Online Appendix to

Learning Multimodal Markov Departure Time Choice

Ben Wang, Qi Luo, and Yafeng Yin

Appendix A: Summary of Notation

Table 4 Summary of Major Notation

Notation	Definition
$a_{i,h}$	Individual action $(m_{i,h}, t_{i,h})$ of the commuter i on day h
c	Daily travel cost function
$\mathbb{C}$	Average daily travel cost per mile
$d_{\min}(X)$	Minimum pairwise distance between distinct points in X
$d_{TV}(\Gamma_1, \Gamma_2)$	Total variation distance between $\Gamma_1 \in \mathbb{G}$ and $\Gamma_2 \in \mathbb{G}$
$\text{diam}(X)$	Diameter of the set X
J	Expected discounted total cost
$k_{m,h}(x)$	System density of the travel mode m at time $x \in \mathcal{T}$ on day h
$\tilde{k}_{m,h}(x)$	Traffic congestion of the travel mode m at time x on day h
$l_{i,h}$	Trip length of the commuter $i \in \mathcal{N}$ on day $h \in \mathcal{H}$
$\triangle l$	Distance interval
ω_h	Demand mass on day h
$m_{i,h}$	Travel mode of the commuter i on day h
$\mathcal{N}_{l,\xi,h}$	Commuters with the trip length $l \in \mathcal{L}$ and the desired arrival time $\xi \in \mathcal{X}$ on day h
$\mathcal{O}_\Gamma$	Estimated travel mode policy given the mean-field distribution Γ
P	State transition function
p	Behavior inertia
$s_{i,h}$	Individual state $(l_{i,h}, \xi_{i,h})$ of the commuter i on day h
$T_m(t)$	Travel time of commuter who departs at $t \in \mathcal{T}^d$ on the mode $m \in \mathcal{M}$
$t_{i,h}$	Departure time of the commuter i on day h
$\bar{t}$	Latest available departure time
$\bar{t}'$	End of the within-day time horizon $\mathcal{T}$
$\underline{t}$	Earliest available departure time
$\triangle t$	Time interval
V_m	Speed function that maps the congestion of mode m to the system velocity
$v_{m,h}(x)$	Velocity of the travel mode m in the system at time x on day h
$\mathbb{W}$	Equilibrium error
Γ_h	Mean-field distribution, the joint state $s \in \mathbb{S}$ and action $a \in \mathbb{A}$ distribution on day h
$\zeta_m(s;\Gamma)$	Learning policy of departure time given the travel mode m and the mean-field distribution Γ
ζ_Γ	Learning policy of departure time and travel mode given the mean-field distribution Γ
$\mu_h(l,\xi)$	Proportion of commuters with the trip length l and the desired arrival time ξ on day h
$\xi_{i,h}$	Desired arrival time of the commuter i on day h
π_i	Stationary Markov policy of the commuter i
π_{-i}	Stationary Markov policy profile except for the commuter i
$\tau_h(m,t)$	Proportion of commuters departing at t with the mode m on day h
$\phi_{m,j}$	Impact of mode j's density on mode m's congestion
$\psi_{m,h}$	Characteristic travel distance of the mode m on day h
Acronym	**Definition**
AMPE	Approximate Markov Perfect Equilibria
HPG	Hierarchical Policy Gradient
MDTMC	Markovian Departure Time and Mode Choice
MPE	Markov Perfect Equilibria
SMFE	Stationary Mean Field Equilibria

The notations are introduced in alphabetical order.

Appendix B: Supplementary Materials for Section 2

B.1. Case Study in Section 2.4

B.1.1. Fitting population distributions. The preprocessing of AACVTE trip data includes the following steps:

1. Removing outliers: The analysis focuses on highest-volume OD pairs during the morning rush hour (7:30 am to 10:00 am). We detect and remove outliers with abnormal trip lengths using the IsolationForest algorithm from the sklearn.ensemble package. Isolation forest algorithm is an effective anomaly detection method that identifies outliers by isolating observations through random partitioning. By applying this technique, we ensure that the dataset excludes extreme trip data that could skew the analysis.

2. Estimate state transition matrices: Since the commuter identity is unobserved within the dataset, it is challenging to directly estimate the probability transition matrix by observing the trip length dynamics of each commuter. Instead, we parameterize the commuter's trip length transition probability matrix by $\boldsymbol{\theta}$ and apply an Expectation Maximization (EM) algorithm to estimate the optimal $\boldsymbol{\theta}^*$.

Denote the empirical trip length distributions and a trip length sample sequence of a representative agent of H days by $\mathbf{X} = \{\hat{\omega}_h\}_{h=1}^H$ and $\mathbf{Z}$, respectively. The EM algorithm aims to find the maximum likelihood estimate of the transition matrix parameter by iteratively applying the expectation step (E-step) and the maximization step (M-step). In the E-step, the expected value of the log-likelihood function $L(\boldsymbol{\theta} \mid X, Z)$, concerning the conditional probability of $\mathbf{Z}$ given $\mathbf{X}$ and the current estimate $\boldsymbol{\theta}^k$, is denoted by $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^k)$. It calculates $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^k) = \mathbb{E}_{\mathbf{Z}\sim p(\cdot\mid\mathbf{X},\boldsymbol{\theta}^k)}\big[L(\boldsymbol{\theta} \mid X, Z)\big]$. In the M-step, the next parameter $\boldsymbol{\theta}^{k+1}$ is determined by $\boldsymbol{\theta}^{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^k)$. The E-step and M-step continue iteratively until convergence.

As the dataset only contains the actual arrival time information of anonymous commuters, we deduce the desired arrival time for each day using a log-normal distribution (Li, Zhou, and Rouphail 2017). We first estimate the scale parameter σ by fitting all actual arrival times over all days. For each day h, we take the most frequent actual arrival time as the mean desired arrival time of that day, denoted by $\hat{\xi}_h$. The location parameter μ_h for day h is then calculated using $\mu_h = \log(\hat{\xi}_h) - \frac{1}{2}\sigma^2$. Subsequently, we apply the EM algorithm to estimate the optimal transition matrix for the dynamics of μ.

