arXiv:2301.05660v1 [physics.data-an] 13 Jan 2023

从有信息数据中学习熵：确保一致识别广义熵的一条公理

Andrea Somazzi[1, 2, *]和 Diego Garlaschelli[1, 3, 4]

[1]*IMT School for Advanced Studies, Piazza S. Francesco 19, 55100 Lucca, Italy*[2]*Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy*[3]*Lorentz Institute for Theoretical Physics, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands*[4]*INdAM-GNAMPA Istituto Nazionale di Alta Matematica, Italy*

香农熵是信息论、统计物理学和推断方法的基石，由香农-金欣公理或 Shore-Johnson公理唯一确定。对香农熵的推广源于对非广延系统或非遍历系统的研究；这些推广放宽了部分公理，并导出由某些“熵”参数索引的熵族。一般而言，选择这些参数需要预先了解系统，或会导致不一致。本文为任意熵族提出一条简单公理：任何熵参数都不能从完全无信息（均匀）的概率分布中推断得出。将该公理应用于 Uffink-Jizba-Korbel 熵和哈内尔-图尔纳熵时，只有 R´enyi 熵被选为可行熵。该公理还使其与最大似然原理保持一致，进而可以仅凭数据估计熵参数，这一点已通过数值计算得到验证。值得注意的是，在广义最大熵框架下，该公理意味着最大化对数似然始终等于负的香农熵，即使推断出的概率分布最大化的是广义熵而非香农熵，从而解决了以往方法中遇到的一系列问题。

I. 引言

熵的概念由克劳修斯在热力学框架中提出 [1] ，后来玻尔兹曼和吉布斯将其应用于统计物理学，用以根据宏观系统处于各种微观状态的概率描述系统 [2, 3]。在信息论中，香农通过公理化推导（重新推导）出熵，将其定义为概率分布所包含的不确定性的量度；熵还可用于衡量遍历概率源生成的符号序列的可压缩性等 [4]。由此，杰恩斯随后提出：在由关于真实系统的可用经验信息所隐含的约束条件下，使香农熵最大的分布，能够以最少偏见（最大限度不作预设）的方式推断描述该系统未知的微观细节；这一构想可用于从信息论视角重新诠释统计物理学 [5]。在现代研究中，基于熵最大化的统计推断和模型识别与最大似然估计方法完全一致，也是若干机器学习技术的核心 [6]。

人们在不同的背景下提出了香农熵的各种推广形式（其中最流行的形式源于非广延和/或非遍历系统的统计物理学） [7–10], ，由此得到了一系列扩展的熵；除通常的参数外，这些熵还取决于额外的“熵”参数。对于这些参数的一种固定取值，仍然可以最大化由此得到的熵，并推广推断过程。当我们对系统有足够的先验知识，从而可以“手动”将熵参数设为正确值时，这是可行的。然而，通常无法同时保持与最大似然原理的兼容性，更关键的是，无法仅从数据中推断熵参数的值而不遇到矛盾；因此，若没有关于正确熵的先验知识，这种广义方法便无法应用。

本文通过引入一条限制参数化熵形式的公理，探讨并缓解这些不一致性。正如我们通过解析结果和数值示例所展示的，这条公理提出了一项简单的信息论要求，并使我们能够仅根据现有数据一致地推断熵参数。本文结构如下。在第II 节中，我们首先回顾熵的公理化定义、最大熵原理和最大似然原理的理论背景。在第III 节中，我们接着讨论本文的主要贡献，即提出这条新公理及其对从若干常见熵族中选择熵的影响、恢复与最大似然原理的一致性，以及对后者进行推广，以便仅根据数据推断熵参数。最后，在第IV节中，我们给出一些总结性意见。

* andrea.somazzi@sns.it

II. 理论背景

A. 香农-欣钦公理

给定一个分布 $P = (p(G_1) \ldots, p(G_\Omega))$，其中 $p(G_i)$ 是随机变量 G 取第 i 个离散结果（或“状态”） G_i 的概率，Ω 是不同结果的总数，并且显然 $\sum_{i=1}^{\Omega} p(G_i) = 1$，香农熵 $S[P]$ 通过以下四条香农–金欣（SK）公理 [11] 进行公理化定义：

- SK1（连续性）： $S[P]$ 在 P的各项取值上连续。

- SK2（极大性）：当 P 为均匀分布 $P_u \equiv (\Omega^{-1}, \ldots, \Omega^{-1})$时， $S[P]$ 取极大值。

- SK3（可扩展性）： $S[P]$ 具有可扩展性，即当为变量 G 添加一个概率为零的（$\Omega + 1$）阶结果（$p(G_{\Omega+1}) = 0$）时，其值不变：

$$S[(p(G_1), \ldots, p(G_\Omega))] = S[(p(G_1), \ldots, p(G_\Omega), 0)].$$

- SK4（可分性）：两个变量 G 和 G' 的联合分布 $R = (r(G_1, G'_1) \ldots, r(G_\Omega\ G'_{\Omega'}))$的熵可分解为其边缘分布 $P = (p(G_1) \ldots, p(G_\Omega))$和 $Q = (q(G'_1) \ldots, q(G'_{\Omega'}))$，其中 $p(G_i) = \sum_{j=1}^{\Omega'} r(G_i, G'_j)$且$q(G'_j) = \sum_{i=1}^{\Omega} r(G_i, G'_j)$，分别满足：

$$S[R] = S[P] + S[Q|P].$$

这里， $S[Q|P]$ 是 Q 在P上的条件熵，定义为 $S[Q|P] = \sum_{k=1}^{\Omega} p(G_k) S[Q_{|k}]$，其中$Q_{|k} = (r(G_k, G'_1)/p(G_k) \ldots, r(G_k, G'_{\Omega'}/p(G_k))$表示$Q$ 中各事件在 P中的第 k个事件条件下的条件分布。请特别注意，如果这两个事件相互独立（对所有 k，$Q = Q_{|k}$），那么 $S[R] = S[P] + S[Q]$，此时可分性便成为可加性。

可以证明，除去一个无关紧要的整体乘法因子，满足四条SK公理的 $S[P]$唯一的函数形式是香农熵：

$$S_1[P] = -\sum_{i=1}^{\Omega} p(G_i) \ln p(G_i), \tag{1}$$

其中，下标1将在后文说明其合理性。按照SK2的要求， $S_1[P]$ 的最大值在均匀分布 P_u处取得，由此得到玻尔兹曼熵：

$$S_1[P_u] = \ln \Omega. \tag{2}$$

不存在这样的分布 P，使得 $S_1[P] > S_1[P_u]$。

B. 最大熵原理

式（ $S_1[P]$ 1）中的信息熵与吉布斯推导出的物理熵相一致 [3], ，而后者又是式（2）中玻尔兹曼熵的推广 [2]。这种等价性并非偶然，而是植根于统计推断，正如杰恩斯在提出最大熵原理（MEP）时所阐明的那样 [5]。最大熵原理指出，在仅掌握一组 I 关于某个系统的经验信息（在物理情境中，这通常意味着知道少数几个宏观守恒量，例如总能量和/或粒子总数）的情况下，应为可能的微观状态指定一个概率分布 P，使其熵达到最大。这个最大熵分布有时记作 $P = \circ I$。换言之，熵可以用作一种推断泛函；对其进行最大化可以使偏差最小，并避免武断性。

特别地，考虑一个具有一组 Ω 个可能微观状态 $\{G_i\}_{i=1}^{\Omega}$ 的系统，并假设可用信息 I 由系统微观状态的某个（标量或向量）函数 C 的经验值$C^* = C(G^*)$编码，其中 G^* 是特定的（不可观测的）经验微观状态。C^* 是唯一可用的观测。由于 G^* 未知，因此将微观状态视为随机变量 G。将 MEP 应用于 $S_1[P]$ ，可确定 G 的最大熵分布，我们将其记作 $P_0 = (p_0(G_1) \ldots, p_0(G_\Omega))$，或者$P_1 = (p_1(G_1) \ldots, p_1(G_\Omega))$，具体取决于是否将 C^* 视为“硬”约束或“软”约束。

在硬约束的情形下（微正则系综），只有数量受限的微观状态 i（共 $\Omega_{C^*} < \Omega$ 个）会被赋予非零概率，这些状态满足 $C(G_i)$ 与 C^* 完全一致。由于 SK2 和 SK3 的要求，该概率在受限支撑集上必须均匀分布，即 $p_0(G_i) = \Omega_{C^*}^{-1}$ if $C(G_i) = C^*$ and $p_0(G_i) = 0$ otherwise。由此得到的熵为

$$S_1[P_0] = \ln \Omega_{C^*} < S_1[P_u]. \tag{3}$$

遗憾的是，计算 Ω_{C^*} 通常是一个困难的组合问题，因此微正则系综不适合进行解析计算。

对于软约束（正则系综），只需约束可观测量的期望值 $\langle C \rangle$ 与 C^*相匹配，即

$$\langle C \rangle \equiv \sum_{i=1}^{\Omega} p(G_i)\, C(G_i) = C^*, \tag{4}$$

从而允许完整的 Ω 个微观状态集合，但其概率 $p(G_i)$ 尚未确定且并非均匀分布。为了求出在上述软约束下使 S_1 最大化的具体概率 $p_1(G_i)$，可以引入

拉格朗日乘子 θ （其量纲与 C 相同），再加上一个用于确保 P 归一化的乘子 α，并求出具体数值（记为 P_1、θ_1、α_1），使得拉格朗日函数的所有导数

$$\mathcal{L}_1[P] \equiv S_1[P] - \alpha\left[\sum_{i=1}^{\Omega} p(G_i) - 1\right] - \theta \cdot [\langle C\rangle - C^*] \quad (5)$$

