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

Haozhe Wang*
Alibaba Group
Beijing, China

Chao Du
Alibaba Group
Beijing, China

Panyan Pang
Alibaba Group
Beijing, China

Li He
Alibaba Group
Beijing, China

Liang Wang
Alibaba Group
Beijing, China

Bo Zheng
Alibaba Group
Beijing, China

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, bidding 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%.

CCS CONCEPTS

• **Information systems** → **Computational advertising**; • **Theory of computation** → **Reinforcement learning**.

KEYWORDS

Constrained Bidding, Reinforcement Learning, Causality

*Correspondence to Haozhe Wang <jasper.whz@outlook.com>

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599254

ACM Reference Format:
Haozhe Wang, Chao Du, Panyan Pang, Li He, Liang Wang, and Bo Zheng. 2023. Adversarial Constrained Bidding via Minimax Regret Optimization with Causality-Aware Reinforcement Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599254

1 INTRODUCTION

The proliferation of the Internet has led to the rise of online advertising as a multi-billion dollar industry. At the heart of online advertising lies online auctions [17], where publishers repeatedly sell ad slots to advertisers seeking brand promotion, greater conversions, etc. Traditionally, incentive-compatible auctions such as second-price auctions are widely adopted, as they possess the desirable property of 'truthful bidding' for myopic bidders – truthfully revealing private values is optimal for these non-strategic bidders in order to maximize their immediate utility [32, 34].

However, the crucial assumption of myopic bidders has become obsolete in recent times and truthful bidding does not optimize advertisers' long-term utilities. Demand-side platforms (DSPs), as an intermediary on behalf of aggregated advertisers, is now the actual entity participating in billions of auctions a day. Rather than truthful bidding, DSP agents employ bidding strategies to satisfy the demands of the diversified advertisers, who often seek to maximize certain utilities while subject to spending constraints [44]. For instance, brand advertisers seek long-term growth and awareness and typically optimize for metrics such as impressions, clicks, and subject to return-on-investment (ROI) constraints that require a minimum ratio of utilities to cost.

In order to cater to the diverse demands of the advertisers, extensive studies have been conducted on designing and learning bidding strategy. The existing literature can be broadly classified based on the setting of constraints. The majority of research has been focused on bidding subject to at most a budget constraint [6, 7, 12, 13, 20, 21, 24], which may not fully capture the diversity of spending constraints in the field. To address this limitation, a few studies [27, 41, 47] have explored optimal bidding with ROI-like constraints. The ROI-Constrained Bidding (RCB) problem, which involves ensuring that the ROI-like constraints, such as acquisition-per-cost and click-per-cost, exceed a prescribed limit while adhering to a budget constraint, is viewed as a prototypical problem that generalize to the diverse advertising objectives [41].

Despite promising results obtained by previous methods [27, 41], they typically follows the Empirical Risk Minimization (ERM) principle [31, 42] that relies on the assumption of independent and identically distributed (i.i.d.) train and test conditions. This contradicts many real-world situations where the sellers and other rival bidders behave adversarially, as all parties seek to optimize their own utilities that are potentially conflicting with one another [34]. For instance, sellers may learn the distribution of bidders' private values and set personalized reserved prices in auctions [14, 18, 35]. Recent studies [16, 37] have introduced neural network-based selling mechanisms learned from data. In addition, rival bidders can also employ complex bidding strategies to optimize their long-term utility [15], leading to a complex distribution of competing bids that can affect the performance of our bidding agent [26]. These considerations point out the inherently adversarial nature of the bidding environment (also summarized as non-stationarity in [41]).

The problem of bidding in adversarial environments has remained largely unexplored, with only a few recent works [26, 33] showing developments. These works focused on adversarial bidding *in the absence of* constraints in *first-price auctions* [26] or relied on *assumptions about the adversaries* [33]. However, there remains an uncharted problem that we aim to investigate in this paper, namely constrained bidding in black-box adversarial environments, which assumes no knowledge about how exogenous factors cause adversarial perturbations to the bidding environment. From a game-theoretic standpoint, a black-box adversary purposefully perturb the bidding environment, such as altering the market dynamics or the value distribution, exploiting its knowledge about the bidder. The bidder is therefore subjected to test environments where it potentially perform worse, and it must behave adaptively in the adversarial environments to achieve optimal performance.

To address the issue of broken i.i.d. assumption, our basic insight is to align the train distribution of environments with the potential test distribution, meanwhile minimizing policy regret, i.e., the performance discrepancy between the policy and a pre-computed optimal policy (referred to as the 'expert' policy). Based on this insight, we derive a Minimax Regret Optimization (MiRO) framework which interleaves between identifying the aligning train distribution in the inner supremum problem and optimizing the regret-minimizing policy under such distribution in the outer infimum problem. While MiRO appears appealing, we discovered that both the inner and the outer problems have practical limitations that necessitate refinement to achieve optimal performance:

- The inner problem (Sec. 3.2). For lack of knowledge about the exact function structure of the environments and the adversarial factors, it is infeasible to obtain a tractable inner supremum problem that directly optimizes for a distribution. To address this issue, we propose a data-driven approach that learns the latent representation of the adversarial factors by reconstructing the causal structure of world model. This renders a differentiable game that can be optimized end-to-end.
- The outer problem (Sec. 3.3). Although regret minimization aims to close the gap between the policy and the expert, policy learning indeed degenerates to a value maximization problem with the experts taking no effects. To circumvent this issue, we seek to explicitly utilize the useful knowledge from the experts to guide policy learning. Surprisingly, we find that the straight-forward behavioral cloning approach does not work due to the unobserved confounding issue [36]. To overcome this challenge, we develop a causality-aware alignment strategy that factorizes into a sub-policy mimicking the causal structure of the experts.

The effectiveness and the generalizability of our method, Minimax Regret Optimization with Causality-aware reinforcement Learning (MiROCL), is validated through both large-scale synthetic data and industrial data.

2 BACKGROUND AND PRELIMINARIES

In this section, we first describe the standard constrained bidding problem, and then introduce the common RL formulation as the foundation of our paper.

Real-time Bidding (RTB) is an important marketing channel in online advertising, which enables advertisers to gain exposure across multiple media and helps publishers to achieve monetization through the effective distribution of their traffic [17]. The advertisers resort to demand-side platforms (DSPs) who buy and display ads on their behalf. DSP agents repeatedly interact with billions of auctions a day and employs bidding strategies to optimize the advertisers' long-term objectives subject to various spending constraints, leading to a surge of research interest in constrained bidding [12, 20, 21, 24, 41].

Conventionally, the constrained bidding problem considers an RTB process comprised of auctions arriving sequentially, with the goal of scheduling bids for each auction to optimize the target utility while satisfying the relevant constraints. Suppose the bidding process consists of T repeated auctions. At each auction triggered for an ad opportunity, the bidding agent is given (partial) information x_i regarding the auction, which summarizes key details such as information about the user, the selected ad and the display context. Based on this information, the agent must decide on a bid price b_i. If the bid exceeds the market price $m_i = \max b_i^-$ (the maximal competing bids), the agent wins the auction, denoted as $\mathbb{1}_{b_i > m_i}$. The winning auction entails a charge of c_i for ad display according to the selling mechanism prescribed by the publisher, and sends feedback about the relevant utilities, e.g., clicks, conversions. Conversely, losing auctions result only in a loose notice. In this work, we assume the online auctions adopt (or stem from) second-price auctions, holding the property of incentive compatibility [32].

