信息论与编码

马啸
maxiao@mail.sysu.edu.cn

计算机学院
中山大学

2021 年春季学期

1. 离散无记忆信源编码定理
2. 典型序列
3. 编码定理证明
   - 定长到定长
   - 定长到变长
4. 编码定理的逆定理证明

离散无记忆信源

Definition 1 (离散无记忆信源)

设信源$\mathbf{X} = (X_1, X_2, ..., X_n, ...)$, 满足：

1. **无记忆的：** $P_{X^n}(x^n) = \prod_{1 \le t \le n} P_{X_t}(x_t)$ 对于任意 $n > 1$
2. **平稳的：** $P_{X_t}(x) \equiv P_{X_1}(x) \triangleq P_X(x)$ 对于任意 $t > 1$

则称该信源为离散平稳无记忆信源，也称作独立同分布信源。

熵

Definition 2 (熵)

离散随机变量X，概率质量函数为 $P_X(x)$， $x \in \mathcal{X}$，则X 的熵，记作 $H(X)$，

$$H(X) = \mathrm{E}(\log \frac{1}{P_X(X)}) = -\sum_{x\in\mathcal{X}} P_X(x) \log P_X(x). \tag{1}$$

对于有K 个时间的离散集X，若干个时间的概率为$P_1, P_2, \cdots, P_K$，则集X 的熵可写成：

$$H_K(P_1, P_2, \cdots, P_K) = -\sum_{k=1}^{K} P_k \log P_k$$

其中，$\sum_{k=1}^{K} P_k = 1, P_k \geqslant 0$。

信源编码基本框架

一般地，一个时间离散的信源可以表示为一个随机变量序列：$X_1, X_2, ..., X_n, ...$，其中X_t 取值在$\mathcal{X}$ 上，其统计规律可以用一族联合分布律$\{P_{\mathbf{X}}(\mathbf{x})\}, n = 1, 2, ...$ 来表征。设$\mathcal{D} = \{0, 1, ..., D-1\}$ 是字符集，我们用$\mathcal{D}^*$ 表示由$\mathcal{D}$ 构成的字符串的全体，包括空字符串，即$\mathcal{D}^* = \bigcup_{\ell \geq 0} \mathcal{D}^\ell$。信源编码的一般框架可以描述为：

$$\text{编码 } \phi_n : \mathcal{X}^n \longmapsto \mathcal{D}^*$$

$$\text{译码 } \psi_n : \mathcal{D}^* \longmapsto \mathcal{X}^n$$

$$\text{码率 } R_n = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} P_{\mathbf{X}}(\mathbf{x}) \ell\left(\phi_n(\mathbf{x})\right)$$

$$\text{译码错误 } \epsilon_n = \Pr\left\{\psi_n\left(\phi_n(\mathbf{X})\right) \neq \mathbf{X}\right\}$$

信源编码基本框架

R_n 中的$\ell\left(\phi_n(\mathbf{x})\right)$ 表示码字$\phi_n(\mathbf{x})$ 的长度。由此，我们知道码率表示在统计意义下每个信源符号所用的码字的平均长度。离散信源编码的问题就是通过证明ϕ_n 与ψ_n 的存在性，寻找满足$\lim_{n\to\infty}\epsilon_n = 0$ 的码率R_n 的下极限。

