中国科学: 数学　2026 年　第 56 卷　第 1 期: 1～24
SCIENTIA SINICA Mathematica
《中国科学》杂志社
SCIENCE CHINA PRESS
CrossMark
click for updates
综　述

熵中心极限定理综述

献给林正炎教授 85 寿辰

马志明[1,2*], 姚柳全[1,2], 原帅[1,2]

1. 中国科学院大学数学科学学院, 北京 100049;
2. 中国科学院数学与系统科学研究院, 北京 100190
E-mail: mazm@amt.ac.cn, yaoliuquan20@mails.ucas.ac.cn, yuanshuai2020@amss.ac.cn

收稿日期: 2024-02-29; 接受日期: 2024-05-10; 网络出版日期: 2024-12-09; * 通信作者
国家重点研发计划 (批准号: 2023YFA1009603) 资助项目

摘要　本文整理和总结自 20 世纪 50 年代至今的熵中心极限定理的相关研究, 包括独立同分布和形式的极限分布、熵中心极限定理的 Berry-Esseen 界以及非独立情形时的收敛结果. 通过梳理熵中心极限定理的发展历程, 我们希望展示出热力学、信息论和概率论这三门不同领域学科间的交叉融合与微妙联系. 本文最后部分简要介绍本文作者在条件熵中心极限定理研究方向的一些最新研究成果及其在 Hadamard 压缩问题上的应用.

关键词　熵中心极限定理　熵不等式　条件中心极限定理　Hadamard 变换

MSC (2020) 主题分类　60F05, 68P30, 62B10

1　引言

1.1　中心极限定理

中心极限定理作为概率论的重要极限理论之一, 始终备受关注. 这一理论揭示了包括独立和在内的诸多随机变量在满足一定条件下, 都可以收敛到 Gauss 分布这一事实. 这一发现不仅丰富和发展了概率论的极限理论, 更为人们生活中的一些现象提供了解释, 并为许多生活实践提供了指导作用. 事实上, 在自然科学、经济科学以及工程理论中, 都常常遇到某些由彼此不相干的随机因素产生的参量, 例如在通信领域中来自信道外界的噪声干扰, 或在线性回归分析中除控制变量以外的随机扰动. 中心极限定理就是使用概率论的方法, 研究这一类随机量的规律, 说明虽然每一个因素个体作用都不大, 但随着数量增加其叠加作用呈现的规律性 (参见文献 [38]).

英文引用格式: Ma Z-M, Yao L Q, Yuan S. A review for entropic central limit theorem (in Chinese). Sci Sin Math, 2026, 56: 1–24, doi: 10.1360/SSM-2024-0049

1.1.1 特征函数方法

经典中心极限定理的研究对象为独立同分布和, 即随机变量

$$W_n := \frac{1}{\sqrt{n}}\sum_{k=1}^{n} X_i, \quad n \in \mathbb{N}, \tag{1.1}$$

其中 $X_1, X_2, \ldots, X_n$ 是独立同分布 (independent and identically distributed, i.i.d.) 的随机变量, 满足 $\mathrm{E}X_1 = 0$, $\mathrm{Var}(X_1) = 1$. 为探究 W_n 的极限分布, 特征函数方法通过随机变量的特征函数来对分布进行研究. 具体而言, 一个随机变量 X 的特征函数 $\varphi_X(t)$ 定义为

$$\varphi_X(t) := \mathrm{E}\mathrm{e}^{\mathrm{i}tX}. \tag{1.2}$$

在独立和情形下, 特征函数可将随机变量卷积转化为乘积形式, 即对于任意两个独立随机变量 X 和 Y, 有

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t), \tag{1.3}$$

因此研究随机变量和情形时, 特征函数可以提供很好的简化作用. 再根据逆转公式

$$\lim_{T\to\infty}(2\pi)^{-1}\int_{-T}^{T}\frac{\mathrm{e}^{-\mathrm{i}ta}-\mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t}\varphi_X(t)dt = \Pr(X\in(a,b)) + \frac{1}{2}\Pr(X\in\{a,b\}), \quad \forall\, a<b, \tag{1.4}$$

可知特征函数可以确定分布 (参见文献 [16]). 事实上可以证明对于独立和情形, 极限分布与特征函数的极限存在着相互唯一决定的关系. 依据上述理论, 可以得到经典 i.i.d. 形式的中心极限定理 (参见文献 [16, 定理 3.4.1]).

定理 1.1 (i.i.d. 形式的中心极限定理) 设 $X_1, X_2, \ldots$ 是 i.i.d. 随机变量列, 均值为 0, 方差为 1, 则

$$W_n := \frac{1}{\sqrt{n}}\sum_{k=1}^{n} X_i \Rightarrow G, \tag{1.5}$$

其中 $\Rightarrow$ 表示弱收敛或依分布收敛, G 表示标准正态随机变量.

事实上, 根据特征函数理论, 可以得到独立和收敛到 Gauss 的充分必要条件 (参见文献 [38, 定理 10.3.10]).

定理 1.2 (Lindeberg-Feller 中心极限定理) 设 X_{nk}, $k = 1, 2, \ldots, k_n$ 为一个三角型 (triangular arrays) 的独立随机变量阵列, 满足 $\mathrm{E}(X_{nk}) = 0$, $\mathrm{Var}(X_{nk}) = \sigma_{nk}^2$, $\forall\, n, k$, 且 $\lim_{n\to\infty}\sum_k \sigma_{nk}^2 = 1$, 则 $\sum_k X_{nk} \Rightarrow G$ 且 $\max_k \sigma_{nk}^2 \to 0$ 的充要条件是

$$\lim_{n\to\infty}\sum_k \int_{|x|\geqslant\varepsilon} x^2 F_{X_{nk}}(dx) = 0, \quad \forall\, \varepsilon > 0. \tag{1.6}$$

该条件称为 Lindeberg 条件.

1.1.2 Stein 方法

Stien 方法最早由 Stein 于 1972 年提出, 该方法的核心是 Gauss 分布的如下性质:

$$\mathrm{E}(Zf(Z)) = \sigma^2\mathrm{E}(f'(Z)) \tag{1.7}$$

对于任意绝对连续函数成立当且仅当 $Z \sim \mathcal{N}(0,\sigma^2)$. 当随机变量 W 靠近 Gauss 随机变量 Z 时, 应有 $\mathrm{E}(Wf(W)) \approx \sigma^2 \mathrm{E}(f'(W))$. 与此同时, 对于一些可测函数 h, 应有

$$\mathrm{E}h(W) - \mathrm{E}h(Z) \approx 0, \tag{1.8}$$

因此获得了两种判断随机变量是否接近 Gauss 随机变量的方法.

给定可测函数 h, 将上述两种方案结合, 得到了 Stein 方程

$$f'(w) - wf(w) = h(w) - \mathrm{E}h(Z), \tag{1.9}$$

其中 Z 是 Gauss 分布, $f = f_h$ 称为 Stein 方程的解. 为了将 $wf(w)$ 写成 f' 的相关形式从而使用 Taylor 展开, 定义 W 的零偏估计 (zero bias) W^*: 给定均值为 0、方差为 σ^2 的随机变量 W, 如果随机变量 W^* 满足

$$\sigma^2 \mathrm{E}f'(W^*) = \mathrm{E}(Wf(W)), \tag{1.10}$$

则称 W^* 是 W 的零偏估计, 并记 $\Delta := W^* - W$. 可以证明零偏估计总是存在, 于是根据 Taylor 展开以及 Stein 方程解的性质, 可以得到如下结论 (参见文献 [11, 命题 2.4]).

命题 1.1 若 h 是绝对连续的, 则

$$|\mathrm{E}h(W) - \mathrm{E}h(Z)| \leqslant \|f_h''\|_\infty \mathrm{E}|\Delta| \leqslant 2\|h'\|_\infty \mathrm{E}|\Delta|. \tag{1.11}$$

最后只需要分析 Δ 的性质以及选择特定的 h 即可得到对应的极限理论. 特别地, 如果 h 选择 Lipschitz 连续函数, W 为归一化部分和, 则 Stein 方法可以得到独立和情形下中心极限定理的 Berry-Esseen 界 (参见文献 [11, 定理 3.6]).

定理 1.3 设 $W_n = \sum_{k=1}^n \xi_k$ 是 n 个独立随机变量的和, 且 $\mathrm{E}\xi_i = 0, \forall i, \mathrm{Var}(W_n) = 1$, 则

$$\sup_x |\Pr(W_n \leqslant x) - \Phi(x)| \leqslant 9.4 \sum_{k=1}^n \mathrm{E}|\xi_k|^3, \tag{1.12}$$

其中 Φ 是标准 Gauss 分布函数.

事实上, 上述界可以改进为

$$\sup_x |\Pr(W_n \leqslant x) - \Phi(x)| \leqslant 4.1 \left[\sum_{k=1}^n \mathrm{E}(|\xi_k|^2 1_{|\xi_k|>1}) + \sum_{k=1}^n \mathrm{E}(|\xi_k|^3 1_{|\xi_k|\leqslant 1}) \right], \tag{1.13}$$

上述界可以推导出 Lindeberg 中心极限定理 1.2 (参见文献 [11, 第 3.2 和 3.4 小节]).

中心极限定理除了前文所述的独立和理论外, 还包括了更多更复杂的模型以及其他衍生理论, 如条件中心极限定理、Poisson 分布收敛定理、中偏差理论、局部收敛定理 (local limit theorem) 以及各式统计量形式的极限理论 (参见文献 [11, 15, 16, 31, 34]). 时至今日, 中心极限定理仍是各领域学者们重点关注的概率论问题.

1.2 熵方法

熵理论最初由 Harry Nyquist 和 Ralph Hartley 于 20 世纪 20 年代建立 (参见文献 [24]), 并在 1948 年由 Shannon 发展成为信息论的奠基理论之一. 在信息论中存在一套独立的熵理论体系, 包括相对熵

及 KL (Kullback-Leibler) 散度相关性质以及熵跳不等式、熵幂不等式和熵连续性等现有理论, 与此同时, 熵作为描述事物所包含的信息量的工具, 与概率论存在着非常紧密的联系, 两者之间也存在着诸多桥梁, 如 Pinsker 不等式 (详见第 3.2 小节). 通过对熵理论的灵活应用, 许多概率论中的经典理论都可以通过熵方法进行证明甚至推广.

熵中心极限定理是熵方法在概率论中的重要应用之一, 也体现了信息论与数学的紧密关系. 熵中心极限定理最早由 Linnik [28] 提出, 随后 Barron [5] 将结论推广至 KL 散度意义下的收敛. Carlen [10] 讨论了熵在密度函数空间的连续性, 给出了非独立情形的极限结果. 在此之后, Johnson [20, 23] 讨论了独立和情形的熵中心极限定理以及收敛速度的问题, 并出版了专著 *Information Theory and the Central Limit Theorem*, 总结了至 21 世纪初为止关于熵中心极限定理的各类研究成果 (参见文献 [21]). Bobkov 等 [6] 提出了熵中心极限定理的 Berry-Esseen 界. 这些经典的熵中心极限定理结果将于第 3 节中做更详细的介绍和讨论.

