_1, \dots, X_n)}{n} \le L_n^* < \frac{H(X_1, \dots, X_n)}{n} + \frac{1}{n} \tag{4.19}$$

此外，如果 $X_1, X_2, \dots, X_n$ 是平稳随机过程，

$$L_n^* \to H(\mathcal{X}) \tag{4.20}$$

其中 $H(\mathcal{X})$ 是过程的熵率。

4.4.3　错误分布的代价

如果我们根据错误的分布 $q(x)$ 设计编码（即取 $l(x) \approx \log \frac{1}{q(x)}$），而真实分布是 $p(x)$，会发生什么？

Theorem 4.6 (Wrong Code)**.** 在真实分布 $p(x)$ 下，基于 $q(x)$ 设计的码长 $l(x) = \lceil \log \frac{1}{q(x)} \rceil$ 的期望长度满足：

$$H(p) + D(p||q) \le \mathbb{E}_p l(X) < H(p) + D(p||q) + 1 \tag{4.21}$$

Proof. 期望码长为：

$$\mathbb{E}_p l(X) = \sum_x p(x) \left\lceil \log \frac{1}{q(x)} \right\rceil \tag{4.22}$$

1. 证明上界：利用不等式 $\lceil \alpha \rceil < \alpha + 1$：

$$\mathbb{E}_p l(X) < \sum_x p(x) \left(\log \frac{1}{q(x)} + 1 \right) \tag{4.23}$$

$$= \sum_x p(x) \log \frac{1}{q(x)} + \sum_x p(x) \tag{4.24}$$

$$= \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \cdot \frac{1}{p(x)} \right) + 1 \quad (\text{引入 } p(x) \text{ 项}) \tag{4.25}$$

$$= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1 \tag{4.26}$$

$$= D(p||q) + H(p) + 1 \tag{4.27}$$

2. 证明下界：利用不等式 $\lceil \alpha \rceil \geq \alpha$：

$$\mathbb{E}_p l(X) \geq \sum_x p(x) \log \frac{1}{q(x)} \tag{4.28}$$

$$= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} \tag{4.29}$$

$$= D(p||q) + H(p) \tag{4.30}$$

□

Remark 4.2. 这表明，如果我们错误地认为分布是 $q(x)$ 而真实分布是 $p(x)$，我们付出的平均额外代价大约是 $D(p||q)$ 比特。这赋予了相对熵 $D(p||q)$ 一个非常具体的物理含义：**由信息不准确导致的描述复杂度的增加量**。

4.5 唯一可译码的 Kraft 不等式

唯一可译码比前缀码类更广，我们能通过使用唯一可译码获得更短的平均长度吗？答案是**不能**。

Theorem 4.7 (McMillan)**.** 任何 D 元唯一可译码的码长都满足 Kraft 不等式：

$$\sum D^{-l_i} \leq 1 \tag{4.31}$$

Proof. 考虑码的 k 次扩展 C^k。定义 $A = \sum_{x \in \mathcal{X}} D^{-l(x)}$。考察 A^k：

$$A^k = \left(\sum_x D^{-l(x)} \right)^k = \sum_{x_1 \ldots x_k} D^{-(l(x_1)+\cdots+l(x_k))} = \sum_{m=1}^{kl_{\max}} a(m) D^{-m} \tag{4.32}$$

其中 $a(m)$ 是总长度为 m 的码字序列的个数。由于是唯一可译码，长度为 m 的不同码字序列必须对应不同的 D 元字符串，而长度为 m 的 D 元串只有 D^m 个。因此 $a(m) \leq D^m$。代入得：

$$A^k \leq \sum_{m=1}^{kl_{\max}} D^m D^{-m} = kl_{\max} \tag{4.33}$$

即 $A \leq (kl_{\max})^{1/k}$。由于该式对任意 k 都成立，当 $k \to \infty$ 时，$(kl_{\max})^{1/k} \to 1$。所以 $A = \sum D^{-l_i} \leq 1$。 □

唯一可译码的码长约束与即时码完全相同。因此，最优的即时码和最优的唯一可译码具有相同的期望长度。既然即时码解码更方便，我们在数据压缩中通常只关注即时码。

4.6 霍夫曼编码 (Huffman Codes)

霍夫曼编码（Huffman Codes）是一种构建最优（即平均码长最短）前缀码的简单算法。其核心思想是贪心算法：总是合并概率最小的两个符号。

4.6.1 算法描述与示例

对于 D 进制编码，每次合并操作会将 D 个节点合并为 1 个父节点，使得总节点数减少 $D-1$ 个。

为了保证算法在最后一步正好能将剩余的 D 个节点合并为 1 个根节点（概率为 1），初始的符号总数 m 必须满足特定的模运算条件。如果不满足，我们需要添加概率为 0 的哑符号 (Dummy Symbols)。

