Transportation Research Part C 166 (2024) 104757

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

# Participatory traffic control: Leveraging connected and automated vehicles to enhance network efficiency

Minghui Wu [a], Ben Wang [b], Yafeng Yin [a,b,*], Jerome P. Lynch [c]

[a] Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, USA
[b] Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, USA
[c] Department of Civil and Environmental Engineering, Duke University, Durham, NC, USA

ARTICLE INFO

*Keywords:*
Connected and automated vehicles
Participatory traffic control
Network efficiency
Mean-field control
Day-to-day traffic dynamics

ABSTRACT

This paper aims to establish a framework of participatory traffic control, wherein connected and automated vehicles (CAVs) subtly influence the day-to-day adjustment process of human drivers, strategically redistributing traffic demand to enhance overall system efficiency. To address this complex challenge, we adopt the mean-field control framework, which enables us to model macroscopic interactions between CAVs and other travelers. After theoretically establishing the existence of the optimal policy, we leverage reinforcement learning algorithms to numerically solve the control problem. Distinct from existing approaches, our proposed method is scalable, model-free, distributed, and does not rely on the convergence properties of the underlying day-to-day traffic dynamics. It helps pave the way for the practical implementation of participatory traffic control.

## 1. Introduction

Participatory traffic control involves engaging a subset of participants in a traffic system to subtly influence the behaviors of others, thereby enhancing the overall system efficiency. The concept draws inspiration from previous strategies where individual travelers are nudged to alter their behaviors, with the expectation that these collective changes will positively impact the behavior of the wider traveling population to improve system performance (Xiong et al., 2020). We envision that the advance of connected and automated vehicle (CAV) technologies will offer new opportunities for participatory control, allowing for more real-time, adaptive, individualized, and flexible control mechanisms. As CAV adoption grows, this control scheme is anticipated to become increasingly appealing and advantageous.

Specifically, driving automation requires drivers to relinquish certain levels of control to their vehicles. In the early stages of development, this involves handing over driving maneuvers such as pedal and brake operations. Such relinquishment enables CAVs to function as "traffic stream regulators". By controlling the real-time speeds of CAVs, traffic authorities can influence the speed and acceleration of following vehicles to manage traffic. Pioneering work by Jin and Orosz (2014, 2016), Cui et al. (2017), Wu et al. (2018) and Zheng et al. (2020) demonstrated the ability of CAVs in maintaining string stability in mixed traffic environments, with field experiments providing further validation (Stern et al., 2018; Jin and Orosz, 2018). Subsequent studies by Vinitsky et al. (2018) and Wu et al. (2021) expanded these controllers and explored broader management objectives such as maximizing network capacity. Čičić et al. (2021) further explored CAVs in coordinating platoons to alleviate bottlenecks on highways. These studies collectively demonstrate the potential of CAVs as control actuators to improve road performance in various local traffic scenarios.

* Corresponding author at: Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, USA.
*E-mail address:* yafeng@umich.edu (Y. Yin).

https://doi.org/10.1016/j.trc.2024.104757
Received 19 February 2024; Received in revised form 26 May 2024; Accepted 7 July 2024
Available online 22 July 2024

As driving automation continues to advance and travelers become more trusting of the technology, they are expected to be more willing to surrender higher levels of control to their CAVs, including choices about routes and departure times (Di and Shi, 2021). Given human drivers' natural tendency to avoid congestion, this higher level of relinquished control enables CAVs to act as "traffic demand distributors", effectively regulating traffic flow across the network. Considering human drivers who aim to minimize individual travel costs while CAV actuators aim to minimize system travel costs, Zhang and Nie (2018) investigated the resulting mixed equilibrium and determined the optimal ratio of these two user types to strike a balance between improving system performance and lowering control intensity. Sharon et al. (2018) and Chen et al. (2020) investigated the minimal ratio of CAVs required to be controlled to achieve the system optimum. Moreover, Chen et al. (2020) demonstrated that CAV-based control can be effectively combined with pricing mechanisms to further enhance traffic management. More recently, Zhang et al. (2022) extended these considerations to include emission reduction as an additional control objective.

In another vein of inquiry, studies by Li et al. (2018), Lazar et al. (2021), Guo et al. (2022), and Liang et al. (2023) have moved beyond the steady-state analysis of network equilibrium to explore how network flow evolves on a day-to-day basis as travelers adjust their choices. In this setting, these works focus on using CAVs to steer the system towards a more efficient state of equilibrium. Although these investigations showcase the potential of CAVs as traffic demand distributors, they may fall short in supporting real-world implementation. For one thing, existing model-based methods assume human drivers following certain response dynamics, which may not align with actual human behavior. Moreover, the reliance on centralized control in previous models limits scalability, particularly in scenarios with a high population of CAVs. Furthermore, the success of previous methods largely hinges on the assumption that travel costs are monotone, a crucial factor for the global stability of the day-to-day traffic dynamics. However, as highlighted by Guo et al. (2018a), the monotonicity requirement may not always hold and flow dynamics do not necessarily converge to equilibria, especially when departure time choices are involved.

Motivated by these issues, this paper proposes a model-free and distributed approach to control a fraction of CAVs to enhance the system performance over time. The proposed approach offers a practical solution for implementing CAVs as traffic demand distributors without relying on exhaustive knowledge of human behaviors. This method employs a distributed control policy, instructing individual CAVs to act based on their local information. In addition, our approach significantly differs from previous studies (Li et al., 2018; Guo et al., 2022; Liang et al., 2023) in its control objective. Instead of driving the system to a desired state of equilibrium, we focus on minimizing the total system cost over time. Therefore, our method has no requirement for the choice of the underlying day-to-day dynamical models and their convergence properties.

To facilitate the presentation of our approach, we first consider a scenario with homogeneous travelers and model it as a finite agent control problem. Noting that the problem quickly becomes intractable as the number of travelers increases, we reformulate the problem within the mean-field control (MFC) framework (Cui et al., 2023) by considering the limiting case with an infinite number of travelers. In the new formulation, the control problem is elevated from the individual level to the population level, thereby significantly alleviating the computational burden. We further extend the model to accommodate traveler heterogeneity, enhancing its applicability to real-world scenarios. After establishing the existence of the optimal control policy, we employ reinforcement learning algorithms to numerically compute the optimal policy. The effectiveness of the proposed method is then tested in various choice scenarios and penetration levels.

Our model-free, scalable, and distributed control scheme offers a flexible and adaptable solution for various traffic management scenarios, capable of accommodating varying levels of CAV penetration. The remainder of this paper is structured as follows. Section 2 presents the model. Section 3 discusses the control algorithm and Section 4 presents numerical examples. Lastly, Section 5 concludes the paper.

## 2. Model

In this section, we present a general model that is applicable to various travel choices, such as route and/or departure time choices. We start with a simplified case where all travelers are assumed to be homogeneous and subsequently extend it to accommodate heterogeneous travelers. The main notations are summarized in the Appendix.

### 2.1. Finite-agent control

Consider a traffic system with $N$ controllable CAVs (their recruitment is out of the scope of this paper) and $M$ uncontrollable vehicles (e.g. human drivers and uncontrolled CAVs). We refer the controllable CAVs as system-optimal (SO) users, focusing on overall system efficiency, and the uncontrollable vehicles as user-optimal (UO) users, prioritizing individual interests.

The day-to-day (DTD) travel decision-making is modeled as a Markov decision process (MDP). The travel demand is $N$ SO users and $M$ UO users, and each traveler's choice is considered as their state of that day. Let $x_t^i \in \mathcal{X}$ represent the travel choice of SO user $i \in [N] = \{1, \dots, N\}$ on day $t$, where $\mathcal{X}$ is the finite set of all travel choices. Similarly, $s_t^j \in \mathcal{X}$ denotes the travel choice of UO user $j \in [M]$ on day $t$. The penetration rate of SO users is defined as $\theta = \frac{N}{M+N}$.

Each traveler's state contributes to a state distribution over the population. We define the empirical distributions $\mu_t^N$ and $\nu_t^M$ for SO and UO users, respectively, as follows:

$$\mu_t^N = \frac{1}{N} \sum_{i \in [N]} \delta_{x_t^i},$$

$$\nu_t^M = \frac{1}{M} \sum_{j \in [M]} \delta_{s_t^j}.$$

where $\delta_x$ denotes a Kronecker delta function at point $x$. It represents a vector of shape $|\mathcal{X}|$ that equals 1 at $x$ and 0 elsewhere. Here, the superscripts, $M$ and $N$, are used to clarify that we are dealing with a finite agent model.

To provide an illustrative example, $x_t^i$ can represent the route choice of SO user $i$ on day $t$. In this case, $\mathcal{X}$ represents the path set, while $\mu_t^N$ and $\nu_t^M$ reflect the path choice distribution of the two groups. Therefore, the path flow of the two groups on day $t$ can be represented by $N\mu_t^N$ and $M\nu_t^M$.

UO users are assumed to follow a certain response dynamic, aggregatedly represented by a DTD model $\nu_{t+1}^M = q(\mu_t^N, \nu_t^M)$, which may not be revealed to the traffic management agency. This model reflects the influence of the previous day's experience, dictated by $\mu_t^N$ and $\nu_t^M$, on the subsequent day's behavior. An example of such response functions is Smith's dynamic (Smith, 1984), a widely used model in literature, which is characterized by the following equation:

$$\nu_{t+1}^M(x) - \nu_t^M(x) = \eta \sum_{x' \in \mathcal{X}} \left( \nu_t^M(x')[c_{x'}(\mu_t^N, \nu_t^M) - c_x(\mu_t^N, \nu_t^M)]^+ - \nu_t(x)[c_x(\mu_t^N, \nu_t^M) - c_{x'}(\mu_t^N, \nu_t^M)]^+ \right), \tag{1}$$

where $\nu_t^M(x)$ represents the proportion of UO users choosing travel choice $x$ on day $t$, $[\cdot]^+ = \max\{0, \cdot\}$, $\eta$ dictates the user inertia level, and $c_x(\mu_t^N, \nu_t^M)$ denotes the travel cost of choice $x$.