消失（记号 $\theta \cdot C$ 表示标量积）。令 $\partial\mathcal{L}[P]/\partial P|_{P_1} = 0$，即 $\partial\mathcal{L}[P]/\partial p(G_i)|_{p_1(G_i)} = 0 \forall i$，可得到 P_1的函数形式，结果是熟知的玻尔兹曼-吉布斯分布，其各项为

$$p_1(G_i,\theta) = \frac{e^{-\theta\cdot C(G_i)}}{Z_1(\theta)}, \quad Z_1(\theta) = \sum_{j=1}^{\Omega} e^{-\theta\cdot C(G_j)}, \quad (6)$$

其中 $Z_1(\theta)$ 是配分函数，它由归一化约束得出

$$\left.\frac{\partial\mathcal{L}[P_1]}{\partial\alpha}\right|_{\alpha_1} = 0 \quad \Rightarrow \quad \sum_{i=1}^{\Omega} p_1(G_i,\theta) = 1 \quad (7)$$

由此得到

$$\alpha_1 = -1 + \ln Z_1(\theta) \quad (8)$$

与 θ 的取值无关。

重要的是， P_1 只有在参数 θ 也确定之后，才算完全确定。要做到这一点，需令其余导数为零，从而确定使公式（ θ_1 4）成立的值：

$$\left.\frac{\partial\mathcal{L}[P_1]}{\partial\theta}\right|_{\theta_1} = 0 \quad \Rightarrow \quad \sum_{i=1}^{\Omega} p_1(G_i,\theta_1)\, C(G_i) = C^*, \quad (9)$$

其中，若 θ 是一个向量，该记号同样表示 $\mathcal{L}[P]$ 对θ 各分量的所有偏导数分别为零。因此，最大熵原理问题的最终解是将 θ_1 代入公式（），并记作 $P_1(\theta_1) = (p_1(G_1, \theta_1)\ldots\ldots, p_1(G_\Omega, \theta_1))$。当可观测量预期会发生波动时，采用软约束的最大熵原理；它已成功应用于物理学以外的许多领域中的推断和模型选择，包括网络理论、神经科学、经济学和生物学 [12, 13]。

C. 最大似然原理

需要特别指出的是，上文所述的最大熵原理（MEP）过程与最大似然（ML）原理有着深刻的联系，二者也具有理想的一致性；最大似然原理适用于更一般的参数概率分布和状态（不一定是最大熵分布）使对数似然在数据 θ^* 上达到最大值的参数值 G^*才是最优参数值。在我们的设定中，最大似然原理会将式（6）视为任意其他参数分布，并选取该值

$$\theta_1^* = \operatorname*{argmax}_{\theta} \ell_1(\theta), \quad \ell_1(\theta) \equiv \ln p_1(G^*,\theta). \quad (10)$$

首先，容易证明，由式（10）定义的值 θ_1^*与由式（9）定义的值 θ_1 相同， [14, 15], 即（在我们的记法中，参数旁边的星号始终表示该参数的最大似然值），也就是说

$$\left.\frac{\partial\ell_1(\theta)}{\partial\theta}\right|_{\theta_1^*} = 0 \quad \Rightarrow \quad \sum_{i=1}^{\Omega} p_1(G_i,\theta_1^*)\, C(G_i) = C^* \quad (11)$$

与式（9类似。这意味着，最大似然原理可以视为拉格朗日函数优化中关于 θ 的部分。

此外，容易证明，最大化对数似然等于负熵：

$$S_1[P_1(\theta_1^*)] = -\ell_1(\theta_1^*), \quad (12)$$

这是软约束情形下式（3的对应结果。这一关系非常重要，因为最大化似然是模型选择准则 [16, 17]的基础：如果将备选模型（即备选参数概率分布）与相同的经验数据进行比较，那么应优先选择最大化似然最高的模型（假设所有模型的复杂度相同，例如参数数量相同）。因此，式（12）保证了基于最大似然的模型排序与基于负熵的排序一致：应优先选择不确定性最低（即信息量最高）的模型。对于参数数量和/或函数形式不同的模型，则需要通过添加一个用于控制模型复杂度变化的项，来修正基于似然/熵的排序，从而得到赤池信息准则、贝叶斯信息准则、最小描述长度等准则。 [16, 17] （为简便起见，此处不讨论这种情况）。另请注意，将最大熵分布 P_1）代入式（5），可得

$$\mathcal{L}_1[P_1(\theta_1^*)] = S_1[P_1(\theta_1^*)] = -\ell_1(\theta_1^*), \quad (13)$$

因此，在 $P_1(\theta_1^*)$处求值的拉格朗日函数与最大化对数似然的负值相一致，因此也可用于对备选模型进行排序。上述所有结果表明，最大熵原理可以像最大似然原理一样用作模型选择准则，根据模型实现的熵对模型进行排序。

还必须考虑这样一种情况：系统存在M 独立观测 $\{C_m^*\}_{m=1}^M$，严格来说，这意味着系统有 M 独立同分布（i.i.d.）的实现值 $\{G_m^*\}_{m=1}^M$

微观状态 G （请回想， G 被视为随机变量）的实现值；对每个实现值，都会观测到量 $C_m^* = C_m\left(G_m^*\right)$ （$m = 1$, M）。显然，由于被多次观测的系统仍是同一个系统，描述它的概率分布仍必须由与量C耦合的单个拉格朗日乘子 θ 来指定。此时需要指出，用于确定如何最优地组合 M 观测 $\{C_m^*\}_{m=1}^M$ 以估计 θ 的原理并非最大熵原理（MEP），而是最大似然（ML）原理。事实上，将最大似然原理应用于联合对数似然 $\sum_{m=1}^M \ln p_1\left(G_m^*,\ \theta\right)$，或等价地应用于平均对数似然$\overline{\ell}_1\left(\theta\right)\ \equiv\ \sum_{m=1}^M \ln\ p_1\left(G_m^*,\ \theta\right)/M$，可通过将式（10）替换为以下表达式来表述

$$\theta_1^* = \underset{\theta}{\operatorname{argmax}}\,\overline{\ell}_1(\theta), \quad \overline{\ell}_1(\theta) \equiv \frac{\sum_{m=1}^M \ln p_1(G_m^*, \theta)}{M}. \quad (14)$$

很容易证明，条件 $\partial\overline{\ell}_1(\theta)/\partial\theta|_{\theta_1^*} = 0$识别 θ_1^* 后会得到这一众所周知的结果

$$\langle C\rangle = \frac{1}{M}\sum_{m=1}^{M} C_m^* \quad (15)$$

其中， M 观测的（算术）样本平均值由此出现。因此，为了求得最大似然参数值 θ_1^*，应将式（4）替换为式（15），或者等价地，将式（ C^* （4）中的定义重新设为 $\{C_m^*\}_{m=1}^M$的样本平均值。简而言之，样本平均值是由最大似然原理“产生”的。相反，在最大熵原理的构造中，我们无法“告知”式（5）和（9），在有 M 个观测的情况下， C^* 的含义和定义应当是什么。因此，在这种情况下，最大似然原理比最大熵原理更具信息量，这也是人们希望熵与最大似然原理所得结果完全一致的另一个原因。特别地，容易证明，由于 M 个样本相互独立，最大化后的平均对数似然$\overline{\ell}_1\left(\theta_1^*\right)$仍等于熵的负值：

$$S_1[P_1(\theta_1^*)] = -\overline{\ell}_1(\theta_1^*), \quad (16)$$

对公式（12）进行推广。注意，公式（16）没有对应的微正则情形，因为公式（3）无法推广到 $M > 1$的情况，除非所有 M 值$\{C_m^*\}_{m=1}^M$ 都相同。事实上，无法构造微正则系综，因为根据定义，它无法处理约束值的不同实现值：若对相同约束有不同观测，则只有正则系综是可行的。

以上讨论表明，熵与最大化的对数似然保持一致至关重要，因为最大似然原理既适用于模型选择，也用于确定如何合并同一系统的多个观测，以便最优地估计参数。

D. Shore-Johnson公理

Shore and Johnson（SJ）提出了一种熵泛函的替代公理化定义：在一组 I 信息片段的约束下对其进行最大化，应得到具有某些性质的概率分布$P = \circ I$ 。该定义由以下公理 [18]构成：

- SJ1（唯一性）：给定 I， $P = \circ I$ 是唯一的。
- SJ2（不变性）：如果 $\Gamma[\cdot]$ 是坐标变换（变量替换），那么 $\Gamma[\circ I] = \circ(\Gamma[I])$。
- SJ3（系统独立性）：给定两个独立系统 A 和 B，无论分别考虑有关它们的不同信息片段（以边缘概率表示），还是联合考虑（以联合概率表示），结果都不应有所不同。这意味着 $\circ(I_A \wedge I_B) = (\circ I_A)(\circ I_B)$，其中 $I_A \wedge I_B$表示关于 A 和 B 的现有信息片段 I_A 和 I_B 的并集。
- SJ4（子集独立性）：对于系统状态的一个独立子集，无论将其视为单独的条件密度，还是视为完整系统密度的一部分，结果都应相同。考虑将系统状态划分为互不相交的子集$\{\Lambda_k\}_k$ ，使得每个 k 都有一条可用信息 I_k ，且 $\bigcup_k \Lambda_k = \Omega$。此时，$(\circ I)_{\Lambda_k} = \circ I_k\ \forall k$，其中 $I = \bigwedge_k I_k$ 是总信息；而 $P_{\Lambda_k} = (p_{\Lambda_k}(G_1)\ldots, p_{\Lambda_k}(G\Omega))$，其中$p_{\Lambda_k}(G_i) = p(G_i|G_i \in \Lambda_k)$表示相对于子集 Λ_k的条件分布。
- SJ5（极大性）[1]：当没有可用信息时（$I = \varnothing$），$P = \circ I$ 为均匀分布P_u。

