短视频推荐的两阶段约束演员-评论家算法

Qingpeng Cai 快手科技 北京，中国
caiqingpeng@kuaishou.com

Zhenghai Xue 快手科技 北京，中国
xuezhenghai@kuaishou.com

Chi Zhang 快手科技 北京，中国
zhangchi08@kuaishou.com

Wanqi Xue 快手科技 北京，中国
xuewanqi@kuaishou.com

Shuchang Liu 快手科技 北京，中国
liushuchang@kuaishou.com

Ruohan Zhan 香港科技大学 中国香港 rhzhan@ust.hk

Xueliang Wang 快手科技 北京，中国
wangxueliang03@kuaishou.com

Tianyou Zuo 快手科技 北京，中国
zuotianyou@kuaishou.com

Wentao Xie 快手科技 北京，中国
xiewentao@kuaishou.com

Dong Zheng 快手科技 北京，中国
zhengdong@kuaishou.com

Peng Jiang*
快手科技
中国北京
jiangpeng@kuaishou.com

Kun Gai 无隶属机构 中国北京
gai.kun@qq.com

arXiv:2302.01680v3 [cs.LG] 9 Jan 2024

摘要

短视频在社交媒体上的广泛流行，为优化视频分享平台上的推荐系统带来了新的机遇与挑战。用户依次与系统互动，并提供复杂且多维的响应，包括观看时长以及与多个视频的各种互动。一方面，平台旨在长期优化用户的累计观看时长（主要目标），而强化学习能够有效地实现这一目标。另一方面，平台还需要满足容纳多种用户互动响应的约束（辅助目标），例如点赞、关注、分享等。本文将短视频推荐问题建模为约束马尔可夫决策过程（CMDP）。我们发现，传统的约束强化学习算法在这一场景中效果不佳。我们提出一种新颖的两阶段约束演员-评论家方法：第一阶段，学习各个策略，分别优化每项辅助信号。第二阶段，学习一种策略，以（i）优化主要信号，并（ii）与第一阶段学到的策略保持接近，从而有效保证该主策略在……上的性能辅助目标。通过大量离线评估，我们证明了与其他方法相比，我们的方法在优化主要目标以及平衡其他目标方面均有效。我们进一步展示了该方法在短视频推荐在线实验中的优势：在观看时长和互动方面均显著优于其他基线方法。我们的方法已全面部署于生产系统，以优化平台上的用户体验。

CCS 概念

• **信息系统** → **推荐系统**； • **计算方法** → **强化学习**。

关键词

约束强化学习、推荐系统、短视频推荐

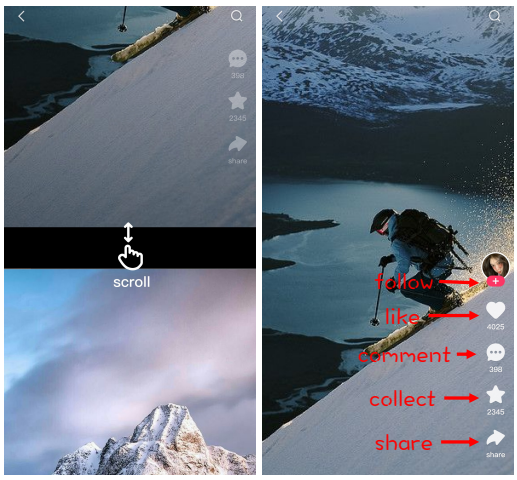
ACM 参考格式：
Qingpeng Cai、Zhenghai Xue、Chi Zhang、Wanqi Xue、ShuchangLiu、Ruohan Zhan、Xueliang Wang、Tianyou Zuo、Wentao Xie、Dong Zheng、Peng Jiang 和 Kun Gai。2023年。短视频推荐的两阶段约束演员-评论家算法。载于*2023*年*ACM*万维网大会论文集（*WWW '23*），*2023*年*5*月*1*日至*5*日，美国得克萨斯州奥斯汀。ACM，纽约州纽约，美国，11页。
https://doi.org/10.1145/3543507.3583259

*通讯作者

允许免费为个人或课堂使用而制作本作品全部或部分内容的数字或纸质副本，前提是这些副本不用于或分发以谋取利润或商业利益，且副本首页保留本声明和完整引文。对于本作品中由作者以外的其他人拥有版权的部分，必须尊重其版权。注明出处后，允许摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，则须事先获得具体许可和/或支付费用。请联系 permissions@acm.org 申请许可。
WWW '23，*2023*年*5*月*1*日至*5*日，美国得克萨斯州奥斯汀© 2023年版权归所有者/作者所有。出版权授权给 ACM。ACM ISBN 978-1-4503-9416-1/23/04...$15.00
https://doi.org/10.1145/3543507.3583259

1 引言

短视频的迅猛普及正在改变社交媒体的现状。短视频消费为组织带来了巨大的商业机会。因此，优化短视频平台的推荐策略日益受到关注 [Gong 等，2022；Lin 等，2022；Wang 等，2022a；Zhan 等。



(a) (b)

图 1：热门短视频平台（如 TikTok、快手等）的示例。

2022]用户通过上下滑动并观看多个视频与平台互动，如图 1(a) 所示。用户会针对每个视频提供多维度响应。如图 1(b) 左侧所示，用户观看视频后的潜在响应包括 WatchTime（观看视频所花费的时间）和多种互动：Follow（关注视频作者）、Like（点赞该视频）、Comment（评论该视频）、Collect（收藏该视频）、Share（与朋友分享该视频）等。

一方面，平台的主要目标是优化多个视频的累积观看时长，因为观看时长反映了用户的注意力，并且与日活跃用户数（DAU）密切相关。近年来，越来越多的研究着眼于将强化学习（RL）应用于推荐系统，因为它能够提升累积奖励 [Afsar 等，2021年；Chen 等，2019b、2018；Gao 等，2022a；Ge 等，2021年；Liu 和 Yang，2019年；Ma 等，2020年；Nemati 等，2016年；Wang 等，2022b；Xian 等，2019年；Xin 等，2022年；Zhao 等，2018年、2017年；Zou 等， 2019]。具体而言，采用强化学习方法可以有效地累积最大化观看时长，从而增加用户观看多个视频所花费的时间。另一方面，点赞／关注／分享等其他响应也反映了用户满意度。因此，平台需要满足用户互动方面的约束。因此，只优化单一目标（例如电子商务平台的商品交易总额 [Pi 等， 2020]）的现有推荐系统已不足以满足需求——所采用的系统应将响应的各个方面都纳入考量，以优化用户体验。

本文将短视频推荐问题建模为约束马尔可夫决策过程：用户充当环境，推荐算法充当智能体；在每个时间步，智能体执行一个动作（向用户推荐一条视频），环境向智能体发送多种奖励（响应）。智能体的目标是在满足其他交互响应（辅助目标）约束的前提下，最大化累积观看时长（主要目标）。我们的目标不同于帕累托最优性，其旨在寻找一个帕累托最优解 [Chen 等人，2021；Lin 等人，2019；Sener 和 Koltun 2018]，，但这可能无法优先实现系统的主要目标。

与无约束情形相比，这种约束策略优化问题要困难得多。一个自然的想法是应用标准的约束强化学习算法，通过预先设定的乘子 [最大化拉格朗日函数 2018]Tessler等人。然而，由于以下两个原因，这种方法并不适用于我们的场景：

首先，由于用户会产生不同类型的响应，仅使用单个策略评估模型来估计拉格朗日对偶目标是不够的。将这些响应组合起来并不合适，尤其是当各类响应具有各自的折扣因子时——基于价值的模型中的时序差分误差公式只允许使用单一折扣值。即使只需使用一个折扣因子，单个价值模型也可能难以准确评估策略，尤其是在不同响应的观测频率各异时；这在短视频推荐中十分常见。观看时长响应是密集的，每次视频观看都会产生观测；而点赞／关注／分享等交互信号则稀疏得多，可能连续观看数十个视频都不会出现。若直接将这些响应相加，密集响应会削弱稀疏响应所提供的信号。为解决这一多响应评估难题，我们通过各自的价值模型分别评估每种响应，这既支持针对不同响应设置专属折扣因子，也能减轻不同响应之间对评估造成的干扰。第 4.1 节中的实验验证了该方法的有效性。

