Distributionally Robust Bidding in First-Price Auctions

Yunfan Zhang, Chen Cheng, Suofei Feng, and Zhengyuan Zhou

Abstract

In this paper, we study bidder-side distributionally robust offline learning in repeated first-price auctions. Using historical data collected under a baseline policy and only censored transaction-price feedback, the bidder seeks a policy from a prescribed policy class that maximizes worst-case expected reward over all joint distributions within a Kullback–Leibler (KL) divergence ball around a stationary nominal model. This formulation hedges against both covariate shifts in item features and opponent shifts in competitors' behavior. We consider two settings depending on whether the feature distribution is known (*informative*) or unknown (*uninformative*). For each setting, we propose bidding policies based on specific estimates of the competitors' highest bids. We further derive a tractable dual reformulation via strong duality and prove that the resulting policy achieves a robust sub-optimality gap of order $O(T^{-1/2})$ in the informative case for a general policy class. In the uninformative case, we obtain an $O(\sqrt{(d+1)/T})$ bound when the policy class has finite pseudo-dimension d. Finally, numerical experiments illustrate the benefit and cost of robustness, learning from censored feedback, and the effect of policy-class complexity.

1 Introduction

In the rapidly expanding realm of e-commerce, propelled by a surge across various sectors [1, 2, 3, 4, 5], digital advertising has emerged as the predominant force in today's market. Historically, second-price auctions, exemplified by the Vickrey auction [6], dominated the online ad auction landscape due to their truthfulness. However, to simplify the auction process and improve transparency for market participants, major ad exchanges like Google, AppNexus, Index Exchange, and OpenX transitioned to first-price auctions by 2019 [7, 8, 9]. This structural shift has sparked a surge of interest in bidder-side learning under first-price auctions, where bidders must strategically shade their bids below their true valuations to optimize outcomes.

We study a contextual first-price auction model from a robust optimization perspective. At each round, the bidder observes a context, assigns a value to the impression, and submits a bid. The bidder's reward depends on the distribution of the highest competing bid, but this distribution is unknown. The key difficulty is informational: after each auction, the bidder observes only the transaction price, so the highest competing bid is revealed only when she loses. This censored feedback is richer than pure win/loss bandit feedback [10], but substantially weaker than full bid

Y. Zhang and Z. Zhou are with the Department of Technologies, Operations and Statistics, New York University. {yz11751, zz26}@stern.nyu.edu

C. Cheng is with the Department of Statistics, University of Illinois Urbana-Champaign. chcheng@illinois.edu

S. Feng is with Meta. suofeif@alumni.stanford.edu

revelation, and it prevents direct empirical estimation of the highest competing bid distribution; see also [11] for learning under censored transaction-price feedback.

Our goal is not to model fully adversarial or nonstationary dynamics as in [12, 13, 14, 15]. Instead, we consider a practically common offline-learning situation: a bidder has historical data generated under a nominal stationary environment, but the deployment environment may differ moderately because the traffic mix, competing bidders, or context distribution changes. A similar motivation underlies distributionally robust contextual bandits [16], where policies learned from historical data can fail when the future environment differs from the past one. We therefore seek a policy that is robust to moderate misspecification of the nominal model.

Main Contributions. We study an offline contextual bidder-side learning problem in repeated first-price auctions. Given historical observations collected under a baseline policy, the learner outputs a robust optimal bidding policy from a prescribed candidate class, whose performance is measured by the distributionally robust sub-optimality gap to the best policy. Building on the strong duality of [17], we demonstrate that the worst-case value of a fixed policy reduces to a one-dimensional dual optimization in Lemma 1. This turns the original infinite-dimensional minimax problem into a statistical estimation problem under censored transaction price feedback, where the main challenge is recovering the competing bid distribution.

We consider two settings in this paper. In the *informative case*, where the context distribution is known, we construct a clipped ratio estimator of winning probabilities and, under the stated coverage and zero-reward-mass assumptions, prove the learned policy achieves a robust sub-optimality gap of $O(T^{-1/2})$ in Theorem 1. In the *uninformative case*, where the context distribution is unknown, we leverage the empirical bid CDF under the baseline policy to establish an $O(\sqrt{(d+1)/T})$ sub-optimality bound for policy classes with finite pseudo-dimension d in Theorem 2.

Related work. *Bidder-side first-price auction learning.* The closest theoretical comparators study bidder-side learning under bandit or adversarial models: contextual bandits with cross-learning [10], stochastic censored-feedback learning [11], hint-assisted bidding [13], and contextual bidding under stochastic or adversarial competition [12, 18, 19]. All target low nominal regret under bandit or adversarial uncertainty; none imposes a KL-robustness constraint on a stationary nominal model. Non-stationary and dynamic-pricing models [14, 15, 20, 21] address temporal drift rather than distributional misspecification, and are thus orthogonal to our setting.

Robust mechanism design and pricing. Mohri and Medina [22] study reserve-price learning in second-price auctions with *full* bid information; Golrezaei et al. [23] studies seller-side robust reserve pricing in repeated contextual second-price auctions; and Si et al. [16] study offline distributionally robust batch contextual bandits under environment shift. The closest first-price robustness paper is [24], which develops KL-based robust bid shading under limited feedback; in contrast, we analyze a policy class and provide nonasymptotic sub-optimality guarantees for a KL-DRO objective.

Independence and contextual competition. A common competing-bid distribution is also used by Han et al. [11]; Kumar and Mouchtaki [25] explicitly assume independence between competition and the bidder's value–bid pair. Our nominal independence assumption is appropriate when x represents bidder-private information within a fixed public market segment; it need not hold across pooled segments. In contrast, [19, 26] model competition through a linear signal and noise, estimating $G(u \mid x) = F_\epsilon(u - \theta^\top x)$. These settings require conditional-distribution estimation; our ratio identity $L = H^{\pi_0} G$ addresses a common marginal G and generally fails without nominal independence.

2 Model and Dual Reformulation

2.1 Model Setup

We study an offline contextual first-price auction problem under a stationary nominal model. Suppose a dataset of T observations is collected under a baseline policy π_0. At each round $t = 1, 2, \ldots, T$, an impression arrives with feature $x_t \in \mathcal{X}$ and associated value $v(x_t) \in [0, 1]$, where $\{x_t\}_{t=1}^T$ are drawn i.i.d. from a distribution Q and the value function $v : \mathcal{X} \to [0, 1]$ is assumed known. Given the feature x_t, the bidder submits a bid $b_t = \pi_0(x_t)$, where $\pi_0 : \mathcal{X} \to [0, 1]$ is the baseline policy. Let $m_t \in [0, 1]$ denote the highest competing bid at round t, where $\{m_t\}_{t=1}^T$ are drawn i.i.d. from an unknown distribution P, independently of $\{x_t\}_{t=1}^T$. The platform runs a *first-price auction*: the bidder wins if and only if $b_t \geq m_t$, in which case she pays b_t; otherwise she loses and pays nothing. Crucially, the bidder does not observe m_t directly; she observes only the censored transaction price $y_t := \max\{b_t, m_t\}$.

Policies and Rewards. To avoid negative payoffs, we consider a policy class Π such that

$$0 \leq \pi(x) \leq v(x) \leq 1, \qquad \forall \pi \in \Pi,\ \forall x \in \mathcal{X}.$$

For any policy $\pi \in \Pi$, the realized reward is defined as

$$r_\pi(x, m) := (v(x) - \pi(x))\mathbf{1}_{\pi(x) \geq m}. \tag{1}$$

Under the nominal joint distribution $J = Q \times P$, the expected reward is given by

$$\bar{r}(\pi) := \mathbb{E}_{(x,m)\sim J}[r_\pi(x, m)] = \mathbb{E}_{x\sim Q}\big[(v(x) - \pi(x))G(\pi(x))\big], \tag{2}$$

where $G(u) := \mathbb{P}_{m\sim P}(m \leq u)$ for $u \in [0, 1]$ is the CDF of the highest competing bid.

The Censored CDF Identity. Evaluating this nominal reward requires knowledge of $G(\cdot)$, which is unknown. To bridge the gap between the unobserved G and our censored dataset $\mathcal{D}_T$, we define the bid CDF induced by a generic policy π as $H^\pi(u) := \mathbb{P}(\pi(x) \leq u) = \mathbb{E}_{x\sim Q}\mathbf{1}_{\pi(x)\leq u}$, and define the CDF of the observed transaction price under the baseline policy π_0 as $L(u) := \mathbb{P}(y_t \leq u)$. Since $y_t = \max\{\pi_0(x_t), m_t\}$ and x_t is independent of m_t, we establish the key identity:

$$L(u) = \mathbb{P}(\pi_0(x_t) \leq u, m_t \leq u) = H^{\pi_0}(u)G(u). \tag{3}$$

This relation forms the foundation for estimating $G(\cdot)$. The corresponding empirical CDFs derived from $\mathcal{D}_T$ are

$$L_T(u) := \frac{1}{T}\sum_{t=1}^T \mathbf{1}_{y_t\leq u}, \qquad H_T^{\pi_0}(u) := \frac{1}{T}\sum_{t=1}^T \mathbf{1}_{b_t\leq u}. \tag{4}$$

Distributionally Robust Objective. To safely leverage these empirical estimates while hedging against misspecification of the nominal model J, we adopt a KL-based distributionally robust optimization framework. For a fixed radius $\rho \geq 0$, we define the ambiguity set as

$$\mathcal{U}_\rho(J) := \{J' \ll J : D_{\mathrm{KL}}(J'\|J) \leq \rho\}.$$

Only the nominal center $J = Q \times P$ is required to factorize. Alternative laws J' may be dependent; the factorization in the dual expectation below uses only the nominal law J. For each policy $\pi \in \Pi$, its worst-case expected value over this ambiguity set is

$$V(\pi) := \inf_{J'\in\mathcal{U}_\rho(J)} \mathbb{E}_{(x,m)\sim J'}[r_\pi(x, m)], \tag{5}$$

and the distributionally robust optimal policy is

$$\pi^{DR} \in \arg\max_{\pi \in \Pi} V(\pi). \tag{6}$$

Given the historical dataset $\mathcal{D}_T$, our learning algorithm outputs a candidate policy $\hat{\pi}_T \in \Pi$. The performance of this learned policy is evaluated via its DRO sub-optimality gap,

$$R_{\mathrm{DRO}}(\hat{\pi}_T) := V(\pi^{DR}) - V(\hat{\pi}_T), \tag{7}$$

which is strictly nonnegative by the optimality of π^{DR}.

