Breaking Determinism: Stochastic Modeling for Reliable Off-Policy Evaluation in Ad Auctions

Hongseon Yeom*
NAVER Corporation
Seongnam-si, Republic of Korea
yeom.j.hs@navercorp.com

Jaeyoul Shin*
NAVER Corporation
Seongnam-si, Republic of Korea
jaeyoul.shin@navercorp.com

Soojin Min
NAVER Corporation
Seongnam-si, Republic of Korea
soojin.min@navercorp.com

Jeongmin Yoon
NAVER Corporation
Seongnam-si, Republic of Korea
jeongmin.yoon@navercorp.com

Seunghak Yu
NAVER Search US
Seattle, Washington, United States
seunghak.yu@navercorp.com

Dongyeop Kang†
NAVER Search US, University of Minnesota
United States
dongyeop@umn.edu

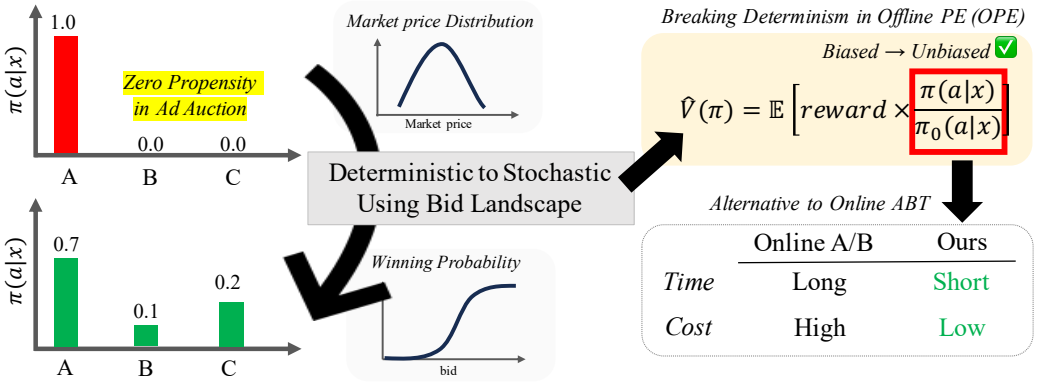

Figure 1: Breaking determinism in offline policy estimation in Ad auction via bid landscape modeling.

Abstract

Online A/B testing, the gold standard for evaluating new advertising policies, consumes substantial engineering resources and risks significant revenue loss from deploying underperforming variations. This motivates the use of Off-Policy Evaluation (OPE) for rapid, offline assessment. However, applying OPE to ad auctions is fundamentally more challenging than in domains like recommender systems, where stochastic policies are common. In online ad auctions, it is common for the highest-bidding ad to win the impression, resulting in a deterministic, winner-takes-all setting. This results in zero probability of exposure for non-winning ads, rendering standard OPE estimators inapplicable. We introduce the first principled framework for OPE in deterministic auctions by repurposing the bid landscape model to approximate the propensity score. This model allows us to derive robust approximate propensity scores, enabling the use of stable estimators like Self-Normalized Inverse Propensity Scoring (SNIPS) for counterfactual evaluation. We validate our approach on the AuctionNet simulation benchmark and against 2-weeks online A/B test from a large-scale industrial platform. Our method shows remarkable alignment with online results, achieving a 92% Mean Directional Accuracy (MDA) in CTR prediction, significantly outperforming the parametric baseline. MDA is the most critical metric for guiding deployment decisions, as it reflects the ability to correctly predict whether a new model will improve or harm performance. This work contributes the first practical and validated framework for reliable OPE in deterministic auction environments, offering an efficient alternative to costly and risky online experiments.

CCS Concepts

• **Information systems → Evaluation of retrieval results**; **Online advertising**.

Keywords

Off-Policy Evaluation;Bid Landscape;Online Ad;Online A/B

ACM Reference Format:
Hongseon Yeom, Jaeyoul Shin, Soojin Min, Jeongmin Yoon, Seunghak Yu, and Dongyeop Kang. 2026. Breaking Determinism: Stochastic Modeling for Reliable Off-Policy Evaluation in Ad Auctions. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*

*Both authors contributed equally to this research.
†Corresponding author.

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

WSDM '26, Boise, ID, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2292-9/2026/02
https://doi.org/10.1145/3773966.3778011

(WSDM '26), February 22–26, 2026, Boise, ID, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3773966.3778011

1 Introduction

Online A/B testing is the gold standard for evaluating new policies in large scale web systems, from user interface changes to new algorithmic models. However, this gold standard comes at a steep price. Deploying a new policy, even for a limited test, requires significant engineering resources and time. More critically, it carries the inherent risk of exposing users to an underperforming model, which can directly harm key business metrics such as user engagement and revenue. The ability to reliably predict a policy's performance *offline*, before deployment, is therefore not just a matter of efficiency, but a crucial component of responsible and cost-effective innovation.

Off-Policy Evaluation (OPE) offers a compelling alternative by estimating the value of a new policy using historical data collected under a different policy. While OPE has achieved notable success in domains such as recommender systems [4, 13, 16], its application to online advertising auctions presents unique challenges. Recommender systems often employ stochastic policies, where multiple items have a non-zero probability of being displayed, enabling computation of the propensity scores required by standard estimators such as Inverse Propensity Scoring (IPS). In contrast, real-time bidding (RTB) ad auctions employ a deterministic, winner-takes-all mechanism: for a given ad request, only the highest-bidding ad is shown. This creates what we term the *curse of the ad auction*: the observed propensity of any non-winning ad being shown is exactly zero, rendering standard OPE estimators mathematically inapplicable.

Limitations of existing methods. A common workaround in the deterministic OPE literature is to *relax* the policy using kernel-based smoothing [6–8], artificially introducing stochasticity by assuming a continuous probability distribution (e.g., a Gaussian kernel) around the chosen action. However, these generic methods are ill-suited for the unique characteristics of ad auctions. First, the action space selecting one winner from N discrete ad candidates lacks the natural distance metric that continuous kernels rely on. Second, and more importantly, such methods ignore the complex, often multimodal nature of the underlying market price distributions and face severe computational scalability challenges when applied to the winner determination problem in auctions.

