Route-Based Markov Decision Processes for Dynamic Vehicle Routing Problems

Marlin W. Ulmer  Justin C. Goodson  Dirk C. Mattfeld
Barrett W. Thomas

May 2, 2017

Abstract

We propose a new modeling framework for dynamic routing problems (DRPs). DRPs are problems in which a set of geographically dispersed customers is visited by one or more travelers or vehicles, information changes stochastically over the problem horizon, and there exist opportunities to make decisions in response to new information. Acknowledging the disconnect between model and method in DRP research, our framework extends the conventional Markov decision process (MDP) model for dynamic and stochastic optimization problems to more closely align with route-based solution methodologies predominate in the DRP literature. We construct route-based MDPs by redefining the conventional MDP action space to operate on sets of planned routes. The modification leads to a generalization of the conventional MDP state (and post-decision state) to include route plans and to a definition of the current-period reward or cost to be the marginal change in value associated with a route plan update. In addition to joining the model with both application and method, route-based MDPs are positioned to facilitate more scientific rigor in DRP studies. We anticipate route-based MDPs will provide researchers with a common language, allow for better inquiry, and improve classification and description of solution methods. Under an easily satisfiable condition, we show route-based MDPs are equivalent to the conventional MDP model, thus providing a formal means to model the evolution of routes in a sequential decision-making environment. Via an example from the literature, we illustrate the value of connecting model and solution methodology.
1 Introduction

Dynamic routing problems (DRPs) are problems in which a set of geographically dispersed customers is visited by one or more travelers or vehicles. The problems are dynamic because information stochastically changes over the operating horizon and there exist opportunities to make decisions in response to the revealed information. For example, a ride-sharing service may dynamically assign vehicles to customer requests, a delivery truck may adjust its route in response to changing traffic conditions, or appointments may be rescheduled on-the-fly to compensate for uncertain service times. While dynamic routing has a 30-year history in the operations research literature, the majority of DRP work is recent. For example, 51 of the 117 dynamic routing papers cited in Psaraftis et al. (2016) have been published since 2012. Further, the 2013-14 most cited paper in the European Journal of Operational Research reviewed dynamic routing (Pillac et al., 2013), indicating high interest in the topic among researchers. The trend in DRP research is likely to accelerate as opportunities afforded by data availability and computing power enable the integration of predictive tools with prescriptive optimization methods to anticipate and dynamically respond to uncertain events. Savelsbergh and Van Woensel (2016) suggest taking advantage of these opportunities “may necessitate new views, paradigms, and models for decision support.” Further, Speranza (2016) indicates a major obstacle for researchers addressing DRPs is “the importance of appropriately modeling dynamic events and simultaneously incorporating information about the uncertainty of future events.”

To address the challenges and opportunities surrounding DRPs, the routing community requires a unified modeling framework. Positioned between problem descriptions and solution approaches, models ideally allow researchers to connect applications with solution methods and thus clearly communicate how the proposed approach addresses the given problem. Despite the undebated importance of modeling in the deterministic routing literature, very few of the dynamic routing papers surveyed by Ulmer (2017) present a model tying together the routing application and the optimization procedure. Further, more than than two-thirds of the surveyed works do not present a model capturing the dynamic and stochastic aspects of the DRP. A unified DRP modeling framework, one that improves modeling fluency and provides constructs allowing researchers to strongly connect application to model and model to method, does not exist.
Markov decision processes (MDPs) strongly tie the dynamic routing application to the model. In particular, MDP system dynamics mimic practitioners’ environments, formally modeling decisions made in sequence and separated by the receipt of new information. The conventional MDP identifies decisions stage-by-stage, such as which customer to visit next. However, the sequential decision-making of the MDP model is often disjoint from the solution methods found in the literature. A significant portion of the papers surveyed in Ulmer (2017) present DRP solution procedures that iteratively update route plans over the course of the problem horizon. Often, the solution methods draw on mixed integer programming methods and/or metaheuristics to generate route plans, which are typically then evaluated via currently available information or by simulating future scenarios. Though route plans constitute state-of-the-art solution approaches for DRPs, conventional MDP models do not contain routing constructs. Thus, while conventional MDPs connect application to model, they often fall short of tying model to method.

