ONLINE LEARNING IN QUOTA-CONSTRAINED ACQUISITION GAMES WITH TRANSIENT IMPACT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study strategic acquisition over a finite horizon T when every agent must fill an exact quota, aggregate demand raises the price, and part of this impact persists before decaying. Our question is what an agent can learn from public prices when the competitors' quotas and actions are hidden. Under a natural stability condition, this game is strongly convex and has a unique pure Nash equilibrium. We show that for every fixed transient impact parameter $\gamma < 1$, equilibrium schedules approach the uniform strategy and the price of anarchy approaches 1 as T grows.

Across K repeated episodes, the public price together with an agent's own schedule determine its cost gradient. Hence, decentralized online projected gradient descent achieves $O(\log K)$ regret and $O(1/K)$ mean-square last-iterate convergence without observing the number, quotas or schedules of the other agents. We obtain similar convergence guarantees for a Bayesian setting with (possibly correlated) private quotas. When the impact state is inherited across episodes, we characterize a self-consistent myopic equilibrium and its stochastic analogue.

We also study learning within a single episode. We prove an $\Omega(T)$ external-regret lower bound against the best schedule in hindsight, which motivates comparison against finite families of scheduling rules. We present a model-free bandit algorithm and show that it achieves $O(T^{2/3})$ policy regret and empirical schedules approach a coarse correlated equilibrium at rate $O(T^{-1/3})$.

1 INTRODUCTION

Consider N agents who must each acquire a prescribed quantity of a common resource within a deadline of T periods. Their combined demand in a period raises the price paid by everyone in that period, and part of this distortion persists before gradually decaying. Each agent must complete its quota exactly, knows neither how many competitors it faces nor how much they must acquire, and observes essentially one common signal: the public price. Scheduled acquisition of computational resources, energy, inventory or storage has this form. Our typical application is the execution of a large market order in online trading (see, e.g., Chriss, 2024a; Campbell & Nutz, 2026; Kearns & Shi, 2025) from which we adapt the basic model and sometimes borrow the terminology.

Two aspects make our acquisition model different from standard online learning and motivate its study. First, the constraint is a *quota* to be met, *not a cap* to be respected. Doing nothing is therefore infeasible: an action postponed must be taken before the deadline and actions across all periods are coupled. Second, the state is *endogenous*. The price that an agent faces in the future is determined by the past actions of all agents. If q_t denotes aggregate acquisition, I_t the impact state, and P_t the public price at time t, the per-unit price in our acquisition game is governed by

$$I_{t+1} = \gamma I_t + \lambda q_t \;\text{ and }\; P_t = I_t + \kappa q_t\,, \;\text{ where }\; \gamma \in (0, 1]. \tag{1}$$

Thus one unit acquired in period s increases the price in period $s + r$ by $\lambda\gamma^{r-1}$, in addition to the impact κ on the present price. Impact is *permanent* at $\gamma = 1$ and *transient* when $\gamma < 1$, i.e., old activity is eventually forgotten and the model's effective memory is governed by $(1 - \gamma)^{-1}$.

Main Question and Relation to Motivating Models. We investigate the following central question:

> *What can an agent learn from public prices in competitive acquisition environments with transient impact, when the competitors' quotas and actions are hidden?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To this end, we adopt a model that simplifies the strategic-acquisition models of Kearns & Shi (2025) and Hossain et al. (2026) in several respects, but employs transient impact. We consider N agents competing over a common time horizon of length T. Agent i must acquire a *quota* $B_i > 0$, choosing $x_{i,t} \geq 0$ units in each period t, subject to $\sum_t x_{i,t} = B_i$. Aggregate demand $q_t = \sum_i x_{i,t}$ determines instantaneous and persistent impact according to (1). In the baseline model, the common impact parameters $(\kappa, \lambda, \gamma)$ are fixed and publicly known; we treat noise and parameter error as extensions.

Relative to the models of Kearns & Shi (2025) and Hossain et al. (2026), we focus on buy-only actions and a fixed acquisition quota, and do not consider more general convex quota constraints and complex agent types. At the same time, we retain the key strategic interaction of acquisition games: agents jointly determine aggregate demand, which affects current and future prices, while the parameter γ controls the impact's persistence over time. We focus on buy-only actions for clarity; under our stability assumption and with bounded per-period sales, our learning results across episodes apply to buy-and-sell actions with the same rates and different, T-dependent constants (Proposition E.2).

So, rather than directly adopting their model, we focus on a mathematically cleaner abstraction that allows us to study (and to obtain strong and clean results in) two essentially orthogonal directions, both listed as directions for future work by Kearns & Shi (2025) and Hossain et al. (2026). The first deals with *how well (in terms of regret and convergence rate guarantees) agents can learn over repeated episodes of the acquisition game with a transient bounded-effective-memory state*. Our second direction deals with the more demanding question of *how well agents can learn to update their schedule within a single episode of the acquisition game with a transient state, when public price is the only available signal*. Table 3 in Appendix B briefly summarizes our results and techniques.

Contributions and Techniques. In Appendix C, we characterize the structure and the inefficiency of the equilibrium under complete information. We make the *stability assumption* $\mu_\gamma := \kappa - \frac{\lambda}{1+\gamma} > 0$, which makes the individual costs $2\mu_\gamma$-strongly convex and the pseudo-gradient μ_γ-strongly monotone. These directly imply existence and uniqueness of the pure Nash equilibrium. With equal quotas B and initial state $I_1 = 0$, each agent's equilibrium schedule approaches the uniform at rate $O(B/(T(1-\gamma)^{3/2}))$ in Euclidean norm (Theorem C.2). A similar analysis yields a price of anarchy of $1 + O(1/(T(1-\gamma)^4))$ (Theorem C.3). In continuous time, Campbell & Nutz (2026) show that the transient excess cost of anarchy vanishes as the product of the horizon and the decay rate grows, with equilibria approaching the uniform schedule; our results give discrete-time, buy-only analogues with explicit rates. These statements complement existing equilibrium analyses for transient impact (Strehle, 2017; Schied & Zhang, 2019; Luo & Schied, 2019; Ohnishi & Shimoshimizu, 2024).

In Appendix D, we study the *uniform policy*, which is oblivious to the game's environment and simply allocates a demand of B_i/T to each period. We show that the uniform policy is close to both the best response (against other uniform agents) and the equilibrium schedule (Theorem D.2). Moreover, its cost is robust to the timing of competing demand (Proposition D.1). The results above are formally stated and proven in Appendix C and Appendix D due to space restrictions.

Regarding learning across repeated episodes, we observe that the gradient of an agent's cost consists of the public price trajectory and information about its own schedule (see also (3)). Hence online projected gradient descent (OPGD) across episodes can be implemented without any knowledge about the number of agents, their quotas or their actions. (Hossain et al. (2026) also build gradients from prices, by reconstructing aggregate demand; here the gradient is exact given $(\kappa, \lambda, \gamma)$). Standard analysis (Theorem 3.1) shows that with a harmonic step size, OPGD achieves $O(\log K)$ regret against arbitrary sequences of competing schedules and $O(1/K)$ last-iterate convergence to equilibrium, when all agents implement the same algorithm with the same step size. If an upper bound on N is available, a constant step size gives geometric convergence to equilibrium. These bounds depend linearly on $s_T(\gamma) \leq \min\{T, (1-\gamma^2)^{-1}\}$. These results can be generalized to the case where private quotas are correlated and drawn i.i.d., for each episode, from a fixed unknown distribution. Then a Bayesian Nash equilibrium exists and can be reached at convergence rate of $O(1/K)$ (Theorem 3.2).

When the impact state is inherited from one episode to the next, a different schedule may leave a different state for future episodes. So the comparison used for external regret does not capture the full counterfactual effect of a deviation. For stationary play, we consider a notion of *self-consistent myopic equilibrium*. Agents best respond to the state at the beginning of the current episode, without assigning a continuation value to the state they leave behind. Under our stability assumption, this schedule-state system has a unique fixed point, which attracts repeated equilibrium play geometrically

(Appendix E.4). With i.i.d. private quotas, the same idea leads to a unique stationary distribution over inherited states. We also derive pseudo-regret and convergence guarantees if white noise is added to the price and/or to the impact state, if γ is not known and aggregate demand in each period is observed, and if the agents have inaccurate knowledge of the game's model in (1) (Appendix E.5).

We next switch to learning within a single episode of the acquisition game. We show (Theorem 4.1) that if an agent's quota is $\Theta(T)$, then no online learning (even randomized) algorithm can guarantee worst-case regret of $o(T)$. The proof is simple, uses only the quota constraint and holds even if $\lambda = 0$ (so the price trajectory is memoryless). The idea is that no online learning algorithm can distinguish between two different opponent profiles, one with zero demand throughout the horizon and another with zero demand in the first half and high demand in the second half.

We also study model-free learning within a single episode, where the agent does not know the impact model or future competing demand. Motivated by our lower bound and by the bounded effective memory $(1-\gamma)^{-1}$, we consider a weaker benchmark: a finite library of acquisition policies. We introduce PB-Exp3, a block-bandit algorithm that selects among these policies using only price feedback and guarantees exact quota completion. Against a bounded-memory response fixed in advance and under a feedback stability assumption, PB-Exp3 achieves $O(T^{2/3})$ policy regret against this library (Theorem 5.1). If all agents use synchronized PB-Exp3, we show that their expected empirical policy converges to a coarse correlated equilibrium of the induced finite policy game at rate $O(T^{-1/3})$ (Theorem 5.2). The key argument of the analysis uses the geometric decay of transient impact and the feedback stability assumption to control the difference between the realized trajectory and the counterfactual trajectory generated by an alternative rule.

Our experimental evaluation, in Section 6, complements our theoretical results in their different regimes. We compare Nash equilibrium and uniform execution as γ increases from the transient towards the permanent regime and evaluate both price-gradient and model-free learning under noise and model misspecification. We also present experimental results on real market data in Appendix I.

Relation to Previous Work. The closest papers on online learning for discrete acquisition games are Kearns & Shi (2025) and Hossain et al. (2026). The former studies a position-building game with temporary and permanent impact. They give an efficient best-response procedure and use no-regret dynamics with full action-profile feedback to obtain an approximate coarse correlated equilibrium. Hossain et al. (2026) study a considerably broader Bayesian strategic acquisition model and show how to learn an approximate Bayesian equilibrium from repeated price paths without a common prior. However, they focus on permanent impact, where $\gamma = 1$, and each round is an independent draw of the game instance, so no impact state carries over. For finite type spaces, they prove convergence to equilibrium with rate $\tilde{O}(K^{-1/2})$, assuming that the algorithm knows (or can estimate) κ and λ with accuracy $o(1/\log^2 K)$. They leave extensions to transient impact, where $\gamma < 1$, and to dynamic adjustment within episodes (i.e., essentially learning within an episode) as directions for future work. Appendix B.1 gives a thorough comparison to Kearns & Shi (2025) and Hossain et al. (2026).

Similar quota-constrained strategic acquisition models have received significant attention in the literature of optimal market-order execution. Optimal execution with temporary, permanent, and transient impact is a classical problem (Bertsimas & Lo, 1998; Almgren & Chriss, 2001; Obizhaeva & Wang, 2013; Gatheral, 2010). Strategic position building and execution have been studied, among others, under complete information by Chriss (2024c;b;a) and under incomplete information by Chriss (2025), while transient-impact games are analyzed in Schied & Zhang (2019); Schied et al. (2017); Strehle (2017); Luo & Schied (2019); Campbell & Nutz (2026); Ohnishi & Shimoshimizu (2024). This literature contains rich equilibrium characterizations, including oscillatory behavior when signed trades are allowed and transaction costs are below a critical level (Schied & Zhang, 2019). Our buy-only restriction simplifies the model and rules out round trips.

Methodologically, the repeated-episode results connect to learning in strongly monotone and aggregative games and to online convex optimization (Rosen, 1965; Koshal et al., 2016; Mertikopoulos & Zhou, 2019; Hazan et al., 2007; Duvocelle et al., 2023; Jordan et al., 2025). The single-episode result instead builds on learning with memory and policy regret (Auer et al., 2002; Arora et al., 2012; Cesa-Bianchi et al., 2013; Anava et al., 2015). The difference between these toolkits reflects the two timescales: after an episode we have a convex gradient problem, while during an episode we have a controlled, stateful counterfactual problem. A detailed and extensive comparison to previous work can be found in Appendix B.

2 Model, Notation and Preliminaries

For a positive integer n, let $[n] = \{1, \dots, n\}$. Throughout, vectors are bold italic, matrices upright, and scalars plain italic; e.g., $\boldsymbol{x}_i \in \mathbb{R}^T$ is a schedule, L_γ a memory matrix, and B_i the quota of agent i. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^T$, write $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_t u_t v_t$ and $\|\boldsymbol{u}\|_2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle^{1/2}$. Table 1 and Table 2, in Appendix A, summarize the notation used in this work.

The Acquisition Game. There are $N \geq 1$ agents and $T \geq 2$ periods. Agent i must acquire exactly $B_i > 0$ units (B_i is referred to as i's *quota*) and chooses a schedule

$$\boldsymbol{x}_i \in \mathcal{X}_i := \{\boldsymbol{x} \in \mathbb{R}_+^T : \mathbf{1}^\top \boldsymbol{x} = B_i\} = B_i \Delta_T\,, \quad \text{where } \Delta_T := \{\boldsymbol{p} \in \mathbb{R}_+^T : \mathbf{1}^\top \boldsymbol{p} = 1\}.$$

Nonnegativity excludes round trips and the equality requires completion by the deadline. We let $\mathrm{X} := (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)^\top$ be a schedule profile, let $q_t = \sum_i x_{i,t}$ and $\boldsymbol{q} = \sum_i \boldsymbol{x}_i$ be the aggregate demand (per time step and as a schedule, respectively), and let $V = \sum_i B_i$, $V_{-i} = V - B_i$, and $\mathcal{X} = \prod_i \mathcal{X}_i$.

The scalar state I_t records the impact of earlier activity on the price. Starting from $I_1 \geq 0$,

$$P_t = I_t + \kappa q_t \qquad \text{and} \qquad I_{t+1} = \gamma I_t + \lambda q_t. \tag{2}$$

Here P_t is the public price, $\kappa > 0$ measures the present demand's instantaneous impact, $\lambda > 0$ measures the increment to the persistent state due to the present demand, and $\gamma \in (0, 1]$ is the fraction of the state surviving one period. Impact is *transient* for $\gamma < 1$, when an action's effect decays geometrically and the effective memory is of order $(1-\gamma)^{-1}$; it is *permanent* at $\gamma = 1$.

All activity in period t is priced at P_t, and agent i aims to minimize $C_i(\boldsymbol{x}_i; \mathrm{X}_{-i}) = \sum_{t=1}^T x_{i,t} P_t$, where X_{-i} consists of the other agents' schedules. A profile $\mathrm{X}^\star \in \mathcal{X}$ is a *pure Nash equilibrium* if no agent can improve her cost by a unilateral deviation; formally, $C_i(\boldsymbol{x}_i^\star; \mathrm{X}_{-i}^\star) \leq C_i(\boldsymbol{x}_i; \mathrm{X}_{-i}^\star)$ for every agent i and every feasible schedule $\boldsymbol{x}_i \in \mathcal{X}_i$.

For convenience, we usually write $z := I_1 \geq 0$. Unrolling the state recursion in the game's price dynamics (2) and differentiating i's cost C_i w.r.t. $x_{i,t}$ gives (see also (13) in Appendix A):

$$g_{i,t} := \frac{\partial C_i}{\partial x_{i,t}} = P_t + \kappa x_{i,t} + \lambda \sum_{s>t} \gamma^{s-t-1} x_{i,s}\,. \tag{3}$$

Thus every term in i's gradient $g_{i,t}$ depending on the other agents is already accounted for in the public price P_t and the remaining terms depend only on i's own schedule $\boldsymbol{x}_i$.

In Appendix C.1 we show that for every $\boldsymbol{v} \in \mathbb{R}^T$, $\boldsymbol{v}^\top \nabla^2_{\boldsymbol{x}_i} C_i\, \boldsymbol{v} \geq 2\left(\kappa - \frac{\lambda}{1+\gamma}\right) \|\boldsymbol{v}\|_2^2$. Hence, we always assume that the price impact parameters κ, λ and γ satisfy the following *stability condition*:

$$\mu_\gamma := \kappa - \frac{\lambda}{1+\gamma} > 0. \tag{A}$$

(A) implies that the agents' individual costs are $2\mu_\gamma$-strongly convex and the game is μ_γ-strongly monotone, i.e., satisfies (4), and hence admits a unique equilibrium.

Theorem 2.1 (Strong Monotonicity and Unique Equilibrium)**.** *Let* $\mathrm{F}(\mathrm{X}) := (\boldsymbol{g}_1(\mathrm{X}), \dots, \boldsymbol{g}_N(\mathrm{X}))^\top$ *be the pseudo-gradient of the acquisition game. Under* (A)*, for every* N*,* T*, quota vector and initial state* $z \geq 0$*, the game has a unique pure Nash equilibrium* $\mathrm{X}^\star(z)$ *and, for all* $\mathrm{X}, \mathrm{Y} \in \mathcal{X}$*,*

$$\langle \mathrm{F}(\mathrm{X}) - \mathrm{F}(\mathrm{Y}), \mathrm{X} - \mathrm{Y} \rangle_{\mathrm{F}} \geq \mu_\gamma \big(\|\mathrm{X} - \mathrm{Y}\|_{\mathrm{F}}^2 + \|\boldsymbol{q}(\mathrm{X}) - \boldsymbol{q}(\mathrm{Y})\|_2^2 \big). \tag{4}$$

In (4), $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ and $\| \cdot \|_{\mathrm{F}}$ are the Frobenius inner product and norm. Equilibrium existence follows from the standard theory of convex games, and strong monotonicity makes it unique (Rosen, 1965). Existence and uniqueness have also been established for similar acquisition games (Strehle, 2017; Schied & Zhang, 2019; Luo & Schied, 2019), but with negative actions and different feasible sets. For completeness, we prove Theorem 2.1 in Appendix C.

The *(external) regret* of agent i with schedule sequence $\mathrm{X}_i = (\boldsymbol{x}_i^1, \dots, \boldsymbol{x}_i^K)$ against a sequence of aggregate schedules $\mathrm{Y} = (\boldsymbol{y}^1, \dots, \boldsymbol{y}^K)$ across $K \geq 1$ episodes of the acquisition game is:

$$\mathrm{Reg}_i(\mathrm{X}_i, \mathrm{Y}, K) := \sum_{k=1}^K C_i(\boldsymbol{x}_i^k; \boldsymbol{y}^k) - \min_{\boldsymbol{x}_i \in \mathcal{X}_i} \sum_{k=1}^K C_i(\boldsymbol{x}_i; \boldsymbol{y}^k).$$

We often write $\mathrm{Reg}_i(K)$, if the schedules are clear from the context. The comparator is the single best feasible schedule in hindsight, while the realized schedule sequence of the other agents is fixed. *Policy regret*, defined in (9), Section 5, instead regenerates the state and response under the comparator.

Algorithm 1 Projected price-gradient learning for agent i

Require: $\boldsymbol{x}_i^1 \in \mathcal{X}_i$, believed parameters $(\kappa, \lambda, \gamma)$, step sizes $(\eta_k)_{k\geq 1}$

1: **for** $k = 1, 2, \ldots, K$ **do**
2: implement $\boldsymbol{x}_i^k$ and observe the price path $\boldsymbol{P}^k = (P_1^k, \ldots, P_T^k)$ ▷ episode k is realized
3: $Z_{i,T}^k \leftarrow 0$; **for** $t = T-1, \ldots, 1$ **do**: $Z_{i,t}^k \leftarrow x_{i,t+1}^k + \gamma Z_{i,t+1}^k$ ▷ one backward pass
4: $\boldsymbol{g}_i^k \leftarrow \boldsymbol{P}^k + \kappa \boldsymbol{x}_i^k + \lambda \boldsymbol{Z}_i^k$; $\boldsymbol{x}_i^{k+1} \leftarrow \mathsf{P}_{\mathcal{X}_i}\big(\boldsymbol{x}_i^k - \eta_k \boldsymbol{g}_i^k\big)$
5: **end for**

3 Learning across Episodes from Public Prices

We now consider episodes $k = 1, \ldots, K$, each containing T acquisition periods. We begin with a fixed quota vector and a common initial state z in every episode. Agent i knows $(\kappa, \lambda, \gamma)$, its own quota B_i and its own schedule. At the end of episode k, it observes the public price vector $\boldsymbol{P}^k$, but it does not observe the number, quotas, states, or actions of the other agents. Using (3) (see also (13) in Appendix A) agent i can compute its exact gradient $\boldsymbol{g}_i^k = (g_{i,1}^k, \ldots, g_{i,T}^k)$ from public prices $\boldsymbol{P}^k = (P_1^k, \ldots, P_T^k)$ and its own schedule $\boldsymbol{x}_i^k$, which is enough for exact first-order learning:

$$g_{i,t}^k = P_t^k + \kappa x_{i,t}^k + \lambda \textstyle\sum_{s=t+1}^{T} \gamma^{s-t-1} x_{i,s}^k \,. \tag{5}$$

Namely, increasing $x_{i,t}^k$ has three effects on i's cost. It pays the current price P_t^k; it increases that price through impact by $\kappa x_{i,t}^k$; and it raises the future state paid on the agent's later actions. Competing demand and the initial state are accounted for in P_t^k; the discounted future sum is computed in Step 3, Algorithm 1. Projection, Step 4, enforces nonnegativity and exact quota completion for every episode.

Gradient Norm. In Proposition E.1, we show that for every feasible profile and initial state $z \geq 0$,

$$\|\boldsymbol{g}_i\|_2 \leq \mathcal{G}_i(z) := z\sqrt{s_T(\gamma)} + V_{-i}\sqrt{\kappa^2 + \lambda^2 s_T(\gamma)} + B_i\sqrt{4\kappa^2 + 2\lambda^2 s_T(\gamma)}\,, \tag{6}$$

where $s_T(\gamma) = \sum_{t=0}^{T-1} \gamma^{2t} \leq \min\{T, (1-\gamma^2)^{-1}\}$. The three terms bound, respectively, the entering state, the aggregate effect of the other agents, and the agent's own direct and future impact.

For fixed quotas and $\gamma < 1$, the gradient bound in (6) is independent of T, while at $\gamma = 1$, it is $O(\sqrt{T})$. For γ very close to 1, the gradient bound depends on how T compares against $1/(1-\gamma)$: $T(1-\gamma) \ll 1$ behaves as if $\gamma = 1$, while $T(1-\gamma) \gg 1$ behaves as if γ is bounded away from 1. In (6), dependence on the quotas of agent i and the other agents remains linear.

Theorem 3.1. *Assume that every episode begins with the same initial state z. Let $\eta_k = 1/(\mu_\gamma(k+1))$.*

1. *Agent i using Algorithm 1 alone satisfies against any sequence of competing schedules:*

$$\mathrm{Reg}_i(K) \leq \frac{\mathcal{G}_i(z)^2}{2\mu_\gamma}(1 + \log(K+1))\,. \tag{7}$$

2. *If all agents use Algorithm 1 with the same step size η_k, the distance to equilibrium $\mathrm{X}^\star$ is:*

$$\|\mathrm{X}^K - \mathrm{X}^\star\|_{\mathrm{F}}^2 \leq \frac{\max\big\{\|\mathrm{X}^1 - \mathrm{X}^\star\|_{\mathrm{F}}^2\,,\ \sum_i \mathcal{G}_i(z)^2/\mu_\gamma^2\big\}}{K}\,.$$

3. *If $\gamma < 1$ and an upper bound on the number N of agents is known, let $L_N = (N+1)(\kappa + \lambda/(1-\gamma))$. For step size $\eta = \mu_\gamma/L_N^2$, each agent has bounded cumulative regret and*

$$\|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}} \leq (1 - \mu_\gamma^2/L_N^2)^{(k-1)/2}\,\|\mathrm{X}^1 - \mathrm{X}^\star\|_{\mathrm{F}}.$$

The first guarantee is adversarial and individual: the competing schedules may change arbitrarily from one episode to the next and other agents need not learn at all. The comparator is the best single fixed feasible schedule in hindsight. The second and third guarantees require all agents to coordinate on their algorithm, which lets them converge simultaneously to the unique Nash equilibrium.

The proof of Theorem 3.1 can be found in Appendix E.2. It follows the standard analysis of online gradient descent for strongly convex losses (Hazan et al., 2007), for (7), and for projected learning in strongly monotone games (Mertikopoulos & Zhou, 2019). The model-specific part is that prices provide the exact gradient, the gradient norm is bounded as above, and that Theorem 2.1 establishes strong convexity of the individual cost and strong monotonicity of the game.

3.1 Private Quotas and Bayesian equilibrium

Next, we let the agent quota vectors be drawn i.i.d. from an unknown joint distribution $\mathcal{D}$ on $\prod_i \mathcal{B}_i$, allowing correlation among agents within an episode. Agent i observes only B_i^k, i.e., its quota for the present episode k. A Bayesian policy π_i maps each own quota $b \in \mathcal{B}_i$ to a feasible schedule $\pi_i(b) \in \mathcal{X}_i(b) := \{\boldsymbol{x}_i \in \mathbb{R}_+^T : \mathbf{1}^\top \boldsymbol{x}_i = b\}$. We write $\Pi = (\pi_1, \ldots, \pi_N)$ for the joint policy profile. Then, given a realized quota vector $\boldsymbol{B} = (B_1, \ldots, B_N)$, the corresponding execution profile is $\mathrm{X}^\Pi(\boldsymbol{B}) := \big(\pi_1(B_1), \ldots, \pi_N(B_N)\big)$. A policy profile Π is a *Bayesian Nash equilibrium* (BNE) if, for every i and every type $b \in \mathcal{B}_i$ that occurs with positive probability, $\pi_i(b)$ minimizes agent i's expected cost conditional on $B_i = b$. With $\mathcal{D}_i$ the marginal distribution of B_i, we define the probability-weighted distance $\big\|\Pi - \Pi'\big\|_{\mathcal{D}}^2 := \sum_i \sum_{b \in \mathcal{B}_i} \mathcal{D}_i(b) \big\|\pi_i(b) - \pi_i'(b)\big\|_2^2$.

For the Bayesian pseudo-gradient, we take the conditional expectation of the ordinary gradient $\boldsymbol{\Phi}_i(\Pi)(b) := \mathbb{E}\big[\boldsymbol{g}_i\big(\mathrm{X}^\Pi(\boldsymbol{B})\big) \,\big|\, B_i = b\big]$. Strong monotonicity of the realized game carries over to policies with the same constant μ_γ, which implies existence and uniqueness of a BNE (Theorem E.3).

One can show that Algorithm 1 with the unbiased gradient oracle $\widehat{\boldsymbol{g}}_{i,b}^k = \mathbb{1}\{B_i^k = b\}\, \boldsymbol{g}_i^k / \mathcal{D}_i(b)$ achieves pseudo-regret and convergence to BNE similar to Theorem 3.1 (Appendix E.3.2). To bound the expected *realized* regret and avoid the assumption that agents know their marginal distribution $\mathcal{D}_i$, we use an *empirical gradient estimator*. Specifically, every agent i maintains a count $n_{i,b}^{k-1}$ of how many times quota b has been observed before episode k. When $B_i^k = b$, the corresponding policy coordinate $\pi_i^{k+1}(b)$ is updated at the end of episode k to $\pi_i^{k+1}(b) = \mathsf{P}_{\mathcal{X}_i(b)}\big[\pi_i^k(b) - \boldsymbol{g}_i^k / (\mu_\gamma(n_{i,b}^{k-1} + 1))\big]$, while all other coordinates of π_i^{k+1} are copied from π_i^k. The following is proven in Appendix E.3.

Theorem 3.2. *The Bayesian game has a unique pure BNE $\Pi^\star$. If all agents use Algorithm 1 with the empirical estimator above in Step 4,*

$$\mathbb{E}\big[\|\Pi^K - \Pi^\star\|_{\mathcal{D}}^2\big] \le O\Big(\max\Big\{\|\Pi^1 - \Pi^\star\|_{\mathcal{D}}^2,\ \textstyle\sum_{i,b} \mathcal{G}_{i,b}^2 / \mu_\gamma^2\Big\}\Big) / K\,.$$

Expected regret against the best fixed policy in hindsight is $O(\log K)$.

3.2 Extensions: Inherited State, Noise and Parameter Uncertainty

It is natural that the terminal state of one episode becomes the initial state of the next. If z_k is the initial state of episode k, then $z_{k+1} = \gamma^T z_k + \lambda \sum_{t=1}^T \gamma^{T-t} q_t^k$. The gradient identity (5) still applies, because the realized state is included in the price vector. Moreover, for $\gamma < 1$ the reachable states are bounded. E.g., if aggregate quota is at most $V_{\max}$, then $z_k \le z_1 + \lambda V_{\max}/(1 - \gamma^T)$. Hence, external regret in Theorem 3.1 remains logarithmic, with $\mathcal{G}_i(z)$ equal to its maximum over this bounded interval of state values. But in the external regret bounds, when evaluating a deviation, we assume that its path is fixed. However, a different schedule in episode k would probably change z_{k+1} and therefore every later cost. External regret above does not recompute this counterfactual trajectory.

For stationary play, we characterize a *self-consistent myopic equilibrium*: a pair (X^*, z^*) such that X^* is the one-episode Nash equilibrium entered at z^* and results in z^* as its terminal state. The difference from a Markov-perfect equilibrium is that the agents "myopically" optimize the present episode given its entering state without attaching a continuation value to the state they leave behind.

We prove that, for $\gamma < 1$, a self-consistent myopic equilibrium exists, is unique, and is globally attracting under repeated equilibrium play, at a geometric rate $\chi_T < 1$ (Lemma E.7 and Remark E.8). For i.i.d. quotas, the inherited state is a random variable, because the equilibrium changes with the realized quotas in each episode. If, given the entering state, the agents play in each episode the BNE for i.i.d. private quotas, this random state process converges geometrically to a unique stationary distribution (Appendix E.4.1, this result concerns repeated equilibrium play, not Algorithm 1).

We also consider additive noise $P_t^k = I_t^k + \kappa q_t^k + \varepsilon_t^k$ and $I_{t+1}^k = \gamma I_t^k + \lambda q_t^k + \xi_{t+1}^k$, where $\{\varepsilon_t^k\}$ and $\{\xi_t^k\}$ are independent zero-mean processes with variance σ_ε^2 and σ_ξ^2 (Appendix E.5). The gradient is unbiased, with noise $\nu_t^k = \varepsilon_t^k + \sum_{s=2}^t \gamma^{t-s} \xi_s^k$. For fixed $\gamma < 1$, $\mathbb{E}[\|\boldsymbol{\nu}^k\|_2^2] \le T\sigma_\varepsilon^2 + \frac{T\sigma_\xi^2}{1-\gamma^2} := \sigma_\nu^2$. We prove we can obtain logarithmic pseudo-regret and convergence to equilibrium with $\mathcal{G}_i(z)^2$ replaced by $\mathcal{G}_i(z)^2 + \sigma_\nu^2$. If γ is unknown, but κ, λ and q_t are observable and there is no price noise, γ can be estimated by least squares regression, under sufficient variation in the state (see also Appendix E.5.1).

4 Linear External Regret for Single-Episode Learning

For learning within a single episode of the acquisition game, suppose agent i chooses $x_{i,t}$ from prices observed before period t and must satisfy its quota by T. We let $y_t = \sum_{j \neq i} x_{j,t}$ for the aggregate competing demand. We compare i's schedule with the best complete schedule in hindsight.

Theorem 4.1. *Suppose $z = 0$. For any even $T \geq 2$, any $(\kappa, \lambda, \gamma)$ such that $\mu_\gamma > 0$, any $b \geq 1$ and $B_i = bT$, there are two competing-demand vectors $\boldsymbol{y}^{(0)}$ and $\boldsymbol{y}^{(1)}$, with quota $O(T)$, such that the schedule $\boldsymbol{x}_i$ computed by any (even randomized) online learning algorithm has*

$$\frac{1}{2} \sum_{\omega \in \{0,1\}} \mathbb{E}\left[C_i(\boldsymbol{x}_i; \boldsymbol{y}^{(\omega)}) - \min_{\boldsymbol{u} \in \mathcal{X}_i} C_i(\boldsymbol{u}; \boldsymbol{y}^{(\omega)}) \right] \geq \frac{9\mu_\gamma B_i^2}{128\, T} = \frac{9\mu_\gamma b^2}{128}\, T\,.$$

We let $\boldsymbol{y}^{(0)}$ and $\boldsymbol{y}^{(1)}$ agree in the first $T/2$ periods. In $\boldsymbol{y}^{(0)}$, competing demand is zero throughout. In $\boldsymbol{y}^{(1)}$, it is zero in the first $T/2$ periods and constant and appropriately large in the remaining periods. In Appendix F, we show that the total regret of any fixed schedule $\boldsymbol{x}_i$ against both $\boldsymbol{y}^{(0)}$ and $\boldsymbol{y}^{(1)}$ is at least $9\mu_\gamma b^2 T/128$.

The construction above can also be regarded as a new agent entering halfway through i's acquisition horizon. By the time the opponent's demand becomes visible, the early part of the quota allocation is irreversible. Hence, obtaining strong worst-case regret and convergence guarantees requires that the agents' acquisition horizons are (at least approximately) synchronized.

5 Model-Free Policy Regret within One Episode

We next focus on learning within an episode based on the running cost of the schedule played so far. The agent does not know the parameters $(\kappa, \lambda, \gamma)$ or the competing quotas, and does not observe the competing demand. After period t, it observes only the price P_t and its cost $c_t = x_t P_t$. Due to Theorem 4.1, we adopt a weaker benchmark, the best scheduling rule chosen from a fixed finite library (with the benchmark policy run on its own counterfactual trajectory).

Response Maps and Learner's Policies. We fix an agent i and write $y_t = \sum_{j \neq i} x_{j,t}$ for the aggregate competing demand at period t. In this section only, we let $x_t \in [0, \bar{x}]$ and $y_t \in [0, \bar{y}]$. The competing demand is generated by *aggregate response maps* ψ_t with *memory* m, i.e., $y_t = \psi_t(x_{\max\{t-m,1\}}, \ldots, x_t)$, where the whole sequence $(\psi_t)_{t \leq T}$ is fixed before the episode and is thus oblivious to i's random choices, as for the bounded-memory adversaries of policy regret (Arora et al., 2012; Cesa-Bianchi et al., 2013). The maps ψ_t may be adversarial, nonlinear and time dependent.

The learner has a library $\mathcal{R} = \{\pi_1, \ldots, \pi_M\}$ consisting of deterministic scheduling policies. A scheduling policy may react to prices, but it does not observe any information about other agents.

The Algorithm and the Benchmark. PB-Exp3 (sketch in Algorithm 2) partitions the time horizon into K blocks of length τ each. We let $T = K\tau$ and $\beta = B/T \leq \bar{x}$. Each block has quota $\tau\beta$. PB-Exp3 first applies the uniform pace β for m periods. This *prefix* resets the finite memory of ψ_t. It then draws one policy according to the Exp3 distribution from $\mathcal{R}$, resets that policy's local memory, and follows it for the rest of the block. To guarantee complete quota execution in every block, if R units remain with s periods left, we project every policy's proposal onto the interval

$$\mathcal{I}(R, s) = \big[\max\{0, R - (s-1)\bar{x}\},\ \min\{\bar{x}, R\}\big]. \tag{8}$$

For each $j \in [M]$, π_j^K denotes the *complete T-period policy* induced by K copies of π_j. In every block, π_j^K follows the uniform pace for the first m periods, initializes a new copy of π_j for the suffix, and applies the projection (8). $J_T(\pi_j^K)$ denotes its total cost when the response and the impact state are regenerated from its own actions on the same initial state. The *policy regret* of an algorithm A is

$$\mathrm{PReg}_T(A) = \mathbb{E}[J_T(A)] - \min_{j \in [M]} J_T(\pi_j^K). \tag{9}$$

The policy regret benchmark does not contain arbitrary schedules or policies outside $\mathcal{R}$ and does not let the comparator change its policy from one block to the next.

Two copies of a policy starting from slightly different impact states may result in slightly different prices. Transient impact alone does not stop a discontinuous policy from amplifying this difference.

Algorithm 2 Pace-Banded Block Exp3 (PB-Exp3)

Require: Horizon $T = K\tau$, block length $\tau > m$, library $\mathcal{R} = \{\pi_1, \dots, \pi_M\}$ of scheduling policies, quota B, a suffix-loss interval $[a_\tau, a_\tau + R_\tau]$, and $\eta = \sqrt{2 \log M/(KM)}$

```
Initialize the Exp3 weights uniformly
for k = 1, ..., K do
    Play β = B/T during the first m periods of block k
    Draw J_k from the current Exp3 distribution, reset π_{J_k}, and set L_k ← 0
    for each remaining period of block k do
        Obtain a proposal from π_{J_k} using its realized (x_s, P_s) history
        Project it onto (8), play it, observe P_t and c_t = x_t P_t, and update L_k ← L_k + c_t
    end for
    Give Exp3 the normalized realized suffix cost min{1, (L_k − a_τ)_+ / R_τ} ∈ [0, 1]
end for
```

Thus, we assume that every policy in $\mathcal{R}$, together with the response process, is *feedback stable*, in the sense that if two coupled copies start a block suffix with the same quota and response history, but their initial states differ by d, then the sum of the absolute differences between their subsequent costs is at most Sd, for a constant S *independent of the suffix length* (Assumption G.2).

The analysis compares each realized block with the block that a fixed comparator π_j^K would have generated along its own counterfactual history. The common prefix removes the m-period memory of the response map, while transient impact and feedback stability bound the resulting difference in block costs by a constant D_γ independent of T and τ (see also the proof of Lemma G.4 for D_γ).

Theorem 5.1 (Policy Regret of PB-Exp3)**.** *Suppose that $\gamma < 1$, the competing process has memory m and every policy in $\mathcal{R}$ satisfies the feedback stability condition. Assume also that the realized cost $c_t = x_t P_t \in [0, \bar{c}]$, for some fixed $\bar{c}$. Then, for every block length $\tau > m$ dividing T, PB-Exp3 ensures that the quota is met by T and satisfies*

$$\mathrm{PReg}_T(\text{PB-Exp3}) \le \bar{c}\sqrt{2T\tau M \log M} + D_\gamma T/\tau\,. \tag{10}$$

Consequently, the model-free choice $\tau = \max\{m+1, \lceil T^{1/3} \rceil\}$ gives $\mathrm{PReg}_T(\text{PB-Exp3}) = O(T^{2/3})$.

For the proof, we regard each block of the time horizon as one bandit round. Conditional on the history before the block, every policy has a well-defined suffix cost. Thus, the standard analysis of Exp3 gives the first term in (10). This is initially a comparison based on the learner's realized trajectory. To obtain an upper bound on policy regret, we compare it with the trajectory generated by the comparator itself using transient impact and feedback stability (see Appendix G.6 for the details).

The restriction to $\gamma < 1$ is crucial. At $\gamma = 1$, state differences do not decay and the analysis fails to guarantee sublinear policy regret. Bounded price and state noise can also be included by evaluating the learner and the comparator under the same shocks. Details are given in Appendix G.7.

Simultaneous learning and CCE. Suppose that every agent i has a fixed, finite library $\mathcal{R}_i$ of policies, with the same properties as $\mathcal{R}$ before. The agents use the same block boundaries and, at the beginning of each block, select their policies simultaneously and independently, conditional on the past. Agents may have different quotas and policy libraries, but we require that each agent's quota and library, together with the within-block dynamics, remain unchanged across blocks.

In the finite *canonical policy game* $\mathcal{G}_\tau$, an action of agent i is a policy from $\mathcal{R}_i$. To define the cost of a joint policy profile, we run one block of length τ from a fixed reference state z° and take the expected *average cost per period* in that block. Actual blocks need not start from z°. Feedback stability allows us to couple every actual block to this reference game with a uniformly bounded error. The proof then combines the standard no-regret-to-CCE argument with this coupling step; see Appendix G.8.

Theorem 5.2 (CCE of the Policy Game)**.** *Let agent i have a library of M_i policies, and suppose that i's realized cost $c_t = x_t P_t \in [0, \bar{c}_i]$, for some fixed $\bar{c}_i$. Let $\widehat{\nu}_K$ be the empirical distribution of the joint policies selected in the $K = T/\tau$ blocks. Then $\mathbb{E}[\widehat{\nu}_K]$ is an $(\varepsilon_{i,K})_{i\in[N]}$-coarse correlated equilibrium of $\mathcal{G}_\tau$, where*

$$\varepsilon_{i,K} \le \bar{c}_i \sqrt{\tfrac{2M_i \log M_i}{K}} + \tfrac{2D_{i,\gamma}\, K}{T}\,. \tag{11}$$

For fixed policy libraries, $K = \Theta(T^{2/3})$ gives $\varepsilon_{i,K} = O(T^{-1/3})$.

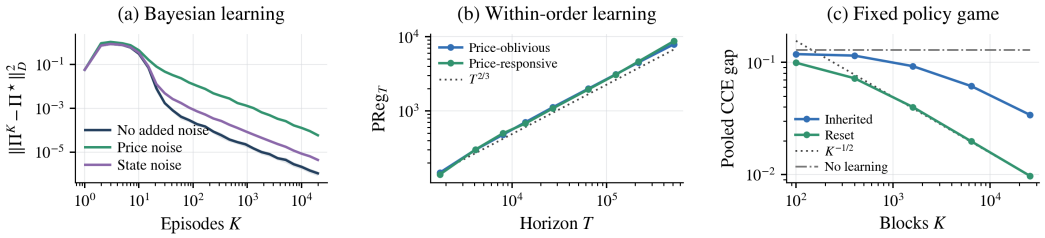


Figure 1: **(a)** Probability-weighted squared distance to BNE. **(b)** Total policy regret within one episode. The dotted line has slope $2/3$. **(c)** Pooled per-period CCE gap in a fixed policy game with reset or inherited impact. The dotted line has slope $-1/2$.

6 Experiments

We use controlled synthetic experiments to complement and validate our theoretical results. Unless stated otherwise, we use $N = 3$, $z = 0$, $T = 12$, and $(\kappa, \lambda, \gamma) = (1, 0.6, 0.7)$. All results are computed independently from their closed-form optimality conditions. Stochastic curves average 24 runs. Appendix H gives the complete protocols and additional structural checks. Appendix I gives a financial market execution study.

Learning across episodes: In Figure 1(a), every agent has a private quota in $\{1, 2\}$, drawn uniformly and independently in each episode. Each agent maintains one schedule per quota and updates only the realized-quota schedule after each episode. The local-clock price-gradient update (87) uses step size $1/(\mu_\gamma m)$ on the agent's m-th observation of that quota. The figure reports $\|\Pi^K - \Pi^\star\|^2_{\mathcal{D}}$ with no added noise, price noise of standard deviation 0.2, and state noise of standard deviation 0.05. Fitted slopes over $K \geq 1{,}000$ are respectively -0.96, -1.01, and -0.97. Noise changes the error at a given K, but the observed decay is consistent with the $O(K^{-1})$ guarantee.

Policy regret within one order (PB-Exp3): Figure 1(b) uses six fixed scheduling rules, one-period response memory, pace $B/T = 0.75$, and block length $\tau = T^{1/3}$. We test a first price-oblivious library and a second one containing two price-responsive rules. Across horizons $T = \tau^3$, the fitted exponents of total policy regret are 0.692 and 0.707. Both are sublinear and close to the $T^{2/3}$ upper bound, although the price-responsive case converges more slowly over the tested range.