On the other hand, each SO user's behavior is modeled individually by an MDP. Each SO user $i$ is assigned an action $a_t^i \in \mathcal{A}$ on day $t$, which represents the suggested travel option for the next day $t+1$. In the homogeneous case, $\mathcal{X}$ and $\mathcal{A}$ are equivalent since they both denote the identical travel choice set. The action $a_t^i$ is drawn from an assignment policy $\pi(\cdot|x_t^i, \mu_t^N, \nu_t^M)$ determined by the management agency, which is a function of the current choice and the empirical distributions. Based on their actions, SO users' states evolve according to the transition kernel $x_{t+1}^i \sim p(\cdot|x_t^i, a_t^i, \mu_t^N, \nu_t^M)$. The kernel can be adapted for various compliance scenarios:

- Full compliance: SO users are perfectly compliant with the assignment. In such case, $p(x|x_t^i, a_t^i, \mu_t^N, \nu_t^M) = p(x|a_t^i) = \begin{cases} 1, & \text{if } x = a_t^i, \\ 0, & \text{otherwise.} \end{cases}$
- Partial compliance due to inertia: SO users may have a preference for retaining their previous choices (Srinivasan and Mahmassani, 2000; Qi et al., 2023), which can be modeled by

$$p(x|x_t^i, a_t^i, \mu_t^N, \nu_t^M) = p(x|x_t^i, a_t^i) = \begin{cases} 1-\epsilon, & \text{if } x = a_t^i, \\ \epsilon, & \text{if } x = x_t^i, \\ 0, & \text{otherwise.} \end{cases}$$

- Partial compliance due to self-interests: we can also manipulate the transition kernel to model the behavior considered in Guo et al. (2022), where SO users are only willing to sacrifice interest within a threshold $\epsilon$

$$p(x|x_t^i, a_t^i, \mu_t^N, \nu_t^M) = \begin{cases} 1, & \text{if } x = a_t^i \text{ and } a_t^i \in \Omega^{\epsilon-BR}_{\mu_t^N, \nu_t^M}, \\ 1, & \text{if } x = x_t^i \text{ and } a_t^i \notin \Omega^{\epsilon-BR}_{\mu_t^N, \nu_t^M}, \\ 0, & \text{otherwise .} \end{cases}$$

where $\Omega^{\epsilon-BR}_{\mu_t^N, \nu_t^M}$ refers to the set of acceptable choices for $\epsilon$-bounded rational travelers. The travel cost of choices within $\Omega^{\epsilon-BR}_{\mu_t^N, \nu_t^M}$ is $\epsilon$-close to the optimal choices determined by $\mu_t^N$ and $\nu_t^M$.

The system's average travel cost follows

$$C(\mu_t^N, \nu_t^M) = \frac{1}{M+N} \sum_{x \in \mathcal{X}} (N\mu_t^N(x) + M\nu_t^M(x)) c_x(\mu_t^N, \nu_t^M),$$

where $\mu_t^N(x)$ is the proportion of SO users that choose travel choice $x$. It is worth noting that the system cost is equivalent to

$$C(\mu_t^N, \nu_t^M) = \sum_{x \in \mathcal{X}} (\theta\mu_t^N(x) + (1-\theta)\nu_t^M(x)) c_x(\mu_t^N, \nu_t^M),$$

which reflects the impact of the penetration rate $\theta$ on the system's total travel cost. Moreover, the penetration rate also affects the UO user behavior as SO users have a higher influence when the penetration rate increases.

We now define the metrics to facilitate the following discussion. We first metrize $\mathcal{X}$ with the discrete metric, i.e. for $x, y \in \mathcal{X}$, $d(x, y) = 0$ if $x = y$, and 1 otherwise. Hence, $(\mathcal{X}, d)$ is a complete metric space. We further use the total variation for probability distributions $\mu, \nu \in \mathcal{P}(\mathcal{X})$, i.e. $d_{TV}(\mu, \nu) = \frac{1}{2}\sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|$.

Two assumptions are made regarding the transition kernel and the cost function, which will be used in later sections.

**Assumption 1.** The transition kernels $p$ and $q$ are Lipschtiz continuous with respect to $\mu$ and $\nu$.

**Assumption 2.** The individual cost $c_x$ is Lipschtiz continuous with respect to $\mu$ and $\nu$ for all states $x \in \mathcal{X}$.

For route choice problems, the Lipschtiz continuous cost assumption is commonly satisfied by a range of link performance functions including the BPR function (Bureau of Public Roads, 1964). When the departure time choice is involved, the assumption

(essentially the Lipschitz continuity of the delay operator (Friesz et al., 2021)) is still widely used in literature, such as in Mounce and Carey (2011) and Friesz et al. (2011). Under Lipschtiz continuous cost functions, it is mild to assume transition kernels also being Lipschtiz continuous, which can be satisfied by many day-to-day dynamical models. Proposition 1 establishes the Lipschtiz continuity of Smith's dynamic (Smith, 1984). Besides, our numerical experiments later demonstrate that the proposed method can achieve good performance even when these assumptions are not strictly satisfied (e.g. experiments with the bathtub model).

**Proposition 1.** *Under Assumption 2, Smith's dynamic $\nu_{t+1} = q(\mu_t, \nu_t)$ characterized by Eq. (1) is Lipschitz continuous*

**Proof.** To simplify notations, we omit the superscripts $M$ and $N$. For every state $x, x'$, $c_{x'}(\mu_t, \nu_t) - c_x(\mu_t, \nu_t)$ is Lipschitz continuous, denoted as $g(\mu_t, \nu_t)$. By definition, we have $|g(\mu_t', \nu_t') - g(\mu_t, \nu_t)| \leq L_1 d_{TV}(\mu_t', \mu_t) + L_2 d_{TV}(\nu_t', \nu_t)$ for all $\mu_t, \nu_t, \mu_t', \nu_t' \in \mathcal{P}(\mathcal{X})$, where $L_1$ and $L_2$ are two Lipschitz constants.

Now let us consider $[g(\mu_t, \nu_t)]^+$. If $g(\mu_t, \nu_t)$ and $g(\mu_t', \nu_t')$ are both non-negative, then $|[g(\mu_t, \nu_t)]^+ - [g(\mu_t', \nu_t')]^+| = |g(\mu_t, \nu_t) - g(\mu_t', \nu_t')| \leq L_1 d_{TV}(\mu_t', \mu_t) + L_2 d_{TV}(\nu_t', \nu_t)$. If one of them is negative, without losing generality, assume $g(\mu_t', \nu_t') < 0$, then $|[g(\mu_t, \nu_t)]^+ - [g(\mu_t', \nu_t')]^+| = g(\mu_t, \nu_t) \leq |g(\mu_t, \nu_t) - g(\mu_t', \nu_t')| \leq L_1 d_{TV}(\mu_t', \mu_t) + L_2 d_{TV}(\nu_t', \nu_t)$. Otherwise, when the two are both negative, $|[g(\mu_t, \nu_t)]^+ - [g(\mu_t', \nu_t')]^+| = 0$. Therefore, $[g(\mu_t, \nu_t)]^+$ is also Lipschitz continuous. As the finite sum and product of Lipschitz continuous functions is also Lipschitz continuous, $\nu_{t+1}(x)$ or $q(\mu_t, \nu_t)(x)$ is Lipschtiz continuous for all states $x \in \mathcal{X}$. Denote $|q(\mu_t, \nu_t)(x) - q(\mu_t', \nu_t')(x)| \leq L_1^x d_{TV}(\mu_t, \mu_t') + L_2^x d_{TV}(\nu_t, \nu_t')$. Furthermore, we have

$$
\begin{aligned}
d_{TV}(q(\mu_t, \nu_t), q(\mu_t', \nu_t')) &= \frac{1}{2} \sum_{x \in \mathcal{X}} |q(\mu_t, \nu_t)(x) - q(\mu_t', \nu_t')(x)| \\
&\leq \frac{1}{2} \left( \sum_{x \in \mathcal{X}} L_1^x \right) d_{TV}(\mu_t, \mu_t') + \frac{1}{2} \left( \sum_{x \in \mathcal{X}} L_2^x \right) d_{TV}(\nu_t, \nu_t'),
\end{aligned}
$$

which proves the Lipschitz continuity of $q(\mu_t, \nu_t)$. □

The control objective of the management agency is to find the optimal policy to minimize the total discounted cost over the infinite horizon, leading to the following optimal control problem:

$$
\begin{aligned}
&\min_{\pi} \; J(\pi) = E\left[ \sum_{t=0}^{\infty} \gamma^t C(\mu_t^N, \nu_t^M) \right] \\
&s.t. \; \nu_{t+1}^M = q(\mu_t^N, \nu_t^M), \\
&\quad a_t^i \sim \pi(\cdot | x_t^i, \mu_t^N, \nu_t^M); \quad \forall i \in [N] \\
&\quad x_{t+1}^i \sim p(\cdot | x_t^i, a_t^i, \mu_t^N, \nu_t^M); \quad \forall i \in [N]
\end{aligned}
$$

where $\gamma$ is the discount factor, and the constraints govern the behavior of SO and UO users.

### 2.2. Mean-field control

Effectively solving the multi-agent control problem above involves the joint state–action space of all agents (Zhang et al., 2021). Due to its exponential nature, solving the optimal control policy becomes intractable as the number of agents increases significantly (Yang et al., 2018). To address this issue, we adopt the mean-field control (MFC) framework proposed by Cui et al. (2021, 2023). This approach, inspired by mean-field game theory (Huang et al., 2006; Lasry and Lions, 2007), simplifies the model by considering an infinite number of agents, thereby making it more scalable and computationally feasible.

In this MFC framework, as the numbers of controllable CAVs ($N$) and uncontrollable vehicles ($M$) approach infinity, the empirical distribution $\mu_t^N$ and $\nu_t^M$ becomes the mean-field (MF) distribution $\mu_t, \nu_t \in \mathcal{P}(\mathcal{X})$ under the law of large numbers, where $\mathcal{P}(\mathcal{X})$ denotes all probability mass functions on the state space. This transition to the limiting case eliminates the necessity of tracking the state of each individual SO user. Instead, we concentrate on a representative agent, whose state $x$ is now regarded as a random variable aligned with the MF distribution. For consistency, we use the same notations as previously introduced. Specifically, $q(\mu_t, \nu_t)$ represents the transition kernel for UO users, while $p(\cdot | x, a, \mu_t, \nu_t)$ refers to the transition kernel for each individual SO user. The assignment policy and the system cost are denoted as $\pi(\cdot | x, \mu_t, \nu_t)$ and $C(\mu_t, \nu_t)$, respectively.

We are now ready to reformulate the control problem at the population level. The two MF distributions $(\mu_t, \nu_t)$ are used to represent the aggregate behavior of the two populations, which will be considered as the population state. The aggregate assignment for all SO users is represented by the joint state–action distribution over all SO users:

$$
h_t = \mu_t \otimes \pi_t(\mu_t, \nu_t) \in \mathcal{H}(\mu_t),
$$

where $\pi_t(\mu_t, \nu_t) = \pi_t(\cdot | \cdot, \mu_t, \nu_t)$ and $\otimes$ denotes the element-wise product. $\mathcal{H}(\mu_t) \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{A})$ denotes the joint distribution whose state marginal distribution matches with $\mu_t$. Note that $h_t$ encompasses the comprehensive information of all individual assignments, hence we regard it as the population action.

It is intriguing to ask how the population state evolves with the population action. Due to the homogeneity assumption, every SO user follows the same transition kernel. Consequently, the subsequent MF distribution of SO users can be calculated deterministically by:

$$
\mu_{t+1} = \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} p(\cdot | x, a, \mu_t, \nu_t) \mu_t(x) \pi_t(a | x, \mu_t, \nu_t). \tag{2}
$$

**Table 1**
Comparison between finite-agent and mean-field control.

| | Finite-agent control | Mean-field control |
|---|---|---|
| State | Current route choice of each traveler | Current route choice distributions of two groups |
| Action | Assigned route choice for each SO user | Joint distribution of current and assigned route choices for all SO users |
| Transition | Each SO user follows assignment<br>UO users follow day-to-day dynamics | SO users follow a population kernel<br>UO users follow day-to-day dynamics |
| Cost | Total cost of all travelers | Total cost of all travelers |

To bring more insights into Eq. (2), $p(\cdot|x,a,\mu_t,\nu_t)$ represents the outcome of implementing state–action pair $(x,a)$. The next MF distribution $\mu_{t+1}$ is the weighted sum of all possible outcomes based on the possibility of each state–action pair, $\mu_t(x)\pi_t(a|x,\mu_t,\nu_t)$. Note that $\mu_t(x)\pi_t(a|x,\mu_t,\nu_t)$ is simply $h_t(x,a)$, and the transition kernel $p(\cdot|x,a,\mu_t,\nu_t)$ is determined by $\mu_t$ and $\nu_t$. Therefore, Eq. (2) can be rewritten as a function of $\mu_t,\nu_t$ and $h_t$:

$$\mu_{t+1} = T(\mu_t,\nu_t,h_t),$$

which, together with $\nu_{t+1} = q(\mu_t,\nu_t)$, provides a population transition kernel.

