arXiv:2603.01309v1 [cs.LG] 1 Mar 2026

强化学习的 PAC 保证：样本复杂度、覆盖与结构

Joshua Steier*独立
研究者

2025年10月3日

摘要

当数据稀缺或失误代价高昂时，平均情形指标并不足够。实践者需要的是一种保证：经过 N个回合后，学习策略以至少 $1-\delta$的概率与最优相差不超过 ε。这就是 PAC 承诺；在2018至2025年之间，强化学习理论界就此类承诺何时能够兑现取得了显著进展。我们综述这一进展。我们的组织工具是本文提出的覆盖—结构—目标（CSO）框架，它将几乎每一项PAC样本复杂度结果分解为三个因素：覆盖（数据如何获得）、结构（内在的马尔可夫决策过程复杂度或函数类复杂度）以及目标（学习者必须交付的内容）。CSO 并非定理，而是一种解释性模板，用以识别瓶颈，并使跨设定比较一目了然。技术核心涵盖紧的表格基准以及通向遗憾的一致 PAC 桥梁；支配函数逼近下可学习性的结构复杂度度量（贝尔曼秩、见证秩、Bellman-Eluder维数）；线性、核/神经切线核与低秩模型的结果；作为前期覆盖投入的无奖励探索；以及以所继承覆盖为约束瓶颈的悲观离线强化学习。我们提供实践者工具：按 CSO 坐标索引的速率查找表、贝尔曼残差诊断、带部署门槛的覆盖估计，以及逐回合策略证书。最后一节梳理开放问题，将近期目标与前沿问题区分开来；在这些前沿问题中，覆盖、结构与计算以当前理论无法化解的方式纠缠在一起。

关键词 PAC强化学习；一致–PAC；样本复杂度；函数逼近；离线强化学习；低秩马尔可夫决策过程；Bellman–埃鲁德维度。

1 引言

如何阅读本综述。 主要关心何时使用何种方法 的读者，可先从表2、CSO 模板（第2节）以及实用决策树（第14节）入手，再跳转到相关设置：表格基准（第4节）、结构复杂度度量（第5节）、函数逼近（第6节）、丰富观测与低秩（第7节）、无奖励探索（第8节）、离线强化学习（第9节）、部分可观测性（第10节）或PAC-贝叶斯（第11节）。寻求技术深度的读者应阅读第3节以及各章节开头的定义块。我们将证明保持在非正式层面（依赖关系与缩放关系），并指向原始文献以获取精确常数。图1和图2连同表2和表3提供了一份快速分流指南。

1.1 强化学习以及对固定置信度保证的需求

在强化学习（RL）中，智能体在离散的时间步内与环境交互。在每一步，智能体观察一个状态 s （例如机器人的传感器读数或患者的生命体征），选择一个动作 a （例如电机指令或治疗选择），获得一个标量奖励 r，并转移到一个新状态s'。智能体的目标是学习一种策略，即一条将状态映射为动作的规则 π ，以便在规划时域 H内尽可能多地累积奖励。核心矛盾在于探索 （尝试不熟悉的动作

*joshsteier@gmail.com

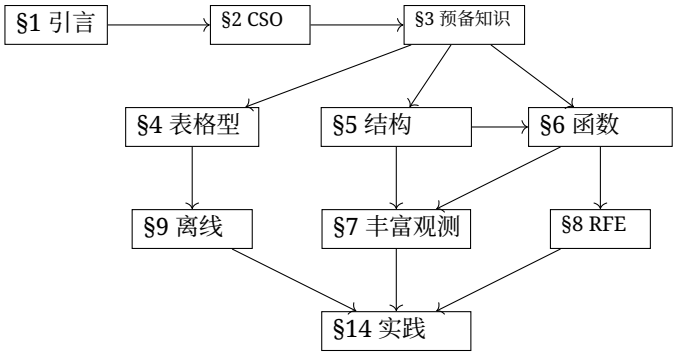

图 1：**综述路线图**。 CSO 框架（§2）统筹所有结果。预备知识（§3）介绍形式化工具。三大支柱（表格型 §4、结构度量 §5、函数逼近 §6）支撑应用（§7–§9）。实践工具包（§14）对其进行综合。**阅读路径：** 从业者从 §2 和 §14 开始，然后阅读各领域章节；理论学家从 §3 开始，接着阅读 §4–§6。

以了解其后果）与利用 （重复已知效果良好的动作）之间的取舍。Sutton和Barto [38], Szepesvári [39], 以及Kaelbling、Littman和Moore [24]对这一问题进行了奠基性研究。

大多数强化学习研究通过遗憾来衡量性能：这是智能体在学习过程中采取次优动作所造成的累积损失。遗憾是一种平均情形指标，非常适合偶尔犯错可以接受、且智能体将与环境长期交互的设置。但许多现实问题并不符合这一模式。临床试验无法承受在算法收敛期间，让数千个回合接受有害治疗组。自动驾驶车辆控制器必须从一开始就可靠运行，而不能仅仅在其整个生命周期内表现平均良好。根据医院记录学习得到的离线策略，除非我们能够以量化置信度证明其优于现状，否则不应部署。

这些设置需要固定置信度保证。形式上，如果某算法以至少 $1-\delta$ 的概率，在指定数量的回合 $N(\varepsilon,\delta)$ 后输出一个价值距最优值不超过 ε 的策略，那么该算法就是 (ε,δ)-PAC（可能近似正确）。PAC框架起源于计算学习理论，源自 Valiant 关于监督分类的研究 [44] ，之后由 Strehl、Li 和 Littman [36], 引入序贯决策领域，他们为有限马尔可夫决策过程（MDP）建立了首批系统性的 PAC 分析。本综述探讨三个问题：强化学习领域有哪些固定置信度保证？哪些假设使这些保证成为可能？这些保证如何随问题参数变化？

1.2 范围与时间窗口

我们聚焦于 2018 至 2025年这一时期，因为在此期间，三项进展汇聚起来，重塑了 PAC强化学习的格局。

首先，Dann、Lattimore 和 Brunskill [4] 提出的一致 *PAC*框架表明，PAC 与遗憾并非相互竞争的目标，而是同一统计挑战的两种视角。一致PAC算法能够同时界定所有精度水平 ε下的 ε 次优回合数，并由此自动推出高概率遗憾界（定理1）。这一联系在 2017 年末得以明晰，并为之后的一切奠定了基础。

其次，结构复杂度度量在 2017 年至 2021 年间逐渐成熟。Jiang 等人 [19] 提出了贝尔曼秩，Sun 等人 [37] 提出了见证秩，Jin 、 Liu 和 Miryooseofi [22] 定义了贝尔曼 - 埃鲁德维度。总体而言，这些度量刻画了函数逼近何时能够实现多项式样本复杂度，并以依赖问题的容量参数取代了对状态数量的蛮力依赖。

第三，以覆盖为中心的 表述重塑了我们对数据访问的思考方式。Jin 等人 [20] 形式化了无奖励探索，其中智能体在指定任何奖励之前构建可复用的数据集。Jin、Yang 和 Wang [23] 与 Shi 等人 [35] 发展了悲观离线强化学习，其显式依赖于训练数据对目标策略重要状态的覆盖程度。Li 等人 [29] 解决了极小极大样本

复杂度，针对表格型设定中基于模型的离线学习，并且 Zhang 等人 [48] 解决了在线表格情形。

2018年前的基础内容已由现有参考文献充分涵盖：Strehl、Li 和 Littman [36]介绍了经典 PAC-MDP 分析，Szepesvári [39] 阐述了算法与分析基础，Kaelbling、Littman和Moore [24] 提供了早期的综合综述，Sutton和Barto [38] 则介绍了更新的算法成果。本文以这些工作为基础，而不再赘述。

1.3 核心概念简述

第 §3 节将详细介绍形式化框架；此处我们先概述关键思想，帮助读者了解全文。

有限时域马尔可夫决策过程 $M=(\mathcal{S},\ \mathcal{A},\ P,\ r,\ H,\ \rho)$具有状态空间 $\mathcal{S}$（其中 $|\mathcal{S}|=S$）、动作空间 $\mathcal{A}$（其中 $|\mathcal{A}|=A$）、阶段相关的转移和奖励、时域 H以及初始分布 ρ。最优策略 $\pi^{\star}$使累积奖励最大化；差距 $V_1^{\star}(\rho)-V_1^{\hat{}\pi}(\rho)$用于衡量候选策略 $\hat{}\pi$的次优性。

$(\varepsilon,\ \delta)$-PAC算法保证，在经过 $N(\varepsilon,\ \delta)$个回合后，这一差距至多为 ε，且该保证以概率 $1-\delta$成立（定义 1）。一致 PAC 进一步将这一保证扩展到所有 ε 同时成立（定义 2），由此可推出高概率遗憾保证（定理1）。

超越表格化情形后，样本复杂度取决于马尔可夫决策过程或函数类的结构属性。相关参数包括特征维度 d （线性模型）、贝尔曼秩 B、见证秩W、贝尔曼-埃鲁德维数 d_{BE}、有效维度 d_{eff}（λ）（核模型），以及潜在秩 r（低秩马尔可夫决策过程）。这些参数分别取代表达式中的表格因子 SA，而它们之间的层级关系（图3、表 3）是反复出现的主题。

在离线强化学习中，智能体无法收集新数据，集中系数 $C_{\star}$（定义 6）用于量化训练分布对目标策略所访问状态的覆盖程度。较大的 $C_{\star}$ 会增加样本需求，不受结构简洁性的影响；当 $C_{\star}$ 为无穷大时，若无进一步假设，就不可能实现一致的策略改进。

1.4 本综述的贡献

四项产出使本综述有别于以往的论述。

第一项是**CSO 框架**（§2），即本综述提出的一种阅读视角，它将每一个PAC界分解为覆盖、结构与目标三类因素。CSO 分解并非定理，而是一种组织原则，它使跨设定比较系统化，并在保证空洞时识别出哪个因素构成瓶颈。我们将在§2中详细展开该框架，并用它来统领本文其余部分。

第二项是**统一技术综合**，涵盖通常被孤立处理的各类设置。我们采用统一的符号表示，陈述表格型、线性、核、低秩、无奖励以及离线场景下的结果，并明确参数依赖关系，同时指向原始文献以获取精确常数。表2按各项主要结果的 CSO 坐标进行交叉索引。

第三项是一套**实践者工具**：用于检验可实现性与贝尔曼完备性的诊断程序（算法1）、带有有限样本部署门槛的覆盖率估计算法（算法 2），以及提供逐回合问责的策略证书（定义 9）。这些工具将理论操作化，供应用研究人员判断某项PAC 保证是否与其问题相关，以及其数据是否支持部署。

第四部分是**对开放问题的结构化梳理**（§13），按时间范围（近期与前沿）组织，并附有不可能性结果，指出哪些方面的进展需要新思路，而非渐进式改进。

1.5 为综述奠定基础的典范结果

为帮助读者尽早把握全貌，我们重点介绍四项贯穿综述始终的结果，展示 CSO 模板的适用范围。精确的表述和证明见相关章节；此处我们着重说明这些结果之间的依赖关系及其揭示的内容。

第一个基准是**表格型极小极大速率**。Zhang 等人 [48] 证实，在具有 S 个状态、 A 个动作且时域为 H 的有限MDP中，学习一个 ε-最优策略需要 $\tilde{\Theta}(SAH^3/\varepsilon^2)$个回合（定理 2）。

这是目前任何强化学习设定中已知最紧的 PAC 界，每个结构化结果都必须将其作为一个特例恢复出来。从 CSO 视角来看，覆盖是平凡的（智能体在线探索），结构为 SA，目标是标准 PAC 控制。

第二个是**一致 PAC 到遗憾桥接**。Dann、Lattimore 和 Brunskill [4] 表明，任何以预算 $N(\varepsilon,\delta)$ 满足一致 PAC 的算法，在经过 K 个回合后可达到高概率遗憾 $\mathcal{O}(\int_0^H \min\{K, N(\varepsilon,\delta)\}\, d\varepsilon)$（定理1）。当 N 对 $(S, A, H, 1/\varepsilon)$ 具有多项式依赖时，这可恢复近极小极大遗憾。该桥接意味着 PAC 分析与遗憾分析相互启发，我们会反复调用这一转换。

第三个锚点是**具有线性特征的函数逼近**。Jin、Yang 和 Wang [21] 表明，在维度为 d 的线性可实现性下，乐观最小二乘价值迭代可达到 PAC样本复杂度 $\tilde{\mathcal{O}}(d^3H^4/\varepsilon^2)$（定理8）。时域指数从 H^3 上升到 H^4，因为共享特征的各状态之间的估计误差是相关的。He、Zhou 和 Gu [13] 在贝尔曼完备性下将其扩展到一致 PAC 设置。Jin、Liu 和 Miryoosefi [22] 的 Bellman-Eluder维数进一步加以推广，为丰富函数类下的可学习性提供了最广泛的充分条件（定理6）。

第四种是**通过悲观性实现离线强化学习**。在不进行进一步交互、仅从固定数据集中学习时，起约束作用的因素会从探索转为覆盖。Shi 等人 [35] 证明，在满足线性可实现性和集中性 $C_\star$ 的条件下，悲观 Q 学习以 $\tilde{\mathcal{O}}(\mathrm{poly}(d,\ C_\star,\ H)/\varepsilon^2)$ 个样本实现控制误差 ε（定理 15）。Li 等人 [29] 则在极小极大层面解决了基于模型的表格情形。在 CSO 模板中，覆盖 poly($C_\star$) 取代了在线情形下平凡的因子 1；而无论结构多么简单，数据支持不佳都可能使保证变得空洞。

1.6 不同设置之间的联系

有几条线索贯穿这篇综述，值得在技术内容展开之前指出。

统一 PAC 框架不仅是理论上的便利工具，也是一种转换工具。第 4–6 节中的任何一致 PAC 结果都能自动导出一个遗憾界；反过来，足够精确的遗憾分析也可以转换为 PAC 保证。这种双向联系意味着，即使读者关注的是遗憾，也能从以 PAC 为重点的章节中有所收获，反之亦然。

覆盖有两种概念上互为对偶的形式。在无奖励探索（§8）中，智能体构建 覆盖，将其作为可复用资源，为此需要预先付出代价（在表格情形下，额外增加一个 S 因子），以支持任意下游奖励。在离线强化学习（第9节）中，覆盖继承 自行为策略，并由 $C_\star$ 量化；智能体无法改善覆盖，只能通过采取悲观策略来防范其局限性。认识到这种对偶性，有助于厘清何时应投入探索，何时应接受既有数据。

结构复杂度度量构成一个严格的层级（图3）。表格型速率最为紧致；线性马尔可夫决策过程速率放宽至维度 d；低秩类和双线性类进一步推广；而 Bellman-Eluder维数则提供了最宽泛的统一框架。沿着这一层级前进，是以更宽的适用范围换取较宽松的常数；实践者的任务是将问题定位在能够适用的最具体层级，从而获得现有最紧致的保证。

依赖实例的识别（§4）带来了另一种改进：用间隙加权和替代最坏情形下的结构因子；当最优策略与备选动作之间区分明显时，间隙加权和可能小得多。我们在 §13 中指出，将这一方法扩展到函数逼近设置是一个开放问题。

1.7 对应用研究人员的实践启示

理论中涌现出若干可付诸实践的经验，我们在此加以陈述，以便从业者在阅读技术章节时牢记这些经验。

应用任何 PAC 保证之前，都应先确认其基本假设成立。线性马尔可夫决策过程界假设奖励和状态转移关于已知特征呈线性；若将这些界用于任意深度网络特征，就会违反可实现性，并可能导致置信度判断产生误导。§14.1（算法1）中的贝尔曼残差诊断提供了一项具体检验：使用随机策略数据通过岭回归拟合价值函数，并检查留出贝尔曼残差是否随时域增长。如果残差确实增长，则说明线性函数类存在模型设定错误，需要采用更丰富的特征或核方法。

在奖励函数并非预先固定的多任务设置中，无奖励探索（§8）为临时起意的数据收集提供了一种有原则的替代方案。预先投入的探索预算会构建一个数据集，支持对任意奖励进行 ε-最优规划，从而将成本分摊到各项任务中。

在离线流程中，覆盖是把关因素。按照 §14.2 中的流程估计覆盖代理（密度比、岭杠杆分数），并在着手进行策略改进之前应用算法 2 的部署门槛。覆盖不足时，应优先进行离策略评估（采用区间估计，而非点估计改进），并弃权，暂不部署，直到能够收集更多数据。

Dann 等人 [6], 提出的策略证书（定义 9）能够实现逐回合问责：它给出一个依赖数据的次优性界，可实时监控。只有当证书值低于预期容差时才部署；如果证书值仍然很大，说明算法需要更多数据。

1.8 开放问题一览

引言最后，我们简要预览 §13 中展开讨论的开放问题，并按 CSO 轴进行组织。

在结构轴上，最亟待解决的缺口是：针对核函数类和过参数化函数类，在条件可由数据验证而非先验假设的情况下，给出统一 PAC 保证。当前结果要求满足 Bellman 完备性，而这一条件在实践中很少能够验证；我们的目标是设计样本复杂度随有效维度 $d_{\text{eff}}(\lambda)$ 缩放的算法，并通过残差检验验证其完备性假设。

在覆盖轴上，模型设定错误情况下的离线强化学习仍缺乏充分理解。现有保证假设精确可实现性；当函数类存在模型设定错误时，误差可分解为逼近、估计和覆盖度分量，但对于这一三方权衡的精确刻画以及可处理算法，目前仍然缺乏。

在结构与目标的交汇处，结合函数逼近的实例依赖识别几乎尚未得到探索。针对线性或低秩模型的基于间隙的最优策略识别，若能达到按可达性加权的表格型下界，将是一个自然而且看起来可行的下一步。

最后，数据驱动的结构选择（根据观测数据在线性、核和低秩类之间进行选择，并提供预言机不等式和统一 PAC 保证）将使理论图景更贴近实际部署；在实际部署中，正确的结构假设往往无法预先得知。

1.9 本综述与先前工作的不同之处

表 1 展示了本综述与最接近的两部先前著作之间的关系。Szepesvári 的专著 [39] 对截至约 2010 年的强化学习进行了深入的算法性论述，涵盖价值迭代、时序差分学习和策略梯度方法，并讨论了 PAC 分析和遗憾分析，主要聚焦于表格设定。Strehl、Li 和 Littman [36] 建立了有限马尔可夫决策过程的 PAC-MDP 框架，并推导出首批系统性的样本复杂度界，确立了此后所有研究所依据的经典基线。

我们的综述从这些研究的未竟之处展开。我们涵盖 2018—2025 年期间，在此期间，一致 PAC、结构复杂度度量、无奖励探索和悲观离线强化学习改变了这一领域。本文提出的 CSO 框架提供了一种贯穿各领域的组织原则，这是前人的工作所不具备的。此外，实践者工具（覆盖度门控、残差诊断、基于证书的部署规则）是旨在弥合理论与实际应用之间差距的新贡献。

2 覆盖-结构-目标框架

2.1 为何需要统一的视角

2018年至2025年的 PAC 强化学习文献涵盖了丰富多样的设置：表格型和连续状态空间、在线和离线数据访问、最坏情况和实例依赖目标、基于价值和基于模型的算法。每篇论文都引入了自己的符号表示、假设和复杂度度量，使读者难以比较不同设置下的结果，也难以判断某项保证是否适用于特定问题。

	Szepesvári (2010)	Strehl等人 (2009年)	本综述 (2018–2025年)
发表场所/类型	专著 (Morgan & Claypool)	JMLR 综述文章	TMLR 综述（2025年）
范围	强化学习算法；主要针对表格型和基础函数近似	PAC-MDP 框架；有限马尔可夫决策过程；样本复杂度	一致 PAC，涵盖表格型、结构度量、函数逼近、RFE，离线、部分可观测马尔可夫决策过程
时间窗口	近似 2010 年之前	2009 年之前	2018–2025年
主要轴线	算法与理论	PAC-MDP 样本复杂度	CSO：覆盖 × 结构 × 目标
保证类型	PAC/遗憾（主要是禁忌搜索-lar)	有限情形下的 PAC-MDP 界 状态-动作空间	一致 PAC，实例-离线/悲观主义，PAC-贝叶斯
实践者工具	概念性算法	理论基础	决策树；覆盖度门控；证书；速率表

表 1：**与奠基性综述的比较**。我们在 Szepesvári 的算法处理和 Strehl 等人的 PAC-MDP 框架基础上，纳入 2018–2025 年 的研究成果，并以统一的 CSO 视角提供面向部署的工具。

现有综述通常按算法类别（基于模型与无模型）或设置（表格型与函数逼近）来组织结果。这些划分维度很自然，但还不够：它们没有解释为何数据来源从在线转为离线时，样本复杂度会发生变化；也没有解释为何同一个结构参数（例如特征维度 d）在无奖励探索中的作用方式与在悲观控制中的不同。想要知道“我有一个包含线性特征的离线数据集，并且需要一个 ε-最优策略；目前可获得的最佳速率是什么？”的读者，必须翻阅符号表示各不相同的论文，拼凑出答案。

本综述提出了一种替代性的组织原则，称为**覆盖—结构—目标（CSO）框架**。CSO 框架并非定理，也未见于任何一篇先前论文。它是本综述作为阅读视角所作的贡献：我们观察到，在 2018–2025年 的文献中，PAC界反复呈现乘法结构，并将这一观察提升为一种有意识的组织工具。

2.2 三个维度与一般速率

我们注意到，在第4至第11节综述的几乎所有 PAC样本复杂度界，都可以解析为三个具有可解释性的因子的乘积，再乘以多对数项和时域多项式：

$$N(\varepsilon,\delta) \;\approx\; \underbrace{\mathsf{Cov}}_{\text{coverage}} \;\times\; \underbrace{\mathsf{Comp}}_{\text{structure}} \;\times \mathrm{poly}(H) \;\times\; \varepsilon^{-2} \;\times\; \log(1/\delta). \tag{1}$$

