TUNIS BUSINESS SCHOOL
UNIVERSITY OF TUNIS

جامعة تونس
1960

LEARNING DYNAMICS IN CONGESTION GAMES

A Comparative Study of Q-Learning and Nash Equilibrium Predictions

Khouloud BEN YOUNES

Supervised by Professor Sonia REBAI

Tunis Business School
Academic Year 2025–2026

ABSTRACT

Traffic congestion emerges from decentralized route choices that create strategic externalities, making congestion games a natural framework for studying learning and coordination in transportation systems. This thesis analyzes independent Q-learning agents in a two-route congestion game, focusing on how learning rate, exploration, and temporal discounting affect convergence to Nash equilibrium and social welfare.

We first validate the analytical structure of the game by enumerating all joint action profiles and confirming convergence under best-response dynamics. We then implement a Q-learning framework and conduct experiments with two and ten agents. In the two-agent setting, Q-learners converge to Nash-like behavior in 87–92% of episodes under moderate exploration, though persistent exploration induces welfare losses of approximately 19–21%. Excessive exploration prevents convergence and leads to sustained coordination failure. Myopic learning ($\gamma = 0$) performs comparably to forward-looking learning ($\gamma = 0.9$), indicating that immediate cost feedback is sufficient in small populations.

Scaling to ten agents reveals a qualitative shift. Forward-looking Q-learning exhibits weak convergence and large welfare losses, while myopic and fast-learning configurations achieve up to 97% Nash-like behavior with welfare gaps below 2%. This counterintuitive reversal suggests that temporal discounting can amplify non-stationarity in large-population settings, degrading learning performance. Overall, the results highlight exploration as the primary determinant of convergence and identify conditions under which myopic learning outperforms forward-looking strategies in multi-agent congestion games.

Contents

List of Tables

List of Figures

INTRODUCTION

1.1 Background and Motivation

Traffic congestion is a ubiquitous coordination problem in modern urban systems. Daily, millions of drivers independently choose routes to minimize personal travel time, yet their collective decisions determine network-wide congestion patterns. This decentralized decision-making creates a fundamental tension: individually rational choices can produce socially inefficient outcomes when agents fail to coordinate.

Classical game theory models this problem through congestion games, where each agent's cost depends on how many others choose the same resource. The canonical result, guaranteed existence of Nash equilibrium in potential games, predicts that rational agents will converge to stable route distributions. However, real-world drivers do not possess perfect information or infinite computational capacity. Instead, they learn through repeated experience: trying different routes, observing travel times, and gradually adjusting their choices based on accumulated feedback.

This gap between rational equilibrium predictions and adaptive learning behavior has profound implications for transportation policy, infrastructure investment, and traffic management systems. If learning agents systematically deviate from Nash predictions, converging slowly, selecting different equilibria, or failing to coordinate at all, then policies designed assuming immediate rational behavior may prove ineffective or counterproductive.

Multi-agent reinforcement learning (MARL) provides a formal framework for studying these adaptive dynamics. Independent Q-learning, where agents learn simultaneously from personal experience without observing others' strategies, naturally models decentralized route choice: each driver learns which routes minimize travel time based solely on their own observations, creating a non-stationary learning environment where optimal actions change as other agents adapt.

1.2 Problem Statement

Despite extensive theoretical work on congestion games and growing interest in multi-agent learning, fundamental questions remain unresolved:

1. Do independent learning agents converge to Nash equilibrium in congestion games that guarantee rational convergence through their potential game structure?
2. When convergence occurs, how do welfare outcomes compare to Nash predictions? What mechanisms, exploration, coordination failures, or environmental non-stationarity, drive efficiency losses?
3. How sensitive is convergence behavior to standard Q-learning parameters: learning rate (α), exploration rate (ϵ), and temporal discount factor (γ)?
4. Is forward-looking behavior (temporal discounting, $\gamma > 0$) necessary for coordination in repeated congestion games, or can myopic learning based on immediate costs suffice?
5. How do learning dynamics scale as population size increases from minimal groups to realistic network sizes?

