非经典关联和量子相干的熵度量

席政军 著

非外借

科学出版社

非经典关联和量子相干的熵度量

席政军 著

科学出版社

北京

内 容 简 介

本书系统介绍用熵来度量和刻画量子系统中非经典关联和量子相干，主要内容是作者近十年来研究工作的总结，同时也兼顾国内外相关领域中最新的研究成果. 全书共 5 章，具体包括非经典关联和量子相干度量过程中用到的量子信息论的基础知识；不同背景下非经典关联的度量定义、性质，不同度量之间的层次关系以及与纠缠之间的联系等；正交基和非正交基下叠加相干的度量理论以及叠加相干和非经典关联之间的关系等.

本书可作为高等院校计算机类、信息类等专业高年级本科生和研究生的参考用书，同时对相关专业的研究人员也有一定的参考价值.

图书在版编目(CIP)数据

非经典关联和量子相干的熵度量/席政军著. —北京：科学出版社，2019.1
ISBN 978-7-03-060020-2

Ⅰ. ①非… Ⅱ. ①席… Ⅲ. ①熵-应用-量子-度量-研究 Ⅳ. ①O4

中国版本图书馆 CIP 数据核字(2018) 第 292922 号

责任编辑：李 萍 / 责任校对：郭瑞芝
责任印制：张 伟 / 封面设计：陈 敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京中石油彩色印刷有限责任公司 印刷
科学出版社发行 各地新华书店经销
*
2019 年 1 月第 一 版 开本：720 × 1000 B5
2019 年 1 月第一次印刷 印张：15
字数：302 000
定价：98.00 元
(如有印装质量问题，我社负责调换)

前　言

量子力学的叠加原理是量子力学与经典力学的重要区别之一, 也是导致非经典特征的必要条件。量子纠缠是一类非经典特征的重要体现, 在量子计算和量子信息处理中扮演着重要的角色, 如量子隐形传态和超密编码. 随着量子计算、量子信息理论以及量子通信的快速发展, 这些非经典的特征更加凸显出其非常重要的作用. 同时, 对非经典特征的认知不仅局限于量子纠缠, 还有其他的非经典特征也在量子计算和量子信息中有非常重要的应用. 例如, 最近研究发现两体量子系统中可分量子态也存在非经典特征, 并得到实验证实且在量子计算中得到应用. 与量子纠缠的研究历史相比, 这些非经典特征只有近二十年的发展历程, 但已经形成了量子信息中的一个重要领域.

量子信息论的发展目前还远没有经典信息论那样繁荣, 仅部分得到实际应用. 目前相对比较成熟的研究是 Shannon 信息论部分基础内容的量子版本的推广, 并借助其成熟的思想来刻画量子系统的非经典特征. 例如, 使用经典相对熵的量子版本即量子相对熵来刻画量子纠缠取得了非常好的成果. 由于经典信息 (或者经典关联) 是通过测量可以获得的, 在量子系统中全部信息 (或全部关联) 是由量子互信息来刻画, 但可观测到的经典信息不超过全部信息. 用信息论的方式研究所不可观测到的信息 (或关联) 相对于经典信息, 称为非经典信息或非经典关联, 也有学者称之为量子关联或者量子信息. 这些非经典关联是由量子力学中的叠加原理导致的, 量子态的叠加现象在经典物理中并不存在, 这是量子力学独有的特征, 也可以理解为非经典信息的一种体现. 如果说非经典关联必须在两体或者两体以上的复合量子系统中出现, 那么这种叠加相干或者量子相干在单个物理系统中都可能会存在. 没有叠加相干就不可能有非经典关联, 存在非经典关联, 可以断言在复合量子系统上一定存在叠加相干. 本书的一个研究重点就是基于信息论的方法探讨叠加相干以及它与非经典关联的关系等.

本书内容共三大部分, 从基本数学概念到实际应用处理. 先介绍量子信息的基础内容, 再介绍与之相关的最新进展. 在主体部分, 先介绍基本度量方式的相关来源, 再重点阐述相关的成果, 最后介绍相关的工作和进展.