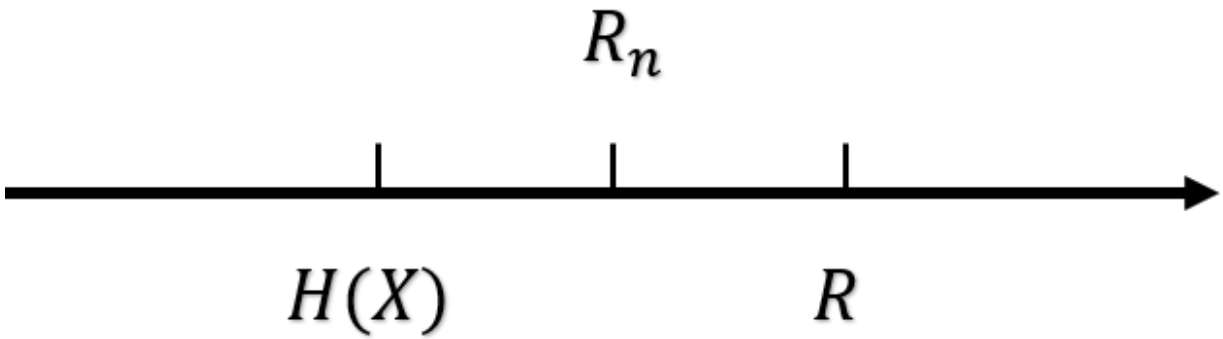
此外，根据序列或码字的长度是否固定，编译码大致可以分为四种类型：

1. 定长 $\longmapsto$ 定长
2. 定长 $\longmapsto$ 变长
3. 变长 $\longmapsto$ 定长
4. 变长 $\longmapsto$ 变长

信源编码定理

Theorem 3 (离散信源无失真编码定理)

给定一个离散无记忆信源，即一个独立同分布（*IID*）的随机变量序列$X_1, X_2, \cdots$，其熵为 $H(X)$。设码率$R > H(X)$，则存在固定码长编码(ϕ_n, ψ_n)，使得 R_n 满足$R_n \leq R$，并且$\lim_{n\to\infty} \epsilon_n = 0$。若允许变码长编码，则可以使得 $\epsilon_n = 0$。



信源编码定理的逆定理

Theorem 4 (离散信源无失真编码逆定理)

设 $R < H(X)$。则对于任何定长编码，若其码率 $R_n \leq R < H(X)$，则必有 $\lim_{n\to\infty} \epsilon_n = 1$。

典型序列

典型序列

Lemma 5 (弱大数定理)

若$X_1, X_2, \cdots, X_n, \cdots$ 是独立同分布的随机变量序列，一维分布是$P_X(x)$，则$X_1, X_2, \cdots, X_n, \cdots$ 的样本熵$-\frac{1}{n}\log P_{X^n}(X^n)$ 依概率收敛于$H(X)$，记作$\lim_{n\to\infty} -\frac{1}{n}\log P_{X^n}(X^n) \overset{P}{=} H(X)$，即，$\forall \epsilon > 0$，我们有

$$\lim_{n\to\infty} \Pr\left\{\left|-\frac{1}{n}\log P_{X^n}(X^n) - H(X)\right| \geq \epsilon\right\} = 0 \tag{2}$$

典型序列

Definition 6 (典型集)

有一离散无记忆信源$X_1, X_2, \cdots, X_n, \cdots$，服从$P_X(x)$，$x \in \mathcal{X}$。对于该信源中的序列$(x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$，如果它满足

$$2^{-n(H(X)+\epsilon)} \leq P_{X^n}(x^n) \leq 2^{-n(H(X)-\epsilon)}$$

则称它是$\epsilon-$典型 的。所有$\epsilon-$典型 序列的全体称为**典型集**，记作$A_\epsilon^{(n)}$

典型序列

例子1.
投掷一枚不均匀硬币，其正面朝上的概率为3/4，反面朝上的概率为1/4。对应有随机变量$X \in \{0,1\}$，其概率质量函数为$P(1) = 3/4$ 和$P(0) = 1/4$。现在独立投掷该硬币n 次，设$n = 5$，则当ϵ等于0.1 时，哪些序列是$\epsilon -$ 典型 的？$\epsilon = 0.01$ 呢？

投掷硬币5 次，则有32(2^5) 个序列：$(0,0,0,0,0)$、$(0,0,0,0,1)$、$\cdots$ 。从典型序列的定义可以看出，给定n，判定一个序列是不是典型的，与设定的ϵ 有关。在本例中，$n = 5$，当$\epsilon = 0.1$ 时，则一个有4 次正面朝上的序列是典型的，而当$\epsilon = 0.01$ 时，则没有序列是典型的。