其次，与 [Tessler等人 2018], 仅考虑一个约束不同，推荐系统中存在多个约束，尤其是在短视频系统中。我们发现，由于多维拉格朗日乘子的搜索空间更大，最大化拉格朗日函数的算法更难进行优化。强化学习算法的训练耗时较长，因此对拉格朗日乘子进行网格搜索会耗费大量时间。鉴于此，我们提出先学习策略以优化各项辅助响应，再对主要响应的策略进行“软”正则化，使其接近其他策略，而不是搜索拉格朗日乘子的最优值。我们从理论上证明了最优解的闭式形式。离线与线上实验结果表明，我们的方法能够更好地最大化主要响应，同时平衡其他响应。

综上，我们将贡献总结如下：

- **短视频推荐中的约束优化**：我们对短视频推荐中的约束策略学习问题进行了形式化，其中不同响应的观测频率可能各不相同，智能体在平衡其他响应的约束下最大化其中一项响应。
- **两阶段约束演员-评论家算法** 我们提出了一种新颖的两阶段约束演员-评论家算法，可有效应对这一挑战：(1) 多评论家策略估计：为了更好地评估可能具有不同折扣因子和观测频率的多重响应的策略，我们提出分别学习一个价值模型来评估每个响应。(2) 两阶段演员学习：我们提出一种

两阶段演员学习方法，首先学习一种策略以优化每个辅助响应，然后对主要响应的策略进行软正则化，使其与其他策略相差不远。我们证明，与其他替代方案相比，这种方法在具有多个约束的约束优化中更为有效。

- **离线与线上实验取得显著成效**：我们通过离线与线上实验展示了该方法的有效性。
- **在真实世界的短视频应用中部署**：我们已在一个热门短视频平台全面上线该方法。

2 相关工作

推荐中的强化学习。 由于强化学习能够优化用户的长期满意度，将其应用于推荐系统的研究日益增多 [Afsar 等 2021]。基于价值的方法会估计向用户推荐候选集中的某个条目所带来的满意度，然后选择预测满意度最高的条目 [Chen 等，2018年；Liu 和 Yang，2019年；Nemati 等，2016年；Zhao 等 2018]。基于策略的方法则直接学习策略（即推荐哪个条目），并通过优化策略来提升用户满意度 [Chen 等，2019b、2019a；Ma 等，2020年；Xian 等 2019]。近年来，研究者日益关注如何将强化学习应用于更复杂的推荐场景，而非仅优化单一目标，例如促进内容项目获得平等曝光机会 [Ge 等 2021], 、提高推荐的多样性和新颖性 [Stamenkovic 等 2021], ，以及通过表征性奖励塑形刻画更全面的用户动态 [Chen 等 2021]。我们的工作与第三类研究互为补充。面对多方面的用户响应，实际应用中的系统通常会对不同类型的用户响应有所偏好。为此，我们提出约束优化问题，而非追求 [Chen 等 2021] 和 [Ge 等 2022]提出的帕累托最优性。

约束强化学习。 我们的工作也与约束强化学习领域的相关研究密切相关。在这类研究中，序贯决策问题被建模为约束马尔可夫决策过程 [Sutton 和 Barto 2018]，，策略学习过程则需要满足约束[Chow 等，2017年；Dalal 等，2018年；Garcıa 和 Fernández，2015；Liu 等，2021年；Tessler等人 2018]。例如， [Tessler等人 2018] 提出交替更新策略和拉格朗日乘子，并证明其算法会收敛到一个不动点。然而，这种方法只对一个约束进行建模，难以扩展到具有多个约束的问题。相比之下，我们针对每种辅助响应分别学习一个策略，以最大化该响应，然后对主策略进行“软”正则化，使其接近其他策略。我们的实证结果表明，在推荐系统中处理多种响应时，这种方法对于约束策略学习更有效。与研究离线强化学习、并将学习到的策略正则化使其接近某一行为策略的 [Nair 等人 2020]不同，我们将策略软性地限制在能够最大化其他辅助响应的策略范围内。

多目标优化。 我们还讨论了多目标优化领域的一类相关研究。为了权衡不同目标，该领域的方法大致可分为两类：帕累托优化，以及通过预先设定权重进行联合优化。帕累托优化旨在寻找这样一种解：不存在其他解能够同时改善所有目标，这称为帕累托最优性 [Chen 等人，2021；Ge 等，2022；Nguyen 等，2020；Sener 和 Koltun 2018]。然而，帕累托最优解未必会优先考虑应用中最受重视的目标。另一种方法是通过预先设定权重，将不同目标合并为单一目标 [Mossalam 等，2016；White 等 1980]。然而，要确定能够准确反映实际应用偏好的权重并不容易 [Tessler等人 2018]。

3 短视频推荐的约束马尔可夫决策过程

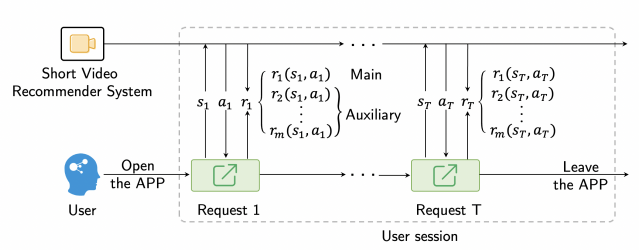


图 2：短视频推荐的马尔可夫决策过程。

我们首先对短视频推荐问题进行建模，如图 2 所示。当用户 u 打开应用时，会开启一个新的会话。一个会话由多个请求组成。每次请求时 t，推荐系统（智能体）会根据用户当前的状态采取一个动作 a_t，向用户推荐一个视频。然后，用户会对展示的视频做出多方面的响应（例如观看时长、点赞、分享和关注），智能体会将这些响应作为向量值奖励信号接收。用户离开应用后，会话结束。推荐系统的目标是在不大幅牺牲其他响应的前提下，优化主要响应（例如观看时长）的累积奖励。

我们将上述过程建模为一个约束马尔可夫决策过程（CMDP） [Sutton and Barto 2018] $(S, A, P, R, C, \rho_0, \Gamma)$，其中 S 是用户当前表示的状态空间 s_t， A 是动作空间（每个动作 a_t 对应一次请求所推荐的视频）， P：$S \times A \to \Delta(S)$描述状态转移， R：$S \times A \to \mathbb{R}^m$ 定义向量值奖励函数，可产生 m 种不同的奖励 $r(s_t, a_t) = (r_1(s_t, a_t) \ldots, r_m(s_t, a_t))$， ρ_0 是初始状态分布，$\Gamma = (\gamma_1, \ldots, \gamma_m) \in (0, 1)^m$ 表示各类响应奖励的折扣因子向量。C 规定了对辅助响应的约束，即其他目标的信号总数下限。

定义向量值折扣累积奖励 R_t 为 $R_t = \sum_{t'=t}^{T} \Gamma^{t'-t} \cdot r(s_{t'}, a_{t'})$，其中 T 表示会话长度（即请求数量），$\Gamma^b = (\gamma_1^b, \ldots, \gamma_m^b)$，且 $\mathbf{x} \cdot \mathbf{y}$ 表示逐元素乘积。令 $V^\pi(s) = (V_1^\pi(s) \ldots, V_m^\pi(s))$ 为状态价值

$E_\pi[R_t|s_t = s]$在遵循策略 π采样的动作下，以及 $Q(s,a) = (Q_1^\pi(s,a)\ldots, Q_m^\pi(s,a))$为其状态-动作价值 $E_\pi[R_t|s_t = s, a_t = a]$。记 ρ_π为策略 π诱导的状态分布。不失一般性，我们将第一个响应设为主要响应。目标是学习推荐策略 $\pi(\cdot|s)$，以求解以下优化问题：

$$\begin{aligned} \max_{\pi} \quad & E_{\rho_\pi}\left[V_1^\pi(s)\right] \\ \text{s.t.} \quad & E_{\rho_\pi}\left[V_i^\pi(s)\right] \geq C_i, \quad i = 2, \ldots, m \end{aligned} \tag{1}$$

