Online Price Competition under Generalized Linear Demands

Daniele Bracale University of Michigan
Moulinath Banerjee University of Michigan
Cong Shi University of Miami
Yuekai Sun University of Michigan

Abstract

We study sequential price competition among N sellers, each influenced by the pricing decisions of their rivals. Specifically, the demand function for each seller i follows the single index model $\lambda_i(\mathbf{p}) = \mu_i(\langle \boldsymbol{\theta}_{i,0}, \mathbf{p}\rangle)$, with known increasing link μ_i and unknown parameter $\boldsymbol{\theta}_{i,0}$, where the vector $\mathbf{p}$ denotes the vector of prices offered by all the sellers simultaneously at a given instant. Each seller observes only their own realized demand – unobservable to competitors – and the prices set by rivals. Our framework generalizes existing approaches that focus solely on linear demand models. We propose a novel decentralized policy, PML-GLUCB, that combines penalized MLE with an upper-confidence pricing rule, removing the need for coordinated exploration phases across sellers – which is integral to previous linear models – and accommodating both binary and real-valued demand observations. Relative to a dynamic benchmark policy, each seller achieves $O(N^2\sqrt{T}\log(T))$ regret, which essentially matches the optimal rate known in the linear setting. A significant technical contribution of our work is the development of a variant of the elliptical potential lemma – typically applied in single-agent systems – adapted to our competitive multi-agent environment.

1 Introduction

Pricing is the strategic engine of competitive markets: firms revise prices not only to track shifting demand, but to anticipate and counter rivals' moves. Algorithmic pricing amplifies this feedback, placing learning at the center of price formation. Yet much of the learning-to-price literature still leans on monopolistic or weakly interactive assumptions (Fan et al., 2024; Javanmard and Nazerzadeh, 2019; Golrezaei et al., 2019; Bracale et al., 2025b), effectively treating demand as if competitors were irrelevant. Recent works have designed policies in multiple seller interaction scenarios, where each seller is influenced by their competitors' strategies. These works propose sequential pricing mechanisms that achieve sublinear regret[1] and converge to a Nash equilibrium (NE) (Li et al., 2024; Bracale et al., 2025a). Yet two issues remain for practice: the demand is often forced to be linear (or narrowly parameterized) (Li et al., 2024; Goyal et al., 2023), and many methods separate "explore" from "optimize," with phase lengths that depend on unknown constants (Li et al., 2024; Bracale et al., 2025a) and i.i.d. pricing policy in the exploration phase.

> *"A natural question arises: can a seller update individual parameter estimators in each period and subsequently compute an optimal price to offer based on their current estimates? [...] Literature in both statistics and operations research has demonstrated that such a myopic pricing policy does not necessarily lead to convergence to the Nash equilibrium or individual optimal prices."* (Li et al., 2024)

In this paper, we address this question within a more flexible and realistic model of competition. Specifically, we adopt a generalized linear (single-index) demand framework to capture the broader relevance of nonlinear demand patterns observed in real-world markets (Gallego et al., 2006; Wan et al., 2022, 2023). Building on this model, we propose a myopic pricing-while-estimating policy that operates without a dedicated i.i.d. pricing design exploration phase. Despite the absence of such a phase and, even if the resulting price trajectory does not converge to a Nash equilibrium (NE), we prove that the individual regret still scales as $\widetilde{O}(\sqrt{T})$ against a dynamic benchmark, where $\widetilde{O}$

[1] In this paper, we use the terms *sublinear regret* and *no regret* interchangeably to indicate that the cumulative regret grows sub-linearly with time, that is, $\mathrm{Reg}(T) = o(T)$ as $T \to \infty$.

Preliminary work. Under review.

suppresses logarithmic factors.

It is important to emphasize that in this work, we do assume that a NE *does* exist; however, the convergence of our algorithm to this equilibrium is not a prerequisite for achieving the sublinear regret guarantee established in this work.

1.1 Related Literature

Demand learning in competitive environments. Over the past two decades, there has been substantial progress in demand learning under competition. Early work by Kirman (1975) analyzed symmetric two-seller competition with linear demand. Asymmetry and noise were incorporated by Cooper et al. (2015), who let each seller learn independently (effectively as a monopolist) and established equilibrium under known price sensitivities. In contrast, Li et al. (2024) studied dynamic pricing with *unknown* price sensitivities under a linear demand model and achieved the optimal $\widetilde{O}(\sqrt{T})$ regret rate. In their setup, firm i's expected demand is

$$\lambda_i(\mathbf{p}) = \alpha_{i,0} - \beta_{i,0}p_i + \textstyle\sum_{j\in[N]\setminus\{i\}} \gamma_{ij,0}p_j,$$

where $[N] = \{1, 2, \ldots, N\}$ indexes the sellers and the parameters $(\alpha_{i,0}, \beta_{i,0}, \gamma_{ij,0})$ are unknown with $\beta_{i,0} > 0$. The condition $\beta_{i,0} > 0$ captures the standard economic assumption that the demand of seller i decreases as their own price p_i increases. However, nonlinear demand models are often more realistic and have been explored in several applications (Gallego et al., 2006; Wan et al., 2022, 2023). For instance, Gallego et al. (2006) considered the mean demand function $\lambda_i(\mathbf{p}) = a_i(p_i)/(\sum_j a_j(p_j) + \kappa)$, for some *known* increasing functions a_i and real value $\kappa \in (0, 1]$. They establish the existence of a NE, but they do not propose an online learning procedure.

In contrast to prior studies (Li et al., 2024; Kachani et al., 2007; Gallego et al., 2006), we analyze a more general *monotone single-index* expected-demand model:

$$\lambda_i(\mathbf{p}) = \mu_i(-\beta_{i,0}p_i + \langle \boldsymbol{\gamma}_{i,0}, \mathbf{p}_{-i}\rangle) = \mu_i(\langle \boldsymbol{\theta}_{i,0}, \mathbf{p}\rangle), \quad (1)$$

for each $i \in [N]$, where $\mathbf{p}_{-i} = (p_j)_{j\in[N]\setminus\{i\}}$, $\mathbf{p} = (p_j)_{j\in[N]}$ and $\boldsymbol{\theta}_{i,0}$ is a vector of dimension N with i-th entry equals to $-\beta_{i,0}$ (where $\beta_{i,0} > 0$) and the rest of the entries are the values of $\boldsymbol{\gamma}_{i,0}$ ordered. The link function μ_i is assumed to be increasing, capturing the same economic intuition that demand decreases with the seller's own price. Here μ_i is known to seller i (but unknown to competitors) and the parameter $\boldsymbol{\theta}_{i,0}$ is unknown (to everyone). This model generalizes the linear demand model by Li et al. (2024); specifically, $\mu_i(u) = u + \alpha_{i,0}$ recovers their formulation.

Multi-Agent Reinforcement Learning (MARL). MARL is a broad framework for sequential decision-making problems involving multiple strategic agents that at each instant select actions based on their local observations, receive feedback in the form of rewards (which depend on personal and competitors' actions), and then update their policy to maximize long-term returns. The ultimate objective is to reach a stable solution concept, such as an NE, where no agent can unilaterally improve its expected return.

A key modeling assumption in MARL algorithms is full or partial observability of other agents' rewards or payoffs (Hu and Wellman, 2003; Zhang et al., 2021; Yang and Wang, 2020). For instance, the canonical formulation of Hu and Wellman (2003) assumes that each agent can observe the immediate reward of all players and use this to update joint policy. Similarly, Zhang et al. (2021) and Yang and Wang (2020) rely on either joint reward observability or access to sufficient statistics (e.g., payoff signals or utility vectors) that allow agents to infer the equilibrium structure. This assumption is critical: observing other agents' realized rewards enables an agent to estimate the full payoff landscape and to learn best-response dynamics with respect to the joint reward function.

Our setting fundamentally differs from this MARL paradigm in both information structure and observability. In the sequential pricing game we study, each seller observes only their own reward (or demand value) and rivals' posted prices – never competitors' demands or revenues. Because demand functions are private and unobservable, a seller cannot recover rivals' payoff functions or utility gradients, so classical MARL techniques are not directly applicable. Nevertheless, we design a policy whose dynamics need not converge to a Nash equilibrium, yet whose individual regret satisfies $\widetilde{O}(\sqrt{T})$. This makes our approach better aligned with real-world competitive pricing, where revenue data is proprietary and competitors' demand is hidden.

Connection to contextual bandits – and a timing asymmetry. One may view $\mathbf{p}_{-i}^{(t)}$ at any instant t as the "context" for seller i, linking our problem to contextual bandits (Chu et al., 2011; Agrawal and Goyal, 2013; Chakraborty et al., 2023; Kim et al., 2023) and to the nonlinear Cournot structure (Cournot, 1838; Saloner, 1987). However, unlike standard contextual bandits – where the context is observed *before* the action is taken – in our market the sellers post prices simultaneously and observe $\mathbf{p}_{-i}^{(t)}$ only *after* choosing $p_i^{(t)}$. This endogeneity/delay in the "context" prevents a direct application of the elliptical potential lemma (Carpentier et al., 2020), a staple in bandit regret analyses. We resolve this by adapting the analysis to our timing, recovering an appropriate potential-type bound (see Lemma C.2) that enables regret control despite the delayed feedback structure, which arises as a natural consequence of the competitive multi-seller set-up we consider in this paper.

1.2 Contributions

We summarize our main contributions and position them relative to prior works.

General monotone single–index demand. We model each seller's mean demand as $\lambda_i(\mathbf{p}) = \mu_i(\langle \boldsymbol{\theta}_{i,0}, \mathbf{p}\rangle)$, where μ_i is a known and increasing link function but $\boldsymbol{\theta}_{i,0}$ is unknown. This framework strictly generalizes the linear model of Li et al. (2024), capturing a broader range of nonlinear demand patterns. In contrast to Li et al. (2024); Bracale et al. (2025a); Goyal et al. (2023), our analysis applies seamlessly to both binary (purchase/no-purchase) and real-valued (continuous) demand outcomes, thereby unifying common revenue models within a single learning procedure.

No forced i.i.d. exploration phase. Prior work relies on a coordinated, front-loaded exploration phase across all sellers (Li et al., 2024; Bracale et al., 2025a), in which i.i.d. randomized prices are posted to facilitate estimation. While effective in theory, the length of these "artificial" exploration phases depends on unknown constants, making them very difficult to calibrate in practice. Our policy removes the need for such an explicit exploration phase: sellers continuously learn while pricing, yet still achieve optimal statistical guarantees. This makes our method more practical and better aligned with realistic market conditions.

No-regret algorithm. We establish an upper bound on the total expected individual regret of order $O(N^2\sqrt{T}\log(T))$ per seller, up to additive constants. Moreover, we show that the price vector $\mathbf{p}^{(t)}$ satisfies the following bound $\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2] = O(N^2\sqrt{T}\log(T))$, where $\mathbf{p}^*$ is the NE. Thus, our analysis not only guarantees near–optimal regret, but also provides quantitative control over deviation of play from equilibrium.

1.3 General Notation

We write $[N] = \{1, \ldots, N\}$ and $[T] = \{1, \ldots, T\}$. We denote by $\mathbf{p} = (p_j)_{j\in[N]}$ the vector of prices set by all sellers, and by $\mathbf{p}_{-i} = (p_j)_{j\in[N]\setminus\{i\}}$ the vector of prices set by seller i's competitors. When we write $\mathbf{p} = (p_i, \mathbf{p}_{-i})$, we do so to emphasize that the price p_i is chosen by seller i; this notation does not imply that p_i is the first component of $\mathbf{p}$ – it remains in its i-th position. We use $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$ for the ℓ^1, Euclidean, and sup norms, respectively, and $\langle \mathbf{u}, \mathbf{v}\rangle = \mathbf{u}^\top \mathbf{v}$ for the inner product. We write $\widetilde{O}(\cdot)$ to suppress logarithmic factors and use $\lesssim$ to hide absolute constants. For a vector $\mathbf{p} \in \mathbb{R}^N$ and a symmetric positive definite matrix $V \in \mathbb{R}^{N\times N}$ we define the norm $\|\mathbf{p}\|_V := \sqrt{\mathbf{p}^\top V \mathbf{p}}$. A function $f(\mathbf{p}) : \mathbb{R}^N \mapsto \mathbb{R}$ is L-Lipschitz (with respect to a norm $\|\cdot\|$), for some $L > 0$ if $|f(\mathbf{p}') - f(\mathbf{p})| \leq L\|\mathbf{p}' - \mathbf{p}\|$ for all $\mathbf{p}, \mathbf{p}'$ in the domain. We use $\nabla f(\mathbf{p})$ and $\nabla^2 f(\mathbf{p})$ to denote the gradient vector and the Hessian matrix of f at $\mathbf{p}$. We denote by $\mathrm{C}^m(\Omega)$ the set of m-times continuously differentiable functions $f : \Omega \to \mathbb{R}$. A vector field $\mathbf{F} : \Omega \mapsto \Omega$ is a contraction w.r.t. some norm $\|\cdot\|$ if it is L-Lipschitz with $L < 1$.

1.4 Organization of the paper

The remainder of the paper is organized as follows. Section 2 introduces the problem formulation and model assumptions, and formally defines regret and Nash equilibrium. Section 3 presents our proposed methods, including the penalized maximum likelihood estimator and the UCB-based pricing policy. Section 4 derives concentration bounds and establishes our main regret guarantees, outlining the key steps of the proofs. Section 5 concludes the paper by discussing limitations and future research directions. Additional technical details and complete proofs are deferred to the supplementary material.

2 Problem Setting

We consider a pricing problem faced by N sellers that sell a single product with unlimited inventory sequentially over a time horizon T, where T is *known* to the sellers.

