Adaptive Pricing and Learning in the Multi-Market Routing Problem

Behrad Koohy[a,*], **Vahid Yazdanpanah**[a], **Sebastian Stein**[a] **and Enrico Gerding**[a]

[a]Southampton University, Southampton, United Kingdom
ORCID (Behrad Koohy): https://orcid.org/0000-0002-8115-2887, ORCID (Vahid Yazdanpanah): https://orcid.org/0000-0002-4468-6193, ORCID (Sebastian Stein): https://orcid.org/0000-0003-2858-8857, ORCID (Enrico Gerding): https://orcid.org/0000-0001-7200-552X

Abstract. In modern urban transportation networks, multiple self-interested travel providers (public transit, micromobility providers, ride-sharing platforms and toll roads) compete for heterogenous transportation users that wish to balance time and cost. Traditional congestion models assume fixed, exogenous costs, while dynamic-pricing frameworks typically focus on a single operator, overlooking the rich strategic interplay among decentralised transportation providers. This paper introduces the Multi-Market Routing Problem (MMRP), a game-theoretic model in which each provider utilises adaptive pricing to maximise profit and heterogeneous transportation users aim to minimise their travel time and cost.

We present the MMRP as an extension of traditional congestion games, and extend it to consider online instances for adaptive pricing under dynamic and stochastic congestion. We demonstrate the computational complexity for game-theoretic and exact solutions to the MMRP, reflecting the computational complexity of coordinating routing in dynamic and uncertain settings. To address this, we propose the use of independent Proximal Policy Optimisation as a decentralised and effective solution to the online MMRP, demonstrating reduced travel times and more equitable and fair outcomes for transportation users, and increased profitability for transportation providers. The MMRP framework and learning algorithms offer a principled foundation for competitive, multimodal routing in modern urban transportation networks.

1 Introduction

Urban transportation networks closely resemble multi-provider marketplaces, where transportation users can select from a variety of available transportation options, such as public transit, micro-mobility services, toll roads, and ride-sharing platforms, to traverse the network to their destination. Transportation users vary in their sensitivity to time and cost, leading to heterogeneous route-choice behaviour that classical congestion models, such as Congestion Games (CGs) [21], which assume fixed exogenous costs and homogeneous users, cannot capture accurately. Moreover, traditional frameworks assume a passive cost structure that scales only with aggregate flow, neglecting the fact that each provider could strategically set its own pricing to influence demand and incentivise the use of their transportation method. Therefore, existing CG models fail to consider the competitive interdependencies among providers that characterise real-world, multimodal transportation networks.

Furthermore, CGs model a setting in which non-atomic players choose resources such as network routes and incur costs dependent on aggregate usage [21]. While this framework guarantees the existence of pure-strategy equilibria under fixed, exogenous cost functions, it assumes homogeneous users and passive cost structures that scale only with flow. Extensions such as atomic and weighted congestion games [13, 1] accommodate player heterogeneity and multiple origin-destination pairs but still treat delays or tolls as a predetermined, monotonic function of the total flow rather than a decision variable controlled by a strategic provider. Bilevel and Stackelberg [2] models introduce tolling but assume a single, monopolistic paradigm, where route prices are determined by a single centralised operator. However, in real-world transportation systems, tolls are set by multiple autonomous operators with limited information and without a central coordinator, so bilevel and Stackelberg formulations cannot capture the asynchronous, strategic interplay and information constraints inherent in truly decentralised and competitive markets. Real-time dynamic tolling frameworks [30, 10, 24, 14, 29] and multiagent reinforcement learning approaches capture online pricing under uncertainty but focus on a centralised pricing mechanism or assume communication across pricing agents.

To address these gaps, we introduce the Multi-Market Routing Problem (MMRP), a novel game-theoretic framework in which each route provider is a strategic agent utilising tolling to maximise its own profit while a heterogenous population of transportation users (or players) travel through the network, selecting routes based on delay, toll and individual value of time. The MMRP is defined as both an offline problem that combines route pricing strategy and delay function, and player strategy spaces and preferences, and an online problem which extends the problem to include decision-making under uncertainty, allowing routes to adapt to evolving congestion and competitive behaviour through adaptive tolling. To this end, this paper provides the following contributions:

1. We rigorously define the MMRP as both an online and offline problem, and provide a game-theoretic framework for modelling the MMRP to understand the strategic interactions and competitive dynamics, including a proof of its NP-Hard nature in the problem's offline form.
2. As there exists an inherent complexity associated with developing efficient algorithms for solving the offline MMRP, we evaluate

* Corresponding Author. Email: behrad.koohy@soton.ac.uk.

heuristic-based solutions for the offline MMRP, utilising offline solutions as a baseline for the online MMRP.
3. We proposed a modified variant of Proximal Policy Optimisation, adapted for use in competitive multiagent scenarios to ensure reliable learning in dynamic and uncertain environments as a learning-based solution for the online MMRP.
4. We conduct extensive empirical evaluations encompassing both offline and online instances of the MMRP, demonstrating improved travel time and equality outcomes for players and profitability for route providers, suggesting that our PPO-based method effectively manages congestion and competition between providers

2 The Multi-Market Routing Problem

Classical CGs [21] model a setting where non-atomic users independently select routes to minimise travel delay, reaching an equilibria in which no-one can unilaterally improve their outcome. In contrast, the Multi-Market Routing Problem elevates each route to the status of a strategic transportation provider: every provider sets tolls or prices to maximise their own profit, while a heterogeneous population of transportation users choose routes based on the combined delay and price. We begin by formalising the MMRP as a six-tuple that brings together provider pricing strategies, user strategy spaces, flow-dependent delay functions, and value-of-time heterogeneity.

