对话式上下文老虎机的高效探索性关键词选择策略

Zhiyong Wang,[1] **Xutong Liu,**[1] **Shuai Li,**[2*] **John C.S. Lui**[1]

香港中文大学，中国香港特别行政区[2]上海交通大学，中国上海 zywang21@cse.cuhk.edu.hk、liuxt@cse.cuhk.edu.hk、shuaili8@sjtu.edu.cn、cslui@cse.cuhk.edu.hk

arXiv:2303.00315v2 [cs.LG] 1 Oct 2023

摘要

对话式上下文老虎机通过偶尔询问用户对关键词的显式反馈来获取用户偏好，从而加速学习。然而，现有方法的某些方面限制了其性能。首先，关键词级对话和臂级推荐中获取的信息未能得到恰当整合，因而无法加快学习。其次，提出探索性关键词以快速发掘用户在不同领域的潜在兴趣，对于加速用户偏好估计的收敛至关重要，但现有研究从未考虑这一点。为解决这些问题，我们首先提出"ConLinUCB"，这是一个适用于对话式老虎机的通用框架，能够更好地整合信息，在每个时刻将臂级和关键词级反馈结合起来，一步估计用户偏好。在此框架基础上，我们进一步设计了两种采用探索式关键词选择策略的老虎机算法：ConLinUCB-BS 和 ConLinUCB-MCR。我们证明了所提出算法具有更紧的遗憾上界。具体而言，ConLinUCB-BS 的遗憾界为 $\tilde{O}(d\sqrt{T}\log T)$，优于此前的结果 $O(d\sqrt{T}\log T)$。在合成数据和真实世界数据上进行的大量实验表明，与经典的 ConUCB 算法相比，我们的算法在学习准确性（最高提升 54%）和计算效率（最高提升 72%）方面具有显著优势，展现出其对推荐系统的潜在益处。

1 引言

如今，推荐系统已广泛应用于各个领域。传统在线推荐系统的学习速度通常较慢，因为发现用户偏好需要进行大量探索。为加快学习过程并提供更加个性化的推荐，对话式推荐系统（对话式推荐系统）应运而生（Christakopoulou、Radlinski 和 Hofmann 2016年；Christakopoulou 等人 2018年；Sun 和 Zhang 2018年；Zhang 等人 2018年；Li 等人 2021年；Gao 等人 2021年）。在对话式推荐系统中，学习智能体会不时询问用户对某些"关键词"的明确反馈，并利用这些额外的对话信息，更好地了解用户偏好（Zhang 等人，2020；Xie et al. 2021）。

尽管对话式推荐系统近期取得了成功，但使用对话式上下文老虎机方法设计推荐系统仍存在一些关键局限。这些局限包括：(a) 从关键词级对话和臂级推荐中获得的信息未能得到妥善整合以加快学习，因为这两个阶段本质上被假定为用户偏好相同，却分别进行估计（Zhang 等人，2020；Xie et al. 2021；Wu et al. 2021）；(b) 使用传统关键词进行的查询具有局限性，探索性也不足。具体而言，如果某个关键词迄今探索不足，且系统对用户在与之相关的物品上的偏好尚不确定，我们就称该关键词具有"探索性"。询问用户对探索性关键词的反馈，可以高效地了解她在不同领域（如体育、科学）的潜在兴趣。这意味着我们能够快速估计用户偏好向量在特征空间各个方向上的取值，从而加快学习速度。因此，设计探索性关键词选择策略至关重要，而现有研究尚未考虑这一问题。

受上述考虑的启发，我们提出设计对话式老虎机算法：(i) 同时利用臂级和关键词级交互来估计用户偏好，从而恰当地整合这两个层面所获得的信息；(ii) 在进行对话时采用有效策略选择探索性关键词，以便快速推断用户偏好。

为了更好地利用来自推荐和对话的交互反馈，我们提出了 ConLinUCB，这是一个适用于对话式老虎机的通用框架，并可灵活采用不同的关键词选择策略。ConLinUCB 通过求解一个单一的优化问题来估计用户偏好向量，同时最小化臂级估计奖励和关键词级估计反馈的均方误差，而不像以往的工作那样在不同层面分别进行估计。这样一来，来自这两个层面的信息便能更好地结合起来，从而指导学习。

基于这一 ConLinUCB 框架，我们设计了两种采用探索性关键词选择策略的新算法：ConLinUCB-BS 和 ConLinUCB-MCR。

- ConLinUCB-BS利用包含线性无关向量的重心张成集，该集合可作为老虎机问题中高效的探索基（Amballa、Gupta 和 Bhat，2021 年）。每当允许进行一次对话时，

*通讯作者。版权所有 © 2023，人工智能促进协会 (www.aaai.org)。保留所有权利。

ConLinUCB-BS便会从给定关键词集合 $\mathcal{K}$的预先计算好的重心张成集 $\mathcal{B}$ 中，均匀随机地选择一个探索性关键词。

•ConLinUCB-MCR 适用于更一般的情形，即关键词集合可以随时间变化，并且能够利用交互历史自适应地选择探索性关键词。请注意，在老虎机环境中，我们通常使用置信半径来自适应地评估某个臂是否已得到充分探索；每当一个臂被选中时，其置信半径就会缩小（Lat- timore 和 Szepesv´ari，2020年）。这意味着探索性关键词应具有较大的置信半径。基于这一推理，ConLinUCB-MCR 在进行对话时，会选择置信半径最大的、最具探索性的关键词。

借助探索性对话，我们的算法能够快速 elicitate 用户偏好，从而提供更好的推荐。例如，如果关键词 体育 在轮次 t具有探索性，这表明智能体目前尚不确定用户是否偏好与体育 相关的物品（例如篮球、排球），那么它会直接询问用户对体育的反馈，并据此进行推荐。通过这种方式，智能体能够快速为用户找到合适的物品。我们证明了所提算法的遗憾上界优于经典的 ConUCB 算法。

总而言之，本文的贡献如下：

• 我们提出了一种新的通用对话式上下文老虎机框架 ConLinUCB，能够高效地整合从推荐和对话中获取的交互信息。

•基于 ConLinUCB，我们设计了两种采用探索性关键词选择策略的新算法：ConLinUCB-BS 和 ConUCB-MCR，它们能够加快用户偏好估计的收敛速度。

•我们证明了所提出的算法能够达到紧致的遗憾上界。特别是，ConLinUCB-BS 的上界为 $O(d\sqrt{T}\log\ T)$，优于对话式老虎机文献中此前的 $O(d\sqrt{T}\log\ T)$。

• 合成数据和真实世界数据上的实验验证了所提出算法在学习准确性（最高提升 54%）和计算效率（最高提升 72%）方面的优势[1]。

2 相关工作

我们的工作与对话式上下文老虎机研究的关系最为密切。

上下文线性老虎机是一种在线序贯决策问题：在每个时间步，智能体都必须选择一个动作，并获得相应的奖励，其期望值是动作的未知线性函数（Li 等人，2010；Chu 等人，2011；Abbasi-Yadkori、P´al 和 Szepesv´ari，2011；Wu 等人，2016）。目标是在 T 轮次中尽可能多地获取奖励。

传统线性老虎机需要进行大量探索，才能捕捉推荐系统中的用户偏好。为了加快在线推荐，对话式上下文老虎机的概念最早由 Zhang 等人，2020 提出。在这种方法中，利用围绕关键词的对话反馈来辅助用户偏好引导。在该研究中，他们提出了 ConUCB 算法，其理论遗憾界为 $O(d\sqrt{T}\log\ T)$。一些后续研究借助额外信息尝试提升 ConUCB 的性能，例如自生成关键词（Wu et al. 2021）、相对反馈（Xie et al. 2021）和知识图（Zhao et al. 2022）。与这些研究不同，我们采用与 ConUCB 相同的问题设定，并在不依赖额外信息的情况下改进其底层机制。不过，在有额外信息可用时，也可以运用本文提出的高效信息整合和探索性对话原则来增强这些研究，这留待未来进一步探讨。

3 问题设定

本节介绍对话式上下文老虎机的问题设定。假设存在一个有限的 $\mathcal{A}$ 臂集合。每个臂 $a \in \mathcal{A}$ 代表一个待推荐的物品，并关联一个特征向量 $\boldsymbol{x}_a \in \mathbb{R}^d$。不失一般性，假设特征向量已归一化，即 $\|\boldsymbol{x}_a\|_2 = 1$、 $\forall a \in \mathcal{A}$。智能体与用户进行交互，共$T \in \mathbb{N}_+$ 轮次；用户对物品的偏好由一个未知的向量 $\boldsymbol{\theta}^* \in \mathbb{R}^d$、 $\|\boldsymbol{\theta}^*\|_2 \leq 1$表示。