Assumption 2.1 (Bounded prices)**.** *For all $t \in [T]$, the i-th seller's price is denoted as $p_i^{(t)} \in \mathcal{P}_i \triangleq [\underline{p_i}, \bar{p}_i]$, with price bounds $\underline{p_i} < \bar{p}_i$ and $\underline{p_i}, \bar{p}_i \in [0, +\infty)$. Every seller knows their own price domain $\mathcal{P}_i$ but not the competing ones $\{\mathcal{P}_j\}_{j\in[N]\setminus\{i\}}$. However, only $B_p = (\sum_{i\in[N]} \bar{p}_i^2)^{1/2}$ is known to all the sellers.*

Let $\mathbf{p}_{-i}^{(t)} \triangleq (p_j^{(t)})_{j\in[N]\setminus\{i\}}$ denote the i-th competitors' prices at time t and $\mathbf{p}^{(t)} \triangleq (p_j^{(t)})_{j\in[N]}$ denote the joint prices. Only for the purpose of the analysis we define $\mathcal{P} \triangleq \prod_{i\in[N]}[\underline{p_i}, \bar{p}_i]$ which denotes the support of joint prices and $\mathcal{P}_{-i} = \prod_{j\in[N]\setminus\{i\}} \mathcal{P}_j$ the support of the competitors' prices. The general dynamic is as follows: at time $t = 0$, each seller i selects any initial price $p_i^{(0)} \in \mathcal{P}_i$ and everyone observes their own realized demand $y_i^{(0)}$ according to (2). The price vector $\mathbf{p}^{(0)}$ is made public. Let $\mathcal{H}_i^{(0)} = \{(\mathbf{p}^{(0)}, y_i^{(0)})\}$ be the data collected by seller i. For each sales round $t = 1, \ldots, T$:

1. Each seller i sets their price $p_i^{(t)}$ based on their

data history $\mathcal{H}_i^{(t-1)}$.

2. Each seller i observes their realized demand $y_i^{(t)}$, sampled by nature according to the model in (2). The vector of prices $\mathbf{p}^{(t)}$ is made public, and then each seller updates their history: $\mathcal{H}_i^{(t)} = \mathcal{H}_i^{(t-1)} \cup \{(\mathbf{p}^{(t)}, y_i^{(t)})\}$.

3. Each seller i observes their (empirical) revenue $r_i = p_i^{(t)} y_i^{(t)}$.

We emphasize that *sellers can observe the historical prices of competitors, but do not know the demand of competitors*. We assume that $y_i^{(t)}$, given the price vector $\mathbf{p}^{(t)}$, is sampled by nature according to a *canonical exponential family* w.r.t. a reference measure ν_i:

$$\frac{d\mathbb{P}_{\boldsymbol{\theta}_{i,0}}(y_i|\mathbf{p})}{d\nu_i(y_i)} = \exp\left\{y_i\langle\boldsymbol{\theta}_{i,0},\mathbf{p}\rangle - b_i(\langle\boldsymbol{\theta}_{i,0},\mathbf{p}\rangle) + c_i(y_i)\right\}, \tag{2}$$

where $c_i(\cdot)$ is a real-valued function and $b_i(\cdot)$ is assumed to be twice continuously differentiable. A random variable y_i with the above density satisfies $\mathbb{E}(y_i|\mathbf{p}) = b_i'(\langle\boldsymbol{\theta}_{i,0},\mathbf{p}\rangle)$ and $\mathrm{var}(y_i|\mathbf{p}) = b_i''(\langle\boldsymbol{\theta}_{i,0},\mathbf{p}\rangle)$, showing that $b_i(\cdot)$ is strictly convex. The inverse link function $\mu_i := b_i'$ is consequently strictly increasing. Thus, we can write

$$\begin{aligned}\lambda_i(\mathbf{p}^{(t)}) &= \mathbb{E}[y_i^{(t)}|\mathbf{p}^{(t)}]\\ &= \mu_i(-\beta_{i,0}p_i^{(t)} + \langle\boldsymbol{\gamma}_{i,0},\mathbf{p}_{-i}^{(t)}\rangle) = \mu_i(\langle\boldsymbol{\theta}_{i,0},\mathbf{p}^{(t)}\rangle),\end{aligned}$$

where μ_i is *known to the seller* i (but not to competitors $j \in [N]\setminus\{i\}$) and $\boldsymbol{\theta}_{i,0}$ is an N dimensional vector *unknown to everyone* with i-th entry equals to $-\beta_{i,0}$ and the rest of the entries are the values of $\boldsymbol{\gamma}_{i,0}$ ordered. The parameters $\beta_{i,0}$ and $\boldsymbol{\gamma}_{i,0}$ are commonly referred to as the i-th seller's own-price sensitivity and the competitors' price sensitivities, respectively. The demands $\{y_j^{(t)}\}_{j\in[N]}$ within a period t can be correlated across sellers conditional on $\mathbf{p}^{(t)}$. The model in (2) generates samples $y_i^{(t)}$ that reflect not only each seller's own price $p_i^{(t)}$ but also the strategies of the competitors $\mathbf{p}_{-i}^{(t)} = (p_j^{(t)})_{j\in[N],j\neq i}$. The exponential model in (2) allows us to generalize linear models and also to accommodate both binary and continuous outcomes: when binary, it might represent the indicator of whether seller i sold the item or not at price $p_i^{(t)}$; when continuous, it can be seen as observed mean demand value at price $\mathbf{p}^{(t)}$ plus a mean-zero error. We present two relevant examples.

Example 2.2 (Binary Response Model)**.** *We consider a generalization of the monopolistic binary response model considered by Fan et al. (2024); Bracale et al. (2025b) (and corresponding references). Each customer arrives at time t and, given the price vector posted by sellers $\mathbf{p}^{(t)}$, samples a valuation for each seller i*

$$v_i^{(t)} = \langle\widetilde{\boldsymbol{\theta}}_{i,0},\mathbf{p}^{(t)}\rangle + \delta_i^{(t)},$$

where $\widetilde{\boldsymbol{\theta}}_{i,0}$ is an unknown parameter and $\delta_i^{(t)}$ are i.i.d. across t with c.d.f. F_i. Here $\delta_i^{(t)}$ are independent across t but might be correlated across i conditional on $\mathbf{p}^{(t)}$. Let $p_i^{(t)}$ be the price posted by seller i. A purchase from i happens if $y_i^{(t)} = \mathbb{I}(p_i^{(t)} \leq v_i^{(t)})$ is equal to 1. Let $\mu_i = 1 - F_i$ and note that

$$\mathbb{E}[y_i^{(t)}|\mathbf{p}^{(t)}] = \mu_i(p_i^{(t)} - \langle\widetilde{\boldsymbol{\theta}}_{i,0},\mathbf{p}^{(t)}\rangle) = \mu_i(\langle\boldsymbol{\theta}_{i,0},\mathbf{p}^{(t)}\rangle),$$

where $\boldsymbol{\theta}_{i,0}$ coincides with $\widetilde{\boldsymbol{\theta}}_{i,0}$ except for the i-th coordinate which instead is equal to $1-[\widetilde{\boldsymbol{\theta}}_{i,0}]_i$, where $[\widetilde{\boldsymbol{\theta}}_{i,0}]_i$ is the i-th coordinate of $\widetilde{\boldsymbol{\theta}}_{i,0}$. This model coincides with the canonical exponential family in (2) with $\mu_i = b_i'$.

Example 2.3 (Linear Regression Model)**.** *Here we describe the model by Li et al. (2024). The observed demand at time t for seller i is assumed to have the form*

$$y_i^{(t)} = \alpha_{i,0} - \beta_{i,0}p_i^{(t)} + \langle\boldsymbol{\gamma}_{i,0},\mathbf{p}_{-i}^{(t)}\rangle + \eta_i^{(t)},$$

$\boldsymbol{\eta}^{(t)} \mid \mathbf{p}^{(t)} \sim \mathcal{N}\left(\mathbf{0},\Sigma\left(\mathbf{p}^{(t)}\right)\right)$, with $\Sigma\left(\mathbf{p}^{(t)}\right)$ positive semidefinite, where $\boldsymbol{\eta}^{(t)} = (\eta_i^{(t)})_{i\in[N]}$. This model matches the canonical exponential family in (2) with $\mu_i(x) = b_i'(x) = \alpha_{i,0} + x$.

We now present the assumptions on μ_i and the parameters $\boldsymbol{\theta}_{i,0}$. We assume that the average demand λ_i is non-negative and $\partial_{p_i}\lambda_i < 0$ among all values of $\{\boldsymbol{\theta}_{i,0}\}_{i\in[N]}$; a similar assumption is found in Birge et al. (2024); Li et al. (2024); Bracale et al. (2025a). The above conditions hold if $\mu_i, \mu_i' > 0$ and $\beta_{i,0} > 0$, which are explicitly stated in Assumption 2.4 and subsequent Assumption 2.5.

Assumption 2.4 (Bounded and feasible parameter space)**.** *The unknown vector $\boldsymbol{\theta}_{i,0} = (-\beta_{i,0},\boldsymbol{\gamma}_{i,0}) \in \Theta_i \subset \mathbb{R}^N$ satisfies $0 < \underline{\beta}_i \leq \beta_{i,0} \leq \bar{\beta}_i$, $\boldsymbol{\gamma}_{i,0} \in K_i \subset \mathbb{R}^{N-1}$ for some K_i compact convex, so that $\Theta_i \triangleq [-\bar{\beta}_i, -\underline{\beta}_i] \times K_i$ is compact convex in $\mathbb{R}^N$. Θ_i is known to seller i.*

We denote by $B_{\theta_i} > 0$ the ℓ^2 upper bound on Θ_i, that is $\|\boldsymbol{\theta}_i\|_2 \leq B_{\theta_i}$ for all $\boldsymbol{\theta}_i \in \Theta_i$.

As far as $\mu_i = b_i'$, we already know that it is non-decreasing because b_i is convex. However, to derive concentration bounds, we require μ_i' to be bounded away from zero. Before making this explicit, we need to define

$$\mathcal{U}_i \triangleq [u_i^{\min}, u_i^{\max}],$$

where $u_i^{\min} \triangleq \inf_{(\boldsymbol{\theta}_i,\mathbf{p})\in\Theta_i\times\mathcal{P}} \boldsymbol{\theta}_i^\top\mathbf{p}$, and $u_i^{\max} \triangleq \sup_{(\boldsymbol{\theta}_i,\mathbf{p})\in\Theta_i\times\mathcal{P}} \boldsymbol{\theta}_i^\top\mathbf{p}$ (known to seller i by Assumption 2.4).

Assumption 2.5 (Smoothness of μ_i). *We assume that $\mu_i \in \mathrm{C}^2(\mathcal{U}_i)$ with*

$$0 \le \mu_i(u) \le B_i, \quad \mu_i'(u) \ge c_{\mu_i}, \quad \forall u \in \mathcal{U}_i,$$

and $\mu_i'(u) \le L_{\mu_i}$, $\forall u \in \mathbb{R}$, for some $B_i, c_{\mu_i}, L_{\mu_i} > 0$. The function μ_i is known to seller i but unknown to competitors $j \in [N] \setminus \{i\}$.

Requiring μ_i' to be bounded away from zero is standard: it prevents degeneracy in the curvature of the log-likelihood (equivalently, keeps the Fisher information well-conditioned) and underpins the concentration inequalities used in our analysis; similar assumptions can be found in Russac et al. (2019, 2020); Balabdaoui et al. (2019); Groeneboom and Hendrickx (2018). Let $B_i'' > 0$ be such that $|\mu_i''(u)| \le B_i''$ for all $u \in \mathcal{U}_i$, which exists because $\mu_i \in \mathrm{C}^2(\mathcal{U}_i)$.

We define the noise term as

$$\eta_i^{(t)} \triangleq y_i^{(t)} - \mu_i(\langle \mathbf{p}^{(t)}, \boldsymbol{\theta}_{i,0} \rangle),$$

which satisfies the following Lemma 2.6, used to derive concentration bounds in our analysis.

Lemma 2.6. *Under Assumption 2.5, for every $i \in [N]$, $\eta_i^{(t)}$ is L_{μ_i}-subgaussian conditionally on $\mathcal{H}_i^{(t-1)}$.*

Remark 2.7 (Impossibility to learn competitors' models). *Before proceeding to the next section, we highlight a key distinction between our framework and the MARL models discussed in Section 1.2. In most MARL settings, the realized rewards of all agents – in our setting the realized demands of competing sellers – are observed by every participant. By contrast, our model assumes that these realized demands remain private. This privacy has an important implication: no seller can infer the policy of their competitors. To see this, recall that the demand model for seller j is $\mathbb{E}[y_j^{(t)} | \mathbf{p}^{(t)}] = \mu_j(\langle \boldsymbol{\theta}_{j,0}, \mathbf{p}^{(t)} \rangle)$ where $\boldsymbol{\theta}_{j,0}$ and μ_j are the parameter vector and link function of seller j, respectively, both unknown to competitors. At each time t, a competitor seller $i \neq j$ only observes $\mathcal{H}_i^{(t)} = \cup_{s \le t} \{(y_i^{(s)}, \mathbf{p}^{(s)})\}$ but does not have access to the demand realizations $\{y_j^{(s)}\}_{s \le t}$ of seller j. Thus, seller i has no information from which to infer their competitor parameters $(\boldsymbol{\theta}_{j,0}, \mu_j)$, and hence predict their pricing policy.*

2.1 Regret and Nash equilibrium

Each seller competes with a dynamic optimal sequence of prices in hindsight while assuming that the other sellers would not have responded differently if this sequence of prices had been offered. Under such a dynamic benchmark, the objective of each seller $i \in [N]$ is to minimize the following regret metric in hindsight:

$$\mathrm{Reg}_i(T) \triangleq \mathbb{E} \textstyle\sum_{t=1}^T \mathrm{rev}_i(\Gamma_i(\mathbf{p}_{-i}^{(t)}) | \mathbf{p}_{-i}^{(t)}) - \mathrm{rev}_i(p_i^{(t)} | \mathbf{p}_{-i}^{(t)})$$

where for every $(p_i, \mathbf{p}_{-i}) \in \mathcal{P}$ we define

$$\mathrm{rev}_i(p_i | \mathbf{p}_{-i}) \triangleq p_i \lambda_i(\mathbf{p}) = p_i \mu_i(\langle \mathbf{p}, \boldsymbol{\theta}_{i,0} \rangle), \tag{3}$$

and for every $\mathbf{p}_{-i} \in \mathcal{P}_{-i}$

$$\Gamma_i(\mathbf{p}_{-i}) \triangleq \arg\max_{p_i \in \mathcal{P}_i} \mathrm{rev}_i(p_i | \mathbf{p}_{-i}), \tag{4}$$

$\Gamma_i : \mathcal{P}_{-i} \to \mathcal{P}_i$ being the i-th seller's *best-response map*.

