L16 代数编码

错误检测与纠正 Error-Detecting and Correcting Codes

一般来说，我们此处只讨论每一位要么为 0，要么为 1 的二进制编码。

基本的编码过程是这样的：

1. 我们首先有一条 m 位的信息（message）
2. 我们通过编码器（encoder）将其编码为一个 n 位的字（word）
3. 接下来发送者（transmitter）将其发送给接收者（receiver），然而这其中可能有噪声（noise）干扰。
4. 接收者得到了 n 位的字（received word），将其放入解码器（decoder）。
5. 解码器给出解码的结果：m 位信息（received message）或错误（error）。

我们称呼这样的情形为发生了一个错误：如果在码字（codeword）中改变了一位或多位。

我们称呼这样的一个理论为一个解码方案（decoding scheme）：对于任意的 n 位元组，要么将其解码为有意义的解码后信息，要么说明其存在错误。

Example 8.2 重复三次编码：对于信息 $(x_1, \cdots, x_n)$，其编码结果为 $3n$ 位的 $(x_1, \cdots, x_n, x_1, \cdots, x_n, x_1, \cdots, x_n)$，如果其存在一位或两位改变，那么其可以检测出来（三个重复位不一致）；如果存在一位改变，还可以纠正（三个重复位取更多的那一位）。

Example 8.3 奇偶校验：对于 n 位信息 $x_1, \cdots, x_n$，我们在最前面额外添加一个 x_0，满足 $x_0, \cdots, x_n$ 中 1 的个数为偶数。如果存在一位的改变，那么我们可以检测出来（1 个数的奇偶性被破坏），但无法纠正；多余一位的改变无法检测。此处 x_0 称为奇偶校验位或奇偶检验位（parity check bit）.

最大似然解码（maximum-likelihood decoding）：我们假设如果我们收到了一条被噪声改变的信息，那么我们应当假设噪声应为最小可能的那个噪声，也就是改变前信息应取所有可能的改变前信息和改变后信息中，差距最小的那条。依照这个假设在对其进行纠正、解码。这一原则便称为最大似然解码原则。

一条**二进制对称信道**（binary symmetric channel）是这样一个模型：其由一个只能发送二进制信号的发送者（transmitter）和一个接受者构成。

Th. 8.7 如果在一条二进制对称信道中，一位信息有 p 的可能不翻转，$q = 1 - p$ 的概率翻转，那么一条 n 位二进制信息恰有 k 位改变的概率是：

$$\binom{n}{k} q^k p^{n-k}$$

码（codes）：在谈论中可能出现的特定二进制串的集合。

块码（block codes）：一般来说，我们对于长度为 km 的串，常常将其每 m 位编为一个单位，然后每个单位分别编码为 n 位、从 n 位解码，有这种特点的串称为 (n, m) 块码。

编码函数、解码函数：我们对于长度为 m 的二进制串，其到一个长度为 n 的二进制串的单射，用这种这样编码的函数称为编码函数，也就有：$E : \mathbb{Z}_2^m \to \mathbb{Z}_2^n$，类似地用于解码的单射函数称为解码函数：$D : \mathbb{Z}_2^n \to \mathbb{Z}_2^m$.

一个**码字**（codeword）是 E 的像中的任何一个元素。

汉明距离（Hamming distance）：两个串的 Hamming 距离为其最少需要修改多少位，使得从一个串变为另一个串，在二进制串的情况下，为二者异或的 1 的个数，记作 $d(\mathbf{x}, \mathbf{y})$.

最小距离（minimum distance）：任意两个码字的 Hamming 距离的最小值，记作 $d_{\min}$.

重量（weight）：串中 1 的个数，记作 $w(\mathbf{x})$.

Prop. 8.11：假设 $\mathbf{x}, \mathbf{y}, \mathbf{z}$ 都为 n 位二进制串，那么：

1. $w(\mathbf{x}) = d(\mathbf{x}, \mathbf{0})$.
2. $d(\mathbf{x}, \mathbf{y}) \geq 0$.
3. $d(\mathbf{x}, \mathbf{y}) = 0$ iff. $\mathbf{x} = \mathbf{y}$.
4. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$.
5. $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$.