Simultaneous learning: Finally, three agents run PB-Exp3 with four price-oblivious rules and common blocks of length $\tau = 9$. Figure 1(c) plots the CCE gap of the pooled empirical joint-policy distribution. When every block starts from the reference state, the gap falls from 0.0994 to 0.00974 and its tail slope is -0.51, consistent with the $K^{-1/2}$ bandit term, which becomes $O(T^{-1/3})$ under Theorem 5.2's $K = \Theta(T^{2/3})$. With inherited impact, it falls from 0.1186 to 0.0341, but more slowly. This difference comes from the fact that, at fixed τ, the state retains a coupling term of order $1/\tau$, whereas that term vanishes under blockwise state resets.

7 Conclusions and Open Questions

We present a rich and diverse collection of structural and learning results for strategic acquisition with transient impact. When the horizon exceeds the state's effective memory, equilibrium is close to the uniform schedule, the price of anarchy is close to 1, and gradient bounds stop growing with T. Across episodes, public prices allow for logarithmic regret and last-iterate convergence without observing the competitors. Within one episode, sublinear regret is impossible against the best schedule in hindsight. PB-Exp3 instead guarantees $O(T^{2/3})$ policy regret over a fixed finite library; under synchronized and stationary play, expected empirical policy approaches a CCE of the induced policy game.

The scale $T(1-\gamma)$ determines whether impact decay is visible within a horizon of length T. Transient impact can still behave as permanent if $T(1-\gamma) = o(1)$. Several questions remain. Can richer policy classes replace finite libraries while preserving exact quota completion? Can simple and easily verifiable conditions guarantee feedback stability for price-responsive policies? What guarantees survive when agents' horizons are not synchronized? Finally, our inherited-state equilibrium is myopic; efficient learning of a dynamic equilibrium that accounts for continuation values is open.

AI DISCLOSURE

The model formulation, the main ideas behind the results about existence, uniqueness and structure of equilibrium, price of anarchy, and online learning across episodes, and some preliminary ideas about learning within an episode predate the use of generative AI in this project. The authors used generative AI for checking several alternative approaches for learning within an episode and for the initial derivations of most of the calculations in appendices C, D, E and G. During our interaction with generative AI, it suggested online learning within an episode against a significantly restricted benchmark as a potential research direction. The authors used generative AI to improve their understanding of the literature, identify relevant papers and to prepare a preliminary draft of this work. The authors also used generative AI to assist with code development and debugging for the experiments and the visualization of results, while the authors reviewed, executed the code, and critically assessed the reported findings.

The authors have reviewed, explicitly verified and significantly improved the presentation of the material and the proofs in the main part and the appendices. The authors take full responsibility for the final content of this work, including the mathematical claims and the theoretical and experimental results, their verification and final presentation.

ETHICS STATEMENT

Our work is mostly theoretical, complemented by experiments with synthetic data, and studies strategic resource acquisition under transient impact and the resulting equilibrium inefficiency and learning. We do not anticipate direct ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

For our theoretical results, full proofs are included in the respective appendix of every section. Our modeling assumptions are explained in Section 2 and in Appendix A, while result-specific assumptions are stated in the statements of the corresponding theorems. All code used for the experiments is available at `https://anonymous.4open.science/r/ICLR39710`.

REFERENCES

Aurélien Alfonsi, Antje Fruth, and Alexander Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10(2):143–157, 2010. doi: 10.1080/14697680802595700. URL https://doi.org/10.1080/14697680802595700.

Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *The Journal of Risk*, 3(2):5–39, 2001. doi: 10.21314/JOR.2001.041. URL https://doi.org/10.21314/JOR.2001.041.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: Price of past mistakes. In *Advances in Neural Information Processing Systems*, volume 28, pp. 784–792, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/38913e1d6a7b94cb0f55994f679f5956-Abstract.html.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: From regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1503–1510, 2012. URL https://arxiv.org/abs/1206.6400.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Dimitris Bertsimas and Andrew W. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50, 1998. doi: 10.1016/S1386-4181(97)00012-8. URL https://doi.org/10.1016/S1386-4181(97)00012-8.

Kush Bhatia and Karthik Sridharan. Online learning with dynamics: A minimax perspective. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15020–15030, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/abb451a12cf1a9d93292e81f0d4fdd7a-Abstract.html.

Jean-Philippe Bouchaud, Yuval Gefen, Marc Potters, and Matthieu Wyart. Fluctuations and response in financial markets: The subtle nature of "random" price changes. *Quantitative Finance*, 4(2): 176–190, 2004. doi: 10.1080/14697680400000022. URL https://doi.org/10.1080/14697680400000022.

Mario Bravo, David S. Leslie, and Panayotis Mertikopoulos. Bandit learning in concave N-person games. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5666–5676, 2018. URL https://arxiv.org/abs/1810.01925.

Steven Campbell and Marcel Nutz. Optimal execution among N traders with transient price impact. *Mathematical Finance*, 2026. doi: 10.1111/mafi.70037. URL https://doi.org/10.1111/mafi.70037. Published online 22 June 2026; Early View. arXiv:2501.09638.

Nicolò Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, volume 26, pp. 1160–1168, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/9cf81d8026a9018052c429cc4e56739b-Abstract.html.

Neil A. Chriss. Competitive equilibria in trading. *arXiv preprint arXiv:2410.13583*, 2024a. URL https://arxiv.org/abs/2410.13583.

Neil A. Chriss. Position-building in competition with real-world constraints. *arXiv preprint arXiv:2409.15459*, 2024b. URL https://arxiv.org/abs/2409.15459.

Neil A. Chriss. Optimal position-building strategies in competition. *arXiv preprint arXiv:2409.03586*, 2024c. URL https://arxiv.org/abs/2409.03586.

Neil A. Chriss. Position building in competition is a game with incomplete information. *arXiv preprint arXiv:2501.01241*, 2025. URL https://arxiv.org/abs/2501.01241.

Qiwen Cui, Zhihan Xiong, Maryam Fazel, and Simon S. Du. Learning in congestion games with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23748–23762, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/47561f5e1dc53c7d119185e217b523d0-Abstract-Conference.html.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of Markov equilibrium in stochastic games. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4180–4234. PMLR, 2023. URL https://proceedings.mlr.press/v195/daskalakis23a.html.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo R. Jovanović. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5166–5220. PMLR, 2022. URL https://proceedings.mlr.press/v162/ding22b.html.

Jing Dong, Baoxiang Wang, and Yaoliang Yu. Convergence to Nash equilibrium and no-regret guarantee in (Markov) potential games. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2044–2052. PMLR, 2024. URL https://proceedings.mlr.press/v238/dong24a.html.

Benoît Duvocelle, Panayotis Mertikopoulos, Mathias Staudigl, and Dries Vermeulen. Multiagent online learning in time-varying games. *Mathematics of Operations Research*, 48(2):914–941, 2023. doi: 10.1287/moor.2022.1283. URL https://doi.org/10.1287/moor.2022.1283.

Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum Markov games. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9343–9373. PMLR, 2023. URL https://proceedings.mlr.press/v202/erez23a.html.

Roy Fox, Stephen M. McAleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in Markov potential games. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4414–4425. PMLR, 2022. URL https://proceedings.mlr.press/v151/fox22a.html.

Jim Gatheral. No-dynamic-arbitrage and market impact. *Quantitative Finance*, 10(7):749–759, 2010. doi: 10.1080/14697680903373692. URL https://doi.org/10.1080/14697680903373692.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000. doi: 10.1111/1468-0262.00153. URL https://doi.org/10.1111/1468-0262.00153.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2–3):169–192, 2007. doi: 10.1007/s10994-007-5016-8. URL https://doi.org/10.1007/s10994-007-5016-8.

Safwan Hossain, Mirah Shi, Andrew Bennett, Neil Andrew Chriss, Michael Kearns, Anderson Schneider, and Yuriy Nevmyvaka. Learning to strategically acquire resources in competition. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2606.06882.

Gur Huberman and Werner Stanzl. Price manipulation and quasi-arbitrage. *Econometrica*, 72(4):1247–1275, 2004. doi: 10.1111/j.1468-0262.2004.00531.x. URL https://doi.org/10.1111/j.1468-0262.2004.00531.x.

Michael I. Jordan, Tianyi Lin, and Zhengyuan Zhou. Adaptive, doubly optimal no-regret learning in strongly monotone and exp-concave games with gradient feedback. *Operations Research*, 73(3):1675–1702, 2025. doi: 10.1287/opre.2022.0446. URL https://doi.org/10.1287/opre.2022.0446. Published online in 2024.

Michael Kearns and Mirah Shi. Algorithmic aspects of strategic trading. *arXiv preprint arXiv:2502.07606*, 2025. URL https://arxiv.org/abs/2502.07606.

Khashayar Khosravi, Renato Paes Leme, Chara Podimata, and Apostolis Tsorvantzis. Preferences evolve and so should your bandits: Bandits with evolving states for online platforms. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 101. Association for Computing Machinery, 2024. doi: 10.1145/3670865.3673554. URL https://doi.org/10.1145/3670865.3673554.

Jayash Koshal, Angelia Nedić, and Uday V. Shanbhag. Distributed algorithms for aggregative games on graphs. *Operations Research*, 64(3):680–704, 2016. doi: 10.1287/opre.2016.1501. URL https://doi.org/10.1287/opre.2016.1501.

Raunak Kumar, Sarah Dean, and Robert D. Kleinberg. Online convex optimization with unbounded memory. In *Advances in Neural Information Processing Systems*, volume 36, pp. 38874–38914, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/531230cfac80c65017ad0f85d3031edc-Abstract-Conference.html.

Mingyang Liu, Asuman Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. Regret minimization with adaptive opponents in repeated games. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 4857–4858. PMLR, 2026. URL https://proceedings.mlr.press/v336/liu26c.html.

Xiangge Luo and Alexander Schied. Nash equilibrium for risk-averse investors in a market impact game with transient price impact. *Market Microstructure and Liquidity*, 5(1–4): 2050001, 2019. doi: 10.1142/S238262662050001X. URL https://doi.org/10.1142/S238262662050001X.

Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1–2):465–507, 2019. doi: 10.1007/s10107-018-1254-8. URL https://doi.org/10.1007/s10107-018-1254-8.

Eyal Neuman and Yufei Zhang. Statistical learning with sublinear regret of propagator models. *The Annals of Applied Probability*, 36(4):2873–2915, 2026. doi: 10.1214/25-AAP2155. URL https://doi.org/10.1214/25-AAP2155.

Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 673–680. Association for Computing Machinery, 2006. doi: 10.1145/1143844.1143929. URL https://doi.org/10.1145/1143844.1143929.

Anna A. Obizhaeva and Jiang Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1):1–32, 2013. doi: 10.1016/j.finmar.2012.09.001. URL https://doi.org/10.1016/j.finmar.2012.09.001.

Masamitsu Ohnishi and Makoto Shimoshimizu. Trade execution games in a Markovian environment. *Applied Mathematical Finance*, 31(4):239–277, 2024. doi: 10.1080/1350486X.2025.2479498. URL https://doi.org/10.1080/1350486X.2025.2479498. Published online 4 April 2025; volume dated 2024.

Beomsoo Park and Benjamin Van Roy. Adaptive execution: Exploration and learning of price impact. *Operations Research*, 63(5):1058–1076, 2015. doi: 10.1287/opre.2015.1415. URL https://doi.org/10.1287/opre.2015.1415.

J. B. Rosen. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica*, 33(3):520–534, 1965. doi: 10.2307/1911749. URL https://doi.org/10.2307/1911749.

Alexander Schied and Tao Zhang. A market impact game under transient price impact. *Mathematics of Operations Research*, 44(1):102–121, 2019. doi: 10.1287/moor.2017.0916. URL https://doi.org/10.1287/moor.2017.0916.

Alexander Schied, Elias Strehle, and Tao Zhang. High-frequency limit of Nash equilibria in a market impact game with transient price impact. *SIAM Journal on Financial Mathematics*, 8(1):589–634, 2017. doi: 10.1137/16M107030X. URL https://doi.org/10.1137/16M107030X.

Elias Strehle. Optimal execution in a multiplayer model of transient price impact. *Market Microstructure and Liquidity*, 3(3–4):1850007, 2017. doi: 10.1142/S2382626618500077. URL https://doi.org/10.1142/S2382626618500077.

Tatiana Tatarenko and Maryam Kamgarpour. Learning Nash equilibria in monotone games. In *2019 IEEE 58th Conference on Decision and Control*, pp. 3104–3109. IEEE, 2019. URL https://arxiv.org/abs/1904.01882.

Symbol	Meaning
N, T, K	number of agents, periods in an episode, and learning episodes (or blocks, in Section 5)
B_i, V, V_{-i}	quota of agent i, total quota $\sum_i B_i$, and competing quota $V - B_i$
$\kappa, \lambda > 0$	instantaneous-impact coefficient, state-impact coefficient
$\gamma \in (0, 1]$	surviving fraction of the state
z, I_t, P_t	entering state in the current episode, state and public price in period t
$\boldsymbol{x}_i, \boldsymbol{q}$	schedule of agent i and aggregate schedule $\sum_i \boldsymbol{x}_i$
$\mathcal{X}_i = B_i \Delta_T$	nonnegative schedules with coordinate sum B_i
$\mu_\gamma > 0$	stability and strong convexity coefficient $\mu_\gamma := \kappa - \lambda/(1+\gamma)$
$\mathrm{Reg}_i, \mathrm{PReg}_T$	external regret across episodes and policy regret within an episode
$\mathcal{R}, M$	finite policy library and its cardinality
m	response memory in Section 5
τ	block length in Section 5

Table 1: Scalar quantities, schedules and benchmarks.

A Notation and Frequently Used Identities

This appendix summarizes the notation used in the proofs. We recall that vectors are bold italic, matrices upright, and scalars plain italic; e.g., $\boldsymbol{x}_i \in \mathbb{R}^T$ is a schedule, L_γ a memory matrix, and B_i the quota of agent i. Unless stated otherwise, all vectors have T coordinates and all matrices are $T \times T$. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^T$, we write $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_t u_t v_t$ and $\|\boldsymbol{u}\|_2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle^{1/2}$. The vectors $\mathbf{1}$ and $\boldsymbol{e}_t$ are respectively the all-ones vector and the t-th standard basis vector; I_T is the identity matrix. Table 1 and Table 2 summarize our notation (all objects are also formally defined when they are first used).

We recall that all activity in period t is priced at P_t, and agent i aims to minimize

$$C_i(\boldsymbol{x}_i; \mathrm{X}_{-i}) = \textstyle\sum_{t=1}^{T} x_{i,t} P_t \,,$$

where X_{-i} consists of the other agents' schedules.

We also recall that for convenience, we let $z := I_1 \geq 0$ throughout. Unrolling the state recursion in the game's price dynamics (2), we obtain that:

$$I_t = \gamma^{t-1} z + \lambda \sum_{s<t} \gamma^{t-s-1} q_s \quad \text{and} \quad P_t = \gamma^{t-1} z + \lambda \sum_{s<t} \gamma^{t-s-1} q_s + \kappa q_t \,.$$

Differentiating i's cost C_i w.r.t $x_{i,t}$ and the form of P_t above gives:

$$\begin{aligned} g_{i,t} := \frac{\partial C_i}{\partial x_{i,t}} &= z\gamma^{t-1} + \lambda \sum_{s=1}^{t-1} \gamma^{t-s-1} q_s + \kappa q_t + \kappa x_{i,t} + \lambda \sum_{s=t+1}^{T} \gamma^{s-t-1} x_{i,s} \\ &= P_t + \kappa x_{i,t} + \lambda \sum_{s=t+1}^{T} \gamma^{s-t-1} x_{i,s} \,. \end{aligned} \tag{12}$$

The terms in (12) have a direct interpretation. Changing $x_{i,t}$ affects the current cost through the current price P_t and through the agent's self-impact $\kappa x_{i,t}$ on the current price. It also changes the impact state faced by the agent's future acquisitions, which gives the final discounted-future term. Interestingly, every term in i's gradient $g_{i,t}$ depending on the other agents is already accounted for in the public price P_t and the remaining terms depend only on i's own schedule $\boldsymbol{x}_i$.

We next derive a succinct description of the model in matrix form, which is useful in the proofs. We let L_γ be strictly lower triangular with $(\mathrm{L}_\gamma)_{ts} = \gamma^{t-s-1}$ for $s < t$, let $\mathrm{U}_\gamma = \mathrm{L}_\gamma^\top$ be its transpose, and let $\boldsymbol{d}_\gamma = (1, \gamma, \ldots, \gamma^{T-1})^\top$. Therefore, for any $\boldsymbol{v} \in \mathbb{R}^T$, $(\mathrm{L}_\gamma \boldsymbol{v})_t = \sum_{s<t} \gamma^{t-s-1} v_s$ is the discounted past of $\boldsymbol{v}$ and $(\mathrm{U}_\gamma \boldsymbol{v})_t = \sum_{s>t} \gamma^{s-t-1} v_s$ is its discounted future. Unrolling the price dynamics of the acquisition game in (2) as above, we obtain that:

$$\boldsymbol{I} = z\, \boldsymbol{d}_\gamma + \lambda\, \mathrm{L}_\gamma\, \boldsymbol{q} \qquad \text{and} \qquad \boldsymbol{P} = z\, \boldsymbol{d}_\gamma + \kappa \boldsymbol{q} + \lambda \mathrm{L}_\gamma \boldsymbol{q} \,.$$

where $q_t = \sum_i x_{i,t}$ and $\boldsymbol{q} = \sum_i \boldsymbol{x}_i$ denote the aggregate demand (per time step and as a schedule).

Symbol	Definition and interpretation
$\mathbf{1}, \boldsymbol{e}_t, \mathrm{I}_T$	all-ones vector, t-th basis vector and identity matrix
C_T	$\mathrm{I}_T - \mathbf{1}\mathbf{1}^\top/T$, the centering matrix; it subtracts the coordinate mean from a vector
$\mathrm{L}_\gamma, \mathrm{U}_\gamma$	$(\mathrm{L}_\gamma)_{ts} = \gamma^{t-s-1}$ for $s < t$, and $\mathrm{U}_\gamma = \mathrm{L}_\gamma^\top$; discounted past and future operators
$\boldsymbol{d}_\gamma, \boldsymbol{w}_\gamma$	$(1, \gamma, \ldots, \gamma^{T-1})^\top$ and its reversal $(\gamma^{T-1}, \ldots, 1)^\top$
$K_\gamma(\boldsymbol{x})$	$\boldsymbol{x}^\top \mathrm{L}_\gamma \boldsymbol{x} = \sum_{1 \le s < t \le T} \gamma^{t-s-1} x_s x_t$; the memory form measures the interaction of a schedule with its own past impact.
A_γ	$\kappa \mathrm{I}_T + \lambda \mathrm{L}_\gamma$, mapping aggregate demand to current price
H_γ	$2\kappa \mathrm{I}_T + \lambda(\mathrm{L}_\gamma + \mathrm{U}_\gamma)$, the Hessian of the individual cost function
$h_m(\gamma)$	$\sum_{r=0}^{m-1} \gamma^r = (1-\gamma^m)/(1-\gamma)$ for $\gamma < 1$, and m for $\gamma = 1$
$s_m(\gamma)$	$\sum_{r=0}^{m-1} \gamma^{2r} = (1-\gamma^{2m})/(1-\gamma^2)$ for $\gamma < 1$, and m for $\gamma = 1$
$\delta_T^2(\gamma)$	$s_T(\gamma) - h_T(\gamma)^2/T = \|\mathrm{C}_T \boldsymbol{d}_\gamma\|_2^2$, the squared norm of the decay profile after subtracting its coordinate mean.
$\mathsf{A}_T(\gamma)$	$\sum_{t=1}^{T} h_{t-1}(\gamma) \le \min\{T^2/2, T/(1-\gamma)\}$, cumulative geometric memory over one episode
$\langle \cdot, \cdot \rangle_{\mathrm{F}}, \|\cdot\|_{\mathrm{F}}$	Frobenius inner product and norm for schedule profiles represented as matrices

Table 2: Vectors, matrices, and geometric sums used in the proofs.

In this form, the individual cost of agent i can be written as:

$$C_i(\boldsymbol{x}_i; \mathrm{X}_{-i}) = z\boldsymbol{x}_i^\top \boldsymbol{d}_\gamma + \kappa \boldsymbol{x}_i^\top \boldsymbol{q} + \lambda \boldsymbol{x}_i^\top \mathrm{L}_\gamma \boldsymbol{q}\,.$$

Differentiating, we express i's gradient in our matrix form:

$$\begin{aligned} \boldsymbol{g}_i(\mathrm{X}) := \nabla_{\boldsymbol{x}_i} C_i(\mathrm{X}) &= \underbrace{z\boldsymbol{d}_\gamma + \kappa \boldsymbol{q} + \lambda \mathrm{L}_\gamma \boldsymbol{q}}_{=\boldsymbol{P}} + \kappa \boldsymbol{x}_i + \lambda \mathrm{U}_\gamma \boldsymbol{x}_i \\ &= \boldsymbol{P} + \kappa \boldsymbol{x}_i + \lambda \mathrm{U}_\gamma \boldsymbol{x}_i \end{aligned} \tag{13}$$

Interestingly, the public price $\boldsymbol{P}$ contains every opponent-dependent term; the remaining future-impact term uses only agent i's own schedule $\boldsymbol{x}_i$.

The Hessian of the individual cost is $\mathrm{H}_\gamma := 2\kappa\, \mathrm{I}_T + \lambda\, (\mathrm{L}_\gamma + \mathrm{U}_\gamma)$. In Appendix C.1, we prove that the minimum eigenvalue of H_γ is $\lambda_{\min}(\mathrm{H}_\gamma) \ge 2\big(\kappa - \frac{\lambda}{1+\gamma}\big) = 2\mu_\gamma$ (and that it converges to $2\mu_\gamma$ as $T \to \infty$). This justifies the stability condition $\mu_\gamma > 0$, stated in (A), which ensures the strong convexity and the strong monotonicity of the acquisition game.

In the analysis, we repeatedly use the geometric sums

$$h_m(\gamma) := \sum_{r=0}^{m-1} \gamma^r \le \min\{m, 1/(1-\gamma)\} \quad \text{and} \quad s_m(\gamma) := \sum_{r=0}^{m-1} \gamma^{2r} \le \min\{m, 1/(1-\gamma^2)\}$$

and the cumulative sum

$$\mathsf{A}_T(\gamma) := \sum_{t=1}^{T} h_{t-1}(\gamma) \le \min\{T^2/2, T/(1-\gamma)\}\,.$$

Intuitively, $h_m(\gamma)$ is the total weight assigned to m consecutive periods of unit demand by the impact state process. If we have a constant unit demand for m consecutive periods, $h_m(\gamma)$ is exactly the resulting increase on the price one period later, before any further decay. It is also, by the same reasoning, the total future exposure generated over the next m periods by a single unit acquired today. $\mathsf{A}_T(\gamma)$ is what you get if you let a uniform agent accumulate this exposure against itself, period after period, for the whole horizon.

These geometric sums explain why the permanent-impact factor T is often replaced under transient impact by $(1-\gamma)^{-1}$, or why $\sqrt{T}$ is replaced by $(1-\gamma^2)^{-1/2}$. These replacements are not uniform as γ increases towards 1. The product $T(1-\gamma)$ essentially determines the number of effective memory lengths included in each episode and plays a crucial role in the distance of the equilibrium to the uniform policy, the price of anarchy, and the dependence of our regret and convergence bounds in T.

Setting	Information and benchmark	Result	Main argument
Complete-information, transient impact	Known parameters, opponents and their quotas; pure Nash equilibrium and social optimum	Unique equilibrium; under transient impact, equilibrium approaches the uniform policy and PoA tends to 1	Memory form inequality, strong convexity and strong monotonicity
Learning across episodes	Own schedule and public prices; arbitrary competing schedules	$O(\log K)$ external regret; $O(1/K)$ last-iterate error under simultaneous learning	Getting gradient from prices; projected gradient descent
Private i.i.d. quotas	Own quota observed; opponents' quotas hidden	Convergence to BNE with similar guarantees	Sensitivity of variational inequalities and importance weighting
Inherited state	Terminal state becomes the next entering state	Self-consistent myopic fixed point; stationary state distribution consistent with BNE for i.i.d. quotas	Equilibrium sensitivity and contraction
Learning within one episode	Public prices and exact completion	Linear external-regret lower bound; $O(T^{2/3})$ policy regret for a finite policy library	Indistinguishable demand paths; PB-Exp3 with block synchronization among agents
Simultaneous policy learning	Stationary policy libraries and synchronized partitioning of the horizon into blocks	Empirical play approaches a CCE of the induced finite policy game	External regret plus coupling to a canonical block game

Table 3: Overview of the paper's questions, information structures, and proof ideas.

B Detailed Comparison to Related Work

Next, we elaborate on the positioning of our work, stated in Section 1. The closest papers overlap with different parts of our contribution. Hence, we separate four axes: impact dynamics, acquisition and position-building models, information revealed to an agent, the solution or learning concept.

B.1 Position Building and Strategic Resource Acquisition

Position-Building Models. The recent position-building papers of Chriss (2024c;b;a; 2025) provide the main game-theoretic by which (Kearns & Shi, 2025; Hossain et al., 2026) are motivated. Agents build long or short positions over a fixed horizon and incur temporary and permanent impact. Previous work in this line of research establishes equilibrium existence and structural characterizations, studies economically relevant constraints, and treats incomplete information. (Chriss, 2024c;b;a; 2025) already discuss much of the economic motivation for strategic resource acquisition and, in the market interpretation, for exact terminal positions.

Our work restricts attention to a finite-dimensional buy-only simplex, which removes round trips and the oscillatory signed strategies that are central in parts of the execution-game literature. Moreover, we take the equilibrium as a benchmark to be learned under restricted feedback. The opponents' quotas and schedules are not inputs to the agent's optimization problem anymore (see e.g., the Bayesian analysis in Section 3.1).

Online Learning and Position-Building. Kearns & Shi (2025) study algorithmic questions in a discrete version of the position-building game. A strategy specifies signed trades over a discrete horizon and must reach a fixed terminal position. They give a dynamic program for best response, show that the temporary-impact game is a potential game, decompose the general temporary–permanent game using potential and constant-sum components. They also present an efficient FTPL implementation to obtain no regret and converge to an approximate CCE.

For the results on online learning, every player observes the full joint action profile at the end of each episode. The game instance is the same across episodes, including terminal positions and constraints. Their impact state is the cumulative position generated by past trades, corresponding to permanent

impact (i.e., to $\gamma = 1$ in our model). Our repeated-execution learner instead observes only the realized price path, and Appendix E.3 allows the private acquisition budgets to vary.

We should also highlight that the CCE equilibria studied in (Kearns & Shi, 2025) and our work concern fundamentally different games and are incomparable. The CCE in (Kearns & Shi, 2025) is a distribution over complete schedules in the repeatedly played acquisition game. In our Theorem 5.2, a single execution partitioned into synchronized blocks, the actions are scheduling rules and the theorem's conclusion is an approximate CCE of the induced finite policy game.

Hossain et al. (2026) are closest to our repeated-learning setting. Their agents acquire a costly divisible resource over time subject to general convex constraints and aim to maximize a concave value minus acquisition cost. Agent types are very general and can encode constraints, valuations and market parameters. With a common prior, they prove existence, uniqueness, computability and price-of-anarchy results for the corresponding Bayesian acquisition game.

They then remove the common-prior assumption: across i.i.d. interactions, agents observe their realized types and price trajectories together with estimates of the market-impact coefficients. Because aggregate demand is not directly observed, it is reconstructed recursively from price increments and the coefficient estimates. The resulting gradient can be biased, and their last-iterate guarantee quantifies the effect of discretization and parameter-estimation error. Estimation of the game parameters within an additive error of $o(1/\log^2 K)$ is necessary for last-iterate convergence to equilibrium, while convergence rate is $O(L(\log K)^{-1} + K^{-1/2})$, where L is the Lipschitz constant of the type to strategy mapping.

Their permanent-impact parameters correspond to $\lambda = \alpha$ and $\kappa = \alpha + \beta$. Since they require $\beta \geq 0$, this gives $\kappa \geq \lambda$; our stability region at $\gamma = 1$ corresponds to the more relaxed requirement where $\kappa > \lambda/2$. Of course, the agent types, the individual cost functions and the strategy feasibility constraints in (Hossain et al., 2026) are substantially more general and their model is strictly more expressive than ours.

Our research directions and contributions can be regarded as orthogonal to those of Hossain et al. (2026). We focus on how well an agent can learn in strategic acquisition, across episodes and within a single episode, under transient impact. Our directions are closely related to two of the directions for further research proposed in (Hossain et al., 2026): transient propagators, which corresponds to the case where $\gamma < 1$ in our work, and dynamic adjustment within episodes, which is closely related to learning within an episode. Moreover, due to the clean, simple and stylized structure of our model, we prove convergence to the Bayesian Nash equilibrium for i.i.d. finite quota types at a rate of $O(1/K)$ from an exact importance-weighted price oracle.

We note that neither state inheritance nor within-execution policy regret have been considered in (Hossain et al., 2026). We deal with state inheritance using a self-consistent myopic target and, for i.i.d. quotas, a stationary state distribution that is consistent with the corresponding Bayesian Nash equilibrium. Our model-free bandit algorithm, in Section 5, does not assume any knowledge of the impact parameters and achieves sublinear policy regret against a fixed finite library of policies.

Regarding the price of anarchy, Hossain et al. (2026) show that without the buy-only restriction, it can be made arbitrarily large. Under the buy-only restriction, they obtain a bound that still grows polynomially in N and T. In the symmetric case with equal quotas, one can show (e.g., using standard techniques) that the price of anarchy under permanent impact is at most $2N/(N+1)$ for every T, with equality in the limit. Chriss (2024a) study a related continuous-time limit. In this work, we show that in the symmetric case, the price of anarchy under transient impact is $1 + O(1/(T(1-\gamma)^4))$.

In a nutshell, relative to Hossain et al. (2026), our contributions are transient state dynamics, the characterization of the transient price of anarchy, impact state inheritance, careful (and essentially tight) analysis of the dependence of our regret and convergence rate guarantees on T and $1-\gamma$ in the standard and the Bayesian setting, the information-theoretical lower bound on external regret for learning within an episode and the policy-regret guarantees for model-free learning within an episode.

B.2 Optimal Execution and Transient-Impact Games

Single-Agent Impact Models. Previous work on mathematical models for market execution usually distinguishes immediate trading costs from the effect of a trade on later prices. Early discrete-time

models include the permanent-impact formulation of Bertsimas & Lo (1998) and the temporary-permanent mean-variance framework of Almgren & Chriss (2001). Propagator models are motivated by the empirical observation that the response to previous orders decays with time (Bouchaud et al., 2004). Obizhaeva & Wang (2013) give a structural interpretation through a limit-order book that replenishes over time. Restrictions on impact kernels are not only technical; Huberman & Stanzl (2004); Gatheral (2010); Alfonsi et al. (2010) show that inappropriate kernels and impact functions can permit price manipulation.

Complete-Information Transient Acquisition Games. Schied & Zhang (2019) analyze two strategic agents who liquidate prescribed inventories on a common discrete grid under transient impact (cf. our impact state term multiplied by γ) and an additional quadratic transaction cost (cf. our quadratic term multiplied by κ, used essentially as a regularizer). They prove existence and uniqueness of equilibrium and identify the low-regularization regime, in which equilibrium trades alternate in sign, producing a "hot-potato" phenomenon. The high-frequency analysis of Schied et al. (2017) shows that these oscillations and the existence of a meaningful limit depend critically on the regularization. Strehle (2017) extend transient impact to multiple strategic investors and obtain existence, uniqueness, and closed-form equilibrium structure with quadratic trading costs. Luo & Schied (2019) incorporate risk aversion and give explicit equilibrium representations in related N-agent models.

Compared to this line of research, our work focuses on the finite buy-only setting, identifies and analyzes the horizon-uniform policy, which is robust w.r.t. individual costs and is quite close to equilibrium in Euclidean distance if $T(1-\gamma) = \Omega(1)$, and shows that under transient impact, the Price of Anarchy tends to 1 as the time horizon T grows. Restriction to buy-only schedules is important: our equilibrium may have inactive periods, but it cannot exhibit the signed round-trip oscillations studied in this line of work.

Campbell & Nutz (2026) provide a complementary continuous-time N-agent analysis in the model of Obizhaeva & Wang (2013). They prove existence and uniqueness of equilibrium under quadratic transaction costs and explain erratic behavior in discrete transient games, when the quadratic part diminishes. Their work motivates our stability assumption (A).

Ohnishi & Shimoshimizu (2024) are closer to our inherited-state discussion. They formulate a two-agent Markov acquisition game with a stochastic, sequentially dependent observed factor, and obtain a closed-form Markov-perfect strategy by backward induction. The expected future payoffs are taken into account in their equilibrium computation. In our setting of learning across episodes with inherited impact state (Appendix E.4), repeated equilibrium play converges to *self-consistent myopic equilibrium*, which does not account for future expected payoffs and is fundamentally different from Markov-perfect equilibrium. This is established through sensitivity bounds for the one-episode equilibrium response and contraction arguments for the induced state map.

B.3 Learning in Games with Aggregates or Evolving State

Monotone and Aggregative Games. The use of the pseudo-gradient and variational inequalities goes back to Rosen (1965). Gradient and mirror-based methods in monotone games provide convergence under deterministic, stochastic or bandit feedback, while distributed algorithms for aggregative games often estimate or communicate the aggregate (Koshal et al., 2016; Mertikopoulos & Zhou, 2019; Bravo et al., 2018; Tatarenko & Kamgarpour, 2019; Jordan et al., 2025). Time-varying monotone games lead naturally to path-variation and tracking guarantees (Duvocelle et al., 2023). Our analysis for learning across episodes, in Appendix E, is based on standard arguments from the analysis of online projected gradient descent. However, we analyze carefully the dependence of the gradient norms on T and γ, and the separation between arbitrary and i.i.d. private quotas.

General Markov and stochastic games are much more expressive than our linear state recursion, and decentralized equilibrium learning can be computationally or statistically difficult without additional structure. Positive results often exploit potential structure or target correlated notions (Daskalakis et al., 2023; Erez et al., 2023; Fox et al., 2022; Ding et al., 2022; Dong et al., 2024; Cui et al., 2022). Our game is strongly convex and strongly monotone, but it does not correspond to a potential game. Our results exploit this special structure and are neither implied nor related to known positive or negative results about equilibrium computation and learning in Markov games.

Online Learning with Memory and Dynamics. In online convex optimization with memory, current loss depends on a finite window of earlier decisions (Anava et al., 2015). Online learning with dynamics and its unbounded-memory extensions compare policies through the states induced by their actions (Bhatia & Sridharan, 2020; Kumar et al., 2023). Khosravi et al. (2024) study a single-agent bandit problem with a deterministically evolving, unobserved state and show that a sequence comparator can be more meaningful than a fixed arm. Our setting shares the central difficulty that actions have delayed effects, but the impact state is jointly created by several strategic agents and exact (and buy-only) acquisition strongly couples all periods across the horizon.

Policy regret makes this counterfactual explicit. Instead of evaluating a comparator on the losses generated by the learner's realized history, it evaluates the history that the comparator itself would have induced (Arora et al., 2012; Cesa-Bianchi et al., 2013). Against an unrestricted adaptive environment, sublinear policy regret is impossible in general, which motivates restriction to a finite library of feedback stable policies in Section 5. We impose a bounded memory restriction on the opponent-response process. The price process may itself have longer memory, but for fixed $\gamma < 1$, its effect contracts geometrically and can be coupled explicitly.

The block reduction behind Theorem 5.1 follows the broad policy-regret principle of mini-batching an adversarial bandit algorithm, here Exp3 (Auer et al., 2002; Arora et al., 2012). The execution-specific work is to make every counterfactual complete exactly the same quota, to erase both response and rule memories, and to retain the contracted impact-state error in the coupling constant. This yields the balance $\sqrt{T\tau} + T/\tau$ and the optimized $T^{2/3}$ rate. At $\gamma = 1$ the residual state discrepancy is not forgotten, so the same proof does not give a sublinear guarantee. Recent work also studies policy regret in games against adaptive opponents (Liu et al., 2026).

Equilibrium Implications of Regret. External no-regret play in a fixed finite game implies convergence of the empirical distribution to CCE (Hart & Mas-Colell, 2000). We use precisely this property, after defining a canonical stationary block game over execution rules. We should highlight that: (i) the convergence is of expected empirical play, not of the last iterate; (ii) deviations select another fixed rule from the library and are not arbitrary schedules; and (iii) the opponents' future Exp3 weights are not recomputed under the deviation. The result is therefore a CCE of the induced policy game, where the agent pure strategies correspond to the members of an agent-dependent fixed finite library with scheduling rules.

B.4 Learning for Market Order Execution

There is a substantial single-agent literature on learning for market order execution. Reinforcement-learning methods treat the market as an environment and learn a policy from state and reward observations (Nevmyvaka et al., 2006). Park & Van Roy (2015) combine exploration with execution when the price impact is unknown. More recently, Neuman & Zhang (2026) study online liquidation with an unknown transient propagator and temporary impact, obtain statistical estimates of the impact law, and prove sublinear regret in a single-agent setting.

Our setting differs in that the opponents are strategic and the relevant benchmarks are therefore game-theoretic (Nash equilibrium, Bayesian Nash equilibrium and coarse correlated equilibrium), rather than a single-agent optimum. With known impact parameters, the public price directly reveals the information needed for gradient learning, an agent need not explore in order to estimate its gradient or observe its opponents' actions or quotas. When γ is unknown, Appendix E.5.1 separates learning from parameter identification and states the additional volume observation and excitation required for the latter.

C Properties of the Acquisition Game and Equilibrium Analysis

In this section we establish the main properties of the acquisition game with transient state impact, stated and briefly discussed in Section 2, and also analyze the structure of the symmetric equilibrium and its price of anarchy.

We first isolate a property of the exponential memory kernel that is used throughout. Although the quadratic form of the exponential memory kernel need not be positive, its negative part is uniformly bounded (Lemma C.1). We then prove strong monotonicity, characterize the symmetric equilibrium

through its KKT conditions, and derive distance-to-uniform and price-of-anarchy bounds. The entering state z is retained in the equilibrium calculations and set to zero only for the price of anarchy.

We let $\boldsymbol{I} = (I_1, \ldots, I_T)^\top$ denote the state vector along times $t = 1, \ldots, T$, and let $\boldsymbol{d}_\gamma = (1, \gamma, \ldots, \gamma^{T-1})^\top$ is the memory impact to the particular timestep. We let $\mathbb{1}\{s < t\}$ be 1, if $s < t$, and 0 otherwise. From Appendix A, we recall that

$$\boldsymbol{I} = z\boldsymbol{d}_\gamma + \lambda \mathrm{L}_\gamma \boldsymbol{q} \qquad \text{and} \qquad C_i(\mathrm{X}) = \boldsymbol{x}_i^\top (z\boldsymbol{d}_\gamma + \lambda \mathrm{L}_\gamma \boldsymbol{q} + \kappa \boldsymbol{q}),$$

where $\boldsymbol{q}$ is the vector of aggregate demand over the time horizon and L_γ is the strictly lower-triangular matrix $(\mathrm{L}_\gamma)_{ts} = \gamma^{t-s-1}\mathbb{1}\{s < t\}$. We also let $\mathrm{U}_\gamma = \mathrm{L}_\gamma^\top$ be its transpose.

C.1 The Memory Form and Strong Convexity of the Individual Cost

For $\boldsymbol{v} \in \mathbb{R}^T$, define the *memory form* of the acquisition game with transient impact:

$$K_\gamma(\boldsymbol{v}) := \boldsymbol{v}^\top \mathrm{L}_\gamma \boldsymbol{v} = \sum_{t=1}^{T} \sum_{s=1}^{t-1} \gamma^{t-s-1} v_s v_t = \sum_{1 \le s < t \le T} \gamma^{t-s-1} v_s v_t. \tag{14}$$

The matrix L_γ is not symmetric, but the scalar quadratic form is unchanged by transposition. Therefore, $(\boldsymbol{v}^\top \mathrm{L}_\gamma \boldsymbol{v})^\top = \boldsymbol{v}^\top \mathrm{U}_\gamma \boldsymbol{v} = K_\gamma(\boldsymbol{v})$.

Lemma C.1 (Memory Inequality)**.** *For every $\boldsymbol{v} \in \mathbb{R}^T$ and every $\gamma \in (0, 1]$,*

$$K_\gamma(\boldsymbol{v}) \ge -\frac{1}{1+\gamma} \|\boldsymbol{v}\|_2^2. \tag{15}$$

Moreover, for $\gamma < 1$, we have that

$$K_\gamma(\boldsymbol{v}) \le \frac{1}{1-\gamma} \|\boldsymbol{v}\|_2^2. \tag{16}$$

Proof. Set $p_1 = 0$ and $p_{t+1} = \gamma p_t + v_t$, for all $t \in [T]$. Then $p_t = (\mathrm{L}_\gamma \boldsymbol{v})_t$ and $K_\gamma(\boldsymbol{v}) = \sum_t p_t v_t$. Since $v_t = p_{t+1} - \gamma p_t$,

$$\begin{aligned}(1+\gamma) p_t v_t + v_t^2 &= (p_{t+1} - \gamma p_t)(p_{t+1} + p_t) \\ &= p_{t+1}^2 + (1-\gamma) p_t p_{t+1} - \gamma p_t^2.\end{aligned}$$

Summing this identity over all $t \in [T]$, we obtain that:

$$(1+\gamma) K_\gamma(\boldsymbol{v}) + \|\boldsymbol{v}\|_2^2 = \frac{1+\gamma}{2} p_{T+1}^2 + \frac{1-\gamma}{2} p_2^2 + \frac{1-\gamma}{2} \sum_{t=2}^{T} (p_t + p_{t+1})^2 . \tag{17}$$

For the derivation of (17), we use that $p_1 = 0$, that

$$\sum_{t=1}^{T} p_{t+1}^2 - \gamma \sum_{t=1}^{T} p_t^2 + (1-\gamma) \sum_{t=1}^{T} p_t p_{t+1} = p_{T+1}^2 + (1-\gamma) \sum_{t=2}^{T} \big(p_t^2 + p_t p_{t+1}\big),$$

because the difference of the two sums of squares telescopes, and that

$$p_t^2 + p_t p_{t+1} = \tfrac{1}{2}(p_t + p_{t+1})^2 + \tfrac{1}{2}(p_t^2 - p_{t+1}^2).$$

We also observe that when we sum the r.h.s. over all $t \in [T]$, the sum corresponding to the second term telescopes.

We observe that every term on the r.h.s of (17) is nonnegative, which concludes the proof of (15).

