A ROBUST AND CONSTRAINED MULTI-AGENT REINFORCEMENT LEARNING METHOD FOR ELECTRIC VEHICLE REBALANCING IN AMOD SYSTEMS

Sihong He[1], **Yue Wang**[2], **Shuo Han**[3], **Shaofeng Zou**[2], **Fei Miao**[1]
Department of Computer Science and Engineering, University of Connecticut[1]
Department of Electrical Engineering, University at Buffalo, The State University of New York[2]
Department of Electrical and Computer Engineering, University of Illinois, Chicago[3]
{sihong.he, fei.miao}@uconn.edu,
{ywang294, szou3}@buffalo.edu, hanshuo@uic.edu

ABSTRACT

Electric vehicles (EVs) play critical roles in autonomous mobility-on-demand (AMoD) systems, but their unique charging patterns increase the model uncertainties in AMoD systems (e.g. state transition probability). Since there usually exists a mismatch between the training and test/true environments, incorporating model uncertainty into system design is of critical importance in real-world applications. However, model uncertainties have not been considered explicitly in EV AMoD system rebalancing by existing literature yet, and the coexistence of model uncertainties and constraints that the decision should satisfy makes the problem even more challenging. In this work, we design a robust and constrained multi-agent reinforcement learning (MARL) framework with state transition kernel uncertainty for EV AMoD systems. We then propose a robust and constrained MARL algorithm (ROCOMA) that trains a robust EV rebalancing policy to balance the supply-demand ratio and the charging utilization rate across the city under model uncertainty. Experiments show that the ROCOMA can learn an effective and robust rebalancing policy. It outperforms non-robust MARL methods in the presence of model uncertainties. It increases the system fairness by 19.6% and decreases the rebalancing costs by 75.8%.

1 INTRODUCTION

Figure 1: Unbalanced demand and supply happen several times a day. For example, at morning peak, there are more commutes from the residential area to the work zone. While at evening peak, there are more needs to leave the work area to recreational area/home.

The autonomous mobility-on-demand (AMoD) system is one of the most promising energy-efficient transportation solutions as it provides people with one-way rides from their origins to destinations (Zardini et al., 2021). Electric vehicles (EVs) are being adopted worldwide for environmental and economical benefits (IEA, 2020), and AMoD systems embrace this trend without exception. However, as shown in Fig. 1, the trips sporadically appear, and the origins and destinations are asymmetrically distributed. Such spatial-temporal nature of urban mobility motivates researchers to study vehicle rebalancing methods (Wen et al., 2017; He et al., 2020), i.e. redistribution of vacant EVs to areas of high demand and assigning low-battery EVs to charging stations.

In real-world AMoD systems, the simulation-to-reality gap remains challenging for vehicle rebalancing solutions calculated based on simulators, since there usually exists a model mismatch between the simulator (training environment) and the real world (test environment). For instance, at time t, with the system state information such as the number of available vehicles and passenger demand in each region of the city, and the action to take as the number of available vehicles to be balanced

among regions according to the mobility demand, it is difficult to accurately predict the state of the system (available vehicle supply and mobility demand) at $t+1$ (Zardini et al., 2021; Miao et al., 2021; Parys et al., 2016). Hence, we usually do not have the true dynamic model of the system, i.e., the state transition probability of the AMoD systems. Thus, existing EV AMoD vehicle rebalancing methods (Yuan et al., 2019; Sadeghianpourhamami et al., 2020; Turan et al., 2020; Wen et al., 2017) may have significant performance degradation in the test (true) environment. One example is provided in Fig. 2. Moreover, in real-world applications, the vehicle rebalancing decisions should satisfy specific constraints such as providing fair mobility service in different regions; when there is model mismatch, the algorithm solution calculated based on a simulator may violate the constraints in real AMoD systems. Despite model-based methods considering prediction errors in mobility demand or vehicle supply (Zhang et al., 2016; He et al., 2020; Miao et al., 2021; Hao et al., 2020; He et al., 2023), how to calculate policies that satisfy the constraints and optimize the objectives under model uncertainty of the dynamic state transition remains largely unexplored for AMoD rebalancing algorithms. More related work is discussed in the appendix due to the page limit.

In this work, to address the simulation-to-reality gap and calculate solutions that satisfy the constraints, we propose a robust and constrained multi-agent reinforcement learning (MARL) framework for EV AMoD systems. The goal is to find robust policies that minimize the rebalancing cost of the vacant and low-battery EVs under model uncertainties and achieve mobility and charging fairness. The advantages of our methodology are two-fold: (i) fairness constraints can still be satisfied even if there exists model mismatch; and (ii) the expected rebalancing cost is still optimized when there is model mismatch. Our *Key Contributions* are as follows:

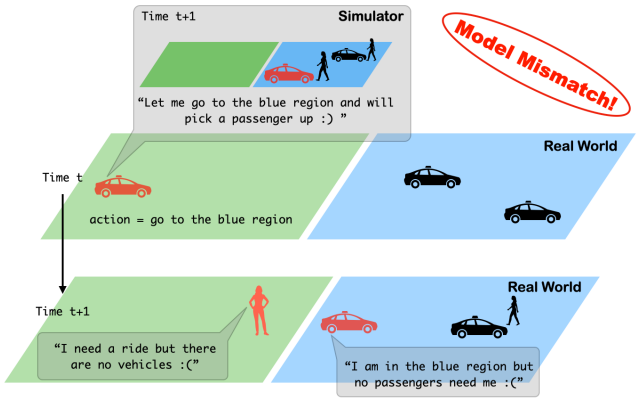

Figure 2: The model mismatch between the simulator and the real world degrades the performance of vehicle rebalancing methods. The red EV chooses to go to the blue region at time t and thinks it can pick up a passenger at time $t+1$ according to the simulator model. However, in the real world, at time $t+1$, the red EV gets no passengers in the blue region and a passenger gets no cars in the green region.

(1) To the best of our knowledge, this work is the first to formulate EV AMoD system vehicle rebalancing as a robust and constrained multi-agent reinforcement learning problem under model uncertainty. Via a proper design of the state, action, reward, cost constraints, and uncertainty set, we set our goal as minimizing the rebalancing cost while balancing the city's charging utilization and service quality, under model uncertainty.

(2) We design a robust and constrained MARL algorithm (ROCOMA) to efficiently train robust policies. The proposed algorithm adopts the centralized training and decentralized execution (CTDE) framework. We also develop the robust natural policy gradient (RNPG) in MARL for the first time.

