Congestion-aware Routing and Rebalancing of Autonomous Mobility-on-Demand Systems in Mixed Traffic

Salomón Wollenstein-Betech[1], Arian Houshmand[1], Mauro Salazar[2,3], Marco Pavone[2], Christos G. Cassandras[1], and Ioannis Ch. Paschalidis[1]

***Abstract*— This paper studies congestion-aware route-planning policies for Autonomous Mobility-on-Demand (AMoD) systems, whereby a fleet of autonomous vehicles provides on-demand mobility under mixed traffic conditions. Specifically, we first devise a network flow model to optimize the AMoD routing and rebalancing strategies in a congestion-aware fashion by accounting for the endogenous impact of AMoD flows on travel time. Second, we capture reactive exogenous traffic consisting of private vehicles selfishly adapting to the AMoD flows in a user-centric fashion by leveraging an iterative approach. Finally, we showcase the effectiveness of our framework with a case-study considering the transportation sub-network in New York City. Our results suggest that for high levels of demand, pure AMoD travel can be detrimental due to the additional traffic stemming from its rebalancing flows, whilst the combination of AMoD with walking or micromobility options can significantly improve the overall system performance.**

I. Introduction

In the past decade, the rapid adoption of smartphone technologies and wireless communications coupled with the emergence of sharing economies has resulted in a widespread use of Mobility-on-Demand (MoD) services. One of the main operational challenges that these services face is deciding the routing and rebalancing policies for their vehicles. Currently, MoD systems use *user-centric* routing services (e.g., Waze and Google Maps) to route their vehicles, and dynamic pricing combined with a real-time heat-map of the users' demand to rebalance their fleets.

Given this *user-centric* approach to route vehicles in which every driver acts selfishly to minimize their own travel time, the network reaches an equilibrium known as the *Wardrop* equilibrium [1]. Unfortunately, these equilibria are in general suboptimal compared to the system optimum, achievable when the vehicles are coordinated by a central controller in a *system-centric* fashion.

Recently, the combination of MoD services with Connected and Automated Vehicles (CAVs) resulting into Autonomous Mobility-on-Demand (AMoD) systems (see Fig. 1) has attracted the interest of academia and industry. These fleets of CAVs providing on-demand mobility are expected to reduce labor costs, accidents, harmful emissions [2], and increase the efficiency of the fleets' operation as they can be *centrally* controlled [3]. Considering high penetration rates of AMoD in the mobility ecosystem, the routing and rebalancing policies designed to centrally control the vehicles will affect the congestion levels and, in turn, the routing decisions of privately owned vehicles. In this context, this paper studies system-optimal routing and rebalancing strategies for AMoD systems in mixed-traffic conditions.

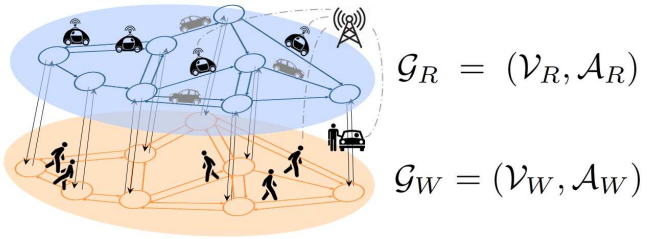


Fig. 1: AMoD network (supergraph) consisting of two digraphs for the road (blue) and the walking (orange) network; the black arrows represent switching arcs. AMoD vehicles are in black and private vehicles in grey.

Related literature: AMoD systems and rebalancing policies have been extensively studied using simulation models [4]–[6], queuing-theoretical models [7], [8], and network-flow models [9], [10]. In [4], the rebalancing of an AMoD system is addressed using a data-driven real-time parametric controller. Conversely, in [9], the rebalancing problem is studied using a steady-state fluid model. Although [4] and [9] seek to find effective rebalancing policies, they do not consider the impact of the AMoD routes on congestion, but rather assume travel times on the road links to be constant.

Little work has been done to solve the congestion-aware routing and rebalancing problem jointly. Most approaches leverage approximations of the travel time function relating traffic density to travel times to address the non-convex nature of the problem. The authors of [10] use a threshold model to show that under relatively mild assumptions rebalancing vehicles does not lead to an increase in congestion, suggesting that the joint problem can be decoupled without a substantial impact on the solution's quality. Furthermore, a piecewise-affine approximation of the travel time function was introduced in [11] in order to relax the problem to a quadratic program. Yet, depending on the congestion levels, both approaches may lack in accuracy. Moreover, [10] and [11] assume a static exogenous traffic flow that does not change for varying AMoD routes. Finally, reactive private traffic was modeled in [12] to show that under a system-centric optimal-routing strategy both CAVs and non-CAVs can achieve better performance in terms of travel time and

*This work was supported by NSF under grants ECCS-1509084, DMS-1664644, CNS-1645681, IIS-1914792, and CMMI-1454737, by AFOSR under grant FA9550-19-1-0158, by ARPA-E's NEXTCAR grant DEAR0000796, by the MathWorks, by the ONR grant N00014-19-1-2571, by the NIH grant 1R01GM135930, and by the Toyota Research Institute (TRI). This article solely reflects the opinions and conclusions of its authors and not NSF, TRI, or any other entity. We thank D. Sverdlin-Lisker, Dr. I. New and Dr. K. Solovey for proofreading this paper.

[1] Division of Systems Engineering, Boston University, USA {salomonw, arianhm, cgc, yannisp}@bu.edu

[2] Department of Aeronautics and Astronautics, Stanford University, USA pavone@stanford.edu

[3] Department of Mechanical Engineering, Eindhoven University of Technology, The Netherlands m.r.u.salazar@tue.nl

energy savings. However, such an approach neither captures rebalancing effects nor intermodal routing possibilities.

Statement of contribution: This paper bridges the gap between [11] and [12]. In particular, we study how system-optimal routing of AMoD services can affect the system-level performance in *mixed-traffic* (presence of AMoD and private vehicles in the road network). Similar to [11], we assume that AMoD users can use multiple modes of transportation, i.e., autonomous taxi rides and walking. In addition, we assume the private vehicle flow to be *reactive*, meaning that private vehicles will choose their routes selfishly considering the congestion stemming from the AMoD flow. To this end, we use the framework developed in [12] for modeling the interaction between AMoD and private vehicles. Moreover, we devise an approximation of the travel time function that is more accurate than the one proposed in [11], whilst still maintaining the quadratic convex structure of the problem. The proposed model can efficiently compute congestion-aware routing and rebalancing strategies for a given demand and road network topology. Finally, with this framework at hand, we analyze the trade-offs between the benefits of system-centric routing and the cost of rebalancing, and investigate the achievable benefits stemming from the combination of AMoD with walking and micromobility options.

