

ROI-Constrained Bidding via Curriculum-Guided Bayesian Reinforcement Learning

Haozhe Wang
jasper.whz@outlook.com
ShanghaiTech University
China

Xuming He
ShanghaiTech University
China

Chao Du
duchao0726@gmail.com
Alibaba Group
China

Liang Wang
Alibaba Group
China

Panyan Fang
Shuo Yuan
Alibaba Group
China

Bo Zheng
Alibaba Group
China

ABSTRACT

Real-Time Bidding (RTB) is an important mechanism in modern online advertising systems. Advertisers employ bidding strategies in RTB to optimize their advertising effects subject to various financial requirements, especially the return-on-investment (ROI) constraint. ROIs change non-monotonically during the sequential bidding process, and often induce a see-saw effect between constraint satisfaction and objective optimization. While some existing approaches show promising results in static or mildly changing ad markets, they fail to generalize to highly dynamic ad markets with ROI constraints, due to their inability to adaptively balance constraints and objectives amidst non-stationarity and partial observability. In this work, we specialize in ROI-Constrained Bidding in non-stationary markets. Based on a Partially Observable Constrained Markov Decision Process, our method exploits an indicator-augmented reward function free of extra trade-off parameters and develops a *Curriculum-Guided Bayesian Reinforcement Learning (CBRL)* framework to adaptively control the constraint-objective trade-off in non-stationary ad markets. Extensive experiments on a large-scale industrial dataset with two problem settings reveal that CBRL generalizes well in both in-distribution and out-of-distribution data regimes, and enjoys superior learning efficiency and stability.

CCS CONCEPTS

- Information systems → Display advertising; • Theory of computation → Reinforcement learning.

KEYWORDS

Online Advertising, Reinforcement Learning, Bayesian Learning

ACM Reference Format:

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 (KDD '22), August 14–18, 2022, Washington, DC, USA*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3534678.3539211>



This work is licensed under a Creative Commons Attribution International 4.0 License.

KDD '22, August 14–18, 2022, Washington, DC, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9385-0/22/08.
<https://doi.org/10.1145/3534678.3539211>

1 INTRODUCTION

Online advertising [12, 40] has become an important business in the modern Internet ecosystem, connecting vast amounts of advertisers and users closely. Through Real-Time Bidding (RTB) systems [41], the online advertising markets manage to process a throughput of billions of ad impression opportunities, each triggering a bidding auction (Fig. 1). During the online sequential bidding process, the advertisers employ bidding strategies to optimize their advertising effects, subject to the budget constraint, and usually with return-on-investment (ROI) requirements. ROI, computed as the ratio of the value obtained to the price paid, is the standard metric to measure the immediate trade-off between (various types of) return and investment. In particular, ROI constraints are widely adopted by *performance advertisers* who concern about the effectiveness of resource used [36].

In recent years, extensive research has been conducted on constrained bidding. Most of these works focus on the budget-only setting [3, 6, 24, 37, 41, 42], and they cannot generalize to deal with the ROI constraints, due to the *non-monotonicity* of ROIs. ROIs can either increase or decrease over time during the sequential bidding process, in contrast to the budget that always decreases. Previous works on budget-constrained bidding derive pacing strategies (c.f. [3] for a survey) that terminate bidding upon depleted budget (c.f. [14]) or exploit the monotonicity of the budget in a Markov Decision Process (MDP) formulation [6, 37]. Neither of these approaches complies with the non-monotonic ROI constraints.

Moreover, ROI-constrained bidding usually witnesses a see-saw effect between constraint satisfaction and objective optimization, which urges the need to balance between constraints and objective. For example, return can increase with ROI decreasing when return and investment grows at different speeds (c.f. Sec. 2). Recent approaches [16, 32] handles the constraint-objective trade-off by *soft combination algorithms*, which introduce extra trade-off parameters to softly combine constraint violations and objective value in the objective functions. Despite their promising results, these methods assume static or mildly changing markets, which are limited in the more *non-stationary markets*. Such application scenarios are common when uncontrollable or unpredictable external forces affect the auction markets. For instance, the external online ad markets are prone to unknown adversaries that interfere with auction winning.

With ROI constraints and non-stationarity intertwined, ROI-Constrained Bidding (RCB) in the general ad markets is challenging. On the one hand, the optimal constraint-objective trade-off can

vary across different market dynamics. As such, soft combination algorithms that employ a static trade-off parameter design fail to adapt constraint-objective trade-off per dynamics, leading to non-responsive and unidentifiable bidding behaviors. In addition, the bidders are generally unobservable to other competing bidders in each auction, with market information leaking only conditionally (c.f. Sec.2). Consequently, such *partial observability* makes it even harder for the bidders to coordinate with the market dynamics.

To address these challenges, we specialize in the problem of ROI-Constrained Bidding (RCB) in non-stationary markets. Based on a Partially Observable Constrained Markov Decision Process (POCMDP) formulation for RCB, we introduce the first hard barrier solution to accommodate non-monotonic constraints(c.f. soft combination solutions). Our method employs the indicator function to render RCB an unconstrained problem, and develops a Curriculum-Guided Bayesian Reinforcement Learning (CBRL) framework to achieve adaptive control of constraint-objective trade-off.

Specifically, to avoid the pitfalls of soft combination algorithms in non-stationary markets, we introduce a reward function that incorporates the indicator function and involves *no extra trade-off parameters*. The indicator-augmented reward function explicitly encourages feasible solutions over infeasible ones by setting a hard barrier, which removes ambiguity in rewards (Sec. 3.2.2). However, as the reward function inherits the final-time sparsity from ROI that may hinder policy learning due to reward sparsity [29], we further develop a curriculum learning procedure to address inefficient policy search. By exploiting the problem structure, the curriculum learning arranges a sequence of proxy problems that provides immediate reward signals with an optimality guarantee, leading to faster convergence and better performance (Sec. 3.2.3).

While the parameter-free property of the hard barrier rewards is intriguing, the adaptive constraint-objective trade-off in partially observable markets amidst non-stationarity inherently lends to policy learning. To achieve this, we embrace a Bayesian approach. In particular, the agent learns to express its uncertainty about the market based on its past trajectory, by approximate inference of the posterior [1, 5, 22]. During deployment, the agent manages to infer the market dynamics, acts towards it, and updates the belief over the market with the latest experience, through an iterative process of posterior sampling [28, 30]. As such, the agent turns out a Bayes-optimal bidder that achieves the exploration-exploitation trade-off in unknown environments, meanwhile balancing the constraint-objective trade-off (Sec. 3.2.4).

We evaluate the proposed CBRL framework on a large-scale industrial dataset, including two problem settings for different advertisers. Extensive experiments verify our superiority over prior methods in both constraint satisfaction and objective maximization, and demonstrate several favorable properties regarding stability and out-of-distribution generalization. Our contributions are:

- We present the first hard barrier solution to deal with non-monotonic constraints, which achieves adaptive control of the constraint-objective trade-off in non-stationary advertising markets, and empirically found to reach a new state-of-the-art.
- We develop a novel curriculum-guided policy search process that promotes efficient policy learning against reward sparsity.
- We propose a Bayesian approach that learns adaptive bidding strategies in the partially observable non-stationary markets.

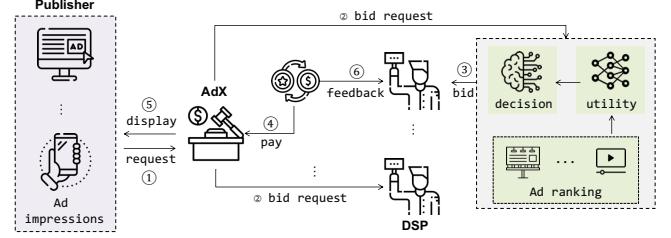


Figure 1: An RTB auction. Ad Exchanger (AdX) broadcasts the ad request to the advertisers. The ad system make decisions and sends the bid. AdX sends win notice, gets paid and diplay the ad. The winner gets delayed feedbacks.

2 PROBLEM STATEMENT

Real-Time Bidding (RTB) has become a prevailing advertising paradigm that allows the advertisers to make decisions for every ad impression opportunity [40, 41]. In RTB, billions of ad impression opportunities arrive sequentially, each triggering an auction. To relieve the advertisers' burden, demand-side platforms (DSPs) offer programmatic buying agents to the advertisers. On behalf of the advertisers, these bidding agents bid for each impression based on the ad context, in an attempt to optimize the hindsight total delivery under financial constraints.