In this paper, we propose a novel framework that sidesteps these limitations by leveraging the inherent structure of the auction mechanism. Rather than modeling the discrete, deterministic action (which ad to show), we focus on the continuous, stochastic variable governing the outcome: the *market price* (i.e., the second-highest bid). To this end, we adapt the Discrete Price Model (DPM), a technique from the bid landscape forecasting literature. While DPM was originally developed for bid optimization, we demonstrate its novel and powerful application for policy evaluation. By modeling the distribution of market prices, our framework can compute the winning probability for any given ad's bid. This probability serves as a principled and effective Approximate Propensity Score (APS), enabling the use of standard OPE estimators, otherwise intractable setting. Figure 1 provides an overview of the process.

Contributions. First, to the best of our knowledge, this is the first attempt to introduce a practical and validated framework for applying OPE in deterministic, winner-takes-all ad auction environments. Second, we demonstrate the novel use of the Discrete Price Model (DPM) to estimate effective propensity scores, providing a domain-aware, computationally tractable alternative to generic smoothing methods. Third, we rigorously validate our framework on both the AuctionNet simulation benchmark and a 2-week online A/B test in a large scale industrial platform, achieving 92% mean directional accuracy (MDA) in CTR prediction. Our results indicate that this offline evaluation framework can closely replicate online experiment outcomes, substantially reducing the cost and risk of model evaluation in advertising setups.

2 Related Works

2.1 Off-Policy Evaluation

Off-Policy Evaluation (OPE) estimates the value of a target policy π using logs generated by a different behavior policy π_0 [21, 22]. Importance sampling (IS) methods such as Inverse Propensity Scoring (IPS) [5] and its self-normalized variant SNIPS [19] correct for distributional shift via importance weights $w(x, a) = \frac{\pi(a|x)}{\pi_0(a|x)}$. These estimators rely on the *common support* assumption [12], requiring that any action chosen by π has non-zero probability under π_0. Deterministic auctions inherently violate this condition, which is the central challenge our work addresses.

2.2 OPE in Recommender Systems and Advertising Auctions

OPE has been extensively applied in recommender systems [17, 20], where stochastic logging, such as ϵ-greedy exploration or randomization of item order, ensures that most items retain a non-zero probability of exposure. This inherent stochasticity satisfies the common support assumption and enables direct use of IPS or Doubly Robust (DR) estimators, though the large action spaces can still induce high variance [17, 20].

In contrast, real-time bidding (RTB) ad auctions are inherently deterministic and *winner-takes-all*: the highest bidder wins with probability one, and all others have zero probability of being shown [1, 11]. This zero-propensity phenomenon means that standard IPS-based estimators cannot evaluate counterfactual outcomes for losing ads.

Prior work on OPE in advertising has focused on orthogonal challenges. For example, delayed feedback in conversion events has been addressed through specialized modeling [2], and Sagtani et al. [15] studied OPE for ad-load balancing, leveraging randomized logging policies to ensure full support. While effective in their respective contexts, these methods rely on policy randomization, which is often infeasible in production RTB systems due to business or performance constraints. As a result, a gap remains for approaches that operate directly on deterministic logs without altering the serving policy.

2.3 Bid Landscape Forecasting

Bid Landscape Forecasting (BLF) predicts the market price distribution $P(z|x)$ to optimize bidding strategies [3, 14]. The task is complicated by censored data when losing bids reveal only that z

exceeded the submitted price. Models such as DLF [14] and ADM [9] address this for bid optimization. Our work repurposes BLF for OPE: instead of selecting optimal bids, we use the market price distribution to compute the probability of winning for a given bid, yielding an *Approximate Propensity Score* (APS). This bridges BLF and OPE, providing a principled way to overcome the zero propensity barrier in deterministic auctions.

3 Problem Formulation

Off-Policy Evaluation (OPE) offers a promising alternative to costly online A/B tests by leveraging historical log data. However, applying standard OPE methods to deterministic environments like real-world ad auctions presents a fundamental challenge. This section formally defines the problem, illustrating why the deterministic, winner-takes-all nature of ad auctions systematically biases OPE results, motivating our need for a new framework.

3.1 Preliminaries

To formally ground our discussion, we first define the key components in the context of both OPE and online advertising auctions.

Off-Policy Evaluation. The primary goal of OPE is to estimate the expected outcome, or value $V(\pi)$, of a new evaluation policy π using a historical log dataset $\mathcal{D} = \{(x_i, a_i, r_i)\}_{i=1}^n$. This dataset was collected by a different logging policy π_0. The value of a policy is the expected reward (e.g., Click-Through Rate, CTR) it would achieve if deployed:

$$V(\pi) = \mathbb{E}_{a\sim\pi(\cdot|x),x\sim p(x)}[r(x,a)] \tag{1}$$

The most fundamental OPE estimator is the Inverse Propensity Scoring (IPS) estimator [1, 4, 5], which re-weights the observed rewards by the ratio of propensity scores between the two policies:

$$\hat{V}_{IPS}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}r_i(x_i,a_i) \tag{2}$$

Here, the ratio $w_i = \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}$ is the importance weight. For this estimator to be unbiased, a critical assumption known as common support is required: if an action a could be taken by the evaluation policy ($\pi(a|x) > 0$), it must also have a non-zero probability of being taken by the logging policy ($\pi_0(a|x) > 0$).

Online Ad Auctions.

- **Context (x)**: A feature vector representing a user, a webpage, and other situational information for an ad impression; the vector can contain both discrete and continuous features.
- **Action (a)**: The specific ad creative chosen to be displayed for the given context x. The set of available actions is the pool of candidate ads, $A(x)$.
- **Reward (r)**: An indicator of a click on the displayed ad ($r = 1$ for a click, $r = 0$ otherwise).
- **Policy (π)**: A function that, given a context x, selects an ad a from the set of candidates $A(x)$. In modern ad systems, this selection is based on a score, often a product of predicted CTR (pCTR) and a bid price.