(3) We run experiments based on real-world E-taxi system data. We show that our proposed algorithm performs better in terms of reward and fairness, which are increased by 19.6%, and 75.8%, respectively, compared with a non-robust MARL-based method when model uncertainty is present.

2 Robust and Constrained MARL Framework for EV Rebalancing

2.1 Problem Statement

We consider the problem of managing a large-scale EV fleet to provide fair and robust AMoD service. The goal is to (i) rebalance vacant EVs among different regions to provide fair mobility service on the passenger's side; (ii) allocate low-battery EVs to charging stations for fair charging service on the EVs' side; (iii) minimize the managing cost of (i) and (ii). These three goals need to be achieved

in the presence of model uncertainties, i.e. uncertainties in the state transition probability model of AMoD systems.

We divide the city into N regions according to a pre-defined partition method (Miao et al., 2019; Turan et al., 2020; He et al., 2020). A day is divided into equal-length time intervals. In each time interval $[t, t+1)$, customers' ride requests and EVs' charging needs are aggregated in each region. After the location and status of each EV are observed, a local trip and charging assignment algorithm matches vacant EVs with passengers and low-battery EVs with charging stations, using existing methods in the literature (Mourad et al., 2019; Chen et al., 2017). Then the state information of each region is updated, including the numbers of vacant EVs and available charging spots in each region. Each region then rebalances both vacant and low-battery EVs according to the well-trained MARL policy. This work focuses on a robust EV rebalancing algorithm design under model uncertainties to maximize the worst-case expected reward of the system while satisfying fairness constraints. For notational convenience, the parameters and variables defined in the following parts of this section omit the time index t when there is no confusion.

2.2 Preliminary: Multi-Agent Reinforcement Learning

We denote a Multi-Agent Reinforcement Learning (MARL) problem by a tuple $G = \langle \mathcal{N}, S, A, r, p, \gamma \rangle$, in which $\mathcal{N}$ is the set of N agents. Each agent i is associated with an action $a^i \in A^i$ and a state $s^i \in S$. We use $A = A^1 \times \cdots \times A^N$ to denote the joint action space, and $S = S^1 \times \cdots \times S^N$ the joint state space. At time t, each agent chooses an action a_t^i according to a policy $\pi^i : S^i \to \Delta(A^i)$, where $\Delta(A^i)$ represents the set of probability distributions over the action set A^i. We use $\pi = \prod_{i=1}^N \pi^i : S \to \Delta(A)$ to denote the joint policy. After executing the joint action is executed, the next state follows the state transition probability which depends on the current state and the joint action, i.e. $p : S \times A \to \Delta(S)$. And each agent receives a reward according to the reward function $r^i : S \times A \to \mathbb{R}$. Each agent aims to learn a policy π^i to maximize its expected total discounted reward, i.e. $\max_{\pi^i} v_r^{\pi,i}(s)$ for all $s \in S$, where $v_r^{\pi,i}(s) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t^i(s_t, a_t) | a_t \sim \pi(\cdot|s_t), s_1 = s]$ which is also known as the state value function for agent i. $\gamma \in (0, 1)$ is the discounted rate. When these agents belong to a team, the objective of all agents is to collaboratively maximize the average expected total discounted reward over all agents, i.e. $\max_\pi v_r^\pi(s)$ for all $s \in S$, where $v_r^\pi(s) = \mathbb{E}_\pi[\sum_{t=1}^{\infty} \gamma^{t-1} \sum_{i \in \mathcal{N}} r_t^i(s_t, a_t)/N | s_1 = s]$.

2.3 Robust and Constrained Multi-Agent Reinforcement Learning Formulation for EV Rebalancing

We formulate the EV rebalancing problem as a robust and constrained MARL problem $G_{rc} = \langle \mathcal{N}, S, A, P, r, c, d, \gamma \rangle$, and we define the agent, state, action, probability transition kernel uncertainty set, reward, and cost and fairness constraints as follows.

Agent: We define a *region agent* for each region, who determines the rebalancing of vacant and low-battery EVs at every time step. This multi-agent setting is more tractable for large-scale fleet management than a single-agent setting because the action space can be prohibitively large if we use a single system-wide agent (Lin et al., 2018a).

State: A state s^i of a region agent i consists two parts that indicate its spatiotemporal status from both the local view and global view of the city. We define the state $s^i = \{s_{loc}^i, s_{glo}^i\}$, where $s_{loc}^i = (V_i, L_i, D_i, E_i, C_i)$ is the state of region i from the local view, denoting the number/amount of vacant EVs, low-battery EVs, mobility demand, empty charging spots, and total charging spots in region i, respectively. And $s_{glo}^i = (t, pos_i)$, where t is the time index (which time interval), pos_i is region location information (longitudes, latitudes, region index). The initial state distribution is ρ.

Action: The rebalancing action for vacant EVs is denoted as $a_v^i = \{a_{v,j}^i\}_{j \in \text{Nebr}_i}$, the charging action for low-battery EVs as $a_l^i = \{a_{l,j}^i\}_{j \in \text{Nebr}_i}$, where $a_{v,j}^i, a_{l,j}^i \in [0, 1]$ is the percentage of currently vacant EVs and low-battery EVs to be assigned to region j from region i, respectively. And Nebr_i is the set consisting of region i and its adjacent regions as defined by the given partition. Therefore $\sum_{j \in \text{Nebr}_i} a_{v,j}^i = 1$ and $\sum_{j \in \text{Nebr}_i} a_{l,j}^i = 1$ for all i. We denote $m_{v,j}^i = h(a_{v,j}^i v^i)$ the actual number of vacant EVs assigned from region i to region j, $m_{l,j}^i = h(a_{l,j}^i l^i)$ the actual number of low-battery

EVs in region i assigned to region j. The function $h(\cdot)$ is used to ensure that the numbers remain as integers and the constraints $\sum_j m^i_{v,j} = v^i, \sum_j m^i_{l,j} = l^i$ hold for all i.

Transition Kernel Uncertainty Set: We restrict the transition kernel p to a δ-contamination uncertainty set P (Ronchetti & Huber, 2009; Prasad et al., 2020), in which the state transition could be arbitrarily perturbed by a small probability δ. Specifically, let $\tilde{p} = \{\tilde{p}^a_s \mid s \in S, a \in A\}$ be the centroid transition kernel, from which training samples are generated. The δ-contamination uncertainty set centered at $\tilde{p}$ is defined as $P := \bigotimes_{s\in S, a\in A} P^a_s$, where $P^a_s := \{(1-\delta)\tilde{p}^a_s + \delta q \mid q \in \Delta(S)\}, s \in S, a \in A$.

