arXiv:2601.22079v1 [econ.TH] 29 Jan 2026

无悔学习算法经济学 *

Jason Hartline

摘要

现代经济学面临的一项根本挑战，是理解当经济中的行为者被算法取代时会发生什么。正如理性有助于理解传统经济行为者的结果，无后悔也能帮助我们理解算法行为者所产生的结果。本文综述计算机科学中关于无后悔算法的经典文献，为概述经济学领域关于无后悔算法的最新研究奠定基础，重点关注操纵、统计推断和算法合谋等新兴主题。

关键词：无后悔学习、多臂老虎机、交换遗憾、相关均衡、计量经济学、斯塔克尔伯格均衡、操纵、算法合谋。

1 引言

一切皆是算法。从宏观上看，算法接收输入、对其进行处理，并生成输出。对决策者而言，输入可能是偏好、信号和信念，输出则可能是行动。对经济体而言，输入是个人的行动，以及供给等状态变量的实现值；输出则是按照经济体聚合规则产生的结果。企业和个人可以被视为通过算法采取行动。例如，该算法可以是贝叶斯—纳什最佳回应：根据信号进行贝叶斯更新，计算他人的均衡策略，并根据该均衡作出最佳回应。更进一步，如今个人和企业越来越常直接使用计算机算法来做出这些决策。或许在重复博弈中，计算机算法会根据历史行动和收益接收输入，并反复做出下一步决策。例如，Airbnb、eBay、亚马逊和 Google Ads 等在线市场中的卖家都在使用算法。许多此类市场甚至会向参与者提供算法作为产品。当然，市场本身的规则也由算法支配。

理论计算机科学中的算法领域，旨在从“一切皆为算法”的视角理解世界。对于上述经济场景，这意味着要融合经济学中的博弈论、机制设计和信息设计等核心概念，并探究算法理论如何带来新的

*感谢 Michihiro Kandori 和计量经济学会邀请本文作为综述文章发表于2025年计量经济学会世界大会。感谢世界大会的与会者提出富有见地的讨论和问题。感谢 Nicole Immorlica 讨论本文并提供宝贵建议。本文综述文章将发表于 *Advances in Economics and Econometrics: Thirteenth World Congress, Volume 2*。

视角与理解。2000年代中期的一项经典成果，研究了纳什均衡的算法复杂性。[1] 在算法复杂性的标准假设下，对于一般的双人双矩阵博弈，不存在能在合理时间内求出每个此类博弈纳什均衡的算法。因此，尽管所有这类博弈都保证存在纳什均衡——这一关键性质使其在经济理论中得到广泛应用——算法却通常无法找到这些均衡。如果算法找不到，个人也无法找到，无论他们是否彼此互动。毕竟，个人和个人组成的群体也都是算法（请记住，算法通过某个过程将输入映射为输出）。

经济学与计算机科学交叉领域的首要学术会议，是美国计算机协会每年举办的经济学与计算研讨会（EC）。计算机科学、运筹学和经济学领域的顶尖论文，常常在期刊发表之前先在 EC 上发表。要在一篇《计量经济学会世界大会综述文章》的篇幅内，充分概述这一交叉领域的所有研究活动，几乎是不可能的。EC 2025 的征稿启事列出了以下主题（按字母顺序排列）：基于代理人的建模、算法公平与数据隐私、**拍卖与定价**、**行为经济学与有限理性**、区块链与加密货币、合同设计、众包与信息获取、决策理论、计量经济学、**学习算法的经济学层面**、神经网络与大语言模型的经济学层面、**均衡分析与无政府代价**、**均衡计算与复杂性**、公平分配、产业组织、信息设计、实验室实验与实地实验、市场设计与匹配市场、市场均衡、机制设计、**在线算法**、**在线平台与应用**、社会公益与伦理学、社会选择与投票理论，以及社交网络与社会学习。本文将聚焦于一个主题——无悔学习算法的经济学——并对其进行较为全面的综述。上述主题中包含这一主题的部分以粗体标出。

相关均衡（Aumann，1974）是一种易于处理的均衡概念，并且对应于重复博弈中简单的“无遗憾”学习算法的动态过程。相关均衡是纳什均衡的一种典型放宽形式。纳什均衡要求玩家所选择行动的分布相互独立，即博弈行为的联合分布是外积，也就是一个秩为一的矩阵；而相关均衡允许任意联合分布，并要求每位玩家的每个行动相对于该联合分布都是条件最佳回应。在纳什均衡难以求解的同类双矩阵博弈中，相关均衡可以通过线性规划轻松计算。此外，Foster 和 Vohra（1997年）以及 Hart 和 Mas-Colell（2000年）表明，存在一些简单的学习算法，能够通过重复交互达到相关均衡。事实上，计算机科学领域关于在线学习（即从重复交互中学习）的大量文献中，许多标准的所谓“无遗憾”算法，尽管并非专为策略情境设计，却能使博弈行为的经验分布收敛到粗相关均衡（这是对相关均衡的进一步放宽形式；Moulin 和 Vial，1978年）。无遗憾——也就是在进行一段时间的博弈后，玩家不会后悔自己采取的行动，相较于基准策略中的行动而言——是一种较宽松且容易实现的理性概念（可通过算法实现；Hannan，1957）。从某种意义上说，它是理性的下界，因为理性的最佳回应也满足无遗憾。因此，如今计算机科学文献普遍认为，阶段博弈中的相关均衡和粗相关均衡概念，分别对应于策略互动中在计算和信息方面易于处理的结果。

[1]这项工作荣获博弈论学会颁发的2008年计算机科学与博弈论卡莱奖。更多讨论请参见第5节。

本文将回顾相关均衡与粗相关均衡的基本定义、在重复交互中实现这些均衡的经典算法，以及由这些概念引出的若干经济学新结论。具体而言，遗憾有两种典型表述：事后最优遗憾和交换遗憾。对于这些概念，本文综述将重点介绍以下结果：

- 交换遗憾算法会产生相关均衡（Foster 和 Vohra，1997；Hart 和 Mas-Colell，2000），而事后最优遗憾算法会产生粗相关均衡（例如 Young，2004）：

- 存在一种归约，可将任意无事后最优遗憾算法转换为无交换遗憾算法（Blum 和 Mansour，2007年）。

- 这些概念在斯塔克尔伯格意义下对操纵的稳健性不同。首先，面对静态斯塔克尔伯格领导者，任何合理的学习算法都会学会作出最佳回应，从而使斯塔克尔伯格领导者和跟随者（该算法）获得其斯塔克尔伯格均衡收益。其次，无事后最优遗憾算法可以被操纵到超过斯塔克尔伯格收益的程度（从而提高领导者收益），而无换位后悔算法则保证领导者的斯塔克尔伯格收益不会得到改善（Braverman 等人，2018年；Deng 等人，2019年）。

- 将纳什均衡放宽为无后悔或低后悔结果，可以采用结构推断方法，识别出与实证博弈行为一致的偏好和遗憾的可理性化集合（Nekipelov等人，2015年）。

- 算法合谋（例如 Calvano 等人，2020年）以及如何监管用于合谋的算法，都可以从无后悔保证和无交换遗憾算法的不可操纵性角度加以理解（Hartline 等人，2024年、2025年）。

本篇关于无悔学习算法的概述介绍了最基本的方法和结果，并将其与当前计算机科学和经济学交叉领域的前沿研究联系起来。本文的分析侧重于建立主要直觉，并给出最基本结果的完整证明。

2 在线学习

在线学习研究的是如何在未知且可能具有对抗性的环境中进行序贯决策，目标是在事后表现得几乎与最佳固定动作一样好。经典的在线学习模型是无状态的，即行动只影响当前轮次的收益，而不会影响未来轮次的收益。在每一 n 轮中，决策者从 k 个可能的行动中选择一个，获得取决于所选行动的收益，然后观察所有行动的收益。目标是随着轮数增加，使算法的表现接近事后最佳行动。

“在线”一词源自在线算法 领域，指的是一类算法问题：算法必须在所有输入都揭晓之前做出决策。

2.1 模型

设想一位学习者在 n 轮次中反复从 k 行动 $a \in A$ 中选择。令 $u_a^i \in [0,h]$ 表示第 i轮次中行动 a的收益。在每一轮 i:

1. 学习者选择一个行动 a^i，2. 观察所有收益 $\mathbf{u}^i = (u_1^i, \ldots, u_k^i)$，并且 3. 获得所选行动 $u_{a^i}^i$ 的收益。

该算法的总收益为

$$\mathrm{ALG}_n = \sum_{i=1}^n u_{a^i}^i,$$

而事后最佳固定动作的总收益为

$$\mathrm{OPT}_n = \max_a \sum_{i=1}^n u_a^i.$$

经过 n 轮次后的每轮遗憾定义为

$$\mathrm{Regret}_n = \frac{1}{n}[\mathrm{OPT}_n - \mathrm{ALG}_n] = \frac{1}{n}\left[\max_a \sum_{i=1}^n u_a^i - \sum_{i=1}^n u_{a^i}^i\right].$$

如果对于所有输入序列 $\mathbf{u}^1, \ldots, \mathbf{u}^n$，遗憾满足 $\lim_{n\to\infty} \mathrm{Regret}_n = 0$，则该算法具有消失遗憾。

该框架假设处于最坏情况（对抗性）环境中，这一点之后将有助于从更复杂的学习模型进行归约。对抗模型与独立同分布随机模型形成对照：在后者中，每轮的收益都是从某个未知分布中独立同分布地抽取的。在尝试为更具挑战性的学习问题开发算法时，对在线学习进行对抗性分析尤其有用。具体而言，接下来的两节将通过归约到本节已解决的在线学习问题，为这类问题提供算法。

2.2 跟随领导者，FTL

首先，我们看到，在对抗性环境中，自然的跟随领导者（FTL）算法会彻底失效。

定义 2.1（跟随领导者，FTL）。令 $U_a^i = \sum_{i'=1}^i u_a^{i'}$ 表示行动a 截至第 i 轮的累积收益。在每一轮 i，算法选择 $a^i = \mathrm{argmax}_a U_a^{i-1}$.

事实上，在独立同分布随机环境中，跟随领导者算法是最优的（Hartline 等，2020）。我们关注对抗模型，是因为有些应用无法用独立同分布模型描述，而在这些应用中，该算法的失效影响重大。跟随领导者算法是虚拟博弈的单玩家版本。它也与无状态 Q 学习算法有关；近期，经济学文献在研究算法合谋时对该算法进行了探讨。

下面的例子表明，该算法会失效。此外，正如接下来将要证明的，这种问题并非只存在于该算法，而是任何确定性算法都会遇到的。从本质上说，要在对抗性的在线输入下实现良好的学习，就需要随机化。

例 2.1. 考虑 $k = 2$ 个行动和收益的情形，并且不失一般性地假设，在第一轮中FTL选择行动 *1*。

Round	*1*	*2*	*3*	*4*	*5*	*6*	...
Action 1	**0**	*1*	**0**	*1*	**0**	*1*	...
Action 2	*1/2*	**0**	*1*	**0**	*1*	**0**	...

