# Learning to Bid with AuctionGym

Olivier Jeunen
Amazon
Edinburgh, United Kingdom

Sean Murphy
Amazon
Edinburgh, United Kingdom

Ben Allison
Amazon
Edinburgh, United Kingdom

## ABSTRACT

Online advertising opportunities are sold through auctions, billions of times every day across the web. Advertisers who participate in those auctions need to decide on a bidding strategy: how much they are willing to bid for a given impression opportunity. Deciding on such a strategy is not a straightforward task, because of the *interactive* and *reactive* nature of the repeated auction mechanism. Indeed, an advertiser does not observe counterfactual outcomes of bid amounts that were not submitted, and successful advertisers will adapt their own strategies based on bids placed by competitors. These characteristics complicate effective learning and evaluation of bidding strategies based on logged data alone.

The *interactive* and *reactive* nature of the bidding problem lends itself to a bandit or reinforcement learning formulation, where a bidding strategy can be optimised to maximise cumulative rewards. Several design choices then need to be made regarding parameterisation, model-based or model-free approaches, and the formulation of the objective function. This work provides a unified framework for such "*learning to bid*" methods, showing how many existing approaches fall under the value-based paradigm. We then introduce novel policy-based and doubly robust formulations of the bidding problem. To allow for reliable and reproducible offline validation of such methods without relying on sensitive proprietary data, we introduce AuctionGym: a simulation environment that enables the use of bandit learning for bidding strategies in online advertising auctions. We present results from a suite of experiments under varying environmental conditions, unveiling insights that can guide practitioners who need to decide on a model class. Empirical observations highlight the effectiveness of our newly proposed methods. AuctionGym is released under an open-source license, and we expect the research community to benefit from this tool.

**ACM Reference Format:**
Olivier Jeunen, Sean Murphy, and Ben Allison. 2022. Learning to Bid with AuctionGym. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining AdKDD Workshop (AdKDD '22), August 15th, 2022, Washington DC, USA.* ACM, New York, NY, USA, 6 pages.

## 1 INTRODUCTION & MOTIVATION

Ad exchanges run advertising auctions, where the opportunity to show an ad is sold off in real-time. Advertisers participate in these auctions, in an attempt to maximise the utility they can obtain from the ads they show. Auctions are well-studied in the economics literature, and several Nobel laureates have contributed to our understanding of them. Indeed, Vickrey showed that the truthful second-price auction maximises social welfare [38], and Myerson showed that with a well-chosen reserve price, this auction format can be revenue-maximising for the auctioneer [26].

Nevertheless, these strong theoretical results rely on assumptions that are seldomly met in present-day advertising scenarios. Indeed, bidders' valuations are *not* symmetrical and repeated auctions are *not* statistically independent. As a result, the second-price format will *not* maximise revenue for the auctioneer, and all major ad exchanges have moved towards first-price auctions where the winner pays their bid amount. It is easy to see that truthful bidding is no longer an optimal strategy here, and that a well-chosen *bidding strategy* should be adopted in order to maximise the surplus a bidder can obtain from participating in the auction. This is not an easy problem to solve, as the only feedback a bidder receives from participating is whether they win, and if they win, what price they need to pay. It is natural to frame such a repeated game with limited information as a *bandit* or *reinforcement learning* problem. This opens up a plethora of design choices that need to be made, regarding parameterisation, model-based or model-free approaches, and the formulation of the objective function. We provide an overview of such design options, and show where existing approaches fit in this framework. This allows us to propose novel approaches for learning to bid, under the policy-based and doubly robust paradigms.

Reliable and reproducible offline validation of "learning to bid" approaches is hard, due to the limitations of logged offline data. Indeed, observational data can only provide limited signal, and experimental data with broad interventions is costly to obtain. Online experiments offer no viable alternative, as they are also prohibitively expensive. Simulations can provide a way forward in such settings, as evidenced by recent strong empirical progress in reinforcement learning [2, 30]. To this end, we propose a novel open-source simulation environment for real-time bidding in computational advertising: **AuctionGym**. AuctionGym allows us to unveil insights that can guide practitioners who need to decide on a "learning to bid" strategy—insights that are not straightforward to extract from logged data alone. We use AuctionGym to empirically illustrate the improvements in bidder surplus that can be attained from our proposed "learning to bid" approaches, leveraging novel policy-based and doubly robust estimators.

In summary, the main contributions of our work are:

(1) We formalise the "learning to bid" problem as a bandit or reinforcement learning task, showing how existing approaches fit into the value-based paradigm.
(2) We introduce novel formulations of the problem, leveraging policy-based and doubly robust estimators.
(3) We present AuctionGym: a simulation environment that enables reproducible and robust validation of "learning to bid" methods without relying on sensitive proprietary data.[1]
(4) We present experimental results that highlight the competitiveness of our newly proposed methods, and uncover insights that can guide practitioners who need to decide which method to use under particular environmental conditions.



[1] AuctionGym is publicly available at github.com/amzn/auction-gym.

