arXiv:2302.03614v1 [cs.GT] 7 Feb 2023

策略行为　与排队系统中的无悔学习

S

Lucas Baudin* Marco Scarsini† Xavier Venel‡

2023年2月8日

摘要

本文研究一种动态离散时间排队模型：每个时期，玩家都会获得一个新作业，并必须将所有作业送入容量有限的队列。玩家有动机尽可能晚地发送作业；然而，如果某个作业未能在固定截止时间前离开队列，其所有者就要受到惩罚，该作业也会被退还给玩家，并在下一时期重新加入队列。因此，系统稳定性（即系统中作业数量有界）无法得到保证。我们证明，如果玩家采取短视策略，那么当惩罚足够高时，系统就是稳定的。此外，如果玩家使用一种源自典型无悔算法（指数权重）的学习算法，那么当惩罚超过一个取决于系统中作业总数的界限时，系统就是稳定的。

1 引言

在排队的经典研究中，代理人在随机时刻到达，按照规定的状态等待，然后接受服务；服务时间也是随机的。在这些模型中，代理人的策略行为没有任何发挥空间。即使代理人拒绝加入队列或放弃排队，这也被建模为随机事件，而非代理人的策略选择。从 Naor（1969）的开创性论文开始，排队模型中便纳入了策略要素。例如，在 Naor 的模型中，代理人到达时会理性地选择是否加入队列或拒绝加入队列。

分析策略性排队模型的合适工具来自博弈论。文献在不同的随机假设、不同的服务机制以及不同的参与者策略集合下，研究了排队系统的多种策略性模型。研究目标各不相同，例如求解博弈的纳什均衡、考察其效率，以及分析各种均衡下的系统稳定性。

*LAMSADE，巴黎多菲内大学，法国巴黎。lucas.baudin@dauphine.eu

†路易斯大学经济与金融系，意大利罗马 Viale Romania 32，邮编 00197。msc arsini@luiss.it, https://orcid.org/0000-0001-6473-794X

‡经济与金融系，路易斯大学，意大利罗马 Viale Romania 32，00197，xvenel@luiss.it, https://orcid.org/0000-0003-1150-9139

Glazer 和 Hassin（1983）最先研究了一类有趣的问题：参与者必须在固定截止时间之前获得服务，否则就要支付高额惩罚。在这一模型的若干变体中，参与者会重复进行博弈；例如，他们每天通勤去办公室，并且需要准时到达。

近期，Gaitonde 和 Tardos（2020、2021）研究了一种离散时间排队模型，其中代理人使用学习算法来决定选择哪台服务器。他们分析的新意在于考虑了不同时期之间的溢出效应。他们的目标之一是为系统稳定建立条件。

1.1 我们的贡献

本文既借鉴了关于具有固定截止时间的队列的文献，也借鉴了有关学习的研究成果，并探讨了一种离散时间模型：每个时期，代理人都会收到一项需要由单台服务器提供服务的新作业，并需要决定何时将作业加入队列，同时权衡等待成本与因延迟而受到的严厉惩罚。模型包含溢出效应，因为截止时间前无法获得服务的作业会返回其所有者，并在下一时期重新发送给服务器；这些延迟任务随后会与当天新收到的作业合并。从排队的角度看，该模型是确定性的：每位玩家在每个时期恰好收到一项新作业。随机性来自玩家的行动（这些行动可以是混合策略），也来自队列的状态：如果多项作业同时加入队列，它们接受服务的顺序将是均匀随机的。

我们在代理人具有策略性但目光短浅的假设下，考察模型的若干方面：也就是说，他们在每个时期都会进行均衡博弈，但不会考虑其行动对未来产生的影响。这促使我们考察单期博弈的均衡。在这一框架下，我们证明，均衡的结构取决于系统中的工作数量是否超过每个时期的时段数；此外，还取决于延迟所需支付的惩罚成本是否足够高。

当系统中的工作数量超过每个时期的时段数，且惩罚成本足够高时，我们证明阶段博弈存在唯一的粗相关均衡（因此也存在唯一的纳什均衡），在该均衡中，所有玩家都会尽早将自己的全部工作送入队列。当系统中的工作数量不超过每个时期的时段数时，阶段博弈则存在多个均衡，我们对其结构进行了研究。令人惊讶的是，即使每个玩家都有一种极小极大策略，能够保证该玩家的工作按时完成，均衡中仍有一些工作会以正概率延迟。这意味着下一时期系统中的工作数量将多于当前时期。

在论文的第二部分，我们研究一个模型，其中参与者使用无悔学习算法来做出选择。正如前面提到的，每位参与者的工作数量可能随时期而变化。为应对这一情况，我们将采用一种考虑环境变化的指数权重算法

考虑到不断变化的环境。

我们使用博弈论的语言描述模型，但也可以从其他角度进行解释。例如，可以采用收益管理解释：在每个时期，代理人购买作业的优先级。优先级等级各不相同，价格随优先级单调变化。高优先级的作业先于低优先级的作业得到处理，但每个时期的处理量不超过固定容量。发生优先级冲突时，则随机决定处理顺序。约束条件是：最多处理一名优先级最低的代理人的作业，最多处理两名优先级最低的代理人的作业，依此类推。

1.2 相关文献

关于排队系统中策略行为的分析可追溯至 Naor（1969）的开创性论文。他研究了采用先到先服务（FIFO）策略的 M/M/1 排队系统，其中代理人的收益由其获得服务时得到的回报减去等待成本构成，而等待成本与其在队列中花费的时间成正比。到达后，代理人可以决定加入队列或拒绝加入队列。如果他们采用纳什均衡策略，其行为在社会层面上是无效率的，也就是说，这种行为不能使社会收益（所有参与者收益之和）最大化。这是因为一个代理人的自利行为没有考虑到其对其他代理人造成的外部性。Hassin（1985）指出，采用后到先服务（LIFO）策略可以实现最优。此后，关于策略性排队系统的文献迅速增长。有关该主题的全面论述，请参阅 Hassin 和 Haviv（2003）以及 Hassin（2016）。

一些模型研究了可以决定何时加入队列的策略性智能体。Glazer 和 Hassin（1983）的开创性论文研究了一种称为 ?/M/1 的模型：设施每天从时间 0 开始提供服务，并按照先进先出原则为某个时间T 之前到达的所有顾客提供服务。每天，代理人都会决定是否前往该设施。如果决定前往，他们会选择到达时间，以尽量缩短预期排队时间。Haviv 和 Ravner（2021）最近综述了这类关于到达时间策略性选择的排队文献。

在交通运输理论的框架下，Vickrey（1969）研究了一个瓶颈模型，其中代理人选择离家时间以便到达办公室，并采用了某种流体近似。Rivera 等（2018）研究了 Vickrey 模型的离散版本，并考察了纳什均衡、相关均衡及其效率。我们的模型是他们模型的一个（变体）重复版本，并包含溢出效应。Kawasaki 等（2023）将这一瓶颈模型的研究扩展到更一般的偏好，并给出了纯纳什均衡存在的条件；此外，他们还考察了该模型与强均衡之间的联系。

最近，Gaitonde 和 Tardos（2020）提出了一种模型，其中多个队列以固定且不随时间变化（但因队列而异）的概率接收数据包，并必须将它们发送到若干服务器。每台服务器都可能以固定且不随时间变化的概率处理收到的数据包。

该论文研究了参与者采用无悔学习程序时系统的行为，尤其确定了系统保持稳定所需的模型参数条件。在后续论文中，Gaitonde 和 Tardos（2021）比较了采用无悔短期学习的参与者与采取长期优化行为的参与者的表现。Sentenac 等人（2021）使用了类似的模型，并提出了一种参与者可以采用的新型协作式分布式算法。与这项工作相比，我们的论文确实假定参与者有意合作，即他们使用标准算法来最大化自身收益。

无悔学习过程算法是一类在线强化学习算法，其表现至少不逊于事后选定的任何固定行动（即在其他玩家的行动或环境行动固定的情况下）。这一概念也称为 Hannan 一致性或普适一致性（Fudenberg 和 Levine，1998），最初由 Hannan（1957）提出，且多种算法都满足这一性质（近期综述见 Perchet，2014）。具体而言，指数权重算法（EWA）（Littlestone 和 Warmuth，1994；Cesa-Bianchi 和 Lugosi，2006）是一种简单而高效的强化学习方法，它会根据过去的经验调整各行动的权重。然而，目前还没有广受认可的方法，将无悔学习的概念推广到具有（内生变化的）状态变量的系统（例如马尔可夫决策过程），如我们的系统（队列状态）。因此，我们提出一种简单的、多层次的 EWA 扩展。Besbes 等人（2015，2019）研究了奖励存在时间不确定性的随机优化和单智能体多臂老虎机问题。在一篇近期颇具启发性的论文中，Crippa 等人（2022）在（非随机）重复博弈的背景下，研究了相对于动态动作序列而言具有无悔性质的策略，这些序列的变化次数随时间次线性增长。此外，Cardoso 等人（2019）、Mertikopoulos 和 Staudigl（2021）、Duvocelle 等人（2022）、Zhang 等人（2022）、Anagnostides 等人（2023）等学者也对时变博弈进行了大量研究。在博弈论框架下，玩家的奖励函数可能因两种原因随时间变化：博弈发生变化，或其他玩家的策略发生变化。在这些文章中，作者探讨了不同的解概念，以将这一区别纳入考量。与我们的模型的一个关键区别在于，博弈的变化在前述研究中是外生的，而在我们的模型中则是内生的。我们采用较为经典的遗憾概念。

1.3 概要

第2节介绍模型以及本文将要探讨的关键概念。第3节分析短视策略性参与者的行为。本节的重点是分析参与者在任意工作分配下进行的一次性博弈。第4节分析学习型玩家参与的重复博弈。

2 模型

我们考虑一种离散时间博弈，其中 I 是参与者集合，时间被划分为长度相等的时期 $L \geqslant 2$。一般时期记为 t。全文假设参与者至少有 $|I| \geqslant 3$ 名。