Organization: The rest of the paper is organized as follows: In Section II we provide preliminaries of the model and its formulation. In Section III we develop a convex approximation of the original problem to overcome its non-convex nature. Then, we present experiments using a New York City case-study in Section IV. Finally, in Section V we conclude the paper and point to future research directions.

Notation: All vectors are column vectors and denoted by bold lowercase letters. We use "prime" to denote transpose, and use $\mathbb{1}$ to denote the indicator function.

II. Problem Formulation

In this section, we present macroscopic models for planning the routing and rebalancing strategies used throughout the paper. First, we introduce the notation and preliminaries of transportation modeling. With this at hand, we model the system-centric routing and rebalancing of AMoD, followed by the user-centric model for the private vehicles. Finally, we formulate the joint problem of congestion-aware routing and rebalancing of AMoD in mixed traffic.

A. Preliminaries

Consider an AMoD system which provides mobility services through two modes of transportation: walking and autonomous taxi-rides. To model the system, let $\mathcal{G}$ be a network (supergraph) composed of two layers, a road and a walking network. We denote by $\mathcal{G}_{\mathrm{R}} = (\mathcal{V}_{\mathrm{R}}, \mathcal{A}_{\mathrm{R}})$ the road network and by $\mathcal{G}_{\mathrm{W}} = (\mathcal{V}_{\mathrm{W}}, \mathcal{A}_{\mathrm{W}})$ the pedestrian graph where $(\mathcal{V}_{\mathrm{R}}, \mathcal{A}_{\mathrm{R}})$ and $(\mathcal{V}_{\mathrm{W}}, \mathcal{A}_{\mathrm{W}})$ are the sets of intersections (vertices) and streets (arcs) in the road and in the pedestrian network, respectively. Then, the supergraph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ is composed of $\mathcal{G}_{\mathrm{R}}$ and $\mathcal{G}_{\mathrm{W}}$, and a set of *switching* arcs $\mathcal{A}_{\mathrm{S}} \subset \mathcal{V}_{\mathrm{R}} \times \mathcal{V}_{\mathrm{W}} \cup \mathcal{V}_{\mathrm{W}} \times \mathcal{V}_{\mathrm{R}}$ that connect the pedestrian and the road network layers to allow AMoD users to change modes (see Fig. 1). Formally $\mathcal{G}$ is composed of the set of vertices $\mathcal{V} = \mathcal{V}_{\mathrm{R}} \cup \mathcal{V}_{\mathrm{W}}$ and arcs $\mathcal{A} = \mathcal{A}_{\mathrm{R}} \cup \mathcal{A}_{\mathrm{W}} \cup \mathcal{A}_{\mathrm{S}}$.

In order to model the demanded trips, let $\mathbf{w} = (w_s, w_t)$ denote an Origin-Destination (OD) pair and $d_{\mathbf{w}} \geq 0$ the demand rate at which customers request service per unit time from origin w_s to destination w_t. Let W be the total number of OD pairs and $\mathcal{W} = \{\mathbf{w}_k : \mathbf{w}_k = (w_{sk}, w_{tk}), k = \{1, ..., W\}\}$ the set of OD pairs. Let a vectorized version of the demand be $\mathbf{g} = (d^{\mathbf{w}} : \mathbf{w} \in \mathcal{W})$, which denotes the demand flows for all OD pairs.

To keep track of AMoD users' flow on an arc, we let $x_{ij}^{\mathbf{w}}$ denote the AMoD flow induced by OD pair $\mathbf{w}$ in link $(i,j) \in \mathcal{A}$. Given that the AMoD needs to rebalance its vehicles to ensure service, we let x_{ij}^r be the *rebalancing flow* of empty vehicles on road (i,j). Finally, to consider the interaction between the AMoD provider and the other vehicles, we let x_{ij}^p be the self-interested *private vehicle* flow on (i,j). We use the term *private* as we assume that self-interested users must arrive at their destination with their vehicle and do not have the option of switching transportation mode (i.e., walking). To simplify notation, we let the AMoD user flow on any edge (road, walking, or switching) to be $x_{ij}^u = \sum_{\mathbf{w} \in \mathcal{W}} x_{ij}^{\mathbf{w}}$, $\forall (i,j) \in \mathcal{A}$, and the total flow on a link to be

$$x_{ij} = x_{ij}^u + x_{ij}^r + x_{ij}^p, \qquad \forall (i,j) \in \mathcal{A}. \tag{1}$$

Note that neither rebalancing nor the private vehicle flow should exist on the switching or walking arcs. Hence, for those arcs we set $x_{ij}^r = x_{ij}^p = 0$, $\forall (i,j) \in \mathcal{A}_{\mathrm{S}} \cup \mathcal{A}_{\mathrm{W}}$.

Let $t_{ij}(x) : \mathbb{R}_+^{|\mathcal{A}|} \mapsto \mathbb{R}_+$ be the *travel time* function, i.e., the time it takes to cross link (i,j) given the flow on that link. Using the same function structure as in [13], we characterize t_{ij} as a function of the flow x_{ij} with

$$t_{ij}(x_{ij}) = t_{ij}^0 f(x_{ij}/m_{ij}), \tag{2}$$

where m_{ij} is the road's capacity, $f(\cdot)$ is a strictly increasing, positive, and continuously differentiable function, and t_{ij}^0 is the free-flow travel time on link (i,j). We consider functions with $f(0) = 1$, which ensures that if there is no flow on the link, the travel time t_{ij} is equal to the free-flow travel time. Typically, travel time functions used by urban planners and researchers are polynomials which are hard to estimate [14]. A widely used function is the *Bureau of Public Roads (BPR)* travel time function [15] denoted by $t_{ij}(x_{ij}) = t_{ij}^0(1 + 0.15(x_{ij}/m_{ij})^4)$. Throughout this paper, we use this function to decide the routes of AMoD users and private vehicles, given the network flow levels. For AMoD users who walk, we consider a constant travel time (independent of the flow) on each link.

