

---

# SEQUENTIAL DECISION-MAKING FOR RIDE-HAILING FLEET CONTROL: A UNIFYING PERSPECTIVE

---

**Stefan Pilot**  
 RWTH Aachen University  
 pilot@cl.rwth-aachen.de

**Murwan Siddig**  
 University of Florida  
 msiddig@ufl.edu

December 29, 2025

## ABSTRACT

This paper provides a unified framework for the problem of controlling a fleet of ride-hailing vehicles under stochastic demand. We introduce a sequential decision-making model that consolidates several problem characteristics and can be easily extended to include additional characteristics. To solve the problem, we design an efficient procedure for enumerating all feasible vehicle-to-request assignments, and we introduce scalable techniques to deal with the exploration-exploitation trade-off. We construct reusable benchmark instances that are based on real-world data and that capture a range of spatial structures and demand distributions. Our proposed modelling framework, policies and benchmark instances allow us to analyze interactions between problem characteristics that were not previously studied. We find no significant difference between revenue generated by internal combustion engine fleets and fast-charging electric fleets, but both significantly outperform slow-charging electric fleets. We also find that pooling increases the revenue, and reduces revenue variability, for all fleet types. Our contributions can help coordinate the significant research effort that this problem continues to receive.

**Keywords** Ridehailing, · Sequential Decision-making, · Approximate Dynamic Programming

## 1 Introduction

Ride-hailing is an on-demand transportation service in which customers request rides, typically through a mobile application or an online platform. The service provider matches the requests with available vehicles that transport the customers to their destinations. Ride-hailing service providers can also offer pooling options, where multiple customers share a ride at the same time, usually at a lower cost. Examples of ride-hailing services include Uber, Lyft, DiDi, and Waymo. Ride-hailing services have become increasingly popular in recent years due to their convenience, affordability, and flexibility compared to traditional taxi services and public transportation. They also have the potential to reduce traffic congestion and emissions by optimizing vehicle usage and reducing the number of empty seats on the road (Ackermann and Rieck, 2023).

Optimal control of ride-hailing systems (RHSs) is a challenging sequential decision-making problem. Every decision made by the service provider impacts not only the current RHS performance, but also the RHS's future performance. There are many possible decisions that the provider can make, e.g., dispatching vehicles to satisfy customer requests, keeping vehicles parked at their current locations, repositioning empty vehicles to different locations, and recharging vehicles if they are electric. The decision of dispatching a vehicle to satisfy customer requests is, in itself, a complex decision with many layers. For example, which vehicle should be assigned to which requests? If a vehicle is assigned to multiple requests, should the requests be pooled together or served one after the other? In what order should they be picked up and dropped off? All of these decisions are interrelated and must be executed dynamically while satisfying operational constraints (e.g., vehicle availability, capacity, driving range) and customer satisfaction constraints (e.g., waiting time until the vehicle arrives, detour time to pick up other passengers).

The complex nature of this optimal control problem has led to a rich and diverse body of literature studying various problem settings, modeling choices, and solution methods (see surveys Mourad et al., 2019; Soeffker et al., 2022;

Hildebrandt et al., 2023). As a result, the literature appears fragmented and would benefit significantly from a more coordinated research effort that facilitates meaningful comparisons and generalization of findings across studies. This work is an attempt towards a unified framework for studying RHS control.

### 1.1 Scope and Terminology

RHSs can take different forms depending on the service provider’s business model, the type of vehicles used, and the level of automation. For instance, the fleet of vehicles used by a RHS can consist of internal combustion engine (ICE) vehicles, electric vehicles (EVs), or a mix of both. We consider both ICE vehicles and EVs. Furthermore, RHSs can be fully autonomous (e.g., Waymo and Zoox) or operated by human drivers. In the latter case, drivers can either be employees of the service provider (e.g., Ioki, Moia, and Via) or independent contractors (e.g., Uber and Lyft). We focus on RHSs that operate a fleet of vehicles owned by the service provider, regardless of whether the vehicles are autonomous or driven by employees.

Different studies use different terminology to refer to the RHSs of our interest (see e.g., Qin et al., 2022). The terminology is especially inconsistent in studies that allow the option of pooling. We emphasize that the focus of this work does not include rental car sharing (Boyaci et al., 2015), sharing rides with parcels (Li et al., 2014), matching private drivers and riders with similar plans (Agatz et al., 2012), and dial-a-ride paratransit with large-capacity vehicles and pre-specified requests (Heitmann et al., 2024). To that end, we use the term *pooling-enabled* RHS to emphasize the taxi-like on-demand nature of RHSs with the option to share rides that we study.

### 1.2 Problem statement

We study a pooling-enabled RHS that controls a fixed-size fleet of homogeneous vehicles. The fleet consists of either ICE vehicles or EVs, and each vehicle has a fixed capacity for passengers and a fixed maximum driving range before it needs to recharge or refuel.

The RHS dispatches vehicles to serve travel requests. Travel requests are not known in advance and arrive over the planning horizon. The specific point in time at which the operator makes decisions is called the decision epoch. At each decision epoch, the RHS chooses which requests to accept, which vehicles to dispatch to serve the accepted requests, and how to dispatch them. Vehicles not serving requests may be relocated, left idle, or refueled (recharged, if EVs).

The suitability of a vehicle to serve a request depends on the request characteristics and the vehicle state. A request is characterized by its origin, destination, headcount, latest response time, latest pickup time, and fare. If a request is not accepted in one epoch, it is reconsidered in the next until its latest response time passes, after which it is treated as rejected. The RHS may pool accepted requests and serve them simultaneously with one vehicle. Pooling can increase travel time relative to serving requests individually, but passengers pay a lower fare. Pooling can also reduce waiting times in the rush hour if it allows the same fleet to serve more requests in a given period. The RHS decides whether to pool and which vehicles to dispatch.

A vehicle state includes its current location, destination (if occupied), remaining capacity (empty seats), driving range (for EVs, the battery level), and actionable time (when it can start a new task). A vehicle can be assigned to a request only if it has enough empty seats to accommodate the request headcount, can reach the request origin before the latest pickup time, and has sufficient driving range to reach the origin and then the destination.

The RHS aims to maximize its total expected reward over the planning horizon. The total reward is the sum of the fares collected from the passengers minus detour penalties and refuelling or recharging costs incurred by the system.

### 1.3 Contributions

The main contributions of this paper are:

- **Model:** We present a unified sequential decision-making model that (1) consolidates many of the previously studied features, decisions, and constraints, and (2) can be easily extended to include additional features, decisions, constraints, and objectives.
- **Policies:** We present a tractable and well-performing value function approximation (VFA) policy that combines the advantages of existing approaches. This involves introducing an efficient procedure for enumerating all feasible vehicle-to-request assignments and scalable techniques for handling the exploration-exploitation tradeoff.

- **Benchmark instances:** We create a set of benchmark instances that (1) are based on real-world data, (2) include a variety of spatial structures and demand patterns, and (3) are easily accessible, and can be directly used for future research.
- **Numerical results:** We perform extensive numerical experiments, which show that our VFA policy significantly outperforms a baseline policy, which is consistent with previous literature. We also analyze previously unstudied interactions between problem characteristics, and find that both ICE fleets and fast-charging EV fleets obtain high rewards compared to slow-charging EV fleets. We also find that pooling increases the revenue of the RHS and reduces revenue variability for all fleet types.

The remainder of this paper is organized as follows. Section 2 surveys relevant literature. Section 3 describes the sequential decision-making model. Section 4 defines our two policies. Section 5 presents our benchmark instances and numerical experiments. Section 6 concludes the paper and discusses future research directions.

## 2 Related Work

In this section, we provide an overview of the literature on the topic by highlighting the different decisions, constraints, and performance metrics that have been considered. We also discuss the different models and solution methods that have been used to solve the problem.

### 2.1 Decisions

The most basic decision that the operator can make is to assign an *empty* vehicle to a single travel request. We refer to this decision as the *single trip* decision ( $d^{\text{sngl}}$ ). The operator can also assign an *empty* vehicle to park at its current location (Al-Kanj et al., 2020; Kullman et al., 2022; Yu et al., 2023). We refer to this as the *idle* decision ( $d^{\text{idle}}$ ). Not all studies consider the  $d^{\text{idle}}$  decision. For instance, Alonso-Mora et al. (2017) assume that empty vehicles are relocated to areas with high demand. We use a separate decision for modeling the relocation of empty vehicles. We refer to this decision as the *relocate* decision ( $d^{\text{relc}}$ ). The  $d^{\text{relc}}$  decision is commonly studied in the literature (e.g., Al-Kanj et al., 2020; Kullman et al., 2022; Tuncel et al., 2023).

Some decisions are only relevant for certain types of problem settings. For instance, the *recharge* decision ( $d^{\text{chr}}$ ) is typically only included in studies that consider electric vehicles (e.g., Al-Kanj et al., 2020; Kullman et al., 2022; Yu et al., 2023). Similarly, two decisions that are only relevant for studies that allow ride pooling are the *multi-trip* decision ( $d^{\text{mlti}}$ ) and the *pool* decision ( $d^{\text{pool}}$ ). The  $d^{\text{mlti}}$  decision assigns an *empty* vehicle to multiple requests, where the passengers of those requests may or may not occupy the vehicle simultaneously at some point during the trip (Alonso-Mora et al., 2017; Tuncel et al., 2023). The  $d^{\text{pool}}$  decision assigns an *occupied* vehicle to a new request, which is picked up before the vehicle drops off its existing occupants at their destination (Yu and Shen, 2020; Heitmann et al., 2023). To distinguish between  $d^{\text{pool}}$  and the decision of assigning an *occupied* vehicle to a new request whose passengers are picked up *after* the vehicle drops off its existing occupants, we introduce a separate decision that we refer to as the *queue* decision ( $d^{\text{queu}}$ ). The  $d^{\text{queu}}$  decision is not always considered in the literature and is sometimes modeled implicitly as a  $d^{\text{sngl}}$  decision (Kullman et al., 2022), or using tentative routes that can be adapted over time (Heitmann et al., 2023).

Some decisions are only relevant for certain types of models. For instance, the specific point in time at which the operator makes decisions is called the decision epoch. Decision epochs can be triggered by events, e.g., arrival of a new request (Kullman et al., 2022), or time-triggered at regular intervals (Alonso-Mora et al., 2017; Al-Kanj et al., 2020). We consider a time-triggered decision epoch, but our model can be easily adapted to consider an event-triggered decision epoch. Modeling time-triggered decision epochs requires special handling of vehicles that are still performing tasks at the time of the decision epoch. We do this by introducing a pseudo decision that we refer to as the *continue* decision ( $d^{\text{cont}}$ ), which ensures that the vehicle continues to perform the task it is currently performing (e.g., servicing requests, relocating or recharging). The  $d^{\text{cont}}$  decision is referred as the *null* decision by Kullman et al. (2022).

### 2.2 Constraints

The most commonly studied type of constraints concern the *waiting* time ( $t^{\text{wai}}$ ), which refers to the time elapsed between the moment at which a customer is assigned a vehicle and the moment at which the vehicle picks up the customer; the *detour* time ( $t^{\text{det}}$ ), which refers to the additional time incurred during one passenger’s trip due to picking up or dropping off other passengers along the way; and the *response* time ( $t^{\text{res}}$ ), which refers to the time elapsed between the moment at which a customer sends a request for a ride and when the system assigns a vehicle to the customer. For instance, Özkan and Ward (2020) and Kullman et al. (2022) impose a maximum waiting time constraint

to ensure that customers are picked up within a certain time after being assigned a vehicle. Some studies do not impose a constraint on the waiting time but penalize it in the objective function (Yu and Shen, 2020). A detour time constraint is only relevant for studies that consider pooling-enabled RHSs. For instance, Alonso-Mora et al. (2017) consider a detour time constraint implicitly by imposing a maximum drop-off time for every request and penalizing the detour time in the objective function. Yu and Shen (2020) do not impose a detour time constraint but penalize it in the objective function. Constraints on the response time (Erdmann et al., 2021) are not as common as the waiting and detour time constraints. However, a key benefit of modeling a response time constraint is that it defines a demand backlog for the requests that the system receives. Studies without a response time constraint typically consider all requests that are not served instantly to be lost (e.g., Al-Kanj et al., 2020) or penalize the lost requests in the objective function (Yu and Shen, 2020; Alonso-Mora et al., 2017). Another type of constraint concerns the recharge and idle decisions. While some studies assume that vehicles can charge and idle everywhere (Al-Kanj et al., 2020), others assume that vehicles can only charge in specific locations (Yu et al., 2023), only idle in specific parking lots (Ackermann and Rieck, 2023), or both (Kullman et al., 2022).

In this paper, we impose the waiting time and response time as hard constraints and penalize the detour time in the objective function. Our reasoning for this is as follows. Customers who share rides with other passengers typically pay cheaper fares, and it is reasonable to interpret a penalty cost in the objective function as a discount or compensation for the inconvenience of a detour time. Additionally, imposing the waiting time as a hard constraint allows us to restrict the time it takes a vehicle to pickup a request to some upper bound, e.g., the time it takes to complete the trip itself. Similarly, imposing the response time as a hard constraint allows us to model a demand backlog for the system that can be interpreted as what Yan et al. (2020) refer to as a “batching window”. Furthermore, we assume that vehicles can idle and fully charge everywhere. We make this assumption to simplify the problem. Nevertheless, our model can be easily adapted to consider dedicated charging stations and parking lots, as well as any combination of hard and soft constraints on the waiting time, response time, and detour time.

### 2.3 Performance metrics

The choice between imposing a constraint or adding a penalty to the objective function is subjective, in part because customers’ tolerance for waiting, response, and detour times can vary significantly. Service providers try to strike a balance between profitability and customer satisfaction. Many performance metrics have been considered in the literature to calibrate this trade-off. The most commonly studied performance metrics are the rewards generated by the system for serving requests (Al-Kanj et al., 2020), operational costs such as the cost of driving (Kullman et al., 2022), charging (Al-Kanj et al., 2020), as well as service quality metrics such as the waiting time (Alonso-Mora et al., 2017; Lyu et al., 2024), response time (Erdmann et al., 2021), and detour time (Alonso-Mora et al., 2017). Some studies consider the number of served requests (Özkan and Ward, 2020), cost of losing requests (Alonso-Mora et al., 2017; Yu and Shen, 2020), the number of available vehicles (Braverman et al., 2019), or geographical unfairness where customers whose origins and/or destinations are far from centralized locations suffer from excessive service rejections (Ben-Gal and Tzur, 2025). Many of these performance metrics are often considered simultaneously. This could be as simple as maximizing a revenue function that comprises rewards and costs (Al-Kanj et al., 2020; Kullman et al., 2022), but can also be modelled as a multi-objective problem (Lyu et al., 2024). In this paper, we use a simple revenue function that comprises the trip fares (gained by serving requests) minus the detour penalty costs and the charging costs. This revenue function can be easily adapted to consider other performance metrics. Service fairness metrics are outside the scope of this paper.

