Contextual Dynamic Pricing with Heterogeneous Buyers

Thodoris Lykouris
MIT
lykouris@mit.edu

Sloan Nietert
EPFL
sloan.nietert@epfl.ch

Princewill Okoroafor
Harvard University
pco9@cornell.edu

Chara Podimata
MIT
podimata@mit.edu

Julian Zimmert
Google
zimmert@google.com

Abstract

We initiate the study of contextual dynamic pricing with a heterogeneous population of buyers, where a seller repeatedly (over T rounds) posts prices that depend on the observable d dimensional context and receives binary purchase feedback. Unlike prior work assuming homogeneous buyer types, in our setting the buyer's valuation type is drawn from an unknown distribution with finite support $K_\star$. We develop a contextual pricing algorithm based on Optimistic Posterior Sampling with regret $\tilde{O}(K_\star\sqrt{dT})$, which we prove to be tight in d and T up to logarithmic terms. Finally, we refine our analysis for the non-contextual pricing case, proposing a variance-aware Zooming algorithm that achieves the optimal dependence on $K_\star$.[1]

1 Introduction

In online learning for contextual pricing, a learner (aka seller) repeatedly sets prices for different products with the goal of maximizing revenue through interactions with agents (aka buyers or customers). Concretely, in each round $t = 1, \dots, T$, nature selects a product with a d-dimensional feature representation u_t (context) and the seller selects a price $p_t \geq 0$. In the simplest variant, the *linear valuation model*, customers have a fixed intrinsic valuation model (type) that is unknown to the learner; this has a d-dimensional representation $\theta^\star$ whose coordinates reflect the valuation that each product feature adds, i.e., the customer's valuation is $v_t = \langle \theta^\star, u_t \rangle + \varepsilon_t$ where ε_t is a noise term. The customer makes a purchase only when their valuation is higher than the price, i.e., $v_t \geq p_t$. The learner's goal is to maximize revenue, i.e., the sum of the prices in rounds when purchases occur. An equivalent objective is to minimize *regret*, which is measured against a benchmark that always selects the customer's valuation as the price for the given round.

One key difficulty is that the learner faces both an infinite action space (i.e., all possible prices) and a discontinuous revenue function, hence causing sharp revenue loss for the learner. However, the problem offers a richer feedback structure than classical multi-armed bandits: a non-purchase indicates that all higher prices would also be rejected by the buyer, while a purchase confirms that all lower prices would be accepted too. The two primary approaches from the literature to tackle this problem involve estimating the unknown parameter $\theta^\star$ through online regression or multi-dimensional binary search (see Section 1.1 for further discussion).

A crucial limitation for both approaches is that they require all customers to behave *homogeneously* according to a single type $\theta^\star$ (see related work for results robust to small deviations from this assumption). Moving beyond this homogeneity assumption, we pose the following question:

[1] The full version of this paper will appear at the NeurIPS'25 main conference track.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

How can one design contextual pricing algorithms with a heterogeneous population of customers?

1.1 Our contribution

Our setting. To study contextual pricing with a heterogeneous buyer population, we assume that the type θ_t in round t is drawn from a fixed, unknown distribution $D_\star$. When $D_\star$ is supported on a single type $\theta_\star$, we recover the homogeneous setting. In our model, the number of distinct buyer types $K_\star = |\operatorname{supp} D_\star| (> 1)$ reflects the *degree of heterogeneity*. There are several obstacles to applying existing algorithms from the literature. First, canonical contextual pricing algorithms based on linear regression either compete against (simple) linear policies or assume context-independent and identically distributed (i.i.d.) valuation noise. In contrast, the optimal policy in our setting may best respond based on a *context-dependent* type rather than a *fixed* type, and the stochasticity due to heterogeneity is inherently context-dependent and thus non-i.i.d. Second, given that *the buyer types are not observable*, one cannot connect the observed feedback to shrinkage of type-dependent uncertainty sets; this rules out running canonical multi-dimensional binary search / contextual pricing algorithms for each buyer type in parallel. Third, since in our setting there is a *continuum of actions*, any canonical contextual bandits algorithm whose regret scales with the discretized action count (e.g., EXP4) will suffer suboptimal performance.

