第 1 章　绪论与概览

信息论解答了通信理论中的两个基本问题：临界数据压缩的值(答案：熵 H)和临界通信传输速率的值(答案：信道容量 C)。因此，有人认为信息论是通信理论的一个组成部分，但我们将竭力阐明信息论远不止于此。其实，信息论在统计物理(热力学)、计算机科学(科尔莫戈罗夫(Kolmogorov)复杂度或算法复杂度)、统计推断(奥克姆剃刀(Occam Razor)：“最简洁的解释最佳”)以及概率和统计(关于最优化假设检验与估计的误差指数)等学科中都具有奠基性的贡献。

本章是“开场白”，通过介绍信息论及其关联的思想的来龙去脉，提纲挈领地给出该书的整体布局。所涉及的术语和内容，将从第 2 章开始逐步给予详细叙述和讨论。图 1-1 揭示了信息论与其他学科之间的关系。如图中所示，信息论与物理学(统计力学)、数学(概率论)、电子工程(通信理论)以及计算机科学(算法复杂度)都有交叉。我们接下来对这些交叉的领域作更详细的说明。

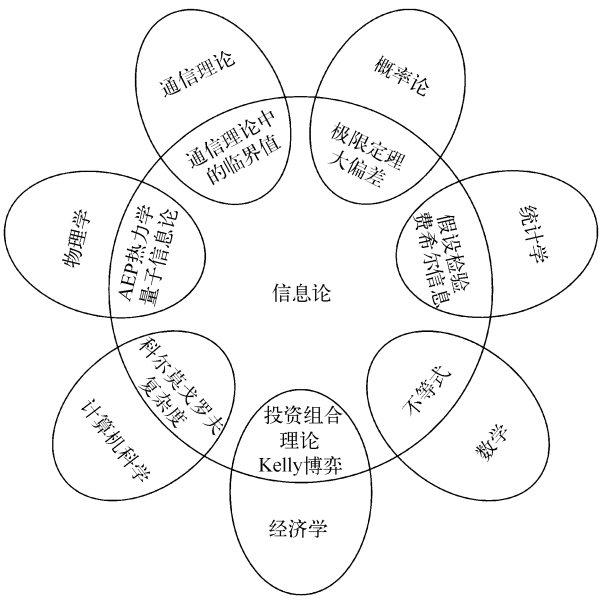

图 1-1　信息论与其他学科的关系

电子工程(通信理论)。20 世纪 40 年代早期，人们普遍认为，以正速率发送信息，而忽略误差概率是不可能做到的。然而，香农(Shannon)证明了只要通信速率低于信道容量，总可以使误差概率接近于零，这个结论震惊了通信理论界。信道容量可以根据信道的噪声特征简单地计算出来。香农还进一步讨论了诸如音乐和语音等随机信号都有一个不可再降低的复杂度，当低于该值时，信号就不可能被压缩。遵从热力学的习惯，他将这个临界复杂度命名为熵，并且讨论了当信源的熵小于信道容量时，可以实现渐近无误差通信。 1

如果将所有可能的通信方案看成一个集合，那么今天的信息论描绘了这个集合的两个临界值，如图 1-2 所示。数据压缩达到最低程度的方案对应的是该集合的左临界值 $I(X;\hat{X})$。所有数据压缩方案所需的描述速率不得低于该临界值。右临界值 $I(X;Y)$ 所对应方案的数据传输速率最大，临界值 $I(X;Y)$就是信道容量。因此，所有调制方案和数据压缩方案都必须介于这两个临界值之间。

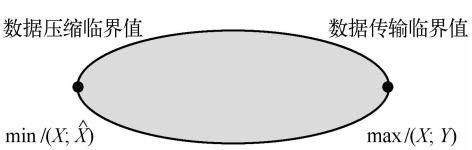

图 1-2　通信理论的信息论临界点 2

信息论也提供能够达到这些临界值的通信方案。从理论上讲，最佳通信方案固然很好，但从计算的角度看，它们往往是不切实际的。惟一的原因是，只有使用简单的调制与解调方案时才具

有计算可行性，而香农信道容量定理的证明过程中所提出的随机编码和最邻近译码规则却不然。集成电路与编码设计方面的进展使得我们能获得香农理论所蕴涵的一些硕果。随着 Turbo 码的诞生，最终实现了计算的实用性。比如，纠错码在光盘和 DVD 中的应用就是信息论的一个绝好实例。