Following Roughgarden and Tardos [22]'s definition of non-atomic congestion games, we formally define the Multi-Market Routing Problem (MMRP) as a 6-tuple:

Definition 1. *An instance of the MMRP M is defined by the 6-tuple:* $M = (R, V, (\Phi_j)_{j\in V}, (\Theta_j)_{j\in V}, (D_i)_{i\in R}, (\Omega_i)_{i\in R})$

The set $R = \{R_1, R_2, \ldots, R_N\}$ is the set of available routes in the network. The set $V = \{V_1, V_2, \ldots, V_M\}$ represents the set of players or transportation users in the game. In contexts where a network graph is available, each route R_i corresponds to a connected link in the network. For each route $R_i \in R$, the delay function $D_i : \{0, \ldots, j\} \to \mathbb{R}^+$ assigns a travel delay based on the number of players on the route. That is, if f_i players select route R_i, the travel time for R_i is $D_i(f_i)$ [1]. We define delay functions as non-negative, continuous and non-decreasing. For each route R_i, there is a toll function $\Omega_i : \mathcal{C}_i \to \mathbb{R}$, where $\mathcal{C}_i$ denotes the conditions relevant to the route (e.g. current flow, historical toll values, or other dynamic factors). In many practical settings, this toll may evolve over time, and consider temporal factors as well as spatial ones (i.e. $\Omega_i(t) = g_i\big(\mathcal{C}_i, \Omega_i(t-1)\big)$). The toll function Ω_i allows the routes to incorporate pricing strategies that could be used to manage congestion whilst generating revenue, and its dynamic form is particularly amenable to time-dependent analyses.

Each player V_j is associated with an origin–destination (OD) pair (o_j, d_j); this specifies the starting point and desired destination. Θ_j denotes the Value of Time (VoT) of player V_j, a value capturing the heterogeneity in players and their amenability to incur a higher toll for arriving at their destination sooner. For every player V_j, the strategy space is defined as $\Phi_j \subseteq R'(o_j, d_j)$, where $R'(o_j, d_j)$ is the set of routes connecting player V_j's origin o_j to destination d_j. We define $x = (x_1, x_2, \ldots, x_M)$ as a strategy profile such that $x_j \in \Phi_j$ is the route chosen by player V_j. The induced flow on any route $R_i \in R$ is then given by $f_i = \#\{j \mid x_j = R_i\}$. We define the utility of a route R_i for player V_j as a combination of the travel time of the route and the associated price of the route. Formally, $u_{R_i,V_j} = \Omega_i(\mathcal{C}_i) + \Theta_j D_i(f_i)$.

In this formulation of the MMRP, each problem instance requires a pricing strategy for every available route as input. Each route provider utilises the respective pricing strategy, and the objective is to select the combination of pricing strategies that maximises the total profit across all routes. However, profit maximisation alone does not capture the full picture of network performance. In addition to route profit, it is crucial to consider the travel time experienced by players, the price they pay, and the fairness and equality in costs among them. Consequently, the best solutions are those that not only optimise the routes' profit margins but also balance efficiency and equity for the players, leading to improved overall network performance.

The MMRP extends classical non-atomic congestion games by incorporating competition amongst routes, OD-specific strategy spaces, heterogeneity in players and dynamic toll mechanisms. In real-world scenarios, however, network conditions and demand patterns evolve over time. Consequently, tolls and travel delays cannot be assumed to remain constant; they respond dynamically to congestion levels and players' route selections. To capture this temporal evolution and allow for adaptive tolling strategies, we extend the MMRP, as described by Definition 1 from an offline framework to an online one. In this dynamic setting, toll decisions are made repeatedly in real time, reflecting the instantaneous state of the network and the current distribution of traffic flows.

3 The Online MMRP and Definition as a Markov Decision Process

Building on the definition of the MMRP as an offline problem, we extend this definition to capture the evolving and stochastic nature of real-world transportation networks. Extending Definition 1 to a 7-tuple OM captures the dynamic evolution of the transportation network by incorporating a state space that reflects real-time toll levels and traffic flows.

Definition 2. *We define an online instance of the MMRP OM as* $\mathcal{OM} = (R, V, (\Phi_j)_{j\in V}, (\Theta_j)_{j\in V}, (\Lambda_j)_{j\in V}, (D_i)_{i\in R}, (\Omega_i)_{i\in R})$

The variables (R, V, Φ) from the offline 6-tuple definition are not changed. In the online definition, Λ_i represents the entry time of a player V_i. The delay and toll functions become time-dependent strategies, denoted $D_i(x, t)$ and $\Omega_i(x, t)$, which allow the model to capture the network dynamics as traffic flows and toll decisions evolve over time.

Building on this extended framework, we illustrate the granular decisions within the online MMRP by defining the problem as a Markov Decision Process (MDP). This definition isolates a single route within the MMRP and models how its state evolves over time in response to toll adjustments and subsequent player behaviour.

Definition 3. *We use an MDP $\mathcal{M}_i$ to represent the environment in which a sequential decision-making agent is operating for route R_i as* $\mathcal{M}_i = \langle S_i, A_i, P_i, Re_i, \gamma \rangle$

Here, each state $s_{i,t} \in S_i$ at time t is a tuple $s_{i,t} = (\Omega_i(t),\ f_i(t))$, where $\Omega_i(t)$ is the toll level on route R_i at time t and $f_i(t)$ is the corresponding traffic flow. The action $a_{i,t} \in A_i$ available to route R_i is an adjustment $\Delta\Omega_i(t)$ to the current toll, and the set A_i includes all feasible toll adjustments. The new route toll is updated via $\Omega_i(t+1) = \Omega_i(t) + \Delta\Omega_i(t)$. The state transition probability $P_i(s_{i,t+1} \mid s_{i,t}, a_{i,t})$ captures how the local state evolves in

[1] Examples of delay functions include volume delay functions [25, 17].

response to a toll adjustment. This evolution reflects both the deterministic update of the toll and the stochastic response of players. The reward function Re_i encourages each route to maximise their profit per timestep: $R_i(s_{i,t}, a_{i,t}, s_{i,t+1}) = \Omega_i(t+1)\, f_i(t+1)$ where $\Omega_i(t+1)\, f_i(t+1)$ represents the total revenue collected on route R_i at time $t+1$. As each route provider acts as a self-interested agent focused solely on maximising its own profit. Since players are more likely to choose a route with reduced travel time, a provider can improve its performance by minimising delays, thereby attracting more players. This increased throughput directly boosts the provider's revenue as reducing travel time serves as a competitive advantage that ultimately maximises profit through higher player counts on the route. The scalar $\gamma \in [0, 1]$ is the discount factor that weights future rewards relative to immediate ones.