B.1.2. Data augmentation for alternative travel modes. Since the dataset only provides information on commuters who choose to drive, we need to perform data augmentation before analyzing the correlation between travel modes and commuter states. Let $\mathcal{F}_1$ denote the observed commuter state distribution over days for those who choose driving, and let $\mathcal{F}_2$ denote the unknown commuter state distribution for those who choose the alternative mode. An iterative Bayesian updating algorithm is employed to estimate $\mathcal{F}_2$, while the probabilities of choosing driving and the alternative mode, denoted by p_1 and p_2, are also unknown. The initial guess on $\mathcal{F}_2$ is set to be beta distribution, using a Beta-Binomial conjugate model. Two main steps are performed: In the first step, we calculated the best response (γ_i), representing the probability of driving for a given commuter state $X_i \sim \mathcal{F}_1$. This is computed using the current estimates of the travel mode probabilities p_1, p_2 and commuter state distribution $\mathcal{F}_2^k$ as follows:

$$\gamma_i = \frac{p_1^k \mathcal{F}_1(X_i)}{p_1^k \mathcal{F}_1(X_i) + p_2^k \mathcal{F}_2^k(X_i)}.$$

In the second step, we updated the state distribution for the alternative mode commuters and the probabilities of choosing each mode based on the computed best response. The expected count of commuters in each state for the alternative mode is computed by summing the best response for each state, $N_{s,2}^{k+1} = \sum_{i:\mathbf{X}_i=s}(1-\gamma_i)$. The state distribution $\mathcal{F}_2^{k+1}$ is then updated by normalizing these counts as $\mathcal{F}_2^{k+1}(s) = N_{s,2}^{k+1}/\sum_s N_{s,2}^{k+1}$. Simultaneously, the travel mode probabilities are updated by averaging the best response across all observed commuter states $p_1^{k+1} = \frac{1}{n}\sum_{i=1}^n \gamma_i$ and $p_2^{k+1} = 1 - p_1^{k+1}$. This iterative process continues until the changes in the commuter state distribution and travel mode probabilities between iterations are below a predefined tolerance. After this data augmentation step, we can generate a comparable number of commuter states for the alternative travel mode and perform logistic regression to analyze the correlation between the travel mode and commuter states.

Appendix C: Supplementary Materials for Section 3

C.1. Proximal Policy Gradient (PPO) Implementation

While any (vanilla) policy-gradient methods will not affect the main theoretical results, implementing a more efficient and robust policy gradient method will certainly improve the convergence performance. PPO is used in each inner-loop iteration of the HPG algorithm. Here, a fully connected neural network with two hidden layers of 256 neurons is implemented for policy representation. We also tried other alternative representation models and found this one to be adequate for the bimodal commute problem.

The policy output layer uses the hyperbolic tangent activation function that takes the desired arrival time ξ and the trip length l as input and outputs logits $o_t \in [-1,1]^{\frac{(\bar{t}-\underline{t})}{\triangle t}+1}$ and logits $o_m \in [-1,1]^2$. The departure time policy and the travel mode policy are then computed as a categorical distribution by a softmax layer:

$$\zeta(t,m) = \text{softmax}_\nu(o_t) \times \text{softmax}_\nu(o_m). \tag{13}$$

Here, the temperature parameter ν takes value 1.

The values of hyperparameters are provided in Table 5:

Table 5 HPG Algorithm Hyperparameter Values

Hyperparameter	Value
Discount factor	0.99
Generalized Advantage Estimator (GAE) lambda	1
Kullback–Leibler coefficient	0.03
Clip parameter	0.2
Learning rate	0.00005
Training batch size	4,000
Mini-batch size	2,000
Stochastic Gradient Descent (SGD) steps per batch	8
PPO training steps (inner-loop step)	1000

As the morning commute problem is investigated under the assumption of an infinite horizon, we limit the planning horizon to 80 days in PPO implementations, which suffices for convergences across all instances. Fifty episodes of 80-day training trajectory samples are collected, and eight steps of SGD update are applied per PPO training step. The training process involves 1000 PPO training steps (inner-loop steps) for a single PPO implementation (outer-loop step).

C.2. Definitions of Distances and Metric Spaces.

W_1-Wasserstein distance. The W_1-Wasserstein distance, denoted as $W_1(\cdot)$, quantifies the dissimilarity between two probability distributions by calculating the minimum cost required to transport the mass from one probability distribution to another. Given that two probability distributions of the generalized mean-field, Γ_1 and Γ_2, are supported on the same metric space $X \subseteq \mathbb{R}^k$, the W_1-Wasserstein distance is expressed as an integral of the Kantorovich-Rubinstein norm:

$$W_1(\Gamma_1, \Gamma_2) = \sup_{f \in \mathcal{F}} \left| \int_X f d\Gamma_1 - \int_X f d\Gamma_2 \right|,$$

where the Wasserstein distance is defined as the supremum of the absolute difference between the integrals of a 1-Lipschitz function, denoted as f, concerning Γ_1 and Γ_2; $\mathcal{F}$ denotes a set of 1-Lipschitz functions, i.e., any function that satisfies the condition that, for all $x, x' \in X$, $|f(x) - f(x')| \leq \|x - x'\|_2$.

Total variation distance. Define the total variation distance as:

$$d_{TV}(\Gamma_1, \Gamma_2) = \frac{1}{2} \int_{x \in X} |\Gamma_1(x) - \Gamma_2(x)| dx. \tag{14}$$

Combining these two metrics, we have the following lemma:

LEMMA 4 **(Gibbs and Su (2002))**. *The W_1-Wasserstein distance between two probability distributions Γ_1 and Γ_2, is bounded by the following inequality:*

$$d_{\min}(X) \cdot d_{TV}(\Gamma_1, \Gamma_2) \leq W_1(\Gamma_1, \Gamma_2) \leq diam(X) \cdot d_{TV}(\Gamma_1, \Gamma_2). \tag{15}$$

Here, $d_{\min}(X)$ is the minimum pairwise distance between distinct points in X such that $d_{\min}(X) \triangleq \min_{x \neq x' \in X} \|x - x'\|_2$; $\mathrm{diam}(X)$ represents the diameter of the set X, defined as $\mathrm{diam}(X) \triangleq \sup_{x, x' \in X} \|x - x'\|_2$.

C.3. Proofs for Lemma 1 to Lemma 3 and Theorem 1.

Proof for Lemma 1: Consider the Markov strategic form of the MPE for the N-player morning commute problem. The commuter i chooses a mixed strategy for $a_{i,h} \in \mathbb{A}$ and induces a cost $c(s_{i,h}, \omega_h, \mathbf{a}_h)$, i.e., a probability over all feasible departure times for each mode. Since there are finitely many states, the Markov strategic form has a finite number of commuters, each of whom has a finite number of pure strategies as $\mathbb{A}$ and Nash equilibrium exists. As the mixed strategies at the equilibrium is a Markov strategy profile of the original game for any commuter i given the state $s_{i,h}$, it is the best response to π_{-i}. Hence, π is an MPE by construction. □

Proof for Lemma 2: We show the existence of AMPE in the non-zero-sum stochastic game with compact state and action spaces. The N-player morning commute problem is a special stochastic game using the Markov policy. Under the assumptions of boundedness and continuity for the commuters' cost functions in Assumption 1, Whitt (1980) shows that this stochastic game possesses a stationary AMPE. Since this extension is a notable result in stochastic games, we focus on proving the second part of this lemma.