信息论中关于通信方面的近期研究集中在网络信息论：存在干扰和噪声的情况下，大量发送器到大量接收器之间的通信同步率理论。目前，多个发送器与多个接收器之间的一些速率协定还无法预料，已有协定也有待于从数学上得到一定程度的简化。因而，一套统一的理论尚待发掘。

计算机科学(科尔莫戈罗夫复杂度)。科尔莫戈罗夫、Chaitin 和 Solomonoff 指出，一组数据串的复杂度可以定义为计算该数据串所需的最短二进制程序的长度。因此，复杂度就是最小描述长度。利用这种方式定义的复杂度是通用的，即与具体的计算机无关，因此该定义具有相当重要的意义。科尔莫戈罗夫复杂度的定义为描述复杂度的理论奠定了基础。更令人愉快的是，如果序列服从熵为 H 的分布，那么该序列的科尔莫戈罗夫复杂度 K 近似等于香农熵 H。所以信息论与科尔莫戈罗夫复杂度二者有着非常紧密的联系。实际上，科尔莫戈罗夫复杂度比香农熵更为基础。它不仅是数据压缩的临界值，而且也可以导出逻辑上一致的推理过程。

算法复杂度与计算复杂度二者之间存在着微妙的互补关系。计算复杂度(也就是时间复杂度)与科尔莫戈罗夫复杂度(也就是程序长度或描述复杂度)可以看成是对应于程序运行时间与程序长度的两条轴。科尔莫戈罗夫复杂度是沿第二条轴的最小化问题，而计算复杂度是沿第一条轴的最小化问题。沿两条轴同时进行最小化的工作几乎没有。

物理学(热力学)。熵与热力学第二定律都诞生于统计力学。对于孤立系统，熵永远增加。热力学第二定律的贡献之一是促使我们抛弃了存在永动机的幻想。我们将在第 4 章中简述该定律。

数学(概率论和统计学)。信息论中的基本量——熵、相对熵与互信息，定义成概率分布的泛函数。它们中的任何一个量都能刻画随机变量长序列的行为特征，使得我们能够估计稀有事件的概率(大偏差理论)，并且在假设检验中找到最佳的误差指数。

科学的哲学观(奥克姆剃刀)。奥克姆居士威廉说过“因不宜超出果之所需。”其意思是“最简单的解释是最佳的”。Solomonoff 和 Chaitin 很有说服力地讨论了这样的推理：谁能获得适合处理数据的所有程序的加权组合，并能观察到下一步的输出值，谁就能得到万能的预测程序。如果是这样，这个推理可以用来解决许多使用统计方法不能处理的问题。例如，这样的程序能够最终预测圆周率 π 的小数点后面遥远位置上的数值。将这个程序应用到硬币的正面出现概率为 0.7 的硬币抛掷问题中，也能得出推断。不仅如此，如果应用到股票市场，程序能从根本上抓住市场的“规律”并做出最优化的推断。这样的程序能够从理论上保证推出物理学中的牛顿三大定律。当然，这样的推理极度的不切实际，因为清除所有不适合生成现有数据的程序需要花费的时间是不可接受的。如果我们按照这种推理来预测明天将要发生的事情，那么需要花一百年的时间。

经济学(投资)。在平稳的股票市场中重复投资会使财富以指数增长。财富的增长率与股票市场的熵率有对偶关系。股票市场中的优化投资理论与信息论的相似性是非常显著的。我们将通过探索这种对偶性来丰富投资理论。

计算与通信。当将一些较小型的计算机组装成较大型的计算机时，会受到计算和通信的双重限制。计算受制于通信速度，而通信又受制于计算速度，它们相互影响、相互制约。因此，通信理论中所有以信息论为基础所开发的成果，都会对计算理论造成直接的影响。

本书概览

信息论最初所处理的问题是数据压缩与传输领域中的问题，其处理方法利用了熵和互信息等基本量，它们是通信过程的概率分布的函数。先给出一些定义，这会有助于开始讨论，在第 2 章中我们会重述这些定义。

如果随机变量 X 的概率密度函数为 $p(x)$，那么 X 的熵定义为

$$H(X) = -\sum_x p(x)\log_2 p(x) \tag{1-1}$$

使用以 2 为底的对数函数，熵的量纲为比特。熵可以看作是随机变量的平均不确定度的度量。在平均意义下，它是为了描述该随机变量所需的比特数。

例 1.1.1 考虑一个服从均匀分布且有 32 种可能结果的随机变量。为确定一个结果，需要一个能够容纳 32 个不同值的标识。因此，用 5 比特的字符串足以描述这些标识。