近年来, 熵中心极限定理仍备受关注. Li 等 [26] 讨论了 Rényi 信息熵下的中心极限定理; Johnson [22] 结合泛函理论, 使用泛函分析理论，对熵跳给出了更精确的估计; Lampros 和 Ioannis [25] 将熵中心极限定理推广至离散随机变量情形. 除中心极限定理外, 熵方法还被应用于大偏差理论, 包括 Cramér 定理、Gärtner-Ellis 定理和 Sanov 定理的证明和推广中. 而在其他数学领域, 如泛函分析、遍历性理论以及图论等, 都有熵方法的身影 (参见文献 [40]).

熵中心极限定理一方面扩充了中心极限定理的证明思路, 是不同于特征函数法和 Stein 方法等数学理论工具的信息论证明方法. 熵中心极限定理不仅进一步拉近了信息论与概率论之间的联系, 也表现出了两学科之间的相互作用所碰撞出的火花. 依靠信息熵理论的工具, 可以进一步得到 KL 散度意义下的收敛定理, 而此结论强于传统的依分布收敛或全变差距离收敛, 即熵方法的使用在一定条件下往往可以得到更强的收敛性.

另外, 正是由于熵中心极限定理主要讨论 KL 散度的收敛, 使得这一极限理论在信息论中存在着很大的应用潜力. 如实数域上的 Hadamard 压缩问题 (参见文献 [32]), 这种压缩方法的极限行为就与熵极限理论密切相关 (可参见第 4 节). 此外, 在其他需要 KL 散度作为度量工具的场景, 如机器学习领域, 熵中心极限定理都有对其模型极限行为进行描述的潜能 (参见文献 [1, 12]).

值得一提的是, 虽然熵方法在中心极限定理中的应用已经较为成熟, 但在其他极限理论中还有待尝试, 如条件分布的中心极限定理, 非求和情形的极限理论. 事实上, 熵方法中重要的关注点是熵增问题, 抓住了特定随机变量序列的熵增性. 在第 2 节中会看到独立和情形下, 序列的熵增性是较为容易验证的, 但独立和并非熵增的本质, 因此继续探究熵方法可以证明中心极限定理的本质原因, 也是进一步推广中心极限定理适用情形的可能思路.

1.3 本文符号与框架

为阅读顺畅, 本小节先整理出一些本文使用的符号. 本文使用大写字母表示随机变量, 如 X 和 Y 等, 使用粗体 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 表示随机向量. 符号 P_X 表示随机变量 X 导出的概率测度. 本文总是用 G 表示标准正态随机变量, 并使用 Φ 和 φ 分别表示其分布函数与密度函数, 具体而言,

$$\Phi(x) = \Pr(G \leqslant x), \tag{1.14}$$

$$\varphi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}. \tag{1.15}$$

本文使用 G_{σ^2} 表示均值为 0、方差为 σ^2 的 Gauss 分布, 并使用 Φ_{σ^2} 和 φ_{σ^2} 分别表示其分布函数与密度函数.

第 2 节对熵理论进行简要的梳理, 从信息熵和气体热力学熵两个角度简要介绍熵的诞生, 并整理出与极限理论相关的一些熵的性质. 第 3 节介绍连接概率论和信息论的 Pinsker 不等式, 并分类展示和描述多种形式的中心极限定理, 整理出熵中心极限定理至今的发展成果以及众多学者的贡献. 第 4 节简要介绍本文作者在条件熵中心极限定理研究方向的一些最新研究成果及其在 Hadamard 压缩问题上的应用.

2　熵理论概述

2.1　信息熵

2.1.1　离散熵

Shannon [36] 在 1948 年从离散随机变量开始对信息进行度量. 假设随机变量 X 的取值空间为可数集 $S=\{x_1,x_2,\ldots\}$, Shannon 认为定义事件 $X=x$ 所包含的信息量 $I(x)$ 应有如下直观性质:

(1) $I(x)\geqslant 0$, 即信息量非负;

(2) 如果 $p(x)=0$, 则 $I(x)=\infty$, 即几乎不可能事件的发生可以提供大量信息;

(3) 如果 $p(x)=1$, 则 $I(x)=0$, 与上一条相反, 几乎必然的事件发生无法提供有效信息;

(4) 如果 $p(x)>p(y)$, 则 $I(x)<I(y)$, 即事件发生概率越小, 能够提供的信息越大;

(5) 如果 $p(x,y)=p(x)p(y)$, 则 $I(x,y)=I(x)+I(y)$, 即两件独立事件提供的信息是各自信息的总和.

根据以上性质, Shannon 提出了信息熵的公理化定义.

定理 2.1 [36]　若信息量 I 满足以上特性, 则 I 形如

$$I(x)=-K\ln p(x), \tag{2.1}$$

其中 K 是正常数.

当 $K=1$ 时, 信息量单位为 nat; 当 $K=\log_2(\mathrm{e})$ 时, 单位为 bit. 随后很自然地, 一个离散随机变量 X 的信息熵定义为所有取值事件信息量的均值, 来表示 X 所携带的信息量.

定义 2.1　设 X 为离散随机变量, 取值空间为可数集 $S=\{x_1,x_2,\ldots\}$, 则 X 的信息熵定义为

$$H(X):=\mathrm{E}I(X)=-\sum_{i=1}^{|S|}p(x_i)\ln p(x_i), \tag{2.2}$$

这里的单位为奈特.

离散熵在信息论中有着奠基作用, 通过对离散熵的公理化定义, Shannon [36] 提出了著名的三大定理, 始终指导着通信技术的发展.

2.1.2　可微熵

由于许多常见的随机变量, 如 Gauss 分布和指数分布等, 都不是离散随机变量, 因此需要定义连续性随机变量的熵, 也被称为可微熵. 由于连续型随机变量有着不可数个的取值, 因此无法直接按照

离散熵的定义进行推广 (其不确定性是 $+\infty$). 事实上, 可微熵的定义方式有很多, 下面给出使用量化方法得到的可微熵定义.

定义 2.2 针对连续性随机变量 X, 定义可微熵 $h(X)$ 表示其量化后的离散随机变量与离散均匀分布随机变量的熵差

$$h(X) := \lim_{n\to\infty} H(X_n) - n\ln 2, \tag{2.3}$$

其中 X_n 为 X 的 2^n 量化, X_n 取值空间为 $\{\frac{k}{2^n}, k\in\mathbb{Z}\}$,

$$\Pr\bigg(X_n = \frac{k}{2^n}\bigg) = \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} p(x)dx.$$

依照上述定义, 不难证明, 如果 X 有密度函数 $p(x)$, $x\in\mathbb{R}$, 则

$$h(X) = -\int_{\mathbb{R}} p(x)\ln p(x)dx, \tag{2.4}$$

其单位为奈特. 拥有可微熵的定义后, 就可以度量 Gauss 分布以及其他连续型随机变量的熵. 而对于两个随机变量 X 和 Y, 熵理论中常用 KL 散度来度量二者的相似度.

定义 2.3 给定两个随机变量 X 和 Y, 定义

$$\mathrm{KL}(X\,\|\,Y) := \int \bigg(\ln\frac{dF_X}{dF_Y}\bigg)dF_X, \tag{2.5}$$

其中 $\frac{dF_X}{dF_Y}$ 为 Radon-Nikodym 导数. 特别地, 如果 X 和 Y 的密度函数分别为 p_X 和 p_Y, 则

$$\mathrm{KL}(X\,\|\,Y) = \int_{\mathbb{R}} p_X(t)\ln\frac{p_X(t)}{p_Y(t)}dt. \tag{2.6}$$

KL 散度熟知的性质如下:

命题 2.1 设 X 和 Y 为两个随机变量, X 有密度函数, 则

$$\mathrm{KL}(X\,\|\,Y) \geqslant 0, \tag{2.7}$$

等号成立当且仅当 $X \overset{d}{=} Y$. 特别地, 若 $Y = G^X$ 是 Gauss 随机变量, 服从 $\mathcal{N}(\mathrm{E}X, \mathrm{Var}(X))$, 则进一步地, 有

$$\mathrm{KL}(X\,\|\,G^X) = h(G^X) - h(X) = \mathrm{KL}\bigg(\frac{X-\mathrm{E}X}{\sqrt{\mathrm{Var}(X)}}\bigg\|\,G\bigg) = h(G) - h\bigg(\frac{X-\mathrm{E}X}{\sqrt{\mathrm{Var}(X)}}\bigg), \tag{2.8}$$

其中 G 是标准 Gauss 分布.

由于熵 h 有平移不变性 $h(X+a) = h(X)$, $\forall\, a\in\mathbb{R}$, 因此在给定方差的条件下, Gauss 分布的熵是最大的, 即如下定理.

命题 2.2 给定连续随机变量 X 的方差为 σ^2, 则

$$h(X) \leqslant h(G_{\sigma^2}) = \frac{1}{2}\ln(2\pi \mathrm{e}\sigma^2). \tag{2.9}$$

为记号简便, 下文使用 $D(X) := \mathrm{KL}(X\,\|\,G^X)$. 我们将看到, 独立和形式的中心极限定理在熵理论框架下就是在描述随机变量序列不断熵增从而达到最大熵, 即达到 Gauss 分布这一过程.

2.2 物理熵

1850 年, 德国物理学家 Clausius 使用熵描述热力学系统的混乱度, 提出了著名的热力学第二定律:

$$dS \geqslant \frac{dQ}{T}, \tag{2.10}$$

其中 dS 为熵的变化量, dQ 为加入系统的热量, T 表示系统的绝对温度. 显然对于孤立系统, $dQ=0$, 因此热力学第二定律也可以表示为 $dS \geqslant 0$, 即孤立系统的热力学过程一定是熵增过程. 熵增规律甚至可以体现在生命体系中, 奥地利物理学家 Schrödinger [35] 于 1944 年出版的著作 *What is Life* 提出了: 生命体中熵增是必然的, 生命体不断从有序走回无序, 最终不可逆地走向老化和死亡.

为研究理想气体, 奥地利物理学家 Boltzmann 引入了 H- 量

$$H := \int f(x,v,t)\lg f(x,v,t)dxdv, \tag{2.11}$$

其中 x 表示位置, v 表示速度, t 表示时间, f 是理想气体的分布密度函数, lg 表示以 10 为底的对数. Boltzmann 运用统计力学的模型求得了理想气体的平衡态分布, 并对比了热力学熵 S 与 H- 量的计算结果, 得到关系式

$$S - S_0 = -\kappa H, \tag{2.12}$$

其中 S_0 是某个固定状态的热力学熵值, κ 称为 Boltzmann 常数 (参见文献 [27]). 观察 (2.4)、(2.11) 和 (2.12), 可以发现 3 种熵量 S、H 和 h 具有统一性.

热力学第二定律说明自然总是熵增的, 熵也被称为时间之矢 (参见文献 [27]). 那么很自然地, 物理系统的熵增过程最后会趋向于如何的状态, 是很多物理学家关心的问题. McKean [30] 讨论了 Kac 气体分子运动速度分布问题. 假设无穷多个气体分子在一个系统中发生碰撞, 每个分子的初始速度都是随机的且独立同分布的, 分布密度为 f, 平均动能为

$$\frac{\sigma^2}{2} = \int \left(\frac{v^2}{2}\right) f(v)dv.$$

假设碰撞过程中能量守恒, 且每个气体分子速度始终保持独立同分布, 则 t 时刻气体速度分布的密度函数 $u(t)$ 是下列 Boltzmann 问题的解:

$$\frac{\partial u(t,a)}{\partial t} = \int [u(t, a\cos\theta - b\sin\theta)u(t, a\sin\theta + b\cos\theta) - u(t,a)u(t,b)]db\frac{d\theta}{2\pi} =: B[u], \tag{2.13}$$

且初始速度分布 $u(0)=f$.

