arXiv:2302.06014v2 [cs.LG] 6 Feb 2024

具有折扣自适应偏好的在线代理推荐

Arpit Agarwal* William Brown†

2024年2月7日

摘要

我们研究一个老虎机推荐问题，其中智能体的偏好（表示对推荐物品的选择概率）会根据过去的选择，按照一个未知*preference model*演化。每一轮，我们向智能体展示一个包含 k 件物品（总共 n 件物品）的菜单，随后智能体选择一件物品；我们的目标是针对智能体选择所产生的对抗性损失，相对于某个 *target set*（项目单纯形的一个子集）最小化遗憾。我们将 Agarwal 和 Brown(2022) 的设定扩展到这里；他们研究了均匀记忆代理，而我们允许非均匀记忆，即在之后的每一轮中，将折扣因子应用于智能体的记忆向量。在“长期记忆”情形下（有效记忆时域按 T 次线性增长），我们证明，对于任意 *everywhere instantaneously realizabledistributions* 偏好模型，相对于 *smooth*（即先前工作中定义的“EIRD 集”），均可实现高效的次线性遗憾。此外，对于由记忆权重的线性函数上下界定的偏好（我们称之为“尺度有界”偏好），我们给出一种算法，能够相对于几乎整个 *entire* 项目单纯形实现高效的次线性遗憾。我们证明，在一般情况下，将目标扩展到 EIRD 集之外是 NP 困难的。在“短期记忆”情形下（记忆时域为常数），我们证明，当损失变化不太频繁时，即使没有光滑性，尺度有界偏好仍能使我们在几乎整个单纯形上实现高效的次线性遗憾；然而，即使损失保持不变，对于任意光滑偏好模型，我们也证明了一个信息论障碍：无法与 EIRD 集相竞争。

1 引言

推荐系统是电子商务、社交网络和内容分享等在线平台不可或缺的一部分。已有充分文献表明，用户偏好会随着时间推移而变化，以响应内容推荐（Flaxman 等人，2016；Curmei 等人，2022；Abdollahpouri 等人，2019；Mansoury 等人，2020），从而形成自我强化的内容消费模式，并可能给用户、平台和社会带来各种意想不到的后果，包括但不限于收入损失，以及形成推动极化的“过滤气泡”和“回音室”。因此，推荐系统应将这些适应性模式纳入用户建模，以捕捉推荐对智能体偏好动态的长期影响。

我们在 Agarwal 和 Brown (2022) 提出的框架基础上，将在线推荐建模为在智能体偏好空间上的非线性强化学习问题。推荐系统的目标通常是优化智能体行为的某种函数，例如商品购买或广告收入，但我们无法强迫智能体选择向其展示的任何特定选项。当偏好具有适应性时，我们在每次交互中向智能体展示的推荐菜单，不仅决定了我们预期获得的即时奖励（取决于智能体在该菜单上的当前选择分布），还会影响智能体对后续推荐的选择概率（通过偏好的更新），进而影响我们未来可获得的奖励范围。这种因果效应针对每个智能体单独发生，打破了经典方法所需的分布假设，例如

*agarpit@outlook.com
†w.brown@columbia.edu

协同过滤等方法。这凸显了在个体智能体层面为此类问题的动态建立严格理论基础的重要性。为了评估长期交互，自然可以考虑遗憾最小化框架；该框架还能够处理逐项奖励随时间发生的对抗性变化。尽管近期一些研究已将遗憾最小化视角用于多智能体推荐问题（Gaitonde 等人，2021；Dean 和 Morgenstern，2022；Jagadeesan 等人，2022），这些研究对推荐场景和偏好演化模型提出了很强的假设：每轮只提供一条推荐，并由此引发偏好的线性更新。与之相反，我们考虑的场景允许偏好动态未知且可能高度非线性，能够表达物品之间常见的复杂交互（例如替代效应和互补效应、相关的序列排序或类型相关性），并能满足许多系统面临的实际约束，即必须向智能体展示包含多个物品的菜单供其选择。即使只有一个智能体，这一问题也并非易解。

尽管 Agarwal 和 Brown (2022)的研究设置通过允许使用多物品推荐菜单和非线性偏好动态，涵盖了其中许多考量，但仍存在若干重大局限。该研究假设，智能体对各物品的相对偏好是过去物品选择经验分布的函数，且这些函数由一个已知参数类中的函数决定，该参数类满足一个相当强的可学习性条件（例如常次数多项式）；其算法还需要显式学习每个函数。此外，用于遗憾最小化的目标集合相当受限，只包含无论智能体当前偏好如何都始终可行的物品分布（前提是选择了恰当的菜单分布）。我们证明，这两项限制几乎都可以消除，同时还能灵活设定智能体偏好的演化速度；所需调整很小，且从偏好建模的角度看有充分依据。

我们的第一项改动是允许智能体具有 非均匀 记忆，即每轮都对过去选择构成的向量施加折扣因子 $\gamma \leq 1$ 。我们首先假设 γ 随着 T 增大而趋近于 1，从而得到一个有效记忆时域（即其影响在记忆中占主导地位的轮次窗口），其长度为 $\Theta(T^c)$，其中 $c \in (0,1]$。对于这类情形，我们证明，只需假设每个函数都是 Lipschitz 连续的，便可取代参数可学习性假设；我们还给出一种算法，相对于 Agarwal 和 Brown (2022)考虑的同一目标集合实现次线性遗憾，并在均匀记忆情形下（当 $T^{3/4}$ 时）恢复其 $\gamma = 1$速率。接下来，我们为智能体的偏好函数确定了一种性质，它刻画了从线性递增函数到任意函数的插值情形，并以一个乘法因子界定各点偏离线性的程度；我们称这类函数为“尺度有界”。对于偏好具有此类性质且记忆时域严格次线性的智能体（即 严格 次线性的记忆时域（即满足 $c \in (0,1)$），我们给出一种算法，使其相对于几乎整个 物品 单纯形实现次线性遗憾。此处，我们的目标集合与单纯形边界之间的距离取决于智能体的“探索项” $\lambda > 0$；在这种情形下，该项可以任意小。这些条件可以解释为要求：(i) 智能体具有近因偏置；(ii) 偏好通常会随熟悉度增加而增强。这两点在智能体偏好建模中都被广泛采用（综述见 Curmei 等人 (2022)）。我们还给出了若干关于高效菜单选择和常数记忆代理的其他结果，并讨论了进一步放宽条件所导致的否定性结果。

1.1 设定

与 Agarwal 和 Brown (2022) 中一样，在每一轮 $t \in [T]$，我们（推荐系统）必须从 k 个 n物品中选择若干项，放入菜单 K_t 供智能体选择；智能体将以概率方式从 k 个物品中选择一项。每个物品的选择概率取决于智能体的 记忆向量 $v_t \in \Delta(n)$，该向量编码了其此前物品选择的历史记录，也取决于其 偏好模型；该模型通过函数将记忆向量映射为偏好 分数 $f_i(v_t) \in [0,1]$ ，每个物品对应一个函数；每个 $i \in K_t$ 的选择概率与这些分数成正比。对于智能体在每一轮 $r_t(i_t) \in [0,1]$ 中的物品选择，我们会收到对抗性老虎机奖励 i_t 。

请注意，对于 $k = 1$ ，这将退化为经典对抗性多臂老虎机问题，因为智能体会确定性地选择唯一展示的物品。在这里，我们将 $k > 1$ 视为问题的固定输入；由于偏好模型所发挥的作用，这会产生比单物品推荐复杂得多的选择动态：我们正在进行一个无法直接拉取臂（物品）的老虎机问题，而且我们能够诱导出的物品瞬时可实现分布（考虑每轮可以选择的菜单上的任意分布）会根据历史记录而变化，具体取决于未知的非线性动力学。

因此，一般而言，在奖励方面，我们不可能做到和最佳单个物品一样好。即使分数始终均匀，智能体也绝不会在每 k 轮中选择同一个物品超过一次，即便该物品出现在每个菜单中；因此，在选择基准目标集（作为项目单纯形的子集）时必须谨慎，以便我们能够在次线性遗憾意义下切实与之竞争。

1.2 我们的结果

我们为 Agarwal 和 Brown(2022) 提出的针对折扣记忆智能体的在线推荐问题，给出了一系列算法结果和障碍结果，其中记忆向量按如下方式更新： $v_{t+1} \propto i_t + \gamma v_t/(1-\gamma)$，其中$\gamma \in [0,1]$。我们在第2节形式化定义了推荐系统与智能体交互模型，并介绍了关键的预备知识。

在第 3 节中，我们给出一种算法，对于任意 $c \in (0,1]$，该算法针对有效记忆时域为 $\Omega(T^c)$ 的智能体，对“EIRD 集”（由 Agarwal 和 Brown（2022）提出）取得 $o(T)$ 遗憾。该目标集合由处处瞬时可实现的物品分布构成，即物品分布$p_t \in \Delta(n)$，满足对于任意记忆向量 $v_t \in \Delta(n)$ 和由此产生的偏好集 $\{f_i(v_t)\}$，都存在某个菜单分布 z_t，使其在期望意义下产生 $i_t \sim p_t$。此外，我们还针对任意 v_t，提出了瞬时可实现分布集合的一种新刻画，从而实现计算高效的菜单选择，并消除了先前工作中线性规划子程序对 k 的指数级依赖。

在第 4 节中，我们阐述了尺度有界偏好模型的定义，并考虑了一个称为 ϕ-平滑单纯形的目标集合，记作 $\Delta^\phi(n)$。该集合对应于单纯形 $\Delta(n)$ 中任意一点与ϕ 均匀噪声混合后得到的结果。对于记忆时域为 $\Theta(T^c)$ 的智能体，在任意 $c \in (0,1)$ 下，我们给出了一种相对于 $\Delta^\phi(n)$ 获得 $o(T)$ 遗憾的算法，其中 ϕ 是其他参数（尤其是智能体的探索项 $\lambda > 0$）的函数，在此可以任意小。这与关于 EIRD 的结果形成对比：为确保目标集合非空，λ 必须与 0 保持正距离。

在第 5 节中，我们证明：给定一个编码智能体偏好模型的电路，确定能够从自身即时实现的最佳分布的奖励，即使在固定奖励的情况下也是 NP-困难问题。我们认为，这充分表明，若要与任何非平凡地大于 EIRD 的目标集合竞争，就需要对偏好作出额外的结构性假设（例如我们的尺度有界性质）。

最后，在第 6 节中，我们考虑有效记忆时域为 $O(1)$ 的智能体。这里我们再次强调尺度有界性质的作用，并给出一种算法：当奖励变化不太频繁时，该算法相对于 $o(T)$ 可获得遗憾界 $\Delta^\phi(n)$；与此相反，对于一般偏好，我们证明了一个关于 EIRD 集合的遗憾下界，该下界关于 n呈准多项式增长，即使奖励固定也是如此（我们的算法遗憾界关于所有参数均为多项式）。

总而言之，除了为在线推荐中的高效遗憾最小化提供若干新算法外，我们的结果还揭示了目标集合的选择、智能体记忆更新速度以及偏好函数的结构性假设之间若干微妙的权衡，这些发现或许也能为更广泛的智能体偏好建模提供参考。

1.3 相关工作

奖励变化的随机多臂老虎机 未来奖励会受到动作因果影响的问题，已在随机多臂老虎机设定中得到研究（Gittins，1979；Heidari 等，2016；Levine 等，2017；Kleinberg 和 Immorlica，2018；Shah 等，2018；Leqi 等，2021；Laforgue 等，2022；Awasthi 等，2022；Papadigenopoulos 等，2022）。关于这类问题的最新研究大多聚焦于奖励演化的特定模型，其动机包括智能体厌倦、智能体无聊和拥塞，并且假设每轮只选择一个动作。相比之下，我们的设定允许奖励具有对抗性，而偏好演化则由我们在每个菜单中选择的多个物品之间的相互作用决定。

偏好动态模型 此外，人们还开展了大量研究，探讨推荐系统中的偏好动态，重点关注线性更新模型。 Hazla 等人 (2019);Gaitonde 等人 (2021)提出了一种政治偏好动态模型，其中向量偏好会朝着随机抽取议题上的一致或分歧方向漂移

，并且 Dean 和 Morgenstern (2022)在单一物品的个性化推荐情境下研究了类似模型。 Jagadeesan 等人 (2022)也研究了相关模型，以探讨推荐对类型形成的影响。此外，Zhou 等人 (2021)在多臂老虎机设定下研究了支付激励与自我强化偏好之间的相互作用。

强化学习与推荐系统 已有研究利用强化学习进行推荐，以最大化长期奖励（Ie 等人，2019；Zhan 等人，2021；Chen 等人，2019），通常侧重于实证评估。此外，还有大量研究从总体层面探讨推荐系统中的偏差、反馈循环和“回音室”效应（Flaxman 等人，2016；Curmei 等人，2022；Abdollahpouri 等人，2019；Mansoury 等人，2020）。

对决老虎机 “对决老虎机”框架研究了一种与我们的问题类似的推荐问题：每次试验中会选择多个物品，并观察到代表智能体选择的相对反馈（Yue 和 Joachims，2009；Yue 等人，2012；Agarwal 等人，2020；Rangi 等人，2021）。然而，与我们的设定不同，这些研究所考虑的偏好模型是先验固定的先验地，不会随物品历史记录而变化。

2 预备知识

全文中，我们用 $\Delta(n)$ 表示 n 个物品上的单纯形，用 u_n 表示 n 个物品上的均匀分布，并用 $d_{TV}(v, v')$ 表示两个分布之间的全变差距离。除非另有说明（例如 $\|x\|_1$），我们使用 ℓ_2 范数；并用 $B_\epsilon(x, \epsilon)$ 表示以 x 为中心、半径为 ϵ 的球。

2.1 交互模型

我们回顾 Agarwal 和 Brown（2022）针对具有自适应偏好的智能体的在线推荐问题所采用的设定。任意时刻，都存在一个记忆向量 $v \in \Delta(n)$，它表示智能体先前选择的某种函数。智能体的偏好模型是一个映射 $M\colon \Delta(n) \to [0,1]^n$，它根据偏好函数 $f_i\colon \Delta(n) \to [0,1]$，为每件物品指定得分 $M(v)_i = f_i$。问题实例由物品全集、菜单大小 $k < n$、偏好模型 M、记忆更新规则 U以及奖励向量序列 $r_1, \ldots, r_T$ 指定。在每一轮 $t \in [T]$:

- 推荐系统选择一个菜单 K_t，其中包含来自 $[n]$ 的 k 个不同物品，并将其展示给智能体；
- 智能体根据以下分布给出的概率，随机选择一个物品 $i_t \in K_t$:

$$p_t(i; K_t, v_t) = \frac{f_i(v_t)}{\sum_{j \in K_t} f_j(v_t)};$$

- 根据更新规则，将记忆向量更新为 $v_{t+1} = U(v_t, i_t, t)$;
- 推荐系统因所选物品获得奖励 $r_t(i_t) \in [0,1]$。

初始的 $v_1 \in \Delta(n)$ 可以任意选择。我们假设推荐系统不知道每个 f_i，但已知 U。推荐系统的目标是在 T 轮次内，针对某个目标集合$S \subseteq \Delta(n)$最小化遗憾。对于任何此类 S，算法 $\mathcal{A}$ 相对于 S 的遗憾为

$$\mathrm{Reg}_S(\mathcal{A}; T) = \mathbb{E}\left[\max_{x \in S} \sum_{t=1}^{T} \langle r_t, x\rangle - r_t(i_t)\right]$$

其中，i_t 是智能体在时间 t 基于 $\mathcal{A}$所做出的物品选择，期望则是对 $\mathcal{A}$ 的内部随机性以及智能体的选择取的。

2.2 可实现分布

对于偏好模型 M和记忆向量 v，令 IRD(v, M)表示给定 v时瞬时可实现分布的集合，其定义如下：

$$\mathrm{IRD}(v, M) = \mathrm{convhull}\left\{p(K, v) : K \in \left[\binom{n}{k}\right]\right\}$$