Reward: Since one of our goals is to minimize the rebalancing cost, we define the shared reward as the negative value of the total rebalancing cost after EVs execute the decisions: $r(s,a) := -[c_v(s,a) + \bar{\alpha} c_l(s,a)]$, where $\bar{\alpha}$ is a positive coefficient, and $c_v(s,a), c_l(s,a)$ are moving distances of all vacant and low-battery EVs under the joint state s and action a, respectively. We then define the worst-case value function of a joint policy π as the worst-case expected total discounted reward under joint policy π over P: $v^\pi_r(s) = \min_{p\in P} \mathbb{E}_\pi \left[\sum_{t=1}^\infty \gamma^{t-1} r_t | s_1 = s\right]$. The notation is the same as MARL without considering uncertainty. By maximizing the shared worst-case value function, region agents are cooperating for the same goal.

Cost and Fairness Constraints: Another goal is to achieve the system-level benefit, i.e., balanced charging utilization and fair service. We define the charging fairness u_c and mobility fairness u_m in Subsection 2.4. If the values of these fairness metrics are higher than some thresholds by applying a rebalancing policy π, we say the policy π provides fair mobility and charging services among the city. We then augment the MARL problem G with an auxiliary cost function c, and a limit d. The function $c : S \times A \to \mathbb{R}$ maps transition tuples to cost, like the usual reward. Similarly, we let $v^\pi_c(s)$ denote the worst-case state value function of policy π with respect to cost function c: $v^\pi_c(s) = \min_{p\in P} \mathbb{E}_\pi[\sum_{t=1}^\infty \gamma^{t-1} c(s_t, a_t) | s_1 = s]$. The cost function c is defined as the system fairness (a weighted sum of city's charging fairness u_c and mobility fairness u_m), i.e., $c(s,a) := u_c(s,a) + \bar{\beta} u_m(s,a)$, where $\bar{\beta}$ is a positive coefficient. Then the set of feasible joint policies for our robust and constrained MARL EV rebalancing problem is $\Pi_C := \{\pi : \forall s \in S, v^\pi_c(s) \geq d\}$.

Goal: The goal of our robust and constrained MARL EV rebalancing problem is to find an optimal joint policy π^* that maximizes the worst-case expected value function subject to constraints on the worst-case expected cost:

$$\max_\pi \mathbb{E}_{s\sim\rho}[v^\pi_r(s)] \text{ s.t. } \mathbb{E}_{s\sim\rho}[v^\pi_c(s)] \geq d \tag{1}$$

We define $v^{\pi_\theta}_{\text{tp}}(\rho) = \mathbb{E}_{s\sim\rho}[v^{\pi_\theta}_{\text{tp}}(s)]$, $\text{tp} \in \{r, c\}$. We then consider policies $\pi(\cdot|\theta)$ parameterized by θ and consider the following equivalent max-min problem based on the Lagrangian (Boyd & Vandenberghe, 2004):

$$\max_\theta \min_{\lambda\geq 0} J(\theta, \lambda) := v^{\pi_\theta}_r(\rho) + \lambda(v^{\pi_\theta}_c(\rho) - d), \tag{2}$$

2.4 Fairness Definition

We consider both the mobility supply-demand ratio (Miao et al., 2021; Pfrommer et al., 2014; Wen et al., 2017) and the charging utilization rate (He et al., 2020; Wan et al., 2019) in each region as service quality metrics. With limited supply volume in a city, keeping the supply-demand ratio of each region at a similar level allows passengers in the city to receive fair service (Iglesias et al., 2019; Zhang et al., 2016). Similarly, given a limited number of charging stations and spots, to improve the charging service quality and charging efficiency with limited infrastructure, balancing the charging utilization rate of all regions across the entire city is usually one objective in the scheduling of EV charging (Wan et al., 2019; Sadeghianpourhamami et al., 2020).

The fairness metrics of the charging utilization rate u_c and supply-demand ratio u_m are designed based on the difference between the local and global quantities:

$$u_c(s,a) = -\sum_{i=1}^N \left| \frac{E_i}{C_i} - \frac{\sum_{j=1}^N E_j}{\sum_{j=1}^N C_j} \right|, u_m(s,a) = -\sum_{i=1}^N \left| \frac{D_i}{V_i} - \frac{\sum_{j=1}^N D_j}{\sum_{j=1}^N V_j} \right|,$$

where V_i is the number of vacant EVs in region i. The fairness metrics $u_s(s,a)$ and $u_m(s,a)$ are calculated given the EVs rebalancing action a, and the larger the better. One advantage of the

proposed robust and constrained MARL formulation is that the form of the reward/cost function does not need to satisfy the requirements as those of the robust optimization methods (Miao et al., 2019; Miao et al., 2021), e.g., the objective/constraints do not need to be convex of the decision variable or concave of the uncertain parameters.

3 ALGORITHM

3.1 ROBUST AND CONSTRAINED MULTI-AGENT REINFORCEMENT LEARNING ALGORITHM (ROCOMA)

We propose a robust and constrained MARL (ROCOMA) algorithm to solve the problem (2) and train robust policies. The proposed algorithm is shown in Algorithm 1. It adopts the centralized training and decentralized execution (CTDE) framework, which enables us to train agents in the simulator using global information but executes well-trained policies in a decentralized manner in the real world. Specifically, we use centralized critic networks to approximate the value functions and decentralized actor networks to represent policies. Besides, we develop a robust natural policy gradient (RNPG) descent ascent to update actor networks and the Lagrange multiplier.