算法步骤：

1. **添加哑符号（如果需要）：**设信源符号总数为 m，编码进制为 D。每次合并减少 $D-1$ 个节点。为了最终剩下一个节点，初始节点总数必须满足：

$$m \equiv 1 \pmod{D-1} \tag{4.34}$$

   如果 $m \not\equiv 1 \pmod{D-1}$，我们需要添加 k 个概率为 0 的哑符号，使得 $m+k \equiv 1 \pmod{D-1}$。
2. **排序：**将所有符号（包括哑符号）按概率从小到大排序。
3. **合并：**取出概率最小的 D 个符号，将它们合并为一个新符号，新符号的概率为这 D 个子节点概率之和。
4. **重复：**将新符号插入列表中重新排序，重复步骤 3，直到只剩下一个符号（根节点）。
5. **分配码字：**在码树中，从根节点出发，给通往 D 个子节点的分支分别赋值 $0, 1, \ldots, D-1$。

Example 4.1 (二进制霍夫曼编码). 考虑随机变量 X 取值于 $\mathcal{X} = \{1,2,3,4,5\}$，概率分别为 $0.25, 0.25, 0.2, 0.15, 0.15$。

构建过程解析：

1. 初始集合：$\{1(0.25), 2(0.25), 3(0.2), 4(0.15), 5(0.15)\}$。
2. 合并最小的 4 和 5 $\rightarrow$ 新节点 N_1 (0.30)。
3. 集合变为：$\{1(0.25), 2(0.25), 3(0.2), N_1(0.3)\}$。
4. 再次合并最小的 3(0.2) 和 2(0.25) $\rightarrow$ 新节点 N_2 (0.45)。*(注：合并顺序可能因排序策略略有不同，但不影响平均长度)*
5. 集合变为：$\{1(0.25), N_1(0.3), N_2(0.45)\}$。
6. 合并 1(0.25) 和 $N_1(0.3)$ $\rightarrow$ 新节点 N_3 (0.55)。
7. 最后合并 $N_2(0.45)$ 和 $N_3(0.55)$ $\rightarrow$ 根节点 (1.0)。

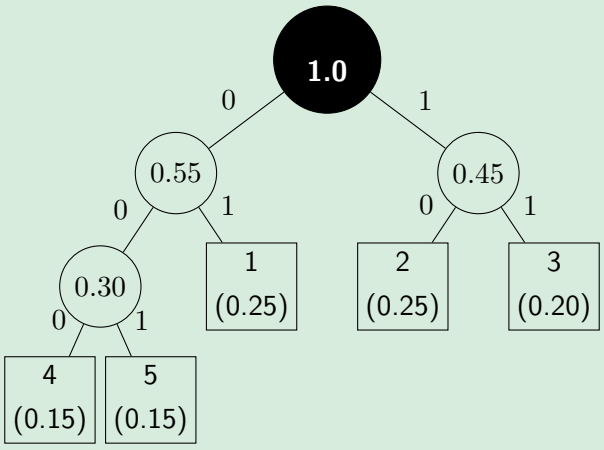

Figure 1: 霍夫曼编码树示例

平均码长计算：

$$L = 0.15 \times 3 + 0.15 \times 3 + 0.25 \times 2 + 0.25 \times 2 + 0.2 \times 2 = 2.3 \text{ bits}$$

Example 4.2 (三进制霍夫曼编码)**.** 假设随机变量 X 有 6 个取值，概率分布为：

$$\mathbf{p} = (0.25, 0.20, 0.15, 0.15, 0.15, 0.10)$$

我们要构建一个三进制 $(0, 1, 2)$ 的最优前缀码。

步骤 1：检查是否需要哑符号

- 符号总数 $m = 6$，进制 $D = 3$。
- 我们需要添加 $k = 1$ 个哑符号，使得总数变为 7。$7 \pmod 2 = 1$，满足条件。

步骤 2 & 3：构建与合并现在的概率列表（排序后）：

$$\{0.10, 0.15, 0.15, 0.15, 0.20, 0.25, \mathbf{0.00}(\text{Dummy})\}$$

注：哑符号概率为 *0*，通常排在最后或最前均可，为了贪心策略，它必须参与第一次合并。

1. **第一次合并**：取最小的三个 $\{0.00, 0.10, 0.15\}$。
   - 新节点概率：$0.00 + 0.10 + 0.15 = 0.25$。
   - 剩余列表：$\{0.15, 0.15, 0.20, 0.25, 0.25(\text{New})\}$。
2. **第二次合并**：取最小的三个 $\{0.15, 0.15, 0.20\}$。
   - 新节点概率：$0.15 + 0.15 + 0.20 = 0.50$。
   - 剩余列表：$\{0.25, 0.25(\text{From Step 1}), 0.50(\text{New})\}$。
3. **第三次合并**：取剩余的三个 $\{0.25, 0.25, 0.50\}$。
   - 新节点概率：1.0。完成。