3.2 Limitation of OPE in Deterministic World

The core challenge arises because the logging policy in a real-world ad auction is not stochastic; it is **deterministic**. The system deterministically selects the single ad with the highest score.

Let $\text{score}(a, x)$ be the score assigned to an ad a for context x. The logging policy π_0 is an 'argmax' function:

$$\pi_0(x) = a^* = \underset{a\in A(x)}{\text{argmax}}(\text{score}_0(a, x)) \tag{3}$$

This means the true probability of an ad being shown is either 1 or 0:

$$\pi_0(a|x) = \begin{cases} 1 & \text{if } a = a^* \\ 0 & \text{if } a \neq a^* \end{cases} \tag{4}$$

Now, consider an evaluation policy π which is also deterministic, based on a new scoring model $\text{score}_{\text{eval}}$:

$$\pi(x) = a' = \underset{a\in A(x)}{\text{argmax}}(\text{score}_{\text{eval}}(a, x)) \tag{5}$$

When we apply the IPS estimator, we only consider the data points where an ad was shown ($a_i = a^*$). For these logged events, the importance weight w_i can only take two values:

$$w_i = \frac{\pi(a^*|x_i)}{\pi_0(a^*|x_i)} = \frac{\pi(a^*|x_i)}{1} = \begin{cases} 1 & \text{if } \pi \text{ also chooses } a^* \\ 0 & \text{if } \pi \text{ chooses some } a' \neq a^* \end{cases} \tag{6}$$

This leads to a critical limitation, as pointed out by previous works on off-policy evaluation in settings with limited support (e.g., [1, 19]). The IPS estimator simplifies to:

$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\pi(x_i) = \pi_0(x_i)\}\, r_i \;\leq\; \frac{1}{n}\sum_{i=1}^{n} r_i = \hat{V}(\pi_0), \tag{7}$$

Under deterministic logging, the estimated value of the evaluation policy cannot exceed that of the logging policy; equality holds only when π agrees with π_0 on all impressions. Any disagreement is counted as a zero contribution, so potential improvements from the new policy are never credited, and $\hat{V}_{\text{IPS}}(\pi)$ systematically underestimates the true value $V(\pi)$. This severe limitation necessitates a new approach that can look beyond the logged deterministic outcome and assign a meaningful, non-zero probability to an ad winning the auction.

4 Proposed Method: DPM-OPE

In the previous section, we established that the deterministic, winner-takes-all nature of online ad auctions renders standard OPE methods systematically biased.

To overcome this, we introduce **DPM-OPE**, a novel framework for Off-Policy Evaluation. Instead of attempting to stochastically approximate the deterministic policy directly, our approach fundamentally shifts the perspective to the inherent stochasticity of the *auction environment* itself. The central concept is to model the competitive landscape that an advertisement encounters. By estimating the probability of an ad winning the auction given its score, we can derive a robust non-zero *Approximate Propensity Score (APS)*. This APS serves as a valid substitute for the propensity score, enabling the use of powerful OPE estimators for reliable offline evaluation. Our framework is illustrated in Figure 2.

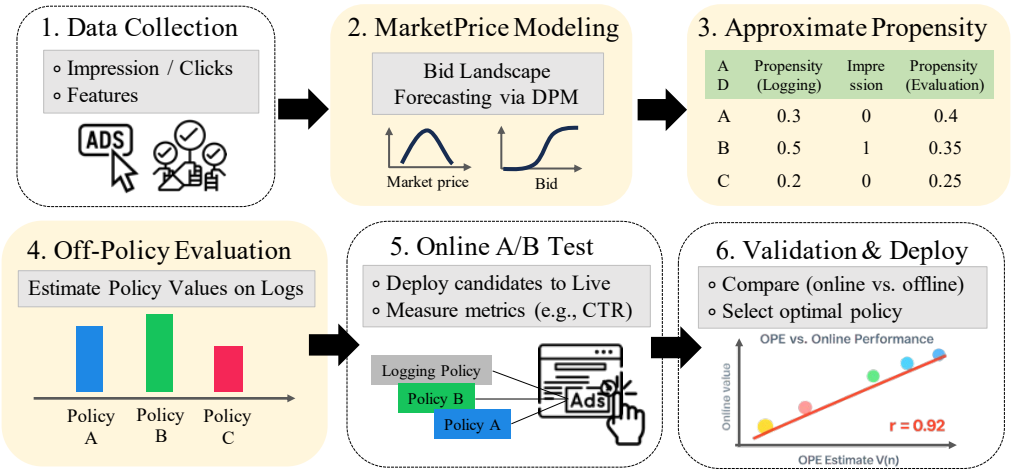

Figure 2: An overview of the end-to-end DPM-OPE framework: the technical steps 2, 3, and 4 (yellow-shaded) are described in Section 4, and the remaining steps are described in Section 5.

4.1 A Shift in Perspective: Modeling the Auction Environment

Our perspective shifts from modeling the policy's choice, $\pi(a|x)$, to modeling the competitive environment where that choice occurs. In an ad auction, an ad's exposure is not a direct choice but the outcome of a competition. Therefore, the probability of an ad being shown equals its probability of winning that competition. This reframes the problem: instead of asking *"What is the probability of the policy choosing ad a?"*, we ask, ***"Given the competitive environment, what is the probability of ad a winning?"*** (Figure 3). This environment-centric view allows us to capture the market's inherent uncertainty.

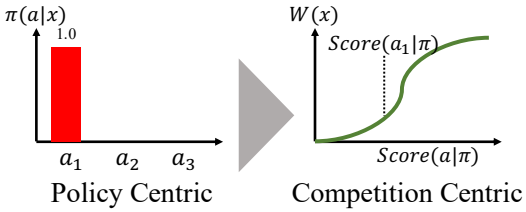

Figure 3: Reframing the Question: From Policy Choice to Winning Probability

4.2 Estimating the Market Price Distribution

To estimate the winning probability, we first represent the entire set of competitors in an auction by a single random variable: the market price, defined as the second-highest score among all other ads [14, 23–25].