It is worth noting that the existence of AMPE can be strengthened to the result of MPE only when introducing norm-continuous assumptions 2 on the transition probability function. For notational brevity, we denote the state $x_i = (s_{i,h}, \omega_h)$. In the separable state space, let $\mathbb{X}_A \subset \mathbb{S} \times \Omega$ denote the countable subset in Assumption 2. Let U_i denote a complete metric space of function $u : \mathbb{X}_A \to \mathbb{R}$ such that $|u(x_i)| \leq L$ for

all $x_i \in \mathbb{S}$. $U = \prod_i U_i$. Let V_i denote the space of all δ_N-equivalence classes of Borel measurable function $v : \mathbb{S} \times \Omega \backslash \mathbb{X}_A \to \mathbb{R}$, such that $|v(x_i)| \le L\delta_N$ almost everywhere. $V = \prod_i V_i$ is a convex and compact space. Recall that, given stationary MPE policy π^*_{-i},

$$\begin{aligned} J^{\pi^*}(s,\omega) &= \min \mathbb{E}^{\pi_i,\pi^*_{-i}}\left[\sum_{h=0}^{\infty} \lambda^h c(s_{i,h},\omega_h,\mathbf{a}_h)\middle| s_{i,0}=s,\omega_0=\omega\right] \\ &= \min_{a_i}\left\{c(s_{i,h},\omega_h,\mathbf{a}_h) + \lambda\int_{\mathbb{S}\times\Omega\backslash\mathbb{X}_A} v(x')P_N(x'|x,\boldsymbol{a})dx' + \lambda\sum_{x'\in\mathbb{X}_A} u(x')P_A(x'|x,\boldsymbol{a})\right\}. \end{aligned}$$

Define the λ-contraction operator T from above. We can define a set-valued mapping from $V \times U$ to a non-empty, compact, convex subset of itself. For every commuter, the fixed point regarding U is uniquely obtained due to countable $\mathbb{X}_A$. Let the value at the fixed point be u^0 and the correspondence of the set-value function for commuter i be U_i^0. On the other hand, we can find a set of all δ_N-equivalence classes of Borel measurable selectors, $M[v,u^0]$, for any $x' \in \mathbb{S}\times\Omega\backslash\mathbb{X}_A$.

According to Nowak (2003), given the norm-continuous assumption, $\Psi = M[v,u^0]\times\prod_i U_i^0$ is convex, compact-valued, and upper semicontinuous. Therefore, for this stochastic game, for a sequence $v^k \to v, u^n \in u$ for every k, we have

$$\sup_{\boldsymbol{a}} |J^{\pi_i^k,\pi_{-i}}(s,\omega) - J^{\pi^*}(s,\omega)| \to 0,$$

as $k \to \infty$. Ψ are closed. Based on Kakutani–Glicksberg's fixed point theorem, there exists a stationary Nash equilibrium in this stochastic game. □

Proof for Lemma 3: Recall the inequalities of W_1-Wasserstein distance introduced in Appendix C.2:

$$W_1(F_2(\pi_1,\Gamma),F_2(\pi_2,\Gamma)) \le \mathrm{diam}(\mathbb{S}\times\mathbb{A}) d_{TV}(F_2(\pi_1,\Gamma),F_2(\pi_2,\Gamma)). \tag{16}$$

Consider the generalized mean-field $\Gamma' = F_2(\pi,\Gamma)$, which can be expanded as:

$$\Gamma'(s',a') = \mu'(s')\pi(a'|s'), \quad \mu'(s') = \sum_{s\in\mathbb{S},a\in\mathbb{A}} \mu(s)P(s'|s,a,\Gamma)\pi(a|s). \tag{17}$$

According to Assumption 4, we can show that:

$$\begin{aligned} &d_{TV}(F_2(\pi_1,\Gamma),F_2(\pi_2,\Gamma)) = d_{TV}\Bigg(\sum_{s\in\mathbb{S},a\in\mathbb{A}} \mu(s)P(s'|s,a,\Gamma)\pi_1(a|s)\pi_1(a'|s') - \\ &\sum_{s\in\mathbb{S},a\in\mathbb{A}} \mu(s)P(s'|s,a,\Gamma)\pi_2(a|s)\pi_2(a'|s')\Bigg) (d_{TV} \text{ defined in Equation (14)}) \\ &= \frac{1}{2}\sum_{s'\in\mathbb{S},a'\in\mathbb{A}}\left|\sum_{s\in\mathbb{S},a\in\mathbb{A}} \mu(s)P(s'|s,a,\Gamma)(\pi_1(a|s)\pi_1(a'|s') - \pi_2(a|s)\pi_2(a'|s'))\right| \\ &\le \frac{1}{2}\max_{s',s,a,\Gamma} P(s'|s,a,\Gamma)\sum_{s',s,a',a} \mu(s)(\pi_1(a|s)+\pi_2(a|s))|\pi_1(a'|s') - \pi_2(a'|s')| \\ &\le \frac{1}{2}\kappa_1\sum_{s',a'} 2\cdot|\pi_1(a'|s') - \pi_2(a'|s')| \\ &= \kappa_1\sum_{s',a'}|\pi_1(a'|s') - \pi_2(a'|s')| = 2\kappa_1\sum_{s'} d_{TV}(\pi_1(s'),\pi_2(s')) \\ &\le 2\kappa_1|\mathbb{S}|\sup_{s'\in\mathbb{S}} d_{TV}(\pi_1(s'),\pi_2(s')) \\ &\le 2\kappa_1|\mathbb{S}|\sup_{s'\in\mathbb{S}}\frac{W_1(\pi_1(s'),\pi_2(s'))}{d_{min}(\mathbb{A})} \\ &= \frac{2\kappa_1|\mathbb{S}|}{d_{min}(\mathbb{A})} D(\pi_1,\pi_2). \end{aligned}$$

Thus, we set

$$d_2 = \frac{2\kappa_1 \text{diam}(\mathbb{S})\text{diam}(\mathbb{A})|\mathbb{S}|}{d_{min}(\mathbb{A})}.$$

Substitute $\text{diam}(\mathbb{S}) = \sqrt{\text{diam}(l)^2 + \text{diam}(\xi)^2}$ and $\text{diam}(\mathbb{A}) = \sqrt{(M-1)^2 + (\bar{t} - \underline{t})^2}$, $|\mathbb{S}| = (\text{diam}(\xi)/\triangle t + 1)(\text{diam}(l)/\triangle l + 1)$. We have $d_{min}(\mathbb{A}) = 1/\sqrt{1+\triangle t^2}$ and obtain the expression of d_2.