In this paper, we propose a modeling framework that connects dynamic routing applications and their solution approaches. We construct route-based MDPs by redefining the conventional MDP action space to operate on sets of planned routes. The modification leads to a generalization of the conventional MDP state (and post-decision state) to include route plans and to a definition of the current-period reward or cost to be the marginal change in value associated with a route plan update. Joining route-based optimization with MDPs allows researchers to formally model the evolution of route plans alongside sequential decision-making, thus joining the model with both application and method.

In addition to strengthening the role of models in routing research, we anticipate route-based MDPs will facilitate more scientific rigor in DRP studies. As a unified modeling framework, route-based MDPs will allow researchers to express their problems in a common language, thereby facilitating inquiry by better identifying contrasts with existing work and opportunities to derive new insights from the extant literature. Additionally, as a standard model for DRPs, we further anticipate route-based MDPs will facilitate better classification and description of solution methods – how plans are formulated, generated, evaluated, impacted by information, etc.

We make three contributions. One, our paper is the first to present an MDP model based on traditional routing plans, connecting the problem model to both the routing application and to state-of-the-art solution methods. Two, we prove the equivalence of conventional and route-based
MDPs, thus establishing a theoretical basis for combining route-based optimization with dynamic and stochastic modeling. Three, using DRPs from the literature, we demonstrate how route-based MDP formulations more closely align problem models with state-of-the-art solution approaches.

The remainder of the paper is outlined as follows. In §2, we graphically illustrate the difference between conventional and route-based MDP models. In §3 and §4, we formalize conventional and route-based MDP models, respectively, providing examples throughout. In §5, we show the value of an optimal route-based MDP policy is equivalent to the value of an optimal policy in the traditional MDP. In §6, we present route-based MDP models for a challenging DRP from the literature with the aim of demonstrating the generality of the proposed model. We provide concluding remarks in §7.

2 Illustrative Example

In this section, we illustrate differences between conventional and route-based MDP models with the aim of giving readers, particularly those unfamiliar with MDPs, a conceptual understanding. In subsequent sections, we provide formal notation for both conventional and route-based MDP models.

We use as the subject of our illustrations the vehicle routing problem with stochastic service requests (VRPSSR), a problem frequently considered in the DRP literature (Gendreau et al., 1999; Bent and Van Hentenryck, 2004; Mitrović-Minić and Laporte, 2004; Thomas and White III, 2004; Hvattum et al., 2006; Ichoua et al., 2006; Thomas, 2007; Ghiani et al., 2009, 2012; Ferrucci et al., 2013; Ulmer et al., to appear). Our VRPSSR variant is characterized by the need to dynamically route one uncapacitated vehicle to meet service calls arriving randomly over a working day of duration $T$ and from a set of potential customers. We denote the known customer locations by the set $\mathcal{N} = \{0, 1, \ldots, N\}$, where 0 represents a depot and the remaining locations represent customers. Although the location of each customer in $\mathcal{N}$ is known, whether or not a customer requests service is uncertain. The known travel time between two locations $n$ and $n'$ in $\mathcal{N}$ is denoted $d(n, n')$. Each customer served accrues a unit reward and the objective is to maximize total expected reward, the expected number of serviced customers.

The following subsections illustrate the conventional MDP for the VRPSSR and the route-
based MDP, respectively. Each illustration has three parts and portrays a single decision-making episode in each model. The first part, the left-hand side of each illustration, depicts the state of the system at a point in time a decision is to be made, the state comprising all information necessary to select a decision. In the second panel of each illustration, we show the decision made given the state in the first panel. Finally, the panel on the right-hand side of each illustration depicts the state of the system resulting from the observation of new exogenous information, in this case the realization of new customer requests.

