arXiv:2603.20274v1 [cs.FL] 17 Mar 2026

所罗门诺夫归纳

TOM F. STERKENBURG

Abstract. 本章探讨所罗门诺夫的通用预测方法。该方法的关键要素是可计算性这一概念；我将其核心思想阐述为一种尝试，旨在满足通用预测器的两项看似合理的可计算性要求。但这一尝试并不成功，Putnam 提出的对角化论证经推广后表明了这一点。随后，我批判性地讨论了该方法所声称的优点，尤其是它为奥卡姆剃刀这一方法论原则提供基础，以及充当机器学习方法发展中的理论理想。

1. Introduction

关于归纳问题的“贝叶斯主流观点”认为，归纳推理的正当性是有条件的或相对的（Huttegger，2017年，第3页）。这一主观主义观点可追溯至 Ramsey、de Finetti 和 Savage；它认为，理性要求信念程度保持一致，并通过条件化进行推理，但对根本性的出发点——先验概率分布——不作任何限制。在归纳推理的语境中，先验体现了我们的归纳假设，而这些假设最终无法在任何绝对意义上得到正当化。尤其是，这些假设必然具有局限性：在某些情形下足够妥当，在另一些情形下却不然（Howson，2000）。从这个意义上说，不存在“通用”的先验。[1]

同样，机器学习中的一个基本教训是，不存在“免费午餐”（Sterkenburg 和 Gr¨unwald，2021年）。每种学习算法都必须带有严格的假设，或归纳偏置；这些假设能使算法在某些可能情形下成功学习，却会使其在另一些情形下无法学习（Shalev-Shwartz 和 Ben-David，2014；Ben-David 等人，2011）。从这个意义上说，不存在“通用”的学习方法，无论它是否基于贝叶斯主义。[2]

然而，理论计算机科学中有一种方法恰好有望做到这一点。其支持者认为，Solomonoff (1964) 提出的方案或许能以如下形式，为归纳问题“提供严谨而令人满意的解决方案”：一种

Date：2026年3月24日。这是一个初步版本，欢迎反馈。[1]这份对主流观点的概述不可避免地忽略了许多尚存的争议。例如，并非所有作者都像 Howson 那样，坚持认为贝叶斯主义是一种逻辑，并包含可靠性与完备性定理（Howson，2000，第7章）；而 Howson 本人实际上并不认为动态一致性（即通过条件化进行更新）是贝叶斯主义的核心原则（同上，第135页），这与Huttegger (2017年，第1.4节）的看法相反。此外，这种观点本身也存在争议：deCanson (2025) 批评了归纳条件性论证的历史渊源及其成功程度。本章无须讨论这些问题；相关的要点是，不存在通用先验。[2]同样，这份概述忽略了许多细节，尤其是，对无免费午餐结果作出精确陈述，需要明确指定学习问题和成功学习的概念；而如果这些规定足够宽松，我们或许就能得到免费午餐（Sterkenburg 和 Gr¨unwald，2021年，第2.5节）。

基于“通用先验”的贝叶斯学习者（Li 和 Vit´anyi，2019，第 354 页）。[3] 近年来，*Solomonoff* 归纳这一方法再次受到了一些关注，因为人们经常将其视为解开现代深度学习之谜的潜在拼图之一（Grau-Moya 等人，2024年；Del´等人，2024年；Goldblum 等人，2024年；Mingard 等人，2025；Young 和 Witbrock，2025；Wilson，2025）。

本章将更深入地探讨 Solomonoff 归纳。我首先在第 2 节中，借助可计算性的形式约束，介绍定义通用贝叶斯预测方法这一关键思想。我指出，这类通用预测器在两项可计算性要求之间存在根本冲突；Putnam 在批评 Carnap 的归纳逻辑时，已经揭示了这一冲突。接着，在第 3 节中，我介绍 Solomonoff（以及 Levin (1971)，他们为这一方法奠定了严谨的数学基础）提出的方案，该方案试图克服这一冲突。我将讨论若干保留意见，涉及这一方案是否成功，以及由此产生的定义所声称带来的（概念和实践上的）优势。在第4节中，我将就这对机器学习中使用 Solomonoff 归纳以及贝叶斯主流观点有何启示，作一些总结性评论。[4]

2. Universal prediction

2.1. 序列预测。 我在此考察的基本归纳学习问题，是对不断生成的二元序列进行反复的一步预测。在每个时刻 t，我们都有一个有序序列 $\boldsymbol{x}^t = x_1 \dots x_t$，其中包含迄今为止观测到的结果 $x_i \in \{0, 1\}$，并试图预测下一个结果 x_{t+1}。我们的预测以对两种可能的下一结果所作的概率分布表示。预测器是一条用于作出此类预测的精确规则：形式上，它是一个函数 $\mathsf{p} : \{0,1\}^* \to \mathcal{P}$，其定义域为所有有限二进制序列，值域为$\{0, 1\}$上的所有分布。我也会将 p 写成二元函数，其中 $\mathsf{p}(\boldsymbol{x}, x)$ 表示 p 在给定观测序列 $\boldsymbol{x}$ 的情况下，对下一结果 x 所预测的概率。[5]

2.2. 普适可靠性。 预测器要具备通用性意味着什么？所罗门诺夫归纳的核心思想是，我们可以从形式可计算性的概念中找到答案。

2.2.1. *可计算性与可靠性。* 如果某个预测器最终能对所有可计算模式给出正确的预测，我们就可以称它为通用预测器，或称其具有通用可靠性。如果 $\boldsymbol{x}^\omega$ 是一个无限二元序列，那么预测方法 p 在$\boldsymbol{x}^\omega$ 上收敛，是指它给出的概率收敛到正确的下一位；也就是说，在下一位为 1 时，预测为 1 的概率收敛到 1；在下一位为 0 时，预测为 1 的概率收敛到 0：

$$\mathsf{p}(\boldsymbol{x}^t, 1) \xrightarrow{t\to\infty} x_{t+1}. \tag{1}$$

[3]类似地，所罗门诺夫归纳通过提出一种“令人信服的理论基础，用于构建 [.....] 理想的通用预测系统”，从而“可以说解决了百年归纳问题”（Grau-Moya 等人，2024年，第1页）。[4]本章大部分内容取材于（Sterkenburg，2018年），其中还可找到更多历史、数学和哲学方面的细节。不过，本章并非只是对该著作的概述：我对结构和重点都作了更新。[5]当然，还可以研究其他形式学习框架，包括限制更多的框架，例如预测器只给出类别式的（0 或 1）预测（Blum和Blum，1975；另见 Belot，2020），以及与序列预测截然不同的学习问题。当前的框架非常一般；只要其他学习框架能够被“编码回”其中，它甚至也可以声称自身是真正“通用”的（参见 Sterkenburg，2018，第I.1节）。

那么，如果某个预测器在所有图灵可计算序列上都收敛，我们就可以称它具有通用可靠性。

但我们仍然可以做得更加一般。也就是说，我们可以考虑结果空间上的所有可计算的概率测度。[6] 无限二进制结果序列上的概率测度由一个函数诱导，μ 该函数定义在所有有限序列上，并满足

$$\begin{aligned} \mu(\varnothing) &= 1; \\ \mu(\boldsymbol{x}0) + \mu(\boldsymbol{x}1) &= \mu(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \{0,1\}^*, \end{aligned} \tag{2}$$

其中，$\varnothing$ 是空序列，而 $\mu(\boldsymbol{x}x)$是 $\boldsymbol{x}$ 与比特 x的串联。（我们总是从空序列开始；对于 $\boldsymbol{x}$ 的两个可能的一比特扩展，观察到任意一个的概率正是观察到 $\boldsymbol{x}$的概率。）例如，均匀测度 λ 由 $\lambda(\boldsymbol{x}^t) = 2^{-t}$ 给出（每个长度相同的序列具有相同的概率）。

一个可计算的概率测度仅由一个图灵可计算的 μ给出。现在，对于给定的测度 μ，当以μ-概率1时，预测器 p 的预测收敛到 μ的“真实”概率时，我们称预测器 p 对 μ 是可靠的，

$$\mathsf{p}(\boldsymbol{x}^t, x_{t+1}) \xrightarrow{t\to\infty} \mu(x_{t+1} \mid \boldsymbol{x}^t). \tag{3}$$

如果一个预测器对所有可计算的测度都是可靠的，那么我们就可以称其为普适可靠。请注意，这包括此前在可计算序列上“确定性”收敛的情形，因为一个可计算序列可由一个将全部概率集中于该单一序列的可计算测度给出。

