Learning to Bid in Repeated Second-Price Auctions with Dynamic Values and Aggregated Feedback

Benjamin Heymann[*]
Criteo AI Lab, Fairplay joint team
b.heymann@criteo.com

Otmane Sakhi[*]
Criteo AI Lab
o.sakhi@criteo.com

Abstract

We study the problem of learning to bid when the bidder's *value is dynamic*, i.e., when the current value depends on past outcomes. Specifically, we consider a bidder participating in repeated second-price auctions whose value depends on the time elapsed since their last successful bid, with auctions arriving in continuous time and aggregate value feedback revealed at the end of each episode. Such a bidder must **(1)** balance the immediate benefit of winning the current auction against its impact on future values and **(2)** learn unknown environmental parameters. We derive regret bounds for a class of learning methods that combine plug-in estimators with a differential-equation characterization of the optimal policy, and show that a specifically designed confidence-bound algorithm learns the optimal policy with near-optimal regret of $\widetilde{\mathcal{O}}(\log N)$ for piecewise-linear primitives, and $\widetilde{\mathcal{O}}(N^{1/5})$ for general smooth primitives, achieving these sublinear regrets without explicit randomization. These theoretical results are supported by numerical experiments.

1 Introduction

1.1 Dynamic values

In a repeated auction, a bidder's value is *dynamic* when it depends on the bidder's past actions. For example, when the auction is a posted-price mechanism and the item being auctioned is a yearly software license, the buyer derives no value from purchasing a second license if they have already bought one. The dynamic value model extends the usual static one by accounting for the fact that winning an auction can temporarily reduce the worth of future opportunities.

Dynamic value matters because it crucially changes the optimal bidding strategy in repeated auctions [1, 2]. Perhaps most strikingly, in a second-price auction, bidding the true value becomes *suboptimal* when values are dynamic. Hence, ignoring the dynamic dependence can lead to significant economic losses.

The dynamic value model for repeated auctions is particularly relevant for digital advertising, where displaying an ad to a user usually diminishes the immediate value of showing another ad to the same user shortly afterward [2–5]. This phenomenon is particularly evident in real-time bidding (RTB), where automated bidders repeatedly compete for ad slots shown to the same users. Research and industry practice both suggest that users experience *ad fatigue* when exposed to too many similar ads in a short period, leading to diminishing returns for the advertiser. For instance, the probability of a user engaging with an ad (e.g., clicking or making a purchase) tends to decrease if they see the same ad multiple times within a brief window [4]. Empirically, it is often observed that a bidder's valuation drops sharply after winning an auction and gradually recovers over time.

[*]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This dynamic nature introduces a planning problem: a bidding policy must carefully balance the immediate benefit of winning an auction against the potential long-term cost of devaluing future opportunities. In practice, automated bidders often implement *frequency capping* to limit the number of ads shown to a single user, further validating the relevance of such dynamic considerations in real-world scenarios. But, as argued in [6], the complexity introduced by dynamic value challenges the common belief that feature engineering and supervised learning are enough to solve an auto-bidding problem up to optimality. Instead, it often requires a more sophisticated approach, such as solving a planning problem with dedicated tools.

Heymann et al. [6] introduced the underlying continuous-time dynamic-value model and characterized an optimal bidding policy through differential equations when the value-recovery function and competition distribution are known. We retain their control characterization and optimal-policy solver, and study how to learn an optimal policy when both primitives are unknown.

The learner observes auction outcomes and winning payments, while the values generated by individual wins are unavailable; only their aggregate is revealed at the end of each episode. Learning must therefore account for both aggregate value feedback and censored observations of the competition.

The central difficulty is that the bidding policy determines which competition prices can be observed, while the relevant bidding range depends on the unknown environment. For example, if the learner never bids above a given threshold, it cannot distinguish competition distributions that agree below that threshold but differ above it. A learning guarantee must therefore establish which part of the competition distribution matters and how the learner obtains sufficient observations of that region.

Our analysis addresses this difficulty by connecting partial identifiability, estimation accuracy, and policy performance. This yields learning algorithms with finite-time regret guarantees, including a confidence-bound method that ensures sufficient coverage without explicit action randomization.

1.2 Contributions

We build on the continuous-time dynamic-value auction model and known-environment control solution of Heymann et al. [6], retaining their Bellman characterization and optimal-policy solver. We study the learning problem in which both the value-recovery function k and the competition distribution q are unknown. Auction outcomes and winning payments are observed, while value feedback is aggregated at the end of each episode.

The central challenge is policy-dependent identifiability: bids determine which competition prices can be observed, while the optimal bidding range depends on the unknown competition distribution. Our contributions are as follows.

1. **Structural learnability.** We establish an extension-invariance theorem showing that admissible changes to the competition distribution above the maximal optimal bid leave the optimal policy and value unchanged (Theorem 4.1). This identifies a sufficient region of the competition distribution to learn and supports asymptotic convergence of iterative plug-in learning under a positive-bid condition (Theorem 5.1).

2. **Estimation and policy stability.** We prove concentration guarantees for estimating k from aggregate episode-level values and q from censored auction observations under adaptively selected policies (Propositions 4.2 and 4.4). We further show that primitive-estimation errors of size ε yield $\mathcal{O}(\varepsilon)$ policy error and $\mathcal{O}(\varepsilon^2)$ value loss for sufficiently small ε (Proposition 5.3), providing the connection between statistical estimation and regret.

3. **Finite-time regret guarantees.** We analyze two exploration-and-commit strategies with $\mathcal{O}(\sqrt{N})$ regret (Theorems 5.4 and 5.5). We then establish comparative statics of the maximal optimal bid and use one-sided confidence estimates to ensure sufficient bid coverage without explicit action randomization (Theorems 5.6 and 5.7). For fixed piecewise-linear function classes with known breakpoints, the resulting algorithm achieves near-logarithmic regret.

4. **Extension to smooth primitives.** For continuously differentiable primitives with Lipschitz derivatives, we establish $\widetilde{\mathcal{O}}(N^{1/3})$ regret using piecewise-linear approximations on learner-chosen grids (Theorem 5.10). This extends the learning guarantees beyond fixed piecewise-linear function classes.

2 Related Work

The interplay between optimal bidding in repeated auctions and learning has been extensively studied, with researchers proposing a range of bandit-based approaches to address the challenge [7, 8]. It is notable that in [8], which corresponds to a static version of our setting, the authors derive a tight logarithmic lower bound under regularity assumptions similar to ours. Over the past few years, a variant of the problem with budget constraints has attracted a lot of attention [9–11].

The problem of sequence effects, where the outcome of one auction influences future bids, has also garnered attention. Early efforts formalized this challenge using a Markov decision process (MDP) framework, though the resulting policy optimization problem proved computationally intractable, necessitating heuristic solutions [2]. Subsequent work has focused on practical improvements: [4] demonstrated a first-order method in production, while [12] introduced a reparameterization technique to map timeline effects back to a contextual bandit setting. This latter approach was refined in [13], offering a more flexible parametrization.

Our approach is philosophically close to model-based reinforcement learning in continuous time [14, 15], which combines ideas from continuous-time RL [16] with model-based approaches. It is notable that the authors of [14] derived an algorithm with logarithmic regret in the number of episodes. Interestingly, our problem does not fit their framework, in particular because of the random, nonlinear rewards and the switching events that depend on the control.

Close to our application domain, [17] also justify the use of "mixed" and "delayed" feedback as a condition to account for causal effects. Following a landmark paper in the field of incrementality for advertising [18], they also use the time gap from the last ad impression as a state. Our results differ in that in our method, we do not need exploration of the states and provide a method with near-logarithmic regret, while they use a UCB algorithm to obtain regret of $O(\sqrt{N})$. Like us, they propose an episodic formulation, where each episode corresponds to a repeated interaction with an internet user. A notable difference is that they use a discrete-time model where the number of auctions is given in advance, which abstracts away a notable practical challenge. A different ad optimization problem was investigated in [19], where the authors consider the problem of reallocating ads over a window of known length. They also motivate their investigation with a behavioral model of the user (different from ours); however, their setting does not consider an auction.

3 Setting

3.1 Dynamic values

Our setting adapts the dynamic-value framework with a concave reward function, stationary competition, and second-price auctions from [6]. The initial input data correspond to an offline reinforcement learning dataset, where each episode consists of a sequence of second-price auctions, each generated by a policy $\pi_0 : t \in \mathbb{R}^+ \to [0, 1]$. The duration of each episode is drawn from an exponential distribution with rate parameter γ, independent of both the arrival process of auction requests and the outcomes of individual auctions. Auction arrivals within an episode follow an independent Poisson process with intensity μ.

The competition faced by the decision maker is stationary and *i.i.d.*, with the highest competing bid (price-to-beat) for each auction drawn independently from the continuous cumulative distribution function (CDF) q. For a bid b, $q(b)$ represents the probability that b exceeds the price-to-beat, thus securing a win. The classic *i.i.d.* assumption is justified in contexts where the bidder's market share is negligible compared to the overall market size, ensuring that individual auction outcomes do not influence the distribution of future competition.

At the end of each episode, the decision maker observes the total value generated by all auctions in that episode. This value is given by:

$$\sum_{i=1}^{\ell} \left(k(t_i - T_{i-1}) + \varepsilon_i - b_i^-\right) \cdot \mathbb{I}[b_i > b_i^-], \tag{1}$$

where ℓ is the random number of auctions in the episode, b_i is the bid placed by π_0 for the ith auction, b_i^- is the price-to-beat for the ith auction, t_i is the time of the ith auction, T_{i-1} is the time of the last

auction won in the episode, so $t_i - T_{i-1}$ is the elapsed time since the last win, $k(\cdot)$ is a concave and increasing function mapping the elapsed time since the last win to the expected reward, ε_i captures fluctuations around the mean reward $k(t_i - T_{i-1})$, and $\mathbb{I}[b_i > b_i^-]$ is the indicator of winning the ith auction. Conditional on the auction trajectory within an episode and the preceding history, the reward noises are independent, centered, and σ^2-sub-Gaussian.

We suppose that each episode starts with an initial $T_0 = 0$. Given a bidding policy π, we denote by $V_\pi(0)$ the resulting expected payoff.

The term $k(t_i - T_{i-1}) - b_i^-$ thus represents the average net profit from winning the auction: the reward $k(t_i - T_{i-1})$ minus the payment b_i^-. In contrast to the setting in [6], where the functions q and k are assumed known, the decision maker must here infer both q and k from observed data. However, even when the data are known, the resulting optimal control problem, with its reset condition, possesses an unusual structure that motivates specific tools for its resolution.

We focus on learning k and q, and assume that μ and γ are known. Estimating these rates is straightforward from observed auction arrivals and episode durations, and does not depend on the bidding policy. Their estimates can be plugged into the solver without changing the structure of our algorithms or requiring additional exploration. We treat them as known to simplify the analysis. We also assume that a fixed finite upper bound $B \geq 0$ is known, with $k(t) \leq B$ for every $t \geq 0$. The terminal value $k(\infty)$ itself need not be known.

3.2 Preliminaries, theoretical solution, and algorithm

We first consider the problem of maximizing criterion (1) given the parameters k and q. Because the auction follows a second-price rule, the price paid on average when bidding b can be inferred from q using the relation $p(b) = q(b)b - \int_0^b q(s)\,\mathrm{d}s$.

The Bellman value $V^\star(\tau)$ represents the maximal expected total value given that the time elapsed since the last won auction is τ. It is a standard method in optimal control to first derive the Bellman value and then use it to derive an optimal bid.

The next result, adapted from [6], provides a characterization of the Bellman value. Instead of solving a Cauchy problem where the initial condition is given, one must derive the unique initial condition that makes the solution bounded, a situation reminiscent of the shooting method in optimal control. The analysis relies on classical results on the dependence of ordinary differential equations on initial conditions and parameters (see Chapter 5 in [20]).

Theorem 3.1 (Adapted from [6])**.** *Let $f(v) = \max_{b\in[0,1]}\{q(b)v - p(b)\}$ and $\Phi(t, v, \lambda) = \gamma v - \mu f(k(t) + \lambda - v)$. The Bellman value function $V^\star(.)$ is the unique bounded element among the solutions of the following family of Cauchy problems parameterized by $y_0 \in \mathbb{R}_+$:*

$$\begin{cases} \dot{Y}(t) = \Phi(t, Y(t), y_0) \\ Y(0) = y_0. \end{cases} \tag{$\mathcal{F}_{y_0}$}$$

In what follows, we denote by Z^{y_0} the solution of $\mathcal{F}_{y_0}$. Next, adapting another result from [6], we note that one can derive an optimal policy if given the Bellman value at 0.

Lemma 3.2 (adapted from [6])**.** *If k is piecewise C^1, then, wherever k admits a derivative, the effective optimal bid $b^\star(t) = k(t) + V^\star(0) - V^\star(t)$ satisfies*

$$\dot{b}^\star(t) = \dot{k}(t) - \gamma(V^\star(0) + k(t)) + \mu f(b^\star(t)) + \gamma b^\star(t).$$

Proof: Appendix B.2.

The solver $\mathcal{S} : k, q \to \pi$ introduced by [6] uses a bisection method to estimate the value of the Bellman function at 0. It iteratively refines a guess by observing the asymptotic behavior of the corresponding ordinary differential equation (ODE) solution: if the solution diverges to $-\infty$, the guess is increased; if it diverges to $+\infty$, the guess is decreased. Combined with Theorem 3.3, this solver allows us to recover the optimal policy $\pi^\star$.

Theorem 3.3 (Adapted from [6])**.** *Let y be the output of the solver as N goes to $+\infty$. Then $Z^y(.) = V^\star(.)$. Moreover, $\pi^\star(\tau) = \min\{1, \max\{0, k(\tau) + V^\star(0) - V^\star(\tau)\}\}$.*

4 Learning tools

While [6] assumes oracle access to q and k and focuses on deriving an optimal bidding policy using a method based on differential equations, here, we suppose that we have access to an oracle that, given q and k, provides us with an optimal solution, but that q and k need to be estimated from the data. In what follows, we denote by $\mathcal{S}(k, q)$ the canonical optimal policy given by the Bellman bid formula for input data k and q. All submitted bids are projected onto the admissible interval $[0, 1]$.

Let $k^\star$ and $q^\star$ be the true primitives. Our goal is to sequentially build the bidding policies $\pi_1 \dots \pi_N$ to minimize the regret defined over a horizon of N episodes as $\mathrm{Regret}(N) = \sum_{n=1}^{N} \big(V_{\pi^\star}(0) - V_{\pi_n}(0)\big)$.

4.1 Further assumptions and notation

An insight from [6] is that for the bid to be well-behaved, k must be *concave*. This requirement prevents us from using alternative hypothesis sets, such as the set of step functions for k. In what follows, we suppose k and q are piecewise linear functions with known breakpoints $(t_i)_{i\in[0,I_k]}$ and $(b_i)_{i\in[0,I_q]}$ to study the core of the problem and extend our results to general primitives in Section 5.5.