2.2 Strong Duality

The minimization in (5) is infinite-dimensional and is hard to analyze directly. By using the KL-divergence properties established in [17], we can reduce this original formulation to a tractable, one-dimensional concave maximization problem. Applying this strong duality to our framework yields the following result.

Lemma 1 (Strong Duality)**.** *For any policy $\pi \in \Pi$,*

$$\begin{aligned} V(\pi) &= \sup_{\alpha \geq 0} \left\{ -\alpha \log \mathbb{E}_{(x,m)\sim J} \left[\exp\left(-\frac{r_\pi(x,m)}{\alpha} \right) \right] - \alpha\rho \right\} \\ &= \sup_{\alpha \geq 0} \left\{ -\alpha \log\big(1 - \phi(\alpha,\pi)\big) - \alpha\rho \right\}, \end{aligned}$$

where the dual index is defined as

$$\phi(\alpha,\pi) := \mathbb{E}_{x\sim Q} \left[G(\pi(x)) \left(1 - \exp\left(-\frac{v(x) - \pi(x)}{\alpha} \right) \right) \right]. \tag{8}$$

At $\alpha = 0$, the dual objective is interpreted by its limit, $\operatorname{ess\,inf}_J r_\pi$. For $\rho = 0$, the supremum equals the nominal mean reward. We use the same endpoint convention for estimated reward laws below.

Denoting the dual objective by

$$\varphi(\alpha,\pi) := -\alpha \log\big(1 - \phi(\alpha,\pi)\big) - \alpha\rho, \tag{9}$$

Lemma 1 effectively reduces the original DRO problem to a one-dimensional dual maximization for each fixed policy. Consequently, the distributionally robust optimal policy is given by:

$$\pi^{DR} \in \arg\max_{\pi \in \Pi} \sup_{\alpha \geq 0} \varphi(\alpha,\pi). \tag{10}$$

The remaining challenge is statistical. Because the competing bid CDF $G(\cdot)$ is unknown, the dual index $\phi(\alpha,\pi)$ and hence the dual objective $\varphi(\alpha,\pi)$ must be estimated from the censored dataset $\mathcal{D}_T$.

2.3 Generic Sub-optimality Bound

Let $\varphi_T^{est}(\alpha, \pi)$ denote a data-driven estimator of the dual objective $\varphi(\alpha, \pi)$ constructed from $\mathcal{D}_T$. The learned policy is obtained via empirical maximization:

$$\hat{\pi}_T \in \arg\max_{\pi \in \Pi} \sup_{\alpha \geq 0} \varphi_T^{est}(\alpha, \pi). \tag{11}$$

In Section 3, φ_T^{est} will be instantiated by the informative estimator $\hat{\varphi}_T$, while in Section 4 it will be replaced by the uninformative estimator $\check{\varphi}_T$.

By (10), (7), and the optimality of $\hat{\pi}_T$ for the estimated objective, we derive the standard sub-optimality decomposition:

$$\begin{aligned} R_{\mathrm{DRO}}(\hat{\pi}_T) &= \sup_{\pi \in \Pi} \sup_{\alpha \geq 0} \varphi(\alpha, \pi) - \sup_{\alpha \geq 0} \varphi(\alpha, \hat{\pi}_T) \\ &\leq \sup_{\pi \in \Pi} \sup_{\alpha \geq 0} \varphi(\alpha, \pi) - \sup_{\alpha \geq 0} \varphi_T^{est}(\alpha, \hat{\pi}_T) \\ &\quad + \sup_{\alpha \geq 0} \varphi_T^{est}(\alpha, \hat{\pi}_T) - \sup_{\alpha \geq 0} \varphi(\alpha, \hat{\pi}_T) \\ &\leq 2 \sup_{\pi \in \Pi} \sup_{\alpha \geq 0} \left| \varphi_T^{est}(\alpha, \pi) - \varphi(\alpha, \pi) \right|. \end{aligned} \tag{12}$$

Thus, bounding the DRO sub-optimality gap reduces to establishing a uniform convergence bound for the estimated dual objective over $\Pi \times [0, \infty)$.

3 The Informative Case

We first consider the informative case, where the context distribution Q is known. In this setting, the bid CDF induced by the baseline policy is available, so the main task is to estimate the competing bid CDF G from censored transaction-price data.

3.1 Estimating G When Q Is Known

Since Q is known, the bidder has access to the CDF of the baseline bid:

$$H^{\pi_0}(u) = \mathbb{E}_{x \sim Q} \mathbf{1}_{\pi_0(x) \leq u}, \quad u \in [0, 1]. \tag{13}$$

The only unknown in the dual objective (9) is $G(\cdot)$. Direct empirical estimation of G is infeasible because m_t is not observed every round; however, by (3), the transaction-price CDF satisfies $L(u) = H^{\pi_0}(u)G(u)$, which suggests a ratio estimator.

We first show that the empirical transaction price CDF, $L_T(\cdot)$ from (4), provides an unbiased estimator.

Lemma 2 (Unbiased Transaction-Price CDF)**.** *For each $u \in [0, 1]$, we have*

$$\mathbb{E}[L_T(u)] = L(u) = H^{\pi_0}(u)G(u).$$

Since $H^{\pi_0}(\cdot)$ is known in the informative setting, we then define the bounded estimate for $G(\cdot)$ as:

$$\hat{G}_T(u) := \mathrm{clip}\left(\frac{L_T(u)}{H^{\pi_0}(u)}, 0, 1 \right). \tag{14}$$

Here $\mathrm{clip}(z, 0, 1) := \min\{1, \max\{0, z\}\}$. Clipping prevents values above one and cannot increase the estimation error. To ensure that its denominator is bounded away from zero, we impose the following assumption.

Assumption 1. *Suppose there is a known constant $0 < \varepsilon_0 \leq 1$ such that $H^{\pi_0}(0) \geq \varepsilon_0$.*

Assumption 2. *There is a constant $\kappa > 0$ such that $\inf_{\pi \in \Pi} \mathbb{P}_J(r_\pi = 0) \geq \kappa$.*

Remark 1. *Assumption 1 ensures $H^{\pi_0}(u) \geq \varepsilon_0$ for all $u \in [0, 1]$, providing coverage for estimating G; it plays a role analogous to the overlap condition in [16, Assumption 1.2]. Assumption 2 controls the stability of the KL-robust value through a uniform lower bound on zero-reward mass. Related positivity conditions are used in [16, Assumption 2.2 and Lemma 6], which lower-bound conditional probabilities at every point of a finite reward support; here only the unconditional mass at zero is required.*

3.2 Learning Rule

Substituting the estimator $\hat{G}_T$ into (8), we construct the empirical dual index:

$$\hat{\phi}_T(\alpha, \pi) := \mathbb{E}_{x \sim Q}\left[\hat{G}_T(\pi(x))\left(1 - \exp\left(-\frac{v(x) - \pi(x)}{\alpha}\right)\right)\right]. \tag{15}$$

The corresponding estimated dual objective is

$$\hat{\varphi}_T(\alpha, \pi) := -\alpha \log\big(1 - \hat{\phi}_T(\alpha, \pi)\big) - \alpha\rho. \tag{16}$$

The complete procedure for identifying the optimal robust policy in the informative setting is summarized in Algorithm 1.

Algorithm 1 Informative-Case Robust Bidding Policy

1: **Input:** Dataset $\mathcal{D}_T = \{(x_t, b_t, y_t)\}_{t=1}^T$, policy class Π, feature distribution Q, radius ρ.
2: Compute $L_T(u)$ by (4) and $H^{\pi_0}(u)$ by (13).
3: Compute the clipped estimate $\hat{G}_T(u)$ by (14).
4: **for** each $\pi \in \Pi$ **do**
5: Form $\hat{\phi}_T(\alpha, \pi)$ by (15) and $\hat{\varphi}_T(\alpha, \pi)$ by (16).
6: Compute $\hat{V}_T(\pi) = \sup_{\alpha \geq 0} \hat{\varphi}_T(\alpha, \pi)$.
7: **end for**
8: **return** $\hat{\pi}_T \in \arg\max_{\pi \in \Pi} \hat{V}_T(\pi)$

3.3 Sub-optimality Bound

To sharpen the bound in (12), we introduce auxiliary reward laws. For a fixed policy $\pi \in \Pi$, let P^π denote the distribution of $r_\pi(x, m)$ under the nominal model $J = Q \times P$. Since $r_\pi(x, m) = (v(x) - \pi(x))\mathbf{1}_{\pi(x) \geq m}$, the reward takes value 0 when $\pi(x) < m$ and value $v(x) - \pi(x)$ when $\pi(x) \geq m$. Therefore, the CDF of P^π, for $u \in [0, 1]$, is given by

$$\begin{aligned} F_{P^\pi}(u) &:= \mathbb{P}(r_\pi(x, m) \leq u) \\ &= 1 - \mathbb{E}_{x \sim Q}\big[G(\pi(x))\mathbf{1}_{v(x) - \pi(x) > u}\big]. \end{aligned} \tag{17}$$

Since $\hat{G}_T$ need not be a CDF, define the reward law directly:

$$\hat{P}^\pi = \mathbb{E}_{x\sim Q}\left[(1-\hat{G}_T(\pi(x)))\delta_0 + \hat{G}_T(\pi(x))\delta_{v(x)-\pi(x)}\right].$$

Its CDF, for $u \in [0,1]$, is

$$F_{\hat{P}^\pi}(u) := 1 - \mathbb{E}_{x\sim Q}\left[\hat{G}_T(\pi(x))\mathbf{1}_{v(x)-\pi(x)>u}\right]. \tag{18}$$

The next lemma rewrites the dual objectives in terms of these auxiliary reward distributions and reduces the sub-optimality analysis to a uniform bound on the CDF estimation gap.

Lemma 3 (Informative-Case Sub-optimality Reduction)**.** *Suppose P^π and $\hat{P}^\pi$ are defined by* (17)–(18)*. Then, it holds that*

$$\varphi(\alpha,\pi) = -\alpha \log \mathbb{E}_{r\sim P^\pi} \exp(-r/\alpha) - \alpha\rho, \tag{19}$$

$$\hat{\varphi}_T(\alpha,\pi) = -\alpha \log \mathbb{E}_{r\sim \hat{P}^\pi} \exp(-r/\alpha) - \alpha\rho. \tag{20}$$

Moreover, for the learning policy $\hat{\pi}_T \in \Pi$ obtained by Algorithm 1, under Assumption 2, we have

$$R_{\mathrm{DRO}}(\hat{\pi}_T) \le \frac{4}{\kappa} \sup_{\pi\in\Pi} \sup_{u\in[0,1]} |F_{\hat{P}^\pi}(u) - F_{P^\pi}(u)|. \tag{21}$$

This reformulation allows us to use tools from statistical learning, such as Rademacher complexity, to derive tight sub-optimality bounds. Theorem 1 offers a specific result based on this approach.