To see why the lower bound of (15) cannot be improved, we consider an even T and let $v_t = (-1)^t$. Then $\|\boldsymbol{v}\|_2^2 = T$ and

$$\frac{K_\gamma(\boldsymbol{v})}{\|\boldsymbol{v}\|_2^2} = \frac{1}{T} \sum_{\ell=1}^{T-1} (T - \ell)(-1)^\ell \gamma^{\ell-1} \longrightarrow -\frac{1}{1+\gamma} \quad \text{as } T \longrightarrow \infty.$$

Finally, for $\gamma < 1$, apply $2v_s v_t \le v_s^2 + v_t^2$ to (14). For any coordinate t, the total geometric weight assigned to the coordinates before and after t is bounded by

$$\sum_{s<t} \gamma^{t-s-1} \le \frac{1}{1-\gamma} \qquad \text{and} \qquad \sum_{s>t} \gamma^{s-t-1} \le \frac{1}{1-\gamma},$$

since both are finite partial sums of the geometric series. The factor $1/2$ therefore gives (16). □

We next use Lemma C.1 in order to show that the individual cost is $2\mu_\gamma$-strongly convex in the agent's schedule. To this end, we fix the other agents' aggregate schedule $\boldsymbol{y} = \sum_{j \neq i} \boldsymbol{x}_j$. Then agent i's cost can be written as

$$C_i(\boldsymbol{x}_i; \boldsymbol{y}) = \frac{1}{2}\boldsymbol{x}_i^\top \mathrm{H}_\gamma \boldsymbol{x}_i + \langle z\boldsymbol{d}_\gamma + \kappa \boldsymbol{y} + \lambda \mathrm{L}_\gamma \boldsymbol{y},\, \boldsymbol{x}_i \rangle .$$

where $\mathrm{H}_\gamma = 2\kappa\, \mathrm{I}_T + \lambda\,(\mathrm{L}_\gamma + \mathrm{U}_\gamma)$ is the Hessian matrix. Hence, for every direction $\boldsymbol{v}$,

$$\boldsymbol{v}^\top \mathrm{H}_\gamma \boldsymbol{v} = 2\kappa \|\boldsymbol{v}\|_2^2 + 2\lambda K_\gamma(\boldsymbol{v}) \ge 2\left(\kappa - \frac{\lambda}{1+\gamma}\right)\|\boldsymbol{v}\|_2^2 = 2\mu_\gamma \|\boldsymbol{v}\|_2^2 .$$

Thus every individual cost is $2\mu_\gamma$-strongly convex in the agent's own schedule, which motivates the use of Assumption A throughout this work.

C.2 Strong Monotonicity and the Unique Equilibrium

Proof of Theorem 2.1. Take two profiles $\mathrm{X}, \mathrm{Y} \in \mathcal{X}$ and write $\boldsymbol{d}_i = \boldsymbol{x}_i - \boldsymbol{y}_i$ and $\boldsymbol{d} = \sum_i \boldsymbol{d}_i = \boldsymbol{q}(\mathrm{X}) - \boldsymbol{q}(\mathrm{Y})$. From (13),

$$\boldsymbol{g}_i(\mathrm{X}) - \boldsymbol{g}_i(\mathrm{Y}) = \kappa(\boldsymbol{d} + \boldsymbol{d}_i) + \lambda \mathrm{L}_\gamma \boldsymbol{d} + \lambda \mathrm{U}_\gamma \boldsymbol{d}_i. \tag{18}$$

The entering state cancels because it does not depend on the schedules. We now take the inner product of (18) with $\boldsymbol{d}_i$ and sum over all agents. By the definition of the pseudo-gradient and the Frobenius inner product,

$$\begin{aligned}
\langle \mathrm{F}(\mathrm{X}) - \mathrm{F}(\mathrm{Y}), \mathrm{X} - \mathrm{Y} \rangle_{\mathrm{F}} &= \sum_{i=1}^{N} \langle \boldsymbol{g}_i(\mathrm{X}) - \boldsymbol{g}_i(\mathrm{Y}), \boldsymbol{d}_i \rangle \\
&= \kappa \sum_{i=1}^{N} \langle \boldsymbol{d} + \boldsymbol{d}_i, \boldsymbol{d}_i \rangle + \lambda \sum_{i=1}^{N} \langle \mathrm{L}_\gamma \boldsymbol{d}, \boldsymbol{d}_i \rangle + \lambda \sum_{i=1}^{N} \langle \mathrm{U}_\gamma \boldsymbol{d}_i, \boldsymbol{d}_i \rangle \\
&= \kappa \left(\|\boldsymbol{d}\|_2^2 + \sum_i \|\boldsymbol{d}_i\|_2^2 \right) + \lambda \left(K_\gamma(\boldsymbol{d}) + \sum_i K_\gamma(\boldsymbol{d}_i) \right).
\end{aligned} \tag{19}$$

We can now apply Lemma C.1 separately to the aggregate difference $\boldsymbol{d}$ and to every individual difference $\boldsymbol{d}_i$. Namely,

$$K_\gamma(\boldsymbol{d}) \ge -\frac{1}{1+\gamma}\|\boldsymbol{d}\|_2^2 \qquad \text{and, for every } i,\ K_\gamma(\boldsymbol{d}_i) \ge -\frac{1}{1+\gamma}\|\boldsymbol{d}_i\|_2^2.$$

Substituting these bounds into (19) yields

$$\begin{aligned}
\langle \mathrm{F}(\mathrm{X}) - \mathrm{F}(\mathrm{Y}), \mathrm{X} - \mathrm{Y} \rangle_{\mathrm{F}} &\ge \left(\kappa - \frac{\lambda}{1+\gamma}\right) \left(\|\boldsymbol{d}\|_2^2 + \sum_{i=1}^{N} \|\boldsymbol{d}_i\|_2^2 \right) \\
&= \mu_\gamma \left(\|\boldsymbol{q}(\mathrm{X}) - \boldsymbol{q}(\mathrm{Y})\|_2^2 + \|\mathrm{X} - \mathrm{Y}\|_{\mathrm{F}}^2 \right),
\end{aligned}$$

which is exactly (4).

Then, equilibrium existence is standard. Each quota simplex is nonempty, compact and convex, and every cost is continuous in the profile and convex in the agent's own schedule. Therefore, Rosen's theorem directly implies existence of a pure Nash equilibrium (Rosen, 1965). Equivalently, an equilibrium corresponds to the solution of the following variational inequality:

$$\langle \mathrm{F}(\mathrm{X}^\star), \mathrm{X} - \mathrm{X}^\star \rangle_{\mathrm{F}} \ge 0, \qquad \mathrm{X} \in \mathcal{X}.$$

If $\mathrm{X}^\star$ and $\mathrm{Y}^\star$ were two different equilibrium solutions, testing each variational inequality with the other solution and combining them would give $\langle \mathrm{F}(\mathrm{X}^\star) - \mathrm{F}(\mathrm{Y}^\star), \mathrm{X}^\star - \mathrm{Y}^\star \rangle_{\mathrm{F}} \le 0$. Then, strong monotonicity forces $\mathrm{X}^\star = \mathrm{Y}^\star$. □

C.3 Symmetric Equilibrium under Transient Impact

We next characterize the symmetric equilibrium structure of the acquisition game. We also analyze, in Appendix D, a useful benchmark that is completely oblivious to the structure of the game: the *uniform policy* $\boldsymbol{u}_i = (B_i/T, \dots, B_i/T)$, which spreads an agent's quota equally across the horizon (and coincides with the social optimum for $\gamma = 1$).

In the symmetric case, agent quotas are $B_i = B$ for all i and $V = NB$. Uniqueness and symmetry imply that the equilibrium $\mathrm{X}^\star = (\boldsymbol{x}_1^\star, \dots, \boldsymbol{x}_N^\star)$ is characterized by its aggregate schedule $\boldsymbol{q}^*$ and every agent i has $\boldsymbol{x}_i^\star = \boldsymbol{q}^\star/N$.

Theorem C.2 (Distance of Equilibrium to Uniform)**.** *Let $\boldsymbol{x}_i^\star = \boldsymbol{q}^\star/N$ be agent i's schedule in the (unique symmetric) Nash equilibrium for the symmetric case, where $B_i = B$ for every agent $i \in [N]$, and let $\boldsymbol{u}_i = (B/T)\mathbf{1}$ be i's uniform policy. Then, for every fixed $\gamma < 1$,*

$$\left\|\boldsymbol{x}_i^\star - \boldsymbol{u}_i\right\|_2 \le \frac{z}{(N+1)\mu_\gamma\sqrt{(1-\gamma^2)}} + \frac{\lambda B}{\mu_\gamma T(1-\gamma)\sqrt{(1-\gamma^2)}} \tag{20}$$

Namely, with initial state $z = 0$, $\boldsymbol{u}_i$ is at most $O(\frac{B}{T(1-\gamma)^{3/2}})$ in Euclidean distance away from the equilibrium (and as we show in Appendix C.4, also from the social optimum). For permanent impact, where $\gamma = 1$, we prove that $\|\boldsymbol{x}_i^* - \boldsymbol{u}_i\|_2 = O(B)$ instead.

Interestingly, transient impact and permanent impact behave differently, even when γ is relatively close to 1. For a fixed $\gamma < 1$, the length of the game's memory (accumulated in its state I_t) is of order $1/(1-\gamma)$. When $T(1-\gamma) \gg 1$, the acquisition horizon is long relative to the game's memory length; hence the non-uniformity of the symmetric equilibrium is primarily confined to two boundary layers, one near the beginning and another near the end of the acquisition horizon, of length $O(1/(1-\gamma))$. The remaining $T - O(1/(1-\gamma))$ coordinates are almost flat, which implies (20).

C.3.1 Complementarity Characterization

We first present some basic facts about the structure of the equilibrium. To this end, we first apply to the gradient identity (13) the fact that in the symmetric case, the equilibrium schedule is symmetric and equal to $\boldsymbol{x}_i^\star = \boldsymbol{q}^\star/N$ for all i. Then,

$$\boldsymbol{g}_i = z\boldsymbol{d}_\gamma + \mathrm{M}_N\boldsymbol{q}, \qquad \text{where}$$
$$\mathrm{M}_N = \alpha_N \mathrm{I}_T + \lambda \mathrm{L}_\gamma + \frac{\lambda}{N}\mathrm{U}_\gamma \qquad \text{and} \qquad \alpha_N = \frac{N+1}{N}\kappa.$$

For $t \in [T]$, we define

$$\Phi_{N,t}(\boldsymbol{q}; z) := z\gamma^{t-1} + \alpha_N q_t + \lambda(\mathrm{L}_\gamma \boldsymbol{q})_t + \frac{\lambda}{N}(\mathrm{U}_\gamma \boldsymbol{q})_t.$$

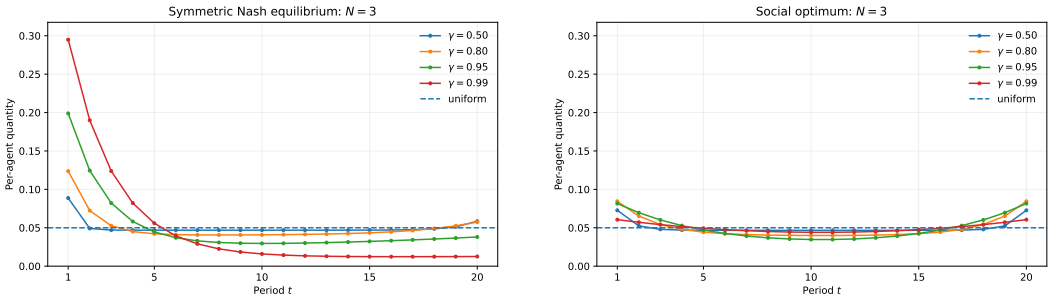
Intuitively, $\Phi_{N,t}(\boldsymbol{q}; z)$ is the common marginal cost at time t under a symmetric profile, expressed directly in terms of the aggregate schedule $\boldsymbol{q}$. It collects the entering-state effect, the impact of the aggregate demand on the current price, the effect of past aggregate demand through L_γ, and the agent's own future-impact correction, which under symmetry contributes the factor $1/N$. This notation reduces the symmetric equilibrium conditions to a T-dimensional problem linear complementarity system with the quota constraint in the aggregate schedule.

More specifically, the aggregate equilibrium is the unique aggregate demand vector of total volume V for which there is a scalar θ satisfying

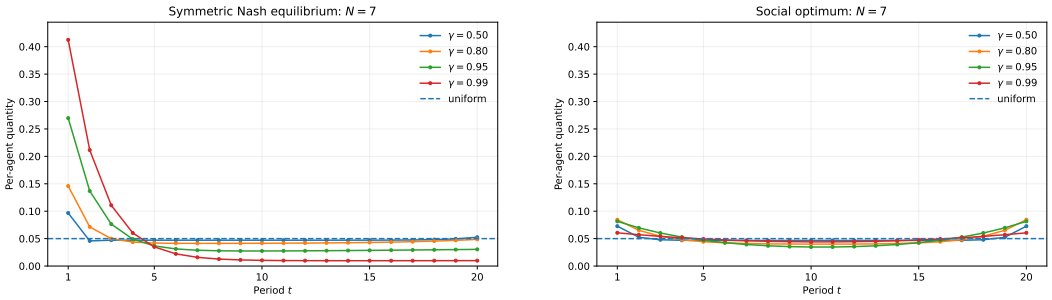
$$q_t^\star \ge 0, \qquad \Phi_{N,t}(\boldsymbol{q}^\star; z) \ge \theta,$$
$$q_t^\star\big(\Phi_{N,t}(\boldsymbol{q}^\star; z) - \theta\big) = 0, \qquad \mathbf{1}^\top \boldsymbol{q}^\star = V.$$

These conditions are precisely the KKT conditions for an agent's quota simplex. They have the usual marginal-cost interpretation: all periods used with positive quantity have the same marginal cost, while any unused period has marginal cost at least as large. This characterization continues to hold when some nonnegativity constraints are binding.

The complementarity system above is the only equilibrium characterization needed for the proof of Theorem C.2. Getting a more detailed characterization of the interior equilibrium is possible, but the derivation is tedious and beyond the scope of this paper.



(a) Nash equilibrium, $N = 3$ (b) Social optimum



(c) Nash equilibrium, $N = 7$ (d) Social optimum

Figure 2: Symmetric Nash equilibrium and social optimum under transient impact. Parameters are $T = 20$, $B = 1$, $\kappa = 1$, $\lambda = 0.6$, $z = 0$, and $\gamma \in \{0.50, 0.80, 0.95, 0.99\}$. The dashed line denotes the uniform schedule B/T. Each row uses the same vertical scale. The corresponding prices of anarchy are $(1.004, 1.018, 1.095, 1.246)$ for $N = 3$ and $(1.009, 1.039, 1.189, 1.430)$ for $N = 7$, in increasing order of γ. Strategic interaction increasingly front-loads execution relative to the social optimum as impact becomes more persistent and as the number of agents increases.

C.3.2 Numerical Illustration

The KKT system is straightforward to solve numerically. Figure 2 illustrates the per-agent symmetric equilibrium and the social optimum for $T = 20$, $B = 1$, $\kappa = 1$, $\lambda = 0.6$, $z = 0$, and $\gamma \in \{0.50, 0.80, 0.95, 0.99\}$. The dashed line is the uniform policy B/T. The plots illustrate the qualitative behavior behind the bounds proved below. When the horizon is long relative to the effective memory of the state, the equilibrium is close to the uniform schedule over most of the horizon and the price of anarchy decreases towards 1. For a fixed horizon, the distortion becomes much more pronounced as γ approaches one.

C.3.3 Proof of Theorem C.2

Distance to Uniform Policy. To establish Theorem C.2, we first bound the distance between the symmetric equilibrium and the uniform schedule from the variational inequality. To this end, we let $\boldsymbol{u} := (V/T)\mathbf{1}$ denote the uniform schedule.

We also write $\boldsymbol{\Phi}_N(\boldsymbol{q}; z) := \big(\Phi_{N,1}(\boldsymbol{q}; z), \dots, \Phi_{N,T}(\boldsymbol{q}; z)\big)^\top$. Then the symmetric equilibrium $\boldsymbol{q}^\star(z)$ is characterized by the variational inequality

$$\langle \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(z); z), \boldsymbol{q} - \boldsymbol{q}^\star(z) \rangle \geq 0 \qquad \text{for every } \boldsymbol{q} \geq \mathbf{0}, \quad \mathbf{1}^\top \boldsymbol{q} = V. \tag{21}$$

The key observation is that the map $\boldsymbol{\Phi}_N$ inherits a strong-monotonicity property very similar to that of the original game. Indeed, for $\boldsymbol{v} = \boldsymbol{q} - \boldsymbol{q}'$,

$$\langle \boldsymbol{\Phi}_N(\boldsymbol{q}; z) - \boldsymbol{\Phi}_N(\boldsymbol{q}'; z), \boldsymbol{q} - \boldsymbol{q}' \rangle = \alpha_N \|\boldsymbol{v}\|_2^2 + \lambda \boldsymbol{v}^\top \mathrm{L}_\gamma \boldsymbol{v} + \frac{\lambda}{N} \boldsymbol{v}^\top \mathrm{U}_\gamma \boldsymbol{v}.$$

Using that $\boldsymbol{v}^\top \mathrm{U}_\gamma \boldsymbol{v} = \boldsymbol{v}^\top \mathrm{L}_\gamma \boldsymbol{v} = K_\gamma(\boldsymbol{v})$ and that $\alpha_N = (N+1)\kappa/N$, we obtain

$$\begin{aligned}\langle \boldsymbol{\Phi}_N(\boldsymbol{q}; z) - \boldsymbol{\Phi}_N(\boldsymbol{q}'; z), \boldsymbol{q} - \boldsymbol{q}' \rangle &= \frac{N+1}{N} \left[\kappa \|\boldsymbol{v}\|_2^2 + \lambda K_\gamma(\boldsymbol{v})\right] \\ &\geq \frac{N+1}{N} \mu_\gamma \|\boldsymbol{v}\|_2^2.\end{aligned}$$

We first consider a clean entering state, $z = 0$. Let $\boldsymbol{e} := \boldsymbol{q}^\star(0) - \boldsymbol{u}$. Both schedules have total volume V, so $\mathbf{1}^\top \boldsymbol{e} = 0$. Strong monotonicity gives

$$\frac{N+1}{N} \mu_\gamma \|\boldsymbol{e}\|_2^2 \leq \langle \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(0); 0) - \boldsymbol{\Phi}_N(\boldsymbol{u}; 0), \boldsymbol{e} \rangle .$$

On the other hand, using $\boldsymbol{u}$ as a feasible comparison point in (21), $\langle \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(0); 0), \boldsymbol{u} - \boldsymbol{q}^\star(0) \rangle \geq 0$, and hence $\langle \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(0); 0), \boldsymbol{e} \rangle \leq 0$. Consequently,

$$\frac{N+1}{N} \mu_\gamma \|\boldsymbol{e}\|_2^2 \leq - \langle \boldsymbol{\Phi}_N(\boldsymbol{u}; 0), \boldsymbol{e} \rangle . \tag{22}$$

To evaluate how far the uniform schedule is from satisfying the equilibrium conditions, we compute its marginal-cost vector explicitly. Recall that $\boldsymbol{u} = (V/T)\mathbf{1}$ and $\boldsymbol{\Phi}_N(\boldsymbol{q}; 0) = \alpha_N \boldsymbol{q} + \lambda \mathrm{L}_\gamma \boldsymbol{q} + \frac{\lambda}{N} \mathrm{U}_\gamma \boldsymbol{q}$. Thus we only need to understand how L_γ and U_γ act on the constant vector $\mathbf{1}$.

For the past-memory term, the t-th coordinate is

$$(\mathrm{L}_\gamma \mathbf{1})_t = \sum_{s=1}^{t-1} \gamma^{t-s-1} = 1 + \gamma + \cdots + \gamma^{t-2} = \frac{1 - \gamma^{t-1}}{1 - \gamma}.$$

Therefore, with $\boldsymbol{d}_\gamma := (1, \gamma, \ldots, \gamma^{T-1})^\top$, we have

$$\mathrm{L}_\gamma \mathbf{1} = \frac{\mathbf{1} - \boldsymbol{d}_\gamma}{1 - \gamma}.$$

Similarly, the t-th coordinate of the future-memory term is

$$(\mathrm{U}_\gamma \mathbf{1})_t = \sum_{s=t+1}^{T} \gamma^{s-t-1} = 1 + \gamma + \cdots + \gamma^{T-t-1} = \frac{1 - \gamma^{T-t}}{1 - \gamma}.$$

Hence

$$\mathrm{U}_\gamma \mathbf{1} = \frac{\mathbf{1} - (\gamma^{T-1}, \ldots, \gamma, 1)^\top}{1 - \gamma}.$$

Substituting $\boldsymbol{u} = (V/T)\mathbf{1}$ into the marginal-cost map now gives

$$\begin{aligned}\boldsymbol{\Phi}_N(\boldsymbol{u}; 0) &= \frac{V}{T} \left[\alpha_N \mathbf{1} + \frac{\lambda}{1-\gamma}(\mathbf{1} - \boldsymbol{d}_\gamma) + \frac{\lambda}{N(1-\gamma)} \left(\mathbf{1} - (\gamma^{T-1}, \ldots, \gamma, 1)^\top\right)\right] \\ &= c_{\mathrm{NE}}\, \mathbf{1} - \frac{\lambda V}{T(1-\gamma)} \left[\boldsymbol{d}_\gamma + \frac{1}{N}(\gamma^{T-1}, \ldots, \gamma, 1)^\top\right],\end{aligned} \tag{23}$$

where

$$c_{\mathrm{NE}} := \frac{V}{T} \left[\alpha_N + \frac{\lambda}{1-\gamma} \left(1 + \frac{1}{N}\right)\right].$$

This decomposition is useful because the first term is constant across periods. We recall that $\mathbf{1}^\top \boldsymbol{e} = 0$. Hence any constant marginal-cost component is orthogonal to $\boldsymbol{e}$, i.e., $\langle c_{\mathrm{NE}} \mathbf{1}, \boldsymbol{e} \rangle = c_{\mathrm{NE}}\, \mathbf{1}^\top \boldsymbol{e} = 0$.

Thus, for comparing the equilibrium with the uniform schedule, only the variation of marginal costs across time matters. To make this temporal variation explicit, we let

$$\bar{d}_\gamma := \frac{1}{T} \mathbf{1}^\top \boldsymbol{d}_\gamma \qquad \text{and} \qquad \delta_T(\gamma) := \|\boldsymbol{d}_\gamma - \bar{d}_\gamma \mathbf{1}\|_2.$$

Since $\mathbf{1}^\top \boldsymbol{e} = 0$, $\langle \boldsymbol{d}_\gamma, \boldsymbol{e}\rangle = \langle \boldsymbol{d}_\gamma - \bar{d}_\gamma \mathbf{1}, \boldsymbol{e}\rangle$. The reversed vector $(\gamma^{T-1}, \dots, \gamma, 1)^\top$ has the same average and the same Euclidean distance from a constant vector. Therefore

$$\left|\left\langle (\gamma^{T-1}, \dots, \gamma, 1)^\top, \boldsymbol{e}\right\rangle\right| \le \delta_T(\gamma)\|\boldsymbol{e}\|_2,$$

and, by Cauchy–Schwarz,

$$|\langle \boldsymbol{d}_\gamma, \boldsymbol{e}\rangle| \le \delta_T(\gamma)\|\boldsymbol{e}\|_2.$$

Applying these two bounds to (23) gives

$$|\langle \boldsymbol{\Phi}_N(\boldsymbol{u}; 0), \boldsymbol{e}\rangle| \le \frac{\lambda V}{T(1-\gamma)}\left(1 + \frac{1}{N}\right)\delta_T(\gamma)\|\boldsymbol{e}\|_2. \tag{24}$$

Intuitively, under a uniform schedule, the current-impact term is the same in every period and therefore does not create an incentive to move volume from one period to another. The only source of nonuniform marginal costs is the finite-horizon memory: early and late periods have different amounts of past and future impact. The quantity $\delta_T(\gamma)$ measures precisely the size of this temporal variation.

Combining (22) and (24), cancelling the common factor $(N+1)/N$, and focusing on the interesting case where $\boldsymbol{e} \neq \mathbf{0}$, which allows division by $\|\boldsymbol{e}\|_2$, yields

$$\|\boldsymbol{q}^\star(0) - \boldsymbol{u}\|_2 \le \frac{\lambda V\, \delta_T(\gamma)}{\mu_\gamma T(1-\gamma)}.$$

Focusing on the equilibrium schedule $\boldsymbol{x}_i^\star$ of a single agent i and using that for any fixed $\gamma < 1$,

$$\delta_T(\gamma) \le \|\boldsymbol{d}_\gamma\|_2 = \left(\sum_{r=0}^{T-1} \gamma^{2r}\right)^{1/2} \le \frac{1}{\sqrt{1-\gamma^2}},$$

we obtain the second term in the r.h.s. of (20).

Thus, under transient impact, the aggregate equilibrium approaches the uniform schedule as the execution horizon grows. Interestingly, this conclusion follows directly from the KKT conditions and strong monotonicity.

The same argument also applies to an inherited state $z > 0$. Let $\boldsymbol{q}^\star(z)$ and $\boldsymbol{q}^\star(0)$ denote the equilibria with and without the entering state and let $\Delta\boldsymbol{q} := \boldsymbol{q}^\star(z) - \boldsymbol{q}^\star(0)$. Applying their two variational inequalities (21) against one another gives

$$\begin{aligned} &\langle \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(z); z), \boldsymbol{q}^\star(0) - \boldsymbol{q}^\star(z)\rangle \ge 0, \qquad \text{and} \\ &\langle \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(0); 0), \boldsymbol{q}^\star(z) - \boldsymbol{q}^\star(0)\rangle \ge 0. \end{aligned}$$

From the two variational inequalities derived above,

$$\langle \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(z); z) - \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(0); z), \Delta\boldsymbol{q}\rangle \le -z\,\langle \boldsymbol{d}_\gamma, \Delta\boldsymbol{q}\rangle\,.$$

On the other hand, strong monotonicity of $\boldsymbol{\Phi}_N(\cdot; z)$ gives

$$\langle \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(z); z) - \boldsymbol{\Phi}_N(\boldsymbol{q}^\star(0); z), \Delta\boldsymbol{q}\rangle \ge \frac{N+1}{N}\mu_\gamma\|\Delta\boldsymbol{q}\|_2^2.$$

Combining the two inequalities, and using $z \ge 0$, yields

$$\frac{N+1}{N}\mu_\gamma\|\boldsymbol{q}^\star(z) - \boldsymbol{q}^\star(0)\|_2^2 \le z\,|\langle \boldsymbol{d}_\gamma, \boldsymbol{q}^\star(z) - \boldsymbol{q}^\star(0)\rangle|\,.$$

The difference of the two equilibrium schedules has zero coordinate sum, so only the non-constant part of $\boldsymbol{d}_\gamma$ matters. Hence

$$\|\boldsymbol{q}^\star(z) - \boldsymbol{q}^\star(0)\|_2 \le \frac{Nz\,\delta_T(\gamma)}{(N+1)\mu_\gamma}.$$

Finally, using $\boldsymbol{u}_i = (B/T)\mathbf{1}$ and that $\boldsymbol{x}_i^\star = \boldsymbol{q}^\star/N$, the triangle inequality yields

$$\|\boldsymbol{x}_i^\star - \boldsymbol{u}_i\|_2 \le \frac{z\,\delta_T(\gamma)}{(N+1)\mu_\gamma} + \frac{\lambda B\,\delta_T(\gamma)}{\mu_\gamma T(1-\gamma)}.$$

Using $\delta_T(\gamma) \le (1-\gamma^2)^{-1/2}$ gives the simpler bound, which concludes the proof of Theorem C.2:

$$\|\boldsymbol{x}_i^\star - \boldsymbol{u}_i\|_2 \le \frac{z}{(N+1)\mu_\gamma\sqrt{1-\gamma^2}} + \frac{\lambda B}{\mu_\gamma T(1-\gamma)\sqrt{1-\gamma^2}}.$$

Hence for $z = 0$ and any fixed $\gamma < 1$, $\|\boldsymbol{x}_i^\star - \boldsymbol{u}_i\|_2 = O(B/T)$ and $O(BT^{-1}(1-\gamma)^{-3/2})$, if dependence on γ is to be taken into account. □

Remark on the permanent endpoint. The estimates above do not consider the case of permanent impact, where $\gamma = 1$. For $\gamma = 1$, quota feasibility implies the following coarse upper bound:

$$\|\boldsymbol{x}_i^\star - \boldsymbol{u}_i\|_2 \le \|\boldsymbol{x}_i^\star\|_2 + \frac{B}{\sqrt{T}} \le B + \frac{B}{\sqrt{T}} = O(B).$$

Thus the $O(B/T)$ transient bound fails when $\gamma = 1$ (or the value of γ is very close to 1).

C.4 Social Optimum and Transient Price of Anarchy

We now use the same idea to study social efficiency of equilibrium, quantified by the price of anarchy. Intuitively, for transient impact, both the Nash equilibrium and the socially optimal aggregate schedule are close to the same uniform schedule. This is enough to show that their costs are also close.

Throughout this section, we consider the symmetric case, where agent quotas are $B_i = B$ for all i, and assume that the initial state is $z = 0$, aiming to isolate the inefficiency generated by the current execution. The total demand is $V = NB$. We let $\boldsymbol{q}^{\mathrm{NE}} := \boldsymbol{q}^\star(0)$ denote the aggregate demand schedule of the Nash equilibrium. We compare social cost at the Nash equilibrium against the social optimum. With $z = 0$, the social cost depends only on the aggregate schedule $\boldsymbol{q}$. Indeed,

$$J(\boldsymbol{q}) := \sum_{i=1}^{N} C_i(\mathrm{X}) = \kappa\|\boldsymbol{q}\|_2^2 + \lambda K_\gamma(\boldsymbol{q}). \tag{25}$$

Thus the social optimum is $\boldsymbol{q}^{\mathrm{opt}} \in \arg\min_{\boldsymbol{q}\ge\boldsymbol{0},\ \mathbf{1}^\top\boldsymbol{q}=V} J(\boldsymbol{q})$, and the price of anarchy is

$$\mathrm{PoA} := \frac{J(\boldsymbol{q}^{\mathrm{NE}})}{J(\boldsymbol{q}^{\mathrm{opt}})}.$$

In this section, we establish the following:

Theorem C.3 (Price of Anarchy)**.** *In the symmetric case, where $B_i = B$ for every agent i, and starting with initial state $z = 0$, for every fixed $\gamma < 1$,*

$$\mathrm{PoA} \le 1 + \frac{\lambda^2}{\mu_\gamma^2 T(1-\gamma)^3}\left(4 + \frac{2\lambda}{\mu_\gamma(1-\gamma)}\right)$$

Hence, in the transient case, $\mathrm{PoA} = 1 + O(1/(T(1-\gamma)^4))$ and tends to 1 as T becomes sufficiently large relative to the memory length. On the other hand, if we have permanent impact (i.e., if $\gamma = 1$), in the standard interior regime where $\lambda < (N+1)\kappa/N$, the equilibrium schedule exhibits a geometric decay and $\mathrm{PoA} \le 2N/(N+1)$ (and tends to the right-hand side as T increases) (Chriss, 2024c; Kearns & Shi, 2025; Hossain et al., 2026).

C.4.1 The Proof of Theorem C.3

Social Cost. We observe that the gradient of the social cost is $\nabla_{\boldsymbol{q}} J(\boldsymbol{q}) = 2\kappa\,\boldsymbol{q} + \lambda(\mathrm{L}_\gamma + \mathrm{U}_\gamma)\,\boldsymbol{q}$. For two aggregate schedules $\boldsymbol{q}, \boldsymbol{q}'$ and $\boldsymbol{v} := \boldsymbol{q} - \boldsymbol{q}'$, we therefore have

$$\langle \nabla J(\boldsymbol{q}) - \nabla J(\boldsymbol{q}'), \boldsymbol{q} - \boldsymbol{q}'\rangle = 2\kappa\|\boldsymbol{v}\|_2^2 + 2\lambda K_\gamma(\boldsymbol{v}) \ge 2\mu_\gamma\|\boldsymbol{v}\|_2^2,$$

where the last inequality follows from Lemma C.1. Hence the social cost $J(\boldsymbol{q})$ is $2\mu_\gamma$-strongly convex and the social optimum is unique.

Applying the memory bound of Lemma C.1 to (25), we obtain that: $J(\boldsymbol{q}) \ge \mu_\gamma\|\boldsymbol{q}\|_2^2$. Since every feasible aggregate schedule satisfies $\mathbf{1}^\top\boldsymbol{q} = V$, Cauchy–Schwarz gives $\|\boldsymbol{q}\|_2^2 \ge V^2/T$. In particular,

$$J(\boldsymbol{q}^{\mathrm{opt}}) \ge \frac{\mu_\gamma V^2}{T}.$$

This lower bound will be the denominator of our price-of-anarchy estimate.

The Social Optimum is also Close to Uniform. We let $\boldsymbol{u} := (V/T)\mathbf{1}$ be the uniform aggregate schedule. We first quantify how far $\boldsymbol{u}$ is from satisfying the planner's first-order conditions.

Using

$$\mathrm{L}_\gamma \mathbf{1} = \frac{\mathbf{1} - \boldsymbol{d}_\gamma}{1-\gamma} \qquad \text{and} \qquad \mathrm{U}_\gamma \mathbf{1} = \frac{\mathbf{1} - (\gamma^{T-1}, \dots, \gamma, 1)^\top}{1-\gamma},$$

the gradient at the uniform schedule $\boldsymbol{u}$ can be written as

$$\nabla J(\boldsymbol{u}) = c_{\mathrm{soc}}\, \mathbf{1} - \frac{\lambda V}{T(1-\gamma)} \left[\boldsymbol{d}_\gamma + (\gamma^{T-1}, \dots, \gamma, 1)^\top\right], \qquad \text{for some scalar } c_{\mathrm{soc}}.$$

Now consider any vector $\boldsymbol{v}$ satisfying $\mathbf{1}^\top \boldsymbol{v} = 0$. The constant term $c_{\mathrm{soc}}\mathbf{1}$ is orthogonal to $\boldsymbol{v}$, and therefore

$$|\langle \nabla J(\boldsymbol{u}), \boldsymbol{v}\rangle| \le \frac{2\lambda V\, \delta_T(\gamma)}{T(1-\gamma)} \|\boldsymbol{v}\|_2. \tag{26}$$

Indeed, both $\boldsymbol{d}_\gamma$ and its reversed version have the same average and the same Euclidean distance $\delta_T(\gamma)$ from a constant vector.

This estimate has the same interpretation as in the equilibrium argument. Because the total demand V is fixed, a component of marginal cost that is constant across time does not affect the allocation of the demand over time. Only the variation of marginal social cost across periods can make the social optimum nonuniform over time.

We can now compare the social optimum with the uniform schedule $\boldsymbol{u} = (V/T)\mathbf{1}$. The argument is the same as the one used in the proof of Theorem C.2 to bound the distance of the symmetric Nash equilibrium to the uniform policy $\boldsymbol{u}$. So we only explain the changes needed for the social optimum.

To compare the social optimum with the uniform schedule $\boldsymbol{u}$, we let $\boldsymbol{e} := \boldsymbol{q}^{\mathrm{opt}} - \boldsymbol{u}$. Since both schedules have total volume V, $\mathbf{1}^\top \boldsymbol{e} = 0$. Moreover, the social-cost gradient satisfies

$$\langle \nabla J(\boldsymbol{q}) - \nabla J(\boldsymbol{q}'), \boldsymbol{q} - \boldsymbol{q}'\rangle \ge 2\mu_\gamma \|\boldsymbol{q} - \boldsymbol{q}'\|_2^2 .$$

So the social cost gradient $\nabla J(\boldsymbol{q})$ plays here the same role as the strongly monotone marginal-cost map $\boldsymbol{\Phi}_N$ did in Appendix C.3.

By the first-order optimality condition for $\boldsymbol{q}^{\mathrm{opt}}$, tested with the feasible schedule $\boldsymbol{u}$, we obtain that $\langle \nabla J(\boldsymbol{q}^{\mathrm{opt}}), \boldsymbol{e}\rangle \le 0$. Combining these two facts yields $2\mu_\gamma \|\boldsymbol{e}\|_2^2 \le -\langle \nabla J(\boldsymbol{u}), \boldsymbol{e}\rangle$.

As in the equilibrium calculation, only the non-constant part of the marginal-cost vector matters because $\mathbf{1}^\top \boldsymbol{e} = 0$. Thus, from (26),

$$|\langle \nabla J(\boldsymbol{u}), \boldsymbol{e}\rangle| \le \frac{2\lambda V\, \delta_T(\gamma)}{T(1-\gamma)} \|\boldsymbol{e}\|_2.$$

Therefore,

$$2\mu_\gamma \|\boldsymbol{e}\|_2^2 \le \frac{2\lambda V\, \delta_T(\gamma)}{T(1-\gamma)} \|\boldsymbol{e}\|_2.$$

Canceling the common factor 2, focusing on the non-trivial case where $\boldsymbol{e} \neq \mathbf{0}$ and dividing by $\mu_\gamma \|\boldsymbol{e}\|_2$, we obtain that:

$$\left\|\boldsymbol{q}^{\mathrm{opt}} - \boldsymbol{u}\right\|_2 \le \frac{\lambda V\, \delta_T(\gamma)}{\mu_\gamma T(1-\gamma)}.$$

From the previous subsection, the Nash equilibrium satisfies exactly the same bound:

$$\left\|\boldsymbol{q}^{\mathrm{NE}} - \boldsymbol{u}\right\|_2 \le \frac{\lambda V\, \delta_T(\gamma)}{\mu_\gamma T(1-\gamma)}.$$

Thus both the social optimum and the strategic agents at equilibrium choose aggregate schedules within the same $O(T^{-1})$ neighborhood of the uniform schedule when $\gamma < 1$ is fixed.

From Distance to Social Cost. We have shown that both the Nash equilibrium and the social optimum are close to the same uniform aggregate schedule. For convenience, we denote this common distance bound by

$$\varepsilon_T := \frac{\lambda V\, \delta_T(\gamma)}{\mu_\gamma T(1-\gamma)}. \tag{27}$$

In this notation, $\|\boldsymbol{q}^{\mathrm{NE}} - \boldsymbol{u}\|_2 \le \varepsilon_T$ and $\|\boldsymbol{q}^{\mathrm{opt}} - \boldsymbol{u}\|_2 \le \varepsilon_T$. We now translate these distance bounds into costs. We use two simple estimates. First, from (26), for every $\boldsymbol{v}$ with $\mathbf{1}^\top \boldsymbol{v} = 0$,

$$|\langle \nabla J(\boldsymbol{u}), \boldsymbol{v}\rangle| \le 2\mu_\gamma \varepsilon_T \|\boldsymbol{v}\|_2. \tag{28}$$

Second, by (25) and the upper bound of Lemma C.1,

$$J(\boldsymbol{v}) \le \left(\kappa + \frac{\lambda}{1-\gamma}\right) \|\boldsymbol{v}\|_2^2. \tag{29}$$

Consider first the Nash equilibrium and write $\boldsymbol{a} := \boldsymbol{q}^{\mathrm{NE}} - \boldsymbol{u}$. Since $J(\boldsymbol{q})$ is a homogeneous quadratic function, its exact quadratic expansion around $\boldsymbol{u}$ gives

$$J(\boldsymbol{u} + \boldsymbol{a}) - J(\boldsymbol{u}) = \langle \nabla J(\boldsymbol{u}), \boldsymbol{a}\rangle + J(\boldsymbol{a}).$$

Moreover, $\mathbf{1}^\top \boldsymbol{a} = 0$. Using (28), (29), and $\|\boldsymbol{a}\|_2 \le \varepsilon_T$, we obtain that

$$J(\boldsymbol{q}^{\mathrm{NE}}) - J(\boldsymbol{u}) \le \left[2\mu_\gamma + \kappa + \frac{\lambda}{1-\gamma}\right] \varepsilon_T^2. \tag{30}$$

We also need to control how much the social optimum improves on the uniform schedule. We let $\boldsymbol{b} := \boldsymbol{u} - \boldsymbol{q}^{\mathrm{opt}}$. As before, since $J(\boldsymbol{q})$ is a homogeneous quadratic function, its exact quadratic expansion around $\boldsymbol{u}$ gives

$$J(\boldsymbol{u}) - J(\boldsymbol{q}^{\mathrm{opt}}) = \langle \nabla J(\boldsymbol{u}), \boldsymbol{b}\rangle - J(\boldsymbol{b}).$$

By (28) and $J(\boldsymbol{b}) \ge \mu_\gamma \|\boldsymbol{b}\|_2^2$, we obtain that

$$J(\boldsymbol{u}) - J(\boldsymbol{q}^{\mathrm{opt}}) \le 2\mu_\gamma \varepsilon_T \|\boldsymbol{b}\|_2 - \mu_\gamma \|\boldsymbol{b}\|_2^2.$$

The right-hand side is maximized at $\|\boldsymbol{b}\|_2 = \varepsilon_T$, and therefore

$$J(\boldsymbol{u}) - J(\boldsymbol{q}^{\mathrm{opt}}) \le \mu_\gamma \varepsilon_T^2. \tag{31}$$

Adding (30) and (31), we get

$$J(\boldsymbol{q}^{\mathrm{NE}}) - J(\boldsymbol{q}^{\mathrm{opt}}) \le \left[3\mu_\gamma + \kappa + \frac{\lambda}{1-\gamma}\right] \varepsilon_T^2. \tag{32}$$

Using $\kappa + \frac{\lambda}{1-\gamma} = \mu_\gamma + \frac{2\lambda}{1-\gamma^2}$, the bracket becomes $4\mu_\gamma + \frac{2\lambda}{1-\gamma^2}$.

Finally, we recall that $J(\boldsymbol{q}^{\mathrm{opt}}) \ge \mu_\gamma V^2/T$. Dividing (32) by this lower bound on the optimum and substituting (27) gives

$$\mathrm{PoA} - 1 \le \frac{\lambda^2 \delta_T^2(\gamma)}{\mu_\gamma^2 T(1-\gamma)^2} \left(4 + \frac{2\lambda}{\mu_\gamma(1-\gamma^2)}\right).$$

Using $\delta_T^2(\gamma) \le \frac{1}{1-\gamma^2} \le \frac{1}{1-\gamma}$, we recover the simpler bound stated in Theorem C.3 and conclude the proof of the theorem. □

Hence, we observe that in the transient case, for any fixed $\gamma < 1$, $\mathrm{PoA} = 1 + O(T^{-1}(1-\gamma)^{-4})$. The reason is relatively simple. Both the equilibrium and the social optimum lie within $O(T^{-1})$ of the same uniform schedule. Since the social cost is quadratic, the resulting difference in their costs is $O(T^{-2})$, whereas the optimal social cost itself is $O(T^{-1})$. Their ratio therefore differs from one by only $O(T^{-1})$. Moreover, the assumption that $z = 0$ isolates the inefficiency generated by strategic interaction during the current episode. If $z > 0$, the social cost includes an additional timing term $z\langle \boldsymbol{d}_\gamma, \boldsymbol{q}\rangle$, so a price of anarchy bound would also reflect the incentive to delay execution while the inherited state decays.

D PROPERTIES OF THE UNIFORM POLICY

In this section, we analyze the properties of the uniform policy $\boldsymbol{u}_i = (B_i/T)\mathbf{1}$, which does not require any knowledge of N, competing quotas or the parameters of the game. Hence, it serves as a useful model-oblivious benchmark. In (20), we proved that $\boldsymbol{u}_i$ is close to the symmetric equilibrium and to the social optimum. Next, we consider its absolute cost, its sensitivity to the timing of competing demand, and its exploitability by a strategic agent.

Throughout this section, we let $V := \sum_{j=1}^{N} B_j$ be the aggregate quota and let $V_{-i} := V - B_i$ denote the aggregate competitive quota of an agent i's opponents. We also repeatedly use the sums $h_m(\gamma)$ and $\mathsf{A}_T(\gamma)$ introduced in Appendix A.