In addition, we introduce a population policy $\hat{\pi} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X}\times\mathcal{A})$, which takes the population state $(\mu_t,\nu_t)$ as input, and outputs the population action $h_t$. This allows us to retrieve the following single-agent control:

$$\begin{aligned}
&\min_{\hat{\pi}} J(\hat{\pi}) = E\left[\sum_{t=0}^{\infty}\gamma^t C(\mu_t,\nu_t)\right]\\
&s.t.\ \nu_{t+1} = q(\mu_t,\nu_t),\\
&\quad h_t \sim \hat{\pi}(\cdot|\mu_t,\nu_t),\\
&\quad \mu_{t+1} = T(\mu_t,\nu_t,h_t).
\end{aligned}$$

To provide more insights into this reformulation, Table 1 compares the model formulation of the finite-agent and mean-field control using the routing example in Section 2.1. Notably, in the new model, the single agent represents the entire population rather than an individual traveler. Since we no longer need to track multiple agents simultaneously, the mean-field control is considerably more scalable than the finite agent model.

The optimal stationary policy $\hat{\pi}$ always exists, as shown in the following proposition:

**Proposition 2.** *Under Assumptions 1 and 2, the MFC model always has an optimal stationary policy $\hat{\pi}$.*

**Proof.** Since we are using the discrete metric on $\mathcal{X}$, for $\mu,\nu \in \mathcal{P}(\mathcal{X})$, the total variation distance is equivalent to 1-Wasserstein distance (De Palma et al., 2021), i.e., $W_1(\mu,\nu) = d_{TV}(\mu,\nu)$.

Meanwhile, the UO user's response dynamic is equivalent to

$$\nu_{t+1} \sim \delta_{q(\mu_t,\nu_t)},$$

where $\delta_{q(\mu_t,\nu_t)}$ is a degenerate distribution located at point $q(\mu_t,\nu_t) \in \mathcal{P}(\mathcal{X})$. Note that for two degenerate distributions located at point $q,q' \in \mathcal{P}(\mathcal{X})$, the 1-Wasserstein distance between the two is equivalent to the $\|q-q'\|_2$, which is further bounded between $\frac{2}{|\mathcal{X}|}d_{TV}(q,q')$ and $2d_{TV}(q,q')$. Therefore, Assumption 1 indicates Assumption B.1 in Cui et al. (2023). Since the finite sum, product, and composition of Lipschitz continuous functions is also Lipschitz continuous, Assumption 2 implies the Lipschitz continuity of the system cost $C(\mu_t,\nu_t)$, which is equivalent to Assumption B.2 in Cui et al. (2023). Hence, Theorem B.4 (Cui et al., 2023) leads to the existence of the optimal stationary policy. □

### 2.3. Relaxing the homogeneity assumption

While previous sections focus on homogeneous travelers, real-world transportation systems exhibit significant heterogeneity among travelers. To more accurately model these systems, we now relax the homogeneity assumption to accommodate varying traveler characteristics.

#### 2.3.1. Different action spaces

One typical heterogeneity arises from the variation in action spaces among travelers. For instance, in routing scenarios, travelers can only choose paths between their origin–destination (OD) pairs.

To address this issue, we first classify travelers into $K$ types, index by type $k \in \mathcal{K} = \{1,\dots,K\}$. Each type is associated with a unique action space $\mathcal{A}^k$, which represents the allowable travel choices for type $k$. We assume these action spaces are disjoint, and the overall state space $\mathcal{X} = \cup_{k=1}^{K}\mathcal{A}^k$. For each state $x \in \mathcal{X}$, denote the type of travelers that can choose it as $k(x)$. To ensure a valid

transition kernel for UO users, it is necessary that $q(\mu,\nu)$ maintains proportionality across types. Specifically, denote $\nu' = q(\mu,\nu)$, then it must satisfy $\sum_{x\in\mathcal{A}^k}\nu'(x) = \sum_{x\in\mathcal{A}^k}\nu(x)$ for all types $k$. Note that this validity is typically ensured as the kernel is derived from existing dynamics such as Smith's dynamic (Smith, 1984).

To facilitate analysis, we introduce a common action space $\bar{\mathcal{A}}$, together with $K$ mappings $\mathcal{G}^k : \bar{\mathcal{A}} \to \mathcal{A}^k$, where $k \in \mathcal{K}$. For example, if a network contains two OD pairs, each connected by three paths, then $\mathcal{X} = \{1,2,\dots,6\}$, $\mathcal{A}^1 = \{1,2,3\}$, and $\mathcal{A}^2 = \{4,5,6\}$. We can define $\bar{\mathcal{A}} = \{1,2,3\}$, and $\mathcal{G}^1(a) = a$, $\mathcal{G}^2(a) = a+3$.

Based on the common action space, the transition kernel for SO users and the individual policy are redefined as $\hat{p}(x'|x,\bar{a},\mu,\nu) = p(x'|x,\mathcal{G}^{k(x)}(\bar{a}),\mu,\nu) : \mathcal{X}\times\mathcal{X}\times\bar{\mathcal{A}}\times\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{X}) \to \mathbb{R}$ and $\pi(\bar{a}|x,\mu,\nu) : \bar{\mathcal{A}}\times\mathcal{X}\times\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{X}) \to \mathbb{R}$. As can be seen, introducing the common action space reduces the dimensionality of the action space and ensures consistency across types. Therefore, the heterogeneous action spaces can be equivalently represented by the common action space $\bar{\mathcal{A}}$ and the $K$ mappings. As a result, considering agent types is no longer necessary, which leads to the following homogeneous optimal control problem:

$$\begin{aligned}
&\min_{\pi}\ J(\pi) = E\left[\sum_{t=0}^{\infty}\gamma^t C(\mu_t^N, \nu_t^M)\right],\\
&s.t.\ \nu_{t+1}^M = q(\mu_t^N, \nu_t^M),\\
&\quad \bar{a}_t^i \sim \pi(\cdot|x_t^i, \mu_t^N, \nu_t^M), \quad \forall i\in[N],\\
&\quad x_{t+1}^i \sim \hat{p}(\cdot|x_t^i, \bar{a}_t^i, \mu_t^N, \nu_t^M), \quad \forall i\in[N].
\end{aligned}$$

As before, letting $N, M \to \infty$ yields the limiting MFC model:

$$\begin{aligned}
&\min_{\hat{\pi}}\ J(\hat{\pi}) = E\left[\sum_{t=0}^{\infty}\gamma^t C(\mu_t, \nu_t)\right],\\
&s.t.\ \nu_{t+1} = q(\mu_t, \nu_t),\\
&\quad h_t \sim \hat{\pi}(\cdot|\mu_t, \nu_t),\\
&\quad \mu_{t+1} \sim \hat{T}(\mu_t, \nu_t, h_t).
\end{aligned}$$

where $\hat{T}(\mu_t,\nu_t,h_t) = \sum_{x\in\mathcal{X}}\sum_{\bar{a}\in\bar{\mathcal{A}}}\hat{p}(\cdot|x,\bar{a},\mu,\nu)\mu_t(x)\pi_t(\bar{a}|x,\mu_t,\nu_t)$.

Since our modification of action space and transition kernel does not influence the MF distributions, Assumptions 1 and 2 still imply the Lipschitz continuity in the revised model. Consequently, we can retain the existence of the optimal policy. We skip the proof as it is trivial.

**Proposition 3.** *Under Assumptions 1 and 2, the MFC model with heterogeneous state spaces always has an optimal stationary policy $\hat{\pi}$.*

#### 2.3.2. Different cost formulations

Besides the state spaces, traveler heterogeneity also manifests in other aspects such as cost functions. For example, travelers may have different values of time or desired arrival times, leading to different costs even with identical travel choices. Moreover, different cost functions also lead to heterogeneity in the transition process as commuters react differently to the same population behavior.

To capture this, we augment the state with an "ID variable" $\hat{x}_t^i = (x_t^i, z^i)$. $x_t^i$ is still the travel choice of traveler $i$ on day $t$ and $z^i \in \mathcal{Z}$ represents a variable sufficient to determine their type, where $\mathcal{Z}$ denotes the finite set for all possible ID variables. For example, $z^i$ refers to the value of time or the desired arrival time for traveler $i$. Consequently, the new state space $\hat{\mathcal{X}}$ is the cartesian product of $\mathcal{X}$ and $\mathcal{Z}$, which is embedded with the discrete metric as before. The action $a_t^i \in \mathcal{X}$ continues to represent the assignment for traveler $i$ on day $t$.

With the augmented state for both SO users and UO users, the empirical distribution $\hat{\mu}_t^N, \hat{\nu}_t^M$ now reflects the joint distribution of travel choices and ID variables. We still metrize $\mathcal{P}(\mathcal{X})$ by the total variation distance. Each commuter's daily travel cost is then a function of this joint distribution $\hat{c}_{\hat{x}}(\hat{\mu}_t^N, \hat{\nu}_t^M)$, leading to a new system cost:

$$\hat{C}(\hat{\mu}_t^N, \hat{\nu}_t^M) = \frac{1}{M+N}\sum_{\hat{x}\in\hat{\mathcal{X}}}(N\mu_t^N(\hat{x}) + M\nu_t^M(\hat{x}))\hat{c}_{\hat{x}}(\hat{\mu}_t^N, \hat{\nu}_t^M).$$

We further assume that each traveler has a fixed ID variable. Thus, the transition kernels for both UO and SO users are modified to reflect these augmented states:

$$\begin{aligned}
&\hat{x}_{t+1}^i = (x_{t+1}^i, z^i) \sim \hat{p}(\cdot|\hat{x}_t^i, a_t^i, \hat{\mu}_t^N, \hat{\nu}_t^M),\\
&\hat{\nu}_{t+1} = \hat{q}(\hat{\mu}_t^N, \hat{\nu}_t^M),
\end{aligned}$$

which leads to the new finite agent control model:

$$\begin{aligned}
&\min_{\pi}\ J(\pi) = E\left[\sum_{t=0}^{\infty}\gamma^t \hat{C}(\hat{\mu}_t^N, \hat{\nu}_t^M)\right],\\
&s.t.\ \hat{\nu}_{t+1}^M = \hat{q}(\hat{\mu}_t^N, \hat{\nu}_t^M),\\
&\quad a_t^i \sim \pi(\cdot|\hat{x}_t^i, \hat{\mu}_t^N, \hat{\nu}_t^M), \quad \forall i\in[N],
\end{aligned}$$

$$\hat{x}^i_{t+1} \sim \hat{p}(\cdot|\hat{x}^i_t, a^i_t, \hat{\mu}^N_t, \hat{\nu}^M_t), \quad \forall i \in [N].$$

and the corresponding MFC model:

$$\begin{aligned} \min_{\hat{\pi}} \ & J(\hat{\pi}) = E\left[\sum_{t=0}^{\infty} \gamma^t \hat{C}(\hat{\mu}_t, \hat{\nu}_t)\right], \\ s.t. \ & \hat{\nu}_{t+1} = \hat{q}(\hat{\mu}_t, \hat{\nu}_t), \\ & h_t \sim \hat{\pi}(\cdot|\hat{\mu}_t, \hat{\nu}_t), \\ & \hat{\mu}_{t+1} \sim \hat{T}(\hat{\mu}_t, \hat{\nu}_t, h_t). \end{aligned}$$

where $\hat{T}(\hat{\mu}_t, \hat{\nu}_t, h_t) = \sum_{\hat{x}\in\hat{\mathcal{X}}} \sum_{a\in\mathcal{A}} \hat{p}(\cdot|\hat{x}, a, \hat{\mu}, \hat{\nu})\hat{\mu}_t(\hat{x})\pi_t(a|\hat{x}, \hat{\mu}_t, \hat{\nu}_t)$.