Within the MMRP, each route independently determines its pricing strategy based solely on its local state metrics without direct coordination with competing providers. Although the overall system is multiagent, modeling each route provider's decision problem as an individual MDP is appropriate because the competitive effects (such as shifts in player flows resulting from competitors' toll changes) are encapsulated in the observed environmental dynamics. This approach allows each provider to perceive the problem as a sequential, single-agent decision task where the impact of other agents is reflected in the state transitions. We can apply single-agent reinforcement learning techniques to optimise pricing strategies, reducing the computational complexity associated with joint multiagent formulations like Dec-POMDPs or stochastic games. This independent MDP formulation preserves the decentralized decision-making and competitive interactions, while remaining tractable and directly focused on the optimisation of local pricing strategies.

The decentralised MDP formulation not only illustrates how each route provider optimises its pricing strategy based on local dynamics but also captures the competitive interactions indirectly through the observed state transitions. The broader MMRP framework lays the groundwork for an exploration of exact solutions to the MMRP to tackle the complexity inherent in decentralised, multiagent environments.

4 Feasibility of Game-Theoretic Solutions

Through existing analysis of congestion games, the formal representation of the MMRP, we consider game theoretic approaches to understand strategic competition among route providers. In this analysis, we define an instance of the 6-tuple MMRP with a simple two-route instance analogous to a parallel two-link network from the literature [28, 13].

We define $R = \{R_1, R_2\}$, where D_i is defined by the Bureau of Public Roads volume delay function [20]:

$$D_i(x) = \mathrm{F}_0 \cdot (1 + \alpha(\frac{x}{\mathrm{F}_c})^{\beta})$$

where x is the number of vehicles on the road at timestep t, F_0 is the free flow travel time of the route, F_c is the route capacity and α, β are calibration parameters to align the function behaviour to real-world road characteristics. In our abstract example, we align our example with calibrated values from literature of $\alpha = 0.68, \beta = 2.73$ [17].

In this example, our population of players, V, are defined with a strategy space $\{\Phi_j\}_{V_j \in V}$ as a probability distribution over the possible routes $R_i \in R$. The probability that a player V_j will take a route R_i is determined through Quantal Response Equilibria (QRE) [12], a framework for modelling bounded rationality in strategic interactions that captures the probabilistic nature of human decision-making and accounts for the possibility that players may not always play their exact best responses. Formally, we define QRE as

$$\Phi(R_i) = \frac{\exp(\lambda u_{R_i,V_j})}{\sum_{R' \in R} \exp(\lambda u'_{R_i,V_j})} \tag{1}$$

where $u_j(R_i)$ is the utility of taking route R_i for player v_j at timestep t. For our theoretical and empirical analysis, we used a value of $\lambda = 1$ for our application of QRE.

To model the heterogeneity of player preferences, we assume player VoT, Θ_j, is drawn from the uniform distribution $\mathbb{U}(0, 1)$. To align our player arrival distribution with real-world conditions, we model Λ using a Beta distribution [16]. This distribution introduces variability in congestion levels throughout the simulation, similar to traffic peaks observed in real-world traffic systems.

We adopt three assumptions for R_i in this instance of the MMRP. Firstly, the pricing function for each route remains fixed each game, $\Omega_i(\cdot) = p_i$ where p_i is set at the start of the simulation. Secondly, routes have infinite capacity (travel time is constant regardless of users). Formally, we assume that for this instance of the MMRP, for $(\mathrm{F}_0, \mathrm{F}_c)$, we set R_1 as $(15, \infty)$ and R_2 as $(30, \infty)$. Thirdly, the total number of players is known in advance. These conditions allow us to identify a Pure Nash Equilibrium (PNE), where no route can unilaterally increase its profit by altering prices.

Table 1. Pure Nash Equilibrium in the Offline MMRP with Infinite Route Capacity.

	MMRP-PPO			**Nash Equilibrium**		
Players	p_1	p_2	**Profit**	p_1	p_2	**Profit**
500	17.66	10.56	6777.92	11.00	6.00	4642.60
600	12.66	7.46	6049.10	11.00	6.00	5571.13
750	12.47	7.62	7399.64	11.00	6.00	6963.91
900	13.08	7.84	9324.62	11.00	6.00	8356.70
1000	12.10	7.40	9588.02	11.00	6.00	9285.22

Table 1 presents the pure-strategy PNE for varying player populations $|V|$, illustrating that there exists a pair of tolls (p_1, p_2) at which neither route provider can unilaterally increase its revenue. However, this equilibrium relies on infinite route capacities. Under these assumptions, finding a PNE in the MMRP becomes a tractable problem as the interactions between route providers are simplified. Furthermore, the fixed pricing paradigm ensures that the number of calculations is limited.

Introducing finite capacities creates discontinuities in the best-response mappings due to nonlinear congestion effects, which destroy the underlying potential-game structure guaranteed by Rosenthal's theorem [21]. Furthermore, allowing providers to adjust tolls in a continuous, dynamic fashion introduces further complexities, yielding a non-cooperative game with no general equilibrium guarantees. Together, these challenges highlight the limitations of exact, static solution methods and motivate the development of adaptive, learning-based approaches for dynamic, competitive routing environments.

5 Computational Complexity

Given the dynamic and real-time decision-making requirements of the MMRP, it is essential to understand the computational challenges posed by the offline version, where all players and routes are known in advance. Where in Section 4 we assumed that the route prices are fixed for the instance of the MMRP and found the existence of

Table 2. PSO and PPO results for the online and offline MMRP at variable congestion levels.

