中国科学技术大学

学士学位论文

不同设定下的多用户多臂老虎机模型研究

作者姓名: 熊伟

学科专业: 数学与应用数学

导师姓名: 兰小红 副教授

完成时间: 二〇二三年六月二十三日

University of Science and Technology of China

A dissertation for bachelor's degree

Algorithms of Multi-player Multi-armed Bandits under different settings

Author: Wei Xiong

Speciality: Mathematics and Applied Mathematics

Supervisor: Associate Prof. Xiaohong Lan

Finished time: June 23, 2023

致　谢

Pass

目 录

中文内容摘要

分布式多用户多臂老虎机模型近年来是多臂老虎机模型的一大热点, 在这个模型中, M 个用户共同在 K 个支撑集为 $[0,1]$ 的分布上进行 T 轮连续的采样, 采样结果被称为用户获得的奖励, 用户的目的是最大化群体的期望奖励。此时, 用户无法显式地进行通信, 不同用户通过碰撞 (多个用户采样同一个分布) 进行交互, 当碰撞发生, 涉及碰撞的用户将会获得奖励 0。在这篇论文里, 我们考虑三种典型的多用户多臂老虎机模型, 并设计一个统一的由初始化、探索-通信、守成三个阶段构成的算法框架, 其在三种模型下的实例化是目前算法设计的最优结果, 特别的, 前两个模型的算法均达到了问题的理论下界, 因此在相差一个常数因子的意义下, 它们是最优的。

关键词: 多臂老虎机; 多用户多臂老虎机; 集中不等式; 编码理论

Abstract

Decentralized multi-player multi-armed bandits (MPMAB) problems have received significant interest in recent years. In this model, M players make decisions simultaneously on K distributions supported on $[0, 1]$. The samples on these distributions are called the rewards for the players and players cooperatively make decisions to maximize the cumulative rewards. Players are not allowed to communicate with each other, but interact through arm collisions where a collision happens when multiple players sample the same distribution and all these players receive reward 0. In this paper, we consider three typical settings of MPMAB problem and propose an algorithmic framework, called, BEACON – *Batched Exploration with Adaptive COmmunicatioN* – to solve the problem. The implementations of this framework achieve the state-of-the-art results in different settings. In particular, the first two algorithms are optimal up to a constant factor since they match the theoretical lower bound of the problems we consider.

Key Words: Multi-armed bandits; Decentralized multi-player multi-armed bandits; Concentration inequality; Coding theory

第一章　简介与问题表述

第一节　简介

多臂老虎机模型 (multi-armed bandit, MAB) 是在线学习与强化学习的一个基本而重要的模型, 它描述了在面临多个选择对象并要进行长期的序列决策时所面对的一个核心的权衡问题: 我们是应该探索 (exploration) 来尝试更多的可能性, 还是应该选择守成, 根据历史信息选择经验最优的对象? 多臂老虎机的名字来源于下面这个非常著名的例子, 假定一个赌徒走进了一家赌场, 赌场有 K 台老虎机 (也称为臂, arms), 我们假定每台老虎机 k 被赌徒选中并摇了之后将会以 p_k 的概率吐出一块钱, 以 $1-p_k$ 的概率不吐钱, 赌徒总共能摇 T 次老虎机, 用什么样的策略才能使赌徒在期望意义下获得最多的金钱? 这就是经典的多臂老虎机模型所研究的问题。将这个问题对应的统计问题是: 我们需要连续对 K 个分布进行 T 次采样, 其中分布 k 的均值为 $\mu_k = p_k \in [0,1]$, 假定在第 t 轮我们选择分布 k, 奖励 X_k^t 将会从分布 $\mathcal{D}_k$ 中采样, 我们假定每一次选择分布 k 的采样是独立同分布的, 并希望最大化 $\sum_{t=1}^{T} X_k^t$。

MAB 的其中一个分布式版本被称为多用户多臂老虎机问题 (multi-player MAB, MPMAB), 在这个模型中, M 个用户将会同时在 K 个分布上进行 MAB 游戏。其中, ” 用户” 这一词是由于这个模型最早来自于无线通信领域, M 个用户将会在 K 个信道中进行动态的选择。模型假定 M 个用户无法进行显式的通信, 相反, 用户之间的交互通过碰撞 (collision) 进行: 当一个信道/分布被多个用户使用/采样, 所有用户将会获得奖励 0。这是由于, 不同用户关于信道资源是竞争关系, 当多个用户同时选择同样的信道, 彼此将互相成为噪声并导致通信质量迅速下降。如何在分布式的情况下完成长期的探索与守成的权衡, 并获得集体奖励的最大化是 MPMAB 模型所关心的问题, 由于碰撞的存在, 问题将比单用户场景更加具有挑战

性。

MAB 模型的研究可以追溯到上世纪的 30 年代, 并在最近二十年间成为研究热点, 其中, 当奖励严格遵从独立同分布的 K 个分布的情况 (称为 stochasitc MAB) 已经有了非常成熟的算法: 1) 证明了算法遗憾 (将在下一小节介绍) 的下界。2) Auer et al.[1] 提出的 UCB1 算法是 MAB 最著名且高效的算法之一, 且对更广泛的 bandit 模型有着深远的影响。3) Sébastien Bubeck 的博士论文以及 Garivier et al.[2] 在改进算法遗憾前的常数系数上进行了努力并获得了严格逼近理论下界的算法。因此, Stochastic MAB 模型的结构已经被很好的了解。另一方面, MPMAB 在 Liu et al.[3] 被系统地表述并研究, 并在最近十年获得了大量的关注。尽管有许多的算法已经被提出, 但是分布式的场景所带来的挑战使得现存的算法的遗憾要比中心化可通信的算法要更加差, 这在很长一段时间被认为是不可改善的。在 2020 年, Boursier et al.[4] 提出一种全新的算法框架 SIC-MMAB, 在分布式的场景下利用" 隐式通信" 来实现部分信息交换, 从而使得算法能够获得类似于中心化可通信的算法的表现, 从而使得 MPMAB 的研究进入新的时代, 但无论是这篇工作所引入的隐式通信的复杂度, 还是其框架所能解决的问题场景都仍然有一定的局限性。在这篇论文中, 我们将在 SIC-MMAB 的基础上给出一个更加通用的算法框架, 其能解决当前最为流行而广泛研究的三类 MPMAB 场景, 并获得与中心化可通信算法类似的表现。

第二节　模型与问题

在这一节中，我们用数学语言建模 MPMAB 模型，并给出衡量一个 MPMAB 算法的准则。

MPMAB 模型包含有 M 个用户与 K 个支撑集为 $[0,1]$ 的分布 $\{\mathcal{D}_1,...,\mathcal{D}_k\}$, 分布对应于均值 $(\mu_1,...,\mu_k)$。我们假定 $K>M$ 且初始时每个用户知道 K 但不知道

Algorithm 1 MPMAB for user m

1: **Input:** K arms, Time horizon: T;

2: **for** t=1,2,...,T **do**

3: user m selects arm $s_m^t \in [K]$;

4: user m observes info.

5: **end for**

M, 不同用户之间无法进行显式的通信。M 个用户将同时 K 个分布上进行 T 轮的决策, 在第 t 轮, 用户 $m \in [M]$ 将会选择一个分布 $s_m = k$, 真实奖励 $X_{k,m}^t$ 将根据分布 $\mathcal{D}_k$ 采样。我们记用户选择的分布为 $S = (s_1, s_2, ..., s_M)$, 则用户 m 最终获得的奖励将会由 $X_{k,m}^t$ 与下面的碰撞示性函数决定:

$$\eta_k(S) := \mathcal{I}\{|C_k(S)| \leqslant 1\} \quad 其中： \ C_k(S) := \{n \in [M] | s_n = k\}. \tag{1.1}$$

用户观察到的奖励形式为:

$$O_{k,m}(t) := X_{k,m}(t)\eta_k(S(t)). \tag{1.2}$$

换言之, 如果 m 是唯一选择这个分布的用户, 她观察到的用户奖励将会是真实的奖励, 否则, 她观察到的奖励将会是 0。由于分布的支撑集为 $[0, 1]$，我们知道在一轮内，碰撞会导致最坏的奖励。我们在伪代码 1中正式描述用户 m 所面临的 MPMAB 游戏:

我们并没有具体给出”info” 这个变量的定义, 这是因为, 根据用户所能观察到的信息内容, 我们可以将 MPMAB 分成不同的类型。

一、可观察与不可观察 (Sensing and no sensing)

如果用户可以观察到用户奖励 $O_{k,m}$ 与碰撞示性函数 $\eta_k(S)$, 我们称此时的 MPMAB 是可观察的; 如果用户只能观察到用户奖励 $O_{k,m}$, 我们称此时的 MPMAB 是不可观察的。可观察与不可观察的最大差别在于, 用户是否可以明确感知到她与其它的用户产生了碰撞。在无线通信的应用中, 许多通信协议返回的信息中是

可以读取这样一个示性函数的, 但在更广泛的应用中, 特别是传感器的代价高昂的场景, 这样的一个示性函数往往并不一定能观察到。此时, 在第 t 轮, 对于可观察与不可观察场景, 用户 m 将分别根据历史信息

$$\begin{aligned}&\{s_m^1, O_{s_m^1,m}, \eta_{s_m^1}(S^1), s_m^2, O_{s_m^2,m}, \eta_{s_m^2}(S^2), ..., s_m^{t-1}, O_{s_m^{t-1},m}, \eta_{s_m^{t-1}}(S^{t-1})\}\\&\{s_m^1, O_{s_m^1,m}, s_m^2, O_{s_m^2,m}, ..., s_m^{t-1}, O_{s_m^{t-1},m}\}\end{aligned} \tag{1.3}$$

来进行决策。

二、同质与非同质 (Homogeneous and heterogeneous)

根据 K 个分布对于不同用户是否相同, 我们将 MPMAB 分为同质与非同质: 如果对每个用户, 分布的均值都是 $(\mu_1, ..., \mu_K)$, 我们称 MPMAB 是同质的, 此时每轮的真实奖励 $X_{k,1} = X_{k,2} = ... = X_{k,M}$; 如果对不同的用户, 分布的均值是不同的, 我们称 MPMAB 是非同质的, 此时分布的均值将被记为一个矩阵 $(\mu_{k,m})_{k\in[K],m\in[M]}$, 注意, 不论是可观察还是不可观察的情况, 都只在有且仅有一个用户采样了某个分布时有可能观察到真实奖励 $X_{k,m}$。

三、衡量 MPMAB 算法的准则

我们记从 K 个分布中 (有放回) 的选择 M 个元素的所有方案为 $\mathcal{S}$, 则 $|\mathcal{S}| = K^M$。我们定义 $S = (s_1, ..., s_m) \in \mathcal{S}$ 的价值为 $V_S := \sum_{m=1}^{M} \mu_{s_m,m} \times \eta_{s_m}(S)$。注意这个定义对同质与非同质模型都是成立的, 因为我们可以把同质模型的均值向量扩展为每一行均为 $(\mu_1, ..., \mu_K)$ 的矩阵。此外, 我们定义最优价值为 $V_* := \max_{S\in\mathcal{S}} V_S$, 那么, 给定一个策略 π, 它在每一轮将会使得 M 个用户选择分布组合 $S(t)$, 我们用最优收益与它的期望收益之差作为衡量这个策略的准则:

$$R(T) = TV_* - \mathbb{E}\left[\sum_{t=1}^{T} V_{S(t)}\right].$$

我们有 $R(T)$ 非负, 且任意总期望收益为 TV_* 的策略被称为最优策略, 记作 π^*。

第三节　相关工作

在这一节, 我们总结与这篇论文紧密相关的工作。

同质化多用户多臂老虎机。Liu et al.[3] 与 Anandkumar et al.[5] 最早提出了多用户多臂老虎机 (MPMAB) 模型, 大部分这个领域的工作集中在同质化模型, 即待采样的 K 个分布均值对 M 个用户是相同的, 例如: Avner et al.[6], Rosenski et al.[7], Besson et al.[8]。特别的, Boursier et al.[4], Wang et al.[9] 利用碰撞进行隐式通信, 并成功获得了与中心化算法类似的理论结果, 因此, 同质化模型在探索层面上已经基本被解决。然而, 由于这两个工作的通信结构并不高效, 它们的最坏情况的损失会由通信损失主导, 而此时通信损失是理论最优结果的高阶无穷大量。

非同质化多用户多臂老虎机。更一般的非同质化模型在 Kalathil et al.[10] 中被提出, 比起同质化模型, 非同质化模型的研究要更少得多, 尽管没有明确表出, 同样的隐式通信思路在 Kalathil et al.[10], Nayyar et al.[11] 以及 Bistritz et al.[12,13] 提出的算法中被采用, 此外 Magesh et al.[14] 提出的 MUMAB, Tibrewal et al.[15] 提出的 ESE1 以及 Boursier et al.[16] 提出的 METC 也都直接或间接地利用碰撞来完成信息交换。这是由于在非同质化的模型下, 用户无法由本地信息来推测全局模型, 碰撞带来的交互在这种情况下是必要的。然而, 这些算法的表现距离中心化算法的理论下界还有很大的距离, 因此对于非同质化模型, 还有大量的挑战需要解决。

非线性奖励函数。在我们的调研结果中, 之前的文献中没有完整考虑一般非线性奖励函数的算法, Bistritz et al.[17] 考虑了 MPMAB 的公平性问题, 但它所考虑的模型与我们这考虑的奖励函数有本质不同。还有一些工作考虑找到一个” 稳定” 的分布分配策略, 包括 Avner et al.[18], Darak et al.[19], 这个模型相比起优化总奖励要简单一些。

碰撞示性函数的可观察性。由于碰撞是不同用户交互的唯一方式, 用户是否

能够观察到碰撞示性函数在 MPMAB 模型中可能带来本质困难。有许多工作考虑了更加困难的不可观察模型, 包括 Lugosi et al.[20], Shi et al.[21], Bubeck et al.[22]。特别地, Shi et al.[21] 提出的 EC-SIC 巧妙地利用了编码理论来给出一般性的推广可观察模型算法到不可观察模型的方法; 然而 EC-SIC 依赖于假设: 分布的均值有严格大于 0 的下界且已知, Huang et al.[23] 证明在最小均值为 0 的情况下, 此时我们仍然进行一定的" 预处理" 来得到类似的理论结果。

组合多臂老虎机问题。组合多臂老虎机问题某种意义上可以看做与 MPMAB 紧密联系的 (中心化) 算法。一般性的组合多臂老虎机框架第一次在[24] 中提出。在那之后, 许多组合模型变种吸引了学界的兴趣, 例如, 线性奖励函数在 Kveton et al.[25], Combes et al.[26], Degenne et al.[27] 中被研究, 拟阵多臂老虎机 (matroid bandits) 在 Kveton et al.[28], Talebi et al.[29] 中被研究。此外, 最近有一些关于 Thompson Sampling 在组合多臂老虎机问题中的应用, 包括 Wang et al.[30], Perrault et al.[31]。此外, 组合多臂老虎机模型的理论下界也是一个活跃的领域。Kveton et al.[32] 证明了线性奖励函数下的理论下界。当分布是彼此独立的, 理论下界可以在包括 Combes et al.[26], Degenne et al.[27] 中找到。对于一般的奖励函数, 其理论下界在最近的一项工作中给出 Merlis et al.[33]。

第四节　论文组织

这篇论文的剩下部分将会按照如下结构组织:

- 章节 1 的余下部分将介绍一些基础知识与相关的基本概念。
- 章节 2 将讨论单用户情况下多臂老虎机的算法。
- 章节 3 将在同质且可观察的情况下发展算法的基本框架。
- 章节 4 将在非同质且可观察的情况下设计新的探索算法。

- 章节 5 将给出编码方案的技巧以及理论的编码长度, 以解决不可观察模型。
- 章节 6 提供了充分的实验结果来验证我们的理论。
- 章节 7 讨论并总结这篇论文。

第五节　基础知识

这一小节将介绍一系列基础知识, 包括次高斯随机变量与集中不等式, 信息论中编码的概念, 以及算法的理论下界。

一、次高斯随机变量与集中不等式

根据马尔可夫不等式以及在概率符号中的不等式两边同时取指数函数的技巧, 我们可以得到如下的第一个概率不等式:

引理 1.1　Chernoff Bound。设 X 是一个有矩函数的随机变量, 那么我们有

$$P(X - \mathbb{E}X \geqslant t) \leqslant \inf_{\lambda \geqslant 0} \mathbb{E}[e^{\lambda(X-EX)}]e^{-\lambda t}$$

这启发我们, 当一个随机变量的矩函数的上界有一些特殊的结构，我们可以得到更加好的概率上界, 因此我们定义如下的次高斯族随机变量:

定义 1.1　一个随机变量被称为是 σ^2-次高斯的, 如果 $\forall \lambda \in R$，我们有

$$\mathbb{E}[\exp(\lambda(X - EX))] \leqslant \exp(\frac{\lambda^2\sigma^2}{2})$$

利用上面的 Chernoff 不等式, 我们立刻有

引理 1.2　。设 X 是 σ^2-次高斯的, 那么我们有

$$P(X - EX \geqslant t) \leqslant \exp(-\frac{t^2}{2\sigma^2})$$
$$P(X - EX \leqslant -t) \leqslant \exp(-\frac{t^2}{2\sigma^2})$$
$$P(|X - EX| \geqslant t) \leqslant 2\exp(-\frac{t^2}{2\sigma^2})$$

同时我们有, 独立的次高斯随机变量的和仍然是次高斯分布的:

引理 1.3 令 $X_1, ..., X_n$ 服从独立的 σ_i^2-次高斯分布, 那么

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}\left(X_i-\mathbb{E}\left[X_i\right]\right)\right)\right]\leqslant\exp\left(\frac{\lambda^2\sum_{i=1}^{n}\sigma_i^2}{2}\right),\forall\lambda\in R,$$

也就是说, $\sum_{i=1}^{n} X_i$ 是 $\sum_{i=1}^{n}\sigma_i^2$-次高斯的. 从而有下面的概率不等式

$$\max\{P(\sum_{i=1}^{n}(X_i-EX_i)\geqslant t), P(\sum_{i=1}^{n}(X_i-EX_i)\leqslant -t)\}\leqslant\exp(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2})$$

引理 1.4 支撑集为 $[0,1]$ 的随机变量是 $\frac{1}{2}$-次高斯的; 若 $X_1, ..., X_n$ 为 i.i.d. 的支撑集为 $[0,1]$, 均值为 μ 的随机变量, 则对任意 $t>0$, $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 满足:

$$\begin{aligned}P(\bar{X}-\mu>t)&\leqslant\exp(-2nt^2)\\P(\bar{X}-\mu<-t)&\leqslant\exp(-2nt^2)\\P(|\bar{X}-\mu|>t)&\leqslant 2\exp(-2nt^2)\end{aligned}\tag{1.4}$$

上面这个引理告诉我们, 对我们考虑的 K 个分布而言, 样本均值偏移真实均值的概率是指数下降的, 因此, 用户能够根据历史信息, 利用这个不等式对各个分布的均值建立起一个高概率的区间估计, 这是一个极其重要的观察。

二、信息论的编码理论

我们将要提出的算法框架中一个非常重要的组成成分是所谓的隐式通信, 它将利用碰撞 (collision) 的信息来完成隐式通信, 从而完成分布式下的信息交换。在可观察的情况下, 由于用户能够观察到碰撞的示性函数, 这对应于没有噪声的信道, 发送方发送比特 0 或者比特 1 时, 接收方将会以概率 1 接收到相应的比特, 因此此时误码率为 0; 在不可观察的 MPMAB 中, 由于此时用户无法观察到碰撞的示性函数, 它将无法分辨奖励 0 是由于碰撞引起的还是由于采样引起的, 而它总是能根据一个非 0 的奖励判断出不存在碰撞, 这对应于信道中存在噪声的情况, 此时误码率可能不为 0, 为了刻画这种情况, 我们介绍如图片 1.1所示的 Z-信道。幸

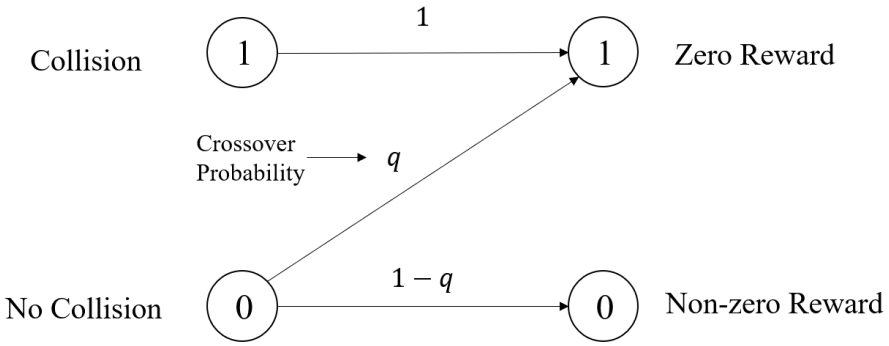


图 1.1　The Z-channel model

运的是, 根据香农定理, 当传输的码率小于信道的容量时, 对任意小的误码率上界我们都可以找到一个相对应的编码方案。对于 Z-信道, 我们有许多成熟的编码技巧, 例如最简单而又高效的是所谓的重复码: 为了传送一个信号"1/0", 我们将会重复地发送 N 个比特 $1/0$。我们会在讨论将可观察模型算法推广到不可观察情形时对编码理论做进一步的阐述。

三、理论下界: 我们能期待的最好结果

在这一小节中, 我们介绍多臂老虎机算法的理论遗憾下界。我们首先在单用户情形提供一个直觉, 如果记 K 个分布的均值满足 $\mu^* = \mu_{(1)} \geqslant \mu_{(2)} \geqslant \cdots \geqslant \mu_{(K)}$, 一个多臂老虎机问题的本质困难在于最优分布与次优分布均值之间的差, 记作 $\Delta_i = \mu^* - \mu_i$, 根据中心极限定理, 样本均值收敛到真实均值的速度在大样本情况下与高斯分布是类似的, 因此对于我们所使用的蒙特卡洛方法, 我们无法期待比高斯分布情况下更好的结果, 由于高斯分布的尾部概率是指数衰减的, 换言之, 直觉上我们不得不对次优分布进行 $\Omega(\log T)$ 的采样才能最终将它与最优分布区分开来。我们介绍来自 Bubeck et al.[34] 的结果:

定理 1.5 (单用户情形的理论下界)　考虑单用户的多臂老虎机问题, 固定决策轮数 T 与分布数 K, 我们考虑这样一个算法 $\mathcal{A}$, 对任何伯努利分布实例, $\mathcal{A}$ 采样所有 $\Delta_i > 0$ 的次优分布 i 的次数 $T_i(T)$ 满足: 对任何 $a > 0$ 满足 $\mathbb{E}T_i(T) = o(T^a)$,

那么, 对任何的伯努利分布实例, $\mathcal{A}$ 的算法遗憾满足:

$$\liminf_{T\to+\infty} \frac{R(T)}{\log T} \geqslant \sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathrm{kl}\left(\mu_i, \mu^*\right)}$$

其中 $kl(\mu_i, \mu^*) =: \mu_i \log \frac{\mu_i}{\mu^*} + (1-\mu_i) \log \frac{(1-\mu_i)}{(1-\mu^*)}$ 是两个伯努利分布的相对熵, $\mu_i = p_i$.

注意我们对算法的假设事实上是在说, $\mathcal{A}$ 在某种程度上是" 好的", 这是为了得到这个对任何伯努利实例都成立的较强的算法下界。事实上, 假设我们不做这样的一个假设, 一个始终采样第一个分布的算法, 在分布 1 是最优分布的情况下将会取得 0 损失, 因此在不做额外假设的情况下, 对于任何分布都成立的下界是不可能得到的。另一方面, 假设如果这个假设不成立, 我们总能找到一个问题实例, 使得算法采样次优分布的次数为 $\Omega(T^a)$ 对某个 $a > 0$ 成立, 此时它的遗憾要远远大于 $\Omega(\log T)$。此外, 同质化模型与非同质化模型有自然的中心化算法的理论下界: Anantharam et al.[35] 证明同质化的理论下界为 $\Omega(\sum_{k>M} \frac{\log T}{\mu_{(M)}-\mu_{(k)}})$; Kveton et al.[25] 中给出的非同质化模型的理论下界为 $\Omega\big(\frac{M^2K}{\Delta_{\min}} \log(T)\big)$。

第二章　MAB: 单机算法理论

在这一章中, 我们将介绍单用户的多臂老虎机模型中的两个著名的探索算法:`Successive Elimination` 与 `UCB1` (Upper Confidence Bound)。根据单用户算法的理论下界, 这两个算法在渐进情况下达到了最优的算法遗憾上界。概况地说, 这两个单用户算法共同的思想是, 在 t 时刻, 对 K 个分布维护 K 个同时置信区间 $[\bar{X}_k^t - r_k(t), \bar{X}_k^t + r_k(t)]$, K 个分布的均值同时落在各自维护的区间内的概率为 $1-\delta_t$, 其中, $\bar{X}_k^t$ 表示分布 k 在第 t 轮的样本均值, $r_k(t)$ 是对应的置信半径。此时, 我们用所维护的区间来近似替代真实的均值做出决策, 当采样逐渐进行, 这个区间估计将会越来越准, 从而我们决策的质量也会越来越高。此外, 一个重要的特点是, 探索 (exploration) 与守成 (exploitation) 在整个决策历史中是同时进行的, 探索也同样基于历史信息序列。

第一节　UCB1

`UCB1` 算法在每一轮中会选择具有最大置信上界的分布, 置信上界的形式为 $\bar{X}_j^t + r_j(t)$, 我们首先需要构造出上述介绍的 K 个置信区间从而获得置信上界的形式, 因此我们需要决定如何选择置信半径。我们记分布 j 在第 t 轮被采样的次数为 $T_j(t)$, 从直觉上讲, 我们在如下两种情况下会希望对某个分布进行采样

- 从探索出发, 我们希望被采样次数较少的分布更有可能被选择;
- 从守成出发, 我们希望具有更大样本均值的分布被选择。

因此, 具有大样本均值与探索次数 $T_j(t)$ 较小的分布将会具有更大的置信上界, 此外, 为了使得置信度的阶为 $\frac{1}{T}$, 我们最终选择 $r_j(t) = \sqrt{\frac{2\log T}{T_j(t)}}$, 完整的 `UCB1` 算法在伪代码 2中描述。

我们不加证明的给出如下的遗憾理论上界, 这个结果来自 Auer et al.[1]:

Algorithm 2 UCB

1: **Input:** K arms, T rounds, $r_j(t) = \sqrt{\frac{2\log T}{T_j(t)}}$

2: Try each arm once.