McKean [30] 研究了极限 $\lim_{t\to\infty} u(t)$, 即时间趋向于无穷后, 气体运动速度呈现怎样的分布. 首先, 根据热力学第二定律, 孤立系统是熵增过程, 在上述气体运动问题中反映为 Boltzmann H- 定理.

定理 2.2 (Boltzmann H- 定理) 速度分布的熵 $h(u(t))$ 关于 t 是递增函数.

而 McKean [30] 通过更精细的分析, 得到了速度分布的极限定理:

定理 2.3 设 φ_{σ^2} 是 Gauss 分布 $\mathcal{N}(0,\sigma^2)$ 的密度函数, 则

$$(4\mathrm{e})^{-1}\|u - \varphi_{\sigma^2}\|_1^2 \leqslant h(\varphi_{\sigma^2}) - h(u) \leqslant ct^{3/2}\delta(t), \tag{2.14}$$

其中 $\|\cdot\|_1$ 表示 L^1 范数, $c=c(f)$ 是只与 f 有关的常数,

$$\delta(t) = \exp\left(\frac{2}{9}\left(\frac{8}{3\pi} - 1\right)t\right).$$

如果 f 有有限三阶矩 $\int |v^3| f(v)dv < \infty$, 则

$$\left| \int gu - \int g\varphi_{\sigma^2} \right| \leqslant \|g\|_\infty \mathrm{e}^{-t} + C\|g'''\|_\infty \mathrm{e}^{(8/3\pi-1)t}, \quad \forall\, g \in C^3(\mathbb{R}), \tag{2.15}$$

其中 $\|\cdot\|_\infty$ 为 L^∞ 范数, $C = C(f)$ 是只与 f 有关的常数.

从上述定理可以看出, Kac 气体的运动最终将趋于最大熵的混乱状态, 即 Gauss 分布. 虽然这里的气体运动模型和研究方法与中心极限定理存在一些不同, 但最后的收敛极限均为 Gauss 分布, 说明熵增现象与趋向于 Gauss 分布的行为, 是在多个问题和领域中均存在的普遍规律.

2.3 熵不等式

熵理论中存在诸多重要的不等式, 其中熵指数不等式 [13] 和熵跳不等式 [10] 与本文的关系最为密切.

定理 2.4 (熵指数不等式) 设 X 和 Y 是两个独立的连续随机变量, 则

$$\mathrm{e}^{2h(X+Y)} \geqslant \mathrm{e}^{2h(X)} + \mathrm{e}^{2h(Y)}, \tag{2.16}$$

等号成立当且仅当 X 和 Y 是方差相同的 Gauss 分布.

由于函数 $f(x) = \mathrm{e}^x$ 是凸函数, 所以容易得到如下推论:

定理 2.5 (熵跳不等式) 设 X 和 Y 是两个独立的连续随机变量, 则

$$h(\lambda X + \sqrt{1-\lambda^2}Y) \geqslant \lambda^2 h(X) + (1-\lambda^2)h(Y), \quad \forall\, \lambda \in [0,1], \tag{2.17}$$

等号成立当且仅当 X 和 Y 是方差相同的 Gauss 分布.

特别地, 当 $\lambda = 1/\sqrt{2}$ 时, 有

$$h\left(\frac{X+Y}{\sqrt{2}}\right) - \frac{h(X)+h(Y)}{2} \geqslant 0. \tag{2.18}$$

(2.18) 的提出最早可追溯到 1948 年 Shannon [36] 的信息论奠基文章, 也是 Barron [5] 证明其熵中心极限定理的出发点.

记熵增为 $\delta := h(\frac{X+Y}{\sqrt{2}}) - \frac{h(X)+h(Y)}{2}$, 对其越精确的估计可以得到越精确的极限理论. Carlen 和 Soffer [10] 在 1991 年讨论了 δ 关于 KL 散度 $D(X)$ 和 $D(Y)$ 的一致性, 从而得到了非独立情形的中心极限定理. Johnson [23] 在 2004 年给出了 δ 更具体的下界, 从而给出了熵中心极限定理在一定条件下的收敛速度 (详见定理 3.5). 由此可见, 在使用熵方法证明中心极限定理时, 熵不等式扮演着重要的角色.

值得一提的是, Stam [37] 在 1959 年对熵指数不等式 (定理 2.4) 进行证明时, 发现了下列不等式 (参见文献 [13]):

$$\frac{1}{2\pi \mathrm{e}} \mathrm{e}^{2h(X)} J(X) \geqslant 1, \tag{2.19}$$

其中 X 有密度函数 p_X 且二阶矩有限,

$$J(X) := \mathrm{E}\left(\frac{p_X'(X)}{p_X(X)}\right)^2 = \int_{\mathbb{R}} \frac{(p_X'(x))^2}{p_X(x)} dx \tag{2.20}$$

称为 X 的 Fisher 信息量. Fisher 信息量是概率和统计理论中的一个重要参量, (2.20) 是它的一种特殊形式. Fisher 信息量在熵理论中拥有重要的地位, 不论是 McKean 对 Kac 气体的讨论, 还是下一节介绍的诸多熵中心极限定理, Fisher 信息量都发挥着不可替代的作用. 对不等式 (2.19) 两边取对数, 并结合命题 2.2, 可以得到熵的上下界.

命题 2.3　设 X 是连续随机变量, 且有绝对连续的密度函数和有限二阶矩, 则

$$\frac{1}{2}\ln\left(2\pi\mathrm{e}\frac{1}{J(X)}\right)=h\left(\mathcal{N}\left(0,\frac{1}{J(X)}\right)\right)\leqslant h(X)\leqslant h(\mathcal{N}(0,\mathrm{Var}(X)))=\frac{1}{2}\ln(2\pi\mathrm{e}\mathrm{Var}(X)). \tag{2.21}$$

于是很自然地可以得到著名的 Cramér-Rao (CR) 下界 (可参见文献 [5,21]).

推论 2.1 (CR 下界)　设 X 是连续随机变量, 且有绝对连续的密度函数和有限二阶矩, 则

$$J(X)\mathrm{Var}(X)\geqslant 1. \tag{2.22}$$

这也是使用熵方法得到经典理论的实例之一. 需要说明的是, 推论 2.1 形式上与统计学中的 CR 下界很相似, 但也存在微妙的差异.

3　熵中心极限定理概述

3.1　最初奠基人—Linnik

Linnik [28] 是最早采用信息论方法研究中心极限定理的学者. 1959 年 Linnik 考虑独立随机变量列 $X_1,X_2,\dots$, 并对该列随机变量作截断 $X_i':=X_i1_{|X_i|\leqslant T_0\sqrt{B_i}}$, 其中 $B_i:=\sum_{k=1}^{i}\mathrm{Var}(X_k)$, Gauss 光滑化 $X_i'':=X_i'+G_{\delta_0^2\mathrm{Var}(X_i')}-\mathrm{E}X_i'$, 其中 δ_0 为充分小的常数, 从而使得该列随机变量可以定义熵且拥有其他良好的正则性. 针对处理过的随机变量列 $X_1'',X_2'',\dots$, 定义归一化求和

$$W_m:=\frac{\sum_{k=1}^{m}X_k''}{\sqrt{\sum_{k=1}^{m}\mathrm{Var}(X_k'')}}. \tag{3.1}$$

Linnik 通过定义函数 (后来被称为 Linnik 函数 [30])

$$I(X)=-\left(\int_{\mathbb{R}}p(x)\ln p(x)dx+\frac{1}{2}\ln\int_{\mathbb{R}}x^2p(x)dx\right) \tag{3.2}$$

来将熵理论引入中心极限定理中, 其中 $p(x)$ 为 X 的密度函数. 事实上, Linnik 试图给出增量 $I(W_{m+1})-I(W_m)$. 为此, Linnik 给出了个体 (individually) Lindeberg 条件:

假设 3.1 (个体 Lindeberg 条件)　给定 X_j'' 和数组 $(\tau,\xi)\in(\mathbb{R}^+)^2$, 若

$$\int_{|x|>\tau}x^2\mathrm{Pr}_{X_j''}(dx)<\frac{\xi\mathrm{Var}(X_j'')}{B_j}, \tag{3.3}$$

则称 X_j'' 满足 (τ,ξ) 个体 Lindeberg 条件.

如此 Linnik 得到了增量表达式

$$I(W_{m+1})-I(W_m)=\frac{\mathrm{Var}(X_{m+1}'')}{\sum_{k=1}^{m}\mathrm{Var}(X_k'')}\left(\frac{1}{2}[J(W_m)-1]+B(\tau+\mu)\right), \tag{3.4}$$

其中 $J(\cdot)$ 为 Fisher 信息量, B 是一个有界常数, 即 $|B|<C<\infty, C$ 是正常数. 如果 X''_{m+1} 满足 (τ,ξ) 个体 Lindeberg 条件, 则 $\mu=\xi$, 否则 $\mu=1$. 使用 CR 下界 (2.22) 并整理 (3.4), 可得

$$0 \leqslant J(W_m)-1 \leqslant 2\frac{\sum_{k=1}^{m}\mathrm{Var}(X''_k)}{\mathrm{Var}(X''_{m+1})}(I(W_{m+1})-I(W_m))+2C(\tau+\mu) \tag{3.5}$$

以及

$$I(W_{m+1})-I(W_m) \geqslant 0, \tag{3.6}$$

这也是 Linnik 称为 “增量” 的原因. 由于 $\mathrm{Var}(W_m)=1, \forall m$, 根据命题 2.2 可知

$$\lim_{m\to\infty} I(W_{m+1})-I(W_m)=0,$$

进一步地, 如果存在两个正实数列 $\tau_m\to 0$ 和 $\xi_m\to 0$, $m\to\infty$, 使得 X''_m 满足 (τ_m,ξ_m) 个体 Lindeberg 条件, 并且

$$\frac{\sum_{k=1}^{m}\mathrm{Var}(X''_k)}{\mathrm{Var}(X''_{m+1})}=o\left(\frac{1}{I(W_{m+1})-I(W_m)}\right), \quad m\to\infty,$$

则可以得到 $J(W_m)-1\to 0$, $m\to\infty$. 文献 [28] 使用了如下引理作为证明中心极限定理的最后一步.

引理 3.1 [28] 设 p_m 为 W_m 的密度函数, 则对于任意 $\epsilon>0$, 存在 η, 使得若 $J(W_m)-1<\eta$, 则

$$\sup_x |p_m(x)-\varphi(x)|<\epsilon, \tag{3.7}$$

其中 φ 是标准 Gauss 分布的密度函数.

于是 Linnik 证明了如下中心极限定理.

定理 3.1 [28] 设 $X_1, X_2, \ldots$ 为一列独立随机变量, 形如 $X_i=\hat{X}_i+G_{\sigma_0^2}$, 其中 $\hat{X}_i$ 为一致有界的随机变量, $\mathrm{E}\hat{X}_i=0, \forall i$, σ_0 充分小. 记

$$W_m := \frac{\sum_{k=1}^{m}X_k}{\sqrt{\sum_{k=1}^{m}\mathrm{Var}(X_k)}}. \tag{3.8}$$

若这一列随机变量满足下列条件:

(1) 存在 $M_0>0$, 以及两个正实数列 $\tau_m\to 0$ 和 $\xi_m\to 0$, $m\to\infty$, 使得 X_m 满足 (τ_m,ξ_m) 个体 Lindeberg 条件, $\forall m\geqslant M_0$;

(2) $\frac{\sum_{k=1}^{m}\mathrm{Var}(X_k)}{\mathrm{Var}(X_{m+1})}=o(\frac{1}{I(W_{m+1})-I(W_m)})$, $m\to\infty$,

则

$$\lim_{m\to\infty}\sup_x |p_m(x)-\varphi(x)|=0, \tag{3.9}$$

其中 p_m 为 W_m 的密度函数, φ 是标准 Gauss 的密度函数.