在每个时期开始时， $t \geqslant 0$ 每位玩家 i 都会收到一项作业，并独立选择一个行动 $a_t^i \in \{0、\ldots\ldots、L-1\}$，该行动表示在时期 t 内 i的作业加入队列的时间。符号 $\boldsymbol{a}_t := (a_t^i)_{i \in I}$ 表示时期 t的行动组合。由于玩家 i的所有作业会在同一时间加入队列，行动集 A 不取决于每位玩家持有的作业数量。

a_t^i 的选择可能取决于博弈的历史。根据所有参与者采取的行动，有些作业会延迟，即无法在该时期结束前离开队列。延迟作业会被返还给各自的所有者，并将在下一时期再次加入队列。这会产生一种溢出效应：在每个时期，玩家 i 处理的作业数量等于上一时期 i的延迟作业数量加一。

时期 t 系统的状态是向量 $\boldsymbol{k}_t = (k_t^i)_{i \in I}$，其中 k_t^i 是玩家i在时期 t开始时的作业数量。状态空间为 $S := \mathbb{N}^I$。时间t 的作业总数为

$$k_t := \sum_{i \in I} k_t^i. \tag{2.1}$$

时期 t 的延迟任务会产生惩罚成本 C_{k_t}，该成本取决于时期 t系统中的作业总数 k_t。在时期 t结束时，每位玩家 i 支付的成本 $c_{\boldsymbol{k}_t}^i(\boldsymbol{\sigma}_t)$ 由两部分组成：等待成本，即 i的每项作业在队列中花费的时间，以及惩罚成本：

$$c_{\boldsymbol{k}_t}^i(\boldsymbol{a}_t) = k_t^i(L - a_t^i) + C_{k_t}\mathbb{E}\left[\text{number of } i\text{'s jobs that are late at } t\right]. \tag{2.2}$$

给定行动组合 $\boldsymbol{a}_t$，假设玩家 i的所有作业发生延迟的概率都相同。因此

$$\mathbb{E}\left[\text{number of } i\text{'s jobs that are late at } t\right] = k_t^i p^i(\boldsymbol{a}_t), \tag{2.3}$$

这意味着

$$c_{\boldsymbol{k}_t}^i(\boldsymbol{a}_t) = k_t^i(L - a_t^i) + C_{k_t} k_t^i p^i(\boldsymbol{a}_t). \tag{2.4}$$

一个动态排队模型 (DQM)由周期长度 L、参与者数量 I 以及惩罚序列 $(C_{k_t})_{t \in \mathbb{N}}$ 指定，该惩罚序列取决于系统中的作业总数 k_t。

鉴于溢出效应，研究工作数量 k_t 的渐近行为是很自然的。特别是，如果 $k_t > L$，在时期 t 有比时间单位更多的工作；因此，至少

必然有一项工作延迟。即使 $k_t \leqslant L$，也可能有一些工作延迟。在本文的其余部分，我们将关注在不同玩家行为下对系统中工作数量的界限进行研究。更正式地说，我们令 $\mathcal{H}_t := (S \times A)^{t-1} \times S$ 表示长度为 t 的历史集合，并令 $\mathcal{H} := \cup_{t \geqslant 1} \mathcal{H}_t$ 表示有限历史集合。

定义 1（策略）。 策略 是一个函数 $x^i \colon \mathcal{H} \to \triangle(A)$。在不存在歧义的情况下，我们将用 $x_t^i := x^i(h_t)$ 表示玩家 i在时期 t的混合行动。

定义 2（策略剖面的稳定性）。 如果 DQM 中的策略组合 $\boldsymbol{x}$ 满足工作数量 k_t 几乎必然有界，即，则称其为稳定 的。

$$\mathbb{P}_{\boldsymbol{x}}\left[\exists M, \forall t \in \mathbb{N}, k_t \leqslant M\right] = 1.$$

以下假设将在全文中始终成立。

假设 3。 以下不等式成立： $L \geqslant I$。

若没有假设 3，则 DQM 中不存在任何稳定的策略组合，因为每个时期至多有 L 项工作离开系统，而有 I 项新任务到达。

我们研究两类不同策略组合的稳定性。在第一种情况下，代理人采取短视策略，即关注当前时期的状况，并针对该时期采取策略行动，也就是说，他们在每个时期都采取相关均衡。我们证明，只要惩罚成本 C_{k_t} 足够大（可能为常数），DQM 在该策略组合下就是稳定的。在第二种情况下，代理人使用取决于其工作数量的无悔策略。我们为这些策略在 DQM 中的稳定性提供了充分条件。

3 短视策略型参与者

本节假设参与者具有策略性且短视。短视假设意味着，在时期 t ，参与者只考虑自己在当前时期可能支付的成本，而不考虑其行动对博弈未来的影响。换言之，参与者反复进行收益函数为 $(c^i_{\boldsymbol{k}_t})_{i \in I}$ 的一次性博弈。

现在我们介绍短视纳什均衡 (NE)和短视粗相关均衡 (CCE)的定义。

定义 4（短视纳什均衡）。 若策略配置 $\boldsymbol{x} \colon \mathcal{H} \to \triangle(A^I)$ 满足：对于任意当前状态为 $\boldsymbol{k}_t$ 的历史 h_t ，混合行动 $\boldsymbol{x}(h_t)$ 都是由工作 $\boldsymbol{k}_t = (k_t^i)_{i \in I}$ 构成的单次博弈的纳什均衡，则称其为短视纳什均衡 。

定义 5（短视粗相关均衡）。 策略分布 $\boldsymbol{x} \colon \mathcal{H} \to \triangle(A^I)$ 称为短视粗相关均衡 ，如果对于任意历史 h_t 及当前状态 $\boldsymbol{k}_t$，分布 $\boldsymbol{x}(h_t)$都是包含工作 $\boldsymbol{k}_t = (k_t^i)_{i \in I}$的单次静态博弈的粗相关均衡。

当工作数量低于 L 时，任何选择 0 的玩家都一定不会支付惩罚成本。即使在这种情况下，也存在一些均衡，使得某些延迟任务以正概率延迟。因此，正如第3.1节中所详述的，工作数量在整个博弈过程中可能并不保持不变。

现在，我们分析由每位玩家拥有的作业数量作为参数的一类单次博弈。

本节其余部分安排如下。我们先聚焦于单次博弈，先讨论 $L \geqslant k_t$ 时的情形，再讨论 $L < k_t$ 时的情形。最后，我们分析全局队列。

3.1 $L \geqslant k_t$ 时单次博弈中的均衡

我们将证明，即使在 $L \geqslant k_t$ 时，均衡中仍有一些作业以正概率延迟。

我们首先证明，均衡中只会使用最后 k_t 个行动；因此，只需考虑 $L = k$ 的情形。

引理 6 (被支配行动，适用于 $L > k_t$)。若 $L > k_t$，则对于每位玩家 i，任意行动 $a_t^i \in \{0, \dots, L - k_t - 1\}$ 都被行动 $L - k_t$ 严格支配。此外，对于任意（混合）行动组合 $\boldsymbol{\sigma}_t$，

$$c^i_{\boldsymbol{k}_t}(L - k_t, \boldsymbol{\sigma}_t^{-i}) = k_t^i k_t. \tag{3.1}$$

式 (3.1) 表明，玩家 i 采取行动 $L - k_t$ 时支付的成本 $k_t^i k_t$ 与其他玩家的行动无关。

引理 6 的证明。如果玩家 i 的作业在时间 a_t^i 加入队列，那么它们最迟会在时间 $a_t^i + l$ 离开队列，其中 l 是截至时间 a_t^i （含该时间）加入队列的作业数量。由于 $l \leqslant k_t$，玩家 i 的作业最迟会在时间 $a_t^i + k_t$ 离开队列；如果 a_t^i 小于或等于 $L - k_t$，那么 $a_t^i + k_t$ 小于或等于 L。因此，玩家 i 的某个特定作业延迟的概率 $p^i(a_t^i, \boldsymbol{\sigma}_t^{-i})$ 为 0，玩家 i 的成本为 $k_t^i(L - a_t^i)$。特别地，选择行动 $L - k_t$ 的成本为 $k_t^i k_t$，严格小于任何 $a_t^i \in \{0, \dots, L - k_t - 1\}$ 的成本。

□

引理 6 表明，只需考虑 $L = k_t$ 的情况。本节其余部分的所有结果都将在这一假设下证明。一般情形 $L \geqslant k_t$ 可以通过重新命名行动得到。

下一个定理表明，在任何均衡中，参与者都可以分为两组：第一组中的参与者选择相同的混合行动，并且只在前两个未被支配行动$L - k_t$和 $L - k_t + 1$上进行混合；第二组中的参与者不在 $L - k_t$ 上进行混合，并且对 $L - k_t + 1$赋予严格正的权重。他们在其余未被支配行动上的混合方式尚未确定。

符号 $\mathsf{SC}(\boldsymbol{\sigma}_t)$表示配置 $\boldsymbol{\sigma}_t$ 的社会成本，即

$$\mathsf{SC}(\boldsymbol{\sigma}_t) := \sum_{i \in I} c^i_{\boldsymbol{k}_t}(\boldsymbol{\sigma}_t). \tag{3.2}$$

定理 7（纳什均衡的结构）。 如果 $L \geqslant k_t$、 $C_{k_t} > k_t^2$ 和 $\boldsymbol{\sigma}_t$ 是纳什均衡，那么：

(i) 对于每个玩家 j和行动 $a_t^j < L - k_t$，我们有 $\sigma_t^j(a_t^j) = 0$;

(ii) 存在一个玩家 i ，使得

(a) $\sigma_t^i(L - k_t) > 0$,

(b) *for all* $a_t^i > L - k_t + 1$, *we have* $\sigma_t^i(a_t^i) = 0$;

(iii) 对于每个玩家 $j \neq i$, $\sigma_t^j(L - k_t + 1) > 0$;

(iv) 对于每个玩家 j，如果 $\sigma_t^j(L - k_t) > 0$，那么 $\sigma_t^j = \sigma_t^i$;

(v) $k_t^2 - k_t + 1 \leqslant \mathsf{SC}(\boldsymbol{\sigma}_t) \leqslant k_t^2$。