With the modified assumptions for the new model, we can recover the existence of the optimal policy. The proof is skipped since it is similar to Proposition 2.

**Assumption 3.** The transition kernels $\hat{p}$ and $\hat{q}$ are Lipschtiz continuous with respect to $\hat{\mu}$ and $\hat{\nu}$.

**Assumption 4.** The individual travel cost function $\hat{c}_{\hat{x}}$ is Lipschtiz continuous with respect to $\hat{\mu}$ and $\hat{\nu}$ for all $\hat{x} \in \hat{\mathcal{X}}$.

**Proposition 4.** *If Assumptions 3 and 4 hold, the MFC model with heterogeneous cost functions always has an optimal stationary policy $\hat{\pi}$.*

## 3. Algorithm

So far, we have developed a single-agent control model for the MFC problem and established the existence of the optimal policy under mild conditions. In this section, we propose a distributed and model-free algorithm to solve for the optimal policy, aligning with the practical requirements for real-world implementation. We utilize the MFC reinforcement learning (RL) approach (Cui et al., 2023) as outlined in Algorithm 1. This approach provides a framework for iteratively refining the population policy based on the observed system behavior.

**Algorithm 1** MFC-RL algorithm framework

1: **input**: Initialize policy $\hat{\pi}^\theta$.
2: **for** iterations $n = 1, 2, \ldots$ **do**
3: Sample population action $h_t \sim \hat{\pi}^\theta(\cdot|\mu_t, \nu_t)$;
4: Execute $h_t$ and observe next population state $\mu_{t+1}, \nu_{t+1}$ (using Algorithm 2 or 3);
5: Observe system cost $C_t$;
6: Update policy $\hat{\pi}^\theta$ (with RL algorithms);
7: **end for**

During training, we can compute the subsequent population state $\mu_{t+1}, \nu_{t+1}$ directly from the current one based on the population kernels $T$ and $q$. This approach is detailed in Algorithm 2.

**Algorithm 2** System transition based on the population kernel

1: **input**: Population state $\mu_t, \nu_t$ and population action $h_t$.
2: Calculate $\mu_{t+1} = \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} p(\cdot|x, a, \mu_t, \nu_t)h_t(x, a)$;
3: Calculate $\nu_{t+1} = q(\mu_t, \nu_t)$;
4: **return**: Next population state $\mu_{t+1}, \nu_{t+1}$

Note that although the population kernel is convenient to use, it accurately captures the system transition only when there are an infinite number of agents. As practical scenarios involve only a finite number of travelers, the actual system transition may deviate from the ideal assumptions made by Algorithm 2. To address this, Algorithm 3 gives an alternative system transition based on individual sampling, which is more aligned with real-world conditions. Here, the population action $h_t$ is first broadcasted to all SO users. As it represents the joint distribution of states and actions, the original individual policy can be recovered (Line 4 in Algorithm 3) through $\pi_t(a|x) = \frac{h_t(x,a)+\epsilon}{\sum_{a'\in\mathcal{A}}(h_t(x,a')+\epsilon)}$ for all state–action pairs. Here, $\epsilon = 10^{-10}$ is introduced to ensure numerical stability (Cui et al., 2023). Subsequently, each SO user independently chooses their action based on their current state, without relying on any population information. In the final step, the management agency observes the behaviors of UO users and computes the empirical distribution of SO users as the next population state.

The proposed algorithm offers several advantages for the practical traffic management: (1) By utilizing the individual sampling-based transition, the algorithm follows the paradigm of centralized training decentralized execution (CTDE) (Zhang et al., 2021).

**Algorithm 3** System transition based on individual sampling

**input**: Population state $\mu_t, \nu_t$ and population action $h_t$.
Broadcast $h_t$;
**for** SO agent $i = 1, ..., N$ **do**
    Recover individual policy $\pi_t$ from $h_t$;
    Observe current state $x_t^i$;
    Sample and execute action $a_t^i \sim \pi_t(\cdot|x_t^i)$;
**end for**
Observe $\nu_{t+1}$;
Calculate empirical distribution $\mu_{t+1}$;
**return**: Next population state $\mu_{t+1}, \nu_{t+1}$

**Table 2**
Hyperparameter values.

| Hyperparameter | Value |
|---|---|
| GAE lambda | 1 |
| KL coefficient | 0.03 |
| Clip parameter | 0.2 |
| Learning rate | 0.00005 |
| Training batch size | 24,000 |
| Mini-batch size | 4,000 |
| Gradient steps per batch | 8 |

This approach efficiently distributes the decision-making process among individual agents, thereby significantly reducing the computational burden; (2) As a model-free approach, the MFC-RL framework learns the optimal policy by observing outcomes of transitions and costs, eliminating the need for requiring specific underlying mechanism; (3) The proposed scheme is flexible in the choice of RL algorithms for updating the policy (Line 6 in Algorithm 1).

## 4. Numerical experiments

In this section, we apply the proposed model and algorithm to a range of scenarios including route choices, departure time choices, and their combination. The experiments span diverse network scales, congestion technologies, and response dynamics, showcasing the versatility of the proposed method.

### 4.1. Experiment setup

For all experiments, the discount factor $\gamma$ is set to 0.99. Considering the practical implausibility of the infinite horizon in the control model, we truncate the horizon length to 200 days, hence the training process consists of iterations of these 200-day episodes. To ensure robustness, we initialize the system randomly at the beginning of each episode, which allows the algorithm to be trained across various network flow conditions. The system costs are normalized against the average cost of 10 random distributions to maintain problem independence in our experiment settings and results. For the RL algorithm component, we employ Proximal Policy Optimization (PPO) (Schulman et al., 2017) with hyperparameters detailed in Table 2. Here we use the same hyperparameter values for all experiments, indicating the generality of our approach without the need for individual fine-tuning for each scenario.

### 4.2. Problems and benchmarks

Five problems are considered in this paper, including two for route choices, two for departure time choices, and one joint choice scenario. The problem details are given as follows.

#### 4.2.1. Routing

In route choice scenarios, travelers are naturally grouped based on their OD pairs. $\mathcal{X}$ refers to the set of all available paths for all OD pairs, while the path set for OD $i$ is $\mathcal{A}^i \subseteq \mathcal{X}$. In addition, the travel time is taken as the individual travel cost.

The following two networks are used as testbeds:

- Nguyen–Dupuis network (Nguyen and Dupuis, 1984) with 4 OD pairs and 19 links;
- Sioux Falls network (LeBlanc et al., 1975) with 518 OD pairs and 76 links.

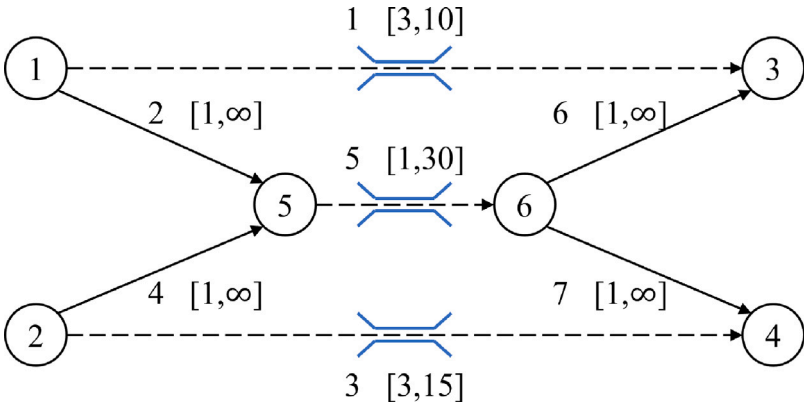


**Fig. 1.** The example road network. Each link is labeled with link number [free-flow travel time, bottleneck capacity].

As discussed in Section 2.3.1, a common space $\bar{\mathcal{A}}$ and the associated population action $h_t \in \mathcal{P}(\mathcal{X} \times \bar{\mathcal{A}})$ are utilized to manage the heterogeneity arising from multiple OD pairs. Given the large number of paths ($|\mathcal{X}|$), we further assume that the assignment policy is the same for all commuters within the same OD, regardless of their current path choices. Then, the population action can be represented by a two-dimensional matrix $O \in [-1, 1]^{K \times |\bar{\mathcal{A}}|}$, where $K$ is the number of types or ODs. Each individual assignment policy can be recovered in a similar way as $\pi(\mathcal{G}^{k(s)}(a)|s) = \frac{O_{k(s)a}+1+\epsilon}{\sum_{x \in \bar{\mathcal{A}}}(O_{k(s)x}+1+\epsilon)}$. This simplification, although sacrificing some flexibility for more personalized control, significantly reduces the problem dimension and lowers the training complexity.

#### 4.2.2. Departure time choices

In the departure time choice problem, $\mathcal{X} = \mathcal{A}$ represents the discretized departure time window, and the MF distribution implies the departure rate profile over the population. The individual travel cost comprises both the travel time cost and the scheduling delay.

Two congestion models, the bottleneck and the bathtub model, are considered:

- The bottleneck model, as described in Guo et al. (2018b), consists of 6,000 commuters and a bottleneck with a capacity of 3,000 vehicles per hour. The penalties for travel time, early arrival, and late arrival are set as $\alpha = 10$, $\beta = 5$, and $\gamma = 15$ respectively. The departure time window for each day is defined as $[0, 3]$ hours, which is further discretized into 60 intervals. The desired arrival time for all commuters is set at 2.
- The bathtub model is based on the downtown rush-hour setting in Arnott and Buli (2018), which takes Greenshields' model (Greenshields et al., 1935) as the network fundamental diagram. All commuters share the same trip length of 2 miles. The penalties for travel time, early arrival, and late arrival are specified as $\alpha = 1, \beta = 0.51, \gamma = 2.06$ (Lamotte and Geroliminis, 2018). The departure time window is defined as [0, 1] hours and is discretized into 20 slices.

#### 4.2.3. Joint route and departure time choices

In this more complex scenario, the state space $\mathcal{X}$ comprises all combinations of departure time and route choices. The action space $\mathcal{A}^i$ for type $i$ refers to all the possible choices associated with OD pair $i$. As in the previous case, the travel cost also contains the scheduling delay.

