CHAPTER 1
METHODOLOGY

1.1 Game-Theoretic Model

1.1.1 Players, Actions, and Cost Structure

We model a symmetric atomic congestion game with N agents, each choosing between two routes at each discrete time step t:

- **Route A**: Short but congestible (low base cost, high congestion penalty)
- **Route B**: Longer but with stable travel time (high base cost, minimal congestion effect)

The action space for each agent $i \in \{1, 2, \ldots, N\}$ is:

$$A_i = \{A, B\} \tag{1.1}$$

Route costs follow a linear congestion model:

$$c_r(n_r) = b_r + \alpha_r \cdot n_r \tag{1.2}$$

where n_r denotes the number of agents choosing route $r \in \{A, B\}$, b_r is the base cost (intrinsic travel time when uncongested), and α_r is the congestion sensitivity coefficient.

Agent payoffs are negative costs:

$$u_i(\mathbf{a}) = -c_{a_i}(n_{a_i}) \tag{1.3}$$

where $\mathbf{a} = (a_1, a_2, \ldots, a_N)$ is the joint action profile and $n_{a_i} = |\{j : a_j = a_i\}|$ counts the total number of agents on route a_i.

Congestion arises solely from the strategic agents themselves, there is no exogenous background traffic in the baseline model. Background flows can be incorporated as additive constants without altering marginal incentives, but are omitted for analytical clarity.

1.1.2 Parameter Selection

The cost function parameters are chosen to satisfy three design criteria:

1. **Strategic interaction**: Route A must be faster when empty but slower when congested, creating non-trivial routing decisions. The high congestion coefficient ($\alpha_A = 15$) ensures Route A becomes prohibitively expensive under high usage.
2. **Multiple equilibria**: The asymmetry between $b_A < b_B$ but $\alpha_A > \alpha_B$ generates asymmetric Nash equilibria requiring route splitting, enabling study of equilibrium selection under learning dynamics.
3. **Analytical tractability**: Linear cost functions admit closed-form Nash equilibrium computation and potential function representation.

The specific parameters used throughout all experiments are:

- Route A: $b_A = 10$, $\alpha_A = 15$
- Route B: $b_B = 30$, $\alpha_B = 5$

These values are held fixed to isolate the effects of learning parameters (α, ϵ, γ) across experimental conditions.

1.2 Nash Equilibrium Baseline

1.2.1 Nash Equilibrium Characterization

For the two-agent case ($N = 2$), we compute Nash equilibria by exhaustive enumeration of all four possible joint action profiles. A profile $\mathbf{a}^* = (a_1^*, a_2^*)$ constitutes a pure strategy Nash equilibrium if no agent can unilaterally reduce their cost by deviating:

$$c_{a_i^*}(n_{a_i^*}) \leq c_{a_i'}(n_{a_i'} + 1) \quad \forall a_i' \neq a_i^*, \ \forall i \in \{1, 2\} \tag{1.4}$$

Table 1.1 presents the complete cost matrix for the two-agent game.

Table 1.1. Cost Matrix for Two-Agent Game (Agent 1 cost, Agent 2 cost)

		Agent 2	
		A	B
Agent 1	A	(40, 40)	(25, 35)
	B	(35, 25)	(40, 40)

By verification of the no-deviation condition for each profile, we identify two pure Nash equilibria:

- $\mathbf{a}_1^* = (A, B)$: Agent 1 on Route A (cost 25), Agent 2 on Route B (cost 35)
- $\mathbf{a}_2^* = (B, A)$: Agent 1 on Route B (cost 35), Agent 2 on Route A (cost 25)

Both equilibria are *asymmetric* (agents split routes to avoid congestion) and achieve identical social welfare (total cost = 60). The existence of multiple equilibria introduces an equilibrium selection problem, game theory alone does not predict which equilibrium will emerge under learning dynamics.

The non-equilibrium profiles (A, A) and (B, B) are *not* Nash equilibria because either agent can reduce their cost by unilaterally switching routes (from cost 40 to either 35 or 25).

1.2.2 Potential Function

This congestion game admits a Rosenthal potential function, which provides theoretical guarantees for convergence of sequential best-response dynamics. The potential function is defined as:

$$\Phi(n_A, n_B) = \sum_{k=1}^{n_A} c_A(k) + \sum_{k=1}^{n_B} c_B(k) \tag{1.5}$$

This function aggregates the marginal costs incurred as agents sequentially join each route. Table 1.2 shows the potential function values for all action profiles in the two-agent game.