2.2.2. 通用混合。理论上，实现这种普适可靠性并不十分困难。首先考虑以下一般性定义。

定义 1（贝叶斯混合）。设 $\mathcal{H} = \{\mu_i\}_{i\in\mathbb{N}}$ 为一个可数的测度类。设 权重函数 w: $\mathbb{N} \to (0,1]$ 满足 $\sum_{i\in\mathbb{N}} w(i) \leq 1$。$w$ 关于 $\mathcal{H}$ 的贝叶斯混合测度定义为

$$\xi_w^{\mathcal{H}}(\boldsymbol{x}) := \sum_{i\in\mathbb{N}} w(i)\mu_i(\boldsymbol{x}), \tag{4}$$

并进而定义一个混合预测器：[7]

$$\mathsf{p}_w^{\mathcal{H}}(\boldsymbol{x}, x) := \xi_w^{\mathcal{H}}(x \mid \boldsymbol{x}) = \frac{\xi_w^{\mathcal{H}}(\boldsymbol{x}x)}{\xi_w^{\mathcal{H}}(\boldsymbol{x})}. \tag{5}$$

我们有如下的一般可靠性结果。

定理 2（贝叶斯一致性）。对于混合预测器 $\mathsf{p}_w^{\mathcal{H}}$，对所有 $\mu \in \mathcal{H}$，以 μ-概率 *1*，$p_w^{\mathcal{H}}(\boldsymbol{x}^t, x_{t+1}) \xrightarrow{t\to\infty} \mu(x_{t+1} \mid \boldsymbol{x}^t)$。

证明。这直接由布莱克韦尔–杜宾斯定理（1962）得出，因为每个 $\mu \in \mathcal{H}$ 都被混合测度 ξ_w 支配（即，对每个事件 A，如果$\mu(A) > 0$ 那么 $\xi(A) > 0$也成立）。□

[6]下文中，我会略过若干形式细节。有关康托尔空间（即无限二进制序列构成的空间）上的（可计算）概率测度的恰当定义，请参见（STERKENBURG，2018年，第 2.1.1 节）。[7]一般而言，预测器与测度之间存在形式上的对应关系，或者至少与那些对每个有限序列都赋予正概率的测度存在这种对应关系（参见STERKENBURG，2018年，第 3.1.1 节）。

现在请注意，由于图灵机有可数多个，所有可计算测度组成的类 $\mathcal{H}^{\text{comp}}$是可数的。因此，存在一个对所有可计算测度的枚举 $\{\mu_i\}_{i\in\mathbb{N}}$，我们可以据此定义相应的混合测度 ξ_w^{comp}和预测器 $\mathrm{p}_w^{\text{comp}}$；根据贝叶斯一致性，它对所有可计算测度都是可靠的。一种通用可靠的预测方法：还有什么不喜欢的呢？[8]

2.3. **通用最优性**。 一个哲学上的疑虑是，将通用可靠性等同于可计算性方面的可靠性，也就是说，对所研究序列所有可能的生成方式都可靠。虽然这无疑具有普遍性，但仍是一种限制，一种归纳假设（参见 Howson，2000，第 77 页）。坚持这一归纳假设确实具有通用性，就意味着认定我们将来或许只能遇到可计算生成的序列；这可能是一种“物理版丘奇–图灵论题”，认为自然是可计算的（见 Copeland and Shagrir，2020）；这种立场难以评估。

2.3.1. 所有方法中的最优性。 统计学家Schervish(1985，第 1274 页）打趣道：“据我所知，自然并未受到与统计学家相同的可计算性限制。”但我们这些统计学家，或者一般而言，预测方法的开发者，很可能受到可计算性的约束。至少，我们能够开发的预测方法是如此。

顺着这个思路，我们可以明确地将可能的预测方法限定为预测算法，也就是说（接受原始的丘奇–图灵论题）限定为图灵可计算的预测器。接着，我们可以说，如果某个预测器在某种意义上与任何预测算法、任何可计算预测器一样可靠，那么它就是通用的，或者说是最优的。

这一做法与Reichenbach试图为归纳辩护的做法颇为相似（例如参见Salmon，1991）。用我们的术语来说，他的想法是：我们必须承认，无法证明“归纳方法”普遍可靠，因为我们无法先验地排除这样一种可能性：存在某些结果序列，使该方法在这些序列上失效。同样，我们也无法先验地排除不可计算的结果序列存在的可能性，而通用混合在这些序列上会失效。不过，我们或许能够证明，在凡是能够取得成功的情况下，归纳都是成功的。

Reichenbach本人关于“归纳直规则”的提议，以及该规则在存在极限相对频率时总能成功的主张，最终未能令人信服（Galavotti，2011）。但或许当前的提议更有希望：我们可能拥有一种通用最优方法，只要某种方法能够成功（也就是只要成功本身是可实现的），这种方法就能成功。

2.3.2. 聚合预测器。 事实上，我们可以证明关于混合预测器的最优性，有一个相当强的结论 $\mathrm{p}_w^{\text{comp}}$。首先请注意，正如混合测度会给出一个混合预测器（定义 1 见上文）一样，任意测度μ 都会给出一个相应的预测器p。因此，一类测度对应于一个预测器集合

[8]在这里，人们或许已经会担心，所有可计算测度构成的类虽然是可数的，却未必能以有效方式枚举。我将在下文第 2.4–3.1 节讨论这一点。

的预测器。[9] 混合预测器$\mathsf{p}_w^{\mathcal{H}}$ 也可以直接写成一个预测器，它聚合相应预测器集合中的所有预测：[10]

$$\mathsf{p}_w^{\mathcal{H}}(\boldsymbol{x}, x) = \sum_{i\in\mathbb{N}} w_{t+1}(i)\mathsf{p}_i(\boldsymbol{x}, x). \tag{6}$$

这里，p_i 是与 μ_i对应的预测器，权重函数按以下方式更新

$$w_{t+1}(i) = \frac{w_t(i)\mathsf{p}(\boldsymbol{x}_{t-1}, x_t)}{Z}, \tag{7}$$

归一化因子为 $Z = \sum_{i\in\mathbb{N}} w_t(i)\mathsf{p}(\boldsymbol{x}_{t-1}, x_t)$，初始权重为 $w_0 = w_\circ$

2.3.3. 最优性。 现在，我们可能会希望将聚合预测器的最优性精确表述如下：对于每个可能的无限结果序列，如果集合中的某个预测器在该序列上收敛，那么聚合器也会收敛。遗憾的是，这个要求过于严格。[11,12]

我们将改用另一种途径，至少尽可能接近这个想法。[13]为此，我们需要引入一个损失函数 ℓ_{p}，用于量化预测器 p 的预测相对于实际结果 x所带来的代价或损失。序列预测中的标准损失函数是对数损失函数，简称对数损失函数，其定义为

$$\ell_{\mathsf{p}}(\boldsymbol{x}, x) := -\log \mathsf{p}(\boldsymbol{x}, x). \tag{8}$$

预测器 p 在序列 $\boldsymbol{x}$ 上的累积损失，是在生成 $\boldsymbol{x}$的过程中，p 的各项预测所产生的瞬时损失之和：

$$L_{\mathsf{p}}(\boldsymbol{x}^s) := \sum_{t=0}^{s-1} \ell_{\mathsf{p}}(\boldsymbol{x}^t, x_{t+1}). \tag{9}$$

最后，预测器 p_1 相对于另一个预测器 p_2 在序列 $\boldsymbol{x}$ 上的累积遗憾，是指 p_1 相较于 p_2,所承受的额外损失

$$R_{\mathsf{p}_1,\mathsf{p}_2}(\boldsymbol{x}) := L_{\mathsf{p}_1}(\boldsymbol{x}) - L_{\mathsf{p}_2}(\boldsymbol{x}). \tag{10}$$

---

[9]再次说明（脚注 7），严格来说，这仅适用于处处为正的测度，因为相应的预测器在遇到零概率结果时会“陷入昏迷”（Kelly，1996，第 305 页）。这在此并非真正的问题，因为发生这种情况时，该预测器只会从预测器集合的混合中退出（前提是并非所有预测器都发生这种情况——在所有可计算元素这一相关情形中，这一点已被排除）。[10]参见（Sterkenburg，2018，第 3.2.2 节）。[11]事实上，对于我将在下文第 3 节讨论的所罗门诺夫-列文半预测器——它们是一个包含自身的预测器集合的聚合器——可以证明，对于任意一对这样的半预测器，都存在一些序列，使它们不会收敛到相同的预测（Sterkenburg，2018，定理 3.2）。关于（贝叶斯）学习者在所有事件上的“确定性”收敛研究，参见（Nielsen，2020）。[12]或者，也可以将一致性定理 2 重新诠释为一个*merger-of-opinion* 结果（例如参见 Huttegger，2015）：每个预测器都确信地预期，聚合器的预测会收敛到它自己的预测。我不会采用这条思路，部分原因是对于我必须在下文第 3 节转而讨论的半预测器，这种诠释更难理解（参见 Sterkenburg，2018，第 3.3.2.1 节）。