对应的三进制霍夫曼树：

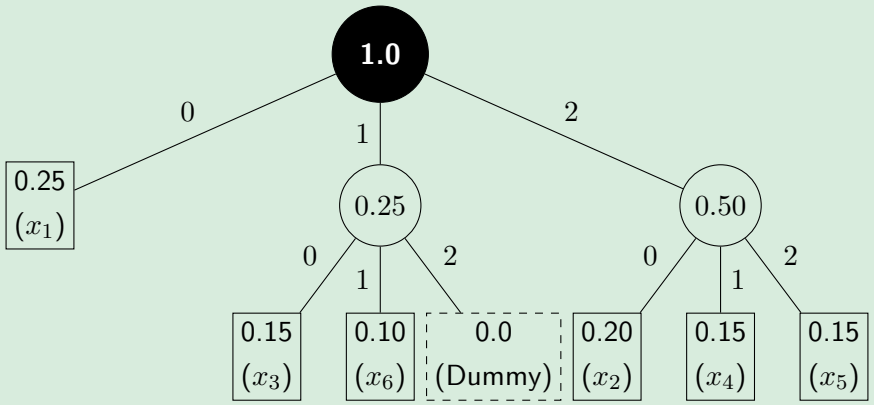

Figure 2: 三进制霍夫曼树

平均码长计算：

$$L = 0.25(1) + 0.20(2) + 0.15(2) + 0.15(2) + 0.15(2) + 0.10(2) = 1.85 \text{ ternary digits}$$

4.7 霍夫曼编码的最优性 (Optimality)

我们通过归纳法证明二进制霍夫曼编码是最优的。需要注意的是，最优码并不是唯一的（例如，反转所有比特，或者交换两个长度相同的码字，依然是最优码）。霍夫曼过程构建的是其中一种特定的最优码。

为了证明其最优性，我们首先证明任何最优码都必须满足的一些结构性质。

Lemma 4.1 (最优码的正则性)**.** 对于任意概率分布，存在一个最优即时码（平均长度最小）满足以下性质：

1. $p_j > p_k \Rightarrow l_j \leq l_k$（大概率对应短码）。
2. 两个最长码字长度相同。
3. 两个最长码字仅在最后一位不同，且对应概率最小的两个符号（即它们在树上是兄弟节点）。

证明. 这个证明通过“交换、修剪、重排”的策略，展示如果一个码不满足这些性质，我们可以修改它使它变得更好（或至少一样好），从而构造出一个满足性质的最优码（称为**正则码**）。

1. 假设 C_m 是一个最优码，且存在两个符号 j 和 k，使得 $p_j > p_k$ 但 $l_j > l_k$。我们构建一个新码 C'_m，交换这两个符号的码字。平均码长的变化量为：

$$\begin{aligned} L(C'_m) - L(C_m) &= \sum p_i l'_i - \sum p_i l_i && (4.35)\\ &= (p_j l_k + p_k l_j) - (p_j l_j + p_k l_k) && (4.36)\\ &= p_j(l_k - l_j) + p_k(l_j - l_k) && (4.37)\\ &= (p_j - p_k)(l_k - l_j) && (4.38) \end{aligned}$$

由于 $p_j - p_k > 0$ 且 $l_k - l_j < 0$，故 $(p_j - p_k)(l_k - l_j) < 0$。这意味着 $L(C'_m) < L(C_m)$，即新码比“最优码”还要短，这与 C_m 的最优性矛盾。因此，必须有 $l_j \leq l_k$。

2. 如果两个最长的码字长度不同，设最长的码字长度为 l_{max}。如果不止一个码字是 l_{max}，则性质成立。如果只有一个码字长度为 l_{max}，意味着它在树的最深层是没有兄弟节点。我们可以直接删掉这个码字的最后一位，它依然是前缀码（因为没有其他码字以此为前缀）。这样操作后，平均码长减小了，这与最优性矛盾。因此，最长码字必须成对出现（即有兄弟），故长度必然相同。