3: At each round, play arm with maximal $\bar{X}_j + r_j$

定理 2.1 `UCB1` 算法的理论遗憾上界。令 $\Delta_i := \mu^* - \mu_i$, 则:

$$\mathbb{E}R(T) \leqslant \left[8 \sum_{i:\Delta_i>0} \left(\frac{\log T}{\Delta_i}\right)\right] + \left(1 + \frac{\pi^2}{3}\right) \left(\sum_{j=1}^{K} \Delta_j\right)$$

注 我们同样可以替换置信半径中的 $\log T$ 为 $\log t$, 这将使得我们不需要知道游戏的总轮数 T, 其代价为一个更加复杂的证明过程。

第二节 Successive Elimination

这一小节给出 `Successive Elimination` 算法, 它与 `UCB1` 算法一样, 对每个分布维护一个区间估计, $[\bar{X}_j - r_j(t), \bar{X}_j + r_j(t)], j \in [K]$, 由于每个分布的真实均值以大概率落在对应的置信区间里, 如果两个分布的置信区间不相交, 我们就可以以大概率判断, 区间在实轴左侧的分布的均值要比区间在右侧的分布低, 因此我们可以" 删除" 这个分布, 这个想法引出了我们的第二个算法, 它在伪代码 3中给出。

Algorithm 3 Successive Elimination

1: **Input:** K arms, T rounds, $r_j(t) = \sqrt{\frac{2\log T}{T_j(t)}}$

2: Try all active arms;

3: Deactivate all arms k s.t. $\exists k' \in [K], UCB(k) < LCB(k')$.

我们同样不加证明的给出如下的遗憾理论上界, 这个结果来自 Slivkins[36]:

定理 2.2 `Successive Elimination` 算法的理论遗憾上界

$$\mathbb{E}R(T) \leqslant O\left(\sum_{i:\mu_i<\mu^*} \left(\frac{\log T}{\Delta_i}\right)\right)$$

第三章　MPMAB: 同质情形

在这一章中, 我们在可观察情况下对同质情形的多用户多臂老虎机问题进行分析, 并给出我们算法框架在同质化模型下的实现, 并证明它是渐进最优的。

第一节　算法发展

一、符号

除了在问题描述部分定义的符号, 在同质模型中我们可以使用更加简单的符号, 考虑从 K 个对应于均值 $(\mu_1, ..., \mu_K)$ 的分布中有放回地取出 M 个的所有可能性构成的集合: $\Gamma = \{\pi | \pi \subseteq [K], |\pi| = M\}$, $\pi = \{\pi_1, ..., \pi_M\}$, 特别的, 我们这里认为 π 是无序的, 这是由于在同质化模型下, 分布对不同用户是相同的, 从群体收益的角度出现, 用户内部的次序是无关紧要的。此外, $V(\pi) = \sum_{m=1}^{M} \mu_{\pi_m}$ 是组合 π 的功效函数。算法的理论遗憾此时被定义为:

$$R_{hm}(T) := TV_* - \mathbb{E}\left[\sum_{t=1}^{T}\sum_{m=1}^{M} r_m(t)\right]$$

这里 $r_m(t)$ 是用户 m 第 t 轮获得的采样奖励, 其中 $V_* := \max_{\pi\in\Gamma} V(\pi)$ 是最优组合的功效函数。除此之外, 我们定义最优分布与次优分布的概念:

定义 3.1　令 σ 为 $\{1, 2, ..., K\}$ 的一个排列, $0 \leqslant l < M$ 且 $M \leqslant n \leqslant K$ 使得 $\mu_{\sigma(1)} \geqslant ... \geqslant \mu_{\sigma(l)} > \mu_{\sigma(l+1)} = ... = \mu_{\sigma(M)} = ... = \mu_{\sigma(n)} > \mu_{\sigma(n+1)} \geqslant ... \geqslant \mu_{\sigma(K)}$。其中分布 $\sigma(1), ..., \sigma(l)$ 被定义为 *distinctly* M*-best* 分布, 分布 $\sigma(n+1), ..., \sigma(K)$ 被定义为 *distinctly* M*-worst* 分布, 最后, 均值等于 $\mu_{\sigma(M)}$ 的分布均值被定义为M*-border* 分布。

简而言之, 我们将 K 个分布按照均值大小进行排序, 之后, 将它们按照均值次序分为三类, 其中 *distinctly* M*-best* 是属于最优组合的 $l < M$ 个分布, *distinctly*

M-worst 是不属于最优组合的 $K-n$ 个分布, 最后剩下的 $n-l$ 个分布拥有相同的均值, 我们从中任意选取 $M-l$ 个分布加入 *distinctly M-best* 分布后就得到了一个最优分布组合。我们强调, 当 $n-l>1$, 也就是最优分布组合不唯一, 此时它将在分布式情况下起到类似于 `MAB` 中 Δ 的作用, 用户将会为了区分这不同的分布组合做出大量无用的努力, 特别的, 由于此时的算法损失不仅仅来源于探索, 它会在分布式情况下带来区别于单机情况的本质困难。

此外, 我们记 K_p 与 M_p 为第 p 个阶段仍然活跃的分布数与用户数, $\mathcal{E}$ 与 $\mathcal{M}_p$ 是对应的仍然活跃的分布与用户集合, 其中活跃指的是此分布还没有被消除, 以及此用户仍然参与到探索中; 此外, $\mathcal{A}_p$ 与 $\mathcal{B}_p$ 分别代表这一阶段与拒绝的分布集合; E_m^p 是用户 m 在第 p 个探索阶段要采样的分布。

二、相关算法

第一个阶段: 单用户算法与碰撞避免机制

在上一章中，我们介绍了单用户多臂老虎机问题中的成熟算法，一个自然的想法是，将单用户的算法推广到多用户的情形。但是如果 M 个用户都采取成熟的单用户算法, 在同质情况下, 他们将会很快聚集在最优的几个分布上, 从而可能产生大量的冲突，严重影响算法表现。因此, 算法必须要引入一定的碰撞避免机制, Rosenski et al.[37] 提出的著名的 `Musical Chair` 的理论算法遗憾为 $O\left(\frac{MK\log(T)}{\Delta^2_{\sigma(M+1)}}\right)$。可以看到, 与中心化算法相比, 由于用户之间并没有直接或间接地在探索采样上进行合作或者交换信息，由探索次优分布导致的损失此时会多一个常数因子 MK, 这是并行执行 M 个单机算法所带来的求和导致的。

第二个阶段: 利用少量碰撞传递信息

Boursier et al.[4] 创造性将碰撞视作一种可能的信号传递方式，提出算法 `SIC-MMAB`, 其以少量刻意产生的碰撞为代价，使得各个用户实现合作。其基本想法是, 通过预先设置好的协议, 不同用户在约定的轮次内按照约定好的方式

进行信息交换: 碰撞代表比特 1, 不碰撞代表比特 0。其交换的信息包括采样得到的样本均值, 以及部分用于在接下来的探索阶段使得各个用户彼此不会发生冲突的信息。在假定最优分布组合唯一的情况下, SIC-MMAB 可以达到的理论算法遗憾为

$$O\left(\sum_{k>M}\frac{\log(T)}{\Delta_{\sigma(k)}}+KM^3\log^2\left(\frac{\log(T)}{\Delta^2_{\sigma(M+1)}}\right)\right).$$

可以看到, SIC-MMAB 成功消除了第一项探索损失前边由于分布式带来的常数因子, 代价是第二项一个更低阶的由于通信带来的损失。

第三个阶段: 高效的信息传递框架

SIC-MMAB 的原始论文只提供了上述理论结果, 它隐式的假定了只有唯一的最优分布组合, 这是由于当有多个最优分布组合的时候, 通信将会在整个游戏过程中进行, 这会导致通信损失最终作为损失的高阶主项, 具体的说, 我们有一般性的结果:

$$O\left(\sum_{k>M}\frac{\log(T)}{\Delta_{\sigma(k)}}+\frac{(n-M)\log(T)}{\Delta_{\sigma(l)}}+KM^3\log^2(T)\right).$$

其本质原因是, 当通信贯穿整个游戏进程, 我们总共需要进行 $O(\log T)$ 次的通信, 而每次通信所需要的长度将逐渐增长到 $O(\log T)$, 这是由于我们需要用更多的比特来表示样本均值, 从而获得更高的精度, 最终导致了通信损失对 $\log T$ 的二次依赖。

三、一般框架结构

我们在这一小节提出我们的框架 BEACON-HM (Batched Exploration with Adaptive COmmunicatioN), 它是针对 SIC-MMAB 算法的改进, 因此也将分为初始化、探索、通信、守成四个阶段, 用户完成初始化阶段后, 将同步地在探索与通信不断切换, 我们称一个探索与一个通信为一个探索-通信阶段对。若在某一次通信中达到终止条件, 则一同进入守成阶段, 探索与通信将持续到游戏结束。在本章中, 我们给出的第一个 BEACON-HM 是在同质模型下基于 Successive

Algorithm 4 BEACON-HM

Input: $T,\ K$;
Initialize $p \leftarrow 1$; $F \leftarrow -1$
`Initialization Phase:`
$(M, m) \leftarrow$ Init(); $E_m^1 \leftarrow (m, m+1, ..., K, 1, ..., m-1)$
$\mathcal{E} \leftarrow [K]$; $\mathcal{A}, \mathcal{B} \leftarrow \emptyset$
while $F = -1$ **do**
    `Exploration Phase:`
    Pull each arm in E_m^p in order for $\lceil 2^p \log(T) \rceil$ times
    Update empirical sample mean $\hat{\mu}_{k,m}^p, \forall k \in [K]$
    `Communication Phase:`
    if $m = 1$ **then**
        (C^{p+1}, E^{p+1}) $\leftarrow$ CommHMLeader()
    else
        (C_1^{p+1}, C_m^{p+1}, E_m^{p+1}) $\leftarrow$ CommHMFollower()
    end if
    if $|E_m^{p+1}| = 1$ **then** $F \leftarrow$ the only element in E_m^{p+1}
    end if
    $p \leftarrow p + 1$
end while
`Exploitation phase:` Pull F until T

`Elimination` 的实现。

初始化 (Initialization)。在这个阶段, 每个用户将会独立地估计出用户数 M 并得到一个指定的序号 $m \in [M]$, 我们将使用来自 Wang et al.[9] 的初始化技巧, 其期望损失不超过 $\frac{K^2M^2}{K-M} + 2KM$。M 的估计与序号的获得对接下来有序的探索与通信至关重要。

探索 (Exploration)。我们假定在第 p 个探索-通信阶段对开始前, 用户将维

护有相同的活跃用户与活跃分布集合 $\mathcal{E}$ 与 $\mathcal{M}_p$。此时, 序号为 m 的用户将从集合 $\mathcal{E}$ 的第 m 个分布开始采样, 由于每个用户的序号 m 不同, 第一次采样将不会产生碰撞, 此后, M 个用户同步地向后采样一个分布, 其中, $K+1$ 将视作回到第一个分布。由此用户将在无碰撞的情况下完成对分布的探索, 探索的长度为 $K_p \times 2^p \log T$, 因此每个用户会采样每个 $\mathcal{E}$ 中的分布 $2^p \log T$ 次。

通信 (Communication)。根据探索部分的讨论, 我们需要传输的内容要能够使得各个用户维护有相同的 $\mathcal{E}$ 与 $\mathcal{M}_p$。概况的说, 用户 1 将自动成为 *Leader* 而其余用户将成为 *Follower*, 每个 *Follower* 将会将用有限比特量化的样本均值发送给 *Leader*, 而 *Leader* 将根据多用户情形的消除分布规则, 给出接受的分布与要消除的分布, 并回传给 *Follower*。*Follower* 将按照" 序号大的先进入守成" 的规则守成于接受的分布, 这样每个 *Follower* 能够根据回传的 $\mathcal{A}_p$ 的长度判断哪些用户进入守成从而维护相同的 $\mathcal{E}$ 与 $\mathcal{M}_p$。特别要强调的是, 即使是进入守成的用户也会参与之后的通信, 这是因为我们此时需要不断减小的量化损失。我们在下一小节具体介绍 `BEACON` 针对 `SIC-MMAB` 的主要改进: 自适应差分通信 (Adaptive Differential Communication), 这是我们能够摆脱通信损失作为主项的原因。特别的, 为了减小通信损失, 我们还会传输当前的经验最优分布组合 C_i^{p+1} 作为下一轮用以通信的分布。

守成 (Exploitation)。对于基于 `Successive Elimination` 的实现, 当用户消除了所有次优的分布, 便会显式的进入守成阶段: 所有用户将固定在仍然未被消除的分布组合上直到游戏结束。

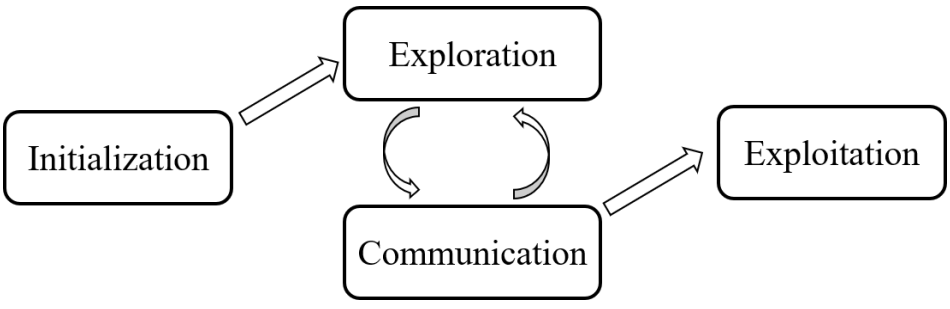


图 3.1　基于 `Successive Elimination` 的 `BEACON-HM` 框架。

四、自适应差分通信 (Adaptive Differential Communication)

我们将在这一小节在全感知模型与 `Successive Elimination` 实现下介绍 `BEACON-HM` 的通信方案。

Algorithm 5 CommHM (ADC-HM)

CommHMLeader

`Gather information from followers:`

$\forall m \in (2, ..., M), \forall k \in \mathcal{E}$, Receive($C_m^p$, the last L_{hm}^p bits of $\delta_{k,m}^p$), $\tilde{\mu}_{k,m}^p \leftarrow \tilde{\mu}_{k,m}^{p-1} + \delta_{k,m}^p$

$\forall k \in \mathcal{E}, \tilde{\mu}_k^p \leftarrow \sum_{m=1}^M \tilde{\mu}_{k,m}^p T_m^p / T^p$ and update $\mathcal{A}_p$ and $\mathcal{B}_p$ as in Eqn. (3.1)

`Assign arms to followers:`

$\forall m \in [M], C_m^{p+1} \leftarrow$ empirically m-th best arm in $\mathcal{A} \cup \mathcal{E}$

$\forall m \in (2, ..., M)$, Send($C_m^p$, $\{|\mathcal{A}_p|, |\mathcal{B}_p|\}$) and Send($C_m^p$, $\{\mathcal{A}_p, \mathcal{B}_p, C_m^{p+1}, C_1^{p+1}\}$)

$\mathcal{E} \leftarrow \mathcal{E} \backslash \mathcal{A}_p \cup \mathcal{B}_p$; $\mathcal{A} \leftarrow \mathcal{A} \cup \mathcal{A}_p$; $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{B}_p$; $M_{p+1} \leftarrow M - |\mathcal{A}|$; $K_{p+1} \leftarrow |\mathcal{E}|$

if $M \leqslant |\mathcal{A}|$ **then** $E_1^{p+1} \leftarrow (\mathcal{A}[M])$

else $E_1^{p+1} \leftarrow (\mathcal{E}[1], ...\mathcal{E}[K_{p+1}])$

end if

CommHMFollower

`Send information to the Leader:`

$\tilde{\mu}_{k,m}^p \leftarrow$ quantize $\hat{\mu}_{k,m}^p$ by Q_{hm}^p bits and $\delta_{k,m}^p \leftarrow \tilde{\mu}_{k,m}^p - \tilde{\mu}_{k,m}^{p-1}$

$\forall k \in \mathcal{E}$, Send($C_1^p$, the last L_{hm}^p bits of $\delta_{k,m}^p$)

`Receive arm assignment from the Leader:`

Receive(C_1^p, $\{|\mathcal{A}_p|, |\mathcal{B}_p|\}$) and Receive($C_1^p$, $\{\mathcal{A}_p, \mathcal{B}_p, C_m^{p+1}, C_1^{p+1}\}$)

$\mathcal{E} \leftarrow \mathcal{E} \backslash \mathcal{A}_p \cup \mathcal{B}_p$; $\mathcal{A} \leftarrow \mathcal{A} \cup \mathcal{A}_p$; $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{B}_p$ and $M_{p+1} \leftarrow M - |\mathcal{A}|$; $K_{p+1} \leftarrow |\mathcal{E}|$

if $M - m + 1 \leqslant |\mathcal{A}|$ **then** $E_m^{p+1} \leftarrow (\mathcal{A}[M - m + 1]) \times K_{p+1}$

else $E_m^{p+1} \leftarrow (\mathcal{E}[m], ..., \mathcal{E}[K_{p+1}], \mathcal{E}[1], ..., \mathcal{E}[m-1])$

end if

比特传输。注意到，在全感知模型下，每一个用户可以感知到她是否与其他用户产生了碰撞，这个示性函数可以看成是一个 0-1 比特。用户 m 可以使用 $\mathcal{E}$

中的第 m 个分布作为它的通信分布, 例如当用户 1 向用户 2 发送信息, 如果用户 1 采样第 2 个分布, 会导致二者发送碰撞, 这就代表比特 1, 反之代表比特 0。更一般的, *Leader* 可以选择当前经验最优的分布组合发送给 *Follower* 作为通信分布。

自适应差分通信: 样本均值的量化长度。首先, *Follower* 将会依次向 *Leader* 发送样本均值, 而实数必须截断量化为有限的 Q^p 比特才能完成传输。我们将选择 Q^p 以使得由量化的误差的阶和我们维护的区间估计半径相同, 从而将量化噪声作为抽样的不确定性进行处理。此时的置信半径为 $\epsilon^p = \sqrt{\frac{\log(T)}{2T^p}}$, 其中 $T^p = \sum_{q=1}^{p} M_q 2^q \lceil \log(T) \rceil$ 是在第 p 轮探索结束时, 分布被采样的总次数, T_m^p 则是用户 m 采样分布的总次数。每个样本均值 $\hat{\mu}_{k,m}^p$ 将会被截断至 $Q^p = \lceil \log_2(1/\epsilon^p) \rceil$ 比特, 真正被传输的是截断后的 $\tilde{\mu}_{k,m}^p$。

自适应差分通信: 差分通信。可以看到 Q^p 最终将会与 $\log T$ 同阶, 由于通信最多可能有 $O(\log T)$ 轮, 我们将会面临 $O(\log^2 T)$ 的通信损失。为了克服这个问题, 根据集中不等式, 我们注意到 $\tilde{\mu}_{k,m}^{p+1}$ 与 $\tilde{\mu}_{k,m}^p$ 在数值上接近以大概率成立, 从信息的角度而言, 后者中蕴含了相当一部分的前者的信息, 因此反复传送整个样本均值/样本总奖励是没有必要的。我们提出自适应差分通信 (Adaptive Differential Communication), 在每次通信中传输二者之间的差值。首先, 由于 $\tilde{\mu}_{k,m}^{p+1}$ 与 $\tilde{\mu}_{k,m}^{p+1}$ 分别由 Q^{p+1}, Q^p 比特表示, 因此它们的差值 $\delta_{k,m}^{p+1}$ 只需要不超过 Q^{p+1} 比特来表示; 此外, 假定 $\tilde{\mu}_{k,m}^p \in [\mu_k - \epsilon^p, \mu_k + \epsilon^p]$ 与 $\tilde{\mu}_{k,m}^{p+1} \in [\mu_k - \epsilon^{p+1}, \mu_k + \epsilon^{p+1}]$ 同时成立, 此时 $|\delta_{k,m}^{p+1}| = |\tilde{\mu}_{k,m}^{p+1} - \tilde{\mu}_{k,m}^p| < \epsilon^p + \epsilon^{p+1}$。因此, $\delta_{k,m}^p$ 以大概率会很小, 我们只需要传输最后部分的非 0 比特即可。我们会证明, 只有最后的 $L^p = O(1)$ 个比特是非 0 比特以大概率成立, 加上一个符号位后, 此时一个样本均值的通信代价为 $O(1)$ 比特。

***Leader* 消除分布的准则**。令 $\mathcal{E}$ 为本轮仍然活跃的分布的集合, 则 *Leader* 在接收到其余 *Follower* 的量化样本均值后, 将会决定要接受的分布集合 $\mathcal{A}_p$ 与要拒

绝的分布集合 $\mathcal{B}_p$, 其准则为:

$$\begin{aligned}\mathcal{A}_p &= \left\{k \in \mathcal{E} : |\{j \in \mathcal{E} | \tilde{\mu}_k^p - \tilde{\mu}_j^p \geqslant 4\epsilon_{hm}^p\}| \geqslant K_p - M_p\right\};\\ \mathcal{B}_p &= \left\{k \in \mathcal{E} : |\{j \in \mathcal{E} | \tilde{\mu}_j^p - \tilde{\mu}_k^p \geqslant 4\epsilon_{hm}^p\}| \geqslant M_p\right\}.\end{aligned} \tag{3.1}$$

之后, *Leader* 首先向每个 *Follower* 发送两个集合的长度, 并随之发送两个集合, 之后仍然活跃的用户中序号最大的 $|\mathcal{A}_p|$ 个用户将会进入守成阶段。因此, 用户将维护相同的活跃分布集合与活跃用户集合, 这使得之后的探索与通信阶段能够继续同步进行。

第二节　理论结果

在这一小节中, 我们将给并证明我们的理论结果:

一、主要结果

定理 3.1　记 $\Delta_{\sigma(k)} = \mu_{\sigma(M)} - \mu_{\sigma(k)}, \forall k > M$ 以及 $\Delta_{\sigma(l)} = \mu_{\sigma(l)} - \mu_{\sigma(M+1)}$, 则 BEACON-HM 的理论遗憾上界为:

$$\begin{aligned}R_{hm}(T) \leqslant 22M\log_2(M)K\log(T) + \sum_{k>n}\frac{96(2+\sqrt{3})\log(T)}{\Delta_{\sigma(k)}}\\ + 96(1+\sqrt{3})(n-M)\frac{\log(T)}{\Delta_{\sigma(l)}} + o\left(\log(T)\right).\end{aligned}$$

特别的, 当最优分布组合唯一, 也就是 $n = M$, 我们有

$$R_{hm,s}(T) \leqslant \sum\nolimits_{k>M}\frac{96\left(2+\sqrt{3}\right)\log(T)}{\Delta_{\sigma(k)}} + o\left(\log(T)\right). \tag{3.2}$$

我们在表 3.1中将现存结果与本文中的结果进行对比, 可以看到, BEACON-HM 是第一个能够在不做任何额外假设的情况下保持 $O(\log T)$ 阶损失的算法。

二、理论分析

BEACON 的理论遗憾可以分为三部分: 由初始化导致的损失, 由探索次优分布组合导致的损失, 以及由通信导致的损失。我们会对每一项进行分析, 并最终

表 3.1　同质化模型下算法理论遗憾上界汇总。

Homogeneous Setting		
Assumptions	Algorithm (Reference)	Asymptotic Upper Bound
$\Delta_{\sigma(M+1)} > 0$ and is known	Musical Chair [37]	$O\left(\frac{MK\log(T)}{\Delta^2_{\sigma(M+1)}}\right)$
$\mu_{\sigma(M)} > \mu_{\sigma(M+1)}$	SIC-MMAB [4]	$O\left(\sum_{k>M}\frac{\log(T)}{\Delta_{\sigma(k)}} + KM^3\log^2\left(\frac{\log(T)}{\Delta^2_{\sigma(M+1)}}\right)\right)$
($\star$)	SIC-MMAB [4]	$O\left(\sum_{k>M}\frac{\log(T)}{\Delta_{\sigma(k)}} + \frac{(n-M)\log(T)}{\Delta_{\sigma(l)}} + KM^3\log^2(T)\right)$
$\mu_{\sigma(1)} > \ldots > \mu_{\sigma(M)} > \mu_{\sigma(M+1)} > \ldots > \mu_{\sigma(K)}$	DPE1 [9]	$O\left(\sum_{k>M}\frac{\Delta_{\sigma(k)}\log(T)}{\mathrm{kl}(\mu_{\sigma(k)}+\delta,\mu_{\sigma(M)}-\delta)} + K^{5/2}M^3\delta^{-2}\right)$
-	this work	$O\left(\sum_{k>n}\frac{\log(T)}{\Delta_{\sigma(k)}} + \frac{(n-M)\log(T)}{\Delta_{\sigma(l)}} + KM\log(M)\log(T)\right)$

Homogeneous Setting: $0 < \delta < \min_{1\leqslant k\leqslant K-1}\frac{\mu_{\sigma(k)}-\mu_{\sigma(k+1)}}{2}; \forall k > M, \Delta_{\sigma(k)} = \mu_{\sigma(M)} - \mu_{\sigma(k)}$, $\Delta_{\sigma(l)} = \mu_{\sigma(l)} - \mu_{\sigma(M+1)}$;

($\star$) 在原始论文中没有给出但可以经过简单推导得到。

通过简单的计算得到总的算法遗憾上界。首先, 我们给出初始化阶段的损失上界, 这个结果来自 Wang et al.[9]。

引理 3.2　BEACON-HM 与 BEACON-HT 的初始化损失有上界:

$$R^{init}_{hm/ht}(T) \leqslant M\left(\frac{K^2M}{K-M} + 2K\right),$$

且在初始化完成后, 所有用户正确地估计出用户数 M 并获得一个唯一的属于 $\{1, 2, ..., M\}$ 中的次序。

之后, 我们证明我们只需要 $L^p_{hm} = O(1)$ 比特进行量化均值差值传输。

引理 3.3　$P\left(L^p_{hm} \leqslant 5 + \frac{1}{2}\log_2(M)\right) \geqslant 1 - \frac{2}{T^2}$.

