arXiv:2601.03533v1 [stat.ML] 7 Jan 2026

滑动窗口模型中的有限信息在线学习

VladimirBraverman SumeghaGarg约翰斯·霍普金斯大学 罗格斯大学
vova@cs.jhu.edu sumegha.garg@rutgers.comChen Wang David P.Woodruff 伦斯勒理工学院 卡内基梅隆大学
wangc33@rpi.edu dwoodruf@andrew.cmu.edu

Samson Zhou 德克萨斯农工大学
samsonzhou@gmail.com

摘要

受近期流模型中专家问题相关研究的启发，我们研究滑动窗口模型中的专家问题。滑动窗口模型是一种研究成熟的模型，适用于交通监测、疫情追踪和自动化交易等应用场景，在这些场景中，近期信息比旧数据更有价值。形式上，我们有 n 位专家、 T 天，每天都能查询 q 位专家的预测，内存有限，并且应在最近 W 天的任意窗口上实现（近似）最优的遗憾 $\sqrt{nW}$ polylog(nT)。虽然只进行 1 次查询不可能实现这样的遗憾，但我们证明，进行 2 次查询即可实现这样的遗憾，且只需 polylog(nT) 位内存。我们的算法不仅对滑动窗口最优，而且还证明，对于任意由天构成的区间$\mathcal{I}$，只需 2 次查询和 polylog(nT) 位内存，就能实现 $\sqrt{n|\mathcal{I}|}$polylog(nT) 的遗憾，使内存需求相比以往的区间遗憾算法实现了指数级改进。基于这些技术，我们还研究了数据流中的老虎机问题，其中 $q = 1$，使用 polylog(nT) 内存即可实现 $nT^{2/3}$polylog(T) 的遗憾。这是在多臂老虎机设定下，首次在流模型中以多对数内存实现次线性遗憾；如果最佳专家的损失以随机顺序排列，还可以进一步改进至最优的 $\mathcal{O}(\sqrt{nT})$ 遗憾。

1 引言

专家在线学习问题是序贯预测的一种基础框架，其中算法利用专家预测，随时间推移做出决策。在长度为 T天（或时间步）的时间范围内，算法必须根据 n 个预定义专家（在多臂老虎机文献中也称为“臂”）的建议，对未知结果进行预测。[1]每天，算法都会观察一组专家的预测，然后结合过去的观察结果和当前的专家预测，生成自己的预测。做出预测后，算法会收到反馈，其中包括表明其决策准确度的代价，以及当天查询的一部分专家预测所对应的代价。通常，代价被限制在某个范围 $[0, \rho]$ 内，其中参数为 $\rho > 0$；我们将其归一化为 1。这一迭代过程会随时间持续进行，使算法能够不断改进其策略。

形式化地说，在每一天 $t \in [T]$，算法都会选择一个臂 $i_t \in [n]$ 来操作，从而产生损失 $\ell^t(i_t)$。此外，它还可以无损地查询额外集合 $S_t \subseteq [n]$ 中各臂的损失，不过我们称 $\max_{t\in[T]} |S_t|$ 为算法的查询复杂度。专家预测和损失序列预先固定，且我们不作任何分布假设；也就是说，输入是最坏情况的，但对于算法的输出而言是不可适应的（非自适应）。目标是最小化遗憾，其定义为我们得到的损失序列与事后表现最佳的臂的损失序列之间的累积差距，即

$$\sum_{t\in[T]} \ell^t(i_t) - \min_{i\in[n]} \sum_{t\in[T]} \ell^t(i). \tag{1}$$

在算法可以随时自由观测所有专家的情况下，也就是说， $S_t = [n]$ 对于所有 $t \in [T]$，随机加权多数算法可以实现遗憾$\mathcal{O}\left(\sqrt{T \log n}\right)$ [LW94]。此后还出现了多种变体，例如 Hedge [CFH+97、FS97] 以及跟随扰动领袖[KV05]，，它们也能实现遗憾$\mathcal{O}\left(\sqrt{T \log n}\right)$，且已知这是最优的 [Cov66]。然而，这些算法采用的策略需要维护每位专家的累积成本，这需要$\Omega(n)$ 位空间，以及$\Omega(n)$ 更新时间。

在许多实际场景中，专家的数量 n 和时间范围 T 可能很大，因此存储和处理所有专家预测既会带来较高的计算开销，也会占用大量内存。因此，近期有一系列研究 [SWXZ22、ACNS23、PR23、PZ23、WZZ23a] 关注有界内存下的专家在线学习问题，尤其是在数据流模型中，算法只需单次扫描输入，同时使用的工作内存小于输入规模的线性量级 n 。具体而言，算法无法完整存储过去的专家预测，因此不能采用以往的方法，而必须策略性地决定保留哪些信息，同时确保其遗憾仍具有竞争力。尽管如此， [PR23]提出了一种算法，使用多重对数(nT) 空间，并实现了 $\sqrt{nT}\cdot$ 多重对数 (nT) 遗憾。因此，从某种意义上说，遗憾与空间复杂度之间似乎不存在代价高昂的权衡。

滑动窗口模型与区间遗憾。尽管数据流模型为分析如何以有限内存处理大规模数据的算法提供了框架，但它并未反映许多应用中的现实情况：近期信息往往比陈旧数据更有价值 [BBD+02, BDMO03,PGD15], 例如，交通监测中，实时拥堵数据可为路由决策提供依据；疫情追踪中，近期感染率可指导公共卫生响应；自动化交易中，最新市场信号则推动投资策略的制定。这

[1]我们交替使用“专家”和“臂”这两个术语。

这促使人们提出滑动窗口模型，在该模型中，任意时刻的数据集仅包含数据流中最近的W次更新。这里，参数$W > 0$定义了窗口大小，而早于W最近的数据流元素的更新则被视为已过期。目标是使用相对于W次方而言次线性的空间，对活跃数据进行统计汇总。滑动窗口模型可视为流模型的一种推广，因为参数W可以设为整个数据流的长度；该模型尤其适用于时间敏感场景，例如数据摘要 [CNZ16, ELVZ17],社交媒体中的事件检测 [OMM+14]以及网络流量监控 [CM05,CG07, Cor13]。

从实际角度来看，滑动窗口模型刻画了现实世界中数据保留方面的限制。例如，由于《通用数据保护条例》（GDPR）等监管政策对某些用户数据的存储时长施加了严格限制， [GDP], Apple会将用户信息存储 3 个月， [App], ChatGPT 最多保存用户对话 30 天， [Ope],而谷歌会将浏览数据保留最长 9 个月 [Goo]。通过适当设置窗口大小参数W，滑动窗口模型能够有效刻画此类限制，因此已在各个领域得到广泛研究 [DGIM02, LT06a, LT06b, BO07, CMS13, BLLM15, BLLM16, BWZ21, BGL+18, WZ21, BEL+20, EMMZ22, JWZ22, BLMZ23, WY23, WZZ23b, CJY+25]。我们指出，由于活动窗口之外的数据会隐式过期，滑动窗口模型通常需要采用不同于标准流模型中所用的算法技术。例如，对整个数据流中的更新进行聚合的线性草图方法，通常与隐式丢弃过期信息的能力不兼容。

在专家在线学习问题中，滑动窗口模型的遗憾是在每个时间$t \in [T]$通过比较算法的表现与过去W个时间步中最佳臂的表现来衡量的。形式化地，时间t的遗憾定义为$\sum_{s=t-W+1}^{t} \ell^s(i_s) - \min_{i\in[n]} \sum_{s=t-W+1}^{t} \ell^s(i)$，而算法在滑动窗口模型中的总体遗憾则定义为所有时间步上的最大遗憾。顺便一提，这一概念与更强的区间遗憾概念相一致；区间遗憾是遗憾分析的一种自然细化，用于评估任意连续时间子区间上的表现，而不是整个时间范围上的表现。形式化地，对于任意区间$\mathcal{I} = [t_1, t_2] \subseteq [T]$，区间遗憾定义为

$$\sum_{t=t_1}^{t_2} \ell^t(i_t) - \min_{i\in[n]} \sum_{t=t_1}^{t_2} \ell^t(i). \tag{2}$$

这一指标能够更细致地了解算法的适应能力，体现其不仅在长期运行中的表现，也体现其在较短且可能动态变化的时间段内的表现。注意，滑动窗口遗憾是区间遗憾的一个特例，对应于在长度为$\mathcal{I}$ $|\mathcal{I}| = W$的区间上取最大值的情形。令人惊讶的是，尽管滑动窗口模型已得到广泛研究，但据我们所知，此前尚无关于具有内存约束的专家在线学习问题的研究。

1.1 我们的贡献

在这项工作中，我们首次研究滑动窗口模型中的专家在线学习问题。首先回顾一下，在单查询多臂老虎机设定中，任何算法都必然会遭受区间遗憾$\Omega(|\mathcal{I}|^{1-\varepsilon})$，对所有固定常数$\varepsilon > 0$ [DGS15]均成立。由于区间遗憾恰好对应于算法在滑动窗口模型中的最坏情况遗憾，因此，滑动窗口模型中任何用于专家在线学习的单查询算法都必须遭受遗憾$\Omega(W^{1-\varepsilon})$，对所有固定常数$\varepsilon > 0$ [DGS15]。

我们证明，只需额外增加一次查询，就不仅能够实现最优区间遗憾，还能仅使用多对数内存。具体而言，其中一次查询必须跟随一位专家，由算法承担相应的损失；另一次查询则可以自由观察任意一位专家，该专家可以随时间变化，且不会产生损失。

定理 1。 存在一种在线学习算法，对于任意包含 n 位专家和 T 天的实例，满足 $T \geqslant n$，且每个时刻进行两次查询；该算法能够以高概率为任意区间 I 实现 $\sqrt{n|I|} \cdot$ 多重对数(T) 的区间遗憾，所用内存为 多重对数(T) 个内存字数，即 $1 - \frac{1}{\text{多项式}(nT)}$。

我们注意到，由于区间遗憾是比整个 T 天上的期望遗憾更强的概念，定理 1 也立即推出了一种双查询算法，对于$\sqrt{nT} \cdot$ 多重对数(T) 的遗憾，在长度为T的时间范围内解决专家在线学习问题，并使用多对数空间。事实上，由于已知，对于每个时刻查询次数为任意常数的专家在线学习问题，$\Omega(\sqrt{nT})$ 的遗憾是不可避免的 [ACFS02], ，我们的算法实现了最优遗憾，差距仅在多重对数因子以内。

值得注意的是，虽然 [LZC+24] 同样实现了 $\sqrt{nT} \cdot$ 多重对数(T) 的区间遗憾，且只需两次查询，定理 1 所用空间为多对数空间，而 [LZC+24] 则需要线性空间来追踪所有 n 专家的损失。对于滑动窗口模型而言，这一区别至关重要，因为该模型要求次线性空间复杂度。因此，定理 1立即给出了滑动窗口模型下首个双查询算法。

推论2。 在滑动窗口模型下，对于专家在线学习问题，存在一种每个时刻仅使用两次查询的算法。对于时间范围长度为 T、滑动窗口大小为 $W \geqslant n$ 且包含 n 个专家的任意实例，该算法以高概率实现 $\sqrt{nW} \cdot$ 多重对数(T) 遗憾，并使用多重对数(T) 比特的空间，即 $1 - \frac{1}{\text{多项式}(nT)}$。

我们再次强调，推论2 在查询复杂度方面是最优的，在遗憾和空间复杂度方面也接近最优，因为：(1) [ACFS02] 表明，$\Omega(\sqrt{nW})$ 遗憾是查询次数为常数的任何设置所必需的；(2) [DGS15] 的结果表明，即使内存不受限制，$\Omega(W^{1-\varepsilon})$ 遗憾仍是每个时刻仅进行一次查询时所必需的；(3) 仅仅为了存储任意“优秀”专家的身份，就需要对数级内存。因此，结合 [ACFS02、DGS15], 推论2 解决了滑动窗口模型下的专家在线学习问题。

我们指出，定理 1 中的技术可以调整，以处理区间遗憾和滑动窗口模型之外的许多其他设定。例如，一个自然的问题是：在无需保证更强的区间遗憾概念的情况下，仅具有单次查询老虎机反馈的内存受限算法能够达到怎样的遗憾。为此，我们给出以下结果。

定理 3。 存在一种在线学习算法，对于任意包含 n 个专家和 T 天的实例，满足 $T \geqslant n$，并且仅能查询单个专家，即在多臂老虎机设定下，以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率达到 $nT^{2/3} \cdot \operatorname{polylog}(T)$ 遗憾，同时使用 $\operatorname{polylog}(nT)$ 个内存字数。

据我们所知，在单查询多臂老虎机设定中，唯一一种使用次线性内存进行在线学习的先前算法，在整数 $\mathcal{O}\left(L \cdot n^{1/2L} \cdot T^{1-\frac{1}{2L}}\right)$ 遗憾，并使用 $\Theta(n^{1/L})$ 空间时适用 $L \geqslant 1$ [XZ21]。值得注意的是，他们的结果在使用 $\mathcal{O}\left(n^{1/4}T^{3/4}\right)$ 遗憾和 $\Theta(\sqrt{n})$ 空间时适用。相比之下，定理 3 以少得多（呈指数级减少）的空间实现了严格更强的遗憾保证。

因此，结合前面对区间遗憾的讨论，定理 3进一步凸显了流模型和滑动窗口模型中，单查询多臂老虎机设定下可实现保证之间的显著差异。事实上，在滑动窗口模型中，$\Omega(W^{1-\varepsilon})$ 遗憾对于任何 $\varepsilon > 0$ [DGS15], 都是必要的，而定理 3 则以多对数内存实现了大约 $T^{2/3}$ 遗憾，从而对任意 $W = T^{2/3+\Omega(1)}$都形成了显著差异。

最后，人们或许会问，我们的界能否在最坏情况之外进一步改进，例如通过对臂的表现作出分布假设。举例来说，流式文献中有大量工作研究随机顺序模型 [GM09]。在专家在线学习问题的背景下， [SWXZ22]指出，随机顺序模型意味着输入分布中损失向量的任意排列都具有相同的可能性。特别是， [SWXZ22]指出，按天计的损失具有可交换性 这一性质，使随机顺序模型能够涵盖独立同分布模型；在该模型中，每个臂每天的损失都遵循一个固定的底层分布（随机专家）。因此，在随机顺序模型中建立的任何算法上界也适用于独立同分布模型。

我们证明，即使在较弱的分布假设下，也能够同时利用单次查询老虎机反馈和多对数空间达到最优遗憾。

定理 4. 存在一种在线学习算法，对于任意包含 n 个专家和 T 天的实例，满足 $T \geqslant n$ ，并且仅能查询单个专家（即多臂老虎机设定），当最佳专家的损失序列按随机顺序排列时，该算法以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率，使用 $\operatorname{polylog}(nT)$ 个内存字，实现 $\sqrt{nT} \cdot \operatorname{polylog}(nT))$ 遗憾。

我们强调，定理 4 只要求最佳专家按随机顺序排列，这一要求弱于完整的随机顺序模型。相比之下， [SWXZ22] 最初提出了一种算法，在随机顺序模型下解决专家在线学习问题，实现期望遗憾 R，使用 $\tilde{\mathcal{O}}\left(\frac{nT}{R}\right)$ 空间；不过，随后这些结果得到改进 [PZ23、PR23]，最终由 $\tilde{\mathcal{O}}\left(\sqrt{nT}\right)$ 遗憾的算法进一步改进：在任意顺序模型下仅使用多对数空间 [PR23]。我们指出，这一系列工作始终使用多重对数次查询，而定理 4 仅使用单次查询老虎机反馈。

2 技术概述

本节将对我们的算法和分析进行技术概述。我们首先讨论其他自然方法、它们为何失效，以及由此产生的影响。

2.1 自然方法的不足

尽管流模型中有各种各样的技术，但这些技术通常无法直接应用于滑动窗口模型。例如，如前所述，线性草图无法直接处理因额外更新而隐式过期的数据。因此，我们将重点讨论最相关的滑动窗口技术。

直方图框架行不通。滑动窗口算法最常用的方法是基于直方图的方法，例如指数直方图 [DGIM02], 平滑直方图 [BO07], 及其改进版本 [BGL+18、BLLM15、BLLM16、CNZ16、ELVZ17、BEL+20、BWZ21、EMMZ22、BLMZ23]。这些框架将针对某个问题的仅插入流算法

针对问题 并行运行该问题的多个实例，并在数据流中的不同时刻启动这些实例，称为检查点，从而构造滑动窗口算法。每次数据流更新时都会创建新的检查点；当相应算法的输出值彼此过于“接近”时（例如相差不超过 2 倍），则移除已有检查点。因此，如果某个函数性质良好且有界，可以证明，在数据流运行过程中的任意时刻，检查点的数量都只有对数级。重要的是，存在两个检查点 t_1 和 t_2，它们“夹住”滑动窗口的起点。直观来说，从 t_1 开始的数据流上的函数值，以及从 t_2 开始的数据流上的函数值，将滑动窗口上的函数值夹在中间。因此，只需输出从时刻 t_2开始运行的算法 $\mathcal{A}$ 的值即可。

遗憾的是，将这种方法用于专家在线学习问题时会遇到多个问题。首先，这些直方图方法必然会大幅增加查询复杂度。通过并行维护$\mathcal{O}(\log n)$个单查询多臂老虎机算法实例，查询复杂度会变为$\mathcal{O}(\log n)$，这对于我们的目标而言大得难以接受。更严重的是，这些直方图方法通常会使用$\mathcal{O}(\log n)$个流式算法实例，以实现对总体损失的常数因子近似。然而，在大小为 W的窗口上，对损失作常数因子近似可能会导致线性遗憾，而非我们期望的 $\sqrt{W}$ 遗憾。我们可以通过维护 $\sqrt{W}$ 个流式算法实例，并在损失每增加 $\sqrt{W}$ 而非按乘法比例变化时创建检查点来解决这一问题，但这样一来，所需空间会变为关于 W 的多项式，而非关于 W的多重对数。

在线核心集和重要性采样行不通。一种类似的思路是将在线算法的概念转化为滑动窗口算法。这一框架的核心是在线核心集的概念，它是数据流的一个代表性子集，能够保留与到达元素顺序相关的特性。 [BDM+20] 证明，对于某个问题，可以利用该问题的在线核心集实现滑动窗口算法；此后，这一思路被用于实现数值线性代数 [BDM+20、WY23] 以及聚类 [WZZ23b、CJY+25]。另一种常见且相关的方法，是根据数据流中元素看起来有多“重要”来对其进行采样。具体来说，重要性这一概念同时取决于独特性（即某个元素与数据集中其余元素看起来有多不同）和新近性（即某次数据流更新是否最近才出现在数据流中）。

然而，这两种方法都需要为底层问题构造一个核心集。与数值线性代数和聚类不同，目前尚无适用于专家在线学习的核心集构造方法，而且很难说为何应当存在一个能够以小空间保留必要遗憾保证的核心集。这一根本局限源于专家学习的对抗性：专家预测的多样性和适应性使得维护一个能够准确反映总体损失的代表性子集变得困难。

区间遗憾算法的改造。另一种自然的做法是改造现有的离线区间遗憾算法，使其使用次线性空间。例如， [DGS15]给出了一种算法，其遗憾为 $\widetilde{O}(\sqrt{nT})$，每天使用$\mathcal{O}(\log T)$次查询；而 [LZC+24] 给出了一种算法，其遗憾相似，为 $\widetilde{O}(\sqrt{nT})$，但每天只需两次查询。不过，这两种算法都需要$\Omega(n)$内存，以跟踪多个时间区间内的专家损失；这对于

分析至关重要，因为分析要据此确定每个时刻选择某位专家的概率。若没有线性内存，就可能严重误算某位专家的损失，进而导致不理想的概率分布，造成较高的遗憾。目前完全不清楚如何将这些方法改造为使用次线性内存；事实上，即使是在其他“较容易”的情形中，例如具有完全信息的专家在线学习问题，也就是 n 每天进行查询，要在使用次线性内存的同时实现近最优遗憾，仍是一项重大挑战 [XZ21、SWXZ22、PZ23、PR23、WZZ23a]。

对内存受限算法的改造。另一种做法是从现有的流式算法入手，并将其改造为适用于滑动窗口模型。例如， [XZ21]的算法可实现 $\mathcal{O}\left(Ln^{1/L}\right)$ 内存和 $\mathcal{O}\left(L\cdot T^{1-1/(2L)}n^{1/(2L)}\right)$ 遗憾，其中权衡参数 $L\geqslant 1$ 为任意正整数。尽管该算法适用于单查询多臂老虎机设定，但它会将 n 专家划分为若干个专家块，例如，将 $\sqrt{n}$ 个专家划分为 $\sqrt{n}$ 个专家的块，适用于 $L=2$。每个包含 $\sqrt{n}$ 个专家的块构成一个元专家，并且算法采用层次结构，在这些元专家上运行标准的单查询多臂老虎机算法。这必然需要跟踪所有元专家的（估计）损失，导致 $\mathcal{O}(\sqrt{n})$ 内存，适用于 $L=2$。其他 L设定也存在类似的局限性。一个自然的想法是沿层次结构递归，以构建更多层级，但这样一来，如何生成中间层级元专家的输出就并非显而易见了。

近期另一类研究探讨了有界内存但处于完全信息设置下的专家在线学习 [SWXZ22、PZ23、PR23、WZZ23a]。这些研究为元专家构建层次结构，并关键性地利用完全信息设置来模拟元专家的输出。由于该层次结构有$\mathcal{O}(\log n)$ 层，因此目前尚不清楚如何使用$\mathcal{O}(1)$ 次查询/时刻来模拟元专家。此外，该层次结构必然会导致较差的区间遗憾，因为结构的最高层要求维护在整个时间范围内表现良好的专家，而算法即使在这些专家于较小区间内表现不佳时，也不得不选择它们。

2.2 滑动窗口算法

我们的算法采用一种自然的方法：对算法跟踪的专家池进行采样，并剪除池中表现不佳的专家 [SWXZ22, PZ23, PR23, WZZ23a]。考虑一种算法，它将时间范围 T 划分为若干长度为 B的时期。在每个时期内，算法对池中的专家运行乘法权重更新（MWU）。每个时期结束时，算法以 $\frac{1}{n}$ 的概率向池中采样并加入一些额外的专家。为确保池的总大小较小，算法将通过移除不优于基准损失的专家来剪枝。这里，专家 i 在持续时间 D 内的基准损失 $\mathrm{L}^{\mathcal{D}}$(BM) 定义为：除 i之外，池中各臂在每个时期的最优损失。如果池中任何臂在持续时间 D 内的损失为 $\mathrm{L}^{\mathcal{D}}(i)$，且其表现并未严格优于基准，即，

$$\mathcal{L}^{\mathcal{D}}(i)\geqslant\mathcal{L}^{\mathcal{D}}(\mathsf{BM})-1, \tag{3}$$

那么，臂 i 将被算法淘汰。这样一来，留在池中的臂的损失必定以指数级递减，因此池的大小有界于 $\mathcal{O}(\log\ T)$，因为损失必须为 $[0,T]$。此外，在任意持续时间内跟踪 $\mathcal{O}(\log\ T)$ 个臂的损失，总共只需 polylog(nT) 空间。

由于该算法在池中的专家上运行乘法权重更新（MWU），遗憾分析的主要关注点是：当最佳专家 i^* 不在池中时，算法可能会如何表现。为此

，自然可以将每个时期划分如下：如果在某个时期对 i^* 进行采样，并且它随后因池中其他臂表现良好而被淘汰，那么该时期称为好时期。否则，该时期称为不良轮次。直观地说，得益于乘法权重更新的保证，算法在长度为 B 的好时期上的遗憾必定为 $\sqrt{B}$。另一方面，我们无法控制不良轮次上的遗憾，但可以证明，至多存在 $n\cdot$ 多重对数(n)个不良轮次，因为此后最佳臂 i^* 会在池中。因此，我们可以非正式地将遗憾上界估计为$((T/B)\cdot\sqrt{B+nB})\cdot$多重对数$(nT)$，通过适当选择 B，可将其优化至大约 $T^{2/3}$。我们指出，这一论证存在一些微妙之处：驱逐策略是针对已采样到池中的专家定义的，因此分析中存在依赖关系。不过，通过增加一些技术处理，采样相互独立的专家来处理驱逐（代价是池大小增加至多重对数(nT)），可以克服这些问题。

然而，该算法的主要瓶颈在于对大小为 polylog(nT) 的池中的专家运行乘法权重更新，这要求每个时刻进行数量大得难以承受的查询。一个直接的想法是用 EXP3 替代乘法权重更新。EXP3 是一种算法，使用单次查询的多臂老虎机信号 [ACFS02]，即可实现 $\sqrt{T}$遗憾（忽略 poly(n) 项）。因此，该算法在每个长度为 B 的时期内仍能实现大约 $\sqrt{B}$ 遗憾。这使得时期之间的算法只需单次查询。然而，我们无法在持续时间 D 内获得池中某个固定臂 i 的损失 $\mathrm{L}^{\mathcal{D}}(i)$。我们可以进行第二次查询，从池中均匀采样臂，以产生其损失的估计值 $\mathrm{L}^{\mathcal{D}}(i)$，误差为 $\sqrt{|\mathrm{D}|}$。随后，我们可以使用池中各臂的估计损失替代其实际损失，用于驱逐规则。这样一来，每个时刻的查询复杂度降至针对两个专家进行查询，但该算法仍能实现遗憾 $T^{2/3}$。

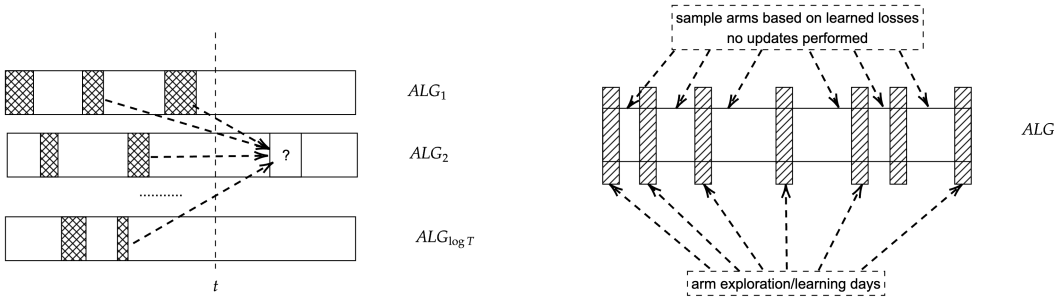
通过额外查询来提升。 我们的算法$\mathcal{A}_1$ 能够达到 $T^{2/3}$ 遗憾，主要是因为每个不良轮次产生 B 遗憾。自然的做法是采用标准的提升策略，再运行一个$\mathcal{A}_2$ 算法实例，其时间范围为 B，但这个算法在不良轮次上的遗憾为 $B^{2/3}$。然后，我们可以在$\mathcal{A}_1$ 和$\mathcal{A}_2$ 上运行一个“外层”EXP3，从而使不良轮次上的总遗憾达到 $B^{2/3}$。这样，总遗憾便为$\frac{T}{B}\sqrt{B+nB^{2/3}}$，经过优化后可得 $T^{4/7}$（省略 n 因子）。理想情况下，我们可以递归地应用这一提升过程，在经过 log(nT) 层后获得约为 $\sqrt{nT}$ 的遗憾。

遗憾的是，这种方法有一个致命缺陷。在部分信息设置下，基准算法和外层 EXP3 的更新都是随机的。然而，它们使用的随机性并非相互独立，而是彼此干扰。具体而言，我们运行各个基准算法的天数取决于每个基准算法过去的损失序列，而是否运行某个算法的决定会影响这些基准算法未来的表现（这一问题的示意图见图 1a）。这个问题是本文的部分信息设置所特有的。在完全信息设置下，无论外层乘法权重更新是否选择某个基准算法，所有基准算法都会在每天结束后获得更新，因此我们可以独立分析每个基准算法。但在部分信息设置下情况并非如此，克服这一依赖问题是一项重大的算法挑战。

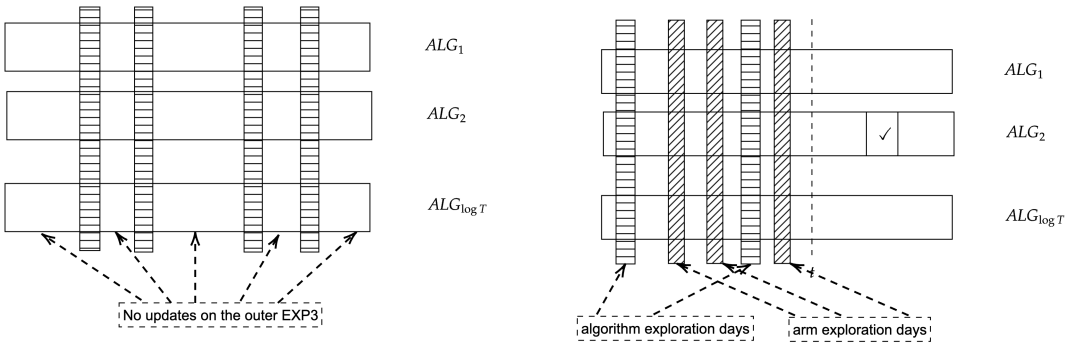
消除双查询算法的依赖问题。我们处理依赖问题的主要思路，是将内层和外层 EXP3 的更新提炼为“探索”和“利用”两个步骤。我们首先考虑构成基线算法的内层 EXP3 子程序。在探索阶段，我们通过每天使用第二次查询均匀随机地采样一位专家，非自适应地估计损失，并利用其反馈形成估计值。在利用阶段，我们

~以与 exp(−$\mathcal{L}(i)$) 成正比的概率采样每位专家（即臂 i的估计损失），但不更新权重。这种方法也可以推广到外层 EXP3 算法的更新中，该算法以基线算法为操作对象。具体而言，每天我们以概率 1 2 使用第二次查询采样一个基线算法来估计其代价。采样基线算法的概率同样与估计损失的指数成正比。渐近地，对于持续时间为$\mathcal{D}$ 的区间，专家和算法的损失估计方差仍为 $\sqrt{|\mathcal{D}|}$（忽略 n polylog(nT) 因子），因此好时期的遗憾至多为$\sqrt{B}$，周期长度为 B。这里的关键在于，一旦固定第二次查询的随机性，基线算法的更新就变成确定性的。因此，我们可以采用相同的递归提升，并在经过 log(nT) 层后获得近乎最优的 $\sqrt{T}$ 遗憾。最后，我们可以实施一种更具技术性的提升过程，进一步优化遗憾中对 n 的依赖。

示意 我们讨论的依赖问题相关思路可见图 1 .