典型序列

固定n，若ϵ 取值很大，则所有序列都是典型的；若ϵ 取值很小，则典型序列的数量（典型集$A_\epsilon^{(n)}$ 的元素个数）会减少。

信息论关心的是另一个问题，任意给定一个$\epsilon > 0$ （无论多小），n 充分大时，哪些序列是典型的？如果硬币是均匀的（正面朝上的概率$P(1)$与反面朝上的概率$P(0)$ 均等于$\frac{1}{2}$），按照我们目前关于典型序列的定义，则所有序列都是典型的。

渐近均分性

Theorem 7 (AEP)

(1) 若$x^n \in A_\epsilon^{(n)}$, 则$H(X) - \epsilon \leq -\frac{1}{n} \log P_{X^n}(x^n) \leq H(X) + \epsilon$。

典型序列的样本熵具有上下界。

(2) 当n 充分大时，$\Pr\{A_\epsilon^{(n)}\} \geq 1 - \epsilon$。

典型集的概率可以接近于 1 （大数定律）

(3) $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$，其中$|A|$ 表示集合A 中元素的个数。

要表示典型序列，至多需要多少二进制数位？

(4) 当n 充分大时，$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$。

若要表示典型序列，至少需要多少二进制数位？

第 3 点与第 4 点给出了典型序列的个数，也即典型集的大小的上下界。

编码定理证明

编码定理证明

离散信源无失真编码定理:

给定一个离散无记忆信源，即一个独立同分布（IID）的随机变量序列$X_1, X_2, \cdots$，其熵为 $H(X)$。设码率$R > H(X)$，则存在固定码长编码(ϕ_n, ψ_n)，使得 R_n 满足$R_n \leq R$，并且$\lim_{n\to\infty} \epsilon_n = 0$。若允许变码长编码，则可以使得 $\epsilon_n = 0$。

我们可以使用三种方法来证明离散信源无失真编码定理：典型序列、随机装箱和型方法。这节课主要介绍其中的典型序列方法。

定长到定长：编码

1. 将$\mathcal{X}^n$中的所有序列划分成两个集合：典型集$A_\epsilon^{(n)}$及其补集$\overline{A_\epsilon^{(n)}} = \mathcal{X}^n - A_\epsilon^{(n)}$。
2. 将所有的典型序列按照某种顺序（比如字典序）编号，并把这些编号转化为二进制，用作对应序列的编码，同时将非典型序列全部映射为全零序列。由于典型序列的个数$|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$，所以典型序列可以用不超过$\lceil n(H(X)+\epsilon) \rceil$比特表示，其中$\lceil x \rceil$表示“向上取整”，即，不小于$x$的最小整数。
3. 这样就得到了$\mathcal{X}^n$的所有序列的一个编码方案，描述序列需要的总长度不超过$\lceil n(H(X)+\epsilon) \rceil$比特。

译码

由于编码长度固定，并且典型序列的编码是一一映射的，所以可以直接根据码字找到对应的典型序列，从而实现对典型序列的译码。

码率

$$R_n = \frac{1}{n}\sum_{x^n} P_{X^n}(x^n)\ell(\phi_n(x^n)) = \frac{1}{n}\lceil n(H(X)+\epsilon)\rceil$$

因为编码方式为定长到定长，所以其码率可以看作是固定的。

译码错误概率

在这个编码方案中，我们只将 $\mathcal{X}^n$ 上的典型序列的编号的二进制进行一对一映射，因而在译码时可以对应恢复，不会发生错误。但非典型序列没有一对一映射的编码，因此，译码错误概率 ϵ_n 不为 0，但小于等于 ϵ。