在每一轮 $t = 1$、2、.....、T，智能体都可以从可用臂子集 $\mathcal{A}_t \subseteq \mathcal{A}$ 中进行选择。根据历史交互，智能体选择一个臂 $a_t \in \mathcal{A}_t$，并获得相应的奖励 $r_{a_t,\ t} \in [0, 1]$。假设该奖励是上下文向量的线性函数

$$r_{a_t,t} = \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}^* + \epsilon_t\,, \tag{1}$$

其中 ϵ_t 是均值为零的 1-次高斯随机噪声。

令 $a_t^* \in \text{argmax}_{a\in\mathcal{A}_t} \boldsymbol{x}_a^\top \boldsymbol{\theta}^*$ 表示在 t 处具有最大期望奖励的最优臂。学习目标是最小化累积遗憾

$$R(T) = \sum_{t=1}^{T} \boldsymbol{x}_{a_t^*}^\top \boldsymbol{\theta}^* - \sum_{t=1}^{T} \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}^*. \tag{2}$$

智能体还可以偶尔询问用户对某些对话关键词的反馈，以帮助了解用户偏好。具体而言，“关键词”是指与某个臂子集相关的关键词或主题。例如，关键词 体育 与篮球、橄榄球、游泳等臂相关。

假设存在一个有限的关键词集合 $\mathcal{K}$。臂与关键词之间的关系由加权二分图 $(\mathcal{A},\ \mathcal{K},\ \boldsymbol{W})$ 给出，其中 $\boldsymbol{W} \triangleq [w_{a,k}]_{a\in\mathcal{A},\ k\in\mathcal{K}}$ 表示臂与关键词之间的关系，即关键词 $k \in \mathcal{K}$ 以权重 $w_{a,\ k} \geq 0$与臂 $a \in \mathcal{A}$ 相关联。我们假设每个关键词 k 都与某些相关臂具有正权重（即 $\sum_{a\in\mathcal{A}} w_{a,\ k} > 0,\ \forall k \in \mathcal{K}$），并且与每个臂相关联的权重之和为 1，即 $\sum_{k\in\mathcal{K}} w_{a,\ k} = 1,\ \ a \in \mathcal{A}$。关键词 k的特征向量定义为 $\tilde{\boldsymbol{x}}_k = \sum_{a\in\mathcal{A}} \frac{w_{a,\ k}}{\sum_{a'\in\mathcal{A}} w_{a',\ k}} \boldsymbol{x}_a$。时刻 t 针对关键词 k 的关键词级反馈定义为

$$\tilde{r}_{k,t} = \tilde{\boldsymbol{x}}_k^\top \boldsymbol{\theta}^* + \tilde{\epsilon}_t\,, \tag{3}$$

[1]代码见 https://github.com/ZhiyongWangWzy/ ConLinUCB。

其中 $\tilde{\epsilon}_t$ 被假设为 1-次高斯随机噪声。需要强调的是，在先前的研究中（Zhang 等人，2020；Wu et al. 2021；Xie et al. 2021；Zhao et al. 2022），基本假设是，未知的用户偏好向量 $\boldsymbol{\theta}^*$ 在臂级别和关键词级别上本质相同。

为避免影响用户体验，智能体不应过于频繁地进行对话。参照（Zhang 等人，2020），我们定义函数 b：$\mathbb{N}_+ \to \mathbb{R}_+$，其中 $b(t)$ 随 t 增大，以控制智能体的对话频率。在每个轮次 t，如果 $b(t) - b(t-1) > 0$，智能体便可以通过询问用户对 $q(t)$ 个关键词的反馈，进行 $q(t) = \lfloor b(t) - b(t-1) \rfloor$ 次对话。采用这种建模方式后，直到轮次 t，智能体将与用户进行 $b(t)$ 次对话交互。

4 算法与理论分析

本节首先介绍 ConLinUCB，这是一种能够更好地整合信息的对话式老虎机框架，适用于“任何”关键词选择策略。基于 ConLinUCB，我们进一步提出了两种老虎机算法：ConLinUCB-BS 和 ConLinUCB-MCR，它们采用了探索式关键词选择策略。

为简化表述，我们在算法1中合并了 ConLinUCB 框架、ConLinUCB-BS 和 ConLinUCB-MCR。我们还从理论上给出了所提出算法的遗憾界。

4.1 通用 ConLinUCB 算法框架

在对话式老虎机中，通常会假设未知的用户偏好向量 $\boldsymbol{\theta}^*$ 在臂级别和关键词级别本质上是相同的（Zhang 等人，2020；Xie 等人，2021；Wu 等人，2021）。然而，现有工作都以不同方式处理$\boldsymbol{\theta}^*$ 在这两个级别上的情况。具体而言，它们采取两种不同的步骤来估计臂级别和关键词级别的用户偏好向量，并使用折扣参数$\lambda \in (0, 1)$来平衡从这两个级别交互中进行的学习。通过这种方式，臂级信息和关键词级信息对估计收敛的贡献分别被 λ 和 $1-\lambda$折扣。因此，这种折扣会造成观测数据的浪费，表明无法充分利用这两个级别的信息来加速学习过程。

为了解决上述问题，我们提出了一种适用于对话式上下文老虎机的通用框架，称为 ConLinUCB。在这一新框架中，为了充分利用来自两个层面的交互信息，我们通过求解一个单一的优化问题，同时估计用户偏好向量，该问题旨在最小化臂级估计奖励和关键词级估计反馈的均方误差。具体而言，在 ConLinUCB 中，轮次 t 的用户偏好向量通过求解以下线性回归来估计

$$\boldsymbol{\theta}_t = \mathop{\arg\min}_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{\tau=1}^{t-1} (\boldsymbol{x}_{a_\tau}^\top \boldsymbol{\theta} - r_{a_\tau,\tau})^2 + \sum_{\tau=1}^{t} \sum_{k \in \mathcal{K}_\tau} (\tilde{\boldsymbol{x}}_k^\top \boldsymbol{\theta} - \tilde{r}_{k,\tau})^2 + \beta \|\boldsymbol{\theta}\|_2^2 , \tag{4}$$

算法1： 通用 ConLinUCB 框架

输入： 图($\mathcal{A}$、$\mathcal{K}$、$\boldsymbol{W}$)、对话频率函数 $b(t)$、关键词选择策略 $\boldsymbol{\pi}$。

初始化：$\boldsymbol{M}_0 = \beta \boldsymbol{I}$、$\boldsymbol{b}_0 = \boldsymbol{0}$。
对于$t = 1$ 到 T，**执行**
  如果 $b(t) - b(t-1) > 0$，**则**
    $q(t) = \lfloor b(t) - b(t-1) \rfloor$;
    当 $q(t) > 0$ **时**
      使用以下方法选择一个关键词 $k \in \mathcal{K}$ 指定的关键词选择策略 $\boldsymbol{\pi}$（例如，ConLinUCB-BS 使用公式（9），而 ConLinUCB-MCR 使用公式（10）），并查询用户对它的偏好;
      接收用户的反馈 $\tilde{r}_{k,t}$;
      $\boldsymbol{M}_t = \boldsymbol{M}_{t-1} + \tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^\top$;
      $\boldsymbol{b}_t = \boldsymbol{b}_{t-1} + \tilde{\boldsymbol{x}}_k \tilde{r}_{k,t}$;
      $q(t) -= 1$;
    end
  else
    $\boldsymbol{M}_t = \boldsymbol{M}_{t-1}, \boldsymbol{b}_t = \boldsymbol{b}_{t-1}$;
  end
  $\boldsymbol{\theta}_t = \boldsymbol{M}_t^{-1} \boldsymbol{b}_t$;
  选择 $a_t = \arg\max_{a \in \mathcal{A}_t} \boldsymbol{x}_a^\top \boldsymbol{\theta}_t + \alpha_t \|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}$;
  询问用户对臂 a_t 的偏好并接收奖励$r_{a_t,t}$;
  $\boldsymbol{M}_t = \boldsymbol{M}_{t-1} + \boldsymbol{x}_{a_t} \boldsymbol{x}_{a_t}^\top$;
  $\boldsymbol{b}_t = \boldsymbol{b}_{t-1} + \boldsymbol{x}_{a_t} r_{a_t,t}$;
结束