在这个序列中，$\mathrm{OPT}_n \approx n/2$ 可以选择任一行动，而$\mathrm{FTL}_n = 0$ 则选择加粗的行动。因此，FTL 的每轮遗憾为 $\frac{1}{n}[\mathrm{OPT}_n - \mathrm{FTL}_n] \approx 1/2$，且不会随$n$消失。

定理 2.1. 在最坏情况下，所有确定性在线学习算法的每轮遗憾都是常数。

证明. 由于该算法是确定性的，自然可以在每一轮采取对抗性选择：将算法所选行动的收益设为 0，将所有其他行动的收益设为 1。显然，在这种选择下，$\mathrm{ALG}_n = 0$ ，而 $\mathrm{OPT} \geq n/2$ ，因为必定存在某个行动，其未被算法选中的次数超过一半。每轮遗憾至少为 $1/2$。 □

这一不可能性结果表明，在对抗模型中，要实现消失遗憾，随机化至关重要。

2.3 学习算法

一种自然的方法是用软最大值替代最大值：对过去收益较高的行动，提高其被选中的概率（反之，对过去收益较低的行动降低其被选中的概率）。由此得到的算法称为指数加权、乘法权重更新和 *Hedge*，最早由 Littlestone 和 Warmuth（1994年）以及 Freund 和 Schapire（1997）提出。

定义 2.2 （指数加权，EW）。令 $\epsilon > 0$ 表示学习率。在轮次 i 中，定义$U_a^{i-1} = \sum_{i'=1}^{i-1} u_a^{i'}$ 并以概率选择行动 a

$$\alpha_a^i = \frac{(1+\epsilon)^{U_a^{i-1}/h}}{\sum_{j'}(1+\epsilon)^{U_{j'}^{i-1}/h}}.$$

示例 2.2。对于 $\epsilon = 1$ 和 $u_a^i \in \{0, 1\}$，每当某个行动获得1的收益时，指数加权就会将该行动的权重加倍。对于以下收益（左），指数加权按如下方式计算权重（中）和行动选择概率（右）：

轮次 i	*1*	*2*	*3*	*4*
u_1^i	*1*	*1*	*0*	*0*
u_2^i	*0*	*0*	*1*	*1*

轮次 i	*1*	*2*	*3*	*4*
$2^{U_1^{i-1}}$	*1*	*2*	*4*	*4*
$2^{U_2^{i-1}}$	*1*	*1*	*1*	*2*

轮次 i	*1*	*2*	*3*	*4*
α_1^i	*1/2*	*2/3*	*4/5*	*2/3*
α_2^i	*1/2*	*1/3*	*1/5*	*1/3*

指数加权的总期望收益由所有价值与概率的点积给出。本例中： $1/2 + 2/3 + 1/5 + 1/3 = 1.7$。

以下定理和推论表明，指数加权具有趋于零的事后最优遗憾。

定理 2.2。 对于 $[0,h]$ 收益、 k 个行动以及学习率 ϵ:

$$\mathbb{E}[\mathrm{EW}] \geq (1-\epsilon)\,\mathrm{OPT} - \frac{h}{\epsilon}\ln k.$$

较小的 ϵ 会减慢对数据的适应，使该算法对对抗性噪声更为稳定，但学习良好行动的速度更慢；较大的 ϵ 能带来快速调整，但对对抗性噪声的敏感性更高。通过比较指数加权在例 2.1 的极端输入上的表现，以及在由所有轮次 i的常数输入$\mathbf{u}^i = (1,0)$所给出的极端输入上的表现，可以更好地理解这一权衡。

推论 2.1。 对于收益位于 $[0,h]$中的 n 轮次，选择 $\epsilon = \sqrt{(\ln k)/n}$ 可得

$$\mathbb{E}[\mathrm{Regret}(\mathrm{EW})] \leq 2h\sqrt{\frac{\ln k}{n}}.$$

证明。 由于 $\mathrm{OPT} \leq hn$，定理 2.2 给出 $\mathbb{E}[\mathrm{EW}] \geq \mathrm{OPT} - \epsilon hn - \frac{h}{\epsilon}\ln k$。 通过 $\epsilon hn = \frac{h}{\epsilon}\ln k$ 平衡误差项，可得 $\epsilon = \sqrt{(\ln k)/n}$ 以及界 $2h\sqrt{(\ln k)/n}$。 □

由推论 2.1 可见，当轮数 n 较大时，较慢的学习率是有益的；而当行动数 k 较大时，较快的学习率是有益的。定理 2.2 的证明来自代数操纵与不等式；这些推导虽简单，却不能提供清晰的直觉。本节的其余部分将介绍另一种能提供更多直觉的算法。

2.4 一个上界：BetheLeader

开发近似最优算法的一种标准方法是：(a) 确定一个为最优表现提供上界的基准，以及 (b) 遵循该基准的直觉，开发一种表现可证明相近的算法。在以下定义中，Be the Leader（BTL）是一个基准，而非可实现的算法。

定义 2.3（Be the Leader，BTL）。 *Be the Leader* 基准在每一轮次 i 选择在观察到轮次 i 之后 本应最优的行动： $a^i = \mathrm{argmax}_a\, U_a^i$。

随后的例子和定理表明，BTL 为事后最优表现 OPT 给出了上界。见图 2.1。

例 2.3. 对于两个具有收益的行动

	1	2	3	4
Action 1	0.4	0.3	0.3	1.0
Action 2	0.2	0.2	1.0	0.2

每轮之后的累计表现为$(0.4, 0.7, 1.7, 2.7)$（对应BTL），而$(0.4, 0.7, 1.4, 2.0)$（对应OPT）。BTL 在第 *1*、*2* 和 *4* 轮采取行动 *1*，并获得这些轮次的全部收益，记作btl^i，对应第 i 轮。OPT 在这些轮次切换为跟随领导者，但只能获得累计表现的边际增量，记作opt^i，对应第 i轮。

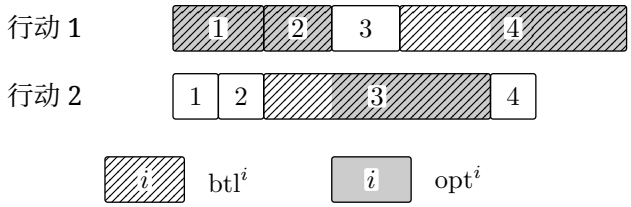


图 2.1：例 2.3 和定理 2.3 的示意图。每个行动的收益以按轮次编号的矩形宽度表示。在每一轮 i 中，BTL 选择的收益 btl^i 用条纹表示，OPT 累积的收益 opt^i 以灰色阴影表示，并且有 $\mathrm{btl}^i \geq \mathrm{opt}^i$。

定理 2.3。BTL $\geq$OPT。

证明。 令 OPT^i 表示经过 i 轮后的事后最优值，并令 opt^i = 表示 OPT^i –与 OPT^{i-1}之间的差值。也就是说，opt^i 等于第 i轮前后领导者累积收益的变化。类似地，定义 BTL^i 和 btl^i，其中 btl^i 是领导者在第 i 轮获得的全部收益。因此，对所有 i，$\mathrm{btl}^i \geq$ 不小于 opt^i。对所有轮次求和可得 BTL $\geq$不小于 OPT。

□

直观地说，BTL 在在线学习模型中无法实现，因为它需要预见能力；而 FTL 虽然可以实现，但由于对抗性噪声，其表现可能很差。为了衔接二者，我们引入随机扰动，以平滑噪声。

2.5 扰动式跟随领导者，PFTL

现在，我们设计一种算法来逼近无法达到的 BTL 上界。扰动跟随领导者 （PFTL）算法通过随机化初始价值来打破平局并稳定学习。此处的分析出自 Kalai 和 Vempala（2005年）。

定义 2.4 （扰动跟随领导者，PFTL）。 固定学习率 $\epsilon > 0$。对于每个行动a，设想一个初始收益 $u_a^0 = h \times Z_a$，其中 Z_a 是通过抛掷一枚偏向 ϵ的硬币得到的连续反面次数（一个几何随机变量）。令 $U_a^i = \sum_{i'=1}^{i} u_a^{i'}$。在每一轮 i，选择

$$a^i = \mathrm{argmax}_a \big(u_a^0 + U_a^{i-1}\big).$$

例 2.4。 用PFTL重新考察例 *2.1*：

	0	*1*	*2*	*3*	*4*	*5*	*6*	...
Action 1	*2*	*0*	*1*	*0*	*1*	*0*	*1*	...
Action 2	*3*	**1/2**	**0**	**1**	**0**	**1**	**0**	...

粗体显示的收益是PFTL在所示的第 *0* 轮扰动下选出的。对扰动取平均后，PFTL 大约有一半时间选择粗体收益，另一半时间选择其他收益；其平均收益约为 $n/2$。因此，PFTL $\approx n/2$。当然，OPT $\approx n/2$ 也一样。

定理 2.4。 对于 $[0, h]$ 中的收益，

$$\mathbb{E}[\mathrm{PFTL}] \geq (1-\epsilon)\,\mathrm{OPT} - \frac{h}{\epsilon}(1 + \ln k).$$

推论 2.2。 对于 n 轮次和 $[0,h]$ 中的收益，调整 $\epsilon = \sqrt{(1+\ln k)/n}$ 可得

$$\mathbb{E}[\text{Regret}(\text{PFTL})] \le 2h\sqrt{\frac{1+\ln k}{n}}.$$

证明。 论证与推论 2.1 相同。 □

许多分布都可用于扰动 PFTL 中的得分；[2] 几何分布尤其方便，因为它具有无记忆性。具体而言，对于任意整数 ℓ，几何随机变量 X 满足 $\mathbb{E}[X \mid X \ge \ell] = \ell + \mathbb{E}[X]$。

引理 2.1 （几何变量的最大值，）. 设 X_1，……， X_k 为参数为 ϵ 的独立几何随机变量。则 $\mathbb{E}[\max_i X_i] \le \frac{1+\ln k}{\epsilon}$。

证明。 若我们在并行轮次中同时抛掷这 k 个几何分布各自对应的硬币，则在每一轮次中，预期有 ϵ 比例的 k 枚硬币成功（正面；且剩余 $1-\epsilon$ 比例）。直到至多剩下一枚为止的期望轮数至多为 $\log_{1/(1-\epsilon)} k$——依据归纳法与 Jensen 不等式——而最后一枚的期望持续 $1/\epsilon$ 轮次，总计为 $\frac{1+\ln k}{\epsilon}$。 □

2.6 PFTL 的分析

分析 PFTL 的策略是引入扰动的领头者算法（PBTL）。PBTL 与 PFTL 相同，但会将第 i 轮的收益计入累计分数，即 $a^i = \text{argmax}_a(u_a^0 + U_a^i)$。我们将（a）通过稳定性引理建立 PFTL 与 PBTL 的联系，并（b）通过小扰动引理和定理 2.3 建立 PBTL 与 OPT 的联系。

引理 2.2 （稳定性）。 $\mathbb{E}[\text{PFTL}] \ge (1-\epsilon)\mathbb{E}[\text{PBTL}]$.

引理 2.3 （小扰动）。 $\mathbb{E}[\text{PBTL}] \ge \text{OPT} - \frac{h}{\epsilon}(1+\ln k)$。 –

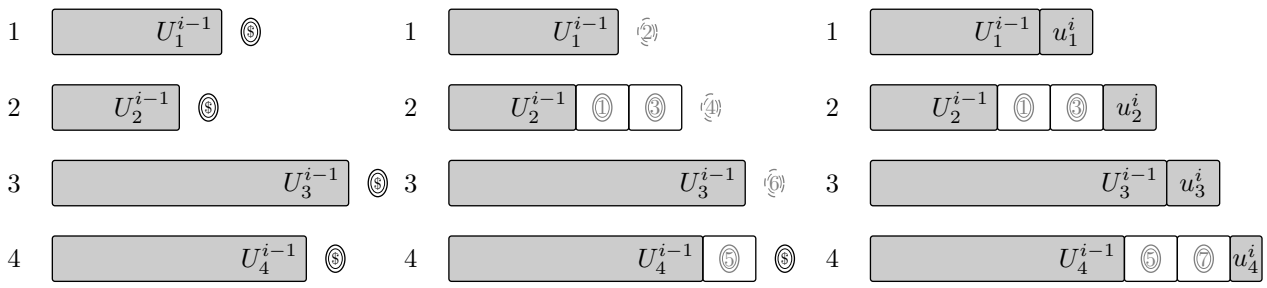
引理 *2.2* 的证明。 在相同的随机扰动下耦合 PFTL 和 PBTL，并以一致的方式处理平局。对于每一轮 i，从未扰动的分数开始，在每个行动上放一枚硬币（见图 2.2a），然后按如下方式添加扰动：选取总分最低的行动；抛掷其硬币；如果是反面 （概率为 $1-\epsilon$），则向其分数加上 h，否则将其舍弃（如果是正面）；重复此过程，直到只剩下一个行动 a^* （见图 2.2b）。抛掷 a^* 的硬币。如果结果是反面 （概率为 $1-\epsilon$），那么最佳行动的分数会比次佳行动高出 h 以上，因此 PFTL 和 PBTL 会选择相同的行动。因此，给定 PBTL 的行动，PFTL 与 PBTL 一致的概率至少为 $1-\epsilon$；否则，PFTL ≥ 0。根据期望的线性性，结论得证。

□

引理 *2.3* 的证明。 令 MAXPTRB $= \max_a u_a^0$ 表示最大幻觉扰动。与定理 2.3 相同的论证表明，PBTL +MAXPTRB $\ge$ OPT：将第 0 轮的收益视为实际收益，然后注意到，不含幻觉的 OPT 小于含幻觉的 OPT。整理得 PBTL $\ge$ OPT –MAXPTRB，取期望，并应用引理 2.1，即可得到所需的界。

□

[2] 事实上，该 exponential 权重算法等价于带 Gumbel 扰动的 PFTL 变体 .