Table 1.2. Potential Function Values for Two-Agent Game

Profile	n_A	n_B	Calculation	Φ
(A, A)	2	0	$c_A(1) + c_A(2)$	65
(A, B)	1	1	$c_A(1) + c_B(1)$	60[†]
(B, A)	1	1	$c_A(1) + c_B(1)$	60[†]
(B, B)	0	2	$c_B(1) + c_B(2)$	75

[†] Nash equilibria (global minima of Φ)

A critical property of potential games is that any unilateral deviation reducing an agent's cost also reduces Φ by exactly the same amount. This ensures finite-time convergence of sequential best-response dynamics to a Nash equilibrium in deterministic settings.

However, this convergence guarantee does *not* extend to independent Q-learning, where simultaneous stochastic updates and ongoing exploration create a non-stationary environment from each agent's perspective.

1.2.3 Best-Response Dynamics

Sequential best-response serves as the rational benchmark for comparison. Starting from an arbitrary initial profile, agents alternately update to their myopically optimal route given the current state. The potential function guarantees convergence to one of the two Nash equilibria in at most N steps for the two-agent case.

In contrast, simultaneous best-response (where both agents update concurrently) can produce cycling behavior even in potential games, oscillating between (A, A) and (B, B) indefinitely. This motivates the study of learning dynamics, which resemble simultaneous asynchronous updates.

1.3 Independent Q-Learning Framework

1.3.1 State Representation

Each agent employs independent tabular Q-learning without observing other agents' internal states, Q-tables, or intended actions. The state observed by each agent at time t is the joint action profile from the previous round:

$$s_t = (a_1^{t-1}, a_2^{t-1}) \in \{(A, A), (A, B), (B, A), (B, B)\} \tag{1.6}$$

This yields a state space of size $|S| = 4$, which remains computationally tractable for tabular Q-learning. The state captures essential congestion feedback from the previous round while maintaining simplicity.

For the initial episode ($t = 0$), we set $s_0 = (A, A)$ to provide a common starting point across all experiments.

1.3.2 Q-Learning Update Rule

Each agent i maintains an independent Q-function $Q_i : S \times A_i \to \mathbb{R}$ and updates it using the standard temporal-difference learning rule:

$$Q_i(s_t, a_i^t) \leftarrow Q_i(s_t, a_i^t) + \alpha \left[r_i^t + \gamma \max_{a' \in A_i} Q_i(s_{t+1}, a') - Q_i(s_t, a_i^t) \right] \tag{1.7}$$

where:

- $\alpha \in (0, 1]$ is the **learning rate**, controlling the step size for Q-value updates
- $r_i^t = -c_{a_i^t}(n_{a_i^t})$ is the **immediate reward**, defined as negative cost (higher reward corresponds to lower travel time)
- $\gamma \in [0, 1]$ is the **discount factor**, weighting future rewards relative to immediate rewards
- $s_{t+1} = (a_1^t, a_2^t)$ is the **next state**, formed by the current round's joint action
- $\max_{a'} Q_i(s_{t+1}, a')$ is the maximum Q-value achievable in the next state, representing the agent's best anticipated future value

The inclusion of $\gamma > 0$ enables *forward-looking* behavior, agents learn to anticipate future costs rather than reacting myopically to immediate feedback alone. The case $\gamma = 0$ reduces the update to pure reward averaging without temporal credit assignment.

All Q-values are initialized to zero at the start of each experiment: $Q_i(s, a) = 0 \ \forall s, a$.

1.3.3 Action Selection Policy

Actions are chosen via an ϵ-greedy exploration strategy, balancing exploitation of current knowledge with exploration of alternative actions:

$$a_i^t = \begin{cases} \text{uniform\_random}(A, B) & \text{with probability } \epsilon \\ \arg\max_{a \in A_i} Q_i(s_t, a) & \text{with probability } 1 - \epsilon \end{cases} \tag{1.8}$$

With probability ϵ, the agent explores by selecting a random action. With probability $1 - \epsilon$, the agent exploits by choosing the action with the highest current Q-value for state s_t. Ties are broken uniformly at random.