Linnik 通过使用熵与 Fisher 信息量, 得到了最初的熵中心极限定理, 为后续熵方法在极限理论的应用奠定了基础. 特别地, Linnik 使用的 Gauss 分量假设 $X_i=\hat{X}_i+G_{\sigma_0^2}$ 被后续多篇文献采用, 可以很好地简化定理证明, 且与热力学熵模型存在联系 (可参见文献 [30]).

3.2 Pinsker 不等式

Linnik 给出的熵中心极限定理最终回归到了 L^∞ 收敛 (3.9), 但随着更多信息论方法的引入, 后续的学者和文章更多关注于 KL 散度形式的收敛. 此时, Pinsker 不等式 [14] 便是一个颇有意义的结论, 它成为了连结信息论与传统概率论的桥梁.

定理 3.2 (Pinsker 不等式) 给定两个定义在 σ- 代数 $\mathcal{F}$ 上的概率测度 P 和 Q, 则

$$\mathrm{KL}(P\|Q) \geqslant \frac{1}{2}\|P-Q\|_{TV}^2, \tag{3.10}$$

其中

$$\mathrm{KL}(P\|Q) := \int_\Omega \ln\left(\frac{dP}{dQ}\right) dP, \tag{3.11}$$

$$\|P-Q\|_{TV} := 2\sup_{A\in\mathcal{F}} |P(A)-Q(A)|. \tag{3.12}$$

注 3.1 如果 P 和 Q 是两个定义在可数集上 B 的离散概率测度, 则

$$\|P-Q\|_{TV} = 2\sup_{A\in\mathcal{F}} |P(A)-Q(A)| = \sum_{x\in B} |P(x)-Q(x)|. \tag{3.13}$$

如果 P 和 Q 是两个有密度函数的概率测度, 密度函数分别为 p 和 q, 则

$$\|P-Q\|_{TV} = 2\sup_{A\in\mathcal{F}} |P(A)-Q(A)| = \int_{\mathbb{R}} |p(x)-q(x)|dx = \|p-q\|_1. \tag{3.14}$$

信息论中许多结论都围绕熵或 KL 散度, 与传统概率论中度量分布距离的手段存在不同. Pinsker 不等式阐明了 KL 散度可以控制全变差的这一事实, 即如果得到 KL 散度收敛, 即可得到全变差的收敛性, 从而进一步得到弱收敛等其他收敛性. 这一性质使得熵方法和概率论的关系变得更加紧密, 如此, 更多的熵工具可以被考虑使用, 熵中心极限定理的目标也可以转化为对 KL 散度收敛性的证明.

3.3 i.i.d. 求和的熵中心极限定理

一个理论的发展往往从易到难, 熵中心极限定理也是如此, 最初的理论主要考虑的是独立同分布 (i.i.d.) 的求和形式, 即考虑随机变量列

$$W_n := \frac{1}{\sigma\sqrt{n}}\sum_{k=1}^n X_k, \quad n\in\mathbb{N}, \tag{3.15}$$

其中 $X_1, X_2, \ldots,$ 是 i.i.d. 序列, $\mathrm{E}X_1 = 0$, $\mathrm{Var}(X_1) = \sigma^2$.

3.3.1 极限理论

1986 年 Barron [5] 通过总结 1959 年 Stam [37] 以及 1982 年 Brown [9] 关于熵与 Fisher 信息量的研究, 建立了 Fisher 信息量与 KL 散度两种意义下的中心极限定理. Barron 首先根据 Fisher 信息量在正交变换下不增的性质, 即对于任意独立随机变量 X_1 和 X_2, 成立

$$J(\sqrt{\alpha_1}X_1+\sqrt{\alpha_2}X_2) \leqslant \alpha_1 J(X_1)+\alpha_2 J(X_2), \quad \forall\, \alpha_1,\alpha_2\geqslant 0, \quad \alpha_1+\alpha_2=1, \tag{3.16}$$

得到 Fisher 信息量意义下的极限结果.

定理 3.3 [5]　考虑 W_n 如 (3.15), 记 $W_n^t := \sqrt{t}W_n + \sqrt{1-t}G$, $0 \leqslant t < 1$, 则

$$\lim_{n\to\infty} J(W_n^t) = 1, \quad \forall t \in [0,1). \tag{3.17}$$

如下引理建立了 KL 散度与 Fisher 信息量的密切关系:

引理 3.2 (熵是 Fisher 信息量的积分 [5])　设 X 为连续随机变量, 方差为 σ^2, 若记 $G^X \sim \mathcal{N}(\mathrm{E}X, \sigma^2)$, 则

$$D(X) := \mathrm{KL}(X\|G^X) = \int_0^1 [\sigma^2 J(\sqrt{t}X + \sqrt{1-t}G^X) - 1]\frac{dt}{2t}. \tag{3.18}$$

利用单调收敛定理和 $nD(W_n)$ 的次可加性, Barron [5] 给出了 KL 散度下的收敛定理.

定理 3.4 [5]　考虑 W_n 如 (3.15), 则

$$\lim_{n\to\infty} D(W_n) = 0 \tag{3.19}$$

当且仅当存在某个 n_0, 使得 $D(W_{n_0}) < \infty$.

定理 3.4 是首个给出 KL 散度收敛性的结论, 为后续的熵中心极限定理研究方向奠定了一定的基础. 而文献 [5] 对熵与 Fisher 信息量的关系总结 (如引理 3.2), 也为后续研究中引入 Ornstein-Uhlenbeck 算子等方法打下了铺垫.

3.3.2　收敛速率理论

在 Barron 给出了 i.i.d. 情形下的收敛定理后, 考虑其收敛速度是一个很自然的问题. 文献 [4, 23] 分别使用了不同的方法得到了在 Poincaré 不等式成立的情形下熵增的精确估计, 并因此得到了 i.i.d. 情形下熵中心极限定理的收敛速度.

Johnson 和 Barron [23] 使用概率空间投影的方法, 建立了如下引理:

引理 3.3 [23]　设 Y_1 和 Y_2 是两个独立随机变量, 且有绝对连续的密度函数, 以及有限约束的 Poincaré 常数 R_1^* 和 R_2^*:

$$R_i^* := \sup_{g\in H_1^*(Y_i)} \frac{\mathrm{E}g^2(Y_i)}{\mathrm{E}g'(Y_i)^2} < \infty, \quad i = 1, 2, \tag{3.20}$$

其中 $H_1^*(Y_i) := \{g \in C_{ab}(\mathbb{R}) : 0 < \mathrm{Var}(g(Y_i)) < \infty, \mathrm{E}g(Y_i) = \mathrm{E}g'(Y_i) = 0\}$, C_{ab} 为所有绝对连续函数构成的集合.

考虑函数 f 满足 $\mathrm{E}f(Y_1+Y_2)^2 < \infty$, $\mathrm{E}f(Y_1+Y_2) = 0$, 并记投影 $g_1(u) := \mathrm{E}_{Y_2}f(u+Y_2)$, $g_2(v) = \mathrm{E}_{Y_1}f(Y_1+v)$, 则存在常数 μ、ν_1 和 ν_2, 使得对于任意 $\beta \in [0,1]$, 都有

$$\begin{aligned}&\mathrm{E}(f(Y_1+Y_2) - g_1(Y_1) - g_2(Y_2))^2\\ &\quad\geqslant \frac{1}{(1-\beta)J(Y_1)+\beta J(Y_2)}\left(\frac{\beta}{R_1^*}\mathrm{E}(g_1(Y_1)-\mu Y_1-\nu_1)^2 + \frac{1-\beta}{R_2^*}\mathrm{E}(g_2(Y_2)-\mu Y_2-\nu_2)^2\right).\end{aligned} \tag{3.21}$$

由此得到了 Fisher 信息量在正交变化下更精确的估计.

命题 3.1 [23]　设 Y_1 和 Y_2 独立同分布, 有绝对连续的密度函数, 方差为 σ^2. 若它们有有限约束的 Poincaré 常数 R_1^*, 则

$$\sigma^2 J\left(\frac{Y_1+Y_2}{\sqrt{2}}\right) - 1 \leqslant (\sigma^2 J(Y_1) - 1)\left(\frac{2R_1^*}{\sigma^2 + 2R_1^*}\right). \tag{3.22}$$

通过迭代计算, 命题 3.1 能够证明定理 3.3 在 $t=1$ 时仍然成立, 并给出如下收敛速度.

定理 3.5 [23]　设 $X_1, X_2, \ldots$ 为 i.i.d. 的随机变量列, 有绝对连续的密度函数, 方差为 σ^2, 且有有限约束的 Poincaré 常数 R^*. 记

$$W_n := \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i - \mathrm{E}X_i, \tag{3.23}$$

则

$$0 \leqslant \sigma^2 J(W_n) - 1 \leqslant \frac{2R^*}{2R^* + (n-1)\sigma^2}(\sigma^2 J(X_1) - 1), \quad \forall\, n. \tag{3.24}$$

进一步地, 结合引理 3.2, 文献 [23] 得到了 KL 散度的收敛速度.

定理 3.6 [23]　设 $X_1, X_2, \ldots$ 为 i.i.d. 的随机变量列, 有绝对连续的密度函数, 方差为 σ^2, 且有有限 Poincaré 常数 R:

$$R := \sup_{g \in H_1(X_1)} \frac{\mathrm{E}g^2(X_1)}{\mathrm{E}g'(X_1)^2} < \infty, \tag{3.25}$$

其中 $H_1(X_1) := \{g \in C_{ab}(\mathbb{R}) : 0 < \mathrm{Var}(g(X_1)) < \infty, \mathrm{E}g(X_1) = 0\}$, C_{ab} 为所有绝对连续函数构成的集合. 记

$$W_n := \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i - \mathrm{E}X_i, \tag{3.26}$$

则

$$0 \leqslant D(W_n) \leqslant \frac{2R}{2R + (n-1)\sigma^2} D(X_1), \quad \forall\, n. \tag{3.27}$$

值得一提的是, 引理 3.1 本身并不要求同分布, 因此通过更细致的分析和计算, 可以将文献 [23] 所得到的熵跳估计推广至独立不同分布的情形.

定理 3.7 [39]　设两个独立随机变量 X 和 Y 有绝对连续的密度函数, 方差分别为 σ_X^2 和 σ_Y^2. 若存在常数 R, 使得

$$\max\{R_X^*, R_Y^*, R_G^*\} \leqslant R, \tag{3.28}$$

则存在常数 $c = c(\sigma_X^2, \sigma_Y^2, R)$, 使得

$$c\left(\frac{1}{2}\ln\left(2\pi\mathrm{e}\frac{\sigma_X^2 + \sigma_Y^2}{2}\right) - \frac{h(X) + h(Y)}{2}\right) \leqslant h\left(\frac{X+Y}{\sqrt{2}}\right) - \frac{h(X) + h(Y)}{2},$$

等号成立当且仅当 X 和 Y 是同方差的 Gauss 分布.

