Multi-Agent Reinforcement Learning in Cournot Competition: The Strategic Value of Information Asymmetry

authored by
Scheubel Eelai (eelai.scheubel@etu.unistra.fr)
Istrati Laurentiu (laurentiu.istrati@etu.unistra.fr)
Marcu Vasile-Stefan (vasile.marcu@etu.unistra.fr)
Rustamli Sayqin (sayqin.rustamli@etu.unistra.fr)

January 20, 2026

Abstract

This report presents a study of oligopolistic market dynamics through multi-agent reinforcement learning (MARL). We train neural network agents to compete in a single-shot Cournot duopoly under three information regimes: perfect information (full observability), partial information (unknown competitor costs), and asymmetric information (minimal observability). Our central finding challenges conventional rationality assumptions: perfectly informed agents who expect rational competitor behavior become vulnerable to boundedly rational agents that have learned robust strategies under uncertainty. In one-shot competition, strategies developed under uncertainty outperform those relying on rationality assumptions, even when the latter possess superior information. Using policy gradient methods with regret minimization over 100,000 training episodes, we demonstrate that perfect information agents converge to Nash equilibrium, while information-constrained agents learn profitable deviations. However, symmetric ignorance leads to industry-wide overproduction and 57% profit losses, illustrating coordination failure.

Contents

1 Introduction

This report presents a study of oligopolistic market dynamics through the lens of multi-agent reinforcement learning (MARL). We develop and train neural network agents to compete in a single-shot Cournot oligopoly market, where firms simultaneously choose production quantities without repeated interaction. Our primary research question investigates how different levels of market information affect strategic behavior and profitability in competitive environments, and which strategies firms learn to adopt under those uncertainties.

We implement three distinct information regimes: perfect information (agents know demand parameters and competitor costs), partial information (agents know demand but not competitor costs), and asymmetric information (agents only know their own costs). Each regime corresponds to different real-world market conditions, from transparent regulated markets to opaque competitive environments. Critically, our experimental design evaluates these learned strategies in one-shot games, where firms cannot condition behavior on past interactions or build reputation.

The central finding of our research challenges conventional assumptions about rationality in strategic settings: when perfectly informed agents assume competitors are equally rational and informed, they become vulnerable to exploitation by boundedly rational agents who have learned robust strategies under uncertainty. In single-shot competition, agents that develop strategies resilient to uncertainty outperform those that rely on rationality assumptions, even when the latter possess superior information. However, when both agents lack information, industry-wide overproduction destroys value for all participants.

Our implementation uses policy gradient reinforcement learning with regret minimization. For perfect information agents, we employ a residual architecture combining linear Nash equilibrium baselines with neural network corrections to stabilize learning. For partial and asymmetric information agents, we use standard neural network policies. Through 100,000 training episodes per experiment, agents learn to optimize production decisions under uncertainty. The results validate theoretical predictions under perfect information while revealing new insights about strategic behavior under information asymmetry in single-shot competition.

2 Basic Model

2.1 Cournot Competition Framework

The Cournot model describes an oligopolistic market where n firms compete by simultaneously choosing production quantities. The market price is determined by the aggregate supply through an inverse demand function. In our implementation, we consider a duopoly (two-firm) market with linear demand.

The inverse demand function is specified as:

$$P = a - b \cdot Q \tag{1}$$

where P represents market price, $Q = \sum_{i=1}^{n} q_i$ is total quantity produced by all firms, a is the

demand intercept (maximum willingness to pay), and b is the demand slope (price sensitivity to quantity).

Each firm i chooses quantity q_i to maximize its profit function:

$$\pi_i = (P - c_i) \cdot q_i = (a - b \cdot \sum_j q_j - c_i) \cdot q_i \tag{2}$$

where c_i represents firm i's marginal cost of production and the summation runs over all firms in the market.

2.2 Nash Equilibrium Solution

The Nash equilibrium in Cournot competition is found by solving each firm's profit maximization problem given competitors' equilibrium quantities. For firm i, the first-order condition is:

$$\frac{\partial \pi_i}{\partial q_i} = a - b \cdot \sum_j q_j - b \cdot q_i - c_i = 0 \tag{3}$$

In the symmetric two-firm case where both firms have identical costs c, the Nash equilibrium quantities are:

$$q^* = \frac{a - c}{3b} \tag{4}$$

This yields equilibrium price $P^* = \frac{a+2c}{3}$ and per-firm profit $\pi^* = \frac{(a-c)^2}{9b}$.

The Nash equilibrium serves as our theoretical benchmark. Rational agents with perfect information should converge to this solution, as no firm can unilaterally improve its profit by deviating from q^*.

2.3 Best Response Functions

A critical concept in Cournot analysis is the best response function, which specifies firm i's optimal quantity given competitor j's quantity choice:

$$BR_i(q_j) = \frac{a - c_i - b \cdot q_j}{2b} \tag{5}$$

The Nash equilibrium occurs at the intersection of best response functions, where each firm's quantity is a best response to the other's. This reaction function framework provides the foundation for our regret-based learning signal, as agents learn to minimize the difference between chosen actions and best responses.