### 2.4 Models and solution methods

The solution methods used in the literature can be distinguished based on the type of policy they study, and whether constructing the policy involves solving an assignment problem. Most policies solve some variant of a linear assignment problem to match vehicles to requests or to any other task (e.g.,  $d^{\text{idle}}$  and  $d^{\text{relc}}$ ). The assignment problem can be solved as a Linear Program (LP) (e.g., Al-Kanj et al., 2020) or an Integer Program (IP) (e.g., Alonso-Mora et al., 2017). There are also policies that do not solve an assignment problem. For example, Kullman et al. (2022) learn Q-value approximations using an artificial neural network (ANN) to make separate, decentralized decisions for each vehicle.

Some studies use myopic policies. For instance, the seminal paper by Alonso-Mora et al. (2017) define a bipartite request-trip-vehicle (RTV) graph whose nodes represent vehicles and potential requests groups (i.e., shared rides) and whose edges represent feasible matches between vehicles and these rides. The authors then solve an assignment problem myopically, in a rolling-horizon fashion, without explicitly considering the downstream impact of the current decisions on the future performance of the system. A myopic policy can also be parameterized (Powell, 2011). For instance, Al-Kanj et al. (2020) solve an assignment problem that includes a parameterized rule to ensure that a vehicle

is assigned to charge if its battery level is below a certain threshold. Other studies create policies that explicitly consider the downstream impact of the current decisions by optimizing the current reward (or cost) plus an approximation of the value of future states, which the system evolves to, given the current decisions. We refer to these policies as VFA-based policies. VFA-based policies can be further distinguished based on the method used to approximate the value of future states. For instance, Al-Kanj et al. (2020) use lookup tables to approximate the value of future states and Wang et al. (2018) learn an ANN-VFA from the perspective of a single vehicle. We refer the reader to Qin et al. (2022) for a more detailed overview of solution methods used in the literature.

In this paper, we study a parameterized myopic (PM) policy and a VFA policy. To design each policy, we use a modified version of the RTV graph algorithm to enumerate the decisions of a linear assignment problem. For the PM policy, we solve the assignment problem as an IP to obtain integer solutions directly. For the VFA policy, we first solve a relaxed version of the assignment problem as an LP to obtain a VFA lookup table, which we then use in the final policy that solves the assignment problem as an IP.

### 3 Problem Formulation

We consider the problem of controlling a pooling-enabled RHS as described in Section 1.2. We model vehicle movements on a weighted directed graph  $G = (\mathcal{V}, \mathcal{A})$  where  $\mathcal{V}$  is a set of nodes and  $\mathcal{A}$  is a set of arcs. Each node  $v \in \mathcal{V}$  represents a location that can act as the origin and destination of requests. Each arc  $(u, v) \in \mathcal{A}$  represents a drivable road segment from location  $u$  to  $v$ . We assume that the graph is connected. We also assume that the time it takes a vehicle to travel the shortest path from node  $u \in \mathcal{V}$  to node  $v \in \mathcal{V}$  is deterministic and is given by  $\tau(u, v)$ . Extending our model to the case where travel times are random and vehicles can enter or leave service at random (e.g., due to accidents) is straightforward.

Furthermore, we consider a finite time horizon  $T$  with a time-triggered decision epoch  $t \in \{1, \dots, T\}$ . In every decision epoch  $t$ , we observe the system *state*, apply a *decision* on each vehicle, receive a *reward*, observe new *exogenous information*, and advance to a new state in the next decision epoch  $t + 1$  via the *transition functions*.

#### 3.1 States

The system state variable describes the state of every vehicle in the fleet and the attributes of the travel requests. The state variable at time  $t$  is given by the vector  $S_t = (R_t, D_t)$  where

- $R_t = (R_{ta})_{a \in \mathcal{A}}$  is the number of vehicles (resources) with attribute  $a \in \mathcal{A}$ , and
- $D_t = (D_{tb})_{b \in \mathcal{B}}$  is the number of travel requests (demand) with attribute  $b \in \mathcal{B}$ ,

where the attribute vectors  $a$  and  $b$  are defined as

$$a = \begin{pmatrix} o_a \\ d_a \\ l_a \\ n_a \\ t_a \end{pmatrix} = \begin{pmatrix} \text{current location} \\ \text{destination} \\ \text{driving range} \\ \text{capacity} \\ \text{actionable time} \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} o_b \\ d_b \\ n_b \\ t_b^{\text{res}} \\ t_b^p \\ f_b \end{pmatrix} = \begin{pmatrix} \text{origin} \\ \text{destination} \\ \text{headcount} \\ \text{latest response time} \\ \text{latest pickup time} \\ \text{fare} \end{pmatrix}.$$

The vehicle's current location  $o_a \in \mathcal{V}$  and destination  $d_a \in \mathcal{V}$  are nodes in the graph  $G$ . Let  $l^{\max}$  be the maximum driving range. We define  $n^{\max}$  to be the maximum capacity of a vehicle, such that  $n_a = n^{\max}$  when the vehicle is empty and  $n_a = 0$  when the vehicle is fully occupied. We distinguish between two types of vehicles in the system: (i) *empty* vehicles  $\mathcal{A}^{\text{empt}} = \{a \in \mathcal{A} \mid n_a = n^{\max} \wedge o_a = d_a\}$ , and (ii) *occupied* vehicles  $\mathcal{A}^{\text{occu}} = \{a \in \mathcal{A} \mid n_a < n^{\max} \wedge o_a \neq d_a\}$ . The actionable time  $t_a \in \mathbb{R}_{\geq t}$  is the time at which a vehicle  $a$  can start performing a new task. When  $t_a > t$ , the vehicle  $a$  is currently performing a task, and  $t_a = t$  means that the vehicle can start performing a new task immediately.

The trip's origin  $o_b \in \mathcal{V}$  and destination  $d_b \in \mathcal{V}$  are nodes in the graph  $G$ . We only consider requests where  $o_b \neq d_b$ . The headcount  $n_b \in \mathbb{Z}_{>0}$  is the number of passengers in request  $b$ , and we only consider requests where  $n_b \leq n^{\max}$ . The latest response time  $t_b^{\text{res}} \in \{1, \dots, T\}$  is the latest time at which the RHS can accept request  $b$  by assigning it to a vehicle. The latest pickup time  $t_b^p \in \{1, \dots, T\}$  is the latest time at which a vehicle dispatched by the RHS can reach the origin of request  $b$  and pick up its passengers. At time  $t$ , the set  $\mathcal{B}$  consists only of requests  $b$  such that  $t \leq t_b^{\text{res}}$  and  $t \leq t_b^p$ . The trip fare  $f_b \in \mathbb{R}_{>0}$  is the amount of money the RHS receives if it accepts request  $b$ .

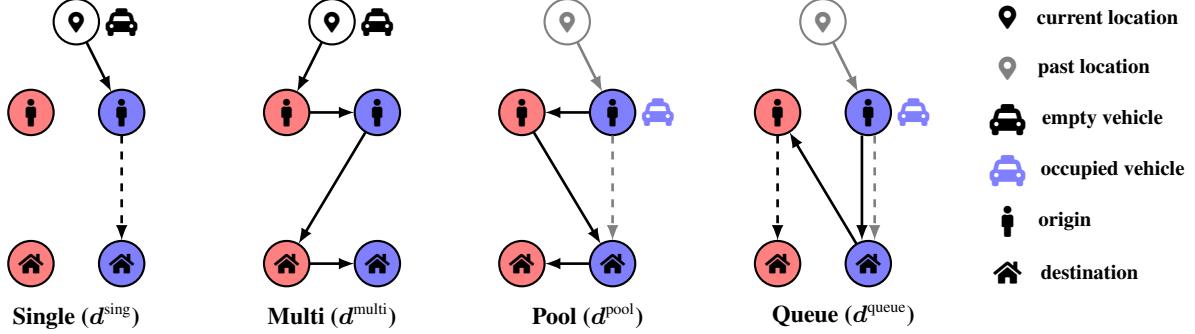


Figure 1: Four types of decisions that assign a vehicle to serve one or more requests. Solid arcs must be traversed and dashed arcs are tentative routes that may be modified later. Black arcs are part of the post-decision state. Gray dashed arcs are part of the pre-decision state, and gray solid arcs are part of the path history.

### 3.2 Decisions

A decision  $d \in \mathcal{D}$  is a task that a vehicle can be assigned to perform. For all vehicles with attribute  $a \in \mathcal{A}$  and requests with attribute  $b \in \mathcal{B}$ , let  $\mathcal{D}(a)$  be the subset of decisions that can be assigned to a vehicle  $a \in \mathcal{A}$ . Let  $\mathcal{D}(a, b) \subseteq \mathcal{D}(a)$  be the set of decisions that assign vehicle  $a$  to serve request  $b$ , see Figure 1 for an illustration. The set of all decisions is given by  $\mathcal{D} = \bigcup_{a \in \mathcal{A}} \mathcal{D}(a)$ , where

$$\mathcal{D}(a) = \mathcal{D}_a^{\text{sngl}} \cup \mathcal{D}_a^{\text{mlti}} \cup \mathcal{D}_a^{\text{pool}} \cup \mathcal{D}_a^{\text{queu}} \cup \mathcal{D}_a^{\text{relc}} \cup \mathcal{D}_a^{\text{cont}} \cup \mathcal{D}_a^{\text{idle}} \cup \mathcal{D}_a^{\text{rchr}}.$$

To define the individual subsets, we introduce the following notation. Let  $P$  be the set of paths in  $G$  and let  $\sigma$  be a function that maps each path  $p \in P$  to the driving range that a vehicle requires to traverse  $p$ . A path  $p \in P$  is *a-feasible* for vehicle  $a \in \mathcal{A}$  if it starts with  $o_a$  and  $l_a \geq \sigma(p)$ . Let  $P_a$  be the set of *a*-feasible paths, for each  $a \in \mathcal{A}$ , and let  $B \subseteq \mathcal{B}$  be a nonempty set of requests. A path  $p \in P_a$  that  $a \in \mathcal{A}$  traverses to satisfy the requests in  $B \in \mathcal{B}$  is *a-B-feasible* if it satisfies the following conditions: (1) for every request  $b \in B$ ,  $p$  visits the origin  $o_b$  before the destination  $d_b$ , and  $p$  reaches the origin  $o_b$  before the latest pickup time  $t_b^p$ ; and (2) vehicle  $a$  has enough seats to pick up and drop off the requests  $B$  in the order of  $p$ . For each  $a \in \mathcal{A}$  and  $B \subseteq \mathcal{B}$ , let  $P_{aB}$  the set of *a-B*-feasible paths and let  $P_{aB}(u, v) \in P_{aB}$  be the set of *a-B*-feasible paths that visit location  $u \in \mathcal{V}$ , then location  $v \in \mathcal{V}$ , and may or may not visit other locations before  $u$ , after  $v$ , and in between.

**Single trip:** We use  $d_b^{\text{sngl}}$  to denote the decision to assign an empty vehicle to serve a single travel request  $b \in \mathcal{B}$ . We define

$$\mathcal{D}_a^{\text{sngl}} = \left\{ d_b^{\text{sngl}} \mid b \in \mathcal{B} \wedge P_{a\{b\}} \neq \emptyset \right\} \quad (1)$$

if  $a \in \mathcal{A}^{\text{empt}}$ . Otherwise,  $\mathcal{D}_a^{\text{sngl}} = \emptyset$ .

**Multi-trip:** We use  $d_{Bp}^{\text{mlti}}$  to denote the decision to assign an empty vehicle to serve a set of requests  $B \subseteq \mathcal{B}$  on the path  $p \in P_{aB}$ . We define

$$\mathcal{D}_a^{\text{mlti}} = \left\{ d_{Bp}^{\text{mlti}} \mid B \subseteq \mathcal{B} \wedge |B| > 1 \wedge p \in P_{aB} \right\} \quad (2)$$

if  $a \in \mathcal{A}^{\text{empt}}$ . Otherwise,  $\mathcal{D}_a^{\text{mlti}} = \emptyset$ .

**Pool:** We use  $d_{bp}^{\text{pool}}$  to denote the decision to assign an occupied vehicle  $a \in \mathcal{A}$  to serve a new travel request  $b \in \mathcal{B}$ , together with the passengers that are already in the vehicle, on the path  $p \in P_{a\{b\}}$ . We define

$$\mathcal{D}_a^{\text{pool}} = \left\{ d_{bp}^{\text{pool}} \mid b \in \mathcal{B} \wedge n_a + n_b \leq n^{\max} \wedge p \in P_{a\{b\}}(o_b, d_a) \right\} \quad (3)$$

if  $a \in \mathcal{A}^{\text{occu}}$ . Otherwise,  $\mathcal{D}_a^{\text{pool}} = \emptyset$ .

**Queue:** We use  $d_b^{\text{queu}}$  to denote the decision to assign an occupied vehicle to serve a new travel request  $b \in \mathcal{B}$  after it drops off its current passengers. We define

$$\mathcal{D}_a^{\text{queu}} = \left\{ d_b^{\text{queu}} \mid b \in \mathcal{B} \wedge P_{a\{b\}}(d_a, o_b) \neq \emptyset \right\} \quad (4)$$

if  $a \in \mathcal{A}^{\text{occu}}$ . Otherwise,  $\mathcal{D}_a^{\text{queu}} = \emptyset$ .

**Relocate:** We use  $d_v^{\text{relc}}$  to denote the decision to relocate a vehicle at location  $v \in \mathcal{V}$ . Vehicles can either relocate to adjacent locations or locations that are reachable before the next decision epoch. Vehicles  $a \in \mathcal{A}$  whose actionable time  $t_a$  is after the next decision epoch  $t$  cannot relocate because the decision would be premature. We define

$$\mathcal{D}_a^{\text{relc}} = \left\{ d_v^{\text{relc}} \mid v \in \mathcal{V} \wedge \tau(o_a, v) \leq l_a \wedge t_a < t + 1 \wedge ((o_a, v) \in A \vee \tau(o_a, v) \leq t + 1) \right\} \quad (5)$$

if  $a \in \mathcal{A}^{\text{empt}}$ . Otherwise,  $\mathcal{D}_a^{\text{relc}} = \emptyset$ .

**Continue:** We use  $d^{\text{cont}}$  to denote the decision to let an already occupied vehicle continue its task with no new instructions. We define  $\mathcal{D}_a^{\text{cont}} = \{d^{\text{cont}}\}$  if  $a \in \mathcal{A}^{\text{occu}}$ , and  $\mathcal{D}_a^{\text{cont}} = \emptyset$  otherwise.

