arXiv:2303.01745v1 [math.OC] 3 Mar 2023

对抗性老虎机学习下的队列调度

预印本

黄家泰清华大学
hjt18@mails.tsinghua.edu.cn

Leana Golubchik南加州大学 leana@usc.edu

Longbo Huang清华大学
longbohuang@tsinghua.edu.cn

2023年3月6日

摘要

本文研究在完全不了解瞬时网络状况的情况下，如何调度排队系统。我们考虑一个由 K 个队列组成的单跳单服务器排队系统，其中每个队列的到达率和服务率均随时间变化且非平稳。我们的调度方法创新性地结合了对抗性老虎机学习与李雅普诺夫漂移最小化，无需了解每个队列的瞬时网络状态（到达率和服务率）。随后，我们提出两种新颖的算法：SoftMW（SoftMaxWeight）和 SSMW（滑动窗口 SoftMaxWeight），二者均能够稳定那些可由某个（可能未知的）随机化策略序列稳定的系统，前提是该序列的时间变化满足一个较弱的条件。我们进一步将结果推广到到达量和离开量仅具有有界矩、而非确定性有界的情形，并提出能够稳定系统的 SoftMW+ 和 SSMW+ 。作为新算法的构建模块，我们还扩展了经典的多臂老虎机算法 EXP3.S [3]，使其能够处理无界反馈信号；这一扩展本身也可能具有独立价值。

关键词 调度、排队 · 赌博机学习 · 李雅普诺夫分析

1 引言

随机网络调度关注的是在动态环境中分配资源以满足需求这一基本问题，并广泛应用于现实网络化系统的建模，包括数据通信 [19, 37], 云计算和服务器集群 [25, 10, 31, 5], 智能电网管理 [14, 18, 24], 供应链管理 [32, 4], 以及交通网络控制 [39, 7, 6]。大多数现有调度方案的一个基本要求是，在采取新的调度动作之前，掌握即时网络状态——即到达流量的数量，以及在任何可行控制动作下的服务量，例如所有链路之间的功率分配。基于这些信息，已有许多成功的网络调度算法，并在多个方面提供了理论性能保证，包括队列稳定性 [38, 34, 23], 时延 [26, 27, 15], 和效用 [16, 15, 29]。

然而，在许多现实场景中，如果测量或估计网络状态知识过于困难或成本过高，这类知识可能并非总是可用。即使能够获得这类知识，它也可能存在偏差且不完善。例如，在物联网系统中，由于传感器温度漂移或设备故障，流量和信道模式可能随时发生意料之外的变化 [11]。在水下通信系统中，要实现精确的信道状态估计极具挑战性 [17]。此外，在通信双方能够快速移动的应用中，例如自动驾驶车辆 [1], , 或者能够以任意方式移动的应用中，例如无线增强现实/虚拟现实设备 [8],, 信道条件也可能快速变化，因此难以准确估计。因此，依赖精确网络状态知识的调度策略可能不适用于许多现实任务；由于信息不准确，采用这类策略可能导致性能显著下降。因此，在没有网络状态即时知识和准确估计的情况下进行网络调度无需，无论在理论上还是实践中都很重要，也就是说，这可以显著提升大规模网络化系统的鲁棒性和可用性，同时降低运营和维护成本。

为此，本文聚焦于一种新颖的无需网络状态知识的调度 建模。具体而言，我们研究一项单跳调度任务，其中单服务器为 K 个队列提供服务，每个队列对应一种作业类型。服务器在每个时隙选择一个队列进行服务。网络动态，即到达率和服务率，以无知对抗式 的方式演变，并且在做出调度决策之前未知。此外，只有在采取动作后才能观察到服务结果，并且反馈采用老虎机反馈的形式，即只有被服务的队列会产生观测。我们的目标是寻求一种高效的调度策略来稳定网络。

为解决这一问题，我们提出了新颖的学习增强型调度算法，其灵感来自著名的 MaxWeight 队列调度算法 [28]以及 EXP3 系列算法在非平稳多臂老虎机（MAB）问题中的成功应用 [3]。只要某种随机化策略能够稳定该系统，且其为每类作业提供服务的概率全变差不太大，所提出的算法就能够稳定非平稳系统。或许令人惊讶的是，我们的算法既不依赖于事先了解网络统计信息，也不依赖于对系统进行复杂的显式实时估计。因此，与依赖网络状态知识的对应方法相比，我们的算法天然更能抵御系统中的抖动以及意外的流量/服务模式。事实上，附录 A 对我们的算法与其依赖精确信息的对应方法进行了数值比较。通过这一比较，我们表明，正如图 1 所示，在存在服务状态噪声的系统中，本文提出的算法确实表现更优。

我们的工作与现有的学习增强型网络控制文献（例如 [9, 21, 13, 40], ）在以下方面有所不同。 [9, 13, 21] 研究的是平稳系统上的调度或负载均衡任务，且事先不知道速率统计信息；而在我们的设定中，系统可以是时变且对抗性的。 [40]也考虑了非平稳系统，但他们假设服务率平滑时变，并使用指数平均和折扣 UCB 奖励项显式估计瞬时服务率。与这些工作相比，我们的方法既不需要显式优化离线问题，也不需要显式探测和估计瞬时信道状态，而是利用对抗性老虎机学习技术，在探索系统的同时使其保持稳定。

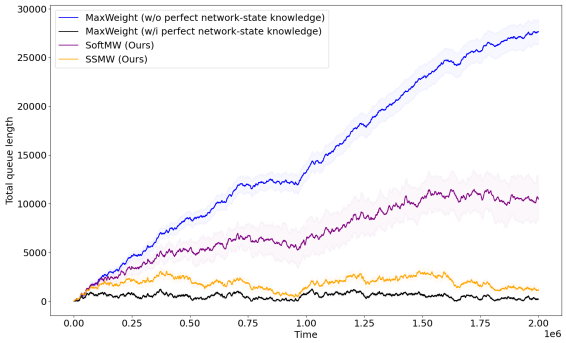


图 1：非平稳系统上的数值评估（详情请参见附录 A）

本工作的贡献可概括如下：

• 我们提出了两种新颖的调度算法 SoftMW（算法2）和 SSMW（滑动窗口 SoftMW，算法3），能够在不了解信道状态的情况下调度单跳排队系统，同时在参考随机策略的时间变化满足温和条件时保证系统稳定。

• 在设计这两种算法时，我们将在线赌博机学习与基于李雅普诺夫漂移的调度方法及其分析技术进行了巧妙结合。赌博机部分能够保证，在有限时间范围内，相对于未知的时变随机策略，我们的算法的“遗憾”较小。随后，我们以一种创新的方式将遗憾保证与李雅普诺夫漂移分析相结合，从而得出稳定性结果（见第 5.3 节和第 6.1 节）。

• 我们扩展了 EXP3.S 算法 [3], ，该算法最初是为具有有界奖励的对抗性多臂赌博机问题设计的，使其能够采用时变学习率和探索率来处理无界反馈（见第5.1节、算法1）。这一扩展后的 EXP3.S 算法（我们称之为 EXP3.S+）被用作 SoftMW 和 SSMW 的构建模块。不过，超出排队问题的范畴后，它本身也具有独立的研究价值。

• 我们进一步将结果推广到到达量和离开量具有有界矩而非确定性有界的情形（参见第 7 节）。我们提出了 SoftMW+ （算法 4 ）和 SSMW+ （算法 5 ），它们能够稳定系统。

表1总结比较了我们提出的算法与密切相关的研究。据我们所知，我们的工作首次将具有动态遗憾保证的对抗性多臂老虎机算法用于排队系统调度。大多数先前的研究基于ε-贪心或上置信界（UCB），并需要假设系统是平稳的，或虽非平稳，但到达（离开）速率具有足够的平滑性。

因此，我们的算法可应用于更一般、更复杂的情形。我们相信，我们的方法能够为最大权重型方法以及其他排队调度算法设计问题带来新颖且有价值的见解。

表1：我们的算法及密切相关工作的概述

算法	可稳定系统	平均队列长度
MaxWeight [36]	同质作业	$\emptyset(\frac{KM^2}{\epsilon})$
	假设 1 + 服务率预测	$\emptyset(\frac{C_W KM^2}{\epsilon})$
MaxWeight 与 折扣 UCB[40]	假设 1， 服务率具有与折扣因子相匹配的平滑性[1]	$\left(MK\epsilon^{-1}\right)^{\emptyset(1/\delta)}$
SoftMW (我们的方法，算法2)	假设 1， 假设 2（参考策略的全变差为 $\emptyset(T^{\frac{1}{2}-\delta})$）	$\emptyset(\frac{C_W KM^2}{\epsilon})$
SSMW (我们的方法，算法 3)	假设 1， 假设3（$\emptyset(T^{1-\delta})$ 参考策略时间齐次全变差）	$\left((1+C_V)MK\epsilon^{-1}\right)^{\emptyset(1/\delta)}$
SoftMW+ (我们的方法，算法 4)	假设 1，假设 2（$\emptyset(T^{\frac{1}{2}-\delta})$ 参考策略全变差）， 到达量和离开量可以无界，但其 α 阶矩有界，$\alpha\cdot\delta>7$	$\emptyset(\frac{C_W KM^2}{\epsilon})$
SSMW+ （我们的方法，算法 5）	假设 1， 假设3（$\emptyset(T^{1-\delta})$ 参考策略时间齐次全变差）， 到达量和离开量可以无界，但其二阶矩有界	$\left((1+C_V)MK\epsilon^{-1}\right)^{\emptyset(1/\delta)}$

[1] [40] 采用了类似的假设，即每个信道的单步服务率漂移都被某个关于 $(1-\gamma)^{-1}$的多项式统一上界。这里，γ是其算法的一个超参数，即上置信界（UCB）中的折扣因子。

2 符号说明

在本文中，对于 $n\geq 1$，我们将集合 $\{1,2,\ldots,n\}$ 记为 $[n]$，将 $[n]$ 上的（$n-1$）维概率单纯形记为 $\triangle^{[n]}$。我们使用粗体英文字母（例如，$\mathbf{Q}_t$、$\mathbf{S}_t$）以及带上箭头的希腊字母（例如，$\vec{\sigma}_t$、$\vec{\lambda}_t$）表示向量值变量。我们用 $\mathbf{0}$ 表示全零向量，用 $\mathbf{1}$ 表示全一向量。我们用 $\mathbf{1}_i$ 表示第 i 个坐标为 $\mathbf{1}$ 的独热向量，即当 $i=j$ 时为（$\mathbf{1}_i$）$_j=1$，否则为 $\mathbf{0}$。我们用 $\mathbb{1}[\text{statement}]$ 表示给定命题的指示函数；若该命题成立，其值为 $\mathbf{1}$，否则为 $\mathbf{0}$。我们用 $\mathbf{x}\odot\mathbf{y}$ 表示两个向量 $\mathbf{x}$ 和 $\mathbf{y}$的逐元素乘积。

设 f 是定义在某个凸域 $A\subseteq\mathbb{R}^K$ 上的严格凸函数。对于任意 $\mathbf{x}$、$\mathbf{y}\in A$，若 $\nabla f(\mathbf{x})$ 存在，则将由 f 诱导的 y 与 x 之间的 Bregman 散度记为

$$D_f(\mathbf{y},\mathbf{x})\triangleq f(\mathbf{y})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle$$

我们使用 $f^*(\mathbf{y})\triangleq\sup_{\mathbf{x}\in\mathbb{R}^K}\{\langle\mathbf{y},\mathbf{x}\rangle-f(\mathbf{x})\}$ 表示 f 的凸共轭。

$\tilde{\ }$ $\tilde{\ }$ $\tilde{\ }$我们使用 $\mathcal{O}$、Ω 或 Θ 来省略 T（决策时域的长度）和 K（队列数量）中的多对数因子。除非另有说明，我们使用

$$\mathcal{F}_t=\sigma\left(a_1,\ldots,a_t,\mathbf{Q}_0,\ldots,\mathbf{Q}_t,\mathbf{A}_1,\ldots,\mathbf{A}_t,S_{1,a_1},\ldots,S_{t,a_t}\right)$$

对于任意 $t\geq 0$，表示 由 σ-代数生成的滤过，用于研究以时间为索引的随机量，即 $\mathcal{F}_t$ 由所有决策生成 调度策略在第 t个时隙结束时可见的信息和数量。

3 问题设定

我们研究在分时隙系统中，使用单个工作保持服务器对 K 种作业类型进行调度的问题。每个到达的作业首先加入与其作业类型 i 相关联的队列，我们将其记为 Q_i。记 $A_{t,i}$ 为第 t 个时隙到达的 i 类型作业数量，记 $S_{t,i}$ 为服务器在第 t 个时隙最多可以服务的 i 类型作业数量。在每个时隙 t 开始时，服务器选择恰好一个作业 $a_t\in[K]$ 类型进行服务。记 $Q_{t,i}$ 为时隙 t 结束时 i 类型作业的队列长度。于是，每个 $Q_{t,i}$ 按照以下方程演化：

$$Q_{t,i}=\max\left\{Q_{t-1,i}+A_{t,i}-S_{t,i}\mathbb{1}[i=a_t],0\right\}$$

其中，$\mathbf{Q}_0 = (Q_{0,1}, ..., Q_{0,K}) = \mathbf{0}$。在每个时隙 t 开始时，最新的队列长度 $Q_{t-1,1}, \ldots, Q_{t-1,K}$ 对服务器可用，以便其作出新的决策。过去动作的最大服务量 $S_{0,a_0}, \ldots, S_{t-1,a_t}$ 对服务器也是可见的。

我们假设存在两个分布序列 $\{\mathcal{A}_1, \mathcal{A}_2, \ldots\}$ 和 $\{\mathcal{S}_1, \mathcal{S}_2, \ldots\}$，它们都在队列过程开始前固定，且其统计特性事先已为调度器所知。所有分布 $\mathcal{A}_t$s 和 $\mathcal{S}_t$s 的支撑集均为 $[0, M]^K$，其中 M 是一个事先已知的常数。我们进一步假设，每个 $\mathbf{A}_t$ 都是从 $\mathcal{A}_t$ 中随机抽样得到的，每个 $\mathbf{S}_t$ 都是从 $\mathcal{S}_t$ 中抽样得到的，并且所有 $\mathbf{A}_t$s 和 $\mathbf{S}_t$s 都是相互独立的随机向量。我们用 λ_t 表示 $\mathcal{A}_t$ 的均值，用 $\vec{\sigma}_t$ 表示 $\mathcal{S}_t$ 的均值。

我们的目标是设计一种调度策略，使得平均期望队列长度满足以下上界：

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{K}\mathbb{E}[Q_{t,i}] = \emptyset(1).$$

如果上述界成立，我们就称调度策略 使系统稳定，或者称系统在某种调度策略下稳定。

平稳系统上的经典调度任务（例如 [9, 21]）对应于 $\mathcal{A}_t = \mathcal{A}_1$ （$\mathcal{S}_t = \mathcal{S}_1$）的情形，即在我们的设定中，分布是时不变的。在我们的问题中，在保证队列稳定性的同时探索并估计时变服务分布十分复杂。

4 稳定系统的充分条件

在本文中，我们对系统作出如下假设。该假设类似于平稳网络调度中的容量区域定义 [28], ，并可视为对抗性环境中调度的一种广义稳定性条件。

假设 1 (分段可稳定性)。 存在 $C_W \geq 0$、 $\epsilon > 0$、 $\vec{\theta}_1$、 $\vec{\theta}_2$、 $\cdots \in \Delta^{[K]}$ 以及将 $\mathbb{N}_+$ 划分为区间 W_0、 W_1、 $\cdots$ 的一个划分，使得对于任意 $T \geq 1$，均有

$$\sum_{i:\min_{t\in W_i} t<T} \left(|W_i| - 1\right)^2 \leq C_W T \tag{1}$$

并且对于任意 $i \geq 0$ 和 $j \in [K]$，均有

$$\frac{1}{|W_i|}\sum_{t\in W_i}\theta_{t,j}\sigma_{t,j} \geq \epsilon + \frac{1}{|W_i|}\sum_{t\in W_i}\lambda_{t,j}. \tag{2}$$

假设 1 可视为 [22] 中（W， ϵ）受约束动态的一种推广。它本质上假设时间范围可划分为若干区间，在每个区间内都存在能够稳定网络的平稳策略（式（2））。作为快速的合理性检验，对于到达率向量位于容量区域内部的平稳实例，假设 1 自动成立，此时 $C_W = 0$（因此所有 W_i 均为单元素集合），且所有有 $\vec{\theta}_i$ 都等于某个固定元素 $\vec{\theta} \in \Delta^{[K]}$；该元素是能够稳定系统的随机化策略。

注。 事实上，在上述假设下，通过二次李雅普诺夫漂移论证（见定理 5.5），我们还可以证明一种策略同样能够稳定系统：在每个时间步 t，根据 $\vec{\theta}_t$ 所示的分布，独立随机地服务一种类型的作业 a_t （需要事先知道 $\vec{\theta}_t$）。我们将 $\{\vec{\theta}_t : t \geq 1\}$ 参考混合动作序列 称为.....，并将由 $\{\vec{\theta}_t : t \geq 1\}$ 诱导的上述随机化策略称为 参考随机策略。

在假设 1 下，通常而言，调度系统仍然是一个具有挑战性的问题。对于第 5 节中的主要结果，我们还需要下面提出的另一项技术假设。

假设 2 (参考策略平稳性)。 对于假设 *1* 中的参考混合动作序列 $\{\theta_t\}$，存在某些 $\delta > 0$ 和 $C_V > 0$，使得

$$\sum_{t=1}^{T-1}\|\vec{\theta}_{t+1} - \vec{\theta}_t\|_1 \leq C_V T^{\frac{1}{2}-\delta}$$

对于任意 $T \geq 1$。

¯直观而言，假设 2 表明，序列 $\{\theta_t\}$（从而环境）不会发生非常 abrupt 的变化。已有研究中也采用了类似的平滑性假设，例如 [40]。 [1] 在第6节中，我们还将研究何时能够处理参考策略变化幅度显著更大的问题。

5 仅凭老虎机反馈进行队列调度

与网络状态信息完备的情形相比，在我们的设定中，调度器无法获得准确的信道条件。具体而言，服务器在每个时间步只能收到关于实际服务的老虎机反馈，即只有在做出服务决策后，才知道$S_{t,\ a_t}$。

本节将介绍一种新颖的算法，它仅利用老虎机反馈$S_{t,\ a_t}$即可稳定系统。我们的核心思路是将合适的多臂老虎机算法嵌入 MaxWeight 调度器 [36], ，从而保证作为 MaxWeight 关键要素的项 $\mathbb{E}[\sum_{t=1}^{T} Q_{t-1,a_t} S_{t,a_t}]$与 $\mathbb{E}[\sum_{t=1}^{T}\langle \mathbf{Q}_{t-1}\odot\mathbf{S}_t,\ \vec{\theta}_t\rangle]$不会相差太远。若能获取 $\vec{\sigma}_t$，MaxWeight 会在每个时间步 t通过贪心地选择 $a_t = \arg\max_i Q_{t-1,\ i}\sigma_{t,\ i}$来实现这一点。然而，当 $\vec{\sigma}_t$ 未知且随时间变化时，很难保证每个加项$Q_{t-1,\ a_t}S_{t,\ a_t}$ 都足够大。因此，我们着重优化总和 $\mathbb{E}[\sum_{t=1}^{T} Q_{t-1,a_t} S_{t,a_t}]$。

本节余下部分将首先介绍 EXP3.S+，这是针对对抗性多臂老虎机的 EXP3.S [3] 算法的扩展版本（第5.1节）。EXP3.S+ 具有足够的灵活性，可作为我们新型调度算法 SoftMW（第5.2节）的重要构件。我们还将介绍其性能保证，因为这对于理解后续分析至关重要。最后，在第5.3节中，我们将概述 SoftMW 的分析，并介绍几种重要的新技术，用于将对抗性多臂老虎机学习与李雅普诺夫漂移分析联系起来。

5.1 EXP3.S+：EXP3.S 的扩展版本

我们首先介绍 EXP3.S+。它扩展了最初为解决对抗性多臂老虎机（MAB）问题而设计的 EXP3.S 算法 [3], ，以应对排队系统中可能出现的无界队列长度；现有的老虎机算法无法直接处理这种情况。

更正式地说，EXP3.S+ 适用于以下情形：一个智能体和一个对手同时在有限长度的时间范围 $t = 1 \dots T$内做出决策。在每个时间 t，智能体根据已观察到的历史确定性地选择一个 $\mathbf{x}_t \in \Delta^{[K]}$，然后按照 $\mathbf{x}_t$进行采样。与此同时（在时间 t），对手根据已观察到的历史确定性地选择 $\mathbf{g}_t \in \mathbb{R}^K$。随后，$g_{t,a_t}$ 会向智能体揭示。智能体的总体目标是最大化累积反馈 $\sum_{t=1}^{T} g_{t,a_t}$。EXP3.S+ 的具体细节见算法1。

算法1： EXP3.S+

输入：动作数量 K，时间范围长度 T，初始混合动作 $\mathbf{x}_1 \in \Delta^{[K]}$输出：动作序列 a_1、a_2、……、$a_T \in [K]$中间变量：学习率序列 η_1、η_2、……、$\eta_T \in \mathbb{R}_+$，隐式探索率序列 β_1、β_2、……、$\beta_T \in [0, 1/K]$，显式探索率序列 γ_1、γ_2、……、$\gamma_T \in [0, 1/2]$，显式探索法向量序列 $\mathbf{e}_1, \mathbf{e}2$、……、$\mathbf{e}_T \in \Delta^{[K]}$1 $\Psi(\mathbf{x}) \triangleq \sum_{i=1}^{K}(x_i \ln x_i - x_i)$ 2，对于 $t = 1$、2、……、T执行 3 选择 β_t、η_t、$\mathbf{e}_t$ 和 γ_i4 以下记 $\Delta^{[K]}$、$_{\beta_t} \triangleq \{\mathbf{x} \in \Delta^{[K]} : \mathbf{x}_i \geq \beta_t \quad \forall i \in [K]\}$5 $\mathbf{p}_t \leftarrow (1-\gamma_t)\mathbf{x}_t + \gamma_t\mathbf{e}_t$6 从 $a_t \sim \mathbf{p}_t$中采样，采取动作 a_t，观察 g_t、$_{a_t}$7 $\tilde{\mathbf{g}}_t \leftarrow \begin{cases} g_t \\ 0 \end{cases}$、$_{a_t}/p_t$、$_{a_t}$ 第 a_t 个坐标及其他坐标 8 $\mathbf{x}_{t+1} \leftarrow \arg\min_{\mathbf{x}' \in \Delta^{[K]}}$、$_{\beta_t}$ $\langle -\eta_t \tilde{\mathbf{g}}_t$、$\mathbf{x}'\rangle + D_\Psi(\mathbf{x}'$、$\mathbf{x}_t)$

[1] 严格来说， [40] 对到达率和服务率而非参考随机策略引入了平滑性假设。

注。 第 6 行中反馈值 $g_{t\text{的幅度}a_t}$ 对于 EXP3.S 的正确性至关重要。 [3] 中的原始 EXP3.S 算法在所有 T时间步中使用恒定的学习率 η 和恒定的探索率 γ。然而，该算法只能支持反馈值不超过 $\eta^{-1}\gamma$的问题，不适用于我们的场景，因为队列长度（即 EXP3.S 的奖励）可能无界。为此，在所提出的算法中，我们将 $Q_{t-1,\,a_t}S_{t,\,a_t}$ 输入 EXP3.S+ 作为奖励值；该值可能任意大（因为 $Q_{t-1,\,a_t}$ 可能任意大）。在 EXP3.S+中，学习率和探索率都可以是时变的，并且探索率甚至可以依赖于动作（它允许指定任意的$\mathbf{e}_t \in \Delta^{[K]}$，而非 $\mathbf{1}/K$）。

EXP3.S+ 的$\sum_{t=1}^{T} g_{t,\,a_t}$ 正式性能保证如下方定理 5.1 所述。

定理 5.1 （EXP3.S+ 动态遗憾保证）。在算法*1*的执行过程中，对于任意固定序列$\vec{\theta}_1$、.....、 $\vec{\theta}_T \in \Delta^{[K]}$，若以概率 *1* 发生以下事件，

(i) $\mathbf{x}_1 \in \Delta^{[K],\,\beta_1}$; *(ii)* $\mathbf{g}_t \le \eta_t^{-1}\gamma_t\mathbf{e}_t$ 对所有 $1 \le t \le T$;
(iii) $\eta_1 \ge \eta_2 \ge \cdots \ge \eta_T$; *(iv)* $\beta_1 \ge \beta_2 \ge \cdots \ge \beta_T$;

(v) $\vec{\theta}_t \in \Delta^{[K],\,\beta_t}$ 对所有 $1 \le t \le T$,

那么令

$$V \triangleq \sum_{t=1}^{T-1} \|\vec{\theta}_{t+1} - \vec{\theta}_t\|_1,$$

我们将得到

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \mathbf{g}_t, \vec{\theta}_t\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T} g_{t,a_t}\right] \le (1+V)\mathbb{E}\left[\eta_T^{-1}\ln\frac{1}{\beta_T}\right] + e\mathbb{E}\left[\sum_{t=1}^{T}\eta_t\|\mathbf{g}_t\|_2^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\langle \mathbf{g}_t, \mathbf{e}_t\rangle\right].$$

在附录B中，我们使用基于在线镜像下降的分析方法，对定理 5.1 给出形式化证明 [12],。与 [3]中经典的指数和势函数方法相比，这种方法更适合处理时变学习率。我们还在附录C中讨论了 argmax 计算（第8行）的实际实现。值得注意的是，算法1及其分析本身也具有独立的研究价值，并可应用于随机网络调度以外的问题。

5.2 使用EXP3.S+的软最大权重调度

现在，我们在算法2中介绍新颖的调度算法 SoftMW。SoftMW 基于精心设计的反馈信号，以及 EXP3.S 中的参数和学习率+。其名称源于 EXP3.S+（算法1）中的计算，该计算大量依赖软最大操作（见附录C）。

算法2有效的直观原因如下。我们以精心设计的方式使用 EXP3.S+ 来驱动调度过程，使得算法2的效果非常接近（甚至优于）假设 1 给出的参考随机策略；具体而言，在算法2下，队列的总二次李雅普诺夫漂移仅略大于（甚至小于）参考随机策略下的漂移。因此，算法2具有相似（甚至更强）的系统稳定能力。

算法2在任何有限时间范围内的平均队列长度上界由定理 5.2 给出。**定理 5.2.** 对于满足假设 *1* 和 *2* 的问题实例， *SoftMW(*算法*2)* 保证 提供

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_t\|_1\right] \le \frac{2(K+1)M^2 + 4C_W(KM^2 + \epsilon KM)}{\epsilon} + o(1).$$

特别是，系统是稳定的。

注。 作为快速的合理性检查，对于平稳问题实例，定理 5.2 给出的平均队列长度上界为 $\emptyset(KM^2/\epsilon)$，这与我们在平稳问题中能够达到的经典结果一致（[28] 第 3.1 节）。事实上，可以证明，对于以下两种情况：(i) 假设能够获得准确的一步服务率预测，并运行原始

算法2： `SoftMW`（**Soft M**ax**W**eight）

输入： 一步到达/服务上界 $M > 0$，作业类型数量 K，问题实例光滑度参数 δ ¿ 0**输出：** 要服务的作业类型序列 a_1， a_2， $\ldots \in [K]$

使用 + 个可用动作初始化一个 EXP3.S K 实例，并令 $\mathbf{x}_1 = \mathbf{1}/K$ **对于** $t = 1$，2，... **执行**

为时隙 t设置以下 EXP3.S+ 参数：

$\beta_t \leftarrow t^{-3}/K$

$\eta_t = \left(t^{-(\frac{1}{4}-\frac{\delta}{2})} M \sqrt{86M^2K^6t^{\frac{3}{2}} + \sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}$

$\mathbf{e}_t = \mathbf{Q}_{t-1}/\|\mathbf{Q}_{t-1}\|_1$

$\gamma_t = M\eta_t\|\mathbf{Q}_{t-1}\|_1 = \|\mathbf{Q}_{t-1}\|_1 \left(t^{-(\frac{1}{4}-\frac{\delta}{2})} \sqrt{86M^2K^6t^{\frac{3}{2}} + \sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}$

从 EXP3.S+中获取新的动作决策输出 a_t ，服务第 a_t 个队列，将$Q_{t-1、a_t}S_{t、a_t}$ 视为新的反馈$g_{t,\ a_t}$ 并将其输入当前的 EXP3.S+ 实例

$\vec{}$MaxWeight，以及 (ii) 运行假设 1 中指定的参考随机策略 $\{\theta_t\}$ ，通过标准的二次李雅普诺夫分析得到的平均队列长度上界为 Ø $\left(\begin{matrix}(C_W K+K+1 \\ )_\epsilon\end{matrix} \quad M^2\right)$。因此，粗略来说，就队列长度上界而言，对于满足假设 2 的问题实例，SoftMW 带来的开销可以忽略不计。

5.3 队列稳定性分析大纲

本节简要概述如何正式证明队列稳定性结果（定理 5.2）。我们首先回顾二次李雅普诺夫漂移分析的一般流程。接着，我们说明 EXP3.S+ 调度能够使终端李雅普诺夫函数值接近假设 1 中的参考策略，其差异与 $\sqrt{\sum}\|\mathbf{Q}_t\|_2^2$ 成正比。最后，我们将该 $\sqrt{\sum}\|\mathbf{Q}_t\|_2^2$ 项与我们希望界定的队列长度（$\sum\|\mathbf{Q}_t\|_1$）联系起来，并说明李雅普诺夫终端值上界可以推出平均队列长度上界。

5.3.1 李雅普诺夫漂移分析回顾

在我们的分析中，我们使用二次李雅普诺夫漂移分析中的标准结果 [28]。按照惯例，我们定义

$$L_t \triangleq \frac{1}{2}\|\mathbf{Q}_t\|_2^2 = \frac{1}{2}\sum_{i=1}^{K} Q_{t,i}^2,$$

为队列长度的二次李雅普诺夫函数。首先，我们有以下关于漂移上界的标准引理。

引理 5.3 （一般二次李雅普诺夫漂移上界 [28]）。 考虑该排队系统的任意调度策略，并假设该策略根据概率分布随机选择一种作业类型 a_t ，其中该概率分布可能取决于系统历史，即，$\mathbf{p}_t$ 是一个 $\mathcal{F}_{t-1}$-可测且取值于 $\Delta^{[K]}$ 的随机向量。令 $\mathbf{Q}_t$ 表示该策略下的队列长度向量。我们有

$$\begin{aligned}\mathbb{E}\left[L_t - L_{t-1} \middle| \mathcal{F}_{t-1}\right] &\leq \frac{(k+1)M^2}{2} + \langle \mathbf{Q}_{t-1}, \vec{\lambda}_t - \vec{\sigma}_t \odot \mathbf{p}_t\rangle \\ &= \frac{(k+1)M^2}{2} + \langle \mathbf{Q}_{t-1}, \vec{\lambda}_t\rangle - \mathbb{E}\left[Q_{t-1,a_t}S_{t,a_t} \middle| \mathcal{F}_{t-1}\right]\end{aligned}$$