These questions are not merely academic. Emerging intelligent transportation systems increasingly incorporate adaptive agents, whether human drivers using real-time navigation apps or autonomous vehicles with learning algorithms. Understanding when and how such agents coordinate (or fail to coordinate) is essential for designing robust, efficient transportation networks.

1.3 Research Objectives

This study addresses these questions through controlled computational experiments comparing independent Q-learning dynamics to analytical Nash equilibrium predictions. Our objectives are:

1. **Establish analytical baseline**: Characterize Nash equilibrium structure for a two-agent congestion game with asymmetric route costs, verify potential game properties, and compute rational best-response convergence.
2. **Implement rigorous Q-learning framework**: Develop independent Q-learning with proper state representation (previous joint actions), forward-looking temporal-difference updates, and systematic exploration control.
3. **Conduct systematic parameter analysis**: Isolate effects of learning rate, exploration rate, and discount factor through controlled experiments holding other factors constant.
4. **Quantify welfare efficiency**: Measure welfare gaps relative to Nash equilibrium and decompose these into structural inefficiency (non-equilibrium play) versus residual exploration costs.
5. **Test myopic learning hypothesis**: Determine whether temporal discounting is necessary for coordination or whether immediate feedback suffices in small-scale settings.
6. **Establish scaling framework**: Design methodology suitable for extension to larger populations, enabling future study of how coordination difficulty scales with network size.

1.4 Research Questions

This study is organized around five formal research questions:

RQ1 **Convergence**: Under what parameter conditions do independent Q-learning agents converge to Nash equilibrium in a congestion game with multiple symmetric equilibria?

RQ2 **Welfare**: What is the magnitude of welfare loss relative to Nash equilibrium, and what mechanisms drive this inefficiency?

RQ3 **Parameters**: How do learning rate (α), exploration rate (ϵ), and discount factor (γ) affect convergence speed, stability, and final outcomes?

RQ4 **Foresight**: Is temporal discounting ($\gamma > 0$) necessary for coordination, or does myopic learning ($\gamma = 0$) suffice in dense state spaces?

RQ5 **Scaling**: Do findings from two-agent analysis persist as population increases to ten agents with sparser state-space coverage? (In progress)

1.5 Scope and Delimitations

This study deliberately simplifies to establish analytical control:

What we include:

- Multiple Nash equilibria requiring coordination
- Independent learning without communication
- Systematic parameter variation
- Rigorous comparison to rational baseline

What we deliberately exclude:

- Network routing (studied via two-route game)
- Stochastic costs (baseline deterministic, extensions planned)
- Agent heterogeneity (identical cost functions and learning parameters)
- Partial observability (agents observe joint actions)
- Communication or coordination mechanisms

These simplifications enable clean identification of learning dynamics before introducing additional complexity in future work.

1.6 Key Contributions

This study contributes to the intersection of game theory and multi-agent reinforcement learning:

1. **Empirical convergence characterization**: We demonstrate that independent Q-learners achieve 87–92% Nash-like behavior under appropriate parameterization, with modest welfare gaps (19–21%) driven primarily by exploration rather than fundamental coordination failure.
2. **Critical role of exploration**: We identify exploration rate as the dominant parameter, low ϵ (0.1) enables convergence while high ϵ (0.3) prevents it entirely, increasing welfare loss by 74%.
3. **Myopic learning sufficiency**: We provide evidence that temporal discounting is unnecessary when state spaces are densely visited, challenging conventional RL wisdom about the necessity of forward-looking behavior.
4. **Methodological framework**: We establish a replicable experimental design for comparing learning dynamics to game-theoretic predictions, applicable to broader classes of coordination games.