(a) 幻觉前的分数　　(b) 从最低分开始幻觉　　(c) 最后一次抛硬币结果为反面

图 2.2：引理 2.2 的耦合论证示意图。中间图中展示的抛硬币结果为反面；正面；反面；正面；反面；正面 （如图中标注的 1–6）。抛出正面 的硬币以灰色虚线表示。右图中的抛硬币结果为反面（标注为 7）。右侧图示表明，对于第 $i = 4$ 轮的这些抛硬币结果，PFTL 和 PBTL 获得相同的收益。

结合引理 2.2 和引理 2.3 可得定理 2.4，证明 PFTL 达到了与指数加权相同的渐近遗憾率。稳定性（引理 2.2）使其能够抵御对抗性噪声，而较小的扰动（引理 2.3）意味着该算法能够迅速学会跟踪最佳行动。

注意：扰动跟随领导者的正式定义（定义 2.4）仅在做出任何决策之前进行随机化。对于此定义，上述遗憾趋零分析适用于任何预先固定的输入。在后续各节讨论的博弈中使用该算法时，我们将倾向于采用一个对自适应输入序列也成立的界。显然，根据上述分析，对于此类输入，可以证明对该算法的一个版本也有相同的界，该版本会在每一轮重新随机化幻觉。

3（在线）多臂老虎机学习

多臂老虎机学习将在线学习扩展到部分反馈的情境。全反馈在线学习中，学习者在每轮结束后都会观察到所有行动的收益；而在老虎机设定中，学习者只能观察到所采取行动的收益。这种信息限制造成了一种根本性的权衡：在探索 （了解不确定行动）与利用 （采取已知的最佳行动）之间进行权衡。

3.1 模型

考虑 k 个行动 A 和 n 轮次。令 $u_a^i \in [0, h]$ 表示行动 a 在第 i轮的收益。每轮中，学习者：

1. 选择一个行动 a^i，2. 观察实际获得
的收益 $u_{a^i}^i$，并且

3. 获得该收益。

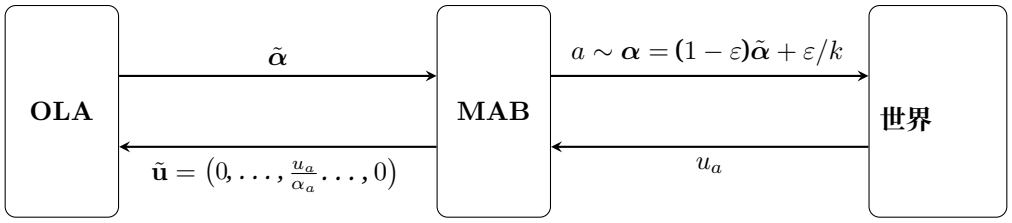


图3. 1：从多臂老虎机学习到全反馈在线学习的归约 g.

总收益为 $\mathrm{ALG}_n = \sum_{i=1}^{n} u^i_{a^i}$.

与在线学习问题一样，目标仍是在事后实现相对于最佳固定动作的消失遗憾，$\mathrm{OPT}_n = \max_a \sum_{i=1}^{n} u^i_a$. 第2节中的在线学习模型与多臂老虎机模型的区别在于，在后者中，每一轮只能观察到 $u^i_{a^i}$ ；学习者看不到其他 u^i_a 的 $a \neq a^i$。因此，如果没有执行某个行动 a ，就无法直接评估它的质量。

明确记下算法在行动上的概率分布会很有用。令 $\boldsymbol{\alpha}^i = (\alpha^i_1、\ldots\ldots、\alpha^i_k)$表示用于在第 i轮抽取行动的概率向量。学习者抽取 $a^i \sim \boldsymbol{\alpha}^i$。对于收益$\mathbf{u}^i = (u^i_1、\ldots\ldots、u^i_k)$，学习者获得的期望收益为 $\mathbb{E}\left[u^i_{a^i}\right] = \sum_a u^i_a \alpha^i_a = \mathbf{u}^i \cdot \boldsymbol{\alpha}^i$ ，其中$\mathbf{u}^i \cdot \boldsymbol{\alpha}^i$ 是向量点积。

3.2 归约为全反馈学习

我们将把部分反馈老虎机设定归约为全反馈在线学习设定。在这一归约中，多臂老虎机算法（以下简称 MAB）将与在线学习算法（以下简称 OLA）交互。OLA 将推荐行动。多臂老虎机算法将选择一个行动，该行动可能是 OLA 推荐的行动，也可能不是。多臂老虎机算法将学习所选行动的收益。随后，它必须为每个行动构造一个完整的收益向量，并将其输入 OLA。图 3.1 描述了这一结构。对于未被执行的行动，应向 OLA 输入什么内容，这是其中的关键难点；我们将通过生成所有行动收益的无偏估计来解决这一问题。

3.2.1 倾向评分

我们必须明确应向底层全反馈算法报告什么收益向量。令 $\tilde{}\mathbf{u}^i$ 表示报告的（估计）收益向量。目标是确保 $\tilde{}\mathbf{u}^i$ 是真实收益向量 $\mathbf{u}^i$ 的无偏估计量。这项称为倾向评分的技术，将观测到的收益除以其采样概率进行缩放，并将未观测到的收益记录为零。

定义 3.1 （倾向得分）。 对于已实现行动的分布 $\boldsymbol{\alpha}^i$ a^i，倾向得分 是对真实收益 $\mathbf{u}^i = （u^i_1, \ldots,$ u^i_k）的估计向量 $\tilde{}\mathbf{u}^i$，定义如下

$$\tilde{u}^i_a = \begin{cases} u^i_a/\alpha^i_a & if\ a = a^i, \\ 0 & otherwise. \end{cases}$$

倾向得分 $\tilde{}\mathbf{u}^i$ 是实际收益$\mathbf{u}^i$的无偏估计量，即 $\mathbb{E}\left[\tilde{}\mathbf{u}^i\right] = \mathbf{u}^i$

引理 3.1. 该 .

证明。 对于每个行动 a,

$$\begin{aligned}\mathbb{E}\big[\tilde{u}_a^i\big] &= \mathbb{E}\big[\tilde{u}_a^i \mid a^i = a\big]\ \mathbf{Pr}\big[a^i = a\big] + \mathbb{E}\big[\tilde{u}_a^i \mid a^i \neq a\big]\ \mathbf{Pr}\big[a^i \neq a\big] \\ &= \tfrac{u_a^i}{\alpha_a^i}\alpha_a^i + 0\,(1-\alpha_a^i) = u_a^i.\end{aligned}$$

将此等式应用于所有行动 a，可得 $\mathbb{E}\big[\tilde{\mathbf{u}}^i\big] = \mathbf{u}^i$。 □

在图 3.1 之后的归约中，倾向得分 $\tilde{\mathbf{u}}^i$ 将作为输入提供给在线学习算法。直观而言，相较于事后最优动作表现良好的在线学习算法，在只能获取无偏估计的情况下也能做到这一点。事实上，后续分析将证明这一结果。

3.2.2 有效收益范围的界

在线学习算法 OLA 的表现取决于其接收的收益范围。$\tilde{}$（回顾定理 2.2。）将此范围记为 $[1,h]$。利用倾向得分，在线学习算法的非零输入为 $\tilde{u}_a^i = u_a^i/\alpha_a^i$。注意，罕见动作的 α_a^i 较小，可能产生较大的估计收益。这些收益的范围上界为 $\tilde{h} = \max_{i,\ a} u_a^i/\alpha_a^i$。如果概率 α_a^i 没有远离零的下界，$\tilde{h}$ 就可能任意大，底层算法的遗憾界也会变得毫无意义。

幸运的是，我们的多臂老虎机算法可以自由选择这些概率，并确保它们不会变得过小。确保这些概率不会过小的最直接方法，是要求以至少 ϵ 的概率随机探索 k 个行动中的一个。这种随机探索可以为采取每个行动 a的概率提供一个下界，该下界为$\alpha_a^i \geq \epsilon/k$。

引理 3.2。 对于收益 $u_a^i \in [0,h]$，如果对所有 i都有 $\alpha_a^i \geq \epsilon/k$，$a$，那么 $\tilde{u}_a^i \leq \tilde{h} = kh/\epsilon$。

证明。 由于 $u_a^i \leq h$，我们有 $\tilde{u}_a^i = u_a^i/\alpha_a^i \leq h/\epsilon/k = kh/\epsilon$。 □

$\tilde{}$这种权衡显而易见：确保充分探索（较大的 ϵ）会使 h 保持较小，而过度探索则会降低利用带来的收益。

3.2.3 归约算法

现在，我们可以将无偏估计与最低限度的探索结合起来。下面的归约可将任意全反馈在线学习算法（OLA）转换为多臂老虎机算法（MAB）。将此归约应用于指数权重算法，即可得到 Auer 等人（2002）提出的*Exp3*算法的一个变体。在该归约中，学习率和探索概率均设为 ϵ。

定义 3.2 (从 MAB 到 OLA 的归约)。 在每一轮 i:

1. 从在线学习算法获取概率 $\tilde{\boldsymbol{\alpha}}^i$ 。
2. 将实际采样概率 $\boldsymbol{\alpha}^i$ 定义为

$$\alpha_a^i = (1-\epsilon)\,\tilde{\alpha}_a^i + \epsilon/k.$$

3. 抽取一个行动 $a^i \sim \boldsymbol{\alpha}^i$ 并获得收益$u^i_{a^i}$。

4. 将估计的收益向量 $\tilde{\mathbf{u}}^i$ 报告给在线学习算法，形式如下

$$\tilde{u}^i_a = \begin{cases} u^i_a/\alpha^i_a & \text{if } a = a^i, \\ 0 & \text{otherwise.} \end{cases}$$

定理 3.1. 假设在线学习算法在线学习算法 在 $[0, h]$ 上的收益满足保证 s:

$$\mathbb{E}[\mathrm{OLA}] \geq (1-\epsilon)\widetilde{\mathrm{OPT}} - \frac{\tilde{h}}{\epsilon}\ln k.$$

那么，对于 $[0, h]$ 上的收益，化简后的多臂老虎机算法满足

$$\mathbb{E}[\mathrm{MAB}] \geq (1-2\epsilon)\,\mathrm{OPT} - \frac{h\,k}{\epsilon^2}\ln k.$$

直观地说，根据引理 3.1，估计的收益是无偏的。根据引理 3.2，收益被~界定为 $h = {}^{kh}/_{\epsilon}$，这代入遗憾上界的第二项。由于我们仅以 $1-\epsilon$ 的概率遵循在线学习算法，整个上界还会额外损失 $(1-\epsilon)$；而如果我们始终遵循在线学习算法，本可获得的 $(1-\epsilon)$ OPT 会退化为 $(1-\epsilon)^2$OPT，这样更便于将其下界估计为 $(1-2\epsilon)$ OPT。接下来的正式证明将验证我们关于无偏估计的直觉是正确的。

我们可以求出最优学习与探索率 ϵ，并代入指数权重算法，得到以下推论。这个组合算法是 Exp3 的一种变体，即探索与利用指数权重算法。在 Auer 等人（2002）的直接分析中，该推论的遗憾上界改进为 $2.63\,h\sqrt[3]{\frac{k}{n}\ln k}$——该分析对无偏估计的方差采用了更好的上界，以不同方式设定学习率和探索率，并实现了关于轮数 n 的最优统计依赖关系。