其中， K 是 $[n]$ 的一个含 k个物品的子集， p(K, v)表示具有记忆v 的智能体在看到菜单 K时的物品选择分布，其定义为

$$p(i; K, v) = \frac{f_i(v)}{\sum_{j \in K} f_j(v)}$$

对于每个物品 i ，其取值为 K （其他情况下为 0），其中 f_i 是任意物品 i的偏好评分函数。请注意，这表示了具有记忆 v 的智能体可能产生的所有物品选择分布，这些分布源自推荐系统的所有可能菜单选择策略，因为菜单上的任意分布都会产生物品分布 p(K, v)的凸组合。处处瞬时可实现分布的集合定义为

$$\mathrm{EIRD}(M) = \bigcap_{v \in \Delta(n)} \mathrm{IRD}(v, M).$$

这是 Agarwal 和 Brown（2022）考虑的目标集合，也是我们将在若干结果中考虑的集合。作为一个玩具示例，考虑偏好对所有 v均固定为 f_i(v) $= 1$ 的情形；此时，每个 IRD 集合都等价于 EIRD，后者是 Δ(n)中所有向量的凸包，其中恰好有 k 个物品的权重为$1/k$ 。我们指出，考虑“物品空间”而非“菜单空间”中的遗憾基准有几个自然的理由：除了先前已证明的、针对最佳固定菜单分布的线性遗憾下界之外，我们的奖励取决于智能体选择的物品，而非我们的推荐；即使固定菜单分布，选择分布也不一定会迅速稳定。我们假设评分函数 f_i 实际上有界于范围 $[\lambda, 1]$ 内，其中某个常数 $\lambda > 0$ 体现了智能体自身的探索；在第 3.3 节中，我们假设 $\lambda \geq k/n$，这确保 EIRD 非空（且包含均匀分布）；而在第 4.1 节中，我们允许λ 任意小。

2.3 折扣记忆代理

在全文中，我们考虑记忆更新规则为 γ-折扣的智能体。

定义 1（折扣记忆更新）。在 γ-折扣记忆更新规则 U_γ下，对于某个$\gamma \in [0, 1]$，当物品 i_t 在第 t轮被选中时，记忆向量 v_t 更新为 $v^{t+1} = U_\gamma$(v_t, i_t, t)，其中

$$v_{t+1}(i) = \frac{\sum_{s=1}^{t} \gamma^{t-s} \cdot \mathbf{1}(i = i_s)}{\sum_{s=1}^{t} \gamma^{t-s}}.$$

取 $\gamma = 1$ 可得到 Agarwal 和 Brown (2022所考虑的均匀记忆更新规则）。与许多涉及折扣因子的情形一样，我们可以将 γ 较接近 1 的取值视为对应着较大的记忆“有效视野”；对于任意 $\gamma \leq 1 - o$(1)，我们将量 $1/(1-\gamma)$ 简称为 有效记忆时域。我们始终假设 γ 是已知的。

2.4 平滑偏好模型

我们的许多结果都考虑满足以下性质的偏好模型：每个评分函数都是 Lipschitz 连续的，且有界于 0 以上；我们将这类模型称为平滑偏好模型。

定义 2（平滑偏好模型）。偏好模型 M 是（λ， L）-平滑的，如果每个评分函数 f_i 的取值都在 $[\lambda, 1]$ 中，且相对于 ℓ_1 范数，在 Δ（n）上是 L-Lipschitz 的。

这一性质允许函数类相当广泛，而且Agarwal 和 Brown (2022中的每一类函数（例如低次多项式）在参数取值适当时都满足这一性质。尽管这一假设相当宽泛，我们在第3 节中证明，仅凭这一假设就足以使我们始终能够保持对模型的准确局部近似，前提是智能体的记忆向量变化不太快；为此，我们定期执行查询学习流程。为方便起见，我们假设偏好分数始终经过归一化，使其总和为常数 $\sum_i f_i(v) = C$ 对于某个 C 以及任意 v，不过，对于我们的每项结果，这一假设都可以放宽至多相差 poly(n, L) 倍。

3 针对长记忆时域智能体的 EIRD

我们首先考虑以下情形： $\gamma = 1 - o(1)$，有效记忆时域为 $\Omega(T^c)$，其中某个$c \in (0, 1]$。此时，记忆向量变化缓慢，且 $\Delta(n$中的每个点都可以由某个物品选择序列很好地近似。在第3.1 节中，我们给出关于 IRD 集合结构的结果，从而能够在我们的 $o(T)$-遗憾算法中高效地进行菜单选择，该算法用于 EIRD ，详见3.3节。

3.1 通过菜单时间刻画 IRD

我们引入菜单时间这一概念，表示每个物品为诱导特定物品分布所需的时间x；借此，我们可以直接刻画 IRD 集合，并通过贪心方法在每一轮高效地构造菜单分布，避免 Agarwal 和 Brown（2022）的线性规划程序对 k 的指数依赖。该程序会枚举所有$\binom{n}{k}$ 菜单。在记忆向量 v下，对于目标物品分布 x，物品 i 的菜单时间定义为

$$\mu_i = \frac{k \cdot \frac{x_i}{f_i(v)}}{\sum_{j=1}^{n} \frac{x_j}{f_j(v)}}.$$

注意到，对于任意 v 和 x，这些量总是满足 $\sum_i \mu_i = k$。我们证明，当且仅当 $\max_i \mu_i \le 1$ 时，分布 x 可以由记忆向量 v 实现。

引理 1. 物品分布 x 属于 IRD(v, M)，当且仅当每个物品的菜单时间 μ_i 至多为 1。若此条件成立，则存在一个运行时间为 poly(n) 的算法 MenuDist(v, x, M)，用于构造菜单分布 z，使得 $\mathbb{E}_{K\sim z}[p(K, v)] = x$。

证明概要. 若 $x \in$ IRD(v, M)，则存在某个菜单分布 z，可产生 x；将此菜单分布转换为 μ_i 值时，按菜单的概率质量及其物品得分总和的倒数对菜单进行“记账”，即可得到满足 $\sum_i \mu_i = k$ 且 $\max_i \mu_i \le 1$ 的菜单时间向量。

给定满足这些条件的菜单时间向量，我们可以通过贪心地选择剩余菜单时间最高的 k 个物品组成菜单，并以相同速率“计费”其剩余菜单时间来构造这样的分布；对于排名第 k 高的并列物品，则按分数比例计费并以分数比例纳入菜单。每个阶段中，剩余菜单时间并列排名第 k 高的物品数量都会增加 1，而初始菜单时间最高的 $k-1$ 个物品（其中 $\mu_i \le 1$）会以非分数比例纳入，直到并列排名第 k 高。最终分布 z 中新增的每个菜单的概率质量，将按菜单内物品的得分总和成比例分配。这样，菜单得分总和中的各项便可相互抵消，从而得到一个菜单分布，其中物品的选择概率与其得分 $f_i(v)$ 以及该物品被纳入菜单的（分数）阶段数成比例。由于物品被纳入菜单的阶段数与其菜单时间成比例，且其菜单时间与 $x_i/f_i(v)$ 成比例，因此诱导出的物品选择分布便与 x_i 成比例。

□

`MenuDist`$(v,\ x,\ M)$ 直接实现了这一菜单分布构造，并被第 3.3 节和第 4.1 节中的算法采用。更多细节见附录 B.1。

3.2 一种实用的对抗性多臂老虎机算法

在此，我们提出一种用于对抗性多臂老虎机问题的新算法，它具有多项实用的鲁棒性性质，是我们在第3.3节所采用方法的核心。我们的算法称为 Deferred Bandit

Gradient，能够容忍每轮中动作分布 ξ_t 受到未观测到的对抗性扰动 x_t （其中 x_t 在抽样前被扰动为 $y_t = x_t + \xi_t$ ），并且能够适应收缩的决策集（每轮选取的动作分布 x_t 必须属于集合 $\mathcal{K}_t$，其中 $\mathcal{K}_t \subseteq \mathcal{K}_{t-1}$）。Agarwal 和 Brown （2022）指出，这两项性质在此设定中很有用，因为偏好评分估计不可避免地存在一定误差，而且我们通常事先并不知道EIRD 集合的形状。对于一般的在线优化而言，采用类似跟随正则化领导者的方法（例如 Hedge 或 EXP3），似乎很难获得收缩域性质；而采用类似（投影）在线梯度下降的方法则要容易得多。与 Agarwal 和 Brown （2022）不同，他们扩展了用于老虎机凸优化的“OGD 风格” FKM 算法，该算法可实现 $O(T^{3/4})$ 遗憾（Flaxman 等人（2004）），我们的算法直接应用于单纯形上线性损失的对抗性老虎机设定，并利用线性性质降低梯度估计的方差（这是 FKM 遗憾的瓶颈）：通过将每次奖励观测的贡献“延后”到之后的若干轮次中，从而提升运行效率。

算法 1：延迟老虎机梯度

输入：奖励序列 r_t、扰动向量 ξ_1、…、ξ_T，其中每轮 t 有 $|\xi_t$、$_i| \le \frac{\epsilon}{n}$，以及收缩凸决策集 $\mathcal{K}_1$、…$\mathcal{K}_T$，其中对于给定的 ϵ， $\mathcal{B}_\epsilon \subseteq \mathcal{K}_T$。令 $x_1 =$u$_n$、 $H = \frac{n}{\epsilon}$，对 $t = 1$，令其取值为 T。对手扰动分布 x_t，使其变为 $y_t = x_t + \xi_t$。采样 $i_t \sim y_t$，观察 i_t 和奖励 $r_t(i_t)$。令 $\tilde{r}_t = \frac{e_{i_t}}{H} \cdot r_t$、$\frac{i_t}{x_t}$、$_{i_t}$ 和 $\tilde{\nabla}_t = \sum_{s=}^{t}max(_{t-H+1}$,1) $\frac{\tilde{r}_t}{H}$。令 $\mathcal{K}_{t+1}$、$_\epsilon = \{x|$u$_n + \frac{1}{1-\epsilon}(x -$ u$_n) \in \mathcal{K}_{t+1}\}^{\sim}$。更新 $x_{t+1} = \Pi_{\mathcal{K}_{t+1}}$、$_\epsilon[x_t + \eta\nabla_t]$。结束循环。

定理 1. 对于奖励序列 $r_t, \ldots, r_T \in [0,1]^n$、收缩的凸决策集 $\mathcal{K}_1, \ldots \mathcal{K}_T \subseteq \Delta(n)$（其中 u$_n \in \mathcal{K}_T$），以及满足 $|\xi_t$、$_i| \le \frac{\epsilon}{n}$的扰动向量 $\xi_1, \ldots, \xi_T$，在每一轮 t给定 ϵ的情况下，Deferred Bandit Gradient 获得的期望遗憾上界为

$$\max_{x^* \in \mathcal{K}_t} \sum_{t=1}^{T} r_t^\top x^* - \sum_{t=1}^{T} r_t^\top y_t \le 2\eta n^2 T + \frac{\sqrt{2}}{\eta} + 3\sqrt{n}\epsilon T + \frac{n}{2\epsilon} + \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{|\xi_{t,i}|}{x_{t,i}}.$$

我们在附录 A中证明定理 1 。我们的分析通过跟踪在线梯度下降的一种变体 $^{\sim}$的遗憾来展开，该变体能够处理序列 ∇_t 向量期望值上的收缩；这些向量的平方范数在期望意义下较小。我们还证明，这一变体的遗憾与我们的算法所得遗憾以及事后最优解 x^* 的奖励都非常接近。注意，若每一轮都满足约束$|\xi_{t,i}| \le \frac{\epsilon x_t}{_n}$ ^{i}，则可以对 DBG 进行校准，使其获得遗憾 $O(n\sqrt{T} + \epsilon\sqrt{n}T)$。

3.3 针对 EIRD

概括而言，我们的方法是引导智能体代表我们，在 $\Delta(n)$ 的一个不断收缩的子集上隐式运行 Deferred Bandit Gradient，该子集始终包含 EIRD。我们会定期暂停，以更新对智能体当前偏好的估计；在此过程中，所有物品都会被足够频繁地展示给智能体，从而能够准确估计当前记忆向量附近的偏好分数。我们利用偏好的光滑性，确定随着 v_t 更新而变化的分数估计精度界限。给定一个供智能体采样的目标物品分布 x_t（由 DBG 选定），我们可以利用分数估计调用 MenuDist 例程，构造一个菜单分布 z_t，使其近似诱导出 x_t 的选择分布（其误差由 ξ_t 对于 DBG 所表示的扰动刻画）。

由于我们的方法依赖于记忆向量在各轮次之间的稳定性，当记忆时域相对于 $\Theta(T$趋近于零时，我们的遗憾会逐渐降至 T；关于短记忆时域所带来的挑战，我们将在第6节进一步讨论。对于任意 T^c 的记忆时域，只要 $c > 0$，我们就能获得严格次线性的遗憾，并且

对于均匀记忆，我们恢复了 $\tilde{O}(T^{3/4}$的速率，该结果来自Agarwal 和 Brown (2022)；现在，这一结果适用于任意平滑偏好，而不再局限于特定的参数类。

定理 2. 对于具有 $(\lambda,\ L)$-光滑偏好模型 M 且 $\lambda \geq k/n$，并具有 γ-折扣记忆 $\gamma \geq 1 - \frac{1}{T^c}$和 $c \in (0,1]$, 算法2的智能体，其遗憾有界于

$$\max_{x^* \in \mathrm{EIRD}(M)} \sum_{t=1}^{T} r_t^\top x^* - \mathbb{E}\left[\sum_{t=1}^{T} r_t(i_t)\right] = \tilde{O}\left((n/\lambda)^{3/2} L^{1/4} \cdot T^{1-c/4}\right).$$

算法 2（针对 EIRD 的平滑模型）。

令 $c^* = \min(c, 3/4)$, $\epsilon = \tilde{O}(nL^{1/4}\lambda^{-3/2}T^{-c/4})$, $Q = \tilde{O}(\frac{n^2}{\lambda^4\epsilon^2})$, and $\eta = (nT)^{-1/2}$。初始化 $q = 0$, $v_0 = v^* = \mathbf{u}_n$, $F_i = \frac{C}{n}$，对于 $i \in [n]$, $M^* = \{F_i\}$ 初始化 DBG，对于 ϵ, η。当 $t \leq T$ 时，若 $t < T^{c^*}$，则向智能体展示任意菜单 K_t；否则，若 $t \geq T^{c^*}$ 且满足 $\|v_t - v^*\|_1 \geq \frac{\lambda\epsilon}{2nL}$ 或 $q = 0$，则对于 $b \in \{0, \ldots, \lceil\frac{n-1}{k-1}\rceil\}$，向智能体展示代理菜单 $K_b = \{1\} \cup \{b(k-1)+2, \ldots, (b+1)(k+1)+1\}$，持续 Q 轮次。令 $F_i = (\#\text{ 次 } i_t = i) / (\#\text{ 次 } i_t = 1)$，在这 Q 轮次内，对于 $i \in K_b$，结束循环。设置 $F_i = \frac{C \cdot F_i}{\sum_{j=1}^{n} F_j}$，对于每个 $i \in [n]$, $M^* = \{F_i\}$，设置 $v^* = v^t$，将 q 增加 $\lceil\frac{n-1}{k-1}\rceil \cdot Q$，并为 DBG 设置 $\mathcal{K}_{t-q+1} = \mathcal{K}_{t-q} \cap \mathrm{IRD}(v^*, M^*)$；否则，从 DBG获取 x_{t-q}。令 $z_t = \mathsf{MenuDist}(v_t, x_{t-q}, M^*)$，采样菜单 $K_t \sim z_t$。向智能体展示 K_t，使用观察到的 i_t 更新 DBG，并更新 $r_t(i_t)$。结束 if。~~对于~~由 q 计数的每一轮，必要时设置 $v_{t+1} = U(v_t, i_t, t)$。结束 while。

定理2 的完整证明推迟至附录2。基于定理 DBG 中的遗憾界1，核心挑战在于证明偏好估计 F_i 与各自的 $f_i(v_t$足够接近，从而能够通过`MenuDist`准确地进行选择定位；同时，还要确保用于更新的非DBG 轮次 F_i (由 q跟踪)不会增长过快。尽管 EIRD 至少包含均匀分布（由引理1 在 $\lambda \geq \frac{k}{n}$ 时可知），但通常它可能并不特别大。在第4节中，我们将确定适用条件，在这些条件下，替代算法方法使我们能够与比 EIRD大得多的物品分布集合竞争。