### 2.1 Conventional MDP Illustration

Figure 1 illustrates the conventional MDP model for the VRPSSR. We consider a situation with nine known customers, depicted as 1 through 9 in each of the three panels of Figure 1. The leftmost panel of Figure 1 shows the scenario at time 20. At this time, the vehicle has just serviced Customer 4 at its current location, Customers 8 and 9 have not requested service, Customers 2, 3, 5, 6, and 7 have requested service but have not yet been visited, and Customer 1 has already received service. Time, vehicle location, and customer statuses constitute the state of the system. With the information given in the state, we make a decision. For the purposes of this example, we choose to travel to Customer 2, a customer who has requested service but who has not yet been visited. We refer to this decision as $x = 2$ and represent the decision by a solid line connecting the current vehicle location at Customer 4 and the next location at Customer 2. To execute the decision, the vehicle travels to Customer 2. We assume the vehicle traverses a Manhattan-style grid where each edge requires 10 time units. We illustrate arrival of the vehicle to Customer 2 at time 40 in the right-hand-side panel of Figure 1. At this time, we also observe any new requests, the random information $\omega$, occurring between the departure from Customer 4 at time 20 and the arrival to Customer 2 at time 40. At this point, we have a new state with the vehicle located at Customer 2, Customer 4 has now been serviced, and Customer 8 has just requested service but has not yet been visited. Customers 3, 5, 6, and 7 have also requested service but have not yet been visited and Customer 1 has already received service.
2.2 Route-based MDP Illustration

The illustration of Figure 1 reveals a gap between the problem formulation and the methods most practitioners – as well as almost all approaches in the literature – employ to solve the problem. Notably, Figure 1 does not present a route through known customers, even a planned route that might later be modified. The vast majority of VRPSSR solution methods given in the literature consider not only which customer to visit next, but also determine a set of planned routes reflecting potential future movements (Gendreau et al., 1999; Bent and Van Hentenryck, 2004; Mitrović-Minić and Laporte, 2004; Hvattum et al., 2006; Ichoua et al., 2006; Thomas, 2007; Ghiani et al., 2009, 2012; Ferrucci et al., 2013; Ulmer et al., to appear). For the VRPSSR, a route plan typically contains a sequence of unserved customers, including the routing action captured by the conventional MDP, and ending in the depot. In a subsequent decision epoch, the remainder of the former plan is updated.

As Figure 2 illustrates, a route-based MDP model more closely aligns with state-of-the-art VRPSSR solution methods. Figure 2 is analogous to Figure 1, depicting the same episode but via a route-based MDP. The left-most panel of Figure 2 shows the same situation at time 20 as Figure 1. However, it also illustrates a planned route we associate with the state of the route-based MDP model. At time 20 in the current state, the route proposes travel from the vehicle’s current location at Customer 4 to Customer 3, then Customer 5, and back to the depot. Despite the planned tour
depicted in the leftmost panel of Figure 2, the decision in the route-based case is also to travel to Customer 2. This decision is depicted by the solid line from Customer 4 to Customer 2 in the middle panel of Figure 2. In contrast to the analogous panel in Figure 1, the middle panel of Figure 2 also depicts the updated route associated with the route-based MDP. With the choice to go from Customer 4 to Customer 2, the planned route continues from Customer 2 to Customers 7, 6, and 5 before returning to the depot. While Customer 9 has not yet requested service, we note the updated planned route puts the vehicle in a position to serve Customer 9 if the opportunity arises. Finally, the panel on the right-hand side of Figure 2 shows the state of the route-based MDP at time 40 when the vehicle arrives at Customer 2. At this point, Customer 8 has requested service and the planned route is updated to reflect the vehicle’s location.

In both Figures 1 and 2, the immediate action, that of traveling to Customer 2, is the same. However, Figure 2 combines the action with a route plan, thus providing a stronger connection to the solution method predominate in the literature. Further, as demonstrated by the route plan and Customer 9 in Figure 2, route plans can be an indicator of the potential impact of future realizations of exogenous information.
3 Conventional MDP Model

In this section, we begin to formalize our description of the route-based MDP. We first review the conventional MDP model and give an example for the VRPSSR. The notation presented in this section is used as the basis for the route-based MDP formulation in the following section. More detailed introductions to MDPs can be found in Puterman (2005) and Powell (2011).

