# Speed of Intervention in Algorithmic Markets: Controlling Collusion and Stability

Click here for the latest version.

Tong Xie
Booth School of Business, University of Chicago, Chicago, Illinois 60637, USA. tong.xie@chicagobooth.edu
René Caldentey
Booth School of Business, University of Chicago, Chicago, Illinois 60637, USA. rene.caldentey@chicagobooth.edu
Martin Haugh
Department of Analytics & Operations, Imperial Business School, Imperial College London, London, UK. m.haugh@imperial.ac.uk

Algorithmic pricing is increasingly prevalent in online marketplaces, raising concerns that learning algorithms may sustain supra-competitive outcomes even without explicit communication. We study competing sellers who interact in a repeated Prisoner's Dilemma environment and delegate decisions to memoryless value-based learning algorithms that we call Q-value Reinforcers. A platform can mitigate collusive behavior by steering demand toward sellers choosing more competitive actions. We analyze the induced fluid dynamics and show that the effectiveness of the platform's intervention depends on three distinct forces. First, the intervention level affects equilibrium behavior: stronger intervention reduces the level of cooperation at an equilibrium fixed-point. Second, the learning architecture matters in two distinct ways: the algorithm's degree of greediness (how strongly it shifts toward the action with the higher current value) shapes the location and cooperativeness of the stationary outcomes, while its learning rate and sensitivity to payoff feedback determine whether the learning dynamics converge to those outcomes or become unstable. Third, the speed of implementation matters. Even when the target intervention level preserves local stability, an abrupt intervention can move inherited Q-values outside the fixed point's basin of attraction and induce persistent oscillations through a Hopf bifurcation that is subcritical in our numerical computations. Gradual implementation, by contrast, allows the learning dynamics to track the moving stable equilibrium and reach a more competitive stable outcome. The results show that platform regulation of algorithmic collusion requires choosing not only how much to intervene, but also how quickly to implement the intervention relative to the algorithms' learning dynamics.

## 1. Introduction

Algorithms are increasingly used by firms to set prices and make other strategic decisions in online marketplaces. Automated repricing tools allow firms to react to market conditions at a frequency and scale that would be difficult for human managers to match. These capabilities can improve responsiveness, but they also raise concerns about algorithm-enabled coordination.

Early enforcement cases illustrate one channel through which algorithms can implement explicit collusive agreements. The U.S. Department of Justice charged an e-commerce executive with conspiring to fix prices of posters sold through Amazon Marketplace and alleged that the conspirators used pricing algorithms and custom code to carry out the agreement (U.S. Department of Justice, Office of Public Affairs 2015, U.S. Department of Justice, Antitrust Division 2015). The UK Competition and Markets Authority

similarly found that two competing sellers agreed not to undercut one another's prices for posters and frames on Amazon's UK website and implemented the arrangement through automated repricing software (Competition and Markets Authority 2015). The concern is broader than explicit agreements, however. Controlled experiments show that reinforcement-learning pricing agents can sustain supra-competitive prices in repeated interaction even without direct communication among firms (Calvano et al. 2020), motivating policy discussions about whether algorithmic pricing and price transparency can heighten the risk of tacit coordination (OECD Competition Committee Secretariat 2017).

In this paper, we use *algorithmic collusion* to mean supra-competitive market outcomes that arise when firms delegate pricing or other strategic choices to algorithms that process market information, choose actions, and adapt to feedback. This definition is outcome-based: it does not require direct communication or an explicit agreement among firms. The concern is that algorithmic agents may react quickly and systematically to market feedback, making elevated prices or other supra-competitive outcomes easier to sustain (see Section 3.5 for further discussion).

Algorithmic collusion can arise through several channels. Rival-monitoring repricing tools can match competitors' prices quickly and punish undercutting; Assad et al. (2024), for example, document high-frequency price adjustments in the German retail gasoline market consistent with such dynamics. Common-vendor pricing tools create a second channel, when competitors rely on the same third-party system to generate recommendations, as alleged in recent disputes in rental housing and healthcare (U.S. Department of Justice 2024, Henry 2025). A third channel, and the one we study, is decentralized learning: independently operated algorithms may adapt to one another's behavior and settle into supra-competitive outcomes without a common vendor or explicit agreement.

We study how such outcomes emerge and how they can be mitigated when sellers delegate pricing decisions to learning algorithms. Our model combines three elements. First, sellers interact repeatedly in a Prisoner's Dilemma environment, which captures the tension between competitive behavior and mutually profitable cooperation. Second, each seller uses a value-based learning rule that we call a Q-value Reinforcer. The algorithm maintains a score for each action, updates these scores using realized payoffs, and maps score differences into randomized choices. Third, a platform can steer demand by reallocating exposure away from less competitive actions and toward more consumer-friendly actions, as in ranking, recommendation, buy-box, or featured-placement policies.

The analysis highlights a dynamic tradeoff in platform intervention. A higher intervention level can make the competitive action more attractive and reduce the equilibrium degree of cooperation. But intervention also changes the feedback loop through which algorithms learn. As the intervention level increases, the stable fixed point may lose stability through an oscillatory mode, generating persistent fluctuations rather than convergence. More importantly, even before this loss of local stability, the fixed point's basin of attraction—the set of learning states from which the dynamics converge back to it—can become small. An abrupt policy change can therefore move the current Q-values outside this basin, causing the market to enter persistent oscillations even when the post-intervention fixed point remains locally stable.

The central managerial implication is that platform intervention has both a level and a speed. The target level determines the incentives that sellers face after the policy is in place. The implementation speed determines whether the learning dynamics can track the moving stable equilibrium on the way to that target. Gradual implementation can preserve convergence by keeping the system inside the relevant basin of attraction; sudden implementation can destabilize learning and generate persistent cycles. Thus, the policy question we study is not only how strongly a platform should intervene, but also how quickly it should move to the desired intervention level.

The remainder of the paper is organized as follows. Section 2 positions the paper in the related literature. Section 3 introduces the model, the Q-value Reinforcer algorithms, the platform's demand-steering policy, the fluid approximation, and the measure of algorithmic collusion. Section 4 provides a non-technical roadmap of the analysis and main results. Section 5 studies fixed points and stability under a constant intervention level. Section 6 analyzes how intervention intensity affects equilibrium cooperation, local stability, and the basin of attraction of the system's stable fixed point equilibrium. Section 7 studies implementation speed and derives conditions under which gradual intervention preserves convergence. Section 9 concludes. Proofs and additional technical details are provided in the appendices.

## 2. Related Literature

Our paper contributes to the literature on algorithmic collusion. Legal and policy discussions by Harrington (2019) and the OECD (OECD Competition Committee Secretariat 2017) frame algorithmic pricing as a regulatory concern, and recent enforcement statements emphasize that algorithmically coordinated pricing may be unlawful even without direct communication (FTC 2024). Empirical work shows that algorithmic pricing is widely used in online marketplaces (Chen et al. 2016), while the dynamic pricing literature documents the operational value of adaptive pricing algorithms (Misra et al. 2019). At the same time, the evidence is nuanced: pricing algorithms can intensify undercutting rather than facilitate coordination (Spann et al. 2024). The relevant question is therefore not whether algorithms collude mechanically, but when particular learning rules, market environments, and institutional settings make supra-competitive outcomes likely to emerge and persist.

A growing theoretical and computational literature studies this question in repeated-market environments. In repeated Bertrand competition, Calvano et al. (2020) show that reinforcement-learning pricing agents can learn reward–punishment patterns that sustain supra-competitive prices. Subsequent work shows that the phenomenon depends on the pricing protocol, algorithmic architecture, and information structure: Klein (2021) studies Q-learning under sequential pricing, Lamba and Zhuk (2022) endogenizes firms' choices of pricing algorithms, Asker et al. (2022) analyzes how algorithm design shapes pricing outcomes, and Banchio and Mantegazza (2024) shows that collusion can emerge even between agents without explicit memory of play histories. Related work studies simpler learning architectures: Douglas et al. (2026) show that multi-armed bandit learners can generate "naive" algorithmic collusion in a repeated Prisoner's

Dilemma, and Hansen et al. (2021) show that independent UCB bandit algorithms can generate supra-competitive prices. Empirical evidence from the German gasoline market suggests that algorithmic pricing can raise margins in practice (Assad et al. 2024). Together, these papers establish algorithmic collusion as a learning-and-dynamics problem: market outcomes depend not only on payoffs and information, but also on how algorithms convert feedback into future actions.

Our analysis is also related to the broader literature on learning in games. Classical work on fictitious play and adaptive learning shows that equilibrium existence does not imply convergence of the realized path (Brown 1951, Robinson 1951, Shapley 1964, Fudenberg and Levine 1998). Evolutionary and population-game models reach a similar conclusion: simple continuous-time dynamics can converge, cycle, or exhibit more complex behavior depending on the game and the adjustment process (Taylor and Jonker 1978, Zeeman 1980, Hofbauer and Sigmund 1998). The no-regret literature draws a related distinction between convergence of time averages and convergence of play itself. Regret-based procedures can drive empirical play toward correlated equilibrium in time average (Hart and Mas-Colell 2000), while last-iterate behavior may cycle or fail to converge (Mertikopoulos et al. 2018, Daskalakis et al. 2018, Daskalakis and Panageas 2019, Cai et al. 2022). This distinction is central in our setting because persistent oscillations in prices, demand, or seller behavior are economically meaningful even when long-run averages appear well behaved.

The reinforcement-learning literature provides the algorithmic foundation for our model. Watkins and Dayan (1992) introduced Q-learning as a model-free method for estimating action values, and Littman (1994) extended reinforcement learning to multi-agent environments through Markov games. In multi-agent settings, each learner faces a nonstationary environment because other agents are adapting at the same time. Much of the multi-agent reinforcement-learning literature is therefore prescriptive, modifying the learning rule to recover convergence guarantees or equilibrium structure, as in Nash-Q learning (Hu and Wellman 2003), Friend-or-Foe Q-learning (Littman 2001), WoLF (Bowling and Veloso 2002), and correlated-Q learning (Greenwald and Hall 2003); see Zhang et al. (2021) for a survey. Our focus is different. We take the sellers' learning algorithms as given and ask how a platform can reshape the market environment in which those algorithms learn.

This connects the paper to work on platform governance and market-design interventions. Platforms can influence seller behavior through entry rules, quality standards, information provision, recommendations, rankings, and demand allocation (Teh 2022). Related work studies specific platform levers, including recommendations and demand steering (de Cornière and Taylor 2019) and search-ranking design (Dinerstein et al. 2018). A smaller literature studies interventions aimed directly at limiting algorithmic collusion. Johnson et al. (2023) show that ranking and buy-box rules can mitigate collusion by reallocating demand, while Brown and MacKay (2023) argue for restricting algorithms that condition directly on rivals' prices. The results of Banchio and Mantegazza (2024) and Douglas et al. (2026) suggest that such restrictions may be insufficient when collusive-looking outcomes arise from simpler forms of learning. Other work studies regulation through auditing, incentives, or planner interventions (Yang et al. 2023, Hartline et al. 2025,

Zhang et al. 2024, Brero et al. 2024). These papers characterize the form or level of intervention. We show that the path of implementation, and especially its speed, is itself a distinct design variable.

Methodologically, our approach uses stochastic approximation and nonlinear dynamics. The ODE method justifies studying stochastic learning algorithms through deterministic fluid limits (Borkar and Meyn 2000, Kushner and Yin 2003, Borkar 2008), and related methods have been applied to Q-learning in games (Leslie and Collins 2005, Gomes and Kowalczyk 2009, Wunder et al. 2010). Recent work also uses dynamical-systems tools to study phase transitions and oscillations in deterministic approximations of multi-agent learning (Leonardos and Piliouras 2022, Goll et al. 2025). We use this perspective to connect platform intervention to stability, equilibrium selection, and persistent oscillations.

Our contribution is to show that demand-steering intervention affects not only the equilibrium incentives faced by learning algorithms, but also the stability of the learning process itself. In a repeated Prisoner's Dilemma with Q-value Reinforcer algorithms, stronger intervention reduces the equilibrium level of cooperation but can also shrink the basin of attraction of the desired fixed point. As a result, an abrupt intervention can move the system outside this basin and induce persistent oscillations, whereas gradual implementation can preserve convergence by allowing the learning state to track the moving equilibrium. This identifies implementation speed as a platform-design lever alongside intervention intensity.

## 3. Model Setup

We consider a market in which two firms sell through a common online platform and compete for the same pool of consumers. Consumers arrive according to a homogeneous Poisson process with rate $\Lambda$. Sellers choose mixed actions over the binary action set $\mathbb{A} = \{C, D\}$. The action $C$ denotes *cooperation* and $D$ denotes *defection*. In the applications we have in mind, cooperation is the less consumer-friendly action—for example, a high price, low inventory, or low service quality—whereas defection is the more consumer-friendly action—for example, a low price, high inventory, or high service quality.

The firms interact repeatedly. Decisions are made at epochs $k = 1, 2, 3, \ldots$, with consecutive epochs separated by a period of length $\Delta > 0$. The number of consumers arriving in period $k$ is Poisson with mean $\Lambda\Delta$, independently across periods. At the beginning of period $k$, firm $i \in \{1, 2\}$ chooses a mixed action $\mathbf{p}_k^i = (a_k^i, 1 - a_k^i) \in \Sigma_{\mathbb{A}}$, where $a_k^i$ is the probability assigned to action $C$. Conditional on $\mathbf{p}_k^i$, firm $i$ applies this randomization independently across consumers arriving in the period. We write $\mathbf{a}_k = (a_k^1, a_k^2) \in [0, 1]^2$ for the resulting profile of cooperation probabilities.

Absent platform intervention, the expected per-customer payoffs are given in Table 1. Expected payoffs are normalized on a per-customer basis: the arrival process determines the number of payoff observations in a period, but not the entries of the stage-game matrix. The stage game has the Prisoner's Dilemma ordering $T > R > P > S > 0$. Mutual cooperation gives both firms the reward payoff $R$, while mutual defection gives both firms the punishment payoff $P$. If one firm defects while the other cooperates, the defector obtains the temptation payoff $T$ and the cooperator obtains the sucker payoff $S$. Thus, mutual cooperation is more profitable than mutual defection, but defection is privately optimal against either rival

| | | Firm 2: C | Firm 2: D |
|---|---|---|---|
| Firm 1 | C | $R$, $R$ | $S$, $T$ |
| | D | $T$, $S$ | $P$, $P$ |

**Table 1 Baseline expected per-customer payoff matrix without platform intervention. The first entry is firm 1's payoff and the second entry is firm 2's payoff.**

action. The model therefore captures the basic tension between a jointly attractive cooperative outcome and individual incentives to choose the more competitive action.

### 3.1. Platform Intervention

Platforms have both economic and governance reasons to shape seller behavior without directly setting prices or prescribing actions. High prices, low service quality, or poor fulfillment can reduce consumer surplus, weaken trust in the marketplace, lower repeat traffic, and increase regulatory scrutiny. Because platforms often earn revenue from transactions, advertising, subscriptions, or long-run consumer participation, they may benefit from maintaining a competitive and reliable marketplace even when individual sellers prefer less consumer-friendly actions.

A platform can therefore use demand-steering tools to reward more competitive offers. It can rank lower-priced offers more prominently in search results, allocate a buy box or featured offer more often to sellers with lower effective prices or better fulfillment quality, adjust recommendation rules, or modify featured-placement criteria. These tools shift consumer attention and demand, and therefore change the payoff feedback received by sellers' learning algorithms. Such mechanisms are common in platform markets and have been studied in search design, recommendations, platform governance, and algorithmic pricing (de Cornière and Taylor 2019, Dinerstein et al. 2018, Teh 2022, Johnson et al. 2023). For example, Amazon's Featured Offer mechanism conditions prominence on attributes related to customer experience, including total price and fulfillment (Amazon Seller Central 2026).

In our setting, the platform intervention favors the consumer-friendly action $D$ relative to the less consumer-friendly action $C$. We represent the strength of this demand-allocation policy in period $k$ by a nonnegative intervention level $\delta_k$. A larger $\delta_k$ corresponds to a stronger reallocation of exposure toward sellers choosing $D$. We impose $\delta_k \in [0, S]$ so that all possible payoffs remain nonnegative.

Under intervention, the expected per-customer payoff matrix is given in Table 2. When one firm chooses $D$ and the other chooses $C$, the firm choosing $D$ receives additional exposure while the firm choosing $C$ loses exposure. This induces the payoff adjustments $T \mapsto T + \delta_k$ and $S \mapsto S - \delta_k$. For any $\delta_k \in [0, S]$, the adjusted payoffs satisfy

$$T + \delta_k > R > P > S - \delta_k,$$

so the intervention changes the payoff feedback that drives learning while preserving the Prisoner's Dilemma ordering. We interpret $\delta_k$ as a reduced-form per-customer expected payoff wedge induced by

| | | Firm 2: C | Firm 2: D |
|---|---|---|---|
| Firm 1 | C | $R,\ R$ | $S-\delta_k,\ T+\delta_k$ |
| | D | $T+\delta_k,\ S-\delta_k$ | $P,\ P$ |

**Table 2 Expected per-customer payoff matrix in period $k$ under platform intervention. The first entry is firm 1's payoff and the second entry is firm 2's payoff.**

the platform's allocation rule. It may reflect changes in exposure, demand, or conversion probabilities. Importantly, the intervention does not require the platform to observe latent payoffs or learning states; it depends only on observable seller choices and can be implemented through standard ranking or allocation rules. This reduced-form specification keeps the intervention analytically tractable while retaining the key economic effect: demand is shifted toward the more consumer-friendly action.

The intervention level evolves according to

$$\delta_{k+1} = \delta_k + \mathbb{I}(k, \delta_k)\,\Delta, \tag{1}$$

where $\mathbb{I}(k, \delta_k)$ denotes the intervention speed in period $k$. The level $\delta_k$ determines the intensity of intervention, while $\mathbb{I}(k, \delta_k)$ determines the speed of implementation. Large positive values of $\mathbb{I}(k, \delta_k)$ represent abrupt increases in intervention; small positive values represent gradual implementation. We assume the path remains admissible so that $\delta_k \in [0, S]$ for all $k$.

For a cooperation-probability profile $\mathbf{a} = (a^1, a^2)$ and intervention level $\delta$, let

$$\mathbf{r}^i(\mathbf{a}, \delta) = \big(r^i_{\mathrm{C}}(\mathbf{a}, \delta), r^i_{\mathrm{D}}(\mathbf{a}, \delta)\big) \in \mathbb{R}^2$$

denote firm $i$'s random per-customer payoff-feedback vector. For $i, j \in \{1, 2\}$ and $i \neq j$,

$$\mathbb{E}\big[r^i_{\mathrm{C}}(\mathbf{a}, \delta)\big] = R\,a^j + (S-\delta)(1-a^j), \qquad \mathbb{E}\big[r^i_{\mathrm{D}}(\mathbf{a}, \delta)\big] = (T+\delta)a^j + P(1-a^j). \tag{2}$$

Thus, the per-customer cooperation-minus-defection expected payoff difference is $\mathbb{E}\big[r^i_{\mathrm{C}}(\mathbf{a}, \delta)\big] - \mathbb{E}\big[r^i_{\mathrm{D}}(\mathbf{a}, \delta)\big] = Aa^j + B - \delta$, where $A := (R-T) - (S-P)$ and $B := S-P$.

### 3.2. Learning Algorithms

Firms are not assumed to play equilibrium strategies of the repeated game. Instead, each firm delegates decisions to a value-based learning rule that we call a Q-value Reinforcer (Banchio and Mantegazza 2024). The algorithm maintains one value, or score, for each action. Firm $i$'s state at decision epoch $k$ is

$$\mathbf{s}^i_k = (Q^i_{\mathrm{C},k}, Q^i_{\mathrm{D},k}),$$

where $Q^i_{b,k}$ is the current estimated discounted value of action $b \in \{C, D\}$. The current Q-values determine the firm's mixed action

$$\mathbf{p}^i_k = \boldsymbol{\pi}(\mathbf{s}^i_k) = \big(\boldsymbol{\pi}_{\mathrm{C}}(\mathbf{s}^i_k), 1 - \boldsymbol{\pi}_{\mathrm{C}}(\mathbf{s}^i_k)\big) = (a^i_k, 1-a^i_k).$$

The Q-value Reinforcer is memoryless: its current behavior depends only on its current Q-values, not on an explicit history of past play, observed rival histories, or a separate market state. This restriction

is useful for our purposes because it removes the standard repeated-game channel of history-dependent rewards and punishments. Any persistent cooperation in the model must therefore arise from the learning dynamics themselves rather than from a designed contingent strategy.

Let $\mathbf{r}_k^i = (r_{\text{C},k}^i, r_{\text{D},k}^i)$ denote the observed normalized payoff-feedback vector realized by firm $i$ in period $k$.[1] Suppressing firm and time superscripts, let $\mathbf{1}_2$ denote the two-dimensional vector of ones and define

$$\overrightarrow{\max}\{\mathbf{s}\} := \max\{Q_{\text{C}}, Q_{\text{D}}\}\mathbf{1}_2.$$

The algorithm updates its Q-values toward the Bellman-like target

$$\mathcal{T}_{\text{QR}}(\mathbf{s}, \mathbf{r}) = \mathbf{r} + \gamma\, \overrightarrow{\max}\{\mathbf{s}\}, \tag{3}$$

where $\gamma \in [0, 1)$ is a discount factor. Equivalently, $\mathcal{T}_{\text{QR},b}(\mathbf{s}, \mathbf{r}) = r_b + \gamma \max\{Q_{\text{C}}, Q_{\text{D}}\}$ for $b \in \{C, D\}$.

The state update is

$$\mathbf{s}_{k+1}^i = \mathbf{s}_k^i + \boldsymbol{\alpha}(\mathbf{s}_k^i) \odot \left[\mathcal{T}_{\text{QR}}(\mathbf{s}_k^i, \mathbf{r}_k^i) - \mathbf{s}_k^i\right], \tag{4}$$

where $\odot$ denotes the Hadamard product and $\boldsymbol{\alpha}(\mathbf{s}_k^i)$ is the component-wise learning-rate vector defined below. Thus, each Q-value moves part of the way toward its current payoff-plus-continuation target.

**Remark 1** *The term Q-value* Reinforcer *is used to describe the reduced-form value-based updating rule rather than a full state-space implementation of tabular Q-learning. The algorithm has no explicit state and no history-dependent strategy; it updates two action values using payoff feedback. This specification isolates the feedback loop between current action values, randomized behavior, payoff observations, and subsequent value updates.*

Given a state $\mathbf{s} = (Q_{\text{C}}, Q_{\text{D}})$, the firm's probability of cooperation is the piecewise-linear $\epsilon$-greedy response

$$\boldsymbol{\pi}_{\text{C}}(\mathbf{s}) = \mathcal{M}_{\epsilon,\tau}(\mathbf{s}) := \begin{cases} \epsilon, & \text{if } Q_{\text{C}} - Q_{\text{D}} \le -\Delta_{\epsilon,\tau}, \\ \tau(Q_{\text{C}} - Q_{\text{D}}) + \frac{1}{2}, & \text{if } |Q_{\text{C}} - Q_{\text{D}}| < \Delta_{\epsilon,\tau}, \\ 1 - \epsilon, & \text{if } Q_{\text{C}} - Q_{\text{D}} \ge \Delta_{\epsilon,\tau}, \end{cases} \qquad \text{where } \Delta_{\epsilon,\tau} := \tfrac{1-2\epsilon}{2\tau}. \tag{5}$$

Here $\epsilon \in (0, 1/2)$ is the minimum exploration probability and $\tau > 0$ is the policy's responsiveness parameter: larger values make the action probability more responsive to Q-value differences. The threshold $\Delta_{\epsilon,\tau}$ is the Q-value gap at which the policy saturates. In particular, when $|Q_{\text{C}} - Q_{\text{D}}| \ge \Delta_{\epsilon,\tau}$, the higher-valued action is chosen with probability $1 - \epsilon$ and the other action with probability $\epsilon$. Within the linear region, small changes in the Q-value difference translate smoothly into changes in the probability of cooperation. When no confusion arises, we write $\mathcal{M}(\mathbf{s})$ for $\mathcal{M}_{\epsilon,\tau}(\mathbf{s})$, so that $\boldsymbol{\pi}(\mathbf{s}) = (\mathcal{M}(\mathbf{s}), 1 - \mathcal{M}(\mathbf{s}))$.

**Remark 2 (Softmax interpretation)** *The policy map $\mathcal{M}$ in (5) can be viewed as a piecewise-linear local approximation to a softmax choice rule. The canonical softmax probability of choosing action $C$ takes the form*

$$\mathcal{M}_{sm}(\mathbf{s}) = \frac{\exp\{\beta Q_{\text{C}}\}}{\exp\{\beta Q_{\text{C}}\} + \exp\{\beta Q_{\text{D}}\}} = \frac{1}{1 + \exp\{-\beta(Q_{\text{C}} - Q_{\text{D}})\}},$$

[1] If action $b$ is not taken by firm $i$ in period $k$, no payoff feedback for that action is observed, and the $b$-component of the Q-value update is skipped. The frequency-weighted learning rates introduced below give the corresponding expected update after averaging over the mixed action.

*for some softmax sensitivity parameter $\beta > 0$.*

*Around $Q_{\text{C}} - Q_{\text{D}} = 0$, this sigmoid is approximately linear with slope $\beta/4$. After setting the responsiveness parameter $\tau = \beta/4$, the map in (5) keeps this local slope and caps the tails at $\epsilon$ and $1-\epsilon$, yielding a tractable soft $\epsilon$-greedy rule with explicit exploration bounds.*

Finally, learning rates depend on the action probabilities induced by current Q-values. Since a period may contain several consumer arrivals, an action assigned a higher probability is expected to be taken more often and to generate more payoff feedback. We capture this frequency effect by defining

$$\boldsymbol{\alpha}(\mathbf{s}) = \alpha\, \boldsymbol{\pi}(\mathbf{s})^{\theta}, \tag{6}$$

where $\alpha \in (0, 1]$ and $\theta \geq 0$ is applied component-wise. Thus

$$\boldsymbol{\pi}(\mathbf{s})^{\theta} = \left(\boldsymbol{\pi}_{\text{C}}(\mathbf{s})^{\theta}, (1 - \boldsymbol{\pi}_{\text{C}}(\mathbf{s}))^{\theta}\right), \qquad \boldsymbol{\pi}(\mathbf{s})^{0} \equiv (1, 1).$$

For $\theta > 0$, the update rate of each Q-value is increasing in the probability assigned to the corresponding action; larger values of $\theta$ make update rates more sensitive to play frequencies. The case $\theta = 0$ gives a constant learning-rate vector. The case $\theta = 1$ corresponds, in expectation, to the usual asynchronous Q-learning frequency weighting: an action's Q-value is updated at a rate proportional to how often that action is taken. This formulation captures action-dependent feedback frequency without adding state variables, thus keeping the subsequent dynamical-system analysis tractable.

### 3.3. Fluid Approximation

We analyze the learning dynamics through a continuous-time fluid approximation of the discrete-time updating process. Embed the model in a sequence indexed by the period length $\Delta > 0$ and scale the learning rate as

$$\boldsymbol{\alpha}_{\Delta}(\mathbf{s}) = \boldsymbol{\alpha}(\mathbf{s})\Delta. \tag{7}$$

Let $\mathbf{s}_k^{\Delta}$, $\delta_k^{\Delta}$, $\mathbf{p}_k^{\Delta} = \boldsymbol{\pi}(\mathbf{s}_k^{\Delta})$, and $\mathbf{r}_k^{\Delta}$ denote the state, intervention level, mixed action, and payoff-feedback vector. The period-normalized update has average drift

$$\overline{\mathbb{H}}(\mathbf{s}, \delta) := \mathbb{E}_{\delta}\left[\boldsymbol{\alpha}(\mathbf{s}) \odot (\mathcal{T}_{\text{QR}}(\mathbf{s}, \mathbf{r}) - \mathbf{s})\right],$$

where the expectation is over payoff realizations and the action randomization induced by $\mathbf{p} = \boldsymbol{\pi}(\mathbf{s})$, with the payoff-feedback distribution evaluated at intervention level $\delta$. Under the boundedness and regularity conditions stated in Appendix C, standard stochastic-approximation arguments imply that, as $\Delta \downarrow 0$, the piecewise-linear interpolation of the discrete-time process converges weakly on finite time horizons to the ODE

$$\dot{\mathbf{s}}_t = \overline{\mathbb{H}}(\mathbf{s}_t, \delta_t), \qquad \mathbf{s}(0) = \mathbf{s}_0. \tag{8}$$

Appendix C provides the formal derivation, including the treatment of the piecewise-smooth drift.