Similarly, we have

$$\begin{aligned} W_1(F_2(\pi,\Gamma_1), F_2(\pi,\Gamma_2)) &\le \text{diam}(\mathbb{S}\times\mathbb{A}) d_{TV}(F_2(\pi,\Gamma_1), F_2(\pi,\Gamma_2)) \\ &= \frac{\text{diam}(\mathbb{S}\times\mathbb{A})}{2} \sum_{s'\in\mathbb{S}, a'\in\mathbb{A}} \left| \sum_{s\in\mathbb{S}, a\in\mathbb{A}} \mu(s)\pi(a|s)\pi(a'|s')[P(s'|s,a,\Gamma_1) - P(s'|s,a,\Gamma_2)] \right| \\ &\le \frac{\kappa_2 \text{diam}(\mathbb{S}\times\mathbb{A})}{2} \end{aligned}$$

Substituting $\text{diam}(\mathbb{S})$ and $\text{diam}(\mathbb{A})$ back gets the expression of d_3. □

Proof for Theorem 1: With the mapping F_1 and F_2 defined above, a profile of (π,Γ) is an SMFE if and only if $\Gamma = F(\Gamma) = F_2(F_1(\Gamma),\Gamma)$ and $\pi = F_1(\Gamma)$. For any $\Gamma_1, \Gamma_2 \in \mathbb{G}$, we have

$$\begin{aligned} W_1(F(\Gamma_1), F(\Gamma_2)) &= W_1(F_2(F_1(\Gamma_1),\Gamma_1), F_2(F_1(\Gamma_2),\Gamma_2)) \\ &\le W_1(F_2(F_1(\Gamma_1),\Gamma_1), F_2(F_1(\Gamma_2),\Gamma_1)) + W_1(F_2(F_1(\Gamma_2),\Gamma_1), F_2(F_1(\Gamma_2),\Gamma_2)) \\ &\le (d_1 d_2 + d_3) W_1(\Gamma_1,\Gamma_2) \end{aligned}$$

When $d_1 d_2 + d_3 \in [0,1)$, we conclude F is a contraction mapping with respect to the W_1-distance. Since the metric space $\mathbb{G}$ is complete, the Banach fixed-point theorem ensures the existence of the mean-field Γ such that $\Gamma = F(\Gamma)$. □

C.4. Convergence of the HPG Algorithm

The HPG algorithm computes the optimal policy, denoted as $\zeta_h = \hat{F}_1(\Gamma_h)$, for the mean-field Γ_h, which is obtained through the two-step procedure introduced in Section 2.3. Let Γ^* be the mean-field distribution at SMFE. Given Assumption 3 (b), we have

$$\begin{aligned} W_1(\Gamma_{h+1}, \Gamma^*) &= W_1(F_2(\hat{F}_1(\Gamma_h),\Gamma_h), F_2(F_1(\Gamma^*),\Gamma^*)) \\ &\le W_1(F_2(F_1(\Gamma_h),\Gamma_h), F_2(F_1(\Gamma^*),\Gamma^*)) + W_1(F_2(\hat{F}_1(\Gamma_h),\Gamma_h), F_2(F_1(\Gamma_h),\Gamma_h)) \\ &\le W_1(F(\Gamma_h), F(\Gamma^*)) + d_2 D(\hat{F}_1(\Gamma_h), F_1(\Gamma_h)) \end{aligned} \tag{18}$$

The first term is bounded by Theorem 1. Denote $d = d_1 d_2 + d_3$, and we have $W_1(F(\Gamma_h), F(\Gamma^*)) \le d W_1(\Gamma_h, \Gamma^*)$. The second term represents the optimality gap between the optimal departure time, travel mode policy, and the estimated policy learned from the HPG.

Note that HPG decomposes the policy ζ_h into the departure time policy $\zeta_{m,h}$, which can be learned by any policy gradient type of method, and the travel mode policy $\mathcal{O}_h$ that follows the Boltzmann policy over the state-mode value induced by $\zeta_{m,h}$. Consider the vanilla policy gradient. (Yuan, Gower, and Lazaric 2022) (Corollary E.5.) gives the guarantee, that is, for any mode $m \in \mathcal{M}$, with probability at least $1 - \delta_{m,h}$, we have $\|Q^{\zeta_{m,h}} - Q^*\| \le \epsilon_{m,h}$ after obtaining $T_{m,h}$ samples:

$$T_{m,h} = \frac{\mathcal{C}_h \log(\frac{1}{\epsilon_{m,h}}) \log(\frac{1}{\delta_{m,h}})}{\epsilon_{m,h}^4 \delta_{m,h}^2}, \tag{19}$$

where $\mathcal{C}_h$ denotes the hyperparameters related to the learning method.

As the departure time policy $\zeta_{m,h}$ for each mode m is learned, setting $\epsilon_{m,h} = \epsilon_h/M$ and $\delta_{m,h} = 1-(1-\delta_h)^{\frac{1}{M}}$ guarantees the convergence to equilibrium. With the probability at least $1-\delta_h$, we have $\|\zeta_H(s), \pi_H(s)\|_2 \le \epsilon_h$ after obtaining $T_h(\delta_h, \epsilon_h)$ samples:

$$T_h(\delta_h, \epsilon_h) = \frac{M^5 \mathcal{C}_h \log(\frac{M}{\epsilon_h}) \log(\frac{1}{1-(1-\delta_h)^{\frac{1}{M}}})}{\epsilon_h^4 (1-(1-\delta_h)^{\frac{1}{M}})^2}, \tag{20}$$

With the HPG learning guarantee, we have the following with the probability of at least $1-\delta_h$ after obtaining $T_h(\delta_h, \epsilon_h)$ samples:

$$\begin{aligned}
d_2 D(\hat{F}_1(\Gamma_h), F_1(\Gamma_h)) &= d_2 D(\zeta_h, \pi_h) && (21)\\
&= d_2 \sup_{s\in\mathbb{S}} W_1(\zeta_h(s), \pi_h(s)) && (22)\\
&\le d_2 \int_{\mathbb{S}} \operatorname{diam}(\mathbb{A}) d_{TV}(\zeta_h(s), \pi_h(s)) && (23)\\
&\le d_2 \int_{\mathbb{S}} \frac{\operatorname{diam}(\mathbb{A})}{2} \|\zeta_h(s), \pi_h(s)\|_1 && (24)\\
&\le d_2 \int_{\mathbb{S}} \frac{\operatorname{diam}(\mathbb{A})\sqrt{|\mathbb{A}|}}{2} \|\zeta_h(s), \pi_h(s)\|_2 \quad \text{(Cauchy–Schwarz inequality)} && (25)\\
&= \frac{d_2 \operatorname{diam}(\mathbb{A})\sqrt{|\mathbb{A}|}|\mathcal{L}||\mathcal{X}|\epsilon_h}{2}, && (26)
\end{aligned}$$

where (21) to (23) and (26) are definitions in the main body of Section 2.1. (24) is from Appendix C.2. (25) is a variation of the Cauchy-Schwarz inequality.