## 2 BACKGROUND & RELATED WORK

Truthful bidding (reporting the expectation of one's own private valuation for the good being sold), is a dominant strategy in second-price auctions under several assumptions [38]. These assumptions include that (1) the bidder *knows* their expected valuation given a context, (2) placed bids do not influence the value of the good, (3) competitors all have access to the same information, and (4) repeated rounds of auctions are statistically independent.

In present day online advertising auctions, many of these assumptions are bound to be violated. As a result, the second-price mechanism will not maximise revenue for the auctioneer, and all major ad exchanges have moved away from the second-price format. Advertisers who wish to participate in such auctions now need to decide on a bidding strategy, as the previously industry-standard strategy of truthful bidding has become sub-optimal.

A common violation of the independence assumption occurs when advertisers have budgets. Wu et al. adopt a model-free reinforcement learning approach to learn a single scalar "pacing" parameter for budget optimisation in second-price auctions [40], and other methods have been proposed to incorporate further KPI constraints into the objective [41]. In contrast, we introduce a bandit-based learning framework for *any* auction mechanism, which is crucial for surplus optimisation in non-second price auctions. Furthermore, the bidding strategies we deal with are dependent on contextual covariates per opportunity, allowing high flexibility.

Lowering one's bid in a first-price auction is often referred to as *bid shading*. When the auctioneer reveals the winning bid to *all* participants, this data can be leveraged to learn optimal strategies [10]. Nevertheless, this information is seldom available. Pan et al. propose a two-step bid shading procedure, consisting of (1) win-rate estimation, and (2) surplus maximisation. They adopt a logistic regression model paired with a bisection search for fast inference [28]. Other work directly models the distribution of the "minimum bid to win" [19]—using a range of estimators and an efficient golden section search at inference time [43]. As we will show, these works are in line with a value-based (also known as *model*-based) view of the "learning to bid" problem. Zhang et al. leverage the flexibility of non-parametric approaches to bid shading when the size of the training sample is large, reporting improvements over parametric approaches [42]. AuctionGym allows us to reproduce these insights, whilst providing an additional view on the performance of parametric approaches under a range of environmental conditions. This allows us to identify empirically optimal methods in low- or high-data regimes, with weak or strong competition and frequent or rare model updates, among other configurable parameters.

We are inspired by the success of simulation environments in the broader reinforcement learning research community [2]. In particular, we draw from the RecoGym simulation environment that aims to enable bandit-based optimisation of the *allocation* step, dictating which ad should be shown in a given context [30]. AuctionGym jointly models this step with the *bidding* problem, deciding how much we should bid for a given ad impression opportunity. We believe this opens up exciting future research directions where both problems can be solved jointly. Indeed, even though the outcome of the auction is independent of the allocated ad—the auction outcome has a strong influence on future training data and *exploration*.

## 3 LEARNING TO BID

This section formalises our problem setting and presents a general framework for bandit-based "learning to bid". We highlight parallels with existing approaches, and present novel ways to learn optimal bidding strategies that maximise alternative estimators of *utility*. In what follows, estimated quantities $Q$ are denoted as $\widehat{Q}$.

An advertiser receives a bid request from an ad exchange, described by contextual features $x \in \mathcal{X}$. The advertiser then needs to make two decisions: (1) Which ad from the inventory do we want to show, given context $x$? (*The Ad Allocation Problem*), and (2) How much should we bid for this ad impression? (*The Bidding Problem*).

### 3.1 The Ad Allocation Problem

From the full catalogue of ads $\mathcal{A}$, we source a subset of ads that are eligible to be shown in this context: $\mathcal{A}_x$. Every ad $a \in \mathcal{A}_x$ is tied to a private valuation $v_a \in \mathbb{R}^+$, detailing the advertiser's willingness-to-pay for a conversion-event after an impression (in USD). Low-probability conversion events like sales might be valued highly, whereas higher-probability events such as clicks or views can be valued lower. We denote with the binary random variable $C$ whether such an event has occurred after an impression. For every eligible ad $a_i \in \mathcal{A}_x$, the advertiser can estimate the expected welfare $\omega$ they will obtain from an ad impression:

$$\widehat{\mathbb{E}}[\omega|A = a_i; X = x] \coloneqq v_{a_i} \cdot \widehat{\mathrm{P}}(C|A = a_i; X = x). \tag{1}$$

The conversion estimator $\widehat{\mathrm{P}}(C|A; X)$ is a crucial part of any online advertising system, as reflected by a substantial research literature [13, 17, 25, 39]. This estimator is typically trained in a supervised manner on a collected log of impression-outcome pairs. Any system features related to *exploration* of allocation are assumed to be encoded at this level, and general heterogeneous "conversion events" are considered. We will assume w.l.o.g. that an advertiser chooses the ad that maximises their estimated expected welfare $\widehat{\omega}$:

$$a^\star = \operatorname*{arg\,max}_{a_i \in \mathcal{A}_x} \widehat{\mathbb{E}}[\omega|A = a_i; X = x]. \tag{2}$$

### 3.2 The Bidding Problem

Now we have decided which ad to show, we need to decide how much we are willing to bid for it. That is, we need to decide on a dollar amount $b \in \mathbb{R}^+$ to submit to the ad exchange in response to the bid request. After placing a bid, two things can happen:

(1) We lose the auction to a competing bidder or a reserve price, and we do not need to provide a payment.
(2) We win the auction, and get charged a price $p \leq b$. The auction rules determine this price. Although first-price auctions are common ($p \coloneqq b$), in the general case the price for a given bid will not be known beforehand.