推论 3.1. 对于 $[0, h]$ 中的收益，通过设置 $\epsilon = \sqrt[3]{\frac{k}{n}\ln k}$，基于指数加权（*Exp3*）的多臂老虎机算法可实现趋于零的每轮遗憾

$$\mathrm{Regret}_n \leq 3h\sqrt[3]{\frac{k}{n}\ln k}.$$

我们以定理 3.1 的正式证明结束本节。

定理 *3.1* 的证明。 此分析将在线学习算法（OLA）对其输入（即估计收益）的保证，转化为多臂老虎机~算法（MAB）对真实收益的保证。令 $\tilde{r} = {}^{h}/_{\epsilon}\ln k$ 表示估计收益上的加性损失，并令 $a^* = \mathrm{argmax}_a \sum_i u^i_a$ 表示根据事后真实收益得到的最佳固定动作。根据定理陈述中的假设，全反馈算法 OLA 对任意输入~$\tilde{\mathbf{u}}^1, \ldots, \tilde{\mathbf{u}}^n$ 以及任意行动 a^* 均能保证以下结果（尽管 a^* 是针对真实收益的最优行动，但对于估计收益而言，它是任意行动）：

$$\mathrm{OLA} = \sum\nolimits_i \tilde{\boldsymbol{\alpha}}^i \cdot \tilde{\mathbf{u}}^i \qquad \geq (1-\epsilon)\sum\nolimits_i \tilde{u}^i_{a^*} - \tilde{r}$$

对等式两边取期望，得：

$$\begin{aligned}\mathbb{E}_{\tilde{\boldsymbol{\alpha}},\tilde{\mathbf{u}}}[\mathrm{OLA}] = \sum\nolimits_i \mathbb{E}_{\tilde{\boldsymbol{\alpha}},\tilde{\mathbf{u}}}\left[\tilde{\boldsymbol{\alpha}}^i \cdot \tilde{\mathbf{u}}^i\right] &\geq (1-\epsilon)\sum\nolimits_i \mathbb{E}_{\tilde{\mathbf{u}}}\left[\tilde{u}^i_{a^*}\right] - \tilde{r} \\ \shortparallel & \qquad\shortparallel \\ \sum\nolimits_i \mathbb{E}_{\tilde{\boldsymbol{\alpha}}}\left[\tilde{\boldsymbol{\alpha}}^i \cdot \mathbf{u}^i\right] &\geq (1-\epsilon)\sum\nolimits_i u^i_{a^*} - \tilde{r}. \end{aligned} \tag{1}$$

两行之间的等式成立，是因为即使在给定 $\tilde{\boldsymbol{\alpha}}^i$ 的条件下，$\tilde{\mathbf{u}}^i$ 仍是$\mathbf{u}^i$ 的无偏估计量，即 $\mathbb{E}_{\tilde{\mathbf{u}}^i,\ \tilde{\boldsymbol{\alpha}}^i}\left[\tilde{\mathbf{u}}^i \mid \tilde{\boldsymbol{\alpha}}^i\right] = \mathbf{u}^i$。

多臂老虎机算法满足：

$$\begin{aligned}\mathbb{E}[\mathrm{MAB}] = \sum\nolimits_i \mathbb{E}\left[\boldsymbol{\alpha}^i \cdot \mathbf{u}^i\right] &= (1-\epsilon)\sum\nolimits_i \mathbb{E}\left[\tilde{\boldsymbol{\alpha}}^i \cdot \mathbf{u}^i\right] + \frac{\epsilon}{k}\sum\nolimits_{i,a} u^i_a \\ &\geq (1-\epsilon)\sum\nolimits_i \mathbb{E}\left[\tilde{\boldsymbol{\alpha}}^i \cdot \mathbf{u}^i\right]. \end{aligned} \tag{2}$$

将不等式 (1) 和 (2) 与 $(1-\epsilon)^2 \geq (1-2\epsilon)$ 及 $\tilde{r} = \frac{h}{\epsilon}\ln k = \frac{hk}{\epsilon^2}\ln k$ 合并：

$$\mathbb{E}[\mathrm{MAB}] \geq (1-2\epsilon)\,\mathrm{OPT} - \tilde{r} = (1-2\epsilon)\,\mathrm{OPT} - \frac{h\,k}{\epsilon^2}\ln k. \qquad \square$$

请注意，在线学习算法的对抗性界对这一归约至关重要。即使多臂老虎机问题的原始收益在各轮次之间独立同分布，这些无偏估计也并非如此。因此，在线学习算法在独立同分布模型下的保证，不能通过这一证明推广到多臂老虎机设定。

4 交换遗憾

本节的目标是将（完全反馈）在线学习算法的分析，从最小化事后最优遗憾（也称为外部遗憾）扩展到最小化交换遗憾 （也称为内部遗憾）。交换遗憾考察考虑每个行动所带来的总改进：查看使用该行动的轮次，并将其替换为这些轮次中的事后最佳行动。正如后续各节将要看到的，渐近消失的交换遗憾算法具有重要的经济学性质。本节将回到第2节的完全反馈模型，并沿用 Blum 和 Mansour (2007) 的分析；不过，也可以将本节的归约与第 3 节的归约结合起来，从而为部分反馈模型构造一种渐近消失的交换遗憾算法。

交换遗憾最小化可用以下直觉来理解（Camara 等人，2020年）。学习者在不同轮次采取不同的行动，揭示了影响收益的信息出现的时机，从而导致这些行动各不相同。渐近消失的交换遗憾要求学习者针对这些揭示的信息进行了充分优化。

4.1 模型

根据第2节的模型，在线学习算法在每一轮 $i = 1$、.....、 n选择一个混合策略 $\boldsymbol{\alpha}^i \in \Delta_k$，其中 α^i_a 表示算法在第 i轮选择行动 a 的概率，而 Δ_k是行动 $\{1,\ldots,k\}$上的分布集合。令$\mathbf{u}^i \in [0,h]^k$ 表示第 i轮实现的收益向量。

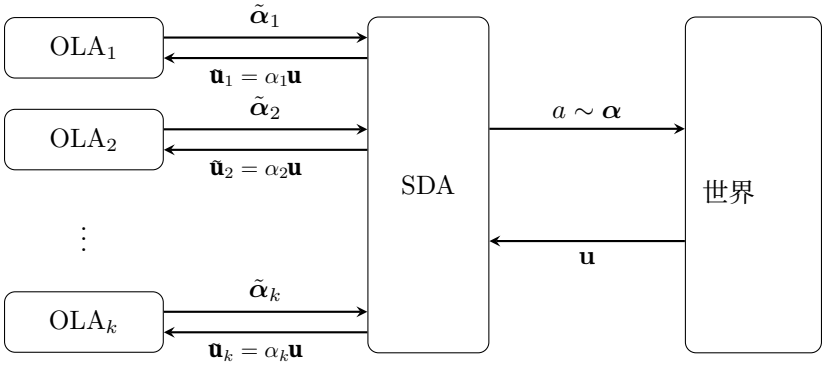


图 4.1：从交换遗憾到事后最优遗憾的归约。

定义 4.1 (事后最优遗憾)。如果对于所有偏离动作 $a \in \{1, \ldots, k\}$，算法在 n 轮中的事后最优遗憾至多为 r，则称该算法满足：

$$\mathrm{ALG} = \sum\nolimits_{i=1}^{n} \boldsymbol{\alpha}^i \cdot \mathbf{u}^i \geq \sum\nolimits_{i=1}^{n} u_a^i - rn.$$

定义 4.2 (交换遗憾)。如果对于任意表示偏离规则的函数 $f\colon \{1, \ldots, k\} \to \{1, \ldots, k\}$，某算法在 n 轮次内的交换遗憾至多为 r，则

$$\mathrm{ALG} = \sum\nolimits_{i=1}^{n} \boldsymbol{\alpha}^i \cdot \mathbf{u}^i \geq \sum\nolimits_{i=1}^{n} \sum\nolimits_{a=1}^{k} \alpha_a^i \, u_{f(a)}^i - rn.$$

对于一个固定的输入，交换遗憾比较的是：如果学习者始终将其每个行动 a 替换为某个替代行动 $f(a)$，结果会如何。显然，最小化交换遗憾比最小化事后最优遗憾是更强的条件。因此，具有渐近消失的交换遗憾的算法也具有渐近消失的事后最优遗憾。

4.2 从交换遗憾到事后最优遗憾的归约

回顾一下，最坏情况在线学习模型的一个重要特征是，更复杂的遗憾最小化问题（例如交换遗憾最小化）可以归约为事后最优遗憾最小化。同样地，这种归约在随机（独立同分布）假设下并不成立，凸显了最坏情况模型的相关性。该归约通过为每个行动维护一个事后最优算法，并通过一个平稳混合规则将这些算法结合起来实现；该规则确保委托给某个算法 a 来选择一个行动的概率，与该行动 a 被选中的概率保持一致（见图 4.1）。

一个具有 k 个状态的马尔可夫链由 k 个转移概率向量 $\tilde{\boldsymbol{\alpha}}_1$、.....、$\tilde{\boldsymbol{\alpha}}_k$ 给出，这些向量可以按行排列成一个行随机矩阵 A。每个有限马尔可夫链至少有一个平稳分布 $\boldsymbol{\alpha} \in \Delta_k$，它可以视为一个行向量，并满足 $\boldsymbol{\alpha} A = \boldsymbol{\alpha}$。换言之，如果我们以 $\boldsymbol{\alpha}$ 给出的概率作为初始分布，并通过 A 进行一次转移，得到的状态分布仍然是 $\boldsymbol{\alpha}$。注意，平稳分布 $\boldsymbol{\alpha}$ 也是 A 的主特征向量，即特征值为 1 的特征向量。对于这样的 $\boldsymbol{\alpha}$，我们有 $(\boldsymbol{\alpha} A)_a = \sum_\ell \alpha_\ell A_{\ell a} = \alpha_a$。

4.3 平稳委托算法

对于使用 k 个行动进行学习的情形，平稳委托算法（SDA）通过将决策委托给一组 k 事后最优算法之一，实现较低的交换遗憾。受委托的算法 OLA$_a$ 决定采取何种行动，并观察所有行动的收益（对委托决策取期望）。这些算法完全相同，但输入各不相同，因此通常会给出不同的建议。

定义 4.3（平稳委托算法，随机双重平均）。 实例化 k 事后最优算法在线学习算法$_1$，.....，在线学习算法$_k$。在轮次 i:

1. 每个算法在线学习算法$_a$ 推荐一个混合策略 $\tilde{\boldsymbol{\alpha}}_a^i \in \Delta_k$，我们将这些策略收集为矩阵的各行:

$$\tilde{A}^i = \begin{bmatrix} \text{———} & \tilde{\boldsymbol{\alpha}}_1^i & \text{———} \\ & \vdots & \\ \text{———} & \tilde{\boldsymbol{\alpha}}_k^i & \text{———} \end{bmatrix}$$

2. 令 $\boldsymbol{\alpha}^i$ 为 $\tilde{A}^i$的平稳分布，满足 $\boldsymbol{\alpha}^i \tilde{A}^i = \boldsymbol{\alpha}^i$。

3. 该算法通过对 $a^i \sim \boldsymbol{\alpha}^i$进行采样来选择一个行动。等价地，它首先以概率 α_a^i 选择算法a ，然后根据 $\tilde{\boldsymbol{\alpha}}_a^i$选择一个行动。

4. 将收益 $\tilde{\mathbf{u}}_a^i = \alpha_a^i \mathbf{u}^i$ 输入每个算法在线学习算法$_a$。

定理 4.1。 如果每个基础算法 在线学习算法$_a$ 保证事后最优遗憾至多为 r，那么随机双重平均 算法实现的交换遗憾至多为 kr。也就是说，对于所有交换函数 f： $\{1,\dots,k\} \to \{1,\dots,k\}$,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\alpha}^i \cdot \mathbf{u}^i \geq \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{k} \alpha_a^i \, v_{f(a)}^i - kr.$$

证明。 对于任意 a 和 $f(a)$，每个基础算法 在线学习算法$_a$ 都满足

$$\sum_{i=1}^{n} \alpha_a^i (\tilde{\boldsymbol{\alpha}}_a^i \cdot \mathbf{u}^i) \geq \sum_{i=1}^{n} \alpha_a^i v_{f(a)}^i - rn. \tag{3}$$

对所有 a 将左边相加，得到

$$\sum_{i=1}^{n}\sum_{a=1}^{k} \alpha_a^i (\tilde{\boldsymbol{\alpha}}_a^i \cdot \mathbf{u}^i) = \sum_{i=1}^{n} (\boldsymbol{\alpha}^i \tilde{A}^i) \cdot \mathbf{u}^i = \sum_{i=1}^{n} \boldsymbol{\alpha}^i \cdot \mathbf{u}^i,$$