Combining this with (18), with the probability at least $1-\delta_h$, we have

$$W_1(\Gamma_{h+1}, \Gamma^*) \le d W_1(\Gamma_h, \Gamma^*) + \frac{d_2 \operatorname{diam}(\mathbb{A})\sqrt{|\mathbb{A}|}|\mathcal{L}||\mathcal{X}|\epsilon_h}{2}. \tag{27}$$

Therefore, with the probability at least $1-\sum_{h=0}^{H-1}\delta_h$,

$$\begin{aligned}
W_1(\Gamma_H, \Gamma^*) &\le d^H W_1(\Gamma_0, \Gamma^*) + \frac{d_2 \operatorname{diam}(\mathbb{A})\sqrt{|\mathbb{A}|}|\mathcal{L}\|\mathcal{X}| \sum_{h=0}^{H-1} d^{H-h}\epsilon_h}{2}\\
&\le d^H W_1(\Gamma_0, \Gamma^*) + \frac{d \cdot d_2 \operatorname{diam}(\mathbb{A})\sqrt{|\mathbb{A}|}|\mathcal{L}\|\mathcal{X}| \sup_h \epsilon_h}{2(1-d)}. && (28)
\end{aligned}$$

Since $d \le 1$ and $W_1(\Gamma_0, \Gamma^*) \le \operatorname{diam}(\mathcal{S}\times\mathcal{A})$, we set $\epsilon = \sup_h \epsilon_h$ and $\delta = \sum_{h=0}^{H-1}\delta_h$. When $H \ge \log_d(\frac{\epsilon}{\operatorname{diam}(\mathcal{S})\operatorname{diam}(\mathcal{A})})$, we have with the probability at least $1-\delta$:

$$W_1(\Gamma_H, \Gamma^*) \le \left(1 + \frac{d \cdot d_2 \operatorname{diam}(\mathbb{A})\sqrt{|\mathbb{A}|}|\mathcal{L}||\mathcal{X}|}{2(1-d)}\right)\epsilon. \tag{29}$$

In sum, the HPG algorithm will achieve convergence with T samples,

$$\begin{aligned}
T &= \sum_{h=0}^{H} T_h(\delta_h, \epsilon_h)\\
&= \sum_{h=0}^{H} \frac{M^5 \mathcal{C}_h \log(\frac{M}{\epsilon_h}) \log(\frac{1}{1-(1-\delta_h)^{\frac{1}{M}}})}{\epsilon_h^4 (1-(1-\delta_h)^{\frac{1}{M}})^2}.
\end{aligned}$$

C.5. Extension to General Settings with Intrinsic and Extrinsic Factors

We can generalize the morning commute problem by incorporating additional factors that influence daily transitions as follows. Let I_h^i denote the intrinsic factors influencing agent i's daily demand profile, explicitly capturing demand fluctuations. Let E_h denote the extrinsic factors observed by all agents that affect the daily travel cost, capturing effects such as weather conditions that impact travel times and mode preferences on day h. The space of these extrinsic factors is $\mathbb{E}$. Both factors can be modeled as Markov chains. The intrinsic factor may also depend on the previous day's decision or realized cost, whereas the extrinsic factor follows an exogenous Markov chain that is independent of the agents' policy.

Regarding the representative agent, we expand the state space $\mathbb{S}$ to $\tilde{\mathbb{S}}$ such that $\tilde{s}_h \triangleq (s_h, I_h) \in \tilde{\mathbb{S}}$. The realized augmented state and extrinsic factors on day h are denoted by $\tilde{s}_h$ and e_h, respectively. The new policy is defined as $\tilde{\pi} : \tilde{\mathbb{S}} \times \Omega \times \mathbb{E} \to \Delta(\mathbb{A})$. The augmented transition kernel $\tilde{P} : \tilde{\mathbb{S}} \times \mathbb{A} \times \mathbb{G} \times \mathbb{E} \to \Delta(\tilde{\mathbb{S}} \times \mathbb{E})$ is given as follows:

$$\tilde{P}(\tilde{s}', e' | \tilde{s}, a, \Gamma, e) = P(\tilde{s}' | \tilde{s}, a, \Gamma) P(e' | e),$$

where $P(\tilde{s}' | \tilde{s}, a, \Gamma)$ is similar to the original transition kernel, and $P(e'|e)$ is the transition probability of the Markovian extrinsic factor. The cost function is redefined on the augmented state as $c(\tilde{s}_h, a_h, \Gamma_h, e_h)$. This cost function can be further introduced to explicitly express how intrinsic and extrinsic factors affect the travel cost. For example, recall that the action includes the departure time such that $a_{i,h} = t_{i,h}, m_{i,h}$, then the trip time is extended as a function $T_m \triangleq T_m(a_{i,h}, a_{-i,h}; \tilde{s}_{i,h}, \omega_h, E_h)$. As a result, the instantaneous cost is defined as

$$c(a_{i,h}, a_{-i,h}; \tilde{s}_{i,h}, \omega_h, E_h) = \alpha T_m + \beta(\xi(I_h^i) - t_{i,h} - T_m)^+ + \gamma(\xi(I_h^i) - t_{i,h} - T_m)^-,$$

where the $\alpha - \beta - \gamma$ daily travel cost function distinguishes between two categories of influencing factors. The intrinsic factor, such as a change in individual schedule preferences, is represented by the desired arrival time ξ, while the extrinsic factors, such as weather conditions, affect the average travel time T_m. Hence, within the mean-field game setting, the cost function is extended to $c(\tilde{s}_h, a_h, \Gamma_h, E_h)$, allowing the remaining optimality conditions and the equilibrium characterization in (10) to be redefined accordingly.

Appendix D: Analysis of Alternative Optimization Benchmarks

In the main text, we focused on a mean-field approximation of commuters' departure-time choices under the assumption that each commuter optimizes long-term outcomes. Yet, real-world commuters often prioritize more immediate costs, making the validation of planning horizon equally relevant. This appendix investigates how system dynamics and equilibrium outcomes change when commuters' planning horizons shift from predominantly long-term to more short-term objectives. We compare the mean-field-game-based model with two benchmarks: multi-agent reinforcement learning (MARL) and fictitious play, which represent general policies with varying and short planning horizons, respectively.

D.1. Benchmark 1: Fictitious Play with Myopic Agents

The first benchmark represents myopic commuters who optimize their daily travel plans based on prior information. Although several studies have proposed various day-to-day mode or departure time switching rules (Guo, Yang, and Huang 2018, Jin 2020b), these approaches either lack theoretical convergence guarantees or are excessively complex for practical implementation. Therefore, we adopt a fictitious play framework in Algorithm 3 in which agents act as purely myopic decision-makers based on empirical estimation of traffic congestion patterns.

Algorithm 3: Fictitious Play with Myopic Agents

Input: Number of agents N and number of days H; MDTMC simulator $\mathcal{G}$ (Algorithm 4).

for *agent* $i \in \{1, \dots, N\}$ **do**
  Initialize the estimated state-action distribution $\hat{\Gamma}_i$ of other agents with initial state s_i;
end
for $h = 1$ **to** H **do**
  for *agent* $i \in \{1, \dots, N\}$ **do**
    Optimize for a_i via Bayesian optimization with simulator $\mathcal{G}$ and estimation $\hat{\Gamma}_i$:
    $a_i \leftarrow \arg\min_a \mathbb{E}_{(s_{-i}, a_{-i}) \sim \hat{\Gamma}_i} \left[c_i(s_i, a, s_{-i}, a_{-i}) \right]$
  end
  Execute joint action: $\left(s'_1, \dots, s'_N\right), \left(r_1, \dots, r_N\right) \leftarrow \mathcal{G}\left(\{s_1, \dots, s_N\}, \{a_1, \dots, a_N\}\right)$;
  for *agent* $i \in \{1, \dots, N\}$ **do**
    Record new observations of other agents: $\hat{\Gamma}_i \leftarrow \text{Update}\left(\hat{\Gamma}_i, \{s_{-i}, a_{-i}\}\right)$;
    State transition: $s_i \leftarrow s'_i$ with $P(s'_i | s_i, a_i)$;
  end
end

Rather than optimizing cumulative rewards over time, each myopic agent tracks historical observations of aggregate states and actions, selecting a one-step action that maximizes its immediate payoff. However, computing the best response using the fictitious play heuristic is challenging in large state spaces and becomes even more intractable with continuous trip lengths or desired arrival times. To address this, the benchmark model calculates policies for N representative agents, each representing a subgroup of commuters. At the initialization stage, each representative agent i maintains an estimate of the state-action distribution of other agents, $\hat{\Gamma}_{i,0}$, typically assumed to be a uniform distribution over the joint state-action space. On day h, all agents select the optimal departure rate per mode, $a_{i,h}$, as a best response to minimize their expected cost, given the state distribution of their respective group, $s_{i,h}$. At the end of day h, they update their empirical estimate, $\hat{\Gamma}_{i,h}$, of the aggregate state-action distribution of other agents.

