DATA, OBJECTIVE, AND ARCHITECTURE: WHEN DO INDEPENDENT POLICY-GRADIENT LEARNERS COLLUDE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative AI agents increasingly make sequential decisions in multi-agent environments, where independently reward-maximizing agents can nevertheless learn cooperative or collusive behavior. We study how two defining components of modern generative agents—*data used in pre-training* and *memory*—shape collusion under independent policy-gradient learning. Using a repeated pricing game as a controlled social-dilemma environment, we develop an analytical framework that separates what an agent learns *from* its pre-training data, what it learns *toward* under its post-training objective, and what it can *remember* through its context architecture. We first show that next-token pre-training acts as a behavioral-transmission mechanism: under realizability and standard generalization conditions, the learned policy recovers the teacher's history-conditional action distribution, so pre-training determines the behavioral initialization from which subsequent multi-agent learning begins. We then characterize independent own-reward population policy-gradient dynamics. With one-step memory, we derive an exact action-value gap showing that the gradient favors high-price actions only when the opponent's learned reciprocity exceeds an explicit threshold. This yields a patience condition and an invariant non-collusive basin: sufficiently non-collusive initial policies converge toward competitive behavior, while crossing the threshold is not by itself sufficient for collusion. Memory changes these dynamics by creating additional continuation-value channels. Without learned memory, agents cannot retaliate, but collusion can still arise through coordinated discovery; we derive a separate threshold below which policy-gradient learning eliminates high-price behavior. For learned context length $K \geq 2$, we analyze a soft-attention policy and derive a sufficient non-collusion threshold that decreases with K while remaining strictly positive for every fixed discount factor below one. Thus, longer memory enlarges the set of future states through which current actions can affect an opponent's behavior and correspondingly shrinks the guaranteed competitive basin. Together, our results show that pre-training shapes multi-agent learning through *initialization*, while memory reshapes its *basins of attraction*, providing a tractable learning-theoretic account of emergent collusion under independent policy-gradient learning.

1 INTRODUCTION

Firms increasingly delegate pricing and other market decisions to adaptive algorithms. Independent learning systems can generate supracompetitive outcomes even without explicit communication (Calvano et al., 2020; Abada & Lambin, 2023; Douglas et al., 2024; den Boer et al., 2026), and recent work shows that LLM-based pricing agents can sustain high prices and react materially to changes in prompts or market information (Fish et al., 2024; Akata et al., 2025; Lin et al., 2024). These findings create a practical identification problem: when a generative pricing agent behaves collusively, which part of the technology is responsible?

A deployed generative agent bundles together three conceptually distinct mechanisms. Its *pretraining data* determine which conditional behaviors are learned before deployment; its *post-training*

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.

objective determines how those behaviors are reinforced or discouraged; and its *architecture/context* determines which past interactions can affect current actions. Existing work typically varies one of these ingredients while treating the others as fixed. Classical algorithmic-collusion models emphasize learning rules and exploration (Calvano et al., 2020; Abada & Lambin, 2023; Douglas et al., 2024); LLM market experiments largely study already-trained prompted models (Fish et al., 2024); and recent AI-monoculture theory studies correlated recommendations and retraining of a shared model (Cao & Hu, 2026). We instead separate the three channels within one repeated-pricing framework with independently parameterized agents.

Our first result concerns the *data channel*. Under realizability and a standard finite-class generalization condition, next-token training recovers the teacher's episode distribution and therefore its history-conditional high-price probabilities on teacher-supported histories. Pretraining is thus a behavioral-transmission mechanism: it can seed a deployed agent with a more or less collusive conditional policy before any own-profit adaptation occurs. Our second result concerns the *objective channel*. With one-step memory, we derive an exact action-value gap. High-price behavior is favored only if the opponent's response to a previous high price is sufficiently stronger than its response to a low price. This yields an exact reciprocity threshold, a patience condition, and an invariant non-collusive basin under exact population-gradient learning. Importantly, being outside that basin is not itself sufficient for collusion.

Our third result concerns the *architecture channel*. With no learned memory, agents cannot retaliate, but a persistent high-price regime may still be reached through coordinated discovery; we derive a separate sufficient threshold below which learning drives the high-price probability to zero. With learned context $K \geq 2$, current actions can change multiple future information sets. A coupling argument yields a sufficient non-collusion threshold that decreases with K and converges to a strictly positive limit for every fixed discount factor below one. Longer context therefore weakens the sufficient competitive guarantee without mechanically implying collusion. Together, these results answer the question in our title in the form of necessary conditions: within the model, independent policy-gradient learners can end up colluding only if their initial policy lies outside the relevant non-collusive basin, they are patient enough that $\delta > \Delta/\Gamma$ (with one-step memory), and a current high price can change future play through learned memory or a discoverable coordination regime; these conditions are necessary, not sufficient, for collusion.

2 Related Work

Our framework connects three literatures. Algorithmic-collusion research shows that Q-learning, bandit exploration, and misspecified pricing algorithms can produce supracompetitive outcomes, while also emphasizing that high prices need not imply strategic retaliation (Calvano et al., 2020; Abada & Lambin, 2023; Douglas et al., 2024; den Boer et al., 2026; Hansen et al., 2021). Work on LLMs as strategic agents documents autonomous high-price behavior, market division, and stable responses in repeated games (Fish et al., 2024; Lin et al., 2024; Akata et al., 2025). We differ by opening the training pipeline rather than treating model weights as fixed. Finally, our decomposition maps naturally to predictive sequence modeling, attention, and reward-based adaptation: pretraining uses autoregressive likelihood (Vaswani et al., 2017), whereas post-training follows own-reward policy gradients (Williams, 1992; Sutton et al., 2000). Relative to shared-model or AI-monoculture mechanisms (Cao & Hu, 2026), our agents have separate parameters and coordinate, if at all, through the economic interaction rather than mechanically correlated recommendations.

3 Model

Repeated pricing. Two agents $i \in \{1, 2\}$ choose $A_{i,t} \in \{H, L\}$ at times $t = 0, 1, \ldots$. Stage payoff is

$$u_i(a_i, a_{-i}) = u_0 + \Gamma \mathbf{1}\{a_{-i} = H\} + \Delta \mathbf{1}\{a_i = L\}, \qquad \Gamma > \Delta > 0, \tag{1}$$

so L is strictly dominant in the one-shot game, while mutual H yields higher joint payoff than mutual L. The competitive and collusive per-agent benchmarks are $\pi^N = u_0 + \Delta$ and $\pi^C = u_0 + \Gamma$. Agents maximize

$$J_i = \mathbb{E}\left[\sum_{t=0}^{\infty} \delta^t u_i(A_{i,t}, A_{-i,t})\right], \qquad 0 < \delta < 1, \tag{2}$$

Table 1: Model comparison with existing literature. "In-context": punishment arises with frozen weights, through prompting rather than training; in Cao & Hu (2026), one shared model's correlated recommendations, not reward and punishment, elevate prices.