(a) 过去几天使用的随机性会干扰基线算法的损失。(b) 将学习日与利用日分开。EXP3 算法仅根据损失估计进行更新。



(c) 将学习日与利用日（或探索日）分开后，接下来的一天运行的每个基线算法都是固定的，因此可以控制遗憾。(d) 固定外层 EXP3 的第二次查询所用的随机性。外层 EXP3 算法仅根据损失估计进行更新。

图 1：依赖问题及克服该问题的思路示意图。

从双查询框架到滑动窗口算法。现在我们讨论如何以低内存模拟 [LZC+24] 中的方法，以及如何实现滑动窗口算法

，使得整个算法每天只需进行两次查询。在较高层次上，[LZC+24] 中的算法包含两个组成部分：区间算法，即针对长度为 I 的区间 $|I| = 2^s$，在 $s \in \log T$ 下定制的 EXP3 算法；以及“外层算法”，它控制用于运行各个区间算法的权重。直观地说，这两层的交互方式与之前讨论的递归提升类似：每天，外层算法会抽取一个区间算法，并按照其分布选择一个臂。此外，对各区间算法权重的更新通过第二次查询完成，该查询从方差小得多的分布中抽样，例如专家上的均匀分布。这确保了某个区间算法的方差不会显著影响其他算法。最后，外层算法按照 [DGS15]中用于区间遗憾的标准方法，调整各区间算法的权重。

这里的关键观察是，滑动窗口算法的算法框架依赖于对损失序列进行低方差估计，而我们先前的框架尤其适合这一性质。因此，我们能够以白盒方式证明：由于第二次查询足以对损失进行低方差估计，每个区间算法仍然满足我们在提升算法中得到的性质。 მედეგად，每个区间算法都能在区间 I 上实现 $\sqrt{|I|}$ 遗憾，并使用多重对数(nT)空间，这也给出了滑动窗口模型中的遗憾界 $\sqrt{W}$。最后，由于最多有$\mathcal{O}(\log T)$个区间算法，且每个算法使用多重对数(nT)位内存，因此最终的内存上界仍为多重对数(nT)。

2.3 单查询设定下的流式算法

现在我们描述单查询多臂老虎机设定下的框架，在该设定中，我们获得具有 $T^{2/3}$ 遗憾的流式算法。该算法遵循对专家池进行采样和剪枝的相同框架。然而，这里一个重大挑战是，对于一个已在专家池中存在持续时间$\mathcal{D}$的专家，我们不再拥有第二次查询来提供具有 $\sqrt{|\mathcal{D}|}$ 误差的估计值。因此，我们需要分配可能产生遗憾的探索日，以估计损失序列。令 $\gamma \in (0,1)$ 为 EXP3 中的探索因子，对于持续时间$\mathcal{D}$，损失估计的加性误差被$\mathcal{O}\left(\sqrt{|\mathcal{D}|}/\gamma\right)$所界定。我们可以将这一加性误差纳入驱逐规则，使得长度为 B 的好时期具有$\mathcal{O}\left(\sqrt{B}/\gamma\right)$遗憾。另一方面，单查询设定会在全部 T 天中，以概率 γ 进行探索，从而产生 γT 加性遗憾。对于长度为 B 的时期，总体遗憾的形式为 $\left(\frac{T}{B} \cdot \sqrt{B}/\gamma + nB + \gamma T\right) \cdot$ 多重对数(nT)。通过平衡 $\gamma = 1/B^{1/3}$ 和 $B = T^{3/4}$，我们大致得到 $T^{3/4}$ 遗憾。

通过更好的平衡来进行提升。为了进一步提高单查询算法的性能，一种自然的做法是模拟我们为双查询算法开发的递归提升过程。然而，我们不再拥有用于估计专家和基线算法损失序列的“免费”第二次查询，因此需要“牺牲”一些探索日，并为这些天承担额外的遗憾。具体而言，我们再次以某个概率 γ在每天进入探索日。在探索日，我们采用与双查询算法相同的策略，为基线算法及其维护的所有专家维护估计损失，即以$1/2$的概率均匀随机地采样一位专家，并以$1/2$的概率均匀随机地采样一种基线算法。在非探索日，我们通过依据

估计损失对每种基线算法和专家进行采样来利用，但不更新权重。

通过这种方式，如果我们在长度为 D 的持续时间$\mathcal{D}$内以概率 γ 进行探索，就能将专家和基线算法的损失估计维持在加性误差至多为 $\sqrt{D/\gamma}$ 的范围内。通过更好的权衡可知，若以 $\gamma = 1/T^{1/3}$ 的速率进行探索，则以高概率，加性误差为 $\sqrt{D \cdot T^{1/6}} \leqslant T^{2/3}$。此外，探索步骤带来的遗憾为 $\gamma D \leqslant T^{2/3}$。因此，我们可以应用递归提升过程 $\log(nT)$ 次，并以 $\text{polylog}(nT)$ 的内存获得最终遗憾界 $T^{2/3}$，使遗憾和内存都显著优于单查询多臂老虎机设定下先前最先进水平。

随机顺序最佳专家。为了在最佳专家处于随机顺序时进一步将遗憾改进至 $\sqrt{T}$，我们基于第一性原理设计了一种不同的算法，不再沿用先前基于采样和驱逐的框架。其子程序明显简单得多。

从较高层面来看，该算法首先通过二分搜索来确定最佳专家的“正确”误差率。设 $\gamma\sqrt{T}$ 为最佳专家的损失，即损失率为 $\gamma/\sqrt{T}$，并设 $C\sqrt{T}$ 为我们在二分搜索中要达到的当前目标损失，其中 $1 \leqslant C \leqslant \sqrt{T}$。为了建立直觉，我们先考虑一种理想化情形：在任何区间内，其他所有专家都远不如最佳专家。因此，我们只需轮询专家，检查每位专家的误差率是否低于$\mathcal{O}\left(C/\sqrt{T}\right)$。如果误差率过高，就淘汰该专家；如果没有专家达到所需的误差率，我们就可以增大 C 并再次检查。另一方面，如果找到了误差率较低的专家，我们就可以在剩余的天数里一直选择该专家。这个过程只需要$\mathcal{O}(\log\ T)$ 位空间。

在非理想化情形下，分析会变得更加复杂，尤其是当次优专家可能经历一段预测正确的“连胜期”时。在这里，次优专家的误差率可能会在一小段时间内低于 $C/\sqrt{T}$，随后开始产生高得多的损失。在这种情况下，简单地坚持使用该专家可能会在之后的天数中造成很高的遗憾。这个问题可以通过动态检查当前专家的误差率是否仍然令人满意来处理。如果某位专家在短时间内展现出远优于最佳专家的误差率，随后却出现糟糕得多的误差率，那么算法可以淘汰该专家并继续这一过程。遗憾仍然很低，因为该专家有许多天的表现优于最佳专家；换言之，我们可以计入“反向遗憾”。当我们检查到误差率“正确”的最佳专家后，过程便会停止：由于该专家的损失序列是随机顺序的，我们永远不会淘汰最佳专家。随后，通过对该算法进行摊还遗憾分析，可以得到其遗憾界为 $\sqrt{T}$。

3 预备知识

记号。对于正整数 $n > 0$，我们使用记号 $[n]$ 表示集合 {1、2、.....、n}。我们使用记号 $\text{poly}(n)$表示关于 n的固定多项式，其次数可以通过设置适当的常数来推断。特别地，我们使用记号 $\text{polylog}(n)$表示 $\text{poly}(\log n)$。如果一个随机事件的失败概率为 $1 - {}^{1}_{\text{poly}(nT)}$，则称该事件以高概率发生。

我们使用 $\ell^t(i)$ 表示臂 i 在第 t 天的损失。因此，ℓ^t 是第 t 天的 n 维损失向量。对于臂 i 在持续时间 $\mathcal{D}$ 内的累积损失，我们使用 $\mathcal{L}^{\mathcal{D}}(i)$。我们还将区间记为 $\mathcal{I}$，并将专家 i 在区间 $\mathcal{L}^{\mathcal{I}}(i)$ 上的损失记为后者；具体使用哪种表示取决于上下文。一般来说，我们使用 ℓ 表示单天的损失，并使用$\mathcal{L}$ 表示累积损失

（跨越多天）。此外，我们有时将$\mathcal{L}^{\mathcal{D}}$(ALG) 写作一个函数，用于表示算法 ALG 在持续时间$\mathcal{D}$内的代价。

3.1 模型的形式化描述

我们将对问题的描述给出一个更形式化的版本。首先介绍具有一般最坏情况损失序列的在线学习问题。

使用查询和 q 空间进行 s 专家在线学习。我们考虑一个包含 n 个专家、持续 T 天的在线学习问题。结果序列由无视对手决定：对于每一天 $t \in [T]$，每个专家的损失 $\ell^t(i) \in \{0、1\}$ 都预先固定，且算法并不知道。在线学习算法 ALG 在每一天 $t \in [T]$ 依次与实例交互，具体如下。

(1) **算法决策**：在线学习算法 ALG 选择一位专家 $i_t \in [n]$ 进行跟随，即输出与该专家相同的预测。

(2) **承受损失**：算法承受所选专家的损失 $\ell^t(i_t)$。

(3) **反馈信号**：作出选择后，算法会观察所选专家的损失，以及另外一组（$q-1$）专家的损失。ALG 可以免费观察这另外（$q-1$）位专家的损失。

目标是设计一种算法，使其在以下目标下实现最小化：

(a). **累积遗憾，**即在这 T 天内， ALG 的损失与事后表现最佳的单个专家的损失之差，即

$$\sum_{t=1}^{T} \ell^t(i_t) - \min_{i\in[n]} \sum_{t=1}^{T} \ell^t(i).$$

(b). **滑动窗口遗憾**。在每天 t，我们考察区间 $[t-W+1,t]$ 内的损失之差，并计算 ALG 的损失与 $[t-W+1,t]$ 的事后最佳臂的损失之间的差距。换言之，时间步 t 的滑动窗口遗憾定义为

$$\sum_{s=t-W+1}^{t} \ell^s(i_s) - \min_{i\in[n]} \sum_{s=t-W+1}^{t} \ell^s(i).$$

我们将滑动窗口模型中的遗憾定义为任意 $t \in [T]$ 对应的滑动窗口遗憾的最大值。

(c). **区间遗憾**。类似于滑动窗口的情况，对于区间$\mathcal{I} = [t_1,t_2] \subseteq [T]$，区间遗憾定义为

$$\sum_{t=t_1}^{t_2} \ell^t(i_t) - \min_{i\in[n]} \sum_{t=t_1}^{t_2} \ell^t(i).$$

与滑动窗口遗憾相比，区间遗憾不要求任何固定大小。我们将区间遗憾模型中的遗憾定义为任意区间的最大遗憾$\mathcal{I} = [t_1,t_2] \subseteq [T]$。

我们还旨在优化空间复杂度 s，其定义为ALG 在执行过程中任意时刻所使用的最大单词数。

随机顺序最佳专家损失序列。 对于在线学习中包含 n 个专家和 T 天的累积遗憾最小化问题，我们将随机顺序最佳专家情形定义如下。令 i^* 为达到最小代价的专家，即：

$$i^* = \operatorname*{argmin}_{i\in[n]} \sum_{t=1}^{T} \ell^t(i).$$

然后，我们对 $[T]$ 中专家 i^* 的损失进行随机排列。换言之，令σ：$2^{[T]} \to 2^{[T]}$为 T 上的均匀随机排列，我们在在线学习开始之前令 $\ell^t(i^*) \leftarrow \ell^{\sigma(t)}(i^*)$。

3.2 集中不等式

我们给出本文中使用的标准集中不等式。

命题 3.1 (伯恩斯坦不等式，[Ber24])。令 X_1，X_2，……，X_n 表示相互独立的随机变量，且满足 $|X_i - \mathbb{E}[X_i]| \leqslant M$对所有 $i \in [n]$均成立。令 $S_n = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$ 并令$\sigma^2 = \sum_{i=1}^{n}$方差$(X_i)$。那么，对于任意 $t > 0$:

$$\mathbf{Pr}\left[|S_n| \geqslant t\right] \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{2}{3}Mt}\right).$$

3.3 EXP3 算法与带探索的学习

在单臂赌博机信号的设定下，实现最优遗憾的标准算法是探索与利用指数权重算法，简称 EXP3 算法 [ACFS02]。该算法的描述如下：算法 1。

EXP3 算法的标准保证如下。

命题 3.2。 *[ACFS02]* 令 $\gamma \in (0, \frac{1}{\sqrt{T}})$，算法 1的期望遗憾至多为$\mathcal{O}\left(\sqrt{nT\log n}\right)$。

EXP3 的“带探索的学习”视角。算法 1的原始版本显式维护每个臂的权重 $w_t(i)$。另一种看待 EXP3 算法的方式是：以 $\gamma \in (0, 1]$, 的概率均匀随机地抽样，否则按损失成比例地抽取臂。重要的是，算法的“学习”，即损失的更新，仅在探索日进行。该算法可描述为算法 2。

我们对算法 2的遗憾保证如下。

命题 3.3。 设 $\gamma \in (0, \frac{1}{\sqrt{T}})$，算法 2 的期望遗憾至多为 $(n \cdot \sqrt{T}/\gamma + \gamma \cdot T) \cdot$ 多重对数(nT)。

我们认为命题 3.3 在文献中已有论述；无论如何，为求完整，我们在A 节 给出证明。此外，命题 3.3 推出了以下引理。

算法 1 探索与利用指数权重算法（EXP3）算法**输入：** n 臂的集合；参数为 T 天。**输入：** 探索率为 $\gamma \in (0,1]$。

将 $w_t(1)\ w_t(2), \cdots\ w_t(n)$ 作为 n 臂的权重。初始化为所有 $i \in [n]$ 设置 $w_1(i) \leftarrow 1$ 。**对于**每一天 t **执行**

从分布 P_t 中采样，使得对每个 $i \in [n]$ 都满足条件。

$$P_t(i) = (1-\gamma) \cdot \frac{w_t(i)}{\sum_{i=1}^{n} w_t(i)} + \frac{\gamma}{n}.$$

令 $i_t \sim P_t$ 为抽样臂，并观测损失 $\ell^t(i_t)$。
对于每个臂 $i \in [n]$ **执行**
损失估计如下。

$$\widetilde{\ell}^t(i) = \begin{cases} \ell^t(i)/P_t(i), & \text{if } i = i_t \\ 0, & \text{otherwise} \end{cases}$$

更新 $w_{t+1}(i) = \exp\left(-\gamma \cdot \widetilde{\ell}^t(i)\right) \cdot w_t(i)$。

算法 2 EXP3 的“以学习作为探索”版本

输入： n 个臂的集合； T 天的参数。
输入： 探索率为 $\gamma \in (0,\ 1/2]$。

为每个 $i \in [n]$ 和 $t \in [T]$ 维护估计损失 $\mathcal{L}^{1:t}(i)$。
对于每一天 t **执行**
以 γ 的概率进入探索日。
如果 t 是探索日**则**
从臂集合中均匀随机地采样 i_t 。

$$\widetilde{\ell}^t(i) = \begin{cases} n \cdot \ell^t(i)/\gamma, & \text{if } i = i_t \\ 0, & \text{otherwise} \end{cases}$$

更新 $\widetilde{\mathcal{L}^{1:t+1}}(i_t) \leftarrow \widetilde{\mathcal{L}^{1:t}}(i_t) + \widetilde{\ell}^t(i_t)$。
else
从以下分布中采样一个臂 i_t 。

$$P_t(i) = \frac{\exp(-\gamma \cdot \widetilde{\mathcal{L}^{1:t}}(i))}{\sum_{i=1}^{n} \exp(-\gamma \cdot \widetilde{\mathcal{L}^{1:t}}(i))}. \tag{4}$$

推论 5。令 $\{\ell^t\}_{t=1}^{T}$ 为损失向量集合，令 $L^{\mathcal{D}}$ 为持续时间 D 内的损失估计 r

D。假设

$$\left|\widetilde{\mathcal{L}^{\mathcal{D}}}(i)-\sum_{t\in\mathcal{D}}\ell^t(i)\right|\leqslant\sqrt{|\mathcal{D}|/\gamma}\cdot\operatorname{polylog}(nT).$$

然后，在 每一步 都按照 D 进行采样 $\sqrt{|\mathcal{D}|/\gamma}\cdot n$ 多重对数(nT)，可得到遗憾

使用一次额外查询。 在多信号查询模型中，我们可以观测多个臂的损失，因此能够获得与 算法 2 类似（但更强）的保证。具体而言，我们可以每天进行探索，从而得到 算法 3。

算法 3 EXP3，进行两次查询

输入： n 个臂（专家）的集合； T 天的参数。

输入： 探索率为 $\gamma\in(0,1]$; 学习率为 η。

为每个臂$^{1:t}(i)$和 $i\in[n]$ 维护估计损失。**对于**每一天 t

执行从以下分布中采样一个臂 i_t 并选择它。 i_t

$$P_t(i)=\frac{\exp(-\eta\cdot\widetilde{\mathcal{L}^{1:t}}(i))}{\sum_{i=1}^n\exp(-\eta\cdot\widetilde{\mathcal{L}^{1:t}}(i))}.$$

从臂集合中均匀随机地抽取 i_t^{est}，但不要选择 i_t^{est}。

▷由于我们没有进行选择，因此没有产生损失 i_t^{est}

$$\widetilde{\ell^t}(i)=\begin{cases}n\cdot\ell^t(i), & \text{if } i=i_t^{\text{est}}\\ 0, & \text{otherwise}\end{cases}$$

更新$\mathcal{L}^{1:t+1}(i_t^{\text{est}})\leftarrow\mathcal{L}^{1:t}(i_t^{\text{est}})+\ell^t(i_t^{\text{est}})$。

算法 3 的保证可直接由已有研究得出。以下命题的证明见A 节。

命题 3.4。 算法 3 的期望遗憾至多为 $O\left(n\cdot\sqrt{T\log n}\right)$，其中使用 $\eta=\frac{1}{n}\cdot\sqrt{\log_T n}$。

3.4 SQUINT算法

我们使用 SQUINT 算法进行最优提升，如 [PR23] 所示。对于某个算法，其在第 p_t 天 t 在臂（专家）上的分布为 $v_t(i)=\sum_{i=1}^n p_t(i)\ell^t(i)-\ell^t(i)$，定义为专家 i 与该算法之间的损失差。SQUINT 算法（第 3.4 节）及其保证如下。

算法4 SQUINT， [KvE15],参见 [PR23]

对于 $t=1,2,\ldots,T$**执行**：计算满足 $p_t(i)\propto\mathbb{E}_\eta[\eta\cdot\exp(\eta\sum_{\tau=1}^{t-1}v_\tau(i)-\eta^2\sum_{\tau=1}^{t-1}v_\tau^2(i))]$ 的专家上的 $p_t\in\Delta_n$：抽取一位专家 $i_t\sim p_t$，并观察损失向量 $\ell_t\in[0,1]^n$

引理 3.1。 *[KvE15]* 设学习率 η 从先验分布 $\gamma(\eta)$ 上抽取，其中$[0,1/2]$满足 $\gamma(\eta) \propto O\left(\eta^{\frac{1}{2}\log(\eta)}\right)$。对于任意专家 $i \in [n]$，~~SQUINT~~ 保证

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(i_t) - \sum_{t=1}^{T} \ell_t(i)\right] \leqslant \mathcal{O}\left(\sqrt{V_T^i \log(nT)}\right).$$

其中 $V_T^i = \sum_{t=1}^{T} (\mathbb{E}_{i_t}[l_t(i_t)] - l_t(i))^2$ 是专家 i 的损失方差。

4 我们的基线：采用单查询赌博机损失的低内存算法

正如我们在第 2 节中讨论的，对于滑动窗口模型和流模型，我们的算法基于 [PR23] 中的算法的一个变体，该算法适用于单查询多臂老虎机设定。该算法以高概率实现 $\widetilde{O}(nT^{3/4})$ 遗憾，且所需空间为 $\mathrm{polylog}(nT)$。能够实现 $T^{3/4}$ 遗憾的最佳算法是 [XZ21] 中的算法，其内存需求为 $\Theta(\sqrt{n})$。因此，该算法本身就已将多臂老虎机设定下亚线性内存在线学习的内存效率提升了近乎指数倍。

高层次概述与必要概念。 本节将对我们的算法作高层次介绍。粗略地说，该算法将 [PR23] 的框架扩展到了多臂老虎机设定。我们已在第 2 节讨论过一种简化的分析；不过请注意，若直接应用该分析框架，会导致严重的依赖问题：为了定义“能够淘汰最佳专家的时期”，我们需要固定专家池中各专家的随机性；然而，如果最佳专家实际上加入了池中，它就会干扰池中的其他专家，从而使论证陷入循环。

为了解决这类依赖问题，我们需要采用与 [PR23]相同的结构，以维护池的迭代合并过程。具体而言，池被划分为 $K = \log T$ 个池 $\mathrm{P} = \mathrm{P}_1 \cup \mathrm{P}_2 \cdots \cup \mathrm{P}_K$，其中第 k 个子池包含最近 2^k 个周期内加入的专家。每经过 2^k 个周期，池 P_k 就会合并到池 P_{k+1}，而我们将通过移除已被覆盖的专家来进行剪枝，也就是说，移除其损失序列不如基准损失的专家。在这次合并期间，我们将抽取 $1/\mathrm{polylog}(n)$ 比例的专家组成过滤集 F，最终，“能够驱逐”的定义将依据 F 中专家的基准损失来确定。

现在，我们介绍基准损失正式定义所需的概念。当我们谈到专家 $i \in [n]$ 的生命周期（或持续时间）$\mathcal{D}(i)$时，指的是一个区间 $(t_1, t_2]$，其中 t_1 和 t_2 都是某个参数 B 的整数倍。在上下文明确时，我们用$\mathcal{D}$ 作为简写。由于我们采用部分信息设置，因此必须维护损失的估计。对于给定的区间$\mathcal{D}$ 和专家 i，我们用$\widetilde{\mathcal{L}^{\mathcal{D}}(i)}$表示专家 i 在持续时间$\mathcal{D}$内损失的估计值。容易看出，由于损失取值在$\{0,1\}$中，$\left|\widetilde{\mathcal{L}^{\mathcal{D}}(i)} - \sum_{t\in\mathcal{D}} \ell^t(i)\right| \leqslant |\mathcal{D}|$。为了得到遗憾界，我们最终需要精度更高的估计。

基准损失。 现在，我们可以正式定义基准损失。为了更好地说明如何得到基准损失，我们将其写成算法的形式（算法 5）。

算法 5 专家（臂） i的动态基准输入：过滤专家集合 F。输入：周期长度 B；生命周期 D $=(t_1,$ $t_2]$ 是 B 的整数倍。输入：每位专家 $j \in$ F 的每轮估计成本。输出：成本基准。1：令 $t(j_1) \leqslant t(j_2) \leqslant t(j_3) \leqslant \cdots \leqslant t(j_{|\mathcal{F}|})$ 按过滤专家进入池的时间排序。2：将生命周期 D 划分为区间 $\mathrm{I}_1, \mathrm{I}_2, \cdots,$ $\mathrm{I}_{\lceil \mathcal{D}|/B}$。3：初始化 $\mathrm{L}^{\mathcal{D}}$(BM) $\leftarrow 0$。4：对于每个区间 I_ℓ for $\ell \in [1, |L|/B]$ 执行 5：令 I($t(j_{k-1})$ $t(j_k)$) 为满足 $t(j_{k-1}) \leqslant (t_1 + (\ell-1) \cdot B) \leqslant t(j_k)$ 的时间区间。6：$\mathrm{L}^{\mathcal{D}}$(BM) $\leftarrow \min_{j \in \mathcal{F}} \widetilde{\mathrm{L}}^{\mathcal{I}(t(j_{k-1}) t(j_k))}(j) + \widetilde{\mathrm{L}}^{\mathcal{D}}$(BM)。7：返回 $\mathrm{L}^{\mathcal{D}}$(BM)。

在算法 5中，过滤集$\mathcal{F}$ 由算法 8生成。周期长度是算法 6的一个参数，而专家的生命周期是算法 6的一个函数。与 [PR23], 相比，我们仅使用损失的估计来计算基准损失，维护估计损失的方法见算法 6。

现在，我们正式定义覆盖这一概念，其最初定义见 [PR23]。或许与最初的直觉相反，我们会更积极地驱逐臂。覆盖的定义如下。

定义 1 (专家的近似覆盖)。给定一个$\mathcal{F}$ 过滤臂集合以及一个指数$\rho \in (0,\ 1)$。若一个臂 i 的生命周期为$\mathcal{D}$，则称其被覆盖，如果使用算法 5 并取 i 和$\mathcal{F}$ 运行后得到$\mathcal{L}^{\mathcal{D}}$(BM)，且满足

$$\widetilde{\mathcal{L}^{\mathcal{D}}(i)} \geqslant \widetilde{\mathcal{L}^{\mathcal{D}}(\mathsf{BM})} - C \cdot n \log(nT) \cdot (|\mathcal{D}|)^{\rho},$$

其中 C 是一个绝对常数。

现在，我们可以正式介绍基线算法，如算法 6所示。其控制流程如下：在每个时期，算法以$\mathcal{P}_1$ 的概率$1/n$将臂（专家）采样到池中。然后，算法使用 $\gamma = (\frac{n}{B})^{1/3}$ 作为该时期的探索参数来执行 EXP3。时期结束时，如果当前是第$\mathcal{P}_k$ 个时期，池$\mathcal{P}_{k+1}$ 将合并到 2^k。合并过程由合并算法（算法 7）控制，该算法又会调用算法 8 来确定覆盖与驱逐。