**Idle:** We use  $d^{\text{idle}}$  to denote the decision to park an empty vehicle. We define  $\mathcal{D}_a^{\text{idle}} = \{d^{\text{idle}}\}$  if  $a \in \mathcal{A}^{\text{empt}}$  at time  $t_a$ . Otherwise,  $\mathcal{D}_a^{\text{idle}} = \emptyset$ .

**Recharge:** We use  $d^{\text{rchr}}$  to denote the decision to refuel or recharge a vehicle. Vehicles  $a \in \mathcal{A}$  whose actionable time  $t_a$  is after the next decision epoch  $t$  cannot recharge because the decision would be premature. We define

$$\mathcal{D}_a^{\text{rchr}} = \{d^{\text{rchr}} \mid l_a < l^{\max} \wedge t_a < t + 1\} \quad (6)$$

if  $a \in \mathcal{A}^{\text{empt}}$ . Otherwise,  $\mathcal{D}_a^{\text{rchr}} = \emptyset$ .

We define a decision variable  $x_{tad}$  as the number of times we apply a decision  $d \in \mathcal{D}(a)$  to a vehicle  $a \in \mathcal{A}$  at time  $t$ . The decision vector  $x_t = (x_{tad})_{a \in \mathcal{A}, d \in \mathcal{D}(a)}$  must respect the constraints

$$\sum_{d \in \mathcal{D}(a)} x_{tad} = R_{ta} \quad \forall a \in \mathcal{A}, \quad (7a)$$

$$\sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a,b)} x_{tad} \leq D_{tb} \quad \forall b \in \mathcal{B}, \quad (7b)$$

$$x_{tad} \in \mathbb{Z}_{\geq 0} \quad \forall a \in \mathcal{A}, d \in \mathcal{D}(a). \quad (7c)$$

Constraint (7a) is a flow-balance constraint. Constraint (7b) ensures that the number of vehicles assigned to a travel request with attribute  $b \in \mathcal{B}$  does not exceed the number of available requests with such attribute. Constraint (7c) ensures that the number of vehicles  $a \in \mathcal{A}$  assigned to a decision  $d \in \mathcal{D}(a)$  is non-negative and integer. Constraints (7a), (7b), and (7c) define the feasible set  $\mathcal{X}(S_t)$ .

### 3.3 Rewards

The reward is given by the function

$$C(x_t) = \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} c_{tad} x_{tad}, \quad (8)$$

where the contribution  $c_{tad}$  is defined as follows. Assigning a vehicle  $a \in \mathcal{A}$  to serve a request  $b \in \mathcal{B}$  via a decision  $d \in \{d_b^{\text{sngl}}, d_b^{\text{queu}}\}$  at time  $t$  produces a contribution that is equal to the trip fare, i.e.,  $c_{tad} = f_b$ . Let  $\rho^{\text{detr}}(a, d) \in \mathbb{R}_{\leq 0}$  be the detour penalty incurred when a vehicle  $a$  is assigned to decision  $d \in \mathcal{D}(a)$ . The penalty is equal to zero if  $d$  does not involve any detour. Assigning a vehicle  $a \in \mathcal{A}$  to serve a request  $b \in \mathcal{B}$  on path  $p \in P_{a\{b\}}$  via decision  $d = d_{bp}^{\text{pool}}$  produces the contribution  $c_{tad} = f_b + \rho^{\text{detr}}(a, d)$ . Assigning a vehicle  $a \in \mathcal{A}$  to serve a set of requests  $B \in \mathcal{B}$  on a path  $p \in P_{aB}$  via decision  $d = d_{Bp}^{\text{milt}}$  produces the contribution

$$c_{tad} = \sum_{b \in B} f_b + \rho^{\text{detr}}(a, d) \quad (9)$$

Let  $\rho^{\text{rchr}}(a) \in \mathbb{R}_{\leq 0}$  be the cost of fully recharging a vehicle  $a \in \mathcal{A}$ , which depends on the vehicle's current range  $l_a$ . Assigning a vehicle  $a \in \mathcal{A}$  to a recharge decision  $d \in \mathcal{D}_a^{\text{rchr}}$  at time  $t$  produces the contribution  $c_{tad} = \rho^{\text{rchr}}(a)$ . Assigning a vehicle  $a \in \mathcal{A}$  to an idle, relocate, or continue decision  $d \in \mathcal{D}_a^{\text{idle}} \cup \mathcal{D}_a^{\text{relc}} \cup \mathcal{D}_a^{\text{cont}}$  at time  $t$  produces no contribution or penalty, i.e.,  $c_{tad} = 0$ .

### 3.4 Exogenous information

The exogenous information is the new information that the RHS receives between decision epochs  $t - 1$  and  $t$ . The exogenous information is given by the demand vector  $\hat{D}_t = (\hat{D}_{tb})_{b \in \mathcal{B}}$  where  $\hat{D}_{tb}$  is the number of travel requests with attribute  $b \in \mathcal{B}$  that first became known between time  $t - 1$  and  $t$ .

### 3.5 Transition functions

The transition functions specify how the system evolves from one state to the next. Let  $a' = a^M(a, d)$  be the future attribute of a vehicle  $a$  to which a decision  $d \in \mathcal{D}(a)$  is applied. The function  $a^M$  defines how the attribute of vehicle  $a$  changes after it is assigned to decision  $d$ . We define a *post-decision state* variable  $S_t^x$ . The post-decision state  $S_t^x$  is the state of the system after we make decision  $x_t$  but before any new information (i.e., exogenous demand  $\hat{D}_{t+1}$ ) has arrived. The post-decision state is defined as  $S_t^x = (R_t^x, D_t^x)$  where

- $R_t^x = (R_{ta'}^x)_{a' \in \mathcal{A}}$  and  $R_{ta'}^x = \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} x_{tad} \mathbf{1}_{\{a^M(a, d) = a'\}}$ .
- $D_t^x = (D_{tb}^x)_{b \in \mathcal{B}}$  and  $D_{tb}^x = \left( D_{tb} - \sum_{a \in \mathcal{A}} \sum_{d_b \in \mathcal{D}(a, b)} x_{tad_b} \right) \mathbf{1}_{\{t_b^{\text{res}} \geq t+1\}}$ .

We can obtain  $D_{tb}^x$  from the vector of slack variables of the demand constraints (7b) by setting every element that corresponds to a request  $b \in \mathcal{B}$  with  $t_b^{\text{res}} \leq t+1$  to 0. We now complete the definition of the transition function as

$$S_{t+1} = S^M(S_t, x_t, \hat{D}_{t+1}) = (R_t^x, D_t^x + \hat{D}_{t+1}) = (R_{t+1}, D_{t+1})$$

where  $R_{t+1}$  and  $D_{t+1}$  are the *resource* and *demand* vectors at time  $t+1$ , respectively. We refer the reader to Section A for a detailed definition of the function  $a^M(a, d)$ .

## 4 Policies

A policy is a function  $\pi : S_t \rightarrow \mathcal{X}(S_t)$  that maps the state  $S_t$  to a decision vector  $x_t \in \mathcal{X}(S_t)$ . Let  $\Pi$  be the set of policies. The optimal policy  $\pi^*$  maximizes the sum of expected rewards over the entire planning horizon, i.e.,

$$\pi^* = \arg \max_{\pi \in \Pi} \sum_{t=1}^T \mathbb{E} [\{C(\pi(S_t)) \mid S_0\}]. \quad (10)$$

### 4.1 Myopic policy

A myopic policy makes decisions by maximizing the reward function (8) subject to the constraints (7a)–(7c), i.e.,

$$\pi^{\text{myo}}(S_t) = \arg \max_{x_t \in \mathcal{X}(S_t)} \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} c_{tad} x_{tad}. \quad (11)$$

The myopic policy in (11) is suboptimal because it does not consider how a decision  $x_t$  affects the next state  $S_{t+1}$  the system will evolve into and, through that, all subsequent rewards. If the fleet of vehicles is electric, a simple way to improve  $\pi^{\text{myo}}$  is to introduce a threshold (parameter)  $\theta \in (0, 1]$  that forces the vehicle to charge only when its battery level is below the threshold. This can be done by modifying Problem (11) to become

$$\begin{aligned} \pi^{\text{PM}}(S_t) &= \arg \max_{x_t \in \mathcal{X}(S_t)} \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} c_{tad} x_{tad} \\ \text{s.t. } x_{tad} &= R_{ta} \cdot \mathbf{1}_{\{l_a < \theta l_a^{\text{max}}\}}, \quad \forall a \in \mathcal{A}, d \in \mathcal{D}^{\text{rchr}}(a). \end{aligned} \quad (12)$$

We refer to the policy in (12) as the PM policy.

### 4.2 Value function approximation policy

To capture the downstream impact of current decisions on future rewards, we use the well-known Bellman equation. We define the value function  $V(S_t)$  as the expected total reward from time  $t$  to  $T$  given that the system is in state  $S_t$  at time  $t$ , i.e.,

$$V(S_t) = \max_{x_t \in \mathcal{X}(S_t)} \{C(x_t) + \mathbb{E}[V(S_{t+1}) \mid S_t, x_t]\}, \quad (13)$$

for  $t \in \{1, \dots, T\}$  and  $V(S_{T+1}) = 0$ . A solution to the Bellman equation (13) gives the optimal policy in (10). Solving the Bellman equation (13) exactly is intractable. A common approach to address this challenge is to approximate the value function  $V(S_t)$  by a function that is easier to compute. We use a linear approximation (Simao et al., 2009) given by

$$\bar{V}^x(S_t^x) = \sum_{a \in \mathcal{A}} \bar{v}_a^R R_{ta}^x + \sum_{b \in \mathcal{B}} \bar{v}_{tb}^D D_{tb}^x \quad (14)$$

where

- $\bar{v}_a^R = \frac{\partial V(S_t)}{\partial R_{ta}}$  is the marginal value of a resource with attribute  $a \in \mathcal{A}$  at any time  $t \in \{1, \dots, T\}$ , and
- $\bar{v}_{tb}^D = \frac{\partial V(S_t)}{\partial D_{tb}}$  the marginal value of a demand with attribute  $b \in \mathcal{B}$  at time  $t$ .

Note that we omit the time index  $t$  in  $\bar{v}_a^R$  since the vehicle's actionable time is already encoded in the attribute  $a$ .

Substituting  $\bar{V}^x(S_t^x)$  into Equation (13), we obtain the VFA policy

$$\pi^{\text{VFA}}(S_t) = \arg \max_{x_t \in \mathcal{X}(S_t)} \left\{ \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} \left( c_{tad} + \bar{v}_{a^M(a,d)}^R \right) x_{tad} + \sum_{b \in \mathcal{B}} \bar{v}_{tb}^D D_{tb}^x \right\}. \quad (15)$$

We refer the reader to Section B for a detailed derivation of Equation (15).

### 4.3 Policies design and implementation

Solving the PM problem in (11) or the VFA problem in (15) is associated with several challenges that require careful design decisions. We discuss and address these challenges in the remainder of this section.

#### 4.3.1 Enumerating the multi-trip decisions

The presence of the multi-trip decision  $d^{\text{mlti}}$  leads to an exponential increase in the complexity of enumerating the elements of the decision set  $\mathcal{D}(a)$  for each  $a \in \mathcal{A}$ . Not only do we need to enumerate all possible subsets of requests  $B \subseteq \mathcal{B}$  that can be assigned to vehicle  $a$ , but we also need to consider the different feasible paths that a vehicle with attribute  $a$  can take to serve the requests in  $B$ .

To enumerate the subsets of requests that a vehicle  $a \in \mathcal{A}$  can serve, we use a modified version of the efficient *RTV graph algorithm* of Alonso-Mora et al. (2017). The original RTV graph algorithm creates subsets of requests  $B \subseteq \mathcal{B}$  that can be served by a virtual empty vehicle located at the same location as the origin of one of the requests in  $B$ , while satisfying all operational constraints (e.g., time windows, vehicle capacity, maximum waiting time). To limit computation, trips are constructed incrementally, starting with single-request trips and adding one request at a time until no more requests can be added without violating any operational constraint or exceeding a pre-specified maximum number of requests per trip. Then, the algorithm discards each subset of requests  $B$ , for which there exists no vehicle that can reach the origin of at least one request in  $B$  before its latest pickup time.

We modify the RTV graph algorithm by adding all constraints of our model. Our modified algorithm enumerates all feasible trips  $B \subseteq \mathcal{B}$  separately for each vehicle attribute  $a \in \mathcal{A}$ , again proceeding incrementally. Because of the range constraint, checking whether vehicle  $a$  can serve trip  $B$  requires solving a shortest path problem with resource constraints (SPPRC) (Irlich and Desaulniers, 2005). Solving this SPPRC yields a shortest path  $p^* \in P_{aB}$  that serves all requests in  $B$  while satisfying all operational constraints. This problem can be solved very efficiently using a labelling algorithm since we have no negative-weight arcs. If we assume that the detour penalty function  $\rho^{\text{detr}}$  is monotonically increasing in the path length, we can limit the number of paths  $|P_{aB}|$  that we need to consider for each subset of requests  $B$ . We formalize this assumption in the following definition.

**Definition 4.1** (Monotonicity of detour penalty). *The detour penalty function  $\rho^{\text{detr}}$  is monotonically increasing in the path length if, and only if, for all vehicle attributes  $a \in \mathcal{A}^{\text{empt}}$ , all sets of requests  $B \subseteq \mathcal{B}$  with  $|P_{aB}| > 0$ , and all pairs of paths  $p, p' \in P_{aB}$ , the following statement is true:*

$$\tau(p) \leq \tau(p') \implies \rho^{\text{detr}}(a, d_{Bp}^{\text{mlti}}) \leq \rho^{\text{detr}}(a, d_{Bp'}^{\text{mlti}}).$$

Since the PM policy does not consider future states, we only need to include one multi-trip decision  $d_{Bp^*}^{\text{mlti}}$  per feasible subset of requests  $B$  in the decision set  $\mathcal{D}(a)$  for each vehicle  $a \in \mathcal{A}$ . In this case,  $p^* \in P_{aB}$  is one of the shortest paths in  $P_{aB}$ . That is, since the PM policy is indifferent to the location, battery level, and time of arrival of vehicle  $a$  after serving the requests in  $B$ , we only need to include the decision that maximizes the immediate contribution  $c_{tad}$  given by Equation (9). Assuming that the detour penalty function  $\rho^{\text{detr}}$  is monotonically increasing in the path length, the shortest path  $p^* \in P_{aB}$  will yield the smallest detour penalty  $\rho^{\text{detr}}(a, d_{Bp^*}^{\text{mlti}})$  among all paths in  $P_{aB}$ , which in turn maximizes the contribution  $c_{tad}$ . This shortest path  $p^* \in P_{aB}$  is the solution of the SPPRC problem that we solve in our modified RTV graph algorithm.

