arXiv:2303.08613v2 [cs.LG] 6 Aug 2023

学习激励信息获取：适当评分规则与委托代理模型的结合

Siyu Chen *[1]，Jibang Wu †[2]，Yifan Wu ‡[3]，以及 Zhuoran Yang §[1]

[1] 耶鲁大学统计与数据科学系[2] 芝加哥大学计算机科学系
[3] 西北大学计算机科学系

摘要

我们研究激励信息获取问题，即委托人聘请代理人为其收集信息。该问题被建模为委托人与代理人之间的斯塔克尔伯格博弈：委托人先公布一项规定支付方式的评分规则，随后代理人选择能使自身利润最大化的努力水平，并报告信息。我们从委托人的角度研究该问题的在线设定，即通过反复与策略性代理人交互来设计最优评分规则。我们设计了一种具有可证明样本效率的算法，将置信上界算法（Auer 等人，2002）适配到我们的模型中；该算法在 $T^{2/3}$ 次迭代后实现次线性 T-遗憾。我们的算法采用精细的估计过程来估计委托人的最优利润，并通过保守的修正方案确保代理人的预期行动得到激励。此外，我们的遗憾界具有一项关键特性：它与环境状态的数量无关。

1 引言

委托信息获取是一种十分常见的情形：一方（称为委托人）希望获取有助于决策的信息，因此聘请另一方（称为代理人）代为收集信息。设想一位投资组合经理，为了决定是否投资某公司的股票，想了解该公司的潜力。经理聘请了一位分析师，由她投入一定努力开展研究、向经理提交报告，并根据所撰写的报告获得支付。经理依据报告作出投资决策；如果股票上涨，便能获得利润。分析师在研究中投入的努力水平会影响所收集信息（即报告）的质量，以及她开展研究所产生的成本。因此，理性的分析师会选择一个能使自身利润最大化的努力水平——也就是支付与成本之差。而

*电子邮箱：siyu.chen.sc3226@yale.edu
†电子邮箱：wujibang@uchicago.edu
‡电子邮箱：yifan.wu@u.northwestern.edu§电子邮箱：zhuoran.yang@yale.edu

经理也希望使自身的期望利润最大化，即投资的预期收益减去支付。既然分析师是理性的，经理的目标就是设计一套支付规则，激励分析师投入适当的努力，从而使获取的信息能够引导投资决策，实现投资组合经理利润最大化。

从数学上讲，这类问题可以在委托代理模型（Laffont 和 Martimort，2009）下建模：委托人希望了解随机环境的状态 ω，而代理人获取的信息是状态空间 Ω 上的一个分布 $\hat{\sigma}$，也称为报告信念。委托人与代理人之间的博弈如下。开始时，状态尚未实现，双方都不知道其取值。委托人选择一种评分规则 S（Savage，1971；Gneiting 和 Raftery，2007）作为支付规则，并将其提供给代理人。代理人从 K 个努力水平中进行选择，通过采取行动 $b_k \in \{b_1, \ldots, b_K\}$ 付出成本 c_k。随后，b_k 决定状态 ω 与观测 o 的联合分布 $p(\omega, o \mid b_k)$。代理人根据这一条件分布以及 o 的实现值，向委托人报告 $\hat{\sigma}$。委托人基于所获取的信息选择行动 $a \in \mathcal{A}$。最后，状态 ω 揭晓，委托人获得效用 $u(a, \omega)$，并向代理人支付 $S(\hat{\sigma}, \omega)$。此处，代理人的支付由评分规则 S 决定；该规则通过将报告信念与实现状态进行比较，量化报告信念的价值。

更一般地，我们的模型是一个存在信息不对称的一般斯塔克尔伯格博弈（von Stackelberg，1952；Mas-Colell 等人，1995），其中代理人知道状态的分布，而委托人并不知道。委托人首先公布一条评分规则 S。随后，代理人选择一个能使自身利润最大化的努力水平 b_k，并报告一个信念 $\hat{\sigma}$。也就是说，b_k 是代理人对 S 的最佳响应。委托人和代理人的期望利润都是 S、b_k 和 $\hat{\sigma}$ 的函数。从这个角度看，设计能够最优地获取信息的评分规则，等价于求解该斯塔克尔伯格博弈的强斯塔克尔伯格均衡。

在这项工作中，我们关注此类委托代理模型的在线设定。具体而言，我们旨在回答：

> 从委托人的角度出发，如何通过与策略性代理人互动来学习最优评分规则？

在线设定带来了一些挑战。首先，由于代理人具有策略性，其报告信念 $\hat{\sigma}$ 可能不真实，甚至完全随意。其次，与多臂老虎机等其他在线学习问题类似（Lattimore 和 Szepesv´ari，2020年），我们需要探索随机环境。更重要的是，在我们的问题中，状态的分布由代理人的行动 b_k 决定，委托人无法控制这一点。因此，任何成功的学习算法都需要实施能够激励代理人探索其行动空间的评分规则。第三，委托人不知道成本 c_k 和分布 $p(\omega, o \mid b_k)$。换言之，委托人和代理人的利润函数都未知，必须根据在线数据进行估计。具体而言，这意味着代理人关于委托人评分规则的最佳响应是未知的。要找到委托人的最优评分规则，我们需要知道如何激励代理人为委托人选择最有利的 b_k，这需要学习最佳响应

函数。

我们通过提出一种新颖的算法 OSRL-UCB 来应对这些挑战。该算法利用了适当评分规则（Gneiting 和 Raftery，2007年）、不确定性下的乐观原则（Auer 等人，2002；Lattimore 和 Szepv´ari，2020年），以及针对委托人模型特有的约束优化形式。

代理人模型。具体而言，为了引出真实信息，我们证明了一个揭示原理 Myerson（1979年），表明只需关注适当评分规则这一类规则（引理 2.2）。接着，我们表明，委托人的利润最大化问题可以表示为一个 K 臂老虎机问题，其中，每个臂 $h^k,*$ 的奖励是代理人的最佳响应固定为 b_k 时委托人的最优利润（方程（3.1））。$h^k,*$ 的价值由一个约束优化问题的最优目标值决定：寻找一个适当评分规则 S^k，使委托人的利润最大化，同时满足代理人对 S^k 的最佳响应为 b_k 这一约束。此外，我们表明，$h^k,*$ 等价于一个约束优化线性规划（LP）的最优值，该线性规划涉及未知的成对成本差和由 b_k 诱导的真实信念分布（方程（3.3））。遵循乐观原则，我们旨在构造每个 $h^k,*$ 的置信上界（ UCB ），并激励代理人选择使置信上界最大的行动。为此，我们基于在线数据构造成对成本差和信念分布的置信集，并通过求解线性规划的乐观变体来获得 $h^k,*$ 的置信上界：将未知量替换为相应置信集中的元素（方程（3.7））。此外，这种乐观线性规划的最优解（即评分规则）可能违反其最佳响应为 b_k 的条件。为解决这一问题，我们设计了一种保守修正方案，以高概率同时保证所需的最佳响应和乐观性。最后，我们证明，所提出的算法在经过 T 轮交互后，实现了 $\widetilde{\mathcal{O}}\left(K^2\mathcal{C}_{\mathcal{O}} \cdot T^{2/3}\right)$ 的次线性遗憾上界。这里，K 是代理人的努力水平数量，$\mathcal{C}_{\mathcal{O}}$ 是所有可能观测的数量，$\widetilde{\ }$，而 $\mathcal{O}$ 忽略了对数项。我们的遗憾界的一个关键特点是，它与环境状态的数量无关。

1.1 相关研究

最优评分规则。 本文基于一系列关于评分规则优化的研究。若干论文考虑了具有多个努力水平的模型。 Neyman 等人(2021年) 在二元状态模型下设计了结果最优评分规则；对于整数努力水平，努力水平表示抽取的样本数量，且这些努力水平按信息量排序。 Hartline 等◯人 (2022) 在多维二元状态模型下设计了努力最优评分规则，其中努力的每个维度对应于相互独立的多维状态之一。相比之下，本文考虑的是一般状态空间，包含多个不一定有序或相互独立的努力水平。 Li 等人 (2022)；Papireddygari 和 Waggoner(2022)；Chen 和 Yu (2021) 为二元努力模型设计了最优评分规则，这与本文的多努力水平模型不同。 Oesterheld 和 Conitzer (2020年) 针对单轮次中的多个代理人设计了遗憾最优评分规则，其中委托人不知道信息结构；而我们的模型只有一名代理人，且我们的学习算法在多轮次中实现遗憾递减。此外，以上

上述论文均将状态建模为外生给定，而在本文中，状态的先验分布可能会受到代理人内生行动的影响。

委托代理问题。我们关于信息获取的模型可以视为一类委托代理问题，它已发展成为经济学中一个重要分支，即契约理论（Grossman 和 Hart，1992年；Smith，2004年；Laffont 和 Martimort，2009年）。随着基于互联网的应用中基于契约的市场日益兴起，委托代理问题近期受到越来越多的关注，尤其是来自计算机科学界的关注（D¨utting 等，2019年；Dutting 等，2021；Guruganesh 等，2021；Alon 等，2021；Castiglioni 等，2021，2022）。正如 Alon 等人（2021）所指出的，这些市场包括众包、赞助内容创作、联盟营销、自由职业等在线市场。这些市场具有可观的经济价值，而数据和计算发挥着关键作用。不同于经典的合同设计问题，我们关注如何设计合同（即评分规则），以最优地激励代理人获取信息，同时由此产生一定成本。

策略环境中的在线学习。更广泛地说，我们的工作补充了策略环境下在线学习的相关研究，这一领域近年来日益受到关注。具体而言，在某种重复博弈中，每轮在线学习者的效用由其自身的行动以及其他参与者的策略性响应共同决定，而学习者的典型目标是在某种均衡下找到自己的最优策略。这些重复博弈取材于有影响力的经济模型，包括但不限于 Stackelberg（安全）博弈（Marecki et al., 2012；Balcan et al., 2015；Haghtalab et al., 2022）、拍卖设计（Amin et al., 2013；Feng et al., 2018；Golrezaei et al., 2019；Guo et al., 2022）、匹配（Jagadeesan et al., 2021）、合同设计（Zhu 等人（2022））以及贝叶斯劝说（Castiglioni et al., 2020；Zu et al., 2021；Wu et al., 2022）。我们的模型可视为在线学习框架下的一种广义信息引出问题。据我们所知，尚无先前研究考虑过类似的学习问题。此外，我们指出，在许多现有研究中，例如 Balcan et al. (2015)、Guo et al. (2022) 和 Wu et al. (2022)，研究假设学习者对其他策略性参与者有充分了解，而其不确定性涉及环境或自身效用。此类假设可以显著简化问题，使其成为标准的在线学习问题，因为学习者几乎能够预测其他参与者的最佳响应。然而，我们的工作不作此类假设；事实上，学习算法设计中最具挑战性的部分，正是在不确定性下确保代理人作出期望的响应。Camara et al. (2020) 研究了在线机制设计问题，其中委托人和代理人都可以随着时间推移，从状态历史中学习。然而，我们的论文假设代理人在每一轮都是短视的。

2 信息获取模型

在本节中，我们介绍委托代理框架下如何最优地获取信息这一问题，并将其形式化为斯塔克尔伯格博弈。遵循揭示原理（Myerson,

1979年），我们通过将问题限定在适当评分规则这一类别中来简化问题，从而实现高效的在线学习。

2.1 信息获取基础

为在委托代理框架下对最优信息获取问题进行形式化描述，我们考虑一个由委托人和全能代理人组成的随机环境。在第 t 轮，存在一个隐藏状态 $\omega_t \in \Omega$，它会影响委托人的效用，但在本轮结束之前，代理人和委托人都不知道该状态。为了获取更精确的信息，即代理人对隐藏状态的信念，委托人先行动，向代理人提供一项评分规则；代理人根据其报告信念的质量获得相应支付。代理人可以从其有限行动空间 $\mathcal{B}$中选择一项行动并承担相应成本，获取与隐藏状态相关的观测，并报告自己的信念，从而使委托人能够更好地做出决策。最后，隐藏状态揭晓，委托人获得效用，并根据评分规则向代理人支付报酬。对于任意 $t \geq 0$，在第 t 轮，委托人与代理人之间的互动如下。

通过评分规则获取信息

在第 t 轮，委托人与代理人按以下方式进行博弈：

1. 委托人承诺采用评分规则 S_t：$\Delta(\Omega) \times \Omega \to \mathbb{R}_+$，其中 $\Delta(\Omega)$ 是 Ω 上的分布空间。2. 代理人基于 S_t选择一个由 k_t 索引的行动 $b_{k_t} \in \mathcal{B}$，并承担成本 $c_{k_t} \geq 0$。委托人可以观测到该行动 b_{k_t}。3. 随机环境随后选择一个隐藏状态 $\omega_t \in \Omega$，并根据 $p(\omega_t, o_t \mid b_{k_t})$ 仅向代理人发出观测$o_t \in \mathcal{O}$。此时，隐藏状态 ω_t 对代理人和委托人双方都未知。4. 代理人向委托人报告关于隐藏状态的信念 $\hat{\sigma}_t \in \Delta(\Omega)$。5. 委托人基于 $(\hat{\sigma}_t, k_t, S_t)$ 作出决策 $a_t \in \mathcal{A}$。6. 最后，隐藏状态 ω_t 被揭示。委托人获得其效用 $u(a_t, \omega_t)$，并按 $S_t(\hat{\sigma}_t, \omega_t)$ 向代理人支付报酬。

在这里，评分规则 $\mathcal{S}_t$ 是一个取决于代理人的报告 $\hat{\sigma}_t$ 和真实状态 ω_t的支付规则。令 $\mathscr{S}$为范数有界于 $\|S\|_\infty \leq B_S$的评分规则类。下文中，我们假设委托人从 $\mathscr{S}$中选择$\mathcal{S}_t$。我们假设奖励函数 u：$\mathcal{A} \times \Omega \to \mathbb{R}$ 也具有有界范数 $\|u\|_\infty \leq B_u$。我们认为代理人的行动集合 $\mathcal{B}$ 和观测集合 $\mathcal{O}$ 均为有限集。具体而言，代理人的行动空间$\mathcal{B} = \{b_1, \ldots, b_K\}$包含 K 个行动。下文中，我们还用行动索引 k_t 表示代理人的行动。代理人选择其行动 k_t和报告 $\hat{\sigma}_t$的策略，以及委托人选择其行动 a_t 的策略，将在稍后介绍。

值得注意的是，我们的建模捕捉了代理人的动作选择可能影响

环境状态这一内生效应。这比假设状态是外生的更为一般，也反映了现实情形：信息获取行为（例如市场调查）会影响随机环境。考虑第§1节介绍的投资组合经理和金融分析师的例子。分析师撰写的关于某只股票的报告一旦公开发布，可能会产生重大影响并左右该股票价格（Lui 等人，2012）。

信息结构。在本小节余下部分，我们暂时忽略下标 t。在此信息获取过程中，我们假设代理人无所不知，对信息结构了如指掌，即完全知晓每个行动的成本 c_k，以及在行动 b_k下隐藏状态和观测的生成过程 $p(\omega, o \mid b_k)$。因此，获得观测 o后，代理人能够通过贝叶斯法则 $\sigma(\omega) := p(\omega \mid o, b_k) = p(\omega, o \mid b_k)/p(o \mid b_k)$，更新她对隐藏状态的信念 $\sigma \in \Delta(\Omega)$。注意，$\sigma$ 是一个从观测空间映射到概率空间 $\Delta(\Omega)$的随机测度，用以刻画 o中的随机性。令 $\Sigma \subset \Delta(\Omega)$ 为 σ的支撑集。定义 M 为 Σ 的基数；根据 $\mathcal{B}$ 和 $\mathcal{O}$ 的离散性，可得 $M \leq K \times C_{\mathcal{O}}$，其中 $C_{\mathcal{O}}$ 为 $\mathcal{O}$的基数。令$q_k(\sigma) \in \Delta(\Sigma)$表示代理人采取行动 $k \in [K]$时 σ 的分布。由于 σ 已经包含了观测中关于隐藏状态的全部信息，我们忽略观测 o，并将成本$\{c_k\}_{k\in[K]}$ 及信念 $\{q_k\}_{k\in[K]}$ 的分布称为信息结构，这些信息对代理人而言是私有的。我们在表 1 中总结了所有类型的信息。

公开信息	代理人私有信息	委托人私有信息	延迟信息
$S_t, k_t, \widehat{\sigma}_t$	$\{c_k\}_{k\in[K]}, \{q_k\}_{k\in[K]}, o_t, \sigma_t$	u	w_t