		Number of Players	500	600	650	700	750	800	900	1000
Offline MMRP	**Constant**	Travel Time	20.93	21.75	22.53	24.04	30.66	56.98	404.42	1244.19
		Profit	53.55	62.71	60.16	83.18	61.15	66.81	50.01	54.37
	Flow-Linear	Travel Time	20.72	21.42	22.04	23.07	27.74	49.32	353.29	1168.15
		Profit	50.24	47.13	47.41	43.67	55.32	57.31	74.54	89.84
	Time-Index	Travel Time	21.09	21.87	22.74	24.44	30.50	58.03	404.9	1247.95
		Profit	66.75	66.91	70.11	75.34	74.85	60.90	58.44	58.85
Online MMRP	**MMRP-PPO**	Travel Time	26.66	30.85	36.24	51.58	117.1	255.65	469.12	1739.88
		Gini Coef.	0.14	0.14	0.17	0.26	0.33	0.25	0.18	0.13
		Profit	13.38	24.65	36.67	47.54	66.04	77.91	86.78	90.41
	PPO	Travel Time	30.45	35.18	38.34	66.95	108.31	257.13	791.88	1863.87
		Gini Coef.	0.15	0.14	0.17	0.44	0.44	0.33	0.25	0.15
		Profit	34.08	29.95	29.95	31.87	30.16	30.16	39.88	40.94
	Random	Travel Time	32.50	34.87	34.87	40.14	151.83	291.61	553.88	2083.49
		Gini Coef.	0.18	0.15	0.18	0.39	0.44	0.37	0.20	0.14
		Profit	46.13	50.80	46.64	49.61	47.28	49.94	48.64	46.55

a PNE, relaxing this constraint to allow adaptive pricing functions leads to increased complexity in finding a PNE.

The theoretical confirmation of the offline MMRP's NP-Hardness serves as a foundation, informing our algorithmic strategies for both offline and online scenarios, and justifies the necessity for the use of heuristic and learning-based approaches to address offline and online instances of the MMRP.

Definition 4. *We cast the MMRP as defined in Definition 1 to a decision problem, MMRP-Decision, where:*

MMRP-Decision

Input: An instance of the offline MMRP M over two routes ($R = (R_1, R_2)$) and a target revenue $K \in \mathbb{B}$.

Question: Does there exist $\{\Omega_i\}_{i \in R}$ such that:

1. $\{\Omega_i\}_{i \in R}$ is a Nash equilibrium where both routes have strictly positive flow (i.e. neither route encounters no demand).
2. The total profit $\Omega_1 f_1 + \Omega_2 f_2 \geq K$?

Theorem 1. *MMRP-Decision is NP-Hard.*

Proof. We reduce from the NP-Complete Partition problem to an instance of the MMRP:

Partition

Input: A multiset of positive integers $\{a_1, \ldots, a_n\}$ with total sum $S = \sum_j a_j$.

Question: Does there exist an equal-sum subset $I \subseteq \{1, \ldots, n\}$ such that $\sum_{j \in I} a_j = \frac{S}{2}$

Given an instance of $\{a_j\}$ of Partition, we construct a polynomial time reduction to an instance of the MMRP with two routes $R = \{R_1, R_2\}$ and S players. The set V has $|V| = \sum_{j=1}^{n} a_j = S$, where for each integer a_j, we introduce a group of exactly a_j identical players to V with $\Theta_j = 1$. We define our delay functions to be equal, $D_1(f) = D_2(f) = 0, \forall f$, and therefore player route choices is wholly and entirely dependent on pricing strategies. We define the pricing function for each route to be constrained within $\{0, 1, \ldots, S\}$, and our total profit $K = \frac{S^2}{2}$.

In the instance where Partition has an equal-sum subset, there is $I \subseteq \{1, \ldots, n\}$ with $\sum_{j \in I} a_j = S/2$. We set $\Omega_1 = \Omega_2 = S/2$. Then at equilibrium exactly $S/2$ players go on each route, and each route collects profit of $(S/2) \times (S/2) = S^2/4$. Hence, total profit $= 2 \cdot (S^2/4) = S^2/2 = K$. This is only feasible if the instance of Partition has an equal-sum subset.

In the instance where total profit $\geq K$ is achievable in the MMRP, any toll assignment (Ω_1, Ω_2) with $\Omega_1 \neq \Omega_2$ causes all players to choose the cheaper route, yielding revenue $\max(\Omega_1, \Omega_2) \times S$. Since $\max(\Omega_i) \leq S$, this profit is at most S^2, but for any strict inequality $\Omega_{\max} > S/2$ the profit $S \cdot \Omega_{\max} > S^2/2$, which exceeds $S^2/2$. The only way to exactly achieve total profit in $[K, S^2/2]$ and have an equilibrium split is to set $\Omega_1 = \Omega_2 = S/2$ and have exactly half the players on each route. This even split in turn corresponds to a choice of which $S/2$ players go on route 1, which is a partition of the a_j into two subsets each summing to $S/2$.

Therefore, **the computation of whether the total profit $\geq K$ is attainable is equivalent to solving Partition.** As Partition is NP-Complete, **MMRP-Decision is NP-Hard**. □

The reduction from Partition to MMRP shows that even in a highly simplified setting, selecting profit-maximising pricing strategies encapsulates the NP-Complete Partition problem. Therefore, computing exact equilibrium pricing strategies for the MMRP is intractable in the general case. To address more complex and stochastic instances of the MMRP that present in real-world settings, we must utilise scalable, approximation-based techniques to produce effective pricing strategies in realistic and large-scale environments.

6 Particle Swarm Optimisation for Offline MMRP

Given the computational complexity of finding exact revenue-maximising tolls under user equilibrium in instances of the MMRP with infinite capacity, we turn to heuristic optimisation to obtain effective solutions for the offline MMRP with constrained capacity. For this, Particle Swarm Optimisation (PSO) [7] is utilised as it explores complex and non-convex search spaces with few hyperparameters. Relaxing assumptions of infinite capacity and static tolls better reflects real-world conditions, as dynamic toll-setting heightens competition by forcing providers to continually adapt their prices in response to congestion and competitor actions. Formally, for $(\mathrm{F}_0, \mathrm{F}_c)$ of our volume-delay functions, we set R_1 as $(15, 20)$ and R_2 as $(30, 20)$. We evaluated three families of static pricing strategies, each differing in expressiveness and complexity of search space:

1. **Constant Toll**: $\Omega_i(f) = p_i$, with a single parameter p_i per route. This policy serves as a baseline and has the least complex search space.
2. **Flow-Linear Toll**: $\Omega_i(f) = \alpha_i + \beta_i f_i$, introducing two parameters (α_i, β_i) per route. This policy provides adaptive tolls in proportion to congestion levels while maintaining a low-dimensional search space.