Fig. 1 shows the workflow of each bidding auction. An ad exchanger (AdX) receives a bid request from the publisher when a user triggers an ad impression. AdX then holds an auction and broadcasts the request to all participating DSPs. The bidding agent of each DSP achieves real-time bidding by a modularized bidding engine. The bidding engine first selects an ad targeting the user based on superior personalization techniques, e.g., utility estimations and ad ranking, and decides the bid based on the valuations. Receiving all the bids, AdX announces the highest-bid bidder as the winner, and sends the winner's ad to the publisher for display. The winner pays a charge for the impression opportunity according to the regulated pricing mechanism (e.g., Generalized Second Price [12]), and receives delayed feedback from the publisher.

Auctions as such take place repeatedly within a period, forming a sequential bidding process for which the advertisers expect to optimize some desired delivery subject to certain constraints. Since RTB is a complex problem that additionally involves personalization techniques [9, 10] and auction mechanism design [12], in this work we focus only on the constrained bidding problem, and assume that utility estimations and mechanism design are given beforehand.

Particularly, in this work, we discuss a class of constrained bidding problems, *ROI-Constrained Bidding (RCB)*, which is a major concern of various advertisers. Formally, suppose a bidder observes a bid request x_i (features about the impression context, and the selected ad) for impression i . A bid b_i is decided based on the bidder's estimated utility u_i for the impression. If the bid b_i is larger than the competing market price m_i (i.e., the highest bid of the competing bidders), the bidder wins, pays a cost c_i , and receives delayed feedback about the delivery d_i (u_i estimates d_i). The RCB problem aims to maximize the total delivery subject to a budget B and a *return-on-investment (ROI)* constraint limit L , within T impressions.

$$\max_b D(\epsilon_T), \quad \text{s.t.} \quad \text{ROI}(\epsilon_T) \geq L, \quad B - C(\epsilon_T) \geq 0 \quad (1)$$

where we denote $\epsilon_t = \{(b_i, m_i, c_i, u_i, d_i)\}_{i=1}^t$ as a t -step episode containing t impressions, and we introduce the following notations to denote the cumulative delivery, cost, and ROI of an episode ϵ_t ,

$$D(\epsilon_t) \stackrel{\text{def}}{=} \sum_{i=1}^t d_i \mathbb{1}_{b_i > m_i}, C(\epsilon_t) \stackrel{\text{def}}{=} \sum_{i=1}^t c_i \mathbb{1}_{b_i > m_i}, \text{ROI}(\epsilon_t) \stackrel{\text{def}}{=} \frac{D(\epsilon_t)}{C(\epsilon_t)} \quad (2)$$

and use the short-hands $D(\epsilon_t) \equiv D_t$, $C(\epsilon_t) \equiv C_t$, and $\text{ROI}(\epsilon_t) \equiv \text{ROI}_t$ if no misconception may arise.

It is noteworthy that, many widely adopted cost-related key performance indicator (KPI) constraints are viewed as a type of ROI constraint. For example, upper-bounded cost per acquisition (CPA) is equivalent to a lower-bounded per-cost acquisition in the context of ROI constraints. Besides, in this work, we treat *delivery* as equal to *return* for simplicity, which is often the case but exceptions do exist, e.g., profit maximization with per-cost income constraints¹. **Challenges of RCB and Related Work.** While the constrained optimization problem (1) appears simply a programming problem, real-world RCB is challenging due to the properties of ROIs, the properties of online ad markets, and both properties intertwined.

Eq. (2) shows that ROI can either increase or decrease during the bidding process, since both D_t and C_t increase at an uncertain rate. Besides non-monotonicity, a see-saw effect often emerges between constraint violations and the delivery value, esp. when the delivery grows with the investment at a different speed. For example, revenue increases as the cost grow, but the per-cost revenue (ROI) may plunge, inducing the demands of constraint-objective trade-off.

Most existing works on constrained bidding focus on the budget-only setting (c.f. [3] for a survey), they cannot generalize to deal with ROI constraints. Based on the primal-dual framework, many works derive pacing strategies that terminate bidding when the budget runs out (c.f. [14]). Alternatively, Reinforcement Learning (RL) formulations have been proposed [6, 37], which encode the budget status in action space or state space.

In contrast to these work that exploit the monotonicity of budget, some works propose to deal with specific non-monotonic constraints [14, 35, 39] or general constraints [16, 32]. Among them, a promising solution [16, 32] adopts a *soft combination* design that softly combines the constraint violations and the delivery value in the objective function with extra trade-off parameters, theoretically grounded by Lagrangian relaxation to achieve a balanced constraint-objective trade-off. These works, however, are typically established in controlled markets, where market dynamics change smoothly as each ad campaign binds to similar types of impressions, and full access to market information can be gained².

By contrast, the external online advertising markets experience more drastic market changes, due to unexpected adversaries and system failures. In addition, partial observability of the market aggravates the difficulty of bidding amidst non-stationarity. The bidders can be observable to the market conditionally, or completely unobservable. In particular, under the second-price auctions [12], the market price m_i equals the cost c_i when the auction is won ($b_i > m_i$), which leaks hindsight information of the market.

¹We discuss a generalized version of RCB in our recent work.

²When the publisher offers programmatic advertising services, e.g., in-station advertising of e-Commerce platforms, market information of all bidders is logged.

In such partially observable markets amidst non-stationarity, adaptive control of the constraint-objective trade-off is hard to achieve, as the optimal trade-off varies across dynamics which is in turn unidentifiable. Previous soft combination solutions rely on static trade-off parameters, logically and empirically found to fail in non-stationary advertising markets (Sec. 4.2). To this end, in this work, we shed light on an alternative hard barrier solution to accommodate non-monotonic constraints, which learns adaptive bidding strategies per dynamics.

3 METHOD

3.1 MDP Formulation

Markov Decision Processes (MDPs) provide a mathematical framework widely used to learn from interaction with the environment [31]. To account for the constraints and the partial observability typical of RCB, we formulate a *Partially Observable Constrained MDP (POCMDP)*, which is finite with T steps, represented as a nine-tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{E}, \mathcal{T}, \mu, \gamma, \mathcal{R}, \mathcal{C})$:

- \mathcal{S} . The state space reflects critical information for each impression. Each state $s_i \in \mathcal{S}$ includes impression-level information $(x_i, d_i, c_i, m_i, u_i)$ and the cumulative statistics $(D_{i-1}, C_{i-1}, \text{ROI}_{i-1}, B, L)$ ³.
- \mathcal{A} . The action $a_i \in \mathcal{A}$ is a real-valued scalar $b_i \in \mathbb{R}^+$.
- \mathcal{O} and \mathcal{E} . The observation space accounts for the partial observability. The emission function $\mathcal{E} : \mathcal{S} \mapsto \mathcal{O}$ maps a state s_i to an observation o_i by removing (d_i, c_i, m_i) .
- \mathcal{T} and μ . Both symbols determine the market dynamics. The transition probability density $\mathcal{T}(s_{i+1}|s_i, a_i)$: (1) explains the transitions between the cumulative statistics, e.g., $D_i = D_{i-1} + d_i \mathbb{1}_{b_i > m_i}$, which are markovian; and (2) induces stochasticity from the market dynamics, i.e., $P(x|s_i), P(d, c, m|x_i)$, which are time-varying. The initial state distribution $\mu(s_1)$ can be arbitrary. While we have no access to the exact form of \mathcal{T}, μ , we approximate these with the empirical distribution using logged dataset.
- γ . The discount factor.
- \mathcal{R} and \mathcal{C} . We define the following reward function and cost function to account for the performance objective and constraint requirements respectively.

$$\mathcal{R}(s_i, a_i) = (D_i - D^-) \mathbb{1}_{i=T} \quad (3)$$

$$\mathcal{C}(s_i, a_i) = \left((L - \text{ROI}_i) \mathbb{1}_{\bar{F}_L} + (C_i - B) \mathbb{1}_{\bar{F}_B} \right) \cdot \mathbb{1}_{i=T} \quad (4)$$

We note the above functions are piece-wise functions that only evaluate at termination. We define $D^- \stackrel{\text{def}}{=} \inf D_T$ so that $\mathcal{R}(s_T, a_T) > 0$. To simplify the notation of feasibility, we use $F_L(\epsilon_t) \stackrel{\text{def}}{=} \{\epsilon_t \mid \text{ROI}(\epsilon_t) \geq L\}$, $F_B(\epsilon_t) \stackrel{\text{def}}{=} \{\epsilon_t \mid C(\epsilon_t) \leq B\}$, $F(\epsilon_t) \stackrel{\text{def}}{=} F_L(\epsilon_t) \cap F_B(\epsilon_t)$ to indicate the feasible solution sets that respect the ROI constraint, the budget constraint and both constraints. By convention, $\bar{F}_L, \bar{F}_B, \bar{F}$ are their negations.

