第二章离散信源信息度量

第一、二、三节

陈兴同

中国矿业大学 数学学院

2021 年 8 月

内容提要

内容提要

通信系统的组成

本章开始将学习研究通信系统中各个环节的信息处理问题。一个实用的通信系统一般由信源、编码器、调制器、信道、译码器、信宿等基本元素构成，如图 2-1.

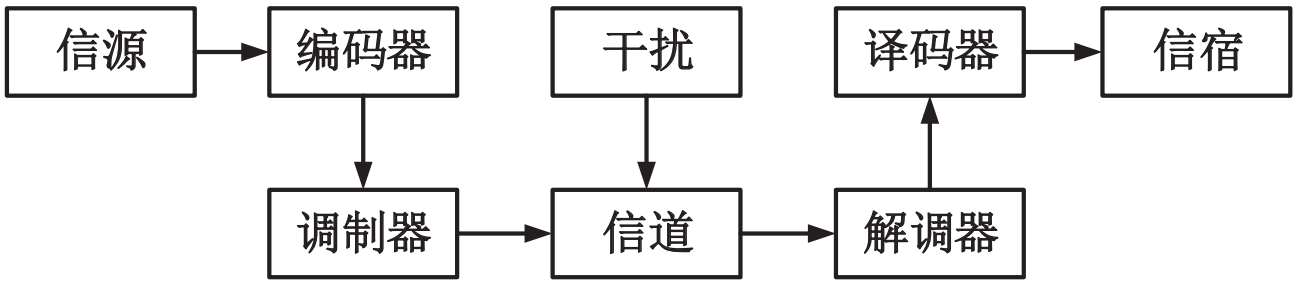

Figure: tu2-1

信源

信源是指产生信息的源泉，如一个被压缩的文件；一个电报通信系统发报机；一个电话通信系统通话的用户；信源发出的字符通常不止一个，每个时刻发出哪个字符也是随机的，因此每个字符发出与否都有一定的概率。信源所能发出的所有字符称为**信源字符集** 或**信源字符空间**，信源字符空间及其上的概率分布构成了信源的两个基本要素。比如：在中文通信中，信源字符空间可以是全体汉字、标点符号、数字、以及英文字母等构成，每个字符被用到的概率互不相同，它们就构成了信源基本要素。信源通常需要连续地发出字符，由信源在不同时刻发出的字符所形成的字符串就构成了**消息**，其中携带了通信双方所需要的信息。

信道

信道是指定传输信息的通道。它包括信道编码器、信道译码器、调制器、通信线路等。比如两台经过五类双绞线连接的计算机，网卡及连线就可以认为是一种信道；连接两地的电话线与电话机也可以认为是一种信道。信道中传输的**信号**是对信源发出的“消息”进行压缩编码后再进行传输编码及调制后所得到的，它作为消息的载体，从甲地传到乙地，然后再进行解调与译码，交给信息接收者即信宿。信号在传输过程中会受到各种干扰，因此输出信号与输入信号之间会有误差。

编码

编码在信息的保存与传输中尤其重要。一方面信源发出的消息包含大量的冗余，进行压缩编码后可以大大提高信息含量，提高传输效率；另一方面为了更好地利用信道特性进行传输，还要对压缩后的消息进传输编码即信道码，这样作可以大大提高信息传输的可靠性。编码方案包括编码与译码两部分，这部分工作由编码器与译码器去做。

信号

在数字通信中消息一般用方波信号来表示，信号的强度只取有限个值比如 0 与 1，每个信号都占有一定延时，其图像基本上都是方波形式，如图 2-2a 是一种典型的数字信号，它的表达式为：

$$f(t,k)=\begin{cases}1 & k\le t<k+0.5\\ -1 & k+0.5\le t<k+1\\ 0 & \text{其它}\end{cases},f(t)=\sum_{k=0}^{5}f(t,k),t\in(0,5)$$

高电压为 1，低电压为 0，用它可以表示二进制串 $1010101010\cdots$。表示消息的方波信号由于能量较小，通常不能直接传输，需要使用能量较大的载波信号进行调制转换成电信号传输，图 2-2b 是上述数字信号叠加幅度为 10 的正弦波信号 $y=10\sin(3\pi t)$ 后的简单调制图像。在输出端需要进行解调恢复原来的方波信号。调制与解调原理一般在通信工程中研究。