In the following, we focus on the ROI-Constrained Bidding (RCB) setup, which serves as a prototypical problem that can generalize to diverse advertising objectives [27, 41]. The problem of RCB aims to maximize the cumulative utility U subject to a budget constraint $\mathbb{1}_{C \le B}$ and a return-on-investment (ROI) constraint $\mathbb{1}_{\mathrm{ROI} \ge L}$:

$$\max_{\mathbf{b}} U_T(\mathbf{b}; \mathbf{x}), \text{ s.t. } \frac{U_T(\mathbf{b}; \mathbf{x})}{C_T(\mathbf{b}; \mathbf{x})} \ge L, \; B - C_T(\mathbf{b}; \mathbf{x}) \ge 0, \tag{1}$$

where the bold letters, $\mathbf{x}$ and $\mathbf{b}$, denote the sequence of auction (features) and bids, the quantity $U_T(\mathbf{b}; \mathbf{x}) \stackrel{\text{def}}{=} \sum_{i=1}^{T} \mathbb{E}\left[u_i | x_i \right] \mathbb{1}_{b_i > m_i}$ denote the cumulative utility and $C_T(\mathbf{b}; \mathbf{x}) \stackrel{\text{def}}{=} \sum_{i=1}^{T} \mathbb{E}\left[c_i | x_i \right] \mathbb{1}_{b_i > m_i}$ denote the cumulative cost.

Existing approaches on RCB have achieved promising results based on a reinforcement learning (RL) formulation, empowered by its ability in long-term planning [27, 41]. Following this trend, we adopt the Partially Observable Constrained MDP (POCMDP) formulation proposed in CBRL [41] as the foundation of our work. In

the following, we briefly summarize the main idea of the POCMDP formulation and we refer the readers to Wang et al. [41] for details.

Many leading DSPs, such as Google [23] and Alibaba [3], experience traffic throughput at the scale of billions, which poses a challenge for RL training due to the excessively lengthy decision sequences if each auction is viewed as a decision step. To mitigate this issue, CBRL adopts a distinctive perspective of the RCB problem at the aggregate level, which disentangles impression-level bid decision as a slot-wise bid ratio controlling problem with impression-level utility prediction, drawn upon the following optimal bidding theorem for constrained bidding problems [6, 27, 41].

Theorem 1. *In second-price auctions, the optimal bidding function for problem* (1) *takes the linear form of* $b_i = a\, u_i$ $(a > 0)$.

This theorem states that the optimal bids (in hindsight) equals the impression value u weighted linearly by a ratio a. Therefore, instead of treating each auction as a decision step, CBRL proposed to control the slot-wise bid ratio a within a time window as a decision step, and the final bids can be computed by multiplying the bid ratio with each impression-level utility u.

Built upon this slot-wise ratio controlling formalism, CBRL proposed to model the bidding process as a finite-horizon episodic RL problem with H time steps. Each time step t represents a time window $[j_t, j_{t+1})$ containing auctions $\{x_i\}_{i\in[j_t, j_{t+1})}$. Given that the market price is only known when an auction is won, the POCMDP introduces an observation space $\mathcal{O}$, in addition to the full state space $\mathcal{S}$, in order to account for this partial observability. Both $\mathcal{S}$ and $\mathcal{O}$ contain slot-wise statistics (e.g., winning rate, ROI, total revenue and cost) but S also includes information that the agent cannot observe (e.g., market price). Within this framework, the action $a \in \mathcal{A}$ is defined as the bid ratio scheduled for each time slot. The dynamics model $P(s', r|s, a) = \mathcal{T}(s'|s, a) \cdot P(r|s, a)$ accounts for the transition and reward function conceptually, but the exact function mappings are unknown.

Specifically, for the transition model, we assume partial observability of the market since the selling mechanisms and rival strategies are not transparent in sealed-bid autions. For the reward model, the step-wise reward should conceptually account for both utility and constraint violations, which involves non-trivial credit assignment to each time slot. Despite the unknown dynamics model, we can still simulate bidding environments as long as the market price is known. In this regard, past bidding logs can construct a large number of bidding environments to serve as our dataset. In addition, we can compute optimal decision sequences for each environment by solving linear programs [41], which we will denote as expert trajectories in the following passages.

In the above POCMDP formulation, we aim to find a policy π which is a member of the policy space $\Pi : \mathcal{O} \times \mathcal{H} \mapsto \mathrm{P}(\mathcal{A})$. The policy inputs past trajectories $h_t = \{o_i, a_i, r_i\}_{i=1}^{t-1}$ in addition to the current observation o_t, as is common practice in partially observable MDPs [45]. The standard objective for identifying a stationary policy is to maximize the policy's value under $\mathcal{M}$, given by the expected cumulative reward $V(\pi; \mathcal{M}) = \mathbb{E}\left[\sum_{t=1}^{H} r_t|\pi, \mathcal{M}\right]$. As the bidding environment can differ from day to day, it is crucial for the bidding policy to act adaptively in varying conditions. To this end, previous methods typically adopt the following RL objective,

$$\max_{\pi} \mathbb{E}_{\mathcal{M}}\left[V(\pi; \mathcal{M})\right], \tag{2}$$

which optimizes the policy over a distribution of MDPs $p(\mathcal{M})$, assuming the test distribution of environments is i.i.d. with the train distribution. Essentially, the objective embodies the principle of meta-reinforcement learning [42, 48], aiming to meta-learn an adaptive policy that generalizes across multiple environments.

3 METHODOLOGY

While it is well agreed recently that online ad markets dynamically changes [27, 41], we emphasize that the bidding environment can be essentially adversarial [26, 33] because online auctions involve multiple parties with conflicting objectives. For instance, the sellers may update their mechanism towards maximal revenue, e.g., by learning personalized reserve prices [14, 18], or even automatically learn mechanisms from data [16, 37]. On the other hand, rival bidders can employ data-driven auto-bidding algorithms to optimize their own utilities. Moreover, the propensity of users to click an ad can vary over time due to exogenous factors, leading to incorrect utility estimations [19, 34]. Observations in real-world data also support these conjectures, as detailed in the appendix.

Unfortunately, none of the adversarial factors are directly observable to our agent, as online auctions typically seal the competing bids and the sellers have little incentive to disclose how they update their selling mechanisms. In light of this, we explore in this paper the uncharted problem of constrained bidding in adversarial environments (i.e., adversarial constrained bidding), which assumes no knowledge about how adversarial factors cause perturbations to the environment. From a game-theoretic perspective, we aim to design a bidding strategy that can effectively resist the black-box adversary in the subsequent round, leveraging the interaction history of previous rounds.

Adversarial constrained bidding is especially challenging because the widely adopted assumption of i.i.d. train and test environments is violated, as the test environments in the adversarial setting can be purposefully manipulated towards unfavorable ones, as depicted in Fig. 1. Consequently, existing works following this assumption is inapt for adversarial constrained bidding.

To address this issue, our main insight is to *align the train distribution of environments with the potential test distribution* instead of relying on the i.i.d. assumption. In what follows, we give a practical solution that realizes this insight. We first illustrate how to mathematically formulate this insight into a Minimax Regret Optimization (MiRO) framework. Then we discuss several practical issues and elaborate on how to improve upon the crude MiRO framework.

3.1 The MiRO Framework

In this section, we first address two fundamental questions related with achieving train-test distribution alignment using collected bidding logs, and finally reach at the proposed minimax regret optimization (MiRO) framework for adversarial constrained bidding. Specifically, to achieve train-test alignment, we must first answer the following questions: what property does the adversarial setting imply about the test distribution? And secondly, how to identify the aligning train distribution given such property?