4.2.1 The Market Price and Winning Condition. An ad a_i with a score $s_i = \text{score}(a_i, x)$ wins an auction if its score is higher than that of all other competing ads. It is important to note that the formulation of score$(\cdot)$ varies depending on the auction mechanism. While it may correspond directly to a raw bid price in classic auctions, modern ad systems typically utilize a composite value, such as bid × pCTR, to rank candidates.

In such environments, the determining factor for winning is this composite score. Consequently, the "effective price" one must beat to win the auction is the highest composite score among competitors. We therefore generalize the definition of market price z as the highest score among all other competitors for a given impression:

$$z = \max_{j \neq i} \{\text{score}(a_j, x)\} \tag{8}$$

The winning condition for ad a_i is simply $s_i > z$. Since the set of competitors and their scores vary for each auction, the market price z is a random variable. The probability of an ad with score s winning is therefore $P(s > z)$, which is the cumulative distribution function (CDF) of the market price, often denoted as the winning function $W(s) = \int_0^s p(z)dz$. Our primary task is to model the probability distribution of the market price, $p(z)$.

4.2.2 The Discrete Price Model (DPM). We employ a Discrete Price Model (DPM), a technique well-suited for this task and widely used in bid landscape forecasting [9, 14]. Since prices and scores in real-time bidding are inherently discrete due to finite precision, modeling them in a discrete space is a natural and effective choice [10].

We discretize the continuous score space into L disjoint intervals, or bins, $V_l = (b_{l-1}, b_l]$ for $l = 1, \ldots, L$. This allows us to define the winning function $W(b_l)$ (the probability of winning with a score up to b_l) and the survival function $S(b_l)$ (the probability of losing with a score of b_l) over this discrete space:

$$\begin{aligned} W(b_l) &\doteq \Pr(z < b_l) = \sum_{j<l} \Pr(z \in V_j) \\ S(b_l) &\doteq \Pr(z \geq b_l) = \sum_{j \geq l} \Pr(z \in V_j) \end{aligned} \tag{9}$$

The probability of the market price z falling into a specific interval V_l is then given by:

$$p_l = \Pr(z \in V_l) = S(b_{l-1}) - S(b_l) \tag{10}$$

4.2.3 Adaptive Binning. Choosing the number of bins L is crucial for balancing bias and variance in DPM. To ensure robust estimation, we determine the number of bins, L, *adaptively from the data*. This process guarantees that the winning-rate estimate for each bin satisfies a predefined confidence interval width.

Setup. Let n be the number of data points. We construct *quantile bins* over the score $s = \text{score}(a, x)$ so that each bin has approximately n/L samples. Under a normal approximation to the binomial proportion p in a bin, the two-sided $100 \times (1-\alpha)\%$ confidence interval has half-width bounded by

$$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n/L}} \;\le\; \varepsilon, \tag{11}$$

where $z_{\alpha/2}$ is the standard normal quantile and $\varepsilon > 0$ is the target accuracy.

Rule for L. We adopt the worst-case variance $p(1-p) \le \frac{1}{4}$ and tie the accuracy to the discretization granularity by $\varepsilon = \frac{1}{2L}$. Substituting into (11) yields

$$z_{\alpha/2}\sqrt{\frac{0.25\,L}{n}} \;\le\; \frac{1}{2L} \quad\Longrightarrow\quad L^3 \;\le\; \frac{n}{z_{\alpha/2}^2}.$$

For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$ and $z_{\alpha/2}^2 = 3.8416$, which gives the closed-form sizing rule

$$L^\star \;=\; \left\lfloor \left(\frac{n}{3.8416}\right)^{1/3} \right\rfloor. \tag{12}$$

L as the maximum. The stochastic approximation introduces *quantization bias*: using fewer bins mixes heterogeneous scores within a bin and flattens the winning function $W(s)$, thereby biasing the approximate propensities. The rule in (12) identifies the *largest* L that still satisfies the per-bin accuracy constraint; selecting $L = L^\star$ (subject to latency/memory caps) therefore minimizes discretization bias *under the variance bound* implied by (11).

Segmentation rationale. We apply adaptive binning *per segment* g (e.g., per period in AuctionNet) because the score distribution can vary substantially across segments. To select a meaningful segmentation feature, we evaluate ANOVA explained variance on scores:

$$\text{SST} \;=\; \sum_g \sum_{i\in g} (s_i - \bar{s})^2, \qquad \text{SSB} \;=\; \sum_g n_g\,(\bar{s}_g - \bar{s})^2, \qquad R^2 \;=\; \frac{\text{SSB}}{\text{SST}},$$

where $\bar{s}$ is the global mean, $\bar{s}_g$ is the group mean, and n_g is the group size. We choose the segmentation that maximizes R^2, indicating that between-group differences explain a large portion of the total variance. After fixing the segmentation, we fit DPM and compute $L_g^\star$ per segment using (12), then form quantile bins $V_\ell = (b_{\ell-1}, b_\ell]$ within each segment.

4.3 Approximate Propensity Score (APS)

With the DPM framework established, we can derive a meaningful propensity score. While the cumulative winning probability $W(b_l)$ tells us the chance of beating the market, it does not represent the probability of a specific score being the winning one. For an importance weight, we need a value analogous to a probability mass function (PMF), not a cumulative distribution function (CDF).

To this end, we use the conditional winning probability, h_l, as our APS. This is defined as the probability that the market price falls within a specific interval V_l, given that it was not in any of the lower intervals:

$$h_l = \Pr(z \in V_l | z \ge b_{l-1}) = \frac{\Pr(z \in V_l)}{\Pr(z \ge b_{l-1})} = \frac{p_l}{S(b_{l-1})} \tag{13}$$

The term h_l represents the probability of just winning the auction with a score in the interval V_l. Therefore, we define the APS for an action a with score $s \in V_l$ as:

$$\pi_{\text{APS}}(a|x) \doteq h_l \quad \text{where score}(a, x) \in V_l \tag{14}$$

This APS is a non-zero, granular probability that reflects the ad's likelihood of being the marginal winner, effectively creating the stochastic approximation we need.