4 超越 EIRD：尺度有界偏好与平滑单纯形

Agarwal 和 Brown （2022）提出考虑 EIRD 的一个动机是：在均匀记忆下，探索很困难，因为当前记忆无法被反复“冲洗掉”而不导致指数级膨胀。然而，考虑 $\gamma \leq 1 - o$的折扣因子（1）带来了这样一种可能性：我们或许能够高效地探索可行向量空间，并与位于 EIRD之外的物品分布竞争；也就是说，这些物品分布仅对所有记忆向量中的一个真子集可行。我们发现了一种能实现这一点的结构性质：偏好评分函数的输出与其对应物品在记忆中的权重之间，乘法差异不能过大。我们将这类函数称为尺度有界。

定义 3（尺度有界函数）。若偏好评分函数 $f_i\colon \Delta(n) \to [\frac{\lambda}{\sigma}, 1]$ 对于 $\sigma \geq 1$ 和 $\lambda > 0$ 是（σ, λ）-尺度有界的，则

$$\sigma^{-1}((1-\lambda)v_i + \lambda) \leq f_i(v) \leq \sigma((1-\lambda)v_i + \lambda).$$

我们称偏好模型 M 是尺度有界的，如果每个 f_i 都是尺度有界的；在这种情况下，得分向量 $M(v$不会偏离其在 v中的取值太远。当这一性质成立时，我们可以证明（使用引理 1中的菜单时间方法），任何不太靠近 $\Delta(n$边界的点都包含在其自身的 IRD 集合中。这启发我们定义由所有此类点构成的目标集合，并称之为 ϕ-平滑单纯形。

定义 4（ϕ-平滑单纯形）。对于任意 $\phi \in [0,1]$，ϕ-平滑单纯形是由 $\Delta^{\phi}(n) = \{(1-\phi)x + \phi \mathbf{u}_n\colon x \in \Delta(n)\}$ 给出的集合。

此外，我们证明，任何此类点周围的一个邻域也包含在 $\Delta^{\phi}(n)$ 中；由于偏好分数不会与物品当前的记忆向量权重相差太远，因此，对于分布 x 中邻近的 v 任何物品，所需的菜单时间不会太长。

引理 2。设 M 是一个 (σ, λ)-尺度有界偏好模型，且满足 $\sigma \leq \sqrt{n/(2k)}$。那么，对于任意 $x \in B_{\lambda\phi}(v) \cap \Delta^{\phi}(n)$ 和任意 $v \in \Delta^{\phi}(n)$，只要 $\phi \geq 4k\lambda\sigma^2$，就有 $x \in \mathrm{IRD}(v, M)$。

这里，我们将 $\Delta^{\phi}(n)$ 视为遗憾最小化的目标集合。我们不再要求 λ 具有显式下界，因此可以通过适当选择 ϕ，使遗憾基准在 λ 趋近于 0 时逼近整个单纯形。这与 EIRD 基准形成鲜明对比：尺度有界这一性质表明，现在或许可以说服智能体几乎每一轮都选择最佳物品，而不是在 $O(T/n$轮中选择最佳物品（例如，在第 3.3 节中，如果某个 $f_i(v)$ 在每个 $v \in \Delta(n)$ 轮次都 $= \lambda = k/n$，就可能出现这种情况）。

4.1 针对 $\Delta^{\phi}(n)$ 的无遗憾算法

不同于算法 2（其中我们将每一轮视为老虎机优化算法在交错学习阶段中的一个步骤），在这里，我们将学习和目标设定的多次迭代合并为在线梯度下降中的 单个 步骤，该步骤在 Δ^{ϕ}（n）上运行，此时智能体的整个记忆向量都会沿期望方向移动。虽然我们无法再即时实现目标集合中的任意分布，但能够在邻近的球内诱导任意选择分布，从而通过智能体的记忆向量在 Δ^{ϕ}（n）中进行探索。此外，尺度有界条件将得分与记忆权重联系起来，从而降低了奖励和偏好估计中的方差。然而，这也会导致当 c 趋近于 1（以及趋近于 0）时出现衰减，因为记忆更新得不够快，无法支持探索。Agarwal 和 Brown 的 定理 1（2022）表明，这是必然的：当我们无法再显著移动整个记忆向量时，对手可能会在后续轮次改变奖励分布，从而导致线性遗憾。

定理 3 (尺度有界折扣遗憾界)。 对于任意具有偏好模型 M 且满足 (σ, λ)-尺度有界且 $(\frac{\lambda}{\sigma}, L)$-光滑，并具有 $\sigma \leq \sqrt{n/(2k)}$，且 γ-折扣记忆，适用于 $\gamma = 1 - \frac{1}{T^c}$，其中 $c \in (0,1)$，算法 3 实现的遗憾为

$$\mathrm{Reg}_{\Delta^{\phi}(n)}(\mathcal{A}_2; T) = \tilde{O}\left((n^4 L\left(T^{1-c/2} + T^{1/2+c/2}\right)\right)$$

相对于 ϕ-平滑单纯形，适用于 $\phi = 4k\lambda\sigma^2$。

证明概要。在“预热期”阶段，我们的目标只是将记忆向量推向 $\mathbf{u}_n$；首先仅在 k 个物品上饱和记忆，持续 T^c 轮次后，我们便能够以近乎均匀的速率将所有低记忆物品推向$\frac{1}{n}$，因为此时记忆变化缓慢，且根据尺度有界条件， k 个最低值将保持彼此接近˜，因此没有物品会“卡”在接近 0 的位置，我们可以在 $O(T^c)$ 轮次内到达 $\mathbf{u}_n$ 。在“初始学习”阶段，我们现在可以假定均匀分布属于我们的 IRD 集合，并通过在 $v_{t,i} < \frac{1}{n}$ 时采用悲观评分、否则采用乐观评分，迫使记忆保持在该处。通过将观察到的选择频率与我们假设的评分所指示的选择频率进行比较，我们得到

算法 3（面向尺度有界模型的 Δ^{ϕ} ）。

令 $\epsilon = \tilde{O}(n^4 L \cdot \max(T^{-c/2}, T^{c/2-1/2}))$，$S = \tilde{O}(n^{3/2} T^c)$，$\eta = \tilde{O}(n^{-1/2} L \cdot T^{c/2-1/2})$。

--- 预热期 ---

对于 $t = 1$ 到 T^c，执行

展示智能体菜单 $K = \{1, \ldots, k\}$

结束循环

当 $\max_i \left| v_{t,\,i} - \frac{1}{n} \right| \geq \frac{\epsilon}{4n^2 L \sigma}$ **do**

显示具有最小 v_t 的智能体 k 个物品，i，在最多为 T^{-c} 的并列中随机选择

结束当

--- 初始学习 ---

进行 T^c 轮次

令 $F_{t,\,i} = \sigma^{-1}((1-\lambda) v_{t,\,i} + \lambda)$，若 $v_{t,\,i} < \frac{1}{n}$，否则 $F_{t,\,i} = \sigma((1-\lambda) v_{t,\,i} + \lambda)$

令 $z_t = \text{MenuDist}(v_t, \mathbf{u}_n, \{F_{t,i}\})$，显示智能体 $K_t \sim z_t$

结束循环

设置 $F_i = \sum_t \mathbf{1}(i_t = i) \cdot F_{t,\,i} / (\frac{1}{n} T^c)) \cdot C(\sum_{j=1}^{n} F_{t,\,j})^{-1}$，设置 $v^* = v_t$

--- 优化 ---

初始化 OGD，在 $\Delta_{\epsilon}^{\phi}(n)$ 上运行 T/S 轮次，使用 η，设置 $x_1 = v^* := v_t$。

对于 $s = 1$ 到 T/S 执行

接收 $x_{\quad s}$ 来自 OGD

执行 S 轮次

如果 $\|v_t - v^*\|_1 \geq \epsilon/(2nL)$，则

令 $\tilde{v} = v_t$

对智能体 $K_t \sim z =$ 展示 $\text{MenuDist}(v_t, \tilde{v}, \{F_i\})$，持续 T^c/L^2 轮次

设置 $F_i = \sum_t \mathbf{1}(i_t = i) \cdot F_i / (\tilde{v}_i T^c / L^2) \cdot C(\sum_{j=1}^{n} F_j)^{-1}$，设置 $v^* = v_t$

结束 if

对智能体 $K_t \sim z_t =$ 展示 $\text{MenuDist}(v^*, x_s, \{F_i\})$

结束 for

设置 $\tilde{\nabla}_s = \sum_{h=t-S+1}^{t} e_{i_h} r_h(i_h) / (x_{h,i} S)$，更新 OGD

结束循环

在 $\mathfrak{u}_n$ 附近对真实的 f_i 值进行无偏估计量。在“优化”阶段，我们将 $O(T^c)$ 轮次批量处理为在线梯度下降的“步骤”，以便在整个过程中维持局部准确的 f_i 估计；同时交替推进至选定目标和更新得分，从而还能在每一步中对平均奖励向量进行集中估计，并得到类似于“减速”在线梯度下降的遗憾界。

□

我们允许 λ 任意小，并且只假设 T 足够大，从而得到 $\lambda \geq T^{-c/4}\mathrm{poly}(n)$；除此之外，我们的界不依赖于 λ 或 ϕ。在 c 上，我们再次达到最优速率 $\tilde{O}(T^{3/4})$，但这次该速率出现在 $c = \frac{1}{2}$ 时，此时学习中的方差缩减效果与在记忆空间中快速探索的需求达到了平衡。本节的完整证明见附录 C。

5 放宽基准的困难性

在此，我们给出一个困难性结果，涉及确定给定损失函数下偏好模型最优奖励的复杂性。我们认为，这表明若要将目标集合扩展到 EIRD 之外，就必须对偏好模型作出精心设计的假设，例如我们在第 4 节中考虑的尺度有界性质。

定理 4。除非 RP $\supseteq$ NP，否则不存在多项式时间算法，能够以偏好模型 M 和线性奖励函数 r 的电路表示作为输入，并在 $1/n$ 因子内近似其自身 IRD 集合中最佳分布 $v \in \mathrm{IRD}(v, M)$ 的奖励 $r(v)$。

我们证明，可以将“最大独立集”问题的一个实例编码为智能体的偏好模型；该问题的近似求解是 NP-困难的，而优化奖励等同于选取任意一个独立集。我们假设只有 1 个物品能获得正奖励，并且它始终位于菜单中；目标函数则是最大化其得分。我们的构造将赋予物品权重的方案解释为可能的独立集提案，然后高效地检查图中对应顶点之间是否存在边；于是，物品 1 的得分与任一有效独立集的大小成正比。

这表明，超越 EIRD 的优化难题，不仅源于可学习性或对抗性损失方面的问题，还源于偏好模型可能编码的最优策略所具有的复杂性。一般而言，若试图与一个 不属于其自身 IRD 集合的分布竞争，似乎毫无希望；而对于光滑偏好模型，只要某个点最初可以到达，那么经过足够长的时间，它就会在其 IRD 集合内保持 稳定 （允许任意近似）。因此，类似引理 2 的性质似乎是识别可行目标所必需的。

6 短记忆时域的智能体

当智能体的折扣因子足够小时，记忆向量可能快速移动，此时我们便失去了实现查询所需的精度，因此无法使用第 3 节中的算法。事实上，可行状态空间可能更像一个离散网格，其中记忆向量编码了在相对于 T 保持恒定的有效时域内所选取的物品序列。尽管如此，对于尺度有界模型，我们给出一种称为 EXP-ϕ 的算法，该算法相对于 $\Delta^{\phi}(n)$ 可获得 $o(T)$ 遗憾，适用于 任意 取值的 $\gamma \in [0, 1)$，前提是奖励的对抗性受到一定限制。具体而言，我们假设在长度为 $o(T)$ 的窗口内，奖励是随机的而非对抗性的，但分布可以在每个窗口之间对抗性地变化；我们要求 ϕ ($O(\lambda)$) 的下界略大一些（但仍然是 ϕ （仍然是 $O(\lambda)$ ））。

6.1 尺度有界模型的无遗憾算法

EXP-ϕ背后的思路是，将平滑单纯形的每个顶点视为多臂老虎机问题中的一个动作，其中每次“拉臂”对应若干轮次。当我们“承诺”在足够长的时间内选择菜单中的某个物品，同时在其他时候选择记忆中权重最小的物品时，尺度有界性质将确保该物品的选择频率逐渐趋近于平滑单纯形中的对应顶点。此外，由于我们不再试图显式学习偏好模型，因此可以放宽平滑性要求。

定理5. 对于任意具有偏好模型的智能体 M，该模型满足$(\sigma,\ \lambda)$-尺度有界性质，且每个 $f_i(v)\ \in[\lambda,1]$ 对于 $\lambda \geq \frac{\sigma^2 k}{n}$以及$\sigma \leq \sqrt{n/(2k)}$均成立；并且该智能体具有 γ-折扣记忆（ $\gamma \in [0,1)$ ，当损失来自一个每隔至多 $t_{hold} = \tilde{O}\left(\frac{T^{2/3}}{1-\gamma}\right)$ 轮次才变化一次的分布时，算法3获得的遗憾至多为

$$\mathrm{Reg}_{\Delta^{\phi}(n)}(\mathcal{A}_3; T) = \tilde{O}(T^{5/6})$$

对于 $\Delta^{\phi}(n)$，当 $\phi = 3\lambda k^3\sigma^6$ 时。

算法 4（EXP-ϕ）。

初始化 EXP3，运行 T/t_{hold} 步
当 $t < T$ 时执行
从 EXP3 中抽取臂 i^*
对于 $\tilde{O}(T^{2/3}/(1-\gamma))$ 轮次，执行
令 $K_t = \{i^*\} + \text{argmin}_{j\neq i}^{k-1} v_j$
结束循环
使用 i^* 的平均奖励更新 EXP3
结束循环

6.2 一般模型的障碍

如果我们不能假设偏好是尺度有界的，那么即使是针对任意光滑模型与门EIRD 进行竞争，似乎也很困难。我们通过构造偏好模型，证明了对于任何算法，在准多项式时间范围内相对于 EIRD 的遗憾下界。这些偏好模型的最优策略微妙地取决于当前记忆向量，同时会在离散状态空间中诱发快速探索。

定理 6. 对于任意 $\gamma \in (0, 1/2)$，存在一组 (λ, L)-光滑偏好模型 $\mathcal{M}$，其满足 $\lambda = O(1/n)$ 且 $L = \text{poly}(n)$，使得对于任意 $T \in O(n^{\log(n)})$，当偏好模型 M 从 $\mathcal{M}$中均匀采样时，任何算法的期望遗憾都至少为 $\Omega(T)$。

我们的方法是观察到，每个可行记忆向量都编码了一个长度为 $O(\log n)$ 的唯一截断历史记录，从而隐式地产生一个大小为 $O(n^{\log(n)})$ 的状态空间。我们设计的偏好模型中，最优策略可通过在每一轮诱导均匀分布来实现，该分布属于 EIRD，但要最大化物品 1 的选择概率，则需要识别特定的一组备选物品，并在每个状态下将其确定性地放入菜单中。我们证明，任何竞争策略都必然会以高概率探索许多状态，因此任何算法都会频繁到达这样的状态：在这些状态下，它无法识别最优菜单分布，而该分布是由一个随机过程按每个状态分别定义的。

参考文献

Himan Abdollahpouri、Robin Burke 和 Bamshad Mobasher。通过个性化重排序管理推荐系统中的流行度偏差，2019。

Arpit Agarwal 和 William Brown。为具有自适应偏好的智能体提供多样化推荐。载于 NeurIPS，2022。

Arpit Agarwal、Nicholas Johnson 和 Shivani Agarwal。选择多臂老虎机。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，神经信息处理系统进展，第 33 卷，第 18399–18410 页。Curran Associates, Inc.，2020。URLhttps://proceedings.neurips.cc/paper/2020/file/d5fcc35c94879a4afad61cacca56192c-Paper.pdf。