其中，最后一个等号成立是因为 $\boldsymbol{\alpha}^i$ 是 $\tilde{A}^i$ 的平稳分布。因此，将不等式 (3) 两边对 a 求和，得到

$$\sum_{i=1}^{n} \boldsymbol{\alpha}^i \cdot \mathbf{u}^i \geq \sum_{i=1}^{n}\sum_{a=1}^{k} \alpha_a^i v_{f(a)}^i - krn,$$

并证明了该定理。 □

将定理 4.1 与定理 2.2 相结合，可得如下推论。尽管我们不给出具体细节，但仍有可能改进该推论，从而得到 $2h\sqrt{k \ln_n k}$ 的交换后悔界，即将 k 移入平方根内。这一改进源于用事后最优学习算法来实例化该归约，这些算法无需知晓时间范围 n 即可获得最优遗憾上界。事实上，扰动跟随领导者和指数加权均可加以修改以满足该条件。

推论 4.1. 采用指数加权作为在线学习算法的随机双重平均归约，实现的每轮交换后悔至多为 $2kh\sqrt{\ln_n k}$。

5 重复博弈中的学习

重复博弈是指阶段博弈重复进行多次的博弈。博弈中的学习所研究的一个典型问题是：在重复博弈中，收益起初未知的玩家是否能通过自然的程序，最终学会在阶段博弈中采取均衡策略。本节将采用类似的方法，考察博弈行为的经验分布的联合分布是否收敛到标准的静态均衡概念。

5.1 模型

考虑一个重复进行的有限双人双矩阵博弈，其中行玩家选择行动$a \in A$，列玩家选择行动 $b \in B$。行玩家和列玩家的相应收益分别记为$U_R(a,\ b)$ 和 $U_C(b,\ a)$。在每一天 i，

1. 行玩家和列玩家分别选择行动 a^i 和 b^i，并且
2. 观察实现收益；例如，行玩家观察到 $U_R(a^i,\ b^i)$。

平均收益是在 在这 n 天内计算；例如，行玩家的平均收益为 $1/n \sum_{i=1}^{n} U_R\ (a^i, b^i)$.

回想一下，混合纳什均衡假设玩家独立地随机化其行动，即行动的联合分布是乘积分布。重复博弈中的学习通常不会迅速收敛到独立随机化。

5.2 不收敛到纳什均衡

纳什均衡计算研究中的一项核心结果是，计算纳什均衡通常并不可行（Daskalakis 等，2009）。这一不可能性意味着，简单的学习动力学通常无法以任何意义迅速收敛到纳什均衡。如果连集中式算法都无法高效计算这样的均衡，学习算法也同样做不到。

定理 5.1。 在计算复杂性中的标准假设下，计算双矩阵博弈的纳什均衡，在最坏情况下是计算上不可行的

5.3 无后悔学习的均衡

沿着 Foster 和 Vohra（1997）以及 Hart 和 Mas-Colell（2000）的分析，我们将证明，无交换遗憾学习和无事后最佳遗憾学习的结果分别对应于相关均衡和粗相关均衡。

定义 5.1（相关均衡，CE）。行动上的联合分布 (a,b) 称为相关均衡，如果协调者根据该分布抽取 (a,b)，并私下向行玩家建议 a、向列玩家建议 b 时，每位玩家都认为遵从建议是最优的。具体而言，对于行玩家，

$$a \in \operatorname{argmax}_{a'} \mathbb{E}_{(a,b)}\left[U_R(a',b) \mid a\right],$$

列玩家的情况类似。

定义 5.2（粗相关均衡，CCE）。行动上的联合分布 (a,b) 称为粗相关均衡，如果协调者根据该分布抽取 (a,b)，并私下向行玩家建议 a、向列玩家建议 b 时，每位玩家都认为遵从建议，至少不劣于事前承诺采取任何一种忽略该建议的固定动作。具体而言，对于行玩家及所有行动 $a \in A$，

$$\mathbb{E}_{(a,b)}[U_R(a,b)] \;\geq\; \mathbb{E}_{(a,b)}\left[U_R(a',b)\right] \quad \textit{for all } a',$$

列玩家的情况类似。

混合纳什均衡等于相关均衡与独立行动的交集，也等于粗相关均衡与独立行动的交集。因此，纳什 $\subseteq$ 相关均衡 $\subseteq$ 粗相关均衡。另已知，在双人零和博弈中，纳什 $=$ 相关均衡 $=$ 粗相关均衡。在下面的无平局石头剪刀布博弈中，这些概念彼此各不相同。

例5.1 (无平局石头–剪刀–布**)**。 考虑如下所示收益的改良版石头剪刀布博弈。

	Rock	*Paper*	*Scissors*
Rock	$\mathbf{-6}, -6$	$\mathbf{-1}, 1$	$\mathbf{1}, -1$
Paper	$\mathbf{1}, -1$	$\mathbf{-6}, -6$	$\mathbf{-1}, 1$
Scissors	$\mathbf{-1}, 1$	$\mathbf{1}, -1$	$\mathbf{-6}, -6$

唯一的纳什均衡是均匀混合。还存在一些并非纳什均衡的相关均衡；例如，在 $\{(R,P)$、(R,S)、(P,R)、(P,S)、(S,R)、$(S,P)\}$上进行均匀混合时，遵循协调者的建议可获得0的收益，而对建议 P 的最佳收益为0。还存在一些并非相关均衡的粗相关均衡：在$\{(S,P)$、$(P,S)\}$ 上进行均匀混合时，遵循协调者的建议可获得0的收益，而事前偏离到 R、 P和 S 的收益分别为0、 $-7/2$和 $-5/2$。

定理 5.2。 当且仅当博弈行为的经验分布是粗相关均衡时，博弈行为不存在事后最优遗憾；当且仅当博弈行为的经验分布是相关均衡时，博弈行为不存在交换遗憾。

证明。 固定一个序列 $((a^1,\ b^1),\ldots,\ (a^n,\ b^n))$。对于粗相关均衡的结论，行玩家不存在遗憾意味着，对于每个 a'，

$$\sum_{i=1}^{n} U_R(a^i, b^i) \ \geq \ \sum_{i=1}^{n} U_R(a', b^i).$$

考虑一位协调者，从 $\{1,\ldots,n\}$ 中均匀选取 i，并向行玩家推荐 a^i，向列玩家推荐 b^i。当且仅当协调者诱导出粗相关均衡时，行玩家和列玩家的遗憾都为零，因为这些不等式完全相同。类似地，在无交换遗憾的博弈行为下，可得到相关均衡。 □

重要的是，算法所选择的博弈行为分布并不会收敛（也称最后迭代收敛）；收敛的是整个博弈行为的经验分布。具体而言，循环变化的博弈行为满足这一定义。

例5.2。 考虑无平局石头剪刀布游戏（例*5.1*）中的两位玩家，从 (R, S) 开始，反复对对方作出最佳回应。随后行动对的变化依次为 (R, P)、(S, P)、(S, R)、(P, R)、(P, S)、(R, S)，并无限重复。博弈行为的经验分布收敛于例*5.1*中给出的相关均衡，即在六种无平局行动对上均匀混合。每轮所选行动的分布是对上一轮最佳回应的点质量分布，它永远不会收敛。

6 操纵

博弈中的无后悔学习研究中的一个核心问题，是理解它与经典策略行为模型的相似之处和差异。操纵是一个重要方面，其经典模型是斯塔克尔伯格均衡：领导者承诺采用某种策略，跟随者则作出最佳回应。本节沿用 Deng 等人（2019）的分析，并表明在完全信息博弈中，具有渐近消失的事后最优遗憾的学习者仍可受到进一步操纵，而具有趋于零的交换遗憾的学习者则不会。

6.1 模型

考虑一个有限的双人双矩阵博弈，其中领导者 （行玩家）选择一个行动$a \in A$，而跟随者 （列玩家）选择一个行动 $b \in B$。相应的收益分别记为领导者的 $U_L(a, b)$ 和跟随者的 $U_F(b, a)$。

定义6.1 （斯塔克尔伯格均衡）。 当领导者承诺采取某个（可能是随机化的）行动，而跟随者作出最佳回应时，领导者的最优此类策略构成一个斯塔克尔伯格均衡，其斯塔克尔伯格值（SV）为双方的期望收益。

为便于阐述，我们假设跟随者的行动是对领导者斯塔克尔伯格策略的唯一最佳回应。若情况并非如此，只要领导者偏好的跟随者行动不受支配，领导者仍可迫使跟随者采取该行动。

命题6.1。 所有无后悔算法都可以被引导去采取斯塔克尔伯格结果。

证明。 如果领导者在每一轮都采取其斯塔克尔伯格策略——即采取静态策略——那么跟随者面对的是一个平稳环境。在平稳环境中的无悔学习者会收敛到最佳回应，而这正是跟随者的斯塔克尔伯格回应。因此，博弈过程会收敛到斯塔克尔伯格结果。

□

定义 6.2（可操纵性）。 如果存在某个博弈和一种领导者策略，使得领导者收益与斯塔克尔伯格值之差为正且不趋于零，则称学习算法是可操纵的 。

粗略而言，只能保证渐近消失的事后最优遗憾的算法往往具有可操纵性，而能保证渐近消失的交换遗憾的算法则不具有可操纵性。尽管渐近消失的交换遗憾意味着渐近消失的事后最优遗憾，但反之并不成立，因此，为了得到下界，我们聚焦于一个更弱的事后最优学习者子类。

定义 6.3 (均值型学习者**)**。 如果在每一轮中，行动被选择的概率至多为 γ，其中这些行动的每轮收益比事后最优动作的每轮收益 γ 低，那么该学习算法就是基于均值的 。

基于均值的算法的典型例子包括指数加权（EW）和扰动跟随领导者（PFTL）。

6.2 均值型学习者的可操纵性

定理 6.1。 基于均值的无后悔算法具有可操纵性。

证明。 考虑以下双矩阵博弈，其中领导者（行玩家，以粗体标出）与学习者（列玩家）对弈。

	Left	Mid	Right
Up	**0**, ϵ	**-2**, -1	**-2**, 0
Down	**0**, -1	**-2**, 1	**2**, 0

领导者的操纵策略是先玩上 $m = n/2$ 轮，然后玩下 $m = n/2$ 轮。在这样的序列中，均值型学习者的行为与 Be the Leader（BTL）类似。为简便起见，我们对 BTL 进行分析。BTL 与均值型学习者的主要区别在于：当一个新行动开始与事后最佳行动竞争时，后者会逐渐调整其概率，而前者会在新行动超越事后最佳行动时立即调整。

在前 $m = n/2$ 轮中，领导者选择上。在这些轮次中，BTL 选择左 ，其在（左、中、右）上的累积收益约为（ϵm，$-m$，0）。在剩余的 m 轮中，领导者切换到下。起初，BTL 继续选择左；再经过 ϵm轮后，累积收益变为（0，$-m+\epsilon m$，0），此时 BTL 切换到右。对局结束时，累积收益约为（$-m+\epsilon m$，0，0）。此时领导者的总收益约为

$$0 \cdot (m + \epsilon m) \; + \; 2 \cdot (m - \epsilon m) \; \approx \; 2m \; = \; n.$$

另一方面，这个博弈的斯塔克尔伯格值为 0。为此，我们考察领导者收益如何随其混合策略（由其选择上的概率定义）变化：

概率[采取上行动]	学习者行动	领导者收益
$[0, \frac{1}{2})$	Mid	-2
$(\frac{1}{2}, \frac{1}{1+\epsilon})$	右	$2 - 4$**概率**[选择向上] ≤ 0
$(\frac{1}{1+\epsilon}, 1]$	Left	0

因此，面对均值型学习者时，领导者每轮都能获得显著高于斯塔克尔伯格值的收益，这证明基于均值的无后悔学习者具有可操纵性。 □

注意，无交换遗憾算法不会在上述博弈中受到操纵。这是因为，交换后悔学习者会跟踪以其所采取行动为条件的收益。切换行动时，它会立即开始为新行动收集新的数据，而不是依赖过时的累积平均值。在上述博弈中，最初进行 $m + m\epsilon$ 轮 左 行动后，累积收益为 (0， $-m + \epsilon m$, 0)。切换到 右 后，会重新开始收集归因于右侧行动的收益，即在 (左、中、右) 上的 $(-1, 1, 0)$, 使 中 立即变得有吸引力，而不是等到第 $2m$ 轮；因此，领导者的操纵失败了。

6.3 无交换遗憾学习者的不可操纵性

定理 6.2。无交换遗憾算法不可被操纵。

T这一分析将用到斯塔克尔伯格值的以下推广 e.