In the bygone era of second-price auctions, a weakly dominant strategy is for all bidders to bid truthfully ($b \coloneqq \widehat{\omega}$). Note that this implicitly assumes that the conversion estimator is wholly unbiased and well-calibrated, which is a strong assumption in real-world systems. For general auctions, bids can be sampled according to some policy $\pi$, where $\mathrm{P}(B = b|A = a; X = x; \Pi = \pi)$ is denoted as $\pi(b|a; x)$. Note that this notation subsumes deterministic bidding strategies when $\pi$ denotes a degenerate distribution.

An advertiser wishes to maximise the expectation of their *utility* $U$, or the surplus in value that they obtain by participating in the auction. Let $W$ be a binary random variable indicating whether the auction was won, let $V \equiv \omega$ be the welfare the advertiser obtains from an ad impression, and let $P$ denote the price paid for participating in the auction. This notation allows us to factorise our utility or surplus as follows:

$$U = W(V - P). \tag{3}$$

Note that, after we have won an auction round, all three components are observable. When we have lost, $W = V = P = 0$. As such, we can write the expected utility obtained from following a bidding strategy $\pi$ over all possible contexts, values and prices as:

$$\begin{aligned} \mathop{\mathbb{E}}_{b\sim\pi(B|A;X)}[U] = \int \mathrm{P}(W = 1|X = x; B = b)(v - p) \\ \mathrm{P}(V = v|A = a; X = x)\mathrm{P}(P = p|X = x; B = b)\mathrm{d}x\,\mathrm{d}v\,\mathrm{d}p. \end{aligned} \tag{4}$$

Here, we assume that (1) the probability of winning and the resulting price are independent of the allocated ad given the bid and context, and (2) welfare is independent of the bid given the allocated ad and context. In some cases, the price $P$ will be known beforehand (for example, when we know that we are participating in a first-price auction). Nevertheless, we can learn a pricing estimator $\widehat{\mathrm{P}}(p|x; b)$ to cover general use-cases with potentially opaque auction mechanisms. There are several ways to approximate the above expectation. In what follows, we explore our options.

*3.2.1 Value-Based Estimation (Model-Based).* By decoupling the *inference* and *decision-making* steps, we can leverage decades of progress in supervised learning to handle bidding. That is, we first derive a utility estimator $\widehat{u}$ for a context-ad-bid triplet:

$$\widehat{u}(x, a, b) \approx \mathbb{E}[U|X = x; A = a; B = b]. \tag{5}$$

Indeed, this regression model can be learned from observed samples in a supervised manner. Naturally, we can leverage the "hurdle" structure in Eq. 3 to factorise the estimator into separate *winrate*, *welfare* and *pricing* estimators [24].[2] When we can obtain an estimate of utility for every bid in every impression opportunity, we can obtain an estimate for the expected utility a bidding policy will obtain. In the bandit literature, such an approach is often dubbed the Direct Method (DM). For a given training sample $\mathcal{D}$:

$$\begin{aligned} \mathop{\mathbb{E}}_{b\sim\pi(B|A;X)}[U] &\approx \widehat{U}_{\mathrm{DM}}(\pi, \mathcal{D}) \\ &= \sum_{(x,a,b,u)\in\mathcal{D}} \int \widehat{u}(x, a, b')\pi(b'|a; x)\mathrm{d}b'. \end{aligned} \tag{6}$$

This integral is maximised by the degenerate distribution:

$$\pi(b'|a; x) = 1 \iff b' = \arg\max_b \widehat{u}(x, a, b). \tag{7}$$

This optimum is typically attained by doing a discretised search at inference time [28, 43]. Note that this can complicate adoption of such methods, due to the latency constraints in real-time bidding environments. Furthermore, it does not handle exploration.

When we enforce a certain family of non-degenerate distributions on $\pi$, we need to explicitly optimise Eq. 6. We can optimise the policy to maximise its expected estimated utility via Monte Carlo samples. That is, given a utility model $\widehat{u}$, we sample from $\pi$ to approximate the integral in $\widehat{U}_{\mathrm{DM}}$, and backpropagate to perform gradient ascent (possibly improving variance by leveraging reparameterisation tricks [20]). To avoid for the scale of the bidding distribution to collapse (as the bandit setting might not incentivise exploration), we can add an entropy regularisation term to the objective, balancing exploitation with exploration [12]. This learning approach for $\pi$ has the added advantage of only needing a single forward pass at inference time to obtain $b \sim \pi(b|a; x)$.