Sanjeev Arora 和 Boaz Barak。《计算复杂性：现代方法》。剑桥大学出版社，美国，第1版，2009年。ISBN 0521424267。

Pranjal Awasthi、Kush Bhatia、Sreenivas Gollapudi 和 Kostas Kollias。拥塞多臂老虎机：通过短期重置实现最优路由。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesv´ari、Gang Niu 和 Sivan Sabato 编，国际机器学习大会，ICML 2022，2022年7月17日至23日，美国马里兰州巴尔的摩，机器学习研究论文集第162卷，机器学习研究论文集，第1078–1100页。PMLR，2022年。Minmin Chen、Alex Beutel、Paul Covington、Sagar Jain、Francois Belletti 和 Ed H. Chi。REINFORCE 推荐系统的 Top-k 离策略校正。载于 J. Shane Culpepper、Alistair Moffat、Paul N. Bennett 和 Kristina Lerman 编，第十二届 ACM 网络搜索与数据挖掘国际会议论文集，WSDM 2019，澳大利亚维多利亚州墨尔本，2019年2月11日至15日，第456–464页。ACM，2019年。DOI：10.1145/3289600.3290999。URL https://doi.org/10.1145/3289600.3290999。Mihaela Curmei、Andreas A. Haupt、Benjamin Recht 和 Dylan Hadfield-Menell。迈向基于心理学的动态偏好模型。载于 Jennifer Golbeck、F. Maxwell Harper、Vanessa Murdock、Michael D. Ekstrand、Bracha Shapira、Justin Basilico、Keld T. Lundgaard 和 Even Oldridge 编，RecSys '22：第十六届 ACM 推荐系统会议，美国华盛顿州西雅图，2022年9月18日至23日，第35–48页。ACM，2022年。Sarah Dean 和 Jamie Morgenstern。个性化推荐下的偏好动态。载于 David M. Pennock、Ilya Segal 和 Sven Seuken 编，EC '22：第23届 ACM 经济学与计算会议，美国科罗拉多州博尔德，2022年7月11日至15日，第795–816页。ACM，2022年。Abraham Flaxman、Adam Tauman Kalai 和 H. Brendan McMahan。多臂老虎机设定下的在线凸优化：无梯度的梯度下降。*CoRR*，cs.LG/0408007，2004年。Seth Flaxman、Sharad Goel 和 Justin M. Rao。过滤气泡、回音室与在线新闻消费。《公众舆论季刊》，80(S1)：298–320，2016年3月。ISSN 0033-362X。DOI：10.1093/poq/nfw006。URL https://doi.org/10.1093/poq/nfw006。´Jason Gaitonde、Jon M. Kleinberg 和 Eva Tardos。几何意见动态中的极化。载于 P´eterBir´o、Shuchi Chawla 和 Federico Echenique 编，EC '21：第22届 ACM 经济学与计算会议，匈牙利布达佩斯，2021年7月18日至23日，第499–519页。ACM，2021年。John C Gittins。多臂老虎机过程与动态分配指数。《英国皇家统计学会会刊：B辑（方法论）》，41(2)：148–164，1979年。Jan Hazla、Yan Jin、Elchanan Mossel 和 Govind Ramnarayan。意见极化的几何模型。*CoRR*，abs/1910.05274，2019年。

Hoda Heidari、Michael J. Kearns 和 Aaron Roth。改进型与衰减型多臂老虎机的紧致策略遗憾界。在 Subbarao Kambhampati 编，《第二十五届国际人工智能联合会议论文集》，IJCAI 2016，美国纽约州纽约市，2016 年 7 月 9—15 日，第 1562–1570 页。IJCAI/AAAI出版社，2016 年。URL：http://www.ijcai.org/Abstract/16/224。

Eugene Ie、Vihan Jain、Jing Wang、Sanmit Narvekar、Ritesh Agarwal、Rui Wu、Heng-Tze Cheng、Tushar Chandra 和 Craig Boutilier。SlateQ：一种用于带推荐集合的强化学习的可处理分解方法。在 Sarit Kraus 编，《第二十八届国际人工智能联合会议论文集》，IJCAI 2019，中国澳门，2019 年 8 月 10—16 日，第 2592–2599 页。ijcai.org，2019 年。DOI：10.24963/ijcai.2019/360。URL：https://doi.org/10.24963/ijcai.2019/360。Meena Jagadeesan、Nikhil Garg 和 Jacob Steinhardt。推荐系统中的供给侧均衡，2022 年。

Robert Kleinberg 和 Nicole Immorlica。可充能多臂老虎机。载于 Mikkel Thorup（编），第59届 IEEE计算机科学基础年会，FOCS 2018，法国巴黎，2018年10月7日至9日，第309–319页。IEEE计算机学会，2018年。DOI：10.1109/FOCS.2018.00037。URLhttps://doi.org/10.1109/FOCS.2018.00037。

Pierre Laforgue、Giulia Clerici、Nicol`o Cesa-Bianchi 和 Ran Gilad-Bachrach。多臂老虎机中饱和效应与季节性的最后切换依赖分析。载于 Gustau Camps-Valls、Francisco J. R. Ruiz 和 Isabel Valera（编），《第25届国际人工智能与统计学会议论文集》，第151卷《机器学习研究论文集》，第971–990页。PMLR，2022年3月28日至30日。

Liu Leqi、Fatma Kilin̡c-Karzan、Zachary C. Lipton 和 Alan L. Montgomery。《用于饱和效应建模的反弹型多臂老虎机》。载于 Marc'Aurelio Ranzato、Alina Beygelzimer、Yann N. Dauphin、Percy Liang 和 Jennifer Wortman Vaughan 编，《神经信息处理系统进展第34卷：2021年神经信息处理系统年会》，NeurIPS 2021，2021年12月6日至14日，线上举行，第4003–4014页，2021年。

Nir Levine、Koby Crammer 和 Shie Mannor。《腐烂型多臂老虎机》。载于 Isabelle Guyon、Ulrike von Luxburg、Samy Bengio、Hanna M. Wallach、Rob Fergus、S. V. N. Vishwanathan 和 Roman Garnett 编，《神经信息处理系统进展第30卷：2017年神经信息处理系统年会》，2017年12月4日至9日，美国加利福尼亚州长滩，第3074–3083页，2017年。

Masoud Mansoury、Himan Abdollahpouri、Mykola Pechenizkiy、Bamshad Mobasher 和 Robin Burke。推荐系统中的反馈回路与偏差放大。*CoRR*，abs/2007.13019，2020。URLhttps://arxiv.org/abs/2007.13019。

Orestis Papadigenopoulos、Constantine Caramanis 和 Sanjay Shakkottai。充能收益下的非平稳多臂老虎机：以次线性遗憾实现改进规划。*CoRR*，abs/2205.14790，2022。DOI：10.48550/ arXiv.2205.14790。URLhttps://doi.org/10.48550/arXiv.2205.14790。

Anshuka Rangi、Massimo Franceschetti 和 Long Tran-Thanh。用于用户体验个性化的带反馈序贯选择多臂老虎机，2021。URL https://arxiv.org/abs/2101.01572。

Virag Shah、Jose H. Blanchet 和 Ramesh Johari。具有正外部性的多臂老虎机学习。*CoRR*，abs/1802.05693，2018。URL http://arxiv.org/abs/1802.05693。

Yisong Yue 和 Thorsten Joachims。将信息检索系统的交互式优化作为对决老虎机问题。载于《第26届国际机器学习年会论文集》，ICML '09，第1201–1208页，纽约，纽约州，美国，2009。计算机协会。ISBN 9781605585161。DOI：10.1145/1553374.1553527。

Yisong Yue、Josef Broder、Robert Kleinberg 和 Thorsten Joachims。k臂对决老虎机问题。《计算机与系统科学杂志》，78(5)：1538–1556，2012。ISSN 0022-0000。DOI：https://doi.org/10.1016/j.jcss.2011.12.028。URLhttps://www.sciencedirect.com/science/article/pii/S0022000012000281。JCSS特刊：2011年云计算。

Ruohan Zhan、Konstantina Christakopoulou、Ya Le、Jayden Ooi、Martin Mladenov、Alex Beutel、Craig Boutilier、Ed H. Chi 和 Minmin Chen。面向内容提供商感知推荐系统：关于用户和提供商效用相互作用的仿真研究。载于 Jure Leskovec、Marko Grobelnik、Marc Najork、Jie Tang 和 Leila Zia（编），WWW '21：2021年万维网大会，线上活动 / 斯洛文尼亚卢布尔雅那，2021年4月19日至23日，第 3872–3883 页。ACM / IW3C2，2021年。DOI：10.1145/3442381. 3449889。URLhttps://doi.org/10.1145/3442381.3449889。

Tianchen Zhou、Jia Liu、Chaosheng Dong 和 Jingyuan Deng。具有自我强化用户偏好的激励式老虎机学习。*CoRR*，abs/2105.08869，2021 年。URL https://arxiv.org/abs/2105.08869。

延迟 bandit 梯度分析

我们首先证明，标准在线梯度下降的分析可直接扩展到对抗性收缩域；在每一轮中，我们选择的动作 x_t 必须位于观测集合 $\mathcal{K}_t$ 中。

算法 5 收缩在线梯度下降。

输入：收缩凸决策集序列 $\mathcal{K}_1, \dots \mathcal{K}_T$, $x_1 \in \mathcal{K}_1$，步长 η

设置 $x_1 = \mathbf{0}$

对于 $t = 1$ 到 T，执行

选择 x_t 并观测代价 $\ell_t(x_t)$

更新并投影：

$$y_{t+1} = x_t - \eta \nabla \ell_t(x_t)$$
$$x_{t+1} = \Pi_{\mathcal{K}_{t+1}}(y_{t+1})$$

结束循环

引理 3. 对于一列收缩的凸决策集 $\mathcal{K}_1$，$\dots \mathcal{K}_T$，$x_1 \in \mathcal{K}_1$，每个集合的直径至多为 D，一列 G-Lipschitz 损失函数 ℓ_1、.....、ℓ_T，以及参数 η，算法 5 相对于 $\mathcal{K}_t$ 的遗憾有界为

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x^* \in \mathcal{K}_T} \sum_{t=1}^{T} \ell_t(x^*) \le \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla_t\|^2 \le GD\sqrt{T}$$

当 $\eta = \frac{D}{G\sqrt{T}}$ 时。

证明。令 $x^* = \arg\min_{x \in \mathcal{K}_T} \sum_{t=1}^{T} \ell_t(x)$，并令 $\nabla_t = \nabla \ell_t(x_t)$。首先注意到

$$\ell_t(x_t) - \ell_t(x^*) \le \nabla_t^\top (x_t - x^*)$$

根据凸性；随后，我们可以将每个点与 x^* 的距离上界为：

$$\|x_{t+1} - x^*\| = \left\|\Pi_{\mathcal{K}_{t+1}}(x_t - \eta \nabla \ell_t(x_t)) - x^*\right\| \le \|x_t - \eta \nabla_t - x^*\|,$$

如 $x^* \in \mathcal{K}_{t+1} \supseteq \mathcal{K}_T$。于是我们有

$$\|x_{t+1} - x^*\|^2 \le \|x_t - x^*\|^2 + \eta^2 \|\nabla_t\|^2 - 2\eta \nabla_t^\top (x_t - x^*)$$

and

$$\nabla_t^\top (x_t - x^*) \le \frac{\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2}{2\eta} + \frac{\eta \|\nabla_t\|^2}{2}.$$

由此我们可以得出：

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(x^*) &\le \sum_{t=1}^{T} \nabla_t^\top (x_t - x^*) \\
&\le \sum_{t=1}^{T} \frac{\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla_t\|^2 \\
&\le \frac{\|x_T - x^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla_t\|^2 \\
&\le \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla_t\|^2 \\
&= GD\sqrt{T}. \qquad \left(\eta = \tfrac{D}{G\sqrt{T}}\right)
\end{aligned}$$

□

证明。借助前面的结果，我们现在可以证明定理 1 的遗憾界。令 $r_t^* = \sum_{s=\max(t-H+1,1)}^{t} r_s \otimes (y_s \oslash x_s)$ H，其中 $\otimes$ 和 $\oslash$ 分别表示逐元素乘法和除法，并令 $\hat{r}_t = \sum_{s=\max(t-H+1,1)}^{t} \frac{r_s}{H}$。此外，令 $x_\epsilon^* = \Pi_{\mathcal{K}_T,\epsilon}[x^*]$。注意，对每个 t，以下关系均成立：

$$\frac{r_{t,i} \cdot y_{t,i}}{x_{t,i}} = r_{t,i}\left(1 + \frac{\xi_{t,i}}{x_{t,i}}\right);$$

$$\mathbb{E}[r_t^* - \hat{r}_t] = \frac{1}{H}\sum_{s=t-H+1}^{t}((r_s \otimes \xi_s) \oslash x_s); \tag{1}$$

$$\mathbb{E}\left[\widetilde{\nabla}_t\right] = r_t^*. \tag{2}$$

注意，由于每个 $\mathcal{K}_{t,\epsilon}$ 上的约束，每个 x_t 都可以表示为 $(1-\epsilon)x + \epsilon \mathfrak{u}_n$，其中 $x \in \mathcal{K}_t \subseteq \Delta(n)$，因此我们始终有 $x_{t,i} \geq \frac{\epsilon}{n}$。为了界定 ∇_t 的平方范数以应用引理 5，考虑最大化情形：除一个元素外，所有元素均满足 $x_t = \frac{\epsilon}{n}$ 和 $y_t = \frac{2\epsilon}{n}$，并且对于所有奖励都有 $r_{t,i} = 1$；当概率质量从臂 $x_{t,i}$ 转移到 $x_{t,j} > x_{t,i}$ 时，$\mathbb{E}\left[\left\|\widetilde{\nabla}_t\right\|^2\right]$ 会递增。因此，我们可以用一个二项随机变量的平方的期望来给出界 X，该随机变量有 $H = \frac{n}{\epsilon}$ 次试验，每次试验以 $\frac{2\epsilon}{(n-1)\,n}$ 的概率取值至多为 1（如果抽取了任一 $n-1$），否则取值为 $\frac{1}{H}$。由此可得：

$$\begin{aligned}\mathbb{E}\left[\left\|\widetilde{\nabla}_t\right\|^2\right] &\leq H\left(\frac{2\epsilon(n-1)}{n}\right)\left(1 - \frac{2\epsilon(n-1)}{n}\right) + \left(\frac{2(H-1)\epsilon(n-1)}{n} + 1\right)^2 \\ &\leq 2(n-1) + (2n-1)^2 \\ &\leq 4n^2. \end{aligned} \tag{3}$$

对于所有 T，对于任意固定的 $x^* \in \mathcal{K}_T$，我们有：

$$\sum_{t=1}^{T} r_t^\top x^* - \hat{r}_t^\top x^* \leq \frac{H}{2} = \frac{n}{2\epsilon}, \tag{4}$$

因为只有最近 H 轮次中的分数奖励未能在 $\sum_t \hat{r}_t$ 中得到适当计数。现在，我们分析算法相对于序列 $\{\hat{r}_t\}$ 的遗憾。对于 $x^* \in \mathcal{K}_T$，我们有：

$$\begin{aligned}\sum_{t=1}^{T}\hat{r}_t^\top x^* - \sum_{t=1}^{T}\mathbb{E}\left[\hat{r}_t^\top y_t\right] &\leq \sum_{t=1}^{T}\hat{r}_t^\top x_\epsilon^* - \sum_{t=1}^{T}\mathbb{E}\left[\hat{r}_t^\top x_t\right] + \sqrt{n}\epsilon T && \text{(each } r_t \text{ is } \tfrac{\sqrt{n}}{2}\text{-Lipschitz)} \\ &\leq \sum_{t=1}^{T} r_t^{*\top} x_\epsilon^* - \sum_{t=1}^{T} r_t^{*\top} x_t + \sqrt{n}\epsilon T + \sum_{t=1}^{T}\sum_{i=1}^{n}\frac{|\xi_{t,i}|}{x_{t,i}} && \text{(by (1))} \\ &\leq \mathbb{E}\left[\mathrm{Reg}_{COGD}(\widetilde{\nabla}_1, \ldots, \widetilde{\nabla}_T)\right] + \sqrt{n}\epsilon T + \sum_{t=1}^{T}\sum_{i=1}^{n}\frac{|\xi_{t,i}|}{x_{t,i}} && (5) \\ &\leq \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widetilde{\nabla}_t\right\|^2\right] + \frac{\sqrt{2}}{\eta} + \sqrt{n}\epsilon T + \sum_{t=1}^{T}\sum_{i=1}^{n}\frac{|\xi_{t,i}|}{x_{t,i}}. && \text{(by Lemma 3)}\end{aligned}$$