Theorem 1 (Informative-Case Sub-optimality Bound)**.** *For any policy class Π, suppose that Assumptions 1 and 2 hold. Then,*

$$R_{\mathrm{DRO}}(\hat{\pi}_T) \le \frac{4}{\kappa}\left[\frac{1}{\varepsilon_0\sqrt{T}} + \delta\right],$$

with probability at least $1-\exp(-T\varepsilon_0^2\delta^2/2)$, for any $\delta \ge 0$.

Proof Sketch. By Lemma 3, bounding the sub-optimality gap reduces to controlling the uniform CDF gap in (21). From (17)–(18), we have

$$\sup_{\pi\in\Pi}\sup_{u\in[0,1]} \left|F_{\hat{P}^\pi}(u) - F_{P^\pi}(u)\right| \le \sup_{s\in[0,1]} \left|\hat{G}_T(s) - G(s)\right|.$$

By Assumption 1 and the monotonicity of H^{π_0},

$$\sup_{s\in[0,1]} \left|\hat{G}_T(s) - G(s)\right| \le \frac{1}{\varepsilon_0} \sup_{s\in[0,1]} |L_T(s) - L(s)|.$$

The final term is the uniform deviation over the threshold function class $\mathcal{F}_{\pi_0,1} := \{(x,b,y) \mapsto \mathbf{1}_{y\le u} : u \in [0,1]\}$. Because $\mathscr{F}_{\pi_0,1}$ has a finite Vapnik-Chervonenkis (VC) dimension, standard symmetrization arguments [27, Theorem 4.10] bound this uniform deviation by a multiple of its Rademacher complexity, $\mathscr{R}_T(\mathscr{F}_{\pi_0,1})$. Applying the VC property [27, Example 5.24] yields $\mathscr{R}_T(\mathscr{F}_{\pi_0,1}) = O(T^{-1/2})$. For an explicit constant, the Dvoretzky–Kiefer–Wolfowitz inequality [28] gives

$$\mathbb{P}\left(\|L_T - L\|_\infty > \frac{1}{\sqrt{T}} + \varepsilon_0\delta\right) \le 2e^{-2(1+\sqrt{T}\varepsilon_0\delta)^2} \le e^{-T\varepsilon_0^2\delta^2/2}.$$

Combining this with the preceding bounds proves the theorem.

4 The Uninformative Case

In this section, we consider an uninformative setting where the bidder lacks knowledge of the feature's distribution Q. Consequently, both H^{π_0} and G must be estimated from the same censored sample, introducing an extra layer of statistical error compared to the informative case.

4.1 Estimating H^{π_0} and G When Q Is Unknown

Since $H^{\pi_0}(\cdot)$ cannot be computed without Q, we replace it by the empirical bid CDF

$$H_T^{\pi_0}(u) := \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{b_t \le u}. \tag{22}$$

It is easy to verify that $H_T^{\pi_0}(\cdot)$ serves as an unbiased estimate of $H^{\pi_0}(\cdot)$. In addition, we adopt the same unbiased estimate $L_T(u) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{y_t \le u}$ for $H^{\pi_0}(\cdot)G(\cdot)$ as defined in (3). Then, we construct a bounded estimate of $G(\cdot)$ by protecting the denominator of the ratio of these two empirical distributions. Specifically, we define $\check{G}_T(\cdot)$ as:

$$\check{G}_T(u) := \frac{L_T(u)}{\max\{H_T^{\pi_0}(u), \varepsilon_0\}}. \tag{23}$$

The ratio (23) is generally biased, and both the numerator and the denominator are random. Its analysis requires separate control of the errors in L_T and $H_T^{\pi_0}$. We use the population coverage bound in Assumption 1, so that

$$\max\{H_T^{\pi_0}(u), \varepsilon_0\} \ge \varepsilon_0, \qquad u \in [0,1]. \tag{24}$$

Moreover, because $y_t \le u$ implies $b_t \le u$, we have $L_T(u) \le H_T^{\pi_0}(u)$, and hence $\check{G}_T(u) \in [0,1]$.

4.2 Learning Rule

Using (23) and the empirical context distribution, we define the estimated dual index as

$$\check{\phi}_T(\alpha, \pi) := \frac{1}{T}\sum_{t=1}^{T} \check{G}_T(\pi(x_t)) \left(1 - \exp\left(-\frac{v(x_t) - \pi(x_t)}{\alpha}\right)\right). \tag{25}$$

The corresponding estimated dual objective is given by

$$\check{\varphi}_T(\alpha, \pi) := -\alpha \log\big(1 - \check{\phi}_T(\alpha, \pi)\big) - \alpha\rho. \tag{26}$$

We summarize the complete procedure for identifying the optimal robust policy in Algorithm 2.

4.3 Sub-optimality Bound

Using the same method to bound the sub-optimality gap (12), we first introduce the distribution $\hat{P}_T^{\pi}$ to reformulate the dual function $\check{\varphi}_T(\alpha, \pi)$. Define the valid reward mixture

$$\hat{P}_T^{\pi} = \frac{1}{T}\sum_{t=1}^{T} \left[(1 - \check{G}_T(\pi(x_t)))\delta_0 + \check{G}_T(\pi(x_t))\delta_{v(x_t)-\pi(x_t)}\right].$$

Algorithm 2 Uninformative-Case Robust Bidding Policy

Input: Dataset $\{(x_t, b_t, y_t)\}_{t=1}^T$, policy class Π, known coverage lower bound ε_0, radius ρ.
Compute $L_T(u)$ by (4) and $H_T^{\pi_0}(u)$ by (22).
Compute the protected ratio $\check{G}_T(u)$ by (23).
for each $\pi \in \Pi$ **do**
    Form $\check{\varphi}_T(\alpha, \pi)$ by (26) and $\check{\phi}_T(\alpha, \pi)$ by (25).
    Compute $\check{V}_T(\pi) = \sup_{\alpha \geq 0} \check{\varphi}_T(\alpha, \pi)$.
end for
return $\hat{\pi}_T = \arg\max_{\pi \in \Pi} \check{V}_T(\pi)$.

Its CDF for $u \in [0, 1]$ is

$$F_{\hat{P}_T^\pi}(u) = 1 - \frac{1}{T} \sum_{t=1}^{T} \check{G}_T(\pi(x_t)) \mathbf{1}_{v(x_t) - \pi(x_t) > u}. \tag{27}$$

Furthermore, to facilitate further analysis, we introduce an auxiliary distribution P_T^π on $[0, 1]$, whose CDF is defined as

$$\begin{aligned} F_{P_T^\pi}(u) :&= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{m \sim P} \mathbf{1}_{r_\pi(x_t, m) \leq u} \\ &= 1 - \frac{1}{T} \sum_{t=1}^{T} G(\pi(x_t)) \mathbf{1}_{v(x_t) - \pi(x_t) > u}. \end{aligned} \tag{28}$$

The following lemma reformulates the estimated dual objectives in terms of $\hat{P}_T^\pi$ and decomposes the sub-optimality gap into two terms.

Lemma 4 (Uninformative-Case Sub-optimality Decomposition)**.** *Suppose that P^π, $\hat{P}_T^\pi$ and P_T^π are defined in* (17)*,* (27) *and* (28)*, respectively. Then, it holds that*

$$\begin{aligned} \varphi(\alpha, \pi) &= -\alpha \log \mathbb{E}_{r \sim P^\pi} \exp(-r/\alpha) - \alpha\rho, \\ \check{\varphi}_T(\alpha, \pi) &= -\alpha \log \mathbb{E}_{r \sim \hat{P}_T^\pi} \exp(-r/\alpha) - \alpha\rho. \end{aligned} \tag{29}$$

Moreover, for the learning policy $\hat{\pi}_T \in \Pi$ obtained by Algorithm 2, under Assumption 2, we have

$$\begin{aligned} R_{\mathrm{DRO}}(\hat{\pi}_T) &\leq \frac{4}{\kappa} \sup_{\pi \in \Pi} \sup_{u \in [0,1]} |F_{\hat{P}_T^\pi}(u) - F_{P^\pi}(u)| \\ &\leq \frac{4}{\kappa} \sup_{\pi \in \Pi} \sup_{u \in [0,1]} |F_{\hat{P}_T^\pi}(u) - F_{P_T^\pi}(u)| \qquad (30) \\ &\quad + \frac{4}{\kappa} \sup_{\pi \in \Pi} \sup_{u \in [0,1]} |F_{P_T^\pi}(u) - F_{P^\pi}(u)|. \qquad (31) \end{aligned}$$

Lemma 4 makes clear why the uninformative case is harder than the informative one. The first term in (30) measures the estimation error induced by the empirical ratio estimator $\check{G}_T$, and the second term in (31) captures the approximation error of P_T^π to P^π.

Theorem 2 (Uninformative-Case Sub-optimality Bound)**.** *Suppose Assumptions 1 and 2 hold, and the policy class* Π *has finite pseudo-dimension* $d = \mathrm{Pdim}(\Pi)$*. Then, for a universal constant* $C > 0$*,*

$$R_{\mathrm{DRO}}(\hat{\pi}_T) \le \frac{4}{\kappa}\left[\frac{2}{\varepsilon_0\sqrt{T}} + 2C\sqrt{\frac{d+1}{T}} + 3\delta\right], \tag{32}$$

with probability at least $1 - \exp(-T\delta^2/2) - \exp(-T\varepsilon_0^2\delta^2/2)$*, for any* $\delta \ge 0$*.*

Proof Sketch. Using the exponential moments in Lemma 4, we separately control the estimation error of G and the empirical context error. For the first term, Assumption 1 and the denominator protection in (24) give

$$|\check{G}_T(u) - G(u)| \le \frac{1}{\varepsilon_0}\left(|L_T(u) - L(u)| + |H_T^{\pi_0}(u) - H^{\pi_0}(u)|\right).$$

The two empirical CDF errors are $O(T^{-1/2})$. For the second term, the reward class has pseudo-dimension at most $5d$. A covering-number bound and Dudley's entropy integral give $O(\sqrt{(d+1)/T})$ control. Combining these bounds with the dual stability factor $4/\kappa$ proves the result. The full proof is given in Appendix C.2.

