Pairwise Ranking Losses of Click-Through Rates Prediction for Welfare Maximization in Ad Auctions

Boxiang Lyu [1] **Zhe Feng** [2] **Zachary Robertson** [3] **Sanmi Koyejo** [2 3]

Abstract

We study the design of loss functions for click-through rates (CTR) to optimize (social) welfare in advertising auctions. Existing works either only focus on CTR predictions without consideration of business objectives (e.g., welfare) in auctions or assume that the distribution over the participants' expected cost-per-impression (eCPM) is known a priori, then use various additional assumptions on the parametric form of the distribution to derive loss functions for predicting CTRs. In this work, we bring back the welfare objectives of ad auctions into CTR predictions and propose a novel weighted rankloss to train the CTR model. Compared to existing literature, our approach provides a provable guarantee on welfare but without assumptions on the eCPMs' distribution while also avoiding the intractability of naively applying existing learning-to-rank methods. Further, we propose a theoretically justifiable technique for calibrating the losses using labels generated from a teacher network, only assuming that the teacher network has bounded ℓ_2 generalization error. Finally, we demonstrate the advantages of the proposed loss on synthetic and real-world data.

1. Introduction

Global online advertising spending is expected to exceed $700 billion in 2023 (Statista, 2022). At the core of online advertising are advertising (ad) auctions, held billions of times per day, to determine which advertisers get the opportunity to show ads (Jeunen, 2022). A critical component of these auctions is predicting the click-through rates (CTR) (Yang and Zhai, 2022). Typically, advertisers submit cost-per-click (CPC) bids, i.e., report how much they are willing to pay if a user clicks. The CTR is the probability that a user clicks the ad when the ad is shown. Combined with the cost-per-click bid, the platform can then calculate the value of *showing* the ad, usually called the cost-per-impression (eCPM). As the CTR needs to be learned, the platform instead uses the predicted click-through rates (pCTRs) to convert the submitted CPC bids to predicted eCPM bids, which then determine the auctions' outcomes.

Due to the importance of predicting the CTRs, a wealth of related literature exists, and we refer interested reader to Choi et al. (2020); Yang and Zhai (2022) for thorough reviews of these advances. Of these works, the majority focus on the various neural network architectures designed for the task, such as DeepFM (Guo et al., 2017), Deep & Cross Network (DCN) (Wang et al., 2017), MaskNet (Wang et al., 2021), among many others. These works propose novel neural network architectures but train these networks using off-the-shelf classification losses with no guarantees on the actual economic performance of the ad auctions, creating a discrepancy between the upstream model training for CTR prediction and downstream model evaluation.

Some works aim to ameliorate these discrepancies by using business objectives such as social welfare (or welfare for short) to motivate the design of loss functions (Chapelle, 2015; Vasile et al., 2017; Hummel and McAfee, 2017). However, these works either lack reproducible experiments on publicly available real-world benchmarks (Hummel and McAfee, 2017), or depend on ad-hoc heuristics with insufficient theoretical guarantees (Vasile et al., 2017). Moreover, many of these works suffer from an unrealistic assumption that bidders submit eCPM bids and the eCPM of the highest competing bid follows a known and fixed distribution. However, in real life, some ad auctions at industry leaders such as Amazon, Meta, and Google only accept CPC bids (Amazon Ads, 2023; Meta Business Help Center, 2023; Google Ads Help, 2023), and adjustments to the CTR prediction model changes the distribution of competing bids' eCPM.

We avoid the pitfalls of existing works by limiting assumptions about the eCPMs' distribution. Since various types of ad auctions with drastically different revenue functions are widely deployed, ranging from Generalized Second

[1]The University of Chicago Booth School of Business, Chicago, IL USA [2]Google Research, Mountain View, CA USA [3]Department of Computer Science, Stanford University, Stanford, CA USA. Correspondence to: Boxiang Lyu <blyu@chicagobooth.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Price (Edelman et al., 2007) to Vickrey-Clarke-Groves (Varian and Harris, 2014), and first price auction (Conitzer et al., 2022), we focus on maximizing the welfare achieved by these auctions, which measures the efficiency of the ad auction in terms of showing the most valuable ads.

Our Contributions. We list our contributions below.

- We propose a learning-to-rank loss with welfare guarantees by drawing a previously underutilized connection between welfare maximization and ranking.
- We propose two surrogate losses that are easy to optimize and theoretically justifiable.
- Inspired by student-teacher learning (Hinton et al., 2015), we construct an approximately calibrated, easy-to-optimize surrogate, whose theoretical guarantees only depend on the ℓ_2-generalization bound of the teacher network.
- We demonstrate the benefits of the proposed losses on both simulated data and the Criteo Display Advertising Challenge dataset[1], arguably the most popular benchmark for CTR prediction in ad auctions.

1.1. Related Works

In this section, we divide the related works into three main categories: applied research in CTR prediction, theoretical analysis of ad auctions, and methods in learning-to-rank.

Applied Research in CTR Prediction. There is an abundance of application oriented literature on CTR prediction (McMahan et al., 2013; Chen et al., 2016; Cheng et al., 2016; Zhang et al., 2016; Qu et al., 2016; Juan et al., 2017; Lian et al., 2018; Zhou et al., 2018; 2019; Wang et al., 2021; Pi et al., 2019; Pan et al., 2018; Li et al., 2020; Chapelle, 2015), and we refer interested readers to Yang and Zhai (2022) for a detailed survey. Two works with well-documented performance on the Criteo dataset are Guo et al. (2017) and Wang et al. (2017). Particularly, Guo et al. (2017) proposes DeepFM, short for deep factorization machines, which combines deep learning with factorization machines. Wang et al. (2017) is similar, where the proposed Deep Cross Network model combines deep neural networks with cross features. These works focus on the development of neural network architectures and use classification losses with little to no theoretical guarantees. Our work is orthogonal to and complements this line of literature by proposing easy-to-optimize loss functions rooted in economic intuition with provable guarantees on economic performance.

A well-known technique in knowledge distillation is student-teacher learning (Hinton et al., 2015), where a smaller network is used to approximate the predictions of a larger one. Recently some attempts have been made at applying the technique in CTR prediction (Zhu et al., 2020) and, as we demonstrate in this manuscript, the technique can even benefit the design of welfare-inspired loss functions, in addition to reducing the computation and memory requirements of the teacher network itself.

Among this line of work, two papers are closer to ours in spirit. Chapelle (2015) studies the design of CTR evaluation metrics that approximate the bidders' expected utility. Similarly, Vasile et al. (2017) uses the utility that the bidder derives from the auction to design a suitable loss function that the bidder should use for CTR prediction. While both works provide empirical justifications for the proposed losses, they only provide heuristic arguments when designing the loss functions themselves and include no theoretical guarantees on the generalization or calibration of the losses. Moreover, they both rely on the assumption that the distribution of the highest competing bid's eCPM is fixed and known a priori.

Theoretical Analysis of Ad Auctions. Many works study the theoretical properties of ad auctions (Fu et al., 2012; Edelman and Schwarz, 2010; Gatti et al., 2012; Aggarwal et al., 2006; Varian, 2009; Dughmi et al., 2013; Bergemann et al., 2022; Lucier et al., 2012), and Choi et al. (2020) offers a detailed survey of a collection of recent advances in the analysis of ad auctions.

Hummel and McAfee (2017) is the most relevant work to ours, as it studies the design of loss functions in ad auctions from the seller's perspective, offering new insights on how to design losses for either welfare maximization or revenue maximization. However, the real-world experiments in the paper rely on proprietary data, and the claims are not verified on widely available benchmarks. Moreover, it again relies heavily on the assumption that the distribution of the highest competing bid's eCPM is known beforehand, which can be unrealistic in practice.

Learning-to-Rank. Our work draws inspiration from a line of research on learning-to-rank (Burges et al., 2005; 2006; Cortes et al., 2010; Burges, 2010; Wang et al., 2018), which incorporates information retrieval performance metrics such as Normalized Discounted Cumulative Gain into the design of the loss functions, resembling our works. However, as we show in Section 3.2, these works do not directly apply to the welfare maximization setting. Moreover, to the best of our knowledge, these works have not been examined in the context of welfare maximization in ad auctions.

2. Models and Preliminaries

We begin with a multi-slot ad auction (Edelman et al., 2007; Varian, 2007) where each ad is associated a cost-per-click (CPC) bid. Let K denote the number of the slots and each slot, indexed by k, is associated with a position multiplier

[1] https://www.kaggle.com/c/criteo-display-ad-challenge

α_k. Without loss of generality assume that $\alpha_1 = 1$ and the weights are decreasing in k, namely $\alpha_1 \geqslant \ldots \geqslant \alpha_K$. Assume there are $n \geqslant K$ ads participating in the auction where each ad has a feature vector $x_i \in \mathbb{R}^d$ and CPC bid b_i. There exists a function $p^* : \mathbb{R}^d \to [0, 1]$ such that the CTR of the ad at slot k is $\alpha_k p_i$, where we let $p_i = p^*(x_i)$ for convenience. The ad's CTR is affected by both the slot it is assigned to and the ad's features. Intuitively, p_i is the ad's base CTR if it were assigned to the first slot, and is scaled according to α_k for any slot k.

Throughout this paper, we assume that the position multipliers are known, and we focus only on learning p^*, i.e., the ad's CTR if it were assigned the first slot. Learning a position-based CTR prediction model requires additional assumptions to model the user's click behavior and is outside of our scope, which focuses on welfare maximization instead. Indeed, we will show it is without loss of generality to focus on single-slot auctions to maximize welfare, which is equivalent to learning the base CTR when shown in the first slot (Proposition 2.2).

More concretely let $\mathcal{H} \subseteq \{f : \mathcal{X} \to [0, 1]\}$ denote the hypothesis space and assume that p^* is realizable, i.e. $p^*(\cdot) \in \mathcal{H}$. Conditioned on a set of n ads $\{(b_i, x_i)\}_{i=1}^n$, let $f(\cdot)$ denote an arbitrary function that the seller uses to predict the CTRs. The function f, combined with the submitted bid b_i and the observed context x_i, yields the predicted eCPMs $b_i f(x_i)$ for all $i \in [n]$. The seller then awards the first slot to the bidder with the highest predicted eCPM, the second slot to the bidder with the second, and so forth, achieving a welfare of

$$\text{Welfare}_f(\{(b_i, x_i)\}_{i=1}^n) = \sum_{k=1}^{K} b_{\pi_f(k)} p_{\pi_f(k)},$$

where for any function f, $\pi_f(k)$ returns the index of the ad with the k-th highest predicted eCPM. The welfare maximization problem is then

$$\max_{f \in \mathcal{H}} \sum_{k=1}^{K} b_{\pi_f(k)} p_{\pi_f(k)}. \tag{1}$$

As we will prove, a solution to the problem is $f = p^*$. For convenience, we let $\pi_*(\cdot) = \pi_{p^*}(\cdot)$, $\text{Welfare}_*(\{(b_i, x_i)\}_{i=1}^n) = \text{Welfare}_{p^*}(\{(b_i, x_i)\}_{i=1}^n)$, and assume there are no ties in $b_i f(x_i)$ or $b_i p_i$.

To better illustrate welfare and advertisement auction, we include an specific instance of ad auction in the following example.

Example 2.1. *Let $\texttt{ad}_1, \texttt{ad}_2, \texttt{ad}_3$ denote three different advertisements, where $\texttt{ad}_1$'s CTR is 0.1 and CPC bid 10, $\texttt{ad}_2$'s CTR 0.4 and CPC bid 2, and $\texttt{ad}_3$'s CTR 0.9 and CPC bid 0.5. Suppose there two advertisement slots where the first slot has multiplier $\alpha_1 = 1$ and the second $\alpha_2 = 0.9$. Assigning the first slot to $\texttt{ad}_1$ and the second to $\texttt{ad}_2$ maximizes welfare, and the maximum welfare is $1 + 0.8 \times 0.9 = 1.62$. Knowing the ads' exact CTR helps us achieve this maximum welfare.*

2.1. Welfare Maximization and Ranking

We first show that we lose no generality by restricting our focus to single-slot ad (e.g., the first slot) auctions.

Proposition 2.2 (Reduction to Single-slot Setting)**.** *The function f maximizes welfare in a K-slot auction only if it maximizes welfare in single-slot ad auctions held over subsets of the participating ads. Moreover, the ground-truth CTR function p^* maximizes welfare.*

Detailed proof of the proposition is deferred to Appendix A.1. Consider the setting in Example 2.1, for instance. Only considering the welfare objective, note that we can auction off the two ad slots one by one, where $\texttt{ad}_1, \texttt{ad}_2, \texttt{ad}_3$ participates in the auction for the first slot and $\texttt{ad}_2, \texttt{ad}_3$ participates in that for the second. In this setting, if we know the ads' ground-truth CTR, then $\texttt{ad}_1$ wins the first slot and $\texttt{ad}_2$ wins the second, achieving the maximum welfare.

By Proposition 2.2, we can see that welfare maximization in multi-slot ad auctions is no harder than welfare maximization in single-slot ad auctions, and this relies on the fact that the position multipliers are independent of advertisers. For the rest of the paper, we then without loss of generality focus only on single-slot auctions.

As welfare is maximized by the ground-truth CTR function, a common approach is to treat the problem as a classification problem, using y_i as feedback for learning p^* (Vasile et al., 2017; Hummel and McAfee, 2017). However, as noted in Section 1, this approach can suffer from a mismatch between the loss function and the business metric (in our case, welfare).