Let $\mathcal{K}$ denote the set of **piecewise linear, concave, and increasing** functions whose breakpoints are in $(t_i)_{i\in[0,I_k]}$. Similarly, let $\mathcal{Q}$ denote the set of **piecewise linear, increasing, and bounded** functions, with breakpoints also restricted to $(b_i)_{i\in[0,I_q]}$. Let $\mathcal{P}$ denote the set of **optimal policies** when the parameters are $k \in \mathcal{K}$ and $q \in \mathcal{Q}$. For planning, the same solver also accepts any bounded, nonnegative, nondecreasing concave piecewise-linear recovery function with finitely many breakpoints, and any continuous nondecreasing function $r : [0, 1] \to [0, 1]$ with $r(0) = 0$. Thus clipping may introduce additional planning breakpoints or flat CDF segments. Only the restriction of r to the admissible bid interval matters; it can be extended to a continuous CDF beyond that interval. The fixed-grid classes above continue to specify the statistical estimators. We assume first that $k \in \mathcal{K}$ and $q \in \mathcal{Q}$ and then extend our results to general primitives. We assume there exists a constant $\alpha > 0$ such that for any bid $b \in (0, k(t_{I_k-1})]$, the cumulative distribution function q satisfies $q(b) > \alpha b$.

Given a policy $\pi : t \in [0, +\infty[\to b \in \mathbb{R}_+$, let $\beta(\pi) = \lim_{t\to+\infty} \pi(t)$ denote its asymptotic bid value. For a given policy π, we denote by $I(\pi)$ the smallest index i such that $\beta(\pi) \le b_i$. For $k \in \mathcal{K}$, $q \in \mathcal{Q}$, $\pi \in \mathcal{P}$ and $t \in \mathbb{R}_+$, we denote by $\mathcal{U}(k, q, \pi, t)$ the expected payoff of policy π when evaluated in the environment (k, q) and starting at time t.

4.2 Extension invariance theorem

We prove a core result below and believe it elegantly captures the interplay between the second-price auction mechanism, the dynamic optimization problem, and the learning problem.

Theorem 4.1 (Extension invariance)**.** *Let $k \in \mathcal{K}$ and $(\tilde{q}, \hat{q}) \in \mathcal{Q}^2$ such that $\tilde{q}(b) = \hat{q}(b)$ for all $b \in [0, \beta(\mathcal{S}(k, \hat{q}))]$. Then we have the identity*

$$\mathcal{U}(k, \hat{q}, \mathcal{S}(k, \hat{q}), 0) = \mathcal{U}(k, \tilde{q}, \mathcal{S}(k, \tilde{q}), 0).$$

Proof: See Appendix B.4.
Theorem 4.1 has significant implications for learning. Specifically, it guarantees that knowledge of $q^\star$ up to the maximal optimal bid $\beta(\mathcal{S}(k^\star, q^\star))$ is sufficient to identify the optimal policy.

4.3 Plug-in Estimator for k

Piecewise Linear Representation. We parameterize k as a concave and increasing piecewise linear function with coefficients $a = (a_0, \dots, a_{I_k})^\top$: $\forall \tau \ge 0\,, k(\tau) = a_0 + \sum_{i=1}^{I_k} a_i \cdot (t_i \wedge \tau - t_{i-1} \wedge \tau)$ with an unknown intercept $a_0 = k(0) \ge 0$, $a_{I_k} = 0$, $t_{I_k-1} = \bar{T}$ and $t_{I_k} = +\infty$.

Linear Regression Formulation. Each episode provides an observation relating the cumulative values of k at different ages τ (time elapsed since the last win) to the total episode value. Aggregating N episodes, we obtain a linear system where each episode n contributes an equation. The *age* τ at each auction is the time elapsed since the last won auction in that episode.

Data Representation. Suppose that after N episodes, we have access to a dataset $D_N^k = \{\tau_n^1, \cdots, \tau_n^\ell, v_n\}_{n\in[N_e]}$, where each n indexes an episode (a sequence of auctions before the episode ends). Here $N_e \leq N$ is the number of episodes with at least one won auction. Let M_n be the number of wins in episode n, and let $v_n = R_n + \sum_{j \text{ won in } n} p_j$ be its observed net payoff with the observed winning payments added back. This dataset can be transformed into a matrix form $\{X_n = [M_n, x_{1,n}, \ldots, x_{I_k,n}]^\top, v_n\}_{n\in[N_e]}$. Here $x_{i,n} = \sum_{j=1}^{M_n}(t_i \wedge \tau_n^j - t_{i-1} \wedge \tau_n^j)$, so the intercept contributes $M_n a_0$. Writing $X = [X_1, \ldots, X_{N_e}]$, with one design vector per column, we can express the relationship between the observed values and the coefficients as:

$$\forall n, \quad v_n = X_n^\top a + \epsilon_n \quad \iff \quad V = X^\top a + \epsilon,$$

where ϵ_n is a centered, conditionally sub-Gaussian noise term.

Estimation and Projection Let $\bar{a} = (XX^\top)^{-1}XV$ be the ordinary least squares (OLS) estimator of a. Define the convex set $\mathcal{C} = \left\{a \in \mathbb{R}^{I_k+1} \mid a_0 \geq 0,\ 0 = a_{I_k} \leq a_{I_k-1} \leq \cdots \leq a_1\right\}$, which encodes the monotonicity and concavity constraints. The true parameter a is assumed to lie in $\mathcal{C}$. We then define the projected estimator $\hat{a} = \text{PROJ}_{\mathcal{C}}(\bar{a})$, where $\text{PROJ}_{\mathcal{C}}$ denotes the projection onto the convex set $\mathcal{C}$. Given a dataset D_N^k, we denote by $\hat{k} = K(D_N^k)$ the resulting estimator $\hat{k}(\tau) = \hat{a}_0 + \sum_{i=1}^{I_k} \hat{a}_i(t_i \wedge \tau - t_{i-1} \wedge \tau)$.

Proposition 4.2 (Concentration of k - Informal)**.** *Let $\delta \in (0,1)$ and $\epsilon > 0$. The bidding policies π_n are allowed to depend on the past. If the bidding policies used to collect the data D_N^k constructed from N episodes satisfy $\pi_n(t) \geq \epsilon \min\{t/\bar{T}, 1\}$ for every $t > 0$, then, for sufficiently large N, with probability at least $1-\delta$, we have:*

$$\max_{\tau\geq 0} |k^\star(\tau) - \hat{k}(\tau)| \leq \mathcal{O}\left(1/\sqrt{N}\right).$$

Proof: Appendix C.5.

Remark 4.3. The exposure condition includes ages before the first positive breakpoint, allowing the intercept and the first slope to be identified separately. Positive constant exploration bids satisfy it.

4.4 Plug-in Estimator for q

As shown in [6], access to q is necessary to achieve the optimal policy. This introduces additional complexity compared to the static setting, where q does not need to be learned. Indeed, as demonstrated in [8], the static problem is truthful, and the optimal policy simply bids the value k.

Representation and Dataset We parameterize the win probability q as a piecewise linear, increasing function with slope coefficients $(c_i)_{i\in[I_q]}$: $\forall v \in [0,1]\,, q(v) = \sum_{i=1}^{I_q} c_i \cdot (b_i \wedge v - b_{i-1} \wedge v)\,$, where $c_i > 0$ for all i to ensure monotonicity, and the b_i are its breakpoints.

In second-price auctions, we observe wins but also the price paid when we win. Our dataset is therefore $\mathcal{D}^q = \{b_n, w_n, p_n\}_{n\in[N_b]}$, where b_n is the bid, $w_n \in \{0,1\}$ indicates a win, p_n is the price paid (with the convention $p_n = 0$ when $w_n = 0$), and N_b is the number of auctions observed.

Maximum Likelihood Estimation Since q is also the cumulative distribution of the competition with density $f_c(v) = \sum_{i=1}^{I_q} c_i \cdot \mathbb{I}[v \in (b_{i-1}, b_i]]$, we estimate the slope vector c via maximum likelihood as the maximizer $\hat{c}$ of

$$\sum_{n\in[N_b]} \left[\mathbb{I}[w_n = 1] \log f_c(p_n) + \mathbb{I}[w_n = 0] \log\left(1 - q_c(b_n)\right)\right].$$

Maximization is over nonnegative slopes satisfying $\sum_i c_i(b_i - b_{i-1}) \leq 1$. This is a *concave optimization problem*. MLE exploits the additional price information p_n from second-price auctions, achieving efficient estimation compared to methods based solely on win/loss outcomes. Given a dataset D_N^q, we denote by $\hat{q} = Q(D_N^q)$ the resulting estimator of q.

As stated in the extension-invariance theorem 4.1, knowledge of q up to the maximal optimal bid $\beta(\mathcal{S}(k^\star, q^\star))$ is sufficient to learn the optimal policy. Therefore, it is enough to focus on this region of the bid space and ensure convergence of the estimator there. The following result provides a sufficient condition for the concentration of our estimator of q.

Proposition 4.4 (Concentration of q - Informal)**.** *Let $\delta \in (0,1)$ and $T > 0$. The bidding policies π_n are allowed to depend on past observations. If the bidding policies used to collect the data D_N^q constructed from N episodes satisfy $\forall t \geq T, \pi_n(t) \geq \beta(\mathcal{S}(k^\star, q^\star))$, then, for sufficiently large N, with probability at least $1-\delta$, we have:*

$$\forall v \in [0, \beta(\mathcal{S}(k^\star, q^\star))],\ |q^\star(v) - \hat{q}(v)| \leq \mathcal{O}\left(1/\sqrt{N}\right).$$

Proof: See Appendix D.6.

Remark 4.5. Constructing bidding policies that cover the maximal optimal bid over a non-negligible time interval is sufficient to ensure efficient learning of q on the region of interest.

5 Learning algorithms

The model's two primitives, q and k, require *distinct treatments* due to differences in their identifiability. On the one hand, the valuation k can be estimated from aggregate episode values together with winning-auction ages, provided that bids ensure sufficient exposure across the recovery intervals. In contrast, the competition primitive q is only identifiable up to the highest observed bid threshold. If the bidding policy never submits bids above a certain value, the competitive environment beyond that threshold remains unobserved. This fundamental asymmetry between the identifiability of k and q, in connection with Theorem 4.1, shapes the design and analysis of our learning algorithms.

Algorithm 1 GENERAL LEARNING$(\mathcal{D})$

Require: Dataset $\mathcal{D} = (D^k, D^q)$

1: Set $\hat{k} = K(D^k)$
2: Set $\hat{q} = Q(D^q)$
3: Set $\hat{\pi} = \mathcal{S}(\hat{k}, \hat{q})$
4: Roll out $\hat{\pi}$

We now present four learning algorithms with different theoretical guarantees and practical merit. They all share the structure described in Algorithm 1: data are generated and used to learn possibly biased estimators, which are then plugged into the solver to produce a policy. This process is then repeated.

5.1 Asymptotic Convergence without Regret Guarantees

Our first result establishes that a simple iterative plug-in approach converges to the optimal policy without requiring explicit exploration, though it does not provide finite-time regret bounds.

Theorem 5.1 (Asymptotic Convergence)**.** *Consider initial datasets $\mathcal{D}_0^k$, $\mathcal{D}_0^q$ generated by a policy π_0, and the iterative sequence:*

$$\begin{aligned}
\pi_n &= \mathcal{S}(K(\mathcal{D}_n^k), Q(\mathcal{D}_n^q)) \\
\mathcal{D}_{n+1}^k &= \mathcal{D}_n^k \cup \mathcal{D}_{\pi_n}^k \\
\mathcal{D}_{n+1}^q &= \mathcal{D}_n^q \cup \mathcal{D}_{\pi_n}^q,
\end{aligned}$$

where $\mathcal{D}_\pi$ is a dataset generated by policy π over a strictly positive number of episodes. If

$$\limsup_{n\to\infty} \beta(\pi_n) > 0,$$

then for any $\varepsilon > 0$, π_n is ε-optimal for n sufficiently large.

Theorem 5.1 is a consequence of Theorem 4.1, through Lemma B.7.

Proof: See Appendix B.5.

Remark 5.2. The condition $\limsup_{n\to\infty} \beta(\pi_n) > 0$ ensures that the sequence of policies explores a region where the competition distribution q is identifiable. This theorem shows that no explicit randomization or exploration is required: the policy naturally converges to optimality through greedy plug-in estimation alone. Another way to understand this result is that self-confirming policies are optimal. However, this result does not provide finite-time regret guarantees, as convergence may be slow in the initial phases.

Guidelines for Algorithms with Sublinear Regret. We give the following core result that links ϵ-optimality of the estimators q and k to the optimality of $\pi = \mathcal{S}(k, q)$ and characterizes the regret.

Proposition 5.3 (ϵ-optimality)**.** *Let $\epsilon > 0$, and let q and k be such that $\|k^\star - k\|_\infty \leq \epsilon$ and $\max_{[0,\beta(\pi^\star)]} |q^\star(v) - q(v)| \leq \epsilon$. For small ϵ, we have:*

$$\|\pi^\star - \pi\|_\infty \leq \mathcal{O}(\epsilon)\,, \qquad\qquad \|V^\star - V^\pi\|_\infty \leq \mathcal{O}(\epsilon^2).$$

Proof: This is the result of combining Lemmas E.1 and E.2.

While Proposition 5.3 does not directly yield a no-regret algorithm, it provides the final ingredient needed to establish regret guarantees. In particular, ϵ-optimality can be achieved with a sufficiently accurate estimate of k and an estimate of q that is accurate up to $\beta(\mathcal{S}(k^\star, q^\star))$. To ensure this, and as stated in Propositions 4.2 and 4.4, the bidding policies must satisfy the recovery-exposure condition and cover the maximal optimal bid $\beta(\mathcal{S}(k^\star, q^\star))$. The difficulty is that this quantity is unknown, which forces sublinear regret strategies to exploit the structure of the problem in order to guarantee adequate coverage of this region.

5.2 Two-phase and Three-phase Algorithms

We next consider two explore-then-commit strategies that separate data collection from deployment of the learned policy. Both use an initial exploration period to estimate the primitives, followed by an exploitation period in which these estimates are plugged into the optimal-policy solver.

The two-phase strategy collects data for learning k and q simultaneously. During exploration, it bids $\min\{1, B\}$, ensuring sufficient observations of both the value dynamics and the relevant competition range. It then deploys the policy computed from the resulting estimates.

Theorem 5.4 (Two-phase algorithm)**.** *Consider the Learn-Then-Rollout algorithm:*

1. ***Exploration Phase** (N_1 **episodes**): Bid $\min\{1, B\}$ to collect data for estimating k and q. Compute $\hat{k} = K(\mathcal{D}^k_{N_1})$ and $\hat{q} = Q(\mathcal{D}^q_{N_1})$.*

2. ***Exploitation Phase** ($N_2 = N - N_1$ **episodes**): Use policy $\pi = \mathcal{S}(\hat{k}, \hat{q})$.*

Let $\delta \in (0, 1)$. There exists $C_0 \geq 9$ such that once $N \geq C_0$, setting $N_1 = \lceil \sqrt{N} \rceil$ yields with probability at least $1 - \delta$ that $\mathrm{Regret}(N) \leq \mathcal{O}(\sqrt{N})$.