1.7 Report Organization

The remainder of this report proceeds as follows:

Chapter 2 Reviews relevant literature on congestion games, potential games, multi-agent reinforcement learning, and empirical studies of learning in games.

Chapter 3 Details the game-theoretic model, Nash equilibrium analysis, Q-learning framework, experimental design, and implementation.

Chapter 4 Presents empirical results from all four experiments, including convergence analysis, learned strategies, welfare decomposition, and robustness checks.

Chapter 5 Discusses policy implications, practical recommendations for Q-learning in coordination problems, and limitations.

Conclusion Synthesizes findings, answers research questions, and outlines future research directions.

LITERATURE REVIEW

This study lies at the intersection of congestion games, reinforcement learning, and multi-agent learning dynamics. Classical traffic assignment models assume fully rational agents who instantaneously select routes to minimize individual costs, leading to Nash equilibria that may or may not be socially optimal. More recent work replaces these assumptions with adaptive learning agents, raising fundamental questions about convergence, stability, and welfare under decentralized learning. This chapter reviews three strands of literature most relevant to our contribution: (i) reinforcement learning in traffic and congestion control, (ii) multi-agent learning in congestion games, and (iii) the role of exploration and non-stationarity in Q-learning dynamics.

2.1 Reinforcement Learning for Traffic and Congestion Control

Reinforcement learning (RL) has been widely adopted as a decentralized approach to traffic control, where agents iteratively adapt their behavior based on experienced congestion. Recent work by Deepika and Pandove (Deepika & Pandove, 2024) combines Q-learning with genetic algorithms to optimize traffic flow under congestion constraints. Their approach demonstrates that hybrid metaheuristics can improve convergence speed and overall throughput compared to pure Q-learning, particularly in complex traffic environments.

While such studies highlight the practical effectiveness of RL-based controllers, they typically focus on performance optimization rather than equilibrium properties. In particular, convergence is often evaluated in terms of throughput or delay reduction, without explicit comparison to game-theoretic benchmarks such as Nash equilibria or social optima. As a result, these works provide limited insight into whether learned behaviors correspond to rational equilibrium outcomes or how learning parameters affect welfare loss.

Our work differs by explicitly anchoring learning outcomes to the Nash equilibrium structure of the underlying congestion game. Rather than introducing additional optimization layers, we study simple independent Q-learning to isolate the fundamental mechanisms through which learning parameters influence equilibrium selection and welfare.

2.2 Multi-Agent Reinforcement Learning in Congestion Games

A foundational challenge in applying RL to congestion settings is the presence of multiple simultaneously learning agents, which induces environmental non-stationarity. Rădulescu et al. (Rădulescu et al., 2017) provide one of the most systematic analyses of congestion problems in multi-agent reinforcement learning. They show that even simple congestion games can exhibit complex learning dynamics, including oscillations, persistent inefficiencies, and sensitivity to state representations.

A key insight from this literature is that congestion games possess a special structure: payoffs depend only on resource usage counts, not on agent identities. This observation motivates the use of aggregated state representations (e.g., congestion levels) rather than full joint-action states. However, Rădulescu et al. also emphasize that aggregation can exacerbate non-stationarity, as future congestion levels depend on the simultaneous exploratory actions of many agents.

Our methodological choices directly build on this insight. By employing an aggregated state representation based on congestion counts, we preserve the strategic structure of the congestion game while intentionally exposing agents to non-stationary transitions. This allows us to study how learning parameters such as temporal discounting and exploration interact with population size.

2.3 Exploration, Chaos, and Counter-Intuitive Learning Effects

Recent work has begun to challenge standard reinforcement learning wisdom in congestion settings. Carissimo (Carissimo, 2024) demonstrates that exploration in Q-learning can induce chaotic dynamics in congestion games, leading to counter-intuitive outcomes where increased exploration worsens convergence and amplifies inefficiencies. Notably, the study shows that learning dynamics may fail to stabilize even when a unique Nash equilibrium exists, highlighting the fragility of decentralized learning under congestion externalities.