B. System-centric Routing and Rebalancing of AMoD

Recall that our goal is to find the system-centric congestion-aware routes and the rebalancing policy of an AMoD provider. The objective consists of minimizing the cost composed of the overall travel time of AMoD users, and a regularizer penalizing rebalancing flow.

We formulate the problem similar to [11] where we address it from an AMoD perspective. Let $d_{\mathbf{w}}^u$ be customer requests to the AMoD provider traveling from origin w_s to destination w_t, and let the total link flow be $\mathbf{x} = \{\mathbf{x}^{\mathbf{w}}\}_{\mathbf{w} \in \mathcal{W}} \cup \mathbf{x}^r$ where we use bold notation $\mathbf{x}$ to represent a vector

containing all the elements of x_{ij}. The problem we aim to solve is then expressed by

$$\min_{\mathbf{x}} \quad J(\mathbf{x}) := \sum_{(i,j)\in\mathcal{A}} t_{ij}(x_{ij})x^u_{ij} + \mathbf{c}'\mathbf{x}^r \tag{3a}$$

$$\text{s.t.} \sum_{i:(i,j)\in\mathcal{A}} x^{\mathbf{w}}_{ij} + \mathbb{1}_{j=w_s} d^u_{\mathbf{w}} = \sum_{k:(j,k)\in\mathcal{A}} x^{\mathbf{w}}_{jk} + \mathbb{1}_{j=w_t} d^u_{\mathbf{w}}, \quad \forall \mathbf{w}\in\mathcal{W}, j\in\mathcal{V}, \tag{3b}$$

$$\sum_{i:(i,j)\in\mathcal{A}_{\mathrm{R}}} \left(x^r_{ij} + x^u_{ij}\right) = \sum_{k:(j,k)\in\mathcal{A}_{\mathrm{R}}} \left(x^r_{jk} + x^u_{jk}\right), \quad \forall j\in\mathcal{V}_{\mathrm{R}}, \tag{3c}$$

$$\mathbf{x} \geq 0. \tag{3d}$$

The constraints (3b) take care of flow conservation and demand compliance as in a multi-commodity transportation problem (including flow conservation on the walking network), constraints (3c) ensure the rebalancing of the AMoD fleet (only on the road network), and (3d) restrict the flows to non-negative values. By solving (3) we find the optimal AMoD user and rebalancing flows. Note that the AMoD users' flow may consist of both walking or vehicle options.

The objective J is composed of two terms. The first term considers the total travel time of AMoD users. This term evaluates the travel time function $t_{ij}(x_{ij})$ with respect to the total flow (see (1)) which includes variables corresponding to private vehicle flow x^p_{ij} (assumed to be fixed), and the rebalancing flow x^r_{ij}. Hence, when taking the product of $t_{ij}(x_{ij})x^u_{ij}$ we obtain a non-convex function. To address the non-convexity issue, we will use a piecewise-affine approximation of $t_{ij}(x_{ij})$, further presented in Section III. The second term, $\mathbf{c}'\mathbf{x}^r$, acts as a linear reguralizer whose purpose is to penalize rebalancing flows. This will ensure that a cost for rebalancing of the fleet is taken into account. In this work, we use $\mathbf{c} = \lambda\mathbf{t}^0$. One can think of this reguralizer as a linear travel time function with respect to the rebalancing flow (since $(\lambda\mathbf{t}^0)'\mathbf{x}^r$). Therefore, if one lets λ be high with respect to the overall travel time, the reguralizer term will dominate the objective. Hence, we use a small λ in order to guide the rebalancing flow through good paths without dominating the AMoD user routing decisions.

C. Private Vehicle Flow Modeling

Aiming to understand the interaction between a system-centric AMoD fleet and self-interested private vehicles, we assume some rationale behind private vehicle decisions. To model this class of vehicles we use the *user-centric* approach as in the Traffic Assignment Problem (TAP) [16]. This model finds, given OD demands, the flows in the network which achieve a Wardrop equilibrium [1].

Given a demand $\mathbf{g}^p$ for this type of vehicle, each private user decides its route such that it minimizes its own travel time. Moreover, we impose that private vehicles can travel exclusively through the road network $\mathcal{G}_{\mathrm{R}}$. In other words, we do not allow private vehicles to change their transportation mode to walking.

Let $x^{p,\mathbf{w}}_{ij}$ be the flow on link (i,j) induced by private vehicle demand $d^p_{\mathbf{w}}$ of OD pair $\mathbf{w}$. Then, we assume private vehicles decide their routes by using the user-centric approach,

$$\min_{\mathbf{x}^p\geq 0} \quad \sum_{(i,j)\in\mathcal{A}_{\mathrm{R}}} \int_{x^u_{ij}+x^r_{ij}}^{x_{ij}} t_{ij}(s)ds \tag{4a}$$

$$\text{s.t.} \sum_{i:(i,j)\in\mathcal{A}_{\mathrm{R}}} x^{p,\mathbf{w}}_{ij} + d^p_{\mathbf{w}}\mathbb{1}_{j=w_s} = \sum_{k:(j,k)\in\mathcal{A}_{\mathrm{R}}} x^{p,\mathbf{w}}_{jk} + d^p_{\mathbf{w}}\mathbb{1}_{j=w_t}, \quad \forall \mathbf{w}\in\mathcal{W}, j\in\mathcal{V}_{\mathrm{R}}. \tag{4b}$$

Notice that this version of the user-centric TAP is slightly different from the classical [16], given that it considers the AMoD flow in its objective (see the integral's limits in (4a)).

To solve this problem we assume that the AMoD flow is fixed and private vehicles plan their routes considering AMoD flows as exogenous. When working with this restriction, we can use any efficient TAP algorithm (e.g. Frank-Wolfe) [16] to solve (4). Let us use the shorthand notation of $\texttt{TAP}(\mathbf{g}, \mathbf{x}^e)$ to indicate the TAP with $\mathbf{x}^e$ being the exogenous flow. We denote a solution to (4) by $\mathbf{x}^p = \min \texttt{TAP}(\mathbf{g}^p, \mathbf{x}^u + \mathbf{x}^r)$.

