2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Mean-Field Learning for Day-to-Day Departure Time Choice with Mode Switching

Ben Wang[1], Qi Luo[2], and Yafeng Yin[3]

Abstract— **Understanding travelers' day-to-day departure time choice (DDTC) is vital for managing traffic congestion, especially in multi-modal transportation systems. While providing real-time traffic information and alternative trip plans brings convenience to travelers, their collective travel patterns may conversely lead to unstable traffic equilibrium states. We investigate a DDTC problem with mode switching in this paper. A group of heterogeneous agents can adaptively choose their modes and departure times to minimize total travel costs in a dynamic game. Using a customized hierarchical soft actor-critic (HSAC) algorithm with a continuum approximation of other agents, the traffic dynamics will converge to an approximate Markovian Perfect Equilibrium (MPE). Our findings also shed light on changes in long-term travel behavior due to the widespread deployment of emerging mobility and travel information technology. This approach serves as a foundation for promoting intelligent travel plans through adaptive traffic control policies.**

I. INTRODUCTION

Transportation agencies are now adopting multi-modal transportation solutions [1] that provide connected travelers with unified access to various modes of transportation [2]. However, assessing the long-term effects of introducing a new mode to an existing traffic network is challenging because the shift of travelers' collective travel decisions will perturb the traffic dynamics at equilibrium. In this process, connected travelers can adaptively observe traffic patterns and strategically modify their schedules to reduce their expected travel costs. In order to understand general patterns in a complex and dynamical transportation system encompassing massive strategic travelers, researchers often investigate day-to-day trip planning in aggregate network models like bottleneck or bathtub models [3], [4]. While these aggregate models do not estimate traffic dynamics at the link level, they prove to be effective surrogate models for drawing systematic implications in early-stage urban planning and policy-making [5]. In aggregate multi-modal networks [6], travelers' long-term mode-switching behavior and DDTC are interdependent decisions. Their mode choices determine the population split and the potential level of traffic congestion; overloading one mode will then stimulate travelers to depart at different times to avoid waiting in traffic jams. The interaction between these two decisions is notably critical for the early adoption of multi-modal transportation systems because travelers may not know the traffic patterns and must gradually learn the most efficient travel plans. Hence, this work focuses on a fundamental question: How will learning agents' mode and departure time choice evolve over days without prior information about system dynamics?

The difficulty of characterizing the day-to-day traffic dynamics when travel schedules vary is well-known [7], particularly due to the instability in finding equilibrium solutions under specific schedule deviation policies [8], [9]. This instability is further compounded by the fact that heterogeneous travelers (agents) do not have prior information about the overall population distribution, which changes daily due to mode switching. In this case, identifying optimal policies for agents to arrive at their desired times relays on the assumption that traffic congestion patterns, which depend on the number of travelers on the road, remain stable at equilibrium. However, since each agent has multiple transportation mode options, and the daily number of on-road agents per mode fluctuates, changes in collective travel behavior disrupt this equilibrium.

This research proposes a novel approach to model day-to-day mode switching and departure time choices through a dynamic game with learning agents. Our approach leverages the extension to a continuum of agents to guarantee stable solutions for the departure time choice, as described in [9]. We approximate other agents in the system through a generalized mean-field representation. This mean-field approximation enables us to use reinforcement learning (RL) techniques to study the evolution of traffic patterns as heterogeneous agents adapt their travel plans. More specifically, this study captures how agents can use RL algorithms to determine their daily mode of transportation and departure times in a simplified multi-modal transportation system.

A. Related Work

The analysis of within-day departure time dynamics based on aggregate network models [10] is a crucial component of traffic congestion predictions. The bathtub model [11], [12] is widely used to describe the macroscopic evolution of traffic patterns [4] affected by travelers' departure time choices. The interaction between travelers is modeled by a delay differential equation, in which a network fundamental diagram (NFD) [13], [14] characterizes the relationship between the average speed and the density. Common strategies

*This work was partially supported by research grants from National Science Foundation (CMMI-1904575 and CMMI-2308750)

[1]Ben Wang is with the Department of Industrial and Operations Engineering, College of Engineering, University of Michigan, 1205 Beal Ave, Ann Arbor, MI 48109 USA papaver@umich.edu

[2]Qi Luo is with the Department of Industrial Engineering, College of Engineering, Computing and Applied Sciences, and School of Mathematical and Statistical Sciences, Clemson University, 277B Freeman Hall, Clemson, SC 29634, USA qluo2@clemson.edu

[3]Yafeng Yin is with the Department of Civil and Environmental Engineering and Industrial and Operations Engineering, University of Michigan, 2350 Hayward St, Ann Arbor, MI 48109 USA yafeng@umich.edu

that treat discontinuities in departure rates at the bathtub model's equilibrium solution include (1) approximating the delay differentiable solution [4], (2) developing numerical methods satisfying unique equilibrium utility conditions [15], and (3) expanding the population to continuum of trip length [10] and relaxing the distributional assumption [16]. Our work aligns with the MFG setting [17] that considers random desired arrival times and trip lengths to obtain departure time equilibrium solutions under the most relaxed setting.