第一部分包括第 1 章和第 2 章, 主要概述非经典关联和量子相干度量过程中

用到的基本数学知识和量子信息论的知识. 第 1 章介绍概率论、Shannon 信息论以及量子力学的背景知识. 第 2 章先介绍量子信息论的基础知识, 再重点讨论带噪量子信道上量子信息的处理方法, 并讨论 Holevo 信息、相干信息、信息增益以及逆相干信息, 最后概述量子纠缠的度量.

第二部分包括第 3 章和第 4 章, 主要阐述非经典关联的描述、度量和刻画. 第 3 章给出非经典关联的度量——量子失协, 重点讨论该量化的基本性质以及最优化过程. 第 4 章讨论对偶的量子失协——单向非局域量子失协, 先从定义到基本性质进行介绍, 再对它与纠缠等其他量子关联之间的关系做详细的阐述.

第三部分是第 5 章, 系统阐述量子相干性的资源理论体系. 首先介绍资源理论框架, 在该理论框架下, 介绍相干态、非相干态以及非相干操作的基本定义. 其次, 基于量子相干性度量的公理化要求, 介绍几类常见的度量工具. 重点阐述相对熵相干度量, 讨论基于相对熵作为统一度量尺度, 较为全面分析多体量子系统上非经典关联、全局量子相干以及局部量子相干之间的关系. 再次, 分析量子操作引起的相干性的变化, 给出相干分布和非经典关联之间的一些关系. 最后, 以实际应用为背景, 阐述非正交基下叠加的度量, 修改并完善已有的叠加资源理论, 提出叠加诱导的相干性资源理论.

在本书撰写过程中, 作者得到了李永明教授的鼓励和支持. 感谢李老师对作者自研究生求学十多年来的谆谆教诲和悉心指导! 同时, 感谢浙江大学的王晓光教授在学术交流中对许多问题的深刻指导! 感谢中国科学院物理所的范桁教授提供多次学术交流和讨论的机会, 并给出启发式指导! 感谢西安邮电大学胡明亮教授给作者提供讲授本书的机会! 感谢陕西师范大学计算机科学学院王小明教授、雷秀娟教授和马苗教授等的关心和支持! 另外, 感谢科学出版社对本书出版的大力支持! 最后, 感谢陕西师范大学优秀著作出版基金、211 学科建设经费、陕西师范大学优秀青年学术骨干资助计划、陕西省创新人才推进计划项目和国家自然科学基金为本书出版所提供的资助!

本书是作者十余年来在非经典特征度量方面研究工作的系统总结, 由于作者水平有限, 书中不足之处在所难免, 恳请读者不吝指正.

目　录

第 1 章　经典信息论引论

本章介绍本书将要用到的一些基础知识, 包括概率论和经典信息论的基本内容. 首先介绍概率空间和随机变量的一些基本定义和概念, 然后介绍经典信息论的相关知识, 但这里并不涉及具体的证明.

1.1　概率论的基础知识

1.1.1　随机变量

概率和随机变量在信息论中是两个非常重要的概念. 粗略地讲, 随机变量可以看作是描述一个经典系统的一些物理自由度的值. 因此, 在经典信息论中，数据就用随机变量来描述. 为了后面内容介绍的方便, 下面给出概率空间和随机变量的公理化解释, 并通过规定概率应具备的基本性质来定义概率, 也就是 Kolmogorov 公理化. 首先介绍下面三个概念.

(1) 样本空间 Ω: 所有可能出现结果的集合;

(2) 事件集 $\mathcal{E}$: 样本空间的一个子集合;

(3) 概率测度 P: 给出任何事件的概率.

事件集 $\mathcal{E}$ 是一个 σ-代数, 即

(1) $\mathcal{E} \neq \varnothing$;

(2) 如果 $E \in \mathcal{E}$, 则 $E^C \in \mathcal{E}$;

(3) 如果 $(E_i)_{i\in N}$ 是有限可数的, 那么 $\bigcup_{i\in N} E_i$ 是事件.

$(\Omega, \mathcal{E})$ 上的概率测度是一个实值函数, 即 $P : \mathcal{E} \to \mathbb{R}^+$. 记 $p(E)$ 是事件 E 的概率, 那么它满足 Kolmogorov 概率公理化要求:

(1) $p(\Omega) = 1$;

(2) $p[\bigcup_{i\in N} E_i] = \sum_{i\in N} p(E_i)$, 其中 E_i 两两不相交.