D.1 COST AND SENSITIVITY TO COMPETING DEMAND

Proposition D.1 (Uniform Cost and Sensitivity)**.** *Suppose $0 < \gamma < 1$. If every agent uses the uniform policy, the cost of an agent i is*

$$C_i^{\boldsymbol{u}} \le \frac{B_i\, z}{T(1-\gamma)} + \frac{\lambda B_i\, V}{T(1-\gamma)} + \frac{\kappa B_i\, V}{T}. \tag{33}$$

If agent i remains uniform while the other agents distribute their aggregate quota V_{-i} arbitrarily over time, then

$$\max C_i^{\boldsymbol{u}} - \min C_i^{\boldsymbol{u}} \le \frac{\lambda B_i V_{-i}}{T(1-\gamma)}. \tag{34}$$

The maximum is attained when the competing demand is placed as early as possible, and the minimum when it is placed as late as possible.

Before proving the result, it is useful to read the two bounds from a conceptual viewpoint. In (33), the three terms correspond to three distinct sources of expenditure. The first is exposure to the initial state z, the second is due to the aggregated state created during the episode, and the third is due to congestion in the present period. For fixed $\gamma < 1$, all three are $O(1/T)$. At $\gamma = 1$, the sum of the first two converges to $B_i z + \lambda B_i V/2$.

The timing range in (34) concerns a different kind of robustness. If a competitor buys early, the state created by that trade remains in the market for many of agent i's later purchases. If the same competing volume is moved toward the end, much less of agent i's schedule is exposed to it. The bound measures exactly how much this timing can matter.

Proof of Proposition D.1. If every agent is uniform, aggregate demand is $q_t = V/T$. Unrolling the state recursion gives $I_t = \gamma^{t-1} z + \frac{\lambda V}{T} h_{t-1}(\gamma)$, and

$$P_t = \gamma^{t-1} z + \frac{\lambda V}{T} h_{t-1}(\gamma) + \kappa \frac{V}{T}.$$

Since agent i buys B_i/T units in every period,

$$C_i^{\boldsymbol{u}} = \frac{B_i}{T} \sum_{t=1}^{T} P_t = \frac{B_i z}{T} h_T(\gamma) + \frac{\lambda B_i V}{T^2} \sum_{t=1}^{T} h_{t-1}(\gamma) + \frac{\kappa B_i V}{T}.$$

For $0 < \gamma < 1$,

$$h_T(\gamma) \le \frac{1}{1-\gamma}, \qquad \text{and} \qquad \sum_{t=1}^{T} h_{t-1}(\gamma) \le \frac{T}{1-\gamma},$$

which immediately gives (33).

At the permanent endpoint, the same analysis gives

$$C_i^{\boldsymbol{u}} = B_i z + \lambda B_i V \frac{T-1}{2T} + \frac{\kappa B_i V}{T}.$$

Thus, unlike the transient case, the uniform cost need not vanish as the horizon grows.

We next keep agent i uniform and allow the competitors to choose any aggregate schedule

$$\boldsymbol{y} := \sum_{j \neq i} \boldsymbol{x}_j, \qquad \boldsymbol{y} \geq \mathbf{0}, \qquad \mathbf{1}^\top \boldsymbol{y} = V_{-i}.$$

Their same-period contribution to agent i's cost is

$$\kappa \sum_{t=1}^{T} \frac{B_i}{T} y_t = \frac{\kappa B_i V_{-i}}{T},$$

which depends only on the total competing quota and not on its timing.

Timing matters through the state. One unit of competing demand placed in period s affects agent i's later purchases with total weight $1 + \gamma + \cdots + \gamma^{T-s-1} = h_{T-s}(\gamma)$. Therefore the timing-dependent part of the cost is

$$\frac{\lambda B_i}{T} \sum_{s=1}^{T} h_{T-s}(\gamma) y_s.$$

The weights decrease with s, from $h_{T-1}(\gamma)$ in the first period to zero in the last period. Because the above expression is linear in $\boldsymbol{y}$, its maximum and minimum over the competing-quota simplex are attained at these two extreme timings. Hence the exact range is

$$\max_{\boldsymbol{y}} C_i(\boldsymbol{u}_i; \boldsymbol{y}) - \min_{\boldsymbol{y}} C_i(\boldsymbol{u}_i; \boldsymbol{y}) = \frac{\lambda B_i V_{-i}}{T} h_{T-1}(\gamma).$$

Using $h_{T-1}(\gamma) \leq (1-\gamma)^{-1}$ proves (34). □

Proposition D.1 highlights the robustness of the uniform policy under transient impact. Then, a unit of competing demand has only a finite effective memory. A uniform agent spreads its own exposure over all T periods, so both its total cost and the effect of adversarial timing vanish at rate T^{-1}. At permanent impact, this protection disappears because early trades continue to affect the price throughout the horizon.

D.2 Best Response to Uniform Competitors

We now suppose that every other agent remains uniform and ask how agent i would optimally deviate. Let $\boldsymbol{u}_{-i} := (V_{-i}/T)\mathbf{1}$, and define agent i's individual cost against $\boldsymbol{u}_{-i}$ as $f_i(\boldsymbol{x}) := C_i(\boldsymbol{x}; \boldsymbol{u}_{-i})$, for $\boldsymbol{x} \in \mathcal{X}_i$. In the matrix form of Appendix A,

$$f_i(\boldsymbol{x}) = \kappa \|\boldsymbol{x}\|_2^2 + \lambda K_\gamma(\boldsymbol{x}) + \langle z \boldsymbol{d}_\gamma + \kappa \boldsymbol{u}_{-i} + \lambda \mathrm{L}_\gamma \boldsymbol{u}_{-i}, \boldsymbol{x} \rangle ,$$

where $\boldsymbol{d}_\gamma := (1, \gamma, \ldots, \gamma^{T-1})^\top$. By Lemma C.1, f_i is $2\mu_\gamma$-strongly convex. Therefore, it has a unique best response, denoted $\boldsymbol{x}_i^{\mathrm{br}}$.

We measure the deviation of i's best response strategy from uniform in two ways:

$$D_i^{\boldsymbol{u}} := \|\boldsymbol{u}_i - \boldsymbol{x}_i^{\mathrm{br}}\|_2 \qquad \text{and} \qquad E_i^{\boldsymbol{u}} := f_i(\boldsymbol{u}_i) - f_i(\boldsymbol{x}_i^{\mathrm{br}}).$$

The first quantity, $D_i^{\boldsymbol{u}}$, is the Euclidean distance of the uniform policy to i's best response when all other agents remain uniform and quantifies how much $\boldsymbol{x}_i^{\mathrm{br}}$ differs from $\boldsymbol{u}_i$. The second quantity, $E_i^{\boldsymbol{u}}$, a.k.a. the *exploitability of the uniform policy*, is the largest possible improvement on i's cost when all other agents remain uniform and the difference of costs between $\boldsymbol{x}_i^{\mathrm{br}}$ and $\boldsymbol{u}_i$. The remainder of this section is devoted to the proof of the following.

Theorem D.2. *Assume that the agent quotas $B_1, \ldots, B_N$ are fixed and independent of T. Let $V = B_1 + \cdots + B_N$ and $V_{-i} = V - B_i$. Then the following hold for the uniform policy's exploitability $E_i^{\boldsymbol{u}}$ and its distance $D_i^{\boldsymbol{u}}$ to agent i's best response.*

	$D_i^{\boldsymbol{u}}$	$E_i^{\boldsymbol{u}}$
$0 < \gamma < 1,\ z = 0$	$O\left(\frac{\lambda V}{\mu_\gamma (1-\gamma)^{3/2}} \frac{1}{T}\right)$	$O\left(\frac{\lambda^2 V^2}{\mu_\gamma (1-\gamma)^3} \frac{1}{T^2}\right)$
$0 < \gamma < 1,\ z > 0$	$O\left(B_i \frac{\sqrt{\log T}}{T}\right)$	$\frac{B_i z}{1-\gamma} \frac{1}{T} + o(T^{-1})$
$\gamma = 1, V_{-i} > 0$	$O\left(B_i^{3/4} \left(\frac{\lambda V_{-i}}{\mu_1 T}\right)^{1/4}\right)$	$\frac{\lambda B_i V_{-i}}{2} + O(T^{-1/2})$.

(35)

The three regimes in Theorem D.2 have different economic mechanisms. With transient impact and $z = 0$, the uniform policy is already nearly optimal: only finite-horizon boundary effects make marginal costs differ across time. With transient impact and $z > 0$, the entering state makes the first few periods more expensive, so the best response waits briefly before behaving almost uniformly. With permanent impact, the market never forgets competing flow. Hence, the best response is genuinely front-loaded and the value of deviating does not vanish.

D.2.1 A Technical Lemma about Quadratic Comparisons

We start with proving a simple technical lemma about strongly convex quadratic functions, which is used repeatedly in the proof of Theorem D.2.

For any vector $\boldsymbol{g} \in \mathbb{R}^T$, write $\bar{g} := \mathbf{1}^\top \boldsymbol{g}/T$ and $\widetilde{\boldsymbol{g}} := \boldsymbol{g} - \bar{g}\,\mathbf{1}$. Thus $\widetilde{\boldsymbol{g}}$ is simply the part of $\boldsymbol{g}$ that varies across time. This is the only part that matters on a quota simplex, because any two feasible schedules have the same coordinate sum.

Lemma D.3 (Quadratic Comparison)**.** *Let f be a quadratic function that is 2μ-strongly convex on a convex set $\mathcal{X}$ contained in a fixed-sum hyperplane $\{\boldsymbol{x} : \mathbf{1}^\top \boldsymbol{x} = B\}$. Let $\boldsymbol{x}^\star$ denote the minimizer of f over $\mathcal{X}$. Then every feasible $\boldsymbol{p} \in \mathcal{X}$ satisfies*

$$f(\boldsymbol{p}) - f(\boldsymbol{x}^\star) \geq \mu \|\boldsymbol{p} - \boldsymbol{x}^\star\|_2^2. \tag{36}$$

Moreover, if

$$\widetilde{\boldsymbol{g}}(\boldsymbol{p}) := \nabla f(\boldsymbol{p}) - \left(\frac{1}{T}\mathbf{1}^\top \nabla f(\boldsymbol{p})\right)\mathbf{1}$$

denotes the time-varying part of the gradient of f at $\boldsymbol{p}$, then

$$\|\boldsymbol{p} - \boldsymbol{x}^\star\|_2 \leq \frac{\|\widetilde{\boldsymbol{g}}(\boldsymbol{p})\|_2}{2\mu} \qquad \textit{and} \qquad f(\boldsymbol{p}) - f(\boldsymbol{x}^\star) \leq \frac{\|\widetilde{\boldsymbol{g}}(\boldsymbol{p})\|_2^2}{4\mu}. \tag{37}$$

Proof. Because f is 2μ-strongly convex,

$$f(\boldsymbol{p}) \geq f(\boldsymbol{x}^\star) + \langle \nabla f(\boldsymbol{x}^\star), \boldsymbol{p} - \boldsymbol{x}^\star \rangle + \mu \|\boldsymbol{p} - \boldsymbol{x}^\star\|_2^2.$$

Since $\boldsymbol{x}^\star$ minimizes f over the convex set $\mathcal{X}$, its first-order condition gives

$$\langle \nabla f(\boldsymbol{x}^\star), \boldsymbol{p} - \boldsymbol{x}^\star \rangle \geq 0.$$

This proves (36).

For the distance bound, strong convexity also implies strong monotonicity of the gradient:

$$2\mu \|\boldsymbol{p} - \boldsymbol{x}^\star\|_2^2 \leq \langle \nabla f(\boldsymbol{p}) - \nabla f(\boldsymbol{x}^\star), \boldsymbol{p} - \boldsymbol{x}^\star \rangle .$$

The first-order condition at $\boldsymbol{x}^\star$ implies $\langle \nabla f(\boldsymbol{x}^\star), \boldsymbol{p} - \boldsymbol{x}^\star \rangle \geq 0$, so

$$2\mu \|\boldsymbol{p} - \boldsymbol{x}^\star\|_2^2 \leq \langle \nabla f(\boldsymbol{p}), \boldsymbol{p} - \boldsymbol{x}^\star \rangle .$$

Now $\mathbf{1}^\top(\boldsymbol{p} - \boldsymbol{x}^\star) = 0$. Therefore subtracting the average component of the gradient does not change this inner product:

$$\langle \nabla f(\boldsymbol{p}), \boldsymbol{p} - \boldsymbol{x}^\star \rangle = \langle \widetilde{\boldsymbol{g}}(\boldsymbol{p}), \boldsymbol{p} - \boldsymbol{x}^\star \rangle .$$

Cauchy–Schwarz gives the first inequality in (37).

Finally, because f is quadratic, its expansion around $\boldsymbol{p}$ is exact. Strong convexity therefore gives

$$f(\boldsymbol{x}^\star) \geq f(\boldsymbol{p}) + \langle \nabla f(\boldsymbol{p}), \boldsymbol{x}^\star - \boldsymbol{p} \rangle + \mu \|\boldsymbol{p} - \boldsymbol{x}^\star\|_2^2.$$

Again the constant component of the gradient disappears because $\mathbf{1}^\top(\boldsymbol{p} - \boldsymbol{x}^\star) = 0$. Hence, with $d := \|\boldsymbol{p} - \boldsymbol{x}^\star\|_2$,

$$f(\boldsymbol{p}) - f(\boldsymbol{x}^\star) \leq \|\widetilde{\boldsymbol{g}}(\boldsymbol{p})\|_2\, d - \mu d^2 \leq \frac{\|\widetilde{\boldsymbol{g}}(\boldsymbol{p})\|_2^2}{4\mu},$$

where the last step maximizes a one-dimensional concave quadratic. □

Lemma D.3 has a useful interpretation. On a fixed-quota simplex, a gradient component that is the same in every period cannot create an incentive to move volume across time. What matters is only the variation of marginal cost across periods. If that variation is small, the minimizer is close to the reference schedule and the possible cost improvement is even smaller.

We will use the lemma three times below: first at the uniform schedule with a clean entering state $z = 0$; second to control the clean-state value gap that appears when $z > 0$; and third at a delayed-uniform comparison schedule in the same $z > 0$ regime. The permanent case is different and is analyzed directly.

D.3 The Proof of Theorem D.2

We are ready to proceed with the proof of Theorem D.2. We treat the three regimes separately.

D.3.1 Transient Impact and Clean Entering State

Assume $0 < \gamma < 1$ and $z = 0$. We first compute how much the marginal cost at the uniform schedule varies across time. At the uniform strategy, the tth marginal cost of agent i is

$$g_{i,t}^{\boldsymbol{u}} = \frac{\kappa(V + B_i)}{T} + \frac{\lambda V}{T} h_{t-1}(\gamma) + \frac{\lambda B_i}{T} h_{T-t}(\gamma).$$

The first term is constant across periods. Using the geometric-sum formula, we obtain that there is a $c_{\mathrm{u},i,T}$, which does not depend on t, such that

$$g_{i,t}^{\boldsymbol{u}} = c_{\mathrm{u},i,T} - \frac{\lambda}{T(1-\gamma)} \left(V\gamma^{t-1} + B_i\gamma^{T-t} \right).$$

Thus all time variation comes from two geometric terms: one decays away from the beginning of the horizon and the other decays away from the end. This is the finite-horizon boundary effect behind the upper bounds in the first line of (35).

Let

$$\bar{d}_\gamma := \frac{1}{T} \sum_{t=1}^{T} \gamma^{t-1} = \frac{h_T(\gamma)}{T},$$

and define the centered geometric vectors

$$\widetilde{\boldsymbol{d}}_\gamma := (1, \gamma, \ldots, \gamma^{T-1})^\top - \bar{d}_\gamma \mathbf{1} \qquad \text{and} \qquad \widetilde{\boldsymbol{w}}_\gamma := (\gamma^{T-1}, \ldots, \gamma, 1)^\top - \bar{d}_\gamma \mathbf{1}.$$

Then the time-varying part of the gradient at $\boldsymbol{u}_i$ is

$$\widetilde{\boldsymbol{g}}_i^{\boldsymbol{u}} = -\frac{\lambda}{T(1-\gamma)} \left(V\widetilde{\boldsymbol{d}}_\gamma + B_i \widetilde{\boldsymbol{w}}_\gamma \right).$$

Both centered geometric vectors have squared norm

$$\delta_T^2(\gamma) := \sum_{r=0}^{T-1} \gamma^{2r} - \frac{h_T(\gamma)^2}{T}$$

and inner product

$$\left\langle \widetilde{\boldsymbol{d}}_\gamma, \widetilde{\boldsymbol{w}}_\gamma \right\rangle = T\gamma^{T-1} - \frac{h_T(\gamma)^2}{T}.$$

For fixed $\gamma < 1$,

$$\delta_T^2(\gamma) \longrightarrow \frac{1}{1-\gamma^2} \qquad \text{and} \qquad T\gamma^{T-1} - \frac{h_T(\gamma)^2}{T} \longrightarrow 0.$$

Hence

$$T^2 \left\| \widetilde{\boldsymbol{g}}_i^{\boldsymbol{u}} \right\|_2^2 \longrightarrow \frac{\lambda^2 (V^2 + B_i^2)}{(1-\gamma)^2(1-\gamma^2)}.$$

We can now invoke Lemma D.3 with $f = f_i$, $\boldsymbol{p} = \boldsymbol{u}_i$, and $\boldsymbol{x}^\star = \boldsymbol{x}_i^{\mathrm{br}}$. Since f_i is $2\mu_\gamma$-strongly convex,

$$D_i^{\boldsymbol{u}} \le \frac{\|\widetilde{\boldsymbol{g}}_i^{\boldsymbol{u}}\|_2}{2\mu_\gamma} \qquad \text{and} \qquad E_i^{\boldsymbol{u}} \le \frac{\|\widetilde{\boldsymbol{g}}_i^{\boldsymbol{u}}\|_2^2}{4\mu_\gamma}.$$

Using that $B_i \le V$ and $\delta_T(\gamma) \le (1-\gamma^2)^{-1/2} \le (1-\gamma)^{-1/2}$, we obtain the asymptotic upper bounds on $D_i^{\boldsymbol{u}}$ and $E_i^{\boldsymbol{u}}$ in the first line of (35).

Intuitively, under transient impact, uniform play is not exactly optimal because the beginning and end of a finite horizon are not symmetric with the middle. Early periods have little accumulated past impact, while late periods have little future self-impact. But these distortions live in geometric boundary layers of fixed effective width of order $(1-\gamma)^{-1}$. As T grows, their contribution to the full T-dimensional marginal-cost vector is only $O(T^{-1})$, and the quadratic comparison lemma turns this directly into the two rates above.

D.3.2 Transient Impact with a Positive Entering State

Now assume $0 < \gamma < 1$ and $z > 0$. The new term $z\boldsymbol{d}_\gamma$ in the marginal cost does not become flat as T grows and its first coordinates may be up to $O(1)$. Thus applying the previous argument to the uniform schedule would only give a coarse $O(1)$ bound. Indeed, the uniform schedule is genuinely poor in the first few periods, because it insists on buying while the inherited state is still large.

The natural comparison schedule waits for a short initial period and then trades uniformly over the remaining horizon. Precisely, we let f_i^0 be the objective with the same uniform competitors but with an entering state equal to 0. Then $f_i^z(\boldsymbol{x}) = f_i^0(\boldsymbol{x}) + z\langle \boldsymbol{d}_\gamma, \boldsymbol{x}\rangle$, where we denote $f_i^z(\boldsymbol{x}) := f_i(\boldsymbol{x})$ to distinguish between the two versions of the cost. The added term is nonnegative on the feasible set.

Our first use of Lemma D.3 in this regime is to control the clean-state benchmark. Applying Lemma D.3 to f_i^0 at $\boldsymbol{u}_i$ and using the clean-state residual bound above gives

$$0 \le f_i^0(\boldsymbol{u}_i) - \min_{\boldsymbol{x}\in\mathcal{X}_i} f_i^0(\boldsymbol{x}) = O(T^{-2}). \tag{38}$$

Now, we define an "almost uniform" schedule $\boldsymbol{y}$ which skips the first $O(\log T)$ periods and is uniform in the rest of the time horizon:

$$L_T := \left\lceil \frac{3\log T}{|\log\gamma|} \right\rceil, \qquad \text{and} \qquad y_t := \begin{cases} 0, & t \le L_T, \\ B_i/(T-L_T), & t > L_T. \end{cases}$$

For large enough T, $L_T < T$, so $\boldsymbol{y}$ is feasible. We first show that the clean-state cost $f_i^0(\boldsymbol{y})$ of the "almost uniform" $\boldsymbol{y}$ is within $O(\log T/T^2)$ from the clean-state cost $f_i^0(\boldsymbol{u}_i)$ of the uniform policy. Since $\boldsymbol{y}$ skips only $O(\log T)$ periods,

$$\|\boldsymbol{y} - \boldsymbol{u}_i\|_2^2 = \frac{B_i^2 L_T}{T(T-L_T)} = O\left(\frac{\log T}{T^2}\right). \tag{39}$$

We next check that this small delay costs little under the clean objective. Because f_i^0 is quadratic, its exact expansion around $\boldsymbol{u}_i$ gives

$$f_i^0(\boldsymbol{y}) - f_i^0(\boldsymbol{u}_i) = \left\langle \nabla f_i^0(\boldsymbol{u}_i), \boldsymbol{y} - \boldsymbol{u}_i \right\rangle + \kappa\|\boldsymbol{y} - \boldsymbol{u}_i\|_2^2 + \lambda K_\gamma(\boldsymbol{y} - \boldsymbol{u}_i).$$

The difference $\boldsymbol{y} - \boldsymbol{u}_i$ has zero coordinate sum, so the constant part of the gradient disappears. The linear term is therefore bounded by

$$\|\widetilde{\boldsymbol{g}}_i^{\boldsymbol{u}}\|_2\|\boldsymbol{y} - \boldsymbol{u}_i\|_2 = O(T^{-1})O\left(\frac{\sqrt{\log T}}{T}\right).$$

For the quadratic term we use the elementary bound

$$\kappa\|\boldsymbol{y} - \boldsymbol{u}_i\|_2^2 + \lambda K_\gamma(\boldsymbol{y} - \boldsymbol{u}_i) \le \left(\kappa + \frac{\lambda}{1-\gamma}\right)\|\boldsymbol{y} - \boldsymbol{u}_i\|_2^2.$$

Together with (39), this gives

$$f_i^0(\boldsymbol{y}) - f_i^0(\boldsymbol{u}_i) = O\left(\frac{\log T}{T^2}\right).$$

The benefit of waiting is that the inherited state has essentially disappeared before trading starts. By construction, $\gamma^{L_T} \le T^{-3}$, so

$$z\langle \boldsymbol{d}_\gamma, \boldsymbol{y}\rangle = \frac{zB_i}{T-L_T}\sum_{t=L_T+1}^{T}\gamma^{t-1} \le \frac{zB_i\gamma^{L_T}}{(T-L_T)(1-\gamma)} = O(T^{-4}).$$

Hence

$$f_i^z(\boldsymbol{y}) \le f_i^0(\boldsymbol{u}_i) + O\left(\frac{\log T}{T^2}\right).$$

Let $\boldsymbol{x}_i^{\mathrm{br}}$ now denote the true best response with entering state z. Optimality gives $f_i^z(\boldsymbol{x}_i^{\mathrm{br}}) \le f_i^z(\boldsymbol{y})$, while the nonnegativity of the inherited-state term and (38) give

$$f_i^z(\boldsymbol{x}_i^{\mathrm{br}}) \ge \min_{\boldsymbol{x}\in\mathcal{X}_i} f_i^0(\boldsymbol{x}) \ge f_i^0(\boldsymbol{u}_i) - O(T^{-2}).$$

Combining these two inequalities,

$$f_i^0(\boldsymbol{u}_i) - O(T^{-2}) \le f_i^z(\boldsymbol{x}_i^{\mathrm{br}}) \le f_i^0(\boldsymbol{u}_i) + O\left(\frac{\log T}{T^2}\right). \tag{40}$$

Interestingly, inequality above determines the value of the unknown best response, up to an $O(\log T/T^2)$ term, without solving for the best response itself.

We now use the basic part of Lemma D.3 for the third time, with the feasible comparison point $\boldsymbol{p} = \boldsymbol{y}$. Since f_i^z is $2\mu_\gamma$-strongly convex,

$$\mu_\gamma\|\boldsymbol{y} - \boldsymbol{x}_i^{\mathrm{br}}\|_2^2 \le f_i^z(\boldsymbol{y}) - f_i^z(\boldsymbol{x}_i^{\mathrm{br}}) = O\left(\frac{\log T}{T^2}\right) \qquad \text{and} \qquad \|\boldsymbol{y} - \boldsymbol{x}_i^{\mathrm{br}}\|_2 = O\left(\frac{\sqrt{\log T}}{T}\right).$$

Combining this with (39) and the triangle inequality yields

$$D_i^{\boldsymbol{u}} = O\left(B_i\frac{\sqrt{\log T}}{T}\right).$$

For exploitability we can identify the leading term. By definition,

$$E_i^{\boldsymbol{u}} = f_i^z(\boldsymbol{u}_i) - f_i^z(\boldsymbol{x}_i^{\mathrm{br}}) = z\langle \boldsymbol{d}_\gamma, \boldsymbol{u}_i\rangle + \left[f_i^0(\boldsymbol{u}_i) - f_i(\boldsymbol{x}_i^{\mathrm{br}})\right].$$

The bracket is $O(\log T/T^2)$ by (40). Moreover,

$$z\langle \boldsymbol{d}_\gamma, \boldsymbol{u}_i\rangle = \frac{zB_i}{T}h_T(\gamma) = \frac{B_i z}{1-\gamma}\frac{1}{T} - \frac{B_i z\,\gamma^T}{1-\gamma}\frac{1}{T}.$$

For fixed $\gamma < 1$, the last term decays exponentially and is, in particular, $O(T^{-2})$. Hence

$$E_i^{\boldsymbol{u}} = \frac{B_i z}{1-\gamma}\frac{1}{T} + O\left(\frac{\log T}{T^2}\right).$$

Intuitively, a positive entering state does not make the entire uniform policy bad. It makes only the first few periods unattractive. After $O(\log T)$ periods the state has decayed to a negligible level. The best response therefore differs from uniform mainly by skipping this short initial window and redistributing that volume over the rest of the horizon. That is why the schedule distance still vanishes, although more slowly than in the clean-state case.

D.4 Permanent Impact

Finally we assume that $\gamma = 1$ and $\mu_1 = \kappa - \frac{\lambda}{2} > 0$. The entering state $z \ge 0$ contributes $z\sum_{t=1}^{T} x_t = zB_i$ to every feasible schedule and thus, it is irrelevant for the best response. Also, if $V_{-i} = 0$, there is no competing flow and the uniform schedule is the unique best response. Hence $D_i^{\boldsymbol{u}} = E_i^{\boldsymbol{u}} = 0$. So, we assume that $V_{-i} > 0$.

For $\gamma = 1$,

$$K_1(\boldsymbol{x}) = \sum_{s<t} x_s x_t = \frac{1}{2}\left(B_i^2 - \|\boldsymbol{x}\|_2^2\right).$$

Therefore, up to an additive term of $\lambda B_i^2/2$, which does not depend on the timing of $\boldsymbol{x}$, the agent's own-impact cost is simply $\mu_1\|\boldsymbol{x}\|_2^2$.

Uniform competitors buy V_{-i}/T in every period. Because their impact is permanent, by period t they have created the state $\lambda V_{-i}(t-1)/T$. The best-response problem therefore reduces to

$$\min_{\boldsymbol{x}\in\mathcal{X}_i}\left\{\mu_1\|\boldsymbol{x}\|_2^2 + c_{\mathrm{perm},i,T}\sum_{t=1}^{T}(t-1)x_t\right\} \qquad \text{where} \qquad c_{\mathrm{perm},i,T} := \frac{\lambda V_{-i}}{T}. \tag{41}$$

The second term is a linear time tilt. Waiting one more period becomes progressively more expensive because competing impact never decays. The best response must therefore be front-loaded.

The KKT conditions give

$$x_{i,t}^{\mathrm{br}} = \left[\frac{\nu - c_{\mathrm{perm},i,T}(t-1)}{2\mu_1}\right]_+,$$

where ν enforces the quota constraint. Since the expression inside the brackets decreases with t, the positive coordinates form an initial segment $\{1,\dots,m\}$. Summing over these m active periods and imposing $\sum_t x_{i,t}^{\mathrm{br}} = B_i$ gives

$$x_{i,t}^{\mathrm{br}} = \frac{B_i}{m} + \frac{c_{\mathrm{perm},i,T}}{4\mu_1}(m+1-2t), \qquad \text{for} \quad t \le m, \tag{42}$$

and $x_{i,t}^{\mathrm{br}} = 0$ for $t > m$. The cutoff m is determined by requiring the m-th coordinate to remain positive while the next unconstrained coordinate is non-positive. This gives

$$c_{\mathrm{perm},i,T}\,m(m-1) < 4\mu_1 B_i \le c_{\mathrm{perm},i,T}\,m(m+1). \tag{43}$$

Since $c_{\mathrm{perm},i,T} = \lambda V_{-i}/T$, it follows that

$$m = (1+o(1))\sqrt{\frac{4\mu_1 B_i T}{\lambda V_{-i}}}. \tag{44}$$

Thus the best response trades only during $\Theta(\sqrt{T})$ early periods. The active window grows with the horizon, but its fraction of the horizon vanishes.

To compute the distance from uniform, we use

$$\sum_{t=1}^{m}(m+1-2t) = 0 \qquad \text{and} \qquad \sum_{t=1}^{m}(m+1-2t)^2 = \frac{m(m^2-1)}{3}.$$

Substituting (42) gives

$$(D_i^{\boldsymbol{u}})^2 = \frac{B_i^2}{m} - \frac{B_i^2}{T} + \frac{c_{\mathrm{perm},i,T}^2\, m(m^2-1)}{48\mu_1^2}.$$

Using (43), $c_{\mathrm{perm},i,T}m^2 \to 4\mu_1 B_i$, and hence

$$(D_i^{\boldsymbol{u}})^2 = \left(\frac{4}{3} + o(1)\right)\frac{B_i^2}{m}.$$

Together with (44),

$$D_i^{\boldsymbol{u}} = (1+o(1))\sqrt{\frac{2}{3}}\,B_i^{3/4}\left(\frac{\lambda V_{-i}}{\mu_1 T}\right)^{1/4}.$$

Finally, evaluate the reduced objective in (41) at the uniform schedule and at the best response. At uniform, its value is $\frac{\mu_1 B_i^2}{T} + \frac{c_{\mathrm{perm},i,T}B_i(T-1)}{2}$. Using (42) and the elementary sums above, we obtain that the value of the reduced objective at $\boldsymbol{x}_i^{\mathrm{br}}$ is:

$$\frac{\mu_1 B_i^2}{m} + \frac{c_{\mathrm{perm},i,T}B_i(m-1)}{2} - \frac{c_{\mathrm{perm},i,T}^2\, m(m^2-1)}{48\mu_1}.$$

Subtracting gives

$$E_i^{\boldsymbol{u}} = \mu_1 B_i^2 \left(\frac{1}{T} - \frac{1}{m}\right) + \frac{c_{\mathrm{perm},i,T} B_i (T-m)}{2} + \frac{c_{\mathrm{perm},i,T}^2 m(m^2-1)}{48\mu_1} = \frac{\lambda B_i V_{-i}}{2} + O(T^{-1/2}).$$

The leading term comes from avoiding the state accumulated by uniform competitors. Unlike the transient cases, this gain does not vanish. □

E Learning across Episodes: Proofs and Extensions

This appendix proves the results of Section 3. We start with the price-based gradient and its norm, and then give the standard projected-gradient arguments for regret and equilibrium convergence. The remaining subsections treat private and varying quotas, inherited state, noise, and parameter error.

E.1 The Price Gradient and Its Size

For convenience in the proofs, we introduce matrix $\mathrm{A}_\gamma := \kappa \mathrm{I}_T + \lambda \mathrm{L}_\gamma$ which maps aggregate demand to current price. From Appendix A, we recall that $\boldsymbol{d}_\gamma = (1, \gamma, \ldots, \gamma^{T-1})^\top$, that $s_T(\gamma) = \sum_{r=0}^{T-1} \gamma^{2r} = \|\boldsymbol{d}_\gamma\|_2^2$ and that the Hessian of the individual cost is $\mathrm{H}_\gamma = 2\kappa\, \mathrm{I}_T + \lambda\, (\mathrm{L}_\gamma + \mathrm{U}_\gamma)$.

Proposition E.1. *Let $\boldsymbol{g}_i$ denote the gradient of agent i's cost in* (5)*. Then*

$$\|\boldsymbol{g}_i\|_2 \le \mathcal{G}_i(z) := |z| \sqrt{s_T(\gamma)} + \|\boldsymbol{q}_{-i}\|_1 \sqrt{\kappa^2 + \lambda^2 s_T(\gamma)} + \|\boldsymbol{x}_i\|_1 \sqrt{4\kappa^2 + 2\lambda^2 s_T(\gamma)}.$$

Proof. Using the matrix form introduced in Appendix A, the gradient of agent i's cost in (5) (see also (13) in Appendix A) can be written as

$$\boldsymbol{g}_i = \boldsymbol{P} + \kappa \boldsymbol{x}_i + \lambda \mathrm{U}_\gamma \boldsymbol{x}_i = z \boldsymbol{d}_\gamma + \mathrm{H}_\gamma \boldsymbol{x}_i + \mathrm{A}_\gamma \boldsymbol{q}_{-i}. \tag{45}$$

By the triangle inequality, $\|\boldsymbol{g}_i\|_2 \le |z|\, \|\boldsymbol{d}_\gamma\|_2 + \|\mathrm{A}_\gamma \boldsymbol{q}_{-i}\|_2 + \|\mathrm{H}_\gamma \boldsymbol{x}_i\|_2$.

For any matrix $\mathrm{A} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{b} \in \mathbb{R}^n$,

$$\|\mathrm{A}\, \boldsymbol{b}\|_2 = \left\| \sum_{s=1}^{n} b_s\, \mathrm{A}\, \boldsymbol{e}_s \right\|_2 \le \|\boldsymbol{b}\|_1 \max_s \|\mathrm{A}\, \boldsymbol{e}_s\|_2.$$

The columns of A_γ and H_γ satisfy $\|\mathrm{A}_\gamma \boldsymbol{e}_s\|_2^2 \le \kappa^2 + \lambda^2 s_T(\gamma)$ and $\|\mathrm{H}_\gamma \boldsymbol{e}_s\|_2^2 \le 4\kappa^2 + 2\lambda^2 s_T(\gamma)$. Combining the three bounds concludes the proof of the proposition. □

Signed Schedules. The buy-only constraint enters Proposition E.1 only through $\|\boldsymbol{x}_i\|_1 = B_i$ and $\|\boldsymbol{q}_{-i}\|_1 = V_{-i}$. Suppose now that the agents may also sell, i.e., agent j chooses $\boldsymbol{x}_j \in \mathcal{X}_j^M := \{\boldsymbol{x} \in \mathbb{R}^T : \mathbf{1}^\top \boldsymbol{x} = B_j,\ x_t \ge -M \text{ for all } t\}$ for some $M \ge 0$. Theorem 2.1 and its proof do not use the sign of the schedules (Lemma C.1 holds for every $\boldsymbol{v} \in \mathbb{R}^T$), and $\mathcal{X}_j^M$ is compact and convex, so the game still has a unique Nash equilibrium.

Proposition E.2. *If $\boldsymbol{x}_j \in \mathcal{X}_j^M$ for every agent j, then*

$$\|\boldsymbol{g}_i\|_2 \le \mathcal{G}_i^M(z) := |z| \sqrt{s_T(\gamma)} + \big(V_{-i} + 2(N-1)(T-1)M\big) \sqrt{\kappa^2 + \lambda^2 s_T(\gamma)} + \big(B_i + 2(T-1)M\big) \sqrt{4\kappa^2 + 2\lambda^2 s_T(\gamma)}.$$

Proof. The proof of Proposition E.1 gives

$$\|\boldsymbol{g}_i\|_2 \le |z| \sqrt{s_T(\gamma)} + \|\boldsymbol{q}_{-i}\|_1 \sqrt{\kappa^2 + \lambda^2 s_T(\gamma)} + \|\boldsymbol{x}_i\|_1 \sqrt{4\kappa^2 + 2\lambda^2 s_T(\gamma)}$$

for arbitrary signed schedules. For $\boldsymbol{x} \in \mathcal{X}_j^M$, write $x_t = x_t^+ - x_t^-$ with $x_t^+, x_t^- \ge 0$. Then

$$\|\boldsymbol{x}\|_1 = \mathbf{1}^\top \boldsymbol{x} + 2 \sum_t x_t^- = B_j + 2 \sum_t x_t^-.$$

Every $x_t^- \le M$, and, if $B_j > 0$, at most $T-1$ coordinates are negative. Hence $\|\boldsymbol{x}\|_1 \le B_j + 2(T-1)M$, and $\|\boldsymbol{q}_{-i}\|_1 \le \sum_{j \ne i} \|\boldsymbol{x}_j\|_1 \le V_{-i} + 2(N-1)(T-1)M$. □

Consequently, Parts 1 and 2 of Theorem 3.1, Theorem 3.2 and the results of Appendix E.5 hold for signed schedules in $\mathcal{X}_i^M$, with the same rates and with $\mathcal{G}_i(z)$ (and $\mathcal{G}_{i,b}$) replaced by $\mathcal{G}_i^M(z)$ (and the analogous bound with $B_i = b$). In Proposition E.10 and in the constant c_{bias} of Theorem E.14, B_i is replaced by $B_i + 2(T-1)M$, since

$$\|\mathrm{D}\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1 \max_s \|\mathrm{D}\boldsymbol{e}_s\|_2.$$

Part 3 of Theorem 3.1 does not use the sign constraint at all. However, the bound is no longer independent of T for $\gamma < 1$, and some growth in T is unavoidable. E.g., for $N = 1$, $z = 0$, even T and $\boldsymbol{x}_1 = (B_1/T)\mathbf{1} + M((-1)^t)_{t=1}^T \in \mathcal{X}_1^M$, the bound $\mathrm{H}_\gamma \succeq 2\mu_\gamma \mathrm{I}_T$ of Appendix C gives

$$\|\boldsymbol{g}_1\|_2 = \|\mathrm{H}_\gamma \boldsymbol{x}_1\|_2 \ge 2\mu_\gamma \|\boldsymbol{x}_1\|_2 \ge 2\mu_\gamma M\sqrt{T}.$$

Without a lower bound on the actions, the gradients are unbounded.

E.2 Fixed Quotas and a Reset State: The Proof of Theorem 3.1

We now prove Theorem 3.1. The proof of (7) follows from the standard analysis of online gradient descent for strongly convex losses (Hazan et al., 2007). We start from the nonexpansiveness of Euclidean projection. For every comparator $\boldsymbol{u}_i$,

$$\|\boldsymbol{x}_i^{k+1} - \boldsymbol{u}_i\|_2^2 \le \|\boldsymbol{x}_i^k - \boldsymbol{u}_i\|_2^2 - 2\eta_k \langle \boldsymbol{g}_i^k, \boldsymbol{x}_i^k - \boldsymbol{u}_i \rangle + \eta_k^2 \|\boldsymbol{g}_i^k\|_2^2. \tag{46}$$

Strong convexity converts the inner product into the one-episode regret plus $\mu_\gamma \|\boldsymbol{x}_i^k - \boldsymbol{u}_i\|_2^2$. With $\eta_k = 1/(\mu_\gamma(k+1))$, the distance terms cancel when summed leaving only nonpositive residual terms and $\sum_{k \le K} 1/(k+1) \le 1 + \log(K+1)$.

For simultaneous play, we follow the standard analysis of online projected gradient-learning for strongly monotone games (Mertikopoulos & Zhou, 2019). Summing (46) over agents and using strong monotonicity of the pseudo-gradient gives a recursion for $\|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}}^2$. A diminishing step yields the $O(1/K)$ last-iterate rate. A constant step needs an upper Lipschitz bound L_N (which depends on the number N of agents) on the joint pseudo-gradient.

The model-specific part concerns that prices provide the exact gradient, the gradient norm is bounded as in Proposition E.1, and that Theorem 2.1 establishes strong convexity of the individual cost and strong monotonicity of the pseudo-gradient.

Proof of Theorem 3.1. We prove the three statements separately.

Part 1. Fix an agent i and an arbitrary comparator $\boldsymbol{u}_i \in \mathcal{X}_i$. Since Euclidean projection is nonexpansive, the update of agent i satisfies

$$\|\boldsymbol{x}_i^{k+1} - \boldsymbol{u}_i\|_2^2 \le \|\boldsymbol{x}_i^k - \boldsymbol{u}_i\|_2^2 - 2\eta_k \langle \boldsymbol{g}_i^k, \boldsymbol{x}_i^k - \boldsymbol{u}_i \rangle + \eta_k^2 \|\boldsymbol{g}_i^k\|_2^2. \tag{47}$$

Moreover, the cost of agent i is strongly convex in its own schedule $\boldsymbol{x}_i$, and therefore

$$C_i^k(\boldsymbol{x}_i^k) - C_i^k(\boldsymbol{u}_i) \le \langle \boldsymbol{g}_i^k, \boldsymbol{x}_i^k - \boldsymbol{u}_i \rangle - \mu_\gamma \|\boldsymbol{x}_i^k - \boldsymbol{u}_i\|_2^2 .$$

Combining the two inequalities and using $\|\boldsymbol{g}_i^k\|_2 \le \mathcal{G}_i(z)$, we obtain that:

$$C_i^k(\boldsymbol{x}_i^k) - C_i^k(\boldsymbol{u}_i) \le \frac{\|\boldsymbol{x}_i^k - \boldsymbol{u}_i\|_2^2 - \|\boldsymbol{x}_i^{k+1} - \boldsymbol{u}_i\|_2^2}{2\eta_k} - \mu_\gamma \|\boldsymbol{x}_i^k - \boldsymbol{u}_i\|_2^2 + \frac{\eta_k}{2} \mathcal{G}_i(z)^2.$$

We set $\eta_k = 1/(\mu_\gamma(k+1))$ and sum over $k = 1, \dots, K$. Then, the distance terms cancel and only nonpositive terms survive. Hence,

$$\sum_{k=1}^K \big(C_i^k(\boldsymbol{x}_i^k) - C_i^k(\boldsymbol{u}_i)\big) \le \frac{\mathcal{G}_i(z)^2}{2\mu_\gamma} \sum_{k=1}^K \frac{1}{k+1}.$$

Using $\sum_{k=1}^K 1/(k+1) \le 1 + \log(K+1)$ and minimizing over $\boldsymbol{u}_i \in \mathcal{X}_i$ yields the result.

Part 2. Let X^k denote the matrix obtained by stacking the agents' strategies $\boldsymbol{x}_i^k$, and similarly let $\mathrm{X}^\star$ denote the Nash equilibrium. Thus,

$$\|\mathrm{X}^k - \mathrm{X}^\star\|_F^2 = \sum_{i=1}^N \|\boldsymbol{x}_i^k - \boldsymbol{x}_i^\star\|_2^2.$$

Let $\mathrm{F}(\mathrm{X}^k)$ be the matrix obtained by stacking the individual gradients $\boldsymbol{g}_i^k$ in the same way. Then

$$\|\mathrm{F}(\mathrm{X}^k)\|_{\mathrm{F}}^2 = \sum_{i=1}^N \|\boldsymbol{g}_i^k\|_2^2 \qquad \text{and} \qquad \langle \mathrm{F}(\mathrm{X}^k), \mathrm{X}^k - \mathrm{X}^\star\rangle_{\mathrm{F}} = \sum_{i=1}^N \langle \boldsymbol{g}_i^k, \boldsymbol{x}_i^k - \boldsymbol{x}_i^\star\rangle,$$

where $\|\mathrm{F}(\mathrm{X}^k)\|_{\mathrm{F}}^2 = \sum_{i=1}^N \|\boldsymbol{g}_i^k\|_2^2 \le \sum_{i=1}^N \mathcal{G}_i(z)^2$.