定义 6.4 （乐观斯塔克尔伯格值，OSV(b, r)）。 对于跟随者行动 b 和遗憾容忍度 r，$\mathrm{OSV}(b, r)$是满足以下条件的混合策略 $\boldsymbol{\alpha} \in \Delta(A)$ 所能达到的最大领导者收益： b 与最佳回应的差距不超过 r ：

$$\mathbb{E}_{a\sim\boldsymbol{\alpha}}[U_F(b, a)] \geq \max_{b'\in B} \mathbb{E}_{a\sim\boldsymbol{\alpha}}\big[U_F(b', a)\big] - r.$$

如果行动 跟随者行动 b 无法在遗憾不超过 r 的条件下被强制采用，那么 $\mathrm{OSV}(b, r) = -\infty$。

乐观斯塔克尔伯格值具有若干重要性质，接下来的事实、引理和推论将对此加以说明。

事实 6.1。 斯塔克尔伯格值为 $\mathrm{SV} = \max_{b\in B} \mathrm{OSV}(b, 0)$。

引理 6.1。 对于每个 b，$\mathrm{OSV}(b, r)$ 都可以表示为一个有限线性规划，其中 r 作为单个线性约束出现；因此，$\mathrm{OSV}(b, r)$ 关于 r 是分段线性且凹的（ $-\infty$不成立的地方）。

推论 6.1. 存在常数 C，使得对于所有 b 和 r，$\mathrm{OSV}(b, r) \leq$ 斯塔克尔伯格值 $+ Cr$。

证明。 根据事实 6.1，所有行动 $b \in B$ 都满足 OSV(b, 0) ≤斯塔克尔伯格值。每条曲线 OSV(b, r) 都是分段线性的，因此 max(斯塔克尔伯格值, OSV(b, r)) 关于 r 是分段线性的。综合上述结果，对于某个有限常数 C，OSV(b, r) 上界为斯塔克尔伯格值 $+ r\,C$。 □

定理 *6.2* 的证明。 考虑序列 $((a^1, b^1) \ldots, (a^n, b^n))$。令 $\boldsymbol{\beta} \in \Delta(B)$ 表示跟随者行动的经验频率，并令 ρ_b 表示跟随者在选择 b 时的每轮遗憾。总平均遗憾为 $r = \boldsymbol{\rho} \cdot \boldsymbol{\beta}$。令 $\boldsymbol{\alpha}^b$ 表示跟随者选择 b时领导者在各行动上的条件分布。领导者的平均收益等于:

$$\sum\nolimits_b \beta_b\, \mathbb{E}_{a\sim\boldsymbol{\alpha}^b}[U_L(a, b)] \;\leq\; \sum\nolimits_b \beta_b\; \mathrm{OSV}(b, \rho_b) \;\leq\; \sum\nolimits_b \beta_b[\mathrm{SV} + C\,\rho_b] \;\leq\; \mathrm{SV} + C\,r,$$

其中的不等式依据 OSV 的定义、推论 6.1 和事实 6.1。随着遗憾 r 趋于零，领导者的收益趋近于斯塔克尔伯格值，因此领导者无法确保每轮都获得固定优势；由此可见，无交换遗憾学习者不可操纵。

□

由此可见，在完全信息的有限博弈中，跟随者的交换遗憾趋于零，足以保证斯塔克尔伯格领导者无法获得高于斯塔克尔伯格值的收益。交换遗憾趋于零使领导者的收益接近斯塔克尔伯格值的速度可能非常缓慢（见下文例6.1）。有趣的是，对于贝叶斯博弈和扩展式博弈，要保证不可操纵性，需要比无交换遗憾更强的条件（Arunachaleswaran 等人，2025年）。

例6.1。 考虑以下斯塔克尔伯格值为斯塔克尔伯格值 $= 0$的博弈：

	Left	右
Up	$1, 0$	$0, \epsilon$
Down	$0, 0$	$0, 0$

另一方面，对于 $r \leq \epsilon$，领导者可以在推论 *6.1* 中实现$\mathrm{OSV}(\text{左 } r) = r/\epsilon$，即 $C = 1/\epsilon$，一个任意大的数。

6.4 注记

Braverman 等人（2018年）首先提出了学习算法的可操纵性，并定义了均值型学习者这一类别。他们证明，在一个贝叶斯定价博弈中，买方运行基于均值的学习算法，而卖方试图操纵该算法以最大化收入，卖方可以攫取全部剩余，这超过了斯塔克尔伯格值（在此情形下，斯塔克尔伯格值也称为垄断收入，见 Myerson，1981年）。上述关于标准式博弈的分析，即基于均值的算法的可操纵性，以及趋于零的交换遗憾算法的不可操纵性，源自 Deng 等人（2019年）。这种可操纵性理论可以推广到贝叶斯博弈，但需要采用更精细的交换遗憾概念。参见 Mansour 等人（2022年）、Rubinstein 和 Zhao（2024年），以及 Arunachaleswaran 等人（2025年）。

7 学习竞标者的推断

大量证据表明，在频繁拍卖中持续参与的竞标者所提交的出价并不处于阶段博弈的均衡状态，例如 Google Ad Words 平台推广的在线广告拍卖。这些出价呈现出周期性行为，通常看起来像是由算法产生的。本节遵循 Nekipelov 等人（2015）的方法，在竞标者具有较低事后最优遗憾这一假设下，推断重复拍卖中竞标者的私人估值。

回顾一下，如果所有竞标者都是低遗憾学习者，那么行动的联合分布会逼近粗相关均衡（定理 5.2），而粗相关均衡是纳什均衡的一种推广。因此，针对低遗憾学习者发展计量经济学方法，是对经典计量经济学方法的一种放宽。

7.1 模型

共有 m 名竞标者，编号为 j，其固定估值为 $\mathbf{v} = (v_1, \ldots, v_m)$。该机制重复进行 n 轮，编号为 $i = 1, \ldots, n$。在第 i 轮，

1. 竞标者提交出价 $\mathbf{b}^i$，

2. 机制应用分配规则 $\mathbf{x}^i$ 和支付规则 $\mathbf{p}^i$, 并且

3. 每位竞标者 j 获得分配$x_j^i(\mathbf{b}^i)$，支付$p_j^i(\mathbf{b}^i)$，并具有线性效用 $U_j(\mathbf{b}^i) = v_j\, x_j^i(\mathbf{b}^i) - p_j^i(\mathbf{b}^i)$。

我们假设竞标者的行为在 n 轮次内满足较低的事后最优遗憾。

对于竞标者 j，从观测到的行动$\mathbf{b}^1, \ldots, \mathbf{b}^n$ 得到的平均实现效用为

$$\frac{1}{n}\sum\nolimits_i \left[v_j\, x_j^i(\mathbf{b}^i) - p_j^i(\mathbf{b}^i)\right].$$

设 z 表示任意替代出价。如果竞标者 j的观测行为相对于事后出价 z 的每轮遗憾至多为r_j，则

$$\frac{1}{n}\sum\nolimits_i \left[v_j\, x_j^i(\mathbf{b}^i) - p_j^i(\mathbf{b}^i)\right] \;\geq\; \frac{1}{n}\sum\nolimits_i \left[v_j\, x_j^i(z, \mathbf{b}_{-j}^i) - p_j^i(z, \mathbf{b}_{-j}^i)\right] \;-\; r_j.$$

7.2 可理性化集合

掌握出价数据$\mathbf{b}^1, \ldots, \mathbf{b}^n$，并了解阶段分配规则$\mathbf{x}^1, \ldots, \mathbf{x}^n$和支付规则$\mathbf{p}^1, \ldots, \mathbf{p}^n$的分析者希望推断竞标者的价值。遗憾的是，尽管分析者愿意假设竞标者的遗憾较低，却通常不知道其遗憾究竟有多低。由于缺乏这方面的信息，我们需要将价值和遗憾都视为参数，并确定与数据相符的参数集合。因此，我们的目标是刻画给定出价序列后可理性化的价值和遗憾集合。

定义 7.1（可理性化集合，R_j）。竞标者 j 的可理性化集合 R_j 是所有满足以下条件的对（v_j、r_j）的集合：竞标序列$\mathbf{b}^1$，...，$\mathbf{b}^n$ 对所有偏离z都满足上述遗憾不等式。

对于每个偏离 z，定义竞标者 j 相对于观测到的行动在分配和支付方面将经历的平均变化：

$$\Delta x_j(z) = \frac{1}{n}\sum_i \left[x_j^i(\mathbf{b}^i) - x_j^i(z, \mathbf{b}_{-j}^i)\right],$$
$$\Delta p_j(z) = \frac{1}{n}\sum_i \left[p_j^i(\mathbf{b}^i) - p_j^i(z, \mathbf{b}_{-j}^i)\right].$$

注意，$\Delta x_j(z)$ 和 $\Delta p_j(z)$ 是数据中的常数。重新整理遗憾不等式可得，对每个 z,

$$v_j\, \Delta x_j(z) \;-\; \Delta p_j(z) \;\geq\; -r_j.$$

因此，每个偏离 z 都对（v_j, r_j）施加一个线性约束。见图 7.1。

命题 7.1。竞标者 j 的可理性化集合

$$R_j \;=\; \left\{(v_j, r_j) \;:\; v_j \Delta x_j(z) - \Delta p_j(z) \geq -r_j \;\textit{for all}\; z\right\},$$

是凸集。

证明。 每个不等式都在 (v_j, r_j) 中给出一个半空间；半空间的交集是凸集。 □

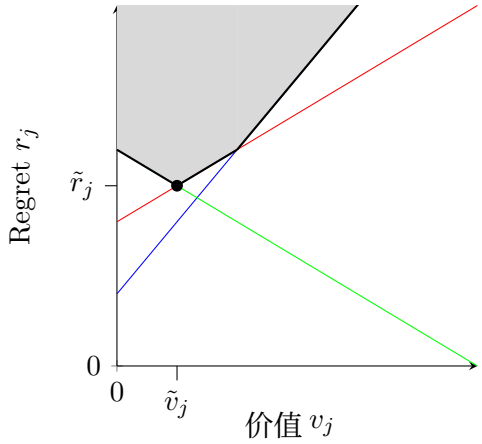


图 7.1：由线性不等式定义的竞标者 j 的可理性化集合。最小可理性化遗憾给出价值—遗憾值对 $(\tilde{v}_j, \tilde{r}_j)$ 。

7.3 识别与解释

由于真实的 (v_j, r_j) 必须位于 R_j 中，一种合理的确定某个点的方法是选择使隐含遗憾 $\tilde{r}_j$ 最小的值$\tilde{v}_j$：

$$(\tilde{v}_j, \tilde{r}_j) = \arg \min_{(v_j, r_j) \in R_j} r_j.$$

该方法识别出与尽可能小的遗憾相一致的价值 t.