The RL objective for the above MDP is:

$$\max_{\pi} \mathbb{E} \left[\sum_{t=1}^T \mathcal{R}(s_t, a_t) \right], \text{ s.t. } \mathbb{E} \left[\sum_{t=1}^T \mathcal{C}(s_t, a_t) \right] \leq 0 \quad (5)$$

We remark that both the objective value and the constraints are studied in expectation. The expectation is taken over different problem instances, which coincides with the fact that real-world

³These statistics cumulates up to the last impression $i-1$ because the feedback of current impression i are received only after the bid.

advertisers run various ad campaigns or consider advertising effects over different time periods. As we encode the stochasticity and variations of the market dynamics in (\mathcal{T}, μ) , the RL objective aligns with RCB on an aggregate level. Moreover, cost function (4) is a subtle yet noteworthy design that doesn't violate constraint satisfaction in the expected sense⁴.

3.2 Curriculum-Guided Bayesian Reinforcement Learning

In this section, we present a *Curriculum-Guided Bayesian Reinforcement Learning (CBRL)* framework (Fig. 2) to solve the MDP. Specifically, we tackle the long sequence with a slot-wise policy design (Sec. 3.2.1), accommodate the constraints with a parameter-free hard barrier reward function (Sec. 3.2.2), promote efficient policy learning by curriculum-guided policy search (Sec. 3.2.3) and achieves adaptive control of constraint-objective trade-off in non-stationary markets with Bayesian RL (Sec. 3.2.4).

3.2.1 A semi impression-level policy to deal with long sequences. While bidding agents aim to bid in real-time for each impression, real-world advertising systems experience a throughput of billions of impressions. This brings about extremely long trajectories for RL that incurs training issues. The optimal bidding theorem comes to the rescue, providing a proxy for converting impression-level bidding to a ratio controlling problem.

THEOREM 3.1. *In second-price auctions, the optimal bidding function for problem (1) is:*

$$b_i = \beta u_i, \quad \beta > 0 \quad (6)$$

The theorem (proof in the appendix) states that the optimal bid for each impression is linear w.r.t. the impression value u_i , governed by a bid ratio β computed given the problem data. Intuitively, the bid ratio β acts like a threshold in the knapsack problem where items are filtered by their value-to-weight ratios [3].

Based on Thm. 3.1, we reduce the online RCB problem into a ratio controlling problem in which we seek to find optimal bid ratios for binned time slots. The reduction makes policy trajectories tractable to train with, though at the cost of precisely impression-level control. Specifically, we set the slot length to the unit time period of dynamics changes (empirically thirty minutes). Indexing each time slot by t , we add an auxiliary action space $a_t \equiv \beta_t \in \mathcal{A}'$ for the slot-wise bid ratios, and an observation space $o_t \in \mathcal{O}'$, which replaces the impression-level information with aggregate-level statistics, e.g., the slot-wise delivery $D_t - D_{t-1}$ (more details in the appendix). We note that this slot-wise policy can still produce impression-level bids with the impression-level utilties:

$$b_t = \beta_t u_i, \quad \text{where } \beta_t \sim \pi(\cdot | o_t) \quad (7)$$

3.2.2 An indicator-augmented reward function to accommodate constraints. Our goal is to develop a parameter-free hard barrier solution to address problem (1). The main idea is to convert the constrained optimization problem to an unconstrained counterpart, which derives a parameter-free reward function that still satisfies Bellman Equation to fit in with conventional policy optimization.

⁴Non-negative entries less than 0 in expectation means each entry must evaluate 0.

Specifically, we note that the RL objective (5) has an equivalent unconstrained form $\max_{\pi} \mathbb{E} [\sum_{t=1}^T \mathcal{R}'(s_t, a_t)]$, with an indicator-augmented reward function defined as follows,

$$\mathcal{R}'(s_t, a_t) = \left(\mathbf{1}_F \sum_{t=1}^T \mathcal{R}(s_t, a_t) - \mathbf{1}_{\bar{F}} \sum_{t=1}^T C(s_t, a_t) \right) \cdot \mathbf{1}_{t=T}. \quad (8)$$

The equivalence holds because $\sum_{t=1}^T \mathcal{R}(s_t, a_t) > 0 \geq -\sum_{t=1}^T C(s_t, a_t)$ strictly holds. Another critical property of Eq. (8) is it satisfies the recursive property of Bellman Equation and thus works with conventional RL algorithms.

While this reward function design appears simple, it renders a parameter-free solution to accommodate constraints, by setting a hard barrier between feasible and infeasible solutions. The underlying philosophy is that, we designate feasibility to weigh heavier than infeasibility in rewards, instead of softly combining the constraint violations and delivery value as adopted in soft combination algorithms [16, 32]. Soft combination solutions are ambiguous in reflecting the quality of different solutions, especially when the trade-off parameters are inappropriate. For example, an infeasible solution with inadequate penalties for constraint violations would be favored over a feasible solution. Another drawback of soft combination algorithms is that static trade-off parameters may be inapt for dynamic markets (Sec. 4.2).

3.2.3 Curriculum-guided policy search to promote efficient learning. While the reward function (8) enjoys several favorable properties, we note that its reward sparsity hinders efficient policy learning. In particular, the agent is only guided with reward signals at termination, leading to inefficient and blinded exploration in the policy search space [29]. To resolve this, our goal is to provide the agent with immediate rewards to guide policy learning, while preserving the optimality of convergence. With proxy problems to constrained problem (1) that provide dense signals, we arrange a sequence of proxies into a curriculum learning procedure that regularizes policy search and guides the policy towards optimality.

We begin by defining the following problem $P_k (k > 0)$ as a proxy to problem (1), denoted as P_0 .

$$\max_b D_T \quad \text{s.t.} \quad \text{ROI}_t \geq L_t^k, B - C_t \geq B_t^k, \forall t \in \{1, \dots, T\}, \quad (9)$$

where we add $T - 1$ constraints in each time slot, except for T -th slot where $L_T^k = L, B_T^k = 0$. These additional constraints admit the potential for immediate rewards.

To arrange a meaningful curriculum sequence which starts with easier problems and converges to the original problem (1), we consider two questions. How do we exploit the dense constraints to provide dense signals? And how do we arrange the curricula?

For the first question, we adopt a greedy approximation that derives a dense reward function. Treat $P_k (k > 0)$ as a recursive stack of sub-problems in each time slot t , we define recursively for each time slot:

$$\begin{aligned} \max_b D_t &= D_{t-1}^* + D_{t-1:t} \\ \text{s.t.} \quad \text{ROI}_t &\geq L_t^k, \quad B - C_t \geq B_t^k, \forall t \in \{1, \dots, T\}. \end{aligned} \quad (10)$$

The objective of the above recursive sub-problem is to greedily maximize $D_{t-1:t}$ under the added slot-wise constraints, based on D_{t-1}^* obtained in the previous slot. As a result, this recursive structure

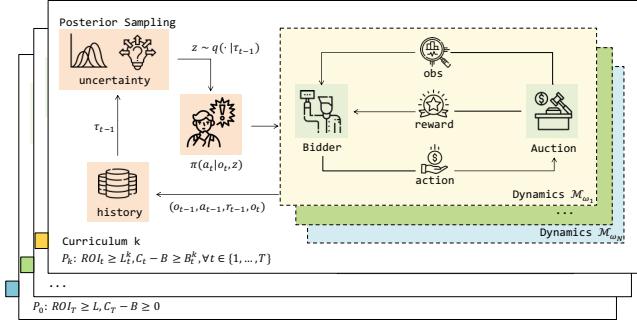


Figure 2: Model Overview. We propose CBRL for RCB. Guided by a curriculum sequence, the agent learns to adaptively trade-off constraints and the objective in non-stationary markets. During deployment, the agent updates its belief over the market dynamics based on its past experience, and acts through posterior sampling.

offers a *dense* reward function:

$$\mathcal{R}'_k(s_t, a_t) = D_{t-1:t} \mathbb{1}_F - \left(L_t^k - \text{ROI}_t \right) \mathbb{1}_{F_{L_t^k}} - \left(B_t^k + C_t - B \right) \mathbb{1}_{F_{B_t^k}}, \quad (11)$$

which credits the agent with the slot-wise delivery if cumulative constraints are satisfied, and penalizes the agent with constraint violations if otherwise.

To arrange the curricula into meaningful sequence, the idea is to evolve the constraints of the proxies from tightened ones to loosened ones, until approaching P_0 . As we deploy each curriculum as a dense reward function, the curriculum sequence $\{P_n, \dots, P_k, \dots, P_0\}$ ($n > k$) is implemented as a dense reward function with constraint limits $\{L_t^k\}_{t=1, k=1}^{T,n}, \{B_t^k\}_{t=1, k=1}^{T,n}$ evolving according to the following principles: (1) The limits increase as k increases, so tighter constraints are levied on earlier curricula; (2) $\lim_{t \rightarrow T} L_t^k = L$ and $\lim_{t \rightarrow T} B_t^k = 0$, so that the constraint limits approach the final constraint as time goes. The specific design of the constraint limits can be found in the appendix.