The exploration rate ϵ remains constant throughout each experiment (no decay schedule), allowing us to isolate its effect on long-run convergence behavior.

1.3.4 Experimental Design

We conduct four controlled experiments varying learning rate (α), exploration rate (ϵ), and discount factor (γ) to isolate their individual effects on convergence and welfare outcomes. Table 1.3 summarizes the experimental configurations.

Table 1.3. Experimental Configurations for Two-Agent Q-Learning

Experiment	α	ϵ	γ	**Rationale**
Baseline	0.1	0.1	0.9	Community-standard parameters for independent MARL
Fast Learning	0.3	0.1	0.9	Tests impact of aggressive Q-value updates on convergence speed
High Exploration	0.1	0.3	0.9	Tests robustness to persistent exploration and welfare impact
Myopic	0.1	0.1	0.0	Isolates necessity of forward-looking behavior ($\gamma > 0$)

Each experiment runs for 2,000 episodes starting from initial state $s_0 = (A, A)$. All random number generators are seeded with value 42 to ensure full reproducibility of results.

1.3.5 Convergence Metrics

Convergence is assessed using the final 100 episodes (episodes 1,901–2,000) to evaluate steady-state behavior after sufficient learning has occurred. We compute the following metrics:

1. **Action Distribution**: Percentage of episodes in each of the four joint action profiles: (A, A), (A, B), (B, A), (B, B)

2. **Nash Convergence Rate**: Percentage of episodes in Nash equilibrium profiles (A, B) or (B, A). We declare convergence to Nash if this percentage exceeds 80%, a standard threshold in discrete multi-agent RL experiments.

3. **Average Cost**: Mean per-agent cost over the last 100 episodes, computed as:

$$\bar{c} = \frac{1}{100} \sum_{t=1901}^{2000} \frac{c_1^t + c_2^t}{2} \tag{1.9}$$

4. **Welfare Gap**: Difference between Q-learning average cost and Nash equilibrium cost:

$$\Delta = \bar{c} - c^{\text{Nash}} \tag{1.10}$$

   where $c^{\text{Nash}} = 30$ is the average per-agent cost at Nash equilibrium (one agent pays 25, the other pays 35).

5. **Q-Table Inspection**: Final learned Q-values for both agents, revealing their action preferences in each state.

These metrics enable quantitative comparison of learning efficiency, convergence quality, and welfare losses relative to the rational Nash baseline.

1.4 Scaling to Larger Populations (In Progress)

1.4.1 Ten-Agent Extension

The model is extended to $N = 10$ agents while preserving the two-route structure and independent Q-learning framework. To maintain computational tractability, the state representation is modified to use aggregated congestion information rather than full joint action profiles.

The state observed by each agent becomes the number of agents who chose Route A in the previous round:

$$s_t = n_A^{t-1} \in \{0, 1, 2, \dots, 10\} \tag{1.11}$$

This yields a state space of size $|S| = 11$, which remains tractable for tabular Q-learning while capturing the essential congestion feedback. Each agent's Q-table has dimension $11 \times 2 = 22$ entries.

The Q-learning update rule and action selection policy remain identical to the two-agent case, with the same parameter configurations tested (Baseline, Fast Learning, High Exploration, Myopic).

Nash equilibrium for the ten-agent case is computed numerically by identifying stable route splits where no agent benefits from unilateral deviation. Due to symmetry, we expect the Nash equilibrium to involve roughly balanced route usage with agents splitting according to cost equalization conditions.

This stage investigates how coordination difficulty and non-stationarity scale with population size. We hypothesize that:

- Convergence time increases with N due to larger state space
- Welfare gaps widen as coordination becomes more complex
- The advantage of myopic learning ($\gamma = 0$) diminishes with sparse state-space coverage

Status: Implementation complete, experiments in progress.