定长到变长：编码

1. 将$\mathcal{X}^n$中的所有序列划分成两个集合：典型集$A_\epsilon^{(n)}$及其补集$\overline{A_\epsilon^{(n)}} = \mathcal{X}^n - A_\epsilon^{(n)}$。
2. 将所有的典型序列按照某种顺序（比如字典序）编号，并把这些编号转化为二进制，用作对应序列的编码。由于典型序列的个数$|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$，所以典型序列可以用不超过$\lceil n(H(X)+\epsilon) \rceil$比特表示，其中$\lceil x \rceil$表示“向上取整”，即，不小于$x$的最小整数。类似地，由于非典型序列的个数显然不超过$|\mathcal{X}|^n$，非典型序列可以用不超过$\lceil n \log |\mathcal{X}| \rceil$比特表示。
3. 分别给典型集$A_\epsilon^{(n)}$中的所有序列前面加上0，非典型集$\overline{A_\epsilon^{(n)}}$的序列前面加上1。这样就得到了$\mathcal{X}^n$的所有序列的一个编码方案，其中，描述典型序列需要的总长度不超过$\lceil n(H(X)+\epsilon)+1 \rceil$比特，非典型序列的编码总长度不超过$\lceil n \log |\mathcal{X}| + 1 \rceil$比特。

译码

起始位 0 或 1，作为标识位，识别是典型序列或非典型序列，标明紧随码字的长度。由于编码是一对一映射，那么根据序列的下标可以找到对应的序列，从而实现译码。

码率

$$
\begin{aligned}
R_n &= \frac{1}{n}\sum_{x^n} P_{X^n}(x^n)\ell(\phi_n(x^n)) \\
&= \frac{1}{n}\left[\sum_{x^n\in A_\epsilon^{(n)}} P_{X^n}(x^n)\ell(\phi_n(x^n)) + \sum_{x^n\in \overline{A_\epsilon^{(n)}}} P_{X^n}(x^n)\ell(\phi_n(x^n))\right] \\
&\le \frac{1}{n}\left[\sum_{x^n\in A_\epsilon^{(n)}} P_{X^n}(x^n)(n(H(X)+\epsilon)+2) + \sum_{x^n\in \overline{A_\epsilon^{(n)}}} P_{X^n}(x^n)(n\log|\mathcal{X}|+2)\right] \\
&= \frac{1}{n}\left[\Pr\{A_\epsilon^{(n)}\}(n(H(X)+\epsilon)+2) + \Pr\{\overline{A_\epsilon^{(n)}}\}(n\log|\mathcal{X}|+2)\right] \\
&\le \frac{1}{n}\left[n(H(X)+\epsilon)+\epsilon n\log|\mathcal{X}|+2\right] = H(X)+\epsilon',
\end{aligned}
$$

码率

$$R_n \leq \frac{1}{n}\left[n(H(X)+\epsilon)+\epsilon n \log|\mathcal{X}|+2\right] = H(X)+\epsilon'$$

由定理条件知，$R > H(X)$，所以我们可以取充分小的正数 ϵ，使得对所有充分大的 n，$\epsilon' \triangleq \epsilon + \epsilon \log|\mathcal{X}| + \frac{2}{n} < \frac{R-H(X)}{2}$。这样，我们得到上述编码的码率$R_n \leq H(X)+\epsilon' \leq R$。

译码错误概率

在这个编码方案中，我们将$\mathcal{X}^n$ 上的所有序列的编号的二进制作为编码，进行一对一映射，因而在译码时可以对应恢复，不会发生错误。即，译码错误概率 ϵ_n 为 0。

例子

我们以不均匀硬币为例对上述证明进行说明。在例子中，$n = 5$，熵 $H(X) \approx 0.8113$ bit/ 符号，共有32 个序列。当$\epsilon = 0.1$ 时，这些序列可以分为两类：典型的与非典型的，分别按照字典序排序及编号如表。