3 Extension

3.1 Information Asymmetry

Our primary extension to the basic Cournot model introduces heterogeneous information structures across agents. While traditional game-theoretic analysis assumes common knowledge of

all parameters, real markets exhibit varying degrees of transparency. We model three distinct information regimes (see Figure 1):

Figure 1: Three information conditions: Perfect (4D input), Partial (3D input), and Asymmetric (1D input). Each regime represents different levels of market observability.

Perfect Information (4D input): Agents observe the complete state vector $s = [a, b, c_i, c_j]$, including demand parameters and both firms' costs. This represents the textbook Cournot setting with complete information.

Partial Information (3D input): Agents observe $s = [a, b, c_i]$, knowing demand parameters and their own cost but not the competitor's cost. This reflects markets where demand is publicly observable (through market research or regulatory disclosure) but competitor cost structures remain private.

Asymmetric Information (1D input): Agents observe only $s = [c_i]$, their own marginal cost. Both demand parameters and competitor information are unknown. This extreme scenario represents highly opaque markets where firms must learn entirely from experience.

3.2 Applicability of Solution Concepts

The classical Nash equilibrium solution concept remains applicable in principle across all information regimes, though its practical computation differs significantly.

Under perfect information, the standard Nash equilibrium $q^* = \frac{a-c}{3b}$ applies directly. Agents should converge to this quantity through rational best-response dynamics or reinforcement learning.

Under partial information where firm i knows (a, b, c_i) but not c_j, computing a Bayesian Nash equilibrium requires specifying beliefs over the competitor's cost distribution. If firm i believes $c_j \sim F(c)$, the equilibrium involves solving:

$$q_i^* = \arg\max \mathbb{E}[\pi_i(q_i, BR_j(q_i; c_j)) \mid c_j \sim F] \tag{6}$$

This expectation integrates over the belief distribution, yielding a modified best response that accounts for uncertainty about the competitor.

Under asymmetric information with only c_i known, a rational equilibrium requires even more complex belief structures. Agents must form beliefs about both demand parameters (a, b) and competitor costs. In principle, a Bayesian Nash equilibrium exists where each agent's strategy is optimal given beliefs, and beliefs are updated via Bayes' rule from observed market

outcomes. However, analytical computation becomes intractable, motivating our reinforcement learning approach where agents learn equilibria through experience.

3.3 Deviation from Rational Expectations

Our extension also examines bounded rationality through neural network policies that may not conform to rational expectations. Unlike classical analysis where agents correctly anticipate competitor behavior, our learned policies emerge from self-play training. This can lead to equilibria that differ from theoretical predictions, particularly when information is incomplete. The strategic value of commitment to non-rational strategies emerges naturally from this framework.

4 Implementation

4.1 Model Structure

Our implementation consists of four primary modules organized as Python classes and utility functions (see Figure 2).

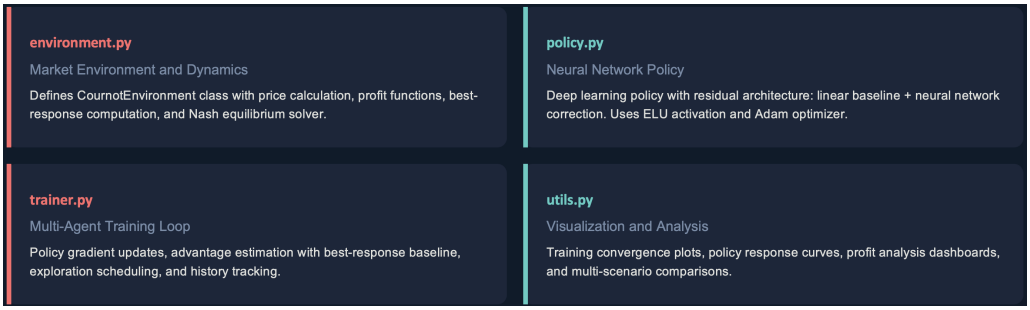


Figure 2: Project structure showing the four main components: environment (market dynamics), policy (neural networks), trainer (learning loop), and utils (visualization).

CournotEnvironment Class (environment.py) encapsulates the market dynamics and game logic. Public methods include:

- `sample_scenario(n_players)`: Randomly samples demand parameters (a, b) and firm costs from configured ranges
- `price(quantities, a, b)`: Computes market clearing price given total production
- `profit(quantities, costs, a, b)`: Calculates profit for each firm
- `best_response(q_others, cost, a, b)`: Computes optimal quantity given competitors' quantities
- `nash_equilibrium(a, b, costs)`: Solves for Nash equilibrium quantities analytically

The environment configuration specifies parameter ranges: $a \in [80, 120]$, $b \in [0.8, 1.2]$, costs $\in [5, 50]$, and maximum feasible quantity $q_{\max} = 100$.

PolicyNetwork Classes (policy.py) implement the neural network decision-making agents. Two architectures are used:

- `PolicyNetwork_Nash`: For perfect information (4D input)
- `PolicyNetwork_asymmetrical`: For partial and asymmetric information (3D or 1D input)

Both implement a residual architecture with public methods:

- `forward(state)`: Takes state observation and returns chosen quantity
- `backward(state, action, advantage)`: Performs gradient descent using advantage signal
- `get_linear_baseline(state)`: Computes Nash equilibrium baseline

CournotTrainer Class (trainer.py) orchestrates the multi-agent learning process with methods for training loops, advantage computation, and opponent synchronization in self-play mode.