The perturbation of the equilibrium state of DDTC has attracted growing interest, but most previous studies relies on simulation-based approaches. Numerical simulations in [18] show the impacts of elastic demand on daily traffic conditions, where travelers may adjust departure times, switch modes, or divert to alternative routes based on the provided information, in which travelers' choices follow a nested logit model and day-to-day adjustments follows a Markovian process. [19] examines the myopic adjustment and the learning heuristics for travelers' mode switching policies. The numerical experiments reveal a counter-intuitive phenomenon that long-memory learning models may lead to unstable day-to-day dynamics. Analytical results for DDTC are limited to the point queue model [7], [20]. For example, [8] proposes a stable day-to-day dynamical system based on a backward choice, cost-balancing, and scheduling cost-reducing principle. [9] proposes an alternative stable dynamical system in which travelers' local switching policy only depends on the departure times. Our work considers DDTC in the bathtub model, which captures hyper-congestion effects and connects to a large body of NFD literature that has been validated by empirical traffic data [14].

B. Contributions

Our main contributions include (1) proposing a generalized mean-field game (GMFG) model for DDTC with mode-switching that warrants the approximate MPE solution; (2) designing an HSAC algorithm that learns the optimal mode and departure time policies based on traveler characteristics; (3) providing insights into the evolution of mode splitting and predicted traffic patterns after the introduction of a new mode into a multi-modal transportation system.

II. DDTC WITH MODE SWITCHING PROBLEM

Problem statement for N-agent games. Consider travelers (agents) indexed by $i \in \mathcal{N} \triangleq \{1, 2, \dots, N\}$ commute to work over a horizon of $h \in \mathcal{H} \triangleq \{0, 1 \dots\}$. Heterogeneous agents with varying trip lengths and desired arrival times make joint travel mode and departure time decisions on the day h to minimize their expected discounted cumulative travel costs over the horizon. Since the network congestion level depends on agents' collective decisions, we model DDTC with mode switching as a Markovian (dynamic) game. For simplicity of analysis, the remainder only considers driving and an alternative on-demand transit mode, which can be generalized to more commuting options. At the beginning of each day, N agents independently make their mode and departure time decisions based on their desired arrival time and trip length without sharing this information with others. Let $m_{i,h}$ denote the travel mode taken by agent i on the day h. $m_{i,h} = 1$ denotes that agent i chooses to drive and $m_{i,h} = 0$ for taking the transit mode. Let $t_{i,h} \in \mathcal{T}^d \triangleq [0, \overline{t}]$ denote the agent i's departure time on the day h. The trip length and the desired arrival time of agent i on the day h is represented by $L_{i,h} \in \mathcal{L} \triangleq [0, \overline{L}]$ and $\xi_{i,h} \in \mathcal{X} \triangleq [0, \overline{\xi}]$ respectively, which may shift according to the observed travel costs from previous days. The *individual state* $s_{i,h} = (L_{i,h}, \xi_{i,h}) \in \mathbb{S} \triangleq \mathcal{L} \times \mathcal{X}$ includes the trip length and the desired arrival time. The individual action $a_{i,h} = (m_{i,h}, t_{i,h}) \in \mathbb{A} \triangleq \{0, 1\} \times \mathcal{T}^d$ includes the mode choice and the departure time. Thus, the state profile on the day h is denoted by $\mathbf{s}_h \triangleq \{s_{i,h}\}_{i \in \mathcal{N}}$ and the joint actions on the day h is denoted by $\mathbf{a}_h \triangleq \{a_{i,h}\}_{i \in \mathcal{N}}$.

An *aggregate state* is introduced to simplify the state space, which would otherwise grow exponentially with the number of agents, while still fully characterizing the trip distributions. Let $\mathcal{N}_{l,h}$ denote a set of agents who drive with a trip length equal to l on the day h, and let $N_{l,h}$ be the number of such agents. The *trip length mass* ℓ_h, is a distribution of driving agents, where $\ell_{l,h} = N_{l,h}$ for any $l \in \mathcal{L}$. It serves as an aggregate state at the beginning of day h, and $\mathbb{L}$ denotes the space of all possible trip length mass distributions. Each agent in this Markovian game only observes the individual state $s_{i,h}$ and the aggregate state ℓ_h to determine their mode choice and departure time on the day h.