对任意 $t \geq 1$均成立。将不等式对 $1 \leq t \leq T$ 求和，取全期望，然后重新整理各项，可得

$$\mathbb{E}\left[\sum_{t=1}^{T} Q_{t-1,a_t}S_{t,a_t} - \langle \mathbf{Q}_{t-1}, \vec{\lambda}_t\rangle\right] \leq \frac{(K+1)M^2T}{2} \tag{3}$$

对于任意时间范围长度 $T \geq 1$。

接下来，我们给出引理 5.4，讨论参考策略下的漂移值。与标准的李雅普诺夫漂移分析一样， [28],这一界将有助于推导基于队列的策略的队列长度结果。

引理 5.4（参考策略下的负李雅普诺夫漂移）。 假设假设 *1* 成立。考虑该排队系统中的任意调度策略，在该策略下，队列长度向量记为 $\{\mathbf{Q}_t\}$。令 $\{\theta_t : \ t \geq 1\}$表示假设 *1* 中定义的为每个队列提供服务的概率序列。则对于任意时间范围长度$T \geq 1$，我们可以找到一个仅依赖于 T的常数 $\mathcal{T}_T$，使得 $T \leq \mathcal{T}_T \leq T + \sqrt{C_W^T} + 1$ 以及

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\langle \mathbf{Q}_{t-1}, \vec{\sigma}_t \odot \vec{\theta}_t - \vec{\lambda}_t\rangle\right] \geq \epsilon\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right] - (KM^2 + \epsilon KM)C_W\mathcal{T}_T$$
$$\geq \epsilon\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right] - (KM^2 + \epsilon KM)C_W\mathcal{T}_T.$$

这里， $\odot$ 是逐元素乘积，即 $\vec{a} \odot b = （a_1 b_1, ..., a_K b_K）$ *</style}*，而 C_W 是假设 *1* 中定义的常数。

证明。 见附录 D。 □

结合定理 5.3 和定理 5.4，我们得到以下对分析至关重要的命题。

命题 5.5 (足够大的权重意味着队列稳定)。 假设假设 *1* 成立，并且调度策略保证以下条件。

$$\mathbb{E}\left[\sum_{t=1}^{T} Q_{t-1,a_t} S_{t,a_t}\right] \geq \mathbb{E}\left[\sum_{t=1}^{T}\langle \mathbf{Q}_{t-1}, \vec{\sigma}_t \odot \vec{\theta}_t\rangle\right] - f(T)$$

对所有 $T \geq \max\{\frac{4}{C_W}, C_W\}$，其中 $f(T)$ 是 T 的某个非负递增函数。那么，我们有

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_t\|_1\right] \leq \frac{(K+1)M^2 + 2C_W(KM^2 + \epsilon KM)}{\epsilon} + \frac{f(2T)}{\epsilon T}.$$

特别地，如果 $f(T)$ 为 $\emptyset(T)$，那么该策略就能使系统稳定。

证明。 参见附录 D。 □

注。 定理 5.5 表明，对于满足假设 1 的问题实例，按照要么 $\{\theta_t\}$ （参考随机策略），要么原始 MaxWeight算法（假设算法在决策时能够获得服务率预测）来服务队列，其平均队列长度不超过 $\frac{(K+1)M^2+2C_W(KM^2+\epsilon KM)}{\epsilon}$，正如我们先前在第 4 节中所述。这是因为在这两种情况下，我们都有 $\mathbb{E}[Q_{t-1,a_t}\overline{S_{t,a_t}}|\mathcal{F}_{t-1}] \geq \langle \mathbf{Q}_{t-1}, \vec{\sigma}_t \odot \vec{\theta}_t\rangle$。因此，对于这两种策略，定理 5.5 中的条件在 $f(T) = 0$下成立。

在接下来的分析中，我们将推导 SoftMW+ 对应的 $f(T)$，从而可以通过与定理 5.5 类似的论证得出队列稳定性。

5.3.2 从 EXP3.S+遗憾界到李雅普诺夫函数值界

为了建立 SoftMW（算法2）的队列稳定性结果，我们的总体思路是推导定理 5.5 中所需的条件，使得 $f(T)$ 也能得到适当控制。由于 SoftMW 基于 EXP3.S+ 做出决策，直观上，我们应利用遗憾上界结果定理 5.1。为此，我们需要验证定理 5.1 中所需的条件 (i)–(iv)[2] 是否成立。

事实上，在 SoftMW 中，我们对 η_ts 和 β_ts 的选择显然是递减的，因此条件 (iii) 和 (iv) 成立。我们选择$\mathbf{x}_1 = (1/K, \ldots, 1/K)$，因此条件 (i) 也成立；对 γ_t 和 $\mathbf{e}_t$ 的选择也保证了条件 (ii)。真正的问题在于 γ_τ 是否大于$_{1\,2}$。以下命题将证明这一点。

[2]参考策略 $\{\vec{\theta}_t\}$ 本身可能不满足条件 (v)，但我们会将每个 $\vec{\theta}_t$ 投影到 $\Delta^{[K],\ \beta_t}$ 作为 $\vec{\theta}'_t$，并且只使用定理 5.1 得到相对于动作序列 $\{\vec{\theta}'_t\}$的遗憾界。

命题 5.6（SoftMW 中探索率的可行性）。 对于所有 $t \geq 1$，我们有 $\gamma_t \leq \frac{1}{2}$ 在 *SoftMW* .

附录 E 给出了定理 5.6 的详细证明。确认该算法可行后，我们现在可以放心地应用定理 5.1，从而得到 SoftMW 的以下性质。

引理 5.7 (SoftMW 大权重保证)。 假设假设 *1* 和 *2* 成立；那么，运行算法 *2* 可保证

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}\left[\langle \mathbf{Q}_{t-1}, \mathbf{S}_t \odot \vec{\theta}_t\rangle - Q_{t-1,a_t}S_{t,a_t}\right]\\ &\leq \mathbb{E}\left\{9M(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(3\ln T+\ln K)\sqrt{86M^2K^6T^{\frac{3}{2}}+\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}+4M^2\right\}\end{aligned} \tag{4}$$

$\vec{}$对于任意时间范围长度 $T \geq 1$。此处，$\{\theta_t\}$ 是假设 *1* 和 *2* 中的参考策略。

证明。 参见附录 E。 □

定理 5.7 给出了 $\mathbb{E}\left[\sum Q_{t-1,a_t}S_{t,a_t} - \sum\langle \mathbf{Q}_{t-1}, \vec{\sigma}_t \odot \vec{\theta}_t\rangle\right]$ 的一个上界，这与定理 5.5 所要求的条件密切相关。然而，这个上界还不是一个仅取决于 T 的量；它仍包含一个因子 $\sqrt{\mathbb{E}[\sum\|\mathbf{Q}_t\|_2^2]}$，该因子取决于实际的排队轨迹。因此，我们无法直接应用定理 5.5 来证明队列稳定性。相反，我们需要处理 $\sqrt{\mathbb{E}[\sum\|\mathbf{Q}_t\|_2^2]}$ 因子，将其转换为累积队列长度 $\mathbb{E}[\sum\|\mathbf{Q}_t\|_1]$，正如我们在定理 5.4 中将 $\mathbb{E}\left[\sum\langle \mathbf{Q}_{t-1}, \vec{\sigma}_t \odot \vec{\theta}_t - \vec{\lambda}_t\rangle\right]$ 转换为队列长度时所做的那样。

5.3.3 关联遗憾 $\sqrt{\mathbb{E}[\sum\|\mathbf{Q}_t\|_2^2]}$ 与队列长度 $\mathbb{E}[\sum\|\mathbf{Q}_t\|_1]$

将式 (4) 代入定理 5.3 中的式 (3)，进一步应用定理 5.4 并重新整理各项后，我们得到以下命题，它给出了 $\sqrt{\mathbb{E}[\sum\|\mathbf{Q}_t\|_2^2]}$ 与 $\mathbb{E}[\sum\|\mathbf{Q}_t\|_1]$ 之间的不等式。

命题 5.8。 给定假设 *1* 和 *2*，算法 *2* 给出

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right]\\ &\leq \frac{(K+1)M^2+2C_W(KM^2+\epsilon KM)}{\epsilon}\mathcal{T}_T+\frac{4M^2}{\epsilon}+\frac{g(\mathcal{T}_T)}{\epsilon}\cdot\sqrt{86M^2K^6\mathcal{T}_T^{\frac{3}{2}}+\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_2^2\right]}\\ &\leq \frac{(K+1)M^2+2C_W(KM^2+\epsilon KM)}{\epsilon}\mathcal{T}_T+\frac{4M^2}{\epsilon}+\frac{\sqrt{86}MK^3\mathcal{T}_T^{\frac{3}{4}}g(\mathcal{T}_T)}{\epsilon}+\frac{g(\mathcal{T}_T)}{\epsilon}\cdot\sqrt{\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_2^2\right]}\end{aligned} \tag{5}$$

对于任意 $T \geq \max\{\frac{4}{C_W}, C_W\}$，其中 $\mathcal{T}_T$ 是不超过 $2T$ 的某个常数，且

$$g(T) = 9M(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(3\ln T+\ln K) = \widetilde{\mathcal{O}}(T^{\frac{1}{4}-\frac{\delta}{2}}).$$

回顾一下，假设所有到达量和离开量都由常数 M 限定。因此，队列长度向量 $\{\mathbf{Q}_t\}$ 的每个维度都是一个非负数序列，其中任意两个相邻项之差都不超过 $\pm M$。于是，我们可以使用以下关于这类有界差分序列的引理。

引理 5.9。 假设对于所有 $1 \leq i < n$，$x_1 = 0$、x_2、.....、$x_n \geq 0$、$|x_{i+1} - x_i| \leq 1$。记为 $S = \sum_{i=1}^{n} x_i$；则有

$$\sum_{i=1}^{n} x_i^2 \leq 4S^{\frac{3}{2}}.$$

证明。 参见附录 E。 □

对于我们的目的，定理 5.9 保证

$$\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2 \le 4\sqrt{M}\sum_{i=1}^{K}\left(\sum_{t=1}^{T}Q_{t-1,i}\right)^{\frac{3}{2}} \le 4\sqrt{M}\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{2}}. \tag{6}$$

然后，将公式(6)代入公式(5)，我们得到以下完全取决于 $\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|_{t-1}\|_1\right]$ **队列**的不等式：

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right] \le h(\mathcal{T}_T) + g(\mathcal{T}_T)\left(\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right]\right)^{\frac{3}{4}} \tag{7}$$

其中

$$g(T) = 18M^{\frac{5}{4}}(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(3\ln T + \ln K) = \widetilde{\mathcal{O}}\left(\frac{T^{\frac{1}{4}-\frac{\delta}{2}}}{\epsilon}\right),$$

$$h(T) = \frac{(K+1)M^2 + 2C_W(KM^2+\epsilon KM)}{\epsilon}T + \widetilde{\mathcal{O}}(T^{1-\frac{\delta}{2}}).$$

为了得到 $\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|_{t-1}\|_1\right]$ **队列**的上界，仍需解公式(7)。为此，我们使用以下引理。

引理5.10。设 y、f、g：$\mathbb{R}_+ \to [1,\infty)$ 是三个非递减函数。若

$$y(x) \le f(x) + y(x)^{\frac{1}{4}}g(x)$$

对所有 $x \ge 0$，则有

$$y(x) \le \left(f(x)^{\frac{1}{4}} + g(x)\right)^4.$$

证明。见附录 E。 □

最后，根据定理 5.10，公式(7)的解给出：

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right] &\le \left(h(\mathcal{T}_T)^{\frac{1}{4}} + g(\mathcal{T}_T)\right)^4 \\ &\le \frac{(K+1)M^2 + 2C_W(KM^2+\epsilon KM)}{\epsilon}\mathcal{T}_T + o(\mathcal{T}_T).\end{aligned}$$

因此，

$$\begin{aligned}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right] \le \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right] &\le \frac{(K+1)M^2 + 2C_W(KM^2+\epsilon KM)}{\epsilon}\frac{\mathcal{T}_T}{T} + o(\mathcal{T}_T/T) \\ &\le \frac{2(K+1)M^2 + 4C_W(KM^2+\epsilon KM)}{\epsilon} + o(1)\end{aligned}$$

这正是所需结果。

6 驯服时间齐次的$\mathcal{O}(T^{1-\delta})$ 参考策略全变差

在本节中，我们提出另一种能够稳定我们的对抗排队系统的新颖算法。具体而言，只要总变差在无限时间范围内在某种程度上“均匀”分布，该算法便能在总变差为 $O(T^{1-\delta})$ 的参考随机策略下保持稳定。下面将这一新条件形式化表述。

假设3 (时间齐次参考策略平稳性)**.** 对于假设 *1* 中的序列 $\{\theta_t\}$，存在某些 $\delta > 0$ 和 $C_V > 0$，使得

$$\sum_{t=T_0+1}^{T_0+T-1}\|\theta_{t+1} - \theta_t\|_1 \le C_V T^{1-\delta}$$

对于任意 $T_0 \ge 0$ 和 $T \ge 1$。

注。 假设3可视为假设 2 的平移不变版本，其中 T 的程度从 $\frac{1}{2}-\delta$ 放宽至 $1-\delta$。粗略地说，只要参考策略变化以线性速率累积的时间段数量有限，该假设就成立。例如，如果 $\sum_{t=0}^{T}\|\theta_{t+1}-\theta_t\|_1=\Theta(T^{1-\delta})$，那么假设3成立。

对于满足假设 1 和 3 的问题实例，我们提出一种用于稳定系统的新算法，即**滑动软MW** (SSMW)，详见算法 3。

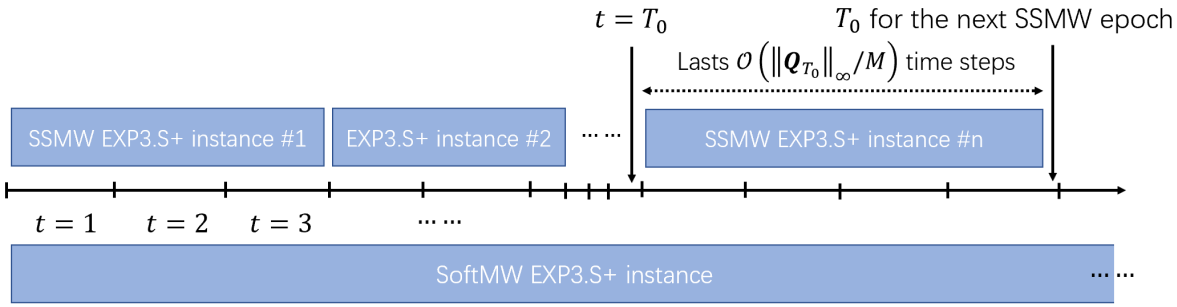


图2：SSMW 与 SoftMW 在 EXP3.S+ 调用上的比较

算法 3： SSMW（**滑动窗口软MW**）

输入： One-s 到达/服务矩上界参数 $M>0$，作业类型数量 K，问题实例 光滑度参数 δ ¿ 0

输出： 要服务的作业类型序列 $a_1,\ a_2,\ \ldots\in[K]$

1 当为真 **时执行**

2 $T_0\leftarrow$ 最近一次做出新决策的时间索引 t a_t // 对于第一次迭代，我们应有 $T_0=0$3 $m\leftarrow \max\left\{\left\lceil\frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M}\right\rceil,1\right\}$4 运行一个新的 EXP3.S+ 实例，持续 m 个时间步，并采用以下配置（下文中的 τ 表示长度为 m的时期内的时间索引，从 1 开始）：5 $\beta=m^{-2}/K$6 x_1 可以是 $\Delta^{[K]}$中的任意元素，$\beta\triangleq\{\mathrm{x}\in\Delta^{[K]}:\ \mathrm{x}_i\geq\beta\quad\forall i\in[K]\}$7 $\eta_\tau=\left(6M^2Km^{1+\frac{\delta}{2}}\right)$ 8 $\mathrm{e}_\tau=\mathrm{Q}_{T_0+\tau-1}/\|\mathrm{Q}_{T_0+\tau-1}\|_1$

9 $\gamma_\tau=M\eta_\tau\|\mathrm{Q}_{T_0+\tau-1}\|_1=\frac{1}{6}K^{-1}M^{-1}m^{-1-\frac{\delta}{2}}\|\mathrm{Q}_{T_0+\tau-1}\|_1$10 从当前 EXP3.S+ 实例获取一个新的动作决策输出，服务此类型的作业（回想一下，我们处于整个时间范围的第 $(T_0+\tau)$ 个时间步），将 $Q_{T_0+\tau-1}$, $a_{T_0+\tau}$ $S_{T_0+\tau},^{a_{T_0+\tau}}$视为新的反馈 g_{τ,a_τ}并将其输入 EXP3.S+

与 SoftMW（算法2）相比，SSMW（算法 3）在开始时不使用历史队列长度来调整 EXP3.S+ 学习率。相反，SSMW 会启动新的 EXP3.S+ 问题实例，其长度与当前队列长度成正比（第 3 行）。因此，SSMW 在整个执行过程中会启动更多的 EXP3.S+ 问题实例，尽管每个 EXP3.S+ 周期可能都很短（如图2所示）。从这个意义上说，SSMW 与 MaxWeight 更为相似，因为 MaxWeight 始终使用当前队列长度向量来做出新的决策，而忽略系统是如何到达当前状态的。定理 6.1 给出了 SSMW 的队列稳定性结果。

定理 em 6.1. 对于满足假设 *1* 和 *3* 的问题实例， *SSMW* （算法 *3*）保证 s

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_t\|_1\right]\leq\left[3KM^2m_0+\frac{(K+1)M^2}{2}+(KM^2+\epsilon KM)C_W+6M^2\right]\cdot\frac{10}{\epsilon}$$

对于任意 ti 长度为 $T\geq{}^{4}_{C_W}+C_W$ 的时间范围。特别是，该系统是稳定的。这里将 m_0 定义为

$$m_0\triangleq\inf\left\{m:m\geq2,f(m')\leq\frac{\epsilon}{2}\forall m'\geq m\right\}\leq\left((1+C_V)MK\ln K\epsilon^{-1}\right)^{\emptyset(1/\delta)}$$

其中

$$f(m) = 88(1 + C_V)MKm^{-\frac{\delta}{2}}(2\ln m + \ln K).$$

注。 与 SoftMW 的 $\emptyset(\epsilon^{-1})$ 队列长度界（定理 5.2）相比，定理 6.1 仅给出了一个 $\epsilon^{\emptyset(1/\delta)}$队列长度保证。尽管如此，附录 A 中的仿真结果表明，SSMW 的实测性能与 SoftMW 相当，甚至更好。

6.1 SSMW 的队列稳定性分析概要

本节概述定理 6.1 的证明。首先，应用定理 5.1，可得到定理 6.2 的结论，即 SSMW 的每个 EXP3.S+实例相较于相应参考策略的性能。

引理 6.2。 假设假设 *1* 和 *3* 成立，则令 T_0 为算法 *3* 中启动一个长度为 m 的新 *EXP3.S+*实例的某个时间步；那么有

$$\begin{aligned}
&\mathbb{1}[T_0 \text{ 结束一个 } \textit{EXP3.S+} \text{ 实例，而新的 } \textit{EXP3.S+} \text{ 实例长度为 } m] \\
&\cdot \sum_{t=1}^{m} \mathbb{E}\left[\langle \mathbf{Q}_{T_0+t-1} \odot \mathbf{S}_{T_0+t}, \vec{\theta}_{T_0+t}\rangle - Q_{T_0+t-1,a_{T_0+t}} S_{T_0+t,a_{T_0+t}} \middle| \mathcal{F}_{T_0}\right] \\
&\leq 6\left(1 + C_V\right) M^2 K m^{2-\frac{\delta}{2}} \cdot (2\ln m + \ln K) + K^{-1} m^{-1-\frac{\delta}{2}} \mathbb{E}\left[\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_2^2 \middle| \mathcal{F}_{T_0}\right] + 6M^2. \qquad (8)
\end{aligned}$$

证明。参见附录 F。 □

与分析 SoftMW 时在得到定理 5.7 后所做的类似，我们将定理 6.2 中遗憾的上界与实际累积队列长度 $\mathbb{E}[\sum\|\mathbf{Q}_t\|_1]$ 联系起来。为此，我们需要若干关于有界增量序列求和的引理。

引理 6.3。 对于任意 $T_0 \geq 1$，假设 $M \geq 0$, $\|\mathbf{Q}_{T_0}\|_\infty \geq 4M$，那么对于任意 $T_0 + 1 \leq t \leq T_0 + \frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M} + 1$，有

$$\frac{1}{4}\|\mathbf{Q}_{T_0}\|_\infty \leq \frac{1}{2}\|\mathbf{Q}_{T_0}\|_\infty - M \leq \|\mathbf{Q}_{t-1}\|_\infty \leq \frac{3}{2}\|\mathbf{Q}_{T_0}\|_\infty + M \leq 2\|\mathbf{Q}_{T_0}\|_\infty.$$

令 $T = \lceil \frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M} \rceil$；则有

$$\frac{1}{4}MT \leq \|\mathbf{Q}_{T_0+t-1}\|_\infty \leq 4MT$$

对于任意 $1 \leq t \leq T$。此外，

$$\frac{1}{16}M^2T^3 \leq \sum_{t=T_0+1}^{T_0+T} \|\mathbf{Q}_{t-1}\|_2^2 \leq 16KM^2T^3,$$

$$\frac{1}{4}MT^2 \leq \sum_{t=T_0+1}^{T_0+T} \|\mathbf{Q}_{t-1}\|_1 \leq 4KMT^2.$$

证明。 请参见附录 F。 □

现在可以应用定理 6.3，将定理 6.2 中的遗憾上界与队列长度和 $\sum\|\mathbf{Q}_t\|_1$) 联系起来，从而得到定理 6.4 中的结果。

引理 6.4。 假设假设 *1* 和 *3* 成立。设 T_0 为我们在算法 *3* 中启动一个长度为 m 的新 *EXP3.S+* 实例时的时间步。那么，我们有

$$\begin{aligned}
&\mathbb{1}\left[T_0 \textit{ ends an EXP3.S+ instance, and the new EXP3.S+ instance is of length } m = \left\lceil \frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M} \right\rceil, m \geq 2\right] \\
&\cdot \sum_{t=1}^{m} \mathbb{E}\left[\langle \mathbf{Q}_{T_0+t-1} \odot \mathbf{S}_{T_0+t}, \vec{\theta}_{T_0+t}\rangle - Q_{T_0+t-1,a_{T_0+t}} S_{T_0+t,a_{T_0+t}} \middle| \mathcal{F}_{T_0}\right]
\end{aligned}$$

$$\leq 6M^2 + \mathbb{E}\left[f(m) \cdot \sum_{t=1}^{m} \|\mathbf{Q}_{T_0+t-1}\|_1 \middle| \mathcal{F}_{T_0} \right],$$

其中

$$f(m) = 88(1+C_V)MKm^{-\frac{\delta}{2}}(2\ln m + \ln K).$$

证明。 参见附录 F。 □

与 SoftMW 相比，SSMW 的一次执行包含多个 EXP3.S+ 执行，其开始时间和结束时间都是随机量。为了在分析中严谨地处理这些随机的 EXP3.S+ 时期，我们将引入一些额外的符号。记 τ_i ($i \geq 0$) 为第 i 个 EXP3.S+ 实例结束的时间。那么，$\tau_0 = 0$ 且 $\{\tau_i\}$ 是一个非递减的 $\{\mathcal{F}_t\}$-适应停止时序列。此外，每个 τ_{i+1} 都是 $\mathcal{F}_{\tau_i}$-可测的。固定任意 $T \geq 1$ 并定义

$$\tau_i' \triangleq \begin{cases} 0 & \text{if } i = 0 \\ \tau_i & \text{if } i > 0 \text{ and } \tau_{i-1}' < T \\ \tau_{i-1}' & \text{otherwise} \end{cases},$$

也就是说，τ_i'可以视为时期结束时间 τ_i，但截断于 T；在考虑截至时间 T 的累积遗憾时，它比 τ_i 更方便。于是，$\{\tau_i'\}$ 也是一个非递减的 $\{\mathcal{F}_t\}$-适应停止时序列，其中每个 τ_{i+1}' 都是 $\mathcal{F}_{\tau_i'}$-可测的，并且 $\tau_{i+1}' = \tau_i'$ 当且仅当 $\tau_i' \geq T$。因此，我们可以将定理 6.4 改述为定理 6.5。

引理 6.5. 假设假设 *1* 和 *3* 成立。那么，我们有

$$\mathbb{1}\left[\|\mathbf{Q}_{\tau_i'}\|_\infty \geq 4M\right] \sum_{t=1}^{\tau_{i+1}'-\tau_i'} \mathbb{E}\left[\langle \mathbf{Q}_{\tau_i'+t-1} \odot \mathbf{S}_{\tau_i'+t}, \vec{\theta}_{\tau_i'+t} \rangle - Q_{\tau_i'+t-1,a_{\tau_i'+t}} S_{\tau_i'+t,a_{\tau_i'+t}} \middle| \mathcal{F}_{\tau_i'} \right]$$
$$\leq h(\tau_{i+1}' - \tau_i') + g(\tau_{i+1}' - \tau_i') \cdot \sum_{t=1}^{\tau_{i+1}'-\tau_i'} \mathbb{E}\left[\|\mathbf{Q}_{\tau_i'+t-1}\|_1 \middle| \mathcal{F}_{\tau_i'} \right]$$

对于任意 $i \geq 0$，其中

$$g(m) = 88(1+C_V)MKm^{-\frac{\delta}{2}}(2\ln m + \ln K), \tag{9}$$
$$h(m) = \mathbb{1}[m > 0] \cdot 6M^2. \tag{10}$$

与定理 6.4 相比，定理 6.5 允许我们自由地对各个时期的遗憾界求和，而无需顾虑有限时间范围长度 T 带来的细微问题。

接下来，我们将利用引理 6.5 来界定 $\mathbb{E}[\sum_{t=1}^{\mathcal{T}_1}\|\mathbf{Q}_{t-1}\|_1]$。首先，固定某个 $T \geq 1$，令 $\mathcal{T}_0 \triangleq \sup\{\tau_i : i \geq 0, \tau_i < T\}$，并令 $\mathcal{T}_1 \triangleq \inf\{\tau_i : i \geq 0, \tau_i \geq T\}$。我们看到，$\mathcal{T}_0$ 和 $\mathcal{T}_1$ 都是 $\{\mathcal{F}_t\}$- 适应停止时，而 $\mathcal{T}_1$ 是 $\mathcal{F}_{\mathcal{T}_0}$- 可测的。注意到 $\mathcal{T}_0 < T \leq \mathcal{T}_1$。此外，由于 $\mathcal{T}_1 - \mathcal{T}_0$ 是前 T 个时间步中最后一个时期的长度，我们有 $\mathcal{T}_1 - \mathcal{T}_0 \leq \|\mathbf{Q}_{\mathcal{T}_0}\|_\infty + 1 \leq \frac{\mathcal{T}_0 \cdot M}{2M} + 1 = \frac{\mathcal{T}_0}{2} + 1 \leq \frac{\mathcal{T}_0}{2} + T$，因此可知 $\mathcal{T}_1 \leq \frac{5}{2}T$。

在本节余下部分，我们结合定理 5.3 和定理 6.5，将 $\mathbb{E}[\sum_{t=1}^{\mathcal{T}_1}\|\mathbf{Q}_{t-1}\|_1]$ 限定为 $\mathcal{O}(\mathbb{E}[\mathcal{T}_1]) = \mathcal{O}(T)$，从而得出 $\mathbb{E}[\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1]$ 也为 $\mathcal{O}(T)$。

为此，回顾一下，$\epsilon > 0$ 是假设 1 中参考策略 $\{\vec{\theta}_t\}$ 的“离开相对于到达的平均优势”的下界。定义

$$m_0 \triangleq \inf\left\{ m : m \geq 2, g(m') \leq \frac{\epsilon}{2} \forall m' \geq m \right\}.$$

那么，m_0 是一个仅取决于 δ 和 ϵ的常数；事实上，它的解为

$$m_0 \leq \left((1+C_V)MK\ln K\epsilon^{-1} \right)^{\mathcal{O}(1/\delta)}.$$

通过考察每个时期长度 $\tau_{i+1}' - \tau_i'$ 是否大于 m_0，我们根据定理 6.5 得出，对于所有 $i \geq 0$,

$$\sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[\langle \mathbf{Q}_{\tau'_i+t-1} \odot \mathbf{S}_{\tau'_i+t}, \vec{\theta}_{\tau'_i+t}\rangle - Q_{\tau'_i+t-1,a_{\tau'_i+t}} S_{\tau'_i+t,a_{\tau'_i+t}} \middle| \mathcal{F}_{\tau'_i}\right]$$

$$\leq \underbrace{h(\tau'_{i+1}-\tau'_i) + \frac{\epsilon}{2}\sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1 \middle| \mathcal{F}_{\tau'_i}\right]}_{\text{when } \tau'_{i+1}-\tau'_i > m_0\text{, apply Theorem 6.5}} + \underbrace{\mathbb{1}[\tau'_{i+1}-\tau'_i \leq m_0] \sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[\langle \mathbf{Q}_{\tau'_i+t-1} \odot \mathbf{S}_{\tau'_i+t}, \vec{\theta}_{\tau'_i+t}\rangle \middle| \mathcal{F}_{\tau'_i}\right]}_{\text{when } \tau'_{i+1}-\tau'_i \leq m_0\text{, simply drop the minus-signed term}}$$

$$\overset{(a)}{\leq} h(\tau'_{i+1}-\tau'_i) + \frac{\epsilon}{2}\sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1 \middle| \mathcal{F}_{\tau'_i}\right] + \mathbb{1}[\tau'_{i+1}-\tau'_i \leq m_0] M \sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1 \middle| \mathcal{F}_{\tau'_i}\right]$$

$$\overset{(b)}{\leq} h(\tau'_{i+1}-\tau'_i) + \frac{\epsilon}{2}\sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1 \middle| \mathcal{F}_{\tau'_i}\right] + \mathbb{1}[\tau'_{i+1}-\tau'_i \leq m_0] KM \sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[\|\mathbf{Q}_{\tau'_i}\|_\infty + Mm_0 \middle| \mathcal{F}_{\tau'_i}\right]$$

$$\overset{(c)}{\leq} h(\tau'_{i+1}-\tau'_i) + \frac{\epsilon}{2}\sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1 \middle| \mathcal{F}_{\tau'_i}\right] + \mathbb{1}[\tau'_{i+1}-\tau'_i \leq m_0] KM \sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[3Mm_0 \middle| \mathcal{F}_{\tau'_i}\right]$$

$$\leq h(\tau'_{i+1}-\tau'_i) + \frac{\epsilon}{2}\sum_{t=1}^{\tau'_{i+1}-\tau'_i} \mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1 \middle| \mathcal{F}_{\tau'_i}\right] + 3KM^2 m_0 (\tau'_{i+1}-\tau'_i) \tag{11}$$

这里， $h(m)$ 是式 (10) 中的函数。在步骤 (*a*) 和 (*b*) 中，我们直接利用队列长度增量有界这一假设。步骤 (*c*) 则是因为，只要 $\tau'_{i+1}-\tau'_i > 0$， $\|\mathbf{Q}_{\tau'_i}\|_\infty \leq 2M(\tau'_{i+1}-\tau'_i) \leq 2Mm_0$；也就是说， i 不是前 T 个时隙之后某个时期的索引。