4.2 Neural Network Architecture

Our neural network policy uses a residual formulation combining a linear baseline with a learned correction:

$$q = \text{Linear}(s) + \text{NN}(s) \tag{7}$$

The linear component computes the Nash equilibrium quantity directly from state features when full information is available. The neural network component learns strategic adjustments.

Network specifications:

- **Input layer:** Dimension varies by information regime (1D, 3D, or 4D)
- **Hidden layers:** Two fully connected layers with 64 units each
- **Activation:** ELU (Exponential Linear Unit) for smooth gradients
- **Output layer:** Single unit producing quantity adjustment
- **Final activation:** Clipping to $[0, q_{\text{max}}]$ to ensure feasible quantities

The residual architecture provides several advantages. The linear baseline initializes agents near rational behavior, accelerating early learning. The additive structure ensures stable gradients even when the neural network component is poorly trained initially.

4.3 Reinforcement Learning Algorithm

We employ a policy gradient algorithm with regret minimization as the learning signal. The training process operates as follows:

Episode Execution:

1. Sample new market scenario (a, b, costs) from configured ranges
2. Each agent i observes state s_i according to its information structure
3. Agents simultaneously choose quantities $q_i = \text{policy}_i(s_i)$ with exploration noise

4. Environment computes market price and resulting profits

Advantage Calculation: The advantage signal represents regret-how much better the agent could have done with hindsight:

$$\text{Advantage}(s_i, q_i) = \text{normalize}(BR_i(q_{-i}) - \pi_i(q_i)) \tag{8}$$

where $BR_i(q_{-i})$ is the profit from best-responding to observed competitor quantities, and $\pi_i(q_i)$ is the actual profit received.

Policy Update: Each agent performs gradient ascent on expected regret:

$$\theta \leftarrow \theta + \alpha \cdot \nabla_\theta \log \pi(q_i|s_i;\theta) \cdot \text{Advantage}(s_i, q_i) \tag{9}$$

Exploration Strategy: To ensure adequate exploration, we add Gaussian noise: $q_{\text{explored}} = q_{\text{policy}} + \mathcal{N}(0, \sigma^2)$

The exploration variance decays over training: $\sigma_t = \max(\sigma_{\text{final}}, \sigma_{\text{init}} \cdot \text{decay}^t)$

4.4 Training Process and Parameters

Self-Play Configuration: For perfect information agents, we use self-play training where the opponent policy is periodically synchronized with the learning agent every 1000 episodes. For partial and asymmetric information, both agents train simultaneously without synchronization.

Learning Rate Scheduling: We implement linear warmup followed by decay:

- Warmup: Episodes 0–10,000 (perfect) or 0–15,000 (partial/asymmetric)
- Decay: Remaining episodes with linear decrease to final learning rate
- Values: $\alpha_{\text{init}} = 0.003$–$0.005$, $\alpha_{\text{final}} = 0.0001$

Training Duration: All experiments run for 100,000 episodes.

Table 1: Hyperparameter configuration for different information regimes

Parameter	Perfect Info	Partial Info	Asymmetric Info
Input dim	4	3	1
Hidden dim	64	64	64
Episodes	100,000	100,000	100,000
Initial LR	0.003	0.005	0.005
Final LR	0.0001	0.0001	0.0001
Warmup	10,000	15,000	15,000
σ initial	12.0	15.0	15.0
σ final	3.0	3.0	3.0
σ decay	0.99996	0.99998	0.99998

Higher initial exploration and warmup periods for partial and asymmetric information compensate for increased learning difficulty.

4.5 Evaluation Methodology

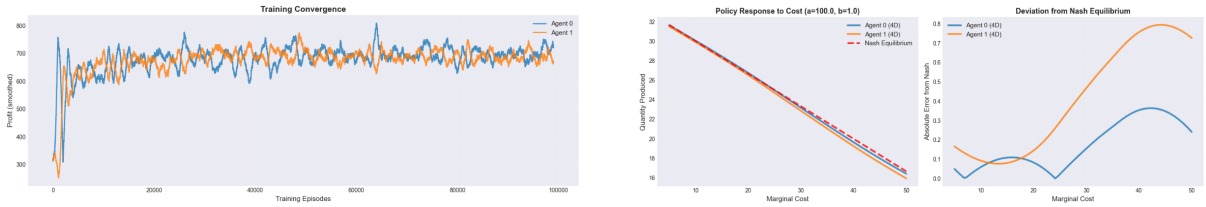
After training, we evaluate agents in head-to-head matchups on a fixed test environment with $a = 90$, $b = 1.1$, and costs $= 10$ for both firms. This yields theoretical Nash equilibrium values $q^* = 24.24$, $P^* = 36.67$, and $\pi^* = 646.46$ per firm.

Each matchup runs 5,000 episodes to obtain stable profit and quantity statistics. We test six configurations: three self-play scenarios and three cross-matchups.

5 Analysis

5.1 Perfect Information: Nash Convergence

Under perfect information conditions where both agents observe the complete state $[a, b, c_1, c_2]$, our reinforcement learning agents successfully converge to the theoretical Nash equilibrium. The trained agents achieve average profit of 650.75 with quantities of 24.08, compared to the theoretical Nash values of 646.46 profit and 24.24 quantity.