Traffic dynamics under the collected mode and departure time decisions are modeled by a generalized bathtub model [16] with heterogeneous travelers. Agents who choose to drive on the day h and depart at $t_{i,h}$ arrive at $t_{i,h} + T(t_{i,h}) \in \mathcal{T} \triangleq [0, \overline{t}']$, and their travel times are affected by the traffic congestion of the capacitated road network. The system density at each time x is given by $k_h(x) = \sum_{i \in \mathcal{N}_h} \mathbb{1}_{[t_{i,h}, t_{i,h} + T(t_{i,h}))}(x)$, where $\mathbb{1}_A(x)$ is an indicator function. The relationship between the traffic velocity $v_h(x)$ and the system density $k_h(x)$ follows a NFD equation, as $v_h(x) = V(k_h(x))$. To calculate the agent i's travel time, we consider a virtual agent who drives from $x = 0$ to $\overline{t}'$ on the day h. Her travel distance at time x is $\psi_h(x; \ell_h) := \int_0^x V(k_h(u))du$. Since ψ_h is invertible, the travel time of agent i on the day h with the departure time $t_{i,h}$ is expressed as $T(t_{i,h}) = \psi_h^{-1}(L_{i,h} + \psi_h(t_{i,h}; \ell_h)) - t_{i,h}$. The daily travel cost is calculated by an $\alpha - \beta - \gamma$ function:

$$c(s_{i,h}, \ell_h, \mathbf{a}_h) = \begin{cases} \overline{c} L_i, & \text{if } m_{i,h} = 0, \\ \alpha T(t_{i,h}) + c_{schedule}(\xi_{i,h}, t_{i,h}), & \text{if } m_{i,h} = 1, \end{cases} \quad (1)$$

where $\overline{c}$ and α are travel time cost coefficients for transit and driving, respectively; the scheduled cost $c_{schedule}(\xi_{i,h}, t_{i,h})$ is given by

$$c_{schedule}(\xi_{i,h}, t_{i,h}) = \beta(\xi_{i,h} - t_{i,h} - T(t_{i,h}))^+ + \gamma(\xi_{i,h} - t_{i,h} - T(t_{i,h}))^-. \quad (2)$$

Here β, and γ are the cost coefficients for early arrival and late arrival, respectively. The implicit assumptions made are

(1) the average transit travel time is independent of their departure time [21], (2) $\alpha > \min\{\beta, \gamma\}$, meaning that agents prefer to avoid traffic congestion. The state transition of each agent i is modeled by a stationary function $P : \mathbb{S} \times \mathbb{L} \times \mathbb{A}^N \to \Delta(\mathbb{S} \times \mathbb{L})$, where $\Delta(X)$ denotes the set of probability distributions on X. Given the joint action $\mathbf{a}_h$ on the day h, the transition probability from the individual state $s_{i,h}$ and the trip length mass ℓ_h to the next individual state $s_{i,h+1}$ and trip length mass ℓ_{h+1} is given by $P(s_{i,h+1}, \ell_{h+1} | s_{i,h}, \ell_h, \mathbf{a}_h)$.

We study stationary Markov policies for DDTC with mode switching in an infinite horizon N-agent stochastic game (a discounted non-zero-sum stochastic game), denoted as $\pi_i : \mathbb{S} \times \mathbb{L} \to \Delta(\mathbb{A})$. Let Π denote the Markov policy space. A *policy profile* $\boldsymbol{\pi} = \pi_1 \times \cdots \times \pi_N$ maps a state profile and a trip length mass to a joint action. Given $\boldsymbol{\pi}$, the transition probability and the daily travel cost are functions of the current individual state $s_{i,h} \in \mathbb{S}$ and the aggregate state $\ell_h \in \mathbb{L}$, denoted as $P^{\boldsymbol{\pi}}(s_{i,h+1}, \ell_{h+1} | s_{i,h}, \ell_h)$ and $c^{\boldsymbol{\pi}}(s_{i,h}, \ell_h)$, respectively. Agents aim to minimize their expected discounted cumulative cost over an infinite horizon by adaptively updating their mode and departure time choices. Denote the initial individual state as s, the initial trip length mass as ℓ, and the policy profile as $\boldsymbol{\pi}$. The expected discounted cumulative cost of the agent i, denoted by $J^{\pi_i, \pi_{-i}} : \mathbb{S} \times \mathbb{L} \to \mathbb{R}$, is defined as:

$$J^{\pi_i,\pi_{-i}}(s,\ell) := \mathbb{E}\Bigg[\sum_{h=0}^{\infty} \lambda^h c(s_{i,h}, \ell_h, \mathbf{a}_h) \Bigg| s_{i,0} = s, \ell_0 = \ell, \mathbf{a}_h \sim \boldsymbol{\pi}(\mathbf{s}_h, \ell_h), (s_{i,h+1}, \ell_{h+1}) \sim P^{\boldsymbol{\pi}}(s_{i,h}, \ell_h)\Bigg], \quad (3)$$

where $\lambda \in (0, 1)$ is the discount factor.