[13]这是一个得到更广泛研究的方法，见 *prediction with expert advice* 或 *competitiveonline learning*（Cesa-Bianchi 和 Lugosi，2006)。这一领域的研究成果是*meta-inductive* 归纳问题方法的核心（Schurz，2019；另见STERKENBURG，2020）。

对于对数损失，我们可以将预测器 p 的累积损失写作，其中 μ 为相应的测度：

$$L_{\mathsf{p}}(\boldsymbol{x}^s) = \sum_{t=0}^{s-1} -\log \mu(x_{t+1} \mid \boldsymbol{x}^t) = -\log \prod_{t=0}^{s-1} \mu_{\mathsf{p}}(x_{t+1} \mid \boldsymbol{x}^t) = -\log \mu_{\mathsf{p}}(\boldsymbol{x}^s), \tag{11}$$

因此，p_1 相对于对应于 μ_1 和 μ_2 的 p_2，的累积遗憾，

$$R_{\mathsf{p}_1,\mathsf{p}_2}(\boldsymbol{x}) = -\log \mu_1(\boldsymbol{x}) - (-\log \mu_2(\boldsymbol{x})) = -\log \frac{\mu_1(\boldsymbol{x})}{\mu_2(\boldsymbol{x})}. \tag{12}$$

我们将$R^{\mathcal{H}}_{w,\ i}$ 记作聚合预测器 $\mathsf{p}^{\mathcal{H}}_w$ 相对于对应于 $\mu_i \in \mathcal{H}$ 的预测器 p_i 的遗憾函数。那么，对于对数损失，我们可以推导出聚合器相对于预测器集合中任意预测器的如下常数界，该界适用于每个结果序列。

定理 3（最优性）。对于聚合预测器 $\mathsf{p}^{\mathcal{H}}_w$、预测器集合中每个对应于 $\mathcal{H}$ 的p_i，以及每个有限序列 $\boldsymbol{x}$，

$$R^{\mathcal{H}}_{w,i}(\boldsymbol{x}) \leq -\log w(i).$$

证明。我们有

$$R^{\mathcal{H}}_{w,i}(\boldsymbol{x}) = -\log \frac{\xi_w(\boldsymbol{x})}{\mu_i(\boldsymbol{x})} \leq -\log \frac{w(i)\mu_i(\boldsymbol{x})}{\mu_i(\boldsymbol{x})} = -\log w(i)$$

这是因为支配性质 $\xi_w(\boldsymbol{x}) = \sum_{j\in\mathbb{N}} w(j)\mu_j(\boldsymbol{x}) \geq w(i)\mu_i(\boldsymbol{x})$。 □

2.3.4. 通用聚合器。 因此，特别是对于我们针对所有可计算预测器构成的预测器集合的通用聚合预测器，通用聚合器相对于该集合中任意给定预测器的对数遗憾，在任何 结果序列上都有一个常数上界。也就是说，再次接受将预测器等同于预测算法这一做法，并考虑可计算的 预测器，我们可以说，只要任何预测器取得成功（损失较低），聚合预测器就能被证明是成功的（损失较低）。由此，我们得到了一种真正通用最优的预测方法。

2.4. **对角论证**。 不过，如果像我们之前那样，将可能的预测器限定为可计算预测器，那么聚合预测器本身就不再是一个可能的预测器。

这一观察可追溯到 Putnam (1963年) 对 Carnap（1950年； 1952年； 1959年）归纳逻辑纲领所提出的一项论证。Putnam 认为，Carnap 同样是在寻求某种用于归纳推理的通用形式方法，并试图通过证明关于这种通用方法的两个合理条件彼此排斥，一劳永逸地说明这一努力徒劳无功。[14]

Putnam 的第一项条件是，作为对我们归纳实践的理性重构，Carnap 的归纳方法至少应能最终检测出所有可计算的模式。用我们的术语来说，这可以被解释为一项可靠性条件，也可以被解释为一项最优性条件，因为其理念是，该方法应能发现我们任何可能的归纳实践所能发现的所有模式。我将仅把这项条件表述如下，涵盖这两种解释。

(I) p 对可计算性而言是可靠的（最优的）。

[14]更多详情请参见（STERKENBURG，2018年，第 1 章）。

他的第二项条件恰恰是，为了仍能算得上一种实际方法，它应满足某种可计算性条件。我将其简述如下：

（II）p 是可计算的。

Putnam 基于 G¨odel 的原始对角化证明有些复杂，但对于我们的两个条件，论证很直接（另见 Kelly，2016年，第 702 页）。

定理 4（Putnam 的对角化）。不存在满足 p 的 （I）和（II）。

证明。假设存在这样的 p。我们构造一个可计算序列 $\boldsymbol{x}^\omega$，使得 p 永远不会收敛于该序列上（其对数损失无界，因此相对于对应于 $\boldsymbol{x}^\omega$的可计算预测器，其对数遗憾也无界），如下所示。对于每个 t，计算并选取使 $\mathrm{p}(\boldsymbol{x}^t,\ x_{t+1}) \leq 0.5$. □

3. The Solomonoff-Levin approach

Solomonoff 在其里程碑式论文（1964年）中提出了若干种用于预测的“模型”。其中一种正是我们对可计算测度的混合，他还指出，这种混合本身不可计算（同上，第 21 页）。他提出的其他模型采用了不同的思路，因此给出了另一种预测方法。不过，Solomonoff 本人的论述并不完全严谨，而恰当的形式化框架直到柯尔莫哥洛夫的学生 Levin （1970；1971）才得到严格的形式化。我现在将首先阐述这一框架。[15]

3.1. 半可计算半测度。

将这一框架的动机视为一种明确避免对角化的尝试，并先考察自然数上的可计算函数这一背景下的类似做法，颇有启发性。

3.1.1. 可计算函数。

可计算性理论中的一个基本事实是，全体全函数可计算（t.c.）函数 $f\colon \mathbb{N}\to\mathbb{N}$ 组成的类可以对角化（例如参见 Soare，2016 年）。更准确地说：若存在一个通用的全函数可计算函数。 $_\circ f$，能够模拟其他所有全函数可计算函数（也就是说，对于全体全函数可计算函数的一个枚举 $\{f_i\}_{i\in\mathbb{N}}$，有 $f(i,x) = f_i(x)$），那么我们便可推出一个对角函数 g°（例如 $g(x) := f(x, x) + 1$），它是全函数可计算的，却与每一个 f_i 都不同（因为对于所有 i，$g(i) = f_i(i) + 1 \neq f_i(i)$），这就造成了矛盾。这也意味着，不可能存在一个由全体全函数可计算函数组成、且自身可计算的枚举 $\{f_i\}_{i\in\mathbb{N}}$：全函数可计算函数类不是可计算可枚举的。

3.1.2. 部分可计算函数。

然而，标准的图灵机模型刻画的是图灵可计算（我们简称为：可计算）函数，并不局限于全函数。也就是说，图灵机可能无法在特定输入上停机，因此它对该输入所计算的函数值未定义。

因此，图灵机对应的是更大的部分可计算（p.c.）函数类 φ。这一扩展避开了前述对角化论证（试想：如果 $\varphi_i(i)$未定义，会怎样？）。事实上，所有 p.c. 函数组成的类是可计算可枚举的；从一个最直接的观察便可理解这一点：我们可以将所有可能的有效描述编码为整数索引，

[15]本节内容取材于（Sterkenburg，2018年，第2章），其中包含更多细节。关于这一框架的其他介绍，参见（Shen等人，2017年；Li and Vit´anyi，2019）。

（图灵机指令集的描述）。具体而言，我们可以定义一台通用图灵机（实际上，对于每一种这样的编码，都可以定义无穷多台这样的机器）。也就是说，给定所有 p.c. 函数的一个可计算枚举 $\{\varphi_i\}_{n\in\mathbb{N}}$，我们可以设计一台机器：它接受一个索引 i 和输入 x，还原出 φ_i，并计算 $\phi_i(x)$。这样一台机器计算出一个通用部分可计算函数 $\mathring{\varphi}: \ i, \ x \mapsto \varphi_i(x)$。

3.1.3. 可计算测度。我们回到康托尔空间上的概率测度；康托尔空间是我们的无限二进制序列结果空间。设由博雷尔函数 $F: \mathbb{B}^\omega \to \mathbb{B}^\omega$ 对均匀测度进行的变换 λ_F 定义为：对每个可测事件 A，$\lambda_F(A) = \lambda(F^{-1}(A))$。康托尔空间上的每个博雷尔测度 μ，都可以通过某个博雷尔函数将 λ 变换得到。