(a) Training convergence

(b) Policy response curve

Figure 3: Perfect information training results showing (a) profit and regret convergence over 100,000 episodes, and (b) learned policy closely tracking Nash equilibrium across cost levels.

The close alignment (within 0.7% for both profit and quantity) validates our implementation and demonstrates that policy gradient methods with regret minimization can discover game-theoretic equilibria through self-play. The training convergence plots (Figure 3a) show rapid initial learning followed by stable performance, with regret EMA declining from over 90 to approximately 12 by episode 100,000.

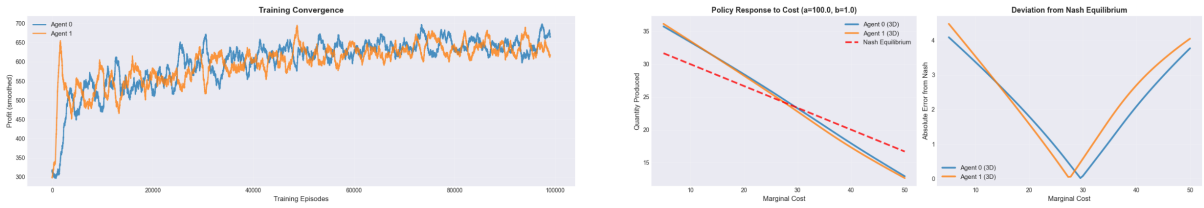
The policy response curves (Figure 3b) reveal that learned strategies closely track the Nash quantity across all marginal cost levels from 5 to 50. This robustness indicates the agents have learned the underlying strategic structure rather than memorizing specific scenarios.

5.2 Partial Information: Strategic Deviation

Having established that our reinforcement learning algorithm successfully converges to Nash equilibrium under perfect information, we now extend the same learning framework to partial information conditions. This allows us to observe how the identical algorithm-policy gradient with regret minimization-produces qualitatively different strategic behavior when agents face uncertainty about competitor costs.

When agents know demand parameters $[a, b]$ but not competitor costs, they adopt markedly different strategies from Nash equilibrium. The learned policies exhibit a characteristic V-

shaped deviation pattern: overproduction when marginal costs are low (above the Nash line) and underproduction when costs are high (below the Nash line).



(a) Training convergence (b) Policy response curve

Figure 4: Partial information results showing (a) slower convergence with higher regret, and (b) V-shaped deviation from Nash equilibrium with overproduction at low costs and underproduction at high costs.

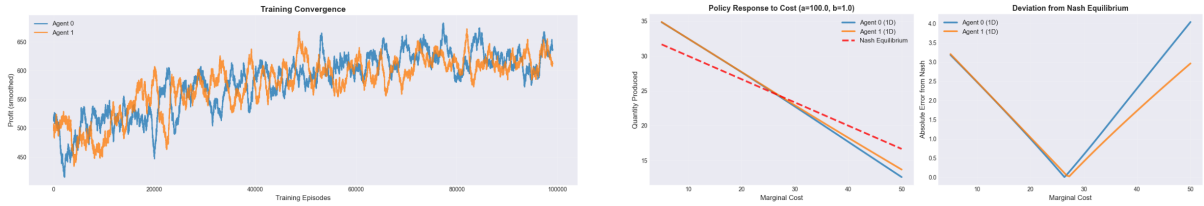
In self-play between two partial-information agents, average quantities reach 27.61 compared to Nash 24.24, representing 14% overproduction. This aggressive stance reduces profits to 531.74, a 17.7% decline from Nash optimal 646.46.

However, the strategic landscape changes dramatically in asymmetric matchups. When a partial-information agent faces a perfect-information agent, the partial agent achieves profit of 638.94 while the perfect agent earns only 557.26-an 81.69 profit advantage despite having less information.

This counter-intuitive result stems from commitment effects. The partial agent learns an aggressive heuristic: "produce high when my cost is low." This commitment to overproduction forces the perfect-information agent to accommodate by reducing its own quantity below Nash levels.

5.3 Asymmetric Information: Maximum Uncertainty

Under extreme information asymmetry where agents observe only their own cost $[c]$, strategic deviations amplify further. The learned policies show even more pronounced V-shaped patterns, with significant overproduction at low costs.



(a) Training convergence (b) Policy response curve

Figure 5: Asymmetric information results showing (a) highest regret levels during training, and (b) most extreme V-shaped deviation with severe overproduction at low costs (quantity 32.48 vs Nash 24.24).

In self-play scenarios, both agents produce 32.48 units (34% above Nash), driving profits down to 277.77-a devastating 57% reduction from Nash optimal. This severe mutual overproduction illustrates market failure where lack of coordination destroys industry value.

In cross-matchups, the asymmetric agent continues to benefit from commitment. Against perfect information, the asymmetric agent earns 577.72 profit while the informed agent gets only 428.35-a 149.36 advantage to ignorance.

5.4 Head-to-Head Matchup Results

Table 2 summarizes all head-to-head matchup results, with Nash equilibrium serving as the reference point.