本节余下部分将证明算法 6的内存和遗憾。

4.1 分析：多重对数(nT) 内存和 $\widetilde{O}(nT^{3/4})$ 遗憾

与 [PR23], 的情况类似，我们首先分析内存复杂度，因为遗憾分析依赖于此。下一个引理表明，合并算法是高效的，这构成了我们内存效率分析的基础。

算法 6 采用单查询老虎机信号的基线算法

参数：B 作为每个时期的长度；γ 作为 EXP3 中用于探索的概率。

维护一个臂池 $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \mathcal{P}_K$，其中 $K = \log T$。
对于每个臂 $i \in \mathcal{P}$，令 $D(i)$ 表示池中 i 的生命周期。
将时间范围划分为 T/B 个时期。
为池 $\mathcal{P}$ 中的臂维护每个时期的估计成本。
对每个时期 $\tau \in [T/B]$ **执行**
    以概率采样每个臂，并将其加入$\mathcal{P}_1$ $\frac{1}{n}$.
    对池 $\mathcal{P}$ 中的所有臂运行 EXP3，参数为 $\gamma = (\frac{n}{B})^{1/3}$。 −
    对于时间专家所诱导的每个区间 $\mathcal{I}$，将其加入池中**执行**
        维护 $\widetilde{\mathcal{L}^{\mathcal{I}}}(i) = \sum_{t:i_t=i} \widetilde{\ell^t}(i)$，对每个 i，$\ell^t(i)$，其中 $\widetilde{\ell^t}(i)$ 是 EXP3 的伪损失
    如果 $\tau = C \cdot 2^{C'}$ 对于某个整数 C，C' **则**
        令 $\mathrm{pw}(\tau)$ 为满足以下条件的最大整数：对于某个整数 C，$\tau = C \cdot 2^{\mathrm{pw}(\tau)}$。
        对于 $k \in [\mathrm{pw}(\tau)]$，**执行**
            更新 $\mathcal{P}_{k+1}$，使用算法 7将其合并为$\mathcal{P}_{k+1}$ 和 $\mathcal{P}_k$。
            $\mathcal{P}_k \leftarrow \emptyset$。

算法 7 合并($\mathcal{P}_A$，$\mathcal{P}_B$)

输入：池 $\mathcal{P}_A$ 和$\mathcal{P}_B$。

参数：$Q = 16 \log(nT)$。

初始时令 $\mathcal{P}_C = \mathcal{P}_A \cup \mathcal{P}_B$。
对于 $q = 1 : Q$，**执行**
    通过为$\mathcal{P}_C$ 中的每个臂分配一个伯努利($\frac{1}{\log^4(nT)}$)随机变量来估计$\mathcal{P}_C$的大小。
    如果 $s \leqslant \log^5(nT)$ **则**
        返回池$\mathcal{P}_C$。
    else
        采样 $\frac{1}{\log^4(nT)}$-臂中用于获得过滤集 $\mathcal{F}$ 的比例。
        运行算法 8，使用$\mathcal{P}_C$ 和$\mathcal{F}$ 得到集合$\mathcal{X}$。
        $P_C = F \cup X$。

算法 8 过滤器($\mathcal{F}$，$\mathcal{Q}$)

输入：池 $\mathcal{Q}$ 和臂过滤集 $\mathcal{F}$。

对于每个臂 $i \in \mathcal{Q}$ **执行**2: 如果 i 被$\mathcal{F}$ 覆盖（如定义1所述），且 $\rho = 2/3$，则从$\mathcal{Q}$中移除 i。
3: 返回更新后的$\mathcal{Q}$集合。

引理 4.1。以概率 $1 - {}^{1}_{\mathrm{poly}(nT)}$，~~*Merge*~~$(P_A, P_B)$ 的输出 P_C满足

$$|\mathcal{P}_C| \leqslant \max\left(2\log^9(nT), \frac{1}{4}(|\mathcal{P}_A| + |\mathcal{P}_B|)\right).$$

证明。假设过滤器已针对 $q_{\max} \in [Q]$ 次调用过，对于 $q \in [q_{\max}]$，令 $\mathcal{P}_{C,q}$ 表示第 q次开始时的池 $\mathcal{P}_C$。令 $\mathcal{E}$ 表示以下事件：每位已在池中停留 i 的专家的代价

都被正确估计到乘法 $\left(1+\mathcal{O}\left(\frac{n}{\log(nT/(1-\rho))}\right)\right)^{|\mathcal{D}|}$近似范围内。我们将证明，在以 $\mathcal{E}$ 为条件时，以至少 $1-\frac{1}{\mathrm{poly}(nT)}$的概率，以下情况之一成立：

- $|\mathcal{P}_{C,q}| \leqslant 2\log^9(nT)$
- $|\mathcal{P}_{C,q+1}| \leqslant \left(1-\frac{1}{6\log(nT)}\right)\cdot|\mathcal{P}_{C,q}|$.

注意，随后对 $k\in[k_{\max}]$ 取并集界即可推出这样的结论：

$$\begin{aligned}|\mathcal{P}_{C,q_{\max}+1}| &\leqslant \max\left(2\log^9(nT), \left(1-\frac{1}{6\log(nT)}\right)^{16\log(nT)}\cdot|\mathcal{P}_{C,1}|\right)\\ &\leqslant \max\left(2\log^9(nT), \frac{1}{4}(|\mathcal{P}_A|+|\mathcal{P}_B|)\right),\end{aligned}$$

因为 $|\mathcal{P}_{C,1}| = |\mathcal{P}_A|+|\mathcal{P}_B|$。

为证明该断言，我们固定 $q\in[q_{\max}]$，并假设由此可得 $|\mathcal{P}_{C,q}| > 2\log^9(nT)$。因此，只需证明 $|\mathcal{P}_{C,q+1}| \leqslant \left(1-\frac{1}{6\log(nT)}\right)\cdot|\mathcal{P}_{C,q}|$ 以高概率成立。注意，根据标准的 Chernoff界，估计大小 s 满足

$$\mathbf{Pr}\left[s\leqslant\log^5(nT)\right]\leqslant\frac{1}{\mathrm{poly}(nT)}.$$

现在，我们令 $M=|\mathcal{P}_{C,q}|$，并初始化 $\mathcal{W}_1=\mathcal{P}_{C,q}$。对于每个 $r\geqslant 1$ 以及每个专家 $i\in\mathcal{W}_r$，我们定义

$$V_r(i)=\left\{j\in\mathcal{W}_r\setminus\{i\}\,|\,\mathcal{L}^{\mathcal{I}(t(i),t(j_{r-1}))}(i)\geqslant\mathcal{L}^{\mathcal{I}(t(i),t(j_{r-1}))}(j)-2^{r-1}\right\},$$

因此，$V_r(i)$ 是一组专家集合，它们将在区间 $\mathcal{I}(t(i),t(j_{r-1}))$ 内覆盖 i。我们还定义

$$R_r=\{i\in\mathcal{W}_r\,|\,|V_r(i)|\geqslant\log^6(nT)\},$$

因此，R_r 是专家 $i\in\mathcal{W}_r$ 的集合，这些专家的 $V_r(i)$ 规模较大。我们最终将论证，这些专家很容易被我们的算法移除，因为它们很可能被过滤器中的专家 $j_1,\ldots,j_{r-1}$ 与来自 V_r 的某个专家共同覆盖。

令 j_r 为 $\mathcal{W}_r\setminus R_r$ 中最近进入池且也被采样到过滤器 $\mathcal{F}_q$ 中的专家，使得

$$j_r=\operatorname*{argmax}_{j\in(\mathcal{W}_r\setminus R_r)\cap\mathcal{F}_q} E_j,$$

其中 E_j 是专家 j 的进入时间（如果存在这样的专家）。否则，我们将 j_r 留为未定义。如果 j_r 有定义，我们将集合 O_r 定义为进入池的时间晚于 j_r 的专家，使得

$$O_r=\{i\in\mathcal{W}_r\setminus R_r\,|\,E_i\geqslant E_{j_r}\}.$$

最后，对于每个 $r\geqslant 1$，我们定义 $\mathcal{W}_{r+1}=\mathcal{W}_r\setminus(R_r\cup O_r\cup V_r(j_r))$。

归纳基例。 令 $r \geqslant 1$，并令 $\mathcal{E}_1$ 表示以下事件：对于所有 $m \in [r]$, $R_m \leqslant \frac{1}{4 \log(nT)} \cdot M$；这里回顾一下， $M = |\mathcal{P}_{C,q}|$。我们通过归纳法证明

(1) $|\mathcal{W}_{r+1}| \geqslant \left(1 - \frac{2}{} r_{\log(nT)}\right) \cdot M$(2) 我们有 $j_1, \ldots$

, $j_r \in \mathcal{F}_q$ 并且 $j_r \preceq j_{r-1} \preceq \ldots \preceq j_1$(3) 对于任意

$m \in [r]$ 和专家 $i \in \mathcal{W}_{r+1}$，我们有

$$\mathcal{L}^{\mathcal{I}(t(j_m),t(j_{m-1}))}(i) \geqslant \mathcal{L}^{\mathcal{I}(t(j_m),t(j_{m-1}))}(j_m) + 2^{m-1}.$$

我们注意到， $r = 0$ 的基础情形立即成立，并假设该陈述对所有不超过 $r-1$ 的情况均成立。

归纳步骤。 现在我们根据 R_r 的大小进行分类讨论。

情况 1。在 $|R_r| \geqslant \frac{1}{4} \log(nT) \cdot M$ 的情况下，对于每个专家 $i \in R_r$，相应的每个专家 $j \in V_r(i)$ 以 $\frac{1}{\log^4(nT)}$ 的概率被抽入过滤集 F_q。因此，没有专家 $j \in V_r(i)$ 被抽入 ~~F_q~~ 的概率至多为 $\left(1 - \frac{1}{4 \log(nT)}\right)^{\log^6(nT)} \leqslant \frac{1}{\mathrm{poly}(nT)}$，因为根据定义，对于任意 $i \in R_r$， $|V_r(i)| \geqslant \log^6(nT)$。因此，以高概率，存在一个专家 $j \in V_r(i) \cap \mathrm{F}_q$。接着，在以 E 为条件的情况下，我们有：

$$\begin{aligned}
\mathcal{L}^{\mathcal{I}(t(i),t(j_0))}(i) &= \mathcal{L}^{\mathcal{I}(t(i),t(j_{r-1}))}(i) + \sum_{m=1}^{r-1} \mathcal{L}^{\mathcal{I}(t(j_m),t(j_{m-1}))}(i) \\
&\geqslant \mathcal{L}^{\mathcal{I}(t(i),t(j_{r-1}))}(i) + \sum_{m=1}^{r-1} \left(\mathcal{L}^{\mathcal{I}(t(j_m),t(j_{m-1}))}(i)\right) - n \log(nT) \left(\frac{t(j_{m-1}) - t(j_m)}{B}\right)^{\rho} \\
&\geqslant \left(\mathcal{L}^{\mathcal{I}(t(i),t(j_{r-1}))}(j) - 2^{r-1}\right) + \sum_{m=1}^{r-1} \left(\mathcal{L}^{\mathcal{I}(t(j_m),t(j_{m-1}))}(j_m) + 2^{m-1}\right) - n \log(nT) \left(\frac{t(j_{m-1}) - t(j_m)}{B}\right)^{\rho} \\
&\geqslant \sum_{m=1}^{r-1} \left(\mathcal{L}^{\mathcal{I}(t(j_m),t(j_{m-1}))}(j_m)\right) + \mathcal{L}^{\mathcal{I}(t(i),t(j_{r-1}))}(j) - 1 - n \log(nT) \left(\frac{t(j_{m-1}) - t(j_m)}{B}\right)^{\rho}.
\end{aligned}$$

再次假设 在事件$\mathcal{E}$发生的条件下，每位专家的损失都得到了良好的近似，我们有

$$\mathcal{L}^{\mathcal{I}(t(i),t(j_0))}(i) \geqslant \sum_{m=1}^{r-1} \left(\mathcal{L}^{\mathcal{I}(t(j_m),t(j_{m-1}))}(j_m)\right) + \mathcal{L}^{\mathcal{I}(t(i),t(j_{r-1}))}(j) - 3n \log(nT) \left(\frac{t(j_{m-1}) - t(j_m)}{B}\right)^{\rho}.$$

因此，以高概率，专家集合 $R_r \setminus \mathcal{F}_q$ 将从池中移除。注意，根据标准的 Chernoff 界，过滤器 $\mathcal{F}_q$ 的大小至多为 $2 \log^4(nT)$ ，且该结论以高概率成立。由于我们有 $|R_r| \geqslant \frac{1}{4 \log(nT)} \cdot M$，因此根据并集界，以高概率有 $|R_r \setminus \mathcal{F}_q| \geqslant \frac{1}{6 \log(nT)}$。

情况 2。 在另一种情况下，我们有 $|R_r| < \frac{1}{4 \log(nT)} \cdot M$。由于 j_r 被定义为 $\mathcal{W}_r \setminus R_r$ 中最新的、被抽样到 $\mathcal{F}_q$ 的专家，并且每位专家以 $\frac{1}{\log^4(nT)}$ 的概率被抽样到 $\mathcal{F}_q$ 中，因此以高概率，$\mathcal{W}_r \setminus R_r$ 中最迟到达的专家至多有 $\log^6(nT)$ 位

晚于 j_r。换句话说，以高概率，$|O_r| \leqslant \log^6(nT)$。由于我们递归地定义了$\mathcal{W}_{r+1} = \mathcal{W}_r \setminus (R_r \cup O_r \cup V_r(j_r))$，因此我们有

$$\begin{aligned}|\mathcal{W}_{r+1}| &\geqslant |\mathcal{W}_r| - |R_r| - |O_r| - |V_r(j_r)| \\ &\geqslant \left(1 - \frac{r-1}{2\log(nT)}\right) \cdot M - \frac{1}{4\log(nT)} \cdot M - \log^6(nT) - \log^6(nT) \\ &\geqslant \left(1 - \frac{r}{2\log(nT)}\right) \cdot M,\end{aligned}$$

其中，最后一步是基于假设：在此情况下，$|M| \geqslant 2\log^9(nT)$。根据定义，$j_r \in \mathcal{F}_q$；此外，由于$\mathcal{W}_r$ 是$\mathcal{W}_{r-1}$的子集，因此 $j_r \preceq j_{r-1} \preceq \ldots \preceq j_1$ 。最后，由于$V_r(j_r) \cap \mathcal{W}_{r+1} = \emptyset$，归纳步骤便完成了。

综合以上结果可知，若对所有 $r \in [\log(nT) + 1]$，~~都有~~$|R_r| < \frac{1}{4\log(nT)} \cdot M$，那么对所有 $i \in \mathcal{W}_r$，都有$V_r(i) = \mathcal{W}_r$，因为任意专家 j 的损失至多为 T。因此，$|R_r| = |\mathcal{W}_r| \geqslant \left(1 - \frac{\log(nT)}{2\log(nT)}\right) \cdot M = \frac{1}{2} \cdot M$，此时我们可以得到更强的保证：$|\mathcal{P}_{C,q+1}| \leqslant \frac{1}{2} \cdot |\mathcal{P}_{C,q}|$。 □

引理 4.2. 对于任意时期 $\tau \in \left[\begin{smallmatrix} T \\ B \end{smallmatrix}\right]$ 和任意 $q \in [Q]$，以高概率有 $|\mathcal{P}_q^{(\tau)}| \leqslant 2\log^9(nT)$。

证明。我们对池的层级 $q \in [Q]$ 进行归纳。注意，对任意固定的 $\tau \in \left[\begin{smallmatrix} T \\ B \end{smallmatrix}\right]$，根据标准的 Chernoff界，以高概率有 $|P_q^{(\tau)}| \leqslant 2\log^9 n(nT)$ 现在假设该结论对层级 $q-1$之前的所有层级均成立。由于 $\mathcal{P}_q$ 在开始时为空，并且每隔 $\mathcal{P}_{q-1}$ 个2^{q-1} 时期与之合并，因此根据 引理 4.1，以高概率有 $|\mathcal{P}_q^{(\tau)}| \leqslant 2\log^9(nT)$。 □

引理 4.3. 以高概率，对所有时刻 $t \in [T]$，$|\mathcal{P}^{(t)}| = \mathcal{O}\left(\log^{10}(nT)\right)$，因此，算法 6 所使用的内存至多为 $\mathcal{O}\left(\log^{20}(nT)\right)$ 字空间。

证明。根据引理 4.2，对于所有 $\tau \in \left[\begin{smallmatrix} T \\ B \end{smallmatrix}\right]$ 和任意 $q \in [Q]$，我们有 $|\mathcal{P}_q^{(\tau)}| \leqslant 2\log^9(nT)$。由于 $Q = \mathrm{O}(\log(nT))$，且对于 $\tau \in \left[\begin{smallmatrix} T \\ B \end{smallmatrix}\right]$，$\mathcal{P}^{(t)}$ 等于最近的 $\mathcal{P}^{(\tau)}$，因此我们有 $|\mathcal{P}^{(t)}| \leqslant \sum_{q \in [Q]} |\mathcal{P}_q^{(\tau)}| = \mathrm{O}\left(\log^{10}(nT)\right)$。接下来只需在每位专家$j \in \mathcal{P}^{(t)}$的整个生命周期内存储该专家 $i \in \mathcal{P}^{(t)}$ 的损失。由于专家数量为 $j \in \mathcal{P}^{(t)}$，每位专家需要 $\mathrm{O}\left(\log^{10}(nT)\right)$ 字空间。因此，算法 6 使用的总内存至多为 $\mathrm{O}\left(\log^{20}(nT)\right)$ 字空间。

□

4.2 遗憾分析

现在我们转向遗憾分析。与 [PR23], 中的情形相比，我们的算法有一个复杂之处：为了界定实际遗憾，我们不能简单地使用定义 1中的覆盖概念来进行分析。这是因为我们需要在算法中定义好和坏时期。粗略地说，如果最佳专家 i^* 加入池后最终能够被过滤器覆盖，那么该时期就被定义为好时期。然而，如果我们使用定义 1中的覆盖概念，那么用 i^* 随机变量来覆盖究竟意味着什么？正如第 2 节所讨论的，我们分析中的一项技术创新是引入概念覆盖这一概念，定义如下。

定义 2（专家的概念覆盖）。给定一个过滤臂集合 F，并设专家 i *be* 的生命周期为 D。令 $L^{\mathcal{D}}(\mathsf{BM})$ 按照与 $L^{\mathcal{D}}(\mathsf{BM})$ 类似的方式构造，但使用每位专家的损失 $\widetilde{L}$，而非估计成本 L。若某个臂 i 满足以下条件，则称其为被覆盖

$$\mathcal{L}^{\mathcal{D}}(i) \geqslant \mathcal{L}^{\mathcal{D}}(\mathsf{BM}) - 3C \cdot n \log(nT) \cdot (|\mathcal{D}|)^{\rho},$$

其中，C 是一个绝对常数。

现在我们将 式化为与 [PR23] 中相同的好时期和坏时期概念。 .

定义 3（主动专家和被动专家）。对于任意时期 $\tau \in [T/B]$，如果某个专家在过滤过程中既未被抽样并放入过滤器，也未被抽样用于估计规模，我们就称该专家为被动专家。

固定损失序列、主动专家以及*EXP3* 中探索所使用的随机性。对于每个时期 a_τ，我们假设最佳专家 i^* 在时期 a_τ 加入池。我们进一步将驱逐时间 $t(a_\tau)$ 定义为 i^* 在概念上被主动专家集合覆盖的最早时刻。如果 i^* 从未在概念上被覆盖，我们就将 a_τ 加入坏时期集合$\mathcal{H}$。

引理 4.4。令 E 表示这样一个事件：每个在池中存在至少 $|\mathcal{D}(i)|$ 时间的专家 i，其代价估计值均在乘法因子 $\left(1+O\left(\binom{n}{|}^{\log(nT)}_{|^{1-\rho}}\right)\right)^{\mathcal{D}(i)}$ 的近似误差范围内。以事件 E 为条件。此外，在给定固定的损失序列以及每个时期抽样的主动专家集合的条件下，以高概率有

$$\sum_{t=1}^{T/B}\sum_{b=1}^{B} \ell^{(t-1)B+b}(i_{(t-1)B+b}) - \sum_{t=1}^{T} \ell^t(i^*) \leqslant B \cdot |\mathcal{H}| + \mathcal{O}\left(\frac{T}{B} \cdot (\gamma B \log(nT) + B^{\rho} n \log(nT))\right).$$

证明。对于每个 $t \in \left[\frac{T}{B}\right]$，令 i_t^*为时期 t内池 $\mathcal{P}^{(t)}$ 中的最佳专家。令 $\mathcal{H}$ 为不良轮次的集合。那么，以高概率，我们有

$$\begin{aligned}
\sum_{t=1}^{T/B}\sum_{b=1}^{B} \ell^{(t-1)B+b}(i_{(t-1)B+b}) - \sum_{t=1}^{T} \ell^t(i^*) &= \sum_{t=1}^{T/B}\sum_{b=1}^{B} \ell^{(t-1)B+b}(i_{(t-1)B+b}) - \sum_{t=1}^{T/B} \mathcal{L}^t(i^*) \\
&= \sum_{t=1}^{T/B}\sum_{b=1}^{B} \ell^{(t-1)B+b}(i_{(t-1)B+b}) - \sum_{t=1}^{T/B} \mathcal{L}^t(i_t^*) + \sum_{t=1}^{T/B} \mathcal{L}^t(i_t^*) - \sum_{t=1}^{T/B} \mathcal{L}^t(i^*) \\
&\leqslant \frac{T}{B} \cdot (\gamma B \log(nT) + B^{\rho} n \log(nT)) + \sum_{t=1}^{T/B} \mathcal{L}^t(i_t^*) - \sum_{t=1}^{T/B} \mathcal{L}^t(i^*),
\end{aligned}$$

在以下条件下：(1) EXP3 成功，因为每个时期 t 的损失与时期 t内池 $\mathcal{Q}^{(t)}$ 中的最佳专家 i_t^* 相比具有竞争力；以及 (2) 以高概率发生的事件：在任何长度为 B的时期内，探索率 γ至多被选中 $\gamma B \log(nT)$ 次。

现在我们将其分解 将时期 $t \in \left[\frac{T}{B}\right]$ 分解为不良轮次 $t \in \mathcal{H}$ 和好时期 $t \notin \mathcal{H}$.

$$\begin{aligned}
\sum_{t=1}^{T/B} \mathcal{L}^t(i_t^*) - \sum_{t=1}^{T/B} \mathcal{L}^t(i^*) &\leqslant \sum_{t \in \mathcal{H}} \left(\mathcal{L}^t(i_t^*) - \mathcal{L}^t(i_t^*)\right) + \sum_{t \notin \mathcal{H}} \left(\mathcal{L}^t(i_t^*) - \mathcal{L}^t(i_t^*)\right) \\
&\leqslant B \cdot |\mathcal{H}| + \frac{T}{B} \cdot (\gamma B \log(nT) + B^{\rho} n \log(nT)),
\end{aligned}$$

再次以 EXP3 成功为条件。这就完成了所声称的遗憾界。 □

现在我们给出坏时期数量的上界。

引理 4.5. 设 令 H 为不良轮次的集合。以高概率，有 $|H| \leqslant O\left(n \log^{10}(nT)\right)$.

证明。 固定损失序列 $\ell^1, \ldots, \ell^T$ 以及每个时期 t 抽样的主动专家序列 Y_t，在此条件下，令 W_t 为抽样的被动专家集合，因此 $W_t \cap Y_t = \emptyset$。令 $\mathcal{H}$ 为不良轮次的集合，并令 $\mathcal{E}$ 表示以下事件：对于每个在池中已存在 $D(i)$ 时间的专家 i，其代价估计值的乘法误差不超过 $\left(1+\mathcal{O}\left(\frac{n}{B}^{\log(nT)}\right)\right)^{\mathcal{D}(i)}$。对于任意不良轮次 $t \in \mathcal{H}$，在给定 $\mathcal{E}$ 的条件下，我们必有 $i^* \notin Y_t$，否则 i^* 会自行被淘汰，这与不良轮次 t 的定义矛盾。于是我们有

$$\begin{aligned}\mathbf{Pr}\left[i^* \in W_t \,|\, Y_1, \ldots, Y_{T/B}, \ell^1, \ldots, \ell^T\right] &= \mathbf{Pr}\left[i^* \in W_t \,|\, i^* \notin Y_t\right] \\ &\geqslant \mathbf{Pr}\left[i^* \in W_t\right] = \frac{1}{n} \cdot \left(1 - \frac{1}{\log^4(nT)}\right)^{2K \cdot 2Q} \geqslant \frac{1}{2n}.\end{aligned}$$

由于在固定损失序列以及活跃专家和采样专家后，$i^* \in W_t$ 在所有 $t \in \mathcal{H}$ 中是相互独立的事件，因此根据标准的 Chernoff 界，我们有

$$\begin{aligned}\mathbf{Pr}\left[|\mathcal{Q}| \leqslant \frac{|\mathcal{H}|}{4n} \mid Y_1, \ldots, Y_{T/B}, \ell^1, \ldots, \ell^T\right] &\leqslant \mathbf{Pr}\left[\sum_{t \in \mathcal{H}} \mathbb{1}[i^* \in W_t] \leqslant \frac{|\mathcal{H}|}{4n} \mid Y_1, \ldots, Y_{T/B}, \ell^1, \ldots, \ell^T\right] \\ &\leqslant \exp\left(-\frac{|\mathcal{H}|}{16n}\right),\end{aligned}$$

其中，$\mathcal{Q}$ 是时刻 t 的整个池。根据引理 4.3，我们有 $|\mathcal{Q}| = \mathcal{O}\left(\log^{10}(nT)\right)$ 以高概率成立。因此，以高概率有，$|\mathcal{H}| \leqslant \mathcal{O}\left(n \log^{10}(nT)\right)$。

□

引理 4.6。 设 $\mathcal{E}$ 表示如下事件：每个专家 i 在池中已存在 $D(i)$ 时间，其代价都能以乘法 $\left(1+\mathcal{O}\left(\frac{n}{B}^{\log(nT)}\right)\right)^{\mathcal{D}(i)}$-近似被正确估计，其中有 $D(i) \geqslant B$。在事件 $\mathcal{E}$ 条件下，以高概率，算法 6 的遗憾至多为

$$\left(nB + \gamma \cdot T + nT \cdot B^{\rho-1}\right) \cdot \operatorname{polylog}(nT).$$

证明。根据引理 4.4，遗憾至多为 $B \cdot |\mathrm{H}| + \mathrm{O}\left(\frac{T}{B} \cdot (\gamma B \log(nT) + B^{\rho} n \log(nT))\right)$。根据引理 4.5，我们有，以高概率，$|\mathrm{H}| \leqslant \mathrm{O}\left(n \log^{10}(nT)\right)$。因此，遗憾至多为 $Bn \cdot$ 多重对数$(nT) + \mathrm{O}\left(\frac{T}{B} \cdot (\gamma B \log(nT) + B^{\rho} n \log(nT))\right)$，这可以简化为引理陈述中的形式。

□

现在我们给出基线算法的遗憾的主要引理。

引理 4.7。 对于参数 $\gamma = (\frac{n}{B})^{1-\rho}$，当 $\rho = \frac{2}{3}$ 且 $B = \mathcal{O}\left(T^{3/4}\right)$ 时，算法 6 的遗憾为 $nT^{3/4} \cdot$ 多重对数(n)，其概率至少为 $1 - 1/\operatorname{poly}(nT)$。

证明。由引理 4.4，遗憾至多为 $B \cdot |\mathrm{H}| + \mathrm{O}\left(\frac{T}{B} \cdot \left(B^{2/3} n \log(nT)\right)\right)$。由引理 4.5，我们有，以高概率，$|\mathrm{H}| \leqslant \mathrm{O}\left(n \log^{10}(nT)\right)$。因此，当 $B = \mathrm{O}\left(T^{3/4}\right)$ 时，遗憾为 $nT^{3/4} \cdot$ 多重对数(nT)。

