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-Wang[1)] Chao-Du[1)] Panyan-Pang[1)] Li-He[1)] Liang-Wang[1)] Bo-Zheng[1)]

[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$mi = maxb\bar{\imath}$（竞争性最高出价），代理赢得拍卖，表示为$1_{bi>mi}$。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]。

接下来，我们将重点关注 ROI 受限出价（ROI-Constrained Bidding，RCB）的设置，这作为一个原型问题，可以概括多样化的广告目标[27, 41]。RCB 问题的目标是在预算约束$1_{C\leq B}$和投资回报率（ROI）约束$1_{\mathrm{ROI}\geq L}$的情况下最大化累积效用U：

$$\max_{b} U_T(\boldsymbol{b};\boldsymbol{x}), s.t.$$

$$\frac{U_T(\boldsymbol{b};\boldsymbol{x})}{C_T(\boldsymbol{b};\boldsymbol{x})} \geq L, B - C_T(\boldsymbol{b};\boldsymbol{x}) \geq 0 \quad (1)$$

其中，粗体字母x和b表示拍卖的序列（特征）和出价，而下面我们将使用$U_T(\boldsymbol{b};\boldsymbol{x}) \stackrel{\text{def}}{=} \sum_{i=1}^{T} E[u_i|x_i]\, 1_{b_i>m_i}$代表累计效用，$C_T(\boldsymbol{b};\boldsymbol{x}) \stackrel{\text{def}}{=} \sum_{i=1}^{T} E[c_i|x_i]\, 1_{b_i>m_i}$代表累计成本。

现有的 RCB 方法已经在强化学习（RL）的公式基础上取得了令人鼓舞的结果，这是因为 RL 具有长期规划的能力[27, 41]。沿着这一趋势，我们采用了 CBRL 中提出的部分可观测受限制 MDP（POCMDP）公式作为我们工作的基础。接下来，我们简要总结 POCMDP 公式的主要思想，并建议读者参考 Wang 等人的详细信息[41]。

许多领先的 DSP，如谷歌[23]和阿里巴巴[3]，经历着数十亿级别的流量吞吐量，如果将每个拍卖视为一个决策步骤，会导致决策序列过长，这对 RL 训练构成了挑战。为了缓解这个问题，CBRL 采用了一个不同寻常的视角，将 RCB 问题视为聚合级别的问题，将印象级别的出价决策看作是一个以广告位为单位的出价比例控制问题，具有印象级别的效用预测，依据以下受限制出价问题的最优出价定理[6, 27, 41]。

定理 1. 在二价拍卖中，问题（1）的最优出价函数采用线性形式，即$bi = aui$（a>0）。

这个定理说明，最优出价（事后看来）等于印象价值u与一个比率a的线性加权。因此，CBRL 提出将每个拍卖视为一个决策步骤，控制时间窗口内的广告位级出价比率a，最终的出价可以通过将出价比率与每个印象级别的效用u相乘来计算。

基于这种广告位级比率控制的形式，CBRL 提出将出价过程建模为具有H个时间步的有限时间段的强化学习问题。每个时间步t代表一个时间窗口$[jt, jt+1)$，其中包含了拍卖$\{xi\} i \in [jt, jt+1)$。鉴于市场价格只有在赢得拍卖时才能知道，POCMDP 引入了一个观测空间O，除了完整状态空间 S，以考虑这种部分可观察性。S和O都包含了广告位级别的统计信息（如获胜率、投资回报率、总收入和成本），但S还包括代理无法观察到的信息（如市场价格）。在这个框架内，行动$a \in A$被定义为安排给每个时间段的出价比例。动态模型$P(s', r|s, a) = \mathbb{T}(s'|s, a) \cdot P(r|s, a)$用于概念上考虑过渡和奖励函数，但确切的函数映射是未知的。

具体来说，对于过渡模型，我们假设市场的部分可观察性，因为密封竞标拍卖中的销售机制和竞争策略不是透明的。对于奖励模型，分步奖励应在概念上考虑效用和约束违规，这涉及为每个时间段分配非平凡的信用分配。尽管动态模型未知，只要市场价格已知，我们仍然可以模拟出价环境。在这方面，过去的出价日志可以构建大量的出价环境，作为我们的数据集。此外，我们可以通过解线性规划[41]来为每个环境计算最优决策序列，我们将在接下来的部分中称之为专家轨迹。

在上述 POCMDP 公式中，我们的目标是找到一个策略π，它属于策略空间$\Pi: O \times \mathcal{H} \to P(\mathcal{A})$。策略的输入包括过去的轨迹$h_t = \{o_i, a_i, r_i\}_{i=1}^{t-1} = 1$以及当前的观察值$ot$，这是部分可观察 MDP 中的常见做法[45]。确定一个稳定策略的标准目标是在给定$\mathcal{M}$下最大化策略的值，其由预期的累积奖励$V(\pi; \mathcal{M}) = E[\sum_{t=1}^{H} r_t|\pi, \mathcal{M}]$给出。由于出价环境可能每天都不同，出价策略需要在不同条件下表现出适应性至关重要。为此，先前的方法通常采用以下强化学习目标：

$$\max_{\pi} E_{\mathcal{M}}[(\pi; \mathcal{M})] \tag{2}$$

这个目标优化了在 MDP 的分布$p(\mathcal{M})$上的策略，假设测试环境的分布与训练分布是独立同分布的。基本上，这个目标体现了元强化学习的原则[42, 48]，旨在元学习一个能够在多个环境中泛化的自适应策略。

3 方法论

虽然最近人们普遍认为在线广告市场动态变化[27, 41]，但我们强调出价环境本质上可能是对抗性的[26, 33]，因为在线拍卖涉及多方利益冲突。例如，卖家可能会调整他们的机制以实现最大收入，例如通过学习个性化的保留价格[14, 18]，甚至可以从数据中自动学习机制[16, 37]。另一方面，竞争对手可以采用数据驱动的自动出价算法来优化他们自己的效用。此外，由于外部因素的影响，用户点击广告的倾向随时间变化，导致效用估计不准确[19, 34]。实际数据观察也支持这些猜测，详见附录。

不幸的是，对于我们的代理来说，这些对抗因素都无法直接观察到，因为在线拍卖通常会封存竞争性出价，而卖家也没有太多动机透露他们如何更新他们的销售机制。鉴于这一情况，我们在本文中探讨了受限制出价在对抗性环境中的未知问题（即对抗性受限制出价），假设对对抗因素如何引起环境的扰动一无所知。从博弈论的角度来看，我们的目标是设计一个出价策略，可以有效地抵制下一轮的黑箱对手，利用之前轮次的互动历史。

对抗性受限制出价尤其具有挑战性，因为广泛采用的独立同分布训练和测试环境的假设被违反了，因为对抗性设置中的测试环境可以被有意地操纵以不利于我们，如图 1 所示。因此，遵循这一假设的现有工作不适用于对抗性受限制出价。

为了解决这个问题，我们的主要见解是将环境的训练分布与潜在的测试分布相匹配，而不是依赖于独立同分布的假设。接下来，我们将提供一个将这一见解数学化为极小后悔优化（MiRO）框架的实际解决方案。然后我们将讨论几个实际问题，并详细说明

如何改进粗糙的 MiRO 框架。

3.1 MiRO 框架

在本节中，我们首先讨论了与使用收集的出价日志实现训练-测试分布对齐相关的两个基本问题，最终提出了用于对抗性受限制出价的极小后悔优化（MiRO）框架。具体来说，要实现训练-测试对齐，我们首先必须回答以下两个问题：对于测试分布，对抗性设置意味着什么性质？其次，在给定这种性质的情况下，如何确定与之对齐的训练分布?