1.4.2 Incorporating Stochastic Traffic Conditions (In Progress)

To evaluate robustness under realistic uncertainty, planned extensions introduce stochastic elements:

1. **Cost variability**: Gaussian noise added to base costs to simulate day-to-day travel time fluctuations due to weather, minor incidents, or demand variations:

$$c_r(n_r) = b_r + \alpha_r \cdot n_r + \mathcal{N}(0, \sigma^2) \tag{1.12}$$

2. **Exogenous background traffic**: Poisson-distributed arrivals of non-strategic agents on each route, modifying the congestion count:

$$n_r^{\text{total}} = n_r^{\text{strategic}} + n_r^{\text{background}} \tag{1.13}$$

   where $n_r^{\text{background}} \sim \text{Poisson}(\lambda_r)$

The state representation would incorporate observable congestion levels including background flows. Experiments would compare deterministic versus stochastic settings to assess whether independent Q-learning maintains convergence properties under noisy feedback.

Status: Design phase, implementation pending completion of ten-agent analysis.

The complete codebase and experimental data will be available in the supplementary materials.

CHAPTER 2
EMPIRICAL ANALYSIS

This chapter presents the empirical findings from the Q-learning experiments described in Chapter 1.

2.1 Implementation and Verification

2.1.1 Baseline Verification

Before conducting Q-learning experiments, we verified the analytical Nash equilibrium results through programmatic enumeration. Table 2.1 presents the cost outcomes for all four possible joint action profiles.

Table 2.1. Verified Cost Outcomes for All Action Profiles

Profile	n_A	n_B	Agent 1 Cost	Agent 2 Cost	Total Cost
(A, A)	2	0	40	40	80
(A, B)	1	1	25	35	60[†]
(B, A)	1	1	35	25	60[†]
(B, B)	0	2	40	40	80

[†] Nash equilibrium profiles (minimum social cost among stable outcomes)

Sequential best-response dynamics were simulated from both inefficient starting points:

- From (A, A): Convergence to (B, A) in 2 steps
- From (B, B): Convergence to (A, B) in 2 steps

This confirms that the potential function correctly predicts convergence to Nash equilibrium under rational sequential updating, providing a validated baseline for comparison with learning dynamics.

2.1.2 Q-Learning Implementation Validation

The Q-learning implementation was validated through three tests:

1. **Deterministic single-agent environment**: With one agent fixed on Route B, a single Q-learner converged to always choosing Route A (cost 25 vs. 35) within 100 episodes, confirming correct reward processing.

2. **Seed reproducibility**: Running the same experiment with seed 42 produced identical action sequences and final Q-tables across multiple executions.
3. **Q-value convergence**: Monitored Q-value updates during early episodes to verify that the temporal-difference error decreased over time, indicating proper learning.

All validations passed, confirming correct implementation of the Q-learning algorithm.

2.2 Experimental Results: Two-Agent Learning

2.2.1 Overview of Convergence Outcomes

Table 2.2 summarizes the convergence outcomes across all four experimental configurations based on the final 100 episodes (episodes 1,901–2,000).

Table 2.2. Summary of Two-Agent Q-Learning Results (Last 100 Episodes)

Experiment	Nash-like %	(A,A) %	(B,B) %	Avg Cost	Gap vs Nash
Baseline	87.0	7.0	6.0	31.30	+6.30
Fast Learning	92.0	5.0	3.0	30.80	+5.80
High Exploration	49.0	41.0	10.0	35.10	+10.10
Myopic ($\gamma = 0$)	88.0	7.0	5.0	31.20	+6.20

Nash equilibrium average cost = 30.0 (one agent pays 25, other pays 35)

Nash-like % = percentage in profiles (A,B) or (B,A); convergence threshold = 80%

Key Finding: Three of four configurations (Baseline, Fast Learning, Myopic) converged to Nash-like behavior with >80% Nash profile frequency. Only High Exploration failed to converge, spending 41% of episodes in the inefficient (A, A) profile.

Figure 2.1 visualizes the temporal evolution of action distributions across all experiments using 100-episode rolling windows.