为了现在在这种测度刻画中施加可计算性限制，我们需要用有限序列上的函数来表达这些变换。略去若干细节后，我们得到具有某种单调性性质的可计算函数；它们恰好由一种称为单调机的图灵机来定义。

可以把单调机想象成在稳定不断的输入符号流上运行，并在此过程中生成输出比特流。若单调机 M 接收 $\boldsymbol{x}$ 作为输入后，生成一个以 $\boldsymbol{y}$ 为初始片段的序列，就称 $\boldsymbol{x}$ 是 M-描述。集合 $D_M(\boldsymbol{y})$ 中针对 $\boldsymbol{y}$ 的最小 M-描述，是所有满足 $\boldsymbol{x}$ 的序列 $\boldsymbol{x}$；但 $\boldsymbol{x}$ 的任何初始片段都不是 $\boldsymbol{y}$ 的 M-描述。

单调图灵机M 对均匀测度 λ_M 的变换定义如下：

$$\lambda_M(\boldsymbol{y}) := \sum_{\boldsymbol{x}\in D_M(\boldsymbol{y})} \lambda(\boldsymbol{x}). \tag{13}$$

可计算测度由单调机 M 的变换给出，这些单调机是“几乎全定义”的。这意味着，对于 λ-几乎所有无限序列 $\boldsymbol{x}^\omega$（即概率为 λ的序列）M，随着 $\boldsymbol{x}^\omega$的初始片段越来越长，它都会返回越来越长的输出序列，从而生成一个无限 $\boldsymbol{y}^\omega$。

3.1.4. 半可计算半测度。一般而言，单调机可能无法生成这样一个无尽的序列。它可能是部分定义的，即对于某些 $\boldsymbol{x}$及其所有扩展 $\boldsymbol{x}$，它只生成一个有限序列 $\boldsymbol{y}$，此后不再生成任何内容$\boldsymbol{y}$。这意味着，在 λ_M的定义中，扩展 $\boldsymbol{y}$的测度可能会“丢失”，并且 $\lambda_M(\boldsymbol{y})$严格大于 $\lambda_M(\boldsymbol{y}0) + \lambda_M(\boldsymbol{y}1)$。

也就是说，针对单调机的变换 λ_M M 是半测度 ν，它们只满足以下不等式

$$\begin{aligned} \nu(\varnothing) &\leq 1; \\ \nu(\boldsymbol{x}0) + \nu(\boldsymbol{x}1) &\leq \mu(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \{0,1\}^*. \end{aligned} \tag{14}$$

此外，这些半测度还满足一种较弱的可计算性概念。具体而言，变换 λ_M 对所有 M 而言，恰好就是（下） 半可计算的 半测度，即满足条件（ ν （15），且存在一个可计算的 f_ν 从下方逼近 ν 的函数。也就是说，对所有 $\boldsymbol{x} \in \{0,1\}^*$和所有 $s \in \mathbb{N}$,

$$\begin{aligned} f_\nu(\boldsymbol{x}, s) &\leq f_\nu(\boldsymbol{x}, s+1); \\ \lim_{s\to\infty} f_\nu(\boldsymbol{x}, s) &= \nu(\boldsymbol{x}). \end{aligned} \tag{15}$$

3.1.5. 通用半可计算半测度。 与通常的图灵机一样，所有单调机组成的类可以被可计算地枚举，因此我们可以定义 通用 单调机 U。这又使我们能够定义通用 变换 U。对于所有单调机的枚举 $\{M_i\}_{i\in\mathbb{N}}$，这样的通用变换可以写成所有变换的混合，

$$\lambda_U(\boldsymbol{y}) = \sum_{i\in\mathbb{N}} \lambda(\boldsymbol{x}_i)\lambda_{M_i}(\boldsymbol{x}), \tag{16}$$

使用 $\{\boldsymbol{x}_i\}_i$ 某种整数编码。由此可直接得出，通用变换是通用半可计算半测度，也就是说，是满足以下条件的测度：对所有 ν，存在某个常数 $c_\nu \in (0,\ 1)$，

$$\mathring{\nu}(\boldsymbol{x}) \geq c_\nu \cdot \nu(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \{0,1\}^*.$$

3.1.6. 表示定理。事实上，定义通用变换有多种不同方式，它们刻画的是同一类 $\mathcal{SL}$ 通用半可计算半测度。

首先，我们可以不用均匀测度 λ，而是用任何其他连续可计算测度 μ 来定义通用变换。[16]

其次，我们可以直接使用贝叶斯混合ξ_w^{semi}的定义，对象是 $\mathcal{H}^{\text{semi}}$ 所有半可计算半测度。

遵循良好的哲学传统（Suppes，2002年），我们可以汇集并总结这些观测结果，将表面上迥异、实则描述同一对象的各种表示联系起来，并以表示定理为题。

表示定理5 (Wood等人；STERKENBURG，2017年)。**对于所有连续可计算测度** μ，以及用于定义 $\{\nu_i\}_i$ 的 $\mathcal{H}^{\text{semi}}$ 可接受枚举 ξ_w^{semi}，

$$\mathcal{SL} = \{\mu_U\}_U = \{\xi_w^{\text{semi}}\}_w, \tag{17}$$

其中 U 遍历与 μ 兼容的通用单调机，而 w遍历可计算权重函数。[17]

证明。 这里需要一些篇幅。参见（Sterkenburg，2018年，第2.2节）。 □

3.1.7. 包含通用性。 定理 4 表明，我们可以对可计算测度类进行对角化：任何支配所有可计算测度的测度，本身必定不属于该类。通过将可计算测度类扩展为半可计算半测度类，我们便摆脱了对角化。也就是说，该类包含通用元素，它们支配类中的所有元素，同时自身仍属于该类。

3.2. **所罗门诺夫-列文半预测器**。类 $\mathcal{SL}$ 的元素是所罗门诺夫-列文半测度。所罗门诺夫-列文半预测器，就是与所罗门诺夫-列文半测度相对应的半预测器——和之前一样，通过条件化得到。

[16]如果一个测度没有任何 *atoms*，则称其为连续测度：即被赋予正概率的单个无限序列。[17]关于半可计算半测度的 *acceptable enumeration* 定义（与部分可计算函数可接受枚举的标准定义一致，参见（Soare，2016年，第21页）），见（Sterkenburg，2018年，第194页）。要求机器为 *compatible*，并且 μ，这是因为需要排除以下情况：机器所用编码 U 从 μ 获得非零概率（同上，第190页）。

3.3. 对角化再次出手。我们能否说，所罗门诺夫-列文半预测器是普遍最优预测器?

为了重现第2.3节中的同类推理——该推理始于将可能的预测器认定为那些可计算的预测器——现在我们必须降低所要求的可计算性水平，并仍允许仅半可计算的对象作为可能的预测器。这样做的希望在于，我们便能统一 Putnam 关于通用预测器的两项要求：预测器既是可能预测器类中的通用预测器，同时本身也是一个合格的预测器。

乍看之下，所罗门诺夫-列文半预测器似乎正好满足这一要求。由此可直接得出：所罗门诺夫-列文半预测器是对应于半可计算半测度的半预测器的聚合预测器，因此它们在对应于半可计算半测度的半预测器中是最优的。此外，所罗门诺夫-列文半预测器本身也是对应于半可计算半测度的半预测器。

然而，这里存在一个问题。可计算性这一较弱的要求——半可计算性——适用于底层的概率测度，但不一定能延伸到相应的半预测器。事实上，所罗门诺夫-列文半-预测器 不再是半可计算的。[18]

定理 6。 所罗门诺夫－列文半预测器不是半可计算的。

证明。假设对于某个 $\xi_w^{\text{semi}} \in \mathcal{SL}$，$\mathrm{p}_w^{\text{SL}}(\boldsymbol{x}, x) := \xi_w^{\text{semi}}(x \mid \boldsymbol{x})$ 是半可计算的。现在我们可以如下构造一个可计算的无限序列 $\boldsymbol{x}^\omega$ 。首先，以交错计算的方式，从下方开始计算 $\mathrm{p}_{\text{SL}}(1^t, 1)$，并让 $t \in \mathbb{N}$ 逐渐增加，直到找到一个满足 $\mathrm{p}_{\text{SL}}(1^{t_0}, 1) > 0.5$ 的 t_0 （因为 1^ω 显然是可计算的，而 p_{SL} 特别地对所有可计算序列都可靠，因此这样的 t_0 必然存在）。接下来，对逐渐增加的 t 计算 $\mathrm{p}_{\text{SL}}(1^{t_0}01^t, 1)$，直到找到一个满足 $\mathrm{p}_{\text{SL}}(1^{t_0}01^{t_1}, 1) > 0.5$ 的 t_1（同样，由于 $1^{t_0}01^\omega$ 是可计算的， t_1 必然存在）。如此继续，我们得到一个位置列表 t_0、t_1、t_2、.....；令 $\boldsymbol{x}^\omega := 1^{t_0}01^{t_1}01^{t_2}1\ldots$。序列 $\boldsymbol{x}^\omega$ 是可计算的，但根据构造，$\mathrm{p}_w^{\text{SL}}(\boldsymbol{x}, x)$ 的预测将在该序列上不收敛，这与预测器对可计算序列的可靠性相矛盾。□