Proof: See Appendix E.3.

A more gradual approach exploits the different information requirements for learning the two primitives: estimating k requires only a fixed positive bid, whereas estimating q requires bids that cover the relevant competition range.

The three-phase strategy therefore begins by estimating k using a fixed positive bid. It then uses the estimated value function as a bidding policy to collect data for estimating q. Once both primitives have been learned, it deploys the policy returned by the solver. This separates the exploration needed for each primitive while preserving the same regret order.

Theorem 5.5 (Three-phase Algorithm)**.** *Consider the Learn-k-Then-q-Then-Rollout algorithm:*

1. ***Exploration Phase for** k (N_1 **episodes**): Use a fixed exploratory policy $\pi_{\mathtt{x}}$ that bids $b_0 > 0$ to collect data for estimating k. Compute $\hat{k} = K(\mathcal{D}^k_{N_1})$.*

2. ***Exploration Phase for** q (N_2 **episodes**): Use the learned policy $\pi_k = \hat{k}$ to collect data for estimating q. Compute $\hat{q} = Q(\mathcal{D}^q_{N_2})$.*

3. ***Exploitation Phase** ($N_3 = N - N_1 - N_2$ **episodes**): Use policy $\pi = \mathcal{S}(\hat{k}, \hat{q})$.*

Let $\delta \in (0, 1)$. There exists $C_0 \geq 9$ such that once $N \geq C_0$, setting $N_1 = N_2 = \lceil \sqrt{N} \rceil$ yields with probability at least $1 - \delta$ that $\mathrm{Regret}(N) \leq \mathcal{O}(\sqrt{N})$.

Proof: See Appendix E.4.

Both strategies achieve $\mathcal{O}(\sqrt{N})$ regret, but differ in how they collect informative observations. The two-phase strategy uses the known bound to learn both primitives simultaneously, whereas the three-phase strategy first learns the value function and then uses it to guide exploration of the competition. The relation between the value function and the optimal bid (Theorem 3.3) underlies this second approach. The common regret order allows the exploration strategy to be chosen according to deployment constraints, although the regret constants and initial exploration costs may differ.

5.3 UCB-Style Algorithm with Biased Estimates

The phased strategies separate exploration from exploitation and commit to a fixed policy once the primitives have been estimated. We now consider an online strategy that updates its estimates after every episode. To maintain sufficient exploration, it uses an upper confidence bound for k and a lower confidence bound for q. The following monotonicity result explains how these one-sided adjustments control the maximal bid of the resulting policy.

Theorem 5.6 (Monotonicity of β)**.** *Let $k_1, k_2 \in \mathcal{K}$ such that $k_1 \leq k_2$ and $(q_1, q_2) \in \mathcal{Q}^2$ such that $q_1 \geq q_2$. Then*

$$\beta(\mathcal{S}(k_1, q_1)) \leq \beta(\mathcal{S}(k_2, q_2))\,.$$

Proof. We combine Lemmas B.2 and B.3. □

The monotonicity result suggests increasing the estimated value function and decreasing the estimated competition CDF within their confidence bounds, as formalized by Corollary B.6. On the confidence event, these adjustments ensure that the resulting policy's maximal bid covers the relevant competition range, without requiring knowledge of its endpoint. As more observations are collected, the confidence bounds narrow, allowing the policy to approach optimality while maintaining sufficient coverage. This motivates the following algorithm.

Theorem 5.7 (Confidence Bounds Algorithm)**.** *Choose $\lambda_k, \lambda_q > 0$. In the first $\min\{2, N\}$ episodes, submit the fixed bid 1 and augment both datasets using the update below. For each remaining episode $n \in \{3, \ldots, N\}$ do:*

- $\hat{k}_n^{\textit{UCB}} = \min(K(\mathcal{D}_n^k) + \lambda_k \sqrt{\log\log n / n}\,,\ B)$
- $\hat{q}_n^{\textit{LCB}} = \max(Q(\mathcal{D}_n^q) - \lambda_q \sqrt{\log\log n / n}\,,\ 0)$
- $\pi_n = \mathcal{S}(\hat{k}_n^{\textit{UCB}}, \hat{q}_n^{\textit{LCB}})$
- $\mathcal{D}_{n+1}^k = \mathcal{D}_n^k \cup \mathcal{D}_{\pi_n}^k$ *and* $\mathcal{D}_{n+1}^q = \mathcal{D}_n^q \cup \mathcal{D}_{\pi_n}^q$ *with* $\mathcal{D}_{\pi_n}^k$ *and* $\mathcal{D}_{\pi_n}^q$ *collected at episode* n.

The confidence-adjusted estimates ensure coverage of the relevant competition range.

Let $\delta \in (0, 1)$. There exist λ_q and λ_k (depending on problem constants and δ) that make the algorithm yield with probability at least $1 - \delta$ the Regret$(N) \leq \tilde{\mathcal{O}}(\log(N))$.

Proof: See Appendix E.6.

In the following section, we establish that this algorithm achieves near-minimax optimality. This algorithm shares structural similarities with the first algorithm; the key distinction lies in the use of confidence-adjusted estimates rather than unmodified point estimates.

In addition, optimizing over q does not require randomization, which is typical when learning to bid in first-price auctions. In contrast, the static second-price case does not provide an interesting point of comparison, since the truthfulness property of static second-price auctions implies that knowledge of the bid distribution is unnecessary [8, 1].

5.4 Lower Bound

We consider two degenerate cases where k is constant. In doing so, the problem reduces to the static case, with the important distinction that feedback arrives in batches and is aggregated. Since allowing the algorithm to de-aggregate the data freely can only improve its performance, any lower bound derived under this relaxation remains valid for the original problem. This reduction maps our setting

to one of the scenarios analyzed in [8], for which a lower bound has already been established. More precisely, we can apply their Theorem 4 with $\alpha = 1$ to get the next result.

Theorem 5.8. *For each fixed $B > 0$, there exists $C > 0$ such that, for every learning algorithm $\mathcal{A}$ and every $N \geq 2$, there is an admissible environment in which*

$$\mathbb{E}[\mathrm{Regret}_{\mathcal{A}}(N)] \geq C \log N.$$

The expectation is over the observations and any randomization of the algorithm.

Theorem 5.8 implies, in particular, that our last algorithm is near-minimax, and, surprisingly, that the learning problem when primitives are piecewise linear is not much harder than the static case.

5.5 General Smooth Primitives

A slight modification of the confidence-bound estimators extends the same plug-in approach to smooth primitives through piecewise-linear approximation. The modified estimators and their proofs are detailed in Appendix F.1; the algorithm still uses one optimal-policy solver call per episode.

Lemma 5.9 (Piecewise linear approximation)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuously differentiable and assume that f' is L-Lipschitz. Let $f_{PL,m}$ be the piecewise linear interpolant of f on a uniform grid of m intervals on $[a, b]$. Then $\|f - f_{PL,m}\|_\infty \leq \mathcal{O}(1/m^2)$.*

Theorem 5.10 (Regret for general smooth primitives)**.** *Assume a known finite cutoff $T > 0$, with $k(t) = k(T) > 0$ for $t \geq T$. Suppose that k' and q' are Lipschitz on $[0, T]$ and $[0, 1]$, respectively, with available upper bounds on their Lipschitz constants and on the reward-noise parameter. On uniform grids with $m \asymp N^{1/5}$ intervals, the modified confidence-bound algorithm of Appendix F.1 satisfies, with high probability,*

$$\mathrm{Regret}(N) \leq \widetilde{\mathcal{O}}(N^{1/5}).$$

Proof. See Appendix F.2.

The planning inputs remain piecewise linear; the true primitives need not be piecewise linear or have known breakpoints.

6 Experiments

In this section, we design a simple experimental setting to validate our algorithms. We consider a single environment with $k(t) = 1 - \exp(-\theta t)$ and $q(v) = v^\alpha$, where $\alpha = 2$ and $\theta = 0.1$. We use a family of piecewise linear functions with 10 uniform breakpoints as an approximation. We set $\mu = 0.5$ and $\gamma = 0.1$, and assume that these parameters are known. This environment illustrates behavior beyond the statistical assumptions of the analysis: $q(v) = v^2$ violates the linear lower-tail condition near zero and the positive survival margin at bid 1. We evaluate the no-regret algorithms on this environment. To compute the optimal policy, we numerically solve the associated ODE using the `SciPy` ODE solver with the default Runge–Kutta method. The results are reported in Figure 1, where each strategy is rerun for every fixed horizon N; they are overall consistent with our theoretical results.

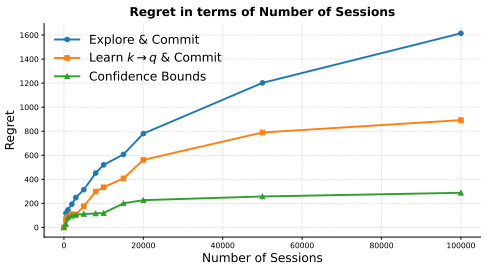


Figure 1: Regret as a function of N.

7 Conclusion and Future Work

In this work, we studied learning to bid with dynamic values, focusing on second-price auctions where the bidder's value depends on the time since their last win. Natural extensions include other auction formats, whose strategic features may interact differently with our resolution method. Another direction is to enrich the value model beyond recency, which may be too restrictive for some applications, allowing values to depend on the full history of won auctions, patterns of wins and losses, past prices paid, or random item-specific components capturing heterogeneity across objects. We believe these directions could make the model richer and relevant beyond computational advertising, including financial applications.

References

[1] Vijay Krishna. *Auction theory*. Academic press, 2009.

[2] Martin Bompaire, Alexandre Gilotte, and Benjamin Heymann. Causal models for real time bidding with repeated user interactions. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 75–85, New York, NY, USA, 2021. Association for Computing Machinery.

[3] Eustache Diemert, Julien Meynet, Pierre Galland, and Damien Lefortier. Attribution modeling increases efficiency of bidding in display advertising. In *Proceedings of the ADKDD'17*, pages 1–6. 2017.

[4] Benjamin Heymann, Alexandre Gilotte, and Rémi Chan-Renous. A pragmatic policy learning approach to account for users' fatigue in repeated auctions. *arXiv preprint arXiv:2407.10504*, 2024.

[5] Martin Bompaire, Antoine Désir, and Benjamin Heymann. Fixed point label attribution for real-time bidding. *Manufacturing & Service Operations Management*, 26(3):1043–1061, 2024.

[6] Benjamin Heymann, Alexandre Gilotte, and Rémi Chan-Renous. Repeated bidding with dynamic value. In Marios Mavronicolas, Qi Qi, and Grant Schoenebeck, editors, *Web and Internet Economics*, pages 548–563, Cham, 2026. Springer Nature Switzerland.

[7] Juliette Achddou, Olivier Cappé, and Aurélien Garivier. Efficient algorithms for stochastic repeated second-price auctions. In *Algorithmic Learning Theory*, pages 99–150. PMLR, 2021.

[8] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583. PMLR, 2016.

[9] Santiago R Balseiro and Yonatan Gur. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968, 2019.

[10] Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. Dual mirror descent for online allocation problems. In *International Conference on Machine Learning*, pages 613–628. PMLR, 2020.

[11] Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Giulia Romano, and Nicola Gatti. A unifying framework for online optimization with long-term constraints. *Advances in Neural Information Processing Systems*, 35:33589–33602, 2022.

[12] Artem Betlei, Mariia Vladimirova, Mehdi Sebbar, Nicolas Urien, Thibaud Rahier, and Benjamin Heymann. Maximizing the success probability of policy allocations in online systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11061–11068, 2024.

[13] Benjamin Heymann and Otmane Sakhi. Non-linear counterfactual aggregate optimization. *arXiv preprint arXiv:2509.03438*, 2025. Presented at Recsys 2025 Consequences workshop.

[14] Xin Guo, Anran Hu, and Yufei Zhang. Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls. *SIAM Journal on Control and Optimization*, 61(2):755–787, 2023.

[15] Cagatay Yildiz, Markus Heinonen, and Harri Lähdesmäki. Continuous-time model-based reinforcement learning. In *International Conference on Machine Learning*, pages 12009–12018. PMLR, 2021.

[16] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34, 2020.

[17] Ashwinkumar Badanidiyuru Varadaraja, Zhe Feng, Tianxi Li, and Haifeng Xu. Incrementality bidding via reinforcement learning under mixed and delayed rewards. *Advances in Neural Information Processing Systems*, 35:2142–2153, 2022.

[18] Randall Lewis and Jeffrey Wong. Incrementality bidding and attribution. *arXiv preprint arXiv:2208.12809*, 2022.

[19] Kailash Gopal Darmasubramanian, Akash Pareek, Arindam Khan, and Arpit Agarwal. Ads that stick: Near-optimal ad optimization through psychological behavior models. *arXiv preprint arXiv:2509.20304*, 2025.

[20] Philip Hartman. *Ordinary differential equations*. SIAM, 2002.

[21] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[22] Joel A. Tropp. Freedman's inequality for matrix martingales, 2011.

[23] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[24] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055 – 1080, 2021.

Appendix Outline

A Concentration Lemmata

B Proofs for the Theoretical Solution

C Proofs and remarks for Section 4.3 (concentration of K)

D Proofs and remarks for Section 4.4 (concentration of Q)

E Analysis of the algorithms

F Results for general primitives

G Analysis of the lower bound

A Concentration Lemmata

In this section, we give the different lemmas that will be essential to prove our regret bounds. We begin with classical concentration inequalities.

For any square matrix $A \in \mathcal{M}_n(\mathbb{R})$ of size n, we denote by $\lambda_{\min}(A)$ the smallest eigenvalue of A. We denote the operator norm of a matrix by $\|\cdot\|_{\text{OP}}$.

Lemma A.1 (Vector Hoeffding–Azuma inequality for martingales)**.** *Let $\{V_k\}_{k\in[n]}$ be a sequence of vectors adapted to a filtration $\mathcal{F}_k$. Assume the vectors are conditionally bounded ($||V_k||_2 \leq G$ almost surely, given $\mathcal{F}_{k-1}$) and have zero conditional mean ($\mathbb{E}[V_k|\mathcal{F}_{k-1}] = 0$) for all $k \in [n]$.*
We have the following result with probability $1-\delta$:

$$\left\| \sum_{k\in[n]} V_k \right\|_2 \leq G\sqrt{2n \log 2/\delta}$$

This lemma is the classical Hoeffding–Azuma result on scalar martingales extended to vector martingales [21].

Lemma A.2 (Matrix Chernoff bound for martingales)**.** *Let $\{X_k\}_{k\in[n]}$ be a sequence of $d \times d$ self-adjoint matrices adapted to a filtration $\mathcal{F}_k$. Assume the matrices are conditionally PSD and bounded:*

$$0 \leq X_k \leq R \cdot \mathcal{I}\,, \quad \text{almost surely, given } \mathcal{F}_{k-1}$$

Define the sum of conditional expectations:

$$M_n = \sum_{k=1}^{n} \mathbb{E}[X_k|\mathcal{F}_{k-1}]$$

For any $\epsilon \in [0,1)$ and a fixed target ν, we have:

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{k=1}^{n} X_k\right) \leq (1-\epsilon)\nu \quad \text{and} \quad \lambda_{\min}(M_n) \geq \nu\right) \leq d \cdot \exp\left(-\frac{\epsilon^2 \nu}{2R}\right)$$

To obtain this lemma, one can use the matrix Bernstein inequality from [22] and simple matrix manipulations to link the variance to the mean.