序列类型	序列	十进制序号	二进制序号	码字
典型序列	(0,1,1,1,1)	0	000	0000
	(1,0,1,1,1)	1	001	0001
	(1,1,0,1,1)	2	010	0010
	(1,1,1,0,1)	3	011	0011
	(1,1,1,1,0)	4	100	0100
非典型序列	(0,0,0,0,0)	0	00000	100000
	(0,0,0,0,1)	1	00001	100001
	(0,0,0,1,0)	2	00010	100010
	(0,0,0,1,1)	3	00011	100011
	(0,0,1,0,0)	4	00100	100100
	(0,0,1,0,1)	5	00101	100101
	(0,0,1,1,0)	6	00110	100110
	(0,0,1,1,1)	7	00111	100111
	(0,1,0,0,0)	8	01000	101000
	(0,1,0,0,1)	9	01001	101001
	(0,1,0,1,0)	10	01010	101010
	(0,1,0,1,1)	11	01011	101011
	(0,1,1,0,0)	12	01100	101100
	(0,1,1,0,1)	13	01101	101101
	(0,1,1,1,0)	14	01110	101110
	(1,0,0,0,0)	15	01111	101111
	(1,0,0,0,1)	16	10000	110000
	(1,0,0,1,0)	17	10001	110001
	(1,0,0,1,1)	18	10010	110010
	(1,0,1,0,0)	19	10011	110011
	(1,0,1,0,1)	20	10100	110100
	(1,0,1,1,0)	21	10101	110101
	(1,1,0,0,0)	22	10110	110110
	(1,1,0,0,1)	23	10111	110111
	(1,1,0,1,0)	24	11000	111000
	(1,1,1,0,0)	25	11001	111001
	(1,1,1,1,1)	26	11010	111010

(1) **编码:** 分别把典型序列和非典型序列的编号转化为二进制，在典型序列的编号前加上 0，非典型序列加上 1，作为对应序列的码字。典型序列的码长 $\ell(x^n) = 4 < \lceil n(H(X) + \epsilon) + 1 \rceil = 6$，非典型序列的码长 $\ell(x^n) = 6 \leq \lceil n \log |\mathcal{X}| + 1 \rceil = 6$。

(2) **译码:** 根据编码输出的起始位识别该序列是典型序列还是非典型序列，若起始位为 0，则该序列为典型序列，紧随码长为 4；若起始位为 1，则该序列为非典型序列，紧随码长为 6。根据码字可以找到对应序列实现译码。更具体地，假设信源输出（即编码器的输入）是 $(1,1,1,1,0)(1,0,1,1,1)(0,1,0,0,1)(1,1,1,1,1)$，则编码输出为 01000001101001111010，而这个码字序列可以唯一译码得到信源输出。如此类推，可以将编码输出 01000001101001111010 译码为典型序列 4 号、典型序列 1 号、非典型序列 9 号、非典型序列 26 号，对应找到信源输出序列。

注：上述例子说明了码表构造与译码的过程，但可以验证，这并不是一个有效的压缩编码，因为码字序列可能比直接表示观察结果还要长。实际上，要达到压缩效果，ϵ 要充分小，码长要充分大。此时，用查表方法是不切实际的。