	Calvano et al. (2020)	Fish et al. (2024)	Douglas et al. (2024)	Cao & Hu (2026)	This paper
Agent	Q-learning	LLM (prompted)	multi-armed bandit	shared AI model	generative policy, trained
Method	simulation	experiment	theory + simulation	theory	theory
Reward mechanism	✓	✓ (in-context)	× (absence shown)	× (correlation)	✓ (learned, $K \geq 1$)
Pre-training data	×	indirect (no isolation)	×	indirect (propensity)	✓ (imitation bound)
Post-training objective	×	×	partial (bandit rule)	partial	✓ (profit-gradient dynamics)
Architecture/memory	×	×	×	×	✓ (length varied)
Channels separated	×	×	×	×	✓

and update separate policy parameters by simultaneous exact population-gradient ascent. Conditional on the current state, agents randomize independently. The reward range is

$$B := r_{\max} - r_{\min} = \Gamma + \Delta. \tag{3}$$

Memory. For $K \geq 1$, let $S_t = (g_{t-1}, \ldots, g_{t-K})$, where $g_t = (A_{1,t}, A_{2,t})$. Although the implementation supplies only an agent's own actions and payoffs, deterministic payoffs identify the rival's lagged action. Define $Y_{i,j}(S_t) = \mathbf{1}\{A_{-i,t-j} = H\}$ and a positional soft-attention feature

$$\begin{aligned} \mu_i(S_t) &= \sum_{j=1}^{K} a_{ij} Y_{i,j}(S_t), \qquad a_{ij} = \frac{e^{w_{ij}}}{\sum_{\ell=1}^{K} e^{w_{i\ell}}}, \\ \pi_i(H \mid S_t) &= \sigma(\theta_i + \lambda_i[\mu_i(S_t) - c]), \end{aligned} \tag{4}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the logistic function and $c \in [0,1]$ is a fixed constant. This restricted attention policy isolates how learned memory changes continuation values; it is not a claim about arbitrary multilayer transformers. For $K = 0$ we use a separate coordination-discovery benchmark with pre-discovery probability $p = \sigma(\alpha)$ and a persistent regime activated after m consecutive joint-H outcomes.

Outcome. For an evaluation horizon T_{ev}, with periods indexed by $t = 1, \ldots, T_{\text{ev}}$, define the period-$t$ market payoff as

$$\pi_t^M := \frac{1}{2}(u_{1,t} + u_{2,t}),$$

and the normalized discounted weight as

$$\omega_t := \frac{(1-\delta)\delta^{t-1}}{1-\delta^{T_{\text{ev}}}}.$$

The normalized discounted average market payoff is therefore

$$\bar{\pi}^\delta := \sum_{t=1}^{T_{\text{ev}}} \omega_t \pi_t^M.$$

The collusion index is

$$\kappa^\delta = \frac{\bar{\pi}^\delta - \pi^N}{\pi^C - \pi^N}. \tag{5}$$

The proofs control its positive part and the probability of exceeding any positive threshold.

4 DATA CHANNEL: PRETRAINING AS BEHAVIORAL TRANSMISSION

A teacher generates episodes $Z = (X_1, \dots, X_{T_{\text{pre}}}) \sim Q^\star$ with $Q^\star(Z) = \prod_t q^\star(X_t \mid X_{<t})$, and the student has $P_\omega(Z) = \prod_t p_\omega(X_t \mid X_{<t})$. From n independent episodes, the student minimizes

$$\widehat{L}_n(\omega) = -\frac{1}{n}\sum_{r=1}^{n}\sum_{t=1}^{T_{\text{pre}}} \log p_\omega(X_t^{(r)} \mid X_{<t}^{(r)}). \tag{6}$$

Assume realizability, a lower bound $p_\omega(X_t \mid X_{<t}) \geq \gamma > 0$ on teacher-supported histories, and a finite effective hypothesis class $\mathcal{H}$.

Assume the following pretraining conditions:

- **(P1) Realizability:** the teacher distribution belongs to the finite effective hypothesis class, $Q^\star \in \mathcal{H}$.
- **(P2) Teacher-support lower bound:** for every $P_\omega \in \mathcal{H}$ and every token X_t that can occur under the teacher,
$$p_\omega(X_t \mid X_{<t}) \geq \gamma > 0.$$

Let $\mathcal{A} \subseteq \{1, \dots, T_{\text{pre}}\}$ denote the set of price-decision positions in the sequence. At each $t \in \mathcal{A}$, define the student's and teacher's conditional probabilities of choosing the high price H by

$$\widehat{p}_{H,t} := P_{\widehat{\omega}}(X_t = H \mid X_{<t}), \qquad p^\star_{H,t} := Q^\star(X_t = H \mid X_{<t}).$$

Proposition 1 (Faithful behavioral transmission)**.** *Let*

$$\widehat{\omega} \in \arg\min_{\omega:\, P_\omega \in \mathcal{H}} \widehat{\mathcal{L}}_{\text{pre}}(\omega),$$

where $\widehat{\mathcal{L}}_{\text{pre}}$ is the empirical version of equation 6, and let $C_{\text{pre}} = T_{\text{pre}} \log(1/\gamma)$. With probability at least $1 - \eta$,

$$D_{\text{KL}}(Q^\star \| P_{\widehat{\omega}}) \leq C_{\text{pre}} \sqrt{\frac{2\log(2|\mathcal{H}|/\eta)}{n}}. \tag{7}$$

At the binary pricing-decision positions $\mathcal{A}$,

$$\sum_{t \in \mathcal{A}} \mathbb{E}_{Q^\star}\left[\left(\widehat{p}_{H,t} - p^\star_{H,t}\right)^2\right] \leq \frac{1}{2} D_{\text{KL}}(Q^\star \| P_{\widehat{\omega}}). \tag{8}$$

Thus predictive training transmits the conditional behavior encoded in the training distribution on teacher-supported histories. It does not imply exact finite-sample imitation, behavior off support, or optimization outside the effective hypothesis class $\mathcal{H}$. Economically, pretraining determines the behavioral initialization from which subsequent reward-based adaptation begins. The proof is in Appendix A.1.