As shown in Algorithm 1, in line 1, we randomly initialize the actor network parameter θ_0 and the Lagrange multiplier parameter λ_0. At each iteration t, in line 3, we estimate the critic networks $v_r^{\theta_t}, v_c^{\theta_t}$ under policy π^{θ_t} using Algorithm 3 in (Wang & Zou, 2022). Line 4 to line 14 are to estimate the robust natural policy gradient (RNPG) $\tilde{g}_{r,t}, \tilde{g}_{c,t}$ for $v_r^{\theta_t}$ and $v_c^{\theta_t}$, respectively. For notational convenience, we omit the subscripts r and c in the value functions when there is no confusion. In lines 5 and 6, we sample an initial state s_1^j following the initial distribution ρ and a time horizon T_j from the geometric distribution *Geom*$(1-\gamma+\gamma\delta)$ at iteration $j = 1, ..., M$. We use these samples to estimate the RNPG according to Corollary 3.1. Specifically, we initialize $\tilde{g}_{t,0}^j = 0$ and use the following stochastic gradient descent (SGD) steps: $\tilde{g}_{t,k+1}^j = \tilde{g}_{t,k}^j - \zeta \nabla_{\tilde{g}} \mathcal{L}(\tilde{g}_{t,k}^j, \theta_t)$, where ζ is the learning rate and $\mathcal{L}(\tilde{g}_{t,k}^j, \theta_t) = \sum_{\mathcal{D}(s_{T_j}^j)} [\tilde{g}^\top \psi^{\theta_t}(s,a) - \phi^{\theta_t}(\tau) - b^{\theta_t}]^2 / D$, $\mathcal{D}(s_{T_j}^j)$ is a set of trajectories τ starting at $s_{T_j}^j$ using policy π^{θ_t}, i.e. $\tau = (s_{T_j}^j, a, r, c, s')$, $D = |\mathcal{D}(s_{T_j}^j)|$. After W steps of SGD iterations, the robust natural policy gradient for $v^{\theta_t}(s_1^j)$ is estimated as $\sum_{k=1}^W \tilde{g}_{t,k}^j / W$.

To reduce the computational complexity, we adopt the centralized training and decentralized execution (CTDE) framework of Lowe & Wu (2017) in ROCOMA and assume all agents share the same policy $\pi^{\theta^i}(a^i|s^i)$, where $\theta^1 = \cdots = \theta^N = \theta$. Then we have $\nabla \pi(a|s) = \sum_i^N \psi_i^\theta(s,a)$ where $\psi_i^\theta(s,a) := \pi^{-i}(a^{-i}|s^{-i}) \nabla \pi^i(a^i|s^i)$, $\pi^{-i}(a^{-i}|s^{-i}) := \prod_{j \neq i} \pi^j(a^j|s^j)$. Therefore, in lines 7 to 12, we address the high-dimensional action and state space issue in computing RNPG by using $\psi_i^\theta(s,a)$ instead of $\psi^\theta(s,a)$ in (5). Finally, we update θ_{t+1} and λ_{t+1} using Gradient Descent Ascent (GDA) (Lin et al., 2020) in lines 15, 16.

3.2 ROBUST NATURAL POLICY GRADIENT

Natural policy gradient (NPG) (Schulman et al., 2015; Lillicrap et al., 2015; Mnih et al., 2015) applies a preconditioning matrix to the gradient, and updates the policy along the steepest descent direction in the policy space(Ding et al., 2020; Kakade, 2001). It has been proved that NPG moves toward choosing a greedy optimal action rather than just a better action in the literature (Kakade, 2001). Generally, for a function L defined on a Riemannian manifold Θ with a metric M, the steepest descent direction of L at θ is given by $-M^{-1}(\theta)\nabla L(\theta)$, which is called the natural gradient of L (Amari, 1998). In the policy parameter space $\{\pi_\theta\}$, the natural gradient of L at θ is given by $\tilde{\nabla} L(\theta) = F(\theta)^{-1} \nabla L(\theta)$, where $F(\theta) := \mathbb{E}_s [F_s(\theta)]$ is the Fisher information matrix at θ and $F_s(\theta) = \mathbb{E}_{\pi(a|s,\theta)} \left[\frac{\partial \log \pi(a|s,\theta)}{\partial \theta_i} \frac{\partial \log \pi(a|s,\theta)}{\partial \theta_j} \right]$ (Kakade, 2001). Although the natural gradient method has been studied in non-robust RL, it is not straightforward to efficiently find the NPG for a robust and constrained MARL problem. We show the robust natural policy gradient for robust and constrained MARL in the following Theorem 3.1.

Algorithm 1: Robust and Constrained Multi-Agent Reinforcement Learning Algorithm (RO-COMA)

Input $\zeta, \alpha, \beta, \gamma, \delta$. Initialize θ_0, λ_0.
for $t = 0$ to T **do**
  Estimate $v_r^{\theta_t}, v_c^{\theta_t}$ using Algorithm 3 in (Wang & Zou, 2022)
  for $j = 1$ to M **do**
    Sample $T_j \sim$ *Geom*$(1 - \gamma + \gamma\delta)$, $s_1^j \sim \rho$
    Sample trajectory from s_1^j: $(s_1^j, a_1^j, \cdots, s_{T_j}^j)$
    for agent $i = 1$ to N **do**
      for $k = 1$ to W **do**
        $\tilde{g}_{t,k+1}^j(i) = \tilde{g}_{t,k}^j(i) - \zeta \nabla_{\tilde{g}} \mathcal{L}(\tilde{g}_{t,k}^j(i), \theta_t)$, $\mathcal{L}$ is defined in (5)
      end for
      $\tilde{g}_{t,k}^j = \sum_{i=1}^N \tilde{g}_{t,k}^j(i)/N$
    end for
  end for
  $\tilde{g}_t = \sum_{j=1}^M \sum_{k=1}^W \tilde{g}_{t,k}^j / MW$
  $\theta_{t+1} = \theta_t + \alpha_t(\tilde{g}_{r,t} + \lambda_t \tilde{g}_{c,t})$
  $\lambda_{t+1} = \max\{\lambda_t - \beta_t(\sum_j v_c^{\theta_t}(s_1^j)/M - d), 0\}$
end for
Output θ_T

Theorem 3.1 (Robust Natural Policy Gradient)**.** *Let $\tilde{g}^*$ minimizes the objective $J(\tilde{g}, \pi_\theta)$ defined as follows:*

$$\sum_{s,a} d_{\gamma,\delta,s_1}^{\pi} \pi(a|s)[\tilde{g}^\top \psi^\pi(s,a) - \phi^\pi(\tau) - b^\pi]^2, \tag{3}$$

where $d_{\gamma,\delta,s_1}^{\pi} \propto \sum_k \gamma^k (1-\delta)^k p^\pi(s_k = s|s_1)$ is the discounted visitation distribution of $s_k = s$ when the initial state is s_1 and policy π is used; $\psi^\pi(s,a)$ denotes $\nabla \log \pi(a|s,\theta)$; τ denotes a trajectory (s,a,r,c,s'); $\phi^\pi(\tau) = r + \gamma\delta \min_s v^\pi(s) + \gamma(1-\delta)v^\pi(s') - v^\pi(s)$ is the TD residual; $b^\pi = \gamma\delta/(1-\gamma+\gamma\delta)\partial_\theta \min_s v^\pi(s)$.