对所有 $i \geq 0$ 对式（11）求和，再取全期望，可得

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1} \langle \mathbf{Q}_{t-1} \odot \mathbf{S}_t, \vec{\theta}_t\rangle - Q_{t-1,a_t} S_{t,a_t}\right] \leq \frac{\epsilon}{2}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1} \|\mathbf{Q}_{t-1}\|_1\right] + 3KM^2 m_0 \mathbb{E}\left[\mathcal{T}_1\right] + \mathbb{E}\left[\sum_{i=0}^{\infty} h(\tau'_{i+1}-\tau'_i)\right].$$

由（10）可知，在任意样本路径上， $\sum_{i=0}^{\infty} h(\tau'_{i+1}-\tau'_i) \leq 6M^2\mathcal{T}_1$。因此，

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1} \langle \mathbf{Q}_{t-1} \odot \mathbf{S}_t, \vec{\theta}_t\rangle - Q_{t-1,a_t} S_{t,a_t}\right] \leq \frac{\epsilon}{2}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1} \|\mathbf{Q}_{t-1}\|_1\right] + (3KM^2 m_0 + 6M^2)\mathbb{E}\left[\mathcal{T}_1\right]. \tag{12}$$

根据定理 5.4，我们还可以找到一个取决于 $\mathcal{T}_1$的常数 $\mathcal{T}_2$，使得 $\mathcal{T}_2 \leq \mathcal{T}_1 + \sqrt{\frac{\mathcal{T}_1}{C_W}} + 1$，并且

$$-\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_2} \langle \mathbf{Q}_{t-1}, \vec{\sigma}_t \odot \vec{\theta}_t - \vec{\lambda}_t\rangle\right] \leq -\epsilon\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1} \|\mathbf{Q}_{t-1}\|_1\right] + (KM^2 + \epsilon KM) C_W \mathbb{E}[\mathcal{T}_2]. \tag{13}$$

若$T \geq \frac{4}{C_W} + C_W$，则 $\mathcal{T}_1 \geq \max\{C_W, \frac{4}{C_W}\}$，因此 $\mathcal{T}_2 \leq 2\mathcal{T}_1 \leq 5T$。此外，定理 5.3 保证

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_2} Q_{t-1,a_t} S_{t,a_t} - \langle \mathbf{Q}_{t-1}, \vec{\lambda}_t\rangle\right] \leq \frac{(K+1)M^2 \mathbb{E}[\mathcal{T}_2]}{2}. \tag{14}$$

将式 (12) 至 (14) 直接相加，得到

$$\begin{aligned}\frac{\epsilon}{2}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1} \|\mathbf{Q}_{t-1}\|_1\right] &\leq \left[\frac{(K+1)M^2}{2} + (KM^2 + \epsilon KM)C_W + 3KM^2 m_0 + 6M^2\right]\mathbb{E}[\mathcal{T}_2] \\ &\leq \left[\frac{(K+1)M^2}{2} + (KM^2 + \epsilon KM)C_W + 3KM^2 m_0 + 6M^2\right] \cdot 5T.\end{aligned}$$

因此，当 $T \geq \frac{4}{C_W} + C_W$ 时，我们有

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \|\mathbf{Q}_{t-1}\|_1\right] \leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1} \|\mathbf{Q}_{t-1}\|_1\right] \leq \left[3KM^2 m_0 + \frac{(K+1)M^2}{2} + (KM^2 + \epsilon KM)C_W + 6M^2\right] \cdot \frac{10}{\epsilon},$$

从而完成定理 6.1 的证明。

7 放宽队列增量的有界性假设

在本节中，我们进一步将 SoftMW 和 SSMW 扩展到如下情形：队列长度增量（单个到达量和离开量）不一定被限制在一个已知范围内，但具有有界矩。形式上，我们将用以下新假设替换问题设定中的有界到达与服务假设。

假设 4 (队列长度增量具有有界矩**)**。 到达与服务分布 $\{\mathcal{A}_1, \mathcal{A}_2, \ldots\}$ 和 $\{\mathcal{S}_1, \mathcal{S}_2, \ldots\}$ 的支持集为 $\mathbb{R}_+^K$，但存在常数 $\alpha \geq 2$ 和 $M > 0$，系统调度器事先已知这两个常数，使得

$$\mathbb{E}\left[A_{t,i}^{\alpha} \middle| \mathcal{F}_{t-1}\right], \mathbb{E}\left[A_{t,i}^{\alpha} \middle| \mathcal{F}_{t-1}\right] \leq M^{\alpha}$$

对所有 $t \geq 1$ 和 $i \in [K]$均成立。由此可立即推出，对所有 $t \geq 1$ 和 $i \in [K]$，也有

$$\mathbb{E}\left[|Q_{t,i} - Q_{t-1,i}|^{\alpha} \middle| \mathcal{F}_{t-1}\right] \leq 2M^{\alpha}.$$

请注意，虽然在标准李雅普诺夫分析中，这一假设通常并不难满足 [28], ，但在学习增强控制分析中却带来了新的挑战，尤其是在算法使用 UCB 奖励项（例如 [9, 13, 21, 40]）时，主要是因为它对估计造成了影响。[3]

下面，我们介绍 SoftMW 和 SSMW 的新变体，它们能够在假设 4 下稳定系统。我们在附录G中解释高层次的设计思路，并在附录H和I中给出详细的队列稳定性证明。

算法 4: 适用于具有有界矩的队列长度增量的 SoftMW（SoftMW+）

输入: 队列 长度增量矩上界参数 $M > 0$、 $\alpha > 14$，作业类型数量 K，
问题实例平滑性 p参数 $0 < \delta \leq$ [1]
输出: </style}待服务的作业类型序列 , $a_2, \ldots \in [K]$

1 $L_0 \leftarrow M$ 初始化一个扩展版 EXP3.S+ 实例。

3 **对于** $t = 1, 2, \cdots$ **执行**

4 为时隙 t选取以下 EXP3.S 参数:

5 $\beta_t \leftarrow t^{-4}/K$

6 $\eta_t = \left(t^{-\left(\frac{1}{4}-\frac{\delta}{2}\right)} L_{t-1} \sqrt{86 L_{t-1}^2 K^6 t^{\frac{3}{2}} + \sum_{s=0}^{t-1} \|\mathbf{Q}_s\|_2^2}\right)^{-1}$

7 $\mathbf{e}_t = \mathbf{Q}_{t-1}/\|\mathbf{Q}_{t-1}\|_1$

8 $\gamma_t = M t^{\frac{\delta}{4}} \eta_t \|\mathbf{Q}_{t-1}\|_1 = t^{\frac{\delta}{4}} M L_{t-1}^{-1} \|\mathbf{Q}_{t-1}\|_1 \left(t^{-\left(\frac{1}{4}-\frac{\delta}{2}\right)} \sqrt{86 L_{t-1}^2 K^6 t^{\frac{3}{2}} + \sum_{s=0}^{t-1} \|\mathbf{Q}_s\|_2^2}\right)^{-1}$

9 从 EXP3.S+中获取新的动作决策输出 a_t，服务第 a_t 个队列，并接收反馈S_{t,a_t}

10 $S_t' \triangleq \begin{cases} S_{t,a_t} & \text{如果 } S_{t,a_t} \leq M t^{\frac{\delta}{4}} \\ 0 & \text{否则} \end{cases}$

11 将 $Q_{t-1,a_t} S_t'$ 视为新的反馈 g_{t,a_t}，并将其输入当前的 EXP3.S+ 实例

12 $L_t \leftarrow \max\left\{L_{t-1}, \|\mathbf{Q}_t - \mathbf{Q}_{t-1}\|_\infty\right\}$

SoftMW+ （算法 4）是针对假设 4 的 SoftMW 泛化版本。与 SoftMW（算法2）相比，SoftMW+ （算法 4）会显式跟踪 L_t （截至时间t遇到的最大队列长度增量），并且算法 2 中大多数对队列长度增量上界 M 的引用，在新算法中都替换为 L_t。我们还将显式探索率提高了 $t^{\frac{\delta}{4}}$ 倍（第 8 行），并在将服务反馈输入 EXP3.S+ 之前对其进行截断（第 10 行）。

下面的定理 7.1 给出了任意有限时间范围内相应的平均队列长度上界。

定理 7.1 。 对于满足假设 *1*、*2* 和 *4* 的问题实例， *SoftMW+* （算法 *4*）保证 *s*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \|\mathbf{Q}_t\|_1\right] \leq \frac{2(K+1)M^2 + 4C_W(KM^2 + \epsilon KM)}{\epsilon} + o(1)$$

只要 $\delta \cdot \alpha > 7$。特别地，该系统是稳定的。

[3]例如，基于 UCB 的估计器通常要求分布为次高斯分布；与我们的假设 4 相比，这一假设的限制严格得多。

证明。 见附录 H。 □

接下来，我们介绍 SSMW+ （算法 5），即 SSMW 的广义版本。与 SoftMW+ （算法 4）相比，SSMW+不需要知道假设 4 中 α 的实际值，并且能够稳定任意新到达任务和服务的二阶矩有界的问题实例（假设 4 中的任意 $\alpha \geq 2$ 也意味着这一点）。相比之下，算法 4 需要预先知道 α ，并要求 δ 与 α 的乘积不能太小。

算法 5： 具有有界矩的队列长度增量的 SSMW（SSMW+）

输入： 队列长度增量矩上界参数 $M > 0$、作业类型数量 K、问题实例光滑度参数 $\delta > 0$

输出： 要服务的作业类型序列 $a_1, a_2, \ldots \in [K]$

1 当条件为真 **时执行**

2 $T_0 \leftarrow$ 我们做出新决策时的最新时间索引 t a_t // 对于第一次迭代，应有 $T_0 = 0$

3 $m \leftarrow$ 最大值 $\left\{\left\lceil \frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M} \right\rceil, 1\right\}$

4 使用以下配置运行一个全新的 EXP3.S+ 实例，持续 m 个时间步（下文中的 τ 表示长度为 m 的时期内的时间索引，从 1 开始计数）：

5 $\beta = m^{-3}/K$

6 $\mathbf{x}_1$ 可以是 $\Delta^{[K]}$ 中的任意元素，$\beta \triangleq \{\mathbf{x} \in \Delta^{[K]} : \mathbf{x}_i \geq \beta \quad \forall i \in [K]\}$

7 $\eta_\tau = \left(4M^3 K m^{1+\frac{2}{3}\delta}\right)$

8 $\mathbf{e}_\tau = \mathbf{1}/K$

9 $\gamma_\tau = m^{\frac{\delta}{3}} KM\eta_\tau \|\mathbf{Q}_{T_0}\|_\infty = \frac{1}{4}M^{-2}m^{-1-\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty$

10 从 EXP3.S+获取一个新的动作决策输出 a_t ，服务第 a_t 个队列，并接收反馈 $S_{T_0+\tau, a_{T_0+\tau}}$

11 $g_{T_0+\tau} \triangleq \begin{cases} Q_{T_0+\tau-1, a_{T_0+\tau}} S_{T_0+\tau,} & \\ 0 & \text{如果 } Q_{T_0+\tau-1, a_{T_0+\tau}} S_{T_0+\tau,} \\ & a_{T_0+\tau} \leq m^{\frac{\delta}{3}} M Q_{T_0,} \end{cases}$

否则 12 将 $g_{T_0+\tau}$ 视为新的反馈，并将其输入 EXP3.S+ 实例

定理 7.2 给出了任意有限时间范围内相应的平均队列长度上界。**定理**

7.2 **7.2.** 对于满足假设 *1*、*3* 和 *4* 的问题实例， *SSMW+ (*算法 *5)* 保证 s

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_t\|_1\right] \leq \left[3KM^2 m_0 + \frac{(K+1)M^2}{2} + (KM^2 + \epsilon KM)C_W + 4M^2\right] \cdot \frac{10}{\epsilon}$$

对于任意时间范围长度 $T \geq \frac{4}{C_W} + C_W$。特别地，系统是稳定的。这里 m_0 定义为

$$m_0 \triangleq \inf\left\{m : m \geq 2, f(m') \leq \frac{\epsilon}{2} \forall m' \geq m\right\} \leq \left((1+C_V)M^2 K \ln K \epsilon^{-1}\right)^{\varnothing(1/\delta)}$$

其中

$$f(m) = 42(1+C_V)M^2 K m^{-\frac{\delta}{3}} \cdot (3\ln m + \ln K)\,.$$

证明。 见附录 I。 □

8 相关工作

近期文献中包括一些基于学习的调度策略，这些策略几乎不需要先验知识，并且能够在运行时收集信道统计信息。

在缺乏完美信道状态知识的情况下，基于学习的方法调度排队系统需要进行大量探索，以探寻系统内所有信道的更多信息，而不是仅仅利用现有的统计信息（例如通过 MaxWeight 风格的规划）。引入充分探索的典型方法包括 ε-贪心，该方法明确分配一个较小的概率来无条件地服务每个信道 [30, 20, 21]；在此方法中，探索对队列大小和历史信道统计信息视而不见，因此几乎与利用相互独立。相比之下，乐观探索通过向当前信道统计信息中添加奖励项来实现，从而使探索与

利用在调度过程中自然耦合 [9, 21, 35, 40]。上置信界（UCB） [2] 是设计奖励项的经典方法。

关于非平稳排队系统中调度的现有研究包括 [40],，该方法使用折扣上置信界估计器来估计每条链路的最新服务率，以替代经典 MaxWeight 中的实际平均服务率。由此得到的策略能够稳定这样的问题实例：在任意长度为 W 的时间窗口内，任意两个时间步之间每条链路的到达率（和服务率）之差都足够小；并且该窗口长度 W 需要与折扣上置信界估计器所使用的折扣因子 γ 相匹配。与 [40], 相比，我们的平滑性假设针对的是参考随机策略，而非真实服务率。

9 结论与未来工作

本文提出了一种新颖的方法，将对抗性赌博机学习技术应用于信道状态未知且时变的排队系统调度。本文提出的新算法 SoftMW 和 SSMW 能够在系统可由某个（可能未知的）随机策略序列稳定，且其时间变化满足某些温和条件时稳定该系统。我们进一步将结果推广到到达量和离开量仅具有有界矩的情形，并提出了两种稳定算法 SoftMW+ 和 SSMW+。

我们认为，我们的方法可以推广到更复杂的随机网络（例如，多跳网络），并用于实现其他任务，例如在保证队列稳定性的前提下进行效用优化。利用对抗性赌博机学习技术设计分布式网络调度算法，也是一个值得探索的未来研究方向。

参考文献

[1] Mohammad Ashjaei、Lucia Lo Bello、Masoud Daneshtalab、Gaetano Patti、Sergio Saponara 和 Saad Mubeen。2021。汽车嵌入式系统中的时间敏感网络：现状与研究机遇。系统架构杂志 117 (2021), 102137。[2] Peter Auer。2002。利用置信界权衡利用与探索。机器学习研究杂志3, Nov (2002), 397–422。[3] Peter Auer、Nicolo Cesa-Bianchi、Yoav Freund 和 Robert E Schapire。2002。非随机多臂老虎机问题。*SIAM*计算杂志 32, 1 (2002), 48–77。[4] Oussama Ben-Ammar、Belgacem Bettayeb 和 Alexandre Dolgui。2019。随机交货期和动态需求下的多周期供应计划优化。《国际生产经济学杂志》218 (2019), 106–117。[5] Benjamin Berg、Mor Harchol-Balter、Benjamin Moseley、Weina Wang 和 Justin Whitehouse。2020。弹性与非弹性作业的最优资源分配。载于第*32*届*ACM*并行算法与架构研讨会论文集。75–87。[6] Anton Braverman、Jim G Dai、Xin Liu 和 Lei Ying。2017。面向现代拼车系统的基于流体模型的车辆路径规划。载于*2017*年*ACM SIGMETRICS*/计算机系统测量与建模国际会议论文集。11–12。[7] Anton Braverman、Jim G Dai、Xin Liu 和 Lei Ying。2019。拼车系统中的空车路径规划。运筹学67, 5 (2019), 1437–1452。[8] Jiangong Chen、Feng Qian 和 Bin Li。2022。利用商用移动设备提升协作式虚拟现实的体验质量。载于*2022*年*IEEE*第*42*届分布式计算系统国际会议（*ICDCS*）。IEEE，1018–1028。[9] Tuhinangshu Choudhury、Gauri Joshi、Weina Wang 和 Sanjay Shakkottai。2021。服务率未知的排队系统的作业派发策略。载于第二十二届移动网络与移动计算理论、算法基础及协议设计国际研讨会论文集。181–190。[10] Said El Kafhali 和 Khaled Salah。2017。云计算数据中心的随机建模与分析。载于*2017*年第*20*届云、互联网与网络创新会议（*ICIN*）。IEEE，122–126。[11] Anuroop Gaddam、Tim Wilkin、Maia Angelova 和 Jyotheesh Gaddam。2020。物联网中的传感器故障、异常和离群值检测：挑战与解决方案综述。电子学9, 3 (2020), 511。[12] Elad Hazan 等。2016。在线凸优化导论。优化基础与趋势®2, 3-4 (2016), 157–325。

[13] Wei-Kang Hsu、Jiaming Xu、Xiaojun Lin 和 Mark R Bell。2022。不确定收益下排队系统中的在线学习与自适应控制集成。运筹学70, 2 (2022), 1166–1181。[14] Shuyan Hu、Xiaojing Chen、Wei Ni、Xin Wang 和 Ekram Hossain。2020。能量采集和智能电网供电无线通信网络的建模与分析：当代综述。IEEE绿色通信与网络汇刊4, 2 (2020), 461–496。[15] Longbo Huang、Scott Moeller、Michael J Neely 和 Bhaskar Krishnamachari。2012。后进先出反压实现近最优效用-时延权衡。IEEE/ACM网络汇刊21, 3 (2012), 831–844。[16] Longbo Huang 和 Michael J Neely。2011。处理网络中的效用最优调度。性能评估68, 11 (2011), 1002–1021。[17] Md Rizwan Khan、Bikramaditya Das 和 Bibhuti Bhusan Pati。2020。水下声学（UWA）通信的信道估计策略：概述。《富兰克林研究所杂志》357, 11 (2020), 7229–7265。[18] Tae Hyun Kim、Hansol Shin、Kyuhyeong Kwag 和 Wook Kim。2020。采用分解跨期约束的微电网储能系统并行多周期最优调度算法。能源202 (2020), 117669。[19] Xiangqi Kong、Ning Lu 和 Bin Li。2019。具有流级动态的无人机网络最优调度。移动计算汇刊20, 3 (2021), 1186–1197。[20] Subhashini Krishnasamy、PT Akhil、Ari Arapostathis、Rajesh Sundaresan 和 Sanjay Shakkottai。2018。通过显式学习增强最大权重算法，以实现具有切换成本的无线调度。IEEE/ACM网络汇刊26, 6 (2018), 2501–2514。[21] Subhashini Krishnasamy、Rajat Sen、Ramesh Johari 和 Sanjay Shakkottai。2021。队列中未知服务率的学习：一种多臂老虎机方法。运筹学69, 1 (2021), 315–330。[22] Qingkai Liang 和 Eytan Modiano。2018。对抗性网络模型下的队列长度遗憾最小化。ACM测量与分析计算系统论文集2, 1 (2018), 1–32。[23] Shihuan Liu、Lei Ying 和 R Srikant。2011。流级动态下的吞吐量最优机会调度。IEEE/ACM网络汇刊19, 4 (2011), 1057–1070。[24] Lingling Lv、Chan Zheng、Lei Zhang、Chun Shan、Zhihong Tian、Xiaojiang Du 和 Mohsen Guizani。2021。基于合同和李雅普诺夫优化的无人机充电站负载调度与能源管理。IEEE绿色通信与网络汇刊5, 3 (2021), 1381–1394。[25] Siva Theja Maguluri、Rayadurgam Srikant 和 Lei Ying。2012。云计算集群中负载均衡与调度的随机模型。载于2012年IEEE Infocom会议论文集。IEEE，702–710。[26] Michael J Neely。2008。多用户无线链路上行和下行机会调度的阶最优时延。IEEE/ACM网络汇刊16, 5 (2008), 1188–1199。[27] Michael J Neely。2009。无线系统中最大权重机会调度的时延分析。IEEE自动控制汇刊54, 9 (2009), 2137–2150。[28] Michael J Neely。2010。随机网络优化及其在通信和排队系统中的应用。通信网络综合讲座3, 1 (2010), 1–211。[29] Michael J Neely。2012。基于时延的网络效用最大化。IEEE/ACM网络汇刊21, 1 (2012), 41–54。[30] Michael J Neely、Scott T Rager 和 Thomas F La Porta。2012。未知环境下调度的最大权重学习算法。IEEE自动控制汇刊57, 5 (2012), 1179–1191。[31] Konstantinos Psychas 和 Javad Ghaderi。2021。云服务资源预留的自动扩缩容理论。ACM SIGMETRICS性能评估评论48, 3 (2021), 27–32。[32] Mohammad Rahdar、Lizhi Wang 和 Guiping Hu。2018。不确定需求和交货期下库存控制的三级优化模型。《国际生产经济学杂志》195 (2018), 96–105。[33] Ralph Tyrell Rockafellar。2015。凸分析。普林斯顿大学出版社。[34] Bilal Sadiq 和 Gustavo De Veciana。2009。流动态无线系统中时延驱动的最大权重调度器的吞吐量最优性。载于2009年第47届Allerton通信、控制与计算年会（Allerton）。IEEE, 1097–1102。[35] Thomas Stahlbuhk、Brooke Shrader 和 Eytan Modiano。2018。信道统计信息未知的无线网络调度学习算法。载于第十八届ACM移动自组网国际研讨会论文集

g 和 *Com* *puting*。31–40。

[36] Leandros Tassiulas 和 Anthony Ephremides。1993。具有随机变化连接性的并行队列动态服务器分配。*IEEE* 信息论汇刊39, 2 (1993), 466–478。[37] Christos Tsanikidis 和 Javad Ghaderi。2021。无线网络中实时流量调度的随机化优势。*IEEE/ACM*网络汇刊29, 4 (2021), 1703–1716。[38] Vagelis Tsibonis、Leonidas Georgiadis 和 Leandros Tassiulas。2003。利用无线信道状态信息实现吞吐量最大化。载于 *IEEE INFOCOM 2003. IEEE* 计算机与通信学会第二十二届年度联合会议（*IEEE Cat. No. 03CH37428*），第 1 卷。IEEE，301–310。[39] Hua Wei、Chacha Chen、Guanjie Zheng、Kan Wu、Vikash Gayah、Kai Xu 和 Zhenhui Li。2019。Presslight：学习最大压力控制以协调干道网络中的交通信号灯。载于第 *25* 届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集。1290–1298。[40] Zixian Yang、R Srikant 和 Lei Ying。2022。采用折扣 UCB 的 MaxWeight：适用于统计特性未知的非平稳多服务器系统的可证明稳定调度策略。*arXiv* 预印本 *arXiv:2209.01126*（2022）。

补充材料

A 仿真结果 20 B EXP3.S+**（算法1）的证明 23 C** EXP3.S+ **（算法1）的实现 26 D 一般二次李雅普诺夫分析的证明 27 E** SoftMW **（算法 2）稳定性分析的证明 30 F** SSMW **（算法 3）稳定性分析的证明 34 G 处理具有有界矩的队列长度增量的高级思路 36 H 算法4的详细分析 40 I 算法5的详细分析 48**

A 仿真结果

本节基于合成问题实例，对所提出的算法 SoftMW（算法 2）和 SSMW（算法 3）的性能进行评估。在实验中，有 $K = 5$ 个队列，每个队列的到达过程和离开过程均服从伯努利分布。具体而言，我们将到达率和离开率设置如下：

• 每个队列的到达过程均为固定的伯努利分布，即 $\vec{\lambda}_t = \vec{\lambda}_{(0)}$ 对所有 $t \geq 1$，对于某个固定的$\vec{\lambda}_{(0)} \in [0,1]^K$。

• 我们将每个队列的伯努利服务率设为$\vec{\sigma}_t = \min\{\max\{\vec{\sigma}_{(0)} + \vec{\zeta}_t,\ \mathbf{0}\},\ \mathbf{1}\}$，其中 $\vec{\sigma}_{(0)} \in [0,1]^K$ 是一个固定向量，而 $\vec{\zeta}_t$ 由某个 K 维随机过程生成，满足$\mathbb{E}[\vec{\zeta}_t] = \mathbf{0}$，即它们的边缘均值为零。

因此，每个所考虑的问题实例都可以视为平稳伯努利到达过程与服务过程对，以及服务上的零均值噪声过程的叠加。

预加噪平稳问题

在实验中，我们考虑以下平稳问题实例：

$$\begin{aligned} \vec{\lambda} &= (0.25, 0.2, 0.15, 0.1, 0.05)\,, \\ \vec{\sigma} &= (0.9, 0.85, 0.8, 0.59, 0.39)\,. \end{aligned} \tag{15}$$

对于这个静态问题，我们求解以下线性规划问题

$$\begin{aligned} \max_{\vec{\theta} \in \mathbb{R}^K, \epsilon} \quad & \epsilon \\ \text{s.t.} \quad & \vec{\theta} \in \Delta^{[K]} \\ & \epsilon + \lambda_i \leq \theta_i \cdot \sigma_i \quad \forall i \in [K] \end{aligned} \tag{16}$$

$\vec{\theta}$以获得 θ，称为基于线性规划的随机策略（例如参见 [28]），以及 ϵ，即问题实例与容量区域边界之间的距离。方程 (16) 的解为

$$\begin{aligned} \vec{\theta} &= (0.27802, 0.23556, 0.18778, 0.16987, 0.12877)\,, \\ \epsilon &= 2.2207 \cdot 10^{-4}. \end{aligned}$$

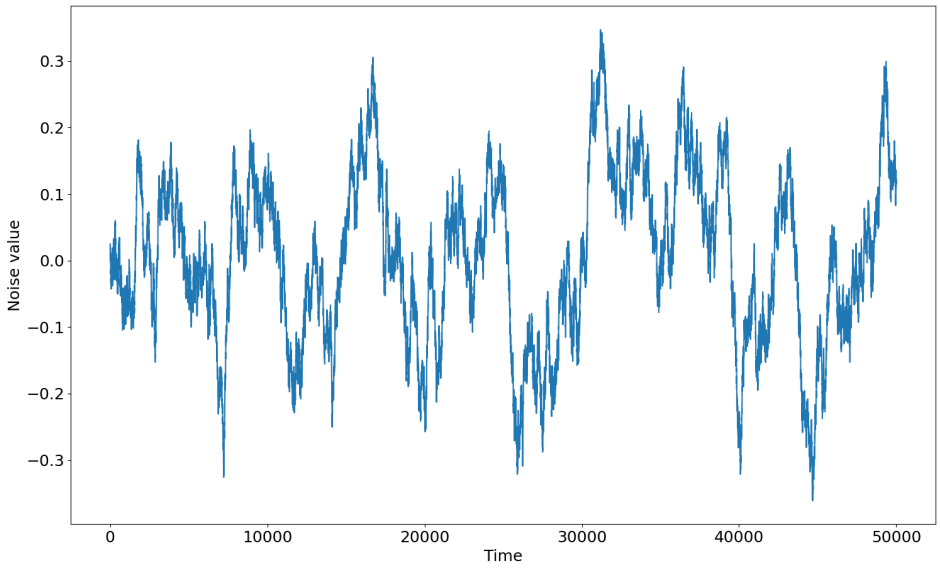


图3：队列1生成的服务噪声序列

噪声过程

我们使用以下　使用 $AR(1)$ 自回归过程生成要添加到平稳服务率中的噪声　s:

$$\zeta_{t,i} = \begin{cases} 0 & t = 0 \\ 0.999 \cdot \zeta_{t-1,i} + \mathcal{N}(0, 0.005^2) & t \geq 1 \end{cases}, \tag{17}$$

对于所有 $t \geq 1$ 和 $i \in [K]$。也就是说，时间 t 的噪声值是 $t-1$ 时的噪声乘以折扣因子 0.999，然后加上一个独立采样的均值为 0 的正态分布随机变量。

对于这种噪声机制，我们生成一条 $\{\vec{\zeta_t}$ 的轨迹： $0 \leq t \leq 2 \cdot 10^6\}$（即，生成由式（17）定义的一维 AR(1) 过程的 $K = 5$ 个独立同分布样本）。我们保存该轨迹，并在后续重复进行的$\vec{}$排队仿真中使用它。换言之，当我们重复模拟系统时，噪声向量 ζ_t将不会被重新采样，因此加噪后服务率 $\vec{\sigma}_t$在预先便固定，每次新的运行中仅重新采样伯努利试验结果。在图3中，我们绘制了为第一个队列生成的服务噪声的前 50000 项。

评估的算法

在实验中，我们评估以下算法实例:

- （基线）MaxWeight算法，以加噪前服务率向量 $\vec{\sigma}(0)$ 作为输入参数。它根据$a_t = \text{argmax}_{i \in [K]} \left\{ Q_{t-1,i} \cdot \sigma_{i}^{(0} \right\})$ 对系统进行调度;
- （有噪声时的预言机基线）MaxWeight算法以所有加噪后的服务率向量$\{\vec{\sigma}_t : \ 1 \leq t \leq T\}$ 作为输入参数，称为MaxWeightGT。它按照$a_t = \text{argmax}_{i \in [K]} \{Q_{t-1,i} \cdot \sigma_{t,i}\}$来调度系统;
- （基线）基于线性规划的随机策略（称为Randomized）。该策略在每个时间步 t独立地根据概率向量 θ 对 $a_t\vec{}$进行采样，以此调度系统。
- （我们的方法）采用参数 $M = 1, \delta = 0.5$ 的SoftMW（算法2）（称为SoftMW-0.5）；
- （我们的方法）采用参数 $M = 1, \delta = 0.1$ 的SoftMW（算法2）（称为SoftMW-0.1）；
- （我们的方法）采用参数 $M = 1, \delta = 0$的SoftMW（算法2）（称为SoftMW-0）；
- （我们的方法）SSMW（算法 3），参数为 $M = 1, \delta = 0.5$ （称为 SSMW-0.5）；
- （我们的方法）SSMW（算法 3），参数为 $M = 1, \delta = 0.1$ （称为 SSMW-0.1）；
- （我们的方法）SSMW（算法 3），参数为 $M = 1$、 $\delta = 0$（称为 SSMW-0）。

注。 我们的主要理论结果——定理 5.2 和 6.1——无法为使用 $\delta = 0$ 运行的两种算法提供队列稳定性保证。不过，在 $\delta = 0$ 条件下，这两种算法都是合理且可行的排队策略。因此，我们也将它们纳入待评估的算法实例列表，以展示当 δ 足够小时，SoftMW 和 SSMW 的极限行为。

在实现 SSMW 时，我们将上一个 EXP3.S+ 时期最后一个时间步的$\mathbf{x}_\tau$ 值，作为新 EXP3.S+ 实例中的初始混合动作$\mathbf{x}_1$ 。