Curriculum learning of this kind starts from proxy problem P_n , with tightened constraints that strongly narrow the policy search space. However, these strong signals might bias toward sub-optimal behaviors, so we proceed with problems with loosened constraints, and finally, approach problem P_0 . The curriculum-guided policy search constructs recursive sub-problems of RCB, and promotes faster convergence by dense reward signals than direct optimization with sparse rewards. Although logically curriculum learning requires multiple stages of training, policy learns efficiently in each curriculum so that the overall training time is less than policy learning with sparse rewards (Sec. 4.3). To relieve hand-tuning of the curriculum settings, we develop an automated curriculum learning process driven by differentiable regret minimization, as detailed in the appendix.

3.2.4 Bayesian reinforcement learning to act optimally amidst non-stationarity and partial observability. In uncontrollable markets, the bidder has no access to the complete auction market information (i.e., (d_i, c_i, m_i)), and the market dynamically changes due to unknown time-varying parameters (i.e., $P(d, c, m | \omega_t)$). We summarize these factors as partial observability in the POCMDP formulation.

Adapting the constraint-objective trade-off per dynamics is challenging. To resolve this, we adopt a Bayesian perspective [11, 13]. In tabular POMDPs, policies that perform posterior sampling [28, 30] given the belief over MDPs have proven Bayes-optimal [11, 25], meaning to balance exploration and exploitation in an unknown environment. Motivated by this, we aim to first infer the posterior of the unobservable market dynamics via Variational Bayes [22] and then act adaptively through posterior sampling.

Specifically, we adopt a latent representation z for the unobservability. Our goal is to learn a variational distribution $q(z | \tau_t)$ that allows the agent to approximately infer the posterior $P(z)$, based on the historic trajectory $\tau_t = \{(o_i, a_i, o_{i+1})\}_{i=1}^t$. In Variational Bayes, the variational distribution is typically learned with an *Evidence Lower Bound (ELBO)* [22]. Recall that the Q-learning objective is:

$$\mathbb{E}_{c_t \sim \mathcal{B}} \left[\left(Q(o_t, a_t) - (r_t + \gamma \max_a Q(o_{t+1}, a)) \right)^2 \right] \quad (12)$$

where \mathcal{B} denotes a replay buffer [27], and $c_t \stackrel{\text{def}}{=} (o_t, a_t, o_{t+1}, r_t)$.

Minimizing Bellman Residual can be interpreted as maximizing the log likelihood of the transition tuple (o_t, a_t, o_{t+1}, r_t) in a model-free manner [34], and thus we maximize the ELBO of Eq. (12) to learn the inference network $q(z | \tau_t)$ (derivations in the appendix).

$$\mathbb{E} [-\mathbb{E}_z [(Q(o_t, z, a_t) - y_t)^2]] - D_{KL}(q(\cdot | \tau_{t-1}) || P(z)) \quad (13)$$

where $z \sim q(\cdot | \tau_{t-1})$ and $c_t, \tau_{t-1} \sim \mathcal{B}$ is omitted, the target value⁵ computes as $y_t \stackrel{\text{def}}{=} r_t + \gamma \max_a Q(o_{t+1}, z_t, a)$. $P(z)$ is set to $\mathcal{N}(0, 1)$.

The learned Gaussian distribution $q(z | \tau_t)$ expresses our uncertainty about the current dynamics based on our experience, and is implemented as a neural network. Since $q(\cdot | \tau_t)$ is iteratively updated along the trajectory τ_t up to step t , its specific network structure should input variable number of transition tuples $\{c_i\}_{i=1}^t$ and effectively exploit the relationships between the tuples $\{c_i\}_{i=1}^t$ that are inherently correlated [34]. We adopt the transformer blocks [33] followed with average pooling.

$$q(z | \tau_t) = \mathcal{N} (\text{AvgPool}(f^\mu(\tau_t)), \text{AvgPool}(f^\sigma(\tau_t))) \quad (14)$$

where f^μ, f^σ represents a three-layer transformer for the mean and standard deviation distribution parameter.

Equipped with the approximate posterior, the agent performs posterior sampling [28, 30] to act in the unknown environment during deployment. Posterior sampling begins with a hypothesis z_{t-1} sampled from $q(z | \tau_{t-1})$, based on past experience τ_{t-1} . The agent $\pi(a_t | z_{t-1}, o_t)$ acts towards the hypothesized MDP characterized as z_t , and collects a new transition tuple c_t . The tuple is used to update the belief $q(z | \tau_t)$, which reflects the current uncertainty of the market given the agent's past experience. Intuitively, this iterative process allows the agent to test its hypothesis in a temporally extended exploration process, leading to Bayes-optimal exploration-exploitation trade-off in the POCMDP [11, 25].

Recall that the reward function (8) induces a parameter-free form that inherently lends the commonly explicit constraint-objective trade-off to policy learning. The Bayesian bidder complements the reward function by learning to trade-off according to the market dynamics. To this end, the proposed hard barrier solution bears

⁵The target value can be computed by other formulas according to the algorithm used.

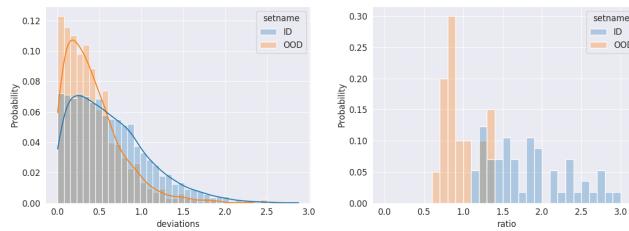


Figure 3: Dataset. The distribution of the slot-wise bid ratio deviations (Left) reveals the high non-stationarity of the marketplace. The distribution of the day-wise bid ratio (Right) reveals the distributional shift between ID and OOD split.

no extra parameters for hand-tuning (c.f. USCB [16]), and performs adaptive control in non-stationary ad markets even with out-of-distribution (OOD) data (Sec. 4.4). The insight is that our agent learns to infer the posterior of market transitions in the in-distribution data, which overlaps with that of the OOD data, despite the data distribution being different.

4 EXPERIMENTS

In this work, we formulate RCB in non-stationary markets as a Partially Observable Constrained MDP (POCMDP) with an *indicator-augmented reward function*, and present a *Curriculum-Guided Bayesian Reinforcement Learning (CBRL)* framework. Therefore, in the experiments, we aim to: (1) verify the superiority of the proposed CBRL, esp. as the first hard barrier approach to handle constraints; (2) validate the learning efficiency of the proposed curriculum learning; and (3) reveal the effects of Bayesian learning on adaptive control of the constraint-objective trade-off. We begin with a brief introduction of the evaluation setups and then detail our findings. *Implementation details are left to the appendix or check out the code.*

4.1 Experimental Setup

Dataset. We use a large-scale logged dataset from the Alibaba display advertising platform for evaluation. The dataset includes 80-day winning bidding logs, with each day two million impressions on average. The daily logs are deliberately selected so that each day represents different market dynamics and is treated as an independent problem instance. The dataset is partitioned into two splits, containing the 60 days and 20 days respectively. The data distribution of the two splits is distinct from each other, with market dynamics changes caused by e-Commerce festivals, adversaries, and ad system changes (Fig. 3). Consequently, the 60-day split is termed *in-distribution (ID)*, while the 20-day split is termed *out-of-distribution (OOD)*. We i.i.d sample from the **ID** split to construct our **ID** train set and the **ID** test set with 30 days each, which follows the conventional machine learning assumptions⁶. In addition, the **OOD** split serves as an extra test set for OOD generalization, and reflects the performance of online serving after extended periods. **Problem settings.** We construct two problem settings for the demands of platforms and performance advertisers. In the single constraint (**SC**) setting, each problem instance has a single and fixed ROI constraint $L = 1$, with no budget constraint. This mimics

⁶Previous works [16, 37, 41] typically use week-wise logged data for training and one consecutive day for testing. We construct our benchmark different from the previous practice to obtain evaluation results that are more statistically reliable.

the platform’s demands to optimize social welfare in a holistic view regardless of individual budgets. In the multiple constraints (**MC**) setting, each problem instance is subject to different budget and ROI requirements, as performance advertisers expect to optimize the effects of different ad campaigns over different time periods, with different budgets and performance efficiency requirements.