3.1.1 测试分布的性质

从博弈论的角度来看，一个拥有完全了解我们之前策略的对手可以扰乱环境，将其变为这个策略的最坏情况。虽然尚不清楚这样的对手如何在实践中可行地利用我们的策略，但严格的条件为我们提供了有关通用对抗设置的有价值的见解。具体来说，我们认为在测试过程中遇到的环境的可能性与策略在该环境中的表现成正比。

为了在数学上表达这个思想，我们首先必须引入策略（即出价策略）的性能度量。在对抗性设置中，没有适用于所有环境的固定策略，因此我们使用遗憾作为相对于每个环境的“神谕”（oracle 或 prophet [40]，但在以下部分中我们称之为专家）的性能度量[9, 26, 40]。遗憾度量了两种策略之间的价值差异，一种是最优策略（即神谕或预言者[40]，但在接下来的部分中我们称之为专家），另一种是正在学习的策略：

$$Reg(\pi, \mathcal{M}) \stackrel{\text{def}}{=} V(\xi^*; \mathcal{M}) - V(\pi; \mathcal{M}) = V^*_{\mathcal{M}} - V(\pi; \mathcal{M}) \quad (3)$$

上面的公式中$\xi^* = argmax_{\xi \in \Xi} V(\xi; M)$代表环境$\mathcal{M}$的专家策略，而$V^*_{\mathcal{M}}$代表其累计价值。值得注意的是，由于我们不知道环境的确切函数结构，我们不能直接求解从任何给定观察到最佳决策的专家策略函数。实际上，我们基于离线出价日志使用近似动态规划来计算专家轨迹，因此概念上，专家示范需要关于未来动态的先见之明。为此，我们在概念上将专家策略空间定义为 $\xi \in \Xi$：$O \times H \times W \mapsto\mapsto P(A)$，它还输入了关于对抗性因素$\omega \in \mathbb{W}$的特权信息。

由于我们必须为测试分布选择一个表示，我们选择使用一般的基于能量的分布来表示上述测试分布的比例性质，

$$P_{test}(\mathcal{M}) = \frac{\exp\left(\frac{1}{\alpha} reg(\pi; \mathcal{M})\right)}{Z(\pi)}, \quad (4)$$

在这里，我们将遗憾函数$Reg(\pi; M)$设置为带有温度α的自由能，分区函数$Z(\pi)$用于规范化分布，尽管它不贡献梯度。

3.1.2 识别与测试分布对齐的训练分布

在建立了测试分布的潜在形式之后，我们现在的重点转向从环境集合中识别合适的训练分布。为了确保对齐的训练分布与可用的训练集保持一致，我们选择将训练分布投影到参数化分布集P（稍后在第 3.2.2 节中定义，它代表了训练集）中，通过KL（Kullback-Leibler）散度，得到一个熵正则化的遗憾最大化目标，

$$\min_{P \in \mathcal{P}} \stackrel{\text{def}}{=} D_{KL}(\mathrm{P} \parallel P_{test}) = \max_{\mathrm{P}(\mathcal{M}) \in \mathcal{P}} E_{\mathcal{M}}[\mathrm{Reg}(\pi; \mathcal{M})] + \alpha \mathcal{H}(\mathrm{P}) + \text{const.} \quad (5)$$

直观地说，这个目标旨在找到在训练集中引起高策略遗憾的环境分布，同时遵循最大熵原则，因为我们对真正的对手一无所知。熵正则化器控制了分布如何朝着具有温度超参数α的最坏情况环境偏移。实际上，这个超参数使得在严格的对抗性设置和 iid 随机设置之间的插值逐渐发生，反映了对对抗性设置的信仰。更具体地说，在$\alpha = 0$时，引发的分布仅关注最坏情况，而在$\alpha \to \infty$时，引发的分布在训练集上均匀分布。

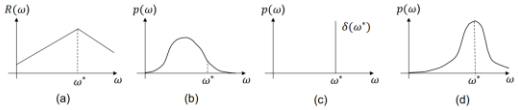

图 1 ERM 与 MiRO 对比。(a)示例了在由一维ω特征表征的不同环境下的$Reg(\pi, \omega)$。(b)展示了训练环境的经验分布。(c)展示了测试环境的分布，违反了 ERM 的 iid 假设。相比之下，MiRO 假设它与$Reg(\pi, \omega)$成比例，如(d)所示。

3.1.3 极小后悔优化

在给定对齐的训练分布的情况下，策略的目标是在该分布下最小化其遗憾。因此，我们得出以下（熵正则化的）极小后悔优化（MiRO）框架：

$$\min_{\pi} \max_{\mathrm{P} \in \mathcal{P}} E_{\mathcal{M}}[Reg(\pi; \mathcal{M})] + \alpha \mathcal{H}(P) \quad (6)$$

MiRO 框架呈现了两个玩家之间的极小极大博弈，其中内部问题寻找一个环境分布$P(M)$，可能与对抗性测试条件对齐，而外部问题优化给定环境下策略的性能。与以前广泛采用的经验风险最小化（ERM）相比，ERM 假设小的训练经验风险可以推广到小的测试风险，MiRO 暗示了一种泛化，因为策略努力最小化（近似）最坏情况的遗憾，从而上界测试遗憾。

虽然 MiRO 看起来很吸引人，但解决这样的双层优化通常是难以处理的。我们的主要思想是将极小极大问题转化为一类“可微分博弈”[4]，以便我们可以借助对偶上升[10]来寻找有用的解决方案，这得到了生成对抗网络[22]及其后续研究的支持。

3.2 可行的 MiRO 算法

在这一部分，我们的目标是使方程（6）成为一个可微分的博弈，由于不可观察的对抗性因素，它受到环境M的未知结构的限制。为了克服这一挑战，我们建议通过重新构建世界模型的因果结构来从出价日志中学习那些对抗性因素的潜在表示$\omega \in W$。通过世界模型重建，我们可以实现两个关键好处。首先，由于对抗性因素ω解释了环境的变化，我们可以在学习的潜在空间中搜索分布$p(\omega)$，替换方程（6）中的环境M为ω。其次，世界模型重建建立了从ω到奖励r的映射，使得通过遗憾函数进行微分成为可能。因此，我们可以直接通过基于梯度的优化来搜索 MiRO 中的分布。

3.2.1 创建一个可微分博弈

为了学习可以反映对环境的因果关系的对抗性因素的表示，我们首先分析环境的因果结构，然后利用变分信息瓶颈（VIB）[2]来进行表示学习，其目标是从输入中学习最大程度压缩的表示，同时保留有关输出的最大信息。

我们首先描述一个折叠时间步的竞价过程的因果图[36]（这是一种用于分析随机变量之间因果关系的增强概率图模型[8]），图 2 显示了两个策略的因果关系：专家$\xi \in \Xi$和策略$\pi \in \Pi$对环境的干预。这两个策略的干预导致了观察到的变量$\{o_t, a_t, h_t\}_{t=1}^{H}$和未观察到的变量$\{s_t\}_{t=1}^{H}$之间的因果关系。然而，在变量$\{\omega_t\}_{t=1}^{H}$和因果关系$\{a_t \longrightarrow \omega_t\}_{t=1}^{H}$（虚线）上存在差异，因为专家在概念上知道特权信息以做出最佳决策。

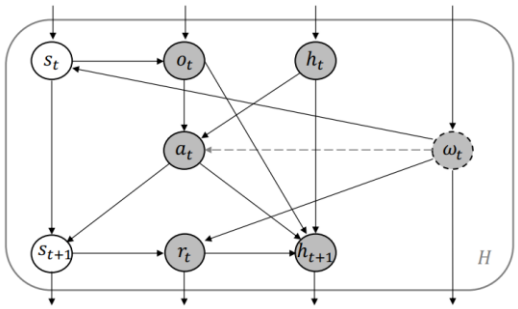

图 2 在时间步骤t上展开的因果图。观察到的变量被着色，而未观察到的变量没有。变量ω_t（虚线边缘）在专家的因果图 G^{ξ}中被观察到，并指向at，而在策略的因果图 G^{π}中未观察到，与a_t没有关联。