数字信号

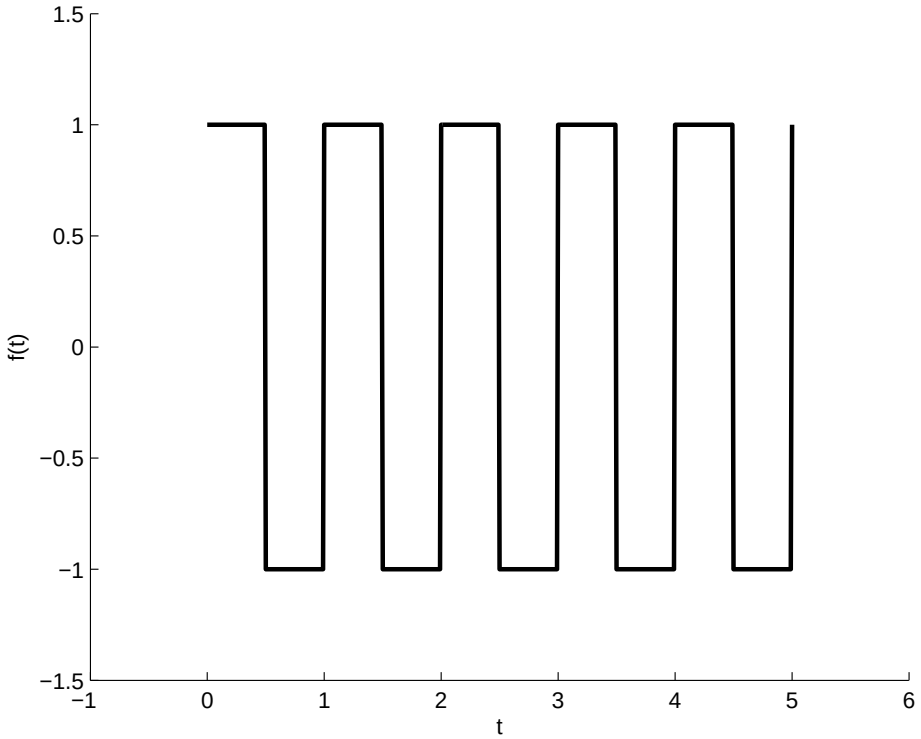

Figure: 数字信号

2.2a

模拟信号

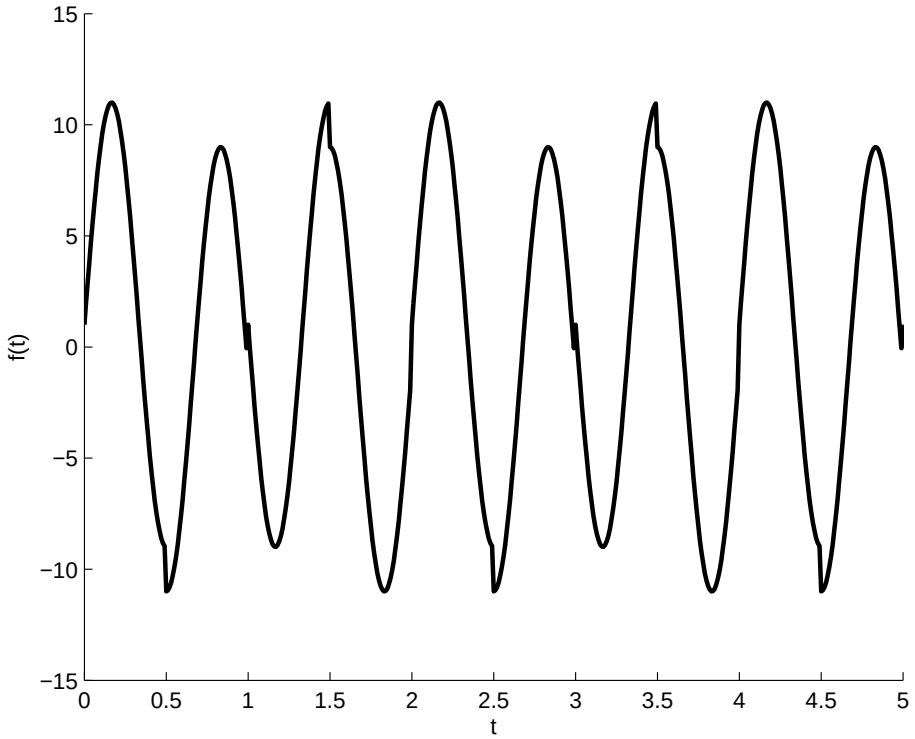

Figure: 模拟信号

2.2b

信源数学模型

信源字符空间及其上的概率分布通常用一个随机变量来表示，如果信源字符空间是离散符号集则用离散型随机变量来表示，这种信源称为离散信源；如果信源字符空间是一个连续符号集就用一个连续型随机变量来表示，这种信源称为连续型信源，比如电视信号或广播信号。本章只讨论有限离散信源。设信源字符空间用 $\mathcal{X} = \{a_1, a_2, \cdots, a_N\}$ 来表示，其上的概率分布为