Table 2: Test results for all agent matchups (Nash reference: $q^* = 24.24$, $P^* = 36.67$, $\pi^* = 646.46$)

Matchup	Agent 1	Profit 1	Qty 1	Agent 2	Profit 2	Qty 2
Test 1	Perfect	650.75	24.08	Perfect	650.75	24.08
Test 2	Perfect	557.26	24.08	Partial	**638.94**	27.61
Test 3	Perfect	428.35	24.08	No-Info	**577.72**	32.48
Test 4	Partial	531.74	27.61	Partial	531.74	27.61
Test 5	Partial	383.94	27.61	No-Info	**451.62**	32.48
Test 6	No-Info	277.77	32.48	No-Info	277.77	32.48

The results reveal a clear pattern: in asymmetric matchups, the less-informed agent consistently earns higher profits (shown in bold). However, in symmetric matchups, less information leads to lower profits for both agents.

5.5 Profit Analysis Across Cost Range

Figure 6 shows profit curves across marginal cost levels for cross-matchups between different agent types.

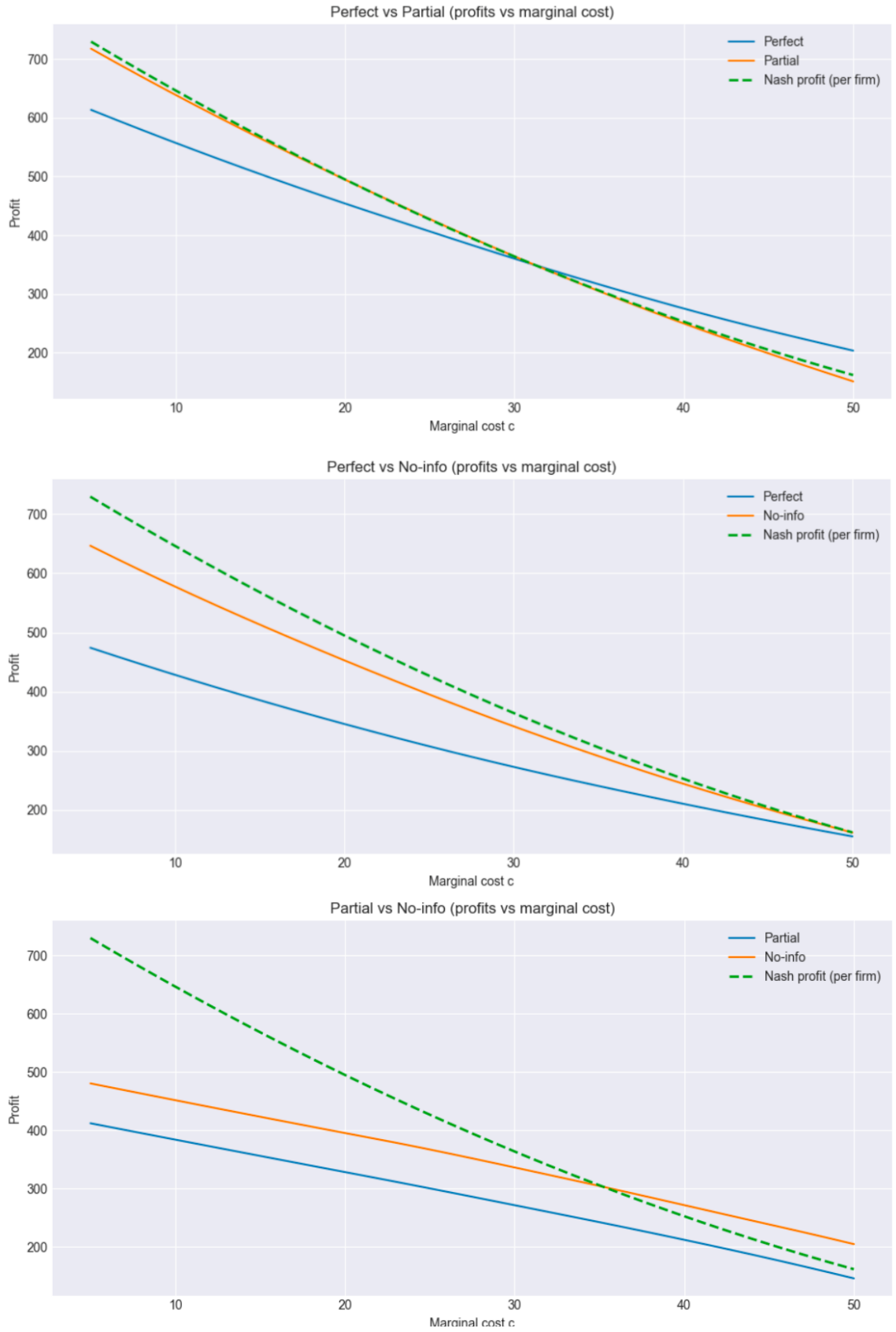


Figure 6: Cross-matchup analysis showing profits versus marginal cost for asymmetric pairings. Less-informed agents (Partial, No-info) consistently outperform Perfect information agents, with the gap largest at moderate costs.

Figure 7 compares self-play performance across information regimes.