The situation for the VFA policy is more involved because (15) maximizes not only the immediate contribution  $c_{tad}$  but also the downstream value captured by the term  $\bar{v}_{a^M(a,d)}^R$ . For example, a path  $p'$  that ends at an airport with high demand may be preferred over a shorter  $p^*$  that ends in a suburban neighborhood with little demand, even if  $p'$  takes

longer to complete and yields a lower immediate contribution  $c_{tad}$  than  $p^*$ . We show that enumerating all possible paths  $p \in P_{aB}$  for each subset of requests  $B \subseteq \mathcal{B}$  in the decision set  $\mathcal{D}(a)$  is not necessary to obtain an optimal solution to (15). Instead, for each subset of requests  $B \subseteq \mathcal{B}$ , it suffices to consider at most  $|B|$  paths from  $P_{aB}$ . Specifically, for each subset of requests  $B$ , we need to consider at most one path  $p_b^* \in P_{aB}$  per each request  $b \in B$ , which is one of the shortest paths in  $P_{aB}$  that end at the request destination  $d_b$ . Fortunately, without requiring any additional enumeration, the labeling algorithm we use to solve the SPPRC yields a set  $\{p_b^* \in P_{aB} \mid b \in B\}$ , which contains all paths that we need to consider. We formalize this result in the following definitions and propositions.

**Definition 4.2** (Sufficient paths set). *Let  $\pi$  be a policy. For all time steps  $t \in \{1, \dots, T\}$  and all states  $S_t$  where a vehicle attribute  $a \in \mathcal{A}^{\text{empt}}$  satisfies  $R_{ta} > 0$  and where the nonempty set of requests  $B \subseteq \mathcal{B}$  satisfies  $\forall b \in B. D_{tb} > 0$  and  $|P_{aB}| > 0$ , the set of paths  $P_{aB}^* \subseteq P_{aB}$  is a-B- $\pi$ -sufficient if, and only if, there exists a solution  $x_t$  in the set of optimal solutions  $\pi(S_t)$  such that all decision variables  $x_{tad}$  with  $d \in \{d_{Bp}^{\text{milti}} \in \mathcal{D}_a^{\text{milti}} \mid p \in P_{aB} \setminus P_{aB}^*\}$  are equal to zero.*

**Definition 4.3** (Last dropoff shortest paths set (LDSPS)). *For all time steps  $t \in \{1, \dots, T\}$  and all states  $S_t$  where a vehicle attribute  $a \in \mathcal{A}^{\text{empt}}$  satisfies  $R_{ta} > 0$  and where the nonempty set of requests  $B \subseteq \mathcal{B}$  satisfies  $\forall b \in B. D_{tb} > 0$  and  $|P_{aB}| > 0$ , the set of paths  $P_{aB}^* \subseteq P_{aB}$  is an LDSPS for  $(a, B)$  if, and only if the following statement is true for all requests  $b \in B$ : Either  $P_{aB}^*$  contains exactly one of the shortest paths in  $P_{aB}$  that end in the destination of  $b$ ,  $d_b$ , or no path in  $P_{aB}$  ends in  $d_b$ .*

**Definition 4.4** (Partial Order). *Let  $a, a' \in \mathcal{A}$  with  $a = (o_a \ d_a \ l_a \ n_a \ t_a)^\top$  and  $a' = (o_{a'} \ d_{a'} \ l_{a'} \ n_{a'} \ t_{a'})^\top$ . Then,  $a \preccurlyeq a'$  if, and only if,  $(o_a, d_a) = (o_{a'}, d_{a'})$ ,  $l_a \leq l_{a'}$ ,  $n_a \leq n_{a'}$  and  $t_a \geq t_{a'}$ .*

Definition 4.4 allows us to adopt and extend the monotonicity results of Al-Kanj et al. (2020) to our setting in the following proposition.

**Proposition 4.1** (Monotonicity). *For every pair of vehicle attributes  $a, a' \in \mathcal{A}$  with  $a \preccurlyeq a'$ , if the statements  $a^M(a, d) \preccurlyeq a^M(a', d)$  and  $c_{tad} \leq c_{ta'd}$  are true for all decisions  $d \in \mathcal{D}(a) \cap \mathcal{D}(a')$  and time steps  $t \in \{1, \dots, T\}$ , then the following monotonic properties hold:*

1.  $v_a^{R*}$  increases monotonically with the  $l_a$  dimension:  $l_a < l_{a'} \implies v_a^{R*} \leq v_{a'}^{R*}$ .
2.  $v_a^{R*}$  increases monotonically with the  $n_a$  dimension:  $n_a < n_{a'} \implies v_a^{R*} \leq v_{a'}^{R*}$ .
3.  $v_a^{R*}$  decreases monotonically with the  $t_a$  dimension:  $t_a > t_{a'} \implies v_a^{R*} \leq v_{a'}^{R*}$ .

*Proof.* See Section B. □

**Proposition 4.2** (Sufficiency of the LDSPS). *For all time steps  $t \in \{1, \dots, T\}$  and all states  $S_t$  where a vehicle attribute  $a \in \mathcal{A}^{\text{empt}}$  satisfies  $R_{ta} > 0$  and where the nonempty set of requests  $B \subseteq \mathcal{B}$  satisfies  $\forall b \in B. D_{tb} > 0$  and  $|P_{aB}| > 0$ , if the approximate value function used by policy  $\pi^{\text{VFA}}$  satisfies Proposition 4.1 and  $\rho^{\text{detr}}$  is monotonically increasing in the path length, all LDSPSs for  $(a, B)$  are a-B- $\pi^{\text{VFA}}$ -sufficient.*

*Proof.* See Section C. □

### 4.3.2 Obtaining the marginal attribute values

To implement the VFA policy in (15), we need to know the marginal attribute values  $\bar{v}_a^R$  for every  $a \in \mathcal{A}$  and  $\bar{v}_{tb}^D$  for every  $b \in \mathcal{B}$  and  $t \in \{1, \dots, T\}$ . In practice, these values are unknown and need to be estimated. We estimate them using a Forward Approximate Dynamic Programming (ADP) algorithm that starts with initial values  $\bar{v}_a^{R,0}$  and  $\bar{v}_{tb}^{D,0}$  and iteratively improves these estimates by simulating the RHS for  $N$  iterations. In the  $n$ -th iteration, we obtain the decision  $x_t^n$  and the estimates  $\hat{v}_{ta}^{R,n}$  and  $\hat{v}_{tb}^{D,n}$  of the slope of the objective function with respect to  $R_{ta}$  and  $D_{tb}$ , respectively, at the points  $R_{ta}^x$  and  $D_{tb}^x$ , respectively. We then update the current marginal attribute value estimates  $\bar{v}_a^{R,n-1}$  and  $\bar{v}_{tb}^{D,n-1}$  to obtain new estimates  $\bar{v}_a^{R,n}$  and  $\bar{v}_{tb}^{D,n}$ . The last step of the iteration is to transition to the next state using  $x_t^n$  and a random demand realization. As the final result of the algorithm, we obtain the values  $\bar{v}_a^{R,N}$  for every  $a \in \mathcal{A}$  and  $\bar{v}_{tb}^{D,N}$  for every  $b \in \mathcal{B}$  and  $t \in \{1, \dots, T\}$ . We then use these values in (15) to obtain our VFA policy. Note that we retain the time index  $t$  in  $\hat{v}_{ta}^{R,n}$  to emphasize that it is an estimate of the objective function's slope at the point  $R_{ta}^x$  at time  $t$  in iteration  $n$ . Section B gives a detailed description of the algorithm.

To use the algorithm, we need a surrogate policy that allows us to obtain the estimates  $\hat{v}_{ta}^{R,n}$  and  $\hat{v}_{tb}^{D,n}$  in the  $n$ -th iteration. If the feasible region  $\mathcal{X}$  of (15) has a totally unimodular coefficient matrix and an integer right-hand side, we can relax the integrality constraints (7c) and solve (15) as an LP. Doing so allows us to obtain the marginal values

$\hat{v}_{ta}^{R,n}$  and  $\hat{v}_{tb}^{D,n}$  directly from the dual variables associated with the resource constraints (7a) and demand constraints (7b), respectively. To obtain a feasible region  $\mathcal{X}^{\text{LP}}$  that has a totally unimodular coefficient matrix, we remove the multi-trip decision  $d^{\text{mlti}}$  from  $\mathcal{X}$ . We only use  $\mathcal{X}^{\text{LP}}$  in the surrogate policy, and we only use the surrogate policy to obtain  $\bar{v}_a^{R,N}$  and  $\bar{v}_{tb}^{D,N}$ . Once we have these values, we can use the final VFA policy that solves (15) with the multi-trip decisions. That is, we include the multi-trip decisions in the final VFA policy but not in the surrogate policy used in the ADP procedure.

### 4.3.3 Exploitation versus exploration trade-off

ADP faces the classic *exploration vs. exploitation trade-off*. On the one hand, we want to *explore* different decisions to visit a diverse set of states and learn their values. On the other hand, we want to *exploit* the current estimates of the value function to make meaningful progress in the learning process. For instance, without exploration, the policy may get stuck in a local optimum that allocates all vehicles to the same location repeatedly because it only knows the values of that location and never visits other locations. Conversely, without exploitation, the policy may wander randomly across the state space without making meaningful progress in learning the values of the states.

A standard approach to address this trade-off is to introduce some randomness in the decision-making process to encourage exploration. We follow this approach by adding some randomness to the relocate decisions made by the VFA policy in (15). Specifically, after solving Problem (15) in each decision epoch  $t$ , we post-process the chosen relocate destinations, as follows. We consider the subset of attributes for which at least one relocate decision variable is non-zero, i.e.,  $\mathcal{A}^{\text{relc}} = \{a \in \mathcal{A} \mid r_a \geq 0\}$ , where  $r_a = \sum_{d \in \mathcal{D}_a^{\text{rel}}} x_{tad}$ . For each attribute  $a \in \mathcal{A}^{\text{relc}}$ , we draw, with replacement,  $r_a$  random samples from the set of feasible relocate destinations of  $a$ . In the  $n$ -th iteration, the probability of drawing a destination  $v \in \mathcal{V}$  is proportional to the value  $\bar{v}_{a'}^{R,n}$  of the post-decision state  $a' = a^M(a, d_v^{\text{relc}})$ .

Another approach to address the exploration vs. exploitation trade-off is to leverage the monotonicity properties introduced in Proposition 4.1 to update some of the  $\bar{v}^{R,n}$  values during the ADP procedure. For instance, suppose we find that our current estimate of the marginal attribute value  $\bar{v}_a^{R,n} > \bar{v}_{a'}^{R,n}$  for two attributes  $a, a' \in \mathcal{A}$ , even though  $a \preccurlyeq a'$ . If we leave these estimates as-is, the policy may wrongly exploit the supposedly higher value of attribute  $a$  by more frequently choosing decisions whose post-decision state is  $a$ . We can correct this inconsistency by setting  $\bar{v}_{ta'}^{R,n} = \bar{v}_{ta}^{R,n}$ . This correction encourages exploration by increasing the value of attribute  $a'$ , which may lead the policy to more frequently choose decisions whose post-decision state is  $a'$ , thereby exploring states associated with  $a'$ .

## 5 Numerical Experiments

In this section, we provide numerical results and analyze the impact of different features, decisions, and constraints on the performance of the RHS. We start by describing the instances we use (Section 5.1) and the problem settings (Section 5.2). We then present the results of our numerical experiments (Section 5.3). We refer the reader to Section D for implementation details.

### 5.1 Benchmark instances

Benchmark instances from the literature are either unpublished or do not cover the full range of problem settings that we consider. Furthermore, most existing benchmarks consider only one area and demand distribution. Ulmer et al. (2018) and Yu and Shen (2020) show that spatial differences in the demand distribution induce differences in the relative performance of policies. A comprehensive numerical study should, therefore, include multiple areas and demand distributions. We create a set of instances. Upon publication of this paper, we will publish the instances, along with the code we used to create them, as readily usable files on Zenodo.

A popular choice for the street network used in the literature is New York City (NYC). We create four street networks using data published by the NYC Taxi and Limousine Commission (TLC). The data comprises real taxi trips from the years 2009 and 2018. Two of our street networks represent the Manhattan borough alone, while the other two represent the four eastern boroughs of NYC, which are Manhattan, The Bronx, Brooklyn, and Queens. We use a fine-grained representation of the street networks with the 2009 TLC data, and a coarse-grained representation with the 2018 TLC data. We associate each street network with a demand distribution. Each demand distribution is represented by a set of sample paths, where each sample path consists of a sequence of travel requests that occur between time  $t = 0$  and  $t = T$ . We refer to every combination of one of the four street networks and its corresponding demand distribution as an instance. Table 1 summarizes the characteristics of each instance. We refer the reader to Section D.1 for a detailed description of how we create the instances.

Instance	TLC data year	Boroughs	$ \mathcal{V} $	$ A $	# of Vehicles	Avg. # of Requests (per sample path)
1	2009	Manhattan	923	3581	6400	395605
2		four eastern	9017	37511	8000	424636
3	2018	Manhattan	64	319	4400	240536
4		four eastern	231	1280	6250	282739

Table 1: Characteristics of the benchmark instances.

## 5.2 Problem settings and parameters

We consider six different problem settings per instance in a full factorial design. Factors are whether the system is pooling-enabled (i.e., whether  $d^{\text{pool}}$  and  $d^{\text{multi}}$  are allowed), and whether the fleet consists of ICE vehicles, EVs that use direct current fast charging (DCFC), or EVs that use slower level 2 charging (L2C).

We use the following vehicle parameters. All types of vehicles can carry up to four passengers at a time. The time it takes to fill the tank or charge the battery (i.e., execute  $d^{\text{rchr}}$ ) is  $\ell(l_a) = (l^{\max} - l_a)\dot{\ell} + 15$  min. The slope  $\dot{\ell}$  depends on the type of vehicle, whereas the intercept is independent of the type of vehicle and accounts for driving to the nearest gas station or charger, queuing, etc. Fully-fueled ICE vehicles can drive for  $l^{\max} = 26$  h and completely filling their tank requires 1 min, i.e.,  $\dot{\ell} = 2.308 \text{ s/h}$ . For EVs, we use the EV parameters from Al-Kanj et al. (2020) and convert them to durations, assuming that vehicles either stay at rest or move at the average NYC taxi speed, 18.2 km/h according to TLC data. We use this constant speed only for determining  $\dot{\ell}$ , not for simulating driving durations (cf. Section D). This results in a maximum range of  $l^{\max} = 17$  h 41 min and a recharging rate of  $\dot{\ell} = 2.262 \text{ min/h}$  for DCFC vehicles. We assume that L2C vehicles have the same range but recharge 20 times slower, i.e.,  $\dot{\ell} = 45.23 \text{ min/h}$ .