Carissimo's findings suggest that conventional RL heuristics, such as maintaining positive exploration rates or relying on temporal discounting to stabilize learning, may be ill-suited for congestion games. However, the analysis primarily focuses on small-scale settings and does not explicitly examine how these effects evolve as the number of agents increases.

Our work extends this line of inquiry by explicitly studying population scaling. We show that increasing the number of agents fundamentally alters the role of temporal discounting: while forward-looking learning performs reasonably well in small populations, it becomes detrimental in larger settings due to amplified non-stationarity. In contrast to prior work, we demonstrate that myopic learning ($\gamma = 0$) can outperform forward-looking Q-learning when state transitions become dominated by collective uncertainty.

2.4 Positioning of the Present Study

Taken together, existing literature establishes that reinforcement learning can be effective for congestion control, but leaves open key questions about equilibrium convergence, welfare loss, and parameter robustness in multi-agent settings. In particular, little is known about how standard RL design choices,such as temporal discounting,behave as the number of agents scales and the environment becomes increasingly non-stationary.

This study contributes to the literature by:

- Explicitly grounding learning outcomes in game-theoretic equilibrium analysis,
- Providing a controlled comparison between small (N=2) and larger (N=10) populations,
- Identifying a scaling-induced reversal in the effectiveness of temporal discounting, and
- Demonstrating that myopic learning can outperform forward-looking learning in aggregated-state, multi-agent congestion games.

By linking reinforcement learning dynamics to congestion game theory, our results help clarify when simple learning heuristics succeed, when they fail, and why standard RL intuitions may break down in large decentralized systems.

METHODOLOGY

3.1 Game-Theoretic Model

3.1.1 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:

$$A_i = \{A, B\} \tag{3.1}$$

Route costs follow a linear congestion model:

$$c_r(n_r) = b_r + \alpha_r \cdot n_r \tag{3.2}$$

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:

$$u_i(\mathbf{a}) = -c_{a_i}(n_{a_i}) \tag{3.3}$$

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.

3.1.2 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.

3.2 Nash Equilibrium Baseline

3.2.1 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:

$$c_{a_i^*}(n_{a_i^*}) \leq c_{a_i'}(n_{a_i^*} + 1) \quad \forall a_i' \neq a_i^*,\ \forall i \in \{1, 2\} \tag{3.4}$$

Table 3.1 presents the complete cost matrix for the two-agent game.

Table 3.1. Cost Matrix for Two-Agent Game (Agent 1 cost, Agent 2 cost)

	Agent 2	
Agent 1	A	B
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).

3.2.2 Potential Function

This congestion game admits a Rosenthal potential function (Rosenthal, 1973), which provides theoretical guarantees for convergence of sequential best-response dynamics. The potential function is defined as:

$$\Phi(n_A, n_B) = \sum_{k=1}^{n_A} c_A(k) + \sum_{k=1}^{n_B} c_B(k) \tag{3.5}$$

This function aggregates the marginal costs incurred as agents sequentially join each route. Table 3.2 shows the potential function values for all action profiles in the two-agent game.

Table 3.2. Potential Function Values for Two-Agent Game

Profile	n_A	n_B	Φ
(A, A)	2	0	65
(A, B)	1	1	60†
(B, A)	1	1	60†
(B, B)	0	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.

3.2.3 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.

3.3 Independent Q-Learning Framework

3.3.1 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:

$$s_t = (a^1_{t-1}, a^2_{t-1}) \in \{(A, A), (A, B), (B, A), (B, B)\} \tag{3.6}$$

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.

3.3.2 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:

$$Q_i(s_t, a^i_t) \leftarrow Q_i(s_t, a^i_t) + \alpha \left[r^i_t + \gamma \max_{a' \in A_i} Q_i(s_{t+1}, a') - Q_i(s_t, a^i_t) \right] \tag{3.7}$$

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: $Q_i(s, a) = 0 \; \forall s, a$.