注意到，我们的算法等价于对序列 $\{\widetilde{\nabla}_t\}$ 运行收缩在线梯度下降，其中 $\mathbb{E}\left[\sum_{t=1}\widetilde{\nabla}_t\right] = \sum_{t=1} r_t^*$ 根据（2）。因此，由（3）和（4）可得

$$\sum_{t=1}^{T} r_t^\top x^* - \sum_{t=1}^{T}\mathbb{E}\left[r_t^\top y_t\right] \leq 2\eta n^2 T + \frac{\sqrt{2}}{\eta} + 3\sqrt{n}\epsilon T + \frac{n}{2\epsilon} + \sum_{t=1}^{T}\sum_{i=1}^{n}\frac{|\xi_{t,i}|}{x_{t,i}}.$$

□

B 第 3 节省略的证明

B.1 引理 1 的证明

证明。令每个物品的菜单时间 μ_i 为

$$\mu_i := \frac{k \cdot \frac{x_i}{f_i(v)}}{\sum_{j=1}^{n} \frac{x_j}{f_j(v)}}.$$

不难看出 $\sum_i \mu_i = k$。直观而言，菜单时间大致对应于某物品必须被纳入菜单的相对频率；其中，得分为 $\mu_i = 1$ 的物品必须始终纳入菜单；分配给某个菜单的"菜单时间"与该菜单中物品得分之和成反比，这使得对最终选择概率进行"苹果对苹果"的比较成为可能。

我们首先证明，任意 $x \in \mathrm{IRD}(v, M)$ 对每个物品产生的 μ_i 至多为 1。对于任意 $x \in \mathrm{IRD}(v, M)$，考虑由物品得分 $f_i(v)$ 给出的菜单条件物品分布的任意凸组合，其中每个菜单的概率为 p_K。按 $p_K / \sum_{i \in K} f_i(v)$ 的比例向每个菜单 K 分配 " 菜单时间单位 " μ_K，使得 $\sum_K \mu_K = k$；此外，对每个 $i \in K$，令 $\mu_{K,i} = \mu_K / k$。注意，物品 i 的选择概率为

$$\begin{aligned} x_i &= \sum_{K: i \in K} p_K \cdot \frac{f_i(v)}{\sum_{j \in K} f_j(v)} \\ &= \frac{1}{Z} \sum_{K: i \in K} \frac{\mu_K}{k} \cdot f_i(v) \\ &= \frac{f_i(v)}{Z} \sum_{K: i \in K} \mu_{K,i} \end{aligned}$$

其中，Z 是一个归一化常数，使得 $\sum_K \mu_K = k$，因此我们有 $\sum_K \mu_{K,i} \le 1$ 因为每个 μ_K 均为正数。此外，我们有

$$\begin{aligned} \sum_{K: i \in K} \mu_{K,i} &= Z \cdot \frac{x_i}{f_i(v)} \\ &= \frac{x_i}{f_i(v)} \cdot \frac{k}{\sum_{j=1}^{n} \frac{x_j}{f_j(v)}} \end{aligned}$$

求解满足 $\sum_K \mu_K = k$ 的 Z 后，可得

$$\sum_{K: i \in K} \mu_{K,i} = \mu_i$$

并得到每个物品的 $\mu_i \le 1$。

接下来，我们介绍用于构造菜单分布 z 的算法 MenuDist(v, x, M)，该算法可为任意满足 $\max_i \mu_i \le 1$ 的 x 和 v 生成 $\mathbb{E}_{K \sim z}[p(K, v)] = x$，从而以构造性方式证明 $x \in \mathrm{IRD}(v, M)$。给定满足 $\max_i \mu_i \le 1$ 的 x、v 和 M，我们通过 $H \le n$ 个阶段迭代地构造菜单分布。令 $\mu_i^1 = \mu_i$ 为每个物品的初始剩余菜单时间。在每个阶段 $h \ge 1$，我们将每个物品的剩余时间递减至 $\mu_i^{h+1} \le \mu_i^h$，并跟踪其变化 $\Delta_i^h = \mu_i^h - \mu_i^{h+1}$，其中 $\Delta^h = \sum_i \Delta_i^h$。此外，令 μ_j^h 为剩余值中第 k 高的值，令 $k_+^h < k$ 为满足 $\mu_i^h > \mu_j^h$ 的物品 i 的数量，令 $k_*^h \ge 1$ 为满足 $\mu_i^h = \mu_j^h$ 的物品 i 的数量。在每个阶段，我们维持以下不变量：

1. $k_*^{h+1} \ge k_*^h + 1$ 若 $k_*^h < n$；
2. $\Delta_i^h = \frac{\Delta^h}{k}$ 若 $\mu_i^h > \mu_j^h$（其中，μ_j^h 是第 k 高的值）；
3. $\Delta_i^h = \frac{(k - k_+^h)_{k_*^h \cdot k}^{\Delta^h}}{}$ 如果 $\mu_i^h = \mu_j^h$；

4. $\Delta_i^h = 0$ if $\mu_i^h < \mu_j^h$; and

5. $\Delta_i^h = \mu_i^h$ if $k_*^h = n$.

直观而言，对于每件物品，我们以相同的速率降低 μ_i，前提是其 μ_i^h 严格大于第 k高的值；这对应于在该阶段以概率 1 将其纳入菜单，并在剩余位置中以分数纳入的方式处理第 k高值的并列情况。我们持续降低，直到此类并列物品的数量至少增加 1（若所有物品都并列，则耗尽所有剩余菜单时间），然后继续下一阶段。注意，当 $\max_i \mu_i \le 1$时，这会使所有物品的 $\mu_i^{H+1} = 0$；如果某件物品的 $\mu_i^h > \mu_j^h$（仅在 $k_+^h > 0$时出现），其剩余菜单时间将以总降低速率的 $1/k$ 比例递减；由于我们只在剩余值最高的物品之间处理并列，因此不会有任何 μ_i^h 降至 0 以下，且耗尽的菜单时间总量至多为 k。一旦 $k_+^h = 0$，我们就只降低剩余值最高且并列的物品的 μ_i^h，从而在维持不变量的同时成功满足 $\mu_i^{H+1} = 0$。

现在说明如何根据数量 Δ_i^h 构造菜单分布 z，使其产生 $p(z, v) = \mathbb{E}_{K\sim z}[p(K, v)] = x$。对于一个阶段 h，我们构造一个菜单（多）集 $S^h = \{K_s^h : s \in [|S^h|]\}$，其大小为$|S_h| = k_*^h$；不失一般性，我们可以假设每个 $K_s^h \in S^h$ 都各不相同（例如，可以用某些辅助标记标出重复项）。令 $J^h = \{i : \mu_i^h = \mu_j^h\}$ 为在剩余物品中并列第 k 高的物品集合；我们通过 k_*^h 个步骤迭代构造 S^h：每一步加入一个菜单 K_s^h，其中包含所有满足 $\mu_i^h > \mu_j^h$ 的物品，以及来自 J^h 且截至目前属于 S^h 中菜单数量最少的 $k - k_+^h$ 个物品，并以任意方式打破并列。总计有 $(k - k_+^h)k_*^h$ 次将某个物品纳入 J^h 中某个菜单 S^h 的操作，因此每个物品 $i \in J^h$ 恰好被纳入 S^h 中的 $k - k_+^h$ 个菜单（因为 $(k - k_+^h)k_*^h$ 可被 $k - k_+^h$ 整除），而在 S^h 上的均匀分布随后满足

$$\Pr_{K_s^h \sim \mathrm{Unif}(S^h)} \left[i \in K_s^h\right] = \begin{cases} 1 & \mu_i^h > \mu_j^h \\ \frac{k-k_+^h}{k_*^h} & \mu_i^h = \mu_j^h \\ 0 & \mu_i^h < \mu_j^h \end{cases}.$$

设菜单分布 z^h 为 S^h 上的分布，其定义为

$$\Pr_{K\sim z_h} \left[K = K_s^h\right] = \frac{\sum_{i\in K} f_i(v)}{\sum_{s\in|S^h|} \sum_{j\in K_s^h} f_j(v)},$$

由此可得

$$\begin{aligned} \Pr_{K\sim z^h}[\text{Agent chooses i}] &= \sum_{K\in S^h : i\in K} \frac{f_i(v)}{\sum_{q\in K} f_q(v)} \cdot \frac{\sum_{q\in K} f_q(v)}{\sum_{s\in|S^h|} \sum_{j\in K_s^h} f_j(v)} \\ &= \frac{k\Delta_i^h}{\Delta^h} \cdot \frac{f_i(v)}{\sum_{s\in|S^h|} \sum_{j\in K_s^h} f_j(v)} \\ &\triangleq \frac{k\Delta_i^h}{\Delta^h} \cdot \frac{f_i(v)}{Z^h} \end{aligned}$$

作为向智能体展示从 z^h 中抽取的菜单时，其选择 i 的概率。定义$Z = \sum_{h=1}^H \frac{Z^h\Delta^h}{k}$，令 z 为分布 z^h 上的混合分布，每个分布的权重为 $\frac{Z^h\Delta^h}{Zk}$。抽取菜单$K \sim z$ 可得

$$\begin{aligned} \Pr_{K\sim z}[\text{Agent chooses i}] &= \sum_{h=1}^H \frac{Z^h\Delta^h}{Zk} \cdot \Pr_{K\sim z^h}[\text{Agent chooses i}] \\ &= \sum_{h=1}^H \frac{Z^h\Delta^h}{Zk} \cdot \frac{k\Delta_i^h}{\Delta^h} \cdot \frac{f_i(v)}{Z^h} \\ &= f_i(v) \cdot \frac{\mu_i}{Z} && (\textstyle\sum_h \Delta_i^h = \mu_i) \\ &= x_i, && (\mu_i \propto x_i/f_i(v)) \end{aligned}$$

其中 $Z=\sum_{j=1}^{n} k_{f_j\ v()}^{x_j}$ 由比例关系可知成立，因为 x 和 $p(z,v)$ 都是$[n]$上的概率分布。根据构造，可得 $x\in \mathrm{IRD}(v,M)$。

实现该构造的算法 MenuDist(v,x,M) 可在 poly(n) 时间内运行，因为数量 Δ_i^h 是在 $H\le n$ 轮次中计算的，每轮需要 $O(n)$ 的计算量（在按 μ_i 值降序进行初始排序之后）；每个集合 S^h 包含 $k_*^h\le n$ 个菜单，每个菜单都可在 $O(k)$ 时间内构造（通过以轮转方式从 J^h 添加元素），同时维护计算最终归一化常数所需的数量。采样也可以高效实现，例如从均匀分布中采样并适当设置阈值。

□

B.2 有效记忆时域

这里给出关于随着轮次推进，记忆变化界限的结果。

引理 4（界定记忆漂移）。对于任意 $\gamma\in\in(0,1)$、 $g\in\in(0,\gamma]$, $t\ge 1$，以及满足 $g^w\ge 1-2\beta$ 和 $g^{t+w-1}\le\frac{1}{2}$ 的 $w\ge 1$，我们有 $d_{TV}(v_t,v_{t+w})$ 至多为 β。

证明。我们可以将记忆向量 v_{t+w} 表示为

$$\begin{aligned}
v_{t+w} &= \frac{\sum_{s=t}^{t+w-1}\gamma^{t+w-s-1}\cdot i_s+\sum_{s=1}^{t-1}\gamma^{t+w-s-1}\cdot i_s}{\sum_{s=1}^{t+w-1}\gamma^{t+w-s-1}}\\
&= \frac{\sum_{s=t}^{t+w-1}\gamma^{t+w-s-1}\cdot i_s}{\sum_{s=1}^{t+w-1}\gamma^{t+w-s-1}}+v_t\cdot\left(1-\frac{\sum_{s=t}^{t+w-1}\gamma^{t+w-s-1}}{\sum_{s=1}^{t+w-1}\gamma^{t+w-s-1}}\right)
\end{aligned}$$

由此可得

$$\begin{aligned}
d_{TV}(v_t,v_{t+w}) &\le \frac{\sum_{s=0}^{w-1}\gamma^s}{\sum_{s=0}^{t+w-2}\gamma^s}\\
&= \frac{1-\gamma^w}{1-\gamma^{t+w-1}}\\
&\le \frac{1-g^w}{1-g^{t+w-1}}\\
&\le \beta.
\end{aligned}\tag{6}$$

步骤 6 由以下事实得出：

$$\begin{aligned}
\frac{\partial}{\partial\gamma}\left(\frac{1-\gamma^w}{1-\gamma^{t+w-1}}\right) &= \frac{-w\gamma^{w-1}(1-\gamma^{t+w-1})+(1-\gamma^w)(t+w-1)\gamma^{t+w-2}}{(1-\gamma^{t+w-1})^2}\\
&= \frac{\gamma^{w-1}\left((t+w-1)\gamma^{t-1}-(t-1)\gamma^{t+w-1}-w\right)}{(1-\gamma^{t+w-1})^2}\\
&\le 0,
\end{aligned}$$

因为函数 $(t+w-1)\gamma^{t-1}-(t-1)\gamma^{t+w-1}-w$ 在 $\gamma\in[0,1]$ 从 $-w$ 到 0 的区间内递增（因此，(6) 的导数在任意 $\gamma\in(0,1)$). □

我们可以利用这一点，得到关于 w 的上限，该上限用 c 表示，使此界成立 s.

引理 5。对于 $\gamma\ge 1-1/T^c$、 $t\ge T^c$ 和 $w\le\beta\cdot T^c$，我们有 $d_{TV}(v_t,v_{t+w})\le\beta$。

证明。令 $g = 1 - \frac{1}{T^c}$，其中我们有 $g^{(T^c)} \leq \frac{1}{e} \leq \frac{1}{2}$。此外，我们有：

$$\begin{aligned}
\log\left(\frac{1}{1-2\beta}\right) &\geq 2\beta \\
&\geq \frac{2w}{T^c} \\
&\geq w \log\left(\frac{1}{1-\frac{1}{T^c}}\right) && (\text{for } \tfrac{1}{T^c} \leq \tfrac{1}{2}) \\
&= \log\left(\frac{1}{g^w}\right).
\end{aligned}$$

由于 $\log(1/x)$ 关于 x 是递减的，因此有 $g^w \geq 1 - 2\beta$，由引理 4即可得到该结论。 □

B.3 算法 2分析：针对 EIRD

算法 2（针对 EIRD 的平滑模型）。

令 $c^* = \min(c, 3/4)$、$\epsilon = \tilde{O}(nL^{1/4}\lambda^{-3/2}T^{-c/4})$、$Q = \tilde{O}(\frac{n^2}{\lambda^4\epsilon^2})$ 和 $\eta = (nT)^{-1/2}$。
初始化 $q = 0$、$v_0 = v^* = \mathbf{u}_n$，其中 $F_i = \frac{C}{n}$ 为 $i \in [n]$、$M^* = \{F_i\}$
为 ϵ、η初始化 DBG 。
当 $t \leq T$ 时
  如果 $t < T^{c^*}$ **then**
    向智能体展示任意菜单 K_t
  否则，如果 $t \geq T^{c^*}$ 与任一 $\|v_t - v^*\|_1 \geq \frac{\lambda\epsilon}{2nL}$ 或 $q = 0$，则
    对于 $b \in \{0, \ldots, \lceil \frac{n-1}{k-1} \rceil\}$
      显示智能体菜单 $K_b = \{1\} \cup \{b(k-1) + 2, \ldots, (b+1)(k+1) + 1\}$，持续 Q 轮次
      Let $\hat{F}_i = (\#\text{ 次 } i_t = i)/(\#\text{ 次 } i_t = 1)$，在 Q 轮次内，针对 $i \in K_b$
    结束循环
    设置 $F_i = \frac{C \cdot \hat{F}_i}{\sum_{j=1}^n \hat{F}_j}$ 对每个 $i \in [n]$, $M^* = \{F_i\}$
    设置 $v^* = v^t$，将 q 增加 $\lceil \frac{n-1}{k-1} \rceil \cdot Q$，设置 $\mathcal{K}_{t-q+1} = \mathcal{K}_{t-q}$ 对 DBG 执行 $\cap \mathrm{IRD}(v^*, M^*)$
  else
    从 DBG 获取 x_{t-q}
    令 $z_t = \mathrm{MenuDist}(v_t, x_{t-q}, M^*)$，采样菜单 $K_t \sim z_t$
    向智能体展示 K_t，使用观测到的 i_t 更新 DBG 和 $r_t(i_t)$
  结束 if
  设置 $v_{t+1} = U(v_t, i_t, t)$，对于每个由 q 计数的轮次（如有必要）
结束循环