其中， C_i是对辅助响应 i的约束。

4 两阶段约束演员-评论家

本节提出一种新颖的两阶段约束演员-评论家方法，用于应对短视频推荐场景中的学习挑战：

多评论家策略估计我们提出分别估计各项响应，以便更准确地估计密集和稀疏信号。

第一阶段 针对每种辅助响应，我们学习一个策略来优化其累积奖励。

第二阶段 针对主要响应，我们学习一个策略来优化其累积奖励，同时通过软性约束使其接近其他为优化辅助响应而学习的策略。

我们首先讨论分别评估不同策略相较于联合估计的优势。其次，我们将在第 4.2 节和第 4.3 节中详细阐述如何在采用随机策略的在线学习设置下应用我们的方法。随后，我们将讨论该方法在离线设置和确定性策略下的扩展。

4.1 多评论家策略估计

我们展示了针对每种响应分别评估，相较于对响应总和进行联合评估的优势。具体而言，对于每次视频观看，我们考虑两种响应：观看时长和互动（即一个指示函数，用于判断观看期间是否发生互动）。

- 对于联合评估，我们学习一个价值模型 V_{joint}，其奖励为观看时长与互动之和。
- 对于单独评估，我们训练两个价值模型 V_w和 V_i，分别以观看时长和互动作为奖励。将单独评估的价值定义为 $V_{separate} = V_w + V_i$

为进行公平比较，我们为所有价值模型使用相同的折扣因子 0.95，并使用从某个热门短视频平台收集的一天数据对它们进行训练。为评估价值模型在观看时长和互动方面的准确性，我们计算模型值 V_{joint}和 $V_{separate}$与每个会话中相应响应总和的蒙特卡罗值之间的相关性。与 V_{joint}相比， $V_{separate}$与观看时长和互动的相关性分别高出 0.19% 和 0.14%（观看时长和互动提高 0.1% 即具有显著性），这表明单独评估比联合学习更能学习不同的奖励响应。

4.2 第一阶段：辅助响应的策略学习

在此阶段，我们分别学习策略，以优化每种辅助响应的累积奖励。为完整起见，我们写出随机策略 [Williams 1992]的过程。对于响应 i，设学习得到的演员和评论家分别由 π_{θ_i}和 V_{ϕ_i}参数化。在迭代 k时，我们观察到由 $\pi_{\theta^{(k)}i}$收集的样本 $(s, a, s\prime)$，即，$s \sim \rho_{\pi_{\theta_i^{(k)}}}, a \sim \pi_{\theta^{(k)}i}(\cdot|s)$ 和 $s\prime \sim P(\cdot|s,a)$。我们通过最小化贝尔曼方程来更新评论家：

$$\phi_i^{(k+1)} \leftarrow \arg\min_{\phi} E_{\pi_{\theta_i^{(k)}}}\left[\left(r_i(s,a) + \gamma_i V_{\phi_i^{(k)}}(s') - V_\phi(s)\right)^2\right]. \tag{2}$$

我们更新演员以最大化优势：

$$\begin{aligned} \theta_i^{(k+1)} &\leftarrow \arg\max_{\theta} E_{\pi_{\theta_i^{(k)}}}\left[A_i^{(k)} \log\left(\pi_\theta(a|s)\right)\right] \\ \text{where} \quad A_i^{(k)} &= r_i(s,a) + \gamma_i V_{\phi_i^{(k)}}(s') - V_{\phi_i^{(k)}}(s). \end{aligned} \tag{3}$$

4.3 第二阶段：主要响应的软约束优化

在预训练策略π_{θ_2}、……、 π_{θ_m}以优化辅助响应后，我们现在进入学习策略的第二阶段，以优化主要响应。我们提出了一种带有多个约束的新型约束策略优化方法。

分别令演员和评论家为π_{θ_1}和 V_{ϕ_1}。在第 k次迭代中，我们同样更新评论家，以最小化贝尔曼方程：

$$\phi_1^{(k+1)} \leftarrow \arg\min_{\phi} E_{\pi_{\theta_1^{(k)}}}\left[\left(r_1(s,a) + \gamma_1 V_{\phi_1^{(k)}}(s') - V_\phi(s)\right)^2\right]. \tag{4}$$

更新演员的原则包含两个方面：(i) 最大化优势；(ii) 将策略限制在与其他策略相差不大的范围内。优化形式化如下：

$$\begin{aligned} \max_{\pi} \quad & E_\pi[A_1^{(k)}] \\ \text{s.t.} \quad & D_{KL}(\pi||\pi_{\theta_i}) \leq \epsilon_i, \quad i = 2, \ldots, m, \\ \text{where} \quad & A_1^{(k)} = r_1(s,a) + \gamma_1 V_{\phi_1^{(k)}}(s') - V_{\phi_1^{(k)}}(s). \end{aligned} \tag{5}$$

我们在以下定理中给出了式 (5) 的拉格朗日函数的闭式解。由于篇幅所限，我们省略证明，请参阅附录A。

定理1。 式 *(5)* 的拉格朗日函数的闭式解为

$$\pi^*(a|s) \propto \prod_{i=2}^{m}\left(\pi_{\theta_i}(a|s)\right)^{\frac{\lambda_i}{\sum_{j=2}^m \lambda_j}} \exp\left(\frac{A_1^{(k)}}{\sum_{j=2}^m \lambda_j}\right), \tag{6}$$

其中 λ_i、 $i = 2$、……、 m 为拉格朗日乘子。

根据由 $\pi_{\theta_1^{(k)}}$ 收集的数据，我们通过最小化其与最优策略 π^*之间的 KL散度来学习策略 π_{θ_1}：

$$\begin{aligned} \theta_1^{(k+1)} &\leftarrow \arg\min_{\theta} E_{\pi_{\theta_1^{(k)}}}\left[D_{KL}(\pi^*(a|s)||\pi_\theta(a|s))\right] \\ &= \arg\max_{\theta} E_{\pi_{\theta_1^{(k)}}}\left[\frac{\prod_{i=2}^m \left(\pi_{\theta_i}(a|s)\right)^{\frac{\lambda_i}{\sum_{j=2}^m \lambda_j}}}{\pi_{\theta_1^{(k)}}(a|s)} \exp\left(\frac{A_1^{(k)}}{\sum_{j=2}^m \lambda_j}\right) \log \pi_\theta(a|s)\right]. \end{aligned} \tag{7}$$

两阶段约束演员-评论家算法的流程见附录 B，我们简称其为 TSCAC。下面我们对式（7）中的演员更新作一些直观解释。项 $\pi_{\theta_i}(a|s)$ 表示策略i 选择该动作的概率，起到重要性权重的作用，可对学习到的策略 π_{θ_1} 进行软约束，使其接近其他策略 π_{θ_i}。较小的拉格朗日乘子 λ_i 意味着约束较弱；当 $\lambda_i = 0$时，我们允许学习到的策略 π_{θ_1} 不受约束策略 π_{θ_i}的影响。请注意，我们将 λ 设为相同的值，这对生产系统来说更为实际。如果采用不同的拉格朗日乘子值进行微调，TSCAC 的性能会更好。不过，正如我们将在后续章节中看到的那样，离线实验和在线实验均验证了 TSCAC 在λ 取相同值时的有效性。

离线学习 现在，我们将讨论如何将约束演员-评论家方法应用于离线设置，即使用固定数据集。从在线学习转向离线学习时，主要变化在于对策略梯度进行偏差校正。演员不再基于当前策略收集的数据进行更新，而是基于另一种行为策略 π_β 收集的数据进行更新，这可能导致被更新策略所诱导的数据分布不同。为解决估计策略梯度时的分布不匹配问题，一种常见策略是通过重要性采样 [Precup 2000；Precup 等人 2001] 使用偏差校正比率。给定一条轨迹 $\tau = (s_1, a_1, s_2, a_2, \ldots\ldots)$，策略$\pi_{\theta_i}$ 的策略梯度偏差校正比率为$w(s_t, a_t) = \prod_{t'=1}^{t} \frac{\pi_{\theta_i}(s_{t'}|a_{t'})}{\pi_\beta(s_{t'}|a_{t'})}$，这能给出无偏估计，但方差可能很大。因此，我们建议使用一阶近似，并在优化辅助响应的演员时使用当前的动作选择比率，

$$\theta_i^{(k+1)} \leftarrow \arg\max_\theta E_{\pi_\beta}\left[\frac{\pi_{\theta_i^{(k)}}(a|s)}{\pi_\beta(a|s)} A_i^{(k)} \log(\pi_\theta(a|s))\right]. \quad (8)$$