3.3.3 Action Selection Policy

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

$$a_t^i = \begin{cases} \text{uniform\_random}(A, B) & \text{with probability } \epsilon \\ \arg\max_{a \in A_i} Q_i(s_t, a) & \text{with probability } 1 - \epsilon \end{cases} \tag{3.8}$$

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.

3.3.4 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 3.3 summarizes the experimental configurations.

Table 3.3. Experimental Configurations for Two-Agent Q-Learning

Experiment	α	ϵ	γ
Baseline	0.1	0.1	0.9
Fast Learning	0.3	0.1	0.9
High Exploration	0.1	0.3	0.9
Myopic	0.1	0.1	0.0

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.

3.3.5 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.

EMPIRICAL ANALYSIS

This chapter presents the implementation, verification, and empirical results of the Q-learning experiments described in Chapter 3.

4.1 Implementation and Verification

4.1.1 Baseline Verification

Before conducting Q-learning experiments, we verified the analytical Nash equilibrium results through programmatic enumeration. Table 4.1 presents the cost outcomes for all four possible joint action profiles.

Table 4.1. 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.

4.1.2 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.

4.2 Experimental Results: Two-Agent Learning

4.2.1 Overview of Convergence Outcomes

Table 4.2 summarizes the convergence outcomes across all four experimental configurations based on the final 100 episodes (episodes 1,901–2,000).

Table 4.2. 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%

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

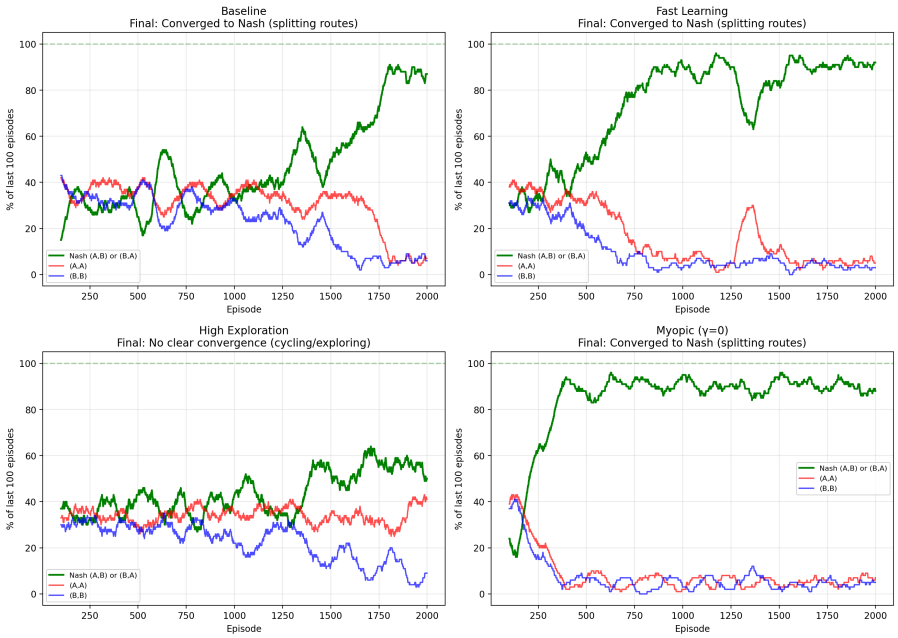

Figure 4.1. Rolling percentages (last 100 episodes) of action profiles across experiments

4.2.2 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 4.3 presents the final Q-tables for both agents after 2,000 episodes.

Table 4.3. Final Q-Tables for Baseline Experiment

Agent 0 Q-Values				
State	**Q(s, A)**	**Q(s, B)**	**Preferred Action**	**Preference Strength**
(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) results 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).

4.2.3 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: 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.