Fictitious play is used as a heuristic due to its theoretical guarantee of convergence to the original single-day traffic equilibrium as the number of days approaches infinity. The best responses are solved iteratively using Bayesian optimization with Gaussian processes. Specifically, Algorithm 3 leverages the MDTMC simulator

introduced in Algorithm 4 as a black-box environment for the multimodal bathtub model and iteratively solves the one-step optimization. Note that the approximate stationary solution obtained through this method is a Nash equilibrium, distinct from the MPE or SMFE defined in this paper. This distinction has also been validated through our numerical experiments.

Algorithm 4: Subroutine for Simulating MDTMC with Representative Agents

Input: (1) Representative agents' states $s_1, \ldots, s_N$; (2) Representative agents' actions $a_1, \ldots, a_N$; (3) Individual commuter state transition $P(\cdot|s, a, \Gamma)$; (4) Number of samples for commuter subgroups K;

Compute the population distribution Γ based on $s_1, \ldots, s_N$ and $a_1, \ldots, a_N$;

Determine system congestion for all modes by Algorithm 1;

for $i \in \{1, \ldots, N\}$ **do**

  for $j \in \{1, \ldots, K\}$ **do**

    Sample the individual state $s_i \sim \Gamma_j$ and the individual action $a_i \sim \pi_j$;

    Observe the group-specific cost c_{ij} and the next individual state $s_i' \sim P(\cdot|s_i, a_i)$;

  end

  Compute the average cost $c_i = \frac{1}{K} \sum_{i \sim j} c_{ij}$ and the next state distribution of the group s_i';

end

D.2. Benchmark 2: MARL with $Q(\lambda)$

Our second benchmark adopts a standard MARL framework in which we model N representative agents with an augmented *state decomposer* for inferring the aggregate daily traffic conditions. Because this state decomposer functions analogously to a generalized mean-field, we also denote it by Γ to avoid notational ambiguity. This MARL formulation with a state decomposer enhances learning stability across day-to-day transitions while maintaining decentralized execution for individual travel decisions. Given the two-dimensional state space, we partition it uniformly by population density so that each representative agent corresponds to a fraction $1/N$ subgroup of the total population. This population discretization is consistent with the finite-difference approximation used to simulate multimodal bathtub dynamics.

Multi-Agent $Q(\lambda)$ with State Decomposer. Building on this discretized N-agent model, we employ a centralized training and decentralized execution paradigm. On day h, every agent $i \in \{1, \ldots, N\}$ observes their own subgroup state $s_{i,h}$ together with a public decomposer state Γ_h. Unlike the conventional value decomposition framework, we propose a direct state decomposition framework (see Figure 10). The decomposer consists of two components: a continuous monotonic function g_ϕ that aggregates per-agent policies, and a mapping function h_ϕ that combines joint aggregates and local states to produce a compact per-agent latent representation $z_{i,h}$ defined as

$$z_{i,h} = h_\phi\big(s_{i,h}, g_\phi(\Gamma_h), \big),$$

During the centralized training phase, the decomposer functions are jointly updated based on agents' temporal-difference losses, and the updated latent state is distributed to all agents at the beginning of day h. In the decentralized execution phase, each agent maintains its own parameterized value function $Q_i(z, a; \theta_i)$, where the latent state depends solely on locally available information. Here, θ_i denotes the hyperparameter vector of the i-th agent's Q-value-network.

Eligibility traces and λ-updates. Let $c(s, a, \Gamma)$ denote the average subgroup-level travel cost produced by the bathtub traffic dynamics as in (2). Define the per-agent temporal difference error

$$\delta_{i,h} \;=\; c\big(s_{i,h}, a_{i,h}, \Gamma_h\big) + \gamma \min_{a'} Q_i\big(z'_{i,h+1},\, a';\, \theta_i^-\big) - Q_i\big(z_{i,h},\, a_{i,h};\, \theta_i\big),$$

where $z'_{i,h+1} = h_{\phi^-}\big(s_{i,h+1},\, g_{\phi^-}(\Gamma_{h+1})\big)$ uses the target decomposer parameters ϕ^-. The update of $Q(\lambda)$ follows a per-agent eligibility trace $\mathbf{e}_i \in \mathbb{R}^{|\theta_i|}$:

$$\mathbf{e}_i \;\leftarrow\; \gamma\lambda\, \mathbf{e}_i \;+\; \nabla_{\theta_i} Q_i\big(z_{i,h}, a_{i,h}; \theta_i\big), \qquad \theta_i \;\leftarrow\; \theta_i \;+\; \alpha\, \delta_{i,h}\, \mathbf{e}_i,$$

with soft target updates for θ_i^- and ϕ^-. We update ϕ using the summed temporal difference loss to ensure stable latent variables $z_{i,h}$ as

$$\phi \;\leftarrow\; \phi \;-\; \alpha_\phi\, \nabla_\phi \sum_{i=1}^{N} \delta_{i,h}^2.$$

Algorithm 5: Multi-Agent $Q(\lambda)$ with State Decomposition (Representative Agents)

Input: (1) MDTMC simulator $\mathcal{G}$ in Algorithm 4; (2) Episodes num_episodes bounded by horizon H; (3) Learning rates α, α_ϕ, discount γ, trace parameter $\lambda \in [0,1]$, target rate τ_{tgt}; (4) Decomposer $D_\phi = (g_\phi, h_\phi)$ with target D_{ϕ^-}; per-agent $Q_i(\cdot;\theta_i)$ with target θ_i^-; (5) Boltzmann temperature τ.