证明　我们首先对差值求取如下上界:

$$\begin{aligned}|\delta^p_{k,m}| &= |\tilde{\mu}^p_{k,m} - \tilde{\mu}^{p-1}_{k,m}| \\ &\leqslant |\tilde{\mu}^p_{k,m} - \hat{\mu}^p_{k,m} - (\tilde{\mu}^{p-1}_{k,m} - \hat{\mu}^{p-1}_{k,m})| + |\hat{\mu}^p_{k,m} - \hat{\mu}^{p-1}_{k,m}|\end{aligned}$$

$$\begin{aligned}
&\leqslant |\tilde{\mu}_{k,m}^{p} - \hat{\mu}_{k,m}^{p}| + |\tilde{\mu}_{k,m}^{p-1} - \hat{\mu}_{k,m}^{p-1}| + |\hat{\mu}_{k,m}^{p} - \hat{\mu}_{k,m}^{p-1}| \\
&\overset{(i)}{\leqslant} \epsilon_{hm}^{p} + \epsilon_{hm}^{p-1} + |\hat{\mu}_{k,m}^{p} - \hat{\mu}_{k,m}^{p-1}|,
\end{aligned}$$

其中不等式 (i) 是因为在 $Q_{hm}^{p} = \lceil \log(1/\epsilon_{hm}^{p}) \rceil$ 条件下, $|\tilde{\mu}_{k,m}^{p} - \hat{\mu}_{k,m}^{p}| \leqslant \epsilon_{hm}^{p}$。此外, 如果用户 m 在 p 轮没有固定在某个分布上, 我们有:

$$\begin{aligned}
|\hat{\mu}_{k,m}^{p} - \hat{\mu}_{k,m}^{p-1}| = |\hat{\mu}_{k,m}^{p} - \hat{\mu}_{k,m}^{p-1}| &= \left| \frac{\sum_{\tau=1}^{T_m^p} \gamma_{k,m}(\tau)}{T_m^p} - \frac{\sum_{\tau=1}^{T_m^{p-1}} \gamma_{k,m}(\tau)}{T_m^{p-1}} \right| \\
&= \left| \frac{\sum_{\tau=1}^{T_m^{p-1}} \gamma_{k,m}(\tau) + \sum_{\tau=T_m^{p-1}+1}^{T_m^p} \gamma_{k,m}(\tau)}{T_m^{p-1} + T_m^p - T_m^{p-1}} - \frac{\sum_{\tau=1}^{T_m^{p-1}} \gamma_{k,m}(\tau)}{T_m^{p-1}} \right| \\
&\leqslant \left| \frac{\sum_{\tau=T_m^{p-1}+1}^{T_m^p} \gamma_{k,m}(\tau)}{T_m^p - T_m^{p-1}} - \frac{\sum_{\tau=1}^{T_m^{p-1}} \gamma_{k,m}(\tau)}{T_m^{p-1}} \right|,
\end{aligned}$$

其中 $\gamma_{k,m}(\tau)$ 是用户 m 在第 τ 次采样分布 k 获得的奖励。进一步有:

$$\begin{aligned}
&P\left(|\hat{\mu}_{k,m}^{p} - \hat{\mu}_{k,m}^{p-1}| \geqslant \sqrt{\frac{2\log(T)}{T_m^{p-1}}} \right) \\
\leqslant &P\left(\left| \frac{\sum_{\tau=T_m^{p-1}+1}^{T_m^p} \gamma_{k,m}(\tau)}{T_m^p - T_m^{p-1}} - \frac{\sum_{\tau=1}^{T_m^{p-1}} \gamma_{k,m}(\tau)}{T_m^{p-1}} \right| \geqslant \sqrt{\frac{2\log(T)}{T_m^{p-1}}} \right) \\
\leqslant &2\exp\left(-\frac{4\frac{\log(T)}{T_m^{p-1}}}{\frac{1}{T_m^p - T_m^{p-1}} + \frac{1}{T_m^{p-1}}} \right) \\
\leqslant &2\exp\left(-2T_m^{p-1}\frac{\log(T)}{T_m^{p-1}} \right) = \frac{2}{T^2}.
\end{aligned}$$

而在阶段 p 固定在某分布的用户不再参与探索, 则 $|\hat{\mu}_{k,m}^{p} - \hat{\mu}_{k,m}^{p-1}| = 0$, 这同样满足上述的不等式。因此, 以概率 $1 - \frac{2}{T^2}$, $\delta_{k,m}^{p}$ 的最末端至多有 L_{hm}^{p} 个非 0 比特, 我们做如下计算:

$$\begin{aligned}
L_{hm}^{p} &\leqslant \lceil \log_2(1/\epsilon_{hm}^{p}) \rceil - \left\lfloor \log_2\left(1/\left(\epsilon_{hm}^{p} + \epsilon_{hm}^{p-1} + \sqrt{\frac{2\log(T)}{T_m^{p-1}}} \right) \right) \right\rfloor \\
&\leqslant 2 + \log_2\left(1 + \frac{\epsilon_{hm}^{p-1}}{\epsilon_{hm}^{p}} + \frac{\sqrt{\frac{2\log(T)}{T_m^{p-1}}}}{\epsilon_{hm}^{p}} \right) \\
&\leqslant 2 + \log_2\left(1 + \sqrt{3} + 2\sqrt{\frac{T^p}{T_m^{p-1}}} \right)
\end{aligned}$$

$$\leqslant 2+\log_2\left(1+\sqrt{3}+2\sqrt{M}\right)$$
$$\leqslant 5+\frac{1}{2}\lceil\log_2(M)\rceil.$$

□

之后, 我们同样利用集中不等式证明, 真实均值落在我们维护的区间估计之中以大概率成立。

引理 3.4 对任何阶段数 p 以及任何仍然活跃的分布 $k\in\mathcal{E}$, 我们有

$$P\left(|\tilde{\mu}_k^p-\mu_k|\geqslant 2\epsilon_{hm}^p\right)\leqslant\frac{2}{T}.$$

如果定义事件 $F_1=\left\{\forall p,\forall k,|\tilde{\mu}_k^p-\mu_k|\leqslant 2\epsilon_{hm}^p\forall m, \text{有少于 } L_{hm}^p \text{ 非 } 0 \text{ 比特于 } \delta_{k,m}^p\right\}$, 我们可以得到

$$P(F_1)\geqslant 1-\frac{2K\log_2(T)}{T}-\frac{2MK\log(T)}{T^2}.$$

证明 我们有

$$\begin{aligned}P\left(|\tilde{\mu}_k^p-\mu_k|\geqslant 2\epsilon_{hm}^p\right)&=P\left(|\tilde{\mu}_k^p-\hat{\mu}_k^p+\hat{\mu}_k^p-\mu_k|\geqslant 2\epsilon_{hm}^p\right)\\&\leqslant P\left(|\tilde{\mu}_k^p-\hat{\mu}_k^p|+|\hat{\mu}_k^p-\mu_k|\geqslant 2\epsilon_{hm}^p\right)\\&\overset{(i)}{\leqslant} P\left(|\hat{\mu}_k^p-\mu_k|\geqslant\epsilon_{hm}^p\right)\\&\overset{(ii)}{\leqslant} 2\exp(-2T^p(\epsilon_{hm}^p)^2)\\&=\frac{2}{T},\end{aligned}$$

其中不等式 (i) 是因为

$$|\tilde{\mu}_k^p-\hat{\mu}_k^p|=\frac{\left|\sum_{m=1}^M(\tilde{\mu}_{k,m}^p-\hat{\mu}_{k,m})T_m^p\right|}{T^p}\leqslant\epsilon_{hm}^p,$$

不等式 (ii) 是因为 Hoeffding's 不等式, 利用 Union Bound, 我们可以获得事件 F_1 的结果。 □

由此, 我们可以利用条件均值将分析分为 F_1 成立与 F_1 不成立两种情况, 在 F_1 不成立的情况下可以直接得到一个 $O(1)$ 的遗憾上界, 而在 F_1 成立的情况下, 我们分别分析由初始化, 探索, 通信带来的期望遗憾。

三、探索引发的损失

首先, 我们证明探索损失满足如下的分解关系:

引理 3.5 假定探索与守成的总轮数为 T^{expl}, 则我们可以做如下分解

$$R_{hm}^{expl}(T) = \sum_{k:M\text{-best}} \left(\mu_k - \mu_{\sigma(M)}\right) \left(T^{expl} - T_k^{expl}\right) + \sum_{k:M\text{-worst}} \left(\mu_{\sigma(M)} - \mu_k\right) T_k^{expl}, \tag{3.3}$$

其中 T_k^{expl} 是所有用户在探索守成期间对分布 k 的总采样次数。

证明 为表达简便起见, 在这个证明里, T^{expl} 与 T_k^{expl} 将被 T 与 T_k 表示, 则

$$\begin{aligned}
&R_{hm}^{expl}(T)\\
=&T\sum_{k=1}^{M}\mu_{\sigma(k)} - \sum_{k=1}^{K}\mu_{\sigma(k)}T_{\sigma(k)}\\
=&\sum_{k=1}^{l}\mu_{\sigma(k)}T + \mu_{\sigma(M)}\sum_{k=l+1}^{M}T - \sum_{k=1}^{l}\mu_{\sigma(k)}T_{\sigma(k)} - \mu_{\sigma(M)}\sum_{k=l+1}^{M}T_{\sigma(k)} - \mu_{\sigma(M)}\sum_{k=M+1}^{n}T_{\sigma(k)} - \sum_{k=n+1}^{K}\mu_{\sigma(k)}T_{\sigma(k)}\\
=&\sum_{k=1}^{l}\mu_{\sigma(k)}\left(T - T_{\sigma(k)}\right) + \sum_{k=n+1}^{K}\left(\mu_{\sigma(M)} - \mu_{\sigma(k)}\right)T_{\sigma(k)} + \mu_{\sigma(M)}\left[\sum_{k=l+1}^{M}\left(T - T_{\sigma(k)}\right) - \sum_{k=M+1}^{K}T_{\sigma(k)}\right]\\
=&\sum_{k=1}^{l}\left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right)\left(T - T_{\sigma(k)}\right) + \sum_{k=n+1}^{K}\left(\mu_{\sigma(M)} - \mu_{\sigma(k)}\right)T_{\sigma(k)} + \mu_{\sigma(M)}\left[\sum_{k=1}^{M}\left(T - T_{\sigma(k)}\right) - \sum_{k=M+1}^{K}T_{\sigma(k)}\right]\\
\overset{(i)}{=}&\sum_{k=1}^{l}\left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right)\left(T - T_{\sigma(k)}\right) + \sum_{k=n+1}^{K}\left(\mu_{\sigma(M)} - \mu_{\sigma(k)}\right)T_{\sigma(k)}\\
=&\sum_{k:M\text{-best}}\left(\mu_k - \mu_{\sigma(M)}\right)\left(T - T_k\right) + \sum_{k:M\text{-worst}}\left(\mu_{\sigma(M)} - \mu_k\right)T_k,
\end{aligned}$$

其中等式 (i) 是因为 $MT = \sum_{k=1}^{M} T_k$。 □

根据上面的结果, 我们只需要求取对次优分布采样的次数上界, 就可以得到探索损失的上界, 我们给出如下引理:

引理 3.6 在 F_1 发生的条件下, distinct M-best 分布 $\sigma(k), k \leqslant l$ 被接受 (也就是在集合 $\mathcal{A}$ 中), 至多只需要 $\frac{96\log(T)}{(\mu_{\sigma(k)} - \mu_{\sigma(M+1)})^2} + 3$ 次采样; distinct M-worst 分布

$\sigma(k), k > n$ 被拒绝 (也就是在集合 $\mathcal{B}$ 中), 至多只需要 $\frac{96\log(T)}{(\mu_{\sigma(M)}-\mu_{\sigma(k)})^2}+3$ 次采样。

证明 我们记 $\Delta_{\sigma(k)} = \mu_{\sigma(k)} - \mu_{\sigma(M+1)}$; 再记 $p_{\sigma(k)}$ 为满足 $T^{p_\sigma(k)} \geqslant s_{\sigma(k)}$ 的最小正整数, 其中 $8\sqrt{\frac{\log(T)}{2s_{\sigma(k)}}} \leqslant \Delta_{\sigma(k)}$。因此, 当事件 F_1 成立, 我们有 $\forall p \geqslant p_{\sigma(k)}, \forall j \in \{j : j \geqslant M+1, \sigma(j) \in \mathcal{E}\}$,

$$\begin{aligned}
\tilde{\mu}^p_{\sigma(k)} - 2\epsilon^p_{hm} &\geqslant \mu_{\sigma(k)} - 4\epsilon^p_{hm} \\
&\geqslant \mu_{\sigma(k)} - \mu_{\sigma(M+1)} + \mu_{\sigma(j)} - 4\epsilon^p_{hm} \\
&\geqslant \Delta_{\sigma(k)} - 8\epsilon^p_{hm} + \mu^p_{\sigma(j)} + 4\epsilon^p_{hm} \\
&\geqslant \hat{\mu}^p_{\sigma(j)} + 2\epsilon^p_{hm}.
\end{aligned}$$

由 $|\{j : j \geqslant M+1, \sigma(j) \in \mathcal{E}\}| \geqslant K - M - |\mathcal{B}| = K_p - M_p$, 分布 $\sigma(k)$ 被放入 $\mathcal{A}$ 不迟于 $p_{\sigma(k)}$, 这意味着在这个阶段的末尾, 有一个用户将会固定于这个分布, 简单计算得到:

$$s_{\sigma(k)} = \left\lceil \frac{32\log(T)}{\Delta^2_{\sigma(k)}} \right\rceil.$$

此外, 我们有 $T^{p+1} = \sum_{q=1}^{p+1} M_q \lceil 2^q \log(T) \rceil \leqslant 3T^p = \sum_{q=1}^{p} 3M_q \lceil 2^q \log(T) \rceil$, 由于 M_p 是非增的, 分布 $\sigma(k)$ 被采样的总次数 $T_{\sigma(k)}$ 满足:

$$T_{\sigma(k)} \leqslant T^{p_{\sigma(k)}} \leqslant 3T^{p_{\sigma(k)}-1} \leqslant 3s_{\sigma(k)} = 3\left\lceil \frac{32\log(T)}{\Delta^2_{\sigma(k)}} \right\rceil \leqslant \frac{96\log(T)}{\Delta^2_{\sigma(k)}} + 3.$$

对于 M-worst 分布是类似的, 只需要使用类似的如下符号即可: $\Delta_{\sigma(k)} = \mu_{\sigma(M)} - \mu_{\sigma(k)}, \forall k > n$. □

引理 3.7 在 F_1 发生的条件下, BEACON-HM 探索的损失有上界:

$$\begin{aligned}
R^{expl}_{hm}(T) \leqslant &\sum_{k>n} 96(2+\sqrt{3})\frac{\log(T)}{\mu_{\sigma(M)} - \mu_{\sigma(k)}} + 96\left(1+\sqrt{3}\right)(n-M)\frac{\log(T)}{\mu_{\sigma(l)} - \mu_{\sigma(M+1)}} \\
&+ 12(K-M)\sqrt{2}\left(1+\sqrt{3}\right)\sqrt{\log(T)} + 3K.
\end{aligned} \tag{3.4}$$

证明 根据引理 3.5, 只需求取 $\forall k > n, \forall 1 \leqslant k \leqslant l, \left(\mu_{\sigma(M)} - \mu_{\sigma(k)}\right) T^{expl}_{\sigma(k)}$ 以及 $\left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right)\left(T^{expl} - T^{expl}_{\sigma(k)}\right)$ 的上界。首先, $\forall k > n$, 有

$$\left(\mu_{\sigma(M)} - \mu_{\sigma(k)}\right) T^{expl}_{\sigma(k)} \overset{(i)}{\leqslant} \frac{96\log(T)}{\mu_{\sigma(M)} - \mu_{\sigma(k)}} + 3\left(\mu_{\sigma(M)} - \mu_{\sigma(k)}\right),$$

其中不等式 (i) 是引理 3.6的结果。利用引理 3.6中相同的符号 p_k, 我们可以得到

$$\begin{aligned}
&\sum_{k=1}^{l}\left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right)\left(T^{expl} - T^{expl}_{\sigma(k)}\right) \\
&\overset{(i)}{\leqslant} \sum_{k=1}^{l}\sum_{p=1}^{p_{\sigma(k)}}\left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right)\left(K_p - M_p\right)\lceil 2^p\log(T)\rceil \\
&\overset{(ii)}{\leqslant} \sum_{k=1}^{l}\sum_{p=1}^{p_{\sigma(k)}}\sum_{j=n+1}^{K}\left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right)\lceil 2^p\log(T)\rceil\mathcal{I}\left\{p-1 \leqslant p_{\sigma(j)}\right\} \\
&\quad + \sum_{k=1}^{l}\sum_{p=1}^{p_{\sigma(k)}}\sum_{j=M+1}^{n}\left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right)\lceil 2^p\log(T)\rceil,
\end{aligned} \tag{3.5}$$

其中不等式 (i) 是因为被接受的 M-best 分布 $\sigma(k)$ 在第 p 个阶段会被采样 $K_p\lceil 2^p\log(T)\rceil$ 次; 没有在阶段 p 被接受的分布只被采样 $M_p\lceil 2^p\log(T)\rceil$ 次; 不等式 (ii) 是因为 $K_p - M_p$ 是第 p 阶段满足 $j \geqslant M+1$ 且 $\sigma(j) \in \mathcal{E}$ 的分布数目, 因此 $K_p - M_p \leqslant n - M + \sum_{j=n+1}^{K}\mathcal{I}\left\{p-1 \leqslant p_{\sigma(j)}\right\}$。我们可以进一步考虑 $\forall n+1 \leqslant j \leqslant K$:

$$\begin{aligned}
&\sum_{k=1}^{l}\sum_{p=1}^{p_{\sigma(k)}}\left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right)\lceil 2^p\log(T)\rceil\mathcal{I}\left\{p-1 \leqslant p_{\sigma(j)}\right\} \\
&= \sum_{k=1}^{l}\sum_{p=1}^{p_{\sigma(j)}+1}\left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right)\lceil 2^p\log(T)\rceil\mathcal{I}\left\{p \leqslant p_{\sigma(k)}\right\} \\
&\overset{(i)}{\leqslant} 8\sum_{p=1}^{p_{\sigma(j)}+1}\epsilon_{hm}^{p-1}\left(T^p - T^{p-1}\right) \\
&\overset{(ii)}{\leqslant} 4\log(T)\sum_{p=1}^{p_{\sigma(j)}+1}\epsilon_{hm}^{p-1}\left(\frac{1}{(\epsilon_{hm}^p)^2} - \frac{1}{(\epsilon_{hm}^{p-1})^2}\right) \\
&\leqslant 4\log(T)\sum_{p=1}^{p_{\sigma(j)}+1}\epsilon_{hm}^{p-1}\left(\frac{1}{\epsilon_{hm}^p} + \frac{1}{\epsilon_{hm}^{p-1}}\right)\left(\frac{1}{\epsilon_{hm}^p} - \frac{1}{\epsilon_{hm}^{p-1}}\right) \\
&\overset{(iii)}{\leqslant} 4\left(1+\sqrt{3}\right)\log(T)\sum_{p=1}^{p_{\sigma(j)}+1}\left(\frac{1}{\epsilon_{hm}^p} - \frac{1}{\epsilon_{hm}^{p-1}}\right) \\
&\leqslant 4\left(1+\sqrt{3}\right)\log(T)\frac{1}{\epsilon_{hm}^{p_{\sigma(j)}+1}} \\
&\overset{(iv)}{\leqslant} 4\left(3+\sqrt{3}\right)\sqrt{\frac{2T^{p_{\sigma(j)}}}{\log(T)}}\log(T) \\
&\overset{(v)}{\leqslant} 4\left(3+\sqrt{3}\right)\sqrt{2\left(\frac{96\log(T)}{(\mu_{\sigma(M)} - \mu_{\sigma(j)})^2} + 3\right)\log(T)} \\
&\leqslant 96\left(1+\sqrt{3}\right)\frac{\log(T)}{\mu_{\sigma(M)} - \mu_{\sigma(j)}} + 12\sqrt{2}\left(1+\sqrt{3}\right)\sqrt{\log(T)},
\end{aligned}$$

其中不等式 (i) 因为 $\forall p \leqslant p_{\sigma(k)}, \mu_{\sigma(k)} - \mu_{\sigma(M)} \leqslant \mu_{\sigma(k)} - \mu_{\sigma(M+1)} \leqslant 8\epsilon_{hm}^{p-1}$ 且 $\sum_{k=1}^{l} \mathcal{I}\left\{p \leqslant p_{\sigma(k)}\right\} \leqslant M_p$; 不等式 (ii) 是由于 ϵ_{hm}^{p} 的定义; 不等式 (iii) 是因为 $\frac{\epsilon_{hm}^{p-1}}{\epsilon_{hm}^{p}} + 1 = \sqrt{\frac{T^p}{T^{p-1}}} + 1 \leqslant \sqrt{3} + 1$; 不等式 (iv) 是由于 $T^{p+1} \leqslant 3T^p$; 且不等式 (v) 是引理 3.6的结果。更近一步, 对 $M+1 \leqslant j \leqslant n$, 我们有

$$
\begin{aligned}
\sum_{k=1}^{l} \sum_{p=1}^{p_{\sigma(k)}} \left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right) \lceil 2^p \log(T) \rceil &= \sum_{k=1}^{l} \sum_{p=1}^{p_{\sigma(k)}} \left(\mu_{\sigma(k)} - \mu_{\sigma(M)}\right) \lceil 2^p \log(T) \rceil \mathcal{I}\{p \leqslant p_{\sigma(k)}\} \\
&\leqslant 8 \sum_{p=1}^{1+\max p_{\sigma(k)}} \epsilon_{hm}^{p-1} (T^p - T^{p-1}) \\
&\leqslant 4(1+\sqrt{3}) \log(T) \sum_{p=1}^{1+\max p_{\sigma(k)}} \left(\frac{1}{\epsilon_1^p} - \frac{1}{\epsilon_1^{p-1}}\right) \\
&\leqslant 4(1+\sqrt{3}) \log(T) \sqrt{\frac{2 \max T^{p_{\sigma(k)}+1}}{\log(T)}} \\
&\leqslant 4\sqrt{2}\left(3+\sqrt{3}\right) \sqrt{\frac{96 \log^2(T)}{(\mu_{\sigma(l)} - \mu_{\sigma(M+1)})^2} + 3\log(T)} \\
&\leqslant 96\left(1+\sqrt{3}\right) \frac{\log(T)}{\mu_{\sigma(l)} - \mu_{\sigma(M+1)}} + 12\sqrt{2}\left(1+\sqrt{3}\right) \sqrt{\log(T)}.
\end{aligned}
$$