更新主要响应的演员时，我们有

$$\theta_1^{(k+1)} \leftarrow \arg\max_\theta E_{\pi_\beta}\left[\frac{\prod_{i=2}^{m}\left(\pi_{\theta_i}(a|s)\right)^{\frac{\lambda_i}{\sum_{j=2}^{m}\lambda_j}}}{\pi_\beta(a|s)} \times \exp\left(\frac{A_1^{(k)}}{\sum_{j=2}^{m}\lambda_j}\right)\log(\pi_\theta(a|s))\right]. \quad (9)$$

确定性策略现在，我们讨论将 TSCAC 扩展到确定性策略[，其灵感来自 Lillicrap 等人 2015]，针对（7）中讨论的受约束策略演员所提出的更新规则。类似地，在第一阶段，对于每个辅助响应 i，我们学习独立的评论家模型 $Q_{\phi_i}(s, a)$和演员模型$\pi_{\theta_i}(s)$。在第二阶段，对于主要响应，我们通过时序学习来学习评论家 $Q_{\phi_1}(s, a)$；对于演员$\pi_{\theta_1}(s)$，其更新规则形式如下：

$$\max_\theta \quad \prod_{i=2}^{m}\left(h(\pi_{\theta_i}(s), \pi_{\theta_1}(s))\right)^{\lambda_i} Q_{\phi_1}(s, \pi_\theta(s)), \quad (10)$$

其中 $h(a_1, a_2)$在两个动作 a_1，a_2 彼此接近时得分较高，反之得分较低；而 $h(\pi_{\theta_i}(s)\ \pi_{\theta_1}(s))$在策略$\pi_{\theta_1}$ 和π_{θ_i}所选动作彼此接近时得分较高。 $\lambda_i \geq 0$的作用与约束拉格朗日乘子类似——λ_i 越大，表示约束越强。例如，给定 n 维动作空间，可以选择 $h(a_1, a_2) = \sum_{d=1}^{n} \exp\left(-\frac{(a_{1d}-a_{2d})_2}{}^2\right)$。TSCAC 的确定性版本可应用于连续动作场景，例如用户偏好的嵌入。

表1：KuaiRand 的统计数据。

维度	数量	稀疏率
用户	26858	-
条目	10,221,515	-
样本	68,148,288	-
点击	25,693,008	37.70%
like	1094434	1.61%
评论	163977	0.24%
hate	32449	0.048%

5 离线实验

本节通过大量离线学习模拟，在一个关于短视频推荐的公开数据集上评估我们的方法。我们展示了与现有基线方法相比，我们的方法在实现主要目标和权衡辅助目标方面的有效性。我们还在另一个公开推荐数据集上测试了该方法的通用性；由于篇幅有限，请参阅附录 C。

5.1 实验设置

我们使用一个名为 *KuaiRand* 的短视频推荐公开数据集（https://kuairand.com/） [Gao 等，2022b]。该数据集采集自一款知名的视频共享移动应用，且由于不存在偏差，适合用于强化学习方法的离线评估。该数据集不仅记录视频的总观看时长，还记录用户的互动行为，包括点击、点赞、评论和仇恨。数据集的统计数据如表1所示。可以看出，点赞、评论和仇恨都是稀疏信号。需要注意的是，仇恨信号极其稀疏。我们将同一用户提供的日志串接起来形成一条轨迹，并选取观看频率最高的前 150 个视频。

MDP.

- 状态 s_t：一个 1044 维向量，由用户特征（用户属性）、用户最近观看的 20 个视频特征（用户历史）以及全部 150 个候选视频特征（上下文）拼接而成。
- 动作 a_t：当前要推荐的视频 ID。
- 奖励 r_t：一个包含五项评分的向量，表示用户针对已观看视频在点击、点赞、评论、仇恨和观看时长方面给出的评分。
- 回合：用户视频观看历史的一个序列。
- 折扣因子 γ：0.99
- 目标：我们将主要目标设为最大化视频观看时长，并将其他目标视为辅助目标。

评估。 我们采用归一化截断重要性采样（NCIS）方法评估不同的策略。该方法是推荐系统中强化学习方法常用的标准离线评估方法 [Zou 等人 2019]。我们还从以下方面评估我们的方法：

其他指标，详见附录D。NCIS 分数定义如下：

$$N(\pi) = \frac{\sum_{s,a\in D} w(s,a)r(s,a)}{\sum_{s,a\in D} w(s,a)}, w(s,a) = \min\{c, \frac{\pi(a|s)}{\pi_\beta(a|s)}\}, \quad (11)$$

其中，D表示数据集，$w(s, a)$表示截断重要性采样比率，π_β表示行为策略，c是一个正常数。

基线方法。我们将 TSCAC 与以下基线方法进行比较。

- **行为克隆**：一种监督式行为克隆策略π_β，用于模仿数据集中的推荐策略，其输入为用户状态，输出为视频 ID。
- **Wide&Deep** [Cheng 等人 2016]：一种利用宽层与深层来平衡记忆能力和泛化能力的监督模型。该模型输入用户状态，输出物品ID，并将每个样本的权重设为该物品所有响应的加权和。
- **DeepFM** [Guo 等人 2017]：一种结合深度神经网络和因子分解机的监督式推荐模型。该模型输入用户状态，输出物品ID，并将每个样本的权重设为该物品所有响应的加权和。
- **RCPO** [Tessler 等人 2018]：一种称为奖励约束策略优化的约束演员-评论家方法，通过优化策略来最大化约束程序的拉格朗日对偶函数。具体而言，奖励函数定义为 $r = r_0 + \sum_{i=1}^{n} \lambda_i * r_i$，其中 r_0 是主要目标，观看时长和 r_i 表示其他反馈，λ_i 是拉格朗日乘子。
- **RCPO-Multi-Critic**：我们测试了 RCPO 的一种改进版本，该版本采用多个评论家。我们分别学习多个评论家模型，以评估每种反馈的累积奖励。随后，在优化演员时，我们最大化由拉格朗日乘子加权的多个评论家的线性组合。
- **帕累托** [Chen 等人 2021]：一种多目标强化学习算法，可为推荐系统找到帕累托最优解。
- **TSCAC**：我们的两阶段约束演员-评论家算法。

5.2 总体性能

表2展示了不同算法在五项得分上的性能。可以看出，我们的 TSCAC 算法显著优于其他算法，包括受约束的强化学习算法和监督学习方法：对于主要目标（观看时长），TSCAC 的性能最高，为13.14(2.23%)；对于辅助目标，TSCAC 在4项得分中的3项（点击、点赞、评论）也排名最高。请注意，TSCAC 在每个维度上的表现均优于行为克隆和 RCPO。帕累托算法确实学到了一个在仇恨指标上性能最佳的帕累托最优解，但其性能得分最低，为11.90(−7.4%)，也就是说，它不满足以优化观看时长为主要目标的设定。RCPO 算法在观看时长上的性能排名第二，为13.07(1.70%)，但其仇恨得分最差，因为在单一评估模型中，稀疏信号被密集信号主导。与 RCPO 相比，RCPO-Multi-Critic 在仇恨指标上的得分高出许多，这证明了多评论家策略估计方法的有效性。TSCAC 在每个维度上的表现也优于 RCPO-Multi-Critic，这表明我们的两阶段演员学习方法具备处理多重响应的能力。