We notice that welfare maximization can be reduced to a learning-to-rank problem instead. Let $i^* = \pi_*(1)$ be the index of the ad with the highest ground-truth eCPM and $j^* = \pi_f(1)$ be that of the ad with the highest predicted eCPM. We note that

$$\begin{aligned}
&\text{Welfare}_*(\{(b_i, x_i)\}_{i=1}^n) - \text{Welfare}_f(\{(b_i, x_i)\}_{i=1}^n) \\
&\quad = \sum_{i=1}^{n} \sum_{j=1}^{n} ((b_i p_i - b_j p_j)\, \mathbb{1}\{i = i^*\}\, \mathbb{1}\{j = j^*\} \\
&\qquad\qquad \times \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}).
\end{aligned} \tag{2}$$

We defer the detailed derivation of (2) to Appendix A.3. Since $b_{i*} p_{i*}$ yields the highest ground-truth eCPM, welfare is maximized if and only if $j^* = i^*$. Consequently, as

long as f correctly *ranks* each pair of observations according to their ground-truth eCPM, it also correctly identifies the ad with the highest ground-truth eCPM and maximizes welfare. The reduction to ranking generalizes to multi-slot auctions, and we defer a formal statement to Lemma A.1 in the appendix.

The same intuition is illustrated by Example 2.1: as long as we can rank the three ads according to their ground-truth eCPM ($\texttt{ad}_1 > \texttt{ad}_2 > \texttt{ad}_3$), then we can maximize the auction's welfare.

To summarize, we must rank the ads according to their ground-truth eCPM using a suitable CTR prediction function to maximize welfare. An approach that follows this observation is to learn a CTR prediction rule to rank the ads, leading to the proposed ranking-inspired losses.

3. Ranking-Inspired Loss Functions for Welfare Maximization

Let $\mathcal{D} = \{(b_i, x_i), y_i\}_{i=1}^n$ be a batch of n ads participating in one round of an ad auction, where $y_i \sim \mathrm{Ber}(p_i)$ indicates whether the ad has been clicked or not. We then call $b_i p_i$ the ad's ground-truth eCPM and $b_i y_i$ its empirical eCPM. Consider the following pairwise loss function (which we propose the seller minimize):

$$\ell(f;\mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{n} (b_i y_i - b_j y_j)\, \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}. \tag{3}$$

Let $\mathcal{R}(f;\mathcal{D}) = \mathbb{E}_{\{y_i\}_{i=1}^n}[\ell(f;\mathcal{D})]$ denote the conditional risk induced by the loss function ℓ. Recalling that $y_i \sim$ Bernoulli(p_i), we know

$$\mathcal{R}(f;\mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{n} (b_i p_i - b_j p_j)\, \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}. \tag{4}$$

Observe the similarities between (2) and (4). The conditional risk $\mathcal{R}(f;\mathcal{D})$ can be viewed as a proxy for the welfare suboptimality of f, where we replace $\mathbb{1}\{i = i^*\}\,\mathbb{1}\{j = j^*\}$ with 1. While j^* is easy to determine once f is given, we do not the index with the highest ground-truth eCPM. Fortunately, as we show in the following proposition, minimizing $\mathcal{R}(f;\mathcal{D})$ via empirical risk minimization is a reasonable proxy for minimizing welfare suboptimality.

Proposition 3.1. *For any $\mathcal{D}$, let $\widehat{f}$ be an arbitrary and fixed minimizer of the conditional risk $\mathcal{R}(f;\mathcal{D})$. We then know $\widehat{f}$ ranks every pair in the sequence correctly, i.e. $b_i p_i \geqslant b_j p_j$ if and only if $b_i \widehat{f}(x_i) \geqslant b_j \widehat{f}(x_j)$. Moreover, the ground-truth CTR function $p^*(\cdot)$ minimizes the conditional risk $\mathcal{R}(f;\mathcal{D})$ for any $\mathcal{D}$.*

Detailed proof for the proposition can be found in Appendix A.2. Proposition 3.1 shows that minimizing the conditional risk $\mathcal{R}(f;\mathcal{D})$ is a surrogate for maximizing welfare and minimizing (3) is a reasonable choice of loss function. In the following theorem, we make explicit the connection between the conditional risk and welfare. With a slight abuse of notation let $\mathrm{Welfare}_f(\mathcal{D}), \mathrm{Welfare}_*(\mathcal{D})$ denote the welfare achieved by f and the optimal (achievable) welfare, respectively, when $\{(b_i, x_i)\}_{i=1}^n$ are given by the dataset $\mathcal{D}$. We emphasize the conditional risk $\mathcal{R}(f;\mathcal{D})$ can be negative, an important fact to bear in mind in the context of the following theorem.

Theorem 3.2. *The following holds for all $f \in \mathcal{H}$ and $\mathcal{D}$*

$$\mathrm{Welfare}_*(\mathcal{D}) \leqslant \mathrm{Welfare}_f(\mathcal{D}) + \frac{1}{2}\mathcal{R}(f;\mathcal{D}) + \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} |b_i p_i - b_j p_j|\,.$$

Moreover, the bound is tight for any minimizer of $\mathcal{R}(f;\mathcal{D})$ and for all $\mathcal{D}$

$$\min_{f\in\mathcal{H}} \mathcal{R}(f;\mathcal{D}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} |b_i p_i - b_j p_j|\,.$$

See Appendix A.3 for detailed proof. We note that the theorem provides a valid lower bound for all possible $f \in \mathcal{H}$. More importantly, for any dataset $\mathcal{D}$, we can show that there is at least one minimizer of $\mathcal{R}(f;\mathcal{D})$ thanks to the realizability assumption, for which $\frac{1}{2}\mathcal{R}(f;\mathcal{D}) + \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} |b_i p_i - b_j p_j| = 0$. Crucially, the theorem implies that minimizing the conditional risk on any dataset $\mathcal{D}$ maximizes welfare, further justifying the use of $\ell(f;\mathcal{D})$.

While we have shown minimizing $\mathcal{R}(f;\mathcal{D})$ suffices for welfare maximization, recovering the ground-truth CTR function $p^*(\cdot)$ remains crucial for real-world ad auctions. For instance, revenue in generalized second price auctions depends on the pCTRs themselves, and functions that correctly rank the ads do not necessarily lead to high revenue. Fortunately, by adding a calibrated classification loss to $\ell(f;\mathcal{D})$, we can ensure that $p^*(\cdot)$ minimizes the (unconditional) risk. Particularly, we have the following proposition.

Proposition 3.3. *Let $h(f;\mathcal{D})$ denote an arbitrary loss function such that p^* is the unique minimizer of $\mathbb{E}_{\mathcal{D}}[h(f;\mathcal{D})]$. For any constant $\lambda > 0$, p^* is the unique minimizer of $\mathbb{E}_{\mathcal{D}}[\ell(f;\mathcal{D}) + \lambda h(f;\mathcal{D})]$.*

See Appendix A.4 for detailed proof. We note that logistic loss and mean squared error are both valid choices for $h(f;\mathcal{D})$ in Proposition 3.3.

3.1. Easy-to-Optimize Surrogates

While $\ell(f;\mathcal{D})$ is attractive as it is closely related to the welfare, the function itself is nondifferentiable and cannot be

efficiently optimized using first-order methods (e.g., SGD) due to the indicator variables. We thus propose two differentiable surrogates with provable performance guarantees

$$\ell_\sigma^{\log}(f;\mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{n}(b_i y_i - b_j y_j) \times \log(1+\exp(-\sigma(b_i f(x_i) - b_j f(x_j)))), \tag{5}$$

and

$$\ell_\sigma^{\text{hinge}}(f;\mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{n}(b_i y_i - b_j y_j) \times (-\sigma(b_i f(x_i) - b_j f(x_j)))_+. \tag{6}$$

For (5), we replace the indicators in $\ell(f;\mathcal{D})$ with the log logistic function $-\log(1+\exp(-\sigma(b_i f(x_i) - b_j f(x_j))))$. Similarly, (6) acts as a surrogate to $\ell(f;\mathcal{D})$ with the indicator replaced by $(-\sigma(b_i f(x_i) - b_j f(x_j)))_+$ instead, where for any $a \in \mathbb{R}$ we let $(a)_+ = \max(0, a)$. While the function $(\cdot)_+$ itself is not differentiable at $x = 0$, it is differentiable almost everywhere and can be easily optimized using its subderivative.

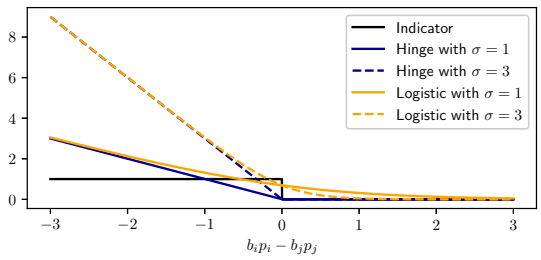

Figure 1. Visualization of the surrogates to $\mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}$ for different values of σ as functions of $b_i f(x_i) - b_j f(x_j)$.

For both surrogates, the term σ is a manually adjustable parameter controlling how much we penalize small margins between a pair of eCPM, $b_i f(x_i) - b_j f(x_j)$. As we can see from Figure 1, for pairs of ads whose predicted eCPMs are close to each other, a larger σ accentuates the difference between them and leads to a surrogate value close to one. However, as σ increases, the surrogate value for ads with large gaps in predicted eCPMs tend to be much larger than one. Adjusting σ is then a balancing act between these two kinds of pairs.

Regardless of the choice of surrogate for the indicator function, the surrogate losses themselves remain closely related to (2), which we highlight in the following theorems.

Theorem 3.4. *Assuming all bids are bounded by some $B \in \mathbb{R}_{>0}$, setting $\sigma = 2/B$ ensures for any $f \in \mathcal{H}$ and $\mathcal{D}$*

$$|\mathbb{E}_{\{y_i\}_{i=1}^n}[\ell_\sigma^{\log}(f;\mathcal{D})] - \mathcal{R}(f;\mathcal{D})| \leqslant \Delta,$$
$$|\mathbb{E}_{\{y_i\}_{i=1}^n}[\ell_\sigma^{\text{hinge}}(f;\mathcal{D})] - \mathcal{R}(f;\mathcal{D})| \leqslant \Delta,$$

where $\Delta = \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n |b_i p_i - b_j p_j|$ is a problem-dependent constant.

See Appendix A.5 for detailed proof. Theorem 3.4 shows that $\ell_\sigma^{\log}(f;\mathcal{D})$ and $\ell_\sigma^{\text{hinge}}(f;\mathcal{D})$ are closely tied to the original loss $\ell(f;\mathcal{D})$. While assuming the CPC bids are bounded implicitly implies the eCPMs' are also bounded, the assumption is mild and does not restrict the parametric form of the eCPMs' distribution. While the surrogates do not exactly match the proposed loss $\ell(f;\mathcal{D})$, the gap is due to approximating the indicators in $\ell(f;\mathcal{D})$ and cannot be avoided.

3.2. Failure of Directly Applying Learning-to-Rank

It may be tempting to further exploit the connection between welfare and ranking over predicted eCPMs by applying a learning-to-rank loss function directly on the observed eCPMs $b_i y_i$. As we show below, the approach, unfortunately, fails, as the inclusion of bids makes the empirical observation $\mathbb{1}\{b_i y_i \geqslant b_j y_j\}$ a poor estimate of $\mathbb{1}\{b_i p_i \geqslant b_j p_j\}$.

Proposition 3.5. *For any $1/2 > \epsilon > 0$, there exists a pair of ads i and j such that*

$$\Pr(\mathbb{1}\{b_i y_i \geqslant b_j y_j\} = \mathbb{1}\{b_i p_i \geqslant b_j p_j\}) = \epsilon,$$

where (b_i, p_i, y_i) and (b_j, p_j, y_j) are the CPC bids, ground-truth CTR, and realized click indicator for the two ads.

See Appendix A.6 for detailed proof. Intuitively, the construction of the counterexample in Proposition 3.5 relies on the fact that the ground-truth eCPM of an ad increases as its corresponding CPC bid increases, but the probability that the ad is clicked does not. In other words, for any ad i, the probability that $b_i y_i$ is non-zero does not depend on b_i while the ground-truth eCPM does, creating a discrepancy between the ground-truth eCPM and the empirical eCPM. We may then strategically manipulate b_i to construct an example satisfying Proposition 3.5.

Crucially, Proposition 3.5 shows that there exist pairs of ads whose empirically observed CPM rankings agree with their ground-truth eCPM rankings with probability arbitrarily close to zero. Unless strong assumptions are made on the distributions of empirically observed CPMs, it is impossible to directly apply off-the-shelf learning-to-rank loss functions for $\mathbb{1}\{b_i y_i \geqslant b_j y_j\}$.

On the other hand, (3) avoids the pitfall by weighing each entry by $(b_i y_i - b_j y_j)$. When conditioned on any CTR prediction rule $f(\cdot)$, by the linearity of expectation, we can

see that the weight is an unbiased estimate of the difference in ground-truth eCPM. The fact that $\ell(f;\mathcal{D})$ is linear in each observed eCPM $b_i y_i$ is crucial, as the linearity ensures that the loss function accurately reflects the differences in $b_i p_i$, enabling us to relate the conditional risk to the actual welfare loss and obtain theoretical guarantees without any assumptions on the empirical eCPMs.

To the best of our knowledge, no existing works on learning-to-rank use loss functions of this form, and our proposed methods are uniquely capable of avoiding the challenge highlighted by Proposition 3.5. While resembling a learning-to-rank loss, (3) is at its core a loss function that resembles the shape of the welfare objective in an ad auction, ensuring that optimizing the loss is closely related to optimizing welfare.

4. Replacing y_i with Predictions from the Teacher Model

A concern for (5) and (6) is that their variance scales with b_i^2, the squared values of the CPC bids. Combined with the noisiness of y_i, the resulting loss may be overly noisy. While the issue might be mitigated by properly pre-processing the CPC bids, we propose a theoretically justifiable alternative inspired by student-teacher learning (Hinton et al., 2015). In fact, distillation loss associated with the prediction from the teacher model is widely used in industrial-scale advertising systems (Anil et al., 2022). This technique is shown to be helpful for stabilizing the training and improving the pCTR accuracy of the student model.