显然, 该公理化的要求与事件的 σ-代数是相容的, 且 $E \bigcap E^C = \varnothing$, 则 $p(E) +$

$p(E^C) = p(\Omega) = 1$. 因此, 称 $(\Omega, \mathcal{E}, P)$ 是一个概率空间 (probability space), 而 $(\Omega, \mathcal{E})$ 是一个可测空间 (measurable space).

在概率理论中, 随机变量是某个集合上的一个函数. 假设有概率空间 $(\Omega, \mathcal{E}, P)$, 且 $(\mathcal{X}, \mathcal{F})$ 是一个可测空间. 那么, 随机变量 X 是从 Ω 到 X 的一个函数:

$$X : \omega \to X(\omega). \tag{1.1}$$

它是关于 σ-代数 $\mathcal{E}$ 和 $\mathcal{F}$ 可测量的. 可测的意义是: 对 $\forall F \in \mathcal{F}$, 有 $X^{-1}(F) \in \mathcal{E}$. 从而事件 $\mathcal{F}$ 从概率空间继承了一个概率测度 P_X, 对所有的 $F \in \mathcal{F}$, 定义

$$P_X[F] = P[X^{-1}(F)]. \tag{1.2}$$

因此, 事件可以用随机变量来定义. 例如, 若 X 的值域是实数, 则

$$E = \{\omega \in \Omega : X(\omega) > x_0\} \tag{1.3}$$

是一个事件.

习惯上, 去掉 ω, 简记作 $X > x_0$. 如果事件就是一个函数的自变量, 记 $P[X > x_0]$ 为事件 $\{X > x_0\}$ 的概率.

在信息论中, 大部分用到的是随机变量, 并非事件. 对一个随机变量 X, 有一些有限的值, 样本空间就取 $\mathcal{X}$, 事件的 σ-代数是它的幂集, 且

$$X : \Omega \to \mathcal{X}, \tag{1.4a}$$

$$X : \omega \to X(\omega). \tag{1.4b}$$

因此, 习惯上记 $\mathcal{X}$ 是随机变量 X 的字母表, 概率密度函数 $P_X(x)$ 表示事件 $\{X = x\}$ 的概率, 且满足完备性条件

$$\sum_{x \in \mathcal{X}} P_X(x) = 1. \tag{1.5}$$

为了方便, 记 $p(x) = P_X(x)$, 这就是一个概率分布.

接下来介绍联合概率分布、边际概率分布和条件概率分布. 对于一个二元随机变量 X 和 Y, 其字母表为 $\mathcal{X}$ 和 $\mathcal{Y}$, 它们的联合概率分布记作 P_{XY}, 则 P_X 和 P_Y 是边际概率. 由于对任意的 $y \in \mathcal{Y}$, 有

$$P_Y(y) = \sum_{x \in \mathcal{X}} P_{XY}(x, y). \tag{1.6}$$

因此, 对任意的 $y \in \mathcal{Y}$ 且 $P_Y(y) > 0$, 随机变量 X 关于事件 $Y = y$ 的概率分布 $P_{X|Y=y}$ 满足 Bayes 定律:

$$p(x|y) = P_{X|Y=y}(x) = \frac{P_{XY}(x,y)}{P_Y(y)}. \tag{1.7}$$

或者说, 对所有的 $x \in \mathcal{X}$, 都有

$$p(x|y) = \frac{p(x,y)}{p(y)}. \tag{1.8}$$

特别地, 如果事件 $\{X = x\}$ 和 $\{Y = y\}$ 是相互独立的, 那么两个随机变量 X 和 Y 是相互独立的. 对于任意的 $(x,y) \in \mathcal{X} \times \mathcal{Y}$, 它们的联合分布满足

$$P_{XY}(x,y) = P_X(x) \cdot P_Y(y). \tag{1.9}$$

1.1.2 随机变量的函数

在 1.1.1 小节介绍了随机变量的基本定义, 本小节将给出随机变量的一些函数, 如期望值、方差以及协方差等.