其中，$\mathcal{K}_\tau$ 表示轮次 τ 提出的关键词集合，系数 $\beta > 0$ 用于控制正则化。该优化问题的闭式解为

$$\boldsymbol{\theta}_t = \boldsymbol{M}_t^{-1} \boldsymbol{b}_t , \tag{5}$$

其中

$$\begin{aligned} \boldsymbol{M}_t &= \sum_{\tau=1}^{t-1} \boldsymbol{x}_{a_\tau} \boldsymbol{x}_{a_\tau}^\top + \sum_{\tau=1}^{t} \sum_{k \in \mathcal{K}_\tau} \tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^\top + \beta \boldsymbol{I} , \\ \boldsymbol{b}_t &= \sum_{\tau=1}^{t-1} \boldsymbol{x}_{a_\tau} r_{a_\tau,\tau} + \sum_{\tau=1}^{t} \sum_{k \in \mathcal{K}_\tau} \tilde{\boldsymbol{x}}_k \tilde{r}_{k,\tau} . \end{aligned} \tag{6}$$

为平衡探索与利用，ConLinUCB 采用上置信界（UCB）策略选择臂

$$a_t = \mathop{\arg\max}_{a \in \mathcal{A}_t} \underbrace{\boldsymbol{x}_a^\top \boldsymbol{\theta}_t}_{\hat{R}_{a,t}} + \underbrace{\alpha_t \|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}}_{C_{a,t}} , \tag{7}$$

其中，$\|\boldsymbol{x}\|_{\boldsymbol{M}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}}$、$\hat{R}_{a,t}$ 和 $C_{a,t}$ 分别表示第 t 轮时臂 a 的估计奖励和置信半径，且

$$\alpha_t = \sqrt{2 \log(\frac{1}{\delta}) + d \log(1 + \frac{t + b(t)}{\beta d})} + \sqrt{\beta} , \tag{8}$$

该结论由以下引理1得出。

ConLinUCB 算法框架如算法1所示。关键词级交互发生在第3至14行。在第 t 轮，智能体首先使用 $b(t)$ 判断是否允许进行对话。进行对话时，智能体会询问用户对 $q(t)$ 个关键词的反馈，并利用反馈更新参数。第15至20行概述了臂级交互。智能体基于历史交互计算估计的 $\boldsymbol{\theta}^*$，选择 UCB 指数最大的臂，接收相应的奖励，并据此更新参数。ConLinUCB 仅维护一组协方差矩阵 $\boldsymbol{M}_t$ 和回归向量 $\boldsymbol{b}_t$，其中包含臂级交互和关键词级交互的反馈。通过这种方式，ConLinUCB 比 ConUCB 更充分地利用了反馈信息。请注意，ConLinUCB 是一个通用框架，具体的关键词选择策略 $\boldsymbol{\pi}$ 尚待确定。

4.2 使用重心张成集中的关键词的 ConLinUCB（ConLinUCB-BS）

基于 ConLinUCB 框架，我们提出了采用探索性关键词选择策略的 ConLinUCB-BS 算法。具体而言，ConLinUCB-BS 从关键词集合$\mathcal{K}$的重心张成集 $\mathcal{B}$ 中选择关键词，以进行探索性对话。重心张成集是在线学习中一种高效的探索基（Amballa、Gupta 和 Bhat，2021年）。下面给出关键词集合 $\mathcal{K}$的重心张成集的正式定义。

定义1（$\mathcal{K}$的重心张成集）子集 $\mathcal{B}=\{k_1,k_2,...,k_d\}\subseteq\mathcal{K}$ 是 $\mathcal{K}$ 的重心张成集，当且仅当对于任意 $k\in\mathcal{K}$，都存在一组系数 $\boldsymbol{c}\in[-1,1]^d$，使得 $\tilde{\boldsymbol{x}}_k=\sum_{i=1}^d \boldsymbol{c}_i\tilde{\boldsymbol{x}}_{k_i}$。

我们假设关键词集合 $\mathcal{K}$ 是有限的，且 $\{\tilde{\boldsymbol{x}}_k\}_{k\in\mathcal{K}}$ 张成 $\mathbb{R}^d$，因此可以保证存在 $\mathcal{K}$ 的重心张成集 $\mathcal{B}$（Awerbuch 和 Kleinberg，2008）。

重心张成集中的对应向量是线性无关的。通过从重心张成集中选择关键词，我们可以沿着不同方向快速探索未知的用户偏好向量 $\boldsymbol{\theta}^*$。基于这一推理，只要允许进行对话，ConLinUCB-BS 就会选择一个关键词

$$k\sim\text{unif}(\mathcal{B}),\tag{9}$$

也就是说，从 $\mathcal{K}$ 的重心张成集 $\mathcal{B}$ 中均匀随机地抽取一个关键词 k。在 ConLinUCB 框架中，ConLinUCB-BS 使用上述策略 $\boldsymbol{\pi}$ 完成（算法1）。如下文引理1和引理2 所示，在 ConLinUCB-BS 中，统计估计不确定性会迅速缩小。此外，由于关键词集合 $\mathcal{K}$ 的重心张成集 $\mathcal{B}$ 可以离线预先计算，ConLinUCB-BS 的计算效率很高，这对实时推荐至关重要。

4.3 采用最大置信半径关键词的 ConLinUCB（ConLinUCB-MCR）

我们可以从以下几个方面进一步改进 ConLinUCB-BS。首先，ConLinUCB-BS 不适用于关键词集合 $\mathcal{K}$ 随时间变化的更一般情形，因为它需要预先计算 $\mathcal{K}$的重心张成集 $\mathcal{B}$。其次，由于关键词的选择独立于过去的观察结果，ConLinUCB-BS 未能充分利用历史信息。例如，假设智能体根据先前的交互已经确定用户是否偏好体育。在这种情况下，它就不必再询问用户对关键词体育的反馈。为解决这些问题，我们提出了 ConLinUCB-MCR 算法，该算法 (i) 适用于关键词集合 $\mathcal{K}$ 随 t 变化的情形，且 (ii) 能够根据历史交互自适应地开展探索性对话。

在多臂老虎机中，置信半径用于衡量某个臂在交互历史中是否已得到充分探索，并且每当该臂被选中时，置信半径就会缩小。受此启发，如果某个关键词的置信半径较大，就意味着系统尚未充分探索用户对相关物品的偏好，这表明该关键词具有探索性。基于这一推理，ConLinUCB-MCR 会选择置信半径最大的关键词，自适应地开展探索性对话。具体而言，当允许在 t 进行对话时，ConLinUCB-MCR 会选择一个关键词，如下所示

$$k\in\arg\max_{k\in\mathcal{K}_t}\alpha_t\left\|\tilde{\boldsymbol{x}}_k\right\|_{\boldsymbol{M}_t^{-1}},\tag{10}$$

其中，α_t 在公式（8）中定义，$\mathcal{K}_t\subseteq\mathcal{K}$ 表示轮次 t 时可用的、可能随时间变化的关键词集合。ConLinUCB-MCR 采用上述策略（公式（10）），作为 ConLinUCB（算法 1）中的 $\boldsymbol{\pi}$ 完成。

4.4 理论分析

我们给出算法遗憾的上界。按照惯例，对话频率满足 $b(t)\le t$，因此我们假设 $b(t)=b\cdot t, b\in(0,1)$。由于篇幅限制，引理1-2和定理3-4的证明放在附录中。

以下引理给出了基于 ConLinUCB 的算法中，沿动作向量 $\boldsymbol{x}_a$ 方向上 $\boldsymbol{\theta}_t$ 与 $\boldsymbol{\theta}^*$ 之差的高概率上界。

引理1 当 $\forall t$ 时，对于任意 $a\in\mathcal{A}$，以至少 $1-\delta$ 的概率，其中某个 $\delta\in(0,1)$

$$\left|\boldsymbol{x}_a^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)\right|\le\alpha_t\left\|\boldsymbol{x}_a\right\|_{\boldsymbol{M}_t^{-1}}=C_{a,t},$$

where $\alpha_t=\sqrt{2\log(\frac{1}{\delta})+d\log(1+\frac{t+b(t)}{\beta d})}+\sqrt{\beta}.$

对于关键词集合 $\mathcal{K}$ 的重心张成集 $\mathcal{B}$，令

$$\lambda_{\mathcal{B}}:=\lambda_{\min}(\boldsymbol{E}_{k\sim\text{unif}(\mathcal{B})}[\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top])>0,\tag{11}$$