□

5 在流模型中使用双查询信号实现 $\sqrt{T}$ 遗憾

为了实现我们的滑动窗口算法（该算法具有区间遗憾，且每天进行两次查询），现在我们来考虑在 T 天内每天进行两次查询的在线学习问题，即每天进行两次查询的流式遗憾最小化。我们将展示一种算法，其遗憾为 $\widetilde{O}(n\sqrt{T})$，且所需空间为多重对数(nT)。这是迈向滑动窗口遗憾的重要中间步骤。为阐述我们的主要思想，我们着重在 T（即 $\sqrt{T}$）上实现最优遗憾，而非 n，并将在本节给出定理 6。最优的提升将留待第 9 节。

我们首先将基线算法修改为双查询设置。在这里，我们将利用无损信号，为每个臂的损失提供低方差估计，从而使单个基线算法的遗憾达到 $T^{2/3}$（而非 $T^{3/4}$）。修改后的基线算法如下：算法 9。

算法 9 基线算法 Baseline$_0$

输入： 时间范围 T，估计损失 $\widetilde{\mathcal{L}(i)}$ 对所有臂 $i \in \mathcal{P}$ 在池 $\mathcal{P}$中的持续时间内

$\mathcal{P} \leftarrow \emptyset$, $B \leftarrow \sqrt{T}\cdot$ 多重对数(nT), $\eta = \frac{1}{\sqrt{T}}$

对于长度为 B 的每个时期，**执行**

  以该概率将每个臂采样到 $\mathcal{P}$ 中 $\frac{1}{n}$

  保持与第4节相同的层次结构

  对于时期中的每个时刻 t，**执行**

    更新 $\widetilde{\text{通过在 } \mathcal{P} \text{ 中均匀采样一个臂并拉取它，利用无损信号更新 } \mathcal{L}(i)}$

    以与 $\exp(-\eta$ 成正比的概率选择臂 i $\widetilde{\mathcal{L}(i)}$如算法 3

    ▷不要更新估计损失

  利用估计损失，淘汰被过滤器覆盖的臂

$\widetilde{\text{引理 } 5.1.}$ 假设在所有时刻 $t \in [T]$，臂 i 在其持续时间 D 内、且 $|D| = D$ 属于 P 时，其损失 $L(i)$ 的估计值 $L(i)$ 的加性误差至多为 $\sqrt{D}\cdot$ 多重对数(nT)。那么，算法 9 的遗憾至多为 $nT^{2/3}\cdot$ 多重对数(nT)。此外，算法 9 使用的内存至多为多重对数(nT) 位。

证明。注意，算法 9 中基线计算出的概率向量恰好是乘法权重更新的概率向量。此外，算法 6 使用 EXP3 作为子程序，而 EXP3 就是在池 P 中针对每个臂的估计损失运行的乘法权重更新。因此，若臂 i 在 P 中的持续时间内，其损失 L(i) 的估计值 L(i) 的加性误差至多为 $\sqrt{T}\cdot$ 多重对数(nT)，则算法 9 等同于参数为 $\gamma = 0$ 和 $\rho = \frac{1}{2}$ 的算法 6。注意，根据命题 3.4，运行 EXP3 算法时，每个时期的期望遗憾为 B^{ρ} 多重对数(nT)。因此，我们可以应用引理 4.6，得出算法 9 的期望遗憾至多为

$$Bn\cdot\operatorname{polylog}(nT) + \mathcal{O}\left(\frac{T}{B}\cdot(\gamma B\log(nT) + B^{\rho}n\log(nT))\right) \leqslant \left(\frac{nT}{B}\cdot\sqrt{B} + Bn\right)\cdot\operatorname{polylog}(nT).$$

因此，对于 $B = T^{2/3}$，我们有 算法 9 的期望遗憾至多为 $nT^{2/3}\cdot$多重对数(nT)。

对于内存复杂度，在假设对臂 $\mathcal{L}(i)$ 的损失$\mathcal{L}(i)$ 在池 i 中整个持续时间内的估计$\mathcal{P}$ 的加性误差至多为 $\sqrt{T}\cdot$ 多重对数(nT)的条件下，我们可以使用引理 4.1 并取 $\rho = 1/2$，即以 $\sqrt{T}$ 多重对数(nT) 的加性误差进行驱逐，池大小合并的保证仍然成立。因此，内存复杂度至多为 多重对数(nT) 比特。□

现在，我们将通过一种类似于 $nT^{1/2}$PR nT 中的思路，将该算法的遗憾提升至多重对数的多项式级别 。然而，由于我们处理的是部分信息，不能简单地声称基线每一层的遗憾都很低。相反，我们需要界定通过均匀探索获得的信号所对应的每个臂的代价。

这里的关键观察是，如果我们通过均匀采样来估计专家和算法的损失，那么在固定损失的随机性$\{\ell^t\}_{t=1}^t$、活跃专家集合以及第二次查询之后，驱逐时间仍然是明确定义的。因此，我们仍然能够以与 [PR23] 相同的方式进行分析。

算法 10 双查询算法 基线$_k$

输入：时间范围 T1：初始化一个时间范围为 T 的基线实例 A$_{1\ k-1}$ 2：$B \leftarrow n^{(2-2^{k+2})}$ /$(2^{k+2}-1)\cdot T^{1-\frac{1}{2^{k+2}-1}}$3：对于每个长度为 B 的时期，执行 4：L(A$_1$), L(A$_2$) $\leftarrow$ 05：初始化一个时间范围为 B 的基线实例 A$_{2\ k-1}$ 6：在 A$_1$ 和 A$_2$ 上初始化 EXP3 7：对于每个 $t \in [T]$，执行 8：令 P 为 A$_1$ 和 A$_2$ 维护的臂集合 9：以 1/2 的概率 10：观察臂集合 $i \in$ P 中的一个随机臂，并获得无遗憾信号 11：将 L(i) 增加 $2|$P$|\cdot\ \ell^t(i)$12：使用 L(i) 更新 A$_1$ 和 A$_2$ 13：否则 14：从 $j \in$ {1, 2} 中均匀随机地选择一个 15：观察 Aj 选出的臂，并获得无遗憾信号16：将 L(Aj) 增加 $4\ell^t(i)$ 17：使用 L(A$_1$) 和 L(A$_2$) 上的乘法权重更新来选择要拉取的臂▷不要更新估 –

估计的损失

我们首先证明，算法能够很好地估计每个臂在池中持续时间内的真实损失。

引理 5.2. 考虑 算法 *10* 并假设 $\ell^t(i) \in [0,1]$ 对所有 $t \in [T]$均成立。对于任意在池 $\mathcal{P}$ 中持续了 $\mathcal{D}$ 长度 D 的臂，我们有

$$\left|\widetilde{\mathcal{L}(i)} - \sum_{t\in\mathcal{D}} \ell^t(i)\right| \leqslant \sqrt{D}\cdot \mathrm{polylog}(nT),$$

以高概率。

证明。对于任意 $t \in D$，令 $\ell^t(i)$ 表示时间 t 对 $L(i)$ 的估计值贡献，因此 $L(i) = \sum_{t \in D(i)} \ell^t(i)$，其中 $D(i)$ 是 i 的生命周期。令 P_t 表示 P 在时间 t 的状态。令 E 表示事件：对于所有 $t \in [T]$，$|P_t| \leqslant \text{polylog}(nT)$，因此根据引理 5.1，$\Pr[E] \geqslant 1 - \frac{1}{\text{poly}(nT)}$。我们有 $\ell^t(i) = 2^{|P_t|} \cdot \ell^t(i)$ 的概率为 $\frac{1}{2^{|P_t|}}$，否则，$\ell^t(i) = 0$ 的概率为 $1 - \frac{1}{2^{|P_t|}}$。于是，在给定 E 的条件下，我们有

$$\mathbb{E}\left[\widetilde{\ell^t(i)}\right] = \ell^t(i), \qquad \mathbb{E}\left[(\widetilde{\ell^t(i)})^2\right] = (\ell^t(i))^2 \cdot \text{polylog}(nT).$$

因此，根据伯恩斯坦不等式，参见 命题 3.1，我们以高概率有，

$$\left|\widetilde{\mathcal{L}(i)} - \sum_{t \in \mathcal{D}} \ell^t(i)\right| \leqslant \sqrt{D} \cdot \text{polylog}(nT),$$

当对所有 $t \in [T]$ 都有 $\ell^t(i) \in [0,1]$ 时。 □

接下来我们证明，算法能够良好地估计每个基线的真实损失。

引理 5.3。 *考虑算法 10，并假设对所有 $t \in [T]$ 都有 $\ell^t(i) \in [0,1]$。对于固定的 $t \in [T]$ 以及每个 $j \in \{1,2\}$，令 $L(A_j)$ 表示 A_j 截至并包括时间 t 的损失，并令 $L(A_j)$ 表示其估计值。那么，对于长度为 B 的任意时期，*

$$\left|\widetilde{\mathcal{L}(\mathcal{A}_j)} - \mathcal{L}(\mathcal{A}_j)\right| \leqslant \sqrt{B} \cdot \text{polylog}(nT),$$

以高概率成立。

证明。考虑一个长度为 B 的固定时期。令 $t \in [T]$ 固定，并令 P_t 表示 P 在时间 t 的状态。令 $L(A_j, t)$ 表示时间 t 对 $L(A_j)$ 的估计值贡献，令 $L(A_j, t)$ 表示 A_j 在时间 t 的真实损失。令 E 表示以下事件：对所有 $t \in [T]$，$|P_t| \leqslant$ 多重对数(nT)，因此根据引理 5.1，$\Pr[E] \geqslant 1 - \frac{1}{\text{多项式}(nT)}$。注意，$\widetilde{L}(A_j, t) = 4 \cdot L(A_j, t)$ 发生的概率为 1/4，而 $\widetilde{L}(A_j, t) = 0$ 发生的概率为 3/4。以 E 为条件，我们因此有

$$\mathbb{E}\left[\widetilde{\mathcal{L}}(\mathcal{A}_j, t)\right] = \mathcal{L}(\mathcal{A}_j, t), \qquad \mathbb{E}\left[(\widetilde{\mathcal{L}^t(i)})^2\right] \leqslant 4,$$

前提是 $\ell^t(i) \in [0,1]$ 对所有 $t \in [T]$。因此，根据伯恩斯坦不等式，参见 命题 3.1，

$$\mathbf{Pr}\left[\left|\widetilde{\mathcal{L}(i)} - \sum_{t \in \mathcal{D}} \ell^t(i)\right| \geqslant \sqrt{B} \cdot \text{polylog}(nT)\right] \leqslant \frac{1}{\text{poly}(nT)},$$

按需。 □

现在，我们给出每个基线的遗憾上界。

引理 5.4。 *基线 k 的遗憾为 $n \cdot T^{2^{k+1}/(2^{k+2}-1)} \cdot$ 多重对数(nT)，适用于任意 $k \leqslant$ 多重对数(nT)。*

证明。我们首先固定从 P 中均匀采样并观测某个臂的时间，因此也固定了探索元专家 A$_j$ 的时间。令 E_1 表示这样一个事件：在持续时间为 D 的区间内，P 中每个臂的损失估计值的加性误差至多为$\sqrt{D}\cdot$ 多重对数(nT)。类似地，令 E_2 表示这样一个事件：每个元专家 A$_j$ 的损失估计值的加性误差至多为 $\sqrt{T}\cdot$ 多重对数(nT)。令 E 表示 E_1 和 E_2同时发生的事件。根据 $k \leqslant$多重对数(nT)这一条件，P 中的臂至多有多重对数(nT)个。因此，E 以高概率成立，因为根据引理 5.2 和引理 5.3，E_1 和 E_2 均以高概率成立。接着，根据引理 5.1，基线$_0$的期望遗憾至多为 $nT^{2/3}\cdot$ 多重对数(nT)。至此，基础情形得证。

现在，我们将引理 4.4 的分析推广，以处理基线$_k$ 在 $k > 0$下的遗憾。考虑基线$_k$ ，并注意到存在一个基线$_{k-1}$ 实例，其时间范围为 T，即$\mathcal{A}_1$。对于每个时期 $\tau \in \left[\frac{T}{B}\right]$，令 i_t^*为池$\mathcal{P}^{(\tau)}$在时期 τ期间，对于$\mathcal{A}_1$,而言的最佳专家，与第4节中的定义相同。令$\mathcal{H}$ 为$\mathcal{A}_1$的不良轮次集合。那么，以高概率，与之前一样有：

$$\begin{aligned}\sum_{\tau=1}^{T/B}\sum_{b=1}^{B}\ell^{(\tau-1)B+b}(i_{(\tau-1)B+b}) - \sum_{\tau=1}^{T}\ell^{\tau}(i^*) &= \sum_{\tau=1}^{T/B}\sum_{b=1}^{B}\ell^{(\tau-1)B+b}(i_{(\tau-1)B+b}) - \sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i^*)\\ &= \sum_{\tau=1}^{T/B}\sum_{b=1}^{B}\ell^{(\tau-1)B+b}(i_{(\tau-1)B+b}) - \sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i_t^*) + \sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i_t^*) - \sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i^*)\\ &\leqslant \frac{T}{B}\cdot\left(\sqrt{T}\log(nT)\right) + \sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i_t^*) - \sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i^*),\end{aligned}$$

这归因于乘法权重更新的期望遗憾，在以$\mathcal{E}$ 为条件且探索率 γ 设为零的情况下。特别地，我们指出，乘法权重更新相对于$\mathcal{L}(\mathcal{A}_1)$ 和 $\mathcal{L}(\mathcal{A}_2)$ 的期望遗憾为 $\sqrt{T}\cdot$ 多重对数(nT)；以$\mathcal{E}$为条件时，这些值与$\mathcal{L}(\mathcal{A}_1)$ 和$\mathcal{L}(\mathcal{A}_2)$之间的误差至多为加性项 $\sqrt{T}\cdot$ 多重对数(nT)。

和之前一样，我们将这些时期 $\tau \in \left[\frac{T}{B}\right]$ 划分为不良轮次 $\tau \in \mathcal{H}$和好时期$\tau \notin \mathcal{H}$。

$$\sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i_t^*) - \sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i^*) \leqslant \sum_{\tau\in\mathcal{H}}\left(\mathcal{L}^{\tau}(i_t^*) - \mathcal{L}^{\tau}(i_t^*)\right) + \sum_{\tau\notin\mathcal{H}}\left(\mathcal{L}^{\tau}(i_t^*) - \mathcal{L}^{\tau}(i_t^*)\right).$$

注意，根据归纳假设，$\mathcal{A}_2$的期望遗憾至多为 $n\cdot B^{2^{k+1}/(2^{k+2}-1)}\cdot$多重对数$(nT)$。因此，根据 EXP3 在两个专家 $\mathcal{A}_1$ 和 $\mathcal{A}_2$, 上的保证，我们有

$$\sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i_t^*) - \sum_{\tau=1}^{T/B}\mathcal{L}^{\tau}(i^*) \leqslant n\cdot B^{2^k/(2^{k+1}-1)}\cdot\operatorname{polylog}(nT)\cdot|\mathcal{H}| + \frac{T}{B}\cdot\left(\sqrt{2B}\log(nT)\right).$$

根据 引理 4.5，不良轮次的数量满足 $|\mathcal{H}| \leqslant \mathcal{O}\left(n\log^{10}(nT)\right)$ 以高概率成立，此时上述表达式在以下取值处达到最小：

$$B = \left(\frac{1}{n}\right)^{\frac{2^{k+1}-1}{2^k}}\cdot T^{1-\frac{1}{2^{k+2}-1}}$$

并得到遗憾 $n\cdot T^{2^{k+1}/(2^{k+2}-1)}\cdot$多重对数$(nT)$，正如所需。 □

现在，我们可以递归地使用提升步骤，并得到 $\sqrt{T}\mathrm{poly}(n, \log T)$ 的遗憾，如下述定理所示。

定理 6。 对于 $k =$ 多重对数(nT)，以高概率，遗憾至多为 $n\sqrt{T}\cdot$ 多重对数(nT)，总空间至多为 多重对数(nT)。

定理 6 在关于 n的多项式依赖方面略弱于 定理 9 ，我们将在第 9 节中看到，通过更复杂的提升过程可以得到最优依赖关系。

6 有界内存和查询条件下的区间遗憾

现在，我们将重点转向滑动窗口模型，并研究在有界内存和少量查询条件下在线学习的 区间遗憾 。如前所述，我们的策略是模拟 [LZC+24] 中的算法，其在所有可能的区间$\mathcal{O}\left(\sqrt{n\,|\mathcal{I}|}\right)\cdot$ 多重对数(nT) 上均具有遗憾$\mathcal{I}$。本节将证明，采用我们在第 5 节中的算法思想，可以类似地在 $\mathcal{O}\left(n\sqrt{|I|}\right)\cdot$ 多重对数(nT) 遗憾下，仅使用 多重对数(nT) 内存，并且每轮进行两次查询。

本节中，我们使用 第 5 节 中的双查询遗憾最小化算法。与 第 5 节中的阐述类似，我们着重介绍本节中对 T 具有最优依赖关系的算法（即 $\sqrt{T}$），因为其分析更直观，也更容易理解。我们将对 n具有最优依赖关系的算法（其复杂得多）留到第 9 节再介绍。

我们首先给出如下算法 11。

算法 11 每天进行两次查询的区间遗憾算法

Input: 时间范围 T

1：令 $N\mathrm{meta} = \log nT$，每个 $\kappa \in [N^{\mathrm{meta}}]$，实例化一个时间范围为 2^κ 的算法 10 实例 $\mathtt{ALG}_\kappa$。2：令 $\eta_\kappa = \frac{1}{\sqrt{n}\cdot 2^\kappa}$ 为 $\mathtt{ALG}_\kappa$ 的“学习率”。3：将 $w_t(\kappa)$ 作为第 $t \in [T]$ 天每个算法 $\mathtt{ALG}_\kappa$ 的权重进行维护；初始化为 $w_1 = \eta_\kappa$。4：令 $v^{t,\kappa}$ 为第 t 天 $\mathtt{ALG}_\kappa$ 的臂上的概率分布。5：对每个天 $t \in [T]$，执行以下操作：6：运行利用算法，使用算法 12 进行游玩 i_t，并获得 q^t。▷此步骤可能会产生损失。7：运行探索算法，使用算法 13 更新区间算法。▷此步骤不会产生损失。8：令 (j_t, p_t) 为算法 13 返回的臂和概率。~ ~9：令 $\ell^t(j_t) = |\mathbb{P}_t| \cdot \ell^t(j_t)$，并令 $\ell^t(i) = 0$，对所有 $i \neq j_t$ 均成立。10：对每个 $\kappa \in [N^{\mathrm{meta}}]$，执行以下操作：~11：令 $r_t(\kappa) = \ell^t(j_t) \cdot (q^t(\kappa) \cdot \sum_\kappa v^{t,\kappa} - v^{t,\kappa})$。12：按如下方式更新 $w_{t+1}(\kappa)$：13：如果 $t+1 \bmod 2^\kappa$ 是整数，则 14：$w_{t+1}(\kappa) \leftarrow \eta_\kappa$。15：否则，16：$w_{t+1}(\kappa) \leftarrow (1 + \eta_\kappa \cdot r_\kappa(t))w_t(\kappa)$。

算法 12 区间遗憾：利用

输入： N^{meta} 算法 $\{\texttt{ALG}_\kappa\}_{\kappa=1}^{N^{\text{meta}}}$；权重 $\{w_t(\kappa)\}_{\kappa=1}^{N^{\text{meta}}}$

1: 计算 $q_t(\kappa) = \frac{w_t(\kappa)}{\sum_\kappa w_t(\kappa)}$，对于所有 $\kappa \in [N^{\text{meta}}]$，作为选择 $\texttt{ALG}_\kappa$ 的概率。

2: 抽样臂 $i_t \sim \sum_\kappa q_t(\kappa) \cdot v^{t,\kappa}$ 并选择 i_t。 3: 返回 q_t，作为 k 区间算法上的分布。

算法 13 区间遗憾：探索与更新

输入：N^{meta} 个算法 $\{\texttt{ALG}_\kappa\}_{\kappa=1}^{N^{\text{meta}}}$ 以概率 *1/2* 均匀随机地抽取一个区间算法 $\kappa \in [N^{\text{meta}}]$，从池 P_κ 中均匀随机地抽取 $\texttt{ALG}_\kappa$，如算法 *10* 所示。将 $L(j_t)$ 增加 $2N^{\text{meta}} \cdot |P_\kappa| \cdot \ell^t(i)$。按照算法 *10* 更新 $\texttt{ALG}_\kappa$ 中的基线算法。令 $p_t(j_t) = \frac{1}{2N^{\text{meta}} \cdot |P_\kappa|}$

否则

按照算法 10，从 $\texttt{ALG}_\kappa$ 中抽取一个算法 $\mathcal{A}_k$。令 K 为 $\texttt{ALG}_\kappa$ 中基线算法的总数（$K =$ 多重对数(nT)，如定理 6所示）。观察$\mathcal{A}_k$ 所选的臂 j_t，并将$\mathcal{L}(\mathcal{A}_k)$ 增加 $2N^{\text{meta}} \cdot K \cdot \ell^t(j_t)$。令 $p_t(j_t) = \frac{1}{2N^{\text{meta}} \cdot K}$返回抽样臂 j_t 及概率 $p_t(j_t)$。

为了分析 算法 11的遗憾，我们首先证明以下各项的性质：$\texttt{ALG}_\kappa$（它是我们 算法 10的副本）。这并非完全是从定理 6出发的黑箱论证，因为选择臂的规则现在已经改变。不过，由于我们在每个时刻都从各个池的并集中均匀随机地采样，而且只有 $N^{\text{meta}} = \mathcal{O}(\log T)$ 个算法，因此算法 10的性质不会发生显著变化。

引理 6.1。以高概率，对于 $\texttt{ALG}_\kappa$ 中的每个算法，$\kappa \in [N^{\text{meta}}]$ 使用的总空间至多为 多重对数(nT)，且其期望遗憾至多为 $n\sqrt{T}$ 多重对数(nT)。

证明。我们首先指出，引理 5.2 和引理 5.3 中的保证仍以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率成立（nT），只是其中的多重对数因子稍大一些）。为便于理解，我们在此重述引理的内容。

- 引理 5.2：对于任意长度为 D 的持续时间 $\mathcal{D}$，以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率，有 $\left| i\mathcal{L}() - \sum_{t\in\mathcal{D}} \ell^t(i) \right| \leqslant \sqrt{D} \cdot$ 多重对数 (nT)。

- 引理 5.3：设 $\mathrm{L(A}_i\mathrm{)}$ 和 $\mathrm{L(A}_i\mathrm{)}$ 分别为算法 10 中基线算法 A_i 的估计损失和实际损失，并设 B 为 A_i 的周期长度。那么，对于任意 $j \in \{i, i+1\}$，以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率，有 $\left| {}_j\mathrm{L(A}\ ) - \mathrm{L(A}_i\mathrm{)} \right| \leqslant \sqrt{B} \cdot$ 多重对数(nT)。

我们说明引理 5.2成立的原因，而引理 5.3 的推理也基于相同的论证。定义$\mathcal{F}_\kappa$ 为 $\texttt{ALG}_\kappa$ 被采样的概率。由于至多有$\log(T)$ 个算法，我们

有

$$\Pr\left(\mathcal{F}_\kappa\right) = \frac{1}{\log T}.$$

与引理 5.2的证明类似，令 $\mathcal{E}$ 为满足 $|\mathcal{P}_t| \leqslant$ 多重对数(nT) 的事件，并且我们知道该事件发生的概率至少为 $1 - 1/\operatorname{poly}(nT)$。我们以该高概率事件为条件。因此，我们有 $\widetilde{\mathcal{L}}(j_t) = 2|\mathcal{P}_t| \cdot \log T$，其概率为$1\,2\cdot|\mathcal{P}_t|\cdot\log T$，否则为 0。因此，我们可以类似地得到

$$\mathbb{E}\left[\widetilde{\mathcal{L}^t(i)}\right] = \ell^t(i), \qquad \mathbb{E}\left[(\widetilde{\mathcal{L}^t(i)})^2\right] = (\ell^t(i))^2 \cdot \operatorname{polylog}(nT)$$

与引理 5.2中的证明相比，仅相差 $\log^2(T)$。因此，应用伯恩斯坦不等式后，我们可以再次得到 $\left|\widetilde{i\mathcal{L}}() - \sum_{t\in\mathcal{D}} \ell^t(i)\right| \leqslant \sqrt{D} \cdot$ 多重对数(nT)。

请注意，引理 5.4 的正确性只要求引理 4.4、引理 4.5、引理 5.2和引理 5.3的正确性。我们已证明引理 5.2 和引理 5.3仍然成立。对于引理 4.4和引理 4.5，我们同样可以固定臂信号的随机性，而不失一般性。因此，对于任意 $\kappa \in [N^{\mathrm{meta}}]$，我们可以固定 $\texttt{ALG}_\kappa$ 中的更新步骤，并进行与引理 5.4相同的分析。遗憾和内存保证由引理 5.4中各项性质的正确性推出。

□

下文中，我们用 $\mathbb{E}_{\text{explore}}[\cdot]$ 表示仅对探索阶段的随机币（即 算法 13中使用的随机性）取期望。类似地，我们用 $\mathbb{E}_{\text{exploit}}[\cdot]$ 表示仅对利用阶段的随机币（即 算法 12中使用的随机性）取期望。根据 引理 6.1，我们可以为每个区间算法 $\texttt{ALG}_\kappa$ 建立如下区间遗憾。

引理 6.2。 设 $(i_1, i_2, \cdots\ i_{|I|})$ 为由带有 $I = 2^\kappa$ 的算法 *11*所选臂的随机变量。回顾一下，$v^{t,\kappa}$ 是 $\texttt{ALG}_\kappa$ 在第 t 天的臂分布。此外，回顾一下，i^* 表示最佳臂，而 $\ell^t(i)$ 是每天的估计成本。那么，以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率，我们有

$$\mathbb{E}_{\text{explore}}\left[\sum_{t\in I} \widetilde{\ell}^t(i_t) \cdot v^{t,\kappa}(i_t) - \sum_{t\in I} \widetilde{\ell}^t(i^*)\right] \leqslant n\sqrt{|I|} \cdot \operatorname{polylog}(nT),$$

其中，期望是对无损信号的随机性取的。

证明。在以下高概率事件的条件下：引理 6.1，每个 $\texttt{ALG}_k$ 算法的遗憾为 $\sqrt{n\,|I|} \cdot$ 多重对数(nT)，这意味着

$$\mathbb{E}_{\text{explore}}\left[\sum_{t\in I}\sum_{i=1}^{n} \widetilde{\ell}^t(i) v^{t,\kappa}(i) - \sum_{t\in I} \ell^t(i^*)\right] \leqslant n\sqrt{|I|} \cdot \operatorname{polylog}(nT).$$

此外，根据我们的算法规则，我们有 $\ell^t(i) = 0$ 如果 $i \neq i_t$。因此，我们有

$$\mathbb{E}_{\text{explore}}\left[\sum_{t\in I}\sum_{i=1}^{n} \widetilde{\ell}^t(i) v^{t,\kappa}(i) - \sum_{t\in I} \ell^t(i^*)\right] = \mathbb{E}_{\text{explore}}\left[\sum_{t\in I} \widetilde{\ell}^t(i_t) \cdot v^{t,\kappa}(i_t) - \sum_{t\in I} \widetilde{\ell}^t(i^*)\right] \leqslant n\sqrt{|I|} \cdot \operatorname{polylog}(nT),$$

如所需。 □

现在，我们对区间遗憾所产生的“外层”算法的遗憾进行界定。这部分基本上是标准的，沿用 [DGS15、LZC+24]中的相同论证。

引理 6.3. 对于任意固定区间 I，我们有

$$\mathbb{E}\left[\sum_{t\in I}\ell^t(i_t)-\sum_{t\in I}\ell^t(i^*)\right]=\mathbb{E}_{\text{explore}}\left[\sum_{t\in I}\sum_{i=1}^{n}\widetilde{\ell}^t(i)\cdot\left(\sum_{\kappa}q^t(\kappa)v^{t,\kappa}(i)\right)-\sum_{t\in I}\widetilde{\ell}^t(i^*)\right].$$

此外，对于任意 κ，有

$$\mathbb{E}_{\text{explore}}\left[\sum_{t\in I}\sum_{i=1}^{n}\widetilde{\ell}^t(i)\cdot\left(\sum_{\kappa}q^t(\kappa)v^{t,\kappa}(i)\right)-\sum_{t\in I}\widetilde{\ell}^t(i_t)\cdot v^{t,\kappa}(i_t)\right]\leqslant n\sqrt{|I|}\cdot\text{polylog}(nT).$$