We employ a simple network with three bottlenecks and four expressways with stable link travel times (Yin et al., 2004), as shown in Fig. 1. This network consists of two OD pairs ($1 \to 3$, $2 \to 4$) with the demand of 120 and 150 respectively. Each OD pair has two routes, and the bottleneck on link 5 is shared by commuters from both OD pairs. We further discretize the departure time window to 20 time periods, and the desired arrival time for all travelers is set to 8. The penalty for travel time, early arrival, and late arrival are $\alpha = 1, \beta = 0.5, \gamma = 3$. Similar to Section 4.2.1, we assume that SO users within the same OD pair share the same assignment to manage computational demands.

#### 4.2.4. Benchmarks

Two benchmarks on the system total costs are considered to evaluate the policy performance:

- Pure UO drivers response: This benchmark is established when the penetration rate of SO users is 0%, where the system performance is entirely dictated by the UO drivers' response. The control is considered effective if the resulting system performance outperforms this baseline.
- Theoretical lower bound (LB): Due to the complexity of the MFC problem, deriving the exact optimal control policy is often impossible. Nevertheless, in a special case with 100% penetration, the optimal system behavior is to attain and maintain system optimum from the second day onwards, regardless of the random initialization. It results in the lowest achievable system cost, hence we aim to explore whether the algorithm can approach this LB in cases with 100% SO users. It is important to note that

**Table 3**
Trained policy performance under 100% penetration.

| | Nguyen–Dupuis | Sioux Falls | Bottleneck | Bathtub | Joint choices |
|---|---|---|---|---|---|
| System cost | 81.18 | 16.34 | 133.09 | 105.58 | 47.92 |
| LB | 81.04 | 13.98 | 131.82 | 103.57 | 47.76 |
| Relative error(%) | 0.17 | 16.88 | 0.96 | 1.94 | 0.34 |

this LB is only attainable at 100% penetration. With a lower penetration level, immediately reaching the system optimum is not feasible due to the inertia in the response of UO drivers. For the routing experiments and the bottleneck experiment, the system optimum state can be calculated analytically. For the other two experiments, we numerically compute the population distribution that minimizes the system cost. The lower bound is then calculated by summing the expected cost from the random initialization and the system optimum cost over the subsequent 199 days.

### 4.3. Experiment results

This section presents the experiment results of the five problems in an idealized scenario with fully compliant SO users, and UO drivers following Smith's dynamic (Smith, 1984). In this dynamic, individuals switch to lower-cost options based on their experience from the previous day, as described by the following equation:

$$v_{t+1}(x) - v_t(x) = \eta \sum_{x' \in \mathcal{X}} \left( v_t(x')[c_{x'}(\mu_t, v_t) - c_x(\mu_t, v_t)]^+ - v_t(x)[c_x(\mu_t, v_t) - c_{x'}(\mu_t, v_t)]^+ \right).$$

As we normalize the daily travel cost, $\eta$ can be chosen independently of the problem at a value of 0.02, consistent with the setting in Guo et al. (2023). In addition, we train the RL control policy with the population-based kernel, as shown in Algorithm 2.

#### 4.3.1. Training results

Training curves for different penetration levels (10%, 50%, and 100%) are depicted in Fig. 2. Each solid curve represents the mean episode cost over three trials, with the shaded region indicating the standard deviation. The colored curves correspond to the different penetration rate settings, with the purple and red dotted lines representing the two benchmarks, respectively. While each figure illustrates the entire training process, the subplot zooms in to display the last quarter of training episodes, providing a clearer view of the final control performance.

Notably, the initial policy performance typically worsens with increasing penetration rates, except for the bottleneck model. This can be attributed to the RL algorithm's initial random policy, which may be less efficient compared to the pure UO user responses guided by Smith's dynamic. It is further investigated in Fig. 3, which offers a comparative analysis of a random action against the pure UO users' response. Under Smith's dynamic, the population behavior is either stabilized at user equilibrium as in Fig. 3(a) and (b), or driven to a more efficient region as in 3(d) and (e). The only exception is the bottleneck model, where even implementing a random action outperforms the chaotic UO user response. This contributes to the different pattern in Fig. 2(c), where the initial system performance improves with a higher penetration rate.

Throughout the training, the costs decrease for all scenarios, showing a significant control performance improvement regardless of choice scenarios and penetration rates. In particular, the green curves (100% penetration) consistently converge to the theoretical lower bound. This demonstrates the potential of the proposed scheme to leverage the control capabilities of the SO users. The quantitative measurements of training outcomes are given in Table 3. While the Sioux Falls network experiment produces an acceptable training result, it exhibits relatively poorer performance compared to other experiments. This is because it involves higher dimensions in the state and action spaces due to its larger scale, which naturally makes the algorithm harder to train. Nevertheless, compared with the pure UO driver response (episodic cost of 48.09), the trained policy is already able to achieve 93% of the maximum possible reduction, showcasing the effectiveness of the proposed method. Reducing the penetration rate to 50% (orange curves) and 10% (blue curves) leads to a decline in system performance, indicating a trade-off between control capabilities and intensity. Notably, even at a low penetration of 10%, the system performance represented by the blue curves consistently surpasses the benchmark with no SO users.

Despite the varying complexity and scale of the experiments, these results demonstrate that our proposed control scheme can effectively enhance system efficiency across a range of choice scenarios, congestion technologies, and penetration rates.

#### 4.3.2. Policy implementation

To visually demonstrate the effectiveness of the trained policy, we apply it on 100% and 50% penetration scenarios and track the system behavior over the first 5 days.

Fig. 4 shows the link flow evolution for the routing experiment on the Nguyen–Dupuis network. The average link flow of three trails is plotted using the solid curve, and the standard deviation is shown in the shaded region. The dotted lines represent the system optimum link flow.

In the 100% penetration scenario (left figure), regardless of the initial flow patterns, the link flows consistently converge towards the system optimum by the second day and maintain these levels on subsequent days. This consistency demonstrates the capability of the proposed control scheme in managing traffic dynamics and directing them towards an optimal state. Conversely, due to the

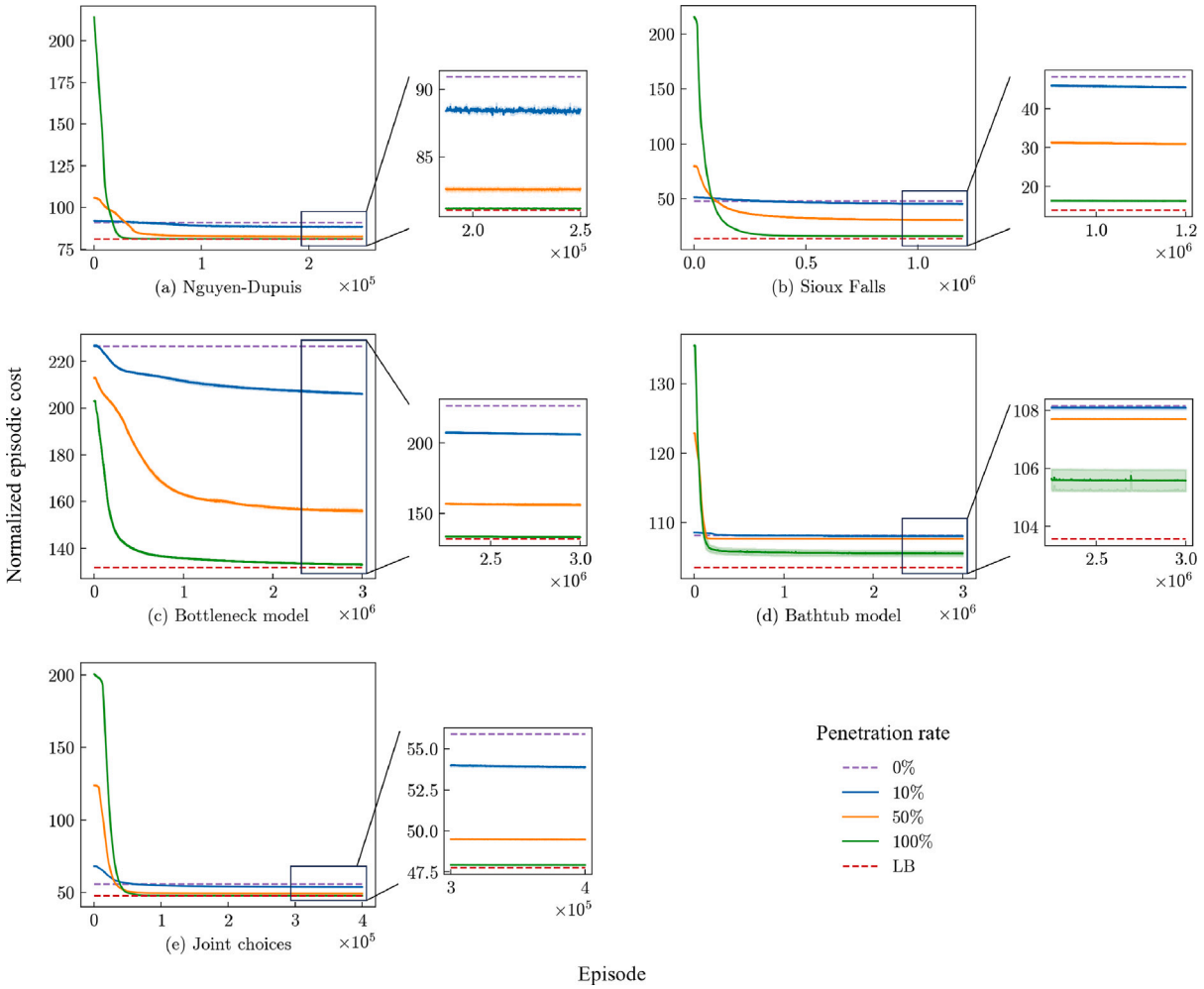


**Fig. 2.** Training curves (normalized episodic cost) under Smith's dynamic and the population kernel with standard deviation (three trials, shaded). (a) Routing on the Nguyen–Dupuis network; (b) Routing on the Sioux Falls network; (c) Departure time choices using the bottleneck model; (d) Departure time choices using the bathtub model; (e) Joint choices of route and departure time.

reduced control intensity at the 50% penetration rate (right figure), the link flow approaches but does not precisely align with the system optimum pattern.

Furthermore, we apply the trained policy to a departure time choice scenario using the bottleneck model. Fig. 5 illustrates the cumulative departure rate evolution of the entire population, with solid curves representing daily departure profiles and the red dotted line marking the system optimum pattern. For a clearer view of convergence, Fig. 6 shows the daily departure profiles' deviation from the system optimum, with a perfectly horizontal curve indicating an optimal system state. The results indicate that under full control of all vehicles, the departure profile aligns with the system optimum after approximately four days. This slower convergence, compared to the Nguyen–Dupuis case, is attributed to the larger state space resulting from the discretized departure time window—a challenge also reflected in the marginally larger error noted in Table 3. At the 50% penetration rate, the control scheme fails to achieve the system optimum, instead maintaining a peak around the desired arrival time.

### *4.3.3. Influence of penetration rates*

To further demonstrate the impact of varying penetration rates on the proposed control scheme, Fig. 7 presents the final control performance across different penetrations, ranging from 0% to 100%, on the Nguyen–Dupuis network. The results indicate a clear monotonic improvement in control performance with higher penetration rates. However, the convex pattern suggests diminishing marginal benefits at higher rates. This finding is crucial for traffic management authorities, as it suggests a threshold beyond which increasing SO user penetration may yield limited additional benefits.