Shore and Johnson 声称，香农熵是唯一与他们的公理相容的推断泛函，这一说法暗示了 SK 公理与 SJ 公理等价。然而，后来人们澄清 [19, 20]，Shore and Johnson 的结论源于他们在推理中正式运用 SJ3 时无意间作出的一个额外隐藏假设。具体而言，他们考虑了这样一种情况：已知关于两个系统 A 和 B 的不同信息片段 I_A 和 I_B ，并暗示由此得到的联合概率分布分解为 $\circ(I_A \wedge I_B) = (\circ I_A)(\circ I_B)$，从而即使不能保证两个系统相互独立，也应用了 SJ3（关于两个系统只有互不重叠的信息片段，并不能保证这两个系统相互独立）

[1] 实际上，Shore 和 Johnson 仅隐含地定义了极大性公理。事实上，他们从最小交叉熵原理出发，将最大熵原理（MEP）引入为先验分布均匀时与之等价的原理。因此，尽管没有明确将其公理化，他们仍认为，在没有可用信息时，后验分布 P 应等于均匀分布（即与先验分布相同）。

。这一额外假设的存在意味着，只有当系统相互独立时，香农熵才确实是所需的泛函：若情况并非如此，那么得到的最大熵分布便不再是“对于缺失信息最大限度地不作预设”，而这正是杰恩斯的最大熵原理所要求的 [5], ，因为实际上并没有关于系统是否独立的“信息”。

E. 广义熵

Uffink [19] 指出，如果在重新审视 Shore and Johnson 的证明时，去除那个缺乏依据的额外假设，那么由 SJ 公理得到的熵并非唯一确定，而是构成一个完整的广义熵族 $S_q^{(f)}[P]$，其中每个熵都可表示为某个泛函 $U_q[P]$ 的任意递增函数 f。我们将这个泛函称为 Uffink 泛函，即

$$S_q^{(f)}[P] = f(U_q[P]), \quad U_q[P] = \left(\sum_{i=1}^{\Omega} p^q(G_i) \right)^{\frac{1}{1-q}} \tag{17}$$

其中参数为 $q > 0$。对于给定的 f，该熵族中的每一种熵都由参数 q标识，因此我们将其称为“熵参数”。请注意，熵参数与熵中其他结构参数所起的作用不同，例如上文讨论的香农情形中的 θ。显然，香农熵必定是该熵族中可能的一种，而事实上，取 $f(x) = \ln x$，可以证明

$$\begin{aligned} \lim_{q\to 1} S_q^{(\ln)}[P] &= \lim_{q\to 1} \ln U_q[P] \\ &= -\sum_{i=1}^{\Omega} p(G_i) \ln p(G_i) \\ &= S_1[P]. \end{aligned} \tag{18}$$

换言之，香农熵在形式上对应于$q = 1$，这也解释了式（1）中采用的下标（但请注意，第二节 B 中用于描述硬约束下微正则分布的均匀分布P_0 的下标与 的情形毫无关系 $q = 0$，后者是不允许的）。值得注意的是，Jizba和Korbel [20, 21] 证明，类型为 f（$U_q[P]$）的熵也可以从SK公理推导出来，前提是将$SK4$ 放宽为一种广义可分性条件，即用所谓的 Kolmogorov-Nagumo 和替代求和[2], ，该和此前已在广义算术的背景 [22, 23]。这表明，SJ 公理实际上等价于 SK 公理的一种特定推广。SK4 的推广数十年来一直是统计物理学文献中的讨论主题，因为它与非广延（或者更确切地说，非可加）热力学 [7] 这一主题相关。我们将形如 $f(U_q[P])$ 的任何熵称为 Uffink-Jizba-Korbel（UJK）熵。

通过放宽 SK 或 SJ 公理而得到的其他若干广义熵族也已有研究 [9, 10]。一个著名的例子是 Hanel 和 Thurner [24]提出的所谓 (c, d)-熵 $S_{c,\ d}[P]$，他们以迹形式（或更一般地说，可组合）熵的假设取代 SK4；也就是说，这类熵可以写成对系统状态 $\{G_i\}_{i=1}^{\Omega}$ 求和（或其函数）的形式。特别地，如果熵 $S(P)$ 可以写成某个函数 g 的求和形式 $\sum_{i=1}^{\Omega} g(p(G_i))$，那么它就是迹形式的。请注意，香农熵属于这一类，其中 $g(x) = -x\ln x$。更一般地，可组合熵可以写成这类求和的函数 h，即

$$S_{c,d}^{(h,g)}[P] = h\left(\sum_{i=1}^{\Omega} g\left(p(G_i)\right) \right), \tag{19}$$

其中，熵参数 (c, d) 由熵随可访问构型数量 Ω 的变化规律决定 [8, 24]。具体而言，考虑变换 $\Omega \to \lambda\Omega$、$\Omega \to \Omega^{1+a}$，并根据以下极限比值确定 c 和 d：

$$\lim_{\Omega\to\infty} \frac{S_{c,d}^{(h,g)}[(p(G_1), ..., p(G_{\lambda\Omega}))]}{S_{c,d}^{(h,g)}[(p(G_1), ..., p(G_{\Omega})]} = \lambda^{1-c}, \tag{20}$$

$$\lim_{\Omega\to\infty} \frac{S_{c,d}^{(h,g)}[(p(G_1), ..., p(G_{\Omega^{1+a}})]}{S_{c,d}^{(h,g)}[(p(G_1), ..., p(G_{\Omega})]} \Omega^{a(c-1)} = (1+a)^d. \tag{21}$$

对 h 和 g 的不同选择，可能会得到相同的熵参数值。在这种情况下，相应的熵被视为渐近等价 [8]。因此，此时熵参数标识出具有相同渐近性质的熵的等价类。我们将满足$SK1$-$SK3$以及式（19）的熵称为 Hanel-Thurner（HT）熵。

F. 如何识别正确的熵?

对于 UJK 熵、HT 熵，原则上也适用于任何广义熵族，重要的是确保最大熵原理能够一致地重新表述为一种

考虑一个双射 $f^{-1}: M \mapsto N \subset \mathbb{R}$，广义算术定义如下:

$$\begin{aligned} x \oplus y &= f(f^{-1}(x) + f^{-1}(y)), \\ x \ominus y &= f(f^{-1}(x) - f^{-1}(y)), \\ x \otimes y &= f(f^{-1}(x) f^{-1}(y)), \\ x \oslash y &= f(f^{-1}(x) / f^{-1}(y)). \end{aligned}$$

根据系统的观测构造概率分布的工具。这一过程有时称为广义最大熵原理（GMEP）。然而，目前仍有许多严重的概念和实践问题尚未解决。

首先，尽管在熵参数取定值时，仍有可能确定在某些“软”约束下使广义熵最大化的概率分布的函数形式，但一般而言，施加这些约束是否仍与对拉格朗日乘子应用最大似然原理相一致，以及熵是否仍像式（12）那样在模型选择中发挥作用，都不再有保证。只有某些广义熵能够恢复这种一致性，并非所有广义熵都如此，稍后我们将通过一些典型例子加以说明。最大似然原理不考虑概率分布的形式，更不关心该分布最大化的是哪种熵，因此这种不一致令人起疑。遗憾的是，文献中对其可能的成因讨论甚少。

其次，在确定熵参数本身，或者说确定参数族中的“正确”熵时，会出现根本性问题。具体而言，目前已有两种主要方法。一种方法要求 先验 了解系统（例如，熵或系统的某些性质如何随可访问构型数量变化 [8, 24, 25]），如式（20）和（21）所示；这意味着，如果缺乏此类知识，就无法像其他参数那样，仅凭数据一致地推导出熵参数。另一种方法则允许从数据中推断熵参数，同样诉诸某种形式的广义熵最大化 [26, 27]。然而，正如下文所示，如果将最大似然原理扩展到对熵参数本身的估计，这一要求就会与最大似然原理相冲突。

最后，当有关于系统的多个独立同分布观测可用时（例如在受控重复实验中），广义熵也应以某种方式与香农熵保持“一致”，因为在独立的情况下，按照 Shore 和 Johnson 最初的解释，公理 SJ3 确实应导出香农熵。那么，当只有单个观测可用时，同一系统如何能用非香农熵描述，而在有多个观测可用时又能用香农熵描述呢？据我们所知，文献中尚未讨论这一难题。

III. 一条统摄一切的公理

上述局限性导致 GMEP 在缺乏正确熵参数先验知识的情况下难以在实践中应用，或者与最大似然原理以及 *SJ3*在独立性条件下的信息论后果不一致。本节其余部分将展示我们所有的研究结果，并说明如何从一个看似不同的视角着手解决这一问题：即引入一条附加公理，以某种方式使给定族中的所有熵“保持一致”，从而仅凭数据（如果数据包含信息）就能选出该族中最可能的成员，而无需先验了解系统的性质。值得注意的是，这一简单要求的引入解决了第IIF节中讨论的所有不一致问题。

A. 无信息性公理

现在我们引入这条公理。不同于 SK 或 SJ 公理，这条公理并不适用于广义参数族中的某个单独的熵，而是适用于整个熵族。事实上，这条公理并不代表对 SK 或 SJ 公理的又一次推广，而是一项“辅助性”要求：恰恰在进行任何此类推广时加入该要求，以限制所得熵族的形式。

- 无信息性公理：在熵参数族中，均匀分布 P_u 所达到的熵值不应取决于熵参数的取值。

显然，如果将该公理应用于包含香农熵 S_1 作为特例的熵族，则意味着该族的所有成员在应用于 P_u时都达到相同的值$S_1[P_u] = \ln \Omega$。这一要求赋予广义熵一种通用尺度和含义。正如我们将在下文展示的，当推断过程扩展到对熵参数本身的识别时，我们的公理能够提供一定的保证。一方面，该公理确保无法从完全无信息的（即均匀的）分布中推断出任何熵参数，而不论参数估计程序是如何构建的。另一方面，当存在信息丰富的（非均匀）数据时，该公理确保与广义最大似然原则及模型选择方法保持一致：在没有关于系统的先验知识的情况下，包括熵参数在内的所有参数都可以从经验观测中得到识别。