此结论可用于证明一般独立和的熵中心极限定理及其收敛速度 (详见文献 [39]).

2004 年, Artstein 等 [4] 也在有限 Poincaré 常数 R 的条件下, 使用全变差不等式对 KL 散度收敛速度进行证明.

定理 3.8 (全变差不等式 [4])　设 $\omega : \mathbb{R}^2 \to (0, \infty)$ 为密度函数, 且

$$\int \frac{|\nabla\omega|^2}{\omega} < \infty, \quad \int |\partial_{xx}\omega| < \infty. \tag{3.29}$$

令 h 为边际概率, 即 $h(x) := \int \omega(x,y)dy$, 则

$$J(h) \leqslant \int \omega \cdot (\partial_y p)^2 dxdy + \int \omega \cdot (1,p)\mathrm{Hess}(-\ln \omega)(1,p)^{\mathrm{T}} dxdy, \tag{3.30}$$

其中 $p:\mathbb{R}^2 \to \mathbb{R}$ 是任意可微函数, 满足

$$\frac{p\partial_x \omega \partial_y \omega}{\omega} \in L^1(\mathbb{R}^2), \quad p\partial_{xy}\omega \in L^1(\mathbb{R}^2), \quad p^2 \partial_{yy}\omega \in L^1(\mathbb{R}^2). \tag{3.31}$$

依据上述不等式, Ball 等 [4] 得到了更一般线性组合下的熵中心极限定理收敛速度.

定理 3.9 [4] 设 $X_1, X_2, \ldots$ 为 i.i.d. 的随机变量列, 有绝对连续的密度函数, 方差为 1, 且有有限 Poincaré 常数 R. 任意给定 $a \in \mathbb{R}^n$, $\sum_i a_i^2 = 1$, 记

$$W_n^a := \sum_{i=1}^n a_i(X_i - \mathrm{E}X_i), \tag{3.32}$$

则

$$0 \leqslant D(W_n) \leqslant \frac{\sum_{i=1}^n a_i^4}{R/2 + (1-R/2)\sum_{i=1}^n a_i^4} D(X_1), \quad \forall n. \tag{3.33}$$

文献 [4, 23] 完成了 i.i.d. 情形的熵中心极限定理速度的证明, 均得到了在有限 Poincaré 常数的条件下, $D(W_n) = O(\frac{1}{n})$ $(n \to \infty)$ 的收敛速率. 虽然一些常见的分布, 如 Gauss 分布或均匀分布, 都满足有限 Poincaré 常数的条件, 但这个条件本身在验证时并不容易. 现有一些文献专注于讨论 Poincaré 常数的性质, 例如, 在卷积意义下不增等特征, 相关理论也与概率空间上的 Sobolev 不等式有着密切联系, 详细内容可以参见文献 [8, 17, 33] 等. 结合一般测度空间的 Sobolev 嵌入等理论, 进一步研究如何满足有限 Poincaré 常数条件, 是推广文献 [4, 23] 结论的可能方案, 使之能够给出非同分布的独立和, 甚至更复杂情形下的收敛速度.

3.4 熵中心极限定理的 Berry-Esseen 界

非同分布的独立和极限理论是中心极限定理的一大议题, 在 Barron 和 Ball 等学者研究了 i.i.d. 情形后, 2000 年 Johnson 研究了独立和情形, 并且求和不仅限于按顺序的 $S_n = \sum_{k=1}^n X_k$, 而且推广至 $U^{(S)} = \sum_{i\in S} X_i$, 其中 $S \subset \mathbb{N}$, $X_i = (X_0)_i + G_{\sigma^2}$ 满足 Gauss 分量假设 (参见文献 [20]).

Bobkov 等 [7] 对更一般的独立和问题进行了更深入的探究, 并给出了熵中心极限定理的 Berry-Esseen 界. 考虑独立随机变量列 $X_1, X_2, \ldots, X_n$, 满足 $D(X_k) \leqslant D < \infty, 1 \leqslant k \leqslant n$, 即这些随机变量与 Gauss 分布的 KL 散度存在一致的上界. 在这种情形下, 文献 [7] 分别给出了全变差距离 $\|F_n - \Phi\|_{\mathrm{TV}}$ 和 KL 散度 $D(F_n)$ 的收敛速度, 其中 F_n 为随机变量

$$W_n := \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{\sum_{k=1}^n \mathrm{Var}(X_k)}} \tag{3.34}$$

的分布函数, Φ 为标准 Gauss 分布的分布函数, 任意两个分布 F_1 和 F_2 的全变差定义为

$$\|F_1 - F_2\|_{\mathrm{TV}} := 2 \sup_{A\in\mathcal{B}(\mathbb{R})} \left| \int_A dF_1 - \int_A dF_2 \right|. \tag{3.35}$$

将文献 [7] 的结论总结为如下定理:

定理 3.10　设 $X_1, X_2, \ldots, X_n$ 为一列独立随机变量, 满足 $D(X_k) \leqslant D < \infty$, $1 \leqslant k \leqslant n$. 若每个 X_i 都有有限三阶矩, 则

$$\|F_n - \Phi\|_{\mathrm{TV}} \leqslant C_1(D) \frac{\sum_{k=1}^n \mathrm{E}|X_k|^3}{(\sum_{k=1}^n \mathrm{Var}(X_k))^{3/2}}, \tag{3.36}$$

其中 $C_1(D)$ 是只与 D 相关的正常数. 特别地, 在 i.i.d. 情形, 可得

$$\|F_n - \Phi\|_{\mathrm{TV}} = O\left(\frac{1}{\sqrt{n}}\right), \quad n \to \infty. \tag{3.37}$$

若每个 X_i 都有有限四阶矩, 则

$$D(W_n) \leqslant C_2 \mathrm{e}^{62D} \frac{\sum_{k=1}^n \mathrm{E}|X_k|^4}{(\sum_{k=1}^n \mathrm{Var}(X_k))^2}, \tag{3.38}$$

其中 C_2 为绝对常数. 特别地, 在 i.i.d. 情形, 可得

$$D(W_n) = O\left(\frac{1}{n}\right), \quad n \to \infty. \tag{3.39}$$

为证明定理 3.10, 文献 [7] 对卷积进行了分解. 首先将若干个 X_i 结合:

$$\begin{aligned} S_n &:= X_1 + X_2 + \cdots + X_n \\ &= \left(\sum_{k=1}^{n_1} X_k\right) + \left(\sum_{k=n_1+1}^{n_2} X_k\right) + \cdots + \left(\sum_{k=n_{N-1}+1}^{n} X_k\right) \\ &=: V_1 + V_2 + \cdots + V_N, \end{aligned} \tag{3.40}$$

使得 V_i 的方差尽可能接近, 并记 V_i 的密度函数为 ρ_i. 分解卷积

$$\begin{aligned} p_n &= \rho_1 * \rho_2 * \cdots * \rho_N \\ &= 2^{-N} \sum (\rho_{10}^{\delta_1} * \rho_{11}^{1-\delta_1}) * \cdots * (\rho_{N0}^{\delta_N} * \rho_{N1}^{1-\delta_N}), \end{aligned} \tag{3.41}$$

其中 p_n 为 S_n 的密度函数, ρ_{ij}, $i = 1, 2, \ldots, N$, $j = 1, 2$ 是密度函数 ρ_i 的分解 (详见文献 [7, 定义 4.2]),

$$\rho_{i0}^{\delta_i} * \rho_{i1}^{1-\delta_i} := \begin{cases} \rho_{i0}, & \delta_i = 1, \\ \rho_{i1}, & \delta_i = 0. \end{cases} \tag{3.42}$$

从而将 p_n 分解为 (详见文献 [7, 定义 8.1])

$$p_n = (1 - \varepsilon_n) p_{n0} + \varepsilon_n p_{n1}, \tag{3.43}$$

且 $\varepsilon_n \leqslant 2^{-N+1} N_0^m$, m_0 为给定常数.

(3.36) 和 (3.38) 的证明是熵方法与特征函数的结合, 其关键在于如下命题所展示的二者关系:

命题 3.2　设 X 为连续随机变量, 方差为 σ^2, 则对于任意 $t \in \mathbb{R}$, 都有

$$|f(t)| \leqslant 1 - c \min\{1, \sigma^2 t^2\} \mathrm{e}^{-4D(X)}. \tag{3.44}$$

由于全变差距离可以由特征函数差的 L^2 范数以及特征函数导数差的 L^2 范数控制, 文献 [7] 使用截断的方法, 将上述 L^2 积分分解为 $[-\sqrt{N},\sqrt{N}]$ 以及外部, 并综合运用 Gauss 随机变量尾分布性质、命题 3.2 以及其他数学方法, 分别估计了特征函数差和特征函数导数差在这两个区域上的上界, 并根据全变差距离与特征函数 L^2 范数的关系获得 (3.36). 针对 (3.38), 文献 [7] 将 $D(W_n)=D(S_n)$ 转化为 $D(p_{n0})$, 原因在于下面两个不等式:

$$D(p_{n0}) \leqslant \alpha^2 + 4(\|f_{n0}-g_\alpha\|_2 + \|f_{n0}'''-g_\alpha'''\|_2) \tag{3.45}$$

和

$$|D(p_{n0}) - D(W_n)| \leqslant 2^{-N+6}N^4(D+1), \tag{3.46}$$

其中 f_{n0} 为 p_{n0} 的特征函数, $g_\alpha(t):=g^{p_{n0}}(t)(1+\alpha\frac{(\mathrm{i}t)^3}{3!})$, $\alpha=\sum_{k=1}^n \mathrm{E}X_k^3$, $g^{f_{n0}}(t)$ 为与 p_{n0} 同均值同方差的 Gauss 分布的密度函数, D 为 $D(X_k)$ 的一致上界. 针对项 $\|f_{n0}-g_\alpha\|_2+\|f_{n0}'''-g_\alpha'''\|_2$, 运用与证明 (3.36) 时相同的截断方法可以得到其上界估计, 再结合 (3.45) 和 (3.46), 可以得到最终结论.

关于文中始终假设的 KL 散度一致上界 $D(X_i)\leqslant D$, 由文献 [13] 可得 $D(X)\leqslant \frac{1}{2}\ln(\mathrm{Var}(X)J(X))$, 因此 $D(X)$ 的一致有界条件可由一致有界的方差和 Fisher 信息量控制.

文献 [7] 很好地结合了熵理论的知识, 包括熵增不等式等以及概率论中的特征函数法和卷积分解, 得到了在相对一般的条件下熵中心极限定理的收敛速度. 证明中所采用的步骤和方法在很大程度上展示了信息论和概率论的和谐统一. 定理 3.10 可以被应用于证明传统概率论中的 Berry-Esseen 界和条件中心极限定理的收敛速度等问题, 拥有广泛的理论价值, 可以说, 正是两个学科的联合, 才能诞生出这一具有出色适用性的结论.