基于这些因果关系，我们构建了一个包含以下组件的世界模型：(1)嵌入模型$p(\omega_t|h_t)$，将历史轨迹h_t映射到时刻-t对手变量ω_t；(2)观察模型$p(o_t, a_t, r_t \mid \omega_t, h_t)$，从历史和对手中恢复观察值$(o_t, a_t, r_t)$；(3)潜在动态模型$p(\omega_t \mid \omega_{t-1}, a_{t-1})$，建模嵌入空间中的转换。这些概率模型假设为均值和方差实现为神经网络函数逼近器的高斯分布。例如，嵌入模型采用形式$p(\omega_t|h_t) = N(\omega_t|f_\theta^\mu(h_t), f_\theta^\sigma(h_t))$。

嵌入模型$p(\omega_t|h_t)$提供了对对手因素的潜在表示，并通过重构环境的观察证据来学习。这导致了以下基于 VIB 的下界：

$$\max \mathop{\mathbb{E}}_{d^{\pi}, d^{\xi}}[\log p(o_t, r_t \mid \omega_t)] + \mathop{\mathbb{E}}_{d^{\xi}}[\log p(a_t \mid \omega_t)] - \beta \mathcal{D}_{KL}\big(p(\omega_t \mid h_t) \,\|\, p(\omega_t \mid \omega_{t-1}, a_{t-1})\big), \quad (7)$$

其中前两项是观察的证据，最后一项作为信息压缩的 KL 正则项，其中超参数β控制其强度。

由于对手因素ω解释了环境的变化，我们在公式(6)中用ω替换环境 M，而在学到的潜在空间中，我们寻找表示对齐的训练环境的分布$p(\omega)$。同时，我们还通过关于ω的可微分遗憾函数实现了可行的基于梯度的搜索。为了展示这一点，我们首先写出遗憾函数，如下所示：

$$Reg(\pi,\omega)=V(\xi^*;M_\omega)-V(\pi;\ M_\omega)=$$
$$\mathbb{E}_{d_\omega^{\xi^*}}[\sum_{t=1}^{H}\mathbb{E}[r_t|s_t,a_t;\omega]]-$$
$$\mathbb{E}_{d_\omega^{\pi}}[\sum_{t=1}^{H}\mathbb{E}[r_t|s_t,a_t;\omega]],\quad (8)$$

其中$d_\omega^{\pi}(s_t,a_t)$（以及$d_\omega^{\xi^*}(s_t,a_t)$）表示 MDP M$\boldsymbol{\omega}$ 中的策略π（和ξ^*）的状态-动作访问。

通过世界模型的重建，奖励估计器 $\mathrm{E}[r_t|s_t,a_t;\boldsymbol{\omega}]$ 本质上是作为观察模型 $p(o_t,a_t,r_t|\omega_t,h_t)$ 的一个组件而学习的。具体而言，我们学习一个神经网络函数逼近器 $r_\theta(o_t,a_t,h_t,\boldsymbol{\omega})$，作为奖励估计器的替代，满足以下的最小二乘目标：

$$\min_\theta \mathbb{E}_D\left[\left(r^H-\sum_{t=1}^{H}r_\theta(o_t,a_t,h_t,\omega_t)\right)^2\right] \quad (9)$$

我们用 D 表示包含多个环境的训练日志，r^H 表示每个回合的奖励。

3.2.2 优化可微分博弈

最终，我们得到了一个可通过同时梯度下降（即对偶上升）进行优化的可微分博弈，如下所示：

$$P^{(t)}(\omega)=\arg\max_{P\in\mathcal{P}}\ \mathop{\mathbb{E}}_{\omega\sim P}\left[\mathrm{Reg}\left(\pi^{(t-1)},\omega\right)\right]+\alpha\mathcal{H}(P), \quad (10)$$

$$\pi^{(t)}=\arg\min_{\pi}\ \mathop{\mathbb{E}}_{\omega\sim P^{(t)}}\left[\mathrm{Reg}\left(\pi^{(t-1)},\omega\right)\right]. \quad (11)$$

直观地说，这个博弈在两个玩家之间交替进行优化 - 一位老师通过找到在学到的潜在空间中的最坏情况环境的分布 $P(\boldsymbol{\omega})$ 来辅导以前的策略，以及一位学习者在给定的环境分布 $P(\boldsymbol{\omega})$ 上元学习其策略 π。

最坏情况的辅导步骤。为了确保训练分布应符合经验数据集，即 $P\in$ P，我们选择基于 Wasserstein 距离定义集合 P，该距离在适当的假设下具有方便的形式[38]。

Wasserstein 距离计算将一个分布转换为另一个分布的最小成本，以捕捉潜在空间的几何特性[1]而闻名。具体而言，我们使用 L2-范数成本函数定义了 Wasserstein 距离 $W\kappa(\cdot,\cdot)$，其中。直观地说，我们的目标是在 Wasserstein 距离下将集合 P 定义为潜在空间中经验分布的 ρ-邻域。为了实现这一点，我们首先将在潜在空间中收集到的环境的经验分布表示为$\bar{P}(\omega)=\frac{1}{M}\sum_{i=1}^{M}\delta(\omega_i)$，其中$\boldsymbol{\omega}_i$表示第$i$个环境。然后我们定义集合为$P=\{P:W_k(P,\bar{P})\le\rho\}$

根据对方程（10）的双重重新表达（详见附录 A.1），我们得到以下目标：

$$\max_{P\in\mathcal{P}}\mathbb{E}_\omega[Reg(\pi,\omega)]=\mathbb{E}_{\widetilde{\omega}}[\max_{\omega}Reg\ (\pi,\omega)-\lambda\parallel\omega-\widetilde{\omega}\parallel_2], \quad (12)$$

其中$\widetilde{\omega}$表征一个收集到的环境$M_{\widetilde{\omega}}$。在实现中，我们采样一组以$\{\widetilde{\omega}_i\}_{i=1}^{n}$表征的日志环境，并通过步长 η 进行基于梯度的更新，

$$\omega'\leftarrow\widetilde{\omega}+\eta\nabla_\omega[\mathrm{Reg}(\pi,\omega)-\lambda\parallel\omega-\widetilde{\omega}\parallel_2], \quad (13)$$

通过从采样的环境中获取一个训练环境的分布，表示为$\{\omega_i'\}_{i=1}^{n}$

所以，通过从最坏情况环境的分布 $P(\omega)=\frac{1}{n}\sum_{i=1}^{n}\delta(\omega_i')$中获取的策略改进步骤，根据公式 (11)，成为标准的值最大化目标，因为专家值对π是常数。

$$\min_\pi \mathrm{Reg}(\pi;\boldsymbol{\omega})$$
$$=\min_\pi \mathop{\mathbb{E}}_{(\widetilde{\omega},\omega)}\left[\mathop{\mathbb{E}}_{d_{\widetilde{\omega},\pi}}\left[\sum_{t=1}^{H}r_\theta(o_t,a_t,h_t,\boldsymbol{\omega})\right]\right]. \quad (14)$$

因此，在实施中，每个敌对环境ω与采样环境$\widetilde{\omega}$关联，因为在公式(13)中有成对的基于梯度的搜索。遵循深度强化学习[25]，我们将策略分布实现为高斯分布，其均值和方差由神经网络函数逼近器参数化，即

$$\pi(\cdot|\ o_t,h_t)=\mathcal{N}\left(\cdot|\ f^\mu(o_t,h_t),f^\sigma(o_t,h_t)\right).$$