Evaluation protocols. In the experiment, budget constraints are satisfied surely by an early termination [16, 37], which terminates the bidding process if the budget runs out. For RCB, we introduce the following three metrics to account for overall performance, constraint satisfaction, and objective optimization⁷: (1) *Average Normalized Score (ANS)* computes the daily score $D_T \mathbb{1}_F$ normalized by the oracle value D_T^* and takes average; (2) *Constraint Satisfaction Rate (CSR)* counts the ratio of feasible solutions to all problems; and (3) *Average Normalized Delivery Regret (ANDR)* normalizes the delivery regret (the difference of actual delivery and best possible delivery) by the oracle delivery D^* on condition of feasibility, and takes average. The formulas are included in the appendix.

4.2 Empirical Results

Competing Methods. We compare recent methods that can (be adapted to) deal with non-monotonic constraints, categorized into three lines of works: (1) primal-dual algorithms **RM** ([35]); (2) slot-wise approximation methods **PID** ([39]) and **CEM** [19]; (3) soft combination (RL-based) algorithms **RCPO** ([32]) and **USCB** ([16]). In addition, we include a baseline model **Hard** (same as **Ind-Sp** in the following passages), as the hard counterpart to the soft combination method **RCPO**. We provide a brief introduction for each method along with the analysis of the results below.

Results. The evaluation results are shown in Fig. 4. On both settings, **CBRL** achieves the best overall performance with stable performance (narrowest box), and performs no worse than other competing methods regarding constraint satisfaction and objective maximization. Specifically, in the **SC** setting, **CBRL** reaches 78.9% of the oracle (median ANS 0.789), satisfies the constraints almost surely (median CSR 1), with feasible solutions falling behind oracle by 20.8% (median ANDR 0.792)⁸. Similar results are obtained in **MC**, with median ANS 0.789, median CSR 1.0, and median ANDR -21.5%. The results of competing methods are analyzed as follows.

- **RCPO** (2018) is a general CMDP policy optimization algorithm based on Lagrangian relaxation, which introduces extra *learning parameters* to control the constraint-objective trade-off. We adapt RCPO to RCB by the proposed POCMDP formulation. While its maximal performance reaches **CBRL**, its average performance is hampered by high variance. The instability is due to (1) the sensitive initializations of the Lagrangian multipliers for stochastic optimization, especially in the non-convex condition; and (2) the soft combination of constraint violations and delivery with improper weights that leads to ambiguity in rewards. Particularly, we notice that the ambiguity makes instability combinatorially more likely in **MC** than in **SC** (wider box in **MC** than in **SC**).

⁷We use the metrics different from USCB [16] as theirs depend on an extra hyper-parameter. We include results using their metric in the appendix.

⁸Note that we use slot-wise oracle policy which reaches higher performance upper bound than in previous studies [16]. Check the appendix for details.

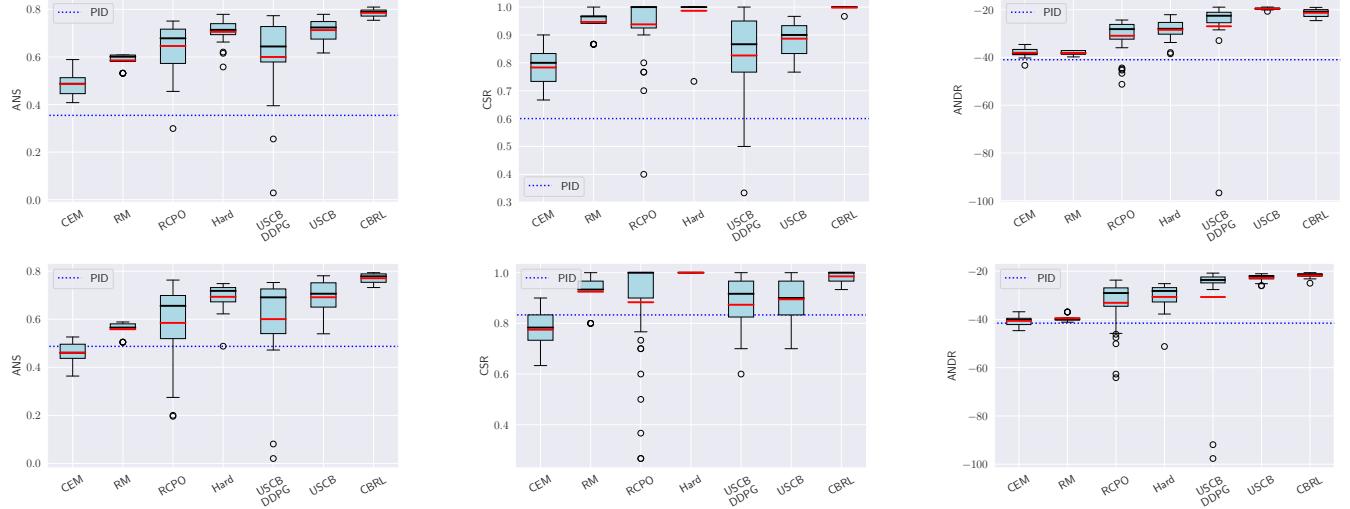


Figure 4: Evaluation results on the ID split. The results of ANS (Left), CSR (Middle), and ANDR (Right) in the SC (Top) and MC (Bottom) settings are shown above. Each boxplot shows the average (red) and median (black) results of 20 independent repeated runs.

- **USCB** (2021) is the prior art that formulates an RL framework and uses extra *hyper-parameters* to non-linearly control the constraint-objective trade-off, which can be treated as a variant of Lagrangian relaxation. For a fair comparison, our method **CBRL** and **USCB** use the same input features and network structures. We provide two specific implementations, **USCB** and **USCB-DDPG**. **USCB** aligns with **CBRL** in entropy regularization [15] and independent action space, while **USCB-DDPG** respects the plain Monte Carlo estimation based actor-critic approach and the temporally correlated action space in [16] (check the appendix). It follows that **USCB** tends to be more stable than **RCPO** (narrower box), as the instability caused by non-convex optimization is alleviated. However **USCB** still suffers from the reward ambiguity. In particular, we remark that while the best **USCB** model (top-rating in ANS) exhibits the best ANDR performance (at the cost of constraint satisfaction), its trade-off parameter design indeed shows a significant see-saw effect, and requires laborious tuning. By contrast, **CBRL** adopts a parameter-free solution, which is user-friendly and turns out best-performing in ANS.
- **CEM** [19] Cross-Entropy Method is a gradient-free stochastic optimization method. Widely used in the industry, CEM attempts to optimize a greedy sub-problem in each time slot and bears the exploration-exploitation trade-off. Since winning is sparse in the data, more exploration is required to obtain a more accurate estimate, which squeezes the space for exploitation. Consequently, the best CEM model achieves decent constraint satisfaction (around 0.8 in CSR) but lower objective optimization, due to the averaging effect of the dominant exploration traffic.
- **PID** (2019) adopts a PID control solution to bidding with CPC constraint and budget constraint. Based on the optimal bidding function (3.1), we adapt **PID** to control the bid ratio that drives the ROI constraint toward the target in each time slot. We note that **PID** itself does not handle changing systems well, and the online adjustment of **PID** parameters to suit the changed systems is non-trivial and beyond the scope of this paper. We find empirically that **PID** cannot balance constraint-objective trade-off well in

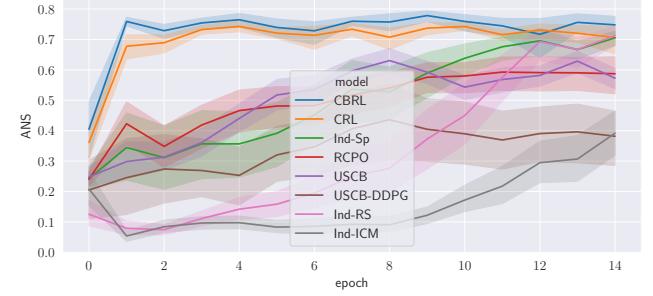


Figure 5: The learning curve on ID train set. Confidence intervals are computed using 20 random trials.

highly non-stationary markets presented in our dataset, and hence the best **PID** model with the best ANS score shows the only moderate status of both constraints satisfaction (CSR) and objective optimization (ANDR).

- **RM** (2019) propose to deal with RCB under a static functional optimization framework, which solves the optimal bid ratio over the train set and applies to the test set. The solved bid ratio achieves the best performance on the training problems on average, but does not adapt to each of them. As a result, in the ID test set, the **RM** model performs far from optimal, although it respects the constraints well (CSR close to 1).