3. **Time-Index Toll:** Each route R_i has an initial price $p_{i,0}$ and a sequence of timestep deltas $\{\delta_{i,1}, \ldots, \delta_{i,T}\}$. Prices evolve via $p_{i,t} = p_{i,t-1} + \delta_{i,t}$, and the policy has $T+1$ parameters per route. This class offers the greatest flexibility, allowing for distinct and dynamic pricing profiles, but at the cost of a high dimensional search for optimal solutions.

For each pricing policy, the mean travel time is used as the fitness metric. Each pricing policy was evaluated over 50 randomised instances, where player count and value-of-time distributions were sampled to reflect realistic demand variability. In this evaluation, we recorded the average travel time and average profit per player. Table 2 shows the mean travel time and profit per player, revealing a modest improvement for adaptive approaches (flow-linear and time-index) compared to constant pricing.

7 Empirical Methodology for the Online MMRP

Despite the insight provided by the offline MMRP and PSO-based adaptive pricing solutions, their assumptions of full knowledge of future demand fall short of real-world variability and stochasticity.

In this instance of the online MMRP, we use Definition 3 of the MDP of the MMRP to define our environment. For each route R_i, we model each route provider's dynamic pricing problem as $\mathcal{M}_i = \langle S_i, A_i, P_i, Re_i, \gamma \rangle$. At each timestep t, $s_{i,t} \in S_i$, agent i's observation, is a 12-dimensional vector $s_{i,t} = \big(n_i(t),\ p_i(t),\ a_i(t),\ \tau_i(t),\ n_{-i}(t),\ p_{-i}(t),\ \tau_{-i}(t),\ t,\ a_i(t-1),\ a_{-i}(t-1),\ V_{\text{rem}}(t),\ V_{\text{on}}(t)\big)$ where:

- $n_i(t)$: number of players queued on route i,
- $p_i(t)$: current price on route i,
- $a_i(t)$: number of new arrivals to route i at t,
- $\tau_i(t)$: current travel time on route i,
- $n_{-i}(t)$, $p_{-i}(t)$, $\tau_{-i}(t)$: same for the competing route(s),
- $a_i(t-1)$, $a_{-i}(t-1)$: the previous pricing behaviour of agents,
- $V_{\text{rem}}(t)$: number of players still to arrive,
- $V_{\text{on}}(t)$: total number of players currently in-system (queued or en-route).

Our routes have the action space $A_i = \{-1, 0, +1\}$ so that $\Omega_i(t+1) = \Omega(t) + A_{i,t}$. The next state $s_{i,t}$ is generated by our simulation environment, which given the current route prices, previous actions and new arrivals into the environment, updates the queue lengths, travel times and player counts for all routes. Algorithm 1 provides a high-level simulation loop used during PPO training. At each timestep, agents construct their local observations, sample their policy for an action, and the environment then progresses through the timestep to simulate the impacts of the chosen actions.

Each route maintains the same reward function as defined in Definition 3. We set our discount factor as $\gamma = 0.99$.

For evaluation of adaptive pricing solutions for the online MMRP, we utilise two additional complex environments (coupled with our two-route environment introduced in Section 6), using volume-delay functions with parameters calibrated from real-world measurements in London, United Kingdom [3].

For Scenario 2, we focus on a five-route network defined in Table 3, with $|V|$ ranging over $\{5000, 6000, \ldots, 9000\}$. In this scenario, vehicle arrival times are sampled from a uniform distribution over the simulation horizon, ensuring consistent temporal randomness and allowing us to evaluate policy robustness under steady but unpredictable demand patterns. In Scenario 3, we utilise a subsection of the Sioux Falls network [27], specifically nodes 10, 11, 14, 15, 16, 17, and 19. We model demand in this scenario from the provided trips and network data provided.

Table 3. Selected VDF parameters from Casey et al. [3].

Route	α	β	f_c	f_0
1	1	2	223.00	25.38
2	1	2	309.80	32.81
3	1	2	276.90	44.05
4	1	2	314.30	40.83
5	1	2	358.80	44.05

Our chosen three empirical evaluation settings are such that we can understand the MMRP under escalating levels of complexity and realism. The two-route synthetic parallel-link benchmark, referred to as Scenario 1, isolates the competitive element of the problem to two agents, allowing us to compare our chosen solution with the calculated equilibrium tolls in an analytically tractable context. The five-route parallel benchmark, referred to as Scenario 2, introduces multiple providers and a broader set of choices for players, using calibrated volume-delay functions to rest policy stability and generalisation as the non-stationarity in the problem increases. Finally, the Sioux Falls subnetwork applies our solution in a realistic urban topology with heterogenous origin-destination flows and varying link capacities within a network where agents' incoming observations are impacted by the chosen policy of other agents.

Algorithm 1 Environment Loop for Online MMRP

Input
  $\mathcal{OM} = (R, V, \Phi_j, \Theta_j, \Lambda_j, D_i, \Omega_i)$
  where $i \in R$, $j \in V$
for each route i **do**
  $toll[i] \leftarrow initial\_price$
  $queue[i] \leftarrow 0$
$vehicles\_rem \leftarrow total\_vehicles$
for $t = 1$ to $T_{\max}$ **do**
  for $R_i \in R$ **do**
    Observe state for each route i:
      $obs[i] \leftarrow$ get_observation(i)
    Agents select actions:
      $action[i] \leftarrow \pi_i(obs[i]) \in \{-1, 0, +1\}$
    Update tolls:
      $toll[i] \leftarrow toll[i] + action[i]$
    Sample arrivals & update remaining vehicles:
      $arrivals \leftarrow$ sample_arrivals()
      $vehicles\_rem \leftarrow vehicles\_rem - \sum arrivals$
    Update queues & compute departures:
      $queue[i] \leftarrow queue[i] + arrivals[i]$
      $departed[i] \leftarrow$ compute_departures$(queue[i])$
      $queue[i] \leftarrow queue[i] - departed[i]$
    Update travel times via VDF:
      $travel\_time[i] \leftarrow$ VDF$(queue[i])$
    Compute rewards (revenue):
      $reward[i] \leftarrow toll[i] \times departed[i]$
    Observe next state & store transition:
      $next\_obs[i] \leftarrow$ (as in step 4 with updated variables)
      store_transition$(obs[i], action[i], reward[i], next_obs[i])$

Table 4. Mean travel time, social cost and combined cost for MMRP-PPO and Random agents in Scenario 2. We define the social cost as the $\Phi_j \cdot D_i(f_i)$ (player j VoT multiplied by the travel time of the route) and the combined cost as $\Omega_i(\mathcal{C}_i) + \Theta_j D_i(f_i)$, the specific utility of the chosen route for the player.