其中 $\lambda_{\min}(\cdot)$ 表示该增广矩阵的最小特征值。我们可以得到以下引理，它给出了 ConLinUCB-BS 的 $\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}$ 的高概率上界。

引理 2 对于 *ConLinUCB-BS*，$\forall a\in\mathcal{A}$，在 $\forall t\ge t_0=\frac{256}{b\lambda_{\mathcal{B}}^2}\log(\frac{128d}{\lambda_{\mathcal{B}}^2\delta})$时，以至少 $1-\delta$ 的概率，对于 $\delta\in(0,\frac{1}{8}]$，

$$\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}}\le\sqrt{\frac{2}{\lambda_{\mathcal{B}}bt}}.$$

下面的定理给出了我们的 ConLinUCB-BS 的高概率遗憾上界。

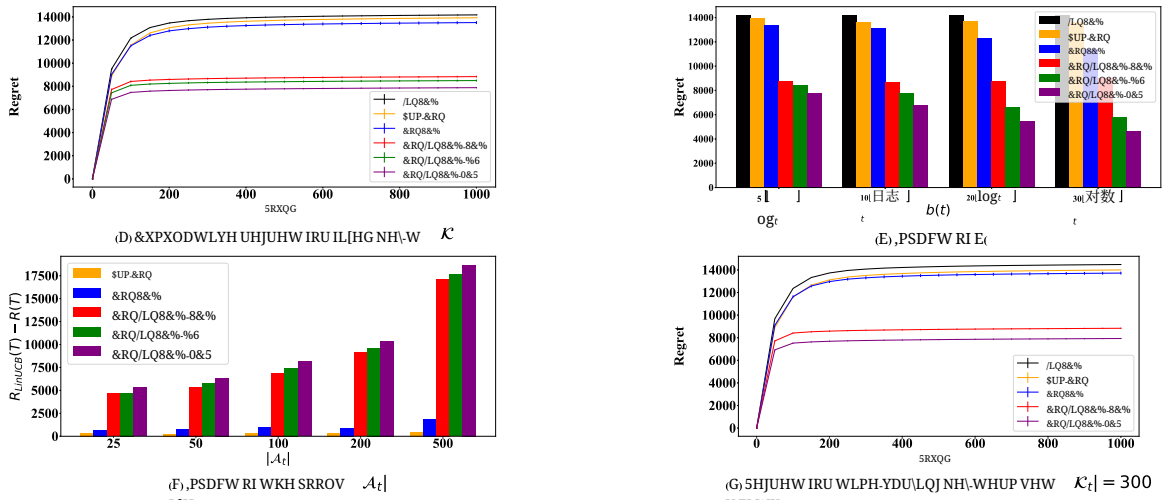


图1：合成数据集上的实验结果

定理3 以至少 $1-\delta$ 的概率，对于某个 $\delta \in (0, \frac{1}{4}]$，*ConLinUCB-BS* 的遗憾 $R(T)$ 满足

$$R(T) \leq 4\sqrt{\frac{2}{b\lambda_{\mathcal{B}}}}\sqrt{T}\left(\sqrt{2\log(\frac{2}{\delta}) + d\log(1+\frac{(1+b)T}{\beta d})} + \sqrt{\beta}\right) + \frac{256}{b\lambda_{\mathcal{B}}^2}\log(\frac{256d}{\lambda_{\mathcal{B}}^2\delta}) + 1\,.$$

回顾一下，ConUCB 的遗憾上界（Zhang等人，2020年）为

$$R(T) \leq 2\sqrt{2Td\log(1+\frac{\lambda(T+1)}{(1-\lambda)d})}\left(\sqrt{\frac{1-\lambda}{\lambda}} + \sqrt{\frac{1-\lambda}{\lambda\beta}}\sqrt{2\log(\frac{2}{\delta}) + d\log(1+\frac{bT}{\beta d})} + \sqrt{2\log(\frac{2}{\delta}) + d\log(1+\frac{\lambda T}{(1-\lambda)d})}\right),$$

其为 $O(d\sqrt{T}\log T)$。定理3中给出的 ConLinUCB-BS 的遗憾界为 $O(d\sqrt{T}\log T)$（因为 $\lambda_{\mathcal{B}}$ 的阶为 $O(\frac{1}{d})$），优于 ConUCB，减少了一个乘法因子 $\sqrt{\log T}$。

接下来，以下定理给出了 ConLinUCB-MCR 的高概率遗憾上界。

定理4 以至少 $1-\delta$ 的概率，对于某个 $\delta \in (0,\ 1)$，*ConLinUCB-MCR* 的遗憾 $R(T)$ 满足

$$R(T) \leq 2\sqrt{2Td\log(1+\frac{T+1}{\beta d})} \times \left(\sqrt{\beta} + \sqrt{2\log(\frac{1}{\delta}) + d\log(1+\frac{(b+1)T}{\beta d})}\right).$$

请注意，通过减少一些加性项，ConLinUCB-MCR 的遗憾上界小于 ConUCB。

5 合成数据集上的实验

本节展示合成数据上的实验结果。为了获得离线预计算的重心张成集 $\mathcal{B}$，我们使用了 (Awerbuch 和 Kleinberg 2008) 提出的方法。

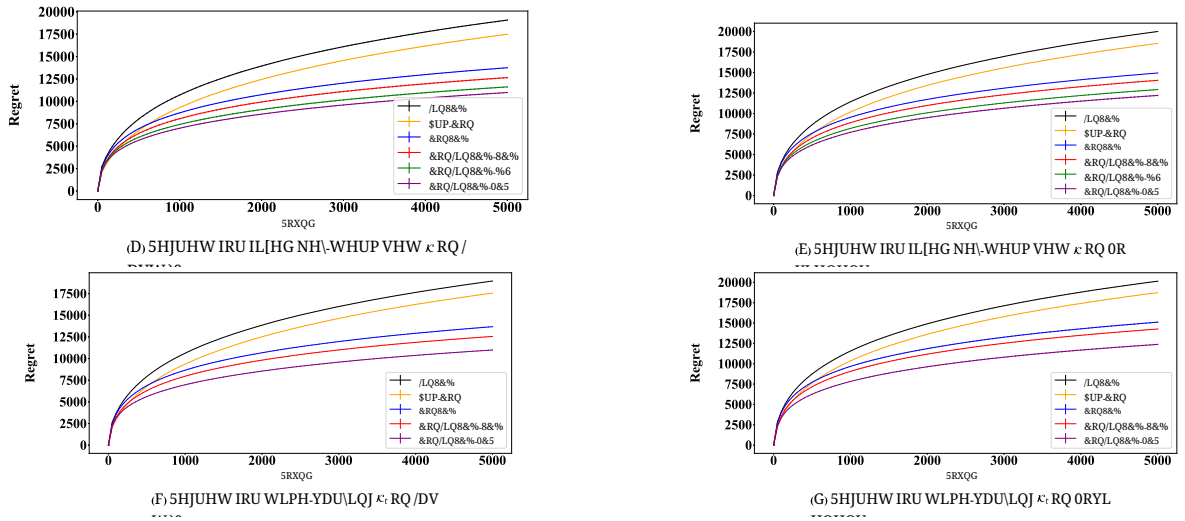
5.1 实验设置

合成数据集的生成。 我们创建一组臂 $\mathcal{A}$，其中包含 $|\mathcal{A}| = 5{,}000$ 个臂，以及一组关键词 $\mathcal{K}$。我们将特征空间的维度设为 $d = 50$，用户数量设为 $N_u = 200$。

对于每个用户偏好向量 $\boldsymbol{\theta}_u^*$ 和每个臂特征向量 $\boldsymbol{x}_a$，每个分量均通过从标准正态分布 $\mathcal{N}(0,1)$ 中独立抽样生成，并将所有这些向量归一化，使得 $\|\boldsymbol{\theta}_u^*\|_2 = 1$, $\|\boldsymbol{x}_a\|_2 = 1$。权重矩阵 $\boldsymbol{W} \triangleq [w_{a,k}]$ 的生成方式如下：首先，对于每个关键词 k，我们均匀随机地选择一个整数 $n_k \in [1, 10]$，然后随机选择一个包含 n_k 个臂 $\mathcal{A}_k$ 的子集，作为关键词 k 的相关臂；其次，对于每个臂 a，如果它与一组包含 n_a 个关键词 $\mathcal{K}_a$ 的集合相关联，我们便赋予相等的权重 $w_{a,k} = \frac{1}{n_a}, \forall k \in \mathcal{K}_a$。遵循 (Zhang 等人，2020)，每个关键词 k 的特征向量通过 $\tilde{\boldsymbol{x}}_k = \sum_{a \in \mathcal{A}} \frac{w_{a,k}}{\sum_{a' \in \mathcal{A}} w_{a',k}} \boldsymbol{x}_a$ 计算得出。臂级奖励和关键词级反馈按照公式 (1) 和公式 (3) 生成。