请注意，根据 *SK2* 和 *SJ5*的要求，对于给定的 q，式（17）中的 Uffink 泛函在 P_u处取得最大值。这项要求采用一种“横向”视角，因为它对该参数族中的每一种 q-熵都成立。另一方面，我们的公理则提供了一种

“纵向”视角：应用于 P_u 时，所有 q-熵都不应被优先选择。换言之，该公理确保均匀分布发挥无信息性作用的范围不仅限于参数族中的某一种特定熵，而是涵盖其中所有熵。由于 SK2 和 SJ5 确保任何熵在 P_u 处的取值都是其上限，该公理建立了一种共同的参照框架或通用尺度，使得参数族中的不同熵能够得到一致比较。具体而言，它确保所有满足 SK2 或 SJ5 且将香农熵作为特例的参数族中的熵，无论熵参数取何值，取值都落在同一区间 $[0, \ln \Omega],$。我们将证明，这一保证意味着，只要输入数据具有信息性（分布不均匀），就可以纯粹根据输入数据，通过模型选择方法估计熵参数。

B. 对重要熵族的应用

现在，我们来讨论对常见熵族施加无信息性公理所带来的若干结果。

我们从 UJK 熵 $S_q^{(f)}[P]$ 开始，并要求该熵族将香农熵作为一个特例。熵 $S_q^{(f)}[P] = f(U_q[P])$ 在均匀概率分布 $P_u = (\Omega^{-1}, \ldots, \Omega^{-1})$ 上取值时，得到

$$S_q^{(f)}[P_u] = f\left(U_q[P_u]\right) = f(\Omega) \qquad \text{for} \quad q \neq 1. \tag{22}$$

我们的公理要求 $S_q^{(f)}[P_u]$ 与 q 相互独立，这意味着 f 应与 q 相互独立。对于 $q = 1$，严格来说，$S_q^{(f)}[P]$ 仅定义为极限

$$\lim_{q\to 1} S_q^{(f)}[P] = f\left(\lim_{q\to 1} U_q[P]\right), \tag{23}$$

这里我们用到了 f 的 q-独立性。如果我们要求当 $P = P_u$ 时，该极限与香农熵在 P_u 上的取值一致，即 $S_1[P_u] = \ln\Omega$，那么就需要一个满足以下条件的函数 f

$$\lim_{q\to 1} S_q^{(f)}[P_u] = f\left(\lim_{q\to 1} U_q[P_u]\right) = \ln \Omega, \tag{24}$$

即 $f\,(x) = \ln x$。因此，结合方程（22）和（24），我们得到 $f\,(x) = \ln x$ 对所有 q，也就是说，唯一可行的 UJK 熵是 R´enyi 熵 [28]

$$S_q[P] \equiv S_q^{(\ln)}[P] = \ln U_q[P] = \frac{1}{1-q} \ln \sum_{i=1}^{\Omega} p^q(G_i), \tag{25}$$

其中，由于上述熵是该族中唯一“幸存”的熵 $S_q^{(f)}[P]$，因此我们移除了所得结果 $S_q^{(\ln)}[P]$ 中的上标。根据式（18），我们可以确认，该熵在极限 $q \to 1$ 下退化为香农熵，这是 Rényi 熵的一个著名结果。对于均匀分布 P_u，该熵满足，

$$S_q[P_u] = \ln \Omega, \tag{26}$$

它不依赖于 q，正如我们的公理所要求的。因此，唯一可行的 UJK 熵是 R´enyi 熵。一般而言，其他 UJK 熵不符合我们的公理。

一个重要的反例是 Tsallis 熵 [29]，其定义为

$$S_q^{\text{Tsallis}}[P] \equiv S_q^{(\ln_q)}[P] = \frac{1}{1-q}\left(\sum_{i=1}^{\Omega} p^q(G_i) - 1\right) \tag{27}$$

它由所谓的“q-对数” $f(x) = \ln_q(x) \equiv (x^{1-q} - 1)/(1-q)$ 得到（不要与以 q 为底的 x 的普通对数混淆）：事实上，将该熵应用于 P_u 时，它取值为一个依赖于 q 的值

$$S_q^{\text{Tsallis}}[P_u] = \frac{\Omega^{1-q} - 1}{1-q} = \ln_q(\Omega). \tag{28}$$

从我们的公理来看，这种对 q 的依赖自相矛盾：不同的 q 值不应人为地赋予本质上没有信息的分布不同程度的信息量。从另一个角度看，这一矛盾源于函数 f 对 q 的依赖；该函数通过 Uffink 泛函 $U_q[P]$ 定义 Tsallis 熵：我们的公理不允许这种对 q 的依赖，因为 $f(U_q[P_u])$ 不应依赖于 q。请注意，定义 UJK 熵 $f(U_q[P_u])$ 的函数 f 对 q 的独立性，是我们公理带来的一个非平凡结果，因为只有在比较不同 q 值所对应的熵时，这种独立性才是必要的（如果只考虑单个 q 值，那么完全可以用该值 q 来指定 f）。具体而言，为了满足 $S_q^{\text{Tsallis}}[P_u] = S_1[P_u]$，我们的公理要求 $q = 1$；也就是说，唯一可行的 Tsallis 熵是香农熵。此处我们应强调，Tsallis 熵不被接受，并不与以下事实矛盾：在固定 q [7] 的情况下，使 Tsallis 熵最大化的分布已成功应用，因为该分布与使 Rényi 熵最大化的分布完全相同，也与任何其他 Uffink 泛函的单调函数最大化所得到的分布相同，正如我们将在本文后文讨论的那样。然而，当该分布“代回”熵中时，只有 Rényi 熵在不确定性的绝对量化以及相关的最大似然估计和模型选择程序方面给出一致的结果。事实上，正如下文所示，基于 Tsallis 熵对模型（或 q 的取值）进行排序，会扰乱基于最大似然的排序；而使用 Rényi 熵则能恢复并扩展其与最大似然原理的一致性。

作为另一个例子，我们将无信息性公理应用于可组合的 HT 族 (c, d)-熵

可写成式（19）中的形式。如果我们要求$S_c^{(h,g)}$，$_d\ [P_u] = S_1[P_u] =$ 类比式（26）中的 ln Ω，那么该公理将式（20）和（21）转化为：

$$\lim_{\Omega\to\infty}\frac{\ln\lambda\Omega}{\ln\Omega}=\lambda^{1-c} \tag{29}$$

$$\lim_{\Omega\to\infty}\frac{\ln\Omega^{1+a}}{\ln\Omega}=(1+a)^d \tag{30}$$

并意味着 $(c,d) = (1,1)$。这一参数选择确定了对独立事件具有可加性的熵的等价类。香农熵和 R´enyi 熵都属于此类。特别是，在 $h(x) = x$ (迹形式熵) 和 $g(x) = -x\ln x$(香农熵) 的情形下，可得 $(c,d) = (1,1)$ [8], 即$S^{(x,-x\ln x)}$ 1, $_1\ [P] = S_1[P]$。因此，香农熵是符合我们公理的可行迹形式 HT 熵。同样，在 $h(x) = \ln(x)/(1-q)$ 和 $g(x) = x^q$(R´enyi 熵) 的情形下，同样可得 $(c,d) = (1,1)$ [8], 即$S^{(\ln(x)/(1-q),x^q)}$ 1,$_1\ [P] = S_q[P]$。因此，R´enyi 熵是可行的可组合 HT 熵。相比之下，$h(x) = x$ 和 $g(x) = (x^q - \Omega^{-1})/(1-q)$（Tsallis 熵）的情形会得到 $(c, d) = (q, 0)$ [8],，这证实 Tsallis 熵（另一种迹形式熵）不符合我们的公理。

对于 UJK 和 HT 族而言，只有 Rényi 熵（或与之渐近等价的熵）能在我们的公理下“存续”这一事实，并不与熵可能具有非广延性相矛盾；过去几十年来，这种可能性促使人们引入了许多熵的变体 [7, 8]。事实上，我们的公理选择了独立系统的熵可加性（香农熵和 Rényi 熵也如此），但当独立性不存在甚至未知时，它并不会带来直接影响。特别需要强调的是，非广延性并非熵本身的性质，而是系统构型数 Ω 随物理尺度（即单元或粒子数）变化的标度性质 [8]。即使是香农熵，如果应用于 Ω（或存在约束时的 Ω_{C^*}）并非关于 n呈指数增长的系统，也可能是非可加的，这从式（2）或式（3）中可以清楚看出。注意，式（3）适用于微正则情形，但在正则情形下，熵也会呈现类似的非广延标度。随机图是这方面的一个重要例子：由 n 个顶点构成的所有二元图的数量为 $\Omega = 2^{\binom{n}{2}}$，因此呈超指数增长 [8, 30]。即使施加了各种约束 C^*，数量 Ω_{C^*} 仍然呈超指数增长，但香农熵依然适用于随机图系综 [13]。在另一个极端，即使对于 Ω 确实随 n呈指数增长的系统，也可能存在某些约束，使得 Ω_{C^*} 关于 n呈次指数增长，从而导致熵具有次广延性。状态空间缩减过程 [24]就是一个例子。因此无信息性公理所蕴含的首个一般性结果是：非广延性或非遍历性（若存在）应完全体现在 Ω_{C^*} 随 n变化的标度中，归根结底也就是体现在对适当（有效）约束 C^*的识别上，而不应体现在熵本身的表达式中。

C. 广义最大熵原理

在 GMEP 的语境下，我们的公理限制了广义熵的可行表达式，由此直接导致相应地限制了在软约束下使这类广义熵达到最大值的概率分布（请注意，在硬约束下，所有最大熵分布都会归结为第 P_0 II B 节中所述的微正则均匀分布）。这种限制可能产生两种（相关的）影响：一是最大熵分布的函数形式，二是该分布与熵本身及可能存在的其他量之间的关联方式。接下来我们将说明，HT 熵和 UJK熵是这两种影响的典型例子。