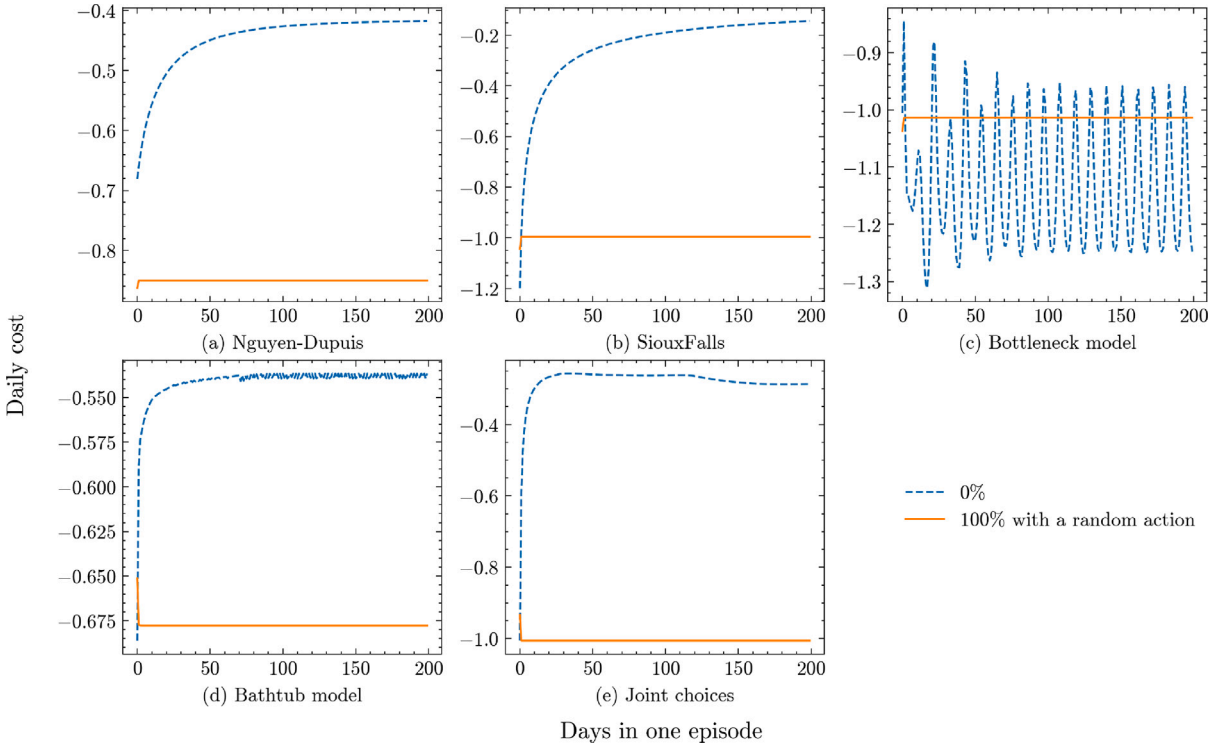


**Fig. 3.** Initial policy performance. (a) Routing on the Nguyen–Dupuis network; (b) Routing on the Sioux Falls network; (c) Departure time choices using the bottleneck model; (d) Departure time choices using the bathtub model; (e) Joint choices of route and departure time.

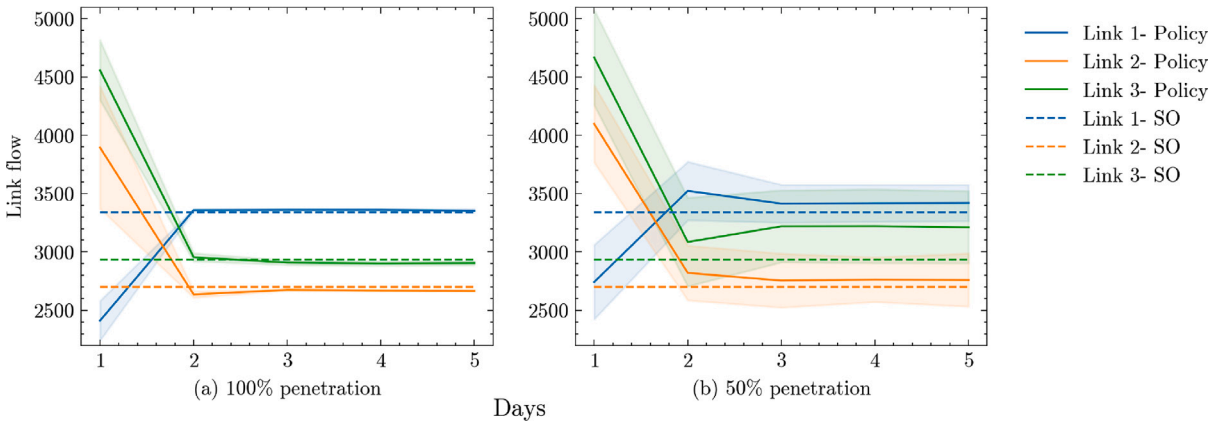


**Fig. 4.** Nguyen–Dupuis network: link flow evolution and standard deviation (three trails, shaded) under the trained policy.

#### *4.3.4. Cost comparison between two groups*

Although the recruitment of SO users is beyond the scope of this paper, analyzing the travel costs for both SO and UO users enhances the understanding of the control scheme and offers valuable insights into the incentive designs. Fig. 8 illustrates the ratio of daily travel costs between SO and UO users over several days, with different colors distinguishing between two penetration rate settings.

Notably, at a high penetration rate of 50%, the high control intensity allows SO users to improve network efficiency while sacrificing only 20% more in travel costs. Conversely, at a low penetration rate (10%), SO users may need to experience up to 70% more travel costs to counteract the selfish driving behavior of the large proportion of UO users. This outcome highlights the dual benefits of increasing SO user recruitment: improving overall system performance and reducing the need for SO users to make significant sacrifices.

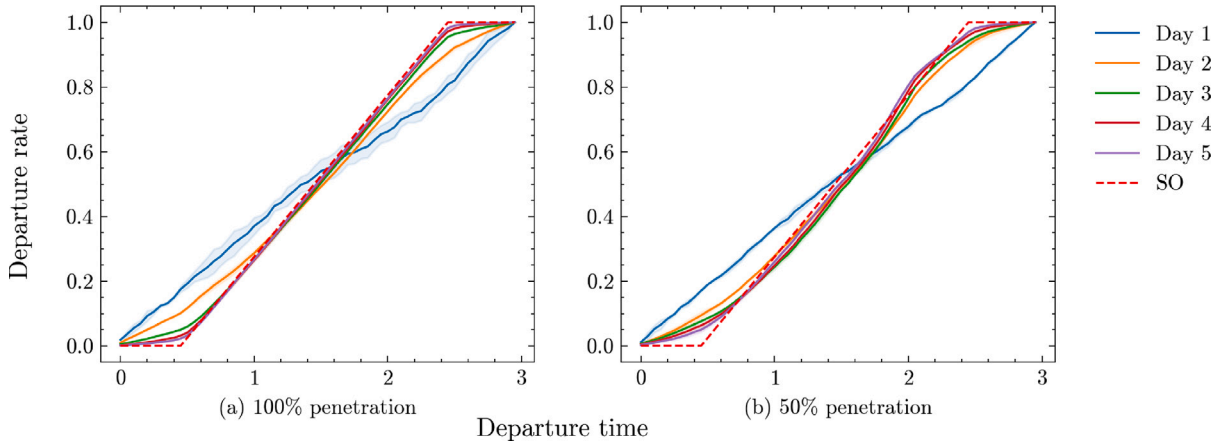


**Fig. 5.** Bottleneck model: cumulative departure rate evolution under the trained policy.

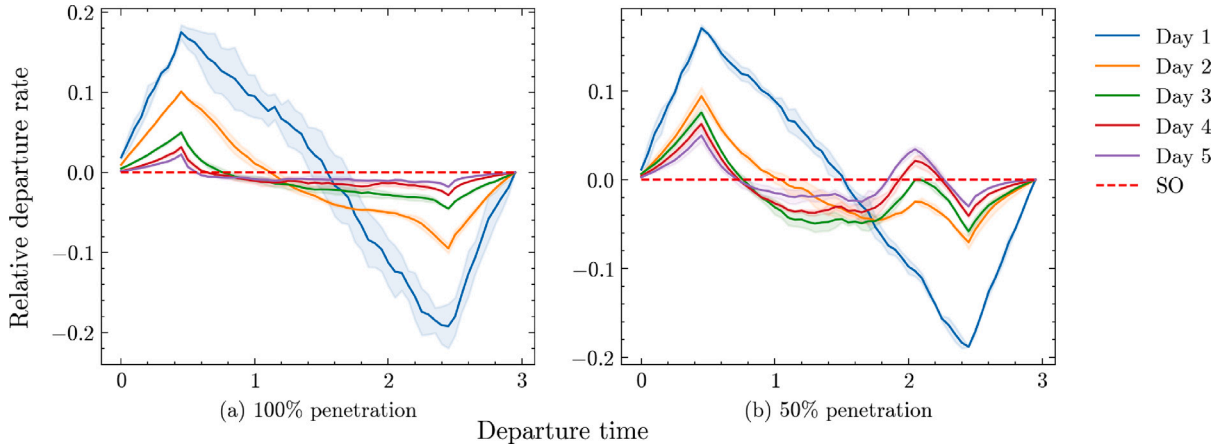


**Fig. 6.** Bottleneck model: deviation from the system optimum with standard deviation (three trails, shaded).

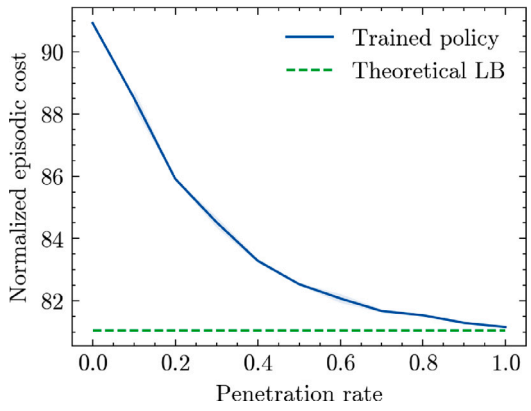


**Fig. 7.** Nguyen–Dupuis network: influence of penetration rate on the final control performance.

### *4.3.5. Influence of partial compliance*

As mentioned in Section 4.3, so far we only consider fully compliant SO users. This section explores the influence of partial compliance due to inertia as mentioned in Section 2.1. Fig. 9 depicts the training curves with 100% and 20% compliant SO users, both at a 50% penetration rate. Here, 20% compliance means that SO users only have 20% probability following the assignment;

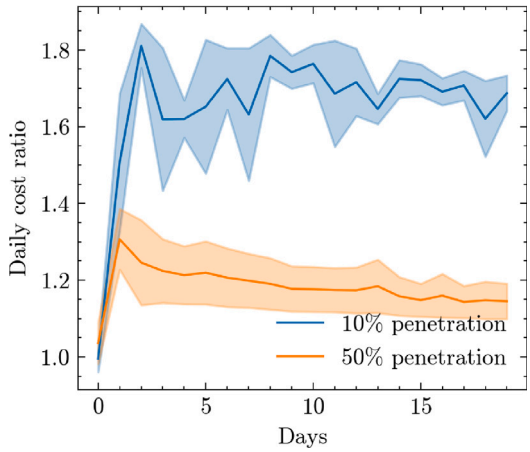


**Fig. 8.** Nguyen–Dupuis network: ratio of daily travel costs between SO and UO users and its standard deviation (3 trails, shaded).