表 1：信息类型表。这里，公开信息是指委托人和代理人在每一轮次都可观察到的信息；代理人私有信息是指仅代理人掌握的信息；委托人私有信息是指仅委托人掌握的信息；延迟信息仅包含隐藏状态 ω_t，该状态在第 t 轮次进行时不可观察，但会在第 t 轮次结束时揭示。

2.2 通过适当评分规则获取信息

我们首先注意到，信息获取过程可以建模为一般的斯塔克尔伯格博弈。令 $\mu: \mathscr{S} \to [K]$ 和 $\nu: \mathscr{S} \times \Sigma \times [K] \to \Delta(\Omega)$ 分别表示代理人的行动选择策略和信念报告策略。这里，报告策略 ν 取决于 σ，而不是 o，因为 σ 涵盖了有关隐藏状态的所有信息。给定任意评分规则 S，代理人（作为该博弈中的跟随者）通过 $k = \mu(S)$选择自己的行动，并报告 $\widehat{\sigma} = \nu(S, \sigma, k)$，以最大化自身的期望利润，即 $g(\mu, \nu; S) := \mathbb{E}[S(\nu(S, \sigma, \mu(S)), \omega) - c_{\mu(S)}]$，其中期望是针对 $q_{\mu(S)}$ 而对 ω, σ 的随机性取值。令 $\iota: \mathscr{S} \times \Delta(\Omega) \times [K] \to \mathcal{A}$ 表示委托人的决策策略。因此，委托人（作为该博弈中的领导者）要找到最优评分规则 S 和最佳决策策略 ι，在给定代理人的最佳响应（μ^*, ν^*）的情况下最大化自身的期望利润，即

$h(\mu^*, \nu^*;\ S, \iota) := \mathbb{E}[u(\iota(S, \nu^*(S, \sigma, \mu^*(S))\ \mu^*(S)), \omega)\ -\ S(\nu^*(S, \sigma, \mu^*(S))\ \omega)],$，其中期望是针对 $q_{\mu^*}(s)$ 而对 ω, σ 的随机性取值。最优领导者策略称为强斯塔克尔伯格均衡，可以表述为双层优化问题的解，见 Conitzer (2016)，即，

$$\max_{\iota, S \in \mathscr{S}} h(\mu^*, \nu^*; S, \iota), \quad \text{s.t. } (\mu^*, \nu^*) \in \operatorname*{argmax}_{\mu, \nu} g(\mu, \nu; S). \tag{2.1}$$

然而，式（2.1）中的双层优化在计算上难以处理，因为代理人的行动空间（尤其是报告方案 ν的空间）难以处理。

为解决这一问题，下面我们将建立一个揭示原理结果，保证在不失一般性的前提下，只需关注代理人如实报告的情形。具体而言，存在一个众所周知的评分规则类（见定义 2.1），它刻画了所有评分规则；在此类评分规则下，如实报告符合代理人的最大利益 Savage (1971)。

定义 2.1（适当评分规则）。评分规则 S： $\Delta(\Omega) \times \Omega \ \to\ \mathbb{R}_+$ 若满足：对于任意信念$\sigma \in \Delta(\Omega)$和任意报告的后验信念 $\hat{\sigma} \in \Delta(\Omega)$，均有 $\mathbb{E}_{\omega \sim \sigma} S(\hat{\sigma}, \omega)\ \leq \mathbb{E}_{\omega \sim \sigma} S(\sigma, \omega)$，则称其为适当评分规则。此外，若该不等式对任意 $\hat{\sigma} \neq \sigma$均严格成立，则评分规则 S 为严格适当评分规则。

根据定义，如实报告信念可使代理人的支付最大化。根据揭示原理 Myerson (1979年），我们可以说明，任何可能存在信念不实报告的均衡，都可以通过一个如实报告信念的均衡来实现。这意味着，委托人的最优评分规则可以假定为适当评分规则，且不失一般性，因此我们可以将报告方案限制为 ν 如实报告的方案——我们在下述引理中陈述适用于信息获取模型的揭示原理，并将其证明推迟至 §B。

引理 2.2（揭示　原理）。 存在一个适当评分规则 S^*，它是（2.1）的最优解　.

在下文中，我们用 $\mathcal{S}$ 表示范数有界于 $\|S\|_\infty \leq B_S$ 的适当评分规则类。当 $S \in \mathcal{S}$ 时，代理人的最优报告方案为 $v^*(S, \sigma) = \sigma$，因为如实报告可使支付最大化；委托人的最优决策策略 ι^* 可简化为 $a^*(\sigma) := \iota^*(S, \sigma, k)$，因为 S，且在给定 σ 的情况下， k 不会为隐藏状态提供额外信息。使用记号 $u(\sigma) = \mathbb{E}_{\omega \sim \sigma} u(a^*(\sigma)\ \omega)$、$S(\sigma) = \mathbb{E}_{\omega \sim \sigma} S(\sigma, \omega)$ 和 $k^*(S) = \mu^*(S)$，我们可以将（2.1）中的优化问题转化为

$$\begin{aligned} &\max_{S \in \mathcal{S}} \quad \mathbb{E}_{\sigma \sim q_{k^*(S)}} \left[u(\sigma) - S(\sigma)\right], \\ &\text{s.t.} \quad k^*(S) \in \operatorname*{argmax}_{k \in [K]} \mathbb{E}_{\sigma \sim q_k} S\left(\sigma\right) - c_k, \end{aligned} \tag{2.2}$$

其中，我们将代理人的效用函数记为 $g(k,\ S)$: $= \mathbb{E}_{\sigma \sim q_k} S\left(\sigma\right),\ - c_k$ 并将代理人采取最佳响应时委托人的效用函数 $k^*(S)$ 记为 $h(S)$: $= \mathbb{E}_{\sigma \sim q_{k^*(S)}} [u(\sigma)\ - S(\sigma)]$。该优化问题可以高效求解，例如，求出能使代理人的每种行动成为其最佳响应的最优适当评分规则。我们将在在线学习过程中重新讨论（2.2），届时委托人并不知道信息结构，即 q_k 和 c_k。

合同设计是评分规则设计的一个特例。我们还指出，我们的模型完全能够刻画标准的合同设计问题。如果信息结构是完全揭示的（例如，σ 是一个点信念），那么包含内生状态的模型就退化为标准合同设计问题，此时评分规则 S 成为一种合同，即将（如实报告的）结果映射为支付。详情请参见附录A。

2.3 在线学习以获取信息

现在，我们将求解信息获取最优评分规则这一在线学习问题形式化。我们考虑这样一种情形：委托人仅了解其效用函数u [1]，并且能够观察代理人的行动 b_k [2]。表1汇总了上述讨论的所有信息类型及其可用性。给定 $H_{t-1} = \{(S_\tau,\ k_\tau,\ \sigma_\tau,\ \omega_\tau)\}_{\tau\in[t-1]} \in \mathcal{H}_{t-1}$ 作为委托人在第 t轮之前观察到的历史记录，委托人可以为下一轮部署一个策略来确定评分规则π_t：$\mathcal{H}_{t-1} \to \mathcal{S}$。因此，数据生成过程可描述如下：

$$p^\pi(S_t, k_t, \sigma_t, \omega_t \mid H_{t-1}) = \mathbb{1}\left(S_t = \pi_t(H_{t-1}), k_t = k^*(S_t)\right) \cdot q_{k_t}(\sigma_t) \cdot \sigma_t(\omega_t). \tag{2.3}$$

在线策略 $\pi = \{\pi_t\}_{t\in[T]}$ 的遗憾定义如下：

$$\mathrm{Reg}^\pi(T) := T \cdot \max_{S\in\mathcal{S}} h(S) - \mathbb{E}_\pi\left[\sum_{t=1}^{T} u(a^*(\sigma_t), \omega_t) - S_t(\sigma_t, \omega_t)\right],$$

其中，期望是相对于数据生成过程取的。我们的目标是开发一种在线策略 π_t，以较小的遗憾学习最优评分规则。

3 OSRL-UCB 算法

在本区域中，我们介绍基于上置信界的在线评分规则学习（OSRL-UCB）算法。我们先在 §3.1 概述该算法，并在 §3.2介绍该算法所必需的动作知情预言机。在 §3.3中，我们详细描述 OSRL-UCB 算法。为简化记号，后文令 k 表示 b_k。

[1]由于每轮结束时委托人都知道效用和隐藏状态，因此，只要代理人如实报告自己的信念，就可以高效地学习效用函数。为简化讨论，我们假设委托人已知 u。[2]如果委托人无法观察代理人的行动，仍有办法区分不同的行动。例如，当 q_k 对不同的 k 具有不同的支撑集，且代理人在适当评分规则下如实报告自己的信念时，委托人可以通过多次重复使用同一评分规则，学习特定代理人所选行动对应的支撑集。下次代理人选择相同行动时，委托人便能察觉。不过，一般而言，在无法观察代理人行动的情况下学习最优评分规则仍然是一个难题。

3.1 算法概述

定义 $\mathcal{V}_k = \{S \in \mathcal{S} \mid g(k,\ S) \geq g(k',\ S)\ \forall k' \in [K]\}$ 为代理人采取行动 b_k作为其最佳响应的第 k 个区域。委托人的优化问题（2.2）可改写为

$$\max_{k\in[K]} h^{k,*} := \sup_{S\in\mathcal{V}_k} \mathbb{E}_{\sigma\sim q_k}\left[u(\sigma) - S(\sigma)\right], \tag{3.1}$$

其中，$h^{k,*}$ 是代理人的最佳响应为 k时委托人的最优利润。令 $S^{k,*}$ 为（3.1）中内层问题的最优解。如果委托人知道能够实现 $h^{k,*}$的最佳评分规则 $S^{k,*}$，那么问题就立即简化为一个多臂老虎机问题，其中 $k \in [K]$是 K 个臂，而 $h^{k,*}$ 是臂 k的奖励。因此，这类问题可以通过标准的 UCB 算法（Auer等人，2002年）来处理。然而，存在两个障碍：（i）行动区域 $\mathcal{V}_k$ 未知；（ii）信念分布 q_k 未知。回顾 $\mathcal{V}_k$的定义：

$$\mathcal{V}_k = \left\{S \in \mathcal{S} \mid \langle q_k - q_{k'}, S\rangle_\Sigma \geq c_k - c_{k'}, \forall k' \in [K]\right\}.$$

要识别 $\mathcal{V}_k$，我们只需估计信念分布 q_k 以及定义为$C(i,j) = c_i - c_j$的成对成本差。具体而言，估计量 $\widehat{q} = (\widehat{q}_k)_{k\in[K]}$ 可以根据 σ_t的经验分布进行更新，使得 $k_t = k$；而估计量 $C = (C(i,j))_{i,j\in[K]}$ 可以利用以下恒等式进行更新

$$C(i,j) = \langle q_i - q_j, S\rangle_\Sigma, \quad \forall S \in \mathcal{V}_i \cap \mathcal{V}_j, \tag{3.2}$$

其中我们代入估计量 $\widehat{q}$。此外，我们必须确定一个评分规则 $S \in \mathcal{V}_i \cap \mathcal{V}_j$，以便成功应用 (3.2)。为此，我们对 S_1、S_2 的凸组合采用二分搜索方法，使得 $k^*(S_1) = i$ 且 $k^*(S_2) = j$。二分搜索的详细过程见算法 2。为了向委托人告知 K 行动，并确保委托人能够找到一个评分规则 S，使得$k^*(S) = i$，以便进行二分搜索，我们在 §3.2 中通过示例介绍一个动作知情预言机，它为委托人提供预先已知的评分规则 S_1、S_1、.....、S_K，使得 $k^*(S_i) = i$。

现在，关于 q_k 和 C（i，j）的估计问题已经解决，(3.1) 的内层问题可通过以下约束线性规划求解：

$$\begin{aligned} \mathrm{LP}_k: \quad h^{k,*} = \max_{S\in\mathcal{S}} \quad & \langle q_k, u - S\rangle_\Sigma, \\ \text{s.t.} \quad & \langle q_k - q_{k'}, S\rangle_\Sigma \geq C(k,k'), \quad \forall k' \in [K]. \end{aligned} \tag{3.3}$$

若（3.3）得到求解，我们就可以将（3.1）的外层问题转化为一个老虎机问题：将 $[K]$ 视为臂的集合，并将 $h^{k,*}$ 视为每个臂 $k \in [K]$的奖励。为说明求解 LP_k的方法，设 $\widetilde{\mathcal{Q}} = \{(\widetilde{q}_k)_{k\in[K]}\}$ 和 $\mathcal{C} = \{(C_{ij})_{i,j\in[K]}\}$ 分别为 $\widehat{q}$和 C的置信集，它们以高概率覆盖真实的 q和 C。为鼓励探索，我们遵循乐观原则，通过将 $(\widetilde{q}, \widetilde{C})$ 代入 LP_k，在置信集 $\mathcal{Q} \times \mathcal{C}$上最大化优化目标（委托人的利润），从而为$h^{k,*}$ 求得置信上界。这个乐观版的 LP_k 在（3.7）中表示为 Opt-LP$_k$。我们不显式构造这些置信集，而是利用 $\widehat{q}$ 和 C来构造用于估计效用、支付和成本差异的置信区间。给定由 Opt-LP$_k$为每个行动确定的乐观奖励后，该算法只需选择使奖励最大化的行动 k_t^*，并获得与 k_t^*对应的评分规则解 S_t^*。

算法 1 基于上置信界的在线评分规则学习（OSRL-UCB）

输入：$\mathcal{S}, (\widetilde{S}_1, \cdots, \widetilde{S}_K), T$.
当 $t \le T$ 时
　估计信念分布 $\widehat{q}_k^t$ 通过 (3.4)，置信区间 $I_q^t(k)$ 通过 (3.5);
　估计成本差异 $\widehat{C}^t(i, j)$，置信区间为 $I_c^t(i, j)$ 由（3.6）得到;
　求解 Opt-LP$_k$，见（3.7），其最优值为 $h_{\mathrm{LP}}^t(k)$，解为 $S^t_{\mathrm{LP},k}$ 对于 $k \in [K]$;
　选择最佳臂 $k_t^* \leftarrow \operatorname{argmax}_{k\in[K]} h_{\mathrm{LP}}^t(k)$，并令 $S_t^* \leftarrow S^t_{\mathrm{LP},\, k_t^*}$;
　公布评分规则 $S_t = \alpha_t \widetilde{S}_{k_t^*} + (1-\alpha_t) S_t^*$，并获取代理人的响应 k_t;
　如果 $k_t \neq k_t^*$，则
　　进行二分搜索 BS($S_t, \widetilde{S}_{k_t^*}, k_t^*, t$) 如算法 2所述;
　结束如果
　进入新一轮 $t \leftarrow t+1$;
</style}end while

不过，若要部署 S_t^* 以激励代理人选择行动k_t^*，仍有一个问题需要解决。（3.3）中的约束在 Opt-LP$_k$ 中因采用乐观原则而有所放宽，这可能导致 S_t^* 超出 $\mathcal{V}_{k_t^*}$。为解决这一问题，我们提出部署经过保守调整的评分规则 $S_t = (1-\alpha_t) S_t^* + \alpha_t S_{k_t^*}$，以高概率保证代理人选择 k_t^*。通过令 α_t 随 t衰减，经过保守调整的评分规则也能保证乐观性。算法总结于算法 1，算法详情见 §3.3。

3.2 行动意识预言机

我们假设存在一个预言机，它向委托人提供一组预先已知的评分规则，这些规则能够促使代理人选择其所有行动。这一假设严格弱于许多现有的策略环境在线学习模型；那些模型假设委托人能够预测代理人的最佳响应，或知道代理人效用函数的某些参数。

假设 3.1（动作知情预言机）。我们假设存在一个预言机，它能够给出 K 评分规则 $(S_1, \cdots, S_K)$，使得在 S_k 下，代理人的最佳响应是针对 $k \in [K]$ 的行动 k。此外，对于代理人的利润 $g(k, S) = \mathbb{E}_{\sigma\sim q_k} S(\sigma) - c_k$，我们假设存在 $\varepsilon > 0$，使得对于任意 $k' \neq k$，都有 $g(k, S_k) - g(k', S_k) > \varepsilon$。