3. 并不是所有最优码都自然满足“最长码字对应概率最小的两个符号”。但是，我们可以对最优码进行调整：

- 由性质 1，最长码字必须属于低概率符号。
- 由性质 2，最长码字必须有兄弟。
- 如果最长码字对应的不是概率**最小**的那两个，我们可以交换树叶的位置（把最小概率的两个符号放到最深层的两个兄弟节点上）。
- 这种交换不会增加平均码长（甚至可能减少，如果原来的分配违反了性质 1）。

因此，总存在一个满足上述所有条件的最优码，我们称之为**正则码 (Canonical Code)**。 □

我们现在证明霍夫曼算法构造的码是最优的。定义：

- $\mathbf{p} = (p_1, \dots, p_m)$：原始概率分布，且 $p_1 \geq \cdots \geq p_m$。
- $\mathbf{p}' = (p_1, \dots, p_{m-2}, p_{m-1} + p_m)$：霍夫曼缩减后的分布（合并了最小的两个）。

Theorem 4.8 (霍夫曼编码最优性)**.** 霍夫曼编码是最优的；即如果 C^* 是霍夫曼码，而 C' 是任何其他唯一可译码，则 $L(C^*) \leq L(C')$。

证明. 我们使用关于字母表大小 m 的归纳法。

1. 基础情况 ($m = 2$)：对于两个符号，霍夫曼编码分配码字 0 和 1，平均长度为 1 bit。显然这是最优的（熵的下界也是 1，如果概率均等）。

2. 归纳假设：假设对于大小为 $m - 1$ 的任何分布，霍夫曼编码都是最优的。

3. 归纳步骤：我们需要建立 $L(\mathbf{p})$ 和 $L(\mathbf{p}')$ 之间的数值关系。

A. 从缩减分布扩展

令 $C^*_{m-1}(\mathbf{p}')$ 为 $\mathbf{p}'$ 的最优码（根据归纳假设，即霍夫曼码）。我们将其扩展为 $\mathbf{p}$ 的码 $C_m(\mathbf{p})$：

- 保持前 $m - 2$ 个符号的码字不变。
- 将对应概率为 $p_{m-1} + p_m$ 的码字 w 扩展为 $w0$（给 p_{m-1}）和 $w1$（给 p_m）。

新码的平均长度为：

$$L(C_m) = \sum_{i=1}^{m-2} p_i l_i + p_{m-1}(l'_{m-1} + 1) + p_m(l'_{m-1} + 1) \tag{4.39}$$

$$= \underbrace{\left(\sum_{i=1}^{m-2} p_i l_i + (p_{m-1} + p_m) l'_{m-1} \right)}_{L(C^*_{m-1})} + p_{m-1} + p_m \tag{4.40}$$

$$L(C_m) = L(C^*_{m-1}) + p_{m-1} + p_m \tag{4.41}$$

B. 从原始分布压缩

假设 $C_m^*(\mathbf{p})$ 是 $\mathbf{p}$ 的某个最优码。根据lemma 4.1，我们可以假设它是正则的（即 p_{m-1} 和 p_m 是兄弟）。我们将这两个兄弟合并为父节点，赋概率 $p_{m-1}+p_m$，得到 $\mathbf{p}'$ 的一个码 C_{m-1}。这个压缩码的平均长度为：

$$L(C_{m-1}) = L(C_m^*) - p_{m-1} - p_m \tag{4.42}$$

C. 矛盾法证明最优性

将 (4.41) 和 (4.42) 联立。由 (4.41) 我们知道霍夫曼生成的码长 $L_{\text{Huff}} = L(C_{m-1}^*) + q$，其中 $q = p_{m-1} + p_m$。

我们想证明 L_{Huff} 就是 $L(C_m^*)$。

根据归纳假设，C_{m-1}^* 是 $m-1$ 个符号的最优码，因此对于任何其他码 C_{m-1}（包括由 C_m^* 压缩而来的那个），都有：

$$L(C_{m-1}^*) \le L(C_{m-1}) \tag{4.43}$$

将 (4.42) 代入 (4.43)：

$$L(C_{m-1}^*) \le L(C_m^*) - (p_{m-1}+p_m) \tag{4.44}$$

$$L(C_{m-1}^*) + (p_{m-1}+p_m) \le L(C_m^*) \tag{4.45}$$

注意左边正是我们在步骤 A 中构建的扩展码（霍夫曼码）的长度 $L(C_m)$。所以：

$$L(C_m^{\text{Huffman}}) \le L(C_m^*) \tag{4.46}$$

另一方面，由于 C_m^* 定义为最优码，显然有 $L(C_m^*) \le L(C_m^{\text{Huffman}})$。

因此，必须有 $L(C_m^{\text{Huffman}}) = L(C_m^*)$。

这意味着：如果我们从 $m-1$ 的最优码开始扩展，我们得到的 m 的码必然也是最优的。归纳证毕。 □

4.8　香农-法诺-伊莱亚斯编码 (SFE Code)