Finally, we give a matrix lemma that will be useful to relate the smallest eigenvalues of PSD matrices.

Lemma A.3 (Weyl's inequality)**.** *For two self-adjoint matrices A and B, we have:*

$$\lambda_{\min}(A) \geq \lambda_{\min}(B) - ||A-B||_{\text{OP}}\,.$$

B Proofs for the Theoretical Solution

B.1 Proof of Theorem 3.1

Proof. The first-auction Bellman map is a sup-norm contraction with factor $\mu/(\mu+\gamma)$. Integrating any bounded solution of ($\mathcal{F}_{y_0}$) gives a fixed point of this map. Conversely, the Bellman value satisfies ($\mathcal{F}_{y_0}$) with $y_0 = V^\star(0)$. Uniqueness of the bounded fixed point proves the claim. □

B.2 Proof of Lemma 3.2

Proof. Differentiate $b^\star(t) = k(t) + V^\star(0) - V^\star(t)$ wherever k is differentiable, and substitute $\dot{V}^\star(t) = \gamma V^\star(t) - \mu f(b^\star(t))$. Using $V^\star(t) = k(t) + V^\star(0) - b^\star(t)$ gives the stated equation. □

B.3 Properties of β

The next simple fact is used implicitly in many places throughout the appendix.

Lemma B.1. *For a canonical optimal policy $\pi \in \mathcal{P}$, $\pi(t) = \pi(\bar{T})$ for $t \geq \bar{T} = t_{I_k-1}$, and hence $\beta(\pi) = \pi(\bar{T})$.*

Proof. Starting from any age at least $\bar{T}$, recovery remains equal to $k(\infty)$ until the first win; afterward the state resets to zero. The control problems from all such initial ages are therefore identical. Their Bellman values agree, and the canonical bid formula gives the same bid. □

Lemma B.2 (Monotonicity of β w.r.t. k)**.** *Let $k_1, k_2 \in \mathcal{K}$ satisfy $k_1 \leq k_2$, and let $q \in \mathcal{Q}$. Then $\beta(\mathcal{S}(k_1, q)) \leq \beta(\mathcal{S}(k_2, q))$.*

Proof. Write V_i for the corresponding optimal value and $A_i = k_i(\infty) + V_i(0) - V_i(\bar{T})$ for its effective tail bid. Increasing rewards increases the value of every fixed policy, so $V_1(0) \leq V_2(0)$. Define $F_q(v) = \max_{b\in[0,1]}\{q(b)v - p(b)\}$. The constant-tail Bellman equation gives

$$\gamma A_i + \mu F_q(A_i) = \gamma\big(k_i(\infty) + V_i(0)\big).$$

The left side is strictly increasing in A_i, whereas the right side is nondecreasing in i. Thus $A_1 \leq A_2$, and projection onto the admissible bid interval gives the asserted ordering of maximal bids. □

Lemma B.3 (Monotonicity of β w.r.t. q)**.** *Let $k \in \mathcal{K}$ and $q_1, q_2 \in \mathcal{Q}$ such that $q_2 \leq q_1$. Then*

$$\beta(\mathcal{S}(k, q_2)) \geq \beta(\mathcal{S}(k, q_1)).$$

Proof. Write V_i for the value in environment (k, q_i), $a_i(t) = k(t) + V_i(0) - V_i(t)$, and $A_i = a_i(\bar{T})$. The canonical policy is the projection of a_i onto $[0, 1]$. The first-win comparison in Lemma C.2, applied to any two initial ages, gives $0 \leq V_i(t+h) - V_i(t) \leq [\mu/(\mu+\gamma)](k(t+h) - k(t))$; hence a_i is nondecreasing. Suppose for contradiction that $\beta_2 < \beta_1$. Then $A_2 = \beta_2 < 1$ and $A_1 > A_2$. Put $d = a_1 - a_2$ and $\bar{q}_i(u) = q_i(\min\{u, 1\})$ for $u \geq 0$. At any contact $a_1(t) = a_2(t) = z$, we have $z \leq A_2$. Subtracting the Bellman equations and using their constant-tail identities gives

$$d'(t) = \mu\left(\int_z^{A_2} \bar{q}_2(u)\,\mathrm{d}u - \int_z^{A_1} \bar{q}_1(u)\,\mathrm{d}u\right) + \gamma(A_2 - A_1) < 0.$$

The common recovery term cancels in d, so this derivative exists also at recovery breakpoints. But $d(0) = 0$, making d negative just after zero, while $d(\bar{T}) = A_1 - A_2 > 0$. An upward crossing of zero would have a nonnegative derivative, a contradiction. □

For $k \in \mathcal{K}$, $q \in \mathcal{Q}$, $\pi \in \mathcal{P}$ and $t \in \mathbb{R}_+$, we denote by $\mathcal{U}(k, q, \pi, t)$ the expected payoff of policy π when evaluated in the environment (k, q) and starting at time t. The function $\mathcal{U}$ enjoys the following regularity property.

(Lemma B.4 is used in the proof of Lemma B.7.)

Write $[v]_{[0,1]}$ for projection onto the admissible bid interval and

$$F_q(v) = \max_{b \in [0,1]} \{q(b)v - p_q(b)\} = \int_0^{\max\{v,0\}} q(\min\{u, 1\})\, \mathrm{d}u.$$

Here $p_q(b) = bq(b) - \int_0^b q(u)\, \mathrm{d}u$, and $F_q = f$ is 1-Lipschitz.

Lemma B.4. *Let $k, \hat{k} \in \mathcal{K}$, $q, \hat{q} \in \mathcal{Q}$, and $\pi = \mathcal{S}(k, q)$. If $\|k - \hat{k}\|_\infty \le \epsilon$ and $\sup_{b \in [0, \beta(\pi)]} |q(b) - \hat{q}(b)| \le \epsilon$, then*

$$\|\mathcal{U}(k, q, \pi, \cdot) - \mathcal{U}(\hat{k}, \hat{q}, \pi, \cdot)\|_\infty \le \frac{\mu}{\gamma}\big(1 + k(\infty)\big)\epsilon.$$

Proof. Let $V = V_{k,q}$ and $a(t) = k(t) + V(0) - V(t)$. Lemma C.2 gives $0 \le a(t) \le k(\infty)$, and $\pi(t) = [a(t)]_{[0,1]}$. At the continuation value V, the single-auction difference between the two fixed-policy Bellman expressions is

$$F_{\hat{q}}(a(t)) - F_q(a(t)) + \hat{q}(\pi(t))(\hat{k}(t) - k(t)).$$

Its absolute value is at most $(1 + k(\infty))\epsilon$: every CDF argument used in the integral up to $a(t)$ is at most $\pi(t)$ and hence at most $\beta(\pi)$. The first-auction Bellman map is a sup-norm contraction with factor $\rho = \mu/(\mu + \gamma)$, because the continuation coefficients $1 - q(b)$ and $q(b)$ are nonnegative and sum to one. Its residual is bounded by $\rho(1 + k(\infty))\epsilon$. Comparing its two fixed points and using $\rho/(1 - \rho) = \mu/\gamma$ proves the result. □

Lemma B.5 (Local optimal-value stability)**.** *Under the preceding assumptions, write $\hat{V} = V_{\hat{k},\hat{q}}$. Then*

$$\|\hat{V} - V_{k,q}\|_\infty \le \frac{\mu}{\gamma}(1 + k(\infty))\epsilon, \qquad \|\mathcal{S}(\hat{k}, \hat{q}) - \mathcal{S}(k, q)\|_\infty \le \left[1 + \frac{2\mu}{\gamma}(1 + k(\infty))\right]\epsilon.$$

Proof. Evaluate both optimal first-auction Bellman maps at $V = V_{k,q}$. Their maximized utilities differ by $F_{\hat{q}}(a(t) + \hat{k}(t) - k(t)) - F_q(a(t))$, whose absolute value is at most $(1 + k(\infty))\epsilon$, by the 1-Lipschitz property of $F_{\hat{q}}$ and the same local CDF bound as above. Both maps contract by $\mu/(\mu+\gamma)$, giving the value bound. The canonical effective bid identity, followed by the 1-Lipschitz projection onto $[0, 1]$, gives the policy bound. □

B.4 Proof of Theorem 4.1

Proof. Let $\hat{V}$ be the optimal value for $(k, \hat{q})$ and set $a(t) = k(t) + \hat{V}(0) - \hat{V}(t)$, so the canonical bid is $[a(t)]_{[0,1]} \le \beta(\mathcal{S}(k, \hat{q}))$. Agreement of the competition CDFs on that bid range implies $F_{\hat{q}}(a(t)) = F_{\tilde{q}}(a(t))$ at every age. Therefore $\hat{V}$ is also a bounded solution of the Bellman problem for $(k, \tilde{q})$. The first-auction Bellman map contracts by $\mu/(\mu+\gamma)$, so its bounded fixed point is unique. Hence $\hat{V} = \tilde{V}$, and the canonical bid formula yields identical policies and equal optimal values. □

The plateau, extension-invariance, maximal-bid monotonicity, and Bellman sensitivity and policy-stability arguments also apply to the planning inputs defined above. Their Bellman maps remain $\mu/(\mu + \gamma)$-contractions. Concavity gives nondecreasing canonical effective bids, and bounded piecewise-linear recovery with finitely many breakpoints is constant after a finite age. In the monotonicity proofs, choose a common age beyond the plateaus of both recovery functions. Neither fixed breakpoint locations nor strictly positive CDF slopes are used in those arguments. The quadratic-loss bound continues to use the Lipschitz constant of the true competition CDF.

Corollary B.6 (Local competition order)**.** *Let (k, q) be the true environment and put $\beta^\star = \beta(\mathcal{S}(k, q))$. Let (h, r) be planning inputs as above. If $h \geq k$ at every age and $r \leq q$ on $[0, \beta^\star]$, then*

$$\beta(\mathcal{S}(h, r)) \geq \beta^\star.$$

Proof. Set $r_+(b) = \max\{r(b), q(b)\}$ on $[0, 1]$. Then $r_+ = q$ on $[0, \beta^\star]$. The proof of Theorem 4.1 gives $\mathcal{S}(k, r_+) = \mathcal{S}(k, q)$. Since $h \geq k$ at every age and $r \leq r_+$ on the whole bid interval, maximal-bid monotonicity gives $\beta(\mathcal{S}(h, r)) \geq \beta(\mathcal{S}(k, r_+)) = \beta^\star$. □

The following lemma uses competition accuracy only on the learned policy's bid range.

Lemma B.7. *Let $k, \hat{k} \in \mathcal{K}$, $q, \hat{q} \in \mathcal{Q}$, and $\hat{\pi} = \mathcal{S}(\hat{k}, \hat{q})$. If $\|k - \hat{k}\|_\infty \leq \epsilon$ and $\sup_{b \in [0, \beta(\hat{\pi})]} |q(b) - \hat{q}(b)| \leq \epsilon$, then*

$$\mathcal{U}(k, q, \hat{\pi}, 0) \geq V_{k,q}(0) - \frac{2\mu}{\gamma}(1 + \hat{k}(\infty))\epsilon.$$

In particular, for $\epsilon \leq 1$ the coefficient can be replaced by $2\mu(2 + k(\infty))/\gamma$, a constant of the true environment.

Proof. Apply Lemma B.5 with the two models interchanged to compare their optimal values. Apply Lemma B.4, again with the models interchanged, to compare the fixed policy $\hat{\pi}$ in the two models. Both comparisons are bounded by $\mu(1 + \hat{k}(\infty))\epsilon/\gamma$, and $\mathcal{U}(\hat{k}, \hat{q}, \hat{\pi}, 0) = V_{\hat{k},\hat{q}}(0)$. Adding the two comparisons proves the claimed true-model payoff bound. □

B.5 Proof of Theorem 5.1

Proof. Let $\bar{\beta} = \limsup_{n \to \infty} \beta(\pi_n)$. By construction, for n large enough, we have an arbitrarily good approximation of $q^\star$ on $[0, \bar{\beta}]$ and of $k^\star$. The conclusion follows from Lemma B.7. □

Lemma B.8. *Let $a^\star(t) = k(t) + V^\star(0) - V^\star(t)$. For any admissible policy π,*

$$||V^\star_\cdot - V^\pi_\cdot||_\infty \leq \frac{\mu}{\gamma} ||D_{F_q}(a^\star_\cdot; \pi_\cdot)||_\infty \tag{2}$$

where $D_{F_q}(u, v) = F_q(u) - F_q(v) - F_q'(v)(u - v)$ is the Bregman divergence of F_q.

We use the subscript notation in the proof for better readability.

Proof. Let $t \geq 0$, let $\tau \sim \text{Exp}(\mu)$, and write $U(v, b) = q(b)v - p_q(b)$. By the dynamic programming relation (see [6]), $V^\star_t - V^\pi_t$ is equal to

$$\mathbb{E}_\tau \left[e^{-\gamma\tau} \left(V^\star_{t+\tau} + U(k_{t+\tau} + V^\star_0 - V^\star_{t+\tau}, \pi^\star_{t+\tau}) - V^\pi_{t+\tau} - U(k_{t+\tau} + V^\pi_0 - V^\pi_{t+\tau}, \pi_{t+\tau}) \right) \right]. \tag{3}$$

After adding and subtracting the term $U\left(k_{t+\tau} + V^\star_0 - V^\star_{t+\tau}, \pi_{t+\tau}\right)$ and observing that

$$U\left(k_{t+\tau} + V^\star_0 - V^\star_{t+\tau}, \pi_{t+\tau}\right) - U(k_{t+\tau} + V^\pi_0 - V^\pi_{t+\tau}, \pi_{t+\tau}) = q(\pi_{t+\tau}) \left(V^\star_0 - V^\pi_0 + V^\pi_{t+\tau} - V^\star_{t+\tau} \right)$$

we can rewrite (3) as

$$\mathbb{E}_\tau \Bigg[e^{-\gamma\tau} \Big(\underbrace{U(k_{t+\tau} + V^\star_0 - V^\star_{t+\tau}, \pi^\star_{t+\tau}) - U(k_{t+\tau} + V^\star_0 - V^\star_{t+\tau}, \pi_{t+\tau})}_{D_{F_q}(a^\star_{t+\tau}, \pi_{t+\tau})}$$
$$+ (V^\star_{t+\tau} - V^\pi_{t+\tau}) + q(\pi_{t+\tau}) \left(V^\star_0 - V^\pi_0 + V^\pi_{t+\tau} - V^\star_{t+\tau} \right) \Big) \Bigg],$$

which is upper bounded by

$$\frac{\mu}{\mu+\gamma}\big(||D_{F_q}(a^\star_\cdot;\pi_\cdot)||_\infty + ||V^\star_\cdot - V^\pi_\cdot||_\infty\big).$$

We have therefore proved that

$$||V^\star_\cdot - V^\pi_\cdot||_\infty \leq \frac{\mu}{\mu+\gamma}\big(||D_{F_q}(a^\star_\cdot;\pi_\cdot)||_\infty + ||V^\star_\cdot - V^\pi_\cdot||_\infty\big),$$

from which we can derive the claim. □

For a plug-in policy, put $\hat{a}(t) = \hat{k}(t) + \hat{V}(0) - \hat{V}(t)$. Since F_q is linear outside $[0,1]$,

$$D_{F_q}(a^\star(t), [\hat{a}(t)]_{[0,1]}) = D_{F_q}(a^\star(t), \hat{a}(t)).$$

C Proofs and remarks for Section 4.3 (concentration of K)

Overall Problem and Result

The value k of winning an auction is modeled by a concave, increasing piecewise linear function with an estimated intercept:

$$\forall \tau \geq 0, \qquad k(\tau) = a_0^\star + \sum_{i=1}^{I_k} a_i^\star (t_i \wedge \tau - t_{i-1} \wedge \tau).$$

Here

$$a^\star \in C = \{a \in \mathbb{R}^{I_k+1} : a_0 \geq 0,\ 0 = a_{I_k} \leq a_{I_k-1} \leq \cdots \leq a_1\}, \qquad t_{I_k-1} = T_f, \quad t_{I_k} = \infty.$$

In particular, $a_0^\star = k(0)$ is not fixed to zero, and constant recovery functions correspond to zero slope coefficients. The known finite cutoff may be taken positive, using redundant knots for a constant function. Mesh-width bounds $0 < \Delta_{\min}^k \leq t_i - t_{i-1} \leq \Delta_{\max}^k < \infty$ refer only to the finite intervals, $i \leq I_k - 1$.