借助假设 3.1 中的预言机，委托人一开始就知道代理人可能做出的 K 行动 $\widetilde{S}_1, \cdots, \widetilde{S}_K$，只需依次尝试（$S_1$，$\cdots$，$S_K$）即可轻松确定。此外，我们假设 S_k位于 k号区域 $\mathcal{V}_k = \{S \in \mathcal{S} \mid g(k, S) \ge g(k', S)\ \forall k' \in [K]\}$ 的内部，且与 $\mathcal{V}_k$的边界保持一定的距离。我们指出，让 S_k 远离边界，对于基于 S_k 但存在一定误差的近似评分规则而言，是在高概率下促使代理人采取预期行为的关键。值得注意的是，代理人通过选择 k 获得边际收益 ε这一假设，在

$\sim S_k$ 还确保了区域 $\mathcal{V}_k$具有最小半径，这对应于策略斯塔克尔伯格博弈中的非退化假设（Letchford 等人，2009）。我们首先给出一个不可能性结果：在没有预言机的情况下，无法通过在线学习获得最优适当评分规则。

引理 3.2（不可能性结果）。假设行动数量为 $K \geq 3$，可能的信念数量为 $M \geq 3$。在没有动作知情预言机的情况下，对于任意在线算法，都存在一个实例，使得 $\mathrm{Reg}(T) = \Omega(T)$。

有关困难实例的构造，请参见 §D.1 。在没有动作知情预言机的情况下，任何在线算法都不可避免地会遭受线性遗憾。困难实例背后的直觉是，没有预言机时，任何算法都可能无法找到能够诱导代理人采取理想行动的评分规则。这是因为针对该理想行动的有界适当评分规则空间中的可行区域，即 $\mathcal{V}_k$ k，可能任意小。

我们给出两个动作知情预言机可实现的例子。第一个例子采用随机抽样，改编自（Letchford 等人，2009）。我们从强适当评分规则类 $\mathcal{S}_\beta$ 中抽样，定义见定义 D.1。

示例 3.3（通过随机抽样实现行动意识）。令 $d_1 = \min_{1\leq i<j\leq M} \|\sigma_i - \sigma_j\|_\infty$ and $d_2 = \min_{1\leq k<k'\leq K} \max_{i\in[M]} [q_k(i) - q_{k'}(i)]$。由于适当评分规则类 $\mathcal{S}$ 具有有界范数，β-强适当评分规则类也具有有界体积 $\mathrm{Vol}(\mathcal{S}_\beta) < \infty$。设 $\kappa = d_1^2\beta/2$，并令 $\widetilde{\mathcal{V}}_k = \{S \in \sim\mathcal{S}_\beta \,|\, g(k,S) \geq g(k',S) + \kappa, \forall k' \neq k, k' \in [K]\}$。我们假设对于 $k \in [K]$，$\mathrm{Vol}(\mathcal{V}_k) \geq \eta \mathrm{Vol}(\mathcal{S}_\beta)$。将 $\mathcal{M} = \varnothing$ 初始化为候选集。每次，我们随机抽取一个 β-强适当评分规则 $S \in \mathcal{S}_\beta$，并获得代理人针对 S 的最佳响应 $k^*(S)$。令 $e_i(\sigma,\omega) = \mathbb{1}(\sigma = \sigma_i)$，其中 $i \in [M]$。通过设置$\kappa = d_1^2 \cdot \beta/2$，我们部署 $\{S - \kappa e_i\}_{i\in[M]}$，并观察代理人是否仍作出响应 $k^*(S)$。若是，则令 $\mathscr{S} = \mathscr{S} \cup \{S\}$；否则，拒绝 S。经过 $\mathcal{O}(M\eta^{-1}K \log K)$ 轮次后，以高概率，$\mathscr{S}$将成为参数为 $\epsilon = d_2 \cdot \kappa$ 的有效行动信息预言机。

详见 §D.2。在例 3.3中，我们通过在 $\mathcal{S}$中随机抽样至多 $\mathcal{O}(MK \log K)$ 次，确保得到一组能够成功诱导每种行动的评分规则。接下来，我们通过一个线性契约的例子进行说明，其中的信息结构满足某些特殊性质。

例3.4（通过线性评分规则实现行动意识）。假设这些 K 行动按照成本递增的顺序排列。定义 $u_k = \mathbb{E}_{\omega\sim\sigma,\ \sigma\sim q_k}[u(a^*(\sigma)\ \omega)]$ 并假设 $u_k > 0$。我们假设边际信息增益严格递减，即存在一个 $\epsilon > 0$ 使得

$$\frac{u_K - u_{K-1}}{c_K - c_{K-1}} > \epsilon, \quad \text{and} \quad \frac{u_k - u_{k-1}}{c_k - c_{k-1}} - \frac{u_K - u_k}{c_K - c_k} > \epsilon, \quad \forall k = 2, \cdots, K-1.$$

此外，我们假设 $(u_2 - u_1)/(c_2 - c_1) \leq b$。委托人将评分规则设为 $S(\sigma,\omega) = \lambda u(a^*(\sigma)\ \omega)$，并在 $\lambda \in [0, 2/\epsilon]$ 上进行二分搜索。具体而言，在所有满足 $k^*(\lambda_1 u) \neq k^*(\lambda_2 u)$ 的线段 (λ_1, λ_2) 上迭代进行二分搜索，其中 λ_1, λ_2 是此前在 $[0, 1/\epsilon]$ 上搜索过的相邻点。凭借最大搜索深度 $m = \lceil \log_2(2b(b-\epsilon)/\epsilon^2) \rceil$，我们可以识别出代理人的所有行动。假设 $(\lambda_1^k, \lambda_2^k)$ 是满足 λ_1^k 且此前已搜索过 λ_2^k 的最大线段

且 $k^*(\lambda_1^k u) = k^*(\lambda_1^k u) = k$。令 $\widetilde{S}_k = (\lambda_1^k + \lambda_2^k)u/2$，我们便得到一个具有 $\varepsilon = \epsilon u_1/4b^2$ 的预言机。该过程需要 $\mathcal{O}(K\log_2(\varepsilon^{-1}))$ 轮次。

详见 §D.3 。在这个例子中，我们利用线性契约（Dütting 等人，2019年）的优势来识别代理人的所有行动，并基于委托人的效用生成一组评分规则。具体而言，递减的边际信息增益这一假设在现实实践中十分常见，而且它还能保证所有行动都可以通过线性契约诱导。为了获得这样的预言机，我们至多只需进行 $\mathcal{O}(K\log_2(\varepsilon^{-1}))$ 次试验，这比随机抽样更有效。我们指出，预言机得到的这些预先已知的评分规则在每个区域 $\mathcal{V}_k$中不必是最优的。对于一般情形，它们甚至可以通过从 β-强适当评分规则中随机抽样得到（见例 3.3）；如果边际信息增益严格递减，则可以在线性评分规则类中找到它们（见例 3.4）。

3.3 OSRL-UCB 算法

OSRL-UCB 算法包含以下四个部分：(i) 学习信念分布；(ii) 学习成对成本差；(iii) 通过乐观线性规划 Opt-LP 求解每 $\mathcal{V}_k$ 的最优支付/评分规则$_k$ ，并通过置信上界 UCB 选择最佳“臂”；(iv) 应用保守评分规则，并在 $k_t \neq k_t^*$时进行二分搜索，以改进对 $\mathcal{V}_k$的估计。本小节将详细介绍 OSRL-UCB 算法。

学习信念分布。令 n_k^t 表示代理人在第 t轮之前对行动 k作出响应的总次数。然后，可以通过经验方法学习 q_k ，如下所示：

$$\widehat{q}_k^t(\sigma) = \frac{1}{n_k^t}\sum_{\tau=1}^{t-1}\mathbb{1}(\sigma_\tau = \sigma, k_\tau = k), \forall \sigma \in \Sigma. \tag{3.4}$$

根据Mardia 等人 （2020年中的标准集中性结果，我们将$\widehat{q}_k^t$ 的置信区间定义为

$$I_q^t(k) = \sqrt{\frac{2\log((K)\cdot 2^M t)}{n_k^t}} \tag{3.5}$$

我们在引理E.6中陈述集中性结果。在 $\widehat{q}_k^t$条件下，如果代理人对行动 k 作出响应，则评分规则 S 的估计支付为 $\widehat{v}_S^t(k) = \langle S(\cdot), \widehat{q}_k^t(\cdot)\rangle_\Sigma$。

学习成对成本差。接下来，我们将说明如何学习成对成本差。对于每个满足 $k_\tau = i$ 和 $j \neq i$ 的 $\tau < t$，定义

$$C_+^t(i,j) = \min_{\tau<t:k_\tau=i} \widehat{v}_{S_\tau}(i) - \widehat{v}_{S_\tau}(j) + B_S\left(I_q^t(i) + I_q^t(j)\right),$$
$$C_-^t(i,j) = \max_{\tau<t:k_\tau=j} \widehat{v}_{S_\tau}(i) - \widehat{v}_{S_\tau}(j) - B_S\left(I_q^t(i) + I_q^t(j)\right).$$

对于每一对 (i,j)，我们还定义

$$\theta^t(i,j)=\frac{C_+^t(i,j)+C_-^t(i,j)}{2},\quad \varphi^t(i,j)=\left[\frac{C_+^t(i,j)-C_-^t(i,j)}{2}\right]_+.$$

直接将 $\theta^t(i,j)$ 用作成对成本差的估计并非最佳选择，原因有二：(i) 要使 $\theta^t(i,j)$ 准确，我们需要检测满足 S_τ 位于边界 $\mathcal{V}_i\cap\mathcal{V}_j$上的 S_τ。然而，可能出现 $\mathcal{V}_i\cap\mathcal{V}_j=\varnothing$的情况，从而产生恒定误差。(ii) 即使 $\mathcal{V}_i\cap\mathcal{V}_j\neq\varnothing$，为所有 (i,j) 对找到 $S_\tau\in\mathcal{V}_i\cap\mathcal{V}_j$也可能代价高昂，并可能增加算法复杂度。相反，我们注意到成本中未知参数的数量至多为 K，因此只需在“树”结构中搜索。具体而言，令 $\mathcal{G}=(\mathcal{B},\mathcal{E})$ 表示一个图，其中节点集 $\mathcal{B}$ 对应于 K 个代理行动，边集 $\mathcal{E}$对应于成对成本差 $C(i,j)=c_i-c_j$。此外，我们令 $\varphi^t(i,j)$ 表示赋予边 $C(i,j)$ 的“长度”，它对应于使用 θ^t 估计成本差时的不确定性。因此，以最小误差估计 $C(i,j)$ 的问题，就转化为在图 $\mathcal{G}$上寻找 b_i 与 b_j 之间的最短路径，这可以通过 Dijkstra 算法或 Bellman-Ford 算法高效求解（Ahuja 等人，1990年）。总之，成本差估计为

$$l_{ij}=\text{shortest path}(\mathcal{G},i,j),\quad \widehat{C}^t(i,j)=\sum_{(i',j')\in l_{ij}}\theta^t(i',j'),\quad I_c^t(i,j)=\sum_{(i',j')\in l_{ij}}\varphi^t(i',j'),\tag{3.6}$$

其中，$I_c^t(i,j)$ 是成对成本估计量 $\widehat{C}^t(i,j)$ 的置信区间。此外，我们很容易验证 $\widehat{C}^t(i,j)=-\widehat{C}^t(j,i)$，因为 l_{ij} 与 l_{ji} 相同，并且根据定义，$\theta^t(i,j)=-\theta^t(j,i)$。引理 E.7 证明了 $I_c^t(i,j)$ 作为 $\widehat{C}^t$ 的置信区间的有效性。

求解 Opt-LP 并选择最佳臂。注意，我们有信念分布的估计量 $\widehat{q}_k^t$ 以及成对成本差异的估计量 $\widehat{C}^t$。现在，我们可以针对代理人行动 b_k，求解式（3.3）给出的线性规划（LP），以找到对应的最佳评分规则。为纳入乐观原则，我们使用先前得到的置信区间I_q^t，I_c^t，放宽式（3.3）中的约束。具体而言，对于代理人行动 b_k，我们考虑以下乐观线性规划 Opt-LP$_k$，

$$\begin{aligned}\text{Opt-LP}_k:\quad &\max_{S\in\mathcal{S}}\ \langle u,\widehat{q}_k^t\rangle_\Sigma+B_uI_q^t(k)-v,\\ &\text{s.t.}\quad \left|v-\widehat{v}_S^t(k)\right|\le B_S\cdot I_q^t(k),\\ &v-\widehat{v}_S^t(i)\ge\widehat{C}^t(k,i)-\left(I_c^t(k,i)+B_S\cdot I_q^t(i)\right),\forall i\neq k.\end{aligned}\tag{3.7}$$

此处的优化目标是在代理人的最佳响应 b_k 下最大化委托人的利润，其中我们加入 $B_uI_q^t(k)$，以高概率为真实价值设定上界。第一个约束实际上为支付 v 构造了置信区间 $B_SI_q^t(k)$，而第二个约束则利用 I_c^t 和 I_q^t放宽了 (3.3) 中的初始边界条件。这些约束中的放宽保证 Opt-LP$_k$ 以高概率具有乐观性，正如引理 E.1 所验证的那样。设 Opt-LP$_k$~的最优值和最优解分别为 $h_{\text{LP}}^t(k)$ 和 $S_{\text{LP},\,k}^t$。如果 Opt-LP$_k$ 不可行，我们就令 $h_{\text{LP}}^t(k)=\langle u,\widehat{q}_k^t\rangle_\Sigma-\mathbb{E}_{\sigma\sim\widehat{q}_k}S_k(\sigma)+\widetilde{\ }(B_S+B_u)I_q^t(k)$ 和 $S_k^t=S_k$。接下来，将该问题视为一个 K 臂老虎机，我们选择使委托人的乐观期望利润最大的“臂”，$k_t^*=\operatorname{argmax}_{k\in[K]}h_{\text{LP}}^t(k)$。为简单起见，我们令

$S_t^* := S_{\text{LP},\ k_t^*}^t$。

构造保守评分规则。不过，有时我们可能会对代理人的最佳响应过于“乐观”。也就是说，当我们放宽 Opt-LP$_k$ 中的边界约束时，代理人可能不会在评分规则 S_t^* 下采取行动 k_t^*。这说明，在设计评分规则时，我们应当在一定程度上保持保守。具体而言，我们将保守评分规则$\widetilde{\ }$定义为，$S_t = (1-\alpha_t)S_t^* + \alpha_t S_{k_t^*}$. 其直觉是，由于代理人严格偏好在$\widetilde{\ }$ $\widetilde{\ }$评分规则 $S_{k_t^*}$下采取行动 k_t^*，将 S_t^*与 $S_{k_t^*}$结合起来会提高代理人选择行动 k_t^* 的相对偏好。在引理 E.2 中，我们证明，选取 $\alpha_t = \mathcal{O}(t^{-1/3})$ 后，可以保证代理人以高概率在保守评分规则 S_t 下采取行动 k_t^*。这也意味着，乐观原则（体现为代理人选择行动 k_t^*）得到了保证。

细化参数估计。一旦 S_t 部署后，我们会考虑两种类型的结果。如果代理人采取行动$k_t = k_t^*$，那么我们对代理人决策边界 $\mathcal{V}_{k_t^*}$ 的估计已足够准确，可以直接进入下一轮。否则，代理人会采取另一行动 $k_t \neq k_t^*$，我们需要通过更新 $\mathcal{V}_{k_t^*}$ 来改进对 $\widehat{q}^t$的估计，$\widehat{C}^t$ 以便在未来成功诱导出期望的行动k_t^*。具体而言，由于 S_t的特殊保守构造，只需更新$\widetilde{\ }$中位于 $\mathcal{V}_{k_t^*}$ 与 S_t 之间的 $S_{k_t^*}$边界。为此，我们在连接$\widetilde{\ }$与$S_{k_t^*}$的线段 S_t 上进行二分搜索，找出代理人的最佳响应从行动k_t^*变为另一行动时的第一个切换点。二分搜索算法的详细信息见算法2。

具体而言，在搜索切换点的过程中，我们可以利用两条有用信息：$\widetilde{\ }$首先，对于 $\mathcal{V}_{k_t^*}$位于线段 $(S_{k_t^*}, S_t)$上的边界，我们至少诱导出该边界所分隔的两种行动(i, j)，因此它们的信念分布估计量 $\widehat{q}_i^t$ 和 $\widehat{q}_j^t$ 都会更新。其次，更重要的是，切换点 S 位于边界附近，因此我们可以通过 $\widehat{C}(i,j)$ $\approx \langle \widehat{q}_i^t - \widehat{q}_j^t$，细化 $S\rangle_\Sigma$ 式（3.2）中的成本差。因此，这种二分搜索可以同时细化信念估计量和成本差估计量，从而更准确地估计 $\mathcal{V}_{k_t^*}$。值得注意的是，这一二分搜索过程只需对数级搜索时间即可达到足够的精度。