for episode $= 1$ **to** num_episodes **do**
  Obtain initial subgroup states $\{s_{1,0}, \ldots, s_{N,0}\}$ and aggregate Γ_0; $c_0 \leftarrow g_\phi(\Gamma_0)$; $z_{i,0} \leftarrow h_\phi(s_{i,0}, c_0)$ for all i; initialize per-agent traces $\mathbf{e}_i \leftarrow \mathbf{0}$;
  for $h = 0$ **to** $H-1$ **do**
    for *agent* $i = 1$ **to** N **do**
      Sample $a_{i,h}$ from Boltzmann$(Q_i(z_{i,h}, \cdot; \theta_i), \tau)$;
    end
    Execute $\{a_{i,h}\}_{i=1}^N$ in $\mathcal{G}$ to obtain next states $\{s_{i,h+1}\}$, aggregate Γ_{h+1}, costs $\{c_{i,h}\}$;
    $c'_{h+1} \leftarrow g_{\phi^-}(\Gamma_{h+1})$; $z'_{i,h+1} \leftarrow h_{\phi^-}(s_{i,h+1}, c'_{h+1})$ for all i;
    for *agent* $i = 1$ **to** N **do**
      $\delta_{i,h} \leftarrow c_{i,h} + \gamma \min_{a'} Q_i(z'_{i,h+1}, a'; \theta_i^-) - Q_i(z_{i,h}, a_{i,h}; \theta_i)$;
      $\mathbf{e}_i \leftarrow \gamma\lambda\, \mathbf{e}_i + \nabla_{\theta_i} Q_i(z_{i,h}, a_{i,h}; \theta_i)$;
      $\theta_i \leftarrow \theta_i + \alpha\, \delta_{i,h}\, \mathbf{e}_i$;
    end
    `// Centralized decomposer update (no traces on` ϕ`)`
    $\phi \leftarrow \phi - \alpha_\phi \nabla_\phi \sum_{i=1}^N \delta_{i,h}^2$;
    `// Soft target updates`
    $\theta_i^- \leftarrow (1 - \tau_{\text{tgt}})\, \theta_i^- + \tau_{\text{tgt}}\, \theta_i$ for all i;;
    $\phi^- \leftarrow (1 - \tau_{\text{tgt}})\, \phi^- + \tau_{\text{tgt}}\, \phi$;;
    `// Roll to next day`
    $c_{h+1} \leftarrow g_\phi(\Gamma_{h+1})$; $z_{i,h+1} \leftarrow h_\phi(s_{i,h+1}, c_{h+1})$ for all i;
  end
end

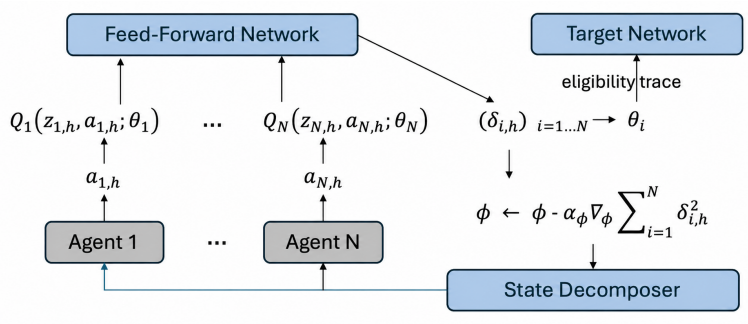


Figure 10 Overview of the Benchmark 2 MARL framework

D.3. Comparative Numerical Experiments

We conduct a series of numerical experiments covering a broad range of look-ahead horizons represented by λ values. Specifically, these experiments reveal how the representative agents' departure-time choices evolve in the early and later stages of learning processes. This offers insight into the congestion patterns that emerge when commuters place different emphases on immediate versus cumulative rewards. By comparing outcomes under strongly short-sighted and long-horizon settings, we show how the interplay between short-term decision-making and overall traffic efficiency can lead to different equilibria. In particular, we highlight how policies that undervalue future congestion effects may result in suboptimal system performance, whereas those that give greater weight to long-term returns tend to yield more stable and efficient traffic flows.

Experiment setting: We train the Multi-Agent $Q(\lambda)$ algorithm, denoted as MA-$Q(\lambda)$ in figures, with eligibility trace decays $\lambda = 0.2, 0.5, 0.8$. The greater value of λ indicates that agents place more emphasis on long-term rewards compared to short-term rewards. This comparative study incorporates the fictitious play method described in Benchmark 1, with all numerical experiments initialized from the same state on day $h = 1$. These models are compared against a fictitious play approach in Benchmark 1 with the same initial state. To ensure consistency in comparison, both setups employ $N = 4$ representative agents in the partition of the state space, given the complexity of repeatedly calculating the equilibrium solutions. We benchmark these methods against the HPG algorithm's training performance, shown in Figure 5c, under the same experiment configurations introduced in Section 4.1 with $\phi_{1,0} = 0.2$, $\phi_{0,1} = 0.4$, and $p = 0.2$. The episode length H is set to 80 to match the HPG outer-loop steps. The learning rate of MA-$Q(\lambda)$ is fixed at 10^{-3}, and the algorithms are trained for 50 episodes. The resulting comparison regarding the average daily travel cost per mile $\mathbb{C}$ of commuters is illustrated in Figure 11.

Over time, higher λ values enable agents to better anticipate congestion patterns in future days and refine their current departure time and mode choice decisions. From the comparison, we can observe that when agents use more long-term learning policies, the system yields more efficient flows with smaller average travel costs. In contrast, more short-sighted policies often fail to internalize future congestion effects related to collective decisions during the current day, leading to less efficient equilibria.

Another important observation is that multi-agent RL training is significantly slower than fictitious play (Benchmark 1) and mean-field-based learning (HPG). The difference in the number of days is over a hundredfold, highlighting the importance of designing efficient learning algorithms for multimodal transportation

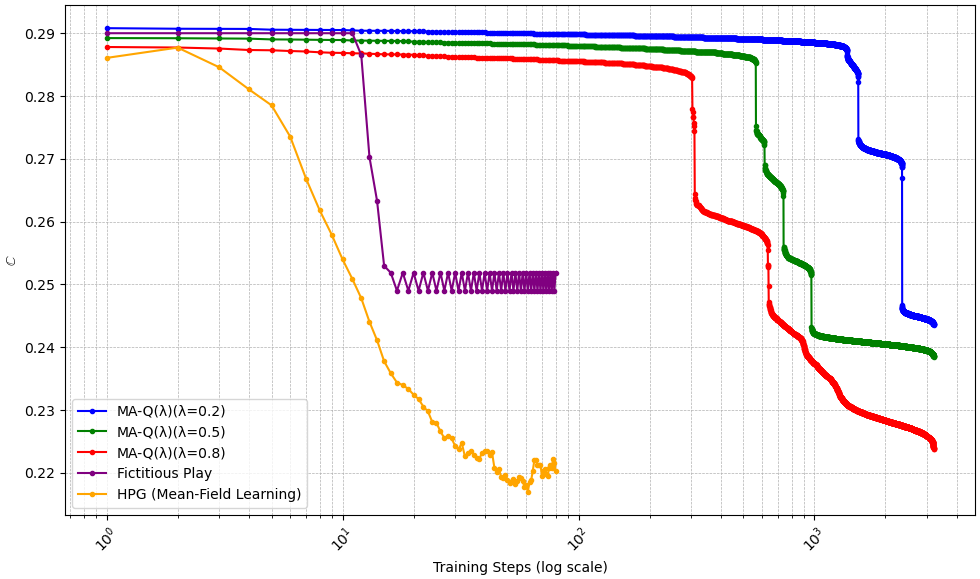


Figure 11 **Performance comparison of the average daily travel cost per mile $\mathbb{C}$ for fictitious play with myopic agents (Benchmark 1), MA-$Q(\lambda)$ with $\lambda \in \{0.2, 0.5, 0.8\}$ (Benchmark 2), and HPG algorithms. The logarithmic x-axis denotes the length of horizon (in days h).**

systems. Indeed, fictitious play (plotted in purple) converges quickly but settles at a higher cost equilibrium solution, illustrating how purely myopic decision-making can trap the system in congested states. Note that this equilibrium solution is different from the MPE or SMFE defined in this paper. Meanwhile, the HPG algorithm (in orange) leverages a mean-field viewpoint that inherently accounts for longer-term value functions, allowing it to converge faster and more robustly to lower travel costs at the stationary state. These comparative results validate the advantages of integrating global information-sharing and extended planning horizons in multi-modal transportation systems.

