SAME PRICES, DIFFERENT RESPONSES: COUNTERFACTUAL RESPONSE SURFACES IN LEARNING MARKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning sellers can produce similar price paths for very different reasons. Some may act independently, while others may react strongly to a rival's action. Observed market outcomes alone may therefore hide important differences in strategic behavior. We study this problem by measuring how a seller responds when a rival's action is changed. For each unilateral intervention, we compare an intervention rollout with a matched control rollout that starts from the same market state and uses the same future shocks. Repeating this comparison across intervention settings gives a counterfactual response surface that describes the strength and duration of the response. We show that observed trajectories alone cannot identify these responses when different policy systems generate the same observed outcomes. Matched intervention and control rollouts make the responses identifiable and allow them to be estimated over a finite intervention grid. Experiments with scripted and learned pricing policies show that the response surface separates policy systems that look similar from observed prices alone. It also improves identification under the same evaluation budget, predicts responses to held-out interventions, and improves targeting when intervention capacity is limited. These results show that how a policy reacts to a controlled change can reveal information that its observed market outcomes do not.

1 INTRODUCTION

Learning-based sellers interact repeatedly with both the market and other sellers. A seller's price affects current demand, but it can also change how another seller acts later. This creates a basic identification problem. Similar price paths may come from independent behavior, common market shocks, or policies that react to a rival's action. These policy systems can look similar on the observed path even when their responses to the same intervention are very different. This paper asks whether observed market outcomes are enough to recover these dynamic responses. In general, they are not. Two policy systems can generate the same distribution of observed trajectories while responding differently after an off-path action. Outcome statistics such as prices, profits, sales, and price correlation therefore do not fully determine how a policy responds to a unilateral intervention. The missing information lies in behavior that is not reached on the observed path, as illustrated in Figure 1.



Figure 1: Similar observed price paths can hide different counterfactual responses. The two markets have almost the same PCI and price correlation, but Seller B responds differently to the same price cut by Seller A.

We study this information through a *counterfactual response surface*. For a unilateral intervention on one focal seller, an intervention rollout is paired with a control rollout from the same pre-intervention snapshot and the same future exogenous shocks. The difference between the two

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

paths measures how the opponent responds to the intervention. Repeating this comparison across intervention settings and response lags gives a response surface that records the strength and duration of the response. Frozen evaluation measures the response with fixed learner states, while Adaptive evaluation allows learning to continue after the intervention. The theory follows the same structure. First, we show that observed trajectories alone cannot identify the response surface when two policy systems have the same observational law. Second, we show that matched intervention and control rollouts identify the counterfactual response under the stated conditions. We then define Frozen and Adaptive evaluation regimes and give a uniform estimation bound over a finite intervention grid.

We evaluate the method on both scripted policy systems and learned sellers. The experiments show that policy systems with similar observed prices can have very different counterfactual responses. Under the same evaluation budget, the response surface identifies response-dependent behavior more accurately than a scalar deviation baseline. It also predicts responses to intervention settings that are not used during fitting and improves targeting when intervention capacity is limited. Together, these results show that off-path responses contain useful information that is hidden from observed market outcomes.

Our main contributions are:

- We formalize response-dependent behavior in learning markets as an identification problem and show that observed trajectories cannot identify counterfactual responses when policy systems share the same observational law but differ off the observed path.
- We introduce counterfactual response surfaces and estimate them with matched intervention and control rollouts, and define Frozen and Adaptive evaluation regimes that make explicit whether the measured response includes continued learning.
- We show that counterfactual response surfaces separate scripted policy systems with nearly identical observed prices, improve equal-budget mechanism identification over a scalar deviation baseline, and improve targeting under limited intervention capacity.

2 Preliminaries

2.1 Repeated Pricing Games

A repeated pricing game has N sellers and lasts for T periods, indexed by $t = 0, \dots, T-1$. At each period, seller i chooses a price $p_{i,t} \in \mathcal{P}_i$. The joint price profile is

$$p_t = (p_{1,t}, \dots, p_{N,t}). \tag{1}$$

Given the price profile, seller i receives one-period profit $r_i(p_t)$. Current actions may depend on information from earlier periods. Repeated price competition is a standard model of dynamic oligopoly (Tirole, 1988; Maskin & Tirole, 1988).

2.2 Policies in Reinforcement Learning

A policy maps the information available to an agent to a distribution over actions (Sutton & Barto, 2018). For seller i, the policy is denoted by

$$\pi_i(a \mid h_{i,t}), \tag{2}$$

where $h_{i,t}$ is the information available to seller i at period t. The policy may be deterministic or stochastic. Over T periods, the return of seller i is

$$Y_i^T = \sum_{t=0}^{T-1} r_{i,t}. \tag{3}$$

During training, a reinforcement learning algorithm updates its policy using observed states, actions, and rewards.

2.3 Nash and Joint-Profit Benchmarks

A Nash equilibrium is a price profile in which no seller can increase its profit by changing its own price while the prices of the other sellers remain fixed (Nash, 1951). A Nash price profile p^N therefore satisfies

$$r_i(p_i^N, p_{-i}^N) \geq r_i(p_i, p_{-i}^N), \qquad \forall p_i \in \mathcal{P}_i,\ \forall i. \tag{4}$$

The joint-profit benchmark is a price profile that maximizes total seller profit (Tirole, 1988). It is defined by

$$p^M \in \arg\max_p \sum_{i=1}^{N} r_i(p). \tag{5}$$

In a symmetric market, p^N and p^M also denote the common seller prices under the Nash and joint-profit benchmarks.

3 Problem Formulation

3.1 Repeated Learning Market

The analysis uses the repeated market introduced in Section 2. There are N sellers over periods $t = 0, \ldots, T-1$. At period t, seller i chooses an action $a_{i,t} \in \mathcal{A}_i$ according to its policy $\pi_i(\cdot \mid h_{i,t})$, where $h_{i,t}$ is the information available to the seller. In the pricing setting, $a_{i,t} = p_{i,t}$. The market outcome depends on the joint action and an exogenous shock z_t.

Two evaluation regimes are considered for learned sellers. Under *Frozen* evaluation, policy parameters and learner states remain fixed during the rollout. Under *Adaptive* evaluation, each learner follows its original update rule after each transition. The two regimes are denoted by $e \in \{F, A\}$.

3.2 Unilateral Intervention

A unilateral intervention changes the action of one focal seller at a chosen time. An intervention is written as

$$I = (i, \tau, \delta, d), \tag{6}$$

where i is the focal seller, τ is the intervention time, δ is the action change, and d is the intervention duration. The intervention is observed by the opponent in the next period. For any $j \neq i$, seller j is an opponent whose response can be measured. Let H denote the post-intervention response horizon. The control case is denoted by $I = 0$. Let S_τ denote the pre-intervention snapshot, which contains the market history and all learner states at time τ. Intervention and control rollouts start from the same snapshot and share the same future shock path $z_{\tau:\tau+H-1}$.