	MMRP-PPO			Random		
V	**Travel Time**	**Social Cost**	**Combined Cost**	**Travel Time**	**Social Cost**	**Combined Cost**
5000	**28.22**	13.08	17.17	34.99	15.82	27.32
6000	**29.76**	13.70	18.26	38.97	17.15	29.81
7000	**31.47**	14.51	19.17	40.51	17.77	29.70
8000	**39.25**	17.45	26.85	45.21	19.31	33.21
9000	**41.65**	18.18	27.97	49.21	20.75	36.35

Table 5. Adaptive Pricing Results for the Online MMRP in the reduced Sioux Falls scenario. Social cost and combined cost are defined the same as in Table 4

		Minimum	**Q1**	**Median**	**Mean**	**Q3**	**Maximum**	**Gini. Coef**
MMRP-PPO	Travel Time	6.343	6.643	6.749	6.734	6.889	7.14	0.016
	Social Cost	0.141	0.307	0.424	0.416	0.54	0.694	0.192
	Combined Cost	86.058	97.519	99.982	100.056	104.134	111.561	0.033
Random	Travel Time	6.619	7.267	7.59	7.55	7.898	8.682	0.032
	Social Cost	0.156	0.345	0.475	0.483	0.625	0.774	0.19
	Combined Cost	23.03	28.271	31.793	31.092	34.843	39.435	0.079

Table 6. Hyperparameters used in our PPO implementation. All other hyperparameters are as seen in [23, 6].

Hyperparameter	[23]	[6]	This paper
Num. Epochs	3	3	2
Minibatch Size	256	256	128
Num. Minibatches	4	8	16
GAE Parameter (λ)	0.95	0.95	0.95
Number of Actors	8	32	64
Clip Parameter (ϵ)	0.1 *	0.2	0.1
VF Coeff. (c_1)	1	0.5	0.1
Action Masking	No	Yes	Yes
Separate Networks	No	Yes	Yes
Reward Norm.	No	Yes	Yes
Prior Reward Norm.	No	No	Yes
Episode Length	128	128	1000
Total Steps	1×10^7	1×10^7	1×10^7

8 Proximal Policy Optimisation for Online MMRP

In the stochastic and uncertain environments presented by the online MMRP, each route provider is part of a highly non-stationary landscape where updates to competitors policies continually reshape observed congestion dynamics and rewards. To achieve stable, scalable learning under these conditions while preserving the competitive and independent nature of the routes within our problem, we deploy independent Proximal Policy Optimisation (PPO). Although PPO was originally designed for single-agent settings, recent studies have shown its surprising robustness and convergence properties in independent multi-agent deployments [31, 4]. The use of a clipped surrogate objective in PPO to bound each update's deviation from the current policy prevents destructive policy swings and balancing exploration with reliable improvement [23]. This clipping is critical in multi-agent contexts, where aggressive updates by one provider can destabilise others and amplify environmental non-stationarity [19, 8].

To stabilise learning in our highly non-stationary setting, we apply reward normalisation so that policy and value networks see consistent signal scales across training. Since the true reward distribution is unknown a priori, we first execute a series of random-policy rollouts to collect a sample of raw rewards $\{R_t\}$, from which we compute the empirical mean μ_R and standard deviation σ_R. During PPO training, each observed reward is then standardised as $\tilde{R}_t = \frac{R_t - \mu_R}{\sigma_R}$, ensuring that the agent's updates operate on zero-mean, unit-variance targets. This simple normalisation prevents runaway gradients and divergent value estimates, yielding more stable convergence even as multiple providers continuously adapt their tolling strategies.

To improve sample efficiency and stabilise learning in our application of PPO, we extend the single-agent PPO setup by running simulation environments in parallel. Each parallel actor instance executes all route-provider agents simultaneously under different random seeds and demand scenarios, collecting trajectories for every policy update. By aggregating experiences across these vectorised rollouts, we obtain a richer, more diverse batch of transitions, reducing gradient variance, accelerating convergence, and yielding more robust policy updates in the inherently non-stationary, competitive pricing environment.

In scenarios 1 and 2, we trained our PPO agents for 1×10^7 steps, with each episode lasting 1000 steps and a full list of hyperparameters used, alongside the two reference implementations guiding our agent design is provided in Table 6. In scenario 3, we trained our PPO agents for 7.2×10^7 steps, and an episode length of 1800 steps. All other hyperparameters were kept as described in Table 6.

9 Empirical Evaluation and Results

This section presents the empirical findings from our experiments evaluating the performance of PPO strategies under different pricing functions in the MMRP. The results of our experiments provide valuable insights into the effectiveness of PPO in addressing the online MMRP, demonstrating increased route profitability, and reduced travel time and more equitable outcomes for players.