We acknowledge that the accuracy of our framework is contingent on the quality of the estimated market price distribution, $P(z)$. Any misestimation in this step could potentially introduce bias into the final OPE result.

To mitigate this, we deliberately chose the DPM for its non-parametric flexibility, which allows it to capture complex, real-world distributions more faithfully than simpler parametric models. While our empirical results in Section 6 demonstrate that this approach is practically effective, we believe that developing more advanced distribution modeling techniques is a valuable direction for future research. Exploring methods that can even more precisely model the nuances of the market price would be a promising step toward further enhancing the accuracy of OPE in deterministic environments.

4.4 DPM-OPE Estimator

Having derived a valid, non-zero APS, we can now substitute it into a standard OPE estimator in place of $\pi_0(a|x)$ and $\pi(a|x)$. To mitigate the high variance often associated with the basic IPS estimator, we use the Self-Normalized Inverse Propensity Scoring (SNIPS) estimator [19], which is known for its stability.

By replacing the problematic $\pi(a_i|x_i)$ with our derived $\text{APS}(a_i|x_i)$, we arrive at our final DPM-SNIPS estimator:

$$\hat{V}_{DPM-SNIPS}(\pi) = \frac{\sum_{i=1}^n r_i \frac{\pi_{\text{APS}}(a_i|x_i)}{\pi_{0\,\text{APS}}(a_i|x_i)}}{\sum_{i=1}^n \frac{\pi_{\text{APS}}(a_i|x_i)}{\pi_{0\,\text{APS}}(a_i|x_i)}} \tag{15}$$

This estimator allows us to reliably evaluate new policies π using historical data from a deterministic logging environment, bridging the gap that previously made such offline evaluation impractical.

5 Experiments

In this section, we present a series of experiments designed to validate the effectiveness and robustness of our proposed DPM-OPE framework. We aim to answer the following key questions:

(1) How accurately does DPM-OPE estimate the performance of new policies compared to the baseline in a controlled simulation environment?
(2) Does the effectiveness of DPM-OPE hold in a large-scale, real-world industrial setting?
(3) How sensitive is our method to its core hyperparameters?

5.1 Datasets & Setup

To thoroughly evaluate our framework, we utilized two distinct datasets: a public benchmark for controlled simulation and a large-scale industrial dataset for real-world validation.

5.1.1 AuctionNet Benchmark. Since there is no public benchmark specifically designed for OPE in ad auctions, we adapted **AuctionNet**[18], a well-established benchmark for bidding decisions derived from a real-world advertising platform. To create an OPE testbed, we independently ran simulations for seven different bidding strategies, generating separate datasets for each.

We designated the "PID" bidding strategy as our logging policy (π_0) and the remaining six strategies ("BC", "BCQ", "CQL", "IQL", "OnlineLP", "TD3-BC") as the evaluation policies (π) to be assessed via OPE. The ground truth performance for each evaluation policy was obtained from its corresponding simulation run. Key statistics for each simulation are presented in Table 1.

Table 1: Ground truth statistics for each policy simulation on the AuctionNet benchmark.

Policy	Impressions	Conversions	CTR
pid (π_0)	9,679,017	10,317	0.001065
bc	9,999,526	14,348	0.001434
bcq	9,999,526	15,107	0.001510
cql	9,999,526	14,730	0.001473
iql	9,999,526	14,406	0.001440
onlinelp	9,999,526	14,718	0.001471
td3_bc	9,999,526	14,639	0.001463

When applying our *adaptive binning* strategy to the AuctionNet benchmark, the resulting number of bins L ranged from approximately 48 to 50 across different segments (period), balancing estimation resolution with statistical reliability.

5.1.2 Real-World Data. To demonstrate the practical applicability of our framework, we collected a 2-week log from a large-scale online advertising platform at NAVER. This dataset contains the results of an online A/B test where two different CTR prediction models were deployed. The environment involves selecting one ad from a multitude of candidates for each ad slot. This extensive dataset provides a challenging and realistic testbed for evaluating our method's performance in an industrial setting.

5.2 Evaluation Metrics

A key goal in business contexts is not just to measure the absolute performance of a new model, but to accurately assess its *relative improvement* over the current production model. Therefore, all our metrics are based on the CTR lift (or difference) of the evaluation policy relative to the logging policy.

- **MDA (Mean Directional Accuracy)**: Measures the percentage of times our offline estimate correctly predicts the direction (improvement or degradation) of the CTR change. High MDA ensures stable day-to-day decision-making and *minimizes the risk of deploying underperforming models.*
- **RMSE (Root Mean Square Error)**: Measures the average of the squares of the error between the ground truth CTR lift and the lift estimated by OPE. We report these in percentage points (%-points) for intuitive comparison.
- **Pearson Correlation**: Measures the linear relationship between the daily performance trends of the ground truth and the OPE estimates, indicating how well our offline metric tracks online performance fluctuations.

Predicting the correct direction of performance change (i.e., whether CTR will improve or degrade) is more important than minimizing numerical error or maximizing correlation. Accordingly, we prioritize our metrics in the following order: **MDA** > **RMSE** > **Pearson Correlation**.

5.3 Baseline

We compare our DPM-OPE framework against a strong, domain-aware baseline. Before introducing it, we first justify the exclusion of a more generic class of methods for deterministic policies: kernel-based smoothing [6–8]. This common technique "relaxes" a deterministic choice by placing a continuous probability kernel (e.g., Gaussian) around the chosen action, thereby creating a synthetic stochastic policy to enable OPE.

However, we argue that this generic approach is fundamentally misaligned with the mechanics of our ad auction problem for two primary reasons. First, the action space consists of a discrete set of competing ad candidates, which lacks the natural distance metric that continuous kernels rely on. Second, and more critically, such methods are agnostic to the underlying auction dynamics. They ignore the complex, often multimodal market price distribution that truly governs the winning probability.