Then $\tilde{g}^* = F(\theta)^{-1}\nabla_\theta v^\pi(s_1)$ being the robust natural policy gradient of the objective function $v^\pi(s_1)$. For notational convenience, we omit the subscripts r and c in the value functions when there is no confusion.

Proof. Considering we have denoted $\psi^\pi(s,a) = \nabla \log \pi(a|s,\theta)$, Fisher information matrix is then given by $F(\theta) = \sum_{s,a} d_{\gamma,\delta,s_1}^{\pi}(s)\pi(a|s)\psi^\pi(s,a)\psi^\pi(s,a)^\top$. The robust policy gradient of the value function is given by $\nabla_\theta v^\pi(s_1) = \sum_{s,a} d_{\gamma,\delta,s_1}^{\pi}(s)\nabla_\theta \pi(a|s)\phi^\pi(\tau) + b^\pi \propto \mathbb{E}_{\pi,s_1}[\phi^\pi(\tau)\nabla \log \pi(a|s) + b^\pi]$ (Wang & Zou, 2022).

Since $\tilde{g}^*$ minimizes (3), it satisfies the condition $\partial J/\partial \tilde{g}_i = 0$, which implies: $\sum_{s,a} d_{\gamma,\delta,s_1}^{\pi} \pi(a|s) \times \psi^\pi(s,a)[\psi^\pi(s,a)^\top \tilde{g}^* - \phi^\pi(\tau) - b^\pi] = 0$. Then we have

$$\begin{aligned} &\sum_{s,a} d_{\gamma,\delta,s_1}^{\pi} \pi(a|s)\psi^\pi(s,a)\psi^\pi(s,a)^\top \tilde{g}^* \\ =&\sum_{s,a} d_{\gamma,\delta,s_1}^{\pi} \pi(a|s)\psi^\pi(s,a)[\phi^\pi(\tau) + b^\pi]. \end{aligned} \tag{4}$$

By the definition of Fisher information: LHS $= F(\theta)\tilde{g}^*$ and RHS $= \nabla_\theta v^\pi(s_1)$, which lead to: $F(\theta)\tilde{g}^* = \nabla_\theta v^\pi(s_1)$. Solving for $\tilde{g}^*$ gives $\tilde{g}^* = F(\theta)^{-1}\nabla_\theta v^\pi(s_1)$ which follows from the definition of the NPG on the worst-case value function of robust and constrained MARL. We name it a robust natural policy gradient in robust and constrained MARL. □

Considering the vanilla policy gradient may suffer from overshooting or undershooting and high variance, which results in slow convergence (Liu et al., 2020b), our proposed robust natural policy

gradient (RNPG) method updates the policy along the steepest ascent direction in the policy space in robust and constrained MARL (Ding et al., 2020).

Corollary 3.1 (Calculating RNPG by SGD)**.** *As shown in Theorem 3.1, we can get the RNPG of* $v^{\pi}(s_1)$ *by minimizing the objective defined in* (3)*. To minimize* (3) *and get the minimizer, we initialize* $\tilde{g}_0 = 0$ *and use the following stochastic gradient descent (SGD) steps:*

$$\tilde{g}_{k+1} = \tilde{g}_k - \zeta \nabla_{\tilde{g}} \mathcal{L}(\tilde{g}_k, \pi),$$

where ζ *is the learning rate and* $\mathcal{L}$ *is defined as follows:*

$$\mathcal{L}(\tilde{g}, \pi) = \sum_{\mathcal{D}(s_1)} [\tilde{g}^{\top} \psi^{\pi}(s, a) - \phi^{\pi}(\tau) - b^{\pi}]^2 / D, \tag{5}$$

where $\mathcal{D}(s_1)$ *is a set of trajectories* τ *starting at* s_1 *using policy* π*, i.e.* (s_1, a, r, c, s')*,* $D = |\mathcal{D}(s_1)|$*. After* W *steps of SGD iterations, the robust natural policy gradient for* $v^{\pi}(s_1)$ *is estimated as* $\sum_{k=1}^{W} \tilde{g}_k / W$*.*

4 Experiment

4.1 Experiment Setup

Three different data sets (He et al., 2020; 2022) including E-taxi GPS data, transaction data, and charging station data are used to build an EV AMoD system simulator as the training and testing environment. We modify the parameters of the simulator model such that the testing environment is different from the training environment, e.g., the parameters of the order generator. The simulated map is set as a grid city. The policy networks and critic networks are two-layer fully-connected networks, both with 32 nodes. We use Softplus as activations to ensure the output is positive. The output of policy networks is used to be the concentration parameters of the Dirichlet distribution to satisfy the action constraints (sum to one). We set the maximal training episode number $= 20000$, the maximal policy/critic estimation number $= 2000$, the NRPG SDG iteration number $= 500$, the discount rate $\gamma = 0.99$, the perturbed rate $\delta = 0.05$, the coefficients $\bar{\alpha} = \bar{\beta} = 1$, the fairness constraint limit $d = -20$ for one simulation step, and use AdamOptimizer with a learning rate of 0.001 for both policy/critic networks.

4.2 Experiment Results

Our goal of the experiments is to validate the following hypothesis: (1) The proposed ROCOMA can learn effective rebalancing policies; (2) Our proposed ROCOMA learns more robust policies than a non-robust MARL algorithm by considering state transition uncertainties and constraints in the MARL problem formulation and the proposed RNPG method for policy training. We compare metrics:*Rebalancing cost:* the total moving distance of vacant and low-battery EVs by using a rebalancing policy (the lower the better); and *System fairness:* the weighted sum of mobility and charging fairness (the higher the better); we also monitor *Number of expired orders:* the total number of canceled orders due to waiting for more than 20 minutes (the lower the better) and *Order response rate:* the ratio between the number of served demands and the number of total passenger demand (the higher the better). All metrics are calculated in every testing period which consists of 25 simulation steps. Then the fairness constraint limit for one testing period is -500. We repeat testing for 10 times and show the average values.

Table 1: Comparison: rocoma VS other rebalancing methods

	rebalancing cost	system fairness	expired order	response rate
ROCOMA	2.06×10^5	-292.14	1.20×10^2	99.82%
COP	1.88×10^5	-383.19	1.61×10^3	93.05%
EDP	2.15×10^5	-409.49	6.90×10^1	99.69%
RDP	2.43×10^5	-629.85	3.68×10^3	84.34%
NO	-	-4317.53	7.64×10^3	66.89%

Compared to no rebalancing, by using our method, the expired orders number is decreased by 98.4%, the system fairness and order response rate are increased by about 93.2% and 32.9%, respectively.