4.3 Effects of Curriculum Learning

Reward sparsity is notorious for inefficient policy learning due to blinded exploration. We propose a curriculum-guided policy search process and use a three-stage curriculum sequence in the experiments, including one cold-start curriculum for three epochs, and one warm-up curriculum for three epochs, followed by the original problem. To verify its effectiveness in promoting efficient learning, we compare with the following baseline models: (1) **Ind-Sp** uses the sparse reward function Eq. (8); (2) **Ind-RS** uses the reward shaping [26, 38] technique, which introduces extra reward signals in each slot based on human expertise; (3) **Ind-ICM** applies

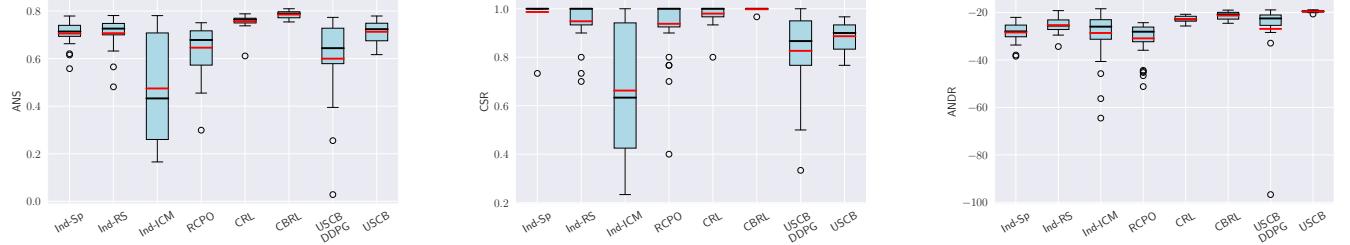


Figure 6: Ablations. ANS, CSR, ANDR (Left to Right) results of ablative models evaluated in SC .

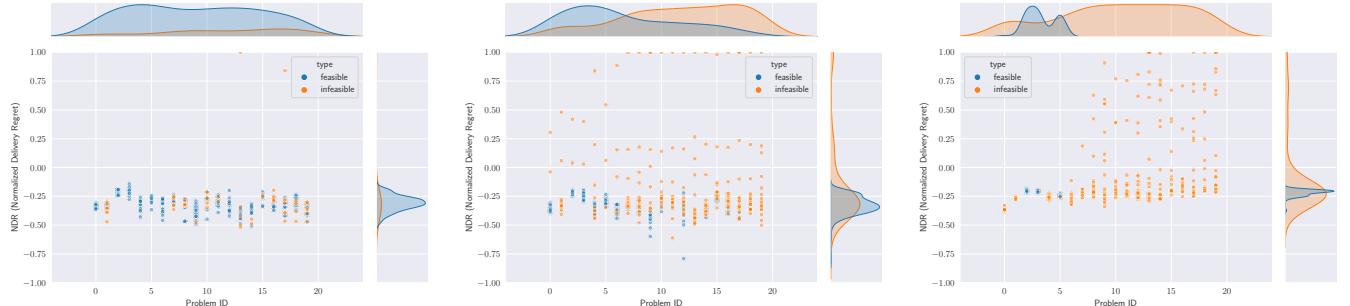


Figure 7: The Regret-PID joint distribution with SC in OOD split. Each column for CBRL, CRL, USCB (Left to Right). Each subplot contains a scatter plot for DIO and days with feasible (Blue) and infeasible (Orange) solutions, and contains the KDE plot for the marginal of regret (Right-most) and the solution distribution over problem IDs (Top-most).

the Intrinsic Curiosity Module (ICM) [29]. Intuitively, ICM uses reconstruction errors as extra reward signals.

Fig. 6 shows that, on the **ID** test set, the proposed curriculum learning baseline **CRL** outperforms reward shaping and ICM, both of which even perform worse than the sparse reward baseline. Learning curves of the baseline models are shown in Fig. 5, indicating **CRL** achieves high performance within three epochs (the first curriculum), which already beats the prior art **USCB**. It can be concluded that curriculum learning (**CRL**) improves training efficiency from sparse rewards (**Ind-Sp**), while the extra signals provided by reward shaping and ICM appear noisy to the policy at the beginning (downward curve) and afterward cause instability during learning (high error band).

4.4 Effects of Bayesian Learning

In the proposed hard barrier solution, Bayesian learning takes the responsibility to adaptively control the constraint-objective trade-off according to market dynamics. Here we examine two scenarios, in-distribution (**ID**) and out-of-distribution (**OOD**) data regimes.

Fig. 6 shows, **CBRL** improves **CRL** in **ID** split (median ANS from 0.764 to 0.789). More significant improvement is witnessed in the challenging **OOD** split (median ANS from 0.24 to 0.54, see the appendix). In Fig. 7, the feasible solution distributions (top KDE plots) show **CBRL** achieves better constraint satisfaction than **CRL** (median CSR 0.775 v.s. 0.35), while **USCB** fails dramatically (median CSR 0.15). Among the feasible solutions, the regret distribution (right-most KDE plots) show both **CBRL** and **CRL** deteriorates by 10% in regret from the **ID** split (median ANDR -30.18% v.s. -32.93%). As **CBRL** is not specifically designed for **OOD**, such performance is sub-optimal yet reasonable. Moreover, the regret does not increase much while respecting constraints, indicating **CBRL** achieves good constraint satisfaction not through a (too) conservative policy, but

through market adaptability. The scatter plots show the specific failure patterns in **OOD** split. We remark that, compared with **CRL**, **CBRL** rarely presents orange points high above, which earns excessive delivery (regret of feasible solutions are below 0) at the cost of severely violated constraints.

The good performance in **OOD** scenario is credited to Bayesian learning. Logically, although the joint data distribution is different between **ID** and **OOD** split, the distribution of market variations can overlap. It follows that the variational distribution $q(z)$ may generalize (partially) to **OOD**. Since the agent expresses its uncertainty about the market by $q(z)$, and eliminates the uncertainty through the iterative process of posterior sampling, correctly inferred posterior $q(z)$ empowers the agent to achieve adaptive control of the constraint-objective trade-off in the unseen environment.

5 RELATED WORK

Reward Sparsity. Learning with sparse reward is challenging since RL relies on the feedback from the environment to narrow down the search space for the optimal policy. Reward shaping [26, 38] relies on expertise and suffers from biased learning. Intrinsic motivations have been proposed to help exploration with extra forces, e.g., surprise [29]. Curriculum Learning has been explored on supervised learning [4], and is extended to RL in [21, 38] to deal with sparse reward. In this work, we exploit the problem structure of constrained bidding for efficient policy learning.

Bayesian RL. Policy learning in unknown environments requires to balance the exploration-exploitation trade-off. A Bayes-optimal policy does this optimally by conditioning on not only the observations but also agent’s uncertainty about the current MDP. Under the framework of *Bayes Adaptive Markov Decision Processes (BAMDP)*, a policy is Bayes-optimal, by augmenting the state space with a belief distribution over possible MDPs [11]. The idea is associated

with belief MDP [20] that represent POMDPs as a belief over MDPs. To achieve tractability, planning on BAMDP is usually achieved by posterior sampling [28, 30] in an MDP periodically sampled from the hypothesis over MDPs. Following this, works in solving POMDPs [18] or meta-RL problems [17, 34, 43], learn approximately Bayes-optimal policies while maintaining a posterior distribution over MDPs, usually via deep variational inference [5, 22]. In this work, we adopt similar ideas to achieve adaptive control in partially observable non-stationary markets.

Constrained RL. Various methods [2, 7, 8, 32] have been proposed to solve CMDPs. Lagrangian relaxation is commonly adopted in [7, 32], which introduces Lagrangian multipliers to control the constraint-objective trade-off, and is shown to have stability issues [8]. RCPO [32] relates with our method closely as we both accommodate constraints into the reward function. However, the proposed indicator augmentation method is parameter-free and exploits the problem structure in contrast to RCPO that degenerates in dynamics-varying environments.

6 CONCLUSION

In this work, we propose the first hard barrier solution to RCB. Based on a Partially Observable Constrained MDP formulation, the indicator-augmented reward function in conjunction with the Curriculum-Guided Bayesian Reinforcement Learning framework achieves adaptive control in partially observable non-stationary markets, without laborious tuning for hyper-parameters. Extensive experiments on a large-scale industrial dataset with two problem settings verify the superior generalization and stability of our method in both in-distribution and out-of-distribution data regimes.