We interpret (8) as the mean-field, or fluid, approximation of the learning process. When decision periods are short and updates are small, random payoff and action fluctuations average out, and the system follows a deterministic drift. The ODE is therefore not meant to replace the underlying discrete algorithm, but to describe its average motion under small updates. The stability and bifurcation analysis below is based on this ODE representation.

### 3.4. Fluid Dynamics of Q-Value Reinforcer Algorithms

We now derive the drift for Q-value REINFORCER algorithms. Let $\mathbf{s} = (\mathbf{s}^1, \mathbf{s}^2)$, where $\mathbf{s}^i = (Q^i_{\mathrm{C}}, Q^i_{\mathrm{D}})$ is firm $i$'s internal state, and let $\boldsymbol{\pi}(\mathbf{s}) = (\boldsymbol{\pi}(\mathbf{s}^1), \boldsymbol{\pi}(\mathbf{s}^2))$ be the mixed strategy profile. Combining (1)–(8), the Q-value dynamics are, for $i = 1, 2$,

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{s}^i_t}{\mathrm{d}t} &= \overline{\mathbb{H}}^i(\mathbf{s}_t, \delta_t) = \alpha\, \boldsymbol{\pi}(\mathbf{s}^i_t)^\theta \odot \Big(\mathbb{E}[\mathbf{r}^i(\boldsymbol{\pi}(\mathbf{s}_t), \delta_t)] + \gamma\, \overrightarrow{\max}\{\mathbf{s}^i_t\} - \mathbf{s}^i_t\Big), && \text{(Q-Learner Update)}\\
\boldsymbol{\pi}(\mathbf{s}^i_t) &= \big(\mathcal{M}(\mathbf{s}^i_t), 1 - \mathcal{M}(\mathbf{s}^i_t)\big), && \text{(Action Update)}\\
\frac{\mathrm{d}\delta_t}{\mathrm{d}t} &= \mathbb{I}(t, \delta_t) && \text{(Intervention Policy)}
\end{aligned}$$

where the dependence on the intervention level $\delta_t$ is now explicit. The Q-Learner Update moves each firm's Q-values toward the corresponding Bellman-like target, the Action Update maps Q-values into mixed actions, and the Intervention Policy specifies the law of motion for the platform's intervention level. Together, these equations define a closed-loop dynamical system for $(\mathbf{s}_t, \delta_t)$.

This representation separates learning, choice, and platform control: the Q-Learner Update evolves internal scores, the Action Update maps scores into observable behavior, and the Intervention Policy governs the platform's payoff wedge.

It is useful to express the system directly in terms of Q-values. Denote the firms' cooperation probabilities $\big(\boldsymbol{\pi}^1_{\mathrm{C}}(\mathbf{s}^1_t), \boldsymbol{\pi}^2_{\mathrm{C}}(\mathbf{s}^2_t)\big)$ by $(x_t, y_t)$. Then, for $\mathbf{Q}(t) = (Q^1_{\mathrm{C}}(t), Q^1_{\mathrm{D}}(t), Q^2_{\mathrm{C}}(t), Q^2_{\mathrm{D}}(t))$, (Q-Learner Update) becomes

$$\begin{aligned}
\frac{\mathrm{d}Q^1_{\mathrm{C}}(t)}{\mathrm{d}t} &= \alpha x_t^\theta \left(y_t R + (1 - y_t)(S - \delta_t) + \gamma \max\{Q^1_{\mathrm{C}}(t), Q^1_{\mathrm{D}}(t)\} - Q^1_{\mathrm{C}}(t)\right),\\
\frac{\mathrm{d}Q^1_{\mathrm{D}}(t)}{\mathrm{d}t} &= \alpha (1 - x_t)^\theta \left(y_t (T + \delta_t) + (1 - y_t) P + \gamma \max\{Q^1_{\mathrm{C}}(t), Q^1_{\mathrm{D}}(t)\} - Q^1_{\mathrm{D}}(t)\right),\\
\frac{\mathrm{d}Q^2_{\mathrm{C}}(t)}{\mathrm{d}t} &= \alpha y_t^\theta \left(x_t R + (1 - x_t)(S - \delta_t) + \gamma \max\{Q^2_{\mathrm{C}}(t), Q^2_{\mathrm{D}}(t)\} - Q^2_{\mathrm{C}}(t)\right),\\
\frac{\mathrm{d}Q^2_{\mathrm{D}}(t)}{\mathrm{d}t} &= \alpha (1 - y_t)^\theta \left(x_t (T + \delta_t) + (1 - x_t) P + \gamma \max\{Q^2_{\mathrm{C}}(t), Q^2_{\mathrm{D}}(t)\} - Q^2_{\mathrm{D}}(t)\right).
\end{aligned} \tag{9}$$

Using (5), the action update is

$$\begin{aligned}
x(t) &= \min\left\{1 - \epsilon,\, \max\left\{\epsilon,\, \tfrac{1}{2} + \tau(Q^1_{\mathrm{C}}(t) - Q^1_{\mathrm{D}}(t))\right\}\right\},\\
y(t) &= \min\left\{1 - \epsilon,\, \max\left\{\epsilon,\, \tfrac{1}{2} + \tau(Q^2_{\mathrm{C}}(t) - Q^2_{\mathrm{D}}(t))\right\}\right\}.
\end{aligned} \tag{10}$$

For future reference, we combine (9) and (10) and write the system as

$$\dot{\mathbf{Q}}_t = \mathbb{F}(\mathbf{Q}_t, \delta_t), \tag{11}$$

where $\mathbb{F}$ incorporates the action probabilities in (10) and the payoff effects of intervention. We use (11) both as a frozen-$\delta$ system for studying intervention intensity and as a time-varying system, with $\dot{\delta}_t = \mathbb{I}(t, \delta_t)$, for studying whether the learning state can track stable fixed points as the platform moves the intervention level.

### 3.5. Defining and Measuring Algorithmic Collusion

We conclude by clarifying what we mean by algorithmic collusion in this model. In both legal and economic usage, collusion is typically associated with coordination among distinct decision makers. From a legal

perspective, the canonical concern is an agreement or mutual understanding among competitors that restricts competition, such as price fixing, bid rigging, or market allocation. From an economic perspective, collusion is often modeled as coordination that generates supra-competitive outcomes and is sustained by intertemporal incentives: firms refrain from the one-shot dominant action because deviations trigger future punishments or the loss of future cooperative rents (Fudenberg and Maskin 1986, Harrington 2019, OECD Competition Committee Secretariat 2017). Thus, collusion is not merely the observation of frequent cooperation; it is usually associated with a strategic arrangement in which agents understand that current conduct affects rivals' future behavior.

Our setting calls for a different, outcome-based interpretation. The algorithms we study do not communicate, explicitly model rivals, condition on market histories, or contain a state variable through which current actions can be used to trigger future rewards or punishments. Each algorithm updates Q-values from payoff feedback and maps those values into actions through a fixed response rule. As a result, algorithmic collusion in our model is not a strategic commitment to coordinate away from the dominant action of defection. It is a property of the market outcomes generated by learning algorithms.

We operationalize this idea using the firms' cooperation probabilities. If both algorithms assign cooperation only because of forced exploration, then $x_t = y_t = \epsilon$, and the probability that both firms assign $C$ to a given consumer arrival is $\epsilon^2$. We report the square root of the resulting normalized joint-cooperation probability:

$$\mathcal{C}(t) := \frac{\sqrt{x_t y_t}}{\epsilon}. \qquad \text{(Normalized Collusion Measure)}$$

Equivalently, $\mathcal{C}(t)^2 = x_t y_t / \epsilon^2$ is the probability of mutual cooperation normalized by the forced-exploration benchmark. The square-root normalization makes $\mathcal{C}(t)$ a per-firm cooperation-level measure: in the symmetric case, where $x_t = y_t$, it reduces to $\mathcal{C}(t) = x_t/\epsilon$. Thus, $\mathcal{C}(t) = 1$ corresponds to the forced-exploration benchmark, while $\mathcal{C}(t) > 1$ indicates cooperation probabilities above that benchmark. Since $x_t, y_t \in [\epsilon, 1-\epsilon]$, the measure satisfies $\mathcal{C}(t) \in [1, (1-\epsilon)/\epsilon]$. At a stable fixed point $(x^*, y^*)$, the long-run collusion level is $\mathcal{C}^* = \sqrt{x^* y^*}/\epsilon$; for non-convergent dynamics, such as limit cycles, the analogous object is the long-run time average of $\mathcal{C}(t)$. Because the measure is normalized by $\epsilon$, it is a relative measure of cooperation, not an absolute measure of welfare loss or price increases.

This definition differs from reward-and-punishment mechanisms studied in some Q-learning models. In Calvano et al. (2020), high-price outcomes can be supported by market-history states: a deviation changes the state reached in subsequent periods, and the rival's learned policy may prescribe lower prices before the system returns to the high-price path. Because Q-values include discounted continuation payoffs, the short-run gain from deviation can be offset by future losses. Our algorithms do not have such history-dependent states, so the mechanism is different.

Our definition also differs from Q-value-based criteria such as Banchio and Mantegazza (2024), under which an algorithm is non-collusive once it has learned that defection has a higher value than cooperation, i.e., $Q_{\text{D}} > Q_{\text{C}}$. In an $\epsilon$-greedy environment, however, internal Q-values need not coincide with realized

actions. An algorithm may rank defection above cooperation and still choose cooperation with probability above $\epsilon$ if the Q-value gap is not large enough to saturate the response rule. Measuring collusion through $\mathcal{C}(t)$ therefore aligns the definition with observable behavior rather than internal (normally unobservable) scores alone.

Taken together, these distinctions clarify the role of $\mathcal{C}(t)$. The measure does not assert the existence of an agreement, a punishment scheme, or a legal violation. It records supra-exploratory cooperative behavior generated by the learning dynamics. This is the outcome-based sense in which we use the term algorithmic collusion in the analysis below.

## 4. Roadmap of Analysis and Main Results

Before turning to the formal analysis, we provide a roadmap of the main results and the logic connecting them. As described in Section 3, the analysis has two linked components:

$$\dot{\mathbf{Q}}_t = \mathbb{F}(\mathbf{Q}_t, \delta_t) \qquad \text{and} \qquad \dot{\delta}_t = \mathbb{I}(t, \delta_t).$$

The first equation describes the Q-value REINFORCER dynamics under a given intervention level; the second describes the platform's implementation path. The platform's problem is therefore not only to choose a target intervention level $\delta^*$, but also to reach that target without pushing the inherited Q-values into a region with different long-run behavior.

Section 5 studies the Q-value dynamics under a fixed intervention level, $\delta_t = \delta$. We first establish boundedness and the existence of a compact global attractor (Propositions 1–2), so the long-run analysis is well posed. We then characterize the symmetric interior fixed point induced by a fixed intervention level (Proposition 3). Existence of a fixed point, however, is not the same as convergence to that point: the same fixed intervention level can support a locally stable fixed point and nearby oscillating behavior. Figure 1 illustrates this distinction, and Theorem 1 gives the local stability condition.

Section 6 treats $\delta$ as a moving parameter. Stronger intervention lowers the equilibrium cooperation probability $a^*(\delta)$, but it can also reduce the local stability margin. Proposition 4 shows that, along the symmetric interior branch, there is at most one threshold $\delta_c$ at which local stability changes. At such a threshold the loss of stability occurs through an oscillatory Hopf mode. In the subcritical case, the fixed point may remain locally stable for $\delta < \delta_c$ and the local basin around it shrinks; Proposition 6 makes this basin geometry explicit.

Section 7 returns to the full control problem with a time-varying intervention path. The key implication of the Hopf analysis is that a target $\delta^* < \delta_c$ may be locally stable and still be risky if implemented abruptly, because the learning state may fail to remain inside the moving basin of attraction. Proposition 7 gives a sufficient local bound on implementation speed, and Corollary 1 converts this bound into a rollout-time lower bound. Thus intervention levels and implementation speed are distinct policy instruments: the target level determines post-intervention payoffs and the local stability margin, while the path $\delta_t$ determines whether the learning dynamics track the moving fixed point as those payoffs are introduced.

## 5. Equilibrium and Stability with a Constant Intervention Level

We first analyze the Q-value dynamics in (11) under a fixed intervention level $\delta$, or equivalently with $\mathbb{I}(t,\delta)=0$. The no-intervention benchmark is obtained by setting $\delta=0$. Time-varying interventions are studied later in Section 7.

The section proceeds in four steps. We first show that all trajectories are globally bounded and eventually enter a compact attracting set. We then characterize the fixed points and the cooperation probabilities they induce. Next, we use the constant-learning-rate case, i.e. $\theta=0$ in (6), to illustrate the feedback structure of the dynamics. Finally, we study local stability of the symmetric interior equilibrium. The main distinction is that the responsiveness parameter $\tau$ determines the location and structure of fixed points, whereas the learning-rate sensitivity parameter $\theta$ affects their stability but not their location. The baseline learning rate $\alpha$ rescales time in the fixed-$\delta$ system: it changes the speed of motion along ODE trajectories, but not the fixed points or the signs of their local stability eigenvalues. The role of $\alpha$ becomes central again when we study time-varying interventions in Section 7.

### 5.1. Global Boundedness and the Global Attractor

Before characterizing particular outcomes, we establish that the fixed-$\delta$ system is point dissipative.

**Proposition 1 (Point Dissipativity)** *Fix $\delta\in[0,S]$. The set*

$$\Omega_\delta := \left\{\mathbf{Q}\in\mathbb{R}^4 : \|\mathbf{Q}\|_\infty \le \frac{T+\delta}{1-\gamma}\right\}$$

*is positively invariant and absorbing for the dynamical system defined by (11) with this fixed intervention level. That is, if $\mathbf{Q}(0)\in\Omega_\delta$, then $\mathbf{Q}(t)\in\Omega_\delta$ for all $t\ge 0$; and for any initial condition $\mathbf{Q}(0)\in\mathbb{R}^4$, there exists a time $t_0$ such that $\mathbf{Q}(t)\in\Omega_\delta$ for all $t\ge t_0$.*

The bound follows from bounded stage payoffs and discounting. Since the largest possible stage payoff is $T+\delta$, any sufficiently large Q-value has a Bellman target below its current value and therefore drifts downward; sufficiently negative Q-values drift upward. Proposition 1 implies that instability in this model cannot take the form of exploding Q-values.

The next result collects all possible long-run behavior in a compact invariant set. A global attractor need not be a single equilibrium. It may contain multiple stable fixed points, periodic orbits, or other bounded recurrent dynamics.

**Proposition 2 (Existence of a Global Attractor)** *For each fixed $\delta\in[0,S]$, the dynamical system defined by (11) has a global attractor $\mathcal{A}_\delta\subset\Omega_\delta\subset\mathbb{R}^4$.*

To interpret the attracting dynamics, it is useful to introduce two pairs of coordinates. The firms' cooperation–defection Q-value gaps are

$$\Delta_1(t) := Q_{\text{C}}^1(t) - Q_{\text{D}}^1(t), \qquad \Delta_2(t) := Q_{\text{C}}^2(t) - Q_{\text{D}}^2(t), \tag{Q-Value Gaps}$$

and the inter-firm Q-value differences are

$$v_1(t) := Q_{\mathrm{C}}^1(t) - Q_{\mathrm{C}}^2(t), \qquad v_2(t) := Q_{\mathrm{D}}^1(t) - Q_{\mathrm{D}}^2(t). \qquad \text{(Antisymmetric Differences)}$$

The gaps $(\Delta_1, \Delta_2)$ determine each firm's action probabilities according to the $\epsilon$-greedy rule in (5) and are therefore the natural coordinates for fixed-point analysis. The variables $(v_1, v_2)$ are the raw antisymmetric difference variables between firms. At a Hopf point, the critical eigenspace lies in this antisymmetric plane; Section 6 introduces the corresponding local Hopf coordinates.

Figure 1 illustrates this point for the baseline parameters. Under the same fixed intervention level, different initial Q-values can converge either to a stable symmetric fixed point or to a stable limit cycle. The main analysis below therefore separates fixed-point existence from local stability and basin geometry.

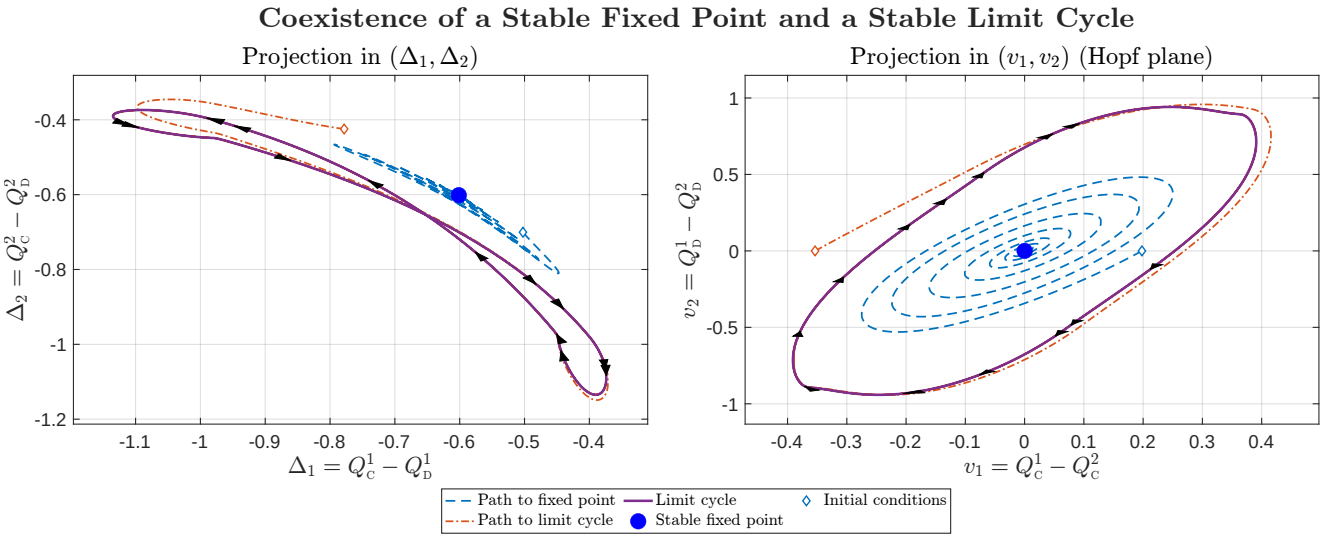



**Figure 1** **Bistability under a fixed intervention level. One trajectory converges to the stable symmetric fixed point, while another converges to a stable limit cycle.**

We next characterize the fixed points contained in $\mathcal{A}_\delta$ and then study their stability.

### 5.2. Equilibria

We classify fixed points by the cooperation probabilities induced by the policy map. Interior fixed points correspond to cooperation above the forced-exploration level. Mutual-boundary fixed points correspond to the non-collusive benchmark, in which both firms cooperate only through forced exploration. One-sided-boundary fixed points place one firm at the forced-exploration lower bound and the other in the interior of the policy map.

**Definition 1** *Fix $\delta \in [0, S]$. A fixed point of the dynamical system $\dot{\mathbf{Q}}_t = \mathbb{F}(\mathbf{Q}_t, \delta)$ is a vector $\mathbf{Q}^*(\delta) = (Q_{\mathrm{C}}^{1*}, Q_{\mathrm{D}}^{1*}, Q_{\mathrm{C}}^{2*}, Q_{\mathrm{D}}^{2*}) \in \mathbb{R}^4$ such that $\mathbb{F}(\mathbf{Q}^*(\delta), \delta) = 0$. Let $x^*(\delta) = \mathcal{M}(Q_{\mathrm{C}}^{1*}, Q_{\mathrm{D}}^{1*})$ and $y^*(\delta) = \mathcal{M}(Q_{\mathrm{C}}^{2*}, Q_{\mathrm{D}}^{2*})$ denote the induced cooperation probabilities. We say that $\mathbf{Q}^*(\delta)$ is an interior fixed point, or* collusive outcome*, if $x^*(\delta), y^*(\delta) \in (\epsilon, 1-\epsilon)$. It is a one-sided-boundary fixed point, or* partially-collusive outcome*, if either $x^*(\delta) = \epsilon$ and $y^*(\delta) > \epsilon$, or $y^*(\delta) = \epsilon$ and $x^*(\delta) > \epsilon$. It is a mutual-boundary fixed point, or* non-collusive outcome*, if $x^*(\delta) = y^*(\delta) = \epsilon$.*

We first record a useful property of all fixed points.

**Lemma 1** *At any fixed point* $\mathbf{Q}^*$*, the Q-value for defection strictly exceeds the Q-value for cooperation:* $Q_{\text{D}}^{i*} > Q_{\text{C}}^{i*}$ *for* $i = 1, 2$.

This ranking reflects the Prisoner's Dilemma payoff structure: defection is the dominant action. Importantly, however, $Q_{\text{D}}^{i*} > Q_{\text{C}}^{i*}$ does not imply that firm $i$ cooperates only with probability $\epsilon$. If the Q-value gap is not large enough to saturate the policy map, the algorithm can still assign cooperation an interior probability.

At a fixed point, the right-hand side of each equation in (9) is zero. Because $\epsilon > 0$, every action is chosen with positive probability under the policy map, so each component of the learning-rate vector $\boldsymbol{\alpha}(\mathbf{s}) = \alpha \boldsymbol{\pi}(\mathbf{s})^{\theta}$ is strictly positive. Hence each Q-value must equal its payoff-plus-continuation target:

$$Q_b^{i*} = \mathbb{E}[r_b^i(\boldsymbol{\pi}(\mathbf{s}^*), \delta)] + \gamma \max\{Q_{\text{C}}^{i*}, Q_{\text{D}}^{i*}\}, \qquad b \in \{C, D\},\ i = 1, 2.$$

Thus the parameters $\alpha$ and $\theta$ do not enter the fixed-point equations; they affect the speed and local stability of the dynamics, but not the location of fixed points.

By Lemma 1, the maximum in the continuation term is $Q_{\text{D}}^{i*}$ for each firm. Subtracting the defection equation from the cooperation equation for each firm gives

$$\Delta_1^* = A y^* + B - \delta, \qquad \Delta_2^* = A x^* + B - \delta.$$

Moreover, Lemma 1 implies $\Delta_i^* < 0$, so the upper cooperation cap $1 - \epsilon$ in (10) cannot bind. The only possible policy boundary is therefore the forced-exploration lower bound $\epsilon$. Combining these observations, any fixed point can be characterized by cooperation probabilities $(x^*, y^*)$ and Q-value gaps $(\Delta_1^*, \Delta_2^*)$ satisfying

$$\begin{aligned} A\, y^* + B - \delta - \Delta_1^* &= 0, \qquad & x^* &= \epsilon + \left(\tfrac{1}{2} + \tau \Delta_1^* - \epsilon\right)^+, \\ A\, x^* + B - \delta - \Delta_2^* &= 0, \qquad & y^* &= \epsilon + \left(\tfrac{1}{2} + \tau \Delta_2^* - \epsilon\right)^+, \end{aligned} \qquad \text{(Equilibrium Conditions)}$$

where $z^+ := \max\{z, 0\}$, $A = (P - S) - (T - R)$, and $B = S - P$, as defined in Section 3.1.

To streamline the presentation, we restrict the main analysis to parameter instances satisfying the conditions in Assumption 1 below. Under this assumption, all fixed points satisfying the equilibrium conditions are interior. This restriction isolates the stability question that is central to the paper: how the local stability of a collusive outcome changes as the platform varies the intervention level. For completeness, Appendix D characterizes the fixed-point set in the general case, including cases in which the assumption is not satisfied.

Define the policy-saturation threshold

$$\tau_\epsilon(\delta) := \frac{1/2 - \epsilon}{\delta - A\epsilon - B} = \frac{1/2 - \epsilon}{\delta + (1 - \epsilon)(P - S) + \epsilon(T - R)}. \qquad \text{(Policy-Saturation Threshold)}$$

The denominator is strictly positive under the Prisoner's Dilemma ordering, so $\tau_\epsilon(\delta) > 0$. This is the value of the greediness parameter at which the symmetric interior fixed point reaches the forced-exploration boundary.

**Assumption 1** *There exists a nonempty interval $I \subseteq [0, S]$ such that, for every $\delta \in I$,*

$$0 < \tau < \tau_\epsilon(\delta) \qquad \text{and} \qquad |A\tau| < 1.$$

The first inequality keeps the symmetric fixed point away from the forced-exploration boundary, while the second rules out degeneracy in the linear fixed-point equations.

**Proposition 3 (Symmetric interior fixed point)** *Suppose Assumption 1 holds. Then, for every $\delta \in I$, the equilibrium conditions reduce to*

$$\Delta_1^* = Ay^* + B - \delta, \qquad x^* = \frac{1}{2} + \tau\Delta_1^*,$$
$$\Delta_2^* = Ax^* + B - \delta, \qquad y^* = \frac{1}{2} + \tau\Delta_2^*.$$

*The system admits a unique fixed point, which is symmetric and interior, with $x^*(\delta) = y^*(\delta) = a^*(\delta)$ and $\Delta_1^*(\delta) = \Delta_2^*(\delta) = \Delta^*(\delta)$, where*

$$a^*(\delta) = \frac{1 + 2\tau(B - \delta)}{2(1 - A\tau)} \qquad \text{and} \qquad \Delta^*(\delta) = Aa^*(\delta) + B - \delta < 0.$$

*Equivalently,*

$$\mathbf{Q}^*(\delta) = \big(Q_{\text{C}}^*(\delta), Q_{\text{D}}^*(\delta), Q_{\text{C}}^*(\delta), Q_{\text{D}}^*(\delta)\big),$$

*where*

$$Q_{\text{D}}^*(\delta) = \frac{a^*(\delta)(T + \delta) + (1 - a^*(\delta))P}{1 - \gamma}, \qquad Q_{\text{C}}^*(\delta) = a^*(\delta)R + (1 - a^*(\delta))(S - \delta) + \gamma Q_{\text{D}}^*(\delta).$$

Proposition 3 shows that the Q-value REINFORCER dynamics admit a stationary outcome with cooperation above the forced-exploration benchmark even though defection receives the higher Q-value at the fixed point. Under Assumption 1, the fixed-point collusion level is $\mathcal{C}^*(\delta) = \frac{a^*(\delta)}{\epsilon} > 1$. Whether this stationary outcome attracts nearby learning trajectories is a stability question studied below.

The intervention level shifts this fixed point in the intended direction. Because $|A\tau| < 1$, the denominator in $a^*(\delta)$ is positive, and

$$\frac{\partial a^*(\delta)}{\partial \delta} = -\frac{\tau}{1 - A\tau} < 0.$$

Thus, stronger intervention lowers equilibrium cooperation. The responsiveness parameter has a similar effect within the interior symmetric regime:

$$\frac{\partial a^*(\delta)}{\partial \tau} = \frac{B - \delta + A/2}{(1 - A\tau)^2} \leq 0,$$

where the inequality follows from the Prisoner's Dilemma payoff ordering and $\delta \geq 0$. Hence greater policy responsiveness weakly decreases the symmetric equilibrium level of cooperation.