我们知道，只要满足 Kraft 不等式，就存在唯一可译码。香农提出的码长 $l(x) = \lceil \log \frac{1}{p(x)} \rceil$ 满足 Kraft 不等式，因此肯定存在对应的码。本节介绍一种利用累积分布函数 (CDF) 的构造性方法来生成这样的码。虽然它在平均长度上不如霍夫曼码（SFE 需要额外多约 1 bit），但其构造过程不需要构建树，且分析起来具有独特的几何直观。

4.8.1　算法定义

不失一般性，设 $X = \{1, 2, \dots, m\}$。

Definition 4.7 (修正累积分布函数). 定义累积分布函数 $F(x) = \sum_{a \le x} p(a)$。定义**修正累积分布函数 (Modified CDF)** $\bar{F}(x)$ 为阶梯函数中 x 对应的中点：

$$\bar{F}(x) = \sum_{a<x} p(a) + \frac{1}{2}p(x) = F(x-1) + \frac{1}{2}p(x) \tag{4.47}$$

其中 $F(0) = 0$。

编码步骤：

1. 计算每个符号 x 的 $\bar{F}(x)$。这是一个 $(0,1)$ 之间的实数。

2. 将 $\bar{F}(x)$ 展开为二进制小数 $0.z_1z_2z_3\ldots$。
3. 截取该二进制小数的前 $l(x)$ 位作为码字，其中码长定义为：

$$l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil + 1 \tag{4.48}$$

我们将截断后的二进制小数记为 $\lfloor \bar{F}(x) \rfloor_{l(x)}$。

4.8.2 前缀性质的证明

为什么这样简单的截断能保证是前缀码？

Proof. 每个码字 $z_1 \ldots z_{l(x)}$ 实际上代表了区间 $[0.z_1 \ldots z_{l(x)}, 0.z_1 \ldots z_{l(x)} + 2^{-l(x)})$。如果所有符号对应的区间互不重叠，那么这就构成了一个前缀码。

考查区间长度与阶梯高度的关系：

$$2^{-l(x)} = 2^{-\lceil \log \frac{1}{p(x)} \rceil - 1} \leq 2^{-(\log \frac{1}{p(x)}) - 1} = \frac{p(x)}{2}$$

即：**码字区间的长度小于该符号概率 $p(x)$ 的一半**。

再看区间的位置。码字的值 $\lfloor \bar{F}(x) \rfloor_{l(x)}$ 是 $\bar{F}(x)$ 向下取整到 $l(x)$ 位的结果，因此：

$$\bar{F}(x) - 2^{-l(x)} < \lfloor \bar{F}(x) \rfloor_{l(x)} \leq \bar{F}(x)$$

这意味着码字区间的下界在 $\bar{F}(x)$ 下方，但距离不超过 $p(x)/2$。码字区间的上界是 $\lfloor \bar{F}(x) \rfloor_{l(x)} + 2^{-l(x)}$，由于下界 $\leq \bar{F}(x)$ 且长度 $\leq p(x)/2$，所以上界 $\leq \bar{F}(x) + p(x)/2$。

综合起来，码字对应的区间完全落在开区间 $(\bar{F}(x) - \frac{p(x)}{2}, \bar{F}(x) + \frac{p(x)}{2})$ 内。

根据 $\bar{F}(x)$ 的定义，这个开区间正好是 CDF 函数图像中对应符号 x 的那个“台阶”的内部（台阶底是 $F(x-1)$，顶是 $F(x)$，中点是 $\bar{F}(x)$，高度是 $p(x)$）。由于不同符号的台阶是互不相交的，因此这些码字区间也互不相交。证毕。 □

4.8.3 平均码长界

$$L = \sum p(x)l(x) = \sum p(x) \left(\left\lceil \log \frac{1}{p(x)} \right\rceil + 1 \right) \tag{4.49}$$

$$< \sum p(x) \left(\log \frac{1}{p(x)} + 2 \right) \tag{4.50}$$

$$= H(X) + 2 \tag{4.51}$$

相比于霍夫曼码的 $H(X)+1$ 上界，SFE 码稍差一点，但它提供了一个无需构建树的闭式解。

4.9 香农码的竞争最优性

我们知道霍夫曼码在**期望**意义下是最优的。但是，对于某一次具体的传输，香农码是否可能比其他码差很多？答案是：几乎不可能。香农码具有极强的鲁棒性。

考虑一个博弈：给定分布，两方分别给出编码。随机抽取一个符号，码短者赢。

Theorem 4.9. 设 $l(x) = \lceil \log \frac{1}{p(x)} \rceil$ 为香农码的长度，设 $l'(x)$ 为**任何**其他唯一可译码的长度。则香农码比对手长 c 位以上的概率呈指数级衰减：