*3.2.2 Policy-Based Estimation (Model-Free).* Modelling the reward process, as value-based methods do, is essentially a way to decrease variance when estimating Eq. 4. Reduced variance often comes at the cost of increased bias, and biases in either the winrate, welfare or pricing estimators can propagate and amplify, leading to suboptimal solutions. In fact, there is no need to explicitly model the reward process. In contrast, we can directly optimise a bidding policy to maximise the integral in Eq. 4 based on observed samples. For this to work, we make use of importance sampling [27]. We now additionally need information about the policy that was in production at the time of data collection, often referred to as the *logging policy* $\pi_0$. Given a training sample $\mathcal{D}$, the optimisation objective can be written as:

$$\mathop{\mathbb{E}}_{b\sim\pi(B|A;X)}[U] \approx \widehat{U}_{\mathrm{IPS}}(\pi, \mathcal{D}) = \sum_{(x,a,b,u)\in\mathcal{D}} u\frac{\pi(b|a; x)}{\pi_0(b|a; x)}. \tag{8}$$

This estimator is often referred to as an Inverse Propensity Score (IPS) estimator, as it effectively weights observed samples by the ratio of the propensities between the learnt and logging policies. When a bid $b$ that was rare under the logging policy leads to positive utility, the learnt and logging policies will tend to diverge—as increasing $\pi(b|a; x)$ increases the objective in Eq. 8. The ratio between the two probability densities (the so-called importance weights) will be high, and such rare samples can bear disproportional weight in the final estimator. Indeed, even though the vanilla IPS estimator is unbiased, its variance is often problematic. In the finite sample scenarios that are relevant to practitioners, variance-reducing extensions are known to yield empirical improvements. Most often, the importance weights are clipped to some maximal value [9, 15, 32], or the learnt policy is disincentivised to deviate from the logging policy through a regularisation term [8, 31, 33, 37].

*3.2.3 Doubly Robust Estimation.* The value- and policy-based families tackle the same estimation problem from a different angle, and can provide complementary advantages. We can combine these advantages in a *doubly robust* estimator, that is provably unbiased when *either* the utility model or the propensity scores are [4].[3]

$$\begin{aligned} \mathop{\mathbb{E}}_{b\sim\pi(B|A;X)}[U] \approx \widehat{U}_{\mathrm{DR}}(\pi, \mathcal{D}) = \\ \sum_{(x,a,b,u)\in\mathcal{D}} \left( \int \widehat{u}(x, a, b')\pi(b'|a; x)\mathrm{d}b' + (u - \widehat{u}(x, a, b))\,\frac{\pi(b|a; x)}{\pi_0(b|a; x)} \right) \end{aligned} \tag{9}$$

[2] Additional structure can be leveraged here to improve predictive performance, such as the monotonicity between the placed bid, and the winrate and impression cost.

[3] Note that this does not guarantee performance improvements in practice [16].

Intuitively, $\widehat{U}_{\text{DR}}$ uses $\widehat{U}_{\text{DM}}$ as a baseline, and corrects for its errors using the importance weights. As in the Direct Method, we can approximate the integral in Eq. 9 by sampling from $\pi$, and subsequently updating $\pi$ via backpropagation. Several extensions to the DR paradigm exist. They either focus on optimising the trade-off between DM and IPS [36], optimise the reward model to minimise the overall variance of the estimator [7], or transform the IPS weights to minimise bounds on the expected error of the estimate [35]. When using DR, we need to make an additional choice. That is, we need to decide which samples to use to learn the utility model, and which samples to use for the policy optimisation process. In low training sample regimes, this could become problematic. Nevertheless, doubly robust learning has led to significant performance improvements in a range of application domains [35].

## 4 AUCTION GYM

Validating the performance of a learnt bidding policy is not a straightforward task. Offline, we can make use of logged data to generate counterfactual estimates of performance. These are in fact exactly the counterfactual estimators proposed in Section 3. This is problematic, as it is reminiscent of Goodhart's Law: "*Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes*" [11].[4] Moreover, existing counterfactual estimators tend to make strong stationarity assumptions about the environment (i.e. the bidding behaviour of competing advertisers) that do not hold in practice. Indeed, competitor bids will react to the chosen bidding policy, and this reactive effect is not sufficiently captured by logged data alone. This renders them irrevocably biased—although useful when learning bidding policies.

Online experiments, as an alternative, are too expensive to be used as a first-line validation tool. Indeed, prototypes for new approaches need to be brought up to standards for production code, A/B-tests typically span at least several days to obtain statistically significant performance estimates, and we risk losing business value by actively exploring suboptimal bidding approaches.

The reinforcement learning research community is well aware of these shortcomings, and reliable simulation environments are at the heart of significant advances in recent years [2]. Their success has led to enthusiasm and advocacy for the use of simulations in related fields like Recommender Systems [6, 14, 30], where they have been accepted and adopted as an alternative evaluation mechanism [1, 16–18]. It is our belief that simulation can open similar doors in the computational advertising and real-time bidding research communities, especially with respect to novel approaches for bandit and reinforcement learning.