At the end of episode n, let R_n denote the observed aggregate net payoff. Since winning payments are observed, the regression response is $v_n = R_n + \sum_{j \text{ won in } n} p_j$. If the episode contains M_n wins with inter-win ages $\tau_1^n, \ldots, \tau_{M_n}^n$, then, conditional on its auction trajectory and the previous history,

$$\mathbb{E}[v_n \mid \text{trajectory, past}] = M_n a_0^\star + \sum_{i=1}^{I_k} a_i^\star \sum_{j=1}^{M_n} (t_i \wedge \tau_j^n - t_{i-1} \wedge \tau_j^n) = \langle a^\star, z_n \rangle,$$

where

$$\phi(\tau) = \left(t_i \wedge \tau - t_{i-1} \wedge \tau\right)_{i=1}^{I_k}, \quad \psi(\tau) = (1, \phi(\tau)^\top)^\top, \quad z_n = \sum_{j=1}^{M_n} \psi(\tau_j^n).$$

Thus the intercept feature is the number of wins. Writing $v_n = z_n^\top a^\star + \eta_n$, the stated conditional reward-noise assumption gives conditional mean zero and variance at most $\sigma^2 M_n$ for η_n, where σ denotes the reward-noise scale. The conditioning is the centered conditional regression model used in Section 4.3; it does not condition on future adaptive designs.

For the analysis, index all N chronological episodes, inserting a zero design and response for zero-win episodes. These rows do not change the ordinary least-squares estimator based on the N_e nonzero rows. With Z having rows $z_n^\top$, the estimator in the main text is

$$\bar{a} = (Z^\top Z)^{-1} Z^\top V, \qquad \hat{a} = \text{PROJ}_C(\bar{a}), \qquad \hat{k}_N(\tau) = \hat{a}_0 + \sum_{i=1}^{I_k} \hat{a}_i (t_i \wedge \tau - t_{i-1} \wedge \tau).$$

The inverse exists on the high-probability event proved below. On singular data, set $\hat{k}_N \equiv 0$; this fixed default leaves the stated high-probability guarantee unchanged. If a known value cap is used operationally, capping $\hat{k}_N$ cannot increase its uniform error relative to the bounded true k.

In this section, we show that:

Let $\{\pi_n\}_{n \in [N]}$ be predictable episode policies, each allowed to depend on the previously collected data.

Condition C.1. *There exists $\epsilon > 0$ such that*

$$\pi_n(t) \geq \epsilon \min\{t/T_f, 1\} \qquad \textit{for every } t > 0 \textit{ and } n \in [N].$$

If Condition C.1 holds, then, for N large enough, with high probability,

$$\max_{\tau \geq 0} |k(\tau) - \hat{k}_N(\tau)| \leq \mathcal{O}(N^{-1/2}).$$

The all-age exposure premise permits the intercept and the first slope to be identified separately. A condition only at the first positive knot need not do so: a policy bidding zero at every earlier age

supplies no observations there, and its observed intercept and first-slope features can be collinear. No separate monotonicity assumption on collection policies is used below. Positive constant exploration bids satisfy the displayed premise.

To prove the result, we retain the same two ingredients: enough informative episodes and a lower bound on the aggregate regression's Gram matrix. The final proposition combines that lower bound with direct control of its centered score.

Lemma C.2 (A lower bound on canonical bids)**.** *For a nonnegative, nondecreasing concave recovery function k, write $b(t) = k(t) + V^\star(0) - V^\star(t)$ for its effective optimal bid. Then*

$$k(0) + \frac{\gamma}{\mu+\gamma}\big(k(t) - k(0)\big) \le b(t) \le k(t), \qquad b(0) = k(0).$$

Proof. Couple processes starting at ages 0 and t, submitting the same bids until their first win and using identical optimal continuations afterward. Concavity bounds the reward difference at that win by $k(t)-k(0)$. The first winning time is no earlier than the first auction, whose discounted probability is $\mu/(\mu+\gamma)$. Imitating each initial state's optimal policy in the other therefore gives $0 \le V^\star(t) - V^\star(0) \le [\mu/(\mu+\gamma)](k(t)-k(0))$. Substitute this inequality into $b(t) = k(t)+V^\star(0)-V^\star(t)$. □

The admissible canonical bid is obtained by projecting $b(t)$ onto the bid interval. If a concave planning primitive h_n dominates k, this lemma implies

$$\pi_n(t) \ge \min\left\{1, \frac{\gamma}{\mu+\gamma}k(\bar{T})\right\}\min\{t/\bar{T}, 1\}.$$

Thus the all-age recovery-exposure condition holds for optimistic canonical policies whenever the existing terminal value is positive.

Lemma making the dependence on the number of episodes explicit

Lemma C.3. *Let N_e be the number of episodes with at least one won auction and $\delta \in (0,1)$. Define*

$$p_w = e^{-\gamma T_f}\frac{\mu q(\epsilon)}{\gamma + \mu q(\epsilon)} > 0.$$

If $N \ge 8\log(1/\delta)/p_w$, then, with probability at least $1-\delta$, $N_e \ge Np_w/2$.

Proof. Before the first win, after T_f the winning hazard is at least $\mu q(\epsilon)$. Independent exponential termination therefore makes the conditional probability of an informative episode at least p_w. For its win indicator Y_n and any $\theta > 0$,

$$\mathbb{E}[e^{-\theta Y_n} \mid \mathcal{F}_{n-1}] \le \exp\{-p_w(1-e^{-\theta})\}.$$

Iteration and Markov's inequality at $\theta = \log 2$ yield $\mathbb{P}(N_e \le Np_w/2) \le e^{-Np_w/8} \le \delta$. □

Lemma to prove learnability

We lower-bound the actual aggregate Gram matrix $Z^\top Z$ without removing any observations from the estimator. Let $p = I_k + 1$, $T_+ = T_f + 1$, $g(t) = \min\{t/T_f, 1\}$, and define

$$B_k = \int_0^{T_+} q(\epsilon g(t))\psi(t)\psi(t)^\top dt, \qquad R = (1+T_+)^2,$$
$$\nu = \frac{\mu\gamma}{\mu+\gamma}e^{-(\mu+\gamma)T_+}\lambda_{\min}(B_k) > 0.$$

Decrease ϵ if necessary so that $0 < \epsilon \le 1$. The strictly positive slopes in the competition parameterization imply $q(\epsilon g(t)) > 0$ for $t > 0$.

Lemma C.4. *For $\delta \in (0,1)$, if $N \ge 8R\log(p/\delta)/\nu$, then, with probability at least $1-\delta$,*

$$\lambda_{\min}(Z^\top Z) \ge N\nu/2.$$

Proof. For a lower bound only, let E_n be the event that episode n contains exactly one auction, that it is won, and that its time s belongs to $[0, T_+]$. On this event $z_n = \psi(s)$. Thus $X_n = \mathbf{1}_{E_n}\psi(s)\psi(s)^\top$ satisfies $0 \preceq X_n \preceq z_n z_n^\top$ and $X_n \preceq RI$. Conditional on the past, the density of this sole winning auction is

$$\frac{\mu\gamma}{\mu+\gamma} e^{-(\mu+\gamma)s} q(\pi_n(s)).$$

The exposure condition and monotonicity of q imply $\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] \succeq \nu I$. The weighted matrix B_k is positive definite: a linear combination of the intercept and spline features that vanishes on the first open interval has zero intercept and first slope; its derivatives on each successive interval then force every remaining coefficient to be zero. This also treats the raw terminal-slope coordinate, since $T_+ > T_f$. Apply Lemma A.2 with relative error $1/2$ and use $Z^\top Z \succeq \sum_{n=1}^N X_n$. The informative event is used only in this proof; the estimator still uses every aggregate response. □

Final Concentration Proposition

There is no deterministic bound on the number of auctions in an exponentially terminated Poisson episode. Instead, let T_n denote its duration and $A_n \geq M_n$ its number of arrivals. Inter-win ages sum to the time of the last win, so

$$\|z_n\|_2 \leq \|z_n\|_1 = M_n + \sum_{j=1}^{M_n} \tau_j^n \leq M_n + T_n.$$

Consequently the score $z_n\eta_n$ is a martingale difference and

$$\mathbb{E}[\|z_n\eta_n\|_2^2 \mid \mathcal{F}_{n-1}] \leq \sigma^2 \mathbb{E}[A_n(A_n + T_n)^2] =: s^2 < \infty,$$

because $T_n \sim \mathrm{Exp}(\gamma)$ and $A_n \mid T_n \sim \mathrm{Poisson}(\mu T_n)$ independently of the past. These are the original arrival and termination assumptions.

Proposition C.5. *Let $\delta \in (0,1)$ and define*

$$N_0^k = \left\lceil \frac{8R}{\nu} \log \frac{2p}{\delta} \right\rceil.$$

If $N \geq N_0^k$, then, with probability at least $1-\delta$,

$$\|\hat{a} - a^\star\|_2 \leq \|\bar{a} - a^\star\|_2 \leq \frac{2\sqrt{2}s}{\nu\sqrt{N\delta}},$$

$$\max_{\tau \geq 0} |k(\tau) - \hat{k}_N(\tau)| \leq \frac{2\sqrt{2}\sqrt{1+T_f^2}\, s}{\nu\sqrt{N\delta}}.$$

In particular, both errors are $\mathcal{O}(N^{-1/2})$ for a fixed grid and fixed confidence level.

Proof. Put $S_N = \sum_{n=1}^N z_n\eta_n$. Martingale orthogonality gives $\mathbb{E}\|S_N\|_2^2 \leq Ns^2$; Markov's inequality therefore gives $\|S_N\|_2 \leq s\sqrt{2N/\delta}$ with probability at least $1 - \delta/2$. Lemma C.4, with failure probability $\delta/2$, gives $\lambda_{\min}(Z^\top Z) \geq N\nu/2$. On their intersection, the exact ordinary least-squares identity

$$\bar{a} - a^\star = (Z^\top Z)^{-1} S_N$$

proves the coefficient bound. Euclidean projection onto C is nonexpansive and fixes $a^\star$. Both projected terminal slopes are zero; the remaining function feature vector has Euclidean norm at most $\sqrt{1+T_f^2}$ at every age. This proves the function bound, and operational capping cannot increase it. □

D Proofs and remarks for Section 4.4 (concentration of Q)

Overall Problem and Result

We model the probability q of winning an auction as an increasing, piecewise linear function:

$$\forall v \in [0,1]\,, \quad q(v) = q_{c^\star}(v) = \sum_{i=1}^{I_q} c_i^\star \cdot (b_i \wedge v - b_{i-1} \wedge v)\,,$$

Write the fixed partition of the admissible bid interval as $0 = b_0 < b_1 < \cdots < b_{I_q} = 1$. We have $c_i^\star > 0$, $\sum_i c_i^\star \cdot \Delta_i^q < 1$ and $\Delta_i^q = b_i - b_{i-1}$.

In second-price auctions, q is also the cumulative distribution of the competition, with density $f_{c^\star}(v) = \sum_{i=1}^{I_q} c_i^\star \cdot \mathbb{I}[v \in (b_{i-1}, b_i]]$.

We add the following technical conditions:

- $\forall i, \quad 0 < c_{\min} \leq c_i^\star \leq c_{\max} < \infty$.
- $\forall i, \quad 0 < \Delta^q_{\min} \leq \Delta_i^q \leq \Delta^q_{\max} < \infty$.
- We have for any bid $b \in [0,1]$, $q(b) < 1 - \eta$, with $\eta > 0$.

In our setting, we suppose that we interact with N sessions. After these N sessions, we participate in N_b auctions and collect data of the form:

$$\mathcal{D}^q_{N_b} = \{b_n, w_n, p_n\}_{n \in [N_b]}\,,$$

with b_n the bid placed, w_n the binary win variable and p_n the price paid, with the convention that $p_n = 0$ when $w_n = 0$.

The feasible set for the CDF estimator is

$$\mathcal{C}_q = \left\{ c \in [0,\infty)^{I_q} : \sum_{i=1}^{I_q} c_i \Delta_i^q \leq 1 \right\}.$$

We use any measurable maximizer on this closed set, including boundary maximizers, with $0 \log 0 = 0$. Such a maximizer exists: the set is compact, the log-likelihood is upper semicontinuous, and it is finite at the true parameter. On an empty dataset, use any fixed element of $\mathcal{C}_q$.

Our objective is to estimate the slope vector $c^\star$ with all observed data, using MLE and finding:

$$\begin{aligned}\hat{c} &= \arg\max_{c \in \mathcal{C}_q} \sum_{n \in [N_b]} \left[\mathbb{I}[w_n = 1] \log f_c(p_n) + \mathbb{I}[w_n = 0] \log\left(1 - q_c(b_n)\right)\right] \\ &= \arg\max_{c \in \mathcal{C}_q} \sum_{n \in [N_b]} L_n(c)\,.\end{aligned}$$

This problem is concave and can be solved by a variety of optimization algorithms [23]. After N sessions, we learn $\hat{c}$ and construct our $\hat{q}_N$ estimator with:

$$\forall v \in [0,1]\,, \quad \hat{q}_N(v) = \sum_{i=1}^{I_q} \hat{c}_i \cdot (b_i \wedge v - b_{i-1} \wedge v)\,,$$

Let $\beta(\pi^\star) = \lim_{t \to +\infty} \pi^\star(t)$ denote the asymptotic bid value of the optimal policy, and let $i_\star = I(\pi^\star)$ be the smallest index i such that $\beta(\pi^\star) \leq b_i$. Using our q extension theorem (Theorem 4.1), we want to accurately estimate q on the bid range $[0, \beta(\pi^\star)]$ if we hope to identify the optimal policy. We also control $\|\hat{c}_{:i_\star} - c^\star_{:i_\star}\|_2$, where $c_{:i}$ denotes the first i coordinates of c. In this section, we show that:

Let $\{\pi_n\}_{n\in[N]}$ be the sequence of bidding policies, with each π_n allowed to depend on the previously collected data.