5.3 消融研究

我们研究拉格朗日乘子的取值如何影响性能。由于在第二阶段，我们将所有约束的 λ 取值设为相同，因此我们改变 λ 的取值，分别设为 $[1e-1, 1e-2, 1e-3, 1e-4, 1e-5]$，并展示 TSCAC 在所有响应方面的性能。请注意，较大的 λ 表示对辅助响应的约束更强。图3显示，随着 λ 增大，主要目标观看时长会下降，因为对辅助响应的约束变得更强。如图3所示，当 λ $1e-5$ 较小时，由于约束较弱，互动方面的性能会下降。有趣的是，λ较大时，互动方面的性能也会下降，这表明过强的约束会影响策略的学习。$1e-4$ 取该值时，互动方面的性能最佳，且与其他基线方法相比，观看时长显著提高。

6 在线实验

为了证明我们算法的有效性，我们在一个热门短视频平台上通过在线实验测试其性能以及其他替代方案。算法被应用于该热门短视频平台生产环境中的候选视频排序系统：当用户进入平台时，这些算法需要对候选视频进行排序，系统随后会向用户推荐排名第一的视频。我们展示了所提出的 TSCAC 算法能够学习一种策略，在最大化主要目标的同时有效平衡辅助目标；具体而言，我们将最大化观看时长设为主要目标，将提升用户与视频之间的互动设为辅助目标。

6.1 实验设置

评估指标。我们使用在线指标评估策略的性能。对于主要目标，我们考察用户观看视频所花费的总时长，称为观看时长。对于辅助目标，用户可以通过多种方式与视频互动，例如将视频分享给朋友、下载视频或发表评论。在这里，我们关注与用户和视频互动相关的三项在线指标——分享、下载和评论互动的总数量。

马尔可夫决策过程。按照第 3 节中的表述，我们将介绍用于短视频推荐的约束马尔可夫决策过程的详细内容。

- 状态 s_t：用户历史交互（之前各轮向用户推荐的条目列表及相应的用户反馈）、用户属性（例如设备和位置），以及时间 t 候选视频的特征（嵌入和统计数据）。
- 动作 a_t：算法预测的用户在不同视频主题上的偏好向量嵌入，它通过下文所述的排序函数决定实际的推荐动作（即要推荐的视频）：**排序函数**：对于每个候选视频，此函数计算预测的用户

表2：KuaiRand上不同算法的性能。

算法	点击↑	点赞↑(e-2)	评论↑(e-3)	仇恨↓(e-4)	观看时长↑
BC	0.5338	1.231	3.225	2.304	12.85
Wide&Deep	0.5544 3。86%	1.244 1.07%	3.344 3.69%	2.011 −12.7%	12.84 −0.08%
DeepFM	0.5549* 3.95%*	1.388* 12.76%*	3.310 2.64%	2.112 −8.31%	12.92 0.53%
RCPO	0.5510 3.23%	1.386 12。57%	3.628* 12.5%*	2.951 28。1%	13.07* 1.70%*
RCPO-Multi-Critic	0.5519 3.41%	1.367 11。04%	3.413 5.83%	2.108 −8.49%	13.00 1.14%
帕累托	0.5438 1。87%	1.171 −4.85%	3.393 5.22%	0.9915* −56.96%*	11.90 −7.4%
TSCAC	**0.5570** **4.35%**	**1.462** **18.80%**	**3.728** **15。6%**	1.870 −18.83%	**13.14** **2。23%**

括号中的数字表示该列的单位；每种算法第一行的数字是 NCIS 分数。第二行的百分比表示该算法与 BC 算法之间的性能差距。带有 ∗ 的数字表示各响应维度上所有基线方法中的最佳性能。当 TSCAC 在某个响应维度上取得最佳性能时，最后一行会以粗体标示。

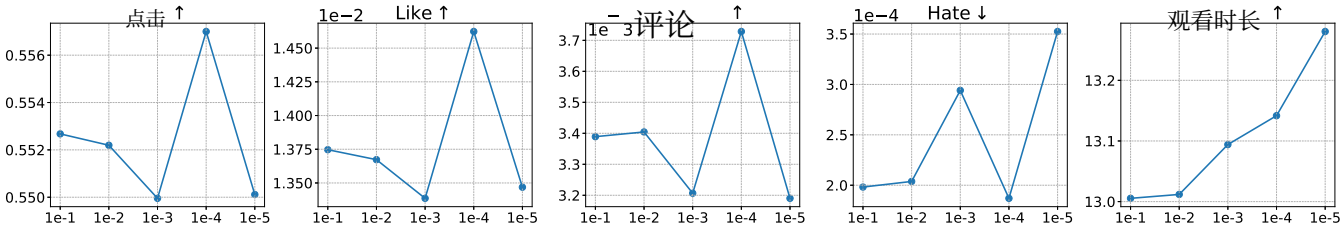


图3：拉格朗日乘子的值对性能的影响。

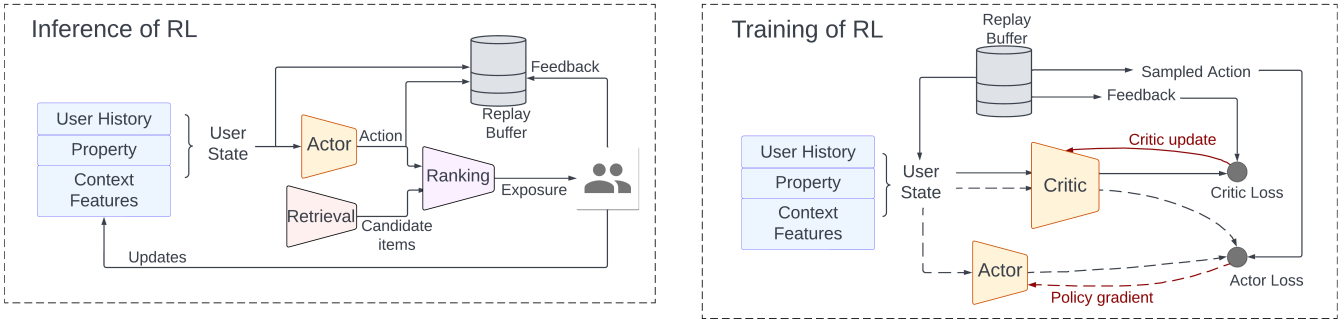


图 4：强化学习在生产系统中的工作流程。

偏好向量（a_t）和视频嵌入（表示其主题和质量），如 [Dulac-Arnold 等人 2015] 所述。然后推荐得分最高的视频。

- 奖励 $r_t = (l_t, i_t)$：每次推荐后，系统都会观察用户在视频上花费了多长时间，即观看时长，记作 l_t，以及用户是否与视频进行了互动（分享/下载/评论），记作 i_t。
- 回合：一条轨迹从用户打开应用时开始，到用户离开时结束。
- 策略：我们选择在在线实验中学习高斯策略。具体而言，动作 a_t 从多元高斯分布中采样，其均值和方差由演员模型输出。

工作流程. 如图 4 所示，强化学习的运行方式如下s:

- **推理** 用户到来时，系统会将用户状态发送给演员网络，演员网络根据高斯分布采样动作。随后，排序函数将这两者作为输入

表3：在线实验中不同算法与 LTR 基线的性能比较。

算法	观看时长	分享	下载	评论
RCPO	+0.309%	−0.707%	0.153%	−1.313%
Interaction-AC	+0.117%	+5.008%	+1.952%	−0.101%
TSCAC	+0.379%	+3.376%	+1.733%	−0.619%