To this end, we propose **AuctionGym**, an open-source environment that simulates the advertising problem end-to-end:

(1) An impression opportunity arises, with features $x \sim \mathrm{P}(X)$,
(2) the auctioneer presents this opportunity to some bidders,
(3) bidders internally decide on an ad to show and a bid to place,
(4) the auctioneer decides on the auction winner and price,
(5) the winning ad is shown and possibly leads to a conversion event that is observable by the winning bidder.

[4] This insight was later paraphrased and popularised as:
"*When a measure becomes a target, it ceases to be a good measure*" [34].

This real-time auction process is repeated into *rounds*, where $\Delta_{\text{r}}$ rounds are repeated into $N_{\text{i}}$ *iterations*. To simulate a delayed batch feedback setting, bidders update their allocation and bidding policies after every iteration, based on the previously observed $\Delta_{\text{r}}$ auction rounds. Naturally, bidders end up paying a price for auction rounds they participate in and win, but only incur *utility* or *reward* when the impressed ad leads to their desired outcome (*welfare*, Eq. 1). Bidders simultaneously need to solve both the *allocation* and *bidding* problems, in order to maximise their own utility.

*Simulating Advertising Outcomes.* AuctionGym not only simulates the auction itself, but also whether an allocation decision leads to a conversion event for the advertiser. As such, for a given context $x$ and an ad $a$, the internal system consists of a stochastic process that simulates this. That is, we draw $c \sim \text{Bernoulli}(\rho_{a,x})$ where:

$$\rho_{a,x} \coloneqq \mathrm{P}(C = 1 | A = a; X = x) = f_\theta(x, a). \tag{10}$$

A design choice needs to be made with respect to the parameterisation of $f_\theta$. No restrictions on the functional form of $f_\theta$ are generally necessary, but this can complicate efficient learning and inference. We can draw on existing work to make reasonable assumptions about how users interact with ads, such as the "*latent factor model*" assumption that is at the foundation of modern recommendation research [21]. For a given dimensionality $D$, we have parameters $\theta = \{\phi, \beta\}$, with ad-specific parameters $\phi_a \in \mathbb{R}^D, \beta_a \in \mathbb{R}$:

$$f_{\phi,\beta}(x, a) = \sigma\left(x\phi_a^\intercal + \beta_a\right). \tag{11}$$

Here, $\sigma$ denotes the logistic sigmoid. This is similar to the parameterisation adopted by RecoGym [30]. One advantage is that it allows the use of fast approximate nearest neighbour techniques in the allocation decision (i.e. the arg max operation in Eq. 2), which are widespread and crucial in real-world large-scale advertising systems [23]. For this reason, simulated bidders adopt the same functional form for $\widehat{\mathrm{P}}(C)$. Naturally, advertisers do not fully observe all the contextual information that influences user behaviour. This confounding effect is simulated by obfuscating the *true* contextual vector: $\tilde{x} \coloneqq x_{[1:k]}$ where $1 \leq k \leq D$, and only $\tilde{x}$ is observable.

Although the ad allocation problem is not the focal point of our work, we believe that it is crucial to jointly study the *allocation* and *bidding* problems, rather than in isolation. Indeed, the value estimates that are used in the allocation step are equally important for bidding, and noisy estimates will propagate and have downstream effects. The auction, in turn, has a strong influence on future training data that is available to train allocation models. AuctionGym includes an implementation of Bayesian logistic regression with Thompson sampling to handle the allocation of ads [3].

*Simulating Bidders.* Every bidder $j$ has a private ad catalogue $\mathcal{A}^j$. Bidders have private valuations $v_a$ they place on a conversion event for a given ad. The ad-specific parameters $\phi_a, \beta_a$ that dictate $\rho_{a,x}$ are not observable by the bidder, and are configurable. That is, they can be fully synthetic and drawn from an arbitrary specified distribution, or they can be instantiated based on real-world data to inform semi-synthetic experiments (as also done by Bendada et al. [1]). AuctionGym includes implementations of all the bidding strategies introduced in Sec. 3, using PyTorch [29]. We expect that existing approaches are easily extendable, and that new approaches can be implemented in the common framework to allow for robust and reproducible validation under varying configurations.