Table 2: Comparison: rocoma VS non-constrained marl method

	rebalancing cost	system fairness	expired order	response rate
ROCOMA	2.06×10^5	-292.14	120	99.82%
Non-constrain	1.98×10^5	-1812.48	1607	93.06%

Our method achieves 83.9% higher in fairness compared to the non-constrained MARL method with 4% extra rebalancing cost.

ROCOMA is effective: In Table 1, we compare ROCOMA with *no rebalancing scenario (NO)* and the following rebalancing algorithms: *(1) Constrained optimization policy (COP):* The optimization goal is to minimize the rebalancing cost under the fairness constraints (He et al., 2023). The fairness limit is the same as that used in ROCOMA. The dynamic models are calculated from the same data sets used in simulator construction. *(2) Equally distributed policy (EDP):* EVs are assigned to their current and adjacent regions using equal probability (20%). *(3) Randomly distributed policy (RDP):* EVs are randomly distributed to their current and adjacent regions.

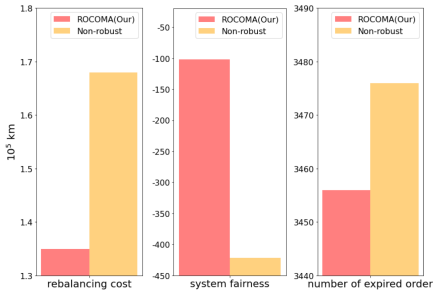

Figure 3: Comparison of ROCOMA and Non-robust MARL method: Compared to the non-robust method, ROCOMA decreases the rebalancing cost and increases the system fairness by 19.6% and 75.8%, respectively, when model uncertainties are present.

In Table 1, compared to the no rebalancing scenario, ROCOMA is effective in rebalancing AMoD systems in terms of fairness, expired orders and response rate. Specifically, ROCOMA policy decreases the number of expired orders by 98.4%, increases the system fairness and order response rate by about 93.2% and 32.9%, respectively. Besides, ROCOMA achieves a higher system fairness and order response rate using less rebalancing cost than EDP and RDP. Though ROCOMA takes more rebalancing costs than COP, it has a better system fairness and order response rate. It is within expectation since the constrained optimization method is a centralized method that aims to optimize the rebalancing cost and it does not consider any uncertainties.

ROCOMA is robust: In Figure 3 and Table 2, we compare ROCOMA with *(1) Non-constrain MARL algorithm:* Instead of considering fairness constraints in MARL, the reward is designed as a weighted sum of negative rebalancing cost and system fairness. The coefficient is 1. And model uncertainty is considered; *(2) Non-robust MARL algorithm:* The model uncertainty is not considered but the fairness constraint is considered in MARL. They use the same network structures and other hyper-parameters as that in ROCOMA.

In Figure 3, we test well-trained robust and non-robust methods in a testing environment (different from the training environment) to show the robustness of the ROCOMA policy. We can see ROCOMA policy achieves better performance in terms of all metrics. Specifically, ROCOMA decreases the rebalancing cost and increases the system fairness by about 19.6% and 75.8% , respectively, when model uncertainty exists, compared to the non-robust method.

In Table 2, ROCOMA achieves 83.9% higher in fairness compared to the non-constrained MARL algorithm with just 4% extra rebalancing cost. Without the fairness constraint design, the non-constrained MARL method falls into a pit that sacrifices fairness to achieve a lower rebalancing cost since its objective is a weighted sum of them. It would take a lot of effort to tune the hyper-parameter to find a policy that performs well in both rebalancing cost and fairness. The constrained MARL design of ROCOMA avoids such extra tuning efforts.

5 Conclusion

It remains challenging to address AMoD system model uncertainties caused by EVs' unique charging patterns and AMoD systems' mobility dynamics in algorithm design. In this work, we design a robust and constrained multi-agent reinforcement learning framework to balance the mobility supply-demand ratio and the charging utilization rate, and minimize the rebalancing cost for EV AMoD systems under state transition uncertainties. We then design a robust and constrained MARL algorithm (ROCOMA) to train robust policies. Experiments show that our proposed robust algorithm can learn effective and robust rebalancing policies.

REFERENCES

Eitan Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

J Andrew Bagnell, Andrew Y Ng, and Jeff G Schneider. Solving uncertain markov decision processes. 2001.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, USA, 2004. ISBN 0521833787.

Eduardo F Camacho and Carlos Bordons Alba. *Model predictive control*. Springer science & business media, 2013.

Ximing Chen, Fei Miao, George Pappas, and Victor Preciado. Hierarchical data-driven vehicle dispatch and ride-sharing. In *Proceedings of the IEEE 56th Conference on Decision and Control*, CDC'17, pp. 4458–4463, 2017.

Ziheng Chen, Fabrizio Silvestri, Jia Wang, He Zhu, Hongshik Ahn, and Gabriele Tolomei. Relax: Reinforcement learning agent explainer for arbitrary predictive models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 252–261, 2022.

Yiming Cui, Zhiwen Cao, Yixin Xie, Xingyu Jiang, Feng Tao, Yingjie Victor Chen, Lin Li, and Dongfang Liu. Dg-labeler and dgl-mots dataset: Boost the autonomous driving perception. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 58–67, 2022.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

Zihan Ding and Hao Dong. Challenges of reinforcement learning. In *Deep Reinforcement Learning*, pp. 249–272. Springer, 2020.

Maxime Guériau and Ivana Dusparic. Samod: Shared autonomous mobility-on-demand using decentralized reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1558–1563. IEEE, 2018.

Songyang Han, Sanbao Su, Sihong He, Shuo Han, Haizhao Yang, and Fei Miao. What is the solution for state adversarial multi-agent reinforcement learning? *arXiv preprint arXiv:2212.02705*, 2022.

Zhaowei Hao, Long He, Zhenyu Hu, and Jun Jiang. Robust vehicle pre-allocation with uncertain covariates. *Production and Operations Management*, 29(4):955–972, 2020.

Sihong He, Lynn Pepin, Guang Wang, Desheng Zhang, and Fei Miao. Data-driven distributionally robust electric vehicle balancing for mobility-on-demand systems under demand and supply uncertainties. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2165–2172. IEEE, 2020.

Sihong He, Yue Wang, Shuo Han, Shaofeng Zou, and Fei Miao. A robust and constrained multi-agent reinforcement learning framework for electric vehicle amod systems. *arXiv preprint arXiv:2209.08230*, 2022.