证明。 (i) 在 $L - k_t$ 之前的行动不会被选择。 根据引理 6，我们知道所有小于 $L - k_t$ 的行动都是严格劣势。结论由纳什均衡不会对被支配行动进行混合这一事实得出。在证明的其余部分，我们假设 $L = k_t$。如前所述，如果 $L < k_t$，则只需通过平移即可轻松调整证明：动作0变为 $L - k_t$，1变为$L - k_t + 1$，依此类推。

(ii)(a) 存在一名玩家 i 使得 $\sigma_t^i(0) > 0$。 反证假设， $\sigma_t^j(0) = 0$ 对每名玩家 j。那么至少有一项工作会延迟。这意味着

$$\mathbb{E}\left[\text{number of late jobs}\right] = \sum_{j \in I} k_t^j p^j(\boldsymbol{\sigma}_t) \geqslant 1 \tag{3.3}$$

并且存在 i 使得$k_t^i p^i(\boldsymbol{\sigma}_t) \geqslant k_t^i / k_t$。于是，

$$c_{\boldsymbol{k}_t}^i(\boldsymbol{\sigma}_t) \geqslant k_t^i C_{k_t} p^i(\boldsymbol{\sigma}_t) \geqslant C_{k_t} k_t^i / k_t.$$

因此，如果 $C_{k_t} > k_t^2$，玩家 i的成本严格高于 $k_t^i k_t$，这就产生了矛盾。因此，一名玩家在0和 $c_{\boldsymbol{k}_t}^i(\boldsymbol{\sigma}_t) = k_t^i k_t$上进行混合。

(iii) 其他每名玩家都在 1上进行混合。 如果情况并非如此，则对于某个 $j \neq i$， $\sigma_t^j(1) = 0$ 。现在我们证明， i 通过选择纯行动1可以有利地偏离。假设事实上 i选择行动1。如果 j 选择纯行动0，那么玩家 i在1处加入队列的工作不会延迟，因为在1处加入队列的工作数量小于 $k_t = L$，因此这些工作会在时期结束前处理完毕。否则，如果 j 不选择0，那么 j的行动不小于2。于是，在0或1处加入队列的工作数量至多为 $L - 1$ ，并且这些工作会在时期结束前处理完毕。

因此，改为选择 1 后， i 的工作不会延迟，成本为 $c_{\boldsymbol{k}_t}^i(1, \boldsymbol{\sigma}_t^{-i})) = k_t^i(k_t - 1)$。因此，这种偏离是有利可图的， $\boldsymbol{\sigma}_t$ 不可能处于均衡。

(iv)（第一步）所有在 0 上混合的参与者都在 1上混合，并且在行动 1上赋予相同的权重。假设至少有两名参与者在 0 上混合。我们先证明他们也在 1 上混合。将这两名参与者称为 i 和 j。参与者 i 满足 (ii)(a)；因此，根据 (iii)，参与者 j 在 1 上混合。类似地，参与者 j 满足 (ii)(a)；因此，根据 (iii)，参与者 i 在 1 上混合。这说明两名参与者都在 1 上混合。

现在我们证明，他们对行动1赋予相同的权重。只有当所有作业都在时间1加入队列时，这些参与者中的任何一人的作业才可能延迟。事实上，如果至少有一个作业在时间0加入队列，那么在时间1加入队列的作业都不会延迟，因为有 $L-1$ 的时间来处理它们。

因此，只有当所有其他作业也在时间1离开时，也就是所有玩家都采取行动1时，在时间1离开的作业才会延迟。如果发生这种情况，作业延迟的概率为 $1/k_t$。因此，在 i采取行动1的条件下，玩家 i 拥有的作业延迟的概率等于

$$p^i(1,\boldsymbol{\sigma}_t^{-i}) = \frac{\mathbb{P}\left[a_j^i = 1\ \forall j \neq i\right]}{k_t} = \frac{\prod_{j\neq i}\sigma_t^j(1)}{k_t}.$$

此外，由于在纳什均衡中，玩家 i 对行动0和行动1无差异，我们有

$$c^i_{\boldsymbol{k}_t}(0,\boldsymbol{\sigma}_t^{-i}) = k_t^i k_t = c^i_{\boldsymbol{k}_t}(1,\boldsymbol{\sigma}_t^{-i}) = k_t^i\left(k_t - 1 + \frac{\prod_{j\neq i}\sigma_t^j(1)}{k_t}C_{k_t}\right). \tag{3.4}$$

因此，我们有

$$\prod_{j\neq i}\sigma_t^j(1)C_{k_t} = k_t. \tag{3.5}$$

这适用于任意满足 $\sigma_t^j(0) > 0$的 j。因此， $\sigma_t^i(1) = \sigma_t^j(1)$。

(ii) (a) 和 (iv) （第二步）在 *0* 上进行混合的参与者不会在严格大于1的任何行动上进行混合。如果玩家 i 在行动 $a_t^i > 1$上进行混合，那么当所有其他参与者都采取 1 时， i 的某项工作会延迟。这意味着采取行动 $a_t^i > 1$ 时发生延迟的概率，大于所有其他参与者都采取 1 的概率，因此利用（3.5）可得

$$c^i_{\boldsymbol{k}_t}(a_t^i,\boldsymbol{\sigma}_t^{-i}) \geqslant k_t^i\left\{k_t - a^i + \prod_{j\neq i}\sigma_t^j(1)C_{k_t}\right\} = k_t^i k_t + k_t^i(k_t - a_t^i) > k_t^i k_t.$$

由于所有在 0 上进行混合的参与者都只在 0 和 1 之间进行混合，并且在 1 上具有相同的权重，因此他们实际上采取相同的混合行动，这证明了 (iv)。

(v) 社会成本。 首先，我们证明社会成本小于 k_t^2。如果 $\boldsymbol{\sigma}_t$ 是一个纳什均衡，那么对于每个玩家 j，我们都有

$$c^j_{\boldsymbol{k}_t}(\boldsymbol{\sigma}_t) \leqslant c^j_{\boldsymbol{k}_t}(0,\boldsymbol{\sigma}_t^{-j}) = k_j^i k_t. \tag{3.6}$$

因此，$\mathsf{SC}(\boldsymbol{\sigma}_t) \leqslant \sum_{j\in I} k_j^i k_t = k_t^2$。现在我们证明社会成本大于 $k_t^2 - k_t + k_t^i$。一名玩家 i 以 0 为策略进行混合，产生的成本等于 $k_t^i k_t$；其他所有玩家 $j \neq i$ 都以 1 为策略进行混合，因此他们的成本至少为 $k_j^i(k_t - 1)$。因此，

$$\mathsf{SC}(\boldsymbol{\sigma}_t) \geqslant (k_t - k_t^i)(k_t - 1) + k_t^i k_t = k_t^2 - k_t + 1. \qquad \square$$

推论 8。 对于任意纳什均衡 $\boldsymbol{\sigma}$，作业都有正概率会延迟，即 i s

$$\mathbb{E}[\text{number of late jobs}] > 0.$$

证明。 根据定理 7 (ii) (a)，我们知道存在一位玩家 i 以正概率在 $L - k_t$ 上混合。我们根据这一概率考虑两种情况。如果 i 以严格小于 1 的概率进行行动，那么根据定理 7 (iii) 和定理 7 (iv)，我们知道所有参与者都有正概率在 $L - k_t + 1$ 中同时行动。因此，其中一位会延迟。

假设 i 以概率 1 选择行动 $L - k_t$。如果两位参与者在 $L - k_t$ 上混合，我们可以分别对他们应用定理 7，因此根据 (iii)，两者都会对 $L - k_t + 1$ 赋予正权重。因此， i 是唯一选择 $L - k_t$ 的玩家，根据定理 7 (iv)，其他每位玩家都有正概率选择 $L - k_t + 1$。令 $j \neq i$，并假设 $\sigma_t^j(L - k_t + 1) = 1$。那么，存在另一位不同于 i 和 j 的 j'（因为 $I \geqslant 3$）。玩家 i 选择 $L - k_t$，j 选择 $L - k_t + 1$，因此 j' 在选择 $L - k_t + 2$ 时不会延迟。因此， j' 严格偏好 $L - k_t + 2$ 而非 $L - k_t + 1$，所以它不可能对 $L - k_t + 1$ 赋予正权重，这就产生了矛盾。

$\square$

3.2 当 $L < k_t$ 时的均衡

现在我们研究周期长度小于工作数量的博弈，即 $L < k_t$。

定理 9 （粗相关均衡）。 若 $k_t > L$ 且 $C_{k_t} > k_t^2 L$，则存在唯一的粗相关均衡；事实上，它是一个纯策略均衡，其中所有玩家都选择0。

为了证明任意粗相关均衡的支撑集都是单元素集合 0，我们采用反证法。假设存在一个支撑集不为 0 的粗相关均衡；我们对每个玩家单边偏离到 0 所产生的成本求和，并证明该总和为负。因此，至少有一个偏离成本为负，这表明该玩家存在有利偏离。这个证明并非“构造性”的，因为它并未直接指出哪位玩家会感到不满。

定理 *9* 的证明。 设 $\boldsymbol{\tau}_t$ 是一个粗相关均衡。我们有

$$c_{\boldsymbol{k}_t}^i(\boldsymbol{\tau}_t) = \sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) c_{\boldsymbol{k}_t}^i(\boldsymbol{a}) = \sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) k_t^i \big\{L - a^i + p^i(\boldsymbol{a}) C_{k_t}\big\}. \tag{3.7}$$

由于 $\boldsymbol{\tau}_t$ 是粗相关均衡， i 通过单方面偏离至 0 所获得的成本

不小于 i 在 $\boldsymbol{\tau}_t$下承担的成本，即:

$$\sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) k_t^i \big\{ L + p^i(0, \boldsymbol{a}^{-i}) C_{k_t} \big\} \geqslant \sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) k_t^i \big\{ L - a^i + p^i(\boldsymbol{a}) C_{k_t} \big\}, \tag{3.8}$$

这将导出

$$\sum_{\boldsymbol{a}\in A^I} \boldsymbol{\tau}_t(\boldsymbol{a}) k_t^i \{ p^i(0, \boldsymbol{a}^{-i}) C_{k_t} - a^i + p^i(\boldsymbol{a}) C_{k_t} \} \geqslant 0. \tag{3.9}$$