编码定理的逆定理证明

定长到定长

定长编码的速率$R_n = L_n/n \leq R = \log\left(2^{nR}\right) < H(X)$

译码正确的概率为：

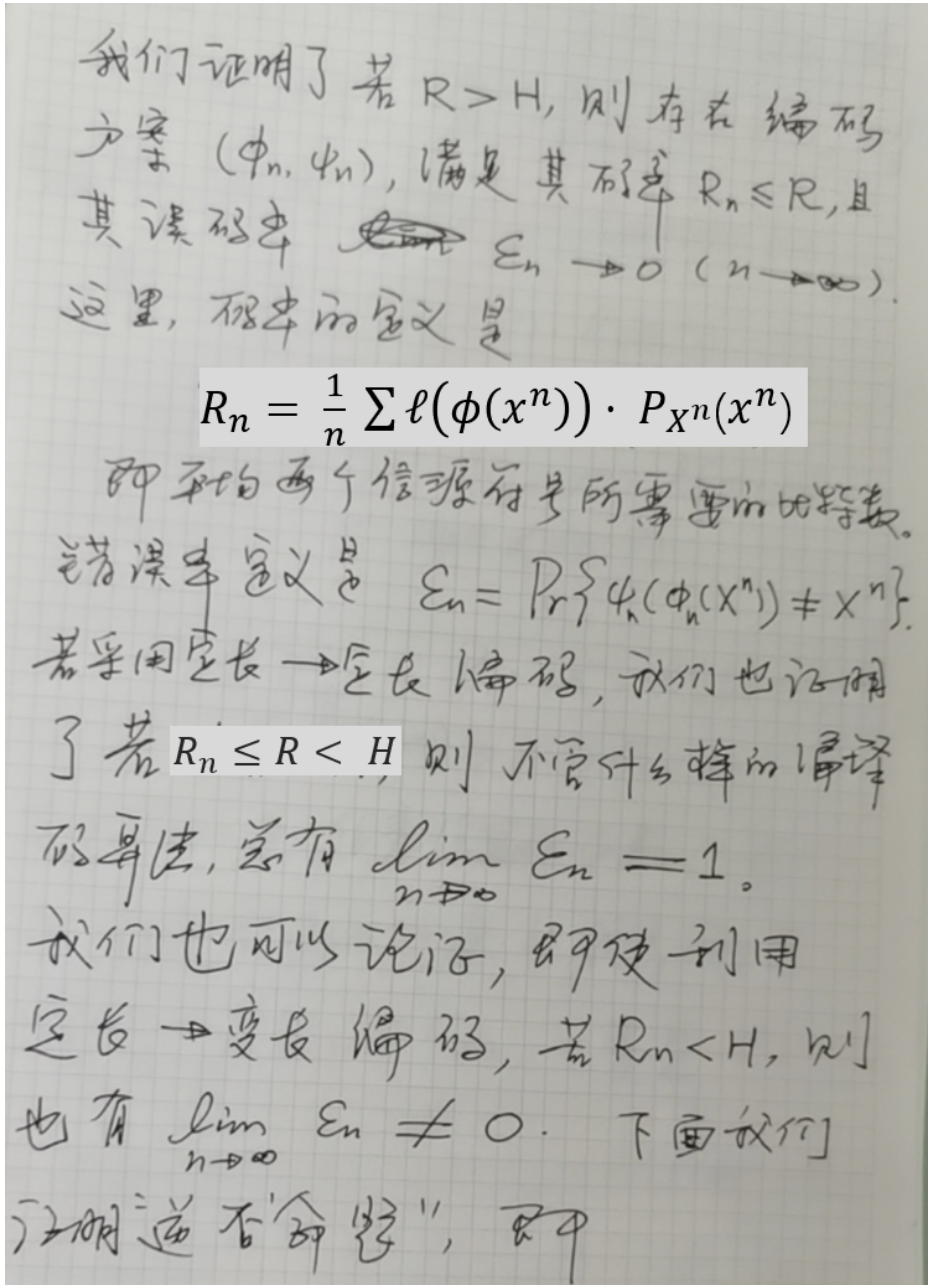
$$P\left\{\psi\left(\phi\left(X^n\right)\right) = X^n\right\} = \sum_{x^n:\psi(\phi(x^n))=x^n} P_{X^n}\left(x^n\right)$$

$$\leq \sum_{x^n\in A_\varepsilon^{(n)}:\psi(\phi(x^n))=x^n} P_{X^n}\left(x^n\right) + \sum_{x^n\notin\overline{A_\epsilon^{(n)}}} P_{X^n}\left(x^n\right)$$

码本大小/码字个数至多为 2^{nR}，至多保证 2^{nR} 个典型序列编码正确。
根据AEP，典型序列中每个 x^n 的概率至多为$2^{-n(H-\epsilon)}$，所以，
当 $n \to \infty$ 时，

$$P\left\{\psi\left(\phi\left(X^n\right)\right) = X^n\right\} \leqslant 2^{nR}2^{-n(H-\epsilon)} + \epsilon \to 0$$

所以，译码错误概率$\epsilon_n = 1 - P\left\{\psi\left(\phi\left(X^n\right)\right) = X^n\right\} \to 1$。


我们证明了若 $R>H$，则存在编码方案 (ϕ_n, ψ_n)，满足其码率 $R_n \le R$，且其误码率 $\varepsilon_n \to 0$ $(n\to\infty)$.