所罗门诺夫-列文半预测器仅仅是极限可计算的：存在一个可计算的逼近函数 $f(\boldsymbol{x}, x, s)$，对于每个 $\boldsymbol{x}$，x 都会收敛到对给定 $\boldsymbol{x}$时 x 的预测概率。与半可计算性的区别在于，在半可计算性中，我们可以确定 $s+1$ 处的逼近比 s处的逼近更准确；而在极限可计算性中，我们连这一点也无法保证。

尽管如此，我们能否尝试将可能的预测方法与所有极限可计算预测器等同起来，并为这一类推导出一个普遍最优预测器？不行，我们做不到：在这个层级上，我们可以用停机集作为预言机，重复定理4中的原始对角线证明。一般来说，借助所谓的“相对化”（计算理论中指将不可计算集用作预言机），前面给出的两个对角线证明足以说明，在可计算性层级的算术层级中的任何层级上，这种做法都行不通。

[18]这一点最早由（Leike 和 Hutter，2015，定理 6）证明。下面这个更为直接的证明出自（STERKENBURG，2019）。

3.4. **我们得到了什么？** 因此，目前我们所能说的最好结论是：如果将可能的预测方法等同于那些与半可计算半测度相对应的半预测器，那么我们就拥有普遍最优的预测方法，即所罗门诺夫-列文半预测器。

3.4.1. 可能的预测方法。 显而易见的问题在于，这看起来几乎不像是对可能的预测方法的一种自然刻画。

仅半测度和半预测器这两个概念本身就至少引发了解释上的疑问。学术文献中已有一些尝试，试图为这些概念提供解释（Campbell；Martin 等人，2016年；Wyeth 和 Hutter，2026年），但归根结底，正如我们在此引入该概念时所说，半测度主要是一种形式上的必要条件，使我们能够得到一种较弱的可计算性概念——半可计算性，而它允许存在通用元素。[19,20]

将半可计算性（或极限可计算性）视为对可能预测器的外部约束，或许看起来颇为合理，因为这类预测器仍然有效地“触手可及”：我们仍能通过可计算的方法对其进行近似。但不妨想想，实际使用一个严格半可计算或极限可计算的半预测器会是什么样子（参见Kelly 等人，1994年，第104页关于部分可计算预测器的讨论）。

为了在任意给定轮次实际作出预测，我们不能永远等待对方法预测结果越来越准确的近似。因此，我们必须决定何时采用当前的近似。现在，要么这些决定可以形式化为一个可计算过程，在这种情况下，所得的半预测器实际上就不再具有通用最优性或可靠性；要么我们的决定在某种程度上不可计算，这就戳破了这样一种想法：可计算近似仍然 somehow 在有效可及的范围内。[21]

3.4.2. 回到可靠性。当然，我们完全可以放弃对最优性的诠释，只保留 Putnam 的第二个条件。也就是说，所罗门诺夫-列文半预测器虽然不可计算，但仍然是通用可靠的：它在所有可计算序列（事实上，所有可计算测度）上收敛。

但是，如果我们不对通用预测方法施加任何可计算性要求，那么就不清楚为何不采用第2.2节中更为直接的定义。该定义直接基于所有可计算测度构造的混合测度，因此得到的是一个实际的预测器，而不是半预测器。此外，这种预测器所编码的归纳假设——数据由某个可计算测度生成——似乎比所罗门诺夫-列文半预测器的归纳假设更自然；后者假定数据由某个半可计算的半测度生成。[22]

---

[19]将半测度中的“概率泄漏”解释为类似于宇宙终结概率的观点，与第 2.1 节的框架具有通用性的理念相矛盾，因为我们可以将任何事物重新编码为一个二元预测问题（见上文脚注 5 ）：为什么不把这种可能性编码进二元序列中呢？[20]Solomonoff 还描述了所罗门诺夫-列文半测度的一种 *normalization* ，由此再次得到一个正规测度（Li 和 Vit´anyi，2019，第 4.5.3 节）。这个测度只是极限可计算的（Leike 和 Hutter，2015），这就引出了一个问题：为什么我们不会满足于对所有极限可计算测度进行直接混合？[21]我将在下文第 3.5.3–3.5.4 节回到这样一种想法：至少我们能够为所罗门诺夫-列文半预测器设计可计算近似。[22]对后者最自然的解释或许是：我们假设数据由某台单调机在随机输入下生成（Li 和 Vit´anyi，2019，第 357 页；Grau-Moya 等人，2024年，第 2 页）。这确实引入了一种可能性：在某个有限时刻之后，不会再有新的结果

3.4.3. 恰到好处？或许我们仍可以认为，所罗门诺夫-列文定义至少在普特南的两个互不相容的条件之间实现了尽可能好的权衡。预测方法并不可计算，但已十分接近；其可靠性或最优性的适用范围略微过宽，但也不至于太过。

但是，当然也完全可以说，这一定义在两个方面都不能令人满意，因为它对其中一项要求达不到，对另一项要求又超出了。事实上，如果第一项被认为是一项要求，那么到目前为止我们的讨论不应得出这样的结论：我们可以为通用预测方法给出一个至少相当接近我们所期望内容的理论定义。结论只能是，我们无法达到这一目标：不可能存在通用预测方法。

3.5. **不过，我们究竟获得了什么？** 让我们暂时搁置先前的担忧，接受所罗门诺夫-列文半预测器作为真正的通用（普遍可靠的）预测器。我们能从所罗门诺夫-列文定义及其性质中学到什么，尤其是着眼于实际的机器学习时？

3.5.1. 奥卡姆剃刀。人们大力强调了 Solomonoff 的原始定义与数据压缩和奥卡姆剃刀之间的联系；奥卡姆剃刀是一项方法论原则，即在归纳推理中偏好简单性。

在通过单调机对均匀测度进行变换的定义（ λ 13）中，我们看到 M，由于 $\lambda_M(\boldsymbol{x})$ 更高，因为 M-描述 $\boldsymbol{y} \in D_M(\boldsymbol{x})$ 具有更高的均匀概率 $\lambda(\boldsymbol{y}) = 2^{-|\boldsymbol{y}|}$，也就是说，$M$-描述更短。由此可以理解为，概率 $\lambda_M(\boldsymbol{x})$ 更高，是因为 $\boldsymbol{x}$ 通过机器更具可压缩性 M。[23]

此外，如果我们使用通用单调机 U，那么相应的变换 λ_U——一种所罗门诺夫-列文半测度——可以说采用了通用的可压缩性概念。事实上，这一定义与 $\boldsymbol{x}$ 的单调柯尔莫哥洛夫复杂度定义密切相关，

$$Km_U(\boldsymbol{x}) = \min\{|\boldsymbol{y}| : \boldsymbol{y} \in D_U(\boldsymbol{x})\}, \tag{18}$$

$\boldsymbol{x}$ 最短单一描述的长度。[24] 要说明这种可压缩性概念在某种意义上是通用的，请注意，如果某个（较长的）序列 $\boldsymbol{x}$ 可以通过某种算法（借助某个机器 M）得到较短的描述（即具有可压缩性），那么，除去一个常数（即 U 模拟 M 所需的索引），它也可以通过 U 得到较短的描述（即具有可压缩性）。

此外，由于不同的通用机也能彼此模拟，因此选择哪一台特定的通用机也无关紧要，至多只差一个常数。这一事实称为不变性定理，并被认为赋予了这一概念

到达，这改变了第2.1节中所述的学习问题。此外，如何理解预测器在半测度上的收敛并不完全清楚。另见（STERKENBURG，2018年，第 3.2.4.3、B.2.2.4 节）。[23]更一般地说，描述长度函数与概率分布之间存在一种形式上的对应关系（Gr¨unwald，2007；Shen 等人，2017年），较短的描述对应较高的概率。重要的是，这凸显了“可压缩性”和“简单性”的说法与某个特定描述长度函数相关，其方式与“高概率”相对于某个特定概率分布而言完全相同。[24]然而，单调柯尔莫哥洛夫复杂度的定义并不 *quite* 对应所罗门诺夫-列文半测度的定义，即便在不变性定理所说的“相差一个常数”意义上也不对应。详情见（STERKENBURG，2018年，A.3.2 节）。