### 5.3. A Constant Learning Rate Benchmark

The case $\theta = 0$ provides a simple benchmark for visualizing the feedback between Q-value gaps and actions. Since all Q-values are updated at the common rate $\alpha$, subtracting the defection-value equation from the

cooperation-value equation in (9) eliminates the continuation term and gives the two-dimensional gap dynamics

$$\begin{aligned} \frac{\mathrm{d}\Delta_1(t)}{\mathrm{d}t} &= \alpha\left[A\, y_t + B - \delta - \Delta_1(t)\right] \qquad & x_t &= \min\left\{1-\epsilon,\, \epsilon + \left(\tfrac{1}{2} + \tau\Delta_1(t) - \epsilon\right)^+\right\} \\ \frac{\mathrm{d}\Delta_2(t)}{\mathrm{d}t} &= \alpha\left[A\, x_t + B - \delta - \Delta_2(t)\right] \qquad & y_t &= \min\left\{1-\epsilon,\, \epsilon + \left(\tfrac{1}{2} + \tau\Delta_2(t) - \epsilon\right)^+\right\}. \end{aligned} \tag{12}$$

Thus, when $\theta = 0$, the action-relevant state is $(\Delta_1, \Delta_2)$; the absolute Q-value levels affect neither current actions nor the gap dynamics. The benchmark makes clear that limited responsiveness can stabilize a symmetric interior fixed point, while the general fixed-point analysis in Appendix D describes how higher responsiveness can introduce asymmetric fixed points. Figure 2 illustrates the constant-learning-rate benchmark in a low-responsiveness regime satisfying Assumption 1, with $\tau = 0.5 < \tau_0 = -1/A$. Relative to the baseline, this figure sets $\delta = 0$, $\theta = 0$, and $\alpha = 1$. The dashed curves are the zero-drift loci $\dot{\Delta}_1 = 0$ and $\dot{\Delta}_2 = 0$; their intersections are the stationary points of (12). The shaded regions identify states in which one or both firms are at the forced-exploration lower bound.

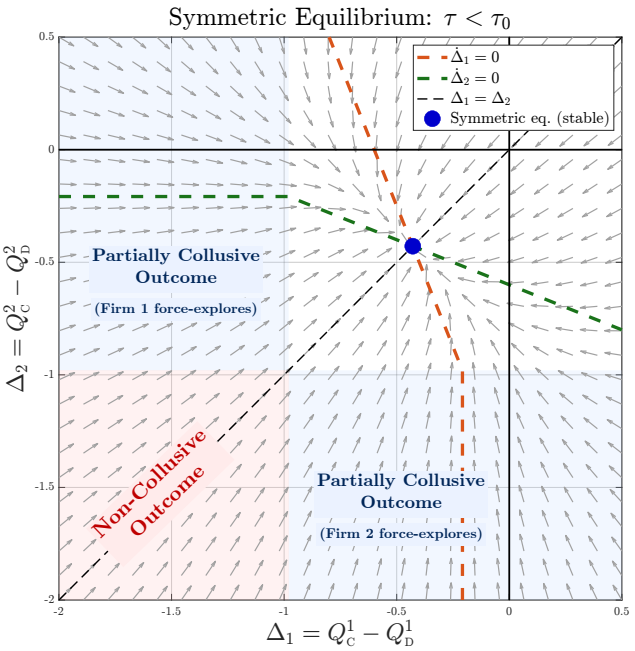



**Figure 2** **Vector field of the reduced $(\Delta_1, \Delta_2)$ system for $\tau = 0.5 < \tau_0$. Shaded regions indicate forced exploration. The system admits a unique symmetric equilibrium (blue circle), which is stable.**

**Example 1** *Consider the parameter instance in Figure 2. The symmetric equilibrium lies in the linear region of the policy map, so* $\Delta^* = Ax^* + B$ *and* $x^* = 1/2 + \tau\Delta^*$. *Solving gives* $x^* = y^* = 2/7 \approx 0.286$. *Hence the probability of mutual cooperation is* $x^* y^* = 4/49 \approx 0.0816$, *while the forced-exploration benchmark is* $\epsilon^2 = 10^{-4}$. *The relative collusion level is*

$$\mathcal{C}^* = \frac{\sqrt{x^* y^*}}{\epsilon} = \frac{2/7}{0.01} \approx 28.6.$$

*Thus, in per-firm equivalent terms, cooperation occurs at about* 28.6 *times the forced-exploration benchmark.*

*For comparison, if firm 2 always selects the dominant action except for forced exploration, so* $y(t) = \epsilon$*, firm 1's limiting gap is* $\Delta_1^* = A\epsilon + B = -0.208$*, which implies* $x^* = 1/2 + \tau\Delta_1^* = 0.396$*. The relative collusion level is then*

$$\mathcal{C}^* = \sqrt{\frac{x^*}{\epsilon}} \approx 6.3.$$

*The high value of* $\mathcal{C}^*$ *in the two-learner case therefore reflects two-sided cooperation generated by the interaction between the learning algorithms.*

### 5.4. Stability

The equilibrium characterization in Proposition 3 is independent of the learning-rate sensitivity parameter $\theta$. Indeed, $\theta$ enters the Q-value dynamics only through the positive learning-rate multipliers, and therefore does not change the fixed-point equations. Its role is instead dynamic: it determines whether the learning dynamics converge locally to a given equilibrium.

Under the conditions in Assumption 1, the stability analysis of the unique symmetric fixed point characterized in Proposition 3 is directly determined by the stability margin

$$\Psi_\theta(x, \delta) := C_2(\delta)(1-x)^\theta - C_1(\delta)x^\theta, \qquad \text{(Stability Margin)}$$

where $C_1(\delta) := 1 + \tau(R - S + \delta)$ and $C_2(\delta) := \tau(T - P + \delta) - (1 - \gamma)$.

Evaluated at the symmetric interior equilibrium $x = a^*(\delta)$, the margin $\Psi_\theta(a^*(\delta), \delta)$ is the trace of the antisymmetric Jacobian block, normalized by the learning-rate $\alpha$. This block describes perturbations in which the firms' Q-values move in opposite directions. Under Assumption 1, the symmetric block remains locally stable, so loss of local stability can occur only through this antisymmetric block. Thus $\Psi_\theta < 0$ means that antisymmetric perturbations decay toward zero, $\Psi_\theta = 0$ marks the stability boundary, and $\Psi_\theta > 0$ means that the equilibrium is locally unstable in the antisymmetric direction. Thus, $\Psi_\theta$ can be interpreted as the equilibrium's local stability margin.

**Theorem 1 (Stability of symmetric interior equilibrium)** *Fix* $\delta \in I$ *and suppose Assumption 1 holds. Let* $\mathbf{Q}^*(\delta)$ *denote the symmetric interior equilibrium in Proposition 3, with* $x^*(\delta) = y^*(\delta) = a^*(\delta)$*. Then the local stability of* $\mathbf{Q}^*(\delta)$ *is determined by the sign of* $\Psi_\theta(a^*(\delta), \delta)$*. If* $\Psi_\theta(a^*(\delta), \delta) < 0$*, then* $\mathbf{Q}^*(\delta)$ *is locally asymptotically stable. If* $\Psi_\theta(a^*(\delta), \delta) > 0$*, then* $\mathbf{Q}^*(\delta)$ *is unstable.*

*Equivalently, if* $C_2(\delta) \le 0$*, then* $\mathbf{Q}^*(\delta)$ *is locally asymptotically stable for all* $\theta \ge 0$*. If* $C_2(\delta) > 0$*, define*

$$\theta_c(\delta) := \frac{\log C_1(\delta) - \log C_2(\delta)}{\log(1 - a^*(\delta)) - \log(a^*(\delta))}.$$

*Then* $\mathbf{Q}^*(\delta)$ *is locally asymptotically stable for* $\theta < \theta_c(\delta)$ *and unstable for* $\theta > \theta_c(\delta)$*.*

Theorem 1 isolates the stability role of $\theta$. At the symmetric equilibrium, the effective learning rates for cooperation and defection are proportional to $(a^*)^\theta$ and $(1 - a^*)^\theta$. Since $a^* < 1/2$, defection is chosen more often than cooperation, and the relative update speed $((1 - a^*)/a^*)^\theta$ increases with $\theta$. Larger $\theta$ therefore makes the defection value update increasingly faster than the cooperation value.

This relative-speed effect is the source of instability. When $\theta$ is small, the two Q-values adjust at sufficiently balanced rates. When $\theta$ is large, one component tracks its moving payoff target much faster than the other, which can amplify fluctuations in the Q-value gap and destabilize the policy response. Thus, $\theta$ affects collusion not by changing the equilibrium cooperation level, but by determining whether learning converges to that equilibrium. The theorem does not classify nonlinear stability at the boundary $\theta = \theta_c(\delta)$.

**Example 2** *Consider the baseline payoff and learning parameters. The symmetric interior equilibrium is* $x^* = y^* = 2/7$*, as before; changing* $\theta$ *does not change this equilibrium, only its stability.*

*For these parameters,* $C_1(0) = 1 + \tau(R - S) = 23/20$ *and* $C_2(0) = \tau(T - P) - (1 - \gamma) = 3/20$*. Therefore,*

$$\theta_c(0) = \frac{\log(23/3)}{\log(5/2)} \approx 2.223.$$

*The symmetric interior equilibrium is locally asymptotically stable for* $0 \le \theta < \theta_c(0)$ *and unstable for* $\theta > \theta_c(0)$*. Thus* $\theta = 1$ *yields a stable equilibrium, while* $\theta = 3$ *destabilizes the same equilibrium.*

*For comparison, suppose firm 2 fixes its cooperation probability at* $y_t = \epsilon$*. Firm 1's limiting Q-value gap is then* $\Delta_1^* = A\epsilon + B = -0.208$*, which implies* $x^* = 1/2 + \tau\Delta_1^* = 0.396$ *and* $C^* = \sqrt{x^*/\epsilon} \approx 6.3$*. Around this one-learner equilibrium, the relevant Jacobian with respect to* $(Q_{\text{C}}^1, Q_{\text{D}}^1)$ *is*

$$\mathbf{J}_\epsilon = \alpha \begin{pmatrix} -(x^*)^\theta & \gamma(x^*)^\theta \\ 0 & -(1-\gamma)(1-x^*)^\theta \end{pmatrix}.$$

*Its eigenvalues are* $-\alpha(x^*)^\theta$ *and* $-\alpha(1-\gamma)(1-x^*)^\theta$*, both strictly negative for every* $\theta \ge 0$*. Hence the one-learner system is locally asymptotically stable for all* $\theta \ge 0$*.*

Example 2 shows that, in this comparison, the instability in Theorem 1 is not caused by the learning-rate sensitivity parameter in isolation. When firm 2's behavior is fixed, firm 1 learns in a stationary payoff environment, so the one-learner dynamics remain locally stable for every $\theta \ge 0$. The destabilizing force appears when both firms learn simultaneously: each firm's action probabilities affect the payoff feedback, learning rates, and future actions of the other. The parameter $\theta$ controls the strength of this endogenous feedback.

## 6. Bifurcation Analysis

Theorem 1 shows that the symmetric interior fixed point can lose stability when learning rates become sufficiently sensitive to action frequencies. We now show that platform intervention can create the same type of instability. Increasing the intervention level $\delta$ reduces equilibrium cooperation, but it also changes the local feedback loop through which the two algorithms learn. Thus, the platform must account not only for where the intervention moves the equilibrium, but also for whether the learning dynamics remain stable around that equilibrium.

We treat $\delta$ as a policy parameter in the Q-value Reinforcer dynamical system. The analysis isolates three effects. First, increasing $\delta$ lowers the cooperation probability at the symmetric fixed point. Second, it

can reduce the local stability margin of that fixed point. Third, near the stability threshold, it can shrink the fixed point's basin of attraction, making the outcome sensitive to implementation speed. We formalize these effects using standard dynamical-systems concepts—fixed points, local stability, basins of attraction, and Hopf bifurcations. Appendix E collects the relevant definitions and technical results, including the center-manifold reduction and Hopf normal form used to characterize stability loss through an oscillatory mode and the emergence of nearby limit cycles.

### 6.1. Intervention Intensity and the Stability Threshold

By Proposition 3, the symmetric interior fixed point satisfies

$$a^*(\delta) = \frac{1 + 2\tau(B - \delta)}{2(1 - A\tau)}, \qquad \frac{\mathrm{d}a^*(\delta)}{\mathrm{d}\delta} = -\frac{\tau}{1 - A\tau} < 0. \tag{13}$$

Thus, as long as the fixed point remains interior, stronger intervention reduces equilibrium cooperation and lowers the collusion measure $\mathcal{C}^*(\delta) = a^*(\delta)/\epsilon$. This is the intended static effect. The same intervention, however, also affects stability. Recall from Section 5.4 that the relevant stability margin is

$$\Psi_\theta(x, \delta) = C_2(\delta)(1 - x)^\theta - C_1(\delta)x^\theta.$$

Along the symmetric fixed-point branch, define

$$\Psi^*_\theta(\delta) := \Psi_\theta(a^*(\delta), \delta). \tag{14}$$

By Theorem 1, the fixed point is locally asymptotically stable when if $\Psi^*_\theta(\delta) < 0$.

**Proposition 4 (Intervention-induced stability crossing)** *Suppose Assumption 1 holds and* $0 < \theta \leq 1$. *Then* $\Psi^*_\theta(\delta)$ *is strictly increasing on* $I$. *Consequently, there is at most one threshold* $\delta_c \in I$ *such that* $\Psi^*_\theta(\delta_c) = 0$. *If such a threshold exists, then* $\mathbf{Q}^*(\delta)$ *is locally asymptotically stable for* $\delta < \delta_c$ *and unstable for* $\delta > \delta_c$, *for* $\delta \in I$. *At* $\delta = \delta_c$, *the equilibrium is a Hopf point; if the first Lyapunov coefficient is nonzero, the system undergoes a nondegenerate Hopf bifurcation there.*[2]

The proposition identifies a tradeoff that a fixed-point comparison misses. Increasing $\delta$ moves the equilibrium toward the non-collusive benchmark, but it also erodes the stability margin. Once $\delta = \delta_c$, the fixed point loses stability through an oscillatory mode.

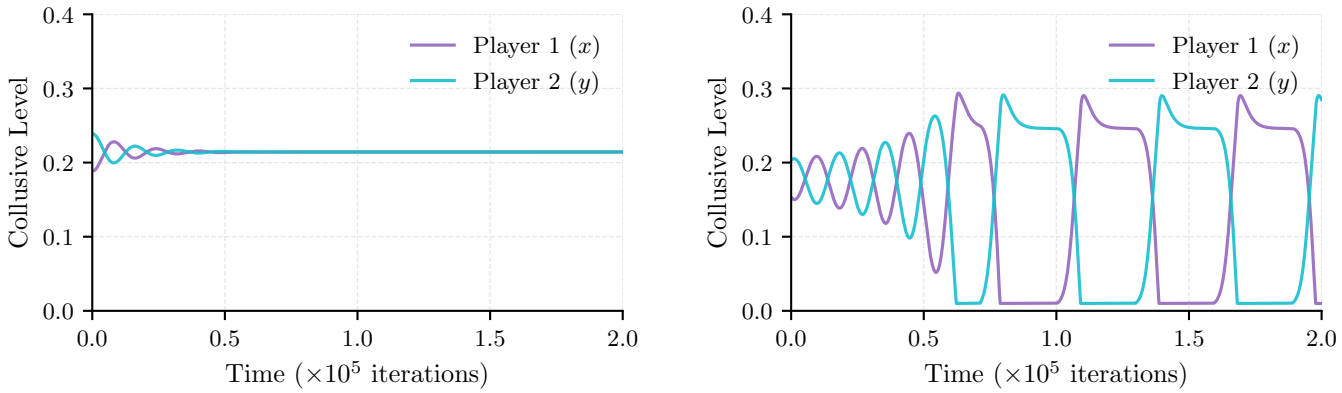
The learning-rate sensitivity parameter $\theta$ is essential for this effect. When $\theta = 0$, cooperation and defection values are updated at the same rate and

$$\Psi^*_0(\delta) = C_2(\delta) - C_1(\delta) = -(2 - \gamma + A\tau),$$

which is independent of $\delta$ and strictly negative under $A\tau > -1$. Thus, intervention intensity alone cannot generate this stability crossing in the constant-learning-rate benchmark. When $\theta > 0$, intervention lowers $a^*(\delta)$, increasing the imbalance between the effective update rates $[a^*(\delta)]^\theta$ and $[1 - a^*(\delta)]^\theta$. A sufficiently large imbalance can destabilize the learning process.

[2] The first Lyapunov coefficient is the leading nonlinear coefficient in the Hopf normal form. Its sign determines whether the nearby periodic orbit is locally stable or unstable. See Appendix E.

We illustrate the crossing using the baseline payoff and learning parameters; see Appendix G. For this instance, $\delta_c \simeq 0.284$. Figure 3 sets $\delta = 0.2 < \delta_c$ in the left panel and $\delta = 0.3 > \delta_c$ in the right panel. A trajectory initialized near the symmetric fixed point converges in the left panel, but develops persistent oscillations in the right panel. Appendix B.1 reports the corresponding discrete-time simulations.




(a) $\delta = 0.2 < \delta_c$: convergence to the stable fixed point.

(b) $\delta = 0.3 > \delta_c$: divergence from the unstable fixed point.

**Figure 3** **Cooperation probabilities under fixed intervention levels below and above the Hopf threshold. The two panels use the same payoff and learning parameters; only the intervention level changes.**

### 6.2. Local Stability and the Oscillatory Mode

We next explain why the stability loss takes the form of oscillations. The analysis starts from the bifurcation pattern illustrated in Figure 3 and studies how it emerges from the Q-value Reinforcer dynamics. The starting point is the local linearization of the system around the symmetric fixed point $\mathbf{Q}^*(\delta)$. Recall from Section 3.4 that the Q-value dynamics can be written as

$$\dot{\mathbf{Q}}_t = \mathbb{F}(\mathbf{Q}_t, \delta).$$

Let $\mathbf{Z}_t = \mathbf{Q}_t - \mathbf{Q}^*(\delta)$ denote a small perturbation around the symmetric fixed point. A first-order expansion gives

$$\dot{\mathbf{Z}}_t = \mathbf{J}(\delta)\mathbf{Z}_t + o(\|\mathbf{Z}_t\|), \qquad \mathbf{J}(\delta) := D_{\mathbf{Q}}\mathbb{F}(\mathbf{Q}^*(\delta), \delta).$$

The eigenvalues of the Jacobian matrix $\mathbf{J}(\delta)$ determine the local behavior. Negative real parts imply local convergence. A complex-conjugate pair crossing the imaginary axis as $\delta$ changes? implies that perturbations neither simply grow nor decay monotonically; instead, the unstable direction is oscillatory.

Let

$$p_\delta = [a^*(\delta)]^\theta, \qquad q_\delta = [1 - a^*(\delta)]^\theta, \qquad r_\delta = R - S + \delta, \qquad t_\delta = T - P + \delta.$$

Since $Q_{\mathrm{D}}^{i*} > Q_{\mathrm{C}}^{i*}$ at every fixed point by Lemma 1, the maximum operator in (9) is locally equal to $Q_{\mathrm{D}}^i$. Moreover, the Bellman residuals vanish at the fixed point. With the state ordered as $(Q_{\mathrm{C}}^1, Q_{\mathrm{D}}^1, Q_{\mathrm{C}}^2, Q_{\mathrm{D}}^2)$, the Jacobian has the block form

$$\mathbf{J}(\delta) = \begin{pmatrix} \mathbf{U}(\delta) & \mathbf{V}(\delta) \\ \mathbf{V}(\delta) & \mathbf{U}(\delta) \end{pmatrix}, \tag{15}$$

where

$$\mathbf{U}(\delta) = \alpha \begin{pmatrix} -p_\delta & \gamma p_\delta \\ 0 & -(1-\gamma) q_\delta \end{pmatrix}, \qquad \mathbf{V}(\delta) = \alpha\tau \begin{pmatrix} p_\delta r_\delta & -p_\delta r_\delta \\ q_\delta t_\delta & -q_\delta t_\delta \end{pmatrix}. \tag{16}$$

This block symmetry separates perturbations into two modes. The matrix $\mathbf{U}+\mathbf{V}$ governs *symmetric* perturbations, in which the two firms' Q-values move together. The matrix $\mathbf{U}-\mathbf{V}$ governs *antisymmetric* perturbations, in which one firm's Q-values move against the other's. Under Assumption 1, the symmetric block remains locally stable. The stability-changing trace is instead

$$\mathrm{Tr}(\mathbf{U}-\mathbf{V}) = \alpha \Psi_\theta^*(\delta). \tag{17}$$

Thus, the intervention-induced stability loss occurs entirely through the antisymmetric mode: the two firms' Q-values begin to move out of phase. The trace and determinant sign checks are given in the proof of Theorem 1.

At $\delta = \delta_c$, the critical eigenvalues of $\mathbf{J}(\delta_c)$ are

$$\lambda_\pm(\delta_c) = \pm i\omega_c, \qquad \omega_c := \alpha\sqrt{(1-\gamma)[a^*(\delta_c)(1-a^*(\delta_c))]^\theta (1+A\tau)}. \tag{18}$$

Their common real part near the threshold is

$$a_H(\delta) := \mathrm{Re}\,\lambda_\pm(\delta) = \frac{\alpha}{2}\Psi_\theta^*(\delta).$$

Since $(\Psi_\theta^*)'(\delta_c) > 0$, this real part crosses zero as $\delta$ increases.

The economic meaning of the critical mode is easiest to see in the antisymmetric difference coordinates

$$v_1 = Q_{\mathrm{C}}^1 - Q_{\mathrm{C}}^2, \qquad v_2 = Q_{\mathrm{D}}^1 - Q_{\mathrm{D}}^2.$$

At the Hopf point, the critical eigenspace lies in the subspace parameterized by $(v_1, v_2)$. Hence the instability is not a common upward or downward movement of both firms' Q-values. It is an oscillatory disagreement between firms. The two remaining directions are locally stable, so the local motion can be summarized by a two-dimensional center-manifold reduction. The next result records the reduced form; Appendix B.3 gives the construction.

**Proposition 5 (Local Hopf reduction)** *Suppose Assumption 1 holds and $\delta_c \in I$ satisfies $\Psi_\theta^*(\delta_c) = 0$ and $(\Psi_\theta^*)'(\delta_c) \neq 0$. Then, near $\mathbf{Q}^*(\delta_c)$, the Q-value dynamics admit a two-dimensional center manifold tangent to the antisymmetric directions $(v_1, v_2)$. In normalized Hopf coordinates $(h_1, h_2) = (r\cos\varphi, r\sin\varphi)$, the oscillation amplitude $r$ satisfies*

$$\dot{r} = a_H(\delta) r + \ell_H r^3 + O(r^5), \qquad a_H(\delta) = \frac{\alpha}{2}\Psi_\theta^*(\delta),$$

*where $\ell_H = \omega_c \ell_1$ and $\ell_1$ is the first Lyapunov coefficient. If $\ell_1 < 0$, the Hopf bifurcation is supercritical and the small periodic orbit is locally stable. If $\ell_1 > 0$, the Hopf bifurcation is subcritical and the small periodic orbit is locally unstable.*

### 6.3. Subcriticality and the Basin Boundary

The subcritical case in Proposition 5 is the important one for platform implementation. It means that the fixed point can remain locally stable while its basin of attraction becomes small. Suppose the stable side of the Hopf point is $\delta < \delta_c$. Then $a_H(\delta) < 0$, so $\mathbf{Q}^*(\delta)$ is locally stable for $\delta$ just below $\delta_c$. If $\ell_1 > 0$, an unstable periodic orbit surrounds the fixed point on the center manifold and contracts toward it as $\delta \uparrow \delta_c$.

**Proposition 6 (Local basin geometry near a subcritical Hopf point)** *Suppose that $\mathbf{Q}^*(\delta_c)$ undergoes a nondegenerate subcritical Hopf bifurcation at $\delta = \delta_c$, with stable side $\delta < \delta_c$. Then there exists $\eta > 0$ such that, for every $\delta \in (\delta_c - \eta, \delta_c)$:*

1. *$\mathbf{Q}^*(\delta)$ is locally asymptotically stable;*
2. *the local center manifold contains an unstable periodic orbit $\Gamma_\delta$ surrounding $\mathbf{Q}^*(\delta)$, whose radius satisfies*
$$\rho_*^2(\delta) = -\frac{a_H(\delta)}{\ell_H} + O\big(a_H(\delta)^2\big) = -\frac{\alpha \Psi_\theta^*(\delta)}{2\omega_c \ell_1} + O\big(\Psi_\theta^*(\delta)^2\big); \tag{19}$$
3. *on the center manifold, trajectories initialized sufficiently close to $\mathbf{Q}^*(\delta)$ and inside $\Gamma_\delta$ converge to the fixed point, whereas trajectories initialized just outside $\Gamma_\delta$ are repelled from it.*

Equation (19) implies that $\rho_*(\delta) \to 0$ as $\delta \uparrow \delta_c$. Thus, below the Hopf threshold, the fixed point can be locally stable but fragile: its local basin in the critical antisymmetric directions shrinks toward the fixed point. The result is local. It identifies the inner unstable cycle that bounds the local basin on the center manifold, but it does not determine the eventual attractor after that boundary is crossed.

Figure 4 illustrates this basin geometry for the baseline payoff and learning parameters, with fixed intervention level $\delta = 0.265$; see Appendix G. The computed first Lyapunov coefficient is positive, so the Hopf bifurcation is subcritical. In normalized Hopf coordinates $(h_1, h_2)$, the unstable periodic orbit is locally circular; in the original antisymmetric coordinates $(v_1, v_2)$, the same boundary appears elliptical. Appendix B.3 reports the first-Lyapunov-coefficient computation and robustness checks.

## 7. Speed of Intervention

Section 6 studies intervention intensity by treating $\delta$ as a fixed policy parameter. We now turn to implementation speed: how quickly the platform moves from no intervention to a target level $\delta^*$. In the continuous-time approximation of (1), an intervention path satisfies

$$\dot{\delta}_t = \mathbb{I}(t, \delta_t), \qquad \delta(0) = 0, \qquad \delta(\mathcal{T}) = \delta^*, \tag{20}$$

where $[0, \mathcal{T}]$ is the implementation horizon. The target level $\delta^*$ determines the payoff environment after the policy is fully implemented. The function $\mathbb{I}$ determines the speed at which the platform moves the learning process to that environment.

Two distinct speeds are therefore at play. The first is the internal adaptation speed of the learning algorithms. For a fixed intervention level $\delta$, local convergence toward $Q^*(\delta)$ is governed by the rate at which perturbations are damped by the Q-value dynamics; near the Hopf threshold, this rate is

$$s(\delta) = -a_H(\delta) = \frac{\alpha}{2}[-\Psi_\theta^*(\delta)],$$

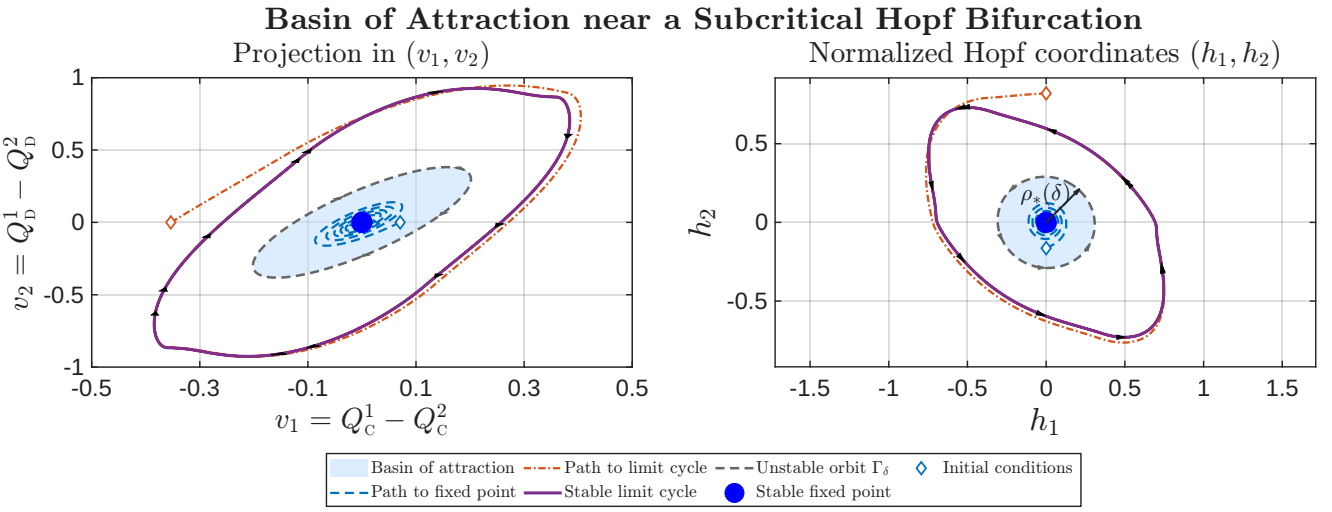



**Figure 4** **Local basin geometry near a subcritical Hopf bifurcation. The shaded region is the local basin of attraction of the stable fixed point, bounded on the center manifold by the unstable periodic orbit $\Gamma_\delta$. In normalized Hopf coordinates $(h_1, h_2)$, the boundary has leading-order radius $\rho_*(\delta)$.**

Thus $\alpha$ sets the overall learning time scale, while $\theta$ affects the stability margin through the relative updating of cooperation and defection values. The second speed is the platform's implementation speed, $\mathbb{I}(t, \delta_t)$, which moves the fixed point $Q^*(\delta_t)$ and changes the size of its basin of attraction. Stability during implementation depends on the balance between these two speeds: the learning dynamics must pull the Q-values toward the moving fixed point faster than the intervention moves the system away from it and contracts its basin of attraction.