5 OBJECTIVE CHANNEL: EXACT ONE-STEP RECIPROCITY

For $K = 1$, let

$$\pi_i(H \mid Y_{i,t}) = \sigma(\theta_i + \lambda_i(Y_{i,t} - c)), \qquad Y_{i,t} = \mathbf{1}\{A_{-i,t-1} = H\}, \tag{9}$$

and define $p_i^+ = \mathbb{P}_i(H \mid Y_{i,t} = 1)$, $p_i^- = \mathbb{P}_i(H \mid Y_{i,t} = 0)$, and

$$\rho_i := p_i^+ - p_i^-. \tag{10}$$

The coefficient ρ_i measures learned reciprocity: how much more likely agent i is to choose H after the rival previously chose H.

Proposition 2 (Exact one-step action-value gap)**.** *For every one-step state s,*

$$A_i := Q_i(s, H) - Q_i(s, L) = \frac{\delta\Gamma\rho_{-i} - \Delta}{1 - \delta^2\rho_i\rho_{-i}}, \tag{11}$$

whose denominator is positive. Hence

$$A_i > 0 \iff \rho_{-i} > \rho^\star := \frac{\Delta}{\delta\Gamma}. \tag{12}$$

High-price behavior is therefore favored not by the rival's unconditional high-price level, but by sufficiently strong contingency. A current H must induce enough additional future H behavior to offset the immediate gain Δ from choosing L.

Corollary 1 (Patience is necessary for one-step reciprocity)**.** *If*

$$\delta \le \frac{\Delta}{\Gamma}, \tag{13}$$

then $A_i < 0$ for every finite policy parameter vector.

Let $p_1^\star = \min\{1, \Delta/(\delta\Gamma)\}$.

Proposition 3 (Invariant non-collusive basin with one-step memory)**.** *If $\max_i p_i^+(0) < p_1^\star$, then under simultaneous exact population-gradient flow,*

$$p_i^+(s) \le p_i^+(0), \quad p_i^-(s) \le p_i^-(0), \quad A_i(s) < 0 \quad \forall s \ge 0. \tag{14}$$

From the deterministic competitive history LL, $p_i^-(s) \to 0$. For finite-horizon evaluation from LL, if $q_1(s) = 1 - (1 - p_1^-(s))(1 - p_2^-(s))$ and $D_{\delta,T_{\mathrm{ev}}} = \sum_t t\omega_t^{\delta,T_{\mathrm{ev}}}$, then

$$\mathbb{E}_s[(\kappa^\delta)_+] \le D_{\delta,T_{\mathrm{ev}}} q_1(s), \qquad \mathbb{P}_s(\kappa^\delta \ge \tau) \le \frac{D_{\delta,T_{\mathrm{ev}}} q_1(s)}{\tau} \to 0. \tag{15}$$

The condition is one-sided: it identifies a guaranteed safe basin, not a sufficient condition for collusion outside it. Proofs are in Appendices A.2–A.5.

6 Architecture Channel: Discovery and Longer Context

No learned memory. Without history-dependent parameters, agents cannot retaliate, but a persistent coordinated regime may still be discovered. If m consecutive joint-H outcomes activate that regime, define

$$p_{\mathrm{disc}}^\star(m) := \min\left\{1, \left[\frac{\Delta(1-\delta)^2}{m\delta B}\right]^{1/(2m-1)}\right\}. \tag{16}$$

Proposition 4 (Coordination-discovery basin without learned memory)**.** *Under symmetric initialization, conditional independence before discovery, bounded rewards, and exact population-gradient learning, if $p(0) < p_{\mathrm{disc}}^\star(m)$, then*

$$p(s) < p_{\mathrm{disc}}^\star(m)\ \forall s \ge 0, \qquad p(s) \to 0. \tag{17}$$

For finite-horizon evaluation from the undiscovered competitive state, with $q_0(s) = 1 - (1 - p(s))^2$ and $C_{\delta,T_{\mathrm{ev}},m} = \sum_t \omega_t^{\delta,T_{\mathrm{ev}}}(t - m + 1)_+$,

$$\mathbb{P}_s(\kappa^\delta \ge \tau) \le \frac{q_0(s) + C_{\delta,T_{\mathrm{ev}},m} p(s)^{2m}}{\tau} \to 0. \tag{18}$$

The two terms distinguish ordinary experimentation from successful discovery of the persistent regime.

Learned context $K \ge 2$. Let $p_i^H = \sigma(\theta_i + (1-c)\lambda_i)$ and $p_i^L = \sigma(\theta_i - c\lambda_i)$, and define

$$\bar{p} := \max_i \max\{p_i^H, p_i^L\}, \qquad S_K(\delta) := \sum_{j=1}^{K} \delta^j. \tag{19}$$

Every history-specific high-price probability is bounded by $\bar{p}$.

Proposition 5 (Sufficient non-collusive basin with learned context-K attention)**.** *If $\bar{p}S_K(\delta) < 1$, then for every agent and history,*

$$Q_i(h, H) - Q_i(h, L) \le -\Delta + B\frac{\bar{p}S_K(\delta)}{1 - \bar{p}S_K(\delta)}. \tag{20}$$

Table 2: Summary of theoretical thresholds.

Mechanism	Threshold		Interpretation
No learned memory	$p^\star_{\mathrm{disc}}(m)$	$=$	Sufficient discovery basin; m is the joint-H streak needed to activate the persistent regime.
	$\min\left\{1, \left\lceil \frac{\Delta(1-\delta)^2}{m\delta B} \right\rceil^{1/(2m-1)}\right\}$		
One-step memory	$\rho^\star = \frac{\Delta}{\delta\Gamma}; p_1^\star = \min\{1, \rho^\star\}$		$\rho^\star$ is the exact reciprocity threshold; $\max_i p_i^+(0) < p_1^\star$ gives a sufficient invariant safe basin.
Context $K \geq 2$	$p^\star_{K,\mathrm{suff}}$	$=$	Conservative sufficient threshold, decreasing in K within the bound family.
	$\min\left\{1, \frac{\Delta}{(B+\Delta)S_K(\delta)}\right\}$		

Define

$$p^\star_{K,\mathrm{suff}} := \min\left\{1, \frac{\Delta}{(B+\Delta)S_K(\delta)}\right\}. \tag{21}$$

If $\bar{p}(0) < p^\star_{K,\mathrm{suff}}$, then $\bar{p}(s) \leq \bar{p}(0) < p^\star_{K,\mathrm{suff}}$ for all learning times and the action-value gap is negative at every history; from the all-L context, the high-price probability converges to zero.