基线方法。 我们将算法与以下基线方法进行比较：

- LinUCB（Li 等人，2010）：一种最先进的上下文线性赌博机算法，仅根据臂级反馈选择臂，而不使用对话反馈。
- Arm-Con（Christakopoulou、Radlinski 和 Hofmann 2016年）：一种对话式多臂老虎机算法，会围绕臂展开对话，但不考虑关键词，并使用 LinUCB 进行臂选择。



(D) 5HJUHW IRU IL[HG NH\-WHUP VHW $ RQ /

(E) 5HJUHW IRU IL[HG NH\-WHUP VHW $ RQ 0R

(F) 5HJUHW IRU WLPH-YDU\LQJ $ RQ /DV

(G) 5HJUHW IRU WLPH-YDU\LQJ $ RQ 0RYL

图 2：真实世界数据集上的实验结果

- ConUCB（Zhang 等人 2020年）：核心对话式多臂老虎机算法，每当允许进行对话时，都会选择一个关键词，以尽量减少某种估计误差。
- ConLinUCB-UCB：一种在我们提出的 ConLinUCB 框架中采用类似 LinUCB 的方法作为关键词选择策略的算法，即在第 t 轮选择关键词 $k \in \text{argmax}_{k \in \mathcal{K}_t} \tilde{\boldsymbol{x}}_k^\top \boldsymbol{\theta}_t + \alpha_t \|\tilde{\boldsymbol{x}}_k\|_{\boldsymbol{M}_t^{-1}}$。

5.2 评估结果

本节首先展示关键词集合$\mathcal{K}$ 固定时的结果。在这种情况下，我们评估所有算法的遗憾 $R(T)$，并研究对话频率函数 $b(t)$ 以及每轮 t可用臂的数量 $|\mathcal{A}_t|$ 所产生的影响。当 $\mathcal{K}$ 随时间变化时，ConLinUCB-BS 不适用，因此我们比较其他算法的遗憾。按照（Zhang 等人，2020）的做法，除非另有说明，我们设置 $T = 1,\ 000, b(t) = 5\lfloor \log(t+1) \rfloor$ 和 $|\mathcal{A}_t| = 50$。

累积遗憾 我们将实验运行 10 次，并计算每种算法中所有用户的平均遗憾。我们将 $\pm std$ 作为误差线，其中 std 表示标准差。结果如图1（a）所示。首先，其他所有算法的表现都优于 LinUCB，体现了对话的优势。此外，在我们提出的 ConLinUCB 框架中，即使使用采用简单 LinUCB 类关键词选择策略的 ConLinUCB-UCB，其性能也已经优于 ConUCB（提升 34.91%），表明其能够更高效地整合信息。通过探索性对话，ConLinUCB-BS 和 ConLinUCB-MCR 实现了显著更低的遗憾（相较于 ConUCB 分别提升 37.00% 和 43.10%），表明其具有更高的学习准确性。ConLinUCB-MCR 进一步利用历史信息，自适应地开展探索性对话，因此实现了最低的遗憾。

对话频率函数的影响 $b(t)$ 较大的$b(t)$意味着智能体可以进行更多对话。我们设置 $b(t) = f_q \cdot \lfloor \log t \rfloor$，并改变 f_q 以调整对话频率，即 $f_q \in \{5, 10, 20, 30\}$。结果如图1 (b)所示。随着 $b(t)$增大，我们的算法遗憾更小，体现了对话的作用。在所有情况下，ConLinUCB-BS 和 ConLinUCB-MCR 的遗憾都低于 ConUCB，而 ConLinUCB-MCR 表现最佳。

.....的影响 $|\mathcal{A}_t|$ 我们将 $|\mathcal{A}_t|$ 分别设为 25、50、100、200 和 500。为了清晰展示我们算法的优势，我们评估 LinUCB 与其他算法之间的遗憾差异，即 $R_{\text{LinUCB}}(T) - R(T)$，以此表示与 LinUCB 相比，对话式老虎机算法在准确率上的提升。请注意，$|\mathcal{A}_t|$ 越大，算法就越难识别出最佳臂。图1（c）的结果显示，随着 $|\mathcal{A}_t|$ 增大，ConLinUCB-BS 和 ConLinUCB-MCR 的优势愈发显著。具体而言，当 $|\mathcal{A}_t|$=25时，ConLinUCB-BS 和 ConLinUCB-MCR 相较于 ConUCB 分别提升了 34.99% 和 40.21%；当 $|\mathcal{A}_t|$=500时，ConLinUCB-BS 和 ConLinUCB-MCR 相较于 ConUCB 分别提升了 50.36% 和 53.77%。在实际应用中，臂集合的规模 $|\mathcal{A}_t|$ 通常很大。因此，我们提出的算法有望在实践中显著优于 ConUCB。

时变累积遗憾 $\mathcal{K}$ 本节研究每轮中智能体只能获得部分关键词 $\mathcal{K}_t \subseteq \mathcal{K}$ 的情况 t。如前所述，ConLinUCB-BS 不适用于这种情况。每个时刻可获得的关键词数量 t 设为 $|\mathcal{K}_t| = 300$。在第 t轮，从 $\mathcal{K}$ 中均匀随机选取 300 个关键词，组成 $\mathcal{K}_t$。我们评估了除 ConLinUCB-BS 之外所有算法的遗憾。结果如图1(d)所示。可以看出，ConLinUCB-MCR 优于所有基线方法，较 ConUCB 提升了 43.02%。

算法	总时间	选择臂的总时间	选择关键词的总时间
ConUCB	11,297	5,217	6,080
ConLinUCB-UCB	5,738	3,060	2,678
ConLinUCB-MCR	4,821	3,030	1,791
ConLinUCB-BS	3,127	3,120	6

表 1：算法在 Movielens（$T = 5{,}000$）上的总运行时间（秒）。

6 在真实世界数据集上的实验

本节展示了在两个真实世界数据集 Last.FM 和 MovieLens 上的实验结果。基线方法、臂级奖励和关键词级反馈的生成方式，以及重心张成集的计算方法，均与上一节相同。参照 Zhang 等人（2020）在真实数据上的实验，除非另有说明，我们设定 $T = 5{,}000$,$b(t) = 5\lfloor \log(t+1) \rfloor$ 和 $|\mathcal{A}_t| = 50$。

6.1 实验设置

Last.FM 和 Movielens 数据集（Cantador、Brusilovsky 和 Kuflik，2011） Last.FM 是一个用于音乐艺术家推荐的数据集，包含 1,892 个用户与 17,632 位艺术家之间的 186,479 条交互记录。Movielens 是一个用于电影推荐的数据集，包含 2,113 个用户与 10,197 部电影之间的 47,957 条交互记录。

数据生成 数据按照（Li 等人，2019；Zhang 等人，2020；Wu et al. 2021）的方法生成。我们将每位音乐艺术家和每部电影视为一个臂。对于两个数据集，我们提取由用户分配标签数量最多的 $|\mathcal{A}| = 2{,}000$个臂，以及分配标签数量最多的 $N_u = 500$ 名用户。对于每个臂，我们最多保留20个与最多臂相关的标签，并将其视为该臂关联的关键词。所有与臂相关且保留的关键词构成关键词集合 $\mathcal{K}$。Last.FM的关键词数量为 $|\mathcal{K}| = 2{,}726$，Movielens的关键词数量为5,585。与同一臂相关的所有关键词的权重均设为相等。基于交互记录，用户反馈构造如下：如果用户曾为该物品分配过标签，则反馈为1；否则反馈为0。按照（Li 等人，2019）的方法生成用户和臂的特征向量：我们基于上述用户反馈构造反馈矩阵 $\boldsymbol{F} \in \mathbb{R}^{N_u \times N}$，并使用奇异值分解（SVD）对其进行分解：$\boldsymbol{F} = \boldsymbol{\Theta} \boldsymbol{S} \boldsymbol{X}^\top$，其中 $\boldsymbol{\Theta} = (\boldsymbol{\theta}_u^*)$、 $u \in [N_u]$ 和 $\boldsymbol{X} = (\boldsymbol{x}_a)$， $a \in [N]$。我们在 $\boldsymbol{S}$中选择奇异值最高的 $d = 50$ 个维度。按照（Zhang 等人，2020），使用 $\tilde{\boldsymbol{x}}_k = \sum_{a \in \mathcal{A}} \frac{w_{a,k}}{\sum_{a' \in \mathcal{A}} w_{a',k}} \boldsymbol{x}_a$ 计算关键词的特征向量。随后按照公式（1）和公式（3）生成臂级奖励和关键词级反馈。