Our contextual pricing algorithm. To tackle these challenges, we employ recent advances in the contextual bandit literature that attain a better scaling with the number of actions, thus evading the shortcomings of EXP4 with naïve discretization. In particular, we build on the *optimistic posterior sampling* (OPS) approach [17] which, in our setting, maintains a posterior μ_t over all candidate type distributions. We call these candidate type distributions *models* and refer to their (possibly infinite) family as $\mathcal{D}$. At a high level, in every round, OPS best responds to a model sampled from μ_t. As typical in online learning, the posterior update penalizes models that *disagree* with the observed feedback (*model mismatch*) aiming to converge to the model $D_\star$. To encourage exploration in the absence of full information, this penalty is reduced by an *optimism bias* term that rewards models with the highest potential to positively contribute to the revenue. The OPS approach enables regret bounds of $\sqrt{T \cdot c \cdot \log |\mathcal{D}|}$, scaling with a *disagreement coefficient* c that measures the structural complexity of the reward functions and captures the tension between exploration and exploitation. Note that c can be significantly smaller than the number of actions.

Our main technical contributions in adapting OPS to heterogeneous contextual pricing are twofold. First, to bound the disagreement coefficient c, we observe that, for any fixed context, the aggregate demand function induced by $D_\star$ has at most $K_\star$ "jumps",[2] thus creating $K_\star + 1$ intervals. Over each interval, we bound the disagreement coefficient by a factor of 2. Combining these arguments with a novel decomposition lemma for the disagreement coefficient of functions with $K_\star$ breakpoints, we show that $c \leq 2(K_\star + 1)$. When $K_\star$ is known, we apply a variant of OPS over a finite covering of the class $\mathcal{D}$ containing *all* possible distributions over $K_\star$ types, of log cardinality $dK_\star \log T$. Second, to extend our sublinear regret guarantee to the infinite model class $\mathcal{D}$, we modify OPS to conservatively perturb its recommended prices (which cannot overly impact regret due to one-sided Lipschitzness of the revenue function). We then construct a coupling between the actual trajectory of OPS and one where $D_\star$ belongs to the finite cover, allowing us to transfer regret bounds. Finally, we adapt to unknown $K_\star$ by initializing OPS with a non-uniform prior over models. These technical contributions enable us to show a regret guarantee of $\widetilde{O}(K_\star \sqrt{dT})$. Finally, we show that this guarantee is optimal (up to logarithmic terms) with respect to the dependence on both the contextual dimension d and the time horizon T, establishing a lower bound of $\Omega(\sqrt{K_\star d T})$ for sufficiently large $T = \Omega(dK_\star^3)$.

Non-contextual improvements. The above upper and lower bounds raise a natural question on the optimal dependence on the number of buyer types $K_\star$; we resolve this question in the non-contextual version of the problem ($d = 1$) by providing an algorithm with an upper bound of $\widetilde{O}(\sqrt{K_\star T})$. Our algorithm, `ZoomV`, combines zooming (i.e., adaptive discretization) methods from Lipschitz bandits [8] with variance-aware confidence intervals [1]. We show that the regret of `ZoomV` scales with a novel variance-aware zooming dimension that can be significantly smaller than the standard measure of complexity for Lipschitz bandits. For pricing, this variance adaptation unlocks our $\widetilde{O}(\min\{\sqrt{K_\star T}, T^{2/3}\})$ bound (versus $O(T^{2/3})$, obtained via the standard zooming analysis).

[2] Each "jump" corresponds to a change of type from (say) type i to type $i + 1$.