Maskin [22] shows that presuming Markov properties in policy profiles can greatly simplify the analysis of equilibrium solutions. Hence, a stationary Markov policy profile $\boldsymbol{\pi}$ is considered a Markovian perfect equilibrium (MPE) of the N-agent DDTC if, for each agent $i \in \mathcal{N}$, any individual state $s \in \mathbb{S}$, and any aggregate state $\ell \in \mathbb{L}$, the following inequality holds for all $\hat{\pi}_i \in \Pi$:

$$J^{\pi_i,\pi_{-i}}(s,\ell) \leq J^{\hat{\pi}_i,\pi_{-i}}(s,\ell). \quad (4)$$

Although using an aggregate state simplifies the decision-making process of each agent by requiring less information about the population, finding an MPE for the N-agent DDTC problem is inherently complex for several reasons. Firstly, the day-to-day and within-day traffic dynamics are coupled, resulting in a complex doubly dynamic system where each agent is influenced by the decisions of all other agents. Secondly, The state space and action space of the stochastic game grow rapidly with the number of agents, and ℓ_h is a combinatorially explosive state of $\mathcal{N}_h$ that depends on each agent's mode choice, making it challenging to solve the problem exactly. Thirdly, ensuring stability in day-to-day departure time dynamics is a difficult task, especially when considering the presence of learning agents [7], [8]. To overcome these challenges, we use the mean-field approximation inspired by a continuum of agents for within-day departure time choice [17].

GMFG formulation for DDTC. The computation of MPE in the complex N-agent stochastic game for DDTC with mode switching can be characterized by assuming indistinguishable and interchangeable agents as a mean field. The *trip length distribution* $\mu_h(l) \in \Delta(\mathcal{L})$ represents the aggregate pattern of the traffic demand as the number of agents N tends to infinity, which is the limit of the *trip length mass* as $\mu_h(l) := \lim_{N\to\infty} \frac{\sum_{j=1, j\neq i}^{N} \mathbb{1}_{L_{j,h}=l, m_{j,h}=1}}{N}$. The *departure time distribution* $\tau_h(t) \in \Delta(\mathcal{T}^d)$ represents the aggregate departure flow of drivers on the day h, which is the limit of the departure time profile as $\tau_h(t) := \lim_{N\to\infty} \frac{\sum_{j=1, j\neq i}^{N} \mathbb{1}_{t_{j,h}=t, m_{j,h}=1}}{N}$. The mean-field $\Gamma_h \in \mathbb{G} \triangleq \Delta(\mathcal{L} \times \mathcal{T}^d)$ is introduced as travelers' joint distribution of trip length and departure time, with marginal distributions μ_h and τ_h. With a large population simultaneously making departure time and mode choice decisions repeatedly, the mean-field of agents converges to a stable distribution as marginal changes in the individual state and action are averaged out. On the day h, the characteristic agent makes near-optimal decisions for $a_h \in \mathbb{A}$ using an *oblivious* policy $\pi(s_h, \Gamma) : \mathbb{S} \times \mathbb{G} \to \Delta(\mathbb{A})$, receives the daily travel cost $c(s_h, a_h, \Gamma)$, and experiences state transitions according to the dynamics $P(\cdot | s_h, a_h, \Gamma)$. Here, both $c : \mathbb{S} \times \mathbb{A} \times \mathbb{G} \to \mathbb{R}$ and $P : \mathbb{S} \times \mathbb{A} \times \mathbb{G} \to \Delta(\mathbb{S})$ are measurable functions and c is bounded. We call the policy *oblivious* [23] because it involves decisions made without the full knowledge of other agents. The expected discounted cumulative cost of the agent under the mean-field game framework is given by

$$J^{\pi}(s,\Gamma) := \mathbb{E}\Bigg[\sum_{h=0}^{\infty} \lambda^h c(s_h, a_h, \Gamma) \Bigg| s_0 = s, s_{h+1} \sim P(s_h, a_h, \Gamma), a_h \sim \pi(s_h, \Gamma)\Bigg], \quad (5)$$

The MPE of the N-agent problem is now approximated in the mean-field game framework:

Definition 1: A profile of (π^*, Γ^*) is an approximate MPE if the following conditions are satisfied.

1. Given the mean-field Γ^*, $\forall \pi$ and $\forall s \in \mathbb{S}$, it holds that
$$J^{\pi^*}(s, \Gamma^*) \leq J^{\pi}(s, \Gamma^*). \quad (6)$$
2. When the agent exercises the policy π^*, the joint distribution of $\{L_h, t_h | m_h = 1\}_{h=0}^{\infty}$ converges to Γ^*, which is generated by the dynamics $L_0 \sim \mu_0$, $a_h \sim \pi^*(s_h, \Gamma_h)$, and $s_{h+1} \sim P(s_h, a_h, \Gamma_h)$. Here, Γ_h is the empirical mean-field distribution on the day h.

III. Solution Method and HSAC Algorithm

Agents in a multimodal transportation system does not have prior information about daily travel costs and the state transition function, especially with newly added on-demand transit modes [1]. This section presents a soft actor-critic RL algorithm incorporating a hierarchical policy to overcome the challenge of computing approximate MPE for the DDTC

problem. Various RL methods for computing approximate equilibrium solutions in mean-field games have been proposed, such as Q-learning [24] proximal policy optimization [25], and actor-critic method [26].