此处的符号 $\approx$表示在 $(S,\ A,\ H,\ 1/\varepsilon,\ 1/\delta)$ 中相差多对数因子，并且在 H 中仅精确到多项式次数；该次数随设置而异。我们写作 poly(H) 时，指的是时域的某个特定次幂（例如，表格情形下的 H^3 、线性马尔可夫决策过程下的 H^4 ）。该模板有意将其抽象化，以便清晰呈现这三个因子的分解。精确指数在相关章节中给出，并汇总于表2。

现在我们来说明每个因子所代表的含义。

覆盖（Cov）反映了数据的获取方式，以及数据对目标策略的支持程度。在线设定中，智能体通过探索自行创造覆盖，因此 Cov $=1$：覆盖不是瓶颈。生成模型设定允许学习者按需查询任意状态-动作对，也能提供 Cov $=1$。在离线强化学习中，覆盖取决于生成数据集的行为策略，并由集中系数 $C_\star$ （定义 6）量化。当 $C_\star$ 较大时，数据集无法为与最优策略相关的状态提供多少信息，而 Cov =poly($C_\star$) 可能主导整个样本复杂度表达式。在无奖励探索中，覆盖既非现成可得，也非免费获得，而是构建为一种可复用资源：探索预算为所有可能的下游奖励提供广泛的支撑集，而表格型无奖励探索界（定理 13）中额外的 S 因子则体现了这项投入。

结构（复杂度）衡量马尔可夫决策过程或用于近似价值或模型的函数类的内在复杂度。在表格情形中，Comp $= SA$，即转移函数和奖励函数中的自由参数数量。在线性特征维度为 d时，Comp $=$poly(d)，对于标准的 LSVI-UCB 界，通常为 d^3 。对于核或再生核希尔伯特空间价值函数类，Comp $= d_{\text{eff}}(\lambda)$，即正则化水平为 λ时的有效维度，该维度受核积分算子的谱衰减控制。对于低秩转移，Comp $=$poly(r)，其中 r 为分解秩。§5 中综述的结构复杂度度量（贝尔曼秩 B、见证秩 W、Bellman-Eluder维数 d_{BE}、双线性秩）都是这一维度的具体实例，形成了从紧致但适用范围窄（表格型）到宽泛但较松（有限 d_{BE}）的层级。

目标（Obj） 规定了学习者必须实现什么。基准是 (ε, δ)-PAC 控制：以 $1-\delta$ 的概率输出一个 ε-最优策略。一致 PAC 将其加强为对所有 ε 同时提供保证，这意味着高概率遗憾（定理1）。依赖实例的识别用间隙加权和取代最坏情况下的结构因子，而间隙加权和可能小得多（§4）。离策略评估要求的是准确的价值估计，而非最优策略。目标的每种变化都会修改 ε^{-2} 核心或时域多项式，但乘积结构得以保留。

这种分解的价值在于组织安排，而非数学层面。它告诉读者，当界限过大时该调整哪个因素：通过使用更好的特征来降低结构项，通过收集更多样化的数据来改善覆盖，或者将目标从控制放宽为评估。此外，它还让跨设定比较变得直接，因为如果两个结果仅在一个 CSO 坐标上有所不同，只需查看相应因子即可进行比较。

2.3 展示该模板如何发挥作用的四个示例

为了让这一模板具体化，我们将其应用于综述后文将介绍的四个典范结果。逐步变化展示了每次改变一个 CSO 坐标时，样本复杂度如何以可预测的方式改变。

表格在线 PAC（§4）。在在线访问条件下，智能体可以自由探索，因此 Cov $= 1$。有限 MDP 的结构复杂度为 Comp $= SA$，时域以 H^3 的形式进入。由此得到的速率为 $N = \tilde{\Theta}(SAH^3/\varepsilon^2)$，Zhang 等人 [48] 证实该速率是极小极大最优的。这是最简单的 CSO 单元，也是所有其他结果的校准基准。

采用 LSVI-UCB 的线性马尔可夫决策过程（§6）。保持在线访问（Cov $= 1$），但将表格结构替换为线性特征后，结构项变为 Comp $= d^3$ ，时域指数升至 H^4。速率为 $N = \tilde{\mathcal{O}}(d^3H^4/\varepsilon^2)$ [21]。从 H^3 到 H^4 的增长，是函数逼近带来的实际代价：共享特征的状态之间存在相关的估计误差，而这些相关性会通过贝尔曼备份不断累积。与表格情形相比，我们用 d^3 换掉了 SA 因子；当 $d \ll SA$ 时，这是显著的提升。

具有悲观性的离线线性马尔可夫决策过程（第9节）。现在将访问模式从在线改为离线，同时保留线性结构。覆盖轴从 Cov $= 1$ 变为 Cov $=$ poly($C_\star$)，其中 $C_\star$ 衡量最优策略的状态访问偏离数据分布的程度。速率变为$n = \tilde{\mathcal{O}}(\text{poly}(d,\ C_\star,\ H)/\varepsilon^2)$ [23, 35]。此时瓶颈在于覆盖：即使有 $d = 5$，如果 $C_\star = 10^4$，界仍然是空泛的，再简洁的结构也无法弥补这一点。这个例子说明，在部署离线策略之前，进行覆盖诊断（第14.2节）至关重要。

无奖励探索，表格型（§8）。回到在线访问和表格结构，但改变目标：智能体必须构建一个数据集，以支持针对 任意奖励函数的 ε-最优规划，而不只是针对某一个奖励函数。这一强化后的目标使覆盖轴膨胀为 Cov $= S$ （覆盖所有可能下游奖励对应的占用测度的成本），从而得到 $N = \tilde{\mathcal{O}}(S^2A\text{poly}(H)/\varepsilon^2)$ [20]。相较于标准在线学习，额外的 S 因子是一项覆盖投入，可在任意多个下游任务之间摊销；在多奖励流水线中，这笔前期成本可能远低于针对每个奖励分别运行探索。

2.4 CSO视角的优势与局限

我们主张 CSO 分解具有三项优势。

首先是可导航性。读者只需指定 CSO 坐标，就能在综述中找到任何结果：确定访问模式（在线、生成式、离线、无奖励），以确定覆盖；描述函数类（表格型、线性、核、低秩、一般），以确定结构；并确定学习目标（PAC控制、一致 PAC、识别、评估），以确定目标。表2通过这些坐标交叉索引了所有主要结果，而 §2.5 中推荐的工作流程将逐步讲解这一过程。

其次是诊断。当保证变得空泛时，CSO解析可以识别问题所在的轴，并提出补救措施。如果结构项占主导地位，就应投入资源改进特征或采用更丰富的函数类。如果覆盖占主导地位，就应收集更多样化的数据，或将目标从控制放宽为评估。如果问题出在时域多项式上，则可考虑采用状态抽象或时间层级，以缩短有效时域。

第三个是前沿识别。§13 中最棘手的开放问题，恰恰是多个维度相互作用的问题。具有设定错误特征的离线强化学习将覆盖与结构耦合在一起：仅改进特征或仅改进数据都无法解决问题。无先验低秩学习将结构与计算耦合在一起：其统计速率已为人所知，但要实现多项式时间算法，还需要额外的假设。采用函数逼近的实例依赖型识别将结构与目标耦合在一起：若没有衡量高维空间中可达性的新工具，表格型设定中的间隙加权速率便无法迁移。

本框架也有明显的局限性，我们在此指出这些局限，以免过度宣称。它没有涵盖计算复杂度：即使两个设置具有相同的 CSO 坐标，是否存在多项式时间算法也可能大不相同。它忽略了不同设置之间有实质差异的时域指数（在实践中，H^3 、 H^4 和 H^6 可能会产生影响）。它不适用于非 PAC 目标，例如贝叶斯遗憾，或未使用 PAC 封装的纯探索简单遗憾。此外，乘法分解只是近似的；在某些设置中，覆盖与结构会通过对数的乘积相互作用，或通过无法清晰分解的共享参数相互作用。我们会在后续章节涉及这些局限时加以指出，并在提供 CSO 解析的同时列出精确指数。

2.5 使用该框架：工作流程

面对新的强化学习问题时，我们建议遵循以下流程。

首先确定访问模式。智能体能否在线与环境交互，还是只有固定数据集？是否有生成模型（模拟器）可用？奖励是否会在之后指定，这意味着采用无奖励设定？答案决定了覆盖轴：在线访问和生成式访问给出 Cov $= 1$；离线访问会引入 Cov $=$poly($C_\star$)；无奖励探索具有更高的覆盖成本，但该成本可以在多个任务之间摊销。

接下来，刻画函数类。问题是否为表格型（状态空间和动作空间较小且有限）？是否有特征可用；如果有，这些特征是线性的、核化的，还是学习得到的？环境是否具有潜在低秩结构？答案决定了结构轴，并指向综述中相关的章节。

然后确定目标。目标是找到单个近最优策略（PAC控制），还是保证在所有精度水平下都具有良好性能（一致 PAC），或者以依赖实例的样本复杂度识别最佳策略，还是在不进行改进的情况下评估给定策略？

掌握这三个坐标后，就可以从表2及相应章节中查出速率。在使用保证之前，先通过第14节中的诊断验证基本假设：通过贝尔曼残差检验（算法1）检查可实现性，通过密度比和杠杆分数估计覆盖，并依据策略证书（定义 9）决定是否部署。

本综述接下来将针对每个 CSO 单元具体展开这一工作流程，首先从第3节的正式定义开始。

3 预备知识：PAC 学习与序贯决策

可能近似正确（PAC）学习这一概念起源于计算学习理论，源自 Valiant 针对监督分类提出的框架 [44]。PAC 学习者保证，以至少

设置	Cov	Comp	多项式(H)	Rate	章节
表格型在线	1	SA	H^3	$\tilde{\Theta}(SAH^3/\varepsilon^2)$	§4
表格型批量策略迭代	1	$\sum \frac{q_h}{\Delta^2}$	变化	依赖于实例	§4
线性马尔可夫决策过程	1	d^3	H^4	$\tilde{\mathcal{O}}(d^3H^4/\varepsilon^2)$	§6
核/再生核希尔伯特空间	1	$d_{\text{eff}}(\lambda)$	H^4–H^6	$\tilde{\mathcal{O}}(d_{\text{eff}}\cdot$多项式$(H)/\varepsilon^2)$	§6
低秩马尔可夫决策过程	1	多项式(r)	有所不同	$\tilde{\mathcal{O}}$(多项式$(r,\ H)/\varepsilon^2)$	§7
块状马尔可夫决策过程	1	poly$(m)\cdot A$	变化	$\tilde{\mathcal{O}}$(多项式$(m)AH/\varepsilon^2)$	§7
Bellman 秩 B	1	poly(B)	各不相同	$\tilde{\mathcal{O}}$(多项式$(B,\ H)/\varepsilon^2)$	§5
RFE（表格型）	S	SA	多项式(H)	$\tilde{\mathcal{O}}(S^2A$多项式$(H)/\varepsilon^2)$	§8
RFE（线性）	1	多项式(d)	变化	$\tilde{\mathcal{O}}$(多项式$(d,\ H)/\varepsilon^2)$	§8
离线线性	多项式($C_\star$)	多项式(d)	各不相同	$\tilde{\mathcal{O}}$(poly$(d, C_\star, H)/\varepsilon^2$)	第9节
离线表格	poly($C_\star$)	SA	各不相同	极小极大最优	§9

表2：**主要结果的 CSO 坐标**。每一行都将一个 PAC 界分解为覆盖（Cov）、结构（复杂度）和时域尺度。“各异”表示依赖于设置的指数，具体见所引用的章节。所有速率均省略了 log$(1/\delta)$ 和多对数因子。

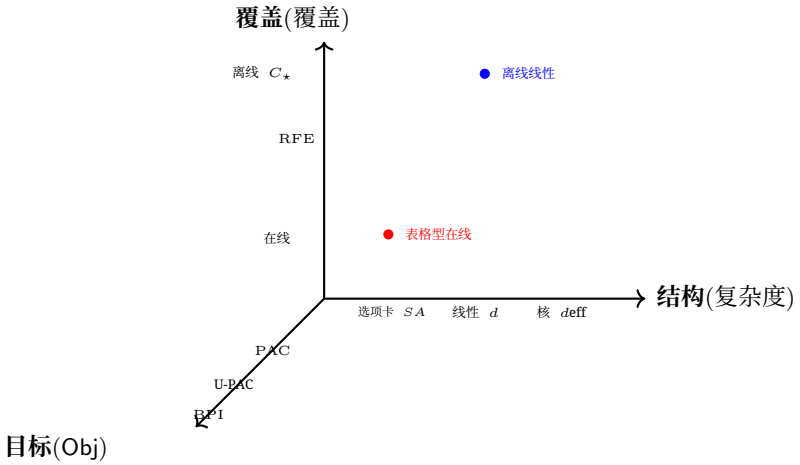

图 2：**CSO 空间**。每个 PAC 保证都对应覆盖 ×结构 ×目标空间中的一个点。沿某一轴移动会改变样本复杂度分解式 **(1)** 中的一个因素。图中标出了两个示例结果。

$1-\delta$ 的概率，在观察到数量关于 $1/\varepsilon$、$1/\delta$ 及问题相关参数呈多项式规模的示例后，其输出的假设误差至多为 ε。将这一思想迁移到序贯决策中，需要应对监督学习场景中不存在的三个挑战：智能体的动作会影响其所观察到的数据，必须将良好结果的功劳分配到一系列决策之中，并且状态分布会随着策略的变化而改变。

序贯决策的形式化框架基于马尔可夫决策过程（MDP）理论，该理论由 Bellman [3] 提出，并由 Howard [14] 和 Puterman [32]加以完善。我们采用有限时域分幕式马尔可夫决策过程作为标准模型，在算法直觉方面遵循 Sutton和Barto [38]，在分析基础方面则遵循 Szepesvári [39]。

本节确定符号表示，定义支撑整篇综述的 PAC 准则与一致 PAC 准则，介绍后续章节将具体化的结构假设与覆盖假设，并陈述用于标定可达成目标的基准结果。已熟悉分幕式马尔可夫决策过程与 PAC 定义的读者，可先略读至 §3.5 的基准结果和 §3.6 的符号总结，并在需要时返回前面的小节。

3.1 情节式马尔可夫决策过程模型

有限时域分幕式马尔可夫决策过程是一个元组 $M=(\mathcal{S},\mathcal{A},\{P_h\}_{h=1}^{H},\{r_h\}_{h=1}^{H},H,\rho)$。状态空间 $\mathcal{S}$的基数为 $|\mathcal{S}|=S$，动作空间 $\mathcal{A}$的基数为 $|\mathcal{A}|=A$。在每个阶段 $h\in\{1,\dots,H\}$，智能体观测到状态 $s_h\in\mathcal{S}$，选择动作 $a_h\in\mathcal{A}$，获得奖励 $r_h(s_h,a_h)\in[0,1]$，并状态转移至新状态 $s_{h+1}\sim P_h(\cdot\mid s_h,a_h)$。初始状态从 ρ 抽取，回合在阶段 H 之后终止。

策略 $\pi=(\pi_1,\dots,\pi_H)$ 在每个阶段将状态映射为动作上的概率分布；策略可以是非平稳的（依赖阶段），这在有限时域设定中是实现最优性所必需的。价值函数$V_h^{\pi}(s)=\mathbb{E}^{\pi}[\sum_{t=h}^{H} r_t(s_t,a_t)\mid s_h=s]$ 衡量在策略 π 下，从阶段 h 的状态 s 开始所能获得的期望累积奖励；而动作价值函数 $Q_h^{\pi}(s,a)=\mathbb{E}^{\pi}[\sum_{t=h}^{H} r_t(s_t,a_t)\mid s_h=s,a_h=a]$ 则进一步以首个动作为条件。这些函数满足 Bellman [3]提出的贝尔曼方程：$Q_h^{\pi}(s,a)=r_h(s,a)+\mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^{\pi}(s')]$ 其中 $V_h^{\pi}(s)=\mathbb{E}_{a\sim\pi_h(\cdot|s)}[Q_h^{\pi}(s,a)]$。

最优价值 $V_h^{\star}(s)=\sup_{\pi}V_h^{\pi}(s)$ 和最优动作价值 $Q_h^{\star}(s,a)=\sup_{\pi}Q_h^{\pi}(s,a)$ 由确定性最优策略 $\pi^{\star}$ 实现，而每个有限MDP [32]中都存在这样的策略。阶段 h 的贝尔曼最优算子作用于函数 $f_{h+1}:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$，定义如下：

$$(\mathcal{T}_h f)(s,a)\;=\;r_h(s,a)+\mathbb{E}_{s'\sim P_h(\cdot|s,a)}\big[\max_{a'} f_{h+1}(s',a')\big],$$

因此，对所有 h，$Q_h^{\star}=\mathcal{T}_h Q_{h+1}^{\star}$ 均成立。候选函数 f_h 在状态-动作对 (s,a) 处的贝尔曼误差，衡量其距离满足这一不动点方程有多远。控制贝尔曼误差是贯穿本综述的一条主线：在线设定中，乐观奖励项会抬高估计值以鼓励探索；离线设置中，悲观惩罚项会压低估计值以防范覆盖不足；两种情况下，样本复杂度都取决于贝尔曼误差集中于零附近的速度。

通过设置有效时域 $H_{\mathrm{eff}}\asymp(1-\gamma)^{-1}$，折扣无限时域问题可纳入这一框架，其中 γ 为折扣因子。我们对有限时域情形陈述所有结果，并在适用时说明如何转换。

3.2 PAC 与一致 PAC 保证

本综述的核心质量标准是固定置信度（PAC）保证。在监督学习场景中，PAC 学习要求经过多项式数量的样本 [44] 后，以高概率得到一个与最优假设相差不超过 ε 的假设。在强化学习中，对应的要求是在有限数量的交互回合后得到一个 ε 最优策略。

定义 1 ((ε,δ)-PAC（固定置信度控制）。 如果算法 Alg 以至少 $1-\delta$ 的概率，在至多 $N(\varepsilon,\delta)$ 个回合后输出策略 $\hat{\pi}$，且满足 $V_1^{\star}(\rho)-V_1^{\hat{\pi}}(\rho)\le\varepsilon$，则称算法 Alg 是 (ε,δ)-*PAC*。

该定义预先固定了单一的准确度目标 ε。然而，在许多应用中，起初并不知道可接受的误差容限；或者，我们希望根据 PAC 风格的结果推导遗憾界。Dann、Lattimore 和 Brunskill [4] 引入了一致 PAC 来满足这一需求，将保证强化为同时适用于所有精度水平。

定义 2 (一致 PAC)。 如果某个算法以至少 $1-\delta$ 的概率，对所有 $\varepsilon>0$ 同时成立：该算法执行次优程度为 ε 的策略的回合数至多为 $N(\varepsilon,\delta)$，则称其为一致 *PAC*算法。

这种区别很重要。单一精度水平下的 PAC算法，在其调优目标以外的其他精度水平上可能表现任意糟糕；而一致 PAC算法则必须在所有精度水平上都表现良好。这种更强的性质使其能够直接联系到累积遗憾，下面我们将陈述这一结论，并在整篇综述中引用它。

定理1 (一致 PAC蕴含高概率遗憾)。 如果某个算法是一致 *PAC*算法，且预算为$N(\varepsilon,\delta)$，那么以概率 $1-\delta$，其执行 K 个回合后的累积遗憾满足

$$\mathrm{Regret}(K)\;=\;\mathcal{O}\!\left(\int_0^{H}\min\{K,\,N(\varepsilon,\delta)\}\,d\varepsilon\right).$$

当 $N(\varepsilon,\delta)$ 对 $(S,A,H,1/\varepsilon)$ 具有多项式依赖时，即可恢复近极小极大表格遗憾率 *[4]*。

这种转换基于以下观察：精度水平 ε 所产生的遗憾贡献至多为 $\varepsilon \cdot N(\varepsilon,\ \delta)$（$N(\varepsilon,\ \delta)$ 个糟糕回合中的每一个最多造成 ε 的代价），对所有精度水平进行积分即可得到所述遗憾界。该结论是紧的：一致 PAC 严格强于单一-ε PAC，且若无额外假设 [4]，则无法改进遗憾积分。对于本综述的其余部分，任何一致 PAC 结果都可通过该定理自动导出遗憾界，从而无需额外工作即可连接固定置信度与在线学习这两种视角。

3.3 函数类、可实现性与完备性

在表格化情形之外，每个状态-动作对都可以单独访问和估计；此时，样本复杂度取决于用于逼近价值或模型的函数类的丰富程度。§5 至 §7 的结构化设定中反复出现三个概念。

定义 3（函数类与可实现性）。设 $\mathcal{F}$ 是候选价值函数或 Q-函数的一个类。可实现性断言 $Q^\star \in \mathcal{F}$（或 $V^\star \in \mathcal{F}$）：真实的最优价值函数属于该假设类。在不可知设定中，$Q^\star$ 不必属于 $\mathcal{F}$，性能则相对于 下确界 $_{f\in\mathcal{F}} \|f - Q^\star\|$ 来衡量，即该类中的最佳近似误差。

可实现性是一项强假设：它意味着建模者选对了函数类。该假设成立时，算法可以专注于统计估计（缩小一个包含 $Q^\star$ 的置信集）；若不成立，近似误差会经由贝尔曼备份不断累积，并可能压过函数逼近带来的统计收益。§14.1 讨论了如何在实践中检验可实现性，即进行误设诊断。

定义 4 (贝尔曼完备性**)**。若对任意 $\mathcal{F}$ 中的元素应用贝尔曼算子后，所得结果仍属于该函数类 $\mathcal{F}$，则称函数类 $\mathcal{F}$ 是贝尔曼完备的：$\mathcal{T}_h\mathcal{F} \subseteq \mathcal{F}$ 对所有 h 均成立。