4 理论结果

在本小节中，我们给出 OSRL-UCB 算法的学习结果。

定理 4.1 (OSRL-UCB 的遗憾)。 在关于动作知情预言机的假设 3.1 下，当 α_t =最小值$\{Kt^{-1/3}, 1\}$ 时，OSRL-UCB 算法 1 的遗憾为

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\big((B_S + B_u)B_S^2\varepsilon^{-2} \cdot K^2 C_{\mathcal{O}} \cdot T^{2/3}\big),$$

其中，B_S 和 B_u 分别界定评分规则和效用函数的幅度，ε是动作知情预言机给出的边际利润增益，K 是代理人的动作数量，$C_{\mathcal{O}}$ 是观测集合的基数。

证明。 详细证明见 §E。 □

~这里，我们用 $\mathcal{O}$ 表示省略对数因子。遗憾关于代理人的动作数量呈二次关系，关于观测集合的基数呈线性关系。值得注意的是，遗憾与隐藏状态 $|\Omega|$的规模无关。此外，在评分规则框架下激励信息获取时，我们实现了关于委托人累积利润的 $\mathcal{O}(T^{2/3})$ 次线性遗憾率。该结果建立在一个温和的假设之上：动作知情预言机能够提供一组评分规则，这些规则具有代理人的边际利润增益，并能诱导代理人采取所有可能的行动。与许多现有研究不同（Balcan 等人，2015；Guo 等人，2022；Wu 等人，2022），我们不假设学习者充分了解其他策略性参与者。此外，我们只假设委托人了解自身效用，并且能够观测代理人的动作选择。关于这两项假设的讨论，请参阅第 §2节的脚注。

对于拥有一组预先已知评分规则的动作知情预言机，这些预先已知的评分规则在每个区域中不必是最优的 $\mathcal{V}_k$。它们甚至可以通过从β-强适当评分规则中随机抽样获得（一般情形见示例 3.3），也可以在边际信息增益严格递减时，从线性评分规则类中发现（见示例 3.4）。我们还给出以下推论，刻画结合寻找动作知情预言机所需努力时的遗憾。

~推论4.2（示例 3.3 中预言机对应的遗憾）。令 $\mathcal{V}_k = \{S \in \mathcal{S}_\beta \,|\, g(k,S) \geq g(k',S) + \kappa, \forall k' \neq k, k' \in [K]\}$ 其中 $\mathcal{S}_\beta \in \mathcal{S}$ 是 β-强适当评分规则类，并假设对于 $k \in [K]$，$\mathrm{Vol}(\mathcal{V}_k) \geq \eta\,\mathrm{Vol}(\mathcal{S}_\beta)$。在部署 OSRL-UCB 算法之前，先运行示例 3.3 中的预言机获取过程 T^γ 轮，则运行 $T - T^\gamma$ 轮后的在线遗憾有如下界：

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\big((d_2 d_1^2 \beta)^{-2} \cdot KM \cdot T^{2/3}\big) + \mathcal{O}(KT\exp(-T^\gamma \eta/M) + T^\gamma),$$

其中 $d_1 = \min_{i\neq j,\ \forall (i,j)\in[M]} \|\sigma_i - \sigma_j\|_\infty$, $d_2 = \min_{k'\neq k,\ (k,\ k')\in[K]} \max_{i\in[M]} [q_k(i) - q_{k'}(i)]$,，且 M 是 Σ 的基数。

此外，我们还刻画了通过线性评分规则获得的动作知情预言机的遗憾。

推论4.3（例 3.4 中使用预言机时的遗憾）。 假设例 3.4 中的模型假设成立，即边际信息增益严格递减。通过在例 3.4 中运行预言机获取过程 $\mathcal{O}(K\log_2(\varepsilon^{-1}))$ 轮次，并在剩余轮次运行 OSRL-UCB 算法，在线遗憾有如下界：

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\big(\varepsilon^{-2} \cdot K^2 C_{\mathcal{O}} \cdot T^{2/3}\big) + \mathcal{O}(K\log_2(\varepsilon^{-1})),$$

其中 $\varepsilon = \epsilon u_1/4b^2$，且 ϵ， u_1， b均为例 3.4中定义的常数。

推论4.2 和推论4.3 均无需使用预言机即可给出遗憾界。具体而言，推论4.2在动作截面体积有下界的假设下，考虑了一个更一般的框架；而推论4.3则假设存在边际信息衰减，这在现实实践中很常见。具体而言，推论4.2 表明，通过对 T^γ 轮次进行$0 < \gamma < 2/3$随机抽样，委托人只需承受 $\widetilde{\mathcal{O}}(T^{2/3})$ 遗憾。此外， γ 可以显著

s 较小，因为第二项随 T^{γ}呈指数级衰减。此外，推论4.3 表明，在预言机获取过程中额外运行固定数量的轮次不会使遗憾界变差。根据Jin 等人(2018) 的讨论，我们的算法也具有如下 PAC 保证。

推论 4.4（PAC 保证）。对于每个 $\zeta > 0$，带有动作知情预言机的 OSRL-UCB 算法使用 $\widetilde{\mathcal{O}}(\varepsilon^{-6}K^{6}C_{\mathcal{O}}^{3}\zeta^{-3})$ 个样本即可找到一个 ζ-最优评分规则。

Zhu 等人（2022）针对在线学习问题提出了一个 $\mathcal{O}$（$T^{2/3}$）朝向最优合同的遗憾下界。尽管标准合同设计是我们模型的特例，但他们的设定与我们的不同，因为代理人可能拥有无限多个行动。因此，我们问题的遗憾下界仍是一个开放问题。为弥合这一差距，我们认为需要回答的关键问题是，能否高效地确定 $\mathcal{V}_{k^*}$ 的决策边界。也就是说，即使已知最佳行动 k^*，学习者仍无法在缺乏足够k^* 相关知识的情况下，从 Opt-LP 求解最优评分规则 $\mathcal{V}_{k^*}$。目前，我们能够构造一类实例，在其中可以通过二分搜索确定 $\mathcal{V}_{k^*}$ 的此类边界，从而避免为每个 q_k 都进行代价高昂的学习 $k \in [K]$。然而，这些高效搜索技术能否推广到任意实例，仍不清楚——对此给出明确答案，应能弥合该问题遗憾上下界之间的差距。我们将这一引人入胜的方向留待未来工作。

5 结论

我们在信息不对称的委托代理框架下，研究如何通过适当评分规则激励信息获取。我们提出 OSRL-UCB 算法，并证明在一个温和的预言机假设下，该算法能够实现 $\mathcal{O}(K^{2}C_{\mathcal{O}}T^{2/3})$ 次线性遗憾。未来的研究方向包括建立遗憾下界，以及将其扩展到上下文和动态设置。

参考文献

Ahuja, Ravindra K, Kurt Mehlhorn, James Orlin, and Robert E Tarjan (1990年)《更快的最短路径问题算法》，ACM期刊（JACM），37（2），213–223。Alon, Tal, Paul D¨utting, and Inbal Talgam-Cohen (2021)《具有单位努力私人化成本的合同》，载于第22届ACM经济学与计算会议论文集，52–69。Amin, Kareem, Afshin Rostamizadeh, and Umar Syed (2013)《面向策略性买家的重复拍卖定价学习》，神经信息处理系统进展，26。Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer (2002)《多臂老虎机问题的有限时间分析》，机器学习，47，235–256。Balcan, Maria-Florina, Avrim Blum, Nika Haghtalab, and Ariel D Procaccia (2015)《无遗憾承诺：斯塔克尔伯格安全博弈中的在线学习》，载于第十六届ACM经济学与计算会议论文集，61–78。Camara, Modibo K, Jason D Hartline, and Aleck Johnsen (2020)《面向无遗憾智能体的机制：超越共同先验》，载于2020年IEEE第61届计算机科学基础年会（FOCS），259–270，IEEE。Castiglioni, Matteo, Andrea Celli, Alberto Marchesi, and Nicola Gatti (2020)《在线贝叶斯劝说》，神经信息处理系统进展，33，16188–16198。Castiglioni, Matteo, Alberto Marchesi, and Nicola Gatti (2021)《贝叶斯代理：线性合同与易处理合同》，载于第22届ACM经济学与计算会议论文集，285–286。（2022）《高效设计合同菜单：随机化的力量》，arXiv预印本 arXiv:2202.10966。Chen, Yiling and Fang-Yi Yu (2021)《最优评分规则设计》，arXiv预印本 arXiv:2107.07420。Conitzer, Vincent (2016)《论斯塔克尔伯格混合策略》，Synthese，193（3），689–703。D¨utting, Paul, Tim Roughgarden, and Inbal Talgam-Cohen (2019)《简单合同与最优合同》，载于2019年ACM经济学与计算会议论文集，369–387。Dutting, Paul, Tim Roughgarden, and Inbal Talgam-Cohen (2021)《合同复杂性》，SIAM计算杂志，50（1），211–254。Feng, Zhe, Chara Podimata, and Vasilis Syrgkanis (2018)《不知道自身价值时的竞价学习》，载于2018年ACM经济学与计算会议论文集，505–522。

Gneiting、Tilmann 和 Adrian E Raftery（2007）《严格适当评分规则、预测与估计》，《美国统计学会期刊》，102（477），359–378。Golrezaei、Negin、Adel Javanmard 和 Vahab Mirrokni（2019）《动态激励感知学习：上下文拍卖中的稳健定价》，《神经信息处理系统进展》，32。Grossman、Sanford J 和 Oliver D Hart（1992）《委托代理问题分析》，载于《保险经济学基础》，302–340：施普林格。Guo、Wenshuo、Michael Jordan 和 Ellen Vitercik（2022）《部分信息拍卖中的无悔学习》，载于《国际机器学习大会》，8039–8055，机器学习研究论文集。Guruganesh、Guru、Jon Schneider 和 Joshua R Wang（2021）《道德风险与逆向选择下的合同》，载于第22届 ACM经济学与计算会议论文集，563–582。Haghtalab、Nika、Thodoris Lykouris、Sloan Nietert 和 Alexander Wei（2022）《与非短视智能体进行斯塔克尔伯格博弈中的学习》，载于第23届 ACM经济学与计算会议论文集，917–918。Hartline、Jason r◯Liren Shan r◯Yingkai Li r◯ Yifan Wu（2022）《多维努力的最优评分规则》，工作论文。Jagadeesan、Meena、Alexander Wei、Yixin Wang、Michael Jordan 和 Jacob Steinhardt（2021）《从老虎机反馈中学习匹配市场的均衡》，《神经信息处理系统进展》，34，3323–3335。Jin、Chi、Zeyuan Allen-Zhu、Sebastien Bubeck 和 Michael I Jordan（2018）《Q学习是否可证明高效？》，《神经信息处理系统进展》，31。Laffont、Jean-Jacques 和 David Martimort（2009）《激励理论》，载于《激励理论》：普林斯顿大学出版社。Lattimore、Tor 和 Csaba Szepesv´ari（2020）《老虎机算法》：剑桥大学出版社。Letchford、Joshua、Vincent Conitzer 和 Kamesh Munagala（2009）《学习并近似最优承诺策略》，载于《国际算法博弈论研讨会》，250–262，施普林格。Li、Yingkai、Jason D Hartline、Liren Shan 和 Yifan Wu（2022）《评分规则优化》，载于第23届 ACM经济学与计算会议论文集，988–989。Lui、Daphne、Stanimir Markov 和 Ane Tamayo（2012）《股票分析师与市场风险评估》，《会计研究杂志》，50（5），1287–1317。

Mardia、Jay、Jiantao Jiao、Ervin T´anczos、Robert D Nowak 和 Tsachy Weissman（2020）《离散分布经验分布的集中不等式：超越类型方法》，《信息与推断：IMA期刊》，9（4），813–850。

Marecki、Janusz、Gerry Tesauro 和 Richard Segal（2012）《与未知对手进行重复斯塔克尔伯格博弈》，载于《第11届自主代理人与多代理系统国际会议论文集》第2卷，821–828。

Mas-Colell、A.、M.D. Whinston 和 J.R. Green（1995）微观经济理论，牛津学生版：牛津大学出版社，https://books.google.com/books?id=sQGDQgAACAAJ。

Myerson、Roger B（1979年）《激励相容与讨价还价问题》，《Econometrica：计量经济学会期刊》，61–73。

Neyman、Eric、Georgy Noarov 和 S Matthew Weinberg（2021）《激励精确度的二元评分规则》，载于《第22届 ACM 经济学与计算会议论文集》，718–733。

Oesterheld、Caspar 和 Vincent Conitzer（2020年）《委托人—专家问题的最小遗憾合同》，载于《Web与互联网经济学：第16届国际会议 WINE 2020，中国北京，2020年12月7日至11日，论文集》第16卷，430–443，施普林格。

Papireddygari, Maneesha 和 Bo Waggoner（2022）“通过评分规则订立信息获取合同”，载于第23届 ACM经济学与计算会议论文集，EC ’22，703–704，纽约州纽约：美国计算机协会，10.1145/3490486.3538261。

Savage, Leonard J（1971）“个人概率与期望的引出”，《美国统计学会期刊》，66(336)，783–801。

Smith, Stephen A（2004）《契约理论》：牛津大学出版社。

von Stackelberg, H.（1952）《市场经济理论》：William Hodge，https://books.google.com/books?id=fjIAtQEACAAJ。

Wu, Jibang、Zixuan Zhang、Zhe Feng、Zhaoran Wang、Zhuoran Yang、Michael I Jordan 和 Haifeng Xu（2022）“序贯信息设计：马尔可夫说服过程及其高效强化学习”，arXiv 预印本 arXiv:2202.10678。

Zhu、Banghua、Stephen Bates、Zhuoran Yang、Yixin Wang、Jiantao Jiao 和 Michael I Jordan（2022）《在线合同设计的样本复杂度》，arXiv 预印本 arXiv:2211.05732。

Zu、You、Krishnamurthy Iyer 和 Haifeng Xu（2021）《即时说服：应对无知的鲁棒性》，EC ’21，927–928，纽约州纽约：美国计算机协会，10.1145/3465456.3467593。

作为评分规则设计特例的合同设计

在本区域中，我们将合同设计框架与评分规则设计框架进行比较，并将标准合同设计问题化简为评分规则设计问题的一个特例。在合同设计问题中，我们将 Ω 称为结果空间。考虑以下合同设计问题：

委托代理框架中的合同设计问题

在第 t 轮，委托人与代理人按以下方式进行：

1. 委托人向代理人宣布一份合同 $C_t: \Omega \to \mathbb{R}_+$。2. 代理人根据 C_t 选择一个由 k_t 索引的行动 $b_{k_t} \in \mathcal{B}$，并承担成本 $c_{k_t} \geq 0$。委托人可以观测到行动 b_{k_t}，但成本 c_{k_t} 对代理人而言是私有信息。3. 随后，随机环境根据 $p(\omega_t \mid b_{k_t})$ 从 Ω 中选取一个结果 $\omega_t \in$。该结果 ω_t 会作为观测揭示出来，但生成过程 $p(\omega_t \mid b_{k_t})$ 对代理人而言是私有信息。4. 委托人根据 ω_t 做出决策 $a_t \in \mathcal{A}$。5. 最后，委托人获得其效用 $u(a_t, \omega_t)$，并通过 $C_t(\omega_t)$ 向代理人支付报酬。

该合同设计问题与评分规则设计问题的区别在于，ω_t具有揭示性，而且代理人的行动只会通过其动作选择影响委托人的效用，而不会提供任何报告。我们指出，如果委托人了解效用函数并且始终采取最佳行动，就可以用 $u(\omega_t) = u(a^*(\omega_t)\ \omega_t)$ 替换 $u(a_t, \omega_t)$。在该合同设计问题中，代理人具有行动策略 $\pi: \mathcal{C} \to [K]$，其中 $\mathcal{C}$ 是合同空间。委托人的目标是设计能够使其利润最大化的最优合同，即最大化效用减去支付，同时满足代理人通过最大化自身利润（即支付减去成本）所作出的最佳响应。该合同设计问题中的斯塔克尔伯格博弈可以表述为

$$\begin{aligned} \max_{C \in \mathcal{C}} \quad & \mathbb{E}_{\omega \sim p(\cdot \mid b_{\pi^*(C)})} \left[u(\omega) - C(\omega) \right], \\ \text{s.t.} \quad & \pi^*(C) \in \operatorname*{argmax}_{k \in [K]} \mathbb{E}_{\omega \sim p(\cdot \mid b_k)} C(\omega) - c_k, \end{aligned} \tag{A.1}$$