例如，使HT 熵 $S_c^{(h,g)}$ $(_d\ [P]$ 迹形式$h(x) = x$最大化的概率分布的一般形式，是所谓朗伯W函数[3] $\mathcal{W}(x)$ [8, 24], 的指数形式；但根据我们的公理，唯一可接受的形式对应于参数取值（c，d）$= (1, 1)$。取这组参数时，$\mathcal{W}(x)$ 函数会化为线性函数，因此最大熵概率便归结为式（6）中的 [8], 玻尔兹曼-吉布斯分布。这与我们的公理所允许的唯一迹形式HT熵是香农熵这一事实相符，正如我们上文所证明的那样。因此，要获得真正广义的最大熵概率，就应考虑非迹形式熵。

特别是，考虑我们的公理从´雷尼熵 $S_q[P]$中选出的熵，它同时属于UJK 和HT 两类，GMEP 可表述为第IIB节所述最大熵原理的一种著名推广，如下所示。给定一个系统未知微观状态 C^* 的某个（标量或向量）函数 $C(G)$ 的经验观测值 G，关于 G的偏差最小的推断由使 P_q 达到最大值、同时满足以下（软）约束的分布 $S_q[P]$ 给出：

$$\langle C\rangle_q \equiv \frac{\sum_{i=1}^{\Omega} p^q(G_i)\,C(G_i)}{\sum_{i=1}^{\Omega} p^q(G_i)} = C^*, \tag{31}$$

[3] 朗伯W函数 $\mathcal{W}(x)$ 无法用闭式表示，它是方程 $x = \mathcal{W}(x)e^{\mathcal{W}(x)}$ 的解。这里所关注的是实数解。

它推广了式 (4) 中通常的香农约束（注意，$\langle C\rangle_1 = \langle C\rangle$）。这里我们使用了记号 $[x]_+^a \equiv 0$ 若 $x<0$，而$[x]_+^a \equiv x^a$ 否则 [7]。注意，方程（36）的分母等于 Uffink 泛函 $U_q[P_q$（θ）]，也必须等于广义配分函数

$$W_q(\theta) \equiv \sum_{i=1}^{\Omega}[1-(1-q)\,\theta\cdot(C(G_i)-\langle C\rangle_q)]_+^{1/(1-q)} \quad (37)$$

因为 $p_q(G_i,\ \theta)$ 已通过式 (35中的条件归一化。换言之，

$$W_q(\theta) = \left[\sum_{i=1}^{\Omega} p_q^q(G_i,\theta)\right]^{1/(1-q)} = U_q[P_q(\theta)]. \quad (38)$$

最后，最大熵概率等于

$$p_q(G_i,\theta) = \frac{[1-(1-q)\,\theta\cdot(C(G_i)-\langle C\rangle_q)]_+^{1/(1-q)}}{W_q(\theta)} \quad (39)$$

其形式为所谓的 q指数 [7] 分布。请注意，方程（32）和（39）分别推广了方程（5）和（6）。还要注意，如果我们形式上引入一个伪状态~，使得 G (C) $G = \langle C\rangle_q$,~ 那么由方程（ 39 ）可得 $p_q(G,\theta) = 1/W_q$ (θ) =$1/U_q[P_q(\theta)]$。接着，由方程（38）可知:

$$p_q^{q-1}(\tilde{G},\theta) = \sum_{i=1}^{\Omega} p_q^q(G_i,\theta) = U_q^{1-q}[P(\theta)]. \quad (40)$$

~我们将在后文讨论 $C(G)$与$C(G^*)$之间的关系。

为对 $S_q[P]$ 进行约束最大化，我们求解 q-拉格朗日函数的导数为零的条件

$$\mathcal{L}_q[P] = S_q[P] - \alpha\left[\sum_{i=1}^{\Omega} p(G_i) - 1\right] - \theta\cdot[\langle C\rangle_q - C^*] \quad (32)$$

关于 P、 α 和 θ 的导数，并从现在起假设 $q\neq 1$。所得数值记为 P_q、 α_q、 θ_q。特别地，令 $\partial\mathcal{L}_q[P]/\partial P|_{P_q}=0$，我们得到

$$0 = \left.\frac{\partial\mathcal{L}_q[P]}{\partial p(G_i)}\right|_{p_q(G_i)} \quad (33)$$
$$= \frac{q}{1-q}\frac{p_q^{q-1}(G_i)}{\sum_j p_q^q(G_j)} - \alpha - q\,p_q^{q-1}(G_i)\frac{\theta\cdot(C(G_i)-\langle C\rangle_q)}{\sum_j p_q^q(G_j)}$$

对所有 i 从 1 到 Ω，由此显然可知 $p_q(G_i)$ 取决于 θ，如同 $q=1$的情形，此外还取决于q。对 $\mathcal{L}_q[P]$ 关于 α 求导，得到与式 (7) 相同的条件:

$$\left.\frac{\partial\mathcal{L}_q[P_q]}{\partial\alpha}\right|_{\alpha_q} = 0 \quad\Rightarrow\quad \sum_{i=1}^{\Omega} p_q(G_i,\theta) = 1, \quad (34)$$

可通过将式 (α_q 两边乘以式 (33，然后对 i求和来确定。由此得到

$$\alpha_q = \frac{q}{1-q} \qquad (q\neq 1), \quad (35)$$

它是方程（8）的对应式。将 α_q 代入（33），并单独提出 p_q（G_i），得到

$$p_q(G_i,\theta) = \frac{[1-(1-q)\,\theta\cdot(C(G_i)-\langle C\rangle_q)]_+^{1/(1-q)}}{\left[\sum_{j=1}^{\Omega} p_q^q(G_j,\theta)\right]^{1/(1-q)}} \quad (36)$$

当 $q\to 1$, $p_q(G_i,\theta)\to Z_1^{-1}(\theta)$ exp$(-\theta\cdot C(G_i))$ 时，可得到式 (6) 中的玻尔兹曼-吉布斯分布。当 $q\neq 1$时，q-指数函数与普通指数函数毫无关系，实际上具有幂律尾部，其在 $C(G_i)$ 取较大值时与 $C(G_i)^{1/(1-q)}$ 成正比。这些重尾广泛存在于若干现实世界的复杂系统中；它们的存在是 q-指数函数受到关注的原因之一，因为从某个合适的熵的最大化推导出这类函数显得方便而简洁 [7, 8]。文献中有些论述强调， q-指数函数是由式 (27) 所给的 Tsallis 熵最大化推导而来的。然而，它们实际上可以由式 (17) 中任意一种 UJK 熵推导得到：最大化 $U_q[P]$ 的分布也必然最大化 $f(U_q[P])$，对任意单调的 f 都成立。事实上，我们上文的推导从 Rényi 熵出发，这种推导也早已广为人知。UJK 熵族各成员之间的真正差异，在于将最大熵的 q-指数函数代回熵本身时才会显现。在这种情况下，无信息公理发挥着重要作用：它从该熵族中选出 Rényi 熵，作为能够解决第 II F 节所讨论的所有不一致问题的成员，正如我们稍后在

该论文。

接下来需要确定参数 θ。此时引入重新参数化很有帮助：

$$\psi(\theta) \equiv \frac{\theta}{1+(1-q)\,\theta\cdot\langle C\rangle_q}, \tag{41}$$

通过这一过程，可以（形式上）从 $p_q(G_i,\theta)$ 的表达式中移除 $\langle C\rangle_q$，并得到

$$p_q(G_i,\psi) = \frac{[1-(1-q)\,\psi\cdot C(G_i)]_+^{1/(1-q)}}{Z_q(\psi)} \tag{42}$$

其中，将 $\psi(\theta)$ 的逆记为 $\theta(\psi)$，

$$Z_q(\psi) \equiv \sum_{i=1}^{\Omega}[1-(1-q)\,\psi\cdot C(G_i)]_+^{1/(1-q)} \tag{43}$$

$$= \frac{W_q\left(\theta(\psi)\right)}{[1+(1-q)\,\theta(\psi)\cdot\langle C\rangle_q]^{1/(1-q)}} \tag{44}$$

是重新参数化的配分函数。注意，除非 $q\to 1$，否则 $Z_q(\psi) \neq W_q(\theta(\psi))$；在这种情况下，$\psi\to\theta$和 $W_1(\theta)\to Z_1(\theta)$。最优值 ψ_q 由以下条件确定

$$\left.\frac{\partial\mathcal{L}[P_q]}{\partial\psi}\right|_{\psi_q} = 0 \quad\Rightarrow\quad \frac{\sum_{i=1}^{\Omega} p_q^q(G_i,\psi_q)\,C(G_i)}{\sum_{i=1}^{\Omega} p_q^q(G_i,\psi_q)} = C^* \tag{45}$$

对应于公式 31中的预期要求，并将公式（9）推广到（ ）$q\neq 1$的情形。一旦通过上述条件确定了 ψ_q，即可将其代入公式（42），从而得到最终的最大熵概率分布 $P_q(\psi_q)$。

最后需要指出的是，如果约束普通均值 $\langle C\rangle$，而不是 q-均值 $\langle C\rangle_q$，并按照上述方法对 $S_q[P]$ 进行相同的最大化，则会得到不同的最大熵分布 $\hat{P}_q(\theta)$：

$$\hat{p}_q(G_i) = \frac{[1-(q-1)\,\hat{\theta}\cdot(C(G_i)-\langle C\rangle)]_+^{\frac{1}{q-1}}}{\hat{W}_q(\hat{\theta})}. \tag{46}$$

根据前面介绍的重新参数化，也可以写成：

$$\hat{p}_q(G_i,\hat{\psi}) = \frac{\left[1-(q-1)\,\hat{\psi}\cdot C(G_i)\right]_+^{1/(q-1)}}{\hat{Z}_q(\hat{\psi})}, \tag{47}$$