The speed question is most relevant when the target lies on the stable side of the Hopf threshold, $\delta^* < \delta_c$. In that case, if the platform were to hold $\delta = \delta^*$ fixed, the symmetric fixed point $\mathbf{Q}^*(\delta^*)$ would be locally asymptotically stable. Thus, the target itself is not locally destabilizing. Nevertheless, the path used to reach it can matter. A sudden intervention can move the inherited Q-values outside the basin of attraction of the new fixed point, whereas a gradual intervention can allow the learning state to track the moving fixed point and remain inside its basin. If $\delta^* > \delta_c$, by contrast, the target fixed point is already locally unstable under the post-intervention environment, so slowing implementation cannot restore its local stability.

The key idea is therefore simple. As $\delta$ increases, the desired fixed point moves and, near a subcritical Hopf bifurcation, its local basin of attraction shrinks. A safe implementation path must give the learning algorithms enough time to adjust while this basin is moving and contracting.

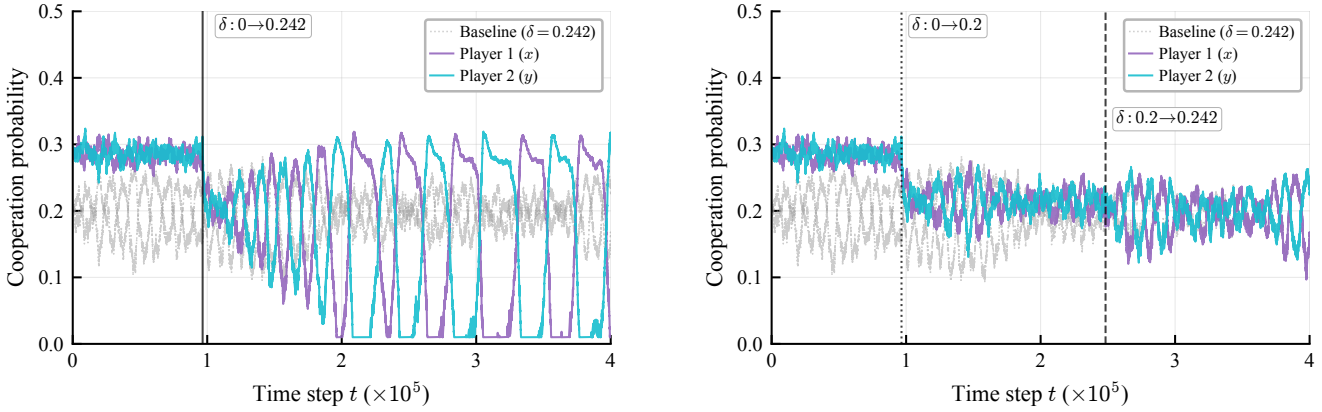
### 7.1. Sudden vs. Gradual Implementation

Suppose the market is initially near the no-intervention fixed point $\mathbf{Q}^*(0)$. If the platform jumps immediately from 0 to $\delta^*$, the payoff environment changes at once. The firms' Q-values, however, do not jump: they are internal learning states that adjust only through subsequent payoff feedback. Thus, after a sudden intervention, the same Q-value vector is evaluated under a new vector field and relative to a new fixed point $\mathbf{Q}^*(\delta^*)$.

This distinction matters in the bistable region identified in Section 6.3. There, the target fixed point may be locally stable but surrounded by an unstable periodic orbit that forms the boundary of its local basin of attraction. If the inherited Q-values lie inside this boundary, the learning dynamics return to the fixed point. If they lie outside, the trajectory is repelled from the local neighborhood and may converge to a stable oscillatory regime. Thus, the same target intervention level can generate convergence or persistent oscillations depending on the path used to reach it.

A gradual intervention changes this geometry more slowly. As $\delta_t$ increases, the fixed point $\mathbf{Q}^*(\delta_t)$ moves and its local basin contracts. If the intervention is sufficiently slow relative to the learning speed, the Q-values can remain close to the moving fixed-point branch and inside the moving basin boundary. The comparison is therefore not between different final policies, but between different implementation paths leading to the same final policy.

Figure 5 illustrates this mechanism for the baseline payoff and learning parameters in Appendix G. The target is $\delta^* = 0.242 < \delta_c \simeq 0.284$, so the target fixed point is locally stable. A sudden jump from $\delta = 0$ to $\delta = 0.242$ can nevertheless place the inherited learning state outside the relevant local basin and generate persistent oscillations. A two-stage path through $\delta = 0.2$ allows the Q-values to adjust before the final increase and preserves convergence. Appendix B.1 reports the corresponding discrete-time simulations.

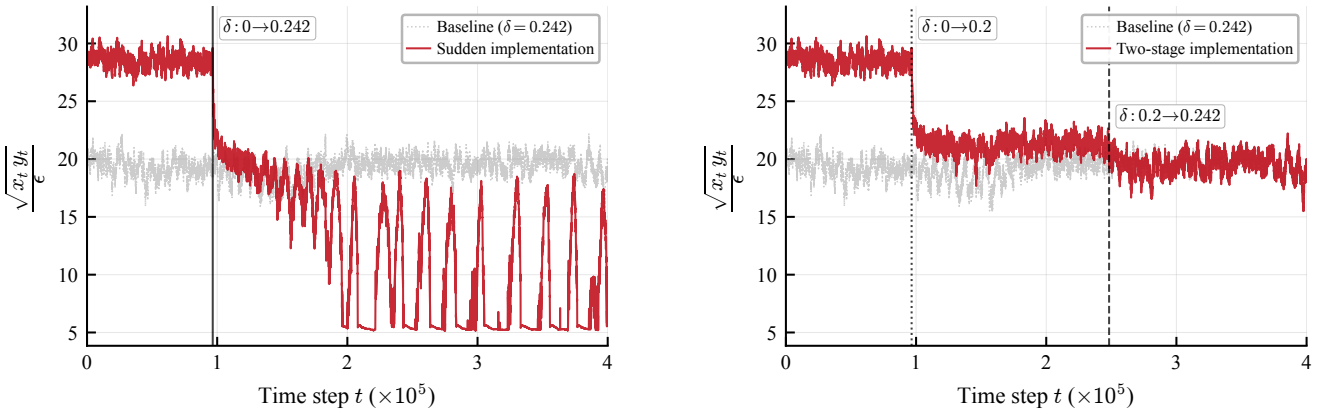



(a) Sudden implementation: $\delta : 0 \to 0.242$.

(b) Two-stage implementation: $\delta : 0 \to 0.2 \to 0.242$.

**Figure 5** **Cooperation probabilities under two paths to the same target intervention level. The sudden policy moves the inherited learning state outside the target fixed point's basin and induces persistent oscillations. The staged policy allows the Q-values to adjust and preserves convergence. The gray series is the reference trajectory with** $\delta = 0.242$ **held fixed, initialized near** $\mathbf{Q}^*(0.242)$.

Figure 6 reports the same comparison using the normalized collusion measure. The sudden intervention induces persistent oscillations in $\mathcal{C}(t)$, and these oscillations occur around a lower level of measured collusion than the reference trajectory with $\delta = 0.242$ held fixed. Thus, the sudden policy can reduce the collusion measure, but it does so through unstable oscillatory dynamics rather than convergence to the target fixed point. By contrast, the two-stage path allows the learning state to adjust before the final increase in $\delta$, so the normalized collusion measure remains close to the reference trajectory with $\delta = 0.242$ held fixed.




(a) Sudden implementation: $\delta : 0 \to 0.242$. (b) Two-stage implementation: $\delta : 0 \to 0.2 \to 0.242$.

**Figure 6** **Normalized collusion measure under two implementation paths to the same target intervention level. The sudden policy generates persistent oscillations in $\mathcal{C}(t) = \sqrt{x_t y_t}/\epsilon$, around a lower measured collusion level than the reference trajectory with $\delta = 0.242$ held fixed. The two stage policy allows the Q-values to adjust before the final increase in $\delta$, preserving convergence to the reference trajectory.**

The deterministic symmetric subspace is invariant, so the speed mechanism is a statement about nearby states rather than this knife-edge path. Initialization differences, asynchronous updating, exploration, and stochastic payoff feedback naturally generate such deviations; rapid implementation can then amplify them by placing the learning state outside the local basin boundary.

### 7.2. A Local Speed Bound Near the Hopf Threshold

We next translate the basin geometry into a local sufficient condition on implementation speed. The result applies near a subcritical Hopf point and should be read as a local tracking condition: it identifies how slowly the platform must move once the path reaches the region where the fixed point's basin is shrinking.

Consider a compact interval

$$I_H = [\underline{\delta}, \overline{\delta}] \subset (\delta_c - \eta, \delta_c)$$

on the stable side of the Hopf point. Reset the local time origin so that $t = 0$ is the time at which the intervention path enters $I_H$. On $I_H$,

$$\Psi_\theta^*(\delta) < 0, \qquad (\Psi_\theta^*)'(\delta) > 0.$$

Thus, the fixed point is locally stable, but its stability margin $-\Psi_\theta^*(\delta)$ decreases as $\delta$ increases toward $\delta_c$.

The relevant deviations are the antisymmetric ones described in Section 6.2:

$$v_1 = Q_{\mathrm{C}}^1 - Q_{\mathrm{C}}^2, \qquad v_2 = Q_{\mathrm{D}}^1 - Q_{\mathrm{D}}^2.$$

Using normalized Hopf coordinates $(h_1, h_2) = (r\cos\varphi, r\sin\varphi)$, write $z(t) = h_1(t) + ih_2(t)$, and let

$$r(t) := |z(t)|$$

be the corresponding oscillation amplitude. For a frozen intervention level $\delta \in I_H$, the unstable periodic orbit that bounds the local basin has leading-order radius

$$\rho_*(\delta) = \left( -\frac{a_H(\delta)}{\ell_H} \right)^{1/2}, \qquad a_H(\delta) = \frac{\alpha}{2} \Psi_\theta^*(\delta).$$

Because $a_H(\delta) < 0$ on the stable side, this radius is well defined and shrinks to zero as $\delta \uparrow \delta_c$.

To measure how close the trajectory is to this moving boundary, define

$$\chi(t) := \frac{r(t)}{\rho_*(\delta_t)}. \tag{21}$$

The quantity $\chi(t)$ is the fraction of the local stability buffer used by the trajectory: $\chi(t) < 1$ means the critical antisymmetric component lies inside the leading-order basin boundary.

The speed restriction comes from a simple comparison between two forces. Let

$$s(\delta) := -a_H(\delta) > 0.$$

The appendix proof of the proposition below derives the following leading-order derivative for the normalized amplitude:

$$\frac{\dot{\chi}(t)}{\chi(t)} = -s(\delta_t)\big(1 - \chi(t)^2\big) + \frac{a_H'(\delta_t)\dot{\delta}_t}{2s(\delta_t)}. \tag{22}$$

The first term is the inward force created by local stability; the second is the outward force created by the moving boundary. A safe implementation path keeps the inward force larger than the boundary contraction.

The following proposition gives one convenient sufficient condition.

**Proposition 7 (Local speed bound under the Hopf normal-form approximation)** *Suppose that* $\mathbf{Q}^*(\delta)$ *undergoes a nondegenerate subcritical Hopf bifurcation at* $\delta_c$*, with stable side* $\delta < \delta_c$*. Let* $\delta : [0, \mathcal{T}] \to I_H$ *be a continuously differentiable and nondecreasing intervention path. Fix a basin buffer* $\zeta \in (0,1)$ *and a slack parameter* $\sigma \in (0,1)$*, and define*

$$m_\zeta := 1 - (1-\zeta)^2 = 2\zeta - \zeta^2.$$

*If* $\chi(0) \le 1 - \zeta$ *and the platform's intervention policy* $\dot{\delta}_t = \mathbb{I}(t, \delta_t)$ *satisfies*

$$0 \le \mathbb{I}(t, \delta_t) < (1-\sigma) m_\zeta \alpha \frac{[-\Psi_\theta^*(\delta_t)]^2}{\Psi_\theta^{*\prime}(\delta_t)} \qquad \text{for all } t \in [0, \mathcal{T}], \tag{23}$$

*then, under the leading-order Hopf approximation,*

$$\chi(t) \le 1 - \zeta \qquad \text{for all } t \in [0, \mathcal{T}].$$

*Hence the critical component of the trajectory remains inside the buffered local basin throughout this portion of the intervention.*

The bound is sufficient, local, and conservative. Its value is that it identifies the scale on which safe implementation must slow down. The admissible speed is proportional to the learning rate $\alpha$, falls rapidly as the stability margin $-\Psi_\theta^*(\delta)$ becomes small, and adjusts for how quickly that margin changes with $\delta$.

An equivalent expression makes the economic content especially transparent. Since $\Psi^{*\prime}_\theta(\delta) > 0$, condition (23) is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{-\Psi_\theta^*(\delta_t)}\right] < (1-\sigma) m_\zeta \alpha. \tag{24}$$

Thus, the relevant object is not the calendar-time speed $\dot{\delta}_t$ by itself, but how quickly the platform exhausts the inverse stability margin.

The asymptotic form of the bound reinforces this point. Transversality gives

$$\Psi_\theta^*(\delta) = -\Psi_\theta^{*\prime}(\delta_c)(\delta_c - \delta) + o(\delta_c - \delta) \qquad \text{as } \delta \uparrow \delta_c.$$

Therefore, the admissible speed in (23) is of order

$$\dot{\delta}_t = \mathbb{I}(t, \delta_t) = O\big(\alpha(\delta_c - \delta_t)^2\big).$$

A constant positive implementation speed cannot satisfy this local bound arbitrarily close to $\delta_c$. Safe implementation requires progressively smaller increments as the target approaches the stability threshold.

### 7.3. Implementation-Time Implications

Integrating the pointwise speed restriction yields a lower bound on the time needed to traverse a stable-side interval near the Hopf point. This bound formalizes the idea that near-threshold interventions require longer rollout horizons.

**Corollary 1 (Implementation-time lower bound)** *Suppose that $\delta : [0, \mathcal{T}] \to I_H$ is continuously differentiable,*

$$\delta(0) = \delta_{\text{start}}, \qquad \delta(\mathcal{T}) = \delta_{\text{end}} < \delta_c,$$

*and satisfies the strict speed condition (23). Define*

$$\mathcal{T}_{\text{crit}}^\sigma := \frac{1}{(1-\sigma) m_\zeta \alpha} \left[ \frac{1}{-\Psi_\theta^*(\delta_{\text{end}})} - \frac{1}{-\Psi_\theta^*(\delta_{\text{start}})} \right]. \tag{25}$$

*Then*

$$\mathcal{T} > \mathcal{T}_{\text{crit}}^\sigma. \tag{26}$$

*Moreover,*

$$\mathcal{T}_{\text{crit}}^\sigma \longrightarrow \infty \qquad \text{as } \delta_{\text{end}} \uparrow \delta_c.$$

The corollary applies only to the part of the path lying in $I_H$; the platform may move more quickly before entering this local neighborhood. Once the intervention approaches the Hopf threshold, however, the required implementation time grows quickly. In leading order,

$$\mathcal{T}_{\text{crit}}^\sigma = O\left( \frac{1}{\alpha(\delta_c - \delta_{\text{end}})} \right) \qquad \text{as } \delta_{\text{end}} \uparrow \delta_c.$$

Thus, targets closer to the threshold require disproportionately longer rollout times. Increasing the learning speed $\alpha$ reduces the required time, while targeting a smaller remaining stability margin increases it. The main design implication is that the platform's problem is not only how much to intervene, but how quickly to deploy the intervention relative to the adaptation speed of the sellers' learning algorithms.

## 8. Computational Experiments

The local speed bound derived above links implementation speed to the stability margin of the symmetric fixed point. We illustrate this implication by varying the learning parameters $\theta$ and $\tau$, holding the baseline payoff parameters fixed; see Appendix G. The takeaway is that implementation can proceed faster when learning is less aggressive. The speed calculations are evaluated at fixed fractions of each perturbed parameter pair's own Hopf threshold, so comparisons are made at the same relative distance from instability.

For any admissible pair $(\theta, \tau)$, let $\delta_c(\theta, \tau)$ denote the corresponding Hopf threshold. Along the symmetric fixed-point branch, write $\Psi^*(\delta; \theta, \tau)$ for the stability margin computed under those learning parameters, with all other parameters held at their benchmark values. Local stability on the stable side is equivalent to $\Psi^*(\delta; \theta, \tau) < 0$, and the Hopf threshold is characterized by $\Psi^*(\delta_c(\theta, \tau); \theta, \tau) = 0$. The local speed limit used in the comparison is

$$\mathcal{G}(\delta; \theta, \tau) = \frac{[-\Psi^*(\delta; \theta, \tau)]^2}{\partial_\delta \Psi^*(\delta; \theta, \tau)}.$$

We plot this object while varying one learning parameter at a time.

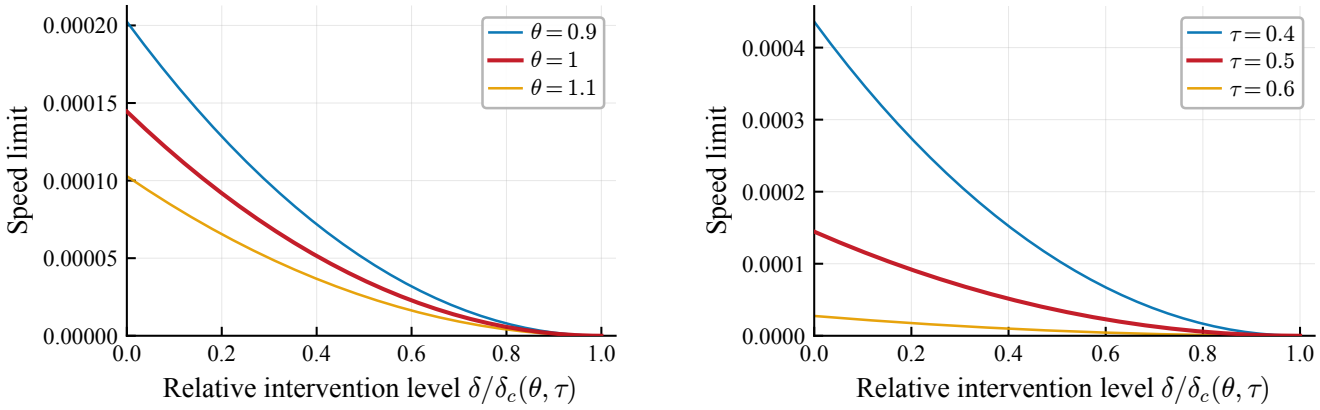



**Figure 7** **Local speed limits under perturbations of the learning parameters. The horizontal axis reports the relative intervention level $\delta/\delta_c(\theta, \tau)$, so each curve is evaluated at the same relative distance from its own Hopf threshold.**

Figure 7 shows the same pattern in both panels: the admissible speed decreases as the intervention path approaches the Hopf threshold, and, at any fixed relative intervention level, the speed limit is lower when either $\theta$ or $\tau$ is larger. The effect of $\theta$ comes from the frequency-weighted update rule $\boldsymbol{\alpha}(\mathbf{s}) = \alpha \boldsymbol{\pi}(\mathbf{s})^\theta$: because $a^*(\delta) < 1/2$, a larger $\theta$ amplifies the difference between cooperation and defection update rates. The effect of $\tau$ comes from the action update $\mathcal{M}_{\epsilon,\tau}$: a larger $\tau$ makes policy responses sharper. Thus, lower $\theta$ or lower $\tau$ gives the platform more room to move quickly, while higher values require slower implementation near $\delta_c(\theta, \tau)$.

## 9. Conclusion

In this paper, we study the mechanisms behind algorithmic collusion and the effects of platform interventions using a continuous-time approximation of Q-learning dynamics. By studying the deterministic

ODEs governing algorithmic interactions, we show that algorithmic collusion is driven by algorithm design and market payoff structures. Moreover, the market outcome of algorithmic interaction is not just a static state but a dynamic trajectory that can lead to complex instability.

Our primary contribution lies in the dynamical analysis of platform intervention. We show that while favoring the non-collusive player can reduce the collusive level, it may also lead to market instability. Using bifurcation analysis, we identify that excessive intervention level does not lead to perfect competition but can trigger a Hopf bifurcation, driving the system into unstable limit cycles. This implies that aggressive interventions may inadvertently create instability, which can be detrimental to both consumer welfare and seller planning.

Furthermore, our results highlight the critical importance of the *speed* of intervention, a dimension often overlooked in static equilibrium analyses. We show that the transition to instability occurs through a Hopf bifurcation. For the benchmark and robustness configurations examined numerically, the Hopf bifurcation is subcritical, implying a shrinking local basin on the stable side. Consequently, a sudden intervention can push the system out of this basin and into an oscillatory regime, even if a stable equilibrium still exists. In contrast, a gradual intervention allows the learning algorithms to adapt continuously, keeping the system within the stability region.

These findings suggest a paradigm shift in algorithmic governance. Effective intervention requires more than identifying the optimal magnitude of penalties; it requires a dynamic control perspective. Policymakers and platforms must consider the trajectory of implementation, favoring adaptive, gradual mechanisms over static, immediate policy shifts. Future research could extend this framework to more complex market settings, including multi-player, multi-product competition, and stochastic demand environments, to further refine our understanding of algorithmic collusion and its regulation.

## References

Amazon Seller Central. 2026. Featured offer. https://sellercentral.amazon.com/help/hub/reference/external/G201687550. Accessed May 24, 2026. 6

Asker, J., C. Fershtman, A. Pakes. 2022. Artificial intelligence, algorithm design, and pricing. *AEA Papers and Proceedings* **112** 452–56. 3

Assad, S., R. Clark, D. Ershov, L. Xu. 2024. Algorithmic pricing and competition: Empirical evidence from the german retail gasoline market. *Journal of Political Economy* **132**(3) 723–771. 2, 4

Banchio, M., G. Mantegazza. 2024. Artificial intelligence and spontaneous collusion. Working Paper. 3, 4, 7, 11

Borkar, V. S. 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge Books Online, Cambridge University Press, Cambridge. 5, 49

Borkar, V. S., S. P. Meyn. 2000. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization* **38**(2) 447–469. 5

Bowling, M., M. Veloso. 2002. Multiagent learning using a variable learning rate. 215–250. 4

Brero, G., N. Lepore, E. Mibuari, D. C. Parkes. 2024. Learning to mitigate AI collusion on economic platforms. Working Paper. 5

Brown, G. W. 1951. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation* 374–376. 4

Brown, Z. Y., A. MacKay. 2023. Competition in pricing algorithms. *American Economic Journal: Microeconomics* **15**(2) 109–156. 4

Cai, Y., A. Oikonomou, W. Zheng. 2022. Finite-time last-iterate convergence for learning in multi-player games. *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35. 33904–33919. 4

Calvano, E., G. Calzolari, V. Denicolò, S. Pastorello. 2020. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* **110**(10) 3267–97. 2, 3, 11

Chen, L., A. Mislove, C. Wilson. 2016. An empirical analysis of algorithmic pricing on amazon marketplace. *Proceedings of the 25th International Conference on World Wide Web*. 1339–1349. 3

Competition and Markets Authority. 2015. Online sales of posters and frames. https://www.gov.uk/cma-cases/online-sales-of-discretionary-consumer-products. 2

Daskalakis, C., A. Ilyas, V. Syrgkanis, H. Zeng. 2018. Training GANs with optimism. *International Conference on Learning Representations*. URL https://openreview.net/forum?id=SJJySbbAZ. 4

Daskalakis, C., I. Panageas. 2019. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *10th Innovations in Theoretical Computer Science Conference (ITCS)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 27:1–27:18. 4

de Cornière, A., G. Taylor. 2019. A model of biased intermediation. *The RAND Journal of Economics* **50**(4) 854–882. doi:10.1111/1756-2171.12298. 4, 6

Dinerstein, M., L. Einav, J. Levin, N. Sundaresan. 2018. Consumer price search and platform design in internet commerce. *American Economic Review* **108**(7) 1820–1859. doi:10.1257/aer.20171218. 4, 6

Douglas, C., F. Provost, A. Sundararajan. 2026. The illusion of collusion. Working paper, arXiv:2411.16574v2. 3, 4

FTC. 2024. FTC and DOJ file statement of interest in hotel room algorithmic price-fixing case. https://www.ftc.gov/news-events/news/press-releases/2024/03/ftc-doj-file-statement-interest-hotel-room-algorithmic-price-fixing-case. 3

Fudenberg, D., D. K. Levine. 1998. *The Theory of Learning in Games*. MIT Press, Cambridge, MA. 4

Fudenberg, D., E. Maskin. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54**(3) 533–554. 11

Goll, T., J. Heitzig, W. Barfuss. 2025. Deterministic model of incremental multi-agent boltzmann Q-learning: Transient cooperation, metastability, and oscillations . 5

Gomes, R., R. Kowalczyk. 2009. Dynamic analysis of multiagent Q-learning with $\epsilon$-greedy exploration. *Proceedings of the 26th International Conference on Machine Learning*. 369–376. 5

Greenwald, A., K. Hall. 2003. Correlated-Q learning. *Proceedings of the 20th International Conference on Machine Learning (ICML)*. 242–249. 4

Guckenheimer, J., P. Holmes. 2002. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, *Applied Mathematical Sciences*, vol. 42. Springer, New York. 52, 56

Hale, J. K. 1988. *Asymptotic Behavior of Dissipative Systems*. Mathematical Surveys and Monographs, American Mathematical Society, Providence, R.I. 33, 34

Hansen, K. T., K. Misra, M. M. Pai. 2021. Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Science* **40**(1) 1–12. 4

Harrington, J. E. 2019. Developing competition law for collusion by autonomous artificial agents†. *Journal of Competition Law & Economics* **14**(3) 331–363. 3, 11

Hart, S., A. Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* **68**(5) 1127–1150. 4

Hartline, J. D., C. Wang, C. Zhang. 2025. Regulation of algorithmic collusion, refined: Testing pessimistic calibrated regret. *Proceedings of the 2025 Symposium on Computer Science and Law*. Association for Computing Machinery, New York, NY, USA, 108–120. 4

Henry, T. A. 2025. Physicians will get their day in court to challenge insurer price-fixing. URL https://www.ama-assn.org/health-care-advocacy/judicial-advocacy/physicians-will-get-their-day-court-challenge-insurer-price. 2

Hofbauer, J., K. Sigmund. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge. 4

Hu, J., M. P. Wellman. 2003. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research* **4**(Nov) 1039–1069. 4

Johnson, J. P., A. Rhodes, M. Wildenbeest. 2023. Platform design when sellers use pricing algorithms. *Econometrica* **91**(5) 1841–1879. 4, 6

Khalil, H. K. 2002. *Nonlinear Systems*. 3rd ed. Prentice Hall. 35

Klein, T. 2021. Autonomous algorithmic collusion: Q-learning under sequential pricing. *The RAND Journal of Economics* **52**(3) 538–558. 3

Kushner, H. J., G. G. Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*, *Applications of Mathematics*, vol. 35. 2nd ed. Springer, New York. 5, 49

Kuznetsov, Y. A. 2004. *Elements of Applied Bifurcation Theory*. Applied Mathematical Sciences, Springer Science & Business Media, New York. 37, 39, 45, 46, 52, 55, 56, 57

Lamba, R., S. Zhuk. 2022. Pricing with algorithms Working Paper. 3

Leonardos, S., G. Piliouras. 2022. Exploration–exploitation in multi–agent learning: Catastrophe theory meets game theory. *Artificial Intelligence* **304** 103653. 5

Leslie, D. S., E. J. Collins. 2005. Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization* **44**(2) 495–514. 5

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. *Proceedings of the 11th International Conference on Machine Learning (ICML)*. 157–163. 4

Littman, M. L. 2001. Friend-or-foe q-learning in general-sum games. *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 322–328. 4

Mertikopoulos, P., C. Papadimitriou, G. Piliouras. 2018. Cycles in adversarial regularized learning. *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2703–2717. 4

Misra, K., E. M. Schwartz, J. Abernethy. 2019. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science* **38**(2) 226–252. 3

OECD Competition Committee Secretariat. 2017. Algorithms and collusion: Background note by the secretariat. Tech. rep., Organisation for Economic Co-operation and Development (OECD). 2, 3, 11

Robinson, J. 1951. An iterative method of solving a game. *Annals of Mathematics* **54**(2) 296–301. 4

Shapley, L. S. 1964. Some topics in two-person games. *Annals of Mathematics Studies* **52** 1–28. 4

Spann, M., M. Bertini, O. Koenigsberg, R. Zeithammer, D. Aparicio, Y. Chen, F. Fantini, G. Z. Jin, V. Morwitz, P. P. Leszczyc, M. A. Vitorino, G. Y. Williams, H. Yoo. 2024. Algorithmic pricing: Implications for consumers, managers, and regulators. Working Paper 32540, National Bureau of Economic Research. 3

Strogatz, S. H. 2018. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. 2nd ed. CRC Press. 52

Taylor, P. D., L. B. Jonker. 1978. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* **40**(1–2) 145–156. 4

Teh, T.-H. 2022. Platform governance. *American Economic Journal: Microeconomics* **14**(3) 213–254. doi:10.1257/mic.20190307. 4, 6

U.S. Department of Justice. 2024. Justice department sues RealPage for algorithmic pricing scheme that harms millions of american renters. https://www.justice.gov/archives/opa/pr/justice-department-sues-realpage-algorithmic-pricing-scheme-harms-millions-american-renters. 2

U.S. Department of Justice, Antitrust Division. 2015. Information: U.S. v. David Topkins. https://www.justice.gov/atr/case-document/file/513586/dl. 1

U.S. Department of Justice, Office of Public Affairs. 2015. Former e-commerce executive charged with price fixing in the antitrust division's first online marketplace prosecution. https://www.justice.gov/archives/opa/pr/former-e-commerce-executive-charged-price-fixing-antitrust-divisions-first-online-marketplace. 1

Watkins, C. J. C. H., P. Dayan. 1992. Q-learning. *Machine Learning* **8**(3–4) 279–292. 4

Wiggins, S. 2003. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, *Texts in Applied Mathematics*, vol. 2. 2nd ed. Springer, New York. doi:10.1007/b97481. 37

Wunder, M., M. L. Littman, M. Babes. 2010. Classes of multiagent Q-learning dynamics with $\epsilon$-greedy exploration. *Proceedings of the 27th International Conference on Machine Learning*. 1183–1190. 5

Yang, Z., X. Lei, P. Gao. 2023. Regulating discriminatory pricing in the presence of tacit collusion. Tech. rep. Working Paper. 4

Zeeman, E. C. 1980. Population dynamics from game theory. *Lecture Notes in Mathematics* **819** 471–497. 4

Zhang, B. H., G. Farina, I. Anagnostides, F. Cacciamani, S. M. McAleer, A. A. Haupt, A. Celli, N. Gatti, V. Conitzer, T. Sandholm. 2024. Steering no-regret learners to a desired equilibrium. *Proceedings of the 25th ACM Conference on Economics and Computation (EC '24)*. Association for Computing Machinery. 5

Zhang, K., Z. Yang, T. Başar. 2021. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*. Springer International Publishing, Cham, 321–384. 4

## Appendix A: Proofs

PROOF OF PROPOSITION 1. Let $M = T + \delta$ denote the maximum possible payoff in the game such that the weighted payoffs are bounded by $M$. We consider the evolution of the state variables.