5 Numerical Studies

We illustrate the benefit of robust learning, its convergence, and the effect of policy-class complexity. Let $x = (s, z)$, where s is uniform on $s_j = (j+1/2)/16$, $j = 0, \ldots, 15$, and $z \sim U[0,1]$ is independent. Set $v(x) = s$ and let the independent competing bid have CDF $G(u) = 0.5u + 0.5u^4$ on $[0,1]$. The logging policy $\pi_0(s,z) = 0.8s\,\mathbf{1}\{z > 0.1\}$ bids zero with probability 0.1 and otherwise bids $0.8s$. Partitioning $[0,1]$ into K equal-width intervals $I_1, \ldots, I_K$, we use

$$\Pi_K = \{\pi_\theta(s,z) = \theta_\ell s \text{ for } s \in I_\ell : \theta \in [0.1, 0.9]^K\}, \qquad K \in \{1, 2, 4, 8, 16\}.$$

These nested classes have $\mathrm{Pdim}(\Pi_K) = K$. Since $H^{\pi_0}(0) = 0.1$ and $\mathbb{P}_J(r_\pi = 0) \ge 1 - G(0.9) = 0.22195$, Assumptions 1 and 2 hold with $\varepsilon_0 = 0.1$ and $\kappa = 0.2$. We report means over 100 independent runs in Experiments 1 and 3 and 200 in Experiment 2, with error bars indicating approximate 95% confidence intervals for the mean. All methods use the same logs within each run. Learned policies are evaluated under the true model.

Experiment 1: robust performance versus ambiguity radius. We fix $T = 1000$ and $K = 4$ and vary $\rho \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1\}$. For each log, the nominal learner is trained once at $\rho = 0$, whereas the uninformative DRO learner is trained separately at each radius. Both are evaluated by the true robust value at the same radius, denoted by V_ρ. Figure 1 shows that the curves coincide at $\rho = 0$, while DRO has a higher mean robust value at every tested positive radius. At $\rho = 0.1$, the values are 0.009440 for DRO and 0.004294 for nominal learning. This protection has a nominal cost: their expected rewards under J are 0.035275 and 0.044138, respectively.

Experiment 2: learning from censored feedback. Fixing $K = 4$ and $\rho = 0.25$, we compare the informative and uninformative learners at $T \in \{100, 250, 500, 1000, 2000, 5000\}$. Figure 2 reports $V_K^* - V(\widehat{\pi})$, where $V_K^* = \sup_{\pi \in \Pi_K} V(\pi)$ for the stated radius. Both mean sub-optimality gaps decrease as T increases. The informative learner exhibits smaller mean gaps at the smaller sample sizes, while the two methods achieve comparable performance with more data.

Experiment 3: policy-class complexity. For the uninformative learner with $\rho = 0.1$, we vary K and compare $T \in \{100, 1000, 10000\}$. All classes are evaluated against the same reference V_{16}^*, so the plotted gap $V_{16}^* - V(\widehat{\pi}_K)$ reflects both the expressiveness of the class and finite-sample learning error. Figure 3 illustrates the trade-off between policy expressiveness and finite-sample learning error. Simpler classes achieve smaller gaps when data are limited, whereas richer classes become more effective as the sample size increases.

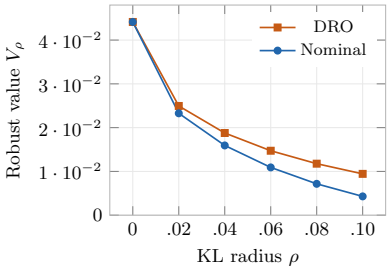

Figure 1: Robust value versus ρ at $T = 1000$, $K = 4$.

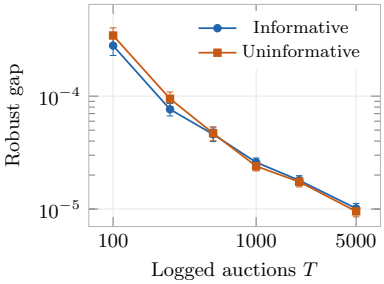

Figure 2: Robust gap versus T at $K = 4$, $\rho = 0.25$.

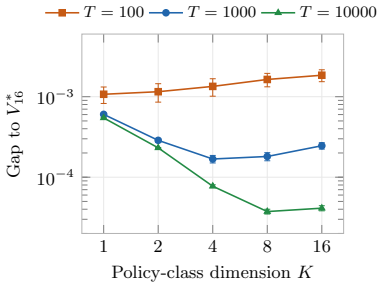

Figure 3: Policy-class comparison at $\rho = 0.1$.

6 Conclusion

We have proposed KL-distributionally robust bidding policies for repeated first-price auctions under censored transaction-price feedback. By combining a strong-duality reduction with bounded ratio estimation, our algorithms achieve, under the stated coverage and zero-reward-mass assumptions, $O(T^{-1/2})$ DRO sub-optimality gap when the feature distribution Q is known, and $O(\sqrt{(d+1)/T})$ when Q is unknown and the policy class has finite pseudo-dimension d. Numerical experiments illustrate the benefit and nominal cost of robustness, learning from censored feedback, and the effect of policy-class complexity. An important extension is to allow the highest competing bid to depend on the context. In that case, the nominal model no longer factorizes as $J = Q \times P$, and the relevant object becomes a conditional CDF $G(\cdot \mid x)$ rather than a single marginal bid CDF, so the ratio-based estimation and analysis developed here do not directly extend.

A Proofs in Section 2

A.1 Proof of Lemma 1

Proof. We first prove that

$$V(\pi) = \inf_{D_{KL}(J' \| J) \le \rho} \mathbb{E}_{(x,m)\sim J'}\ r_\pi(x,m) = \sup_{\alpha \ge 0} \left\{ -\alpha \log \mathbb{E}_{(x,m)\sim J}[\exp(-r_\pi(x,m)/\alpha)] - \alpha\rho \right\},$$

which follows [17, Theorems 1–2].

Theorem 3 ([17], Theorems 1–2)**.** *Suppose for every $x \in X$, set $S = \{s \in \mathbb{R} : s > 0, M_H(s) = \mathbb{E}_{P_0}[e^{sH(x,\xi)}] < +\infty\}$ is a nonempty set. Then, for a positive KL radius, problem* (33)

$$\max_{P \in \mathcal{P}} \quad \mathbb{E}_P\left[H(x,\xi)\right], \tag{33}$$

is equivalent to the following one-layer optimization problem

$$\min_{\alpha\geq 0} h_x(\alpha) := \alpha \log \mathbb{E}_{P_0}\left[e^{H(x,\xi)/\alpha}\right] + \alpha\eta, \tag{34}$$

where $\mathcal{P} = \{P \in \mathcal{D} : D_{KL}(P\|P_0) \leq \eta\}$, ξ *denotes the vector of parameters, and* x *is the vector of decision variables.*

Since $\mathbb{E}_{(x,m)\sim J}[e^{-sr_\pi(x,m)}] \leq 1 < +\infty$ for $0 < s < +\infty$, the set S is nonempty. Apply the Theorem 3 to our setting, where $H(x,\xi) = -r_\pi(x,m) = -(v(x)-\pi(x))\mathbf{1}_{\pi(x)\geq m}$, $P = J'$, $P_0 = J$, we have

$$\begin{aligned}
\inf_{D_{KL}(J'\|J)\leq\rho} \mathbb{E}_{(x,m)\sim J'}\ r_\pi(x,m) &= -\sup_{D_{KL}(J'\|J)\leq\rho} \mathbb{E}_{(x,m)\sim J'}\ -r_\pi(x,m) \\
&= -\inf_{\alpha\geq 0}\left\{\alpha\log\mathbb{E}_{(x,m)\sim J}[\exp(-r_\pi(x,m)/\alpha)] + \alpha\rho\right\} \\
&= \sup_{\alpha\geq 0}\left\{-\alpha\log\mathbb{E}_{(x,m)\sim J}[\exp(-r_\pi(x,m)/\alpha)] - \alpha\rho\right\}.
\end{aligned}$$

The second equation holds because

$$\begin{aligned}
\mathbb{E}_{(x,m)\sim J}\left[\exp\left(-r_\pi(x,m)/\alpha\right)\right] &= \mathbb{E}_{x\sim Q}\mathbb{E}_{m\sim P}\left[\exp\left(-\frac{(v(x)-\pi(x))\mathbf{1}_{\pi(x)\geq m}}{\alpha}\right)\right] \\
&= \mathbb{E}_{x\sim Q}\mathbb{E}_{m\sim P}\left[\exp\left(-\frac{(v(x)-\pi(x))\mathbf{1}_{\pi(x)\geq m}}{\alpha}\right)\left(\mathbf{1}_{\pi(x)\geq m}+\mathbf{1}_{\pi(x)<m}\right)\right] \\
&= \mathbb{E}_{x\sim Q}\mathbb{E}_{m\sim P}\left[\exp\left(-\frac{v(x)-\pi(x)}{\alpha}\right)\mathbf{1}_{\pi(x)\geq m}+\mathbf{1}_{\pi(x)<m}\right] \\
&= \mathbb{E}_{x\sim Q}\left[\exp\left(-\frac{v(x)-\pi(x)}{\alpha}\right)G(\pi(x)) + 1 - G(\pi(x))\right] \\
&= \mathbb{E}_{x\sim Q}\left[1-G(\pi(x))\left(1-\exp\left(-\frac{v(x)-\pi(x)}{\alpha}\right)\right)\right],
\end{aligned}$$

where the first equation follows from the definition of $r_\pi(x,m)$, the second equation expands 1 as $\mathbf{1}_{\pi(x)\geq m} + \mathbf{1}_{\pi(x)<m}$, the third equation results from simple computations: $(\mathbf{1}_{\pi(x)\geq m})^2 = \mathbf{1}_{\pi(x)\geq m}$, $\mathbf{1}_{\pi(x)\geq m}\mathbf{1}_{\pi(x)<m} = 0$, and the fourth equation follows the definition of CDF, $G(\pi(x)) = \mathbb{E}_{m\sim P}\mathbf{1}_{m\leq\pi(x)}$. For $\rho = 0$, the ambiguity set contains only J and the dual supremum equals $\mathbb{E}_J r_\pi$ by taking $\alpha\to\infty$. The $\alpha\downarrow 0$ limit is the essential infimum of the bounded reward. □

B Proofs in Section 3

B.1 Proof of Lemma 2.