动作和候选项的嵌入，计算动作与视频嵌入之间的点积作为分数，并将得分最高的项目推荐给用户。之后，（状态、动作、奖励、下一状态）会被保存到回放缓冲区中。

- **训练** 演员网络和评论家网络使用从回放缓冲区中采样的小批量数据（状态、动作、奖励、下一状态）进行训练。

对比算法。 我们还使用监督式学习排序（LTR）基线模型进行评估，该模型是平台上默认运行的模型。

- **RCPO**：遵循 [Tessler等人 2018], 的方法，我们定义一个组合奖励 $l_t + \lambda i_t$，并学习一种策略，以折扣因子为 0.95 最大化累积组合奖励，其中 λ 为拉格朗日乘子。
- **TSCAC**：我们首先学习一种策略 π_2 来优化辅助目标。然后，我们学习一种策略 π_1 来优化主要目标，并施加软约束，使 π_1 接近 π_2。– **Interaction-AC**：在第一阶段，我们学习一种策略 π_2来最大化交互奖励，评论家按照 (2) 更新，演员按照 (3) 更新。– **TSCAC** 在第二阶段，我们学习一种主策略 π_1 来最大化观看时长的累积奖励，并通过软正则化使 π_1 接近 π_2，评论家按照 (4) 更新，演员按照 (7) 更新。• **LTR（基线）**：Liu等人 2009]提出的学习排序模型 [以用户状态嵌入和视频嵌入为输入，并拟合响应总和。

实验详情。 为了测试不同的算法，我们将平台上的用户随机划分为几个组。第一组运行 LTR 基线模型，其余各组分别运行 RCPO、Interaction-AC 和 TSCAC 模型。模型训练几天后便固定下来，以测试其一天内的性能。

6.2 结果

表3展示了各算法相较于 LTR 基线在观看时长、分享、下载和评论等指标上的性能提升。可以看出，与基线相比，RCPO 能够学习并提升观看时长；但相对于观看时长而言，交互信号过于稀疏，因此将这些响应结合起来时，无法有效地平衡

互动效果良好。Interaction-AC 算法的性能符合预期：仅凭交互奖励的信号，它就能学会提升与互动相关的指标（分享、下载、评论）；用户与视频之间的此类互动也能改善

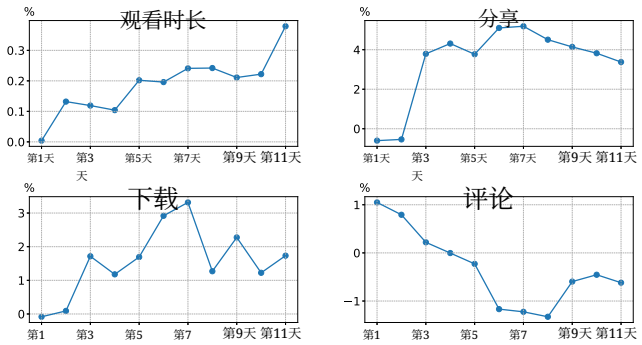


图 5：TSCAC 相较于每日 LTR 基线的在线性能差距。

用户的观看时长，因为推荐了更多有趣且很可能引发互动的视频，从而优化了用户的整体体验。最后，TSCAC 算法取得了最佳性能：与 RCPO 相比，它的观看时长更长，互动指标也有显著提升，这得益于训练过程中有效的软正则化，使其不会偏离 Interaction-AC 策略太远。请注意，在短视频平台上，观看时长提升 0.1%、互动提升 1% 均具有统计显著性。也就是说，我们提出的方法相较于基线方法取得了显著的性能提升。所有强化学习方法的评论指标普遍下降，是观看时长与评论之间自然权衡的结果。

为了解 TSCAC 算法如何学会平衡主要目标和辅助目标，图 5 展示了第二阶段相较于 LTR 基线，在观看时长和互动方面的在线性能差距。如图所示，在 Interaction-AC 策略的约束下，算法在初期迅速学会提升分享和评论这两项互动指标。随后，模型逐渐学会随着时间推移提升观看时长，同时只牺牲少量互动。值得注意的是，TSCAC 的线上性能在每个维度上都显著优于 RCPO，这证明了我们方法的有效性。

7 结论

本文研究了短视频平台中如何在多个辅助稀疏约束下优化主要累积响应的问题。为应对多个约束带来的挑战，我们提出了一种新颖的约束强化学习方法，称为 TSCAC，旨在优化短视频平台的主要目标，同时兼顾其他目标。我们的方法包含多重评论家估计和两个学习阶段。在第一阶段，我们针对每个辅助响应分别学习一种策略，以优化其累积奖励。在第二阶段，我们学习主要策略，以优化主要响应的累积值，并通过软约束限制该策略，使其接近针对其他响应进行优化的策略。我们通过大量离线评估和在线实验，展示了本方法相较于现有替代方案的优势。未来，将本方法应用于其他推荐系统颇有前景。研究 TSCAC 确定性版本的性能也是一项有趣的后续工作。