这里，码率的定义是

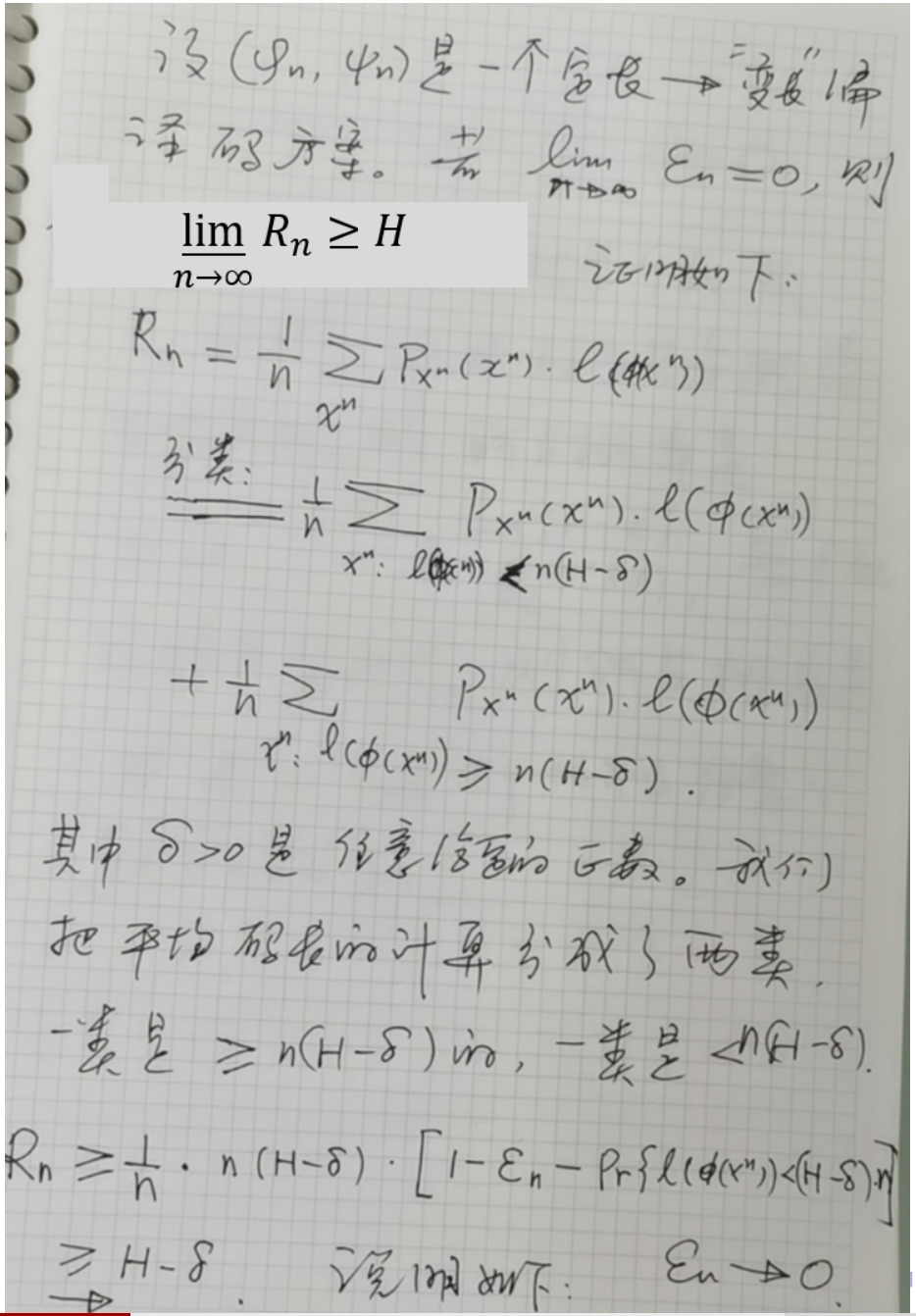
$$R_n = \frac{1}{n}\sum \ell\big(\phi(x^n)\big)\cdot P_{X^n}(x^n)$$

即平均每个信源符号所需要的比特数。

错误率定义是 $\varepsilon_n = \Pr\{\psi_n(\phi_n(X^n)) \neq X^n\}$.