Assumption 2.8 (Strong Concavity). *For every $i \in [N]$, $\mathrm{rev}_i(\cdot | \mathbf{p}_{-i})$ is strongly concave in $\mathcal{P}_i$, uniformly in $\mathbf{p}_{-i} \in \mathcal{P}_{-i}$.*[2]

Assumption 2.8 is common in the literature (Bracale et al., 2025a; Scutari et al., 2014; Li et al., 2024; Tsekrekos and Yannacopoulos, 2024) and includes linear demand models (Li et al., 2024), concave demand specifications, and s_i-concave demand functions with $s_i \ge 1$ (see Bracale et al. (2025a), Appendix I).

Assumption 2.8 guarantees that the best-response map $\Gamma_i(\cdot)$ in (4) is well defined, i.e., admits a unique value for every $\mathbf{p}_{-i}$, which can be found by solving the first-order condition $\partial_{p_i} \mathrm{rev}_i(p_i | \mathbf{p}_{-i}) = 0$, w.r.t. p_i and projecting the solution on $\mathcal{P}_i$. Additionally, Assumption 2.8 guarantees that the learning dynamics converge with sublinear sellers' regrets. Indeed, as in Li et al. (2024); Bracale et al. (2025a), we control the individual regret upper bound by controlling the L^2 difference of the joint policies with respect to the NE, i.e. $\mathrm{Reg}_i(T) \lesssim \sum_{t=1}^T \|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2$ where $\mathbf{p}^*$ is a NE (see Equation (15) for a detailed bound). A Nash equilibrium $\mathbf{p}^* = (p_i^*)_{i \in [N]} \in \mathcal{P}$ is defined as a price vector under which unilateral deviation is not profitable for any seller. Specifically,

$$\mathrm{rev}_i(p_i^* | \mathbf{p}_{-i}^*) \ge \mathrm{rev}_i(p_i | \mathbf{p}_{-i}^*), \quad \forall p_i \in \mathcal{P}_i, \quad \forall i \in [N],$$

or, equivalently, $\mathbf{p}^*$ is a solution to the following fixed point equation:

$$\mathbf{p}^* = \boldsymbol{\Gamma}(\mathbf{p}^*) = (\Gamma_1(\mathbf{p}_{-1}^*), \ldots, \Gamma_N(\mathbf{p}_{-N}^*)), \tag{5}$$

where $\boldsymbol{\Gamma} : \mathcal{P} \to \mathcal{P}$ is called *best-response operator*. The next result is an immediate consequence of Theorem 3 in Scutari et al. (2014).

Lemma 2.9 (Existence of NE). *Under assumptions (2.1), (2.4), (2.5) and (2.8), there exists a $\mathbf{p}^* \in \mathcal{P}$ satisfying the fixed point equation in (5).*

Our next assumption guarantees the uniqueness of the NE.

Assumption 2.10 (Sufficiency for unique NE). *μ_i is log-concave in $\mathcal{U}_i$, and $\sup_{i \in [N]} \frac{\|\boldsymbol{\gamma}_{i,0}\|_1}{\beta_{i,0}} < 1$.*

[2] There exists $\xi_i > 0$ independent of $\mathbf{p}_{-i}$ such that for all $x, y \in \mathcal{P}_i$ and all $\mathbf{p}_{-i}$, $\mathrm{rev}_i(y | \mathbf{p}_{-i}) - \mathrm{rev}_i(x | \mathbf{p}_{-i}) \le \langle \nabla_x \mathrm{rev}_i(x | \mathbf{p}_{-i}), y - x \rangle - \frac{\xi_i}{2} \|y - x\|_2^2$.

The log-concavity assumption covers the linear demand model by Li et al. (2024) for which $\mu_i(u) = \alpha_{i,0} + u$ is log-concave. This assumption can be found in several works with $N = 1$ (monopolistic setting); for example, Assumption 2.1 in Fan et al. (2024), Assumption 1 in Chen and Farias (2018), Equation 1 in Cole and Roughgarden (2014) and Golrezaei et al. (2019); Javanmard (2017); Javanmard and Nazerzadeh (2019) which assume μ_i and $1-\mu_i$ to be log-concave, where μ_i is a survival function. The assumption in the model parameter constraint coincides with Assumption 2 in Li et al. (2024), and similar assumptions can be found in Bracale et al. (2025a); Kachani et al. (2007).

Lemma 2.11 (Contraction of the best-response operator)**.** *If assumption (2.10) holds in addition to the assumptions in Lemma 2.9, then* $\mathbf{\Gamma}$*, defined in (5), is a contraction under the supremum norm with contraction constant* $L_{\mathbf{\Gamma}} := \sup_{i\in[N]} \|\gamma_{i,0}\|_1/\beta_{i,0} < 1$.

Lemma 2.11 was proved by Bracale et al. (2025a) in the more general case of s_i-concave demand model for $s_i > -1$ (log-concavity is covered for $s_i = 0$) and implies that the NE is unique by the Banach Contraction theorem. As in Li et al. (2024); Bracale et al. (2025b) the uniqueness of the NE entails sublinear individual regret for all sellers, as we will see in the proof of our main Theorem 4.1.

3 Algorithm

We adopt a similar algorithm to Russac et al. (2020) based on UCB of the revenue functions $\mathrm{rev}_i(\cdot|\mathbf{p}_{-i})$ for every $\mathbf{p}_{-i} \in \mathcal{P}_{-i}$. This method relies on a penalized log-likelihood approach that has a regularizing effect and avoids the need for common initial exploration across the sellers.

3.1 Penalized MLE

By the exponential model in (2), the (penalized) maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_i^{(t)}$ based on the observed $y_i^{(1)}, \dots, y_i^{(t-1)}$ and the selected prices $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(t-1)}$ is defined as the maximizer of

$$\sum_{s=1}^{t-1} \ln(\mathbb{P}_{\boldsymbol{\theta}_i}(y_i^{(s)}|\mathbf{p}^{(s)})) - \frac{\lambda}{2}\|\boldsymbol{\theta}_i\|_2^2 \tag{6}$$
$$= \sum_{s=1}^{t-1} y_i^{(s)}\langle \boldsymbol{\theta}_i, \mathbf{p}^{(s)}\rangle - b_i(\langle \boldsymbol{\theta}_i, \mathbf{p}^{(s)}\rangle) + c_i(y_i^{(s)}) - \frac{\lambda}{2}\|\boldsymbol{\theta}_i\|_2^2$$

which is strictly concave because, by convexity of b_i, $\sum_{s=1}^{t-1} \ln(\mathbb{P}_{\boldsymbol{\theta}_i}(y_i^{(s)}|\mathbf{p}^{(s)}))$ is concave in $\boldsymbol{\theta}_i$. Define $\widehat{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}$ as the unique solution of

$$\sum_{s=1}^{t-1} \left\{ y_i^{(s)} - \mu_i(\langle \mathbf{p}^{(s)}, \boldsymbol{\theta}_i \rangle) \right\} \mathbf{p}^{(s)} - \lambda \boldsymbol{\theta}_i = 0, \tag{7}$$

where the LHS is the derivative of the penalized log-likelihood in (6). Following the notation in Russac et al. (2020), we introduce

$$V_i^{(t-1)} \triangleq \sum_{s=1}^{t-1} \mathbf{p}^{(s)}\mathbf{p}^{(s)\top} + \tfrac{\lambda}{c_{\mu_i}} I_N, \tag{8}$$

and

$$g_i^{(t-1)}(\boldsymbol{\theta}_i) \triangleq \sum_{s=1}^{t-1} \mu_i(\langle \mathbf{p}^{(s)}, \boldsymbol{\theta}_i \rangle)\mathbf{p}^{(s)} + \lambda \boldsymbol{\theta}_i.$$

For fixed i and t, $g_i^{(t-1)}$ denotes a surrogate function introduced to obtain an upper bound on the matrix $V_i^{(t-1)}$ – in the sense of partial ordering on the non-negative matrices. This bound plays a key role in controlling the upper confidence bound that will be formally defined in Section 3.2. At every time t, the i-th seller estimator is $\widetilde{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}$, defined by

$$\widetilde{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}} = \underset{\|\boldsymbol{\theta}_i\|_2 \le B_{\theta_i}}{\operatorname{argmin}} \|g_i^{(t-1)}(\widehat{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}) - g_i^{(t-1)}(\boldsymbol{\theta}_i)\|_{V_i^{(t-1),-1}}, \tag{9}$$

where $V_i^{(t-1),-1}$ is the inverse of $V_i^{(t-1)}$. We need to consider $\widetilde{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}$ because $\widehat{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}$ is not guaranteed to satisfy $\|\widehat{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}\|_2 \le B_{\theta_i}$. Here, $\widetilde{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}$ should be understood as a "projection" onto the admissible parameter set.

3.2 Pricing Policy via UCB

In this section, we describe how each seller i determines their prices $p_i^{(t)}$ at a given time t. Fix $t \in [T]$ and let $\widetilde{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}} = (-\widetilde{\beta}_i^{(t)\mathtt{PML}}, \widetilde{\boldsymbol{\gamma}}_i^{(t)\mathtt{PML}})$ be the parameter estimate defined in (9). For any $(p_i, \mathbf{p}_{-i}) \in \mathcal{P}$, we define the estimated revenue as the plug-in estimator:

$$\begin{aligned}\widehat{\mathrm{rev}}_i^{(t)}(p_i|\mathbf{p}_{-i}) &= p_i\mu_i(-\widetilde{\beta}_i^{(t)\mathtt{PML}} p_i + \langle \widetilde{\boldsymbol{\gamma}}_i^{(t)\mathtt{PML}}, \mathbf{p}_{-i}\rangle) \\ &= p_i\mu_i(\langle \mathbf{p}, \widetilde{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}\rangle).\end{aligned}$$

Following standard UCB-based approaches (Russac et al., 2020, 2019), we define the upper confidence bound (UCB) of the estimated revenue as

$$\mathrm{UCB}_i^{(t)}(p_i|\mathbf{p}_{-i}^{(t-1)}) = \widehat{\mathrm{rev}}_i^{(t)}(p_i|\mathbf{p}_{-i}^{(t-1)}) + \sigma_i^{(t-1)}(p_i, \delta), \tag{10}$$

where for all $s = 0, 1, \dots, T-1$ we define

$$\begin{aligned}\sigma_i^{(s)}(p_i, \delta) &\triangleq \rho_i^{(s)}(\delta) p_i \|(p_i, \mathbf{p}_{-i}^{(s)})\|_{V_i^{(s),-1}}, \\ \rho_i^{(s)}(\delta) &\triangleq \frac{2L_{\mu_i}}{c_{\mu_i}}\left(\mathrm{c}_i^{(s)}(\delta) + B_{\theta_i}\sqrt{c_{\mu_i}\lambda}\right), \\ \mathrm{c}_i^{(s)}(\delta) &\triangleq L_{\mu_i}\sqrt{2\log\left(\frac{T}{\delta}\right) + N\log\left(1 + \frac{B_p^2 s}{N\lambda}\right)}.\end{aligned} \tag{11}$$

Algorithm 1 `PML-GLUCB` (Penalized Maximum Likelihood Generalized Linear Upper Confidence Bound)

Input: Probability δ, regularization λ, upper bounds B_p, B_{θ_i}, B_i, c_{μ_i}, and L_{μ_i}.
Initialize: $\widehat{\boldsymbol{\theta}}_i^{(0)\texttt{PML}} \in \Theta_i$, each seller starts with any $p_i^{(0)} \in \mathcal{P}_i$ and observes the corresponding $y_i^{(0)}$. At the end of time $t=0$ each seller i observes $\mathbf{p}^{(0)}$ and sets $V_i^{(0)} = \mathbf{p}^{(0)}\mathbf{p}^{(0)\top} + \lambda/c_{\mu_i} I_N$.
for $t = 1$ **to** T **do**
▶ Each seller i computes $\widehat{\boldsymbol{\theta}}_i^{(t)\texttt{PML}}$ according to (7).
if $\|\widehat{\boldsymbol{\theta}}_i^{(t)\texttt{PML}}\|_2 \le B_{\theta_i}$ **let** $\widetilde{\boldsymbol{\theta}}_i^{(t)\texttt{PML}} = \widehat{\boldsymbol{\theta}}_i^{(t)\texttt{PML}}$ **else** compute $\widetilde{\boldsymbol{\theta}}_i^{(t)\texttt{PML}}$ defined in (9).
▶ Each seller i set $p_i^{(t)} \leftarrow \widehat{\Gamma}_i^{(t)}(\mathbf{p}_{-i}^{(t-1)})$ as defined in (12).
▶ Each seller i observes $y_i^{(t)}$.
▶ Each seller i observes $\mathbf{p}^{(t)}$.
▶ Each seller i updates $V_i^{(t)} \leftarrow V_i^{(t-1)} + \mathbf{p}^{(t)}\mathbf{p}^{(t)\top}$.
end for

We finally define $p_i^{(t)} = \widehat{\Gamma}_i^{(t)}(\mathbf{p}_{-i}^{(t-1)})$, where

$$\widehat{\Gamma}_i^{(t)}(\mathbf{p}_{-i}^{(t-1)}) \in \operatorname{argmax}_{p_i \in \mathcal{P}_i} \text{UCB}_i^{(t)}(p_i|\mathbf{p}_{-i}^{(t-1)}), \quad (12)$$

which represents an optimistic approximation of the best-response map Γ_i at $\mathbf{p}_{-i}^{(t-1)}$.

Using this notation, we are now ready to present our `PML-GLUCB` algorithm (Penalized Maximum Likelihood Generalized Linear Upper Confidence Bound), which is fully described in Algorithm 1.