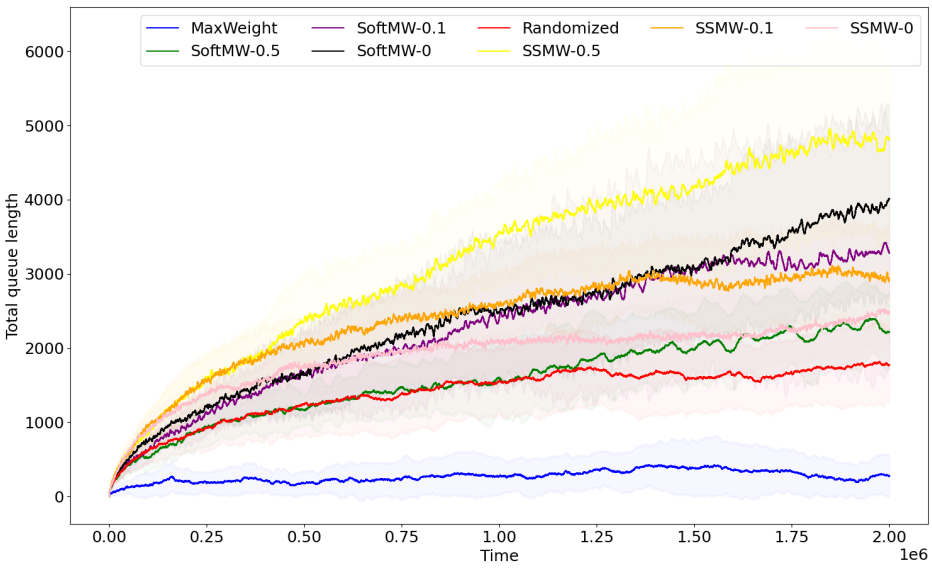
仿真结果

我们基于公式（15）中规定的静态问题，对两个问题实例进行数值评估。$\vec{}$ $\vec{}$在一个实例中，我们将所有服务噪声向量 ζ_ts 设置为$\mathbf{0}$；在另一个实例中，我们根据公式（17）设置 ζ_t 。在这两个问题实例中，我们均仿真 $T = 2 \cdot 10^6$ 个时间步。

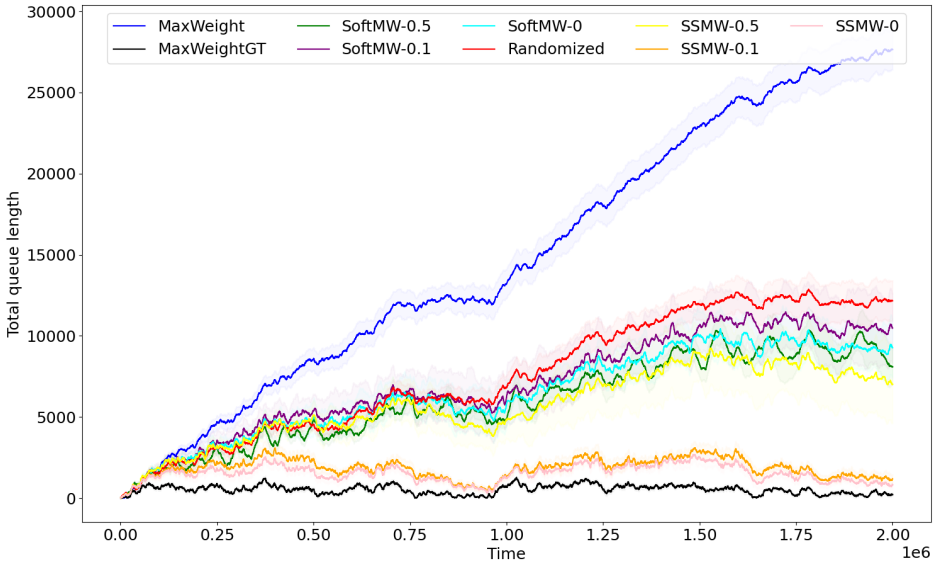
在图 4 中，我们绘制了两个问题实例的仿真结果，其中分别评估了 8 个和 9 个算法实例（在无噪声问题中，MaxWeightGT 和 MaxWeight 完全相同）；每个算法的曲线均取 20 次独立仿真的平均值。下面，我们简要总结仿真结果。

• 在无噪声的问题实例中（图 4a），MaxWeight 使平均队列长度非常短，优于其他受评估的算法。这并不令人意外，因为已知在获得精确的信道条件时，MaxWeight 的表现很好。值得注意的是，尽管我们的算法没有这类信息，也能稳定网络。

• 在含噪的问题实例中（图 4b），MaxWeight 的性能显著下降，系统似乎变得不稳定。而预言机基线 MaxWeightGT 仍能保持相当小的队列长度。我们的 SoftMW 和 SSMW 算法的表现都优于 Rando mized。特别是，SSMW-0.1 和 SSMW-0 产生的平均队列长度接近 MaxWeightGT。



(a) 无噪声的问题。此时 MaxWeight 可以获取准确的信道信息。



(b) A 的问题 R(1) 噪声。此处 MaxWeightGT 可以获取准确的信道信息，而 MaxWeight 只能获取 s
不完美的信道信息 ，且这些信息不计入噪声。

图4：队列总长度图

B EXP3.S+（算法1）的证明

除非另有说明，对于任意严格凸函数 f，我们使用 f 表示其在某个 $\triangle^{[K]}$上的限制，$^{\beta}$，其中β 应根据上下文推断。具体而言，

$$\overline{f}(\mathbf{x}) \triangleq \begin{cases} f(\mathbf{x}), & \mathbf{x} \in \triangle^{[K],\beta} \\ \infty, & \mathbf{x} \notin \triangle^{[K],\beta}. \end{cases}$$

下面列出勒让德函数和 Bregman 散度的一些重要性质。相关证明可在许多凸分析文献中找到，例如 [33]。

引理 B.1。 设 $\mathcal{C} \subseteq \mathbb{R}^n$ 为凸集， $f:\ \mathcal{C} \to \mathbb{R}$ 为勒让德函数。则，

1. ∇f 是 $int(\mathcal{C})$ 与 $int(dom(f^*))$ 之间的双射，其逆为 $(\nabla f)^{-1} = \nabla f^*$。

2. $D_f(\mathbf{y},\mathbf{x}) = D_{f^*}(\nabla f(\mathbf{x}), \nabla f(\mathbf{y}))$ *for all* $\mathbf{x},\mathbf{y} \in int(\mathcal{C})$.

3. 凸共轭 f^* 是勒让德函数。

4.（广义毕达哥拉斯定理）设 W 是 $\mathcal{C}$ 的闭凸子集，对于任意 $x \in \mathcal{C}$，记 $\Pi_f(x\ W)$ 为 $\triangleq arg\ min_{x' \in W} D_f(x', x)$。极小值点保证存在且唯一，并且对于任意 $y \in W$，有 $D_f(y, x) \geq D_f(y\ \Pi_f(x\ W)) + D_f(\Pi_f(x\ W), x)$。

下面给出一个将单步遗憾与单步 OMD 更新联系起来的技术引理。

引理 B.2. 对于任意 $\beta \in [0, 1/K]$, $\eta > 0$, $\mathbf{x}, \mathbf{y} \in \triangle^{[K],\beta}$, $\mathbf{g} \in \mathbb{R}_+^K$ 以及勒让德函数 $\Psi:\ \mathbb{R}_+^K \to \mathbb{R}$，有

$$\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq \eta^{-1} D_\Psi(\mathbf{y},\mathbf{x}) - \eta^{-1} D_\Psi(\mathbf{y},\mathbf{z}) + \eta^{-1} D_\Psi(\mathbf{x},\widetilde{\mathbf{z}}) \tag{18}$$

其中

$$\mathbf{z} = \operatorname*{arg\,min}_{\mathbf{x}' \in \triangle^{[K],\beta}} \langle -\eta\mathbf{g}, \mathbf{x}' \rangle + D_\Psi(\mathbf{x}',\mathbf{x}), \quad \widetilde{\mathbf{z}} = \operatorname*{arg\,min}_{x' \in \mathbb{R}_+^K} \langle -\eta\mathbf{g}, x' \rangle + D_\Psi(\mathbf{x}',\mathbf{x}), \tag{19}$$

或者等价地，

$$\widetilde{\mathbf{z}} = \nabla\Psi^*(\nabla\Psi(x) + \eta\mathbf{g}) \tag{20}$$

证明。 我们首先证明公式 (19) 与公式 (20) 等价。 Ψ 是勒让德，因此 $\nabla\Psi$ 在 $\partial\mathbb{R}_+^K$上趋于无穷，而由公式 (19) 定义的极小值点 $\widetilde{\ }\mathbf{z}$ 将位于 $int(\mathbb{R}_+^K)$ 中；此外，我们必有 $\frac{\partial}{\partial \widetilde{\ }\mathbf{z}'}[\langle -\eta\mathbf{g}\ \widetilde{\ }\mathbf{z}\rangle + D_\Psi(\widetilde{\ }\mathbf{z},\mathbf{x})] =$

$-\eta\mathbf{g} + \nabla\Psi(\tilde{\mathbf{z}}) - \nabla\Psi(\mathbf{x}) = \mathbf{0}$，因此 $\nabla\Psi(\tilde{\mathbf{z}}) = \nabla\Psi(\mathbf{x}) + \eta\mathbf{g}$。接着，定理 B.1 中的双射性质表明，$\tilde{\mathbf{z}} = \nabla\Psi^*(\nabla\Psi(\tilde{\mathbf{z}})) = \nabla\Psi^*(\nabla\Psi(x) + \eta\mathbf{g})$，这正是公式 (20)。

由公式 (20) 可知

$$\mathbf{g} = \eta^{-1}(\nabla\Psi(\widetilde{\mathbf{z}}) - \nabla\Psi(\mathbf{x})) \tag{21}$$

公式 (19) 中 $\widetilde{z}$ 的一阶最优性条件意味着 $\langle -\eta\mathbf{g} + \nabla\Psi(\mathbf{z}) - \nabla\Psi(\mathbf{x}), \mathbf{y} - \mathbf{z}\rangle \geq 0$ 对任意 $y \in \triangle^{[K],\beta}$，因此

$$\langle \mathbf{g}, \mathbf{y} - \mathbf{z}\rangle \leq \eta^{-1}\langle \nabla\Psi(\mathbf{z}) - \nabla\Psi(\mathbf{x}), \mathbf{y} - \mathbf{z}\rangle \tag{22}$$

因此，我们可以写成

$$\begin{aligned}
\langle \mathbf{g}, \mathbf{y} - \mathbf{x}\rangle &= \langle \mathbf{g}, \mathbf{y} - \mathbf{z}\rangle + \langle \mathbf{g}, \mathbf{z} - \mathbf{x}\rangle \\
&\overset{(a)}{\leq} \eta^{-1}\langle \nabla\Psi(\mathbf{z}) - \nabla\Psi(\mathbf{x}), \mathbf{y} - \mathbf{z}\rangle + \eta^{-1}\langle \nabla\Psi(\widetilde{\mathbf{z}}) - \nabla\Psi(\mathbf{x}), \mathbf{z} - \mathbf{x}\rangle \\
&\overset{(b)}{=} \eta^{-1}(D_\Psi(\mathbf{y}, \mathbf{z}) + D_\Psi(\mathbf{z}, \mathbf{x}) - D_\Psi(\mathbf{y}, \mathbf{x})) + \eta^{-1}(D_\Psi(\mathbf{z}, \mathbf{x}) + D_\Psi(\mathbf{x}, \widetilde{\mathbf{z}}) - D_\Psi(\mathbf{z}, \widetilde{\mathbf{z}})) \\
&= \eta^{-1}(D_\Psi(\mathbf{y}, \mathbf{x}) - D_\Psi(\mathbf{y}, \mathbf{z}) + D_\Psi(\mathbf{x}, \widetilde{\mathbf{z}}) - D_\Psi(\mathbf{z}, \widetilde{\mathbf{z}})) \\
&\leq \eta^{-1}(D_\Psi(\mathbf{y}, \mathbf{x}) - D_\Psi(\mathbf{y}, \mathbf{z}) + D_\Psi(\mathbf{x}, \widetilde{\mathbf{z}})).
\end{aligned}$$

在这里，步骤（a）中，我们在第一项代入公式 (22)，在第二项代入公式 (21)；步骤（b）中，我们使用以下布雷格曼散度的“三点恒等式”：

$$D_\Psi(\mathbf{a}, \mathbf{b}) + D_\Psi(\mathbf{b}, \mathbf{c}) - D_\Psi(\mathbf{a}, \mathbf{c}) = \langle \nabla\Psi(\mathbf{c}) - \nabla\Psi(\mathbf{b}), \mathbf{a} - \mathbf{b}\rangle,$$

通过展开所有 Bregman 散度即可验证。 □

直观而言，定理 B.2 给出了混合动作*sample-path* 样本路径单步遗憾的 Bregman 散度上界，其中混合动作 $\mathbf{x}$ 相对于基准混合动作$\mathbf{y}$，奖励向量为$\mathbf{g}$。我们证明定理 5.1 的总体思路是对定理 B.2 中的上界求和，然后适当地取期望并对期望进行控制。首先，我们可以得到总动态遗憾的样本路径界。

引理 B.3。 对于任意固定序列 $\vec{\theta}_1 \in \Delta^{[K],\beta_1}, \ldots, \vec{\theta}_T \in \Delta^{[K],\beta_T}$，算法*1*保证

$$\sum_{t=1}^{T}\langle \widetilde{\mathbf{g}}_t, \vec{\theta}_t - \mathbf{x}_t\rangle \leq \left(1 + \sum_{t=1}^{T-1}\|\vec{\theta}_{t+1} - \vec{\theta}_t\|_1\right)\eta_T^{-1}\ln\frac{1}{\beta_T} - \eta_T^{-1}D_\Psi(\vec{\theta}_T, \mathbf{z}_T) + \sum_{t=1}^{T}\eta_t^{-1}D_\Psi(\mathbf{x}_t, \widetilde{\mathbf{z}}_t) \tag{23}$$

其中 $\Psi(\mathbf{x}) \triangleq \sum_{i=1}^{K}(x_i\ln x_i - x_i)$, $\tilde{\mathbf{z}}_t \triangleq \nabla\Psi^*(\nabla\Psi(\mathbf{x}_t) + \eta_t\tilde{\mathbf{g}}_t)$, $\mathbf{z}_t \triangleq \arg\min_{\mathbf{x}'\in\triangle^{[K],\beta_t}}\langle -\eta_t\tilde{\mathbf{g}}_t, \mathbf{x}'\rangle + D_\Psi(\mathbf{x}', \mathbf{x}_t)$。

证明。 我们将对 T 进行归纳，以证明式 (23)。当 $T = 0$ 时，式 (23) 显然成立（这里我们采用约定 $\eta_0^{-1} = 0$）。因此，只需在假设.....的情况下验证定理 B.3，

$$\sum_{t=1}^{T-1}\langle \widetilde{\mathbf{g}}_t, \vec{\theta}_t - \mathbf{x}_t\rangle \leq \left(1 + \sum_{t=1}^{T-2}\|\vec{\theta}_{t+1} - \vec{\theta}_t\|_1\right)\eta_{T-1}^{-1}\ln\frac{1}{\beta_{T-1}} - \eta_{T-1}^{-1}D_\Psi(\vec{\theta}_{T-1}, \mathbf{z}_{T-1}) + \sum_{t=1}^{T-1}\eta_t^{-1}D_\Psi(\mathbf{x}_t, \widetilde{\mathbf{z}}_t).$$

在算法1中， Ψ 函数选为负熵，因此我们有

$$D_\Psi(\mathbf{y}, \mathbf{x}) = \sum_{t=1}^{K} y_i\ln\frac{y_i}{x_i}. \tag{24}$$

因此，对于任意 $\mathbf{y} \in \Delta^{[K]}$，$\mathbf{x} \in \Delta^{[K],\beta}$，由于对所有 $i \in [K]$ 都有 $x_i \geq \beta$，我们有

$$D_\Psi(\mathbf{y}, \mathbf{x}) \leq \sum_{i=1}^{K} y_i\ln\frac{1}{x_i} \leq \ln\frac{1}{\beta}. \tag{25}$$

注意，在算法1中选择$\mathbf{x}_t$ 只是$\mathbf{z}_{t-1}$，因此我们可以写成

$$\begin{aligned}
&\eta_T^{-1}D_\Psi(\vec{\theta}_T, \mathbf{x}_T) \\
&= \eta_T^{-1}D_\Psi(\vec{\theta}_T, \mathbf{z}_{T-1})
\end{aligned}$$

$$
\begin{aligned}
&= \eta_{T-1}^{-1} D_\Psi(\vec{\theta}_{T-1}, \mathbf{z}_{T-1}) + \eta_{T-1}^{-1}\left(D_\Psi(\vec{\theta}_T, \mathbf{z}_{T-1}) - D_\Psi(\vec{\theta}_{T-1}, \mathbf{z}_{T-1})\right) + (\eta_T^{-1} - \eta_{T-1}^{-1}) D_\Psi(\vec{\theta}_T, \mathbf{z}_{T-1}) \\
&\overset{(a)}{\leq} \eta_{T-1}^{-1} D_\Psi(\vec{\theta}_{T-1}, \mathbf{z}_{T-1}) + \eta_{T-1}^{-1} \sum_{t=1}^{K}\left(\theta_{T,i} \ln \frac{\theta_{T,i}}{z_{T-1,i}} - \theta_{T-1,i} \ln \frac{\theta_{T-1,i}}{z_{T-1,i}}\right) + (\eta_T^{-1} - \eta_{T-1}^{-1}) \ln \frac{1}{\beta_{T-1}} \\
&\overset{(b)}{\leq} \eta_{T-1}^{-1} D_\Psi(\vec{\theta}_{T-1}, \mathbf{z}_{T-1}) + \eta_{T-1}^{-1} \sum_{i=1}^{K}\left((1 - z_{T-1,i}) \ln \frac{1}{z_{T-1,i}} |\theta_{T,i} - \theta_{T-1,i}|\right) + (\eta_T^{-1} - \eta_{T-1}^{-1}) \ln \frac{1}{\beta} \\
&\leq \eta_{T-1}^{-1} D_\Psi(\vec{\theta}_{T-1}, \mathbf{z}_{T-1}) + \eta_{T-1}^{-1} \|\vec{\theta}_T - \vec{\theta}_{T-1}\|_1 \ln \frac{1}{\beta_{T-1}} + (\eta_T^{-1} - \eta_{T-1}^{-1}) \ln \frac{1}{\beta_{T-1}},
\end{aligned}
$$

其中，在步骤（a）中，我们对第二项代入公式（24），对第三项代入公式（25）；在步骤（b）中，我们使用以下事实：令 $f(x) = x \ln \frac{x}{y}$，则 $f'(x) = (1-y) \ln \frac{x}{y} \leq (1-y) \ln \frac{1}{y}$，只要 $x,\ y \in (0,\ 1)$。
然后，我们可以在第 T 个时间步应用定理 B.2，并写作

$$
\begin{aligned}
\langle \widetilde{\mathbf{g}}_T, \vec{\theta}_T - \mathbf{x}_T \rangle &\leq \eta_T^{-1} D_\Psi(\vec{\theta}_T, \mathbf{x}_T) - \eta_T^{-1} D_\Psi(\vec{\theta}_T, \mathbf{z}_T) + \eta_T^{-1} D_\Psi(\mathbf{x}_T, \widetilde{\mathbf{z}}_T) \\
&\leq \eta_{T-1}^{-1} D_\Psi(\vec{\theta}_{T-1}, \mathbf{z}_{T-1}) - \eta_T^{-1} D_\Psi(\vec{\theta}_T, \mathbf{z}_T) + \eta_T^{-1} D_\Psi(\mathbf{x}_T, \widetilde{\mathbf{z}}_T) \\
&\quad + \eta_{T-1}^{-1} \|\vec{\theta}_T - \vec{\theta}_{T-1}\|_1 \ln \frac{1}{\beta_{T-1}} + (\eta_T^{-1} - \eta_{T-1}^{-1}) \ln \frac{1}{\beta_{T-1}},
\end{aligned}
$$

回想一下，η_t 和 β_t 都是递减的，这足以完成归纳步骤。 □

引理 B.4. 对于定理 *B.3* 中定义的每一项 $\eta_t^{-1} D_\Psi(\mathbf{x}_t, \widetilde{\mathbf{z}}_t)$，我们有

$$
\mathbb{E}\left[\eta_t^{-1} D_\Psi(\mathbf{x}_t, \widetilde{\mathbf{z}}_t) \middle| \mathcal{F}_{t-1}\right] \leq e \eta_t \|\mathbf{g}_t\|_2^2.
$$

证明。事实上，对于任意选择的 Ψ，我们有

$$
\begin{aligned}
\eta_t^{-1} D_\Psi(x_t, \widetilde{z}_t) &\overset{(a)}{=} \eta_t^{-1} D_{\Psi^*}(\nabla\Psi(\widetilde{\mathbf{z}}_t), \nabla\Psi(\mathbf{x}_t)) \\
&= \eta_t^{-1} D_{\Psi^*}(\nabla\Psi(\mathbf{x}_t) + \eta_t \widetilde{\mathbf{g}}_t, \nabla\Psi(\mathbf{x}_t)) \\
&= \eta_t^{-1} \left(\Psi^*(\nabla\Psi(\mathbf{x}_t) + \eta_t \widetilde{\mathbf{g}}_t) - \Psi^*(\nabla\Psi(\mathbf{x}_t)) - \langle \mathbf{x}_t, \eta_t \widetilde{\mathbf{g}}_t \rangle\right) \\
&\overset{(b)}{=} \frac{\eta_t}{2} \|\widetilde{\mathbf{g}}_t\|^2_{\nabla^2\Psi^*(\mathbf{w}_t)},
\end{aligned} \tag{26}
$$

其中，在步骤 (a) 中我们使用定理 B.1 的对偶性质；在步骤 (b) 中，我们将 Bregman 散度视为二阶拉格朗日余项，$\mathbf{w}_t$ 是连接 $\nabla\Psi(\mathbf{x}_t) + \eta_t \widetilde{\mathbf{g}}_t$ 和 $\nabla\Psi(\mathbf{x}_t)$ 的线段上的某个元素。即，$w_{t,i} = \nabla\Psi(\mathbf{x}_t)_i$ 对所有 $i \neq a_t$，$w_{t,a_t} \in [\nabla\Psi(\mathbf{x}_t)_{a_t}, \nabla\Psi(\mathbf{x}_t)_{a_t} + \eta_t \frac{g_{t,a_t}}{p_{t,a_t}}]$ 均成立。由于我们假设对所有 i，$\eta_t^{-1}\gamma_t e_{t,i} \geq g_{t,i}$ 均成立，并且使用显式的探索机制来保证 $p_{t,i} = (1-\gamma_t) x_{t,i} + \gamma_t e_{t,i} \geq \gamma_t e_{t,i}$，因此现在有 $\eta_t \frac{g_{t,a_t}}{p_{t,a_t}} \leq 1$。

现在，对于将 Ψ 选为负熵函数这一特定情况，我们有 $\nabla\Psi(\mathbf{x}) = (\ln x_1, \ldots, \ln x_K)$，且 $\nabla^2\Psi(\mathbf{x}) = \mathrm{Diag}(x_1^{-1}, \ldots, x_K^{-1})$。利用定理 B.1 中性质 (1)，我们可以看到

$$
\begin{aligned}
\nabla^2\Psi^*(\mathbf{w}_t) &= \left(\nabla^2\Psi(\nabla\Psi^*(\mathbf{w}_t))\right)^{-1} \\
&= \mathrm{Diag}(\exp(w_{t,1})^{-1}, \ldots, \exp(w_{t,K})^{-1})^{-1} \\
&= \mathrm{Diag}(\exp(w_{t,1}), \ldots, \exp(w_{t,K})),
\end{aligned}
$$

因此，我们有

$$
\mathrm{Diag}(x_{t,1}, \ldots, x_{t,K}) \preceq \nabla^2\Psi^*(\mathbf{w}_t) \preceq e\,\mathrm{Diag}(x_{t,1}, \ldots, x_{t,K}) \tag{27}
$$

然后，我们可以将式（27）代入式（26），得到

$$
\begin{aligned}
\eta_t^{-1} D_\Psi(x_t, \widetilde{z}_t) &\leq \frac{e\eta_t}{2} \widetilde{g}_{t,a_t}^2 x_{t,a_t} \\
&= \frac{e\eta_t}{2} \frac{g_{t,a_t}^2}{p_{t,a_t}^2} x_{t,a_t}.
\end{aligned}
$$

取期望后可得

$$\begin{aligned}\mathbb{E}\left[\eta_t^{-1} D_\Psi(x_t, \widetilde{z}_t)\middle|\mathcal{F}_{t-1}\right] &\le \frac{e\eta_t}{2}\sum_{i=1}^{K} g_{t,i}^2 \frac{x_{t,i}}{p_{t,i}} \\ &\le e\eta_t \sum_{i=1}^{K} g_{t,i}^2\end{aligned}$$

其中最后一步由 $\eta_t \le \frac{1}{2}$ 得出，因此 $p_{t,\ i} \ge (1-\gamma_i) x_{t,\ i} \ge \frac{1}{2} x_{t,\ i}$。 □

现在我们可以证明定理 5.1。

定理 *5.1* 的证明。 对定理 B.3 中公式 (23) 两边取期望，注意到

$$\mathbb{E}\left[\langle \widetilde{\mathbf{g}}_t, \vec{\theta}_t - \mathbf{x}_t\rangle \middle| \mathcal{F}_{t-1}\right] = \langle \mathbf{g}_t, \vec{\theta}_t - \mathbf{x}_t\rangle,$$

and

$$\begin{aligned}\mathbb{E}\left[g_{t,a_t} | \mathcal{F}_{t-1}\right] &= \langle \mathbf{g}_t, \mathbf{p}_t\rangle \\ &= \langle \mathbf{g}_t, \mathbf{x}_t\rangle + \langle \mathbf{g}_t, \mathbf{p}_t - x_t\rangle \\ &= \langle \mathbf{g}_t, \mathbf{x}_t\rangle + \langle \mathbf{g}_t, -\gamma_t \mathbf{x}_t + \gamma_t \mathbf{e}_t\rangle \\ &\le \langle \mathbf{g}_t, \mathbf{x}_t\rangle + \gamma_t \langle \mathbf{g}_t, \mathbf{e}_t\rangle.\end{aligned}$$

然后应用定理 B.4。 □

C 关于 EXP3.S+ （算法1）的实现

实现我们的 EXP3.S+ （算法1）时，核心操作是第 8 行的 argmin 表达式。换言之，我们需要求解以下优化问题：

$$\begin{aligned}\min_{\mathbf{y}\in\mathbb{R}^K} \quad & -\langle \mathbf{g}, \mathbf{y}\rangle + D_\Psi(\mathbf{y}, \mathbf{x}) \\ \text{s.t.} \quad & y_i \ge \beta \quad \forall 1 \le i \le K \\ & \sum_{i=1}^{K} y_i = 1\end{aligned} \tag{28}$$

其中 $\mathbf{g} \in \mathcal{R}_+^K$ 和 $x \in \Delta^{[K]}$ 是两个固定向量，$0 \le \beta < \frac{1}{K}$ 是一个常数。对于这个优化问题，我们可以引入一组拉格朗日乘子 $\lambda_1, \ldots, \lambda_K, \mu$ 以及拉格朗日函数

$$L(\mathbf{y}; \vec{\lambda}, \mu) \triangleq -\sum_{i=1}^{K} g_i y_i + D_\Psi(\mathbf{y}, \mathbf{x}) + \sum_{i=1}^{K} \lambda_i (y_i - \beta) + \mu\left(\sum_{i=1}^{K} y_i - 1\right).$$

令 $\mathbf{y}^* \in \Delta^{[K],\beta}$ 为优化问题式 (28) 的最优解，并令 $\vec{\lambda}^*, \mu^*$ 为相应的乘子。于是我们可以写出$\mathbf{y}^*$的 KKT 条件：

$$\begin{aligned}\left.\frac{\partial L}{\partial y_i}\right|_{\mathbf{y}^*;\vec{\lambda}^*,\mu^*} &= -g_i + \ln(y_i^*) - \ln(x_i) + \lambda_i^* + \mu^* = 0 \quad \forall 1 \le i \le K, \\ & y_i^* > \beta \Rightarrow \lambda_i^* = 0 \quad \forall 1 \le i \le K.\end{aligned}$$

记 $\mathcal{I} \triangleq \{i \in [K] : y_i^* > \beta\}$，则 KKT 条件意味着

$$\ln(y_i^*) - \ln(x_i) - g_i = \ln(y_j^*) - \ln(x_j) - g_j \tag{29}$$

对于所有 $i,\ j \in \mathcal{I}$。另一方面，我们知道

$$\sum_{i\in\mathcal{I}} y_i^* = 1 - \beta \cdot (K - |\mathcal{I}|) \tag{30}$$

因为我们假设对于所有 $y_i^* = \beta\ i \in [K] \setminus \mathcal{I}$。事实上，公式（29）和（30）具有显式的"软最大值"解

$$y_i^* \propto x_i \exp(g_i)$$

因此

$$y_i^* = [1 - \beta \cdot (K - |\mathcal{I}|)] \cdot \frac{x_i \exp(g_i)}{\sum_{j \in \mathcal{I}} x_j \exp(g_j)}$$

对所有 $i \in \mathcal{I}$均成立。因此，为了完整描述 y^*，还需要确定 $\mathcal{I}$。注意，优化问题 Eq. (28) 满足斯莱特条件，因此其最优解完全由 KKT 条件确定。可以验证，以下问题

$$\begin{aligned} \min_{\mathbf{y} \in \mathbb{R}^K} \quad & D_\Psi(\mathbf{y}, \mathbf{x}') \\ \text{s.t.} \quad & y_i \geq \beta \quad \forall 1 \leq i \leq K \\ & \sum_{i=1}^K y_i = 1 \end{aligned} \tag{31}$$

其中 $x_i' = x_i \exp(g_i)$，与 Eq. (28) 具有完全相同的 KKT 条件。此外，不失一般性，我们假设 Eq. (31) 中 $x_1' \leq x_2' \leq \cdots \leq x_K'$（否则，由于 $D_\Psi(\cdot,\ \cdot)$ 和约束对所有坐标均对称，我们可以对 $\mathbf{x}$ 的坐标进行置换和重新排列）。那么，我们可以断言 $y_1^* \leq y_2^* \leq \cdots \leq y_K^*$，也就是说，$y^*$ 的坐标也按顺序排列。原因是，我们可以将 Eq. (31) 的目标函数展开为：

$$\begin{aligned} OBJ &= \Psi(\mathbf{y}) - \Psi(\mathbf{x}') - \langle \nabla \Psi(\mathbf{x}'), \mathbf{y} - \mathbf{x}' \rangle \\ &= \sum_{i=1}^K y_i \ln(y_i) - \sum_{i=1}^K x_i' \ln(x_i') - \sum_{i=1}^K \ln(x_i')(y_i - x_i') - 1 + \|\mathbf{x}'\|_1 \\ &= \sum_{i=1}^K y_i \cdot (-\ln(x_i')) + \text{some permutation-invariant quantity of } \mathbf{y} \end{aligned}$$