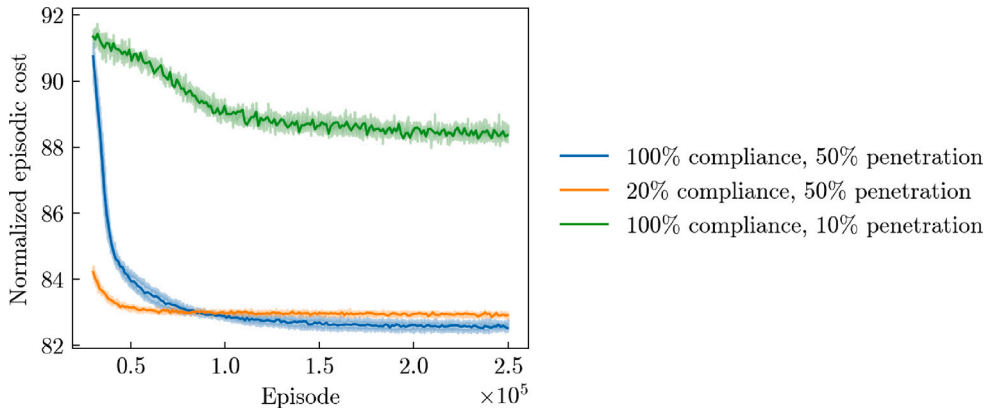


**Fig. 9.** Nguyen–Dupuis network: training curves (normalized episodic cost) and standard deviation (3 trails, shaded) under 50% penetration rate and different compliance rates.

otherwise, they will stick to the previous choice. Surprisingly, at a 50% penetration level, despite the orange curve (20% compliant) ending slightly higher than the blue curve (100% compliant), the difference is minor, less than 0.4%. This small discrepancy highlights that compliance rates have only a negligible impact on the final control performance. This robustness is due to the adaptability of the RL algorithm, which efficiently tailors policies to accommodate varying levels of user compliance.

In addition, Fig. 9 plots the training curve with 100% compliant SO users, but only at a 10% penetration rate. Interestingly, the system with 50% partially compliant SO users greatly outperforms the one with 10% fully compliant users. This is because, while SO users might prefer to retain their previous choices, these choices are still influenced by earlier assignments, providing the control policy more room to impact the system. This result also indicates that, compared to the compliance rate, the penetration rate is a more important factor influencing control performance.

### *4.4. Extension with real-world conditions*

Our initial experiments utilize Smith's dynamic and the population kernel to establish the foundational applicability of the model. However, real-world systems often exhibit more complexity and noise, which poses implementation challenges. To address this, we conduct three additional experiments on the Nguyen–Dupuis network to test the robustness of the proposed method under more realistic conditions:

- System transition variation: The actual travelers are always finite, which may impact the accuracy of the population kernel in fully capturing system transitions. To address this limitation, we substitute the current kernel with the sampling-based kernel outlined in Algorithm 3, while keeping the other components unchanged.
- Information limitation: The previous experiments utilize MF distributions, essentially path flow, as the policy input. In practice, agencies usually only have access to the link flow data. To capture this, we modify the model into a partially observable Markov Decision Process (POMDP), where the population state $(\mu_t, \nu_t)$ is now unobservable. Instead, travelers and management agencies

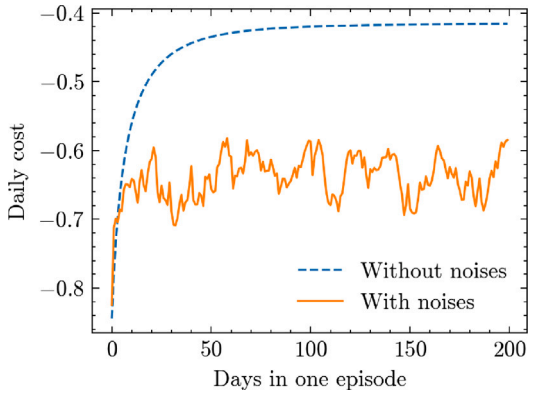


**Fig. 10.** Influence of the noises in the response dynamic.

**Table 4**
Comparison of training results in different scenarios.

| Scenario | Original | Sampling-based kernel | | Link flow-based policy | | Noisy response dynamics | |
|---|---|---|---|---|---|---|---|
| | | Results | Error (%) | Results | Error (%) | Results | Error (%) |
| 10% | 88.53 | 88.49 | −0.05 | 88.10 | −0.49 | 110.2 | 24.48 |
| 50% | 82.63 | 82.92 | 0.35 | 83.24 | 0.74 | 84.02 | 1.68 |
| 100% | 81.18 | 82.89 | 2.11 | 81.20 | 0.02 | 81.19 | 0.01 |

only observe the compressed information, i.e. link flow $\Delta(\theta\mu_t + (1-\theta)\nu_t)$, where $\Delta$ denotes the link-path incidence matrix. The population policy $\hat{\pi}$ now maps from observations, rather than states, to population actions $h_t$. Despite these modifications, we still use the population kernel and Smith's dynamic in this experiment.

- Response dynamics with noises: While Smith's dynamic is commonly used in literature, it remains an idealized representation of human behavior. In reality, human behavior may exhibit more randomness and unpredictability than the model assumes. To address this, we introduce random noises into the transition process, while leaving the other components unchanged:

$$c'(\mu_t, \nu_t) = c(\mu_t, \nu_t) + \bar{c},$$
$$\nu_{t+1}(x) - \nu_t(x) = \eta \sum_{x' \in \mathcal{X}} \left( \nu_t(x')[c'_{x'}(\mu_t, \nu_t) - c'_x(\mu_t, \nu_t)]^+ - \nu_t(x)[c'_x(\mu_t, \nu_t) - c'_{x'}(\mu_t, \nu_t)]^+ \right),$$

where $\bar{c}$ is a random vector with each element following a standard normal distribution. This dynamic introduces random residue to the daily travel cost, which influences the evolution process. Fig. 10 displays the impact of such noise on a system with solely UO drivers, highlighting increased fluctuation and imperfect convergence.

Fig. 11 shows the training curves of three experiments. The initial 50,000 episodes are excluded to better compare the control performance against the original experiment in Section 4.3. The training outcomes, detailed in Table 4, reveal that while modifications on the transition kernel or the information state introduce more fluctuation, they do not significantly affect final control performance. All the relative errors are below 2.11%, which underscores the applicability and robustness of the proposed method. Interestingly, using the link flow-based policy slightly improves training performance in the 10% penetration experiment, due to a trade-off between information richness and trainability. While using link flow as input reduces the information obtained by the RL agent, it lowers the state dimension from 50 to 19, making the policy easier to train. This improvement becomes more significant for complex training tasks, such as the low penetration scenario.

In contrast, adding noise to Smith's dynamics alters the system's behavior. In a 100% controlled environment, this change is negligible as there is no UO user. However, as the proportion of UO users increases, the added noise prevents SO users from perfectly influencing the system, which creates a notable gap in the outcomes. It is important to note that this gap is not a failure of the control policy but an inherent consequence of the noisy dynamic.

## 5. Conclusion and future work

In the transition to the connected and automated mobility era, a significant opportunity arises to harness the shift in travel agency to enhance the efficiency of our transportation systems. We can accomplish this by encouraging travelers to willingly relinquish certain aspects of their agency for the distributed control. In this research, we have proposed a pioneering traffic control scheme that leverages CAVs to subtly influence the day-to-day adjustment process of human drivers, thereby enhancing overall system performance. Our approach began with a finite-agent control model, which was subsequently reformulated as a mean-field control problem by considering the limiting case. We incorporated traveler heterogeneity by introducing a common action space

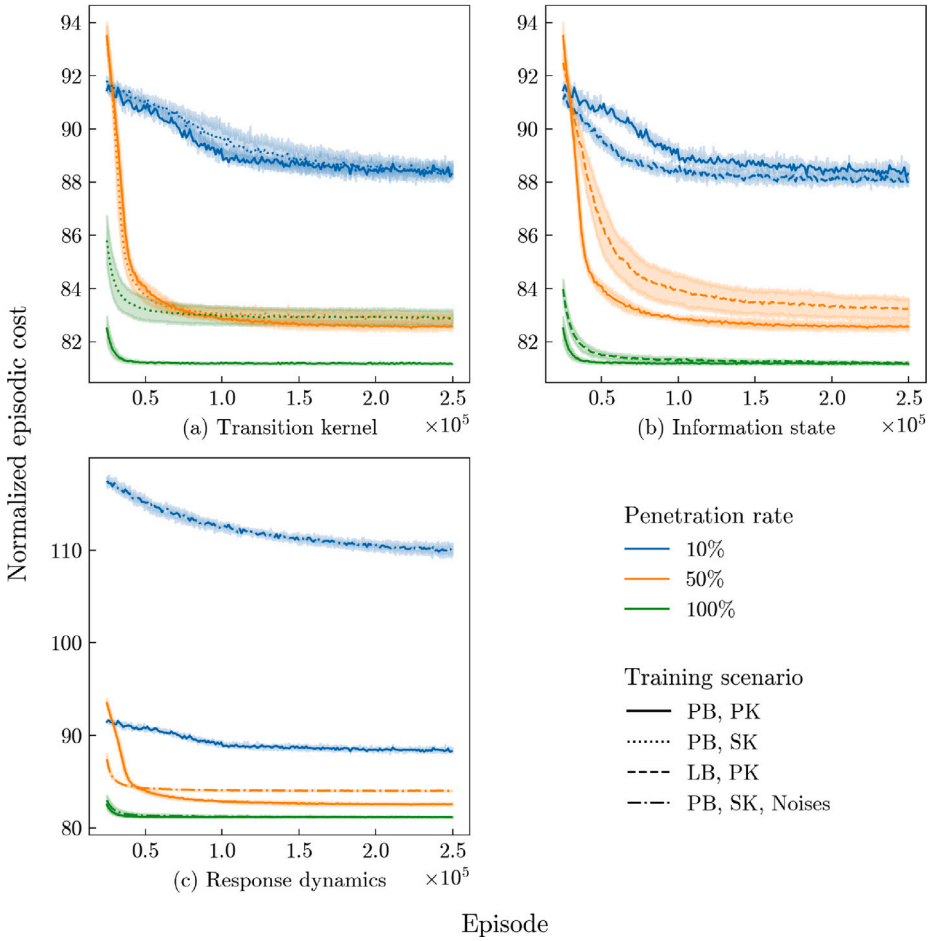


**Fig. 11.** Training curves (normalized episodic cost) with standard deviation (three trails, shaded) under different settings: Path-based (PB) policy against Link-based (LB) policy; Transition under population kernel (PK) against sampling-based kernel (SK); Response dynamics with or without noises.

and augmenting the state. After theoretically establishing the existence of the optimal policy, we employed reinforcement learning to numerically solve the control problem.

Distinct from traditional approaches, our proposed method is distributed, model-free, and does not rely on the convergence properties of day-to-day traffic dynamics. By formulating the problem as a mean-field control model, we address the issue of tractability even in scenarios involving a large number of travelers, thereby significantly increasing the scalability of the control scheme. Through various numerical examples, we have demonstrated that the proposed control scheme can effectively enhance system efficiency across a range of choice scenarios, congestion technologies, and penetration rates.