贝尔曼完备性是一种闭包条件，可防止误差在反向归纳过程中放大。如果 $\mathcal{F}$ 包含 $Q^\star_H$（最后一个阶段），且在 $\mathcal{T}_h$ 下封闭，那么从阶段 H 开始进行反向归纳时，每个阶段的结果都仍处于 $\mathcal{F}$ 中，标准的基于回归的算法也会收敛。如果不具备完备性，函数类 $\mathcal{F}$ 中某个函数的贝尔曼备份可能落在 $\mathcal{F}$ 之外，迫使算法在每个阶段都投影回 $\mathcal{F}$，并引入沿时域累积的投影误差。这种累积是函数逼近强化学习中最常见的失败模式：算法看似收敛，却停留在次优不动点，因为不断累积的投影误差使其偏离 $Q^\star$。对于给定的函数类和马尔可夫决策过程，贝尔曼完备性很少能预先验证，因此 §14.1 中的残差诊断是实际应用中的必要手段，而非理论上的锦上添花。

定义 5（覆盖数）。覆盖数 $\mathcal{N}(\epsilon,\ \mathcal{F},\ \|\cdot\|)$ 是 $\mathcal{F}$在范数 $\|\cdot\|$下的最小 ϵ-网的大小：覆盖 $\mathcal{F}$所需的半径为 ϵ 的球的最少数量。

覆盖数通过并集界出现在 PAC 分析中：要保证置信集包含 $Q^\star$，算法必须在 $\mathcal{F}$上统一控制估计误差，而这种统一控制的代价随 $\mathcal{N}$对数增长。对于参数类（线性、低秩），覆盖数随维度和 $1/\epsilon$呈多项式增长；对于非参数类（核、神经），其增长取决于谱衰减或架构复杂度，而第 6 节中的有效维度 $d_{\text{eff}}(\lambda)$ 提供了一个更精细的替代指标。

3.4 覆盖、访问模型与占用

数据的收集方式，以及这些数据对目标策略相关状态-动作空间的覆盖程度，构成了 CSO 框架的第三个维度。其形式化对象是占用测度和集中系数。

对于策略 π 和初始分布 ρ，占用测度 $d_h^\pi(s,\ a)$ 表示从 ρ 开始遵循 π 时，在阶段 h 访问状态-动作对（$s,\ a$）的概率。在线强化学习中，智能体可以控制哪些策略生成数据，并通过探索奖励将占用引导至有信息量的区域。离线强化学习中，行为策略的占用 $\mu_h(s,\ a)$ 是固定的，智能体必须利用 μ 所提供的支撑集。

定义 6 (集中系数**)**。 集中系数衡量最优策略的访问分布偏离数据分布的程度：

$$C_\star \;=\; \max_{h\in[H]} \left\| \frac{d_h^{\pi^\star}}{\mu_h} \right\|_\infty .$$

当 $C_\star$ 较小时，数据集在最优策略访问的所有位置都具有合理的支撑集，因此离线学习是可行的。当 $C_\star$ 较大时，数据集几乎无法提供关于与 $\pi^\star$相关状态的信息，保证会退化，甚至变得毫无意义。当 $C_\star$ 为无穷大时（行为策略从不访问$\pi^\star$ 会访问的某些状态），仅凭数据不可能实现一致的策略改进。

集中系数是根据未知的 $\pi^\star$定义的，因此无法直接从数据中精确计算。实践中，替代量（密度比估计、岭杠杆分数、有效样本量）可提供近似的覆盖诊断；§14.2将介绍用于该估计的具体流程。

定义 7 （访问模型）。 本综述始终涉及三种访问模式。在在线交互中，智能体根据不断变化的策略依次收集轨迹，选择动作并实时观测状态转移。生成模型允许针对任意查询（s, a, h）从 $P_h(\cdot \mid s, a)$ 进行独立同分布采样；这是一种比在线访问更强的预言机，因为它允许进行有针对性的查询，而无需受限于某条轨迹。在离线设置中，给定一个由未知行为策略收集的固定数据集 $\mathcal{D}$ ，且不允许再与环境进行交互。

在线访问和生成模型访问使 CSO 模板中的 Cov $= 1$ 不再受限，因为智能体可以自行创建覆盖。离线访问会引入 Cov $=$poly($C_\star$)：覆盖是继承而来，而非构建出来的，其质量直接决定样本复杂度。无奖励探索（§8）处于中间位置：智能体可以在线访问，但必须构建足以应对所有可能奖励的覆盖，因此要付出额外代价，这一代价体现在覆盖项中。

在丰富观测设置（§5、§7）中，即使底层动力学由一个具有 $m \ll |\mathcal{X}|$的低维潜在状态 $z \in [m]$ 决定，观测 $x \in \mathcal{X}$ 也可能是高维的（如图像、传感器流）。观测复杂度与潜在复杂度之间的差距，正是表示学习在这些设置中既必要又可行的原因。

3.5 基准结果

符号表示和定义就绪后，我们给出用于校准后续所有结果的基准。这些是在最简单的（表格型）设置中已知的最优保证；后续章节将展示结构如何降低 (S, A) 因子，同时保留 ε^{-2} 和 poly(H) 的尺度关系。

定理 2 （表格型极小极大样本复杂度）。 在有限时域表格型 *MDP* 中，(ε, δ)-*PAC* 学习的极小极大样本复杂度为

$$N(\varepsilon, \delta) \;=\; \tilde{\Theta}\!\left(\frac{SAH^3}{\varepsilon^2} \right),$$

而经过 K 个回合后的极小极大遗憾为 $\tilde{\Theta}(\sqrt{SAH^3K})$，二者均至多多对数因子。上界由 *UCB-VI* *[2]*等乐观价值迭代变体达到；下界表明，SAH^3 这一乘积在信息论意义上是必要的 *[8, 48]*。

H^3 指数是综述中所有设置里最紧的，它来自时域的三项独立贡献：回合长度会放大回报的总方差，方差会跨阶段累积，而逐阶段并合界又贡献一个额外因子。Zhang 等人 [48] 确定了最优在线样本复杂度，证实 H^3 为正确的指数，而 Domingues 等人 [8] 给出了更精细的下界，精确刻画了每个参数的贡献。

其余基准结果（无奖励探索、线性马尔可夫决策过程、Bellman-Eluder维数、离线强化学习）已在引言（§1）中预览，并在各自章节中完整展开。此处不再复述，我们指出它们共有的模式：每一项都分解为方程（1）的乘积形式，其中结构参数取代 SA，覆盖项进入离线或无奖励设置，而ε^{-2}log($1/\delta$)作为不可约的统计核心。理解每个因子何时以及为何可以被降低，是本综述其余部分的主题。

3.6 符号总结

我们用 $\Phi_{\mathrm{tab}}(S,A,H) \equiv SAH^3$ 表示典范表格因子。符号表示 $\tilde{\mathcal{O}}(\cdot)$抑制了$(S,A,H,1/\varepsilon,1/\delta)$中的多重对数因子，而 $\Theta(\cdot)$与$\Omega(\cdot)$表示相应的紧界版本与下界版本。当我们说“多项式依赖”时，是指对所述参数的多项式，其次数有界但并不总是被显式追踪；当精确指数已知时，我们会给出它们或指向原始文献。汇总本综述所用全部符号的完整符号表见附录 A。

4 表格基准：极小极大与依赖具体实例的保证

在表格设定中，状态空间与动作空间均为有限且小到足以枚举，此处已知最尖锐的 PAC 与统一 PAC 保证。其中不涉及函数逼近：算法可以在每个阶段为每个状态-动作对维护显式估计。后续章节中的每一项结果都必须将表格型速率作为一种特例加以恢复（例如，在 $d = SA$ 处具有独热特征的线性马尔可夫决策过程），从而使这些界成为通用的校准基准。核心问题是：以置信度 $1-\delta$ 学习一个 ε-最优策略的极小极大样本复杂度是多少，以及依赖具体实例的结构何时能带来显著更快的识别?

4.1 极小极大图景

定理 2 将极小极大样本复杂度确立为 $\tilde{\Theta}(SAH^3/\varepsilon^2)$。这一结果是在一系列日益严密的分析之上逐步建立的。Strehl、Li 和 Littman [36] 早期的 PAC-MDP 界确立了有限马尔可夫决策过程中的多项式样本复杂度，随后 Azar、Osband 和 Munos [2] 引入了具有 $\sqrt{H}$ 型遗憾的 UCB-VI，Jaksch、Ortner 和 Auer [17] 则处理了平均奖励情形。现代图景随着 Domingues 等人 [8], 为分回合表格强化学习导出更精细的下界与上界，以及 Zhang 等人 [48],确定了对所有参数的最优依赖关系而得以成形。

H^3 指数值得讨论，因为它是本综述所有设置中最紧的。它源于时域的三个独立贡献：回合长度决定回报总方差的规模，方差会在各阶段累积，而对 H 个阶段使用并集界的代价还会额外增加一个因子。在结构化设定（§6）中，这些贡献会因相关估计误差而彼此交织，从而产生更高的指数（线性马尔可夫决策过程为 H^4，核模型最高可达 H^6）。

一致 PAC 框架将这些固定置信度结果直接与遗憾联系起来。根据定理1，任何预算为 $N(\varepsilon,\ \delta) = \tilde{\mathcal{O}}(SAH^3/\varepsilon^2)$ 的一致PAC算法，都能自动实现高概率遗憾 $\tilde{\mathcal{O}}(\sqrt{SAH^3K})$，且除对数因子外达到极小极大最优 [4]。这一桥梁表明，PAC 分析与遗憾分析并非相互竞争的框架，而是对同一统计挑战的互补视角。

4.2 依赖实例的识别

极小极大速率较为悲观：它们反映的是最困难的可能实例。当马尔可夫决策过程在可达状态上的最优动作有明显差距时，识别所需的成本可能会大幅降低。

定义 8（最优策略识别、间隙与可达性）。对于每个阶段 h 和状态 s，动作间隙 $\Delta_h(s,a) = Q_h^\star(s,a_h^\star(s)) - Q_h^\star(s,a)$ 衡量动作 a 的次优程度。可达性因子 $q_h(s) = \sup_\pi d_h^\pi(s)$ 反映了在阶段 h到达状态 s 的难易程度。最优策略识别（*BPI*）旨在设计算法，使其样本复杂度随这些实例依赖量而变化，而非仅随(S,A)变化。

直观上，这一点很简单。如果每个可达状态下的最优动作与备选动作之间都有较大的间隔 Δ，那么算法就不必分辨相似动作之间的细微差异；只需获取足够的数据，确认哪个动作最好即可。反之，如果某个状态在任何策略下都不可达（或几乎不可达），那么浪费样本去识别该状态下的最优动作也无济于事。实例依赖速率同时体现了这两种影响：它根据状态的可达程度以及动作与最优动作的区分难度，对每个状态-动作对的难度进行加权。

定理 3 （实例依赖BPI）。存在能够达到以下结果的乐观识别算法

$$N_{\mathrm{BPI}}(\varepsilon,\delta) \;=\; \tilde{\mathcal{O}}\Big(\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} q_h(s) \sum_{\substack{a\in\mathcal{A}:\\ \Delta_h(s,a)>\varepsilon}} \frac{1}{\Delta_h(s,a)^2}\Big),$$

以及相匹配的下界表明，对间隙倒数平方和可达性的依赖是不可避免的，至多相差对数因子。

Wagenmaker、Simchowitz 和 Jamieson [45] 确立了这些速率，揭示了在某些机制下，识别远比最坏情况学习容易。当所有可达状态都存在较大的间隙时， N_{BPI} 可小至$\tilde{\mathcal{O}}(H^2/\Delta_{\min}^2)$，而不受（$S$， A）影响，因为算法只需在每个状态确认显而易见的选择。当不同状态的间隙差异很大时，这个和会集中在少数几个动作几乎势均力敌的“困难”状态上，其余状态的贡献则微不足道。Tirinzoni、Al-Marjani 和 Kaufmann [43] 借助乐观算法改进了分析，并阐明了可达性与间隙结构如何相互作用：只有当某个策略经常访问间隙极小的状态时，该状态才会按访问频率相应地增加样本复杂度。

从 CSO 视角看，依赖实例的识别用一个可能小得多的间隙加权和，取代了最坏情况结构因子 Comp $=SA$ 。覆盖项仍为 Cov $=1$ （在线访问），目标则从一致 PAC 转向目标识别。

4.3 策略证书

策略证书为问责性提供了一个互补视角：它是依赖数据的、逐回合的次优性界。

定义 9 (策略证书**)**。第 t 回合的策略证书是一个量 U_t，根据迄今收集的数据计算得出，并且以至少 $1-\delta$ 的概率满足 $V_1^\star(\rho)-V_1^{\pi_t}(\rho)\;\le U_t$。证书无需等待算法终止，就能提供可审计保证。

证书的价值在于，它们将回顾式保证（“经过 N 个回合后，最终策略表现良好”）转变为前瞻式保证（“就在现在，第 t 回合，当前策略的次优性至多为 U_t”）。Dann 等人 [6] 表明，在表格型马尔可夫决策过程中可以高效地计算证书，而且证书的累积和能够在低阶项范围内跟踪遗憾。因此，证书自然适合作为部署门槛：运行强化学习流水线的实践者可以实时监控 U_t ，并且仅在 U_t 低于期望容差时才部署当前策略；如果证书仍然较大，则可继续收集数据。在安全敏感场景中，证书方法尤其具有吸引力，因为事后声称“算法已收敛”并不足够，审计人员要求逐决策问责。

4.4 实践考量

在表格型问题中，从业者应优先选择具有显式置信集（乐观奖励项或后验采样）的算法，以确保一致 PAC 行为，而不是采用无法量化不确定性的启发式方法。当识别是首要目标时，探索应优先关注估计可达性高且存在模糊间隙的状态；在间隙 $\Delta_h(s,a)$ 已有较高置信度表明较大的状态-动作对上花费样本，并不能改进输出策略，反而会减缓收敛。

所有表格 PAC 声明都以表格假设本身成立为前提。当 S 或 A 增长到无法显式枚举时，速率将回归函数逼近（§6）所对应的速率，而本章节的保证仅可作为校准基准：任何结构化界都必须在结构退化为表格型时恢复 $\tilde{\Theta}(SAH^3/\varepsilon^2)$。

诊断表格型方法何时不够用：一个完整示例。考虑状态为（x， $\dot{x}$， θ， $\dot{\theta}$）、动作是 {左，右}的 CartPole。由于状态是连续的， $S=\infty$ 无法应用表格型方法。可以将每个维度离散为若干区间，但得到的 S 会随状态维度呈指数增长：即使每个维度划分为10个区间，也会得到 $S=10^4$，而表格方法样本复杂度 $\tilde{\Theta}(10^4\cdot 2\cdot H^3/\varepsilon^2)$在状态维度 H适中时就会变得难以承受。更好的方法是假设采用线性特征 $\phi(s,\ a)\ =[x,\dot{x},\theta,\dot{\theta},\mathbf{1}\{a=\text{右}\}]^\top$，从而将问题简化为 $d=5$。为了验证这种线性结构是否足够，收集随机策略下的状态转移

数据，在每个阶段通过岭回归拟合$\hat{} \; Q_h = \phi^\top w_h$，并计算留出贝尔曼残差。如果残差在各阶段始终较小且稳定，则线性函数类足够，§6中的界在 $d = 5$条件下适用。如果残差随 h 增大，或呈现系统性规律（例如，在动力学呈非线性的 $\theta \approx \pm\pi/2$ 附近出现较大误差），则说明线性函数类存在模型设定错误。此时，实践者应扩充特征（例如，添加多项式项或 θ的三角函数），改用核方法（§6），或探究低秩结构（§7）。这一诊断方法的正式版本见§14.1中的算法1。

5 结构复杂度：是什么让强化学习在表格型方法之外仍然可学习?

当状态-动作空间过于庞大，无法使用表格型方法时，PAC 学习是否可行，取决于马尔可夫决策过程的结构性质，以及用于逼近价值或模型的函数类。2017年至2021年间，一系列复杂度度量相继出现，用以刻画在丰富观测下，强化学习何时具有多项式样本复杂度。这些度量（贝尔曼秩、见证秩、Bellman-Eluder维数和双线性秩）以问题相关量取代表格型参数（S，A），这些量可能小得多，因此得出的保证取决于问题的内在难度，而非观测空间的大小。

本章节追溯这些度量的发展历程，解释它们彼此之间以及与表格基准的关系，并探讨将统计可学习性与可处理算法区分开来的计算假设（预言机效率）。

5.1 从表格型到结构化：为何需要复杂度度量 s

在表格型强化学习中，样本复杂度由 SA决定，即描述转移函数和奖励函数所需的参数数量。采用函数逼近时，相关量不再是状态数量，而是价值类或模型类表示不同贝尔曼回传结构的能力。挑战在于确定恰当的容量概念：覆盖数能够反映假设空间的大小，却无法体现强化学习的序贯结构；在这种结构中，误差会沿着贝尔曼方程向后传播。

考虑一个具体例子。假设状态空间有 10^6 个状态，但每个阶段的 Q-函数都可以写成 $d = 20$ 个已知特征的线性组合。线性函数类的覆盖数随 d 增长，而非随 S 增长，因此并集界论证表明，样本复杂度与 d 成正比，而非与 S 成正比。但这种推理忽略了 Bellman 结构：即使每个阶段的回归都很简单，阶段 $h+1$ 的误差也会传递到阶段 h 的回归目标中，而控制这种误差传播所需的不只是较小的覆盖数。本章节中的复杂度度量，正是为了明确形式化控制这种误差传播所需的额外结构。

Jiang 等人 [19] 在上下文决策过程（CDP）的背景下，通过引入 Bellman 秩弥补了这一缺口。其关键洞见是：如果任意候选价值函数在任意允许分布下评估得到的期望贝尔曼误差都具有低秩分解，那么消除过程就能识别出一个 ε-最优策略，其样本复杂度关于秩和时域为多项式。

定义 10 (贝尔曼秩 [19]**).** 若对每个阶段 h，存在特征映射 $\phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^B$ 与 $\psi_h : \mathcal{F} \to \mathbb{R}^B$（范数有界），使得任意$f \in \mathcal{F}$在任意允许分布 ν_h 下的期望贝尔曼误差可因式分解为 $\mathcal{E}_h(f; \nu_h) = \langle \mathbb{E}_{(x,a)\sim\nu_h} \phi_h(x, a) \;\; \psi_h(f) \rangle$，则称一个 *CDP* 具有贝尔曼秩 B。

该因式分解意味着，贝尔曼误差位于一个 B维空间中，而非整个函数空间。消除算法可根据观测数据检验候选价值函数，并排除贝尔曼误差不一致的那些函数，从而使候选集每轮次至少缩小一个维度。经过 poly(B) 轮次后，仅近最优候选得以保留。

定理 4（贝尔曼秩下的学习）。在可实现性和贝尔曼秩 B 的条件下，消除式算法（例如 *OLIVE*）能够达到 $N(\varepsilon, \delta) = \tilde{O}\left(\text{poly}(B, H) \cdot \varepsilon^{-2} \cdot \log(1/\delta)\right)$，并使用函数类 $\mathcal{F}$ 进行评估和贪心改进 *[19]*。

5.2 基于模型的识别：见证秩

贝尔曼秩是一个基于价值的概念：它刻画了价值函数的贝尔曼误差复杂度。与之互补的基于模型视角则提出了另一个问题：错误的动力学模型能以多高的效率被证伪？如果两个候选模型对状态转移作出不同预测，就应当存在一种能够区分它们的“见证”测试，而此类测试空间的维度决定了算法缩小正确模型范围的速度。

Sun 等人 [37] 通过见证秩对其进行了形式化，该秩衡量竞争模型之间可观测差异的维度。

定义 11 (见证秩 [37]**)。** 设 $\mathcal{M}$ 为一个模型类，$\mathcal{G}$ 为一个判别函数类。见证秩W是满足以下条件的最小整数：对于每个阶段 h，任意两个模型 $M, M' \in \mathcal{M}$ 之间由任意测试函数 $g \in \mathcal{G}$ 衡量的差异，都可通过状态-动作对和模型对的 W 维嵌入进行分解。

定理 5 (见证秩下的学习**)。** 对于见证秩为 W 的可实现模型类，基于模型的乐观算法通过迭代排除那些被数据 *[37]* 证明不一致的模型，可实现 $N(\varepsilon, \delta) = \tilde{\mathcal{O}}(\mathrm{poly}(W, H) \cdot \varepsilon^{-2} \cdot \log(1/\delta))$。

一般而言，贝尔曼秩与见证秩不可比较：两者互不涵盖。在表格情形下，两者都退化为S（贝尔曼秩采用指示特征；若模型类包含所有表格型马尔可夫决策过程，则适用于见证秩），但它们衡量的是问题的不同方面。选择哪一种反映了建模决策。基于价值的方法（贝尔曼秩）无需显式的动力学模型即可进行假设消除，因此当价值函数类定义良好、但转移模型定义不充分时，这类方法很自然。基于模型的方法（见证秩）使用针对判别函数类 $\mathcal{G}$ 的矩匹配测试；当存在良好的判别函数类且模型类定义良好时，这类方法在统计上可能更高效。在实践中，当一种表示（价值或动力学）比另一种更简单时，两者的区别最为重要。

5.3 贝尔曼-埃鲁德维度：信息论综合

Jin、Liu 和 Miryoosefi [22] 通过 Bellman-Eluder（BE）维度统一并推广了前述概念。该维度衡量价值函数类相对于贝尔曼误差的序列复杂度。此定义将多臂老虎机文献中的埃鲁德维度改编到强化学习设定中，用以刻画新观测消除对贝尔曼备份的不确定性的速度。