该随机变量的熵为

$$H(X) = -\sum_{i=1}^{32} p(i)\log p(i) = -\sum_{i=1}^{32} \frac{1}{32}\log\frac{1}{32} = \log 32 = 5 \text{ 比特} \tag{1-2}$$

这个值恰好等于描述该随机变量 X 所需要的比特数。在此情形中，所有结果都有相同长度的表示。

下面考虑一个非均匀分布的例子。

例 1.1.2 假定有 8 匹马参加的一场赛马比赛。设 8 匹马的获胜概率分布为 $\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64}\right)$。我们可以计算出该场赛马的熵为

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{16}\log\frac{1}{16} - 4\,\frac{1}{64}\log\frac{1}{64} = 2 \text{ 比特} \tag{1-3}$$

假定我们要把哪匹马会获胜的消息发送出去，其中一个策略是发送胜出马的编号。这样，对任何一匹马，描述需要 3 比特。但由于获胜的概率不是均等的，因此，明智的方法是对获胜可能性较大的马使用较短的描述，而对获胜可能性较小的马使用较长的描述。这样做，我们会获得一个更短的平均描述长度。例如，使用以下的一组二元字符串来表示 8 匹马：0，10，110，1110，111100，111101，111110，111111。此时，平均描述长度为 2 比特，比使用等长编码时所用的 3 比特小。注意，此时的平均描述长度 2 正好等于熵。在第 5 章中，我们将证明任何随机变量的熵必为表示这个随机变量所需要的平均比特数的一个下界。另外，在“20 问题”的游戏中，将所需问题的数目看成随机变量，那么它的熵也是所需问题数目的平均值的下界。我们也将说明如何构造一些表示法使其平均长度与熵相比较不超过 1 比特。

信息论中的熵与统计力学中的熵概念有着紧密的联系。如果抽出一个包含 n 个独立同分布(i.i.d.)的随机变量的序列，我们将证明该序列是“典型”序列的概率大约为 $2^{-nH(X)}$，而且大约只能抽出 $2^{nH(X)}$ 个典型序列。这个性质(著名的渐近均分性，AEP)是信息论中许多证明的基础。随后我们将介绍利用熵自然地解答的一些问题(例如，生成一个随机变量所需的抛掷均匀硬币的次数)。

随机变量的描述复杂度的概念可以推广到定义单个字符串的描述复杂度。二元字符串的科尔莫戈罗夫复杂度定义为输出该字符串所需的最短计算机程序的长度。如果字符串确实是随机的，那么其科尔莫戈罗夫复杂度接近于它的熵。从统计推断和建模问题的角度考虑，科尔莫戈罗夫复杂度是一个自然的框架，使我们对奥克姆剃刀“最简洁的解释最佳”有更加透彻的理解。我们将在第 14 章中叙述科尔莫戈罗夫复杂度的一些简单性质。

单个随机变量的熵为该随机变量的不确定度。我们还可以定义涉及两个随机变量的条件熵 $H(X|Y)$，即一个随机变量在给定另外一个随机变量的条件下的熵。由另一随机变量导致的原
6 随机变量不确定度的缩减量称为*互信息*。具体地讲，设 X 和 Y 是两个随机变量，那么这个缩减量为互信息

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1-4}$$

互信息 $I(X;Y)$是两个随机变量相互之间独立程度的度量，它关于 X 和 Y 对称，并且永远为非负值，当且仅当 X 和 Y 相互独立时，等于零。

*通信信道*是一个系统，系统的输出信号按概率依赖于输入信号。该系统特征由一个转移概率矩阵 $p(y|x)$决定，该矩阵决定在给定输入情况下输出的条件概率分布。对于输入信号为 X 和输出信号为 Y 的通信信道，定义它的信道容量 C 为

$$C = \max_{p(x)} I(X;Y) \tag{1-5}$$

以后我们将证明容量是可以使用该信道发送信息的最大速率，而且在接收端以极低的误差概率恢复出该信息。下面用一些例子来说明这点。

例 1.1.3（*无噪声二元信道*） 对于无噪声二元信道，二元输入信号在输出端精确地恢复出来，如图 1-3 所示。此信道中，任何传输的信号都会毫无误差地被接收。因此，在每次传输中，可以将 1 比特的信息可靠地发送给接收端，从而信道容量为 1 比特，也可以计算得出信道容量为 $C = \max I(X;Y) = 1$ 比特。