Th. 8.13 设码字集合为 C，且最小距离 $d_{\min} = 2n + 1$，那么任何不超过 n 位错误都可以被纠正，任何 $2n$ 位错误都可被探测。

线性编码 Linear Codes

群码（group code）是 $\mathbb{Z}_2^n$ 的子群的码字集合。

Lem. 8.17 设 $\mathbf{x}$ 和 $\mathbf{y}$ 都是 n 位二进制串，那么 $w(\mathbf{x} + \mathbf{y}) = d(\mathbf{x}, \mathbf{y})$. 注意，此处的 $+$ 应当是二进制串的异或。

Lem. 8.18 设群码 C 的最小距离是 $d_{\min}$，那么 $d_{\min} = \min\{w(\mathbf{x}) | \mathbf{x} \neq \mathbf{0}\}$.

内积（inner product）：$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \cdots + x_n y_n$. 注意，此处的按位乘法应视作按位与。

零空间：设 $\mathbb{M}_{m \times n}(\mathbb{Z}_2)$ 为 $\mathbb{Z}_2$ 元素及运算构成的所有 $m \times n$ 矩阵的集合，对于 $H \in \mathbb{M}_{m \times n}(\mathbb{Z}_2)$，其零空间 $\mathrm{Null}(H)$ 为这样的 $\mathbf{x}$ 的集合：$H\mathbf{x} = \mathbf{0}$.

Th. 8.21 对于 $H \in \mathbb{M}_{m \times n}(\mathbb{Z}_2)$，其零空间 $\mathrm{Null}(H)$ 为一个群码。

如果一个码为某个 H 的零空间，那么称该码为一个线性编码。

奇偶校验矩阵和生成矩阵 Parity-Check and Generator Matrices

规范的奇偶校验矩阵（canonical parity-check matrix）：设 $H \in \mathbb{M}_{m \times n}(\mathbb{Z}_2)$，如果 $H = (A, I_m)$，那么 H 称为规范的奇偶校验矩阵。

标准的生成矩阵（standard generator matrix）：与规范的奇偶校验矩阵对应，$G = (I_{n-m}; A)$ 这个 $n \times (n - m)$ 矩阵称为其标准的生成矩阵。

这样定义是因为：如果 $G\mathbf{x} = \mathbf{y}$ 成立，当且仅当 $H\mathbf{y} = \mathbf{0}$ 成立。

Th. 8.25 如果 $H \in \mathbb{M}_{m \times n}(\mathbb{Z}_2)$ 是一个规范的奇偶校验矩阵，那么 $\mathrm{Null}(H)$ 就有所有这样的 $\mathrm{x} \in \mathbb{Z}_2^n$ 组成：前 $n - m$ 位任意，而后 m 位由前 $n - m$ 位和 $H\mathbf{x} = \mathbf{0}$ 决定。此时，我们这样理解：最后的 m 位实际上是对前 $n - m$ 位的奇偶校验，也就是说 H 给出了一个 $(n, n - m)$ 块码。进一步地，我们称 $\mathbf{x}$ 的前 $n - m$ 位为信息位（information bits），后 m 位称为校验位（check bits）。

Th. 8.26 设 G 位一个 $n \times k$ 的标准的生成矩阵，那么 $C = \{\mathbf{y} | G\mathbf{x} = \mathbf{y}, \mathbf{x} \in \mathbb{Z}_2^k\}$ 是一个 (n, k) 块码，进一步地，C 也是一个群码。

Lem. 8.27 设 $H = (A, I_m)$ 是一个 $m \times n$ 的规范的奇偶校验矩阵，并且 $G = (I_{n-m}; A)$ 是对应的 $n \times (n - m)$ 的标准的生成矩阵，那么我们有 $HG = O_{m \times (n-m)}$.