Applying (47) to each agent with $\boldsymbol{u}_i = \boldsymbol{x}_i^\star$ and summing over i gives

$$\|\mathrm{X}^{k+1} - \mathrm{X}^\star\|_{\mathrm{F}}^2 \le \|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}}^2 - 2\eta_k \langle \mathrm{F}(\mathrm{X}^k), \mathrm{X}^k - \mathrm{X}^\star\rangle_{\mathrm{F}} + \eta_k^2 \|\mathrm{F}(\mathrm{X}^k)\|_{\mathrm{F}}^2 . \tag{48}$$

Since $\mathrm{X}^\star$ is a Nash equilibrium, it satisfies the corresponding variational inequality,

$$\langle \mathrm{F}(\mathrm{X}^\star), \mathrm{X}^k - \mathrm{X}^\star\rangle_{\mathrm{F}} \ge 0.$$

Moreover, setting $\mathrm{X} = \mathrm{X}^k$ and $\mathrm{Y} = \mathrm{X}^\star$ in Theorem 2.1 and combining these two inequalities yields

$$\langle \mathrm{F}(\mathrm{X}^k), \mathrm{X}^k - \mathrm{X}^\star\rangle_{\mathrm{F}} \ge \mu_\gamma \|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}}^2 .$$

Therefore replacing $\eta_k = \frac{1}{\mu_\gamma (k+1)}$ in (48) and applying the bounds above we obtain

$$\|\mathrm{X}^{k+1} - \mathrm{X}^\star\|_{\mathrm{F}}^2 \le \frac{k-1}{k+1} \|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}}^2 + \frac{1}{(k+1)^2} \frac{\sum_i \mathcal{G}_i(z)^2}{\mu_\gamma^2}.$$

A direct induction[1] on the recursion gives

$$\|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}}^2 \le \frac{1}{k} \max\left\{ \|\mathrm{X}^1 - \mathrm{X}^\star\|_{\mathrm{F}}^2, \frac{\sum_i \mathcal{G}_i(z)^2}{\mu_\gamma^2} \right\}.$$

Part 3. Assume $\gamma < 1$ and that an upper bound N on the number of agents is known. The pseudo-gradient is μ_γ-strongly monotone and L_N-Lipschitz, where

$$L_N = (N+1)\left(\kappa + \frac{\lambda}{1-\gamma}\right).$$

To make the Lipschitz estimate explicit, recall that the pseudo-gradient is

$$\mathrm{F}(\mathrm{X}) = \big(\boldsymbol{g}_1(\mathrm{X}), \ldots, \boldsymbol{g}_N(\mathrm{X})\big).$$

Using the expression for the individual gradients, its block derivatives are

$$\frac{\partial \boldsymbol{g}_i}{\partial \boldsymbol{x}_i} = 2\kappa\, \mathrm{I}_T + \lambda(\mathrm{L}_\gamma + \mathrm{U}_\gamma) \qquad \text{and} \qquad \frac{\partial \boldsymbol{g}_i}{\partial \boldsymbol{x}_j} = \kappa\, \mathrm{I}_T + \lambda\, \mathrm{L}_\gamma, \quad \text{for all } j \ne i,$$

where $\mathrm{U}_\gamma = \mathrm{L}_\gamma^\top$. For $\gamma < 1$, $\|\mathrm{L}_\gamma\|_2 = \|\mathrm{U}_\gamma\|_2 \le \frac{1}{1-\gamma}$. Hence every diagonal block has norm at most $2\left(\kappa + \frac{\lambda}{1-\gamma}\right)$, while every off-diagonal block has norm at most $\kappa + \frac{\lambda}{1-\gamma}$. Thus the block row and column sums of the Jacobian are bounded by

$$L_N := (N+1)\left(\kappa + \frac{\lambda}{1-\gamma}\right).$$

[1] For the induction, let $D_k := \|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}}^2$ and $G := \frac{\sum_i \mathcal{G}_i(z)^2}{\mu_\gamma^2}$. Then, the recursion becomes $D_{k+1} \le \frac{k-1}{k+1} D_k + \frac{G}{(k+1)^2}$. Setting $M := \max\{D_1, G\}$, $D_1 \le M$ and we can show, by induction, that $D_k \le M/k$.

Since the pseudo-gradient is affine, we have that $\|\mathrm{F}(\mathrm{X})-\mathrm{F}(\mathrm{Y})\|_{\mathrm{F}} \le L_N\|\mathrm{X}-\mathrm{Y}\|_{\mathrm{F}}$, which concludes the proof of the claim that the pseudo-gradient is L_N-Lipschitz.

Since $\mathrm{X}^\star$ is the Nash equilibrium, it satisfies the corresponding variational inequality and therefore, for every $\eta > 0$, $\mathrm{X}^\star = \mathsf{P}_{\mathcal{X}}\big(\mathrm{X}^\star - \eta\mathrm{F}(\mathrm{X}^\star)\big)$. The projected-gradient iteration is

$$\mathrm{X}^{k+1} = \mathsf{P}_{\mathcal{X}}\big(\mathrm{X}^k - \eta\mathrm{F}(\mathrm{X}^k)\big).$$

Using non-expansiveness of the Euclidean projection,

$$\begin{aligned}\|\mathrm{X}^{k+1} - \mathrm{X}^\star\|_{\mathrm{F}}^2 &\le \big\|(\mathrm{X}^k - \mathrm{X}^\star) - \eta\big(\mathrm{F}(\mathrm{X}^k) - \mathrm{F}(\mathrm{X}^\star)\big)\big\|_{\mathrm{F}}^2 \\ &= \|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}}^2 - 2\eta\left\langle \mathrm{F}(\mathrm{X}^k) - \mathrm{F}(\mathrm{X}^\star), \mathrm{X}^k - \mathrm{X}^\star\right\rangle + \eta^2\|\mathrm{F}(\mathrm{X}^k) - \mathrm{F}(\mathrm{X}^\star)\|_{\mathrm{F}}^2 .\end{aligned}$$

By μ_γ-strong monotonicity,

$$\left\langle \mathrm{F}(\mathrm{X}^k) - \mathrm{F}(\mathrm{X}^\star), \mathrm{X}^k - \mathrm{X}^\star\right\rangle \ge \mu_\gamma\|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}}^2 ,$$

while L_N-Lipschitz continuity gives

$$\|\mathrm{F}(\mathrm{X}^k) - \mathrm{F}(\mathrm{X}^\star)\|_{\mathrm{F}} \le L_N\|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}} .$$

Combining the two bounds yields

$$\|\mathrm{X}^{k+1} - \mathrm{X}^\star\|_{\mathrm{F}}^2 \le \big(1 - 2\eta\mu_\gamma + \eta^2 L_N^2\big)\|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}}^2.$$

Since $\mathrm{X}^\star$ is a fixed point of the projected-gradient map, non-expansiveness of the projection gives

$$\|\mathrm{X}^{k+1} - \mathrm{X}^\star\|_{F}^2 \le \big(1 - 2\eta\mu_\gamma + \eta^2 L_N^2\big)\|\mathrm{X}^k - \mathrm{X}^\star\|_{F}^2.$$

Hence the iteration is a contraction if $0 < \eta < 2\mu_\gamma/L_N^2$, with contraction factor $1 - 2\eta\mu_\gamma + \eta^2 L_N^2 < 1$. Choosing $\eta = \mu_\gamma/L_N^2$ yields

$$\|\mathrm{X}^{k+1} - \mathrm{X}^\star\|_{F}^2 \le \left(1 - \frac{\mu_\gamma^2}{L_N^2}\right)\|\mathrm{X}^k - \mathrm{X}^\star\|_{F}^2.$$

Iterating this inequality and taking square roots gives the result.

It remains to show that geometric convergence to $\mathrm{X}^\star$ implies bounded cumulative regret. Let

$$\rho := 1 - \frac{\mu_\gamma^2}{L_N^2} \in (0, 1),$$

Fix an agent i and any comparator $\boldsymbol{u}_i \in \mathcal{X}_i$. For every episode k, decompose

$$\begin{aligned}C_i(\boldsymbol{x}_i^k, \mathrm{X}_{-i}^k) - C_i(\boldsymbol{u}_i, \mathrm{X}_{-i}^k) = \\ \Big[C_i(\mathrm{X}^k) - C_i(\mathrm{X}^\star)\Big] + \Big[C_i(\mathrm{X}^\star) - C_i(\boldsymbol{u}_i, \mathrm{X}_{-i}^\star)\Big] + \Big[C_i(\boldsymbol{u}_i, \mathrm{X}_{-i}^\star) - C_i(\boldsymbol{u}_i, \mathrm{X}_{-i}^k)\Big].\end{aligned}$$

Since $\mathrm{X}^\star$ is a Nash equilibrium, the middle term is nonpositive. Moreover, the feasible strategy sets are compact and the cost functions are continuously differentiable. Hence, the remaining two terms are uniformly Lipschitz in the strategy profile. Therefore, there exists a constant $c_{\mathrm{reg},i} > 0$, independent of k and $\boldsymbol{u}_i$, such that

$$C_i(\boldsymbol{x}_i^k, \mathrm{X}_{-i}^k) - C_i(\boldsymbol{u}_i, \mathrm{X}_{-i}^k) \le c_{\mathrm{reg},i}\|\mathrm{X}^k - \mathrm{X}^\star\|_{\mathrm{F}} .$$

Using the geometric convergence of the iterates,

$$C_i(\boldsymbol{x}_i^k, \mathrm{X}_{-i}^k) - C_i(\boldsymbol{u}_i, \mathrm{X}_{-i}^k) \le c_{\mathrm{reg},i}\,\rho^{(k-1)/2}\|\mathrm{X}^1 - \mathrm{X}^\star\|_{\mathrm{F}} .$$

Summing over $k = 1, \dots, K$ gives

$$\sum_{k=1}^{K}\Big(C_i(\boldsymbol{x}_i^k, \mathrm{X}_{-i}^k) - C_i(\boldsymbol{u}_i, \mathrm{X}_{-i}^k)\Big) \le \frac{c_{\mathrm{reg},i}\|\mathrm{X}^1 - \mathrm{X}^\star\|_{\mathrm{F}}}{1 - \sqrt{\rho}}.$$

The right-hand side is independent of K. Since the bound holds for every fixed $\boldsymbol{u}_i \in \mathcal{X}_i$, it also holds for the best fixed comparator in hindsight. Consequently, $\mathrm{Reg}_i(K) = O(1)$. □

E.3 Bayesian Nash Learning with Private Quotas: The Proof of Theorem 3.2

We now prove Theorem 3.2. We recall that quota vectors are drawn i.i.d. across episodes from a joint distribution $\mathcal{D}$ on the finite set $\prod_i \mathcal{B}_i$. Quotas may be correlated within an episode. We write $p_{i,b} = \mathcal{D}_i(b) > 0$ for the marginal probability that agent i has quota b. We recall that in this subsection, every episode starts from the same fixed impact state $I_1 = z$. Hence, we suppress the dependence on z in the notation.

For agent i, let $\mathcal{P}_i := \left\{\pi_i : \mathcal{B}_i \to \mathbb{R}_+^T : \boldsymbol{\pi}_i(b) \in \mathcal{X}_i(b) \text{ for every } b \in \mathcal{B}_i\right\}$ be the set of feasible policies, and let $\mathcal{P} := \prod_i \mathcal{P}_i$. Thus, a policy π_i specifies one feasible schedule $\pi_i(b) \in \mathcal{X}_i(b)$ for each possible quota b of agent i. For two policy profiles $\Pi = (\pi_1, \dots, \pi_N)$ and $\Sigma = (\sigma_1, \dots, \sigma_N)$, define

$$\langle \Pi, \Sigma \rangle_{\mathcal{D}} := \sum_{i=1}^{N} \sum_{b \in \mathcal{B}_i} p_{i,b} \left\langle \boldsymbol{\pi}_i(b), \boldsymbol{\sigma}_i(b) \right\rangle \text{ and}$$
$$\|\Pi\|_{\mathcal{D}}^2 := \langle \Pi, \Pi \rangle_{\mathcal{D}}.$$

We use the analogous notation $\langle \cdot, \cdot \rangle_{\mathcal{D}_i}$ and $\| \cdot \|_{\mathcal{D}_i}$ for a single agent.

For a realized quota vector $\boldsymbol{B}$ and policy profile Π, we let $\mathrm{X}^{\Pi}(\boldsymbol{B})$ denote the induced schedule profile and define agent i's gradient by

$$\boldsymbol{g}_i\big(\mathrm{X}^{\Pi}(\boldsymbol{B})\big) := \nabla_{\boldsymbol{x}_i} C_i\big(\boldsymbol{x}_i; \mathrm{X}_{-i}\big)\Big|_{\mathrm{X}=\mathrm{X}^{\Pi}(\boldsymbol{B})}.$$

For every agent i and quota realization $b \in \mathcal{B}_i$, we let $\mathcal{G}_{i,b} < \infty$ be an upper bound on i'a gradient such that

$$\big\|\boldsymbol{g}_i\big(\mathrm{X}^{\Pi}(\boldsymbol{B})\big)\big\|_2 \le \mathcal{G}_{i,b} \qquad \text{whenever } B_i = b,$$

uniformly over all feasible policy profiles Π and quota realizations $\boldsymbol{B}$ consistent with $B_i = b$.

For every $b \in \mathcal{B}_i$, define

$$n_{i,b}^{k-1} := \sum_{s=1}^{k-1} \mathbb{1}\{B_i^s = b\},$$

the number of times agent i has observed quota b before episode k. If $B_i^k = b$, only the policy coordinate associated with b is updated:

$$\pi_i^{k+1}(b) = \mathsf{P}_{\mathcal{X}_i(b)}\left(\pi_i^k(b) - \frac{1}{\mu_\gamma(n_{i,b}^{k-1}+1)}\boldsymbol{g}_i^k\right), \tag{49}$$

while

$$\pi_i^{k+1}(b') = \pi_i^k(b'), \qquad \text{for all} \quad b' \neq b.$$

Thus, if episode k is the m-th occurrence of quota b, the step size used by this coordinate is simply

$$\eta_m = \frac{1}{\mu_\gamma m}.$$

Realized regret is defined as

$$\mathrm{Reg}_i(K) := \sum_{k=1}^{K} C_i(\boldsymbol{x}_i^k; \mathrm{X}_{-i}^k) - \min_{\pi_i \in \mathcal{P}_i} \sum_{k=1}^{K} C_i\big(\pi_i(B_i^k); \mathrm{X}_{-i}^k\big), \tag{50}$$

The comparator is chosen after the realized sequence of quotas and opponent schedules has been observed. The agent's objective is to minimize expected regret with respect to the realized quota sequences.

We analyze the conclusions of Theorem 3.2 separately in the following two subsections. Theorem E.3 establishes the existence and uniqueness of the equilibrium, while Theorem E.4 provides the regret and equilibrium-convergence guarantees.

E.3.1 Existence and Uniqueness of Equilibrium

Theorem E.3. *Suppose that the quota vectors are drawn i.i.d. across episodes from a joint distribution $\mathcal{D}$ on the finite set $\prod_i \mathcal{B}_i$. Let $p_{i,b} := \mathcal{D}_i(b) > 0$ for every i and $b \in \mathcal{B}_i$. Then, the Bayesian game admits a unique pure Bayesian Nash equilibrium $\Pi^\star$.*

Proof. Conceptually, the proof follows the basic argument of Theorem 3.1, but now the agents choose policies rather than a single schedule. From a technical viewpoint, we have to show that the strong monotonicity of every *realized acquisition game* survives after taking conditional expectations over the quota vector (which is now a random variable).

Let $\Pi = (\pi_1, \dots, \pi_N)$ and $\Sigma = (\sigma_1, \dots, \sigma_N)$ be two feasible policy profiles. For a realized quota vector $\boldsymbol{B} = (B_1, \dots, B_N)$, they generate the schedule profiles $\mathrm{X}^\Pi(\boldsymbol{B})$ and $\mathrm{X}^\Sigma(\boldsymbol{B})$, respectively.

For agent i and type $b \in \mathcal{B}_i$, we denote the conditional Bayesian pseudo-gradient by

$$\boldsymbol{\Phi}_i(\Pi)(b) := \mathbb{E}\big[\boldsymbol{g}_i\big(\mathrm{X}^\Pi(\boldsymbol{B})\big) \,\big|\, B_i = b\big]. \tag{51}$$

Thus, $\boldsymbol{\Phi}_i(\Pi)(b)$ is the expected gradient faced by agent i when its own quota is b, averaging over the uncertainty about the other agents' quotas.

Recall that the probability-weighted inner product is

$$\langle \Pi, \Sigma \rangle_{\mathcal{D}} := \sum_i \sum_{b \in \mathcal{B}_i} p_{i,b} \, \langle \pi_i(b), \sigma_i(b) \rangle,$$

with corresponding norm $\|\cdot\|_{\mathcal{D}}$. Using (51),

$$\langle \Phi(\Pi) - \Phi(\Sigma), \Pi - \Sigma \rangle_{\mathcal{D}} = \sum_i \sum_{b \in \mathcal{B}_i} p_{i,b} \Big\langle \mathbb{E}\big[\boldsymbol{g}_i(\mathrm{X}^\Pi(\boldsymbol{B})) - \boldsymbol{g}_i(\mathrm{X}^\Sigma(\boldsymbol{B})) \,\big|\, B_i = b\big], \pi_i(b) - \sigma_i(b) \Big\rangle.$$

Since $\pi_i(b) - \sigma_i(b)$ is fixed conditional on $B_i = b$, we may move it inside the conditional expectation. Summing over b and using the law of total expectation gives

$$\begin{aligned} \langle \Phi(\Pi) - \Phi(\Sigma), \Pi - \Sigma \rangle_{\mathcal{D}} &= \sum_i \mathbb{E}\Big[\big\langle \boldsymbol{g}_i(\mathrm{X}^\Pi(\boldsymbol{B})) - \boldsymbol{g}_i(\mathrm{X}^\Sigma(\boldsymbol{B})), \pi_i(B_i) - \sigma_i(B_i) \big\rangle \Big] \\ &= \mathbb{E}\Big[\big\langle \mathrm{F}(\mathrm{X}^\Pi(\boldsymbol{B})) - \mathrm{F}(\mathrm{X}^\Sigma(\boldsymbol{B})), \mathrm{X}^\Pi(\boldsymbol{B}) - \mathrm{X}^\Sigma(\boldsymbol{B}) \big\rangle_{\mathrm{F}} \Big]. \end{aligned}$$

By Theorem 2.1, for every realization of $\boldsymbol{B}$, the execution-game pseudo-gradient F is μ_γ-strongly monotone. Applying μ_γ-strong monotonicity of F to the r.h.s. of the equalities above, we get:

$$\langle \Phi(\Pi) - \Phi(\Sigma), \Pi - \Sigma \rangle_{\mathcal{D}} \geq \mu_\gamma \mathbb{E}\big[\|\mathrm{X}^\Pi(\boldsymbol{B}) - \mathrm{X}^\Sigma(\boldsymbol{B})\|_{\mathrm{F}}^2\big].$$

Finally,

$$\begin{aligned} \mathbb{E}\big[\|\mathrm{X}^\Pi(\boldsymbol{B}) - \mathrm{X}^\Sigma(\boldsymbol{B})\|_{\mathrm{F}}^2\big] &= \sum_i \mathbb{E}\big[\|\pi_i(B_i) - \sigma_i(B_i)\|_2^2\big] \\ &= \sum_i \sum_{b \in \mathcal{B}_i} p_{i,b} \, \|\pi_i(b) - \sigma_i(b)\|_2^2 = \|\Pi - \Sigma\|_{\mathcal{D}}^2. \end{aligned}$$

Hence

$$\langle \Phi(\Pi) - \Phi(\Sigma), \Pi - \Sigma \rangle_{\mathcal{D}} \geq \mu_\gamma \|\Pi - \Sigma\|_{\mathcal{D}}^2.$$

Thus the Bayesian pseudo-gradient is μ_γ-strongly monotone with the same constant as the realized execution game.

The policy space is a finite product of compact convex sets and the Bayesian pseudo-gradient is continuous. Standard existence results for convex games therefore imply that the associated variational inequality admits a solution (Rosen, 1965); see also the variational formulation of monotone games in Mertikopoulos & Zhou (2019). Strong monotonicity of the pseudo-gradient makes this solution unique. Moreover, since for every type b agent i's conditional expected cost is convex in its own schedule, the variational-inequality conditions are equivalent to the typewise best-response conditions. Hence the unique solution is a pure Bayesian Nash equilibrium, which we denote by $\Pi^\star$. □

E.3.2 Unbiased Gradient Estimation and Pseudo-Regret Bounds

If agent i knows $\mathcal{D}_i$, it may use

$$\widehat{\boldsymbol{g}}_{i,b}^k := \frac{\mathbb{1}\{B_i^k = b\}\boldsymbol{g}_i^k}{p_{i,b}}.$$

Since $\boldsymbol{B}^k$ is independent of $\mathcal{F}_{k-1}$,

$$\mathbb{E}\Big[\widehat{\boldsymbol{g}}_{i,b}^k \mid \mathcal{F}_{k-1}\Big] = \mathbb{E}\Big[\boldsymbol{g}_i^k \mid B_i^k = b, \mathcal{F}_{k-1}\Big] = \boldsymbol{\Phi}_i(\Pi^k)(b) \qquad \text{and}$$
$$\mathbb{E}\Big[\|\widehat{\boldsymbol{g}}^k\|_{\mathcal{D}}^2 \mid \mathcal{F}_{k-1}\Big] = \sum_{i,b} \mathbb{E}\Big[\|\boldsymbol{g}_i^k\|_2^2 \mid B_i^k = b, \mathcal{F}_{k-1}\Big] \le \sum_{i,b} \mathcal{G}_{i,b}^2$$

Thus, if every agent applies Step 4 of Algorithm 1 to every policy coordinate with this oracle, the update is projected stochastic gradient descent in $(\mathcal{P}, \langle\cdot,\cdot\rangle_{\mathcal{D}})$ for the μ_γ-strongly monotone operator Φ. Taking expectations in the proof of Theorem 3.1, Part 2, with $\eta_k = 1/(\mu_\gamma(k+1))$, gives

$$\mathbb{E}\Big[\|\Pi^K - \Pi^\star\|_{\mathcal{D}}^2\Big] \le \frac{\max\{\|\Pi^1 - \Pi^\star\|_{\mathcal{D}}^2, \sum_{i,b} \mathcal{G}_{i,b}^2/\mu_\gamma^2\}}{K}.$$

Similarly, if agent i uses this oracle with $\eta_k = 1/(2\mu_\gamma k)$, the argument of Theorem 3.1, carried out in $\|\cdot\|_{\mathcal{D}_i}$, Part 1, for the expected cost

$$\sum_b p_{i,b}\mathbb{E}\Big[C_i(\pi_i(b); \mathrm{X}_{-i}^k) \mid B_i^k = b, \mathcal{F}_{k-1}\Big],$$

which is $2\mu_\gamma$-strongly convex in this norm and whose gradient in this norm is $\mathbb{E}[\widehat{\boldsymbol{g}}_{i,\cdot}^k \mid \mathcal{F}_{k-1}]$, gives expected pseudo-regret at most $\sum_b \mathcal{G}_{i,b}^2(1+\log K)/(4\mu_\gamma)$ against every fixed own-quota policy. In the proof of Theorem E.4 below, this oracle is used only as an analytical device.

E.3.3 Online Learning with Local Estimators: Realized Regret and Convergence to BNE

Theorem E.4. *Suppose that the quota vectors are drawn i.i.d. across episodes from a joint distribution $\mathcal{D}$ on the finite set $\prod_i \mathcal{B}_i$. Let $p_{i,b} := \mathcal{D}_i(b) > 0$ for every i and $b \in \mathcal{B}_i$. Then:*

1. *If agent i employs the empirical update rule* (49)*, then its expected regret satisfies*

$$\mathbb{E}[\mathrm{Reg}_i(K)] \le \sum_{b \in \mathcal{B}_i} \frac{\mathcal{G}_{i,b}^2}{2\mu_\gamma}\,(1 + \log(1 + Kp_{i,b})) = O(\log K).$$

2. *If all agents employ the empirical update rule* (49)*, then their policy profile converges to the unique pure Bayesian Nash equilibrium $\Pi^\star$. More precisely, there exists a finite constant $c_{\mathrm{BNE}} > 0$ such that for any $K \ge 2$,*

$$\mathbb{E}\big[\|\Pi^K - \Pi^\star\|_{\mathcal{D}}^2\big] \le \frac{c_{\mathrm{BNE}} \max\Big\{\|\Pi^1 - \Pi^\star\|_{\mathcal{D}}^2,\ \sum_{i,b} \mathcal{G}_{i,b}^2/\mu_\gamma^2\Big\}}{K},$$

where $\mathcal{G}_{i,b}$ is the gradient bound (6) *with $B_i = b$ and V_i replaced by the largest competing quota.*

The following lemma provides the moment bounds needed to control the discrepancy between the random local estimator and the corresponding deterministic global estimator.

Lemma E.5. *Let $n \sim \mathrm{Binomial}(k-1, p)$ and let $r := \frac{pk}{n+1}$. Then $\mathbb{E}[r^2] \le 2$ and $\mathbb{E}[(r-1)^2] \le \frac{c_{\mathrm{clk}}}{pk}$, for a universal constant $c_{\mathrm{clk}} > 0$.*

Proof of Lemma E.5. For the second moment, observe first that

$$\frac{1}{(n+1)^2} \le \frac{2}{(n+1)(n+2)}.$$

For $n \sim \mathrm{Binomial}(k-1, p)$, the binomial generating function gives

$$\mathbb{E}\left[\frac{1}{(n+1)(n+2)}\right] = \frac{1-(1-p)^{k+1}-(k+1)p(1-p)^k}{k(k+1)p^2} \leq \frac{1}{k(k+1)p^2}.$$

Therefore

$$\mathbb{E}[r^2] = p^2k^2\mathbb{E}\left[\frac{1}{(n+1)^2}\right] \leq \frac{2k}{k+1} \leq 2.$$

We next control the deviation of r from one. Put $m := pk$.

If $m < 1$, then $0 \leq r \leq m < 1$, so

$$(r-1)^2 \leq 1 \leq \frac{1}{m}.$$

It therefore remains to consider $m \geq 1$. Split according to the event $A := \{n+1 \geq m/2\}$.

On A,

$$(r-1)^2 = \frac{(m-(n+1))^2}{(n+1)^2} \leq \frac{4(m-(n+1))^2}{m^2}.$$

Now

$$\mathbb{E}[n+1] = (k-1)p + 1 = m + (1-p),$$

and hence

$$\mathbb{E}\big[(m-(n+1))^2\big] = \mathrm{Var}(n) + \big(m - \mathbb{E}[n+1]\big)^2 \leq m+1.$$

Thus the contribution on A is $O(1/m)$.

On the complementary event A^c, the binomial count is substantially smaller than its mean. A standard lower-tail Chernoff bound gives $\Pr(A^c) \leq e^{-c_{\mathrm{tail}}m}$ for a universal $c_{\mathrm{tail}} > 0$, after adjusting the constant for the finitely many small values of m. Since $(r-1)^2 \leq (m+1)^2$, the contribution of A^c is also $O(1/m)$. Combining the two parts gives

$$\mathbb{E}[(r-1)^2] \leq \frac{c_{\mathrm{clk}}}{m} = \frac{c_{\mathrm{clk}}}{pk}$$

for a universal constant c_{clk}. □

Proof of Theorem E.4. We prove the statements separately.

Part 1. The proof exploits the fact that the policy space and the hindsight comparator decompose across quota types. For each type, the local estimator update reduces to standard projected gradient descent on the subsequence where that type occurs, allowing the strongly convex regret bound to be applied separately and then summed across types.

Fix an agent i. The first observation is that the hindsight comparator in (50) separates across the agent's quota types. Indeed, a feasible policy π_i chooses one schedule $\pi_i(b) \in \mathcal{X}_i(b)$ independently for each $b \in \mathcal{B}_i$. Therefore

$$\min_{\pi_i \in \mathcal{P}_i} \sum_{k=1}^{K} C_i\big(\pi_i(B_i^k); \mathrm{X}_{-i}^k\big) = \sum_{b \in \mathcal{B}_i} \min_{\boldsymbol{u} \in \mathcal{X}_i(b)} \sum_{\substack{1 \leq k \leq K \\ B_i^k = b}} C_i(\boldsymbol{u}; \mathrm{X}_{-i}^k). \tag{52}$$

Hence it is enough to analyze separately the subsequence on which a fixed quota b occurs.

Fix such a type b, and let $\tau_1 < \tau_2 < \cdots < \tau_M$ be the episodes up to time K in which $B_i^k = b$, where $M = n_{i,b}^K$. For the m-th occurrence, write

$$\boldsymbol{x}_m := \pi_i^{\tau_m}(b), \qquad \boldsymbol{g}_m := \boldsymbol{g}_i^{\tau_m}.$$

Because no other episode changes this policy coordinate, the local estimator update restricted to this subsequence is exactly

$$\boldsymbol{x}_{m+1} = \mathsf{P}_{\mathcal{X}_i(b)}\left(\boldsymbol{x}_m - \eta_m \boldsymbol{g}_m\right), \qquad \eta_m = \frac{1}{\mu_\gamma m}.$$

Thus each type runs an ordinary projected-gradient algorithm on the episodes in which that type is actually observed.

Fix any comparator $\boldsymbol{u} \in \mathcal{X}_i(b)$. In episode τ_m, the realized cost

$$C_i(\,\cdot\,; \mathrm{X}_{-i}^{\tau_m})$$

is $2\mu_\gamma$-strongly convex in agent i's own schedule. Therefore

$$C_i(\boldsymbol{x}_m; \mathrm{X}_{-i}^{\tau_m}) - C_i(\boldsymbol{u}; \mathrm{X}_{-i}^{\tau_m}) \le \langle \boldsymbol{g}_m, \boldsymbol{x}_m - \boldsymbol{u}\rangle - \mu_\gamma \|\boldsymbol{x}_m - \boldsymbol{u}\|_2^2. \tag{53}$$

On the other hand, non-expansiveness of Euclidean projection on the $(m+1)$-th iterate gives

$$\begin{aligned}\|\boldsymbol{x}_{m+1} - \boldsymbol{u}\|_2^2 &\le \|\boldsymbol{x}_m - \eta_m \boldsymbol{g}_m - \boldsymbol{u}\|_2^2 \\ &= \|\boldsymbol{x}_m - \boldsymbol{u}\|_2^2 - 2\eta_m \langle \boldsymbol{g}_m, \boldsymbol{x}_m - \boldsymbol{u}\rangle + \eta_m^2 \|\boldsymbol{g}_m\|_2^2.\end{aligned}$$

Rearranging,

$$\langle \boldsymbol{g}_m, \boldsymbol{x}_m - \boldsymbol{u}\rangle \le \frac{\|\boldsymbol{x}_m - \boldsymbol{u}\|_2^2 - \|\boldsymbol{x}_{m+1} - \boldsymbol{u}\|_2^2}{2\eta_m} + \frac{\eta_m}{2}\|\boldsymbol{g}_m\|_2^2.$$

Combining the above with (53), we obtain

$$\begin{aligned}&C_i(\boldsymbol{x}_m; \mathrm{X}_{-i}^{\tau_m}) - C_i(\boldsymbol{u}; \mathrm{X}_{-i}^{\tau_m}) \\ &\le \frac{\|\boldsymbol{x}_m - \boldsymbol{u}\|_2^2 - \|\boldsymbol{x}_{m+1} - \boldsymbol{u}\|_2^2}{2\eta_m} - \mu_\gamma \|\boldsymbol{x}_m - \boldsymbol{u}\|_2^2 + \frac{\eta_m}{2}\|\boldsymbol{g}_m\|_2^2.\end{aligned}$$

We now sum over $m = 1, \dots, M$. With step size $\eta_m = \frac{1}{\mu_\gamma m}$, the coefficient of every intermediate squared-distance term is nonpositive. More explicitly, for $m \ge 2$,

$$\frac{1}{2\eta_m} - \mu_\gamma - \frac{1}{2\eta_{m-1}} = -\frac{\mu_\gamma}{2},$$

and the initial coefficient is also negative. Dropping these nonpositive terms gives

$$\sum_{m=1}^{M} \left[C_i(\boldsymbol{x}_m; \mathrm{X}_{-i}^{\tau_m}) - C_i(\boldsymbol{u}; \mathrm{X}_{-i}^{\tau_m})\right] leq \frac{1}{2\mu_\gamma} \sum_{m=1}^{M} \frac{\|\boldsymbol{g}_m\|_2^2}{m}.$$

The right-hand side does not depend on $\boldsymbol{u}$, so the above also holds for the best type-b comparator chosen in hindsight. Denoting the resulting type-b regret by $\mathrm{Reg}_{i,b}(K)$,

$$\mathrm{Reg}_{i,b}(K) \le \frac{1}{2\mu_\gamma} \sum_{m=1}^{n_{i,b}^K} \frac{\|\boldsymbol{g}_i^{\tau_m}\|_2^2}{m}. \tag{54}$$

The assumed conditional second-moment bound is uniform over episodes and histories. Hence, at every occurrence of type b,

$$\mathbb{E}\big[\|\boldsymbol{g}_i^{\tau_m}\|_2^2 \,\big|\, \tau_m \le K\big] \le \mathcal{G}_{i,b}^2.$$

Taking expectations in (54) therefore gives

$$\mathbb{E}\big[\,\mathrm{Reg}_{i,b}(K)\big] \le \frac{\mathcal{G}_{i,b}^2}{2\mu_\gamma} \mathbb{E}\Big[H_{n_{i,b}^K}\Big],$$

where $H_n := \sum_{m=1}^{n} \frac{1}{m}$ and $H_0 := 0$.

We can bound $H_n \le 1 + \log(1+n)$ and using the concavity of $\log$ we obtain

$$\begin{aligned}\mathbb{E}[H_{n_{i,b}^K}] &\le 1 + \mathbb{E}\big[\log(1 + n_{i,b}^K)\big] \\ &\le 1 + \log\big(1 + \mathbb{E}[n_{i,b}^K]\big) = 1 + \log(1 + K p_{i,b}),\end{aligned}$$

where $\mathbb{E}[n_{i,b}^K] = Kp_{i,b}$ since the quota vectors are i.i.d. across episodes. Finally, summing the typewise regrets in (52) gives

$$\mathbb{E}[\mathrm{Reg}_i(K)] \le \sum_{b\in\mathcal{B}_i} \frac{\mathcal{G}_{i,b}^2}{2\mu_\gamma}\left(1+\log(1+Kp_{i,b})\right).$$

This proves part 1.

Part 2. The main difficulty is that each policy coordinate is updated only when its corresponding quota type occurs and therefore follows its own random update counter. We show that these random counters concentrate around their expected values $p_{i,b}k$; together with the unbiasedness of the observed gradients and the strong monotonicity of the Bayesian pseudo-gradient, this yields contraction toward the equilibrium at rate $O(1/k)$.

Suppose that all agents employ the local estimator update. The Bayesian pseudo-gradient is defined coordinatewise by

$$\mathbf{\Phi}_i(\Pi)(b) := \mathbb{E}\big[\boldsymbol{g}_i\big(\mathrm{X}^\Pi(B)\big) \,\big|\, B_i = b\big].$$

By Theorem E.3, it is μ_γ-strongly monotone:

$$\langle \Phi(\Pi) - \Phi(\Pi'), \Pi - \Pi'\rangle_{\mathcal{D}} \ge \mu_\gamma \|\Pi - \Pi'\|_{\mathcal{D}}^2.$$

Let $\Pi^\star$ denote the unique Bayesian Nash equilibrium. It satisfies the variational inequality

$$\langle \Phi(\Pi^\star), \Pi - \Pi^\star\rangle_{\mathcal{D}} \ge 0$$

for every feasible policy profile Π.

For every agent–type pair (i,b), define

$$I_{i,b}^k := \mathbb{1}\{B_i^k = b\} \qquad \text{and} \qquad n_{i,b}^{k-1} := \sum_{s=1}^{k-1} \mathbb{1}\{B_i^s = b\}.$$

Thus, $n_{i,b}^{k-1}$ counts the occurrences of type b before episode k. Since quota vectors are drawn independently across episodes, $n_{i,b}^{k-1} \sim \mathrm{Binomial}(k-1, p_{i,b})$. Define the random estimator multiplier

$$r_{i,b}^k := \frac{p_{i,b}k}{n_{i,b}^{k-1}+1}.$$

This multiplier depends only on quota realizations before episode k and is therefore $\mathcal{F}_{k-1}$-measurable. By Lemma E.5,

$$\mathbb{E}[(r_{i,b}^k)^2] \le 2 \qquad \text{and} \qquad \mathbb{E}[(r_{i,b}^k - 1)^2] \le \frac{c_{\mathrm{clk}}}{p_{i,b}k}.$$

For the analysis, define the importance-weighted gradient coordinate

$$\widehat{\boldsymbol{g}}_{i,b}^k := \frac{I_{i,b}^k}{p_{i,b}} \boldsymbol{g}_i^k.$$

We first establish its conditional unbiasedness. Since the current quota vector is independent of $\mathcal{F}_{k-1}$,

$$\begin{aligned}\mathbb{E}\big[\widehat{\boldsymbol{g}}_{i,b}^k \,\big|\, \mathcal{F}_{k-1}\big] &= \frac{1}{p_{i,b}} \mathbb{E}\big[I_{i,b}^k \boldsymbol{g}_i^k \,\big|\, \mathcal{F}_{k-1}\big] \\ &= \mathbb{E}\big[\boldsymbol{g}_i^k \,\big|\, B_i^k = b, \mathcal{F}_{k-1}\big] \\ &= \mathbf{\Phi}_i(\Pi^k)(b).\end{aligned}$$

The last equality follows because Π^k is determined by the history before episode k.

Set

$$\eta_k := \frac{1}{\mu_\gamma k}.$$

Then

$$\eta_k r_{i,b}^k \widehat{\boldsymbol{g}}_{i,b}^k = \frac{1}{\mu_\gamma k} \frac{p_{i,b}k}{n_{i,b}^{k-1}+1} \frac{I_{i,b}^k}{p_{i,b}} \boldsymbol{g}_i^k = \frac{I_{i,b}^k}{\mu_\gamma (n_{i,b}^{k-1}+1)} \boldsymbol{g}_i^k.$$

This is exactly the step used by the local estimator update: the coordinate (i, b) is updated when $B_i^k = b$ and remains unchanged otherwise.

Let R_k be the block-diagonal operator whose (i, b) block is $r_{i,b}^k \mathrm{I}_T$, and let $\widehat{\boldsymbol{g}}^{\,k}$ collect all coordinates $\widehat{\boldsymbol{g}}_{i,b}^k$. The complete update can therefore be written as

$$\Pi^{k+1} = \mathsf{P}_{\mathcal{P}} \left(\Pi^k - \eta_k \mathrm{R}_k \widehat{\boldsymbol{g}}^{\,k} \right).$$

Because $\mathcal{P}$ is a Cartesian product and each weight $p_{i,b}$ is a positive constant, coordinatewise Euclidean projection is nonexpansive in the norm $\|\cdot\|_{\mathcal{D}}$.

Define

$$E^k := \Pi^k - \Pi^\star \qquad \text{and} \qquad e_k := \mathbb{E}\big[\|E^k\|_{\mathcal{D}}^2\big].$$

Projection nonexpansiveness gives

$$\|E^{k+1}\|_{\mathcal{D}}^2 \le \|E^k\|_{\mathcal{D}}^2 - 2\eta_k \left\langle \mathrm{R}_k \widehat{\boldsymbol{g}}^{\,k}, E^k \right\rangle_{\mathcal{D}} + \eta_k^2 \|\mathrm{R}_k \widehat{\boldsymbol{g}}^{\,k}\|_{\mathcal{D}}^2.$$

We bound the drift and squared-gradient terms separately.

Since R_k and E^k are $\mathcal{F}_{k-1}$-measurable, conditional unbiasedness gives

$$\mathbb{E}\big[\left\langle \mathrm{R}_k \widehat{\boldsymbol{g}}^{\,k}, E^k \right\rangle_{\mathcal{D}} \,\big|\, \mathcal{F}_{k-1}\big] = \left\langle \mathrm{R}_k \Phi(\Pi^k), E^k \right\rangle_{\mathcal{D}}.$$

Decompose the right-hand side as

$$\left\langle \mathrm{R}_k \Phi(\Pi^k), E^k \right\rangle_{\mathcal{D}} = \left\langle \Phi(\Pi^k), E^k \right\rangle_{\mathcal{D}} + \left\langle (\mathrm{R}_k - \mathrm{I}) \Phi(\Pi^k), E^k \right\rangle_{\mathcal{D}}.$$

For the first term, strong monotonicity and the equilibrium variational inequality imply

$$\begin{aligned} \left\langle \Phi(\Pi^k), E^k \right\rangle_{\mathcal{D}} &= \left\langle \Phi(\Pi^k) - \Phi(\Pi^\star), E^k \right\rangle_{\mathcal{D}} + \left\langle \Phi(\Pi^\star), E^k \right\rangle_{\mathcal{D}} \\ &\ge \mu_\gamma \|E^k\|_{\mathcal{D}}^2. \end{aligned}$$

The second term is the error caused by the random local estimators. By Cauchy–Schwarz and Young's inequality,

$$\left\langle (\mathrm{R}_k - \mathrm{I}) \Phi(\Pi^k), E^k \right\rangle_{\mathcal{D}} \ge -\frac{\mu_\gamma}{4} \|E^k\|_{\mathcal{D}}^2 - \frac{1}{\mu_\gamma} \|(\mathrm{R}_k - \mathrm{I}) \Phi(\Pi^k)\|_{\mathcal{D}}^2.$$

Combining these inequalities gives

$$\left\langle \mathrm{R}_k \Phi(\Pi^k), E^k \right\rangle_{\mathcal{D}} \ge \frac{3\mu_\gamma}{4} \|E^k\|_{\mathcal{D}}^2 - \frac{1}{\mu_\gamma} \|(\mathrm{R}_k - \mathrm{I}) \Phi(\Pi^k)\|_{\mathcal{D}}^2.$$

It remains to control the last term. By Jensen's inequality and the assumed conditional second-moment bound,

$$\begin{aligned} \|\boldsymbol{\Phi}_i(\Pi^k)(b)\|_2^2 &= \big\| \mathbb{E}[\boldsymbol{g}_i^k \mid B_i^k = b, \mathcal{F}_{k-1}] \big\|_2^2 \\ &\le \mathbb{E}\big[\|\boldsymbol{g}_i^k\|_2^2 \,\big|\, B_i^k = b, \mathcal{F}_{k-1}\big] \le \mathcal{G}_{i,b}^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \|(\mathrm{R}_k - \mathrm{I}) \Phi(\Pi^k)\|_{\mathcal{D}}^2 &= \sum_i \sum_{b \in \mathcal{B}_i} p_{i,b} (r_{i,b}^k - 1)^2 \|\boldsymbol{\Phi}_i(\Pi^k)(b)\|_2^2 \\ &\le \sum_i \sum_{b \in \mathcal{B}_i} p_{i,b} (r_{i,b}^k - 1)^2 \mathcal{G}_{i,b}^2. \end{aligned}$$

Taking expectations and using the estimator-moment bound gives

$$\mathbb{E}\big[\|(\mathrm{R}_k - \mathrm{I}) \Phi(\Pi^k)\|_{\mathcal{D}}^2\big] \le \sum_i \sum_{b \in \mathcal{B}_i} p_{i,b} \frac{c_{\mathrm{clk}}}{p_{i,b} k} \mathcal{G}_{i,b}^2 = \frac{c_{\mathrm{clk}}}{k} \sum_i \sum_{b \in \mathcal{B}_i} \mathcal{G}_{i,b}^2.$$

The factor $p_{i,b}$ from the probability-weighted norm cancels the factor $p_{i,b}^{-1}$ in the estimator-multiplier bound.