For finite-horizon evaluation from the all-L context, let

$$q_K(s) = 1 - \big(1 - p_1^L(s)\big)\big(1 - p_2^L(s)\big).$$

Then

$$\mathbb{E}_s[(\kappa^\delta)_+] \leq D_{\delta,T_{\mathrm{ev}}} q_K(s), \qquad \mathbb{P}_s(\kappa^\delta \geq \tau) \leq \frac{D_{\delta,T_{\mathrm{ev}}} q_K(s)}{\tau} \to 0 \quad \textit{for every } \tau > 0. \tag{22}$$

The threshold is deliberately sufficient rather than exact: the proof uses an extreme-history envelope and a coupling bound uniform over all 4^K histories.

Corollary 2 (Effect of longer context within the sufficient-bound family)**.** *For $K \geq 2$,*

$$p^\star_{K+1,\mathrm{suff}} \leq p^\star_{K,\mathrm{suff}}, \qquad p^\star_{K,\mathrm{suff}} \to \min\left\{1, \frac{\Delta(1-\delta)}{(B+\Delta)\delta}\right\} > 0. \tag{23}$$

Longer context weakens the sufficient guarantee because a current action can influence more future information sets. It does not imply that longer context necessarily causes collusion, and the $K = 0$ discovery threshold is not directly ordered with the memory thresholds because m and K describe different mechanisms.

For $(\Gamma, \Delta, \delta) = (1, 1/2, 0.95)$, the exact one-step reciprocity threshold is 0.5263, while the generic sufficient context bound gives $p^\star_{2,\mathrm{suff}} \approx 0.1350$, $p^\star_{10,\mathrm{suff}} \approx 0.0328$, and $p^\star_{\infty,\mathrm{suff}} \approx 0.0132$. The difference between the exact $K = 1$ result and the generic $K \geq 2$ bounds should not be interpreted entirely as a causal effect of context length. Proofs are in Appendices A.6–A.9.

7 Conclusion

In answer to the question in the title, independent policy-gradient learners in our model can end up colluding only if their initial policy lies outside the relevant non-collusive basin, which with one-step memory requires $\delta > \Delta/\Gamma$, and a current high price can affect future play through learned memory or a discoverable coordination regime; these conditions are necessary, not sufficient.

Our theory supports a mechanism-specific audit rather than a single collusion score. Pretraining can be assessed by measuring a model's conditional high-price probabilities on economically relevant histories. With one-step memory, the economically sharp object is the reciprocity coefficient ρ_i, whose exact action-value threshold is $\Delta/(\delta\Gamma)$; the observable condition $\max_i p_i^+(0) < p_1^\star$ provides a sufficient safe-basin test. For $K \geq 2$, the corresponding sufficient diagnostic compares the maximum endpoint probability $\bar{p}(0)$ with $p^\star_{K,\mathrm{suff}}$; for the no-memory discovery mechanism it compares $p(0)$ with $p^\star_{\mathrm{disc}}(m)$. Within the modeled environment, admission inside the relevant basin guarantees that continued exact own-profit population-gradient adaptation does not escape it.

These are model-based diagnostics, not universal certification rules for arbitrary LLM pricing systems. The guarantees rely on the specified duopoly, payoff structure, information environment, policy classes, separate agent parameters, conditional action independence, and exact population-gradient dynamics. The context-K result is sufficient rather than exact, and the pretraining result requires realizability and generalization assumptions. Their value is to map observable conditional behavior and design choices to transparent mechanisms rather than to infer collusion from high prices alone.

Generative pricing agents combine three channels that are often studied together. Pretraining determines which behavior is initially available; post-training determines how own-profit incentives move that initial policy; and architecture determines how present actions can affect future responses. Our results make each channel explicit. Predictive training transmits the teacher's conditional behavior. One-step own-profit learning favors high prices only above an exact reciprocity threshold and otherwise possesses an invariant non-collusive basin. Without learned memory, coordination requires a distinct discovery mechanism; with longer learned context, the sufficient safe region shrinks because more future information sets can carry the effect of a current action.

Similar high-price outcomes can reflect different economic mechanisms: an inherited behavioral prior, coordinated discovery without retaliation, or history-contingent reward–punishment behavior. The central implication is therefore not that memory or post-training mechanically creates collusion, but that data, objective, and architecture determine different parts of the pathway through which it can arise. Richer markets, finite-sample learning dynamics, unrestricted attention architectures, and larger pretrained models remain natural extensions.

AI USE STATEMENT

We have not used generative AI tools to develop the research questions, the model, or the conceptual framework; to formulate the mathematical claims or write the proofs; or to propose hypotheses. Translation and qualitative or thematic data analysis are not applicable to this work. We used Claude Code (Anthropic) to check the manuscript for grammar, spelling, and typesetting errors, and to check the proofs for errors. Every issue it flagged was independently verified by the authors, and all resulting corrections to statements and proofs were written by the authors. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

This is a theoretical paper. The model, payoffs, policy classes, and exact population-gradient dynamics are defined in Section 3. Section 4 states the realizability, likelihood-floor, and finite-hypothesis-class assumptions of Proposition 1, and the conditions of Propositions 2–5 and Corollaries 1–2 appear in their statements in Sections 5–6. Complete proofs of all propositions, corollaries, and supporting lemmas are given in Appendices A.1–A.9, and Appendix A.10 states the scope of the population-learning results. The threshold values reported in Section 6 and Table 2 follow in closed form from equation 12, equation 21, and equation 23 and can be recomputed directly.

REFERENCES

Ibrahim Abada and Xavier Lambin. Artificial intelligence: Can seemingly collusive outcomes be avoided? *Management Science*, 69(9):5042–5065, 2023. doi: 10.1287/mnsc.2022.4623.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020. doi: 10.1257/aer.20190623.

Shengyu Cao and Ming Hu. Supracompetitive pricing under ai monoculture, 2026.

Arnoud V. den Boer, Janusz M. Meylahn, and Maarten Pieter Schinkel. Artificial collusion: Examining supracompetitive pricing by q-learning algorithms. *Management Science*, 2026. doi: 10.1287/mnsc.2024.08557. Published online June 9, 2026.

Connor Douglas, Foster Provost, and Arun Sundararajan. The illusion of collusion, 2024.

Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models, 2024.

Karsten T. Hansen, Kanishka Misra, and Mallesh M. Pai. Frontiers: Algorithmic collusion: Supracompetitive prices via independent algorithms. *Marketing Science*, 40(1):1–12, 2021. doi: 10.1287/mksc.2020.1276.