D. Nested Problem for AMoD in Mixed Traffic

Critically, AMoD flows react to the decisions made by private vehicles and these, in turn, react to private vehicles' flows. Hence, whenever private vehicles make their routing decisions, the AMoD fleet adjusts theirs, and vice versa. This creates a nested optimization problem between these two classes of vehicles. To give a formal definition of this game-theoretical problem we use the following bi-level optimization problem formulation

$$\min_{\{\mathbf{x}^{\mathbf{w}}\}_{\mathbf{w}\in\mathcal{W}}, \mathbf{x}^r, \mathbf{x}^p} J(\mathbf{x}) \tag{5a}$$

$$\text{s.t.} \quad \text{(3b)} - \text{(3d)}, \tag{5b}$$

$$\mathbf{x}^p \in \arg\min \texttt{TAP}(\mathbf{g}^p, \mathbf{x}^u + \mathbf{x}^r), \tag{5c}$$

which has the same structure as (3) with the additional constraint (5c). The latter constraint refers to the TAP (the lower-level problem), which depends on the solution of the full problem (upper-level). Note that the upper-level problem is minimizing over the AMoD users, rebalancing, and privately-owned vehicle flows.

This phenomenon has been identified and is often described in a *Stackelberg game* framework. In this setting, there is a *leader* agent (in our case the AMoD manager) and a *follower* (the private vehicles). In transportation networks, Korilis et al. [17] derived sufficient conditions to solve this problem when the network has parallel links. Under a similar setting, Lazar et al. [18] have analyzed the links' capacity and price of anarchy for mixed traffic. Although these models enable a better understanding of the phenomenon, they are not applicable to general networks and one can hardly assess the benefits of system-centric routing in realistic networks. To address this limitation, we will leverage the iterative approach [12] to compute an equilibrium between the private vehicles' and AMoD flows.

Discussion: A few comments are in order. First, we assume the demand to be time-invariant. This assumption is in line with densely populated urban environments, where requests change more slowly compared to the average duration of a trip. Second, we use the BPR function to relate traffic flows to travel time and allow flows to be fractional. While not capturing microscopic traffic phenomena, these approximations stem from established modeling assumptions suiting the macroscopic perspective of our study.

III. AMoD Routing and Rebalancing Problem

As mentioned earlier, the problem of routing and rebalancing stated in (3) is non-convex for typical travel time

functions such as BPR. This happens due to the term $t(x_{ij})x^r_{ij}$ in the objective function which takes products of the form $k(x^u_{ij})^n x^r_{ij}$ with k and n being a constant and the order of the polynomial, respectively. To overcome this issue, we take the suggested piecewise-affine approximation in [19] and extend it to a 3-line approximation. We present the derivation of the 3-line segment case (CARS3) followed by a disjoint formulation of the problem which will serve as a benchmark for comparison.

A. 3-line Piecewise-affine Approximation (CARS3)

We approximate the latency function (Eq. (2)) using a piecewise-affine function as shown in Fig. 2. Note that the 2-line approximation (CARS) presented in [11] might not provide a very accurate estimate of travel times when the flow is around the capacity level (Fig. 2), therefore, we approximate the travel time function using a 3-line piecewise-affine function. To construct this approximation, we follow a similar approach as in the 2-line case [11]. Let the piecewise-linear function be

$$\hat{t}_{ij}(x) = \begin{cases} at^0_{ij}, & \text{if } x < \theta_{ij} \\ at^0_{ij} + b_{ij}(x - \theta^{(1)}_{ij}), & \text{if } \theta^{(1)}_{ij} \le x \le \theta^{(2)}_{ij} \\ at^0_{ij} + b_{ij}(\theta^{(2)}_{ij} - \theta^{(1)}_{ij}) + c_{ij}(x - \theta^{(2)}_{ij}), & \text{if } \theta^{(2)}_{ij} \le x, \end{cases}$$

where a, b_{ij} and c_{ij} are constant values with $a = 1$; $b_{ij} = \beta/m_{ij}$; and $c_{ij} = \sigma/m_{ij}$. The slope of the function is β for $x_{ij} \in (\theta^{(1)}_{ij}, \theta^{(2)}_{ij})$ and σ for $x_{ij} > \theta^{(2)}_{ij}$. Moreover, $\theta^{(1)}$ and $\theta^{(2)}$ are the normalized ($\theta^{(1)} = \theta^{(1)}_{ij}/m_{ij}$), non-smooth thresholds of the travel time function. Assuming $\theta^{(2)}_{ij} \ge \theta^{(1)}_{ij}$ and $\sigma, \beta > 0$ we define two new sets of slack variables as

$$\varepsilon^{(1)}_{ij} = \max\{0, x_{ij} - \theta^{(1)}_{ij} - \varepsilon^{(2)}_{ij}\}, \tag{7a}$$
$$\varepsilon^{(2)}_{ij} = \max\{0, x_{ij} - \theta^{(2)}_{ij}\}, \tag{7b}$$

where $\varepsilon^{(1)}_{ij}$ is the excess flow after $\theta^{(1)}_{ij}$ and up to $\theta^{(2)}_{ij} - \theta^{(1)}_{ij}$, and $\varepsilon^{(2)}_{ij}$ is the excess flow after $\theta^{(2)}_{ij}$. Note that $\varepsilon^{(1)}_{ij}$ is defined in terms of $\varepsilon^{(2)}_{ij}$ to ensure that it is upper-bounded by $\theta^{(2)}_{ij} - \theta^{(1)}_{ij}$. Using these definitions we are ready to analyze and propose a tractable cost function. To this end, we focus our attention on an element-wise analysis of the first term (non-convex part) of objective (3a) using $\hat{t}$ instead of t, which we call $\hat{J}_{ij}$.