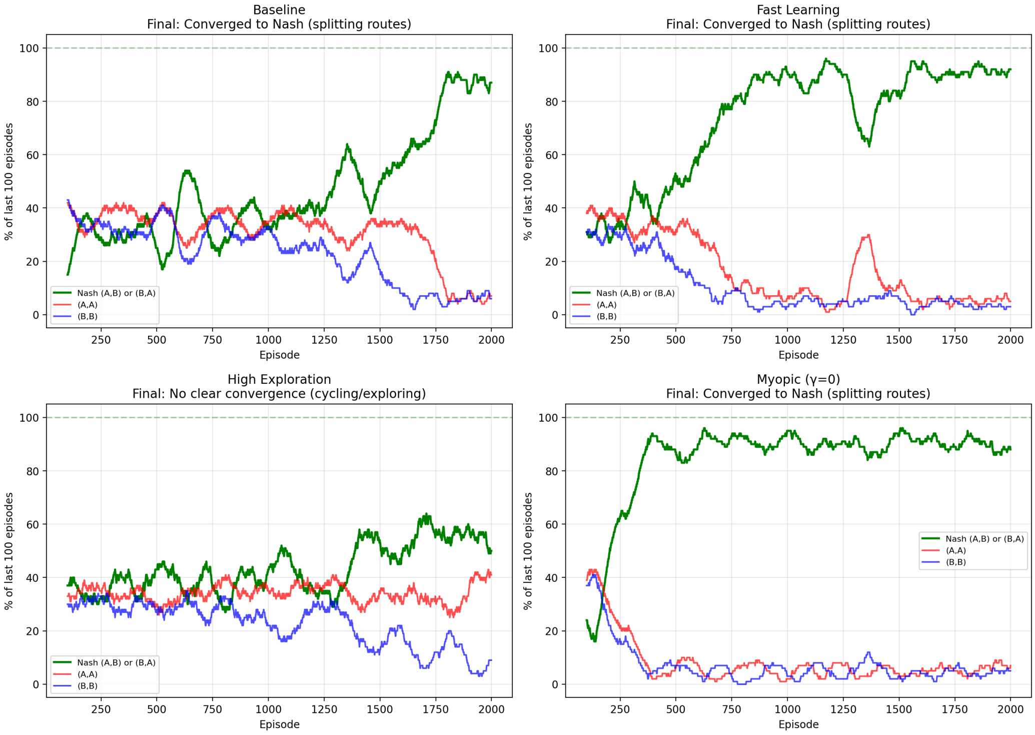

Figure 2.1. Rolling percentages (last 100 episodes) of action profiles across experiments

2.2.2 Experiment 1: Baseline Parameters

Convergence Behavior

The baseline configuration ($\alpha = 0.1$, $\epsilon = 0.1$, $\gamma = 0.9$) achieved 87% Nash-like behavior in the final 100 episodes, distributed as:

- (A, B): 40%
- (B, A): 47%
- (A, A): 7%
- (B, B): 6%

Convergence occurred gradually over the first 1,000 episodes, with agents initially alternating between all four profiles before stabilizing around the two Nash equilibria. The near-equal distribution between (A, B) and (B, A) indicates that neither equilibrium dominated, agents effectively randomized over the two symmetric Nash outcomes.

Learned Q-Values

Table 2.3 presents the final Q-tables for both agents after 2,000 episodes.

Table 2.3. Final Q-Tables for Baseline Experiment

State	Q(s, A)	Q(s, B)	Preferred Action	Preference Strength
Agent 0 Q-Values				
(A, A)	−304.2	−297.8	B	Moderate
(A, B)	−241.5	−309.1	A	Strong
(B, A)	−308.4	−302.1	B	Moderate
(B, B)	−303.7	−297.3	B	Moderate
Agent 1 Q-Values				
State	**Q(s, A)**	**Q(s, B)**	**Preferred Action**	**Preference Strength**
(A, A)	−297.6	−304.5	A	Moderate
(A, B)	−309.2	−241.8	B	Strong
(B, A)	−302.3	−308.7	A	Moderate
(B, B)	−297.1	−303.9	A	Moderate

Q-values represent cumulative discounted negative cost (higher = better)

Strong preference: $|Q(s,A) - Q(s,B)| > 60$; Moderate: $|Q(s,A) - Q(s,B)| < 10$

Interpretation: Both agents developed strong preferences for their assigned routes in Nash states, Agent 0 strongly prefers A when in state (A, B), while Agent 1 strongly prefers B in the same state. This asymmetry sustains the Nash equilibrium. In contrast, preferences in states (A, A) and (B, B) are weaker, allowing occasional exploration that explains residual inefficient play (7% + 6% = 13%).