接下来可以看到，如果$\mathbf{y}$ 有一对坐标 $i < j$ 但 $y_i > y_j$，那么交换 y_i 和 y_j 会使目标函数更小。因此，将优化问题化简为式 (31)，并使 x_i'按升序排列后，存在$0 \leq i \leq K$ 使得 $\beta = y_1^* = \cdots = y_i^* = \beta < y_{i+1}^* \leq y_{i+2}^* \leq \cdots \leq y_K^*$。要解式 (31)，只需枚举这个边界索引 $i = 0 \ldots K$，然后计算 $y_{i+1}^*, \ldots, y_K^*$，方法如下：

$$y_j^* = (1 - \beta \cdot i) \cdot \frac{x_j'}{\sum_{k=i+1}^K x_k'}.$$

如果 $y_{i+1}^* \geq \beta$，我们可以接受当前的 i 作为最优边界索引，并得出 $(\beta, \ldots, \beta, y_{i+1}^*, \ldots, y_K^*)$ 是式 (31) 的最优解。

D 一般二次李雅普诺夫分析的证明

定理 *5.3* 的证明。回想一下，在本文中，我们用 $\mathbf{Q}_t$ 表示第 t 个时隙结束时的队列长度，用 $\mathbf{A}_t$、$\mathbf{S}_t$ 分别表示时间 t的到达量和（最大）服务量。定义

$$\mathbf{U}_t = \max\{-\mathbf{Q}_{t-1} - \mathbf{A}_t + \mathbf{S}_t \odot \mathbf{1}_{a_t}, \mathbf{0}\}$$

表示时间 t未使用的服务量，则显然有

$$\mathbf{Q}_t = \mathbf{Q}_{t-1} + \mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t} + \mathbf{U}_t \tag{32}$$

and

$$\langle \mathbf{Q}_t, \mathbf{U}_t \rangle = 0, \tag{33}$$

$$\langle \mathbf{Q}_{t-1} + \mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t}, \mathbf{U}_t \rangle \leq 0 \tag{34}$$

对所有 $t \geq 1$均成立。

利用式 (32) 表示 L_t，可以看出

$$L_t - L_{t-1} = \frac{1}{2}\|\mathbf{Q}_{t-1} + \mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t} + \mathbf{U}_t\|_2^2 - \frac{1}{2}\|\mathbf{Q}_{t-1}\|_2^2$$

$$
\begin{aligned}
&= \frac{1}{2}\|\mathbf{Q}_{t-1} + \mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t}\|_2^2 + \frac{1}{2}\|\mathbf{U}_t\|_2^2 + \langle \mathbf{Q}_{t-1} + \mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t}, \mathbf{U}_t\rangle - \frac{1}{2}\|\mathbf{Q}_{t-1}\|_2^2 \\
&\overset{(a)}{\leq} \frac{1}{2}\|\mathbf{Q}_{t-1} + \mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t}\|_2^2 + \frac{1}{2}\|\mathbf{U}_t\|_2^2 - \frac{1}{2}\|\mathbf{Q}_{t-1}\|_2^2 \\
&= \frac{1}{2}\|\mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t}\|_2^2 + \frac{1}{2}\|\mathbf{U}_t\|_2^2 + \langle \mathbf{Q}_{t-1}, \mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t}\rangle \\
&\overset{(b)}{\leq} \frac{(K+1)M^2}{2} + \langle \mathbf{Q}_{t-1}, \mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t}\rangle,
\end{aligned}
$$

其中，步骤 (a) 由式 (34) 得出，(b) 由 $\|\mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t}\|_\infty$、$\|\mathbf{U}_t\|_\infty \leq M$ 得出，而 $\mathbf{U}_t$ 至多有一个正值项。因此

$$
\begin{aligned}
\mathbb{E}\left[L_t - L_{t-1} \middle| \mathcal{F}_{t-1}\right] &\leq \frac{(K+1)M^2}{2} + \mathbb{E}\left[\langle \mathbf{Q}_{t-1}, \mathbf{A}_t - \mathbf{S}_t \odot \mathbf{1}_{a_t}\rangle \middle| \mathcal{F}_{t-1}\right] \\
&\leq \frac{(K+1)M^2}{2} + \langle \mathbf{Q}_{t-1}, \vec{\lambda}_t - \vec{\sigma}_t \odot \mathbf{p}_t\rangle,
\end{aligned}
$$

其中，最后一步是因为$\mathbf{Q}_{t-1}$ 是 $\mathcal{F}_{t-1}$-可测的，并且$\mathbf{A}_t$，$\mathbf{S}_t$ 均与 a_t 和 $\mathcal{F}_{t-1}$ 独立，$\mathbb{P}[a_t = i|\mathcal{F}_{t-1}] = p_{t,\ i}$。 □

定理 *5.4* 的证明。对于假设 1 中每个包含 W_j 的时间步区间，设 T_0 为 W_j 中第一个时隙的索引，则我们可以写出

$$
\sum_{t\in W_j} \langle \mathbf{Q}_{t-1}, \vec{\theta}_t \odot \vec{\sigma}_t - \vec{\lambda}_t\rangle = \sum_{t\in W_j} \langle \mathbf{Q}_{T_0-1}, \vec{\theta}_t \odot \vec{\sigma}_t - \vec{\lambda}_t\rangle + \sum_{t\in W_j} \langle \mathbf{Q}_{t-1} - \mathbf{Q}_{T_0-1}, \vec{\theta}_t \odot \vec{\sigma}_t - \vec{\lambda}_t\rangle,
$$

接着逐一对这两个求和项进行界定。首先，对于含$\mathbf{Q}_{T_0-1}$ 因子的求和项，我们有

$$
\begin{aligned}
\sum_{t\in W_j} \langle \mathbf{Q}_{T_0-1}, \vec{\theta}_t \odot \vec{\sigma}_t - \vec{\lambda}_t\rangle &\overset{(a)}{\geq} \epsilon|W_j| \cdot \|\mathbf{Q}_{T_0-1}\|_1 \\
&\geq \epsilon \sum_{t\in W_j} \left(\|\mathbf{Q}_{t-1}\|_1 - \|\mathbf{Q}_{t-1} - \mathbf{Q}_{T_0-1}\|_1\right) \\
&\overset{(b)}{\geq} \epsilon \sum_{t\in W_j} \left(\|\mathbf{Q}_{t-1}\|_1 - KM(t-T_0)\right) \\
&\geq \epsilon \sum_{t\in W_j} \|\mathbf{Q}_{t-1}\|_1 - \epsilon KM\left(|W_j| - 1\right)^2
\end{aligned}
$$

其中，步骤（a）依据假设 1 中的公式（2），步骤（b）是因为队列长度增量有界于 M。对于含$\mathbf{Q}_{t-1} - \mathbf{Q}_{T_0-1}$ 的求和项，我们有

$$
\begin{aligned}
\sum_{t\in W_j} \langle \mathbf{Q}_{t-1} - \mathbf{Q}_{T_0-1}, \vec{\theta}_t \odot \vec{\sigma}_t - \vec{\lambda}_t\rangle &\geq -\sum_{t\in W_j} \|\mathbf{Q}_{t-1} - \mathbf{Q}_{T_0-1}\|_1 \cdot \|\vec{\theta}_t \odot \vec{\sigma}_t - \vec{\lambda}_t\|_\infty \\
&\geq -\sum_{t\in W_j} KM(t-T_0) \cdot M \\
&\geq -KM^2\left(|W_j| - 1\right)^2.
\end{aligned}
$$

将这两部分相加，我们得到

$$
\sum_{t\in W_j} \langle \mathbf{Q}_{t-1}, \vec{\theta}_t \odot \vec{\sigma}_t - \vec{\lambda}_t\rangle \geq \sum_{t\in W_j} \|\mathbf{Q}_{t-1}\|_1 + \left(KM^2 + \epsilon KM\right)\left(|W_j| - 1\right)^2. \tag{35}
$$

对于任意给定的时间范围 $T \geq 1$，我们可以找到最小的 n 使得 $\bigcup_{j=0}^{n} W_j \supseteq [T]$，那么 W_n 最左端点小于 T。根据假设 1 中的式 (1)，我们有 $(|W_n|-1)^2 \leq C_W T$，因此 $|W_n| \leq \sqrt{\frac{T}{C_W}} + 1$。令 $\mathcal{T}_T \triangleq \min_{t\in W_n}$ 为 W_n的最右端点，那么 W_0、.....、W_n 是 $[\mathcal{T}_T]$的一个划分，并且我们有

$$
\mathcal{T}_T \leq T + |W_n| \leq T + \sqrt{\frac{T}{C_W}} + 1.
$$

对 $j=0\dots n$ 求和式 (35)，得到

$$\begin{aligned}\sum_{t=1}^{\mathcal{T}_T}\langle \mathbf{Q}_{t-1},\vec{\theta}_t\odot\vec{\sigma}_t-\vec{\lambda}_t\rangle &\geq \sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1+(KM^2+\epsilon KM)\sum_{j=0}^{n}(|W_j|-1)^2\\ &\overset{(a)}{\geq}\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1+(KM^2+\epsilon KM)C_W\mathcal{T}_T\\ &\geq\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1+(KM^2+\epsilon KM)C_W\mathcal{T}_T\end{aligned}$$

其中在 (a) 中我们应用了式 (1)。 □

定理 *5.5* 的证明。 只需注意，当 $T\geq\max\{\frac{4}{C_W},C_W\}$ 时，我们有 $\sqrt{\frac{T}{C_W}}\geq\sqrt{\frac{C_W}{C_W}}=1$。此外，我们还有 $T^2\geq\frac{4T}{C_W}$，因此 $T\geq 2\sqrt{\frac{T}{C_W}}$。所以，定理 5.4 中的常数 $\mathcal{T}_T$ 满足

$$\begin{aligned}\mathcal{T}_T&\leq T+\sqrt{\frac{T}{C_W}}+1\\ &\leq T+2\sqrt{\frac{T}{C_W}}\\ &\leq 2T.\end{aligned}$$

然后，将时间范围长度 $\mathcal{T}_T$ 代入定理 5.3，我们得到

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}Q_{t-1,a_t}S_{t,a_t}-\langle\mathbf{Q}_{t-1},\vec{\lambda}_t\rangle\right]\leq\frac{(K+1)M^2\mathcal{T}_T}{2}. \tag{36}$$

通过重新排列定理 5.4 给出的界中的各项，我们得到

$$\begin{aligned}\epsilon\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right]&\leq\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\langle\mathbf{Q}_{t-1},\vec{\sigma}_t\odot\vec{\theta}_t-\vec{\lambda}_t\rangle\right]+(KM^2+\epsilon KM)C_W\mathcal{T}_T\\ &=\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}Q_{t-1,a_t}S_{t,a_t}-\langle\mathbf{Q}_{t-1},\vec{\lambda}_t\rangle\right]+(KM^2+\epsilon KM)C_W\mathcal{T}_T\\ &\quad-\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\langle\mathbf{Q}_{t-1},\vec{\sigma}_t\odot\vec{\theta}_t\rangle-Q_{t-1,a_t}S_{t,a_t}\right]\\ &\leq\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}Q_{t-1,a_t}S_{t,a_t}-\langle\mathbf{Q}_{t-1},\vec{\lambda}_t\rangle\right]+(KM^2+\epsilon KM)C_W\mathcal{T}_T+f(\mathcal{T}_T),\end{aligned} \tag{37}$$

其中，最后一个不等式源于假设 $\mathbb{E}\left[\sum_{t=1}^{T}Q_{t-1,a_t}S_{t,a_t}\right]\geq\mathbb{E}\left[\sum_{t=1}^{T}\langle\mathbf{Q}_{t-1},\vec{\sigma}_t\odot\vec{\theta}_t\rangle\right]-f(T)$。将式 (36) 代入式 (37)，可得

$$\epsilon\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right]\leq\frac{(K+1)M^2\mathcal{T}_T}{2}+(KM^2+\epsilon KM)C_W\mathcal{T}_T+f(\mathcal{T}_T)$$

因此

$$\begin{aligned}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_t\|_1\right]&\leq\frac{(K+1)M^2+2C_W(KM^2+\epsilon KM)}{2\epsilon}\cdot\frac{\mathcal{T}_T}{T}+\frac{f(\mathcal{T}_T)}{\epsilon T}\\ &\leq\frac{(K+1)M^2+2C_W(KM^2+\epsilon KM)}{\epsilon}+\frac{f(2T)}{\epsilon T}\end{aligned}$$

其中，最后一步是由 $\mathcal{T}_T\leq 2T$ 以及假设 $f(\cdot)$ 是递增的得出的。

□

E SoftMW（算法2）稳定性分析的证明

下面给出一个与定理 5.9 类似但方向相反的引理。

引理 E.1. 假设对于所有 $1 \le i < n$， $x_1 = 0, x_2, \dots, x_n \ge 0, |x_{i+1} - x_i| \le 1$，则有

$$\sum_{i=1}^{n} x_i^2 \ge \frac{1}{3} x_n^3.$$

证明。对于 $0 \le i \le \lceil x_n \rceil$，有 $x_{n-i} \ge x_n - i \ge 0$，因此

$$\begin{aligned}\sum_{i=1}^{n} x_i^2 &\ge \sum_{i=0}^{\lceil x_n \rceil} x_{n-i}^2 \ge \sum_{i=0}^{\lceil x_n \rceil} (x_n - i)^2 \\ &= \frac{1}{6} x_n (x_n + 1)(2x_n + 1) - \frac{1}{6} \{x_n\}(\{x_n\} + 1)(2\{x_n\} + 1) \\ &\ge \frac{1}{3} x_n^3\end{aligned}$$

其中 $\{x_n\} \triangleq x_n - \lceil x_n \rceil$。 □

推论 y E.2. 设 $\{\mathbf{Q}_t\}$ 为队列长度，其中 $\mathbf{Q}_0 = \mathbf{0}$，且增量有界于 M，则

$$\sum_{t=0}^{T-1} \|\mathbf{Q}_t\|_2^2 \ge \frac{1}{3MK^3} \|\mathbf{Q}_{T-1}\|_1^3$$

对于任意 $T \ge 1$。

证明。设 $i^* \in \arg\max_{i \in [K]} Q_{T-1,i}$，则

$$\frac{Q_{0,i^*}}{M}, \frac{Q_{1,i^*}}{M}, \dots, \frac{Q_{T-1,i^*}}{M}$$

这是一个从 0 开始且差值有界于 1 的非负序列。将定理 E.1 应用于该序列，可得

$$\begin{aligned}\sum_{t=0}^{T-1} \|\mathbf{Q}_t\|_2^2 &\ge M^2 \sum_{t=0}^{T-1} \left(\frac{Q_{t,i^*}}{M} \right)^2 \\ &\ge \frac{M^2}{3} \left(\frac{Q_{T-1,i^*}}{M} \right)^3 \\ &= \frac{1}{3M} Q_{T-1,i^*}^3 \\ &\ge \frac{1}{3M} \left(\frac{\|\mathbf{Q}_{T-1}\|_1}{K} \right)^3 \\ &= \frac{1}{3MK^3} \|\mathbf{Q}_{T-1}\|_1^3\end{aligned}$$

如需证明。 □

定理 *5.6* 的证明。只需验证在每个时间步

$$\begin{aligned}2\|\mathbf{Q}_{t-1}\|_1 &\le t^{-(\frac{1}{4} - \frac{\delta}{2})} \sqrt{86M^2K^6 t^{\frac{3}{2}} + \sum_{s=0}^{t-1} \|\mathbf{Q}_s\|_2^2}, \\ &\Updownarrow \\ 4\|\mathbf{Q}_{t-1}\|_1^2 &\le t^{-\frac{1}{2} + \delta} \left(86M^2K^6 t^{\frac{3}{2}} + \sum_{s=0}^{t-1} \|\mathbf{Q}_s\|_2^2 \right),\end{aligned}$$

$$
\begin{aligned}
&\Uparrow (a)\\
4\|\mathbf{Q}_{t-1}\|_1^2 &\le t^{-\frac{1}{2}+\delta}\left(86M^2K^6t^{\frac{3}{2}}+\frac{1}{3MK^3}\|\mathbf{Q}_{t-1}\|_1^3\right)\\
&\Updownarrow\\
4\|\mathbf{Q}_{t-1}\|_1^2 &\le t^{-\frac{1}{2}+\delta}\left(\frac{1}{3}\cdot 258M^2K^6t^{\frac{3}{2}}+\frac{2}{3}\cdot\frac{1}{2MK^3}\|\mathbf{Q}_{t-1}\|_1^3\right),\\
&\Uparrow (b)\\
4\|\mathbf{Q}_{t-1}\|_1^2 &\le t^{-\frac{1}{2}+\delta}\cdot 258^{\frac{1}{3}}2^{-\frac{2}{3}}t^{\frac{1}{2}}\|\mathbf{Q}_{t-1}\|_1^2\\
&\Updownarrow\\
4 &\le t^{\delta}\left(\frac{258}{4}\right)^{\frac{1}{3}},
\end{aligned}
$$

且最后一个结论显然成立。这里在步骤 (a) 中应用定理 E.2，在步骤 (b) 中应用算术-几何平均不等式 $\frac{1}{3}x+\frac{2}{3}y\ge x^{\frac{1}{3}}y^{\frac{2}{3}}$。 □

定理 *5.7* 的证明。对于每个 $\vec{\theta}_t$，我们可以找到一个 $\vec{\theta}'_t\in\Delta^{[K],\ \beta_t}$，使得 $\|\vec{\theta}_t-\vec{\theta}'_t\|_1\le 2K\beta_t$ 且 $\|\vec{\theta}'_s-\vec{\theta}'_t\|_1\le\|\vec{\theta}_s-\vec{\theta}_t\|_1$。例如，我们可以选择

$$\vec{\theta}'_t=(1-\beta_t)\vec{\theta}_t+\beta_t\mathbf{1}.$$

那么，我们可以写为

$$
\begin{aligned}
\langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}_t\rangle &= \langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}'_t\rangle+\langle\mathbf{Q}_{t-1}\odot\mathbf{S}_t,\vec{\theta}_t-\vec{\theta}'_t\rangle\\
&\le \langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}'_t\rangle+\|\mathbf{Q}_{t-1}\odot\mathbf{S}_t\|_\infty\cdot\|\vec{\theta}_t-\vec{\theta}'_t\|_1\\
&\le \langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}'_t\rangle+M^2t\cdot 2K\beta_t\\
&= \langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}'_t\rangle+2M^2t^{-2}.
\end{aligned}
$$

那么，可以看出该量

$$\sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}'_t\rangle-Q_{t-1,a_t}S_{t,a_t}\right]=\sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbf{Q}_{t-1}\odot\mathbf{S}_t,\vec{\theta}'_t\rangle-(\mathbf{Q}_{t-1}\odot\mathbf{S}_t)_{a_t}\right]$$

满足应用定理 5.1 的条件。定理 5.1 断言

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}'_t\rangle-Q_{t-1,a_t}S_{t,a_t}\right]\\
\le&\left(1+\sum_{t=1}^{T-1}\|\vec{\theta}'_{t+1}-\vec{\theta}'_t\|_1\right)\mathbb{E}\left[\eta_T^{-1}\ln\frac{1}{\beta_T}\right]+e\mathbb{E}\left[\sum_{t=1}^{T}\eta_t\|\mathbf{Q}_{t-1}\odot\mathbf{S}_t\|_2^2\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\left\langle\mathbf{Q}_{t-1}\odot\mathbf{S}_t,\frac{\mathbf{Q}_{t-1}}{\|\mathbf{Q}_{t-1}\|_1}\right\rangle\right].
\end{aligned}
\tag{38}
$$

下面，我们对公式 (38) 右端的每一项进行界定。首先，根据假设 2，我们有 $\sum_{t=1}^{T-1}\|\vec{\theta}'_{t+1}-\vec{\theta}'_t\|_1\le\sum_{t=1}^{T-1}\|\vec{\theta}_{t+1}-\vec{\theta}_t\|_1\le C_VT^{\frac{1}{2}-\delta}$，因此

$$
\begin{aligned}
&\left(1+\sum_{t=1}^{T-1}\|\vec{\theta}'_{t+1}-\vec{\theta}'_t\|_1\right)\mathbb{E}\left[\eta_T^{-1}\ln\frac{1}{\beta_T}\right]\\
\le&\left(1+C_VT^{\frac{1}{2}-\delta}\right)\mathbb{E}\left[\eta_T^{-1}\ln\frac{1}{\beta_T}\right]\\
\le&\left(1+C_VT^{\frac{1}{2}-\delta}\right)(3\ln T+\ln K)\,\mathbb{E}\left[T^{-(\frac{1}{4}-\frac{\delta}{2})}M\sqrt{86M^2K^6T^{\frac{3}{2}}+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right]
\end{aligned}
$$

$$\leq (1+C_V)\, T^{\frac{1}{4}-\frac{\delta}{2}} \left(3\ln T + \ln K\right) M\mathbb{E}\left[\sqrt{86M^2K^6T^{\frac{3}{2}} + \sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right].$$

对于第二项，我们有

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\mathbf{Q}_{t-1}\odot\mathbf{S}_t\|_2^2 \\
&\leq M^2\sum_{t=1}^{T}\eta_t\|\mathbf{Q}_{t-1}\|_2^2 \\
&= M\sum_{t=1}^{T} t^{\frac{1}{4}-\frac{\delta}{2}}\left(\sqrt{86M^2K^6t^{\frac{3}{2}} + \sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}\cdot\|\mathbf{Q}_{t-1}\|_2^2 \\
&\leq M\sum_{t=1}^{T} t^{\frac{1}{4}-\frac{\delta}{2}}\left(\sqrt{1 + \sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}\cdot\|\mathbf{Q}_{t-1}\|_2^2 \\
&\leq MT^{\frac{1}{4}-\frac{\delta}{2}}\sum_{t=1}^{T}\left(\sqrt{1 + \sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}\cdot\|\mathbf{Q}_{t-1}\|_2^2 \\
&\leq 2MT^{\frac{1}{4}-\frac{\delta}{2}}\sqrt{1 + \sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2},
\end{aligned}$$

其中在最后一步，我们利用了以下事实:

$$\sum_{i=1}^{n}\frac{x_i}{\sqrt{1+\sum_{j=1}^{i}x_j}} \leq 2\sqrt{1+\sum_{i=1}^{n}x_i}$$

对于非负的 $x_1\ldots,\ x_n$。

对于第三项，我们有

$$\begin{aligned}
&\sum_{t=1}^{T}\gamma_t\left\langle\mathbf{Q}_{t-1}\odot\mathbf{S}_t, \frac{\mathbf{Q}_{t-1}}{\|\mathbf{Q}_{t-1}\|_1}\right\rangle \\
&\leq M\sum_{t=1}^{T}\gamma_t\|\mathbf{Q}_{t-1}\|_1^{-1}\|\mathbf{Q}_{t-1}\|_2^2 \\
&= M\sum_{t=1}^{T} t^{\frac{1}{4}-\frac{\delta}{2}}\left(\sqrt{1 + \sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}\cdot\|\mathbf{Q}_{t-1}\|_2^2
\end{aligned}$$

那么它可以用与第二项完全相同的方式进行界定，因此它也不超过

$$2MT^{\frac{1}{4}-\frac{\delta}{2}}\sqrt{1 + \sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}.$$

将所有内容合并后再取期望，我们可以看出

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbf{Q}_{t-1}, \mathbf{S}_t\odot\vec{\theta}_t'\rangle - Q_{t-1,a_t}S_{t,a_t}\right] \\
&\leq T^{\frac{1}{4}-\frac{\delta}{2}}M\mathbb{E}\left[\sqrt{86M^2K^6T^{\frac{3}{2}} + \sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right]\cdot\left\{(1+C_V)\left(3\ln T+\ln K\right)+2e+2\right\} + 2M^2\cdot\frac{\pi^2}{6}
\end{aligned}$$

$$\leq (2e+3)\cdot(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(3\ln T+\ln K)M\mathbb{E}\left[\sqrt{86M^2K^6T^{\frac{3}{2}}+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right]+4M^2$$

$$\leq 9(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(3\ln T+\ln K)M\mathbb{E}\left[\sqrt{86M^2K^6T^{\frac{3}{2}}+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right]+4M^2.$$

□

定理 *5.9* 的证明。将 $x_1,\ldots,x_n$ 按升序排列为 $y_1\leq y_2\leq\cdots\leq y_n$。接下来我们断言，$y_1,\ldots,y_n$ 也是一个非负数序列，且相邻项之差不超过 1，即对于所有 $1\leq i<n$，$|y_{i+1}-y_i|\leq 1$。

令 $\sigma\colon[n]\to[n]$ 是 $[n]$ 上的一个排列，使得对任意 $1\leq i\leq n$，都有 $y_i=x_{\sigma(i)}$。对于任意 $1\leq i<n$，不失一般性，假设 $\sigma(i+1)>\sigma(i)$，令

$$j\triangleq\min\left\{k:\sigma(i)<k\leq\sigma(i+1),x_k\geq x_{\sigma(i)}\right\}.$$

由于我们已假设 $\sigma(i+1)>\sigma(i)$，因此可以保证 j 定义良好。根据 j 的定义，以及 $\{y_i\}$ 是 $\{x_i\}$ 的有序列表这一事实，我们有

$$x_{j-1}\overset{(a)}{\leq}x_{\sigma(i)}=y_i\leq y_{i+1}=x_{\sigma(i+1)}\overset{(b)}{\leq}x_j,$$

其中，不等式 (a) 根据 j 的定义得出（j 是 $\{x_i\}$ 中位于 $x_{\sigma(i)}$ 之后且大于 y_i 的第一个元素的索引），不等式 (b) 根据 $x_j\geq y_i$ 得出，并且 y_{i+1} 是大于 y_i 的数中最小的元素。因此，

$$|y_{i+1}-y_i|=y_{i+1}-y_i\leq x_j-x_{j-1}\leq|x_j-x_{j-1}|\leq 1.$$

$\sigma(i+1)<\sigma(i)$ 的情形也可用类似方法验证。

定义

$$s\triangleq\lceil y_n\rceil=\left\lceil\max_{1\leq i\leq n}x_i\right\rceil,$$

那么，对于所有 $0\leq i<s$，我们有 $y_{n-i}\geq y_n-i\geq 0$，因此

$$S=\sum_{i=1}^{n}y_i\geq\sum_{i=0}^{s-1}y_{n-i}\geq\sum_{i=0}^{s-1}(y_n-i)=\frac{s(2y_n-s+1)}{2}\geq\frac{y_n^2}{2},$$

其中，最后一个不等式由 $y_n\leq s\leq y_n+1$ 得出，因此 $2y_n-s+1\geq y_n$。因此 $y_n\leq\sqrt{2S}$。然后，很容易得出

$$\sum_{i=1}^{n}x_i^2\leq y_n\sum_{i=1}^{n}x_i\leq\sqrt{2S}\cdot S=\sqrt{2}S^{\frac{3}{2}}.$$

□

定理 *5.10* 的证明。令 $z(x)\triangleq y(x)^{\frac{1}{4}}$，那么我们有

$$z(x)^4\leq f(x)+z(x)g(x)$$

对于所有 $x\geq 0$。注意，对于任意固定的 $x\geq 0$，函数

$$h(z)\triangleq z^4-g(x)z-f(x)$$

在 $\left[\left(\frac{g(x)}{4}\right)^{\frac{1}{3}},\infty\right)$ 上递增。为了证明定理 5.10，只需证明，$z=f(x)^{\frac{1}{4}}+g(x)$ 的特定选择能够保证 $z\geq\left(\frac{g(x)}{4}\right)^{\frac{1}{3}}$ 以及 $h(z)\geq 0$。条件 $z\geq\left(\frac{g(x)}{4}\right)^{\frac{1}{3}}$。首先，条件 $z\geq\left(\frac{g(x)}{4}\right)^{\frac{1}{3}}$ 显然成立，因为我们已假设 $f(x)$ 和 $g(x)$ 都不小于 1。然后，我们可以直接计算

$$z^4=\left(f(x)^{\frac{1}{4}}+g(x)\right)^4$$

$$= f(x) + 4f(x)^{\frac{3}{4}}g(x) + 6f(x)^{\frac{1}{2}}g(x)^2 + 4f(x)^{\frac{1}{4}}g(x)^3 + g(x)^4,$$

因此，

$$\begin{aligned} h(z) &= z^4 - g(x)\left(f(x)^{\frac{1}{4}} + g(x)\right) - f(x) \\ &= \left(4f(x)^{\frac{3}{4}}g(x) - f(x)^{\frac{1}{4}}g(x)\right) + \left(6f(x)^{\frac{1}{2}}g(x)^2 - g(x)^2\right) + 4f(x)^{\frac{1}{4}}g(x)^3 + g(x)^4 \end{aligned}$$

其中可以看出，所有项均为非负，因为 $f(x)\ g(x) \geq 1$。 □

F 算法 3 的证明：稳定性分析

定理 *6.2* 的证明。与定理 5.7 的证明类似，只需应用定理 5.1，但我们需要先验证所有 γ_τ 都不超过 $\frac{1}{2}$。

我们选择 m，保证了 $m \geq \|\mathbf{Q}_{2M}^{T_0}\|_\infty$，因此 $\|\mathbf{Q}_{T_0}\|_\infty \leq 2Mm$，并且对于此时期内的任意 τ，都有 $\|\mathbf{Q}_{T_0+\tau}\|_\infty \leq \|\mathbf{Q}_{T_0}\|_\infty + M\tau \leq 3Mm$。因此

$$\begin{aligned} \gamma_\tau &= \frac{1}{6}K^{-1}M^{-1}m^{-1-\frac{\delta}{2}}\|\mathbf{Q}_{T_0+\tau-1}\|_1 \\ &\leq \frac{1}{6}K^{-1}M^{-1}m^{-1-\frac{\delta}{2}} \cdot 3KMm \\ &= \frac{1}{2}m^{-\frac{\delta}{2}} \\ &\leq \frac{1}{2}. \end{aligned}$$

对于每个 $\vec{\theta}_{T_0+\tau}$，我们都可以找到一个 $\vec{\theta}'_{T_0+\tau} \in \Delta^{[K],\ \beta_\tau}$，使得 $\|\vec{\theta}_{T_0+\tau} - \vec{\theta}'_{T_0+\tau}\|_1 \leq 2K\beta_\tau$，并且对于任意 $1 \leq s, t \leq m$，都有 $\|\vec{\theta}'_s - \vec{\theta}'_t\|_1 \leq \|\vec{\theta}_{T_0+s} - \vec{\theta}_{T_0+t}\|_1$。例如，我们可以选择

$$\vec{\theta}'_{T_0+\tau} = (1-\beta_\tau)\vec{\theta}_{T_0+\tau} + \beta_\tau \mathbf{1}.$$

于是，我们可以写成

$$\begin{aligned} \langle \mathbf{Q}_{T_0+\tau-1}, \mathbf{S}_{T_0+\tau} \odot \vec{\theta}_{T_0+\tau}\rangle &= \langle \mathbf{Q}_{T_0+\tau-1}, \mathbf{S}_{T_0+\tau} \odot \vec{\theta}'_{T_0+\tau}\rangle + \langle \mathbf{Q}_{T_0+\tau-1} \odot \mathbf{S}_{T_0+\tau}, \vec{\theta}_{T_0+\tau} - \vec{\theta}'_{T_0+\tau}\rangle \\ &\leq \langle \mathbf{Q}_{T_0+\tau-1}, \mathbf{S}_{T_0+\tau} \odot \vec{\theta}'_{T_0+\tau}\rangle + \|\mathbf{Q}_{T_0+\tau-1} \odot \mathbf{S}_{T_0+\tau}\|_\infty \cdot \|\vec{\theta}_{T_0+\tau} - \vec{\theta}'_{T_0+\tau}\|_1 \\ &\leq \langle \mathbf{Q}_{T_0+\tau-1}, \mathbf{S}_{T_0+\tau} \odot \vec{\theta}'_{T_0+\tau}\rangle + 3Mm \cdot M \cdot 2K\beta_\tau \\ &= \langle \mathbf{Q}_{T_0+\tau-1}, \mathbf{S}_{T_0+\tau} \odot \vec{\theta}'_{T_0+\tau}\rangle + 6M^2m^{-1}. \end{aligned}$$