$$\begin{aligned} \hat{J}_{ij} &= \hat{t}_{ij}(x_{ij})x^u_{ij} && \text{(8a)} \\ &= (at^0_{ij} + b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij} + c_{ij}t^0_{ij}\varepsilon^{(2)}_{ij})x^u_{ij} && \text{(8b)} \\ &= at^0_{ij}x^u_{ij} + b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}(\varepsilon^{(1)}_{ij} + \varepsilon^{(2)}_{ij} + \theta^{(1)}_{ij} - x^r_{ij} - x^e_{ij}) \\ &\quad + c_{ij}t^0_{ij}\varepsilon^{(2)}_{ij}(\varepsilon^{(2)}_{ij} + \theta^{(2)}_{ij} - x^r_{ij} - x^e_{ij}) && \text{(8c)} \\ &\le at^0_{ij}x^u_{ij} + b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}(\varepsilon^{(1)}_{ij} + \varepsilon^{(2)}_{ij} + \theta^{(1)}_{ij} - x^e_{ij}) \\ &\quad + c_{ij}t^0_{ij}\varepsilon^{(2)}_{ij}(\varepsilon^{(2)}_{ij} + \theta^{(2)}_{ij} - x^e_{ij}), && \text{(8d)} \end{aligned}$$

where in (8c) we express x^u_{ij} by using a combination of (1) and (7); in the last step (8d), we add to $\hat{J}_{ij}$ the term $b_{ij}t^0_{ij}\varepsilon_{ij}x^r_{ij}$. By adding this term to $\hat{J}$, we consider a relaxation of the original problem (i.e., minimizing an upper bound of $\hat{J}$ (8d) as opposed to the original $\hat{J}$ in (8a)). This relaxation allows the proposed objective to be quadratic. Moreover, even though the quadratic term $b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}\varepsilon^{(2)}_{ij}$ is not guaranteed to be convex, we have that $\varepsilon^{(1)}_{ij}\varepsilon^{(2)}_{ij} = 0$ if $x_{ij} < \theta^{(2)}_{ij}$. Additionally, notice that when $x_{ij} > \theta^{(2)}_{ij}$ the residual flow $\varepsilon^{(1)}_{ij} = (\theta^{(2)}_{ij} - \theta^{(1)}_{ij})$. Therefore, we can replace $b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}\varepsilon^{(2)}_{ij}$ with $b_{ij}t^0_{ij}(\theta^{(2)}_{ij} - \theta^{(1)}_{ij})\varepsilon^{(2)}_{ij}$ and write the objective function of the QP as

$$\begin{aligned} J^{\text{QP}}_{ij} &= at^0_{ij}x^u_{ij} + b_{ij}t^0_{ij}\varepsilon^{(1)}_{ij}(\varepsilon^{(1)}_{ij} + \theta^{(1)}_{ij} - x^e_{ij}) \\ &\quad + c_{ij}t^0_{ij}\varepsilon^{(2)}_{ij}(\varepsilon^{(2)}_{ij} + \theta^{(2)}_{ij} - x^e_{ij}) \\ &\quad + b_{ij}t^0_{ij}(\theta^{(2)}_{ij} - \theta^{(1)}_{ij})\varepsilon^{(2)}_{ij} \\ &= \hat{t}_{ij}(x_{ij})x^u_{ij} + \hat{t}^{a=0}_{ij}(x_{ij})x^r_{ij}, \end{aligned} \tag{9}$$

where $\hat{t}^{a=0}(x)$ is equal to $\hat{t}(x)$ with parameter $a = 0$, and where $\varepsilon^{(1)}_{ij}$ and $\varepsilon^{(2)}_{ij}$ are linearly constrained as follows:

$$\varepsilon^{(1)}_{ij} \ge 0, \quad \varepsilon^{(1)}_{ij} \ge x_{ij} - \theta^{(1)}_{ij} - \varepsilon^{(2)}_{ij}, \tag{10a}$$
$$\varepsilon^{(2)}_{ij} \ge 0, \quad \varepsilon^{(2)}_{ij} \ge x_{ij} - \theta^{(2)}_{ij}. \tag{10b}$$

By analyzing this convex approximation J^{QP} with both J and $\hat{J}$, we observe that the implication of adding the extra term is taking into account congestion-aware rebalancing when the flow is greater than $\theta^{(1)}_{ij}$. Nevertheless, this congestion-aware routing of the rebalancing vehicles has a lower impact in J^{QP} than the AMoD users flows since $a = 0$ in $\hat{t}^{a=0}_{ij}(x_{ij})x^r_{ij}$, i.e., the function starts to increase from an initial point equal to zero instead of t^0_{ij}. Considering that the number of rebalancing vehicles has a minor impact on J in comparison to road congestion, and the fact that it converges to zero for perfectly symmetric demand distributions [10], J^{QP} can be used as a model to estimate the total travel time on road arcs. Our empirical studies show that, when no rebalancing is considered, the difference between the solution J^* and J evaluated with the optimal solution of the Quadratic Program (QP) is typically less than 5% (Fig. 5). In contrast with the previous method to the original CARS model in [19], we get a better convex approximation of the original problem. To summarize, the QP problem is to minimize (9) subject to (3b)-(3d), and (10a)-(10b).

An important trade-off worth noting is the difference between CARS and CARS3. Even though CARS3 provides a better approximation of the cost function and hence a better solution to the problem, it requires $|\mathcal{A}|$ additional variables and linear constraints.

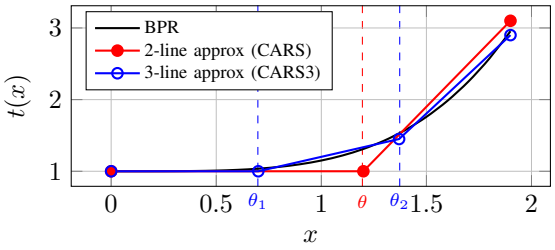


Fig. 2: Travel time function approximation.

B. Disjoint Strategy

Another way of addressing the system-centric routing and re-balancing problem is to solve the problem using a disjoint method instead of the joint approach. That is, to solve the system-centric problem for AMoD users first, and then solve the rebalancing problem formulated as a linear program (LP).

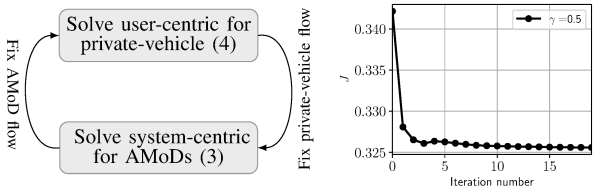


Fig. 3: (a): A sketch of the procedure for solving the bi-level problem (5). (b): An example of the total cost converging for an AMoD penetration rate of 0.5 on the NYC sub-network.

A formal definition of this problem is first solving

$$\min_{\{\mathbf{x}^w\}_{w\in\mathcal{W}}} \sum_{(i,j)\in\mathcal{A}} t_{ij}(x_{ij})x^u_{ij}, \quad \text{s.t.} \quad (3b), (3d), \tag{11}$$

and then using the resulting optimal $\mathbf{x}^{u*}$ as an input to

$$\min_{\mathbf{x}^r} \quad \mathbf{c}'\mathbf{x}^r, \quad \text{s.t.} \quad (3c), (3d). \tag{12}$$

It is important to point out that the system-centric Problem (11) is a constrained nonlinear program (NLP) which might take time to solve. In contrast to the disjoint formulation, the methodology we propose offers the possibility to solve the problem as a QP, which is usually faster than a higher order NLP and provides global optimality guarantees.