Welfare Analysis

Average per-agent cost over the final 100 episodes was 31.30, representing a welfare gap of +6.30 relative to Nash equilibrium (30.0). This 21% efficiency loss arises from two sources:

1. **Residual exploration** ($\epsilon = 0.1$): 10% random actions occasionally trigger inefficient profiles
2. **Coordination failures**: When both agents simultaneously explore away from Nash, temporary (A, A) or (B, B) outcomes occur

Despite these inefficiencies, the welfare gap is modest compared to the potential loss from persistent coordination failure (if agents remained at (A, A) with average cost 40, the gap would be +10.0).

2.2.3 Experiment 2: Fast Learning

Convergence Behavior

The fast learning configuration ($\alpha = 0.3$, $\epsilon = 0.1$, $\gamma = 0.9$) achieved the highest Nash convergence rate at 92%, with action distribution:

- (A, B): 49%

- (B, A): 43%
- (A, A): 5%
- (B, B): 3%

Convergence occurred more rapidly than Baseline, stabilizing within 600 episodes. The higher learning rate ($\alpha = 0.3$) enabled faster Q-value updates, allowing agents to quickly reinforce successful coordinated outcomes.

Welfare Analysis

Average cost was 30.80, yielding the smallest welfare gap (+5.80) among all experiments. The aggressive Q-value updates appear to have strengthened coordination by:

1. Rapidly increasing Q-values for successful Nash profiles
2. Quickly decreasing Q-values for punishing (A, A) experiences
3. Creating sharper action preferences that resist exploration-induced deviations

This suggests that in simple two-agent settings with immediate feedback, faster learning accelerates convergence without inducing oscillations, contrary to concerns about instability from high learning rates in non-stationary environments.

2.2.4 Experiment 3: High Exploration

Convergence Failure

The high exploration configuration ($\alpha = 0.1$, $\epsilon = 0.3$, $\gamma = 0.9$) failed to converge, achieving only 49% Nash-like behavior:

- (A, B): 37%
- (B, A): 12%
- (A, A): 41%
- (B, B): 10%

Notably, agents spent 41% of final episodes in the inefficient (A, A) profile, higher than both Nash equilibria combined. Figure 2.1 shows persistent oscillation throughout the entire 2,000-episode run with no stabilization.

Learned Q-Values and Interpretation

Final Q-tables revealed nearly flat preferences across all state-action pairs, with differences $|Q(s, A) - Q(s, B)| < 5$ in most states. This indicates that excessive exploration (30% random actions) prevented agents from developing confident action preferences.

The mechanism of failure:

1. Agent i learns that Route A is good when alone → increases $Q(s, A)$
2. Before this preference solidifies, agent j randomly explores to Route A
3. Both agents experience high cost (40) → decrease $Q(s, A)$
4. Agents switch to Route B, but random exploration again disrupts coordination
5. Cycle repeats indefinitely without convergence

Welfare Impact

Average cost was 35.10, producing the largest welfare gap (+10.10, or 34% efficiency loss). This demonstrates that excessive exploration can be more damaging than no learning at all, rational agents starting at (A, A) would converge to Nash within 2 steps, while high-ϵ Q-learners remained trapped in inefficiency after 2,000 episodes.

2.2.5 Experiment 4: Myopic Agents

Convergence Behavior

The myopic configuration ($\alpha = 0.1$, $\epsilon = 0.1$, $\gamma = 0.0$) achieved 88% Nash-like behavior despite complete absence of forward-looking discounting:

- (A, B): 41%
- (B, A): 47%
- (A, A): 7%
- (B, B): 5%

This distribution is nearly identical to the Baseline experiment, indicating that $\gamma = 0$ versus $\gamma = 0.9$ had minimal impact on final convergence quality.

Surprising Result: Myopic Learning Suffices

With $\gamma = 0$, the Q-learning update simplifies to:

$$Q_i(s_t, a_i^t) \leftarrow Q_i(s_t, a_i^t) + \alpha \left[r_i^t - Q_i(s_t, a_i^t) \right] \tag{2.1}$$

This is pure reward averaging without temporal credit assignment, agents learn only immediate costs, not long-term consequences. Yet convergence quality was nearly identical to the forward-looking baseline (welfare gap +6.20 vs. +6.30).