3.5 非独立分布求和的结论

Carlen 和 Soffer [10] 使用 Ornstein-Uhlenbeck (O-U) 算子对熵关于 KL 散度的连续性 (在 Gauss 分布处) 进行了证明. 具体而言, 考虑 O-U 变换 $P_t^*X:=\mathrm{e}^{-t}X+(1-\mathrm{e}^{-2t})^{1/2}G^X$, $t>0$, 其中 G^X 是与 X 同均值同方差的 Gauss 分布. Carlen 和 Soffer [10] 证明了 O-U 变换的关键性质:

命题 3.3 [10] 设 $X_1,X_2,\dots,X_n$ 是独立随机变量, 方差均为 1, 则对于任意 $0<\lambda_1,\lambda_2,\dots,\lambda_n\leqslant 1$, $\sum_{k=1}^n \lambda_k^2=1$, 都有

$$0 \leqslant h\left(\sum_{k=1}^n \lambda_k P_t^* X_k\right) - \sum_{k=1}^n \lambda_k^2 h(P_t^*X_k) \leqslant h\left(\sum_{k=1}^n \lambda_k X_k\right) - \sum_{k=1}^n \lambda_k^2 h(X_k). \tag{3.47}$$

命题 3.3 说明原卷积的熵增量不小于 O-U 算子作用后的熵增量. 接下来的引理也被称为 de Bruijn 恒等式, 它给出了微分熵和 Fisher 信息量的关系.

引理 3.4 [10] 对于有有限方差的连续型, 随机变量 X, 有

$$\frac{d}{dt}h(X+\sqrt{t}G) = \frac{1}{2}J(X+\sqrt{t}G). \tag{3.48}$$

其中 G 是标准 Gauss 随机变量.

如下引理给出了 Fisher 信息量正交不增的性质, 也是 (3.16) 的推广.

引理 3.5 [10]　对于任意同方差的独立随机变量列 $X_1, X_2, \ldots, X_n$ 以及任意 $0 < \lambda_1, \lambda_2, \ldots, \lambda_n \leqslant 1$, $\sum_{k=1}^n \lambda_k^2 = 1$, 有

$$J\bigg(\sum_{k=1}^n \lambda_k X_k\bigg) \leqslant \sum_{k=1}^n \lambda_k^2 J(X_k), \tag{3.49}$$

等号成立当且仅当这些 X_k 都是 Gauss 的.

通过结合 de Bruijn 恒等式以及 Fisher 信息量正交不增的性质, 文献 [10] 证明了熵关于 KL 散度具有一定的连续性, 并找到了包含所有形如 $P_t^* X$ 的随机变量的紧集, 从而证明了熵跳的下界关于 KL 散度的一致性定理.

定理 3.11　设 X_1 和 X_2 为两个独立随机变量, 均值都为 0, 方差都为 1. 假设下列两条件成立:

(1) 存在 $J_0 < \infty$, 使得

$$J(X_1) \vee J(X_2) \leqslant J_0;$$

(2) 存在单调递减函数 $l(R)$, $R \geqslant 0$, 且 $\lim_{R\to\infty} l(R) = 0$, 使得

$$L_{X_1}(R) := \mathrm{E}(X_1^2 1_{|X_1|\geqslant R}) \leqslant l(R), \quad L_{X_2}(R) := \mathrm{E}(X_2^2 1_{|X_2|\geqslant R}) \leqslant l(R), \quad \forall\, R \geqslant 0. \tag{3.50}$$

则如果记

$$\varepsilon := D(X_1) \vee D(X_2) > 0,$$

则对于任意满足 $a \leqslant \lambda^2 \leqslant 1 - a$ 的实数 λ, 存在 $\delta > 0$ 仅依赖于 J_0、l、a 和 ε, 使得

$$h(\lambda X_1 + (1-\lambda^2)^{1/2} X_2) - \lambda^2 h(X_1) - (1-\lambda^2) h(X_2) \geqslant \delta. \tag{3.51}$$

这说明熵跳的增量只与 X_1 和 Gauss 分布的 KL 散度, X_2 和 Gauss 分布的 KL 散度, 以及一些正则条件有关, 与 X_1 和 X_2 具体的分布无关. 在文献 [10] 中, 定理 3.11 适用于 n 维随机向量, 此时方差设为 n.

Carlen [10] 应用定理 3.11 不仅证明了独立同分布情形的熵中心极限定理, 还将结论推广到了非独立的情形.

定理 3.12　考虑一列随机变量 $X_1, X_2, \ldots$, 均值都为 0, 且有非退化的协方差. 定义

$$Y_{1,k} = X_k; \quad Y_{n,k} := Y_{n-1,2k-1} + Y_{n-1,2k}, \quad Z_{n,k} := (\mathrm{Var}(Y_{n,k}))^{-1/2} Y_{n,k}. \tag{3.52}$$

记 $\tilde{Z}_{n,k} = P_t^* Z_{n,k}$, $\tilde{W}_{n,k}$ 为 $\tilde{Z}_{n,k}$ 的独立复制. 如果任意给定一族 $\lambda_{n,k}$ 满足

$$a \leqslant \lambda_{n,k}^2 \leqslant 1 - a, \tag{3.53}$$

其中 $a > 0$, 都有

$$\lim_{n\to\infty} \sup\{[h(\lambda_{n,k}\tilde{Z}_{n,2k-1} + (1-\lambda_{n,k}^2)\tilde{W}_{n,2k}) - h(\tilde{Z}_{n+1,k})] \vee 0 \mid k \in \mathbb{N}^+\} = 0, \quad \forall\, t > 0, \tag{3.54}$$

则对于任意 Borel 集 A, 都有

$$\lim_{n\to\infty} \Pr(Z_{n,k} \in A) = \Pr(G \in A), \quad \forall\, k, \tag{3.55}$$

其中 G 是标准 Gauss 随机变量. 事实上, 可以证明

$$\lim_{n\to\infty} \sup_k D(Z_{n,k}) = 0, \tag{3.56}$$

即收敛关于 k 是一致的.

可以发现, 定理的核心要求是熵的收敛性, 而这在 i.i.d. 情形自然成立, 因为此时熵是递增有上界的. 在下一节中也将再次看到, 随机变量列熵的极限存在性, 是证明其收敛到 Gauss 分布的一项重要条件, 我们有理由猜测, 熵极限的存在性可能成为熵中心极限定理的充分必要条件. 挖掘熵中心极限定理的充要假设, 特别是 KL 散度收敛的本质条件, 是未来熵极限理论的一个可能研究方向.

值得提出的是, 定理 3.11 可以推广至方差不同的情形, 但需要对文献 [10] 中所选择的紧集以及熵的连续性进行重新的梳理和证明. 下一节将展示推广后的结论, 以及进一步地推导出的条件熵中心极限定理的相关结果.

4 条件熵中心极限定理

正如文献 [41] 中提到的, 应用科学中出现的许多问题都是在给定一定条件下的随机数学模型, 如先验估计、预测模型或压缩感知等实际问题, 都是在给定一些信息的情形下进行问题处理, 因此有必要研究条件分布的性质. 而作为中心极限定理的衍生问题, 条件分布的极限理论也始终备受关注. Holst [18] 研究了离散的独立同分布随机变量和的条件分布, 并证明了一些特殊条件分布的中心极限定理, 他的结果在 2001 年被进一步改进 (参见文献 [19]). Rubshtein [34] 应用鞅极限理论证明了平稳分布的条件中心极限定理. 近年来, 文献 [41] 给出了条件独立的情形下的中心极限定理; 而文献 [15] 则是针对条件中心极限定理问题, 使用了 Stein 方法进行探究.

本节余下的部分将简要介绍我们在条件熵中心极限定理研究方向的一些最新研究成果, 以及其在 Hadamard 压缩问题上的应用, 详细内容可参见文献 [29].

4.1 问题背景

我们对条件中心极限定理的研究, 起源于通信中的实数域信道极化问题. 经典的信息编解码方法主要针对有限数域上的信号, 经过通信技术多年的发展和积累, 有限数域上的编解码技术已趋于成熟. 例如, 引导 3G 技术革命的 Turbo 码以及目前在 5G 标准中广泛使用的 LDPC 码和 Polar 码, 都已经能够高效地逼近香农极限. 然而, 新一代的通信网络有着更高的性能要求和更广泛的应用场景, 这对信号的编解码方案提出了更加严峻的挑战. 面对这些新一代通信网络中出现的问题, 现有的有限数域上经典的信息编解码技术难以满足实际的需要, 这就促使我们思考如何针对实数域上的模拟信号设计高效的编解码方案.

与有限数域不同, 实数域具有连续的码字空间, 这意味着模拟信号中蕴含着更多的信息, 但也同时为实数域上的编解码问题带来了更大的难度, 有限数域上的经典信息编解码方案很难直接推广到实数域中. 因此, 目前实数域广义编解码技术领域的研究仍处于初步阶段, 针对模拟信号的编解码, 至今仍没有高效可行的方案. 然而, 许多实际应用问题中对实数域上的编解码已经提出了需求, 如压缩感知和信道预测等问题. 在实数域广义编解码这一领域, 目前的理论储备远远不能满足实际应用的需要.

Polar 码之父 Arıkan [3] 在 2021 年提出了实数域上的压缩问题, 并提出了 Hadamard 变换可以作为 Polar 矩阵在实数域上的一种可能的推广形式. 而早在 2010 年, Pilancı [32] 就对 Hadamard 压缩能力进行了仿真模拟, 并得到了良好的效果. 具体而言, 考虑输入信号

$$\boldsymbol{\xi}+\boldsymbol{G}=(\xi_1,\xi_2,\ldots,\xi_{2^n})+(G_1,G_2,\ldots,G_{2^n}), \tag{4.1}$$

其中 ξ_i, $i=1,2,\ldots,2^n$ 是独立同分布的某个随机变量, G_i, $i=1,2,\ldots,2^n$ 是独立同分布的 Gauss 分布. 选择压缩矩阵为 Hadamard 矩阵

$$H_n := \left(\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\right)^{\otimes n}, \tag{4.2}$$

其中 $\otimes n$ 表示 n 次 Kronecker 积. 文献 [32] 指出, 在对变换后的信号

$$\boldsymbol{\eta} = H_n\boldsymbol{\xi} + \boldsymbol{G} = H_n(\xi_1,\xi_2,\ldots,\xi_{2^n})^{\mathrm{T}} + (G_1,G_2,\ldots,G_{2^n})^{\mathrm{T}} \overset{d}{=} H_n(\boldsymbol{\xi}+\boldsymbol{G}), \tag{4.3}$$

进行压缩时, 可以通过计算条件熵 $h(\eta_k \mid \eta_1,\eta_2,\ldots,\eta_{k-1})$, 并保留熵较高的信号位, 即可实现良好的压缩和恢复性能 (详见文献 [32, 第 4 章]). 然而这个方案的表现仅通过实验证明, 再无充足的理论结果对其进行支撑.

通过分析条件分布 $[\eta_k \mid \eta_1,\eta_2,\ldots,\eta_{k-1}]$ 的极限性质, 欲将其熵与信号的不确定性关联起来, 为 Hadamard 压缩方案提供有力的理论依据.

4.2　定理内容

为了使用数学语言描述 Hadamard 压缩信号分布 $[\eta_k \mid \eta_1,\eta_2,\ldots,\eta_{k-1}]$ 的性质, 定义一个概率空间和一个映射.

定义 4.1　定义概率空间 $(\Omega,\mathcal{F},\Pr)$, 其中 $\Omega=\{0,1\}^\infty$, $\mathcal{F}$ 是由所有柱集 $S(b_1,b_2,\ldots,b_n):=\{\omega\in\Omega:(\omega_1,\omega_2,\ldots,\omega_n)=(b_1,b_2,\ldots,b_n)\}$, $n\geqslant 1$, $(b_1,b_2,\ldots,b_n)\in\{0,1\}^n$ 生成的 σ- 代数; Pr 满足对于任意柱集 $S(b_1,b_2,\ldots,b_n)$, 有 $\Pr(S(b_1,b_2,\ldots,b_n))=1/2^n$.