原本选择行动 0 的玩家实际上并未偏离。因此，求和中对应于选择行动 0 的玩家的项等于零。因此，在 (3.9) 中，我们可以对所有未选择行动 0 的玩家求和，得到

$$\sum_{\boldsymbol{a}\in A^I\setminus\{\mathbf{0}\}} \boldsymbol{\tau}_t(\boldsymbol{a}) \left\{ \sum_{i\in I} k_t^i (p^i(0, \boldsymbol{a}^{-i}) - p^i(\boldsymbol{a})) C_{k_t} + k_t^i a^i \right\} \geqslant 0. \tag{3.10}$$

在时间 0 之后加入队列的作业，延迟的概率更高:

$$p^i(\boldsymbol{a}) \geqslant p^i(0, \boldsymbol{a}^{-i}) \tag{3.11}$$

下面的论断对上述陈述作了进一步细化。

论断 10. 给定行动组合 $\boldsymbol{a} \neq \mathbf{0}$，若

$$a^j = \max_{i\in I}(a^i), \tag{3.12}$$

then

$$p^j(\boldsymbol{a}) \geqslant \frac{1}{N_{a^j} k_t} + p^j(0, a^j), \tag{3.13}$$

其中

$$N_{a^j} := \sum_{i\in I\,:\; a^i = a^j} k_t^i \tag{3.14}$$

是在时间 a^j 加入队列的作业数量。

证明。 首先注意到，在（3.12）中有 $a^j > 0$。记 M_{a^j} 为在时间 a^j 加入队列并在截止时间 L之前离开系统的作业数。由于 $L \leqslant k_t$，有些作业会延迟；更准确地说，在 a^j 加入队列的 $N_{a^j} - M_{a^j}$ 个作业会延迟。于是

$$p^j(\boldsymbol{a}) = \frac{N_{a^j} - M_{a^j}}{N_{a^j}}. \tag{3.15}$$

此外， $p^j(0,\ \boldsymbol{a}^{-j}) \leqslant (k_t - L)/k_t$, 因此

$$p^j(\boldsymbol{a}) - p^j(0, \boldsymbol{a}^{-j}) \geqslant \frac{N_{a^j} - M_{a^j}}{N_{a^j}} - \frac{k_t - L}{k_t} = \frac{-M_{a^j} k_t + L N_{a^j}}{N_{a^j} k_t}. \tag{3.16}$$

然而， 不延迟而离开系统的工作数量最多为 L。

在行动组合 $\boldsymbol{a}$ 下， $k_t - N_{a^j}$ 个工作严格早于时间 a^j 加入队列。为完成证明，我们考虑与 $k_t - N_{a^j}$ 的取值有关的几种情况:

- 若 $0 < k_t - N_{a^j} \leqslant L$，则在时间 a^j 加入队列的工作中，按时到达的至多有 $L - (k_t - N_{a^j})$ 个，即 $M_{a^j} \leqslant L - (k_t - N_{a^j})$。将此不等式代入（3.16），得到

$$\begin{aligned} p^j(\boldsymbol{a}) - p^j(0, \boldsymbol{a}^{-j}) &\geqslant \frac{-(L - k_t + N_{a^j})k_t + L N_{a^j}}{N_{a^j} k_t} \\ &= \frac{L}{k_t} - \frac{L - k_t}{N_{a^j}} - 1 \\ &= \{L - k_t\}\left\{\frac{1}{k_t} - \frac{1}{N_{a^j}}\right\} \\ &= \{k_t - L\}\frac{k_t - N_{a^j}}{k_t N_{a^j}} \\ &\geqslant \frac{1}{k_t N_{a^j}}, \end{aligned}$$

因为在这种情况下 $k_t - N_{a^j} \geqslant 1$。

- 若 $k_t - N_{a^j} = 0$，则所有工作都在大于 1 的时间加入队列；因此 $M_{a^j} \leqslant L - 1$。由于在这种情况下 $k_t = N_{a^j}$，经过一些化简后，（3.16）变为

$$p^j(\boldsymbol{a}) - p^j(0, \boldsymbol{a}^{-j}) \geqslant \frac{-M_{a^j} + L}{k_t} \geqslant \frac{-(L-1) + L}{k_t} = \frac{1}{k_t} \geqslant \frac{1}{k_t N_{a^j}}. \tag{3.17}$$

- 如果 $k_t - N_{a^j} > L$，那么没有任何工作能避免延迟，且 $M_{a^j} = 0$。因此，(3.16) 变为 s

$$p^i(\boldsymbol{a}) - p^i(0, \boldsymbol{a}^{-i}) \geqslant \frac{L}{k_t} \geqslant \frac{1}{k_t N_{a^j}}.$$

至此，断言的证明结束。 □

断言 11. 对于任意 $\boldsymbol{a} \neq \boldsymbol{0}$，我们有

$$-\frac{1}{k_t} \geqslant \sum_{i \in I} k_t^i (p^i(0, \boldsymbol{a}^{-i}) - p^i(\boldsymbol{a})). \tag{3.18}$$

证明。我们将参与者分成两组：其作业在时间 a^j 加入队列的参与者（定义为

式 (3.12)）以及其余参与者。我们有

$$\begin{aligned}\sum_{i\in I} k_t^i(p^i(0,\boldsymbol{a}^{-i})-p^i(\boldsymbol{a})) = &\sum_{i\in I\,:\,a^i\neq a^j} k_t^i(p^i(0,\boldsymbol{a}^{-i})-p^i(\boldsymbol{a}))\\ &+\sum_{i\in I\,:\,a^i=a^j} k^i(p^i(0,\boldsymbol{a}^{-i})-p^i(\boldsymbol{a})).\end{aligned} \tag{3.19}$$

第一项因（3.11）而非正，第二项因（3.13）而非正，因此得到

$$\sum_{i\in I} k_t^i(p^i(0,\boldsymbol{a}^{-i})-p^i(\boldsymbol{a})) \leqslant -\sum_{i\in I\,:\,a^i=a^j} k_t^i\frac{1}{N_{a^j}k_t}. \tag{3.20}$$

由（3.20）中的不等式可得

$$\sum_{i\in I} k_t^i(p^i(0,\boldsymbol{a}^{-i})-p^i(\boldsymbol{a})) \leqslant -\frac{N_{a^j}}{N_{a^j}k_t} = -\frac{1}{k_t},$$

这就证明了该断言。 □

结合断言 11 和式（3.10）可得

$$\begin{aligned}0 &\leqslant \sum_{\boldsymbol{a}\in A^I\setminus\{\mathbf{0}\}} \boldsymbol{\tau}_t(\boldsymbol{a})\left\{\sum_{i\in I} k_t^i(p^i(0,\boldsymbol{a}^{-i})-p^i(\boldsymbol{a}))C_{k_t}+k_t^i a^i\right\}\\ 0 &\leqslant \sum_{\boldsymbol{a}\in A^I\setminus\{\mathbf{0}\}} \boldsymbol{\tau}_t(\boldsymbol{a})\left\{-\frac{1}{k}C_{k_t}+\sum_{i\in I} k_t^i a^i\right\}\\ 0 &\leqslant \sum_{\boldsymbol{a}\in A^I\setminus\{0\}} \boldsymbol{\tau}_t(\boldsymbol{a})\left\{-\frac{1}{k_t}C_{k_t}+Lk_t\right\}.\end{aligned} \tag{3.21}$$

根据定理的假设， $C_{k_t} > k^2 L$，因此

$$-\frac{1}{k_t}+Lk_t<0. \tag{3.22}$$

因此，（3.21）只有在对所有 $\boldsymbol{a}\in A^I\setminus\{\mathbf{0}\}$ 都满足 $\boldsymbol{\tau}_t(\boldsymbol{a})=0$ 时才成立，这意味着 $\boldsymbol{\tau}_t$ 的支撑集为 0。 □

3.3 全局队列

以下定理是前述结果的推论。

定理 12 （策略性重复的稳定性）。 考虑一个具有 $I \leqslant L$ 的 *DQM*。若 下确界$_k\, C_k$ $>(L+I)^2 L$, 则任何短视粗相关均衡都具有稳定性。

推论 13（策略性重复的稳定性）。 考虑一个具有 $I \leqslant L$ 的 *DQM*。若 下确界$_k\, C_k$ $>(L+I)^2 L$, 则任何短视纳什均衡都具有稳定性。

特别要注意的是，如果惩罚成本是常数 $C > (L+I)^2 L$，那么任何短视粗相关均衡都是稳定的。该推论是定理的直接结果，因为任何纳什均衡也是粗相关均衡。

正如定理 12 的证明中将会清楚看到的，队列会在两种可能的状态之间交替。第一种状态对应于 $k_t \leqslant L$。在这种状态下，阶段均衡具有非平凡的结构，而且在均衡中，有些作业可能会延迟。因此，系统中的工作数量 k_t 可能随时间波动。当这一数量超过水平 L时，系统便进入另一种状态，对应于 $k_t > L$。在这种状态下，唯一的阶段均衡是纯策略组合 0。因此，如果 $I = L$，系统中的工作数量 k_t 保持不变；如果 $I < L$，工作数量就会减少，直到系统回到第一种状态。

注释 14. 如果惩罚成本较低，则可能存在不稳定的短视纳什均衡。例如，如果对于所有 $k \in \mathbb{N}$、$C_k < 1$，抢占其他参与者的成本都远高于潜在收益，那么在时期的最后阶段行动对每个参与者而言都是严格占优的。唯一的短视纳什均衡是每个参与者都等到时期的最后阶段行动，并且有 $k_t - 1$ 个延迟任务。因此， k_t 几乎必然无界。

定理 *12* 的证明。 假设 $I \leqslant L$，且对于所有 k、 $C_k > (L+I)^2 L$。我们通过归纳法证明，对于每个 t，都有 $k \leqslant L + I$。