定理5。 对于具有 (λ, L)-光滑偏好模型 M（针对 $\lambda \geq k/n$）以及 γ-折扣记忆（针对 $\gamma \geq 1 - \frac{1}{T^c}$和 $c \in (0, 1]$, 算法 2所获得的遗憾有如下界

$$\max_{x^* \in \mathrm{EIRD}(M)} \sum_{t=1}^T r_t^\top x^* - \mathbb{E}\left[\sum_{t=1}^T r_t(i_t)\right] = \tilde{O}\left((n/\lambda)^{3/2} L^{1/4} \cdot T^{1-c/4}\right).$$

证明。在每一轮中，除去用于更新偏好估计的轮次（由 q 计数），我们会从 DBG 收到目标分布 x_t，并构造菜单分布 z_t，旨在代表智能体诱导出选择分布 x_t。回顾一下，我们假设 γ 已知，因此可以准确跟踪智能体在各轮中的记忆向量 v_t；我们注意到，通过在每一轮检查 $\|v_t - v^*\|_1$ 对 γ 所有可能取值的条件，该结果可以扩展到仅知道 γ 下界的情形。请注意，如果我们的偏好估计 F_i 始终能准确跟踪智能体的真实偏好f_i（v_t），并在每一轮对 IRD（v_t）进行精确表征，那么我们就会得到扰动 $\xi_t = \mathbf{0}$

对于每个目标分布 x_t，根据引理 1 的保证，以及一个决策集 $\mathcal{K}_T$，其中包含 EIRD。因此，我们的遗憾立即有以下界

$$\max_{x^* \in \mathrm{EIRD}(M)} \sum_{t=1}^{T} r_t^\top x^* - \mathbb{E}\left[\sum_{t=1}^{T} r_t(i_t)\right] \le \mathrm{Reg}_{\mathrm{EIRD}}(\mathrm{DBG}; T, \epsilon, \eta) + T^{1-c/4} + q = \tilde{O}(T^{1-c/4} + q)$$

由于 $T^{c^*} \le T^{1-c/4}$，对于任意 $c \in (0, 1]$，忽略非-T项）。这里请注意， EIRD 非空，且根据引理 [1], 包含 $\mathbf{u}_n$，因为 $\lambda \ge \frac{k}{n}$。接下来的分析将着重说明，偏好估计不精确所导致的扰动 ξ_t仍然很小，并且估计时间 q 不会增长得过大。

情况 1： $c^* = c$。对于任意 $c \le 3/4$，当我们结束展示任意菜单并开始在学习与优化之间交替时，智能体的记忆已经“饱和”。令 ϵ 满足

$$\epsilon \ge 8nL^{1/4}\lambda^{-3/2}\log(2T/\delta)^{1/4}T^{-c/4}$$

其中 $\delta = 1/T$。注意，以下结论成立：

$$\frac{\lambda^2\epsilon}{8nkL} \cdot T^c \ge \frac{64n^2(n-1)\log(2T/\delta)}{(k-1)\lambda^4\epsilon^2} \qquad (\epsilon^3 \ge \epsilon^4 \ge 512n^4L\lambda^{-6}\log(2T/\delta)T^{-c})$$

$$\frac{\lambda\epsilon^2}{4nL} \cdot T^c \ge \frac{64n^2(n-1)\log(2T/\delta)}{(k-1)\lambda^4\epsilon^2} \qquad (\epsilon^4 \ge 256n^4k^{-1}L\lambda^{-5}\log(2T/\delta) \cdot T^{-c})$$

根据引理 5，在每个$Q \cdot \lceil \frac{n-1}{k-1} \rceil$ 更新分数估计之前的 F_i中，我们有

$$\left| \frac{f_i(v^*)}{\sum_{j \in K_b} f_i(v^*)} - \frac{f_i(v_t)}{\sum_{j \in K_b} f_i(v_t)} \right| \le \frac{\lambda^2\epsilon}{8n},$$

由（λ, L）-光滑性适用于 $\{f_i\}$，因为 $Q \cdot \lceil \frac{n-1}{k-1} \rceil \le \frac{\lambda^2\epsilon}{8nkL} \cdot T^c$。根据霍夫丁界，我们于是有 e

$$\Pr\left[\left|(\#\text{ times } i \in K_b \text{ chosen }) - \frac{Qf_i(v^*)}{\sum_{j \in K_b} f_i(v^*)}\right| > \frac{Q\lambda^2\epsilon}{4n}\right] \le 2\exp\left(-\frac{Q^2\lambda^4\epsilon^2}{64n^2Q}\right) \le \frac{\delta}{T},$$

为 $Q \ge \frac{64n^2}{\log(2T/\delta)\,\lambda^4\epsilon^2}$。若此结论对每个 K_b 中的所有 i 都成立（包括 $i = 1$的每个实例），则我们有

$$\left| \hat{F}_i - \frac{f_i(v^*)}{f_1(v^*)} \right| \le \frac{\lambda\epsilon}{2n},$$

为 $f_1(v^*) \ge \lambda$。注意，将其归一化为 $\{F_i\}$ 等价于对每个物品观测到的经验频率重新缩放，使它们的得分在 $i = 1$ 上保持一致，且所有物品中的最高得分至多为 1；这不会增加任何物品的相对误差。因此，我们有

$$|F_i - f_i(v^*)| \le \frac{\lambda\epsilon}{2n}$$

同样成立。此外，在任何后续轮次中，若 $\|v_t - v^*\|_1 \le \frac{\lambda\epsilon}{2nL}$，我们有

$$\begin{aligned} |F_i - f_i(v_t)| &\le |F_i - f_i(v^*)| + |f_i(v_t) - f_i(v^*)| \\ &\le \frac{\lambda\epsilon}{n}, \end{aligned}$$

由于每个 f_i 都 L关于 ℓ_1 范数是 Lipschitz 的，因此对于所有轮次 t 其中 DBG 被采用时，都有$|F_i - f_i(v_t)| \le \frac{\lambda\epsilon}{n}$ 对每个 i 以概率 $1 - \delta$成立；若$\delta = 1/T$，则这对我们的期望遗憾至多贡献 1。根据引理 5，这至少在随后 $\frac{\lambda\epsilon}{4nL} \cdot T^c$ 轮次中成立。现在我们证明，这使得 $|\xi_i| \le \frac{\epsilon}{n}x_i$ 在每一轮中的界为 DBG。

引理 6。假设对于每个 i， $|F_i - f_i(v_t)| \le \frac{\lambda\epsilon}{n}$。那么，对于任意实现了得分为 $\{F_i\}$ 时的选择分布 x 的菜单分布 z，得分为 $\{f_i(v_t)\}$ 时的选择分布 y 满足 $|x_i - y_i| \le \frac{\epsilon}{n} x_i$。

证明。对每个 i，令 $y_i = {}^{f_i}\!/_{(v_t)\,F_i}\ x_i$。注意，$(x, \{F_i\})$ 和 $(y, \{f_i(v_t)\})$ 得到相同的菜单时间值：

$$\frac{k \cdot \frac{x_i}{F_i}}{\sum_{j=1}^{n} \frac{x_j}{F_j}} = \frac{k \cdot \frac{y_i}{f_i(v_t)}}{\sum_{j=1}^{n} \frac{y_j}{f_j(v_t)}}.$$

根据引理 1 中的第一种构造，这意味着任何在 $\{F_i\}$ 下实现 x 的菜单分布 z 都满足：

$$\frac{x_i}{F_i} = \sum_{K \in z: i \in K} \frac{z_K}{\sum_{j \in K} F_j}$$

因此，相同的分布也满足

$$\begin{aligned}\sum_{K \in z: i \in K} \frac{z_K}{\sum_{j \in K} f_i(v_t)} &= \frac{\frac{f_i(v_t)}{F_i} x_i}{f_i(v_t)} \\ &= \frac{y_i}{f_i(v_t)}\end{aligned}$$

在 $\{f_i(v_t)\}$ 下，得到选择分布 y。因此，我们有

$$\begin{aligned}|x_i - y_i| &\le \left| \frac{F_i - f_i(v_t)}{F_i} \right| x_i \\ &\le \frac{\epsilon}{n} x_i.\end{aligned}$$

□

给定一组分数 $M^* = \{F_i\}$，可行分布集合可以通过线性约束表示为

$$\mathrm{IRD}(M^*) = \{x \in \Delta(n) : \frac{k x_i}{F_i} \le \sum_{j=1}^{n} \frac{k x_j}{F_j}\}.$$

为了确保不会因 IRD 估计不精确而从目标集合中移除属于 EIRD 的点，我们可以沿每个维度将每个目标集合放宽 $\frac{\epsilon}{n}$：

$$\mathrm{IRD}_\epsilon(M^*) = \left\{ \left[\left(1 - \frac{\epsilon}{n}\right) x, \left(1 - \frac{\epsilon}{n}\right) x \right] \cap \Delta(n) : x \in \mathrm{IRD}(M^*) \right\}.$$

不过，由于在 DBG 中， $\mathcal{K}_t$ 到$\mathcal{K}_{t,\epsilon}$ 的 ϵ-收缩，这些点无论如何都不会被我们的算法选中。根据 ϵ 的界限，用于更新我们估计值的总时间 F_i （按 q 计）至多为ϵT，因为至少

$$\frac{\lambda\epsilon}{4nL} T^c \ge \frac{64n^2(n-1)\log(2T/\delta)}{(k-1)\lambda^4\epsilon^3}$$

每经过若干轮次，便会进入一个长度为 $Q \cdot \lceil \frac{n-1}{k-1} \rceil = {}^{64n^2}\!/_{(n-1)\log(2T/\delta)\,(k-1)\lambda^4\epsilon^2}$ 的学习阶段。因此，我们的总期望遗憾可界定为

$$\max_{x^* \in \mathrm{EIRD}(M)} \sum_{t=1}^{T} r_t^\top x^* - \mathbb{E}\left[\sum_{t=1}^{T} r_t(i_t)\right] \le \tilde{O}\left((n/\lambda)^{3/2} L^{1/4} T^{-c/4}\right)$$

假设在前 T^c 个阶段以及用于估计 F_i 的 $q \le \epsilon T$ 个阶段中，每一阶段都获得最坏情况奖励。

情况 2： $c > c^*$。当 $c > c^*$ 时，我们的分析与之前类似，唯一的例外是，我们不再能对 $\|v_t - v_1\| \le \frac{\lambda\epsilon}{2nL}$ 在对 F_i 的更新之间所经历的轮次数给出统一上界。不过，我们证明摊销学习时间相近，并且 q 仍可由 $\tilde{O}(T^{1-c/4})$ 界定。令 $c^*(t) = \min(c, \log(t)/\log(T))$ 为满足 $t = T^{c^*(t)}$（对于 $c^*(t) \le c$）的 c^* 参数化；注意，根据 $c^*(t)$ 而非 c，我们仍可应用引理 5。在任意 $t \ge T^{3/4}$，$c^*(t) \ge 3/4$ 足以界定 v_t 在 $Q \cdot \lceil \frac{n-1}{k-1} \rceil =$ 个学习轮次中发生的必要变化，正如我们在情况 1 中所做的那样，

$$\frac{\lambda^2\epsilon}{8nkL} \cdot T^{c^*(t)} \ge \frac{64n^2(n-1)\log(2T/\delta)}{(k-1)\lambda^4\epsilon^2}. \qquad (\epsilon^3 \ge 512n^4L\lambda^{-6}\log(2T/\delta)T^{-3/4})$$

对于任意范围 $t = [2^h, 2^{h+1}]$ for $h \ge \log(T^{3/4})$ 和 $h \le \log(T^c)$，我们的 v_t 界至少在以下

$$\frac{\lambda\epsilon}{4nL} \cdot T^{c^*(t)} \ge \frac{\lambda\epsilon}{4nL} \cdot 2^h$$

轮次内成立，因此在此窗口内最多会出现 $\frac{4nL}{\lambda\epsilon}$ 个学习阶段。对于 $h \ge \log(T^c)$，我们可以应用与情况 1 相同的界。对于所有 $h \le \log(T)$，我们有

$$\begin{aligned} q &\le \frac{4nL}{\lambda\epsilon} \cdot \frac{64n^2(n-1)\log(2T/\delta)}{(k-1)\lambda^4\epsilon^2} \log(T) \\ &\le \epsilon T \log(T), \end{aligned}$$

由此得到的总遗憾界与情况 1 中的界相差至多一个 $\log(T)$ 因子。 □

C 省略的证明，自第4

C.1 引理的证明2

证明。 由引理 1，只需证明对于任意这样的点 μ_i，菜单时间 x 至多为 1。鉴于 v 和 x 位于 $\Delta^\phi(n)$ 内，且每个 f_i 都是伪递增的，我们有

$$\frac{x_i}{((1-\lambda)v_i + \lambda)\sigma} \le \frac{x_i}{f_i(v)} \le \frac{\sigma x_i}{(1-\lambda)v_i + \lambda} \tag{7}$$

并且当 $x \in B_{\lambda\phi}(v)$ 时，$x_i \le \min(v_i + \lambda\phi/\sqrt{2}, 1 - \phi(n-1)/n)$。回想一下，$x$ 的菜单时间 μ_i 为

$$\mu_i(x) = \frac{k \cdot \frac{x_i}{f_i(v)}}{\sum_{j=1}^n \frac{x_j}{f_j(v)}}.$$

首先，我们证明该分子至多为 $k\sigma$。我们有

$$\begin{aligned} x_i &\le v_i + \lambda\phi/\sqrt{2} \\ &\le (1-\lambda)v_i + \lambda(1 - \phi(n-1)/n) + \lambda\phi/\sqrt{2} \\ &\le (1-\lambda)v_i + \lambda \qquad (\text{for } n \ge 4) \end{aligned}$$

由此可得 $kx_i/f_i(v) \le k\sigma$ 由（7）得到。我们还可以将菜单分母的下界表示为 $k\sigma$。令$\alpha_j = v_j - x_j$，其中有：

$$\sum_{j=1}^n \frac{x_j}{f_j(v)} \ge \sum_{j=1}^n \frac{v_j - \alpha_j}{((1-\lambda)v_j + \lambda)\sigma}.$$

对 v_j 中的任意项进行微分，可得：

$$\frac{\partial}{\partial v_j}\frac{v_j-\alpha_j}{((1-\lambda)v_i+\lambda)\sigma}=\frac{((1-\lambda)v_i+\lambda)\sigma-(v_j-\alpha_j)(1-\lambda)\sigma}{\sigma^2((1-\lambda)v_i+\lambda)^2}$$
$$=\frac{\lambda+\alpha_j(1-\lambda)}{\sigma((1-\lambda)v_i+\lambda)^2}$$

对于任意 v_j，其均为正，因此对于任意固定的 α_j （因为 $|\alpha_j|\le\lambda$），每个分子项都是递增的。因此，在有效的 v_j 和 x_j 范围内，当 $x_j=\phi/n$ 且 $v_j=\phi/n+\lambda\phi/\sqrt{2}$ 时，每一项均达到最小值，从而得到

$$\begin{aligned}\sum_{j=1}^{n}\frac{x_j}{f_j(v)}&\ge\frac{\phi}{((1-\lambda)(\phi/n+\lambda\phi/\sqrt{2})+\lambda)\sigma}\\&\ge\frac{(k\sigma)\cdot\phi}{k\sigma^2\lambda+k\sigma^2(1-\lambda)\phi/n+k\sigma^2(1-\lambda)\phi\lambda/\sqrt{2}}\\&\ge\frac{(k\sigma)\cdot\phi}{\phi/4+\phi/2+\phi^2/(4\sqrt{2})}&(\phi\ge4k\lambda\sigma^2,\ \sigma^2\le n/(2k))\\&\ge k\sigma.&(\phi\le1)\end{aligned}$$