The idea is straightforward. Let $\widehat{p}(\cdot)$ be a teacher network trained on the same dataset and we replace y_i with $\widehat{p}(x_i)$. We show that doing so leads to an empirical loss that is close to $\mathcal{R}(f;\mathcal{D})$, the conditional risk, as long as the teacher network itself is sufficiently accurate. We begin with the following theorem for replacing the y_i's in (3).

Theorem 4.1. *Let $\widehat{p}$ be an estimate of p^* such that $\mathbb{E}_x[(\widehat{p}(x)-p^*(x))^2]\leqslant\epsilon$. Let $\widehat{\ell}(f;\mathcal{D})$ be* (3) *but with each y_i replaced by $\widehat{p}_i=\widehat{p}(x_i)$, i.e.,*

$$\widehat{\ell}(f;\mathcal{D})=\sum_{i=1}^{n}\sum_{j=1}^{n}(b_i\widehat{p}_i-b_j\widehat{p}_j)\,\mathbb{1}\{b_i f(x_i)\geqslant b_j f(x_j)\}.$$

Assuming all bids are upperbounded by positive constant $B\in\mathbb{R}_{>0}$, for any $f\in\mathcal{H}$ we have

$$\mathbb{E}_{\mathcal{D}}[|\widehat{\ell}(f;\mathcal{D})-\mathcal{R}(f;\mathcal{D})|]\leqslant(n-1)nB\sqrt{\epsilon}$$

where n is the number of ads.

See Appendix A.7 for proof. As $\widehat{\ell}(f;\mathcal{D})$ sums over all pairs of ads, the bound necessarily grows in $\mathcal{O}(n^2)$, and the factor can be removed if we use the average over the pairs instead.

While the teacher network may be used in ad auctions as-is, student networks still offer several benefits in addition to the theoretical guarantee in Theorem 4.1. First, teacher networks may be costly to deploy, thus student networks offer efficiency benefits from knowledge distillation. Second, the ranking losses may help the student network better differentiate the eCPMs of pairs of ads, leading to higher welfare, as we observe in experiments.

It is also reasonable to suggest directly learn-to-rank with $\mathbb{1}\{b_i\widehat{p}(x_i)\geqslant b_j\widehat{p}(x_j)\}$ as the labels. However, theoretical guarantees for the approach require additional assumptions on the distribution of the gaps between pairs of predicted eCPM, which is not needed for Theorem 4.1.

Recalling Theorem 3.4 and Theorem 3.4, it is not hard to see that replacing y_i with $\widehat{p}(x_i)$ in (5) and (6) lead to losses that are also sufficiently close to $\mathcal{R}$. We instead focus on using the teacher network to improve calibration.

4.1. Improving Calibration with the Teacher Network

A drawback shared by (5) and (6) is that they are not calibrated. While both penalizes pCTR functions for incorrectly ranking pairs of ads, they also reward pCTR functions that overestimate the margin between pairs of ads. As the minimizers of their expected values are not necessarily the ground-truth CTR function, using these losses may have negative consequences on other important metrics such as revenue or area under the curve. Fortunately, we show that using a teacher network also improves the calibration of the loss function. We propose the following loss function.

$$\widehat{\ell}_\sigma^{\text{hinge},+}(f;\mathcal{D})=\sum_{i=1}^{n}\sum_{j=1}^{n}(b_i\widehat{p}_i-b_j\widehat{p}_j)_+ \times(-\sigma(b_i f(x_i)-b_j f(x_j)))_+, \tag{7}$$

Intuitively, $\widehat{\ell}_\sigma^{\text{hinge},+}(f;\mathcal{D})$ no longer punishes f for having a small margin between predicted eCPMs, as long as f ranks the pair the same way $\widehat{p}$ does. When the teacher network is sufficiently close to the ground-truth, the loss function eliminates the bias that (5) and (6) have towards functions with larger margins between pairs. Additionally, compared to directly using $\widehat{p}(\cdot)$, (7) better reflects the impact that the pCTRs have on welfare and has theoretical guarantees in terms of welfare performance.

We now present theoretical justification for the approach. Recall from Vasile et al. (2017) that calibration in ad auctions is defined as follows.

Definition 4.2 (Calibration)**.** A loss function $\ell'(f;\mathcal{D})$ is calibrated if its expected value $\mathbb{E}_{\mathcal{D}}[\ell'(f;\mathcal{D})]$ is minimized by the ground-truth CTR function p^*.

Based off of Definition 4.2, we first define a slightly relaxed

notion of calibration, ϵ-approximate calibration.

Definition 4.3 (ϵ-Approximate Calibration)**.** A loss function $\ell'(f;\mathcal{D})$ is said to be ϵ-approximately calibrated if the expected value of the loss achieved by the ground-truth CTR function p^* is at most ϵ greater than the minimum, namely

$$\mathbb{E}_{\mathcal{D}}[\ell'(p^*;\mathcal{D})] - \min_{f\in\mathcal{H}} \mathbb{E}_{\mathcal{D}}[\ell'(f;\mathcal{D})] \leqslant \epsilon.$$

We then have the following guarantee for $\widehat{\ell}_{\sigma}^{\text{hinge},+}$.

Theorem 4.4. *Let $\widehat{p}$ be an estimate of p^* such that $\mathbb{E}[(\widehat{p}(x) - p^*(x))^2] \leqslant \epsilon$. Assuming all bids are upper bounded by some $B \in \mathbb{R}_{>0}$, for any $f \in \mathcal{H}$ we have*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\text{Welfare}_*(\mathcal{D})] \\
&\quad\leqslant \mathbb{E}_{\mathcal{D}}[\text{Welfare}_f(\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[\widehat{\ell}_{\sigma}^{\text{hinge},+}(f;\mathcal{D})] \\
&\qquad + \frac{n(n-1)}{2} B \max\{1, \sigma B - 1\} \\
&\qquad + n(n-1)\sigma B^2 \sqrt{\epsilon}.
\end{aligned}$$

Moreover, the loss function $\widehat{\ell}_{\sigma}^{\text{hinge},+}(f;\mathcal{D})$ is $\mathcal{O}(\sqrt{\epsilon})$-approximately calibrated.

See Appendix A.8 for detailed proof. An important feature Theorem 4.1 and Theorem 4.4 share is that they depend only on the ℓ_2 generalization error of the teacher network, and not on the explicit parametric assumptions on the distribution of eCPM. In other words, for any $\widehat{p}$ we can simply use off-the-sheif results on its generalization error to show that using the induced $\widehat{\ell}_{\sigma}^{\text{hinge},+}(f;\mathcal{D})$ is approximately calibrated, with only the mild assumption that the CPC bids are bounded.

4.2. Weighing with Teacher Networks

The inclusion of the teacher network further guides us in developing theoretically-inspired weights for the proposed losses. The goal of the weight for the pair i, j is to mimic the indicator product $\mathbb{1}\{i = i^*\}\,\mathbb{1}\{j = j^*\}$, where $i^* = \text{argmax}_{i\in[n]}\, b_i p_i$ and $j^* = \text{argmax}_{i\in[n]}\, b_j f(x_j)$, so that the resulting loss better resembles the welfare suboptimality in (2). The first indicator corresponds to the event that the ad i has the highest ground-truth eCPM and the second the event that the ad j has the highest predicted eCPM. The weight should then be increasing in both $b_i\widehat{p}(x_i)$ and $b_j f(x_j)$, with $\widehat{p}$ being the teacher network.

5. Experiments

We now demonstrate the advantages of our proposed losses on both simulated data and the Criteo Display Advertising Challenge dataset, a popular real-world benchmark for CTR prediction in ad auctions. Recalling Proposition 3.3, we use the weighted sum of the logistic loss and the proposed ranking losses for all experiments to ensure the learned CTR model is sufficiently close to the ground truth.

5.1. Synthetic Dataset

For the simulation setting, we assume that the ads' features are 50-dimensional random vectors where each component is i.i.d. drawn from the standard normal distribution, namely $x_i \sim \mathcal{N}(0, I_{50})$, where I_{50} denotes the 50-dimensional identity matrix. For training, we generate 10,000 x_i's from the $\mathcal{N}(0, I_{50})$ distribution and generate the corresponding ground-truth CTR from a logistic model and the CPC bids from a log-normal distribution. We then draw the click indicators $y_i \sim \text{Ber}(p_i)$. We defer a more detailed description of the data-generating process to Appendix B.

A two-layer neural network is used, where the hidden layer has 50 nodes with ReLU activation, and the output layer has one node with sigmoid activation. For evaluation, we simulate 2,000 auctions with 50 ads each. The training and evaluation processes are then repeated 30 times.

We begin by introducing the baselines we consider: logistic loss (also referred to as cross-entropy) and two versions of weighted logistic loss. Logistic loss is commonly used for training models for predicting CTRs, and is used by Guo et al. (2017); Lian et al. (2018); Chen et al. (2016) among many other works. Existing works (Vasile et al., 2017; Hummel and McAfee, 2017) suggest the usage of a weighed logistic loss, with each entry weighted by the CPC bid. Finally, Vasile et al. (2017) propose weighing the logistic loss by the square root of the CPC bid.

We focus on three loss function representative of what we proposed: $\ell_{\sigma=1}^{\log}$, $\widehat{\ell}_{\sigma=1}^{\log}$, and $\widehat{\ell}_{\sigma=1}^{\text{hinge},+}$. The first and the third correspond to (5) and (7), respectively. The second, $\widehat{\ell}_{\sigma=1}^{\log}$, replaces the y_i's in $\ell_{\sigma=1}^{\log}$ with $\widehat{p}_i$ obtained from a teacher network.

As discussed immediately after Theorem 3.2, we add binary cross entropy loss to $\ell_{\sigma=1}^{\log}$, $\widehat{\ell}_{\sigma=1}^{\log}$, and $\widehat{\ell}_{\sigma=1}^{\text{hinge},+}$ and optimize over the composite loss. Additionally, motivated by Section 4.2, we use logistic functions to weigh each pair in $\widehat{\ell}_{\sigma=1}^{\text{hinge},+}$ and $\widehat{\ell}_{\sigma=1}^{\log}$. Both $\widehat{\ell}_{\sigma=1}^{\text{hinge},+}$ and $\widehat{\ell}_{\sigma=1}^{\log}$ use the model trained with logistic loss as the teacher network. We defer a more detailed discussion to Appendix B.

As we can see from Figure 2, all three proposed pairwise ranking losses achieve higher test time welfare than the naive baselines. As we use the same model structure and optimizer for all models, it is further possible that with more careful tuning, the advantages of the pairwise ranking losses may be more pronounced.

Student-Teacher Learning. Comparing the performance of $\ell_{\sigma=1}^{\log}$ and $\widehat{\ell}_{\sigma=1}^{\log}$ shows that student-teacher learning overall beneficial for simulated data. Moreover, while $\widehat{\ell}_{\sigma=1}^{\text{hinge},+}$ is theoretically proven to be calibrated by Theorem 4.4, in the simulated task we found that the loss does perform well compared to other proposed methods. We conjecture that

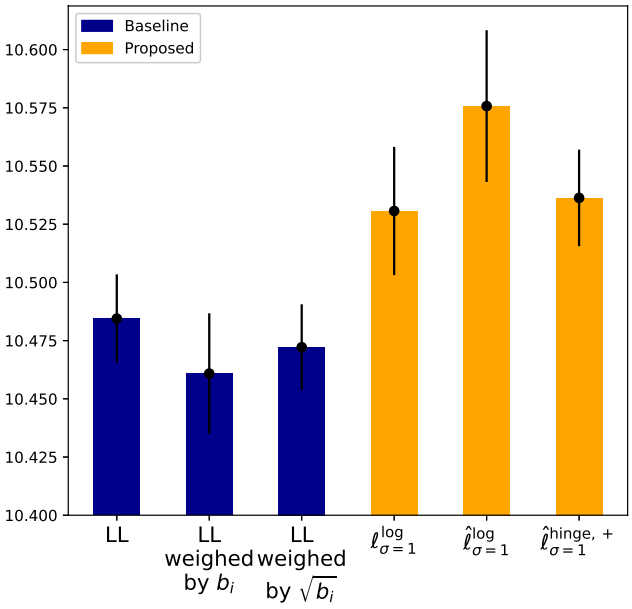

Figure 2. Test welfare on simulated data (higher is better). From left to right: (**In Blue**) models trained with logistic loss; logistic loss weighted by b_i (Hummel and McAfee, 2017); logistic loss weighted by $\sqrt{b_i}$ (Vasile et al., 2017), (**In Yellow**) proposed $\ell^{\log}_{\sigma=1}$ indicator replaced by logistic function (defined in (5)); $\hat{\ell}^{\log}_{\sigma=1}$ indicator replaced by logistic function + student-teacher learning; $\hat{\ell}^{\text{hinge},+}_{\sigma=1}$ indicator replaced by hinge function + student-teacher learning (defined in (7)).

the relatively modest performance is due to the fact that hinge function is not as smooth as logistic function, and thus is not well-suited for training neural networks.

Comparison with Existing Works. The experiment results also show that the loss functions derived in earlier works may depend on unrealistic assumptions and may be lacking in empirical justification, as can be seen in the performance of both weighted logistic losses. Regardless, we have shown that our proposed methods significantly outperform existing baselines.

5.2. Criteo Dataset

We use the popular Criteo Display Advertising Challenge dataset. We follow standard data preprocessing procedures and use a standard 8-1-1 train-validation-test split commonly found in the literature. We defer to Appendix C for a more detailed description of the setup.