$$X \sim p(x) = \begin{pmatrix} a_1 & a_2 & \cdots & a_N \\ p_1 & p_2 & \cdots & p_N \end{pmatrix}. \tag{2.1}$$

离散信源

对于离散信源来说，信源总是按一定的时间间隔不断地发出字符，形成时间域上无限长的符号序列。在时刻 $n = 1, 2, \cdots$ 发出什么字符是随机的，但都是字符空间 $\mathcal{X}$ 中的字符，用随机变量 X_n 来表示信源在第 n 时刻发出的字符，不同时刻的随机变量就形成时间域上随机变量序列，因此将离散信源的数学模型抽象成取值在符号空间 $\mathcal{X}$ 中的离散型随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$，其中的下标可以看成是不同的时刻。根据信源序列的不同统计特征，可以将离散信源划分成不同类型。

离散信源的熵率

信源的熵通常定义为信源发出字符序列中平均每个字符包含信息量。

设离散信源 $X_1, X_2, \cdots$，如果极限

$$\lim_{n\to\infty} \frac{H(X_1, X_2, \cdots, X_n)}{n}$$

*存在，则称它为该离散信源的**熵率**，记成 $H_\infty(X)$。*

熵率描述了离散信源的每个信源字符所包含的平均信息量，可以看成信源信息量大小的度量，也可以看成是离散随机序列的一个数字特征。

离散无记忆信源

如果离散随机序列 $X_1, X_2, \cdots, X_n, \cdots$ 中任意有限个随机变量相互独立, 并且每个随机变量有相同的分布（2-1），则称它是一个**离散无记忆信源**。在离散无记忆信源中，不同时刻信源发出的字符之间没有任何依赖关系。

离散无记忆信源联合分布

n 长输出序列 $X_1, X_2, \cdots, X_n$ 的联合分布函数是

$$F(x_1, x_2, \cdots, x_n) = P\{X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n\} = \prod_{i=1}^{n} F_{X_i}(x_i)$$

联合分布律为

$$p(x_1, x_2, \cdots, x_n) = P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} = \prod_{i=1}^{n} p(x_i),$$

这里 $x_1, x_2, \cdots, x_n$ 表示信源输出的一个 n 长消息，这种消息共有 N^n 种可能的取值。

离散无记忆信源熵率

离散无记忆信源的熵率容易计算。事实上：由于随机变量之间相互独立性，故联合熵：

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i) = nH(X_1),$$

$$H_\infty(X) = H(X_1) = H(X).$$

在以后的编码原理中主要学习这种信源的编码。

离散平稳信源

如果离散随机序列 $X_1, X_2, \cdots, X_n, \cdots$ 中任意有限个随机变量 $X_{i_1}, X_{i_2}, \cdots, X_{i_m}$ 与经过任意时间 h 后的新序列 $X_{i_1+h}, X_{i_2+h}, \cdots, X_{i_m+h}$ 有相同的联合分布函数即

$$F_{i_1 i_2 \cdots i_m}(x_1, x_2, \cdots, x_m) = F_{i_1+h i_2+h \cdots i_m+h}(x_1, x_2, \cdots, x_m),$$

则称它是一个**离散平稳信源**。
因为这种信源的任意有限维联合分布函数不随时间推移而变化，因此信源的统计特性具有时间不变性或平稳性。像联合分布、边缘分布、条件分布、数字特征比如均值、方差、协方差等都不随时间推移而变化。另外由 h 的任意性可知，$F_i(x) = F_{i+h}(x) = F_1(x)$，故不同时刻信源发出同一个字符的概率相同，不随机时间变化，但字符之间可能互相影响。