因此，我们有

$$\frac{k\cdot\frac{x_i}{f_i(v)}}{\sum_{j=1}^{n}\frac{x_j}{f_j(v)}}\le1,$$

因而 $x\in\mathrm{IRD}(v,M)$。 □

C.2 Anal算法 3的分析：针对 $\Delta^{\Phi}(n)$ 的尺度有界模型

定理9 (尺度有界折扣遗憾界)。 对于任意具有偏好模型 M 且该模型是 $(\sigma,\ \lambda)$-尺度有界且 $(\frac{\lambda}{\sigma},\ L)$-光滑的智能体，其中 $\sigma\le\sqrt{n/(2k)}$，并且具有 γ-折扣记忆，适用于 $\gamma=1-\frac{1}{T^c}$，对于 $c\in(0,\ 1)$，算法 3 可获得遗憾

$$\max_{x^*\in\Delta^{\phi}(n)}\sum_{t=1}^{T}r_t^{\top}x^*-\mathbb{E}\left[\sum_{t=1}^{T}r_t(i_t)\right]=\tilde{O}\left((n^4L\left(T^{1-c/2}+T^{1/2+c/2}\right)\right)$$

相对于 ϕ-平滑单纯形，对于 $\phi=4k\lambda\sigma^2$。

˜预热期。这里我们分析算法在最初 $O(T^c)$ 轮次中的行为，并证明每个$v_{t,\ i}$ 都以高概率趋近于 $\frac{1}{n}$。在 $t=T^c$，记忆向量 v_t 完全集中在 k 个物品上，且至少有 $n-k$ 个其他物品满足 $v_{t,\ i}=0$。我们通过证明，在接下来的 $\tilde{O}(T^c)$ 轮次中，每轮选取 $v_{t,\ i}$ 最小的 k 个物品，就能以高概率达到记忆向量 $v_t\in\left[\frac{1}{n}\pm\frac{\epsilon}{4n^2L\sigma}\right]^n$。假设$v_{t,\ i}$ 至少比其他至多 $k-1$ 个物品的值低 T^{-c}，且原先最小的 $n-k$ 个物品的值均在 $\lambda/2$ 范围内满足 $v_{t,\ j}$。那么， i 在第 t 轮被选中的概率至少为

$$y_{t,i}\ge\frac{\lambda}{k\sigma^2(\lambda+(1-\lambda)(\lambda/2))}$$
$$\ge\frac{4}{3n}.$$

因此，$v_{t,i}$ 比平均的v_t更快地趋近于 $\frac{1}{n}$，直到它不再属于最小的 k 个之列。对于$\lambda>T^{-c/2}$，v_t的$_i$比其他任一 $n-k$低超过 $\lambda/2$（或任何常数）的概率会指数地衰减；因此，在经过足够多轮次、使前 T^c 轮的记忆权重降至 $O(T^{-c})$ 后， $2T^c$时每个 $v_{t,i}$ 的期望值为$\frac{1}{n}+O(T^{-c})$。根据引理 [...14]的分析，这会在之后 $O(T^c\log(T^c))$ 轮次内发生。根据Azuma不等式，考虑追踪 $v_{t,i}$ 与其条件期望之间偏差的鞅，这一结论也以高概率成立，误差至多为 $\frac{1}{n}+\tilde{O}(T^{-c/2})=\pm\frac{\epsilon}{4n^2L\sigma}$。若该结论在 $O(T^c)$ 轮次内未能成立，此事件对总期望遗憾的贡献为 $O(1)$。˜在整个证明过程中，我们将对数项和一些常数隐藏在 O 记号中。

算法 2（针对 Δ^{ϕ} 的尺度有界模型）。

令 $\epsilon = \tilde{O}(n^4 L \cdot \max(T^{-c/2}, T^{c/2-1/2}))$、 $S = \tilde{O}(n^{3/2} T^c)$、 $\eta = \tilde{O}(n^{-1/2} L \cdot T^{c/2-1/2})$。

--- 预热期 ---

对于 $t = 1$ 到 T^c，执行

展示代理菜单 $K = \{1, \ldots, k\}$

结束循环

当 $\max_i \left| v_{t,i} - \frac{1}{n} \right| \geq \frac{\epsilon}{4n^2 L \sigma}$ **do**

展示具有最小 $v_{t,i}$ 的智能体 k 个物品，并在并列时随机选择，最多至 T^{-c}

结束当

--- 初始学习 ---

执行 T^c 轮次

若 $v_{t,i} <$，则令 $F_{t,i} = \sigma^{-1}((1-\lambda)v_{t,i} + \lambda)\ \frac{1}{n}$，否则 $F_{t,\ i} = \sigma((1-\lambda)v_{t,\ i} + \lambda)$

令 $z_t = \text{MenuDist}(v_t, \mathbf{u}_n, \{F_{t,i}\})$，显示智能体 $K_t \sim z_t$

结束循环

设置 $F_i = \sum_t \mathbf{1}(i_t = i) \cdot F_{t,\ i} / (\frac{1}{n} T^c)) \cdot C(\sum_{j=1}^{n} F_{t,\ j})^{-1}$，设置 $v^* = v_t$

--- 优化 ---

初始化 OGD 在 $\Delta_{\epsilon}^{\phi}(n)$ 上运行 T/S 轮次，使用 η，设置 $x_1 = v^* := v_t$。

对于 $s = 1$ 到 T/S，执行

接收 $x_{\ s}$ 来自 OGD

执行 S 轮次

如果 $\|v_t - v^*\|_1 \geq \epsilon/(2nL)$，则

令 $\tilde{v} = v_t$

为智能体 $K_t \sim z = \text{MenuDist}(v_t, \tilde{v}, \{F_i\})$ 显示 T^c / L^2 轮次

设置 $F_i = \sum_t \mathbf{1}(i_t = i) \cdot F_i / (\tilde{v}_i T^c / L^2) \cdot C(\sum_{j=1}^{n} F_j)^{-1}$，设置 $v^* = v_t$

结束 if

为智能体 $K_t \sim z_t = \text{MenuDist}(v^*, x_s, \{F_i\})$ 显示

结束 for

设置 $\tilde{\nabla}_s = \sum_{h=t-S+1}^{t} e_{i_h} r_h(i_h) / (x_{h,\ i} S)$，更新 OGD

结束循环

初始学习。在此，我们利用尺度有界偏好结构，为接近当前记忆向量的得分获取高效的估计量。给定 $v_t \in \left[\frac{1-\alpha}{n}, \frac{1+\alpha}{n}\right]^n$，对于任意 $\lambda > 0$ 且 α足够小时，我们有

$$\frac{f_i(v_t)}{f_j(v_t)} \leq \frac{n(\lambda + (1-\lambda)\frac{1+\alpha}{n})}{2k(\lambda + (1-\lambda)\frac{1-\alpha}{n})} \leq \frac{n}{3k}$$

因此，对于任意分布 $x \in \left[\frac{1}{2n}, \frac{3}{2n}\right]^n$ 均匀分布，我们有菜单时间

$$\mu_i \leq \frac{k \cdot 3n/2}{n \cdot 3k/2} \leq 1$$

对于每个 i，根据真实分数 $f_i(v_t)$。我们的目标是学习这些分数的准确估计值。注意，我们提出的每组分数 $\{F_{t,i}\}$ 都满足尺度有界条件，并包含 $\mathrm{u}_n \in \mathrm{IRD}(v_t, \{F_{t,i}\})$；与之前一样，我们再次可以令 $v_t \in \left[\frac{1}{n} \pm \frac{\epsilon}{4n^2 L\sigma}\right]^n$ 在每个 T^c 轮次上以高概率成立，因为每个 $v_{t,i}$ 在每一轮的期望中都会向 $\frac{1}{n}$ 靠近（最多达到 T^{-c}），因此，在此过程中，鞅跟踪记忆向量在任意轮次相对于期望的最大偏差，都由 $\tilde{O}(T^{c/2})$界定。基于此，我们可以获得每个 $f_i(v_t)$的无偏估计量；我们首先假设对于每次试验，$\sum_i f_i(v_t) = \sum_i F_{t,i}$，随后将在样本聚合中通过重新归一化进行校正，使得 $\sum_i F_{t,i} = C$。

引理 7. 对任意 t，如果从菜单分布中抽样，该分布根据分数 $\{F_{t,i}\}$在 v_t处生成选择分布 x_t，且满足 $|x_{t,i} - v_{t,i}| \leq \lambda$，那么每个真实偏好分数$f_i(v_t)$的一个无偏估计量为

$$\mathbb{E}\left[\frac{F_{t,i}}{x_{t,i}} \cdot \mathbf{1}(i \text{ is chosen})\right] = f_i(v_t),$$

其取值范围有界于 $\frac{n}{2k\sigma}$，如果 $x_t \in \Delta^{\phi}(n)$ 且 $F_{t,i}$ 在 v处对 x_t 满足尺度有界约束 $_t$.

证明。 回顾引理 6，我们有，期望选择分布 y_t 满足 s

$$y_{t,i} = \frac{f_i(v_t)x_{t,i}}{F_{t,i}},$$

重新排列后即可得到该估计量，因为此处 $x_{t,i}$ 和 $F_{t,i}$ 均为固定值。我们可以利用尺度有界函数的性质以及 ϕ的条件来界定取值范围：

$$\begin{aligned}
\frac{F_{t,i}}{x_{t,i}} &\leq \frac{\sigma(\lambda + (1-\lambda)(x_i + \lambda))}{x_i} \\
&\leq \frac{\sigma(\lambda n + (1-\lambda)4k\lambda\sigma^2)}{4k\lambda\sigma^2} && (x_i \geq \phi/n \geq 4k\lambda\sigma^2/n) \\
&\leq \frac{n}{2k\sigma}.
\end{aligned}$$

□

根据尺度有界条件，量 $\sum_i f_i(v_t)$ 和 $\sum_i F_{t,i}$ 之间相差不超过 σ^2倍，因此将其重新归一化为 C 只会增大每个估计量的取值范围至 $\frac{n\sigma}{2k}$。将假设的每次试验 $\sum_i F_{t,i}$ 之和归一化为 C，正如我们在聚合以估计 F_i时所做的那样，可得到一组随机变量之和，其中每个变量的均值为 f_i n $T^c \geq \tilde{O}$ $n^8 L^2/\epsilon^2$ $\geq$(u)。对 ( ) ($n^6\sigma^4/$ $\tilde{O}$ $k\epsilon$ ^{2} v^* ( )) 的试验序列应用 Azuma不等式，便足以得到 $|f_i(\ ) - F_i| \leq \frac{\epsilon}{4n^2\sigma}$；此外，鉴于 f_i 满足 Lipschitz 条件且 $v^* \in \left[\frac{1}{n} \pm \frac{\epsilon}{4n^2 L\sigma}\right]^n$，我们也有 $|f_i(\mathrm{u}_n) - F_i| \leq \frac{\epsilon}{2n^2\sigma}$。在接下来的轮次中，只要 $\|v^* - v_t\| \leq \frac{\epsilon}{2n^2 L\sigma}$，这就能让我们准确地以 $\mathrm{IRD}(v^*)$ 中的任意分布为目标。

引理 8. 假设我们有估计值 F_i，它们在 v_t处满足关于 x_t 的尺度有界约束，其中$|x_{t,i} - v_{t,i}| \leq \lambda$，并且对于每个 i，$|f_i(v_t) - F_i| \leq \frac{\epsilon}{\sigma n^2}$。那么，按照 $\{F_{t,i}\}$ 以 x_t 为目标时，生成分布 y_t 满足 $\|y_t - x_t\|_1 \leq \frac{\epsilon}{n}$。

证明. 生成分布定义为

$$y_i = x_i + \frac{f_i(v_t) - F_i}{F_i} x_i$$

因此

$$|y_i - x_i| \le \frac{\epsilon}{\sigma n} \cdot \frac{x_i}{F_i} \le \frac{\epsilon}{n}.$$

□

在线梯度下降。我们将每批 S 轮次视为 OGD 的一步（其中 OGD选择 x_s 作为点），并证明以下各不变量以高概率在每一步中都能保持：

1. 我们完成每个学习阶段时，得到的估计值 F_i 满足 $|f_i(v_t) - F_i| \le \frac{\epsilon}{2\sigma n^2}$；2. 在每个以 x_s 为目标的轮次中，$|f_i(v_t) - F_i| \le \frac{\epsilon}{\sigma n^2}$均成立；4. 每一步开始时，我们都有$\|v_t - x_s\|_1 \le \frac{4\epsilon}{n}$；5. 每一步结束时，我们都有 $\|v_t - x_s\|_1 \le \frac{2\epsilon}{n}$；6. 该步骤每一轮的期望选择分布 x_t 满足$\|x_t - x_s\|_1 \le \frac{4\epsilon}{n}$ 和 $x_t \in \Delta^\phi(n)$。
3. Each gradient estimate satisfies $\left\|\tilde{\nabla}_t - r_s\right\|_2 \le \frac{4\epsilon}{n}$, $\left\|\tilde{\nabla}_t\right\|_2 \le 2\sqrt{n}$, and $\mathbb{E}\left[\left\|\tilde{\nabla}_s\right\|^2\right] \le O(n^2)$.

(1.) 只要 $\|v^* - v_t\| \le \frac{\epsilon}{2n^2 L\sigma}$成立，且每次更新都会使 $|f_i(v^*) - F_i| \le \frac{\epsilon}{2n^2\sigma}$在完成时成立，以上结论便成立。后者的推导与初始学习阶段类似；虽然我们的初始精度可能只有 $\frac{\epsilon}{n^2\sigma}$ 精度，但学习会在 $\tilde{O}(n^8/\epsilon^2)$ $\le T^c/L^2$ 轮次内达到目标精度 $\frac{\epsilon}{4\sigma n^2}$，对于 $\tilde{v}$而言，因此整个学习过程中的总漂移可界定为 $\frac{\epsilon}{4n^2 L\sigma}$，因为引理 5 表明，漂移所涉及的记忆占比有上界（假设 L有足够大的常数下界，不失一般性）。这进一步意味着更新后的 v^*处的 F_i 具有所需精度。

(2.) 只要我们在$\|v^* - v_t\| \ge \frac{\epsilon}{2n^2 L\sigma}$时重新学习，这一点就成立。

(3.) 每个 x_s 都位于 $\Delta^\phi_\epsilon(n)$ 内，并且记忆向量在 $\Delta^\phi_\epsilon(n)$ 之外的任何总漂移都可以由 $\frac{\epsilon}{2n} \cdot v_t$ 界定，因此我们的目标分布始终位于 $\Delta^\phi(n)$ 内；根据引理 2，对于 $\lambda\phi \gg \epsilon$，它也位于 $\mathrm{IRD}(v_t, M)$ 内。根据 (6.)，我们将有 $\mathbb{E}\left[\left\|\tilde{\nabla}_s - r_s\right\|\right] \le \frac{4\epsilon}{\sqrt{n}}$，与对 DBG 的分析类似，其中 $r_s = \sum_S r_t/S$。此外，对于每个 $x_t \ge \frac{\phi}{n} \ge \frac{\epsilon}{n}$，我们也将有 $\mathbb{E}\left[\left\|\tilde{\nabla}_s\right\|^2\right] \le O(n^2)$，正如对 DBG 所得到的结果一样；以高概率，范数界为 $2\sqrt{n}$ 。

(4.) 这在 $\left\|\tilde{\nabla}_s\right\|_2 \eta \le 2\sqrt{n}\eta \le \frac{\epsilon}{2n}$时成立。

(5.) 当 $\|v_t - x_s\|_2 \le \frac{4\epsilon}{n}$但 $\|v_t - x_s\|_2 \ge \frac{2\epsilon}{n}$时，我们用于定位 x_s（根据引理，最多耗时 8）足以使距离至少缩短 $\frac{3\epsilon}{8n^2 L\sigma}$，并且在漂移超过 $\frac{\epsilon}{2n^2 L\sigma}$之前，以高概率实现这一点；在每次重新学习 F_i时，遵循势函数 $\frac{\epsilon}{4n^2 L\sigma}$ 的漂移后，我们仍会以高概率更接近 $\frac{\epsilon}{8n^2 L\sigma}$。因此，最多需要 $16n\sigma$ 个阶段，即可达到距 x_s不超过 $\frac{2\epsilon}{n}$ 的状态；这满足$S = \tilde{O}(n^{3/2}T^c)$；并且在每次更新到 $v^*(T^c/L^2)$ 的总时间为 $\tilde{O}$。