4 Regret Analysis

In this section, we present our main result, Theorem 4.1, which shows the $\widetilde{O}(\sqrt{T})$ bound on the individual regret.

Theorem 4.1 (Regret of `PML-GLUCB`)**.** *Suppose assumptions (2.1), (2.4), (2.5), (2.8) and (2.10) hold. For every $i \in [N]$ we have that, if $\delta = \frac{1}{T^\gamma}$ for some fixed $\gamma > 1$, then, as long as λ varies between $O(1/\sqrt{T})$ and $O(1)$,*

$$\text{Reg}_i(T) \le C + O(N^2\sqrt{T}\log(T)),$$

for some $C > 0$ independent of T. An upper bound expression for $\text{Reg}_i(T)$ depending on λ, N and T can be found in (30). Additionally we have

$$\textstyle\sum_{t=1}^T \mathbb{E}[\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2] = O(N^2\sqrt{T}\log(T)), \quad (13)$$

where $\mathbf{p}^{(t)} = \widehat{\boldsymbol{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) = (\widehat{\Gamma}_i^{(t)}(\mathbf{p}_{-i}^{(t-1)}))_{i\in[N]}$, i.e. $p_i^{(t)} = \widehat{\Gamma}_i^{(t)}(\mathbf{p}_{-i}^{(t-1)})$, $\forall i \in [N]$.

Proof Sketch. Using Taylor expansion around $\Gamma_i(\mathbf{p}_{-i}^{(t)})$, we can write

$$\begin{aligned}&\text{rev}_i\left(\Gamma_i(\mathbf{p}_{-i}^{(t)})|\mathbf{p}_{-i}^{(t)}\right) - \text{rev}_i\left(p_i^{(t)}|\mathbf{p}_{-i}^{(t)}\right)\\ &\lesssim (\Gamma_i(\mathbf{p}_{-i}^{(t)}) - p_i^{(t)})^2 \le \|\boldsymbol{\Gamma}(\mathbf{p}^{(t)}) - \mathbf{p}^{(t)}\|_2^2,\end{aligned}$$

and summing over t and taking $\mathbb{E}$ both sides we have

$$\text{Reg}_i(T) \lesssim \textstyle\sum_{t=1}^T \mathbb{E}\|\boldsymbol{\Gamma}(\mathbf{p}^{(t)}) - \mathbf{p}^{(t)}\|_2^2. \quad (14)$$

Adding and subtracting the fixed point $\mathbf{p}^*$ from the RHS of (14) and using that $\mathbf{p}^* = \boldsymbol{\Gamma}(\mathbf{p}^*)$ along with Lipschitzianity of $\boldsymbol{\Gamma}$ we get that $\text{Reg}_i(T)$ is bounded by

$$\textstyle\sum_{t=1}^T \mathbb{E}\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2 = \sum_{t=1}^T \mathbb{E}\|\widehat{\boldsymbol{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \mathbf{p}^*\|_2^2. \quad (15)$$

Now let $\sigma^{(t)}(\delta) \triangleq \sum_{i\in[N]} \sigma_i^{(t)}(p_i^{(t+1)}, \delta)$. Adding and subtracting $\boldsymbol{\Gamma}(\mathbf{p}^{(t-1)})$ from $\|\widehat{\boldsymbol{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \mathbf{p}^*\|_2$ and iterating this inequality T times, we get

$$\|\widehat{\boldsymbol{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \mathbf{p}^*\| \lesssim \left(\textstyle\sum_{j=0}^{t-1} L_{\boldsymbol{\Gamma}}^j \sqrt{\sigma^{(t-j-1)}(\delta)}\right) + L_{\boldsymbol{\Gamma}}^{t-1}, \quad (16)$$

which holds with high probability by the following Proposition 4.2.

Proposition 4.2. *Let the assumptions of the theorem hold. For every $0 < \delta < 1$, the event*

$$\cap_{t\in[T]}\{\|\widehat{\boldsymbol{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \boldsymbol{\Gamma}(\mathbf{p}^{(t-1)})\|_2^2 \lesssim \sigma^{(t-1)}(\delta)\},$$

holds with probability at least $1 - 2N\delta$.

Plugging the bound from (16) into the expectation (15) (and ignoring the expectation over the complementary probability set, which can be shown to be negligible), and using that $L_{\boldsymbol{\Gamma}} < 1$, we obtain

$$\text{Reg}_i(T) \lesssim \mathbb{E}\textstyle\sum_{t=0}^{T-1} \sigma^{(t)}(\delta).$$

Now, for simplicity of the proof's sketch, consider λ a fixed constant and choose $\delta = 1/T^2$. From the definition of $\rho_i^{(t)}(\delta)$ in Equation (11) it is easy to see that $\rho_i^{(t)}(\delta) \lesssim \sqrt{N\log(T)}$, hence

$$\begin{aligned}\sigma^{(t)}(\delta) &= \sum_{j\in[N]} \rho_j^{(t)}(\delta) p_j^{(t+1)} \|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t),-1}}\\ &\lesssim \sqrt{N\log(T)} \sum_{j\in[N]} \|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t),-1}}.\end{aligned}$$

Therefore we obtain

$$\text{Reg}_i(T) \lesssim \sqrt{N\log(T)} \cdot K(T), \quad (17)$$

where $K(T) = \mathbb{E}\sum_{t=0}^{T-1}\sum_{j\in[N]}\|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t),-1}}$. The reader can immediately recognize the correspondence of $K(T)$ with the elliptical potential lemma used in the bandit literature, with the only difference being that, in a multi-agent setting, the presence of $p_j^{(t+1)}$ does not allow a direct application of this lemma. However, in Lemma C.2 we prove that it is still possible to bound $K(T)$ by $O(N^{3/2}\sqrt{T\log(T)})$ without additional assumptions. This completes the regret convergence since, by Equation (17), we obtain $\mathrm{Reg}_i(T) \lesssim \sqrt{N\log(T)}\cdot K(T) = O(N^2\sqrt{T}\log(T))$. The convergence of $\sum_{t=1}^{T}\mathbb{E}\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2$ comes from (15). □

Remark 4.3 (Optimality of the `PML-GLUCB` algorithm). *The algorithm of Li et al. (2024) hits the minimax-optimal $\widetilde{\Theta}(\sqrt{T})$ regret under linear demand. Because our framework strictly generalizes the linear case and* `PML-GLUCB` *still achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret, our mechanism is order-optimal – minimax optimal up to logarithmic factors – in this more general setting.*

Remark 4.4 (Computational complexity.). *At round t, seller i updates the matrix $V_i^{(t)} = V_i^{(t-1)} + \mathbf{p}^{(t)}\mathbf{p}^{(t)\top}$ and its inverse. Using the Sherman–Morrison formula to update $V_i^{(t),-1}$ costs $O(N^2)$ time (and forming $V_i^{(t)}$ itself is also $O(N^2)$). To evaluate the UCB bonus for a candidate price p_i, one needs $\|(p_i, \mathbf{p}_{-i})\|_{V_i^{(t),-1}}$, which requires a matrix–vector product $V_i^{(t),-1}\mathbf{p}$ in $O(N^2)$ time and then a dot product in $O(N)$. If the price is chosen by evaluating a grid of G candidate prices (or an equivalent one-dimensional search that uses G evaluations), the per-round cost per seller is $O(GN^2)$. Over T rounds, the time is $O(TGN^2)$ per seller, and $O(TGN^3)$ if all N sellers are run. Space usage per seller is $O(N^2)$ to store $V_i^{(t)}$ and (optionally) $V_i^{(t),-1}$ plus $O(N)$ for current vectors; if past histories are not stored, $O(N^2)$ dominates.*

4.1 Convergence to Nash Equilibrium

The bound in (13) alone does not imply convergence of $\mathbb{E}[\|\mathbf{p}^{(T)} - \mathbf{p}^*\|_2^2]$. This happens of course if $\mathbb{E}[\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2]$ is monotonically decreasing in t, as in Li et al. (2024) and Bracale et al. (2025a) due to the initial i.i.d. pricing exploration in their algorithm. Under such an i.i.d. exploration phase, those works establish that $\|\widehat{\boldsymbol{\theta}}_i^{(t)} - \boldsymbol{\theta}_{i,0}\|_2 = O_P(t^{-1/2})$ under standard regularity conditions, which in turn implies $\mathbb{E}[\|\mathbf{p}^{(T)} - \mathbf{p}^*\|_2^2] = O(T^{-1/2})$. Unfortunately, this is not the case in our setting due to the absence of i.i.d. pricing estimation, since the prices $\{\mathbf{p}^{(t)}\}_{t\in[T]}$ are inherently *adaptive* rather than i.i.d., as their selection depends on past noise realizations. In the absence of i.i.d. pricing excitation, estimation error can only be controlled *along the directions that are sufficiently explored.* As a result, $\|\widehat{\boldsymbol{\theta}}_i^{(t)} - \boldsymbol{\theta}_{i,0}\|_2$ does not generally decay at the parametric rate $O_P(t^{-1/2})$. Nevertheless, as formally established by our Theorem 4.1, achieving *optimal* (or near-optimal) regret remains possible, which is consistent with the general bandit literature, where the same phenomenon is observed – for example, in OFU and UCB-type algorithms for linear and generalized linear bandits (Dani et al., 2008; Abbasi-Yadkori et al., 2011; Filippi et al., 2010; Russac et al., 2019).

5 Conclusions

Limitations. We assume a *known* increasing link μ_i with mild smoothness/curvature, bounded prices and compact parameter sets, and strong concavity of per-round revenues together with a contraction-type condition that guarantees a unique NE. None of these is new: they are standard in dynamic pricing and learning-in-games and *strictly include* the linear-demand setting of Li et al. (2024) as a special case. These conditions ensure well-posed best responses and well-conditioned information for concentration, but they also circumscribe our scope to stationary parameters and synchronous price moves under a unique-equilibrium regime.

Future directions. A natural avenue is to relax link knowledge by replacing it with shape constraints (e.g., monotonicity or s-concavity) and to derive calibration-free confidence bonuses that adapt to local curvature while maintaining a non-cooperative initial exploration phase. On the game-theoretic side, moving beyond global contractions – so as to allow multiple or set-valued equilibria – suggests a variational-inequality/monotone-operator treatment to obtain no-regret convergence. Algorithmically, Bayesian or Thompson sampling approaches to GLMs under our information structure (where rivals' revenues remain unobserved) are particularly appealing, as they tend to yield less conservative confidence bounds in practice compared to UCB-based methods. Finally, incorporating nonstationarity in $\boldsymbol{\theta}_{i,0}$, inventory constraints, and asynchronous observability of rivals' prices would bring the model even closer to more real market scenarios.

Overall, `PML-GLUCB` shows that UCB-based policies are efficient in competitive markets with nonlinear demand, even when rivals' payoffs are unobserved. We view our elliptical-potential variant (see Lemma C.2) and concentration results as building blocks for broader learning-in-games settings.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits.

Advances in neural information processing systems, 24.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR.

Balabdaoui, F., Durot, C., and Jankowski, H. (2019). Least squares estimation in the monotone single index model. *Bernoulli*, 25(4B):3276–3310.

Birge, J. R., Chen, H., Keskin, N. B., and Ward, A. (2024). To interfere or not to interfere: Information revelation and price-setting incentives in a multiagent learning environment. *Operations Research*.

Bracale, D., Banerjee, M., Shi, C., and Sun, Y. (2025a). Revenue maximization under sequential price competition via the estimation of s-concave demand functions. *arXiv preprint arXiv:2503.16737*.

Bracale, D., Banerjee, M., Sun, Y., Stoll, K., and Turki, S. (2025b). Dynamic pricing in the linear valuation model using shape constraints. *arXiv preprint arXiv:2502.05776*.

Carpentier, A., Vernade, C., and Abbasi-Yadkori, Y. (2020). The elliptical potential lemma revisited. *arXiv preprint arXiv:2010.10182*.

Chakraborty, S., Roy, S., and Tewari, A. (2023). Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning*, pages 3979–4008. PMLR.

Chen, Y. and Farias, V. F. (2018). Robust dynamic pricing with strategic customers. *Mathematics of Operations Research*, 43(4):1119–1142.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

Cole, R. and Roughgarden, T. (2014). The sample complexity of revenue maximization. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 243–252.

Cooper, W. L., Homem-de Mello, T., and Kleywegt, A. J. (2015). Learning and pricing with models that do not explicitly incorporate competition. *Operations research*, 63(1):86–103.

Cournot, A. (1838). Recherches sur les principes mathématiques de la théorie des richesses par augustin cournot (paris, france: chez l. hachette).

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, number 101, pages 355–366.

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

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23.

Gallego, G., Huh, W. T., Kang, W., and Phillips, R. (2006). Price competition with the attraction demand model: Existence of unique equilibrium and its stability. *Manufacturing & Service Operations Management*, 8(4):359–375.

Golrezaei, N., Javanmard, A., and Mirrokni, V. (2019). Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Advances in Neural Information Processing Systems*, 32.

Goyal, V., Li, S., and Mehrotra, S. (2023). Learning to price under competition for multinomial logit demand. *Available at SSRN 4572453*.

Groeneboom, P. and Hendrickx, K. (2018). Current status linear regression. *The Annals of Statistics*, 46(4):1415–1444.

Hu, J. and Wellman, M. P. (2003). Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069.

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

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

Kachani, S., Perakis, G., and Simon, C. (2007). Modeling the transient nature of dynamic pricing with demand learning in a competitive environment. *Network science, nonlinear science and infrastructure systems*, pages 223–267.

Kim, W., Lee, K., and Paik, M. C. (2023). Double doubly robust thompson sampling for generalized linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8300–8307.

Kirman, A. P. (1975). Learning by firms about demand conditions. In *Adaptive economic models*, pages 137–156. Elsevier.

Li, S., Shi, C., and Mehrotra, S. (2024). LEGO: Optimal online learning under sequential price competition. *Major Revision at Operations Research. Available at SSRN 4803002*.