因此, 引理 3.7可以如下证明:

$$
\begin{aligned}
& R_{hm}^{expl}(T) \\
\leqslant & \sum_{k>n} \left[\frac{96 \log(T)}{\mu_{\sigma(M)} - \mu_{\sigma(k)}} + 3\left(\mu_{\sigma(M)} - \mu_{\sigma(k)}\right)\right] + \sum_{j>n} \left[\frac{96\left(1+\sqrt{3}\right) \log(T)}{\mu_{\sigma(M)} - \mu_{\sigma(j)}} + 12\sqrt{2}\left(1+\sqrt{3}\right) \sqrt{\log(T)}\right] \\
& + 96\left(1+\sqrt{3}\right)(n-M) \frac{\log(T)}{\mu_{\sigma(l)} - \mu_{\sigma(M+1)}} + 12(n-M)\sqrt{2}\left(1+\sqrt{3}\right) \sqrt{\log(T)} \\
\leqslant & \sum_{k>n} \frac{96\left(2+\sqrt{3}\right) \log(T)}{\mu_{\sigma(M)} - \mu_{\sigma(k)}} + \frac{96\left(1+\sqrt{3}\right)(n-M) \log(T)}{\mu_{\sigma(l)} - \mu_{\sigma(M)}} + 12(K-M)\sqrt{2}\left(1+\sqrt{3}\right) \sqrt{\log(T)} + 3K.
\end{aligned}
$$

□

四、通信损失

引理 3.8 (BEACON-HM 的通信损失) 在 F_1 发生的条件下, BEACON-HM 通信的损失有上界:

$$
R_{hm}^{comm}(T) \leqslant 20M \log_2(M) K \log_2(T) + \frac{20M^2}{\sqrt{2}-1} \log_2(M) K + \left\lceil M^2 K \log_2(K) \right\rceil. \tag{3.6}
$$

特别的, 当 $n = M$, 也就是最优分布组合唯一, 我们有

$$R_{hm,s}^{comm}(T) \leqslant 20M\log_2(M)K\log_2\left(\frac{96}{\Delta_{hm}^2}\right) + \frac{20M^2}{\sqrt{2}-1}\log_2(M)K + \left\lceil M^2K\log_2(K)\right\rceil. \tag{3.7}$$

证明 在任何阶段,$M-1$ 个 *followers* 至多为 K 个分布发送 6 比特的信息, 这会导致至多 $6(M-1)K$ 的通信步; 此外, 至多有 $4\lceil\log_2(K)\rceil$ 比特来传输 $\mathcal{A}_p, \mathcal{B}_p$, 以及用以通信的分布 C_m^{p+1}, C_1^{p+1}。在阶段 p, 对仍然活跃的分布, 基于差分通信构造的估计量的量化损失至多为:$2\epsilon_{hm}^{p-1}$, 以经验最优分布组合作为通信分布, 一次采样至多导致 $2\epsilon_{hm}^{p-1}$ 的损失。此外, 在每一个通信的时隙, 至多只有两个用户碰撞, 对每个通信步, 损失至多为 $2+2(M-2)\epsilon_{hm}^{p-1}$ 损失。

当有多个最优分布组合时, 至多会有 $\log_2(T)$ 个探索-通信阶段, 因此, 此时通信损失至多为

$$\begin{aligned}
&\sum_{p=1}^{\log_2(T)}\left[\left(5+\frac{1}{2}\lceil\log_2(M)\rceil\right)(M-1)K + 4(M-1)\lceil\log_2(K)\rceil\right](2+2(M-2)\epsilon_{hm}^{p-1}) \\
&\leqslant \sum_{p=1}^{\log_2(T)} 10M\log_2(M)K\left(2+2M\sqrt{\frac{\log_2(T)}{2T^{p-1}}}\right) \\
&\overset{(i)}{\leqslant} 20M\log_2(M)K\log_2(T) + \sum_{p=1}^{\log_2(T)}\frac{20M^2\log_2(M)K}{\sqrt{2^p}} \\
&\leqslant 20M\log_2(M)K\log_2(T) + \frac{20}{\sqrt{2}-1}M^2\log_2(M)K,
\end{aligned}$$

其中不等式 (i) 是因为 $T^p \geqslant 2^p\log(T)$。此外, 至多有 K 个分布被接受或者拒绝, 传输这些信息至多需要 $M^2K\lceil\log_2(K)\rceil$ 的损失, 因此, 总的通信损失为:

$$R_{hm}^{comm}(T) \leqslant 20M\log_2(M)K\log_2(T) + \frac{20}{\sqrt{2}-1}M^2\log_2(M)K + M^2K\lceil\log_2(K)\rceil.$$

当的最优分布组合唯一, 所有 M-worst 分布被消除后系统进入守成; 其需要的采样次数在引理 3.6中刻画。具体的说, 在阶段 $p_{\sigma(M+1)}$, 要使得条件: $T^{p_{\sigma(M+1)}} \geqslant \frac{96\log(T)}{(\mu_{\sigma(M)}-\mu_{\sigma(M+1)})^2}+3$ 满足。通过 $T^{p_{\sigma(M+1)}} = \sum_{p=1}^{p_{\sigma(M+1)}} M_p 2^p\lceil\log(T)\rceil \geqslant 3+2^{p_{\sigma(M+1)}}\log(T)$,

我们可以反解出阶段数:$p_{\sigma(M+1)} \leqslant \log_2\left(\frac{96}{(\mu_{\sigma(M)}-\mu_{\sigma(M+1)})^2}\right) = \log_2\left(\frac{96}{\Delta_{hm}^2}\right)$。此后,所有用户固定在最优分布上,不再有损失,因此:

$$
\begin{aligned}
&\sum_{p=1}^{p_{\sigma(M+1)}} \left[\left(5+\frac{1}{2}\lceil\log_2(M)\rceil\right)(M-1)K + 4(M-1)\lceil\log(K)\rceil\right](2+2(M-2)\epsilon_{hm}^p) \\
&\leqslant \sum_{p=1}^{p_{\sigma(M+1)}} 10M\log_2(M)K\left(2+2(M-2)\sqrt{\frac{\log_2(T)}{2T^p}}\right) \\
&\overset{(i)}{\leqslant} 20M\log_2(M)Kp_{\sigma(M+1)} + \frac{20}{\sqrt{2}-1}M^2K \\
&\leqslant 20M\log_2(M)K\log_2\left(\frac{96}{\Delta_{hm}^2}\right) + \frac{20}{\sqrt{2}-1}M^2\log_2(M)K.
\end{aligned}
$$

总的来说,我们有:

$$
R_{hm,s}^{comm}(T) \leqslant 20M\log_2(M)K\log_2\left(\frac{96}{\Delta_{hm}^2}\right) + \frac{20}{\sqrt{2}-1}M^2\log_2(M)K + M^2K\lceil\log_2(K)\rceil.
$$

□

第四章　MPMAB: 可观察与非同质情况

在这一章中, 我们将针对非同质多臂老虎机模型设计算法。和同质模型相比, 推广到非同质模型的本质困难来自于指数增长的探索空间, 因此, 除了上一章所提出的自适应差分通信算法外, 我们还需要更加高效的探索算法。此外, 我们在本章将把线性奖励函数推广到更广的满足一定假设的非线性函数, 许多现实中具有应用的例子都被我们考虑的函数族所涵盖。

第一节　奖励函数与假设

在这一节中, 我们引入一般的 (非线性) 奖励函数, 以及我们对它所做的假设; 一般性的非同质模型在第一章中已经给出。

奖励函数。至今为止, 我们考虑的奖励函数都是简单的线性函数, 也就是 M 个用户的奖励相加, 这对应于简单的最大收益化的想法。但是, 现实中出于公平性或者风险控制的想法, 我们往往会考虑更加一般的优化目标, 此时, 线性奖励函数就无法描述用户所具有的共同目标。假定时刻 t, 用户选择的分布组合为 S, 一个非负的随机系统奖励被记为 $V(S,t)$, 我们用以下例子进一步说明这个概念:

- 线性奖励函数: $V(S,t)=\sum_{m\in[M]}O_{s_m,m}(t)$;
- 均衡公平性函数: $V(S,t)=\sum_{m\in[M]}\omega_m\log(\epsilon+O_{s_m,m}(t))$, 其中 $\epsilon>0$ 且 $\omega_m>0$ 为常数;
- 最小奖励函数: $V(S,t)=\min_{m\in[M]}\{O_{s_m,m}(t)\}$, 这个奖励函数对应于系统奖励由最低的奖励决定, 也就是系统的短板。

假设 4.1 (奖励函数族的假设)　我们做如下假设:

- 存在一个期望奖励函数 $v(\cdot)$ 使得 $V_S := \mathbb{E}[V(S,t)] = v(\boldsymbol{\mu}_S \odot \boldsymbol{\eta}_S)$;[①]
- 对任意有重复分布选择的分布组合, 也就是 $\exists m \neq n \in [M]$ 使得 $s'_m = s'_n$, 总存在一个不发生碰撞的分布组合 $S \in \mathcal{S}$ 使得 $V_S \geqslant V_{S'}$;
- 期望奖励函数是关于 $\boldsymbol{\Lambda} = \boldsymbol{\mu}_S$ 是单调非降的, 也就是, 如果 $\boldsymbol{\Lambda} \preceq \boldsymbol{\Lambda}'$, 则 $v(\boldsymbol{\Lambda}) \leqslant v(\boldsymbol{\Lambda}')$;
- 存在一个严格递增的函数 $f(\cdot)$ 使得 $\forall \boldsymbol{\Lambda}, \boldsymbol{\Lambda}', |v(\boldsymbol{\Lambda}) - v(\boldsymbol{\Lambda}')| \leqslant f(\|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}'\|_\infty)$。

第二节　算法发展

非同质情形下的 BEACON 算法实现与同质情况下有类似的框架[38], 我们将避免大段的重复, 相反, 我们集中于此时更加高效的探索算法设计上, 完成的算法在算法 6与算法 7中给出。

一、基于 UCB 的分布式探索机制

在这一小节, 我们讨论如何在分布式情况下设计 UCB 类的探索算法。Boursier et al.[16] 提出的 METC 基于 Successive Elimination, 探索以阶段进行, 用户不断探索所有未消除的分布, 这忽视了这些分布的采样历史, 从而导致次优的探索损失。Boursier et al.[16] 证明, 基于 Successive Elimination 的 METC 具有理论遗憾上界: 1) 当最优分布唯一且 T 已知, 上界为 $\tilde{O}\big(\frac{M^3K}{\Delta_{\min}}\log(T)\big)$; 2) 当最优分布未必唯一, 但 T 已知, 此时上界为 $\tilde{O}\big(MK\big(\frac{M^2}{\Delta_{\min}}\log(T)\big)^{1+1/c_2}\big)$。根据 Kveton et al.[25] 中的结果, 对于线性奖励函数, 理论下界为 $\Omega\big(\frac{M^2K}{\Delta_{\min}}\log(T)\big)$, 可见 METC 并没有达到最优。

批量 UCB 探索: 分布组合的选择。我们将提出的探索方案如下, 在每一个阶段 r 的起始阶段, 每个用户会为每个分布 (k,m) 维护一个计数器 $p^r_{k,m}$, 计数器的

[①]对于 $x = [x_1, ..., x_N], y = [y_1, ..., y_N]$, 我们记 $x \odot y := [x_1y_1, ..., x_Ny_N]$.

Algorithm 6 BEACON-HT: Leader

Set epoch counter $r \leftarrow 0$; arm counter $[p^r_{k,m}] \leftarrow [-1]$; sample time $[T^r_{k,m}] \leftarrow [0]$; received statistics $[\tilde{\mu}^r_{k,m}] \leftarrow [0]$

In the order of $k \in [K]$, play arm k and $T^{r+1}_{k,1} \leftarrow T^r_{k,1} + 1$

while not reaching the time horizon **do**

  $r \leftarrow r+1$; $\forall (k,m) \in [K] \times [M], p^r_{k,m} \leftarrow \left\lfloor \log_2(T^r_{k,m}) \right\rfloor$

  Update $\hat{\mu}^r_{k,1}$ with the first $2^{p^r_{k,1}}$ samples from arm $(k,1)$ during exploration phases

  ▷ *Communication Phase*

  for $(k,m) \in [K] \times [M]$ **do**

    if $p^r_{k,m} > p^{r-1}_{k,m}$ **then**: $\tilde{\delta}^r_{k,m} \leftarrow \texttt{Receive}(\tilde{\delta}^r_{k,m}, m)$; $\tilde{\mu}^r_{k,m} \leftarrow \tilde{\mu}^{r-1}_{k,m} + \tilde{\delta}^r_{k,m}$

    else: $\tilde{\mu}^r_{k,m} \leftarrow \tilde{\mu}^{r-1}_{k,m}$

    end if

  end for

  $\forall (k,m) \in [K] \times [M], \bar{\mu}^r_{k,m} \leftarrow \tilde{\mu}^r_{k,m} + \sqrt{3 \log t^r / 2^{p^r_{k,m}+1}}$

  $S^r = [s^r_1, ..., s^r_M] \leftarrow \texttt{Oracle}(\bar{\boldsymbol{\mu}}^{\boldsymbol{r}})$ and $\forall m \in [M], \texttt{Send}(s^r_m, m)$

  ▷ *Exploration Phase*

  $p^r \leftarrow \min_{m \in [M]} p^r_{s^r_m, m}$; Play arm s^r_1 for 2^{p^r} times

  Signal followers to stop exploration and update $\forall m \in [M], T^{r+1}_{s_m,m} \leftarrow T^r_{s_m,m} + 2^{p^r}$

end while

更新规则为 $p^r_{k,m} = \lfloor \log_2(T^r_{k,m}) \rfloor$, 其中 $T^r_{k,m}$ 是分布 (k,m) 截止到阶段 r 被采样的总次数。之后, *Leader* 将会得到一个置信上界矩阵 $\bar{\boldsymbol{\mu}}^r = [\bar{\mu}^r_{k,m}]_{(k,m) \in [K] \times [M]}$。特别地, 我们令 $\bar{\mu}^r_{k,m} = \tilde{\mu}^r_{k,m} + \sqrt{3 \log t^r / 2^{p^r_{k,m}+1}}$, 其中 t^r 是阶段 r 探索结束那一刻的时间, $\tilde{\mu}^r_{k,m}$ 是 *Leader* 接收到的分布 (k,m) 的量化后的样本均值 $\hat{\mu}^r_{k,m}$。之后, 置信上界矩阵将会被输入一个 `Oracle`, 它将输出一个分布组合 $S^r = [s^r_1, ..., s^r_M] \leftarrow \texttt{Oracle}(\bar{\boldsymbol{\mu}}^{\boldsymbol{r}})$, 基于我们对奖励函数的假设, 我们可以假定这个分布组合是无碰撞的, *Leader* 将会将对应的 S^r_m 发送给用户 m, 之后, 这个分布组合将会在下一个阶段被采样。

Algorithm 7 BEACON-HT: Follower m

Set epoch counter $r \leftarrow 0$; counter $[p^r_{k,m}]_{k\in[K]} \leftarrow 0$; sample time $[T^r_{k,m}]_{k\in[M]} \leftarrow 0$; communicated statistics $[\tilde{\mu}^r_{k,m}]_{k\in[M]} \leftarrow 0$
In order $k \in [K]$, play each $[(m-1+k) \bmod K]$ once; update sample time $T^{r+1}_{k,m} \leftarrow T^r_{k,m} + 1$
while not reaching the time horizon T **do**
    $r \leftarrow r+1$; $\forall k \in [K], p^r_{k,m} \leftarrow \left\lfloor \log_2(T^r_{k,m}) \right\rfloor$
    Update $\hat{\mu}^r_{k,m}$ from the first $2^{p^r_{k,m}}$ samples from arm (k,m) during exploration
    ▷ *Communication Phase*
    for $k \in [K]$ **do**
        if $p^r_{k,m} > p^{r-1}_{k,m}$ **then**
            $\hat{\mu}^r_{k,m} \leftarrow \texttt{ceil}(\tilde{\mu}^r_{k,m})$ with $1 + p^r_{k,m}$ bits and $\tilde{\delta}^r_{k,m} \leftarrow \tilde{\mu}^r_{k,m} - \tilde{\mu}^{r-1}_{k,m}$
            $\texttt{Send}(\tilde{\delta}^r_{k,m}, 1)$
        else: $\tilde{\mu}^r_{k,m} \leftarrow \tilde{\mu}^{r-1}_{k,m}$
        end if
    end for
    $s^r_m \leftarrow \texttt{Receive}(s^r_m, 1)$
    ▷ *Exploration Phase*
    Play arm s^r_m until signaled
    Update $T^{r+1}_{s^r_m,m} \leftarrow T^r_{s^r_m,m} + 2^{p^r}$
end while

批量 UCB 探索: 探索阶段的长度。一个很重要的点在于, 探索的长度是由待采样的分布组合中, 被采样次数最少的分布决定的。具体的说, 对分布组合 S^r, 我们记 $p^r = \arg\min_{m\in[M]} p^r_{s^r_m,m}$, 则下一个探索的阶段长度为 2^{p^r}, 在整个探索阶段中, 用户们都将采样 S^r。这是由于, 一个分布组合的不确定性是由被采样最少的分布决定的。我们所选择的探索长度 2^{p^r} 保证了充分而不过于浪费的探索。特别的, 我们选择的探索长度与 Auer et al.[1] 提出的单机算法 UCB2 具有相同的设计思想, 而单机情况已经证明, UCB2 与 UCB1 有类似的理论保证。

批量 UCB 探索: 利用收集的数据。当第 $r+1$ 个阶段完成后, 如果 $p_{k,m}^{r+1} > p_{k,m}^{r}$, 则 *Leader* 将会从 *Follower* m 处收集新的数据 $\tilde{\mu}_{k,m}^{r+1}$; 否则, $\tilde{\mu}_{k,m}^{r+1}$ 将维持不变, 也就是新的一个阶段收集的数据不会被使用。也就是说, 只有新的数据累积到一定程度, 我们才会去更新的, 这有助于降低通信频率从而降低通信损失。收集数据完成后, *Leader* 进行同样的置信上界矩阵计算, 并获取新的探索分布组合 S^{r+1}, 决定新的阶段长度 p^{r+1}。

二、自适应差分通信与 Stop-upon-signal 机制

BEACON 在非同质情况下的通信及结构与同质情况下大体相似, 其核心仍然是自适应差分通信, 然而, 由于我们此时探索算法是基于 UCB 算法实现, 在通信内容与细节部分有所不同. 我们在这里进行描述。

通信的内容。根据探索阶段的讨论, 我们需要传输三部分的信息: 1) 分布的样本均值 $\tilde{\mu}_{k,m}^{r}$; 2) 选择的分布组合 S^r; 3) 探索阶段的长度 p^r。其中, 分布的组合与探索阶段的长度是容易传输的, 因为它们只需要用**固定而有界**的比特数进行传输。[②] 而样本均值 $\tilde{\mu}_{k,m}^{r}$ 的传输要更加具有挑战性, 我们仍然利用自适应差分通信算法完成。

Stop-upon-signal。出于理论分析的简便, *Leader* 将不会传输给 *Follower* 探索长度, 相反, 当探索结束时, *Leader* 将会使用 $M-1$ 步来制造碰撞作为停止信号。类似的, 下一阶段采样的分布, 通信所用的分布传输也是基于类似的信号机制。此外, 样本均值将会以: " 以碰撞作为开始传输信号, 一个无碰撞信号, 一数据位, ..., 一个无碰撞信号, 一数据位, 以碰撞作为结束信号" 的方式传输, 换言之, 为传出 L 比特的信息, 我们需要 $2L+2$ 长度的时间。

[②] 事实上, 直接传输 p^r 可能导致通信损失成为主项, 因此 BEACON 事实上采用的是 stop-upon-signal 协议。

第三节　理论结果

在这一章中, 我们给出线性奖励函数与我们所考虑的奖励函数族情况下的算法理论遗憾分析。我们定义如下记号: $\mathcal{S}_c = \{S|\exists m \neq n, s_m = s_n\}$ 是存在碰撞的分布组合; $\mathcal{S}_b = \mathcal{S}\backslash(\mathcal{S}_*\cup\mathcal{S}_c)$ 是没有碰撞的次优分布组合; $\Delta_{\min}^{k,m} = V_* - \max\{V_S|S \in \mathcal{S}_b, s_m = k\}$; $\Delta_{\max}^{k,m} = V_* - \min\{V_S|S \in \mathcal{S}_b, s_m = k\}$; $\Delta_{\min} = \min\{\Delta_{\min}^{k,m}|(k,m) \in [K] \times [M]\}$; 以及 $\Delta_c = f(1)$ 是由碰撞引起的损失, 也是一次采样导致的最大损失。

一、主要结果

对于我们考虑的一般奖励函数族, BEACON-HT 的理论遗憾上界在下面的定理中给出 [③]

定理 4.1 (一般奖励函数族)　在本章所做的假设下, 对于一般的奖励函数族, BEACON-HT 的理论遗憾上界为:

$$\begin{aligned} R(T) &\leqslant \sum_{(k,m)} \left[\frac{28\Delta_{\min}^{k,m}}{(f^{-1}(\Delta_{\min}^{k,m}))^2} + \int_{\Delta_{\min}^{k,m}}^{\Delta_{\max}^{k,m}} \frac{28}{(f^{-1}(x))^2} dx \right] \log(T) \\ &+ \frac{6M^2K\log_2(K)}{\log 2}\Delta_c \log(T) + \frac{15MK}{\log 2}\Delta_c \log(T) + o(\log(T)) \\ &\leqslant \tilde{O}\left(\sum\nolimits_{(k,m)} \frac{\Delta_{\max}^{k,m}\log(T)}{(f^{-1}(\Delta_{\min}^{k,m}))^2} + M^2K\Delta_c \log(T) \right), \end{aligned} \tag{4.1}$$

其中求和关于 $(k,m) \in [K] \times [M]$。

现存文献并没有对一般的奖励函数建立理论下界, 因此我们并不知道这个结果距离理论最优有多远; 我们现在将 BEACON-HT 放回到线性奖励函数情况, 特别地, 我们强调, 我们将给出的线性函数情况并不是上述结论的特殊情况, 事实上, 线性函数的特殊结构使得我们可以给出更精细的分析, 从而获得更好的结果:

[③] 渐进符号 $\tilde{O}(\cdot)$ 表示关于 K 的对数项被忽略。

定理 4.2 对于线性奖励函数族, BEACON-HT 的理论遗憾上界为:

$$
\begin{aligned}
&R_{\text{linear}}(T)\\
\leqslant&\sum\nolimits_{(k,m)} \frac{3727M}{\Delta_{\min}^{k,m}}\log(T)+\frac{38M^2K}{\log 2}\log(T)+\frac{4M^2K}{\log 2}\log_2(K)\log(T)+o(\log(T))\\
\leqslant&\tilde{O}\left(\sum\nolimits_{(k,m)}\frac{M}{\Delta_{\min}^{k,m}}\log(T)+M^2K\log(T)\right)\\
\leqslant&\tilde{O}\left(\frac{M^2K}{\Delta_{\min}}\log(T)+M^2K\log(T)\right),
\end{aligned}
\tag{4.2}
$$

其中求和关于 $(k,m)\in[K]\times[M]$。

线性函数的理论下界在 Kveton et al.[25] 中给出为 $\Omega\big(\frac{M^2K}{\Delta_{\min}}\log(T)\big)$, 可以看到 BEACON-HT 在线性奖励函数情况下渐进最优。我们总结非同质情况下的理论结果于表 4.1。可以看到对于线性奖励函数情形, BEACON-HT 是第一个不需要额外假设就能近似有中心化算法表现的 MPMAB 算法, 此外, BEACON-HT 提供了第一个针对一般奖励函数的结果。

表 4.1 非同质化模型算法理论遗憾上界表。

Algorithm/Reference	Reward function	Assumptions	Regret
GoT[13]	Linear	Unique optimal matching	$\tilde{O}(c_1 M\log^{1+\kappa}(T))$
MUMAB[14]	Linear	Known gap $\Delta_{\min}$	$\tilde{O}\big(\frac{K^3}{(\Delta_{\min})^2}\log(T)\big)$
ESE1[15]	Linear	Unique optimal matching	$\tilde{O}\big(\frac{M^2K}{(\Delta_{\min})^2}\log(T)\big)$
METC[16]	Linear	Unique optimal matching; Known T	$\tilde{O}\big(\frac{M^3K}{\Delta_{\min}}\log(T)\big)$
METC[16]	Linear	Known T	$\tilde{O}\big(MK\big(\frac{M^2}{\Delta_{\min}}\log(T)\big)^{1+1/c_2}\big)$
BEACON (this work, Theorem 4.1)	General		$\tilde{O}\big(\frac{MK\Delta_{\max}}{(f^{-1}(\Delta_{\min}))^2}\log(T)\big)$
BEACON (this work, Theorem 4.2)	Linear		$\tilde{O}\big(\frac{M^2K}{\Delta_{\min}}\log(T)\big)$
Lower bound[25]	Linear		$\Omega\big(\frac{M^2K}{\Delta_{\min}}\log(T)\big)$

c_1: 一个与 $M, K, \Delta_{\min}$ 相关的常数; $\frac{1}{c_2}$, κ: 任意小的正常数。

第四节 BEACON-HT 一般奖励函数情形的证明

在这一节中, 我们给出一般奖励函数下 BEACON-HT 理论遗憾上界的证明。我们首先定义一系列我们需要用到的符号:

$\hat{\mu}_{k,m}^r$: 第 r 阶段中分布 (k,m) 的样本均值;

$\tilde{\mu}_{k,m}^r$: 第 r 阶段中分布 (k,m) 的量化样本均值;

$\bar{\mu}_{k,m}^r$: 第 r 阶段中分布 (k,m) 的置信上界 (UCB);

$V_* = \max\{V_S|S \in \mathcal{S}\} = \max\{v(\boldsymbol{\mu}_S \odot \boldsymbol{\eta}_S)|S \in \mathcal{S}\}$: 最优分布组合对应的价值;

$\mathcal{S}_* = \{S|S \in \mathcal{S}, V_S = V_*\}$: 最优分布组合构成的集合;

$\mathcal{S}_c = \{S|\exists m \neq n, s_m = s_n\}$: 存在碰撞的分布组合;

$\mathcal{S}_b = \mathcal{S}\backslash(\mathcal{S}_* \cup \mathcal{S}_c)$: 没有碰撞的次优分布组合;

$\Delta_{\min}^{k,m} = V_* - \max\{V_S|S \in \mathcal{S}_b, s_m = k\}$: 含有分布 (k,m) 的无碰撞分布中, 与 V_* 最小的价值差;

$\Delta_{\max}^{k,m} = V_* - \min\{V_S|S \in \mathcal{S}_b, s_m = k\}$: 含有分布 (k,m) 的无碰撞分布中, 与 V_* 最大的价值差;

$\Delta_{\min} = \min\{\Delta_{\min}^{k,m}|(k,m) \in [K] \times [M]\}$: 无碰撞分布中, 与 V_* 最小的价值差;

$\Delta_{\max} = \max\{\Delta_{\max}^{k,m}\}$: 无碰撞分布中, 与 V_* 最大的价值差;

$\Delta_c = f(1)$: 由碰撞引起的损失, 也是一次采样导致的最大损失,

定理 4.1的证明 类似同质化模型的情况, 总的算法遗憾 $R(T)$ 可以分解为三部分: (1) 探索损失 $R_e(T)$; (2) 通信损失 $R_c(T)$; 以及 (3) 其它损失 $R_o(T)$ (主要来源于初始化), 也就是:

$$R(T) = R_e(T) + R_c(T) + R_o(T).$$

我们直接给出 $R_o(T)$ 的上界, 而 $R_e(T)$ 与 $R_c(T)$ 的分析将会在下面几个引理中给出。$R_o(T)$ 来源于阶段 1 开始之前, 则我们有:

$$R_o(T) \leqslant \left(\frac{K^2M}{K-M} + 2K\right)\Delta_c + K\Delta_{\max}, \tag{4.3}$$

其中第一项来自于引理 3.2以及第二项来自最开始对 MK 个分布的一次采样 (算

法 6与 7的第二行)。利用下面给出的引理 4.3与 4.4, 结合上述 $R_o(T)$ 的上界, 我们完成了定理的证明。 □

一、通信损失: 一般奖励函数

我们给出如下的通信损失上界:

引理 4.3 考虑总决策轮数为 T, BEACON-HT 的通信阶段数至多为:

$$\mathbb{E}[D_c] \leqslant \frac{6}{\log 2} M^2 K \log_2(K) \log(T) + \frac{15}{\log 2} MK \log(T) + MK,$$

通信损失 $R_c(T)$ 至多为:

$$R_c(T) \leqslant \mathbb{E}[D_c]\Delta_c \leqslant \frac{6}{\log 2} M^2 K \log_2(K)\Delta_c \log(T) + \frac{15}{\log 2} MK\Delta_c \log(T) + MK\Delta_c.$$

证明 根据我们对通信部分的讨论以及算法 6与 7, 通信阶段包含三部分的信息交换: (1) 量化样本均值 $\tilde{\mu}^r_{k,m}$; (2) 下一轮采样的分布组合 S^r; (3) 通信阶段长度 p^r。我们分别考虑这三部分导致的通信损失:

Part I: 量化样本均值。我们以 $(k,m), m \neq 1$ 为例。在阶段 1, $\tilde{\mu}^0_{k,m}$ 被初始化为 0, $\hat{\mu}^1_{k,m}$ 是一次对分布 (k,m) 随机采样得到的奖励。由于 $p^1_{k,m} = \lfloor \log_2(T^1_{k,m}) \rfloor = \lfloor \log_2(1) \rfloor = 0$, $\tilde{\mu}^1_{k,m} = \texttt{ceil}(\hat{\mu}^1_{k,m})$ 将被量化为 $1+p^1_{k,m} = 1$ 比特, 特别的 `ceil` 代表当有限比特无法完整表示 $\hat{\mu}^1_{k,m}$, 最后一个比特会被令为 1。差分 $\tilde{\delta}^1_{k,m} = \tilde{\mu}^1_{k,m} - \tilde{\mu}^0_{k,m} = \tilde{\mu}^1_{k,m}$ 同样将被以 1 比特传输。

在阶段 $r > 1$, 如果 $p^r_{k,m} > p^{r-1}_{k,m}$, 也就是 $p^r_{k,m} = p^{r-1}_{k,m} + 1$, 分布 (k,m) 的样本均值将会被传输, 我们将会传输差分 $\tilde{\delta}^r_{k,m} = \tilde{\mu}^r_{k,m} - \tilde{\mu}^{r-1}_{k,m}$ 最后不为 0 的比特, 因此, 我们需要分析 $\tilde{\delta}^r_{k,m}$ 有多少个不为 0 的比特。考虑差值:

$$\begin{aligned}
|\tilde{\delta}^r_{k,m}| &= |\tilde{\mu}^r_{k,m} - \tilde{\mu}^{r-1}_{k,m}| \\
&= |\tilde{\mu}^r_{k,m} - \hat{\mu}^r_{k,m} - (\tilde{\mu}^{r-1}_{k,m} - \hat{\mu}^{r-1}_{k,m}) + (\hat{\mu}^r_{k,m} - \hat{\mu}^{r-1}_{k,m})| \\
&\leqslant |\tilde{\mu}^r_{k,m} - \hat{\mu}^r_{k,m}| + |\tilde{\mu}^{r-1}_{k,m} - \hat{\mu}^{r-1}_{k,m}| + |\hat{\mu}^r_{k,m} - \hat{\mu}^{r-1}_{k,m}|
\end{aligned}$$

$$\overset{(a)}{\leqslant} \frac{1}{2^{p_{k,m}^r}} + \frac{1}{2^{p_{k,m}^r-1}} + |\hat{\mu}_{k,m}^r - \hat{\mu}_{k,m}^{r-1}|,$$

其中不等式 (a) 是因为 $\tilde{\mu}_{k,m}^r$ 被 $1+p_{k,m}^r$ 量化为 $\hat{\mu}_{k,m}^r$, 这会导致至多为 $\frac{1}{2^{p_{k,m}^r}}$ 的量化误差。此外, 令 $\gamma_\tau^{k,m}$ 为探索阶段第 τ 次对分布 (k,m) 采样获得的奖励, 我们将差分 $\hat{\mu}_{k,m}^r - \hat{\mu}_{k,m}^{r-1}$ 重写为:

$$\begin{aligned}
\hat{\mu}_{k,m}^r - \hat{\mu}_{k,m}^{r-1} &= \frac{\sum_{\tau=1}^{2^{p_{k,m}^r}} \gamma_\tau^{k,m}}{2^{p_{k,m}^r}} - \frac{\sum_{\tau=1}^{2^{p_{k,m}^r-1}} \gamma_\tau^{k,m}}{2^{p_{k,m}^r-1}} \\
&= \frac{\sum_{\tau=1}^{2^{p_{k,m}^r-1}} \gamma_\tau^{k,m} + \sum_{\tau=1+2^{p_{k,m}^r-1}}^{2^{p_{k,m}^r}} \gamma_\tau^{k,m}}{2^{p_{k,m}^r}} - \frac{\sum_{\tau=1}^{2^{p_{k,m}^r-1}} \gamma_\tau^{k,m}}{2^{p_{k,m}^r-1}} \\
&= \frac{\sum_{\tau=1+2^{p_{k,m}^r-1}}^{2^{p_{k,m}^r}} \gamma_\tau^{k,m} - \sum_{\tau=1}^{2^{p_{k,m}^r-1}} \gamma_\tau^{k,m}}{2^{p_{k,m}^r}},
\end{aligned}$$

则它是 $\frac{1}{2\sqrt{2^{p_{k,m}^r}}}$-sub-Gaussian 随机变量, 这是因为每一次采样都是独立的。考虑 $x \geqslant \sqrt{\log 2}$, 我们有

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{\mu}_{k,m}^r - \hat{\mu}_{k,m}^{r-1}\right| \geqslant \frac{x}{2^{p_{k,m}^r}}\right) \leqslant 2\exp\left[-2^{p_{k,m}^r+1}\left(\frac{x}{2^{p_{k,m}^r}}\right)^2\right] \leqslant 2\exp[-x^2] \\
\Rightarrow &\mathbb{P}\left(|\tilde{\delta}_{k,m}^r| \geqslant \frac{1}{2^{p_{k,m}^r}} + \frac{1}{2^{p_{k,m}^r-1}} + \frac{x}{2^{p_{k,m}^r}}\right) \leqslant 2\exp[-x^2] \\
\overset{(a)}{\Rightarrow} &\mathbb{P}\left(L_{k,m}^r \geqslant 2 + p_{k,m}^r + \log_2\left(\frac{3+x}{2^{p_{k,m}^r}}\right)\right) \leqslant 2\exp[-x^2] \\
\Rightarrow &\mathbb{P}\left(L_{k,m}^r \geqslant 2 + \log_2(3+x)\right) \leqslant 2\exp[-x^2] \\
\Rightarrow &\mathbb{P}\left(L_{k,m}^r \leqslant 2 + \log_2(3+x)\right) \geqslant 1 - 2\exp[-x^2] \\
\overset{(b)}{\Rightarrow} &\mathbb{P}\left(L_{k,m}^r \leqslant l\right) \geqslant 1 - 2\exp\left[-(2^{l-2}-3)^2\right]
\end{aligned}$$

其中 (a) 是因为 $|\tilde{\delta}_{k,m}^r|$ 的非零比特数 $L_{k,m}^r$ 有上界:

$$\begin{aligned}
L_{k,m}^r &\leqslant 1 + p_{k,m}^r - \lfloor \log_2(1/|\tilde{\delta}_{k,m}^r|) \rfloor \\
&\leqslant 2 + p_{k,m}^r + \log_2(|\tilde{\delta}_{k,m}^r|).
\end{aligned}$$

在 (b) 部分, 我们替换 $2+\log_2(3+x)$ 为 l, 其满足 $l \geqslant 2+\log_2(3+\sqrt{\log 2})$; 等价的, 我们有 $x = 2^{l-2}-3$。如果我们把 $L_{k,m}^r$ 看成随机变量, 那么我们考虑它的

CDF$F_{L_{k,m}^r}(l)$:

$$\forall l \geqslant 4 > 2 + \log_2(3 + \sqrt{\log 2}), F_{L_{k,m}^r}(l) = \mathbb{P}\left(L_{k,m}^r \leqslant l\right) \geqslant 1 - 2\exp\left[-(2^{l-2} - 3)^2\right].$$

我们有:

$$\begin{aligned}
\mathbb{E}\left[L_{k,m}^r\right] &= \sum_{l=0}^{\infty}(1 - F_{L_{k,m}^r}(l)) \\
&\leqslant 5 + \sum_{l=5}^{\infty} 2\exp\left[-(2^{l-2} - 3)^2\right] \\
&\leqslant 5 + \int_{l=4}^{\infty} 2\exp\left[-(2^{l-2} - 3)^2\right] \mathrm{d}l \\
&\leqslant 6.
\end{aligned}$$

因此, 在期望意义下, 少于 6 比特的信息比特将用来传输差分 $|\tilde{\delta}_{k,m}^r|$, 此外我们还需要传输 1 比特的符号位。在整个 T 的决策轮数中, 分布 (k, m) 至多被更新 $\log_2(T)$ 次, 用以传输量化样本均值的期望的通信长度 D_s 因此有上界:

$$\begin{aligned}
\mathbb{E}\left[D_s\right] = &\underbrace{MK}_{\text{epoch } r = 1} + \underbrace{\mathbb{E}\left[\sum_{r} \sum_{(k,m):p_{k,m}^r > p_{k,m}^{r-1}} (1 + 1 + 2L_{k,m}^r + 1)\right]}_{\text{epoches } r > 1} \\
&\overset{(a)}{\leqslant} MK + (3 + 2 \times 6)MK\log_2(T) \\
&\leqslant 15MK\log_2(T) + MK \\
&= \frac{15}{\log 2} MK\log(T) + MK,
\end{aligned} \tag{4.4}$$

第一个等式中的三个"1" 分别代表一个符号位, 一个开始传输信号, 以及一个传输完成信号; 不等式 (a) 是因为每个分布 (k, m) 至多被传输 $\log_2 T$ 次而我们有 MK 个分布。

Part II & III: 分布选择以及阶段长度。在阶段 r, *Leader* 会告知每个 *Follower m* 它要采样的分布, 也是它用以通信的分布, 以及 *Leader* 用以通信的分布 (s_m^r 和 s_1^r), 考虑到额外的一个开始通信的信号位, 分布组合传输的通信长度 D_m 可以如

下给出上界:

$$D_m = \sum_r (M-1)(1+2\lceil \log_2(K) \rceil) \tag{4.5}$$

$$\leqslant (M-1)(2\log_2(K)+3)MK\log_2(T) \tag{4.6}$$

$$< \frac{1}{\log 2} M^2 K (2\log_2(K)+3)\log(T). \tag{4.7}$$

对阶段长度传输所需要的通信长度 D_b, 我们并不直接传输探索长度, 而是采用 stop-upon-signal 的方式, 也就是说, *leader* 将使用一个比特的停止信号 (碰撞) 来告知 *Follower* 来停止探索, 此外, 我们最多有 $MK\log_2(T)$ 的探索阶段, 因此:

$$D_b = \sum_r (M-1) \leqslant (M-1)MK\log_2(T) < \frac{1}{\log 2} M^2 K \log(T). \tag{4.8}$$

总的来说, 我们有:

$$\begin{aligned}
\mathbb{E}[D_c] &= \mathbb{E}[D_s] + \mathbb{E}[D_m] + \mathbb{E}[D_b] \\
&\leqslant \frac{15}{\log 2} MK\log(T) + MK + \frac{1}{\log 2} M^2 (2\log_2(K)+3) K\log(T) + \frac{1}{\log 2} M^2 K \log(T) \\
&= \frac{6}{\log 2} M^2 K \log_2(K)\log(T) + \frac{15}{\log 2} MK\log(T) + MK.
\end{aligned}$$

□

二、探索损失: 一般奖励函数

引理 4.4 考虑总决策轮数为 T, BEACON-HT 的探索损失至多为:

$$\begin{aligned}
R_e(T) &\leqslant \sum_{(k,m)\in[K]\times[M]} \left[\frac{28\Delta_{\min}^{k,m}\log(T)}{(f^{-1}(\Delta_{\min}^{k,m}))^2} + \int_{\Delta_{\min}^{k,m}}^{\Delta_{\max}^{k,m}} \frac{28\log(T)}{(f^{-1}(x))^2} dx + 8KM\Delta_{\max}^{k,m} \right] \\
&\leqslant \sum_{(k,m)\in[K]\times[M]} \frac{28\Delta_{\max}^{k,m}\log(T)}{(f^{-1}(\Delta_{\min}^{k,m}))^2} + 8K^2M^2\Delta_{\max}.
\end{aligned}$$

证明 对探索损失的分析主要受到 Chen et al.[24] 对 CUCB 的证明, 但是我们这里需要考虑的是以阶段形式进行的探索, 因此面临着额外的困难。我们首先给出下列符号:

$$\mathcal{S}_b^{k,m} = \{S | S \in \mathcal{S}_b, s_m = k\} = \{S_1^{k,m}, ..., S_{N(k,m)}^{k,m}\};$$

$$\Delta_n^{k,m} = V_* - V_{S_n^{k,m}}, \forall n \in \{1, ..., N(k,m)\},$$

其中 $\mathcal{S}_b^{k,m}$ 是包含分布 (k,m) 且无碰撞的次优分布组合构成的集合, 它的大小被记作 $N(k,m)$; $\Delta_n^{k,m}$ 是相应次优分布组合与最优分布组合价值的差。我们假定集合 $\mathcal{S}_b^{k,m} = \{S_1^{k,m}, ..., S_{N(k,m)}^{k,m}\}$ 关于 $\Delta_n^{k,m}$ 是降序排列的, 也就是说, 如果 $n_1 \geqslant n_2$, 则 $\Delta_{n_1}^{k,m} \leqslant \Delta_{n_2}^{k,m}$。此外, 为了符号的简便, 我们令 $\Delta_0^{k,m} = 1$ 以及 $\Delta_{N(k,m)+1}^{k,m} = 0$。我们有 $\Delta_{\min}^{k,m} = \Delta_{N(k,m)}^{k,m}$ 与 $\Delta_{\max}^{k,m} = \Delta_1^{k,m}$。我们再定义如下整数: $\forall n \in \{1, ..., N(k,m)\}$, 我们定义 $q_n^{k,m}$ 使得

$$2^{q_n^{k,m}-1} \leqslant \frac{14\log(T)}{(f^{-1}(\Delta_n^{k,m}))^2} < 2^{q_n^{k,m}} < \frac{28\log(T)}{(f^{-1}(\Delta_n^{k,m}))^2}.$$

除此之外, 我们定义 $q_0^{k,m} = 0$, $q_{N(k,m)+1}^{k,m} = \lceil \log_2(T) \rceil$, 由于我们假设 f 是严格递增的, 我们有 $\forall p \geqslant q_n^{k,m}$,

$$f\left(2\sqrt{\frac{3\log t^r}{2^{p+1}}} + \frac{1}{2^{p+1}}\right) \leqslant f\left(3\sqrt{\frac{3\log t^r}{2^{p+1}}}\right) \leqslant f\left(3\sqrt{\frac{3\log T}{2^{p+1}}}\right) \leqslant \Delta_n^{k,m}. \quad (4.9)$$

在阶段 r, 我们定义所谓的 "representative distribution" $\rho^r = (s_m^r, m)$ 为分布组合 S^r 中满足 $p_{s_m^r,m}^r = p^r$ 的分布, 如果不止一个分布满足条件, 我们只需要随机选取一个。因此每个探索阶段有且仅有一个 "representative distribution"。直觉的说, "representative distribution" 是分布组合中计数器最小的, 也就是采样最不充分的分布, 整个探索的长度就由它决定。根据我们更新计数器的规则, 在阶段 r 结束后, 分布 ρ^r 的计数器会自然地增长 1。

Step I: 遗憾分解。我们关于 "representative distribution" 分解理论遗憾如下:

$$\begin{aligned} R_e(T) &= \mathbb{E}\left[\sum_r 2^{p^r}(V_* - V_{S^r})\right] \\ &= \mathbb{E}\left[\sum_r \sum_{(k,m)\in[K]\times[M]} 2^{p^r}(V_* - V_{S^r})\mathcal{I}\{\rho^r = (k,m)\}\right] \\ &\overset{(a)}{=} \mathbb{E}\left[\sum_r \sum_{(k,m)\in[K]\times[M]} 2^{p_{k,m}^r}(V_* - V_{S^r})\mathcal{I}\{\rho^r = (k,m)\}\right] \end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=} \mathbb{E}\left[\sum_{r}\sum_{(k,m)\in[K]\times[M]}\sum_{n=1}^{N(k,m)} 2^{p^r_{k,m}}\Delta_n^{k,m}\mathcal{I}\left\{\rho^r=(k,m), S_r=S_n^{k,m}\right\}\right]\\
&\overset{(c)}{=} \mathbb{E}\left[\sum_{(k,m)\in[K]\times[M]}\sum_{p_{k,m}\geqslant 0}\sum_{n=1}^{N(k,m)} 2^{p_{k,m}}\Delta_n^{k,m}\mathcal{I}\left\{S_{k,m,p_{k,m}}=S_n^{k,m}\right\}\right]\\
&\overset{(d)}{=} \sum_{(k,m)\in[K]\times[M]} R_e^{k,m}(T),
\end{aligned}
$$

等式 (a) 来源于 “representative distribution” 的定义: 如果 $\rho^r=(k,m)$, 则 $p^r=p^r_{k,m}$。等式 (b) 将每个探索阶段的损失与探索的特定次优组合联系起来。$S_{k,m,p_{k,m}}$ 是关于 “representative distribution” (k,m) 和相应的计数器 $p_{k,m}$ 的探索组合。等式 (c) 是因为 $\rho^r=(k,m)$ 说明它的计数器会增加 1。等式 (d) 是符号变化: $R_e^{k,m}(T)=\mathbb{E}\left[\sum_{p_{k,m}>0}\sum_{n=1}^{N(k,m)} 2^{p_{k,m}}\Delta_n^{k,m}\mathcal{I}\left\{S_{k,m,p_{k,m}}=S_n^{k,m}\right\}\right]$。

对 $R_e^{k,m}(T)$, 我们相当于固定分布 (k,m) 作如下的分析:

$$
\begin{aligned}
R_e^{k,m}(T) =&\mathbb{E}\left[\sum_{p_{k,m}\geqslant 0}\sum_{n=1}^{N(k,m)} 2^{p_{k,m}}\Delta_n^{k,m}\mathcal{I}\left\{S_{k,m,p_{k,m}}=S_n^{k,m}\right\}\right]\\
=&\sum_{p_{k,m}\geqslant 0}\sum_{n=1}^{N(k,m)} 2^{p_{k,m}}\Delta_n^{k,m}\mathbb{P}\left(S_{k,m,p_{k,m}}=S_n^{k,m}\right)\\
\overset{(a)}{\leqslant}&\sum_{p_{k,m}>0}\sum_{n=1}^{N(k,m)} 2^{p_{k,m}}\Delta_n^{k,m}\mathbb{P}\left(S_{k,m,p_{k,m}}=S_n^{k,m}|\mathcal{E}_{k,m,p_{k,m}}\right)\mathbb{P}\left(\mathcal{E}_{k,m,p_{k,m}}\right)\\
&+\sum_{p_{k,m}\geqslant 0}\sum_{n=1}^{N(k,m)} 2^{p_{k,m}}\Delta_n^{k,m}\mathbb{P}\left(S_{k,m,p_{k,m}}=S_n^{k,m}|\bar{\mathcal{E}}_{k,m,p_{k,m}}\right)\mathbb{P}\left(\bar{\mathcal{E}}_{k,m,p_{k,m}}\right)\\
\leqslant&\sum_{p_{k,m}\geqslant 0}\sum_{n=1}^{N(k,m)} 2^{p_{k,m}}\Delta_n^{k,m}\mathbb{P}\left(S_{k,m,p_{k,m}}=S_n^{k,m}|\mathcal{E}_{k,m,p_{k,m}}\right)\\
&+\sum_{p_{k,m}\geqslant 0} 2^{p_{k,m}}\Delta_{\max}^{k,m}\mathbb{P}\left(\bar{\mathcal{E}}_{k,m,p_{k,m}}\right)\\
\leqslant&\underbrace{\sum_{h=0}^{N(k,m)}\sum_{q_h^{k,m}\leqslant p_{k,m}<q_{h+1}^{k,m}}\sum_{n=1}^{N(k,m)} 2^{p_{k,m}}\Delta_n^{k,m}\mathbb{P}\left(S_{k,m,p_{k,m}}=S_n^{k,m}|\mathcal{E}_{k,m,p_{k,m}}\right)}_{\text{term (A)}}\\
&+\underbrace{\sum_{p_{k,m}\geqslant 0} 2^{p_{k,m}}\Delta_{\max}^{k,m}\mathbb{P}\left(\bar{\mathcal{E}}_{k,m,p_{k,m}}\right)}_{\text{term (B)}},
\end{aligned}
$$

其中等式 (a) 引入了"nice event" $\mathcal{E}_{k,m,p_{k,m}}$, 在阶段 r, $\mathcal{E}_r$ 定义为:

$$\mathcal{E}_r = \left\{ \forall (k,m) \in [K] \times [M], -\sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}} < \tilde{\mu}_{k,m}^r - \mu_{k,m} < \sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}} + \frac{1}{2^{p_{k,m}^r+1}} \right\}.$$

此外, 当某个阶段的"representative distribution" 为 (k,m), 其对应计数器为 (p,m), 这时候 $\mathcal{E}_{k,m,p_{k,m}}$ 是 $\mathcal{E}_r$ 事件的另外一个记号。