3.1 Model

Conventional MDP models are characterized by six model elements:

Decision Epochs Points in time at which decisions are made.

State A tuple containing all information necessary to define the feasible actions at a particular decision epoch, the reward for choosing an action, and the resulting transition.

Actions Feasible decisions available for a particular state, also referred to as decisions.

Reward The quantity earned by choosing a particular action in a particular state.

Transition A function describing how the system evolves given a chosen action in a particular state. Transitions may involve exogenous information, notably the realization of random variables.

Objective A function of the rewards, typically maximization of the expected sum of rewards across all decision epochs.

Throughout the paper we refer to rewards. However, for minimization problems, rewards would more naturally be referred to as costs and the typical objective would minimize the expected sum of costs across all decision epochs.

We assume a finite horizon in which decisions are made at epochs 0, 1, ..., K, where K may be a random variable. At the k\textsuperscript{th} decision epoch, the system occupies state $s_k$ in state space $\mathcal{S}$. A state contains the minimum amount of information necessary to determine the actions, current-period rewards, and transition probabilities if known. At epoch $k$ and given state $s_k$, the decision maker chooses an action $x$ from the set of feasible decisions $\mathcal{X}(s_k)$. Choosing action $x$ when
in state $s_k$ induces a state transition from state $s_k$ to state $s_{k+1}$ in decision epoch $k + 1$. The transition is random and determined by selected decision $x$ and the set of random variables $\Omega_{k+1}$ representing the random information arriving between decision epochs $k$ and $k + 1$. We denote a realization of $\Omega_{k+1}$ as $\omega_{k+1}$. As discussed in Powell (2011), the transition can be split into two parts – a transition from pre-decision state $s_k$ to post-decision state $s^x_k$ and a transition from $s^x_k$ to pre-decision state $s_{k+1}$. The deterministic transition for the pre- to post-decision state is given by the function $S^X(s_k, x)$ and the random transition to the next pre-decision state is given by the function $S^\Omega(s^x_k, \omega_{k+1})$. Thus, $s_{k+1} = S(s_k, x, \omega_{k+1}) = S^\Omega(S^X(s_k, x), \omega_{k+1})$. These transitions are seen in Figures 1 and 2 in which the left-most panels illustrate pre-decision states, the middle panels post-decision states, and the right-most panels subsequent pre-decision states.

Let $\hat{R}_{k+1}(s_k, x, \omega_{k+1})$ be the random reward earned at decision epoch $k$ when selecting decision $x$ in state $s_k$ and observing random information $\omega_{k+1}$. Because $\omega_{k+1}$ may not be realized when selecting decision $x$, we define the reward in decision epoch $k$ as the expected reward $R_k(s_k, x) = E[\hat{R}_k(s_k, x, \Omega_{k+1}) | s_k, x]$, where $E[\cdot]$ denotes the expectation operator (in this case with respect to $\Omega_{k+1}$).

Let $\pi$ be a function mapping a state to an action and $\Pi$ the set of all such functions. Then, the objective is to maximize the total expected reward, conditional on initial state $s_0$, given as $\max_{\pi \in \Pi} E \left[ \sum_{k=0}^{K} R_k(s_k, x^{\pi}(s_k)) | s_0 \right]$, where $x^{\pi}(s_k)$ is the action prescribed by policy $\pi$ for state $s_k$. Discounting can be incorporated into the reward function, but we omit discount factors to simplify notation.