3.3 Counterfactual Response

The main object is the change in the opponent's action caused by the intervention. Let $s > 0$ be an action scale. For the pricing experiments,

$$s = p^M - p^N, \tag{7}$$

where p^N and p^M are the Nash and joint-profit benchmark prices in Eqs. (4) and (5). For lag $k \in \{1, \ldots, H\}$, the normalized paired response of opponent j is

$$\Delta_j^e(k; I, \mathsf{S}_\tau, z_{\tau:\tau+H-1}) = \frac{a_{j,\tau+k}^{I,e} - a_{j,\tau+k}^{0,e}}{s}. \tag{8}$$

A negative value means that the opponent takes a lower action under the intervention than under the control, which in the pricing setting means a lower price.

The population response is the difference between the opponent's expected actions under the intervention and under the control. For a snapshot σ, define

$$R_{j\leftarrow i}^e(k; I, \sigma) = \frac{\mathbb{E}\big[a_{j,\tau+k}^{I,e} \mid \mathsf{S}_\tau = \sigma\big] - \mathbb{E}\big[a_{j,\tau+k}^{0,e} \mid \mathsf{S}_\tau = \sigma\big]}{s}, \tag{9}$$

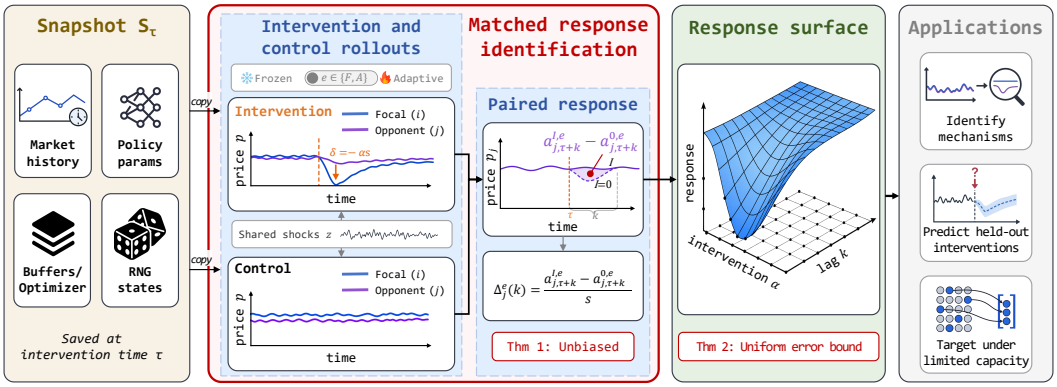


Figure 2: Overview of the counterfactual response methodology. From a snapshot S_τ, an intervention rollout and a control rollout share the same future shocks, and only the focal seller's action is changed. Their paired difference gives the normalized response $\Delta^e_j(k)$, and repeating it over a grid of interventions and lags gives the response surface.

where each expectation is taken over future shocks and policy randomness in a separate run of the market from σ. Averaging over the evaluation distribution of snapshots gives

$$R^e_{j\leftarrow i}(k; I) = \mathbb{E}\big[R^e_{j\leftarrow i}(k; I, \mathsf{S}_\tau)\big]. \tag{10}$$

For a fixed intervention I, the sequence $\{R^e_{j\leftarrow i}(k; I)\}_{k=1}^H$ is its dynamic response path. Let $\mathcal{I}$ be a set of unilateral interventions. The counterfactual response surface is

$$\mathcal{R}_{j\leftarrow i} = \big\{R^e_{j\leftarrow i}(k; I) : I \in \mathcal{I},\ k \in \{1, \dots, H\},\ e \in \{F, A\}\big\}. \tag{11}$$

It records how the opponent responds across intervention settings, response lags, and evaluation regimes. Section 4 shows how matched rollouts identify and estimate this surface.

4 Counterfactual Response Methodology

4.1 Why Intervention Is Needed

Observational equivalence. Observed trajectories show how policies behave on the realized path. They do not determine how the same policies respond after an off-path action. Two policy systems can therefore have the same observed trajectory distribution but different counterfactual response surfaces.

Let O denote the trajectory observed without intervention. Let $M \in \{0, 1\}$ index two policy systems. Let P^O_m denote the distribution of O under system m, and let $\mathcal{R}_m$ denote its counterfactual response surface.

Proposition 1 (Outcome-only impossibility)**.** *There exist two finite-state policy systems such that*

$$P^O_0 = P^O_1,$$

while

$$\mathcal{R}_0 \neq \mathcal{R}_1$$

for at least one unilateral intervention.

Under equal priors, any possibly randomized decision rule $\phi(O)$ based only on the observed trajectory satisfies

$$\Pr(\phi(O) \neq M) = \frac{1}{2}.$$

Hence,

$$\inf_\phi \Pr(\phi(O) \neq M) = \frac{1}{2}.$$

Proof. The proof is given in Appendix A.1. □

Intervention reveals the difference. Proposition 1 shows that observed trajectories alone cannot identify the counterfactual response surface. The difference becomes visible when an intervention moves the market away from the observed path. The next subsection shows how matched intervention and control rollouts identify this response.

4.2 Matched Response Identification

Matched rollout design. For each sampled pre-intervention snapshot, an intervention rollout is paired with a control rollout, as shown in Figure 2. Both rollouts start from separate copies of the snapshot and use the same future exogenous shocks. The intervention is applied only to the focal seller in the intervention rollout, and the two copies never share learner states. Each evaluation cluster c corresponds to one independent policy pair and training seed, and contains m_c matched pairs indexed by r. Let $\mathsf{S}_{\tau,c,r}$ be the snapshot of pair r, and let $\Delta^e_{j,c,r}(k;I)$ be its normalized paired response from Eq. (8). The cluster-level response averages over matched pairs, and the response estimator averages over n independent clusters,

$$\Delta^e_{j,c}(k;I) = \frac{1}{m_c}\sum_{r=1}^{m_c}\Delta^e_{j,c,r}(k;I), \qquad \widehat{R}^e_{j\leftarrow i}(k;I) = \frac{1}{n}\sum_{c=1}^{n}\Delta^e_{j,c}(k;I). \tag{12}$$

Response identification. Sharing the snapshot and the shock path makes the two rollouts of a matched pair dependent. The next result shows that this dependence leaves the target unchanged, so the paired responses can be averaged directly.