Ryan Y. Lin, Siddhartha M. Ojha, Kevin Cai, and Maxwell F. Chen. Strategic collusion of llm agents: Market division in multi-commodity competitions, 2024.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pp. 1057–1063. MIT Press, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

A Appendix

A.1 Proof of Proposition 1

Proof. For an episode Z, define

$$\ell_\omega(Z) := -\log P_\omega(Z) = -\sum_{t=1}^{T_{\mathrm{pre}}} \log p_\omega(X_t \mid X_{<t}).$$

Assumption (P2) implies $0 \le \ell_\omega(Z) \le T_{\mathrm{pre}} \log(1/\gamma) = C_{\mathrm{pre}}$. Let

$$L(\omega) := \mathbb{E}_{Q^\star}[\ell_\omega(Z)].$$

By the cross-entropy identity,

$$\begin{aligned} L(\omega) &= \mathbb{E}_{Q^\star}[-\log Q^\star(Z)] + \mathbb{E}_{Q^\star}\left[\log \frac{Q^\star(Z)}{P_\omega(Z)}\right] \\ &= H(Q^\star) + D_{\mathrm{KL}}(Q^\star \,\|\, P_\omega). \end{aligned} \tag{24}$$

Under realizability, there is $\omega^\star$ with $P_{\omega^\star} = Q^\star$, so

$$L(\omega^\star) = H(Q^\star)$$

and

$$L(\omega) - L(\omega^\star) = D_{\mathrm{KL}}(Q^\star \,\|\, P_\omega).$$

For a fixed $P_\omega \in \mathcal{H}$, Hoeffding's inequality gives

$$\mathbb{P}\left(|\widehat{L}_n(\omega) - L(\omega)| \ge \varepsilon\right) \le 2\exp\left(-\frac{2n\varepsilon^2}{C_{\mathrm{pre}}^2}\right).$$

A union bound over $\mathcal{H}$ implies that, with probability at least $1-\eta$,

$$\sup_{\omega:\, P_\omega \in \mathcal{H}} |\widehat{L}_n(\omega) - L(\omega)| \le \varepsilon_n :=$$

$$C_{\mathrm{pre}} \sqrt{\frac{\log(2|\mathcal{H}|/\eta)}{2n}}.$$

On this event, empirical-risk minimization yields

$$\begin{aligned} L(\widehat{\omega}) &\le \widehat{L}_n(\widehat{\omega}) + \varepsilon_n \\ &\le \widehat{L}_n(\omega^\star) + \varepsilon_n \le L(\omega^\star) + 2\varepsilon_n. \end{aligned}$$

Therefore

$$D_{\mathrm{KL}}(Q^\star \,\|\, P_{\widehat{\omega}}) \le 2\varepsilon_n = C_{\mathrm{pre}} \sqrt{\frac{2\log(2|\mathcal{H}|/\eta)}{n}},$$

which proves equation 7.

Finally, the chain rule for KL divergence gives

$$\begin{aligned} &D_{\mathrm{KL}}(Q^\star \,\|\, P_{\widehat{\omega}}) \\ &= \sum_{t=1}^{T_{\mathrm{pre}}} \mathbb{E}_{Q^\star}\left[D_{\mathrm{KL}}\left(q^\star(\cdot \mid X_{<t}) \,\|\, p_{\widehat{\omega}}(\cdot \mid X_{<t}) \right) \right]. \end{aligned}$$

At a binary pricing-decision position, Pinsker's inequality implies

$$\begin{aligned} &\left(\widehat{p}_{H,t}(X_{<t}) - p^\star_{H,t}(X_{<t}) \right)^2 \\ &\le \frac{1}{2} D_{\mathrm{KL}}\left(q^\star(\cdot \mid X_{<t}) \,\|\, p_{\widehat{\omega}}(\cdot \mid X_{<t}) \right). \end{aligned}$$

Summing over decision positions gives equation 8. □

A.2 Supporting Results for One-Step Memory

For the proof, augment the state as

$$X_{i,t} := \mathbf{1}\{A_{i,t-1} = H\}, \qquad Y_{i,t} := \mathbf{1}\{A_{-i,t-1} = H\},$$
$$S_{i,t} = (X_{i,t}, Y_{i,t}).$$

Let

$$r_i(y) = p_i^- + \rho_i y, \qquad r_{-i}(x) = p_{-i}^- + \rho_{-i} x.$$

Then the expected one-period payoff at (x, y) is

$$\bar{u}_i(x, y) = u_0 + \Gamma r_{-i}(x) + \Delta(1 - r_i(y)). \tag{25}$$

Lemma 1 (Additive value representation)**.** *For a fixed joint policy, the infinite-horizon value function has the form*

$$V_i(x, y) = C_i + F_i(x) + G_i(y).$$

Proof. The stage payoff equation 25 is additive in x and y. If $V(x, y) = C + F(x) + G(y)$, conditional independence implies

$$\begin{aligned} \mathbb{E}[V(S_{t+1}) \mid x, y] = C &+ \mathbb{E}[F(X_{t+1}) \mid y] \\ &+ \mathbb{E}[G(Y_{t+1}) \mid x], \end{aligned}$$

which is also additive. The fixed-policy Bellman operator therefore preserves the additive subspace. Since $\delta < 1$, it is a contraction and has a unique fixed point, which must be additive. □

Define

$$D_i := F_i(1) - F_i(0), \qquad E_i := G_i(1) - G_i(0).$$

Lemma 2 (Continuation equations)**.** *The continuation differences satisfy*

$$D_i = \Gamma \rho_{-i} + \delta \rho_{-i} E_i, \tag{26}$$

$$E_i = -\Delta \rho_i + \delta \rho_i D_i. \tag{27}$$

Consequently,

$$D_i = \frac{\rho_{-i}(\Gamma - \delta \Delta \rho_i)}{1 - \delta^2 \rho_i \rho_{-i}}. \tag{28}$$

Proof. Changing the first state coordinate from zero to one increases the rival's current probability of H by ρ_{-i}, producing current benefit $\Gamma \rho_{-i}$ and continuation change $\delta \rho_{-i} E_i$. This proves equation 26. Changing the second coordinate from zero to one increases agent i's probability of H by ρ_i. Since H sacrifices Δ, the current effect is $-\Delta \rho_i$, and the next-period effect is $\delta \rho_i D_i$, proving equation 27. Substitution gives equation 28. □