By the definition of the probability-weighted norm,

$$\|\mathrm{R}_k \widehat{\boldsymbol{g}}^{\,k}\|_{\mathcal{D}}^2 = \sum_i \sum_{b \in \mathcal{B}_i} p_{i,b} (r_{i,b}^k)^2 \frac{I_{i,b}^k}{p_{i,b}^2} \|\boldsymbol{g}_i^k\|_2^2.$$

Conditioning on $\mathcal{F}_{k-1}$ and using the independence of the current quota vector from the past,

$$\mathbb{E}\big[\|\mathrm{R}_k\widehat{\boldsymbol{g}}^k\|_{\mathcal{D}}^2 \,\big|\, \mathcal{F}_{k-1}\big] = \sum_i \sum_{b\in\mathcal{B}_i} (r_{i,b}^k)^2 \mathbb{E}\big[\|\boldsymbol{g}_i^k\|_2^2 \,\big|\, B_i^k = b, \mathcal{F}_{k-1}\big] \\ \le \sum_i \sum_{b\in\mathcal{B}_i} (r_{i,b}^k)^2 \mathcal{G}_{i,b}^2.$$

Taking expectations again and using $\mathbb{E}[(r_{i,b}^k)^2] \le 2$ gives

$$\mathbb{E}\|\mathrm{R}_k\widehat{\boldsymbol{g}}^k\|_{\mathcal{D}}^2 \le 2\sum_i \sum_{b\in\mathcal{B}_i} \mathcal{G}_{i,b}^2.$$

For brevity, define

$$S_{\mathcal{D}}^2 := \sum_i \sum_{b\in\mathcal{B}_i} \mathcal{G}_{i,b}^2.$$

Taking expectations in the one-step distance inequality and applying the preceding drift and second-moment bounds gives

$$e_{k+1} \le e_k - \frac{3\mu_\gamma\eta_k}{2} e_k + \frac{2c_{\mathrm{clk}}\eta_k}{\mu_\gamma k} S_{\mathcal{D}}^2 + 2\eta_k^2 S_{\mathcal{D}}^2.$$

Since $\eta_k = 1/(\mu_\gamma k)$,

$$e_{k+1} \le \left(1 - \frac{3}{2k}\right) e_k + \frac{2(c_{\mathrm{clk}}+1)S_{\mathcal{D}}^2}{\mu_\gamma^2 k^2}.$$

Set

$$c_{\mathrm{rec}} := \frac{2(c_{\mathrm{clk}}+1)S_{\mathcal{D}}^2}{\mu_\gamma^2}.$$

We prove by induction that $e_k \le c_{\mathrm{ind}}/k$ for every $k \ge 2$, where

$$c_{\mathrm{ind}} \ge \max\{2e_2, 2c_{\mathrm{rec}}\}.$$

The claim holds at $k = 2$. If it holds at some $k \ge 2$, then

$$e_{k+1} \le \left(1 - \frac{3}{2k}\right)\frac{c_{\mathrm{ind}}}{k} + \frac{c_{\mathrm{rec}}}{k^2} \\ = \frac{c_{\mathrm{ind}}}{k} - \frac{3c_{\mathrm{ind}}/2 - c_{\mathrm{rec}}}{k^2}.$$

Since $c_{\mathrm{ind}} \ge 2c_{\mathrm{rec}}$, we have $3c_{\mathrm{ind}}/2 - c_{\mathrm{rec}} \ge c_{\mathrm{ind}}$, and hence

$$e_{k+1} \le \frac{c_{\mathrm{ind}}}{k} - \frac{c_{\mathrm{ind}}}{k^2} \le \frac{c_{\mathrm{ind}}}{k+1}.$$

This completes the induction. Therefore,

$$\mathbb{E}\big[\|\Pi^K - \Pi^\star\|_{\mathcal{D}}^2\big] = e_K \le \frac{c_{\mathrm{ind}}}{K}, \qquad \text{for all } K \ge 2.$$

□

E.4 Inherited State and the Equilibrium State Map

Throughout this subsection, assume $0 < \gamma < 1$. For a fixed quota vector, let $\mathrm{X}^\star(z)$ denote the unique Nash equilibrium of the one-episode acquisition game when the entering state is $z \ge 0$, and let

$$\boldsymbol{q}^\star(z) := \sum_{i=1}^N \boldsymbol{x}_i^\star(z)$$

be the corresponding aggregate schedule.

The state left at the end of the episode is

$$I_{T+1} = \gamma^T z + \lambda \sum_{t=1}^{T} \gamma^{T-t} q_t.$$

It is therefore convenient to define

$$\boldsymbol{w}_\gamma := (\gamma^{T-1}, \dots, \gamma, 1)^\top, \qquad s_T(\gamma) := \|\boldsymbol{w}_\gamma\|_2^2 = \|\boldsymbol{d}_\gamma\|_2^2 = \sum_{r=0}^{T-1} \gamma^{2r},$$

and the *equilibrium state map*

$$\Psi(z) := \gamma^T z + \lambda \boldsymbol{w}_\gamma^\top \boldsymbol{q}^\star(z).$$

Thus $\Psi(z)$ is the state entering the next episode when the current episode starts from z and the agents play the one-episode Nash equilibrium.

Definition E.6 (Self-consistent myopic equilibrium)**.** A pair $(\mathrm{X}^\star, z^\star)$ is a *self-consistent myopic equilibrium* if

$$\mathrm{X}^\star = \mathrm{X}^\star(z^\star)$$

and the state generated at the end of the episode coincides with the entering state:

$$z^\star = \Psi(z^\star) = \gamma^T z^\star + \lambda \boldsymbol{w}_\gamma^\top \boldsymbol{q}^\star(z^\star).$$

The term "myopic" is important: the agents optimize only their cost in the current episode. The fixed-point condition makes the entering state consistent across episodes, but it does not introduce a continuation value into the agents' optimization problem.

Lemma E.7 (Existence and Uniqueness of a Self-Consistent State)**.** *A self-consistent myopic equilibrium exists. If*

$$\widehat{\chi}_T := \gamma^T + \frac{\lambda N}{N+1} \frac{s_T(\gamma)}{\mu_\gamma} < 1,$$

then the self-consistent myopic equilibrium is unique.

Proof. By Theorem 2.1, for every $z \geq 0$ the one-episode game has a unique Nash equilibrium $\mathrm{X}^\star(z)$.

We first show that the equilibrium aggregate depends continuously, in fact Lipschitz continuously, on the entering state. Let

$$\mathrm{X} := \mathrm{X}^\star(z), \qquad \mathrm{X}' := \mathrm{X}^\star(z'),$$

and define

$$\Delta\mathrm{X} := \mathrm{X} - \mathrm{X}', \qquad \Delta\boldsymbol{q} := \boldsymbol{q}^\star(z) - \boldsymbol{q}^\star(z').$$

The two equilibrium variational inequalities are

$$\langle \mathrm{F}(\mathrm{X}; z), \mathrm{X}' - \mathrm{X} \rangle_{\mathrm{F}} \geq 0,$$

and

$$\langle \mathrm{F}(\mathrm{X}'; z'), \mathrm{X} - \mathrm{X}' \rangle_{\mathrm{F}} \geq 0.$$

Adding them and then adding and subtracting $\mathrm{F}(\mathrm{X}'; z)$ gives

$$\langle \mathrm{F}(\mathrm{X}; z) - \mathrm{F}(\mathrm{X}'; z), \Delta\mathrm{X} \rangle_{\mathrm{F}} \leq - \langle \mathrm{F}(\mathrm{X}'; z) - \mathrm{F}(\mathrm{X}'; z'), \Delta\mathrm{X} \rangle_{\mathrm{F}}. \tag{55}$$

The entering state enters every agent's gradient only through the term $z\boldsymbol{d}_\gamma$. Hence the ith row of $\mathrm{F}(\mathrm{X}'; z) - \mathrm{F}(\mathrm{X}'; z')$ is $(z - z')\boldsymbol{d}_\gamma^\top$, and therefore

$$|\langle \mathrm{F}(\mathrm{X}'; z) - \mathrm{F}(\mathrm{X}'; z'), \Delta\mathrm{X} \rangle_{\mathrm{F}}| = |z - z'| \left| \sum_{i=1}^{N} \langle \boldsymbol{d}_\gamma, \Delta\boldsymbol{x}_i \rangle \right| = |z - z'| \, |\langle \boldsymbol{d}_\gamma, \Delta\boldsymbol{q} \rangle|.$$

Applying the strong-monotonicity bound of Theorem 2.1 to the left-hand side of (55) gives

$$\mu_\gamma \left(\|\Delta\mathrm{X}\|_{\mathrm{F}}^2 + \|\Delta\boldsymbol{q}\|_2^2 \right) \leq |z - z'| \, |\langle \boldsymbol{d}_\gamma, \Delta\boldsymbol{q} \rangle|. \tag{56}$$

Now

$$\|\Delta \boldsymbol{q}\|_2^2 = \left\| \sum_{i=1}^{N} \Delta \boldsymbol{x}_i \right\|_2^2 \le N \sum_{i=1}^{N} \|\Delta \boldsymbol{x}_i\|_2^2 = N \|\Delta \mathrm{X}\|_{\mathrm{F}}^2,$$

and so

$$\|\Delta \mathrm{X}\|_{\mathrm{F}}^2 \ge \frac{1}{N} \|\Delta \boldsymbol{q}\|_2^2.$$

Combining this with (56) and applying Cauchy–Schwarz,

$$\mu_\gamma \left(1 + \frac{1}{N}\right) \|\Delta \boldsymbol{q}\|_2^2 \le |z - z'| \|\boldsymbol{d}_\gamma\|_2 \|\Delta \boldsymbol{q}\|_2.$$

If $\Delta \boldsymbol{q} \neq \mathbf{0}$, cancelling one factor $\|\Delta \boldsymbol{q}\|_2$ yields

$$\|\boldsymbol{q}^\star(z) - \boldsymbol{q}^\star(z')\|_2 \le \frac{N}{N+1} \frac{\|\boldsymbol{d}_\gamma\|_2}{\mu_\gamma} |z - z'|. \tag{57}$$

The inequality is trivial when $\Delta \boldsymbol{q} = \mathbf{0}$. Thus $\boldsymbol{q}^\star(z)$ is Lipschitz continuous in z.

We can now prove existence of a fixed point of Ψ. Let $V := \sum_{i=1}^{N} B_i$ be the total quota. Since every feasible aggregate schedule is nonnegative and has total mass V, $0 \le \boldsymbol{w}_\gamma^\top \boldsymbol{q}^\star(z) \le V$. Hence

$$0 \le \Psi(z) \le \gamma^T z + \lambda V.$$

Define

$$\bar{z} := \frac{\lambda V}{1 - \gamma^T}.$$

For every $z \in [0, \bar{z}]$,

$$\Psi(z) \le \gamma^T \bar{z} + \lambda V = \bar{z}.$$

Thus Ψ maps the compact interval $[0, \bar{z}]$ into itself. Since Ψ is continuous by (57), it has a fixed point.

It remains to study uniqueness. For any $z, z' \ge 0$,

$$\begin{aligned} |\Psi(z) - \Psi(z')| &\le \gamma^T |z - z'| + \lambda \|\boldsymbol{w}_\gamma\|_2 \|\boldsymbol{q}^\star(z) - \boldsymbol{q}^\star(z')\|_2 \\ &\le \left[\gamma^T + \frac{\lambda N}{N+1} \frac{\|\boldsymbol{w}_\gamma\|_2 \|\boldsymbol{d}_\gamma\|_2}{\mu_\gamma} \right] |z - z'| \\ &= \widehat{\chi}_T |z - z'|, \end{aligned}$$

because $\|\boldsymbol{w}_\gamma\|_2 = \|\boldsymbol{d}_\gamma\|_2 = \sqrt{s_T(\gamma)}$.

Therefore, if $\widehat{\chi}_T < 1$, Ψ is a contraction and its fixed point is unique. Since the one-episode Nash equilibrium is unique at every entering state, the pair $(\mathrm{X}^\star, z^\star)$ is then unique as well.

A simpler, slightly stronger sufficient condition is obtained by dropping the factor $N/(N+1) < 1$. Since

$$s_T(\gamma) = \frac{1 - \gamma^{2T}}{1 - \gamma^2},$$

the condition

$$\gamma^T + \frac{\lambda s_T(\gamma)}{\mu_\gamma} < 1$$

is equivalent to

$$\mu_\gamma > \lambda \frac{1 + \gamma^T}{1 - \gamma^2}.$$

Using $\mu_\gamma = \kappa - \lambda/(1+\gamma)$, this becomes

$$\kappa > \lambda \frac{2 - \gamma + \gamma^T}{1 - \gamma^2}. \tag{58}$$

Hence (58) implies $\widehat{\chi}_T < 1$. □

Remark E.8. Although the proof of Lemma E.7 holds without knowledge of the exact cost function, leveraging the exact formula lets us derive an alternative contraction factor. Moreover, this analysis proves a stronger result: the map is a contraction under the stability condition (A) alone.

Define

$$c := \min\left\{\frac{\kappa + \mu_\gamma/N}{\lambda}, \frac{1}{2\gamma}\right\}, \qquad \rho := \sqrt{\frac{1 + 2\gamma - 2\gamma(1+\gamma)c}{2(1+\gamma)c - 1}},$$

with $c := 1/(2\gamma)$ when $\lambda = 0$. Under (A), $c(1+\gamma) > 1$ and $c \le 1/(2\gamma)$, so $\rho < 1$.

Indeed, consider the equilibrium paths starting from z and z', and let

$$J_t := I_t - I'_t, \qquad u_t := J_{t+1} - \gamma J_t = \lambda \Delta q_t.$$

Adding the corresponding equilibrium variational inequalities gives

$$\sum_{t=1}^{T} u_t(J_t + cu_t) \le 0.$$

A direct expansion and $J_t J_{t+1} \ge -(J_t^2 + J_{t+1}^2)/2$ show that

$$u_t(J_t + cu_t) \ge \beta(J_{t+1}^2 - J_t^2) + m(J_{t+1}^2 + J_t^2),$$

where

$$\beta := \frac{c(1-\gamma^2) + \gamma}{2}, \qquad m := \frac{(1+\gamma)(c(1+\gamma) - 1)}{2} > 0.$$

Summing over t and dropping the nonnegative intermediate terms yields

$$(\beta + m) J_{T+1}^2 \le (\beta - m) J_1^2.$$

Since $J_1 = z - z'$, $J_{T+1} = \Psi(z) - \Psi(z')$, and $(\beta - m)/(\beta + m) = \rho^2$, we obtain

$$|\Psi(z) - \Psi(z')| \le \rho |z - z'|.$$

Thus, the self-consistent state is unique and globally geometrically attracting with parameter $\chi_T := \min\{\rho, \widehat{\chi}_T\} < 1$. Neither ρ nor $\widehat{\chi}_T$ uniformly dominates the other.

E.4.1 Random Quotas and the Stationary Inherited-State Law

We now allow the quota vector to vary across episodes. For each entering state $z \ge 0$, let

$$\Pi^\star(z) = \big(\pi_1^\star(z), \dots, \pi_N^\star(z)\big)$$

denote the unique pure Bayesian Nash equilibrium of the one-episode game. After a quota vector $\boldsymbol{B} = (B_1, \dots, B_N)$ is drawn, agent i plays $\pi_i^\star(z)(B_i)$, and the resulting aggregate schedule is

$$\boldsymbol{q}_z^\star(\boldsymbol{B}) := \sum_{i=1}^{N} \pi_i^\star(z)(B_i).$$

The inherited state generated by this realization is therefore

$$\Psi(z; \boldsymbol{B}) := \gamma^T z + \lambda \boldsymbol{w}_\gamma^\top \boldsymbol{q}_z^\star(\boldsymbol{B}). \tag{59}$$

Before the quota vector is observed, the expected state entering the next episode is described by the conditional-mean map

$$\overline{\Psi}(z) := \mathbb{E}_{\boldsymbol{B}\sim\mathcal{D}}[\Psi(z; \boldsymbol{B})] = \gamma^T z + \lambda \boldsymbol{w}_\gamma^\top \mathbb{E}_{\boldsymbol{B}\sim\mathcal{D}}[\boldsymbol{q}_z^\star(\boldsymbol{B})].$$

Thus, iterating $\overline{\Psi}$ produces a deterministic conditional-mean dynamics, whereas using the realized value of (59) produces a Markov chain.

Let $\mathcal{P}_1(\mathbb{R}_+)$ be the set of probability measures on $\mathbb{R}_+$ with finite first moment. For $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}_+)$, their 1-Wasserstein distance is

$$W_1(\nu, \nu') := \inf_{(Z, Z')} \mathbb{E}|Z - Z'|,$$

where the infimum is taken over all couplings satisfying $Z \sim \nu$ and $Z' \sim \nu'$.

Proposition E.9. *Suppose that $(\boldsymbol{B}^k)_{k\geq 1}$ is drawn i.i.d. from $\mathcal{D}$, whose support is finite, and let*

$$V(\boldsymbol{B}) := \sum_{i=1}^{N} B_i \leq V_{\max} \qquad \textit{almost surely.}$$

For every $z \geq 0$, let $\Pi^\star(z)$ be the unique pure Bayesian Nash equilibrium, and define

$$\widehat{\chi}_T := \gamma^T + \frac{\lambda N}{N+1}\frac{s_T(\gamma)}{\mu_\gamma}.$$

If $\widehat{\chi}_T < 1$, then the following statements hold.

1. *The conditional-mean map $\overline{\Psi}$ has a unique fixed point $\overline{z}^\star$. Moreover, the deterministic recursion $\overline{z}_{k+1} = \overline{\Psi}(\overline{z}_k)$ satisfies*

   $$|\overline{z}_k - \overline{z}^\star| \leq \widehat{\chi}_T^k |\overline{z}_0 - \overline{z}^\star|.$$

   Consequently, its unique stationary distribution is $\delta_{\overline{z}^\star}$.

2. *The realized inherited-state process*

   $$Z_{k+1} = \Psi(Z_k; \boldsymbol{B}^{k+1})$$

   admits a unique stationary distribution $\varpi^\star \in \mathcal{P}_1(\mathbb{R}_+)$. If $\mathcal{P}$ denotes its transition kernel, then, for every initial law $\nu \in \mathcal{P}_1(\mathbb{R}_+)$,

   $$W_1(\nu\mathcal{P}^k, \varpi^\star) \leq \widehat{\chi}_T^k W_1(\nu, \varpi^\star).$$

Proof. The main step is to control how the Bayesian Nash equilibrium changes with the entering state. For $z, z' \geq 0$, write

$$\Pi := \Pi^\star(z), \qquad \Pi' := \Pi^\star(z'), \qquad \Delta\Pi := \Pi - \Pi',$$

and, for each quota realization $\boldsymbol{B}$, define

$$\Delta \mathrm{X}(\boldsymbol{B}) := \mathrm{X}^{\Pi}(\boldsymbol{B}) - \mathrm{X}^{\Pi'}(\boldsymbol{B}), \qquad \Delta\boldsymbol{q}(\boldsymbol{B}) := \boldsymbol{q}_z^\star(\boldsymbol{B}) - \boldsymbol{q}_{z'}^\star(\boldsymbol{B}).$$

Let Φ_z denote the Bayesian pseudo-gradient when the entering state is z. The variational inequalities satisfied by Π and Π' are

$$\langle \Phi_z(\Pi), \Pi' - \Pi\rangle_{\mathcal{D}} \geq 0, \qquad \langle \Phi_{z'}(\Pi'), \Pi - \Pi'\rangle_{\mathcal{D}} \geq 0.$$

Adding these inequalities and adding and subtracting $\Phi_z(\Pi')$ gives

$$\langle \Phi_z(\Pi) - \Phi_z(\Pi'), \Delta\Pi\rangle_{\mathcal{D}} \leq -\langle \Phi_z(\Pi') - \Phi_{z'}(\Pi'), \Delta\Pi\rangle_{\mathcal{D}}\,. \tag{60}$$

For every realization $\boldsymbol{B}$, Theorem 2.1 gives

$$\begin{aligned}\Big\langle \mathrm{F}_z(\mathrm{X}^{\Pi}(\boldsymbol{B})) - \mathrm{F}_z(\mathrm{X}^{\Pi'}(\boldsymbol{B})), \Delta \mathrm{X}(\boldsymbol{B})\Big\rangle_{\mathrm{F}} \\ \geq \mu_\gamma \left(\|\Delta \mathrm{X}(\boldsymbol{B})\|_{\mathrm{F}}^2 + \|\Delta\boldsymbol{q}(\boldsymbol{B})\|_2^2\right).\end{aligned}$$

Taking expectations and using the definition of the probability-weighted inner product therefore bounds the left-hand side of (60) from below by

$$\mu_\gamma \mathbb{E}\big[\|\Delta \mathrm{X}(\boldsymbol{B})\|_{\mathrm{F}}^2 + \|\Delta\boldsymbol{q}(\boldsymbol{B})\|_2^2\big]\,.$$

The entering state appears in every agent's gradient only through the term $z\boldsymbol{d}_\gamma$. Hence the right-hand side of (60) is bounded above by

$$|z - z'|\,|\langle \boldsymbol{d}_\gamma, \mathbb{E}[\Delta\boldsymbol{q}(\boldsymbol{B})]\rangle|\,.$$

Combining the preceding bounds yields

$$\mu_\gamma \mathbb{E}\big[\|\Delta \mathrm{X}(\boldsymbol{B})\|_{\mathrm{F}}^2 + \|\Delta\boldsymbol{q}(\boldsymbol{B})\|_2^2\big] \leq |z - z'|\,|\langle \boldsymbol{d}_\gamma, \mathbb{E}[\Delta\boldsymbol{q}(\boldsymbol{B})]\rangle|\,. \tag{61}$$

For every $\boldsymbol{B}$,

$$\|\Delta \boldsymbol{q}(\boldsymbol{B})\|_2^2 \le N \|\Delta \mathrm{X}(\boldsymbol{B})\|_{\mathrm{F}}^2.$$

Moreover, Jensen's and Cauchy–Schwarz inequalities give

$$\|\mathbb{E}[\Delta \boldsymbol{q}(\boldsymbol{B})]\|_2 \le \left(\mathbb{E}\|\Delta \boldsymbol{q}(\boldsymbol{B})\|_2^2\right)^{1/2}.$$

Applying these two inequalities to (61) and cancelling the common factor when it is nonzero gives

$$\left(\mathbb{E}\|\Delta \boldsymbol{q}(\boldsymbol{B})\|_2^2\right)^{1/2} \le \frac{N}{N+1} \frac{\|\boldsymbol{d}_\gamma\|_2}{\mu_\gamma} |z - z'|. \tag{62}$$

The inequality is immediate when the left-hand side is zero.

We now compare the next states generated from z and z' using the same quota realization. By (59), Cauchy–Schwarz, and (62),

$$\begin{aligned} \mathbb{E}_{\boldsymbol{B}} |\Psi(z; \boldsymbol{B}) - \Psi(z'; \boldsymbol{B})| &\le \gamma^T |z - z'| + \lambda \|\boldsymbol{w}_\gamma\|_2 \left(\mathbb{E}\|\Delta \boldsymbol{q}(\boldsymbol{B})\|_2^2\right)^{1/2} \\ &\le \widehat{\chi}_T |z - z'|, \end{aligned}$$

where $\|\boldsymbol{w}_\gamma\|_2 = \|\boldsymbol{d}_\gamma\|_2 = \sqrt{s_T(\gamma)}$. In particular,

$$|\overline{\Psi}(z) - \overline{\Psi}(z')| \le \widehat{\chi}_T |z - z'|.$$

Thus $\overline{\Psi}$ is a contraction on $\mathbb{R}_+$. It maps $\mathbb{R}_+$ into itself and is finite because

$$0 \le \Psi(z; \boldsymbol{B}) \le \gamma^T z + \lambda V_{\max}.$$

Banach's fixed-point theorem gives a unique fixed point $\overline{z}^\star$ and the stated geometric convergence of the deterministic conditional-mean recursion. The only stationary distribution of this deterministic recursion is therefore $\delta_{\overline{z}^\star}$.

It remains to consider the realized process. Its transition kernel is

$$\mathcal{P}(z, \cdot) := \mathrm{Law}(\Psi(z; \boldsymbol{B})).$$

Let $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}_+)$, let (Z, Z') be any coupling of ν and ν', and draw $\boldsymbol{B} \sim \mathcal{D}$ independently of (Z, Z'). Using the same quota realization in the two transitions and conditioning on (Z, Z') gives

$$\mathbb{E}[|\Psi(Z; \boldsymbol{B}) - \Psi(Z'; \boldsymbol{B})| \,|\, Z, Z'] \le \widehat{\chi}_T |Z - Z'|.$$

Taking expectations and then the infimum over all couplings yields

$$W_1(\nu\mathcal{P}, \nu'\mathcal{P}) \le \widehat{\chi}_T W_1(\nu, \nu'). \tag{63}$$

The bound $\Psi(z; \boldsymbol{B}) \le \gamma^T z + \lambda V_{\max}$ also shows that $\mathcal{P}$ maps $\mathcal{P}_1(\mathbb{R}_+)$ into itself. Since $(\mathcal{P}_1(\mathbb{R}_+), W_1)$ is complete and $\widehat{\chi}_T < 1$, the map $\nu \mapsto \nu\mathcal{P}$ is a contraction. Banach's fixed-point theorem therefore gives a unique stationary distribution $\varpi^\star$. Finally, iterating (63) gives

$$W_1(\nu\mathcal{P}^k, \varpi^\star) = W_1(\nu\mathcal{P}^k, \varpi^\star\mathcal{P}^k) \le \widehat{\chi}_T^k W_1(\nu, \varpi^\star),$$

which proves geometric convergence to the stationary inherited-state law. □

The fixed point $\overline{z}^\star$ and the stationary law $\varpi^\star$ describe different dynamics. The former corresponds to repeatedly replacing the next state by its conditional expectation, whereas the latter describes the distribution of the realized inherited state under random quota draws. In general, $\varpi^\star$ need not be a point mass.

The algebraic argument from Remark E.8 can also be applied in the Bayesian case. Thus, Proposition E.9 holds for $\chi_T = \min\{\rho, \widehat{\chi}_T\}$, without any further assumptions except for (A).

E.5 Misspecified Parameters under Price and State Noise

We consider parameter misspecification in the presence of the price and state noise. In episode k, the observed price and impact state satisfy

$$P_t^k = I_t^k + \kappa q_t^k + \varepsilon_t^k, \qquad I_{t+1}^k = \gamma I_t^k + \lambda q_t^k + \xi_{t+1}^k.$$

Both $\boldsymbol{\varepsilon}^k$ and $\boldsymbol{\xi}^k$ are conditionally zero-mean and independent of the schedule profile selected for episode k. For the explicit second-moment bounds, we assume that they are zero-mean, mutually independent, and satisfy

$$\mathbb{E}[(\varepsilon_t^k)^2] = \sigma_\varepsilon^2, \qquad \mathbb{E}[(\xi_t^k)^2] = \sigma_\xi^2.$$

Unrolling the state recursion shows that the noise entering the public-price gradient oracle is the vector $\boldsymbol{\nu}^k$ with coordinates

$$\nu_t^k := \varepsilon_t^k + \sum_{s=2}^{t} \gamma^{t-s} \xi_s^k.$$

The sum is empty for $t = 1$. Conditional on the information available before episode k and on X^k,

$$\mathbb{E}[\boldsymbol{\nu}^k \mid \mathcal{F}_{k-1}, \mathrm{X}^k] = \mathbf{0}.$$

Moreover,

$$\mathbb{E}\|\boldsymbol{\nu}^k\|_2^2 = T\sigma_\varepsilon^2 + \sigma_\xi^2 \sum_{r=0}^{T-2} (T-1-r)\gamma^{2r} \le T\sigma_\varepsilon^2 + T\sigma_\xi^2 s_T(\gamma). \tag{64}$$

We denote the right-hand side of the equality by σ_ν^2.

Let $\boldsymbol{g}_i^k$ be the gradient of the corresponding deterministic game under the true parameters. If the true parameters are used in the own-schedule correction, the public-price oracle returns $\boldsymbol{g}_i^k + \boldsymbol{\nu}^k$. If agent i instead uses parameters $(\widehat{\kappa}, \widehat{\lambda}, \widehat{\gamma})$, its computed gradient is

$$\widetilde{\boldsymbol{g}}_i^k = \boldsymbol{g}_i^k + \boldsymbol{\nu}^k + \boldsymbol{b}_i^k,$$

where $\boldsymbol{b}_i^k$ is the systematic error caused by parameter misspecification. We replace $\boldsymbol{g}_i^k$ and $\widetilde{\boldsymbol{g}_i^k}$ with their expressions from (45). The observed price appears identically in the correctly specified and misspecified gradient oracles. Subtracting their own-schedule corrections, the noise vector $\boldsymbol{\nu}^k$ is canceled and therefore we obtain

$$\boldsymbol{b}_i^k = (\widehat{\kappa} - \kappa)\boldsymbol{x}_i^k + \left(\widehat{\lambda}\mathrm{U}_{\widehat{\gamma}} - \lambda \mathrm{U}_\gamma\right) \boldsymbol{x}_i^k.$$

We note that $\boldsymbol{\nu}^k$ is conditionally mean-zero oracle noise, whereas $\boldsymbol{b}_i^k$ is generally biased.

Proposition E.10. *Suppose that agent i computes its gradient using parameters $(\widehat{\kappa}, \widehat{\lambda}, \widehat{\gamma})$, while the true parameters are $(\kappa, \lambda, \gamma)$. For every feasible schedule $\boldsymbol{x}_i^k \in \mathcal{X}_i$,*

$$\|\boldsymbol{b}_i^k\|_2 \le B_i \left(|\widehat{\kappa} - \kappa| + |\widehat{\lambda} - \lambda| \sqrt{s_{T-1}(\widehat{\gamma})} + \lambda |\widehat{\gamma} - \gamma| \sqrt{\frac{1+\rho^2}{(1-\rho^2)^3}} \right),$$

where $\rho := \max\{\gamma, \widehat{\gamma}\} < 1$.

Proof. Let

$$\mathrm{D} := \widehat{\lambda}\mathrm{U}_{\widehat{\gamma}} - \lambda \mathrm{U}_\gamma.$$

By the triangle inequality,

$$\|\boldsymbol{b}_i^k\|_2 \le |\widehat{\kappa} - \kappa| \, \|\boldsymbol{x}_i^k\|_2 + \|\mathrm{D}\boldsymbol{x}_i^k\|_2.$$

Since $\boldsymbol{x}_i^k \ge \mathbf{0}$ and $\mathbf{1}^\top \boldsymbol{x}_i^k = B_i$, $\|\boldsymbol{x}_i^k\|_2 \le B_i$. Moreover,

$$\|\mathrm{D}\boldsymbol{x}_i^k\|_2 \le \sum_{s=1}^{T} x_{i,s}^k \|\mathrm{D}\boldsymbol{e}_s\|_2 \le B_i \max_{1 \le s \le T} \|\mathrm{D}\boldsymbol{e}_s\|_2.$$

The nonzero entries of $\mathrm{D}\boldsymbol{e}_s$ are indexed by $r = 0, \dots, s-2$ and satisfy

$$\widehat{\lambda}\widehat{\gamma}^r - \lambda\gamma^r = (\widehat{\lambda} - \lambda)\widehat{\gamma}^r + \lambda(\widehat{\gamma}^r - \gamma^r).$$

Minkowski's inequality therefore gives

$$\|\mathrm{D}\boldsymbol{e}_s\|_2 \le |\widehat{\lambda} - \lambda| \left(\sum_{r=0}^{s-2} \widehat{\gamma}^{2r}\right)^{1/2} + \lambda \left(\sum_{r=0}^{s-2} (\widehat{\gamma}^r - \gamma^r)^2\right)^{1/2}.$$

For every $r \ge 1$, factoring the difference of powers yields

$$\widehat{\gamma}^r - \gamma^r = (\widehat{\gamma} - \gamma) \sum_{\ell=0}^{r-1} \widehat{\gamma}^{r-1-\ell}\gamma^\ell.$$

Every term in the sum is at most ρ^{r-1}, so

$$|\widehat{\gamma}^r - \gamma^r| \le r\rho^{r-1}|\widehat{\gamma} - \gamma|.$$

Consequently,

$$\left(\sum_{r=0}^{s-2} (\widehat{\gamma}^r - \gamma^r)^2\right)^{1/2} \le |\widehat{\gamma} - \gamma| \left(\sum_{r=1}^{\infty} r^2\rho^{2r-2}\right)^{1/2}.$$

Differentiating the geometric-series identity twice gives

$$\sum_{r=1}^{\infty} r^2\rho^{2r-2} = \frac{1+\rho^2}{(1-\rho^2)^3}.$$

The remaining geometric sum is maximized at $s = T$, which proves the claim. □

We next show that both learning guarantees tolerate the simultaneous presence of $\boldsymbol{\nu}^k$ and $\boldsymbol{b}_i^k$. The noise changes only the second-moment constants, whereas the systematic error contributes through its cumulative squared norm.

Theorem E.11. *Assume that $\boldsymbol{b}_i^k$ is determined before the noise of episode k is realized, and let agent i update as*

$$\boldsymbol{x}_i^{k+1} = \mathsf{P}_{\mathcal{X}_i}\left(\boldsymbol{x}_i^k - \eta_k \widetilde{\boldsymbol{g}}_i^k\right), \qquad \eta_k = \frac{1}{\mu_\gamma k},$$

where

$$\widetilde{\boldsymbol{g}}_i^k = \boldsymbol{g}_i^k + \boldsymbol{\nu}^k + \boldsymbol{b}_i^k.$$

Then, against any sequence of competing schedules, the pseudo-regret for the underlying deterministic game satisfies

$$\begin{aligned}\mathrm{Reg}_i^{\mathrm{ps}}(K) &:= \max_{\boldsymbol{u}_i \in \mathcal{X}_i} \mathbb{E}\left[\sum_{k=1}^{K} \left(C_i(\boldsymbol{x}_i^k; \mathrm{X}_{-i}^k) - C_i(\boldsymbol{u}_i; \mathrm{X}_{-i}^k)\right)\right] \\ &\le \frac{\mathcal{G}_i(z)^2 + \sigma_\nu^2/2}{\mu_\gamma}(1 + \log K) + \frac{3}{2\mu_\gamma}\sum_{k=1}^{K} \mathbb{E}\|\boldsymbol{b}_i^k\|_2^2.\end{aligned}$$

Proof. Fix $\boldsymbol{u}_i \in \mathcal{X}_i$ and write $D_k := \|\boldsymbol{x}_i^k - \boldsymbol{u}_i\|_2^2$. Since the deterministic one-episode cost is $2\mu_\gamma$-strongly convex in $\boldsymbol{x}_i$, we have

$$C_i(\boldsymbol{x}_i^k; \mathrm{X}_{-i}^k) - C_i(\boldsymbol{u}_i; \mathrm{X}_{-i}^k) \le \langle \boldsymbol{g}_i^k, \boldsymbol{x}_i^k - \boldsymbol{u}_i\rangle - \mu_\gamma D_k.$$

Substituting $\boldsymbol{g}_i^k = \widetilde{\boldsymbol{g}}_i^k - \boldsymbol{\nu}^k - \boldsymbol{b}_i^k$ and applying Young's inequality to the bias term gives

$$\begin{aligned}&C_i(\boldsymbol{x}_i^k; \mathrm{X}_{-i}^k) - C_i(\boldsymbol{u}_i; \mathrm{X}_{-i}^k) \\ &\quad \le \langle \widetilde{\boldsymbol{g}}_i^k, \boldsymbol{x}_i^k - \boldsymbol{u}_i\rangle - \langle \boldsymbol{\nu}^k, \boldsymbol{x}_i^k - \boldsymbol{u}_i\rangle - \frac{\mu_\gamma}{2} D_k + \frac{\|\boldsymbol{b}_i^k\|_2^2}{2\mu_\gamma}.\end{aligned}$$

Projection nonexpansiveness implies

$$\langle \widetilde{\boldsymbol{g}}_i^k, \boldsymbol{x}_i^k - \boldsymbol{u}_i \rangle \le \frac{D_k - D_{k+1}}{2\eta_k} + \frac{\eta_k}{2} \|\widetilde{\boldsymbol{g}}_i^k\|_2^2.$$

Summing over $k = 1, \dots, K$ telescopes the squared-distance terms and leaves the nonpositive terminal term $-D_{K+1}/(2\eta_K)$.

Moreover, since $\boldsymbol{x}_i^k - \boldsymbol{u}_i$ is measurable with respect to $\sigma(\mathcal{F}_{k-1}, \mathrm{X}^k)$ and

$$\mathbb{E}[\boldsymbol{\nu}^k \mid \mathcal{F}_{k-1}, \mathrm{X}^k] = \mathbf{0},$$

we have

$$\mathbb{E}\big[\big\langle \boldsymbol{\nu}^k, \boldsymbol{x}_i^k - \boldsymbol{u}_i \big\rangle \,\big|\, \mathcal{F}_{k-1}, \mathrm{X}^k\big] = 0.$$

The tower property therefore implies

$$\mathbb{E}\big[\big\langle \boldsymbol{\nu}^k, \boldsymbol{x}_i^k - \boldsymbol{u}_i \big\rangle\big] = 0.$$

Taking expectations, dropping the nonpositive terminal term, and maximizing over $\boldsymbol{u}_i \in \mathcal{X}_i$ yields

$$\mathrm{Reg}_i^{\mathrm{ps}}(K) \le \frac{1}{2\mu_\gamma} \sum_{k=1}^{K} \frac{1}{k} \mathbb{E}\|\widetilde{\boldsymbol{g}}_i^k\|_2^2 + \frac{1}{2\mu_\gamma} \sum_{k=1}^{K} \mathbb{E}\|\boldsymbol{b}_i^k\|_2^2. \tag{65}$$

Conditionally on $\mathcal{H}_k := \sigma(\mathcal{F}_{k-1}, \mathrm{X}^k)$, the vectors $\boldsymbol{g}_i^k$ and $\boldsymbol{b}_i^k$ are measurable, while $\mathbb{E}[\boldsymbol{\nu}^k \mid \mathcal{H}_k] = \mathbf{0}$. Hence,

$$\begin{aligned}\mathbb{E}\big[\|\widetilde{\boldsymbol{g}}_i^k\|_2^2 \,\big|\, \mathcal{H}_k\big] &= \|\boldsymbol{g}_i^k + \boldsymbol{b}_i^k\|_2^2 + \mathbb{E}\big[\|\boldsymbol{\nu}^k\|_2^2 \,\big|\, \mathcal{H}_k\big] \\ &\le 2\mathcal{G}_i(z)^2 + 2\|\boldsymbol{b}_i^k\|_2^2 + \sigma_\nu^2.\end{aligned}$$

Taking expectations once more and applying the tower property gives

$$\mathbb{E}\|\widetilde{\boldsymbol{g}}_i^k\|_2^2 \le 2\mathcal{G}_i(z)^2 + 2\mathbb{E}\|\boldsymbol{b}_i^k\|_2^2 + \sigma_\nu^2.$$

Substituting this bound into (65) and using $\sum_{k=1}^{K} 1/k \le 1 + \log K$ proves the claimed bound. □

Thus, if $\mathbb{E}\|\boldsymbol{b}_i^k\|_2^2 = O(1/k)$, the pseudo-regret remains $O(\log K)$ despite both price and state noise.

Theorem E.12. *Let $\mathrm{X}^\star$ be the Nash equilibrium of the deterministic game under the true parameters. Suppose that all agents use*

$$\widetilde{\mathrm{F}}^{\,k} = \mathrm{F}(\mathrm{X}^k) + \mathcal{E}^k + \mathcal{N}^k,$$

where the i-th rows of $\mathcal{E}^k$ and $\mathcal{N}^k$ are $\boldsymbol{b}_i^k$ and $\boldsymbol{\nu}^k$, respectively.

If $\mathrm{X}^{k+1} = \mathsf{P}_{\mathcal{X}}\left(\mathrm{X}^k - \eta_k \widetilde{\mathrm{F}}^{\,k}\right)$ and $\eta_k = \frac{2}{\mu_\gamma(k+1)}$, then

$$\mathbb{E}\|\mathrm{X}^{K+1} - \mathrm{X}^\star\|_{\mathrm{F}}^2 \le \frac{8G^2 + 4N\sigma_\nu^2}{\mu_\gamma^2(K+1)} + \frac{6}{\mu_\gamma^2 K(K+1)} \sum_{k=1}^{K} k\, \mathbb{E}\|\mathcal{E}^k\|_{\mathrm{F}}^2,$$

where $G^2 := \sum_{i=1}^{N} \mathcal{G}_i(z)^2$.