We note there are several limitations to the dataset. Firstly, the Criteo dataset only includes ads that are shown. In an ad auction setting, this means that all ads have won their respective multi-slot auction. Moreover, the Criteo dataset only includes anonymous features, which means we have no access to key attributes such as the CPC bid or the slot for each ad. Lastly, we do not know the specific auction each ad belongs to. Unfortunately, these limitations are shared by all openly available benchmarks to the best of our knowledge.

For the first limitation, we note that it is near-impossible to learn accurate CTR models without assuming the CTRs of the shown ads follow the same distribution as those of the unshown ads. To handle the intrinsic bias between shown ads and unshown ads is very challenging and out of the scope of this paper. While the slot each ad belongs to is unavailable, as we argued previously, learning a position-based CTR model is not the focus of this work, and here we learn the CTR of each ad, assuming that it is assigned to the first slot. Finally, while we do not know the exact auction round, from Proposition 2.2, we know maximizing the welfare of multi-slot ad auctions requires maximizing the welfare of single-slot auctions over subsets of participating ads (given the position multipliers are independent wrt. advertisers). Thus, it remains viable to treat each minibatch as a specific instance of single-slot auction.

We generate the CPC bids using the outputs from a DeepFM model (Guo et al., 2017) with randomly initialized weights, ensuring that the generated bids follow a log-normal distribution. Particularly, let $h(\cdot)$ denote a randomly initialized DeepFM model, we set $b_i = \exp(c \cdot h(x_i) + \epsilon_i)$, with c being a scaling factor and ϵ_i a Gaussian noise.

We experiment with both DeepFM (Guo et al., 2017) and DCN (Wang et al., 2017), two popular models with great performance on the Criteo dataset whose parameter choices are well-documented. To ensure that our loss functions benefit welfare *holding all else constant*, we did not perform any parameter tuning or architecture search and used the model architectures and training protocols specified in the papers.

In this setting, we omit $\hat{\ell}^{\text{hinge},+}_{\sigma}$ as it is non-smooth and not well suited for complex neural network architectures considered here, given our empirical study. Logistic loss is chosen as the baseline and $\ell^{\log}_{\sigma=3}$ and $\hat{\ell}^{\log}_{\sigma=3}$ are the proposed candidates due to smoothness. Here we set $\sigma = 3$ to better mimic the shape of the indicator variable. We omit weighted logistic losses proposed in Vasile et al. (2017); Hummel and McAfee (2017) due to their poor performance on the synthetic dataset. We compare the losses based on three metrics: test-time welfare, area-under-curve (AUC) loss, and logistic loss, where AUC loss is defined as $1 - \text{AUC}$.

For both DeepFM and DCN, we repeat the following procedure 10 times. We fit the models using logistic loss and $\ell^{\log}_{\sigma=3}$. The model fitted using logistic loss is then used as the teacher network, whose outputs are used to construct $\hat{\ell}^{\log}_{\sigma=3}$. We then evaluate the welfare, AUC loss, and logistic loss of the three models on the test set.

We report the welfare and AUC loss for DeepFM in Table 1and those for DCN in Figure 3. Additional results

	DeepFM	
	Welfare	AUC Loss
LL (baseline)	1.4448 ± 0.0025	0.2200 ± 0.0004
$\ell^{\log}_{\sigma=3}$	1.4622 ± 0.0021	0.2169 ± 0.0003
$\widehat{\ell}^{\log}_{\sigma=3}$	$\mathbf{1.4660 \pm 0.0022}$	0.2229 ± 0.0004

Table 1. Welfare (higher is better) and AUC loss (lower is better) for DeepFM. Top to bottom: (**Baseline**) logistic loss; (**Proposed**) $\ell^{\log}_{\sigma=3}$ indicator replaced by logistic function (defined in (5)); $\widehat{\ell}^{\log}_{\sigma}$ indicator replaced by logistic function + student-teacher learning.

including comparisons on the wall-clock run time can be found in Appendix C.

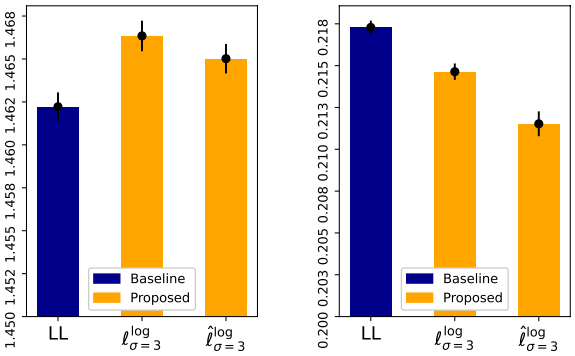

Figure 3. Experimental results for DCN. **Left: Welfare** (higher is better). **Right: AUC loss** (lower is better). For each plot, from left to right: (**Baseline, Blue**) logistic loss, (**Proposed, Yellow**) $\ell^{\log}_{\sigma=3}$ indicator replaced by *logistic* function (defined in (5)); $\widehat{\ell}^{\log}_{\sigma}$ indicator replaced by *logistic* function + *student-teacher learning*.

As we observe from Table 1 and Figure 3, the proposed losses significantly improve test time welfare at a minimal cost (if any) to classification performance. Moreover, the improvement does not depend on the specific underlying model structure, and student-teacher learning continues to prove to be beneficial. Surprisingly, the proposed losses may also benefit AUC, a classification metric. We conjecture the improvement is due to the ranking loss formulation, which forces the model to better differentiate the ads' CTRs.

6. Conclusion and Future Work

We propose surrogates that improve welfare for ad auctions with theoretical guarantees and good empirical performance. We hypothesize that the improvements will be more pronounced if we further tune the model architecture for the proposed losses and we leave architecture search as a future direction.

Acknowledgements

Part of the work was completed while Boxiang Lyu and Zachary Robertson were Student Researchers at Google Research Mountain View. We would like to thank for Phil Long for the initial discussions, Ashwinkumar Badanidiyuru, Zhuoshu Li, and Aranyak Mehta for the insightful feedback.

References

Gagan Aggarwal, Ashish Goel, and Rajeev Motwani. Truthful auctions for pricing search keywords. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 1–7, 2006.

Amazon Ads. How does bidding work for Sponsored Products?, 2023. URL `https://advertising.amazon.com/library/videos/campaign-bidding-sponsored-products`. Accessed on 01/08/2023.

Rohan Anil, Sandra Gadanho, Da Huang, Nijith Jacob, Zhuoshu Li, Dong Lin, Todd Phillips, Cristina Pop, Kevin Regan, Gil I. Shamir, Rakesh Shivanna, and Qiqi Yan. On the factory floor: ML engineering for industrial-scale ads recommendation models. In João Vinagre, Marie Al-Ghossein, Alípio Mário Jorge, Albert Bifet, and Ladislav Peska, editors, *Proceedings of the 5th Workshop on Online Recommender Systems and User Modeling co-located with the 16th ACM Conference on Recommender Systems, ORSUM@RecSys 2022, Seattle, WA, USA, September 23rd, 2022*, volume 3303 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

Dirk Bergemann, Paul Dütting, Renato Paes Leme, and Song Zuo. Calibrated click-through auctions. In *Proceedings of the ACM Web Conference 2022*, pages 47–57, 2022.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

Christopher Burges, Robert Ragno, and Quoc Le. Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems*, 19, 2006.

Christopher JC Burges. From RankNet to LambdaRank to LambdaMART: An overview. *Learning*, 11(23-581):81, 2010.

Olivier Chapelle. Offline evaluation of response prediction in online advertising auctions. In *Proceedings of the 24th international conference on world wide web*, pages 919–922, 2015.

Junxuan Chen, Baigui Sun, Hao Li, Hongtao Lu, and Xian-Sheng Hua. Deep CTR prediction in display advertising. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 811–820, 2016.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

Hana Choi, Carl F Mela, Santiago R Balseiro, and Adam Leary. Online display advertising markets: A literature review and future directions. *Information Systems Research*, 31(2):556–575, 2020.

Vincent Conitzer, Christian Kroer, Debmalya Panigrahi, Okke Schrijvers, Nicolas E Stier-Moses, Eric Sodomka, and Christopher A Wilkens. Pacing equilibrium in first price auction markets. *Management Science*, 2022.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

Shaddin Dughmi, Nicole Immorlica, and Aaron Roth. Constrained signaling for welfare and revenue maximization. *ACM SIGecom Exchanges*, 12(1):53–56, 2013.

Benjamin Edelman and Michael Schwarz. Optimal auction design and equilibrium selection in sponsored search auctions. *American Economic Review*, 100(2):597–602, 2010.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review*, 97(1):242–259, 2007.

Hu Fu, Patrick R. Jordan, Mohammad Mahdian, Uri Nadav, Inbal Talgam-Cohen, and Sergei Vassilvitskii. Ad auctions with data. In *INFOCOM Workshops*, pages 184–189, 2012.

Nicola Gatti, Alessandro Lazaric, and Francesco Trovo. A truthful learning mechanism for contextual multi-slot sponsored search auctions with externalities. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 605–622, 2012.

Google Ads Help. How the Google Ads auction works, 2023. URL `https://support.google.com/google-ads/answer/6366577?hl=en`. Accessed on 01/08/2023.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. DeepFM: A factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 1725–1731, Melbourne, Australia, August 2017. AAAI Press. ISBN 978-0-9992411-0-3.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Patrick Hummel and R Preston McAfee. Loss functions for predicted click-through rates in auctions for online advertising. *Journal of Applied Econometrics*, 32(7): 1314–1328, 2017.

Olivier Jeunen. Amazon scientists win best-paper award for ad auction simulator, September 2022. URL `https://www.amazon.science/blog/amazon-scientists-win-best-paper-award-for-ad-auction-simulator`. Accessed on 01/08/2023.

Yuchin Juan, Damien Lefortier, and Olivier Chapelle. Field-aware factorization machines in a real-world online advertising system. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 680–688, 2017.

Zeyu Li, Wei Cheng, Yang Chen, Haifeng Chen, and Wei Wang. Interpretable click-through rate prediction through hierarchical attention. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 313–321, 2020.

Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1754–1763, 2018.

Brendan Lucier, Renato Paes Leme, and Éva Tardos. On revenue in the generalized second price auction. In *Proceedings of the 21st international conference on World Wide Web*, pages 361–370, 2012.

H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230, 2013.

Meta Business Help Center. About ad auctions, 2023. URL `https://www.facebook.com/business/help/430291176997542`. Accessed on 01/08/2023.

Junwei Pan, Jian Xu, Alfonso Lobos Ruiz, Wenliang Zhao, Shengjun Pan, Yu Sun, and Quan Lu. Field-weighted factorization machines for click-through rate prediction in display advertising. In *Proceedings of the 2018 World Wide Web Conference*, pages 1349–1357, 2018.

Qi Pi, Weijie Bian, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2671–2679, 2019.

Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. Product-based neural networks for user response prediction. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1149–1154. IEEE, 2016.

Weichen Shen. DeepCTR: Easy-to-use, modular and extendible package of deep-learning based ctr models. `https://github.com/shenweichen/deepctr`, 2017.

Statista. Digital Advertising - Worldwide, 2022. URL `https://www.statista.com/outlook/dmo/digital-advertising/worldwide`. Accessed on 01/08/2023.

Hal R. Varian. Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178, December 2007. URL `https://ideas.repec.org/a/eee/indorg/v25y2007i6p1163-1178.html`.

Hal R Varian. Online ad auctions. *American Economic Review*, 99(2):430–34, 2009.

Hal R Varian and Christopher Harris. The vcg auction in theory and practice. *American Economic Review*, 104(5): 442–45, 2014.

Flavian Vasile, Damien Lefortier, and Olivier Chapelle. Cost-sensitive learning for utility optimization in online advertising auctions. In *Proceedings of the ADKDD'17*, pages 1–6. Association for Computing Machinery, 2017.

Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*, pages 1–7. Association for Computing Machinery, 2017.

Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. The lambdaloss framework for ranking metric optimization. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1313–1322, 2018.

Zhiqiang Wang, Qingyun She, and Junlin Zhang. Masknet: introducing feature-wise multiplication to ctr ranking models by instance-guided mask. *arXiv preprint arXiv:2102.07619*, 2021.

Yanwu Yang and Panyu Zhai. Click-through rate prediction in online advertising: A literature review. *Information Processing & Management*, 59(2):102853, 2022.

Weinan Zhang, Tianming Du, and Jun Wang. Deep learning over multi-field categorical data. In *European conference on information retrieval*, pages 45–57. Springer, 2016.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1059–1068, 2018.

Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5941–5948, 2019.

Jieming Zhu, Jinyang Liu, Weiqi Li, Jincai Lai, Xiuqiang He, Liang Chen, and Zibin Zheng. Ensembled CTR prediction via knowledge distillation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2941–2958, 2020.

A. Omitted Proofs

A.1. Proof of Proposition 2.2

In order to prove Proposition 2.2, we begin with the following lemma that reduces welfare maximization to ranking.

Lemma A.1. *The function f maximizes welfare in a K-slot auction only if it $\pi_f(k) = \pi_*(k)$ for all $k = 1, \dots, K$.*

Proof of Lemma A.1. We begin by showing a basic fact that ads with higher ground-truth eCPM should be ranked higher. Let $1 \leqslant k_1 \leqslant k_2 \leqslant K$ be two arbitrary and fixed slots and i, j be a pair of arbitrary and fixed ads, and assume without loss of generality that $b_i p_i \geqslant b_j p_j$. Immediately, we know that

$$\alpha_{k_1} - \alpha_{k_2} \geqslant 0 \qquad b_i p_i - b_j p_j \geqslant 0$$

where the first inequality comes from the assumption that α_k is decreasing in k and $k_1 \leqslant k_2$. The product of two nonnegative reals is nonnegative, and therefore

$$(\alpha_{k_1} - \alpha_{k_2})(b_i p_i - b_j p_j) \geqslant 0$$

which rearranges to

$$\alpha_{k_1} b_i p_i + \alpha_{k_2} b_j p_j \geqslant \alpha_{k_1} b_j p_j + \alpha_{k_2} b_i p_i.$$

In other words, for pair of ads (i, j) in the top k slots, if $b_i p_i \geqslant b_j p_j$, then ad i should be assigned to a slot that is closer to the first. As $p_i = p^*(x_i)$, the function p^* correctly ranks each ad and therefore achieves maximum welfare.