离散平稳信源转移概率

既然统计特性与时间的流逝无关，那么（1）每个随机变量都有相同的分布（2-1）；（2）联合分布与时间推移无关；（3）条件概率具有平稳性。事实上：

$$
\begin{aligned}
p_{j|i}^{(l)}(m) &= P\{X_{m+l}=x_j|X_m=x_i\} \\
&= \frac{P\{X_{m+l}=x_j, X_m=x_i\}}{P\{X_m=x_i\}} \\
&= \frac{P\{X_{l+1}=x_j, X_1=x_i\}}{P\{X_1=x_i\}} \\
&= P\{X_{l+1}=x_j|X_1=x_i\} = p_{j|i}^{(l)}(1).
\end{aligned}
$$

续：离散平稳信源转移概率

这种条件概率表示：在时刻 m 发出字符 x_i 的条件下经过 l 长时间后在时刻 $m+l$ 发出字符 x_j 的可能性大小，称为消息序列 $X_1X_2\cdots X_n\cdots$ 在时刻 m 处的 l 步**转移概率**。但是它与时刻 1 时的 l 步转移概率相同，所以是平稳的。另外多个条件时的条件概率也是平稳的即

$$\begin{aligned} & P\{X_{m+l}=x_{l+1}|X_m=x_1,X_{m+1}=x_2,\cdots,X_{m+l-1}=x_l\} \\ = \quad & P\{X_{l+1}=x_{l+1}|X_1=x_1,X_2=x_2,\cdots,X_l=x_l\}. \end{aligned}$$

需要用两个结论

(命题 2.3.1)

平稳信源的 n 长序列的联合熵与条件熵也具有平稳性即

$$H(X_m, X_{m+1}, \cdots, X_{m+n-1}) = H(X_1, X_2, \cdots, X_n),$$
$$H(X_{m+n}|X_m, X_{m+1}, \cdots, X_{m+n-1}) = H(X_{n+1}|X_1, X_2, \cdots, X_n).$$

这可由平稳信源的联合分布与条件分布都具有平稳性得知。

(引理 2.3.1)

对离散平稳信源 $X_1, X_2, \cdots$，若 $H(X_1)$ 存在，则下面极限存在。

$$\lim_{n\to\infty} H(X_n|X_{n-1}, \cdots, X_2, X_1).$$

证明：

只需证明条件熵序列
$h_n = H(X_n|X_1, X_2, \cdots, X_{n-1}), n = 1, 2, \cdots$ 是单调有界即可。
事实上：

（1） $H(X_{n+1}|X_n, \cdots, X_2, X_1) \leq$
$H(X_{n+1}|X_n, \cdots, X_2) = H(X_n|X_{n-1}, \cdots, X_2, X_1)$
即 $h_{n+1} \leq h_n \quad n = 1, 2, \cdots$。

（2） $0 \leq h_n = H(X_n|X_{n-1}, \cdots, X_2, X_1) \leq H(X_n) =$
$H(X_1)$。

引理 2.3.2

(引理 2.3.2)

如果实数列 $a_1, a_2, \cdots, a_n, \cdots$ *存在极限* a，*则前* n *项和的算术平均仍有极限* a *即*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} a_k = a.$$

练习：

证明引理 2.3.2。

定理 2.3.4：平稳信源的熵率

对离散平稳信源 $X_1, X_2, \cdots$，若 $H(X_1)$ 存在，则它的熵率为

$$H_\infty(X) = \lim_{n\to\infty} H(X_n|X_{n-1}, \cdots, X_2, X_1). \tag{2.2}$$

证明：根据熵的链式法则及引理（3.3.1）、（2.3.2）可得

$$\begin{aligned} H_\infty(X) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} H(X_k|X_1, \cdots, X_{k-1}) \\ &= \lim_{n\to\infty} H(X_n|X_{n-1}, \cdots, X_2, X_1) \end{aligned}$$

存在。

定理 2.3.4 的意义：

这个定理说明，离散平稳信源的熵率依赖于信源从初始时刻发出的所有字符，这相当于要获得信源的任意多维的联合分布与条件分布，这在实际过程中是不可能的，通常字符的依赖关系都有一定的长度，这是一种特殊的信源即马氏信源。

定理 2.3.4：平稳信源的熵率

对离散平稳信源 $X_1, X_2, \cdots$，若 $H(X_1)$ 存在，则它的熵率为

$$H_\infty(X) = \lim_{n\to\infty} H(X_n|X_{n-1}, \cdots, X_2, X_1). \tag{2.3}$$

这个定理说明，离散平稳信源的熵率依赖于信源从初始时刻发出的所有字符，这相当于要获得信源的任意多维的联合分布与条件分布，这在实际过程中是不可能的，通常字符的依赖关系都有一定的长度，这是一种特殊的信源即马氏信源。