该结果在时期 1 成立。现在我们证明它在每个时期都成立。从时期 t 到时期 $t+1$，我们需要考虑两种情况。如果 $k_t \leqslant L$，那么 $k_{t+1} \leqslant L + I$，因为在下一时期I 新任务到达，且至多有 k_t 个延迟任务。如果 $L < k_t \leqslant L + I$，那么

$$C_{k_t} > (L+I)^2 L \geqslant k_t^2 L. \tag{3.23}$$

根据定理 9，存在唯一的阶段粗相关均衡，即所有玩家都选择 0。这意味着至少有 L 项工作离开系统。由于 $I \leqslant L$，且下一时期有 I 个新任务，因此我们有$k_{t+1} \leqslant k_t \leqslant L + I$。

□

4 学习型玩家

本节研究这样一种模型：每位玩家都独立地采用无悔策略。无遗憾性质最初由 Hannan（1957）提出，是在线强化学习中多种算法所具有的一种性质（例如参见 Perchet，2014）。它指出，事后看来，玩家采取的行动至少（渐近地）与任何固定行动一样好。形式上，在单人情形下，给定一列按时间 t 索引的成本函数 l_t 以及一列

行动 a_t^i，玩家 i 的遗憾定义为

$$R_t^i = \max_{b^i \in A} \sum_{u=1}^{t} l_u(b^i) - l_u(a_u^i). \tag{4.1}$$

若 $R_t^i = o(t)$，则策略满足无悔 性质。著名的无悔策略包括遗憾匹配（Hart 和 Mas-Colell，2013）、随机虚拟博弈（Fudenberg 和 Levine，1998）以及指数权重算法（EWA）（Littlestone 和 Warmuth，1994；Cesa-Bianchi 和 Lugosi，2006），我们将在下文研究这些策略。

指数权重算法 以下我们使用指数权重算法（EWA）。已知当收益有界时，该算法能够保证无悔。遗憾的是，此处不满足有界性，因为系统中的工作数量可能无限增长，从而导致惩罚无界。因此，我们需要更仔细地研究该算法的效率。

在重复博弈的背景下，权重通常定义如下：

$$w_t^i(b^i) = \exp\left(\sum_{u=1}^{t-1} -\eta c^i(b^i, \boldsymbol{a}_u^{-i})\right), \tag{4.2}$$

其中 η 是一个正数常量。式（4.2）可以递归地改写为

$$w_{t+1}^i(b^i) = w_t^i(b^i) \exp\left(-\eta c^i(b^i, \boldsymbol{a}_t^{-i})\right). \tag{4.3}$$

Th时，EWA 策略规定在时间 t以概率选择行动 b^i y

$$x_t^i(b^i) = \frac{w_t^i(b^i)}{\sum_{a^i \in A} w_t^i(a^i)}. \tag{4.4}$$

多层指数权重算法（MLEWA）EWA 并非为应对不断变化的环境而设计。在此情境中，每位参与者持有的工作数量会随时间变化。因此，我们需要说明如何利用这类信息。我们设计了一种名为 MLEWA 的新协议，其中每位参与者都使用多个 EWA 副本。该协议由参数 n 索引，我们称其为层级。当参与者 i 拥有的工作数量首次达到一个新的层级时，该参与者会启动一个新的 EWA，并根据过去的信息初始化权重。当参与者 i 的工作数量达到一个已访问过的层级时，该参与者会遵循该层级的 EWA 所给出的建议，并按照 EWA 更新权重。请注意，在任一给定时期 t，不同参与者可能使用不同层级的 EWA。

令 τ_n^i 为玩家 i 拥有 n 份工作的首次时间；约定若玩家 i 从未拥有过 n 份工作，则 $\tau_n^i = +\infty$。玩家 i 的权重 $w_{t,\ n}^i(b^i)$ 现在由两个参数确定：时期 t 和

算法1： 多层指数权重算法（MLEWA）

$\forall b^i \in A$，初始化 $w_1^i(b^i)$;
$\forall i \in I$，$k^i \leftarrow 1$；// 第1级
对于 每个步骤 $t \geqslant 1$ **执行**
  对每个玩家 i **执行**
    选择一个行动 $a_t^i \sim x_{k^i}^i$；// 与 $w_{k^i}^i$ 成正比
  end
  对每个玩家 i **执行**
    $\forall b^i \in A, w_{k^i}^i(b^i) \leftarrow w_{k^i}^i(b^i)\exp(-\eta c_{\boldsymbol{k}_t}^i(b^i, \boldsymbol{a}_t^{-i}))$；// 延迟作业数量 + 1
    $k^i \leftarrow k^i + 1$;
    如果$w_{k^i}^i$ 未定义，**则**
      $\forall b^i \in A$，初始化 $w_{k^i}^i(b^i)$；// 层级 k^i
    end
  end
end

层级 n。层级 n 的算法对所有 $t \geqslant \tau_n^i$ 均通过归纳法定义。我们先介绍归纳步骤，其定义如下：

$$w_{t+1,n}^i(b^i) = \begin{cases} w_{t,n}^i(b^i)\exp\left(-\eta c_{\boldsymbol{k}_t}^i(b^i, \boldsymbol{a}_t^{-i})\right) & \text{if } k_t^i = n, \\ w_{t,n}^i(b^i) & \text{otherwise.} \end{cases} \tag{4.5}$$

公式（4.5）意味着，$w_{t,\ n}^i(b^i)$ 当且仅当 $k_t^i = n$ 时更新。

现在我们描述初始化过程。在 τ_n^i，我们启动一个新的 EWA 协议并定义初始权重

$$w_{\tau_n^i,n}^i := w_{\tau_n^i,k_{\tau_n^i-1}^i}^i,$$

也就是说，新遇到的状态的权重定义为与此前访问层级的权重相等。

算法1总结了我们整个过程的所有步骤。由于变量数量固定，且算法不会访问完整历史，因此变量不以 t 为索引。相反，它会在每个步骤计算新值，并更新相应的变量。

MLEWA 基于一种适用于变化状态的无悔算法。公式（4.1）并未处理变化状态。这是 Gaitonde 和 Tardos 明确指出的一项局限：计算遗憾时假定状态路径不会因行动变化而改变。当状态内生变化时，情况会复杂得多。关于我们的算法，我们能说的是，它在每个状态下都具有无遗憾性质。

我们可以 现在陈述关于参与者学习时系统稳定性的主要结果 .

定理 15 （亚临界情形下联合无悔策略的稳定性）。 如果 $I < L$ 且 $C_k > 4kL$对所有 k_t都成立，那么所有参与者均使用 *MLEWA* 的策略组合是稳定的。

证明定理 15 需要几个引理。首先，我们证明，在静态情境下，对于持有足够多工作的一名参与者，动作0严格优于其他任何行动。这一支配关系在我们的动态模型中的含义是：当一名参与者持有足够多工作时，动作0的权重会趋近于1。最后，我们介绍一些关于强化随机游走的结果。

4.1 通过动作0实现支配0

下面的引理表明，在静态情形下，对于拥有足够多作业的玩家 i ，动作0严格占优。

引理 16 （静态模型中当 $k_t^i > 2L^2$ 时由0严格支配）。若 $k_t^i > 2L^2$， $C_{k_t} > 4k_tL$，且 $\boldsymbol{a}$ 是一个行动组合，使得 $a^i \neq 0$，则 $c^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) < c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) - k_t^i$。

证明。设 $\boldsymbol{a}$ 是一个纯行动组合，满足 $a^i \neq 0$。记

$$k_t^{-i}(0) = \sum_{j \neq i} k_j^i 1_{a^j=0} \tag{4.6}$$

在动作0处加入队列的工作数量，不包括玩家 i 的作业。

于是:

$$c^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) = k_t^i \left\{ L + \frac{k_t^i + k_t^{-i}(0) - L}{k_t^i + k_t^{-i}(0)} C_{k_t} \right\} = k_t^i \left\{ L + \left\{ 1 - \frac{L}{k_t^i + k_t^{-i}(0)} \right\} C_{k_t} \right\}.$$

对于每项作业， i需要承担等待成本 L ，以及因延迟概率而产生的额外成本。在时期 0，$k_t^i + k_t^{-i}(0) > L$ 项作业加入队列；因此，有些作业必然会延迟。此外，某项作业不承担惩罚的概率，等于该作业加入队列时排在前 L 项作业之内的概率，即 $L/(k_t^i + k_t^{-i}(0))$。

- 如果 $k_t^{-i}(0) \geqslant L$，那么在配置 $\boldsymbol{a}$ 下，队列从阶段 0 起就已满，由 i发送的作业都会延迟，即 $c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) = k_t^i(L - a^i) + k_t^i C_{k_t}$。于是

$$c^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) = -\frac{k_t^i L}{k_t^i + k_t^{-i}(0)} C_{k_t} + k_t^i a^i.$$

关于 C_{k_t} 的假设意味着 $C_{k_t} > k_t(L+1)/L,$ ，因此:

$$c^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) \leqslant -\frac{k_t^i}{k_t^i + k_t^{-i}(0)} k_t(L+1) + k_t^i a^i.$$

由于根据定义， $k_t \geqslant k_t^i + k_t^{-i}(0)$，因此:

$$c^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) \leqslant -k_t^i(L+1) + k_t^i a^i \leqslant k_t^i(a^i - L) - k_t^i \leqslant -k_t^i.$$

• 现在考虑情况 $k_0^{-i} < L$。玩家 i 为每项作业支付等待成本 $L - a^i$；玩家 i的作业中，至多有 $L - a^i$ 项可以不延迟地离开系统。因此，以下界限成立：

$$c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) \geqslant k^i_t(L - a^i) + (k^i_t - L + a^i)C_{k_t}.$$

Then

$$\begin{aligned} c^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) &= k^i_t L + k^i_t C_{k_t} - \frac{L k^i_t}{k^i_t + k_t^{-i}(0)} C_{k_t} \\ &\quad - k^i_t L + k^i_t a^i - (k^i_t - L + a^i)C_{k_t}, \\ &\leqslant -\frac{k^i_t L}{k^i_t + k_t^{-i}(0)} C_{k_t} + k^i_t a^i + L C_{k_t} - a^i C_{k_t}, \\ &= \left(\frac{k_t^{-i}(0) L}{k^i_t + k_t^{-i}(0)} - a^i\right) C_{k_t} + k^i_t a^i. \end{aligned} \tag{4.7}$$