Related work. The closest line of work to ours is *contextual pricing/search*, where a learner interacts with nature to learn a hidden vector $\theta^\star \in \mathbb{R}^d$ while receiving single-bit feedback [3, 12, 14, 11, 7, 6, 4, 13]. Under an appropriate pricing loss, this setting reduces to ours with a *homogeneous* buyer population, i.e., $K_\star = 1$. However, the deterministic nature of the feedback leads to aggressive search policies via cutting planes that do not lend themselves to the heterogeneous case. Although some works tolerate i.i.d., context-independent valuation noise [7, 6, 4, 13], their methods do not treat our non-i.i.d., context-dependent noise due to heterogeneity. Next is the case of *non-contextual dynamic pricing* where $d = 1$, originally treated with multi-armed bandits methods by [9]. The closest non-contextual work is [2], whose finite-types model reduces to ours with $d = 1$. We improve upon their regret bound, but neither our improvement nor existing methods generalize readily to the contextual setting. Finally, our setting relates to *Lipschitz bandits*. Although revenue is *not* fully Lipschitz, it satisfies a *one-sided Lipschitzness*, enabling the use of zooming [8] when $d = 1$ (see, e.g., [15]). We successfully refine these for the non-contextual case, but contextual variants of zooming [16, 10] scale with complexity parameters which admit no direct bounds for heterogeneous pricing.

2 Setup and Preliminaries

Notation. Let $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product on $\mathbb{R}^d$. Let $\mathbb{S}^{d-1}, \mathbb{B}^d \subseteq \mathbb{R}^d$ denote the unit sphere and ball, respectively. Let $\Delta(S)$ denote the set of all probability measures on a measurable set $S \subseteq \mathbb{R}^d$, and let $\operatorname{supp}(D)$ denote the support of $D \in \Delta(\mathbb{R}^d)$. We use $\Delta_k(S)$ for those $D \in \Delta(S)$ with $|\operatorname{supp}(D)| \leq k$. For a positive integer m, let $[m] := \{1, 2, \ldots, m\}$.

Setup. We consider T rounds of repeated interaction between a seller, a population of buyers, and an adversary. At each round $t \in [T]$, the seller posts a price $p_t \in [0, 1]$ for an item to be sold and a buyer, sampled from the population, decides whether or not to buy the item based on their valuation $v_t \in [0, 1]$. We denote the indicator of their purchase by $y_t = \mathbb{1}\{v_t \geq p_t\}$. The valuation of the buyer is determined by two factors: their *type* θ_t (encoding their intrinsic preferences), and an external *context* u_t, which describes the current item to be sold and any relevant environmental factors. The learner does *not* know θ_t, but they *do* know u_t. We use a linear valuation model: i.e., θ_t and u_t lie in d-dimensional spaces $\Theta \subseteq [0, 1]^d$ and $\mathcal{U} \subseteq \mathbb{S}^{d-1}$, respectively, and take $v_t = \langle\theta_t, u_t\rangle$. We assume that $\langle\theta, u\rangle \in [0, 1]$ for all $\theta \in \Theta$ and $u \in \mathcal{U}$. We impose no further assumptions on the contexts, allowing them to be generated (potentially adaptively) by the adversary. On the other hand, we assume that each θ_t is sampled independently from a fixed distribution $D_\star \in \Delta(\Theta)$ that describes the buyer population, unknown to the seller. All together, the following occur at each round $t \in [T]$:

1. the adversary selects a context $u_t \in \mathcal{U}$;
2. a buyer arrives with type $\theta_t \in \Theta$ sampled independently from $D_\star$, with valuation $v_t = \langle u_t, \theta_t\rangle$;
3. the seller observes u_t and posts price $p_t \in [0, 1]$ for the item;
4. the seller observes the purchase decision $y_t = \mathbb{1}\{v_t \geq p_t\}$ and receives revenue $p_t y_t$.

Benchmark. The seller's goal is to maximize their cumulative revenue compared to that which they could have achieved with knowledge of $D_\star$. To express this concisely, we introduce some additional notation. Each distribution Q over valuations in $[0, 1]$ induces the following:

- a demand function $\mathsf{dem}_Q(p) := \mathbb{P}_{v\sim Q}[v \geq p]$,
- an expected revenue function $\mathsf{rev}_Q(p) := p \cdot \mathsf{dem}_Q(p)$,
- a revenue-maximizing best response $\mathsf{br}_Q := \arg\max_{p\in[0,1]} \mathsf{rev}_Q(p)$ (breaking ties arbitrarily).

