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𝑏（竞争性最高出价），代理赢得拍卖，表示为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𝐶≤𝐵和投资回报率（ROI）约束ROI≥L的情况下最大化累积效用𝑈：

其中，粗体字母𝑥和𝑏表示拍卖的序列（特征）和出价，而下面我们将使用代表累计效用，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，其中包含了拍卖。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，除了完整状态空间S，以考虑这种部分可观察性。和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，行动被定义为安排给每个时间段的出价比例。动态模型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。策略的输入包括过去的轨迹以及当前的观察值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下最大化策略的值，其由预期的累积奖励给出。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：

(2)

这个目标优化了在MDP的分布上的策略，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：

上面的公式中代表环境的专家策略，而代表其累计价值。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 ，它还输入了关于对抗性因素𝜔 ∈ 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，

在这里，我们将遗憾函数设置为带有温度的自由能，分区函数用于规范化分布，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（稍后在第3.2.2节中定义，它代表了训练集）中，通过（Kullback-Leibler）散度，得到一个熵正则化的遗憾最大化目标，

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，在时，引发的分布仅关注最坏情况，而在时，引发的分布在训练集上均匀分布。

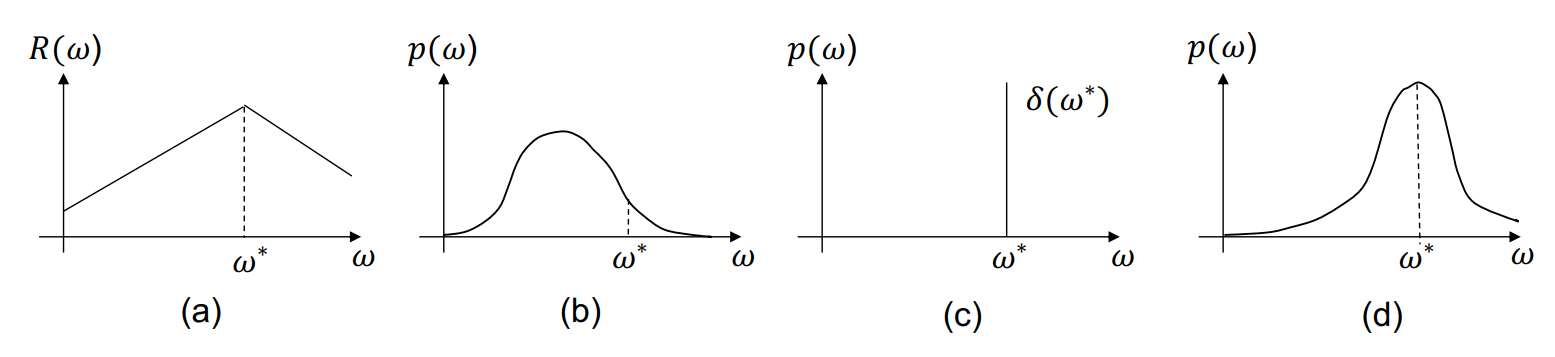


图1 ERM与MiRO对比。(a)示例了在由一维𝜔特征表征的不同环境下的。(b)展示了训练环境的经验分布。(c)展示了测试环境的分布，违反了ERM的iid假设。相比之下，MiRO假设它与成比例，如(d)所示。

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：

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，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。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，替换方程（6）中的环境为。其次，世界模型重建建立了从到奖励的映射，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：专家和策略对环境的干预。这两个策略的干预导致了观察到的变量和未观察到的变量之间的因果关系。然而，在变量和因果关系（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。

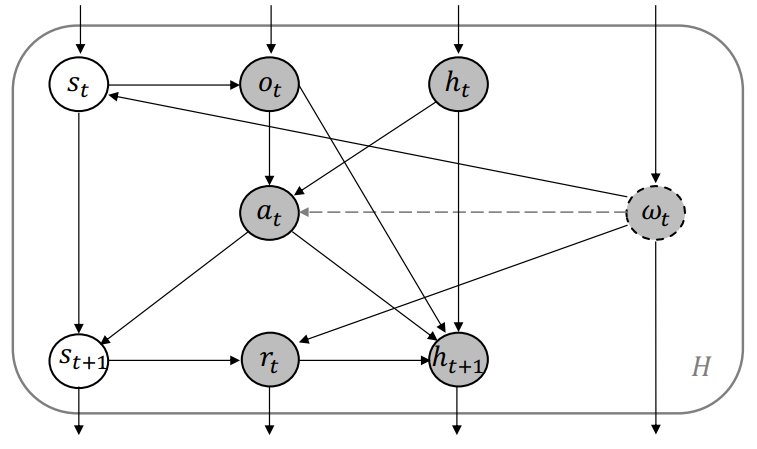


图2 在时间步骤𝑡上展开的因果图。观察到的变量被着色，而未观察到的变量没有。变量𝜔𝑡（虚线边缘）在专家的因果图G𝜉中被观察到，并指向𝑎𝑡，而在策略的因果图G𝜋中未观察到，与𝑎𝑡没有关联。