We implement one PM policy and one VFA policy. For each instance, we create 3000 training and 30 test sample paths, along with an initial state for each path. To enable a fair comparison, we reuse the same sample paths and initial states across policies and problem settings.

We use a planning horizon of  $T = 24$  h, where the RHS operator makes a decision every two minutes. This means that the planning horizon consists of 720 decision epochs. We assume energy prices and detour penalties are constant and negligible, i.e.,  $\rho^{\text{rchr}} = \rho^{\text{detr}} = \$0$ . We consider multi-trip decisions up to  $|B| \leq 2$ . This is inspired by Alonso-Mora et al. (2017), who report that the average number of passengers per vehicle rarely exceeds two and increased seat capacities yield diminishing returns.

## 5.3 Results

We report the reward fulfillment ratio (RFR), which we define as the ratio between the reward obtained by a policy and the sum of the rewards of all requests in the sample path. Figure 2 shows kernel density plots of the RFR obtained by the PM and VFA policies across different instances and problem settings. Section D.4 contains tables of reward and RFR statistics for each instance, problem setting, and policy.

**Result 1: The VFA policy performs significantly better than the PM policy.** The median RFR (interquartile range (IQR)) of the PM and VFA policies is 76.8% (24.2%) and 88.4% (4.9%), respectively. Figure 2 shows that the difference in performance cannot be fully explained by controlling for instance or problem characteristics. Nevertheless, the difference is more pronounced in instances 3 and 4, and when using L2C EVs. The advantage of the VFA policy is consistent with the literature. Since the PM policy is clearly suboptimal, we ignore it for the remainder of this analysis.

**Result 2: Both ICE vehicles and DCFC EVs achieve the highest RFR, while L2C EVs perform worse.** The median RFR (IQR) of ICE and DCFC vehicles is 89.2% (3.7%) and 89.6% (2.3%), respectively. By contrast, the median RFR (IQR) of L2C vehicles is 84.3% (5.7%). This is illustrated in the fleet types subplot in Figure 2. The relatively weak performance of L2C vehicles is expected, because one would expect the vehicle range and the recharging or refueling rate  $\dot{\ell}$  to have a positive correlation with the RFR. Interestingly, the performance of DCFC vehicles is on par with ICE vehicles despite the significant differences in their range and  $\dot{\ell}$  values. One possible explanation for this is that the comparatively small distances of metropolitan areas such as NYC remove the significance of the vehicle range.

**Result 3: Pooling enables the RHS to achieve high rewards more consistently.** When pooling is not enabled, the median RFR is 86.4%, which improves to 89.2% with pooling. More noteworthy than the improvement in the median

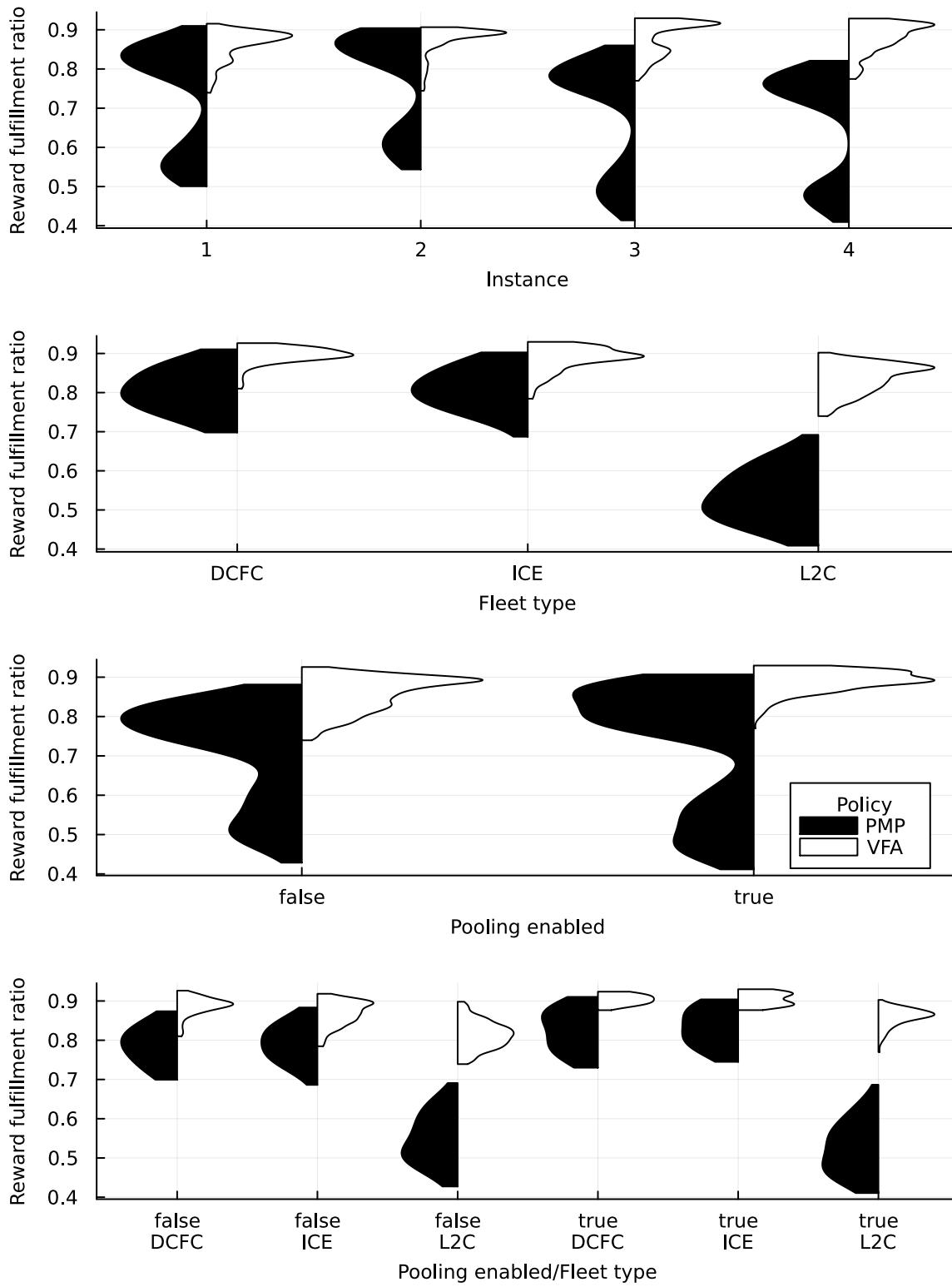


Figure 2: Kernel density plots of the RFR obtained by the PM and VFA policies across different instances and problem settings.

RFR is the change in the shape of the RFR distribution, as visualized in the third subplot from the top in Figure 2. The change in the shape of the distribution is associated with a decrease in the IQR from 6.9% to 3.9%. When we focus on the DCFC and ICE settings (see Result 4 below for a detailed analysis of the L2C setting), the bottom subplot shows that pooling makes their RFR distributions especially consistent, without significantly affecting their best-case RFR. One possible explanation for this is the following. Without pooling, the VFA policy already achieves the maximum RFR in some “easy” sample paths, and this maximum is around 90% RFR. The remaining requests in the gap to 100% RFR are not profitable because their reward is low, they take very long, or they connect low-demand areas. As a result, enabling pooling has no effect on an “easy” sample path. However, there are other, more difficult sample paths where the policy needs pooling to achieve maximum RFR.

**Result 4: Pooling alleviates the disadvantage of L2C vehicles.** Pooling improves the median RFR (IQR) of L2C fleets from 81.3% (5.0%) to 86.4% (2.3%). By contrast, the median RFR (IQR) of DCFC vehicles merely improves from 89.1% (2.4%) to 90.4% (2.1%), with similar numbers for ICE vehicles. Figure 3 illustrates the influence of pooling on L2C fleets. Comparing the top subplots shows that the pooling (right) policy executes multi-trip and pool decisions, which the non-pooling (left) policy does not. However, this is not the only difference: The pooling policy also executes significantly more relocate decisions. Comparing the middle subplots shows that towards the end of the day, the pooling (right) fleet’s average range is higher than the non-pooling (left) fleet’s average range, even though both fleets have similar recharging behavior. The bottom subplots show that the pooling (left) fleet is able to transport much more passengers in the high-demand hours between 6 p.m. and midnight than the non-pooling (right) fleet. The data suggests that the pooling decisions allow the RHS to serve the same passengers with fewer vehicles and relocate the others to more promising locations, while using the range of all vehicles more efficiently. The effect of this on the total reward is most noticeable when the need to make efficient use of the vehicles’ range is highest. This makes the L2C setting stand out.

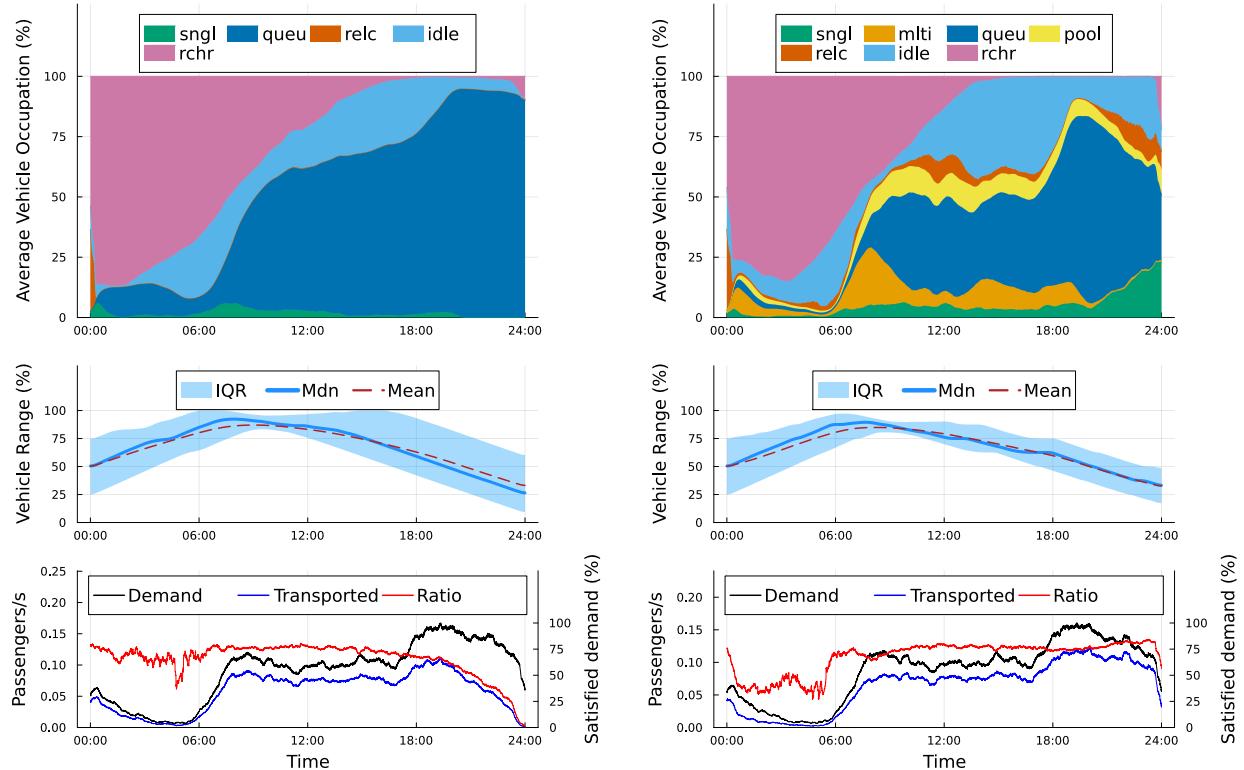


Figure 3: Behavior of the VFA policy in instance 1, using L2C vehicles, without (left) and with pooling (right), averaged across 30 sample paths.

## 6 Conclusion

Optimal control of ride-hailing systems is a challenging problem that involves several interdependent decisions, which must be made in a dynamic and stochastic environment while adhering to various operational constraints. Although the problem has been studied extensively in the literature, even in comparable settings, studies differ significantly in the types of problem features, decisions, constraints, and performance metrics they consider. This lack of consistency makes it difficult to draw generalizable conclusions and undermines the practical relevance of the findings. This is exacerbated by the fact that many studies do not publish their benchmark instances, making it hard to replicate or build upon their findings.

To address this gap, we proposed a modular, unified modeling framework that integrates various problem features, including pooling, repositioning, EVs and ICEs fleets, as well as various constraints regarding customer response time, waiting time, and detour time. To solve the problem, we designed a PM policy that charges (or fuels) vehicles when their batteries (or fuel tanks) are below a certain threshold, and a VFA policy that captures the long-term impact of current decisions. Both policies utilize an efficient procedure for enumerating all feasible vehicle-to-request assignments. Furthermore, we introduced a scalable technique that introduces randomness in the decision-making process, as well as monotonicity properties of the value function, to effectively address the exploration-exploitation tradeoff in ADP. We also created reusable benchmark instances that capture a range of spatial structures and demand distributions, based on real-world data from NYC. Our computational results provide insights into previously unstudied interactions between problem characteristics. Specifically, we found no significant difference between revenue generated by ICE fleets and fast-charging EV fleets, but both significantly outperformed slow-charging EV fleets. We also found that pooling increases the revenue, and reduces revenue variability, for all fleet types.

Our unified modeling framework can serve as a foundation for several promising future research directions. Relevant extensions include the integration of pre-scheduled trips (as seen in airport shuttles or paratransit services), heterogeneous fleets (with varying seat capacities or energy sources), stochastic travel times and congestion, uncertain passenger headcounts and destinations that are only revealed upon pickup, infrastructure limitations (e.g., scarce parking at key locations and queuing at charging or fueling stations), and explicit modeling of customers' varying preferences (e.g., willingness to pool and tolerance for detours). On the solution side, future work may benefit from spatial aggregation techniques that enhances scalability without compromising solution quality. Additionally, exploring alternative policy architectures, such as regression-based value function approximations, direct lookahead policies, and more expressive representations using deep learning models, including large-scale transformers, all present promising avenues for further investigation.