Note that each type distribution $D \in \Delta(\Theta)$ and context $u \in \mathcal{U}$ induce a value distribution $Q = \mathsf{proj}(D, u)$, where $\mathsf{proj}(D, u)$ is defined as the probability law of $\langle\theta, u\rangle$ when $\theta \sim D$. We then set $\mathsf{dem}_D(p, u) := \mathsf{dem}_Q(p)$, $\mathsf{rev}_D(p, u) := \mathsf{rev}_Q(p)$, and $\mathsf{br}_D(u) := \mathsf{br}_Q$, accordingly. We abbreviate a subscript of $D_\star$ by "$\star$" alone, writing, e.g., $\mathsf{dem}_\star(p, u)$ and $\mathsf{br}_\star(u)$.

A seller policy $\mathcal{A}$ is a (potentially randomized) map from a history $\{u_\tau, p_\tau, y_\tau\}_{\tau=1}^{t-1}$ and the current context u_t to a posted price p_t. An adversary policy $\mathcal{B}$ is a (potentially randomized) map from a history $\{u_\tau, \theta_\tau, p_\tau, y_\tau\}_{\tau\in[t-1]}$ to the next context u_t. We then define the seller's *pricing regret* by $R_{\mathcal{A},\mathcal{B}}(T) := \sum_{t\in[T]} \big(\mathsf{rev}_\star(\mathsf{br}_\star(u_t), u_t) - \mathsf{rev}_\star(p_t, u_t)\big)$, where $\{u_t, p_t\}_{t\in[T]}$ are selected according to $\mathcal{A}$ and $\mathcal{B}$. We will omit the policies from the subscript when clear from context. We seek seller policies which control the pricing regret in expectation, for all adversaries $\mathcal{B}$.

Our guarantees will scale with context dimension d and the *degree of heterogeneity*, which we quantify via the support size $K_\star := |\operatorname{supp}(D_\star)|$, *unknown* to the seller.

3 Main Results

We first sketch our main result for the contextual setting. We employ a perturbed variant of optimistic posterior sampling (OPS), originally studied for contextual bandits under the name "Feel-Good Thompson Sampling" [17]. As outlined in Section 1, our analysis controls a *disagreement coefficient* for the class of relevant demand functions, along with a metric entropy bound for the class of candidate type distributions, under the suitable Levy metric. In broad strokes, OPS allows us to improve the loose $\sqrt{\#\text{ actions} \cdot T \cdot \log \#\text{ policies}}$ bound from EXP4 to $\sqrt{K_\star \cdot T \cdot K_\star d}$ (up to log factors). Vanilla OPS can achieve this bound if $K_\star$ is known and $D_\star$ belongs to a large but finite cover. To remove these restrictions, we introduce a non-uniform prior over type distributions, favoring those with small support size, and apply a small perturbation to the price recommended by OPS.

Theorem 3.1. *A perturbed variant of OPS (*`POPS`*) achieves expected regret $\widetilde{O}(K_\star\sqrt{dT})$, without prior knowledge of $K_\star$. Moreover, even for known $K_\star \geq 2$ and stochastic contexts, no contextual pricing policy can achieve expected regret $o(\sqrt{K_\star dT})$ for all instances if $T \geq d \cdot K_\star^3$.*

For the lower bound, we modify a construction of [2] for the non-contextual case so that it can be tensored into d-dimensions without leaking information between orthogonal contexts.

Theorem 3.1 leaves a key open question: what is the optimal regret dependence on $K_\star$? We resolve this question for the non-contextual setting, where $d = 1$ and, without loss of generality, $u_t \equiv 1$ for all t. Our algorithm, `ZoomV`, builds upon adaptive discretization methods for Lipschitz bandits (aka zooming [8]) with two key adjustments. First, since revenue is only one-sided Lipschitz, we use a dyadic price selection rule inspired by [15]. Second, our confidence intervals incorporate empirical variance, a method previously used for variance-aware K-armed bandits [1].

Theorem 3.2. `ZoomV` *achieves minimax-optimal expected regret $\widetilde{O}\big(\min\big\{\sqrt{K_\star T}, T^{2/3}\big\}\big)$ for non-contextual pricing with heterogeneous buyers, without knowledge of $K_\star$.*

The performance of standard zooming is controlled by the so-called *zooming dimension*, which can equal 1 for worst-case non-contextual pricing instances. We introduce a smaller, variance-aware dimension which is always 0, leading to our tight bound. This setting was previously considered by [2], but they are only able to achieve our bound when the locations of the $K_\star$ types are well-separated.