定义1.1.1 设 X 为离散随机变量, 其概率密度函数为 $p(x)$, 则随机变量 X 的期望函数 (expectation function) 或期望值 $E(X)$ 定义为

$$E(X) = \sum_{x:p(x)>0} xp(x). \tag{1.10}$$

期望值反映的是随机变量平均取值的大小, 并不一定包含于变量的输出值集合. 期望值 $E(X)$ 是随机变量 X 可能值的加权平均, 但并不能说明这些值的分散程度. 例如, 随机变量 W, Y 以及 Z 的概率密度函数定义如下:

$$\begin{aligned} W &= 0, \quad p = 1; \\ Y &= \begin{cases} -1, & p = \dfrac{1}{2}, \\ +1, & p = \dfrac{1}{2}; \end{cases} \\ Z &= \begin{cases} -100, & p = \dfrac{1}{2}, \\ +100, & p = \dfrac{1}{2}. \end{cases} \end{aligned} \tag{1.11}$$

容易计算这三个随机变量的期望值均为 0. 但显然 Y 分散比 W 大, Z 分散比 Y 大. 因此, 希望 X 在均值 $E(X)$ 范围内取值, 这就又产生了如何合理地度量 X

的可能变差的问题, 即如何刻画随机变量 X 与均值偏离程度. 下面引入方差的概念来解决此问题.

定义1.1.2　若随机变量 X 的期望值为 $E(X)=\mu$, 那么随机变量 X 的方差 (variance) $V(X)$ 定义为

$$V(X)=E((X-\mu)^2). \tag{1.12}$$

通过简单的代数运算就得到一个紧凑的表达式, 即

$$V(X)=E(X^2)-E(X)^2. \tag{1.13}$$

方差有一些物理解释. 例如, 若把均值看成质量分布的重心, 方差表示力学中的转动惯量. 有时需要标准方差, 它是方差 $V(X)$ 的平方根, 记为 $\mathrm{SD}(X)$, 即 $\mathrm{SD}(X)=\sqrt{V(X)}$. 也可以将方差的定义推广到两个随机变量的情形, 也就是协方差 (covariance).

定义1.1.3　对于两个随机变量 X 和 Y, 其协方差 $C(X,Y)$ 定义为

$$C(X,Y)=E[(X-E(X))(Y-E(Y))]. \tag{1.14}$$

经过计算, 协方差 $C(X,Y)$ 有一个更直观的表达式, 即

$$C(X,Y)=E(XY)-E(X)E(Y). \tag{1.15}$$

当然也可以定义条件的数学期望. 由于 $E(X|Y=y)$ 表示随机变量 Y 在 $Y=y$ 的期望值, 那么 $E(X|Y=y)$ 也是一个随机变量, 从而给出条件的数学期望, 即

$$E(X|Y)=E(E(X|Y=y))=\sum_{y\in\mathcal{Y}}E(X|Y=y)p(Y=y). \tag{1.16}$$

本小节最后介绍 Jensen 不等式, 该不等式是信息论中许多基本结论的基础. 设 X 是一个随机变量, 其字母表为 $\mathcal{X}$, 设 $f:\mathcal{X}\to\mathcal{Y}$, 记 $\mathcal{Y}=f(x)$, 则随机变量 Y 的概率分布为

$$P_Y(y)=\sum_{x\in f^{-1}(\{y\})}P_X(x). \tag{1.17}$$

对于凸集 $\mathcal{X}$ 上的一个凸函数 f (实值), 随机变量 X 和 $f(X)$ 的期望值由 Jensen 不等式给出

$$E(f(X))\geqslant f(E(X)), \tag{1.18}$$

其中,

$$E(f(X)) = \sum_{x\in\mathcal{X}} f(x)p(x). \tag{1.19}$$

1.1.3 概率分布之间的距离

设 P 和 Q 是字母表 $\mathcal{X}$ 上的两个概率分布, 它们之间的迹距离 (trace distance) 定义为

$$\delta(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |p(x) - q(x)|. \tag{1.20}$$

迹距离也叫做统计距离或 Kolmogorov 距离. 易证 $\delta(P,Q)$ 是一个数学上的距离, 满足下面三个条件:

(1) $\delta(P,Q) \geqslant 0$, $\delta(P,Q) = 0$ 当且仅当 $P = Q$;

(2) $\delta(P,Q) = \delta(Q,P)$;

(3) $\delta(P,Q) \leqslant \delta(P,R) + \delta(R,Q)$, 其中 R 也是字母表 $\mathcal{X}$ 上的概率分布.