3.1.1 The Property of the Test Distribution. Through the lens of game theory, an adversary with complete knowledge of our prior strategy can perturb the environment into a worst-case scenario

for this strategy. While it remains uncertain as to how such an adversary could feasibly exploit our strategy in practice, the strict condition provides us with a valuable insight into the generic adversarial setting. SSpecifically, we are led to believe that the likelihood of an environment encountered during testing is proportional to the performance a policy performs in that environment.

To mathematically formulate this idea, we must first introduce the performance measure of policies (i.e., bidding strategies). In the adversarial setting, there are no stationary policies for all environments, so we use regret as a relative performance metric w.r.t. an oracle for each environment [9, 26, 40]. Regret measures the value difference of two policies, an optimal policy (known as oracle or prophet [40] but we mention as the expert in the following passages), and the policy being learned:

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

where $\xi^* = \arg\max_{\xi\in\Xi} V(\xi;\mathcal{M})$ denotes an expert policy for environment $\mathcal{M}$ and $V^*_{\mathcal{M}}$ shorthands for its cumulative value. It should be noted that, since we do not know the exact function structure of the environment, we cannot directly solve for the expert policy function that maps from any given observations to the optimal decisions. Indeed, we compute the expert trajectories based on the offline bidding logs using approximate dynamic programming, and hence the expert demonstrations conceptually requires the knowledge of anticipatory information about future dynamics. To this end, we conceptually define the expert policy space as $\xi \in \Xi : \mathcal{O}\times\mathcal{H}\times\mathcal{W} \mapsto \mathrm{P}(\mathcal{A})$, which additionally inputs the privileged information of adversarial factors $\omega \in \mathcal{W}$.

Since we must choose a representation for the test distribution, we opt for using the general energy-based distribution to express the aforementioned proportional property of the test distribution,

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

where we set the regret function $\mathrm{Reg}(\pi;\mathcal{M})$ as the free energy with temperature α, and the partition function $Z(\pi)$ serves to normalize the distribution, albeit without contributing gradients.

3.1.2 Identifying the Aligning Train Distribution. Having established the potential form of the test distribution, our focus now turns to identifying an appropriate train distribution from the collection of environments. To ensure that the aligning train distribution adheres to the available training set, we choose to project the train distribution into the set of parameterized distributions $\mathcal{P}$ (defined later in Sec. 3.2.2, which delegate the training set) in terms of the Kullback-Leibler (KL) divergence, which turns out an entropy-regularized regret maximization objective,

$$\begin{aligned}&\min_{P\in\mathcal{P}} \mathcal{D}_{KL}\left(P\|P_{\text{test}}\right)\\ =&\max_{P(\mathcal{M})\in\mathcal{P}} \mathbb{E}_{\mathcal{M}}[\mathrm{Reg}(\pi;\mathcal{M})] + \alpha\mathcal{H}(P) + \text{const.}\end{aligned} \quad (5)$$

Intuitively, the objective aims to find a distribution of environments within the training set that induces high policy regret, while adhering to the maximum entropy principle since we have no knowledge about the real adversary. The entropy regularizer controls how the distribution bias towards the worst-case environments

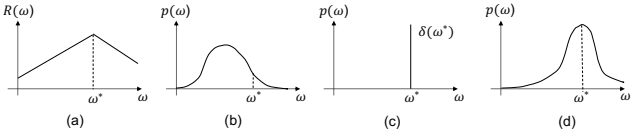

Figure 1: ERM v.s. MiRO. (a) exemplifies $\mathrm{Reg}(\pi,\omega)$ under different environments characterized by one-dimensional ω. (b) shows the empirical distribution of training environments. (c) shows the distribution of the test environment, which violates the iid assumption of ERM. In contrast, MiRO assumes it proportional to $\mathrm{Reg}(\pi,\omega)$ as (d) shows.

with the temperature hyperparameter α. In effect, this hyperparameter anneals the interpolation between the strict adversarial setting and the iid stochastic setting, reflecting the belief of the adversarial setting. More specifically, at $\alpha = 0$, the induced distribution solely focuses on worst-case scenarios, while at $\alpha \to \infty$, the induced distribution puts a uniform mass over the training set.

3.1.3 Minimax Regret Optimization. Given the aligning train distribution, the policy aims to minimize its regret under this distribution. Therefore, we reach at the following (entropy-regularized) minimax regret optimization (MiRO) framework,

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

The MiRO framework presents a minimax game between two players, where the inner problem searches for a distribution of environments $P(\mathcal{M})$ likely to align with the adversarial test conditions, and the outer problem optimizes the policy's performance under the given environments. Compared with Empirical Risk Minimization (ERM) widely adopted in previous works, which assumes small training empirical risk could generalize to small test risk, MiRO implies generalization because the policy strives to minimize the (approximately) worst-case regret that upper bounds test regret.

While MiRO appears appealing, solving such a bi-level optimization is generally intractable. Our main idea is to convert the minimax problem into a class of 'differentiable games' [4] so that we can resort to dual ascent [10] to search for a useful solution, as supported by generative adversarial networks [22] and its follow-up works.

3.2 A Practical Algorithm for MiRO

In this section, our goal is to make Eq. (6) a differentiable game, which is circumscribed by the unknown structure of the environment M due to unobservable adversarial factors. To overcome this challenge, we propose to learn the latent representation $\omega \in \mathcal{W}$ of those adversarial factors from bidding logs by reconstructing the causal structure of the world model. Through world model reconstruction, we can achieve two key benefits. Firstly, since the adversarial factors ω explain the variations of the environment, we can instead search for the distribution $p(\omega)$ supported in the learned latent space, replacing the environment $\mathcal{M}$ with ω in Eq. (6). And secondly, world model reconstruction establishes the map from ω to the rewards r, which enables the differentiation through the regret function. To this end, we can directly search for a distribution in MiRO with gradient-based optimization.

3.2.1 Rendering a Differentiable Game. To learn representations for the adversarial factors that can reflect the cause-effects on the

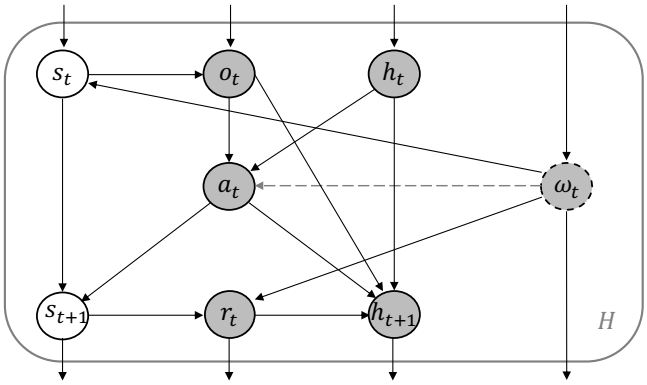

Figure 2: Causal Diagram un-rolled at step t. Observed variables are shaded, while unobserved variables are not. The variable ω_t (doted edge) is observed and directed to a_t in the expert's causal diagram $\mathcal{G}^{\xi}$ while unobserved with no link to a_t in the policy's causal diagram $\mathcal{G}^{\pi}$.

environment, we first analyze the causal structure of the environment, and then leverage the Variational Information Bottleneck (VIB) [2] for representation learning, which aims to learn maximally compressed representations from the input that retain maximal informativeness about the output.