6.2 评估结果

本节首先展示在两种情况下的两个数据集上的结果：$\mathcal{K}$ 固定，$\mathcal{K}$ 随时间 t变化。我们还比较了所有算法在 MovieLens 数据集上的运行时间，因为该数据集的关键词数量多于 Last.FM。

累积遗憾 我们在固定的生成的数据集上运行实验 10 次，并计算所有用户在 $T = 5{,}000$ 轮次中的平均遗憾。实验的随机性来自随机选择的 $\mathcal{A}_t$（也包括 $\mathcal{K}_t$ 在时变关键词集合情况下）以及 ConLinUCB-BS 算法中的随机性。我们还将 $\pm std$ 作为误差条。对于时变关键词集合情况，我们设置 $|\mathcal{K}_t| = 1{,}000$，并在第 t 轮从 $\mathcal{K}$ 中随机选择 $|\mathcal{K}_t|$ 个关键词，形成 $\mathcal{K}_t$。固定关键词集合下 Last.FM 和 MovieLens 的结果如图 2(a) 和图 2(b) 所示。在两个数据集上，ConLinUCB-BS 和 ConLinUCB-MCR 的遗憾都远小于 ConUCB（在 Last.FM 上分别提升 13.28% 和 17.12%，在 MovieLens 上分别提升 13.08% 和 16.93%），甚至基于我们的 ConLinUCB 框架的简单 ConLinUCB-UCB 也优于 ConUCB。时变关键词集合下 Last.FM 和 MovieLens 的结果如图 2(c) 和图 2(d) 所示。ConLinUCB-MCR 在两个数据集上的表现都显著优于 ConUCB（在 Last.FM 和 MovieLens 上分别提升 19.66% 和 17.85%）。

运行时间 我们在具有代表性的 MovieLens 数据集上评估所有对话式老虎机算法的运行时间，以比较它们的计算效率。为清晰起见，我们报告臂和关键词选择的总运行时间。我们将 $T = 5{,}000$，结果汇总于表 1。显然，与 ConUCB 相比，我们的算法在关键词选择和臂选择上耗时都少得多。具体而言，ConLinUCB-BS 和 ConLinUCB-MCR 的总运行时间分别比 ConUCB 缩短了 72.32% 和 57.32%。主要原因是，我们的算法只需一步即可估计未知的 用户偏好向量，而 ConUCB 则如前所述，需要分两步完成。对于 ConLinUCB-BS，关键词选择所耗费的时间几乎可以忽略，因为每当允许进行一次对话时，它只需从预先计算的重心张成集中随机选取一个关键词。

7 结论

本文提出 ConLinUCB，这是一个用于对话式老虎机、能够高效整合信息的通用框架。基于该框架，我们提出了 ConLinUCB-BS 和 ConLinUCB-MCR，它们采用探索式关键词选择策略，能够快速引出用户的潜在兴趣。我们证明了所提算法的紧遗憾界。特别是，ConLinUCB-BS 达到的遗憾界为$O(d\sqrt{T}\log\ T)$，远优于经典 ConUCB 的 $O(d\sqrt{T}\log\ T)$。在实证评估中，我们的算法表现远胜经典 ConUCB。未来值得进一步研究结合知识图谱（Zhao et al. 2022）、层次化物品树（Song 等，2022）、相对反馈（Xie et al. 2021）或不同反馈选择策略（Letard 等，2020、2022）的情形，并运用我们的框架和原则来提升现有算法的性能。

8 致谢

通讯作者 Shuai Li 获得中国国家自然科学基金（62006151）和上海市启明星计划的资助。John C.S. Lui 的研究部分获得研究资助局的 GRF 14215722 资助。

参考文献

Abbasi-Yadkori, Y.; P´al, D.; 和 Szepesv´ari, C. 2011年。改进的线性随机老虎机算法。神经信息处理系统进展，24。

Amballa, C.; Gupta, M.K.; 和 Bhat, S.P. 2021年。利用一元多项式特征进行学习的高效探索基计算。载于*AAAI*人工智能会议论文集，第35卷，6636–6643。

Awerbuch, B.; 和 Kleinberg, R. 2008年。在线线性优化与自适应路由。计算机与系统科学期刊，74(1):97–114。

Cantador, I.; Brusilovsky, P.; 和 Kuflik, T. 2011年。第二届信息异质性与推荐系统融合研讨会（HetRec 2011）。载于第*5*届*ACM*推荐系统会议论文集，RecSys 2011。美国纽约州纽约市：ACM。

Christakopoulou, K.; Beutel, A.; Li, R.; Jain, S.; and Chi, E. H. 2018年。Q&R：一种面向交互式推荐的两阶段方法。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，139–148。

Christakopoulou, K.; Radlinski, F.; and Hofmann, K. 2016年。迈向对话式推荐系统。载于第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，815–824。

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011。具有线性收益函数的上下文老虎机。载于第十四届国际人工智能与统计学会议论文集，208–214。JMLR研讨会与会议论文集。

Gao, C.; Lei, W.; He, X.; de Rijke, M.; and Chua, T.-S. 2021年。对话式推荐系统的进展与挑战：一项综述。*AIOpen*，2:100–126。

Ikebe, Y.; Inagaki, T.; and Miyamoto, S. 1987. 单调性定理、Cauchy交错定理和Courant-Fischer定理。美国数学月刊，94(4): 352–354.

Lattimore, T.; and Szepesv´ari, C. 2020年。老虎机算法。剑桥大学出版社。

Letard, A.; Amghar, T.; Camp, O.; and Gutowski, N. 2020年。部分老虎机与半老虎机：充分利用有限的用户反馈。载于*2020*年*IEEE*第*32*届人工智能工具国际会议（*ICTAI*），1073–1078。IEEE。

Letard, A.; Amghar, T.; Camp, O.; and Gutowski, N. 2022年. COM-MABs：从用户反馈到推荐。载于*The InternationalFLAIRS Conference Proceedings*，第 35 卷。

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010年. 一种用于个性化新闻文章推荐的上下文老虎机方法。载于第 *19* 届万维网国际会议论文集，661–670。Li, S.; Chen, W.; Li, S.; and Leung, K.-S. 2019年. 改进的老虎机在线聚类算法。载于第 *28* 届国际人工智能联合会议论文集，IJCAI'19，2923–2929。AAAI出版社。ISBN 9780999241141。Li, S.; Lei, W.; Wu, Q.; He, X.; Jiang, P.; and Chua, T.-S. 2021年. 无缝统一属性与物品：面向冷启动用户的对话式推荐。*ACM* 信息系统汇刊（*TOIS*），39(4)：1–29。

Li, S.; and Zhang, S. 2018年。上下文级联老虎机在线聚类。载于*AAAI*人工智能会议论文集，第32卷。

Song, Y.; Sun, S.; Lian, J.; Huang, H.; Li, Y.; Jin, H.; and Xie, X. 2022年。展示整个世界：迈向交互式个性化推荐的整个物品空间探索。载于第十五届*ACM*国际网络搜索与数据挖掘会议论文集，947–956。

Sun, Y.; and Zhang, Y. 2018年. 对话式推荐系统. In 第*41*届*ACM SIGIR*信息检索研究与发展国际会议, 235–244.

Woodbury, M. A. 1950年. 求逆修正矩阵. 统计研究组。

Wu, J.; Zhao, C.; Yu, T.; Li, J.; and Li, S. 2021年. 对话式老虎机聚类，用于用户偏好学习和偏好引导。In 第*30*届*ACM*信息与知识管理国际会议论文集, 2129–2139.

Wu, Q.; Wang, H.; Gu, Q.; and Wang, H. 2016年. 协作环境中的上下文老虎机。In 第*39*届*ACM SIGIR*信息检索研究与发展国际会议, 529–538.

Xie, Z.; Yu, T.; Zhao, C.; and Li, S. 2021年. 基于比较的对话式推荐系统与相对老虎机反馈。In第*44*届*ACM SIGIR*国际信息检索研究与发展会议, 1400–1409.