Step II: Bounding term (B). 我们考虑任意一个阶段对应的 $\bar{\mathcal{E}}_r$ 发生的概率:

$$\begin{aligned}
\mathbb{P}\left(\bar{\mathcal{E}}_r\right) &= \mathbb{P}\left(\exists (k,m) \in [K] \times [M], \tilde{\mu}_{k,m}^r - \mu_{k,m} \leqslant -\sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}} \text{ or } \tilde{\mu}_{k,m}^r - \mu_{k,m} \geqslant \sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}} + \frac{1}{2^{p_{k,m}^r+1}} \right) \\
&\leqslant \sum_{(k,m)\in[K]\times[M]} \mathbb{P}\left(\tilde{\mu}_{k,m}^r - \mu_{k,m} \leqslant -\sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}} \right) \\
&+ \sum_{(k,m)\in[K]\times[M]} \mathbb{P}\left(\tilde{\mu}_{k,m}^r - \mu_{k,m} \geqslant \sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}} + \frac{1}{2^{p_{k,m}^r+1}} \right) \\
&= \sum_{(k,m)\in[K]\times[M]} \mathbb{P}\left(\tilde{\mu}_{k,m}^r - \hat{\mu}_{k,m}^r + \hat{\mu}_{k,m}^r - \mu_{k,m} \leqslant -\sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}} \right) \\
&+ \sum_{(k,m)\in[K]\times[M]} \mathbb{P}\left(\tilde{\mu}_{k,m}^r - \hat{\mu}_{k,m}^r + \hat{\mu}_{k,m}^r - \mu_{k,m} \geqslant \sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}} + \frac{1}{2^{p_{k,m}^r+1}} \right) \\
&\overset{(a)}{\leqslant} \sum_{(k,m)\in[K]\times[M]} 2\mathbb{P}\left(\left|\hat{\mu}_{k,m}^r - \mu_{k,m}\right| \geqslant \sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}} \right) \\
&\leqslant \sum_{(k,m)\in[K]\times[M]} \sum_{p_{k,m}=0}^{\lfloor \log_2(t^r) \rfloor} 2\mathbb{P}\left(\left|\hat{\mu}_{k,m}^r - \mu_{k,m}\right| \geqslant \sqrt{\frac{3\log t^r}{2^{p_{k,m}^r+1}}}, p_{k,m}^r = p_{k,m} \right) \\
&\leqslant \sum_{(k,m)\in[K]\times[M]} \sum_{p_{k,m}=0}^{\lfloor \log_2(t^r) \rfloor} 2\mathbb{P}\left(\left| \frac{\sum_{\tau=1}^{2^{p_{k,m}}} \gamma_\tau^{k,m}}{2^{p_{k,m}}} - \mu_{k,m} \right| \geqslant \sqrt{\frac{3\log t^r}{2^{p_{k,m}+1}}} \right) \\
&\overset{(b)}{\leqslant} \sum_{(k,m)\in[K]\times[M]} \sum_{p_{k,m}=0}^{\lfloor \log_2(t^r) \rfloor} 4\exp\left[-2\cdot 2^{p_{k,m}} \frac{3\log t^r}{2^{p_{k,m}+1}} \right] \\
&\leqslant 4KM \frac{\lceil \log_2(t^r) \rceil}{(t^r)^3} \\
&\leqslant 4KM \frac{1}{(t^r)^2} \\
&\overset{(c)}{\leqslant} 4KM \frac{1}{(2^{p^r})^2},
\end{aligned}$$

其中不等式 (a) 是因为 $\tilde{\mu}_{k,m}^r = \texttt{ceil}(\hat{\mu}_{k,m}^r)$ 在 $p_{k,m}^r + 1$ 比特下量化, 因此量化误差小于 $\frac{1}{2^{p_{k,m}^r+1}}$ 以及我们的量化是不对称的, 最后, 我们用了事件的蕴含关系得到了概率的不等式。(b) 是因为 Hoeffding's inequality。不等式 (c) 是因为 $t^r \geqslant 2^{p^r}$, 其中 t^r 是当前总的时间步长。

利用上面的式子, 我们进一步求取 (B) 的上界

$$\begin{aligned}\text{term (B)} =& \sum_{p_{k,m}\geqslant 0} 2^{p_{k,m}}\Delta_{\max}^{k,m}\mathbb{P}\left(\bar{\mathcal{E}}_{k,m,p_{k,m}}\right)\\ \overset{(a)}{\leqslant}& 4\sum_{p_{k,m}\geqslant 0} 2^{p_{k,m}}\Delta_{\max}^{k,m}\cdot KM\frac{1}{(2^{p_{k,m}})^2}\\ =& 4\sum_{p_{k,m}\geqslant 0}\Delta_{\max}^{k,m}\cdot KM\frac{1}{2^{p_{k,m}}}\\ \leqslant& 8KM\Delta_{\max}^{k,m},\end{aligned}$$

其中不等式 (a) 是因为我们对 $\mathbb{P}\left(\bar{\mathcal{E}}_r\right)$ 的分析且 $p^r = p_{k,m}$。

Step III: Bounding term (A). 我们首先证明如下事件蕴含关系: 在阶段 r, 如果 $\rho^r = (k,m)$ 且 $p^r = p_{k,m}^r = p_{k,m}$, 我们令 $\bar{\boldsymbol{\mu}}^r$ 与 S^r 代表 $\bar{\boldsymbol{\mu}}^{k,m,p_{k,m}}$ 和 $S_{k,m,p_{k,m}}$, 如果事件 $\mathcal{E}_{k,m,p_{k,m}}$ 发生, 我们有:

$$\begin{aligned}&p_{k,m}\geqslant q_h^{k,m}, \text{the oracle outputs } S_{k,m,p_{k,m}} = S_n^{k,m}\\ \Rightarrow& p_{k,m}\geqslant q_h^{k,m}, \forall S\in\mathcal{S}_*\backslash\mathcal{S}_c, v(\bar{\boldsymbol{\mu}}_{S_n^{k,m}}^{k,m,p_{k,m}}\odot\boldsymbol{\eta}_{S_n^{k,m}})\geqslant v(\bar{\boldsymbol{\mu}}_S^{k,m,p_{k,m}}\odot\boldsymbol{\eta}_S)\\ \Rightarrow& p_{k,m}\geqslant q_h^{k,m}, \forall S\in\mathcal{S}_*\backslash\mathcal{S}_c, v(\bar{\boldsymbol{\mu}}_{S_n^{k,m}}^{k,m,p_{k,m}})\geqslant v(\bar{\boldsymbol{\mu}}_S^{k,m,p_{k,m}})\\ \overset{(a)}{\Rightarrow}& p_{k,m}\geqslant q_h^{k,m}, \forall S\in\mathcal{S}_*\backslash\mathcal{S}_c, v(\boldsymbol{\mu}_{S_n^{k,m}}) + f\left(\left\|\bar{\boldsymbol{\mu}}_{S_n^{k,m}}^{k,m,p_{k,m}} - \boldsymbol{\mu}_{S_n^{k,m}}\right\|_\infty\right)\geqslant v(\bar{\boldsymbol{\mu}}_{S_n^{k,m}}^{k,m,p_{k,m}})\\ &\geqslant v(\bar{\boldsymbol{\mu}}_S^{k,m,p_{k,m}})\\ \overset{(b)}{\Rightarrow}& p_{k,m}\geqslant q_h^{k,m}, \forall S\in\mathcal{S}_*\backslash\mathcal{S}_c, V_{S_n^{k,m}} + f\left(2\sqrt{\frac{3\log t^r}{2^{p_{k,m}+1}}} + \frac{1}{2^{p_{k,m}+1}}\right)\geqslant v(\bar{\boldsymbol{\mu}}_{S_n^{k,m}}^{k,m,p_{k,m}})\\ &\geqslant v(\bar{\boldsymbol{\mu}}_S^{k,m,p_{k,m}}) > v(\boldsymbol{\mu}_S) = V_*\\ \overset{(c)}{\Rightarrow}& p_{k,m}\geqslant q_h^{k,m}, V_{S_n^{k,m}} + \Delta_h^{k,m} > V_*,\end{aligned}$$

其中 (a) 是因为我们假设 $\forall\boldsymbol{\Lambda},\boldsymbol{\Lambda}', |v(\boldsymbol{\Lambda}) - v(\boldsymbol{\Lambda}')|\leqslant f(\|\boldsymbol{\Lambda}-\boldsymbol{\Lambda}'\|_\infty)$; (b) 是奖励函数单调性的假设以及 $\mathcal{E}_{k,m,p_{k,m}}$ 的定义, 还有 $S_{k,m,p_{k,m}}$ 中的分布的计数器至少为 $p_{k,m}$; (c) 是因为 $q_h^{k,m}$ 的定义以及式子 (4.9)。

由于上面的事件蕴含关系, 我们知道如果 $p_{k,m}\geqslant q_h^{k,m}$, 分布组合 $S_n^{k,m}, n\leqslant h$ 不会是被选中的组合 S^r; 否则, 根据我们对 $\Delta_n^{k,m}$ 关于 n 降序的假设, 我们有

$V_{S_n^{k,m}} + \Delta_h^{k,m} < V_*$, 它与上面的事件蕴含矛盾。所以, 我们可以进一步求取上界:

$$
\begin{aligned}
\text{term (A)} &= \sum_{h=0}^{N(k,m)} \sum_{q_h^{k,m} \leqslant p_{k,m} < q_{h+1}^{k,m}} \sum_{n=1}^{N(k,m)} 2^{p_{k,m}} \Delta_n^{k,m} \mathbb{P}\left(S_{k,m,p_{k,m}} = S_n^{k,m} | \mathcal{E}_{k,m,p_{k,m}}\right) \\
&= \sum_{h=0}^{N(k,m)} \sum_{q_h^{k,m} \leqslant p_{k,m} < q_{h+1}^{k,m}} \sum_{n=h+1}^{N(k,m)} 2^{p_{k,m}} \Delta_n^{k,m} \mathbb{P}\left(S_{k,m,p_{k,m}} = S_n^{k,m} | \mathcal{E}_{k,m,p_{k,m}}\right) \\
&\overset{(a)}{\leqslant} \sum_{h=0}^{N(k,m)} \sum_{q_h^{k,m} \leqslant p_{k,m} < q_{h+1}^{k,m}} \sum_{n=h+1}^{N(k,m)} 2^{p_{k,m}} \Delta_{h+1}^{k,m} \mathbb{P}\left(S_{k,m,p_{k,m}} = S_n^{k,m} | \mathcal{E}_{k,m,p_{k,m}}\right) \\
&\overset{(b)}{\leqslant} \sum_{h=0}^{N(k,m)} \sum_{q_h^{k,m} \leqslant p_{k,m} < q_{h+1}^{k,m}} 2^{p_{k,m}} \Delta_{h+1}^{k,m} \\
&= \sum_{h=0}^{N(k,m)} (2^{q_{h+1}^{k,m}} - 2^{q_h^{k,m}}) \Delta_{h+1}^{k,m} \\
&= \sum_{h=0}^{N(k,m)-1} (2^{q_{h+1}^{k,m}} - 2^{q_h^{k,m}}) \Delta_{h+1}^{k,m} \\
&= 2^{q_{N(k,m)}^{k,m}} \Delta_{N(k,m)}^{k,m} + \sum_{h=0}^{N(k,m)-1} 2^{q_h^{k,m}} \left(\Delta_h^{k,m} - \Delta_{h+1}^{k,m}\right) - 2^{q_0^{k,m}} \Delta_0^{k,m} \\
&\overset{(c)}{\leqslant} \frac{28 \Delta_{N(k,m)}^{k,m} \log(T)}{(f^{-1}(\Delta_{N(k,m)}^{k,m}))^2} + \sum_{h=0}^{N(k,m)-1} \frac{28 \log(T)}{(f^{-1}(\Delta_h^{k,m}))^2} \left(\Delta_h^{k,m} - \Delta_{h+1}^{k,m}\right) \\
&\overset{(d)}{\leqslant} \frac{28 \Delta_{N(k,m)}^{k,m} \log(T)}{(f^{-1}(\Delta_{N(k,m)}^{k,m}))^2} + \int_{\Delta_{N(k,m)}^{k,m}}^{\Delta_1^{k,m}} \frac{28 \log(T)}{(f^{-1}(x))^2} dx \\
&= \frac{28 \Delta_{\min}^{k,m} \log(T)}{(f^{-1}(\Delta_{\min}^{k,m}))^2} + \int_{\Delta_{\min}^{k,m}}^{\Delta_{\max}^{k,m}} \frac{28 \log(T)}{(f^{-1}(x))^2} dx,
\end{aligned}
$$

其中不等式 (a) 是因为, $\forall n \geqslant h+1, \Delta_n^{k,m} \leqslant \Delta_{h+1}^{k,m}$,; 不等式 (b) 是因为 $\sum_{n=h+1}^{N(k,m)} \mathbb{P}\left(S_{k,m,p_{k,m}} = S_n^{k,m} | \mathcal{E}_{k,$ 1; 不等式 (c) 是因为 $q_n^{k,m}$ 的定义; 不等式 (d) 是因为 $\frac{28 \log(T)}{(f^{-1}(x))^2}$ 在 $[\Delta_{N(k,m)}^{k,m}, \Delta_1^{k,m}]$ 严格递减。

根据对 (A), (B) 的分析, 我们有

$$
\begin{aligned}
R_e^{k,m}(T) &\leqslant \frac{28 \Delta_{\min}^{k,m} \log(T)}{(f^{-1}(\Delta_{\min}^{k,m}))^2} + \int_{\Delta_{\min}^{k,m}}^{\Delta_{\max}^{k,m}} \frac{28 \log(T)}{(f^{-1}(x))^2} dx + 4KM\Delta_{\max}^{k,m} \\
&\leqslant \frac{28 \Delta_{\max}^{k,m} \log(T)}{(f^{-1}(\Delta_{\min}^{k,m}))^2} + 8KM\Delta_{\max}^{k,m}.
\end{aligned}
$$

最终我们有:

$$
R_e(T) = \sum_{(k,m) \in [K] \times [M]} R_e^{k,m}(T)
$$

$$\leqslant \sum_{(k,m)\in[K]\times[M]} \left[\frac{28\Delta_{\min}^{k,m}\log(T)}{(f^{-1}(\Delta_{\min}^{k,m}))^2} + \int_{\Delta_{\min}^{k,m}}^{\Delta_{\max}^{k,m}} \frac{28\log(T)}{(f^{-1}(x))^2}dx + 4KM\Delta_{\max}^{k,m} \right]$$

$$\leqslant \sum_{(k,m)\in[K]\times[M]} \frac{28\Delta_{\max}^{k,m}\log(T)}{(f^{-1}(\Delta_{\min}^{k,m}))^2} + 8K^2M^2\Delta_{\max}.$$

□

第五节 BEACON-HT 线性奖励函数情形的证明

定理 4.2的证明 类似的, 我们将总的算法遗憾 $R_{\text{linear}}(T)$ 分解为探索损失 $R_{e,\text{linear}}(T)$, 通信损失 $R_{c,\text{linear}}(T)$, 以及其余部分 $R_{o,\text{linear}}(T)$, i.e.,

$$R_{\text{linear}}(T) = R_{e,\text{linear}}(T) + R_{c,\text{linear}}(T) + R_{o,\text{linear}}(T).$$

最后一项类似可以求得:

$$R_{o,\text{linear}}(T) \leqslant \left(\frac{K^2M}{K-M} + 2K \right) \Delta_c + K\Delta_{\max},$$

通信损失和探索损失分别在接下来的引理 4.5和引理 4.6中给出, 利用这三个结果即证定理。 □

一、通信损失: 线性奖励函数

引理 4.5 考虑线性奖励函数与总决策轮数 T, BEACON-HT 的通信损失 $R_{c,\text{linear}}(T)$ 有上界:

$$\begin{aligned}
&R_{c,\text{linear}}(T) \\
\leqslant& M^2K + (19M + 2M\log_2(K)) \left[2MK\log_2(T) + MK \left(\frac{3M\sqrt{3\log(T)}}{\sqrt{2}-1} + \frac{16KM^2}{3} \right) \right] \\
\leqslant& \frac{4}{\log 2} M^2K\log_2(K)\log(T) + \frac{38}{\log 2} M^2K\log(T) + o(\log(T)).
\end{aligned}$$

证明 我们阐述如下结论:

(i) 在阶段 1, 传输 $\tilde{\delta}_{k,m}^1$ 需要一步;

(ii) 对每个阶段 $r > 1$, 如果 $p_{k,m}^r > p_{k,m}^{r-1}$, 则 $\tilde{\delta}_{k,m}^r$ (期望意义下) 需要不多于 $3 + 2 \times \mathbb{E}[L_{k,m}^r] \leqslant 15$ 步传输;

(iii) 对每个阶段 $r > 1$, 传输下一轮要探索的分布组合和阶段长度需要不多于 $M(3 + 2\log_2(K)) + M$ 步传输。

这些结论来自一般奖励函数的证明, 故而自然对线性奖励函数成立。**然而, 对于线性奖励函数, 我们可以更精细的分析通信损失**:

$$
\begin{aligned}
R_{c,\text{linear}}(T) &\overset{(a)}{\leqslant} MK \times M \\
&\quad + \mathbb{E}\left[\sum_r (2 + V_* - V_{S^r}) \mathcal{I}\{\mathcal{E}_r\} \left[\sum_{(k,m)} 15\mathcal{I}\left\{p_{k,m}^r \geqslant p_{k,m}^{r-1}\right\} + M(3 + 2\log_2(K)) + M\right]\right] \\
&\quad + \mathbb{E}\left[\sum_r M\mathcal{I}\{\bar{\mathcal{E}}_r\} \left[\sum_{(k,m)} 15\mathcal{I}\left\{p_{k,m}^r \geqslant p_{k,m}^{r-1}\right\} + M(3 + 2\log_2(K)) + M\right]\right] \\
&\overset{(b)}{\leqslant} M^2K + \sum_r \mathbb{E}\left[(2 + V_* - V_{S^r})\mathcal{I}\{\mathcal{E}_r\} + M\mathcal{I}\{\bar{\mathcal{E}}_r\}\right](19M + 2M\log_2(K)) \\
&\overset{(c)}{\leqslant} M^2K + \sum_r \left(2 + 3M\sqrt{\frac{3\log(T)}{2^{p^r+1}}} + 4M\frac{KM}{(2^{p^r})^2}\right)(19M + 2M\log_2(K)) \\
&\leqslant M^2K + (19M + 2M\log_2(K))\left[2MK\log_2(T) + MK\sum_{p_r=0}^{\lceil\log_2 T\rceil}\left(3M\sqrt{\frac{3\log(T)}{2^{p^r+1}}} + 4\frac{KM^2}{(2^{p^r})^2}\right)\right] \\
&\leqslant M^2K + (19M + 2M\log_2(K))\left[2MK\log_2(T) + MK\left(3M\sqrt{3\log(T)}\frac{1}{\sqrt{2}-1} + \frac{16KM^2}{3}\right)\right]
\end{aligned}
$$

其中不等式 (a) 是因为在 $\mathcal{E}_r$ 事件发生的情况下, 至多有 2 用户 (*leader* 和一个 *follower*) 同时碰撞, 当 $\mathcal{E}_r$ 不发生, 一轮损失最大为 M。特别的, 如果使用 S^r 在第 r 个阶段进行通信, 一个通信步最多产生 $2 + V_* - V_{S^r}$ 的损失; 不等式 (b) 是因为在阶段 $r > 1$, 至多有 M 个分布的数据需要传输; 不等式 (c) 是因为, 当事件 $\mathcal{E}_r$ 发生, 我们有:

$$
\begin{aligned}
&\forall S \in \mathcal{S}_* \backslash \mathcal{S}_c, v(\bar{\boldsymbol{\mu}}_{S^r}^r) \geqslant v(\bar{\boldsymbol{\mu}}_S^r) \\
\Rightarrow &\forall S \in \mathcal{S}_* \backslash \mathcal{S}_c, V_{S^r} + M\left(2\sqrt{\frac{3\log t^r}{2^{p^r+1}}} + \frac{1}{2^{p^r+1}}\right) \geqslant v(\boldsymbol{\mu}_{S^r}) + \sum_{m\in[M]}\left(2\sqrt{\frac{3\log t^r}{2^{p^r_{s^r_m,m}+1}}} + \frac{1}{2^{p^r_{s^r_m,m}+1}}\right) \\
&\geqslant v(\bar{\boldsymbol{\mu}}_{S^r}^r) \geqslant v(\bar{\boldsymbol{\mu}}_S^r) > v(\boldsymbol{\mu}_S) = V_* \\
\Rightarrow &V_* - V_{S^r} \leqslant M\left(2\sqrt{\frac{3\log t^r}{2^{p^r+1}}} + \frac{1}{2^{p^r+1}}\right) \leqslant 3M\sqrt{\frac{3\log(T)}{2^{p^r+1}}};
\end{aligned}
$$

否则, 当事件不发生, 在一般奖励函数的证明中, 我们已经证明概率 $\mathbb{P}(\bar{\mathcal{E}}_r) \leqslant$

$\frac{4KM}{(2^{p^r})^2}$。 □

二、探索损失: 线性奖励函数

引理 4.6 考虑线性奖励函数与总决策轮数 T, BEACON-HT 的探索损失 $R_{e,\text{linear}}(T)$ 有上界:

$$R_{e,\text{linear}}(T) \leqslant \sum_{(k,m)} \frac{3727M}{\Delta_{\min}^{k,m}} \log(T) + 4K^2M^2\Delta_{\max}.$$

证明 我们首先介绍如下记号:

$$\begin{aligned} &S^* = [s_1^*, ..., s_M^*] \in \mathcal{S}_* \backslash \mathcal{S}_c \text{: 一个特定的无碰撞最优分布组合;} \\ &\Delta_{S^r} := V_* - V_{S^r}; \\ &[\tilde{M^r}] := \{m | m \in [M], s_m^r \neq s_m^*\}. \end{aligned} \tag{4.10}$$

Step I: 遗憾分解。首先, 我们将探索损失 $R_{e,\text{linear}}(T)$ 分解为:

$$\begin{aligned} R_{e,\text{linear}}(T) &= \mathbb{E}\left[\sum_r 2^{p^r}(V_* - V_{S^r})\right] \\ &= \mathbb{E}\left[\sum_r 2^{p^r}\Delta_{S^r}\mathcal{I}\left\{\mathcal{E}_r, \Delta_{S^r} > 0\right\}\right] + \mathbb{E}\left[\sum_r 2^{p^r}\Delta_{S^r}\mathcal{I}\left\{\bar{\mathcal{E}}_r, \Delta_{S^r} > 0\right\}\right] \\ &\overset{(a)}{\leqslant} \underbrace{\mathbb{E}\left[\sum_r 2^{p^r}\Delta_{S^r}\mathcal{I}\left\{\sum_{m\in[\tilde{M}^r]}\left(2\sqrt{\frac{3\log t^r}{2^{p^r_{s^r_m,m}+1}}} + \frac{1}{2^{p^r_{s^r_m,m}+1}}\right) \geqslant \Delta_{S^r}, \Delta_{S^r} > 0\right\}\right]}_{\text{term (C)}} \\ &\quad + \underbrace{\mathbb{E}\left[\sum_r 2^{p^r}\Delta_{S^r}\mathcal{I}\left\{\bar{\mathcal{E}}_r\right\}\right]}_{\text{term (D)}}, \end{aligned}$$

其中不等式 (a) 是因为当 $\mathcal{E}_r$ 发生, 选择一个次优的分布组合 S^r, 也就是 $\Delta_{S^r} > 0$,

表明

$$
\begin{aligned}
&\forall S \in \mathcal{S}_*, v(\bar{\boldsymbol{\mu}}^r_{S^r}) \geqslant v(\bar{\boldsymbol{\mu}}^r_S) \\
\Rightarrow & v(\bar{\boldsymbol{\mu}}^r_{S^r}) \geqslant v(\bar{\boldsymbol{\mu}}^r_{S*}) \\
\Rightarrow & \sum_{m\in[\tilde{M}^r]} \bar{\mu}^r_{s^r_m,m} \geqslant \sum_{m\in[\tilde{M}^r]} \bar{\mu}^r_{s^*_m,m} \\
\Rightarrow & \sum_{m\in[\tilde{M}^r]} \mu_{s^r_m,m} + \sum_{m\in[\tilde{M}^r]} \left(2\sqrt{\frac{3\log t^r}{2^{p^r_{s^r_m,m}+1}}} + \frac{1}{2^{p^r_{s^r_m,m}+1}}\right) \geqslant \sum_{m\in[\tilde{M}^r]} \bar{\mu}^r_{s^r_m,m} \\
& \geqslant \sum_{m\in[\tilde{M}^r]} \bar{\mu}^r_{s^*_m,m} \geqslant \sum_{m\in[\tilde{M}^r]} \mu_{s^*_m,m} \\
\Rightarrow & \sum_{m\in[\tilde{M}^r]} \left(2\sqrt{\frac{3\log t^r}{2^{p^r_{s^r_m,m}+1}}} + \frac{1}{2^{p^r_{s^r_m,m}+1}}\right) \geqslant \sum_{m\in[\tilde{M}^r]} \mu_{s^*_m,m} - \sum_{m\in[\tilde{M}^r]} \mu_{s^r_m,m} = V_* - V_{S^r} = \Delta_{S^r}.
\end{aligned}
\tag{4.11}
$$

Step II: Bounding term (D). 这里用到的技术和在一般奖励函数情形控制 (B) 是类似的, 我们可以得到

$$
\text{term (D)} = \mathbb{E}\left[\sum_r 2^{p^r}\Delta_{S^r}\mathcal{I}\left\{\bar{\mathcal{E}}_r\right\}\right] \leqslant 8K^2M^2\Delta_{\max} \leqslant 8K^2M^3.
$$