若采用定长→定长编码，我们也证明了若 $R_n \le R < H$，则不管什么样的编译码算法，总有 $\lim_{n\to\infty} \varepsilon_n = 1$。

我们也可以说，即使利用定长→变长编码，若 $R_n<H$，则也有 $\lim_{n\to\infty} \varepsilon_n \neq 0$. 下面我们证明这个命题，即


设(φ_n, ψ_n)是一个定长→"变长"编
译码方案。若 $\lim_{n\to\infty} \varepsilon_n = 0$，则
$$\underline{\lim_{n\to\infty}} R_n \geq H$$
证明如下：
$R_n = \frac{1}{n} \sum_{x^n} P_{X^n}(x^n) \cdot l(\varphi(x^n))$
分类：
$\frac{1}{n} \sum_{x^n: l(\varphi(x^n)) < n(H-\delta)} P_{X^n}(x^n) \cdot l(\varphi(x^n))$
$+ \frac{1}{n} \sum_{x^n: l(\varphi(x^n)) \geq n(H-\delta)} P_{X^n}(x^n) \cdot l(\varphi(x^n))$.
其中$\delta > 0$是任意给定的正数。我们
把平均码长的计算分成了两类，
一类是$\geq n(H-\delta)$的，一类是$< n(H-\delta)$.
$R_n \geq \frac{1}{n} \cdot n(H-\delta) \cdot [1 - \varepsilon_n - \Pr\{l(\varphi(x^n)) < (H-\delta)n\}]$
$\geq H-\delta$. 说明如下：$\varepsilon_n \to 0$.

$\Pr\{l(\phi(x^n)) < n(H-\delta)\} \not\to 0.$

而长 $< n(H-\delta)$ 的码字个数共

$2^{n(H-\delta)}$. 这些码字若与所给

典型序列, 则所占概率是

$\leq 2^{n(H-\delta)} \cdot 2^{-n(H-\epsilon)} = 2^{-n(\delta-\epsilon)}$, 若分配

给非典型序列, 则所占概率是

$\leq \varepsilon$. 这两个都可以任意小,

(取 $\delta = 2\epsilon$).

作业

Exercise 1.

一张百元人民币（第五套）约重 1.15 克。由于流通过程中多种因素影响，我们可以假定一张百元人民币重量在 1.10 ~ 1.20 克之间。办案人员从某贪官（后已处死刑）住处查出大量百元现金，称其重量在 3 ~ 3.05 吨之间。问这些贪污的钱数额在什么范围？

作业

Exercise 2. [Cover(2006)]

Entropy of functions of a random variable. Let X be a discrete random variable. Show that the entropy of a function of X is less than or equal to the entropy of X by justifying the following steps:

$$\begin{aligned} H(X, g(X)) &\overset{(a)}{=} H(X) + H(g(X) \mid X) \\ &\overset{(b)}{=} H(X), \\ H(X, g(X)) &\overset{(c)}{=} H(g(X)) + H(X \mid g(X)) \\ &\overset{(d)}{\geq} H(g(X)) \end{aligned} \tag{3}$$

Thus, $H(g(X)) \leq H(X)$.

作业

Exercise 3. [Cover(2006)]

Zero conditional entropy. Show that if $H(Y|X) = 0$, then Y is a function of X [i.e., for all x with $p(x) > 0$, there is only one possible value of y with $p(x, y) > 0$].

Exercise 4.

证明：1) $H(Y|X) \geq 0$；2) $H(XY) \geq H(X)$，等号成立当且仅当 Y 可以写作 X 的函数，即存在函数 g 使得 $Y = g(X)$。

谢谢！