3.3 因果感知专家学习

虽然 MiRO 的设计目标是最小化策略遗憾（见方程 6），但方程（14）实际上退化为一个无需专家参与的值最大化问题。有趣的是，先前关于有约束出价的研

究也忽略了专家在他们学习目标中的作

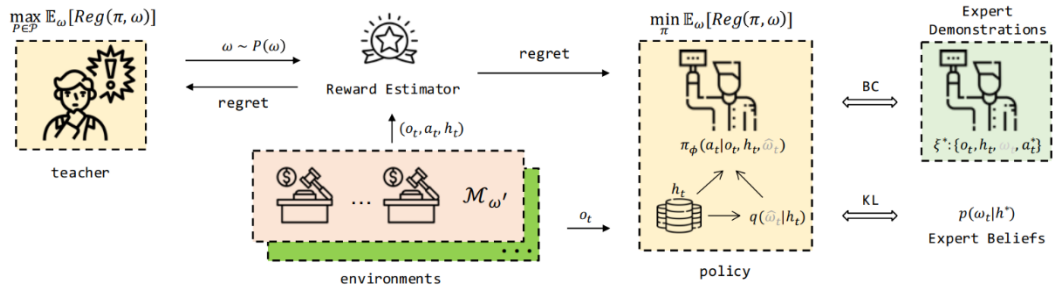

图 3 MiROCL 概述。我们的方法解决一个在教师和学习者之间交替的可微分博弈。教师找到最坏情况环境的分布 $P(\boldsymbol{\omega})\in P$，并且学习者在给定环境分布上元学习其策略$\pi$。为了在因果结构方面与专家保持一致，策略$\pi$被设计为$\pi_\phi$，其取决于通过推理模型$q(\hat{\omega}|h_t)$获得的$\hat{\omega}$。除了来自值最大化的监督之外，子策略$\pi_\phi$和推理模型还从专家演示和专家的后验信念中获得指导。

用。然而，我们的目标是通过学习专家演示来改进方程（14），因为我们认为专家可能包含在不同环境中如何最优行动的宝贵知识。然而，令人惊讶的是发现直接的行为克隆方法，即模仿专家演示，只会导致性能下降（见第 4.2.2 节）。为了理解这个问题，我们将策略学习视为因果推断问题。我们确定了未观察到的混杂问题，使得无法从观测数据中唯一计算策略。在接下来的部分中，我们首先阐述这一现象，然后提出了一种基于因果关系的对齐策略来解决这个问题。

为了说明在策略学习中未观察到的混淆问题，考虑一个拍卖环境，其中销售机制被修改，使得获胜成本大于第二价格。在这种情况下，由于成本增加，专家策略 $\xi*$ 的投标（比例）较第二价格拍卖要低。这导致a_t与 r_t之间存在（虚假的）相关性（为了清晰起见，省略了$(o_t, h_t) = (o<t, a<t)$ 的条件），这表明较小的a_t值与较高的奖励r_t相关联。不幸的是，通过行为克隆学到的策略会捕捉到这样的虚假相关性，因为它只学习统计依赖关系。因此，这样的策略将无法推广到具有不同销售机制的环境。

通过因果镜头，决策问题自然地可以被制定为因果查询，我们的目标是在行动干预下推断结果。基于这一点，从专家演示中学习可以被转化为估计对未来奖励的干预$do(a_t)$的因果效应，即$p\left(\sum_{i=t}^{H} r_i \mid o_t, h_t, do(a_t)\right)$，使用专家收集的观察数据$\{o_i, h_i, a_i, r_i\}\xi$。

正如因果图中所示的那样，混淆变量$\boldsymbol{\omega}$为观测数据中的$a_t \leftarrow \omega_t \rightarrow r_t$的因果结构做出了贡献，但对于策略$\pi(a_t|o_t, h_t)$来说是未观察到的。因此，当$\omega_t$未观察到时，条件独立性$(a_t, r_t) \perp\!\!\!\perp \omega_t$被打破，意味着观察数据呈现$a_t$与$r_t$之间的因果关系和虚假相关性。因此，策略无法从数据中唯一地恢复所需的因果查询，这被称为不可识别性问题。

为了缓解这个问题，我们的想法是对齐专家和策略的因果结构。这通过使用额外的输入$\hat{\omega}_t$对策略进行条件设计来实现，该输入被设计为对于在线服务的策略π不可用的真实ω_t的替代。在这方面，策略可以模仿专家，具有相同的因果结构，从而消除了策略学习中的虚假相关性。因此，我们采用以下策略设计：

$$\pi(a_t \mid o_t, h_t) = \int_{\hat{\omega}_t} \pi_\phi(\cdot \mid o_t, h_t, \hat{\omega}_t) \cdot p(\hat{\omega}_t \mid h_t) d\hat{\omega}_t, \quad (15)$$

这分解成一个子策略$\pi_\phi \in \Xi$ 和一个用于推断$\hat{\omega}_t$的推断模型。我们注意到，这个推断模型正是我们根据方程(7)学到的，我们的目标是进一步利用专家轨迹的指导。因此，我们从最小化策略差异中导出以下约束：

$$\min_{\pi} \mathcal{D}_{KL}(\xi^* \| \pi) \leq$$
$$\min_{\pi_\phi, q} \mathbb{E}\left[-\log \pi_\phi(a \mid o_t, h_t, \widehat{\omega}_t)\right]$$
$$+\beta_2 \mathcal{D}_{KL}\left(p(\omega_t \mid h_t^*) \| p(\widehat{\omega}_t \mid h_t)\right). (16)$$

第一项训练了一个子策略，该子策略输入推断的$\widehat{\omega}_t$，通过行为克隆来模拟专家演示。第二项通过利用专家轨迹的后验信念，为推断模型添加了额外的 KL 正则性。详细的推导包含在 A.2 节中，表明上述目标确实最小化了遗憾的上界。

4 实验

在这项工作中，我们提出了一种用于对抗性有约束出价的 Minimax Regret Optimization (MiRO)框架，并提供了一个端到端优化的实用算法。此外，我们是首批提倡使用专家演示进行策略学习的研究者，从而将 MiRO 提升为 MiROCL（带因果感知强化学习的 MiRO）。因此，我们的实验旨在探讨以下问题：

问题 1（与先前工作的比较）：在黑盒对抗环境中，所提出的方法与先前方法相比在实证上表现如何？

问题 2（割舍实验）：所提出的每个组件的有效性如何？

对于这些问题，我们使用一个名为“Industrial”的工业数据集，该数据集展示了真实世界的对抗情况。由于对抗因素纠结在现实世界的数据中，我们还创建了一个名为“Synthetic”的合成数据集，其中涉及已知结构的不同销售机制。关于数据集和实现的详细信息可在附录中找到，数据集和代码可以在 https://github.com/HaozheJasper/MiROCL 上公开获取。

4.1 实验设置

4.1.1 数据集

在实验中，我们使用了两个数据集。工业数据集来自阿里巴巴展示广告平台，包括 80 天的竞价日志，每天平均有 200 万个请求。每个请求xi包括：市场价格mi，效用估计$\mathbb{E}[ui|xi]$ 和真实的随机反馈ui。

由于我们假设效用估计 $\mathbb{E}[ui|xi]$ 已经预先计算，因此在这项工作中我们未使用上下文特征。由于在 RTB 系统中，未赢得拍卖的情况下市场价格 mi 是未知的，我们采用了一种特殊的策略，即以尽可能高的价格进行出价以获取市场价格。因此，我们认为工业数据集中的每个竞价日志代表一个独特的竞价环境。对于问题 Q1，我们将工业数据集分为前 60 天和后 20 天，基于我们观察到这两个集合在市场动态和/或价值分布方面存在差异（图 5）。从中分布集合中随机选择 30 天形成训练集，从另一集合中选择 30 天形成独立同分布（IID）测试集，最后的 20 天作为离分布测试集。

合成数据集基于公共的合成数据集 AuctionGym[30]。与规模相对较小的 AuctionGym 相比，我们的合成数据集包括 80 天的竞价日志，每天 100 万次曝光，旨在用于对具有专家策略的受限竞价进行研究。为了模拟类似于工业数据的对抗环境中的受限竞价，我们假设现实世界中的黑盒拍卖可以近似为线性混合的第二价格和第一价格拍卖格式，即成本$c_i = k \cdot b_i + (1-k) \cdot m_i$是出价$b_i$和市场价格$m_i$的线性组合，其中可能存在动态比率$k$。这个假设是基于数据的见解，即某些媒体渠道上的费用c_i取决于出价b_i，导致观察到相同流量分布的获胜概率随着出价的变化而变化（在二价拍卖中，相同流量分布的费用随着出价的变化而不变，获胜概率也是如此）。因此，对于合成实验，我们模拟一个动态混合的第二价格和第一价格拍卖环境，以检验算法在对抗环境中的效果。在这种情况下，训练集包括 10 天的 GSP 竞价日志和 20 天的动态混合竞价，其中混合比率k随机采样在(0, 1)范围内，而测试集包括 20 天的 GSP 和 30 天的随机采样混合竞价日志。