C. Iterative Solution Nested Problem

To compute an equilibrium for the nested problem (5) outlined in Section II-D, we leverage the framework developed in [12] which uses an iterative approach to reach an equilibrium between the private and AMoD flows (Fig. 3). Instead of solving the bi-level Problem (5), we solve Problem (3) with one of the methods presented in this Section (CARS, CARS3 or Disjoint) and (4) iteratively and use the output of each problem as the input to the other one. In other words, consider a private vehicle demand $\mathbf{g}^u$ and solve $\mathbf{x}^p = \min \texttt{TAP}(\mathbf{g}^p, \mathbf{0})$. Then, solve the AMoD routing and rebalancing problem (3) for AMoD demand $\mathbf{g}^u$ with fixed input $\mathbf{x}^p$ (the solution of the previously solved TAP). Since private vehicles were unaware of AMoDs in the system while solving the TAP, we again solve the problem considering a fixed flow equal to $\mathbf{x}^u + \mathbf{x}^r$, i.e., $\mathbf{x}^p = \texttt{TAP}(\mathbf{g}^p, \mathbf{x}^u + \mathbf{x}^r)$, and iterate this process until it converges as shown in Fig. 3b.

Also, note that both the disjoint problem in Section III-B and the iterative model allow for updating the component $\mathbf{t}^0$ in $\mathbf{c}$ for the travel times $\mathbf{t}(\mathbf{x})$ from the solution of (11) or previous iteration of the iterative method. This results in a more accurate cost function in terms of the travel time weight for the rebalancing problem.

We do not provide theoretical arguments on the uniqueness or stability of the players' (AMoD and private vehicles) equilibria, due to the non-separability of the cost functions with respect to their individual players' strategies [20]. Yet, empirically, this iterative algorithm always converged in a few iterations to results that are consistent for different penetration rates. We leave the theoretical study of the properties of the equilibria found to future research.

IV. Experiments

In order to validate our proposed routing algorithms, we consider a data-driven case study on a sub-network of New York City (NYC). The network was built using two data sources: OpenStreetMaps [21], from which we retrieve the network topology and road characteristics, and the recently released *Uber Movement Speed Data set* [22], needed to assign speed data to road segments (available hourly). We build a sub-network (Fig. 4) consisting of 28 nodes, 90 edges and 8 zones (green dots). We use the three methodologies described in Section III (CARS, CARS3 and Disjoint) to solve the fleet routing and rebalancing problem and compare their results against each other. Our experiments reveal that using CARS and CARS3 result in accurate solutions with low running times for these networks.

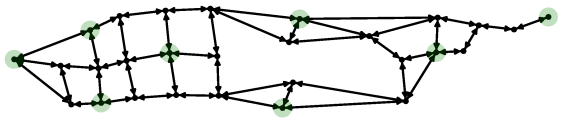


Fig. 4: NYC subnetwork

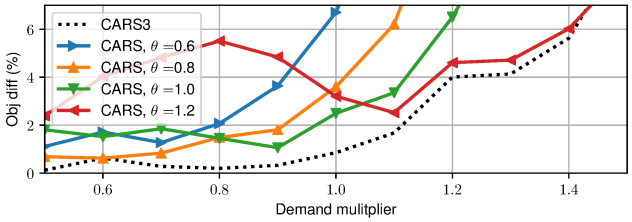


Fig. 5: Deviation in percentage terms between the approximated model and the optimal solution of the non-rebalanced system-centric problem.

A. Accuracy of CARS and CARS3

With the use numerical examples, we show how the optimal solution of CARS and CARS3 compare with the optimal solution of the system-centric problem. To achieve this, we consider the case in which rebalancing is not required, i.e., constraints (3c) are excluded and variables $\mathbf{x}^r$ are set to zero. Then, the non-rebalanced routing problem becomes the system-centric traffic assignment problem with exogenous flow (problem (11)). This problem is convex [16] and can be solved using nonlinear programming (NLP) algorithms.

This experiment assesses the offset of the total cost between the approximate models (CARS, CARS3) and the optimal solution considering the non-rebalancing system-centric model. To make a fair comparison, the solution of CARS and CARS3 is evaluated in the original cost function $J(\mathbf{x})$ from (3a). We gather results for different traffic intensities (Figure 5). The purpose of working with different demands is to investigate the approximation quality of $\hat{t}(\cdot)$ (Fig. 2) at different flow levels.

Our empirical results show that the CARS3 model outperforms the CARS method for different thresholds θ and demand rates $\mathbf{g}$. We attribute this behavior to the fact that the 3-line model yields a more precise approximation to the travel time function than the 2-line one. In addition, we observe that the deviation increases as demand increases. This arises because the piecewise linear function is less precise for high traffic intensities given its linear nature. We conclude the CARS3 model to be a good approximation as its deviations are relatively low.

B. Computational Time and Evaluation of the Cost

We compare the running times of CARS, CARS3, and Disjoint as well as the quality of their solutions. For each

TABLE I: Computational times and objective function for different models and networks. μ_τ and σ_τ are the average computational time (seconds) and variance over 30 samples, respectively. The average cost is denoted with $\bar{J}$.

Model	Type	NYC		
		μ_τ [s]	σ_τ [s]	$\bar{J}$
CARS	QP	**0.170**	8e-4	0.324
CARS3	QP	**0.215**	3e-3	0.317
Disjoint		**24.88**	1.72	0.31
System-centric	NLP	24.88	1.72	
Rebalance	LP	4e-5	2e-10	

approach, we solve 30 problems by multiplying the OD demand vector $\mathbf{g}$ by a uniform distributed random variable in the range of $[0.8, 1.2]$. For each run i, we collect the computational time τ_i as well as J_i which is computed by applying (3a) to each solution. Table I reports the mean μ_τ and variance σ_τ of the computational time. In addition, we report the average objective function divided by the total demand $\bar{J}_i = J_i/(\sum_{\mathbf{w}\in W} d_{\mathbf{w}})$.

All the scenarios studied were performed using an Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz and 32 GB of RAM memory. To solve the NLP we used the IPOPT solver [23], whereas the QP and LP programs were solved using Gurobi 8.1.1 [24].