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(𝜔𝑡|ℎ𝑡)，将历史轨迹ℎ𝑡映射到时刻-𝑡对手变量 𝜔𝑡；(2)观察模型 𝑝(𝑜𝑡, 𝑎𝑡, 𝑟𝑡 | 𝜔𝑡, ℎ𝑡)，从历史和对手中恢复观察值 (𝑜𝑡, 𝑎𝑡, 𝑟𝑡)；(3)潜在动态模型 𝑝(𝜔𝑡 | 𝜔𝑡−1, 𝑎𝑡−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。

嵌入模型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：

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(𝜔)。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：

, (8)

其中（以及）表示 MDP M𝝎 中的策略（和）的状态-动作访问。

通过世界模型的重建，奖励估计器 E[𝑟𝑡|𝑠𝑡, 𝑎𝑡; 𝝎] 本质上是作为观察模型 𝑝(𝑜𝑡, 𝑎𝑡, 𝑟𝑡|𝜔𝑡, ℎ𝑡) 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 𝑟𝜃(𝑜𝑡, 𝑎𝑡, ℎ𝑡, 𝝎)，作为奖励估计器的替代，满足以下的最小二乘目标：

我们用 𝐷 表示包含多个环境的训练日志，𝑟𝐻 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：

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 (𝝎) 来辅导以前的策略，以及一位学习者在给定的环境分布 𝑃 (𝝎) 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 ∈ 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 𝑊𝜅 (·, ·)，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，其中𝝎𝑖表示第𝑖个环境。然后我们定义集合为

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：

其中表征一个收集到的环境。在实现中，我们采样一组以表征的日志环境，并通过步长 𝜂 进行基于梯度的更新，

通过从采样的环境中获取一个训练环境的分布，表示为

所以，通过从最坏情况环境的分布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。

因此，在实施中，每个敌对环境与采样环境关联，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

**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

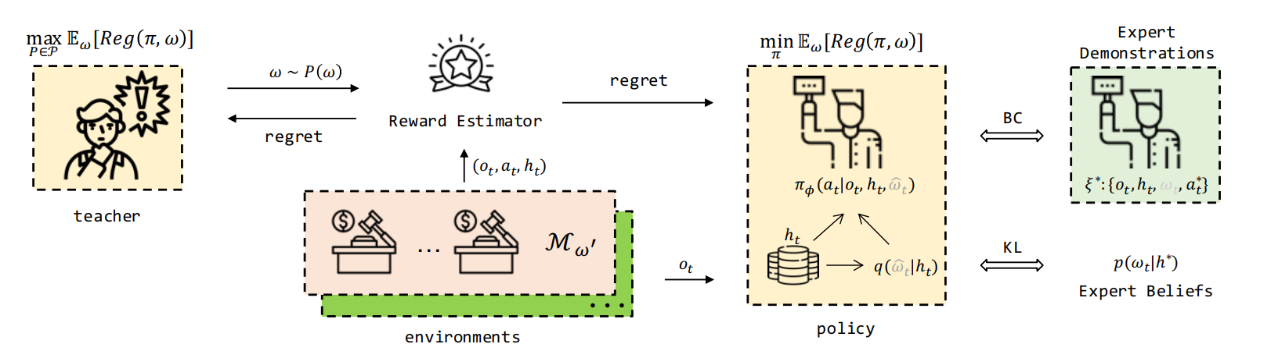
用。然而，我们的目标是通过学习专家演示来改进方程（14），因为我们认为专家可能包含在不同环境中如何最优行动的宝贵知识。然而，令人惊讶的是发现直接的行为克隆方法，即模仿专家演示，只会导致性能下降（见第 4.2.2 节）。为了理解这个问题，我们将策略学习视为因果推断问题。我们确定了未观察到的混杂问题，使得无法从观测数据中唯一计算策略。在接下来的部分中，我们首先阐述这一现象，然后提出了一种基于因果关系的对齐策略来解决这个问题。

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(𝝎)∈𝑃，并且学习者在给定环境分布上元学习其策略𝜋。为了在因果结构方面与专家保持一致，策略𝜋被设计为𝜋𝜙，其取决于通过推理模型获得的。除了来自值最大化的监督之外，子策略𝜋𝜙和推理模型还从专家演示和专家的后验信念中获得指导。

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 𝜉∗ 的投标（比例）较第二价格拍卖要低。这导致𝑎𝑡与 𝑟𝑡之间存在（虚假的）相关性（为了清晰起见，省略了(𝑜𝑡, ℎ𝑡) = (𝑜<𝑡, 𝑎<𝑡) 的条件），这表明较小的𝑎𝑡值与较高的奖励𝑟𝑡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(𝑎𝑡)的因果效应，即，使用专家收集的观察数据 {𝑜𝑖, ℎ𝑖, 𝑎𝑖, 𝑟𝑖}𝜉。

正如因果图中所示的那样，混淆变量𝝎为观测数据中的𝑎𝑡←𝜔𝑡→𝑟𝑡的因果结构做出了贡献，但对于策略𝜋(𝑎𝑡|𝑜𝑡, ℎ𝑡)来说是未观察到的。因此，当𝜔𝑡未观察到时，条件独立性被打破，意味着观察数据呈现𝑎𝑡与𝑟𝑡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𝜔𝑡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：

这分解成一个子策略𝜋𝜙∈Ξ和一个用于推断的推断模型。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：

第一项训练了一个子策略，该子策略输入推断的，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。