其中

$$\hat{\psi}(\hat{\theta}) \equiv \frac{\hat{\theta}}{1+(q-1)\,\hat{\theta}\cdot\langle C\rangle}. \tag{48}$$

请注意，变换 $q\to 2-q$ 在形式上将这两类约束联系起来。具体而言，可以看出

$$\hat{p}_q(G_i,\hat{\psi}) = p_{2-q}(G_i,\hat{\psi}). \tag{49}$$

D. 与最大似然原理和模型选择的联系

现在我们证明，由无信息性公理选出的熵恢复了与最大似然原理的一致性，并保留了模型选择的解释，与香农情形完全相同。其他熵并不能保证这两项性质。同时，我们还将说明如何处理关于同一系统的多个独立观测。

类比第 IIC 节，我们先考虑 $M=1$ 的情形，并定义参数 ψ_q 的最大似然估计程序如下：

$$\psi_q^* = \operatorname*{argmax}_{\psi}\ell_q(\psi), \quad \ell_q(\psi)\equiv\ln p_q(G^*,\psi). \tag{50}$$

要求 $\partial\ell_q(\psi)/\partial\psi|_{\psi_q^*}=0$，可得

$$\sum_{i=1}^{\Omega} C(G_i)\,p_q^q(G_i,\psi_q^*) = C(G^*)\,p_q^{q-1}(G^*,\psi_q^*) \tag{51}$$

并将两项同时除以 $\sum_G p_q^q(G^*,\psi_q^*)$，得

$$\langle C\rangle_q = \frac{C(G^*)\,p_q^{q-1}(G^*,\psi_q^*)}{\sum_G p_q^q(G^*,\psi_q^*)}. \tag{52}$$

人们可能会认为，上述等式右边与“期望的”值不同 $C^*=C(G^*)$，但事实并非如此。事实上，再次考虑一个满足 $\tilde{}$ $\tilde{}$伪状态 G 的条件 $C(G)$，$=\langle C\rangle_q$ 并利用式（40），我们可以将式（52）改写为

$$\frac{C(\tilde{G})}{C(G^*)} = \frac{1-(1-q)\,\psi_q^*\cdot C(\tilde{G})}{1-(1-q)\,\psi_q^*\cdot C(G^*)}, \tag{53}$$

$\tilde{}$由此得到 $C(G)=C(G^*)$。换言之，由式（ψ_q^*50）定义的值与由式（ψ_q45）定义的值相同，即 $\psi_q^*\equiv\psi_q$，也就是说

$$\left.\frac{\partial\ell_q(\psi)}{\partial\psi}\right|_{\psi_q^*} = 0 \quad\Rightarrow\quad \frac{\sum_{i=1}^{\Omega} p_q^q(G_i,\psi_q^*)\,C(G_i)}{\sum_{i=1}^{\Omega} p_q^q(G_i,\psi_q^*)} = C^* \tag{54}$$

与式（45）类似。这意味着，最大似然原理仍可视为与关于 ψ 的拉格朗日优化部分等价。此外，对式（40）两边取对数可得

$$\ell_q(\psi_q^*) = -S_q[P_q(\psi_q^*)], \tag{55}$$

这表明，对于 $M=1$，观测的对数似然等于负的 R´enyi 熵。这将式（12）推广到 $q\neq 1$的情形，进一步推广了这样一个结果：在模型选择框架中，可以根据最大化似然对不同模型进行排序，或者等价地，根据其实现的 R´enyi 熵进行排序。值得注意的是，

UJK 族的其他熵，包括 Tsallis 熵，并不具有这一性质。式（13）中的关系也可推广如下：

$$\mathcal{L}_q[P_q(\psi_q^*)] = S_q[P_q(\psi_q^*)] = -\ell_q(\psi_q^*), \tag{56}$$

将 $P_q(\psi_q^*)$ 取值时的拉格朗日函数与最大化对数似然联系起来。因此，至少到目前为止，R´enyi 熵似乎保留了 Shannon 熵的所有理想性质。

现在我们考虑 $M > 1$ 系统的独立同分布实现 $\{G_m^*\}_{m=1}^M$，由此得到 M 约束的独立观测 $\{C_m^*\}_{m=1}^M$，其中$C_m^* \equiv C(G_m^*)$ 对所有 m均成立。我们在第 IIC节已经看到，在这种情况下，确定如何组合 M 观测值的是最大似然原理，而不是最大熵原理。再次引入平均对数似然 $\overline{\ell}_q(\psi)$，则 ψ 的最大似然条件可直接推广式（14）：

$$\psi_q^* = \underset{\psi}{\operatorname{argmax}}\, \overline{\ell}_q(\psi), \quad \overline{\ell}_q(\psi) \equiv \frac{\sum_{m=1}^M \ln p_q(G_m^*, \psi)}{M}. \tag{57}$$

不难证明，要求 $\partial\overline{\ell}_q(\psi)/\partial\psi|_{\psi_q^*} = 0$ 就意味着：

$$\sum_{i=1}^{\Omega} C(G_i)\, p_q^q(G_i, \psi_q^*) = \frac{1}{M}\sum_{m=1}^{M} C(G_m^*)\, p_q^{q-1}(G_m^*, \psi_q^*) \tag{58}$$

或者等价地

$$\langle C\rangle_q = \frac{\sum_{m=1}^M C(G_m^*)\, p_q^{q-1}(G_m^*, \psi_q^*)}{M\sum_{i=1}^{\Omega} p_q^q(G_i, \psi_q^*)} \tag{59}$$

这将式（15）中的经典（$q = 1$）结果推广到了广义的非香农情形。因此我们得知，算术平均值不再是结合 M 可用观测以确定参数 ψ的最佳方式。事实上，回想一下，先验地说，我们甚至不知道生成这些 M 数值 $\langle C_m^*\}_{m=1}^M$ 的分布的一阶矩是否有限，摒弃算术平均值也就合情合理了。事实上，作为 GMEP 解的 q-指数分布在 $q \neq 1$下呈现幂律行为。因此，原则上，根据 q的取值，它们的所有矩都可能发散。假设q 事先未知，而是由推断过程本身确定（正如我们稍后所假设的那样），那么在有多个观测时，用算术平均值约束 q-均值就完全没有意义，因为当 $q > 3/2$时，该平均值在 $M \to \infty$ 极限下可能变为无穷大；而只要分布可归一化， q-均值按定义就是有限的。原则上，任何高阶矩 $\langle C^n\rangle$ 在 $n > 1$下也可能存在同样的问题；而根据式（42）计算的任何 q-广义矩 $\langle C^n\rangle_q$ 都会收敛，只要 q 满足使分布可归一化（这是该过程保持一致性的基本要求） [7]。由式（59）确定的最大似然估计量能够识别分布参数，而不论任何矩是否收敛。

由于 $\langle C\rangle_q$ 不再等于 M 观测的算术平均值，因此一个重要的后果是，一般而言，对于 $q \neq 1$ 和 $M > 1$,

$$S_q[P_q(\psi_q^*)] \neq -\overline{\ell}_q(\psi_q^*), \tag{60}$$

因此无法将方程（16）推广到情形 $q \neq 1$，也无法将方程（55）推广到情形 $M > 1$。类似地，方程（13）和（56）在此也无法推广。相反，仍然成立的关系是

$$S_q[P_q(\psi_q^*)] = -\tilde{\ell}_q(\psi_q^*), \tag{61}$$

其中，$\tilde{\ell}_q(\psi) \equiv \ln p_q(G, \psi)$ 是一种“伪似然”，涉及伪状态 G，使得 $C(G) = \langle C\rangle_q$ 如上所述。不幸的是，$\tilde{\ell}_q(\psi)$ 不再等于基于这些 M 观测的实际对数似然$\overline{\ell}_q(\psi)$。这是否意味着，当系统的同一量存在多个独立同分布观测时，对数似然与熵之间的对应关系便不复存在？对这个问题的回答，将在考察一个看似无关的问题——熵参数 q最优值的选择——时显现，并将在下文给出。

E. 熵参数的推断

现在我们来谈谈无信息公理所蕴含的最后一项益处；从许多方面来看，这也是最关键的一项，即能够仅凭数据一致地识别熵参数，而无需预先假定关于系统的知识——例如式（20）和（21）所示类型的标度律 [8, 24, 25]。

为此，我们从一般情形$M \geq 1$直接出发，再次运用最大似然原理，并以式（54）证明该原理在估计最大熵分布的其他参数时恢复了一致性为基础，将其扩展至对熵参数本身的识别。事实上，最大似然原理对待任何参数都一视同仁，不赋予其特定解释，因此并不知道q 与其他结构参数在信息论框架中扮演着不同角色。再次将 Rényi 熵视为来自 UJK和 HT 两类熵中唯一可行的熵，应用于熵参数 q 的最大似然原理可正式表述如下：

$$q^* = \underset{q}{\operatorname{argmax}}\, \overline{\ell}_q(\psi), \quad \overline{\ell}_q(\psi) \equiv \frac{\sum_{m=1}^M \ln p_q(G_m^*, \psi)}{M}. \tag{62}$$

另一方面，将上述表达式与式（57）结合起来，可以清楚地看出，对 q 的估计与

ψ的估计相耦合，因此扩展最大似然原理的实际表述为

$$(\psi^*_{q^*}, q^*) = \underset{(\psi,q)}{\operatorname{argmax}} \overline{\ell}_q(\psi), \quad \overline{\ell}_q(\psi) \equiv \frac{\sum_{m=1}^{M} \ln p_q(G^*_m, \psi)}{M}. \tag{63}$$

这个表达式立即告诉我们，一旦将最大似然原理扩展到确定 q，我们在第IIID 节讨论的结果便只代表了问题的一面。现在，若要求两者同时满足

$$\left.\frac{\partial \overline{\ell}_q(\psi)}{\partial \psi}\right|_{(\psi^*_{q^*}, q^*)} = 0, \quad \left.\frac{\partial \overline{\ell}_q(\psi)}{\partial q}\right|_{(\psi^*_{q^*}, q^*)} = 0, \tag{64}$$