柯尔莫哥洛夫复杂度一定程度上的客观性或稳健性。[25] Solomonoff 首次提出了类似柯尔莫哥洛夫复杂度的概念，并指出了这种不变性，因此他与 Kolmogorov、1965 和 Chaitin、1969一起，成为算法信息论的三位独立创始人之一。

按照这种解释，所罗门诺夫-列文半测度 λ_U 会为更易压缩（在这个意义上，更简单）的序列赋予更高的概率，我们还可以得出，对于 $\boldsymbol{x}$ 的单比特扩展 $\boldsymbol{x}x$，若前者更易压缩（更简单），则 $\lambda_U(x \mid \boldsymbol{x}) = \lambda_U(\boldsymbol{x}x)/\lambda_U(\boldsymbol{x})$的值更高。因此，所罗门诺夫－列文半预测器体现了某种简单性偏差，这与奥卡姆剃刀一致。

3.5.2. 一些怀疑。 首先应当指出，所罗门诺夫“算法概率”的不变性，也就是所罗门诺夫-列文半测度之间的不变性，与两种所罗门诺夫-列文半测度会作出相同预测这一说法相去甚远。[26] 方差问题，或者说选择通用机（或者等价地，根据定理 5，为通用混合选择可计算权重函数）所涉及的“主观性”，普遍被视为所罗门诺夫理论面临的主要挑战之一（参见STERKENBURG，2018年，第3.2.5节）。

接下来谈谈它与奥卡姆剃刀的关联，问题在于，这种简单性诠释究竟有多稳健。定理5 还表明，所罗门诺夫-列文半测度不一定要写成均匀测度的通用变换。具体来说，我们可以使用任意连续测度 μ来代替 λ，但这样就失去了对描述长度和可压缩性这种一目了然的诠释。[27]

我们表示所罗门诺夫-列文半测度的方式具有极大的灵活性，这指向了一个更重要的问题：人们所感知到的简单性偏差实际上发挥了什么作用，并不明确。所罗门诺夫－列文半预测器的独特优点是其普适可靠性。而经证明能够保证其普适可靠性的性质，是它们的通用支配性——当然，这一性质并不取决于我们具体如何表示它们。因此，如果问题是这里建立的理论能告诉我们什么才是良好预测的关键，答案就是：通用支配性。奥卡姆剃刀究竟对此补充了什么，并不明确。[28]

这又与一件更重要的事情联系在一起。推想，从所罗门诺夫-列文通用预测方法中推断方法论原则，最终是为了指导我们设计或理解实际可实现或已经实现的机器学习算法。因此，尤其是，

---

[25]不过，“差一个常数”并非无关紧要。这意味着，如果我们先固定两台不同的机器，那么随着我们考察的序列越来越长，通过这两台不同机器得到的两个平均柯尔莫哥洛夫复杂度（即用序列长度除复杂度）将会收敛；在这一精确意义上，机器的选择无关紧要。然而，如果我们转而考虑某个给定特定序列的复杂度，那么通过随后选择的两台不同机器得到的柯尔莫哥洛夫复杂度之差实际上可以是任意值（另见 Sterkenburg，2018年，第 5.2.2 节）。[26]呼应脚注 25：对于任意序列，随后选择的两个不同所罗门诺夫-列文半测度赋予该序列的概率之比实际上可以是任意值。我们至多可以说，对于固定的两种机器选择，随着我们考察长度不断增加的t 序列，这些概率赋值的 t次方根会收敛（见 Sterkenburg，2018年，第 3.2.5.6 节）。这甚至不能推出对于每个序列， *predictions* 都会收敛（见上文脚注11 ）。[27]通过等长描述所具有的相等均匀概率，与无差别原则之间的关联也是如此（Solomonoff，1964；Rathmannerand Hutter，2011）。[28]（Sterkenburg，2016年；2018年，第 5.1 节）中也提出了类似的观点。

将奥卡姆剃刀与之联系起来，推想是为了为在学习算法中实现简单性偏差提供某种依据，或者解释为何实现了某种简单性偏差的学习算法能够表现良好。但所罗门诺夫-列文定义真的能提供这样的指导吗?

3.5.3. 黄金标准。 由于形式上的不可计算性，所罗门诺夫-列文半预测器无法实际实现。尽管如此，文献中仍有这样的观点：“算法概率可以充当归纳系统的一种‘黄金标准’”（Solomonoff，1997年，第83页）。

这种情况与理想认识论 (卡尔， 2021年；另见内特， 2023年，第298页）这一研究计划颇为相似。在这里，我们试图找出稳健的理性规范；实际中受认知能力限制的推理者永远无法满足这些规范，但它们可以作为一种理想，供人们近似实现(卡尔， 2021年，第1156页）。同样，所罗门诺夫-列文定义可以作为预测的理想标准；事实上，我们可以设计一些方法，明确地近似所罗门诺夫－列文半预测器，从而近似这一理想。通常的做法是，对定义中可能无界的量施加有限上限，尤其是对计算步骤数施加上限（例如参见Grau-Moya 等人， 2024年）。随着这些量趋于无穷，我们便能在极限意义下恢复原始定义。因此，这些具有有限上限的定义显然是近似（而且上限越大，近似就越精确）。

此外，通过扩大资源投入而逐步接近预测黄金标准的过程，或许可以恰当地概括机器学习领域的发展历程。将逼近所罗门诺夫－列文半预测器视为理解该领域现状的一种模型，或许会有所帮助。也许一些重要进展甚至明确受到逼近所罗门诺夫－列文半预测器这一目标的推动。而且，理论或许确实能提供规范性指导，即研究人员应该 明确朝着更好地逼近所罗门诺夫－列文半预测器这一目标迈进。

3.5.4. 一些质疑。 这里有一个描述性问题和一个规范性问题。先从第一个问题说起：将机器学习研究建模或描述 为正朝着所罗门诺夫－列文半预测器这一理想迈进的过程，有帮助吗?

描述性论述面临的一个问题是，从这里似乎相关的视角来看，近似这一概念毫无意义。也就是说，对于任何 实际的（即可计算的）预测方法，我们都可以选取一台通用图灵机和有限界限，使该方法被视为对所罗门诺夫-列文定义的近似。[29,30] 这意味着，无论该领域当前处于什么状态（特别是迄今为止已经开发出的、必然可计算的方法），我们都可以将其描述为正处于逼近所罗门诺夫-列文预测器的某个阶段。

当然，这并不排除这样一种可能：著名的机器学习研究者在工作中曾经或如今仍受到动机驱使，试图近似所罗门诺夫－列文

[29]也就是说，我们可以定义一台通用图灵机，使得在这一有限近似中，它唯一模拟的机器就是对应于这种可计算方法的机器。[30]这一观察与如下主张有关：贝叶斯主义是空洞的，因为任何结论事后都可以被重构为贝叶斯推断的结果（Albert，2001）。当然，当我们关心如何进行推断这一规范性问题时，这并不是正确的视角。但对于当前的描述性问题而言，情况则是如此。

半预测器。也许如此；这是一个经验或心理学问题。更重要的问题是规范性问题：他们是否应该这样做。

这里的问题在于，要求我们以所罗门诺夫－列文半预测器为近似对象的规范性建议似乎极其薄弱，甚至几乎毫无用处。随即便会产生一个问题：无数个不同的所罗门诺夫-列文半预测器中，我们究竟应当尝试近似哪一个？这就是方差问题（见上文第 3.5.2 节）。或许存在一些成功的独立论据，能够支持某个“最自然”的通用图灵机，进而支持相应的所罗门诺夫-列文半预测器。但要提出这样的论据，我们就必须超越迄今为止的理论叙述；该叙述正是依据普适可靠性，将所罗门诺夫-列文半预测器——并且 一概 平等地——确立为黄金标准。更一般地说，迄今为止的理论叙述并未告诉我们，在实践中我们大概会关心什么，也就是短期内哪些近似比其他近似更好。无论我们在任何有限阶段采用何种可计算的近似，它都必然会失去所罗门诺夫-列文半预测器独有的优良性质：普适可靠性。[31]