注 4.1　根据 Kolmogorov 相容性定理, 上述概率空间存在.

定义 4.2　给定 $n>0$ 和 $\omega\in\Omega:=\{0,1\}^\infty$, 定义

$$k := k(n,\omega) = [\omega_1\omega_2\cdots\omega_n]_{2\to 10}+1, \tag{4.4}$$

其中 $[\cdot]_{2\to 10}$ 表示从二进制到十进制的转换, 记

$$\eta_k^n(\omega) = \eta_{k(n,\omega)}^n. \tag{4.5}$$

事实上, ω 表示极化路径 (详见文献 [29], 也可参见文献 [2]), $k=k(n,\omega)$ 表示 ω 这条路径下, 第 n 层极化对应的信号位. 可以发现, 对于任意 $n\in\mathbb{N}^+$ 和 $\omega\in\Omega$, 都有

$$\begin{aligned}&|\mathrm{E}h(\eta^{n+1}_{k(n+1,\omega)} \mid \eta^{n+1}_{1:k(n+1,\omega)-1}) - \mathrm{E}h(\eta^n_{k(n,\omega)} \mid \eta^n_{1:k(n,\omega)-1})|\\ &\quad = \mathrm{E}h(\eta^n_{k(n,\omega)} *_\lambda (\eta^n_{k(n,\omega)})' \mid \eta^n_{1:k(n,\omega)-1}, (\eta^n_{1:k(n,\omega)-1})') - \mathrm{E}h(\eta^n_{k(n,\omega)} \mid \eta^n_{1:k(n,\omega)-1}),\end{aligned} \tag{4.6}$$

其中 X' 表示 X 的 i.i.d. 复制. 如果固定 ω 并记

$$X_n = \eta^n_{k(n,\omega)},\quad Y_n = (\eta^n_1,\eta^n_2,\ldots,\eta^n_{k(n,\omega)-1}) = \eta^n_{1:k(n,\omega)-1}, \tag{4.7}$$

则可以将 (4.6) 改写为

$$|\mathrm{E}_{Y_{n+1}}h(X_{n+1}\mid Y_{n+1}) - \mathrm{E}_{Y_n}h(X_n\mid Y_n)| = \mathrm{E}_{Y_n,Y_n'}h(X_n *_\lambda X_n' \mid Y_n,Y_n') - \mathrm{E}_{Y_n}h(X_n\mid Y_n). \tag{4.8}$$

我们认为, (4.8) 是 Hadamard 信号的核心性质.

4.2.1 主定理

正如 Carlen 等 [10] 将他们考虑的模型最终总结为一个熵收敛条件 (3.54), 本文作者也将结论仅聚焦在满足 (4.8) 的更一般的情形. 另外, (4.8) 的右侧实际上为熵增形式, 因此我们的问题与 Carlen 等所研究的熵跳一致下界存在着非常密切的联系. 在文献 [10] 中, 定理 3.11 的核心证明思路是连续函数在紧集上达到最小值, 以及熵增为 0 当且仅当所有随机变量都为 Gauss 分布的这两个性质. 本文对 Carlen 等 [10] 的证明进行了扩展, 得到了在方差不同情形下的熵增估计.

定理 4.1 设 X 和 Y 为两个有绝对连续密度函数的独立随机变量, 并假设下面列条件成立:

(1) 存在实数 J_0, 使得 $\min\{J(X), J(Y)\} \leqslant J_0$;

(2) 存在实数 a, 使得 $\max\{V(X), \mathrm{Var}(Y)\} \geqslant a$;

(3) 存在单调递减函数 $l(R), \lim_{R\to\infty} l(R) = 0$, 使得

$$\min\{L_X(R), L_Y(R)\} \leqslant l(R), \quad \forall\, R > 0, \tag{4.9}$$

其中 L_X 和 L_Y 为 X 和 Y 的尾方差, 定义参见 (3.50).

则当

$$\max\{D(X), D(Y)\} + |\mathrm{Var}(X) - \mathrm{Var}(Y)| \geqslant \epsilon > 0 \tag{4.10}$$

时, 对于任意 $\lambda \in (0,1)$, 存在 $\delta = \delta(\lambda, J_0, a, l, \varepsilon)$ (而不依赖于 X 和 Y 的具体分布), 使得

$$h(\lambda X + \sqrt{1-\lambda^2} Y) \geqslant \lambda^2 h(X) + (1-\lambda^2) h(Y) + \delta. \tag{4.11}$$

定理 4.1 将定理 3.11 推广至了方差有界的情形, 但由于失去了方差相等的条件, 我们在证明过程中需要对 Carlen 等 [10] 所选取的紧集和熵连续性进行对应的修改, 以达到证明定理 4.1 的目的.

下面给出我们的条件熵中心极限定理, 首先定义条件 KL 散度.

定义 4.3 设条件随机变量 $X \mid Y = y$ 是有绝对连续密度函数的连续随机变量, 其与 Gauss 分布的差距可以使用条件 KL 散度

$$D(X \mid Y = y) := \mathrm{KL}[(X \mid Y = y) \,\|\, \mathcal{N}(\mu_y, \sigma_y^2)] = \int f(x \mid y) \log \frac{f(x \mid y)}{g(x)} dx \tag{4.12}$$

来度量, 其中

$$\mu_y = \mathrm{E}(X \mid Y = y), \quad \sigma_y^2 = \mathrm{Var}(X \mid Y = y),$$

$f(x \mid y)$ 是 X 在给定条件 $Y = y$ 时的密度函数, g 是 Gauss 分布 $\mathcal{N}(\mu_y, \sigma_y^2)$ 的密度函数.

定理 4.2 设随机序列 $\{(X_n, Y_n)\}_{n=1}^{\infty}$ (其中 X_n 为随机变量, Y_n 可以是随机向量) 满足对于几乎处处的 y_n, 条件分布 $X_n \mid Y_n = y_n$ 有绝对连续的密度函数. 假设下列条件成立:

(1) 存在 $a > 0$ 使得 $\mathrm{Var}(X_n \mid Y_n) \geqslant a$, a.s., $\forall\, n \in \mathbb{N}^+$;

(2) 存在 $J_0 < \infty$ 使得 $\sup_n \mathrm{E}J(X_n \mid Y_n) \leqslant J_0$;

(3) 存在单调递减函数 $l(R)$, $\lim_{R\to\infty} l(R) = 0$, 使得 $\mathrm{E}[L_{X_n|Y_n}(R)] \leqslant l(R)$, $\forall\, n \in \mathbb{N}^+, R \geqslant 0$.

如果对于某个 $\lambda \in (0,1)$,

$$\lim_{n\to\infty} \mathrm{E}h(X_n \mid Y_n) \text{ 存在}, \tag{4.13}$$

且

$$|\mathrm{E}h(X_{n+1}\mid Y_{n+1})-\mathrm{E}h(X_n\mid Y_n)|=\mathrm{E}[h(X_n *_\lambda X_n'\mid Y_n,Y_n')-h(X_n\mid Y_n)], \tag{4.14}$$

其中 (X_n',Y_n') 是 (X_n,Y_n) 的独立复制, 则

$$D(X_n\mid Y_n)\xrightarrow{\mathrm{Pr.}}0,\quad n\to\infty, \tag{4.15}$$

其中

$$D(X_n\mid Y_n)|_{Y_n=y_n}=\mathrm{KL}[(X_n\mid Y_n=y_n)\,\|\,\mathcal{N}(\mathrm{E}(X_n\mid Y_n=y_n),\mathrm{Var}(X_n\mid Y_n=y_n))]. \tag{4.16}$$

更进一步地, 方差渐近达到其均值

$$\lim_{n\to\infty}\mathrm{E}|\mathrm{Var}(X_n\mid Y_n)-\mathrm{EVar}(X_n\mid Y_n)|=0. \tag{4.17}$$

特别地, 如果 $\{h(X_n\mid Y_n),n\geqslant 1\}$ 一致可积, 则存在实数 σ, 使得

$$\lim_{n\to\infty}\mathrm{E}|\mathrm{Var}(X_n\mid Y_n)-\sigma^2|=0, \tag{4.18}$$

$$\lim_{n\to\infty}\mathrm{E}h(X_n\mid Y_n)=\frac{1}{2}\log 2\pi\mathrm{e}\sigma^2. \tag{4.19}$$

与文献 [10] 相比, 我们除了推广了其定理, 还在证明定理 4.2 时, 不仅运用了熵增为 0 时随机变量为 Gauss 随机变量的这一性质, 而且运用了方差相等这一要求, 从而得到方差渐近性 (4.17).

通过考虑不同的 ω, 我们的定理可以应用于求和形式的条件中心极限定理以及 Hadamard 压缩模型. 下面将给出定理 4.2 在这两个问题上对应的推论.

4.2.2　求和形式的条件中心极限定理

推论 4.1　设 (ξ,η) 为二维随机变量, 并设 $(\xi_1,\eta_1),(\xi_2,\eta_2),\dots,(\xi_n,\eta_n)$ 是独立同分布随机变量列, 分布与 (ξ,η) 相同. 记 $W_n=(\xi_1+\xi_2+\cdots+\xi_n)/\sqrt{n}$ 以及 $\boldsymbol{\eta}_n=(\eta_1,\eta_2,\dots,\eta_n)$. 如果下列条件成立:

(1) $\mathrm{Var}(\xi)<\infty,(\xi\mid\eta)\prec\mathcal{L},\mathrm{Pr}_\eta$-a.s., 其中 $\mathcal{L}$ 表示 Lebesgue 测度, $\mu\prec\lambda$ 表示 μ 关于 λ 绝对连续;

(2) $\mathrm{E}J(\xi\mid\eta)<\infty$, $\mathrm{E}h(\xi\mid\eta)>-\infty$;

(3) 存在 $a>0$ 使得 $\mathrm{Var}(\xi\mid\eta)\geqslant a$, Pr_η-a.s.

记 $h_n=\mathrm{E}h(W_n\mid\boldsymbol{\eta}_n)$, $\sigma^2=\mathrm{EVar}(\xi\mid\eta)$, 则有

$$\lim_{n\to\infty}h_n=\frac{1}{2}\log(2\pi\mathrm{e}\sigma^2) \tag{4.20}$$

和

$$\lim_{n\to\infty}\mathrm{E}D(W_n\mid\boldsymbol{\eta}_n)=0. \tag{4.21}$$

如果取 ξ 和 η 是独立的, 则本文的结果就回归了 i.i.d. 求和情形的熵中心极限定理. 此外, 虽然定理 4.1 阐述的问题与文献 [34] 中的模型相似, 但 Rubshtein 的结论是弱收敛到 Gauss 分布, 而本文的结论则是在 KL 散度意义下的收敛性, 依据 Pinsker 不等式, 本文的收敛性略强.

4.2.3 Hadamard 压缩问题

为了进一步描述 Hadamard 信号的分布特征, 定义下面 3 个随机序列.