$$\Pr(l(X) \geq l'(X) + c) \leq \frac{1}{2^{c-1}} \tag{4.52}$$

例如，香农码比另一个码长 5 位以上的概率小于 $1/16$。

Proof. 我们需要计算满足 $l(x) \geq l'(x) + c$ 的所有 x 的概率之和。

首先展开 $l(x)$ 的定义：

$$\left\lceil \log \frac{1}{p(x)} \right\rceil \geq l'(x) + c$$

由于 $\lceil y \rceil < y + 1$，上述不等式成立意味着：

$$\log \frac{1}{p(x)} + 1 > l'(x) + c \implies \log \frac{1}{p(x)} > l'(x) + c - 1$$

两边取指数：

$$\frac{1}{p(x)} > 2^{l'(x)+c-1} \implies p(x) < 2^{-l'(x)-c+1}$$

现在我们计算概率：

$$\begin{aligned}
\Pr(l(X) \geq l'(X) + c) &= \sum_{x: l(x) \geq l'(x)+c} p(x) && (4.53)\\
&< \sum_{x: l(x) \geq l'(x)+c} 2^{-l'(x)-c+1} && (4.54)\\
&= 2^{-(c-1)} \sum_{x: l(x) \geq l'(x)+c} 2^{-l'(x)} && (4.55)\\
&\leq 2^{-(c-1)} \sum_{\text{all } x} 2^{-l'(x)} && (4.56)
\end{aligned}$$

注意步骤 (4.56) 中，我们将求和范围扩大到了所有 x。因为 $l'(x)$ 对应一个唯一可译码，根据 Kraft 不等式，$\sum 2^{-l'(x)} \leq 1$。因此：

$$\Pr(\cdot) \leq 2^{-(c-1)} \cdot 1 = \frac{1}{2^{c-1}}$$

证毕。 □

4.10 从公平硬币生成离散分布

本章的前半部分讨论如何将随机变量压缩为比特流（去除冗余）。本节讨论对偶问题：如何利用公平的比特流（抛硬币，Bernoulli(1/2)）来生成具有特定分布 $\mathbf{p}$ 的随机变量 X。

我们将生成算法表示为一棵二叉树，树的叶子节点被标记为输出符号 X。路径上的分支对应硬币的正面 (0) 或反面 (1)。

Example 4.3 (生成简单分布). 假设我们要生成分布 $X = \{a(1/2), b(1/4), c(1/4)\}$。这是一个二元分布（Dyadic），我们可以构建一棵有限的完全二叉树。算法如下：

- 抛第一枚硬币。如果是 0，输出 a。

- 如果是 1，抛第二枚硬币。
- 第二枚是 0 输出 b，是 1 输出 c。

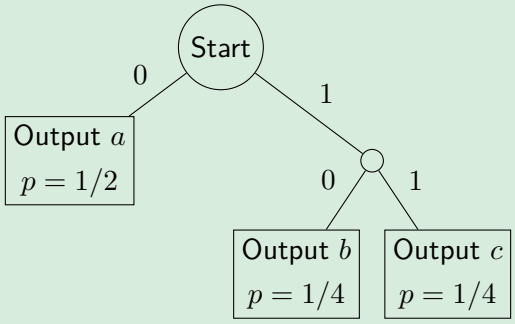

Figure 3: 生成分布 $(1/2, 1/4, 1/4)$ 的算法树。期望抛硬币次数 $\mathbb{E}T = 1\cdot\frac{1}{2}+2\cdot\frac{1}{4}+2\cdot\frac{1}{4} = 1.5$ bits，恰好等于熵 $H(X)$。

4.10.1 引理与下界

Lemma 4.2 (树的深度与熵)**.** 对于任意一棵完全二叉树，如果叶子节点 y 的概率分布为 $2^{-k(y)}$（$k(y)$ 为叶子 y 的深度），则该树的期望深度 $\mathbb{E}T$ 等于该叶子分布的熵 $H(Y)$。

Proof.

$$\mathbb{E}T = \sum_y p(y)k(y) = \sum_y p(y)\log\frac{1}{2^{-k(y)}} = \sum p(y)\log\frac{1}{p(y)} = H(Y)$$