Although `ZoomV` can be implemented efficiently, `POPS` requires running time exponential in d and $K_\star$. In the full version of this paper (appearing in the main conference track), we show how to remedy this under the additional assumption that the buyer type (either a unique identifier or its full coordinates) is revealed to the learner at the end of each round.

4 Discussion

In this work, we have introduced contextual dynamic pricing with heterogeneous buyers. Our main algorithm achieves a regret bound of $\tilde{O}(K_\star\sqrt{dT})$, which is optimal up to $\sqrt{K_\star}$. Our analysis bounds the disagreement coefficient by leveraging a novel decomposition lemma for aggregate demand functions with $K_\star$ breakpoints. Additionally, we propose a variance-aware zooming algorithm for the non-contextual pricing case and improve regret dependence on $K_\star$ by incorporating adaptive discretization methods from the Lipschitz bandits literature.

There are several research questions that stem from our work. The first question revolves around *computation*. Our algorithms are dependent on the size of the model class, which is exponential in $K_\star$ and d. It would be interesting to know if polynomial dependence can be achieved at the same regret rate. The second question is around the optimal dependence on $K_\star$ for the general, contextual case. While we showed how to optimize the dependence of our bounds on $K_\star$ for the non-contextual setting, it is unclear how to scale this approach for the contextual version of the problem. One starting point could be the results on Zooming techniques for contextual bandits (see e.g., [16]). Finally, it would be interesting to see if our results can be applied to broader settings where a learner tries to learn from heterogeneous agents while receiving only binary feedback: for example, it is unknown if the approach presented in this work generalizes to general contextual search settings (i.e., with ε-ball or symmetric loss) or if it generalizes for settings that share some core properties with pricing, but differ in the fundamental techniques used to address them (see e.g., [5]).

References

[1] J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[2] N. Cesa-Bianchi, T. Cesari, and V. Perchet. Dynamic pricing with finitely many unknown valuations. In A. Garivier and S. Kale, editors, *Algorithmic Learning Theory (ALT)*, 2019.

[3] M. C. Cohen, I. Lobel, and R. Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

[4] J. Fan, Y. Guo, and M. Yu. Policy optimization using semiparametric models for dynamic pricing. *Journal of the American Statistical Association*, 119(545):552–564, 2024.

[5] C.-J. Ho, A. Slivkins, and J. W. Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 359–376, 2014.

[6] A. Javanmard. Perishability of data: dynamic pricing under varying-coefficient models. *Journal of Machine Learning Research*, 18(53):1–31, 2017.

[7] A. Javanmard and H. Nazerzadeh. Dynamic pricing in high-dimensions. *Journal of Machine Learning Research*, 20(9):1–49, 2019.

[8] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, page 681–690, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580470. doi: 10.1145/1374376.1374475. URL https://doi.org/10.1145/1374376.1374475.

[9] R. D. Kleinberg and F. T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Symposium on Foundations of Computer Science (FOCS)*, pages 594–605, 2003.

[10] A. Krishnamurthy, J. Langford, A. Slivkins, and C. Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *Journal of Machine Learning Research*, 21(137): 1–45, 2020.

[11] A. Liu, R. P. Leme, and J. Schneider. Optimal contextual pricing and extensions. In *Symposium on Discrete Algorithms (SODA)*, 2021.

[12] I. Lobel, R. Paes Leme, and A. Vladu. Multidimensional binary search for contextual decision-making. *Operations Research*, 66(5):1346–1361, 2018.

[13] Y. Luo, W. W. Sun, and Y. Liu. Distribution-free contextual dynamic pricing. *Mathematics of Operations Research*, 49(1):599–618, 2024.

[14] R. Paes Leme and J. Schneider. Contextual search via intrinsic volumes. *SIAM Journal on Computing*, 51(4):1096–1125, 2022.

[15] C. Podimata and A. Slivkins. Adaptive discretization for adversarial lipschitz bandits. In *Conference on Learning Theory*, pages 3788–3805. PMLR, 2021.

[16] A. Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 679–702. JMLR Workshop and Conference Proceedings, 2011.

[17] T. Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.