However, the joint mode and departure time choices present a unique hierarchical structure of policies, which require the development of new RL algorithms for solving the approximate MPE. We propose an RL method called the Mean-Field Hierarchical Soft Actor-Critic (MF-HSAC) algorithm. MF-HSAC consists of an inner-loop RL step and an outer-loop mean-field updating step, which has two features: (1) utilizing the hierarchical policy of mode and departure time choices, MF-HSAC can significantly speed up the overall learning process; (2) the actor-critic method in the inner loop only accesses to the driver's trip length and departure time distributions compared to traditional RL methods requiring the entire state-action distributional information.

MF-HSAC Algorithm. Recall that the daily travel cost for transit modes is independent of the departure time decision, the inner loop of MF-HSAC focuses on finding the optimal departure time policy for driving agents through the Soft Actor-Critic (SAC) algorithm [27], [28]. Given the mean-field distribution Γ, the departure time policy (actor) for the driving mode is represented by $\pi_{\Gamma,\theta}(s)$, and the corresponding state-action value function (critic) is represented by $Q_{\Gamma,w}(s,t)$. The policy and the value function are parameterized by θ and w, respectively. After the critic value gets stable after several training episodes, the expected critic value, which represents the future value of driving, is compared with the future value of the alternative mode, denoted as $\overline{Q}$, and the mode is decided accordingly. Let $\nu_\Gamma(s)$ denote the mode policy, which is a Bernoulli distribution with the probability mass function (PMF) computed by $\textbf{softmax}_\omega\left(\mathbb{E}_{t\sim\pi_{\Gamma,\theta}}(Q_{\Gamma,w}(s,t)),\overline{Q}\right)$. Here, $\text{softmax}_\omega(x)_i = \frac{exp(x_i/\omega)}{\sum_j exp(x_j/\omega)}$ is the Boltzmann exploration operator with a temperature parameter ω. Without loss of generality, we assume the agent also follows $\pi_{\Gamma,\theta}(s)$ in the transit mode. The oblivious policy is then represented in a hierarchical form, with $\pi_\Gamma = \nu_\Gamma \circ \pi_{\Gamma,\theta}$. The traffic congestion $k(x)$ and the characteristic trip length $\psi(x)$ are computed in the discretized bathtub model given Γ, further determining the inner-loop traffic dynamics. We refer to this process as $\{c(\cdot,\cdot,\Gamma), P(\cdot,\cdot,\Gamma)\} \sim \Theta(\Gamma)$ and omit details of calculating traffic dynamics (see [17] for more information).

The outer loop of MF-HSAC involves sampling a state s from the population, making an action a according to the policy $\pi_\Gamma(s)$ learned from the previous step, observing the next state $s' \sim P(\cdot|s,a,\Gamma)$, and taking another action $a' \sim \pi_\Gamma(s')$. The mean-field distribution, Γ, is then updated to Γ' as an aggregate state-action distribution. Let $\Gamma' \sim \Phi(\pi_\Gamma,\Gamma)$ represent the whole updating process. The MF-HSAC algorithm is fully described in Algorithm 1, where α_θ and α_w are learning rates of SAC. Policy π_{Γ_h} is substituted by π^h for brevity of notation.

Algorithm 1: MF-HSAC

Input: Initial mean-field Γ_0, initial parameter of actor θ, initial parameter of critic w

for $h = 0, 1, \ldots:$ **do**
- 1. Determine the traffic dynamics: $c(\cdot,\cdot,\Gamma_h), P(\cdot,\cdot,\Gamma_h) \sim \Theta(\Gamma_h)$
- 2. Learn the departure time policy:
- $s \leftarrow s_0$
- $m \leftarrow 1$
- **for** $j = 0, 1, \ldots:$ **do**
  - choose departure time $t \sim \pi_{\Gamma_h,\theta}(s)$
  - Take action $a = (m,t)$ in s
  - Observe $s' \sim P(s,a,\Gamma_h)$ and $r = -c(s,a,\Gamma_h)$
  - $\sigma \leftarrow r + \lambda \max_t\{Q_{\Gamma_h,w}(s',\cdot)\} - Q_{\Gamma_h,w}(s,t)$
  - $\theta \leftarrow \theta + \alpha_\theta \frac{\delta \log \pi_{\Gamma_h,\theta}(t|s)}{\delta\theta}\sigma$
  - $w \leftarrow w + \alpha_w \frac{\delta Q_{\Gamma_h,w}(s,t)}{\delta w}\sigma$
- **end**
- 3. Determine the mode choice policy:
- $\nu_{\Gamma_h}(s) = \textbf{softmax}_\omega\left(\mathbb{E}_{t\sim\pi_{\Gamma_h,\theta}}(Q_{\Gamma_h,w}(s,t)),\overline{Q}\right)$
- 4. Update the mean-field distribution: $\Gamma_{h+1} \sim \Phi(\pi^h,\Gamma_h)$
- where $\pi^h = \nu_{\Gamma_h} \circ \pi_{\Gamma_h,\theta}$