(6.) 这源于每个学习阶段的准确性保证，以及应用于每个 $\tilde{v}$ 的漂移界。

遗憾界。基于这些结果，我们可以根据在线梯度下降的界，以及上述近似执行 OGD 所产生的误差，分析我们的遗憾。这给出了我们算法的总遗憾界：

$$\begin{aligned}\mathrm{Reg}_{\Delta^\phi(n)} = & \underbrace{S\left(\eta\cdot\sum_{s=1}^{T/S}\left\|\tilde{\nabla}_s\right\|^2+\frac{\sqrt{2}}{\eta}\right)}_{\text{OGD regret over } \Delta_\epsilon^\phi(n),\ T/S \text{ steps}} + \underbrace{\sum_{s=1}^{T/S}\sum_{h=t}^{t+S-1}\sqrt{n}\left(\left\|r_s-\tilde{\nabla}_s\right\|+\|x_h-x_s\|\right)}_{\text{gradient and reward error}} \\ & + \underbrace{\left(\max_{\Delta^\phi(n)\times\Delta_\epsilon^\phi(n)}\sum_t r_t^\top x^* - \sum_t r_t^\top x\right)}_{\text{target set imprecision vs. } \Delta_n^\phi} + \underbrace{\tilde{O}\left(T^c\right)}_{\text{burn-in and initial learning}} \\ & = \tilde{O}\left(n^2T^{1/2+c/2}+n^{7/2}L(T^{1/2+c/2}+T^{1-c/2})+n^4LT^{1-c/2}+T^c\right).\end{aligned}$$

定理的证明4

证明。 我们将其归约为最大独立集（MIS）问题的一个实例。MIS 问题是 Poly-APX-难的（例如参见 Arora 和 Barak （2009））， 因此， 除非 P = NP， 否则不存在常数因子多项式时间近似算法。给定一个图 G， 其中包含 $n-1$ 个顶点（我们可以假设其具有唯一的最大独立集 V^*， 且该独立集至多包含（$n-1$）/2 个顶点）， 我们为一个偏好模型 M 构造一个电路， 该模型包含 n 个物品；识别其 IRD 集合中所包含的最优物品分布，相当于识别最大独立集 S^*。对于奖励函数，我们假设物品 1 产生恒定奖励 1，而所有其他物品产生奖励 0。因此，目标函数是最大化物品 1 被选中的概率，但评分 $f_1(v) \geq \lambda$ 只有在未分配给物品 i 的记忆质量在对应于最大独立集的物品之间近似均匀分配时，才会达到最大值（其他评分函数可以恒定为 λ）。

我们用一个同时包含算术门和布尔门的电路来描述针对 f_1 的偏好模型，其中假定所有其他函数 f_i 在 λ处均为常数，并指出可以将其转换为纯布尔电路，且最多只会产生多项式级膨胀。令 $N=\{2,\ldots,n\}$ 对应于 G的顶点。对于任意记忆向量 v，令 $V=\{j>1:v_j\geq\frac{1}{n}\}$。我们的函数 f_1 定义如下：

$$f_1(v)=\begin{cases}\frac{|V|(1-\lambda)}{n-1}+\lambda & V \text{ is an independent set in } G\\ \lambda & V \text{ is not independent in } G\end{cases}.$$

要构造针对 f_1的电路，首先加入门 g_j：当且仅当 $v_j\geq\frac{1}{n}$ 时，这些门对每个 j 输出 1（否则输出 0）。有了这些门之后，我们还为图中的每条边加入门 e_{ij}；当且仅当 i 与 j 之间有边时，该门输出 1（其中 $1-e_{ij}$ 表示没有边）。我们可以通过对所有门取与门（$e_{ij}\oplus(g_i\wedge g_j)$），构造一个门 g_{valid}，使其当且仅当集合 V是独立集时输出 1；还可以构造一个门 g_{count}，通过对门 g_i求和来计算所提议的独立顶点数。将 g_{valid} 和 g_{count} 相乘后，如果 V 是独立集，则得到一个计数器，计数 V 的大小；否则结果为 0。随后可通过算术缩放得到 $f_1(v)$。

令 $\lambda=1/((n-1)^2/k+1)$。给定一个独立集 V，与 V相对应的记忆向量 $v\in\mathrm{IRD}(v,M)$ 在 f_1 下可获得的最高奖励可以用菜单时间表示，其中必须满足

$$\frac{k\frac{v_1}{f_1(v)}}{\frac{v_1}{f_1(v)}+\frac{(1-v_1)}{\lambda}}\leq 1$$

根据引理 1。求解 $(k-1)v_1 = (1-v_1)\left(1 + \frac{|V|}{\lambda}\frac{(1-\lambda)}{(n-1)}\right)$，得到 v_1 的表达式为

$$\begin{aligned} v_1 &= \frac{\frac{1}{k-1} + \frac{|V|(1-\lambda)}{\lambda(n-1)(k-1)}}{1 + \frac{1}{k-1} + \frac{|V|(1-\lambda)}{\lambda(n-1)(k-1)}} \\ &= \frac{\frac{k|V|}{n-1} + 1}{\frac{k|V|}{n-1} + k}. \qquad (\lambda(n-1)^2/k = 1-\lambda) \end{aligned}$$

对于任意 $|V| < (n-1)/2$，该值至多为 $1/2$，因此仍然可以分配菜单时间，使每个 $i \in V$ 都获得 $v_i \geq \frac{1}{n}$，从而当 $|V|$ 为最大独立集的大小时，对于某个 $v \in \mathrm{IRD}(v, M)$，v_1 是可行的最大奖励。v_1 关于 $|V|$ 的梯度为

$$\frac{\partial}{\partial |V|} \frac{|V| + (n-1)/k}{|V| + (n-1)} = \frac{(n-1)(1-1/k)}{(|V| + (n-1))^2}$$

对于 $|V| \in [1, n/2]$，其为 $\Theta(1/n)$。任何在 $O(1/n)$ 因子范围内近似 v_1 的多项式时间算法，都必须得到一个记忆向量 v'，该向量对应于一个独立集 V'，并且以常数因子近似 V^*；这将产生一个针对最大独立集的多项式时间常数因子近似算法。 □

E 省略的证明：第 6

E.1 定理 5 的证明

证明。我们分析的关键在于，考察菜单中的目标项目 i 保持不变时，长度为 t_{hold} 的窗口内项目频率的收敛情况。对于任意这样一个长度为$t_{\text{hold}} = O(\alpha^4{}^{1}(1-\gamma))$的窗口，我们可以确保累积奖励与当前奖励分布下的期望之差不超过 α。由于我们选择记忆中权重最小的 $k-1$ 个物品，因此记忆中除 i 之外的物品的总权重至多为 $(k-1)(1-v_i)n$；给定当前记忆向量 v，从菜单 K_t 中选取物品 i 的概率为：

$$\begin{aligned} \Pr[i^* \text{ selected}] &= \frac{f_i(v)}{\sum_{K_t} f_j(v_i)} \\ &\geq \frac{(1-\lambda)v_i + \lambda}{(1-\lambda)v_i + \lambda + \left(\frac{(1-v_i)(1-\lambda)}{n} + \lambda\right)(k-1)\sigma^2} \end{aligned}$$

根据伪递增性质。我们将分析$v_{t,i}$ 随时间变化的期望，其中$E_t = \mathbb{E}[v_{t,i}]$，并证明其趋近于 $1-\phi(n-1)/n$，即平滑单纯形对应顶点处的概率。一个挑战在于，给定当前期望，将概率分配给$v_{t,i}$ 的值有许多可能方式，而这些方式都会产生 E_t。二阶导数检验表明，当 $\sigma^2 \leq n/(2k)$时，上述概率函数关于正 v_i 是凹函数；注意，分子和分母均为正数，并且都随 v_i线性递增，而分子始终较小，但在 v_i中的增长更快。因此，我们可以应用詹森不等式，将考察范围限制在极端情形：期望E_t 完全由以下试验构成：其中$v_{t,i} = 0$ 或$v_{t,i} = 1$；这确实发生在 $\gamma = 0$处。我们可以将P_0 和 P_1 定义为每种情形下选择概率的下界：

$$\begin{aligned} \Pr[i^* \text{ selected} | v_{t,i} = 1] &\geq \frac{1}{1 + (k-1)\lambda\sigma^2} \\ &\geq 1 - k\lambda\sigma^2 \\ &:= P_1; \end{aligned}$$

$$\begin{aligned}\Pr[i^* \text{ selected}|v_{t,i}=0] &\geq \frac{\lambda}{\lambda+(\frac{1-\lambda}{n}+\lambda)(k-1)\sigma^2}\\ &= \frac{1}{1+\left(\frac{1-\lambda}{n\lambda}+1\right)(k-1)\sigma^2}\\ &\geq \frac{1}{1+(k\sigma^2+1)(k-1)\sigma^2}\\ &\geq \frac{1}{2\sigma^2k^4}\\ &:= P_0.\end{aligned}$$

因此，我们有

$$\begin{aligned}E_{t+1} &= \mathbb{E}[v_{t+1}|E_t]\\ &\geq (1-\gamma)\left(E_t\cdot\Pr[i^* \text{ selected}|v_{t,i}=1]+(1-E_t)\cdot\Pr[i^* \text{ selected}|v_{t,i}=0]\right)+\gamma E_t\\ &\geq (1-\gamma)\left(E_t\cdot P_1+(1-E_t)\cdot P_0\right)+\gamma E_t.\end{aligned}$$

我们可以求解满足 $E_{t+1}=E_t$ 的 E_t^*，即 $E_t\cdot P_1+(1-E_t)\cdot P_0=E_t$ 的情形，如下：

$$\begin{aligned}E_t^* &= \frac{1}{1+2\lambda\sigma^6k^3}\\ &\geq 1-2\sigma^6k^3\lambda\\ &\geq 1-\phi(n-1)/n \qquad (n\geq 3)\end{aligned}$$

并进一步求出满足 $E_{t+1}\geq(1-\gamma)(E_t+\alpha)+\gamma E_t$ 的值 E_t^α，如下：

$$E_t^\alpha = E_t^* - 2\sigma^4k^2\alpha.$$

请注意， E_{t+1} 的增长率随 t 增大而递减，并最终达到一个不动点；鉴于当 E_t 距离 E_t^* 在 $O(\alpha)$ 范围内时，其增长率关于 α 呈线性，因此达到 $E_t^*-O(\alpha)$ 所需的累计轮数至多为 $O(\alpha^{-1}(1-\gamma))$。如果我们在此之后继续进行 $O(\alpha^{-2}(1-\gamma))$ 轮，那么这些最初的轮次对总期望的贡献至多为 α；物品 i 被选中的轮次比例至少为 $E_t^*-\alpha$；从期望来看，每种其他物品的出场比例也会迅速趋近于 $\frac{1-E_t^*}{n-1}$。

将每批这样的 $O(\alpha^{-2}(1-\gamma))$ 轮次视为一次试验，并重复 $\tilde{O}(1/\alpha^2)$ 次试验，总计得到 $t_{\text{hold}}=O(\alpha^{-4}(1-\gamma))$ 步。我们可以将每次试验视为相互独立，因为在我们考虑的悲观下界下，将记忆向量重置为某个较低的 v_i 值只会降低期望奖励。根据霍夫丁界，奖励与当前分布下的期望值之间的差距不超过 α，而该当前分布对应于 i 在 ϕ-平滑单纯形中的“臂”。为完成分析，注意到我们的总遗憾（使用 EXP3 的 $\tilde{O}(T^{1/2})$ 界）为：

$$\begin{aligned}\text{Reg}_{\Delta^\phi(n)}(\mathcal{A}_3;T) &= \tilde{O}(t_{\text{hold}}\cdot\left(\frac{T}{t_{\text{hold}}}\right)^{1/2}+\alpha T)\\ &= \tilde{O}(\frac{T^{1/2}}{\alpha^2}+\alpha T)\\ &= \tilde{O}(T^{5/6})\end{aligned}$$

在设定 $\alpha=O(T^{-1/6})$ 后。

□

E.2 定理6的证明

证明。我们考虑一组模型和奖励函数：物品 1 在每一轮的奖励均为 1，其他所有物品的奖励均为 0。对于 $\gamma=\frac{1}{2^c}$，令某个常数为 $c>1$。注意，记忆中任一步的权重都大于此前所有步骤的权重之和，因此记忆向量 v_t

恰好编码了前 $t-1$ 轮的物品选择历史记录。令 $h = \log_2(n)$；当 t 远大于 h 时，步骤 1 到 $t-h$ 的权重之和为 $\Theta(1/n)$。我们将考虑状态 s，它们是记忆向量空间的子集，分别对应于截断为此前 h 步的每一种可能历史记录，并且彼此之间的距离至少为 $O(1/n)$。我们稍微滥用一下记号，将每个记忆向量 v_t 表示为其舍入后的状态 s_t。记忆模型在每个状态内的行为保持不变，并在状态之间进行光滑插值；对于不可行的记忆向量，可任意定义模型，使其满足 $L = \text{poly}(n)$ 的 Lipschitz 连续性。我们生成 $\mathcal{M}$ 的过程如下：

- 令 $k = n/2$;
- 令 $\lambda = \frac{1}{n-k+1}$;
- 对于每个 $s \in [n]^h$，令 G_s 为从 $\{2, \ldots, n\}$中均匀随机抽取的 $k-2$ 个物品组成的集合；
- 若 $i = 1$ 或 $i \in G_s$，则令 $f_i(s) = \lambda$；否则令 $f_i(s) = 1$。

注意，在 s 处的最优策略 π^* 是选入物品 1、G_s中的每个物品，以及任意一个最终物品。注意，得分为 λ 的 $k-1$ 个物品中，每个被选中的概率为

$$\begin{aligned}\Pr[i \text{ chosen} | f_i(s) = \lambda, \pi^* \text{ played}] &= \frac{\lambda}{1 + (k-1)\lambda} \\ &= \frac{1}{n},\end{aligned}$$

因此，每轮的期望奖励也为 1。注意，π^* 与这样一种菜单分布相一致：该分布在物品 1 和 G_s之后，均匀随机选择最终物品，从而生成均匀分布。因此，均匀分布属于 EIRD （可以直接为不可行的记忆向量定义得分，使得对任意 $v \in \Delta(n)$，可行性均成立）。我们还可以看出，任何与 π^*不一致的菜单，其期望奖励至多为：

$$\begin{aligned}\frac{\lambda}{2 + (k-2)\lambda} &= \frac{\frac{1}{n-k+1}}{\frac{2n-2k+2}{n-k+1} + \frac{k-2}{n-k+1}} \\ &= \frac{3}{4n},\end{aligned}$$

因为必须包含某个不在 G_s 中的物品。为了给任何算法的遗憾提供下界，考虑任意时间 t 以及物品选择历史记录。截至 t，目前至多观察到 个不同的状态。考虑以下情况：

- 算法从 t 到 $t+h-1$ 的每一轮都选择与 π^* 一致的菜单；
- 算法在 t 到 $t+h-1$ 之间的某一轮选择了与 π^* 不一致的菜单。

假设 t 小于 $T = \frac{1}{2} \cdot (\frac{n}{2})^h = O(n^{\log(n)})$。在前一种情况下，智能体在每一轮选择的 $\frac{n}{2}$ 个物品服从均匀分布，因此任意给定状态的最大概率至多为 $(\frac{n}{2})^h$。由于集合 G_s 是针对每个状态独立随机生成的，算法对于未访问状态的 G_s 没有任何信息，因此无法通过为 G_s 随机选择一个假设来获得更高的期望奖励；这种选择在第 $t+h$ 轮会产生 $O(\frac{1}{n})$ 的遗憾。在后一种情况下，选择与 π^* 不一致的菜单的那一轮还会额外产生 $O(\frac{1}{n})$ 的遗憾。当 $t < T$ 时，每个事件在期望上每 h 轮至少发生一次，因此截至 T，任何算法每轮都必须承担 $O(\frac{1}{nh})$ 的期望遗憾。

□