We begin by describing the causal diagram [36] (a type of augmented probabilistic graphical models [8] used to analyze the cause-effects between random variables) of a folded time step of the bidding process. Fig. 2 shows the cause-effects of two policies: the expert $\xi \in \Xi$ and the policy $\pi \in \Pi$ intervening with the environment. Both policies' interventions give rise to the observed variables $\{o_t, a_t, h_t\}_{t=1}^{H}$ and the unobserved variables $\{s_t\}_{t=1}^{H}$, with the causal relationships between these variables. However they differ at the variables $\{\omega_t\}_{t=1}^{H}$ and the cause-effects $\{a_t \rightarrow \omega_t\}_{t=1}^{H}$ (dotted line), since the expert conceptually knows the privileged information to give optimal decisions.

Based on these cause-effect relationships, we construct a world model including the following components: (1) the embedding model $p(\omega_t|h_t)$ that maps the history trajectory h_t to step-t adversary variable ω_t; (2) the observation model $p(o_t, a_t, r_t|\omega_t, h_t)$ that recovers the observations (o_t, a_t, r_t) from the history and adversary; (3) the latent dynamics model $p(\omega_t|\omega_{t-1}, a_{t-1})$ that models the transition in the embedding space. These probabilistic models are assumed Gaussian distributions with mean and variance implemented as neural network function approximators. For example, the embedding model takes the form of $p(\omega_t|h_t) = \mathcal{N}(\omega_t|f_\theta^\mu(h_t), f_\theta^\sigma(h_t))$.

The embedding model $p(\omega_t|h_t)$ provides the latent representation for adversarial factors, and is learned by reconstructing the observed evidence of the environment, which derives the following lower bound based on VIB:

$$\begin{aligned}\max_{d^\pi, d^\xi} \mathbb{E}_{d^\pi}\left[\log p(o_t, r_t|\omega_t)\right] + \mathbb{E}_{d^\xi}\left[\log p(a_t|\omega_t)\right] \\ -\beta \mathcal{D}_{KL}\left(p(\omega_t|h_t) \| p(\omega_t|\omega_{t-1}, a_{t-1})\right),\end{aligned} \tag{7}$$

where the first two terms are the evidence of the observations, and the last term serves as a KL regularity for information compression with a hyperparameter β controlling its strength.

Since the adversarial factors ω explain the variations in the environments, we replace the environment $\mathcal{M}$ with ω in Eq. (6) and instead search for a distribution $p(\omega)$ that represents the aligning train environments in the learned latent space. Meanwhile, we also achieve a tractable gradient-based search with a differentiable regret function w.r.t. ω. To show this, we first write out the regret function as follows,

$$\begin{aligned}&\mathrm{Reg}(\pi, \omega) = V(\xi^*; \mathcal{M}_{\boldsymbol{\omega}}) - V(\pi; \mathcal{M}_{\boldsymbol{\omega}}) \\ &= \mathbb{E}_{d_{\boldsymbol{\omega}}^{\xi^*}}\left[\sum_{t=1}^{H} \mathbb{E}\left[r_t|s_t, a_t; \boldsymbol{\omega}\right]\right] - \mathbb{E}_{d_{\boldsymbol{\omega}}^{\pi}}\left[\sum_{t=1}^{H} \mathbb{E}\left[r_t|s_t, a_t; \boldsymbol{\omega}\right]\right],\end{aligned} \tag{8}$$

where $d_{\boldsymbol{\omega}}^{\pi}(s_t, a_t)$ (and $d_{\boldsymbol{\omega}}^{\xi^*}(s_t, a_t)$) denotes the policy π's (and ξ^*'s) state-action visitations in MDP $\mathcal{M}_{\boldsymbol{\omega}}$.

We note that, through world model reconstruction, the reward estimator $\mathbb{E}\left[r_t|s_t, a_t; \boldsymbol{\omega}\right]$ is inherently learned as a component of the observation model $p(o_t, a_t, r_t|\omega_t, h_t)$. Specifically, we learn a neural network function approximator $r_\theta(o_t, a_t, h_t, \boldsymbol{\omega})$ as a surrogate for the reward estimator, per the following least square objective,

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

where D is the training logs containing multiple environments, and we r^H denotes the episode-level reward.

3.2.2 Optimizing the Differentiable Game. Finally, we reach at a differentiable game that can be optimized by an simultaneous gradient descent procedure (i.e., dual ascent) as follows,

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

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

Intuitively, the game alternates between the optimization of two players simultaneously – a teacher who tutors previous policy by finding a distribution $P(\boldsymbol{\omega})$ of worst-case environments in the learned latent space, and a learner who meta-learns its policy π over the given distribution of environments $P(\boldsymbol{\omega})$.

The worst-case tutoring step. To ensure that the train distribution should adhere to the empirical dataset, i.e., $P \in \mathcal{P}$, we opt for defining the set $\mathcal{P}$ based on the Wasserstein distance, which has been shown to derive convenient forms under proper assumptions [38].

The Wasserstein distance computes the minimum cost of transform one distribution into the other, known for the property to capture the geometry of the latent space [1]. Specifically, we define the Wasserstein metric $W_\kappa(\cdot, \cdot)$ with an L2-norm cost function $\kappa(\boldsymbol{\omega}, \boldsymbol{\omega}') = \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2$. Intuitively, we aim to define the set $\mathcal{P}$ as the ρ-neighborhood of the empirical distribution in the latent space under the Wasserstein metric. To realize this, we first denote the empirical distribution of collected environments in the latent space as $\tilde{P}(\boldsymbol{\omega}) = \frac{1}{M}\sum_{i=1}^{M} \delta(\boldsymbol{\omega}_i)$ where $\boldsymbol{\omega}_i$ denotes the i-th environment. Then we define the set as $\mathcal{P} = \{P : W_\kappa\left(P, \tilde{P}\right) \leq \rho\}$.

Based on a dual re-formulation of Eq. (10) (detailed in the appendix A.1), we reach at the following objective,

$$\max_{P \in \mathcal{P}} \mathbb{E}_{\boldsymbol{\omega}}\left[\mathrm{Reg}(\pi, \boldsymbol{\omega})\right] = \mathbb{E}_{\tilde{\boldsymbol{\omega}}}\left[\max_{\boldsymbol{\omega}} \mathrm{Reg}(\pi, \boldsymbol{\omega}) - \lambda\|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_2\right], \tag{12}$$

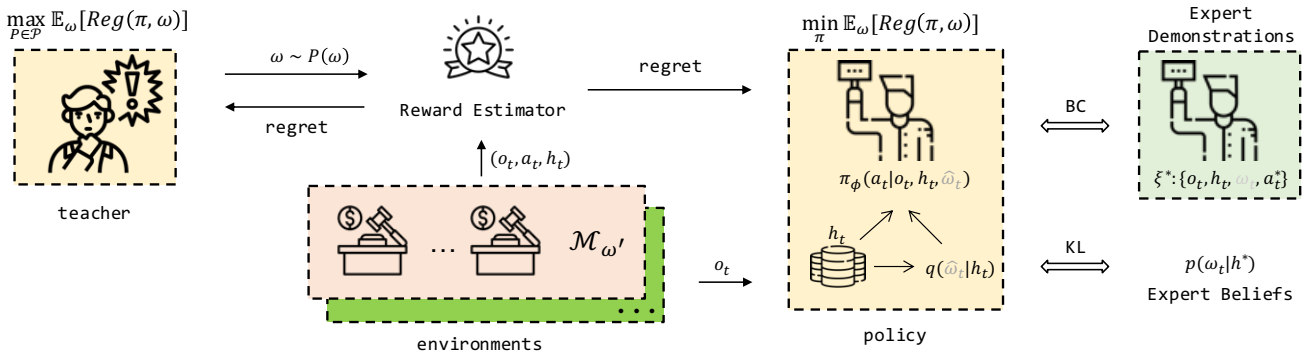

Figure 3: Overview for MiROCL. Our method solves a differentiable game alternating between a teacher and a learner. The teacher finds a distribution $P(\omega) \in \mathcal{P}$ of worst-case environments, and the learner meta-learns its policy π over the given distribution of environments. In order to align with the experts in terms of causal structure, the policy π is designed as π_ϕ conditioning on $\hat{\omega}$ obtained via the inference model $q(\hat{\omega}_t|h_t)$. Besides supervision from value maximization, the sub-policy π_ϕ and the inference model receive guidance from expert demonstrations and experts' posterior beliefs.