end

IV. NUMERICAL EXPERIMENT

Experiment setup. We take a discrete setting for simplicity. Our numerical experiment is based on the downtown rush-hour setting in [15], [29]. The desired arrival time space $\mathcal{X}$ and the departure time space $\mathcal{T}^d$ are both defined on a discrete-time interval ranging from 0 to $\frac{1}{3}$-hour with an interval of 1-minute. The actual arrival time space $\mathcal{T}$ is defined as a time interval ranging from 0 to 1 hour. The trip lengths are either 0.5 miles or 1.0 miles; the space interval dL is 0.5 miles. The trip length for agents is revealed at the beginning and is assumed to be constant throughout the experiment. The desired arrival time ξ is adjusted accordingly as follows: (1) If the agent ran late on the previous day, $\xi' = \xi + dt$; (2) If the agent arrived early in the previous day, $\xi' = \xi - dt$; (3) If the agent arrives on time, ξ' remains the same. This adjustment reflects the alteration of the agent's expectation for the next arrival time based on their observations. A similar design is used in [9] to study the local stability of the bathtub model in DDTC, whereas our RL framework can handle any adjustment strategies. We assume the NFD equation $V(\cdot)$ takes Greenshields' relationship as it in [15] and has the minimal traffic velocity $v_{\min}$, described by $V(k) = \max\{v_f(1-k/\overline{k}), v_{\min}\}$, where $\overline{k}$ is the jam density. We set the free-flow speed v_f in urban networks to 15 miles per hour (mph), which refers to the morning rush hour free-flow speed in [15]. The minimum speed $v_{\min} = 5$ mph and the jam density $\overline{k} = 0.2$, indicating that the maximum capacity of the traffic system is 20% of the population on the road. Note that the congestion is measured

as the fraction of the total population rather than the number of agents.

The daily travel cost is defined as $\alpha - \beta - \gamma$ based on the study of Lamotte et al. [30], where α, β, and γ are set to 1 \$/h, 0.51 \$/h, and 2.06 \$/h, respectively. The unit cost of on-demand transit is $\bar{c} = 0.2$ \$/mile, which represents the travel cost at the velocity of $v_{\min}$. This indicates that transit mode may result in a higher travel time cost, it is considered a safer option in terms of schedule cost compared to driving. The state-action value of the transit mode, $\overline{Q}$, given the trip length L, is calculated as $\overline{Q}(L) := \frac{-\bar{c}L}{1-\lambda}$.

We conduct numerical experiments with both MF-HSAC and MF-HSAC-Tabular algorithms with 50 outer-loop iterations. In each iteration, the departure time policy is trained separately as there is no transition between trip lengths. The learning process consists of 10000 episodes, each lasting 10 days. The numerical experiment assumes that the population has a uniform distribution over trip length and desired arrival time. Furthermore, we assume that the initial mean-field distribution, Γ_0, is also uniformly distributed. The discount factor λ is 0.99. The departure time policy (actor), $\pi_{\Gamma,\theta}$, is parameterized using a Boltzmann policy with a linear approximation based on polynomial bases in the order of 3. The Q function in the driving mode (critic), $Q_{\Gamma,w}$, is parameterized using a linear approximation based on polynomial bases in the second order of 2 [31]. All weights are randomly initialized at the start of training. The learning rates of critic and actor modes are set to 0.1 and 0.01, respectively, and constant learning rate decay is applied every 100 episodes. The mode choice policy ν_Γ uses the temperature coefficient $\omega = 1.2$. The benchmark model deploys a tabular representation for the critic or actor and does *not* consider hierarchical policies, in which $Q_{\Gamma,w}(s,t)$ is replaced with a tabular function and referred to as MF-HSAC-Tabular.

In the outer loop, the mean-field distribution is updated with $N = 1000$ samples (i.e., the number of characteristic agents). Two metrics are used to assess the equilibrium, namely the average daily travel cost per mile of the driver $\mathbb{C}$ and the equilibrium error $\mathbb{W}$. The average daily travel cost per mile of the driver $\mathbb{C}$ is represented as $\mathbb{C}_h := \frac{\sum_i^N c_{i,h} \mathbb{1}_{m_{i,h}=1}}{\sum_i^N L_{i,h} \mathbb{1}_{m_{i,h}=1}}$, where $c_{i,h}$ and $L_{i,h}$ are simulated from $\Phi(\pi^h, \Gamma_h)$ at each step h.

The equilibrium error $\mathbb{W}$, which is a measure of the difference between the policy-induced expected state value in iteration h and $h-1$, is expressed as

$$\mathbb{W}_h = \mathbb{E}_{s\sim\zeta_h}[V^{\pi^h}(s)] - \mathbb{E}_{s\sim\zeta_{h-1}}[V^{\pi^{h-1}}(s)]. \quad (7)$$

To compute the policy-induced state value V^π, we first determine the invariant transition matrix P^π. This transition matrix is calculated by the balance equation $P^\pi(s,s') = \sum_{a\in\mathbb{A}} P(s'|s,a,\Gamma^\pi)\pi(a|s)$ for all $s, s' \in \mathbb{S}$, as well as the policy-induced mean-field Γ^π conditional on $\zeta(s)$ and $\pi(a|s)$. Here, ζ represents the invariant state distribution of the transition matrix P^π. Due to these equations are implicit, we can only solve the invariant transition matrix using an iterative method. Given the invariant transition matrix, the state-value function can be calculated by value iteration.

