通过因果感知强化学习的极小后悔最优化实现对抗性受限出价

王昊哲1) 杜超1) 庞攀原1) 贺李1) 王梁1) 郑波1)

1）(北京阿里巴巴团队,北京市，中国，100000)

摘 要 互联网的普及导致了在线广告的出现，其发展受在线拍卖的机制推动。在这些重复的拍卖中，软件代理代表聚合广告商参与，以优化其长期效益。为满足多样化的需求，出价策略被用来优化广告目标，同时受到不同的支出限制约束。现有的受限制出价方法通常依赖于i.i.d.（独立同分布）的训练和测试条件，这与在线广告市场的对抗性特质相矛盾，其中不同的各方可能拥有潜在的冲突目标。在这方面，我们探讨了在对抗性出价环境中的受限制出价问题，假设对对抗因素没有任何知识。我们的见解是，不依赖于i.i.d.假设，而是将环境的训练分布与潜在的测试分布相一致，同时最小化策略后悔。基于这一见解，我们提出了一种实用的极小后悔优化（MiRO）方法，该方法交替进行，一方面是为了找到供教学的对抗环境的教师，另一方面是为了在给定环境分布上元学习其策略的学习者。此外，我们还首创了将专家演示纳入学习出价策略的方法。通过一个关注因果关系的策略设计，我们通过从专家那里获得知识来改进MiRO。在工业数据和合成数据上进行的大量实验表明，我们的方法，因果感知强化学习（MiROCL），表现优于先前的方法，提高了超过30%。

关键词 受限出价，强化学习，因果关系

**Adversarial Constrained Bidding via Minimax Regret Optimization with Causality-Aware Reinforcement Learning**

Haozhe-Wang1) Chao-Du1) Panyan-Pang1) Li-He1) Liang-Wang1) Bo-Zheng1)

1)(Alibaba Group, City beijing, 100000)

**Abstract** The proliferation of the Internet has led to the emergence of online advertising, driven by the mechanics of online auctions. In these repeated auctions, software agents participate on behalf of aggregated advertisers to optimize for their long-term utility. To fulfill the diverse demands, idding strategies are employed to optimize advertising objectives subject to different spending constraints. Existing approaches on constrained bidding typically rely on i.i.d. train and test conditions, which contradicts the adversarial nature of online ad markets where different parties possess potentially conflicting objectives. In this regard, we explore the problem of constrained bidding in adversarial bidding environments, which assumes no knowledge about the adversarial factors. Instead of relying on the i.i.d. assumption, our insight is to align the train distribution of environments with the potential test distribution meanwhile minimizing policy regret. Based on this insight, we propose a practical Minimax Regret Optimization (MiRO) approach that interleaves between a teacher finding adversarial environments for tutoring and a learner meta-learning its policy over the given distribution of environments. In addition, we pioneer to incorporate expert demonstrations for learning bidding strategies. Through a causality-aware policy design, we improve upon MiRO by distilling knowledge from the experts. Extensive experiments on both industrial data and synthetic data show that our method, MiRO with Causality-aware reinforcement Learning (MiROCL), outperforms prior methods by over 30%.

**Key Words** Constrained Bidding, Reinforcement Learning, Causality

**1** 介绍

互联网的广泛普及导致在线广告成为一个价值数十亿美元的多元产业。在线广告的核心是在线拍卖[17]，在这里，发布商反复出售广告位给寻求品牌推广、提高转化率等的广告商。传统上，广泛采用了激励兼容拍卖，如二价拍卖，因为它们具备“真实出价”的理想特性，对于短视的投标者来说，真实地揭示私人价值是为了最大化他们的即时效益是最优选择[32, 34]。

然而，最近时代已经过时，短视的投标者的关键假设已经过时，真实出价不再优化广告商的长期效益。作为代表聚合广告商的中介，需求方平台（DSPs）现在是每天参与数十亿次拍卖的实际实体。与真实出价不同，DSP代理采用出价策略来满足各种广告商的需求，这些广告商通常在受到支出限制的情况下寻求最大化某些效益[44]。例如，品牌广告商寻求长期增长和知名度，通常会针对印象、点击等指标进行优化，同时受到投资回报率（ROI）的限制，要求效益与成本的最低比率。

为了满足广告商的多样化需求，已经进行了广泛的研究，以设计和学习出价策略。现有文献可以根据约束设置广泛分类。大部分研究集中在出价受到最多预算约束的情况下[6, 7, 12, 13, 20, 21, 24]，这可能无法完全涵盖该领域中不同支出约束的多样性。为了解决这一限制，一些研究[27, 41, 47] 探讨了在类似ROI的约束下的最优出价。ROI受限出价（ROI-Constrained Bidding，RCB）问题涉及确保类似ROI的约束（如成本收益比和点击成本比）超过预定限制的同时遵守预算约束，被视为概括了多样化广告目标的典型问题[41]。