□

Theorem 4.10 (生成分布的下界)**.** 对于任何生成随机变量 X 的算法，其消耗的平均公平比特数 $\mathbb{E}T$ 必须大于等于 X 的熵：

$$\mathbb{E}T \geq H(X) \tag{4.57}$$

Proof. 生成算法可以看作一棵树，叶子节点集合为 $\mathcal{Y}$。每个叶子 $y \in \mathcal{Y}$ 有固定的概率 $2^{-k(y)}$。算法的输出 X 是叶子 Y 的函数（即 $X = f(Y)$，可能有多个叶子映射到同一个输出符号 x）。

由 lemma 4.2，期望抛掷次数 $\mathbb{E}T = H(Y)$。由

$$H(Y) = H(Y) + \underbrace{H(X|Y)}_{=0} = H(X,Y) = H(X) + H(Y|X) \geq H(X)$$

我们有：

$$H(X) \leq H(Y)$$

因此 $H(X) \leq \mathbb{E}T$。证毕。 □

Theorem 4.11 (二元分布的最优性)**.** 如果 X 是二元分布（dyadic distribution，即所有 $p(x)$ 都是 2^{-k} 形式），则存在算法使得 $\mathbb{E}T = H(X)$。（证明：直接构建霍夫曼树即可，霍夫曼树的叶子概率正好匹配二元分布）。

4.10.2　一般分布的生成算法

如果 $\mathbf{p}$ 不是二元分布（例如 $2/3, 1/3$），我们无法用有限的二叉树完美生成它。但是，我们可以通过**二进制展开**的方法，用无限树逼近。

构造方法：将每个概率 p_i 写成二进制小数形式：

$$p_i = \sum_{j \geq 1} p_i^{(j)} 2^{-j}, \quad p_i^{(j)} \in \{0, 1\}$$

这相当于把概率 p_i 拆分成许多概率为 2^{-j} 的“原子”。我们将这些原子挂在一棵完全二叉树的对应深度上。由于 $\sum p_i = 1$，所有原子的概率和为 1，根据扩展 Kraft 不等式，我们可以完美地将它们安排在树的叶子上。

Theorem 4.12 (生成分布的上界)**.** 利用上述二进制展开算法生成 X，所需的期望抛掷次数满足：

$$H(X) \leq \mathbb{E}T < H(X) + 2 \tag{4.58}$$

Proof. 我们已经知道下界 $\mathbb{E}T \geq H(X)$。现在详细证明上界。

设 T 为生成算法抛掷硬币的总次数。我们可以将 $\mathbb{E}T$ 写成每个符号 i 对期望深度的贡献之和：

$$\mathbb{E}T = \sum_{i=1}^{m} T_i \tag{4.59}$$

其中 T_i 是符号 i 的所有原子在树中的深度加权和。根据二进制展开 $p_i = \sum_{j \geq 1} p_i^{(j)} 2^{-j}$，我们有：

$$T_i = \sum_{j \geq 1} p_i^{(j)} \cdot j \cdot 2^{-j} \tag{4.60}$$

(如果 p_i 的二进制第 j 位是 1，则对应一个深度为 j 的叶子，其概率贡献是 2^{-j}，深度贡献是 $j \cdot 2^{-j}$)。

我们的目标是证明 $\mathbb{E}T < H(X) + 2$。这等价于证明对于每个符号 i，其贡献 T_i 满足某个界。具体来说，我们要证明：

$$T_i < -p_i \log p_i + 2p_i \tag{4.61}$$

一旦证明了 (4.61)，对所有 i 求和即可得：

$$\mathbb{E}T = \sum_i T_i < \sum_i (-p_i \log p_i) + 2 \sum_i p_i = H(X) + 2$$

证明不等式 (4.61)

对于任意概率 p_i，我们可以找到一个整数 n，使得：

$$2^{-n} \leq p_i < 2^{-(n-1)} \implies n - 1 < -\log p_i \leq n \tag{4.62}$$

这也意味着 p_i 的二进制展开中，前 $n-1$ 位必须都是 0。即：当 $j < n$ 时，$p_i^{(j)} = 0$。

我们考察差值 $\Delta = T_i - (-p_i \log p_i + 2p_i)$。为了利用 n 的整数性质，我们放缩对数项：由 (4.62) 可知 $-\log p_i > n - 1$，所以 $-p_i \log p_i > p_i(n-1)$。我们要证明上界，所以考察：

$$T_i + p_i \log p_i - 2p_i$$

利用 $\log p_i < -(n-1)$，我们有：

$$T_i + p_i \log p_i - 2p_i < T_i - p_i(n-1) - 2p_i \tag{4.63}$$

$$= T_i - p_i(n+1) \tag{4.64}$$

现在代入 T_i 和 p_i 的级数表达式（注意求和从 $j=n$ 开始）：

$$T_i - (n+1)p_i = \sum_{j=n}^{\infty} p_i^{(j)} j 2^{-j} - (n+1)\sum_{j=n}^{\infty} p_i^{(j)} 2^{-j} \tag{4.65}$$