where $\tilde{\boldsymbol{\omega}}$ characterizes a collected environment $\mathcal{M}_{\tilde{\omega}}$. In implementation, we sample a set of logged environments characterized as $\{\tilde{\omega}_i\}_{i=1}^n$ and by gradient-based updates with stepsize η,

$$\boldsymbol{\omega}' \leftarrow \tilde{\boldsymbol{\omega}} + \eta \nabla_{\boldsymbol{\omega}} \left[\mathrm{Reg}(\pi, \boldsymbol{\omega}) - \lambda \|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}\|_2 \right], \tag{13}$$

we obtain an train distribution of environments from the sampled environments, denoted as $\{\boldsymbol{\omega}_i'\}_{i=1}^n$.

The policy improvement step. Given the distribution of worst-case environments $P(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^n \delta(\boldsymbol{\omega}_i')$, the learner's policy improvement step according to Eq. (11) becomes the standard value maximization objective, as the expert value is constant w.r.t. π,

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

Here we note that, in implementation, each adversarial environment $\boldsymbol{\omega}$ associates with a sampled environment $\tilde{\boldsymbol{\omega}}$ due to the paired graident-based search in Eq. (13). Following deep RL [25], we implement the policy distribution as a Gaussian distribution, with its mean and variance parameterized by neural network function approximators, i.e., $\pi(\cdot|o_t, h_t) = \mathcal{N}(\cdot|f^\mu(o_t, h_t), f^\sigma(o_t, h_t))$.

3.3 Causality-aware Learning with Experts

Although MiRO is designed to minimize the policy regret (c.f. Eq. 6), Eq. (14) indeed degenerates to a value maximization problem without the experts' involvement. Interestingly, previous works on constrained bidding have also ignored the role of experts in their learning objectives. However, we aim to improve upon Eq. (14) by also learning from expert demonstrations, as we believe the experts could entail valuable knowledge on how to optimally act in different environments. Nonetheless, It is surprising to find that the straight-forward behavioral cloning approach, which involves imitating expert demonstrations, only results in decreased performance (Sec. 4.2.2). To understand this issue, we view policy learning as a causal inference problem. We identify the issue of unobserved confounding that makes the policy not uniquely computable from observational data. In the following passages, we first illustrate this phenomenon and then propose a causality-based alignment strategy to remedy this issue.

To exemplify the issue of unobserved confounding in policy learning, consider an auction environment where the selling mechanism is modified such that the winning cost is greater than the second price. In this scenario, the expert policy ξ^* would exhibit lower bid (ratios) in comparison to second-price auctions due to the increased cost. This results in a (spurious) correlation between a_t and r_t (the conditioning of $(o_t, h_t) = (o_{<t}, a_{<t})$ is omitted for clarity), which suggests that smaller values of a_t associate with higher rewards r_t. Unfortunately, a policy learned through behavioral cloning would capture such spurious correlations as it learns only statistical dependency. As a result, such a policy would fail to generalize to environments with different selling mechanisms.

Through a causal lens, decision problems can naturally be formulated as causal queries where we aim to infer the outcome under the intervention of actions. In light of this, learning from expert demonstrations can be translated into estimating the causal effects of interventions $do(a_t)$ on future rewards, i.e., $p(\sum_{i=t}^H r_i | o_t, h_t, do(a_t))$, using the observational data collected by the experts $\{o_i, h_i, a_i, r_i\}_\xi$. As shown in the causal diagram Fig. 2, the confounding variable $\boldsymbol{\omega}$ contributes to the causal structure of $a_t \leftarrow \omega_t \rightarrow r_t$ in the observational data, but is unobserved for the policy $\pi(a_t|o_t, h_t)$. Consequently, the conditional independence $(a_t, r_t) \perp\!\!\!\perp \omega_t$ is broken when ω_t unobserved, implying that the observational data presents both the causal association and the spurious correlation between a_t and r_t. Hence, the policy cannot uniquely recover from data the desired causal query, known as the un-identifiability issue [36].

To mitigate this issue, our idea is to align the causal structure of both the expert and the policy. This is achieved by conditioning the policy with an additional input $\hat{\omega}_t$, which is designed as a surrogate for the truth ω_t unavailable to the policy π during online serving. In this regard, the policy can imitate the expert with identical causal structure so that spurious correlations is eliminated from policy learning. Therefore we adopt the following policy design,

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

which factorizes into a sub-policy $\pi_\phi \in \Xi$ and a inference model required to infer $\hat{\omega}_t$. We note that this inference model is exactly what we have learned per Eq. (7), and we aim to further leverage

the guidance from expert trajectories. Therefore, we derive the following bound from minimizing policy discrepancy,

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

The first term trains the sub-policy that inputs the inferred $\hat{\omega}_t$ to imitate the expert demonstration via behavioral cloning. The second term adds an additional KL regularity for the inference model by leveraging the posterior beliefs of the expert trajectories. The detailed derivation is included in Sec. A.2, which suggests that the above objective indeed minimizes an upper bound on the regret.

4 EXPERIMENTS

In this work, we propose a Minimax Regret Optimization (MiRO) framework for adversarial constrained bidding, with a practical algorithm for end-to-end optimization. In addition, we are the first to advocate policy learning with expert demonstrations, which enhances MiRO into *MiROCL* (MiRO with Causality-aware reinforcement Learning). Therefore, we aim to examine the following questions in the experiments.

- **Q1 (Comparison with Prior Works)**: How does the proposed method empirically perform versus prior methods in black-box adversarial environments?
- **Q2 (Ablation)**: The effectiveness of each component proposed?

For these questions, we use an industrial dataset, denoted as **Industrial**, which presents real-world adversarial situations. Since adversarial factors are entangled in real-world data, we also create a synthetic dataset, denoted as **Synthetic**, which involves varying selling mechanisms of known structure. We put details about the dataset and the implementation in the appendix, and the dataset and code is publicly available at https://github.com/HaozheJasper/MiROCL.

4.1 Experimental Setup

4.1.1 Dataset. We use two datasets in the experiments. The **Industrial** dataset is collected from the Alibaba display advertising platform, including 80 days of bidding logs with each day 2 million requests on average. Each request x_i includes: the market price m_i, the utility estimations $\mathbb{E}[u_i|x_i]$ and the real stochastic feedback u_i. Contextual features are not used in this work as we assume estimations $\mathbb{E}[u_i|x_i]$ are pre-computed. As market price m_i is not revealed in lost auctions in the RTB system, we use a special strategy that bids a price as high as possible to obtain the market price. Hence, *we consider each bidding log in the industrial dataset represents a distinct bidding environment.* For Q1, we split the industrial dataset into the first 60 days and the last 20 days, based on our observation that the two sets differ in their market dynamics and/or value distributions (Fig. 5). Thirty days are sampled from the in-distribution set to form the training set, the remaining 30 days from the set form the **IID** test set and the last 20 days as the **OOD** test set.