Russac, Y., Cappé, O., and Garivier, A. (2020). Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113*.

Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32.

Saloner, G. (1987). Cournot duopoly with two production periods. *Journal of Economic Theory*, 42(1):183–187.

Scutari, G., Facchinei, F., Pang, J.-S., and Palomar, D. P. (2014). Real and complex monotone communication games. *IEEE Transactions on Information Theory*, 60(7):4197–4231.

Tsekrekos, A. E. and Yannacopoulos, A. N. (2024). *Variational Inequalities in Management Science and Finance*. Springer.

Wan, Y., Kober, T., and Densing, M. (2022). Nonlinear inverse demand curves in electricity market modeling. *Energy Economics*, 107:105809.

Wan, Y., Kober, T., Schildhauer, T., Schmidt, T. J., McKenna, R., and Densing, M. (2023). Conditions for profitable operation of p2x energy hubs to meet local demand with energy market access. *Advances in Applied Energy*, 10:100127.

Yang, Y. and Wang, J. (2020). An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*.

Zhang, K., Yang, Z., and Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes: Remark 4.4]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [/Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary for
Online Price Competition under Generalized Linear Demands

A Proof of Proposition 4.2

We want to prove that, under the assumptions of (4.1), for every $0 < \delta < 1$,

$$\mathbb{P}\left(\cap_{t\in[T]}\left\{\|\widehat{\mathbf{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \mathbf{\Gamma}(\mathbf{p}^{(t-1)})\|_2^2 \le C_2\sigma^{(t-1)}(\delta)\right\}\right) \ge 1 - 2N\delta,$$

where $\sigma^{(t)}(\delta) \triangleq \sum_{i\in[N]} \sigma_i^{(t)}(p_i^{(t+1)}, \delta)$ and $C_2 = \frac{4}{\min_{i\in[N]}\xi_i}$, with $\xi_i > 0$ such that for all $x, y \in \mathcal{P}_i$ and all $\mathbf{p}_{-i}$,

$$\mathrm{rev}_i(y|\mathbf{p}_{-i}) - \mathrm{rev}_i(x|\mathbf{p}_{-i}) \le \langle \nabla_x \mathrm{rev}_i(x|\mathbf{p}_{-i}), y - x\rangle - \frac{\xi_i}{2}\|y - x\|^2$$

For every $i \in [N]$ the constant ξ_i is independent of $\mathbf{p}_{-i}$ by (2.8). We recall the definitions

$$\mathrm{rev}_i(p_i|\mathbf{p}_{-i}) = p_i\mu_i(-\beta_{i,0}p_i + \langle \boldsymbol{\gamma}_{i,0}, \mathbf{p}_{-i}\rangle), \qquad \forall (p_i, \mathbf{p}_{-i}) \in \mathcal{P},$$

and

$$\widehat{\mathrm{rev}}_i^{(t)}(p_i|\mathbf{p}_{-i}) = p_i\mu_i(-\widetilde{\beta}_i^{(t)\mathsf{PML}}p_i + \langle \widetilde{\boldsymbol{\gamma}}_i^{(t)\mathsf{PML}}, \mathbf{p}_{-i}\rangle), \qquad \forall (p_i, \mathbf{p}_{-i}) \in \mathcal{P}.$$

For simplicity of notation, we also define

$$p_i^{(t)\star} = \Gamma_i(\mathbf{p}_{-i}^{(t)}) = \operatorname*{argmax}_{p_i\in\mathcal{P}_i} \mathrm{rev}_i(p_i|\mathbf{p}_{-i}^{(t)}), \qquad \forall \mathbf{p}_{-i}^{(t)} \in \mathcal{P}_{-i},$$

and

$$p_i^{(t+1)} = \widehat{\Gamma}_i^{(t+1)}(\mathbf{p}_{-i}^{(t)}) \in \operatorname*{argmax}_{p_i\in\mathcal{P}_i} \mathrm{UCB}_i^{(t+1)}(\cdot|\mathbf{p}_{-i}^{(t)}), \qquad \forall \mathbf{p}_{-i}^{(t)} \in \mathcal{P}_{-i},$$

where we recall from Equation (10) that

$$\mathrm{UCB}_i^{(t+1)}(p_i|\mathbf{p}_{-i}^{(t)}) = \widehat{\mathrm{rev}}_i^{(t+1)}(p_i|\mathbf{p}_{-i}^{(t)}) + \sigma_i^{(t)}(p_i, \delta).$$

For the first part of the proof, we follow similar steps to Russac et al. (2020), Corollary 1. We start by upper-bounding the following difference:

$$\begin{aligned}
\mathrm{rev}_i(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)}) - \mathrm{rev}_i(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)}) &= \underbrace{\mathrm{rev}_i(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)}) - \widehat{\mathrm{rev}}_i^{(t+1)}(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)})}_{A1} \\
&+ \underbrace{\widehat{\mathrm{rev}}_i^{(t+1)}(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)}) - \widehat{\mathrm{rev}}_i^{(t+1)}(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)})}_{A2} \\
&+ \underbrace{\widehat{\mathrm{rev}}_i^{(t+1)}(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)}) - \mathrm{rev}_i(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)})}_{A3}.
\end{aligned}$$

To upper bound $A1$ and $A3$ we use the following Lemma A.1, whose proof can be found in Appendix B.

Lemma A.1. *Let assumptions (2.1), (2.4), (2.5) and (2.8) hold. Recall that for every $t = 0, 1, \ldots, T-1$ we have defined*

$$\sigma_i^{(t)}(p_i, \delta) = \rho_i^{(t)}(\delta)p_i\|(p_i, \mathbf{p}_{-i}^{(t)})\|_{V_i^{(t),-1}},$$

$$\rho_i^{(t)}(\delta) = \frac{2L_{\mu_i}}{c_{\mu_i}}\left(\mathrm{c}_i^{(t)}(\delta) + B_{\theta_i}\sqrt{c_{\mu_i}\lambda}\right), \quad \text{and} \quad \mathrm{c}_i^{(t)}(\delta) = L_{\mu_i}\sqrt{2\log\left(\frac{T}{\delta}\right) + N\log\left(1 + \frac{B_p^2 t}{N\lambda}\right)}.$$

Let $0<\delta<1$ and $t\in[T]$. Let $\mathbf{p}$ be any $\mathcal{P}$-valued (possibly random) price vector. Then, simultaneously for all $t\in[T]$,

$$|\operatorname{rev}_i(p_i|\mathbf{p}_{-i})-\widehat{\operatorname{rev}}_i^{(t)}(p_i|\mathbf{p}_{-i})|=\left|p_i\mu_i(\langle\mathbf{p},\boldsymbol{\theta}_{i,0}\rangle)-p_i\mu_i(\langle\mathbf{p},\widetilde{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}\rangle)\right|\le\rho_i^{(t-1)}(\delta)p_i\|\mathbf{p}\|_{V_i^{(t-1)},-1}$$

holds with probability higher than $1-\delta$.

Thanks to Lemma A.1, we can give an upper bound for the term $A1$ and for the term $A3$. With a union bound, we can simultaneously upper bound $A1$ and $A3$ for all $t\in[T]$ and the following holds

$$\mathbb{P}\left(\forall t\in[T],A1\le\rho_i^{(t)}(\delta)p_i^{(t)\star}\|(p_i^{(t)\star},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}\cap A3\le\rho_i^{(t)}(\delta)p_i^{(t+1)}\|(p_i^{(t+1)},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}\right)\ge1-2\delta \tag{18}$$

Let E be the event in (18). We now upper-bound $A2$.

$$\begin{aligned}
A2&=\widehat{\operatorname{rev}}_i^{(t+1)}(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)})-\widehat{\operatorname{rev}}_i^{(t+1)}(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)})\\
&=\underbrace{\widehat{\operatorname{rev}}_i^{(t+1)}(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)})+\rho_i^{(t)}(\delta)p_i^{(t)\star}\|(p_i^{(t)\star},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}}_{=\mathrm{UCB}_i^{(t+1)}(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)})}\\
&\quad-\rho_i^{(t)}(\delta)p_i^{(t)\star}\|(p_i^{(t)\star},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}+\rho_i^{(t)}(\delta)p_i^{(t+1)}\|(p_i^{(t+1)},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}\\
&\quad\underbrace{-\rho_i^{(t)}(\delta)p_i^{(t+1)}\|(p_i^{(t+1)},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}-\widehat{\operatorname{rev}}_i^{(t+1)}(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)})}_{=-\mathrm{UCB}_i^{(t+1)}(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)})}\\
&=\underbrace{\mathrm{UCB}_i^{(t+1)}(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)})-\mathrm{UCB}_i^{(t+1)}(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)})}_{\le0\text{ by def of }p_i^{(t+1)}}\\
&\quad-\rho_i^{(t)}(\delta)p_i^{(t)\star}\|(p_i^{(t)\star},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}+\rho_i^{(t)}(\delta)p_i^{(t+1)}\|(p_i^{(t+1)},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}\\
&\le\rho_i^{(t)}(\delta)p_i^{(t+1)}\|(p_i^{(t+1)},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}-\rho_i^{(t)}(\delta)p_i^{(t)\star}\|(p_i^{(t)\star},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}.
\end{aligned}$$

Now we put $A1, A2, A3$ together. Under the event E (which occurs with a probability higher than $1-2\delta$), we have

$$\begin{aligned}
\operatorname{rev}_i(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)})-\operatorname{rev}_i(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)})&\le\underbrace{\rho_i^{(t)}(\delta)p_i^{(t)\star}\|(p_i^{(t)\star},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}}_{\text{coming from}A1}\\
&\quad+\underbrace{\rho_i^{(t)}(\delta)p_i^{(t+1)}\|(p_i^{(t+1)},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}-\rho_i^{(t)}(\delta)p_i^{(t)\star}\|(p_i^{(t)\star},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}}_{\text{coming from}A2}\\
&\quad+\underbrace{\rho_i^{(t)}(\delta)p_i^{(t+1)}\|(p_i^{(t+1)},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}}_{\text{coming from}A3}\\
&\le2\rho_i^{(t)}(\delta)p_i^{(t+1)}\|(p_i^{(t+1)},\mathbf{p}_{-i}^{(t)})\|_{V_i^{(t)},-1}=2\sigma_i^{(t)}(p_i^{(t+1)},\delta).
\end{aligned}$$

Up to here we proved that for $0<\delta<1$ and for every $i\in[N]$,

$$\mathbb{P}\left(\cap_{t\in[T]}\left\{\operatorname{rev}_i(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)})-\operatorname{rev}_i(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)})\le2\sigma_i^{(t)}(p_i^{(t+1)},\delta)\right\}\right)\ge1-2\delta. \tag{19}$$

Now, by Assumption 2.8 there exists $\xi_i>0$ independent of $\mathbf{p}_{-i}^{(t)}$ such that for all $x,y\in\mathcal{P}_i$ and all $\mathbf{p}_{-i}\in\mathcal{P}_{-i}$,

$$\operatorname{rev}_i(y|\mathbf{p}_{-i}^{(t)})-\operatorname{rev}_i(x|\mathbf{p}_{-i}^{(t)})\le\langle\nabla_x\operatorname{rev}_i(x|\mathbf{p}_{-i}^{(t)}),y-x\rangle-\frac{\xi_i}{2}\|y-x\|^2.$$

replacing x with $p_i^{(t)*}$ and y with $p_i^{(t+1)}$, together with the optimality of $p_i^{(t)*}$, we recover that

$$(p_i^{(t)*} - p_i^{(t+1)})^2 \le \frac{2}{\xi_i}[\text{rev}_i(p_i^{(t)\star}|\mathbf{p}_{-i}^{(t)}) - \text{rev}_i(p_i^{(t+1)}|\mathbf{p}_{-i}^{(t)})] \tag{20}$$

Now define $C_{i,2} \triangleq \frac{4}{\xi_i}$, and the event

$$\mathcal{B}_i \triangleq \cap_{t=1}^T \left\{(\Gamma_i(\mathbf{p}_{-i}^{(t)}) - \widehat{\Gamma}_i^{(t+1)}(\mathbf{p}_{-i}^{(t)}))^2 \le C_{i,2}\sigma_i^{(t)}(p_i^{(t+1)}, \delta)\right\}.$$

Since $(\Gamma_i(\mathbf{p}_{-i}^{(t)}) - \widehat{\Gamma}_i^{(t+1)}(\mathbf{p}_{-i}^{(t)}))^2 = (p_i^{(t)*} - p_i^{(t+1)})^2$, then (19) and (20) imply that for $\delta \in (0,1)$ and for every $i \in [N]$, $\mathbb{P}(\mathcal{B}_i) \ge 1 - 2\delta$. Now define $\mathcal{B} = \cap_{i\in[N]}\mathcal{B}_i$. Note that $\mathcal{B}$ holds with probability at least $1 - 2N\delta$:

$$\mathbb{P}(\mathcal{B}^c) = \mathbb{P}(\cup_{i\in[N]}\mathcal{B}_i^c) \le \sum_{i\in[N]} \mathbb{P}(\mathcal{B}_i^c) = \sum_{i\in[N]}(1 - \mathbb{P}(\mathcal{B}_i)) \le \sum_{i\in[N]} 2\delta = 2N\delta.$$

Now define

$$C_2 \triangleq \max_{i\in[N]}\{C_{i,2}\} = \frac{4}{\min_{i\in[N]}\xi_i}, \qquad \sigma^{(t)}(\delta) \triangleq \sum_{i\in[N]} \sigma_i^{(t)}(p_i^{(t+1)}, \delta), \qquad \forall t \ge 1.$$

In $\mathcal{B}$ we have that

$$\begin{aligned}
\|\widehat{\mathbf{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \mathbf{\Gamma}(\mathbf{p}^{(t-1)})\|_2 &= \sqrt{\sum_{i\in[N]} (\widehat{\Gamma}_i^{(t)}(\mathbf{p}_{-i}^{(t-1)}) - \Gamma_i(\mathbf{p}_{-i}^{(t-1)}))^2} \\
&\le \sqrt{\max_{i\in[N]}\{C_{i,2}\}}\sqrt{\sum_{i\in[N]} \sigma_i^{(t-1)}(\delta)} \\
&= \sqrt{C_2\sigma^{(t-1)}(\delta)}.
\end{aligned}$$

Let

$$\mathcal{G} = \cap_{t\in[T]} \left\{\|\widehat{\mathbf{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \mathbf{\Gamma}(\mathbf{p}^{(t-1)})\|_2^2 \le C_2\sigma^{(t-1)}(\delta)\right\}.$$

Since $\mathcal{B}$ implies $\mathcal{G}$, we have

$$\mathbb{P}(\mathcal{G}) \ge \mathbb{P}(\mathcal{B}) \ge 1 - 2N\delta.$$

This completes the proof.