例 1.1.4（*有噪声四字符信道*） 观察如图 1-4 所示的信道。在该信道中，传输每个输入字符时，能够正确地接收到该字符的概率为$\frac{1}{2}$，误判为它的下一个字符的概率也为$\frac{1}{2}$。如果将 4 个输入字符全部考虑进去，那么在接收端，仅凭输出结果根本不可能确切地判定原来传输的是哪个字
7 符。另一方面，如果仅使用 2 个输入(比如 1 和 3)，我们立即可以根据输出结果知道传输的是哪个输入字符。于是，这种信道相当于例 1.1.3 中的无噪声信道，该信道上每传输一次可以毫无误差地发送 1 比特信息。此时，可以计算出信道容量 $C = \max I(X;Y)$，亦等于 1 比特/传输，这符合上述分析。

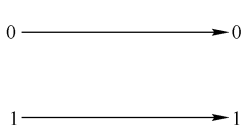

图 1-3　无噪声二元信道，$C = 1$ 比特

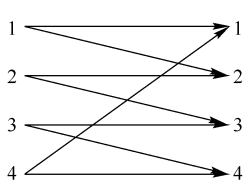

图 1-4　有噪声信道

一般，通信信道的结构不会像我们所举的例子这样简单，所以并不总能准确无误地识别出所发送的信息的某个子集。但是，如果考虑一系列传输，那么任何信道看起来都会像此例一样，并且均可以识别出输入序列集合(码字集)的一个子集，其传输信息的方式是：对应于每个码字的所有可能输出序列构成的集合近似不相交。此时，我们可以观察输出序列，能够以极低的误差概率识别出相应的输入码字。

例 1.1.5（*二元对称信道*） 二元对称信道是有噪声通信系统的一个基本例子，如图 1-5 所
8 示。此信道有一个二元输入，输出字符与输入字符相同的概率为 $1-p$。另外，0 被接收为 1 的概率为 p，1 被接收为 0 的概率也是 p。此时，可以计算得到信道容量为 $C = 1 + p\log p + (1-p)$

$\log(1-p)$比特/传输。如何达到该信道容量已经不再明显了。然而，如果多次使用该信道，那么该信道就会开始类似于例 1.1.4 所示的四字符信道，从而能以 C 比特/传输的速率发送信息而几乎不发生误差。

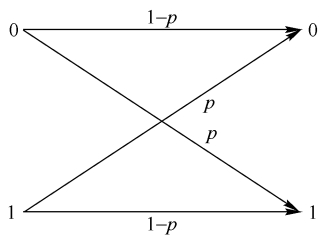

图 1-5　二元对称信道

信道上的信息通信速率的临界值由信道容量决定。信道编码定理证明该临界值可利用较长的分组编码达到。在实际的通信系统中，由于能够使用的编码的复杂度是有限制的，因此我们一般无法达到该信道容量。

互信息实际上是更广泛量的相对熵 $D(p \parallel q)$的特殊情形。相对熵是两个概率密度函数 p 和 q 之间的“距离”度量，定义为

$$D(p \parallel q) = \sum_x p(x)\log\frac{p(x)}{q(x)} \tag{1-6}$$

尽管相对熵并不是一个真正的度量，但它有着度量的某些性质。特别是相对熵总是非负的，且它为 0 的充分必要条件为 $p=q$。在两个分布 p 和 q 之间的假设检验中，相对熵就是误差概率的指数。它也可以用来定义概率分布的几何结构，使得我们能够解释大偏差理论中的许多结论。

信息论和股票市场的投资理论有许多相似之处。可将股票市场定义为一个随机向量 $\mathbf{X}$，其分量是非负的数值，等于某只股票当天的收盘价与当天的开盘价的比值。若股票市场的分布为 $F(\mathbf{x})$，那么我们定义双倍率 W 为

$$W = \max_{\mathbf{b}:\, b_i\geqslant 0,\, \sum b_i = 1} \int \log \mathbf{b}^t \mathbf{x}\, \mathrm{d}F(\mathbf{x}) \tag{1-7}$$

双倍率是财富增长的最大渐近指数。双倍率有一系列性质与熵的对应性质类似。在第 16 章将探讨这些性质。

H，I，C，D，K，W 这些量自然出现在以下领域中：

- *数据压缩*。随机变量的熵 H 是该随机变量的最短描述平均长度的下界。可以构造一个平均长度不超出熵 1 比特的描述。如果放宽完全恢复信源信息的限制，那么此时问：如果不计较失真 D 的话，需要多大的通信速率来描述信源？另外，需要多大的信道容量，才能让信源信息在信道上充分传输，并且在失真不超过 D 的情况下重构信源？这是率失真理论的研究课题。