归根结底，仅仅提出应当尝试逼近通用性的理论理想，这个想法似乎并不比“应尽可能多地将假设纳入先验”这一建议更有实质内容。就方法论建议而言，这似乎并不深刻（我们真的需要这么多数学来得出这个结论吗？）。至少，人们希望能得到关于具体做法的进一步指引，类似于针对某种特定近似方法或通用图灵机的进一步论证。但这类论证看起来可疑地像是在为特定的归纳偏置辩护。事实上，与其试图逼近通用预测，不如认真对待这样一个教训：每一种（实际的、可计算的）学习方法都必须具有归纳偏置，并深入思考你的学习算法不可避免地具有什么样的偏置，或者应该具有什么样的偏置。[32]

4. Conclusion

Mingard 等人(2025) 提供了一个近期案例，展示了如何在探寻对深度学习泛化能力的更深入理解，尤其是对深度神经网络归纳偏置的理解时，提及所罗门诺夫归纳。[33] 作者的总体思路是运用贝叶斯分析来识别一种偏向简单性的归纳偏置，并指出“如果先验中的这种归纳偏置与结构化数据的简单性相匹配，就有助于解释 DNN 为何能如此出色地泛化”（Mingard 等人，2025，第 2 页；另见 Valle-P´erez 等人，2019）。

[31]因此，对于可计算的数据生成过程，两种不同的有限近似并不一定μ几乎必然收敛到相同的预测结果 μ。 内特(2023年)利用这一观察指出，所罗门诺夫归纳的支持者面临一个两难：他们不能一方面声称通过使用可计算的近似解决了不可计算性问题，另一方面又声称通过极限中的几乎必然收敛解决了方差问题。 [32]支持者还会补充说，应当依据简单性对这些假设加权。但即便我们接受所罗门诺夫－列文半预测器实现了客观的简单性偏差，因为它基于柯尔莫哥洛夫复杂度，任何实际的预测器都必须对应于某个可计算的描述长度函数。由于任何预测器都对应于 $some$ 描述长度函数，因此可以认为它实现了 $some$ 简单性偏差（见上文脚注 23 ）；若不进一步说明哪些压缩器是合理的，这条建议依然毫无意义。[33]这类研究所处的背景是“泛化之谜”：经典机器学习理论显然无法解释机器学习（尤其是深度神经网络）为何能成功泛化（Berner 等人，2022年）。

具体而言，作者研究了全连接神经网络利用 ReLU激活函数学习布尔函数的过程，并考察了一种先验 P，该先验基于对网络参数值进行随机采样而得。他们指出，这种先验对采样分布的不同选择并不敏感，而且表现出一种“偏向描述复杂度较低的简单函数”的倾向˜（同上，第 2 页）。更准确地说，先验概率 $P(f)$ 与以下各项 $2^{-\tilde{K}(f)}$相对应，其中 K 是一种基于 Lempel-Ziv（LZ）压缩算法的复杂度测度，可作为“真实（但不可计算的）柯尔莫哥洛夫复杂度”的替代指标（同上）。

鉴于作者强调对柯尔莫哥洛夫复杂度进行近似，他们得出这样的结论也就不足为奇了：“一个值得探索的方向，是算法信息论中关于 Solomonoff 归纳最优性的更形式化论证”（同上，第 7 页）；他们还指出：“Solomonoff 归纳与采用简单性偏置深度神经网络进行学习之间，存在一种有趣的定性相似性。两种情形下，原则上都纳入了所有假设（伊壁鸠鲁），同时又会指数级地偏好较简单的假设（奥卡姆）”（同上，补充材料，第 10 页）。[34]

尽管对 Solomonoff 归纳能带来进一步洞见的期待表述得颇为谨慎，[35] 我们的讨论表明，我们有理由对此持怀疑态度。研究发现，（某些）深度神经网络在（某些情况下）可以被视为依据一种先验运行，而这种先验类似于某种流行压缩算法所对应的分布；[36] 并且这种先验通常与现实世界的数据相吻合。当然，这是有趣且可能具有重要意义的。编码理论也可能有助于进一步分析这一现象。[37] 但认为这个特定先验是“通用简单性先验”的替代物，也就是 Solomonoff 的先验（以及它与“最优性”和奥卡姆剃刀的关联），似乎并没有什么用处，甚至可能毫无意义。与其认为 Solomonoff 归纳或许能在此发挥作用，从而补足对泛化成功的解释，不如更准确、更有用地认识到：这种与这个特定归纳偏置相吻合的特定压缩算法类型（以及，如果你愿意的话，这种特定简单性偏差）只能是一个起点。

也就是说，对于任何实际可行、可计算的学习程序，本章开篇所描绘的图景依然成立。每一种学习方法都必然具有严格受限的归纳偏置和归纳假设。贝叶斯分析有助于我们找出这些限制性假设，即它们存在于先验之中。贝叶斯哲学观点认为，这种先验绝无充分依据；这一观点落实到实践中，就是我们应当努力理解其中涉及的归纳假设。在什么样的学习情境下、针对什么样的数据，这些假设才合适——又有哪些情况下并不合适？尽管通用归纳理论颇具吸引力，深度神经网络中的归纳偏置也同样如此。

---

[34]对伊壁鸠鲁的提及源自（Li and Vit´anyi，1992，第344页及后续页；另见 Li and Vit´anyi，2019，第5.1.1节）。他们将“伊壁鸠鲁的多重解释原则（或无差别原则）”表述为：“[i]f 不止一个理论与观测结果相符，就保留所有理论。”他们还进一步将这一原则与为每种理论赋予等概率的原则联系起来（见注27）。[35]又或者，这不过是“装饰性 prose”（Keinath-Esmail and Louis，2025）。.....[36]或者确实是这类算法的类别：作者在引用（Valle-P´erez 等人，2019；Bhattamishra 等人，2023）时写道：“[o]其他复杂度度量会得出相似结果”（Mingard 等人，2025，第2页）。[37]尤其是（通过任意给定的压缩器）描述复杂度与具有该复杂度的对象数量之间的关系（Mingard 等人，2025，第4页）。

References

M. Albert. 贝叶斯学习与预期形成：一切皆有可能。载于 D. Corfield 和 J. Williamson 编，*Foundations of Bayesianism*，第 24 卷，*Applied LogicSeries*，第 341–362 页。施普林格，2001年。G. Belot. 绝对没有免费午餐！ *Theoretical Computer Science*，845：159–180，2020。S. Ben-David、N. Srebro 和 R. Urner. 通用学习与无免费午餐结果。于 *25th InternationalConference on Neural Information Processing Systems (NIPS)* 举办的研讨会 *Philosophy and Machine Learning* 上的报告，2011年。J. Berner、P. Grohs、G. Kutyniok 和 P. Petersen. 深度学习的现代数学。载于 P. Grohs 和 G. Kutyniok 编，*Mathematical Aspects of Deep Learning*，第 1–111 页。剑桥大学出版社，2022年。S. Bhattamishra、A. Patel、V. Kanade 和 P. Blunsom. Transformer模型中的简单性偏差及其学习稀疏布尔函数的能力。载于 A. Rogers、J. Boyd-Graber 和 N. Okazaki 编，*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*。计算语言学协会，2023年。D. Blackwell 和 L. Dubins. 随着信息增加而实现的意见融合。 *The Annals of Mathematical Statistics*，33：882–886，1962年。L. Blum 和 M. Blum. 迈向归纳推理的数学理论。 *Informationand Control*，28(2)：125–155，1975年。D. Campbell. 算法概率的半测度性质——“特征”还是“缺陷”？第 79–90 页。R. Carnap. *Logical Foundations of Probability*。芝加哥大学出版社，1950年。R. Carnap. *The Continuum of Inductive Methods*。芝加哥大学出版社，1952年。R. Carnap 和 W. Stegm¨uller. *Induktive Logik und Wahrscheinlichkeit*。施普林格，1959年。J. R. Carr. 理想认识论为何重要？ *Mind*，131(524)：1131–1162，2021年。N. Cesa-Bianchi 和 G. Lugosi. *Prediction, Learning and Games*。剑桥大学出版社，2006。G. J. Chaitin. 关于计算有限二进制序列的程序长度：统计考量。*Journal of the Association for Computing Machinery*，16：145–159，1969年。B. J. Copeland 和 O. Shagrir. 物理可计算性论题。载于 *Quantum, Probability, Logic: The Work and Influence of Itamar Pitowsky*，《耶路撒冷科学哲学与科学史研究丛书》，第 217–231 页。施普林格，2020年。A. P. Dawid. 基于校准的经验概率。*The Annals of Statistics*，13(4)：1251–1274，1985年。C. de Canson. 贝叶斯主义与休谟问题的推论式解决方案。 *Philosophers'Imprint*，25(39)，2025年。G. Del´etang、A. Ruoss、P. Duquenne、E. Catt、T. Genewein、C. Mattern、J. Grau-Moya、L. K. Wenliang、M. Aitchison、L. Orseau、M. Hutter 和 J. Veness. 语言建模即压缩。载于 *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*，2024年。M. C. Galavotti. 论 Hans Reichenbach 的归纳主义。*Synthese*，181(1)：95–111，2011年。M. Goldblum、M. A. Finzi、K. Rowan 和 A. G. Wilson. 观点：无免费午餐定理、柯尔莫哥洛夫复杂度及归纳偏置在机器学习中的作用。载于 *Proceedings of Forty-first International Conference on Machine Learning (ICML 2024)*，2024年。J. Grau-Moya、T. Genewein、M. Hutter、L. Orseau、G. Del´etang、E. Catt、A. Ruoss、L. K. Wenliang、C. Mattern、M. Aitchison 和 J. Veness. 学习通用预测器。载于 *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*，2024年。P. D. Gr¨unwald. *The Minimum Description Length Principle*。MIT自适应计算与机器学习丛书。麻省理工学院出版社，2007年。