根据 $\delta(P,Q)$ 的定义, 有

$$\begin{aligned}
\delta(P,Q) &= \frac{1}{2}\left[\sum_{p>q}(p(x)-q(x)) + \sum_{q>p}(q(x)-p(x))\right] \\
&= \frac{1}{2}\left[1 - \sum_{q>p} p(x) - \sum_{p>q} q(x) + 1 - \sum_{p>q} q(x) - \sum_{q>p} p(x)\right] \\
&= 1 - \sum_{q>p} p(x) - \sum_{p>q} q(x) \\
&= 1 - \sum_{q>p} \min[p(x),q(x)] - \sum_{p>q} \min[p(x),q(x)] \\
&= 1 - \sum_{x\in\mathcal{X}} \min[p(x),q(x)].
\end{aligned} \tag{1.21}$$

从而有

$$\delta(P,Q) = 1 - \sum_{x\in\mathcal{X}} \min[p(x),q(x)]. \tag{1.22}$$

迹距离的一个很好的应用就是给出了如何分辨两个概率分布的一个直接的界, 即

$$\delta(P,Q) = \max_{E\in\mathcal{X}} |P_X(E) - Q_X(E)|, \tag{1.23}$$

其中, 最大取遍所有可能的事件.

证明　由于

$$\begin{aligned}|P(E)-Q(E)| &= \left|\sum_{x\in E} p(x) - \sum_{x\in E} q(x)\right| \\ &\leqslant \left|\sum_{x\in E'} p(x) - \sum_{x\in E'} q(x)\right|, \end{aligned} \tag{1.24}$$

其中, $E'=\{x|p(x)\geqslant q(x), x\in E\}$, 且 $E'\subseteq E$. 又

$$\begin{aligned}\delta(P,Q) &= \frac{1}{2}\sum_{x\in\mathcal{X}} |p(x)-q(x)| \\ &= \frac{1}{2}\left[\sum_{p>q}(p(x)-q(x)) + \sum_{q>p}(q(x)-p(x))\right] \\ &= \sum_{p>q}(p(x)-q(x)) \\ &= \sum_{x\in E^\dagger} p(x) - \sum_{x\in E^\dagger} q(x), \end{aligned} \tag{1.25}$$

其中, $E^\dagger=\{x|p(x)\geqslant q(x), x\in\mathcal{X}\}$, 且 $E^\dagger\subseteq\mathcal{X}$, $E'\subseteq E^\dagger$.

从而对于每一个事件 E, 都有下面不等式成立, 即

$$|P(E)-Q(E)|\leqslant\delta(P,Q). \tag{1.26}$$

再由式 (1.25) 可知

$$\delta(P,Q)\leqslant\max|P_X(E)-Q_X(E)|. \tag{1.27}$$

联立式 (1.26) 和式 (1.27), 有

$$\delta(P,Q)=\max_{E\in\mathcal{X}}|P_X(E)-Q_X(E)|. \tag{1.28}$$

从而可知事件 $E^\dagger$ 在某种意义上就是试图区分概率分布 P 和 Q 时所要检测的最优事件, 而迹距离表示这种区分的程度. 也可以去掉迹距离中的绝对值符号, 即

$$\delta(P,Q)=\max_{E\in\mathcal{X}}[P(E)-Q(E)]=\max_{E\in\mathcal{X}}\left[\sum_{x\in E}p(x)-\sum_{x\in E}q(x)\right].$$

由于 $\forall E \in \mathcal{X}$, 有

$$\begin{aligned} P(E) - Q(E) &= \sum_{x \in E} p(x) - \sum_{x \in E} q(x) \\ &= \left[\sum_{p>q} p(x) - \sum_{p>q} q(x)\right] - \left[\sum_{q>p} q(x) - \sum_{q>p} p(x)\right] \\ &\leqslant \sum_{x \in E^\dagger} p(x) - \sum_{x \in E^\dagger} q(x), \end{aligned}$$