接下来，我们旨在说明评分规则设计问题中的以下结论：(i) 如果隐藏状态能够被完全揭示，即代理人采取行动后，$o_t = \omega_t$ 是其观测，那么存在一类评分规则，使得上述合同设计问题等价于评分规则设计问题。(ii) 使用适当评分规则时，委托人在评分规则框架下的最优利润不低于其在契约框架下的最优利润。

为证明（i），考虑评分规则类

$$\mathscr{S}^C = \{ S \in \mathscr{S} \mid S(\widehat{\sigma}, \omega) = \mathbb{1}(\widehat{\sigma} = e_\omega) \cdot C(\omega), \forall C \in \mathcal{C} \},$$

其中 $e_{\omega'}(\omega) = \mathbb{1}(\omega = \omega') \in \Delta(\Omega)$。尽管 $S \in \mathscr{S}^C$ 可能不是适当评分规则，但代理人

始终会如实报告，即 $\hat{\sigma} = e_\omega$，因为只有如实报告，她才能获得非零支付。因此，这个隐藏状态 ω_t 也会通过代理人的报告向委托人揭示。评分规则问题中的（2.2）所示的斯塔克尔伯格博弈因此可以写为：

$$\begin{aligned} &\max_{S\in\mathscr{S}^C} \quad \mathbb{E}_{\omega\sim p(\omega\,|\,b_{k^*(S)})}\left[u(\omega) - S(e_\omega,\omega)\right], \\ &\text{s.t.} \quad k^*(S) \in \operatorname*{argmax}_{k\in[K]} \mathbb{E}_{\omega\sim p(\cdot\,|\,b_k)} S\left(e_\omega,\omega\right) - c_k, \end{aligned} \tag{A.2}$$

注意到 $S(e_\omega,\ \omega)\ = C(\omega)$，可知（A.2）和（A.1）实际上是同一个问题。因此，我们得出结论：合同设计问题完全归约为这个评分规则设计问题。

我们还指出，即使隐藏状态完全揭示信息，委托人也未必事先知情。采用适当评分规则时，代理人总会如实报告。此外，根据引理 2.2 中陈述的揭示原理，对于任意 $S \in \mathscr{S}^C$，总存在一个适当评分规则 $\widetilde{S} \in \widetilde{\mathcal{S}}$，为代理人带来相同的期望支付 $\mathbb{E}_{\omega\sim p(\cdot\,|\,b_k)}S(e_\omega,\omega) = \mathbb{E}_{\omega\sim p(\cdot\,|\,b_k)}C(\omega)$，尽管$\sim S($ $e_\omega,\omega)$ 在逐点意义上可能不等于 $C(\omega)$。因此，命题（ii）也得到了证明，我们由此得出结论：（适当）评分规则框架比契约框架功能更强，因为它多询问了代理人一个关于其信念的问题。

B 关于揭示原理的更多细节

在本区域中，我们将对模型中的揭示原理给出正式论证。也就是说，只设计能够激励代理人如实报告的适当评分规则，不失一般性。

定义 B.1（适当评分规则）。评分规则 $S:\ \Delta(\Omega)\ \times\Omega\ \to\ \mathbb{R}_+$ 若满足以下条件，则称为适当：对于任意信念$\sigma \in \Delta(\Omega)$和任意报告的后验信念 $\hat{\sigma} \in \Delta(\Omega)$，我们有 $\mathbb{E}_{\omega\sim\sigma}S(\hat{\sigma},\omega)\ \leq \mathbb{E}_{\omega\sim\sigma}S(\sigma,\omega)$。此外，若对于任意 $\hat{\sigma} \neq \sigma$，该不等式都严格成立，则称该评分规则 S 为严格适当。

令 S 为一个适当评分规则，并固定代理人用于 $\mu\ (\cdot)$ 行动选择的策略。对于报告方案ν 以及任意真实信念 σ，由定义2.1可得

$$g^{\mu,\nu}(S) = \mathbb{E}_{\omega\sim\sigma}S(\nu(S,\sigma,\mu(S)),\omega) - c_{\mu(S)} \leq \mathbb{E}_{\omega\sim\sigma}S(\sigma,\omega) - c_{\mu(S)}.$$

因此，在适当评分规则这一类规则下，代理人总是如实报告自己的信念，从而使其期望支付最大化。下文中，我们令 $S(\hat{\sigma},\sigma)\ = \mathbb{E}_{\omega\sim\sigma}S(\hat{\sigma},\omega)$。为举例说明适当评分规则，考虑二元隐状态空间 $\Omega\ = \{0,1\}$，在此空间中，适当评分规则这一类规则具有 Schervish 表示（Gneiting 和 Raftery，2007），即 $S(p,1)\ = G(p)\ + (1-p)G'(p)$，且 $S(p,0)\ = G(p)-pG'(p)$，其中 $p\in[0,1]$ 和 $G:\ [0,1]\to\mathbb{R}_+$ 是一个凸函数。直观地说，信念为 σ、报告为 p 时，适当评分规则 S 的期望支付为 $S(p,\sigma)\ = G(p)+(\sigma-p)G'(p)$，这对应于 G在 p处的支撑线。在这个例子中，G的凸性保证了$S(p,\sigma)\ = G(p)+(\sigma-p)G'(p)\ \leq G(\sigma)\ = S(\sigma,\sigma)$。此外，引理 2.2 中的下一个观测表明，将范围限制在适当评分规则这一类规则内，并不会对委托人的目的造成任何一般性损失。

引理 B.2（揭示原理的重述）。 存在一个适当评分规则 S^*，如果代理人在任何适当评分规则下都如实报告，该规则就是（2.1）的一个最优解。

证明。我们首先证明，对于任意满足 $\|S\|_\infty \le B_S$ 的评分规则 S，总存在一个适当评分规则 $S'(\hat{\sigma}, \omega) = S(\nu^*(S, \hat{\sigma}, k)\ \omega)$，使得对于任意 $\sigma \in \Delta(\Omega)$ 和任意代理人的动作选择，它们向代理人支付的金额相同。为证明 S' 是一个范数有界于 B_S 的适当评分规则，我们有 $B_S \ge S(\nu^*(S, \hat{\sigma}, k)\ \omega) = S'(\hat{\sigma}, \omega) \ge 0$，并且

$$S'(\hat{\sigma}, \sigma) = S(\nu^*(S, \hat{\sigma}, k), \sigma) \le S(\nu^*(S, \sigma, k), \sigma) = S'(\sigma, \sigma).$$

将 $\hat{\sigma} = \sigma$ 代入 S' 的定义，即可验证 S' 的支付金额相同，因为在适当评分规则下，代理人会如实报告；再对 $\omega \sim \sigma$ 取期望，即 $S'(\sigma, \sigma) = S(\nu^*(S, \sigma, k)\ \sigma)$，由此证明第一部分。

其次，我们证明，鼓励代理人报告真实信念 σ 会使委托人的收益不减。注意

$\max_\iota \mathbb{E}_{\omega\sim\sigma, \sigma\sim q_k}[u(\iota(S, \nu^*(S, \sigma, k)\ k)\ \omega)] \le \max_\iota \mathbb{E}_{\omega\sim\sigma, \sigma\sim q_k}[u(\iota(S, \sigma, k)\ \omega)] = \mathbb{E}_{\omega\sim\sigma, \sigma\sim q_k}[u(a^*(\sigma)\ \omega)]$，其中 $a^*(\sigma) = \operatorname{argmax}_{a\in\mathcal{A}} \mathbb{E}_{\omega\sim\sigma} u(a, \omega)$。这里，不等式成立是因为 $\nu^*(S, \sigma, k)$ 是 σ 的函数；等式成立是因为 $\omega \perp\!\!\!\perp (S, k) \mid \sigma$。总之，选择 S' 而不是 S，支付金额完全相同，同时委托人的收益不减。因此，选择 S' 可使委托人的利润不减，并且必定存在一个既是适当评分规则、又是最优评分规则的评分规则。

□

根据引理2.2，委托人的最优评分规则属于有界范数的适当评分规则$\mathcal{S}$，其范数有界 $\|S\|_\infty \le B_S$。使用适当评分规则时，一个顾虑是，如实报告可能并非代理人效用的唯一最大化选择。不过，我们注意到，适当评分规则类是一个凸包，其内部为严格适当评分规则。因此，在任意适当评分规则中加入无穷小比例的严格适当评分规则，始终会得到一条适当评分规则，并且该规则严格适当。从这个意义上说，我们可以放心地假设，在适当评分规则下，代理人总是报告真实的后验概率。下表总结了我们模型中的所有不同信息类型。

C 二分搜索算法的更多细节

本区域中，我们对二分搜索算法作如下总结：在步骤 t，二分搜索算法作用于连接两个评分规则 S_0 和 S_1 的线段上，其中代理人在 S_1 下的最佳响应为 $k^*(S_1)$。该二分搜索的目标是在这条线段上找到代理人的最佳响应从 k_1 切换到另一行动的第一个切换点。二分搜索算法持续更新 $\lambda_{\max}$ 和 $\lambda_{\min}$，将其作为包含第一个切换点的候选区间。注意，任意 $\lambda \in [0, 1]$ 都对应于连接 S_0 和 S_1 的线段上的一个评分规则。每次，算法都会采用与（$\lambda_{\min} + \lambda_{\max}$）/2 对应的评分规则，因此候选区间的长度减半。此外，我们考虑将二分搜索进行到有限深度 m，使搜索误差满足最小不确定性

算法 2 二分搜索 BS($S_0, S_1, k^*(S_1)$, t) 1：输入： $S_0, S_1, k^*(S_1)$ $t, k_t, \{I_q^t(k)\}_{k\in[K]}$；2：输出： t；3：如果 $k_t = k^*(S_1)$，则 4：终止二分搜索算法；5：结束 if；6：初始化 $\lambda_{\min} \leftarrow 0, \lambda_{\max} \leftarrow 1$, $t_0 \leftarrow t$；7：当 $\lambda_{\max} - \lambda_{\min} \geq I_q^{t_0}(k_t) \wedge I_q^{t_0}(k^*(S_1))$ 时，8：开始新一轮 $t \leftarrow t+1$；9：选取 $\lambda \leftarrow (\lambda_{\min} + \lambda_{\max})/2$ 作为中点；10：委托人宣布评分规则 $S_t = (1-\lambda)S_0 + \lambda S_1$，并获得代理人的响应 k_t；11：如果 $k_t = k^*(S_1)$，则 12：更新 $\lambda_{\max} \leftarrow \lambda$；13：否则 14：更新 $\lambda_{\min} \leftarrow \lambda$；15：结束 if；16：结束 while

的 $\widehat{q}^t_{k'}$，$\widehat{q}^t_{k^*_t}$，即 $2^{-m} \leq I_q^t(k') \wedge I_q^t(k_t^*)$ 在算法 2中。因此，二分搜索能以对数级搜索时间达到足够的精度。

D 动作知情预言机在区域 3.2中的证明

本节中，我们将证明区域 3.2中示例所述的论断。

D.1 引理 3.2的证明

不失一般性，我们只需考虑 $K = 3$ 和$M = 3$的情形来构造一个困难实例。为简化讨论，我们考虑两个隐藏状态 $\Omega = \{\omega_1, \omega_2\}$，并将 $B_S = 1$ 视为评分规则的装箱条件。构造困难实例的思路如下：

(i) 确保行动 a_2 是最优评分规则下代理人的最优响应，而任何不能诱使代理人选择 a_2 的评分规则都会产生至少为 1 的常数遗憾。(ii) 令 $\mathcal{V}_2$ 为 $\mathcal{S}$ 边界上的一个“单点”，并证明在没有预言机的情况下，任何算法都有很大概率无法找到位于 $\mathcal{V}_2$ 中的正确评分规则。

由于 $\Omega = \{\omega_1, \omega_2\}$，我们也可以用分配给 ω_1 的概率质量来等价地表示 σ_1、σ_2、σ_3。稍微滥用一下记号，我们令 σ_i 表示 $\sigma_i(\omega_1)$，令 $1-\sigma_i$ 表示 $\sigma_i(\omega_2)$。我们考虑 $\Sigma = \{\sigma_1 =$的情形

(0, 1) $\sigma_2 = (.5, .5)$ $\sigma_3 = (1, 0)\}$，适用于 $M = 3$。为简便起见，对于 $i = 1$，2，3，我们令 $S_i = \mathbb{E}_\omega S(\sigma_i, \omega)$。因此，适当评分规则的集合由以下条件规定：

$$0 \le S_2 \le \frac{S_1 + S_3}{2}, \quad B_S \ge S_1 \ge 0, \quad B_S \ge S_3 \ge 0. \tag{D.1}$$

我们指出，第一个条件要求 $G(\sigma) = \mathbb{E}_\omega[S(\sigma, \omega)]$ 为凸曲线，从而保证该评分规则是适当的（Gneiting 和 Raftery，2007）；其余条件只是盒约束。事实上，对于满足上述约束的任意 $S = (S_1, S_2, S_3)$，都可以利用 Schervish 表示（Gneiting 和 Raftery，2007），通过拟合一个定义在 $[0, 1]$ 上的凸函数 G，使其经过 $(0, S_1)$、$(.5, S_2)$ 和 $(1, S_3)$，并考察 G 在这三个点处的支撑超平面，从而轻易构造出一个适当评分规则 $S(\sigma, \omega)$。

为确保 $\mathcal{V}_2$ 是 (ii) 中边界上的一个单点，我们令 $q_2 - q_1$ 和 $q_2 - q_3$ 为

$$q_2 - q_1 = \beta \cdot (0, -1, 1)^\top, \quad q_2 - q_3 = \beta \cdot (1, 1, -2)^\top,$$

其中，β 应视为一个固定常数，而 q_1、q_2、a_3 是我们要设计的量。这些条件推出了 $\mathcal{V}_2$的以下表达式：

$$\mathcal{V}_2 = \{S \in \mathcal{S} \,|\, (0, -1, 1)S \ge \beta^{-1}(c_2 - c_1), \quad (1, 1, -2)S \ge \beta^{-1}(c_2 - c_3)\}.$$

进一步令 $e_1 = \beta^{-1}(c_2 - c_1)$ 且 $e_2 = \beta^{-1}(c_2 - c_3)$。容易验证，只要 $e_1 + e_2 = 1$且 $-1/2 \le e_1 \le 0$，$\mathcal{V}_2$ 就会收缩为由 $S^* = (1, -e_1, 0)$ 给出的单点，并且方程（）中的条件始终成立。因此，我们可以放心地将讨论范围限制在由参数e_1 和 c_1所指定的实例上，c_2，c_3 可相应调整。因此，我们实际上考虑的是一个单参数线性系统。

为确保 (i)，即 a_2 的最优性，我们首先需要确保在 S^* 处，a_2 以某个常数 $\gamma \ge 1$ 的幅度优于 a_1 和 a_2：

$$\begin{aligned} \langle q_2 - q_1, u - S^* \rangle &\ge \gamma, \\ \langle q_2 - q_3, u - S^* \rangle &\ge \gamma, \end{aligned} \tag{D.2}$$

其中 $S^* \in \mathcal{V}_2$。注意，由于 $q_2 - q_1$与 $q_2 - q_3$不在同一直线上，我们总能选取u 使得无论 S^* 为何，（D.2）均成立。对于任意 $S \in \mathcal{V}_1$，我们可以验证，委托人的利润为，

$$\langle q_2, u - S^* \rangle - \langle q_1, u - S \rangle = \langle q_2 - q_1, u - S^* \rangle + \langle q_1, S - S^* \rangle \ge \gamma - 1,$$

其中，最后一个不等式直接由（D.2）以及 S 的每个元素都应在 $[0, 1]$范围内这一事实得出。对于 $S \in \mathcal{V}_3$也同样如此。因此，只需找到一个满足（D.2）且 $\gamma = 2$的实例。

困难实例的构造：基于上述讨论，我们通过令$\gamma = 2$、$\beta = 1/16$和

$$\begin{aligned} q_1 &= (3/4, 3/16, 1/16), \\ q_2 &= (3/4, 1/8, 1/8), \\ q_3 &= (11/16, 1/16, 1/4) \\ u &= (96, 0, 32) + (1, -e_1, 0). \end{aligned}$$

这里，唯一未知的是成本 c_1、c_2、c_3，未知成本的影响被编码在$e_1 \in [-1/2, 0]$中，而$e_1 \in [-1/2, 0]$只会影响 S^*的位置。