Reduced Travel Time in Synthetic, High Demand Settings: Our PPO-based strategy consistently outperformed the Random Agent baseline across most metrics and player numbers. Table 2 represents the travel time, Gini coefficient and profit results for Scenario 1. Specifically, Our PPO achieved significantly lower average travel times compared to the Random Agent in the 2-route synthetic environment, with reductions from 26.66 timesteps at 500 players to 1739.88 timesteps at 1000 players, in contrast to the Random Agent's 32.5 to 2083.49 timesteps, respectively. Additionally, our PPO agent consistently outperformed the original PPO in both average travel time and profit metrics. For instance, at 500 players, our PPO achieved an average travel time of 21.09 timesteps, compared to the original PPO's 32.5 timesteps, resulting in a 35% reduction. This indicates that our PPO pricing strategies effectively mitigate congestion, enhancing overall traffic flow efficiency. When our PPO-based

agents are trained on a modified Scenario 1 but with routes of infinite capacity, as seen in table 1, our agents tend towards strategies near the computed PNE, suggesting that the model converges to stable pricing decisions consistent with classical game-theoretic predictions. This alignment reinforces the applicability of reinforcement learning for effectively capturing equilibrium-like behaviour, even under simplified infinite-capacity assumptions. In terms of the Gini coefficient, our PPO agent maintained lower values (0.13–0.33) across varying player counts, whereas the Random agent reached a peak of 0.44 at 750 players before a slight decline, indicating that PPO strategies promote a more equitable congestion distribution. Regarding route profit, PPO's earnings rose consistently from 13.38 (income per car) at 500 players to 90.41 at 1000 players, surpassing the Random agent, which plateaued between 46.13 and 46.55, by approximately 800 players. Moreover, PPO's profitability aligns with the PSO-based baselines in the offline setting, underscoring its effectiveness under heightened congestion.

Effective Management of Congestion under Increased Competition: Across the tested vehicle volumes in Scenario 2, shown in Table 4, our PPO-based approach yields lower travel times than the Random agent, demonstrating its effectiveness in managing congestion under increasing traffic loads. Notably, at 5000 vehicles, PPO records an average travel time of 28.22, compared to 34.99 with Random. Similarly, at 6000 and 7000 vehicles, PPO achieves travel times of 29.76 and 31.47, respectively, outperforming the Random agent by 9.21 and 9.04 timesteps in each case. The results in this scenario demonstrate the ability of our PPO-based solution to adapt in scenarios with greater environment complexity.

Reduced Inequality and Travel Time in Complex, Sequential Applications: On the reduced Sioux Falls network, as seen in Table 5, MMRP-PPO consistently outperforms random tolling across all key metrics. Median travel time drops from 7.59 under random pricing to 6.749 with PPO, and the variability shows a marked reduction (range of travel time narrows from 6.62–8.68 to 6.34–7.14 and the Gini coefficient halves from 0.032 to 0.016). These improvements demonstrate that independent PPO agents yield smoother traffic flows and more equitable outcomes for users, validating the MMRP's ability to drive emergent equilibrium pricing in a fully decentralised, competitive routing setting.

10 Related Work

The MMRP intersects three areas of research; game-theoretic approaches to congestion modelling, algorithmic pricing in transportation networks and decentralised multiagent learning for real-time control. Classical CG literature provides foundational insights into equilibrium existence under fixed costs [21, 13, 1] and enable the simulation of user behaviours and decision-making processes in congested environments [32, 11], while bilevel and Stackelberg formulations [2] provide explore centralised pricing and toll optimisation, and recent multiagent RL literature demonstrates the promise of decentralised learning in non-stationary environments [31, 4, 8].

Dynamic tolling research has largely focused on centralised or monopolistic settings, in contrast to our fully decentralised MMRP approach. Sharon et al. [24] introduced Δ-tolling, which adaptively adjusts link tolls based on marginal latency changes to improve network throughput in a simulation environment. Mirzaei et al. [14, 15] extended Δ-tolling with policy-gradient RL, demonstrating reduced average travel times in empirical studies. Pandey and Boyles [18] introduced a multiagent RL algorithm for dynamic pricing, utilising a joint action space to attain revenue gains under varied entry-exit configurations. Zhu and Ukkusuri [33] applied R-Markov Average Reward Technique, an off-policy RL algorithm which combines observed experiences into a unified trajectory with distance-based rewards, to learn tolls under uncertainty. Wang et al. [29] applied deep RL to utilise signal control and toll adjustments within a cooperative RL setting to outperform existing adaptive baselines. Unlike these works, which assume a single decision-maker or cooperative agents, the MMRP places each route provider in a non-cooperative, self-interested MDP, resulting in equilibrium pricing behaviours driven solely by local observations and reward signals.

Extending beyond single decision-making frameworks for adaptive tolling, Stier-Moses and Acemoglu [26] explore atomic games with market-power players demonstrate the benefit of coordination or regulation in instances where centralised control is not feasible. Harks et al. [5] analyse price-cap regulation in privatised road networks, limited to parallel networks, and derive bounds on the inefficiency induced by uniform caps under affine latencies. This has further been extended to a sequential decision making problem and formulated as MDP congestion games [9], where tolls can be utilised as reward signals to steer equilibria towards social objectives. While these works address equilibrium existence, efficiency and regulatory intervention, they do not consider adaptive and decentralised learning by self-interested providers. A comprehensive review of congestion pricing can be found in [10]. In contrast, the MMRP embeds each route owner's pricing problem within its own MDP, employing independent learning agents to converge towards effective pricing strategies to maximise provider profitability whilst reducing travel time and increasing fairness and equitability outcomes for players.

11 Conclusion

In this work, we introduced the Multi-Market Routing Problem (MMRP), a dynamic extension of classical congestion games that captures decentralised, competitive pricing by multiple route providers under demand from heterogeneous players. We gave a six-tuple offline formulation and its seven-tuple online counterpart, including a definition as an MDP, enabling rigorous analysis of adaptive route pricing and micro-tolling in real time. Our NP-hardness proof for offline profit maximisation underscores the intrinsic computational challenges, motivating the development of scalable, approximate methods. Finally, we demonstrated that decentralised deep-RL agents, each solving its local MDP, learn near-equilibrium pricing policies, providing improvements in profit and travel-time efficiency across synthetic and real-world benchmark scenarios.

We extended independent PPO to improve stability and sample efficiency while mitigating the problem of non-stationarity through reward normalisation, action masking, clipped policy updates and parallelised rollouts. Our results showed marked improvements when compared to single-agent implementation of PPO, and demonstrate that RL-based adaptive pricing and micro-tolling can balance profit, efficiency and equity in competitive routing markets through real-time control, offering a path towards responsive, scalable traffic management solutions aligned with evolving transportation demands.