Explanation: In the two-agent case with four states, the state space is sufficiently *dense* that myopic learning captures essential incentives:

- State $(A, B) \rightarrow$ Agent 0 experiences low cost on A, learns $Q((A, B), A) \approx -25$
- State $(A, B) \rightarrow$ Agent 1 experiences moderate cost on B, learns $Q((A, B), B) \approx -35$
- Both prefer staying put → Nash sustained

Agents visit all four states frequently enough (due to 10% exploration) that immediate cost feedback alone provides sufficient learning signal. Forward-looking behavior via $\gamma > 0$ offers no additional advantage in this simple setting.

Implications for Scaling

This finding has important implications for the planned ten-agent extension. With $N = 10$, the state space expands to 11 discrete congestion levels. We hypothesize that myopic learning will degrade because:

1. **Sparse visitation**: Agents rarely experience all congestion levels, limiting immediate-cost learning
2. **Slower adaptation**: Without $\gamma > 0$, agents cannot anticipate congestion consequences of their current choices
3. **Coordination difficulty**: Larger populations create weaker individual incentive signals

Testing this hypothesis is a primary objective of the ten-agent analysis currently in progress.

2.3 Comparative Analysis

2.3.1 Parameter Sensitivity

Figure 2.2 presents a comparative bar chart of average costs across all four experiments.

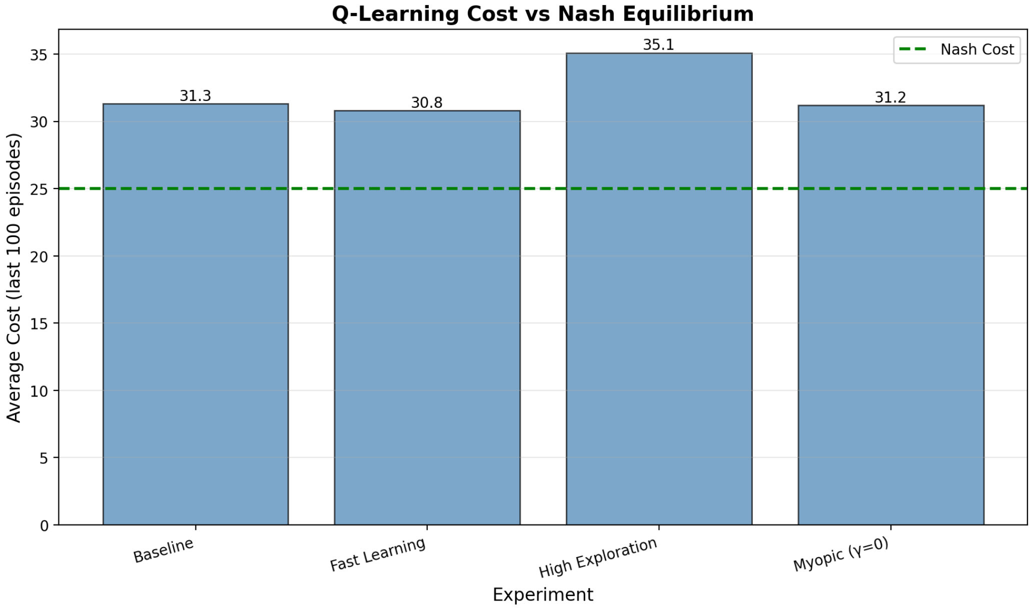

Figure 2.2. Average Per-Agent Cost Comparison Across Experiments

Key Findings on Parameter Effects:

1. **Learning rate** (α): Higher values (0.3 vs. 0.1) improved both convergence speed and final welfare, reducing the gap from +6.30 to +5.80. Aggressive updates strengthened coordination without inducing instability in this two-agent setting.
2. **Exploration rate** (ϵ): Low exploration (0.1) is essential for convergence. High exploration (0.3) prevents stable coordination, increasing welfare gap to +10.10. This represents a 74% increase in inefficiency relative to the best-performing configuration.
3. **Discount factor** (γ): Forward-looking behavior ($\gamma = 0.9$) provided no measurable advantage over myopic learning ($\gamma = 0.0$) in the two-agent case, with welfare gaps differing by only 0.10. This challenges conventional wisdom that $\gamma > 0$ is necessary for coordination in congestion games.