由于我们考虑的是一个第二个坐标 S^* 未确定的单参数线性系统，因此不失一般性，可以考虑在直线 $l = (1, \lambda, 0)$ 上寻找最优评分规则，其中 $0 \le \lambda \le 1/2$。假设我们在步骤 t 采用评分规则 $S^t = (1, b^t, 0)$，且代理人的响应为 $a^t \in \mathcal{A} = \{a_1, a_2, a_3\}$。事实上，对于任意 $b^t < e_1$，代理人都会响应 a_1；对于任意 $b^t > e_1$，代理人都会响应 a_3。于是我们断言，委托人的任何在线策略都会导致线性遗憾，因为在仅获得代理人行动反馈的情况下，任何算法都无法确定 e_1 的确切位置（或许能做的最好选择是对 e_1 进行二分搜索，但这也无济于事）。此外，任何 $b^t \neq e_1$ 在每轮都会造成至少为 1 的次优性。我们还指出，根据有利于委托人的平局打破规则，即使令 $\mathcal{V}_2$ 收缩至一个点，S^* 仍然是最优评分规则。尽管如此，我们总能使 $\mathcal{V}_2$ 成为围绕 e_1 的尽可能小的区间。因此，我们得出结论：对于任何算法，总存在一个困难实例，使其产生线性遗憾 $\mathrm{Reg} \ge \Omega(T)$。引理 3.2 的证明至此完成。

D.2 示例 3.3 的证明3.3

为了确保我们采样的评分规则在诱导动作上有 ϵ 更好的表现，我们从强适当评分规则中采样。

定义 D.1（强适当评分规则）。若评分规则 $S: \Delta(\Omega) \times \Omega \to \mathbb{R}$ 满足：对所有 $p, q \in \Delta(\Omega)$，均有 $\mathbb{E}_{\omega \sim q}[S(q, \omega)] - \mathbb{E}_{\omega \sim q}[S(p, \omega)] \ge \frac{\beta}{2} \|q - p\|_1^2$，则称其为 β-强适当。

示例 D.2（示例 3.3 的重述）。令 $d_1 = \min_{1 \le i < j \le M} \|\sigma_i - \sigma_j\|_\infty$ and $d_2 = \min^{1 \le k < k' \le K} \max_{i \in [M]} [q_k(i) - q_{k'}(i)]$。由于适当评分规则类 $\mathcal{S}$ 具有有界范数，β-强适当评分规则类也具有有界体积 $\mathrm{Vol}(\mathcal{S}_\beta) < \infty$。设 $\kappa = d_1^2 \beta / 2$，并令 $\widetilde{\mathcal{V}}_k = \{S \in \mathcal{S}_\beta \mid g(k, S) \gtrsim g(k', S) + \kappa, \forall k' \neq k, k' \in [K]\}$。我们假设对于 $k \in [K]$，$\mathrm{Vol}(\mathcal{V}_k) \ge \eta \mathrm{Vol}(\mathcal{S}_\beta)$。将 $\mathcal{M} = \varnothing$ 初始化为候选集。每次随机采样一个 β-强适当评分规则 $S \in \mathcal{S}_\beta$，并获得代理人关于 S 的最佳响应 $k^*(S)$。令 $e_i(\sigma, \omega) = \mathbb{1}(\sigma = \sigma_i)$ for $i \in [M]$。通过设置 $\kappa = d_1^2 \cdot \beta / 2$，我们部署 $\{S - \kappa e_i\}_{i \in [M]}$，并观察代理人是否仍然响应 $k^*(S)$。如果是，则令 $\mathscr{S} = \mathscr{S} \cup \{S\}$；否则，拒绝 S。经过 $\mathcal{O}(M \eta^{-1} K \log K)$ 轮后，以高概率，$\mathscr{S}$ 是参数为 $\epsilon = d_2 \cdot \kappa$ 的有效动作知情预言机。

~ ~注意，$\mathcal{V}_k$ 的体积是连续的，随 β递减，并且当 $\beta \to 0$时趋于 $\mathrm{Vol}(\mathcal{V}_k)$。因此，如果 $\mathcal{V}_k$占总体积的一个固定比例，我们就能找到 β，使得 $\mathcal{V}_k$ 占总体积的一个固定比例。

证明。首先，如果$\kappa = d_1^2 \cdot \beta / 2$，$\{S - \epsilon e_i\}_{i \in [M]}$ 也是适当的，因为 S 是 β-强适当的。

接下来，令 $d_2 = \min_{k'} \max_i [q_k(i) - q_{k'}(i)]$。我们证明，每次命中 $\mathcal{V}_k$ 中一个~评分规则的概率至少为 η。假设对于所有 $i \in [M]$，若代理人更偏好行动 k 而非 k'，则 S 不会被拒绝：

$g(k, S-\kappa e_i) - g(k', S-\kappa e_i) \geq 0$。根据 $S-\kappa e_i$ 的适当性，

$$\begin{aligned} g(k, S-\kappa e_i) - g(k', S-\kappa e_i) &= \langle S-\kappa e_i, q_k - q_{k'} \rangle - (c_k - c_{k'}) \\ &= \langle S, q_k - q_{k'} \rangle - \kappa(q_k(i) - q_{k'}(i)) - (c_k - c_{k'}) \geq 0. \end{aligned}$$

若这对任意的 $i \in [M]$ 和 $k' \in [K]$ 都成立，则意味着 $\langle S, q_k - q_{k'} \rangle - (c_k - c_{k'}) \geq \max^{i\in[M]} \kappa(q_k(i) - q_{k'}(i)) \geq \sim \kappa d_2$。另一方面，如果 S 属于集合 $\mathcal{V}_k$，它就永远不会被拒绝。

经过 $\frac{c}{\eta} K \log K$ 轮次后，某个行动未被诱导的概率为 $(1-\eta)^{\frac{c}{\eta} K \log K} = (\frac{1}{K})^{cK}$。根据并集界，任意行动未被诱导的概率至多为 $(\frac{1}{K})^{cK-1}$。将 c 设得足够大，这个概率就会足够小。总之，经过 $\mathcal{O}(M\eta^{-1} K \log K)$ 轮次后，我们得到一个参数为 $\epsilon = d_2 \cdot \kappa$ 的预言机。

□

D.3 示例3.4 的证明

示例 D.3（示例 3.4 的重述）。假设这些 K 个信号按成本递增顺序排列。定义 $u_k = \mathbb{E}_{\omega\sim\sigma,\sigma\sim q_k}[u(a^*(\sigma), \omega)]$ 并假设 $u_k > 0$。我们假设边际信息增益严格递减，即存在一个 $\epsilon > 0$ 使得

$$\begin{aligned} \frac{u_K - u_{K-1}}{c_K - c_{K-1}} &> \epsilon, \quad \text{and} \\ \frac{u_k - u_{k-1}}{c_k - c_{k-1}} - \frac{u_{k+1} - u_k}{c_{k+1} - c_k} &> \epsilon, \quad \forall k = 2, \cdots, K-1. \end{aligned}$$

此外，我们假设 $(u_2 - u_1)/(c_2 - c_1) \leq b$。委托人将评分规则设为 $S(\sigma, \omega) = \lambda u(a^*(\sigma), \omega)$，并在 $\lambda \in [0, 2/\epsilon]$ 上进行二分搜索。具体而言，我们在所有满足 $k^*(\lambda_1 u) \neq k^*(\lambda_2 u)$ 的线段 (λ_1, λ_2) 上迭代进行二分搜索，其中 λ_1, λ_2 是此前在 $[0, 1/\epsilon]$ 上搜索过的相邻点。通过最大搜索深度 $m = \lceil \log_2(2b(b-\epsilon)/\epsilon^2) \rceil$，我们可以识别代理人的所有行动。假设 $(\lambda_1^k, \lambda_2^k)$ 是最大的线段，且此前已搜索过 λ_1^k 和 λ_2^k，并且 $k^*(\lambda_1^k u) = k^*(\lambda_2^k u) = k$。令 $\widetilde{S}_k = (\lambda_1^k + \lambda_2^k)u/2$，即可得到一个搜索深度为 $\varepsilon = \epsilon u_1 / 4b^2$ 的预言机。该过程需要 $\mathcal{O}(K \log_2(\varepsilon^{-1}))$ 轮次。

证明。首先，我们说明，对于任意 $i = 2, \ldots, K$，令 $\lambda = \lambda_i = (c_i - c_{i-1})/(u_i - u_{i-1})$，代理人在采取行动 i 和 $i-1$ 之间无差异。此外，我们令 $\lambda_1 = 0$ 和 $\lambda_{K+1} = 2/\epsilon$。对于任意 $\lambda \in (\lambda_i, \lambda_{i+1})$，代理人的最优回应是采取行动 b_i。此外，$\lambda_2 = (c_2 - c_1)/(u_2 - u_1) \geq 1/b$ 和 $\lambda_K = (c_K - c_{K-1})/(u_K - u_{K-1}) \leq 1/\epsilon$。因此，通过在 $\lambda \in [0, 2/\epsilon]$ 中搜索，可以保证线性契约诱导出所有行动。

由于序列 $(u_i - u_{i-1})/(c_i - c_{i-1})$ 严格递减，λ_i 严格递增，且 $\lambda_i - \lambda_{i-1}$ 以下界 $1/(b-\epsilon) - 1/b = \epsilon/b(b-\epsilon)$ 为下界，其中该下界由 $\lambda_2 = 1/b$ 和 $\lambda_3 = 1/(b-\epsilon)$ 取得。结合搜索是在 $[0, 2/\epsilon]$ 上进行这一事实，我们得到最大搜索深度 $m = \lceil \log_2(2b(b-\epsilon)/\epsilon^2) \rceil$，使得相邻搜索点之间的间隔不超过 $\epsilon/2b(b-\epsilon)$。因此，利用 $\lambda_i - \lambda_{i-1} \geq \epsilon/b(b-\epsilon)$ 这一事实，我们可以保证所有行动都是由该二分搜索诱导的。

对于所选的 λ_1^k 和 λ_2^k，满足 $k^*(\lambda_1^k u) = k^*(\lambda_2^k u) = k$，我们可以保证 $\lambda_k \le \lambda_1^k \le \lambda_k + \epsilon/2b(b-\epsilon)$ 且 $\lambda_{k+1} - \epsilon/2b(b-\epsilon) \le \lambda_2^k \le \lambda_{k+1}$。因此，$\lambda_k + \epsilon/4b(b-\epsilon) \le (\lambda_1^k + \lambda_2^k)/2 \le \lambda_{k+1} - \epsilon/4b(b-\epsilon)$。所以，通过设定 $\lambda = (\lambda_1^k + \lambda_2^k)/2$，选择 b_k 相较于其他行动的边际利润下界为 $\epsilon/4b(b-\epsilon) \cdot u_k \ge \epsilon/4b(b-\epsilon) \cdot u_1 \ge \epsilon u_1/4b^2$。因此，我们得到带有 $\varepsilon = \epsilon u_1/4b^2$ 的预言机。

□

E 定理4.1 的证明

定理 *4.1* 的证明。这里我们对算法 1 的 $\widetilde{\mathcal{O}}(K^2T^{-2/3})$ 遗憾进行高层次分析。我们考察 T 轮次中出现的两类事件：$\mathbf{1}(k_t = k_t^*)$ 或 $\mathbf{1}(k_t \neq k_t^*)$。对于第一类事件，根据 S_t 的构造，每轮的次优性差距可以分解为由$\widetilde{}$$(1-\alpha_t)S_t^*$或 $\alpha_t S_{k_t^*}$贡献的若干部分。由 $(1-\alpha_t)S_t^*$ 导致的累积遗憾遵循 UCB 的乐观原则，量级为 $T^{1/2}$。由 $\alpha_t \widetilde{S}_{k_t^*}$导致的累积遗憾量级为 $T^{2/3}$，因为 $\alpha_t = t^{-1/3}$，且 $S_{k_t^*}$ 的次优性差距为常数。另一方面，对于第二类事件，我们可以在最坏情况下将此类事件发生的总轮次数上界为 $\widetilde{\mathcal{O}}(K^2T^{2/3})$。这需要一些较为复杂的论证，不过直观而言，这就是对每个 $k \in [K]$学习 $\|\widehat{q}_k - q_k\| \le t^{-1/3}$ 所需的样本复杂度。对应的最坏情况是这些信号彼此具有竞争性，且对于任意 $k \in [K]$，$n_{k,\,t}$ 以相同速率增长，这意味着我们必须精确学习所有行动的信念分布和成本。因此，根据占主导地位的项，总遗憾为 $\widetilde{\mathcal{O}}(KT^{2/3})$。

我们通过以下两种情况分析遗憾：(i) 代理人以行动 $k_t = k_t^*$作出响应的轮次 t；(ii) 代理人以行动 $k_t \neq k_t^*$作出响应的轮次 t。令 $\mathrm{SubOpt}(k, S_t) = \max_{S\in\mathcal{S}} h(S) - \mathbb{E}_{\omega\sim\sigma,\ \sigma\sim q_k}[u(a^*(\sigma_t)\ \omega_t) - S_t(\sigma_t, \omega_t)]$ 表示在单轮中实施评分规则 S_t 且代理人选择行动 b_k 时的遗憾。对于在线遗憾，我们有：

$$\begin{aligned}\mathrm{Reg}^{\pi}(T) &= \mathbb{E}_{\pi}\left[\sum_{t=1}^{T}\left(\max_{S\in\mathcal{S}} h(S) - \mathbb{E}_{\omega_t\sim\sigma_t,\sigma_t\sim q_{k_t}}\left[u(a^*(\sigma_t),\omega_t) - S_t(\sigma_t,\omega_t)\right]\right)\right] \\ &= \sum_{t=1}^{T}\underbrace{\mathbb{E}\left[\mathbb{1}(k_t = k_t^*)\mathrm{SubOpt}(k_t, S_t)\right]}_{A_t} + \sum_{t=1}^{T}\underbrace{\mathbb{E}\left[\mathbb{1}(k_t \neq k_t^*)\mathrm{SubOpt}(k_t, S_t)\right]}_{B_t}.\end{aligned}$$

在此处及下文中，我们始终使用 $\mathbb{E}$表示关于（2.3）中所述数据生成过程的期望。

$\widetilde{}$遗憾的上界 A_t。由于 $S_t = \alpha_t S_{k_t^*} + (1-\alpha_t)S_t^*$，根据 $\mathrm{SubOpt}(k_t, S_t)$ 的线性性可得

$$\mathrm{SubOpt}(k_t, S_t) = \alpha_t \mathrm{SubOpt}(k_t, \widetilde{S}_{k_t^*}) + (1-\alpha_t)\mathrm{SubOpt}(k_t, S_t^*). \tag{E.1}$$

$\widetilde{}$在此，第一项 $\mathrm{SubOpt}(k_t, S_{k_t^*})$ 被常数 $C_1 = 2(B_S + B_u)$ 所界定，而第二项 $\mathrm{SubOpt}(k_t, S_t^*)$ 被 $2(\|S_t^*\|_\infty + B_u)\, I_{q,\ k_t^*}^t$ 以至少 $1 - 1/t$ 的概率被以下引理所界定。

引理 E.1. 将事件 $\mathcal{E}_t$ 定义为 $\mathcal{E}_t = \left\{ \left\| q_k - \widehat{q}_k^t \right\|_1 \le I_q^t(k) \ \forall k \in [K]^+ \right\}$.。那么，事件 $\mathcal{E}_t$ 以至少 $1 - 1/t$ 的概率成立，并且在事件 $\mathcal{E}_t$ 上有

$$u_{k_t^*} - v_{S_t^*}(k_t^*) + 2\left(B_S + B_u\right) I_q^t(k_t^*) \ge \max_{S \in \mathcal{S}} h(S),$$

其中， $h(S) = \mathbb{E}_{\omega \sim \sigma,\ \sigma \sim q_{k^*(S)}} \left[u(a^*(\sigma)\ \omega) \ - S(\sigma, \omega) \right]$ 是最优委托人的利润， $k_t^* = \mathrm{argmax}_{k \in [K]}\, h_{\mathrm{LP}}^t(k)$, 且 $S_t^* := S_{\mathrm{LP},\ k_t^*}^t$。

证明。详细证明见 §E.1。 □

(E.1) 右侧的第一项可以直接由 $\alpha_t C_1$界定。对于事件 $\mathcal{E}_t$ 上的第二项，根据引理 $C_1 I_q^t(k_t^*)$，可以将其界定为 E.1。若事件 $\mathcal{E}_t$ 不成立，第二项仍可由 C_1 界定，其概率至多为 $1/t$。因此，对于 $\alpha_t = K t^{-1/3} \wedge 1$ ，有

$$\begin{aligned}
\sum_{t=1}^{T} A_t &\le \sum_{t=1}^{T} C_1 \left(\alpha_t + \frac{1}{t} + I_q^t(k_t^*) \right) \\
&\le C_1 \left(K \left(\frac{3}{2} T^{2/3} + 1 \right) + (\log T + 1) + \sum_{k \in [K]^+} \sum_{t : k_t^* = k} \sqrt{\frac{2 \log(K \cdot 2^M t)}{n_k^t}} \right) \\
&\le C_1 \left(K \left(\frac{3}{2} T^{2/3} + 1 \right) + \log T + 1 \right) + C_1 \sqrt{2 \log(K \cdot 2^M T)} \cdot \sum_{k \in [K]^+} \left(2 \sqrt{n_k^T} + 1 \right) \\
&\le \mathcal{O}(T^{2/3}) + \mathcal{O}(\sqrt{|M| + \log(KT)} \cdot \sqrt{KT}), \qquad \text{(E.2)}
\end{aligned}$$

其中，第一个等号由我们此前关于 (E.1) 的讨论得出，第二个不等式根据 I_q^t 的定义成立，最后一个不等式则由詹森不等式 $\sum_{k \in [K]^+} \sqrt{n_k^T / K} \ \le$ $\sqrt{\sum_{k \in [K]^+} n_k^T / K}$.