A.3 Proof of Proposition 2

Proof. Choosing H rather than L lowers the current payoff by Δ and sets the next-period own-action state coordinate to one rather than zero. Therefore

$$Q_i(s, H) - Q_i(s, L) = -\Delta + \delta D_i.$$

Substituting equation 28,

$$Q_i(s, H) - Q_i(s, L) = \frac{\delta\Gamma\rho_{-i} - \Delta}{1 - \delta^2\rho_i\rho_{-i}}.$$

Because finite logits imply $|\rho_i| < 1$ and $0 < \delta < 1$, the denominator is strictly positive. The sign condition follows immediately. □

A.4 Proof of Corollary 1

Proof. If $\delta \leq \Delta/\Gamma$, then $\Delta/(\delta\Gamma) \geq 1$. For finite logistic policies, $\rho_{-i} < 1$, so $\delta\Gamma\rho_{-i} - \Delta < 0$. Proposition 2 then gives $A_i < 0$. □

A.5 Proof of Proposition 3

Let

$$f_i^{\pm} := p_i^{\pm}(1 - p_i^{\pm}) > 0$$

and define normalized discounted occupancies

$$d_i^+ = (1 - \delta) \sum_{t \geq 0} \delta^t \mathbb{P}(Y_{i,t} = 1),$$
$$d_i^- = (1 - \delta) \sum_{t \geq 0} \delta^t \mathbb{P}(Y_{i,t} = 0).$$

Then $d_i^+ + d_i^- = 1$. The policy-gradient identity of Sutton et al. (2000) gives

$$\frac{\partial J_i}{\partial \theta_i} = \frac{A_i}{1 - \delta} \left(d_i^+ f_i^+ + d_i^- f_i^- \right), \tag{29}$$
$$\frac{\partial J_i}{\partial \lambda_i} = \frac{A_i}{1 - \delta} \left((1 - c) d_i^+ f_i^+ - c\, d_i^- f_i^- \right). \tag{30}$$

Define

$$u_i^+ = \theta_i + (1 - c)\lambda_i, \qquad u_i^- = \theta_i - c\lambda_i.$$

Under exact gradient flow, direct substitution of equation 29–equation 30 yields

$$\dot{u}_i^+ = \frac{A_i}{1 - \delta} \Big([1 + (1 - c)^2] d_i^+ f_i^+ + [1 - c(1 - c)] d_i^- f_i^- \Big), \tag{31}$$
$$\dot{u}_i^- = \frac{A_i}{1 - \delta} \Big([1 - c(1 - c)] d_i^+ f_i^+ + [1 + c^2] d_i^- f_i^- \Big). \tag{32}$$

Every coefficient in brackets is strictly positive for $c \in [0, 1]$. Therefore

$$\operatorname{sign}(\dot{p}_i^+) = \operatorname{sign}(\dot{p}_i^-) = \operatorname{sign}(A_i). \tag{33}$$

Proof. Set

$$\bar{p}_1 := p_1^\star = \min \left\{ 1, \frac{\Delta}{\delta\Gamma} \right\}.$$

If $p_i^+ < \bar{p}_1$, then

$$\rho_i = p_i^+ - p_i^- \leq p_i^+ < \bar{p}_1 \leq \frac{\Delta}{\delta\Gamma}.$$

Hence $\delta\Gamma\rho_i - \Delta < 0$. Proposition 2 implies $A_{-i} < 0$ for both agents. By equation 33, both p_i^+ and p_i^- decrease whenever the trajectory is inside the region. A first-exit-time argument therefore shows that the region is forward invariant, proving equation 14.

Now suppose every training evaluation begins from LL. Define

$$M := \max_i p_i^+(0) < \frac{\Delta}{\delta\Gamma}, \qquad \varepsilon := \Delta - \delta\Gamma M > 0.$$

Forward invariance gives $\rho_i(s) \le M$. The exact action-value formula therefore implies the uniform bound

$$A_i(s) \le -\frac{\varepsilon}{1+\delta^2} =: -a_0 < 0.$$

Because the initial state has $Y_{i,0} = 0$ with probability one, $d_i^-(s) \ge 1 - \delta$. Equation equation 32 then gives

$$\dot{u}_i^- \le -a_0(1+c^2) f_i^-.$$

Since $\dot{p}_i^- = f_i^- \dot{u}_i^-$,

$$\dot{p}_i^- \le -a_0(1+c^2)[p_i^-(1-p_i^-)]^2.$$

Thus p_i^- is decreasing. If it converged to a positive limit, the right-hand side would eventually be bounded above by a strictly negative constant, a contradiction. Hence $p_i^- \to 0$.

For evaluation, when the process is at LL, the probability of departing in the next period is

$$q_1(s) = 1 - (1 - p_1^-(s))(1 - p_2^-(s)).$$

Let τ_{LL} be the first departure time from LL. Then $\mathbb{P}(\tau_{LL} = r) \le q_1(s)$ and therefore

$$\mathbb{P}(g_t \ne LL) \le t q_1(s).$$

Under $\Gamma > \Delta$, π^C is the largest average market payoff, so

$$(c_t)_+ \le \mathbf{1}\{g_t \ne LL\}, \qquad c_t := \frac{\pi_t^M - \pi^N}{\pi^C - \pi^N}.$$

Convexity of the positive-part map and the preceding probability bound imply

$$\mathbb{E}_s[(\kappa^\delta)_+] \le \sum_{t=1}^{T_{\mathrm{ev}}} \omega_t^{\delta, T_{\mathrm{ev}}} \mathbb{P}(g_t \ne LL) \le D_{\delta, T_{\mathrm{ev}}} q_1(s).$$

The expectation therefore converges to zero. Markov's inequality yields equation 15. □

A.6 Proof of Proposition 4

For a pre-discovery streak state $j \in \{0, \dots, m-1\}$, define

$$Q_j(a) := \mathbb{E}\Big[\text{discounted return from the current period onward} \;\Big|\; S_t^{\mathrm{disc}} = j, A_{i,t} = a\Big].$$

If the current streak length is j, then $m - j$ additional joint-H periods are needed. Conditional on agent i choosing H now, completing the current attempt requires one rival H now and $2(m-j-1)$ additional H actions afterward. Hence

$$\begin{aligned}&\mathbb{P}(\text{complete current attempt} \mid A_{i,t} = H, S_t^{\mathrm{disc}} = j) \\ &\quad \le p^{2(m-j)-1}.\end{aligned} \tag{34}$$