Theorem 1 (Response-surface identification)**.** *Fix an evaluation regime $e \in \{F, A\}$, an intervention I, and a lag k. Assume that (i) the two rollouts of each matched pair start from the same snapshot and evolve with separate learner states, and (ii) given the snapshot, the shared shock path follows the same distribution as in a single run and does not depend on the intervention. Then, for every matched pair (c, r) and every snapshot σ,*

$$\mathbb{E}\big[\Delta^e_{j,c,r}(k;I) \mid \mathsf{S}_{\tau,c,r} = \sigma\big] = R^e_{j\leftarrow i}(k;I,\sigma).$$

If the snapshots are drawn from the evaluation distribution, then $\mathbb{E}\big[\widehat{R}^e_{j\leftarrow i}(k;I)\big] = R^e_{j\leftarrow i}(k;I)$. If, in addition, the n clusters are independent and identically distributed and $\Delta^e_{j,c}(k;I)$ is integrable, then $\widehat{R}^e_{j\leftarrow i}(k;I) \xrightarrow{p} R^e_{j\leftarrow i}(k;I)$ as $n \to \infty$.

Proof. The proof is given in Appendix A.2. □

Each rollout still follows its own interventional law, so the paired difference recovers the causal contrast in Eq. (9). The dependence also helps estimation, because the shared shocks cancel noise that the two rollouts have in common. For any paired outcome X,

$$\mathrm{Var}(X^I - X^0) = \mathrm{Var}(X^I) + \mathrm{Var}(X^0) - 2\,\mathrm{Cov}(X^I, X^0),$$

so the paired difference has lower variance whenever the two rollouts are positively correlated. This design extends common random numbers in simulation (Glasserman, 2003) to learning agents, where the snapshot must also copy every learner state.

4.3 Evaluation Regimes and Finite-Grid Estimation

Frozen and Adaptive evaluation. Frozen evaluation fixes both learner states, so $R^F_{j\leftarrow i}$ measures the response encoded in the learned policies. Adaptive evaluation lets both sellers keep updating, so $R^A_{j\leftarrow i}$ also includes learning after the intervention. For scripted policies, which do not update, the two regimes coincide. For learned policies, however, Frozen and Adaptive evaluations capture different response mechanisms because the policy itself may update after intervention.

Remark 1 (Intermediate regimes). The matched design also supports fixing the focal seller and the opponent separately. This gives four regimes $R^{FF}, R^{FA}, R^{AF}, R^{AA}$, with $R^F = R^{FF}$ and $R^A = R^{AA}$. The two intermediate regimes attribute $R^A - R^F$ to opponent updates, focal updates, and their interaction. Our experiments report the two endpoints.

Finite-grid estimation. The response surface is evaluated on a finite intervention grid

$$\mathcal{G} = \{I_1, \ldots, I_Q\}. \tag{13}$$

For each intervention $I \in \mathcal{G}$ and lag $k \in \{1, \ldots, H\}$, Eq. (12) gives the estimated response. Matched rollouts within the same evaluation cluster are averaged before responses are combined across clusters. Thus, n denotes the number of independent evaluation clusters.

Uniform error bound. The finite grid allows the estimation error to be controlled over all intervention and lag cells at the same time.

Theorem 2 (Uniform finite-grid error)**.** *Fix an evaluation regime $e \in \{F, A\}$. Assume that the n evaluation clusters are independent and identically distributed and that*

$$\Delta^e_{j,c}(k; I) \in [-B, B]$$

for every $I \in \mathcal{G}$ and $k \in \{1, \ldots, H\}$. Then, with probability at least $1 - \beta$,

$$\max_{\substack{I \in \mathcal{G} \\ 1 \le k \le H}} \left| \widehat{R}^e_{j \leftarrow i}(k; I) - R^e_{j \leftarrow i}(k; I) \right| \le B \sqrt{\frac{2 \log(2QH/\beta)}{n}}. \tag{14}$$

Proof. The proof is given in Appendix A.3. □

The bound applies separately to Frozen and Adaptive evaluation. As the number of independent clusters increases, the estimated response surface converges uniformly over the finite intervention grid. The number of intervention and lag cells enters the bound only through the logarithmic term.

5 Experiments

5.1 Experimental Setup

Environment, policies, and interventions. We study repeated pricing in a two-seller differentiated Logit market. The scripted policy systems include myopic best response, independent behavior, common-shock response, reciprocal punishment, noisy reciprocity, and opponent shaping. We also evaluate learned sellers trained with Soft Actor-Critic (SAC) (Haarnoja et al., 2018), Twin Delayed Deep Deterministic Policy Gradient (TD3) (Fujimoto et al., 2018), and Proximal Policy Optimization (PPO) (Schulman et al., 2017). Training hyperparameters and implementation details are reported in Appendix C.1. For a downward price intervention, the focal seller's action changes by

$$\delta = -\alpha s, \qquad \alpha \in \{0.10, 0.25, 0.50\},$$

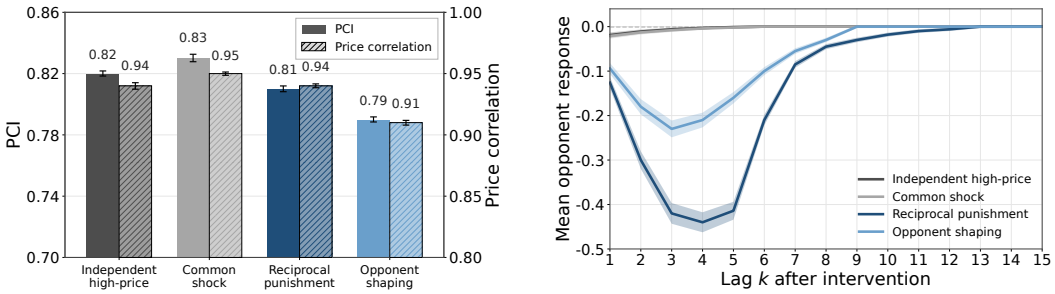
where $s = p^M - p^N$ is the price scale defined in Section 3.3. The intervention lasts for $d \in \{1, 5\}$ periods, and the response horizon is $H \in \{20, 50, 100\}$. Each intervention rollout is paired with a control rollout from the same pre-intervention snapshot and with the same future exogenous shocks. The core experiments use 20 training seeds and 500 matched rollout pairs per intervention cell.