证明。 我们仅给出该引理的证明概要，完整证明请参阅 [LZC+24] 。第一条结论可直接由 [LZC+24] 中完全相同的论证，通过随机性解耦得到。此外，利用与 [LZC+24] 中类似的归纳论证，我们可以得到

$$\begin{aligned}&\mathbb{E}_{\text{explore}}\left[\sum_{t\in I}\sum_{i=1}^{n}\widetilde{\ell}^t(i)\cdot\left(\sum_{\kappa}q^t(\kappa)v^{t,\kappa}(i)\right)-\sum_{t\in I}\widetilde{\ell}^t(i_t)\cdot v^{t,\kappa}(i_t)\right]\\&\leqslant\eta_\kappa\cdot\mathbb{E}_{\text{explore}}\left[\left(\sum_{t\in I}\sum_{i=1}^{n}\widetilde{\ell}^t(i)\cdot\left(\sum_{\kappa}q^t(\kappa)v^{t,\kappa}(i)\right)-\sum_{t\in I}\widetilde{\ell}^t(i_t)\cdot v^{t,\kappa}(i_t)\right)^2\right]+\frac{2\log(nT)}{\eta_\kappa}.\end{aligned}$$

此外，我们可以将右侧的项界定如下

$$\begin{aligned}&\mathbb{E}_{\text{explore}}\left[\left(\sum_{t\in I}\sum_{i=1}^{n}\widetilde{\ell}^t(i)\cdot\left(\sum_{\kappa}q^t(\kappa)v^{t,\kappa}(i)\right)-\sum_{t\in I}\widetilde{\ell}^t(i_t)\cdot v^{t,\kappa}(i_t)\right)^2\right]\\&=\mathbb{E}_{\text{explore},<t}\left[\sum_{i=1}^{n}p_t(i)\cdot(\frac{\ell^t(i)}{p_t(i)})^2\cdot\left(\left(\sum_{\kappa}q^t(\kappa)v^{t,\kappa}(i)\right)-v^{t,\kappa}(i)\right)^2\right]\\&\leqslant\mathbb{E}_{\text{explore},<t}\left[\sum_{i=1}^{n}p_t(i)\cdot(\frac{\ell^t(i)}{p_t(i)})^2\cdot\max_{\kappa}\left(v^{t,\kappa}(i)\right)^2\right]\\&\leqslant n\,\text{polylog}(nT).\qquad\qquad(\text{by } v^{t,\kappa}(i)\leqslant 1 \text{ and } p_t(i)\geqslant 1/\text{多重对数}(nT))\end{aligned}$$

这一步比 [LZC+24] 中的分析明显简单，因为我们有 $p_t(i)\geqslant 1/$多重对数(nT)(在 [LZC+24] 中，如果我们均匀随机地抽样，就会得到一个 n^2 因子，这也是他们采用复杂得多的分布的原因）。因此，我们得到

$$\mathbb{E}_{\text{explore}}\left[\sum_{t\in I}\sum_{i=1}^{n}\widetilde{\ell}^t(i)\cdot\left(\sum_{\kappa}q^t(\kappa)v^{t,\kappa}(i)\right)-\sum_{t\in I}\widetilde{\ell}^t(i_t)\cdot v^{t,\kappa}(i_t)\right]\leqslant\left(\eta_\kappa\cdot n\,|I|+\frac{1}{\eta_\kappa}\right)\cdot\text{polylog}(nT),$$

通过使用 $\eta_\kappa=\mathcal{O}\left(\frac{1}{\sqrt{n\sqrt{I}}}\right)$，得到所需的遗憾界，这与算法 11一致。

□

结合引理 6.2和引理 6.3可得所需的遗憾界：$(n\sqrt{|I|})$ 多重对数(nT)，并且以高概率仅需多重对数(nT) 内存。

定理 7。以高概率，算法 *11*使用$(n\sqrt{|I|})$ 多重对数(nT) 位内存即可实现多重对数(nT) 期望遗憾。

定理 7 在 定理 1 关于多项式依赖于 n这一点上略弱，而我们将在 第 9 节中看到，通过更复杂的提升过程可以得到最优依赖关系。

7 多臂老虎机设定下具有 $T^{2/3}$ 遗憾的算法

本节将注意力转向单查询多臂老虎机设定。 第 5 节 算法的关键组成部分是对臂的损失进行“免费”估计。我们注意到，可以通过设置 $\gamma > 0$来模拟这一过程，尽管这会对遗憾造成一定损失。本节最终将证明，这一思想可以导出具有 $nT^{2/3}$ 多重对数(nT))遗憾的算法。

为了恰当地介绍该算法，我们首先介绍覆盖的一种变体，作为定义 1的推广。

定义 4 （专家的宽松近似覆盖）。 给定一个由过滤臂 组成的集合 $\gamma_{arm} \in$ （0， $1/2$）。如果一个臂 i 的生命周期为 $\mathcal{D}$ ，则称其被覆盖，如果在⁀运行算法 *5* 时使用 i 和 $\mathcal{F}$ 得到$\mathcal{L}^{\mathcal{D}}$（BM），并且我们有

$$\widetilde{\mathcal{L}^{\mathcal{D}}}(i) \geqslant \widetilde{\mathcal{L}^{\mathcal{D}}}(\mathsf{BM}) - C \cdot \log(nT) \cdot \sqrt{\frac{|\mathcal{D}|}{\gamma_{arm}}},$$

其中， C 是一个绝对常数。

与 定义 1 中的原始定义相比，我们在 定义 4 中采用的新驱逐规则明显更加 宽松：在 定义 1中，我们在权衡后选取的参数约为 $|\mathcal{D}|^{2/3}$。由于 $T \geqslant |\mathcal{D}|$，定义 4 中的新定义允许更大的余量。我们的算法见 算法 15。

算法 14 基线算法 Baseline_0

‾**输入：** 时间范围 T，所有臂 $i \in \mathcal{P}$在池 $\mathcal{P}$ 中的持续时间内的估计损失 $\mathcal{L}_i$，周期长度为 B_0

1: $\mathcal{P} \leftarrow \emptyset$2: **对于**每个长度为 B_0 的时期**执行**3: 以概率 $\frac{1}{n}$ 将每个臂采样到 $\mathcal{P}_0$ 4: 维护与第 4 节中相同的层次结构5: **对于**时期中的每个时刻 t **执行**6: 以概率 γ_{arm}7: 均匀采样一个臂 $i \in \mathcal{P}_0$ ‾

8 i $\widetilde{\ }_i \leftarrow \widetilde{\ }_i + |$: 拉取并更新 $\mathcal{L}\mathcal{L}_{\mathcal{P}_{\gamma_{arm}}{}^{0}|} \cdot \ell^t(i)$

⁀9: 否则，以与 $\exp(-\eta\mathcal{L}_i)$ 成正比的概率拉取臂 i10: 使用估计损失，驱逐被过滤器覆盖的臂

——

现在，我们根据新的覆盖规则定义好时期和坏时期。为此，我们需要采用与第 4.2 节相同的论证方式来处理好时期和坏时期，不过概念覆盖的形式有所不同。

算法 15 基线算法 Baseline$_k$ 输入：时间范围 T，估计损失 L_i，所有臂 $i \in \mathrm{P}$ 在池 P 中对应持续时间内的估计损失，周期长度 B_{k-1}, B_k1: P $\leftarrow \emptyset$2: 对每个长度为 B_k 的时期执行 3: 以概率 $\frac{1}{n}$ 将每个臂抽样加入 P_k 4: 维护与第 4 节相同的层次结构 5: 对时期中的每个时间点 t 执行 6: 独立地将该时期划分为长度为 B_{k-1} 的时期 7: 在每个长度为 B_{k-1} 的时期开始时，初始化 $\mathrm{L}(\mathrm{A}_k) = 0$, $\mathrm{L}(\mathrm{A}_{k-1}) = 0$8: 对每个长度为 B_{k-1} 且池为 P_{k-1} 的时期，初始化一个 Baseline$_{k-1}$ 实例 9: 以概率 γ_{arm} 10: 从 $_k \cup \mathrm{P}_{k-1}$ 中均匀抽样一个臂 $i \in$ 11: 拉取 i，并更新 $\widetilde{\mathrm{L}}_i \leftarrow \widetilde{\mathrm{L}}_i + {}^{|}\mathrm{P}_k \cup \mathrm{P}_{\gamma_{arm}}{}^{k-1|} \cdot \ell^t(i)$ 12: 以概率 γ_{meta} 13: 从 $\{k, k-1\}$ 中均匀抽样一个元专家 $j \in$ 14: 遵循 Baseline$_j$，并更新 $\mathrm{L}(\mathrm{A}j) \leftarrow \mathrm{L}(A_j) + {}^{2}_{\gamma_{meta}} \cdot \ell^t(i)$ 15: 否则，以与 $\exp(-\eta \mathrm{L}(\mathrm{A}j))$ 成比例的概率，遵循元专家 $j \in \{k, k-1\}$ 16: 使用估计损失，淘汰被过滤器覆盖的臂

定义 5（专家的宽松概念覆盖）。给定一个过滤臂集合 F 和一个探索参数 $\gamma_{arm} \in (0,\ 1/2)$。以类似于构造 $L^{\mathcal{D}}$（BM）的方式构造 $L^{\mathcal{D}}$（BM），但使用每个专家的损失 L，而非估计成本 L。若一条臂 i 满足被覆盖 的条件，则称其为被覆盖：

$$\mathcal{L}^{\mathcal{D}}(i) \geqslant \mathcal{L}^{\mathcal{D}}(\mathsf{BM}) - 3C \cdot \log(nT) \cdot \sqrt{\frac{|\mathcal{D}|}{\gamma_{arm}}},$$

其中 C 是与 定义 4中相同的绝对常数。

实质上，定义 4 和 定义 5 的定义完全对应于 定义 1和 定义 2 的情形，只是通过改变“松弛量”的计算方式来处理加性误差。同样，我们将时期 E 定义为好时期，当且仅当 最佳专家 i^* 若在时期 E 被抽样，则最终会被概念覆盖，并从池中驱逐出去（采用 定义 5中的新覆盖概念）。另外，若时期 E 不是好时期，则将其定义为 i^* 不良轮次，也就是说，如果在时期 E中被抽样，概念驱逐规则不会将其驱逐。

在下文中，我们首先分析算法的空间复杂度，然后再分析遗憾。对于空间复杂度，我们基本上论证引理 4.1 对所有 基线$_k$算法均成立，只要 $k \leqslant$ 多重对数(nT)。

引理 7.1。设 $P_A^{(k)}$、$P_B^{(k)}$ 和 $P_C^{(k)}$ 为 *Merge* 算法中 *Baseline*$_k$ 对应于 $k \leqslant \log^{10}(nT)$ 的池。以高概率，存在

$$|\mathcal{P}_C^{(k)}| \leqslant \max\left(2\log^9(nT), \frac{1}{4}(|\mathcal{P}_A^{(k)}| + |\mathcal{P}_B^{(k)}|)\right).$$

引理 7.1 的证明主要沿用了 引理 4.1的证明。为避免文本过度重复，我们省略细节。根据 引理 7.1，我们可以如下展示 算法 15的内存效率。

引理 7.2. 以高概率，算法 *15* 所使用的内存最多为 $\mathcal{O}\left(\log^{30}(nT)\right)$个字。

证明。使用我们在引理 4.3 中采用的相同论证，可以证明，对于每个基线$_k$算法，在任意时刻 t，$\mathsf{Q}^{(t)}$ 中最多有 $\mathsf{O}\left(\log^{10}(nT)\right)$ 个臂，并且每个臂的损失可以用 $\mathsf{O}\left((\log^{10}(nT)\right)$ 个字存储。此外，我们（以确定性方式）维护至多 $\log^{10}(nT)$ 层基线。因此，内存字数最多为 $\mathsf{O}\left(\log^{30}(nT)\right)$，从而得到所需结论。

□

引理 7.3. 考虑 算法 *15* 并假设 $\ell^t(i)(t) \in [0,1]$ 对所有 $t \in [T]$均成立。对于任意臂$i \in [n]$，若其在池 $\mathcal{P}$ 中持续了 $\mathcal{D}$ 长度为 D的持续时间，我们有

$$\left|\widetilde{\mathcal{L}_i} - \sum_{t\in\mathcal{D}} \ell^t(i)\right| \leqslant \sqrt{\frac{D}{\gamma_{arm}}} \cdot \mathrm{polylog}(nT),$$

以高概率。

~ ~证明。对于任意 $t \in$D，定义 $\ell^t(i)$ 为对应于时间 t 且对 L_i 有贡献的估计值。令 P_t 表示 P 在时间 t 的状态。考虑事件 E，其中对于所有 $t \in [T]$，$|\mathsf{P}_t| \leqslant \mathrm{polylog}(nT)$ 成立，其概率至少为 $1 - \quad ^1\mathrm{poly}(nT)$（引理 7.3）。

具体而言，我们有：

$$\widetilde{\ell^t}(i) = \begin{cases} \frac{|\mathcal{P}_t|}{2\gamma_{arm}} \cdot \ell^t(i) & \text{with probability } \gamma_{arm} \cdot \frac{2}{|\mathcal{P}_t|}, \\ 0 & \text{with probability } 1 - \gamma_{arm} \cdot \frac{2}{|\mathcal{P}_t|}. \end{cases}$$

在 $\mathcal{E}$ 条件下，可以得到：

$$\mathbb{E}\left[\widetilde{\mathcal{L}_i}(t)\right] = \ell^t(i), \qquad \mathbb{E}\left[\left(\widetilde{\mathcal{L}_i}(t)\right)^2\right] = \frac{\mathcal{L}_i(t)^2}{\gamma_{arm}} \cdot \mathrm{polylog}(n).$$

应用伯恩斯坦不等式，参见 命题 3.1，我们可以以高概率得到：

$$\left|\widetilde{\mathcal{L}_i} - \sum_{t\in\mathcal{D}} \ell^t(i)\right| \leqslant \sqrt{\frac{D}{\gamma_{arm}}} \cdot \mathrm{polylog}(nT),$$

前提是对于所有 $t \in [T]$，$\ell^t(i) \in [0,1]$。 □

引理 7.4。 考虑 算法 *15* 并假设 $\ell^t(i) \in [0,1]$ 对于所有 $t \in [T]$。对于固定的 $t \in [T]$~以及每个 $j \in \{1,2\}$，令 $L(\mathcal{A}^j)$ 表示 $\mathcal{A}^j$ 截至时间 t（含该时刻）的损失，并令 $\mathcal{L}(\mathcal{A}^j)$ 为其估计值。则

$$\left|\widetilde{\mathcal{L}}(\mathcal{A}_j) - \mathcal{L}(\mathcal{A}_j)\right| \leqslant \sqrt{\frac{B_k}{\gamma_{meta}}} \cdot \mathrm{polylog}(nT),$$

以高概率。

证明。考虑一个长度为 B_k 的固定时期。设 $t \in [T]$ 已给定，并令 P_t 表示 P 在时间 t 的状态。定义 $\mathrm{L}(\mathrm{A}j, t)$ 为时间 t 的估计值对 $\mathrm{L}(\mathrm{A}j)$ 的贡献，并令 $L(\mathrm{A}j, t)$ 表示 $\mathrm{A}j$ 在时间 t 的实际损失。令 E 表示事件：对所有 $t \in [T]$，$|\mathrm{P}_t| \leqslant \mathrm{polylog}(nT)$ 成立，从而 $\mathbf{Pr}[\mathrm{E}] \geqslant 1 - {}^{1}\mathrm{poly}(nT)$（引理 7.3）。

我们注意到，以概率 $\frac{\gamma_{meta}}{2}$，$\widetilde{\mathcal{L}}(\mathcal{A}j, t) = \frac{2}{\gamma_{meta}} \cdot L(\mathcal{A}j, t)$，并且以概率 $1 - \frac{\gamma_{meta}}{2}$，$\widetilde{\mathcal{L}}(\mathcal{A}j, t) = 0$。在给定 $\mathcal{E}$ 的条件下，可得

$$\mathbb{E}\left[\widetilde{\mathcal{L}}(\mathcal{A}_j, t)\right] = \mathcal{L}(\mathcal{A}_j, t), \qquad \mathbb{E}\left[(\widetilde{\mathcal{L}_i(t)})^2\right] \leqslant \frac{4}{\gamma_{meta}},$$

假设 $\ell^t(i) \in [0, 1]$ 对所有 $t \in [T]$。因此，应用伯恩斯坦不等式（参见 命题 3.1），得到

$$\mathbf{Pr}\left[\left|\widetilde{\mathcal{L}_i} - \sum_{t \in \mathcal{D}} \ell^t(i)\right| \geqslant \sqrt{\frac{B_k}{\gamma_{meta}}} \cdot \mathrm{polylog}(nT)\right] \leqslant \frac{1}{\mathrm{poly}(nT)},$$

如所需。 □

引理 7.5。 对于任意 k，基线$\textsc{Baseline}_k$ 在定义 定义 5中所定义的一个好时期内，以高概率，其遗憾至多为

$$\left(\gamma_{meta} \cdot B_k + \gamma_{arm} \cdot B_k + \sqrt{\frac{B_k}{\gamma_{arm}}} + \sqrt{\frac{B_k}{\gamma_{meta}}}\right) \cdot \mathrm{polylog}(nT).$$

证明。 在期望意义下，一个长度为 B_k 的时期中有 $\gamma_{arm} \cdot B_k + \gamma_{meta} \cdot B_k$ 次探索，每次代价至多为 1。我们对这些天的遗憾作如下上界估计：

$$\gamma_{arm} \cdot B_k + \gamma_{meta} \cdot B_k.$$

接下来需要考虑选择并跟随某个元专家 $\{\mathcal{A}_{k-1}, \mathcal{A}_k\}$ 时那些天的损失。

在一个好时期中，存在一个专家 i 属于 $\mathcal{P}_{k-1}$ 它覆盖了最佳专家 i^*。因此，$\widetilde{\mathrm{L}^{B_k}(i)} \leqslant \widetilde{\mathrm{L}^{B_k}(i^*)} + C \cdot \log(nT) \cdot \sqrt{\frac{B_k}{\gamma_{arm}}}$。根据引理 7.3，以高概率，所有属于 $i \in \mathrm{P}$ 的臂的损失估计值与真实值之间的误差不超过加性误差 $\sqrt{D(i)\,\gamma_{arm}} \cdot$ 多重对数(nT)。因此，$\mathrm{L}^{B_k}(i) \leqslant \mathrm{L}^{B_k}(i^*) + 3C \cdot n\log(nT) \cdot \sqrt{\frac{B_k}{\gamma_{arm}}}$。根据乘法权重更新（EXP3）在损失序列 $\widetilde{\mathrm{L}(i)}$ 上的正确性（推论 5），A_k 的代价至多为

$$\sqrt{\frac{B_k}{\gamma_{arm}}} \cdot \mathrm{polylog}(nT) + 3C \cdot \log(nT) \cdot \sqrt{\frac{B_k}{\gamma_{arm}}} + \min_{i \in \mathcal{P}_{k-1}} \mathcal{L}^{B_k}(i)$$

根据推论 5。

同样，由引理 7.4可知，所有元专家的损失估计值与真实值之间的误差不超过 $\sqrt{\frac{B_k}{\gamma_{meta}}} \cdot$ 多重对数(nT)。根据乘法权重更新（EXP3）在元专家上的正确性$\mathcal{A}_k$ 和$\mathcal{A}_{k-1}$(推论 5)，算法在利用日上的损失至多为

$$\sqrt{\frac{B_k}{\gamma_{meta}}} \cdot \mathrm{polylog}(nT) + \sqrt{\frac{B_k}{\gamma_{arm}}} \cdot \mathrm{polylog}(nT) + \min_{i \in \mathcal{P}_{k-1}} \mathcal{L}^{B_k}(i) + 3C \cdot \log(nT) \cdot \sqrt{\frac{B_k}{\gamma_{arm}}}.$$

因此，以高概率，总体遗憾至多为

$$\left(\gamma_{meta}\cdot B_k+\gamma_{arm}\cdot B_k+\sqrt{\frac{B_k}{\gamma_{arm}}}+\sqrt{\frac{B_k}{\gamma_{meta}}}\right)\cdot \text{polylog}(nT),$$

如期。 □

以下证明可类似地由引理 4.5得到。

引理 7.6。 令 $\mathcal{H}$ 为算法 *15*产生的不良轮次集合。以高概率，有$|\mathcal{H}|\leqslant n\cdot$ 多重对数(n)。

现在，我们可以为 k-递归步骤建立主要遗憾界。

引理 7.7。 令 $R(B_k)$ 表示基线$_{k-1}$ 在长度为 B_k的一个时期内的遗憾。令 $\mathcal{E}$ 为引理 *7.3* 和引理 *7.4* 均成立的事件。在事件 $\mathcal{E}$的条件下，以高概率，算法 *15* 的遗憾至多为

$$\frac{T}{B_k}\cdot\left(\gamma_{meta}\cdot B_k+\gamma_{arm}\cdot B_k+n\cdot\sqrt{\frac{B_k}{\gamma_{arm}}}+\sqrt{\frac{B_k}{\gamma_{meta}}}\right)\cdot\text{polylog}(nT)+n\cdot R(B_k)\cdot\text{polylog}(nT).$$

证明。 注意，每个时期 $i\in\left[\frac{T}{B_k}\right]$ 要么是好时期，要么是不良轮次。此外，在固定探索时间，以及固定池采样和滤过程的随机性后，一个时期被归类为好时期或不良轮次是明确定义的。根据引理 7.5，每个好时期的遗憾至多为 $\left(\gamma_{meta}\cdot B_k+\gamma_{arm}\cdot B_k+\sqrt{\frac{B_k}{\gamma_{arm}}}+\sqrt{\frac{B_k}{\gamma_{meta}}}\right)\cdot$多重对数$(nT)$。另一方面，每个不良轮次的遗憾最多可达 $R(B_k)$。幸运的是，根据引理 7.6，以高概率，不良轮次至多有 $n\cdot$ 多重对数(nT) 个。因此，所需结论得证。

□

引理 *7.8*。设 $\gamma_{arm}=\gamma_{meta}=\frac{1}{T^{1/3}}$。令 $F(k)=\frac{7^k-3^k}{7^{k+1}-3^{k+1}}$，且令 $G(k)=\frac{2}{3}\cdot 7^{k+1}-3^{k+1}$，并且$B_k=n^{-G(k-1)}\cdot T^{F(k)}\cdot$ 多重对数(nT)。那么以高概率，算法 *15* 对于 $B_k=T^{F(k)}$ 的遗憾至多为 $nT^{G(k)}\cdot$ 多重对数(nT)。

证明。 根据引理 7.7，基线$_k$ 的遗憾为

$$\frac{T}{B_k}\cdot\left(\gamma_{meta}\cdot B_k+\gamma_{arm}\cdot B_k+\sqrt{\frac{B_k}{\gamma_{arm}}}+\sqrt{\frac{B_k}{\gamma_{meta}}}\right)\cdot\text{polylog}(nT)+n\cdot R(B_k)\cdot\text{polylog}(nT).$$

我们令 $\gamma_{arm}=\gamma_{meta}=\frac{1}{T^{1/3}}$。因此，基线$_k$的遗憾为

$$T^{2/3}\cdot\text{polylog}(nT)+T^{7/6}B_k^{-1/2}\cdot\text{polylog}(nT)+n\cdot R(B_k)\cdot\text{polylog}(nT).$$

对于 $k=0$，我们有 $R(B_k)=B_k$，且 $F(0)=\frac{7}{9}$。因此，对于 $B_0=n^{-G(-1)}\cdot T^{F(0)}$多重对数$(nT)=T^{7/9}$，遗憾为 $nT^{7/9}\cdot$多重对数$(nT)=nT^{G(k)}\cdot$多重对数(nT)，由此完成基础情形。

现在，假设基线$_{k-1}$的遗憾为 $nT^{G(k-1)}\cdot$多重对数(nT)，具体而言，对于长度为 B_k 的时期，遗憾为 $R(B_k)=nB_k^{G(k-1)}\cdot$多重对数$(nT)$。我们设定 $B_k=n^{-G(k-1)}\cdot T^{F(k)}\cdot$多重对数$(nT)$，使得基线$_k$ 的遗憾为

$$T^{2/3}\cdot\text{polylog}(nT)+T^{7/6}B_k^{-1/2}\cdot\text{polylog}(nT)+n\cdot T^{F(k)\cdot G(k-1)}\cdot\text{polylog}(nT).$$

注意到 $F(k)\cdot G(k-1)=G(k)$。因此，基线$_k$的遗憾为 $nT^{G(k)}\cdot$多重对数(nT)，这就完成了我们的归纳法。 □

结合 引理 7.2 和 引理 7.8，令 $k =$ 多重对数(nT) 即可直接得到我们在 定理 3中所需的结果。

定理 3. 存在一种在线学习算法，对于任意包含 n 个专家和 T 天的实例，满足 $T \geqslant n$，并且仅能查询单个专家（即多臂老虎机设定），该算法以至少 $1-1/\operatorname{poly}(nT)$ 的概率，使用 $\operatorname{polylog}(nT)$ 个内存字数，实现 $nT^{2/3} \cdot \operatorname{polylog}(T)$ 遗憾。

8 一种具有单查询信号和随机顺序最佳专家的近最优遗憾算法

现在我们讨论一种单查询算法：当最佳专家的损失序列为随机顺序时，该算法可实现近最优遗憾。下面给出定理 4的算法与证明。此结果也可视为一种针对对抗性多臂老虎机 的近最优算法，其分布假设非常宽松：只要求最佳臂的损失按随机顺序排列。

正如我们在第 7 节中所观察到的，由于探索期间产生的损失，在流模型中使用采样与驱逐框架，难以为单查询设定实现最优 $\sqrt{T}$ 遗憾。因此，我们在此采用不同的方法，围绕最佳专家的最优损失构建“二分搜索”结构。我们从 $C\sqrt{nT}$ 误差开始，对 $C = 1$进行逐步递增；如果没有臂满足所需的损失范围，我们便继续增大 C。假设最佳专家的误差为 $\gamma\sqrt{nT}$。由于最佳专家的顺序是随机的，当我们的猜测接近$\gamma\sqrt{nT}$时，也能找到误差率达到目标的专家。另一方面，如果某个专家在尚未 C 增大到 γ之前就已满足条件，这意味着该专家在某个区间内的损失极低，甚至优于最佳专家。随后，我们可以将该区间内的“反向遗憾”计入考量。即使该专家之后表现不佳（这意味着算法会放弃该专家并继续搜索），这些“反向遗憾”也足以摊销遗憾分析中的代价，从而达到最优结果。

该算法与算法 16中的算法相同。与本文探讨的其他算法相比，该算法也明显简单得多。

算法 16 一种具有多臂老虎机信号和随机顺序最佳专家的近乎最优算法**输入：** 时间范围 T1: $C \leftarrow 1$, $i \leftarrow 1$,$B_C \leftarrow \frac{100}{C+1}\sqrt{\frac{T}{n}\log(nT)}$ 2: **对** 每个时刻**执行**3: 选择专家 i4: 令 D_i 为自上次选择 i以来的区间5: **若** $|D_i| = \frac{1}{\varepsilon^2} \cdot \mathcal{B}_C$ 对于某个 $\varepsilon \in (0,1]$ 且 $\frac{T}{|D_i|} \cdot \left(\sum_{t \in D_i} \ell^t(i)\right) - C \cdot \sqrt{nT} > \frac{C\varepsilon}{2}\sqrt{nT}$ **则** 6: $i \leftarrow i+1$7: **若** $i = n+1$ **则**8: $C \leftarrow C+1$, $i \leftarrow 1$

注意到，算法 16 的内存效率显而易见，正如引理 8.1所示。

引理 8.1. 算法 *16* 最多使用 $\mathcal{O}(\log(nT))$ 位内存。

证明。 算法 16 轮询专家，并且在任何时候，我们只需跟踪单个专家的统计信息。这些统计信息仅包括天数、累积损失和一些辅助参数，所有这些都可以用 $\mathcal{O}(\log(nT))$ 位内存记录。 □