尽管之前的方法[27, 41]取得了令人鼓舞的结果，但它们通常遵循经验风险最小化（Empirical Risk Minimization，ERM）原则[31, 42]，依赖于独立同分布（i.i.d.）的训练和测试条件的假设。这与现实世界中的许多情况相矛盾，其中卖家和其他竞争投标者的行为是对抗性的，因为所有各方都寻求优化自己的效益，这些效益可能相互冲突[34]。例如，卖家可能会了解投标者的私人价值分布，并在拍卖中设置个性化的保留价格[14, 18, 35]。最近的研究[16, 37]引入了基于神经网络的销售机制，通过数据学习而来。此外，竞争对手投标者也可以采用复杂的出价策略来优化他们的长期效益[15]，导致竞争性出价的复杂分布，可能影响我们的出价代理的性能[26]。这些考虑指出了出价环境的固有对抗性特质（也在[41]中总结为非平稳性）。

在对抗性环境中进行出价的问题一直未被广泛探讨，仅有少数最近的研究[26, 33]显示了一些进展。这些工作集中于在一等拍卖中没有约束条件下的对抗性出价[26]，或者依赖于对对手的假设[33]。然而，在这篇论文中，我们要研究一个未被探讨的问题，即在黑匣子对抗性环境下的受限制出价，这假定对于外部因素如何导致对抗性扰动对出价环境没有任何知识。从博弈论的角度来看，黑匣子对手有意扰乱出价环境，如改变市场动态或价值分布，利用对投标者的了解。因此，投标者将受到可能表现更差的测试环境的影响，必须在对抗性环境中表现出适应性，以实现最佳性能。

为了解决独立同分布（i.i.d.）假设不成立的问题，我们的基本见解是将环境的训练分布与潜在的测试分布相一致，同时最小化策略后悔，即策略与预先计算的最优策略（称为“专家”策略）之间的性能差异。基于这一见解，我们提出了一个极小后悔优化（Minimax Regret Optimization，MiRO）框架，该框架交替进行，既在内部最高问题中确定对齐的训练分布，又在外部最低问题中优化在这种分布下最小化后悔的策略。虽然MiRO看似吸引人，但我们发现内部和外部问题都存在实际限制，需要改进以实现最佳性能：

* 内部问题（第3.2节）：由于对环境的确切功能结构和对抗性因素的缺乏了解，直接优化分布的可行性问题变得不可行。为解决这个问题，我们提出了一种数据驱动方法，通过重构世界模型的因果结构来学习对抗性因素的潜在表示。这使得可以进行端到端优化的可微分博弈。
* 外部问题（第3.3节）：尽管后悔最小化的目标是缩小策略与专家之间的差距，策略学习实际上会退化为一个没有专家影响的价值最大化问题。为了解决这个问题，我们试图明确利用来自专家的有用知识来引导策略学习。令人惊讶的是，我们发现直接的行为克隆方法由于未观察到的混淆问题而无法奏效[36]。为克服这一挑战，我们开发了一种因果感知的对齐策略，它可以分解为一个子策略，模拟专家的因果结构。

我们的方法，基于因果感知的极小后悔优化（MiROCL），在大规模合成数据和工业数据上进行了验证，证实了其有效性和通用性。

2 背景和基础知识

在这一部分，首先我们描述标准的受限制出价问题，然后介绍了强化学习的一般公式作为我们研究的基础。

实时出价（Real-time Bidding，RTB）是在线广告中的一个重要营销渠道，它使广告商能够跨多种媒体获得曝光，并帮助发布商通过有效分发流量来实现变现[17]。广告商依赖需求方平台（Demand-Side Platforms，DSPs），这些平台代表他们购买和展示广告。DSP代理每天与数十亿次拍卖互动，并采用出价策略来优化广告商的长期目标，同时受到各种支出约束的限制，这引发了受限制出价的大量研究兴趣[12, 20, 21, 24, 41]。

传统上，受限制出价问题考虑了一个由顺序到达的拍卖组成的RTB过程，其目标是为每次拍卖安排出价，以优化目标效益同时满足相关约束。假设出价过程包括𝑇次重复的拍卖。在为广告机会触发的每次拍卖中，出价代理获得（部分）关于拍卖的信息 𝑥𝑖，其中包括有关用户、所选广告和显示上下文的关键详细信息。基于这些信息，代理必须决定出价价格𝑏𝑖。如果出价高于市场价格𝑚𝑖 = max𝑏i（竞争性最高出价），代理赢得拍卖，表示为1𝑏𝑖>𝑚𝑖。获胜的拍卖会根据发布商规定的销售机制收取广告展示费用𝑐𝑖，并发送有关效益的反馈，如点击和转化。相反，输掉的拍卖只会导致一条宽松的通知。在这项工作中，我们假设在线拍卖采用（或源自）二价拍卖，具有激励兼容性的特性[32]。