界定遗憾 B_t。我们通过界定 $\sum_{t=1}^{T} B_t$ 的预期轮次数来刻画遗憾$k_t \ne k_t^*$。为此，我们使用以下引理。

引理 E.2。对于任意固定的 $i \ne k_t^*$，在以下条件下

$$\alpha_t \ge 2\varepsilon^{-1} \left(I_c^t(k_t^*, i) + B_S \left(I_q^t(i) + I_q^t(k_t^*) \right) \right) := \Delta_t(k_t^*, i),$$

~代理人必须在事件 $\mathcal{E}_t$ 上按照评分规则 $S_t = \alpha_t S_{k_t^*} + (1 - \alpha_t) S_t^*$ $k_t \ne i$ 作出响应。

证明。详细证明见 §E.2。 □

在条件 $k_t \ne k_t^*$ 下，该算法必定在进行二分搜索。现在，我们引入用于界定 B_t 的符号。与算法 2 中的定义相比，我们改为将一次二分搜索（BS）定义为 ~ ~BS($S_0, S_1, \lambda_{\min}, \lambda_{\max}, \tau_0, \tau_1, k_0$, k_1, t_0, t_1)，其中 (S_0, S_1) 是该 BS 所针对的初始线段 ~ ~，$(\lambda_{\min}, \lambda_{\max})$ 对应于该 BS 最终得到的 $(\lambda_{\min}, \lambda_{\max})$ 的价值，~ ~ ~ ~τ_0, τ_1 分别对应于 $(1 - \lambda_{\min})$ $S_0 + \lambda_{\min} S_1$ 和 $(1 - \lambda_{\max}) S_0 + \lambda_{\max} S_1$ 被执行的轮次， ~ ~ ~ ~， k_0, k_1 分别是在 $(1 - \lambda_{\min}) S_0 + \lambda_{\min} S_1$ 和 $(1 - \lambda_{\max}) S_0 + \lambda_{\max} S_1$ 下的最佳响应，并且

t_0, t_1 是该 BS 的起始轮次和结束轮次。值得注意的是， k_0 不一定是在 S_0 下的最佳响应 $\tilde{}$ $\tilde{}$，但可以保证它在 S_1 和 $(1-\lambda_{\max})S_0 + \lambda_{\max} S_1$ 下均为最佳响应 k_1。

对于持续到第 t 轮的二分搜索，我们令 $t_0(t)$ 表示二分搜索的第一轮（如果第 t 轮没有进行二分搜索，我们就令 $t_0(t) = t$），$(k_0(t)\ k_1(t))$ 表示该二分搜索结束时得到的最佳响应，$(S_0(t)\ S_1(t))$ 表示该二分搜索所搜索的线段。对于情形 $k_t \neq k_t^*$，我们考虑以下两种情况：(i)$\Delta_{t_0(t)}(k_0(t)\ k_1(t)) \leq \alpha_{t_0(t)}$；(ii) $\Delta_{t_0(t)}(k_0(t)\ k_1(t)) > \alpha_{t_0(t)}$。沿着这一思路，我们得到

$$\begin{aligned}
\sum_{t=1}^{T} B_t &= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(k_t \neq k_t^*)\mathrm{SubOpt}(k_t, S_t)\right] \\
&\leq C_1 \left[\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathrm{BS}=1, \Delta_{t_0(t)}(k_0(t), k_1(t)) \leq \alpha_{t_0(t)})\right] + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathrm{BS}=1, \Delta_{t_0(t)}(k_0(t), k_1(t)) > \alpha_{t_0(t)})\right] \right] \\
&\leq C_1 \Bigg(\underbrace{\sum_{t=1}^{T} \frac{1}{t_0(t)}}_{(B.1)} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathrm{BS}=1, \Delta_{t_0(t)}(k_0(t), k_1(t)) > \alpha_{t_0(t)})\right]}_{(B.2)} \Bigg),
\end{aligned}$$

其中，第一个不等式由 $\mathrm{SubOpt}(k_t, S_t) \leq C_1$ 得出，第二个不等式由引理 E.2 得出。具体而言，进行二分搜索时，所部署的评分规则 S_t 必须位于线段 $(S_0(t)\ S_1(t))$ 上，其中$\sim$ $\sim S_0(t) = \alpha_{t_0}(t) S_{k_{t_0}^*}(t) + (1-\alpha_{t_0}(t)) S_{t_0}^*(t)$，且 $S_1(t) = S_{k_{t_0}^*}(t)$。因此， S_t 可以等价地表示为

$$S_0(t) = \alpha_t' \widetilde{S}_{k_{t_0(t)}^*} + (1-\alpha_t') S_{t_0(t)}^*,$$

其中 $\alpha_t' \geq \alpha_{t_0(t)}$。因此，在 $\Delta_{t_0(t)}(k_0(t)\ k_1(t)) \leq \alpha_{t_0(t)}$ 且代理人在此次二分搜索中作出 $k_0(t) \neq k_1(t)$ 响应的条件下，我们可以使用引理 E.2将概率上界限定为 $1/t_0(t)$。

为了给 $(B.2)$ 设定上界，我们定义一个称为关键二分搜索的概念。

$\tilde{}$ $\tilde{}$定义 E.3（关键二分搜索）。如果二分搜索 $\mathrm{BS}(S_0, S_1, \lambda_{\min}, \lambda_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$ 满足 $\alpha_{t_0} < \Delta_{t_0}(k_0, k_1)$，其中 $\alpha_t = Kt^{-1/3} \wedge 1$，则称其为关键二分搜索（关键 BS）。

以下引理对关键 BS 的总轮次数给出上界。

$\tilde{}$ $\tilde{}$引理 E.4。考虑二分搜索 $\mathrm{BS}(S_0, S_1, \lambda_{\min}, \lambda_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$。假设在 T 轮次中，关键二分搜索的总数为 N。那么，我们有

$$N \leq K \cdot \left(12\varepsilon^{-1} B_S\right)^2 2\log\left(K 2^M T\right) T^{2/3}. \tag{E.3}$$

证明。 详见第§E.3，其中有详细证明。 □

设 m 为二分搜索的最大轮次数，即 $t_0(t) \geq t - m$。根据二分搜索的终止准则，我们有

$$m \leq \max_{t,k} -\log_2\left(I_q^t(k)\right) = \mathcal{O}(\log(T)).$$

项 $(B.1)$ 直接受到.....的界定

$$(B.1) \leq \log(T) + m = \mathcal{O}(\log(T)).$$

项 $(B.2)$ 利用引理E.4,

$$(B.2) \leq m \cdot K \cdot \left(12\varepsilon^{-1}B_S\right)^2 2\log\left(K2^M T\right) T^{2/3} = \mathcal{O}(\varepsilon^{-2}B_S^2(M+\log(KT))mKT^{2/3}。$$

由此完成证明。

□

E.1 引理 E.1 的证明E.1

引理 E.5 （引理 E.1 的重述）。 定义事件 $\mathcal{E}_t$ 为 $\mathcal{E}_t = \left\{\left\|q_k - \widehat{q}_k^t\right\|_1 \leq I_q^t(k) \quad \forall k \in [K]^+\right\}$.则事件 $\mathcal{E}_t$ 以至少 $1-1/t$ 的概率成立，且在事件 $\mathcal{E}_t$ 上满足

$$\langle q_{k_t^*}, u - S_t^* \rangle_\Sigma + 2\left(B_S + B_u\right) I_q^t(k_t^*) \geq \max_{S \in \mathcal{S}} h(S),$$

其中 $h(S) = \mathbb{E}_{\omega\sim\sigma,\sigma\sim q_{k^*(S)}}[u(a^*(\sigma)\ \omega) - S(\sigma,\omega)]$, $v_S(k) = \langle q_k, S\rangle_\Sigma$ 和 $k^*(S)$ 是代理人的最佳响应。

*P*引理 E.1 的证明*E.1*。 我们首先陈述 Mardia 等人（2020年）中的集中性结果。

引理 E.6 (Mardia 等人 (2020年), 引理 1，经验分布的集中性)。 设 p 为一个概率分布，其支撑集是一个有限集合，基数至多为 M，且 p_n 为n个来自 p的独立同分布样本所构成的经验分布。那么，对于任意样本量 $n \in \mathbb{N}_+$ 和 $0 < \delta < 1$,

$$\mathbb{P}\left(\|p - p_n\|_1 \geq \sqrt{\frac{2\log\left(2^M/\delta\right)}{n}}\right) \leq \delta.$$

根据引理 E.6并应用并集界，以至少 $1-1/t$的概率，有 $\left\|\widehat{q}_k^t - q_k\right\|_1 \leq I_q^t(k)$对所有 $k \in [K]^+$均成立，由此得到事件 $\mathcal{E}_t$。下文将在事件 $\mathcal{E}_t$ 成立的条件下进行讨论。我们可以验证，对于 $v_S(k) = \langle q_k,\ S\rangle_\Sigma$ 有

$$\left|\widehat{v}_S^t(k) - v_S(k)\right| = \left|\langle \widehat{q}_k^t - q_k, S\rangle_\Sigma\right| \leq B_S I_q^t(k). \tag{E.4}$$

为了研究 $\widehat{C}^t(i,j)$ 的置信区间，我们引入以下引理。

引理 E.7。在事件 $\mathcal{E}_t = \left\{\left\|q_k - \widehat{q}_k^t\right\|_1 \leq I_q^t(k) \forall k \in [K]^+\right\}$ 下，有

$$\left|\widehat{C}^t(i,j) - C(i,j)\right| \leq I_c^t(i,j), \quad \forall (i,j) \in [K].$$

证明。 (E.4) 的直接推论是

$$\begin{aligned} C_+^t(i,j) &\geq \min_{\tau<t:k_\tau=i} v_{S_\tau}(i) - v_{S_\tau}(j) \geq C(i,j), \\ C_-^t(i,j) &\leq \max_{\tau<t:k_\tau=j} v_{S_\tau}(i) - v_{S_\tau}(j) \leq C(i,j). \end{aligned} \tag{E.5}$$

利用 (E.5) 中的事实，我们有

$$\begin{aligned}\left|\widehat{C}^t(i,j)-C(i,j)\right| &= \left|\sum_{(i',j')\in l_{ij}}\frac{C_+^t(i',j')+C_-^t(i',j')}{2}-C(i',j')\right| \\ &\le \frac{1}{2}\sum_{(i',j')\in l_{ij}}\left(\left|C_+^t(i',j')-C(i',j')\right|+\left|C_-^t(i',j')-C(i',j')\right|\right) \\ &= \frac{1}{2}\sum_{(i',j')\in l_{ij}}\left|C_+^t(i',j')-C_-^t(i',j')\right| = I_c^t(i,j), \end{aligned} \tag{E.6}$$

其中，不等式由三角不等式成立，第二个等号由（E.5）成立，最后一个等号则由 I_C^t 的定义成立。

□

令 k^* 和 S^* 分别为最佳响应和达到最优目标值的最优评分规则$\sup_{S\in\mathcal{S}} h(S)$ （2.2）中的值。回顾线性规划 Opt-LP$_k$,

$$\begin{aligned}\text{Opt-LP}_k: \quad \max_{S\in\mathcal{S}} \quad & \langle u,\widehat{q}_k^t\rangle_\Sigma + B_u I_q^t(k) - v, \\ \text{s.t.} \quad & \left|v-\widehat{v}_S^t(k)\right| \le B_S\cdot I_q^t(k), \\ & v-\widehat{v}_S^t(i) \ge \widehat{C}^t(k,i)-\left(I_c^t(k,i)+B_S\cdot I_q^t(i)\right), \quad \forall i\neq k.\end{aligned}$$

容易验证，利用（E.4）和（E.6），并取 $v=v_{S^*}(k^*)$和 $k=k^*$，S^* 满足 Opt-LP$_{k^*}$ 中的所有约束。此外，$\langle u,\widehat{q}_k^t\rangle_\Sigma + B_u I_q^t(k) \ge \langle u,q_k\rangle_\Sigma$ 对于 $\forall k\in[K]^+$。因此，注意到 $k_t^*=\text{argmax}_{k\in[K]^+} h_{\text{LP}}^t(k)$, 我们得出$h_{\text{LP}}^t(k_t^*)\ge h_{\text{LP}}^t(k^*) \ge h(S^*)$。另外，令$v_{\text{LP}}^*$ 为 Opt-LP$_{k_t^*}$ 的解，并回顾$S_t^*=S_{\text{LP},\,k_t^*}^t$我们有

$$\begin{aligned}h_{\text{LP}}^t(k_t^*) &= \langle u,\widehat{q}_{k_t^*}^t\rangle_\Sigma + B_u I_q^t(k_t^*) - v_{\text{LP}}^* \\ &\le (B_u+B_S) I_q^t(k^*) + \mathbb{E}_{\omega\sim\sigma,\sigma\sim\widehat{q}_{k_t^*}}\left[u(a^*(\sigma),\omega)-S_t^*(\sigma,\omega)\right] \\ &\le 2(B_u+B_S) I_q^t(k^*) + \mathbb{E}_{\omega\sim\sigma,\sigma\sim q_{k_t^*}}\left[u(a^*(\sigma),\omega)-S_t^*(\sigma,\omega)\right],\end{aligned}$$

其中，第一个不等式由注意到$\left|v_{\text{LP}}^*-\widehat{v}_{S_t^*}(k_t^*)\right|\le B_S I_q^t(k_t^*)$ 得出，第二个不等式则根据事件 $\mathcal{E}_t$ 的定义成立。由此，证明完成。

□

E.2 引理证明 E.2

引理 E.8 (引理 E.2 的重述)。 对于 $\forall i\neq k_t^*$，在以下条件下

$$\alpha_t \ge 2\varepsilon^{-1}\left(I_c^t(k_t^*,i)+B_S\left(I_q^t(i)+I_q^t(k_t^*)\right)\right) := \Delta_t(k_t^*,i),$$

代理人 ~在事件 $\mathcal{E}_t$ 发生时，不会在评分规则 $S_t=\alpha_t S_{k_t^*}+(1-\alpha_t)S_t^*$ 下作出响应。

引理 E.2的证明。 在事件 $\mathcal{E}_t$上，对于任意 $i\neq k_t^*$，若 $\alpha_t\ge\Delta_t(k_t^*,\ i)$，我们要证明代理人更倾向于选择~行动 $b_{k_t^*}$而非行动 b_i。回顾 $S_t=\alpha_t S_{k_t^*}+(1-\alpha_t)S_t^*$。代理人采取行动 b_i 所产生的期望利润为

$$v_{S_t}(i)-c_i \le \alpha_t v_{\widetilde{S}_{k_t^*}}(i)+(1-\alpha_t)\widehat{v}_{S_t^*}^t(i)-c_i+(1-\alpha_t)B_S I_q^t(i),$$

而采取行动 $b_{k_t^*}$ 所产生的期望效用为

$$v_{S_t}(k_t^*) - c_{k_t^*} \geq \alpha_t v_{\widetilde{S}_{k_t^*}}(k_t^*) + (1-\alpha_t)\widehat{v}^t_{S_t^*}(k_t^*) - c_{k_t^*} - (1-\alpha_t)B_S I_q^t(k_t^*).$$