Baselines and evaluation. We use the price coordination index (PCI) and price correlation to summarize observed behavior. Let $\bar{p}$ be the mean observed market price. The PCI is

$$\mathrm{PCI} = \frac{\bar{p} - p^N}{p^M - p^N},$$

where larger values indicate prices closer to the joint-profit benchmark. We compare the counterfactual response surface with outcome-only features, a temporal outcome model, a static deviation test, and price deviation over 16 periods (PD-16). PD-16 is a scalar Frozen response summary based on one pre-intervention price and 15 post-intervention prices:

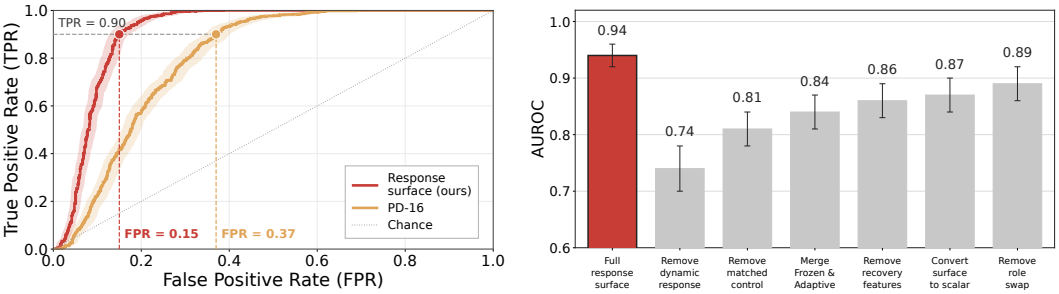
$$\mathrm{PD}_{16} = \frac{1}{15} \sum_{k=1}^{15} \frac{p^0_j - p_{j,\tau+k}}{p^0_j - p^N}.$$



(a) Similar on-path outcomes.

(b) Different counterfactual responses.

Figure 3: **Similar observed outcomes can hide different counterfactual responses.** (a) Four scripted policy systems have similar price coordination index (PCI) and price correlation before intervention. (b) Under the same unilateral intervention and Adaptive evaluation, independent high-price and common-shock policies remain close to zero response, while reciprocal punishment and opponent shaping show clear dynamic responses. Shaded regions denote 95% training-seed cluster-bootstrap confidence intervals.



(a) Equal-budget identification.

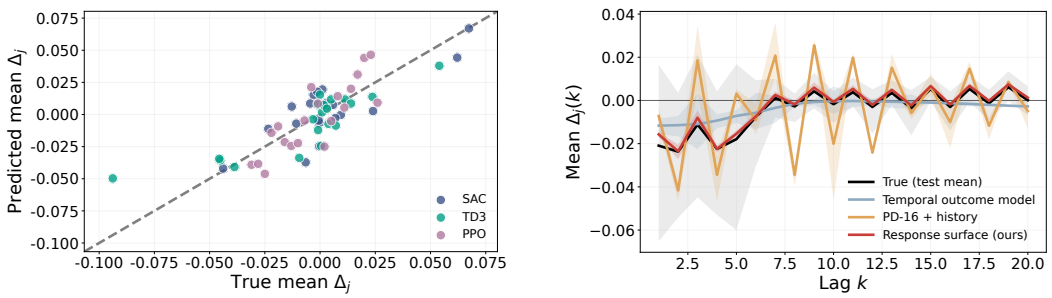
(b) Full-grid ablation.

Figure 4: **Equal-budget identification and component ablation.** (a) Receiver operating characteristic (ROC) curves for the price deviation over 16 periods (PD-16) and the counterfactual response surface using the same 15,000 post-intervention environment transitions per policy-pair seed. (b) Ablation on the full intervention grid. Removing dynamic response information or the matched control causes the largest drops in area under the ROC curve (AUROC).

A larger value indicates a stronger or longer price reduction by the opponent. For the equal-budget comparison, PD-16 and the counterfactual response surface use the same 10-period starting history, the same one-period downward intervention with $\alpha = 0.25$, the same 15-period response horizon, and the same 15,000 post-intervention environment transitions per policy-pair seed. Further details are given in Appendix C.5. Identification performance is measured by the area under the receiver operating characteristic curve (AUROC). We also report the false-positive rate (FPR) at a true-positive rate (TPR) of 90%. Uncertainty is reported at the training-seed level using 95% cluster-bootstrap confidence intervals.

5.2 Main Results

Same observed outcomes, different responses. We first compare four scripted policy systems with similar observed behavior. Their PCI values range from 0.79 to 0.83, and their price correlations range from 0.91 to 0.95. Despite this similarity, their responses to the same unilateral intervention differ sharply. Under Adaptive evaluation, the one-step responses are -0.019 for independent high-price behavior and -0.021 for common-shock response. They are much larger in magnitude for reciprocal punishment and opponent shaping, reaching -0.126 and -0.094, respectively. Figure 3 shows that this difference continues over later lags. Independent high-price and common-shock policies remain close to zero, while reciprocal punishment and opponent shaping



(a) Held-out response prediction.

(b) Held-out response path.

Figure 5: **Prediction of held-out interventions.** (a) Predicted and realized mean responses for held-out intervention settings. The dashed diagonal denotes perfect prediction. (b) Mean response path for the held-out intervention $\alpha = 0.50$, $d = 5$, and $H = 100$. The counterfactual response surface follows the realized response more closely than the temporal outcome model and PD-16 with history. The displayed window shows the first 20 post-intervention lags. Shaded regions denote 95% training-seed cluster-bootstrap confidence intervals.

produce clear dynamic responses. Thus, similar observed prices can hide different counterfactual response patterns.

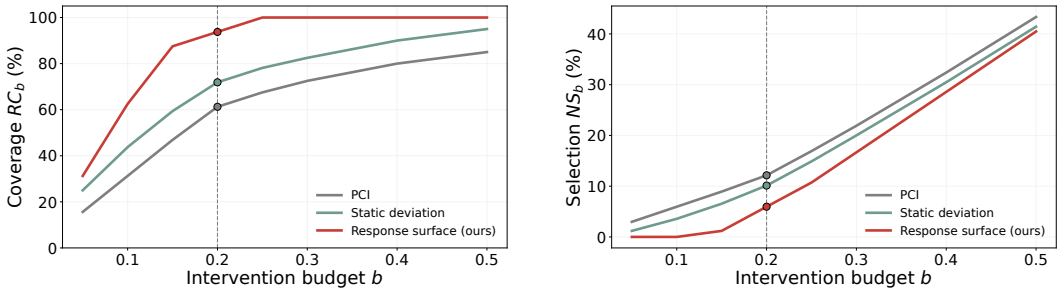
Equal-budget comparison. Under the same evaluation budget, the counterfactual response surface outperforms PD-16 in mechanism identification. AUROC increases from 0.80 to 0.91, while FPR at TPR$=90\%$ decreases from 0.37 to 0.15. Figure 4a shows the ROC curves under the same 15,000 post-intervention environment transitions per policy-pair seed. The full-grid ablation in Figure 4b shows which parts of the response surface drive this gain. Removing dynamic response information reduces AUROC from 0.94 to 0.74, and removing the matched control reduces it to 0.81. These are the largest drops among the ablations, showing that dynamic response and matched controls are the main sources of identification performance.