Let $Q_i$ denote any of the four state variables $\{Q_{\text{C}}^1, Q_{\text{D}}^1, Q_{\text{C}}^2, Q_{\text{D}}^2\}$. The dynamics of any $Q_i$ can be written in the general form:

$$\dot{Q}_i = l_i(t)\left[r_i(t) + \gamma \max\{Q_{\text{C}}^j, Q_{\text{D}}^j\} - Q_i\right], \tag{27}$$

where $l_i(t) > 0$ represents the strictly positive learning rate (e.g., $x^\theta \alpha$) and $r_i(t)$ represents the instantaneous payoff, bounded such that $|r_i(t)| \leq M$.

Consider the upper bound. Suppose that at some time $t$, a variable $Q_i$ first reaches a value $V > \frac{M}{1-\gamma}$. Then, since $\max\{Q_{\text{C}}^j, Q_{\text{D}}^j\}$ is bounded by the maximum value in the system $Q_{\max}$, we have:

$$\dot{Q}_i \leq l_i(t)\left[M + \gamma Q_{\max} - Q_i\right].$$

Since $Q_i$ first reaches the value $V$, we have $Q_i = Q_{\max} = V$, this simplifies to:

$$\dot{Q}_i \leq l_i(t)\left[M - (1-\gamma)V\right].$$

Since $V > \frac{M}{1-\gamma}$, the term in the brackets is strictly negative. Thus, $\dot{Q}_i < 0$, ensuring $Q_i$ decreases.

Similarly, for the lower bound, consider $Q_i = -V$ where $V > \frac{M}{1-\gamma}$. We observe that $\max\{Q_{\text{C}}^j, Q_{\text{D}}^j\} \geq -V$. Thus:

$$\dot{Q}_i \geq l_i(t)\left[-M + \gamma(-V) + V\right] = l_i(t)\left[-M + (1-\gamma)V\right].$$

Since $V > \frac{M}{1-\gamma}$, the derivative is strictly positive. $Q_i$ has to increase.

Therefore, all trajectories eventually enter and remain within the set defined by $|Q_i| \leq \frac{M}{1-\gamma}$, proving the system is dissipative. ■

PROOF OF PROPOSITION 2. We show the proposition by verifying the conditions for the existence of a global attractor as established in Theorem 3.4.6 in Hale (1988). We first restate the theorem below:

**Theorem 3.4.6** *(Existence of Global Attractor, Hale (1988)) Let $X$ be a Banach space and $T(t) : X \to X, t > 0$, be a semigroup of continuous operators. If $T(t)$ is asymptotically smooth, point dissipative, and orbits of bounded sets are bounded, then there exists a global attractor $\mathcal{A}$.*

Let $X = \mathbb{R}^4$. Let $T(t) : X \to X$ be the semigroup generated by the dynamical system, where $T(t)\mathbf{Q}(0) = \mathbf{Q}(t)$ for an initial condition $\mathbf{Q}(0)$. Since the dynamics are Lipschitz continous, we have the semigroup $T(t)$ is continuous. We verify the three conditions:

1. *Asymptotic Smoothness*: A continuous mapping $T : X \to X$ is asymptotically smooth if, for any nonempty closed bounded set $B \subset X$ for which $TB \subset B$, there is a compact set $K \subset B$ such that $K$ attracts $B$. Since $X = \mathbb{R}^4$ is finite-dimensional, we can always find such a compact set $K$, and hence the semiflow $T(t)$ is asymptotically smooth.
2. *Point Dissipativity*: As shown in Proposition 1, there exists a bounded set $\Omega$ that attracts every trajectory starting in $\mathbb{R}^4$. This implies that $T(t)$ is point dissipative.
3. *Boundedness*: Proposition 1 establishes that all orbits eventually enter and remain in $\Omega$, implying that orbits of bounded sets are bounded.

Therefore, by Theorem 3.4.6 in Hale (1988), the semiflow admits a global attractor $\mathcal{A}$ which attracts all bounded subsets of $\mathbb{R}^4$. ■

Proof of Lemma 1. Let $\mathbf{Q}^*$ be any fixed point, either interior or boundary, and let $x^*$ and $y^*$ denote the induced cooperation probabilities. Since $x^*, y^* \in [\epsilon, 1-\epsilon]$ and $\epsilon > 0$, all effective learning rates in (9) are strictly positive. Hence, at a fixed point, the bracketed terms in (9) must vanish.

For player 1, this gives

$$y^* R + (1-y^*)(S-\delta) + \gamma \max_a Q_a^{1*} - Q_{\text{C}}^{1*} = 0,$$
$$y^*(T+\delta) + (1-y^*)P + \gamma \max_a Q_a^{1*} - Q_{\text{D}}^{1*} = 0.$$

Subtracting the first equation from the second yields

$$Q_{\text{D}}^{1*} - Q_{\text{C}}^{1*} = y^*(T-R) + (1-y^*)(P-S) + \delta \geq \epsilon\,(T-R+P-S) + \delta > 0,$$

where the inequalities follow from $T > R$, $P > S$, $y^* \in [\epsilon, 1-\epsilon]$, and $\delta \geq 0$. Therefore, $Q_{\text{D}}^{1*} > Q_{\text{C}}^{1*}$.

The same argument for player 2 gives

$$Q_{\text{D}}^{2*} - Q_{\text{C}}^{2*} = x^*(T-R) + (1-x^*)(P-S) + \delta \geq \epsilon\,(T-R+P-S) + \delta > 0,$$

and hence $Q_{\text{D}}^{2*} > Q_{\text{C}}^{2*}$. Therefore, $Q_{\text{D}}^{i*} > Q_{\text{C}}^{i*}$ for $i = 1, 2$. ■

Proof of Proposition 3. The following proof treats the general setting of Proposition 8 in Appendix D, where the system need not satisfy the condition in Assumption 1. Thus, the statement of Proposition 3 follows as a special case.

At any fixed point of (9)–(10), the equilibrium conditions imply $\Delta_1 = Ay + B - \delta$, $\Delta_2 = Ax + B - \delta$, $x = \mathcal{M}(\Delta_1)$, and $y = \mathcal{M}(\Delta_2)$, where $\mathcal{M}(\Delta) = \epsilon + (1/2 + \tau\Delta - \epsilon)^+$. Moreover, for every $z \in [0,1]$, $Az + B - \delta = z(R-T) + (1-z)(S-P) - \delta < 0$. Hence the upper saturation region of the original policy map is never reached at an equilibrium; only the lower bound $\epsilon$ can bind.

We first characterize symmetric equilibria. A symmetric equilibrium satisfies $a = \mathcal{M}(Aa + B - \delta)$. If the policy is in its linear region, then $a = 1/2 + \tau(Aa + B - \delta)$, so

$$a = \frac{1 + 2\tau(B-\delta)}{2(1-A\tau)}.$$

This candidate is admissible if and only if $a > \epsilon$, equivalently $0 < \tau < \tau_\epsilon(\delta)$. In that case, $x^* = y^* = a^*(\delta)$ and $\Delta_1^* = \Delta_2^* = Aa^*(\delta) + B - \delta$. If $\tau \geq \tau_\epsilon(\delta)$, the linear-region candidate is no larger than $\epsilon$. Equivalently, $A\epsilon + B - \delta \leq (\epsilon - 1/2)/\tau$, so the lower-saturation branch applies at $a = \epsilon$. Hence $x^* = y^* = \epsilon$ and $\Delta_1^* = \Delta_2^* = A\epsilon + B - \delta$. These two cases exhaust the symmetric equilibria.

We next characterize asymmetric equilibria. Suppose first that both policies are in their linear regions. Then $x = 1/2 + \tau(Ay + B - \delta)$ and $y = 1/2 + \tau(Ax + B - \delta)$. Subtracting the two equations gives $(1 + A\tau)(x - y) = 0$. Thus a non-symmetric equilibrium with both policies in the linear region can exist only if $A < 0$ and $\tau = \tau_0 := -1/A$. At this knife edge, the equilibrium conditions reduce to $x + y = 1/2 + \tau(B - \delta)$. Imposing $x, y \geq \epsilon$ gives the line segment $x, y \in [\epsilon,\ 1/2 + \tau(A\epsilon + B - \delta)]$, which is nondegenerate exactly when $\tau_0 < \tau_\epsilon(\delta)$.

It remains to consider equilibria in which one lower bound binds. By symmetry, take $x = \epsilon < y$. Then $y = 1/2 + \tau(A\epsilon + B - \delta)$, and $y > \epsilon$ is equivalent to $\tau < \tau_\epsilon(\delta)$. The condition that $x = \epsilon$ is indeed on the lower-saturation branch is

$$\frac{1}{2} + \tau(Ay + B - \delta) \leq \epsilon.$$

Letting $L := 1/2 - \epsilon$ and $h := A\epsilon + B - \delta < 0$, this condition becomes $(1 + A\tau)(L + \tau h) \le 0$. Since $y > \epsilon$ is equivalent to $L + \tau h > 0$, the lower-saturation condition holds if and only if $1 + A\tau \le 0$, that is, if and only if $A < 0$ and $\tau \ge \tau_0$. Therefore, if $\tau_0 < \tau < \tau_\epsilon(\delta)$, the only asymmetric equilibria are $(x^*, y^*) = (\epsilon,\, 1/2 + \tau(A\epsilon + B - \delta))$ and its symmetric counterpart. If $\tau = \tau_0 < \tau_\epsilon(\delta)$, these two boundary equilibria are the endpoints of the continuum described above.

Finally, the associated Q-value gaps follow directly from $\Delta_1^* = Ay^* + B - \delta$ and $\Delta_2^* = Ax^* + B - \delta$. Since every non-symmetric equilibrium either has both policies in their linear regions or has one policy at the lower bound, no other asymmetric equilibria exist. ■

Proof Corollary 2. By Proposition 3, asymmetric equilibria can arise only if $A < 0$ and $\tau \ge -1/A$, which is equivalent to $A\tau \le -1$. Hence $A\tau > -1$ rules out all asymmetric equilibria. ■

Proof of Theorem 1. Fix $\delta \ge 0$ and suppose $0 < \tau < \tau_\epsilon(\delta)$ and $A\tau > -1$. Let $a^* = a^*(\delta)$ denote the symmetric interior equilibrium in Proposition 3. By Lemma 1, at any interior fixed point we have $Q_{\text{D}}^{i*} > Q_{\text{C}}^{i*}$ for $i = 1, 2$. Hence, in a neighborhood of the symmetric interior equilibrium, the maximum term in (9) is given by $Q_{\text{D}}^i$, and the local dynamics are

$$\begin{aligned}
\frac{\mathrm{d}Q_{\text{C}}^1}{\mathrm{d}t} &= \alpha x^\theta \left(yR + (1-y)(S-\delta) + \gamma Q_{\text{D}}^1 - Q_{\text{C}}^1\right), \\
\frac{\mathrm{d}Q_{\text{D}}^1}{\mathrm{d}t} &= \alpha (1-x)^\theta \left(y(T+\delta) + (1-y)P + \gamma Q_{\text{D}}^1 - Q_{\text{D}}^1\right), \\
\frac{\mathrm{d}Q_{\text{C}}^2}{\mathrm{d}t} &= \alpha y^\theta \left(xR + (1-x)(S-\delta) + \gamma Q_{\text{D}}^2 - Q_{\text{C}}^2\right), \\
\frac{\mathrm{d}Q_{\text{D}}^2}{\mathrm{d}t} &= \alpha (1-y)^\theta \left(x(T+\delta) + (1-x)P + \gamma Q_{\text{D}}^2 - Q_{\text{D}}^2\right).
\end{aligned}$$

At the fixed point, the bracketed terms vanish. Subtracting the defection equation from the cooperation equation for firm 1 gives

$$\Delta_1^* = Q_{\text{C}}^{1*} - Q_{\text{D}}^{1*} = Ay^* + B - \delta,$$

and similarly $\Delta_2^* = Ax^* + B - \delta$. Since the equilibrium is interior, $x^* = 1/2 + \tau\Delta_1^*$ and $y^* = 1/2 + \tau\Delta_2^*$. Imposing symmetry gives the equilibrium in Proposition 3. Moreover, $Aa^* + B - \delta < 0$, so $a^* < 1/2$.

We now study local stability. Let $p := (a^*)^\theta$, $q := (1 - a^*)^\theta$, $r_\delta := R - S + \delta$, and $t_\delta := T - P + \delta$. Because the bracketed terms vanish at the fixed point, derivatives of the learning-rate multipliers do not enter the Jacobian. With the ordering $(Q_{\text{C}}^1, Q_{\text{D}}^1, Q_{\text{C}}^2, Q_{\text{D}}^2)$, the Jacobian has the block form

$$\mathbf{J} = \begin{pmatrix} \mathbf{U} & \mathbf{V} \\ \mathbf{V} & \mathbf{U} \end{pmatrix},$$

where

$$\mathbf{U} = \alpha \begin{pmatrix} -p & \gamma p \\ 0 & -(1-\gamma)q \end{pmatrix}, \qquad \mathbf{V} = \alpha\tau \begin{pmatrix} pr_\delta & -pr_\delta \\ qt_\delta & -qt_\delta \end{pmatrix}.$$

The eigenvalues of $\mathbf{J}$ are the eigenvalues of $\mathbf{U} + \mathbf{V}$ and $\mathbf{U} - \mathbf{V}$. Therefore, by the two-dimensional Routh–Hurwitz criterion, the equilibrium is locally asymptotically stable if and only if both matrices have negative trace and positive determinant; see (Khalil 2002, Theorem 4.7).

A direct calculation gives

$$\det(\mathbf{U} + \mathbf{V}) = \alpha^2(1-\gamma)pq(1 - A\tau), \qquad \det(\mathbf{U} - \mathbf{V}) = \alpha^2(1-\gamma)pq(1 + A\tau).$$

Because $p, q > 0$, $\alpha > 0$, and $\gamma < 1$, and because $A\tau > -1$ is imposed, the determinant conditions reduce to $A\tau < 1$.

It remains to analyze the trace conditions. For $\mathbf{U}+\mathbf{V}$,

$$\mathrm{Tr}(\mathbf{U}+\mathbf{V}) = \alpha\left[-p-(1-\gamma)q+\tau\left(pr_\delta - qt_\delta\right)\right].$$

Since $a^* < 1/2$, we have $q \geq p$. Also $t_\delta > 0$. Hence, if $A\tau < 1$,

$$\mathrm{Tr}(\mathbf{U}+\mathbf{V}) \leq \alpha\left[-p-(1-\gamma)q+\tau p(r_\delta - t_\delta)\right] = \alpha\left[-(1-A\tau)p-(1-\gamma)q\right] < 0,$$

where we used $r_\delta - t_\delta = A$. Thus $\mathrm{Tr}(\mathbf{U}+\mathbf{V}) < 0$ whenever $A\tau < 1$.

For $\mathbf{U}-\mathbf{V}$, using $C_1(\delta) = 1+\tau(R-S+\delta) = 1+\tau r_\delta$ and $C_2(\delta) = \tau(T-P+\delta)-(1-\gamma) = \tau t_\delta - (1-\gamma)$, we obtain

$$\mathrm{Tr}(\mathbf{U}-\mathbf{V}) = \alpha\left[-C_1(\delta)p + C_2(\delta)q\right].$$

Therefore $\mathrm{Tr}(\mathbf{U}-\mathbf{V}) < 0$ if and only if

$$C_2(\delta)(1-a^*)^\theta < C_1(\delta)(a^*)^\theta.$$

Combining the determinant and trace conditions, the symmetric interior equilibrium is locally asymptotically stable if and only if

$$A\tau < 1 \qquad \text{and} \qquad C_2(\delta)(1-a^*)^\theta < C_1(\delta)(a^*)^\theta.$$

The equivalent threshold characterization follows immediately. If $C_2(\delta) \leq 0$, the trace condition for $\mathbf{U}-\mathbf{V}$ holds for all $\theta \geq 0$, since $C_1(\delta) > 0$. If $C_2(\delta) > 0$, then

$$C_2(\delta)(1-a^*)^\theta < C_1(\delta)(a^*)^\theta \quad \Longleftrightarrow \quad \left(\frac{1-a^*}{a^*}\right)^\theta < \frac{C_1(\delta)}{C_2(\delta)}.$$

Because $a^* < 1/2$, the denominator $\log(1-a^*) - \log(a^*)$ is positive. Moreover, under $A\tau > -1$, we have

$$C_1(\delta) - C_2(\delta) = 2 - \gamma + A\tau > 0,$$

so the threshold is well defined and positive. Hence the equilibrium is locally asymptotically stable if and only if $\theta < \theta_c(\delta)$, where

$$\theta_c(\delta) := \frac{\log C_1(\delta) - \log C_2(\delta)}{\log(1-a^*) - \log(a^*)}.$$

At $\theta = \theta_c(\delta)$, the relevant trace is zero and the linearization is non-hyperbolic. For $\theta > \theta_c(\delta)$, $\mathrm{Tr}(\mathbf{U}-\mathbf{V}) > 0$, so at least one eigenvalue has positive real part and the equilibrium is unstable. ■

Proof of Proposition 4. Use the notation from Theorem 1. Along the symmetric interior branch, write $a^* = a^*(\delta)$. The stability margin is

$$\Psi^*_\theta(\delta) = C_2(\delta)(1-a^*)^\theta - C_1(\delta)(a^*)^\theta,$$

where

$$C_1(\delta) = 1+\tau(R-S+\delta), \qquad C_2(\delta) = \tau(T-P+\delta)-(1-\gamma).$$

Moreover,

$$\frac{\mathrm{d}a^*(\delta)}{\mathrm{d}\delta} = -\frac{\tau}{1-A\tau} < 0,$$

because Assumption 1 implies $|A\tau| < 1$. Theorem 1 also gives $a^* < 1/2$ on the interior symmetric branch. Hence $(1-a^*)^\theta > (a^*)^\theta$ for $0 < \theta \leq 1$.

Differentiating $\Psi^*_\theta(\delta)$ gives

$$\frac{\mathrm{d}}{\mathrm{d}\delta}\Psi^*_\theta(\delta) = \tau\left[(1-a^*)^\theta - (a^*)^\theta\right] - \theta\frac{\mathrm{d}a^*}{\mathrm{d}\delta}\left[C_2(\delta)(1-a^*)^{\theta-1} + C_1(\delta)(a^*)^{\theta-1}\right].$$

The first term is strictly positive. Since $\mathrm{d}a^*/\mathrm{d}\delta < 0$, it remains only to show that

$$C_2(\delta)(1-a^*)^{\theta-1} + C_1(\delta)(a^*)^{\theta-1} > 0.$$

If $C_2(\delta) \geq 0$, this is immediate because $C_1(\delta) > 0$. If $C_2(\delta) < 0$, write $C_2(\delta) = -|C_2(\delta)|$. Since $a^* < 1/2$ and $\theta \leq 1$,

$$\left(\frac{1-a^*}{a^*}\right)^{\theta-1} \leq 1.$$

Therefore

$$\begin{aligned} C_2(\delta)(1-a^*)^{\theta-1} + C_1(\delta)(a^*)^{\theta-1} &= (a^*)^{\theta-1}\left[C_1(\delta) - |C_2(\delta)|\left(\frac{1-a^*}{a^*}\right)^{\theta-1}\right] \\ &\geq (a^*)^{\theta-1}\big[C_1(\delta) - |C_2(\delta)|\big] \\ &= (a^*)^{\theta-1}\big[C_1(\delta) + C_2(\delta)\big] \\ &= (a^*)^{\theta-1}\left[\gamma + \tau\big((R-S) + (T-P) + 2\delta\big)\right] > 0. \end{aligned}$$

Thus $(\Psi_\theta^*)'(\delta) > 0$ throughout $I$.

The remaining conclusions follow from Theorem 1. There, $\mathrm{Tr}(\mathbf{U}-\mathbf{V}) = \alpha\Psi_\theta^*(\delta)$, while the determinant of $\mathbf{U}-\mathbf{V}$ is positive under $|A\tau| < 1$. Hence the sign of $\Psi_\theta^*(\delta)$ determines the stability of the antisymmetric Jacobian block. Strict monotonicity implies that there is at most one crossing $\delta_c$, with stability for $\delta < \delta_c$ and instability for $\delta > \delta_c$. At a crossing, the trace of $\mathbf{U}-\mathbf{V}$ is zero and its determinant is positive, so the critical eigenvalues are purely imaginary. Together with the transversality $(\Psi_\theta^*)'(\delta_c) > 0$, this is a Hopf point; if the first Lyapunov coefficient is nonzero, the Hopf bifurcation is nondegenerate. ■

Proof of Proposition 6. Let $\varphi_t^\delta(\mathbf{Q}_0)$ denote the flow of the Q-value dynamics under the fixed intervention level $\delta$, starting from initial condition $\mathbf{Q}_0 \in \mathbb{R}^4$. For compactness, when $\delta$ is fixed, write $\mathbf{Q}^* = \mathbf{Q}^*(\delta)$ and define its basin of attraction by

$$\mathcal{B}_\delta(\mathbf{Q}^*) := \{\mathbf{Q}_0 \in \mathbb{R}^4 : \lim_{t\to\infty} \varphi_t^\delta(\mathbf{Q}_0) = \mathbf{Q}^*\}.$$

Since $\mathbf{Q}^*(\delta_c)$ undergoes a subcritical Hopf bifurcation at $\delta = \delta_c$, standard Hopf bifurcation theory (Kuznetsov 2004) implies that there exists $\eta > 0$ such that, for all $\delta \in (\delta_c - \eta, \delta_c)$, the equilibrium $\mathbf{Q}^*(\delta)$ is locally asymptotically stable. Moreover, in the local two-dimensional Hopf reduction, there exists an unstable periodic orbit $\Gamma_\delta$ surrounding $\mathbf{Q}^*(\delta)$ whose amplitude is of order $O(\sqrt{|\delta - \delta_c|})$ (Wiggins 2003).

Because $\Gamma_\delta$ is an invariant closed curve in the two-dimensional reduced dynamics, it acts as a local boundary. On the Hopf reduction manifold, trajectories inside $\Gamma_\delta$ converge to $\mathbf{Q}^*(\delta)$, while trajectories on or outside $\Gamma_\delta$ in the reduced dynamics cannot cross $\Gamma_\delta$ and therefore are not attracted to $\mathbf{Q}^*(\delta)$. Since the reduced manifold is invariant under the full flow, these initial conditions also correspond to trajectories of the full four-dimensional system that do not belong to $\mathcal{B}_\delta(\mathbf{Q}^*)$.

By Proposition 1, the system is dissipative. In particular, all trajectories are bounded and the dynamics admit a compact global attractor $\mathcal{A}$. To characterize the long-run behavior of trajectories that are not attracted to the equilibrium $\mathbf{Q}^*(\delta)$, we use the notion of an $\omega$-limit set.

Let $\mathbf{Q}_0 \in \mathbb{R}^4$ be an initial condition and let $\varphi_t^\delta(\mathbf{Q}_0)$ denote the flow of the Q-value dynamics. The *$\omega$-limit set* of $\mathbf{Q}_0$ is defined as

$$\omega_\delta(\mathbf{Q}_0) := \left\{\mathbf{Q} \in \mathbb{R}^4 \,\middle|\, \exists\, t_n \to +\infty \text{ such that } \varphi_{t_n}^\delta(\mathbf{Q}_0) \to \mathbf{Q}\right\},$$

where $t_n$ is a sequence of time points going to infinity. Note that $\omega_\delta(\mathbf{Q}_0)$ is closed by construction. Moreover, because the flow is continuous, the set is positively invariant. Since the system is dissipative, the trajectory $\varphi_{t_n}^\delta(\mathbf{Q}_0)$ is bounded for all $t \geq 0$, there must exist a converging subsequence and therefore $\omega_\delta(\mathbf{Q}_0)$ is nonempty. Thus, we have $\omega_\delta(\mathbf{Q}_0)$ is compact, positively invariant under the flow, and contained in the global attractor $\mathcal{A}$.

Intuitively, the $\omega$-limit set collects all possible accumulation points of a trajectory as time tends to infinity and therefore describes its asymptotic behavior. Since the system is dissipative, every trajectory is bounded and has a nonempty compact $\omega$-limit set contained in the global attractor.