定义 4.4 在概率空间 $(\Omega, \mathcal{F}, \Pr)$ 上, 定义

$$\begin{aligned} h_n(\omega) &= \int_{\mathbb{R}^{k-1}} f_{1:k(n,\omega)-1}(y) h(f_{k(n,\omega)\,|\,y}) dy = h(\eta^n_{k(n,\omega)} \mid \eta^n_{1:k(n,\omega)-1}), \\ V_n(\omega) &= \int_{\mathbb{R}^{k-1}} f_{1:k(n,\omega)-1}(y) \mathrm{Var}(f_{k(n,\omega)\,|\,y}) dy = \mathrm{E}[\mathrm{Var}(\eta^n_{k(n,\omega)} \mid \eta^n_{1:k(n,\omega)-1})], \\ J_n(\omega) &= \int_{\mathbb{R}^{k-1}} f_{1:k(n,\omega)-1}(y) J(f_{k(n,\omega)\,|\,y}) dy = J(\eta^n_{k(n,\omega)} \mid \eta^n_{1:k(n,\omega)-1}). \end{aligned}$$

以上 3 式分别称为可微熵过程、方差过程和 Fisher 信息量过程.

定理 4.2 在 Hadamard 压缩模型中对应以下推论:

推论 4.2 假设输入信号 ξ 具有形式

$$\xi = \hat{\xi}_0 + G_{\sigma_0^2}, \tag{4.22}$$

其中 σ_0 为某个非零实数, $G_{\sigma_0^2} \sim \mathcal{N}(0, \sigma_0^2)$, $\hat{\xi}_0$ 与 $G_{\sigma_0^2}$ 独立, $\mathrm{E}\hat{\xi}_0 = 0$, $\mathrm{E}\hat{\xi}_0^4 < \infty$, 则存在集合 $O \subset \Omega$, $\Pr(O) = 1$, 使得

$$h_\infty(\omega) = \frac{1}{2}\log(2\pi \mathrm{e} V_\infty(\omega)), \quad \forall\, \omega \in O, \tag{4.23}$$

$$\lim_{n\to\infty} \mathrm{E}_y |\mathrm{Var}(\eta^n_k(\omega) \mid \eta^n_{1:k(\omega)-1} = y) - V_\infty(\omega)| = 0, \quad \forall\, \omega \in O, \tag{4.24}$$

$$J_\infty(\omega) V_\infty(\omega) = 1, \quad \forall\, \omega \in \Omega, \tag{4.25}$$

其中

$$h_\infty(\omega) = \lim_{n\to\infty} h_n(\omega), \quad V_\infty(\omega) = \lim_{n\to\infty} V_n(\omega), \quad J_\infty(\omega) = \lim_{n\to\infty} J_n(\omega).$$

更进一步地, 有

$$\frac{1}{J(\xi)} \leqslant V_\infty(\omega) \leqslant \mathrm{Var}(\xi), \quad \forall\, \omega \in \Omega. \tag{4.26}$$

定理 4.2 说明, 当信号长度足够大时, 对于几乎所有的信号位 k, 如果已知前 $k-1$ 位的信号结果, 其分布都是渐近 Gauss 的, 再根据 Gauss 分布的熵满足

$$h(G_{\sigma^2}) = \frac{1}{2}\ln(2\pi \mathrm{e}\sigma^2), \tag{4.27}$$

可知熵的大小可以直接对应方差的大小, 即熵越小, 信号位的确定性越高, 这说明了文献 [32] 提供的压缩方案在理论上也是合理的.

此外, 关于方差的结果 (4.24) 说明了信号的方差与信号本身的内容是渐近无关的, 即 Hadamard 压缩模型的性能与发送的信号在码长足够长时是无关的, 这一点说明了 Hadamard 压缩是具有一定的泛用性的, 并非需要根据特定信号内容对压缩方案进行更改.

本文作者对 Hadamard 变换的模拟压缩也有一些最新的研究成果, 有兴趣的读者可以参见文献 [42, 43].

致谢 作者感谢南方科技大学邵启满和华为技术有限公司李源, 在研究条件熵中心极限定理 (第 4 节内容) 时与他们进行了有益的讨论, 邵启满的宝贵建议极大地改进了第 4 节的研究成果. 李源帮助作者修改了第 4 节内容的早期版本. 感谢华为技术有限公司张华滋, 张华滋建议作者研究通信中的实数域信道极化问题, 并参与了条件熵中心极限定理的合作研究.

参考文献

1 Andriamanalimanana B, Tekeoglu A, et al. Symmetric Lullback-Leibler divergence of softmaxed distributions for anomaly scores. In: IEEE Conference on Communications and Network Security. Washington: IEEE, 2019, 1–6

2 Arıkan E. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. IEEE Trans Inform Theory, 2009, 55: 3051–3073

3 Arıkan E. Entropy polarization in butterfly transforms. Digit Signal Process, 2021, 119: 103207

4 Artstein S, Ball K M, Barthe F, et al. On the rate of convergence in the entropic central limit theorem. Probab Theory Related Fields, 2004, 129: 381–390

5 Barron A R. Entropy and the central limit theorem. Ann Probab, 1986, 14: 336–342

6 Bobkov S G, Chistyakov G P, Götze F. Fisher information and the central limit theorem. Probab Theory Related Fields, 2014, 159: 1–59

7 Bobkov S G, Chistyakov G P, Götze F. Berry-Esseen bounds in the entropic central limit theorem. Probab Theory Related Fields, 2014, 159: 435–478

8 Borovkov A A, Utev S A. On an inequality and a related characterization of the normal distribution. Theory Probab Appl, 1984, 28: 219–228

9 Brown L D. A proof of the central limit theorem motivated by the Cramer-Rao inequality. In: Statistics and Probability: Essays in Honor of C. R. Rao. Amsterdam: North-Holland, 1982

10 Carlen E A, Soffer A. Entropy production by block variable summation and central limit theorems. Comm Math Phys, 1991, 140: 339–371

11 Chen H Y, Goldstein L, Shao Q M. Normal Approximation by Stein's Method. Heidelberg: Springer-Verlag, 2011

12 Chen J, Matzinger H, Zhai H Y, et al. Centroid estimation based on symmetric KL divergence for multinomial text classification problem. In: Proceedings of the 17th IEEE International Conference on Machine Learning and Applications. Orlando: IEEE, 2018, 1174–1177

13 Courtade T A. A strong entropy power inequality. IEEE Trans Inform Theory, 2018, 64: 2173–2192

14 David R B, Mark S P, Amiel F. Information and information stability of random variables and processes. J Roy Statist Soc Ser C (Appl Statist), 1964, 13: 134–135

15 Dey P S, Terlov G. Stein's method for conditional central limit theorem. Ann Probab, 2023, 51: 723–773

16 Durrett R. Probability: Theory and Examples, 5th ed. Cambridge: Cambridge Univ Press, 2019

17 Heinonen J, Koskela P, Shanmugalingam N, et al. Sobolev Spaces on Metric Measure Spaces: An Approach Based on Upper Gradients. Cambridge: Cambridge Univ Press, 2015

18 Holst L. Two conditional limit theorems with applications. Ann Statist, 1979, 7: 551–557

19 Janson S. Moment convergence in conditional limit theorems. J Appl Probab, 2001, 38: 421–437

20 Johnson O. Entropy inequalities and the central limit theorem. Stochastic Process Appl, 2000, 88: 291–304

21 Johnson O. Information Theory and the Central Limit Theorem. London: Imperial College Press, 2004

22 Johnson O. Maximal correlation and the rate of Fisher information convergence in the central limit theorem. IEEE Trans Inform Theory, 2020, 66: 4992–5002

23 Johnson O, Barron A. Fisher information inequalities and the central limit theorem. Probab Theory Related Fields, 2004, 129: 391–409

24 Kristian L. Information Theory for Complex Systems. Berlin-Heidelberg: Springer, 2024

25 Lampros G, Ioannis K. The entropic central limit theorem for discrete random variables. In: IEEE International Symposium on Information Theory. Espoo: IEEE, 2022, 708–713

26 Li J, Marsiglietti A, Melbourne J. Entropic central limit theorem for Rényi entropy. In: IEEE International Symposium on Information Theory. Paris: IEEE, 2019, 1137–1141

27 Li X D. Entropy helps us understand chaos and disorder. In: Discover Mathematics, vol. 3 (in Chinese). Editor-in-chief Xi N H. Beijing: Science Press, 2022 [李向东. 熵助我们理解混乱与无序. 收录于《认识数学》第 3 辑. 席南华主编. 北京: 科学出版社, 2022]

28 Linnik J V. An information-theoretic proof of the central limit theorem with Lindeberg conditions. Theory Probab Appl, 1959, 4: 288–299

29 Ma Z-M, Yao L-Q, Yuan S, et al. Entropic conditional central limit theorem and Hadamard compression. arXiv:

2401.11383, 2024
30 McKean H P Jr. Speed of approach to equilibrium for Kac's caricature of a Maxwellian gas. Arch Ration Mech Anal, 1966, 21: 343–367
31 Petrov V V. Sums of Independent Random Variables. Heidelberg: Springer-Verlag, 1975
32 Pilancı M. Uncertain linear equations. Master Thesis. Ankara: Bilkent University, 2010
33 Romanovskiĭ N N. Sobolev spaces on an arbitrary metric measure space: Compactness of embeddings. Sib Math J, 2013, 54: 353–367
34 Rubshtein B-Z. A central limit theorem for conditional distributions. In: Convergence in Ergodic Theory and Probability. Berlin-New York: De Gruyter, 1996, 373–380
35 Schrödinger E. What Is Life? Cambridge: Cambridge Univ Press, 1992
36 Shannon C E. A mathematical theory of communication. Bell Syst Tech J, 1948, 27: 623–656
37 Stam A J. Some inequalities satisfied by the quantities of information of Fisher and Shannon. Inf Control, 1959, 2: 101–112
38 Yan S J, Liu X F. Measure and Probability, 2nd ed (in Chinese). Beijing: Beijing Normal Univ Publ, 2003 [严士健, 刘秀芳. 测度与概率 (第二版). 北京: 北京师范大学出版社, 2003]
39 Yao L-Q, Yuan S. Entropy jump and entropic central limit theorem for independent sum. arXiv:2402.08953, 2024
40 Yu L. The entropy method in large deviation theory. arXiv:2210.13121, 2022
41 Yuan D M, Wei L R, Lei L. Conditional central limit theorems for a sequence of conditional independent random variables. J Korean Math Soc, 2014, 51: 1–15
42 Yuan S, Yao L-Q, Li Y, et al. Lossless analog compression via polarization. In: Proceedings of the 2023 IEEE Global Communications Conference. Kuala Lumpur: IEEE, 2023, 7001–7006
43 Yuan S, Yao L-Q, Li Y, et al. Achieving the fundamental limit of lossless analog compression via polarization. arXiv: 2312.06200, 2024

A review for entropic central limit theorem

Zhi-Ming Ma, Liuquan Yao & Shuai Yuan

Abstract In this paper, we review the research on the entropic central limit theorem starting from the 1950s, including the analysis of the limit distribution of the sum of independent identically distributed random variables using the entropy method, the Berry-Esseen type bound for entropic central limit theorem and the convergence results in the non-independent case. By investigating the development of the entropic central limit theorem, we hope to exhibit the subtle relationship among thermodynamics, information theory, and probability theory. In the last part of this paper, the latest research results of authors about the entropic conditional central limit theorem and its application to Hadamard compression are briefly introduced.

Keywords **entropic central limit theorem, entropic inequality, conditional central limit theorem, Hadamard transform**

MSC(2020) **60F05, 68P30, 62B10**

doi: 10.1360/SSM-2024-0049