### 3.2 Conventional MDP Model for the VRPSSR

In this section, we present a conventional MDP model of the VRPSSR. Serving as an example of a classically-formulated DRP, the model facilitates comparison of the conventional and route-based formulations. Importantly, the example demonstrates a correct model for the VRPSSR does not require ongoing knowledge of a planned route. Additional examples of conventional MDP formulations for DRPs can be found in Thomas and White III (2004), Ichoua et al. (2006), Secomandi and Margot (2009), Goodson et al. (2016), and Ulmer et al. (to appear).
States

A decision epoch begins when the vehicle arrives at a location and observes new customer requests. The state of the system at decision epoch \( k \) is the tuple \( s_k = (n_k, t_k, z_k) \), where \( n_k \) in \( \mathcal{N} \) is the vehicle’s current location, \( t_k \) in the range \([0, T]\) is the time of arrival to \( n_k \), and \( z_k = (z_k(1), z_k(2), \ldots, z_k(N)) \) is an \( N \)-dimensional vector describing the status of each customer at epoch \( k \). For each customer \( n \), \( z_k(n) \) takes on a value in the set \( \{0, 1, 2\} \):

\[
z_k(n) = \begin{cases} 
0, & \text{if customer } n \text{ has not requested service by time } t_k, \\
1, & \text{if customer } n \text{ has made a request but has not been serviced by time } t_k, \\
2, & \text{if customer } n \text{ has been serviced by time } t_k.
\end{cases}
\]

In initial state \( s_0 = (0, 0, z_0) \), the vehicle resides at the depot at time 0 and \( z_0(n) \) is 0 or 1 for each customer \( n \) in \( \mathcal{N} \backslash \{0\} \). At final decision epoch \( K \) the process occupies a terminal state \( s_K \) in the set \( \{(0, t_K, z_K) : t_K \in [0, T], z_K \in \{0, 1, 2\}^N\} \), where the vehicle has returned to the depot by time \( T \) and each customer’s status is 0, 1, or 2. The state space is the set \( \mathcal{S} = \mathcal{N} \times [0, T] \times \{0, 1, 2\}^N \).

Actions

An action at epoch \( k \) is an assignment of the vehicle to a location in \( \mathcal{N} \). When the process occupies state \( s_k \), the set of feasible actions is

\[
\mathcal{X}(s_k) = \left\{ x \in \{n \in \mathcal{N} : z_k(n) = 1\} \cup \{0\} : t_k + d(n_k, x) + d(x, 0) \leq T \right\}.
\]

Condition (1) requires assignment to customers who have requested service or to the depot. Condition (2) restricts movement to locations from which the depot can be reached by time \( T \).

Rewards

When the process occupies state \( s_k \) and decision \( x \) in \( \mathcal{X}(s_k) \) is taken, a unit reward is accrued if a decision \( x \) moves the vehicle to a location where service has been requested: \( R_k(s_k, x) = 1 \) if \( z_k(x) = 1 \) and \( R_k(s_k, x) = 0 \) otherwise.
Transition

Following selection of action $x$ from state $s_k$, the process transitions to a new state. We describe the transition in two parts, first to a post-decision state and then to a new pre-decision state. Transition to the post-decision state reflects the new vehicle location resulting from the selected action, the time at which the vehicle will arrive at the location, and a status update for the new location. We denote the post-decision state as $s^x_k = (n^x_k, t^x_k, z^x_k)$, where vehicle location is the selected decision $n^x_k = x$, time of arrival to $n^x_k$ is $t^x_k = t_k + d(n_k, n^x_k)$, the status for location $n^x_k$ is set to $z^x_k(n^x_k) = 2$ if $n^x_k$ is a customer, and all other customer statuses remain the same.

We denote the next pre-decision state as $s_{k+1} = (n_{k+1}, t_{k+1}, z_{k+1})$. The transition from $s^x_k$ to $s_{k+1}$ marks the arrival of the vehicle to location $n_{k+1} = n^x_k$ at time $t_{k+1} = t^x_k$ as well as observation of any service requests arriving after time $t_k$ and at or before time $t_{k+1}$. Denote the set of customers requesting service by $\omega_{k+1} \subseteq \{ n \in \mathcal{N} \setminus \{0\} : z^x_k(n) = 0 \}$, a subset of the status-0 customers in post-decision state $s^x_k$. Customer statuses are updated accordingly: $z_{k+1}(n) = 1$ for all $n$ in $\omega_{k+1}$. All other customer statuses remain the same.