4.1.2 评估协议。我们在实验中使用竞争比率（CR）来评估方法。CR 是策略价值与专家价值的比率，直接反映了在线遗憾。此外，我们在评估中引入了一种容忍度的概念，这是由实际实践中的情况所驱动的，即如果策略获得了可观的回报，我们可以容忍策略违反约束。具体而言，我们定义了在一个包含N天数据集上的平均容忍感知比率（TACR），其中预定义了最大容忍度γ和基

准回报率ζ，如下所示，

$$\text{TACR} = \frac{1}{N}\sum_{i=1}^{N}\frac{U(i)/(1+\zeta)^{\lambda(i)}\cdot \mathbb{1}_{\text{ROI}(i)\geq L(1-\gamma)}}{U^*(i)}, (17)$$

这里我们使用缩写 $U(i)=U^T(\pi,x_i)$ 表示策略π对于第i天请求序列x^i的累积效用。同样，$C(i)$表示总成本，$U^*(i)$表示基准值。

我们计算 $\lambda(i)=max\{ceil\left(max\left\{1-\frac{\text{ROI}(i)}{L},0\right\}\right),\gamma\}$，作为策略$\pi$对于第$i$天的解决方案的容忍水平。直观地说，该数量度量了允许在最大公差 γ 范围内违反约束的竞争比率，如果违反了约束，则将其值以基线收益率ζ折扣。特别地，在我们的实验中，我们设置 $\gamma=2\%$，$\zeta=5\%$，这表示我们认为每下降 1%的 ROI 可以用 5%的效用增加来交换，且容许的最大违规容忍度为 2%。此外，我们还显示了在最大公差水平 2%下的竞争比率$(CR@2\%)$，它将违反约束低于 2%的解决方案视为可行。我们将该指标表示为$(CR@2\%)$。

4.2 实验结果

4.2.1 与先前方法的比较

本研究旨在展示在对抗性环境中的受限投标问题的挑战，为了比较，我们选择了可以处理或适应于受限投标问题的代表性方法（1）具有预算约束和 ROI 约束：（1）PID 控制方法[47]和交叉熵方法 CEM ［29］被视为在线学习方法；（2）USCB（2021 年）和 CBRL（2022 年）是两种最近提出的基于 RL 的方法，使用收集到的投标日志进行训练。其中，CBRL 建立在 ERM 原则的基础上，不使用专家演示进行学习。

工业数据集和合成数据集的评估结果分别显示在表 1 和表 2 中。我们实验证明，所提出的 MiROCL 在所有性能指标上都在两个数据集上表现最好。为了澄清，我们注意，虽然单次独立运行的 TACR 结果应该在其 IID-TACR 和 OOD-TACR 之间，但 mTACR 不一定在 IID-mTACR 和 OOD-mTACR 分数之间，因为它们报告了 20 次运行的中位数。此外，mTACR 对于一些模型（例如 MiROCL）来说接近于 mCR@2%，这意味着在违反约束时这些模型可能享有较高的回报（因此它们减少的幅度较小，并且接近于 mCR@2%）。此外，在表 2 中，一些模型（遵循批量学习范 paradigm）在 GSP 和合成的对抗性销售机制上显示出类似的性能，因为这些模型是在不同拍卖格式的联合集上进行训练的，从而在不同拍卖格式上产生平均效果。以下是竞争方法的结果：

PID[47]基于实时反馈控制运行。表 1 显示，在 OOD 环境中，PID 保持相对稳定的性能。然而，我们观察到在动态混合二价一价拍卖中，PID 的表现不如在 GSP 中好（在表 2 中，GSP-mTACR 为 0.3817，而 MIX-mTACR 为 0.2837），可能是因为 PID 不太能够提前预测，以调整其被转换为由发布商收费的某些出价。

CEM[29]是一种零阶随机优化方法，理论上不受分布变化的影响。然而，我们观察到 CEM 的超参数对剧烈的分布变化敏感（表 1 中 OOD 分数为 0）。例如，在 OOD 环境中，最优的出价比率可能超出 CEM 的精英分布的范围（图 5）。该算法在合成数据集中表现更为稳定（表 2），因为训练和测试条件是独立同分布的。

USCB[27]采用基于蒙特卡洛值估计的 actor-critic 强化学习方法，使用软奖励函数。表 1 显示，在 OOD 环境中，USCB 无法进行泛化，很可能是由于未观察到的混淆问题。在 i.i.d.测试条件下（表 2），USCB 的性能甚至比 CEM 还差。这是因为 USCB 的整体性能在很大程度上由控制效用约束权衡的超参数决定。由于不同的销售机制需要不同的权衡配置，单一的静态超参数将导致整个数据集的次优性能。

CBRL[41]提出了 POCMDP 公式，为我们的工作奠定了基础。CBRL 和 MiROCL 之间存在两个主要区别。首先，CBRL 采用 ERM 原则，遵循等式(2)的目标，这在对抗性环境中不提供泛化保证。其次，CBRL 不考虑在策略学习中包括专家演示。

CBRL 在具有挑战性的 OOD 环境中表现优于 USCB，从 OOD-mTACR 的角度来

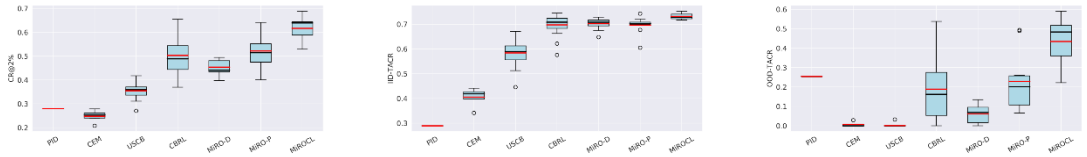

图 2 上述显示的是工业数据集中 CR@2%（左）、IID-TACR（中）OOD-TACR（右）的结果。每个箱形图显示 20 次独立重复运行的平均值（红色）和中位数（黑色）的结果。

看，因为 CBRL 采用了旨在适应环境的贝叶斯机制。然而，CBRL 仍然表现不如我们的方法，因为 MiRO 潜在地对齐了训练和测试分布，因此能够更好地进行泛化。在 i.i.d.训练和测试条件下（见表 2），CBRL 优于 USCB，因为它隐含地学会了为不同环境推断效用约束权衡。然而，所提出的方法仍然优于 CBRL，主要是因为 MiROCL 从专家那里提取了知识。

表 1 在具有现实世界对抗条件的 Industrial 数据集上的中位测试分数。报告的 mTACR 和 mCR 分数是对所有模型进行 20 次随机试验的中位数。TACR 是在最大容忍水平 2%和回报率 5%的情况下计算的，*CR*@2%显示了 2%容忍度水平上的未打折 CR。 IID-mTACR 和 OOD-mTACR 报告了两个集合的详细结果，即分布内集合和分布外集合。

	mTACR	mCR@2%	IID-mTACR	OOD-mTACR
MiROCL	0.6359	0.6391	0.7285	0.4833
MiRO-P	0.5015	0.5146	0.702	0.2013
MiRO-D	0.4337	0.4406	0.7079	0.0684
CBRL	0.4793	0.489	0.708	0.1622
USCB	0.3537	0.359	0.5895	0.0
CEM	0.2515	0.2529	0.4192	0.0
PID	0.2753	0.2795	0.2894	0.2542

4.2.2 消融研究

首先回顾我们方法中提出的设计和组件。

ERM 对比 MiRO。在第 3.1 节中，我们指出了 ERM 的主要缺陷，即在对抗性设置中违反了独立同分布的假设。为了解决这个问题，我们在第 3.2 节提出了一个实用的 MiRO 算法，实现了一个详细的训练-测试对齐。我们选择了目前最先进的方法 CBRL，它遵循 ERM 原则，作为 ERM 方法的代表，并将其与实用的 MiRO 算法进行比较，表示为 MiRO-P。