核心思想是贝尔曼误差意义下的独立性。若一个状态-动作查询序列在序列的每个位置上都存在两个价值函数 $\mathcal{F}$，它们在所有先前查询上的取值一致（累积贝尔曼误差很小），但在当前查询上的取值却有显著差异，那么该序列就是 BE-独立的。此类序列的最长长度就是 BE 维度：它表示在算法必须识别出正确价值函数之前，环境最多能产生多少次“意外”观测。

定义 12 (贝尔曼-埃鲁德维度 [22]**)。给定价值函数类 和尺度，若查询序列 (**$x_1, a_1) \ldots, (x_n, a_n)$ 满足：对每个 t，都存在 $f, f' \in \mathcal{F}$，使它们在 (x_t, a_t) 处的贝尔曼预测相差至少 α，而在所有更早查询处的累积贝尔曼误差仍然很小，则该序列为 *BE*-独立序列。*BE* 维度 $d_{\mathrm{BE}}(\alpha)$ 是此类序列的最大长度；记作 $d_{\mathrm{BE}} = \text{上确界}_{\alpha \in (0,1]}\, d_{\mathrm{BE}}(\alpha)$。

定理6 (有限 BE 维度下的学习**)。** 如果价值函数类 $\mathcal{F}$ 具有 $d_{\mathrm{BE}} < \infty$，则 *PAC* 界和遗憾界关于(H, d_{BE}) 呈多项式尺度，且对 ε^{-2} 的依赖为 *[22]*。

BE 维度是综述中最通用的单一数值复杂度度量：只要贝尔曼秩、见证秩或双线性秩有限，它就有限；对于表格型价值函数类，它等于 SA。包含层级是严格的（存在 d_{BE} 有限、但贝尔曼秩或见证秩无限的函数类），因此它是带函数逼近时 PAC 可学习性的最宽泛充分条件。代价是精确性较低：由于 d_{BE} 涵盖了范围如此广泛的函数类，所得界对问题参数的依赖不如贝尔曼秩界或线性 MDP 界明确。

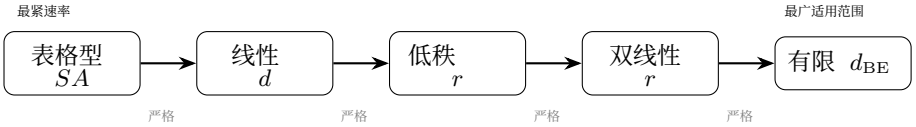

图3：**结构复杂度层级**。每个包含关系都是严格的：向右移动意味着以更紧的常数换取更广的适用范围。替代 SA 的容量参数显示在每个类别名称下方。

测度	Type	表格型价值 Keyassumption		参考文献
贝尔曼秩 B	基于价值	S	可实现性	Jiang 等人 [19]
见证秩 W	基于模型	S	可实现模型类	Sun 等人 [37]
双线性秩 r	统一	S	秩-r 因式分解	Du 等人 [10]
BE 维度 d_{BE}	信息论	SA	有限 BE 维度	Jin 等人 [22]

表 3：**结构复杂度度量比较**。 在表格情形下，它们都简化为 S 或 SA 。BE 维度最为通用： B、 W 或双线性秩有限，都能推出 d_{BE} 有限，但反之不成立。

5.4 双线性类：一种统一的因式分解

Du 等人 [10] 提出了双线性类这一结构框架，在单一代数条件下统一了多种基于秩的设定。

定义 13 （双线性类 [10]）。 **</style}若贝尔曼备份残差可通过阶段相关嵌入：** $\mathcal{X} \times \mathcal{A} \to \mathbb{R}^r$ 和$\psi_h : \mathcal{X} \to \mathbb{R}^r$ 进行秩为 r 的因式分解，且算子范数有界，则称维度为 r 的序贯决策模型是双线性的。

双线性条件是对状态-动作特征与下一状态特征之间交互施加的矩阵秩约束。当该秩较小时，算法可以使用低维内积来估计贝尔曼备份，而不必进行完整的非参数回归；此时样本复杂度随 r 变化，而非随观测空间的大小变化。

定理 7 （双线性类中的学习）。 **</style}维度为 的双线性类存在多项式时间算法，具有** （ε, δ） $= \tilde{\mathcal{O}}$（多项式（r, H） $\cdot \varepsilon^{-2} \cdot$ 对数（$1/\delta$））*</style}*保证，从而复现并统一了若干先前研究的基于特征和低秩模型的保证 *[10]*。

双线性框架将线性马尔可夫决策过程（其中 ϕ_h 是已知特征映射）、低秩马尔可夫决策过程（其中 ϕ_h 和 μ_h 提供因式分解）以及其他若干特殊情形都纳入其中。它的价值主要在于概念层面：它表明，许多表面上各不相同的可处理性结果，其背后都存在同一种代数结构。

5.5 层级及其实际含义

图3和表 3总结了这些关系。包含层级如下：

$$\text{Tabular} \ \subset \ \text{Linear MDP} \ \subset \ \text{Low-rank MDP} \ \subset \ \text{Bilinear class} \ \subset \ \text{Finite } d_{\mathrm{BE}},$$

其中每个包含关系都是严格的。在这个层级中向右移动，意味着以具体性换取一般性：表格型界具有最紧的常数和时域指数（H^3），而基于 d_{BE}的界适用于范围最广的问题类别，但其参数依赖性不够明确。

对于遇到新问题的实践者而言，这一层级提供了一套诊断流程。首先检查问题是否为表格型（有限且规模较小的 S， A）；如果是，则使用 §4 中的极小极大界。如果S 很大，但存在特征，则使用贝尔曼残差诊断（§14.1）检验线性可实现性，并应用带有 $d \ll SA$的线性马尔可夫决策过程界。如果状态转移可通过潜在状态进行因式分解，则检查是否具有低秩或块 MDP 结构（§7）。对于不具备上述任何特殊结构的一般函数类，计算或界定d_{BE} ，以获得最一般的保证。

诊断序列是单调的：每一步都会放宽假设并扩大适用范围，但代价是常数变弱。实践者应停留在最符合条件的具体层级，因为结构假设越严格，速率也越紧。如果线性残差检验通过，就没有理由采用更一般的 d_{BE} 框架；但若检验未通过，更广泛的框架则可作为后备方案。

5.6 计算方面的考量：速率无法带来什么 u

统计可学习性（ B、 W、 d_{BE} 或 r 较小）并不会自动带来多项式时间算法。信息论可行性与计算可行性之间的差距，取决于预言机效率：算法的运行时间是否关于标准监督学习子程序（在函数类 $\mathcal{F}$ 上进行代价敏感分类或回归）的开销呈多项式增长。

Dann 等人 [5] 将这一概念形式化并应用于丰富观测强化学习。贝尔曼秩消除通常在监督子问题为凸问题（例如岭回归）时具有预言机效率，但当 $\mathcal{F}$ 涉及非凸约束时，计算可能变得困难。见证秩检验可化简为针对判别函数类 $\mathcal{G}$进行矩匹配，其优化难度随 $\mathcal{G}$的丰富程度而增加。在不可知设定中，可实现性不成立，算法必须与 $\mathcal{F}$中的最佳函数竞争；此时预计会出现统计计算差距，而要实现多项式时间学习，可能需要超越有限秩或维度的额外结构假设。

其实际意义很明确。函数类为线性模型集合时，预言机就是岭回归：成本低、广为人知且为凸问题。函数类为神经网络集合时，预言机就是针对非凸损失执行随机梯度下降：在许多情况下实用，但没有最坏情形保证。函数类以隐式方式定义时（例如所有 RKHS范数有界的函数），预言机可能需要使用核方法，每次调用的成本为 $O(n^3)$，这会限制其可扩展性。读者应将本节的样本复杂度界视为信息论目标，并针对具体函数类和预言机，另行评估计算可行性。

6 函数逼近：从线性特征到神经切线核

上一章节的结构度量刻画了函数逼近何时能够实现 PAC 学习；本章节则探讨在三种逐渐一般化的可实现性假设下如何实现这一目标。我们将梳理这一演进过程：从特征已知且界最为紧确的线性马尔可夫决策过程，到以有效维度取代特征数量的再生核希尔伯特空间（RKHS），再到在神经切线核（NTK）范式下运行的过参数化神经网络，此时核理论可提供替代性保证。

共同之处在于，样本复杂度将表格因子 SA替换为维度参数（线性特征对应 d ，核函数对应 $d_{\mathrm{eff}}(\lambda)$），同时保留 ε^{-2} 核心；由于特征空间中的估计误差通常彼此相关，时域指数也会相应提高。

6.1 线性马尔可夫决策过程：最简洁的结构化设置

Jin、Yang 和 Wang [21], 提出的线性马尔可夫决策过程框架认为，奖励和转移期望均为已知特征映射的线性函数。这是最简单的非平凡函数逼近范式：特征已知，回归问题是凸的，由此得到的算法具有明确的多项式时间实现。

定义 14 (线性马尔可夫决策过程可实现性 [21])。 存在一个已知的特征映射 $\phi:\ \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ ，满足$\|\phi(s,a)\|_2\le 1$ ，使得 $r_h(s,a)\ =\phi(s,a)^\top\theta_h^{(r)}$ 且 $\mathbb{E}[g(s_{h+1})\mid s,a]=\phi(s,a)^\top\theta_{h,\ g}$ 对所有有界的 g，并满足$\|\theta\|_2\le 1$。

在这一假设下，LSVI-UCB（带上置信界的最小二乘值迭代）算法的工作方式如下：在每个回合中，它通过岭回归估计每个阶段的 Q-函数，计算椭圆置信奖励项，在特征空间探索不足的方向上抬高估计值，然后依据乐观估计采取贪心策略。椭圆加成确保算法探索特征空间中不确定性较大的方向，从而在后续回合中逐步降低整个特征空间上的估计误差。

定理8 （适用于线性马尔可夫决策过程的 LSVI-UCB）。 在线性马尔可夫决策过程可实现性成立、维度为 d 的条件下，*LSVI-UCB* 达到

$$\mathrm{Regret}(T)=\tilde{\mathcal{O}}\big(\sqrt{d^3H^3T}\big)\quad and\quad N(\varepsilon,\delta)=\tilde{\mathcal{O}}\Big(\frac{d^3H^4}{\varepsilon^2}\Big),$$

通过岭回归实现多项式时间计算 *[21]*。

H^4 指数高于表格型设置中的 H^3，其原因颇具启发性。在表格设定中，不同状态-动作对的估计误差相互独立：了解一个状态并不能让你了解另一个状态。在线性设定中，估计误差彼此相关：特征空间中某个估计不佳的方向，会影响所有特征在该方向上有投影的状态-动作对。当这些相关误差在下一阶段进入贝尔曼备份时，其累积方式不同于独立误差，会使样本复杂度额外增加一个 H 因子。方差感知估计器可以在一定程度上缓解这一影响，但在线性设定中缩小 H^4 与 H^3 之间的差距仍是一个开放问题。

He、Zhou 和 Gu [13] 将这些结果推广到了统一 PAC 准则：

定理9 (线性马尔可夫决策过程的一致 PAC)。 在线性可实现性和贝尔曼完备性条件下，存在预算为 $N(\varepsilon,\delta)$ $=\tilde{\mathcal{O}}(\mathrm{poly}(d,H)\cdot\varepsilon^{-2})$的一致 *PAC* 算法，根据定理*1*可推出高概率遗憾。

贝尔曼完备性（定义4），即 $\mathcal{T}_h\mathcal{F}_{\mathrm{lin}}\subseteq\mathcal{F}_{\mathrm{lin}}$这一要求，在此起着重要作用。若不具备该性质，线性函数的贝尔曼备份可能不再是线性的，算法必须在每个阶段将其投影回线性函数类，从而引入会沿时域累积的投影误差。完备性是否成立取决于特征映射与转移核之间的交互：对某些马尔可夫决策过程和特征而言它精确成立，对另一些则严重失效，而对大多数实际问题则介于两者之间。§14.1中的残差诊断提供了一种经验方法，用以检验完备性是否足够成立，从而使线性保证具有参考价值。

6.2 核与再生核希尔伯特空间模型：以维度换取谱 m

当价值函数光滑、但相对于已知的低维特征映射并非线性时，再生核希尔伯特空间（再生核希尔伯特空间）方法提供了一种自然的泛化。其思想是在由核函数隐式定义的潜在无限维特征空间中进行运算，并通过价值函数在该空间中的范数而非特征数量来控制复杂度。

Yang 等人 [47] 分析了基于核的强化学习，并引入了有效维度作为相关的容量度量。

定义 15 （再生核希尔伯特空间价值函数类与有效维度 [47]）。 对于一个核 κ 及其再生核希尔伯特空间 $\mathcal{H}_\kappa$，定义价值函数类 $\mathcal{F}_\kappa=\{f:\ f_h\in\mathcal{H}_\kappa,\|f_h\|_{\mathcal{H}_\kappa}\le B\}$。给定一个分布 μ，其定义在 $\mathcal{X}$ 上，并具有核积分算子$\Sigma=\mathbb{E}_{x\sim\mu}[\kappa$ $(x,\cdot)\ \otimes\kappa(x,\cdot)],$，则正则化参数 $\lambda>0$下的有效维度为 $d_{\mathrm{eff}}(\lambda)$ $=\mathrm{Tr}(\Sigma(\Sigma+\lambda I)^{-1})$。

通过其谱解释可以最好地理解有效维度。如果核积分算子具有特征值 $\sigma_1\ge\sigma_2\ge\cdots$，则 $d_{\mathrm{eff}}(\lambda)$ $=\sum_j\sigma_j/(\sigma_j+\lambda)$。如果 $\sigma_j\gg\lambda$（相应方向在数据中得到良好表示），每个特征值的贡献接近 1；如果 $\sigma_j\ll\lambda$（该方向实际上已被正则化消除），其贡献则接近 0。因此，有效维度衡量了正则化水平 λ下再生核希尔伯特空间中“活跃”方向的数量：对于有限维核，其值介于 d 之间；而对于谱衰减迅速的核，该数量可能要小得多。对于 Matérn 核， $d_{\mathrm{eff}}(\lambda)=\mathcal{O}(\lambda^{-d_x/\nu})$，其中 d_x 为输入维数， ν 为平滑度参数；对于高斯核， d_{eff} 在 $1/\lambda$中仅以多对数速度增长。

定理10 （具有有效维度的核强化学习）。 在可实现性（$V^\star\in\mathcal{F}_\kappa$）、贝尔曼完备性以及控制 $d_{\mathrm{eff}}(\lambda)$ 的谱条件下，乐观算法可达到

$$N(\varepsilon,\delta)=\tilde{\mathcal{O}}\big(\mathrm{poly}(H)\cdot d_{\mathrm{eff}}(\lambda)\cdot\varepsilon^{-2}\big)\quad and\quad \mathrm{Regret}(K)=\tilde{\mathcal{O}}\big(\mathrm{poly}(H)\cdot\sqrt{d_{\mathrm{eff}}(\lambda)\cdot K}\big).$$

时域指数范围为 H^4 至 H^6，具体取决于谱假设；最优时域指数仍是一个开放问题 *[47]*。

时域指数的广泛范围（H^4 到 H^6）反映出核强化学习的发展尚不如其线性对应方法成熟。这一差距源于处理无限维特征空间的方式不同：更严格的谱假设可以得到更低的时域指数，但适用的核函数更少；而更宽松的假设能够涵盖更多核函数，却要以更差的时域依赖性为代价。确定最优时域指数是§13中讨论的开放问题之一。

在实践中，选择正则化参数 λ 和 RKHS 范数预算 B 需要进行验证。一个有用的替代指标是岭杠杆分数 $\tau_i = x_i^\top (X^\top X + \lambda I)^{-1} x_i$，它衡量每个数据点对核回归拟合结果的影响程度。杠杆分数之和等于 $d_{\mathrm{eff}}(\lambda)$，因此，监测杠杆分数的分布可以对有效维度进行数据驱动的估计，并指导算法调优和覆盖评估（§14.2）。

6.3 通过 NTK 区域实现过参数化网络

一个自然的问题是，基于神经网络的强化学习是否继承核保证。答案是有条件地继承，是通过神经切线核（NTK）联系实现的。

当神经网络足够宽，并从随机初始化开始采用小步长梯度下降进行训练时，其预测会随参数近似线性地演化。在这一背景下，网络的行为与特定核下的核回归相一致，该特定核即神经切线核 κ_{NTK}，由 Jacot、Gabriel 和 Hongler [16] 提出，并由 Yang 等人 [47] 在强化学习背景下进行了分析。其结果是，定理10中的核强化学习保证可直接迁移，其中 $d_{\mathrm{eff}}(\lambda)$ 针对 κ_{NTK} 进行计算。

定理 11（NTK 区域保证）。 对于处于 *NTK* 区域的网络，定理*10*的保证适用，其中 $d_{\mathrm{eff}}(\lambda)$ 针对 κ_{NTK} 进行计算，同时还存在会随着宽度增加而消失的逼近误差和优化误差。

这种迁移具有重要的理论锚定价值，但在实践中的适用性有限。NTK 区域要求网络宽度远大于训练集大小、学习率较小，且不使用激进的优化技巧（动量、学习率预热、权重衰减调度）。大多数实际的深度强化学习系统都不满足这些条件。网络宽度适中或采用激进优化时，网络会离开 NTK 区域，其有效核会在训练过程中发生变化，而定理10所依赖的假设（可实现性、贝尔曼完备性、固定的有效维度）可能同时全部失效。

坦率地说，对于远离 NTK 极限运行的网络，PAC 保证基本缺失。理论指出了需要满足哪些条件（较小的 d_{eff}、近似完备性、受控的贝尔曼误差），但无法确认这些条件在特定的深度强化学习流程中是否成立。使用标准宽度网络的从业者应将模型视为可能存在模型设定错误，并依赖其他保障措施：离线场景中的悲观性、用于部署决策的离策略评估、用于分布保证的 PAC-贝叶斯界（§11），以及 §14.1 中的残差诊断，以标示保证可能失效的情况。

6.4 函数逼近下的无奖励探索

无奖励探索（§8）自然扩展到线性设定。Wang 等人 [46] 证明，在满足线性可实现性的条件下，两阶段探索只需 $\tilde{O}(\mathrm{poly}(d,\ H) \cdot \varepsilon^{-2})$ 个回合，即可保证对任意下游奖励达到 ε-最优性；Kaufmann 和 Tirinzoni [27] 则提出了自适应策略，在良性实例上改进了最坏情况预算。从 CSO 视角看，无奖励探索创造了一种可重复使用的覆盖资源：探索预算（第一阶段）是对覆盖的投资，可在任意多个奖励之间分摊。

核扩展则更为微妙。由于 $d_{\mathrm{eff}}(\lambda)$ 取决于数据分布（通过核积分算子），探索期间的有效维度可能不同于规划期间的有效维度；而两阶段保证要求探索策略产生一种分布，使得 d_{eff} 得到良好控制。探索设计与谱性质之间的这种相互作用尚未得到充分解决，并与 §13 中关于核一致 PAC 的开放问题相关。

6.5 从函数逼近到离线强化学习

线性可实现性也是第9节悲观离线强化学习保证的基础。结构参数 d与覆盖系数 $C_\star$共同影响离线样本复杂度，而特征质量与数据支持之间的相互作用决定了离线策略改进是否可行。

要理解这两个因素为何都重要，请考虑两种失效模式。第一种情况下，特征是正确的（$Q^\star$ 关于 ϕ是线性的），但数据覆盖不足（$C_\star$ 很大）：算法无法准确估计 $Q^\star$，在

对 $\pi^\star$至关重要的方向上，而悲观性会产生过于保守的策略。第二种情况下，数据具有极佳的覆盖（$C_\star$ 很小），但特征是错误的（$Q^\star$ 关于 ϕ不是线性的）：算法会自信地拟合错误的函数，而悲观性无法防范错误方向上的系统性偏差。只有当特征与覆盖都充分时，悲观离线强化学习才能兑现其保证。当特征模型设定错误时，离线保证会以额外数据无法修复的方式退化，这一主题我们将在§13的开放问题中再次讨论。

6.6 函数逼近的实践指导

本节的理论引出了若干可为实际工作提供指导的原则。

最重要的原则是在使用界之前验证假设。线性马尔可夫决策过程的保证假设奖励和状态转移对于给定特征是线性的，并且函数类是贝尔曼完备的。如果这些条件不成立，那么这些界不仅仅是不够紧，还可能产生误导，让人对表现不佳的策略产生错误的信心。§14.1 中的贝尔曼残差检验（算法1）是首先要运行的诊断：使用随机策略数据通过岭回归拟合价值函数，计算各阶段的留出贝尔曼残差，并检查残差是否始终较小且稳定。残差不断增大或出现系统性偏差，则表明存在模型设定错误。

对于核方法流程，核和正则化参数 λ 的选择决定了有效维度，进而决定样本复杂度。通过交叉验证选择 λ 并跟踪岭杠杆分数的分布，可基于数据估计 d_{eff}，从而指导算法设计和覆盖评估。如果许多数据点的杠杆分数都集中在 1 附近，那么有效维度就很大，可能需要更多数据；如果分数迅速衰减，则说明核成功压缩了问题。

在所有函数逼近设置中，保持设计矩阵条件良好对于稳定估计至关重要。LSVI-UCB 中的椭圆加成不只是理论工具，也是实践中的必要手段：它能确保算法不会反复访问特征空间中的相同方向，同时忽略其他方向。在智能体无法控制探索的离线场景中，每个阶段检查设计矩阵的条件数，可以作为衡量数据对特征空间覆盖程度的替代指标。

对于长时域问题，应监测置信半径随 H 的增长情况。如果界限随时域迅速膨胀，说明误差正在通过贝尔曼备份不断累积。可采取的补救措施包括状态抽象（缩小有效状态空间）、时间层级（将问题拆分为较短的子回合），或加大折扣（缩短有效时域 H_{eff}）。