B Proof of Lemma A.1

For simplicity of notation, we write $\widehat{\boldsymbol{\theta}}_i^{(t)}$ for $\widehat{\boldsymbol{\theta}}_i^{(t)\text{PML}}$ and $\widetilde{\boldsymbol{\theta}}_i^{(t)}$ for $\widetilde{\boldsymbol{\theta}}_i^{(t)\text{PML}}$. This proof follows a similar structure to the proof of Proposition 1 in Russac et al. (2020). Recall the definitions

$$\text{rev}_i(p_i|\mathbf{p}_{-i}) = p_i\mu_i(-\beta_{i,0}p_i + \langle\boldsymbol{\gamma}_{i,0}, \mathbf{p}_{-i}\rangle) = p_i\mu_i(\langle\mathbf{p}, \boldsymbol{\theta}_{i,0}\rangle),$$

and

$$\widehat{\text{rev}}_i^{(t)}(p_i|\mathbf{p}_{-i}) = p_i\mu_i(-\widetilde{\beta}_i^{(t)\text{PML}}p_i + \langle\widetilde{\boldsymbol{\gamma}}_i^{(t)\text{PML}}, \mathbf{p}_{-i}\rangle) = p_i\mu_i(\langle\mathbf{p}, \widetilde{\boldsymbol{\theta}}_i^{(t)}\rangle).$$

We recall the definition $g_i^{(t-1)} : \mathbb{R}^d \mapsto \mathbb{R}^d$:

$$g_i^{(t-1)}(\boldsymbol{\theta}_i) = \sum_{s=1}^{t-1} \mu_i(\mathbf{p}^{(s)\top}\boldsymbol{\theta}_i)\mathbf{p}^{(s)} + \lambda\boldsymbol{\theta}_i.$$

Let $J_i^{(t-1)}$ denote the Jacobian matrix of $g_i^{(t-1)}$. We have

$$J_i^{(t-1)}(\boldsymbol{\theta}_i) = \sum_{s=1}^{t-1} \mu_i'(\mathbf{p}^{(s)\top}\boldsymbol{\theta}_i)\mathbf{p}^{(s)}\mathbf{p}^{(s)\top} + \lambda I_N.$$

Thanks to the definition of the estimator $\widehat{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}$ defined in Equation (7), we have

$$g_i^{(t-1)}(\widehat{\boldsymbol{\theta}}_i^{(t)\mathtt{PML}}) = \sum_{s=1}^{t-1} \mathbf{p}^{(s)} y_i^{(s)}.$$

We also introduce the martingale

$$S_i^{(t-1)} = \sum_{s=1}^{t-1} \mathbf{p}^{(s)} \eta_i^{(s)}.$$

We define the $G_i^{(t-1)}(\boldsymbol{\theta}_{i,0}, \widetilde{\boldsymbol{\theta}}_i^{(t)})$ matrix as follows,

$$G_i^{(t-1)}(\boldsymbol{\theta}_{i,0}, \widetilde{\boldsymbol{\theta}}_i^{(t)}) = \int_0^1 J_i^{(t-1)}(u\boldsymbol{\theta}_{i,0} + (1-u)\widetilde{\boldsymbol{\theta}}_i^{(t)})du.$$

The Fundamental Theorem of Calculus gives

$$g_i^{(t-1)}(\boldsymbol{\theta}_{i,0}) - g_i^{(t-1)}(\widetilde{\boldsymbol{\theta}}_i^{(t)}) = G_i^{(t-1)}(\boldsymbol{\theta}_{i,0}, \widetilde{\boldsymbol{\theta}}_i^{(t)})(\boldsymbol{\theta}_{i,0} - \widetilde{\boldsymbol{\theta}}_i^{(t)}). \tag{21}$$

Knowing that both $\boldsymbol{\theta}_{i,0}$ and $\widetilde{\boldsymbol{\theta}}_i^{(t)}$ have an L2-norm smaller than B_{θ_i}, $\forall u \in [0,1]$, $\|u\boldsymbol{\theta}_{i,0} + (1-u)\widetilde{\boldsymbol{\theta}}_i^{(t)})\|_2 \leq B_{\theta_i}$. This implies in particular that

$$G_i^{(t-1)}(\boldsymbol{\theta}_{i,0}, \widetilde{\boldsymbol{\theta}}_i^{(t)}) \geq c_{\mu_i} \left(\sum_{s=1}^{t-1} \mathbf{p}^{(s)} \mathbf{p}^{(s)\top} + \frac{\lambda}{c_{\mu_i}} I_N \right) = c_{\mu_i} V_i^{(t-1)}, \tag{22}$$

which in turn ensures $G_i^{(t-1)}(\boldsymbol{\theta}_{i,0}, \widetilde{\boldsymbol{\theta}}_i^{(t)})$ is invertible. Let $\mathbf{p}$ be any price vector in $\mathcal{P}$ (possibly random) and t be a fixed time instant,

$$|\operatorname{rev}_i(p_i|\mathbf{p}_{-i}) - \widehat{\operatorname{rev}}_i^{(t)}(p_i|\mathbf{p}_{-i})| \leq p_i |\mu_i(\langle \mathbf{p}, \boldsymbol{\theta}_{i,0}\rangle) - \mu_i(\langle \mathbf{p}, \widetilde{\boldsymbol{\theta}}_i^{(t)}\rangle)| \tag{23}$$

where

$$\begin{aligned}
&|\mu_i(\langle \mathbf{p}, \boldsymbol{\theta}_{i,0}\rangle) - \mu_i(\langle \mathbf{p}, \widetilde{\boldsymbol{\theta}}_i^{(t)}\rangle)| \\
&\leq L_{\mu_i} |\langle \mathbf{p}, \boldsymbol{\theta}_{i,0} - \widetilde{\boldsymbol{\theta}}_i^{(t)}\rangle| \quad \text{(by Assumption 2.5)} \\
&= L_{\mu_i} |\mathbf{p}^\top G_i^{(t-1),-1}(\boldsymbol{\theta}_{i,0}, \widetilde{\boldsymbol{\theta}}_i^{(t)})(g_i^{(t-1)}(\boldsymbol{\theta}_{i,0}) - g_i^{(t-1)}(\widetilde{\boldsymbol{\theta}}_i^{(t)}))| \quad \text{(by (21))} \\
&\leq L_{\mu_i} \|\mathbf{p}\|_{G_i^{(t-1),-1}(\boldsymbol{\theta}_{i,0}, \widetilde{\boldsymbol{\theta}}_i^{(t)})} \|g_i^{(t-1)}(\boldsymbol{\theta}_{i,0}) - g_i^{(t-1)}(\widetilde{\boldsymbol{\theta}}_i^{(t)})\|_{G_i^{(t-1),-1}(\boldsymbol{\theta}_{i,0}, \widetilde{\boldsymbol{\theta}}_i^{(t)})} \quad \text{(by Cauchy-Schwartz inequality)} \\
&\leq \frac{L_{\mu_i}}{c_{\mu_i}} \|\mathbf{p}\|_{V_i^{(t-1),-1}} \|g_i^{(t-1)}(\boldsymbol{\theta}_{i,0}) - g_i^{(t-1)}(\widetilde{\boldsymbol{\theta}}_i^{(t)})\|_{V_i^{(t-1),-1}} \quad \text{(by (22))} \\
&\leq \frac{2L_{\mu_i}}{c_{\mu_i}} \|\mathbf{p}\|_{V_i^{(t-1),-1}} \|g_i^{(t-1)}(\boldsymbol{\theta}_{i,0}) - g_i^{(t-1)}(\widehat{\boldsymbol{\theta}}_i^{(t)})\|_{V_i^{(t-1),-1}} \quad \text{(by definition of } \widetilde{\boldsymbol{\theta}}_i^{(t)}).
\end{aligned}$$

Now we calculate

$$
\begin{aligned}
\|g_i^{(t-1)}(\boldsymbol{\theta}_{i,0}) - g_i^{(t-1)}(\widehat{\boldsymbol{\theta}}_i^{(t)})\|_{V_i^{(t-1),-1}} &\le \left\| \sum_{s=1}^{t-1} \mu_i(\mathbf{p}^{(s)\top}\boldsymbol{\theta}_{i,0})\mathbf{p}^{(s)} + \lambda\boldsymbol{\theta}_{i,0} - \sum_{s=1}^{t-1}\mathbf{p}^{(s)} y_i^{(s)} \right\|_{V_i^{(t-1),-1}} \\
&\le \left\| - \sum_{s=1}^{t-1} \mathbf{p}^{(s)}\eta_i^{(s)} + \lambda\boldsymbol{\theta}_{i,0} \right\|_{V_i^{(t-1),-1}} \\
&= \left\| -S_i^{(t-1)} + \lambda\boldsymbol{\theta}_{i,0} \right\|_{V_i^{(t-1),-1}} \\
&\le \|S_i^{(t-1)}\|_{V_i^{(t-1),-1}} + \|\lambda\boldsymbol{\theta}_{i,0}\|_{V_i^{(t-1),-1}} \quad \text{(Triangle inequality)} \\
&\le \|S_i^{(t-1)}\|_{V_i^{(t-1),-1}} + \sqrt{\lambda c_{\mu_i}}\|\boldsymbol{\theta}_{i,0}\|_2 \quad \left(V_i^{(t-1)} \ge \frac{\lambda}{c_{\mu_i}} I_N \right) \\
&\le \underbrace{L_{\mu_i}\sqrt{2\log(T/\delta) + N\log\left(1 + \frac{B_p^2(t-1)}{N\lambda}\right)} + B_{\theta_i}\sqrt{\lambda c_{\mu_i}}}_{=\mathrm{c}_i^{(t-1)}(\delta)} \quad \text{(w.p.} \ge 1-\delta).
\end{aligned}
$$

In the last inequality, we have used the assumption that $\|\boldsymbol{\theta}_{i,0}\|_2 \le B_{\theta_i}$ and the concentration result established in Proposition 7 of Russac et al. (2019) for the self-normalized quantity $\|S_i^{(t-1)}\|_{V_i^{(t-1),-1}}$ (which is applicable thanks to the conditional L_{μ_i}-subgaussianity of $\eta_i^{(t)}$ established in Lemma 2.6), which establishes that

$$
\mathbb{P}\left(\exists t \le T, \|S_i^{(t)}\|_{V^{(t)},-1} \ge \mathrm{c}_i^{(t)}(\delta)\right) \le \delta, \quad \text{equivalently} \quad \mathbb{P}\left(\|S_i^{(t)}\|_{V^{(t)},-1} \le \mathrm{c}_i^{(t)}(\delta), \forall t \le T\right) \ge 1-\delta.
$$

We established that, for any $\mathbf{p} \in \mathcal{P}$,

$$
\mathbb{P}\left(|\,\mathrm{rev}_i(p_i|\mathbf{p}_{-i}) - \widehat{\mathrm{rev}}_i^{(t)}(p_i|\mathbf{p}_{-i})| \le p_i\|\mathbf{p}\|_{V_i^{(t-1),-1}} \cdot \underbrace{\frac{2L_{\mu_i}}{c_{\mu_i}}\left(\mathrm{c}_i^{(t-1)}(\delta) + B_{\theta_i}\sqrt{\lambda c_{\mu_i}}\right)}_{=\rho_i^{(t-1)}(\delta)}, \quad \forall t \le T \right) \ge 1-\delta.
$$

This completes the proof.