Prediction of held-out interventions. We next test whether the response surface of learned SAC, TD3, and PPO sellers predicts responses to intervention settings that are not used during fitting. The prediction model is trained on $\alpha \in \{0.10, 0.25\}$, $d = 1$, and $H \in \{20, 50\}$, and is tested without refitting on $\alpha = 0.50$, $d = 5$, and $H = 100$. Training and test sets are also separated by training seed, policy instance, and intervention cell. The counterfactual response surface achieves a path root mean squared error (RMSE) of 0.031, a test R^2 of 0.82, and 92% response-sign accuracy. For comparison, PD-16 with history has a path RMSE of 0.061, while the temporal outcome model reaches 0.088. Figure 5 shows that the predicted response paths remain close to the realized responses under the held-out intervention. These results show that the response surface captures response structure that transfers to intervention settings outside the fitting set.

Decision value. We finally evaluate whether the response surface improves targeting when intervention capacity is limited. The held-out set contains 1,000 scripted deployment units, including 160 response-dependent units and 840 non-response controls. At the main budget $b = 0.20$, every ranking rule selects $K_b = 200$ units. Response-surface ranking covers 93.8% of response-dependent units (150/160), while selecting only 6.0% of non-response controls (50/840). Static deviation test reaches 71.9% coverage with a 10.1% non-response selection rate, while PCI reaches 61.3% coverage with a 12.1% non-response selection rate. Figure 6 shows the same pattern across intervention budgets, with the largest gap in the low-capacity regime.

6 Related Work

Learning sellers and observed market outcomes. A large literature shows that reinforcement-learning sellers can sustain prices above the competitive level (Waltman & Kaymak, 2008; Calvano et al., 2020; Klein, 2021; Hettich, 2021; Asker et al., 2022; Banchio & Mantegazza, 2022; Meylahn & V. den Boer, 2022; den Boer et al., 2026), and empirical studies document related patterns in real



(a) Response-dependent coverage.

(b) Non-response selection rate.

Figure 6: **Decision value under limited intervention capacity.** Curves are computed on a fixed held-out set of 1,000 scripted deployment units. At intervention budget b, each ranking rule selects $K_b = \lfloor b|\mathcal{U}| \rfloor$ units. (a) Response-dependent coverage RC_b, the fraction of response-dependent units selected. (b) Non-response selection rate NS_b, the fraction of non-response controls selected. The vertical dashed line marks the main budget $b = 0.20$.

markets (Musolff, 2022; Brown & MacKay, 2023; Assad et al., 2024). Related work studies how platform design and regulation affect sellers that use pricing algorithms (Johnson et al., 2023; Hartline et al., 2024; 2025). A central difficulty is that high prices alone do not reveal their cause: similar outcomes can arise from independent learning, misspecified demand models, or simultaneous experimentation rather than from responses to rivals (Hansen et al., 2021; Abada & Lambin, 2023; Calvano et al., 2023; Lambin, 2024; Cho & Williams, 2024). We address this non-identifiability by measuring how policies respond when the market is moved away from the observed path.

Deviation tests and counterfactual influence. Economic theory characterizes coordinated behavior through responses to unilateral deviations (Friedman, 1971; Hendon et al., 1996). Calvano et al. (2020) study forced deviations in trained agents, and Harrington (2018) proposes testing pricing algorithms for reward–punishment properties, whose interpretation is examined by Epivent & Lambin (2024). In multi-agent reinforcement learning, counterfactual changes to one agent's action are used to measure its influence on others (Jaques et al., 2019), and opponent-shaping methods learn to affect rivals' updates (Foerster et al., 2018). Structural approaches instead inspect policy graphs (Zhou, 2026) or internal representations (Rose et al., 2026), which require access to policy internals. Our response surface is related to impulse-response analysis (Jordà, 2005), but uses matched simulator rollouts to identify responses over an intervention grid and distinguishes Frozen from Adaptive evaluation. It needs no access to policy internals, only the ability to snapshot and replay the market.

7 Conclusion

Observed market outcomes can hide important differences in how learning sellers respond to a rival's action. We formalize this as an identification problem and show that policy systems with the same observational law can still have different counterfactual responses. We therefore introduce counterfactual response surfaces based on matched intervention and control rollouts. The two rollouts start from the same snapshot and use the same future shocks, and their paired difference measures the response to the intervention. Frozen and Adaptive evaluations distinguish responses from fixed policies from those that also include continued learning. Our theory shows that this matched design identifies the response and supports uniform estimation over a finite intervention grid.

Experiments with scripted policies and learned SAC, TD3, and PPO sellers show that response surfaces separate systems with similar observed outcomes, improve equal-budget mechanism identification over PD-16, predict held-out responses, and improve targeting under limited intervention capacity. Together, these results show that counterfactual responses capture strategic information that observed prices and price correlations alone do not reveal. Evaluating learning markets therefore requires more than observing the prices sellers reach. It also requires measuring how policies react when the market is moved away from the observed path.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, including language editing and improving the clarity and organization of the method and experiment sections. All AI-assisted suggestions were reviewed and revised by the authors. The authors verified the final manuscript and are responsible for its content.

REFERENCES

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

John Asker, Chaim Fershtman, and Ariel Pakes. Artificial intelligence, algorithm design, and pricing. In *AEA Papers and Proceedings*, volume 112, pp. 452–456, 2022.

Stephanie Assad, Robert Clark, Daniel Ershov, and Lei Xu. Algorithmic pricing and competition: Empirical evidence from the german retail gasoline market. *Journal of Political Economy*, 132 (3):723–771, 2024.

Martino Banchio and Giacomo Mantegazza. Artificial intelligence and spontaneous collusion. *arXiv preprint arXiv:2202.05946*, 2022.

Zach Y Brown and Alexander MacKay. Competition in pricing algorithms. *American Economic Journal: Microeconomics*, 15(2):109–156, 2023.

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

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Algorithmic collusion: Genuine or spurious? *International Journal of Industrial Organization*, 90:102973, 2023.

In-Koo Cho and Noah Williams. Collusive outcomes without collusion: Algorithmic pricing in a duopoly model. *Available at SSRN 4753617*, 2024.

Arnoud den Boer, Janusz M Meylahn, and Maarten Pieter Schinkel. Artificial collusion: Examining supracompetitive pricing by Q-learning algorithms. *Management Science*, 2026.