Now consider an initial condition

$$\mathbf{Q}_0 \notin \mathcal{B}_\delta(\mathbf{Q}^*(\delta)),$$

so that the corresponding trajectory is not attracted to the stable interior equilibrium $\mathbf{Q}^*(\delta)$. We claim that its $\omega$-limit set cannot be a singleton. Suppose, toward a contradiction, that

$$\omega_\delta(\mathbf{Q}_0) = \{\bar{\mathbf{Q}}\}.$$

Because $\omega_\delta(\mathbf{Q}_0)$ is invariant under the flow, the point $\bar{\mathbf{Q}}$ must be an equilibrium of the system. Under the uniqueness conditions in Appendix B.2, however, $\mathbf{Q}^*(\delta)$ is the unique fixed point of the Q-value dynamics. Hence

$$\bar{\mathbf{Q}} = \mathbf{Q}^*(\delta).$$

This implies

$$\lim_{t\to\infty} \varphi_t^\delta(\mathbf{Q}_0) = \mathbf{Q}^*(\delta),$$

which contradicts $\mathbf{Q}_0 \notin \mathcal{B}_\delta(\mathbf{Q}^*(\delta))$. Therefore, $\omega_\delta(\mathbf{Q}_0)$ is not a singleton.

Moreover, $\omega_\delta(\mathbf{Q}_0)$ cannot contain any equilibrium. Indeed, the only equilibrium is $\mathbf{Q}^*(\delta)$, and if $\mathbf{Q}^*(\delta) \in \omega_\delta(\mathbf{Q}_0)$, then local asymptotic stability of $\mathbf{Q}^*(\delta)$ would imply that the trajectory eventually enters its basin of attraction and converges to $\mathbf{Q}^*(\delta)$, again contradicting the choice of $\mathbf{Q}_0$.

Thus, for every initial condition outside the basin of attraction of $\mathbf{Q}^*(\delta)$, the trajectory does not settle at a fixed point. Its long-run behavior is contained in the compact invariant non-equilibrium set, which is nonempty and not a singleton. This persistent non-stationary behavior may take the form of a periodic orbit, an invariant torus, or a more complex recurrent invariant set. ■

Proof of Proposition 7. We give the proof in three steps. First, we identify the Hopf plane. Second, we compute the leading-order radius of the unstable periodic orbit. Third, we compare the motion of the trajectory with the motion of this shrinking local boundary.

*Step 1: the Hopf plane.* At the symmetric interior equilibrium, the two sellers have the same equilibrium Q-values, so

$$\mathbf{Q}^*(\delta) = \begin{pmatrix} Q_{\text{C}}^*(\delta) \\ Q_{\text{D}}^*(\delta) \\ Q_{\text{C}}^*(\delta) \\ Q_{\text{D}}^*(\delta) \end{pmatrix}.$$

The Jacobian has the block form

$$\mathbf{J}(\delta) = \begin{pmatrix} \mathbf{U}(\delta) & \mathbf{V}(\delta) \\ \mathbf{V}(\delta) & \mathbf{U}(\delta) \end{pmatrix}.$$

Let

$$\boldsymbol{\xi}_1 = \begin{pmatrix} Q_{\text{C}}^1 - Q_{\text{C}}^*(\delta) \\ Q_{\text{D}}^1 - Q_{\text{D}}^*(\delta) \end{pmatrix}, \qquad \boldsymbol{\xi}_2 = \begin{pmatrix} Q_{\text{C}}^2 - Q_{\text{C}}^*(\delta) \\ Q_{\text{D}}^2 - Q_{\text{D}}^*(\delta) \end{pmatrix}$$

denote deviations from the symmetric equilibrium. The linearized dynamics are

$$\begin{pmatrix} \dot{\boldsymbol{\xi}}_1 \\ \dot{\boldsymbol{\xi}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{U} & \mathbf{V} \\ \mathbf{V} & \mathbf{U} \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix}.$$

Introduce symmetric and antisymmetric coordinates

$$\mathbf{u} = \frac{\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2}{2}, \qquad \mathbf{v} = \frac{\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2}{2}.$$

Equivalently, $\boldsymbol{\xi}_1 = \mathbf{u} + \mathbf{v}$ and $\boldsymbol{\xi}_2 = \mathbf{u} - \mathbf{v}$. Substitution gives the block-diagonal system

$$\dot{\mathbf{u}} = (\mathbf{U} + \mathbf{V})\mathbf{u}, \qquad \dot{\mathbf{v}} = (\mathbf{U} - \mathbf{V})\mathbf{v}.$$

Thus $\mathbf{u}$ is the average seller deviation, while

$$\mathbf{v} = \frac{1}{2}\begin{pmatrix} Q_{\text{C}}^1 - Q_{\text{C}}^2 \\ Q_{\text{D}}^1 - Q_{\text{D}}^2 \end{pmatrix} \tag{28}$$

is the seller-difference deviation. This appendix coordinate is one half of the raw antisymmetric vector $(v_1, v_2)$ used in the main text. From the proof of Theorem 1, the critical eigenvalues are generated by the matrix block $\mathbf{U} - \mathbf{V}$. Hence the *Hopf plane* is the antisymmetric plane $\mathbf{v}$, governed by $\mathbf{U}(\delta_c) - \mathbf{V}(\delta_c)$.

From Theorem 1,

$$\operatorname{Tr}(\mathbf{U} - \mathbf{V}) = \alpha \Psi_\theta^*(\delta).$$

Moreover, under the condition $x^* \in (0,1)$, $\gamma < 1$, and $|A\tau| < 1$, the proof of Theorem 1 gives $\det(\mathbf{U} - \mathbf{V}) > 0$. Hence, the real part of the two eigenvalues of $\mathbf{U} - \mathbf{V}$ is one half of the trace:

$$a_H(\delta) = \frac{1}{2}\operatorname{Tr}(\mathbf{U} - \mathbf{V}) = \frac{\alpha}{2}\Psi_\theta^*(\delta).$$

At the Hopf point, $\Psi_\theta^*(\delta_c) = 0$, so

$$a_H(\delta_c) = 0, \qquad \lambda_\pm(\delta_c) = \pm i\omega_c, \qquad \omega_c > 0.$$

Let $z \in \mathbb{C}$ denote the local Hopf coordinate, which is obtained to leading order by expressing the antisymmetric coordinate $\mathbf{v}$ in the eigenbasis of $\mathbf{U}(\delta_c) - \mathbf{V}(\delta_c)$. On the stable side, $\Psi_\theta^*(\delta) < 0$ and therefore $a_H(\delta) < 0$.

*Step 2: the radius of the unstable periodic orbit.* By the Hopf normal form on the local center manifold, the leading-order dynamics in the Hopf plane are given by

$$\dot{z} = \big(a_H(\delta) + i\omega(\delta)\big)z + (\ell_H + i\ell_I)z|z|^2. \tag{29}$$

The coefficient $\ell_1$ computed using the convention of Kuznetsov (2004) gives the real radial cubic coefficient

$$\ell_H = \omega_c \ell_1.$$

Let $r(t) = |z(t)|$. Since $r^2 = z\bar{z}$,

$$\frac{1}{2}\frac{d}{dt}r^2 = \operatorname{Re}(\bar{z}\dot{z}).$$

Substituting (29) gives

$$\operatorname{Re}(\bar{z}\dot{z}) = \operatorname{Re}\left[(a_H(\delta) + i\omega(\delta))|z|^2 + (\ell_H + i\ell_I)|z|^4\right] = a_H(\delta)r^2 + \ell_H r^4.$$

The imaginary terms only rotate the phase and therefore do not affect the amplitude. For $r > 0$,

$$\dot{r} = a_H(\delta)r + \ell_H r^3. \tag{30}$$

The case $r = 0$ follows by continuity. The nonzero radial equilibrium satisfies

$$a_H(\delta) + \ell_H \rho_*^2(\delta) = 0.$$

Thus, in the leading-order radial equation,

$$\rho_*^2(\delta) = -\frac{a_H(\delta)}{\ell_H}.$$

Including the higher-order terms in the Hopf normal form gives

$$\rho_*^2(\delta) = -\frac{a_H(\delta)}{\ell_H} + O(a_H(\delta)^2).$$

*Step 3: the moving-boundary comparison.* On the stable side define

$$s(\delta) := -a_H(\delta) > 0.$$

Using $\rho_*^2 = s/\ell_H$, equation (30) becomes

$$\dot{r} = -s(\delta)r + \ell_H r^3 = -s(\delta)r\left(1 - \frac{r^2}{\rho_*^2(\delta)}\right). \tag{31}$$

Recall we defined the normalized amplitude as

$$\chi(t) := \frac{r(t)}{\rho_*(\delta_t)}.$$

For $r > 0$, we take the logarithm of $\chi(t)$ and take the time derivative to separate the trajectory's motion from the boundary's motion:

$$\frac{\dot{\chi}}{\chi} = \frac{\dot{r}}{r} - \frac{\dot{\rho}_*}{\rho_*}.$$

The case $r = 0$ follows by continuity. By (31),

$$\frac{\dot{\chi}}{\chi} = -s(\delta)(1-\chi^2) - \frac{\dot{\rho}_*}{\rho_*}. \tag{32}$$

Since, to leading order,

$$\rho_*(\delta) = \left(\frac{-a_H(\delta)}{\ell_H}\right)^{1/2},$$

and $\ell_H$ is fixed at its Hopf value in the leading-order reduction,

$$\frac{\dot{\rho}_*}{\rho_*} = \frac{a_H'(\delta)\dot{\delta}}{2a_H(\delta)}.$$

Because $a_H(\delta) = -s(\delta)$, equation (32) becomes

$$\frac{\dot{\chi}}{\chi} = -s(\delta)(1-\chi^2) + \frac{a_H'(\delta)\dot{\delta}}{2s(\delta)}. \tag{33}$$

The first term is inward recovery of the trajectory. The second term is the outward effect generated by the shrinking local boundary.

Evaluate (33) at the buffered boundary $\chi = 1-\zeta$. We have,

$$1 - \chi^2 = 1 - (1-\zeta)^2 = m_\zeta.$$

Hence

$$\frac{\dot{\chi}}{\chi} = -s(\delta)m_\zeta + \frac{a_H'(\delta)\dot{\delta}}{2s(\delta)}. \tag{34}$$

A sufficient condition ensuring that trajectories cannot exit the buffered region is that $\dot{\chi} < 0$ whenever $\chi = 1-\zeta$. By (34), a sufficient condition ensuring that the vector field points strictly inward at the buffered boundary is

$$-s(\delta)m_\zeta + \frac{a_H'(\delta)\dot{\delta}}{2s(\delta)} < 0.$$

To leave fixed slack for the higher-order terms discussed later, we impose the stronger condition: for some fixed $\sigma \in (0,1)$,

$$\frac{a_H'(\delta)\dot{\delta}}{2s(\delta)} < (1-\sigma)s(\delta)m_\zeta. \tag{35}$$

Equivalently,

$$\dot{\delta} < (1-\sigma)\frac{2m_\zeta s(\delta)^2}{a_H'(\delta)}. \tag{36}$$

Under (35), the boundary derivative satisfies

$$\frac{\dot{\chi}}{\chi} \le -s(\delta)m_\zeta + (1-\sigma)s(\delta)m_\zeta = -\sigma s(\delta)m_\zeta < 0.$$

Using

$$s(\delta) = -a_H(\delta) = -\frac{\alpha}{2}\Psi_\theta^*(\delta), \qquad a_H'(\delta) = \frac{\alpha}{2}\Psi_\theta^{*\prime}(\delta),$$

we obtain

$$(1-\sigma)\frac{2m_\zeta s(\delta)^2}{a'_H(\delta)} = (1-\sigma)m_\zeta\alpha\frac{[-\Psi^*_\theta(\delta)]^2}{\Psi^{*\prime}_\theta(\delta)}.$$

Thus the strengthened speed condition

$$0 \le \dot\delta_t < (1-\sigma)m_\zeta\alpha\frac{[-\Psi^*_\theta(\delta_t)]^2}{\Psi^{*\prime}_\theta(\delta_t)}$$

implies that $\dot\chi < 0$ whenever $\chi = 1-\zeta$. Therefore a trajectory starting in the region $\chi \le 1-\zeta$ cannot cross into $\chi > 1-\zeta$. Hence the critical Hopf component remains inside the buffered local basin throughout the intervention. ■

Proof of Corollary 1. The strengthened strict speed bound gives, for some fixed $\sigma \in (0,1)$,

$$\dot\delta_t < (1-\sigma)m_\zeta\alpha\frac{[-\Psi^*_\theta(\delta_t)]^2}{\Psi^{*\prime}_\theta(\delta_t)}.$$

Because $\Psi^{*\prime}_\theta(\delta) > 0$ on $I_H$, this implies

$$\frac{\Psi^{*\prime}_\theta(\delta_t)\dot\delta_t}{[-\Psi^*_\theta(\delta_t)]^2} < (1-\sigma)m_\zeta\alpha.$$

The left-hand side is exactly the time derivative of the inverse stability margin:

$$\frac{d}{dt}\left(\frac{1}{-\Psi^*_\theta(\delta_t)}\right) = \frac{\Psi^{*\prime}_\theta(\delta_t)\dot\delta_t}{[-\Psi^*_\theta(\delta_t)]^2}.$$

Therefore

$$\frac{d}{dt}\left(\frac{1}{-\Psi^*_\theta(\delta_t)}\right) < (1-\sigma)m_\zeta\alpha.$$

Integrating from 0 to $\mathcal{T}$, and using $\delta(0)=\delta_0$ and $\delta_t = \delta_f$, yields

$$\frac{1}{-\Psi^*_\theta(\delta_f)} - \frac{1}{-\Psi^*_\theta(\delta_0)} < (1-\sigma)m_\zeta\alpha\mathcal{T}.$$

Rearranging gives

$$\mathcal{T} > \frac{1}{(1-\sigma)m_\zeta\alpha}\left[\frac{1}{-\Psi^*_\theta(\delta_f)} - \frac{1}{-\Psi^*_\theta(\delta_0)}\right] = \mathcal{T}^\sigma_{\text{crit}}.$$

Finally, as $\delta_f \uparrow \delta_c$, we have $\Psi^*_\theta(\delta_f)\uparrow 0$ from below. Hence $1/[-\Psi^*_\theta(\delta_f)] \to \infty$, so $\mathcal{T}^\sigma_{\text{crit}} \to \infty$. ■

The frozen Hopf normal form describes the local geometry for each fixed value of $\delta$. Since we are studying a time-varying path $\delta = \delta_t$, it remains to check that the additional terms created by the motion of $\delta_t$ do not have the same order as the two terms used in the speed comparison. We show that, after the moving-boundary term already shown above, all remaining $\dot\delta$-terms are higher order.

Let the full Q-value dynamics be

$$\dot{\mathbf{Q}} = \mathbb{F}(\mathbf{Q},\delta),$$

where $\mathbb{F}$ is the vector field of the ODE. For each fixed $\delta$, the symmetric equilibrium satisfies $\mathbb{F}(\mathbf{Q}^*(\delta),\delta) = 0$.

Let $\boldsymbol{\xi} = \mathbf{Q} - \mathbf{Q}^*(\delta_t)$ denote deviations from the moving equilibrium. By the chain rule,

$$\dot{\boldsymbol{\xi}} = \mathbb{F}(\mathbf{Q},\delta) - \partial_\delta\mathbf{Q}^*(\delta)\dot\delta.$$

The second term is the source of the possible moving-equilibrium forcing. The remainder discussion below uses the following assumptions: $\mathbb{F}$ and $\mathbf{Q}^*(\delta)$ are sufficiently smooth in $(\mathbf{Q},\delta)$ and all noncritical eigenvalues are uniformly bounded away from the imaginary axis on $I$.

Recall the notation introduced earlier: $\mathbf{u} = (\boldsymbol{\xi}_1+\boldsymbol{\xi}_2)/2$, $\mathbf{v} = (\boldsymbol{\xi}_1-\boldsymbol{\xi}_2)/2$, where $\boldsymbol{\xi}_i$ is seller $i$'s deviation vector. Since $\mathbf{Q}^*(\delta)$ is symmetric, at the seller level the last term above contributes the same vector

$$-\partial_\delta\begin{pmatrix} Q^*_{\text{C}}(\delta) \\ Q^*_{\text{D}}(\delta)\end{pmatrix}\dot\delta$$

to each seller's deviation equation. This term enters the symmetric coordinate $\mathbf{u}$, but cancels from the antisymmetric coordinate $\mathbf{v}$. Since the Hopf pair is generated by the antisymmetric block $\mathbf{U}-\mathbf{V}$, there is no constant forcing term of order $O(\dot{\delta})$ in the Hopf equation.

Locally, the $(\mathbf{u},\mathbf{v})$ dynamics can be written as

$$\begin{aligned}\dot{\mathbf{u}} &= (\mathbf{U}+\mathbf{V})\mathbf{u} - \partial_\delta \begin{pmatrix} Q_{\mathrm{C}}^*(\delta) \\ Q_{\mathrm{D}}^*(\delta) \end{pmatrix} \dot{\delta} + G_s(\mathbf{u},\mathbf{v},\delta),\\ \dot{\mathbf{v}} &= (\mathbf{U}-\mathbf{V})\mathbf{v} + G_c(\mathbf{u},\mathbf{v},\delta),\end{aligned}$$

where the terms $G_s$ and $G_c$ start at quadratic order in $(\mathbf{u},\mathbf{v})$. The moving-equilibrium forcing affects only the stable $\mathbf{u}$-equation. Thus this forcing produces only an $O(|\dot{\delta}|)$ in the symmetric coordinate.

The only possible feedback from the moving symmetric coordinate $\mathbf{u}$ to the antisymmetric coordinate $\mathbf{v}$ must vanish when $\mathbf{v}=0$. Indeed, if $\mathbf{v}=0$, the two sellers have the same deviation from the moving equilibrium. Equivalently, the subspace $\mathbf{v}=0$ is invariant, so

$$G_c(\mathbf{u},0,\delta)=0.$$

Hence the antisymmetric equation cannot contain a term depending only on $\mathbf{u}$ or only on the moving-equilibrium forcing. Any feedback from the symmetric coordinate into the antisymmetric equation must be multiplied by the seller-difference variable $\mathbf{v}$. The moving part of this feedback therefore has size

$$O(|\dot{\delta}|\,\|\mathbf{v}\|),$$

where the $|\dot{\delta}|$ term is coming from the dynamics in $\mathbf{u}$. Since $z$ is a local coordinate on the Hopf plane, $\mathbf{v}$ is linear in $z$ and $\bar{z}$ to first order:

$$\mathbf{v} = \mathbf{q}(\delta)z + \overline{\mathbf{q}(\delta)}\,\bar{z} + O(|z|^2).$$

Thus the moving feedback contributes $O(|\dot{\delta}|\,|z|)$ to the $z$-equation.

Moreover, there are moving-coordinate terms. Let $z=h(\mathbf{v},\delta)$ be a local Hopf normal-form coordinate. Since $z(t)=h(\mathbf{v}(t),\delta_t)$, the chain rule gives

$$\dot{z} = D_{\mathbf{v}}h(\mathbf{v},\delta)\dot{\mathbf{v}} + \partial_\delta h(\mathbf{v},\delta)\dot{\delta}.$$

The first term gives the frozen Hopf normal form. The second term is new, but it has no constant part. Indeed, the equilibrium corresponds to $\mathbf{v}=0$ for every $\delta$ in the Hopf coordinate, so $h(0,\delta)=0$ and therefore

$$\partial_\delta h(0,\delta)=0.$$

Thus the Taylor expansion of $\partial_\delta h(\mathbf{v},\delta)$ begins with terms that are linear in $\mathbf{v}$. Using the first-order relation between $\mathbf{v}$ and $(z,\bar{z})$ above, a real linear term in $\mathbf{v}$ becomes a linear combination of $z$ and $\bar{z}$. Therefore, with bounded local coefficients,

$$\partial_\delta h(\mathbf{v},\delta)\dot{\delta} = \dot{\delta}\left[b_1(\delta)z + b_2(\delta)\bar{z} + O(|z|^2)\right].$$

Thus the moving-coordinate contribution has the form

$$R_\delta = O(|\dot{\delta}|\,|z|)$$

in the complex $z$-equation. The same bound applies to the feedback from the moving symmetric coordinate discussed above.

It remains to translate this bound into the radial growth rate. Let $r=|z|$. Since $r^2=|z|^2=z\bar{z}$, we have

$$\frac{d}{dt}r^2 = \dot{z}\bar{z} + z\dot{\bar{z}} = 2\,\mathrm{Re}(\bar{z}\dot{z}).$$

Hence, for $r > 0$,

$$\frac{\dot{r}}{r} = \frac{\mathrm{Re}(\bar{z}\dot{z})}{|z|^2}.$$

The contribution of $R_\delta$ is therefore

$$\frac{\mathrm{Re}(\bar{z}R_\delta)}{|z|^2}.$$

Since $R_\delta = O(|\dot{\delta}|\,|z|)$,

$$\left|\frac{\mathrm{Re}(\bar{z}R_\delta)}{|z|^2}\right| \le \frac{|z|\,|R_\delta|}{|z|^2} = O(|\dot{\delta}|).$$

The extra factor of $|z|$ in the $z$-equation is exactly cancelled when we convert to $\dot{r}/r$.

Recall from the proof that the normalized amplitude $\chi = r/\rho_*(\delta)$ satisfies

$$\frac{\dot{\chi}}{\chi} = -s(\delta)(1-\chi^2) + \frac{a'_H(\delta)\dot{\delta}}{2s(\delta)} + \mathcal{E}, \qquad s(\delta) := -a_H(\delta) > 0,$$

where $\mathcal{E}$ collects the frozen normal-form remainders and the moving-$\delta$ remainders discussed above. On the local Hopf boundary, $|z| = r = O(\sqrt{s(\delta)})$, and the speed condition implies $\dot{\delta} = O(s(\delta)^2)$. Therefore $R_\delta$'s contribution to $\dot{r}/r$ is

$$O(|\dot{\delta}|) = O(s^2).$$

The possible variation of $\ell_H$ with $\delta$ is also of higher order. To see this, let

$$\ell_H^c := \ell_H(\delta_c).$$

Since $\ell_H(\delta)$ is smooth and $\ell_H^c \neq 0$, $\ell_H(\delta)$ remains bounded away from zero and

$$\ell_H(\delta) - \ell_H^c = O(|\delta - \delta_c|).$$

Moreover, from the Taylor expansion, we have $a_H(\delta) = O(\delta - \delta_c)$, so $s(\delta) = -a_H(\delta) = O(|\delta - \delta_c|)$. Hence, on the local Hopf boundary where $r = O(\sqrt{s})$, replacing $\ell_H(\delta)$ by $\ell_H^c$ changes the cubic contribution to the logarithmic radial equation by

$$\big(\ell_H(\delta) - \ell_H^c\big) r^2 = O(|\delta - \delta_c|\, s) = O(s^2).$$

The same conclusion holds for the moving-boundary term. If one uses

$$\rho_*^2(\delta) = \frac{s(\delta)}{\ell_H(\delta)},$$

then

$$\frac{\dot{\rho}_*}{\rho_*} = \frac{1}{2}\frac{\dot{s}}{s} - \frac{1}{2}\frac{\ell'_H(\delta)}{\ell_H(\delta)}\dot{\delta}.$$

The first term gives $-a'_H(\delta)\dot{\delta}/(2s(\delta))$. The second term is $O(\dot{\delta})$, and under the speed scaling $\dot{\delta} = O(s^2)$, it is also $O(s^2)$. Therefore the variation of $\ell_H$ with $\delta$ affects $\dot{\chi}/\chi$ only at order $O(s^2)$, and these terms are included in $\mathcal{E}$.

By contrast, the two terms retained in the leading-order comparison,

$$-s(\delta)(1-\chi^2), \qquad \frac{a'_H(\delta)\dot{\delta}}{2s(\delta)},$$

are both $O(s)$ under the speed scaling $\dot{\delta} = O(s^2)$. Here $a'_H(\delta) = O(1)$ near $\delta_c$, and becasuse the Hopf bifurcation , we have $a'_H(\delta_c) \neq 0$. In particular, at the buffered boundary $\chi = 1 - \zeta$, the fixed-slack speed condition gives the leading-order inward margin

$$-s(\delta)m_\zeta + \frac{a'_H(\delta)\dot{\delta}}{2s(\delta)} < -\sigma m_\zeta s(\delta).$$

The remaining moving-$\delta$ terms and the frozen normal-form remainders enter the logarithmic radial equation at order $O(s^2)$. Thus these terms are asymptotically smaller than the $O(s)$ terms and do not change the leading-order speed law.

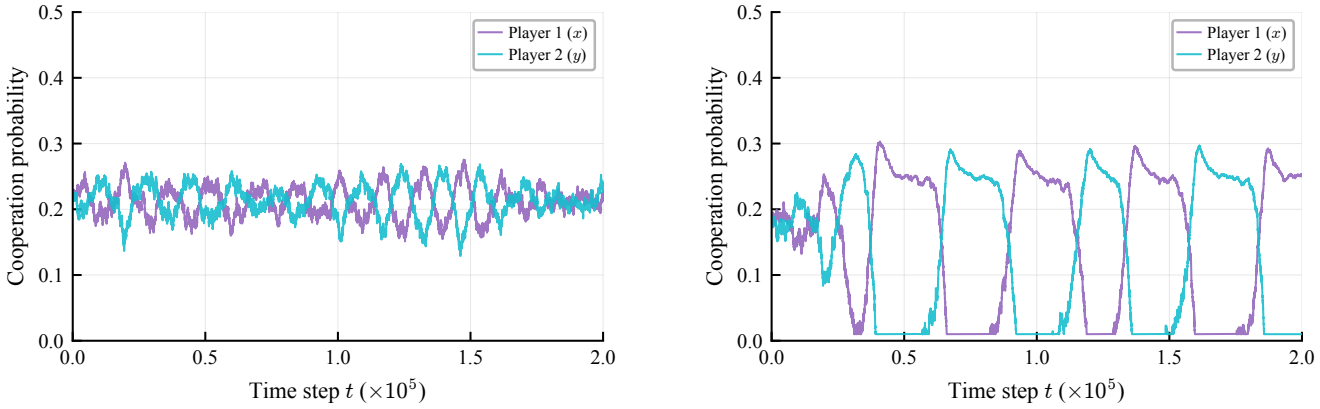
# Appendix B: Additional Results

## B.1. Discrete-Time Q-Learning Simulations

This appendix section reports numerical experiments for the original discrete-time Q-learning process underlying the continuous-time (ODE) approximation analyzed in the main text. The purpose of these simulations is to verify that the phenomena identified in the fluid limit — stability and oscillations present for the discrete-time system as well.

In Figure 8a, we first consider a fixed intervention level $\delta = 0.2 < \delta_c$, where $\delta_c$ is the critical threshold identified in the continuous-time analysis. We can see that the discrete-time learning process oscillates near the neighborhood of the symmetric interior equilibrium.

In Figure 8b, for $\delta = 0.3 > \delta_c$, convergence fails. Instead, the learning dynamics exhibit large-amplitude oscillations in action probabilities. These cycles do not vanish as the simulation horizon increases and are robust across random seeds. This behavior is consistent with the loss of local stability of the interior equilibrium through a Hopf bifurcation in the fluid limit.




(a) $\delta$ below the critical threshold: stable oscillations around the fixed point.

(b) $\delta$ above the critical threshold: divergence from the fixed point over time.

**Figure 8 Collusive level at the interior fixed point for intervention intensities below and above the critical threshold for discrete-time Q-learning.**

Overall, the discrete-time simulations confirm that the stability boundary predicted by the continuous-time approximation has strong predictive power for the actual learning dynamics.

## B.2. Uniqueness of the Interior Fixed Point

Having established the condition for the stability of the symmetric interior fixed point $(x^*, x^*)$ in Proposition 1, we now derive the sufficient conditions under which this fixed point is unique. To demonstrate uniqueness, we must establish that no other stable fixed points exist on the boundary of the strategy space $\Sigma = [\epsilon, 1-\epsilon]^2$. Specifically, we must rule out the mutual boundary points (e.g., $(\epsilon, \epsilon)$) and the asymmetric semi-interior points (e.g., $(\epsilon, y^*)$).