The **Synthetic** dataset is synthesized based on the public synthetic dataset AuctionGym [30]. In contrast to the relatively small-scale AuctionGym, our synthetic dataset includes 80 days of bidding logs with each day 10 million impressions, designed for research on constrained bidding with expert strategies. To simulate constrained bidding in adversarial environments similar to the industrial data, we assume that the real-world black-box auctions can be approximated by a format of linearly mixed second-first price auctions, i.e., the cost $c_i = k \cdot b_i + (1-k) \cdot m_i$ is a linear combination of bid b_i and market price m_i, with possibly dynamic ratio k. The assumption owes to the data insights that the charge c_i on some media channels is dependent on the bid b_i, leading to the observation that the winning probability of the same traffic distribution varies as the bids change (the charge of the same traffic distribution does not change as the bids vary in second-price auctions, and so does the winning probability). Consequently, for synthetic experiments we simulate a dynamic mixed second-first price auction environment to examine the effect of algorithms in the adversarial setting. In this case, the train set includes 10 days of GSP bidding logs, and 20 days of the dynamic mixed auction with randomly sampled ratio $k \in (0, 1)$, whereas the test set includes 20 days of GSP and 30 days of randomly sampled mixed auction logs.

4.1.2 Evaluation Protocols. We use the *competitive ratio* (CR) to evaluate methods in our experiments. CR is the ratio of the policy value versus the expert value, which directly reflects the online regret. Moreover, we introduce a notion of *tolerance* in our evaluation, and it is motivated by real practices in which we can tolerate the policy violating the constraints if it gains a sizeable payoff. Specifically, we define the metric of the average tolerance-aware comparative ratio (TACR) on a dataset of N days, with a pre-defined max tolerance γ and baseline payoff rate ζ as follows,

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

where we use abbreviations $U(i) = U_T(\pi, \mathbf{x}^i)$ to denote the cumulative utility obtained by π for request sequence $\mathbf{x}^i$ of day i. Similarly, $C(i)$ denotes the total cost, and $U^*(i)$ denotes the benchmark value.

We compute $\lambda(i) = \max\{\text{ceil}(\max\{1 - \text{ROI}(i)/L, 0\}), \gamma\}$ as the tolerance level of π's solution for day i. Intuitively, this quantity measures the competitive ratio allowing the violation of constraints within a maximum tolerance of γ, and the values will be discounted by a baseline payoff rate ζ if it violates the constraints. In particular, we set $\gamma = 2\%, \zeta = 5\%$ in our experiment, which indicates that we consider it worthy to exchange each 1% drops in ROI for 5% increase in utility, within the max tolerance of violation 2%. In addition, we also show the competitive ratio (CR) at the max tolerance level 2%, which considers solutions that violate constraints below 2% as feasible. We denote the metric as *CR@2%*.

4.2 Empirical Results

4.2.1 Comparison with Prior Methods. This work aims to present the challenge of constrained bidding in adversarial environments, for comparisons, we select representative methods that can or can be adapted to handle the constrained bidding problem (1) with both budget constraint and ROI constraint: (1) The PID control method [47] and the cross-entropy method CEM [29] are viewed as online learning approaches; (2) USCB (2021) and CBRL (2022) are two recently proposed RL-based methods trained using collected bidding logs. Among them, CBRL establishes with the ERM principle and learns without expert demonstrations.

The evaluation results on the industrial dataset and the synthetic dataset are shown in Tab. 1 and Tab. 2. We empirically show that

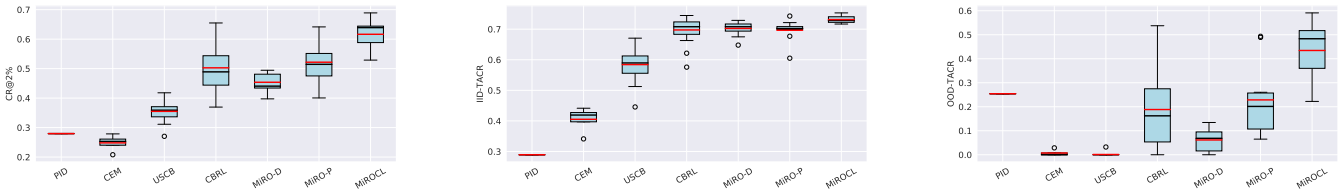

Figure 4: The results of CR@2% (Left), IID-TACR (Middle), and OOD-TACR (Right) in the Industrial dataset are shown above. Each boxplot shows the average (red) and median (black) results of 20 independent repeated runs.

the proposed MiROCL performs the best on both datasets in terms of all performance metrics. For clarification, we note while the TACR result for one independent run shall be between its IID-TACR and OOD-TACR, mTACR are not necessarily between IID-mTACR and OOD-mTACR scores because they report the median over 20 runs. Besides, mTACR is close to mCR@2% for some models (e.g., MiROCL), which implies that the models likely enjoy high payoff when violating constraints (so they are less discounted and is close to mCR@2%). In addition, some models (that follows the batch learning paradigm) show similar performance in the GSP and synthesized adversarial selling mechanisms in Tab. 2, because the models are trained on the joint set of different aution formats, leading to an averaging effect on different auction formats. The results of competing methods are as follows:

- **PID** [47] operates based on real-time feedback control. Tab. 1 shows that PID preserves relatively stable performance in OOD environments. However, we observe that PID performs worse in dynamic mixed second-first price auctions than in GSP (with GSP-mTACR 0.3817 versus MIX-mTACR 0.2837 in Tab. 2), possibly because PID is not so anticipatory as to shade some of its bid which is transformed into its cost charged by the publisher.
- **CEM** [29] is a zeroth order stochastic optimization method supposedly not affected by the distributional shift. However, we observe that the hyper-parameters of CEM are sensitive to drastic distributional shifts (OOD scores 0 in Tab. 1). For example, the optimal bid ratios in OOD environments can lie beyond the touch of CEM's elite distribution (Fig. 5). The algorithm achieves more stable performance in the synthetic dataset (Tab. 2), because the train and test conditions are i.i.d.
- **USCB** [27] adopts a Monte-Carlo value estimation-based actor-critic RL method with a soft reward function. Tab. 1 shows that USCB fails to generalize in OOD environments, highly likely due to the unobserved confounding issue. In i.i.d. test conditions (Tab. 2), USCB performs even worse than CEM. This is because USCB's overall performance is largely determined by the hyperparameter controlling the utility-constraint trade-off. Since different selling mechanisms require a different configuration of such trade-off, a single static hyperparameter will lead to sub-optimal performance over the whole dataset.
- **CBRL** [41] proposed the POCMDP formulation that lays the foundation of our work. There are two major differences between CBRL and MiROCL. Firstly, CBRL adopts the ERM principle following the objective of Eq. (2), which preserves no generalization guarantee in the adversarial setting. Secondly, CBRL do not consider including expert demonstrations for policy learning.