最后，对于无法明确归入线性区域或神经切线核区域的深度强化学习流程，最稳妥的态度是保持认知上的谨慎。应将网络视为可能存在模型设定错误的函数类，在离线场景中优先采用悲观目标，在部署前使用离策略评估检验策略质量，并考虑将 PAC-贝叶斯界（§11）作为一种无需可实现性假设即可获得分布保证的方法。本章节中的保证提供的是目标和诊断依据，而非一概适用的保障。

7 丰富观测与低秩结构

许多实际的强化学习问题向智能体呈现的高维观测，远比底层动力学所需的更为丰富。机械臂接收到包含数百万像素的摄像头图像，但其任务相关状态（关节角度、夹爪位置、物体位置）所处的空间或许只有二十个连续变量。重症监护病房中的患者会产生连续的实验室检验面板、生命体征和临床记录流，但与治疗相关的状态可能归结为少数几种潜在生理状态（稳定、恶化、脓毒症、恢复）。在这两种情形中，观测空间 $\mathcal{X}$ 都极为庞大，但转移动态依赖于紧凑的潜在表示，而观测复杂度与动力学复杂度之间的差距，正是使结构化学习成为可能的原因。

本节介绍这一思想的两种形式化方式：块状马尔可夫决策过程，其中每个观测都通过可解码映射揭示一个离散潜在状态；以及低秩马尔可夫决策过程，其中转移算子通过低维特征空间进行因子分解，而不要求存在离散潜在状态。两者都将有效状态数量从 $|\mathcal{X}|$ 降至小得多的数量（m 或 r），并且都与 §5 的结构复杂度层级相关联。

7.1 块状马尔可夫决策过程：可解码的潜在状态

块状马尔可夫决策过程框架假设观测由少量潜在状态生成，并且智能体可以仅凭交互数据学会将观测解码回潜在状态。

设定如下。环境具有潜在状态空间 $[m] = \{1, \dots, m\}$，其中 $m \ll |\mathcal{X}|$。在每个阶段，潜在状态 $z_h \in [m]$ 同时决定到 z_{h+1} 的转移以及观测x_h上的分布。智能体看到 x_h 但看不到 z_h；然而，观测分布在潜在状态ˆ之间的变化足够大，使得解码器 $\psi : \mathcal{X} \to [m]$ 可以在受控误差下被学习。

定义 16（块状马尔可夫决策过程）**.** 块状马尔可夫决策过程具有满射 $\psi : \mathcal{X} \to [m]$，使得转移核 $P(x' \mid x, a)$ 仅通过 $\psi(x)$依赖于 x，并且奖励依赖于$(\psi(x)\ a)$。以不同潜在状态为条件的观测分布是可分离的：对于 $z \neq z'$，$P(x \mid z)$ 与$P(x \mid z')$ 的支撑集互不相交（或在放宽的版本中，在统计上可区分）。

Du 等人 [9] 率先研究了丰富观测下可证明高效的强化学习，并证明在满足可解码条件时，智能体可以从数据中学习解码器，随后在包含 m 个状态的潜在马尔可夫决策过程中进行规划，而不是在完整的观测空间中进行规划。Misra 等人 [30] 开发了 HOMER，这是一种表示学习算法，它将对比学习（用于发现解码器）与乐观探索（用于确保覆盖所有潜在状态）相结合，从而实现了关于（m、A、H、$1/\varepsilon$）的多项式样本复杂度。

关键洞见在于，一旦获得良好的解码器，块状马尔可夫决策过程就会化简为包含m 个状态的表格型马尔可夫决策过程，§4 中的 PAC 保证也适用，只需用 m 替换 S。学习解码器的成本被纳入对 m的多项式依赖中：表示学习会增加一个乘法因子，但不会改变基本的缩放规律。

7.2 低秩马尔可夫决策过程：连续潜在结构

块状马尔可夫决策过程要求潜在状态是离散的，且不同机制之间存在硬边界。低秩马尔可夫决策过程则将这一要求放宽为对转移算子的连续因式分解，从而适用于潜在结构平滑而非离散的设置。

定义 17 (低秩马尔可夫决策过程**)**。对于每个阶段 h，低秩马尔可夫决策过程都具有一个因式分解 $P_h(s' \mid s, a) = \langle \phi_h(s, a)\ \mu_h(s') \rangle$，其中特征映射 $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^r$ 和 $\mu_h : \mathcal{S} \to \mathbb{R}^r$有界，且 r 为秩。

这种因式分解意味着，在每个阶段，将状态-动作对映射到下一状态分布的转移算子视为矩阵（或算子）时，其秩为 r。特征 ϕ_h 和 μ_h 可能未知，必须从数据中学习；这与第 6 节的线性马尔可夫决策过程设定不同，在那里 ϕ 是已知的。

Agarwal 等人 [1] 通过 FLAMBE算法分析了利用学习到的低秩特征进行基于模型的探索，表明表示学习与乐观规划可以结合起来，使样本复杂度关于（r，H，$1/\varepsilon$）呈多项式增长。Dann 等人 [7] 研究了无关情形，即秩-r 假设仅近似成立，并指出现有算法会产生加性模型设定错误惩罚，且这种惩罚不会随着数据增多而消失。Huang 等人 [15] 提出了密度特征方法，利用来自下一状态侧（μ_h）的因式分解，为状态-动作特征方法提供了另一种选择。

定理12 (块状与低秩样本复杂度**)**。在可实现性和可识别性（块状马尔可夫决策过程的可解码性，或低秩马尔可夫决策过程的秩-r 因式分解）条件下，多项式时间算法能够实现

$$N(\varepsilon, \delta) = \tilde{\mathcal{O}}\big(\mathrm{poly}(H) \cdot \mathrm{poly}(m) \cdot A\, \varepsilon^{-2}\big) \quad \textit{(Block MDPs)}$$

$$\textit{or} \quad \tilde{\mathcal{O}}\big(\mathrm{poly}(H) \cdot \mathrm{poly}(r)\, \varepsilon^{-2}\big) \quad \textit{(low-rank MDPs)},$$

运用表示学习和乐观规划 *[9, 30, 1, 7, 15]*。

7.3 与结构复杂度和函数逼近的联系

块状马尔可夫决策过程和低秩马尔可夫决策过程均与第5节的结构复杂度层级及第6节的函数逼近结果相关。

当存在低维潜在结构时，贝尔曼秩和见证秩通常都很小。具有 m 个潜在状态的块状马尔可夫决策过程的贝尔曼秩至多为 m（使用潜在状态指示函数作为特征），并且

秩为 r 的低秩马尔可夫决策过程的双线性秩至多为 r。这解释了为什么样本复杂度按 m 或 r 而不是 $|\mathcal{X}|$ 进行缩放：即使观测空间十分庞大，结构复杂度度量仍能检测出潜在的简单性。

线性函数逼近与核函数逼近之间的关系是互补的，而非层级关系。线性马尔可夫决策过程特征（定义 14）是给定且固定的；低秩特征则必须通过学习获得。当低秩特征恰好能由已知基线性表示时，低秩设定便退化为线性马尔可夫决策过程设定，§6 中的保证可直接适用。当特征是非线性的或未知时，表示学习会带来线性理论未计入的成本，但最终的保证仍随内在维度 r 而非环境维度 $|\mathcal{X}|$缩放。

从 CSO 视角来看，在在线访问条件下，Block 和低秩马尔可夫决策过程的复杂度均为 $=\text{poly}(m)$ 或 $\text{poly}(r)$，覆盖为 $=1$。观测空间的部分可观测性（或高维性）完全由结构项吸收：它会影响所需的算法（即采用表示学习的算法），但不会影响样本复杂度的基本缩放规律。

7.4 潜在结构的实用诊断

面对高维观测空间时，应用研究人员在决定采用 Block 或低秩模型之前，应先寻找潜在结构存在的证据。

最容易实施的诊断方法是考察转移算子的经验秩。收集一个由状态转移组成的数据集（x_h、a_h、x_{h+1}），计算每个元组的核嵌入或神经网络嵌入，并检查所得矩阵的奇异值谱。如果谱快速衰减，且大部分方差由少数几个奇异值捕获，则说明存在低秩结构；如果谱较为平坦，则说明观测确实包含高维信息，块状或低秩模型可能并不适用。

如果谱证据令人鼓舞，可以进一步检验在压缩空间中，简单的函数逼近是否足够。在学习得到的嵌入之上拟合一个线性 Q-函数头，并在留出数据上评估预测准确率。如果线性头能很好地预测回报，则有效潜在维度较低，本章节或 §6 中的保证很可能适用。如果线性头效果不佳而非线性头表现良好，则说明潜在结构确实存在，但可能需要借助核方法或神经网络表示学习才能加以利用。

对于特别假设存在块 MDP 结构的设置，对比式状态解码可以提供有针对性的检验。训练一个分类器，利用时间邻接关系或奖励相似性作为监督信号，预测两个观测 x、x' 是否来自同一潜在状态。如果采用简单架构便能达到较高的分类器准确率，就能确认块结构确实存在且可解码；如果准确率较低，则说明仅凭观测可能无法辨识潜在状态，实践者应考虑更通用的低秩框架，或退而采用基于 d_{BE}的 §5 保证。

当诊断结果表明存在潜在结构时，应优先选择先学习解码器或拟合因子化模型、再执行价值迭代的算法，并使用逐回合策略证书（定义 9）作为部署的准入依据。当诊断结果不明确时，第 §5 章的 Bellman-Eluder 维度框架无需假设特定的潜在模型，也能提供保证，但代价是参数依赖性不那么明确。

8 无奖励探索

前几章反复提到，样本复杂度取决于数据收集过程与目标策略之间的匹配程度。在标准在线强化学习中，智能体在追求单一奖励信号的同时收集数据，探索也会围绕该奖励进行调整。但如果数据收集期间无法获得奖励，或者同一数据集必须支持针对多种不同奖励进行规划，又会怎样呢?

这一问题自然会出现在多个实际场景中。在多任务机器人学中，工程师可能希望训练单个机械臂执行数十种操作任务，每种任务都由不同的奖励函数定义，而无需针对每项任务重新探索环境。在安全审计中，监管机构可能需要在数据收集完成后，根据多项安全标准评估已部署系统的行为。在药物开发中，按照一个主要终点收集的临床试验数据，可能需要针对设计阶段未指定的次要终点重新分析。在所有这些情况下，数据收集成本都很高，而且在收集数据时，所关注的奖励要么未知，要么有多个。

无奖励探索（RFE）对这一问题进行了形式化。Jin 等人 [20] 提出了一种将数据收集与奖励利用分开的两阶段协议，并首次为表格情形建立了近乎最优的样本复杂度界。

8.1 两阶段协议

RFE 算法分为两个阶段。在探索阶段，智能体与环境交互，但无法获取任何奖励信号。其目标是收集足够的数据（或构建足够准确的模型），使所得信息能够支持针对之后可能指定的任意奖励函数进行近最优规划。在规划阶段，奖励函数 $r \in [0,1]$ 被揭示，智能体必须仅利用第一阶段收集的数据或构建的模型，针对该奖励输出一个 ε-最优策略，且不得再进行交互。

与标准在线探索相比，探索阶段面临的任务本质上更难。在标准强化学习中，智能体只需访问与其正在优化的特定奖励相关的状态和动作。而在 RFE 中，由于奖励尚未知晓，智能体必须访问可能与任意奖励相关的状态和动作。这意味着探索阶段必须实现对状态-动作占用测度的广泛覆盖：必须确保，在某个策略下可达的每个状态-动作对都被访问足够多次，以便准确估计其转移动力学。

这一更宽泛要求的代价，体现为样本复杂度中的一个额外因子。在表格设定中，标准的在线 PAC 学习针对单个奖励需要 $\tilde{\Theta}(SAH^3/\varepsilon^2)$ 个回合。无奖励探索则需要 $\tilde{\mathcal{O}}(S^2A\mathrm{poly}(H)/\varepsilon^2)$，多出一个 S 因子。额外的 S 源于智能体必须覆盖所有 S 个可能的“目标状态”对应的占用测度，因为未来的奖励可能会使这些状态变得重要，而不只是与某个特定奖励相关的状态。这是一种覆盖投入：探索阶段先支付 S，以便规划阶段能够免费应对任意奖励。

8.2 表格型保证

定理 13 (表格型无奖励探索)。 存在一些算法，无需访问奖励即可收集 $\tilde{\mathcal{O}}(S^2A\,\mathrm{poly}(H)\,\varepsilon^{-2})$ 个回合，随后对于任意奖励 $r \in [0,1]$，以概率 $1-\delta$ *[20]* 计算出一个 ε-最优策略。

匹配的下界 $\Omega(S^2AH^2/\varepsilon^2)$ 证实，在无奖励设置 [20]中， S^2 因子在信息论意义上是必要的。没有任何算法能够规避这一代价：如果探索阶段未能充分访问某个状态 s，那么就存在一个将全部权重放在 s 上的奖励函数，从而导致规划阶段失败。

从 CSO 视角来看，表格型 RFE 界可分解为 Cov $= S$ （支持所有奖励所需的覆盖投入）、Comp $= SA$ （表格结构）和标准的 ε^{-2} 核心项。与单奖励在线学习相比，唯一的变化在于覆盖轴：额外的 S 因子就是奖励通用性所付出的代价。

值得思考的是，何时值得付出这一代价。如果智能体只会面对一种奖励，标准在线学习的成本会低 S 倍。但如果智能体将依次面对 k 种不同的奖励，那么标准在线学习的成本为 $k \cdot \tilde{\Theta}(SAH^3/\varepsilon^2)$，而无奖励探索的成本为 $\tilde{\mathcal{O}}(S^2A\mathrm{poly}(H)/\varepsilon^2)$，无论 k如何。两者的交叉点大约为 $k \approx S$：一旦下游任务的数量超过状态数量，前期投入的 RFE 成本就能摊销，因而比反复进行单任务探索更便宜。

8.3 扩展到线性函数逼近

两阶段协议可自然扩展到函数逼近设置，此时复杂度界中的状态数量 S 由特征维度d取代。

定理 14 （线性无奖励探索）。 在线性马尔可夫决策过程可实现性和特征维度为 d 的条件下，两阶段 *RFE* 通过 $\tilde{\mathcal{O}}\left(\mathrm{poly}(H)\,\mathrm{poly}(d)\ \varepsilon^{-2}\right)$ 个样本，以概率 $1-\delta$ *[46, 27]* 保证对于任意奖励达到 ε-最优性。

Wang 等人 [46] 通过证明探索阶段能够构建准确的转移动力学线性模型（使用与 LSVI-UCB 中类似的椭圆探索奖励），且规划阶段能够通过对估计模型求解线性规划，为任意奖励提取近似最优策略，从而确立了这一结果。样本复杂度关于 d 和 H 呈多项式关系，但不依赖于 S，而后者在线性设定中可以是无限的。

Kaufmann 和 Tirinzoni [27] 通过采用自适应策略改进了最坏情况预算，这些策略会根据特征空间不同区域表现出的难度分配探索投入。他们的算法在充分了解的方向上收集较少样本，并将更多精力集中在模型仍不确定的方向上；对于较为简单的问题，其实现的依赖实例的速率可能显著优于最坏情况界。

RFE 的核扩展更为微妙。由于有效维度 $d_{\text{eff}}(\lambda)$ 取决于数据分布（通过核积分算子），探索期间的有效维度可能不同于特定奖励下规划所涉及的有效维度。探索阶段必须生成一种分布，使得对于所有可能的规划分布， d_{eff} 都能得到良好控制；这需要谨慎设计探索策略。探索设计与谱性质之间的相互作用尚未得到彻底解决，并且与 §13 中关于核一致 PAC 的开放问题相关。

8.4 与其他设置的联系

无奖励探　　索以一种颇具启发性的方式处于在线学习与离线强化学习的交汇处　　y.

在标准在线强化学习中，智能体会隐式地构建覆盖：探索奖励将其引向信息丰富的状态，而覆盖则随着奖励优化而逐渐形成。在离线强化学习中，覆盖由行为策略固定，智能体必须适应数据所提供的支撑集。RFE 处于两者之间：智能体可以在线访问（能够自由探索），但必须明确且有目的地构建覆盖，因为决定哪些状态算作“信息丰富”的奖励尚未确定。

这种定位使 RFE 成为离线强化学习流程中自然而然的预处理步骤。如果某个组织能够在确定具体奖励函数之前通过主动探索来收集数据，那么 RFE 协议就能确保最终数据集具备下游离线学习所需的覆盖属性。探索阶段带来广泛覆盖，而规划阶段（或后续的离线强化学习算法）则加以利用。在 CSO 框架中，RFE 将一个具有 $\mathrm{Cov} = \mathrm{poly}(C_\star)$（离线覆盖不确定性）的问题，转变为具有 $\mathrm{Cov} = 1$ （保证覆盖）的问题，代价是消耗探索预算。

8.5 实际考量

在应用场景中使用 RFE 时，有几个实际因素需要考虑。

最重要的问题是，下游任务的数量是否足以证明前期探索成本是值得的。如果智能体只需针对单个已知奖励进行优化，那么标准在线学习更有效率。如果智能体需要面对许多奖励，或者奖励会在部署期间修订（例如迭代式产品设计或自适应临床试验方案），那么 RFE 就能摊销探索成本，并避免重复收集数据。

在探索阶段，保持设计矩阵条件良好对于高效构建覆盖至关重要。在线性设定中，椭圆探索奖励（与 LSVI-UCB 中使用的机制相同）可确保算法不会对熟悉的方向过度探索，同时忽视不确定的方向。实际操作中，监测各阶段的设计矩阵 $X_h^\top X_h + \lambda I$ 的特征值可作为一种诊断手段：如果经过许多回合后，最小特征值仍接近 λ （正则化下限），则说明特征空间中仍有一些方向尚未探索，应继续探索阶段。

对于长时域问题，探索预算会随 $\mathrm{poly}(H)$ 增长，这可能造成难以承受的开销。管理这项成本的策略包括时间抽象（将问题拆分为较短的子回合，并分别探索）、状态抽象（在探索前缩减有效状态空间），以及分阶段探索（先进行粗略探索，再针对早期规划发现存在缺口的区域细化覆盖）。

当探索阶段结束并揭示奖励后，最终策略的质量取决于探索数据对与该特定奖励相关的状态和动作的覆盖程度。如果奖励恰好集中在状态空间中一个探索充分的小区域，规划阶段只需极少的额外计算，就能得到近最优策略。如果奖励取决于一个覆盖不足的区域，尽管算法已尽最大努力（或许是因为该区域难以到达），策略仍可能不是最优的，但仍满足定理给出的 ε 保证。监测奖励相关区域中的估计模型不确定性，可以作为诊断依据：不确定性低，说明探索预算针对这一特定奖励得到了充分利用；不确定性高，则表明增加探索预算或采用更有针对性的探索策略会更有帮助。

9 离线强化学习：从固定数据中学习

在迄今为止讨论的所有设置中，智能体都可以通过与环境交互来收集新数据。在线探索和无奖励探索都假设智能体能够控制自己的数据生成过程，通过乐观奖励项或有意构建覆盖，引导数据流向信息丰富的状态-动作空间区域。离线强化学习则完全不再作此假设。智能体会收到一个由某种行为策略 μ收集的固定轨迹数据集 $\mathcal{D}$ ，并且必须在不再与环境交互的情况下，尽可能学到最优策略。

只要无法收集更多数据，或数据收集成本高昂、存在危险，这种设置就很有现实意义。医院无法针对既往患者开展新实验来改进治疗策略；它必须从已有的医疗记录中学习。自动驾驶公司无法重新部署可能不安全的控制器来收集更好的训练数据；它必须从人工驾驶日志中提取策略。推荐系统无法反复向用户展示糟糕的推荐来探索奖励景观；它必须利用历史点击数据进行改进。

离线强化学习的核心挑战在于覆盖。数据集 $\mathcal{D}$ 反映了行为策略 μ，而该策略可能访问与最优策略 $\pi^\star$ 截然不同的状态和动作。如果 μ 从未访问过 $\pi^\star$ 会访问的某个状态，数据集中就没有关于该状态下应采取何种行动的信息，任何算法都无法在该区域可靠地改进 μ 。集中系数 $C_\star$ （定义 6）量化了这种不匹配：当 $C_\star$ 较小时，数据对最优策略的轨迹覆盖充分；当 $C_\star$ 较大时，关键区域的代表性不足；当 $C_\star$ 为无穷大时，若无进一步假设，就不可能实现改进。

9.1 悲观性作为设计原则

在线强化学习中的标准策略是乐观主义：在不确定区域抬高价值估计，以鼓励智能体访问这些区域并收集有用的数据。在离线强化学习中，智能体无法访问不确定区域，因为它不能收集新数据。在这种设置下，乐观主义适得其反：它会高估数据所知甚少的动作的价值，导致策略自信地选择支撑集薄弱的动作，并在部署时失效。

悲观性则反转了这一逻辑。悲观算法不会抬高不确定区域的估计，而是压低这些估计，即针对每个状态-动作对，减去一个与估计不确定性成正比的惩罚项。这样一来，算法会在支撑集充足的地方信任数据，而在支撑集薄弱的地方默认采用保守（低价值）估计。由此得到的策略会避开数据无法提供有效信息的区域，即使这些区域中的真实价值可能很高。

更具体地说，Jin、Yang 和 Wang 引入的悲观价值迭代（PEVI） [23], 通过减去置信奖励项，而非将其加上，对标准贝尔曼备份进行了修改：

$$\hat{Q}_h(s,a) = \hat{r}_h(s,a) + \hat{P}_h \hat{V}_{h+1}(s,a) - \Gamma_h(s,a),$$

其中，$\hat{r}_h$ 和 $\hat{P}_h$ 是根据数据对奖励和转移的估计，而 $\Gamma_h(s,\ a)$是一个惩罚项，其大小随$(s,\ a)$处估计的不确定性增加。在覆盖良好的区域（即 μ 频繁访问$(s,\ a)$的区域）， Γ_h 较小，估计值接近真实值。在覆盖较差的区域（即 μ 很少访问$(s,\ a)$的区域）， Γ_h较大，估计值会被压低，从而抑制策略依赖这些动作。