## References

- Ackermann, C. and Rieck, J. (2023). A novel repositioning approach and analysis for dynamic ride-hailing problems. *EURO Journal on Transportation and Logistics*, 12:100109.
- Agatz, N., Erera, A., Savelsbergh, M., and Wang, X. (2012). Optimization for dynamic ride-sharing: A review. *European Journal of Operational Research*, 223(2):295–303.
- Al-Kanj, L., Nascimento, J., and Powell, W. B. (2020). Approximate dynamic programming for planning a ride-hailing system using autonomous fleets of electric vehicles. *European Journal of Operational Research*, 284(3):1088–1106.
- Alonso-Mora, J., Samaranayake, S., Wallar, A., Fazzoli, E., and Rus, D. (2017). On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, 114(3):462–467.
- Ben-Gal, S. and Tzur, M. (2025). Data-driven policies for the online ride-hailing problem with fairness. *Transportation Science*, 0(0). Published online: March 26, 2025.
- Boyaci, B., Zografos, K. G., and Geroliminis, N. (2015). An optimization framework for the development of efficient one-way car-sharing systems. *European Journal of Operational Research*, 240(3):718–733.
- Braverman, A., Dai, J. G., Liu, X., and Ying, L. (2019). Empty-car routing in ridesharing systems. *Operations Research*, 67(5):1437–1452.
- Dunning, I., Huchette, J., and Lubin, M. (2017). Jump: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320.
- Erdmann, M., Dandl, F., and Bogenberger, K. (2021). Combining immediate customer responses and car-passenger reassignments in on-demand mobility services. *Transportation Research Part C: Emerging Technologies*, 126:103104.
- Gurobi Optimization, LLC (2024). Gurobi Optimizer Reference Manual.

- Heitmann, R.-J. O., Soeffker, N., Klawonn, F., Ulmer, M. W., and Mattfeld, D. C. (2024). Accelerating value function approximations for dynamic dial-a-ride problems via dimensionality reductions. *Computers & Operations Research*, 167:106639.
- Heitmann, R.-J. O., Soeffker, N., Ulmer, M. W., and Mattfeld, D. C. (2023). Combining value function approximation and multiple scenario approach for the effective management of ride-hailing services. *EURO Journal on Transportation and Logistics*, 12:100104.
- Hildebrandt, F. D., Thomas, B. W., and Ulmer, M. W. (2023). Opportunities for reinforcement learning in stochastic dynamic vehicle routing. *Computers & Operations Research*, 150:106071.
- Irnich, S. and Desaulniers, G. (2005). Shortest path problems with resource constraints. In Desaulniers, G., Desrosiers, J., and Solomon, M. M., editors, *Column Generation*, pages 33–65. Springer.
- Kullman, N. D., Cousineau, M., Goodson, J. C., and Mendoza, J. E. (2022). Dynamic ride-hailing with electric vehicles. *Transportation Science*, 56(3):775–794.
- Li, B., Krushinsky, D., Reijers, H. A., and Van Woensel, T. (2014). The share-a-ride problem: People and parcels sharing taxis. *European Journal of Operational Research*, 238(1):31–40.
- Lyu, G., Cheung, W. C., Teo, C.-P., and Wang, H. (2024). Multiobjective stochastic optimization: A case of real-time matching in ride-sourcing markets. *Manufacturing & Service Operations Management*, 26(2):500–518.
- Mourad, A., Puchinger, J., and Chu, C. (2019). A survey of models and algorithms for optimizing shared mobility. *Transportation Research Part B: Methodological*, 123:323–346.
- New York City Department of Transportation (2018). New York City Mobility Report.
- Özkan, E. and Ward, A. R. (2020). Dynamic matching for real-time ride sharing. *Stochastic Systems*, 10(1):29–70.
- Powell, W. B. (2011). *Approximate dynamic programming: Solving the curses of dimensionality*. Wiley, second edition.
- Powell, W. B. (2022). *Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions*. Wiley.
- Qin, Z. T., Zhu, H., and Ye, J. (2022). Reinforcement learning for ridesharing: An extended survey. *Transportation Research Part C: Emerging Technologies*, 144:103852.
- Schneider, T. W. (2018). Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance.
- Simao, H. P., Day, J., George, A. P., Gifford, T., Nienow, J., and Powell, W. B. (2009). An approximate dynamic programming algorithm for large-scale fleet management: A case application. *Transportation Science*, 43(2):178–197.
- Soeffker, N., Ulmer, M. W., and Mattfeld, D. C. (2022). Stochastic dynamic vehicle routing in the light of prescriptive analytics: A review. *European Journal of Operational Research*, 298(3):801–820.
- Tuncel, K., Koutsopoulos, H. N., and Ma, Z. (2023). An integrated ride-matching and vehicle-rebalancing model for shared mobility on-demand services. *Computers & Operations Research*, 159:106317.
- Ulmer, M. W., Goodson, J. C., Mattfeld, D. C., and Hennig, M. (2018). Offline–Online Approximate Dynamic Programming for Dynamic Vehicle Routing with Stochastic Requests. *Transportation Science*.
- Wang, Z., Qin, Z., Tang, X., Ye, J., and Zhu, H. (2018). Deep reinforcement learning with knowledge transfer for online rides order dispatching. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 617–626. IEEE.
- Yan, C., Zhu, H., Korolko, N., and Woodard, D. (2020). Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics*, 67(8):705–724.
- Yu, X. and Shen, S. (2020). An integrated decomposition and approximate dynamic programming approach for on-demand ride pooling. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3811–3820.
- Yu, X., Zhu, Z., Mao, H., Hua, M., Li, D., Chen, J., and Xu, H. (2023). Coordinating matching, rebalancing and charging of electric ride-hailing fleet under hybrid requests. *Transportation Research Part D: Transport and Environment*, 123:103903.

This appendix is structured as follows. Section A defines the transition function of our model. Section B explains how to derive and implement our VFA policy, and establishes some definitions. Section C gives a proof of correctness of the multi-trip decision enumeration algorithm that we use in our VFA policy. Section D contains implementation details and additional results of our numerical studies.

## A Definition of the transition function

To define  $a^M(a, d)$ , we first introduce the following notation:

- $\tau(v, w)$  is the time it takes to traverse the shortest path from node  $u \in \mathcal{V}$  to node  $v \in \mathcal{V}$ .
- $\tau(p)$  is the time it takes to traverse the entire path  $p \in P$ .
- $\sigma(u, v)$  is the driving range required to travel the shortest path from node  $u \in \mathcal{V}$  to node  $v \in \mathcal{V}$ .
- $\sigma(p)$  is the driving range required to traverse path  $p \in P$ .
- $p_{\text{end}}$  is the last node in the path  $p \in P$ .
- $\ell(l_a)$  is the time it takes a vehicle  $a$  to recharge its battery level from  $l_a$  to  $l^{\max}$ .
- $f_a$  is the first node on the shortest path from  $o_a$  to  $d_a$  that vehicle  $a \in \mathcal{A}$  reaches after  $\lfloor t_a \rfloor + 1$ .
- $m_a(h) = \max \{t_a + h, \lfloor t_a \rfloor + 1\}$ , where  $h$  is a time span.

The function  $a^M(a, d)$  is defined for every attribute  $a \in \mathcal{A}$ , decision  $d \in \mathcal{D}(a)$ , and decision epoch  $t \in \{1, \dots, T\}$ . Below we specify the definition of  $a^M(a, d)$  for every type of decision  $d \in \mathcal{D}(a)$ .

**Continue:** A vehicle  $a \in \mathcal{A}^{\text{occu}}$  that gets assigned to a decision  $d^{\text{cont}} \in \mathcal{D}_a^{\text{cont}}$  at time  $t$  spends the time between its actionable time  $t_a$  and the next decision epoch  $t + 1$  driving on the shortest path from its current location  $o_a$  to its destination  $d_a$ . If  $t_a \geq t + 1$ , the future attribute is equal to  $a$ . If  $t_a > t + 1$  and the vehicle reaches  $d_a$  before time  $\lfloor t_a \rfloor + 1$ , it parks until then. Otherwise, it becomes actionable in the first location between  $o_b$  and  $d_b$  that it reaches after  $\lfloor t_a \rfloor + 1$ . The future attribute is

- $a^M(a, d^{\text{cont}}) = a$  if  $t_a \geq t + 1$ ,
- $a^M(a, d^{\text{cont}}) = (d_a, d_a, l_a - \sigma(o_a, d_a), n_{\max}, \lfloor t_a \rfloor + 1)^T$  if  $t_a < t + 1 \wedge t_a + \tau(o_a, d_a) \leq \lfloor t_a \rfloor + 1$ , and
- $a^M(a, d^{\text{cont}}) = (f_a, d_a, l_a - \sigma(o_a, f_a), n_a, t_a + \tau(o_a, f_a))^T$  otherwise.

**Single trip:** A vehicle  $a \in \mathcal{A}^{\text{empt}}$  that gets assigned to a decision  $d_b^{\text{sngl}} \in \mathcal{D}_a^{\text{sngl}}$  to serve a request  $b \in \mathcal{B}$  starts moving from its current location  $o_a$  to the origin  $o_b$  to pickup the passengers. After pickup, the vehicle acts like under a *continue* decision. Its future attribute is  $a^M(a, d_b^{\text{sngl}}) = a^M(a', d^{\text{cont}})$ , where

$$a' = \begin{pmatrix} o_b \\ d_b \\ l_a - \sigma(o_a, o_b) \\ n_{\max} - n_b \\ t_a + \tau(o_b, d_b) \end{pmatrix}$$

**Multi-trip and pool:** A vehicle  $a \in \mathcal{A}$  that gets assigned to a decision  $d_{Bp}^{\text{mlti}} \in \mathcal{D}_a^{\text{mlti}}$  or  $d_{bp}^{\text{pool}} \in \mathcal{D}_a^{\text{pool}}$  must drive the entire path  $p$  to serve requests  $B$ , or  $b$ , respectively. If this does not take until the next decision epoch, the vehicle idles at the last drop off location until then. The future attribute of these vehicles is  $a^M(a, d_{Bp}^{\text{mlti}}) = a^M(a, d_{bp}^{\text{pool}}) = (p_{\text{end}}, p_{\text{end}}, l_a - \sigma(p), n_{\max}, m_a(\tau(p)))^T$ .

**Queue:** A vehicle  $a \in \mathcal{A}^{\text{occu}}$  that gets assigned to a decision  $d_b^{\text{queu}} \in \mathcal{D}_a^{\text{queu}}$  drives to  $d_a$ , drops off its passengers, then drives directly to the origin  $o_b$  of request  $b \in \mathcal{B}$  and picks up the new passengers. After pickup, the vehicle acts like under a *continue* decision. Its future attribute is  $a^M(a, d_b^{\text{queu}}) = a^M(a', d^{\text{cont}})$ , where

$$a' = \begin{pmatrix} o_b \\ d_b \\ l_a - \sigma(o_a, d_a) - \sigma(d_a, o_b) \\ n_{\max} - n_b \\ t_a + \tau(o_a, d_a) + \tau(d_a, o_b) \end{pmatrix}$$

**Relocate and recharge:** The future attribute of a vehicle  $a \in \mathcal{A}^{\text{empt}}$  that gets assigned to a relocate or recharge decision is  $a^M(a, d_v^{\text{relc}}) = (v, v, l_a - \sigma(o_a, v), n^{\max}, t + 1)^T$  and  $a^M(a, d_v^{\text{relc}}) = (o_a, o_a, l^{\max}, n^{\max}, m_a(\ell(l_a)))^T$ , respectively.

**Idle:** A vehicle  $a \in \mathcal{A}^{\text{empt}}$  that gets assigned to an idle decision parks in its current location until the next decision epoch. If the actionable time  $t_a \geq t + 1$ , the future attribute stays the same. This allows  $t$  to catch up to  $t_a$ . The future attribute of such a vehicle is  $a^M(a, d^{\text{idle}}) = a$  if  $t_a \geq t + 1$ , and  $a^M(a, d^{\text{idle}}) = (o_a, d_a, l_a, n_{\max}, \lfloor t_a \rfloor + 1)^T$  otherwise.

## B Policies Supplement

This section is structured as follows. Section B.1 shows the derivation of the Simao et al. (2009) linear VFA in the context of our model. Section B.2 specifies the details of the Forward ADP algorithm. Section B.3 discusses the monotonicity properties we use to learn our approximate value function.

### B.1 Derivation of linearized value function

The value function is defined recursively as

$$V(S_t) = \max_{x_t \in \mathcal{X}(S_t)} \{C(x_t) + \mathbb{E}[V(S_{t+1}) | S_t, x_t]\}, \quad (16)$$

for  $t \in \{1, \dots, T\}$  and  $V(S_{T+1}) = 0$ .

The first difficulty in dealing with the Bellman equation (16) is the expectation term, which is taken with respect to the random exogenous demand  $\hat{D}_{t+1}$ . We can move the expectation outside the max operator by using the post-decision state  $S_t^x$  and replacing  $\mathbb{E}[V(S_{t+1}) | S_t, x_t]$  with  $V^x(S_t^x)$  in (16), where  $V^x(S_t^x) = \mathbb{E}[V(S_{t+1}) | S_t^x]$ . Readers familiar with other reinforcement learning algorithms may find it helpful to note that approximating  $V^x(S_t^x)$  is essentially the same as Q-learning, except that we estimate the value of post-decision states rather than state-action pairs.

The second difficulty in dealing with the Bellman equation (16) is the exponentially large number of possible states, which makes it intractable to obtain the exact value of  $V^x(S_t^x)$ . A typical approach to deal with this difficulty is to approximate  $V^x(S_t^x)$  by a function  $\bar{V}^x(S_t^x)$  that is easier to compute. We use a linear approximation (Simao et al., 2009) given by

$$\bar{V}^x(S_t^x) = \sum_{a \in \mathcal{A}} \bar{v}_a^R R_{ta}^x + \sum_{b \in \mathcal{B}} \bar{v}_{tb}^D D_{tb}^x \quad (17)$$

where

- $\bar{v}_a^R = \frac{\partial V(S_t)}{\partial R_{ta}}$  is the marginal value of a resource with attribute  $a \in \mathcal{A}$  at any time  $t \in \{1, \dots, T\}$ , and
- $\bar{v}_{tb}^D = \frac{\partial V(S_t)}{\partial D_{tb}}$  the marginal value of a demand with attribute  $b \in \mathcal{B}$  at time  $t$ .