Proof. Let

$$\Delta\mathrm{X}^k := \mathrm{X}^k - \mathrm{X}^\star, \qquad e_k := \mathbb{E}\|\Delta\mathrm{X}^k\|_{\mathrm{F}}^2.$$

Projection nonexpansiveness gives

$$\begin{aligned}\|\Delta\mathrm{X}^{k+1}\|_{\mathrm{F}}^2 \le \|\Delta\mathrm{X}^k\|_{\mathrm{F}}^2 &- 2\eta_k \left\langle \mathrm{F}(\mathrm{X}^k), \Delta\mathrm{X}^k \right\rangle_{\mathrm{F}} \\ &- 2\eta_k \left\langle \mathcal{E}^k, \Delta\mathrm{X}^k \right\rangle_{\mathrm{F}} - 2\eta_k \left\langle \mathcal{N}^k, \Delta\mathrm{X}^k \right\rangle_{\mathrm{F}} + \eta_k^2 \|\widetilde{\mathrm{F}}^{\,k}\|_{\mathrm{F}}^2.\end{aligned}$$

The equilibrium variational inequality and the strong monotonicity of F imply

$$\left\langle \mathrm{F}(\mathrm{X}^k), \Delta\mathrm{X}^k \right\rangle_{\mathrm{F}} \ge \mu_\gamma \|\Delta\mathrm{X}^k\|_{\mathrm{F}}^2.$$

Moreover, Young's inequality gives

$$2\left|\langle \mathcal{E}^k, \Delta \mathrm{X}^k \rangle_{\mathrm{F}}\right| \le \mu_\gamma \|\Delta \mathrm{X}^k\|_{\mathrm{F}}^2 + \frac{1}{\mu_\gamma}\|\mathcal{E}^k\|_{\mathrm{F}}^2.$$

Because $\Delta \mathrm{X}^k$ is measurable with respect to $\sigma(\mathcal{F}_{k-1}, \mathrm{X}^k)$, the conditional mean-zero property of $\mathcal{N}^k$ gives

$$\mathbb{E}\big[\langle \mathcal{N}^k, \Delta \mathrm{X}^k \rangle_{\mathrm{F}} \,\big|\, \mathcal{F}_{k-1}, \mathrm{X}^k\big] = 0.$$

It remains to bound the squared-gradient term. Conditionally on the past and the selected profile,

$$\begin{aligned}\mathbb{E}\Big[\|\widetilde{\mathrm{F}}^k\|_{\mathrm{F}}^2 \,\Big|\, \mathcal{F}_{k-1}, \mathrm{X}^k\Big] &= \|\mathrm{F}(\mathrm{X}^k) + \mathcal{E}^k\|_{\mathrm{F}}^2 + \mathbb{E}\big[\|\mathcal{N}^k\|_{\mathrm{F}}^2 \,\big|\, \mathcal{F}_{k-1}, \mathrm{X}^k\big] \\ &\le 2G^2 + 2\|\mathcal{E}^k\|_{\mathrm{F}}^2 + N\sigma_\nu^2.\end{aligned}$$

Taking expectations and combining the preceding bounds yields

$$e_{k+1} \le (1 - \mu_\gamma \eta_k) e_k + \left(\frac{\eta_k}{\mu_\gamma} + 2\eta_k^2\right)\mathbb{E}\|\mathcal{E}^k\|_{\mathrm{F}}^2 + \eta_k^2\left(2G^2 + N\sigma_\nu^2\right).$$

Substituting $\eta_k = 2/(\mu_\gamma(k+1))$ gives

$$e_{k+1} \le \frac{k-1}{k+1} e_k + \left(\frac{2}{\mu_\gamma^2(k+1)} + \frac{8}{\mu_\gamma^2(k+1)^2}\right)\mathbb{E}\|\mathcal{E}^k\|_{\mathrm{F}}^2 + \frac{4(2G^2 + N\sigma_\nu^2)}{\mu_\gamma^2(k+1)^2}.$$

Iterating this recursion and using

$$\prod_{\ell=k+1}^{K} \frac{\ell - 1}{\ell + 1} = \frac{k(k+1)}{K(K+1)}$$

gives

$$e_{K+1} \le \frac{4(2G^2 + N\sigma_\nu^2)}{\mu_\gamma^2 K(K+1)} \sum_{k=1}^{K} \frac{k}{k+1} + \frac{2}{\mu_\gamma^2 K(K+1)} \sum_{k=1}^{K} k\left(1 + \frac{4}{k+1}\right)\mathbb{E}\|\mathcal{E}^k\|_{\mathrm{F}}^2.$$

Since $\sum_{k=1}^{K} \frac{k}{k+1} \le K$, we obtain

$$e_{K+1} \le \frac{8G^2 + 4N\sigma_\nu^2}{\mu_\gamma^2(K+1)} + \frac{2}{\mu_\gamma^2 K(K+1)} \sum_{k=1}^{K} k\left(1 + \frac{4}{k+1}\right)\mathbb{E}\|\mathcal{E}^k\|_{\mathrm{F}}^2.$$

Finally, $1 + 4/(k+1) \le 3$ for every $k \ge 1$, concludes the proof. □

The conditional second-moment assumption is automatic when the parameter estimates remain in a compact admissible set. Indeed, Proposition E.10 then bounds $\|\mathcal{E}^k\|_{\mathrm{F}}$ uniformly, while (64) bounds the contribution of $\mathcal{N}^k$. In particular, if $\mathbb{E}\|\mathcal{E}^k\|_{\mathrm{F}}^2 = O(1/k)$, then the weighted sum in Theorem E.12 is $O(K)$ and

$$\mathbb{E}\|\mathrm{X}^K - \mathrm{X}^\star\|_{\mathrm{F}}^2 = O(1/K).$$

E.5.1 Estimation of Parameter γ from Demand Observations

We conclude by showing that the required $O(1/k)$ parameter-error bound can be obtained by ordinary least squares (OLS) when κ and λ are known, the learning algorithm can observe the demand volume q_t in each round, the impact state is sufficiently informative, and there is no price noise. The restriction $\varepsilon_t^k = 0$ is essential for ordinary least squares because it makes $I_t^k = P_t^k - \kappa q_t^k$ an exactly observed regressor.

Lemma E.13. *Suppose that κ and λ are known, $\varepsilon_t^k = 0$, and*

$$I_{t+1}^k = \gamma I_t^k + \lambda q_t^k + \xi_{t+1}^k.$$

Assume that $\{\xi_{t+1}^k\}$ is a martingale-difference sequence with respect to the within-episode history $\mathcal{F}_{k,t}$ and satisfies

$$\mathbb{E}[\xi_{t+1}^k \mid \mathcal{F}_{k,t}] = 0, \qquad \mathbb{E}[(\xi_{t+1}^k)^2 \mid \mathcal{F}_{k,t}] \le \sigma_\xi^2.$$

Suppose further that $|I_t^k| \le I_{\max}$ *and that there exists* $c_{\mathrm{exc}} > 0$ *such that*

$$\sum_{t=1}^{T-1} (I_t^k)^2 \ge c_{\mathrm{exc}} \tag{66}$$

in every episode. Define

$$S_k := \sum_{\ell=1}^{k} \sum_{t=1}^{T-1} (I_t^\ell)^2$$

and

$$\widehat{\gamma}_k^{\mathrm{LS}} := \frac{\sum_{\ell=1}^{k} \sum_{t=1}^{T-1} I_t^\ell (I_{t+1}^\ell - \lambda q_t^\ell)}{S_k}.$$

Let $\widehat{\gamma}_k$ *be the projection of* $\widehat{\gamma}_k^{\mathrm{LS}}$ *onto* $[0, \bar{\gamma}]$*, where* $\gamma \le \bar{\gamma} < 1$*. Then*

$$\mathbb{E}[(\widehat{\gamma}_k - \gamma)^2] \le \frac{c_{\mathrm{LS}}}{k}, \qquad c_{\mathrm{LS}} := \frac{(T-1) I_{\max}^2 \sigma_\xi^2}{c_{\mathrm{exc}}^2}.$$

Proof. Since $\varepsilon_t^k = 0$, the impact state is recovered exactly as $I_t^k = P_t^k - \kappa q_t^k$. Substituting the state dynamics into the OLS estimator gives

$$\widehat{\gamma}_k^{\mathrm{LS}} - \gamma = \frac{\sum_{\ell=1}^{k} \sum_{t=1}^{T-1} I_t^\ell \xi_{t+1}^\ell}{S_k}.$$

Assumption (66) implies $S_k \ge c_{\mathrm{exc}} k$. Hence

$$\mathbb{E}[(\widehat{\gamma}_k^{\mathrm{LS}} - \gamma)^2] \le \frac{1}{c_{\mathrm{exc}}^2 k^2} \mathbb{E}\left[\left(\sum_{\ell=1}^{k} \sum_{t=1}^{T-1} I_t^\ell \xi_{t+1}^\ell \right)^2 \right].$$

Since I_t^ℓ is $\mathcal{F}_{\ell,t}$-measurable and $\mathbb{E}[\xi_{t+1}^\ell \mid \mathcal{F}_{\ell,t}] = 0$, we have

$$\mathbb{E}\big[I_t^\ell \xi_{t+1}^\ell \,\big|\, \mathcal{F}_{\ell,t} \big] = I_t^\ell \mathbb{E}\big[\xi_{t+1}^\ell \,\big|\, \mathcal{F}_{\ell,t} \big] = 0.$$

Thus, the variables $\{I_t^\ell \xi_{t+1}^\ell\}_{\ell,t}$ form a martingale-difference sequence. In particular, distinct terms are orthogonal in expectation, so all cross terms vanish when expanding the square of their sum. Therefore,

$$\begin{aligned} \mathbb{E}\left[\left(\sum_{\ell=1}^{k} \sum_{t=1}^{T-1} I_t^\ell \xi_{t+1}^\ell \right)^2 \right] &= \sum_{\ell=1}^{k} \sum_{t=1}^{T-1} \mathbb{E}[(I_t^\ell)^2 (\xi_{t+1}^\ell)^2] \\ &\le k(T-1) I_{\max}^2 \sigma_\xi^2. \end{aligned}$$

The claimed bound follows because projection onto an interval containing γ cannot increase the estimation error. □

Theorem E.14. *Suppose that the conditions of Lemma E.13 hold. Assume that all agents use* $\widehat{\gamma}_k$ *in episode* $k+1$*, and define*

$$\boldsymbol{b}_i^{k+1} := \lambda \left(\mathrm{U}_{\widehat{\gamma}_k} - \mathrm{U}_\gamma \right) \boldsymbol{x}_i^{k+1}.$$

Let $\mathcal{E}^{k+1}$ *be the matrix whose* i*-th row is* $\boldsymbol{b}_i^{k+1}$*. Then*

$$\mathbb{E}\|\mathcal{E}^{k+1}\|_{\mathrm{F}}^2 \le \frac{c_{\mathrm{bias}}}{k},$$

where

$$c_{\mathrm{bias}} := \lambda^2 c_{\mathrm{LS}} \frac{1 + \bar{\gamma}^2}{(1 - \bar{\gamma}^2)^3} \sum_{i=1}^{N} B_i^2.$$

Consequently, under the updates of Theorems E.11,

$$\mathrm{Reg}_i^{\mathrm{ps}}(K) = O(\log K)$$

for every agent i*, and under the updates of E.12*

$$\mathbb{E}\|\mathrm{X}^K - \mathrm{X}^\star\|_{\mathrm{F}}^2 = O(1/K).$$

Proof. For $u, v \in [0, \bar{\gamma}]$, the argument of Proposition E.10 gives

$$\max_{1 \le s \le T} \|(\mathrm{U}_u - \mathrm{U}_v)\boldsymbol{e}_s\|_2^2 \le |u - v|^2 \frac{1 + \bar{\gamma}^2}{(1 - \bar{\gamma}^2)^3}.$$

Since $\boldsymbol{x}_i^{k+1} \ge \mathbf{0}$ and $\mathbf{1}^\top \boldsymbol{x}_i^{k+1} = B_i$,

$$\|\boldsymbol{b}_i^{k+1}\|_2^2 \le \lambda^2 B_i^2 \frac{1 + \bar{\gamma}^2}{(1 - \bar{\gamma}^2)^3} (\widehat{\gamma}_k - \gamma)^2.$$

Taking expectations, summing over the agents, and applying Lemma E.13 yields

$$\mathbb{E}\|\mathcal{E}^{k+1}\|_{\mathrm{F}}^2 = \sum_{i=1}^{N} \mathbb{E}\|\boldsymbol{b}_i^{k+1}\|_2^2 \le \frac{c_{\mathrm{bias}}}{k}.$$

It follows that

$$\sum_{k=1}^{K} \mathbb{E}\|\boldsymbol{b}_i^k\|_2^2 = O(\log K),$$

so Theorem E.11 gives the pseudo-regret claim. Similarly,

$$\sum_{k=1}^{K} k\, \mathbb{E}\|\mathcal{E}^k\|_{\mathrm{F}}^2 = O(K),$$

and Theorem E.12 gives the $O(1/K)$ mean-square convergence rate. The direct state-noise contribution is already included in σ_ν^2 and therefore changes only the constants in both conclusions. □

F Linear External Regret for Single-Episode Learning: The Proof of Theorem 4.1

In this section, we give a detailed proof of Theorem 4.1. The construction is simple and consists of two competing demand schedules, $\boldsymbol{y}^{(0)}$ and $\boldsymbol{y}^{(1)}$, which are identical in the first half of the episode and significantly different in the second half.

For some positive integer $\tau \ge 1$, we let $T = 2\tau$ and recall that $B_i = bT$. For simplicity, we assume that $z = 0$ throughout this proof. Consider the two deterministic competing-demand schedules

$$y_t^{(0)} = 0 \quad \text{and} \quad y_t^{(1)} = \begin{cases} 0, & t \le \tau, \\ 4b, & t > \tau. \end{cases}$$

The second schedule has total demand $4b\tau = 2bT$ and both schedules have total demand $O(T)$ for fixed b.

To give the intuition, we first focus on the simplest special case where $\lambda = 0$ and there is no persistent impact on the price. Since $z = 0$ and $\lambda = 0$, the cost of any feasible schedule $\boldsymbol{x}$ is

$$C_i(\boldsymbol{x}; \boldsymbol{y}) = \kappa\big(\|\boldsymbol{x}\|_2^2 + \langle \boldsymbol{x}, \boldsymbol{y} \rangle\big). \tag{67}$$

Against $\boldsymbol{y}^{(0)}$, the hindsight optimum is the uniform schedule $\boldsymbol{x}^{\star,0} = b\mathbf{1}$ and the regret of any feasible schedule $\boldsymbol{x}$ is:

$$C_i(\boldsymbol{x}; \boldsymbol{y}^{(0)}) - C_i(\boldsymbol{x}^{\star,0}; \boldsymbol{y}^{(0)}) = \kappa \|\boldsymbol{x} - b\mathbf{1}\|_2^2.$$

Against $\boldsymbol{y}^{(1)}$, the hindsight optimum $\boldsymbol{x}^{\star,1}$ acquires $2b$ in every period of the first half and nothing thereafter, i.e., $x_t^{\star,1} = 2b\, \mathbb{1}\{t \le \tau\}$, which is feasible because $2b\tau = bT$. The regret of any feasible schedule $\boldsymbol{x}$ is:

$$C_i(\boldsymbol{x}; \boldsymbol{y}^{(1)}) - C_i(\boldsymbol{x}^{\star,1}; \boldsymbol{y}^{(1)}) = \kappa\, \|\boldsymbol{x} - \boldsymbol{x}^{\star,1}\|_2^2.$$

Moreover, the first half restrictions of $\boldsymbol{x}^{\star,0}$ and $\boldsymbol{x}^{\star,1}$ differ by the constant vector $b\mathbf{1}_\tau$, whose squared norm is $\tau b^2 = B_i^2/(2T)$.

Any (even randomized) online learning algorithm has exactly the same distribution of first-half actions in both environments, since it has observed exactly the same prices. For any realized first-half vector $\boldsymbol{a}$, we have

$$2\big(\|\boldsymbol{a} - b\mathbf{1}_\tau\|_2^2 + \|\boldsymbol{a} - 2b\mathbf{1}_\tau\|_2^2\big) \ge \|b\mathbf{1}_\tau\|_2^2 = B_i^2/(2T).$$

Finally, our objective in (67) is 2κ-strongly convex, so its suboptimality is at least κ times the squared distance to its minimizer. Multiplying the inequality above by $\kappa/4$ and taking expectations, we conclude that

$$\frac{1}{2}\sum_{\omega\in\{0,1\}}\mathbb{E}\Big[C_i(\boldsymbol{x}^{(\omega)};\boldsymbol{y}^{(\omega)})-C_i(\boldsymbol{x}^{\star,\omega};\boldsymbol{y}^{(\omega)})\Big]\;\geq\;\frac{\kappa B_i^2}{8T}\,.$$

F.1 Extension to Persistent Impact

The same argument extends to every $\lambda>0$ and $\gamma\in(0,1]$. The optimal schedules are no longer available in closed form, but we only need to know how much of their quota lies in the first half. As before, we consider $\boldsymbol{y}^{(0)}=\mathbf{0}$ and let

$$y_t^{(1)}=\begin{cases}0, & t\leq\tau,\\ Y, & t>\tau,\end{cases}$$

where the constant Y is specified below. Denote the corresponding hindsight-optimal schedules by $\boldsymbol{x}^{\star,0}$ and $\boldsymbol{x}^{\star,1}$. We shall prove the following three facts:

$$\sum_{t\leq\tau}x_t^{\star,0}=\frac{B_i}{2}, \tag{68}$$

$$\sum_{t\leq\tau}x_t^{\star,1}\geq\frac{7B_i}{8}, \tag{69}$$

$$C_i(\boldsymbol{x};\boldsymbol{y}^{(\omega)})-C_i(\boldsymbol{x}^{\star,\omega};\boldsymbol{y}^{(\omega)})\geq\mu_\gamma\,\|\boldsymbol{x}-\boldsymbol{x}^{\star,\omega}\|_2^2\qquad\text{for all }\omega\in\{0,1\}\,. \tag{70}$$

We show how to complete the lower bound assuming these facts. By Yao's lemma, it is enough to consider an arbitrary deterministic online algorithm against the uniform distribution over $\boldsymbol{y}^{(0)}$ and $\boldsymbol{y}^{(1)}$. The two price histories are identical through period τ, and hence this algorithm chooses the same first-half vector $\boldsymbol{a}$ in both environments. Using (70) and discarding the cost of the second half,

$$\begin{aligned}\mathbb{E}_{\omega\sim\mathcal{D}}[\mathrm{Reg}_\omega]&\geq\frac{\mu_\gamma}{2}\Big(\|\boldsymbol{a}-\boldsymbol{x}_{1:\tau}^{\star,0}\|_2^2+\|\boldsymbol{a}-\boldsymbol{x}_{1:\tau}^{\star,1}\|_2^2\Big)\\&\geq\frac{\mu_\gamma}{4}\big\|\boldsymbol{x}_{1:\tau}^{\star,1}-\boldsymbol{x}_{1:\tau}^{\star,0}\big\|_2^2.\end{aligned} \tag{71}$$

The second inequality follows from the identity $2\big(\|\boldsymbol{a}-\boldsymbol{p}\|_2^2+\|\boldsymbol{a}-\boldsymbol{q}\|_2^2\big)\geq\|\boldsymbol{p}-\boldsymbol{q}\|_2^2$.

By (68) and (69),

$$\sum_{t\leq\tau}\big(x_t^{\star,1}-x_t^{\star,0}\big)\geq\frac{3B_i}{8}\,.$$

Hence, Cauchy-Schwarz gives

$$\big\|\boldsymbol{x}_{1:\tau}^{\star,1}-\boldsymbol{x}_{1:\tau}^{\star,0}\big\|_2^2\geq\frac{9B_i^2}{64\tau}=\frac{9B_i^2}{32T}\,.$$

Substituting into (71), we conclude the proof of Theorem 4.1.

It remains to verify (68) - (70). To this end, we first recall the definition of the memory form

$$K_\gamma(\boldsymbol{x}):=\boldsymbol{x}^\top\mathrm{L}_\gamma\boldsymbol{x}=\sum_{s<t}\gamma^{t-s-1}x_sx_t\,.$$

Then, the self-generated part of the cost is

$$\mathcal{S}_\gamma(\boldsymbol{x}):=\kappa\|\boldsymbol{x}\|_2^2+\lambda K_\gamma(\boldsymbol{x})=\frac{1}{2}\boldsymbol{x}^\top\mathrm{H}_\gamma\boldsymbol{x}\,,$$

where $\mathrm{H}_\gamma=2\,\kappa\,\mathrm{I}_T+\lambda\,(\mathrm{L}_\gamma+\mathrm{U}_\gamma)$ is the Hessian of the individual cost. For $z=0$, the individual cost can be decomposed as

$$C_i(\boldsymbol{x};\boldsymbol{y})=\mathcal{S}_\gamma(\boldsymbol{x})+\langle\kappa\boldsymbol{y}+\lambda\mathrm{L}_\gamma\boldsymbol{y},\,\boldsymbol{x}\rangle\,,$$

We also recall that the stability condition implies $\mathrm{H}_\gamma \succeq 2\mu_\gamma \mathrm{I}_T$ and that the individual cost is $2\mu_\gamma$-strongly convex.

We start with (70), which follows directly from the $2\mu_\gamma$-strong convexity of the individual cost and the first-order optimality condition for the minimizer $\boldsymbol{x}^{\star,\omega}$ against $\boldsymbol{y}^{(\omega)}$ over the quota simplex.

To prove (68), we observe that against $\boldsymbol{y}^{(0)} = \mathbf{0}$, the individual cost is simply $\mathcal{S}_\gamma(\boldsymbol{x})$. Reversing a schedule does not change this objective, due to the symmetry of H_γ , and preserves the quota constraint. Since $\mathcal{S}_\gamma(\boldsymbol{x})$ is strongly convex, its minimizer $\boldsymbol{x}^{\star,0}$ against $\boldsymbol{y}^{(0)}$ is unique and symmetric w.r.t. the mid point of the horizon. Therefore, the first and second halves of $\boldsymbol{x}^{\star,0}$ contain equal total volume, which proves (68) (note that $x^{\star,0}$ may not be uniform within either half).

Finally, we establish (69). Consider the feasible half-uniform schedule $u_t = 2b\mathbb{1}\{t \le \tau\}$. We use $\boldsymbol{u}$ only as a comparator; it does not need to be optimal. We next calculate $C_i(\boldsymbol{u}; \boldsymbol{y}^{(1)})$.

Suppose first that $\gamma < 1$ and define

$$r_{\tau,\gamma} := \frac{1}{\tau}\sum_{d=1}^{\tau-1}(\tau - d)\gamma^{d-1} \le \frac{1}{1-\gamma}.$$

Since $\boldsymbol{u}$ and $\boldsymbol{y}^{(1)}$ have disjoint supports, $\boldsymbol{u}$ incurs no opponent-induced cost. Hence,

$$C_i(\boldsymbol{u}; \boldsymbol{y}^{(1)}) = \mathcal{S}_\gamma(\boldsymbol{u}) = \kappa\|\boldsymbol{u}\|_2^2 + \lambda K_\gamma(\boldsymbol{u})$$

Then, $\|\boldsymbol{u}\|_2^2 = 4b^2\tau = 2b^2T$. For the memory term, grouping pairs (s,t) according to their distance $d = t - s$ results in

$$K_\gamma(\boldsymbol{u}) = 4b^2\sum_{d=1}^{\tau-1}(\tau - d)\gamma^{d-1} = 2b^2T r_{\tau,\gamma}.$$

Consequently,

$$C_i(\boldsymbol{u}; \boldsymbol{y}^{(1)}) = 2b^2T\big(\kappa + \lambda r_{\tau,\gamma}\big). \tag{72}$$

For an arbitrary feasible schedule $\boldsymbol{x}$, let $R(\boldsymbol{x}) = \sum_{t>\tau} x_t$ be its second-half volume. Since $K_\gamma(\boldsymbol{x}) \ge 0$, $\|\boldsymbol{x}\|_2^2 \ge B_i^2/T = b^2T$, and every unit acquired in the second half pays at least the instantaneous cost $\kappa\, Y$ due to competing demand Y,

$$C_i(\boldsymbol{x}; \boldsymbol{y}^{(1)}) \ge \kappa b^2T + \kappa\, Y R(\boldsymbol{x}). \tag{73}$$

The optimal cost is no larger than the cost of $\boldsymbol{u}$. Comparing (72) and (73) therefore yields

$$\kappa\, Y R(\boldsymbol{x}^{\star,1}) \le b^2T\big(\kappa + 2\lambda r_{\tau,\gamma}\big).$$

We choose $Y = 8b(\kappa + 2\lambda r_{\tau,\gamma})/\kappa$. Then $R(\boldsymbol{x}^{\star,1}) \le bT/8 = B_i/8$, which proves (69).

At the endpoint $\gamma = 1$, since the quota B_i must be met by T, we have that:

$$\mathcal{S}_1(\boldsymbol{x}) = \frac{\lambda B_i^2}{2} + \mu_1\|\boldsymbol{x}\|_2^2.$$

The minimum squared norm on the quota simplex is $B_i^2/T = b^2T$, attained by the uniform schedule, whereas $\|\boldsymbol{u}\|_2^2 = 2b^2T$. Hence

$$\mathcal{S}_1(\boldsymbol{u}) - \min_{\boldsymbol{x}\in\mathcal{X}_i}\mathcal{S}_1(\boldsymbol{x}) = \mu_1 b^2T.$$

The self-generated cost of $\boldsymbol{u}$ exceeds the minimum possible self-generated cost by $\mu_1 b^2T$. Hence

$$\kappa\, Y R(\boldsymbol{x}^{\star,1}) \le \mu_1 b^2T.$$

Choosing $Y = 8\mu_1 b/\kappa$ again gives $R(\boldsymbol{x}^{\star,1}) \le B_i/8$ and proves (69).

For any fixed $\gamma \in (0,1]$, the total volume τY of the competing demand is $O(T)$, as required by Theorem 4.1, which completes the proof. □

Table 4: Notation for policy-regret learning used throughout Section G.

Symbol	Meaning
$\bar{x}, \bar{y}$	per-period bounds: $x_t \in [0, \bar{x}]$, $y_t \in [0, \bar{y}]$
m	memory of the aggregate response
τ, $K = T/\tau$	block length and number of blocks; block k starts just after $t_k = (k-1)\tau$
$\beta = B/T$	pace, used in every block prefix; standing requirement $\beta \le \bar{x}$
$\mathcal{R} = \{\pi_1, \dots, \pi_M\}$	finite policy library
J_k, $L_k(j)$	arm drawn in block k; cost of policy j in suffix of block k
$[a_\tau, a_\tau + R_\tau]$	a known interval containing every feasible suffix cost
S	cumulative suffix sensitivity to the entering state (Assumption G.2)
d_k, $\Xi_\gamma(d)$	state-gap and per-block coupling bounds, Lemma G.4
$J_T(\sigma)$, PReg_T	total cost of a complete policy and policy regret

G Policy-Regret Learning: Definitions and Proofs

This section contains the technical details supporting the claims of Section 5. The notation specific to this section is summarized in Table 4. We repeatedly use the standard geometric sum $h_r(\gamma) = \sum_{s=0}^{r-1} \gamma^s$ introduced in Appendix A, where $h_0(\gamma) = 0$ and $h_r(\gamma) = (1-\gamma^r)/(1-\gamma)$ for any fixed $\gamma < 1$. We recall that in this section only, we assume upper bounds $\bar{x}$ and $\bar{y}$ on the agent's and the competing demand in each period.

G.1 Information, Policies and the Counterfactual Benchmark

Information. We fix an agent and restrict our attention to its policy. The agent chooses $x_t \in [0, \bar{x}]$ at the beginning of period t. The environment then produces $y_t \in [0, \bar{y}]$, the public price is $P_t = I_t + \kappa(x_t + y_t)$, the agent pays $c_t = x_t P_t$, and observes only (x_t, P_t). In particular, the agent does not have any information about the impact parameters $(\kappa, \lambda, \gamma)$. Since $c_t = x_t P_t$, recording (x_t, P_t) is the same as recording (x_t, c_t), i.e., the agent's action and the corresponding cost, except when $x_t = 0$, in which case the public price P_t must be obtained otherwise.

The Response. We assume that the response maps ψ_t generating the aggregate response y_t have *memory* m. More specifically, the response maps ψ_t may vary with time and may be adversarial, but need to satisfy two restrictions. First, the response ψ_t at time t depends only on the current and previous m actions,

$$y_t = \psi_t\big(x_{\max\{t-m,1\}}, \dots, x_t\big)\,, \tag{74}$$

and carries no additional hidden state. Second, the whole sequence $(\psi_t)_{t \le T}$ of response maps is fixed before the episode starts, i.e., it is oblivious to the agent's random choices, as for the m-memory adversaries of Arora et al. (2012). The comparator is evaluated under the same maps. The second restriction makes our bandit approach valid, because it ensures that conditional on the history up to the end of block k's m-prefix, the entire vector $(L_k(1), \dots, L_k(M))$ of counterfactual suffix costs incurred by the policies in the agent's library $\mathcal{R}$ is fully determined. So, even though the environment reacts, through ψ_t, to the actions generated by the selected policy, it cannot adapt its response maps to the realized draws $J_1, J_2, \dots$. (If the maps of block k could depend on the draws in earlier blocks, the present conditional Exp3 argument would not apply.)

More specifically, in Algorithm 3 (a more detailed description of Algorithm 2), the common (for all blocks) m-period prefix removes the dependence of the responses on actions taken before the current block. After the prefix, all candidate policies face the same entering impact state, remaining quota and response-memory window. Once a candidate policy π_j is fixed, its actions, the responses generated by the (fixed in advance) maps ψ_t, the resulting prices, and hence its suffix cost $L_k(j)$ are fully determined. Consequently, conditional on the history up to the end of the prefix and before J_k is sampled, the vector $\big(L_k(1), \dots, L_k(M)\big)$ is fully determined.

Policies. Block k occupies periods $t_k + 1, \dots, t_k + \tau$ with $t_k = (k-1)\tau$. Its first m periods are the *prefix* and the remaining $\tau - m$ are the *suffix*, so we require $\tau > m$. So, the first suffix period of block k is $t_k + m + 1$. During the prefix, the agent follows the fixed pace β. The library policy is not active during these periods. At the beginning of the suffix, PB-Exp3 selects a policy $\pi_j \in \mathcal{R}$ and

Algorithm 3 Detailed description of Pace-Banded Block Exp3 (PB-Exp3) Algorithm

Require: Horizon $T = K\tau$, block length $\tau > m$, quota B, library $\mathcal{R} = \{\pi_1, \dots, \pi_M\}$ of scheduling policies, action bound $\bar{x} \geq B/T$, suffix-cost bounds $[a_\tau, a_\tau + R_\tau]$, and $\eta = \sqrt{2 \log M/(KM)}$ and $w_{1,j} \leftarrow 1$ for every $j \in [M]$

for $k = 1, \dots, K$ **do**
  $R \leftarrow \tau\beta$ ▷ quota remaining in block k
  for the first m periods of block k **do**
    Play $x_t \leftarrow \beta$, observe P_t and $c_t = x_t P_t$, and set $R \leftarrow R - x_t$
  end for
  Set $p_{k,j} \leftarrow w_{k,j} / \sum_{\ell=1}^{M} w_{k,\ell}$ and draw $J_k \sim \boldsymbol{p}_k$
  Initialize π_{J_k} with an empty within-block observation history and set $L_k \leftarrow 0$
  for each of the remaining $\tau - m$ periods **do**
    Let r be the number of periods remaining, including the current period
    Obtain the proposal $u_t \leftarrow \pi_{J_k}(r, R; (x_s, P_s)_{s \text{ in suffix}})$
    Set $x_t \leftarrow \mathsf{P}_{\mathcal{I}(R,r)}(u_t)$ and play x_t
    Observe P_t, incur $c_t = x_t P_t$, and update $L_k \leftarrow L_k + c_t$ and $R \leftarrow R - x_t$
  end for
  $\ell_k \leftarrow \min\{1, (L_k - a_\tau)_+ / R_\tau\}$ ▷ $\ell_k \in [0, 1]$
  for $j = 1, \dots, M$ **do**
    $w_{k+1,j} \leftarrow w_{k,j} \exp(-\eta\, \ell_k\, \mathbb{1}\{J_k = j\}/p_{k,j})$
  end for
end for

starts it from a *reset* internal state. At each suffix period t, the policy may use the remaining block quota, the number of periods left and the observations generated since the beginning of the suffix:

$$u_t = \pi_{j,r}\Big(R_t;\ (x_s, P_s)_{s=t_k+m+1}^{t-1}\Big), \tag{75}$$

where r is the number of periods remaining in the block and R_t is the remaining block quota. Thus the policy carries no memory from earlier blocks or from the common prefix. The executed action x_t is the projection of u_t onto the interval

$$\mathcal{I}(R, r) = \big[\max\{0, R - (r-1)\bar{x}\},\ \min\{\bar{x}, R\}\big],$$

which ensures block quota's completion and can be computed with information owned by the agent.

The Benchmark. We recall that for every $j \in [M]$, π_j^K denotes the *complete T-period policy* induced by K copies of π_j. In every block, π_j^K follows the uniform pace during the first m periods, initializes a new copy of π_j for the suffix, and applies the feasibility projection. For a complete T-period policy σ, let $J_T(\sigma)$ be its total cost when the response, the price and the impact state are regenerated from σ's own actions on the same initial state. Then

$$\mathrm{PReg}_T(\sigma) = \mathbb{E}\big[J_T(\sigma)\big] - \min_{j \in [M]} J_T(\pi_j^K).$$

We highlight that in the definition of the policy regret, the cost $J_T(\pi_j^K)$ of the comparator π_j^K is also computed along the trajectory it would have produced itself.

G.2 Exact Quota Completion

Lemma G.1 (Feasibility and Exact Completion)**.** *Suppose that $r \geq 1$ periods remain for the completion of a block and that the remaining quota R satisfies $0 \leq R \leq r\bar{x}$. Then the feasibility interval $\mathcal{I}(R, r)$ is non-empty and $x \in \mathcal{I}(R, r)$ leaves a feasible remaining quota $0 \leq R - x \leq (r-1)\bar{x}$. Consequently, if $\beta = B/T \leq \bar{x}$, PB-Exp3 completes every block quota $\tau\beta$ exactly.*

Proof. Because $R \leq r\bar{x}$, $R - (r-1)\bar{x} \leq \bar{x}$. Clearly, $R - (r-1)\bar{x} \leq R$. Hence the lower endpoint of $\mathcal{I}(R, r)$ does not exceed its upper endpoint and the interval $\mathcal{I}(R, r)$ is nonempty.

Now let $x \in \mathcal{I}(R, r)$. Since $x \leq R$, $R - x \geq 0$. Moreover, $x \geq R - (r-1)\bar{x}$ implies that $R - x \leq (r-1)\bar{x}$. Thus, after every feasible action, the same feasibility condition holds for the remaining $r - 1$ periods.

During the common prefix, if r periods remain then the remaining quota is $R = r\beta$. Since $\beta \le \bar{x}$, the uniform action $x = \beta$ belongs to $\mathcal{I}(r\beta, r)$. Hence the invariant is preserved through the prefix and, by the projection step, throughout the suffix.

Finally, when $r = 1$, $\mathcal{I}(R, 1) = \{R\}$, so the last action is exactly the remaining quota. Therefore every block quota, and hence the total quota B, is completed exactly. □

G.3 Feedback Stability

Finite feedback memory is an information restriction, not a stability property. The analysis of PB-Exp3 is based on a coupling argument that requires stability. The following states the stability condition used by the coupling argument.

Assumption G.2 (Feedback Stability)**.** *There exists a finite constant S, independent of the block length, the suffix length and the selected policy $j \in [M]$, such that the following holds:*

Consider two executions of the same suffix under the same policy π_j, the same response maps (74)*, and the same external randomness used by the response process and the environment. Suppose that, at the start of the suffix, the two executions have the same remaining quota, the same response-memory window (i.e., identical actions in the m periods of the prefix), and the same reset policy memories, but their impact states differ by d. Then,*

$$\sum_{t \text{ in the suffix}} |c_t - c'_t| \le S\,|d|.$$

In other words, even though a policy may react to the prices and the response maps may react to the policy, the interaction between the policy and the response maps does not amplify a small difference in the inherited impact state into an arbitrarily large difference in future costs.

Lemma G.3 (Sufficient Condition)**.** *Suppose that prices are uniformly bounded by $\bar{P}$, i.e., there exists a $\bar{P} < \infty$ such that $P_t \le \bar{P}$ along every possible trajectory. Assume further that there exist constants $A < \infty$ and $\theta < 1$ such that, if two suffixes start with impact-state difference d, then in the r-th suffix period,*

$$|I_t - I'_t| + |x_t - x'_t| + |y_t - y'_t| \le A\theta^r |d|, \quad \textit{for all } r \ge 0. \tag{76}$$

Then Assumption G.2 holds, with

$$S = \frac{A\big[\bar{P} + \bar{x}(1+\kappa)\big]}{1-\theta}.$$

Proof. Since $c_t = x_t P_t$, $|c_t - c'_t| \le \bar{P}|x_t - x'_t| + \bar{x}|P_t - P'_t|$. Moreover,

$$|P_t - P'_t| \le |I_t - I'_t| + \kappa|x_t - x'_t| + \kappa|y_t - y'_t| \le (1+\kappa)A\theta^r|d|.$$

Using also that $|x_t - x'_t| \le A\theta^r|d|$, we obtain that

$$|c_t - c'_t| \le \big[\bar{P} + \bar{x}(1+\kappa)\big]A\theta^r|d|.$$

Summing over the suffix and using $\sum_{r\ge 0}\theta^r = 1/(1-\theta)$ gives the stated bound. □

Condition (76) says that if two suffix trajectories differ initially only in their impact state by an amount d, then the resulting differences in the impact state, the agent's action, and the competing demand eventually all decay geometrically over time, at rate $\theta < 1$. The condition allows a bounded (and possibly substantial) initial amplification, controlled by A, before the initial state perturbation starts vanishing geometrically at rate $\theta < 1$. E.g., condition (76) holds whenever the one-step map sending the triple $(|\Delta I|, |\Delta x|, |\Delta y|)$ to its next value is a contraction in some norm.

A generalization of the sufficient condition in Lemma G.3 may consider geometric decay after $\ell \ge 1$ rounds. More formally, assume that there exist an integer $\ell \ge 1$ and constants $A < \infty$ and $\theta < 1$ such that, if two suffix trajectories start with impact state difference d, then

$$|I_t - I'_t| + |x_t - x'_t| + |y_t - y'_t| \le A\theta^{\lfloor r/\ell \rfloor}|d|\,, \quad \text{for all } r \ge 0,$$

where r is the number of periods elapsed since the beginning of the suffix. Then Assumption G.2 holds with

$$S = \frac{A\ell[\bar{P} + \bar{x}(1+\kappa)]}{1-\theta}.$$

We highlight that transient impact or restriction to a finite policy library $\mathcal{R}$ do not necessarily imply Assumption G.2. A policy may react discontinuously or amplify small differences in the inherited impact state. Hence, our analysis (and any analysis similar to ours) has to impose feedback stability separately.

G.4 One-Block Coupling

Lemma G.4 (One-Block Coupling). *Fix a block k and a policy $j \in [M]$, and consider two executions of policy π_j in this block. The first starts from PB-Exp3's realized history at the beginning of block k, while the second starts from the history generated by the comparator π_j^K. As in PB-Exp3, both executions use the common pace β on the m-prefix of k and π_j on the suffix. Let $d_k := |I_{s_k} - I'_{s_k}|$ be their impact-state difference at the beginning of the block, where s_k is the first period of block k.*

Under Assumption G.2, their block costs, $C_k^A(j)$ for the former and $C_k^{\mathrm{cf}}(j)$ for the latter, satisfy

$$|C_k^A(j) - C_k^{\mathrm{cf}}(j)| \le \Xi_\gamma(d_k),$$

where

$$\Xi_\gamma(d) := \beta \sum_{r=0}^{m-1} \left(\gamma^r d + \lambda\,\bar{y}\,h_r(\gamma) + \kappa\,\bar{y}\right) + S\left(\gamma^m d + \lambda\,\bar{y}\,h_m(\gamma)\right).$$

Proof. We follow (and compare) the two executions of π_j from the beginning of the block.

Cost Gap in the Common Prefix. Let $\delta_t := I_t - I'_t$. During the m prefix periods, both executions use the same action $x_t = x'_t = \beta$. Hence $\delta_{t+1} = \gamma\delta_t + \lambda(y_t - y'_t)$. Since $|y_t - y'_t| \le \bar{y}$, after r prefix periods, $|\delta_{s_k+r}| \le \gamma^r d_k + \lambda\,\bar{y}\,h_r(\gamma)$, for all $r = 0, \dots, m$ (where we use that $h_0(\gamma) = 0$).

At the $(r+1)$-th prefix period $t = s_k + r$, $r = 0, \dots, m-1$, the two executions both acquire β, so $|c_t - c'_t| = \beta|P_t - P'_t|$. Because their own actions coincide,

$$|P_t - P'_t| \le |\delta_t| + \kappa|y_t - y'_t| \le \gamma^r d_k + \lambda\,\bar{y}\,h_r(\gamma) + \kappa\,\bar{y}.$$

Therefore the total difference in prefix costs is at most

$$\beta \sum_{r=0}^{m-1} \left(\gamma^r d_k + \lambda\,\bar{y}\,h_r(\gamma) + \kappa\,\bar{y}\right). \tag{77}$$

Cost Gap in the Suffix. After the m common prefix actions, the two executions have the same remaining quota and the same m preceding actions. Moreover, according to the reset convention in (75), policy π_j is reset in both executions. Thus the only difference relevant to Assumption G.2 is the impact state.

By the analysis of the state difference in the prefix, the state difference at the beginning of the suffix is $|\delta_{s_k+m}| \le \gamma^m d_k + \lambda\,\bar{y}\,h_m(\gamma)$. Feedback stability therefore gives

$$\sum_{t \text{ in suffix}} |c_t - c'_t| \le S\left(\gamma^m d_k + \lambda\,\bar{y}\,h_m(\gamma)\right). \tag{78}$$

Adding (77) and (78) proves the result. □

G.5 The Bandit Step – Applying the Analysis of Exp3

We condition on the history at the end of block k's common prefix. For each policy $j \in [M]$, we let $L_k(j)$ be the suffix cost that would result from using π_j from that point onward. As explained in Section G.1, the vector $(L_k(1), \dots, L_k(M))$ of counterfactual suffix costs incurred by the policies in $\mathcal{R}$ is fully determined before PB-Exp3 randomly selects J_k. Thus, we can regard each block as an ordinary adaptive bandit round for Exp3.

Lemma G.5 (Exp3 over blocks)**.** *Suppose that every suffix cost lies in* $[a_\tau, a_\tau + R_\tau]$*. With step size* $\eta = \sqrt{2\log(M)/KM}$*, PB-Exp3 satisfies, for every fixed* $j \in [M]$*,*

$$\mathbb{E}\left[\sum_{k=1}^{K}\big[L_k(J_k) - L_k(j)\big]\right] \le R_\tau\sqrt{2KM\log M}.$$

Proof. We first divide by R_τ in order to normalize the suffix losses:

$$\ell_k(j) := \frac{L_k(j) - a_\tau}{R_\tau} \in [0,1].$$

If $p_{k,j}$ is the probability assigned to policy j, Exp3 uses the importance-weighted estimate

$$\widehat{\ell}_{k,j} = \frac{\ell_k(J_k)\mathbb{1}\{J_k = j\}}{p_{k,j}}.$$

Conditional on the history before the random selection of J_k, $\mathbb{E}[\widehat{\ell}_{k,j}] = \ell_k(j)$, i.e., our suffix loss estimator is unbiased. Moreover, using the fact that $\ell_k(j) \in [0,1]$,

$$\sum_{j=1}^{M} p_{k,j}\mathbb{E}\left[\widehat{\ell}_{k,j}^{\,2}\right] = \sum_{j=1}^{M}\ell_k(j)^2 \le M.$$

Therefore, the standard potential argument in the analysis of Exp3 (see e.g, (Auer et al., 2002)) gives

$$\mathbb{E}\left[\sum_{k=1}^{K}\big[\ell_k(J_k) - \ell_k(j)\big]\right] \le \frac{\log M}{\eta} + \frac{\eta\, K\, M}{2}.$$

With the stated value of η, the right-hand side becomes $\sqrt{2KM\log M}$. Multiplying by R_τ concludes the proof of the lemma. □

G.6 Putting Everything Together: The Proof of Theorem 5.1

Proof of Theorem 5.1. Exact quota completion follows directly from Lemma G.1. For the rest of the proof, fix a comparator policy $j^\star \in \arg\min_{j\in[M]} J_T(\pi_j^K)$.

Exp3 regret on the realized block histories. For a block k, let C_k^{pre} denote the cost on its prefix, where the schedule is uniform and does not depend on the policy, and let $L_k(j)$ be the cost on its suffix that would result from running π_j after that realized prefix. Since the prefix is the same for every policy, the total cost $C_k^A(j)$ of an execution of π_j on block k which starts from PB-Exp3's realized history at the beginning of k, is $C_k^A(j) = C_k^{\mathrm{pre}} + L_k(j)$. Then, the realized cost of PB-Exp3 is:

$$J_T(\text{PB-Exp3}) = \sum_{k=1}^{K}\big[C_k^{\mathrm{pre}} + L_k(J_k)\big].$$

By Lemma G.5,

$$\mathbb{E}\left[J_T(\text{PB-Exp3}) - \sum_{k=1}^{K} C_k^A(j^\star)\right] \le R_\tau\sqrt{2KM\log M}. \tag{79}$$

However, this comparison concerns the execution of $\pi_{j^\star}$ from PB-Exp3's realized history, based on the policies chosen by PB-Exp3, at the beginning of each block k.