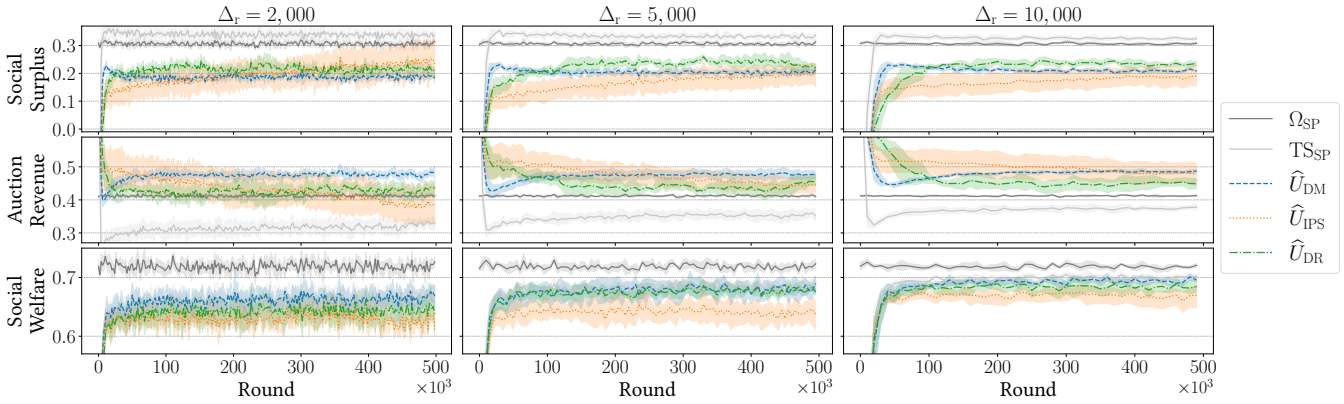


**Figure 1: Evolution of key metrics (95% C.I., y-axis) in repeated auction rounds (x-axis), when all competing bidders optimise their bidding strategy according to the same utility estimator. We vary the number of rounds between model updates $\Delta_r$, increasing from left to right. We observe that compared to the widespread model-based approach, model-free learning leads to high variance, whereas our doubly robust estimator improves upon existing methods, increasing bidders' surplus.**

*Simulating Auctions.* AuctionGym includes implementations of the most often used auction formats: first-price and second-price auctions, possibly with *hard* or *soft* floors. This is however no restriction, and more complex alternatives can be included to test a range of hypotheses. In particular, even though we focus on "learning to bid", we believe that AuctionGym can provide a common framework for evaluating learnt auction mechanisms as well. Indeed, this emerging research area can also be framed as a reinforcement learning problem, where the auctioneer needs to decide (1) who wins the auction, and (2) how much they will be charged [5, 22].

*Metrics of Interest.* AuctionGym allows us to track multiple metrics of interest, such as the auctioneer's revenue, and bidders' welfare and surplus. We also consider *Return On Ad Spend (ROAS)* — an industry standard KPI to evaluate advertising efficiency. We can define multiple notions of bidders' *regret*. That is, how much value bidders are missing out on due to suboptimal allocation or bidding decisions. *Allocation regret* is the loss in welfare incurred due to suboptimal ad allocation. *Estimation regret* is the loss in welfare incurred due to biased value estimation. *Overbid regret* is the loss in surplus incurred due to overbidding, and *underbid regret* is analogously defined as loss in surplus due to underbidding.

## 5 EXPERIMENTAL RESULTS & DISCUSSION

In what follows, we provide a non-exhaustive list of research questions that AuctionGym can help answer. We assume 1st price auctions unless explicitly mentioned otherwise. All models are shallow multi-layer perceptrons, all policies are parametrised Gaussians.

**RQ1** What is the effect of moving from 2nd to 1st price auctions?
**RQ2** What is the effect of learnt (vs. optimal) ad allocation?
**RQ3** How does the choice of estimator affect *social* measures?
**RQ4** How does the choice of estimator affect *individual* measures?

***RQ1-3:*** *Learning to Bid and Social Effects.* Fig. 1 shows results from repeated auction rounds with two out of six competing bidders per round, all having twelve ads in their catalogue. We repeat this process for five different random seeds (whilst keeping the catalogues fixed), and report 95% C.I.'s for the evolution of relevant measures per auction round as all bidders continuously learn and update their allocation and bidding strategies. The aggregated results summarise more than 37 million simulated auctions, and a combined 36 000 distinct learnt bidding strategies over all bidders and configurations. Social welfare indicates the overall value that is generated through the advertising auction: either for the auctioneer (auction revenue) or the bidders (social surplus). This decomposition is at the heart of what learnt bidding strategies aim to influence: they aim to maximise surplus, which inevitably leads to a decrease in revenue for the auctioneer. We include measurements for an oracle that knows the true parameters $\rho_{a,x}$ and bids truthfully in a second-price auction as $\Omega_{SP}$, and the Thompson sampling approach in similar settings as $TS_{SP}$. This allows us to quantify the effects of "learning to bid" approaches on welfare, revenue and surplus.