Condition D.1. *There exists $T_0 > 0$ s.t. for all $t \geq T_0$, $\pi_n(t) \geq \beta(\pi^\star)$ for all $n \in [N]$.*

Under Condition D.1, for N large enough, we have with high probability:

$$\max_{v\in[0,\beta(\pi^\star)]} |q(v) - \hat{q}_N(v)| \leq \mathcal{O}\left(\sqrt{1/N}\right) .$$

We suppose that Condition D.1 holds throughout this section.

Lemma making the dependence on the number of auctions explicit

Lemma D.2. *Let N_b be the random variable representing the number of auctions. For any $\delta \in (0,1)$, if $N \geq 32(1+\gamma/\mu)\log(2/\delta)$, then we have with probability at least $1-\delta$:*

$$N_b \geq N\frac{\mu}{2\gamma} .$$

Proof. Let $T_n \sim \mathrm{Exp}(\gamma)$ be the termination time of session n and, conditional on T_n, let $B_n \mid T_n \sim \mathrm{Poisson}(\mu T_n)$ be the number of arriving auctions. Define $N_b = \sum_{n=1}^N B_n$ and $S_T = \sum_{n=1}^N T_n$. Conditionally on (T_n), we have $N_b \sim \mathrm{Poisson}(\Lambda)$ with $\Lambda = \mu S_T$. A Chernoff bound gives

$$\mathbb{P}\left(S_T < \frac{3N}{4\gamma}\right) \leq \exp\left[-N\left(\log(4/3) - 1/4\right)\right] \leq \frac{\delta}{2}$$

under the stated sample-size condition. On the complementary event, $\Lambda \geq 3\mu N/(4\gamma)$, and the Poisson lower-tail bound gives

$$\mathbb{P}\left(N_b < \frac{\mu N}{2\gamma} \,\middle|\, (T_n)\right) \leq \exp(-\Lambda/18) \leq \exp\left(-\frac{\mu N}{24\gamma}\right) \leq \frac{\delta}{2}.$$

A union bound proves the result. □

A separable representation of the likelihood

Set $p_i = c_i\Delta_i^q$, $S_0 = 1$, and $S_i = 1 - \sum_{j=1}^i p_j$. Define the bin hazards by

$$h_i = \frac{p_i}{S_{i-1}}, \qquad p_i = h_i \prod_{j<i}(1-h_j).$$

Together with $c_i = p_i/\Delta_i^q$, the inverse formula maps $[0,1]^{I_q}$ onto $\mathcal{C}_q$. If $S_{i-1} = 0$, subsequent hazards can be chosen arbitrarily; the corresponding bin probabilities are all zero.

A winning price in bin r contributes

$$\log f_c(p) = \log h_r + \sum_{j<r} \log(1-h_j) - \log \Delta_r^q.$$

For a losing bid $b \in (b_{r-1}, b_r)$, put $\alpha_b = (b - b_{r-1})/\Delta_r^q$. Its contribution is

$$\log(1 - q_c(b)) = \sum_{j<r} \log(1-h_j) + \log(1-\alpha_b h_r).$$

Bids at grid endpoints are handled by the same identities with $\alpha_b = 0$ or 1.

Let W_i be the number of winning prices in $(b_{i-1}, b_i]$. Let R_i count the observations consisting of either a winning price greater than b_i or a loss with bid at least b_i, and let $\mathcal{J}_i$ index losses with bid strictly between b_{i-1} and b_i. The full log-likelihood is therefore

$$L(h) = \text{constant} + \sum_{i=1}^{I_q} \ell_i(h_i), \qquad \ell_i(h) = W_i \log h + R_i \log(1-h) + \sum_{a\in\mathcal{J}_i} \log(1-\alpha_{b_a} h).$$

Thus each hazard coordinate of a full MLE maximizes ℓ_i on $[0,1]$. All the observations in the original likelihood are retained.

Lemma D.3. *Let $h_i^\star$ be the true hazard and let $\hat{h}_i$ be a maximizing hazard coordinate. Whenever $W_i > 0$,*

$$|\hat{h}_i - h_i^\star| \le \frac{2|\ell_i'(h_i^\star)|}{W_i}.$$

Proof. For $h \in (0,1)$,

$$-\ell_i''(h) = \frac{W_i}{h^2} + \frac{R_i}{(1-h)^2} + \sum_{a \in \mathcal{J}_i} \frac{\alpha_{b_a}^2}{(1-\alpha_{b_a} h)^2} \ge W_i.$$

Concavity and integration of this bound give

$$0 \le \ell_i(\hat{h}_i) - \ell_i(h_i^\star) \le \ell_i'(h_i^\star)(\hat{h}_i - h_i^\star) - \frac{W_i}{2}(\hat{h}_i - h_i^\star)^2.$$

This inequality extends to maximizing endpoints by a limit. Taking absolute values proves the claim. □

Exposure and score concentration

Write $\beta^\star = \beta(\pi^\star)$. If $\beta^\star = 0$, then $q(0) = \hat{q}_N(0) = 0$, and the required CDF bound is immediate. Suppose below that $\beta^\star > 0$, and write $d = i_\star$ and

$$u_i = \min\{b_i, \beta^\star\} - b_{i-1} > 0, \qquad r_i = \frac{\mu}{\mu+\gamma} e^{-(\mu+\gamma)T_0} c_i^\star u_i, \qquad r = \min_{1 \le i \le d} r_i > 0.$$

Lemma D.4 (Covered winning counts)**.** *For every N,*

$$\mathbb{P}\left(\min_{1 \le i \le d} W_i < \frac{rN}{2} \right) \le d e^{-rN/8}.$$

Proof. In episode n, let $A_n \sim \mathrm{Exp}(\mu)$ be its first auction-arrival time and let $T_n \sim \mathrm{Exp}(\gamma)$ be its independent termination time. Consider the event

$$A_n \ge T_0, \quad A_n < T_n, \quad P_n \in (b_{i-1}, \min\{b_i, \beta^\star\}],$$

where P_n is the first auction's competition price. This event has probability

$$\int_{T_0}^{\infty} \mu e^{-(\mu+\gamma)t} \, \mathrm{d}t \; c_i^\star u_i = r_i.$$

There has been no previous auction in the episode, so its age is A_n. Condition D.1 ensures that this auction is won. These events are independent across episodes, including when the episode policies depend on past observations. Consequently, W_i dominates a $\mathrm{Binomial}(N, r_i)$ random variable. A multiplicative Chernoff bound and a union bound prove the result. □

Define

$$\theta = \log\left(1 + \frac{\gamma}{2\mu}\right), \qquad C_b = \frac{1 + \log 2}{\theta}, \qquad G = \max_{1 \le i \le d} \max\left\{ \frac{1}{h_i^\star}, \frac{1}{1 - h_i^\star} \right\}.$$

The standing assumptions imply $G \le \max\{(c_{\min}\Delta_{\min}^q)^{-1}, \eta^{-1}\} < \infty$.

Lemma D.5 (Upper auction count and hazard scores)**.** *For every N and every $x > 0$,*

$$\mathbb{P}(N_b > C_b N) \le e^{-N},$$

and

$$\mathbb{P}\left(\max_{1 \le i \le d} |\ell_i'(h_i^\star)| > G\sqrt{2C_b N x}, \quad N_b \le C_b N \right) \le 2de^{-x}.$$

Proof. The auction count B_n in one episode satisfies

$$\mathbb{E}[e^{\theta B_n}] = \mathbb{E}[e^{\mu T_n(e^\theta - 1)}] = \frac{\gamma}{\gamma - \mu(e^\theta - 1)} = 2.$$

Episode clocks are independent, so Markov's inequality gives $\mathbb{P}(N_b > C_b N) \le 2^N e^{-\theta C_b N} = e^{-N}$.

Condition on all the exogenous auction-arrival and termination clocks. The number N_b is then fixed. Current competition prices remain independent of the past, and bids are predictable. The contribution of one auction to the true hazard score is one of

$$\frac{1}{h_i^\star}, \qquad -\frac{1}{1-h_i^\star}, \qquad -\frac{\alpha_b}{1-\alpha_b h_i^\star}, \qquad 0.$$

The contributions are conditionally centered scores of the observed likelihood, and their absolute values are at most G. Scalar Hoeffding–Azuma therefore gives, conditionally on these clocks,

$$\mathbb{P}\left(|\ell_i'(h_i^\star)| > G\sqrt{2N_b x} \,\middle|\, \text{clocks}\right) \le 2e^{-x}.$$

On clock realizations with $N_b \le C_b N$, the threshold $G\sqrt{2C_b N x}$ is no smaller. Integrating over the clocks and taking a union bound over $i \le d$ proves the second assertion. □

The Final Concentration Result for q

Proposition D.6. *Let $\delta \in (0,1)$ and suppose that Condition D.1 holds. If $\beta^\star = 0$, the CDF error on $[0, \beta^\star]$ is zero, and we take $N_0^q = 1$. If $\beta^\star > 0$, define the constants above, put $L_\delta = \log(6d/\delta)$, and set*

$$N_0^q = \left\lceil \max\left\{1, \log(3/\delta), \frac{8}{r}\log(3d/\delta)\right\}\right\rceil.$$

For $N \ge N_0^q$, with probability at least $1-\delta$,

$$\sup_{v\in[0,\beta^\star]} |q(v) - \hat{q}_N(v)| \le \frac{4dG}{r}\sqrt{\frac{2C_b L_\delta}{N}},$$

and

$$\|\hat{c}_{:d} - c^\star_{:d}\|_2 \le \frac{4d^{3/2}G}{r\Delta^q_{\min}}\sqrt{\frac{2C_b L_\delta}{N}}.$$

Proof. Lemma D.4, the upper count bound, and Lemma D.5 with $x = L_\delta$ yield, with total failure probability at most δ,

$$W_i \ge rN/2, \qquad |\ell_i'(h_i^\star)| \le G\sqrt{2C_b N L_\delta} \quad \text{for all } i \le d.$$

The three failure probabilities are bounded by $de^{-rN/8} \le \delta/3$, $e^{-N} \le \delta/3$, and $2de^{-L_\delta} = \delta/3$, respectively. Lemma D.3 now gives

$$\max_{i\le d} |\hat{h}_i - h_i^\star| \le \frac{4G}{r}\sqrt{\frac{2C_b L_\delta}{N}}.$$

The covered winning counts are positive, so every maximizing CDF has positive mass in every covered bin. Its hazards through bin d are therefore well defined and satisfy this bound.

For $v \in [b_{j-1}, b_j]$ with $j \le d$, put $\alpha = (v - b_{j-1})/\Delta_j^q$. The identity

$$q_h(v) = 1 - \prod_{i<j}(1-h_i)(1-\alpha h_j)$$

and a telescoping product bound imply

$$|q_{\hat{h}}(v) - q_{h^\star}(v)| \le \sum_{i\le j} |\hat{h}_i - h_i^\star| \le d \max_{i\le d} |\hat{h}_i - h_i^\star|.$$

Similarly, $c_i = h_i \prod_{j<i}(1-h_j)/\Delta_i^q$ gives

$$|\hat{c}_i - c_i^\star| \leq \frac{1}{\Delta_i^q} \sum_{j \leq i} |\hat{h}_j - h_j^\star|.$$

Taking a supremum in the first bound and a Euclidean norm in the second proves both assertions. $\square$

E Analysis of the algorithms

The concentration results require Conditions C.1 and D.1, which ensure sufficient exposure at all ages to learn k, including its intercept, and over the relevant bid range to learn q. Under these conditions, for sufficiently large N, we obtain the high-probability bounds:

$$\|k - \hat{k}_N\|_\infty \le C_k/\sqrt{N}$$
$$\max_{v\in[0,\beta(\pi^\star)]} |q(v) - \hat{q}_N(v)| \le C_q/\sqrt{N}\,.$$

We take $C = \max\{C_k, C_q\}$ and define $\epsilon_N = C/\sqrt{N}$.

We have by definition $\pi^\star = \mathcal{S}(k, q)$, and we define $\pi_N = \mathcal{S}(\hat{k}_N, \hat{q}_N)$. We start by giving the following lemma:

Lemma E.1. *Write $a^\star(t) = k(t) + V^\star(0) - V^\star(t)$ and $P(v) = \min\{1, \max\{v, 0\}\}$. For any bounded $\tilde{a}$ and $\pi = P \circ \tilde{a}$, let $F_q(v) = \max_{b\in[0,1]}\{q(b)v - p(b)\}$. Then*

$$\|V^\star - V^\pi\|_\infty \le \frac{\mu}{\gamma}\|D_{F_q}(a^\star; \tilde{a})\|_\infty \le \frac{\mu c_{\max}}{2\gamma}\|a^\star - \tilde{a}\|_\infty^2.$$

If $a^\star$ takes values in $[0, 1]$, choosing $\tilde{a} = \pi$ recovers $\|V^\star - V^\pi\|_\infty \le (\mu c_{\max}/(2\gamma))\|\pi^\star - \pi\|_\infty^2$.

Proof. The first inequality follows from Lemma B.8. The derivative $F_q'(v) = q(P(v))$ is $c_{\max}$-Lipschitz because q is $c_{\max}$-Lipschitz on $[0, 1]$ and P is 1-Lipschitz. Integrating this derivative bound gives $D_{F_q}(x; y) \le c_{\max}(x - y)^2/2$, including when effective values lie outside the bid interval. □

We now prove the following result:

Lemma E.2. *Define $a_N(t) = \hat{k}_N(t) + \hat{V}_N(0) - \hat{V}_N(t)$ and $a^\star(t) = k(t) + V^\star(0) - V^\star(t)$. Under the estimation bounds above, with $L = 1 + (2\mu/\gamma)(1 + k(\infty))$,*

$$\|a_N - a^\star\|_\infty \le L\epsilon_N, \qquad \|\pi_N - \pi^\star\|_\infty \le L\epsilon_N.$$

Proof. The Bellman contraction in Lemma B.5, using competition accuracy only on $[0, \beta(\pi^\star)]$, gives

$$\|\hat{V}_N - V^\star\|_\infty \le \frac{\mu}{\gamma}(1 + k(\infty))\epsilon_N.$$

Taking absolute values in the effective-bid identity gives $\|a_N - a^\star\|_\infty \le \|\hat{k}_N - k\|_\infty + 2\|\hat{V}_N - V^\star\|_\infty$. Projection onto $[0, 1]$ is 1-Lipschitz, proving the policy bound. Lemma E.1, applied with $\tilde{a} = a_N$, also gives $\|V^\star - V^{\pi_N}\|_\infty = \mathcal{O}(\epsilon_N^2)$. □

We now have all the ingredients needed to analyze our algorithms. We work in the fixed-horizon setting.