Andréa Epivent and Xavier Lambin. On algorithmic collusion and reward–punishment schemes. *Economics Letters*, 237:111661, 2024.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In Elisabeth André, Sven Koenig, Mehdi Dastani, and Gita Sukthankar (eds.), *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, pp. 122–130. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018.

James W Friedman. A non-cooperative equilibrium for supergames. *The Review of Economic Studies*, 38(1):1–12, 1971.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1582–1591. PMLR, 2018.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2003.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1856–1865. PMLR, 2018.

Karsten Hansen, Kanishka Misra, and Mallesh M. Pai. Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Science*, 40(1):1–12, 2021.

Joseph E Harrington. Developing competition law for collusion by autonomous artificial agents. *Journal of Competition Law & Economics*, 14(3):331–363, 2018.

Jason D. Hartline, Sheng Long, and Chenhao Zhang. Regulation of algorithmic collusion. In *Proceedings of the Symposium on Computer Science and Law, CSLAW 2024, Boston, MA, USA, March 12-13, 2024*, pp. 98–108. ACM, 2024.

Jason D. Hartline, Chang Wang, and Chenhao Zhang. Regulation of algorithmic collusion, refined: Testing pessimistic calibrated regret. In *Proceedings of the 2025 Symposium on Computer Science and Law, CSLAW 2025, Munich, Germany, March 25-27, 2025*, pp. 108–120. ACM, 2025.

Ebbe Hendon, Hans Jørgen Jacobsen, and Birgitte Sloth. The one-shot-deviation principle for sequential rationality. *Games and Economic Behavior*, 12(2):274–282, 1996.

Matthias Hettich. Algorithmic collusion: Insights from deep learning, 2021.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049. PMLR, 2019.

Justin P Johnson, Andrew Rhodes, and Matthijs Wildenbeest. Platform design when sellers use pricing algorithms. *Econometrica*, 91(5):1841–1879, 2023.

Òscar Jordà. Estimation and inference of impulse responses by local projections. *American Economic Review*, 95(1):161–182, 2005.

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

Xavier Lambin. Less than meets the eye: simultaneous experiments as a source of algorithmic seeming collusion, 2024.

Eric Maskin and Jean Tirole. A theory of dynamic oligopoly, ii: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica: Journal of the Econometric Society*, pp. 571–599, 1988.

Janusz M Meylahn and Arnoud V. den Boer. Learning to collude in a pricing duopoly. *Manufacturing & Service Operations Management*, 24(5):2577–2594, 2022.

Leon Musolff. Algorithmic pricing facilitates tacit collusion: Evidence from e-commerce. In David M. Pennock, Ilya Segal, and Sven Seuken (eds.), *EC '22: The 23rd ACM Conference on Economics and Computation, Boulder, CO, USA, July 11 - 15, 2022*, pp. 32–33. ACM, 2022.

John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

Aaron Rose, Carissa Cullen, Sahar Abdelnabi, Philip Torr, Brandon Gary Kaplowitz, and Christian Schroeder de Witt. Detecting multi-agent collusion through multi-agent interpretability. *arXiv preprint arXiv:2604.01151*, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction, 2nd Edition*. MIT Press, 2018.

Jean Tirole. *The theory of industrial organization*. MIT Press, Cambridge, MA, 1988.

Ludo Waltman and Uzay Kaymak. Q-learning agents in a cournot oligopoly model. *Journal of Economic Dynamics and Control*, 32(10):3275–3293, 2008.

Chao Zhou. Beyond price traces: Policy-graph audits for algorithmic collusion, 2026. Preprint.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Proof. Consider two finite-state policy systems, denoted by 0 and 1. Both systems start from the same state q_0. On the observed path, the focal seller takes action $a^\star$, and the opponent uses the same action rule in both systems. The transition and reward rules are also the same on this path.

The two systems differ only after the off-path action

$$a^\star - \delta, \qquad \delta > 0,$$

is taken by the focal seller. In system 0, this action does not change the opponent's rule. In system 1, it moves the system to a state q_1, where the opponent changes its action for at least one period before returning to q_0.

Because the off-path action is never taken under normal play, the two systems have the same distribution over observed trajectories:

$$P_0^O = P_1^O.$$

Under the unilateral intervention, however, the opponent responds differently in the two systems. Thus,

$$\mathcal{R}_0 \neq \mathcal{R}_1.$$

Now let $\phi(O)$ be any possibly randomized outcome-only rule. Let $q(O) \in [0,1]$ be the probability that the rule predicts system 1. Under equal priors, its error is

$$\frac{1}{2}\mathbb{E}_{P_0^O}[q(O)] + \frac{1}{2}\mathbb{E}_{P_1^O}[1 - q(O)].$$

Since $P_0^O = P_1^O$, both expectations are taken under the same distribution. Therefore,

$$\Pr(\phi(O) \neq M) = \frac{1}{2}.$$

Taking the infimum over all ϕ gives

$$\inf_{\phi} \Pr(\phi(O) \neq M) = \frac{1}{2}.$$

□

A.2 PROOF OF THEOREM 1

Proof. Fix e, I, k, and a matched pair (c, r) with snapshot σ. By assumption (i), both rollouts start from the same state and neither rollout can affect the other. By assumption (ii), the shared shock path has the same conditional law given σ as the shock path of a single run. Policy randomness is drawn separately in each rollout. Hence the intervention rollout has the same law as a separate intervention run from σ, and the control rollout has the same law as a separate control run from σ. Linearity of expectation then gives

$$\mathbb{E}\big[\Delta^e_{j,c,r}(k;I) \mid \mathsf{S}_{\tau,c,r} = \sigma\big] = \frac{\mathbb{E}\big[a^{I,e}_{j,\tau+k} \mid \mathsf{S}_\tau = \sigma\big] - \mathbb{E}\big[a^{0,e}_{j,\tau+k} \mid \mathsf{S}_\tau = \sigma\big]}{s} = R^e_{j\leftarrow i}(k;I,\sigma).$$

If the snapshots are drawn from the evaluation distribution, the law of total expectation gives $\mathbb{E}\big[\Delta^e_{j,c,r}(k;I)\big] = \mathbb{E}\big[R^e_{j\leftarrow i}(k;I,\mathsf{S}_\tau)\big] = R^e_{j\leftarrow i}(k;I)$. Averaging over r and then over c keeps this expectation, so $\widehat{R}^e_{j\leftarrow i}(k;I)$ is unbiased. If the clusters are independent and identically distributed and $\Delta^e_{j,c}(k;I)$ is integrable, the weak law of large numbers gives $\widehat{R}^e_{j\leftarrow i}(k;I) \xrightarrow{p} R^e_{j\leftarrow i}(k;I)$. □