Focusing on social surplus, we observe that the model-based approach stabilises quickly but suboptimally, as is expected for a biased low-variance estimator. The model-free importance sampling estimator has high variance, and is able to improve upon the model-based estimator when sufficient learning steps are allowed.[5] The instability of this approach, however, can lead to significant reductions in attainable welfare as it impacts training data collection for subsequent updates to the allocation model. Our novel doubly robust estimator leads to improved surplus over all bidders participating in the auction, with much lower variance than IPS.[6]

***RQ4:*** *Individual Effects.* Fig. 2 shows results from repeated auction rounds in the same configuration as Fig. 1, where all bidders optimise their bidding strategy using $\widehat{U}_{DM}$ — from prior existing work, this can be interpreted as an optimistic industry status quo. The goal here is to get an actionable recommendation: *which learning strategy should a single bidder adopt in order to maximise their profit, and why*? We plot ROAS, overbid and underbid regret respectively over time. This reinforces the observations obtained from research questions 1–3: the Direct Method has low variance but high bias, leading to fast convergence with considerable overbid regret as a result. Importance sampling is promising but entails high variance, effectively reducing overbid regret for a slight increase in underbid regret after 1 000 000 auction rounds. Doubly robust estimation consistently improves ROAS and overbid regret, making it a strong contender for adoption in real-world systems.

[5] Because we implement weight-clipped IPS as in PPO [32], this is expected behaviour.
[6] Note that we clip the importance weights in $\widehat{U}_{DR}$ as well, resembling the recently proposed doubly robust estimator with pessimistic shrinkage [35].

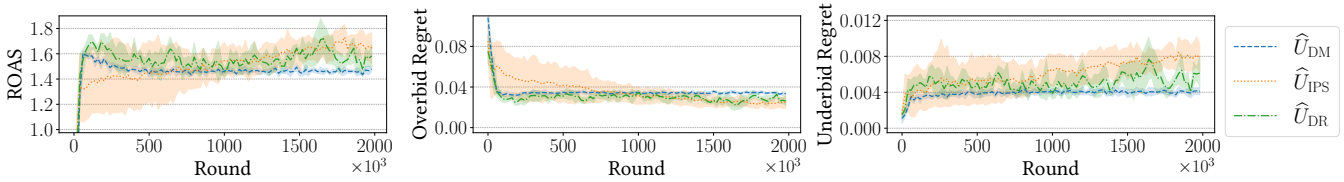


**Figure 2: Evolution of key metrics (95% C.I., y-axis) in repeated auction rounds (x-axis), focused on a single bidder where $\Delta_{\text{r}} = 20\,000$. We observe that compared to the widespread model-based approach, model-free learning leads to high variance, whereas our doubly robust estimator improves upon existing methods, increasing bidders' ROAS due to reduced overbidding.**

## 6 CONCLUSIONS & OUTLOOK

We have advocated for the "*learning to bid*" problem, that is prevalent in present-day online advertising, to be cast as a bandit learning problem. To this end, we have presented a general framework for bandit-based "learning to bid", allowing us to frame existing methods and propose novel approaches that leverage policy-based and doubly robust estimators. We have presented the AuctionGym simulation environment that can be used to reliably validate such approaches in a reproducible manner, without relying on sensitive proprietary data. AuctionGym can be used to unveil insights that cannot be straightforwardly extracted from logged data — and we expect the research community to benefit from this tool.

In future work, we wish to consider full reinforcement learning instantiations of the bidding problem, where current actions influence future states and a notion of *planning* can further improve bidder surplus. Naturally, we wish to further validate the insights presented in this work on real-world data, and to better understand the benefits and limitations of doubly robust "learning to bid". Finally, we wish to extend the simulation environment to support advertiser budgets, multi-item and learnt auction mechanisms.