Proof. For every u, independence of the logging bid b and competing bid m gives

$$\mathbb{P}(\max\{b,m\}\leq u) = \mathbb{P}(b\leq u, m\leq u) = H^{\pi_0}(u)G(u).$$

Taking expectations of the empirical indicators yields $\mathbb{E}L_T(u) = L(u) = H^{\pi_0}(u)G(u)$. This argument also applies to CDFs with atoms and randomized logging. □

B.2 Proof of Lemma 3

We first prove the reformulation of two dual functions and then apply them to establish an upper bound of the regret of the learning policy.

Reformulation of Dual Functions. To prove equation (19), note that

$$\mathbb{P}_{r\sim P^\pi}(r \le u) = F_{P^\pi}(u) = \mathbb{E}_{(x,m)\sim J}\mathbf{1}_{r_\pi(x,m)\le u} = \mathbb{P}_{(x,m)\sim J}(r_\pi(x,m) \le u),$$

which implies that

$$\mathbb{E}_{r\sim P^\pi} \exp\left(-r/\alpha\right) = \mathbb{E}_{(x,m)\sim J}\left[\exp\left(-r_\pi(x,m)/\alpha\right)\right].$$

Then, using Lemma 1, we obtain that

$$\begin{aligned}\varphi(\alpha,\pi) &= -\alpha \log \mathbb{E}_{(x,m)\sim J}\left[\exp\left(-r_\pi(x,m)/\alpha\right)\right] - \alpha\rho \\ &= -\alpha \log \mathbb{E}_{r\sim P^\pi} \exp\left(-r/\alpha\right) - \alpha\rho.\end{aligned}$$

For equation (20), the direct reward mixture gives

$$\mathbb{E}_{r\sim \hat{P}^\pi} e^{-r/\alpha} = \mathbb{E}_Q[1 - \hat{G}_T(\pi(x)) + \hat{G}_T(\pi(x))e^{-(v(x)-\pi(x))/\alpha}] = 1 - \hat{\phi}_T(\alpha,\pi).$$

Substituting this identity into (16) proves the claim.

Upper Bound of the Regret. By the policy comparison (12),

$$R_{\mathrm{DRO}}(\hat{\pi}_T) \le 2 \sup_{\pi\in\Pi}\sup_{\alpha\ge 0} |\hat{\varphi}_T(\alpha,\pi) - \varphi(\alpha,\pi)|. \tag{35}$$

It remains to control the exponential moments uniformly in α. For two reward CDFs $F, \widetilde{F}$ on $[0,1]$, put $\Delta = \|F - \widetilde{F}\|_\infty$ and $M_F(\alpha) = \mathbb{E}_F e^{-r/\alpha}$. Tonelli's theorem gives

$$M_F(\alpha) = e^{-1/\alpha} + \frac{1}{\alpha}\int_0^1 e^{-u/\alpha} F(u)\, du,$$

so $|M_F - M_{\widetilde{F}}| \le (1 - e^{-1/\alpha})\Delta$. Suppose $F(0) \ge \kappa$. If $\Delta \le \kappa/2$, then $\widetilde{F}(0) \ge \kappa/2$ and both exponential moments are at least $\kappa/2$. Thus

$$\alpha|\log M_F - \log M_{\widetilde{F}}| \le \frac{2\alpha(1 - e^{-1/\alpha})}{\kappa}\Delta \le \frac{2}{\kappa}\Delta.$$

If $\Delta > \kappa/2$, both entropic values $-\alpha \log M$ lie in $[0,1]$, so their difference is at most $1 < 2\Delta/\kappa$. The same bound holds at $\alpha = 0$ by taking limits. This proof allows both atoms and continuous reward components. Applying it with $F = F_{P^\pi}$ and $\widetilde{F} = F_{\hat{P}^\pi}$, and using Assumption 2 in (35), yields

$$R_{\mathrm{DRO}}(\hat{\pi}_T) \le \frac{4}{\kappa}\sup_{\pi\in\Pi} \|F_{\hat{P}^\pi} - F_{P^\pi}\|_\infty. \tag{36}$$

The coefficient $4/\kappa$ is independent of the sample and the dual optimizer.

B.3 Proof of Theorem 1

To prove Theorem 1, we analyze the bounds on the distance between two cumulative distribution functions (CDFs). Specifically, we are interested in deriving an upper bound for term $\sup_{\pi\in\Pi}\sup_{u\in[0,1]}|F_{\hat{P}^\pi}(u)-F_{P^\pi}(u)|$. Expanding it, we have

$$\begin{aligned}\sup_{\pi\in\Pi}\sup_{u\in[0,1]}\left|F_{\hat{P}^\pi}(u)-F_{P^\pi}(u)\right| &= \sup_{\pi\in\Pi}\sup_{u\in[0,1]}\left|\mathbb{E}_{x\sim Q}\left[(\hat{G}_T(\pi(x))-G(\pi(x)))\mathbf{1}_{v(x)-\pi(x)>u}\right]\right| \\ &\le \sup_{\pi\in\Pi}\mathbb{E}_{x\sim Q}\left|\hat{G}_T(\pi(x))-G(\pi(x))\right| \\ &\le \sup_{u\in[0,1]}\left|\hat{G}_T(u)-G(u)\right|\end{aligned}$$

The first inequality follows from $|\mathbb{E}[Z\mathbf{1}_A]|\le\mathbb{E}|Z|$. The second inequality holds because $\pi(x)\in[0,1]$ for $x\sim Q$ and $\pi\in\Pi$, which implies that

$$\sup_{\pi\in\Pi}\mathbb{E}_{x\sim Q}\left|\left[\hat{G}_T(\pi(x))-G(\pi(x))\right]\right|\le\sup_{\pi\in\Pi}\sup_{x}\left|\hat{G}_T(\pi(x))-G(\pi(x))\right|\le\sup_{u\in[0,1]}\left|\hat{G}_T(u)-G(u)\right|.$$

Additionally, clipping cannot increase the error from $G(u)\in[0,1]$. By the definition of $\hat{G}_T(u)$ and Lemma 2, we have that

$$\begin{aligned}\sup_{u\in[0,1]}|\hat{G}_T(u)-G(u)| &\le \sup_{u\in[0,1]}\left|\frac{1}{H^{\pi_0}(u)}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\max\{m_t,b_t\}\le u}-\mathbb{E}_{(x,m)\sim J}\left[\mathbf{1}_{\max\{m,b\}\le u}\right]\right]\right| \\ &= \sup_{u\in[0,1]}\left|\frac{1}{H^{\pi_0}(u)}\left[L_T(u)-L(u)\right]\right|.\end{aligned}$$

If Assumption 1 holds, i.e. $H^{\pi_0}(0)\ge\varepsilon_0$ is satisfied, then it follows that

$$\sup_{u\in[0,1]}|\hat{G}_T(u)-G(u)|\le\frac{1}{\varepsilon_0}\sup_{u\in[0,1]}|L_T(u)-L(u)|,$$

which implies that

$$\sup_{\pi\in\Pi}\sup_{u\in[0,1]}\left|F_{\hat{P}^\pi}(u)-F_{P^\pi}(u)\right|\le\frac{1}{\varepsilon_0}\sup_{u\in[0,1]}|L_T(u)-L(u)|. \tag{37}$$

Therefore, we consider the function class $\mathscr{F}_{\pi_0,1}=\left\{f_u(\boldsymbol{x})=\mathbf{1}_{\max\{m,b\}\le u};u\in[0,1]\right\}$, and analyze its Rademacher complexity to derive an upper bound for the term $\sup_{u\in[0,1]}|L_T(u)-L(u)|$. In the following, we first introduce key definitions and lemmas related to Rademacher complexity and then apply them to derive the desired bounds for our specific case.

B.3.1 Auxiliary Lemmas

A key quantity in the subsequent analysis is the Rademacher complexity of the function class $\mathscr{F}$. Let $\boldsymbol{x}_1^n:=\{x_1,\ldots,x_n\}$ be a sample of points and consider a function class $\mathscr{F}$. Then, the *empirical Rademacher complexity* of $\mathscr{F}(\boldsymbol{x}_1^n)$ is given by

$$\mathscr{R}\left(\mathscr{F}(\boldsymbol{x}_1^n)/n\right):=\mathbb{E}_\varepsilon\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right], \tag{38}$$

where $\mathscr{F}(\boldsymbol{x}_1^n) := \{(f(x_1), \ldots, f(x_n)) \mid f \in \mathscr{F}\}$, and ε_i are i.i.d. Rademacher variables (taking values 1 and -1 with equal probability $\frac{1}{2}$ each).

Note that the empirical Rademacher complexity $\mathscr{R}(\mathscr{F}(\boldsymbol{x}_1^n)/n)$ is a random variable. Taking its expectation yields the *Rademacher complexity of the function class* $\mathscr{F}$

$$\mathscr{R}_n(\mathscr{F}) := \mathbb{E}_x[\mathscr{R}(\mathscr{F}(\boldsymbol{x}_1^n)/n)] = \mathbb{E}_{x,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right]. \tag{39}$$

We now consider the uniform laws of large numbers. Let $\mathscr{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution $\mathbb{P}$ over $\mathcal{X}$. Consider the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right| \tag{40}$$

which measures the absolute deviation between the sample avergae $\frac{1}{n}\sum_{i=1}^n f(X_i)$ and the population average $\mathbb{E}[f(X)]$, uniformly over the class $\mathscr{F}$. We now introduce a Theorem that bounds the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ by a multiple of the Rademacher complexity. This result applies to a b-uniformly bounded function class $\mathscr{F}$ satisfying $\|f\|_\infty \le b$ for all $f \in \mathscr{F}$.

Theorem 4 ([27], Theorem 4.10)**.** *For any b-uniformly bounded class of function $\mathscr{F}$, any positive integer $n \ge 1$ and any scalar $\delta \ge 0$, we have*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \le 2\mathscr{R}_n(\mathscr{F}) + \delta \tag{41}$$

with $\mathbb{P}$-probability at least $1 - \exp(-\frac{n\delta^2}{2b^2})$.

The next step is to obtain upper bounds on the Rademacher complexity $\mathscr{R}_n(\mathscr{F})$. Several techniques are available for this purpose, depending on the properties of the function class under consideration. In the following, we focus on a method that applies to function classes with finite Vapnik–Chervonenkis (VC) dimension.

Definition 1. *(Shattering and VC dimension) Given a class $\mathscr{F}$ of binary-valued functions, we say that the set $x_1^n = (x_1, \ldots, x_n)$ is shattered by $\mathscr{F}$ if card$(\mathscr{F}(x_1^n)) = 2^n$. The VC dimension $\nu(\mathscr{F})$ is the largest integer n for which there is some collection $x_1^n = (x_1, \ldots, x_n)$ of n points that is shattered by $\mathscr{F}$.*

When the quantity $\nu(\mathscr{F})$ is finite, the the function class $\mathscr{F}$ is said to be a VC class.