于是，可以看出这个量

$$\begin{aligned} &\sum_{\tau=1}^{m} \mathbb{E}\left[\langle \mathbf{Q}_{T_0+\tau-1}, \mathbf{S}_{T_0+\tau} \odot \vec{\theta}'_{T_0+\tau}\rangle - Q_{T_0+\tau-1,a_{T_0+\tau}} S_{T_0+\tau,a_{T_0+\tau}} \middle| \mathcal{F}_{T_0}\right] \\ =& \sum_{\tau=1}^{m} \mathbb{E}\left[\langle \mathbf{Q}_{T_0+\tau-1} \odot \mathbf{S}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle - (\mathbf{Q}_{T_0+\tau-1} \odot \mathbf{S}_{T_0+\tau})_{a_{T_0+\tau}} \middle| \mathcal{F}_{T_0}\right] \end{aligned}$$

满足应用定理 5.1 的条件。定理 5.1 断言

$$\begin{aligned} &\sum_{\tau=1}^{m} \mathbb{E}\left[\langle \mathbf{Q}_{T_0+\tau-1}, \mathbf{S}_{T_0+\tau} \odot \vec{\theta}'_{T_0+\tau}\rangle - Q_{T_0+\tau-1,a_{T_0+\tau}} S_{T_0+\tau,a_{T_0+\tau}} \middle| \mathcal{F}_{T_0}\right] \\ \leq& \left(1 + \sum_{\tau=1}^{m-1} \|\vec{\theta}'_{T_0+\tau+1} - \vec{\theta}'_{T_0+\tau}\|_1\right) \mathbb{E}\left[\eta_m^{-1} \ln \frac{1}{\beta_m} \middle| \mathcal{F}_{T_0}\right] + e\mathbb{E}\left[\sum_{\tau=1}^{m} \eta_\tau \|\mathbf{Q}_{T_0+\tau-1} \odot \mathbf{S}_{T_0+\tau}\|_2^2 \middle| \mathcal{F}_{T_0}\right] \\ &+ \mathbb{E}\left[\sum_{\tau=1}^{m} \gamma_\tau \left\langle \mathbf{Q}_{T_0+\tau-1} \odot \mathbf{S}_{T_0+\tau}, \frac{\mathbf{Q}_{T_0+\tau-1}}{\|\mathbf{Q}_{T_0+\tau-1}\|_1} \right\rangle \middle| \mathcal{F}_{T_0}\right] \end{aligned}$$

$$
\begin{aligned}
&\leq \left(1+\sum_{\tau=1}^{m-1}\|\vec{\theta}'_{T_0+\tau+1}-\vec{\theta}'_{T_0+\tau}\|_1\right)\cdot 6M^2Km^{1+\frac{\delta}{2}}\cdot(2\ln m+\ln K)\\
&\quad+(e+1)\frac{1}{6}K^{-1}m^{-1-\frac{\delta}{2}}\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{Q}_{T_0+\tau-1}\|_2^2\middle|\mathcal{F}_{T_0}\right]\\
&\leq 6\left(1+C_Vm^{1-\delta}\right)\cdot M^2Km^{1+\frac{\delta}{2}}\cdot(2\ln m+\ln K)\\
&\quad+(e+1)\frac{1}{6}K^{-1}m^{-1-\frac{\delta}{2}}\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{Q}_{T_0+\tau-1}\|_2^2\middle|\mathcal{F}_{T_0}\right],
\end{aligned}
$$

其中在最后一步中，我们使用了假设3中对 $\sum_{\tau=1}^{m-1}\|\vec{\theta}'_{T_0+\tau+1}-\vec{\theta}'_{T_0+\tau}\|_1$ 的界。因此，我们有

$$
\begin{aligned}
&\sum_{\tau=1}^{m}\mathbb{E}\left[\langle\mathbf{Q}_{T_0+\tau-1},\mathbf{S}_{T_0+\tau}\odot\vec{\theta}_{T_0+\tau}\rangle-Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}}\middle|\mathcal{F}_{T_0}\right]\\
&\leq 6\left(1+C_V\right)M^2Km^{2-\frac{\delta}{2}}\cdot(2\ln m+\ln K)+K^{-1}m^{-1-\frac{\delta}{2}}\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{Q}_{T_0+\tau-1}\|_2^2\middle|\mathcal{F}_{T_0}\right]+6M^2.
\end{aligned}
$$

□

定理 *6.3* 的证明。 这些不等式都是有界增量假设的直接推论。记$i^*\in\arg\max_{i\in[K]}Q_{T_0,\ i}$，则有

$$
\begin{aligned}
Q_{t,i^*}&\overset{(a)}{\geq}Q_{T_0,i^*}-(t-T_0)\cdot M\\
&\geq Q_{T_0,i^*}-\left(\frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M}+1\right)\cdot M\\
&=\frac{\|\mathbf{Q}_{T_0}\|_\infty}{2}-M\\
&\overset{(b)}{\geq}\frac{\|\mathbf{Q}_{T_0}\|_\infty}{4}
\end{aligned}
$$

对所有 $T_0+1\leq t\leq T_0+\frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M}+1$均成立。此处步骤 (a) 源于队列长度增量的有界性，步骤 (b) 源于 $\|\mathbf{Q}_{T_0}\|_\infty\geq 4M$这一假设。因此，对于这些 t，有

$$
\|\mathbf{Q}_t\|_\infty\geq Q_{t,i^*}\geq\frac{\|\mathbf{Q}_{T_0}\|_\infty}{4}.
$$

此外，对于任意 $T_0+1\leq t\leq T_0+\frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M}+1$ 和 $i\in[K]$，有

$$
\begin{aligned}
Q_{t,i}&\overset{(a)}{\leq}Q_{T_0,i}+(t-T_0)\cdot M\\
&\leq\|\mathbf{Q}_{T_0}\|_\infty+\left(\frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M}+1\right)\cdot M\\
&=\frac{3}{2}\|\mathbf{Q}_{T_0}\|_\infty+M\\
&\overset{(b)}{\leq}2\|\mathbf{Q}_{T_0}\|_\infty.
\end{aligned}
$$

同理，步骤（a）是由于队列长度增量有界，步骤（b）是由于假设$\|\mathbf{Q}_{T_0}\|_\infty\geq 4M$。因此，

$$
\|\mathbf{Q}_t\|_\infty=\max_{i\in[K]}Q_{t,i}\leq 2\|\mathbf{Q}_{T_0}\|_\infty.
$$

□

定理 *6.4* 的证明。根据定理 6.3，当 $m = \lceil \frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M} \rceil$ 时，我们有

$$Mm^2 \le 4\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_1 \tag{39}$$

and

$$\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_2^2 \le 16KM^2m^3 = 16KMm \cdot Mm^2 \le 64KMm \sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_1. \tag{40}$$

然后，我们可以将这两个不等式应用于定理 6.2 的公式 (8)，从而得到一个关于 $\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_1$ 的上界。具体来说，对于公式 (8) 右端的 $6\,(1+C_V)\;M^2Km^{2-\frac{\delta}{2}} \cdot (2\ln\, m + \ln\, K)$ 项，我们可以应用公式 (39)，将其中一个 Mm^2 因子上界为 $4\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_1$，因此该项不超过

$$24\,(1+C_V)\,MKm^{-\frac{\delta}{2}} \cdot (2\ln m + \ln K) \cdot \mathbb{E}\left[\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_1 \middle| \mathcal{F}_{T_0}\right].$$

对于 $K^{-1}m^{-1-\frac{\delta}{2}}\mathbb{E}\left[\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_2^2 \middle| \mathcal{F}_{T_0}\right]$ 项，我们可以将（在每条样本路径上）$\sqrt{\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_2^2}$ 因子上界为 $64KMm\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_1$。因此，该项不超过

$$64Mm^{-\frac{\delta}{2}}\mathbb{E}\left[\sum_{\tau=1}^{m} \|\mathbf{Q}_{T_0+\tau-1}\|_1 \middle| \mathcal{F}_{T_0}\right].$$

将这两项的上界合并，即可得到所述上界。 □

G 处理有界矩下队列长度增量的高层思路

这里，我们解释 SoftMW+ 和 SSMW+设计背后的直觉。回顾一下，在 SoftMW 和 SSMW 的设计与分析中，有界增量这一原始假设主要用于以下两个步骤:

- （EXP3.S+ **反馈信号尺度**）为了应用定理 5.1，必须确保输入 EXP3.S+ 的每个反馈值$g_{t,\ a_t}$ 都不超过 $\eta_t^{-1}\gamma_t e_{t,\ a_t}$。这正是我们将 $\mathbf{e}_t$ 设为归一化的当前队列长度向量，并在有界增量情形下将 $\gamma_t = \eta_t \cdot M\|\mathbf{Q}_t\|_1$ 设为.....的主要原因。

- （通过 $\sum\|\mathbf{Q}_t\|_1$**界定 EXP3.S+ 的遗憾**）成功应用定理 5.1（并可能对 EXP3.S+ 的多个实例求和）后，我们得到时间 T结束时 Lyapunov 二次函数期望值的一个上界。与通过标准李雅普诺夫分析在平稳情形下使用最大权重所能得到的界相比，我们的界多出一项，其中含有一个 $\sqrt{\sum\|\mathbf{Q}_t\|_2^2}$ 因子。随后，我们对 SoftMW 应用定理 5.9，对 SSMW 应用定理 6.3，从而将 $\sqrt{\sum\|\mathbf{Q}_t\|_2^2}$ 上界化为关于 $\sum\|\mathbf{Q}_t\|_1$的表达式。

因此，如果我们希望在只有有界矩的情况下处理可能无界的队列长度增量，就需要对上述两个步骤的论证进行推广。下面介绍克服无界增量所带来挑战的总体思路。

跳过 EXP3.S 中较大的反馈值+

假设当前处于时间步 t，并在下一步 EXP3.S+中选择 $\mathbf{e}_t = \mathbf{Q}_{t-1}/\|\mathbf{Q}_{t-1}\|_1$ 和 $\gamma_t = C_t \cdot M\eta_t\|\mathbf{Q}_{t-1}\|_1$，其中 $C_t > 0$ 是一个常数，将在时间步 $t-1$ 结束时确定。这样，EXP3.S+ 可以处理不超过 $C_t \cdot MQ_{t-1,a_t}$ 的新反馈信号，其中 a_t 是我们在此时间步最终选择服务的队列索引。换言之，仅当 $S_{t,a_t} \le C_tM$ 时，我们才能将该信号输入 EXP3。由于应用 EXP3.S+ 的目标是使 $\langle \mathbf{Q}_{t-1} \odot \mathbf{S}_t, \vec{\theta}_t\rangle - Q_{t-1,a_t}S_{t,a_t}$ 的总和尽可能小，因此可以写为

$$\langle \mathbf{Q}_{t-1} \odot \mathbf{S}_t, \vec{\theta}_t\rangle - Q_{t-1,a_t}S_{t,a_t} \le \mathbb{1}[S_{t,a_t} > C_tM]\langle \mathbf{Q}_{t-1} \odot \mathbf{S}_t, \vec{\theta}_t\rangle + \langle \mathbf{Q}_{t-1} \odot \mathbf{S}'_t, \vec{\theta}_t\rangle - Q_{t-1,a_t}S'_{t,a_t}$$

其中

$$S'_{t,i} = \begin{cases} S_{t,i} & \text{if } S_{t,i} \le C_tM \\ 0 & \text{otherwise} \end{cases}$$

是 $S_{t,i}$ 的截断。然后，$\langle \mathbf{Q}_{t-1} \odot \mathbf{S}'_t, \vec{\theta}_t\rangle - Q_{t-1,a_t}S'_{t,a_t}$ 是一个具有期望累积和的量，可视为多臂老虎机问题的遗憾，因此，如果我们取 $Q_{t-1,a_t}S'_{t,a_t}$ 作为

反馈值。现在，所得遗憾上界中会有一个乘法因子 $\emptyset(\max_{1\le t\le T} C_t)$。另请注意，

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}[S_{t,a_t} > C_t M]\langle \mathbf{Q}_{t-1}\odot \mathbf{S}_t, \vec{\theta}_t\rangle \middle| \mathcal{F}_{t-1}\right]\\
&= \mathbb{E}\left[\mathbb{1}[S_{t,a_t} > C_t M]\sum_{i\neq a_t} Q_{t-1,i}S_{t,i}\theta_{t,i}\middle|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\mathbb{1}[S_{t,a_t} > C_t M]Q_{t-1,a_t}S_{t,a_t}\theta_{t,a_t}|\mathcal{F}_{t-1}\right]\\
&\overset{(a)}{=} \mathbb{E}\left[\mathbb{P}\left[S_{t,a_t} > C_t M|\mathcal{F}_{t-1}, a_t\right]\sum_{i\neq a_t} Q_{t-1,i}\sigma_{t,i}\theta_{t,i}\middle|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\mathbb{1}[S_{t,a_t} > C_t M]Q_{t-1,a_t}S_{t,a_t}\theta_{t,a_t}|\mathcal{F}_{t-1}\right]\\
&\le \mathbb{P}\left[S_{t,a_t} > C_t M|\mathcal{F}_{t-1}\right]\langle \mathbf{Q}_{t-1}\odot\vec{\sigma}_t, \vec{\theta}_t\rangle + \mathbb{E}\left[\mathbb{1}[S_{t,a_t} > C_t M]Q_{t-1,a_t}S_{t,a_t}\theta_{t,a_t}|\mathcal{F}_{t-1}\right]\\
&\le \mathbb{P}\left[S_{t,a_t} > C_t M|\mathcal{F}_{t-1}\right]\langle \mathbf{Q}_{t-1}\odot\vec{\sigma}_t, \vec{\theta}_t\rangle + \|\mathbf{Q}_{t-1}\|_\infty \mathbb{E}\left[\mathbb{1}[S_{t,a_t} > C_t M]S_{t,a_t}|\mathcal{F}_{t-1}\right]\\
&\overset{(b)}{\le} C_t^{-\alpha}\langle \mathbf{Q}_{t-1}\odot\vec{\sigma}_t, \vec{\theta}_t\rangle + C_t^{-\alpha+1}M\|\mathbf{Q}_{t-1}\|_\infty\\
&\le 2C_t^{-\alpha+1}M\|\mathbf{Q}_{t-1}\|_\infty.
\end{aligned}
$$

在步骤（a）中，我们先对 $\mathbb{1}[S_{t,a_t} > C_t M]\sum_{i\neq a_t} Q_{t-1,\ i}S_{t,\ i}\theta_{t,\ i}$ 关于 a_t 取条件期望，然后再对 $\mathcal{F}_{t-1}$ 取条件期望。在步骤（b）中，我们利用了以下事实：$\mathbb{P}\left[S_{t,a_t} > C_t M|\mathcal{F}_{t-1}\right] \le C_t^{-\alpha}$ 且 $\mathbb{E}\left[\mathbb{1}[S_{t,a_t} > C_t M]S_{t,a_t}|\mathcal{F}_{t-1}\right] \le \mathbb{E}[\mathbb{1}[S_{t,a_t} > C_t M]S_{t,a_t}^\alpha/(C_t M)^{\alpha-1}|\mathcal{F}_{t-1}] \le (C_t M)^{-\alpha+1}\mathbb{E}[S_{t,a_t}^\alpha|\mathcal{F}_{t-1}] \le MC_t^{-\alpha+1}$。因此，只要选择足够小的 C_ts，$\mathbb{1}[S_{t,a_t} > C_t M]\langle \mathbf{Q}_{t-1}\odot\mathbf{S}_t,\ \vec{\theta}_t\rangle$ 的期望累积和也能得到良好控制。

通过选择适当的 C_ts，我们可以权衡裁剪误差 $\sum\mathbb{E}\left[\mathbb{1}[S_{t,a_t} > C_t M]\langle \mathbf{Q}_{t-1}\odot\mathbf{S}_t,\ \vec{\theta}_t\rangle\right]$ 与裁剪后多臂老虎机遗憾 $\sum\mathbb{E}\left[\langle \mathbf{Q}_{t-1}\odot\mathbf{S}'_t,\ \vec{\theta}_t\rangle - Q_{t-1,\ a_t}S'_{t,\ a_t}\right]$，使二者之和（即裁剪前多臂老虎机遗憾 $\sum\mathbb{E}\left[\langle \mathbf{Q}_{t-1}\odot\mathbf{S}_t, \vec{\theta}_t\rangle - Q_{t-1,\ a_t}S_{t,\ a_t}\right]$）不至于过大。

在 $\sum\|\mathbb{Q}_t\|_1$ 和 $\sum\|\mathbb{Q}_t\|_2^2$ 之间的一般转换引理

到目前为止，我们已经处理了在应用 EXP3.S+的遗憾界时，无界反馈值所带来的问题。接下来还需要用 $\sum\|\mathbb{Q}_t\|_1$ 来界定 $\sum\|\mathbb{Q}_t\|_2^2$，从而最终得到整个时间范围内 $\sum\|\mathbb{Q}_t\|_1$ 的上界。

事实上，定理 5.9 和定理 6.3 都是针对 $\sum\|\mathbb{Q}_t\|_2^2$ 的样本路径界，并且在任意调度策略下均成立。具体而言，定理 5.9 指出，在整个时间范围内，$\sum\|\mathbb{Q}_t\|_2^2$ 为 $\emptyset(\sum\|\mathbb{Q}_t\|_1)^{\frac{3}{2}}$；而定理 6.3 则说明，在任意连续的 $m = \Theta(\|\mathbf{Q}_{T_0}\|_\infty/M)$ 个时间步内，如何利用 $\sum\|\mathbb{Q}_t\|_1$ 和 $\sum\|\mathbb{Q}_t\|_2^2$ 相互界定对方（其中，T_0 是所考察的 m 个时间步中的第一个时间步）。另一方面，我们的分析实际上只需要建立期望值 $\mathbb{E}\left[\sqrt{\sum\|\mathbb{Q}_t\|_2^2}\right]$ 和 $\mathbb{E}[\sum\|\mathbb{Q}_t\|_1]$ 之间的关系。因此，对于具有有界矩的队列大小增量，这一结果或许可以推广。

令人惊讶的是，尝试为定理 5.9 和定理 6.3 推导期望形式时，会发现它们的表现截然不同。定理 6.3 有一个直接的期望形式替代版本，但定理 5.9 似乎并非如此。具体而言，我们可以将定理 6.3 推广为定理 G.1。

引理 G.1. 在假设 *4* 下，对于任意 $T_0 \ge 0$，令 $m = \lceil \frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M} \rceil$；此外，假设 $\|\mathbf{Q}_{T_0}\|_\infty \ge 8M$；则

$$\frac{M}{2}m^2 \le \mathbb{E}\left[\sum_{t=1}^{m}\|\mathbf{Q}_{T_0+t-1}\|_1\middle|\mathcal{F}_{T_0}\right] \le 3KMm^2,$$

$$M^2m^3 \le \mathbb{E}\left[\sum_{t=1}^{m}\|\mathbf{Q}_{T_0+t-1}\|_2^2\middle|\mathcal{F}_{T_0}\right] \le 5KM^2m^3.$$

因此，我们有

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{T_0+t-1}\|_2^2}\middle|\mathcal{F}_{T_0}\right] \le \sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_{T_0+t-1}\|_2^2\middle|\mathcal{F}_{T_0}\right]} \le \frac{2\sqrt{5K}}{\sqrt{m}}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_{T_0+t-1}\|_1\middle|\mathcal{F}_{T_0}\right].$$

证明。 见附录 H。 □

然而，我们不能指望以同样的方式推广定理 5.9。严格来说，现在我们的目标是得到

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right] \leq \emptyset\left(T^a\left(\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1}\right]\right)^b\right) \tag{41}$$

对于某个 $a+b<1$。回想一下，为了得到定理 G.1 中的最后一个结论，我们只需应用詹森不等式，得到 $\mathbb{E}\left[\sqrt{\sum\|\mathbf{Q}_t\|_2^2}\right] \leq \sqrt{\mathbb{E}\left[\sum\|\mathbf{Q}_t\|_2^2\right]}$。但现在我们希望式 (41) 在任意调度策略下都成立；如果同样应用詹森不等式，问题就会归结为无条件地界定 $\mathbb{E}\left[\sum\|\mathbf{Q}_t\|_2^2\right]$ 无条件地，这是不可能的，因为在无法稳定系统的平凡调度策略下，我们可能有 $\|\mathbf{Q}_t\|_2^2=\Theta(t)$，因此 $\mathbb{E}\left[\sum\|\mathbf{Q}_t\|_2^2\right]$ 是 $\Theta(T^2)$。

若不交换平方根与期望的顺序，似乎很难直接得到 $\sqrt{\sum\|\mathbf{Q}_t\|_2^2}$ 的期望界。不过，我们仍可以用一种较为直接的方法应用定理 5.9。令

$$L \triangleq \max_{1\leq t\leq T, i\in[K]} |Q_{t,i}-Q_{t-1,i}|,$$

也就是说，L 是该样本路径上的最大队列长度增量；因此，它是一个随机变量。于是，我们可以写成

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right] &\leq \mathbb{E}\left[\mathbb{1}\left[L>C\right]\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right] + \mathbb{E}\left[\mathbb{1}\left[L\leq C\right]\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left[L>C\right]\sqrt{KT^3L^2}\right] + \mathbb{E}\left[\mathbb{1}\left[L\leq C\right]\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right] \\
&\leq \sqrt{K}T^{\frac{3}{2}}\mathbb{E}\left[\mathbb{1}\left[L>C\right]L\right] + \mathbb{E}\left[\mathbb{1}\left[L\leq C\right]\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right].
\end{aligned}$$

对于 $\mathbb{E}\left[\mathbb{1}\left[L>C\right]L\right]$ 项，回想一下，L 是 KT 个随机变量（它们是鞅差）的最大值，且这些变量的 α 阶矩有界；可以证明，L 的 α 阶矩不超过 $2KTM^\alpha$，并且 $\mathbb{E}\left[\mathbb{1}\left[L>C\right]L\right] \leq 2KTM^\alpha C^{1-\alpha}$。至于另一项，我们可以应用定理 5.9，得到样本路径界 $\mathbb{1}\left[L\leq C\right]\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2} \leq \emptyset(C^{\frac{1}{4}}(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1)^{\frac{3}{4}})$。因此，当 α 充分大时，我们可以将 C 取为 T 的某个幂，从而得出 $\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right] \leq \bar{\emptyset}\left(T^a\left(\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1}\right]\right)^b\right) + T^c$，其中

$0<a,b,c<1$ and $a+b<1$.

定理 *G.1* 的证明。根据队列长度增量的有界 α 阶矩假设，我们还可以得到队列长度增量本身的一个界：

$$\begin{aligned}
\mathbb{E}\left[|Q_{t,i}-Q_{t-1,i}||\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\left(|Q_{t,i}-Q_{t-1,i}|^\alpha\right)^{\frac{1}{\alpha}}\middle|\mathcal{F}_{t-1}\right] \\
&\overset{(a)}{\leq} \left(\mathbb{E}\left[|Q_{t,i}-Q_{t-1,i}|^\alpha|\mathcal{F}_{t-1}\right]\right)^{\frac{1}{\alpha}} \\
&\leq \left(M^\alpha\right)^{\frac{1}{\alpha}} \\
&= M,
\end{aligned}$$

其中，(a) 是由于 $x\mapsto x^{\frac{1}{\alpha}}$ 的凹性。类似地，我们有

$$\mathbb{E}\left[\left(Q_{t,i}-Q_{t-1,i}\right)^2\middle|\mathcal{F}_{t-1}\right] \leq M^2.$$

那么，对于任意的 $t\geq T_0$ 和 $i\in[K]$，我们可以写成

$$\mathbb{E}\left[Q_{t,i}|\mathcal{F}_{T_0}\right] \leq Q_{T_0,i} + \mathbb{E}\left[\sum_{s=T_0}^{t-1}\mathbb{E}\left[|Q_{s+1,i}-Q_{s,i}||\mathcal{F}_s\right]\middle|\mathcal{F}_{T_0}\right]$$

$$\leq Q_{T_0,i} + (t - T_0)M.$$

如果 $T_0 \leq t \leq T_0 + m$，那么我们有 $\mathbb{E}\left[Q_{t,i} \middle| \mathcal{F}_{T_0}\right] \leq Q_{T_0,\ i} + Mm$，因此

$$\begin{aligned}
\sum_{t=T_0}^{T_0+m-1} \mathbb{E}\left[\|Q_t\|_1 \middle| \mathcal{F}_{T_0}\right] &\leq \sum_{t=T_0}^{T_0+m-1} \sum_{i\in[K]} (Q_{T_0,i} + Mm) \\
&= m\|Q_{T_0}\|_1 + KMm^2 \\
&\overset{(a)}{\leq} m \cdot 2KMm + KMm^2 \\
&\leq 3KMm^2,
\end{aligned}$$

其中，步骤（a）源于我们的选择 $m = \left\lceil \frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M} \right\rceil$。

类似地，对于任意 $T_0 \leq t \leq T_0 + m$ 和 $i \in [K]$，我们有

$$\begin{aligned}
\mathbb{E}\left[Q_{t,i} \middle| \mathcal{F}_{T_0}\right] &\geq Q_{T_0,i} - \mathbb{E}\left[\sum_{s=T_0}^{t-1} \mathbb{E}\left[|Q_{s+1,i} - Q_{s,i}|\middle| \mathcal{F}_s\right] \middle| \mathcal{F}_{T_0}\right] \\
&\geq Q_{T_0,i} - (t - T_0)M \\
&= Q_{T_0,i} - Mm,
\end{aligned}$$

因此，记 $i^* = \arg\max_{i\in[K]} Q_{T_0,i}$，我们有

$$\begin{aligned}
\sum_{t=T_0}^{T_0+m-1} \mathbb{E}\left[\|Q_t\|_1 \middle| \mathcal{F}_{T_0}\right] &\geq \sum_{t=T_0}^{T_0+m-1} Q_{t,i^*} \\
&\geq \sum_{t=T_0+1}^{T_0+m} (Q_{T_0,i^*} - Mm) \\
&= m\|Q_{T_0}\|_\infty - Mm^2 \\
&\geq m \cdot (2Mm - 2M) - Mm^2 \\
&= Mm^2 - 2Mm \\
&\geq \frac{1}{2}Mm^2,
\end{aligned}$$

其中，最后一步利用了假设 $\|\mathbf{Q}_{T_0}\|_\infty \geq 8M$，这意味着 $m \geq 4$。因此，我们得出结论：

$$\frac{M}{2}m^2 \leq \mathbb{E}\left[\sum_{t=1}^{m} \|\mathbf{Q}_{T_0+t-1}\|_1 \middle| \mathcal{F}_{T_0}\right] \leq 3KMm^2.$$

对于 $\sum\|\mathbf{Q}_t\|_2^2$的界，我们有

$$\begin{aligned}
&\mathbb{E}\left[Q_{t,i}^2 \middle| \mathcal{F}_{T_0}\right] \\
&\leq Q_{T_0,i}^2 + \mathbb{E}\left[\sum_{s=T_0}^{t-1} \mathbb{E}\left[\left(Q_{s+1,i} - Q_{s,i}\right)^2 \middle| \mathcal{F}_s\right] + 2\sum_{T_0\leq s<s'<t} \mathbb{E}\left[|Q_{s+1,i} - Q_{s,i}|\middle| \mathcal{F}_s\right] \cdot \mathbb{E}\left[|Q_{s'+1,i} - Q_{s',i}|\middle| \mathcal{F}_{s'}\right] \middle| \mathcal{F}_{T_0}\right] \\
&\leq Q_{T_0,i}^2 + (t - T_0)M^2 + 2\binom{t - T_0}{2}M^2 \\
&= Q_{T_0,i}^2 + (t - T_0)^2M^2 \\
&\leq Q_{T_0,i}^2 + M^2m^2
\end{aligned}$$

对所有 $T_0 \leq t \leq T_0 + m$ 均成立。因此，我们得到

$$\sum_{t=T_0}^{T_0+m-1} \mathbb{E}\left[\|Q_t\|_2^2 \middle| \mathcal{F}_{T_0}\right] \leq \sum_{t=T_0}^{T_0+m-1} \sum_{i\in[K]} \left(Q_{T_0,i}^2 + M^2m^2\right)$$

$$
\begin{aligned}
&= m\|Q_{T_0}\|_2^2 + KM^2m^3 \\
&\overset{(a)}{\leq} m \cdot K \cdot (2Mm)^2 + KM^2m^3 \\
&< 5KM^2m^3,
\end{aligned}
$$

其中，步骤（a）是由我们选择 $m = \left\lceil \frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M} \right\rceil$ 得出的。对于另一个方向，我们有

$$
\begin{aligned}
&\mathbb{E}\left[Q_{t,i}^2 \middle| \mathcal{F}_{T_0}\right] \\
&\geq Q_{T_0,i}^2 - \mathbb{E}\left[\sum_{s=T_0}^{t-1} \mathbb{E}\left[\left(Q_{s+1,i} - Q_{s,i}\right)^2 \middle| \mathcal{F}_s\right] - 2\sum_{T_0 \leq s < s' < t} \mathbb{E}\left[|Q_{s+1,i} - Q_{s,i}| \,|\, \mathcal{F}_s\right] \cdot \mathbb{E}\left[|Q_{s'+1,i} - Q_{s',i}| \,|\, \mathcal{F}_{s'}\right] \middle| \mathcal{F}_{T_0}\right] \\
&\geq Q_{T_0,i}^2 - (t-T_0)M^2 - 2\binom{t-T_0}{2}M^2 \\
&= Q_{T_0,i}^2 - (t-T_0)^2M^2 \\
&\geq Q_{T_0,i}^2 - M^2m^2.
\end{aligned}
$$

and thus, noting $i^* = \operatorname{argmax}_{i\in[K]} Q_{T_0,i}$, we have

$$
\begin{aligned}
\sum_{t=T_0}^{T_0+m-1} \mathbb{E}\left[\|Q_t\|_2^2 \middle| \mathcal{F}_{T_0}\right] &\geq \sum_{t=T_0}^{T_0+m-1} \left(Q_{T_0,i^*}^2 - M^2m^2\right) \\
&= m\|Q_{T_0}\|_\infty^2 - M^2m^3 \\
&\overset{(a)}{\geq} m \cdot (2Mm - 2M)^2 - M^2m^3 \\
&\overset{(b)}{\geq} m \cdot \left(\frac{3}{2}Mm\right)^2 - M^2m^3 \\
&\geq M^2m^3
\end{aligned}
$$

其中，步骤（a）源于我们的选择 $m = \left\lceil \frac{\|\mathbf{Q}_{T_0}\|_\infty}{2M} \right\rceil$，步骤（$b$）使用了 $m \geq 4$ 的假设，因此 $2M \leq \frac{1}{2}Mm$。所以，

$$
M^2m^3 \leq \mathbb{E}\left[\sum_{t=1}^{m} \|\mathbf{Q}_{T_0+t-1}\|_2^2 \middle| \mathcal{F}_{T_0}\right] \leq 5KM^2m^3.
$$