我们再次得到式（59）（其中 q 替换为 q^*），并附加条件

$$\sum_{i=1}^{\Omega} p_{q^*}(G_i, \psi^*_{q^*}) \ln\left[1 - (1-q^*)\,\psi^*_{q^*} \cdot C(G_i)\right] \tag{65}$$
$$= \frac{1}{M} \sum_{m=1}^{M} p_{q^*}(G^*_m, \psi^*_{q^*}) \ln\left[1 - (1-q^*)\,\psi^*_{q^*} \cdot C(G_m)\right].$$

回顾式（42，可知

$$1-(1-q^*)\,\psi^*_{q^*} \cdot C(G_i) = [p_{q^*}(G_i, \psi^*_{q^*}) Z_{q^*}(\psi^*_{q^*})]^{1-q^*} \tag{66}$$

我们得到条件

$$\sum_{i=1}^{\Omega} p_{q^*}(G_i, \psi^*_{q^*}) \ln p_{q^*}(G_i, \psi^*_{q^*}) = \frac{1}{M} \sum_{m=1}^{M} \ln p(G^*_m, \psi^*_{q^*}). \tag{67}$$

换言之，用于确定q^* 的附加最大似然条件要求最大化对数似然等于负的香农熵，即

$$S_1[P_{q^*}(\psi^*_{q^*})] = -\overline{\ell}_{q^*}(\psi^*_{q^*}), \tag{68}$$

这恢复了与式（16）的类比。此前在考察 $q \neq 1$时，这种类比似乎消失了，取而代之的是式（61）。实际上，我们现在认识到，当最大似然原理扩展到 q时，与式（16）的对应关系并未被取代，而是伴随着式（61）一同出现。值得注意的是，在特定参数值（$\psi^*_{q^*}$，q^*）处，香农熵与对数似然之间的联系仍是一个普遍结果，即使对于 $q \neq 1$ 和 $M > 1$也是如此。这或许看起来相当令人惊讶，因为对于 $q \neq 1$，对数似然基于使 R $q$ényi 熵而非香农熵达到最大值的′-指数分布。

尽管这一结果令人意外，但它完全合乎情理，因为我们假设了 M 独立观测。实际上，这解决了我们在第IIF节指出的最后一个不一致之处：假设观测相互独立，为 Shore and Johnson 对公理$SJ3$ 最初的有限解释提供了依据，并使香农熵成为数据不确定性的度量。事实上，式（60）中的不等式应与我们最初关于系统独立性的公理$SJ3$ 的讨论联系起来。请回想一下，假设这些 M 值$\{C^*_m\}_{m=1}^M$ 来自独立观测，等价于假设存在 M 同一系统的多个相同且相互独立的副本，每个副本恰好被观测一次。在这一独立性假设下，Shore and Johnson 最初的推理便适用，因此应当认为，描述由 M 个副本组成的组合系统的恰当熵是香农熵，而非 Rényi 熵。因此，平均对数似然与 Rényi 熵之间对应关系的失效，可以视为观测相互独立这一假设的一个征兆。当 $M = 1$时，我们可以使用式（55），并将其与式（68）结合，得到

$$S_{q^*}[P_{q^*}(\psi^*_{q^*})] = S_1[P_{q^*}(\psi^*_{q^*})] \tag{69}$$

这表明，在这一特定情况下，最大熵概率分布给出的香农熵与 Rényi 熵数值相同，尽管它最大化的是后者而非前者。对于 $M > 1$，这一结果通常并不成立。

式（68）中的这一显著结果对模型选择具有重要意义。具体而言，为了确定 q^* 和 $\psi^*_{q^*}$，可以考虑 q 的一系列取值，并针对范围内的每个取值，根据式（59）计算 ψ^*_q 。这样，对于 q 的每个取值，都会得到一个仅部分最大化的对数似然 $\overline{\ell}_q(\psi^*_q)$；所谓部分最大化，是指仅针对 ψ_q进行了最大化，尚未针对 q进行最大化。然后，可以从所有这些部分最大化的对数似然中选出数值最大的一个。由此便能确定取值 q^* 及其对应的 $\psi^*_{q^*}$，二者最终对应于完全最大化的对数似然 $\overline{\ell}_{q^*}(\psi^*_{q^*})$。只有在这一参数选择（$q^*, \psi^*_{q^*}$）下，对数似然才等于负的香农熵。因此，香农熵从最大似然条件中自然涌现：虽然概率 P_q 最大化的是 Rényi 熵而非香农熵，但在模型选择中，后者才是考虑独立性的正确熵。因此，引入我们的公理会带来一种基于熵的模型选择准则：通过最大化 Rényi 熵来确定概率分布的函数形式，并运用最大似然原理估计其参数，包括熵参数。为说明上述方法的性能，下面将考察两个简单的数值示例。

我们的第一个例子是一个由可观测量 $C(G)$ 描述的系统，该可观测量只取正实数值，即 $C(G) \in [0, +\infty)$。此外，我们假设对于所有 C，$\Omega_C = 1$，这意味着对于可观测量的每个取值 $C(G)$，只有一个状态 G 能够实现该取值。因此，对系统状态的求和可简化为对可观测量的积分

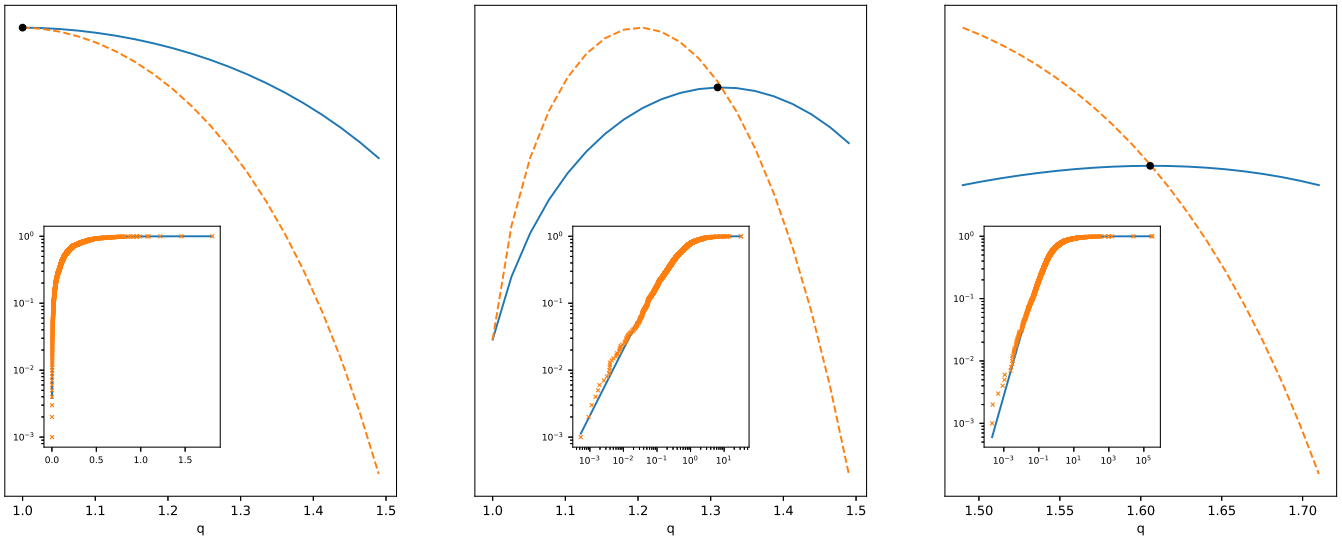


图 1. 对于由式 (70) 中的概率分布 $P_q(\psi)$ 生成的三组 $M = 10^3$ 偏差样本，平均部分最大化对数似然 $\overline{\ell}_q(\psi_q^*)$（实线）与负香农熵 $-S_1[P_q(\psi_q^*)]$（虚线）随 q 变化的比较，具体包括：$q_{\rm true} = 1$ 和 $\psi_{\rm true} = 5.0$ 时的指数分布（左），一阶矩有限且 $q_{\rm true} = 1.3$ 和 $\psi_{\rm true} = 3.0$ 的 q- 指数（幂律）分布（中），以及一阶矩发散且 $q_{\rm true} = 1.6$ 和 $\psi_{\rm true} = 7.0$ 的 q- 指数（幂律）分布（右）。插图比较了 M 个实现值的经验累积分布（叉号）与使用推断值 $(q^*, \psi_{q^*}^*)$ 得到的最大熵分布（实线）。

值：$\sum_G \to \int_0^\infty dC(G)$。由 GMEP 得到的概率分布为：

$$p_q(G_i, \psi) = (2-q)\,\psi\,[1-(1-q)\,\psi\cdot C(G_i)]_+^{\frac{1}{1-q}}, \quad (70)$$