Shi 等人分析的悲观 Q 学习（PQL） [35], 在无模型设置中应用了相同原则：直接从 Q 值估计中减去惩罚项，而不构建显式的转移模型。

9.2 具有线性结构的保证

悲观离线强化学习的形式化保证结合了结构参数 d （来自函数类）和覆盖系数 $C_\star$ （来自数据分布）。

定理 15 （线性马尔可夫决策过程：悲观控制）。 在线性马尔可夫决策过程可实现性成立、特征维度为 d且集中性为 $C_\star$的条件下，悲观算法（*PEVI/PQL*）在数据集大小为以下条件时，能够达到控制误差 ε

$$n \;=\; \tilde{\mathcal{O}}\Big(\mathrm{poly}(H)\,\mathrm{poly}(d)\,\mathrm{poly}(C_\star)\,\varepsilon^{-2}\Big),$$

采用多项式时间计算 *[23, 35]*。

这个界揭示了 CSO 轴之间一个重要的相互作用。结构项 poly(d) 反映了函数类的复杂度，这与在线设定中的情况相同。覆盖项 poly($C_\star$) 则是新增的：它衡量数据分布偏离最优策略需求的程度，并且可能完全主导这个界。即使使用 $d = 5$ （一个简单的线性模型），如果行为策略与 $\pi^\star$（例如 $C_\star = 10^4$）几乎没有重叠，所需的数据集大小也会变得极其庞大。反过来，如果行为策略恰好接近最优策略（$C_\star$ 接近 1），离线学习的效率几乎与在线学习相当。

实践上的启示是，在离线强化学习中，通常真正的瓶颈是覆盖，而非模型复杂度。若实践者拥有一个由多样化行为策略收集的丰富数据集（小 $C_\star$），那么无论模型类如何，都处于有利位置。若实践者拥有一个狭窄的数据集（大 $C_\star$），那么无论算法多么复杂，都会举步维艰。

9.3 表格设置下基于模型的离线强化学习

当状态空间和动作空间均为有限时，基于模型的方法为基于价值的悲观性提供了另一种选择。其思路是根据数据估计转移模型 P_h 和奖励模型 $\hat{r}_h$，量化这些估计中的不确定性，并在估计模型中进行悲观规划，对经过估计不准确区域的状态转移施加惩罚。

定理 16 (表格离线样本复杂度**)。** 存在一些基于模型的离线算法，其达到 ε-最优控制所需的样本复杂度是极小极大最优的（相差至多对数因子），并显式依赖于(S, A, H) 及覆盖 *[29]*。

Li 等人 [29] 确定了这一设置下的极小极大样本复杂度，并表明基于模型的悲观主义不仅合理，而且是最优的：在最坏情况下，没有任何算法能做得更好。这一结果证实，在表格离线情形中，集中系数是衡量覆盖难度的恰当指标，同时也表明 CSO 模板的乘法结构（覆盖 × 结构 × 准确度）是紧的。

9.4 离策略评估

在部署离线学习策略之前，实践者可能希望先估计其价值，而不必真的在环境中运行该策略。离策略评估（OPE）正是为此而设：给定在 μ 下收集的数据集和目标策略 π，估计 $V_1^\pi(\rho)$。

离策略评估（OPE）严格来说比策略优化更容易。评估要求对价值作出准确估计；优化则要求找到能够实现高价值的策略。即使无法完成优化，评估仍有可能成功，因为估计特定策略的价值，只需沿着该策略的轨迹获得覆盖，而不必覆盖最优策略的轨迹。这一区别在实践中很重要：当覆盖不足以支持安全的策略改进时，离策略评估仍可能为候选策略提供有用信息。

两类估计量得到了广泛应用。重要性加权估计量通过目标策略与行为策略概率的比值 π/μ 对观测回报重新加权。Thomas、Theocharous 和 Ghavamzadeh [41, 42] 利用重要性加权和双重稳健校正开发了高置信度离策略评估，从而降低方差。双重稳健估计量和半参数有效估计量将基于模型的估计与重要性加权相结合，在标准正则条件下实现尽可能低的渐近方差。Kallus 和 Uehara [25] 为这一方法奠定了效率理论基础，而 Jiang 和 Huang [18]则展示了如何构造具有有限样本有效性的极小极大价值区间。

定理 17 (高效离策略评估**)。**在温和的正则条件下，存在能够实现统计效率的离策略评估估计量，并且可以构造具有有限样本有效性的极小极大价值区间 *[25, 18]*。

离线策略评估的实践价值不止于点估计。区间估计（ $V_1^\pi(\rho)$ 的置信区间）不仅能告诉实践者策略的表现可能如何，还能说明估计的不确定性有多大。若区间较窄且中心高于行为策略的价值，则表明可以安全部署；若区间较宽或涵盖零值，则说明数据不足以支持对策略质量作出有把握的判断。因此，离线策略评估自然可以充当部署门槛，补充 §14.2 中的覆盖诊断和定义 9 中的策略证书。

9.5 离线强化学习的 CSO 视角

从 CSO 视角来看，离线强化学习是覆盖轴占主导地位的设置。

在线性情形中，Comp =poly(d) 与在线学习中出现的结构项相同，而 Obj 是悲观控制或评估。新的要素是 Cov = poly($C_\star$)，它可能是界中最大的因子。在表格型基于模型的情形中，Comp = SA 和 Cov = poly($C_\star$)，且该界是极小极大最优的。

这种解析方式为离线强化学习提供了一套诊断工作流程。首先，估计覆盖度代理指标 $\widehat{C}_\star$ 使用 §14.2 中的方法（密度比、岭杠杆分数、有效样本量）。如果代理指标较小（比如，$\widehat{C}_\star < 30$），则数据具有合理的覆盖，PEVI 或 PQL 等悲观控制算法是合适的。如果代理指标处于中等水平（$30 < \widehat{C}_\star < 100$），则应谨慎行事：采用悲观性，但在部署前使用离线策略评估验证所得策略。如果代理指标较大（$\widehat{C}_\star > 100$），或密度比具有重尾，则数据不足以支撑策略改进。在这种情况下，应优先考虑离线策略评估区间估计，而不是策略优化；在收集到更多数据，或将目标策略向行为策略收缩（以最优性为代价降低 $\widehat{C}_\star$）之前，应弃权，不予部署。

9.6 两种失效模式以及何时需要担心

两种不同的失 效模式会影响离线强化学习，区分它们对于诊 is.

第一种是覆盖不足：函数类是正确的（$Q^\star \in \mathcal{F}$），但数据分布与 $\pi^\star$对齐不佳。在这种情况下，算法无法在重要方向上准确估计 $Q^\star$，而悲观性会产生过于保守的策略，从而避开高价值但支撑不足的动作。补救方法是获取更多或更好的数据：要么收集覆盖缺失区域的额外轨迹，要么修改行为策略，使其更具探索性。

第二种是模型设定错误：数据具有充分的覆盖（$C_\star$ 很小），但函数类是错误的（$Q^\star \notin \mathcal{F}$）。在这种情况下，算法会自信地拟合错误的函数，而悲观性无法防范系统性偏差，因为该偏差不会在已覆盖的方向上表现为不确定性。补救方法是使用更丰富的函数类，或采用对价值函数设定错误不太敏感的基于模型的方法。

当两种失效模式（覆盖不足和设定错误的特征）同时出现时，情况尤其棘手。总误差可分解为近似误差（源于模型设定错误）、估计误差（源于有限数据）和覆盖误差（源于分布不匹配），但这三个组成部分会以当前理论尚未完全刻画的方式相互作用。如何利用可处理算法，对这种三方权衡给出紧确界，是 §13 中指出的前沿开放问题之一。

9.7 离线流程的实践指导

本章节中的理论为离线强化学习部署提供了一套具体的工作流程。

运行任何离线算法之前，请使用 §14.2 中的诊断方法估计覆盖情况。在每个阶段计算候选策略与行为策略之间的密度比，并对极端值进行截断，以稳定估计结果。报告有效样本量和重要性权重的高分位数。如果覆盖情况充足，则进行悲观学习；否则，评估仅靠离线策略评估是否就能回答相关问题，而无需进行策略改进。

运行悲观算法后、部署之前，使用离线策略评估验证学习策略。计算该策略价值的置信区间，并与行为策略的估计价值进行比较。如果置信区间较窄，且整体高于行为策略的价值，则数据支持部署。如果置信区间较宽，或与行为策略的价值重叠，则改进效果尚不确定；实践者应收集更多数据，或继续采用行为策略。

在整个过程中，使用 §14.1 的贝尔曼残差诊断检查是否存在模型设定错误。如果残差较小，则函数类足够充分，悲观保证适用。如果残差较大或随阶段索引增加而增大，则说明函数类存在模型设定错误，实践者应扩充特征、改用基于模型的方法，或使用保守Q学习 [28] 作为一种对近似误差更具容忍性的启发式方法（但它不具备同等的形式化保证）。

算法 2 的部署门槛和定义 9 中的策略证书将这些检查形式化。应将它们用作顺序筛选条件：覆盖率门控决定是否尝试进行策略改进，证书门控则决定改进后的策略是否足够优秀，可以部署。

10 部分可观测性与混杂

前面的章节假设智能体在每一步都能观测到环境的真实状态。在许多实际问题中，这一假设并不成立。医疗决策支持系统观测到的是实验室检查结果和症状，而非潜在的疾病状态。对话智能体观测到的是用户话语，而非用户意图或满意度。通过摄像头图像导航的机器人观测到的是像素，而非自身的精确位置和朝向。当真实状态被隐藏在带噪声或不完整的观测通道之后，问题就变成了部分可观测马尔可夫决策过程（POMDP），标准的马尔可夫决策过程理论也不再能直接适用，除非加以修改。

从一般意义上说，部分可观测马尔可夫决策过程在计算和统计方面都比马尔可夫决策过程难得多。一般部分可观测马尔可夫决策过程中的规划是 PSPACE 完全的 [31], ，而在没有结构假设的情况下进行学习，所需样本复杂度会随可能的信念状态空间而增长；该空间的规模呈指数级。因此，PAC 强化学习文献一直关注具有结构的子类，在这些问题中，部分可观测性可以得到控制：尽管观测具有高维或带噪声等特点，但仍包含足够的信息来恢复（或近似恢复）紧凑的潜在状态。

本节有意写得简短。我们的综述聚焦于马尔可夫决策过程框架下的 PAC 保证，而部分可观测性是一个相邻领域，我们只在它与前文讨论的覆盖、结构和函数逼近主题直接交汇之处加以涉及。全面讨论部分可观测马尔可夫决策过程学习理论需要另行撰写一篇综述。

10.1 可识别潜在状态模型

对于 PAC 分析而言，最易处理的一类部分可观测马尔可夫决策过程，是观测能够通过解码获得潜在表示的情形。这与 §7 讨论的块状马尔可夫决策过程框架直接相关：如果观测空间$\mathcal{X}$ 很大，但动力学取决于一个规模较小的潜在状态 $z \in [m]$，并且可以从数据中学习观测到潜在状态的映射，那么智能体就可以在具有 m 个状态的潜在马尔可夫决策过程中进行规划，而不必在完整的观测空间中规划。

在能确保不同潜在状态产生可区分的观测分布的分离条件和混合条件下，HOMER 等算法 [30] 可实现关于 $(m,\ A,\ H,\ 1/\varepsilon)$ 的多项式样本复杂度。关键要求是可识别性：不同潜在状态下的观测分布必须有足够差异，才能以可控误差学习解码器。可识别性成立时，部分可观测性会增加表示学习成本（即找到解码器），但相较于潜在状态规模相同的表格型马尔可夫决策过程，并不会从根本上改变 PAC 保证的缩放规律。

从 CSO 视角来看，在在线访问下，可辨识的潜在状态部分可观测马尔可夫决策过程具有 $\mathsf{Comp} = \mathrm{poly}(m) \cdot A$（与 §7 中块状马尔可夫决策过程相同）和 $\mathsf{Cov} = 1$。部分可观测性被吸收到结构项中：m 取代了 S，而学习解码器的成本则计入对 m 的多项式依赖。

10.2 低秩可观测性

除了精确解码之外，低秩转移分解还提供了一种更灵活的潜在结构形式。如果转移算子存在秩为 r 的分解（定义 17 ），表示学习便可利用这种因式分解，而无需要求观测到离散潜在状态之间存在严格映射 $\psi : \mathcal{X} \to [m]$。

§7 中的保证（定理12）适用于此设置：样本复杂度随 $\mathrm{poly}(r,\ H)$ 增长，而非随 $|\mathcal{X}|$ 增长；算法还能从数据中学习低秩特征。相较于块状马尔可夫决策过程，低秩结构的优势在于灵活性。低秩结构能够适应连续潜在空间以及对观测进行软聚类，而块状马尔可夫决策过程则要求离散潜在状态之间具有硬边界。这种灵活性的代价是，低秩保证通常对问题参数的依赖不那么明确，而且在 r 和 H 上可能涉及更高次的多项式。

对于实践者而言，选择块状马尔可夫决策过程还是低秩表述，取决于领域知识。如果有理由相信环境只包含少数几种不同状态（例如疾病阶段、交通模式、市场状况），那么块 MDP 假设便是合适的，并能给出更紧的界限。如果

潜在结构是连续的，或者机制的数量尚不明确，那么低秩假设会更稳妥，也更具通用性。

10.3 混杂离线策略评估

当数据生成过程涉及未观测混杂因素时，离线场景中的部分可观测性会以另一种形式出现。例如，在临床数据集中，医生的治疗选择可能取决于病历中记录的患者特征（可观测），也可能取决于医生的临床直觉或私人信息（未观测）。如果未观测因素同时影响治疗和结果，标准的离线策略评估估计量就会对目标策略的价值产生有偏估计，因为它们无法区分治疗的因果效应与混杂因素的影响。

Shi、Zhou 和 Gu [34] 研究了存在混杂的部分可观测马尔可夫决策过程中的离线策略评估，并证明在结构假设（代理变量、工具变量式条件）下，即使存在混杂，也可以构造有效的价值区间。与无混杂离线策略评估所得区间相比，这些区间更宽（反映了未观测变量带来的额外不确定性），但仍能维持有限样本覆盖保证。

敏感性分析提供了一种互补的方法 [26]。实践者无需假设特定的混杂结构，而是设定一个界限，说明未观测混杂因素对治疗与结果的关系可能产生多大影响。给定这一敏感性参数，分析会得出一系列可能的策略价值，其范围宽度量化了混杂可能在多大程度上扭曲评估结果。范围较窄时，即使存在混杂，离线策略评估的估计仍值得信赖；范围较宽时，未观测变量可能是表面治疗效应的驱动因素，因此，除非收集能消除混杂的额外数据，否则实践者不应信任该评估。

10.4 部分可观测性对部署意味着什么

本节得出了两条实践经验。

首先，如果存在可辨识的潜在结构，部分可观测性并非不可逾越的障碍。如果观测空间维度很高，但能够恢复出紧凑的潜在状态（通过解码器学习或低秩表示），那么马尔可夫决策过程场景中的 PAC 保证也同样适用，只需用 m 或 r 替代 S。§7 中的诊断方法（谱分析、线性 Q 头、对比解码）在此同样适用：在决定采用部分可观测马尔可夫决策过程算法之前，先用这些方法评估潜在结构是否存在且能否恢复。

其次，在可能存在混杂因素的离线设置中，应谨慎看待策略价值的点估计。宜采用能够考虑未观测变量可能影响的区间估计，并通过敏感性分析校准区间宽度。如果区间足够窄，能够支持有把握的决策（例如，下界比行为策略的价值高出有意义的幅度），则可以考虑部署。如果区间较宽，则应暂缓策略改进，并着重收集能够消除混杂影响的数据，例如记录此前未观测到的变量，或设计采用随机化的前瞻性研究。

11 强化学习的 PAC-贝叶斯界

迄今综述的所有保证都假设学习者选定单一策略（或价值函数），并考察该策略是否接近最优。PAC-贝叶斯理论采用了不同的方法。它不对单个假设进行认证，而是对假设分布进行认证，界定从学习得到的后验分布 Q 中抽取的随机化预测器相对于固定先验分布 P的期望损失。后验分布的复杂度通过 Kullback-Leibler散度 KL($Q\|P$) 来衡量：与先验分布保持接近的后验分布只需付出较小的复杂度惩罚，而偏离较大的后验分布则必须以有力的经验性证据证明这种偏离合理。

PAC-贝叶斯界最初用于监督学习，为神经网络和其他过参数化模型提供了目前已知最紧致的泛化保证。它们的优势在于依赖数据（经验损失越低，界越紧）、无分布假设（除独立同分布采样外，不对数据生成过程作任何假设），且算法无关（任何能生成后验分布 Q 的方法都能获得保证）。将这些界应用于强化学习会带来监督学习场景中不存在的难题：数据并非独立同分布（轨迹在回合内彼此相关），损失函数

（负回报）通过状态分布取决于策略，而回报估计量的方差可能很大。

11.1 用于策略评估的通用 PAC-贝叶斯界

PAC-贝叶斯应用于强化学习最简单的情形是策略评估：给定一个由行为策略收集、包含 n 条独立同分布轨迹的数据集，对从后验分布 Q 中抽取的策略的期望回报进行界定。

设 $R(\pi)$ 是一个策略类，P 是在观测数据之前为 π 下观测到的回报进行重要性加权平均）$\hat{}$。Fard、Pineau 和 Szepesvári [11] 为此设定确立了基本的 PAC-贝叶斯不等式：以至少 $1-\delta$ 的概率，对 Π 上的所有后验分布 Q，以下不等式同时成立：

$$\mathbb{E}_{\pi\sim Q}\big[R(\pi)\big] \;\le\; \mathbb{E}_{\pi\sim Q}\big[\hat{R}(\pi)\big] \;+\; \sqrt{\frac{\mathrm{KL}(Q\|P)+\ln(1/\delta)}{2n}} \;+\; \text{bias and variance corrections.}$$

第一项是后验分布下的经验期望回报。第二项是复杂度惩罚：它随 Q 与 P 之间的散度增大而增大，并随数据增多而减小。第三项涵盖了在强化学习设定中估计回报所产生的额外误差（重要性加权方差、依赖时域的偏差、回合内相关性）。

该界对所有后验分布 Q 均一致成立，这意味着学习者可以在看到数据后优化 Q （以最小化右侧），而保证仍然成立。这是它相较于单策略 PAC 界的关键优势：学习者无需预先选定策略类或可实现性假设。任何经验损失较低且与先验分布接近的后验分布，都能获得有效的保证。

11.2 超越标准界

基本的 PAC-贝叶斯不等式假设损失有界且数据独立同分布，而在典型的强化学习设定中，这两个假设都不成立。Rivasplata 等人 [33] 扩展了这一框架，使其能够处理无界损失；当回报可能具有高方差时（例如在长时域或稀疏奖励问题中），这类损失会自然出现。他们的分析用矩条件取代了有界损失假设，因此随着回报方差增大，界会逐渐变松，而不会变得毫无意义。

Flynn 等人 [12]综述了多臂老虎机问题（强化学习的单阶段特例）中的 PAC-贝叶斯界，并通过实验比较表明，当后验分布集中在策略空间的一小块区域时，PAC-贝叶斯界可能显著紧于经典的一致收敛界。他们的研究结果表明，当策略类很大（一致界因此较松），但优良策略彼此聚集（集中后验的 KL 项代价较低）时，PAC-贝叶斯方法最有价值。

Tasdighi等人 [40] 将该框架进一步推进到具有深度探索的序贯设置中，使用PAC-贝叶斯界来正则化连续控制任务中的策略搜索。他们的方法将PAC-贝叶斯界视为目标函数：算法并非先优化期望回报再对结果进行认证，而是直接最小化真实回报的PAC-贝叶斯上界，同时联合优化后验分布与经验性能。由此得到的策略内置泛化证书，不过由此产生的优化问题比标准策略梯度方法更难。

11.3 PAC-贝叶斯如何补充本综述的其余部分

PAC-贝叶斯在PAC强化学习工具集中占据一个独特的位置，理解它的适用之处有助于明确何时使用它。

第4节至第9节中的保证是基于模型或基于价值的：它们假设一个函数类 $\mathcal{F}$，要求可实现性（$Q^\star \in \mathcal{F}$）或完备性（$\mathcal{T}_h\mathcal{F} \subseteq \mathcal{F}$），并推导出随 $\mathcal{F}$的容量而缩放的样本复杂度界。若这些假设成立，这些界是紧的，且算法是高效的。若假设不成立（函数类存在模型设定错误，完备性被违反），这些界将变得空洞，算法可能在没有警告的情况下产生糟糕的策略。

PAC-贝叶斯界以紧致性换取对模型设定错误的鲁棒性。它们不要求可实现性：即使 Π 中没有任何单一策略是最优的，该界仍能保证后验分布 Q 的期望性能，而后验分布可能是对许多表现足够好的策略进行平均得到的。它们不要求贝尔曼完备性：该界

直接针对回报进行计算，完全绕过贝尔曼方程。这种普适性的代价是：当可实现性成立时，PAC-贝叶斯界通常比基于模型的界更宽松，因为它们没有利用马尔可夫决策过程的序贯结构。

这种权衡体现了一种实用的分工。当函数类设定恰当且可验证时（例如，通过 §14.1 的残差诊断的线性特征），应采用前面章节中更紧的基于模型或基于价值的保证。当函数类不确定或可能存在模型设定错误时（例如，在复杂环境中使用神经网络策略），PAC-贝叶斯界可作为一种后备方案，因为它不依赖结构假设。在离线强化学习中，PAC-贝叶斯可以补充悲观性：先使用悲观算法学习候选策略，再在部署之前应用 PAC-贝叶斯界对该候选策略进行认证（或无法认证）。