E.1 Two-Phase Algorithm: Learn then Rollout

Theorem E.3. *Consider the two-phase learning algorithm:*

1. ***Exploration Phase*** *(N_1* ***episodes):*** *Use a fixed exploratory policy π_x that bids $\min\{1, B\}$ to collect data for estimating k and q. Compute $\hat{k} = K(\mathcal{D}^k_{N_1})$ and $\hat{q} = Q(\mathcal{D}^q_{N_1})$.*
2. ***Exploitation Phase*** *($N_2 = N - N_1$* ***episodes):*** *Use policy $\pi = \mathcal{S}(\hat{k}, \hat{q})$.*

The algorithm explores for N_1 episodes and exploits for the remaining N_2 episodes.
Let $\delta \in (0, 1)$. There exists $C_0 \geq 9$ such that once $N \geq C_0$, setting $N_1 = \lceil \sqrt{N} \rceil$ yields with probability at least $1 - \delta$:

$$\textit{Regret}(N) \leq \mathcal{O}(\sqrt{N}). \tag{4}$$

Proof. If $B = 0$, then $k \equiv 0$, exploration produces no wins, and the singular-design default gives $\hat{k} \equiv 0$. The committed policy also bids zero, so regret is zero. For $B > 0$, the argument below also covers $k \equiv 0$.

Let $N_1 = \lceil \sqrt{N} \rceil$ and $C_0 = \max\{9, (N_0^k)^2, (N_0^q)^2\}$, where N_0^k and N_0^q are the thresholds in Propositions C.5 and D.6, each evaluated at failure probability $\delta/2$. Since $N \geq C_0 \geq 9$, the remaining phase length $N_2 = N - N_1$ is nonnegative.

In the exploration phase, $\pi_{\mathtt{x}} = \min\{1, B\} > 0$ satisfies Condition C.1. Since $\beta(\pi^\star) \leq \min\{1, k(\infty)\} \leq \min\{1, B\}$, it also satisfies Condition D.1.

If $N \geq C_0$, then $N_1 \geq \max\{N_0^k, N_0^q\}$ and we can use both Proposition C.5 and Proposition D.6.

These ensure that, with probability at least $1 - \delta$, we have:

$$\|k - \hat{k}\|_\infty \leq \mathcal{O}(1/\sqrt{N_1}), \qquad \sup_{v \in [0, \beta(\pi^\star)]} |q(v) - \hat{q}(v)| \leq \mathcal{O}(1/\sqrt{N_1})$$

Lemma E.2 gives $\|a^\star - \hat{a}\|_\infty = \mathcal{O}(N_1^{-1/2})$ and the same bound for the policies. Applying Lemma E.1 with $\tilde{a} = \hat{a}$ yields $\|V^\star - V^{\hat{\pi}}\|_\infty \leq \mathcal{O}(1/N_1)$ (Lemma E.1) and a total regret of:

$$\text{Regret}(N) \leq \underbrace{\mathcal{O}(N_1)}_{\text{exploration}} + \underbrace{\mathcal{O}(N_2/N_1)}_{\text{exploitation}} \tag{5}$$

$$\leq \mathcal{O}(\sqrt{N}). \tag{6}$$

□

E.2 Three-Phase Algorithm: Learn k then q then Rollout

Theorem E.4. *Consider the three-phase learning algorithm:*

1. ***Exploration Phase for*** *k (N_1* ***episodes):*** *Use a fixed exploratory policy $\pi_{\mathtt{x}}$ that bids $b_0 > 0$ to collect data for estimating k. Compute $\hat{k} = K(\mathcal{D}^k_{N_1})$.*
2. ***Exploration Phase for*** *q (N_2* ***episodes):*** *Use the learned policy $\pi_k = \hat{k}$ to collect data for estimating q. Compute $\hat{q} = Q(\mathcal{D}^q_{N_2})$.*
3. ***Exploitation Phase*** *($N_3 = N - N_1 - N_2$* ***episodes):*** *Use policy $\pi = \mathcal{S}(\hat{k}, \hat{q})$.*

The algorithm learns $\hat{k}$, uses $\hat{k}$ to learn $\hat{q}$, and then exploits with $\pi = \mathcal{S}(\hat{k}, \hat{q})$ for the remaining episodes.
Let $\delta \in (0, 1)$. There exists $C_0 \geq 9$ such that once $N \geq C_0$, setting $N_1 = N_2 = \lceil \sqrt{N} \rceil$ yields with probability at least $1 - \delta$:

$$\textit{Regret}(N) \leq \mathcal{O}(\sqrt{N}).$$

Proof. Let $N_1 = N_2 = \lceil \sqrt{N} \rceil$ and initially take $C_0 = \max\{9, (N_0^k)^2\}$, where N_0^k is evaluated at failure probability $\delta/2$. In the nonconstant case below, also include $(N_0^q)^2$ and $(C_k/\Delta_k)^4$ in C_0, with N_0^q evaluated at $\delta/2$. The constant case does not use these additional thresholds. Since $N \geq C_0 \geq 9$, we have $N_1 + N_2 \leq N$, so $N_3 = N - N_1 - N_2$ is nonnegative. In the first phase, we bid $b_0 > 0$ for N_1 episodes. Since $N_1 \geq N_0^k$, we have, with probability at least $1 - \delta/2$:

$$\|k - \hat{k}\|_\infty \leq C_k/\sqrt{N_1}\,,$$

with C_k a constant. Put $\varepsilon_1 = C_k/\sqrt{N_1}$. If $k \equiv c$ is constant, the optimal admissible bid is the projection of c onto $[0, 1]$. For the committed policy $\hat{\pi} = \mathcal{S}(\hat{k}, \hat{q})$, Lemma C.2, applied to $\hat{k}$ and any admissible $\hat{q}$, places its effective bid between $\hat{k}(0)$ and $\hat{k}(\tau)$. On the first-phase event both endpoints belong to $[c - \varepsilon_1, c + \varepsilon_1]$. Thus $\|\hat{a} - c\|_\infty \leq \varepsilon_1$ independently of competition estimation. Since the true effective bid is c, Lemma E.1 gives committed per-episode loss $\mathcal{O}(N_1^{-1})$ for every $c \geq 0$. Together with the two exploration costs, this gives $\mathcal{O}(N_1 + N_2 + N_3/N_1) = \mathcal{O}(\sqrt{N})$ regret. This includes $c = 0$ and requires neither knowledge of c nor a positive tail gap.

Suppose now that k is nonconstant. Lemma C.2 shows that optimal bids are strictly positive on an interval where $0 < k(\tau) < k(\infty)$. A first win occurs there with positive probability. Imitating the age-zero optimal control from age $\bar{T}$ strictly increases its reward at that win and leaves the continuation unchanged. Hence $V^\star(\bar{T}) > V^\star(0)$ and

$$\beta(\pi^\star) \leq k(\infty) + V^\star(0) - V^\star(\bar{T}) < k(\infty).$$

Define $\Delta_k = (k(\infty) - \beta(\pi^\star))/2 > 0$. If $N_1 \geq (C_k/\Delta_k)^2$, then for every $\tau \geq \bar{T}$,

$$\hat{k}(\tau) \geq k(\infty) - C_k/\sqrt{N_1} \geq k(\infty) - \Delta_k > \beta(\pi^\star).$$

Thus the second-phase value policy $\pi_k = \hat{k}$, projected onto the admissible bid interval, satisfies Condition D.1 with $T_0 = \bar{T}$. Applying Proposition D.6 to the N_2 second-phase episodes gives, conditionally on this event and with probability at least $1 - \delta/2$,

$$\sup_{v \in [0, \beta(\pi^\star)]} |q(v) - \hat{q}(v)| \leq \mathcal{O}(1/\sqrt{N_2})\,.$$

A union bound gives a joint event of probability at least $1 - \delta$. On this event, Lemma E.2 gives $\|a^\star - \hat{a}\|_\infty \leq \mathcal{O}(N_1^{-1/2} + N_2^{-1/2})$ and the same bound for the policies and $\|V^\star - V^{\hat{\pi}}\|_\infty \leq \mathcal{O}(N_1^{-1} + N_2^{-1})$ (Lemma E.1), giving total regret

$$\begin{aligned} \text{Regret}(N) &\leq \underbrace{\mathcal{O}(N_1)}_{\text{exploration}} + \underbrace{\mathcal{O}(N_2)}_{\text{exploration with } \hat{k}} + \underbrace{\mathcal{O}\big(N_3\,(N_1^{-1} + N_2^{-1})\big)}_{\text{exploitation}} && (7) \\ &\leq \mathcal{O}(\sqrt{N})\,. && (8) \end{aligned}$$

□

E.3 Confidence Bounds Algorithm

Here $\mathcal{D}_n^k$ and $\mathcal{D}_n^q$ denote the data available before episode n. The one-episode shift between n and the number of completed learning episodes is absorbed into the confidence constants.

Lemma E.5 (Time-uniform finite-LIL confidence envelopes)**.** *Assume that the estimators $K(\mathcal{D}_n^k)$ and $Q(\mathcal{D}_n^q)$ satisfy the martingale/sub-Gaussian regularity conditions used in Propositions C.5 and D.6. Then there exist problem-dependent constants $C_k, C_q, c_{\mathrm{LIL}} > 0$ and $C_0 \in \mathbb{N}$ such that, for every $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $n \geq C_0$,*

$$\left\|K(\mathcal{D}_n^k) - k\right\|_\infty \leq C_k \sqrt{\frac{1 + \log\log(n) + \log(c_{\mathrm{LIL}}/\delta)}{n}},$$

and

$$\sup_{v \in [0, \beta(\pi^\star)]} |Q(\mathcal{D}_n^q)(v) - q(v)| \leq C_q \sqrt{\frac{1 + \log\log(n) + \log(c_{\mathrm{LIL}}/\delta)}{n}}.$$

Equivalently, the confidence envelopes hold uniformly over time without taking a union bound over $n = 1, \ldots, N$.

Proof. This is a direct application of the finite-LIL confidence sequence of [24, Theorem 1 and Eq. (3.4)].

Under the martingale/sub-Gaussian regularity assumptions of Propositions C.5 and D.6, the centered estimation errors associated with $K(\mathcal{D}_n^k) - k$ and $Q(\mathcal{D}_n^q) - q$ satisfy time-uniform sub-Gaussian martingale concentration with variance proxy of order n. Applying the two-sided polynomial stitched confidence sequence of [24] and allocating failure probability $\delta/2$ to the k-estimator and $\delta/2$ to the q-estimator, yields constants $C_k, C_q, c_{\mathrm{LIL}} > 0$ such that, with probability at least $1-\delta$, simultaneously for all $n \geq C_0$,

$$\left\|K(\mathcal{D}_n^k) - k\right\|_\infty \leq C_k \sqrt{\frac{1 + \log\log(n) + \log(c_{\mathrm{LIL}}/\delta)}{n}},$$

and

$$\sup_{v \in [0, \beta(\pi^\star)]} |Q(\mathcal{D}_n^q)(v) - q(v)| \leq C_q \sqrt{\frac{1 + \log\log(n) + \log(c_{\mathrm{LIL}}/\delta)}{n}}.$$

The constants absorb the numerical constants in the stitched boundary, the two-sided symmetrization, the split of the failure probability between the two estimators, and the constants appearing in Propositions C.5 and D.6. Since the confidence sequence is already uniform over all times, no union bound over $n = 1, \ldots, N$ is needed. □

Theorem E.6 (Confidence Bounds Algorithm)**.** *Choose $\lambda_k, \lambda_q > 0$. In the first $\min\{2, N\}$ episodes, submit the fixed bid 1 and augment both datasets using the update below. For each remaining episode $n \in \{3, \ldots, N\}$ do:*

- $\hat{k}_n^{\mathit{UCB}} = \min(K(\mathcal{D}_n^k) + \lambda_k \sqrt{\log\log n / n}\,,\, B)$
- $\hat{q}_n^{\mathit{LCB}} = \max(Q(\mathcal{D}_n^q) - \lambda_q \sqrt{\log\log n / n}\,,\, 0)$
- $\pi_n = \mathcal{S}(\hat{k}_n^{\mathit{UCB}}, \hat{q}_n^{\mathit{LCB}})$
- $\mathcal{D}_{n+1}^k = \mathcal{D}_n^k \cup \mathcal{D}_{\pi_n}^k$ *and* $\mathcal{D}_{n+1}^q = \mathcal{D}_n^q \cup \mathcal{D}_{\pi_n}^q$ *with* $\mathcal{D}_{\pi_n}^k$ *and* $\mathcal{D}_{\pi_n}^q$ *the datasets generated at episode n.*

The confidence-adjusted estimates guide policy updates while ensuring coverage of the relevant competition range.
Let $\delta \in (0,1)$. There exist λ_q and λ_k (depending on problem constants and δ) such that the algorithm satisfies, with probability at least $1-\delta$:

$$\textit{Regret}(N) \leq \tilde{\mathcal{O}}(\log(N)). \tag{9}$$

Proof. If $B = 0$, then $k \equiv 0$ and all bids after the first two episodes are zero, giving $\mathcal{O}(1)$ regret. For $B > 0$, the following argument is conditional on a joint anytime event for the two-sided estimation bounds in Lemma E.5. Let $C_0 \geq 3$ be its starting index. Choose the constants large enough that, for every $3 \leq n < C_0$, the adjusted estimates equal B and 0, respectively; this is possible because $\sqrt{\log\log n/n} > 0$ on this finite set. The resulting policy bids $\min\{1, B\}$ and covers the optimal bid range. The first $C_0 - 1$ episodes contribute $\mathcal{O}(C_0)$ regret. For $n \geq C_0$, the one-sided consequences used for coverage are:

$$k(\tau) \leq K(\mathcal{D}_n^k)(\tau) + C_k\sqrt{\log\log n/n}, \qquad \tau \geq 0,$$
$$q(v) \geq Q(\mathcal{D}_n^q)(v) - C_q\sqrt{\log\log n/n}, \qquad v \in [0, \beta(\pi^\star)]\,,$$

On this joint event, taking $\lambda_k \geq C_k$ and $\lambda_q \geq C_q$ gives $\hat{k}_n^{\mathtt{UCB}} \geq k$ at every age and $\hat{q}_n^{\mathtt{LCB}} \leq q$ on $[0, \beta(\pi^\star)]$. Corollary B.6 therefore gives $\beta(\pi_n) \geq \beta(\pi^\star)$. The clipped planning recovery is constant after $\bar{T}$, so $\pi_n(t) = \beta(\pi_n)$ for $t \geq \bar{T}$, verifying Condition D.1. When $k(\infty) > 0$, Lemma C.2 and $\hat{k}_n^{\mathtt{UCB}} \geq k$ also verify Condition C.1 with an exposure constant independent of n. Conditional on the joint confidence event, the effective-bid stability and quadratic-loss bounds give $\|a^\star - a_n\|_\infty \leq \mathcal{O}(\sqrt{\log\log n/n})$ and $\|V^\star - V^{\pi_n}\|_\infty \leq \mathcal{O}(\log\log n/n)$. Summing gives

$$\mathrm{Regret}(N) \leq \mathcal{O}(C_0) + \sum_{n=C_0}^{N} \mathcal{O}(\log\log n/n) \leq \tilde{\mathcal{O}}(\log(N))\,, \tag{10}$$

The remaining probabilistic step is to establish the joint anytime confidence event under these adaptively generated policies. This requires a confidence argument valid up to the first coverage failure and a separate treatment of $k \equiv 0$ when the known bound B is positive; the learner cannot identify that case from B alone. □

F Results for general primitives

Lemma F.1 (Piecewise linear approximation)**.** *If f' is L-Lipschitz on $[a, b]$, its interpolant on m uniform intervals satisfies*

$$\|f - f_{\mathrm{PL},m}\|_\infty \leq \frac{L(b-a)^2}{8m^2}.$$

Proof. On a cell of width h, the chord error is at most $Lh^2/8$. □

F.1 Confidence-calibrated plug-in estimates

Retain the observation model, nonnegative concave nondecreasing k, and a known cutoff $T > 0$, with $k(t) = K = k(T)$ for $t \geq T$. The intercept is unrestricted. For $K > 0$, retain the standing condition $q(b) \geq \alpha b$ on $[0, \min\{K, 1\}]$. Assume known upper bounds on B, σ, and let q be Λ-Lipschitz with $q(0) = 0$. In the smooth case, let k' and q' be L_k- and L_q-Lipschitz on their respective compact domains, with known upper bounds on L_k, L_q. Use recovery nodes $0 = x_0 < \cdots < x_m = T$, cell widths h_j, and a price grid on $[0, 1]$ of maximal width h_q. Put

$$\ell = \log\frac{C(m+1)(N+1)^3}{\delta}, \qquad r(s) = \sigma\sqrt{2\ell/s}.$$
$$a_k = \frac{L_k \max_j h_j^2}{8}, \quad a_q = \frac{L_q h_q^2}{8}.$$

For exact piecewise-linear primitives on their known grids, the corresponding approximation allowances are zero. Index n denotes completed episodes.