Instead of artificially smoothing the action space, our work posits that a more principled approach is to directly model the source of the environmental stochasticity, the market price. Therefore, we select a baseline that adheres to this principle:

- **Parametric-OPE**: This baseline assumes that the market price distribution follows a known parametric form. For each auction, we fit the observed bid prices of competing ads to several distributions (Normal, Log-Normal, Beta, Gamma, Exponential) and select the one with the best fit. The APS is then calculated from the Cumulative Distribution Function (CDF) of the chosen parametric distribution.

5.4 Analysis of Framework Components

To validate the specific design choices within our DPM-OPE framework, we designed a series of analytical experiments. This section outlines the setup for these studies, which investigate the impact of two key components: (1) the DPM's binning strategy and (2) the final OPE estimator. The results of these comparisons will be presented in detail in Section 6.

5.4.1 Impact of Binning Strategy. The discretization of the score space into bins is a core component of the DPM. To understand its impact, we will compare two distinct quantile-based strategies:

- **Static Binning**: This approach uses quantile binning to ensure each bin contains an equal number of samples, but the total number of bins (L) is a fixed hyperparameter (e.g.,

Table 2: Performance comparison on the AuctionNet benchmark and the 2-week real-world dataset. (↓ Lower is better, ↑ Higher is better)

Metric	AuctionNet		Real-World (2-week)	
	Parametric	DPM-OPE	Parametric	DPM-OPE
MDA (↑)	100.0%	100.0%	78.6%	**92.9%**
RMSE (↓)	25.585	**4.870**	**1.187**	1.197
Pearson Correlation (↑)	**0.520**	-0.112	0.575	**0.653**

100, 1000). While effective, the choice of L requires manual tuning and may not be optimal across different datasets.

- **Adaptive Binning (Our Method)**: This strategy determines the number of bins L adaptively from the data. The goal is to set L such that the winning-rate estimate within each bin achieves a target statistical precision at a prescribed confidence level. This removes the need to manually specify the number of bins.

5.4.2 Choice of OPE Estimator. While our primary contribution is the derivation of the Approximate Propensity Score (APS), the choice of the final estimator that consumes this score is crucial for stability. We will compare three variants based on the Inverse Propensity Scoring (IPS) family:

- **IPS**: The standard, unbiased estimator. However, it is known to suffer from high variance, especially if some importance weights are large.
- **SNIPS (Self-Normalized IPS)**: This estimator normalizes the weights, which is known to reduce variance and improve stability compared to the standard IPS.
- **Capped SNIPS (Our Final Choice)**: Even with SNIPS, a few outlier events with extremely small APS values can lead to excessively large importance weights, destabilizing the final estimate. To address this, we employ a common technique of capping the importance weights at the 99th percentile of their distribution. This trades a small amount of potential bias for a significant reduction in variance.

6 Results

6.1 Simulation Results: AuctionNet

Table 2 and Figure 4 summarize the performance on the AuctionNet benchmark. Our DPM-OPE method demonstrates a clear advantage in estimation accuracy, achieving a substantially lower RMSE than the parametric baseline. This indicates that DPM-OPE predicts the true CTR lift with significantly higher precision, which is the primary goal of offline evaluation.

While the parametric baseline shows a higher Pearson Correlation, this metric should be interpreted with caution in this specific simulation. The ground truth CTR differences between several evaluation policies are not statistically significant (e.g. *BC vs IQL or BC vs TD3-BC*). Consequently, a high correlation derived from the ranking of these near identical outcomes is less indicative of true performance than the absolute error metrics, where our method clearly excels.

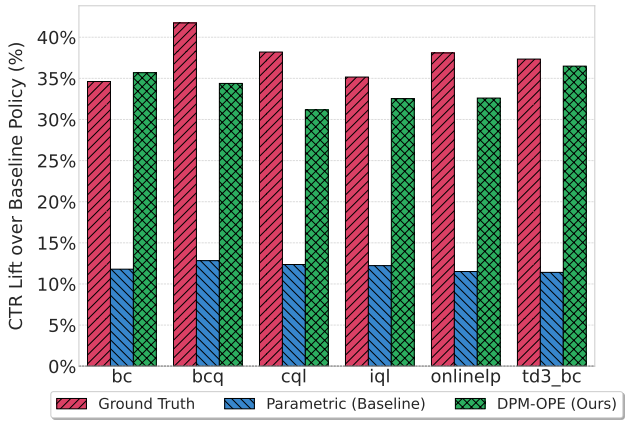

Figure 4: Comparison of estimated CTR lift vs. ground truth on AuctionNet. DPM-OPE's estimates (green bars) are consistently closer to the ground truth (red bars) than the parametric baseline's estimates (blue bars).

6.2 Real-world Online AB Testing: NAVER Platform

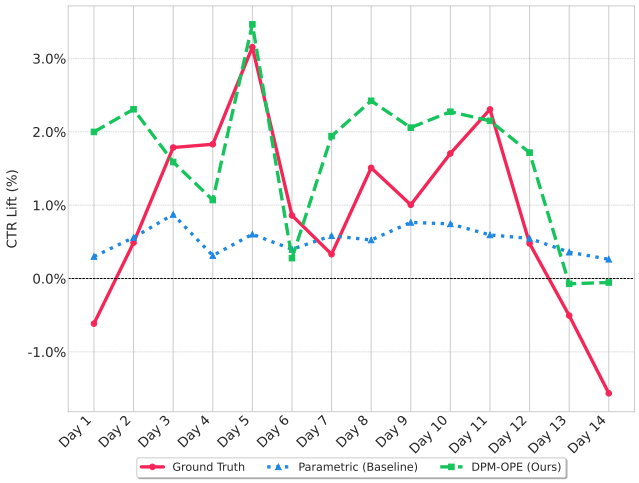

Figure 5: Daily trend comparison on real-world data. The DPM-OPE estimates (green squares) closely follow the ground truth online A/B test results (red circles), while the baseline (blue triangles) fails to capture the trend.