Note that we omit the time index  $t$  in  $\bar{v}_a^R$  since the vehicle's actionable time is already encoded in the attribute  $a$ , and we estimate a single marginal value  $\bar{v}_a^R$  across all  $t \in \{1, \dots, T\}$ . By substituting  $\bar{V}^x(S_t^x)$  into Equation (16), we obtain

$$\begin{aligned} \hat{V}(S_t) &\approx \max_{x_t \in \mathcal{X}(S_t)} \left\{ C(x_t) + \sum_{a \in \mathcal{A}} \bar{v}_a^R R_{ta}^x + \sum_{b \in \mathcal{B}} \bar{v}_{tb}^D D_{tb}^x \right\} \\ &= \max_{x_t \in \mathcal{X}(S_t)} \left\{ \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} c_{tad} x_{tad} + \sum_{a' \in \mathcal{A}} \bar{v}_{a'}^R \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} x_{tad} \cdot \mathbf{1}_{\{a^M(a, d) = a'\}} + \sum_{b \in \mathcal{B}} \bar{v}_{tb}^D D_{tb}^x \right\} \\ &= \max_{x_t \in \mathcal{X}(S_t)} \left\{ \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} \left( c_{tad} + \sum_{a' \in \mathcal{A}} \bar{v}_{a'}^R \cdot \mathbf{1}_{\{a^M(a, d) = a'\}} \right) x_{tad} + \sum_{b \in \mathcal{B}} \bar{v}_{tb}^D D_{tb}^x \right\} \\ &= \max_{x_t \in \mathcal{X}(S_t)} \left\{ \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} \left( c_{tad} + \bar{v}_{a^M(a, d)}^R \right) x_{tad} + \sum_{b \in \mathcal{B}} \bar{v}_{tb}^D D_{tb}^x \right\}, \end{aligned}$$

which gives us our VFA policy

$$\pi^{\text{VFA}}(S_t) = \arg \max_{x_t \in \mathcal{X}(S_t)} \left\{ \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}(a)} \left( c_{tad} + \bar{v}_{a^M(a,d)}^R \right) x_{tad} + \sum_{b \in \mathcal{B}} \bar{v}_{tb}^D D_{tb}^x \right\}. \quad (18)$$

## B.2 Forward ADP algorithm

---

### Algorithm 1 Forward Approximate Dynamic Programming

---

**Input:** Initial resource values  $\bar{v}_a^{R,0}$  for all  $a \in \mathcal{A}$ . Initial demand values  $\bar{v}_{tb}^{D,0}$  for all  $b \in \mathcal{B}, t \in \{1, \dots, T\}$ . Initial states  $S_0^n$ , sample paths  $(\hat{D}_1^n, \hat{D}_2^n, \dots, \hat{D}_T^n)$ , and learning rates  $\alpha_n$  for all  $n \in 1, \dots, N$ .

- 1: **for**  $n \in \{1, \dots, N\}$  **do**
- 2:   **for**  $t \in \{1, \dots, T\}$  **do**
- 3:     Solve the LP
- $$\max_{x_t^n \in \mathcal{X}^{\text{LP}}(S_t^n)} \left\{ C(x_t^n) + \sum_{a \in \mathcal{A}} \bar{v}_a^{R,n-1} R_{ta}^x + \sum_{b \in \mathcal{B}} \bar{v}_{tb}^{D,n-1} D_{tb}^x \right\}$$
- to obtain  $x_t^n, \hat{v}_{ta}^{R,n}$  for all  $a \in \mathcal{A}$  and  $\hat{v}_{tb}^{D,n}$  for all  $b \in \mathcal{B}$ .
- 4:     Update the lookup tables
- $$\bar{v}_a^{R,n} := (1 - \alpha_n) \bar{v}_a^{R,n-1} + \alpha_n \hat{v}_{ta}^{R,n},$$

$$\bar{v}_{tb}^{D,n} := (1 - \alpha_n) \bar{v}_{tb}^{D,n-1} + \alpha_n \hat{v}_{tb}^{D,n}.$$
- 5:     Transition to the next state  $S_{t+1}^n = S^M(S_t^n, x_t^n, \hat{D}_{t+1}^n)$ .
- 6:   **end for**
- 7: **end for**
- 8: **return**  $\bar{v}_a^{R,N}$  and  $\bar{v}_{tb}^{D,N}$  for  $t \in \{1, \dots, T\}$ .

---

We adapt the well-known Forward ADP algorithm (see Algorithm 1) to calculate estimates of  $\bar{v}_a^R$  for every  $a \in \mathcal{A}$  and  $\bar{v}_{tb}^D$  for every  $b \in \mathcal{B}$  and  $t \in \{1, \dots, T\}$ , which we use in our VFA policy (18). The idea of the Forward ADP algorithm is to start from an initial approximation  $\bar{v}_a^{R,0}$  and  $\bar{v}_{tb}^{D,0}$ , simulate  $N$  sample paths of the system, and iteratively improve these approximations by solving an LP at every time step of every sample path. Specifically, in each iteration  $n \in \{1, \dots, N\}$ , we simulate a sample path of the system from time  $t = 0$  to  $t = T$ . Then, at every time step  $t \in \{1, \dots, T\}$ , we evaluate a surrogate policy by removing all multi-trip decisions from the feasible region  $\mathcal{X}(S_t)$ , relaxing the integrality constraint, and solving an LP that uses the current approximations  $\bar{v}_a^{R,n-1}$  and  $\bar{v}_{tb}^{D,n-1}$  (Line 3 in Algorithm 1). Solving this LP is guaranteed to yield an integer solution, due to the total unimodularity of the constraint matrix and the integrality of the right-hand side. Therefore, solving this LP we can obtain the shadow prices

- $\hat{v}_{ta}^{R,n} = \frac{\partial \hat{V}(S_t^n)}{\partial R_{ta}}$  for the constraint of vehicle attribute  $a \in \mathcal{A}$ , and
- $\hat{v}_{tb}^{D,n} = \frac{\partial \hat{V}(S_t^n)}{\partial D_{tb}}$  for the constraint of request attribute  $b \in \mathcal{B}$ .

where  $S_t^n$  is the system state at time  $t$  in the  $n$ -th iteration, and  $\hat{v}_{ta}^{R,n}$  and  $\hat{v}_{tb}^{D,n}$  are estimates of the slope of  $\hat{V}(S_t^n)$  with respect to  $R_{ta}$  and  $D_{tb}$  at the point  $R_{ta}^x$  and  $D_{tb}^x$ , respectively. Note that we retain the time index  $t$  in  $\hat{v}_{ta}^{R,n}$  to emphasize that the partial derivative is taken with respect to the resource vector  $R_{ta}$  at time  $t$ . Finally, we use these estimates to update the current approximations  $\bar{v}_a^{R,n-1}$  and  $\bar{v}_{tb}^{D,n-1}$  as shown in Line 4 of Algorithm 1, transition to the next state  $S_{t+1}^n$ , and repeat this process until we reach the end of the sample path.

## B.3 Monotonicity of the value function

This section extends Section 4.3.3 of the paper. We first repeat the relevant definition:

**Definition B.1** (Partial Order). Let  $a, a' \in \mathcal{A}$  with  $a = (o_a \ d_a \ l_a \ n_a \ t_a)^\top$  and  $a' = (o_{a'} \ d_{a'} \ l_{a'} \ n_{a'} \ t_{a'})^\top$ . Then,  $a \preceq a'$  if, and only if,  $(o_a, d_a) = (o_{a'}, d_{a'})$ ,  $l_a \leq l_{a'}$ ,  $n_a \leq n_{a'}$  and  $t_a \geq t_{a'}$ .

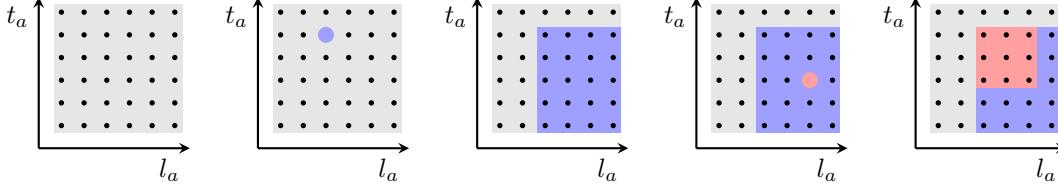


Figure 4: Monotonicity properties are used to update a value function twice. The  $o_a$ ,  $d_a$  and  $n_a$  dimensions are not drawn to allow for a two-dimensional representation. Initially, the value function is at a low value (left). The first update increases it to a high value for all attributes  $a$  with  $a \preceq a'$  (center). The second update decreases it to a medium value for all attributes  $a$  with  $a'' \preceq a$  (right). Figure adapted from Powell (2022).

Definition B.1 allows us to show three monotonicity properties with respect to an optimal value function for vehicle attributes,  $v_a^{R*}$ . The first property applies when we compare two vehicles  $a \preceq a'$  and  $a'$  has a higher range than  $a$ . In this case, the higher-range vehicle cannot have a lower value than the lower-range vehicle. This is because the higher-range vehicle can serve the same trips as the lower-range vehicle and therefore the higher-range vehicle can obtain the same reward as the lower-range vehicle. The second property applies when we compare two vehicles  $a \preceq a'$  and  $a'$  has a higher capacity than  $a$ . In this case, the higher-capacity vehicle cannot have a lower value because it can serve the same trips as the lower-capacity vehicle. The third property applies when we compare two vehicles  $a \preceq a'$  and  $a'$  can start performing new tasks (i.e., is actionable) at an earlier time than  $a$ . In this case, the earlier vehicle cannot have a lower value because it has more time to serve travel requests and therefore can earn the same reward as the later vehicle.

**Proposition B.1** (Monotonicity). *For every pair of vehicle attributes  $a, a' \in \mathcal{A}$  with  $a \preceq a'$ , if the statements  $a^M(a, d) \preceq a^M(a', d)$  and  $c_{tad} \leq c_{ta'd}$  are true for all decisions  $d \in \mathcal{D}(a) \cap \mathcal{D}(a')$  and time steps  $t \in \{1, \dots, T\}$ , then the following monotonic properties hold:*

1.  $v_a^{R*}$  increases monotonically with the  $l_a$  dimension:  $l_a < l_{a'} \implies v_a^{R*} \leq v_{a'}^{R*}$ .
2.  $v_a^{R*}$  increases monotonically with the  $n_a$  dimension:  $n_a < n_{a'} \implies v_a^{R*} \leq v_{a'}^{R*}$ .
3.  $v_a^{R*}$  decreases monotonically with the  $t_a$  dimension:  $t_a > t_{a'} \implies v_a^{R*} \leq v_{a'}^{R*}$ .

*Proof.* For every pair of vehicle attributes  $a, a' \in \mathcal{A}$ , every decision  $d \in \mathcal{D}(a) \cap \mathcal{D}(a')$ , and every time step  $t \in \{1, \dots, T\}$ , the statement  $a \preceq a' \wedge a^M(a, d) \preceq a^M(a', d) \wedge c_{tad} \leq c_{ta'd}$  is satisfied by definition of  $a^M$  and  $c_{tad}$ . All three monotonic properties can be proved using the same backward induction argument as in Al-Kanj et al. (2020).  $\square$

We note that the following two assumptions made by Al-Kanj et al. (2020) for their monotonicity properties are not necessary for us since the resource vector  $R_t$  is deterministic in our model. The first assumption is that for every  $a, a' \in \mathcal{A}$ , if  $a \preceq a'$  is satisfied in some  $S_t^x$ , it is also satisfied in  $S_{t+1}^x$ . The second assumption is that the demand component and the vehicle component of the exogenous information are stochastically independent.

We leverage the monotonicity of  $v_a^{R*}$  to address the exploration-exploitation trade-off, and to accelerate learning  $\bar{v}_a^{R,n}$ , as follows. We initialize the lookup table such that it is monotonic in all dimensions (e.g.  $\bar{v}_a^{R,0} = 0, \forall a \in \mathcal{A}$ ). After updating the lookup table (lines 7 and 8 of Algorithm 1),  $\bar{v}_a^{R,n}$  may violate some monotonicity property. To restore monotonicity, we force all entries in the lookup table that violate monotonicity to a larger or smaller value so that monotonicity is restored. See Figure 4 for an illustration.

## C Proof of multi-trip decision enumeration algorithm

We first repeat the relevant definitions, then state the proposition, and finally give the proof.

**Definition C.1** (Monotonicity of detour penalty). *The detour penalty function  $\rho^{detr}$  is monotonically increasing in the path length if, and only if, for all vehicle attributes  $a \in \mathcal{A}^{empt}$ , all sets of requests  $B \subseteq \mathcal{B}$  with  $|P_{aB}| > 0$ , and all pairs of paths  $p, p' \in P_{aB}$ , the following statement is true:*

$$\tau(p) \leq \tau(p') \implies \rho^{detr}(a, d_{Bp}^{m\text{lti}}) \leq \rho^{detr}(a, d_{Bp'}^{m\text{lti}}).$$

**Definition C.2** (Sufficient paths set). *Let  $\pi$  be a policy. For all time steps  $t \in \{1, \dots, T\}$  and all states  $S_t$  where a vehicle attribute  $a \in \mathcal{A}^{empt}$  satisfies  $R_{ta} > 0$  and where the nonempty set of requests  $B \subseteq \mathcal{B}$  satisfies  $\forall b \in B. D_{tb} > 0$*

and  $|P_{aB}| > 0$ , the set of paths  $P_{aB}^* \subseteq P_{aB}$  is a-B- $\pi$ -sufficient if, and only if, there exists a solution  $x_t$  in the set of optimal solutions  $\pi(S_t)$  such that all decision variables  $x_{tad}$  with  $d \in \{d_{Bp}^{\text{milti}} \in \mathcal{D}_a^{\text{milti}} \mid p \in P_{aB} \setminus P_{aB}^*\}$  are equal to zero.

**Definition C.3 (LDSPS).** For all time steps  $t \in \{1, \dots, T\}$  and all states  $S_t$  where a vehicle attribute  $a \in \mathcal{A}^{\text{empt}}$  satisfies  $R_{ta} > 0$  and where the nonempty set of requests  $B \subseteq \mathcal{B}$  satisfies  $\forall b \in B. D_{tb} > 0$  and  $|P_{aB}| > 0$ , the set of paths  $P_{aB}^* \subseteq P_{aB}$  is an LDSPS for  $(a, B)$  if, and only if the following statement is true for all requests  $b \in B$ : Either  $P_{aB}^*$  contains exactly one of the shortest paths in  $P_{aB}$  that end in the destination of  $b$ ,  $d_b$ , or no path in  $P_{aB}$  ends in  $d_b$ .

**Proposition C.1** (Sufficiency of the LDSPS). For all time steps  $t \in \{1, \dots, T\}$  and all states  $S_t$  where a vehicle attribute  $a \in \mathcal{A}^{\text{empt}}$  satisfies  $R_{ta} > 0$  and where the nonempty set of requests  $B \subseteq \mathcal{B}$  satisfies  $\forall b \in B. D_{tb} > 0$  and  $|P_{aB}| > 0$ , if the approximate value function used by policy  $\pi^{\text{VFA}}$  satisfies Proposition 4.1 and  $\rho^{\text{detr}}$  is monotonically increasing in the path length, all LDSPSs for  $(a, B)$  are a-B- $\pi^{\text{VFA}}$ -sufficient.

*Proof.* Suppose, for the sake of contradiction, that the proposition is false. The negation of the proposition is as follows.