当我们试图对非随机性目标的最短描述的概念进行严格定义时，科尔莫戈罗夫复杂度 K 的定义就应运而生了。在后面，我们将证明科尔莫戈罗夫复杂度的普适性并且满足最短描述理论的许多直观要求。

- *数据传输*。考虑信息传输问题是希望接收器能够以很小的误差概率将消息译码。从本质上讲，我们希望找到的码字(信道的输入字符序列)彼此之间离得足够远，目的是当它们在信道中被噪声污染后依然能够区分开来。这等价于高维空间中的填球问题。对任何码字集，要计算出接收器可能出错(换言之，将传送过来的码字做了错误的判断)的概率是可以办到的。然而，在绝大多数情形下，这种计算很繁琐。

使用随机生成的编码方案，香农证明了，如果码率不超过信道容量 C，就能够以任意小的误差概率发送信息。随机生成码的思想非同寻常，为简化难解问题打下了基础。香农在该证明过程中所使用的关键思想之一是所谓的典型序列概念。容量 C 是可以区分的输入信号个数的对数。

- *网络信息理论*。前面所提到的每一个主题涉及的均是单一信源或单一信道。如果我们希望压缩众多信源信息中的每一个，然后将压缩好的描述放在一起进行信源联合重构，情

况将如何？该问题由 Slepian-Wolf 定理解决。如果希望更多的发送器独立地对一个公共接收器发送信息，情况又如何？该信道的信道容量应该是多少？这样的信道称为多接入
10 信道，已由 Liao 和 Ahlswede 给予了解答。如果有一个发送器和多个接收器，同时发送相同或不相同的信息给每个接收器，该如何处理？这样的信道就是广播信道。最后，如果希望在存在噪声和干扰的背景下，任意多个发送器与任意多个接收器之间可以随意互通信息，又该如何处理？从各发送器到各接收器，可达码率的容量区域是什么？这是一般网络信息论中的问题。所有上述问题都可以归结于多用户或网络信息论这个一般化的领域。虽然要获得一个全面的网络理论超出了现有的研究水平，但我们仍然希望对上述问题的所有解答只涉及互信息和相对熵的完美形式。

- 遍历理论。渐近均分定理表明，遍历过程的绝大多数长度为 n 的样本序列的概率近似为 2^{-nH}，并且大约有 2^{nH}个是这样的典型序列。
- 假设检验。相对熵 D 在两个分布之间的假设检验中，可以表征误差概率的指数，它是两个分布之间距离的自然度量。
- 统计力学。在统计力学中，熵 H 度量一个物理系统的不确定程度或混乱程度。粗略地讲，熵是一个物理系统成形后的状态数的对数值。热力学第二定律说明，一个封闭系统的熵永不减少。后面我们会对第二定律做出一定的解释。
- 量子力学。在量子力学中，冯·诺伊曼(von Neumann)熵 $S = \mathrm{tr}(\rho \ln \rho) = \sum_i \lambda_i \log \lambda_i$ 扮演着经典的香农－玻尔兹曼(Shannon-Boltzmann)熵 $H = -\sum_i p_i \log p_i$ 的角色。由此获得数据压缩和信道容量的量子力学形式。
- 推理。我们可以运用科尔莫戈罗夫复杂度 K 的概念找到数据的最短描述，也可以将它作为模型预测下一个数据是什么。使不确定度或熵最大化的模型可导出最大熵推理方法。
- 博弈与投资。财富增长率的最佳指数由双倍率 W 决定。对于具有均匀收益机会的赛马，双倍率 W 与熵 H 之和为常数。而双倍率在边信息作用下的增量恰好是赛马与边信息之间的互信息 I。股票市场中的投资行为也有类似的结论。
- 概率论。渐近均分性(AEP)证明绝大部分序列是典型的，它们的样本熵接近于 H。因此，
11 我们可以把注意力集中在大约 2^{nH}个典型序列上。在大偏差理论中，考虑任何一个由分布构成的集合，如果真实分布到这个集合最近元的相对熵距离为 D，那么它的概率大约为 2^{-nD}。
- 复杂度理论。科尔莫戈罗夫复杂度 K 是对象的描述复杂度的度量。它与计算复杂度有一定的关系，但不尽相同，因为计算复杂度度量的是计算所需要的时间或空间大小。

信息论中的量(例如熵和相对熵)解决了通信理论和统计学中的许多基本问题而频频出现在该两门学科中。在研究这些问题之前，我们将先研究这些量的一些性质。在第 2 章中，我们开始
12 从熵、相对熵和互信息的定义及其基本性质切入正题。