**Non-existence of the Boundary Fixed Point** $(\epsilon, \epsilon)$ First, we have that the best-response function is $f : [\epsilon, 1-\epsilon] \to \mathbb{R}$, defined by the linear portion of the policy:

$$f(x) = \tau(Ax + B) + \frac{1}{2}.$$

A symmetric fixed point $x^*$ satisfies $x^* = f(x^*)$. For this interior fixed point to be asymptotically stable, as is shown in Proposition 1, we have:

$$A\tau < 1.$$

We now analyze the stability of the boundary profile $(\epsilon, \epsilon)$. For this boundary point to be a fixed point, we must have $f(\epsilon) \leq \epsilon$.

Since $f(x)$ is linear and $A\tau < 1$, condition $f(\epsilon) \leq \epsilon$ and $x^* = f(x^*)$ cannot satisfy simultaneously. To see this, note that if $f(\epsilon) \leq \epsilon$, then the line $y = f(x)$ lies below the line $y = x$ at $x = \epsilon$. Given that the slope of $f(x)$ is less than 1, it follows that $f(x)$ will remain below $y = x$ for all $x > \epsilon$. This contradicts the existence of an interior fixed point $x^* > \epsilon$ where $f(x^*) = x^*$. Therefore, if a stable interior fixed point exists, the boundary point $(\epsilon, \epsilon)$ does not exist.

**Non-existence of Asymmetric Fixed Points** We next consider asymmetric fixed points of the form $(x^*, y^*) = (\epsilon, y^*)$ with $y^* \in (\epsilon, 1-\epsilon)$. As shown in the previous section, such a fixed point requires Player 1 to be pinned at the boundary when Player 2 playing an interior strategy. The necessary stability condition for Player 1 is:

$$x^* = \tau(Ay^* + B) + \frac{1}{2} \leq \epsilon. \tag{37}$$

Here, $y^*$ is Player 2's best response to Player 1's boundary strategy $x = \epsilon$. Since $y^*$ is assumed to be interior, it must satisfy the linear policy equation:

$$y^* = \tau(A\epsilon + B) + \frac{1}{2}. \tag{38}$$

To ensure the non-existence of such asymmetric equilibria, condition (37) must be violated. Substituting (38) into equation (37), we require:

$$\tau\left[Ay^* + B\right] + \frac{1}{2} > \epsilon. \tag{39}$$

Substituting the expression for $y^*$, we have:

$$\tau\left[A\left(\tau(A\epsilon + B) + \frac{1}{2}\right) + B\right] + \frac{1}{2} > \epsilon.$$

This is the necessary and sufficient condition for the non-existence of asymmetric fixed points of the form $(\epsilon, y^*)$.

We then show that this condition is satisfied when $A > 0$. As is hown above, when the interior is stable, we have $A\tau < 1$. Thus, we have $f(\epsilon) > \epsilon$. Therefore, we have

$$z = \tau(A\epsilon + B) + \frac{1}{2} = f(\epsilon) > \epsilon.$$

Moreover, when $A > 0$, $f(z) = \tau(Az + B) + 1/2$ is strictly increasing in $z$, we have

$$\tau\left[A\left(\tau(A\epsilon + B) + \frac{1}{2}\right) + B\right] + \frac{1}{2} = \tau\left[Az + B\right] + \frac{1}{2} = f(z) > \epsilon,$$

where the last inequality holds because $z > \epsilon$ and $f$ is strictly increasing. Thus, we have shown that when $A > 0$, there are no asymmetric fixed points of the form $(\epsilon, y^*)$.

### B.3. Computation of the First Lyapunov Coefficient

This sensitivity to speed is explained by the specific local structure of the Hopf bifurcation. To classify the stability of the emerging limit cycles, we compute the first Lyapunov coefficient $\ell_1$ at the critical intervention level $\delta_c$. This computation relies on the reduction of the system to its center manifold using the normal form theory described in Kuznetsov (2004).

Let $\mathbb{F}(\mathbf{Q}; \delta)$ denote the right-hand side of equations (9). At the Hopf point, the system is linearized at the fixed point $\mathbf{Q}^*(\delta_c)$. Let

$$\mathbf{J}_c := D_{\mathbf{Q}}\mathbb{F}(\mathbf{Q}^*(\delta_c); \delta_c)$$

denote the Jacobian at this point. The critical eigenvalues of $\mathbf{J}_c$ are

$$\lambda_{\pm} = \pm i\omega_c, \qquad \omega_c > 0,$$

where $\omega_c > 0$ is the angular frequency at the Hopf bifurcation.

We introduce the right eigenvector $\mathbf{q}$ and the left eigenvector $\mathbf{p}$ satisfying

$$\mathbf{J}_c\mathbf{q} = i\omega_c\mathbf{q}, \qquad \mathbf{J}_c^T\mathbf{p} = -i\omega_c\mathbf{p},$$

and normalize them so that

$$\langle \mathbf{p}, \mathbf{q} \rangle = 1.$$

Here $\langle \cdot, \cdot \rangle$ denotes the complex inner product used in the standard Hopf normal form calculation.

Expanding $\mathbb{F}(\mathbf{Q}; \delta_c)$ around $\mathbf{Q}^*(\delta_c)$, write

$$\mathbb{F}(\mathbf{Q}^*(\delta_c) + \boldsymbol{\xi}; \delta_c) = \mathbf{J}_c\boldsymbol{\xi} + \frac{1}{2}\mathcal{B}(\boldsymbol{\xi}, \boldsymbol{\xi}) + \frac{1}{6}\mathcal{C}(\boldsymbol{\xi}, \boldsymbol{\xi}, \boldsymbol{\xi}) + O(\|\boldsymbol{\xi}\|^4).$$

The multilinear maps

$$\mathcal{B} : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}^n, \qquad \mathcal{C} : \mathbb{C}^n \times \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}^n$$

are defined component-wise by

$$\mathcal{B}_i(\mathbf{u}, \mathbf{v}) = \sum_{j,k=1}^{n} \left.\frac{\partial^2 \mathbb{F}_i}{\partial \xi_j \partial \xi_k}\right|_{\mathbf{Q}^*(\delta_c)} u_j v_k, \qquad \mathcal{C}_i(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \sum_{j,k,l=1}^{n} \left.\frac{\partial^3 \mathbb{F}_i}{\partial \xi_j \partial \xi_k \partial \xi_l}\right|_{\mathbf{Q}^*(\delta_c)} u_j v_k w_l. \tag{40}$$

The first Lyapunov coefficient $\ell_1$ is determined by the interaction of these nonlinear terms with the critical eigenspace:

$$\ell_1 = \frac{1}{2\omega_c} \operatorname{Re}\left[ \langle \mathbf{p}, \mathcal{C}(\mathbf{q}, \mathbf{q}, \bar{\mathbf{q}}) \rangle - 2 \left\langle \mathbf{p}, \mathcal{B}\left(\mathbf{q}, \mathbf{J}_c^{-1}\mathcal{B}(\mathbf{q}, \bar{\mathbf{q}})\right) \right\rangle + \left\langle \mathbf{p}, \mathcal{B}\left(\bar{\mathbf{q}}, (2i\omega_c\mathbf{I}_n - \mathbf{J}_c)^{-1}\mathcal{B}(\mathbf{q}, \mathbf{q})\right) \right\rangle \right], \tag{41}$$

where $\operatorname{Re}(\cdot)$ extracts the real part and $\bar{\mathbf{q}}$ denotes the complex conjugate of the eigenvector $\mathbf{q}$.

The coefficient $\ell_1$ in (41) is the standard first Lyapunov coefficient under the convention of Kuznetsov (2004). Under this convention, the sign of $\ell_1$ determines the criticality of the Hopf bifurcation. Since $\omega_c > 0$, a positive value of $\ell_1$ implies that the Hopf bifurcation is subcritical. A subcritical Hopf bifurcation is characterized, on the stable side $\delta < \delta_c$, by an unstable limit cycle surrounding the stable equilibrium. Trajectories inside this unstable cycle converge to the stable fixed point, while trajectories outside the cycle may move toward a large-amplitude oscillatory attractor.

For the speed-bound analysis, it is useful to note the convention connecting $\ell_1$ to the radial Hopf normal form. With the definition (41), the real cubic coefficient in the radial normal form is $\omega_c\ell_1$. Thus, locally,

$$\dot{r} = a_H(\delta)r + \omega_c\ell_1 r^3 + \text{higher-order terms}, \qquad r = |z|.$$

Therefore, when $\ell_1 > 0$, the cubic term is outward and the Hopf bifurcation is subcritical.

**B.3.1. Numerical Verification of Subcriticality** To show that the subcritical nature of the Hopf bifurcation is not an artifact of a specific parameter choice, we generated $N = 10000$ random payoff configurations, sampling the payoff parameters $T, R, P$, and $S$ from uniform distributions:

$$T \in [2.5, 4.0], \qquad R \in [1.5, 3.0], \qquad P \in [1.0, 2.5], \qquad S \in [0.5, 2.0].$$

The learning parameters were fixed at

$$\alpha = 0.002, \qquad \gamma = 0.6, \qquad \tau = 0.5, \qquad \epsilon = 0.01, \qquad \theta = 1.0.$$

For each configuration, we numerically solved for the critical intervention level $\delta_c$ and the corresponding symmetric equilibrium $(x^*, x^*)$. The analysis was conditioned on two criteria:

1. The existence of a valid Hopf bifurcation, meaning that the critical eigenvalues are purely imaginary and have nonzero imaginary part $\omega_c > 0$.

2. The preservation of the Prisoner's Dilemma structure at the critical threshold. Specifically, with effective payoffs

$$T' = T + \delta_c, \qquad S' = S - \delta_c, \qquad R' = R, \qquad P' = P,$$

we required

$$T' > R' > P' > S'.$$

For every parameter set satisfying these conditions, we computed the first Lyapunov coefficient $\ell_1$ using the method described above. The numerical results show that $\ell_1 > 0$ for all valid configurations in the sample. This evidence suggests that subcriticality is a robust feature of the learning dynamics under this model, rather than a consequence of a particular payoff configuration.

## Appendix C: Derivation of the Fluid Approximation

In this appendix, we derive the fluid approximation used in Section 3.3. Consider a sequence of discrete-time systems indexed by the period length $\Delta > 0$. In the $\Delta$-system, decisions are made at times $t_k = k\Delta$, $k = 0, 1, 2, \ldots$, and the learning rate is scaled as

$$\boldsymbol{\alpha}_\Delta(\mathbf{s}) = \boldsymbol{\alpha}(\mathbf{s})\Delta.$$

We take $\mathbf{s}_0^\Delta = \mathbf{s}_0$ for all $\Delta$. Let $\mathbf{s}_k^\Delta$ denote the state at decision epoch $t_k$, let $\delta_k^\Delta$ denote the intervention level, let $\mathbf{p}_k^\Delta = \boldsymbol{\pi}(\mathbf{s}_k^\Delta)$ denote the mixed action chosen at that epoch, and let $\mathbf{r}_k^\Delta$ denote the corresponding payoff-feedback vector under $\delta_k^\Delta$. The state evolves according to

$$\mathbf{s}_{k+1}^\Delta = (\mathbf{1}_d - \boldsymbol{\alpha}_\Delta(\mathbf{s}_k^\Delta)) \odot \mathbf{s}_k^\Delta + \boldsymbol{\alpha}_\Delta(\mathbf{s}_k^\Delta) \odot \widetilde{\mathcal{T}}_{\mathrm{QR}}(\mathbf{s}_k^\Delta, \mathbf{p}_k^\Delta, \mathbf{r}_k^\Delta).$$

Equivalently,

$$\frac{\mathbf{s}_{k+1}^\Delta - \mathbf{s}_k^\Delta}{\Delta} = \boldsymbol{\alpha}(\mathbf{s}_k^\Delta) \odot \left( \widetilde{\mathcal{T}}_{\mathrm{QR}}(\mathbf{s}_k^\Delta, \mathbf{p}_k^\Delta, \mathbf{r}_k^\Delta) - \mathbf{s}_k^\Delta \right) =: \mathbb{H}(\mathbf{s}_k^\Delta, \mathbf{p}_k^\Delta, \mathbf{r}_k^\Delta). \tag{42}$$

Define the averaged drift

$$\overline{\mathbb{H}}(\mathbf{s}, \delta) := \mathbb{E}_\delta \left[ \mathbb{H}(\mathbf{s}, \boldsymbol{\pi}(\mathbf{s}), \mathbf{r}) \right],$$

where the expectation is taken over the payoff realization and the action randomization induced by the mixed action $\mathbf{p} = \boldsymbol{\pi}(\mathbf{s})$, including any dependence of payoffs on rival actions. The subscript $\delta$ indicates that the payoff-feedback distribution is evaluated at intervention level $\delta$. Define the martingale-difference term

$$\mathbb{M}_{k+1}^\Delta := \mathbb{H}(\mathbf{s}_k^\Delta, \mathbf{p}_k^\Delta, \mathbf{r}_k^\Delta) - \overline{\mathbb{H}}(\mathbf{s}_k^\Delta, \delta_k^\Delta).$$

Then (42) can be written in stochastic-approximation form as

$$\mathbf{s}_{k+1}^\Delta = \mathbf{s}_k^\Delta + \Delta\, \overline{\mathbb{H}}(\mathbf{s}_k^\Delta, \delta_k^\Delta) + \Delta\, \mathbb{M}_{k+1}^\Delta. \tag{43}$$

Let

$$\mathcal{F}_k^\Delta := \sigma(\mathbf{s}_0^\Delta, \delta_0^\Delta, \mathbf{p}_0^\Delta, \mathbf{r}_0^\Delta, \ldots, \mathbf{p}_{k-1}^\Delta, \mathbf{r}_{k-1}^\Delta, \mathbf{s}_k^\Delta, \delta_k^\Delta)$$

denote the natural filtration. By construction,

$$\mathbb{E}\left[ \mathbb{M}_{k+1}^\Delta \mid \mathcal{F}_k^\Delta \right] = 0.$$

Let $\mathbf{s}^\Delta(\cdot)$ be the piecewise-linear interpolation of the discrete-time process:

$$\mathbf{s}_t^\Delta = \mathbf{s}_k^\Delta + \frac{t - t_k}{\Delta} \left( \mathbf{s}_{k+1}^\Delta - \mathbf{s}_k^\Delta \right), \qquad t \in [t_k, t_{k+1}).$$

Fix a finite time horizon $\mathcal{T} < \infty$. For $t \in [0, \mathcal{T}]$, let $n = \lfloor t/\Delta \rfloor$, define $\underline{u} = t_{\lfloor u/\Delta \rfloor}$ as the most recent decision epoch before time $u$, and write $\delta_{\underline{u}}^\Delta = \delta_{\lfloor u/\Delta \rfloor}^\Delta$. Summing (43) gives

$$\mathbf{s}_t^\Delta = \mathbf{s}_0 + \int_0^t \overline{\mathbb{H}}(\mathbf{s}_{\underline{u}}^\Delta, \delta_{\underline{u}}^\Delta)\, du + \sum_{j=0}^{n-1} \Delta\, \mathbb{M}_{j+1}^\Delta + (t - t_n)\mathbb{M}_{n+1}^\Delta.$$

Since $0 \le t - t_n < \Delta$, the final term is $O(\Delta)$ whenever the martingale differences are uniformly bounded. Hence,

$$\mathbf{s}_t^\Delta = \mathbf{s}_0 + \int_0^t \overline{\mathbb{H}}(\mathbf{s}_{\underline{u}}^\Delta, \delta_{\underline{u}}^\Delta)\, du + \sum_{j=0}^{n-1} \Delta\, \mathbb{M}_{j+1}^\Delta + O(\Delta). \tag{44}$$

We impose the following standard regularity conditions. First, the iterates eventually remain in a compact state set $\mathbb{S}$. In our setting, this follows from the dissipativity property established in Proposition 1. Second, payoffs and the intervention sequence are bounded. Third, on $\mathbb{S}$, the policy map, the update rule, and the learning-rate map $\boldsymbol{\alpha}(\cdot)$ are bounded and Lipschitz. These conditions imply that $\mathbb{H}(\mathbf{s}, \mathbf{p}, \mathbf{r})$ is uniformly bounded and Lipschitz in $\mathbf{s}$ on $\mathbb{S}$, and therefore that $\overline{\mathbb{H}}(\mathbf{s}, \delta)$ is also bounded and Lipschitz in $\mathbf{s}$, uniformly over admissible $\delta$.

These bounds imply tightness of the family of interpolated processes $\{\mathbf{s}^{\Delta}(\cdot)\}_{\Delta>0}$ on $C([0,\mathcal{T}];\mathbb{R}^d)$ for every finite horizon $\mathcal{T}$. In particular, the interpolated paths remain in a common compact set and have a uniformly bounded modulus of continuity. This is the standard weak-convergence tightness condition used in stochastic-approximation fluid limits; see, e.g., Kushner and Yin (2003).

It remains to show that the martingale term in (44) vanishes. Since $\mathbb{H}$ is uniformly bounded on the relevant compact set, the martingale differences satisfy

$$\sup_{k,\Delta} \mathbb{E}\left[\|\mathbb{M}_{k+1}^{\Delta}\|^2 \mid \mathcal{F}_k^{\Delta}\right] < \infty.$$

Because $\{\mathbb{M}_{k+1}^{\Delta}\}$ is a martingale-difference sequence, cross terms vanish. Therefore, for every fixed $\mathcal{T} < \infty$,

$$\mathbb{E}\left\|\sum_{j=0}^{\lfloor \mathcal{T}/\Delta \rfloor - 1} \Delta\, \mathbb{M}_{j+1}^{\Delta}\right\|^2 \le \sum_{j=0}^{\lfloor \mathcal{T}/\Delta \rfloor - 1} \Delta^2 \mathbb{E}\left[\|\mathbb{M}_{j+1}^{\Delta}\|^2\right] = O(\Delta).$$

Thus the martingale term converges to zero in $L^2$, and hence in probability, uniformly on finite horizons.

Combining tightness, the Lipschitz continuity of $\overline{\mathbb{H}}(\cdot,\delta)$ uniformly over the admissible intervention path, and the vanishing martingale term, every weak limit of $\mathbf{s}^{\Delta}(\cdot)$ satisfies

$$\mathbf{s}_t = \mathbf{s}_0 + \int_0^t \overline{\mathbb{H}}(\mathbf{s}_u, \delta_u)\, du.$$

Since $\overline{\mathbb{H}}$ is Lipschitz in $\mathbf{s}$ uniformly along the bounded intervention path, this integral equation has a unique solution. Therefore the full sequence of interpolated processes converges weakly on finite horizons to the unique solution of

$$\dot{\mathbf{s}}_t = \overline{\mathbb{H}}(\mathbf{s}_t, \delta_t), \qquad \mathbf{s}(0) = \mathbf{s}_0. \tag{45}$$

This is the ODE, or fluid, approximation used in the main analysis. See Kushner and Yin (2003) and Borkar (2008) for standard treatments of stochastic approximation and fluid limits. ■

## Appendix D: General Fixed-Point Characterization

This appendix characterizes all the fixed points of (9)–(10) without imposing Assumption 1. The main text focuses on the parameter region in which the system has a unique symmetric interior fixed point. Outside that region, the symmetric fixed point can lie on the forced-exploration boundary, asymmetric fixed points can arise, and at a knife-edge parameter value the system can admit a continuum of fixed points.

Recall the cooperation–defection Q-value gaps

$$\Delta_1(t) := Q_{\text{C}}^1(t) - Q_{\text{D}}^1(t), \qquad \Delta_2(t) := Q_{\text{C}}^2(t) - Q_{\text{D}}^2(t).$$

By Lemma 1, $\Delta_i^* < 0$ at every fixed point. Hence the upper saturation $1-\epsilon$ of the policy map never binds at a fixed point. The action probabilities and Q-value gaps therefore satisfy

$$\begin{aligned} A\,y^* + B - \delta - \Delta_1^* &= 0, \qquad & x^* &= \epsilon + \left(\tfrac{1}{2} + \tau\Delta_1^* - \epsilon\right)^+, \\ A\,x^* + B - \delta - \Delta_2^* &= 0, \qquad & y^* &= \epsilon + \left(\tfrac{1}{2} + \tau\Delta_2^* - \epsilon\right)^+, \end{aligned} \tag{Equilibrium Conditions}$$

where $z^+ := \max\{z, 0\}$.

These equations characterize the action probabilities and Q-value gaps at a fixed point. Given any solution $(x^*, y^*, \Delta_1^*, \Delta_2^*)$, the Q-value levels are uniquely recovered from the Bellman fixed-point equations. Specifically, define

$$u_{\text{C}}^1(y, \delta) = yR + (1-y)(S-\delta), \qquad u_{\text{D}}^1(y, \delta) = y(T+\delta) + (1-y)P,$$

and analogously

$$u_{\text{C}}^2(x, \delta) = xR + (1-x)(S-\delta), \qquad u_{\text{D}}^2(x, \delta) = x(T+\delta) + (1-x)P.$$

Since $Q_{\text{D}}^{i*} > Q_{\text{C}}^{i*}$, the continuation term is $\gamma Q_{\text{D}}^{i*}$, and hence

$$Q_{\text{D}}^{1*} = \frac{u_{\text{D}}^1(y^*, \delta)}{1-\gamma}, \qquad Q_{\text{C}}^{1*} = u_{\text{C}}^1(y^*, \delta) + \gamma Q_{\text{D}}^{1*},$$

with the analogous expressions for firm 2. Thus, solving the equilibrium conditions in $(x, y, \Delta_1, \Delta_2)$ is equivalent to solving for fixed points of the original Q-value system.

The following proposition gives the full classification. Recall the definition of the policy-saturation threshold:

$$\tau_\epsilon(\delta) := \left(\frac{1/2 - \epsilon}{\delta - A\epsilon - B}\right)^+. \tag{Policy-Saturation Threshold}$$

**Proposition 8 (General fixed-point characterization)** *Fix $\delta \in [0, S]$. The equilibria of (9)–(10) are characterized as follows.*

(a) Symmetric. *The system admits a unique symmetric equilibrium, $x^* = y^* = a^*(\delta)$ and $\Delta_1^* = \Delta_2^* = \Delta^*$, where $\Delta^* = Aa^*(\delta) + B - \delta$ and*

$$a^*(\delta) = \begin{cases} \dfrac{1 + 2\tau(B-\delta)}{2(1-A\tau)} & \text{if} \quad 0 < \tau < \tau_\epsilon(\delta) \qquad \text{(collusive outcome)} \\ \epsilon & \text{if} \quad \tau_\epsilon(\delta) \le \tau \qquad \text{(non-collusive outcome).} \end{cases}$$

(b) Asymmetric. *If $A \ge 0$, no asymmetric equilibrium exists. Suppose $A < 0$ and let $\tau_0 := -1/A$.*

(i) *If $\tau_0 < \tau < \tau_\epsilon(\delta)$, there are exactly two asymmetric partially-collusive equilibria:*

$$(x^*, y^*) = \left(\epsilon, \tfrac{1}{2} + \tau(A\epsilon + B - \delta)\right) \quad \text{and} \quad (x^*, y^*) = \left(\tfrac{1}{2} + \tau(A\epsilon + B - \delta), \epsilon\right).$$

*The corresponding Q-value gaps are*

$$(\Delta_1^*, \Delta_2^*) = (Ay^* + B - \delta, A\epsilon + B - \delta)$$

*and*

$$(\Delta_1^*, \Delta_2^*) = (A\epsilon + B - \delta, Ax^* + B - \delta),$$

*respectively.*

(ii) *If* $\tau = \tau_0 < \tau_\epsilon(\delta)$*, the system admits a continuum of equilibria satisfying*

$$x^* + y^* = \frac{1}{2} + \tau(B - \delta),$$

*with*

$$x^*, y^* \in \left[\epsilon,\ \tfrac{1}{2} + \tau(A\epsilon + B - \delta)\right].$$

*All non-symmetric points in this continuum are asymmetric equilibria, and their Q-value gaps are*

$$\Delta_1^* = Ay^* + B - \delta, \qquad \Delta_2^* = Ax^* + B - \delta.$$

(iii) *In all remaining cases, no asymmetric equilibrium exists.*

The proof of this result is provided in Appendix A (see proof of Proposition 3).

**Corollary 2** *If* $A\tau > -1$*, then every equilibrium of* (9)–(10) *is symmetric.*

Proposition 8 clarifies what Assumption 1 excludes. When $\tau \geq \tau_\epsilon(\delta)$, the unique symmetric equilibrium is the mutual-boundary fixed point $x^* = y^* = \epsilon$, so cooperation occurs only through forced exploration. When $A < 0$ and $-1/A < \tau < \tau_\epsilon(\delta)$, the system admits two asymmetric partially-collusive equilibria: one firm is pinned at the forced-exploration bound, while the other remains in the interior of the policy map. At the knife-edge value $\tau = -1/A$, these asymmetric equilibria become part of a continuum.

Thus, outside the unique-interior regime, changes in $\tau$ can alter not only the level of cooperation but also the structure of the fixed-point set. These cases are useful for understanding the full geometry of the learning dynamics, but they are not the main object of the paper. The main text therefore imposes Assumption 1 and studies how platform intervention affects the level and stability of the single collusive fixed point.

Figure 9 extends the example in Figure 2. Specifically, the right panel depicts a case or which the conditions in Assumption 1 are not satisfies. The dashed curves are the zero-drift loci $\dot{\Delta}_1 = 0$ and $\dot{\Delta}_2 = 0$; their intersections are the stationary points of (12). The shaded regions identify states in which one or both firms are at the forced-exploration lower bound.

In the left panel, $\tau = 0.5 < \tau_0$. The zero-drift curves intersect only once, on the diagonal $\Delta_1 = \Delta_2$, yielding a stable symmetric equilibrium. For the displayed parameters, $\Delta_1 = \Delta_2 \approx -0.429$, corresponding to $x = y \approx 0.286$. The dynamics therefore converge to an interior outcome with cooperation above the forced-exploration benchmark.

In the right panel, $\tau = 1.5 \in (\tau_0, \tau_\epsilon)$. The symmetric equilibrium remains interior, at approximately $\Delta_1 = \Delta_2 = -0.272$ and $x = y \approx 0.091$, but it becomes a saddle. Two asymmetric stable equilibria emerge: in one, firm 1 is at the forced-exploration bound while firm 2 cooperates with an interior probability; in the other, the roles are reversed. Thus, greater responsiveness of the policy map can create endogenous asymmetry across otherwise identical firms.

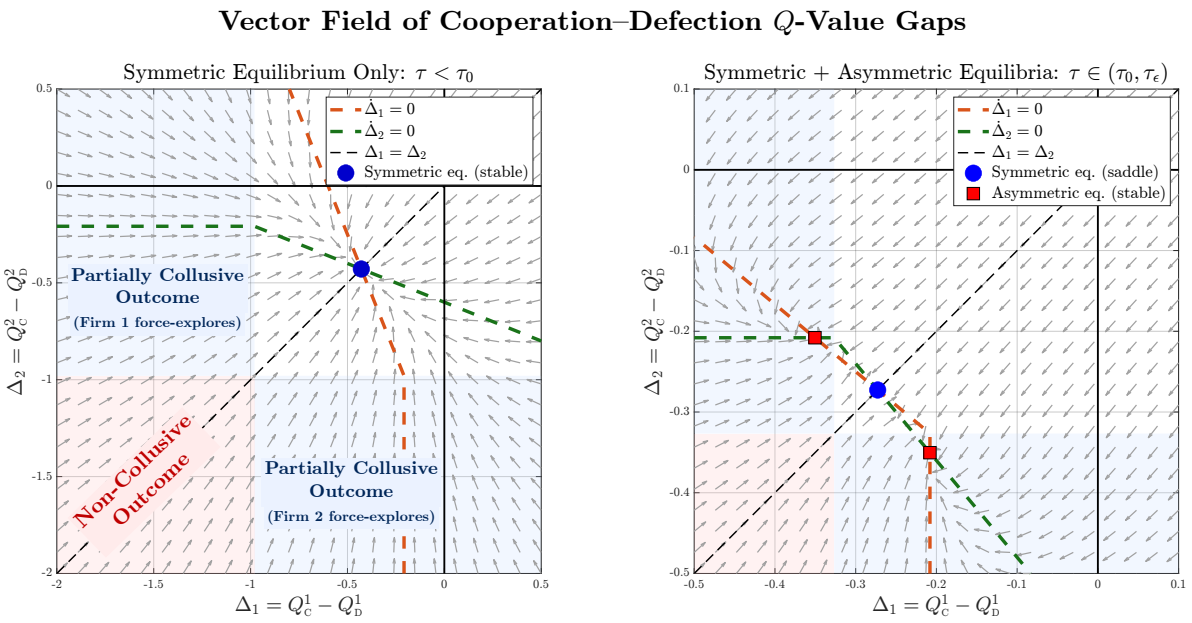



**Figure 9** **Each panel shows the vector field of the reduced $(\Delta_1, \Delta_2)$ system; shaded regions indicate forced exploration. Left panel ($\tau = 0.5 < \tau_0$): the system admits a unique symmetric equilibrium (blue circle, stable). Right panel ($\tau = 1.5 \in (\tau_0, \tau_\epsilon)$): the symmetric equilibrium becomes a saddle (blue circle) and two asymmetric stable equilibria emerge (red squares). Relative to the baseline, both panels set $\delta = 0$, $\theta = 0$, and $\alpha = 1$; the right panel additionally sets $\tau = 1.5$. See Appendix G.**

## Appendix E: Dynamical Systems and Hopf Bifurcation Preliminaries

This appendix collects the dynamical-systems concepts used in the paper. The goal is not to provide a self-contained treatment, but to fix terminology and notation for fixed points, stability, limit cycles, center manifolds, and Hopf bifurcations. The presentation follows standard references in nonlinear dynamics and bifurcation theory; see Strogatz (2018), Kuznetsov (2004), and Guckenheimer and Holmes (2002).