We now show f maximizes welfare only if $\pi_f(k) = \pi_*(k)$ for all k. Let some CTR prediction function f be arbitrary and fixed and assume there exists some k_0 where $\pi_f(k_0) \neq \pi_*(k_0)$. We can then divide the problem into two cases.

1. When $b_{\pi_f(k_0)} p_{\pi_f(k_0)}$ does not have top K ground-truth eCPM. In other words, there is no $1 \leqslant k \leqslant K$ such that

   $$\pi_f(k_0) = \pi_*(k).$$

   In this case, the function f has erroneously selected an ad who does not have top K ground-truth eCPM and assigned it to the k_0-th slot. Moreover, the inclusion of the ad in the K-slots must imply that one of the ads with top K ground-truth eCPM must be omitted by f. It is easy to see that replacing ad $\pi_f(k_0)$ with the ad that is erroenously left out improves f's welfare.

2. When $b_{\pi_f(k_0)} p_{\pi_f(k_0)}$ has top K ground-truth eCPM. In other words, there is some $1 \leqslant k_1 \leqslant K$ such that $\pi_f(k_0) = \pi_*(k_1)$ where $k_1 \neq k_0$. The case can be further divided into two subcases.

   - When $k_0 < k_1$. In this case, the ad is ranked higher by f than it actually is. Similar to our reasoning for case 1, there must be an ad with top k_0 ground-truth eCPM that is erroneously left out of top k_0 by f. Switching ad $\pi_f(k_0)$ and the ad that is left out increases welfare.
   - When $k_0 > k_1$. In this case, the ad is ranked lower by f than it actually is. Therefore, there must be an ad with lower ground-truth eCPM in front of ad $\pi_f(k_0)$, and switching the two ads also increase welfare.

As we can see, whenever π_f and π_*, there is always a method to improve the welfare achieved by f. Hence, f maximizes welfare only if $\pi_f(k) = \pi_*(k)$ for $1 \leqslant k \leqslant K$. □

We then proceed with the proof of the proposition itself.

Proof of Proposition 2.2. Let $\{(b_i, x_i)\}_{i=1}^n$ be the ads participating in the K-slot auction. Let f denote an arbitrary and fixed CTR function. For all $1 \leqslant k \leqslant K - 1$, we define the set

$$S_k = \{(b_i, x_i) : b_i p^*(x_i) < b_{\pi_*(k)} p_{\pi_*(k)}\}.$$

In other words, S_k is the set of ads whose ground-truth eCPMs are *outside* of top-k, excluding the ad with the k-th highest ground-truth eCPM. With a slight abuse of notation let

$$S_0 = \{(b_i, x_i)\}_{i=1}^n.$$

Observe that f maximizes the welfare of the *single*-slot auction over S_k if and only if

$$\pi_f(k+1) = \pi_*(k+1)$$

for all $k = 1, \ldots, K-1$, as we recall the construction of S_k and Lemma A.1. Additionally, note that maximizing the welfare of the single-slot ad auction over S_0 requires $\pi_f(1) = \pi_*(1)$. Combine both observations, we know that f maximizes the welfare of single-slot ad auctions over $S_0, \ldots, S_K$ if and only if $\pi_f(k) = \pi_*(k)$ for $k = 1, \ldots, K$.

Additionally, we know by Lemma A.1 that f maximizes the welfare of the K-slot auction over S_0 if and only if

$$\pi_f(k) = \pi_*(k), \quad \forall\, 1 \leqslant k \leqslant K.$$

The two claims are thus equivalent and we complete the proof. □

A.2. Proof of Proposition 3.1

Proof. Let $\{(b_i, x_i)\}_{i=1}^n$ be arbitrary and fixed. Consider some pair (i, j) and an arbitrary, not necessarily optimal, pCTR function f. We write out the parts of the conditional risk that depend only on the pair

$$\begin{aligned}&(b_i p_i - b_j p_j)(\mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\} - (1 - \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\})) \\ &\quad = (b_i p_i - b_j p_j)(2\,\mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\} - 1)\end{aligned} \tag{8}$$

The conditional risk for the pair is then minimized when $\mathbb{1}\{b_i p_i \leqslant b_j p_j\} = \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}$. By the range of indicator variable, we know that

$$(b_i p_i - b_j p_j)(2\,\mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\} - 1) \geqslant -|b_i p_i - b_j p_j|.$$

Summing over all pairs i, j, the bound above implies

$$\mathcal{R}(f; \mathcal{D}) \geqslant -\sum_{i \neq j} |b_i p_i - b_j p_j|.$$

As we focus on the realizable setting, $p^* \in \mathcal{H}$, and therefore

$$\min_{f \in \mathcal{H}} \mathcal{R}(f; \mathcal{D}) \leqslant \mathcal{R}(p^*) \leqslant -\sum_{i \neq j} |b_i p_i - b_j p_j|,$$

which immediately implies that for any $\{(b_i, x_i)\}_{i=1}^n$,

$$\min_{f \in \mathcal{H}} \mathcal{R}(f; \mathcal{D}) = -\sum_{i \neq j} |b_i p_i - b_j p_j|.$$

From realizability, there always exist at least one hypothesis in $\mathcal{H}$ that minimize the conditional risk. Moreover, as we can see from the equation above, the ground-truth CTR function minimizes the conditional risk. Lastly, note that the ground-truth CTR ranks every pair correctly, which then implies the minimizer of the conditional risk must also rank each pair correctly. □

A.3. Proof of Theorem 3.2

Proof. Let $f \in \mathcal{H}$ be arbitrary and fixed. We proceed by first writing out the welfare loss in a pairwise fashion. For convenience, let $i^* = \operatorname{argmax}_{i' \in [n]} b_{i'} p_{i'}$ denote the index with the highest ground-truth eCPM. For an auction with n participants, the optimal welfare is $\max_{i' \in [n]} b_{i'} p_{i'}$, which expands to

$$\text{Welfare}_*(\mathcal{D}) = \sum_{i=1}^n b_i p_i\, \mathbb{1}\{i = i^*\}.$$

Similarly, the welfare achieved by pCTR rule f expands to

$$\text{Welfare}_f(\mathcal{D}) = \sum_{j=1}^n b_j p_j\, \mathbb{1}\{j = j^*\},$$

where we let $j^* = \operatorname{argmax}_{j' \in [n]} b_{j'} f(x_{j'})$. The difference between them is

$$\text{Welfare}_*(\mathcal{D}) - \text{Welfare}_f(\mathcal{D}) = \sum_{i \in [n]} \sum_{j \in [n]} \mathbb{1}\{i = i^*\} \mathbb{1}\{j = j^*\}(b_i p_i - b_j p_j).$$

Since $\mathbb{1}\{j = j^*\} = 1$ implies $b_j f(x_j) \geqslant b_i f(x_i)$, the difference equals to

$$\text{Welfare}_*(\mathcal{D}) - \text{Welfare}_f(\mathcal{D}) = \sum_{(i,j) \in [n]^2} \mathbb{1}\{i = i^*\} \mathbb{1}\{j = j^*\} \mathbb{1}\{b_j f(x_j) \geqslant b_i f(x_i)\}(b_i p_i - b_j p_j).$$

Consider an arbitrary and fixed unordered pair (i, j) . As the summation contains both (i, j) and (j, i), we write out the parts of ground-truth welfare loss that depend on only the pair,

$$\begin{aligned} &\mathbb{1}\{i = i^*\} \mathbb{1}\{j = j^*\} \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}(b_i p_i - b_j p_j) \\ &\quad + \mathbb{1}\{j = i^*\} \mathbb{1}\{i = j^*\} \mathbb{1}\{b_j f(x_j) \leqslant b_i f(x_i)\}(b_j p_j - b_i p_i). \end{aligned} \tag{9}$$

When $b_i p_i \geqslant b_j p_j$, the sum is upper bounded by

$$\mathbb{1}\{i = i^*\} \mathbb{1}\{j = j^*\} \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}(b_i p_i - b_j p_j) \leqslant \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}(b_i p_i - b_j p_j).$$

When $b_j p_j \geqslant b_i p_i$, the sum is upper bounded by

$$\mathbb{1}\{j = i^*\} \mathbb{1}\{i = j^*\} \mathbb{1}\{b_j f(x_j) \leqslant b_i f(x_i)\}(b_j p_j - b_i p_i) \leqslant \mathbb{1}\{b_j f(x_j) \leqslant b_i f(x_i)\}(b_j p_j - b_i p_i).$$

Combining the two halves, we know that omitting ties, (9) is upper bounded by

$$\max\{\mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}(b_i p_i - b_j p_j), \mathbb{1}\{b_j f(x_j) \leqslant b_i f(x_i)\}(b_j p_j - b_i p_i)\}.$$

Ignoring ties, we may divide the problem to four cases: when $b_i f(x_i) > b_j f(x_j)$ and $b_i p_i > b_j p_j$, when $b_i f(x_i) < b_j f(x_j)$ and $b_i p_i > b_j p_j$, when $b_i f(x_i) > b_j f(x_j)$ and $b_i p_i < b_j p_j$, and finally when $b_i f(x_i) < b_j f(x_j)$ and $b_i p_i < b_j p_j$. We can show that in all four cases,

$$\begin{aligned} &\max\{\mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}(b_i p_i - b_j p_j), \mathbb{1}\{b_j f(x_j) \leqslant b_i f(x_i)\}(b_j p_j - b_i p_i)\} \\ &\quad = \frac{1}{2}(\mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}(b_i p_i - b_j p_j) + \mathbb{1}\{b_j f(x_j) \leqslant b_i f(x_i)\}(b_j p_j - b_i p_i) + |b_i p_i - b_j p_j|). \end{aligned}$$

Summing over all $(i, j) \in [n]^2$ and dividing by two shows

$$\text{Welfare}_*(\mathcal{D}) - \text{Welfare}_f(\mathcal{D}) \leqslant \frac{1}{2}\mathcal{R}(f; \mathcal{D}) + \frac{1}{4} \sum_{(i,j) \in [n]^2} |b_i p_i - b_j p_j|.$$

Rearranging the terms gives us

$$\text{Welfare}_f(\mathcal{D}) \geqslant -\frac{1}{2}\mathcal{R}(f; \mathcal{D}) + \text{Welfare}_*(\mathcal{D}) - \frac{1}{4} \sum_{(i,j) \in [n]^2} |b_i p_i - b_j p_j|.$$

Notice that the term $\text{Welfare}_*(\mathcal{D}) - \frac{1}{4}\sum_{(i,j) \in [n]} |b_i p_i - b_j p_j|$ is independent of f. Thus, we have shown that the conditional risk lower bounds Welfare_f up to some problem dependent constants and scaling.

Recall from Appendix A.2 $\min_{f \in \mathcal{H}} \mathcal{R}(f; \mathcal{D}) = -\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n |b_i p_i - b_j p_j|$. Let $\widehat{f}$ be an arbitrary minimizer of $\mathcal{R}(f; \mathcal{D})$ and we know

$$\text{Welfare}_{\widehat{f}}(\mathcal{D}) \geqslant -\frac{1}{2}\mathcal{R}(\widehat{f}; \mathcal{D}) + \text{Welfare}_*(\mathcal{D}) - \frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n |b_i p_i - b_j p_j| = \text{Welfare}_*(\mathcal{D}).$$

Recall that $\text{Welfare}_*(\mathcal{D})$ is the maximum welfare achievable by definition. Thus the inequality is tight for any maximizer of $\mathcal{R}(f; \mathcal{D})$. □

A.4. Proof of Proposition 3.3

Proof. We begin by showing p^* is a minimizer of $\mathbb{E}_{\mathcal{D}}[\ell(f;\mathcal{D}) + \lambda h(f;\mathcal{D})]$. By Proposition 3.1, we know that p^* minimizes $\mathcal{R}(f;\mathcal{D})$ for all $\mathcal{D}$ and therefore, p^* minimizes $\mathbb{E}[\ell(f;\mathcal{D})]$ by extension. The claim then easily hold.

We then show that p^* is unique. Let $f' \neq p^*$ be arbitrary and fixed. Because p^* is the unique minimizer of $\mathbb{E}_{\mathcal{D}}[h(f;\mathcal{D})]$

$$\mathbb{E}_{\mathcal{D}}[h(f';\mathcal{D})] > \mathbb{E}_{\mathcal{D}}[h(p^*;\mathcal{D})],$$

and we emphasize that the inequality is strict. Moreover, because p^* is the minimizer of $\mathcal{R}(f;\mathcal{D})$ for all $\mathcal{D}$,

$$\mathbb{E}_{\mathcal{D}}[\ell(f';\mathcal{D})] \geqslant \mathbb{E}_{\mathcal{D}}[\ell(p^*;\mathcal{D})].$$

As $\lambda > 0$, we know p^*

$$\mathbb{E}_{\mathcal{D}}[\ell(f';\mathcal{D}) + \lambda h(f';\mathcal{D})] > \mathbb{E}_{\mathcal{D}}[\ell(p^*;\mathcal{D}) + \lambda h(p^*;\mathcal{D})]$$

and hence p^* is the unique minimizer. □

A.5. Proof of Theorem 3.4

We first show the claim holds for $\ell_\sigma^{\log}(f;\mathcal{D})$.