$$= \sum_{j=n}^{\infty} p_i^{(j)} (j-n-1) 2^{-j} \tag{4.66}$$

我们将求和项拆开分析：

- 当 $j=n$ 时：项为 $p_i^{(n)}(n-n-1)2^{-n} = -p_i^{(n)}2^{-n}$。
- 当 $j=n+1$ 时：项为 $p_i^{(n+1)}(n+1-n-1)2^{-(n+1)} = 0$。
- 当 $j \geq n+2$ 时：项为正数。

所以：

$$T_i - (n+1)p_i = -p_i^{(n)}2^{-n} + \sum_{j=n+2}^{\infty} p_i^{(j)}(j-n-1)2^{-j} \tag{4.67}$$

为了使上式最大（求最坏情况上界），我们假设 p_i 的所有位 $p_i^{(j)}$ 都取最大值 1（但这必须受限于 p_i 的范围，不过这里我们直接放缩系数）。首先，因为 $p_i \geq 2^{-n}$，所以二进制第 n 位必须是 1，即 $p_i^{(n)} = 1$。对于 $j \geq n+2$ 的项，我们也放缩为 $p_i^{(j)} \leq 1$。

于是：

$$T_i - (n+1)p_i \leq -2^{-n} + \sum_{j=n+2}^{\infty} 1 \cdot (j-n-1)2^{-j} \tag{4.68}$$

$$\text{令} k = j-(n+1) \implies j = k+n+1 \quad \text{求和变为} \tag{4.69}$$

$$= -2^{-n} + \sum_{k=1}^{\infty} k 2^{-(k+n+1)} \tag{4.70}$$

$$= -2^{-n} + 2^{-(n+1)} \underbrace{\sum_{k=1}^{\infty} k(1/2)^k}_{\text{已知级数和为 } 2} \tag{4.71}$$

$$= -2^{-n} + 2^{-(n+1)} \cdot 2 \tag{4.72}$$

$$= -2^{-n} + 2^{-n} \tag{4.73}$$

$$= 0 \tag{4.74}$$

因此，我们证明了 $T_i + p_i \log p_i - 2p_i < 0$，即 $T_i < -p_i \log p_i + 2p_i$。

对所有 i 求和：

$$\mathbb{E}T = \sum_i T_i < \sum_i (-p_i \log p_i) + 2\sum_i p_i = H(X) + 2$$

证毕。 □

Example 4.4 (生成非二元分布：$2/3, 1/3$)**.** 分布 $\mathbf{p} = (2/3, 1/3)$ 不是二元的（dyadic），无法用有限树完美生成。二进制展开：

$$2/3 = 0.101010\ldots_2 = 1/2 + 1/8 + 1/32 + \ldots$$

$$1/3 = 0.010101\ldots_2 = 1/4 + 1/16 + 1/64 + \ldots$$

我们将概率视为“原子”挂在树上。$2/3$ 占据深度为 1, 3, 5... 的节点；$1/3$ 占据深度为 2, 4, 6... 的节点。

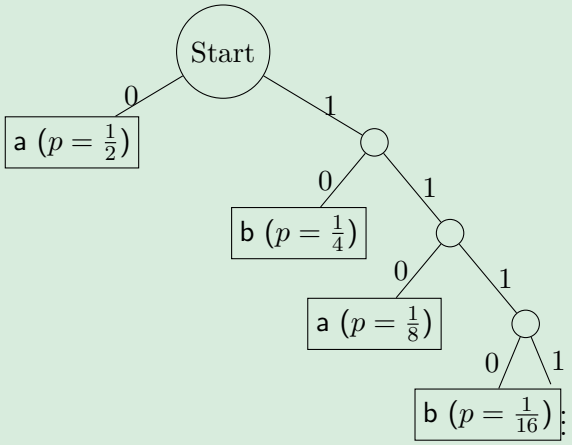

Figure 4: 生成分布 (2/3, 1/3) 的无限树示意图。左侧分支被分配给原子概率。

在这个算法中，如果抛出序列 0，输出 a（概率 1/2）；序列 10，输出 b（概率 1/4）；序列 110，输出 a（概率 1/8），以此类推。a 的总概率 $= 1/2 + 1/8 + 1/32 + \cdots = 2/3$。$b$ 的总概率 $= 1/4 + 1/16 + 1/64 + \cdots = 1/3$。