Recovery confidence intervals. In a sole-auction winning episode, aggregate net payoff plus payment is $Y = k(t) + \varepsilon$. In the first and last quarters of each recovery cell, let $s_{j,G}, \bar{t}_{j,G}, \bar{Y}_{j,G}$ be the counts and mean ages and responses, for $G \in \{L, R\}$. Set

$$g_j(t) = \bar{Y}_{j,L} + \frac{t - \bar{t}_{j,L}}{\bar{t}_{j,R} - \bar{t}_{j,L}}(\bar{Y}_{j,R} - \bar{Y}_{j,L}),$$
$$\mathcal{R}_{j,n} = \min\{B, 2\max_G r(s_{j,G}) + 2L_k h_j^2\}.$$

Use the clipped estimates $\hat{v}_{0,n} = [g_1(0)]_{[0,B]}$ and $\hat{v}_{j,n} = [g_j(x_j)]_{[0,B]}$ for $1 \le j < m$, with $R_{0,n} = \mathcal{R}_{1,n}$ and $R_{j,n} = \mathcal{R}_{j,n}$. Missing groups give estimate zero and radius B; omit the deterministic $2L_k h_j^2$ term on exact affine cells. Sole-auction wins at ages at least T instead estimate K: with count M_n, set

$$\begin{aligned} \hat{v}_{m,n} &= [\bar{Y}_n^T]_{[0,B]}, & R_{T,n} &= \min\{B, r(M_n)\}, \\ U_n &= \min\{B, \hat{v}_{m,n} + R_{T,n}\}, & t_{T,n} &= \min\{B, 2R_{T,n}\}. \end{aligned}$$

With no tail sample, use $\hat{v}_{m,n} = 0$ and $R_{T,n} = B$. Selected-noise concentration gives simultaneous bounds

$$|\hat{v}_{j,n} - k(x_j)| \le R_{j,n} \ (j < m), \qquad 0 \le U_n - K \le t_{T,n}. \tag{11}$$

Indeed, the local fit coefficients have total absolute value at most two, its deterministic bias is at most $2L_k h_j^2$, and a count-indexed martingale bound controls each selected noise mean.

Constrained aggregate fit. Let v_e be aggregate payoff plus winning payments in episode e. Minimize the original loss $\sum_{e \le n}(v_e - \sum_{a \text{ won in } e} H(\tau_{e,a}))^2$ over concave nondecreasing grid interpolants, constant after T, subject to

$$\begin{aligned} 0 \le H \le U_n, \qquad H(T) = U_n, \\ \hat{v}_{j,n} - R_{j,n} \le H(x_j) \le \hat{v}_{j,n} + R_{j,n} + t_{T,n} \quad (j < m). \end{aligned}$$

Use the minimizer of least Euclidean nodal norm, or $H_n \equiv U_n$ if infeasible. All episodes enter the fit; the local intervals only calibrate its constraints. Define

$$\|f\|_P^2 = \int_0^T f(t)^2 \, \mathrm{d}t + f(T)^2,$$

$$\omega_0 = h_1/2, \quad \omega_j = (h_j + h_{j+1})/2 \ (0 < j < m), \quad \omega_m = 1 + h_m/2,$$

$$E_n = \left[\sum_{j<m} \omega_j (2R_{j,n} + t_{T,n})^2 + \omega_m t_{T,n}^2 \right]^{1/2} + \sqrt{T}\, a_k. \tag{12}$$

On (11), $k_{\mathrm{PL},m} + U_n - K$ is feasible. Every feasible fit has nodal error at most $2R_{j,n} + t_{T,n}$, so interpolation gives $\|H_n - k\|_P \le E_n$.

Competition estimate. Retain first auctions arriving after T and before termination. Eligibility has probability $p_T = \mu e^{-(\mu+\gamma)T}/(\mu+\gamma)$ and depends only on the clocks. Their latent prices are therefore i.i.d. with CDF q. With s observations, let b^- be their smallest historical bid. All latent prices strictly below b^- are observed. Interpolate their empirical left CDF counts on the price grid truncated at b^-, including b^- as a node, to obtain Q_n. With $a_n = \min\{1, \sqrt{\ell/(2s)}\}$, put

$$q_n^- = [Q_n - a_n - a_q]_+$$

on $[0, b^-]$, extended constantly afterward; use zero if $s = 0$ or $b^- = 0$. Uniform empirical-CDF concentration and interpolation give

$$q_n^- \le q \text{ on } [0,1], \qquad \|q - q_n^-\|_{[0,b^-]} \le 2a_n + 1/s + 2a_q. \tag{13}$$

The displayed error bound applies when $s > 0$. Left counts handle adaptive bids equal to earlier observed prices. The global lower bound does not require a coverage assumption.

Confidence correction and policy. Let

$$\begin{aligned} \rho &= \frac{\mu}{\mu+\gamma}, & c &= 1 - \rho, \quad C_V = \frac{\mu\sqrt{T} + \rho}{1-\rho}, \\ D &= 1 + 2C_V, \quad C_\beta = D/c. \end{aligned}$$

Use the same single-solver plug-in update

$$d_n = \min\{C_\beta E_n, U_n\}, \qquad k_n^+ = H_n + d_n, \qquad \pi_{n+1} = \mathcal{S}(k_n^+, q_n^-).$$

The planning curve is bounded by $2B$; the true recovery remains bounded by B. The correction targets the maximal-bid functional, rather than pointwise recovery optimism.

F.2 Regret guarantee and proof

Theorem F.2 (Piecewise-linear plug-in extension)**.** *For the uniform recovery grid $h = T/m$ and every fixed environment with $K > 0$, the rule above satisfies, with probability at least $1 - \delta$,*

$$\mathrm{Regret}(N) \le C_K \left[1 + m \log(m+1)\ell \log(N+1) + N(m^{-4} + h_q^4)\right]. \tag{14}$$

Thus $m \asymp N^{1/5}$ and $h_q \asymp N^{-1/5}$ give $\widetilde{\mathcal{O}}_K(N^{1/5})$. On fixed known piecewise-linear grids, all approximation terms vanish and the bound is $C_{K,\mathrm{grid}}[1 + \ell \log(N+1)]$. Constants may diverge as $K \downarrow 0$; this is an instance-dependent guarantee, rather than a uniform minimax assertion.

Proof. We give the three ingredients of the approximation argument.

First, at a common planning CDF r, first-win renewal and the effective-bid identity imply

$$\begin{aligned} \|V^\star_{h,r} - V^\star_{k,r}\|_\infty &\le C_V \|h - k\|_P, \\ |\beta(\mathcal{S}(h,r)) - \beta(\mathcal{S}(k,r))| &\le D \|h - k\|_P. \end{aligned} \tag{15}$$

The first bound follows by dividing $\mu \int_0^T |h-k| + \rho |h(T) - k(T)|$ by $1 - \rho$. Moreover, the effective tail bid A solves $\gamma A + \mu F_r(A) = \gamma(h(T) + V^\star_{h,r}(0))$. A constant shift d raises its right side by at least γd, and the left side is $(\gamma + \mu)$-Lipschitz. Consequently $\beta(\mathcal{S}(h+d,r)) \ge \min\{1, \beta(\mathcal{S}(h,r)) + cd\}$. If $d_n < U_n$, the chosen bonus compensates the maximal-bid error; if $d_n = U_n$, $H_n + U_n \ge K \ge k$ pointwise. Together with $q_n^- \le q$, either case gives $\beta(\pi_{n+1}) \ge \beta^\star$ and $\|k_n^+ - k\|_P \le C E_n$.

Second, $U_n \ge K$ ensures the all-age bid profile $\pi_{n+1}(t) \ge c_0 K \min\{t/T, 1\}$, where $c_0 = \min\{c, 1/(2B)\}$. A sole-auction win in either quarter of uniform cell j therefore has conditional probability at least $c_K h^2 j$. Selected-noise bounds and multiplicative count concentration yield, including small counts through the trivial radius B,

$$\begin{aligned} \mathcal{R}^2_{j,n} &\le C_K \left(\frac{\ell}{n h^2 j} + h^4 \right), & t^2_{T,n} &\le C_K \frac{\ell}{n}, \\ E_n^2 &\le C_K \left(\frac{m \log(m+1)\ell}{n} + m^{-4} \right). \end{aligned}$$

Counts are analyzed up to the first tail-confidence failure; the noise confidence event itself needs no exposure assumption. Since all historical tail bids cover $\beta^\star$, the exogenous eligible-first-auction count and (13) give $\epsilon^2_{q,n} \le C(\ell/n + a_q^2)$ on the relevant bid range. These events hold jointly by a union bound with the displayed ℓ.

Finally, the Bellman comparison in the weighted norm and the true CDF's Lipschitz property give the quadratic policy loss

$$V^\star(0) - V^{\mathcal{S}(h,r)}(0) \le C \left(\|h - k\|_P^2 + \|r - q\|^2_{[0,\beta^\star]} \right). \tag{16}$$

Indeed, the first-auction residual uses CDF error only up to the true optimal bid. Squared effective-bid error controls the surplus Bregman loss, and the discounted occupation of capped age obeys $\mathbb{E} \int e^{-\gamma s} g(\min\{A_s, T\})\, ds \le (1 + \mu/\gamma)(\int_0^T g + g(T)/\gamma)$ for $g \ge 0$. Summing the preceding bounds proves (14). On any fixed exact grid, each group has a fixed positive sampling probability and the same calculation gives the stated logarithmic factors. □

Zero recovery. If $K = 0$, then $k \equiv 0$. The same algorithm still satisfies $\widetilde{\mathcal{O}}(N^{1/3})$ regret. To see this, let X_{n+1} indicate a sole-auction tail win and p_{n+1} its conditional probability. Then $p_{n+1} = \eta q(\beta(\pi_{n+1}))$, where $\eta = \mu\gamma e^{-(\mu+\gamma)T}/(\mu+\gamma)^2$. On the tail confidence event, $U_n \le C_0/\sqrt{M_n \vee 1}$ with $C_0 = B + 2\sigma\sqrt{2\ell}$, and every bid is at most $2U_n$. The episode regret therefore satisfies $R_{n+1} \le C \min\{U_n^2, U_n p_{n+1}\}$. For a fixed $\varepsilon > 0$, a bounded weighted-Bernoulli supermartingale gives, with high probability,

$$\begin{aligned} \sum_{U_n > \varepsilon} U_n p_{n+1} &\le 2 \sum_{U_n > \varepsilon} U_n X_{n+1} + 2B \log(C/\delta) \\ &\le C \left(C_0 + C_0^2/\varepsilon + B \log(C/\delta) \right). \end{aligned}$$

Each success increments M_n, which proves the second inequality. The remaining episodes cost at most $CN\varepsilon^2$. Taking $\varepsilon = N^{-1/3}$ proves the claim without a lower-tail condition on q. If $B = 0$, simply bid zero.

Other regularity classes. The same argument applies whenever a known regularity bound L gives local bias and chord error $a_{k,m} = O(m^{-s})$: replace the local deterministic allowance by $3Lh^s$ and the chord allowance by Lh^s. It yields

$$\mathrm{Regret}(N) \le \widetilde{\mathcal{O}}_K(m + Nm^{-2s} + Na_q^2). \tag{17}$$

Taking $a_q = O(m^{-s})$ and $m \asymp N^{1/(2s+1)}$ gives $\widetilde{\mathcal{O}}_K(N^{1/(2s+1)})$: bounded slopes, including unknown recovery knots, have $s = 1$; a Hölder-α derivative has $s = 1 + \alpha$. A fine price grid suffices even when only a known Lipschitz bound for q is available, using $a_q = \Lambda h_q/2$ and $h_q = O(m^{-s})$. The empirical-CDF statistical term is independent of the number of price-grid cells.

G Analysis of the lower bound

Consider the subclass of constant recovery functions. In doing so, the problem reduces to the static case, with the important distinction that feedback arrives in batches and is aggregated. Since allowing the algorithm to disaggregate the data freely can only improve its performance, any lower bound derived under this relaxation remains valid for the original problem. This reduction maps our setting to one of the scenarios analyzed in [8], for which a lower bound has already been established. More precisely, we can apply Theorem 4 of that reference with $\alpha = 1$ to obtain the following result.

Theorem G.1. *For each fixed $B > 0$, there exists $C > 0$ such that, for every learning algorithm $\mathcal{A}$ and every $N \geq 2$, there is an admissible environment in which*

$$\mathbb{E}[\mathrm{Regret}_{\mathcal{A}}(N)] \geq C \log N.$$

The expectation is over the observations and any randomization of the algorithm.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims made in the abstract are supported by results in the paper, from studying the learning problem, algorithms and regret guarantees.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Last section enumerates limitations

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All statements are formally proven

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental section contains all the information needed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No],

Justification: The code is not shared with the submission.

Guidelines

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes],

   Justification: The experimental section contains all the details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The simulation is only an illustration of the theoretical results obtained

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The experiment is small enough to be run on a laptop

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer:[Yes]

   Justification: We followed the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is theoretical work that does not pose any direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: the paper poses no such risks

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: the paper does not use existing assets

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[N/A].

    Justification: the paper does not release new assets

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer:[N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.