根据假设 3.1，我们已经有 $(v_{\widetilde{S}_{k_t^*}}(k_t^*) - c_{k_t^*}) - (v_{\widetilde{S}_{k_t^*}}(i) - c_i) > \varepsilon$。由于 S_t^* 是 Opt-LP 的解 k_t^*，根据 Opt-LP 的约束 k_t^*，我们有

$$\widehat{v}^t_{S_t^*}(k_t^*) - \widehat{v}^t_{S_t^*}(i) \geq \widehat{C}^t(k_t^*, i) - I_c^t(k_t^*, i) - B_S(I_q^t(i) + I_q^t(k_t^*)).$$

结合上述不等式，我们得到

$$\begin{aligned}
\left[v_{S_t}(k_t^*) - c_{k_t^*}\right] - \left[v_{S_t}(i) - c_i\right] &\geq \alpha_t \cdot \varepsilon + (1-\alpha_t)\left(\widehat{C}^t(k_t^*, i) - C(k_t^*, i) - I_c^t(k_t^*, i) - B_S(I_q^t(i) + I_q^t(k_t^*))\right) \\
&\quad - (1-\alpha_t)B_S\left(I_q^t(k_t^*) + I_q^t(i)\right) \\
&\geq \alpha_t \cdot \varepsilon + 2(1-\alpha_t)\left[-I_c^t(k_t^*, i) - B_S(I_q^t(i) + I_q^t(k_t^*))\right] \\
&> \alpha_t \cdot \varepsilon - 2\left[I_c^t(k_t^*, i) + B_S(I_q^t(i) + I_q^t(k_t^*))\right],
\end{aligned}$$

其中，第二个不等式成立，是因为回顾我们已在 $\mathcal{E}_t$ 条件下于 (E.6) 中证明了 $C(k_t^*, i) \leq \widehat{C}^t(k_t^*, i) + I_c^t(k_t^*, i)$；最后一个不等式成立，是因为注意到 $0 < \alpha_t \leq 1$。当 $\alpha_t \geq 2\varepsilon^{-1}\left(I_c^t(k_t^*, i) + B_S\left(I_q^t(i) + I_q^t(k_t^*)\right)\right)$ 时，上述 $\left[v_{S_t}(k_t^*) - c_{k_t^*}\right] - \left[v_{S_t}(i) - c_i\right] > 0$，这意味着代理人更偏好行动 k_t^* 而非 i。

□

E.3 引理证明 E.4

接下来，命题 E.9表明，必要的二分搜索会缩小置信区间，而命题 E.10 给出了必要二分搜索的停止准则。结合命题 E.9 和命题 E.10，我们可以界定必要二分搜索的总轮次数。

Pr ~ ~命题 E.9。对于二分搜索 BS($S_0, S_1, \lambda_{\min}, \lambda_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1$)，我们有 e

$$I_c^t(k_0, k_1) \leq 2\sqrt{2\log(K \cdot 2^M T)}B_S\left(\frac{1}{\sqrt{n_{k_0}^{t_1}}} + \frac{1}{\sqrt{n_{k_1}^{t_1}}}\right). \tag{E.7}$$

命题证明 *E.9*。根据 $I_c^t(k_0, k_1)$的定义，对于任意 $t > t_1$，

$$I_c^t(k_0, k_1) = \sum_{(i',j') \in l_{k_0 k_1}} \varphi^t(i', j') \leq \varphi^t(k_0, k_1),$$

其中，不等式成立是因为 $l_{k_0 k_1}$ 是“最短路径”，而 φ^t 是每条边的“长度”。根据 φ^t 的定义，我们得出

$$\begin{aligned}
I_c^t(k_0, k_1) &\leq \frac{1}{2}\left[C_+^t(k_0, k_1) - C_-^t(k_0, k_1)\right]_+ \\
&= \frac{1}{2}\left[\min_{\tau' < t : k_{\tau'} = k_0}\left(\widehat{v}^t_{S_{\tau'}}(k_0) - \widehat{v}^t_{S_{\tau'}}(k_1)\right) - \max_{\tau' < t : k_{\tau'} = k_0}\left(\widehat{v}^t_{S_{\tau'}}(k_0) - \widehat{v}^t_{S_{\tau'}}(k_1)\right)\right]_+ \\
&\qquad + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) \\
&\leq \frac{1}{2}\left[\widehat{v}^t_{S_{\tau_0}}(k_0) - \widehat{v}^t_{S_{\tau_0}}(k_1) - \widehat{v}^t_{S_{\tau_1}}(k_0) + \widehat{v}^t_{S_{\tau_1}}(k_1)\right]_+ + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right),
\end{aligned}$$

其中，最后一个不等式成立是因为 $t > t_1 \geq (\tau_0, \tau_1)$。因此，根据 $\widehat{v}_S^t$ 的定义，我们有：

$$\begin{aligned} I_c^t(k_0,k_1) &\leq \frac{1}{2}\left|\langle S_{\tau_0} - S_{\tau_1}, \widehat{q}_{k_0}^t - \widehat{q}_{k_1}^t\rangle_\Sigma\right| + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) \\ &= \frac{(\widetilde{\lambda}_{\max} - \widetilde{\lambda}_{\min})}{2}\left|\langle S_1 - S_0, \widehat{q}_{k_0}^t - \widehat{q}_{k_1}^t\rangle_\Sigma\right| + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) \\ &\leq \frac{\min\left(I_q^{t_1}(k_0), I_q^{t_1}(k_1)\right)}{2}\left(\left|\langle S_1 - S_0, \widehat{q}_{k_0}^t\rangle_\Sigma\right| + \left|\langle S_1 - S_0, \widehat{q}_{k_1}^t\rangle_\Sigma\right|\right) + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right), \end{aligned}$$

其中，$\widetilde{\lambda}_{\max}$ 和 $\widetilde{\lambda}_{\min}$ 分别是本次二分搜索终止时 $\lambda_{\max}$ 和 $\lambda_{\min}$ 的值。这里，最后一个不等式由二分搜索的终止准则得出。此外，我们可以将右侧上界估计为

$$\begin{aligned} I_c^t(k_0,k_1) &\leq \min\left(I_q^{t_1}(k_0), I_q^{t_1}(k_1)\right) B_S + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) \\ &\leq \left[\sqrt{2\log(K\cdot 2^M T)}\cdot\left(\frac{1}{\sqrt{n_{k_0}^{t_1}}} + \frac{1}{\sqrt{n_{k_1}^{t_1}}} + \frac{1}{\sqrt{\max\left\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\right\}}}\right)\right] B_S \\ &\leq 2B_S\left[\sqrt{2\log(K\cdot 2^M T)}\cdot\left(\frac{1}{\sqrt{n_{k_0}^{t_1}}} + \frac{1}{\sqrt{n_{k_1}^{t_1}}}\right)\right], \end{aligned}$$

其中，第一个不等式由以下事实得出：$\widehat{q}_k^t$ 是一个经验概率，且由于 S 非负，$\|S_1 - S_0\|_\infty \leq B_S$；第二个不等式则由以下事实得出：$n_k^t$ 单调不减，且 $t_1 < t$。

□

命题 E.10. 对于二分搜索 BS($S_0, S_1, \lambda_{\min}, \lambda_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1$)，若

$$\min\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\} \geq \left(12\varepsilon^{-1}B_S\right)^2 2\log\left(K2^M T\right) T^{2/3} =: N^*, \tag{E.8}$$

则对于任意 $t_1 < t \leq T$，都有 $t^{-1/3} \geq \Delta_t(k_0,k_1)$，且不可能再有以相同的$(k_0,k_1)$结束的更本质的二分搜索。

命题证明 *E.10*。对于任意 $t_1 < t \leq T$，我们有

$$\begin{aligned} t^{-1/3} \geq T^{-1/3} &= \left(12\varepsilon^{-1}B_S\right)\sqrt{\frac{2\log\left(K2^M T\right)}{N^*}} \\ &\geq 2\left(\varepsilon^{-1}B_S\right)\sqrt{2\log\left(K2^M T\right)}\left(2\frac{1}{\sqrt{n_{k_0}^{t_1}}} + 2\frac{1}{\sqrt{n_{k_1}^{t_1}}} + \frac{1}{\sqrt{n_{k_0}^{t}}} + \frac{1}{\sqrt{n_{k_1}^{t}}}\right) \\ &\geq 2\varepsilon^{-1}\left(B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) + I_c^t(k_0,k_1)\right) = \Delta_t(k_0,k_1), \end{aligned}$$

其中等号根据 N^*的定义成立，第二个不等式根据我们的条件最小值$\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\} \geq N^*$成立，最后一个不等式则利用命题E.9成立。因此，我们得出结论：令 $\alpha_t = Kt^{-1/3} \wedge 1 \geq t^{-1/3}$后，本质二分搜索的条件，即 $\alpha_t < \Delta_t(k_0,k_1)$ 对任何满足 t 的 $t_1 < t \leq T$都不再成立。

□

现在，我们准备好证明引理E.4。

~ ~引理 E.11 （引理 E.4的重述）。 考虑二分搜索 BS（S_0， S_1， $\lambda_{\min}$， $\lambda_{\max}$， τ_0， τ_1， k_0， k_1， t_0， t_1）。假设在 T轮次中，关键二分搜索的总数为 N。那么，我们有

$$N \leq \left(12\varepsilon^{-1}B_S\right)^2 2\log\left(K2^M T\right) KT^{2/3} = KN^*. \tag{E.9}$$

引理的证明 *E.4*。 为证明引理 E.4，我们首先引入另一个概念，称为 临界二分搜索（临界 BS）。

~ ~定义 E.12（临界二分搜索）。对于关键二分搜索 BS($S_0, S_1, \lambda_{\min}, \lambda_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1$)，若 $\min\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\} \geq N^*$，则称其为临界的。

临界基站。我们声称，临界 BS 的数量不超过 K。假设相反，即临界 BS 的数量不少于 $K+1$。考虑图 $\mathcal{G} = (V, E)$，其中 $V = [K]^*$ 和 E 由以 $(k_0, k_1) \in V$结尾的临界基站集合构成。我们声称，以下两种情况必有其一成立

(i) 存在 $(i, j) \in [K]^+$，使得至少有两个以 $(k_0, k_1) = (i, j)$ 结尾的临界基站；

(ii) 由关键基站构成的 $\mathcal{G}$ 中存在一个环。

我们证明 (i) 和 (ii) 都不可能发生。对于 (i)，根据关键基站的定义，$\min\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\} \geq N^*$，其中 t_1 是关键基站结束的轮次；再根据命题 E.10，在 t_1 之后，不应再有以 (k_0, k_1) 结束的关键 BS，因此我们可以直接得出 (i) 不可能。对于 (ii)， ~ ~假设 BS($S_0, S_1, \lambda_{\min}, \lambda_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1$) 是第一个形成环的基站。那么在 t_0 之前，必定存在一条由关键基站构成的路径 $\ell'_{k_0 k_1}$。对于 $\ell'_{k_0 k_1}$ 上的任意一条边 (k'_0, k'_1)，必定存在一个满足 $t'_1 < t_0$ 的临界二分搜索~ ~BS($S'_0, S'_1, \lambda'_{\min}, \lambda'_{\max}, \tau'_0, \tau'_1, k'_0, k'_1, t'_0, t'_1$)。再次使用命题 E.10，我们得到

$$t_0^{-1/3} \geq \Delta_{t_0}(k'_0, k'_1), \quad \forall (k'_0, k'_1) \in \ell'_{k_0, k_1}.$$

由于 $\mathcal{G}$ 中的一条路径长度至多为 K，我们得出

$$\begin{aligned}
\alpha_{t_0} = K t_0^{-1/3} &\geq \sum_{(k'_0, k'_1) \in \ell'_{k_0 k_1}} \Delta_{t_0}(k'_0, k'_1) \\
&= \sum_{(k'_0, k'_1) \in \ell'_{k_0 k_1}} 2\varepsilon^{-1}\left(B_S\left(I_q^{t_0}(k'_0) + I_q^{t_0}(k'_1)\right) + I_c^{t_0}(k'_0, k'_1)\right) \\
&\geq 2\varepsilon^{-1}\left(B_S\left(I_q^{t_0}(k_0) + I_q^{t_0}(k_1)\right) + I_c^{t_0}(k_0, k_1)\right) = \Delta_{t_0}(k_0, k_1),
\end{aligned}$$

其中，最后一个不等式成立，是因为注意到 $I_c^{t_0}(k'_0, k'_1) \leq \varphi^{t_0}(k'_0, k'_1)$，且 $I_c^{t_0}$是通过边长为 φ^{t_0}的最短路径算法得到的。这一事实再次表明，不可能存在一条以 (k_0, k_1) 结尾且从 t_0出发的本质 BR。因此，也不可能形成循环，从而证明了我们的结论：关键基站的总数不超过 K。

重要但非关键基站。接下来，我们统计重要但非关键基站的数量。回顾图 $\mathcal{G} = (V, E)$，其中 $V = [K]^*$ 和 E 由以 $(k_0, k_1) \in V$ 结尾的关键基站集合构成。令 $w^t(i, j)$ 表示第 t 轮之前，以 $(k_0, k_1) = (i, j)$ 结尾的重要基站总数，并

令 $w(i,j) = w^T(i,j)$。令 $d_k^t = \sum_{k' \neq k} w(k,k')$ 表示节点 $k \in V$ 的总权重。对于任意一对 $(k_0, k_1) \in V$，如果 $d_{k_1}^t \wedge d_{k_2}^t \geq N^*$，根据命题 E.10，在 t 之后就不应再有重要基站。另一方面，任何以 (k_0, k_1) 结尾的基站都能保证 $d_{k_0}^t$ 和 $d_{k_1}^t$ 至少增加 1。因此，统计重要但非关键基站数量的问题可以描述为以下权重放置游戏，

(i) 为 $\forall k \in V$ 初始化 $d_k = 0$；

(ii) 在每一轮中，选择一条边 $(k_0, k_1) \in V$，使得 $d_{k_0} \wedge d_{k_1} < N^*$，并将 d_{k_0} 和 d_{k_1} 都加 1；

(iii) 若无法再选择边，则游戏结束。

该权重放置游戏的总轮次数为重要但非关键基站总数的上界。以下命题给出了该权重放置游戏的最大轮次数。

命题 E.13（权重放置游戏）。对于上述权重放置游戏，总轮次数不应超过 $(|V|-1)(N^*-1)$。

证明。我们通过对 $|V|$ 进行归纳法来证明这个命题。$|V| = 1$ 的情况显然成立。对于 $|V| = 2$，结果显而易见，因为只有一条边可供选择。假设结果对 $|V| \leq v-1$ 成立。对于 $|V| = v$，假设进行了 $(v-1)(N^*-1) + 1$ 轮次。我们研究量 $d^- = \min_{k \in V} d_k$。

- 如果存在 $k \in V$ 使得 $d_k \leq N^* - 2$，那么对于子图 $\mathcal{G}^k = \mathcal{G} \backslash \{k\}$，我们有至少 $(v-2)(N^*-1) + 2$ 轮次在 $\mathcal{G}^k$ 上进行。然而，由于 $\mathcal{G}^k$ 只有 $v-1$ 个节点，根据归纳法，在 $\mathcal{G}^k$ 上进行的轮次不应超过 $(v-2)(N^*-1)$，这就产生了矛盾。因此，如果进行了 $(v-1)(N^*-1) + 1$ 轮次，我们有 $d^- \geq N^* - 1$。
- 如果 $d^- \geq N^*$，那么这一局的最后一轮必定是错误地按照规则 (ii) 和 (iii) 进行的。

因此，我们得出结论：$d^- = N^* - 1$，并且能够选择一个节点 $k \in V$，使得 $d_k = N^* - 1$。根据归纳法，在子图 $\mathcal{G}^k = \mathcal{G} \backslash \{k\}$ 上，最多可以进行 $(v-2)(N^*-1)$ 轮。因此，在 $\mathcal{G}$ 上进行的总轮数不超过

$$(v-2)(N^*-1) + d_k = (v-1)(N^*-1),$$

这与我们的假设矛盾：进行的轮数为 $(v-1)(N^*-1) + 1$。因此，对于 $|V| = v$，最多可以进行 $(v-1)(N^*-1)$ 轮。根据归纳法，我们得出结论：该命题对任意$|V|$均成立。

□

因此，重要但非关键基站的总数不超过 $K(N^*-1)$。鉴于关键基站的数量不超过 K，关键 BS 的总数因而不超过 KN^*。 □