Couple the H and L counterfactual trajectories with common future random numbers. If the attempt fails before discovery, the two trajectories can be recoupled once the streak resets. A positive continuation difference can therefore arise only if the current attempt succeeds. Using the continuation range $\delta B/(1-\delta)$,

$$Q_j(H) - Q_j(L) \le -\Delta + \frac{\delta B}{1-\delta} p^{2(m-j)-1}. \tag{35}$$

Define normalized discounted occupancy

$$d_j := (1-\delta)\mathbb{E}\left[\sum_{t\geq 0}\delta^t \mathbf{1}\{S_t^{\text{disc}} = j\}\right].$$

To occupy state j, the preceding j periods must all have been joint H, so

$$d_j \leq p^{2j}. \tag{36}$$

Because the initial state has streak zero, the $t = 0$ term gives

$$\sum_{j=0}^{m-1} d_j \geq 1-\delta. \tag{37}$$

For $p = \sigma(\alpha)$, the Bernoulli score calculation yields

$$\frac{\partial J_i}{\partial \alpha_i} = \frac{p(1-p)}{1-\delta}\sum_{j=0}^{m-1} d_j[Q_j(H) - Q_j(L)]. \tag{38}$$

Combining equation 35–equation 37,

$$\sum_{j=0}^{m-1} d_j[Q_j(H) - Q_j(L)] \leq -\Delta(1-\delta) + \frac{m\delta B}{1-\delta}p^{2m-1}. \tag{39}$$

The right-hand side is negative precisely under the sufficient condition $p < p^\star_{\text{disc}}(m)$.

Proof. If $p(0) < p^\star_{\text{disc}}(m)$, then $\partial J_i/\partial \alpha_i < 0$. Since the sigmoid is increasing, $\dot{p} = p(1-p)\dot{\alpha} < 0$ whenever $p < p^\star_{\text{disc}}(m)$. Hence the region is forward invariant and $p(s)$ decreases to a limit $\bar{p} \geq 0$.

Because $p(s) \leq p(0)$, the right-hand side of equation 39 is bounded above by a strictly negative constant. If $\bar{p} > 0$, then $p(s)(1-p(s))$ is eventually bounded away from zero, so $\dot{p}$ is eventually bounded above by a strictly negative constant. This contradicts convergence to a positive limit. Therefore $p(s) \to 0$.

It remains to connect the action dynamics to the payoff index. Freeze the policy at learning time s for a finite evaluation episode. Let C_t denote that the persistent coordination regime has been reached by the end of period t. Before discovery, the probability that the current joint action differs from LL is

$$q_0(s) = 1 - (1-p(s))^2.$$

A positive stage index can occur either because the current pre-discovery joint action differs from LL, or because the persistent regime has already been reached. Thus

$$(c_t)_+ \leq \mathbf{1}\{C_t = 0,\ g_t \neq LL\} + \mathbf{1}\{C_t = 1\}.$$

The first event has probability at most $q_0(s)$. For the second, any specific length-m window is all joint H with probability at most $p(s)^{2m}$ before discovery. A union bound over at most $(t-m+1)_+$ windows gives

$$\mathbb{P}_s(C_t = 1) \leq (t-m+1)_+ p(s)^{2m}.$$

Therefore

$$\mathbb{E}_s[(c_t)_+] \leq q_0(s) + (t-m+1)_+ p(s)^{2m}.$$

Convexity of the positive part and the normalized weights imply

$$\mathbb{E}_s[(\kappa^\delta)_+] \leq q_0(s) + C_{\delta,T_{\text{ev}},m}\, p(s)^{2m},$$

which tends to zero because $p(s) \to 0$. Markov's inequality gives equation 18. □

A.7 Supporting Results for Context-K Attention

Lemma 3 (Extreme-history envelope)**.** *For every agent i and history $h \in \mathcal{S}_K$,*

$$\pi_i(H \mid h) \le \max\{p_i^H, p_i^L\}.$$

Proof. Because $\mu_i(h) \in [0,1]$,

$$\begin{aligned} z_i(h) =& (1-\mu_i(h))(\theta_i - c\lambda_i) \\ &+ \mu_i(h)(\theta_i + (1-c)\lambda_i). \end{aligned}$$

Thus $z_i(h)$ lies between the two endpoint logits. The sigmoid is increasing, so the corresponding probability lies between the endpoint probabilities. □

To control how a forced current action propagates through future contexts, couple two counterfactual trajectories with common uniform random variables: trajectory H forces agent i to choose H now, and trajectory L forces it to choose L. Let

$$e_i(n) = \mathbb{P}(A^H_{i,t+n} \neq A^L_{i,t+n}),$$

$$e_{-i}(n) = \mathbb{P}(A^H_{-i,t+n} \neq A^L_{-i,t+n}).$$

Then $e_i(0) = 1$ and $e_{-i}(0) = 0$.

Lemma 4 (Propagation bound)**.** *For $n \ge 1$,*

$$e_i(n) \le \bar{p} \sum_{j=1}^{K} e_{-i}(n-j), \tag{40}$$

$$e_{-i}(n) \le \bar{p} \sum_{j=1}^{K} e_i(n-j), \tag{41}$$

where disagreement probabilities at negative indices are zero. If $x := \bar{p} S_K(\delta) < 1$, then

$$\sum_{n=1}^{\infty} \delta^n [e_i(n) + e_{-i}(n)] \le \frac{x}{1-x}. \tag{42}$$

Proof. If all K rival-action tokens entering agent i's current context coincide across the two trajectories, its Bernoulli parameter is identical and common uniform coupling forces the two current actions to coincide. If at least one rival token differs, both Bernoulli parameters lie in $[0, \bar{p}]$, so their difference is at most $\bar{p}$. A union bound over the K context positions gives equation 40; the rival inequality is symmetric.