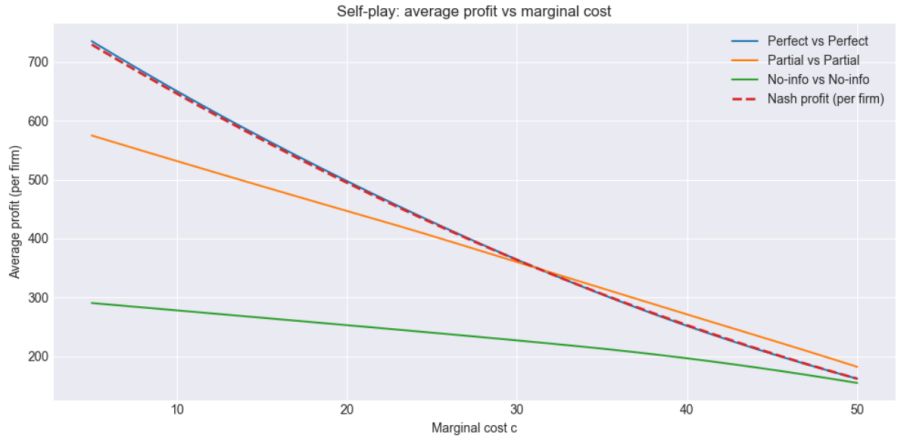


Figure 7: Self-play analysis: average profit versus marginal cost. Perfect information tracks Nash closely, while Partial and No-info show 15–20% and 50–60% profit losses respectively due to mutual overproduction.

Perfect information agents maintain near-Nash performance across all cost levels. Partial and asymmetric agents show non-monotonic profit curves in cross-matchups, with peak commitment advantage at moderate costs. In self-play, profit losses from mutual overproduction are most severe at low costs where overproduction is maximal.

5.6 Rationality Assumptions vs. Robust Strategies

The central empirical finding challenges the conventional superiority of rationality and perfect information in strategic settings. The critical insight is not that information itself creates disadvantage, but rather that assuming competitor rationality and symmetric information becomes a strategic vulnerability in single-shot games.

Perfect information agents operate under two key assumptions: (1) they know all market parameters, and (2) they expect competitors to behave rationally given those parameters. These agents face no uncertainty about the game structure and best-respond optimally to expected rational play. However, this approach fails when competitors are boundedly rational or have learned strategies under different information structures.

In contrast, agents with partial or asymmetric information face inherent uncertainty during training. Unable to perfectly predict competitor behavior, they develop strategies that are robust to uncertainty-aggressive heuristics that perform reasonably well across many possible competitor responses. In single-shot competition, this robustness proves more valuable than precise optimization based on rationality assumptions.

The phenomenon resembles findings in behavioral game theory where boundedly rational strategies can outperform fully rational ones when opponents deviate from Nash equilibrium. Studies of level-k reasoning and quantal response equilibria suggest that assuming perfect rationality can be exploited by agents using simpler decision rules. Our results provide computational evidence that reinforcement learning naturally discovers such exploitable rigidity in rational agents.

Critically, the perfect information agent's disadvantage stems from its flexibility-knowing the competitor's cost, it adjusts its quantity to best-respond to expected rational play. This adaptive behavior, while optimal against rational opponents, becomes exploitable by committed strategies. The uninformed agent, lacking information to fine-tune its response, commits to a fixed aggressive policy that the informed agent must accommodate.

This interpretation reframes "strategic ignorance" not as a benefit of lacking information per se, but as a benefit of developing strategies that don't rely on specific rationality assumptions about opponents. In single-shot games where firms cannot learn about each other through repeated interaction, robustness to opponent uncertainty dominates precise optimization based on potentially incorrect assumptions.

The Trade-off: Exploitation vs. Efficiency: An important qualification emerges from our symmetric matchup results. When two partial-information agents compete, they achieve only 531.74 profit each-17.7% below the Nash equilibrium of 646.46. Similarly, two asymmetric-information agents earn merely 277.77 each-a 57% reduction from optimal. This demonstrates that while uncertainty-robust strategies excel at exploiting rationality-assuming opponents in asymmetric matchups, they do not maximize efficiency when both players employ them.

However, this efficiency loss may conceal an unexplored benefit: resilience to irrationality and external shocks. Strategies developed under uncertainty might be inherently more robust to unexpected competitor behaviors-true irrationality, mistakes, or departures from learned patterns-and to market shocks that alter demand or cost parameters unpredictably. While Nash-optimizing agents are calibrated to specific parameter values and rational opponent responses, uncertainty-robust strategies may maintain more stable performance when these assumptions are violated. This potential resilience-efficiency trade-off merits investigation in future extensions, particularly in environments with stochastic shocks or boundedly rational opponents who deviate unpredictably from equilibrium play.

5.7 Implications for Market Design

These findings offer nuanced insights for market regulation and information policy that go beyond simple prescriptions for transparency.

Our results demonstrate that information disclosure policies are most effective when they create symmetric information environments. When all firms have access to and utilize market information rationally, the market converges to Nash equilibrium with high industry profits (650.75 per firm in our tests). The benefit of transparency emerges from coordination, not from information per se.

However, information disclosure becomes problematic when access or utilization is heterogeneous. If some firms optimize based on disclosed information and rationality assumptions while others-whether due to limited analytical capacity, strategic choice, or information barriers-develop simpler heuristics, the optimizing firms become vulnerable to exploitation. The policy implication is not that disclosure is harmful, but that incomplete or asymmetric adoption of information-based optimization can destabilize markets in single-shot competitive settings.

Markets naturally contain firms with varying levels of analytical sophistication. Large firms may employ teams of economists and data scientists to optimize production decisions, while

smaller firms may rely on simpler rules of thumb. Our results suggest that in one-shot competitive interactions, this heterogeneity can produce outcomes where sophisticated optimization underperforms simple heuristics-not because sophistication is inherently worse, but because it relies on assumptions about competitor behavior that may not hold.