C. Howson。 *Hume's Problem: Induction and the Justification of Belief*。牛津大学出版社，2000年。S. M. Huttegger。意见融合与概率运动学。 *Review of Symbolic Logic* 8 4 611 648, 2015 ，( ): –。S. M. Huttegger。 *The Probabilistic Foundations of Rational Learning*。剑桥大学出版社，2017年。Z. Keinath-Esmail 和 A. A. Louis。奥卡姆算法的用途与局限：对 Herrmann 的回应。PhilSci-Archive 预印本 24910，2025年。K. T. Kelly。 *The Logic of Reliable Inquiry*。哲学中的逻辑与计算。牛津大学出版社，1996年。K. T. Kelly。学习理论与认识论。载于 H. Arl´o-Costa、V. F. Hendricks 和 J. F. A. K. van Benthem 编， *Readings in Formal Epistemology*，计算机科学教材丛书第1卷，第695–716页。施普林格，2016年。K. T. Kelly、C. F. Juhl 和 C. Glymour。可靠性、实在论与相对主义。载于 P. Clark 和 B. Hale 编， *Reading Putnam*，第98–160页。布莱克威尔，1994年。A. N. Kolmogorov。信息定量定义的三种途径。 *Problems of Information Transmission* 1 1 1 7, 1965 ，( ): –。俄文原文译文见*Problemy Peredachi Informatsii* 1(1):3-11，1965年。重印于 Shiryaev (1993)，184-193页。J. Leike 和 M. Hutter。论所罗门诺夫归纳与求知的可计算性。载于 K. Chaudhuri、C. Gentile 和 S. Zilles 编， *Algorithmic LearningTheory: Proceedings of the Twenty-Sixth International Conference (ALT 2015)*，计算机科学教材丛书第9355卷，第364–378页，施普林格，2015。L. A. Levin。 *Nekotorye teoremy ob algoritmicheskom podkhode teorii veroyatnostej i teorii informatsii* 1971 。博士论文，莫斯科国立大学，。译为：《概率论与信息论算法方法的若干定理》。*Annals of Pure and Applied Logic*，162:224-35，2010年。M. Li 和 P. M. B. Vit´anyi。归纳推理与柯尔莫哥洛夫复杂度。， ( ): –。M. Li 和 P. M. B. Vit´anyi。 *An Introduction to Kolmogorov Complexity and Its Applications* 2019 。计算机科学教材丛书。施普林格，第四版，。J. Martin、T. Everitt 和 M. Hutter。通用人工智能中的死亡与自杀。载于 B. Steunebrink、P. Wang 和 B. Goertzel 编， *Proceedings of the Ninth International Conference on Artificial General Intelligence (AGI2016)* 9782 *Lecture Notes in Computer Science* 23 32 2016 ，第 卷，的，第 –页。施普林格，。C. Mingard、H. Rees、G. Valle-P´erez 和 A. A. Louis。深度神经网络内置奥卡姆剃刀。 *Nature Communications*，16，220:1–9，2025年。S. Neth。所罗门诺夫预测的两难困境。 *Philosophy of Science* 90 2 288 306, 2023 ，( ): –。M. Nielsen。确定性收敛与强正则性。 *The British Journal for the Philosophy of Science* 71 4 1461 1491, 2020 ，( ): –。H. Putnam。“确证度”与归纳逻辑。载于 P. A. Schilpp 编， *ThePhilosophy of Rudolf Carnap*， *The Library of Living Philosophers*第 XI 卷，第761–783页。Open Court，1963年。S. Rathmanner 和 M. Hutter。通用归纳的哲学论述。 *Entropy*，13 6 1076 1136，2011 ( ): –。W. C. Salmon。Hans Reichenbach 对归纳的辩护。 *Erkenntnis*，35:99–122，1991年。M. J. Schervish。关于 Dawid (1985 *The Annals of Statistics* 13 4 1274 1282,) 的评论。， ( ): –。1985 G. Schurz。 *Hume's Problem Solved: The Optimality of Meta-Induction*。麻省理工学院出版社，2019年。S. Shalev-Shwartz 和 S. Ben-David。 *Understanding Machine Learning: From Theory to Algorithms* 2014 。剑桥大学出版社，。A. K. Shen、V. A. Uspensky 和 N. K. Vereshchagin。 *Kolmogorov Complexity and Algorithmic Randomness* 220 ， *Mathematical Surveys and Monographs*第 卷。美国

数学学会，2017年。A. N. Shiryaev 编。

Selected Works of A.N. Kolmogorov. Volume III: Information Theory and the Theory of Algorithms 27 *Mathematics and Its Applications*，第 卷，。*(Soviet Series)* 1993 Kluwer，。R. I. Soare。

Turing Computability: Theory and Applications 4 *Theory and*，第 卷，。*Applications of Computability* 2016 施普林格，。R. J. Solomonoff。归纳推理形式理论。第一部分和第二部分。*Information and Control* 7 1 22, 224 254, 1964 ，：––。

R. J. Solomonoff。算法概率的发现。*Journal of Computer and System Sciences* 55 1 73 88, 1997 ，():–。

T. F. Sterkenburg。所罗门诺夫预测与奥卡姆剃刀。 *Philosophy of Science*，83 (4):459–479，2016年。T. F. Sterkenburg。算法概率的广义刻画。*Theory of Computing Systems* 61 4 1337 1352, 2017 ，():–。

T. F. Sterkenburg。 *Universal Prediction: A Philosophical Investigation*。博士论文，格罗宁根大学，2018年。T. F. Sterkenburg。普特南对角论证与通用学习机的不可能性。 *Erkenntnis*，84(3):633–656，2019。T. F. Sterkenburg。归纳的元归纳论证。*Episteme* 17 4 519 541, ，():–。2020

T. F. Sterkenburg 和 P. D. Gr¨unwald。监督学习的无免费午餐定理。*Synthese*，199(3):9979–10015，2021年。P. Suppes。 *Representation and Invariance of Scientific Structures*。CSLI出版社，2002年。G. Valle-P´erez、C. Q. Camargo 和 A. A. Louis。深度学习之所以具有泛化能力，是因为参数-函数映射偏向于简单函数。arXiv 预印本 1805 08522，2019年。。A. G. Wilson。深度学习并没有那么神秘或与众不同。arXiv 预印本 arXiv:2503.02113，2025年。I. Wood、P. Sunehag 和 M. Hutter。通用先验的（非）等价性。第417–425页。C. Wyeth 和 M. Hutter。通用人工智能中的无知情形下的价值。载于 M. Ikl´e、A. Kolonin 和 M. Bennett 编，*Proceedings of the Eighteenth International Conference on Artificial General Intelligence (AGI2025)* 16058 *Lecture*，第 卷，*Notes in Computer Science* 338 349 2026 的，第 – 页。施普林格，。N. Young 和 M. Witbrock。Transformer模型作为所罗门诺夫归纳的近似。载于 M. Mahmud、M. Doborjeh、K. Wong、A. C. S. Leung、Z. Doborjeh 和 M. Tanveer 编， *Proceedings of the 31st International Conference on Neural Information Processing (ICONIP 2024), Part I*，计算机科学教材丛书第15286卷，第16–25页。施普林格，2025年。A. K. Zvonkin 和 L. A. Levin。有限对象复杂性，以及借助算法理论发展信息与随机性概念。*Russian Mathematical Surveys*，26(6):83–124，1970年。俄文原文译文见 *Uspekhi Matematicheskikh Nauk*，25(6):85-127，1970年。

Munich Center for Mathematical Philosophy (MCMP), LMU Munich
Munich Center for Machine Learning (MCML)*Email address*:
tom.sterkenburg@lmu.de