Proof for $\ell_\sigma^{\log}(f;\mathcal{D})$. Consider an arbitrary pair (i,j). For simplicity, let $\Delta_{ij} = b_i p_i - b_j p_j$ and $\Delta_{ij}^f = b_i f(x_i) - b_j f(x_j)$. We begin by writing out the parts of $\ell_\sigma^{\log}(f)$ and $\ell(f)$ that depend only on the pair (i,j), which are

$$\Delta_{ij}\log(1+\exp(-\sigma\Delta_{ij}^f)) - \Delta_{ij}\log(1+\exp(\sigma\Delta_{ij}^f)) \tag{10}$$

and

$$-\Delta_{ij}\,\mathbb{1}\{\Delta_{ij}^f \geqslant 0\} + \Delta_{ij}\,\mathbb{1}\{\Delta_{ij}^f \leqslant 0\}, \tag{11}$$

respectively. Without loss of generality we assume throughout the rest of this proof that $\Delta_{ij}^f \geqslant 0$ and know that

$$(10) - (11) = \Delta_{ij}\left(\log(1+\exp(-\sigma\Delta_{ij}^f)) - \log(1+\exp(\sigma\Delta_{ij}^f)) + 1\right),$$

which in turn implies

$$|(10) - (11)| \leqslant |\Delta_{ij}|\left|\log(1+\exp(-\sigma\Delta_{ij}^f)) - \log(1+\exp(\sigma\Delta_{ij}^f)) + 1\right|.$$

We now focus on the function $g(x) = \log(1+\exp(-\sigma x)) - \log(1+\exp(\sigma x)) + 1$. We quickly note that the function is monotonically decreasing in x and hence

$$\sup_{x\in[0,B]}|g(x)| \leqslant \max\{1, |\log(1+\exp(-\sigma B)) - \log(1+\exp(\sigma B)) + 1|\}$$

Since $\sigma, B > 0$, we always have $\log(1+\exp(\sigma B)) - \log(1+\exp(-\sigma B)) \geqslant 0$. Divide the problem to two cases.

1. When $\log(1+\exp(\sigma B)) - \log(1+\exp(-\sigma B)) \in [0,2)$. In this case we always have

$$\begin{aligned}&|\log(1+\exp(-\sigma B)) - \log(1+\exp(\sigma B)) + 1| \\ &\quad= |1 - (\log(1+\exp(\sigma B)) - \log(1+\exp(-\sigma B)))| \\ &\quad\leqslant 1\end{aligned}$$

   and thus $|g(x)| \leqslant 1$ for all $x \in [0,B]$.

2. When $\log(1+\exp(\sigma B)) - \log(1+\exp(-\sigma B)) \geqslant 2$, we have

$$\sup_{x\in[0,B]}|g(x)| = \log(1+\exp(\sigma B)) - \log(1+\exp(-\sigma B)) - 1.$$

Combine the two cases above, and we have for all $x \in [0, B]$, $g(x) \leqslant \max\{1, \log(1+\exp(\sigma B) - \log(1+\exp(-\sigma B)) - 1\}$. Recall that all bids are upper bounded by B and $f : \mathcal{X} \to [0, 1]$, and we know $\Delta_{ij}^f \in [0, B]$ under our assumption. Therefore, for any arbitrary pair (i, j) we always have

$$\begin{aligned}(10) - (11) &\leqslant \Delta_{ij} \left(\log(1 + \exp(-\sigma\Delta_{ij}^f)) - \log(1 + \exp(\sigma\Delta_{ij}^f)) + 1\right) \\ &\leqslant |\Delta_{ij}| \max\{1, \log(1 + \exp(\sigma B)) - \log(1 + \exp(-\sigma B)) - 1\},\end{aligned}$$

and summing the equation over all $\frac{n(n-1)}{2}$ unique pairs gives us

$$|\ell_\sigma^{\log}(f) - \ell(f)| \leqslant \frac{1}{2} \max\{1, \log(1 + \exp(\sigma B)) - \log(1 + \exp(-\sigma B)) - 1\} \sum_{i=1}^{n} \sum_{j=1}^{n} |b_i p_i - b_j p_j|.$$

Finally, rearrange the term $\log(1 + \exp(\sigma B)) - \log(1 + \exp(-\sigma B))$ and we have the equation

$$\log\left(\frac{1 + \exp(\sigma B)}{1 + \exp(-\sigma B)}\right) = 2,$$

which solves to $\sigma B = 2$, namely $\sigma = 2/B$. □

The proof for $\ell_\sigma^{\text{hinge}}(f; \mathcal{D})$ is similar.

Proof for $\ell_\sigma^{\text{hinge}}(f; \mathcal{D})$. Consider an arbitrary pair (i, j). For simplicity, let $\Delta_{ij} = b_i p_i - b_j p_j$ and $\Delta_{ij}^f = b_i f(x_i) - b_j f(x_j)$. We begin by writing out the parts of $\ell_\sigma^{\text{hinge}}(f)$ that depend only on the pair (i, j), which are

$$\Delta_{ij}(-\sigma\Delta_{ij}^f)_+ - \Delta_{ij}(\sigma\Delta_{ij}^f)_+ \tag{12}$$

and recall from (11) that the corresponding part of $\ell(f)$ is

$$-\Delta_{ij}\, \mathbb{1}\{\Delta_{ij}^f \geqslant 0\} + \Delta_{ij}\, \mathbb{1}\{\Delta_{ij}^f \leqslant 0\}.$$

Without loss of generality we assume that $\Delta_{ij}^f \geqslant 0$ and know that

$$(12) - (11) = \Delta_{ij} \left((-\sigma\Delta_{ij}^f)_+ - (\sigma\Delta_{ij}^f)_+ + 1\right) = \Delta_{ij} \left(1 - \sigma\Delta_{ij}^f\right).$$

Recall that $\Delta_{ij}^f \in [0, B]$. Therefore, for any arbitrary pair (i, j) we always have

$$|(12) - (11)| \leqslant |\Delta_{ij}| \left|1 - \sigma\Delta_{ij}^f\right| \leqslant |\Delta_{ij}| \max\{1, 1 - \sigma B\}$$

and summing the equation over all $\frac{n(n-1)}{2}$ unique pairs and applying triangle inequality gives us

$$|\ell_\sigma^{\log}(f) - \ell(f)| \leqslant \frac{1}{2} \max\{1, \sigma B - 1\} \sum_{i=1}^{n} \sum_{j=1}^{n} |b_i p_i - b_j p_j|.$$

Setting $\sigma = 2/B$ completes the proof. □

A.6. Proof of Proposition 3.5

Proof. Without loss of generality we restrict our focus to a pair of ads i, j where $b_i p_i \geqslant b_j p_j$. We then know that

$$\begin{aligned}\Pr(\mathbb{1}\{b_i y_i \geqslant b_j y_j\} \neq \mathbb{1}\{b_i p_i \geqslant b_j p_j\}) &= \Pr(b_i y_i \leqslant b_j y_j) \\ &= \Pr(y_i = 0 \wedge y_j = 1) = (1 - p_i) p_j.\end{aligned}$$

Set $p_j = 1 - \frac{1}{2}\epsilon$ and $p_i = 1 - \frac{\epsilon}{p_j}$. We first verify that $p_i, p_j \in (0, 1)$.

- For p_j, because $\epsilon \in \left(0, \frac{1}{2}\right)$, $p_j \in \left(\frac{3}{4}, 1\right)$ and is valid.
- For p_i, because $p_j \in \left(\frac{3}{4}, 1\right)$, $\frac{\epsilon}{p_j} \in \left(\epsilon, \frac{4}{3}\epsilon\right) \subset \left(0, \frac{2}{3}\right)$, which in turn shows $p_i \in (0, 1)$ and is valid.

Moreover, plug in the choices of p_i, p_j and we have

$$\Pr(\mathbb{1}\{b_i y_i \geqslant b_j y_j\} = \mathbb{1}\{b_i p_i \geqslant b_j p_j\}) = 1 - (1 - p_i)p_j = 1 - (1 - \epsilon) = \epsilon,$$

which is exactly what we wanted.

Lastly, we note that setting $b_i = \frac{2C}{p_i}, b_j = \frac{C}{p_j}$ ensures that $b_i p_i \geqslant b_j p_j$ for any $C \in \mathbb{R}_{>0}$. □

A.7. Proof of Theorem 4.1

Proof. Let $\mathcal{D}$ be an arbitrary and fixed dataset. We drop the notation $\mathcal{D}$ for the rest of the proof. Consider some pair (i, j) and an arbitrary, not necessarily optimal, pCTR function f. We write out the parts of the conditional risk $\mathcal{R}(f; \mathcal{D})$ that depend only on the pair,

$$-(b_i p^*(x_i) - b_j p^*(x_j))\, \mathbb{1}\{b_i f(x_i) > b_j f(x_j)\} - (b_j p^*(x_j) - b_i p^*(x_i))\, \mathbb{1}\{b_j f(x_j) > b_i f(x_i)\}.$$

Similarly, the parts of $\widehat{\ell}(f)$ that depend only on the pair (i, j) is

$$-(b_i \widehat{p}(x_i) - b_j \widehat{p}(x_j))\, \mathbb{1}\{b_i f(x_i) > b_j f(x_j)\} - (b_j \widehat{p}(x_j) - b_i \widehat{p}(x_i))\, \mathbb{1}\{b_j f(x_j) > b_i f(x_i)\}.$$

Omitting ties, we know one and exactly one of $\mathbb{1}\{b_i f(x_i) > b_j f(x_j)\}$ and $\mathbb{1}\{b_j f(x_j) > b_i f(x_i)\}$ is non-zero. Without loss of generality we assume $b_i f(x_i) > b_j f(x_j)$. The absolute value of the difference between the two pairs is then

$$\begin{aligned} |b_i \widehat{p}(x_i) - b_j \widehat{p}(x_j) - b_i p^*(x_i) - b_j p^*(x_j)| &\leqslant b_i |\widehat{p}(x_i) - p^*(x_i)| + b_j |\widehat{p}(x_j) - p^*(x_j)| \\ &\leqslant B|\widehat{p}(x_i) - p^*(x_i)| + B|\widehat{p}(x_j) - p^*(x_j)|, \end{aligned}$$

where for the second inequality we use the assumption that all bids are upperbounded by B. Summing the inequality above over all $\frac{n(n-1)}{2}$ pairs gives us

$$\mathbb{E}_{\mathcal{D}}[|\widehat{\ell}(f) - \mathcal{R}(f; \mathcal{D})|] \leqslant n(n-1)B\mathbb{E}_{x_i}[|\widehat{p}(x_i) - p^*(x_i)|] \leqslant n(n-1)B\sqrt{\epsilon},$$

where we use Jensen's inequality for the second inequality, completing the proof. □

A.8. Proof of Theorem 4.4

We begin with a slight detour and first prove the following mathematical proposition.

Proposition A.2. *For any pair of real numbers $a, b \in \mathbb{R}$, we have the following*

$$|a_+ - b_+| + |(-a)_+ - (-b)_+| = |a - b|.$$

Proof. We divide the proposition into four cases.

- When $a \geqslant 0, b \geqslant 0$. In this case left-hand side evaluates to $|a - b|$ and the equation holds.
- When $a < 0, b \geqslant 0$. In this case left-hand side evaluates to $|b| + |a| = b - a = |a - b|$ and the equation holds.
- When $a \geqslant 0, b < 0$. In this case left-hand side evaluates to $|a| + |b| = a - b = |a - b|$ and the equation holds.
- When $a < 0, b < 0$. In this case left-hand side evaluates to $|-a-(-b)| = |a - b|$ and the equation holds.

□

We also make use of the following helper function. Let

$$\ell^+(f;\mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{n}(b_i p_i - b_j p_j)_+ \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}. \tag{13}$$

At a high level, $\ell^+(f;\mathcal{D})$ is the version of $\ell(f;\mathcal{D})$ when we know exactly what $p^*(\cdot)$ is. The loss function is, unfortunately, difficult to estimate (recall Proposition 3.5), but remains tightly related to welfare. We have the following proposition, which remains useful for the rest of the section.

Proposition A.3. *We have the following inequality for all $f \in \mathcal{H}$ and $\mathcal{D}$*

$$\mathrm{Welfare}_f(\mathcal{D}) \geqslant \mathrm{Welfare}_*(\mathcal{D}) - \ell^+(f;\mathcal{D}).$$

Moreover, the inequality is tight for any minimizer of $\ell^+(f;\mathcal{D})$.

Proof. The proof is nearly the same as that of Theorem 3.2, which we provided in Appendix A.3. We detail the proof below for completeness.

Condition on an arbitrary and unordered pair of indices (i,j) and assume without loss of generality that $b_i p_i \geqslant b_j p_j$. The parts of $\ell^+(f;\mathcal{D})$ that depend only on the pair

$$(b_i p_i - b_j p_j)_+ \mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\} + (b_j p_j - b_i p_i)_+ \mathbb{1}\{b_j f(x_j) \leqslant b_i f(x_i)\} = (b_i p_i - b_j p_j)\mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}. \tag{14}$$

Recalling (9), the welfare suboptimality induced by the pair is exactly

$$(b_i p_i - b_j p_j)\mathbb{1}\{i = i^*\}\mathbb{1}\{j = j^*\}\mathbb{1}\{b_i f(x_i) \leqslant b_j f(x_j)\}, \tag{15}$$

where we note the assumption that $b_i p_i \geqslant b_j p_j$ implies $j \neq i^*$, where we recall $i^* = \mathrm{argmax}_{i' \in [n]} b_{i'} p_{i'}$. Immediately we note that

$$(15) \leqslant (14) \leqslant (15) + (b_i p_i - b_j p_j)_+.$$

Particularly (14) = (15), when $b_i f(x_i) \geqslant b_j f(x_j)$. We then sum over all pairs of (i,j) and know that

$$\mathrm{Welfare}_*(\mathcal{D}) - \mathrm{Welfare}_f(\mathcal{D}) \leqslant \ell^+(f;\mathcal{D}). \tag{16}$$

and the inequality is tight when $b_i f(x_i) \geqslant b_j f(x_j)$ whenever $b_i p_i \geqslant b_j p_j$. Since $p^* \in \mathcal{H}$, it is possible to exactly minimize $\ell^+(f;\mathcal{D})$ for all $\mathcal{D}$, thus any minimizer of $\ell^+(f;\mathcal{D})$ must rank each pair correctly. Therefore, the inequality is tight for any minimizer of $\ell^+(f;\mathcal{D})$ for any $\mathcal{D}$. □

We now proceed with the proof of Theorem 4.4 itself.