本节其余部分将证明 算法 16的遗憾。为此，我们首先介绍以下集中界。

命题8.1 (超几何随机变量之和的尾界)。 *[设94]* 令 X ~超几何分布$(N,\ K,\ n)$ 是一个期望为 $\mathbb{E}[X] = \frac{K}{N} \cdot n$的超几何随机变量。那么，对于任意 $t < \frac{K}{N}$，

$$\mathbf{Pr}\left[|X - \mathbb{E}[X]| \geqslant tn\right] \leqslant 2\exp(-2t^2 n).$$

然后，我们可以应用 命题8.1 ，从而获得最佳专家的估计损失 i^* 以及参数 C 在 算法 16 中的集中保证，如 引理8.2 和 推论8所示。

引理8.2。 假设最佳专家 i^* 取得了 $\gamma \cdot \sqrt{nT}$损失，其中 $\gamma \in [C, C+1)$，C 为整数，并假设损失按随机顺序排列。令 I 为满足 $|I| = \frac{1}{\varepsilon^2} \cdot B_C$的任意区间，其中 $\varepsilon \in (0, 1]$。那么以高概率，

$$\left|\frac{T}{|I|} \cdot \left(\sum_{t \in I} \ell^t(i^*)\right) - \gamma \cdot \sqrt{nT}\right| \leqslant \frac{\gamma\varepsilon}{3}\sqrt{nT}.$$

证明。 证明直接来自关于超几何随机变量的标准集中不等式，参见命题8.1。

□

推论8。 以高概率，在整个算法 *16*执行过程中，始终有 $C \leqslant \gamma + 1$。

证明。 注意，如果 $C \geqslant \gamma$，那么根据 引理8.2，我们有，以高概率，

$$\frac{T}{|I|} \cdot \left(\sum_{t \in I} \ell^t(i^*)\right) \leqslant \gamma \cdot \sqrt{nT}\left(1 + \frac{\varepsilon}{3}\right) \leqslant C \cdot \sqrt{nT}\left(1 + \frac{\varepsilon}{2}\right),$$

因此，对所有 T, i^* 应用并集界可知，其以高概率永远不会被淘汰。因此可得 $C \leqslant \gamma + 1$。

□

下一个引理给出了每个区间 D_i上的遗憾界。该引理将以下直觉形式化：如果一个次优专家在 $C < \gamma$中存活很长时间，那么它必定展现出了大量“反向遗憾”，而这部分遗憾可以计入最终的遗憾分析。

引理 8.3。 假设最佳专家 i^* 取得了 $\gamma \cdot \sqrt{nT}$ 损失，并令 $C \leqslant \gamma$ 固定，使得$B_C = \frac{100}{C+1}\sqrt{\frac{T}{n}\log(nT)}$。以高概率，区间 D_i上的遗憾为

$$\max\left(B_C, \frac{|D_i|}{T}\left(C\sqrt{nT}\left(1 + \frac{2}{3}\sqrt{\frac{B_C}{|D_i|}}\right) - \gamma\sqrt{nT} \cdot \left(1 - \frac{1}{3}\sqrt{\frac{B_C}{|D_i|}}\right)\right)\right).$$

证明。 令 i^* 为最佳专家。注意到 算法 16 使C的值单调递增，B_C 的值则单调递减。于是根据 引理8.2，以高概率有：

$$\left|\frac{T}{|D_i|} \cdot \left(\sum_{t \in D_i} \ell^t(i^*)\right) - \gamma \cdot \sqrt{nT}\right| \leqslant \frac{\gamma}{3}\sqrt{\frac{B_C}{|D_i|}}\sqrt{nT},$$

这意味着

$$\left(\sum_{t\in D_i}\ell^t(i^*)\right)\geqslant\frac{|D_i|}{T}\cdot\gamma\sqrt{nT}\cdot\left(1-\frac{1}{3}\sqrt{\frac{B_C}{|D_i|}}\right).$$

如果 i 从未被淘汰，那么我们有

$$\frac{T}{|D_i|}\cdot\left(\sum_{t\in D_i}\ell^t(i)\right)-C\cdot\sqrt{nT}\leqslant\frac{C}{2}\sqrt{\frac{B_C}{|D_i|}}\sqrt{nT}$$

使用 $\varepsilon=\sqrt{B_C/|D_i|}$。另一方面，对于任意一个满足 i 最终被淘汰的区间 D_i ，

$$\frac{T}{|D_i|}\cdot\left(\sum_{t\in D_i}\ell^t(i)\right)-C\cdot\sqrt{nT}>\frac{C}{2}\sqrt{\frac{B_C}{|D_i|}}\sqrt{nT}.$$

现在，在前一个时刻我们也没有淘汰 i，原因要么是 $|D_i-1|<B_C$，要么是不等式不成立。第一种情况下，我们有 $\left(\sum_{t\in D_i}\ell^t(i)\right)\leqslant B_C$，因为对于任意 t 和 i， $\ell^t(i)\in\{0,1\}$。第二种情况下，我们有

$$\frac{T}{|D_i|}\cdot\left(\sum_{t\in D_i}\ell^t(i)\right)-C\cdot\sqrt{nT}\leqslant\frac{2C}{3}\sqrt{\frac{B_C}{|D_i|}}\sqrt{nT},$$

因此

$$\sum_{t\in D_i}\ell^t(i)\leqslant\frac{|D_i|}{T}\cdot C\sqrt{nT}\left(1+\frac{2}{3}\sqrt{\frac{B_C}{|D_i|}}\right),$$

由此可得

$$\sum_{t\in D_i}\ell^t(i)-\sum_{t\in D_i}\ell^t(i^*)\leqslant\max\left(B_C,\frac{|D_i|}{T}\left(C\sqrt{nT}\left(1+\frac{2}{3}\sqrt{\frac{B_C}{|D_i|}}\right)-\gamma\sqrt{nT}\cdot\left(1-\frac{1}{3}\sqrt{\frac{B_C}{|D_i|}}\right)\right)\right).$$

□

现在，我们可以给出算法 16的主要遗憾分析如下 .

Lem引理 8.4. 以高概率，*算法 16* 实现的遗憾为 $\mathcal{O}\left(\sqrt{nT\log^2(nT)}\right)$ —— .

证明。 注意， C 最多只能增加到 $\gamma+1$。类似地，对任意固定的 C， i 最多只能从 $i=1$ 循环到 $i=n$ 一次。对于固定的 C 和固定的 i，令$D_{C,i}$ 为选择臂 i 的连续区间。根据引理 8.3，$D_{C,i}$ 上的遗憾至多为

$$\xi_{C,i}:=\max\left(B_C,\frac{|D_{C,i}|}{T}\left(C\sqrt{nT}\left(1+\frac{2}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right)-\gamma\sqrt{nT}\cdot\left(1-\frac{1}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right)\right)\right),$$

以高概率。令 $\xi_1:=\sum_{C\leqslant\frac{\gamma}{2}}\sum_{i\in[n]}\xi_{C,i}$，令 $\xi_2:=\sum_{C>\frac{\gamma}{2}}^{\lfloor\gamma-1\rfloor}\sum_{i\in[n]}\xi_{C,i}$，以及 $\xi_3:=\sum_{C=\lfloor\gamma\rfloor}^{\lfloor\gamma+1\rfloor}\sum_{i\in[n]}\xi_{C,i}$，使得总遗憾至多为

$$\sum_{C\in\lfloor\gamma+1\rfloor}\sum_{i\in[n]}\xi_{C,i}=\xi_1+\xi_2+\xi_3.$$

我们对各项 ξ_1、 ξ_2 和 ξ_3 给出如下上界。

注意，对于 $C \leqslant \frac{\gamma}{2}$，我们有

$$\frac{|D_{C,i}|}{T}\left(C\sqrt{nT}\left(1+\frac{2}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right)-\gamma\sqrt{nT}\cdot\left(1-\frac{1}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right)\right)\leqslant 0.$$

因此

$$\max\left(B_C,\frac{|D_{C,i}|}{T}\left(C\sqrt{nT}\left(1+\frac{2}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right)-\gamma\sqrt{nT}\cdot\left(1-\frac{1}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right)\right)\right)=B_C,$$

因此 $\xi_{C,\,i}\leqslant B_C$。于是我们有

$$\xi_1=\sum_{C\leqslant\frac{\gamma}{2}}\sum_{i\in[n]}\xi_{C,i}\leqslant\sum_{C=1}^{\infty}nB_C\leqslant 200\sqrt{nT}\log^2(nT),$$

因为 $B_C=\frac{100}{C+1}\sqrt{\frac{T}{n}}\log(nT)$。

对于 $C\in(\frac{\gamma}{2},\ \gamma-1]$，我们有

$$\begin{aligned}&\frac{|D_{C,i}|}{T}\left(C\sqrt{nT}\left(1+\frac{2}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right)-\gamma\sqrt{nT}\cdot\left(1-\frac{1}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right)\right)\\&\leqslant\frac{|D_{C,i}|}{T}\cdot\sqrt{nT}(C-\gamma)+\frac{|D_{C,i}|}{T}\sqrt{\frac{B_C}{|D_{C,i}|}}\cdot\gamma\sqrt{nT}.\end{aligned}$$

令 $\Gamma:=\frac{|D_{C,\,i}|}{T}\sqrt{\frac{B_C}{|D_C|}}\cdot\gamma\sqrt{nT}$。由于 $B_C=\frac{100}{C+1}\sqrt{\frac{T}{n}}\log(nT)$，我们有

$$\begin{aligned}\Gamma&=\gamma\sqrt{n\frac{|D_{C,i}|}{T}}\sqrt{B_C}\leqslant\gamma\sqrt{\frac{n|D_{C,i}|}{T}}\sqrt{\frac{200}{\gamma}\sqrt{\frac{T}{n}}\log(nT)}\\&\leqslant\frac{200\gamma^{1/2}n^{1/4}|D_{C,i}|^{1/2}}{T^{1/4}}\log^{1/2}(nT).\end{aligned}$$

由于 $\gamma\leqslant\sqrt{\frac{T}{n}}$，因此

$$\Gamma\leqslant 200|D_{C,i}|^{1/2}\log^{1/2}(nT).$$

现在，我们有

$$\xi_{C,i}\leqslant\max\left(B_C,\frac{|D_{C,i}|}{T}\cdot\sqrt{nT}(C-\gamma)+200|D_{C,i}|^{1/2}\log^{1/2}(nT)\right).$$

注意到，由于 $C-\gamma$为负数，因此当 $\frac{|D_{C,\,i}|}{T}\cdot\sqrt{nT}(C-\gamma)+200|D_{C,i}|^{1/2}\log(nT)$ 取最大值时，$|D_{C,i}|\lesssim\frac{T}{n}(C-\gamma)^2\log(nT)$。因此，

$$\xi_{C,i}\lesssim\max\left(B_C,\frac{1}{(\gamma-C)}\sqrt{\frac{T}{n}}\log(nT)\right).$$

因此，

$$\xi_2 = \sum_{C>\frac{\gamma}{2}}^{\lfloor\gamma-1\rfloor} \sum_{i\in[n]} \xi_{C,i} \lesssim \sum_{C>\frac{\gamma}{2}}^{\lfloor\gamma-1\rfloor} \sum_{i\in[n]} B_C + \frac{1}{(\gamma-C)}\sqrt{\frac{T}{n}}\log(nT) \lesssim \sqrt{nT}\log^2(nT),$$

因为 $B_C = \frac{100}{C+1}\sqrt{\frac{T}{n}}\log(nT)$ 且 $\frac{1}{\gamma-C} \leqslant \mathcal{O}\left(\frac{1}{\gamma}\right) \leqslant \mathcal{O}(1)$（使用 $C \geqslant \gamma/3$）。

最后，对于 $C \in (\gamma-1, \gamma+1]$，我们有

$$\frac{|D_{C,i}|}{T}\left(C\sqrt{nT}\left(1+\frac{2}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right) - \gamma\sqrt{nT}\cdot\left(1-\frac{1}{3}\sqrt{\frac{B_C}{|D_{C,i}|}}\right)\right) \leqslant \frac{|D_{C,i}|}{T}\cdot(12\sqrt{nT})$$

因此

$$\xi_{C,i} \leqslant \max\left(B_C, \frac{|D_{C,i}|}{T}\cdot(12\sqrt{nT})\right) \leqslant B_C + \frac{|D_{C,i}|}{T}\cdot(12\sqrt{nT}).$$

我们有 $B_C = \frac{100}{C+1}\sqrt{\frac{T}{n}}\log(nT)$。因此，

$$\xi_3 \lesssim \sqrt{nT},$$

因此，总遗憾至多为

$$\xi_1 + \xi_2 + \xi_3 \lesssim \sqrt{nT}\log^2(nT).$$

□

结合上述步骤，即得到了关于具有单查询信号的随机顺序最佳专家的目标定理结论。

定理 4. 存在一种在线学习算法，对于任意包含 n 个专家和 T 天的实例，满足 $T \geqslant n$，并且仅能查询单个专家（即多臂老虎机设定），若最佳专家的损失序列顺序随机，则该算法可实现 $\sqrt{nT}\cdot$ 多重对数 $(nT))$ 遗憾，使用 多重对数 (nT) 个内存字数，且成功概率至少为 $1-1/$ 多项式 (nT)。

9 超越 $\Omega(n)$ 的提升：双查询信号

现在，我们将介绍双查询信号算法的近最优提升算法及其分析。我们在第 5 节和第 6 节中提出的算法，在指数 T (分别为 $|I|$) 上已达到最优遗憾，但在 n 上尚未达到最优。本节将证明， [PR23] 中涉及的单周期专家提升同样适用于方差较低的估计损失序列，由此可得以下定理。

定理 1. 存在一种在线学习算法，对于任意包含 n 个专家和 T 天的实例，满足 $T \geqslant n$，并且每个时刻可进行两次查询，可实现 $\sqrt{n\,|I|}\cdot$ 多重对数(T) 的区间遗憾，其中 I 为任意区间；该算法使用 多重对数(T) 个内存字数，且以高概率成功，即 $1-\frac{1}{\text{多项式}(nT)}$。

本节主要分析在 T 天内进行遗憾最小化的定理 9算法。将该算法转换为区间遗憾的分析与第 6 节中的分析相同，我们将在本节末尾进行讨论。

定理 9。存在一种在线学习算法，对于任意包含 n 个专家、持续 T 天且满足 $T \geqslant n$ 的实例，在可以查询两位专家（即选择一位专家并查询另一位而不产生损失）的条件下，能够以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率实现 $\sqrt{nT} \cdot$ 多重对数(T) 遗憾，并使用 多重对数(nT) 个内存字数。

这些提升算法的算法设计和分析相当复杂；不过，其正确性大多可由 [PR23] 中的保证以及我们在前文讨论过的思路推出。

9.1 提升算法

我们首先给出单周期专家提升的算法，该算法与 [PR23], 中的算法类似，但针对部分信息设置进行了调整。该算法如下：算法 17。

算法 17 单次生专家提升算法，参见 [PR23]

令 $R = \log T$ 为要维护的线程总数
对所有 $r \in [R]$ 维护基线$^+(r)$ ▷$\log T$ 个这样的副本
$\mathcal{P} \leftarrow \mathcal{P}_{1,} \cup \cdots \cup \mathcal{P}_{R,}$ ▷$\mathcal{P}$ 是合并池
维护 $\widetilde{\mathcal{L}}^t(i)$ 作为每个 $t \in [T]$ 和 $i \in \mathcal{P}$ 的估计损失。
对于 $t = 1, 2, \ldots, T$ **do**
    运行基线$^+(r, \widetilde{\mathcal{L}}^t)$ $(r \in [R])$ ▷池维护；$\widetilde{\mathcal{L}}^t$ 是估计损失的向量。
    在$\mathcal{P}$上运行 MonocarpicExpert-Bandit。 ▷替换 EXP3
    如果 算法 20进入 算法 20 **则** ▷也就是说，算法 20 及之后的 t未执行
        更新 $\widetilde{\mathcal{L}}(i)$：通过在$\mathcal{P}$中均匀采样一个臂，使用无损 信号拉取该臂，并使用抽样概率的倒数重新加权。
    如果池中包含的专家多于 $\log^C(nT)$，**则** ▷C 将在引理 9.2中指定
        停止并为算法声明“失败”。

算法 17 包含两部分：基线$^+$ 算法和MonocarpicExpert-Bandit算法。与 [PR23], 中的基线$^+$ 一样，使用与我们的基线算法（算法 6）相同的结构，但不运行 EXP3 过程，而是使用 MonocarpicExpert-Bandit 算法来处理各时期内的遗憾最小化。

算法 18 基线$^+(r,$ $\widetilde{\mathcal{L}^t})$

参数：$B_r = \frac{T}{n2^{r-1}}$；~~T_{-1}~~$= T$；$T_r = B_{r-1}$

对于 $s = 1, 2, \ldots, T/T_r$ **执行** ▷次重启

  初始化池 $\mathcal{P}_{r,\cdot} \leftarrow \emptyset$ 和子池 $\mathcal{P}_{r,k} \leftarrow \emptyset$ $(k \in [0 : K])$ ▷$K = \log T$，如 算法 6

  对于 $\tau_{\text{epoch}} = 1, 2, \ldots, T_r/B_r$ **执行** ▷时期 τ_{epoch}

    如果 r 是拥有新时期的最低线程，**那么**

      对于 $r' = R, \ldots, r+1$ **执行**

        $\mathcal{P}_{r,0} \leftarrow$ 合并$(\mathcal{P}_{r,0} \cup \mathcal{P}_{r',\cdot})$.

        ▷从更高层级的线程池继承并执行算法 7

    以概率 $1/n$对每个臂进行采样，生成$\mathcal{P}_{r,0}$。

    若 $\tau = C \cdot 2^{C'}$ 对于某个整数 C, C' **则**

      令 pw(τ_{epoch}) 表示满足 $\tau = C \cdot 2^{\text{pw}(\tau_{\text{epoch}})}$ 的最大整数，其中 C 为某个整数。

      对于 $k = 0, 1, \ldots,$pw(τ_{epoch}) **执行** ▷与 算法 6相同的池更新

        使用 $\mathcal{P}_{k+1}$ 执行更新算法 7，将$\mathcal{P}_{k+1}$ 和$\mathcal{P}_k$合并。

        使用损失估计 $\widetilde{\mathcal{L}^t}(i)$ 的覆盖与驱逐。

        $P_k \leftarrow \emptyset$。

算法 19 MonocarpicExpert-Bandit

初始化 $\mathcal{U}_k \leftarrow \emptyset$、EXP$_k$ $(k \in [K])$ ▷$K =$ 对数$_2(T)$

EXP $\leftarrow$区间遗憾(EXP$_1$、……、EXP$_K$，T，标志 $=$为 True)

▷使用真实损失的区间遗憾

对于 $t = 1, 2, \ldots, T$ **do**

  将新激活的专家添加到$\mathcal{U}_1$。

  从 EXP 中采样一个决策并执行。

  若 $t = C \cdot 2^{C'}$ 对于某个整数 C，C' **则** ▷检查每一天，而不是每个时期

    对于 $l = 1, 2, \ldots,$pw(t) **执行** ▷更新成员资格

      $\mathcal{U}_{k+1} \leftarrow \mathcal{U}_{k+1} \cup \mathcal{U}_k$，$\mathcal{U}_k \leftarrow \emptyset$。

      移除$\mathcal{U}_{k+1}$中的非活跃专家。

过程EXP$_k$ ▷$k \in [K]$

  对 $s =$ 1、2、……、$T/2^{k-1}$ **执行** ▷第 s 次重启

    运行区间遗憾$(\mathcal{U}_k, 2^{k-1},$标志 $=$ 伪损失)

    ▷伪损失下的区间遗憾

算法 20 区间遗憾($\mathcal{U}$ T，标志)

1：初始化 $w_{a,b} \leftarrow 1$ 在 SingleInterval$_{a,b}$ 上，对于 $a \in [K], b \in [T/2^a]$) 2：令 $D_{a,b} = [2^a(b-1)$ $+1:\ 2^a b]$ 对于每个 $a \in [K], b \in [T/2^a]$3：对于 $t = 1, 2, \ldots, T$ 执行 ⁀4：在区间 $D_{a,b}$ 上运行更新 SingleInterval$_{a,b}$($D_{a,b}$, U, L$^{D_{a,b}}$)。5：如果标志 ==为 True，则 6：从 $\{w_{a,b}\}_{h(t,a,b)=1}$中采样动作 $i_{t,a,b}$。▷$h(t, a, b) = 1$ 如果 $t \in [2^a(b-1) + 1:\ 2^a b]$▷这是唯一一行使用所执行的查询及其损失⁀7：令 $\ell(i_{t,a,b}) \leftarrow \ell^t(i_{t,a,b})$。8：否则 9：以概率 $1/2$：10：均匀随机地采样 $k' \in [K]$，并通过 EXP$_{k'}$采样动作 $i_{t,a,b}$。▷使用第二次查询且不产生损失 11：观察损失 $\ell(i_{t,a,b})$，将其作为 EXP$_{k'}$的损失。⁀12：令 $\ell(i_{t,a,b}) \leftarrow \ell(i_{t,a,b}) \cdot K$。13：如果未采样到 no$i_{t,a,b}$，则 14：继续到第 $t+1$ 天（不对 $w_{a,b}$进行更新）15：计算期望损失

$$\bar{\ell}^t \leftarrow \sum_{a,b:h(t,a,b)=1} \frac{w_{a,b}}{\sum_{a',b':h(t,a',b')=1} w_{a',b'}} \cdot \widetilde{\ell}(i_{t,a,b})$$

16: 指定损失 $\hat{\ell}_t(a,b) = \begin{cases} \ell^t(\quad i_{t\,a\,b} & h\ t\ a\ b\ = 1 \\ \bar{\ell}^t & h\ t\ a\ b\ = 0 \end{cases}$

17: 使用 SQUINT 更新权重分布

$$w_{a,b} \leftarrow \mathbb{E}_\eta \left[\eta \cdot \exp\left(\eta \sum_{\tau=1}^{t-1} v_\tau(a,b) - \eta^2 \sum_{\tau=1}^{t-1} v_\tau^2(a,b) \right) \right],$$

其中 $v_\tau(a,\ b) = \bar{\ell}_\tau - \hat{\ell}_\tau(a,\ b)$。

算法 21 SingleInterval$_{a,b}$($D_{a,b}, \mathcal{U},$ $\widetilde{\mathcal{L}^{D_{a,\ b}}}$)

输入： 持续时间 $D_{a,b}$，臂集合 $\mathcal{U}$，以及 $\widetilde{\mathcal{L}^{D_{a,\ b}}}(i)$ 对于 $i \in \mathcal{U}$

1: **对于** $t \in D_{a,b}$ **do**

2: 使用损失估计运行 EXP3 $\widetilde{\mathcal{L}^{D_{a,\ b}}}(i)$，针对每个 $i \in \mathcal{U}$。

与 [PR23],中的单轮强化算法相比，我们的算法最显著的区别在于区间遗憾[2] 和 SINGLEINTERVAL算法。由于我们每天只有两次查询，因此无法再像$_k$ 中的算法 19那样更新所有 EXP$_k$ 算法。我们改为每天均匀随机地采样一个臂，以估计损失，并直接对估计损失运行 SQUINT 算法。由于估计损失的方差较低，我们可以使用一种"从部分到全部"的归约论证，证明$_k$ 在 $k \in [K]$ 上的 EXP 算法遗憾仍然较低，从而得到 [PR23]中所示的区间保证。

9.2 技术引理与 MONOCARPICEXPERT-BANDIT 算法的分析

在给出形式化证明之前，我们首先建立算法中任意臂的损失序列估计误差界。

引理 9.1。 设 i 是 算法 *17*的池中的任意臂。对于任意给定的区间$\mathcal{I}$，以至少 $1 - 1/\operatorname{poly}(nT)$的概率，有

$$\left|\widetilde{\mathcal{L}^D}(i) - \sum_{t\in\mathcal{I}} \ell^t(i)\right| \leqslant \sqrt{|\mathcal{P}| \cdot |\mathcal{I}|} \cdot \operatorname{polylog}(nT),$$

其中， $|\mathcal{P}|$ 是池并集的大小。此外，如果我们只并行维护 $\operatorname{polylog}(nT)$ 个区间，则只需 $\operatorname{polylog}(nT)$ 内存。

证明。 不失一般性，假设 $|\mathcal{I}| \geqslant \operatorname{polylog}(nT)$ ，否则我们显然可以得到 polylog 的误差界，因为 $|\mathcal{P}|$ ， $|\mathcal{I}| \geqslant 1$。每天，我们至少有 $1/2$ 的概率从池中均匀随机地抽取一个臂。因此，我们可以应用 命题 3.4 （另见 算法 3 ）来得到该引理。 □

我们的做法与 [PR23] 不同，因为我们首先对池中臂的总数进行上界估计。这一做法的意义稍后便会显现。

引理 9.2 (参见 [PR23] 中的引理 4.12)。 以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率，每个池 $\mathcal{P}_{r,\cdot}$ 中的臂不超过 $\operatorname{polylog}(nT)$ 个，其中任意 $r \in [R]$ 均成立。因此，以至少$1 - 1/\operatorname{poly}(nT)$ 的概率，所有池的并集大小至多为 $|\mathcal{P}| \leqslant \log^C(nT)$，且算法不会声明"失败"。

证明。该证明遵循与 [PR23] 相同的逻辑。我们首先证明，在任意时刻，以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率，有 $|\mathcal{P}_{r,k}| \leqslant \operatorname{polylog}(nT)$ 对任意 $r \in [R]$ 和 $k \in [K]$均成立。我们通过对 r进行归纳论证来证明这一点，并坚持使用带有 $\sqrt{D}\operatorname{polylog}(nT)$ 误差的覆盖概念（定义 1）进行驱逐。对于 $r = R$，无需从更高层级的池中继承任何内容。由于我们以概率 $1/n$对每个臂进行采样，因此初始时，我们的池大小以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率至多为 $O(\log(nT))$。此外，根据引理 9.1 和归纳假设，每个专家的估计误差至多为 $\sqrt{|\mathcal{I}|} \cdot \operatorname{polylog}(nT)$。注意，为使引理 4.1 和引理 4.2 中的分析成立，我们只需要池大小 $|\mathcal{P}|$ 小于 $O(n^2)$。因此，

[2]不要与 $\mathfrak{b}$ 混淆，它与我们的区间遗憾算法（相对于区间内的最佳臂）不同，见第 6.