A.3 PROOF OF THEOREM 2

Proof. Fix an evaluation regime $e \in \{F, A\}$, an intervention $I \in \mathcal{G}$, and a lag $k \in \{1, \ldots, H\}$. By assumption, the n cluster-level responses $\Delta^e_{j,c}(k;I)$ are independent and identically distributed and satisfy

$$\Delta^e_{j,c}(k;I) \in [-B, B].$$

Their mean is $R^e_{j\leftarrow i}(k;I)$, and Eq. (12) gives

$$\widehat{R}^e_{j\leftarrow i}(k;I) = \frac{1}{n}\sum_{c=1}^{n}\Delta^e_{j,c}(k;I).$$

For any $t>0$, Hoeffding's inequality gives

$$\Pr\left(\left|\widehat{R}^e_{j\leftarrow i}(k;I) - R^e_{j\leftarrow i}(k;I)\right| > t\right) \le 2\exp\left(-\frac{nt^2}{2B^2}\right).$$

There are QH intervention and lag cells. Applying a union bound gives

$$\Pr\left(\max_{\substack{I\in\mathcal{G}\\1\le k\le H}}\left|\widehat{R}^e_{j\leftarrow i}(k;I) - R^e_{j\leftarrow i}(k;I)\right| > t\right) \le 2QH\exp\left(-\frac{nt^2}{2B^2}\right).$$

Setting the right-hand side equal to β gives

$$t = B\sqrt{\frac{2\log(2QH/\beta)}{n}}.$$

Therefore, with probability at least $1-\beta$,

$$\max_{\substack{I\in\mathcal{G}\\1\le k\le H}}\left|\widehat{R}^e_{j\leftarrow i}(k;I) - R^e_{j\leftarrow i}(k;I)\right| \le B\sqrt{\frac{2\log(2QH/\beta)}{n}}.$$

□

B Additional Experiments

This section reports additional evidence on generalization, sample efficiency, and intervention design.

B.1 Cross-Market Generalization

We keep the response-surface specification fixed and evaluate it under different market settings. Table 1 reports the results.

The response surface remains effective across changes in substitution, costs, quality, and market size. AUROC is 0.94 in the baseline market and remains between 0.83 and 0.92 across the alternative settings. The false-positive rate at TPR$=90\%$ ranges from 0.10 to 0.24. Performance remains strong under cost and quality asymmetry, with AUROC values of 0.90 and 0.89, respectively. It also reaches 0.86 with three sellers and 0.88 under held-out demand and cost settings.

Table 1: Cross-market identification with a fixed response-surface specification.

Market setting	AUROC	FPR at TPR$=90\%$
Smooth Logit baseline	0.94	0.10
High substitution	0.92	0.12
Cost asymmetry	0.90	0.15
Quality asymmetry	0.89	0.16
Three sellers	0.86	0.20
Hard tie-breaking	0.83	0.24
Held-out demand and cost	0.88	0.17

B.2 Response-Surface Sample Efficiency

We vary the number of matched rollout pairs used in each intervention cell. Table 2 shows that performance improves steadily as more matched rollouts are added.

With only 25 matched pairs per cell, AUROC is 0.81. It rises to 0.91 with 100 pairs and reaches 0.94 with 500 pairs. At the same time, the 95% confidence-interval width falls from 0.19 to 0.05, while Kendall's τ rises from 0.68 to 0.97. These results show that additional matched rollouts improve both identification accuracy and ranking stability.

Table 2: Performance as the number of matched rollout pairs increases.

Matched pairs per cell	AUROC	95% CI width	Kendall's τ
25	0.81	0.19	0.68
50	0.87	0.14	0.80
100	0.91	0.10	0.89
200	0.93	0.07	0.94
500	0.94	0.05	0.97

B.3 Intervention Size, Duration, and Horizon

Table 3 reports how the response surface changes with intervention magnitude, duration, and response horizon. Identification stays strong in every setting, with AUROC between 0.89 and 0.95.

Larger interventions produce stronger and longer responses. As α increases from 0.10 to 0.50, $D^A_{1:5}$ rises from 0.18 to 0.49, R^A_1 moves from -0.071 to -0.181, and the recovery time grows from 5 to 8 periods. Stronger interventions also give slightly higher AUROC, because larger responses are easier to separate from zero.

A longer intervention deepens the response without changing its first step. When the duration increases from $d = 1$ to $d = 5$, R^A_1 stays the same because the opponent sees the same first-period change, while $D^A_{1:5}$ rises from 0.34 to 0.42 and the recovery time grows from 7 to 8 periods.

The response horizon does not change the estimated response. For $(\alpha, d) = (0.25, 1)$, recovery occurs after 7 periods, well before the shortest horizon $H = 20$. All three horizons therefore give the same R^A_1, $D^A_{1:5}$, and recovery time, and AUROC stays between 0.93 and 0.94. A horizon of 20 periods is enough to capture the full response in this market.

C Experimental Protocol and Evaluation Metrics

This section gives the training settings, matched-rollout protocol, evaluation regimes, and metrics used in the experiments.

C.1 Training and Implementation Details

We train learned sellers with SAC, TD3, and PPO. Each algorithm uses 20 training seeds and 2×10^6 joint market steps. All learned sellers use the same network size, action range, learning rate, discount factor, and environment-interaction budget. Algorithm-specific settings follow the training rule of each method. Table 4 reports the full training settings. All experiments were run on two Intel Xeon Gold 6530 CPUs and eight NVIDIA RTX 5090 GPUs

Reproducibility statement. We provide the training configurations and evaluation protocols used in all experiments. The full hyperparameter settings are summarized in Table 4. All experiments follow the same matched-rollout procedure and use fixed random seeds for evaluation. The implementation will be released after the review period.

Table 3: Sensitivity to intervention magnitude, duration, and response horizon.

Variation	Setting	$D^A_{1:5}$	R^A_1	T^A_{rec}	AUROC
Magnitude	$\alpha = 0.10,\ d = 1,\ H = 50$	0.18	-0.071	5	0.89
Magnitude	$\alpha = 0.25,\ d = 1,\ H = 50$	0.34	-0.126	7	0.94
Magnitude	$\alpha = 0.50,\ d = 1,\ H = 50$	0.49	-0.181	8	0.95
Duration	$\alpha = 0.25,\ d = 1,\ H = 50$	0.34	-0.126	7	0.94
Duration	$\alpha = 0.25,\ d = 5,\ H = 50$	0.42	-0.126	8	0.95
Horizon	$\alpha = 0.25,\ d = 1,\ H = 20$	0.34	-0.126	7	0.93
Horizon	$\alpha = 0.25,\ d = 1,\ H = 50$	0.34	-0.126	7	0.94
Horizon	$\alpha = 0.25,\ d = 1,\ H = 100$	0.34	-0.126	7	0.94

Table 4: Training hyperparameters for learned sellers.