□

H 算法 4 的详细分析

算法 4 的分析与算法2类似。首先，我们需要验证所选的探索率 γ_t's 确实保证不会超过 $\frac{1}{2}$。

命题 H.1。 对于所有 $t \geq 1$，我们有 $\gamma_t \leq \frac{1}{2}$–在算法 *4* 中。

证明。 注意到

$$
\begin{aligned}
\gamma_t &= t^{\frac{\delta}{4}} M L_{t-1}^{-1} \|\mathbf{Q}_{t-1}\|_1 \left(t^{-(\frac{1}{4}-\frac{\delta}{2})} \sqrt{86L_{t-1}^2K^6t^{\frac{3}{2}} + \sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1} \\
&\leq t^{\frac{\delta}{4}} \|\mathbf{Q}_{t-1}\|_1 \left(t^{-(\frac{1}{4}-\frac{\delta}{2})} \sqrt{86L_{t-1}^2K^6t^{\frac{3}{2}} + \sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1},
\end{aligned}
$$

因为对所有 $t \geq 0$，$L_t \geq M$。因此，只需验证

$$
2t^{\frac{\delta}{4}} \|\mathbf{Q}_{t-1}\|_1 \leq t^{-(\frac{1}{4}-\frac{\delta}{2})} \sqrt{86L_{t-1}^2K^6t^{\frac{3}{2}} + \sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2},
$$

$$
\begin{aligned}
&\Updownarrow\\
4\|\mathbf{Q}_{t-1}\|_1^2 &\le t^{-\frac{1}{2}+\frac{\delta}{2}}\left(86L_{t-1}^2K^6t^{\frac{3}{2}}+\sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2\right),\\
&\Uparrow (a)\\
4\|\mathbf{Q}_{t-1}\|_1^2 &\le t^{-\frac{1}{2}+\frac{\delta}{2}}\left(86L_{t-1}^2K^6t^{\frac{3}{2}}+\frac{1}{3L_{t-1}K^3}\|\mathbf{Q}_{t-1}\|_1^3\right)\\
&\Updownarrow\\
4\|\mathbf{Q}_{t-1}\|_1^2 &\le t^{-\frac{1}{2}+\frac{\delta}{2}}\left(\frac{1}{3}\cdot 258L_{t-1}^2K^6t^{\frac{3}{2}}+\frac{2}{3}\cdot\frac{1}{2L_{t-1}K^3}\|\mathbf{Q}_{t-1}\|_1^3\right),\\
&\Uparrow (b)\\
4\|\mathbf{Q}_{t-1}\|_1^2 &\le t^{-\frac{1}{2}+\frac{\delta}{2}}\cdot 258^{\frac{1}{3}}2^{-\frac{2}{3}}t^{\frac{1}{2}}\|\mathbf{Q}_{t-1}\|_1^2\\
&\Updownarrow\\
4 &\le t^{\frac{\delta}{2}}\left(\frac{258}{4}\right)^{\frac{1}{3}},
\end{aligned}
$$

而最后一个结论显然成立。这里在步骤（a）中，我们应用定理 E.2（只要将 M替换为 L_{t-1}，该定理即可适用）；在步骤（b）中，我们应用算术-几何平均不等式 $\frac{1}{3}x+\frac{2}{3}y\ge x^{\frac{1}{3}}y^{\frac{2}{3}}$。 □

在验证了 $\gamma_t\ge\frac{1}{2}$ 后，还请注意，我们在算法 4 中选择了 $\gamma_t=Mt^{\frac{\delta}{4}}\eta_t\|\mathbf{Q}_{t-1}\|_1$，这使得 EXP3.S+能够处理一个新的反馈 g_{t,a_t}，其大小最大可达 $Q_{t-1,a_t}\cdot Mt^{\frac{\delta}{4}}$。与算法2相比，在算法 4 中，我们将新的服务 S_{t,a_t}截断至 $Mt^{\frac{\delta}{4}}$，得到 S'_t，并将 $Q_{t-1,a_t}S'_t$ 输入 EXP3.S+。因此，定理 5.1 现在适用，我们可以得到

引理 H.2。 假设假设 *1*、*2* 和 *4* 成立，则运行算法 *4* 可保证

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbf{Q}_{t-1},\mathbf{S}'_t\odot\vec{\theta}_t\rangle-Q_{t-1,a_t}S'_{t,a_t}\right]\\
&\le\mathbb{E}\left[L_{T-1}(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(4\ln T+\ln K)\sqrt{86L_{T-1}^2K^6T^{\frac{3}{2}}+\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]\\
&\quad+\mathbb{E}\left[8MT^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{1+\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}+4ML_{T-1}\right]
\end{aligned}
\tag{42}
$$

对于任意时间范围长度 $T\ge 1$。其中 $S_{t,i}$ 定义为

$$
S'_{t,i}\triangleq\begin{cases}S_{t,i} & \textit{if } S_{t,i}\le Mt^{\frac{\delta}{4}}\\ 0 & \textit{otherwise}\end{cases},
$$

且 $\{\theta_t\}$ 是假设 *1* 和 *2* 中的参考策略。

证明。 对于每个 $\vec{\theta}_t$，我们都可以找到一个 $\vec{\theta}'_t\in\Delta^{[K],\beta_t}$，使得 $\|\vec{\theta}_t-\vec{\theta}'_t\|_1\le 2K\beta_t$且 $\|\vec{\theta}'_s-\vec{\theta}'_t\|_1\le\|\vec{\theta}_s-\vec{\theta}_t\|_1$。例如，我们可以选择

$$
\vec{\theta}'_t=(1-\beta_t)\vec{\theta}_t+\beta_t\mathbf{1}.
$$

然后，我们可以写成

$$
\begin{aligned}
\langle\mathbf{Q}_{t-1},\mathbf{S}'_t\odot\vec{\theta}_t\rangle&=\langle\mathbf{Q}_{t-1},\mathbf{S}'_t\odot\vec{\theta}'_t\rangle+\langle\mathbf{Q}_{t-1}\odot\mathbf{S}'_t,\vec{\theta}_t-\vec{\theta}'_t\rangle\\
&\le\langle\mathbf{Q}_{t-1},\mathbf{S}'_t\odot\vec{\theta}'_t\rangle+\|\mathbf{Q}_{t-1}\odot\mathbf{S}'_t\|_\infty\cdot\|\vec{\theta}_t-\vec{\theta}'_t\|_1\\
&\le\langle\mathbf{Q}_{t-1},\mathbf{S}'_t\odot\vec{\theta}'_t\rangle+L_{t-1}t\cdot Mt^{\frac{\delta}{4}}\cdot 2K\beta_t\\
&=\langle\mathbf{Q}_{t-1},\mathbf{S}'_t\odot\vec{\theta}'_t\rangle+2L_{t-1}Mt^{\frac{\delta}{4}-3}
\end{aligned}
$$

$$\leq \langle \mathbf{Q}_{t-1}, \mathbf{S}'_t \odot \vec{\theta}'_t \rangle + 2L_{t-1}Mt^{-2}.$$

然后，可以看出该量

$$\sum_{t=1}^{T} \mathbb{E}\left[\langle \mathbf{Q}_{t-1}, \mathbf{S}'_t \odot \vec{\theta}'_t \rangle - Q_{t-1,a_t} S'_{t,a_t}\right] = \sum_{t=1}^{T} \mathbb{E}\left[\langle \mathbf{Q}_{t-1} \odot \mathbf{S}'_t, \vec{\theta}'_t \rangle - (\mathbf{Q}_{t-1} \odot \mathbf{S}'_t)_{a_t}\right]$$

$\vec{}$满足应用定理 5.1 的条件。另一方面，针对 $\{\theta_t\}$ 的遗憾与针对 $\{\vec{\theta}'_t\}$ 的遗憾之间的总差值不超过

$$\begin{aligned}\sum_{t=1}^{T} 2L_{t-1}Mt^{-2} &\leq 2ML_{T-1}\sum_{t=1}^{T} t^{-2} \\ &\leq \frac{\pi^2}{3} ML_{T-1}.\end{aligned}$$

定理 5.1 断言

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}\left[\langle \mathbf{Q}_{t-1}, \mathbf{S}'_t \odot \vec{\theta}'_t \rangle - Q_{t-1,a_t} S'_{t,a_t}\right] \\ \leq &\left(1 + \sum_{t=1}^{T-1} \|\vec{\theta}'_{t+1} - \vec{\theta}'_t\|_1\right) \mathbb{E}\left[\eta_T^{-1} \ln \frac{1}{\beta_T}\right] + e\mathbb{E}\left[\sum_{t=1}^{T} \eta_t \|\mathbf{Q}_{t-1} \odot \mathbf{S}'_t\|_2^2\right] \\ &+ \mathbb{E}\left[\sum_{t=1}^{T} \gamma_t \left\langle \mathbf{Q}_{t-1} \odot \mathbf{S}'_t, \frac{\mathbf{Q}_{t-1}}{\|\mathbf{Q}_{t-1}\|_1} \right\rangle\right]. \end{aligned} \tag{43}$$

下面，我们对公式（43）右端的各项进行界定。首先，根据假设 2，有 $\sum_{t=1}^{T-1} \|\vec{\theta}'_{t+1} - \vec{\theta}'_t\|_1 \leq \sum_{t=1}^{T-1} \|\vec{\theta}_{t+1} - \vec{\theta}_t\|_1 \leq C_V T^{\frac{1}{2}-\delta}$，因此

$$\begin{aligned}&\left(1 + \sum_{t=1}^{T-1} \|\vec{\theta}'_{t+1} - \vec{\theta}'_t\|_1\right) \mathbb{E}\left[\eta_T^{-1} \ln \frac{1}{\beta_T}\right] \\ \leq &\left(1 + C_V T^{\frac{1}{2}-\delta}\right) \mathbb{E}\left[\eta_T^{-1} \ln \frac{1}{\beta_T}\right] \\ \leq &\left(1 + C_V T^{\frac{1}{2}-\delta}\right) (4\ln T + \ln K)\, \mathbb{E}\left[T^{-\left(\frac{1}{4}-\frac{\delta}{2}\right)} L_{T-1} \sqrt{86L_{T-1}^2 K^6 T^{\frac{3}{2}} + \sum_{s=0}^{T-1} \|\mathbf{Q}_s\|_2^2}\right] \\ \leq &\left(1 + C_V\right) T^{\frac{1}{4}-\frac{\delta}{2}} (4\ln T + \ln K)\, L_{T-1} \mathbb{E}\left[\sqrt{86L_{T-1}^2 K^6 T^{\frac{3}{2}} + \sum_{s=0}^{T-1} \|\mathbf{Q}_s\|_2^2}\right].\end{aligned}$$

对于第二项，我们有

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}\left[\eta_t \|\mathbf{Q}_{t-1} \odot \mathbf{S}'_t\|_2^2 \middle| \mathcal{F}_{t-1}\right] \\ = &\sum_{t=1}^{T} \eta_t \sum_{i \in [K]} Q_{t-i,i}^2 \mathbb{E}\left[S_{t,i}'^2 \middle| \mathcal{F}_{t-1}\right] \\ \overset{(a)}{\leq} &M^2 \sum_{t=1}^{T} \eta_t \|\mathbf{Q}_{t-1}\|_2^2 \\ = &M \sum_{t=1}^{T} ML_{t-1}^{-1} t^{\frac{1}{4}-\frac{\delta}{2}} \left(\sqrt{86L_{t-1}^2 K^6 t^{\frac{3}{2}} + \sum_{s=0}^{t-1} \|\mathbf{Q}_s\|_2^2}\right)^{-1} \cdot \|\mathbf{Q}_{t-1}\|_2^2\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} M\sum_{t=1}^{T} t^{\frac{1}{4}-\frac{\delta}{2}}\left(\sqrt{1+\sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}\cdot\|\mathbf{Q}_{t-1}\|_2^2\\
&\leq MT^{\frac{1}{4}-\frac{\delta}{2}}\sum_{t=1}^{T}\left(\sqrt{1+\sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}\cdot\|\mathbf{Q}_{t-1}\|_2^2\\
&\overset{(c)}{\leq} 2MT^{\frac{1}{4}-\frac{\delta}{2}}\sqrt{1+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2},
\end{aligned}
$$

其中，在步骤（a）中，我们对 $S_{t,i}$ 使用二阶矩界，而不是截断阈值 $Mt^{\frac{\delta}{4}}$。步骤（b）是因为对所有 t，$L_t \geq M$。在步骤（c）中，我们利用了这一事实：

$$
\sum_{i=1}^{n}\frac{x_i}{\sqrt{1+\sum_{j=1}^{i}x_j}} \leq 2\sqrt{1+\sum_{i=1}^{n}x_i}
$$

对于非负的 $x_1\ldots,\ x_n$。

第三项可以用与第二项非常类似的方法进行界定，具体而言，我们有

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\gamma_t\left\langle\mathbf{Q}_{t-1}\odot\mathbf{S}'_t,\frac{\mathbf{Q}_{t-1}}{\|\mathbf{Q}_{t-1}\|_1}\right\rangle\middle|\mathcal{F}_{t-1}\right]\\
&\overset{(a)}{\leq} M\sum_{t=1}^{T}\gamma_t\|\mathbf{Q}_{t-1}\|_1^{-1}\|\mathbf{Q}_{t-1}\|_2^2\\
&\leq M\sum_{t=1}^{T}ML_{t-1}^{-1}t^{\frac{\delta}{4}}\cdot t^{\frac{1}{4}-\frac{\delta}{2}}\left(\sqrt{1+\sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}\cdot\|\mathbf{Q}_{t-1}\|_2^2\\
&\leq M\sum_{t=1}^{T}t^{\frac{1}{4}-\frac{\delta}{4}}\left(\sqrt{1+\sum_{s=0}^{t-1}\|\mathbf{Q}_s\|_2^2}\right)^{-1}\cdot\|\mathbf{Q}_{t-1}\|_2^2\\
&\leq 2MT^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{1+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2},
\end{aligned}
$$

其中，在步骤（a）中，我们对 $S_{t,\ i}$ 应用了一阶矩界。

将所有内容合并后再取期望，可以得到

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbf{Q}_{t-1},\mathbf{S}'_t\odot\vec{\theta}'_t\rangle-Q_{t-1,a_t}S'_{t,a_t}\right]\\
&\leq(1+C_V)\,T^{\frac{1}{4}-\frac{\delta}{2}}\,(4\ln T+\ln K)\,L_{T-1}\mathbb{E}\left[\sqrt{86L_{T-1}^2K^6T^{\frac{3}{2}}+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right]\\
&\quad+2MT^{\frac{1}{4}-\frac{\delta}{2}}\mathbb{E}\left[\sqrt{1+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right]+2eMT^{\frac{1}{4}-\frac{\delta}{4}}\mathbb{E}\left[\sqrt{1+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right]+\frac{\pi^2}{3}M\mathbb{E}\left[L_{T-1}\right]\\
&\leq(1+C_V)\,T^{\frac{1}{4}-\frac{\delta}{2}}\,(4\ln T+\ln K)\,L_{T-1}\mathbb{E}\left[\sqrt{86L_{T-1}^2K^6T^{\frac{3}{2}}+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right]\\
&\quad+8MT^{\frac{1}{4}-\frac{\delta}{4}}\mathbb{E}\left[\sqrt{1+\sum_{s=0}^{T-1}\|\mathbf{Q}_s\|_2^2}\right]+4M\mathbb{E}\left[L_{T-1}\right].
\end{aligned}
$$

□

接下来，我们可以将定理 H.2 中关于（$\mathbf{Q}_{t-1}\cdot\mathbf{S}'_t$）目标遗憾的上界转化为（$\mathbf{Q}_{t-1}\cdot\mathbf{S}_t$）目标遗憾的上界：

引理 H.3。 假设假设 *1*、*2* 和 *4* 成立，则运行算法 *4* 可保证

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}_t\rangle-Q_{t-1,a_t}S_{t,a_t}\right]\\
&\le\mathbb{E}\left[L_{T-1}(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(4\ln T+\ln K)\sqrt{86L_{T-1}^2K^6T^{\frac{3}{2}}+\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]\\
&\quad+\mathbb{E}\left[8MT^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{1+\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}+4ML_{T-1}\right]+o(T)
\end{aligned}\tag{44}$$

对于任意时间范围长度 $T\ge 1$。此处，$\{\theta_t\}$ 是假设 *1* 和 *2* 中的参考策略 $\vec{\ }$.

证明。 注意，对于每个 t，单步遗憾 $\langle\mathbf{Q}_{t-1},\ \mathbf{S}_t\odot\theta_t\rangle-Q_{t-1,\ a_t}S_{t,\ a_t}$ 可以被上界限定为

$$\langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}_t\rangle-Q_{t-1,a_t}S_{t,a_t}\le\underbrace{\langle\mathbf{Q}_{t-1},\mathbf{S}'_t\odot\vec{\theta}_t\rangle-Q_{t-1,a_t}S'_{t,a_t}}_{(\mathbf{Q}_{t-1}\cdot\mathbf{S}'_t)\text{-objective single step regret}}+\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]\langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}_t\rangle$$

其中

$$S'_{t,i}\triangleq\begin{cases}S_{t,i} & \text{if } S_{t,i}\le Mt^{\frac{\delta}{4}}\\ 0 & \text{otherwise}\end{cases}.$$

因此，只需对 $\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]\langle\mathbf{Q}_{t-1},\ \mathbf{S}_t\odot\vec{\theta}_t\rangle\right]$ 求上界。事实上，

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]\langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}_t\rangle\middle|\mathcal{F}_{t-1}\right]\\
&\le\|\mathbf{Q}_{t-1}\|_\infty\cdot\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]\|\mathbf{S}_t\|_1\right]\\
&=\|\mathbf{Q}_{t-1}\|_\infty\left(\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]\sum_{i\in[K]:i\ne a_t}S_{t,i}\right]+\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]S_{t,a_t}\right]\right).
\end{aligned}$$

对于第一项 $\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]\sum_{i\in[K]:\ i\ne a_t}S_{t,\ i}\right]$，我们可以写成

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]\sum_{i\in[K]:i\ne a_t}S_{t,i}\right]&=\mathbb{E}\left[\mathbb{P}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\middle|a_t\right]\cdot\mathbb{E}\left[\sum_{i\in[K]:i\ne a_t}S_{t,i}\middle|a_t\right]\right]\\
&=\mathbb{E}\left[\mathbb{P}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\middle|a_t\right]\cdot\sum_{i\in[K]:i\ne a_t}\sigma_{t,i}\right]\\
&\le KM\cdot\mathbb{P}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]\\
&\le KM\cdot M^{\alpha}\cdot(Mt^{\frac{\delta}{4}})^{-\alpha}\\
&=KMt^{-\frac{\delta\alpha}{4}},
\end{aligned}$$

其中最后一个不等式是根据 S_{t,a_t} 的 α 阶矩上界以及切比雪夫不等式得出的。对于另一项 $\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]S_{t,a_t}\right]$，我们有

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t}>Mt^{\frac{\delta}{4}}\right]S_{t,a_t}\right]&\le\mathbb{E}\left[S_{t,a_t}^{\alpha}\cdot(Mt^{\frac{\delta}{4}})^{-(\alpha-1)}\right]\\
&\le Mt^{-\frac{\delta(\alpha-1)}{4}}.
\end{aligned}$$

因此,

$$\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t} > Mt^{\frac{\delta}{4}}\right]\langle \mathbf{Q}_{t-1}, \mathbf{S}_t \odot \vec{\theta}_t\rangle \middle| \mathcal{F}_{t-1}\right] \le KMt^{-\frac{\delta(\alpha-1)}{4}} \cdot \|\mathbf{Q}_{t-1}\|_\infty,$$

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}\left[S_{t,a_t} > Mt^{\frac{\delta}{4}}\right]\langle \mathbf{Q}_{t-1}, \mathbf{S}_t \odot \vec{\theta}_t\rangle\right] \\
&\le KM\sum_{t=1}^{T} t^{-\frac{\delta(\alpha-1)}{4}}\mathbb{E}\left[\|\mathbf{Q}_{t-1}\|_\infty\right] \\
&\le KM\sum_{t=1}^{T} t^{1-\frac{\delta(\alpha-1)}{4}}\mathbb{E}\left[L_t\right] \\
&\le KM\sum_{t=1}^{T} t^{1-\frac{\delta(\alpha-1)}{4}}\mathbb{E}\left[L_t^\alpha\right]^{\frac{1}{\alpha}} \\
&\le 2K^{1+\frac{1}{\alpha}}M^2\sum_{t=1}^{T} t^{1-\frac{\delta(\alpha-1)}{4}}\cdot t^{\frac{1}{\alpha}} \\
&\le o(T)
\end{aligned}$$

其中最后一步是由 $\delta\cdot\alpha > 7$ 的假设得出的。 □

与定理 5.7 相比，定理 H.3 中界的右端现在涉及 L_{T-1}，即样本路径最大队列长度增量。接下来，我们将进一步计算以消去这些 L_{T-1} 因子。

回顾一下， L_t 定义为

$$L_t \triangleq \max\left\{M, \max_{1\le s\le t, i\in[K]}|Q_{s,i} - Q_{s-1,i}|\right\}$$

在算法 4 中，有界矩队列长度增量假设为 L_t 的 α 阶矩给出了一个上界

$$\begin{aligned}
\mathbb{E}\left[L_t^\alpha\right] &\le \max\left\{M^\alpha, \sum_{s=1}^{t}\mathbb{E}\left[|Q_{s,i} - Q_{s-1,i}|^\alpha\right]\right\} \\
&\le \max\left\{M^\alpha, 2KTM^\alpha\right\} \\
&= 2KTM^\alpha.
\end{aligned}$$

这个矩界 对于 L_t，使我们能够处理定理 H.3 中的 L_{T-1} 因子。例如，我们有

$$\mathbb{E}\left[L_{T-1}\right] \le \mathbb{E}\left[L_{T-1}^\alpha\right]^{\frac{1}{\alpha}} \le 2K^{\frac{1}{\alpha}}MT^{\frac{1}{\alpha}} = o(T).$$

为界定 $\mathbb{E}\left[L_{T-1}(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(4\ln T + \ln K)\sqrt{86L_{T-1}^2K^6T^{\frac{3}{2}} + \sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]$ 项，我们首先直接写成

$$\begin{aligned}
&\mathbb{E}\left[L_{T-1}(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(4\ln T + \ln K)\sqrt{86L_{T-1}^2K^6T^{\frac{3}{2}} + \sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right] \\
&\le \mathbb{E}\left[L_{T-1}(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(4\ln T + \ln K)\cdot\left(\sqrt{86L_{T-1}^2K^6T^{\frac{3}{2}}} + \sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right)\right] \\
&= \mathbb{E}\left[K^3(1+C_V)T^{1-\frac{\delta}{2}}(4\ln T + \ln K)L_{T-1}^2\right] + \mathbb{E}\left[(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(4\ln T + \ln K)L_{T-1}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]
\end{aligned}$$

然后逐一界定这两项。L_{T-1}^2 项很容易：

$$\begin{aligned}\mathbb{E}\left[K^3(1+C_V)T^{1-\frac{\delta}{2}}(4\ln T+\ln K)L_{T-1}^2\right] &\le K^3(1+C_V)T^{1-\frac{\delta}{2}}(4\ln T+\ln K)\mathbb{E}\left[L_{T-1}^{\alpha}\right]^{\frac{2}{\alpha}}\\ &\le K^3(1+C_V)T^{1-\frac{\delta}{2}}(4\ln T+\ln K)\cdot 2K^{\frac{2}{\alpha}}M^2T^{\frac{2}{\alpha}}\\ &\le \emptyset(T^{1-\frac{\delta}{2}+\frac{2}{\alpha}})\\ &\le o(T)\end{aligned}$$

其中，最后一步是基于假设 $\delta\cdot\alpha>7$ 得出的。对于另一项，我们可以写成

$$\begin{aligned}&L_{T-1}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\\ &\le \mathbb{1}\left[L_{T-1}\le MT^{\frac{\delta}{4}}\right]L_{T-1}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}+\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]L_{T-1}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}.\end{aligned}\tag{45}$$

对于式（45）中含因子 $\mathbb{1}\left[L_{T-1}\le MT^{\frac{\delta}{4}}\right]$ 的项，我们可以应用定理 5.9，得到

$$\begin{aligned}\mathbb{1}\left[L_{T-1}\le MT^{\frac{\delta}{4}}\right]L_{T-1}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2} &\le \mathbb{1}\left[L_{T-1}\le MT^{\frac{\delta}{4}}\right]2L_{T-1}^{\frac{5}{4}}\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{4}}\\ &\le 2M^{\frac{5}{4}}T^{\frac{5}{16}\delta}\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{4}}.\end{aligned}$$

对于另一项含因子 $\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]$ 的项，我们可以写成

$$\begin{aligned}\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]L_{T-1}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2} &\le \mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]L_{T-1}\sqrt{\sum_{t=1}^{T}K(L_{T-1}T)^2}\\ &\le \sqrt{K}T^{\frac{3}{2}}\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]L_{T-1}^2,\end{aligned}$$

此外，我们可以将其期望值上界估计为

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]L_{T-1}^2\right] &\le \mathbb{E}\left[L_{T-1}^{\alpha}/(MT^{\frac{\delta}{4}})^{\alpha-2}\right]\\ &\le 2KTM^{\alpha}\cdot(MT^{\frac{\delta}{4}})^{-(\alpha-2)}\\ &\le 2KM^2T^{1-\frac{\alpha-2}{4}\delta},\end{aligned}$$

因此

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]L_{T-1}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right] &\le \sqrt{K}T^{\frac{3}{2}}\cdot 2KM^2T^{1-\frac{\alpha-2}{4}\delta}\\ &= 2K^{\frac{3}{2}}M^2T^{\frac{5}{2}-\frac{\alpha-2}{4}\delta}\\ &= 2K^{\frac{3}{2}}M^2T^{\frac{10+2\delta-\delta\alpha}{4}}\\ &\le 2K^{\frac{3}{2}}M^2T^{\frac{11-\delta\alpha}{4}}\\ &\le o(T)\end{aligned}$$

其中，最后一个不等式是由假设 $\delta\cdot\alpha>7$ 得出的。

综合所有项的界，我们可以得出

$$\mathbb{E}\left[(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(4\ln T+\ln K)L_{T-1}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]$$

$$\leq o(T)+\mathbb{E}\left[(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(4\ln T+\ln K)\cdot\mathbb{1}\left[L_{T-1}\leq MT^{\frac{\delta}{4}}\right]L_{T-1}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]$$

$$\leq o(T)+(1+C_V)T^{\frac{1}{4}-\frac{\delta}{2}}(4\ln T+\ln K)\cdot 2M^{\frac{5}{4}}T^{\frac{5}{16}\delta}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{4}}\right]$$

$$=o(T)+2(1+C_V)M^{\frac{5}{4}}T^{\frac{1}{4}-\frac{3}{16}\delta}(4\ln T+\ln K)\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{4}}\right].$$

接下来只需界定定理 H.2 中的 $\mathbb{E}\left[8MT^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{1+\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]$ 项。由于 $T^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{1+\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}+T^{\frac{1}{4}-\frac{\delta}{4}}\cdot\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}=o(T)+T^{\frac{1}{4}-\frac{\delta}{4}}\cdot\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}$，只需界定 $\mathbb{E}\left[MT^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]$。同样地，我们先写出

$$T^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\leq\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]T^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}+\mathbb{1}\left[L_{T-1}\leq MT^{\frac{\delta}{4}}\right]T^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}$$

并逐一对这两项的期望进行界定。对于含有 $\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]$ 因子的项，我们有

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]T^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]&\leq T^{\frac{1}{4}-\frac{\delta}{4}}\cdot\sqrt{K}T^{\frac{3}{2}}\mathbb{E}\left[\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]L_{T-1}\right]\\&\leq T^{\frac{1}{4}-\frac{\delta}{4}}\cdot\sqrt{K}T^{\frac{3}{2}}\cdot 2KTM\cdot T^{-(\alpha-1)\frac{\delta}{4}}\\&=2qK^{\frac{3}{2}}MT^{\frac{11}{4}-\frac{\alpha\delta}{4}}\\&\leq o(T)\end{aligned}$$

其中，最后一步是基于假设 $\delta\cdot\alpha>7$ 得出的。对于含有 $\mathbb{1}\left[L_{T-1}>MT^{\frac{\delta}{4}}\right]$ 因子的项，我们可以将其界定为

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}\left[L_{T-1}\leq MT^{\frac{\delta}{4}}\right]T^{\frac{1}{4}-\frac{\delta}{4}}\sqrt{\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_2^2}\right]\\&\leq T^{\frac{1}{4}-\frac{\delta}{4}}\cdot\mathbb{E}\left[\mathbb{1}\left[L_{T-1}\leq MT^{\frac{\delta}{4}}\right]2L_{T-1}^{\frac{1}{4}}\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{4}}\right]\\&\leq 2M^{\frac{1}{4}}T^{\frac{1}{4}-\frac{3}{16}\delta}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{4}}\right].\end{aligned}$$

因此，我们可以得出结论：

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbf{Q}_{t-1},\mathbf{S}_t\odot\vec{\theta}_t\rangle-Q_{t-1,a_t}S_{t,a_t}\right]\\&\leq o(T)+2(1+C_V)M^{\frac{5}{4}}T^{\frac{1}{4}-\frac{3}{16}\delta}(4\ln T+\ln K)\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{4}}\right]\\&\quad+16M^{\frac{5}{4}}T^{\frac{1}{4}-\frac{3}{16}\delta}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{4}}\right]\end{aligned}$$

$$\leq o(T)+18(1+C_V)M^{\frac{5}{4}}T^{\frac{1}{4}-\frac{3}{16}\delta}(4\ln T+\ln K)\mathbb{E}\left[\left(\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right)^{\frac{3}{4}}\right]$$

$$\leq o(T)+18(1+C_V)M^{\frac{5}{4}}T^{\frac{1}{4}-\frac{3}{16}\delta}(4\ln T+\ln K)\left(\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right]\right)^{\frac{3}{4}}. \tag{46}$$

将式 (46)代入定理 5.3 中的式 (3)，可得

命题 H.4。 在假设 *1*、*2* 和 *4* 下，如果 $\delta\cdot\alpha>7$，则运行算法 *4* 可保证

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right]\leq f(\mathcal{T}_T)+h(\mathcal{T}_T)\cdot\left(\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right]\right)^{\frac{3}{4}} \tag{47}$$

对于任意 $T\geq$最大值$\{\frac{4}{C_W},C_W\}$，其中 $\mathcal{T}_T$是一个不超过 $2T$ 的常数，

$$f(T)=\frac{(K+1)M^2+2C_W(KM^2+\epsilon KM)}{\epsilon}T+o(T),$$

and

$$h(T)=\frac{18(1+C_V)M^{\frac{5}{4}}T^{\frac{1}{4}-\frac{3}{16}\delta}(4\ln T+\ln K)}{\epsilon}=\widetilde{\mathcal{O}}(T^{\frac{1}{4}-\frac{3}{16}\delta}\epsilon^{-1}).$$

将定理 5.10 应用于式 (47)，可解得式 (47) 为

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right]\leq\left(h(\mathcal{T}_T)^{\frac{1}{4}}+f(\mathcal{T}_T)\right)^4$$
$$\leq\frac{(K+1)M^2+2C_W(KM^2+\epsilon KM)}{\epsilon}\mathcal{T}_T+o(\mathcal{T}_T),$$

thus

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right]\leq\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_T}\|\mathbf{Q}_{t-1}\|_1\right]\leq\frac{(K+1)M^2+2C_W(KM^2+\epsilon KM)}{\epsilon}\frac{\mathcal{T}_T}{T}+o(\mathcal{T}_T/T)$$
$$\leq\frac{2(K+1)M^2+4C_W(KM^2+\epsilon KM)}{\epsilon}+o(1)$$

正如所要求的。

I 算法 5 的详细分析

与算法 3 的分析类似，我们首先构建每个 EXP.3 时期的遗憾界引理（定理 6.2）的广义版本:

引理 I.1. 假设假设 *1*、*3* 和 *4* 成立，则令 T_0 为我们在算法 *5* 中启动一个长度为 m 的新 *EXP3.S+*实例的时间步，有

$$\mathbb{1}[T_0\text{ 结束一个 }\textit{EXP3}\text{ 实例，而新 }\textit{EXP3}\text{ 实例的长度为 }m,m\geq 2]$$
$$\cdot\sum_{t=1}^{m}\mathbb{E}\left[\langle\mathbf{Q}_{T_0+t-1}\odot\mathbf{S}_{T_0+t},\vec{\theta}_{T_0+t}\rangle-Q_{T_0+t-1,a_{T_0+t}}S_{T_0+t,a_{T_0+t}}\middle|\mathcal{F}_{T_0}\right]$$
$$\leq 21(1+C_V)M^3Km^{2-\frac{\delta}{3}}\cdot(3\ln m+\ln K)+4M^2. \tag{48}$$

证明。高 层次上的思路是应用定理 5.1，但我们需要先验证所有 γ_τ 都不超过1 2 – .