参考文献

M Mehdi Afsar、Trafford Crump 和 Behrouz Far。2021。基于强化学习的推荐系统：综述。*arXiv*预印本 *arXiv:2101.06286*（2021）。Md Hijbul Alam、Woo-Jong Ryu 和 SangKeun Lee。2016。联合多粒度主题情感：在线评论语义方面的建模。信息科学 339（2016），206–223。Haokun Chen、Xinyi Dai、Han Cai、Weinan Zhang、Xuejian Wang、Ruiming Tang、Yuzhou Zhang 和 Yong Yu。2019b。基于树结构策略梯度的大规模交互式推荐。载于人工智能 *AAAI* 会议论文集，第33卷。3312–3320。Minmin Chen、Alex Beutel、Paul Covington、Sagar Jain、Francois Belletti 和 Ed H Chi。2019a。REINFORCE 推荐系统的 Top-k 离策略校正。载于第十二届 *ACM* 网络搜索与数据挖掘国际会议论文集。456–464。Shi-Yong Chen、Yang Yu、Qing Da、Jun Tan、Hai-Kuan Huang 和 Hai-Hong Tang。2018。利用强化学习稳定动态环境，并将其应用于在线推荐。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议。1187–1196。Xu Chen、Yali Du、Long Xia 和 Jun Wang。2021。基于用户多方面偏好的强化推荐。载于*2021*年万维网大会论文集。425–435。Heng-Tze Cheng、Levent Koc、Jeremiah Harmsen、Tal Shaked、Tushar Chandra、Hrishi Aradhye、Glen Anderson、Greg Corrado、Wei Chai、Mustafa Ispir 等。2016。用于推荐系统的宽度与深度学习。载于首届推荐系统深度学习研讨会论文集。7–10。Yinlam Chow、Mohammad Ghavamzadeh、Lucas Janson 和 Marco Pavone。2017。采用百分位风险准则的风险约束强化学习。机器学习研究期刊 18，1（2017），6070–6120。Gal Dalal、Krishnamurthy Dvijotham、Matej Vecerik、Todd Hester、Cosmin Paduraru 和 Yuval Tassa。2018。连续动作空间中的安全探索。*arXiv*预印本 *arXiv:1801.08757*（2018）。Gabriel Dulac-Arnold、Richard Evans、Hado van Hasselt、Peter Sunehag、Timothy Lillicrap、Jonathan Hunt、Timothy Mann、Theophane Weber、Thomas Degris 和 Ben Coppin。2015。大规模离散动作空间中的深度强化学习。*arXiv*预印本 *arXiv:1512.07679*（2015）。Chongming Gao、Wenqiang Lei、Jiawei Chen、Shiqi Wang、Xiangnan He、Shijun Li、Biao Li、Yuan Zhang 和 Peng Jiang。2022a。CIRS：通过反事实交互式推荐系统打破过滤气泡。*arXiv*预印本 *arXiv:2204.01266*（2022）。Chongming Gao、Shijun Li、Yuan Zhang、Jiawei Chen、Biao Li、Wenqiang Lei、Peng Jiang 和 Xiangnan He。2022b。KuaiRand：包含随机曝光视频的无偏序列推荐数据集。载于第*31*届*ACM*信息与知识管理国际会议论文集（美国佐治亚州亚特兰大）*(CIKM '22)*。5页。https://doi.org/10.1145/3511808.3557624 Javier García 和 Fernando Fernández。2015。安全强化学习综合综述。机器学习研究期刊 16，1（2015），1437–1480。Yingqiang Ge、Shuchang Liu、Ruoyuan Gao、Yikun Xian、Yunqi Li、Xiangyu Zhao、Changhua Pei、Fei Sun、Junfeng Ge、Wenwu Ou 等。2021。迈向推荐中的长期公平性。载于第*14*届*ACM*网络搜索与数据挖掘国际会议论文集。445–453。Yingqiang Ge、Xiaoting Zhao、Lucia Yu、Saurabh Paul、Diane Hu、Chu-Cheng Hsieh 和 Yongfeng Zhang。2022。通过强化学习实现推荐中帕累托高效的公平性-效用权衡。*arXiv*预印本 *arXiv:2201.00140*（2022）。Xudong Gong、Qinlin Feng、Yuan Zhang、Jiangling Qin、Weijie Ding、Biao Li 和 Peng Jiang。2022。移动设备上的实时短视频推荐。载于第*31*届*ACM*信息与知识管理国际会议论文集（美国佐治亚州亚特兰大）*(CIKM '22)*。Huifeng Guo、Ruiming Tang、Yunming Ye、Zhenguo Li 和 Xiuqiang He。2017。DeepFM：用于点击率预测的基于因子分解机的神经网络。*arXiv*预印本 *arXiv:1703.04247*（2017）。Timothy P Lillicrap、Jonathan J Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。2015。基于深度强化学习的连续控制。*arXiv*预印本 *arXiv:1509.02971*（2015）。Xiao Lin、Hongjie Chen、Changhua Pei、Fei Sun、Xuanji Xiao、Hanxiao Sun、Yongfeng Zhang、Wenwu Ou 和 Peng Jiang。2019。用于电子商务推荐多目标优化的帕累托高效算法。载于第*13*届*ACM*推荐系统会议论文集。20–28。Zihan Lin、Hui Wang、Jingshu Mao、Wayne Xin Zhao、Cheng Wang、Peng Jiang 和 Ji-Rong Wen。2022。利用解耦表示实现面向特征的多样化重排序，以提升相关性推荐。载于第*28*届*ACM SIGKDD*知识发现与数据挖掘大会。3327–3335。Dong Liu 和 Chenyang Yang。2019。一种面向移动用户的主动内容推送与推荐深度强化学习方法。*IEEE Access* 7（2019），83120–83136。Tie-Yan Liu 等。2009。信息检索中的学习排序。基础与趋势® 信息检索 3，3（2009），225–331。Yongshuai Liu、Avishai Halev 和 Xin Liu。2021。无模型强化学习中的带约束策略学习：综述。载于第三十届国际人工智能联合会议论文集。Jiaqi Ma、Zhe Zhao、Xinyang Yi、Ji Yang、Minmin Chen、Jiaxi Tang、Lichan Hong 和 Ed H Chi。2020。两阶段推荐系统中的离策略学习。载于*2020*年万维网大会论文集。463–473。Hossam Mossalam、Yannis M Assael、Diederik M Roijers 和 Shimon Whiteson。2016。多目标深度强化学习。*arXiv*预印本 *arXiv:1610.02707*（2016）。Ashvin Nair、Murtaza Dalal、Abhishek Gupta 和 Sergey Levine。2020。AWAC：利用离线数据集加速在线强化学习。（2020）。Shamim Nemati、Mohammad M Ghassemi 和 Gari D Clifford。2016。基于次优临床实例的最佳用药剂量：一种深度强化学习方法。载于*2016*年第*38*届*IEEE*医学与生物学工程学会年会国际会议（*EMBC*）。IEEE，2978–2981。Thanh Thi Nguyen、Ngoc Duy Nguyen、Peter Vamplew、Saeid Nahavandi、Richard Dazeley 和 Chee Peng Lim。2020。一个多目标深度强化学习框架。《人工智能工程应用》96（2020），103915。Qi Pi、Guorui Zhou、Yujing Zhang、Zhe Wang、Lejian Ren、Ying Fan、Xiaoqiang Zhu 和 Kun Gai。2020。利用终身序列行为数据进行基于搜索的用户兴趣建模，以实现点击率预测。载于第*29*届*ACM*信息与知识管理国际会议论文集。2685–2692。Doina Precup。2000。用于离策略策略评估的资格迹。计算机科学系教职员工出版系列（2000），80。Doina Precup、Richard S Sutton 和 Sanjoy Dasgupta。2001。结合函数逼近的离策略时序差分学习。载于*ICML*。417–424。Ozan Sener 和 Vladlen Koltun。2018。将多任务学习视为多目标优化。*arXiv*预印本 *arXiv:1810.04650*（2018）。Dusan Stamenkovic、Alexandros Karatzoglou、Ioannis Arapakis、Xin Xin 和 Kleomenis Katevas。2021。两全其美：通过多目标强化学习提供多样且新颖的推荐。*arXiv*预印本 *arXiv:2110.15097*（2021）。Richard S Sutton 和 Andrew G Barto。2018。强化学习：导论。MIT出版社。Chen Tessler、Daniel J Mankowitz 和 Shie Mannor。2018。奖励约束策略优化。*arXiv*预印本 *arXiv:1805.11074*（2018）。Jiayin Wang、Weizhi Ma、Jiayu Li、Hongyu Lu、Min Zhang、Biao Li、Yiqun Liu、Peng Jiang 和 Shaoping Ma。2022a。让公平性更公平：公平的物品效用估计与曝光量再分配。载于第*28*届*ACM SIGKDD*知识发现与数据挖掘大会论文集。1868–1877。Yuyan Wang、Mohit Sharma、Can Xu、Sriraj Badam、Qian Sun、Lee Richardson、Lisa Chung、Ed H Chi 和 Minmin Chen。2022b。推荐系统中的长期用户体验替代指标。载于第*28*届*ACM SIGKDD*知识发现与数据挖掘大会论文集。4100–4109。C Ch White、CC III WHITE 和 KIM KW。1980。向量准则马尔可夫决策过程的求解方法。（1980）。Ronald J Williams。1992。用于联结主义强化学习的简单统计梯度跟随算法。强化学习（1992），5–32。Yikun Xian、Zuohui Fu、Shan Muthukrishnan、Gerard De Melo 和 Yongfeng Zhang。2019。用于可解释推荐的强化学习知识图谱推理。载于第*42*届*ACM SIGIR*信息检索研究与发展国际会议论文集。285–294。Xin Xin、Alexandros Karatzoglou、Ioannis Arapakis 和 Joemon M Jose。2022。面向推荐系统的监督式优势演员-评论家方法。载于第*15*届*ACM*网络搜索与数据挖掘国际会议论文集。1186–1196。Ruohan Zhan、Changhua Pei、Qiang Su、Jianfeng Wen、Xueliang Wang、Guanyu Mu、Dong Zheng、Peng Jiang 和 Kun Gai。2022。消除视频推荐观看时长预测中的时长偏差。载于第*28*届*ACM SIGKDD*知识发现与数据挖掘大会论文集。4472–4481。Xiangyu Zhao、Liang Zhang、Zhuoye Ding、Long Xia、Jiliang Tang 和 Dawei Yin。2018。通过成对深度强化学习利用负反馈进行推荐。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。1040–1048。Xiangyu Zhao、Liang Zhang、Long Xia、Zhuoye Ding、Dawei Yin 和 Jiliang Tang。2017。用于列表式推荐的深度强化学习。*arXiv*预印本 *arXiv:1801.00209*（2017）。Lixin Zou、Long Xia、Zhuoye Ding、Jiaxing Song、Weidong Liu 和 Dawei Yin。2019。通过强化学习优化推荐系统中的长期用户参与度。载于第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。2810–2818。

定理1的证明

定理2。 方程 (5) 的拉格朗日函数具有闭式解

$$\pi^*(a|s) \propto \prod_{i=2}^{m} \left(\pi_{\theta_i}(a|s)\right)^{\frac{\lambda_i}{\sum_{j=2}^{m}\lambda_j}} \exp\left(\frac{A_1^{(k)}}{\sum_{j=2}^{m}\lambda_j}\right), \tag{12}$$

其中 λ_i、 $i = 2$、...、 m 是拉格朗日乘子。

证明。式 (5) 的拉格朗日函数为

$$\mathcal{L}(\pi, \lambda_2, \ldots, \lambda_m) = -E_\pi[A_1^{(k)}] + \sum_{i=2}^{m} \lambda_i \left(D_{KL}(\pi||\pi_{\theta_i}) - \epsilon_i\right). \tag{13}$$