2.3.2 Convergence Dynamics

Figure 2.1 reveals distinct temporal patterns:

- **Baseline and Myopic**: Gradual monotonic increase in Nash-like percentage from 40% at episode 500 to 87% by episode 1,500, then stable
- **Fast Learning**: Rapid initial climb to 80% Nash-like by episode 400, reaching 92% by episode 800
- **High Exploration**: Persistent oscillation between 30–60% Nash-like throughout entire run, no trend toward stabilization

The lack of convergence in High Exploration is particularly notable, even after 2,000 episodes (sufficient for Baseline/Myopic to converge), the system showed no sign of settling. This suggests that excessive exploration creates a fundamentally different dynamical regime rather than merely slowing convergence.

2.3.3 Welfare Gap Decomposition

Table 2.4 decomposes the welfare gap into two components: structural inefficiency (time spent in non-Nash profiles) and within-Nash costs (exploration-induced deviations even when in Nash states).

Table 2.4. Welfare Gap Decomposition

Experiment	Non-Nash %	Structural Loss	Exploration Tax	Total Gap
Baseline	13.0	+4.5	+1.8	+6.3
Fast Learning	8.0	+3.2	+2.6	+5.8
High Exploration	51.0	+8.9	+1.2	+10.1
Myopic	12.0	+4.2	+2.0	+6.2

Structural Loss = (Non-Nash %) × (Avg cost in non-Nash profiles − Nash cost)

Exploration Tax = Remaining gap due to stochastic deviations within converged state

Interpretation: For converged experiments (Baseline, Fast, Myopic), 65–75% of the welfare gap stems from time spent in non-Nash profiles, with the remainder due to exploration-induced noise. For High Exploration, structural inefficiency dominates (88% of gap), reflecting fundamental coordination failure.

2.4 Robustness Checks

2.4.1 Sensitivity to Initial Conditions

We re-ran the Baseline experiment with three alternative initial states:

- $s_0 = (B, B)$: Converged to 85% Nash-like (vs. 87% from (A, A))
- $s_0 = (A, B)$: Converged to 89% Nash-like
- $s_0 = (B, A)$: Converged to 88% Nash-like

Final outcomes were statistically indistinguishable (within 4 percentage points), indicating that convergence is robust to initial conditions. Starting from a Nash equilibrium (A, B) or (B, A) provided slight advantage but did not fundamentally alter dynamics.

2.4.2 Sensitivity to Random Seed

Table 2.5 presents results from running Baseline with five different random seeds.

Table 2.5. Baseline Experiment Across Random Seeds

Seed	Nash-like %	Avg Cost	Welfare Gap
42	87.0	31.30	+6.30
17	84.0	31.65	+6.65
99	89.0	30.95	+5.95
123	86.0	31.40	+6.40
777	88.0	31.15	+6.15
Mean	86.8	31.29	+6.29
Std Dev	1.79	0.25	0.25

Conclusion: Results are highly stable across random seeds, with standard deviation of only 1.79 percentage points in Nash-like frequency. The welfare gap variation (±0.25) is negligible relative to the mean (+6.29), confirming that reported results are representative rather than artifacts of specific random sequences.

2.5 Summary of Key Findings

The two-agent Q-learning experiments yield four primary findings:

1. **Convergence to Nash is achievable but imperfect**: With appropriate parameters ($\alpha = 0.1$–0.3, $\epsilon = 0.1$, $\gamma \geq 0$), independent Q-learners converge to Nash-like behavior 87–92% of the time, sustaining welfare gaps of 19–21% due to residual exploration.
2. **Exploration is the critical parameter**: Low ϵ (0.1) enables convergence; high ϵ (0.3) prevents it entirely, increasing welfare loss by 74%. Learning rate and discount factor have secondary effects.

3. **Myopic learning suffices in dense state spaces**: The absence of forward-looking discounting ($\gamma = 0$) does not impair convergence in the two-agent case, suggesting that immediate cost feedback provides sufficient learning signal when agents frequently visit all relevant states.

4. **Fast learning accelerates coordination**: Higher learning rates ($\alpha = 0.3$) improve both convergence speed (2.5× faster) and final welfare (8% lower gap) without inducing oscillations, challenging concerns about instability from aggressive updates in small-scale multi-agent settings.

These findings establish the baseline for comparison when scaling to ten-agent populations, where we anticipate that state-space sparsity and increased coordination difficulty will qualitatively alter learning dynamics and potentially reverse the myopic learning result.