Zhang, X.; Xie, H.; Li, H.; and CS Lui, J. 2020年. 对话式上下文老虎机：算法与应用. 载于*2020*年万维网大会论文集，662–672.

Zhang, Y.; Chen,X.; Ai, Q.; Yang, L.; and Croft, W. B. 2018年. 迈向对话式搜索与推荐：系统询问，用户作答. 载于第*27*届 *ACM* 信息与知识管理国际会议论文集，177–186.

Zhao, C.；Yu, T.；Xie, Z.；和 Li, S. 2022年。基于老虎机反馈的知识感知对话式偏好引导。载于*2022*年*ACM Web*会议论文集，483–492。

引理1的证明

证明。根据　根据公式 (5) 中 $\boldsymbol{\theta}_t$ 的闭式解，我们可以计算估计误差如下　s

$$
\begin{aligned}
\boldsymbol{\theta}_t - \boldsymbol{\theta}_* &= \boldsymbol{M}_t^{-1}\boldsymbol{b}_t - \boldsymbol{\theta}_* \\
&= \left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\boldsymbol{x}_{a_\tau}^\top + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top + \beta\boldsymbol{I}\right)^{-1}\left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}r_{a_\tau,\tau} + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{r}_{k,\tau}\right) - \boldsymbol{\theta}_* \\
&= \left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\boldsymbol{x}_{a_\tau}^\top + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top + \beta\boldsymbol{I}\right)^{-1}\left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\left(\boldsymbol{x}_{a_\tau}^\top\boldsymbol{\theta}_* + \epsilon_\tau\right) + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\left(\tilde{\boldsymbol{x}}_k^\top\boldsymbol{\theta}_* + \tilde{\epsilon}_\tau\right)\right) - \boldsymbol{\theta}_* \\
&= \left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\boldsymbol{x}_{a_\tau}^\top + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top + \beta\boldsymbol{I}\right)^{-1}\left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\boldsymbol{x}_{a_\tau}^\top + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top + \beta\boldsymbol{I} - \beta\boldsymbol{I}\right)\boldsymbol{\theta}_* - \boldsymbol{\theta}_* \\
&\quad + \boldsymbol{M}_t^{-1}\left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\epsilon_\tau + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\epsilon}_\tau\right) \\
&= -\beta\boldsymbol{M}_t^{-1}\boldsymbol{\theta}_* + \boldsymbol{M}_t^{-1}\left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\epsilon_\tau + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\epsilon}_\tau\right).
\end{aligned}
$$

We c然后，界定估计误差在动作向量 $\boldsymbol{x}_a$方向上的投影:

$$
\begin{aligned}
\left|\boldsymbol{x}_a^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)\right| &\le \beta\left|\boldsymbol{x}_a^\top\boldsymbol{M}_t^{-1}\boldsymbol{\theta}_*\right| + \left|\boldsymbol{x}_a^\top\boldsymbol{M}_t^{-1}\left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\epsilon_\tau + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\epsilon}_\tau\right)\right| \\
&\le \beta\left\|\boldsymbol{x}_a^\top\boldsymbol{M}_t^{-\frac{1}{2}}\right\|_2\left\|\boldsymbol{M}_t^{-\frac{1}{2}}\boldsymbol{\theta}_*\right\|_2 + \left\|\boldsymbol{x}_a^\top\boldsymbol{M}_t^{-\frac{1}{2}}\right\|_2 \times \left\|\boldsymbol{M}_t^{-\frac{1}{2}}\left(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\epsilon_\tau + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\epsilon}_\tau\right)\right\|_2 \qquad (12) \\
&\le \beta\left\|\boldsymbol{x}_a\right\|_{\boldsymbol{M}_t^{-1}}\left\|\boldsymbol{M}_t^{-\frac{1}{2}}\right\|_2\left\|\boldsymbol{\theta}_*\right\|_2 + \left\|\boldsymbol{x}_a\right\|_{\boldsymbol{M}_t^{-1}}\left\|\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau,\tau}\epsilon_\tau + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\epsilon}_\tau\right\|_{\boldsymbol{M}_t^{-1}} \qquad (13) \\
&\le \left\|\boldsymbol{x}_a\right\|_{\boldsymbol{M}_t^{-1}}\left(\sqrt{\beta}\left\|\boldsymbol{\theta}_*\right\|_2 + \left\|\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\epsilon_\tau + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\epsilon}_\tau\right\|_{\boldsymbol{M}_t^{-1}}\right), \qquad (14)
\end{aligned}
$$

其中，公式（12）由柯西-施瓦茨不等式得出，公式（13）由矩阵算子范数的不等式得出，而公式（14）是因为 $\lambda_{min}(\boldsymbol{M}_t) \ge \beta$，$\left\|\boldsymbol{M}_t^{-\frac{1}{2}}\right\|_2 = \sqrt{\lambda_{max}(\boldsymbol{M}_t^{-1})} = \sqrt{\lambda_{min}(\boldsymbol{M}_t)^{-1}} \le \sqrt{\frac{1}{\beta}}$。

定理 1（Abbasi-Yadkori、P´al 和 Szepesv´ari，2011）表明，以至少 $1-\delta$ 的概率

$$
\left\|\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\epsilon_\tau + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\epsilon}_k\right\|_{\boldsymbol{M}_t^{-1}} \le \sqrt{2\log\left(\frac{det(\boldsymbol{M}_t)^{\frac{1}{2}}det(\beta\boldsymbol{I})^{\frac{1}{2}}}{\delta}\right)}, \qquad (15)
$$

其中，$det(\cdot)$ 表示该参数的行列式。我们有

$$
\begin{aligned}
det(\boldsymbol{M}_t) = \prod_{i=1}^{d}\lambda_i &\le \left(\frac{\sum_{i=1}^{d}\lambda_i}{d}\right)^d \qquad (16) \\
&= \left(\frac{trace(\boldsymbol{M}_t)}{d}\right)^d \qquad (17) \\
&= \left(\frac{trace(\sum_{\tau=1}^{t-1}\boldsymbol{x}_{a_\tau}\boldsymbol{x}_{a_\tau}^\top + \sum_{\tau=1}^{t}\sum_{k\in\mathcal{K}_\tau}\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^\top + \beta\boldsymbol{I})}{d}\right)^d \le \left(\frac{t + b(t) + \beta d}{d}\right)^d,
\end{aligned}
$$

其中，$\lambda_i, i = 1, 2, \ldots, d$ 表示矩阵 $\boldsymbol{M}_t$ 的特征值，$trace(\boldsymbol{M}_t)$ 表示 $\boldsymbol{M}_t$ 的迹。公式（16）由算术平均数与几何平均数的不等式可得；公式（17）则是因为矩阵的迹等于其特征值之和。

将上述不等式和 $det(\beta \boldsymbol{I}) = \beta^d$ 代入公式（15），可得

$$\left\| \sum_{\tau=1}^{t-1} \boldsymbol{x}_{a_\tau} \epsilon_\tau + \sum_{\tau=1}^{t} \sum_{k \in \mathcal{K}_\tau} \tilde{\boldsymbol{x}}_k \tilde{\epsilon}_k \right\|_{\boldsymbol{M}_t^{-1}} \leq \sqrt{2 \log(\frac{1}{\delta}) + d \log(1 + \frac{b(t)+t}{\beta d})} \,. \tag{18}$$

将公式（18）代入公式（14），再结合 $\|\boldsymbol{\theta}^*\|_2 \leq 1$和 □，即可得出该结果。

B 引理 2 的证明

证明。 回顾一下，在 ConLinUCB-BS 中，关键词是从预计算重心张量 $\mathcal{B}$ 中均匀采样的，即$k \sim \text{unif}(\mathcal{B})$。因此我们有

$$\lambda_{\mathcal{B}} \coloneqq \lambda_{\min}(\boldsymbol{E}_{k \sim \text{unif}(\mathcal{B})}[\tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^\top]) > 0 \,. \tag{19}$$

利用 (Li和Zhang（2018）) 中引理7的公式（11），并结合 $b_t = b \cdot t$，则以至少 $1-\delta$ 的概率，对于 $\delta \in (0, \frac{1}{8}]$, ，我们有

$$\lambda_{\min}(\sum_{\tau=1}^{t} \sum_{k \in \mathcal{K}_\tau} \tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^\top) \geq \frac{\lambda_{\mathcal{B}} bt}{2} \,, \tag{20}$$