The results on the 2-week real-world dataset, shown in Table 2 and Figure 5, further highlight the practical advantages of our framework. While the parametric baseline achieved a marginally lower RMSE, the difference between the RMSE of the parametric baseline and that of our DPM-OPE method was found to be statistically insignificant (p=0.9751 via a paired t-test at a 95 % confidence level). As mentioned in Section 5.2, metrics that directly inform business decisions are often more critical in real-world environments. Specifically, DPM-OPE achieved a much higher MDA (92.9% vs. 78.6%), a 14.3 percentage point improvement. This demonstrates a superior ability to make correct directional predictions about whether a new model will improve or harm performance, which is paramount for practical deployment.

Furthermore, the significantly higher Pearson Correlation (0.653 vs. 0.575) shows that our method more reliably tracks the day-to-day fluctuations of the online A/B test, as visualized in Figure 5. For industrial applications, this trend-following capability is often more valuable than marginal gains in average error, as it provides a more trustworthy signal for monitoring and analysis. These results strongly suggest that DPM-OPE offers a more robust and practical solution for real-world offline evaluation.

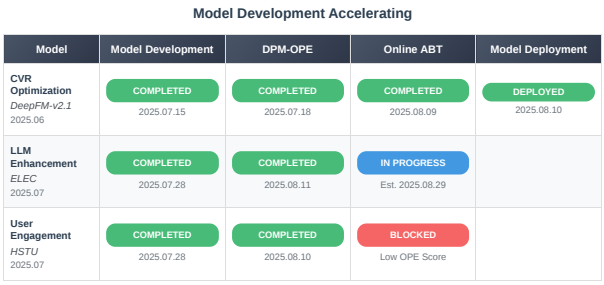

Figure 6: Industrial adoption of DPM-OPE showing its role in accelerating model development.

Beyond controlled experiments, DPM-OPE has seen industrial adoption in a large-scale advertising platform, where it is routinely used to pre-screen candidate models offline before any online A/B testing. In one case, a third candidate model with a low OPE score bypassed the ABT stage entirely, avoiding the cost of an unnecessary online experiment. This practice has measurably reduced the volume of costly online experiments and accelerated the overall model development cycle, further validating the framework's practical value in real-world settings (Figure 6).

6.3 Ablation Studies

To validate the specific design choices within our DPM-OPE framework, we conducted further analysis on the AuctionNet dataset. We investigated the impact of two key components: the DPM's binning strategy and the final OPE estimator. As mentioned in Section 6.1, the Pearson correlation for AuctionNet was not statistically significant, so it was excluded from the ablation studies.

6.3.1 Impact of Binning Strategy. Table 3 and Figure 7 compare our proposed adaptive binning strategy against static quantile binning with a fixed number of bins (100, 1,000, and 10,000). Our adaptive binning approach achieves the lowest RMSE, providing the most stable and accurate estimates. This result validates our choice of this method for robust performance without manual hyperparameter tuning.

Table 3: Metrics per binning strategy on AuctionNet.

Method	RMSE (↓)	MDA (↑)
Bin=100	5.509	100 %
Bin=1,000	6.461	100 %
Bin=10,000	18.504	100 %
Bin=Adaptive	**4.870**	100 %

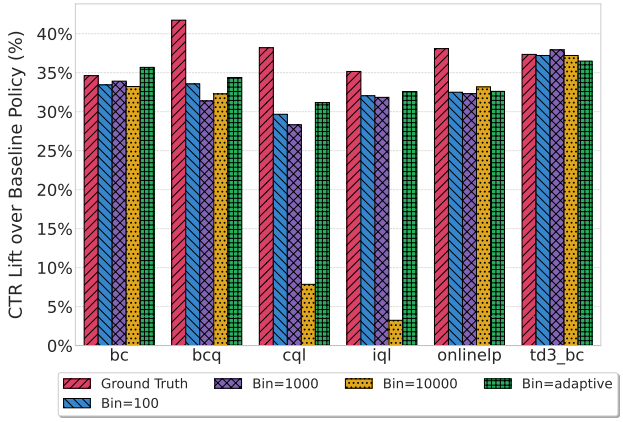

Figure 7: Performance comparison of different binning strategies on AuctionNet.

Table 4: Metrics per OPE estimator on AuctionNet.

Estimator	RMSE (↓)	MDA (↑)
IPS	355.532	100 %
SNIPS	20.448	100 %
Capped SNIPS	**4.870**	100 %

6.3.2 Choice of OPE Estimator. Table 4 and Figure 8 illustrate the importance of the estimator choice. The standard IPS estimator suffers from extremely high variance, making its estimates unusable in practice. Although SNIPS offers a significant improvement, its error rates remain high. Our final choice, Capped SNIPS, significantly reduces the variance by clipping outlier importance weights. This results in the lowest RMSE by a large margin, confirming that capping is an essential step to achieve a stable and reliable OPE framework.

7 Discussion and Future Directions

In this paper, we addressed the critical challenge of performing Off-Policy Evaluation in deterministic, winner-takes-all ad auction environments, a setting where standard estimators fail due to the zero propensity problem. We introduced DPM-OPE, a framework that transforms the deterministic environment into a stochastic

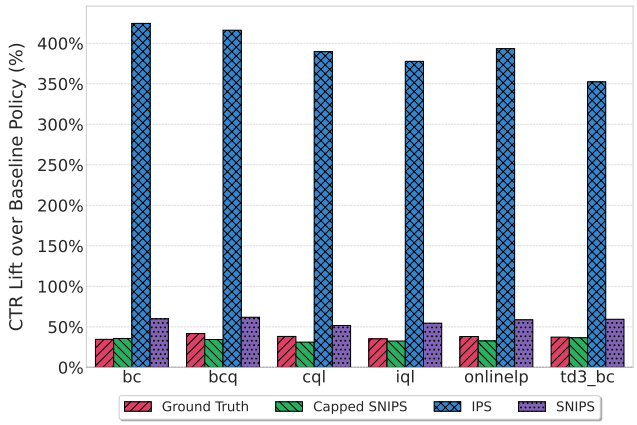

Figure 8: Performance comparison of different OPE estimators on AuctionNet.

one by modeling the market price distribution, producing a robust Approximate Propensity Score (APS).