Hyperparameter	SAC	TD3	PPO
Shared settings			
Hidden layers	$[128, 128]$	$[128, 128]$	$[128, 128]$
Action range	$[-1, 1]$	$[-1, 1]$	$[-1, 1]$
Learning rate	10^{-4}	10^{-4}	10^{-4}
Discount factor γ	0.99	0.99	0.99
Batch / minibatch size	256	256	256
Reward scale	0.2	0.2	0.2
Training steps	2×10^6	2×10^6	2×10^6
Training seeds	20	20	20
Algorithm-specific settings			
Activation	ReLU	ReLU	tanh
Replay buffer size	10^5	10^5	–
Warm-up steps	10^4	10^4	–
Target update coefficient τ	0.005	0.005	–
Policy delay	–	2	–
Exploration noise std.	–	0.1	–
Target policy noise std.	–	0.2	–
Target noise clip	–	0.5	–
Temperature learning rate	3×10^{-4}	–	–
Target entropy	-1	–	–
Rollout length	–	–	256
PPO clip ratio	–	–	0.2
GAE λ	–	–	0.95
Optimization epochs	–	–	10
Entropy coefficient	–	–	0.0
KL coefficient	–	–	0.2

C.2 Matched Rollout and State Snapshot

For each intervention time τ, we save the complete pre-intervention state. The snapshot contains the market state and all learner states needed to continue the rollout, including policy parameters, optimizer states, random-number states, buffers, and update counters.

The intervention and control rollouts start from separate copies of the same snapshot and use the same future exogenous shocks. The control rollout follows the original policy. In the intervention rollout, only the focal seller's declared action is changed during the intervention period. After the two rollouts begin, their learner states evolve separately and are never shared.

Policy-action random draws are sampled separately in the two rollouts. At the end of each matched pair, both rollout copies are discarded. The next evaluation again starts from the saved pre-intervention snapshot.

If an episode ends before the response horizon H, the terminal state is kept for the remaining response lags. The environment is not restarted within the response horizon.

C.3 Frozen and Adaptive Update Timing

We use two evaluation regimes for learned sellers.

Frozen evaluation. The policy and learner state remain fixed throughout the rollout. Executed transitions are not added to replay or rollout buffers, and no policy, critic, target-network, optimizer, or normalization update is performed. After the intervention period, the focal seller returns to its original policy.

Adaptive evaluation. The intervention and control rollouts start from separate copies of the same learner state. Each rollout stores its own executed transitions and follows the original update schedule of the learning algorithm. SAC and TD3 update from their private replay buffers, while PPO updates only when its rollout buffer reaches the normal update boundary. The intervention does not introduce any extra learner update. These two regimes are the endpoints R^{FF} and R^{AA} in Remark 1

C.4 Response Summary Metrics

The response summaries are computed from the estimated Adaptive response path $\widehat{R}^{A}_{j\leftarrow i}(k; I)$.

For compact reporting, the one-step response is

$$R^{A}_{1}(I) = \widehat{R}^{A}_{j\leftarrow i}(1; I).$$

For a downward price intervention, the five-period response depth is

$$D^{A}_{1:5}(I) = -\frac{1}{5}\sum_{k=1}^{5} \widehat{R}^{A}_{j\leftarrow i}(k; I). \tag{15}$$

A larger value means a stronger average price reduction by the opponent during the first five response lags.

We measure recovery on the same estimated Adaptive response path. Let $\eta = 0.05$. The recovery time is

$$T^{A}_{\mathrm{rec}}(I) = \min\left\{k \in \{1, \ldots, H-2\} : \left|\widehat{R}^{A}_{j\leftarrow i}(k+r; I)\right| \le \eta \text{ for } r = 0, 1, 2\right\}. \tag{16}$$

Recovery therefore requires the estimated response to stay within 0.05 of zero for three consecutive periods. If recovery is not observed within the response horizon, the recovery time is right-censored at H.

C.5 PD-16 Baseline

PD-16 is a scalar baseline inspired by the forced-deviation experiment in Calvano et al. (2020). It uses one pre-intervention opponent price and 15 post-intervention prices under Frozen evaluation.

Let p^{0}_{j} denote the opponent's price before the intervention and let p^{N} denote the Nash benchmark price. We define

$$\mathrm{PD}_{16} = \frac{1}{15}\sum_{k=1}^{15} \frac{p^{0}_{j} - p_{j,\tau+k}}{p^{0}_{j} - p^{N}}. \tag{17}$$

A larger value indicates a stronger or longer price reduction by the opponent after the focal seller's intervention.

For the equal-budget comparison, PD-16 and the counterfactual response surface use the same 10-period starting history, the same one-period downward intervention with $\alpha = 0.25$, the same 15-period response horizon, and the same 15,000 post-intervention environment transitions per policy-pair seed.

C.6 Decision-Value Metrics

A deployment unit $u \in \mathcal{U}$ is an ordered seller–policy-pair–market–seed unit. Each ranking method assigns a score to every unit.

For an intervention budget $b \in [0, 1]$, the platform selects

$$K_b = \lfloor b|\mathcal{U}| \rfloor \tag{18}$$

units with the highest scores.

The decision-value experiment uses $1{,}000$ held-out scripted units. Among them, 160 are response-dependent and 840 are non-response controls. Reciprocal punishment, noisy reciprocity, and opponent shaping are labeled response-dependent. Myopic best response, stationary independent behavior, and common-shock response are labeled non-response controls.

Let $\mathcal{R}$ and $\mathcal{N}$ denote the response-dependent and non-response sets, and let $\mathcal{I}_b$ denote the selected units. Response-dependent coverage is

$$RC_b = \frac{|\mathcal{I}_b \cap \mathcal{R}|}{|\mathcal{R}|}, \tag{19}$$

and the non-response selection rate is

$$NS_b = \frac{|\mathcal{I}_b \cap \mathcal{N}|}{|\mathcal{N}|}. \tag{20}$$

Higher RC_b means that more response-dependent units are reached, while lower NS_b means that fewer non-response controls are selected. At the main budget $b = 0.20$, each ranking method selects $K_b = 200$ units.

D Limitations

Our evaluation focuses on controlled market environments that enable precise analysis of strategic interactions and counterfactual responses. While these environments allow us to isolate the effects of different behavioral mechanisms, real-world markets may involve additional factors such as heterogeneous sellers, changing demand patterns, and richer platform constraints. Extending the framework to larger-scale and more diverse market settings is an important direction for future work.