Step III: Bounding term (C). 首先我们记事件

$$
\mathcal{F}_r = \left\{\sum_{m\in[\tilde{M}^r]} \left(2\sqrt{\frac{3\log t^r}{2^{p^r_{s^r_m,m}+1}}} + \frac{1}{2^{p^r_{s^r_m,m}+1}}\right) \geqslant \Delta_{S^r}, \Delta_{S^r} > 0\right\},
$$

直觉上说, 这个事件代表 $\tilde{M}^r$ 中这些用户这轮采样的分布被采样的还不够多, 因此其置信半径与量化误差比起 Δ_{S^r} 还要更大。

$$
\begin{aligned}
\text{term (C)} &= \mathbb{E}\left[\sum_r 2^{p^r}\Delta_{S^r}\mathcal{I}\left\{\sum_{m\in[\tilde{M}^r]} \left(2\sqrt{\frac{3\log t^r}{2^{p^r_{s^r_m,m}+1}}} + \frac{1}{2^{p^r_{s^r_m,m}+1}}\right) \geqslant \Delta_{S^r}, \Delta_{S^r} > 0\right\}\right] \\
&= \mathbb{E}\left[\sum_r 2^{p^r}\Delta_{S^r}\mathcal{I}\left\{\mathcal{F}_r\right\}\right]
\end{aligned}
$$

利用 Kveton et al.[25] 中类似的技术, 我们引入两个降序的常数数列:

$$
1 = \beta_0 > \beta_1 > \beta_2 > \cdots > \beta_i > \cdots
$$

$$
\alpha_1 > \alpha_2 > \cdots > \alpha_i > \cdots
$$

使得 $\lim_{i\to\infty}\alpha_i = \lim_{i\to\infty}\beta_i = 0$; 进一步, 我们定义 q_{i,S^r} 为满足如下条件的整数:

$$2^{q_{i,S^r}-1} \leqslant \alpha_i \frac{M^2}{(\Delta_{S^r})^2}\log(T) < 2^{q_{i,S^r}} \leqslant 2\alpha_i \frac{M^2}{(\Delta_{S^r})^2}\log(T).$$

为简便起见, 我们记 $q_{0,S^r} = 0$ 且 $q_{\infty,S^r} = \infty$, 此外, 我们令 H_i^r 为

$$H_i^r = \left\{ m | m \in [\tilde{M}^r], p^r_{s^r_m,m} < q_{i,S^r} \right\},$$

其含义是, 相比起 q_{i,S^r}, 分布被采样的还不够。利用这些符号, 我们阶段 r 的事件:

$G_1^r = \{|H_1^r| \geqslant \beta_1 M\}$;

$G_2^r = \{|H_1^r| < \beta_1 M\} \cap \{|H_2^r| \geqslant \beta_2 M\}$;

$\cdots$

$G_i^r = \{|H_1^r| < \beta_1 M\} \cap \{|H_2^r| < \beta_2 M\} \cap \cdots \cap \left\{\left|H_{i-1}^r\right| < \beta_{i-1} M\right\} \cap \{|H_i^r| \geqslant \beta_i M\}$;

$\cdots$

这些事件是互斥的, 我们考虑如下的结论:

命题 4.7 令

$$\sqrt{14}\sum_{i=1}^{\infty}\frac{\beta_{i-1}-\beta_i}{\sqrt{\alpha_i}} \leqslant 1. \tag{4.12}$$

如果事件 $\mathcal{F}_r$ 在阶段 r 发生, 那么存在 G_i^r 发生。

这个结论可以利用反证法证明。我们现在考虑 $\bar{G}^r = \overline{\cup_i G_i^r}$, 我们有

$$\begin{aligned}
\bar{G}^r &= \overline{\cup_{i=1}^{\infty} G_i^r} \\
&= \cap_{i=1}^{\infty} \bar{G}_i^r \\
&= \cap_{i=1}^{\infty} \left[\overline{\left(\cap_{j=1}^{i-1}\left\{\left|H_j^r\right| < \beta_j M\right\}\right)} \cup \overline{\{|H_i^r| \geqslant \beta_i M\}} \right] \\
&= \cap_{i=1}^{\infty} \left[\left(\cup_{j=1}^{i-1}\overline{\left\{\left|H_j^r\right| < \beta_j M\right\}}\right) \cup \overline{\{|H_i^r| \geqslant \beta_i M\}} \right] \\
&= \cap_{i=1}^{\infty} \left[\left(\cup_{j=1}^{i-1}\left\{\left|H_j^r\right| \geqslant \beta_j M\right\}\right) \cup \{|H_i^r| < \beta_i M\} \right]
\end{aligned}$$

$$= \cap_{i=1}^{\infty} \left\{ |H_i^r| < \beta_i M \right\}.$$

如果 $\bar{G}^r$ 发生, 令 $\tilde{H}_i^r = [\tilde{M}^r] \backslash H_i^r$, 我们有 $\tilde{H}_{i-1}^r \subseteq \tilde{H}_i^r$ 以及 $[\tilde{M}^r] = \cup_i (\tilde{H}_i^r \backslash \tilde{H}_{i-1}^r)$, 这使得

$$\begin{aligned}
&\sum_{m \in [\tilde{M}^r]} \left(2\sqrt{\frac{3\log t^r}{2^{p^r_{s^r_m,m}}+1}} + \frac{1}{2^{p^r_{s^r_m,m}}+1} \right) \\
\leqslant & 3\sqrt{3\log t^r} \sum_{m \in [\tilde{M}^r]} \frac{1}{\sqrt{2^{p^r_{s^r_m,m}}+1}} \\
= & 3\sqrt{3\log t^r} \sum_{i=1}^{\infty} \sum_{m \in \tilde{H}_{i-1}^r \backslash \tilde{H}_i^r} \frac{1}{\sqrt{2^{p^r_{s^r_m,m}}+1}} \\
= & 3\sqrt{3\log t^r} \sum_{i=1}^{\infty} \frac{|\tilde{H}_i^r \backslash \tilde{H}_{i-1}^r|}{\sqrt{2^{q_{i-1,S^r}}+1}} \\
\leqslant & 3\sqrt{3\log t^r} \sum_{i=1}^{\infty} \frac{|\tilde{H}_i^r \backslash \tilde{H}_{i-1}^r|}{\sqrt{2\alpha_i \frac{M^2}{(\Delta_{S^r})^2}\log(T)}} \\
\leqslant & 3\sqrt{3/2} \frac{\Delta_{S^r}}{M} \sum_{i=1}^{\infty} \left(|H_{i-1}^r| - |H_i^r| \right) \frac{1}{\sqrt{\alpha_i}} \\
= & 3\sqrt{3/2} \frac{\Delta_{S^r}}{M} |H_0^r| \frac{1}{\sqrt{\alpha_1}} + 3\sqrt{3/2} \frac{\Delta_{S^r}}{M} \sum_{i=1}^{\infty} |H_i^r| \left(\frac{1}{\sqrt{\alpha_{i+1}}} - \frac{1}{\sqrt{\alpha_i}} \right) \\
\leqslant & 3\sqrt{3/2} \frac{\Delta_{S^r}}{M} \beta_0 M \frac{1}{\sqrt{\alpha_1}} + 3\sqrt{3/2} \frac{\Delta_{S^r}}{M} \sum_{i=1}^{\infty} \beta_i M \left(\frac{1}{\sqrt{\alpha_{i+1}}} - \frac{1}{\sqrt{\alpha_i}} \right) \\
< & \sqrt{14} \sum_{i=1}^{\infty} \frac{\beta_{i-1} - \beta_i}{\sqrt{\alpha_i}} \Delta_{S^r} \\
\leqslant & \Delta_{S^r},
\end{aligned} \tag{4.13}$$

这与 $\mathcal{F}_r$ 的定义矛盾: $\left\{ \sum_{m \in [\tilde{M}^r]} \left(2\sqrt{\frac{3\log t^r}{2^{p^r_{s^r_m,m}}+1}} + \frac{1}{2^{p^r_{s^r_m,m}}+1} \right) \geqslant \Delta_{S^r}, \Delta_{S^r} > 0 \right\}$。

利用命题 4.7, 我们分解 (C)

$$\text{term (C)} = \mathbb{E}\left[\sum_r 2^{p^r} \Delta_{S^r} \mathcal{I}\left\{ \mathcal{F}_r \right\} \right] \leqslant \mathbb{E}\left[\sum_r \sum_{i=1}^{\infty} 2^{p^r} \Delta_{S^r} \mathcal{I}\left\{ G_i^r, \Delta_{S^r} > 0 \right\} \right].$$

我们定义如下事件,

$$G_{i,k,m}^r = G_i^r \cap \left\{ m \in [\tilde{M}^r], s_m^r = k, p_{k,m}^r < q_{i,S^r} \right\},$$

这蕴含了

$$\mathcal{I}\left\{ G_i^r, \Delta_{S^r} > 0 \right\} \leqslant \frac{1}{\beta_i M} \sum_{(k,m): s_m^* \neq k} \mathcal{I}\left\{ G_{i,s_m^r,m}^r, \Delta_{S^r} > 0 \right\}$$

这是因为为了使得 G_i^r 发生, 至少有 $\beta_i M$ 与 $G_{i,k,m}^r$ 发生相关的分布。因此, 由于 $\mathcal{S}_b^{k,m} = \{S | S \in \mathcal{S}_b, s_m = k\} = \{S_1^{k,m}, ..., S_{N(k,m)}^{k,m}\}$, 我们可以得到

$$
\begin{aligned}
\text{term (C)} &= \mathbb{E}\left[\sum_r \sum_{i=1}^{\infty} 2^{p^r} \Delta_{S^r} \mathcal{I}\left\{G_i^r, \Delta_{S^r} > 0\right\}\right] \\
&\leqslant \mathbb{E}\left[\sum_r \sum_{i=1}^{\infty} 2^{p^r} \Delta_{S^r} \frac{1}{\beta_i M} \sum_{(k,m):s_m^* \neq k} \mathcal{I}\left\{G_{i,k,m}^r, \Delta_{S^r} > 0\right\}\right] \\
&\leqslant \mathbb{E}\left[\sum_r \sum_{i=1}^{\infty} 2^{p^r} \Delta_{S^r} \frac{1}{\beta_i M} \sum_{(k,m):s_m^* \neq k} \mathcal{I}\left\{m \in [\tilde{M}^r], s_m^r = k, p_{k,m}^r < q_{i,S^r}, \Delta_{S^r} > 0\right\}\right] \\
&= \mathbb{E}\left[\sum_{(k,m):s_m^* \neq k} \sum_{n=1}^{N(k,m)} \sum_r \sum_{i=1}^{\infty} 2^{p^r} \frac{1}{\beta_i M} \mathcal{I}\left\{s_m^r = k, p_{k,m}^r < q_{i,S_n^{k,m}}, S^r = S_n^{k,m}\right\} \Delta_n^{k,m}\right] \\
&= \mathbb{E}\left[\sum_{(k,m):s_m^* \neq k} \sum_{i=1}^{\infty} \underbrace{\sum_r \sum_{n=1}^{N(k,m)} 2^{p^r} \frac{1}{\beta_i M} \mathcal{I}\left\{s_m^r = k, p_{k,m}^r < q_{i,S_n^{k,m}}, S^r = S_n^{k,m}\right\} \Delta_n^{k,m}}_{\text{term (E)}}\right] \\
&\overset{(a)}{\leqslant} \mathbb{E}\left[\sum_{(k,m):s_m^* \neq k} \left[\sum_{i=1}^{\infty} \frac{6\alpha_i}{\beta_i}\right] \frac{M}{\Delta_{N(k,m)}^{k,m}} \log(T)\right]
\end{aligned}
\tag{4.14}
$$

其中不等式 (a) 是因为 (E) 有如下上界:

$$
\begin{aligned}
\text{term (E)} &= \sum_r \sum_{n=1}^{N(k,m)} 2^{p^r} \frac{1}{\beta_i M} \mathcal{I}\left\{s_m^r = k, p_{k,m}^r < q_{i,S_n^{k,m}}, S^r = S_n^{k,m}\right\} \Delta_n^{k,m} \\
&\leqslant 3 \times 2^{q_{i,S_1^{k,m}} - 1} \frac{\Delta_1^{k,m}}{\beta_i M} + \frac{1}{\beta_i M} \sum_{n=2}^{N(k,m)} \left(3 \times 2^{q_{i,S_n^{k,m}} - 1} - 3 \times 2^{q_{i,S_{n-1}^{k,m}} - 1}\right) \Delta_n^{k,m} \\
&\leqslant \frac{3\alpha_i M}{\beta_i \Delta_1^{k,m}} \log(T) + \frac{3\alpha_i M}{\beta_i} \sum_{n=2}^{N(k,m)} \left(\frac{1}{(\Delta_n^{k,m})^2} - \frac{1}{(\Delta_{n-1}^{k,m})^2}\right) \Delta_n^{k,m} \log(T) \\
&= \frac{3\alpha_i M}{\beta_i} \log(T) \left[\sum_{n=1}^{N(k,m)-1} \frac{\Delta_n^{k,m} - \Delta_{n+1}^{k,m}}{(\Delta_n^{k,m})^2} + \frac{1}{\Delta_{N(k,m)}^{k,m}}\right] \\
&\leqslant \frac{3\alpha_i M}{\beta_i} \log(T) \left[\sum_{n=1}^{N(k,m)-1} \frac{\Delta_n^{k,m} - \Delta_{n+1}^{k,m}}{\Delta_n^{k,m} \Delta_{n+1}^{k,m}} + \frac{1}{\Delta_{N(k,m)}^{k,m}}\right] \\
&\leqslant \frac{3\alpha_i M}{\beta_i} \log(T) \frac{2}{\Delta_{N(k,m)}^{k,m}}.
\end{aligned}
\tag{4.15}
$$

最后, 我们考虑如下的优化问题来给出 α_i 和 β_i 的选择:

$$
\text{minimize} \sum_{i=1}^{\infty} \frac{6\alpha_i}{\beta_i}
$$

subject to $\lim_{i\to\infty}\alpha_i = \lim_{i\to\infty}\beta_i = 0$

Monotonicity: $1 = \beta_0 > \beta_1 > \cdots > \beta_i > \cdots ; \alpha_1 > \alpha_2 > \cdots > \alpha_i > \cdots$

Eqn. (4.12): $\sqrt{14}\sum_{i=1}^{\infty}\frac{\beta_{i-1}-\beta_i}{\sqrt{\alpha_i}} \leqslant 1.$

如果我们选择 α_i 和 β_i 为几何分布序列 (Kveton et al.[25]), 具体的说 $\alpha_i = d(\alpha)^i$ 且 $\beta_i = (\beta)^i$, 其中 $0 < \alpha, \beta < 1$ 与 $d > 0$ 成立, 进一步, 如果 $\beta \leqslant \sqrt{\alpha}$, 为了使得式子 (4.12)成立, 我们需要

$$\sqrt{14}\sum_{i=1}^{\infty}\frac{\beta_{i-1}-\beta_i}{\sqrt{\alpha_i}} = \sqrt{14}\sum_{i=1}^{\infty}\frac{(\beta)^{i-1}-(\beta)^i}{\sqrt{d(\alpha)^i}} = \sqrt{\frac{14}{d}}\frac{1-\beta}{\sqrt{\alpha}-\beta} \leqslant 1 \Rightarrow d \geqslant 14\left(\frac{1-\beta}{\sqrt{\alpha}-\beta}\right)^2.$$

因此 d 的最佳选择为: $d = 14\left(\frac{1-\beta}{\sqrt{\alpha}-\beta}\right)^2$, 此时问题重写为

$$\text{minimize } \sum_{i=1}^{\infty}\frac{6(\alpha)^i}{(\beta)^i} = 84\left(\frac{1-\beta}{\sqrt{\alpha}-\beta}\right)^2\frac{\alpha}{\beta-\alpha}$$

$$\text{conditioned on } 0 < \alpha < \beta < \sqrt{\alpha} < 1.$$

数值计算得到:$\alpha = 0.1459$ 与 $\beta = 0.2360$, 我们有 $\sum_{i=1}^{\infty}\frac{6(\alpha)^i}{(\beta)^i} \leqslant 3727$。因此

$$\text{term (C)} \leqslant \mathbb{E}\left[\sum_{(k,m):s_m^*\neq k}\left[\sum_{i=1}^{\infty}\frac{6\alpha_i}{\beta_i}\right]\frac{M}{\Delta_{N(k,m)}^{k,m}}\log(T)\right] \leqslant \sum_{(k,m):s_m^*\neq k}\frac{3727M}{\Delta_{N(k,m)}^{k,m}}\log(T)$$
$$\leqslant \sum_{(k,m)}\frac{3727M}{\Delta_{\min}^{k,m}}\log(T).$$

利用 (C), (D) 的上界, 我们能够证明引理 4.6。 □

第五章　No sensing MPMAB: 推广到无碰撞示性函数情况

在这一章中, 我们将讨论如何将之前两章所设计的可观察模型下的算法推广到不可观察情形。出于避免重复的考虑, 我们并不完整地重复整个框架, 相反, 我们专注于通信部分, 考虑如何在无法观察到碰撞的示性函数时 (对应于有噪声信道), 利用编码理论完成可靠的信息传输。

第一节　算法发展

我们将在同质模型下将模型推广到不可观察的情形, 对于非同质模型, 其推广与证明的技术是类似的。

一、带噪声信道的可靠通信

在可观察情况下, 用户利用碰撞的示性函数来实现 $0-1$ 比特的传输; 在不可观察的情况下, 我们不再能够观察到碰撞的示性函数, 此时我们事实上是利用反馈的奖励是否为 0 来区分 $0-1$ 比特。具体的说, 当反馈大于 0, 我们可以无差错的识别出此时没有发生碰撞, 当反馈等于 0, 我们则无法区分出它是由于采样结果为 0 或者是发生了碰撞。如果我们将传输一个比特这个过程看成一个黑箱, 我们知道在可观察情况下相当于每传输一比特需要一次采样的代价, 而如果我们能够在不可观察情况下等价的实现这个过程, 或者至少以大概率实现这个过程, 那么类似的理论结果自然将得到保持。这个替代品我们需要引入的编码技巧, 在有噪声情况下, 通过信息论中的编码技巧, 我们能够以高概率实现一个信息比特的正确传输, 此时通信代价从 1 次采样增长到编码长度 N。

二、信道建模

如果假定 $\mu_{\min} = \min_{k=1,\dots,K} \mu_k$, 则观察到采样 0 的概率小于等于 $(1-\mu_{\min})$, 如果考虑长度为 N 连续的 0 反馈, 则它们由采样导致的概率小于等于 $(1-\mu_{\min})^N$。这启发我们, 如果当 $\mu_{\min} > 0$ 的时候, 我们可以用连续的比特串来区分反馈 0 来自采样还是碰撞。

事实上, 从编码角度来说, 这个假设是本质的。我们可以将这样一个通信的过程建模成一个 Z-信道, 其在图片 1.1中描述。可以看到, 当反馈大于 0 时, 我们可以无差错的解码出无碰撞这一信息 (对应于比特 0), 但是当反馈为 0 时, 它可能来自碰撞也可能来自采样。根据香农定理, 在码率小于信道容量时我们可以以任意小的误码率进行数据传输, 因此我们首先给出 Z-信道的信道容量:

引理 5.1 对于一个由 $0 \to 1$ 的交叉概率为 q 的 Z-信道, 其信道容量为:

$$C_Z(q) = \log_2(1 + (1-q)q^{q/(1-q)}). \tag{5.1}$$

可以看到, 当 $q = 0/1$ 的时候, 信道容量均 0, 此时任何编码方案都不能以任意小的误码率进行数据传输, 因此从编码理论来看, $\mu_{\min} > 0$ 是本质的。但是, 从最小化算法遗憾的角度来考虑, 这个假设是不必要的, 我们会在最后一小节回到这个话题, 这里我们假定 $\mu_{\min} > 0$ 且对于每个用户都是已知的。

三、编码方案

Z-信道是一个熟知的信道, 许多可靠的编码方案已经因此被提出, 我们在这里介绍三种典型的编码方案并推导其所需要的编码长度, 我们假设要发送的信息为 $\boldsymbol{m}$, 它是一个 $0/1$ 比特串; 编码后的 $0/1$ 比特串为 X; 接收方接到的比特串为 Y; 编码的最小单位长度为 Q, 例如, $Q = 1$ 代表我们对每一个比特的信息进行编码, $Q = 2$ 代表我们把信息按 2 比特的长度分组后再进行编码。

重复码 (repetition code)。重复码是一个非常简单但是极其高效的编码方案,

Chen et al.[39] 证明在 $Q=1$ 的时候它是最优的, 其编码方案与解码方案如下:

- 编码: 将信息 $\boldsymbol{m}$ 中比特 $0/1$ 重复 A 次获得编码后的比特串 X。
- 解码: 对于待解码的 Y, 如果存在 $Y[i]\neq 0$, 则解码为 0, 否则解码为 1。

当交叉概率不大于 $1-\mu_{\min}$ 时, 重复码的比特差错概率为

$$P(Y_i\neq X_i)<(1-\mu_{\min})^A.$$

对一个长度为 Q 比特的信息, 其差错概率为

$$\begin{aligned}P_e&=P(\exists i,Y_i\neq X_i)\\&=1-P(Y_i=X_i)^Q\\&\leqslant 1-(1-(1-\mu_{\min})^A)^Q\\&\leqslant Qe^{-\mu_{\min}A}.\end{aligned}$$

因此, 当我们选择 $A=\lceil\frac{\log(QT)}{\mu_{\min}}\rceil$, 则 $P_e<\frac{1}{T}$, 此时 Q 比特信息的编码总长度为

$$N_{rep}=Q\left\lceil\frac{\log(QT)}{\mu_{\min}}\right\rceil.$$

反转码 (flip code)。相比于重复码, Chen et al.[39] 提出的反转码在 $Q>1$ 时能更好地利用 Z-信道的特征, 我们在 $Q=2$ 的情况下给出其编码方案与解码方案如下:

- 编码: 我们将每 2 比特编码为 $2A$ 比特, 编码函数如下:

$$\begin{aligned}&(0,0)\to(\underbrace{1,...,1}_{A},\underbrace{1,...,1}_{A});\ (0,1)\to(\underbrace{1,...,1}_{A},\underbrace{0,...,0}_{A});\\&(1,0)\to(\underbrace{0,...,0}_{A},\underbrace{1,...,1}_{A});\ (1,1)\to(\underbrace{0,...,0}_{A},\underbrace{0,...,0}_{A}).\end{aligned}$$

- 解码: 反转码的解码方案类似重复码, 假定待解码的码字 $\boldsymbol{m}$ 长度为 $2A$, 我们把它分解为长度均为 A 的码字 $\boldsymbol{m_1},\boldsymbol{m_2}$, 之后解码函数为:

- 如果 $\boldsymbol{m_1}$ 与 $\boldsymbol{m_2}$ 的所有比特都是 1, 则解码为 $(0,0)$;
- 如果 $\boldsymbol{m_1}$ 的所有比特都是 1, 且 $\boldsymbol{m_2}$ 中存在 0, 则解码为 $(0,1)$;
- 如果 $\boldsymbol{m_1}$ 中存在 0, 且 $\boldsymbol{m_2}$ 的所有比特都是 1, 则解码为 $(1,0)$;
- 否则, 解码为 $(1,1)$;

当交叉概率不大于 $1-\mu_{\min}$ 时, 反转码的比特差错概率为

$$P(Y_i \neq X_i) \leqslant (1-\mu_{\min})^A - \frac{1}{4}(1-\mu_{\min})^{2A}$$

不等式成立是因为 $q^A - \frac{1}{4}q^{2A}$ 在 $q \in [0,1]$ 时单调上升, 对于长度为偶数 Q 的信息 (我们可以添加一个比特 0 使得奇数长信息变为偶数长), 其差错概率为

$$\begin{aligned}
P_e &= P(\exists i, Y_i \neq X_i) \\
&= 1 - P(Y_i = X_i)^{\frac{Q}{2}} \\
&\leqslant 1 - (1 - (1-\mu_{\min})^A + \frac{1}{4}(1-\mu_{\min})^{2A})^{\frac{Q}{2}} \\
&= 1 - (1 - \frac{1}{2}(1-\mu_{\min})^A)^Q \\
&\leqslant \frac{Q}{2}(1-\mu_{\min})^A \\
&\leqslant \frac{Q}{2}e^{-\mu_{\min}A}.
\end{aligned}$$

因此, 当我们选择 $A = \lceil \frac{\log(QT/2)}{\mu_{\min}} \rceil$, 则 $P_e < \frac{1}{T}$, 此时 Q 比特信息的编码总长度为

$$N_{flip} = Q\lceil \frac{\log(QT/2)}{\mu_{\min}} \rceil.$$

汉明码 (Hamming code)。在 $Q = 4$ 的情况下, 我们可以设计一个修改后的 $(7,4)$ 汉明码, 它是一个复合的编码, 其内部编码为标准的 $(7,4)$ 汉明码, 外部编码是重复码。

- 编码: 我们首先用标准的 $(7,4)$ 汉明编码矩阵 $\boldsymbol{G}$ 来将 4 比特的信息编码为 7 比特串, 之后我们把这个 7 比特串按照重复码方法编码为 $7A$ 长度的比特串。