There exists a time step  $t \in \{1, \dots, T\}$  and a state  $S_t$  where a vehicle attribute  $a \in \mathcal{A}^{\text{empt}}$  satisfies  $R_{ta} > 0$  and where the nonempty set of requests  $B \subseteq \mathcal{B}$  satisfy  $\forall b \in B. D_{tb} > 0$  and  $|P_{aB}| > 0$  such that (i) the approximate value function used by policy  $\pi^{\text{VFA}}$  satisfies Proposition 4.1, (ii)  $\rho^{\text{detr}}$  is monotonically increasing in the path length, and (iii) there exists an LDSPS for  $(a, B)$  that is not a-B- $\pi^{\text{VFA}}$ -sufficient.

Statement (iii) is equivalent to the following statement: There exists a set of paths  $P_{aB}^* \subseteq P_{aB}$  such that (iv) the following statement is true for all requests  $b \in B$ : Either  $P_{aB}^*$  contains exactly one of the shortest paths in  $P_{aB}$  that end in the destination of  $b$ ,  $d_b$ , or no path in  $P_{aB}$  ends in  $d_b$ , and (v) for all solutions  $x_t$  in the set of optimal solutions  $\pi^{\text{VFA}}(S_t)$ , there exists a decision variable  $x_{tad}$  with  $d \in \{d_{Bp}^{\text{milti}} \in \mathcal{D}_a^{\text{milti}} \mid p \in P_{aB} \setminus P_{aB}^*\}$  that is not equal to zero.

Analyzing Statement (v), let  $p \in P_{aB} \setminus P_{aB}^*$  be the path associated with  $x_{tad}$ , i.e.,  $d = d_{Bp}^{\text{milti}}$ . Since  $B$  is nonempty, there exists a request  $b \in B$  that is dropped off last in  $p$ . Using Statement (iv), this implies that the set  $P_{aB}^*$  is nonempty, so there exists a shortest path  $p^* \in P_{aB}^*$  with  $\tau(p^*) \leq \tau(p)$ . By Definition B.1, this is equivalent to  $a^M(a, d) \preccurlyeq a^M(a, d^*)$ , where  $d^* = d_{Bp^*}^{\text{milti}}$ . Because of Statements (i) and (ii), this implies that

$$\bar{v}_{a^M(a,d)}^R \leq \bar{v}_{a^M(a,d^*)}^R \quad \text{and} \quad \rho^{\text{detr}}(a, d) \leq \rho^{\text{detr}}(a, d^*). \quad (19)$$

We have shown that the supposition implies that both Statement (v) and (19) are true. We will now show that the supposition also implies that either Statement (v) or (19) is false.

By construction, the optimization problem (18) contains the decision variable  $x_{tad^*}$ , and the objective function contains both

$$(c_{tad} + \bar{v}_{a^M(a,d)}^R)x_{tad} \quad \text{and} \quad (c_{tad^*} + \bar{v}_{a^M(a,d^*)}^R)x_{tad^*}$$

as summands. Because of the nonnegativity constraint in  $\mathcal{X}$ ,  $x_{tad} \geq 0$  and  $x_{tad^*} \geq 0$  in all solutions  $\pi^{\text{VFA}}(S_t)$ . Using Statement (v), this implies that  $x_{tad} > 0$  in all solutions. Two cases remain:  $x_{tad^*} > 0$  or  $x_{tad^*} = 0$ .

If  $x_{tad^*} > 0$ , the objective coefficients of both variables must be equal. This implies that there is an alternative optimal solution with  $x_{tad} = 0$ . This contradicts Statement (v).

If  $x_{tad^*} = 0$ , this is equivalent to

$$\begin{aligned} c_{tad} + \bar{v}_{a^M(a,d)}^R &> c_{tad^*} + \bar{v}_{a^M(a,d^*)}^R \\ \Leftrightarrow \rho^{\text{detr}}(a, d) + \bar{v}_{a^M(a,d)}^R &> \rho^{\text{detr}}(a, d^*) + \bar{v}_{a^M(a,d^*)}^R \end{aligned}$$

which contradicts (19).  $\square$

## D Supplement to numerical experiments

This section outlines the process of creating the benchmark instances we use in our numerical experiments, discusses the problem settings parameters, explains relevant implementation details, and gives detailed results.

## D.1 Street networks

The 2009 TLC data provides the latitude and longitude coordinates of every trip’s origin and destination. This allows us to adopt a fine-grained representation of the street network where we discretize NYC into rectangles of equal size, each approximately 215 m wide by 280 m high. We define a directed graph that has one vertex per rectangle and that contains an arc  $(u, v)$  if and only if vertices  $u$  and  $v$  represent rectangles that are directly connected by a street according to OpenStreetMap data. The TLC data does not contain a trip between each adjacent pair of vertices that we have in our graph. To define the driving duration of each arc, we fit the simple linear regression model  $t = \beta s$ , where  $t$  is the trip duration in seconds and  $s$  is the haversine distance in meters between the origin and destination of the trip. The estimated coefficients are  $\hat{\beta} = 0.216$  for the Manhattan instance and  $\hat{\beta} = 0.205$  for the four eastern boroughs. We define the driving durations of all arcs in the 2009 instances as the predictions of this model.

The 2018 TLC data does not provide the latitude and longitude coordinates of the trips’ origins and destinations. Instead, it provides the TLC zone of each trip’s origin and destination. We use this information to create a coarse-grained representation of the street network. We define the set of vertices as the set of all 263 TLC zones in the four eastern boroughs. We define the set of arcs such that there is one pair of arcs between each pair of vertices whose zones border each other. Additionally, we manually add arcs to represent the bridges and tunnels that cross the East River and the Bronx River. We define the driving duration of each arc  $(u, v)$  as the median duration of all trips between zones  $u$  and  $v$ .

## D.2 Demand distribution

A demand realization at time  $t$  is a customer request given by the vector

$$b = (o_b, d_b, n_b, t_b^{\text{res}}, t_b^p, f_b)^\top.$$

The 2009 TLC data contains the latitude and longitude coordinates of the origin  $o_b$  and destination  $d_b$  of every trip  $b$ . The 2018 TLC data provides the TLC zones of the origin and destination. For both years, the TLC data contains the number of passengers  $n_b$  and the trip fare  $f_b$  of each request, as well as the pickup time, which we use as the time at which the RHS receives the request. The latest pickup time  $t_b^p = T$ , i.e., all requests must be picked up before the end of the planning horizon. We assume that the latest response time  $t_b^{\text{res}}$  is 5 min after the request occurs.

We define a sample path as a set of demand realizations that occur between time  $t = 0$  and  $t = T$ . We associate each street network with a set of sample paths that we derive from its respective TLC data (2009 and 2018) as follows. Similarly to Kullman et al. (2022), we use the data of March and April, a span of time that contains no public holidays. We take several steps to clean the data. We only consider the trips on Mondays to Fridays to reduce the variance. From these, we keep only those trips whose origin and destination are distinct, distance is between 16.09 m and 80 km, duration is greater than 1 min, passenger count is positive, and whose fare is not below the base fare for taxis in NYC. We then remove the trips with the 5% largest fares, which tend to be outliers. Finally, we define the speed of a trip as its distance divided by its duration, and we remove all trips that are either slower than the average taxi speed in the most crowded part of NYC or faster than the maximum speed limit on any street in NYC, according to the New York City Department of Transportation (2018).

The TLC data contains millions of requests. To keep the computational effort manageable, we define the number of requests in a sample path as 1 % of the trips in a randomly chosen Tuesday, Thursday or Friday in March or April of the year of the instance. We choose these trips randomly, as samples from a uniform distribution, separately for each sample path. To keep the ratio between requests and vehicles realistic, we define the fleet size as the number of simultaneously active taxis, scaled down by 1 % again. We gather taxi count estimates from Schneider (2018). All in all, our sampling method closely resembles that of Kullman et al. (2022), except for the request count, which in their case is a parameter and constant across sample paths, but in our case is sampled from the TLC data and can be different in every sample path.

## D.3 Implementation details

We train the VFA policy on 3000 sample paths, and we test both policies on 30 sample paths. For each instance, we create 3000 training sample paths and 30 test sample paths and reuse them across problem settings and policies. Similarly, we sample the initial state, which is the location and driving range of each vehicle, once per sample path from a uniform distribution. Every vehicle starts empty. As a result, all RHS operators face exactly the same conditions in terms of initial state and demand, which enables a fair comparison.

We use  $\theta = 0.1$  as the charging threshold of the PM policy. This means that the PM policy recharges a vehicle if and only if it has less than or equal to 10 % range left. We found that this threshold, which is the same as the one Al-Kanj et al. (2020) use, provides the best performance based on preliminary experiments.

The VFA policy we implement uses a look-up table with monotonic properties as described in Section B.3 of this online supplement to approximate the value of each vehicle attribute. We aggregate the driving range  $l_a$  into nine equidistant levels, the capacity  $n_a$  into two levels (vehicle is empty or not), and we aggregate the actionable time  $t_a$  into 288 time levels, each of which corresponds to five minutes. This does not affect the decision epoch. We do not aggregate the location dimensions, i.e., the vehicle's origin  $o_a$  and destination  $d_a$ . We use generalized harmonic learning rate decay with the parameter  $a = 300$  (Powell, 2011). For the demand attributes, instead of a lookup table, we use the estimate  $\bar{v}_{tb}^D = 0$  if  $t_b^{\text{res}} \leq t$ , and  $\bar{v}_{tb}^D = 0.9f_b$  otherwise.

We implement our numerical experiments in Julia 1.10 using JuMP version 1.20 (Dunning et al., 2017) with the solver Gurobi 11.0 (Gurobi Optimization, LLC, 2024) to model and solve the LP and MIP problems. We performed the simulations on Intel Xeon 8468 Sapphire CPUs using up to 450 GB of memory, depending on the instance. The code developed for this paper will be made available upon publication.

#### D.4 Tables

The two tables in this section contain detailed results of our numerical experiments. In Table 3, we report the average RFR, the median, the IQR, and the margin of error (MOE) for the 95% confidence interval given by  $1.96\sigma\sqrt{N}^{-1}$  where  $\sigma$  is the standard deviation of the cumulative rewards and  $N$  is the number of sample paths in the test set. In Table 2, we report the same statistics regarding the absolute rewards.

Instance	Pooling-enabled	Fleet type	PM Reward (\$)				VFA Reward (\$)			
			Mean	Median	IQR	MOE	Mean	Median	IQR	MOE
1	false	EV (DCFC)	18990	19017	1296	342	20508	20771	1456	365
		EV (L2C)	13426	13465	836	221	18949	19065	1048	291
		ICE	19170	19459	1711	379	20408	20820	1808	392
	true	EV (DCFC)	19882	19998	1459	363	21240	21403	2022	534
		EV (L2C)	12896	12884	1070	244	20741	20860	1907	506
		ICE	20272	20497	921	436	21146	21298	1991	534
2	false	EV (DCFC)	22382	22226	1410	395	24347	24006	2093	573
		EV (L2C)	16589	16650	788	227	21812	21874	834	282
		ICE	22439	22267	1407	446	24326	24218	1570	494
	true	EV (DCFC)	23415	23352	1601	406	24305	23990	1882	541
		EV (L2C)	15840	15787	889	214	23423	23077	1699	518
		ICE	23579	23581	1720	430	24263	23893	1685	540
3	false	EV (DCFC)	12349	12335	466	198	14849	14829	1006	330
		EV (L2C)	8249	8286	688	162	13733	13754	644	198
		ICE	12469	12472	799	201	13700	13794	526	202
	true	EV (DCFC)	12617	12685	539	183	14895	14826	1252	370
		EV (L2C)	7719	7641	698	171	13599	13736	531	208
		ICE	12973	13094	472	203	14938	14919	1261	387
4	false	EV (DCFC)	16700	16736	819	220	20191	20126	1525	289
		EV (L2C)	11035	11006	487	136	18668	18698	696	191
		ICE	16872	16648	793	243	20288	20164	1063	242
	true	EV (DCFC)	17254	17227	536	152	20824	20538	1594	357
		EV (L2C)	10394	10305	484	190	19576	19438	1005	225
		ICE	17636	17537	811	184	20990	20900	1791	375

Table 2: Total reward statistics.

Instance	Pooling enabled	Fleet type	PM RFR (%)				VFA RFR (%)			
			Mean	Median	IQR	MOE	Mean	Median	IQR	MOE
1	false	EV (DCFC)	80.724	80.581	4.224	1.027	85.808	86.620	4.199	0.989
		EV (L2C)	57.195	56.895	5.664	1.537	79.366	79.597	4.334	1.091
		ICE	81.372	81.458	3.490	1.009	85.249	84.947	3.419	0.825
	true	EV (DCFC)	85.611	85.672	2.093	0.823	89.331	89.204	1.154	0.311
		EV (L2C)	55.438	55.245	4.870	1.369	87.419	87.186	1.768	0.395
		ICE	86.057	86.155	2.822	1.010	89.027	89.044	1.316	0.290
2	false	EV (DCFC)	83.358	83.016	3.515	0.767	89.280	89.280	0.460	0.154
		EV (L2C)	61.867	62.399	5.535	1.331	79.831	79.764	4.066	1.117
		ICE	83.629	83.463	2.616	0.848	89.113	89.444	1.132	0.435
	true	EV (DCFC)	88.160	88.237	2.264	0.478	89.443	89.422	0.609	0.181
		EV (L2C)	59.856	59.589	4.476	1.376	86.471	86.530	0.993	0.324
		ICE	88.660	89.087	2.070	0.421	89.080	89.074	0.639	0.175
3	false	EV (DCFC)	76.639	76.776	3.234	1.215	91.007	91.222	1.411	0.348
		EV (L2C)	51.258	51.229	5.545	1.616	84.188	84.098	3.370	1.152
		ICE	77.411	76.903	3.843	1.309	83.781	84.408	3.794	1.051
	true	EV (DCFC)	79.034	78.778	3.820	1.077	91.596	91.607	0.732	0.209
		EV (L2C)	48.251	47.339	5.619	1.452	83.736	83.883	3.615	0.997
		ICE	81.032	81.267	3.577	1.105	91.959	92.100	1.078	0.247
4	false	EV (DCFC)	73.798	72.832	3.535	0.929	88.318	88.402	1.567	0.454
		EV (L2C)	48.723	49.351	2.872	0.921	81.839	81.998	4.296	1.017
		ICE	74.625	75.098	2.328	0.844	88.891	89.005	2.805	0.662
	true	EV (DCFC)	77.213	77.093	3.486	0.908	91.215	91.230	0.621	0.236
		EV (L2C)	46.072	46.350	2.817	0.910	85.891	86.138	3.365	0.760
		ICE	78.876	78.779	2.979	0.658	91.879	91.963	0.788	0.241

Table 3: Reward fulfillment ratio (RFR) statistics.