Industry Efficiency vs. Distributional Effects: An important caveat emerges from our symmetric information results. When both firms have partial information, industry profits decline to 531.74 per firm-17.7% below Nash equilibrium. When both lack information entirely, profits collapse to 277.77 per firm-a 57% reduction. This indicates that while uncertainty-robust strategies can exploit rationality-assuming opponents in asymmetric matchups, industry-wide uncertainty reduces total welfare.

Therefore, from an efficiency perspective, policies that promote symmetric information access and rational utilization remain desirable-they maximize total industry profits and enable coordination. The strategic advantage of bounded rationality documented here is primarily a distributional effect in asymmetric competitive environments, not an efficiency improvement.

Aggressive production by uncertainty-robust firms increases total quantity and reduces prices in single-shot markets, potentially benefiting consumers. However, this benefit must be weighed against lower industry profits that may reduce investment, innovation, or market sustainability.

These implications apply primarily to single-shot or infrequent competitive interactions where firms cannot learn about each other through repeated play. In markets with frequent interaction, information revelation through observed behavior, and opportunities for tacit coordination, the dynamics differ substantially. Repeated games likely favor rational, information-based strategies over simple heuristics as firms learn to anticipate and coordinate with competitor behavior.

6 Limitations and Conclusion

6.1 What Worked Well

Our implementation successfully demonstrated that neural network agents can learn strategic behavior in oligopolistic competition through policy gradient reinforcement learning. The residual architecture combining Nash baselines with learned corrections proved highly effective, enabling stable training and convergence to equilibrium under perfect information.

The regret minimization learning signal provided intuitive and powerful gradients. By framing learning as reducing the difference between realized and best-response profits, agents directly optimize the game-theoretic objective.

The three-way information structure comparison yielded clear, interpretable, and theoretically interesting results. The emergence of commitment value from information asymmetry was not explicitly programmed but arose naturally from the learning dynamics, validating the approach as a discovery tool for strategic phenomena.

6.2 Limitations

Several limitations constrain the generality of our findings:

Static Competition: Our model considers repeated training episodes but evaluates performance in **single-shot games** where firms make one-time decisions without observing competitor responses or building reputation. In real markets, firms often compete repeatedly, allowing them to learn about competitor costs, signal intentions, and coordinate strategies over time. The strategic advantage of uncertainty-robust strategies documented here may diminish or disappear in repeated games where information revelation occurs naturally through observed play. Extending the framework to truly dynamic, multi-period competition where current decisions affect future states would test whether our findings generalize beyond one-shot interactions.

Linear Demand: The assumption of linear inverse demand simplifies analysis but may not capture realistic demand structures with non-linearities.

Symmetric Costs in Evaluation: While training exposes agents to varied cost distributions, our head-to-head evaluations fix both firms at identical cost $c = 10$.

Two-Agent Restriction: Real oligopolies often involve three or more competitors. Extending to n-player games would test whether commitment effects scale with market structure.

Exploration Strategy: Our Gaussian noise exploration is simple but potentially suboptimal. More sophisticated methods might accelerate learning or discover different strategies.

No Strategic Information Acquisition: Agents cannot choose what information to observe or invest in market research.

6.3 Future Improvements

If time and computational resources were unlimited, several extensions would strengthen the research:

Dynamic Games: Implementing multi-period competition where firms observe price realizations and update beliefs would better reflect real markets.

Collusion and Communication: Allowing agents to observe each other's actions or engage in cheap talk could reveal whether learned strategies support tacit collusion.

Heterogeneous Costs and Capacities: Extending beyond symmetric firms to include capacity constraints and economies of scale would test robustness.

Alternative Learning Algorithms: Comparing policy gradient methods to Q-learning, actor-critic architectures, and evolutionary strategies would identify which algorithms best capture oligopolistic learning.

Empirical Calibration: Calibrating model parameters to real market data and comparing learned behaviors to actual firm strategies would validate empirical relevance.

Welfare Analysis: Computing consumer surplus, producer surplus, and deadweight loss across information regimes would provide complete welfare accounting.

Robustness to Shocks and Irrationality: A critical unexplored dimension is whether uncertainty-robust strategies maintain superior performance under environmental volatility or opponent irrationality. While our results show that partial-information agents achieve lower profits in symmetric matchups (531.74 vs Nash 646.46), these strategies might exhibit greater resilience when demand parameters shift unexpectedly, costs fluctuate, or competitors make errors. Experiments introducing stochastic demand/cost shocks, boundedly rational opponents

who deviate randomly from learned policies, or model misspecification could test whether the efficiency losses of uncertainty-robust strategies are compensated by superior stability and adaptability. This resilience-efficiency trade-off has important implications for market design in volatile or unpredictable environments.

Robustness to Non-Stationarity: Testing how agents adapt when market parameters shift (demand shocks, cost changes, entry of new competitors) would assess the practical value of learned strategies. Transfer learning and meta-reinforcement learning methods could enable rapid adaptation to new environments.

6.4 Lessons Learned

This project provided valuable lessons spanning game theory, reinforcement learning, and empirical methodology.