Policy regret on the comparator's own trajectory. Let $d_k := |I_{s_k} - I'_{s_k}|$ be the impact state difference between the realized and comparator trajectories at the beginning of each block k. Let $C_k^{\mathrm{cf}}(j^\star)$ denote the cost incurred in block k when a reset copy of $\pi_{j^\star}$ is run from the history generated by the complete comparator policy $\pi_{j^\star}^K$. Then Lemma G.4 gives $\big|C_k^A(j^\star) - C_k^{\mathrm{cf}}(j^\star)\big| \le \Xi_\gamma(d_k)$.

Both complete trajectories start from the same initial impact state. Since their actions and responses satisfy $0 \le x_t, x'_t \le \bar{x}$ and $0 \le y_t, y'_t \le \bar{y}$, for any fixed $\gamma < 1$, their state difference is uniformly bounded by

$$d_k \le \frac{\lambda(\bar{x} + \bar{y})}{1-\gamma} =: \bar{d}_\gamma.$$

Using that $\Xi_\gamma(d)$ is increasing in d, we define $D_\gamma := \Xi_\gamma(\bar{d}_\gamma)$. Then, $|C_k^A(j^\star) - C_k^{\mathrm{cf}}(j^\star)| \le D_\gamma$, for every block k, which implies that:

$$\sum_{k=1}^{K} C_k^A(j^\star) \le J_T(\pi_{j^\star}^K) + KD_\gamma \,. \tag{80}$$

Combining (79) and (80) gives

$$\mathrm{PReg}_T(\text{PB-Exp3}) \le R_\tau \sqrt{2KM \log M} + KD_\gamma \,. \tag{81}$$

Regret bound in terms of T and τ. By assumption, $0 \le c_t \le \bar{c}$. The suffix of each block contains $\tau - m$ periods, so $R_\tau \le \tau\bar{c}$. Since $K = T/\tau$, (81) gives:

$$\begin{aligned}\mathrm{PReg}_T(\text{PB-Exp3}) &\le \tau\bar{c}\sqrt{2(T/\tau)M \log M} + D_\gamma(T/\tau) \\ &= \bar{c}\sqrt{2T\tau M \log M} + D_\gamma T/\tau.\end{aligned}$$

Selecting $\tau = \max\{m+1, \lceil T^{1/3} \rceil\}$ (which does not assume any knowledge of the environment parameters) gives, for every fixed m, M, $\bar{c}$ and $\gamma < 1$, $\mathrm{PReg}_T(\text{PB-Exp3}) = O(T^{2/3})$. □

Remark G.6. The analysis above applies only to the case of transient impact, where $\gamma < 1$. The proof of Theorem 5.1 relies on a block-coupling constant Ξ_γ, which is independent of the horizon length. For $\gamma < 1$, differences in the inherited impact state decay geometrically, which makes such a bound possible. At $\gamma = 1$, state differences do not decay, so the present coupling argument does not provide a horizon-independent per-block bound and therefore does not imply sublinear policy regret.

G.7 Policy Regret under Price and State Noise

We next consider a noisy version of our model, with additive noise in both the observed price and the impact state, identical to that in Appendix E.5. More specifically, we let

$$P_t = I_t + \kappa(x_t + y_t) + \varepsilon_t \quad \text{and} \quad I_{t+1} = \gamma I_t + \lambda(x_t + y_t) + \xi_{t+1} \,,$$

where $\{\varepsilon_t\}$ and $\{\xi_t\}$ are mutually independent zero-mean white noise processes with variance σ_ε^2 and σ_ξ^2, respectively.

When comparing PB-Exp3 with a counterfactual policy, we use the same realization of (ε_t, ξ_t) in both trajectories. Thus the only difference between the two trajectories comes from their actions and from the responses generated by those actions.

The one-block coupling argument in Lemma G.4 holds unchanged. Consider the two executions in Lemma G.4, and let $\delta_t = I_t - I'_t$. Subtracting their state equations gives

$$\begin{aligned}\delta_{t+1} &= \gamma\delta_t + \lambda\big[(x_t - x'_t) + (y_t - y'_t)\big] + \xi_{t+1} - \xi_{t+1} \\ &= \gamma\delta_t + \lambda\big[(x_t - x'_t) + (y_t - y'_t)\big].\end{aligned}$$

Hence the state-difference recursion is exactly the same as in the noise-free model.

Similarly, for the price difference,

$$\begin{aligned}P_t - P'_t &= \delta_t + \kappa\big[(x_t - x'_t) + (y_t - y'_t)\big] + \varepsilon_t - \varepsilon_t \\ &= \delta_t + \kappa\big[(x_t - x'_t) + (y_t - y'_t)\big].\end{aligned}$$

Therefore the prefix calculation in Lemma G.4 is unchanged.

If Assumption G.2 holds for all possible noise realizations under consideration, the suffix cost calculation in the proof of Lemma G.4 is unchanged as well. Consequently, the same one-block bound holds: $|C_k^A(j) - C_k^{\mathrm{cf}}(j)| \le \Xi_\gamma(d_k)$. In particular, the uniform coupling constant D_γ used in Theorem 5.1 is the same as in the model without noise.

Thus noise affects the policy-regret bound only through the range of the losses supplied to Exp3.

Bounded noise assumption. Suppose $|\varepsilon_t| \le u_\varepsilon$ and $|\xi_t| \le u_\xi$ for every period t. Since $0 \le x_t \le \bar{x}$ and $0 \le y_t \le \bar{y}$, for $\gamma < 1$, unrolling the state recursion gives

$$|I_t| \le |I_1| + \frac{\lambda(\bar{x}+\bar{y}) + u_\xi}{1-\gamma}.$$

Hence

$$|P_t| \le |I_1| + \frac{\lambda(\bar{x}+\bar{y}) + u_\xi}{1-\gamma} + \kappa(\bar{x}+\bar{y}) + u_\varepsilon.$$

Since $c_t = x_t P_t$ and $x_t \le \bar{x}$, for any period t,

$$|c_t| \le c_{\mathrm{bd}} := \bar{x}\left[|I_1| + \frac{\lambda(\bar{x}+\bar{y}) + u_\xi}{1-\gamma} + \kappa(\bar{x}+\bar{y}) + u_\varepsilon\right].$$

Therefore, the learner's cost in any period is at most c_{bd} and the updated $R_\tau \le \tau c_{\mathrm{bd}}$.

Substituting $2c_{\mathrm{bd}}$, instead of $\bar{c}$ in the proof of Theorem 5.1 shows that bounded price and state noise leave $\mathrm{PReg}_T(\text{PB-Exp3}) = O(T^{2/3})$, for any fixed $\gamma < 1$ and fixed noise bounds.

G.8 Convergence to Coarse Correlated Equilibrium of the Policy Game

We now suppose that all N agents learn simultaneously. Agent i has a fixed finite policy library $\mathcal{R}_i = \{\pi_{i,1}, \dots, \pi_{i,M_i}\}$. All agents use the same block boundaries and the same block length τ. In every block they draw their policies simultaneously and independently, conditional on the past. Quotas, policy libraries, and the within-block dynamics are the same in every block.

We aim to show that the expected empirical distribution of the selected policies converges to a coarse correlated equilibrium of the induced *canonical finite policy game*.

The Canonical Policy Game. We fix a reference entering impact state z°. For a joint policy profile

$$\boldsymbol{j} = (j_1, \dots, j_N) \in [M_1] \times \cdots \times [M_N],$$

we run one block from z° with policy π_{i,j_i} for every agent i and let $C_i^\circ(\boldsymbol{j})$ denote agent i's expected total cost in this block. We define i's cost in the canonical policy game $\mathcal{G}_\tau$ as the average cost per period, $f_i(\boldsymbol{j}) := C_i^\circ(\boldsymbol{j})/\tau$. Thus $\mathcal{G}_\tau$ is an ordinary finite normal-form game whose pure actions are the policies in the agents' libraries $\mathcal{R}_i$.

Joint feedback stability. There is no exogenous response map in this section. The policies of the other agents may react to prices, and hence to the state, so Assumption G.2, which fixes the response maps, does not apply. Instead, we assume the following joint condition.

Assumption G.7 (Joint Feedback Stability)**.** *For every agent i there exists a finite constant S_i, independent of the block length, the suffix length and the joint policy profile, such that the following holds for every joint policy profile $\boldsymbol{j}$ and every block:*

Consider two executions of the suffix under $\boldsymbol{j}$, with the same external randomness, the same remaining quotas and the same reset policy memories, whose impact states at the start of the suffix differ by d. Then,

$$\sum_{t \text{ in the suffix}} |c_{i,t} - c'_{i,t}| \le S_i |d|.$$

Coupling an actual block to the corresponding block of the canonical game. Let z_k be the actual state entering block k, and let $C_{i,k}(\boldsymbol{j})$ denote the expected total cost of agent i, if the joint policy profile $\boldsymbol{j}$ is used in that block starting from z_k. For convenience, we refer to a policy's execution on k starting from its actual entering state z_k (resp. reference initial state z°) as an *actual block* (resp. a *block of the canonical game*).

The feedback-stability argument of Lemma G.4 implies that under Assumption G.7, there is a finite constant $D_{i,\gamma}$, independent of k, τ, T and $\boldsymbol{j}$, such that

$$\left|C_{i,k}(\boldsymbol{j}) - C_i^\circ(\boldsymbol{j})\right| \le D_{i,\gamma}. \tag{82}$$

To establish this claim, we compare two executions of the same joint policy profile, one starting from z_k and one from z°. During the common prefix all agents take the same actions in both executions,

so the state difference (between the state trajectories initialized by z_k and z°) decays geometrically. At the beginning of the suffix, feedback stability bounds the remaining effect of this state difference on future costs. As in the proof of Lemma G.4, because $\gamma < 1$ and reachable states are bounded, the resulting total difference is bounded by a constant independent of the block length.

For example, if the common prefix has length m and $\bar{d} := \sup_k |z_k - z^\circ| < \infty$, then one may take $D_{i,\gamma} = \beta_i \bar{d}\, h_m(\gamma) + S_i \gamma^m \bar{d}$, where β_i is agent i's prefix pace and S_i is its feedback-stability constant. The exact value of $D_{i,\gamma}$ is not important below, only its independence of T and τ is used. We also note that without exogenous responses, the prefix is not needed and we may set $m = 0$. Then $D_{i,\gamma} = S_i \max\{z^\circ, z_1 + \lambda \sum_j \bar{x}_j/(1-\gamma)\}$.

Proof of Theorem 5.2. Let $\boldsymbol{J}_k = (J_{1,k}, \dots, J_{N,k})$ be the joint policy profile drawn in block k. Its random empirical distribution after K blocks is

$$\widehat{\nu}_K(\boldsymbol{j}) := \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\{\boldsymbol{J}_k = \boldsymbol{j}\}, \qquad \text{where } \boldsymbol{j} \in [M_1] \times \cdots \times [M_N].$$

More formally, $\widehat{\nu}_K$ is the random empirical measure of the realized joint policy profiles. Its expectation is the deterministic distribution

$$\mathbb{E}\big[\widehat{\nu}_K\big](\boldsymbol{j}) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{P}(\boldsymbol{J}_k = \boldsymbol{j}),$$

which we refer to as the *expected empirical distribution* and use in the proof.

For the proof, we fix an arbitrary agent i and a possible deviation $j_i' \in [M_i]$ from i's chosen policy.

Exp3's external regret across the actual blocks. We condition on the history before the policy draws in block k and on the policies drawn by the other agents. Because, conditional on the above, the policy selections are independent, as explained in Appendix G.1, every alternative policy $j_i \in [M_i]$ has a well-defined expected block cost before agent i draws $J_{i,k}$. Therefore the standard Exp3 applies to the expected cost of agent i.

If $R_{i,\tau}$ is a range containing all possible expected block costs of agent i, Lemma G.5 gives

$$\mathbb{E}\left[\sum_{k=1}^{K} [C_{i,k}(\boldsymbol{J}_k) - C_{i,k}(j_i', \boldsymbol{J}_{-i,k})]\right] \le R_{i,\tau} \sqrt{2KM_i \log M_i}. \tag{83}$$

Replace actual blocks by blocks of the canonical game. By (82), $C_i^\circ(\boldsymbol{J}_k) \le C_{i,k}(\boldsymbol{J}_k) + D_{i,\gamma}$ and $C_{i,k}(j_i', \boldsymbol{J}_{-i,k}) \le C_i^\circ(j_i', \boldsymbol{J}_{-i,k}) + D_{i,\gamma}$. Using these two inequalities in (83) gives

$$\mathbb{E}\left[\sum_{k=1}^{K} [C_i^\circ(\boldsymbol{J}_k) - C_i^\circ(j_i', \boldsymbol{J}_{-i,k})]\right] \le R_{i,\tau} \sqrt{2KM_i \log M_i} + 2KD_{i,\gamma}. \tag{84}$$

Convert regret into the CCE inequality. We divide (84) by $K\tau$. Since $f_i = C_i^\circ/\tau$,

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[f_i(\boldsymbol{J}_k) - f_i(j_i', \boldsymbol{J}_{-i,k})] \le \frac{R_{i,\tau}}{\tau} \sqrt{\frac{2M_i \log M_i}{K}} + \frac{2D_{i,\gamma}}{\tau}\,. \tag{85}$$

For every fixed function g on the finite joint-policy space, the definition of the empirical distribution $\widehat{\nu}_K$ and linearity of expectation give

$$\mathbb{E}_{\boldsymbol{J} \sim \mathbb{E}[\widehat{\nu}_K]}\big[g(\boldsymbol{J})\big] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\big[g(\boldsymbol{J}_k)\big].$$

We note that this equality does not require the policy draws to be independent or identically distributed across blocks.

Hence (85) becomes

$$\mathbb{E}_{\boldsymbol{J} \sim \mathbb{E}[\widehat{\nu}_K]}\big[f_i(\boldsymbol{J})\big] \le \mathbb{E}_{\boldsymbol{J} \sim \mathbb{E}[\widehat{\nu}_K]}\big[f_i(j_i', \boldsymbol{J}_{-i})\big] + \varepsilon_{i,K},$$

where

$$\varepsilon_{i,K} := \frac{R_{i,\tau}}{\tau}\sqrt{\frac{2M_i \log M_i}{K}} + \frac{2D_{i,\gamma}}{\tau}. \tag{86}$$

Since j_i' is an arbitrary possible deviation of agent i, this is precisely the coarse-correlated-equilibrium inequality for agent i.

Theorem 5.2 assumes that $0 \le c_{i,t} \le \bar{c}_i$. Since a block contains at most τ periods, $R_{i,\tau} \le \tau \bar{c}_i$.

Using $K = T/\tau$ in (86) gives

$$\varepsilon_{i,K} \le \bar{c}_i\sqrt{\frac{2M_i \log M_i}{K}} + \frac{2D_{i,\gamma}K}{T},$$

which is (11). Choosing $K = \Theta(T^{2/3})$ bounds both terms in (86) by $O(T^{-1/3})$. Therefore $\varepsilon_{i,K} = O(T^{-1/3})$ for any fixed policy libraries and fixed problem parameters. □

H Experimental Details and Additional Results

This appendix gives more details behind Section 6 along with additional experiments. See table 5 for an overview of the motivation behind our experiments.

For experiments with a single agent, the simulator uses

$$P_t = I_t + \kappa(x_t + y_t), \qquad I_{t+1} = \gamma I_t + \lambda(x_t + y_t), \qquad c_t = x_t P_t,$$

where y_t is aggregate competing demand. With multiple learning agents, $x_t + y_t$ is replaced by $q_t = \sum_i x_{i,t}$. Unless a sweep is specified, we use $(\kappa, \lambda, \gamma) = (1, 0.6, 0.7)$. Results such as best responses, and social optima are computed independently from their closed-form optimality conditions. Stochastic curves average 24 independent runs and use pointwise 95% Student-t intervals.

H.1 Uniform execution and equilibrium structure

Learning a best response: We fix $N = 3$, $T = 12$, $z = 0$, and unit quotas. Two agents use the uniform policy in every episode, while the third starts uniformly and runs Algorithm 1. Figure 3 compares its schedule after 20,000 episodes with the independently computed best response. We note that the curves overlap: the final schedule reduces the cost of one execution from 0.613 to 0.578, a saving of 5.79%, and shows a remaining best-response gain of 6.77×10^{-15}.

Horizon sweep: We next solve the two-agent game directly with unit quotas, $z = 0$, $(\kappa, \lambda) = (1, 0.6)$, $T \in \{20, 40, 80, 160, 320, 640\}$, and $\gamma \in \{0.3, 0.7, 0.95, 1\}$.

For each (T, γ), a solver computes the symmetric Nash equilibrium, the social optimum, and the best response to uniform. These systems follow from the constraints and cost gradient introduced in the main text, and there is no learning.

Figure 4 reports the price of anarchy and the exploitability of the uniform policy. For every fixed $\gamma < 1$, both values decrease with the horizon. At $\gamma = 0.7$, observations over $T \in \{160, 320, 640\}$ give slopes -0.99 for PoA -1 and -1.96 for uniform exploitability, which shows consistency with

Table 5: Experiments retained in the appendix and the theoretical point each is meant to examine.

Experiment	Quantity measured	Main point
Uniform and structural sweeps	Best response, PoA, exploitability	Transient-impact distortions decrease with T
Repeated-episode learning	Nash/BNE error and regret	Price gradients recover the predicted learning rates
PB-Exp3	Policy regret and sensitivity	Regret is sublinear when the block coupling is stable
Simultaneous PB-Exp3	Policy-game CCE gap	Average joint play improves over random rule selection

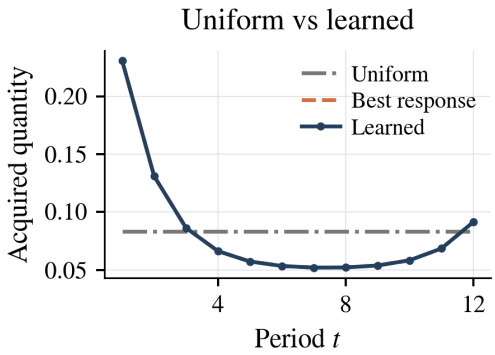


Figure 3: The uniform initialization, the final price-gradient schedule, and the independently computed best response against two uniform competitors.

the $O(T^{-1})$ and $O(T^{-2})$ predictions. At $\gamma = 1$ and $T = 640$, the corresponding values remain 1.3285 and 0.2669. The $\gamma = 0.95$ curve also shows that a transient model can look nearly permanent until $T(1-\gamma)$ becomes large.

H.2 Bayesian learning across repeated episodes

For Figure 1(a), $N = 3$, $T = 12$, and the state is reset to zero before every episode. Each agent independently receives quota one or two with probability $1/2$, observes its own quota and the public price path, and maintains one schedule per quota, initialized as $\pi_i^1(b) = (b/T)\mathbf{1}$.

Let $n_{i,b}^{k-1}$ count how often agent i has received quota b before episode k. If $B_i^k = b$, after completing the episode it applies the local-clock update

$$\pi_i^{k+1}(b) = \mathsf{P}_{\mathcal{X}_i(b)}\left(\pi_i^k(b) - \frac{\boldsymbol{g}_i^k}{\mu_\gamma\left(n_{i,b}^{k-1}+1\right)}\right), \tag{87}$$

where $\boldsymbol{g}_i^k$ is computed from the episode's public prices and the agent's executed schedule as in Algorithm 1. The denominator uses the quota's occurrence count, including the current episode (i.e., how many times did it occur since the beginning of the game). The other schedule is unchanged, and all agents update after the same completed episode. The joint quota distribution is used to generate quotas and to compute the reference BNE and the probability-weighted error $\|\Pi^K - \Pi^\star\|_{\mathcal{D}}^2$.

Price noise adds independent Gaussian innovations of standard deviation 0.2 to P_t, while state noise adds innovations of standard deviation 0.05 to I_{t+1}. Table 6 also includes correlated quotas: with

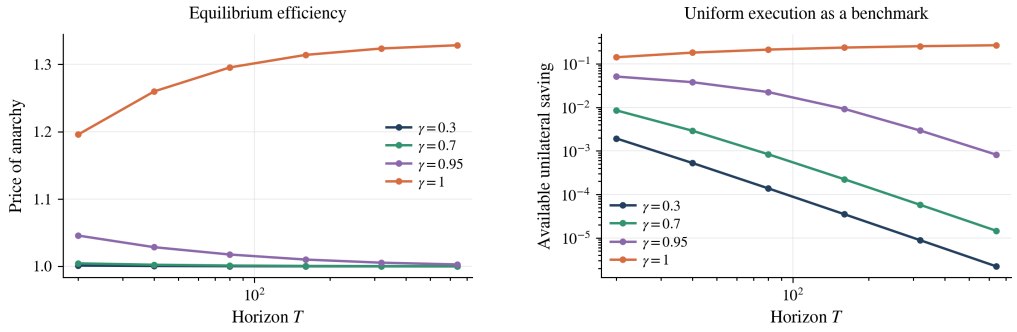


Figure 4: Price of anarchy and uniform-policy exploitability for two fixed unit quotas. The permanent endpoint remains separated from every fixed transient value as the horizon grows.

Table 6: Bayesian learning with the local-clock update at recorded iterate $K = 20{,}000$. Errors are means with 95% Student-t interval half-widths over 24 runs. Slopes fit log mean error against log K using checkpoints with $K \geq 1{,}000$. The first three rows use independent equiprobable quotas while the remaining rows use the correlated distribution described in the text.

Quotas	Noise	$\|\Pi^K - \Pi^\star\|_{\mathcal{D}}^2$	slope
independent	none	$1.07\times10^{-6} \pm 2.46\times10^{-7}$	-0.96
independent	price	$5.84\times10^{-5} \pm 4.78\times10^{-6}$	-1.01
independent	state	$4.38\times10^{-6} \pm 5.22\times10^{-7}$	-0.97
correlated	none	$4.73\times10^{-7} \pm 1.18\times10^{-7}$	-0.97
correlated	price	$5.60\times10^{-5} \pm 5.54\times10^{-6}$	-0.98
correlated	state	$3.63\times10^{-6} \pm 4.12\times10^{-7}$	-0.96

probability $1/2$, the indicators $B_i - 1$ are independent Bernoulli variables of mean 0.3, and otherwise all agents share one such draw. The joint quota vector is redrawn independently in every episode. Across both quota distributions, the fitted exponents lie between -1.01 and -0.96, consistent with the $O(K^{-1})$ mean-square guarantee for local-clock learning. We note that noise mainly changes the level of the error.

H.3 Policy regret within one order

We use $z = 0$, $B/T = 0.75$, and $0 \leq x_t, y_t \leq 1.5$, with

$$\tau \in \{12, 16, 20, 24, 30, 40, 50, 60, 80\}, \qquad T = \tau^3.$$

Every block begins with one uniform purchase. PB-Exp3 then samples one of six rules, resets that rule, and applies it to the remaining periods. Every proposal is projected onto

$$[\max\{0, R - (r-1)1.5\},\ \min\{1.5, R\}],$$

where R is the remaining block quota and r the number of periods left. Thus every block finishes its quota exactly.

The first library consists of uniform, front-loaded, back-loaded, alternating, contrarian, and momentum rules, all using only their own action history. The competing response is $y_t = \mathsf{P}_{[0,1.5]}(0.3 + 0.5x_t + 0.3x_{t-1})$. The second library replaces the last two rules by a pair that, in each three-period group, buys 0.75, observes the price P, and then buys $0.75 + u$ and $0.75 - u$, where $u = \pm 0.2\tanh((P - 2.5)/2)$. Its response is $y_t = 0.3 + 0.15\sin(2\pi(t-1)/12) + 0.3x_t$.

For every rule j, a separate complete-order simulation uses j in every block and regenerates demand, prices, and impact from that rule's own actions. The comparator is therefore the counterfactual policy benchmark from Section 5. Figure 5 reports total and per-period regret. The fitted exponents of total regret are 0.692 and 0.707 for the two libraries while average regret falls from 0.0851 to 0.0154 and from 0.0804 to 0.0170. Both are sublinear and close to the $T^{2/3}$ upper bound over the tested range.

Uniform is the best fixed rule in both libraries, so finite-horizon savings relative to uniform are negative. Figure 6 makes this distinction explicit. We remark that a valid but unnecessarily wide loss interval produces almost linear fitted regret (exponents 1.010 and 1.023), whereas the reachable-state interval gives the sublinear curves above.

Feedback sensitivity: We test whether a small initial impact difference has a bounded cumulative effect on costs, using one fixed price-responsive rule without learning. Every three periods, the rule buys 0.75 units and observes the resulting price P. It computes $u = 0.2\tanh((P - 2.5)/2)$, then buys $0.75 + u$ and $0.75 - u$ in the next two periods. This shifts at most 0.2 units between those purchases while preserving the group's total of 2.25.

With $(\kappa, \lambda) = (1, 0.6)$, we execute this rule from impact states 1 and 1.001, recalculating competing demand $y_t = 0.3 + 0.3x_t$ and prices along each execution. Figure 7 plots $\sum_{t=1}^{L} |c_t - c'_t|/0.001$ against execution length L. Sensitivity levels off for $\gamma \in \{0.3, 0.6, 0.9\}$ but continues growing at $\gamma = 1$. For this rule and response, the impact difference contracts over each three-period group by a factor at most $\gamma^3 + 0.078(1 - \gamma) < 1$ when $\gamma < 1$, supporting the required feedback stability.

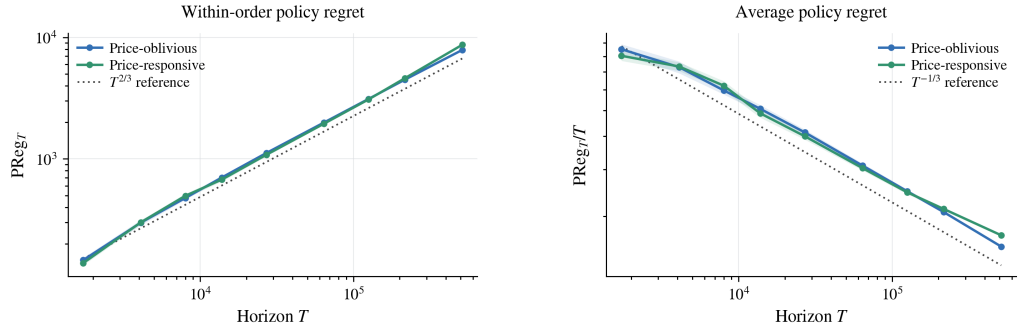


Figure 5: Total policy regret and regret per period. Bands are pointwise 95% intervals over 24 runs; dotted lines have slopes $2/3$ and $-1/3$.

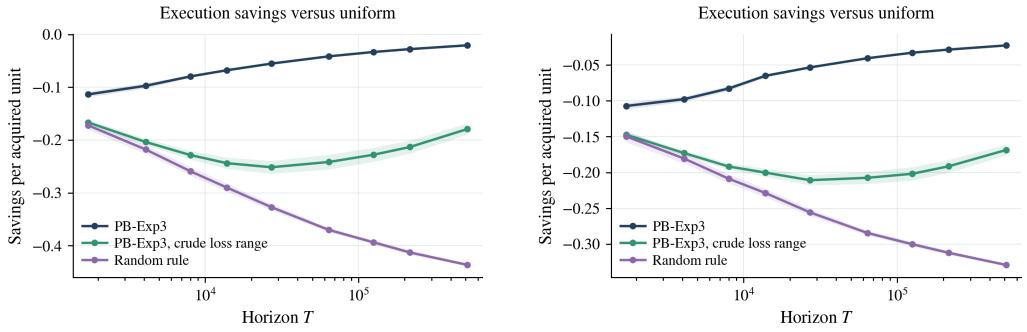


Figure 6: Complete-order savings per acquired unit relative to uniform execution. Negative values are excess cost; uniform is the best library rule in both environments.

H.4 Simultaneous policy learning

Three agents use common blocks, quota 0.75τ per block, purchase cap 1.5, and four price-oblivious rules: uniform, front-loaded, back-loaded, and alternating. For each of the 4^3 joint rule profiles, we compute one canonical block from the reference state

$$z^\circ = \frac{\lambda N(0.75)}{1-\gamma} = 4.5.$$

Let $\bar{\nu}_K$ be the joint-policy frequencies pooled over the 24 runs. The reported per-period CCE gap is

$$G(\bar{\nu}_K) = \max\left\{0, \max_{i,r} \sum_{\boldsymbol{j}} \bar{\nu}_K(\boldsymbol{j})[f_i(\boldsymbol{j}) - f_i(r, \boldsymbol{j}_{-i})]\right\}.$$

Pooling before taking the maximum estimates the gap of the expected empirical distribution in Theorem 5.2.

The fixed-game experiment in Figure 1(c) uses $\tau = 9$. With blockwise resets, the gap decreases from 0.0994 to 0.00974 and has tail slope -0.51. With inherited impact it decreases from 0.1186 to 0.0341, but more slowly. The inherited suffix-loss interval is about 3.17 times wider because it must cover all reachable entering states, and the theoretical comparison retains an $O(1/\tau)$ coupling term at fixed τ.

Figure 8 reports the complementary growing-block experiment: $\tau \in \{12, 20, 30, 40, 60, 90\}$, $K = \tau^2$, and $T = \tau^3$, with inherited impact. The absolute CCE gap falls to 0.0529 and has tail slope -0.264. The displayed relative gap divides by the gap of uniformly random rule selection in the corresponding game and falls from 0.891 to 0.175, with tail slope -0.335. Because the random baseline itself grows with τ, this relative slope should not be read as an estimate of the absolute CCE rate.

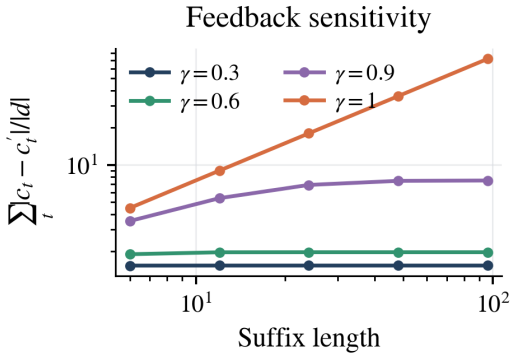


Figure 7: Cumulative cost sensitivity to a 0.001 perturbation of the entering state. The transient trajectories stabilize while the permanent endpoint does not.

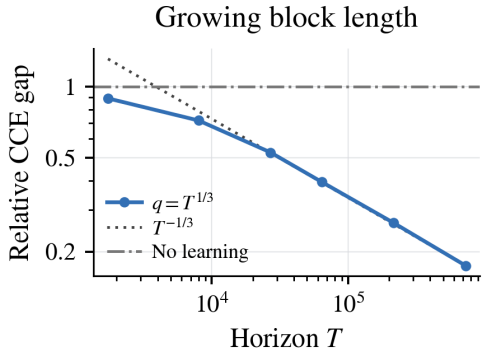


Figure 8: Growing blocks with inherited impact. The plotted quantity is the absolute CCE gap divided by the uniformly random-rule gap in the same policy game; the absolute tail decreases more slowly, as explained in the text.

I Financial Market Illustration

We apply Algorithm 1 to learn a reusable one-minute schedule per stock: we estimate the impact parameters from morning trades, hold them fixed while learning purchase quantities on later morning orders, then evaluate on afternoon prices. Following Hossain et al. (2026), historical quotes supply the background and the simulator adds hypothetical impact. A second experiment prescribes stronger impact to examine learning when timing benefits are larger.

Data and orders. We use LOBSTER's public Nasdaq samples for AAPL, AMZN, GOOG, INTC and MSFT on 21 June 2012. Level-one quotes give the best bid and ask: the highest quoted buying price and lowest selling price. The chronological split is 10:00–12:00 for training, 12:00–13:00 reserved for validation, and 13:00–15:30 for testing, in New York time. After filtering, these contain 600, 300 and 746 one-minute stock-windows. We call a "stock-window" one minute for one stock. Test observations are excluded from fitting and training.

Each parent order buys a total of B shares in $T = 12$ five-second periods. Let V_{60} be a stock's mean executed shares per training minute. We set $B = 0.10V_{60}$ for fitted impact and $B = 0.01V_{60}$ for the stronger scenarios. Each stock's order size stays fixed during testing. One simulated competitor buys the same quantity uniformly, at $y_t = B/T$ per period.

Estimating impact from trades and prices. Each stock has its own fitted triple $(\widehat{\kappa}, \widehat{\lambda}, \widehat{\gamma})$: immediate impact, the increment to persistent impact, and the fraction of that impact surviving five seconds. Let m_t be the last valid bid-ask midpoint before a five-second interval ends, and s the training median

of half the ask-bid difference. Let q_t be buyer-initiated shares minus seller-initiated shares in that interval. Positive q_t means more shares bought by taking standing sell orders than sold by taking standing buy orders. We fit

$$\frac{m_t - m_{t-1}}{s} = \kappa q_t + (\lambda - \kappa) q_{t-1} - \lambda(1-\gamma) \sum_{\ell=2}^{24} \gamma^{\ell-2} q_{t-\ell} + \varepsilon_t,$$

where ε_t is the residual. The two-minute lookback is fixed at 24 lags, giving 25 volume inputs including the current interval. We label a comparison "Free lags", which fits a separate unrestricted coefficient b_ℓ for each $q_{t-\ell}$ by ordinary least squares, using the same observations.

For each stock, bounded nonlinear least squares jointly fits κ, λ, γ on 1,416 usable training intervals, imposing $\kappa, \lambda \geq 0$ and $0 \leq \gamma \leq 0.9999$. Free-lag coefficients b_0 and $b_0 + b_1$, floored at a small positive value, initialize κ, λ. Three runs initialize γ at $0.3, 0.8, 0.99$. All three parameters are optimized in each run, retaining the smallest training error. Volume is rescaled for numerical conditioning and reported $\widehat{\kappa}, \widehat{\lambda}$ are in half-spreads per share.

Without refitting, we score both models on 1,800 afternoon intervals. Test R^2 is the fractional reduction in squared error relative to predicting the test-period mean midpoint change. The impact fit scores 0.202-0.537, matching or slightly exceeding free lags for each stock (Figure 9, left). This evaluates contemporaneous price-volume fit. AAPL has $\widehat{\gamma} = 0.9673$ while the other four fits reach the imposed limit 0.9999, leaving their decay weakly resolved. For a one-minute order, $T(1-\widehat{\gamma}) = 0.0012$-$0.3929$.

Learning the twelve purchase fractions. For each stock and seed, Algorithm 1 starts with $w_t^1 = 1/T$. Here w_t^k is the fraction bought in period t, $x_t^k = B w_t^k$, and $\boldsymbol{w}^k \in \Delta_T = \{\boldsymbol{w} \geq 0 : \sum_t w_t = 1\}$. At training episode k, we sample one of that stock's 120 morning windows uniformly with replacement and execute the already chosen schedule. The simulator uses its historical ask path a_t^k and calculates

$$\begin{aligned} P_t^k &= a_t^k + I_t^k + \kappa_{\text{share}}(B w_t^k + B/T), \\ I_{t+1}^k &= \widehat{\gamma} I_t^k + \lambda_{\text{share}}(B w_t^k + B/T), \qquad I_1^k = 0, \end{aligned} \tag{88}$$

where a_t^k is the last recorded ask before the decision, $\kappa_{\text{share}} = s\widehat{\kappa}$ and $\lambda_{\text{share}} = s\widehat{\lambda}$ are in dollars per share. The impact state is reset for each order while the competitor remains uniform.

After all twelve purchases, the learner updates the whole schedule. For numerical scaling, write $\bar{\kappa} = B\widehat{\kappa}$, $\bar{\lambda} = B\widehat{\lambda}$ and $\widetilde{P}_t^k = (P_t^k - a_1^k)/s$. The price-gradient update of Algorithm 1 becomes

$$\begin{aligned} g_t^k &= \widetilde{P}_t^k + \bar{\kappa} w_t^k + \bar{\lambda} \sum_{r=t+1}^{T} \widehat{\gamma}^{\,r-t-1} w_r^k, \\ \boldsymbol{w}^{k+1} &= \text{proj}_{\Delta_T}\left(\boldsymbol{w}^k - \frac{\boldsymbol{g}^k}{2\bar{\mu}k} \right), \qquad \bar{\mu} = \bar{\kappa} - \frac{\bar{\lambda}}{1+\widehat{\gamma}}. \end{aligned} \tag{89}$$

The last gradient term accounts for an earlier purchase raising prices paid on later purchases by using the trader's executed schedule. Projection keeps quantities nonnegative and their total equal to B. We perform 2,000 updates for each of eight seeds per stock, holding the fitted parameters fixed. Only the twelve fractions are learned, and updates occur between complete orders. Each final schedule is then frozen.

Reference and held-out evaluation. We compare the frozen learned schedule with uniform buying and an impact-only reference: the best response to the same uniform competitor, computed independently by minimizing execution cost under the fitted parameters with a constant background price. This reference is not supplied to the learner.

For every afternoon window, each method uses its fixed purchase fractions and the same historical ask path. Equation (88) is rerun from zero impact for each method, so its own purchases generate its own impact and execution prices. There are no test-time learning updates. For method A, total expenditure is $C_A = \sum_t x_t^A P_t^A$. With m_0 the midpoint at order arrival, we report

$$\text{Saving}_A = 10^4 \frac{C_{\text{uniform}} - C_A}{B m_0} \quad \text{basis points (bps).}$$

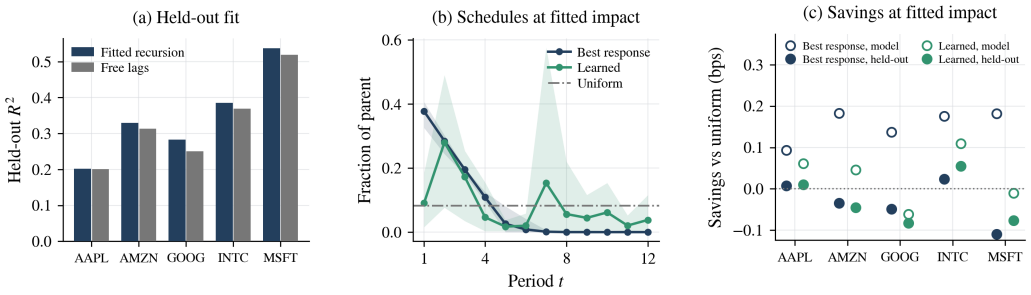


Figure 9: Fitted impact, $B = 0.10V_{60}$. Left: test R^2. Right: final fractions, averaged over seeds then stocks; shading spans stocks. "Best response" is the impact-only reference.

Figure 10: Fitted-parameter savings against uniform. Open markers: impact only. Filled markers: afternoon ask paths included.

We average seeds within each stock-window, then windows within each of ten 15-minute test blocks and stocks equally within each block. Resampling these matched blocks 1,000 times gives descriptive 95% intervals.

Results at fitted parameters. The reference buys early, using four to seven of twelve periods. We observe that learning recovers this shape unevenly across stocks (Figure 9, right). Impact alone gives reference savings of 0.094–0.183 bps, converted using the mean training midpoint. Adding afternoon ask paths gives -0.0322 bps, with interval $[-0.2551, 0.1715]$, for the reference and -0.0284 bps, with interval $[-0.1134, 0.0405]$, for learning. Both intervals include zero: this sample does not resolve positive savings at the fitted impact scale (Figure 10).

Stronger impact scenarios. We repeat the training and evaluation with $B = 0.01V_{60}$ and prescribed parameters

$$\kappa_{\text{share}} = \frac{\nu s T}{B}, \qquad \lambda_{\text{share}} = 0.5\kappa_{\text{share}}, \qquad \nu \in \{0.5, 1, 2\}, \quad \gamma \in \{0.3, 0.7, 0.95, 1\}.$$

A uniform purchase B/T now adds νs of immediate impact, substantially more than the fitted model. Each setting trains its own schedule, using $\bar{\kappa} = B\kappa_{\text{share}}/s$, $\bar{\lambda} = B\lambda_{\text{share}}/s$ and the prescribed γ in the equations above. The "No forward term" learner omits the final sum in (89); "Permanent belief" uses $\gamma = 1$ in its correction and step size. All methods are evaluated under the scenario's actual parameters on the same afternoon windows.

At $(\gamma, \nu) = (0.7, 1)$, full-learner savings range from 0.164 to 0.342 bps across stocks. The cross-stock mean is 0.2328 bps (95% interval $[0.1983, 0.2586]$), close to the reference's 0.2361 bps. Omitting the forward term gives 0.1449 bps; assuming permanent impact gives -0.5316 bps. Savings increase with impact strength and persistence in this grid (Figure 11). The learned schedule stays slightly worse than the reference; Figure 12 also shows AAPL's final allocation. All schedules complete their quota to numerical precision.

Scope. Learning closely recovers the model's timing benefit under stronger impact. The calibration is descriptive and the evidence covers five stocks on one date. The simulator assumes purchases execute at its calculated prices, omitting fill uncertainty, replenishment, routing, latency and fees. Some purchases exceed displayed best-ask depth, especially for AAPL. The intervals describe within-session variation.

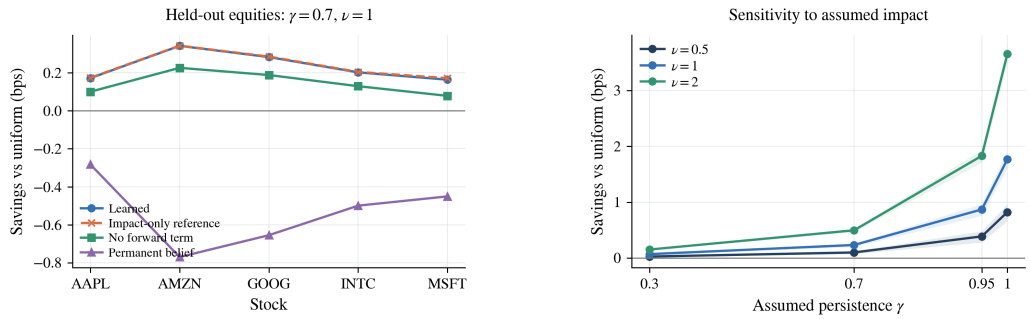


Figure 11: Prescribed impact, $B = 0.01V_{60}$. Left: test savings against uniform at $(\gamma, \nu) = (0.7, 1)$. Right: mean learned savings across stocks as persistence γ and strength ν vary.

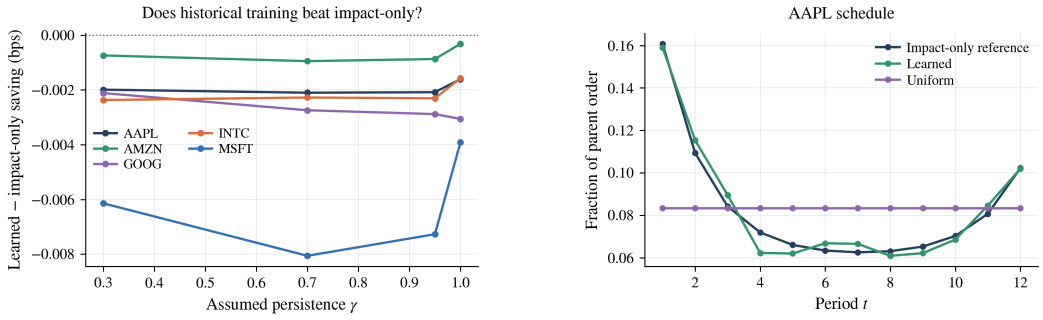


Figure 12: Left: learned savings minus reference savings at $\nu = 1$; negative values favour the reference. Right: AAPL's final purchase fractions at $(\gamma, \nu) = (0.7, 1)$.