Sihong He, Zhili Zhang, Shuo Han, Lynn Pepin, Guang Wang, Desheng Zhang, John A Stankovic, and Fei Miao. Data-driven distributionally robust electric vehicle balancing for autonomous mobility-on-demand systems under demand and supply uncertainties. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

Suining He and Kang G Shin. Spatio-temporal capsule-based reinforcement learning for mobility-on-demand coordination. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

John Holler, Risto Vuorio, Zhiwei Qin, Xiaocheng Tang, Yan Jiao, Tiancheng Jin, Satinder Singh, Chenxi Wang, and Jieping Ye. Deep reinforcement learning for multi-driver vehicle dispatching and repositioning problem. In *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 1090–1095. IEEE, 2019.

Bin Huang and Jianhui Wang. Deep-reinforcement-learning-based capacity scheduling for pv-battery storage system. *IEEE Transactions on Smart Grid*, 12(3):2272–2283, 2020.

Bin Huang and Jianhui Wang. Applications of physics-informed neural networks in power systems-a review. *IEEE Transactions on Power Systems*, 2022.

IEA. Iea (2020), global ev outlook 2020, iea, paris, 2020.

Ramon Iglesias, Federico Rossi, Rick Zhang, and Marco Pavone. A bcmp network approach to modeling and controlling autonomous mobility-on-demand systems. *The International Journal of Robotics Research*, 38(2-3):357–374, 2019. doi: 10.1177/0278364918780335. URL https://doi.org/10.1177/0278364918780335.

Ramón Iglesias, Federico Rossi, Kevin Wang, David Hallac, Jure Leskovec, and Marco Pavone. Data-driven model predictive control of autonomous mobility-on-demand systems. In *IEEE International Conference on Robotics and Automation*, volume abs/1709.07032, 2018.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. Efficient large-scale fleet management via multi-agent deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining*, KDD '18. Association for Computing Machinery, 2018a. ISBN 9781450355520. doi: 10.1145/3219819.3219993. URL https://doi.org/10.1145/3219819.3219993.

Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. Efficient large-scale fleet management via multi-agent deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1774–1783, 2018b.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Dongfang Liu, Yiming Cui, Yingjie Chen, Jiyong Zhang, and Bin Fan. Video object detection for autonomous driving: Motion-aid feature calibration. *Neurocomputing*, 409:1–11, 2020a.

Dongfang Liu, Yiming Cui, Xiaolei Guo, Wei Ding, Baijian Yang, and Yingjie Chen. Visual localization for autonomous driving: Mapping the accurate location in the city maze. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3170–3177. IEEE, 2021a.

Wei Liu, Xin Xia, Lu Xiong, Yishi Lu, Letian Gao, and Zhuoping Yu. Automated vehicle sideslip angle estimation considering signal measurement characteristic. *IEEE Sensors Journal*, 21(19): 21675–21687, 2021b.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020b.

Zhiguang Liu, Tomio Miwa, Weiliang Zeng, Michael GH Bell, and Takayuki Morikawa. Dynamic shared autonomous taxi system considering on-time arrival reliability. *Transportation Research Part C: Emerging Technologies*, 103:281–297, 2019.

Ryan Lowe and Yi I Wu. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*, pp. 6379–6390, 2017.

Xiaoling Luo, Xiaobo Ma, Matthew Munden, Yao-Jan Wu, and Yangsheng Jiang. A multisource data approach for estimating vehicle queue length at metered on-ramps. *Journal of Transportation Engineering, Part A: Systems*, 148(2):04021117, 2022.

Xiaobo Ma. *Traffic Performance Evaluation Using Statistical and Machine Learning Methods*. PhD thesis, The University of Arizona, 2022.

Xiaobo Ma, Abolfazl Karimpour, and Yao-Jan Wu. Statistical evaluation of data requirement for ramp metering performance assessment. *Transportation Research Part A: Policy and Practice*, 141:248–261, 2020.

F. Miao, S. Han, S. Lin, Q. Wang, J. A. Stankovic, A. Hendawi, D. Zhang, T. He, and G. J. Pappas. Data-driven robust taxi dispatch under demand uncertainties. *IEEE Transactions on Control Systems Technology*, 27(1):175–191, Jan 2019. ISSN 1063-6536. doi: 10.1109/TCST.2017.2766042.

Fei Miao, Sihong He, Lynn Pepin, Shuo Han, Abdeltawab Hendawi, Mohamed E Khalefa, John A Stankovic, and George Pappas. Data-driven distributionally robust optimization for vehicle balancing of mobility-on-demand systems. *ACM Transactions on Cyber-Physical Systems*, 2021.

J. Miller and J. P. How. Predictive positioning and quality of service ridesharing for campus mobility on demand systems. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1402–1408, May 2017. doi: 10.1109/ICRA.2017.7989167.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Abood Mourad, Jakob Puchinger, and Chengbin Chu. A survey of models and algorithms for optimizing shared mobility. *Transportation Research Part B: Methodological*, 123:323 – 346, 2019. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2019.02.003.

Arnab Nilim and Laurent Ghaoui. Robustness in markov decision problems with uncertain transition matrices. *Advances in neural information processing systems*, 16, 2003.

B. P. G. Van Parys, D. Kuhn, P. J. Goulart, and M. Morari. Distributionally robust control of constrained stochastic systems. *IEEE Transactions on Automatic Control*, 61(2):430–442, Feb 2016. ISSN 0018-9286.

J. Pfrommer, J. Warrington, G. Schildbach, and M. Morari. Dynamic vehicle redistribution and online price incentives in shared mobility systems. *IEEE Transactions on Intelligent Transportation Systems*, 15(4):1567–1578, Aug 2014. ISSN 1524-9050. doi: 10.1109/TITS.2014.2303986.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017.

Adarsh Prasad, Vishwak Srinivasan, Sivaraman Balakrishnan, and Pradeep Ravikumar. On learning ising models under huber's contamination model. *Advances in neural information processing systems*, 33:16327–16338, 2020.

Elvezio M Ronchetti and Peter J Huber. *Robust statistics*. John Wiley & Sons, 2009.

N. Sadeghianpourhamami, J. Deleu, and C. Develder. Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning. *IEEE Transactions on Smart Grid*, 11(1):203–214, 2020. doi: 10.1109/TSG.2019.2920320.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Sanbao Su, Yiming Li, Sihong He, Songyang Han, Chen Feng, Caiwen Ding, and Fei Miao. Uncertainty quantification of collaborative detection for self-driving. *arXiv preprint arXiv:2209.08162*, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Berkay Turan, Ramtin Pedarsani, and Mahnoosh Alizadeh. Dynamic pricing and fleet management for electric autonomous mobility on demand systems. *Transportation Research Part C: Emerging Technologies*, 121:102829, 2020. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2020.102829.