PAC-贝叶斯界中的先验分布 P 是一种设计选择，体现了实践者的归纳偏置。强化学习中的自然选择包括：在离散策略类上采用均匀先验分布；采用以行为策略为中心的高斯先验（体现出对接近数据生成过程的策略的偏好）；或采用根据预训练模型得到的先验分布。先验分布与优质策略所在区域越匹配，KL惩罚项就越小，界也越紧。选好先验分布是将 PAC-贝叶斯应用于实践的关键技巧；即使数据充足，糟糕的先验分布（远离所有优质策略）也可能使界变得无效。

11.4 局限性与开放问题

与前面章节综述的基于模型理论相比，强化学习中的 PAC-贝叶斯理论仍不够成熟。

主要的技术缺口在于对时间依赖的处理。强化学习中的大多数 PAC-贝叶斯结果，要么假设轨迹独立同分布（这排除了在线学习和单轨迹场景），要么将时间相关性吸收到方差项中，这可能导致时域较长时界过于宽松。如何在不付出过高代价的情况下，将 PAC-贝叶斯理论扩展到能够处理回合内依赖，是一个活跃的研究方向。

第二个缺口在于计算方面。优化 PAC-贝叶斯界（找到使右侧 Q 最小的后验分布）是一个变分问题，在实践中需要进行逼近。界的质量取决于变分族对真实最优后验分布的逼近程度，而标准的变分推断可能不足以应对强化学习中出现的复杂损失景观。

尽管存在这些局限，PAC-贝叶斯仍提供了一项本综述中其他框架都不具备的独特能力：为策略分布提供依赖数据、对假设要求较少的证书。对于使用复杂函数类、且无法验证可实现性的从业者而言，这是一个宝贵的备选方案。

12 综合：CSO 视角揭示了什么

纵览不同设置下的 PAC 保证后，不妨退后一步，看看其中呈现出哪些规律。本节通过 CSO 框架梳理这一领域，指出哪些方面的理论已趋于成熟、哪些尚未成熟，并探讨将问题分解为覆盖、结构和目标后，我们能从中看出未来进展的方向。

12.1 理论已尘埃落定之处

表格型在线场景基本上已经得到解决。Zhang 等人 [48] 证实， $\tilde{\Theta}(SAH^3/\varepsilon^2)$ 速率是紧的，而统一 PAC 框架 [4] 为高概率遗憾提供了清晰的衔接。在无奖励表格型场景中，Jin 等人 [20] 确立了所有参数上的近乎最优依赖关系，其中包括覆盖所有可能下游奖励所必需的 S^2 因子。对于线性马尔可夫决策过程，已知其 PAC 界和遗憾界对 (d, H) 呈多项式依赖 [21], ，并且在贝尔曼完备性条件下存在统一 PAC 扩展 [13]。表格型基于模型场景中的离线强化学习已在极小极大层面得到解决 [29]，而具有显式 $C_\star$ 依赖关系的悲观线性马尔可夫决策过程控制也已得到充分理解 [23, 35]。

通过 CSO 来看，这些已解决的结果有一个共同点：每个轴都被独立且清晰地处理。结构项是精确已知的（或已知其在 H中的多项式次数）。覆盖项要么是平凡的（在线时为 $\mathrm{Cov} = 1$ ），要么由单个系数刻画（离线时为$C_\star$ ）。目标项采用标准 PAC 或一致 PAC。困难恰恰出现在这些清晰的分离不再成立时。

12.2 坐标轴如何相互作用

最棘手的开放问题是多个 CSO 轴相互耦合的情形，因此若不同时处理其他因素，仅改善其中一个因素是不够的。

模型设定错误下的离线强化学习将覆盖与结构联系起来。如果函数类不正确（$Q^\star \notin \mathcal{F}$），增加数据无法消除近似误差，而更好的覆盖也无法弥补系统性偏差。总误差可分解为近似误差、估计误差和覆盖度分量，但这些分量会相互作用：模型设定错误会改变哪些状态对覆盖至关重要，而覆盖不足会放大模型设定错误的影响。目前尚无对这种三方交互的精确刻画，也没有同时具备悲观性且能抵御模型误差的算法。

无先验低秩学习将结构与计算联系起来。低秩马尔可夫决策过程在近似秩-r 结构下的统计速率已有所了解 [7], ，但要通过多项式时间算法达到这些速率，还需要额外假设（精确可实现性、特定秩结构），这限制了其适用性。统计上可能实现的结果与计算上可实现的结果之间仍存在巨大差距。

依赖实例的函数逼近识别将结构与目标联系起来。在表格设定中，间隙加权速率用按可达性加权的间隙倒数平方之和 [45, 43]取代最坏情况 SA 因子。要将这些结果推广到线性或低秩模型，需要新的工具来衡量特征空间中的可达性；在这种情况下，必须用能够考虑特征几何的概念，取代“某个状态的间隙”这一概念。

核与 NTK 的统一 PAC 保证将结构与可验证性结合起来。当前结果要求满足 Bellman 完备性，这一条件涉及函数类与马尔可夫决策过程动态之间的交互。对于线性类，可以通过残差诊断进行检查（§14.1）。对于核类，目前尚无实用的验证程序，而开发此类程序有赖于理解再生核希尔伯特空间中 Bellman 算子的谱结构。

12.3 CSO 能捕捉和不能捕捉的内容

在本综述中，CSO 分解一直是很有用的组织工具，但有必要明确说明它遗漏了什么。

它无法体现计算复杂度。CSO 坐标完全相同的两种设置，在是否存在多项式时间算法这一点上可能大相径庭。双线性类与一般有限-d_{BE} 类的统计速率相近，但前者可以采用高效算法，而后者可能需要进行指数时间搜索。

它忽略了时域指数。该模板将 poly(H) 写成一个单独的因子，但 H^3 与 H^4 及H^6 的区别在实践中很重要，而精确指数往往体现了深层的结构差异（独立与相关的估计误差、逐阶段与全局置信集）。

它不适用于 PAC 类之外的目标。贝叶斯遗憾、不使用 PAC 封装的简单遗憾，以及约束优化（智能体必须在优化奖励的同时满足安全约束）都不属于 CSO 模板。

这些局限并未推翻这一框架，但限定了它的适用范围。CSO 是一种解读 2018年至2025年 PAC 强化学习文献的视角，而非强化学习的普适理论。

12.4 给实践者的启示

对于应用研究人员来说，综合起来就是一份简短的检查清单。

识别你的访问模式（在线、离线或无奖励），以确定覆盖轴。刻画你的函数类（表格型、线性、核、低秩或神经），以确定结构轴。指定你的目标（单一优良策略、一致保证、策略评估或识别），以确定目标轴。查阅表2和相关章节，找到对应的速率。在采信保证之前，先验证其依据的假设：运行贝尔曼残差检验（算法1），检查可实现性和完备性；估算覆盖代理（算法 2），评估离线数据的充分性；并依据策略证书（定义 9）决定是否允许部署。如果假设不成立或界是空泛的，CSO解析会指出问题所在的轴，并提出补救办法：为结构轴采用更好的特征，为覆盖轴收集更多样化的数据，或为目标轴放宽目标（用评估代替控制）。

13 个开放问题

§12汇总的既有成果仍留下了大片未探索的领域。本节梳理这些领域，区分哪些问题看起来可用现有技术解决，哪些则需要新的思路。其组织原则是CSO分解：最容易处理的开放问题只涉及单一轴，而最棘手的问题则涉及不同轴之间的相互作用，现有理论尚无法将其厘清。

13.1 有望解决的问题

有三个开放问题似乎可以通过扩展现有工具来解决，我们预计未来几年将取得显著进展。

第一个问题是在实践者能够验证的条件下，为核类和再生核希尔伯特空间价值函数类提供一致 PAC 保证。当前的核强化学习结果（定理10）实现了速率随有效维度 $d_{\mathrm{eff}}(\lambda)$ 缩放，但它们假设贝尔曼完备性，而这取决于函数类与马尔可夫决策过程动态的联合行为。对于线性类，可以使用 §14.1 中的残差诊断通过经验检验完备性。对于核类，目前还没有类似的检验方法。目标是设计一种算法，在在线访问条件下达到 $N = \tilde{O}(d_{\mathrm{eff}}(\lambda) \cdot \mathrm{poly}(H) \cdot \varepsilon^{-2})$，并通过基于数据的残差检验来验证（或绕过）完备性，而非将其作为假设。技术挑战在于，核函数类是无限维的，因此残差检验必须控制再生核希尔伯特空间范数下的近似误差，而不是有限参数数量下的误差。近期关于贝尔曼算子的谱方法研究表明，特征值阈值化或可提供一种实用的替代指标，但要实现具有有限样本保证的完整解决方案，仍有待研究。

第二个方向是函数逼近下依赖实例的识别。在表格设定中，Wagenmaker等人 [45] 和Tirinzoni等人 [43] 证明，间隙加权速率可以取代最坏情况SA因子，从而使样本复杂度与 $\sum_{h,s} q_h(s) \sum_a \Delta_h(s,a)^{-2}$ 成正比（定理 3）。将这一结果推广到线性或低秩模型是顺理成章的，但并非易事。难点在于，表格设定中的间隙是逐个状态定义的，而在函数逼近设定中，相关的“间隙”概念必须考虑特征几何：特征相似的两个状态-动作对，其间隙无法独立估计；有效间隙是特征空间中各个方向的属性，而非单个状态的属性。一种简洁的表述方式是根据设计矩阵的特征结构定义间隙，并推导依赖实例的速率；当特征为独热指示时，这些速率应能还原表格型界。

第三个问题是为离线强化学习部署开发可扩展的覆盖估计方法，并提供有限样本保证。§14.2 中的覆盖诊断（密度比、岭杠杆分数、有效样本量）在实践中很有用，但在函数逼近设置下，尚缺乏严格的有限样本置信区间。目标是构造一个 $\widehat{}$ $\widehat{}$估计量 $C_\star$ 及其置信区间 $[C_\star \pm \Delta]$，使算法 2 的部署门槛能够在控制第一类和第二类错误率的前提下，做出可证明正确的接受或拒绝决策。主要障碍在于，覆盖估计需要在高维空间中估计密度比，而在光滑性或结构约束下，密度比估计的极小极大速率尚未得到充分刻画。

13.2 前沿问题

三个问题似乎　　乎需要全新的技术或见解，若能在这些问题上取得进展，
将代表一次质的　　飞跃。

第一个前沿问题是不可知低秩与隐状态 PAC 学习。当前关于低秩马尔可夫决策过程的保证（定理12）假设秩-r 结构精确成立，或精确可解码。当这些假设仅近似成立时（转移算子的秩为 $r + \Delta r$，或解码器的误差为 $\varepsilon_{\mathrm{decode}}$），现有算法会产生加性误设惩罚，且不会随着数据增多而消失 [7]。目标是获得关于（r、H、$1/\varepsilon$）呈多项式关系、且随误设水平 $\varepsilon_{\mathrm{approx}}$ 平滑退化的速率，并通过 oracle 高效算法实现这一目标。困难有两个方面。在统计方面，算法必须同时估计表示（特征或解码器）与价值函数，而表示学习中的误差会传导至价值估计，难以将二者分离。在计算方面，对表示与策略进行联合优化是非凸的；即使在良性情形下，多项式时间算法能否达到相应的统计速率也尚不明确。

第二个问题是在模型设定错误与部分覆盖同时存在时的离线强化学习。如第9节所述，离线强化学习中的总误差可分解为近似误差（源于模型设定错误）、估计误差（源于有限数据）和覆盖误差（源于分布不匹配）。当前理论分别处理各个组成部分：在精确可实现的条件下，悲观算法控制估计误差和覆盖误差；在覆盖良好的条件下，无关假设界控制近似误差。目标是对三方权衡作出精确刻画，明确近似误差 $\varepsilon_{\text{approx}}$、样本量 n 和集中性 $C_\star$ 如何共同决定可达到的策略质量，同时构造计算上可行的悲观算法，使其既能抵御模型误差，又能适应覆盖质量。该问题处于研究前沿，因为模型设定错误与覆盖之间的相互作用并非简单的加性关系：模型设定错误会改变最优策略所关注的状态，进而改变所需的覆盖，形成线性分解无法刻画的循环依赖。

第三是核强化学习与神经正切核强化学习的信息论下界。核强化学习的上界随 $d_{\text{eff}}(\lambda)$缩放（定理10），但能够确认 d_{eff} 为正确复杂度度量的匹配下界基本缺失。若没有下界，则不清楚基于 d_{eff}的速率是否紧致，也不清楚是否可能得到更紧的刻画（或许涉及更精细的谱量）。目标是在适当条件下给出核强化学习的$\Omega(d_{\text{eff}}(\lambda)\cdot \mathrm{poly}(H)\cdot\varepsilon^{-2})$下界，并同时解决最优时域指数问题（目前在上界中的范围从 H^4 到 H^6 ）。困难在于，核问题的下界构造需要在再生核希尔伯特空间内构建困难的马尔可夫决策过程实例，而核所施加的光滑性约束使得嵌入标准下界技术所依赖的组合构造颇具挑战。

13.3 已知不可能之事

并非每一种理想的保证都可实现，而且区分开放问题（其中仍可能取得进展）与不可能性结果（其中困难是固有的）十分重要。

缺乏覆盖假设的离线强化学习就是这样一种障碍。如果无法对集中系数 $C_\star$ 或类似量给出界限，那么仅凭离线数据，不可能持续地改进相对于行为策略的策略。原因是根本性的：如果行为策略从未访问过最优策略会访问的状态，那么数据集中就完全没有关于该状态下该如何行动的信息；没有结构假设，任何算法都无法可靠地外推到未见状态。在这种情况下，合适的对象不是近最优策略，而是策略价值的区间估计（如离线策略评估中所用），它会如实量化覆盖缺失带来的不确定性，而不是假装这种不确定性不存在。

缺乏贝尔曼完备性的核强化学习也面临着相关障碍。在一般的上下文决策过程中，若再生核希尔伯特空间价值函数类在贝尔曼更新下不封闭，就可能出现混叠：两个价值函数在观测数据上相同，却在未见状态上存在差异，而反向归纳无法检测到这种差异。这种混叠使得在缺少额外结构（例如允许查询任意状态的生成模型）或经验证的贝尔曼残差界限时，无法给出 PAC 保证。因此，完备性条件并非仅仅是技术上的便利，而是实现本综述所讨论的这类保证所必需的条件。

缺乏计算结构的无先验函数逼近是第三道障碍。在函数类不受限制且没有可实现性假设的情况下，信息论意义上可以达到多项式样本复杂度——通过对 $\mathcal{F}$ 穷举搜索即可实现——但目前尚无已知的多项式时间算法，而且此类算法普遍而言不太可能存在。监督学习领域的计算困难性结果（例如，半空间的无先验学习的困难性）表明，类似障碍也适用于强化学习设定；要推动无先验 PAC 强化学习的发展，就必须限制函数类（如双线性或低秩设定），或接受超多项式计算。

14 实践工具包：覆盖、诊断与部署

只有当实践者能够判断本综述中的理论结果是否适用于某个具体问题，以及若适用，手头的数据是否足以支持部署时，这些结果才有用。本章节介绍了进行这些判断的具体流程。我们依次探讨三个问题：函数类是否适合该问题（模型错设诊断）、数据是否充分支撑目标策略（覆盖估计），以及学习策略是否足够出色、可以部署（部署门槛和证书）？

这里介绍的流程并非新的定理，而是将前面章节提出的理论付诸实践的操作化方法。这些流程旨在离线强化学习流水线运行前后、部署在线学习策略之前，以及部署期间作为持续监控手段使用。

14.1 诊断模型设定错误

第6节的保证假设价值函数类 $\mathcal{F}$ 包含（或能近似）真实的最优价值函数，且 $\mathcal{F}$ 在 Bellman 更新下封闭。任一假设不成立时，这些保证便不再适用，依赖这些假设的算法可能会产生自信却错误的策略。因此，在使用任何界之前检查这些假设，是严谨的强化学习流水线中的第一步。

贝尔曼残差检验提供了一种简单实用的诊断方法。其思路是使用候选函数类拟合价值函数，并检查拟合出的函数是否在留出数据上满足贝尔曼方程。如果满足，则该函数类与马尔可夫决策过程的动态相一致，这些保证很可能适用。如果不满足，则说明该函数类存在模型设定错误，实践者应当扩充特征、改用更灵活的模型类，或转而采用不要求可实现性的方法（例如第11节中的 PAC-贝叶斯界）。

算法1 BellmanResidualTest(ϕ, Alg ε)

1：使用随机（或多样化）策略收集 n 个回合。2：将数据拆分为训练集和留出集。3：对于每个阶段 $h = H$、$H-1$，……，1：通过在训练数据上进行岭回归，拟合 $f_h(s,a) = \phi(s,a)^\top w_h$，并使用$r_h + \max_{a'} f_{h+1}(s',a')$作为回归目标。4：计算留出贝尔曼残差：对每个留出的转移，计算 $\mathcal{E}_{h,i} = |f_h(s_i,a_i) - (r_{h,i} + \max_{a'} f_{h+1}(s'_i,a'))|$。5：报告每个阶段的残差均方 $\bar{\mathcal{E}}_h^2 = \frac{1}{n_{\text{held}}} \sum_i \mathcal{E}_{h,i}^2$。6：若对所有 h 都满足 $\bar{\mathcal{E}}_h^2 \le \varepsilon^2$，则通过；否则标记为模型设定错误。

解读测试输出时需要谨慎。所有阶段的残差都很小，表明该函数类近似可实现且近似满足贝尔曼完备性：拟合得到的价值函数几乎满足贝尔曼方程，因此在 $\mathcal{F}$ 内进行反向归纳时表现良好。随着 h 减小，残差增大（即后期阶段的残差较小，而早期阶段的残差较大），这表明贝尔曼完备性不成立：该函数类能够表示终端价值，却无法表示反复进行贝尔曼备份的结果；随着反向归纳的推进，误差会不断累积。残差普遍较大则表明可实现性本身不成立： $Q^\star$ 无法由 $\mathcal{F}$ 中的任何函数良好近似，因此需要更丰富的函数类。

当测试发现模型设定错误时，可根据残差的模式采取多种补救措施。如果残差仅在特定阶段或状态空间的特定区域较大，添加依赖阶段的特征或非线性基函数（多项式项、径向基函数、交互项）或许可以解决问题。如果残差普遍较大，则可能需要采用核方法（第6节）、学习到的表示（第7节）或完全不同的建模方法。如果残差在训练数据上较小、但在留出数据上较大，则函数类可能过参数化，此时适合使用正则化或更简单的模型。

14.2 估计离线数据的覆盖

函数类通过误设诊断后，离线强化学习接下来要考虑的问题是，数据集 $\mathcal{D}$ 是否为目标策略提供了充分的支撑集。第9节的理论表明，集中系数 $C_\star$ 决定了离线样本复杂度，但 $C_\star$ 取决于未知的最优策略，无法精确计算。三种互补的诊断方法可以提供近似答案。

最直接的方法是密度比估计。对于每个阶段 h，训练一个分类器，用于区分从数据集（行为策略 μ）中抽取的状态-动作对与在学习得到的模型下从候选策略 π 的滚动展开中抽取的状态-动作对。拟合后的分类器概率可得出重要性权重估计 $\hat{w}_h(x,\ a) \approx d_h^\pi(x,\ a)/\mu_h(x,\ a)$，为提高稳定性，应将其截断于最大值 $w_{\max}$。

估计值。根据截断后的权重，计算有效样本量 $\mathrm{ESS}_h = (\sum_i \hat{w}_{h,\,i})^2 / \sum_i \hat{w}^2_{h,\,i}$、权重分布的第 99 百分位数 $w_{h,\,0.99}$，以及最大权重 $\max_i \hat{w}_{h,\,i}$。有效样本量衡量重要性加权数据集相当于多少个“未加权”样本：对于大小为 10,000 的数据集，若 ESS 为 500，则意味着 95% 的数据与目标策略无关；而若 ESS 接近数据集大小，则说明行为策略与目标策略的支撑集相似。

一种互补的诊断方法是使用岭杠杆分数，它衡量的是特征空间而非状态-动作空间中的覆盖。对于线性或核价值估计，根据数据集计算逐阶段设计矩阵 X_h，并为每个数据点计算杠杆分数 $\tau_{h,\,i} = x^\top_{h,\,i}(X_h^\top X_h + \lambda I)^{-1} x_{h,\,i}$。杠杆分数之和等于有效维度 $d_{\mathrm{eff}}(\lambda)$，而各个分数的分布则揭示了哪些区域覆盖充分（杠杆分数低，方向表示充分）以及哪些区域覆盖不足（杠杆分数高，方向表示不充分）。如果许多数据点的杠杆分数接近 1，说明设计矩阵的条件较差，且这些方向上的岭回归估计不可靠。

第三种诊断方法更偏定性，但往往很有启发性，那就是直观检查状态-动作占用情况。将状态-动作对（或其嵌入）划分到网格中，并绘制各阶段的占用热图。行为策略频繁访问的区域会显示为亮区；目标策略会访问、但行为策略不会访问的区域则会显示为暗色空白。状态空间中奖励相关区域的大块暗色空白表明覆盖不足，悲观算法将会产生过于保守的策略，或者不应尝试进行策略改进。

算法 2 $\mathrm{CoverageGate}(\mathcal{D}, \pi, \lambda, w_{\max})$

对于 $h = 1 \dots H$ 执行：拟合逻辑回归分类器，以区分 $(x_h, a_h) \sim \mathcal{D}$ 与 $(x_h, a_h) \sim$ 在学习得到的模型下 π 的滚动展开；获得重要性权重估计 $\hat{w}_h$。截断：$\hat{w}_h \leftarrow \min(\hat{w}_h, w_{\max})$。计算 ESS_h、$q_{0.99}(\hat{w}_h)$、$\max \hat{w}_h$。由 X_h 和 λ 计算岭杠杆分数；汇总 $\bar{\tau}_h$、$q_{0.95}(\tau_h)$。结束循环：若 $\min_h \mathrm{ESS}_h \geq 200$ 且 $q_{0.99}(\hat{w}_h) \leq 30$ 且 $q_{0.95}(\tau_h) \leq 0.5$，则返回通过；否则警告/弃权。