- 解码: 与编码相反, 我们首先按照重复码的解码方式将 $7A$ 长的待解码信息解码为 7 比特串, 之后我们用标准的汉明解码矩阵 $\boldsymbol{H}$ 来进行解码。

$$\boldsymbol{G}=\begin{pmatrix}1&1&0&1\\1&0&1&1\\1&0&0&0\\0&1&1&1\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix},\ \boldsymbol{H}=\begin{pmatrix}1&0&1&0&1&0&1\\0&1&1&0&0&1&1\\0&0&0&1&1&1&1\end{pmatrix}$$

外部的重复码将交叉概率从 q 改善为 q^A, 根据 Barbero et al.[40], 当交叉概率为 q^A, 解码的比特差错率为

$$P(Y_i\neq X_i)=\frac{7}{2}(q^A)^2+o((q^A)^3).$$

之后, 我们忽略 $o((q^A)^3)$ 小项, 此时传输 Q 比特 (类似的, 我们假设 Q 模 4 为 0) 的信息的误码率为

$$\begin{aligned}P_e&=P(\exists i,Y_i\neq X_i)\\&=1-P(Y_i=X_i)^{\frac{Q}{4}}\\&=1-(1-\frac{7}{2}q^{2A})^{\frac{Q}{4}}\\&\leqslant 1-(1-\frac{7}{2}(1-\mu_{\min})^{2A})^{\frac{Q}{4}}\\&\leqslant\frac{7Q}{8}(1-\mu_{\min})^{2A}\\&\leqslant\frac{7Q}{8}e^{-2\mu_{\min}A}.\end{aligned}\tag{5.2}$$

因此, 当我们选择 $A=\frac{1}{2}\lceil\frac{\log(7QT/8)}{\mu_{\min}}\rceil$, 则 $P_e<\frac{1}{T}$, 此时 Q 比特信息的编码总长度为

$$N_{ham}=\frac{7Q}{8}\left\lceil\frac{\log(7QT/8)}{\mu_{\min}}\right\rceil,$$

第二节　理论分析

一、带有额外假设时的理论结果

在本节中, 我们讨论如何把编码理论在不可观察情况下与之前的算法结合, 我们介绍带有额外假设的已知结果, 并讨论如何弱化与消除所作的假设, 这一节的结果来自 Shi et al.[21] 与 Huang et al.[23]。为了使得结果对于一般的最优编码方案成立, 我们首先介绍来自信息论的概念 *error exponent*, Gallager[41]:

定理 5.2　考虑一个离散无记忆信道, 如果码率 R 小于信道容量 C, 则存在一个码字长度为 N 的无反馈编码方案, 使得它的误码率满足:

$$P_e \leqslant \exp[-NE_r(R)],$$

其中 $E_r(R)$ 是关于码率 R 的随机误码率指数。

将 BEACON-HM 中的通信替换为编码通信, 在不使用 ADC 通信技巧的情况下, 我们得到了来自论文 Shi et al.[21] 的 EC-SIC, 其理论结果为:

定理 5.3　假定 $\mu_{\min} > 0$ 以及 $\Delta = \mu_{(M+1)} - \mu_{(M)} > 0$ 且均已知, 对于交叉概率为 $1 - \mu_{\min}$ 的 Z-信道, 任意使得其随机误码率指数达到 $E(\mu_{\min})$ 的最优编码方案, 在 $\epsilon \in (0, \frac{\Delta}{4})$ 的情况下, 我们有

$$\begin{aligned} R(T) \leqslant\ & c_1 MK \frac{\log(T)}{\mu_{\min}} \\ & + c_2 \frac{\Delta}{4\epsilon} \left(\sum_{k>M} \min \left\{ \frac{\log(T)}{\mu_{(M+1)} - \mu_{(k)} + 4\epsilon}, \sqrt{T \log(T)} \right\} \right) \\ & + c_3 N' \left(M^2(K+2) \log \left(\min \left\{ \frac{1}{4\epsilon}, T \right\} \right) + M^2 K \right) \end{aligned} \tag{5.3}$$

其中 c_1, c_2 与 c_3 是常数, 且 $N' = \max\{\frac{Q}{C_Z(1-\mu_{\min})}, \frac{1}{E(\mu_{\min})} \log(T)\}$.

二、假设的不必要性

在这一小节中, 我们讨论为何我们可以去除 $\Delta > 0$ 与 $\mu_{\min} > 0$ 的假设。

$\Delta > 0$。在 EC-SIC 的原始论文中, 作者假设 $\Delta > 0$ 是为了确定量化比特的个数, 使得由于截断实数样本均值为有限比特表示所引起的误差不会对区分两

个均值差别为 Δ 的分布带来影响。然而这是不必要的, 由于抽样本身带有不确定性, 其由置信半径表征, 我们只需要简单地将每轮的量化比特数控制到使得量化误差与置信半径同阶即可, 由此引入的量化误差将能够被当作抽样的不确定性的一部分一同进行分析处理。

$\mu_{\min} > 0$。在上一节中, 我们指出, 对于 Z-信道, $\mu_{\min} > 0$ 是我们能够进行编码的必要条件。然而, 我们真正想要优化的是算法的理论遗憾, 我们讨论以下两种情况。1) $\mu_{(1)} = 0$。此时任何分布都无法作为通信的分布, 然而, 此时的理论遗憾将恒为 0, 此时无论是什么算法, 其都是最优的。2) $\mu_{(L)} = 0, K \leqslant L < 1$。此时有一部分的分布无法作为通信的分布, Huang et al.[23] 巧妙地在初始化阶段进行了一项分布选择, 以 $O(\log T)$ 的代价找出能够作为通信的分布, 此后只需要让所有用户以这些被挑选的分布进行通信即可。

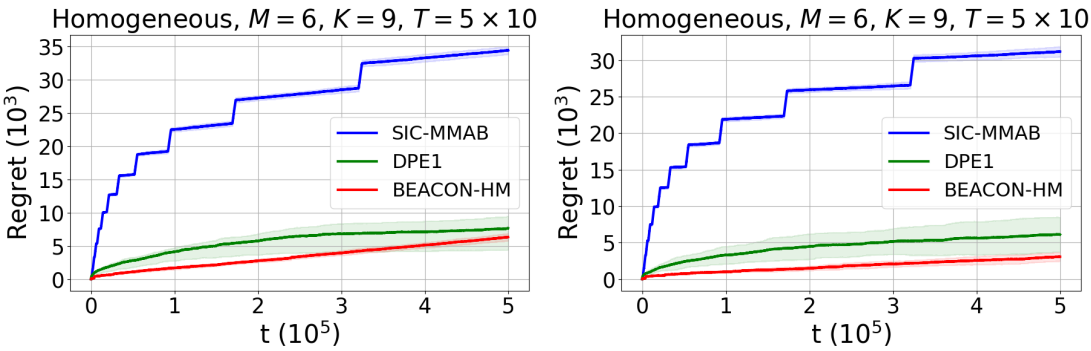
第六章　数值实验

第一节　实验结果

在这一节中我们利用数值实验来比较我们所提出的算法与当前最优的算法，所有实验结果基于 100 次独立重复实验求取平均。

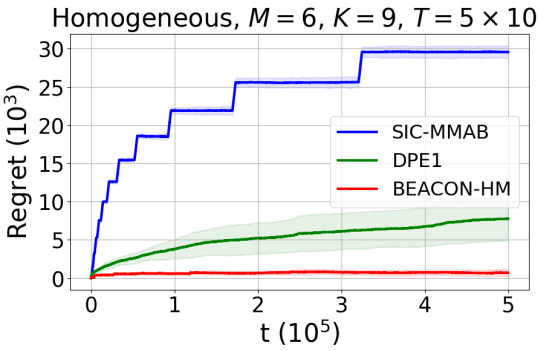
一、同质化模型实验

在这一节中，我将 `BEACON-HM` 与 Boursier et al.[4] 提出的 `SIC-MMAB`，Wang et al.[9] 提出的 `DPE1` 进行比较。



(a) 最优分布组合唯一。

(b) 多个最优分布组合。



(c) 均为最优分布组合。

图 6.1　同质化模型下算法遗憾比较。

我们考虑三个实例，图片 6.1(a)对应的均值向量均匀分布于 0.89 到 0.9 之间，因此只有一个唯一的最优分布组合；图片 6.1(b)对应的均值向量为 $(0.91, 0.9, ..., 0.9, 0.89)$，

也就是只有第 1 和第 K 个分布均值不同, 其余分布的均值都为 0.9; 最后, 图片 6.1(c)对应的分布的均值都为 0.9, 此时所有分布组合都是最优的。

可以看到, 即使在最优分布组合不唯一的图片 6.1(a)中, 由于自适应差分通信能够显著减小通信损失, BEACON-HM 要显著的优于 SIC-MMAB, 且与当前的最优算法 DPE1 表现类似; 在图片 6.1(b)与图片 6.1(c)中, 此时最优分布组合不唯一, 通信将始终进行, 此时 BEACON-HM 的优势更加明显, 特别的, 在最后一个实验中, 此时探索损失始终为 0, BEACON-HM 的算法遗憾亦接近于 0, 相反, 由于通信损失的原因, DPE1 与 SIC-MMAB 没能逼近中心化算法的结果 (严格的 0 损失)。

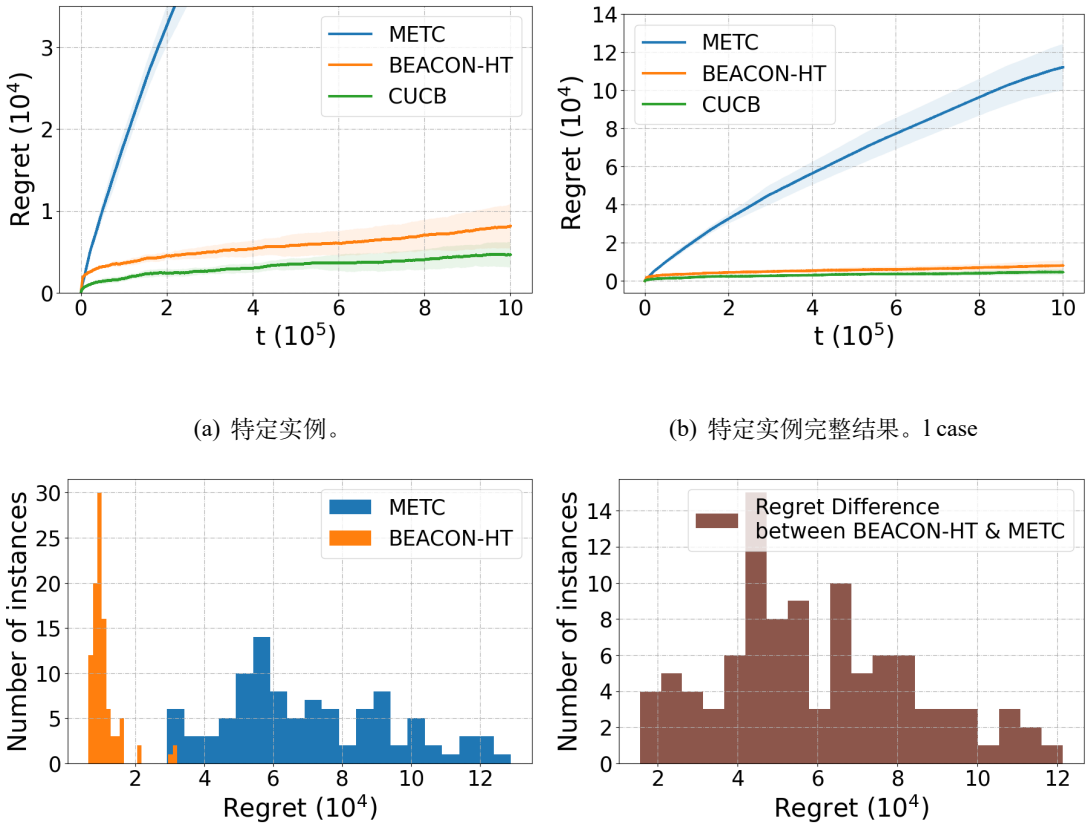
二、非同质化模型实验

在这一节中, 我们比较 BEACON-HT 与 Boursier et al.[16] 提出的 METC, 特别的, 为了说明我们能够接近中心化算法, 我们还与 Chen et al.[24] 提出的 CUCB 进行比较。

线性奖励函数。我们首先考虑一个具体的问题实例, 其样本的均值矩阵为:

$$\begin{bmatrix} 0.5 & 0.49 & 0.39 & 0.29 & 0.5 & 0.39 \\ 0.5 & 0.49 & 0.39 & 0.29 & 0.1 & 0.39 \\ 0.29 & 0.19 & 0.5 & 0.499 & 0.3 & 0.39 \\ 0.29 & 0.49 & 0.5 & 0.5 & 0.3 & 0.39 \\ 0.49 & 0.49 & 0.49 & 0.49 & 0.5 & 0.49 \end{bmatrix}$$

图片 6.2(a)与 6.2(b)汇报了这个实验的结果, 为了使得比较结果更清晰, 在前者中我们截去了 $y > 3.5 \times 10^4$ 的部分。可以看到, METC 的算法遗憾是 BEACON-HT 的十五倍左右, 特别的, 我们看到 BEACON-HT 的表现接近了中心化算法 CUCB, 这验证了 BEACON-HT 的探索高效性以及自适应差分通信算法的损失与探索近似同阶。除此之外, 我们随机生成了 100 个 $M = 5, K = 6$ 的问题实例, 并统计了两个算法在这些实例上的算法遗憾统计, 其频数统计图为图片 6.2(c), 此外, 我们



(a) 特定实例。

(b) 特定实例完整结果。1 case

(c) 随机生成实例。

(d) 随机生成实例下的算法遗憾差。

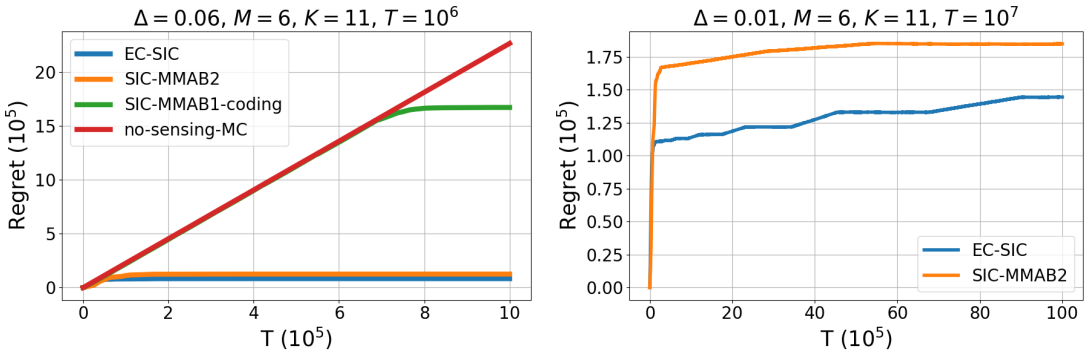
图 6.2　非同质化模型下算法遗憾比较。

将对应的实例二者的遗憾作差, 并绘制差分的频数统计图于图片 6.2(d)。可以看见, BEACON-HT 具有极好的鲁棒性, 在所有生成的实例都要优于 METC, 这再次显示了 BEACON-HT 的优越性。

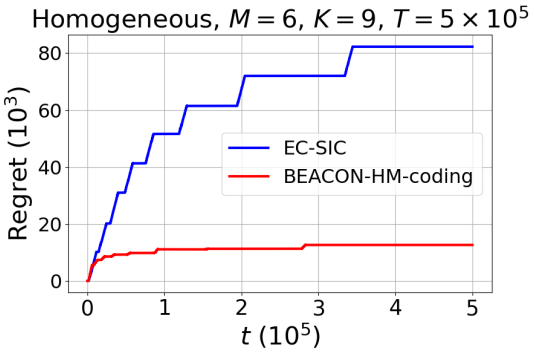
三、不可观察模型实验

在这一小节里, 我们汇报 Shi et al.[21] 提出的 EC-SIC, Boursier et al.[4] 提出的 SIC-MMAB2 算法的结果, 此外, 我们将编码技巧与 SIC-MMAB1 以及 Rosenski et al.[37] 提出的 Musical Chair 结合起来得到 SIC-MMAB1-coding 与 MC-coding。最后, 我们比较 BEACON-HM-coding 与 EC-SIC 来说明差分通信算法的重要性。

我们首先考虑了一个简单的和一个较为困难实例, 其中 $\Delta = \mu_{(M)} - \mu_{(M+1)}$ 分别等于 0.06 与 0.01; 之后, 我们考虑所有分布具有相同的均值 0.9 的示例。图



(a) 较容易的实例。　　(b) 困难的实例。



(c) 所有分布均为最优。

图 6.3　不可观察模型下算法遗憾比较。

片 6.3(a)与图片 6.3(b)分别汇报了简单实例与困难实例的结果, 我们可以看到使用编码技巧的 EC-SIC 在简单实例中与 SIC-MMAB2 表现类似而在困难问题中更优, 此外, EC-SIC 与 SIC-MMAB2 要明显比其余算法更有优势。图片 6.3(c)汇报了当所有分布均值均为 0.9 的结果, 此时算法遗憾由通信损失决定, 可以看到在使用了自适应差分通信后, BEACON-HM-coding 的算法遗憾只有 EC-SIC 的七分之一。

第七章　结论

在这篇论文中,我们研究了三种典型的分布式多用户多臂老虎机模型(MPMAB),并提出一个统一的算法框架: BEACON,其在三种设置下的实例化不论是在理论还是数值实验中都取得了最优的结果。BEACON 的主要算法设计思想在于通过高效的隐式通信与批量探索结构来模仿逼近中心化算法,这种算法设计思路在分布式算法设计中具有广泛的应用可能性;特别的,批量探索结构可以反过来启发中心化算法的设计,例如,应用批量探索结构的 CUCB 能够将 Oracle 复杂度从 $O(T)$ 降低到 $O(\log T)$。

在理论结果之外,我们仍然想讨论这个算法在算法设计与对这个领域本身的意义。从算法哲学上讲,我们的算法依赖于对中心化算法的模仿,其核心部分为我们可以利用有意的碰撞进行信息交换,而精细的通信结构设计使得这部分的损失比起探索损失是同阶或低阶的,这样的算法设计思路在分布式算法设计领域有广泛的应用,例如 Ye et al.[42] 所提出的分布式优化算法 PMGT-VR 就是基于类似的算法设计思想;从另一方面讲,这个算法没有太多的实际应用意义,我们可以看到算法依赖于严苛的同步与无差错假定,一旦隐式通信存在差错,整个算法都会因此而错乱从而导致线性损失。它更大的意义是作为一个理论算法证明: (1) 当前的理论下界是紧的; (2) 不能显式通信的假定是非本质的。我们希望这个算法能够启发同领域的研究者考虑如何设计在能够显式通信情况下设计分布式探索的方案以及如何在隐式通信不可能实现的情况下,进行算法的设计。

参 考 文 献

[1] Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002, 47(2-3): 235–256.

[2] Garivier A, Cappé O. The KL-UCB algorithm for bounded stochastic bandits and beyond// Computational Learning Theory. 2011: 359–376.

[3] Liu K, Zhao Q. Distributed learning in multi-armed bandit with multiple players. *IEEE Trans. Signal Process.*, 2010, 58(11): 5667–5681.

[4] Boursier E, Perchet V. SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits//Advances in Neural Information Processing Systems. 2019: 12048–12057.

[5] Anandkumar A, Michael N, Tang A K, et al. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 2011, 29(4): 731–745.

[6] Avner O, Mannor S. Concurrent bandits and cognitive radio networks//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2014: 66–81.

[7] Rosenski J, Shamir O, Szlak L. Multi-player bandits–a musical chairs approach//International Conference on Machine Learning. 2016: 155–163.

[8] Besson L, Kaufmann E. Multi-player bandits revisited//Algorithmic Learning Theory. 2018: 56–92.

[9] Wang P A, Proutiere A, Ariu K, et al. Optimal algorithms for multiplayer multi-armed bandits// Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. 2020.

[10] Kalathil D, Nayyar N, Jain R. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 2014, 60(4): 2331–2345.

[11] Nayyar N, Kalathil D, Jain R. On regret-optimal learning in decentralized multiplayer multi-armed bandits. *IEEE Transactions on Control of Network Systems*, 2016, 5(1): 597–606.

[12] Bistritz I, Leshem A. Distributed multi-player bandits-a game of thrones approach//Advances in Neural Information Processing Systems. 2018: 7222–7232.

[13] Bistritz I, Leshem A. Game of thrones: Fully distributed learning for multiplayer bandits. *Mathematics of Operations Research*, 2020.

[14] Magesh A, Veeravalli V V. Multi-user mabs with user dependent rewards for uncoordinated spectrum access//2019 53rd Asilomar Conference on Signals, Systems, and Computers. 2019: 969–972.

[15] Tibrewal H, Patchala S, Hanawal M K, et al. Multiplayer multi-armed bandits for optimal

assignment in heterogeneous networks. *arXiv:1901.03868*, 2019.

[16] Boursier E, Kaufmann E, Mehrabian A, et al. A practical algorithm for multiplayer bandits when arm means vary among players//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. 2020.

[17] Bistritz I, Baharav T, Leshem A, et al. My fair bandit: Distributed learning of max-min fairness with multi-player bandits//International Conference on Machine Learning. 2020: 930–940.

[18] Avner O, Mannor S. Multi-user lax communications: a multi-armed bandit approach//The 35th Annual IEEE International Conference on Computer Communications. 2016: 1–9.

[19] Darak S J, Hanawal M K. Multi-player multi-armed bandits for stable allocation in heterogeneous ad-hoc networks. *IEEE Journal on Selected Areas in Communications*, 2019, 37(10): 2350–2363.

[20] Lugosi G, Mehrabian A. Multiplayer bandits without observing collision information. *arXiv:1808.08416*, 2018.

[21] Shi C, Xiong W, Shen C, et al. Decentralized multi-player multi-armed bandits with no collision information//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. 2020.

[22] Bubeck S, Budzinski T. Coordination without communication: optimal regret in two players multi-armed bandits. *Conference on Learning Theory*, 2020.

[23] Huang W, Combes R, Trinh C. Towards optimal algorithms for multi-player bandits without collision sensing information. *arXiv:2103.13059*, 2021.

[24] Chen W, Wang Y, Yuan Y. Combinatorial multi-armed bandit: General framework and applications//International Conference on Machine Learning. 2013: 151–159.

[25] Kveton B, Wen Z, Ashkan A, et al. Tight regret bounds for stochastic combinatorial semi-bandits//Artificial Intelligence and Statistics. 2015: 535–543.

[26] Combes R, Talebi Mazraeh Shahi M S, Proutiere A, et al. Combinatorial bandits revisited//Advances in neural information processing systems: volume 28. 2015: 2116–2124.

[27] Degenne R, Perchet V. Combinatorial semi-bandit with known covariance//Advances in Neural Information Processing Systems. 2016: 2972–2980.

[28] Kveton B, Wen Z, Ashkan A, et al. Matroid bandits: fast combinatorial optimization with learning//Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence. 2014: 420–429.

[29] Talebi M S, Proutiere A. An optimal algorithm for stochastic matroid bandit optimization//Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems. 2016: 548–556.

[30] Wang S, Chen W. Thompson sampling for combinatorial semi-bandits//International Confer-

ence on Machine Learning. 2018: 5114–5122.

[31] Perrault P, Boursier E, Perchet V, et al. Statistical efficiency of thompson sampling for combinatorial semi-bandits//Advances in Neural Information Processing Systems. 2020.

[32] Kveton B, Wen Z, Ashkan A, et al. Combinatorial cascading bandits//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1. 2015: 1450–1458.

[33] Merlis N, Mannor S. Tight lower bounds for combinatorial multi-armed bandits. *arXiv:2002.05392*, 2020.

[34] Bubeck S, Cesa-Bianchi N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Machine Learning*, 2012, 5(1): 1–122.

[35] Anantharam V, Varaiya P, Walrand J. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—part i: I.i.d. rewards. *IEEE Transactions on Automatic Control*, 1987, 32: 968 – 976.

[36] Slivkins A. Introduction to multi-armed bandits. *Found. Trends Machine Learning*, 2019, 12 (1-2): 1–286.

[37] Rosenski J, Shamir O, Szlak L. Multi-player bandits - a musical chairs approach//Proceedings of the 33nd International Conference on Machine Learning: volume 48. 2016: 155–163.

[38] Shi C, Xiong W, Shen C, et al. Heterogeneous multi-player multi-armed bandits: Closing the gap and generalization. *Advances in neural information processing systems*, 2021, 34: 22392–22404.

[39] Chen P N, Lin H Y, Moser S M. Optimal ultrasmall block-codes for binary discrete memoryless channels: volume 59. 2013: 7346–7378.

[40] Barbero A, Ellingsen P, Spinsante S, et al. Maximum likelihood decoding of codes on the Z-channel//IEEE International Conference on Communications: volume 3. 2006: 1200–1205.

[41] Gallager R G. Information theory and reliable communication. USA: John Wiley & Sons, 1968.

[42] Ye H, Xiong W, Zhang T. PMGT-VR: A decentralized proximal-gradient algorithmic framework with variance reduction. *arXiv:2012.15010*, 2020.