While our enhanced PPO method has significantly improved congestion management and profitability in the MMRP, further work could explore scalability with more agents, reduced training costs, and integration with advanced techniques (e.g. opponent modelling). To ensure practical applicability, it is also vital to enhance explainability; interpretable models that clarify pricing adjustments can foster trust and support informed decision-making among stakeholders.

Acknowledgements

This research is supported by an ICASE studentship from EPSRC and Yunex Traffic, the EPSRC AutoTrust platform grant (EP/R029563/1), and an EPSRC Turing AI Acceleration Fellowship on Citizen-Centric AI Systems (EP/V022067/1).

References

[1] H. Ackermann, H. Röglin, and B. Vöcking. Pure nash equilibria in player-specific and weighted congestion games. *Theoretical Computer Science*, 410(17):1552–1563, 2009.

[2] L. Brotcorne, M. Labbé, P. Marcotte, and G. Savard. A bilevel model for toll optimization on a multicommodity transportation network. *Transportation science*, 35(4):345–358, 2001.

[3] G. Casey, B. Zhao, K. Kumar, and K. Soga. Context-specific volume–delay curves by combining crowd-sourced traffic data with automated traffic counters: A case study for london. *Data-Centric Engineering*, 1: e18, 2020.

[4] C. S. De Witt, T. Gupta, D. Makoviichuk, V. Makoviychuk, P. H. Torr, M. Sun, and S. Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

[5] T. Harks, M. Schröder, and D. Vermeulen. Toll caps in privatized road networks. *arXiv preprint arXiv:1802.10514*, 2018.

[6] S. Huang, R. F. J. Dossa, A. Raffin, A. Kanervisto, and W. Wang. The 37 implementation details of proximal policy optimization. In *ICLR Blog Track*, 2022. URL https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/. https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

[7] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. ieee, 1995.

[8] B. Koohy, S. Stein, E. Gerding, and G. Manla. Reward function design in multi-agent reinforcement learning for traffic signal control. *ATT'21@International Joint Conference on AI 2021 (IJCAI'21)*, 2021.

[9] S. H. Q. Li, Y. Yu, D. Calderone, L. Ratliff, and B. Açıkmeşe. Tolling for constraint satisfaction in markov decision process congestion games. *arXiv preprint arXiv:1903.00747*, 2019.

[10] C. Lombardi, L. Picado-Santos, and A. M. Annaswamy. Model-based dynamic toll pricing: An overview. *Applied Sciences*, 11(11):4778, 2021.

[11] O. Massicot and C. Langbort. Public signals and persuasion for road network congestion games under vagaries. *IFAC-PapersOnLine*, 51 (34):124–130, 2019.

[12] R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

[13] I. Milchtaich. Congestion games with player-specific payoff functions. *Games and economic behavior*, 13(1):111–124, 1996.

[14] H. Mirzaei, G. Sharon, S. Boyles, T. Givargis, and P. Stone. Enhanced delta-tolling: Traffic optimization via policy gradient reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 47–52. IEEE, 2018.

[15] H. Mirzaei, G. Sharon, S. D. Boyles, T. Givargis, and P. Stone. Link-based parameterized micro-tolling scheme for optimal traffic management. In *Proc. of the 17th Int'l Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2013–2015, Stockholm, Sweden, 2018.

[16] S. Mondal and A. Gupta. Queue-based headway distribution models at signal controlled intersection under mixed traffic. *Transportation Research Record*, 2674(11):768–778, 2020.

[17] R. Neuhold and M. Fellendorf. Volume delay functions based on stochastic capacity. *Transportation research record*, 2421(1):93–102, 2014.

[18] V. Pandey and S. D. Boyles. Multiagent reinforcement learning algorithm for distributed dynamic pricing of managed lanes. In *2018 IEEE Int'l Conf. on Intelligent Transportation Systems (ITSC)*, pages 1–7, 2018.

[19] G. Papoudakis, F. Christianos, A. Rahman, and S. V. Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

[20] J. Paszkowski, M. Herrmann, M. Richter, and A. Szarata. Modelling the effects of traffic-calming introduction to volume–delay functions and traffic assignment. *Energies*, 14(13):3726, 2021.

[21] R. W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, Dec 1973. ISSN 1432-1270. doi: 10.1007/BF01737559. URL https://doi.org/10.1007/BF01737559.

[22] T. Roughgarden and É. Tardos. Bounding the inefficiency of equilibria in nonatomic congestion games. *Games and economic behavior*, 47(2): 389–403, 2004.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[24] G. Sharon, J. Hanna, T. Rambha, M. Albert, P. Stone, and S. D. Boyles. Delta-tolling: Adaptive tolling for optimizing traffic throughput. In *ATT@ IJCAI*, 2016.

[25] H. Spiess. Conical volume-delay functions. *Transportation Science*, 24 (2):153–158, 1990.

[26] D. S. Stier-Moses and D. Acemoglu. The impact of oligopolistic competition in networks. *Operations Research*, 56(5):1193–1208, 2008. doi: 10.1287/opre.1080.0610.

[27] C. Suwansirikul, T. L. Friesz, and R. L. Tobin. Equilibrium decomposed optimization: a heuristic for the continuous equilibrium network design problem. *Transportation science*, 21(4):254–263, 1987.

[28] H. Tavafoghi and D. Teneketzis. Informational incentives for congestion games. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1285–1292. IEEE, 2017.

[29] Y. Wang, H. Jin, and G. Zheng. Ctrl: Cooperative traffic tolling via reinforcement learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3545–3554, 2022.

[30] H. Yang and H. Hai-Jun. Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory. *Transportation Research Part B: Methodological*, 31(6):425–440, 1997.

[31] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

[32] J. Zhang, J. Lu, J. Cao, W. Huang, J. Guo, and Y. Wei. Traffic congestion pricing via network congestion game approach. *Discrete & Continuous Dynamical Systems: Series A*, 41(7), 2021.

[33] F. Zhu and S. V. Ukkusuri. A reinforcement learning approach for distance-based dynamic tolling in the stochastic network environment. *Journal of Advanced Transportation*, 49(2):247–266, 2015.