D.4. Robustness under relaxed stationarity assumptions

The baseline MDTMC formulation assumes a stationary day-to-day environment: the transition kernel $P: \mathbb{S} \times \Omega \times \mathbb{A}^N \to \Delta(\mathbb{S} \times \Omega)$ is fixed, and the travel cost function is time-invariant. Under these assumptions, the generalized mean field Γ converges to a stationary distribution when commuters follow the optimal policy, yielding a stationary MPE. In practice, however, cost parameters and transition probabilities may vary across days due to demand fluctuations, weather, disruptions, or other evolving traffic conditions. These deviations can be represented by augmenting the state space with additional time-varying variables, such as cost coefficients or exogenous factors. If these variables follow stationary dynamics and are included in the augmented state, the model remains Markovian. If the underlying process is genuinely nonstationary, convergence to a stationary SMFE is no longer guaranteed by the current theory.

To examine the numerical robustness of HPG under such deviations, we introduce controlled time-varying perturbations to both the daily cost coefficients and the transition kernel. We use the same baseline setting

as in Figure 5, with $\phi_{1,0}=0.2$, $\phi_{0,1}=0.4$, $p=0$, and $H=80$. Let $\theta_h=(\alpha_h,\beta_h,\gamma_h)$ and $\theta_0=(\alpha_0,\beta_0,\gamma_0)$. The perturbed cost coefficients evolve as

$$\theta_h=0.75\theta_{h-1}+0.25\theta_0+\epsilon_h^\theta(\lambda_c),\qquad \epsilon_h^\theta(\lambda_c)=\lambda_c\left(0.03\theta_0\odot r_h+0.04\theta_0\odot\eta_h\right),$$

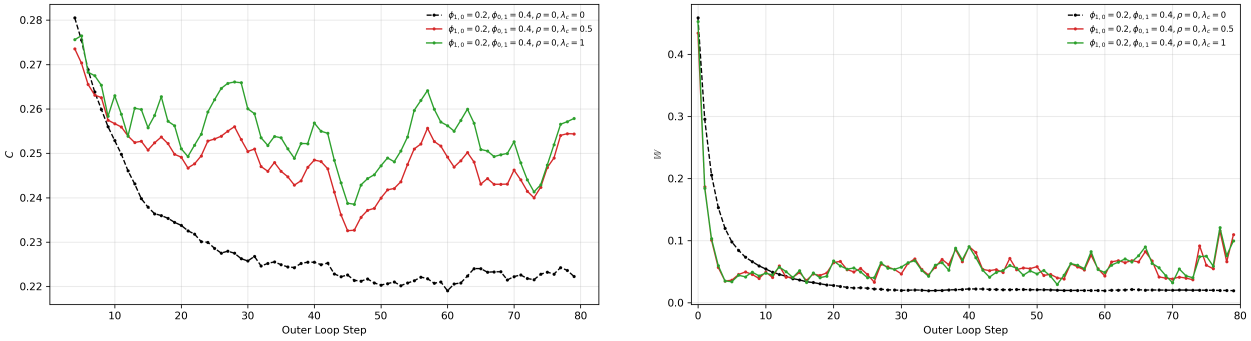
where $\lambda_c\in\{0,0.5,1.0\}$ controls perturbation intensity, r_h is a deterministic seasonal component, η_h is a zero-mean Gaussian innovation, and $\odot$ denotes element-wise multiplication. The case $\lambda_c=0$ recovers the original stationary baseline.

We also perturb each baseline transition probability vector by adding a zero-mean Gaussian residual $\epsilon_h^P(\lambda_c)$ with scale $0.02\lambda_c$:

$$P_h(\cdot\mid s_h,a_h,\Gamma_h)=\text{Normalize}\left[P_0(\cdot\mid s_h,a_h,\Gamma_h)+\epsilon_h^P(\lambda_c)\right],$$

where negative entries are clipped before normalization to preserve the probability simplex constraint.

Figure 12 reports the average daily travel cost per mile $\mathbb{C}$ and the equilibrium error $\mathbb{W}$, measured by the W_1-Wasserstein distance between mean-field distributions on consecutive days. When $\lambda_c=0$, the average cost decreases smoothly and the equilibrium error converges to a low level. When $\lambda_c=0.5$ or 1.0, both metrics exhibit persistent fluctuations, indicating that time-varying costs and transitions affect local learning performance. Nevertheless, the equilibrium error remains bounded and does not diverge, while the average cost stays within a comparable range across perturbation levels. These results suggest that, although HPG is developed under stationary assumptions, it can continue to track evolving mean-field dynamics under moderate day-to-day perturbations.



(a) Average daily travel cost per mile $\mathbb{C}$. (b) Equilibrium error $\mathbb{W}$.

Figure 12 **Performance of HPG under time-varying perturbations. The case $\lambda_c=0$ recovers the original stationary baseline, while $\lambda_c=0.5$ and $\lambda_c=1.0$ introduce increasing residual perturbations in both cost coefficients and transition probabilities.**

References

Gibbs AL, Su FE, 2002 *On choosing and bounding probability metrics. Intern. Stat. Rev.* 70(3):419–435.

Guo RY, Yang H, Huang HJ, 2018 *Are we really solving the dynamic traffic equilibrium problem with a departure time choice? Transp. Sci.* 52(3):603–620.

Jin WL, 2020b *Stable day-to-day dynamics for departure time choice. Transp. Sci.* 54(1):42–61.

Li M, Zhou X, Rouphail NM, 2017 *Quantifying travel time variability at a single bottleneck based on stochastic capacity and demand distributions. J. Intell. Transp. Syst.* 21(2):79–93.

Nowak AS, 2003 *On a new class of nonzero-sum discounted stochastic games having stationary nash equilibrium points. Int. J. Game Theory* 32(1):121.

Whitt W, 1980 *Representation and approximation of noncooperative sequential games. SIAM J. Control and Opt.* 18(1):33–48.

Yuan R, Gower RM, Lazaric A, 2022 *A general sample complexity analysis of vanilla policy gradient. AISTATS*, 3332–3380 (PMLR).