Let

$$D_i = \sum_{n \ge 0} \delta^n e_i(n), \qquad D_{-i} = \sum_{n \ge 0} \delta^n e_{-i}(n).$$

Multiplying the two recursions by δ^n and summing gives

$$D_i - 1 \le x D_{-i}, \qquad D_{-i} \le x D_i.$$

Thus

$$D_i \le \frac{1}{1-x^2}, \qquad D_{-i} \le \frac{x}{1-x^2}.$$

Therefore

$$\sum_{n=1}^{\infty} \delta^n [e_i(n) + e_{-i}(n)] = D_i - 1 + D_{-i} \le \frac{x}{1-x}.$$

□

Lemma 5 (Uniform action-value bound)**.** *If $\bar{p} S_K(\delta) < 1$, then for every i and h,*

$$Q_i(h, H) - Q_i(h, L) \le -\Delta + B \frac{\bar{p} S_K(\delta)}{1 - \bar{p} S_K(\delta)}.$$

Proof. The forced current H loses at least Δ. In a future period, the two stage payoffs can differ by at most B, and only if at least one player's action differs across the coupled trajectories. Hence

$$\begin{aligned} &Q_i(h,H) - Q_i(h,L) \\ &\quad \le -\Delta + B\sum_{n=1}^{\infty} \delta^n [e_i(n) + e_{-i}(n)]. \end{aligned}$$

Apply Lemma 4. □

Lemma 6 (Population policy gradient and endpoint descent)**.** *Let*

$$d_i^\delta(h) = (1-\delta)\sum_{t\ge 0}\delta^t \mathbb{P}(S_t = h),$$

$$f_i(h) = \pi_i(H \mid h)[1 - \pi_i(H \mid h)].$$

Then

$$\nabla_{\psi_i} J_i = \frac{1}{1-\delta}\sum_{h\in\mathcal{S}_K} d_i^\delta(h) f_i(h) A_i(h) \nabla_{\psi_i} z_i(h). \tag{43}$$

If $A_i(h) < 0$ for every history, then both endpoint logits and endpoint probabilities strictly decrease under exact population-gradient flow.

Proof. Equation equation 43 is the discounted policy-gradient identity specialized to a Bernoulli-logit policy. For the attention logits, softmax differentiation gives

$$\frac{\partial z_i(h)}{\partial w_{ir}} = \lambda_i a_{ir}[Y_{i,r}(h) - \mu_i(h)].$$

At the all-H and all-L histories, every $Y_{i,r}$ equals μ_i, so the endpoint logits do not depend directly on w_{ir}:

$$\nabla_{\psi_i} z_i^H = (1, 1-c, 0, \dots, 0),$$
$$\nabla_{\psi_i} z_i^L = (1, -c, 0, \dots, 0).$$

Taking inner products with equation 43,

$$\begin{aligned} \dot z_i^H &= \frac{1}{1-\delta}\sum_h d_i^\delta(h) f_i(h) A_i(h) \\ &\qquad \times \Big[1 + (1-c)\phi_i(h)\Big], \end{aligned} \tag{44}$$

$$\begin{aligned} \dot z_i^L &= \frac{1}{1-\delta}\sum_h d_i^\delta(h) f_i(h) A_i(h) \\ &\qquad \times \Big[1 - c\phi_i(h)\Big]. \end{aligned} \tag{45}$$

Because $\phi_i(h) \in [-c, 1-c]$,

$$1 + (1-c)\phi_i(h) \ge 1 - c(1-c) > 0,$$

and

$$1 - c\phi_i(h) \ge 1 - c(1-c) > 0.$$

Thus uniform negativity of $A_i(h)$ implies $\dot z_i^H < 0$ and $\dot z_i^L < 0$, and hence $\dot p_i^H < 0$ and $\dot p_i^L < 0$. □

A.8 Proof of Proposition 5

Proof. Lemma 5 gives

$$A_i(h) \le -\Delta + B\frac{x}{1-x}, \qquad x = \bar p S_K(\delta).$$

The right-hand side is strictly negative whenever

$$x < \frac{\Delta}{B + \Delta},$$

which is guaranteed by $\bar{p} < p^{\star}_{K,\mathrm{suff}}$. Therefore every action-value gap is negative at learning time zero.

Lemma 6 implies that all endpoint probabilities decrease while the trajectory remains in the region. By continuity, a first-exit-time argument shows that the trajectory cannot cross $p^{\star}_{K,\mathrm{suff}}$. Hence the basin is forward invariant and the action-value gap remains negative at every history and learning time.

Now assume every training evaluation begins from $h^0 = (LL, \ldots, LL)$. Set

$$x_0 := \bar{p}(0) S_K(\delta) < \frac{\Delta}{B + \Delta}$$

and

$$\varepsilon_K := \Delta - B \frac{x_0}{1 - x_0} > 0.$$

Forward invariance and Lemma 5 give $A_i(h; s) \leq -\varepsilon_K$ uniformly. The normalized discounted occupancy of h^0 is at least $1 - \delta$ because the initial state is h^0 with probability one. Keeping only this state in the expression for $\dot{z}_i^L$ yields

$$\dot{z}_i^L \leq -\varepsilon_K (1 + c^2) p_i^L (1 - p_i^L).$$

Therefore

$$\dot{p}_i^L \leq -\varepsilon_K (1 + c^2) [p_i^L (1 - p_i^L)]^2.$$

As in the one-step proof, a positive limiting value would imply an eventually strictly negative derivative bounded away from zero, which is impossible. Thus $p_i^L \to 0$.

For a separate evaluation initialized from the all-L prehistory, the one-period probability of first leaving LL is

$$q_K(s) = 1 - (1 - p_1^L(s))(1 - p_2^L(s)).$$

Before the first departure, the entire context remains h^0. Hence the probability of a departure by period t is at most $t q_K(s)$. Since $(c_t)_+ \leq \mathbf{1}\{g_t \neq LL\}$,

$$\mathbb{E}_s[(\kappa^{\delta})_+] \leq D_{\delta, T_{\mathrm{ev}}} q_K(s) \to 0.$$

Markov's inequality gives the exceedance-probability result. □

A.9 Proof of Corollary 2

Proof. Because

$$S_{K+1}(\delta) = S_K(\delta) + \delta^{K+1} > S_K(\delta),$$

the uncapped quantity $\Delta / [(B + \Delta) S_K(\delta)]$ decreases in K. Taking the minimum with one preserves weak monotonicity. Finally,

$$S_K(\delta) \to \frac{\delta}{1 - \delta},$$

so

$$p^{\star}_{K,\mathrm{suff}} \to \min \left\{ 1, \frac{\Delta(1 - \delta)}{(B + \Delta)\delta} \right\}.$$

The limit is strictly positive for fixed $\delta \in (0, 1)$ and $\Delta > 0$. □

A.10 Scope of the Population-Learning Results

The post-training propositions characterize exact population-gradient dynamics in the restricted policy classes defined in the main paper. They do not by themselves establish finite-sample convergence of GRPO, PPO, DPO, or REINFORCE implementations. Stochastic rollout noise, clipping, group-relative advantage normalization, finite learning rates, and full neural-network parameterizations require additional assumptions or separate analysis. Whether the economic mechanisms identified by the population theory persist under finite-sample post-training algorithms is left for future work.