Proof of Theorem 4.4. Consider an arbitrary pair (i,j). For simplicity, let $\Delta_{ij} = b_i p_i - b_j p_j$, $\widehat{\Delta}_{ij} = b_i \widehat{p}_i - b_j \widehat{p}_j$, and $\Delta^f_{ij} = b_i f(x_i) - b_j f(x_j)$. We begin by writing out the part of $\widehat{\ell}^{\mathrm{hinge},+}_\sigma(f;\mathcal{D})$ that depend only on the pair (i,j)

$$(\widehat{\Delta}_{ij})_+(-\sigma\Delta^f_{ij})_+ + (-\widehat{\Delta}_{ij})_+(\sigma\Delta^f_{ij})_+. \tag{17}$$

We also introduce the loss function

$$\ell^{\mathrm{hinge},+}_\sigma(f;\mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{n}(\Delta_{ij})_+(-\sigma\Delta^f_{ij})_+,$$

and the part that corresponds to the pair (i,j) would be

$$(\Delta_{ij})_+(-\sigma\Delta^f_{ij})_+ + (-\Delta_{ij})_+(\sigma\Delta^f_{ij})_+. \tag{18}$$

We then know

$$\begin{aligned}|(17)-(18)| &= |(-\sigma\Delta_{ij}^f)_+((\widehat{\Delta}_{ij})_+ - (\Delta_{ij})_+) + (\sigma\Delta_{ij}^f)_+((-\widehat{\Delta}_{ij})_+ - (-\Delta_{ij})_+)| \\ &\leqslant \max\{(-\sigma\Delta_{ij}^f)_+, (\sigma\Delta_{ij}^f)_+\}(|(\widehat{\Delta}_{ij})_+ - (\Delta_{ij})_+| + |(-\widehat{\Delta}_{ij})_+ - (-\Delta_{ij})_+|) \\ &= \max\{(-\sigma\Delta_{ij}^f)_+, (\sigma\Delta_{ij}^f)_+\}|\widehat{\Delta}_{ij} - \Delta_{ij}| \\ &\leqslant \sigma B|\widehat{\Delta}_{ij} - \Delta_{ij}|,\end{aligned}$$

where for the second equality we use Proposition A.2 and the last inequality the fact that the bids are bounded by B. As $\mathbb{E}[(\widehat{p}(x) - p^*(x))^2] \leqslant \epsilon$,

$$\begin{aligned}\mathbb{E}[(\widehat{\Delta}_{ij} - \Delta_{ij})^2] &\leqslant 2\mathbb{E}[(b_i p_i - b_i\widehat{p}(x_i))^2] + 2\mathbb{E}[(b_j p_j - b_j\widehat{p}(x_j))^2] \\ &\leqslant 2B^2\mathbb{E}[(\widehat{p}(x) - p^*(x))^2] + 2B^2\mathbb{E}[(\widehat{p}(x) - p^*(x))^2] \leqslant 4B^2\epsilon\end{aligned}$$

where for the first inequality we recall the fact that $(a+b)^2 \leqslant 2a^2 + 2b^2$. By Jensen's inequality, we know

$$\mathbb{E}[|\widehat{\Delta}_{ij} - \Delta_{ij}|] \leqslant 2B\sqrt{\epsilon}.$$

Summing over all unique pairs of (i, j) and we know for any f

$$\mathbb{E}_{\mathcal{D}}[|\widehat{\ell}_\sigma^{\text{hinge},+}(f;\mathcal{D}) - \ell_\sigma^{\text{hinge},+}(f;\mathcal{D})|] \leqslant n(n-1)\sigma B^2\sqrt{\epsilon}. \tag{19}$$

We then control $\mathbb{E}[|\ell^{\text{hinge},+}(f) - \ell^+(f)|]$ similar to how we proved Theorem 3.4. Without loss of generality we assume that $\Delta_{ij} \geqslant 0$ and have

$$(18) - (14) = \Delta_{ij}((-\sigma\Delta_{ij}^f)_+ - 1)$$

and therefore

$$|(18) - (14)| \leqslant |\Delta_{ij}||(-\sigma\Delta_{ij}^f)_+ - 1| \leqslant |b_i p_i - b_j p_j| \max\{1, \sigma B - 1\}.$$

Summing the inequality across all pairs gives us

$$\begin{aligned}|\ell_\sigma^{\text{hinge},+}(f;\mathcal{D}) - \ell^+ f(;\mathcal{D})| &\leqslant \frac{1}{2}\max\{1, \sigma B - 1\}\sum_{i=1}^n\sum_{j=1}^n |b_i p_i - b_j p_j| \\ &\leqslant \frac{n(n-1)}{2}\max\{1, \sigma B - 1\}B.\end{aligned}$$

Taking the expectation over $\mathcal{D}$ and applying Jensen's inequality, we then know that

$$\mathbb{E}_{\mathcal{D}}[|\widehat{\ell}_\sigma^{\text{hinge},+}(f;\mathcal{D}) - \ell^+(f;\mathcal{D})|] \leqslant \frac{n(n-1)B}{2}(2\sqrt{\epsilon} + \max\{1, \sigma B - 1\}).$$

By Proposition A.3, we then know for any $f \in \mathcal{H}$

$$\mathbb{E}_{\mathcal{D}}[\text{Welfare}_f(\mathcal{D})] \geqslant \mathbb{E}_{\mathcal{D}}[\text{Welfare}_*(\mathcal{D})] - \mathbb{E}_{\mathcal{D}}[\widehat{\ell}_\sigma^{\text{hinge},+}(f;\mathcal{D})] - n(n-1)\sigma B^2\sqrt{\epsilon} - \frac{n(n-1)}{2}B\max\{1, \sigma B - 1\}.$$

We now show that p^*'s suboptimality tends to 0 as ϵ goes to 0. Recall the definition of $\ell_\sigma^{\text{hinge},+}(f;\mathcal{D})$, we have

$$\ell_\sigma^{\text{hinge},+}(p^*;\mathcal{D}) = \sum_{i=1}^n\sum_{j=1}^n(\Delta_{ij})_+(-\sigma\Delta_{ij})_+ = 0.$$

Plug the result back to (19), and we know

$$\mathbb{E}_{\mathcal{D}}[\widehat{\ell}_\sigma^{\text{hinge},+}(p^*;\mathcal{D})] \leqslant n(n-1)\sigma B^2\sqrt{\epsilon}.$$

Since $\widehat{\ell}_{\sigma}^{\text{hinge},+}(f;\mathcal{D}) \geqslant 0$ for any $f \in \mathcal{H}$ and $\mathcal{D}$, we have

$$\mathbb{E}_{\mathcal{D}}[\widehat{\ell}_{\sigma}^{\text{hinge},+}(p^*;\mathcal{D}) - \min_f \widehat{\ell}_{\sigma}^{\text{hinge},+}(f;\mathcal{D})] \leqslant n(n-1)\sigma B^2\sqrt{\epsilon}.$$

Since $\min_f \mathbb{E}_{\mathcal{D}}[\widehat{\ell}_{\sigma}^{\text{hinge},+}(f;\mathcal{D})] \leqslant \mathbb{E}_{\mathcal{D}}[\min_f \widehat{\ell}_{\sigma}^{\text{hinge},+}(f;\mathcal{D})]$, we know that

$$\mathbb{E}_{\mathcal{D}}[\widehat{\ell}_{\sigma}^{\text{hinge},+}(p^*;\mathcal{D})] - \min_f \mathbb{E}_{\mathcal{D}}[\widehat{\ell}_{\sigma}^{\text{hinge},+}(f;\mathcal{D})] \leqslant n(n-1)\sigma B^2\sqrt{\epsilon}$$

and the loss is $\mathcal{O}(\sqrt{\epsilon})$-approximately calibrated. □

B. Detailed Description for Experiments on Synthetic Dataset

Here we include a detailed discussion on how we conducted the experiments on synthetic data.

Data Generating Process. To generate the ground-truth CTRs, we draw a weight vector $w_p \in \mathbb{R}^{50}$ where each entry is follows a Unif$(-\sqrt{10}, \sqrt{10})$ distribution. We use a logistic model and set $p_i = \frac{1}{1+\exp(w_p^T x_i)+\xi_i^{\text{CTR}}}$, where ξ^{CTR} is a random noise following $\mathcal{N}(0, 0.1^2)$.

Similarly, in order to generate the bids, we draw a weight vector $w_b \in \mathbb{R}^{50}$ where each entry is drawn from Unif$(-2, 2)$. Inspired by the empirical observations made in Vasile et al. (2017), we assume the CPC bids, when conditioned on the feature x_i, follow a log-normal distribution and set $b_i = \exp(w_b^T x_i + \xi_i^{\text{bid}})$ where $\xi_i^{\text{bid}} \sim N(0, 0.1^2)$. The choice of parameters avoids overflowing due to the exponential term used for generating the bids.

We visualize in Figure 4 the distribution of the CTRs and CPC bids generated when $\sigma_{\text{CTR}} = 0.1$ as a sanity check. We can verify that the generated CTRs is unimodal and centered around 0.5, and hence does not follow a degenerate distribution. The distribution of the generated CPC bids roughly follows a log-normal distribution, agreeing with the empirical observations made in Vasile et al. (2017).

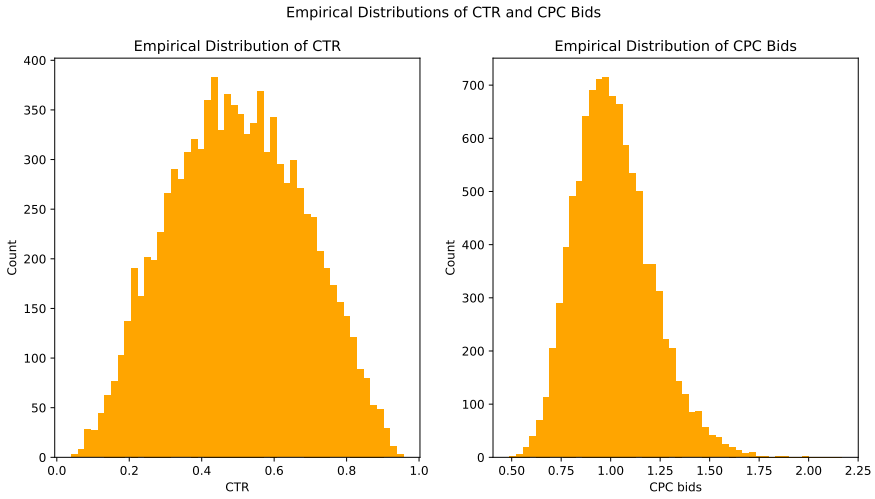

Figure 4. Distribution of the Generated CTR and CPC.

Training. We used Adam with default parameters. Batch size is set to 256. All models are trained till convergence.

Baselines. We use the logistic loss ℓ^{LL} and weighted logistic loss ℓ^{WLL} as baselines. Particularly, for any function f and advertisement $((b_i, x_i), y_i)$, the logistic loss is

$$\ell^{\text{LL}}(f(x_i), y_i) = -(y_i \log(f(x_i)) + (1-y_i)\log(1-f(x_i)))$$

and the weighted logistic loss is

$$\ell^{\mathrm{WLL}}(f(x_i), y_i, b_i) = -b_i(y_i \log(f(x_i)) + (1 - y_i)\log(1 - f(x_i))).$$

Summing them over all $((b_i, x_i), y_i) \in \mathcal{D}$ yields the empirical loss.

Proposed Losses. We begin by introducing the version of $\ell^{\log}_{\sigma}$ we used for the experiment setting, $\ell^{\log}_{\sigma=1}$.

$$\ell^{\log}_{\sigma=1}(f; \mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{n}(b_i y_i - b_j y_j)\log(1 + \exp(-(b_i f(x_i) - b_j f(x_j)))) + 3\sum_{i=1}^{n}\ell^{\mathrm{LL}}(f(x_i), y_i).$$

Intuitively, one can view $\ell^{\mathrm{LL}}(f(x_i), y_i)$ as a regularizer, avoiding the loss function to place too much emphasis on minimizing the ranking loss, and ensures our model can still adequately predict the CTRs. Similar to our choice of $\sigma = 1$, choosing 3 as the regularization strength of ℓ is arbitrary and we did not optimize over the space of possible regularization strengths.

We then write out the version of $\widehat{\ell}^{\text{hinge},+}_{\sigma}$ we employ, which, in addition to a logistic loss function as a regularizer, also uses logistic functions to weigh each pair. More formally, we have

$$\widehat{\ell}^{\text{hinge},+}_{\sigma=1}(f; \mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{1 + \exp(-3b_i\widehat{p}(x_i))}\frac{1}{1 + \exp(-3b_j f(x_j))}(b_i\widehat{p}_i - b_j\widehat{p}_j)_+(-(b_i f(x_i) - b_j f(x_j)))_+ + 3\sum_{i=1}^{n}\ell^{\mathrm{LL}}(f(x_i), y_i).$$

Particularly, here we use $\frac{1}{1+\exp(-3b_i\widehat{p}(x_i))}$ as a proxy to $\mathbb{1}\{i = i^*\}$, which we recall indicates whether the i-th ad has the highest ground-truth eCPM or not. Similarly, $\frac{1}{1+\exp(-3b_j f(x_j))}$ is the proxy to $\mathbb{1}\{j = j^*\}$. The choice to multiply $b_i\widehat{p}(x_i)$ by 3 in the denominator is ad hoc and arbitrary. We did not tune over the space of possible indicators and leave a rigorous treatise of the parameter choice, under additional assumptions on the eCPMs' distribution, as an interesting future direction.