其中, $E^\dagger \subseteq E$ 是事件. 显然 $\delta(P, Q)$ 中的事件是最优的, 也是可达的. □

1.2 Shannon 熵及其性质

熵的概念首先在热力学中引入, 用于表述热力学第二定律, 通常定义为物理系统的微观状态数的对数值. 在 20 世纪 30 年代, Hartley 在通信系统中引入了信息的对数度量. 本节介绍的熵的定义是由 Shannon 在 1948 年首次给出的. Shannon 在文献 [1] 中首次建立了数学模型, 完整地解决了通信速度上限的问题, 奠定了现代信息论的基础. 信息论最初所处理的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的基本量. 信息论中的熵与统计力学中的熵概念有着密切的联系. 如果抽出一个包含 n 个独立同分布的随机变量的序列, 那么该序列是“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. 这就是著名的渐近均分性, 是信息论中许多理论证明的基础. Shannon 三大定理 (可变长无失真信源编码定理、有噪信道编码定理、保失真度准则下的有失真信源编码定理) 是信息论的基础理论, 虽然是存在性定理, 并没有提供具体的编码实现方法, 但是为通信信息的研究提供了方法. 信息论已有相当大的发展, 在信息科学、通信理论、统计物理、计算机科学、统计推断以及概率论等学科中都有非常重要的贡献. 限于篇幅, 本节只给出一些最基本的概念和定义.

1.2.1 Shannon 熵的定义

信息是个相当泛的概念, 很难用一个简单的定义将其完全准确地描述. 然而对于任何一个概率分布, 可定义一个称为熵的量, 它具有许多符合度量信息的直观要求的特性. 下面首先介绍单个随机变量的 Shannon 熵定义, 再将该定义推广到两个

随机变量的情形.

定义1.2.1　设随机变量 X 的概率密度函数为 $p(x)$, 那么随机变量 X 的 Shannon 熵 $H(X)$ 定义为

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x). \tag{1.29}$$

也可写成 $H(X) = -\sum_x p_x \log_2 p_x$, 有时也简写为 $H(p)$. 这里约定 $0\log_2 0 = 0$.

结合期望的定义, 可以将 Shannon 熵表示为期望. 如果 $X \sim p(x)$, 记 $g(X) = \log_2 \dfrac{1}{p(X)}$, 则随机变量 X 的 Shannon 熵可解释为随机变量 $\log_2 \dfrac{1}{p(X)}$ 的期望, 即

$$H(X) = E\left(\log_2 \frac{1}{p(X)}\right). \tag{1.30}$$

假定要获取随机变量 X 的值, 它的 Shannon 熵是对平均值在获取 X 的值的过程中得到信息多少的度量. 换个说法, 随机变量 X 的 Shannon 熵是在得到 X 的值之前关于 X 的不确定性的测度. 这两种说法互补, 既可以把熵看作得到 X 的值之前不确定性的一种测度, 又可以把它视为得到 X 值之后获得信息多少的一种测度. 简言之, 熵是随机变量不确定的度量, 也是平均意义上描述随机变量所需的信息量的度量. 熵可以作为有效描述长度的一个自然度量, 用于量化存储信息所需要的资源. 随机变量包含信息不依赖于随机变量可取值的符号.

定义1.2.2　若 $(X,Y) \sim p(x,y)$, 则其联合熵 $H(X,Y)$ 定义为

$$H(X,Y) = -\sum_{x,y} p(x,y)\log_2 p(x,y). \tag{1.31}$$

类似地, 可以定义一个随机变量在给定另一个随机变量下的条件熵, 它是条件分布熵关于起条件作用的那个随机变量取平均之后的期望值.