## REFERENCES


[1] W. Bendada, G. Salha, and T. Bontempelli. 2020. Carousel Personalization in Music Streaming Apps with Contextual Bandits. In *RecSys '20*.
[2] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. 2016. OpenAI Gym. https://arxiv.org/abs/1606.01540
[3] O. Chapelle and L. Li. 2011. An Empirical Evaluation of Thompson Sampling. In *NeurIPS '11*.
[4] M. Dudík, J. Langford, and L. Li. 2011. Doubly Robust Policy Evaluation and Learning. In *ICML '11*.
[5] P. Duetting, Z. Feng, H. Narasimhan, D. Parkes, and S. S. Ravindranath. 2019. Optimal Auctions through Deep Learning. In *ICML '19*.
[6] M. D. Ekstrand, A. Chaney, P. Castells, R. Burke, D. Rohde, and M. Slokom. 2021. SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research. In *RecSys '21*.
[7] M. Farajtabar, Y. Chow, and M. Ghavamzadeh. 2018. More Robust Doubly Robust Off-policy Evaluation. In *ICML '18*.
[8] L. Faury, U. Tanielian, F. Vasile, E. Smirnova, and E. Dohmatob. 2020. Distributionally Robust Counterfactual Risk Minimization. In *AAAI '20*.
[9] A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. 2018. Offline A/B Testing for Recommender Systems. In *WSDM '18*.
[10] D. Gligorijevic, T. Zhou, B. Shetty, B. Kitts, S. Pan, J. Pan, and A. Flores. 2020. Bid Shading in The Brave New World of First-Price Auctions. In *CIKM '20*.
[11] C. A. E. Goodhart. 1984. *Problems of Monetary Management: The UK Experience*. Macmillan Education UK, 91–121.
[12] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *ICML '18*.
[13] X. He, O. Pan, J.and Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. Candela. 2014. Practical Lessons from Predicting Clicks on Ads at Facebook. In *KDD '14 AdKDD Workshop*.
[14] E. Ie, C. Hsu, M. Mladenov, V. Jain, S. Narvekar, J. Wang, R. Wu, and C. Boutilier. 2019. RecSim: A Configurable Simulation Platform for Recommender Systems. https://arxiv.org/abs/1909.04847
[15] E. L. Ionides. 2008. Truncated Importance Sampling. *Journal of Computational and Graphical Statistics* 17, 2 (2008), 295–311.
[16] O. Jeunen and B. Goethals. 2020. An Empirical Evaluation of Doubly Robust Learning for Recommendation. In *RecSys '20 REVEAL Workshop*.
[17] O. Jeunen and B. Goethals. 2021. Pessimistic Reward Models for Off-Policy Learning in Recommendation. In *RecSys '21*.
[18] O. Jeunen, D. Rohde, F. Vasile, and M. Bompaire. 2020. Joint Policy-Value Learning for Recommendation. In *KDD '20*.
[19] N. Karlsson and Q. Sang. 2021. Adaptive Bid Shading Optimization of First-Price Ad Inventory. In *ACC '21*.
[20] D. P. Kingma and M. Welling. 2013. Auto-Encoding Variational Bayes. https://arxiv.org/abs/1312.6114
[21] Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (Aug. 2009), 30–37.
[22] X. Liu, C. Yu, Z. Zhang, Z. Zheng, Y. Rong, H. Lv, D. Huo, Y. Wang, D. Chen, J. Xu, F. Wu, G. Chen, and X. Zhu. 2021. Neural Auction: End-to-End Learning of Auction Mechanisms for E-Commerce Advertising. In *KDD '21*.
[23] Y. A. Malkov and D. A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE TPAMI* (2020).
[24] A. McDowell. 2003. From the Help Desk: Hurdle Models. *The Stata Journal* 3, 2 (2003), 178–184.
[25] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, et al. 2013. Ad click prediction: a view from the trenches. In *KDD '13*.
[26] R. B. Myerson. 1981. Optimal Auction Design. *Mathematics of Operations Research* 6, 1 (1981), 58–73.
[27] A. B. Owen. 2013. *Monte Carlo theory, methods and examples*.
[28] S. Pan, B. Kitts, T. Zhou, H. He, B. Shetty, a. Flores, D. Gligorijevic, J. Pan, T. Mao, S. Gultekin, and J. Zhang. 2020. Bid Shading by Win-Rate Estimation and Surplus Maximization. In *KDD '20 AdKDD Workshop*.
[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS '19*.
[30] D. Rohde, S. Bonner, T. Dunlop, F. Vasile, and A. Karatzoglou. 2018. RecoGym: A Reinforcement Learning Environment for the problem of Product Recommendation in Online Advertising. In *RecSys '18 REVEAL Workshop*.
[31] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. 2015. Trust Region Policy Optimization. In *ICML '15*.
[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. Proximal Policy Optimization Algorithms. https://arxiv.org/abs/1707.06347
[33] N. Si, F. Zhang, Z. Zhou, and J. Blanchet. 2020. Distributionally Robust Policy Evaluation and Learning in Offline Contextual Bandits. In *ICML '20*.
[34] M. Strathern. 1997. 'Improving ratings': audit in the British University system. *European Review* 5, 3 (1997), 305–321.
[35] Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudik. 2020. Doubly robust off-policy evaluation with shrinkage. In *ICML '20*.
[36] Y. Su, L. Wang, M. Santacatterina, and T. Joachims. 2019. CAB: Continuous Adaptive Blending for Policy Evaluation and Learning. In *ICML '19*.
[37] A. Swaminathan and T. Joachims. 2015. Batch learning from logged bandit feedback through counterfactual risk minimization. *JMLR* (2015).
[38] W. Vickrey. 1961. Counterspeculation, Auctions, and Competitive Sealed Tenders. *The Journal of Finance* 16, 1 (1961), 8–37.
[39] R. Wang, B. Fu, G. Fu, and M. Wang. 2017. Deep & Cross Network for Ad Click Predictions. In *KDD '17 AdKDD Workshop*.
[40] 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. In *CIKM '18*.
[41] X. Yang, Y. Li, H. Wang, D. Wu, Q. Tan, J. Xu, and K. Gai. 2019. Bid Optimization by Multivariable Control in Display Advertising. In *KDD '19*.
[42] W. Zhang, B. Kitts, Y. Han, Z. Zhou, T. Mao, H. He, S. Pan, A. Flores, S. Gultekin, and T. Weissman. 2021. MEOW: A Space-Efficient Nonparametric Bid Shading Algorithm. In *KDD '21*.
[43] T. Zhou, H. He, S. Pan, N. Karlsson, B. Shetty, B. Kitts, D. Gligorijevic, S. Gultekin, T. Mao, J. Pan, J. Zhang, and A. Flores. 2021. An Efficient Deep Distribution Network for Bid Shading in First-Price Auctions. In *KDD '21*.