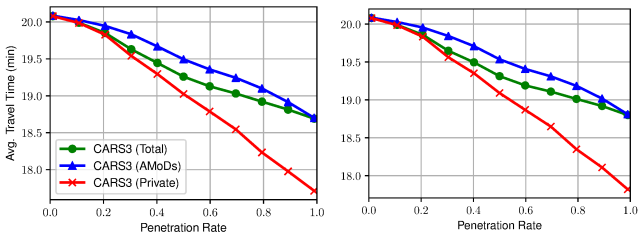
As expected, we observe that the disjoint model is the slowest, given that its first step requires solving a NLP (followed by a significantly faster solution of an LP). This method takes about 100 times more time than solving CARS3. Moreover, given that CARS3 requires more variables and constraints, it takes about 30% more time than CARS to solve.

Furthermore, our results of $\bar{J}$ show that the Disjoint method finds the best solution between the three models. The reason for this is that its model for routing is not an approximation. Nevertheless, the solutions of CARS and CARS3 are less than 4% and 2% away from the Disjoint solution, respectively. Arguably, this result might suggest that the benefit of solving the problem jointly is not as valuable as assumed, which coincides with the results of [10]. However, it is worth mentioning that these results are sensitive to different OD demand distributions. As an example, for perfectly symmetrical OD demands, rebalancing plays no role in the optimization process.

C. System-optimal Routing and Rebalancing Trade-off

Considering the existence of selfish privately-owned vehicles and centrally-controlled AMoD vehicles, we analyze the trade-off that exists between system-optimal AMoD routing and the additional traffic due to AMoD rebalancing in terms of average travel times. We tackle the bi-level Problem (5) following the iterative methodology presented in Section III-C. We use different penetration rates of AMoD customers with respect to the total demand. More specifically, we let $\gamma \in [0, 1]$ be the penetration rate and $\mathbf{g}$ the total OD demand. Then, we assume that $\mathbf{g}^u = \gamma\mathbf{g}$ and $\mathbf{g}^p = (1-\gamma)\mathbf{g}$ are the AMoD's and private vehicles' demand, respectively. In this paper, we choose the same demand distribution for AMoD and private vehicles. Yet, different demand separation criteria can be readily implemented in this framework.

As shown in Fig. 6a, the introduction of AMoD users into the system not only improves the overall travel time of AMoD users themselves, but reduces the travel time of private vehicles even more. This is because smart routing



(a) NYC, no rebalancing **(b)** NYC rebalancing

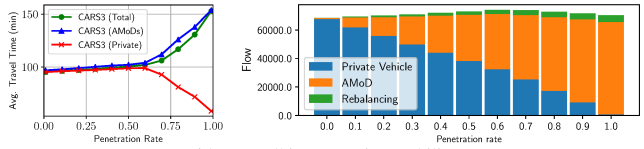
Fig. 6: Travel times for AMoD users, private vehicles and all vehicles (total) for different penetration rates of AMoDs in the network.

decisions of AMoD vehicles reduce the traffic intensity on congested roads, which consequently allows private vehicles to travel faster. As AMoD users begin to enter the system, we see that the average travel time per vehicle decreases compared to the uncontrolled traffic scenario. Moreover, the travel time of commuting through the fastest route (private vehicles) decreases as more AMoD users are in the system.

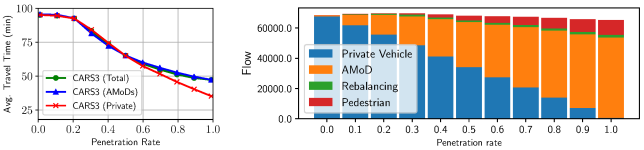
Fig. 6 shows the interaction between the two classes of vehicles when rebalancing is used or not. For NYC, the impact of rebalancing is negligible, and increasing the percentage of AMoD users in the network allows to reduce travel time by up to 10%. Notably, these results are in line with the low-to-medium congestion cases in the peak hour presented in [10, Sec. 5.2]. In particular, we note that for a 100% AMoD penetration, rebalancing slightly increases the overall travel times. Yet, in general, the impact of rebalancing on the system-level performance depends on the network topology, and on the symmetry of the OD demand distribution.

D. Walking and Micromobility Options

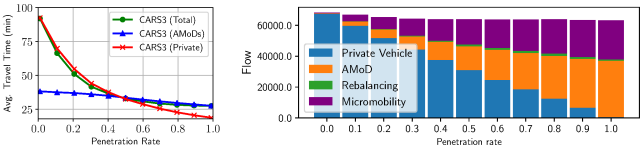
In order to study the impact of centralized routing under high congestion levels, we run experiments with a higher overall demand level (2.5 times higher than in Fig. 6). As in the previous experiment, we run the analysis for different penetration rates. Notably, the initial travel times shown in Fig. 7 are in line with the high congestion case in the peak hour in [10, Sec. 5.2]. We observe in Fig. 7a that without considering the walking or micromobility options, the injection of AMoD users to the network increases travel times. This is a result of the additional rebalancing flow needed to operate the system in high demand periods, and happens due to the evaluation of $t(\cdot)$ at those points. Every additional flow increases travel times quartically when congestion is high. In contrast, by leveraging the possibility of walking (Fig. 7b), the decrease in the overall travel time is much higher for higher AMoD penetration rates. In fact, for a 100% penetration rate, the overall travel time is halved compared to a 0% penetration rate. Additionally, we consider the possibility of using micromobility vehicles. In particular, we analyze the case when electric scooters are available to AMoD users everywhere, for which we assume an average speed of 10 mph and the same network as the walking network $\mathcal{G}_{\mathrm{W}}$. Fig. 7c shows that the average travel time for an AMoD user is lower than for selfish users when penetration rates are low. This happens because even for a 0% penetration rate, AMoD users resort to e-scooters which are not available to private vehicles' owners. Similar to other examples, the travel times for both e-scooters and private vehicles decrease as the penetration rate increases. In



(a) No walking nor micromobility.



(b) Walking case (speed considered at 3.1 mph).



(c) Micromobility case (e-scooters at 10 mph).

Fig. 7: Effect of alternative mode of transport in NYC when demand is high. We increase demand by a 2.5 factor, i.e., we use 2.5**g**.

conclusion, by comparing Fig. 7a with Fig. 7b and 7c, we see that pure AMoD systems might decrease the system-level performance due to the additional congestion resulting from rebalancing the AMoD vehicles. Yet, combining centralized-routing with the possibility of walking or using micromobility solutions such as e-scooters can significantly improve the overall travel times.