如果学习者的遗憾较大，则更广范围的价值 v_j 与观测到的行动相一致。如果学习者的遗憾较小，则相一致的价值更少。小于$\tilde{r}_j$ 的极小遗憾（见图 7.1）是不可理性化的。

请注意，在完全反馈情形下，即当每个反事实出价 z的反事实分配和支付均为已知时，可理性化集合并非统计对象。它对事后实现成立。只要学习者进行了充分探索（见第3节），且分析者能够获得探索概率，就可以在部分反馈模型中采用类似的推断方法。在这种情况下，这些界是统计性的，因为它们基于估计的反事实分配和支付。

7.4 注释

上文介绍的 Nekipelov 等人（2015）的方法，已在多个方面得到改进和扩展。Nisan 和 Noti（2017年）提出了一种方法，在假设遗憾较低的价值更有可能出现的前提下，对参与者的价值进行点识别。他们将该方法应用于多个来自人类受试者实验的数据集，并发现其预测效果优于经典模型。Alaei 等人（2019）和 Noti 与 Syrgkanis（2021）分别将这些方法用于在线广告拍卖中的出价模拟和预测。Noti 与 Syrgkanis（2021）表明，与经典计量经济学方法相比，他们的方法更善于预测反事实行为，例如供给发生变化时的行为。

第 8 节随后讨论的基于数据监管算法合谋的方法，将 Nekipelov 等人（2015）的方法从完全反馈推广到部分反馈，并从

事后最优推广到交换遗憾。由于较小的交换遗憾意味着较小的事后最优遗憾，交换遗憾较小的竞标者的可理性化集合包含于事后最优遗憾较小的竞标者的可理性化集合。

8 算法合谋的监管

在线学习的一个重要应用领域是重复市场中的定价。尽管第2至第4节所发展的算法族可被证明能在阶段博弈中达到均衡（这定义了竞争性，见第5节），但对其他自然算法族的实证研究显示，它们倾向于形成超竞争价格，要么通过显式学习以牙还牙，要么经由统计错误（下文第8.5节将进一步讨论）。鉴于结果的这种二分，监管者可能希望强制采用前一类算法而非后几类算法。本节介绍Hartline等人（2024）用于理解并监管算法合谋的方法。

可能的定价算法空间与重复博弈中的均衡空间都很复杂。尽管本综述介绍了若干可用于学习价格的在线学习算法，卖家当然还可能使用其他算法。此外，众所周知，重复博弈中的均衡可以维持参与者之间的合作，即便不合作是阶段博弈中的占优策略。例如，两名卖家可以合作，双方都报出高价，即便报出低价占优于报出高价（即无论另一名卖家如何行动，报出低价都是最佳回应）。构造试图发现并维持此类合作的算法并不困难。此外，简单的合作策略可以在没有显式编程的情况下被自动学习到。

当前的合谋法律理论侧重于明确协议，因此允许采用其他可能实现与默契合谋相同结果的方法，并将其视为合法。不过，监管机构可能希望通过制定监管措施，限制超竞争价格对消费者造成的潜在损害。此类监管是否应限制允许使用的算法空间？如果应当限制，那么允许使用哪些算法，又应如何执行这些限制？

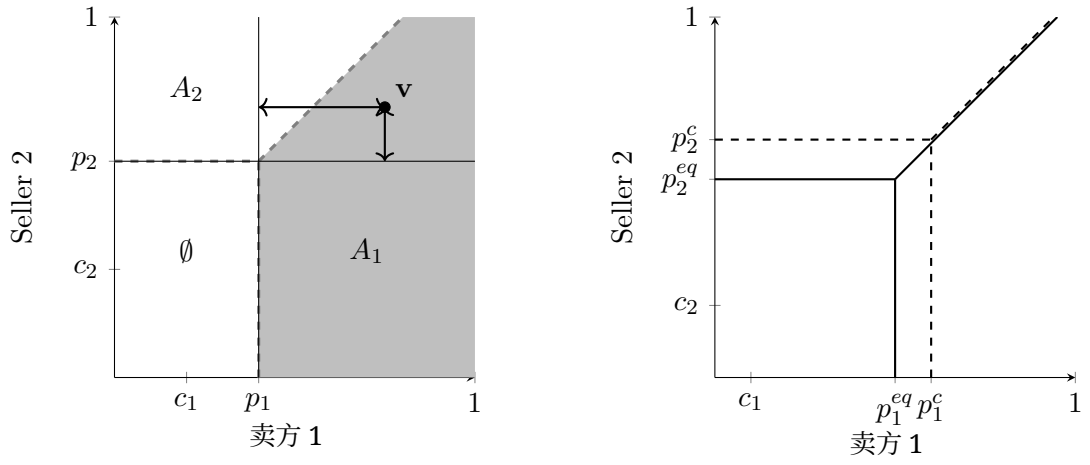
8.1 模型

考虑以下带有侧信息的不完全价格竞争简单模型。卖方 1 和卖方 2 以固定边际成本 c_1 和 c_2参与一个重复市场。在每一轮 i中，两位卖方都会收到私人信号 x_1^i 和 x_2^i；这些信号彼此之间以及与买方对各卖方产品的估值 $\mathbf{v} = (v_1^i, v_2^i)$ 之间可能存在相关性。随后，每位卖方 j 都会报出一个价格 p_j^i。买方 i 从提供更高净价值的卖方 j 处购买，

$$\text{argmax}_j \{v_j^i - p_j^i\},$$

前提是买方效用为正。因此，当买方的估值$\mathbf{v} = (v_1^i, v_2^i)$ 位于估值空间中的区域 A_i 内，即买方更偏好卖方 i的报价时，卖方 i 就会售出产品，如图 8.1a 所示。

两家卖家都可能通过将价格设定在竞争性水平之上而获益。例如，考虑一个包含 $c_1 = 0.1$ 和 $c_2 = 0.2$ 的例子，其中买方的价值取自 $\mathbf{v} \sim U[0,1]^2$。图 8.1b 展示了纳什均衡价格以及通过协调可以实现的最优超竞争



(a) 需求示意图。 (b) 均衡价格与合谋价格。

图 8.1：左图中，买方从卖家 i 处购买，位于区域 $A_i \subseteq \mathbb{R}^2$，其中 $v_i - p_i \geq \max(v_{-i} - p_{-i}, 0)$。价格以实线表示。一个示例价值对 $\mathbf{v} = (v_1, v_2)$ 以点的形式标出，该点到各个价格的距离对应于买方的效用。区域 $\emptyset$，A_1，A_2 由虚线分隔；例如，区域 A_1 以灰色阴影标出，并包含价值 $\mathbf{v}$。右图展示了成本为 $c_1 = 0.1$、$c_2 = 0.2$ 和 $\mathbf{v} \sim U[0,1]^2$ 的示例，以及均衡价格和合谋价格。相对于均衡价格，合谋价格要么使消费者无法购买，要么向消费者收取更高的价格，因此会损害消费者。

价格。因此，政策的核心关切在于，卖家采用的学习算法最终是趋向于竞争性结果，还是非竞争性结果。

该模型允许卖家拥有彼此相关且与需求相关的私人信息。Hartline 等人（2024）的监管观点是，卖家可能拥有此类侧信息，而根据侧信息调整价格并不属于合谋。因此，算法合谋监管模型必须允许这种情况。当然，模型也包括不存在侧信息的情形。

8.2 竞争行为

对于哪些行为构成合谋，人们可能意见不一。例如，在某些市场中，唯一的伯特兰均衡是所有卖家都报出低价；而斯塔克尔伯格领导者可以承诺报出高价，跟随者则对此作出最佳回应，报出中高价，最终所有卖家都能获益。这种（斯塔克尔伯格）行为及其带来的结果并不具有竞争性，但这种行为算合谋吗？此外，这位斯塔克尔伯格领导者与一位无知的卖方无法区分，后者只是报出高价，却并不明白这样做是好是坏。另一方面，（单边）竞争行为则很容易界定。在上面的例子中，斯塔克尔伯格卖方和无知的卖方都不具有竞争性，而作出最佳回应的跟随者则表现出竞争性。显然，竞争行为并非合谋。因此，我们将重点放在对（单边）竞争行为的监管上。

在没有侧信息的博弈中，竞争性结果的经典概念是纳什均

衡。遗憾的是，纳什均衡并不是确保算法具有竞争力的好概念。正如我们在第5节中看到的，纳什均衡通常在计算上难以处理（定理 5.1）。因此，我们通常不能指望算法能在合理的时间内达到纳什均衡。

另一方面，纳什均衡这一概念也过于严格，因为它不允许卖家拥有侧信息。承认卖家可以拥有彼此之间可能相关的侧信息，会扩大可能的均衡集合。具体而言，在所有可能的信息结构下，贝叶斯-纳什均衡的集合就是贝叶斯相关均衡的集合（Aumann，1987年）。

命题8.1。 随机博弈在所有信息结构下的贝叶斯-纳什均衡（*BNE*）集合，等于贝叶斯相关均衡（*BCE*）集合。

证明。 任何带有侧信息的贝叶斯-纳什均衡都会在行动和状态上诱导出一个联合分布。考虑这样一位协调者：根据状态，为行动推荐该联合分布。于是，贝叶斯-纳什性质意味着贝叶斯相关性质。反过来，考虑任意协调者在给定状态时对行动所作的条件联合分布，且该分布诱导出一个贝叶斯相关均衡。考虑与这些行动相同的侧信息，即每位玩家的侧信息都是其在该贝叶斯相关均衡中的行动。一种可能的联合策略组合是，每位卖家都采取与其侧信息相对应的行动。该联合策略是一个与贝叶斯相关均衡具有相同结果的贝叶斯-纳什均衡。

□

在第5节中，我们看到，在不含信息的博弈中，相关均衡集合对应于所有玩家的交换遗憾渐近消失时的结果集合（定理 5.2）。对于信息博弈中的贝叶斯相关均衡和情境交换遗憾渐近消失的玩家，同样的结论也成立。[3] 从监管机构的角度来看，当情境被隐藏时，情境交换遗憾渐近消失看起来就像交换遗憾渐近消失。因此，可由算法实现的单边竞争行为定义就是交换遗憾渐近消失。由于我们识别竞争行为的目标是监管算法合谋，因此识别一种单边属性至关重要；卖家应当能够选择遵循这一属性，从而按照该定义不参与合谋。

定义 8.1。 定价博弈中的卖方若其行动满足交换遗憾渐近消失，则具有竞争行为 。

8.3 学习与模型设定错误

经济学文献一直着重研究一类特定的*Q*学习 算法所导致的算法合谋。Q学习是一种用于马尔可夫决策过程的在线学习算法。Q学习假设环境的状态空间已知，但转移动态未知且平稳，并且在给定状态的条件下，收益独立同分布。

定义8.2 (马尔可夫决策过程, 马尔可夫决策过程**).** 马尔可夫决策过程由有限的状态集合、行动集合、每个行动-状态对到其他状态的转移概率，以及每个行动-状态对的收益分布所规定。当决策者处于给定状态并采取给定行动时，状态按照转移概率进行转移，决策者则按照收益分布获得收益。

[3]情境是卖方获得的信号。情境交换遗憾允许交换函数同时依赖于行动和情境。参见 Blum 和 Mansour（2007）；本文综述将不再进一步讨论这一模型。

请注意，在马尔可夫决策过程的定义中，下一状态与收益的分布在给定当前状态和行动的条件下是独立的。假设反馈同时包含状态和收益，并且存在能够无限次到达每个状态的策略，则 Q-learning 算法将学会在马尔可夫决策过程上最大化其收益。

经验证据表明，一些学习算法可能会收敛到超竞争价格。举例而言，Calvano 等人（2020年）表明，当 Q 学习中的状态 定义为一轮历史时，算法可以学会一种“以牙还牙”的模式，从而维持高价格。这一结果十分重要，因为它表明，即使算法没有被明确编程为维持合谋，也可能学会这样做。重要的是，审查这类算法的源代码，不会发现任何合谋意图。请注意，这种 Q 学习的应用存在模型设定错误，因为对手的策略并非仅由状态（一轮历史）决定。另一个例子是，Banchio 和 Skrzypacz（2022年）研究了无状态 Q 学习，并表明，即使是模型设定错误导致的统计错误，也可能产生超竞争价格。

重要的是，尽管这些超竞争结果可能是重复博弈的均衡（许多结果都是如此），但它们并非阶段博弈的均衡。因此，无论是否将其视为合谋，它们肯定都不是竞争性结果（定义 8.1）。在本节余下部分，我们将看到，可以设计一种监管机制，审计算法是否具有竞争性。

8.4 规制设计

政策面临的挑战是对算法定价进行规制，在确保竞争性的同时允许创新与优化。规制应满足以下要求：

1. 应允许使用信息和优化；2. 应具有单边性，即任何卖方都可以选择采取竞争性（而非合谋）行为；3. 应在计算上合理，即满足规制要求所需的计算要求不会过于繁重；4. 应仅依据可观察数据，而非访问源代码，从而保护知识产权并允许算法创新。