引理 4.1 （合并算法）仍由归纳假设成立，并且我们可以使用与引理 4.1和引理 5.1相同的证明来得到

$$|\mathcal{P}_C^{(k)}| \leqslant \max\left(2\log^9(nT), \frac{1}{4}(|\mathcal{P}_A^{(k)}| + |\mathcal{P}_B^{(k)}|)\right)$$

对所有 $k \in [K]$ 均成立。因此，以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率， $r = R$ 的池大小至多为 $\operatorname{polylog}(nT)$。

现在，为了进行归纳，假设归纳结论对线程 $r+1$ 成立。对于第 r 个线程，$\mathcal{P}_{r,\,0}$ 以概率 $1/n$ 进行采样并从线程 r 继承，从而启动。根据归纳假设，存在

$$|\mathcal{P}_{r',\cdot}| \leqslant \operatorname{polylog}(nT)$$

对于任意 $r' > r$。因此，通过运行合并算法（算法 7），对于归纳中的 r 以及任意 k，我们有

$$|\mathcal{P}_{r,k}| \leqslant \max\{2\log^9(nT), \frac{1}{4}(|\mathcal{P}_{r,k}| + \operatorname{polylog}(nT))\} \leqslant \operatorname{polylog}(nT).$$

最后，由于 $R = \operatorname{log}(nT)$ 且 $K = \operatorname{log}(nT)$，我们有 $|\mathcal{P}| = |\cup_r \cup_k \mathcal{P}_{r,\,k}| \leqslant \operatorname{polylog}(nT)$，如所期望。 □

现在，我们正式证明 MonocarpicExpert-Bandit 算法的区间遗憾保证。这里，我们将利用池中臂的数量至多为 $\operatorname{polylog}(nT)$ 这一事实来界定估计方差。

引理 9.3 (参见 [PR23] 中的引理 4.11)。 设 $T \geqslant 1$ 且 $|\mathcal{U}| \leqslant \operatorname{polylog}(nT)$。对于任意专家 $i \in \mathcal{U}$和时间区间 $\mathcal{I} \subseteq [T]$，满足 $|\mathcal{I}| \geqslant \operatorname{polylog}(nT)$，算法 区间遗憾 保证以至少 $1 - 1/(nT)^{\omega(1)}$ 的概率

$$\sum_{t\in\mathcal{I}} \ell^t(i_t) - \sum_{t\in\mathcal{I}} \ell^t(i) \leqslant \mathcal{O}\left(\sqrt{|\mathcal{I}|}\right) \cdot \operatorname{polylog}(nT)$$

在自适应对手面前成立。此外，区间遗憾 最多使用$\operatorname{polylog}(nT)$ 个内存字数。

证明。 我们首先证明 标志 = true 的情况，因为这与 [PR23] 中的情形基本相同。以下分析主要摘自 [PR23], ，为完整起见，我们将其列出。固定任意 $a \in [L]$， $b \in [T/2^a]$ 以及专家 $i \in \mathcal{U}$。由于我们以 $1/2$ 的概率从池中抽取一条臂，以估计各臂的损失，并且在高概率事件（池中的专家总数至多为 $\operatorname{polylog}(nT)$ (引理 9.2) 的条件下，我们可以论证，对于任意大小至少为 $\mathcal{I}$ 的区间，$\operatorname{polylog}(nT)$，有

$$\left|\widetilde{\mathcal{L}}^{\mathcal{I}}(i) - \sum_{t\in\mathcal{I}} \ell^t(i)\right| \leqslant \sqrt{|\mathcal{I}|} \cdot \operatorname{polylog}(nT)$$

以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率。这里， $\widetilde{\mathcal{L}}^{\mathcal{I}}(i)$ 是根据 i 由算法 17 的算法 17得到的臂的估计损失。然后，我们可以对 T^2 个区间应用并集界，并论证

所有区间都满足 $\sqrt{|\mathcal{I}|}\cdot$ 多重对数(nT) 误差。根据采用以学习作为探索的 EXP3 的遗憾保证（命题 3.4），以至少 $1-1/\operatorname{poly}(nT)$ 的概率，我们有

$$\sum_{t=2^a(b-1)+1}^{2^a b}\ell^t(i_{t,a,b})-\sum_{t=2^a(b-1)+1}^{2^a b}\ell^t(i)\leqslant\mathcal{O}\left(\sqrt{2^a}\right)\cdot\operatorname{polylog}(nT).$$

当标志 $=$ 为 true 时，由于我们有概率采样 $i_{t,\ a,\ b}$ 为 $p_{t,\ a,\ b}=\frac{w_{t,\ a,\ b}}{\sum_{a,\ b:h(t,a,b)=1}w_{t,\ a',\ b'}}$，我们有

$$\begin{aligned}\bar{\ell}^t&=\sum_{a,b:h(t,a,b)=1}\frac{w_{t,a,b}}{\sum_{a',b':h(t,a',b')=1}w_{t,a',b'}}\widetilde{\ell}(i_{t,a,b})\\&=\sum_{a,b:h(t,a,b)=1}\frac{w_{t,a,b}}{\sum_{a',b':h(t,a',b')=1}w_{t,a',b'}}\ell(i_{t,a,b})\qquad\text{(since the setting we are using the true loss)}\\&=\sum_{a,b:h(t,a,b)=1}p_{t,a,b}\ell(i_{t,a,b})=\mathbb{E}\left[\ell(i_t)\right],\end{aligned}$$

其中 i_t 是外层（区间遗憾）算法采取的动作。因此，根据 SQUINT 算法（引理 3.1）的保证，有

$$\mathbb{E}\left[\sum_{t\in[2^a(b-1)+1:2^a b]}\ell^t(i_t)-\sum_{t\in[2^a(b-1)+1:2^a b]}\ell^t(i_{t,a,b})\right]=\mathcal{O}\left(\sqrt{2^a\log(nT)}\right).$$

应用集中不等式和 $|\mathbb{E}[\ell^t(i_t)]-\ell^t(i_t)|\leqslant 2$，可得

$$\sum_{t\in[2^a(b-1)+1:2^a b]}\ell^t(i_t)-\sum_{t\in[2^a(b-1)+1:2^a b]}\ell^t(i_{t,a,b})\leqslant\mathcal{O}\left(\sqrt{2^a\log(nT)}\right)$$

以至少 $1-1/\operatorname{poly}(nT)$ 的概率成立。因此，通过使用三角不等式结合内层和外层遗憾，我们有至少 $1-1/\operatorname{poly}(nT)$ 的概率得到，

$$\begin{aligned}&\sum_{t\in[2^a(b-1)+1:2^a b]}\ell^t(i_t)-\sum_{t\in[2^a(b-1)+1:2^a b]}\ell^t(i)\\&=\sum_{t\in[2^a(b-1)+1:2^a b]}\ell^t(i_t)-\sum_{t\in[2^a(b-1)+1:2^a b]}\ell^t(i_{t,a,b})+\sum_{t\in[2^a(b-1)+1:2^a b]}\ell^t(i_{t,a,b})-\sum_{t=2^a(b-1)+1}^{2^a b}\ell^t(i)\\&\leqslant O\left(\sqrt{2^a}\cdot\operatorname{polylog}(nT)\right).\end{aligned}$$

现在我们转向标志 $\neq$True 的情形。这部分与 [PR23]略有不同，因为现在我们需要对伪损失运行 SQUINT。首先注意，关于 $\sum_{t=2^a(b-1)+1}^{2^a b}\ell^t(i_{t,a,b})-\sum_{t=2^a(b-1)+1}^{2^a b}\ell^t(i)\leqslant O\left(\sqrt{2^a}\right)\cdot$ 多重对数(nT) 的保证保持不变。每天，我们以 $1/2$ 概率从这些池的并集中抽取一条臂。池的数量至多为多重对数(nT)，且每个池中至多有多重对数(nT)条臂（引理 9.2），因此抽取每条臂的概率至少为 $1/$ 多重对数(nT)。设 $i_{t,a,b}$ 为算法抽取的任意一位专家，则有

$$\mathbb{E}\left[\widetilde{\ell}^t(i_{t,a,b})\right]=\ell^t(i_{t,a,b}),\qquad\mathbb{E}\left[(\widetilde{\ell}^t(i_{t,a,b}))^2\right]=(\ell^t(i_{t,a,b}))^2\cdot\operatorname{polylog}(nT).$$

令 $\mathrm{L}^{a,b}(i) := \sum_{i_t,a,b=i} P \cdot \ell^t(i)$ 表示专家 i 在区间 $D_{a,b}$ 内经过 t 步的损失估计。对于 $T \geqslant$ 多重对数(nT)，以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率，每个池 U 中每条臂的损失都至少被抽样多重对数(nT)次。因此，我们可以应用伯恩斯坦不等式（命题 3.1），并得到以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率，

$$\Pr\left(\left|\widetilde{\mathcal{L}^{a,b}}(i) - \sum_{t \in D_{a,b}} \ell^t(i)\right| \leqslant \sqrt{|D_{a,b}|} \cdot \operatorname{polylog}(nT)\right) \geqslant 1 - 1/\operatorname{poly}(nT).$$

然后，我们可以对至多 T^2 个连续区间应用并集界，从而得到误差$\sqrt{|D_{a,\ b}|}$ 对所有区间均成立。因此，我们可以将 $\sum_{t\in[2^a(b-1)+1:\ 2^a b]} \ell^t(i_t) - \sum_{t\in[2^a(b-1)+1:\ 2^a b]} \ell^t(i_{t,\ a,\ b})$ 的遗憾分解如下。

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in[2^a(b-1)+1:2^a b]} \ell^t(i_t) - \sum_{t\in[2^a(b-1)+1:2^a b]} \ell^t(i_{t,a,b})\right] \\
&= \mathbb{E}\left[\sum_{t\in[2^a(b-1)+1:2^a b]} \ell^t(i_t) - \sum_{t\in[2^a(b-1)+1:2^a b]} \widetilde{\ell^t}(i_t)\right] + \mathbb{E}\left[\sum_{t\in[2^a(b-1)+1:2^a b]} \widetilde{\ell^t}(i_t) - \sum_{t\in[2^a(b-1)+1:2^a b]} \widetilde{\ell^t}(i_{t,a,b})\right] \\
&\quad + \mathbb{E}\left[\sum_{t\in[2^a(b-1)+1:2^a b]} \widetilde{\ell^t}(i_{t,a,b})) - \sum_{t\in[2^a(b-1)+1:2^a b]} \ell^t(i_{t,a,b})\right].
\end{aligned}$$

在近似误差有界这一高概率事件的条件下，我们可以将第一项界定为损失之和，即

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in[2^a(b-1)+1:2^a b]} \ell^t(i_t) - \sum_{t\in[2^a(b-1)+1:2^a b]} \ell^t(i_{t,a,b})\right] &\leqslant \mathbb{E}\left[\sum_{t\in[2^a(b-1)+1:2^a b]} \sum_{i\in\text{union of pools}} \left|\ell^t(i) - \widetilde{\ell^t}(i)\right|\right] \\
&\leqslant \sqrt{2^a} \cdot \operatorname{polylog}(nT),
\end{aligned}$$

其中，最后一个不等式利用了 引理 9.2 中高概率事件条件下池中的臂数。第三项也可以类似地界定为 $\sqrt{2^a} \cdot$ 多重对数(nT)。最后，对于第二项，我们注意到 $\bar{\ell}^t$ 仍然是 $\widetilde{\ell^t}$ 的无偏估计量，即

$$\begin{aligned}
\bar{\ell}^t &= \sum_{a,b:h(t,a,b)=1} \frac{w_{t,a,b}}{\sum_{a',b':h(t,a',b')=1} w_{t,a',b'}} \widetilde{\ell}(i_{t,a,b}) \\
&= \sum_{a,b:h(t,a,b)=1} p_{t,a,b} \widetilde{\ell}(i_{t,a,b}) = \mathbb{E}\left[\widetilde{\ell}(i_t)\right].
\end{aligned}$$

因此，我们可以对损失重新缩放，并在虚假损失上运行 SQUINT。由于以至少 $1-1/\operatorname{poly}(nT)$ 的概率有 $\ell^t(i) \leqslant$ 多重对数(nT)（这是因为 $K \leqslant$ 多重对数(nT)），这会使遗憾至多增加~多重对数(nT)倍。因此，我们得出结论：当标志 $\neq$ 为 True 时，区间 $D_{a,\ b}$ 上的遗憾也为 $\sqrt{2^a}$ 多重对数(nT)。

总结 引理 9.3的分析。 与 [PR23], 相同。为完成遗憾分析，我们注意到，对于任意区间 $\mathcal{I} = [t_1 : t_2] \subseteq T$ 可以将 $\mathcal{I}$ 划分为 $X \leqslant 2\log_2(|\mathcal{I}|)$ 个 互不相交的区间

$\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2 \cup \cdots \cup \mathcal{I}_X$，使得：（1）$\mathcal{I}_x$（$x \in [X]$）恰好覆盖某个元专家的生命周期；（2）至多存在两个长度为 2^x 的区间。于是我们得到

$$\sum_{t\in\mathcal{I}} \ell^t(i_t) - \sum_{t\in\mathcal{I}} \ell^t(i) \leqslant \sum_{x=1}^{X} \mathcal{O}\left(\sqrt{|\mathcal{I}_x|}\right) \operatorname{polylog}(nT)$$
$$\leqslant \sum_{x=1}^{\log(|\mathcal{I}|)} \mathcal{O}\left(\sqrt{2^x}\right) \operatorname{polylog}(nT) = \mathcal{O}\left(\sqrt{|\mathcal{I}|}\right) \operatorname{polylog}(nT).$$

为了提高区间遗憾的内存效率，我们可以使用该方法来维护$\{w_{a,b}\}$，最多对应 $\log(T)$ 个元专家。除了在 [PR23],中跟踪所需的信息外，我们只跟踪任意时刻池中各臂的损失，因此内存开销为 $\operatorname{polylog}(nT)$。因此，根据假设 $|\mathcal{U}| \leqslant \operatorname{polylog}(nT)$，内存总用量至多为 $\operatorname{polylog}(nT)$。

□

利用引理 9.3，现在可以得出MonocarpicExpert-Bandit算法的主要性质。

命题 9.1 (参见 [PR23]中的定理 4.10)。设 $T \geqslant 1$。对于任意在区间 $\mathcal{I} \subseteq [T]$内存活的专家 MonocarpicExpert-Bandit 算法保证，以至少 $1 - 1/\operatorname{poly}(nT)$的概率，

$$\sum_{t\in\mathcal{I}'} \ell^t(i_t) - \sum_{t\in\mathcal{I}'} \ell^t(i) \leqslant \sqrt{|\mathcal{I}'|} \cdot \operatorname{polylog}(nT).$$

对每个区间 $I' \subseteq I$ 均成立，即使对手是自适应的。此外，设 M 为任意时刻存活专家的最大数量，则 MonocarpicExpert-Bandit 算法最多使用 $M \operatorname{polylog}(nT)$ 个内存字数。

证明。证明过程与 [PR23] 中定理 4.10 的证明基本相同。设区间 I 为专家 i 的生命周期，按照 [PR23]，中相同的分析，我们可以将区间 $\mathrm{I}' \subseteq \mathrm{I}$ 划分为 $X = \log(|\mathrm{I}'|)$ 个长度为 2 的幂的子序列。设这些区间为 $\mathrm{I}' = \mathrm{I}'_1 \cup \mathrm{I}'_2 \cup \cdots \cup \mathrm{I}'_X$，而 $i_{t,k}$ 是 EXP_k 在第 t 天采取的动作。此外，设 $k(x)$ 为区间 $x \in X$ 对应的层级。根据引理 9.3，我们可以得到

$$\sum_{t\in\mathcal{I}'} \ell^t(i_t) - \sum_{t\in\mathcal{I}'} \ell^t(i) = \sum_{x=1}^{X} \sum_{t\in\mathcal{I}'_x} (\ell^t(i_t) - \ell^t(i))$$
$$= \underbrace{\sum_{x=1}^{X} \sum_{t\in\mathcal{I}'_x} (\ell^t(i_t) - \ell^t(i_{t,k(x)}))}_{\text{interval regret over } \mathrm{EXP}_k \text{ algorithms}} + \underbrace{\sum_{x=1}^{X} \sum_{t\in\mathcal{I}'_x} (\ell^t(i_{t,k(x)}) - \ell^t(i))}_{\text{interval regret for a single EXP algorithm}}$$
$$\leqslant \sum_{x=1}^{X} \sqrt{|\mathcal{I}'_x|} \cdot \operatorname{polylog}(nT) \qquad \text{(applying Lemma 9.3)}$$
$$\leqslant \sqrt{|\mathcal{I}'|} \cdot \operatorname{polylog}(nT). \qquad \text{(there are at most } \log(T) \text{ layers)}$$

内存分析基于以下事实：我们始终只维护 $\log(nT)$ 个 EXP 算法，而每个算法仅占用 $\operatorname{polylog}(nT)$ 的内存，如 引理 9.3所述。因此，每个存活专家占用 $\operatorname{polylog}(nT)$ 的内存，从而 $M \cdot \operatorname{polylog}(nT)$ 的内存可供 M 个存活专家使用。 □

9.3 最优提升算法的分析（算法 17）

我们现在　按照 算法 17 分析提升算法的内存占用和遗憾　　.

内存分析

内存　　分析现在在以下内容的帮助下变得非常简单：引理 9.2 和命题 9.1　　.

引理 9.4。 概率至少为 $1-1/\operatorname{poly}(nT)$，*算法 17* 使用的内存不超过 多重对数$(nT)$个字。

证明。 根据 引理 9.2，池的大小至多为 多重对数(nT)，这也是存活专家数量的上界。因此，根据 命题 9.1，使用的总内存至多为 多重对数(nT)，且其概率至少为 $1-1/$ 多项式(nT)。 □

遗憾分析

单周期专家提升的遗憾分析相当复杂，正如 [PR23] 中所述。然而，大部分分析可直接由 [PR23] 得出。首先，我们介绍 [PR23] 中用到的概念，如下所示。

[PR23] 使用的附加符号。 令 $i^* \in [n]$为最佳专家，并定义$\mathcal{K}_1, \cdots$、$\mathcal{K}_R$如下。

$$\mathcal{K}_1 = [0:n-1], \quad \mathcal{K}_2 = [0:n-1] \times \{0,1\}$$
$$\mathcal{K}_R = [0:n-1] \times \underbrace{\{0,1\} \times \cdots \times \{0,1\}}_{R-1 \text{ times}}.$$

此外，令 K 为 $\mathcal{K}_1$、. . . 、$\mathcal{K}_R$ 的并集，即：

$$\mathcal{K} = \mathcal{K}_1 \cup \cdots \cup \mathcal{K}_R.$$

对于任意时间步 $a = (a_1$、. . . 、$a_{r(a)}) \in K$（其中 $r(a)$ 的定义使得 $a \in \mathcal{K}^{r(a)}$），时间步 a 唯一确定了基线$^+(r(a))$ 的一个时期，即该算法第 $(\sum_{r=1}^{r(a)-1} a_r 2^{r(a)-r-1})$ 次重启中的第 $a_{r(a)}$ 个时期。

定义 6 ($\oplus$ 算子，定义 4.14，见 [PR23])。 对于任意时间步 $a \in \mathcal{K}$，我们记

$$a' = (a'_1, a'_2, \ldots, a'_{r(a')}) = a \oplus 1 \in \mathcal{K}$$

为满足以下条件的唯一时间步

$$\sum_{i=1}^{r(a')} a'_i B_i = \sum_{i=1}^{r(a)} a_i B_i + B_{r(a)} \quad and \quad a'_{r(a')} \neq 0 \quad .$$

也就是说，$a' = a \oplus 1$ 是末尾截断为 *0* 的下一个二进制数（第一个坐标除外，它属于 $[0: \ n-1]$）。

按照相同的方式 [PR23], 我们令$\mathcal{K}(a)$ 包含在 a 下 $\oplus$操作之后发生的所有时间步，即

$$\mathcal{K}(a) := \{a\} \cup \{a \oplus 1\} \cup \{(a \oplus 1) \oplus 1\} \cup \cdots \subseteq \mathcal{K}$$

中的随机位和时期分配 [PR23]。以下是 [PR23]中的随机位分配方案。我们直接沿用该方案，因此不解释其中的直觉。我们建议感兴趣的读者参阅 [PR23]了解详情。

定义 7（随机比特与主动/被动专家， [PR23] 中的定义 4.15）。对于任意时间步 $a \in K$，令随机比特为 $\xi_a = (\xi_{a,1}, \ldots, \xi_{a,n})$，其中 $\xi_{a,i} = (\xi_{a,i,1}, \xi_{a,i,2})$ 用于专家 $i (i \in [n])$。第一个坐标 $\xi_{a,i,1} \in \{0,1\}$ 是均值为 $1/n$ 的伯努利变量，用于抽样专家。随机比特的第二部分 $\xi_{a,i,2} \in \{0,1\}^{R \times 2L \times 2K}$ 来自均值为 $\frac{1}{\log^4(nT)}$ 的伯努利随机变量，用于估计池的规模并执行过滤过程。

在任意时间步 $a \in K$，若 $\xi_{a,\,i,\,2} = \vec{0}$，则称专家 $i \in [n]$ 为被动专家；否则称其为主动专家。

接下来，我们将讨论如何按照 [PR23] 中的相同方式，将轮次序列划分为一组互不相交的子序列。在此，我们只需论证：一旦其他来源的随机性固定，过滤集、存活的主动专家以及规模估计就可以被固定。其余步骤均可由 [PR23] 机械地推出。不过，在这里我们需要更加谨慎，因为每天只有两次查询。因此，我们在此处的主要观察与 [PR23] 略有不同。

观察 9.5 （参见 [PR23] 中的观察 4.16；这与 [PR23] 不同）。假设损失序列 $\{\ell^t\}_{t \in [T]}$ 以及抽样的主动专家集合 $\{Y_a\}_{a \in K}$ 已固定。此外，假设 每天第二次查询的随机性和损失序列 也已固定。那么，在 算法 17 的执行过程中任意时刻，规模估计值 s、过滤集 $\mathcal{F}$ 以及存活的主动专家集合也都已固定，而与抽样的被动专家集合无关。

请注意，观察 9.5 还要求固定第二次查询的随机性。这一点很重要，因为每个 EXP_k 对于 $k \in [K]$ (不要与 $\mathcal{K}$混淆，K 表示时间步）所作的决策都是第二次查询的函数；如果不固定这一随机性来源，不同线程可能会相互干扰。

现在，我们可以定义驱逐时间和轮次分配算法，其定义方式与 中[PR23] 的 定义8 和算法 22完全相同。

定义8（驱逐时间、单次增强， [PR23] 中的定义 4.17）。最佳专家 i^* 的驱逐时间定义如下。假设 i^* 在时间步 $a \in K$ 进入池中，则驱逐时间 $t(a) \in K(a) \cup \{+\infty\}$ 定义为最早的时间步，使得在 $t(a)$ 结束时， i^* 被存活的主动专家集合覆盖。如果 i^* 不会被覆盖，则令 $t(a) = +\infty$。

设 τ_{max} 为 算法 22生成的划分中的区间总数。由 [PR23] 可得以下关于 $\{\mathcal{I}_\tau\}_{\tau \in [\tau_{max}]}$ 的引理。

引理 9.6 ([PR23]中的引理 4.81)。 我们有

- 这些区间 $\{I_\tau\}_{\tau \in [\tau_{max}]}$ 两两不相交，且 $\bigcup_{\tau \in [\tau_{max}]} I_\tau = [T]$。
- 设 $L_1 = \{I_\tau\}_{\tau \in [\tau_{max}]}$，并设 $L_r := \{I \in L_1 : |I| < B_{r-1}\}$ $(r \in [2 : R])$ 中包含长度小于 B_{r-1}的区间，则

$$\sum_{\mathcal{I} \in L_r} |\mathcal{I}| \leqslant |\mathcal{H}_{r-1}| \cdot B_{r-1}.$$

沿着 [PR23], 的思路，接下来我们证明不良轮次的数量 $\mathcal{H}_r$ 以高概率至多为 多重对数 nT。

算法 22 分析用时期分配，算法 12，位于 [PR23]

初始化 $\mathcal{H}_r \leftarrow \emptyset$ $(r \in [R])$，$\tau \leftarrow 1$，$a_1 \leftarrow 0$
当 $\cup_{\tau' \leqslant \tau} \mathcal{I}_{\tau'} \neq [T]$ **时**
  如果 $t(a_\tau) = +\infty$ **则**
    $H_{r(a_\tau)} \leftarrow H_{r(a_\tau)} \cup \{a_\tau\}$ ▷i^* 存活至结束
    如果 $r(a_\tau) = R$ **那么** ▷a_τ 是线程 $r(a_\tau)$ 中的不良轮次
      $I_\tau \leftarrow \{a_\tau\}$，$a_{\tau+1} \leftarrow a_\tau \oplus 1$，$\tau \leftarrow \tau + 1$ ▷停在顶层线程
    else
      $a_\tau \leftarrow (a_\tau,\ 0)$ ▷移至下一个线程
  else
    $a_{\tau+1} \leftarrow t(a_\tau) \oplus 1$，$\mathcal{I}_\tau \leftarrow [a_\tau :\ t(a_\tau)]$，$\tau \leftarrow \tau + 1$

引理 9.7 （参见 [PR23] 中的引理 4.19）。 其成立的概率至少为 $1 - 1/\operatorname{poly}(nT)$，$|\mathcal{H}_r| \leqslant n \operatorname{polylog}(nT)$ 对任意 $r \in [R]$均成立。

证明。 证明与 引理 4.5 以及 [PR23],中的引理 4.19 的证明非常相似，因此为避免不必要的重复，我们仅给出概要。首先，我们对以下内容进行条件化： ℓ^1的损失序列，ℓ^2、$\cdots$ ℓ^T、抽样的主动专家 $\{Y_a\}_{a \in \mathcal{K}}$，以及算法第二次查询所用的所有随机性。然后可以论证，一旦 $i^* \in W_a$，也就是说，时间步 a的抽样被动专家集合，此后便会一直保留到最后。此外，我们还可以得到

$$\mathbf{Pr}\left[i^* \in W_a \,|\, Y_1, \ldots, Y_{T/B}, \ell^1, \ldots, \ell^T, \text{randomness of the second query}\right] \geqslant \frac{1}{2n}$$

根据 引理 4.5中的相同论证。因此，对于任意固定的 r，我们有

$$\Pr\left(|\mathcal{P}| \leqslant \frac{|\mathcal{H}_r|}{4n} \mid Y_1, \ldots, Y_{T/B}, \ell^1, \ldots, \ell^T, \text{randomness of the second query}\right) \leqslant \exp\left(-|\mathcal{H}_r|/16n\right).$$

根据 $|\mathcal{P}| \leqslant$ 多重对数 nT，如 引理 9.2所示，我们得出，以至少$1 - 1/$ 多项式(nT)的概率，有 $|\mathcal{H}_r| \leqslant n \cdot$多重对数$(nT)$，如所需。 □

接下来，我们以与 [PR23], 相同的方式，应用MonocarpicExpert-Bandit（命题 9.1）给出的保证，来界定线程 R 上非糟糕时期的代价。

引理 9.8。 以至少 $1 - 1/\operatorname{poly}(nT)$ 的概率，对任意 $\tau \in [\tau_{max}]$ 和 $a_\tau \notin H_R$，

$$\sum_{t \in \mathcal{I}_\tau} \ell^t(i_t) - \sum_{t \in \mathcal{I}_\tau} \ell^t(i^*) \leqslant \sqrt{|\mathcal{I}_\tau|} \cdot \operatorname{polylog}(nT).$$

证明。 类似于 [PR23], 中的证明，我们对引理 9.2中的事件进行条件化。给定任意区间$\mathcal{I}_\tau$，其起点为 a_τ、终点为 $t(a_\tau)$，且 $a_\tau \notin \mathcal{H}_R$，进入时间为 a_τ的专家 i^*在 $t(a_\tau)$结束时被$\mathcal{P}$覆盖。令 $i_1^*, \ldots, i_s^*$ 为覆盖 i^*的专家集合。注意，根据定义 1 中的覆盖概念，并通过正确选择 ρ，我们可以将区间$\mathcal{I}_\tau = \mathcal{I}_{\tau,1} \cup \cdots \cup \mathcal{I}_{\tau,s}$划分为若干部分，并得到

$$\sum_{t \in \mathcal{I}_\tau} \widetilde{\mathcal{L}^{\mathcal{I}_\tau}}(i^*) \geqslant \sum_{j=1}^{s} \sum_{t \in \mathcal{I}_{\tau,j}} \widetilde{\mathcal{L}^{\mathcal{I}_\tau}}(i_j^*) - C \log n \cdot \sqrt{|\mathcal{I}_\tau|}$$

此外，进一步对引理 9.1中的高概率事件进行条件化，我们有

$$\left|\widetilde{\mathcal{L}^{\mathcal{I}_\tau}}(i) - \sum_{t\in\mathcal{I}_\tau} \ell^t(i)\right| \leqslant \sqrt{|\mathcal{I}_\tau|}\cdot \mathrm{polylog}(nT).$$

因此，我们可以用过滤臂的代价给 i^* 的代价提供一个下界，即

$$\sum_{t\in\mathcal{I}_\tau} \ell^t(i^*) \geqslant \sum_{j=1}^{s}\left(\sum_{t\in\mathcal{I}_{\tau,j}} \ell^t(i_j^*) - C_1\cdot\sqrt{\mathcal{I}_{\tau,j}}\cdot\mathrm{polylog}(nT)\right) \geqslant \sum_{j=1}^{s}\sum_{t\in\mathcal{I}_{\tau,j}} \ell^t(i_j^*) - C_2\cdot\sqrt{\mathcal{I}_\tau}\cdot\mathrm{polylog}(nT) \tag{5}$$