Our contributions are both methodological and practical:

- **the first validated application of DPM to deterministic-auction OPE**, by repurposing the DPM from bid landscape forecasting for OPE. Unlike its original use for future bid optimization, we employ DPM to estimate the probability of *past* actions, enabling OPE where it was previously inapplicable. To our knowledge, this is the first comprehensive, validated solution to the zero propensity problem in online advertising logs.
- **strong empirical validation on both public benchmark and industrial AB test**, achieving higher directional accuracy and better trend tracking than the baseline.
- This work is not only a **validated and deployable solution for the advertising industry**, but also a **truly practical contribution** that can be easily and flexibly adapted to other industries, operating under deterministic policies, such as logistic routing and deterministic ranking systems.

Despite its effectiveness, DPM-OPE has the following limitations and possible directions for future work:

- The current DPM treats the market price distribution as dependent solely on aggregated scores, ignoring other contextual features (e.g., ad type, user attributes). More advanced, feature-aware models such as Deep Landscape Forecasting (DLF) or Arbitrary Distribution Modeling (ADM), could improve personalization and accuracy.
- The framework's accuracy hinges on how well the market price distribution $P(z)$ is estimated. Misestimation can introduce bias into the final OPE results. While DPM mitigates this through non-parametric flexibility, further theoretical study is needed to quantify and control error propagation.
- Although adaptable to other deterministic-policy domains, applying DPM-OPE outside ad auctions may require domain-specific calibration, particularly for defining an equivalent to "market price." Also, our approach focuses on propensity correction rather than reward modeling. Combining APS estimation with model-based or doubly robust estimators could potentially reduce variance further.

References

[1] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* 14, 1 (2013), 3207–3260.

[2] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. *Advances in neural information processing systems* 24 (2011).

[3] Aritra Ghosh, Saayan Mitra, Somdeb Sarkhel, Jason Xie, Gang Wu, and Viswanathan Swaminathan. 2019. Scalable bid landscape forecasting in real-time bidding. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 451–466.

[4] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline a/b testing for recommender systems. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 198–206.

[5] Daniel G Horvitz and Donovan J Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47, 260 (1952), 663–685.

[6] Nathan Kallus and Masatoshi Uehara. 2020. Doubly robust off-policy value and gradient estimation for deterministic policies. *Advances in Neural Information Processing Systems* 33 (2020), 10420–10430.

[7] Haanvid Lee, Tri Wahyu Guntara, Jongmin Lee, Yung-Kyun Noh, and Kee-Eung Kim. 2024. KERNEL METRIC LEARNING FOR IN-SAMPLE OFF-POLICY EVALUATION OF DETERMINISTIC RL POLICIES. In *12th International Conference on Learning Representations, ICLR 2024*. International Conference on Learning Representations, ICLR.

[8] Haanvid Lee, Jongmin Lee, Yunseon Choi, Wonseok Jeon, Byung-Jun Lee, Yung-Kyun Noh, and Kee-Eung Kim. 2022. Local metric learning for off-policy evaluation in contextual bandits with continuous actions. *Advances in Neural Information Processing Systems* 35 (2022), 3913–3925.

[9] Xu Li, Michelle Ma Zhang, Zhenya Wang, and Youjun Tong. 2022. Arbitrary distribution modeling with censorship in real-time bidding advertising. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3250–3258.

[10] Mengjuan Liu, Zhengning Hu, Zhi Lai, Daiwei Zheng, and Xuyun Nie. 2022. Real-time bidding strategy in display advertising: An empirical analysis. *arXiv preprint arXiv:2212.02222* (2022).

[11] Mehryar Mohri and Andrés Munoz Medina. 2016. Learning algorithms for second-price auctions with reserve. *Journal of Machine Learning Research* 17, 74 (2016), 1–25.

[12] Yusuke Narita, Kyohei Okumura, Akihiro Shimizu, and Kohei Yata. 2023. Counterfactual learning with general data-generating policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9286–9293.

[13] Yusuke Narita, Shota Yasui, and Kohei Yata. 2021. Debiased off-policy evaluation for recommendation systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 372–379.

[14] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, and Yong Yu. 2019. Deep landscape forecasting for real-time bidding advertising. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 363–372.

[15] Hitesh Sagtani, Madan Gopal Jhawar, Rishabh Mehrotra, and Olivier Jeunen. 2024. Ad-load balancing via off-policy learning in a content marketplace. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 586–595.

[16] Yuta Saito, Himan Abdollahpouri, Jesse Anderton, Ben Carterette, and Mounia Lalmas. 2024. Long-term off-policy evaluation and learning. In *Proceedings of the ACM Web Conference 2024*. 3432–3443.

[17] Yuta Saito, Takuma Udagawa, Haruka Kiyohara, Kazuki Mogi, Yusuke Narita, and Kei Tateno. 2021. Evaluating the robustness of off-policy evaluation. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 114–123.

[18] Kefan Su, Yusen Huo, Zhilin Zhang, Shuai Dou, Chuan Yu, Jian Xu, Zongqing Lu, and Bo Zheng. 2024. Auctionnet: A novel benchmark for decision-making in large-scale games. *Advances in Neural Information Processing Systems* 37 (2024), 94428–94452.

[19] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems* 28 (2015).

[20] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. 2017. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems* 30 (2017).

[21] Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. 2022. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355* (2022).

[22] Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. 2019. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854* (2019).
[23] Wush Wu, Mi-Yen Yeh, and Ming-Syan Chen. 2018. Deep censored learning of the winning price in the real time bidding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2526–2535.
[24] Wush Chi-Hsuan Wu, Mi-Yen Yeh, and Ming-Syan Chen. 2015. Predicting winning price in real time bidding with censored data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1305–1314.
[25] Weinan Zhang, Tianxiong Zhou, Jun Wang, and Jian Xu. 2016. Bid-aware gradient descent for unbiased learning with censored data in display advertising. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 665–674.