MiRO 对比行为复制的 MiRO。在第 3.3 节中，我们旨在通过包含来自专家演示的显式学习来改进 MiRO，因为我们相信专家包含有关在不同环境中做出最佳决策的宝贵知识。我们首先检查行为克隆的简单想法，该方法要求策略模仿专家演示，遵循最大似然估计（MLE）原则。这种方法被标记为 MiRO-D，显示出令人惊讶的性能下降，这激发了提出的因果关系感知的对齐策略。

MiRO 对比 MiROCL。基于因果分析，我们确定了未观察到的混杂问题，该问题使得直观的 MLE 原则失败。为了解决这个问题，我们提出了一种考虑因果关系的方法，将策略解开为一个模仿专家因果结构的子策略和推理模型。这个整体方法被标记为 MiROCL。

表 2 合成数据的中位数分数。所有模型的 mTACR 和 mCR 得分取 20 次随机试验的中位数。我们还报告了两种拍卖格式的结果。

	mTACR	mCR@2%	GSP-mTACR	MIX-mTACR
MiROCL	0.8094	0.8114	0.7932	0.8098
MiRO-P	0.7231	0.7469	0.7307	0.7116
CBRL	0.6856	0.7003	0.696	0.6793
USCB	0.5171	0.5304	0.5114	0.5256
CEM	0.5811	0.6094	0.5972	0.5438
PID	0.3343	0.3556	0.3728	0.2766

综合实验结果，我们得出以下结论。

ERM 对比 MiRO. 如表 1 所示，MiRO-P 在整体上的性能优于 CBRL。虽然 CBRL 和 MiRO-P 在 IID 性能上相当，但根据图 4 中的箱线图，MiRO-P 在实际的 OOD 情况下表现出更好的平均性能和稳定性。这是因为 MiRO 框架允许策略在更可能与测试分布一致的分布下进行训练，从而导致更好的平均

性能。然而，我们也注意到 MiRO 在稳定性方面仍存在限制（箱线图中方差较大），这是由于测试条件的不可预测性和训练过程中的随机性，这也是未来研究的一个有趣方向。

MiRO 对比行为复制的 MiRO。 与与 ERM 的比较类似，MiRO-P 的整体性能优于 MiRO-D，主要得益于更好的 OOD 性能。在实际的 OOD 情况中，根据图 4，MiRO-D 在平均性能较差，但稳定性较高。这是因为行为克隆涉及到未观察到的混淆问题，学习到的虚假统计依赖可能在分布转移下不会泛化。我们推测 MiRO-D 表现出更好的稳定性是由于行为克隆的记忆效应，这意味着决策对不同环境的适应性较差。

MiRO 对比 MiROCL. 如表 1 所示，MiROCL 在 IID 和 OOD 性能方面均显著优于 MiRO-P，特别是在 OOD-mTACR 方面增加了 21%。IID 性能的提高主要是由于对专家演示的有效提炼。OOD 性能的提高是因为 MiROCL 遵循专家政策的因果结构，并利用推断模型在不同环境中自适应地推断特权信息。值得注意的是，表 2 还显示 MiROCL 在对抗性拍卖机制方面比 MiRO-P 提高了很大的幅度（12%），表明专家演示有效地引导了政策朝向最优方向。

5 相关工作

受限出价。我们讨论了关于二价拍卖的相关工作。大多数关于受限出价的研究都遵循独立同分布的假设或最小风险最大期望（ERM）原则。其中，大多数研究侧重于具有预算约束的出价（参见[5]进行的调查），而一些研究[27, 41, 47]进一步提出处理更具挑战性的与成本相关的约束，即类似 ROI 的约束。我们的工作研究 ROI 受限出价（RCB）问题，基于 CBRL[41]中的 POCMDP 公式化。

与重复拍卖中的对抗学习的联系。最近一些关于对抗性学习投标问题的研究 [26, 33]。 [26] 讨论了对抗性一价拍卖的在线学习方法，处理没有约束的情况，这与我们的工作无关。 [33] 研究了具有对抗性卖方的场景，并假设卖方采用数据驱动的机制 [16]。然而，在现实中，有多个对方因素，而没有一个对于代理人是可观察的。为了解决这个局限性，我们探索了先验自由的对抗性设置。

与极小极大博弈公式化的联系。极小极大博弈的公式化在生成对抗网络[1, 22]、在线学习中的遗憾分析[9, 26, 43]、监督学习中的分布鲁棒优化[28, 39]和对抗训练[38]中也有体现。生成对抗网络旨在实现逼真的生成，将最小化分布差异与极小化最大化目标相联系。分布鲁棒优化和对抗性训练与在分布偏移和对抗性攻击下的强健泛化有关。遗憾分析旨在在最坏在线条件下证明性能差距的上限。特别是，先前针对出价的在线学习方法[9, 11, 19, 43]通常局限于小规模问题，并且需要有关市场的知识。我们的方法可以看作是将在线学习的最小最大优化原则和离线学习的鲁棒性考虑因素结合起来的产物，但其动机是基于训练-测试分布对齐的见解。

6 结论

在这项工作中，我们探讨了在对抗环境中受限出价的一个未知问题，而且对于对抗因素如何扰动环境并没有先验知识。先前的受限出价方法通常依赖于在对抗性设置中被违反的经验风险最小化（ERM）原则。为了解决这个局限性，我们提出了一种最小化遗憾优化（MiRO）框架，该框架在教师识别对齐训练分布和学习者优化在给定环境分布下的策略之间交替进行。为了使最小最大问题变得可行，我们通过变分学习对抗因素的表示，通过重建世界模型的因果结构，使最小最大问题可微分化，并通过双重梯度下降优化可微分化博弈。此外，我们首次将专家演示纳入到策略学习中。我们发现了未观察到的混淆问题，使得从专家那里直接克隆行为的直观想法不可行，因此我们开发了一种考虑因果关系的方法，旨在模仿专家策略的因果结构并从专家演示中提取知识。在大规模工业和合成数据集上的实证结果表明，我们的方法 MiROCL 在性能上超过了先前的方法超过 30%。

参 考 文 献

[1] Jonas Adler and Sebastian Lunz. 2018. Banach wasserstein gan. Advances in neural information processing systems 31 (2018).

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. arXiv preprint arXiv:1612.00410 (2016).

[3] Alimama 2022. Alimama. Retrieved 2022 from https://www.alimama.com/

[4] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. 2018. The mechanics of n-player differentiable games. In International Conference on Machine Learning. PMLR, 354–363.

[5] S. Balseiro, A. Kim, M. Mahdian, and V. Mirrokni. 2021. Budget-Management Strategies in Repeated Auctions. Operations Research 69, 3 (2021).

[6] Santiago R Balseiro, Omar Besbes, and Gabriel Y Weintraub. 2015. Repeated auctions with budgets in ad exchanges: Approximations and design. Management Science 61, 4 (2015), 864–884.

[7] Santiago R Balseiro and Yonatan Gur. 2019. Learning in repeated auctions with budgets: Regret minimization and equilibrium. Management Science 65, 9 (2019), 3952–3968.

[8] Christopher M Bishop and Nasser M Nasrabadi. 2006. Pattern recognition and machine learning. Vol. 4. Springer.

[9] Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu. 2004. Online learning in online auctions. Theoretical Computer Science 324, 2-3 (2004), 137–146.

[10] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. 2004. Convex optimization. Cambridge university press.

[11] Sébastien Bubeck, Nikhil R Devanur, Zhiyi Huang, and Rad Niazadeh. 2017. Multi-scale online learning and its applications to online auctions. arXiv preprint arXiv:1705.09700 (2017).

[12] Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. 2017. Real-time bidding by reinforcement learning in display advertising. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 661–670.

[13] Dragos Florin Ciocan and Vivek Farias. 2012. Model predictive control for dynamic resource allocation. Mathematics of Operations Research 37, 3 (2012), 501–525.