C Proof of Theorem 4.1

We recall the definitions

$$
\mathrm{rev}_i(p_i|\mathbf{p}_{-i}) = p_i\mu_i(-\beta_{i,0}p_i + \langle\boldsymbol{\gamma}_{i,0}, \mathbf{p}_{-i}\rangle), \qquad \forall (p_i, \mathbf{p}_{-i}) \in \mathcal{P},
$$

and

$$
\widehat{\mathrm{rev}}_i^{(t)}(p_i|\mathbf{p}_{-i}) = p_i\mu_i(-\widetilde{\beta}_i^{(t)\mathtt{PML}}p_i + \langle\widetilde{\boldsymbol{\gamma}}_i^{(t)\mathtt{PML}}, \mathbf{p}_{-i}\rangle), \qquad \forall (p_i, \mathbf{p}_{-i}) \in \mathcal{P}.
$$

For simplicity of notation, we also define

$$
p_i^{(t)\star} = \Gamma_i(\mathbf{p}_{-i}^{(t)}) = \operatorname*{argmax}_{p_i \in \mathcal{P}_i} \mathrm{rev}_i(p_i|\mathbf{p}_{-i}^{(t)}), \qquad \forall \mathbf{p}_{-i}^{(t)} \in \mathcal{P}_{-i},
$$

and

$$
p_i^{(t)} = \widehat{\Gamma}_i^{(t)}(\mathbf{p}_{-i}^{(t-1)}) \in \operatorname*{argmax}_{p_i \in \mathcal{P}_i} \mathrm{UCB}_i^{(t)}(\cdot|\mathbf{p}_{-i}^{(t-1)}), \qquad \forall \mathbf{p}_{-i}^{(t-1)} \in \mathcal{P}_{-i},
$$

where we recall from Equation (10) that

$$
\mathrm{UCB}_i^{(t)}(p_i|\mathbf{p}_{-i}^{(t-1)}) = \widehat{\mathrm{rev}}_i^{(t)}(p_i|\mathbf{p}_{-i}^{(t-1)}) + \sigma_i^{(t-1)}(p_i, \delta).
$$

The goal is to upper bound the individual regret

$$\mathrm{Reg}_i(T) = \mathbb{E}\sum_{t=1}^{T} R_i^{(t)} = \mathbb{E}\sum_{t=1}^{T}[\mathrm{rev}_i(\Gamma_i(\mathbf{p}_{-i}^{(t)})|\mathbf{p}_{-i}^{(t)}) - \mathrm{rev}_i(p_i^{(t)}|\mathbf{p}_{-i}^{(t)})].$$

We have

$$R_i^{(t)} = \mathrm{rev}_i(\Gamma_i(\mathbf{p}_{-i}^{(t)})|\mathbf{p}_{-i}^{(t)}) - \mathrm{rev}_i(p_i^{(t)}|\mathbf{p}_{-i}^{(t)}) \le C_{i,1}(\Gamma_i(\mathbf{p}^{(t)}) - p_i^{(t)})^2, \tag{24}$$

where we used that $R_i^{(t)} \le M_i$ and that

$$\begin{aligned}|\partial^2_{p^2}\,\mathrm{rev}_i(p|\mathbf{p}_{-i}^{(t)})_{|p=p'}| &= |-2\beta_{i,0}\mu_i'(-\beta_{i,0}p' + \boldsymbol{\gamma}_{i,0}^\top\mathbf{p}_{-i}^{(t)}) + \beta_{i,0}^2 p'\mu_i''(-\beta_{i,0}p' + \boldsymbol{\gamma}_{i,0}^\top\mathbf{p}_{-i}^{(t)})| \\ &\le 2\beta_{i,0}L_{\mu_i} + \beta_{i,0}^2\bar{p}_i B_i'' \\ &\le 2\overline{\beta}_i L_{\mu_i} + \overline{\beta}_i^2\bar{p}_i B_i'' \triangleq C_{i,1},\end{aligned}$$

for some p' in the segment between the points $p_i^{(t)}$ and $\Gamma_i(\mathbf{p}_{-i}^{(t)})$. Continuing from (24) we have

$$\begin{aligned}(\Gamma_i(\mathbf{p}_{-i}^{(t)}) - p_i^{(t)})^2 &\le \|\boldsymbol{\Gamma}(\mathbf{p}^{(t)}) - \mathbf{p}^{(t)}\|_2^2 \\ &\le 2\|\boldsymbol{\Gamma}(\mathbf{p}^{(t)}) - \mathbf{p}^*\|_2^2 + 2\|\mathbf{p}^* - \mathbf{p}^{(t)}\|_2^2 \\ &= 2(\|\boldsymbol{\Gamma}(\mathbf{p}^{(t)}) - \boldsymbol{\Gamma}(\mathbf{p}^*)\|_2^2 + \|\mathbf{p}^* - \mathbf{p}^{(t)}\|_2^2), && [\boldsymbol{\Gamma}(\mathbf{p}^*) = \mathbf{p}^* \text{ by } \textit{Lemma } 2.9] \\ &\le 2(L_{\boldsymbol{\Gamma}}\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2 + \|\mathbf{p}^* - \mathbf{p}^{(t)}\|_2^2) && [\text{by } \textit{Lemma } 2.11] \\ &\le 2(L_{\boldsymbol{\Gamma}} + 1)\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2.\end{aligned} \tag{25}$$

STEP 1: Bound for $\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2$. Consider the event

$$\mathcal{G} = \cap_{t\in[T]}\left\{\|\widehat{\boldsymbol{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \boldsymbol{\Gamma}(\mathbf{p}^{(t-1)})\|_2^2 \le C_2\sigma^{(t-1)}(\delta)\right\}, \tag{26}$$

that has probability at least $1 - 2N\delta$ by Proposition 4.2. Continuing from (25), we get that, on $\mathcal{G}$,

$$\begin{aligned}\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2 &= \|\widehat{\boldsymbol{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \mathbf{p}^*\|_2 \\ &\overset{(\star)}{\le} \|\widehat{\boldsymbol{\Gamma}}^{(t)}(\mathbf{p}^{(t-1)}) - \boldsymbol{\Gamma}(\mathbf{p}^{(t-1)})\|_2 + \|\boldsymbol{\Gamma}(\mathbf{p}^{(t-1)}) - \boldsymbol{\Gamma}(\mathbf{p}^*)\|_2 \\ &\le \sqrt{C_2\sigma^{(t-1)}(\delta)} + L_{\boldsymbol{\Gamma}}\|\mathbf{p}^{(t-1)} - \mathbf{p}^*\|_2 \\ &\le \sqrt{C_2\sigma^{(t-1)}(\delta)} + L_{\boldsymbol{\Gamma}}\left(\sqrt{C_2\sigma^{(t-2)}(\delta)} + L_{\boldsymbol{\Gamma}}\|\mathbf{p}^{(t-2)} - \mathbf{p}^*\|_2\right) \\ &= \sqrt{C_2}\left(\sqrt{\sigma^{(t-1)}(\delta)} + L_{\boldsymbol{\Gamma}}\sqrt{\sigma^{(t-2)}(\delta)}\right) + L_{\boldsymbol{\Gamma}}^2\|\mathbf{p}^{(t-2)} - \mathbf{p}^*\|_2 \\ &= \sqrt{C_2}\left(\sum_{j=0}^{\ell} L_{\boldsymbol{\Gamma}}^j\sqrt{\sigma^{(t-j-1)}(\delta)}\right) + L_{\boldsymbol{\Gamma}}^{\ell}\|\mathbf{p}^{(t-\ell)} - \mathbf{p}^*\|_2, \qquad \ell \in \{1,\dots,t-1\} \\ &\le \sqrt{C_2}\left(\sum_{j=0}^{t-1} L_{\boldsymbol{\Gamma}}^j\sqrt{\sigma^{(t-j-1)}(\delta)}\right) + L_{\boldsymbol{\Gamma}}^{t-1}\|\mathbf{p}^{(0)} - \mathbf{p}^*\|_2.\end{aligned} \tag{27}$$

Summing over t we obtain, on $\mathcal{G}$, the following inequality

$$\begin{aligned}
\sum_{t=1}^{T} \|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2 &\le \sum_{t=1}^{T} \left\{ 2C_2 \left(\sum_{j=0}^{t-1} L_{\mathbf{\Gamma}}^j \sqrt{\sigma^{(t-j-1)}(\delta)} \right)^2 + 2L_{\mathbf{\Gamma}}^{2(t-1)} \|\mathbf{p}^{(0)} - \mathbf{p}^*\|_2^2 \right\} \\
&\overset{(\star\star)}{\le} \frac{2C_2}{(1-L_{\mathbf{\Gamma}})^2} \sum_{t=0}^{T-1} \sigma^{(t)}(\delta) + 2\left(\sum_{t=1}^{T} L_{\mathbf{\Gamma}}^{2(t-1)} \right) \|\mathbf{p}^{(0)} - \mathbf{p}^*\|_2^2 \\
&\le \frac{2C_2}{(1-L_{\mathbf{\Gamma}})^2} \sum_{t=0}^{T-1} \sigma^{(t)}(\delta) + \frac{2\|\mathbf{p}^{(0)} - \mathbf{p}^*\|_2^2}{1 - L_{\mathbf{\Gamma}}^2} \\
&= C_4 \sum_{t=0}^{T-1} \sum_{j\in[N]} 2\rho_j^{(t)}(\delta) p_j^{(t+1)} \|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t)},-1} + C_3, \qquad (28)
\end{aligned}$$

where $C_3 = 2\frac{\|\mathbf{p}^{(0)} - \mathbf{p}^*\|_2^2}{1-L_{\mathbf{\Gamma}}^2}$ and $C_4 = \frac{2C_2}{(1-L_{\mathbf{\Gamma}})^2}$. In $(\star\star)$ we used the following Lemma C.1, by substituting $q \leftarrow L_\Gamma$ and $a_k \leftarrow \sqrt{\sigma^{(k)}(\delta)}$ (the proof is deferred to Appendix D).

Lemma C.1. *Let $(a_k)_{k\ge 0}$ be any real sequence, $0 < q < 1$, and $T \ge 1$. Then*

$$\sum_{t=1}^{T} \left(\sum_{j=0}^{t-1} q^j a_{t-j-1} \right)^2 \le \frac{1}{(1-q)^2} \sum_{k=0}^{T-1} a_k^2.$$

STEP 2: Regret Bound. Consider the set $\mathcal{G}$ defined in (26) and the partitioning of the sampling space: $\mathcal{G} \cup \mathcal{G}^c$. We can write

$$\begin{aligned}
\mathbb{E}\left[R_i^{(t)}\right] &= \mathbb{E}\left[R_i^{(t)} \cdot (\mathbb{I}(\mathcal{G}) + \mathbb{I}(\mathcal{G}^c))\right] \\
&\le \mathbb{E}\left[R_i^{(t)} \cdot \mathbb{I}(\mathcal{G})\right] + M_i \mathbb{P}(\mathcal{G}^c) \\
&\le \mathbb{E}\left[R_i^{(t)} \cdot \mathbb{I}(\mathcal{G})\right] + M_i 2N\delta.
\end{aligned}$$

Where M_i is an upper bound of $R_i^{(t)}$. For example, using that $0 \le \mu_i \le B_i$ and $0 \le p_i \le \bar{p}_i$, we can take $M_i = 2B_i\bar{p}_i$. Then

$$\mathbb{E}\left[\sum_{t=1}^{T} R_i^{(t)}\right] \le \mathbb{E}\left[\sum_{t=1}^{T} R_i^{(t)} \cdot \mathbb{I}(\mathcal{G})\right] + M_i 2N\delta T.$$

Now, choosing $\delta = 1/T^\gamma$ with $\gamma > 1$, the second term converges to zero, so we only need to bound the first term on $\mathcal{G}$. Merging (24) and (25) we have that, on $\mathcal{G}$, up to a multiplicative constant C_4 and an additive constant C_3

$$\begin{aligned}
\mathrm{Reg}_i(T) = \mathbb{E}\left[\sum_{t=1}^{T} R_i^{(t)} \cdot \mathbb{I}(\mathcal{G})\right] &\lesssim \mathbb{E}\left[\sum_{t=1}^{T} \|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2 \cdot \mathbb{I}(\mathcal{G})\right] \\
&\overset{(28)}{\lesssim} \mathbb{E}\left[\sum_{t=0}^{T-1} \sum_{j\in[N]} \rho_j^{(t)}(\delta) p_j^{(t+1)} \|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t)},-1} \cdot \mathbb{I}(\mathcal{G})\right] \\
&\le \bar{p}\,\mathbb{E}\left[\sum_{t=0}^{T-1} \sum_{j\in[N]} \rho_j^{(t)}(\delta) \|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t)},-1} \cdot \mathbb{I}(\mathcal{G})\right], \qquad (29)
\end{aligned}$$

where $\bar{p} = \max_{j\in[N]} \bar{p}_j > 0$ and where we used that $\rho_j^{(t)}$ (defined in (11)) is increasing in t. Moreover, for $\delta = 1/T^\gamma$

with $\gamma > 1$, (retaining the dependence on λ, N and T), we have

$$\rho_j^{(t)}(\delta) = O\left(\mathrm{c}_i^{(t)}(\delta) + \sqrt{\lambda}\right) = O\left(\sqrt{\lambda} + L_{\mu_i}\sqrt{2\log\left(\frac{T}{\delta}\right) + N\log\left(1 + \frac{B_p^2 T}{N\lambda}\right)}\right)$$
$$= O\left(\sqrt{\lambda} + \sqrt{\log(T) + N\log\left(1 + \frac{T}{N\lambda}\right)}\right).$$

It only remains to bound

$$\sum_{t=0}^{T-1}\sum_{j\in[N]} \|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t),-1}},$$

for which we use the following Lemma C.2, proved in Appendix E.

Lemma C.2 (Variant of the Elliptical Potential Lemma)**.** *Retaining only the dependence on λ, N and T, it holds*

$$\sum_{t=0}^{T-1}\sum_{j\in[N]} \|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t),-1}} = O\left(\frac{N}{\sqrt{\lambda}} + N\sqrt{\lambda + T} + N^{3/2}\sqrt{T\log\left(\frac{T}{N\lambda} + 1\right)}\right).$$

We then conclude that the regret, in terms of λ, N and T is

$$\mathrm{Reg}_i(T) = O\left(\left[\sqrt{\lambda} + \sqrt{\log(T) + N\log\left(1 + \frac{T}{N\lambda}\right)}\right]\cdot\left[\frac{N}{\sqrt{\lambda}} + N\sqrt{\lambda + T} + N^{3/2}\sqrt{T\log\left(\frac{T}{N\lambda} + 1\right)}\right]\right). \quad (30)$$

As long as λ varies between $O(1/\sqrt{T})$ and $O(1)$, then

$$\mathrm{Reg}_i(T) = O\left(\sqrt{N\log(T)}\cdot N^{3/2}\sqrt{T\log(T)}\right) = O\left(N^2\sqrt{T}\log(T)\right).$$

Note that from (29) we can also recover convergence to NE.

$$\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{p}^{(t)} - \mathbf{p}^*\|_2^2] = O(N^2\sqrt{T}\log(T)).$$

This completes the proof.