Finally, we write out $\widehat{\ell}^{\log}_{\sigma=1}(f; \mathcal{D})$, which uses the same weighing scheme used in $\widehat{\ell}^{\text{hinge},+}_{\sigma=1}(f; \mathcal{D})$. We have

$$\widehat{\ell}^{\log}_{\sigma=1}(f; \mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{1 + \exp(-3b_i\widehat{p}(x_i))}\frac{1}{1 + \exp(-3b_j f(x_j))}(b_i\widehat{p}(x_i) - b_j\widehat{p}(x_j))\log(1 + \exp(-(b_i f(x_i) - b_j f(x_j)))) + 3\sum_{i=1}^{n}\ell^{\mathrm{LL}}(f(x_i), y_i).$$

For $\widehat{\ell}^{\text{hinge},+}_{\sigma=1}(f; \mathcal{D})$ and $\widehat{\ell}^{\log}_{\sigma=1}(f; \mathcal{D})$, the plug-in estimator $\widehat{p}(\cdot)$ is model trained with logistic loss. We train on the baseline losses first, and then use the model outputs as the plug-in estimates for $\widehat{\ell}^{\text{hinge},+}_{\sigma=1}(f; \mathcal{D})$ and $\widehat{\ell}^{\log}_{\sigma=1}(f; \mathcal{D})$.

Evaluation. As we know the exact CTR for each advertisement, we can directly calculate the social welfare achieved by the model trained with each of the five losses. For each model, we first pick out the ad with the highest predicted eCPM for each of the 2000 randomly generated auctions. We then calculate the actual CPM of the selected ad and record it as the social welfare achieved on that particular round of auction.

Calculating the Standard Error. Since the bulk of the randomness in test time social welfare lies in the data generation process, reporting the standard error of each model's average test time social welfare would not be informative, as the data generating process' noise dominates. We instead report the standard error of *the difference* between the particular model's social welfare and that of the average social welfare across the 5 models. The difference term is then able to account for the randomness within the data generating process, and we are measuring the standard error of the *comparative* performance of each model.

C. Detailed Description of Experiments on Criteo Dataset

Data Preprocessing. A standard practice is to preprocess the Criteo dataset using the approach used by the winners of the Criteo Display Advertising Challenge[2] (Chen et al., 2016; Guo et al., 2017; Lian et al., 2018; Cheng et al., 2016). The Criteo data contains 13 numerical feature columns with positive integer observations and we replace all observations with value greater than 2 with its log instead, namely

$$f(x) = \begin{cases} x & \text{if } x \leqslant 2 \\ \lfloor \log_2(x) \rfloor & \text{otherwise} \end{cases}.$$

For the remaining 26 categorical features, we replace the values that appear less than 10 times with a special value. We also use a 8-1-1 train-validation-test split, commonly used by existing works.

Model Implementation. To ensure our results can be easily reproduced, we use the open source implementation found in the DeepCTR package for both DeepFM and Deep & Cross Network (Shen, 2017).

Bid Generation. We use the DeepFM implementation found in Shen (2017) with default parameters. Each categorical variable is cast to a 4 dimensional embedding. The weights are initialized randomly. The deep network has 3 hidden layers with decreasing size, consisting of 256, 128, and 64 nodes each. ReLU activation is used for all layers save for the output layer, which uses a sigmoid activation function. The processed data is then fed into the network and then re-scaled to the range $[0, 1]$ (otherwise the outputs are too close to zero). We then add a $\mathcal{N}(0, 0.1^2)$ random noise and take the exponential over the scores. Letting x denote the feature, we use

$$b_i = \exp(c \cdot f(x_i) + \xi_i^{\text{bid}})$$

as the CPC bids, where c is a scaling constant ensuring $cf(x_i) \in [0, 1]$ and $\xi_i^{\text{bid}} \sim \mathcal{N}(0, 1)$.

Choice of Loss Functions. We begin by introducing the version of $\ell_\sigma^{\log}$ used for the Criteo experiments, $\ell_{\sigma=3}^{\log}$.

$$\ell_{\sigma=3}^{\log}(f; \mathcal{D}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{1 + \exp(-3b_i \widehat{p}(x_i))} \frac{1}{1 + \exp(-3b_j f(x_j))} (b_i y_i - b_j y_j) \log(1 + \exp(-3(b_i f(x_i) - b_j f(x_j)))) + \lambda \sum_{i=1}^{n} \ell^{\text{LL}}(f(x_i), y_i).$$

Compared to the version used for the simulation studies, for the Criteo simulations we further incorporate the weighing schemes discussed in Section 4.2. The value of σ is also adjusted slightly to better approximate the indicator function on the range of eCPMs in our setting.

The version of $\widehat{\ell}_\sigma^{\log}$ used is defined as follows. We have

$$\widehat{\ell}_{\sigma=3}^{\log}(f; \mathcal{D}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{1 + \exp(-3b_i \widehat{p}(x_i))} \frac{1}{1 + \exp(-3b_j f(x_j))} (b_i \widehat{p}(x_i) - b_j \widehat{p}(x_j)) \log(1 + \exp(-3(b_i f(x_i) - b_j f(x_j)))) + \lambda \sum_{i=1}^{n} \ell^{\text{LL}}(f(x_i), y_i),$$

and we slightly adjusted the choice of σ.

We perform a grid search to adjust the value of λ over the set $\{0.1, 1, 3, 5, 10\}$ based on the welfare, AUC loss, and logistic loss achieved by the different choices. We set $\lambda = 3$ for both DeepFM and DCN as the choice achieves significant increases in welfare with minor impact on AUC loss and logistic loss for both $\ell_{\sigma=3}^{\log}$ and $\widehat{\ell}_{\sigma=3}^{\log}$.

Evaluation. We take the test set and partition it into auctions with 256 bidders each. We use the models trained on the three losses to determine the winner of each auction, and record the *realized eCPM* (the observed $b_i y_i$'s) as the welfare of the auction. As the ground-truth CTR of the ads are unknown, we are forced to use the realized eCPM as proxies for measuring welfare.

[2] https://www.csie.ntu.edu.tw/~r01922136/kaggle-2014-criteo.pdf

Parameter Settings for DeepFM. We follow the optimal parameters specified in Guo et al. (2017) and construct a DeepFM with 3 hidden layers, each with 400 nodes using ReLU activation. Embedding dimension for the categorial variables is set to 10. Dropout rate is set to 0.5. For optimizing the model we used Adam with default parameters and set batch size to 256. The model is trained for 3 epochs.

Parameter Settings for DCN. We use the parameters set in Wang et al. (2017) and construct a DCN with 2 hidden layers, each with 1024 nodes using ReLU activation. Batch normalization is applied to the network and the number of cross layers is set to 6. For the categorical features, we set the dimensionality of the embedding as $6 \times (\text{feature cardinality})^{1/4}$. For optimizing the model we used Adam and set the batch size to 512. The model is trained for 150,000 steps.

Additional Experimental Results. We plot DeepFM's welfare, AUC loss, and logistic loss in Figure 5.

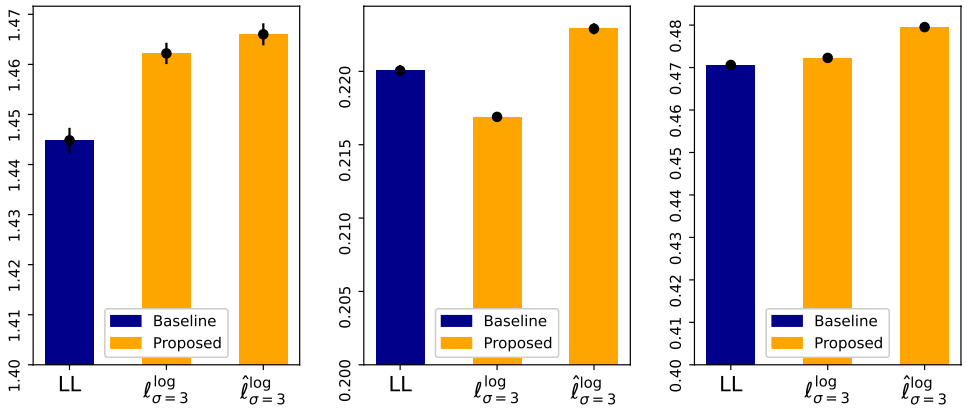

Figure 5. Detailed Performance Metrics for DeepFM. **Left: Welfare** (higher is better). **Center: AUC loss** (lower is better). **Right: Logistic loss** (lower is better). FFor each plot, from left to right: (baselines, blue) logistic loss, (proposed, yellow) $\ell^{\log}_{\sigma=3}$ (defined in (5)), and $\widehat{\ell}^{\log}_{\sigma}$ ((5) with y_i's replaced by outputs from teacher network).

We also plot DCN's welfare, AUC loss, and logistic loss in Figure 6.

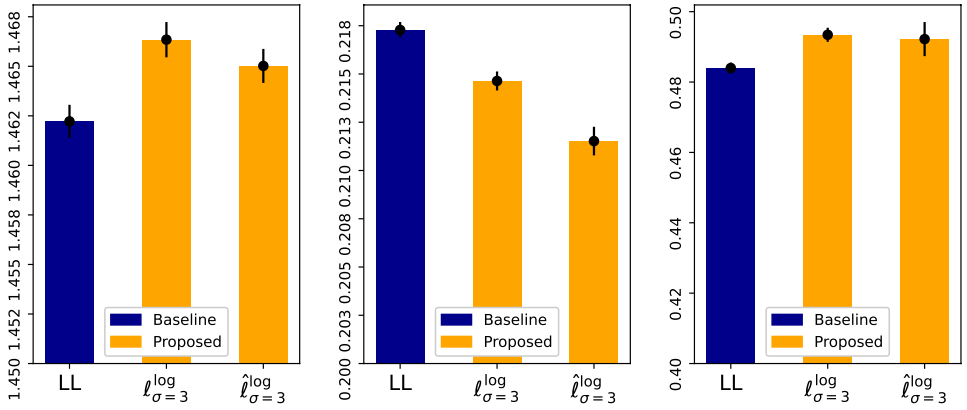

Figure 6. Detailed Performance Metrics for DCN. **Left: Welfare** (higher is better). **Center: AUC loss** (lower is better). **Right: Logistic loss** (lower is better). For each plot, from left to right: (baselines, blue) logistic loss, (proposed, yellow) $\ell^{\log}_{\sigma=3}$ (defined in (5)), and $\widehat{\ell}^{\log}_{\sigma}$ ((5) with y_i's replaced by outputs from teacher network).

The exact values of the metrics and the associated standard errors for DeepFM can be found in Table 2. The exact values of the metrics and the associated standard errors for DCN can be found in Table 3.

DeepFM			
	Welfare	AUC Loss	Logloss
Logistic Loss	1.4448 ± 0.0025	0.2200 ± 0.0004	0.4706 ± 0.0005
$\ell^{\log}_{\sigma=3}$	1.4622 ± 0.0021	0.2169 ± 0.0003	0.4723 ± 0.0009
$\widehat{\ell}^{\log}_{\sigma=3}$	$\mathbf{1.4660 \pm 0.0022}$	0.2229 ± 0.0004	0.4795 ± 0.0009

Table 2. Results for DeepFM.

DCN			
	Welfare	AUC Loss	Logloss
Logistic Loss	1.4622 ± 0.0026	0.2173 ± 0.0004	0.4840 ± 0.0015
$\ell^{\log}_{\sigma=3}$	$\mathbf{1.4663 \pm 0.0028}$	0.2146 ± 0.0005	0.4934 ± 0.0020
$\widehat{\ell}^{\log}_{\sigma=3}$	1.4650 ± 0.0027	0.2115 ± 0.0007	0.4922 ± 0.0048

Table 3. Results for DCN.

As we can see from the results, the proposed methods significantly improve welfare at a minimal cost (if any) to AUC loss. While logistic loss seems to be negatively affected, the impact is relatively small and we conjecture that better tuning the model architecture could resolve the issue.

Finally, we report below the per epoch runtime of the methods.

DCN			
	Logloss	Pairwise Loss ($\ell^{\log}_{\sigma=3}$)	Pairwise Loss + Student-Teacher ($\widehat{\ell}^{\log}_{\sigma=3}$)
Absolute	2635.13 ± 10.86	2700.87 ± 8.91	2700.4 ± 6.59
Relative	100%	102.5%	102.5%

Table 4. Runtime (in seconds) comparison for DCN. We take the average over 15 epochs and report the standard deviation.

These results show that for the more complex DCN model the added cost of the pairwise losses is relatively negligible. For DeepFM, the added computation cost is around 2.5% (60 seconds), which is a reasonable price to pay for significantly improved welfare performance. It is possible that our implementation of the loss is not the most efficient, and additional optimizations may further decrease the overhead.

DCN			
	Logloss	Pairwise Loss ($\ell^{\log}_{\sigma=3}$)	Pairwise Loss + Student-Teacher ($\widehat{\ell}^{\log}_{\sigma=3}$)
Absolute	5292.2 ± 57.39	5333.67 ± 56.62	5330 ± 48.11
Relative	100%	100.8%	100.7%

Table 5. Runtime (in seconds) comparison for DCN. We take the average over 15 epochs and report the standard deviation.