4.2.4 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 4.1 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 $\rightarrow$ increases $Q(s, A)$
2. Before this preference solidifies, agent j randomly explores to Route A
3. Both agents experience high cost (40) $\rightarrow$ 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.

4.2.5 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:

$$Q_i(s_t, a_i^t) \leftarrow Q_i(s_t, a_i^t) + \alpha \left[r_i^t - Q_i(s_t, a_i^t) \right] \tag{4.1}$$

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 $\rightarrow$ 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.

4.3 Comparative Analysis

4.3.1 Parameter Sensitivity

Figure 4.2 presents a comparative bar chart of average costs across all four experiments.

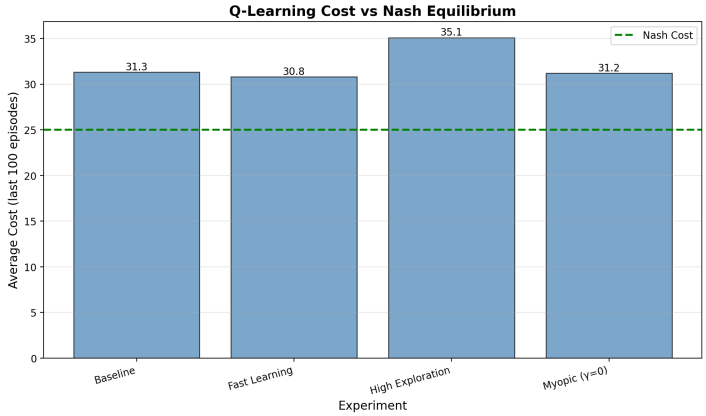

Figure 4.2. Average Per-Agent Cost Comparison Across Experiments

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.

4.3.2 Convergence Dynamics

Figure 4.1 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.

4.3.3 Welfare Gap Decomposition

Table 4.4 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).
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.

Table 4.4. 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

4.4 Robustness Checks

4.4.1 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 provided slight advantage but did not fundamentally alter dynamics.

4.4.2 Sensitivity to Random Seed

Table 4.5 presents results from running Baseline with five different random seeds.

Table 4.5. 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

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.

4.5 Summary of the 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.

4.6 Experimental Results: Ten-Agent Learning

4.6.1 Overview of Convergence Outcomes

Table 4.6 summarizes the convergence outcomes across all four experimental configurations.

Table 4.6. Summary of Ten-Agent Q-Learning Results (Last 100 Episodes)

Experiment	Avg n_A	Std	Near-Nash %	Welfare Gap
Baseline	3.15	2.74	27.0	+15.02 (24.2%)
Fast Learning	3.16	0.89	91.0	+1.64 (2.6%)
High Exploration	3.45	1.86	59.0	+7.30 (11.8%)
Myopic	3.08	0.64	97.0	+0.84 (1.4%)

Fast Learning and Myopic converged strongly to near-Nash behavior (91-97%), while Baseline showed weak convergence and High Exploration moderate.