A. Wallar, M. Van Der Zee, J. Alonso-Mora, and D. Rus. Vehicle rebalancing for mobility-on-demand systems with ride-sharing. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4539–4546, Oct 2018. doi: 10.1109/IROS.2018.8593743.

Z. Wan, H. Li, H. He, and D. Prokhorov. Model-free real-time ev charging scheduling based on deep reinforcement learning. *IEEE Transactions on Smart Grid*, 10(5):5246–5257, 2019. doi: 10.1109/TSG.2018.2879572.

Guang Wang, Shuxin Zhong, Shuai Wang, Fei Miao, Zheng Dong, and Desheng Zhang. Data-driven fairness-aware vehicle displacement for large-scale electric taxi fleets. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 1200–1211. IEEE, 2021.

Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. *arXiv preprint arXiv:2205.07344*, 2022.

J. Wen, J. Zhao, and P. Jaillet. Rebalancing shared mobility-on-demand systems: A reinforcement learning approach. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 220–225, 2017. doi: 10.1109/ITSC.2017.8317908.

Jian Wen, Jinhua Zhao, and Patrick Jaillet. Rebalancing shared mobility-on-demand systems: A reinforcement learning approach. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*, pp. 220–225. Ieee, 2017.

Dongmei Wu, Yuying Guan, Xin Xia, Changqing Du, Fuwu Yan, Yang Li, Min Hua, and Wei Liu. Coordinated control of path tracking and yaw stability for distributed drive electric vehicle based on ampc and dyc. *arXiv preprint arXiv:2304.11796*, 2023.

Yukun Yuan, Desheng Zhang, Fei Miao, Jiming Chen, Tian He, and Shan Lin. p2charging proactive partial charging for electric taxi systems. In *IEEE International Conference on Distributed Computing Systems*, ICDCS'19, 2019.

Gioele Zardini, Nicolas Lanzetti, Marco Pavone, and Emilio Frazzoli. Analysis and control of autonomous mobility-on-demand systems: A review. *arXiv preprint arXiv:2106.14827*, 2021.

Dan Zhang, Fangfang Zhou, Yuwen Jiang, and Zhengming Fu. Mm-bsn: Self-supervised image denoising for real-world with multi-mask based on blind-spot network. *arXiv preprint arXiv:2304.01598*, 2023.

Rick Zhang, Federico Rossi, and Marco Pavone. Model predictive control of autonomous mobility-on-demand systems. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1382–1389. IEEE, 2016.

A Related Work

As the technologies on autonomous vehicles (Cui et al., 2022; Liu et al., 2020a; 2021b) are getting mature, they are becoming essential parts of a transportation system. They can be used to provide transportation services, such as taxi or shuttle services, to passengers in a shared and on-demand manner (Wu et al., 2023). However, these autonomous vehicles need to be rebalanced due to the unbalanced supply and demand distributions in AMoD systems (He et al., 2023). AMoD system vehicle rebalancing algorithms re-allocate vacant vehicles, sometimes considering charging constraints. Heuristics can lead to sub-optimal rebalancing solutions (Liu et al., 2019). Other major categories of AMoD system rebalancing methods include optimization-based algorithms (Miao et al., 2021), Model Predictive Control (MPC) (Camacho & Alba, 2013) and Reinforcement Learning (RL) (Sutton & Barto, 2018; Chen et al., 2022).

Optimization and MPC-based approaches usually formulate the AMoD system vehicle rebalancing problem as an optimization problem, where the objective is to improve service quality (Miller & How, 2017; Pfrommer et al., 2014) or maximize the number of served passengers with fewer vehicles (Zhang et al., 2016; Wallar et al., 2018; Iglesias et al., 2018). These model-based approaches usually rely on knowledge of the probability transition model of the complex dynamics of AMoD systems. Though robust and distributionally robust optimization-based methods have been designed to consider uncertainties caused by mobility demand, supply, or covariates predictions (He et al., 2020; Hao et al., 2020; He et al., 2023), the probability transition error or uncertainty in system dynamics has not been addressed yet. *Various RL-based methods* include DQN, A2C and their variants (Mnih et al., 2015; Konda & Tsitsiklis, 1999; Wen et al., 2017; Guériau & Dusparic, 2018; Holler et al., 2019; Lin et al., 2018b; He & Shin, 2020; Wang et al., 2021) have been proposed to solve the vehicle rebalancing problem. However, RL suffers from the sim-to-real gap; that is, the gap between the simulator and the real world often leads to unsuccessful implementation if the learned policy is not robust to model uncertainties (Ding & Dong, 2020; Pinto et al., 2017). None of the above RL-based rebalancing strategies consider this gap.

As Machine Learning methods have been proposed to advance Smart City (Huang & Wang, 2022; Ma, 2022; Liu et al., 2021a), Reinforcement Learning (RL)-based methods are getting a lot of attention (Huang & Wang, 2020). However, uncertainties caused by sensor errors, noise, malicious attacks, and inaccurate predictions can undermine these RL-based methods (Luo et al., 2022; Ma et al., 2020; Su et al., 2022; Zhang et al., 2023). Therefore, Robust RL has been proposed to find a policy that maximizes the worst-case cumulative reward over an uncertainty set of MDPs (Bagnell et al., 2001; Pinto et al., 2017; Nilim & Ghaoui, 2003; Han et al., 2022). To achieve a desired level of system fairness while minimizing rebalancing cost under model uncertainty, we put the fairness constraints in our RL formulation, which is known as *Constrained RL* that aims to find a policy that maximizes an objective function while satisfying certain cost constraints (Altman, 1999; Wang & Zou, 2022). However, it remains challenging to design a robust EV rebalancing algorithm under model uncertainties and policy constraints, since the problem of robust constrained RL itself is already difficult to solve even in the simple tabular case. A robust and constrained RL for AMoD rebalancing cannot directly apply existing robust constrained RL solutions due to the high-dimensional state and action spaces commonly present in transportation systems. Our proposed robust and constrained MARL formulation and algorithm explicitly consider model uncertainties and policy constraints to learn robust rebalancing solutions for AMoD systems. And we derive a robust natural policy gradient for robust and constrained MARL to improve the efficiency of policy training.