Definition 2. *(Dudley's entropy integral)The Dudley's entropy integral is defined for a metric space (T, d) equipped with a probability measure μ. Given a set T and an ϵ-covering, the entropy of T is the logarithm of the minimum number of balls of radius ϵ required to cover T. Dudley's entropy integral is then given by the formula:*

$$\int_0^\infty \sqrt{\log N(\epsilon; T, d)}\, d\epsilon$$

where $N(\epsilon; T, d)$ is the covering number, i.e. the minimum number of balls of radius ϵ with respect to the metric d that cover the space T.

The significance of Dudley's entropy integral lies in providing a metric entropy measure to assess the complexity of a space with respect to a given probability distribution.

Lemma 5 ([27], Example 5.24)**.** *Let $\mathscr{F}$ be a b-uniformly bounded class of binary-valued functions with finite VC dimension $\nu \geq 1$. Let $\mathscr{F}^{\pm} = \mathscr{F} \cup (-\mathscr{F}) \cup \{0\}$. By Dudley's entropy integral for this signed class, we have*

$$\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right] \leq \frac{24}{\sqrt{n}}\int_0^{2b}\sqrt{\log N(\epsilon;\mathscr{F}^{\pm},\|\cdot\|_{\mathbb{P}_n})}\,\mathrm{d}\epsilon \tag{42}$$

where we have used the fact that $\sup_{f,g\in\mathscr{F}^{\pm}}\|f-g\|_{\mathbb{P}_n} \leq 2b$. Now by known results on VC classes and metric entropy, the bound $N(\epsilon;\mathscr{F}^{\pm}) \leq 2N(\epsilon;\mathscr{F})+1$ can be absorbed into a universal constant C', so that

$$N(\epsilon;\mathscr{F}^{\pm},\|\cdot\|_{\mathbb{P}_n}) \leq C'\nu(16e)^{\nu}\left(\frac{2b}{\epsilon}\right)^{2\nu}, \qquad 0<\epsilon\leq 2b, \tag{43}$$

Substituting the metric entropy bound (43) *into the entropy integral* (42)*, we find that there are universal constants c_0 and c_0', depending on b but not on (ν, n), such that*

$$\begin{aligned}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right] &\leq c_0\sqrt{\frac{\nu}{n}}\left[1+\int_0^{2b}\sqrt{\log(2b/t)}\,dt\right]\\ &= c_0'\sqrt{\frac{\nu}{n}},\end{aligned}$$

since the integral is finite.

B.3.2 Application to Regret Analysis

We now apply Theorem 4 in our settings. To bound term $\sup_{u\in[0,1]}|L_T(u)-L(u)|$, we consider the function class

$$\mathscr{F}_{\pi_0,1} = \left\{f_u(\boldsymbol{x}) = \mathbf{1}_{\max\{m,b\}\leq u}; u\in[0,1]\right\},$$

where $b=\pi_0(x)$ and $\boldsymbol{x}=(x,b,m\mathbf{1}_{b<m})$. Note that given $\boldsymbol{x}_t=(x_t,b_t,m_t\mathbf{1}_{b_t<m_t})$, we have $f_u(\boldsymbol{x}_t) = \mathbf{1}_{\max\{m_t,b_t\}\leq u} \leq 1$. Suppose $\{\boldsymbol{x}_t=(x_t,b_t,m_t\mathbf{1}_{b_t<m_t})\}_{t=1}^T$ are i.i.d. samples from distribution $\mathbb{P}$. Then, by definition

$$\begin{aligned}\|\mathbb{P}_T-\mathbb{P}\|_{\mathscr{F}_{\pi_0,1}} &:= \sup_{u\in[0,1]}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\max\{m_t,b_t\}\leq u} - \mathbb{E}_{(x,m)\sim J}\left[\mathbf{1}_{\max\{m,b\}\leq u}\right]\right|\\ &= \sup_{u\in[0,1]}\left|L_T(u)-L(u)\right|.\end{aligned}$$

In order to apply the lemma 5 in our setting, we first show that the function class $\mathscr{F}_{\pi_0,1}$ is a VC class.

Lemma 6. $\nu(\mathscr{F}_{\pi_0,1}) = 1$.

Proof. For function class $\mathscr{F}_{\pi_0,1} = \{f_u(\boldsymbol{x}) = \mathbf{1}_{\max\{m,b\}\le u}; u \in [0,1]\}$, we note that, for a point $\boldsymbol{x}_1$ with $0 < \max\{m_1, b_1\} \le 1$,

$$\mathscr{F}_{\pi_0,1}(\boldsymbol{x}_1) = \{(f_u(\boldsymbol{x}_1)|f_u \in \mathscr{F}_{\pi_0,1}\} = \{(\mathbf{1}_{\max\{m_1,b_1\}\le u})|u \in [0,1]\} = \{(0),(1)\},$$

which implies that $\text{card}(\mathscr{F}_{\pi_0,1}(\boldsymbol{x}_1)) = 2$. But given two distinct points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ satisfying $\max\{m_1, b_1\} < \max\{m_2, b_2\}$,

$$\begin{aligned}\mathscr{F}_{\pi_0,1}(\boldsymbol{x}_1^2) &= \{(f_u(\boldsymbol{x}_1), f_u(\boldsymbol{x}_2))|f_u \in \mathscr{F}_{\pi_0,1}\} \\ &= \{(\mathbf{1}_{\max\{m_1,b_1\}\le u}, \mathbf{1}_{\max\{m_2,b_2\}\le u})|u \in [0,1]\} \subseteq \{(0,0),(1,0),(1,1)\},\end{aligned}$$

which implies that $\text{card}(\mathscr{F}_{\pi_0,1}(\boldsymbol{x}_1^2)) \le 3 < 2^2$. If the two maxima agree, there are at most two patterns. Therefore, we conclude that $\nu(\mathscr{F}_{\pi_0,1}) = 1$. □

It is evident that $\|f_u\|_\infty \le 1$ for all $f_u \in \mathscr{F}_{\pi_0,1}$, which indicates that $\mathscr{F}_{\pi_0,1}$ is uniformly bounded by 1. Then, applying Lemma 5, we obtain that, for some universal constant c_0,

$$\mathscr{R}_T(\mathscr{F}_{\pi_0,1}) = \mathbb{E}_{\boldsymbol{x},\varepsilon}\left[\sup_{f_u\in\mathscr{F}_{\pi_0,1}}\left|\frac{1}{T}\sum_{i=1}^{T}\varepsilon_i f_u(\boldsymbol{x}_i)\right|\right] \le \mathbb{E}_{\boldsymbol{x}}\frac{c_0}{\sqrt{T}} = \frac{c_0}{\sqrt{T}},$$

To make the CDF bound explicit, the Dvoretzky–Kiefer–Wolfowitz inequality [28] gives, for any $\delta \ge 0$,

$$\mathbb{P}\left(\|L_T - L\|_\infty > \frac{1}{\sqrt{T}} + \delta\right) \le 2e^{-2(1+\sqrt{T}\delta)^2} \le e^{-T\delta^2/2}.$$

Thus, with probability at least $1 - \exp(-T\delta^2/2)$,

$$\|\mathbb{P}_T - \mathbb{P}\|_{\mathscr{F}_{\pi_0,1}} = \sup_{u\in[0,1]}|L_T(u) - L(u)| \le \frac{1}{\sqrt{T}} + \delta.$$

Therefore, by combining this result with Lemma 3 and inequality (37), and setting δ as $\varepsilon_0\delta$, the following holds with probability at least $1 - \exp(-\frac{T\varepsilon_0^2\delta^2}{2})$,

$$R_{DRO}(\hat{\pi}_T) \le \frac{4}{\kappa}\sup_{\pi\in\Pi}\sup_{u\in[0,1]}\left|F_{\hat{P}^\pi}(u) - F_{P^\pi}(u)\right| \le \frac{4}{\kappa}\left[\frac{1}{\varepsilon_0\sqrt{T}} + \delta\right].$$

C Proofs in Section 4

C.1 Proof of Lemma 4

We first prove the reformulation of dual function (29). By the direct definition of $\hat{P}_T^\pi$,

$$\mathbb{E}_{r\sim\hat{P}_T^\pi}e^{-r/\alpha} = \frac{1}{T}\sum_{t=1}^{T}[1 - \check{G}_T(\pi(x_t)) + \check{G}_T(\pi(x_t))e^{-(v(x_t)-\pi(x_t))/\alpha}] = 1 - \check{\phi}_T(\alpha,\pi).$$

It follows that $\check{\varphi}_T(\alpha,\pi) = -\alpha\log\mathbb{E}_{r\sim\hat{P}_T^\pi}e^{-r/\alpha} - \alpha\rho$.

Then we apply the dual functions (19) and (29) to establish an upper bound on the regret of the learning policy. The proof is analogous to the proof of Lemma 3. By following the methods

applied in Appendix B.2 and substituting the distribution $\hat{P}^\pi$ with $\hat{P}_T^\pi$, we obtain the following inequalities under Assumption 2,

$$\begin{aligned} R_{\mathrm{DRO}}(\hat{\pi}_T) &\le \frac{4}{\kappa} \sup_{\pi\in\Pi} \sup_{u\in[0,1]} \left| F_{\hat{P}_T^\pi}(u) - F_{P^\pi}(u) \right| \\ &\le \frac{4}{\kappa} \left(\sup_{\pi\in\Pi} \sup_{u\in[0,1]} \left| F_{\hat{P}_T^\pi}(u) - F_{P_T^\pi}(u) \right| + \sup_{\pi\in\Pi} \sup_{u\in[0,1]} \left| F_{P_T^\pi}(u) - F_{P^\pi}(u) \right| \right). \end{aligned}$$

where the last inequality follows from the triangle inequality.

C.2 Proof of Theorem 2

The pseudo-dimension d is the VC dimension of $\{(x,s) \mapsto \mathbf{1}_{\pi(x)\ge s} : \pi \in \Pi\}$. We first derive a bound for a general policy class, and then bound its complexity using $d < \infty$. Define the function class

$$\begin{aligned} \mathscr{H}_\Pi := &\left\{ (x,m) \mapsto \alpha \left(1 - e^{-r_\pi(x,m)/\alpha} \right) : \pi \in \Pi,\ \alpha > 0 \right\} \\ &\cup \left\{ (x,m) \mapsto \mathbf{1}_{r_\pi(x,m)>0} : \pi \in \Pi \right\}. \end{aligned} \tag{44}$$

This class is 1-uniformly bounded. Write $\mathscr{R}_T(\mathscr{H}_\Pi)$ for its expected absolute Rademacher complexity under T i.i.d. samples from $J = Q \times P$, as defined in Appendix B.3.1.