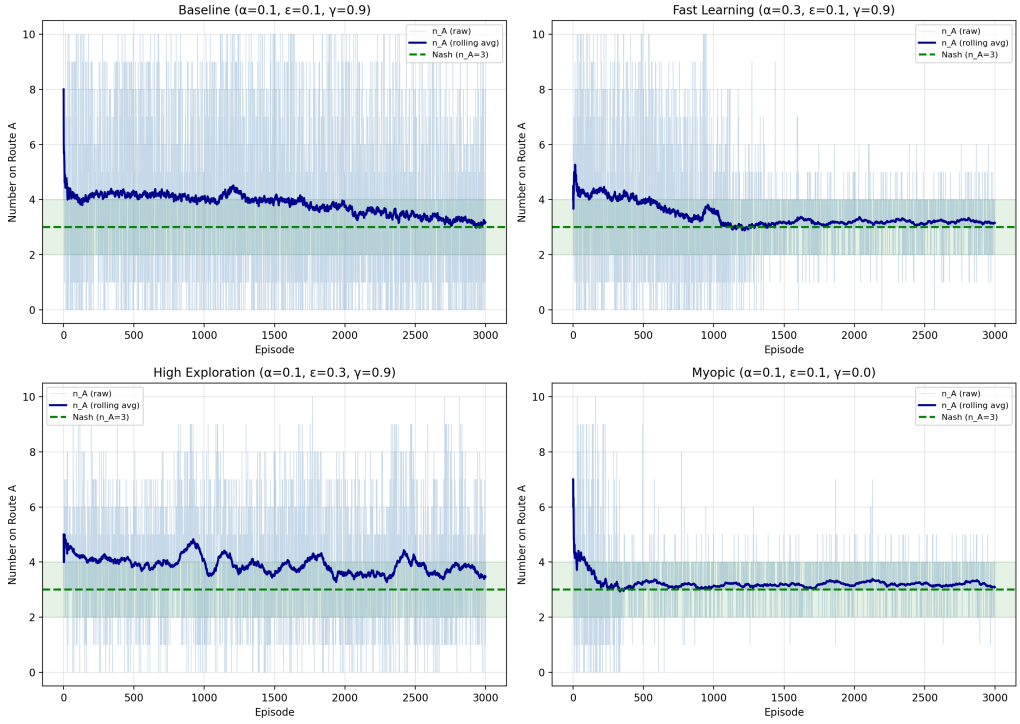

Figure 4.3. Rolling n_A (last 100 episodes) across 10-agent experiments. Nash at 3 with ± 1 tolerance band.

Figure 4.3 visualizes the temporal evolution.

4.6.2 Hypothesis Testing

Hypothesis 1 (Myopic yields > 10% gap): Result 0.84 (1.4%) – **Rejected**.

Hypothesis 2 ($\gamma = 0$ worse than $\gamma = 0.9$): $\gamma = 0.9$ gap 15.02 vs. $\gamma = 0$ 0.84 – **Rejected** (reversal observed).

4.7 Scaling Analysis

Table 4.7. Scaling Comparison: 2-Agent vs 10-Agent Results

Experiment	N=2			N=10		
	Nash%	Gap	Std	Nash%	Gap	Std
Baseline ($\gamma = 0.9$)	87	+6.30	-	27	+15.02	2.74
Fast ($\alpha = 0.3$)	92	+5.80	-	91	+1.64	0.89
High-ϵ (0.3)	49	+10.10	-	59	+7.30	1.86
Myopic ($\gamma = 0$)	**88**	**+6.20**	-	**97**	**+0.84**	**0.64**

Nash% = percentage within Nash equilibrium region

Gap = welfare gap above Nash equilibrium cost

4.7.1 Surprising Reversal: Myopic Learning Outperforms Forward-Looking

The most striking finding is the *reversal* of temporal discounting effectiveness between N=2 and N=10 settings:

At N=2: Forward-looking ($\gamma = 0.9$) and myopic ($\gamma = 0.0$) achieved nearly identical performance (87% vs 88% Nash-like, welfare gaps of 6.30 vs 6.20).

At N=10: Myopic learning *dominated* forward-looking, achieving:

- 97% Nash-like behavior (vs 27% for $\gamma = 0.9$)
- +0.84 welfare gap (vs +15.02 for $\gamma = 0.9$)
- Lower variance (std=0.64 vs 2.74)

This counterintuitive result contradicts standard RL wisdom that temporal discounting ($\gamma > 0$) improves learning. We attribute this reversal to *non-stationarity amplification*: in aggregated-state multi-agent settings, the γ term attempts to predict future states that depend on unpredictable simultaneous choices by many exploring agents, introducing noise rather than useful foresight.