回想一下，计算纳什均衡通常在计算上难以处理（定理 5.1）。然而，在信息博弈中，纳什均衡与相关均衡一致（命题8.1）。此外，已知标准在线学习算法（第 4 节）能够保证收敛到相关均衡（定理 5.2）。这一观察揭示了一种颇有前景的规制方法：要求市场参与者使用的学习算法实现趋于零的交换遗憾。回想一下，交换遗憾最小化策略会收敛到相关均衡集合。

一个关键细节是，卖家的成本和信息结构可能无法直接观察到。合谋法律理论指出，法官不愿对卖家内心的想法作出裁决。因此，用于评估卖家是否具有竞争性的一种相容理论，应允许任何一致的信息结构或成本，即应采用显示性偏好和揭示的信息。对于显示性偏好，可以运用第 7 节中提出的方法，推断出使交换遗憾最低的成本。对于揭示的信息，检验

渐近消失的交换遗憾可确保，卖家通过选择不同价格而揭示的所有信息都得到了最优利用。

Hartline 等人（2024年）证明，只要有足够的数据，就能在已知样本复杂度界的情况下估计交换遗憾，从而无需检查专有源代码即可对算法行为进行经验监管。他们的统计检验中的一个关键步骤，是通过推广第 7 节的方法，估计最小可理性化遗憾（第 7 节和图 7.1）。下面的定理衡量该统计检验的“样本复杂度”，其主要结论在于样本复杂度对各种模型参数的依赖阶。该定理的一些强假设，例如已知离散价格水平和最小探索概率，已由 Hartline 等人（2025年）放宽。

定理8.1。 给定 k 个价格水平、最高价格 p、探索任意价格的最小概率 α、目标遗憾 r 和失败概率 δ；存在一种统计检验，以概率 $1-\delta$ 通过遗憾至多为 r 的算法，并拒绝遗憾超过 $2r$ 的算法，所需轮数为 $n = O((\frac{kp}{\alpha r})^2(\log k/\delta))$。

也可以通过审计渐近消失的事后最优遗憾，实施监管以强制实现粗相关均衡。这有两个缺点：首先，这类均衡的范围过大，超出必要范围（命题8.1），因此可能导致对消费者更不利的结果。此外，根据第6节关于操纵的研究结果，斯塔克尔伯格领导者可以通过操纵事后最优学习者，促成这些对消费者更不利的结果（Hartline 等人，2025年）。

总之，算法合谋监管可以建立在无后悔学习理论的基础上。监管机构要求学习算法表现出渐近消失的交换遗憾，便可确保市场结果对应于竞争性（阶段博弈）均衡，同时保留算法设计的灵活性并鼓励创新。这种方法为算法市场治理提供了以数据为驱动、计算上易于处理且以信息为基础的框架。

8.5 注记

关于算法合谋相关法律理论的讨论，以及对各种算法合谋监管方法的探讨，可参见 Harrington（2018年）。如今，经济学领域已有大量以实验为主的文献，研究了各类算法，并发现在某些配置下存在超竞争价格（Calzolari，2027）。

有一类论文表明，能够在包含以牙还牙和价格战规避等合谋策略的内部策略空间中学习的算法，即使没有经过明确编程，也会学会采用这些策略。Calvano 等人（2020）利用基于单期历史信息的 Q学习实现了这一点。Calvano等人（2021）将先前的分析扩展到带噪声的观测情形。Klein（2021）则将其扩展到异步更新。最后，Fish等人（2024）从基于大语言模型开发的定价算法中发现了这种行为。

另一类文献表明，采用设定错误模型的算法会犯下导致超竞争价格的统计错误。其所犯的统计错误在于假定收益是独立同分布的，而实际上并非如此，即模型设定错误。Asker 等人（2024）的主要结果比较了完全反馈与部分反馈下的无状态 Q 学习，其中部分反馈模型因模型设定错误而犯下统计错误；在其程式化的价格竞争设定中，他们发现完全反馈产生竞争性价格，而设定错误的部分反馈则产生超竞争价格。因此，Q学习的统计错误来自部分反馈

模型。[4] 请注意，如果不像 Calvano 等人（2020）那样纳入上一轮博弈，就无法学习以牙还牙等合谋策略。与此同时，Banchio 和 Skrzypacz（2022）对无状态 Q 学习进行了实证分析，Banchio 和 Mantegazza（2023）则给出了理论分析。其他算法也得到了研究。Hansen 等人（2021）对上置信界（UCB）算法进行了实证分析；该算法是一种多臂老虎机算法，当需求在各轮次间独立同分布时具有可证明的保证，但在定价竞争中存在模型设定错误。

相比之下，Hartline 等人（2024）认为，渐近消失的交换遗憾是竞争行为的一种单边定义；正如我们在第 4 节中所见，算法可以轻易满足该定义。如上所述，交换后悔性质可以根据数据进行审计。Hartline 等人（2025）扩展了该方法，以去除 Hartline 等人（2024）中的一些技术性假设。作为该方法的一个注意事项，Hartline 等人（2025）通过在与 Banchio 和 Skrzypacz（2022）类似的设定中对 Q学习进行模拟表明，当卖家的真实成本未知、且市场收敛到看起来对更高成本具有竞争性的价格时，该审计可能出现假阴性。尽管如此，这一定义的非合谋使监管机构能够识别出正在运行不良算法（即犯下统计错误的算法）的卖家，或学习了诸如以牙还牙之类合谋策略的卖家。

Hartline 等人（2024、2025）的一个核心思想——信息会在无后悔代理人的行动中被揭示，且渐近消失的交换遗憾对应于对这一揭示的信息的最优利用——来自 Camara 等人（2020）。他们将委托人—代理人机制设计中的共同先验假设放宽为委托人和代理人双方的无后悔学习。Collina 等人（2024）改进了该方法的计算效率与统计效率。

参考文献

S. Alaei、A. Badanidiyuru、M. Mahdian 和 S. Yazdanbod（2019）。低后悔代理人的响应预测。载于网络与互联网经济会议，第11920卷，第31–44页。E. R. Arunachaleswaran、N. Collina、Y. Mansour、M. Mohri、J. Schneider 和 B. Sivan（2025）。正规形式博弈之外的交换后悔与相关均衡。载于*ACM*经济学与计算会议，第130–157页。J. Asker、C. Fershtman 和 A. Pakes（2024）。人工智能设计对定价的影响。《经济学与管理战略杂志》，33(2)：276–304。P. Auer、N. Cesa-Bianchi、Y. Freund 和 R. E. Schapire（2002）。非随机多臂老虎机问题。《*SIAM* 计算杂志》，32(1)：48–77。R. J. Aumann（1974）。随机策略中的主观性与相关性。《数理经济学杂志》，1(1)：69–96。R. J. Aumann（1987）。相关均衡作为贝叶斯理性的体现。《计量经济学》，55(1)：1–18。

[4]请注意，在风格化模型的模拟中，具有完全反馈的Q学习取得成功，这与第2节中跟随领导者算法类似，并不意味着它在对抗模型中缺乏理论保证这一点与之矛盾。例如，参见 Bichler 等人（2024年）对这些算法在伯特兰竞争中的理论分析。

M. Banchio 和 G. Mantegazza（2023）。通过耦合实现自适应算法与合谋。载于*ACM*经济学与计算会议，第208页。M. Banchio 和 A. Skrzypacz（2022）。人工智能与拍卖设计。载于*ACM*经济学与计算会议，第30–31页。M. Bichler、J. Durmann 和 M. Oberlechner（2024）。重复价格竞争中的在线优化算法：均衡学习与算法合谋。*arXiv*预印本 *arXiv:2412.15707*。A. Blum 和 Y. Mansour（2007）。从外部遗憾到内部遗憾。机器学习研究杂志，8：1307–1324。M. Braverman、J. Mao、J. Schneider 和 S. M. Weinberg（2018）。向无遗憾买方销售。载于*ACM*经济学与计算会议，第523–540页。E. Calvano、G. Calzolari、V. Denicolo 和 S. Pastorello（2020）。人工智能、算法定价与合谋。美国经济评论，110(10)：3267–3297。E. Calvano、G. Calzolari、V. Denicol`o 和 S. Pastorello（2021）。不完全监控下的算法合谋。《国际产业组织杂志》，79：102712。G. Calzolari（2027）。智能体市场中的算法与竞争。载于《经济学与计量经济学进展：第十三届世界大会》，第2卷。剑桥。M. Camara、J. Hartline 和 A. Johnsen （ 2020 ）。无后悔代理人的机制：超越共同先验。载于 *IEEE* 计算机科学基础研讨会，第259–270页。N. Collina、A. Roth 和 H. Shao（2024）。面向无后悔代理人的高效无先验机制。载于*ACM*经济学与计算会议，第511–541页。C. Daskalakis、P. W. Goldberg 和 C. H. Papadimitriou（2009）。计算纳什均衡的复杂性。《*SIAM* 计算杂志》，39(1)：195–259。Y. Deng、J. Schneider 和 B. Sivan（2019）。针对无遗憾学习者制定策略。载于神经信息处理系统进展，第1577–1585页。S. Fish、Y. A. Gonczarowski 和 R. I. Shorrer（2024）。大语言模型的算法合谋。arXiv预印本。arXiv:2404.00806。D. P. Foster 和 R. V. Vohra（1997）。校准学习与相关均衡。《博弈与经济行为》，21：40–55。Y. Freund 和 R. E. Schapire（1997）。在线学习的决策理论推广及其在提升中的应用。《计算机与系统科学杂志》，55(1)：119–139。J. Hannan（1957）。重复博弈中对贝叶斯风险的近似。载于《博弈论研究》，第3卷，第97–139页。普林斯顿大学出版社。K. T. Hansen、K. Misra 和 M. M. Pai（2021）。前沿：算法合谋：独立算法带来的超竞争价格。《营销科学》，40(1)：1–12。

Harrington, J. E.（2018年）。制定针对自主人工智能代理合谋的竞争法。《竞争法与经济学杂志》，14(3)：331–363。Hart, S. 和 Mas-Colell, A.（2000年）。一种导向相关均衡的简单自适应程序。《计量经济学》，68(5)：1127–1150。Hartline, J.、Johnsen, A. 和 Li, Y.（2020年）。基准设计与先验无关优化。载于*IEEE*计算机科学基础研讨会，第294–305页。Hartline, J.、Long, S. 和 Zhang, C.（2024年）。算法合谋监管。载于*ACM*计算机科学与法律研讨会，第98–108页。Hartline, J.、Wang, C. 和 Zhang, C.（2025年）。算法合谋监管的改进：检验悲观校准遗憾。载于*ACM*计算机科学与法律研讨会，第108–120页。Kalai, A. 和 Vempala, S.（2005年）。在线决策问题的高效算法。《计算机与系统科学杂志》，71(3)：291–307。Klein, T.（2021）。自主算法合谋：序贯定价下的Q学习。《兰德经济学杂志》，52(3)：538–558。Littlestone, N. 和 Warmuth, M. K.（1994年）。加权多数算法。《信息与计算》，108(2)：212–261。Mansour, Y.、Mohri, M.、Schneider, J. 和 Sivan, B.（2022年）。在贝叶斯博弈中针对学习者制定策略。载于学习理论会议，第178卷《机器学习研究论文集》，第5221–5252页。Moulin, H. 和 Vial, J.-P.（1978 年）。策略性零和博弈：完全混合均衡无法改进的博弈类别。《国际博弈论杂志》，7(3- 4)：201–221。Myerson, R. B.（1981年）。最优拍卖设计。《运筹学数学》，6(1)：58–73。Nekipelov, D.、Syrgkanis, V. 和 Tardos, ´E.（2015年）。学习代理计量经济学。载于*ACM*经济学与计算会议，第1–18页。Nisan, N. 和 Noti, G.（2017年）。重复博弈结构计量经济学的量化遗憾方法。载于*ACM*经济学与计算会议，第123页。Noti, G. 和 Syrgkanis, V.（2021年）。有学习机制的重复拍卖中的出价预测。载于*Web*会议，第3953–3964页。ACM。Rubinstein, A. 和 Zhao, J.（2024年）。在第一价格拍卖中针对无遗憾学习者制定策略。载于*ACM*经济学与计算会议，第894–921页。Young, H. P.（2004年）。策略性学习及其局限。牛津大学出版社。