D Proof of Lemma C.1

We want to prove that for a given sequence $(a_k)_{k\geq 0}$ of real values, $0 < q < 1$, and $T \geq 1$, then

$$\sum_{t=1}^{T}\left(\sum_{j=0}^{t-1} q^j a_{t-j-1}\right)^2 \leq \frac{1}{(1-q)^2}\sum_{k=0}^{T-1} a_k^2.$$

To show this, we expand the square and re-index with $k = t - j - 1$ and $m = t - \ell - 1$, which yields

$$\sum_{t=1}^{T}\left(\sum_{j=0}^{t-1} q^j a_{t-j-1}\right)^2 = \sum_{k=0}^{T-1}\sum_{m=0}^{T-1} a_k a_m \sum_{t=\max\{k,m\}+1}^{T} q^{2t-k-m-2}.$$

For $0 < q < 1$ we have

$$\sum_{t=\max\{k,m\}+1}^{T} q^{2t-k-m-2} = q^{2\max\{k,m\}-k-m}\cdot\sum_{r=0}^{T-\max\{k,m\}-1} q^{2r} \leq \frac{q^{2\max\{k,m\}-k-m}}{1-q^2}.$$

Hence

$$\sum_{t=1}^{T}\left(\sum_{j=0}^{t-1} q^j a_{t-j-1}\right)^2 \le \frac{1}{1-q^2}\sum_{k=0}^{T-1}\sum_{m=0}^{T-1} q^{2\max\{k,m\}-k-m} a_k a_m.$$

Observe that $q^{2\max\{k,m\}-k-m} = q^{|k-m|}$, so the right-hand side equals

$$\frac{1}{1-q^2}\mathbf{a}^\top K\mathbf{a}, \qquad K_{k,m} = q^{|k-m|}, \qquad \mathbf{a} = (a_0, \dots, a_{T-1})^\top.$$

By the Schur test,

$$\|K\|_2 \le \max_k \sum_{m=0}^{T-1} q^{|k-m|} \le 1 + 2\sum_{r=1}^{\infty} q^r = \frac{1+q}{1-q}.$$

Therefore,

$$\sum_{t=1}^{T}\left(\sum_{j=0}^{t-1} q^j a_{t-j-1}\right)^2 \le \frac{1}{1-q^2}\|K\|_2\|\mathbf{a}\|_2^2 \le \frac{1}{1-q^2}\cdot\frac{1+q}{1-q}\sum_{k=0}^{T-1} a_k^2 = \frac{1}{(1-q)^2}\sum_{k=0}^{T-1} a_k^2,$$

since $1-q^2 = (1-q)(1+q)$. This proves the claim.

E Proof of Lemma C.2

Recall from (8) that

$$V_j^{(t)} = \sum_{s=1}^{t} \mathbf{p}^{(s)}\mathbf{p}^{(s)\top} + \frac{\lambda}{c_{\mu_j}} I_N,$$

First, note:

$$(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)}) = (p_j^{(t)}, \mathbf{p}_{-j}^{(t)}) + (\underbrace{p_j^{(t+1)} - p_j^{(t)}}_{\Delta p_j^{(t)}}, \mathbf{0}_{-j}) = \mathbf{p}^{(t)} + \Delta p_j^{(t)}\mathbf{e}_j,$$

where $\mathbf{e}_j$ is the null vector with only entry 1 in the j-th coordinate. So:

$$\|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t),-1}} \le \|\mathbf{p}^{(t)}\|_{V_j^{(t),-1}} + \|\Delta p_j^{(t)}\mathbf{e}_j\|_{V_j^{(t),-1}} \le \|\mathbf{p}^{(t)}\|_{V_j^{(t),-1}} + 2\bar{p}_j\sqrt{[V_j^{(t),-1}]_{jj}}.$$

Hence

$$\sum_{t=0}^{T-1}\sum_{j\in[N]} \|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t),-1}} \le \underbrace{\sum_{j\in[N]}\sum_{t=0}^{T-1}\|\mathbf{p}^{(t)}\|_{V_j^{(t),-1}}}_{=A} + 2\bar{p}\underbrace{\sum_{j\in[N]}\sum_{t=0}^{T-1}\sqrt{[V_j^{(t),-1}]_{jj}}}_{=B},$$

where $\bar{p} = \max_{j\in[N]}\{\bar{p}_j\}$.

Bound B. Fix $j \in [N]$. Let $\underline{p} = \min_{j\in[N]} \underline{p}_j > 0$. Then for every $\mathbf{p} \in \mathcal{P}$ we have $\mathbf{p} \ge \underline{p}\mathbf{1}$ element-wise. For $t \ge 1$ and any $\mathbf{x} \in \mathbb{R}^N$,

$$\mathbf{x}^\top\left(\sum_{s=1}^{t}\mathbf{p}^{(s)}\mathbf{p}^{(s)\top}\right)\mathbf{x} = \sum_{s=1}^{t}(\mathbf{x}^\top\mathbf{p}^{(s)})^2 \ge \sum_{s=1}^{t}(\underline{p}x_j)^2 = t\underline{p}^2x_j^2 = \mathbf{x}^\top(t\underline{p}^2\mathbf{e}_j\mathbf{e}_j^\top)\mathbf{x},$$

where $\mathbf{e}_j$ is the j-th vector of the canonical euclidean basis in $\mathbb{R}^N$. For simplicity of notation, let $\lambda_j = \frac{\lambda}{c_{\mu_j}}$. We have

$$\sum_{s=1}^{t}\mathbf{p}^{(s)}\mathbf{p}^{(s)\top} \succeq t\underline{p}^2\mathbf{e}_j\mathbf{e}_j^\top \quad\Rightarrow\quad V_j^{(t)} = \lambda_j I_N + \sum_{s=1}^{t}\mathbf{p}^{(s)}\mathbf{p}^{(s)\top} \succeq \lambda_j I_N + t\underline{p}^2\mathbf{e}_j\mathbf{e}_j^\top.$$

Hence

$$V_j^{(t),-1} \preceq \left(\lambda_j I_N + t\underline{p}^2\mathbf{e}_j\mathbf{e}_j^\top\right)^{-1}.$$

And the (j,j) element satisfies

$$[V_j^{(t),-1}]_{jj} = \mathbf{e}_j^\top V_j^{(t),-1}\mathbf{e}_j \le \mathbf{e}_j^\top \left(\lambda_j I_N + t\underline{p}^2\mathbf{e}_j\mathbf{e}_j^\top\right)^{-1}\mathbf{e}_j = \frac{1}{\lambda_j + t\underline{p}^2}, \implies \sqrt{[V_j^{(t),-1}]_{jj}} \le \frac{1}{\sqrt{\lambda_j + t\underline{p}^2}}$$

Because $x \mapsto (\lambda_j + \underline{p}^2 x)^{-1/2}$ is decreasing on $[0,\infty)$,

$$\sum_{t=0}^{T-1}[V_j^{(t),-1}]_{jj}^{1/2} \le \frac{1}{\sqrt{\lambda_j}} + \int_0^T \frac{dx}{\sqrt{\lambda_j + \underline{p}^2 x}} = \frac{1}{\sqrt{\lambda_j}} + \frac{2}{\underline{p}^2}\left(\sqrt{\lambda_j + \underline{p}^2 T} - \sqrt{\lambda_j}\right).$$

Summing over $j \in [N]$ and retaining the dependence on λ, N and T we obtain $B = O\left(\frac{N}{\sqrt{\lambda}} + N\sqrt{\lambda + T}\right)$.

Bound A. We use the Elliptical Potential Lemma in Carpentier et al. (2020) Proposition 1 (by replacing $p \leftarrow 1$ and $\lambda \leftarrow \frac{\lambda}{c_{\mu_j}}$):

$$\sum_{j\in[N]}\sum_{t=0}^{T-1}\|\mathbf{p}^{(t)}\|_{V_j^{(t),-1}} \le \sum_{j\in[N]}\sqrt{TN\log\left(\frac{c_{\mu_j}T}{N\lambda}+1\right)} \le N\sqrt{TN\log\left(T\frac{c_\mu}{N\lambda}+1\right)}.$$

Merging the two bounds. Retaining only the dependence on λ, N and T, we recover

$$\sum_{t=0}^{T-1}\sum_{j\in[N]}\|(p_j^{(t+1)}, \mathbf{p}_{-j}^{(t)})\|_{V_j^{(t),-1}} = O\left(\frac{N}{\sqrt{\lambda}} + N\sqrt{\lambda+T} + N\sqrt{TN\log\left(\frac{T}{N\lambda}+1\right)}\right).$$

This completes the proof.

F Proof of Lemma 2.6

We want to prove that, under Assumption 2.5, for every $i \in [N]$, $\eta_i^{(t)}$ is L_{μ_i}-subgaussian conditionally on $\mathcal{H}_i^{(t-1)}$. Fix $i \in [N]$. We first prove that $\mathbb{E}[\eta_i^{(t)}|\mathcal{H}_i^{(t-1)}] = 0$, where

$$\eta_i^{(t)} = y_i^{(t)} - \mu_i(\langle \mathbf{p}^{(t)}, \boldsymbol{\theta}_{i,0}\rangle).$$

We recall that $p_i^{(t)}$ is computed using information $\mathcal{H}_i^{(t-1)} = \{(y_i^{(s)}, \mathbf{p}^{(s)})\}_{s\le t-1}$. Let

$$\mathcal{F}_i^{(t-1)} = \sigma\left(\mathcal{H}_i^{(t-1)}, \mathbf{p}^{(t)}\right).$$

Since $\mathcal{H}_i^{(t-1)} \subseteq \mathcal{F}_i^{(t-1)}$, we have by the law of total expectation

$$\begin{aligned}
\mathbb{E}[\eta_i^{(t)}|\mathcal{H}_i^{(t-1)}] &= \mathbb{E}[\mathbb{E}[\eta_i^{(t)}|\mathcal{F}_i^{(t-1)}]|\mathcal{H}_i^{(t-1)}] \\
&= \mathbb{E}[\mathbb{E}[y_i^{(t)} - \mu_i(\langle \mathbf{p}^{(t)}, \boldsymbol{\theta}_{i,0}\rangle)|\mathcal{F}_i^{(t-1)}]|\mathcal{H}_i^{(t-1)}] \\
&= \mathbb{E}[\mu_i(\langle \mathbf{p}^{(t)}, \boldsymbol{\theta}_{i,0}\rangle) - \mu_i(\langle \mathbf{p}^{(t)}, \boldsymbol{\theta}_{i,0}\rangle)|\mathcal{H}_i^{(t-1)}] \\
&= 0.
\end{aligned}$$

In the third equality we used that, by construction $y_i^{(t)}$ is independent of $\mathcal{H}_i^{(t-1)}$ given $\mathbf{p}^{(t)}$, i.e. $\mathbb{E}[y_i^{(t)} \mid \mathcal{H}_i^{(t-1)}, \mathbf{p}^{(t)}] = \mathbb{E}[y_i^{(t)} \mid \mathbf{p}^{(t)}] = \mu_i(\langle \mathbf{p}^{(t)}, \boldsymbol{\theta}_{i,0}\rangle)$. Now let $w_i^{(t)} \equiv \langle \boldsymbol{\theta}_{i,0}, \mathbf{p}^{(t)}\rangle$ (this is $\mathcal{F}_i^{(t-1)}$-measurable), and recall from (2) that we model

$$\frac{d\mathbb{P}_{\boldsymbol{\theta}_{i,0}}(y_i^{(t)}|\mathbf{p}^{(t)})}{d\nu_i(y_i^{(t)})} = \exp\left\{y_i^{(t)}w_i^{(t)} - b_i(w_i^{(t)}) + c_i(y_i^{(t)})\right\},$$

Then, with $\mu_i(w_i^{(t)}) = b_i'(w_i^{(t)})$ and $\eta_i^{(t)} = y_i^{(t)} - \mu_i(w_i^{(t)})$,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\eta_i^{(t)}}|\mathcal{F}_i^{(t-1)}\right] &= e^{-\lambda\mu_i(w_i^{(t)})}\int e^{\lambda y_i^{(t)}}\exp\left\{y_i^{(t)}w_i^{(t)} - b_i(w_i^{(t)}) + c_i(y_i^{(t)})\right\}d\nu_i(y_i^{(t)}) \\ &= e^{-\lambda\mu_i(w_i^{(t)})}e^{-b_i(w_i^{(t)})}\int \exp\left\{y_i^{(t)}(w_i^{(t)}+\lambda) + c_i(y_i^{(t)})\right\}d\nu_i(y_i^{(t)}) \\ &= e^{-\lambda\mu_i(w_i^{(t)})}e^{-b_i(w_i^{(t)})}e^{b_i(w_i^{(t)}+\lambda)} = \exp\left\{b_i(w_i^{(t)}+\lambda) - b_i(w_i^{(t)}) - b_i'(w_i^{(t)})\lambda\right\}.\end{aligned}$$

Taking logs gives the identity

$$\log\mathbb{E}\left[\exp\left(\lambda\eta_i^{(t)}\right)\middle|\mathcal{F}_i^{(t-1)}\right] = b_i(w_i^{(t)}+\lambda) - b_i(w_i^{(t)}) - b_i'(w_i^{(t)})\lambda.$$

By Taylor's expansion, there exists a $\xi \in [0,1]$ such that

$$b_i(w_i^{(t)}+\lambda) - b_i(w_i^{(t)}) - b_i'(w_i^{(t)})\lambda = \tfrac{1}{2}b_i''(w_i^{(t)}+\xi\lambda)\lambda^2 = \tfrac{1}{2}\mu_i'(w_i^{(t)}+\xi\lambda)\lambda^2.$$

Assumption 2.5 yields $\mu_i'(\cdot) \le L_{\mu_i}$ on $\mathbb{R}$, hence

$$\log\mathbb{E}\left[\exp\left(\lambda\eta_i^{(t)}\right)\middle|\mathcal{F}_i^{(t-1)}\right] \le \frac{L_{\mu_i}\lambda^2}{2} \quad\Longrightarrow\quad \mathbb{E}\left[\exp\left(\lambda\eta_i^{(t)}\right)\middle|\mathcal{F}_i^{(t-1)}\right] \le \exp\left(\frac{L_{\mu_i}\lambda^2}{2}\right).$$

And consequently

$$\mathbb{E}\left[\exp\left(\lambda\eta_i^{(t)}\right)\middle|\mathcal{H}_i^{(t-1)}\right] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\eta_i^{(t)}\right)\middle|\mathcal{F}_i^{(t-1)}\right]\middle|\mathcal{H}_i^{(t-1)}\right] \le \exp\left(\tfrac{L_{\mu_i}\lambda^2}{2}\right) \quad \text{a.s. for all } \lambda \in \mathbb{R}.$$

This proves that $\eta_i^{(t)}$ is conditionally sub-gaussian with variance proxy L_{μ_i}. This completes the proof.