REFERENCES

- [1] N. Abramson, D. J. Braverman, and G. S. Sebestyen. 2006. Pattern Recognition and Machine Learning. *Publications of the American Statistical Association* (2006).
- [2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained policy optimization. In *International conference on machine learning*. PMLR.
- [3] S. Balseiro, A. Kim, M. Mahdian, and V. Mirrokni. 2021. Budget-Management Strategies in Repeated Auctions. *Operations Research* 69, 3 (2021).
- [4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curricular learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.
- [5] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association* (2017).
- [6] 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.
- [7] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. 2017. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research* 18, 1 (2017), 6070–6120.
- [8] Yinlam Chow, Ofer Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. 2019. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031* (2019).
- [9] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, 191–198.
- [10] Chao Du, Zhifeng Gao, Shuo Yuan, Lining Gao, Ziyuan 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.
- [11] Michael O’Gordon Duff. 2002. *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*. Ph.D. Dissertation.
- [12] 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* (2007).
- [13] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. 2016. Bayesian Reinforcement Learning: A Survey. *CoRR* abs/1609.04436 (2016).
- [14] Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. 2021. Bidding and Pricing in Budget and ROI Constrained Markets. (2021).
- [15] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR.
- [16] 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*. Association for Computing Machinery, New York, NY, USA, 2993–3001. <https://doi.org/10.1145/3447548.3467199>
- [17] Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A Ortega, Yee Whye Teh, and Nicolas Heess. 2019. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424* (2019).
- [18] Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. 2018. Deep variational reinforcement learning for POMDPs. In *International Conference on Machine Learning*. PMLR, 2117–2126.
- [19] Antoine Jamin and Anne Humeau-Heurtier. 2019. (Multiscale) Cross-Entropy Methods: A Review. *Entropy* 22 (12 2019). <https://doi.org/10.3390/e2210045>
- [20] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* (1998).
- [21] Andrej Karpathy and Michiel van de Panne. 2012. Curriculum learning for motor skills. In *Canadian Conference on Artificial Intelligence*. Springer, 325–330.
- [22] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [23] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).
- [24] Chi-Chun Lin, Kun-Ta Chuang, Wush Chi-Hsuan Wu, and Ming-Syan Chen. 2016. Combining powers of two predictors in optimizing real-time bidding strategy under constrained budget. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2143–2148.
- [25] James John Martin. 1967. *Bayesian decision problems and Markov chains*. Wiley.
- [26] Maja J Mataric. 1994. Reward functions for accelerated learning. In *Machine learning proceedings 1994*. Elsevier, 181–189.
- [27] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
- [28] Ian Osband, Daniel Russo, and Benjamin Van Roy. 2013. (More) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems* 26 (2013).
- [29] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*. PMLR, 2778–2787.
- [30] Malcolm Strens. 2000. A Bayesian framework for reinforcement learning. In *ICML*, Vol. 2000, 943–950.
- [31] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
- [32] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. 2018. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074* (2018).
- [33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [34] Haozhe Wang, Jiale Zhou, and Xuming He. 2020. Learning Context-Aware Task Reasoning for Efficient Meta Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*.
- [35] T. Wang, H. Yang, H. Yu, W. Zhou, and H. Song. 2019. A Revenue-Maximizing Bidding Strategy for Demand-Side Platforms. *IEEE Access PP*, 99 (2019), 1–1.
- [36] Christopher A Wilkens, Ruggiero Cavallo, Rad Niazadeh, and Samuel Taggart. 2016. Mechanism Design for Value Maximizers.
- [37] D. Wu, X. Chen, X. Yang, H. Wang, Q. Tan, X. Zhang, J. Xu, and K. Gai. 2018. Budget Constrained Bidding by Model-free Reinforcement Learning in Display Advertising. *ACM* (2018).
- [38] Yuxin Wu and Yuandong Tian. 2016. Training agent for first-person shooter game with actor-critic curriculum learning. (2016).
- [39] 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.
- [40] Shuai Yuan, Jun Wang, and Xiaoxue Zhao. 2013. Real-time bidding for online advertising: measurement and analysis. In *Proceedings of the seventh international workshop on data mining for online advertising*, 1–8.
- [41] W. Zhang. 2016. Optimal Real-Time Bidding for Display Advertising. In *UCL (University College London)*.
- [42] Y. Zhou, D. Chakrabarty, and Rajan M Lukose. 2008. Budget constrained bidding in keyword auctions and online knapsack problems. In *Internet and Network Economics, 4th International Workshop, WINE 2008, Shanghai, China, December 17-20, 2008. Proceedings*.
- [43] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. 2019. Varibad: A very good method for bayes-adaptive rl via meta-learning. *arXiv preprint arXiv:1910.08348* (2019).

A ADDITIONAL THEORETICAL RESULTS

A.1 Proof of the Optimal Bidding Theorem

THEOREM A.1. *The optimal bidding function for problem (1) is:*

$$b_i = \frac{\lambda_0 + \lambda_1}{\lambda_1 L + \lambda_2} u_i \quad (15)$$

where $[\lambda_0, \lambda_1, \lambda_2]$ is a non-zero vector, with $\lambda_i \geq 0, i \in \{0, 1, 2\}$.

PROOF. Assume the market price m for an impression x follows a distribution $p_{M|X}(m|x)$. The expected cost and the expected delivery are as follows:

$$\begin{aligned} \mathbb{E}[c|x] &= \int_0^b c(m) \cdot p_{M|X}(m|x) dm \\ \mathbb{E}[d|x] &= d \int_0^b p_{M|X}(m|x) dm \end{aligned} \quad (16)$$

Particularly, in second price auctions, $c_i \equiv m_i$, and we write $c_i \equiv c(m_i)$ to indicate that cost is a function of the market price. Note that, $\nabla_b \mathbb{E}[c|x] = b_i \cdot p_{M|X}(b_i|x)$.

Problem (1) is re-phrased as follows considering the stochasticity from $p_{M|X}(m|x)$:

$$\begin{aligned} \max_b \quad & h(b) \\ \text{s.t.} \quad & f(b) \leq 0, g(b) \leq 0 \end{aligned} \quad (17)$$

where

$$\begin{aligned} h(b) &\stackrel{\text{def}}{=} \sum_{i=1}^T \mathbb{E}[d|x_i], f(b) \stackrel{\text{def}}{=} \sum_{i=1}^T -\mathbb{E}[d|x_i] + L \cdot \mathbb{E}[c|x_i] \\ g(b) &\stackrel{\text{def}}{=} \sum_{i=1}^T \mathbb{E}[c|x_i] - B \end{aligned} \quad (18)$$

By the Fritz John conditions, the optimal solution satisfies

$$\lambda_0 \nabla_b h - \lambda_1 \nabla_b f_j - \lambda_2 \nabla_b g = 0 \quad (19)$$

where $\lambda = [\lambda_0, \dots, \lambda_2]$ is a non-zero vector such that $\lambda_k \geq 0, \forall k \in \{0, 1, 2\}$. That means, $\forall i \in \{1, \dots, T\}$

$$(\lambda_0 d_i - \lambda_2 b_i + \lambda_1 (d_i - L b_i)) \cdot p_{M|X}(b_i|x) = 0 \quad (20)$$

which gives the optimal bidding function Eq. (A.1), with $u_i = \mathbb{E}[d_i]$. \square

A.2 Derivation of the ELBO objective

The objective of Q-learning is to minimize the Bellman Residual:

$$\mathbb{E}_{(o_t, a_t, o'_t, r_t) \sim \mathcal{B}} \left[\left(Q(o_t, a_t) - \left(r_t + \gamma \max_a Q(o_{t+1}, a) \right) \right)^2 \right] \quad (21)$$

where \mathcal{B} denotes a replay buffer.

Minimizing the Bellman Residual is equivalent to maximizing the log likelihood of the transition tuple (o_t, a_t, s_{t+1}, r_t) with proper assumptions⁹:

$$\begin{aligned} & \max \log P(o_t, a_t, y_t) \\ &= \max \log P(y_t | o_t, a_t) P(o_t, a_t) \\ &= \max - (Q(o_t, a_t) - \hat{y}_t)^2 + \log P(o_t, a_t) \end{aligned} \quad (22)$$

⁹The distribution is Gaussian with constant std-dev, and the mean function as $Q(o_t, a_t)$.

where $\hat{y}_t = r_t + \gamma \max_a Q(o_{t+1}, a)$ is the target value computed with the full state s_{t+1} . Note that when maximizing the Q function, the second term is treated as constant.