V. Conclusions

In this paper we studied the achievable benefits of centrally controlling an Autonomous Mobility-on-Demand (AMoD) system under mixed traffic conditions. With the goal of minimizing the customers' travel time, we extended a previously presented quadratic model [11] by improving its accuracy and included reactive exogenous traffic flows. Assuming the exogenous traffic (private vehicles) to act selfishly, we leveraged an iterative method [12] to study the interaction between AMoD and private cars. Finally, we presented numerical experiments to compare the proposed method with a disjoint strategy, and to gain insights on the achievable benefits for different AMoD penetration rates and micromobility options. Our results showed that the proposed method outperforms the disjoint strategy in terms of computational time, and revealed that combining AMoD rides with walking and micromobility options can significantly improve the overall system-level performance.

This work can be extended as follows. First, given the large computational time of the disjoint problem (NLP) we would like to propose a MSA-type method to solve the AMoD system-centric TAP considering exogenous flow, possibly leveraging computationally efficient algorithms such as in [25]. Second, we would like to generalize the approximation model to n line segments, and provide theoretical bounds on the model error. Third, given that the solution of these models are in terms of flow, we would like to include route-recovery strategies and apply this framework to larger networks through high-fidelity simulations. Finally, we would like to consider a more general intermodal setting as in [19], [26] by including public transportation options.

References

[1] J. G. Wardrop, "Road paper. some theoretical aspects of road traffic research." *Proceedings of the institution of civil engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[2] J. Guanetti, Y. Kim, and F. Borrelli, "Control of connected and automated vehicles: State of the art and future challenges," *Annual Reviews in Control*, vol. 45, pp. 18–40, 2018.

[3] S. Wollenstein-Betech, I. C. Paschalidis, and C. G. Cassandras, "Joint pricing and rebalancing of autonomous mobility-on-demand systems," *arXiv preprint arXiv:2003.13614*, 2020.

[4] R. M. Swaszek and C. G. Cassandras, "Load balancing in mobility-on-demand systems: Reallocation via parametric control using concurrent estimation," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 2148–2153.

[5] S. Hörl, C. Ruch, F. Becker, E. Frazzoli, and K. W. Axhausen, "Fleet control algorithms for automated mobility: A simulation assessment for Zurich," in *Annual Meeting of the Transportation Research Board*, 2018.

[6] M. W. Levin, K. M. Kockelman, S. D. Boyles, and T. Li, "A general framework for modeling shared autonomous vehicles with dynamic network-loading and dynamic ride-sharing application," *Computers, Environment and Urban Systems*, vol. 64, pp. 373 – 383, 2017.

[7] R. Zhang and M. Pavone, "Control of robotic Mobility-on-Demand systems: A queueing-theoretical perspective," *Int. Journal of Robotics Research*, vol. 35, no. 1–3, pp. 186–203, 2016.

[8] R. Iglesias, F. Rossi, R. Zhang, and M. Pavone, "A BCMP network approach to modeling and controlling Autonomous Mobility-on-Demand systems," in *Workshop on Algorithmic Foundations of Robotics*, 2016.

[9] M. Pavone, S. L. Smith, E. Frazzoli, and D. Rus, "Robotic load balancing for Mobility-on-Demand systems," *Int. Journal of Robotics Research*, vol. 31, no. 7, pp. 839–854, 2012.

[10] F. Rossi, R. Zhang, Y. Hindy, and M. Pavone, "Routing autonomous vehicles in congested transportation networks: Structural properties and coordination algorithms," *Autonomous Robots*, vol. 42, no. 7, pp. 1427–1442, 2018.

[11] M. Salazar, M. Tsao, I. Aguiar, M. Schiffer, and M. Pavone, "A congestion-aware routing scheme for autonomous mobility-on-demand systems," in *European Control Conference*, 2019.

[12] A. Houshmand, S. Wollenstein-Betech, and C. G. Cassandras, "The penetration rate effect of connected and automated vehicles in mixed traffic routing," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1755–1760.

[13] M. J. Beckmann, C. B. McGuire, and C. B. Winsten, "Studies in the Economics of Transportation," p. 359, 1955.

[14] S. Wollenstein-Betech, C. Sun, J. Zhang, and I. C. Paschalidis, "Joint estimation of od demands and cost functions in transportation networks from data*," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 5113–5118.

[15] Bureau of Public Roads, "Traffic assignment manual," U.S. Dept. of Commerce, Urban Planning Division, Tech. Rep., 1964.

[16] M. Patriksson, *The traffic assignment problem: models and methods*. Courier Dover Publications, 2015.

[17] Y. A. Korilis, A. A. Lazar, and A. Orda, "Achieving network optima using stackelberg routing strategies," *IEEE/ACM transactions on networking*, vol. 5, no. 1, pp. 161–173, 1997.

[18] D. A. Lazar, S. Coogan, and R. Pedarsani, "Capacity modeling and routing for traffic networks with mixed autonomy," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017.

[19] M. Salazar, F. Rossi, M. Schiffer, C. H. Onder, and M. Pavone, "On the interaction between autonomous mobility-on-demand and the public transportation systems," in *Proc. IEEE Int. Conf. on Intelligent Transportation Systems*, 2018, Extended Version, Available at https://arxiv.org/abs/1804.11278.

[20] P. T. Harker, "Multiple equilibrium behaviors on networks," *Transportation science*, vol. 22, no. 1, pp. 39–46, 1988.

[21] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org , 2017.

[22] Data retrieved from Uber Movement, "2019 Uber Technologies, Inc," 2019. [Online]. Available: https://movement.uber.com

[23] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical programming*, vol. 106, no. 1, pp. 25–57, 2006.

[24] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2020. [Online]. Available: http://www.gurobi.com

[25] K. Solovey, M. Salazar, and M. Pavone, "Scalable and congestion-aware routing for autonomous mobility-on-demand via frank-wolfe optimization," in *Robotics: Science and Systems*, 2019.

[26] M. Salazar, N. Lanzetti, F. Rossi, M. Schiffer, and M. Pavone, "Intermodal autonomous mobility-on-demand," *IEEE Transactions on Intelligent Transportation Systems*, 2019.