For future work, we plan to further enhance the model and apply it to larger scale networks to demonstrate the broader applicability of the proposed method. In addition, it is interesting to explore the potential synergies between our proposed control scheme and other traditional traffic management methods, such as congestion pricing. Investigating how these different control strategies can be integrated will help achieve even greater improvements in traffic efficiency. Furthermore, it is also important to design incentive mechanisms to encourage CAV owner participation.

## CRediT authorship contribution statement

**Minghui Wu:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Ben Wang:** Writing – review & editing, Software, Methodology. **Yafeng Yin:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Jerome P. Lynch:** Writing – review & editing, Supervision.

## Acknowledgments

The work described in this paper was partly supported by research grants from National Science Foundation, United States (CMMI-1904575, 2233057 and 2240981).

## Appendix. Notations

| | |
|---|---|
| ***Sets*** | |
| $\mathcal{X}$ | Travel choices |
| $\mathcal{K}$ | Types |
| $\mathcal{A}^k$ | Travel choices for type $k$ |
| $\bar{\mathcal{A}}$ | The common action space |
| $\mathcal{Z}$ | ID variables |
| ***Variables*** | |
| $x_t^i$ | State of SO user $i$ on day $t$ |
| $s_t^j$ | State of UO user $j$ on day $t$ |
| $\mu_t^N$ | Empirical state distribution among SO users on day $t$ |
| $\nu_t^M$ | Empirical state distribution among UO users on day $t$ |
| $a_t^i$ | Assigned option for SO user $i$ on day $t$ |
| $\mu_t$ | Mean-field distribution among SO users on day $t$ |
| $\nu_t$ | Mean-field distribution among UO users on day $t$ |
| $h_t$ | Joint state–action distribution among SO users on day $t$ |
| $k$ | Type $k$ |
| $z^i$ | ID variable for traveler $i$ |
| ***Parameters*** | |
| $N$ | Number of SO users |
| $M$ | Number of UO users |
| $\theta$ | Penetration rate of SO users |
| $\gamma$ | Discount factor |
| ***Functions*** | |
| $q(\mu_t, \nu_t)$ | UO users' response dynamics |
| $\pi(\cdot\|x_t, \mu_t, \nu_t)$ | Assignment policy |
| $p(\cdot\|x_t, a_t, \mu_t, \nu_t)$ | Transition kernel for each SO user |
| $c_x(\mu_t, \nu_t)$ | Travel cost of choice $x$ |
| $C(\mu_t, \nu_t)$ | System's average travel cost |
| $T(\mu_t, \nu_t, h_t)$ | Population transition kernel |
| $\hat{\pi}(\cdot\|\mu_t, \nu_t, h_t)$ | Population policy |
| $\mathcal{G}^k(a)$ | Mapping for type $k$ to incorporate heterogeneity |
| $k(x)$ | The type of travelers that can choose state $x \in \mathcal{X}$ |
| $\hat{p}(\cdot\|x_t, a_t, \mu_t, \nu_t)$ | Modified transition kernel to incorporate heterogeneity |
| $\hat{T}(\mu_t, \nu_t, h_t)$ | Modified population kernel to incorporate heterogeneity |

## References

Arnott, R., Buli, J., 2018. Solving for equilibrium in the basic bathtub model. Transp. Res. B 109, 150–175.

Chen, Z., Lin, X., Yin, Y., Li, M., 2020. Path controlling of automated vehicles for system optimum on transportation networks with heterogeneous traffic stream. Transp. Res. C 110, 312–329.

Čičić, M., Xiong, X., Jin, L., Johansson, K.H., 2021. Coordinating vehicle platoons for highway bottleneck decongestion and throughput improvement. IEEE Trans. Intell. Transp. Syst. 23 (7), 8959–8971.

Cui, K., Fabian, C., Koeppl, H., 2023. Multi-agent reinforcement learning via mean field control: Common noise, major agents and approximation properties. arXiv preprint arXiv:2303.10665.

Cui, S., Seibold, B., Stern, R., Work, D.B., 2017. Stabilizing traffic flow via a single autonomous vehicle: Possibilities and limitations. In: 2017 IEEE Intelligent Vehicles Symposium. IV, IEEE, pp. 1336–1341.

Cui, K., Tahir, A., Sinzger, M., Koeppl, H., 2021. Discrete-time mean field control with environment states. In: 2021 60th IEEE Conference on Decision and Control. CDC, IEEE, pp. 5239–5246.

De Palma, G., Marvian, M., Trevisan, D., Lloyd, S., 2021. The quantum Wasserstein distance of order 1. IEEE Trans. Inform. Theory 67 (10), 6627–6643.

Di, X., Shi, R., 2021. A survey on autonomous vehicle control in the era of mixed-autonomy: From physics-based to AI-guided driving policy learning. Transp. Res. C 125, 103008.

Friesz, T.L., Han, K., Bagherzadeh, A., 2021. Convergence of fixed-point algorithms for elastic demand dynamic user equilibrium. Transp. Res. B 150, 336–352.

Friesz, T.L., Kim, T., Kwon, C., Rigdon, M.A., 2011. Approximate network loading and dual-time-scale dynamic user equilibrium. Transp. Res. B 45 (1), 176–207.

Greenshields, B.D., Bibbins, J., Channing, W., Miller, H., 1935. A study of traffic capacity. In: Highway Research Board Proceedings, vol. 14, (no. 1), Washington, DC, pp. 448–477.
Guo, Z., Wang, D.Z., Wang, D., 2022. Managing mixed traffic with autonomous vehicles–A day-to-day routing allocation scheme. Transp. Res. C 140, 103726.
Guo, R.-Y., Yang, H., Huang, H.-J., 2018a. Are we really solving the dynamic traffic equilibrium problem with a departure time choice? Transp. Sci. 52 (3), 603–620.
Guo, R.-Y., Yang, H., Huang, H.-J., 2023. The day-to-day departure time choice of heterogeneous commuters under an anonymous toll charge for system optimum. Transp. Sci..
Guo, R.-Y., Yang, H., Huang, H.-J., Li, X., 2018b. Day-to-day departure time choice under bounded rationality in the bottleneck model. Transp. Res. B 117, 832–849.
Huang, M., Malhamé, R.P., Caines, P.E., 2006. Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. Commun. Inform. Syst. 6 (3), 221–252.
Jin, I.G., Orosz, G., 2014. Dynamics of connected vehicle systems with delayed acceleration feedback. Transp. Res. C 46, 46–64.
Jin, I.G., Orosz, G., 2016. Optimal control of connected vehicle systems with communication delay and driver reaction time. IEEE Trans. Intell. Transp. Syst. 18 (8), 2056–2070.
Jin, I.G., Orosz, G., 2018. Connected cruise control among human-driven vehicles: Experiment-based parameter estimation and optimal control design. Transp. Res. C 95, 445–459.
Lamotte, R., Geroliminis, N., 2018. The morning commute in urban areas with heterogeneous trip lengths. Transp. Res. B 117, 794–810.
Lasry, J.-M., Lions, P.-L., 2007. Mean field games. Japn. J. Math. 2 (1), 229–260.
Lazar, D.A., Bıyık, E., Sadigh, D., Pedarsani, R., 2021. Learning how to dynamically route autonomous vehicles on shared roads. Transp. Res. C 130, 103258.
LeBlanc, L.J., Morlok, E.K., Pierskalla, W.P., 1975. An efficient approach to solving the road network equilibrium traffic assignment problem. Transp. Res. 9 (5), 309–318.
Li, R., Liu, X., Nie, Y.M., 2018. Managing partially automated network traffic flow: Efficiency vs. stability. Transp. Res. B 114, 300–324.
Liang, Q., Li, X.-a., Chen, Z., Pan, T., Zhong, R., 2023. Day-to-day traffic control for networks mixed with regular human-piloted and connected autonomous vehicles. Transp. Res. B 178, 102847.
Mounce, R., Carey, M., 2011. Route swapping in dynamic traffic networks. Transp. Res. B 45 (1), 102–111.
Nguyen, S., Dupuis, C., 1984. An efficient method for computing traffic equilibria in networks with asymmetric transportation costs. Transp. Sci. 18 (2), 185–202.
Bureau of Public Roads, 1964. Traffic Assignment Manual. U.S. Dept. of Commerce, Urban Planning Division.
Qi, H., Jia, N., Qu, X., He, Z., 2023. Investigating day-to-day route choices based on multi-scenario laboratory experiments, Part I: Route-dependent attraction and its modeling. Transp. Res. A 167, 103553.
Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
Sharon, G., Albert, M., Rambha, T., Boyles, S., Stone, P., 2018. Traffic optimization for a mixture of self-interested and compliant agents. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, (no. 1).
Smith, M.J., 1984. The stability of a dynamic model of traffic assignment—an application of a method of Lyapunov. Transp. Sci. 18 (3), 245–252.
Srinivasan, K.K., Mahmassani, H.S., 2000. Modeling inertia and compliance mechanisms in route choice behavior under real-time information. Transp. Res. Rec. 1725 (1), 45–53.
Stern, R.E., Cui, S., Delle Monache, M.L., Bhadani, R., Bunting, M., Churchill, M., Hamilton, N., Pohlmann, H., Wu, F., Piccoli, B., et al., 2018. Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments. Transp. Res. C 89, 205–221.
Vinitsky, E., Kreidieh, A., Le Flem, L., Kheterpal, N., Jang, K., Wu, C., Wu, F., Liaw, R., Liang, E., Bayen, A.M., 2018. Benchmarks for reinforcement learning in mixed-autonomy traffic. In: Conference on Robot Learning. PMLR, pp. 399–409.
Wu, C., Bayen, A.M., Mehta, A., 2018. Stabilizing traffic with autonomous vehicles. In: 2018 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 6012–6018.
Wu, C., Kreidieh, A.R., Parvate, K., Vinitsky, E., Bayen, A.M., 2021. Flow: A modular learning framework for mixed autonomy traffic. IEEE Trans. Robot. 38 (2), 1270–1286.
Xiong, C., Shahabi, M., Zhao, J., Yin, Y., Zhou, X., Zhang, L., 2020. An integrated and personalized traveler information and incentive scheme for energy efficient mobility systems. Transp. Res. C 113, 57–73.
Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., Wang, J., 2018. Mean field multi-agent reinforcement learning. In: International Conference on Machine Learning. PMLR, pp. 5571–5580.
Yin, Y., Lam, W.H., Ieda, H., 2004. New technology and the modeling of risk-taking behavior in congested road networks. Transp. Res. C 12 (3–4), 171–192.
Zhang, F., Lu, J., Hu, X., 2022. Integrated path controlling and subsidy scheme for mobility and environmental management in automated transportation networks. Transp. Res. E 167, 102906.
Zhang, K., Nie, Y.M., 2018. Mitigating the impact of selfish routing: An optimal-ratio control scheme (ORCS) inspired by autonomous driving. Transp. Res. C 87, 75–90.
Zhang, K., Yang, Z., Başar, T., 2021. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In: Handbook of Reinforcement Learning and Control. Springer, pp. 321–384.
Zheng, Y., Wang, J., Li, K., 2020. Smoothing traffic flow via control of autonomous vehicles. IEEE Internet Things J. 7 (5), 3882–3896.