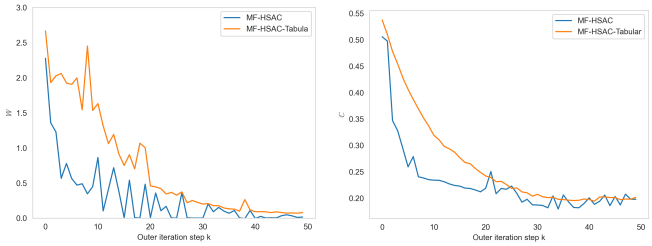

Fig. 1. Convergence Results of MF-HSAC: (a) Equilibrium Error, (b) Cost Comparison

Numerical results. As shown in Figure 1(a), the equilibrium error $\mathbb{W}$ of MF-HSAC and MF-HSAC-Tabular methods decrease to near-zero over the course of 50 iterations. The MF-HSAC method has a faster convergence rate compared to MF-HSAC-Tabular, particularly in the first 10 iterations. This trend is more evident in Figure 1(b), where the average daily travel cost per mile of the driver $\mathbb{C}$ decreases by roughly half of its initial value within the first 5 iterations, and remains stably small at around 0.2 after 50 iterations. The converged cost is close to that of transit cost $\bar{c}$. This observation matches the intuition that driving and the alternative mode are indifferent at the approximate MPE. Despite the faster convergence rate of MF-HSAC, MF-HSAC-Tabular appears to perform more stably.

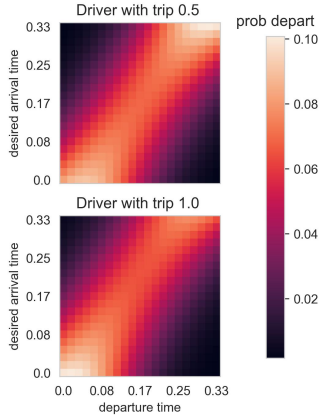

Fig. 2. DDTC solutions at approximate MPE

Recall that DDTC with bathtub model is not tractable [9], so no previous literature has discovered general equilibrium solutions except in special cases. Therefore, our results in Figure 2 present the departure time policy learned by the MF-HSAC algorithm for driving agents at the approximate MPE. The heat map shows the probability distribution of departure times based on the desired arrival time and trip length. The results indicate that for short trips ($L = 0.5$), the agent tends to depart close to their desired arrival time. However, the agent prefers to depart earlier for longer trips

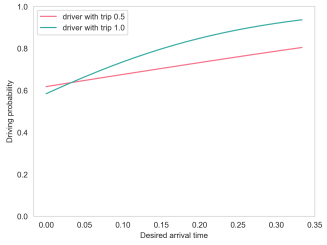

Fig. 3. Mode split at equilibrium solution

($L = 1.0$). Additionally, the probability of driving is higher for long-trip agents (see Figure 3), because transit with linear increasing cost in trip length becomes less favorable when traffic conditions are not overly congested. On the other hand, if long-trip agents aim to arrive earlier or undertake the risk of incurring high late penalties, on-demand transit becomes a more conservative and reliable option. The probability of driving increases when agents prefer delayed arrivals, as their longer planning horizon allows them to search for more cost-effective departure times in the face of unpredictable traffic.

V. CONCLUSIONS

In this study, we explore the DDTC problem with mode switching that captures the long-term travel behavior changes in multimodal transportation systems. The daily traffic dynamics affected by travelers' collective mode and departure time choices are modeled as a bathtub model, where agents' decision-making processes are formulated as a Markovian game with N heterogeneous agents. We introduce a generalized mean-field game approximation where the characteristic agent adopts a Markov policy. A continuum approximation of agents converges to a stationary mean-field when the number of agents and the planning horizon reach infinity. Hence, Markov mode and departure time policies admit approximate MPE solutions. Since travelers (agents) react distinctly to newly introduced modes, such as on-demand transit, we propose an MF-HSAC algorithm. Our numerical experiments demonstrated its effectiveness in finding approximate MPE, at which all modes are cost-indifferent but driving is favored by long-trip travelers.

REFERENCES

[1] S. Banerjee, C. Hssaine, Q. Luo, and S. Samaranayake, "Plan your system and price for free: Fast algorithms for multimodal transit operations," *Working paper, Cornell University, Ithaca, NY, USA*, 2021.

[2] C. Xiong, M. Shahabi, J. Zhao, Y. Yin, X. Zhou, and L. Zhang, "An integrated and personalized traveler information and incentive scheme for energy efficient mobility systems," *Transportation Research Part C: Emerging Technologies*, vol. 113, pp. 57–73, 2020.

[3] T. Yao, T. L. Friesz, M. M. Wei, and Y. Yin, "Congestion derivatives for a traffic bottleneck," *Transportation Research Part B: Methodological*, vol. 44, no. 10, pp. 1149–1165, 2010.