Table 1: Median test scores on Industrial which features real-world adversarial conditions. The reported mTACR and mCR scores take a median across 20 random trials for all models. TACR is computed with max tolerance level 2% and payoff rate 5%, and *CR*@2% shows the un-discounted CR at tolerance level 2%. IID-mTACR and OOD-mTACR reports detailed results on two sets, the in-distribution set and the out-of-distribution set.

	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

CBRL performs better than USCB in the challenging OOD environments in terms of OOD-mTACR, because CBRL employed a Bayesian mechanism that aims to adapt to the environment. However, CBRL still performs worse than our method, since MiRO potentially aligns the train and test distribution and can thus generalize better. In i.i.d. train and test conditions (Tab. 2), CBRL outperforms USCB because it implicitly learns to infer the utility-constraint trade-off for different environments. However, The proposed method still outperforms CBRL mainly because MiROCL has distilled knowledge from the experts.

4.2.2 Ablation Study. We first recap the design and components proposed in our methodology.

- ERM v.s. MiRO. In Sec. 3.1, we pointed out the major pitfall of ERM that the i.i.d. assumption is broken in the adversarial setting. To remedy this, we propose a practical MiRO algorithm in Sec. 3.2 to achieve an elaborated train-test alignment. We choose the state-of-the-art method, **CBRL**, which follows the ERM principle as a representative of ERM-based methods, and we compare it with the practical MiRO algorithm, denoted as **MiRO-P**.
- MiRO v.s. MiRO with behavioral cloning. In Sec. 3.3, we aim to improve upon MiRO by including explicit learning from expert demonstrations due to the belief that experts entail valuable knowledge about making optimal decisions in different environments. We first examine the straight-forward idea of behavioral cloning, which requires the policy to imitate the expert demonstrations following the Maximum Likelihood Estimation (MLE) principle. This approach, denoted as **MiRO-D**, shows surprising performance degradation, which motivates the proposed causality-aware alignment strategy.

Table 2: Median scores on Synthetic. The reported mTACR and mCR scores take a median across 20 random trials for all models. We also report results on two types of auction formats.

	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

- MiRO v.s. MiROCL. Based on a causal analysis, we identify the issue of unobserved confounding that fails the straight-forward MLE principle. To address this issue, we proposed a causality-aware approach that disentangles the policy to a sub-policy that mimics the expert's causal structure and the inference model. This overall method is denoted as **MiROCL**.

We then give a summarization of the empirical results.

- ERM v.s. MiRO. As shown in Tab. 1, MiRO-P has better overall performance than CBRL. While CBRL and MiRO-P has comparable IID performance, MiRO-P shows better average performance and stability than CBRL in terms of real-world OOD situations, according to the boxplots in Fig. 4. This is because the MiRO framework allows the policy to train under a distribution more likely to align with the test distribution, thus resulting in better average performance. However, we also witnessed that MiRO is still limited in stability (large variance in boxplots) due to the unpreditability of test conditions and stochasticity during training, implying a interesting direction for future research.
- MiRO v.s. MiRO with behavioral cloning. Similar to the comparision with ERM, MiRO-P has better overall performance than MiRO-D mainly conducive to better OOD performance. In real-world OOD situations, MiRO-D shows worse average performance with higher stability as compared with MiRO-P according to Fig. 4. This is because behavioral cloning involves with the unobserved confounding issue, which learns the spurious statistical dependency that might not generalize under distributional shifts. We conjecture that MiRO-D shows better stability due to the memorization effect of behavioral cloning, implying that the decisions are less adaptive to different environments.
- MiRO v.s. MiROCL. As shown in Tab. 1, MiROCL has significant improvement over MiRO-P in terms of both IID and OOD performance, with 21% increase in OOD-mTACR in particular. The IID performance improves mainly due to the effective distillation from expert demonstrations. The OOD performance improves because MiROCL follows the causal structure of the expert policy and leverages the inference model that adaptively infer the privileged information in different environments. Notably, Tab. 2 also shows that MiROCL improves MiRO-P by a large margin (12%) in adversarial selling mechanisms, indicating that expert demonstrations effectively guide the policy toward optimality.

5 RELATED WORK

Constrained Bidding. We discuss related works on the second-price auctions. A majority of research on constrained bidding follows the iid assumption or the ERM principle. Among them, most works focus on bidding with at most a budget constraint (c.f. [5] for a survey), while some works [27, 41, 47] further propose to deal with the more challenging cost-related constraints, i.e., ROI-like constraints. Our work investigate the problem of ROI-Constrained Bidding (RCB), based on the POCMDP formulation in CBRL [41].

Connections with adversarial learning in repeated auctions. Recently, a few recent works [26, 33] discuss the adversarial learning to bid problem. [26] discusses an online learning approach in adversarial first-price auctions and deals with no constraints, which is orthogonal to our work. [33] investigate the scenario with adversarial sellers, and assumes the seller adopts data-drive mechanism [16]. In reality, however, there are multiple sources of adversarial factors none of which are observable to the agents. To address this limitation, we explore the prior-free adversarial setting.

Connections with the minimax game formulation. The formulation of a minimax game is also seen in generative adversarial networks [1, 22], regret analysis in online learning [9, 26, 43], distributionally robust optimization (DRO) in supervised learning [28, 39] and adversarial training [38]. GANs aims to achieve realistic generation that links minimizing distribution discrepancy with the minimax objective. DRO and adversarial training relates with robust generalization under distributional shift and adversarial attacks. Regret analysis aims to provably bound the performance gap under the worst online conditions. In particular, previous online learning approaches for bidding [9, 11, 19, 43] are typically limited to small-scale problems and require knowledge about the market. Our approach can be seen as combining the merits of minimax-optimality principle of online learning and robustness considerations of offline learning, but is motivated by the insight of train-test distribution alignment.

6 CONCLUSION

In this work, we explore an uncharted problem of constrained bidding in adversarial environments without knowledge about how adversarial factors perturb the environments. Previous methods on constrained bidding typically rely on the Emipirical Risk Minimization (ERM) principle which is violated in the adversarial setting. To address this limitation, we propose a Minimax Regret Optimization (MiRO) framework, which interleaves between a teacher identifying the aligning train distribution and a learner optimizing the policy under the given distribution of environments. To make the minimax problem tractable, we renders a differentiable game by variational learning the representation of adversarial factors by reconstructing the causal structure of the world model, and optimizes the differentiable game via dual gradient descent. In addition, we are the first to incorporate expert demonstrations for policy learning. We identify the unobserved confounding issue that fails the straight-forward idea of behavioral cloning from experts, and develop a causality-aware approach that aims to mimic the causal structure of the expert policy and distill knowledge from expert demonstrations. Empirical results on both large-scale industrial and synthetic dataset show that our method, MiROCL, outperforms prior methods more than 30%.

REFERENCES

[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.

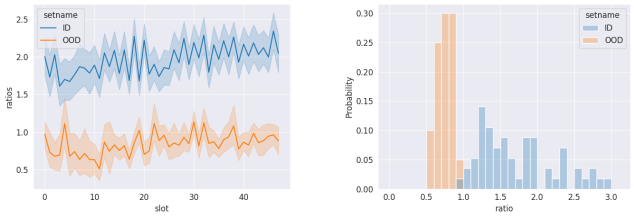

Figure 5: The distribution shift of the Industrial dataset. (Left) the average slot-wise expert policy on the IID and OOD set. (Right) The distribution of the bid ratio of a day-wise expert policy.