Th. 8.28 设 $H=(A,I_m)$ 是一个 $m\times n$ 的规范的奇偶校验矩阵，并且 $G=(I_{n-m};A)$ 是对应的 $n\times(n-m)$ 的标准的生成矩阵。设 C 为 G 生成的群码，那么 $\mathbf{y}\in C$ iff. $H\mathbf{y}=\mathbf{0}$，进一步地，C 是一个有规范的奇偶校验矩阵 H 的群码。

Prop. 8.30 设 $\mathbf{e}_i$ 为只有第 i 位为 1，其余为全 0 的长度为 n 的串，并假设 $H\in\mathbb{M}_{m\times n}(\mathbb{Z}_2)$，那么 $H\mathbf{e}_i$ 就是 H 的第 i 列。

Th. 8.31 设 $H\in\mathbb{M}_{m\times n}(\mathbb{Z}_2)$，那么 $\mathrm{Null}(H)$ 是一个一位错误可检测的码当且仅当 H 中不存在全 0 列。

Th. 8.34 设 $H\in\mathbb{M}_{m\times n}(\mathbb{Z}_2)$，那么 $\mathrm{Null}(H)$ 是一个一位错误可检测的码当且仅当 H 中不存在全 0 列或相等列。

因相关限制，对于 $n=2^m$，信息位最多有 $n-(1+m)$ 位，其中一列零列，m 列 e_i，从而最大可能的编码率为：$(n-1-m)/n=1-\dfrac{1+m}{2^m}$.（P141）

高效解码 Efficient Decoding

如果 $H\in\mathbb{M}_{m\times n}(\mathbb{Z}_2)$ 且 $\mathbf{x}\in\mathbb{Z}_2^n$，那么称 $H\mathbf{x}$ 为 $\mathbf{x}$ 的校验子（syndrome），以下方法可以帮助我们快速解码和纠正错误。

Prop. 8.36 设 $H\in\mathbb{M}_{m\times n}(\mathbb{Z}_2)$ 其决定了一个线性码，$\mathbf{x}$ 是一个接收到的 n 位串，并设 $\mathbf{x}=\mathbf{c}+\mathbf{e}$，其中 $\mathbf{c}$ 为发送信息，$\mathbf{e}$ 为噪声，那么 $\mathbf{x}$ 的校验子等于 $\mathbf{e}$ 的校验子。

Th. 8.37 设 $H\in\mathbb{M}_{m\times n}(\mathbb{Z}_2)$ 且其决定的线性是一位错误可纠正的。设 $\mathbf{r}$ 为一个接收到的 n 位串，其至多只有一位错误，那么：如果 $\mathbf{r}$ 的校验子为 $\mathbf{0}$，那么不存在错误；否则，查找 $\mathbf{r}$ 的校验子是 H 的那一列，假设是第 i 列，那么一位错误就在第 i 位。

陪集解码、标准解码（coset decoding, standard decoding）：其使用 C 的陪集来实现最大似然解码，设 C 是一个 (n,m) 线性码，那么一个 C 对于 $\mathbf{x}\in\mathbb{Z}_2^n$ 的陪集可以写作 $\mathbf{x}+C$，通过 Lagrange 定理（群论），C 总共有 2^{n-m} 个不同的陪集。

陪集代表（coset leader）：陪集中重量最小的 n 位串称为这个陪集的陪集代表。

解码表（decoding table）：因 Prop. 8.43，一个校验子对应一个陪集，将其列出一张表，称为解码表。

Prop. 8.43 设 C 是一个 (n,k) 线性码，其由 H 给定，并且假设 $\mathbf{x},\mathbf{y}\in\mathbb{Z}_2^n$，那么 $\mathbf{x}$ 和 $\mathbf{y}$ 有相同的陪集当且仅当 $H\mathbf{x}=H\mathbf{y}$，也就是说两个串的陪集相同当且仅当其校验子相同。

显然这里存在一个解码方法：对于存在错误的串，取其校验子对应的陪集的陪集代表作为错误，由此算出原发送信息。这一方法由最大似然解码原则给出。