Theoretical Insights: The finding that uncertainty-robust strategies can outperform rationality-based optimization in single-shot games deepened understanding of strategic interaction. Classical game theory emphasizes rationality and common knowledge, but bounded rationality arising from partial observability-combined with the absence of repeated learning opportunities-creates qualitatively different equilibria. The result challenges the assumption that perfect information and rationality are always advantageous, highlighting the strategic value of robustness when opponent behavior is uncertain and cannot be refined through repeated play.

Implementation Skills: Developing a modular, extensible codebase for multi-agent learning required careful software design. Debugging multi-agent systems proved challenging-coordinating exploration, ensuring consistent state representations, and validating gradient calculations demanded systematic testing.

Hyperparameter Sensitivity: The success of policy gradient methods depends critically on well-tuned learning rates, exploration schedules, and normalization schemes. Developing intuition for these sensitivities through experimentation was essential.

Visualization and Interpretation: Translating learned policies into interpretable insights required thoughtful visualization design. Policy response curves proved more informative than raw profit plots.

6.5 Conclusion

This research demonstrates that multi-agent reinforcement learning provides a powerful framework for studying strategic behavior in oligopolistic markets. Our neural network agents successfully learned to compete in a single-shot Cournot duopoly under varying information conditions, validating theoretical predictions while uncovering new insights about the interaction between information, rationality assumptions, and strategic performance.

The central finding challenges standard assumptions about rationality and information value in strategic settings: in single-shot games, agents that develop uncertainty-robust strategies through learning under incomplete information outperform agents that optimize based on perfect information and rationality assumptions. While perfect information enables agents to converge to Nash equilibrium and play optimally against rational opponents, these same agents

become vulnerable when facing boundedly rational competitors who have learned aggressive strategies that don't depend on specific information about opponent types.

This is not a simple story of "ignorance is strength." Rather, it reveals that assuming competitor rationality and symmetric information-core tenets of classical game theory-creates strategic rigidity that can be exploited. Perfect information agents best-respond to expected rational play, but this flexibility becomes a weakness when competitors commit to strategies learned under uncertainty. The key insight is that robustness to opponent uncertainty dominates precise optimization when repeated learning opportunities are absent.

The symmetric ignorance case provides an important counterpoint: when all agents lack information and develop aggressive strategies, mutual overproduction produces coordination failure and industry-wide value destruction (57% profit reduction from Nash). Partial information self-play similarly yields 17.7% profit losses. This reveals a critical trade-off: uncertainty-robust strategies excel at exploiting rational opponents but reduce efficiency when universally adopted.

However, this apparent inefficiency may mask an important benefit worthy of future investigation. Strategies developed under uncertainty could exhibit greater resilience to irrationality and external shocks-maintaining more stable performance when competitors behave unpredictably or when market parameters shift unexpectedly. Nash-optimizing agents are calibrated to specific conditions and rational responses; disruptions to these assumptions may degrade performance more severely than for strategies that never relied on such assumptions. Whether the efficiency losses documented in symmetric low-information matchups are offset by superior robustness to model misspecification, competitor mistakes, or environmental volatility remains an open and important question for future research.

These results have important implications for market design and regulation. Information disclosure policies maximize efficiency when they create symmetric information access and rational utilization by all firms. However, heterogeneous strategic sophistication-where some firms optimize rationally while others use simple heuristics-can create exploitable asymmetries in single-shot competitive settings. The documented effects likely diminish in repeated interactions where learning and coordination become possible.

Beyond specific findings about Cournot competition, this work illustrates the potential of reinforcement learning as a tool for discovering emergent phenomena in strategic interactions. Unlike analytical game theory, which requires strong assumptions about rationality and common knowledge, learned strategies emerge from experience and naturally exhibit bounded rationality, satisficing, and commitment to simple heuristics-properties that may better reflect real firm behavior than perfect rationality.

Future research extending this framework to repeated games would clarify whether uncertainty-robust strategies maintain their advantage when firms can learn about competitors through multiple interactions. Equally important is investigating whether the efficiency losses of uncertainty-robust strategies in symmetric matchups are offset by superior resilience to environmental shocks, demand/cost volatility, or boundedly rational competitor behavior. Extensions incorporating stochastic market dynamics, competitor irrationality, and model misspecification-alongside studies of larger oligopolies and empirically calibrated settings-could generate actionable in-

sights for competition policy, market design, and firm strategy in both stable and volatile competitive contexts.

References

[1] Buşoniu, L., Babuška, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2), 156–172.

[2] Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.

[3] Cournot, A. A. (1838). *Recherches sur les principes mathématiques de la théorie des richesses*. L. Hachette.

[4] Fudenberg, D., & Tirole, J. (1991). *Game theory*. MIT Press.

[5] Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. *Machine Learning Proceedings 1994*, 157–163.

[6] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

[7] Vives, X. (1999). *Oligopoly pricing: Old ideas and new tools*. MIT Press.

[8] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3), 229–256.

[9] Shi, Y., & Zhang, B. (2020). *MARL-Cournot-Games* [Source code]. GitHub. `https://github.com/Yuanyuan-Shi/MARL-Cournot-Games`

[10] Bapuram, H. (2022). *Cournot model simulation* [Source code]. Kaggle. `https://www.kaggle.com/code/hrishitabapuram/cournot-model-simulation`