We retain the two-term decomposition and compare exponential moments directly. Define

$$\mathscr{F}_\Pi := \{ x \mapsto \mathbb{E}_{m\sim P} h(x,m) : h \in \mathscr{H}_\Pi \}, \qquad \tau_T := \sup_{u\in[0,1]} |\check{G}_T(u) - G(u)| + \|Q_T - Q\|_{\mathscr{F}_\Pi}.$$

By (27) and (17),

$$\begin{aligned} |F_{\hat{P}_T^\pi}(0) - F_{P^\pi}(0)| &= \left| \frac{1}{T} \sum_{t=1}^T \check{G}_T(\pi(x_t)) \mathbf{1}_{v(x_t)-\pi(x_t)>0} \right. \\ &\qquad \left. - \mathbb{E}_{x\sim Q} G(\pi(x)) \mathbf{1}_{v(x)-\pi(x)>0} \right| \\ &\le \sup_{u\in[0,1]} |\check{G}_T(u) - G(u)| + \|Q_T - Q\|_{\mathscr{F}_\Pi} = \tau_T. \end{aligned}$$

Similarly, since $0 \le \alpha(1 - e^{-s/\alpha}) \le s \le 1$ for $s \in [0,1]$,

$$\begin{aligned} &\alpha \left| \mathbb{E}_{r\sim \hat{P}_T^\pi} e^{-r/\alpha} - \mathbb{E}_{r\sim P^\pi} e^{-r/\alpha} \right| \\ &= \left| \frac{1}{T} \sum_{t=1}^T \check{G}_T(\pi(x_t)) \alpha \left(1 - e^{-(v(x_t)-\pi(x_t))/\alpha} \right) - \mathbb{E}_{x\sim Q} G(\pi(x)) \alpha \left(1 - e^{-(v(x)-\pi(x))/\alpha} \right) \right| \\ &\le \sup_{u\in[0,1]} |\check{G}_T(u) - G(u)| + \|Q_T - Q\|_{\mathscr{F}_\Pi} = \tau_T. \end{aligned}$$

If $\tau_T \le \kappa/2$, Assumption 2 implies $F_{P^\pi}(0) \ge \kappa$ and $F_{\hat{P}_T^\pi}(0) \ge \kappa/2$ for all $\pi \in \Pi$. Hence, for every $\alpha > 0$,

$$\begin{aligned}|\check{\varphi}_T(\alpha,\pi) - \varphi(\alpha,\pi)| &\le \frac{2\alpha}{\kappa}\left|\mathbb{E}_{r\sim\hat{P}_T^\pi} e^{-r/\alpha} - \mathbb{E}_{r\sim P^\pi} e^{-r/\alpha}\right| \\ &\le \frac{2}{\kappa}\tau_T.\end{aligned}$$

The same bound holds at $\alpha = 0$ by continuity. Applying (12) gives $R_{\mathrm{DRO}}(\hat{\pi}_T) \le 4\tau_T/\kappa$. If $\tau_T > \kappa/2$, then $R_{\mathrm{DRO}}(\hat{\pi}_T) \le 1 < 2\tau_T/\kappa$. Thus, in both cases,

$$R_{\mathrm{DRO}}(\hat{\pi}_T) \le \frac{4}{\kappa}\underbrace{\sup_{u\in[0,1]}|\check{G}_T(u) - G(u)|}_{\text{(a)}} + \frac{4}{\kappa}\underbrace{\|Q_T - Q\|_{\mathscr{F}_\Pi}}_{\text{(b)}}. \tag{45}$$

Upper bound on Term (a). Write $h_T(u) = \max\{H_T^{\pi_0}(u), \varepsilon_0\}$. Assumption 1 implies $|h_T - H^{\pi_0}| \le |H_T^{\pi_0} - H^{\pi_0}|$. Using $L = H^{\pi_0}G$ and $0 \le G \le 1$,

$$\left|\check{G}_T - G\right| = \left|\frac{L_T - L + G(H^{\pi_0} - h_T)}{h_T}\right| \le \frac{|L_T - L| + |H_T^{\pi_0} - H^{\pi_0}|}{\varepsilon_0}.$$

The triangle inequality for the reward CDFs therefore gives

$$\begin{aligned}\sup_{\pi\in\Pi}\|F_{\hat{P}_T^\pi} - F_{P_T^\pi}\|_\infty &\le \sup_{u\in[0,1]}|\check{G}_T(u) - G(u)| \\ &\le \frac{\|L_T - L\|_\infty + \|H_T^{\pi_0} - H^{\pi_0}\|_\infty}{\varepsilon_0}.\end{aligned} \tag{46}$$

Then, we introduce a lemma that provides the bounds for $|L_T - L|$ and $|H^{\pi_0} - H_T^{\pi_0}|$.

Lemma 7. *For any $\delta \ge 0$, the following inequalities hold jointly with probability at least $1 - \exp\left(-T\delta^2/2\right)$:*

$$\sup_{u\in[0,1]}|L_T(u) - L(u)| \le \frac{1}{\sqrt{T}} + \delta, \tag{47}$$

$$\sup_{u\in[0,1]}\left|H_T^{\pi_0}(u) - H^{\pi_0}(u)\right| \le \frac{1}{\sqrt{T}} + \delta. \tag{48}$$

Proof. The Dvoretzky–Kiefer–Wolfowitz inequality [28] gives $\mathbb{P}(\|L_T - L\|_\infty > z) \le 2e^{-2Tz^2}$ and the same bound for $H_T^{\pi_0}$. A union bound gives joint failure probability at most

$$4e^{-2T(1/\sqrt{T}+\delta)^2} \le e^{-T\delta^2/2}.$$

Thus both (47) and (48) hold on the stated event. Independence of these two empirical CDFs is not required. □

By applying Lemma 7, with probability at least $1 - \exp\left(-T\varepsilon_0^2\delta^2/2\right)$, we have,

$$\sup_{u\in[0,1]}|\check{G}_T(u) - G(u)| \le \frac{2}{\varepsilon_0}\left(\frac{1}{\sqrt{T}} + \varepsilon_0\delta\right) = \frac{2}{\varepsilon_0\sqrt{T}} + 2\delta. \tag{49}$$

Upper bound on Term (b). Given that $\{x_t\}_{t=1}^T$ are i.i.d. samples from Q,

$$\|Q_T - Q\|_{\mathscr{F}_\Pi} = \sup_{h \in \mathscr{H}_\Pi} \left| \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{m \sim P} h(x_t, m) - \mathbb{E}_{(x,m)\sim J} h(x, m) \right|.$$

The class $\mathscr{F}_\Pi$ is 1-uniformly bounded. For independent auxiliary draws $m_1, \ldots, m_T \sim P$, Jensen's inequality gives

$$\begin{aligned} \mathscr{R}_T(\mathscr{F}_\Pi) &= \mathbb{E}_{x,\varepsilon} \left[\sup_{h \in \mathscr{H}_\Pi} \left| \frac{1}{T} \sum_{t=1}^T \varepsilon_t \mathbb{E}_{m_t \sim P} h(x_t, m_t) \right| \right] \\ &\leq \mathbb{E}_{x,m,\varepsilon} \left[\sup_{h \in \mathscr{H}_\Pi} \left| \frac{1}{T} \sum_{t=1}^T \varepsilon_t h(x_t, m_t) \right| \right] = \mathscr{R}_T(\mathscr{H}_\Pi). \end{aligned}$$

By applying Theorem 4, we obtain

$$\|Q_T - Q\|_{\mathscr{F}_\Pi} \leq 2\mathscr{R}_T(\mathscr{F}_\Pi) + \delta \leq 2\mathscr{R}_T(\mathscr{H}_\Pi) + \delta, \tag{50}$$

with probability at least $1 - \exp(-T\delta^2/2)$.

Regret Upper Bound. Combining (45), (49), and (50), we obtain

$$\begin{aligned} R_{\mathrm{DRO}}(\hat{\pi}_T) &\leq \frac{4}{\kappa} \left[\sup_{u \in [0,1]} |\check{G}_T(u) - G(u)| + \|Q_T - Q\|_{\mathscr{F}_\Pi} \right] \\ &\leq \frac{4}{\kappa} \left[\frac{2}{\varepsilon_0 \sqrt{T}} + 2\mathscr{R}_T(\mathscr{H}_\Pi) + 3\delta \right], \end{aligned} \tag{51}$$

with probability at least $1 - \exp(-T\delta^2/2) - \exp(-T\varepsilon_0^2\delta^2/2)$ by a union bound; the events are based on the same sample.

The bound (51) holds for a general policy class, without requiring finite cardinality or finite pseudo-dimension.

C.2.1 Policy-Class Complexity Bound

To complete the proof when $d < \infty$, we establish

$$\mathscr{R}_T(\mathscr{H}_\Pi) \leq C \sqrt{\frac{d+1}{T}}. \tag{52}$$

We bound the covering numbers of $\mathscr{H}_\Pi$ and apply the same entropy-integral argument as in Lemma 5.

Pseudo-dimension of the reward class. Let $\mathscr{U}_\Pi := \{r_\pi : \pi \in \Pi\}$. For any fixed (x_i, m_i, t_i) with $t_i \geq 0$,

$$r_\pi(x_i, m_i) > t_i \quad \Longleftrightarrow \quad m_i \leq \pi(x_i) < v(x_i) - t_i. \tag{53}$$

For $t_i < 0$, the label is always one. Thus each reward label is determined by the two policy comparisons at thresholds m_i and $v(x_i) - t_i$. Sauer's lemma gives, for $2n \geq d \geq 1$,

$$\operatorname{card}\left\{\left(\mathbf{1}_{r_\pi(x_i,m_i)>t_i}\right)_{i=1}^n : \pi \in \Pi\right\} \leq \sum_{j=0}^{d} \binom{2n}{j} \leq \left(\frac{2en}{d}\right)^d.$$

At $n = 5d$, $(10e)^d < 2^{5d}$, so $\operatorname{Pdim}(\mathscr{U}_\Pi) \leq 5d$. Strict and weak subgraph conventions have the same pseudo-dimension: for a finite shattered set, the thresholds can be shifted to preserve all labels of its finitely many witnessing functions. If $d = 0$, all policies coincide pointwise, and the reward class is a singleton.