定义1.2.3　若 $(X,Y) \sim p(x,y)$, 则其条件熵 $H(Y|X)$ 定义为

$$H(Y|X) = \sum_x p(x)H(Y|X=x). \tag{1.32}$$

联合熵和条件熵的自然定义可由一个链式法则表示, 一对随机变量联合熵等于其中一个随机变量的熵加上另一个随机变量的条件熵, 即

$$H(X,Y) = H(X) + H(Y|X). \tag{1.33}$$

同理, 有

$$H(X,Y) = H(Y) + H(X|Y). \tag{1.34}$$

由于 $H(X|Y)$ 表示在 Y 下的随机变量的不确定性, 且 $H(X|Y) \geqslant 0$, 从而

$$H(Y) \leqslant H(X,Y). \tag{1.35}$$

同理,

$$H(X) \leqslant H(X,Y). \tag{1.36}$$

当然, 可以将熵的概念推广到多个随机变量上, 利用条件熵之和来表示, 这就是熵的链式法则.

定理1.2.1 若随机变量 $(X_1, X_2, \cdots, X_n) \sim p(x_1, x_2, \cdots, x_n)$, 则

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1). \tag{1.37}$$

该定理的证明重复利用两个随机变量的熵展开即可. 由于本书后面会涉及强次可加性, 下面仅考虑三变量的情形. 上述定理在证明过程中的一个副产品就是条件联合熵. 对于随机变量 X, Y 和 Z, 有

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z). \tag{1.38}$$

1.2.2 相对熵与互信息

在 1.1.3 小节中, 介绍了概率分布之间的距离的度量. 现在, 将介绍另一种在信息论中非常重要的距离的度量, 它描述同一指标上两个概率分布接近的程度. 相对熵 (relative entropy) 是两个随机分布之间距离的度量. 在统计学中, 它对应的是似然比的对数期望. 当真实分布为 $p(x)$ 而假定分布为 $q(x)$ 时, 相对熵度量具有无效性. 例如, 已知随机变量的真实分布为 $p(x)$, 平均构造描述长度为 $H(P)$ 的码. 但如果使用针对分布 $q(x)$ 的编码, 那么平均意义上就需要 $H(P) + H(P||Q)$ 比特态来描述这个随机变量.

定义1.2.4　设 $p(x)$ 和 $q(x)$ 是字母表 $\mathcal{X}$ 上的两个概率分布, 则它们的相对熵或 Kullback-Leibler 距离定义为

$$H(P||Q) = -\sum_{x\in\mathcal{X}} p(x)\log_2\frac{p(x)}{q(x)}. \tag{1.39}$$

约定 $0\log_2\frac{0}{0}=0, p\log_2\frac{p}{0}=\infty$.

下面的定理给出了相对熵非负性的性质, 但是它并不满足对称性和三角不等式. 因此, 相对熵并不是数学意义上的距离.

定理1.2.2　相对熵是非负的, 即

$$H(P||Q) \geqslant 0, \tag{1.40}$$

当且仅当对任意的 x, $p(x)=q(x)$ 时等号成立.

证明　设 $A=\{x|p(x)>0\}$ 是 $p(x)$ 的支撑集, 则

$$\begin{aligned} H(P||Q) &= \sum_{x\in A} p(x)\log_2\frac{p(x)}{q(x)} && (1.41)\\ &= -\sum_{x\in A} p(x)\log_2\frac{q(x)}{p(x)} && (1.42)\\ &\geqslant -\log_2\sum_{x\in A} p(x)\frac{q(x)}{p(x)} && (1.43)\\ &= -\log_2\sum_{x\in A} q(x) && (1.44)\\ &\geqslant -\log_2\sum_{x\in\mathcal{X}} q(x) && (1.45)\\ &= 0. && (1.46) \end{aligned}$$

式 (1.43) 使用了 Jensen 不等式. 由于 $\log_2 x$ 是关于 x 的严格凹函数, 当且仅当 $q(x)/p(x)$ 为常数时, 不等式 (1.43) 中等号成立. 同时, 只有 $\sum_{x\in A} q(x)=\sum_{x\in A} p(x)=1$ 时, 不等式 (1.45) 中等号才成立. 从而完成证明. □

相对熵的非负性的证明也可以由信息论中另一个重要的不等式 $\ln x \leqslant x-1$ (x 取所有的正实数, 当且仅当 $x=1$ 时等号成立) 来证明. 虽然相对熵不能作为数学上的距离, 但相对熵在信息论中有许多重要的应用. 其中, 一个直接的应用就是给出熵的一个上界.