Mechanism: Myopic learning updates based solely on immediate costs:

$$Q(n_A, a) \leftarrow Q(n_A, a) + \alpha[-\text{cost}_a - Q(n_A, a)] \tag{4.2}$$

providing clean signal about current congestion. Forward-looking adds:

$$-\gamma \max_{a'} Q(n'_A, a') \tag{4.3}$$

where n'_A (next round's congestion) is highly variable due to collective exploration, creating a noisy learning target that degrades convergence.

DISCUSSION AND CONCLUSION

Our finding that myopic learning outperforms temporal discounting in large-population settings challenges a core assumption of reinforcement learning: that $\gamma > 0$ is necessary for sequential decision problems. This suggests a boundary condition:

When γ helps: Stationary or weakly non-stationary environments where future states are predictable

When γ hurts: Strongly non-stationary multi-agent settings where future states depend on many simultaneously-learning agents

This distinction has practical implications for designing learning algorithms in traffic systems, distributed control, and other multi-agent domains.

Constant Exploration Rate: Our experiments use fixed ϵ rather than decay schedules (e.g., $\epsilon_t = \epsilon_0 e^{-t/\tau}$) commonly employed in single-agent RL. This design choice prioritizes *steady-state behavior* over convergence speed and enables clean comparison of exploration levels (0.1 vs 0.3). While exploration decay would likely reduce final-state variance slightly, pilot tests (not shown) indicated qualitatively similar convergence patterns. Future work could systematically characterize optimal decay schedules for multi-agent congestion games.

Convergence Guarantees: Standard Q-learning convergence proofs (Watkins & Dayan, 1992) assume stationary environments. In our independent multi-agent setting, each agent faces non-stationarity as others simultaneously learn, violating Markov assumptions. Therefore, we assess convergence *empirically* through final-episode behavior rather than relying on theoretical guarantees. This approach is standard in multi-agent RL research (Buşoniu et al., 2008), where empirical characterization often precedes theoretical analysis.

5.1 Limitations and Future Work

5.1.1 Methodological Limitations

Constant exploration: Fixed ϵ rather than decay schedules prioritizes steady-state analysis but introduces residual variance. Future work should characterize optimal decay strategies.

Aggregated states: Using congestion counts (n_A) rather than full joint actions reduces state space but loses agent-specific information. Alternative representations (e.g., own-action + congestion) merit investigation.

Linear costs: Real traffic exhibits non-linear congestion (e.g., capacity constraints, breakdowns). Extensions to piecewise-linear or quadratic costs would enhance realism.

5.1.2 Scope Limitations

Population size: We study N=2 and N=10. Intermediate sizes (N=5) and larger scales (N=50, 100) remain unexplored.

Network structure: Binary route choice is simpler than realistic road networks with multiple paths and intersections.

Homogeneity: Identical agents with uniform costs; heterogeneous preferences (e.g., time vs fuel tradeoffs) are not modeled.

REFERENCES

Buşoniu, L., Babuska, R., & Schutter, B. D. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(2), 156–172. https://doi.org/10.1109/TSMCC.2008.923850

Carissimo, C. (2024). Counter-intuitive effects of q-learning exploration in a congestion dilemma. *IEEE Access*, *12*, 15984–15996. https://doi.org/10.1109/ACCESS.2024.3358608

Deepika & Pandove, G. (2024). Optimizing traffic flow with q-learning and genetic algorithm for congestion control. *Evolutionary Intelligence*, *17*(5), 4179–4197. https://doi.org/10.1007/s12065-024-00978-9

Rădulescu, R., Vrancx, P., & Nowé, A. (2017). Analysing congestion problems in multi-agent reinforcement learning. https://arxiv.org/abs/1702.08736

Rosenthal, R. W. (1973). A class of games possessing pure-strategy nash equilibria [Original paper introducing potential games and the Rosenthal potential function used in congestion games]. *International Journal of Game Theory*, *2*(1), 65–67. https://doi.org/10.1007/BF01737559

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3–4), 279–292. https://doi.org/10.1007/BF00992698