计算 $\mathrm{L}(\pi, \lambda_2, \ldots, \lambda_m)$ 关于 π 的梯度，

$$\frac{\partial \mathcal{L}}{\partial \pi} = -A_1^{(k)} + \sum_{i=2}^{m} \lambda_i \left(1 + \log(\pi(a|s)) - \log(\pi_{\theta_i}(a|s))\right). \tag{14}$$

设定 $\frac{\partial \mathcal{L}}{\partial \pi} = 0$，我们得到满足以下条件的解 π^*

$$\pi^*(a|s) = \frac{1}{Z(s)} \prod_{i=2}^{m} \left(\pi_{\theta_i}(a|s)\right)^{\frac{\lambda_i}{\sum_{j=2}^{m}\lambda_j}} \exp\left(\frac{A_1^{(k)}}{\sum_{j=2}^{m}\lambda_j}\right), \tag{15}$$

其中 $Z(s)$ 是配分函数，使得 $\int_a \pi^*(a|s) = 1$。 □

B 两阶段约束演员-评论家算法

算法 1: 两阶段 约束 演员-评论家 (TSCAC)

第一阶段： 对于每个辅助响应 $i = 2, \ldots\ldots, m$，学习一个策略以优化响应 i，其中 π_{θ_i} 表示演员， V_{ϕ_i} 表示评论家。

当尚未收敛时，在迭代 k:

$$\phi_i^{(k+1)} \leftarrow \arg\min_{\phi} E_{\pi_{\theta_i^{(k)}}} \left[\left(r_i(s,a) + \gamma_i V_{\phi_i^{(k)}}(s') - V_\phi(s)\right)^2\right],$$

$$\theta_i^{(k+1)} \leftarrow \arg\max_{\theta} E_{\pi_{\theta_i^{(k)}}} \left[A_i^{(k)} \log\left(\pi_\theta(a|s)\right)\right].$$

第二阶段： 对于主要响应，学习一个策略以优化主要响应，并将其作用域限制在辅助响应的策略 $\{\pi_{\theta_i}\}_{i=2}^{m}$ 附近，其中 π_{θ_1} 表示演员， V_{ϕ_1} 表示评论家。当尚未收敛时

, 在第 k次迭代时:

$$\phi_1^{(k+1)} \leftarrow \arg\min_{\phi} E_{\pi_{\theta_1^{(k)}}} \left[\left(r_1(s,a) + \gamma_1 V_{\phi_1^{(k)}}(s') - V_\phi(s)\right)^2\right],$$

$$\theta_1^{(k+1)} \leftarrow \arg\max_{\theta} E_{\pi_{\theta_1^{(k)}}} \Bigg[\frac{\prod_{i=2}^{m}\left(\pi_{\theta_i}(a|s)\right)^{\frac{\lambda_i}{\sum_{j=2}^{m}\lambda_j}}}{\pi_{\theta_1^{(k)}}(a|s)} \times \exp\left(\frac{A_1^{(k)}}{\sum_{j=2}^{m}\lambda_j}\right) \log \pi_\theta(a|s)\Bigg].$$

输出： 受约束的策略 π_1 .

C TripAdvisor 离线实验

我们的代码参考了 https://github.com/AIDefender/TSCAC。

在本节中，我们通过大量离线学习仿真，在另一个公开数据集上评估我们的方法。我们展示了与现有基线方法相比，我们的方法在实现主要目标和兼顾辅助目标方面均具有有效性。

数据集。 我们采用一个名为 *TripAdvisor* 的酒店评论数据集，该数据集是研究具有多种响应的推荐系统策略优化的标准数据集 [Chen etal. 2021]。在该数据集中，顾客不仅会为酒店提供总体评分，还会从多个方面对酒店进行评分，包括服务、商务、清洁度、入住、性价比、客房和位置 [Alam et al. 2016]。[1]同一用户提供的评论按时间顺序连接起来，形成一条轨迹；我们筛除了长度小于 20 的轨迹。总计有 20277 位顾客、150 家酒店和 257932 条评论。

马尔可夫决策过程。 一条轨迹记录了顾客的酒店评论历史。对于每条评论，我们有状态 s_t：顾客 ID、最近评论过的三家酒店的酒店 ID，以及相应的多方面评论评分；动作 a_t：当前评论的酒店 ID；奖励 r_t：顾客针对所评论酒店提供的八项评分组成的向量，分别对应服务、商务、清洁度、入住、性价比、客房、位置和总体评分；折扣因子 γ：0.99。我们将最大化累计总体评分设为主要目标，并将其他目标视为辅助目标。

性能。 表 4 展示了不同算法在八项评分上的结果。可以看出，我们的 TSCAC 算法在所有算法中表现最佳：对于主要目标，TSCAC 的总体评分最高，达到 3。99；对于辅助目标，TSCAC 在 7 项评分中的 5 项（服务、清洁度、性价比、客房和位置）也排名最高。帕累托算法确实学到了一个帕累托最优解，在入住评分和商务评分上表现最佳，但这并不符合本文的设定，因为本文的主要目标是优化总体评分。
RCPO-Multi-Critic 在 6 项评分上优于 RCPO，这验证了相较于联合评论家估计，多评论家估计的效果。
RCPO-Multi-Critic 算法取得了与我们的方法相同的最佳总体评分，但在其他评分上有所牺牲，尤其是其位置评分甚至低于行为克隆算法。

D 其他指标下的离线评估

我们还使用折损累计增益（DCG）指标评估 TSCAC 和基线方法的性能，如表 5 所示。

[1]该数据集包含主要目标和其他响应，也可用于评估推荐系统中的约束策略优化。

表 4：不同算法在 TripAdvisor 上的性能

算法	服务	商务	清洁度	入住	性价比	客房	位置	总体评分
BC	3.38	-1.86	3.57	-0.73	3.32	2.92	2.93*	3.92
Wide&Deep	3.41*	-1.86	3.62*	-0.75	3.36*	2.96	2.88	3.98
RCPO	3.40	-1.82	3.61	-0.71	3.34	2.95	2.91	3.97
RCPO-Multi-Critic	3.41*	-1.82	3.62*	-0.68	3.35	2.97*	2.87	3.99*
帕累托	3.36	−1.79*	3.57	−0.62*	3.29	2.93	2.86	3.95
TSCAC	**3.43**	-1.82	**3.64**	-0.68	**3.37**	**3.00**	**2.98**	**3.99**

∗ 表示在每个响应维度上所有基线方法中的最佳性能；当 TSCAC 达到最佳性能时，最后一行的数据以粗体标出。

表 5：不同算法在 DCG 方面的性能。

算法	点击↑ (e+2)	点赞↑	评论↑	仇恨↓(e-2)	观看时长↑(e+3)
BC	4.617	8.492	2.320	8.137	1.079
Wide&Deep	4.576 −0.88%	8.158 −3.93%	2.274 −1.97%	7.503 −7.79%*	1.073 −0.56%
DeepFM	4.594 −0.50%	8.439 −0.63%	2.274 −1.97%	8.334 2.43%	1.083 0。41%*
RCPO	4.603 −0.30%	8.537 0.53%*	2.282 −1.60%	8.336 2.45%	1.080 0.12%
RCPO-Multi-Critic	4.569 −1.03%	8.168 −3.81%	2.272 −2.09%	7.538 −7.36%	1.069 −0.88%
帕累托	4.605 −0.26%*	8.494 0。02%	2.304 −0.69%*	8.449 3.83%	1.074 −0.48%
TSCAC	**4.617** **0.01%**	**8.652** **1.88%**	**2.378** **2。49%**	8.036 −1.24%	1.083 0。39%

↑：越高越好；↓：越低越好。括号中的数字表示该列的单位；每种算法所在首行的数字为 DCG 分数。第二行中的百分比表示该算法与行为克隆算法之间的性能差距。带有 ∗ 的数字表示各响应维度中所有基线方法的最佳性能。若 TSCAC 在某个响应维度上达到最佳性能，则最后一行以粗体标示。