[14] Alexey Drutsa. 2020. Reserve pricing in repeated second-price auctions with strategic bidders. In International Conference on Machine Learning. PMLR, 2678– 2689.

[15] Chao Du, Zhifeng Gao, Shuo Yuan, Lining Gao, Ziyan Li, Yifan Zeng, Xiaoqiang Zhu, Jian Xu, Kun Gai, and Kuang-Chih Lee. 2021. Exploration in Online Advertising Systems with Deep Uncertainty-Aware Learning. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2792–2801.

[16] Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. 2019. Optimal auctions through deep learning. In International Conference on Machine Learning. PMLR, 1706–1715.

[17] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. American economic review 97, 1 (2007), 242–259.

[18] Zhe Feng, Sébastien Lahaie, Jon Schneider, and Jinchao Ye. 2021. Reserve price optimization for first price auctions in display advertising. In International Conference on Machine Learning. PMLR, 3230–3239.

[19] Zhe Feng, Chara Podimata, and Vasilis Syrgkanis. 2018. Learning to bid without knowing your value. In Proceedings of the 2018 ACM Conference on Economics
and Computation. 505–522.

[20] Joaquin Fernandez-Tapia. 2019. An analytical solution to the budget-pacing problem in programmatic advertising. Journal of Information and Optimization
Sciences 40, 3 (2019), 603–614.

[21] Joaquin Fernandez-Tapia, Olivier Guéant, and Jean-Michel Lasry. 2017. Optimal real-time bidding strategies. Applied mathematics research express 2017, 1 (2017), 142–183.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. Commun. ACM 63, 11 (2020), 139–144.

[23] Google 2022. Google. Retrieved 2022 from https://ads.google.com/ [24] Ramki Gummadi, Peter Key, and Alexandre Proutiere. 2013. Optimal bidding strategies and equilibria in dynamic auctions with budget constraints. Available at SSRN 2066175 (2013).

[25] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. 2018. Soft actor-critic algorithms and applications. arXiv preprint arXiv:1812.05905(2018).

[26] Yanjun Han, Zhengyuan Zhou, Aaron Flores, Erik Ordentlich, and Tsachy Weissman. 2020. Learning to bid optimally and

efficiently in adversarial first-price auctions. arXiv preprint arXiv:2007.04568 (2020).

[27] Yue He, Xiujun Chen, Di Wu, Junwei Pan, Qing Tan, Chuan Yu, Jian Xu, and Xiaoqiang Zhu. 2021. A Unified Solution to Constrained Bidding in Online Display Advertising. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2993–3001.

[28] Zhaolin Hu and L Jeff Hong. 2013. Kullback-Leibler divergence constrained distributionally robust optimization. Available at Optimization Online (2013), 1695–1724.

[29] Antoine Jamin and Anne Humeau-Heurtier. 2019. (Multiscale) Cross-Entropy Methods: A Review. Entropy 22 (12 2019). https://doi.org/10.3390/e22010045

[30] Olivier Jeunen, Sean Murphy, and Ben Allison. 2022. Learning to bid with AuctionGym. (2022).

[31] Sascha Lange, Thomas Gabel, and Martin Riedmiller. 2012. Batch reinforcement learning. Reinforcement learning: State-of-the-art (2012), 45–73.

[32] Roger B Myerson. 1981. Optimal auction design. Mathematics of operations research 6, 1 (1981), 58–73.

[33] Thomas Nedelec, Jules Baudet, Vianney Perchet, and Noureddine El Karoui. 2021. Adversarial Learning in Revenue-Maximizing Auctions. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. 955–963.

[34] Thomas Nedelec, Clément Calauzènes, Noureddine El Karoui, Vianney Perchet, et al. 2022. Learning in repeated auctions. Foundations and Trends® in Machine Learning 15, 3 (2022), 176–334.

[35] Michael Ostrovsky and Michael Schwarz. 2011. Reserve prices in internet advertising auctions: A field experiment. In Proceedings of the 12th ACM conference on Electronic commerce. 59–60.

[36] Judea Pearl et al. 2000. Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress 19, 2 (2000).

[37] Jad Rahme, Samy Jelassi, and S Matthew Weinberg. 2020. Auction learning as a two-player game. arXiv preprint arXiv:2006.05684 (2020).

[38] Aman Sinha, Hongseok Namkoong, and John Duchi. 2017. Certifiable distributional robustness with principled adversarial training. arXiv preprint arXiv:1710.10571 2 (2017).

[39] Matthew Staib and Stefanie Jegelka. 2019. Distributionally robust optimization and generalization in kernel methods. Advances in Neural Information Processing Systems 32 (2019).

[40] Alberto Vera, Siddhartha Banerjee, and Itai Gurvich. 2021. Online allocation and pricing: Constant regret via bellman inequalities. Operations Research 69, 3 (2021), 821–840.

[41] Haozhe Wang, Chao Du, Panyan Fang, Shuo Yuan, Xuming He, Liang Wang, and Bo Zheng. 2022. ROI-Constrained Bidding via Curriculum-Guided Bayesian Reinforcement Learning. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 4021–4031.

[42] Haozhe Wang, Jiale Zhou, and Xuming He. 2020. Learning context-aware task reasoning for efficient meta-reinforcement learning. arXiv preprint arXiv:2003.01373 (2020).

[43] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. 2016. Online learning in repeated auctions. In Conference on Learning Theory. PMLR, 1562–1583.

[44] Christopher A Wilkens, Ruggiero Cavallo, Rad Niazadeh, and Samuel Taggart. 2016. Mechanism design for value maximizers. arXiv preprint arXiv:1607.04362 (2016).

[45] Annie Xie, James Harrison, and Chelsea Finn. 2020. Deep reinforcement learning amidst lifelong non-stationarity. arXiv preprint arXiv:2006.10701 (2020).

[46] Tian Xu, Ziniu Li, and Yang Yu. 2020. Error bounds of imitating policies and environments. Advances in Neural Information Processing Systems 33 (2020), 15737–15749.

[47] Xun Yang, Yasong Li, Hao Wang, Di Wu, Qing Tan, Jian Xu, and Kun Gai. 2019. Bid optimization by multivariable control in display advertising. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 1966–1974.

[48] Luisa Zintgraf, Sebastian Schulze, Cong Lu, Leo Feng, Maximilian Igl, Kyriacos Shiarlis, Yarin Gal, Katja Hofmann, and Shimon Whiteson. 2021. VariBAD: variational Bayes-adaptive deep RL via meta-learning. The Journal of Machine Learning Research 22, 1 (2021), 13198–13236.

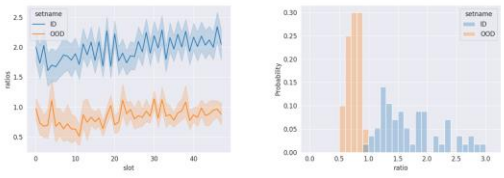

图 5 工业数据集的分布偏移。（左）IID 和 OOD 集上的平均按时槽分的专家策略。（右）按天的专家策略的投标比率分布。

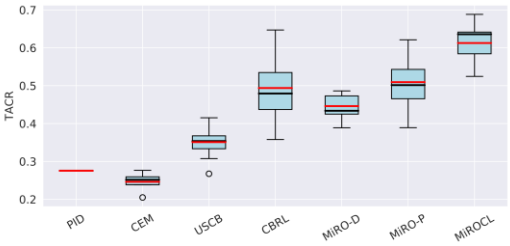

图 6 工业数据集上的 TACR 结果。每个箱线图显示了 20 次随机试验的中位数（红色条）和均值（黑色条）分数。

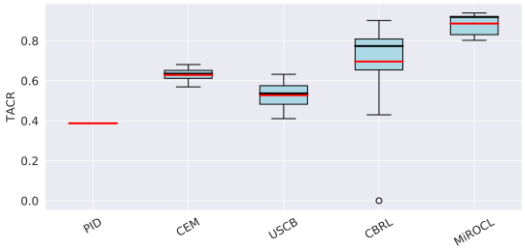

图 7 合成数据集上的 TACR 结果。每个箱线图显示了 20 次随机试验的中位数（红色条）和均值（黑色条）分数。

附录 A

首先，我们的目标是找到一个环境分布，该分布在嵌入空间中的 Wasserstein 球内，最大化当前策略的遗憾，而该球以经验分布为中心。对于支持在潜在空间 W 上的概率测度 $P1$ 和 $P2$，以及联合分布 $\Pi(P1,P2)$，在度量空间 W 上定义的 Wasserstein 距离为：

$$W_\kappa(P_1,P_2) \overset{\text{def}}{=} \sup_{H\in\Pi(P_1,P_2)} \mathbb{E}_H[\kappa(\omega_1,\omega_2)], \qquad (18)$$