算法 2 中的阈值（$\mathrm{ESS} \geq 200$、$q_{0.99} \leq 30$、$q_{0.95}(\tau) \leq 0.5$）并非严格的理论分界，而是为避免最常见失效模式而校准的实用默认值。ESS 低于 200 意味着为重要性加权估计提供信息的有效样本少于 200 个，这在中等 ε 下通常不足以进行可靠的策略评估。第 99 百分位权重高于 30 意味着少数数据点主导了重要性加权估计，使其对这些点上的噪声敏感。第 95 百分位杠杆分数高于 0.5 意味着许多数据点所贡献的信息几乎与单个无约束观测一样多，表明相应特征方向上的覆盖较弱。从业者应将这些阈值视为起点，并根据部署的利害程度加以调整：高风险应用需要更保守的阈值。

需要特别强调一个实现细节：对所有干扰模型（行为策略估计、学习动力学、密度比分类器）使用交叉拟合。在用于评估覆盖的数据上拟合干扰模型会引入乐观偏差，因为模型会对评估样本过拟合。交叉拟合（在一个折上拟合，在另一个折上评估，并对各折结果取平均）几乎不会增加计算成本，却能消除这种偏差。对于杠杆分数计算成本较高的基于核的流水线（精确核矩阵为 $O(n^3)$），Nyström近似提供了具有 k 个地标点的 $O(nk^2)$ 备选方案，以较小的近似误差换取显著的加速。

14.3 部署门槛

模型设定错误诊断和覆盖估计器共同构成一个两阶段部署门槛，用于确定是否应部署学习策略。

第一阶段是覆盖率门控（算法 2）。如果覆盖不足（门控返回“警告”或“弃权”），实践者就不应基于离线数据尝试改进策略。相反，应使用离策略评估（第9节），在不部署候选策略的情况下估计其价值，并报告

区间估计，而非点估计改进。如果可以收集更多数据，应根据杠杆分数和占用热图的识别结果，将数据收集引导至覆盖不足的区域。

第二阶段是证书门控。即使覆盖率充足，学习策略也应在部署前使用定义 9 中的逐回合策略证书进行认证。证书 U_t 提供了当前策略次优性的依赖数据的上界：只有在满足所需容差时才进行部署。如果 U_t 在覆盖率充足的情况下仍然很大，算法可能需要更多数据、更好的函数类，或不同的优化过程。证书不是一次性检查，而是持续监控：在线设定中，应跨回合跟踪 U_t，并在其超过容差时停止部署，因为这可能表明存在分布偏移或模型退化。

这两道门控各有不同用途，能够捕捉不同的失效模式。覆盖率门控会在算法运行前发现数据不足，避免在数据无法支撑改进的问题上浪费计算资源。证书门控则会在算法运行后发现算法失效，防止部署那些在训练期间表现良好、但无法泛化的策略。依次启用这两道门控，可为离线强化学习中最常见的两种部署失败提供分层保护：基于狭窄数据的过度自信改进，以及对训练分布的过拟合。

14.4 一个完整示例

为了具体说明诊断工作流程，我们来看看一个时域为 $H = 30$ 的 10×10 网格世界。随机行为策略 μ（在每个状态下对动作均匀采样）生成了 5,000 个回合的离线数据。一个离线强化学习算法（采用表格特征的悲观 Q 学习）生成了一个候选策略 π。

在此设置下运行算法 2，得到 $\min_h \mathrm{ESS}_h = 74$，$q_{0.99}(\hat{w}_h) = 115$，以及 $q_{0.95}(\tau_h) = 0.81$。这三项指标均未达到各自的阈值。有效样本量较低，说明候选策略 π 会访问随机行为策略很少到达的网格世界区域，这在意料之中：训练得到的策略会集中在高奖励路径上，而随机策略会将概率质量均匀分布，包括分布到训练策略会避开的低奖励区域。较高的权重百分位数证实，少数状态-动作对的重要性远超其他样本，使价值估计变得脆弱。杠杆分数偏高，表明某些特征方向的覆盖不足。该门控机制返回“弃权”：此数据集不足以支持对 π 进行可信部署。

随后，实践者使用带有 ε-贪心扰动的悲观策略 π 收集了另外 200 个在线回合，以探索覆盖较弱的区域附近。再次运行覆盖率门控后，得到 $\min_h \mathrm{ESS}_h = 312$，$q_{0.99}(\hat{w}_h) = 21$，以及 $q_{0.95}(\tau_h) = 0.42$。现在，三个指标均已通过。定向数据收集填补了覆盖缺口：有效样本量增加了两倍，权重尾部缩小至原来的五分之一，杠杆分数也降至阈值以下。门控返回“通过”，实践者随即进行证书检查。

计算策略证书得到 $U_t = 0.12$（以每回合最大回报为单位），低于 $\varepsilon = 0.15$ 的部署容差。两个门控均已通过，该策略获准部署。

这个示例说明了两点实践经验。初始数据集虽然包含 5,000 个回合，但仍然不足，因为覆盖质量比数据集大小更为重要。此外，在覆盖诊断的指导下进行定向补充数据收集，效率远高于再收集 5,000 个随机回合：200 个定向回合实现了数千个随机回合无法实现的效果。

15 伦理声明

本综述汇总现有理论，而非提出新算法，但其论述可能影响医疗保健、机器人学和教育等安全关键领域的实践。主要伦理风险在于超出假设范围误用保证，例如在特征设定错误时援引线性 MDP 界，或在覆盖不足时部署离线策略；这两种情况都可能损害用户。本综述始终强调的缓解措施包括：明确列出涵盖可实现性、时域尺度和覆盖代理的假设检查清单；通过逐回合策略证书把关部署；在覆盖较弱时采用弃权和区间值离策略评估；以及在报告结果时同时提供结构和覆盖诊断。若离线数据集包含历史偏差，就会产生公平性风险；覆盖分析和敏感性检查可以揭示性能不可靠的区域。我们建议采用保守默认设置（悲观性、改进前先进行离线策略评估

），公开记录假设，并在审计流程中纳入本综述归纳的失效模式。

可复现性声明

本综述涵盖 2018–2025年，并纳入发表于 JMLR、PMLR 旗下会议（COLT、ALT、ICML）和 NeurIPS 的同行评审成果；此外，对于奠定常用界限的研究，也纳入广受引用的 arXiv 预印本；2018年前的研究则用于介绍基础理论。所有结果均采用 §3 中引入的统一 MDP、POMDP 和 CDP 符号表示，并在附录 A 中汇总术语表，同时为每个参数提供首次出现时的定义（H， S， A， d， r， B， W， d_{BE}， $d_{\mathrm{eff}}(\lambda)$ $C_\star$）。

全文对经典保证均明确重述其参数依赖关系；速率表会标明省略对数因子的情况，并引用原始文献以提供精确常数和指数。文献检索使用的查询词包括 统一*PAC*强化学习、有限时域*PAC*强化学习、贝尔曼秩、*Bellman-Eluder*维数、线性马尔可夫决策过程、低秩马尔可夫决策过程 / 块状马尔可夫决策过程、无奖励探索、离线强化学习悲观性 / 集中性和*PAC*-贝叶斯强化学习，检索范围涵盖 JMLR、PMLR（COLT/ALT/ICML）、NeurIPS，以及部分后来发表于这些会议的 arXiv 预印本；我们优先选取假设清晰且具有有限样本定理的论文。

本文未引入新的数据集或代码；所有图示均为概念性内容，并在适用时引用了标准实现。对于不同论文中的速率差异（例如时域指数），我们会说明其依赖类别，并引用给出具体常数的定理；在相关情况下，也会提及不同观点。所有引用均对应截至2025年10月的版本。

局限性

本文聚焦于情景式有限时域以及固定置信度（PAC / uniform-PAC）保证；不涉及连续时间控制、平均奖励/遍历设定，也不详细探讨线性、核或神经切线核（NTK）机制之外的深度强化学习实践。若干总结省略了多重对数因子；常数和精确的时域指数留待原始文献讨论。对于部分可观测性，本文仅讨论可辨识的潜在状态／低秩子类以及混杂离线策略评估。许多结果假设可实现性或贝尔曼完备性；模型设定错误下的鲁棒性被列为开放问题。最后，尽管本文提供了诊断方法和决策规则，但并未发布经过审计、用于覆盖估计或证书计算的工具链。

参考文献组织。参考文献按主题分组：基础理论 (Valiant、Bellman、Howard、Puterman、Strehl、Szepesvári、Kaelbling、Sutton–Barto)；统一 *PAC* 与表格型 (Dann–Lattimore–Brunskill；Domingues；Zhang)；结构复杂度 (Jiang；Sun；Jin–Liu–Miryoosefi；Du–Kakade)；函数逼近 (Jin–Yang–Wang；He–Zhou–Gu；Yang 等人)；丰富观测 (Du；Misra；Agarwal；Dann–agnostic)；无奖励 (Jin；Kaufmann；Wang)；离线 (Jin–PEVI；Shi–PQL；Li–AOS；Kallus；Jiang)；*PAC*–贝叶斯(Fard；Rivasplata；Flynn；Tasdighi)。

参考文献

[1] Alekh Agarwal、Sham Kakade、Akshay Krishnamurthy 和 Wen Sun。《Flambe：低秩马尔可夫决策过程的结构复杂度与表示学习》。发表于*NeurIPS*，2020 年。[2] Mohammad Gheshlaghi Azar、Ian Osband 和 Rémi Munos。《强化学习的极小极大遗憾界》。发表于*ICML*，第 70 卷，2017 年。[3] Richard Bellman。《动态规划》。普林斯顿大学出版社，1957 年。[4] Christoph Dann、Tor Lattimore 和 Emma Brunskill。《统一 PAC 与遗憾：回合制强化学习的一致 PAC 界》。发表于*NeurIPS*，2017 年。网址：https://arxiv.org/abs/1703.07710。[5] Christoph Dann、Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal、John Langford 和 Robert E. Schapire。《关于丰富观测下的 oracle 高效 PAC 强化学习》。发表于*NeurIPS*，2018 年。

[6] Christoph Dann、Lihong Li、Wei Wei 和 Emma Brunskill。《策略证书：迈向可问责的强化学习》。发表于 ICML，第 97 卷，第 1507–1516 页，2019 年。[7] Christoph Dann、Yishay Mansour、Mehryar Mohri、Ayush Sekhari 和 Karthik Sridharan。《具有低秩马尔可夫决策过程和丰富观测的无先验强化学习》。发表于 NeurIPS，2021 年。[8] Omar Darwiche Domingues、Pierre Ménard、Emilie Kaufmann 和 Michal Valko。《有限马尔可夫决策过程中的回合制强化学习：重新审视极小极大下界》。发表于 ALT，PMLR 第 132 卷，第 578–598 页，2021 年。[9] Simon S. Du、Akshay Krishnamurthy、Nan Jiang、Alekh Agarwal、Miroslav Dudík 和 John Langford。《通过潜在状态解码，在丰富观测下实现可证明高效的强化学习》。发表于 ICML，第 97 卷，第 1665–1674 页，2019 年。[10] Simon S. Du、Sham M. Kakade、Jason D. Lee、Shachar Lovett、Gaurav Mahajan、Wen Sun 和 Ruosong Wang。《双线性类：用于强化学习中可证明泛化的结构框架》。发表于 ICML，第 139 卷，第 2826–2836 页，2021 年。[11] Mahdi Milani Fard、Joelle Pineau 和 Csaba Szepesvári。《PAC-贝叶斯策略评估》。发表于 AISTATS，第 22 卷，2012 年。[12] Hamish Flynn、David Reeb、Melih Kandemir 和 Jan Peters。《多臂老虎机问题的 PAC-贝叶斯界：综述与实验比较》。IEEE TPAMI，2023 年。[13] Jiafan He、Dongruo Zhou 和 Quanquan Gu。《线性函数逼近强化学习的一致 PAC 界》。发表于 NeurIPS，2021 年。[14] Ronald A. Howard。《动态规划与马尔可夫过程》。麻省理工学院出版社，1960 年。[15] Audrey Huang、Jinglin Chen 和 Nan Jiang。《具有密度特征的低秩马尔可夫决策过程中的强化学习》。发表于 ICML，第 202 卷，第 13710–13752 页，2023 年。[16] Arthur Jacot、Franck Gabriel 和 Clément Hongler。《神经切线核：神经网络中的收敛与泛化》。发表于 NeurIPS，2018 年。[17] Thomas Jaksch、Ronald Ortner 和 Peter Auer。《强化学习的近最优遗憾界》。JMLR，11:1563–1600，2010 年。[18] Nan Jiang 和 Jiawei Huang。《离策略评估与策略优化的极小极大价值区间》。发表于 NeurIPS，第 2747–2758 页，2020 年。[19] Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal、John Langford 和 Robert E. Schapire。《具有低贝尔曼秩的上下文决策过程是 PAC 可学习的》。发表于 ICML，第 70 卷，2017 年。[20] Chi Jin、Akshay Krishnamurthy、Max Simchowitz 和 Yu Tian。《强化学习的无奖励探索》。发表于 ICML，第 119 卷，第 4870–4879 页，2020 年。[21] Chi Jin、Zhuoran Yang、Zhaoran Wang 和 Michael I. Jordan。《通过线性函数逼近实现可证明高效的强化学习》。发表于 COLT，第 125 卷，2020 年。[22] Chi Jin、Qinghua Liu 和 Sobhan Miryoosefi。《贝尔曼-埃鲁德维数：强化学习问题的新型丰富类别及样本高效算法》。发表于 NeurIPS，第 34 卷，第 13406–13418 页，2021 年。[23] Ying Jin、Zhuoran Yang 和 Zhaoran Wang。《悲观性对于离线强化学习是否可证明高效？》。发表于 ICML，第 139 卷，2021 年。[24] Leslie Pack Kaelbling、Michael L. Littman 和 Andrew W. Moore。《强化学习：综述》。JAIR，4:237–285，1996 年。[25] Nathan Kallus 和 Masatoshi Uehara。《强化学习的统计高效离策略评估》。JMLR，21(167): 1–63，2020 年。

[26] Nathan Kallus、Xiaojie Mao 和 Masatoshi Uehara。利用负对照在未测量混杂下进行因果推断：一种极小极大学习方法。arXiv 预印本 arXiv:2103.14029，2021年。[27] Emilie Kaufmann 和 Andrea Tirinzoni。自适应无奖励探索。载于 ALT，第 132 卷，2021年。[28] Aviral Kumar、Aurick Zhou、George Tucker 和 Sergey Levine。用于离线强化学习的保守Q学习。载于 NeurIPS，2020。[29] Gen Li、Laixi Shi、Yuxin Chen、Yuejie Chi 和 Yuting Wei。确定基于模型的离线强化学习的样本复杂度。《统计年鉴》，2024。[30] Dipendra Misra、Mikael Henaff、Akshay Krishnamurthy 和 John Langford。运动学状态抽象与可证明高效的丰富观测强化学习。载于 ICML，第 119 卷，第 6961–6971 页，2020。[31] Christos H. Papadimitriou 和 John N. Tsitsiklis。马尔可夫决策过程的复杂度。《运筹研究数学》，12(3)：441–450，1987。[32] Martin L. Puterman。《马尔可夫决策过程：离散随机动态规划》。约翰·威立父子出版公司，1994。[33] Omar Rivasplata、Ilja Kuzborskij、Csaba Szepesvári 和 John Shawe-Taylor。超越常见界限的 PAC-贝叶斯分析。载于 NeurIPS，2020。[34] Jikai Shi、Dongruo Zhou 和 Quanquan Gu。存在混杂的部分可观测马尔可夫决策过程中的离策略评估。载于 ICML，第 162 卷，2022。[35] Laixi Shi、Gen Li、Yuting Wei、Yuxin Chen 和 Yuejie Chi。用于离线强化学习的悲观 Q 学习：迈向最优样本复杂度。载于 ICML，第 162 卷，第 19967–20025 页，2022。[36] Alexander L. Strehl、Lihong Li 和 Michael L. Littman。一类强化学习算法的 PAC 分析。JMLR，10：2413–2444，2009年。[37] Wen Sun、Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal 和 John Langford。上下文决策过程中的基于模型的强化学习。载于 COLT，第 99 卷，第 2898–2933 页，2019。[38] Richard S. Sutton 和 Andrew G. Barto。《强化学习：导论》（第 2 版）。麻省理工学院出版社，2018。[39] Csaba Szepesvári。《强化学习算法》。Morgan & Claypool，2010。[40] Bahareh Tasdighi、Manuel Haussmann、Nicklas Werge、Yi-Shan Wu 和 Melih Kandemir。利用 PAC-贝叶斯进行深度探索。arXiv 预印本，2024。[41] Philip S. Thomas、Georgios Theocharous 和 Mohammad Ghavamzadeh。高置信度离策略评估。载于 AAAI，2015。[42] Philip S. Thomas、Georgios Theocharous 和 Mohammad Ghavamzadeh。强化学习的数据高效离策略策略评估。载于 ICML，第 48 卷，2016。[43] Andrea Tirinzoni、Mohammed Al-Marjani 和 Emilie Kaufmann。乐观PAC强化学习：实例依赖视角。载于 ALT，第 201 卷，2023。[44] Leslie G. Valiant。可学习性理论。《ACM通讯》，27(11)：1134–1142，1984。doi：10.1145/1968.1972。[45] Andrew J. Wagenmaker、Max Simchowitz 和 Kevin Jamieson。超越无遗憾：实例依赖的 PAC 强化学习。载于 COLT，第 178 卷，2022。[46] Ruosong Wang、Simon S. Du、Lin F. Yang 和 Ruslan Salakhutdinov。关于线性函数逼近下的无奖励强化学习。载于 NeurIPS，2020。

[47] Zhuoran Yang、Chi Jin、Zhaoran Wang、Mengdi Wang 和 Michael I. Jordan。利用核与神经函数逼近实现可证明高效的强化学习。载于 *NeurIPS*，2020。

[48] Zihan Zhang、Yuxin Chen、Jason D Lee 和 Simon S Du。《确定在线强化学习的样本复杂度》。发表于 *COLT*，*PMLR* 第 247 卷，2024 年。

符号表示与缩写

本附录将所有符号表示汇总于表格中，以便快速查阅。

马尔可夫决策过程核心符号

符号	含义	首次定义
$\mathcal{S}, \mathcal{A}, \mathcal{X}$	状态、动作、观测/上下文空间	预备知识
$S = \lvert\mathcal{S}\rvert$, $A = \lvert\mathcal{A}\rvert$	基数（有限情形）	预备知识
H	时域（以时间步计的回合长度）	预备知识
γ	折扣因子； $H_{\text{eff}} \approx (1-\gamma)^{-1}$	预备知识
ρ	初始状态分布； $s_0 \sim \rho$	预备知识
P_h, r_h	阶段-h转移核和奖励	预备知识
π	策略（可能是非平稳的）	预备知识
V_h^π， Q_h^π	价值函数和动作价值函数	预备知识
$\pi^\star$， $V^\star$， $Q^\star$	最优策略和价值	预备知识

样本复杂度和置信度

符号	含义	首次定义
ε	目标次优性（准确度）	定义 1（预备知识）
δ	失败概率（置信度 $1-\delta$）	定义 1（预备知识）
$N(\varepsilon, \delta)$	（ε， δ）-PAC 的回合预算	定义 1（预备知识）
K	总回合数（遗憾时域）	预备知识
$\tilde{O}(\cdot)$ $\tilde{\Theta}(\cdot)$ $\tilde{\Omega}(\cdot)$	忽略多对数因子	分类章节
多项式(x)	x 中的多项式（次数未指定）	始终

结构复杂度参数

参数	含义	章节
d	特征维度（线性马尔可夫决策过程）	函数逼近
r	低秩分解秩	丰富观测
m	潜在状态规模（分块MDP）	丰富观测
B	贝尔曼秩	结构复杂度
W	见证秩	结构复杂度
d_{BE}	Bellman-Eluder维数	结构复杂度
$d_{\text{eff}}(\lambda)$	有效维度（再生核希尔伯特空间/神经切线核）	函数逼近
$\Phi_{\text{tab}}(S, A, H)$	表格因子 $\equiv SAH^3$	预备知识

覆盖与占用

符号	含义	首次定义
$d_h^\pi(s, a)$	π 在第 h 阶段的状态-动作占用	预备知识
$\mu_h(s, a)$	阶段 h的行为/数据占用度	预备知识
$C_\star$	集中系数	预备知识

函数类与算子

符号	含义	首次定义
$\mathcal{F}$	价值函数或 Q 函数类	预备知识
$\mathcal{M}$	模型类（动力学假设）	结构复杂度
$\mathcal{G}$	判别函数类（见证测试）	结构复杂度
$\mathcal{T}_h$	贝尔曼最优算子	预备知识
$\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$	$\mathcal{F}$的覆盖数	预备知识

访问模式

Mode	描述
在线	顺序交互；智能体选择动作并观察状态转移
生成模型	对查询（s，a，h）进行独立同分布采样，采样自 $P_h(\cdot \mid s,\ a)$
离线	固定数据集 $\mathcal{D}$；不再进行交互

缩写

缩写	含义
RL	强化学习
MDP	马尔可夫决策过程
部分可观测马尔可夫决策过程	部分可观测马尔可夫决策过程
CDP	上下文决策过程
PAC	可能近似正确
RFE	无奖励探索
OPE	离策略评估
BPI	最优策略识别
BE	贝尔曼-埃鲁德（维度）
NTK	神经切线核
RKHS	再生核希尔伯特空间
LSVI	最小二乘值迭代
UCB	上置信界
PEVI	悲观情节价值迭代
PQL	悲观 Q 学习
CSO	覆盖—结构—目标框架