由于 $k^i_t > 2L^2$，可得 $k^i_t + k_t^{-i}(0) \geqslant 2L^2$，因此

$$\frac{k_t^{-i}(0) L}{k^i_t + k_t^{-i}(0)} \leqslant \frac{L^2}{2L^2} < \frac{1}{2}.$$

因此，

$$c^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) \leqslant \left(\frac{1}{2} - a^i\right) C_{k_t} + k^i a^i < -\frac{C_{k_t}}{2} + k^i_t L,$$

因为 $1 \leqslant a^i \leqslant L$。

假设 $C_{k_t} > 4kL$，则意味着

$$c^i_{\boldsymbol{k}_t}(0, \boldsymbol{a}^{-i}) - c^i_{\boldsymbol{k}_t}(\boldsymbol{a}) < -2k_t L + k^i_t L < -k_t L < -k_t < -k^i_t. \qquad \square$$

4.2 工作数量增加时，动作 0 越来越受青睐

现在我们使用引理 16 来证明，当系统中的工作数量足够多时，对于按式 (4.4) 定义的策略 $x^i_{t,\ n}$，动作0的权重 $w^i_{t,\ n}(0)$ 比其他权重增长得更快。由于 $x^i_{t,\ n}$ 与该权重成正比，因此，当状态再次被访问时，玩家越来越倾向于选择动作0。

截至时间 $t-1$，玩家 i 的等级为 n 的次数：

$$\lambda_{t,n} = \#\big\{k^i_u = n \mid u \in \{0, \ldots, t-1\}\big\}. \tag{4.8}$$

我们有

引理 17（偏好 0）。 如果 $n > 2L^2$ 且 $C_{k_t} > 4kL$，那么对于所有 $t \geqslant \tau_n^i$，

$$x_{t,n}^i(0) \geqslant \frac{x_{\tau_n^i,n}^i(0)}{x_{\tau_n^i,n}^i(0) + \left(1 - x_{\tau_n^i,n}^i(0)\right) \exp\left(-\eta n \lambda_{t,n}\right)}.$$

证明。 我们首先对 t 进行归纳证明：对于每个 $b^i \neq 0$，都有：

$$\frac{w_{t,n}^i(b^i)}{w_{\tau_n^i,n}^i(b^i)} \leqslant \exp\left(-\eta n \lambda_{t,n}\right) \frac{w_{t,n}^i(0)}{w_{\tau_n^i,n}^i(0)}, \tag{4.9}$$

如果 $t = \tau_n^i$，那么根据定义，玩家 i 之前从未有过 n 个作业，并且 $\lambda_{t,\ n} = 0$。

因此，等式两边均为 1，结论成立。

现在我们证明，如果结论对 $t-1$成立，那么对 t也成立。分两种情况讨论。

如果 $k_{t-1}^i \neq n$，那么在阶段 t 和 $t-1$，所有权重都相等，即对于所有行动 b^i，都有 $w_{t,\ n}^i(b^i) = w_{t-1,\ n}^i(b^i)$。此外，$\lambda_{t,\ n} = \lambda_{t-1,\ n}$，因此在 t 和 $t-1$ 处的不等式相同，结论成立。如果 $k_{t-1}^i = n$，那么 $\lambda_{t,\ n} = \lambda_{t-1,\ n} + 1$。根据递推假设，我们知道

$$\frac{w_{t-1,n}^i(b^i)}{w_{\tau_n^i,n}^i(b^i)} \leqslant \exp\left(-\eta n \lambda_{t-1,n}\right) \frac{w_{t-1,n}^i(0)}{w_{\tau_n^i,n}^i(0)}. \tag{4.10}$$

利用引理 16 和公式（4.5），对于 $b^i \neq 0$，我们得到

$$\begin{aligned}
\frac{w_{t,n}^i(b^i)}{w_{t-1,n}^i(b^i)} &= \exp\left(-\eta c_{\boldsymbol{k}}^i(b^i, \boldsymbol{a}_{t-1}^{-i})\right) && (4.11)\\
&\leqslant \exp\left(-\eta n\right) \exp\left(-\eta c_{\boldsymbol{k}}^i(0, \boldsymbol{a}_{t-1}^{-i})\right), && (4.12)\\
&= \exp(-\eta n) \frac{w_{t,n}^i(0)}{w_{t-1,n}^i(0)}, && (4.13)
\end{aligned}$$

将式（4.10）和式（4.13）相乘，可得 t 的结果。

现在固定 t。由式（4.9）可知

$$w_{t,n}^i(b^i) \leqslant \exp\left(-\eta n \lambda_{t,n}\right) \frac{w_{t,n}^i(0)}{w_{\tau_n^i,n}^i(0)} w_{\tau_n^i,n}^i(b^i).$$

因此，

$$
\begin{aligned}
x_{t,n}^i(0) &= \frac{w_{t,n}^i(0)}{\sum_{b^i\in A} w_{t,n}^i(b^i)} \\
&= \frac{w_{t,n}^i(0)}{w_{t,n}^i(0) + \sum_{b^i\neq 0} w_{t,n}^i(b^i)} \\
&\geqslant \frac{w_{t,n}^i(0)}{w_{t,n}^i(0) + \sum_{b^i\neq 0} \exp\left(-\eta n \lambda_{t,n}\right) \frac{w_{t,n}^i(0)}{w_{\tau_n^i,n}^i(0)} w_{\tau_n^i,n}^i(b^i)} \\
&= \frac{w_{\tau_n^i,n}^i(0)}{w_{\tau_n^i,n}^i(0) + \sum_{b^i\neq 0} \exp\left(-\eta n \lambda_{t,n}\right) w_{\tau_n^i,n}^i(b^i)} \\
&= \frac{x_{\tau_n^i,n}^i(0)}{x_{\tau_n^i,n}^i(0) + \sum_{b^i\neq 0} \exp\left(-\eta n \lambda_{t,n}\right) x_{\tau_n^i,n}^i(b^i)} \\
&= \frac{x_{\tau_n^i,n}^i(0)}{x_{\tau_n^i,n}^i(0) + \left(1 - x_{\tau_n^i,n}^i(0)\right) \exp\left(-\eta n \lambda_{t,n}\right)},
\end{aligned}
$$

这就证明了该引理。 □

引理 18。 存在 $B^i > 0$ 使得

$$
x_{\tau_n^i,n}^i(0) \geqslant \frac{1}{1 + B^i \exp(-\eta n)}. \tag{4.14}
$$

证明。 令 $i \in I$。我们可以定义

$$
Z^i := \max_{n\leqslant 2L^2 \text{ s.t. } \tau_n^i<+\infty} \frac{1}{x_{\tau_n^i,n}^i(0)} - 1.
$$

对于 $n = 1$，我们将算法均匀初始化，使每个行动的初始权重都严格为正。根据 EWA 的定义，如果某个行动在初始化期间的权重严格为正，那么它始终以严格为正的概率被选择。当到达层级 $n = 2$ 时，通过复制层级 $n = 1$ 的算法当前分布来完成初始化，因此每个行动的权重也都严格为正。由归纳法可知，在每个阶段、每个层级，选择每个行动的概率都严格为正且严格小于 1。因此，Z^i 作为有限个严格为正数中的最小值，也严格为正。

根据定义，对于所有满足 $\tau_n^i < +\infty$ 的 $n \leqslant 2L^2$，都有

$$
x_{\tau_n^i,n}^i \geqslant \frac{1}{1 + Z^i} \geqslant \frac{1}{1 + Z^i \exp(\eta 2L^2) \exp(-\eta n)},
$$

因此令 $B^i := Z^i \exp(\eta 2L^2)$。

现在我们证明，这对于 $n > 2L^2$ 也成立。证明采用归纳法。假设它对于 $n > 2L^2$ 成立，并且考虑 ，使得 。

由于作业数量的增量至多为一，这意味着 $\tau_n^i < +\infty$。

根据引理 17，对于所有 $t > \tau_n^i$，0 上的权重满足

$$\begin{aligned} x_{t,n}^i(0) &\geqslant \frac{x_{\tau_n^i,n}^i(0)}{x_{\tau_n^i,n}^i(0) + \left(1 - x_{\tau_n^i,n}^i(0)\right)\exp\left(-\eta n \lambda_{t,n}\right)} \\ &= \frac{1}{1 + \left(\frac{1}{x_{\tau_n^i,n}^i(0)} - 1\right)\exp\left(-\eta n \lambda_{t,n}\right)} \\ &\geqslant \frac{1}{1 + (1 + B^i\exp(-\eta n) - 1)\exp\left(-\eta n \lambda_{t,n}\right)} \qquad (4.15) \\ &\geqslant \frac{1}{1 + B^i\exp\left(-\eta n(1 + \lambda_{t,n})\right)} \\ &\geqslant \frac{1}{1 + B^i\exp\left(-\eta(n+1)\right)} \qquad (4.16) \end{aligned}$$

其中，式（4.15）由递推假设得出，而式（4.16）则由 $t > \tau_n^i$ 推出，因此$\lambda_{t,\ n} \geqslant 1$。

首次到达 $n+1$ 时的初始权重等于当前 n个包的权重，因此

$$x_{\tau_{n+1}^i,n+1}^i(0) = x_{\tau_{n+1}^i,n}^i(0) \geqslant \frac{1}{1 + B^i\exp(-\eta(n+1))}. \qquad (4.17)$$

这证明了 $n+1$ 的情形。因此，归纳过程得以完成，并证明了该引理。 □

4.3 强化随机游走

引理 17 表明，每当该过程到达某个给定水平时，选择所有玩家都采取 0 的行动组合的概率就会得到强化。接下来，我们要了解这种强化如何影响系统动态。为此，我们证明一些关于强化随机游走的结果。我们遵循 Menshikov 等（2017，第 47 页）对一维最近邻随机游走的论述。他们研究的是在空间上非齐次、但在时间上齐次的随机游走。不同之处在于，我们假设漂移中存在一个强化因子，因此得到的随机游走在时间和空间上均非齐次，但有界。此外，我们假设随机游走的权重具有更强的异质性，也就是说，向上或向下移动的确切概率高度依赖于过去，但仍然有界。