For any probability measure ν on $\mathcal{X} \times [0,1]$ and $r, r' \in \mathscr{U}_\Pi$, we have

$$\begin{aligned}\|r - r'\|_{L_2(\nu)}^2 &\leq \int |r - r'| \, d\nu \\ &= \int\int_0^1 \left|\mathbf{1}_{r(x,m)>t} - \mathbf{1}_{r'(x,m)>t}\right|^2 \, dt \, d\nu(x,m).\end{aligned}$$

Applying Haussler's VC covering bound [29] to these binary subgraphs under $\nu \times \mathrm{Unif}[0,1]$ yields

$$N(\epsilon; \mathscr{U}_\Pi, L_2(\nu)) \leq e(5d+1)\left(\frac{2e}{\epsilon^2}\right)^{5d}, \qquad 0 < \epsilon \leq 1, \tag{54}$$

The class $\{\mathbf{1}_{r_\pi>0} : \pi \in \Pi\}$ is a restriction of the same subgraph class to $t = 0$, and hence satisfies the same bound.

Covering the exponential functions. Write $k_\alpha(s) := \alpha(1-e^{-s/\alpha})$, with $k_0(s) = 0$ and $k_\infty(s) = s$ denoting its uniform limits on $[0,1]$. For $\alpha > 0$,

$$0 \leq k_\alpha(s) \leq s, \qquad \partial_s k_\alpha(s) = e^{-s/\alpha} \leq 1, \qquad \partial_\alpha k_\alpha(s) = 1 - (1 + s/\alpha)e^{-s/\alpha} \geq 0.$$

For $0 < \alpha < \beta$, the difference $k_\beta - k_\alpha$ is nonnegative and increasing in s, since

$$\partial_s(k_\beta - k_\alpha)(s) = e^{-s/\beta} - e^{-s/\alpha} \geq 0.$$

Consequently,

$$\|k_\beta - k_\alpha\|_\infty = k_\beta(1) - k_\alpha(1).$$

As α increases from zero to infinity, $k_\alpha(1)$ increases continuously from zero to one. Therefore, allowing the two limiting functions as covering centers,

$$N(\epsilon; \{k_\alpha : \alpha \in [0,\infty]\}, \|\cdot\|_\infty) \leq 1 + \lceil 1/\epsilon \rceil.$$

For any $r, r' \in \mathscr{U}_\Pi$,

$$\|k_\alpha \circ r - k_\beta \circ r'\|_{L_2(\nu)} \leq \|r - r'\|_{L_2(\nu)} + \|k_\alpha - k_\beta\|_\infty.$$

Combining an $\epsilon/2$-cover of $\mathscr{U}_\Pi$ with an $\epsilon/2$-cover of the kernels, and then adding the positive-reward indicator class, gives

$$\begin{aligned}N(\epsilon; \mathscr{H}_\Pi, L_2(\nu)) \leq\ & N(\epsilon/2; \mathscr{U}_\Pi, L_2(\nu))\,(1 + \lceil 2/\epsilon \rceil) \\ &+ N(\epsilon; \{\mathbf{1}_{r_\pi>0} : \pi \in \Pi\}, L_2(\nu)).\end{aligned}$$

Covering centers in the closure are admissible, since the empirical Rademacher process is continuous in the empirical L_2 metric. By (54), the signed class $\mathscr{H}_\Pi^{\pm} := \mathscr{H}_\Pi \cup (-\mathscr{H}_\Pi) \cup \{0\}$ therefore satisfies, for $0 < \epsilon \le 1$,

$$N(\epsilon; \mathscr{H}_\Pi^{\pm}, L_2(\nu)) \le \frac{11e(5d+1)}{\epsilon} \left(\frac{8e}{\epsilon^2}\right)^{5d},$$

where we used $1 + \lceil 2/\epsilon \rceil \le 4/\epsilon$. Since $5d + 1 \le 6^d$, it follows that

$$\begin{aligned} \log N(\epsilon; \mathscr{H}_\Pi^{\pm}, L_2(\nu)) &\le \log(11e) + d \log\bigl(6(8e)^5\bigr) + (10d+1)\log(1/\epsilon) \\ &\le 12(d+1)\log(8/\epsilon). \end{aligned} \tag{55}$$

Rademacher Complexity and Regret. Let ν_T be the empirical distribution of any T points (x_t, m_t). Since $0 \in \mathscr{H}_\Pi^{\pm}$ and $\|h\|_{L_2(\nu_T)} \le 1$ for all $h \in \mathscr{H}_\Pi^{\pm}$, the covering number equals one for $\epsilon > 1$. Applying Dudley's entropy integral as in (42) and Jensen's inequality,

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\sup_{h \in \mathscr{H}_\Pi} \left| \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t h(x_t, m_t) \right| \right] &\le \frac{24}{\sqrt{T}} \int_0^1 \sqrt{\log N(\epsilon; \mathscr{H}_\Pi^{\pm}, L_2(\nu_T))}\, d\epsilon \\ &\le 24 \sqrt{\frac{12(d+1)}{T}} \int_0^1 \sqrt{\log(8/\epsilon)}\, d\epsilon \\ &\le 24 \sqrt{12(\log 8 + 1)} \sqrt{\frac{d+1}{T}} \\ &< 150 \sqrt{\frac{d+1}{T}}. \end{aligned}$$

Taking expectation over the sample proves (52) with $C = 150$. Substituting this bound into (51) proves (32).

References

[1] A. G. Khan, "Electronic commerce: A study on benefits and challenges in an emerging economy," *Global Journal of Management and Business Research*, 2016.

[2] T. Y. Kim, R. Dekker, and C. Heij, "Cross-border electronic commerce: Distance effects and express delivery in european union markets," *International Journal of Electronic Commerce*, vol. 21, no. 2, pp. 184–218, 2017.

[3] H. Hallikainen and T. Laukkanen, "National culture and consumer trust in e-commerce," *International Journal of Information Management*, vol. 38, no. 1, pp. 97–106, 2018.

[4] M. Faraoni, R. Rialti, L. Zollo, and A. C. Pellicelli, "Exploring e-loyalty antecedents in b2c e-commerce," *British Food Journal*, vol. 121, no. 2, pp. 574–589, 2019.

[5] G. Wagner, H. Schramm-Klein, and S. Steinmann, "Online retailing across e-channels and e-channel touchpoints: Empirical studies of consumer behavior in the multichannel e-commerce environment," *Journal of Business Research*, vol. 107, pp. 256–270, 2020.

[6] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," *The Journal of finance*, vol. 16, no. 1, pp. 8–37, 1961.

[7] S. Sluis, "Big changes coming to auctions, as exchanges roll the dice on first-price," https://www.adexchanger.com/platforms/big-changes-coming-auctions-exchanges-roll-dice-first-price/, 2017, published: September 5, 2017.

[8] AppNexus, "Demystifying auction dynamics for digital buyers and sellers," AppNexus white paper, 2018.

[9] J. Davies, "What to know about google's implementation of first-price ad auctions," https://digiday.com/media/buyers-welcome-auction-standardization-as-google-finally-goes-all-in-on-first-price/, 2019, published: September 6, 2019.

[10] S. Balseiro, N. Golrezaei, M. Mahdian, V. Mirrokni, and J. Schneider, "Contextual bandits with cross-learning," 2021. [Online]. Available: https://arxiv.org/abs/1809.09582

[11] Y. Han, Z. Zhou, and T. Weissman, "Optimal no-regret learning in repeated first-price auctions," 2024. [Online]. Available: https://arxiv.org/abs/2003.09795

[12] Y. Han, Z. Zhou, A. Flores, E. Ordentlich, and T. Weissman, "Learning to bid optimally and efficiently in adversarial first-price auctions," 2020. [Online]. Available: https://arxiv.org/abs/2007.04568

[13] W. Zhang, Y. Han, Z. Zhou, A. Flores, and T. Weissman, "Leveraging the hints: Adaptive bidding in repeated first-price auctions," 2022. [Online]. Available: https://arxiv.org/abs/2211.06358

[14] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," *Advances in neural information processing systems*, vol. 27, 2014.

[15] N. B. Keskin and A. Zeevi, "Chasing demand: Learning and earning in a changing environment," *Mathematics of Operations Research*, vol. 42, no. 2, pp. 277–307, 2017.

[16] N. Si, F. Zhang, Z. Zhou, and J. Blanchet, "Distributionally robust batch contextual bandits," *Management Science*, vol. 69, no. 10, pp. 5772–5793, 2023.

[17] Z. Hu and L. J. Hong, "Kullback-leibler divergence constrained distributionally robust optimization," *Available at Optimization Online*, 2013.

[18] Q. Wang, Z. Yang, X. Deng, and Y. Kong, "Learning to bid in repeated first-price auctions with budgets," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 36 494–36 513.

[19] A. Badanidiyuru, Z. Feng, and G. Guruganesh, "Learning to bid in contextual first price auctions," 2021.

[20] Q. Chen, N. Golrezaei, and D. Bouneffouf, "Non-stationary bandits with auto-regressive temporal dependency," 2023. [Online]. Available: https://arxiv.org/abs/2210.16386

[21] A. Javanmard, "Perishability of data: dynamic pricing under varying-coefficient models," *Journal of Machine Learning Research*, vol. 18, no. 53, pp. 1–31, 2017.

[22] M. Mohri and A. M. Medina, "Learning theory and algorithms for revenue optimization in second-price auctions with reserve," 2014. [Online]. Available: https://arxiv.org/abs/1310.5665

[23] N. Golrezaei, A. Javanmard, and V. Mirrokni, "Dynamic incentive-aware learning: Robust pricing in contextual auctions," 2020. [Online]. Available: https://arxiv.org/abs/2002.11137

[24] Y. Qu, R. Kant, Y. Chen, B. Kitts, S. Gultekin, A. Flores, and J. Blanchet, "Double distributionally robust bid shading for first price auctions," 2024.

[25] R. Kumar and O. Mouchtaki, "Prior-independent bidding strategies for first-price auctions," 2025. [Online]. Available: https://arxiv.org/abs/2502.09907

[26] Z. Fu, J. Jiang, and Y. Zhou, "Online bidding for contextual first-price auctions with budgets under one-sided information feedback," 2026. [Online]. Available: https://arxiv.org/abs/2603.07207

[27] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[28] P. Massart, "The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality," *The Annals of Probability*, vol. 18, no. 3, pp. 1269–1283, 1990.

[29] D. Haussler, "Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension," *Journal of Combinatorial Theory, Series A*, vol. 69, no. 2, pp. 217–232, 1995.