其中我们使用了 $Z_q(\psi) = 1/(2-q)\psi$。对于不同的 ψ 和 q，我们从上述分布中抽取了包含$M = 10^3$ 个实现值的独立同分布样本，旨在仅根据由此生成的数据推断这些参数的真实值。具体而言，我们生成了来自指数分布（即 $q_{\rm true} = 1$）、具有有限一阶矩 $\langle C\rangle$ ($q_{\rm true} = 1.3$) 的 q- 指数分布，以及一阶矩发散的 q- 指数分布（$q_{\rm true} = 1.6$）的样本。图1针对这三种情况，展示了作为 q 函数的 $\overline{\ell}_q(\psi_q^*)$（蓝线）和 $-S_1[P_q(\psi_q^*)]$（橙线）。黑点表示两条曲线的交点，它确定了满足方程（68）的估计值 q^*。参数的真实值及其推断出的最大似然估计 $(q^*,\ \psi_{q^*}^*)$ 列于表I。由于左图对应 $q_{\rm true} = 1$，因此它是标准指数分布。在这种情况下，两条曲线仅在 $q = 1$ 处相交。相比之下，另外两种情况对应于 $q_{\rm true} \neq 1$，两条曲线在两点相交，即 $q = 1$ 和 $q = q_{\rm true}$。在这些情况下，两个交点都是方程（68）的解，但解 $q \neq 1$ 对应更高的对数似然（和更低的熵）。这个例子非常简单，却很有说明性：它直接展示了香农熵如何在模型选择中发挥作用，即使所采用的分布来自 GMEP，并且使 Rényi 熵最大化，而非香农熵。我们再次强调，在最后一种情况下，约束通常意义上的均值而不是 q-均值并不合适，因为对于 $q > 1.5$，通常意义上的均值会随着 $M \to \infty$ 而发散；相反，使用 q-平均值，就能够一致地刻画原始的无限均值幂律分布。

第二个也是最后一个例子是一个简单的系统，它由伯努利随机变量 $C(G)$ 表征，该变量取值为 $C(G) = 1$，其真实概率为 $p_{\rm true}$；取值为 $C(G) = 0$，其概率为 $1 - p_{\rm true}$。对 q 平均值施加约束可得

$$p_q(G_i, \psi) = \frac{[1-(1-q)\,\psi\cdot C(G_i)]_+^{1/(1-q)}}{1+[1-(1-q)\,\psi]^{1/(1-q)}}. \quad (71)$$

现在，我们将 $p_q(\psi)$称为在$C(G) = 1$ 时的概率 $p_q(G, \psi)$，并将 $1 - p_q(\psi)$称为在$C(G) = 0$时的概率 $p_q(G, \psi)$。很容易验证

$$\langle C\rangle = p_q(\psi) \quad (72)$$

表I. 真实参数值与最大似然估计的比较。

$q_{\rm true}$	$\psi_{\rm true}$	q^*	$\psi_{q^*}^*$
1.0	5.0	1.0	5.0
1.3	3.0	1.3	2.9
1.6	7.0	1.6	7.3

and

$$\langle C\rangle_q = \frac{p_q^q(\psi)}{p_q^q(\psi) + [1 - p_q(\psi)]^q}. \tag{73}$$

现在，若考虑 M 独立同分布实现 $\{C_m^*\}_{m=1}^M$ 的C并应用式（58），则得到 $p_{q^*}^{q^*}(\psi_{q^*}^*) = p_{q^*}^{q^*-1}(\psi_{q^*}^*)$ f_1其中 $f_1 = \sum_{m=1}^M C_m^*/M$ 是所观测到的、满足 $C_m^* = 1$ 的实例的经验频率。该关系显然可化简为

$$f_1 = p_{q^*}(\psi_{q^*}^*). \tag{74}$$

由于满足最大似然条件、且产生完全相同的最大化对数似然的 $(\psi_{q^*}^*, q^*)$ 组合有无穷多个，因此无需优先选择其中任何一个。根据我们的方法，可以得到一个令人想起香农情形的结果：对于伯努利随机变量，最大熵分布的参数必须设置为使估计概率与经验频率相匹配。任意取 q 都可以做到这一点，因此这是一个退化情形：无法从数据中学习到 q 的特定取值，因为由此得到的最大熵分布彼此完全相同。这并不意外：事实上，我们在这里所做的，实际上是试图用一个依赖两个参数的分布，来捕捉一个只有一个参数的二元随机变量的性质。

IV. 结论

大量文献探讨了通过放宽（或不受限制地解释）SK 和 SJ 公理中的某些公理（尤其是 SK4 和 SJ3）而得出的广义熵公理化定义。众所周知，按这种方式推广后，熵的定义会导出参数化的熵族，其中熵参数取某个特定值时，通常会得到普通的香农泛函。在最大熵方法中，每个熵族都会导出相应的最大熵概率分布族，这些分布同样由熵参数索引；对于仅掌握有限信息的系统，这些分布能够提供偏差最小的推断，而有限信息的形式是对某个被视为软约束的量所作的经验观测。遗憾的是，当把估计得到的最大熵分布“代回”其定义所依据的广义熵时，通常会出现若干不一致之处，包括与最大似然原理不相容、无法仅凭经验数据确定熵参数的取值，以及在能够获得同一系统的多个独立观测时与香农熵脱节。

本文基于这样一个事实：熵族中的每个成员，归根结底都是为了量化输入概率分布所编码的不确定性。我们提出了一条无信息性公理，要求不确定性最大的（即均匀）概率分布始终返回相同的熵最大值，而不受熵参数取值的影响。这条简单的公理意味着所有熵的取值都落在同一区间 $[0, \ln \Omega]$, 内，其中 Ω 是系统可能的（无约束）微观状态数，从而赋予广义熵通用的尺度和含义。该公理大幅限制了熵族中可接受的成员。特别是，对于 UJK 熵和 HT 熵，该公理只选出 R´ényi 熵作为可行熵。Tsallis 熵则是一个显著的反例，因不符合该公理而被排除。从推断视角来看，该公理保证完全无信息的数据（或等价地，完全缺乏经验信息）不能用于学习熵参数的取值。同时，我们还表明，在有信息数据可用时，对最大似然原理作直接扩展，就能仅凭经验观测、无需作出任何假设，最优地估计熵参数。

由此得到的广义最大似然方法，将熵参数的确定与最大熵分布中其他结构参数（拉格朗日乘子）的确定结合起来。具体而言，拉格朗日乘子的最大似然条件指出，应将哪种特定的 M 独立观测组合设为约束的广义平均值；而熵参数的最大似然条件则恰好等同于要求数据的对数似然等于负的香农熵。这一显著结果表明，香农熵与对数似然之间的联系对广义熵（在熵参数取适当的最大似然值时）同样成立，并且与假定的 M观测独立性一致。当 $M = 1$时，最大熵概率给出的 Rényi 熵与香农熵数值相同，尽管它最大化的是前者而非后者。对于多个独立观测（$M > 1$），对数似然与香农熵之间的联系依然成立，而与 Rényi 熵之间的联系则因独立性而消失。因此，在同时对熵参数进行最大化时，对数似然会自动找到适用于模型拟合与选择的正确熵。

我们认为，引入无信息公理有益于统计推断及其众多应用，并为构建仍具有可控且一致性质的广义熵提供了一种途径。

[1] R. Clausius, 《伦敦、爱丁堡和都柏林哲学杂志与科学杂志》12, 241 (1856)。[2] L. Boltzmann, *Über die Beziehung zwischen dem zweiten Hauptsatze des mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht* (Kk Hof-und Staatsdruckerei, 1877)。[3] J. W. Gibbs, *Elementary principles in statistical mechanics* 2014 (Courier Corporation,)。[4] C. E. Shannon, 《贝尔系统技术杂志》27, 379 (1948)。[5] E. T. Jaynes, 《物理评论》106, 620 (1957)。[6] K. P. Murphy, *Probabilistic machine learning: an introduction* 2022 (MIT press,)。[7] C. Tsallis, *Introduction to nonextensive statistical mechanics: approaching a complex world* (Springer Science & Business Media, 2009)。[8] S. Thurner, R. Hanel, and P. Klimek, *Introduction tothe theory of complex systems* (牛津大学出版社, 2018)。[9] J. M. Amig´o, S. G. Balogh, and S. Hern´andez, 熵 20, 813 (2018)。[10] A. M. Lopes and J. A. T. Machado, 熵22, 1374 (2020)。[11] A. Khinchin, 纽约 (1957)。[12] J. Karmeshu, *Entropy measures maximum entropy principle and emerging applications* 119 (Springer Science & Business Media, 2003)。[13] T. Squartini and D. Garlaschelli, *Maximum-Entropy Networks: Pattern Detection Network Reconstruction and Graph Combinatorics* 2017 (施普林格,)。[14] D. Garlaschelli and M. I. Loffredo, 《物理评论 E》 78, 015101 2008( )。[15] D. Garlaschelli and M. I. Loffredo, 《物理评论E》78, 015101 2008( )。[16] D. R. A. Kenneth P. Burnham, *Model Selection and Multimodel Inference* (施普林格，纽约，NY，2002). [17] P. D. Gr¨unwald、I. J. Myung 和 M. A. Pitt, *Advances in minimum description length: Theory and applications* (MIT 出版社，2005). [18] J. Shore 和 R. Johnson，《IEEE 信息论汇刊》26，26（1980）. [19] J. Uffink，《现代物理学史与哲学研究》26，223（1995）. [20] P. Jizba 和 J. Korbel，《物理评论快报》122，120601（2019）. [21] P. Jizba 和 J. Korbel，《物理评论E》101，042126（2020）. [22] A. N. Kolmogorov 和 G. Castelnuovo, *Sur la notion de la moyenne* (G. Bardi，R. Accad. dei Lincei 印刷所，1930). [23] M. Nagumo，载于 *Japanese journal of mathematics: transactions and abstracts* 7 ，第卷（日本数学会，1930）第 71–79 页。[24] R. Hanel 和 S. Thurner，《欧洲物理快报》（EPL）93，20006（2011）. [25] S. G. Balogh、G. Palla、P. Pollner 和 D. Cz´egel，《科学报告》10，1（2020）. [26] A. Plastino、H. Miller 和 A. Plastino，《连续介质力学与热力学》16，269（2004）. [27] A. Bashkirov，《物理评论快报》93，130601（2004）. [28] A. R´enyi，载于 *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (加利福尼亚大学出版社，1961) 第 547–561 页。[29] C. Tsallis，《统计物理学杂志》52，479（1988）. [30] Q. Zhang 和 D. Garlaschelli，《新物理学杂志》24，043011（2022）. [31] C. Beck 和 F. Sch¨ogl, *Thermodynamics of chaotic systems: an introduction*，4（剑桥大学出版社，1995）。