对于所有 $t \geq t_0 = \frac{256}{b\lambda_{\mathcal{B}}^2} \log(\frac{128d}{\lambda_{\mathcal{B}}^2 \delta})$。

根据 Courant–Fischer定理（Ikebe、Inagaki和Miyamoto（1987）），结合 $\|\boldsymbol{x}_a\|_2 = 1$和公式（20），可得对于任意 $t \geq t_0$，以至少 $1-\delta$ 的概率，对于 $\delta \in (0, \frac{1}{8}]$,

$$\begin{aligned}
\|\boldsymbol{x}_a\|_{\boldsymbol{M}_t^{-1}} &= \sqrt{\boldsymbol{x}_a^\top \boldsymbol{M}_t^{-1} \boldsymbol{x}_a} \\
&\leq \max_{\substack{\boldsymbol{x} \in \mathbb{R}^d \\ \|\boldsymbol{x}\|_2 = 1}} \sqrt{\boldsymbol{x}^\top \boldsymbol{M}_t^{-1} \boldsymbol{x}} \\
&= \sqrt{\lambda_{\max}(\boldsymbol{M}_t^{-1})} \\
&= \sqrt{\lambda_{\max}\left((\sum_{\tau=1}^{t-1} \boldsymbol{x}_{a_\tau} \boldsymbol{x}_{a_\tau}^\top + \sum_{\tau=1}^{t} \sum_{k \in \mathcal{K}_\tau} \tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^\top + \beta \boldsymbol{I})^{-1} \right)} \\
&= \sqrt{\frac{1}{\lambda_{\min}\left(\sum_{\tau=1}^{t-1} \boldsymbol{x}_{a_\tau} \boldsymbol{x}_{a_\tau}^\top + \sum_{\tau=1}^{t} \sum_{k \in \mathcal{K}_\tau} \tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^\top + \beta \boldsymbol{I} \right)}} \\
&\leq \sqrt{\frac{1}{\lambda_{\min}(\sum_{\tau=1}^{t} \sum_{k \in \mathcal{K}_\tau} \tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^\top)}} \\
&\leq \sqrt{\frac{2}{\lambda_{\mathcal{B}} bt}} \,.
\end{aligned}$$

□

C 定理3的证明

证明。 我们将第 t 轮的即时遗憾记为 R_t。根据公式（2）中给出的累积遗憾定义、公式（7）所示的臂选择策略以及引理1，我们可以将每一轮 $t = 1$、2、3、4、..... T的遗憾 R_t 界定如下

$$\begin{aligned}
R_t &= \boldsymbol{x}_{a_t^*}^\top \boldsymbol{\theta}^* - \boldsymbol{x}_{a_t}^\top \boldsymbol{\theta}^* \\
&= \boldsymbol{x}_{a_t^*}^\top (\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) + (\boldsymbol{\theta}_t^\top \boldsymbol{x}_{a_t^*} + C_{a_t^*,t}) - (\boldsymbol{\theta}_t^\top \boldsymbol{x}_{a_t} + C_{a_t,t}) \\
&\quad + \boldsymbol{x}_{a_t}^\top (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) + C_{a_t,t} - C_{a_t^*,t} \\
&\leq 2 C_{a_t,t} \,.
\end{aligned} \tag{21}$$

根据引理 2，并结合假设：对于任意 t， $r_t \leq 1$ 以至少 $1-\delta$ 的概率成立，其中某个 $\delta \in (0, \frac{1}{4}]$, ，我们可以得到

$$
\begin{aligned}
R(T) &= R(\lceil t_0 \rceil) + \sum_{t=\lceil t_0 \rceil + 1}^{T} R_t \\
&\leq t_0 + 1 + 2 \sum_{t=\lceil t_0 \rceil}^{T} C_{a_t, t} \\
&\leq t_0 + 1 + 2\alpha_t \sum_{t=\lceil t_0 \rceil}^{T} \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}} \\
&\leq t_0 + 1 + 2\alpha_T \sum_{t=\lceil t_0 \rceil}^{T} \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}} \qquad (22) \\
&\leq t_0 + 1 + 2\alpha_T \sum_{t=\lceil t_0 \rceil}^{T} \sqrt{\frac{2}{\lambda_{\mathcal{B}} b t}} \\
&\leq t_0 + 1 + 2\alpha_T \sqrt{\frac{2}{\lambda_{\mathcal{B}} b}} \int_{t_0}^{T} \sqrt{\frac{1}{t}} dt \\
&=\leq t_0 + 1 + 4\alpha_T \sqrt{\frac{2}{\lambda_{\mathcal{B}} b}} (\sqrt{T} - \sqrt{t_0}) \\
&\leq t_0 + 1 + 4\alpha_T \sqrt{\frac{2}{\lambda_{\mathcal{B}} b}} \sqrt{T}\,,
\end{aligned}
$$

其中，式（22）成立，因为 α_t 关于 t单调不减。

将 t_0 和 α_T的定义代入即可得出该结果。 □

D 定理 4 的证明

证明。 我们首先证明以下结果：对于任意两个正定矩阵 $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d \times d}$以及任意向量 $\boldsymbol{x} \in \mathbb{R}^d$，有：

$$
\|\boldsymbol{x}\|^2_{(\boldsymbol{A}+\boldsymbol{B})^{-1}} \leq \|\boldsymbol{x}\|^2_{\boldsymbol{A}^{-1}} \,. \qquad (23)
$$

该结果可通过以下论证证明：

$$
\begin{aligned}
\|\boldsymbol{x}\|^2_{(\boldsymbol{A}+\boldsymbol{B})^{-1}} &= \boldsymbol{x}^\top (\boldsymbol{A}+\boldsymbol{B})^{-1} \boldsymbol{x} \\
&= \boldsymbol{x}^\top \big(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1})^{-1} \boldsymbol{A}^{-1}\big) \boldsymbol{x} \qquad (24) \\
&= \boldsymbol{x}^\top \boldsymbol{A}^{-1} \boldsymbol{x} - (\boldsymbol{A}^{-1}\boldsymbol{x})^\top (\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1})^{-1} (\boldsymbol{A}^{-1}\boldsymbol{x}) \\
&\leq \boldsymbol{x}^\top \boldsymbol{A}^{-1} \boldsymbol{x} = \|\boldsymbol{x}\|^2_{\boldsymbol{A}^{-1}} \,, \qquad (25)
\end{aligned}
$$

其中，公式（24）由伍德伯里矩阵恒等式（Woodbury 1950）可得，而公式（25）是因为 $(\boldsymbol{B}^{-1} + \boldsymbol{A}^{-1})^{-1}$ 是一个正定矩阵。

根据上述结果，结合公式（21）和柯西-施瓦茨不等式，我们可以得到

$$
\begin{aligned}
R(T) &\leq 2\sum_{t=1}^{T} C_{a_t,t} \\
&= 2\alpha_t \sum_{t=1}^{T} \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}} \\
&\leq 2\alpha_T \sum_{t=1}^{T} \|\boldsymbol{x}_{a_t}\|_{\boldsymbol{M}_t^{-1}} \\
&\leq 2\alpha_T \sqrt{T\sum_{t=1}^{T} \|\boldsymbol{x}_{a_t}\|^2_{\boldsymbol{M}_t^{-1}}} \\
&\leq 2\alpha_T \sqrt{T\sum_{t=1}^{T} \|\boldsymbol{x}_{a_t}\|^2_{\boldsymbol{V}_t^{-1}}}\,.
\end{aligned}
\tag{26}
$$

使用 Lemm a 11（Abbasi-Yadkori, P´al 和 Szepesv´ari 2011）中，至少以 $1-\delta$ 的概率，我们可以 g et

$$
\sum_{t=1}^{T} \|\boldsymbol{x}_{a,t}\|^2_{\boldsymbol{V}_t^{-1}} \leq 2\log\left(\frac{det(\boldsymbol{V}_T)}{det(\beta\boldsymbol{I})}\right). \tag{27}
$$

按照与式（17）类似的步骤，我们可以得到

$$
\frac{det(\boldsymbol{V}_T)}{det(\beta\boldsymbol{I})} \leq \left(\frac{T+\beta d}{\beta d}\right)^d. \tag{28}
$$

因此我们有

$$
\sum_{t=1}^{T} \|\boldsymbol{x}_{a,t}\|^2_{\boldsymbol{V}_t^{-1}} \leq 2d\log(1+\frac{T+1}{\beta d}). \tag{29}
$$

将 α_T 的定义和式（29）代入式（26）即可得出结果。 □