定理 15 的证明需要以下引理，其证明见附录 B。

引理19 (强化随机游走)。 设 $M > 0$ 和 $(X_t,\ Z_t,\ t \geqslant 0)$ 是 $\mathbb{N}$ 中的一列随机变量，使得 $X_t/d \leqslant Z_t \leqslant X_t$，其中 $d > 1$、$|X_{t+1} - X_t| \leqslant M$ 和 $\mathcal{F}_t = \sigma(X_0,\ Z_0、\ldots\ldots、\ Z_t,\ X_t)$。假设存在函数 $r:\ \mathbb{N}^2 \to \mathbb{R}^+$、实数 z_0 和 $A > 0$，使得:

- 对所有 $t \geqslant 0$ 和 $Z_t \geqslant z_0$，$\mathbb{P}\left[X_{t+1} > X_t \mid \mathcal{F}_t\right] \leqslant r(Z_t,\ \lambda_{t,\ z_t})$ 几乎必然，其中 $\lambda_{t,\ z}$ 是事件 $Z_t = z$ 对于 $t' \leqslant t$ 的发生次数，
- 对于 $z \geqslant z_0$，$\sum_m r(z,\ m) < \frac{A}{z}$

那么 X_t 几乎必然有界。

定理 *15* 的证明。令 $X_t := Ik_t$ 和 $Z_t := \max_{j \in I}\left(Ik_t^j + j\right)$ 为随机变量，它们是 k_t, k_t^j, j 的一种编码，其中 j 使 k_t^j 最大，并且在所有最大化者中取最大值。事实上，$k_t = X_t/I$, $k_t^j = [Z_t/I]$ 且 $j = Z_t \bmod I$。

根据定义，

$$Z_t = \max_{j \in I}\left(Ik_t^j + j\right) \leqslant \max_{j \in I}\left(I(k_t - 1) + j\right) \leqslant Ik_t = X_t. \tag{4.18}$$

此外，

$$Z_t \geqslant \max_{j \in I} Ik_t^j \geqslant I \max_{j \in I} k_t^j \geqslant I\frac{k_t}{I} = \frac{X_t}{I}. \tag{4.19}$$

令 $z_0 = 2IL^2$。假设 $Z_t \geqslant z_0$，并令 j，使得 $Z_t = Ik_t^j + j$。下文中，我们用 k^j 表示 k_t^j，用 τ^j 表示 $\tau_{k^j}^j$。那么，j 取 0 的概率为 $x_{t,\ k^j}^j(0)$ 根据引理 17，该概率满足

$$x_{t,k^j}^j(0) \geqslant \frac{x_{\tau^j,k^j}^j(0)}{x_{\tau^j,k^j}^j(0) + \left(1 - x_{\tau^j,k^j}^j(0)\right)\exp\left(-\eta k^j \lambda_{k^j,t}\right)} \tag{4.20}$$

$$= \frac{1}{1 + \left(\frac{1}{x_{\tau^j,k^j}^j(0)} - 1\right)\exp\left(-\eta k^j \lambda_{k^j,t}\right)}. \tag{4.21}$$

此外，根据引理 18，存在 $B > 0$，使得 $x_{\tau^j,\ k^j}^j(0) \geqslant {}_{1+B}{}^{1}{}_{\exp(-\eta k^j)}$，因此由式 (4.21) 可得

$$x_{t,k^j}^j(0) \geqslant \frac{1}{1 + B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{k^j,t})}. \tag{4.22}$$

每个时期都有 I 个新任务。根据假设 3，新任务少于 L 个。由于 j 的工作多于 $[z_0/I] = 2L^2$ 个，当 j 选择 0 时，我们知道至少有 L 个作业没有延迟。因此，下一个时期的工作数量必定小于或等于当前时期。于是，

$$\mathbb{P}\left[X_{t+1} > X_t \mid \mathcal{F}_t\right] \leqslant 1 - x_{t,k^j}^j(0) \leqslant \frac{B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{k^j,t})}{1 + B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{k^j,t})} \tag{4.23}$$

使用公式 (4.22)。

这提示了以下定义，

$$r(Z_t, \lambda_{Z_t,t}) := B\exp(-\eta[Z_t/I])\exp(-\eta[Z_t/I]\lambda_{Z_t,t}) \tag{4.24}$$

它等于

$$B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{Z_t,t}),$$

因为 $j = Z_t$ 模 I, $k^j = [Z_t/I]$ 因此，$\lambda_{Z_t,\ t}$ （ Z_t 等于当前值的次数）小于 $\lambda_{k^j,\ t}$ （ j 具有当前作业数的次数）。因此，

$$\exp(-\eta k^j \lambda_{k^j,t}) \leqslant \exp(-\eta k^j \lambda_{Z_t,t}). \tag{4.25}$$

利用前面的方程，

$$\begin{aligned}
\mathbb{P}\left[X_{t+1} > X_t \mid \mathcal{F}_t\right] &\leqslant B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{k^j,t}) && (4.26)\\
&\leqslant B\exp(-\eta k^j)\exp(-\eta k^j \lambda_{Z_t,t}) && (4.27)\\
&\leqslant r(Z_t, \lambda_{Z_t,t}) && (4.28)
\end{aligned}$$

其中，4.26 来自方程（4.23），4.27 来自方程（4.25），4.28 来自方程（4.24）).

对于所有 $z \geqslant z_0$，关于 m 的 $r(z,m)$ 之和为

$$B\exp(-\eta[z/I])\frac{1}{1-\exp(-\eta[z/I])},$$

因此它是有界的 由 A/z 和某个 $A > 0$ 界定，且引理19适用，因此我们证明了该定理 . □

5 结论

我们研究了一个存在跨时期溢出效应的重复策略性排队模型。我们重点关注参与者采用学习策略时系统的稳定性。该模型中仍有若干问题尚待解决。

多层遗憾我们使用了一种多层指数权重算法。尽管 MLEWA 基于一种无悔算法，但要证明它本身是一种无悔算法，我们需要为内生变化状态下的遗憾给出恰当的定义。采用合适的遗憾定义来证明 MLEWA 的无遗憾性质，并进而用其证明系统稳定性，将是一个有趣的推广。另一个有前景的研究方向，是基于不同的无悔算法定义其他多层算法。特别值得研究的是，哪些稳定性性质取决于所使用的具体算法，哪些其他性质则具有普遍性，适用于

每一种无悔算法。

模型模型还有多种可能的扩展。例如，可以研究包含多个服务器的模型。在这种情况下，每个玩家的策略由两个部分组成：选择哪台服务器，以及选择何时将工作加入所选服务器的队列。一个看似简单、实则并非易事的推广，是考虑较低的惩罚成本。

惩罚值的重要性 C_k 本文的若干结果基于惩罚值C_k。如果惩罚值足够大，系统便具有稳定性。我们猜想，在学习情境中，较大但固定的惩罚不足以保证稳定性；也就是说，惩罚必须取决于系统中的工作数量，否则工作数量可能以正概率无界增长。这与短视策略情形形成对比：在该情形下，只要惩罚成本足够大，固定的惩罚成本就能保证稳定性。

致谢

Marco Scarsini 和 Xavier Venel 是 GNAMPA-INdAM 的成员。Lucas Baudin 访问路易斯大学期间完成了本研究的部分工作。本研究项目获得了意大利教育、大学与研究部 PRIN 2017 项目 ALGADIMAR“算法、博弈与数字市场”的部分资助。Lucas Baudin 感谢 COST 行动 GA MENET CA16228 提供的经费支持。Xavier Venel 感谢国家研究署 CIGNE 项目（ ANR-15-CE38-0007-01）提供的经费支持。

参考文献

Anagnostides, I., Panageas, I., Farina, G., and Sandholm, T. (2023). 时变博弈中无悔学习动态的收敛性。技术报告，arXiv:2301.11241。Besbes, O., Gur, Y., and Zeevi, A. (2015). 非平稳随机优化。《运筹学》, 63(5):1227–1244。Besbes, O., Gur, Y., and Zeevi, A. (2019). 非平稳奖励下多臂老虎机问题中的最优探索-利用。*Stoch. Syst.*, 9(4):319–337。Cardoso, A. R., Abernethy, J., Wang, H., and Xu, H. (2019). 在对抗性变化的零和博弈中与纳什均衡竞争。载于 Chaudhuri, K. and Salakhutdinov, R. 编，第*36*届国际机器学习会议论文集，第97卷《机器学习研究论文集》，第921–930页。PMLR。Cesa-Bianchi, N. and Lugosi, G. (2006). 《预测、学习与博弈》。剑桥大学出版社，英国剑桥。

Crippa, L., Gur, Y., and Light, B. (2022). 重复博弈中采用动态基准的遗憾最小化。技术报告，arXiv:2212.03152。Duvocelle, B., Mertikopoulos, P., Staudigl, M., and Vermeulen, D. (2022). 时变博弈中的多智能体在线学习。《数学运筹与研究》，即将出版。Fudenberg, D. and Levine, D. K. (1998). 《博弈中的学习理论》。麻省理工学院出版社，马萨诸塞州剑桥。Gaitonde, J. and Tardos, E. (2020). 策略性排队系统中的稳定性与学习。载于第*21*届*ACM*经济学与计算会议论文集，EC '20，第319–347页，美国纽约州纽约市。计算机协会。Gaitonde, J. and Tardos, E. (2021). 策略性排队系统中耐心的优势。载于第*22*届*ACM*经济学与计算会议，EC '21，第520–540页，美国纽约州纽约市。计算机协会。Glazer, A. and Hassin, R. (1983). ?$/M/1$：关于顾客到达的均衡分布。《欧洲运筹学杂志》, 13(2):146–150。Hannan, J. (1957). 重复对局中贝叶斯风险的近似。载于《博弈理论文集》第*3*卷，《数学研究年刊》第39号，第97–139页。普林斯顿大学出版社，新泽西州普林斯顿。Hart, S. and Mas-Colell, A. (2013). 《简单自适应策略：从遗憾匹配到非耦合动态》。世界科技出版公司，新泽西州哈肯萨克。Hassin, R. (1985). 关于先到后服务队列最优性的研究。《计量经济学》, 53(1):201–202。Hassin, R. (2016). 《理性排队》。CRC出版社，佛罗里达州博卡拉顿。Hassin, R. and Haviv, M. (2003). 《排队还是不排队：排队系统中的均衡行为》。克鲁维尔学术出版社，马萨诸塞州波士顿。Haviv, M. and Ravner, L. (2021). 关于到达时机具有策略性的排队系统综述。《排队系统》, 99(1-2):163–198。Kawasaki, R., Konishi, H., and Yukawa, J. (2023). 瓶颈博弈中的均衡。《国际博弈理论杂志》，即将出版。Littlestone, N. and Warmuth, M. K. (1994). 加权多数算法。《信息与计算》, 108(2):212–261。Menshikov, M., Popov, S., and Wade, A. (2017). 《非齐次随机游走》。剑桥大学出版社，英国剑桥。