我们对 m 的选择保证 $m\geq{}^{\|}\mathbf{Q}_{2M}^{T_0}\|_\infty$, 因此 $\|\mathbf{Q}_{T_0}\|_\infty\leq 2Mm$，故

$$\gamma_\tau=\frac{1}{4}M^{-2}m^{-1-\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty$$

$$
\begin{aligned}
&\leq \frac{1}{4}M^{-2}m^{-1-\frac{\delta}{3}} \cdot 2Mm \\
&= \frac{1}{2}M^{-1}m^{-\frac{\delta}{3}} \\
&\leq \frac{1}{2}.
\end{aligned}
$$

对于任意 $1 \leq \tau \leq m$，定义

$$
g_{T_0+\tau,i} \triangleq \begin{cases} Q_{T_0+\tau-1,i}S_{T_0+\tau,i} & \text{if } Q_{T_0+\tau-1,i}S_{T_0+\tau,i} \leq m^{\frac{\delta}{3}}MQ_{T_0,i} \\ 0 & \text{otherwise} \end{cases}.
$$

下面，我们首先给出一个上界

$$
\mathbb{E}\left[\sum_{\tau=1}^{m}\left(\left\langle \mathbf{g}_{T_0+\tau}, \vec{\theta}_{T_0+\tau}\right\rangle - g_{T_0+\tau,a_{T_0+\tau}}\right)\middle| \mathcal{F}_{T_0}\right],
$$

即，$\mathbf{g}_{T_0+\tau}$-目标遗憾，然后推导原始$\mathbf{Q}_{T_0+\tau-1} \odot \mathbf{S}_{T_0+\tau}$-目标遗憾的一个上界。

对于每个 $\vec{\theta}_{T_0+\tau}$，我们都可以找到一个 $\vec{\theta}'_{T_0+\tau} \in \Delta^{[K],\ \beta_\tau}$使得 $\|\vec{\theta}_{T_0+\tau}-\vec{\theta}'_{T_0+\tau}\|_1 \leq 2K\beta_\tau$且对于任意 $1 \leq s,\ t \leq m$，都有 $\|\vec{\theta}'_s-\vec{\theta}'_t\|_1 \leq \|\vec{\theta}_{T_0+s}-\vec{\theta}_{T_0+t}\|_1$。例如，我们可以选择

$$
\vec{\theta}'_{T_0+\tau} = (1-\beta_\tau)\vec{\theta}_{T_0+\tau} + \beta_\tau \mathbf{1}.
$$

于是，我们可以写成

$$
\begin{aligned}
\langle \mathbf{g}_{T_0+\tau}, \vec{\theta}_{T_0+\tau}\rangle &= \langle \mathbf{g}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle + \langle \mathbf{g}_{T_0+\tau}, \vec{\theta}_{T_0+\tau} - \vec{\theta}'_{T_0+\tau}\rangle \\
&\leq \langle \mathbf{g}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle + \|\mathbf{g}_{T_0+\tau}\|_\infty \cdot \|\vec{\theta}_{T_0+\tau} - \vec{\theta}'_{T_0+\tau}\|_1 \\
&\leq \langle \mathbf{g}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle + m^{\frac{\delta}{3}}M\|\mathbf{Q}_{T_0}\|_\infty \cdot \|\vec{\theta}_{T_0+\tau} - \vec{\theta}'_{T_0+\tau}\|_1 \\
&\leq \langle \mathbf{g}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle + m^{\frac{\delta}{3}}M \cdot 2Mm \cdot \|\vec{\theta}_{T_0+\tau} - \vec{\theta}'_{T_0+\tau}\|_1 \\
&\leq \langle \mathbf{g}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle + 2M^2m^{1+\frac{\delta}{3}} \cdot 2K\beta_\tau \\
&= \langle \mathbf{g}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle + 4M^2m^{\frac{\delta}{3}-2} \\
&\leq \langle \mathbf{g}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle + 4M^2m^{-1}.
\end{aligned}
$$

于是，可以看出该量

$$
\sum_{\tau=1}^{m}\mathbb{E}\left[\langle \mathbf{g}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle - g_{T_0+\tau,a_{T_0+\tau}}\middle| \mathcal{F}_{T_0}\right]
$$

满足应用定理 5.1 的条件。定理 5.1 断言

$$
\begin{aligned}
&\sum_{\tau=1}^{m}\mathbb{E}\left[\langle \mathbf{g}_{T_0+\tau}, \vec{\theta}'_{T_0+\tau}\rangle - g_{T_0+\tau,a_{T_0+\tau}}\middle| \mathcal{F}_{T_0}\right] \\
\leq &\left(1+\sum_{\tau=1}^{m-1}\|\vec{\theta}'_{T_0+\tau+1} - \vec{\theta}'_{T_0+\tau}\|_1\right)\mathbb{E}\left[\eta_m^{-1}\ln\frac{1}{\beta_m}\middle| \mathcal{F}_{T_0}\right] + e\mathbb{E}\left[\sum_{\tau=1}^{m}\eta_\tau\|\mathbf{g}_{T_0+\tau}\|_2^2\middle| \mathcal{F}_{T_0}\right] \\
&+\mathbb{E}\left[\sum_{\tau=1}^{m}\gamma_\tau\left\langle \mathbf{g}_{T_0+\tau}, \frac{\mathbf{1}}{K}\right\rangle\middle| \mathcal{F}_{T_0}\right] \\
\leq &\left(1+\sum_{\tau=1}^{m-1}\|\vec{\theta}'_{T_0+\tau+1} - \vec{\theta}'_{T_0+\tau}\|_1\right)\cdot 4M^3Km^{1+\frac{2}{3}\delta}\cdot(3\ln m + \ln K) \\
&+e\frac{1}{4}M^{-1}K^{-1}m^{-1-\frac{2}{3}\delta}\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{g}_{T_0+\tau}\|_2^2\middle| \mathcal{F}_{T_0}\right] + \frac{1}{4}K^{-1}m^{-1-\frac{1}{3}\delta}\|\mathbf{Q}_{T_0}\|_\infty\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{g}_{T_0+\tau}\|_1\middle| \mathcal{F}_{T_0}\right] \\
\leq &\,4\left(1+C_Vm^{1-\delta}\right)\cdot M^3Km^{1+\frac{2}{3}\delta}\cdot(3\ln m + \ln K)
\end{aligned}
$$

$$+\frac{e}{4}M^{-1}K^{-1}m^{-1-\frac{2}{3}\delta}\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{g}_{T_0+\tau}\|_2^2\middle|\mathcal{F}_{T_0}\right]+\frac{1}{4}K^{-1}m^{-1-\frac{1}{3}\delta}\|\mathbf{Q}_{T_0}\|_\infty\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{g}_{T_0+\tau}\|_1\middle|\mathcal{F}_{T_0}\right],$$

其中在最后一步中，我们使用了假设3中关于 $\sum_{\tau=1}^{m-1}\|\vec{\theta}'_{T_0+\tau+1}-\vec{\theta}'_{T_0+\tau}\|_1$ 的界。为界定涉及 $\mathbf{g}_{T_0+\tau}$ 因子的两个期望项，回顾一下，$\mathbf{g}_{T_0+\tau}$ 在裁剪之前仍满足矩界：

$$\begin{aligned}\mathbb{E}\left[g_{T_0+\tau,i}^2\middle|\mathcal{F}_{T_0}\right]&\le\mathbb{E}\left[Q_{T_0+\tau-1,i}^2S_{T_0+\tau,i}^2\middle|\mathcal{F}_{T_0}\right]\\&=\mathbb{E}\left[Q_{T_0+\tau-1,i}^2\middle|\mathcal{F}_{T_0}\right]\cdot\mathbb{E}\left[S_{T_0+\tau,i}^2\middle|\mathcal{F}_{T_0}\right]\\&\overset{(a)}{\le}(Q_{T_0,i}^2+M^2(\tau-1)^2)\cdot M^2\\&\le((2Mm)^2+M^2m^2)\cdot M^2\\&=5M^4m^2,\end{aligned}$$

其中，在步骤（a）中，我们以与证明定理 G.1 时相同的方式，将 $\mathbb{E}\left[Q_{T_0+\tau-1,i}^2\middle|\mathcal{F}_{T_0}\right]$ 上界限定为 $Q_{T_0,i}^2+M^2(\tau-1)^2$。类似地，

$$\begin{aligned}\mathbb{E}\left[g_{T_0+\tau,i}\middle|\mathcal{F}_{T_0}\right]&\le\mathbb{E}\left[Q_{T_0+\tau-1,i}S_{T_0+\tau,i}\middle|\mathcal{F}_{T_0}\right]\\&=\mathbb{E}\left[Q_{T_0+\tau-1,i}\middle|\mathcal{F}_{T_0}\right]\cdot\mathbb{E}\left[S_{T_0+\tau,i}\middle|\mathcal{F}_{T_0}\right]\\&\le(Q_{T_0,i}+M(\tau-1))\cdot M\\&\le(2Mm+Mm)\cdot M\\&=3M^2m.\end{aligned}$$

因此，两个期望因子可以界定为

$$\begin{aligned}\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{g}_{T_0+\tau}\|_2^2\middle|\mathcal{F}_{T_0}\right]&\le5KM^4m^3,\\\|\mathbf{Q}_{T_0}\|_\infty\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{g}_{T_0+\tau}\|_1\middle|\mathcal{F}_{T_0}\right]&\le2Mm\cdot3KM^2m\\&=6KM^3m^2.\end{aligned}$$

因此，我们可以得出

$$\begin{aligned}&\sum_{\tau=1}^{m}\mathbb{E}\left[\langle\mathbf{g}_{T_0+\tau},\vec{\theta}'_{T_0+\tau}\rangle-g_{T_0+\tau,a_{T_0+\tau}}\middle|\mathcal{F}_{T_0}\right]\\&\le4\left(1+C_Vm^{1-\delta}\right)\cdot M^3Km^{1+\frac{2}{3}\delta}\cdot(3\ln m+\ln K)\\&\quad+\frac{5e}{4}M^3m^{2-\frac{2}{3}\delta}+\frac{3}{2}M^3m^{2-\frac{1}{3}\delta}\\&\le4\left(1+C_Vm^{1-\delta}\right)\cdot M^3Km^{1+\frac{2}{3}\delta}\cdot(3\ln m+\ln K)\\&\quad+\frac{7}{2}M^3m^{2-\frac{2}{3}\delta}+\frac{3}{2}M^3m^{2-\frac{1}{3}\delta}\\&\le9(1+C_V)M^3Km^{2-\frac{\delta}{3}}\cdot(3\ln m+\ln K).\end{aligned}$$

最后，请注意，对于每个 τ，单步遗憾 $\langle\mathbf{Q}_{T_0+\tau-1},\ \mathbf{S}_{T_0+\tau}\odot\theta_{T_0+\tau}\rangle-Q_{T_0+\tau-1,\ a_{T_0+\tau}}S_{T_0+\tau,\ a_{T_0+\tau}}$可以上界限定为

$$\begin{aligned}&\langle\mathbf{Q}_{T_0+\tau-1},\mathbf{S}_{T_0+\tau}\odot\vec{\theta}_{T_0+\tau}\rangle-Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}}\\&\le\underbrace{\langle\mathbf{g}_{T_0+\tau},\vec{\theta}_{T_0+\tau}\rangle-g_{T_0+\tau,a_{T_0+\tau}}}_{\mathbf{g}_{T_0+\tau}\text{-objective single step regret}}+\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}}>Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]\langle\mathbf{Q}_{T_0+\tau-1},\mathbf{S}_{T_0+\tau}\odot\vec{\theta}_{T_0+\tau}\rangle.\end{aligned}$$

并且每个差分项都可以通过以下方式加以控制

$$\mathbb{E}\left[\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}}>Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]\langle\mathbf{Q}_{T_0+\tau-1},\mathbf{S}_{T_0+\tau}\odot\vec{\theta}_{T_0+\tau}\rangle\middle|\mathcal{F}_{T_0}\right]$$

$$\leq \mathbb{E}\left[\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]\sum_{i=1}^{K}Q_{T_0+\tau-1,i}S_{T_0+\tau,i}\middle|\mathcal{F}_{T_0}\right]$$
$$= \mathbb{E}\left[\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]\sum_{i\in[K]:i\neq a_t}Q_{T_0+\tau-1,i}S_{T_0+\tau,i}\middle|\mathcal{F}_{T_0}\right]$$
$$+\mathbb{E}\left[\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]Q_{T_0+\tau-1,a_t}S_{T_0+\tau,a_t}\middle|\mathcal{F}_{T_0}\right].$$

对于第一项 $\mathbb{E}\left[\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}}Q_{T_0,a_t} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]\sum_{i\in[K]:\ i\neq a_t}Q_{T_0+\tau-1,\ i}S_{T_0+\tau,\ i}\middle|\mathcal{F}_{T_0}\right]$，我们可以写成

$$\mathbb{E}\left[\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]\sum_{i\in[K]:i\neq a_t}Q_{T_0+\tau-1,i}S_{T_0+\tau,i}\middle|\mathcal{F}_{T_0}\right]$$
$$= \mathbb{E}\left[\mathbb{P}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\middle|\mathcal{F}_{T_0},a_t\right]\cdot\mathbb{E}\left[\sum_{i\in[K]:i\neq a_t}Q_{T_0+\tau-1,i}S_{T_0+\tau,i}\middle|\mathcal{F}_{T_0},a_t\right]\mathcal{F}_{T_0}\right].$$

回想一下对于每个 $1\leq\tau\leq m$ 和 $i\in[K]$，我们可以将$Q_{T_0+\tau-1,i}S_{T_0+\tau,i}$ 的二阶矩界定为

$$\begin{aligned}\mathbb{E}\left[Q^2_{T_0+\tau-1,i}S^2_{T_0+\tau,i}\middle|\mathcal{F}_{T_0}\right] &= \mathbb{E}\left[Q^2_{T_0+\tau-1,i}\middle|\mathcal{F}_{T_0}\right]\cdot\mathbb{E}\left[S^2_{T_0+\tau,i}\middle|\mathcal{F}_{T_0}\right]\\ &\overset{(a)}{\leq}(Q^2_{T_0,i}+M^2(\tau-1)^2)\cdot M^2\\ &\leq((2Mm)^2+M^2m^2)\cdot M^2\\ &=5M^4m^2,\end{aligned}$$

因此，根据切比雪夫不等式，我们有

$$\begin{aligned}\mathbb{P}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\middle|\mathcal{F}_{T_0},a_t\right] &\leq 5M^4m^2\cdot(Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty)^{-2}\\ &\overset{(a)}{\leq}5M^4m^2\cdot(Mm^{\frac{\delta}{3}}\cdot Mm)^{-2}\\ &=5m^{-\frac{2}{3}\delta},\end{aligned}$$

其中，步骤（a）基于假设 $m\geq 2$，因此 $\|\mathbf{Q}_{T_0}\|_\infty\geq 2M(m-1)\geq Mm$。对于后续因子$\mathbb{E}\left[\sum_{i\in[K]:\ i\neq a_t}Q_{T_0+\tau-1,i}S_{T_0+\tau,i}\middle|\mathcal{F}_{T_0},a_t\right]$，我们有

$$\mathbb{E}\left[\sum_{i\in[K]:i\neq a_t}Q_{T_0+\tau-1,i}S_{T_0+\tau,i}\middle|\mathcal{F}_{T_0},a_t\right]\leq(K-1)\cdot\sqrt{5}M^2m.$$

对于另一个项 $\mathbb{E}\left[\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]Q_{T_0+\tau-1,\ a_t}S_{T_0+\tau,\ a_t}\middle|\mathcal{F}_{T_0}\right]$，我们有

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]Q_{T_0+\tau-1,a_t}S_{T_0+\tau,a_t}\middle|\mathcal{F}_{T_0}\right]\\ &\leq\mathbb{E}\left[Q^2_{T_0+\tau-1,a_t}S^2_{T_0+\tau,a_t}\middle|\mathcal{F}_{T_0}\right]\cdot\left(Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right)^{-1}\\ &\leq 5M^4m^2\cdot(Mm^{\frac{\delta}{3}}\cdot Mm)^{-1}\\ &=5M^2m^{1-\frac{\delta}{3}}.\end{aligned}$$

因此，结合界各部分的界，我们得到

$$\begin{aligned}&\sum_{\tau=1}^{m}\mathbb{E}\left[\mathbb{1}\left[Q_{T_0+\tau-1,a_{T_0+\tau}}S_{T_0+\tau,a_{T_0+\tau}} > Mm^{\frac{\delta}{3}}\|\mathbf{Q}_{T_0}\|_\infty\right]\langle\mathbf{Q}_{T_0+\tau-1},\mathbf{S}_{T_0+\tau}\odot\vec{\theta}_{T_0+\tau}\rangle\middle|\mathcal{F}_{T_0}\right]\\ &\leq 5\sqrt{5}(K-1)M^2m^{2-\frac{2}{3}\delta}+5M^2m^{2-\frac{\delta}{3}}\end{aligned}$$

$$\leq 12KM^2m^{2-\frac{\delta}{3}}.$$

将该遗憾差值上界与$\mathbf{g}_{T_0+\tau}$-目标遗憾上界结合，可得

$$\begin{aligned}
&\sum_{t=1}^{m}\mathbb{E}\left[\langle \mathbf{Q}_{T_0+t-1}\odot \mathbf{S}_{T_0+t},\vec{\theta}_{T_0+t}\rangle - Q_{T_0+t-1,a_{T_0+t}}S_{T_0+t,a_{T_0+t}}\middle|\mathcal{F}_{T_0}\right]\\
&\leq 9(1+C_V)M^3Km^{2-\frac{\delta}{3}}\cdot(3\ln m+\ln K)+4M^2+12KM^2m^{2-\frac{\delta}{3}}\\
&\leq 21(1+C_V)M^3Km^{2-\frac{\delta}{3}}\cdot(3\ln m+\ln K)+4M^2.
\end{aligned}$$

□

结合定理 I.1 和定理 G.1，可得到 $\sum_{\tau=1}^{m}\|\mathbf{Q}_{T_0+\tau-1}\|_1$中的遗憾上界：

引理 I.2。 假设假设 *1*、*3* 和 *4* 成立，则设 T_0 为我们在算法 *5* 中启动一个长度为 m 的新 *EXP3.S+*实例的某个时间步，有

$$\begin{aligned}
&\mathbb{1}[T_0 \text{ 结束一个 } \textit{EXP3} \text{ 实例，新的 } \textit{EXP3} \text{ 实例长度为 } m,\ m\geq 2]\\
&\cdot\sum_{t=1}^{m}\mathbb{E}\left[\langle \mathbf{Q}_{T_0+t-1}\odot \mathbf{S}_{T_0+t},\vec{\theta}_{T_0+t}\rangle - Q_{T_0+t-1,a_{T_0+t}}S_{T_0+t,a_{T_0+t}}\middle|\mathcal{F}_{T_0}\right]\\
&\leq 42(1+C_V)M^2Km^{-\frac{\delta}{3}}\cdot(3\ln m+\ln K)\cdot\mathbb{E}\left[\sum_{\tau=1}^{m}\|\mathbf{Q}_{T_0+\tau-1}\|_2^2\middle|\mathcal{F}_{T_0}\right]+4M^2. \qquad (49)
\end{aligned}$$

根据定理 I.2，剩余步骤与算法 3 的分析相同。

将第 i 个 EXP3.S+ 实例结束的时间记为 τ_i $(i\geq 0)$。那么， $\tau_0=0$, $\{\tau_i\}$ 是一个非递减的、 $\{\mathcal{F}_t\}$-适应的停止时序列。此外，每个 τ_{i+1} 都是 $\mathcal{F}_{\tau_i}$-可测的。任取一个 $T\geq 1$，定义

$$\tau_i' \triangleq \begin{cases} 0 & \text{if } i=0\\ \tau_i & \text{if } i>0 \text{ and } \tau_{i-1}'<T\\ \tau_{i-1}' & \text{otherwise}\end{cases},$$

那么， $\{\tau_i'\}$ 是一个非递减的、 $\{\mathcal{F}_t\}$-适应的停止时序列，每个 τ_{i+1}' 都是 $\mathcal{F}_{\tau_i'}$-可测的， $\tau_{i+1}'=\tau_i'$ 当且仅当 $\tau_i'\geq T$。因此，我们可以将定理 I.2 重述如下：

引理 I.3。 假设假设 *1*、*3* 和 *4* 成立，则有

$$\begin{aligned}
&\mathbb{1}\left[\|\mathbf{Q}_{\tau_i'}\|_\infty\geq 4M\right]\sum_{t=1}^{\tau_{i+1}'-\tau_i'}\mathbb{E}\left[\langle \mathbf{Q}_{\tau_i'+t-1}\odot \mathbf{S}_{\tau_i'+t},\vec{\theta}_{\tau_i'+t}\rangle - Q_{\tau_i'+t-1,a_{\tau_i'+t}}S_{\tau_i'+t,a_{\tau_i'+t}}\middle|\mathcal{F}_{\tau_i'}\right]\\
&\leq h(\tau_{i+1}'-\tau_i')+f(\tau_{i+1}'-\tau_i')\cdot\sum_{t=1}^{\tau_{i+1}'-\tau_i'}\mathbb{E}\left[\|\mathbf{Q}_{\tau_i'+t-1}\|_1\middle|\mathcal{F}_{\tau_i'}\right]
\end{aligned}$$

对于任意 $i\geq 0$，其中

$$f(m)=42(1+C_V)M^2Km^{-\frac{\delta}{3}}\cdot(3\ln m+\ln K),$$
$$h(m)=\mathbb{1}[m>0]\cdot 4M^2.$$

固定某个 $T\geq 1$，令 $\mathcal{T}_0\triangleq\sup\{\tau_i\colon\ i\geq 0,\ \tau_i<T\}$, $\mathcal{T}_1\triangleq\inf\{\tau_i\colon\ i\geq 0,\ \tau_i\geq T\}$，则 $\mathcal{T}_0$和 $\mathcal{T}_1$都是 $\{\mathcal{F}_t\}$-适应的停止时间，$\mathcal{T}_1\quad\mathcal{F}_{\mathcal{T}_0}\qquad\mathcal{T}_0<T\leq\mathcal{T}_1\qquad\mathcal{T}_1-\mathcal{T}_0\leq\frac{\|\mathbf{Q}_{\mathcal{T}_0}\|_\infty}{2M}+1\leq$是 -可测的。注意，。此外，由于 $\frac{\mathcal{T}_0\cdot M}{2M}+1=\frac{\mathcal{T}_0}{2}+1\leq\frac{\mathcal{T}_0}{2}+T\qquad\mathcal{T}_1\leq\frac{5}{2}T$ Q，因此我们有。

下面，我们将结合定理 5.3 和定理 I.3，对 $\mathbb{E}[\sum_{t=1}^{\mathcal{T}_1}\|\mathbf{Q}_{t-1}\|_1]$ 在 $\mathcal{O}(\mathbb{E}[\mathcal{T}_1])=\mathcal{O}(T)$ 中进行界定，从而得出 $\mathbb{E}[\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1]$ 也属于 $\mathcal{O}(T)$。

回顾一下， $\epsilon>0$ 是假设 1 中参考策略 $\{\theta_t\}$ 的“离开相对于到达的平均优势”的下界，定义

$$m_0\triangleq\inf\left\{m:m\geq 2, f(m')\leq\frac{\epsilon}{2}\forall m'\geq m\right\},$$

那么， m_0 是一个仅取决于 δ 和 ϵ的常数，事实上，

$$m_0 \leq \left((1+C_V)M^2K\ln K\epsilon^{-1}\right)^{\varnothing(1/\delta)}.$$

通过讨论每个时期的轮次长度 $\tau'_{i+1}-\tau'_i$是否大于 m_0，我们根据定理 I.3 得出

$$\begin{aligned}
&\sum_{t=1}^{\tau'_{i+1}-\tau'_i}\mathbb{E}\left[\langle \mathbf{Q}_{\tau'_i+t-1}\odot\mathbf{S}_{\tau'_i+t},\vec{\theta}_{\tau'_i+t}\rangle - Q_{\tau'_i+t-1,a_{\tau'_i+t}}S_{\tau'_i+t,a_{\tau'_i+t}}\middle|\mathcal{F}_{\tau'_i}\right]\\
&\leq h(\tau'_{i+1}-\tau'_i)+\frac{\epsilon}{2}\sum_{t=1}^{\tau'_{i+1}-\tau'_i}\mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1\middle|\mathcal{F}_{\tau'_i}\right]+\mathbb{1}[\tau'_{i+1}-\tau'_i\leq m_0]\sum_{t=1}^{\tau'_{i+1}-\tau'_i}\mathbb{E}\left[\langle \mathbf{Q}_{\tau'_i+t-1}\odot\mathbf{S}_{\tau'_i+t},\vec{\theta}_{\tau'_i+t}\rangle\middle|\mathcal{F}_{\tau'_i}\right]\\
&\leq h(\tau'_{i+1}-\tau'_i)+\frac{\epsilon}{2}\sum_{t=1}^{\tau'_{i+1}-\tau'_i}\mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1\middle|\mathcal{F}_{\tau'_i}\right]+\mathbb{1}[\tau'_{i+1}-\tau'_i\leq m_0]M\sum_{t=1}^{\tau'_{i+1}-\tau'_i}\mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1\middle|\mathcal{F}_{\tau'_i}\right]\\
&\overset{(a)}{\leq} h(\tau'_{i+1}-\tau'_i)+\frac{\epsilon}{2}\sum_{t=1}^{\tau'_{i+1}-\tau'_i}\mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1\middle|\mathcal{F}_{\tau'_i}\right]+\mathbb{1}[\tau'_{i+1}-\tau'_i\leq m_0]M\cdot 3KM(\tau'_{i+1}-\tau'_i)^2\\
&\leq h(\tau'_{i+1}-\tau'_i)+\frac{\epsilon}{2}\sum_{t=1}^{\tau'_{i+1}-\tau'_i}\mathbb{E}\left[\|\mathbf{Q}_{\tau'_i+t-1}\|_1\middle|\mathcal{F}_{\tau'_i}\right]+3KM^2m_0(\tau'_{i+1}-\tau'_i) \qquad (50)
\end{aligned}$$

对所有 $i\geq 0$均成立。这里在步骤 (a) 中，我们应用定理 G.1。

将式 (50) 对所有 $i\geq 0$ 求和，然后取全期望，可得

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1}\langle \mathbf{Q}_{t-1}\odot\mathbf{S}_t,\vec{\theta}_t\rangle - Q_{t-1,a_t}S_{t,a_t}\right]\leq\frac{\epsilon}{2}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1}\|\mathbf{Q}_{t-1}\|_1\right]+3KM^2m_0\mathbb{E}\left[\mathcal{T}_1\right]+\mathbb{E}\left[\sum_{i=0}^{\infty}h(\tau'_{i+1}-\tau'_i)\right].$$

对于任意样本路径，我们有 $\sum_{i=0}^{\infty}h(\tau'_{i+1}-\tau'_i)\leq 4M^2\mathcal{T}_1$，因此

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1}\langle \mathbf{Q}_{t-1}\odot\mathbf{S}_t,\vec{\theta}_t\rangle - Q_{t-1,a_t}S_{t,a_t}\right]\leq\frac{\epsilon}{2}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1}\|\mathbf{Q}_{t-1}\|_1\right]+(3KM^2m_0+4M^2)\mathbb{E}\left[\mathcal{T}_1\right]. \qquad (51)$$

理 5.4，我们还可以找到一个依赖于 $\mathcal{T}_1$的常数 $\mathcal{T}_2$，使得 $\mathcal{T}_2\leq\mathcal{T}_1+\sqrt{\frac{\mathcal{T}_1}{C_W}}+1$an
根据定 d

$$-\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_2}\langle \mathbf{Q}_{t-1},\vec{\sigma}_t\odot\vec{\theta}_t-\vec{\lambda}_t\rangle\right]\leq-\epsilon\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1}\|\mathbf{Q}_{t-1}\|_1\right]+(KM^2+\epsilon KM)C_W\mathbb{E}[\mathcal{T}_2]. \qquad (52)$$

如果 $T\geq\frac{4}{C_W}+C_W$，我们有 $\mathcal{T}_1\geq\max\{C_W,\frac{4}{C_W}\}$，因此 $\mathcal{T}_2\leq 2\mathcal{T}_1\leq 5T$。此外，定理 5.3 保证

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_2}Q_{t-1,a_t}S_{t,a_t}-\langle \mathbf{Q}_{t-1},\vec{\lambda}_t\rangle\right]\leq\frac{(K+1)M^2\mathbb{E}[\mathcal{T}_2]}{2}. \qquad (53)$$

结合式 (51) 至 (53)，我们得到

$$\begin{aligned}
\frac{\epsilon}{2}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1}\|\mathbf{Q}_{t-1}\|_1\right]&\leq\left[\frac{(K+1)M^2}{2}+(KM^2+\epsilon KM)C_W+3KM^2m_0+4M^2\right]\mathbb{E}[\mathcal{T}_2]\\
&\leq\left[\frac{(K+1)M^2}{2}+(KM^2+\epsilon KM)C_W+3KM^2m_0+4M^2\right]\cdot 5T.
\end{aligned}$$

因此，当 $T\geq\frac{4}{C_W}+C_W$ 时，我们有

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{Q}_{t-1}\|_1\right]\leq\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_1}\|\mathbf{Q}_{t-1}\|_1\right]\leq\left[3KM^2m_0+\frac{(K+1)M^2}{2}+(KM^2+\epsilon KM)C_W+4M^2\right]\cdot\frac{10}{\epsilon}.$$