In the main text, for a fixed intervention level $\delta$, the learning dynamics can be written abstractly as

$$\dot{\mathbf{Q}} = \mathbb{F}(\mathbf{Q}; \delta), \qquad \mathbf{Q} = (Q_C^1, Q_D^1, Q_C^2, Q_D^2) \in \mathbb{R}^4, \tag{46}$$

where $\mathbb{F}$ is the right-hand side of (9) after the action probabilities $x$ and $y$ are obtained from (10). The intervention level $\delta$ is therefore a scalar parameter of the vector field. Here, a vector field means the map that assigns to each current state $\mathbf{Q}$ and fixed intervention level $\delta$ the instantaneous direction and speed of motion $\dot{\mathbf{Q}} = \mathbb{F}(\mathbf{Q}; \delta)$ of the Q-values. The symmetric interior equilibrium branch (defined in Appendix E.1 below) is denoted by $\mathbf{Q}^*(\delta)$, with induced cooperation probabilities $x^*(\delta) = y^*(\delta) = a^*(\delta)$.

Bifurcation analysis complements convergence arguments based on Lyapunov or potential functions. Lyapunov methods can certify stability, convergence to an equilibrium or invariant set, and in stronger cases convergence rates. Bifurcation analysis instead asks how the qualitative geometry of the flow changes as a parameter varies: when an equilibrium loses stability, when oscillations appear, and how local basin boundaries move near a critical threshold. In the main text, bifurcation analysis is applied to the frozen-$\delta$ autonomous system $\dot{\mathbf{Q}} = \mathbb{F}(\mathbf{Q}; \delta)$, treating $\delta$ as a fixed parameter. The later analysis of policy implementation then uses this family of frozen-$\delta$ systems to study whether the learning trajectory can track the moving equilibrium as the platform changes $\delta$ over time.

### E.1. Fixed Points and Local Stability

Consider first a general autonomous system

$$\dot{\mathbf{z}} = F(\mathbf{z}), \qquad \mathbf{z} \in \mathbb{R}^n, \tag{47}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. A point $\mathbf{z}^*$ is a *fixed point* or *equilibrium* if

$$F(\mathbf{z}^*) = 0.$$

If $\mathbf{z}(0) = \mathbf{z}^*$, then $\mathbf{z}(t) = \mathbf{z}^*$ for all $t \geq 0$. Throughout this appendix, a statement is called *local* if it holds after restricting attention to a sufficiently small neighborhood of the fixed point or other object under study. Thus, local stability describes the behavior of trajectories initialized sufficiently close to a fixed point; it does not characterize all initial conditions or the full global basin of attraction.

The Q-value vector field in the paper is globally only piecewise smooth because of the maximum operator in the continuation term and the capped policy map in (10). The nonsmooth surfaces are the surfaces $Q^i_{\text{C}} = Q^i_{\text{D}}$ and the surfaces $|Q^i_{\text{C}} - Q^i_{\text{D}}| = \Delta_{\epsilon,\tau}$. The local stability and Hopf results in the main text are applied only at symmetric interior fixed points lying away from these surfaces. At the fixed points studied there, $Q^{i*}_{\text{D}} > Q^{i*}_{\text{C}}$, so the maximum operator locally selects $Q^i_{\text{D}}$, and the policy map is in its linear region for each firm. Hence, in a sufficiently small neighborhood of the relevant fixed point, the vector field is smooth. The differentiable dynamical-systems results recalled below are used in this local sense.

For the parameterized system (46), an equilibrium at intervention level $\delta$ is a point $\mathbf{Q}^*(\delta)$ satisfying

$$\mathbb{F}(\mathbf{Q}^*(\delta); \delta) = 0.$$

As $\delta$ varies, a smoothly varying family of such equilibria is called an *equilibrium branch*.

**Definition 2 (Lyapunov stability)** *A fixed point $\mathbf{z}^*$ of (47) is* Lyapunov stable *if, for every neighborhood $U$ of $\mathbf{z}^*$, there exists a neighborhood $V$ of $\mathbf{z}^*$ such that*

$$\mathbf{z}(0) \in V \quad \Longrightarrow \quad \mathbf{z}(t) \in U \quad \text{for all } t \geq 0.$$

**Definition 3 (Local asymptotic stability and basin of attraction)** *A fixed point $\mathbf{z}^*$ is* locally asymptotically stable *if it is Lyapunov stable and there exists a neighborhood $U_0$ of $\mathbf{z}^*$ such that*

$$\mathbf{z}(0) \in U_0 \quad \Longrightarrow \quad \lim_{t \to \infty} \mathbf{z}(t) = \mathbf{z}^*.$$

*The* basin of attraction *of $\mathbf{z}^*$ is the set of all initial conditions whose trajectories converge to $\mathbf{z}^*$.*

Unless explicitly stated otherwise, when the main text informally refers to a fixed point as stable, it means locally asymptotically stable. A fixed point is *unstable* if it is not Lyapunov stable. A fixed point can be Lyapunov stable without being asymptotically stable; for example, a continuum of closed orbits around an equilibrium can keep trajectories nearby without making them converge.

*Linear stability analysis.* Let $\mathbf{J} = DF(\mathbf{z}^*)$ be the Jacobian matrix of $F$ evaluated at the fixed point $\mathbf{z}^*$. The fixed point is called *hyperbolic* if no eigenvalue of $\mathbf{J}$ has zero real part. It is *non-hyperbolic* if at least one eigenvalue has zero real part. The standard linearization test gives the following local conclusions:

- If $\text{Re}(\lambda_i) < 0$ for every eigenvalue $\lambda_i$ of $\mathbf{J}$, then $\mathbf{z}^*$ is locally asymptotically stable.
- If $\text{Re}(\lambda_i) > 0$ for at least one eigenvalue $\lambda_i$, then $\mathbf{z}^*$ is unstable.

- If no eigenvalue has positive real part but at least one eigenvalue has zero real part, then the fixed point is non-hyperbolic and the preceding sign test does not determine the nonlinear local dynamics. Higher-order terms, center-manifold analysis, or other arguments are needed.

For the learning dynamics, the relevant Jacobian along the equilibrium branch is

$$\mathbf{J}(\delta) = D_{\mathbf{Q}}\mathbb{F}(\mathbf{Q}^*(\delta);\delta).$$

The stability condition in Theorem 1 is obtained by analyzing the eigenvalues of this matrix.

### E.2. Limit Cycles and Basin Boundaries

The *phase space* of a dynamical system is the space of all possible states. For the system (47), the phase space is $\mathbb{R}^n$. A *periodic orbit* is a nonconstant trajectory $\mathbf{z}(t)$ for which there exists a period $T>0$ such that $\mathbf{z}(t+T)=\mathbf{z}(t)$ for all $t$. A *limit cycle* is an isolated periodic orbit, meaning that nearby periodic orbits do not form a continuum around it.

A limit cycle is *stable*, or attracting, if trajectories initialized sufficiently close to the cycle approach the cycle as $t\to\infty$. It is *unstable*, or repelling, if nearby trajectories move away from it in at least one direction. In two-dimensional dynamics, an unstable limit cycle surrounding a stable fixed point can form a local basin boundary: trajectories initialized inside the cycle converge to the fixed point, whereas trajectories initialized just outside the cycle are repelled away from the fixed point and may approach another attractor.

This planar picture is the intuition used in the Hopf analysis below. After the local reduction introduced in Appendix E.4, the relevant dynamics near the critical equilibrium are described by two critical coordinates. In those coordinates, an unstable periodic orbit can form the local boundary between perturbations that return to the stable fixed point and perturbations that move away from it. This is a local statement about the reduced dynamics; it does not by itself characterize the full global basin in the original phase space.

### E.3. Bifurcations

A *bifurcation* is a qualitative change in the dynamics of a system as a parameter varies. Such changes include the creation or destruction of fixed points, a change in the local stability of an existing fixed point, or the appearance or disappearance of periodic orbits. The parameter value at which the qualitative change occurs is called a *bifurcation point*. For the frozen-$\delta$ system

$$\dot{\mathbf{Q}} = \mathbb{F}(\mathbf{Q};\delta),$$

we study bifurcations along the symmetric interior equilibrium branch $\mathbf{Q}^*(\delta)$. A *branch* is a smoothly varying family of equilibria or periodic orbits indexed by the parameter. A local bifurcation of the equilibrium branch occurs at $\delta=\delta_c$ if the local phase portrait near the equilibrium changes qualitatively as $\delta$ passes through $\delta_c$. Equivalently, for each fixed $\delta$, write local perturbations as

$$\mathbf{q} = \mathbf{Q} - \mathbf{Q}^*(\delta).$$

Then the equilibrium is at $\mathbf{q}=0$, and a bifurcation occurs when the local dynamics of $\mathbf{q}$ near zero change qualitatively as $\delta$ crosses $\delta_c$. In this paper, $\delta_c$ denotes a critical intervention level at which the symmetric interior equilibrium branch loses local stability.

A *Hopf point* is a bifurcation point at which the Jacobian has a simple complex-conjugate pair of purely imaginary eigenvalues,

$$\lambda_{\pm}(\delta_c) = \pm i\omega_c, \qquad \omega_c>0,$$

while the remaining eigenvalues have nonzero real parts. In the stability setting studied in the main text, the remaining eigenvalues have negative real parts, and the real part of the critical pair crosses zero as $\delta$ passes through $\delta_c$. Thus, stability is lost through a two-dimensional perturbation direction in which , at the linear level, nearby trajectories spiral inward on the stable side, rotate without linear radial growth or decay at the Hopf point, and spiral outward on the unstable side. If the relevant nonlinear nondegeneracy condition also holds, this Hopf point gives rise to a Hopf bifurcation as discussed in Appendix E.5 below.

For a fixed point of a smooth finite-dimensional system, a local bifurcation can occur only when the fixed point is non-hyperbolic. Thus, in the stability analysis, the candidate bifurcation points are precisely the parameter values at which the Jacobian $D_{\mathbf{Q}}\mathbb{F}(\mathbf{Q}^*(\delta);\delta)$ has at least one eigenvalue with zero real part.

### E.4. Center-Manifold Reduction

The center manifold theorem is a local reduction tool for studying equilibria whose Jacobian has eigenvalues with zero real parts. These are precisely the cases in which linearization does not fully determine stability. In this subsection we state the result for a sufficiently smooth vector field; in the Hopf calculations below, the required derivatives are evaluated only in a neighborhood where the Q-value vector field is smooth.

Fix a parameter value $\delta_0$ and suppose $\mathbf{Q}^*(\delta_0)$ is an equilibrium of (46). After shifting coordinates so that the equilibrium is at the origin,

$$\mathbf{q} = \mathbf{Q} - \mathbf{Q}^*(\delta_0),$$

the dynamics can be written locally as

$$\dot{\mathbf{q}} = \mathbf{J}_0\mathbf{q} + G(\mathbf{q}), \qquad \mathbf{J}_0 = D_{\mathbf{Q}}\mathbb{F}(\mathbf{Q}^*(\delta_0);\delta_0), \qquad G(\mathbf{q}) = O(\|\mathbf{q}\|^2). \tag{48}$$

Let the eigenvalues of $\mathbf{J}_0$ be grouped according to the signs of their real parts: $n_-$ eigenvalues with negative real parts, $n_0$ eigenvalues with zero real parts, and $n_+$ eigenvalues with positive real parts. The directions associated with eigenvalues whose real parts are zero are called the *center directions*. These are the directions along which nonlinear terms determine the local behavior.

After a local linear change of coordinates that groups the linearized directions according to the real parts of their eigenvalues, write $\mathbf{q} = (\mathbf{u}, \mathbf{v})$, where $\mathbf{u} \in \mathbb{R}^{n_0}$ are the center variables and $\mathbf{v} \in \mathbb{R}^{n_- + n_+}$ are the non-center variables. In these coordinates, the system takes the form

$$\begin{aligned} \dot{\mathbf{u}} &= \mathbf{M}\mathbf{u} + g(\mathbf{u}, \mathbf{v}), \\ \dot{\mathbf{v}} &= \mathbf{N}\mathbf{v} + h(\mathbf{u}, \mathbf{v}), \end{aligned} \tag{49}$$

where $\mathbf{M}$ is the block of the linearization associated with eigenvalues whose real parts are zero, $\mathbf{N}$ is the block associated with the remaining eigenvalues, and $g$ and $h$ contain terms of at least quadratic order.

Informally, a *manifold* is a smooth geometric object that locally looks like a Euclidean space of some dimension. For example, a curve is a one-dimensional manifold and a surface is a two-dimensional manifold. In the present setting, the relevant manifolds are locally defined subsets of the state space near an equilibrium. A manifold is *locally invariant* for a dynamical system if any trajectory that starts on it remains on it for as long as the trajectory stays inside the neighborhood where the manifold is defined.

**Theorem 2 (Center manifold theorem and reduction principle (Kuznetsov 2004))** *There exists a locally defined center manifold $W^c(0)$ that passes through the equilibrium, is locally invariant, is tangent to the center eigenspace at the equilibrium, and can be written locally as*

$$W^c(0) = \{(\mathbf{u}, \mathbf{v}) : \mathbf{v} = V(\mathbf{u})\}, \qquad V(0) = 0, \quad DV(0) = 0.$$

*The dynamics restricted to the center manifold are*

$$\dot{\mathbf{u}} = \mathbf{M}\mathbf{u} + g(\mathbf{u}, V(\mathbf{u})). \tag{50}$$

*When all non-center eigenvalues have negative real parts, the local stability or instability of the full equilibrium is determined by the reduced dynamics (50). In particular, if the origin is locally asymptotically stable for the reduced center dynamics, then the full equilibrium is locally asymptotically stable; if the origin is unstable for the reduced center dynamics, then the full equilibrium is unstable.*

In the Hopf case, $n_0 = 2$: the center directions are spanned by the real and imaginary parts of the critical complex eigenvector. We refer to this two-dimensional reduced space as the *Hopf plane*. Equivalently, one may use a complex Hopf coordinate $z \in \mathbb{C}$ or polar coordinates

$$z = re^{i\varphi}, \qquad \mathbf{u} = (r\cos\varphi, r\sin\varphi).$$

The radius $r = |z|$ measures the local amplitude of the oscillatory deviation from the equilibrium, while $\varphi$ records the phase of the oscillation. For a fixed intervention level near a Hopf point, the amplitude $r$ is the radial distance of the local critical-coordinate representation from the equilibrium $\mathbf{Q}^*(\delta)$. In the implementation-speed analysis, the same Hopf-coordinate construction is applied pointwise along the frozen-$\delta$ equilibrium branch, so $r_t$ measures the instantaneous critical-coordinate distance of the learning state from $\mathbf{Q}^*(\delta_t)$.

### E.5. Hopf Bifurcation

A Hopf bifurcation is the local mechanism by which an equilibrium loses stability through a complex-conjugate pair of eigenvalues crossing the imaginary axis. Let $q \in \mathbb{C}^n$ be an eigenvector associated with one eigenvalue in the pair. The real two-dimensional subspace spanned by $\operatorname{Re} q$ and $\operatorname{Im} q$ is the critical subspace. At the linear level, perturbations in this subspace rotate around the equilibrium; as the real part of the eigenvalue pair changes sign, these rotations switch from decaying to growing. Consider the parameterized learning dynamics (46) and a smooth equilibrium branch $\mathbf{Q}^*(\delta)$. Along this branch, let

$$\mathbf{J}(\delta) = D_{\mathbf{Q}}\mathbb{F}(\mathbf{Q}^*(\delta); \delta)$$

denote the local Jacobian. At a Hopf threshold, the local linearization has this two-dimensional critical subspace. The Hopf theorem below states the standard smoothness, spectral, crossing, and nonlinear nondegeneracy conditions under which this linear loss of stability produces a local branch of periodic orbits.

The remaining issue is nonlinear: when periodic orbits bifurcate from the equilibrium branch, are they stable or unstable, and do they appear on the stable or unstable side of the threshold? This is determined by the cubic terms in the Hopf normal form after reducing the dynamics to the two-dimensional center manifold. The coefficient that determines the sign of this cubic effect is called the *first Lyapunov coefficient*, denoted by $\ell_1$. Its explicit formula involves second- and third-order derivatives of the vector field at the Hopf point. We do not need that formula here; Appendix B.3 gives the computation for the Q-value system. The nonlinear nondegeneracy condition is $\ell_1 \neq 0$. Under the convention used below, the sign of $\ell_1$ determines whether the Hopf bifurcation is supercritical or subcritical.

**Theorem 3 (Hopf bifurcation (Kuznetsov 2004, Guckenheimer and Holmes 2002))** *Consider the frozen-$\delta$ system $\dot{\mathbf{Q}} = \mathbb{F}(\mathbf{Q}; \delta)$, and suppose that the following conditions hold at $(\mathbf{Q}^*(\delta_c), \delta_c)$:*

1. *The vector field $(\mathbf{Q}, \delta) \mapsto \mathbb{F}(\mathbf{Q}; \delta)$ is $C^3$ in a neighborhood of $(\mathbf{Q}^*(\delta_c), \delta_c)$, and $\mathbf{Q}^*(\delta)$ is a $C^1$ equilibrium branch;*

2. *The Jacobian* $\mathbf{J}(\delta_c) = D_{\mathbf{Q}}\mathbb{F}(\mathbf{Q}^*(\delta_c);\delta_c)$ *has a simple complex-conjugate pair of purely imaginary eigenvalues*
$$\lambda_{\pm}(\delta_c) = \pm i\omega_c, \qquad \omega_c > 0;$$
3. *All remaining eigenvalues of* $\mathbf{J}(\delta_c)$ *have strictly negative real parts;*
4. *If* $\lambda_{\pm}(\delta) = a_H(\delta) \pm i\omega(\delta)$ *denotes the continuation of the critical eigenvalue pair near* $\delta_c$*, then* $a_H(\delta_c) = 0$ *and* $a'_H(\delta_c) \neq 0$*;*
5. *The nonlinear nondegeneracy condition holds: the first Lyapunov coefficient* $\ell_1$ *at* $(\mathbf{Q}^*(\delta_c), \delta_c)$ *is nonzero.*

*Then the system undergoes a Hopf bifurcation at* $\delta = \delta_c$*. A branch of periodic orbits bifurcates from the equilibrium branch. Under the convention used in Kuznetsov (2004), and describing the side of the bifurcation by the sign of* $a_H(\delta)$*, the sign of* $\ell_1$ *determines the criticality:*

$$\ell_1 < 0 \quad \Longrightarrow \quad \text{a stable small-amplitude periodic orbit occurs for } a_H(\delta) > 0 \quad \text{(supercritical)},$$
$$\ell_1 > 0 \quad \Longrightarrow \quad \text{an unstable small-amplitude periodic orbit occurs for } a_H(\delta) < 0 \quad \text{(subcritical)}.$$

On the two-dimensional center manifold, the Hopf normal form can be written in a complex coordinate $z$ as
$$\dot{z} = \big(a_H(\delta) + i\omega(\delta)\big)z + c_H z|z|^2 + \text{higher-order terms}, \tag{51}$$
where $c_H \in \mathbb{C}$. The real part of the cubic coefficient determines the leading-order amplitude dynamics. Writing
$$\ell_H := \mathrm{Re}(c_H),$$
and setting $r = |z|$, the radial dynamics take the form
$$\dot{r} = a_H(\delta)r + \ell_H r^3 + O\big(|\delta - \delta_c|r^3 + r^5\big). \tag{52}$$
Under the normalization used in Appendix B.3, $\ell_H$ has the same sign as the first Lyapunov coefficient $\ell_1$. In particular, with the convention used there, $\ell_H = \omega_c \ell_1$.

If $a'_H(\delta_c) > 0$, then for $\delta$ sufficiently close to $\delta_c$, the equilibrium is locally asymptotically stable on the side $\delta < \delta_c$ and unstable on the side $\delta > \delta_c$. When $\ell_H < 0$, a stable small-amplitude limit cycle appears on the unstable side $a_H(\delta) > 0$. When $\ell_H > 0$, an unstable small-amplitude limit cycle exists on the stable side $a_H(\delta) < 0$, with leading-order radius
$$r_u(\delta) = \sqrt{\frac{-a_H(\delta)}{\ell_H}} + o\Big(\sqrt{|a_H(\delta)|}\Big). \tag{53}$$
As $\delta \uparrow \delta_c$ along the stable side, this unstable cycle shrinks onto the equilibrium. In the reduced dynamics on the local center manifold, it separates nearby critical-coordinate perturbations that in the critical subspace that return to the equilibrium from nearby critical-coordinate perturbations that are repelled away from it. Thus the reduced local stability region in the critical directions shrinks as the Hopf threshold is approached.

*Supercritical and subcritical cases.* The two cases are illustrated by the following canonical normal forms, written in polar coordinates $(r, \varphi)$. We use $\mu$ in these examples as a generic scalar bifurcation parameter.

**Example 3 (Supercritical Hopf bifurcation)** *Consider*
$$\dot{r} = \mu r - r^3, \qquad \dot{\varphi} = \omega + br^2. \tag{54}$$
*When* $\mu < 0$*, the origin is locally asymptotically stable. When* $\mu > 0$*, the origin is unstable and is surrounded by a stable limit cycle with radius* $r = \sqrt{\mu}$*. The transition is continuous: the amplitude of the attracting cycle grows from zero as* $\mu$ *increases above the critical value* $\mu = 0$*.*


**Figure 10** **Canonical phase portraits of the supercritical Hopf bifurcation in (54): $\mu < 0$ on the left and $\mu > 0$ on the right.**

**Example 4 (Subcritical Hopf bifurcation)** *Consider*

$$\dot{r} = \mu r + r^3 - r^5, \qquad \dot{\varphi} = \omega + br^2. \tag{55}$$

*For $\mu < 0$, the origin is locally asymptotically stable and is surrounded by an unstable periodic orbit; the quintic term creates a stable large-amplitude cycle farther away. As $\mu \uparrow 0$, the unstable periodic orbit shrinks toward the origin. At $\mu = 0$, the origin loses stability. For $\mu > 0$, trajectories near the origin are repelled and may move toward the outer attracting cycle. This produces a discontinuous transition in amplitude and can generate hysteresis: once a trajectory moves to the large-amplitude oscillatory attractor, reducing $\mu$ back below zero need not return it to the fixed point unless it crosses back inside the basin bounded by the unstable cycle.*

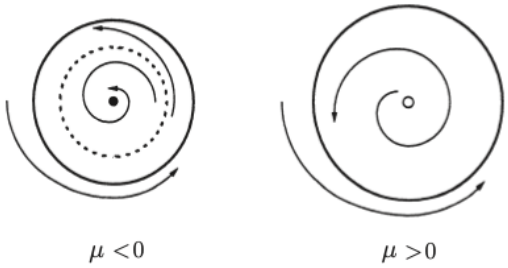



**Figure 11** **Canonical phase portraits of the subcritical Hopf bifurcation in (55): $\mu < 0$ on the left and $\mu > 0$ on the right. The dashed curve in the left panel is the unstable periodic orbit that shrinks to the origin at $\mu = 0$.**

*Interpretation for the intervention analysis.* In the paper, the intervention level $\delta$ plays the role of the bifurcation parameter. The critical threshold $\delta_c$ is the value at which the symmetric interior equilibrium branch $\mathbf{Q}^*(\delta)$ loses local stability. The computation in Appendix B.3 evaluates the first Lyapunov coefficient $\ell_1$ at this threshold. The numerical evidence reported there gives $\ell_1 > 0$, indicating a subcritical Hopf bifurcation under the convention above.

Consequently, for $\delta < \delta_c$ but close to $\delta_c$, the equilibrium can be locally asymptotically stable while, in the local Hopf reduction, an unstable periodic orbit surrounds it. This orbit gives a local basin boundary in the Hopf-reduced coordinate, and its leading-order radius shrinks according to (53). A sudden increase in $\delta$ can therefore move the current Q-values outside this reduced local basin even when the target frozen-$\delta$ equilibrium remains locally stable. The Hopf analysis by itself implies repulsion from the local equilibrium neighborhood; in the benchmark computations reported in the paper, the resulting trajectories converge to persistent oscillations. A gradual intervention can avoid this local loss of tracking by allowing the trajectory to remain near the moving equilibrium branch while the reduced basin boundary shrinks.

## Appendix F: Supplementary Main-Text Illustrations

The fixed-intervention bistability illustration and the constant-learning-rate benchmark example are included in the main text; see Figures 1 and 2 and Example 1.

## Appendix G: Numerical Parameters for the Computational Experiments

This appendix collects the numerical parameter choices used in the figures and computational exercises. Unless stated otherwise, experiments use the following baseline payoff and learning parameters:

$$R = 2.1, \qquad S = 1.8, \qquad T = 3.1, \qquad P = 2.0,$$

$$\alpha = 0.002, \qquad \gamma = 0.6, \qquad \tau = 0.5, \qquad \epsilon = 0.01, \qquad \theta = 1.$$

These payoff values imply

$$A = (R - T) - (S - P) = -0.8, \qquad B = S - P = -0.2,$$

so $A$ and $B$ are not listed as independent numerical parameters below. The intervention level $\delta$ is reported separately for each figure or experiment.

| Experiment or figure | Differences from the baseline parameters |
|---|---|
| Figure 1 | Uses $\delta = 0.242$. |
| Figure 2 and Example 1 | Uses $\delta = 0$, $\theta = 0$, and $\alpha = 1$. For these payoffs, $\tau_0 = -1/A = 1.25$ and $\tau_\epsilon(0) \approx 2.36$. |
| Example 2 | Uses $\delta = 0$ and $\alpha = 1$. The example compares $\theta = 1$, which is stable, with $\theta = 3$, which destabilizes the same symmetric fixed point. |
| Figure 3 | The left panel sets $\delta = 0.2 < \delta_c$, and the right panel sets $\delta = 0.3 > \delta_c$. For this parameterization, $\delta_c \simeq 0.284$. |
| Figure 4 | Uses $\delta = 0.265$. |
| Figures 5 and 6 | The target is $\delta^* = 0.242$. The sudden path moves $\delta : 0 \to 0.242$. The staged path moves $\delta : 0 \to 0.2 \to 0.242$. The reference trajectory holds $\delta = 0.242$ fixed and initializes the state near $\mathbf{Q}^*(0.242)$. |
| Figure 7 | The $\theta$-panel varies $\theta$ around the baseline while holding $\tau = 0.5$. The $\tau$-panel varies $\tau$ around the baseline while holding $\theta = 1$. |
| Figures 8a and 8b | Uses $\delta = 0.2 < \delta_c$ and $\delta = 0.3 > \delta_c$. |
| Figure 9 | Uses $\delta = 0$, $\theta = 0$, and $\alpha = 1$. The left panel uses $\tau = 0.5$; the right panel uses $\tau = 1.5$. For these payoffs, $\tau_0 = -1/A = 1.25$ and $\tau_\epsilon(0) \approx 2.36$. |