该距离计算从分布 $P1$ 到 $P2$ 的变换的最小成本，其中成本函数为 $\kappa(\cdot,\cdot)$。

这个距离使用成本函数捕捉了空间 W 的几何特征，我们假设 $\kappa(x,y) = \| x - y \| 2$，即两个分布之间的距离与它们的样本之间的欧几里得距离相关。我们将一组分布定义为围绕经验分布 $\bar{P}$ 的 ρ-球，从中寻找对抗性环境：

$$\mathcal{P} = \{P: W_\kappa(P,\bar{P}) \le \rho\}, \qquad (19)$$

老师的目标可归纳如下：

$$\begin{aligned} &\sup_{P(\omega)} \mathbb{E}_\omega[\mathrm{Reg}\,(\pi,\omega)] \\ &= \sup_{P(\omega)} \{\mathbb{E}_\omega[\mathrm{Reg}(\pi,\omega)] - \lambda W_\kappa(P,\bar{P})\}' \end{aligned} \qquad (20)$$

基于 $\mathrm{Reg}(\pi,\boldsymbol{\omega})$和$\kappa(\cdot,\cdot)$连续的假设，我们可以得到以下对偶重构[38]：

$$\begin{aligned} &\sup_{P:W_\kappa(P,\bar{P})\le\rho} \mathbb{E}_\omega[\mathrm{Reg}\,(\pi,\omega)] \\ &= \inf_{\gamma\ge 0}\left\{\lambda\rho + \mathop{\mathbb{E}}_{\bar{P}(\omega')}\left[\sup_\omega \mathrm{Reg}(\pi,\omega) - \lambda\kappa(\omega,\omega')\right]\right\}. \end{aligned} \qquad (21)$$

$\lambda \ge 0$时，有如下公式：

$$\begin{aligned} &\sup_{P(\omega)} \{\mathbb{E}_\omega[\mathrm{Reg}(\pi,\omega)] - \lambda W_\kappa(P,\bar{P}) \\ &= \mathop{\mathbb{E}}_{\bar{P}(\omega')}\left[\sup_\omega \mathrm{Reg}(\pi,\omega) - \lambda\kappa(\omega,\omega')\right]. \end{aligned} \qquad (22)$$

我们的目标是展示方程（16）中的目标等于最小化对策与专家之间 KL 散度差异的上界。我们首先展示了如何通过 KL 散度中的差异导出下界。

$$\mathcal{D}_{KL}(\xi,\pi) = \mathbb{E}_{\xi(a|o_t,h_t)}\left[\log\frac{\xi(a\mid o_t,h_t)}{\pi(a\mid o_t,h_t)}\right], \qquad (23)$$

其中，

$$\begin{aligned} \log\frac{\xi(a\mid o_t,h_t)}{\pi(a\mid o_t,h_t)} &= \int_\omega q(\omega\mid h_t)\log\frac{\pi(a\mid o_t,h_t)}{\xi(a\mid o_t,h_t)}d\omega \\ &= \int_\omega q(\omega\mid h_t)\log\left(\frac{\xi(a\mid o_t,h_t,\omega)}{\pi(a\mid o_t,h_t)}\cdot\frac{p(\omega\mid h_t)}{p(\omega\mid h^*)}\right)d\omega \end{aligned} \qquad (24)$$

$$\begin{aligned} &= \mathbb{E}_{q(\omega|h_t)}[-\log\pi(a\mid o_t,h_t,\omega)] - \mathbb{E}_{q(\omega|h_t)}[-\log\xi(a\mid o_t,h_t,\omega)] \\ &+\mathcal{D}_{KL}\big(q(\omega\mid h_t)\,\|\,p(\omega\mid h^*)\big) - \mathcal{D}_{KL}\big(q(\omega\mid h_t)\,\|\,p(\omega\mid h_t)\big) \end{aligned}$$

我们可以类似地推导出方程（7）的上界，就像上面展示的那样。

据此，差异可以分解为专家和策略之间的交叉熵项，专家熵的常数项，以及两个后验之间的 KL 散度项。

$$\begin{aligned} &\min_\pi \mathcal{D}_{KL}(\xi\,\|\,\pi) = \min_\pi \mathbb{E}_{\xi,q}\big[-\log\pi_\phi(a\mid o_t,h_t,\omega)\big] + \text{ const} \\ &+\mathcal{D}_{KL}\big(q(\omega\mid h_t)\,\|\,p(\omega\mid h^*)\big) - \mathcal{D}_{KL}\big(q(\omega\mid h_t)\,\|\,p(\omega\mid h_t)\big) \end{aligned} \qquad (25)$$

我们注意到，第一个 KL 项的目的是与专家的因果结构对齐。经验上发现第二个 KL 项的表现较差，因此我们只使用第一个 KL 项进行学习，从而得到上界（16）。

然后我们介绍了一篇关于模仿学习的论文中关于遗憾和基于 KL 散度的专家-策略差异的关系的定理[46]。

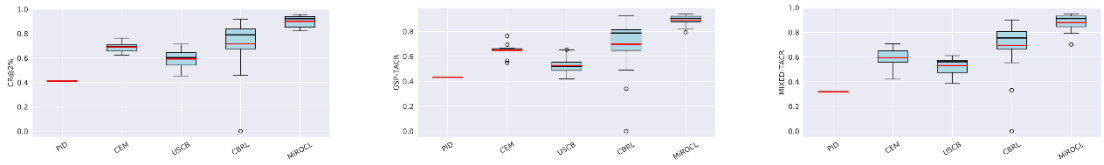

图 8 上方为工业数据集（顶部）和合成数据集（底部）的结果，包括 CR@2%（左）、IID-TACR/GSP-TACR（中）和 OOD-TACR/MIX-TACR（右）。每个箱线图展示了 20 次独立重复运行的平均值（红色）和中位数（黑色）结果。

定理 2 在给定 $MDP\ \mathcal{M}_\omega$ 的情况下，如果 $\mathbb{E}_{d_\omega^{\xi^*}}[\mathcal{D}_{KL}(\xi^*,\pi)] \le \epsilon$，则有 $Reg(\pi;\mathcal{M}_\omega) \le 2\sqrt{2}H^2\sqrt{\epsilon}$。

这个定理 1 表明，后悔可以由专家策略差异度度量来界定。因此，最小化目标（16）直接是在最小化对后悔的一个上界，这可以表示后悔上界的保证。

附录 B

实验设置，如下：

数据集。图 5 显示了工业数据集中分布的转变，分为内分布和外分布两部分。对于合成数据集，我们模拟市场竞争和投资回报率的高低时期，将其视为不同振幅和相位的正弦函数。动态混合拍卖是通过采样拐点然后在这些点之间插值而构建的。

实现。策略被实现为 BERT transformer，包括一个用于$q()$的编码器和一个用于$\pi\phi()$的解码器。地面实际后验$p()$与$q()$共享编码器，但是是双向的，即没有未来屏蔽。我们采用熵正则化的 RL 目标，该目标最小化策略和状态-动作值函数的 Boltzmann 分布之间的 KL 散度。状态-动作值还以推断的ωt为输入，同时通过由离分布奖励估计器返回的奖励进行学习。

附加结果。在图 7 中显示了对合成数据集的 TACR 结果，其他指标显示在图 8 中。