[4] R. Arnott, "A bathtub model of downtown traffic congestion," *Journal of Urban Economics*, vol. 76, pp. 110–121, 2013.

[5] Z.-C. Li, H.-J. Huang, and H. Yang, "Fifty years of the bottleneck model: A bibliometric review and future research directions," *Transportation research part B: methodological*, vol. 139, pp. 311–342, 2020.

[6] L. Balzer, M. Ameli, L. Leclercq, and J.-P. Lebacque, "Dynamic tradable credit scheme for multimodal urban networks," *Transportation Research Part C: Emerging Technologies*, vol. 149, p. 104061, 2023.

[7] R.-Y. Guo, H. Yang, and H.-J. Huang, "Are we really solving the dynamic traffic equilibrium problem with a departure time choice?," *Transportation Science*, vol. 52, no. 3, pp. 603–620, 2018.

[8] W.-L. Jin, "Stable day-to-day dynamics for departure time choice," *Transportation Science*, vol. 54, no. 1, pp. 42–61, 2020.

[9] W.-L. Jin, "Stable local dynamics for day-to-day departure time choice," *Transportation Research Part B: Methodological*, vol. 149, pp. 463–479, 2021.

[10] M. Fosgerau, "Congestion in the bathtub," *Economics of Transportation*, vol. 4, no. 4, pp. 241–255, 2015.

[11] W. Vickrey, "Congestion in midtown manhattan in relation to marginal cost pricing," *Technical Report. Columbia University, New York, NY*, 1991.

[12] W. Vickrey, "Congestion in midtown manhattan in relation to marginal cost pricing," *Economics of Transportation*, vol. 21, p. 100152, 2020.

[13] C. F. Daganzo and N. Geroliminis, "An analytical approximation for the macroscopic fundamental diagram of urban traffic," *Transportation Research Part B: Methodological*, vol. 42, no. 9, pp. 771–781, 2008.

[14] N. Geroliminis and J. Sun, "Properties of a well-defined macroscopic fundamental diagram for urban traffic," *Transportation Research Part B: Methodological*, vol. 45, no. 3, pp. 605–617, 2011.

[15] R. Arnott and J. Buli, "Solving for equilibrium in the basic bathtub model," *Transportation Research Part B: Methodological*, vol. 109, pp. 150–175, 2018.

[16] W.-L. Jin, "Generalized bathtub model of network trip flows," *Transportation Research Part B: Methodological*, vol. 136, pp. 138–157, 2020.

[17] M. Ameli, M. S. S. Faradonbeh, J.-P. Lebacque, H. Abouee-Mehrizi, and L. Leclercq, "Departure time choice models in urban transportation systems based on mean field games," *Transportation Science*, 2022.

[18] M. Ben-Akiva, A. De Palma, and P. Kanaroglou, "Dynamic model of peak period traffic congestion with elastic arrival rates," *Transportation Science*, vol. 20, no. 3, pp. 164–181, 1986.

[19] H. S. Mahmassani and G.-L. Chang, "Experiments with departure time choice dynamics of urban commuters," *Transportation Research Part B: Methodological*, vol. 20, no. 4, pp. 297–320, 1986.

[20] R.-Y. Guo, H. Yang, and H.-J. Huang, "The day-to-day departure time choice of heterogeneous commuters under an anonymous toll charge for system optimum," *Transportation Science*, 2023.

[21] M. Young, J. Allen, and S. Farber, "Measuring when uber behaves as a substitute or supplement to transit: An examination of travel-time differences in toronto," *Journal of Transport Geography*, vol. 82, p. 102629, 2020.

[22] E. Maskin and J. Tirole, "Markov perfect equilibrium: I. observable actions," *Journal of Economic Theory*, vol. 100, no. 2, pp. 191–219, 2001.

[23] G. Y. Weintraub, C. L. Benkard, and B. Van Roy, "Computational methods for oblivious equilibrium," *Operations research*, vol. 58, no. 4-part-2, pp. 1247–1265, 2010.

[24] X. Guo, A. Hu, R. Xu, and J. Zhang, "Learning mean-field games," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[25] Q. Xie, Z. Yang, Z. Wang, and A. Minca, "Provable fictitious play for general mean-field games," *arXiv preprint arXiv:2010.04211*, 2020.

[26] D. Mguni, J. Jennings, and E. M. de Cote, "Decentralised learning in systems with many, many strategic agents," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[27] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, pp. 1861–1870, PMLR, 2018.

[28] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[29] R. Arnott, A. Kokoza, and M. Naji, "Equilibrium traffic dynamics in a bathtub model: A special case," *Economics of transportation*, vol. 7, pp. 38–52, 2016.

[30] R. Lamotte and N. Geroliminis, "The morning commute in urban areas with heterogeneous trip lengths," *Transportation Research Part B: Methodological*, vol. 117, pp. 794–810, 2018.

[31] G. Konidaris, S. Osentoski, and P. Thomas, "Value function approximation in reinforcement learning using the fourier basis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, pp. 380–385, 2011.