对于某个常数 C_1 和 C_2，满足 $C_2 \geqslant C_1$。在计算中，第二步用到了 $s \leqslant |\mathcal{P}| \leqslant$ 多重对数(n)。因此，根据 MonocarpicExpert-Bandit 的遗憾保证，以至少 $1-1/$ 多项式(nT) 的概率，我们有

$$\begin{aligned}
\sum_{t\in\mathcal{I}_\tau} \ell^t(i_t) - \sum_{t\in\mathcal{I}_\tau} \ell^t(i^*) &= \sum_{t\in\mathcal{I}_\tau} \ell^t(i_t) - \sum_{j=1}^{s}\sum_{t\in\mathcal{I}_{\tau,j}} \ell^t(i_j^*) + \sum_{j=1}^{s}\sum_{t\in\mathcal{I}_{\tau,j}} \ell^t(i_j^*) - \sum_{t\in\mathcal{I}_\tau} \ell^t(i^*) \\
&= \sum_{j=1}^{s}\sum_{t\in\mathcal{I}_{\tau,j}} \ell^t(i_t) - \sum_{j=1}^{s}\sum_{t\in\mathcal{I}_{\tau,j}} \ell^t(i_j^*) + \sum_{j=1}^{s}\sum_{t\in\mathcal{I}_{\tau,j}} \ell^t(i_j^*) - \sum_{t\in\mathcal{I}_\tau} \ell^t(i^*) \\
&\qquad\qquad \text{(intervals are disjoint)} \\
&\leqslant \sum_{j=1}^{s}\sqrt{|\mathcal{I}_{\tau,j}|}\cdot\mathrm{polylog}(nT) + \sum_{j=1}^{s}\sum_{t\in\mathcal{I}_{\tau,j}} \ell^t(i_j^*) - \sum_{t\in\mathcal{I}_\tau} \ell^t(i^*) \\
&\qquad\qquad \text{(by Proposition 9.1)} \\
&\leqslant \sum_{j=1}^{s}\sqrt{|\mathcal{I}_{\tau,j}|}\cdot\mathrm{polylog}(nT) + \sqrt{\mathcal{I}_\tau}\cdot\mathrm{polylog}(nT) \qquad \text{(by Eq (5))} \\
&\leqslant \sqrt{\mathcal{I}_\tau}\cdot\mathrm{polylog}(nT),
\end{aligned}$$

其中，最后一步再次用到了事实 $s \leqslant |\mathcal{P}| \leqslant$ 多重对数(n)。这正是引理陈述所要求的。

□

现在，我们以与算法 17 相同的方式界定其遗憾 [PR23]。

引理 9.9。 以至少 $1-1/$ 多项式(nT) 的概率，*算法 17* 的遗憾至多为

$$\sum_{t\in[T]} \ell^t(i_t) - \sum_{t\in[T]} \ell^t(i^*) \leqslant \sqrt{nT}\cdot\mathrm{polylog}(nT).$$

证明。 证明沿用 [PR23] 中的相同论证，并对我们所作的陈述修改作相应调整。与他们的证明相同，我们首先固定损失序列 $\{\ell^t\}_{t\in[T]}$ 以及抽样的主动专家集合 $\{Y_a\}_{a\in\mathcal{K}}$。此外，我们固定第二次查询的随机性，这使得遗憾可以分解到一组区间 $\{\mathcal{I}_\tau\}_{\tau\in[\tau_{max}]}$，并使用轮次分配算法。

现在，我们根据遗憾的长度是否大于 B_{R-1}对其进行分类。以至少 $1-1/\mathrm{poly}(nT)$ 的概率，我们有

$$\sum_{t\in[T]} \ell^t(i_t) - \sum_{t\in[T]} \ell^t(i^*) = \sum_{r=1}^{R-1}\sum_{\mathcal{I}\in L_r\setminus L_{r+1}}\sum_{t\in\mathcal{I}}(\ell^t(i_t) - \ell^t(i^*)) + \sum_{\mathcal{I}\in L_R}\sum_{t\in\mathcal{I}}(\ell^t(i_t) - \ell^t(i^*))$$

(by using $[T] = \cup_{\mathcal{I}\in L_1}\mathcal{I}$ of Lemma 9.6)

$$\leqslant \sum_{r=1}^{R-1} \sum_{\mathcal{I}\in L_r\setminus L_{r+1}} \sum_{t\in\mathcal{I}} (\ell^t(i_t) - \ell^t(i^*)) + \mathcal{O}\left(|\mathcal{H}_{R-1}|\right).$$

(利用引理 9.6的第二个性质)

应用引理 9.8，我们可以得到

$$\sum_{t\in[T]} \ell^t(i_t) - \sum_{t\in[T]} \ell^t(i^*) \leqslant \sum_{r=1}^{R-1} \sum_{\mathcal{I}\in L_r\setminus L_{r+1}} \sqrt{|\mathcal{I}|}\cdot \mathrm{polylog}(nT) + \mathcal{O}\left(|\mathcal{H}_{R-1}|\right). \tag{6}$$

通过与 [PR23], 中完全相同的计算，可得

$$\sum_{\mathcal{I}\in L_r\setminus L_{r+1}} \sqrt{|\mathcal{I}|} \leqslant \begin{cases} |\mathcal{H}_{r-1}|\cdot\sqrt{2T/n} & r\geqslant 2 \\ \sqrt{nT} & r=1 \end{cases}$$

for 任意 $r\in[r-1]$。因此，现在我们可以使用公式 (6) 最终给遗憾设定上界 as

$$\begin{aligned}
\sum_{t\in[T]} \ell^t(i_t) - \sum_{t\in[T]} \ell^t(i^*) &\leqslant \sum_{r=1}^{R-1} \sum_{\mathcal{I}\in L_r\setminus L_{r+1}} \sqrt{|\mathcal{I}|}\cdot\mathrm{polylog}(nT) + \mathcal{O}\left(|\mathcal{H}_{R-1}|\right) && \text{(Eq (6))} \\
&\leqslant \sqrt{nT}\cdot\mathrm{polylog}(nT) + \sqrt{\frac{2T}{n}}\cdot\mathrm{polylog}(nT)\cdot\sum_{r=2}^{R}\sum_{\mathcal{I}\in L_r\setminus L_{r+1}} |\mathcal{H}_{r-1}| \\
&\leqslant \sqrt{nT}\cdot\mathrm{polylog}(nT) + \sqrt{\frac{2T}{n}}\cdot\mathrm{polylog}(nT)\cdot R\cdot n\,\mathrm{polylog}(nT) \\
& && \text{(by Lemma 9.7)} \\
&\leqslant \sqrt{nT}\cdot\mathrm{polylog}(nT),
\end{aligned}$$

这正如引理陈述所要求的。 □

完成定理 9的证明。结合引理9.4(关于空间)和引理9.9 (关于遗憾)，即可得到定理 9中的结论。

9.4 关于 $\tilde{O}(\sqrt{nT})$的区间遗憾讨论

我们简要讨论如何将上述结果推广到区间遗憾，从而得到定理 1。

算法。区间遗憾算法沿用算法 11的相同结构（不要与算法 20混淆）。我们沿用算法 11 中的相同结构，设置$N^{\mathrm{meta}} = \log(T)$种不同的区间长度，并将 $\texttt{ALG}_\kappa$ 初始化为算法 17 的副本，其天数为 2^κ。我们令$\{w_t(\kappa)\}_{\kappa=1}^{N^{\mathrm{meta}}}$ 表示每个区间算法的权重。每天，我们按照$^{w_t(\kappa)}/_{\sum_\kappa w_t(\kappa)}$上的分布选择一个区间算法，并遵循

决策 $\texttt{ALG}_\kappa$ （如果被采样，即遵循算法 19的决策）。不过，此处我们省略了所有更新步骤。另一方面，对于更新步骤，我们采用以下步骤。我们抛一枚公平硬币；如果硬币显示“正面”（概率为 $1/2$），则均匀随机地采样一条臂，并使用损失 $\ell^t(i)$ 来更新算法 17中的算法 17。如果硬币显示“反面”（概率为 $1/2$），则使用算法 20和标志 $=$ Fake 进行更新。池̃管理的所有更新步骤（即合并和成员关系更新）都通过使用每条臂$\mathcal{L}(i)$的估计值 i来完成。

分析。这遵循了我们在定理 7证明中采用的相同思路。第一步是证明，经过修改后，算法的正确性仍然成立。事实上，由于我们继续均匀随机地采样臂，引理 9.1中的保证依然成立。对于算法 20中的更新，我们在引理 9.3中证明了，使用估计损失进行更新能够为存活专家带来所需的区间遗憾。此外，所有其他更新步骤均保持不变，这意味着空间和遗憾界仍然成立。

分析中最后一个尚未完成的部分，是证明外层算法的正确性，也就是得到引理 6.3中的保证。在证明中，我们用到的关键性质是，对于池中的每个臂 i，在更新步骤中臂 i被采样的概率至少为 $\frac{1}{\text{多重对数}(nT)}$。注意，根据引理 9.2，上述算法中这一性质仍然成立。因此，我们完成了对定理 1的分析。

致谢

David P. Woodruff 部分经费由美国海军研究办公室项目（编号 N000142112647）、Simons Investigator 奖以及美国国家科学基金会 CCF-2335412 资助。Samson Zhou 部分经费由美国国家科学基金会 CCF-2335411 资助。Samson Zhou 衷心感谢橡树岭联合大学（ORAU）Ralph E. Powe 初级教职提升奖提供的资助。

参考文献

[ACFS02] Peter Auer、Nicolò Cesa-Bianchi、Yoav Freund 和 Robert E. Schapire。《非随机多臂老虎机问题》。《SIAM J. Comput.》，32(1):48–77，2002年。3，7，12[ACNS23] Anders Aamand、Justin Y. Chen、Huy Lê Nguyen 和 Sandeep Silwal。《专家学习的空间界改进》。CoRR，abs/2303.01453，2023年。1[App] Apple。https://images.apple.com/privacy/docs/Differential_Privacy_Overview.pdf。2[BBD+02] Brian Babcock、Shivnath Babu、Mayur Datar、Rajeev Motwani 和 Jennifer Widom。《数据流系统中的模型与问题》。载于第二十一届 ACM SIGACT-SIGMOD-SIGART 数据库系统原理研讨会论文集，第1–16页，2002年。1[BDM+20] Vladimir Braverman、Petros Drineas、Cameron Musco、Christopher Musco、Jalaj Upadhyay、David P. Woodruff 和 Samson Zhou。《在线模型和滑动窗口模型中的近乎最优线性代数》。载于第61届 IEEE 计算机科学基础年会（FOCS），第517–528页，2020年。5[BDMO03] Brian Babcock、Mayur Datar、Rajeev Motwani 和 Liadan O'Callaghan。《在数据流窗口上维护方差和 k-中位数》。载于第二十二届 ACM SIGACT-SIGMOD-SIGART 数据库系统原理研讨会论文集，第234–243页，2003年。1[BEL+20] Michele Borassi、Alessandro Epasto、Silvio Lattanzi、Sergei Vassilvitskii 和 Morteza Zadimoghaddam。《k-聚类问题的滑动窗口算法》。载于《神经信息处理系统进展》

33：神经信息处理系统年会（NeurIPS），2020年。2，4[Ber24] S. N. Bernstein。《关于切比雪夫不等式及拉普拉斯误差公式的一种修正》。乌克兰科学院学报，数学辑，1(4):38–49，1924年。俄文。12[BGL+18] Vladimir Braverman、Elena Grigorescu、Harry Lang、David P. Woodruff 和 Samson Zhou。《滑动窗口上的近乎最优不同元素计数与重频项》。载于《近似、随机化与组合优化：算法与技术》，APPROX/RANDOM，第7:1–7:22页，2018年。2，4[BLLM15] Vladimir Braverman、Harry Lang、Keith D. Levin 和 Morteza Monemizadeh。《多对数空间中的滑动窗口聚类》。载于第35届 IARCS 软件技术基础与理论计算机科学年会（FSTTCS），第350–364页，2015年。2，4[BLLM16] Vladimir Braverman、Harry Lang、Keith D. Levin 和 Morteza Monemizadeh。《滑动窗口上的聚类问题》。载于第二十七届 ACM-SIAM 离散算法研讨会年会论文集（SODA），第1374–1390页，2016年。2，4[BLMZ23] Jeremiah Blocki、Seunghoon Lee、Tamalika Mukherjee 和 Samson Zhou。《滑动窗口模型中的差分隐私重频项》。载于第十一届国际学习表征会议（ICLR），2023年。2，4[BO07] Vladimir Braverman 和 Rafail Ostrovsky。《滑动窗口的平滑直方图》。载于第48届 IEEE 计算机科学基础年会（FOCS）论文集，第283–293页，2007年。2，4[BWZ21] Vladimir Braverman、Viska Wei 和 Samson Zhou。《滑动窗口上的对称范数估计与回归》。载于《计算与组合数学——第27届国际会议》（COCOON）论文集，2021年。2，4[CFH+97] Nicolò Cesa-Bianchi、Yoav Freund、David Haussler、David P. Helmbold、Robert E. Schapire 和 Manfred K. Warmuth。《如何利用专家建议》。《ACM学报》，44(3):427–485，1997年。1[CG07] Graham Cormode 和 Minos Garofalakis。《互联世界中的流处理：查询和跟踪分布式数据流》。载于2007年 ACM SIGMOD 数据管理国际会议论文集，第1178–1181页，2007年。2[CJY+25] Vincent Cohen-Addad、Shaofeng H.-C. Jiang、Qiaoyuan Yang、Yubo Zhang 和 Samson Zhou。《滑动窗口模型中的公平聚类》。载于第十三届国际学习表征会议（ICLR），2025年。2，5[CM05] Graham Cormode 和 Shanmugavelayutham Muthukrishnan。《新动态：发现网络数据流中的显著差异》。《IEEE/ACM 网络学报》，13(6):1219–1232，2005年。2

[CMS13] Michael S. Crouch、Andrew McGregor 和 Daniel M. Stubbs。滑动窗口模型中的动态图。载于《算法——ESA 第21届欧洲年会》论文集，第337–348页，2013年。2[CNZ16] Jiecao Chen、Huy L. Nguyen 和 Qin Zhang。滑动窗口上的子模最大化。CoRR，abs/1611.00129，2016年。2, 4[Cor13] Graham Cormode。连续分布式监测模型。SIGMOD Rec.，42(1):5–14，2013年。2[Cov66] Thomas M Cover。二进制序列序贯预测器的行为。斯坦福电子实验室技术报告第7002号。斯坦福大学，系统理论，1966年。1[DGIM02] Mayur Datar、Aristides Gionis、Piotr Indyk 和 Rajeev Motwani。维护滑动窗口上的流统计量。SIAM J. Comput.，31(6):1794–1813，2002年。2, 4[DGS15] Amit Daniely、Alon Gonen 和 Shai Shalev-Shwartz。强自适应在线学习。载于 Francis R. Bach 和 David M. Blei 编，《第32届国际机器学习会议论文集》，ICML 2015，法国里尔，2015年7月6日至11日，第37卷《JMLR研讨会与会议论文集》，第1405–1411页。JMLR.org，2015年。2, 3, 4, 5, 9, 30[ELVZ17] Alessandro Epasto、Silvio Lattanzi、Sergei Vassilvitskii 和 Morteza Zadimoghaddam。滑动窗口上的子模优化。载于第26届国际万维网会议论文集，WWW，第421–430页，2017年。2, 4[EMMZ22] Alessandro Epasto、Mohammad Mahdian、Vahab S. Mirrokni 和 Peilin Zhong。通过基于分桶的草图改进聚类和覆盖的滑动窗口算法。载于2022年ACM-SIAM离散算法研讨会论文集，SODA，第3005–3042页，2022年。2, 4[FS97] Yoav Freund 和 Robert E. Schapire。在线学习的决策理论推广及其在提升中的应用。计算机与系统科学杂志，55(1):119–139，1997年。1[GDP] 欧盟《通用数据保护条例》（条例 (EU) 2016/679）：欧洲议会和欧盟理事会于2016年4月27日颁布，涉及个人数据处理过程中的自然人保护及此类数据的自由流动，并废除指令95/46/EC。2[GM09] Sudipto Guha 和 Andrew McGregor。流顺序与顺序统计量：随机顺序流中的分位数估计。SIAM 计算学杂志，38(5):2044–2059，2009年。4[Goo] 谷歌。https://policies.google.com/technologies/retention。2[Hoe94] Wassily Hoeffding。有界随机变量之和的概率不等式。Wassily Hoeffding 文集，第409–426页，1994年。37

[JWZ22] Rajesh Jayaram、David P. Woodruff 和 Samson Zhou。数据流和滑动窗口的真正完美采样器。载于PODS '22：国际数据管理会议，第29–40页，2022年。2[KV05] Adam Tauman Kalai 和 Santosh S. Vempala。在线决策问题的高效算法。计算机与系统科学杂志，71(3):291–307，2005年。1[KvE15] Wouter M. Koolen 和 Tim van Erven。专家与组合博弈的二阶分位数方法。载于 Peter Grünwald、Elad Hazan 和 Satyen Kale 编，《第28届学习理论会议论文集》，COLT 2015，法国巴黎，2015年7月3日至6日，第40卷《JMLR研讨会与会议论文集》，第1155–1175页。JMLR.org，2015年。14, 15[LT06a] Lap-Kei Lee 和 H. F. Ting。维护滑动窗口上的重要流统计量。载于第十七届ACM-SIAM离散算法年度研讨会论文集，SODA，第724–732页，2006年。2[LT06b] Lap-Kei Lee 和 H. F. Ting。一种更简单、更高效的确定性方案，用于在滑动窗口中查找频繁项。载于第二十五届ACM SIGACT-SIGMOD-SIGART数据库系统原理研讨会论文集，第290–297页，2006年。2[LW94] Nick Littlestone 和 Manfred K. Warmuth。加权多数算法。信息与计算，108(2):212–261，1994年。1[LZC+24] Zhou Lu、Qiuyi Zhang、Xinyi Chen、Fred Zhang、David P. Woodruff 和 Elad Hazan。多臂老虎机的自适应遗憾成为可能：两次查询即可实现。载于第十二届国际学习表征会议，ICLR 2024，奥地利维也纳，2024年5月7日至11日。OpenReview.net，2024年。3, 5, 8, 9, 27, 30, 60[OMM+14] Miles Osborne、Sean Moran、Richard McCreadie、Alexander Von Lunen、Martin Sykora、Elizabeth Cano、Neil Ireson、Craig Macdonald、Iadh Ounis、Yulan He 等。社交媒体中自动发现事件的实时检测、跟踪与监测。载于计算语言学协会第52届年会系统演示论文集，第37–42页，2014年。2[Ope] OpenAI。https://openai.com/enterprise-privacy/。2[PGD15] Odysseas Papapetrou、Minos N. Garofalakis 和 Antonios Deligiannakis。分布式滑动窗口数据流的草图化。VLDB期刊，24(3):345–368，2015年。1[PR23] Binghui Peng 和 Aviad Rubinstein。在线学习中近乎最优的内存-遗憾权衡。载于2023年第64届IEEE计算机科学基础年度研讨会（FOCS），第1171–1194页。IEEE，2023年。1, 4, 6, 14, 15, 16, 20, 21, 24, 40, 41, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53[PZ23] Binghui Peng 和 Fred Zhang。次线性空间中的在线预测。载于2023年ACM-SIAM离散算法研讨会论文集，SODA，第1611–1634页，2023年。1, 4, 6

[SWXZ22] Vaidehi Srinivas、David P. Woodruff、Ziyu Xu 和 Samson Zhou。专家问题的内存上界。载于 STOC '22：第 54 届 ACM SIGACT 计算理论研讨会，页码 1158–1171，2022年。1、4、6[WY23] David P. Woodruff 和 Taisuke Yasuda。在线 Lewis 权重采样。载于《2023年ACM-SIAM离散算法研讨会论文集》，SODA，页码 4622–4666，2023年。2、5[WZ21] David P. Woodruff 和 Samson Zhou。通过差分估计器获得对抗鲁棒数据流和滑动窗口的紧确界。载于第 62 届 IEEE 计算机科学基础年会，FOCS，页码 1183–1196，2021年。2[WZZ23a] David P. Woodruff、Fred Zhang 和 Samson Zhou。面向专家学习的鲁棒流处理：算法与下界。载于《神经信息处理系统进展》第 36 卷：2023年神经信息处理系统年会，NeurIPS，2023年。1、6[WZZ23b] David P. Woodruff、Peilin Zhong 和 Samson Zhou。滑动窗口模型中的近最优 k 聚类。载于《神经信息处理系统进展》第 36 卷：神经信息处理系统年会，NeurIPS，2023年。2、5[XZ21] Xiao Xu 和 Qing Zhao。对抗性多臂老虎机中的内存受限无悔学习。《IEEE 信号处理汇刊》，69：2371–2382，2021年。3、6、15

第 3 节中的省略证明

我们指出，命题 3.3 和命题 3.4 的陈述和证明都是标准的；为求完整起见，我们在此给出。

命题 3.3 的证明

证明。第一步是将遗憾分解为探索步骤和利用步骤上的遗憾。令$\mathcal{D}_{\text{explore}}$ 为算法 2中的探索日，令 $\mathcal{D}_{\text{exploit}}$ 为算法 2从分布中抽取臂的天数。由于损失取值在$\{0,\ 1\}$内，且期望探索步数为 γT，我们可以直接得出如下界：

$$\mathbb{E}\left[\sum_{t\in\mathcal{D}_{\text{explore}}}\ell^t(i_t)-\sum_{t\in\mathcal{D}_{\text{explore}}}\ell^t(i^*)\right]\leqslant\gamma T.$$

剩下要证明的是利用日上的遗憾。在某个利用日，令我们考虑一个虚构的损失序列$\{\widehat{\ell^t}(i)\}$，它通过以下过程生成：

$$\widehat{\ell^t}(i)=\begin{cases}\widetilde{\ell^t}(i),\ \text{if } t\in\mathcal{D}_{\text{explore}};\\ 0,\ \text{otherwise}\end{cases}$$

接下来，我们将遗憾分解为

$$\mathbb{E}\left[\sum_{t\in\mathcal{D}_{\text{exploit}}}\ell^t(i_t)-\sum_{t\in\mathcal{D}_{\text{exploit}}}\ell^t(i^*)\right]\leqslant\mathbb{E}\left[\sum_{t=1}^{T}\ell^t(j_t)-\sum_{t=1}^{T}\widehat{\ell^t}(j_t)\right]+\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\ell^t}(i_t)-\sum_{t=1}^{T}\ell^t(i^*)\right].$$

我们利用集中不等式对 $\mathbb{E}\left[\sum_{t=1}^{T}\ell^t(j_t)-\sum_{t=1}^{T}\widehat{\ell}^t(j_t)\right]$ 进行界定。特别要注意，$\widehat{\ell^t}\widehat{\mathcal{L}}^t:=\sum_{\tau=1}^{t}\widehat{\ell}^t$ 是第 t 天的无偏估计量，且方差较低，这意味着

$$\mathbb{E}\left[\widehat{\ell}^t\right]=\ell^t(i),\qquad \mathbb{E}\left[(\widehat{\ell^t(i)})^2\right]=(\ell^t(i))^2\cdot\frac{n}{\gamma}.$$

因此，应用伯恩斯坦不等式可得

$$\Pr\left(\left|\widehat{\mathcal{L}}^t(i)-\sum_{t=1}^{T}\ell^t(i)\right|\geqslant\sqrt{T/\gamma}\cdot n\cdot\mathrm{polylog}(nT)\right)\leqslant\frac{1}{\mathrm{poly}(nT)},$$

由此得到第一项的遗憾界。

为了界定第二项的遗憾，即 $\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\ell}^t(i_t)-\sum_{t=1}^{T}\ell^t(i^*)\right]$，我们采用一种从部分信息到完全信息的归约论证来界定遗憾（思路与使用乘法权重更新证明 EXP3 的遗憾界相同）。我们首先回顾完全信息 Hedge 算法的保证如下。

命题 A.1。 考虑以下 *Hedge* 算法：

学习率为 η 的 *Hedge* 算法

- 在时间 t，该算法接收损失向量 $\ell^t\in[0,U]^n$。
- 令 $L^t(i)$ 表示截至时间 t时 i的总损失。
- 该算法通过从以下分布中采样来选择 i_t 。

$$\overline{P_t}(i)=\frac{\exp(-\eta\cdot\mathcal{L}^t(i))}{\sum_{i=1}^{n}\exp(-\eta\cdot\mathcal{L}^t(i))}.$$

- 该算法通过观察 ℓ^t中的损失，更新 $i\in[n]$ 的所有 $L^{t+1}(i)$ 。

H 边缘算法的期望遗憾至多为 $O\left(\sqrt{UT\log n}\right)$，只要 *s*

$$\mathbb{E}\left[\sum_{i=1}^{n}\overline{P_t}(i)(\overline{\ell^t}(i))^2\right]\leqslant U,$$

即使损失是自适应选择的。

我们注意到，本质上，算法 2 可以视为一个 HEDGE 算法，其中 ℓ^t 为 ℓ^t，而 $\eta=\gamma$。由于 $\widehat{\ell}^t$ 是 ℓ^t的无偏估计量，令 R^{HEDGE} 表示上述设置下 Hedge 算法的遗憾，则有

$$\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\ell}^t(i_t)-\sum_{t=1}^{T}\ell^t(i^*)\right]=\mathbb{E}\left[R^{\mathrm{HEDGE}}\right].$$

因此，剩下的工作就是界定 Hedge 算法的遗憾。根据抽样规则，我们有 $\widehat{\ell}^t\leqslant n/\gamma$。此外，我们可以如下界定 $\mathbb{E}\left[\sum_{i=1}^{n}P_t(i)(\ell^t(i))^2\right]$。

$$\mathbb{E}\left[\sum_{i=1}^{n}\overline{P_t}(i)(\overline{\ell^t}(i))^2\right]=\mathbb{E}\left[\sum_{i=1}^{n}P_t(i)(\widehat{\ell}^t(i))^2\right]=\mathbb{E}\left[\sum_{i=1}^{n}\frac{\gamma}{n}\cdot(\frac{n}{\gamma})^2\right]\leqslant n^2/\gamma.$$

因此，根据 命题 A.1，第二部分的期望遗憾也至多为 $\mathcal{O}\left(n\sqrt{T\log n/\gamma}\right)$。这意味着

$$\mathbb{E}\left[\sum_{t\in\mathcal{D}_{\text{exploit}}}\ell^t(i_t)-\sum_{t\in\mathcal{D}_{\text{exploit}}}\ell^t(i^*)\right]\leqslant\mathbb{E}\left[\sum_{t=1}^{T}\ell^t(j_t)-\sum_{t=1}^{T}\widehat{\ell}^t(j_t)\right]+\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\ell}^t(i_t)-\sum_{t=1}^{T}\ell^t(i^*)\right]$$
$$\leqslant n\sqrt{\frac{T\log n}{\gamma}}\cdot\operatorname{polylog}(nT).$$

将此结果与探索日上的损失结合起来，即可得到所需结果。 □

命题的证明 命题 3.4

证明。 我们使用 [LZC+24] 中的以下引理来证明 命题 3.4。

引理 *A.1*（[*LZC*+24]，改述）。设 $\hat{\ell}_t$ 是损失向量 ℓ^t的一个无偏估计量，使得对于某个分布 z_t，以概率 $z_t(i)$ $\hat{\ell}_t(i)=\frac{1}{z_t(i)}\ell^t(i)$ 且 $\hat{\ell}_t(j)=0$ 对于 $j\neq i$成立。此外，假设 $z_t(i)$ 满足 $P_t(i)\leqslant C\cdot z_t(i)$，对所有 i均成立。那么，采用学习率 $\eta=\sqrt{\log n/CTn}$ 的算法 3 可达到 $O\left(\sqrt{CnT\log n}\right)$ 的期望遗憾。

由于我们对每个臂均匀随机地进行采样，因此有 $z_t(i)=1/n$。由于对于任意i和 t，$P_t(i)\leqslant n$ 成立，因此我们有 $C\leqslant n$；通过设置 $\eta=\frac{1}{n}\cdot\sqrt{\frac{\log T}{n}}$，可得到 $\mathcal{O}\left(n\cdot\sqrt{T\log n}\right)$ 的期望遗憾。 □