We use z as the real-valued vector representation for the unobserved state. We have the following *Evidence Lower Bound* for the log likelihood:

$$\begin{aligned} & \log P(o_t, a_t, y_t) \\ & \geq \mathbb{E}_{z \sim q} [\log P(o_t, a_t, y_t | z)] - \mathcal{D}_{KL}(q(z) \| P(z)) \\ &= \mathbb{E}_{z \sim q} [\log P(y_t | o_t, a_t, z)] + \log P(o_t, a_t) - \mathcal{D}_{KL}(q(z) \| P(z)) \end{aligned} \quad (23)$$

Similar to Eq. (22):

$$\log P(y_t | o_t, a_t, z) \Leftrightarrow (Q(o_t, a_t, z) - y_t)^2 \quad (24)$$

Accordingly, minimizing Eq. (21) amounts to maximizing the following ELBO:

$$\max_q \mathbb{E} [-\mathbb{E}_{z \sim q} [(Q(o_t, a_t, z) - y_t)^2] - \mathcal{D}_{KL}(q(z) \| P(z))] \quad (25)$$

B IMPLEMENTATION DETAILS

B.1 Curriculum Design

We have mentioned the design principles for L_t^k in Sec. 3.2.3. We implement each curriculum as a dense reward function of the form (11) with the constraint limits L_t^k evolving along time following the power law:

$$\begin{aligned} L_t^k &= (1 - b_k \cdot (1 - t/T)^g) \cdot L \\ B_t^k &= (h_k \cdot (1 - t/T)^g) \cdot B \end{aligned} \quad (26)$$

where $b_k \in [0, 1]$ determines the relaxation of the original constraint L . For example, $b_k = 0.5$ indicates the maximal relaxation of L is by one half, at the beginning of the bidding process. b_k, h_k depends on data, and for fixed curriculum we set b_k to 0.1 and 0.2 for the first two curriculum, and h_k fixed to 0.95. We empirically set $g = 3$. To achieve automated curriculum learning, the objective is regret minimization,

$$\min_{b_k} \mathbb{E}_{\epsilon_T} \left[\left(D_T - \sum_{t=1}^T \mathcal{R}'(s_t, a_t) \right) \mathbf{1}_F \right]. \quad (27)$$

The insight is to learn b_k such that the proxy cumulative return $\sum_{t=1}^T \mathcal{R}'(s_t, a_t)$ is close to the ground truth D_T given that the episode ϵ_T is feasible. While reward function (11) is a function of b_k , it is not differentiable due to the indicator function. Accordingly, we use the following smooth approximation for $\mathbf{1}_{F_{L_t^k}}$:

$$I(\text{ROI}_T; v) = \frac{1}{\exp(-v(x + \sqrt{v}))} \quad (28)$$

where v controls the slope of the above function transitioning from 0 to 1. We empirically set to 10 with learning rate $3e-3$.

B.2 Implementation

Policy design. The proposed model includes a variational encoder parameterized by a three-layer bi-directional transformer, a conditional policy, a critic that comprises two Q networks and two target Q networks, all implemented as MLPs. For more configurations please refer to our code.

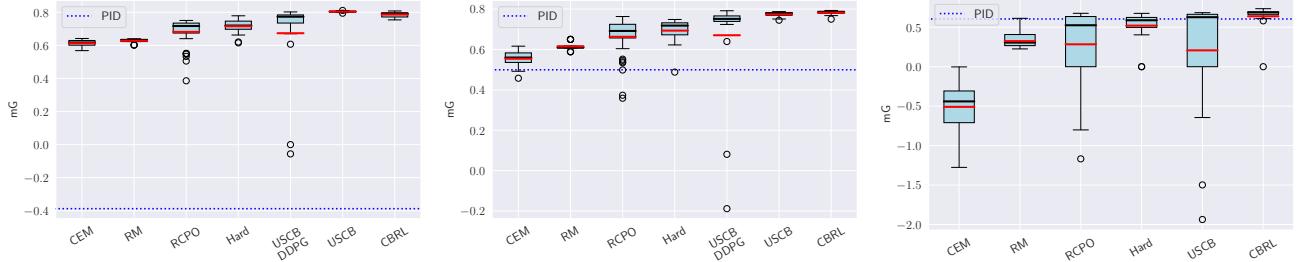


Figure 8: The G-metric performance of SC (Left) setting and MC (Middle) setting on ID split, and SC setting on OOD split (Right).

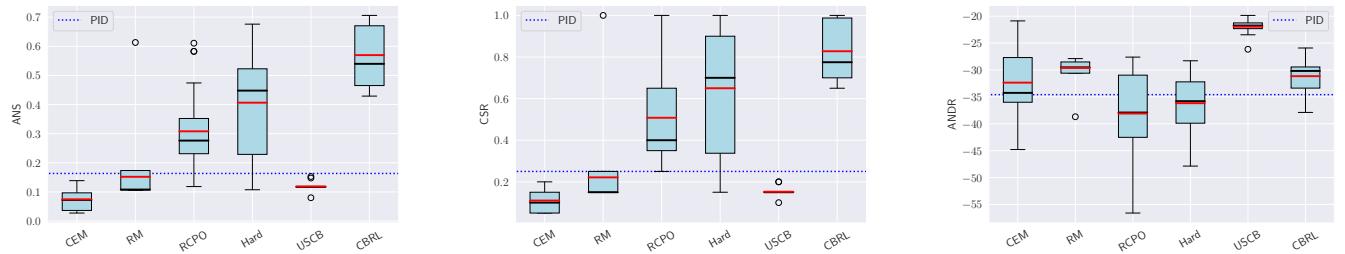


Figure 9: Results of competing methods on OOD split.

The input to the policy includes the following statistics: (1) the time slot t , (2) the bid ratio in previous slot b_{t-1} , (3) the current ROI difference $\text{ROI}_{t-1} - L$ and the current budget consumption rate C_{t-1}/B , (4) the ROI difference of previous slot $\frac{D_{t-2:t-1}}{C_{t-2:t-1}} - L$, (5) the normalized delivery of previous slot $T \times D_{t-2:t-1}/D_T^*$, (6) the current surplus $D_{t-1} - L \times C_{t-1}$. Clipping is adopted to ensure the statistics remain in the proper scale. The time slot length is empirically set to half an hour, and thus $T = 48$. The output space is set to $[0, 4]$ with tanh Gaussian trick to bound the action.

In [16], temporally correlated action space is used, i.e., policy output is added to the previous bid ratio. Models in our experiments use independent action space except for **USCB-DDPG**. Besides, we note that **USCB-DDPG** is different from the standard DDPG [23] as it fits the Q-function with Monte Carlo return estimates instead of the (bootstrapped) Q-learning.

Policy Learning. We adopt SAC [15] for policy optimization, an actor-critic method [23] that uses entropy regularization. We normalize the objective value by the oracle and the constraint violations by the limits, to balance the scale of the two parties. The learning rate is set $3e-4$ for all networks, and is stepped at $\{4000, 8000, 12000\}$ with decay rate 0.5.

Slot-wise Oracle. We solve the following programming problem,

$$\max \sum_{t=1}^T D(\epsilon_{[t]}; \beta_t), \quad \text{s.t. } \text{ROI}_T \geq L, B - C_t \geq 0, \quad (29)$$

where $D(\epsilon_{[t]}; \beta_t, u_{[t]})$ is the total delivery obtained by bidding β_t to the impressions in slot t with utilities $u_{[t]}$.

If we discretize the bid ratio space, the problem (29) can be treated as a group knapsack problem. Suppose we have T groups of items, in group t , an item β_t is with value $D(\epsilon_{[t]}; \beta_t) = \sum_{i \in [t]} d_i \mathbb{1}_{\beta_t u_i > m_i}$ and weight $C(\epsilon_{[t]}; \beta_t) = \sum_{i \in [t]} c_i \mathbb{1}_{\beta_t u_i > m_i}$. Problem (29) equals

to solving the optimal item set such that, (1) in each group, one and only one item is selected, (2) the total weight of selected items does not exceed B , and (3) the total value-to-weight ratio should be above L . We solve this knapsack problem by an linear programming toolbox¹⁰ to obtain the slot-wise oracle. This oracle performs better than a day-wise oracle when there are utility prediction error. Intuitively, slot-wise policy lends more space to correct the prediction error by adjusting the bid ratio, while the day-wise policy offers no such flexibility.

Evaluation Protocols. The metrics used in our experiments are computed as:

$$\text{ANS} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^N \frac{D(\epsilon_T^{(i)})}{D^*(\epsilon_T^{(i)})} \cdot \mathbb{1}_F(\epsilon_T^{(i)}; L^{(i)}, B^{(i)}) \quad (30)$$

$$\text{CSR} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^T \mathbb{1}_F(\epsilon_T^{(i)}; L^{(i)}, B^{(i)}) \quad (31)$$

$$\text{ANDR} \stackrel{\text{def}}{=} \frac{1}{|F|} \sum_{\epsilon_T^{(i)} \in F} \left(\frac{D(\epsilon_T^{(i)})}{D^*(\epsilon_T^{(i)})} - 1 \right) \times 100\% \quad (32)$$

C ADDITIONAL EMPIRICAL RESULTS

The performance of competing methods are shown in Fig. 9.

USCB [16] propose a G-metric, which non-linearly combines the constraint violations and the performance objective with a hyper-parameter λ .

The main drawback of the G-metric is how to choose the hyper-parameter. We follow [16] to use the hyper validated by USCB, i.e., the hyper-parameter with which USCB achieves the best performance in ANS. The results in G-metric are shown in Fig. 8.

¹⁰<https://github.com/coin-or/pulp>