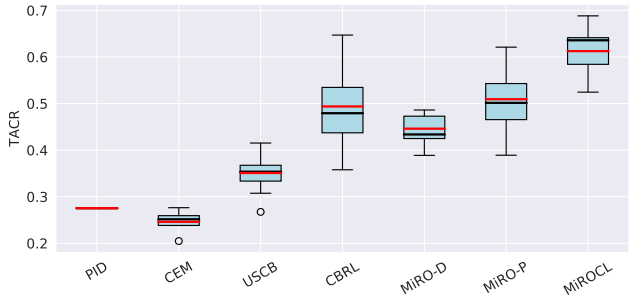

Figure 6: TACR result on the Industrial dataset. Each boxplot shows the median (red bar) and mean (black bar) scores over 20 random trials.

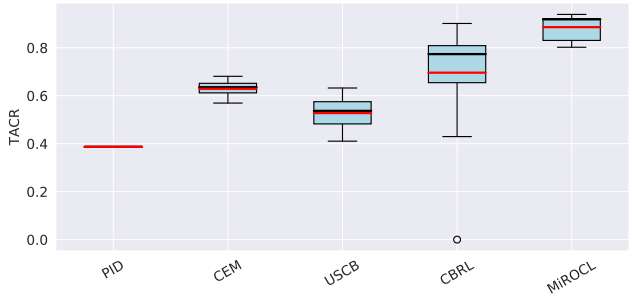

Figure 7: TACR result on the Synthetic dataset. Each boxplot shows the median (red bar) and mean (black bar) score over 20 random trials.

A DERIVATIONS

A.1 Derivation for the teacher's objective

To begin with, we aim to find a distribution of environments that maximizes the regret of the current policy, within the Wasserstein ball around the empirical distribution in the embedding space. For probability meassures P_1 and P_2 supported on the latent space $\mathcal{W}$, and the couplings $\Pi(P_1, P_2)$, the Wasserstain distance over the metric space $\mathcal{W}$ is defined as:

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

which finds the minimum cost to morph from the distribution P_1 to P_2, with the cost function $\kappa(\cdot, \cdot)$.

The distance capture the geometry of the space $\mathcal{W}$ with the cost function, and we assume that $\kappa(x, y) = \|x - y\|_2$, i.e., the distance between two distributions relates with the euclidean distance between the samples of the two distributions. We define a set $\mathcal{P}$ of distributions as the ρ-ball around the empirical distribution $\bar{P}$, from which we search for the adversarial environments:

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

The teacher's objective amounts to the following:

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

with Lagrangian relaxation parameter $\lambda \ge 0$.

Based on the assumption that $\mathrm{Reg}(\pi, \boldsymbol{\omega})$ and $\kappa(\cdot, \cdot)$ are continuous, we have the following dual re-formulation [38],

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

where the dual variable $\lambda \ge 0$. Then we have

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

A.2 Derivation for the variational bounds

We aim to show that the objective in Eq. (16) amounts to minimizing an upper bound on the regret. We first show how we derive the lower bound from the discrepancy between the expert and the policy in the KL divergence.

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

where

$$\begin{aligned} &\log \frac{\xi(a|o_t, h_t)}{\pi(a|o_t, h_t)} = \int_\omega q(\omega|h_t) \log \frac{\pi(a|o_t, h_t)}{\xi(a|o_t, h_t)} d\omega \\ &= \int_\omega q(\omega|h_t) \log \left(\frac{\xi(a|o_t, h_t, \omega)}{\pi(a|o_t, h_t)} \cdot \frac{p(\omega|h_t)}{p(\omega|h^*)}\right) d\omega \\ &= \mathbb{E}_{q(\omega|h_t)} \left[-\log \pi(a|o_t, h_t, \omega)\right] - \mathbb{E}_{q(\omega|h_t)} \left[-\log \xi(a|o_t, h_t, \omega)\right] \\ &+ \mathcal{D}_{KL}\left(q(\omega|h_t) \| p(\omega|h^*)\right) - \mathcal{D}_{KL}\left(q(\omega|h_t) \| p(\omega|h_t)\right) \end{aligned} \tag{24}$$

We can derive the bound of Eq. (7) similarly as the above shows.

It follows that the discrepancy can factorizes into a cross entropy term between the expert and the policy, a constant term on the entropy of the expert, and two KL terms between the posteriors.

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

We note that the first KL term aims to align with the causal structure of the expert. The second KL term is empirically found to perform worse, so we learn with only the first KL term, resulting in the upper bound (16).

Then we introduce a theorem from the imitation learning literature that states the relationship of regret and KL divergence-based expert-policy discrepancy [46].

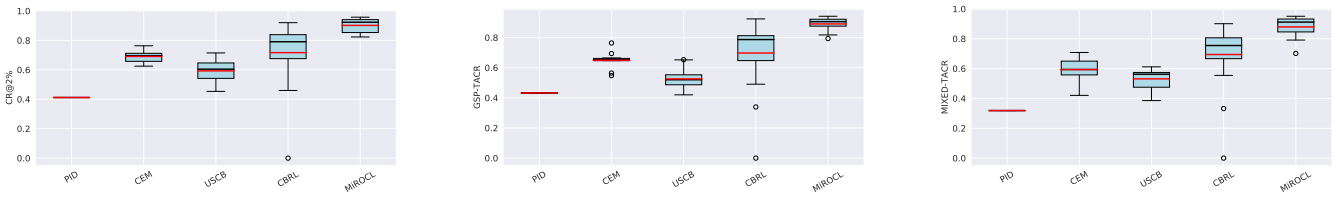

Figure 8: The results of CR@2% (Left), IID-TACR/GSP-TACR (Middle), and OOD-TACR/MIX-TACR (Right) in the Industrial (Top) and Synthetic (Bottom) settings are shown above. Each boxplot shows the average (red) and median (black) results of 20 independent repeated runs.

Theorem 2. *Under a given MDP* $\mathcal{M}_{\boldsymbol{\omega}}$, *if* $\mathbb{E}_{d_{\boldsymbol{\omega}}^{\xi^*}}\left[\mathcal{D}_{KL}\left(\xi^*, \pi\right)\right] \leq \epsilon$, *we have that* $Reg(\pi; \mathcal{M}_{\boldsymbol{\omega}}) \leq 2\sqrt{2}H^2\sqrt{\epsilon}$.

The theorem[1] states that the regret is bounded by the expert-policy discrepancy measure. Accordingly, minimizing the objective (16) is directly minimizing an upper bound on the regret, which can indicate regret bound guarantee.

B EXPERIMENTS

Dataset. Fig. 5 shows the distribution shift in the in-distribution and out-of-distribution split of **Industrial** dataset. For **Synthetic** dataset, We simulate the high and low periods of market competitions and ROI as sinusoidal functions with different amplitudes and phases. The dynamic mixed auction is constructed by sampling inflection points and then interpolating between the points.

Implementations The policy is implemented as a BERT transformer, which consists of an encoder for $q()$ and a decoder for $\pi_\phi()$. The ground-truth posterior $p()$ shares the encoder with $q()$ but is bi-directional, i.e., without future masking. We adopt entropy-regularized RL objective, which minimizes the KL divergence between the policy and the Boltzmann distribution of a state-action value function. The state-action value also takes as input the inferred ω_t, meanwhile learned with the reward returned by the off-distribution reward estimator $r_\theta()$.

Additional Results The TACR result on **Synthetic** dataset is shown in Fig. 7, and other metrics are shown in Fig. 8.

[1]The bound is under the infinite sample situation, and can be further bounded using classical learning theory.