Mertikopoulos, P. 和 Staudigl, M.（2021）。动态博弈中的均衡跟踪与收敛。载于*2021*年第*60*届*IEEE*决策与控制会议（*CDC*），第930–935页。Naor, P.（1969）。通过征收通行费调节队列规模。《计量经济学》，37(1)：15–24。Pemantle, R.（2007）。强化随机过程综述。《概率调查》，4：1–79。Perchet, V.（2014）。可接近性、遗憾与校准：影响与等价性。《动态博弈杂志》，1(2)：181–254。Rivera, T.、Scarsini, M. 和 Tomala, T.（2018）。瓶颈博弈中相关性的效率。技术报告，SSRN3219767。Sentenac, F.、Boursier, E. 和 Perchet, V.（2021）。在线排队系统中的去中心化学习。《神经信息处理系统进展》，34：18501–18512。Vickrey, W. S.（1969）。拥堵理论与交通投资。《美国经济评论》，59(2)：251–260。Zhang, M.、Zhao, P.、Luo, H. 和 Zhou, Z.-H.（2022）。时变零和博弈中的无悔学习。载于 Chaudhuri, K.、Jegelka, S.、Song, L.、Szepesvari, C.、Niu, G. 和 Sabato, S.（编），第*39*届国际机器学习会议论文集，第162卷《机器学习研究论文集》，第26772–26808页。PMLR。

A 符号和缩略语

符号

$\boldsymbol{a}$, $\boldsymbol{b}$ 典型的纯行动组合，其中 a^i， b^i 表示 i所采取的行动。5a_t^i 在重复场景中， i 在时间 t采取的行动。5C， C_{k_t} 在一个时期结束后到达的作业所产生的惩罚。5η EWA 中权重更新的平滑度。15h_t 历史：前几个时期的状态—动作组合对。6L 每个时期的长度。5$\lambda_{t,\ n}$ 在 1 到 t 之间， k_t^i 等于 n的次数。18

S 系统的状态，即每个玩家持有的工作数量。5$\boldsymbol{\sigma}$ 典型的混合行动组合，其中 σ^i 表示 i采取的混合行动， $\sigma^i(a^i)$ 表示 i 采取 a^i的概率。7 SC($\boldsymbol{\sigma}_t$) 行动组合 $\boldsymbol{\sigma}$的社会成本，即 $\sum_{i\in I} c^i(\boldsymbol{\sigma})$。7$\tau_n^i$ 玩家 i 首次拥有 n 个工作并进行分派的时间。15$w_t^i(b^i)$ EWA 中玩家 i 的行动 b^i 在时间 t的权重。15$w_{t,n}^i(b^i)$ MLEWA 中玩家 i 在 n 级的行动 b^i 于时间 t的权重。16x^i 重复情境中玩家 i 的策略，即一个函数 $\mathcal{H} \to \triangle(A)$。6$x_t^i$ 玩家 i在时间 t的策略，即 $x^i(h_t)$。6

缩略语

先进先出 先进先出**后进先出** 后进先出**粗相关均衡** 粗相关均衡**DQM** 动态排队模型**EWA** 指数权重算法**MLEWA** 多层指数权重算法**NE** 纳什均衡

B 强化随机游走引理的证明

引理*19*的证明。 以下证明受 Pemantle（2007）的启发。

我们首先证明：对于所有满足 $x \geqslant x_0 + M$ 的 x，在已知 $[x - M$ 已达到 $x]$ 的条件下，$\sup_{t\in\mathbb{N}} X_t \in [x - M, x]$ 的概率有一个严格大于 0 的下界，且该下界不依赖于 x_0。正如我们将要证明的，这意味着几乎必然存在 x，使得 $\sup_{t\in\mathbb{N}} X_t = x_0$。

记事件 使得 $[x-M, x]$ 通过 X_t 经由 $A(x)$ 达到。现在我们证明，对于所有 $x \geqslant x_0+M$ ：

$$\mathbb{P}\left[\sup_{t\in\mathbb{N}} X_t \in [x - M, x] \mid A(x)\right] \geqslant \prod_{z\in\left[\frac{x-M}{d}, x\right]} \prod_{m\in\mathbb{N}} (1 - r(z, m)) \tag{B.1}$$

为证明这一结果，我们固定 $x \in \mathbb{N}$ 使得 $x > dz_0 + M$，并引入一个扩展的随机过程。定义新的状态空间 $\Omega = \mathbb{N}^2 \times \mathbb{N} \times \mathbb{N}^x \times \{0,1\}$。状态 $(x, z, n_0, ..., n_x, i)$ 的含义如下：

- 当前状态为 (x, z),
- Z_t 的路径经过状态 r 已有 n_r 次,

当且仅当 X_t 从介于 x 和 $x - M$之间的状态向上跳跃时，• i 等于 1。

形式上，设 $(X, Y, N_0, ..., N_x, I)_{t\geqslant 1}$ 为 Ω 上的随机过程。第一个坐标等于 X_t，而所有其他坐标均由它推导得出。根据构造，我们知道 X_t 的最大增量为 M，因此，为使上确界严格大于 x，必须从介于 $x - M$ 和 x之间的状态发生正跳跃，因此

$$\{\sup_{t\in\mathbb{N}} X_t > x\} \subset \{\exists t \geqslant 1, I_t = 0\}.$$

由此可得

$$\mathbb{P}\left[\sup_{t\in\mathbb{N}} X_t \in [x - M, x] \mid A(x)\right] \geqslant \mathbb{P}\left[\forall t \geqslant 1, I_t = 0 \mid A(x)\right].$$

此外,

$$\mathbb{P}\left[\forall t \geqslant 1, I_t = 0 \mid A(x)\right] \geqslant \prod_{z\in[\frac{x-M}{d},x]} \prod_{m\in\mathbb{N}} (1 - r(z, m)).$$

事实上，根据辅助随机过程的构造，我们知道：

- 对于每个 $r \in \{0, ..., M\}$, N_r 都只增不减,
- 在给定 $Z_t = z \geqslant \frac{x-M}{d}$, $N_z = n$ 及过去的条件下， i 保持为 0 的概率至少为 $(1 - r(z, n))$,
- 在给定 $z < \frac{x-M}{d}$ 的条件下， i 保持为 0 的概率为 1。

由此可得

$$\mathbb{P}\left[\sup_{t\in\mathbb{N}} X_t \in [x - M, x] \mid A(x)\right] \geqslant \mathbb{P}\left[\forall t \geqslant 1, I_t = 0 \mid A(x)\right] \tag{B.2}$$

$$\geqslant \prod_{z\in[\frac{x-M}{d},x]} \prod_{m\in\mathbb{N}} (1 - r(z, m)). \tag{B.3}$$

那么，右端的对数为

$$\begin{aligned}
\sum_{z\in[\frac{x-M}{d},x]}\sum_{m\in\mathbb{N}}\log(1-r(z,m)) &= \sum_{z\in[\frac{x-M}{d},x]}\sum_{m\in\mathbb{N}}-\log(1+\frac{r(z,m)}{1-r(z,m)}) \\
&\geqslant \sum_{z\in[\frac{x-M}{d},x]}\sum_{m\in\mathbb{N}}-\frac{r(z,m)}{1-r(z,m)} \\
&\geqslant \sum_{z\in[\frac{x-M}{d},x]}\sum_{m\in\mathbb{N}}-\frac{r(z,m)}{1-\rho} \\
&= \sum_{z\in[\frac{x-M}{d},x]}-\frac{A}{z(1-\rho)} \\
&\geqslant -\frac{A}{1-\rho}\sum_{z\in[\frac{x-M}{d},x]}\frac{d}{x-M} \\
&= -\frac{A}{1-\rho}\left(x-\frac{x-M}{d}+1\right)\frac{d}{x-M} \\
&= -\frac{A}{1-\rho}\frac{(d-1)x+M+d}{x-M} \geqslant -B > 0,
\end{aligned}$$

其中， $B>0$ 是一个不依赖于 x的正数常量。

由此可得

$$\mathbb{P}\left[\sup_{t\in\mathbb{N}} X_t \in [x-M,x] \mid A(x)\right] \geqslant \exp(-B) > 0. \tag{B.4}$$

因此，在已到达该区间的条件下， X_t 的上界属于 $[x-M,x]$ 的概率有一个不依赖于 x 的常数下界。

$$\begin{aligned}
\mathbb{P}\left[\sup_t X_t < \infty\right] &= \sum_{k\geqslant 1}\mathbb{P}\left[\sup_t X_t \in [(k-1)M, kM]\right] \\
&\geqslant \sum_{k\geqslant[\frac{x_0}{M}]+1}\mathbb{P}\left[\sup_t X_t \in [(k-1)M,kM] \mid A(kM)\right]\mathbb{P}\left[A(kM)\right].
\end{aligned}$$

然而，如果 X_t无界，那么 $A(kM)$就会发生，因此 $\mathbb{P}\left[A(kM)\right] \geqslant \mathbb{P}\left[\sup_t X_t = +\infty\right]$。因此，使用 g (B.4):

$$\mathbb{P}\left[\sup_t X_t < \infty\right] \geqslant \sum_{k\geqslant[\frac{x_0}{M}]}\exp(-B)\,\mathbb{P}\left[\sup_t X_t = +\infty\right].$$

当 $\mathbb{P}\left[\sup_t X_t = +\infty\right] > 0$时，右端等于8，因此必然有 $\mathbb{P}\left[\sup_t X_t = +\infty\right] = 0$。 □