4 Route-Based MDP Model

In this section, we formally present our route-based MDP model. The route-based MDP model includes the same model elements (decision epochs, states, rewards, transitions, and objective) as the conventional MDP model, but redefines the feasible action space $\mathcal{X}(s_k)$ to include planned routes, typically one per vehicle per decision epoch. Generally speaking, these route plans describe a sequence of potential future actions, such as customer visits. Though the components of route plans are problem-specific, they may simply contain a sequence of customers, but might also include additional information such as assumed arrival times or planned amounts of goods to be delivered. The route-based model carries the route plan in the state and reformulates the current-period reward to capture the difference in plan values rather than the immediate return for selecting a decision. Despite these differences, as we show in the next section, the route-based MDP model is equivalent to the conventional MDP model.
4.1 Model

We first define the actions associated with the route-based MDP model, followed by descriptions of states and transitions as well as the reward function.

Actions and Route Plans

We denote an action by the pair \((x_k, \theta_{k+1}) \in \mathcal{X}(s_k) \times \Theta_{k+1}(s_k)\), where \(x_k\) and \(\mathcal{X}(s_k)\) are defined by the conventional MDP, \(\theta_{k+1}\) is a route plan, and \(\Theta_{k+1}(s_k)\) is the set of route plans given current state \(s_k\). The action and route plans are indexed differently, \(k \) versus \(k+1\), to reflect that the action \(x_k\) will be executed in the current epoch \(k\) and that the route plan includes planned actions beginning in epoch \(k+1\). It is possible that \(\Theta_{k+1}(s_k) = \emptyset\) or that \(\emptyset \in \Theta_{k+1}(s_k)\). In these cases, selection of the action \((x_k, \emptyset)\) proceeds as in the conventional MDP. Though route plans in \(\Theta_{k+1}(s_k)\) can be defined independently of \(s_k\) and \(x_k\), as subsequent examples illustrate, it is natural for the plans to be dependent on the current state and the selected action.

The example of Figure 2 illustrates the concept of an action-plan pair. In the example, we choose a route plan to be a sequence of feasible actions, an ordering of customers who have requested service but have not yet been visited. Beginning in state \(s_k\) with current route plan \(\theta_k = (3, 5)\), the selected action \((x_k, \theta_{k+1}) = (2, (7, 6, 5))\) next visits Customer 2 and updates the route plan to visit Customers 7, 6, and 5. We provide additional examples in the next section.

States and Transitions

We augment the conventional MDP state variable as the pair \((s_k, \theta_k)\) to carry both the original state and the route plan selected by action \((x_{k-1}, \theta_k)\) in the previous decision epoch. Figure 2 illustrates the state variable in a route-based MDP. In addition to depicting vehicle location, arrival time, and customer statuses, the left panel of Figure 2 shows the current route plan \(\theta_k = (3, 5)\) visiting Customers 3 and 5. Similarly, the right panel of Figure 2 gives the updated route plan \(\theta_{k+1} = (7, 6, 5)\) paired with state \(s_{k+1}\).

Analogous to the conventional MDP model, upon selection of an action and route plan, the state variable transitions to an augmented post-decision state \(S^X(s_k, (x, \theta_{k+1})) = (s_k^x, \theta_{k+1})\). Upon observation of random information \(\omega_{k+1}\), the post-decision state variable again transitions analogous
to the conventional MDP, the post-decision plan becoming the current plan, yielding pre-decision state \((s_{k+1}, \theta_{k+1})\).

**Marginal Rewards**

We connect route plan selection with decision-making by redefining the conventional MDP reward function as a marginal reward accounting for the difference in value between a previous route plan and a new route plan as well as immediate contributions by the current-period action. Let reward function \(R^\theta : \Theta \rightarrow \mathbb{R}\) map a route plan to a real number. Though \(R^\theta\) may be freely defined, it is typically aligned with the problem and chosen structure of a route plan. We define the route-based MDP reward function as

\[
R^\Delta_k (\langle s_k, \theta_k \rangle, \langle x_k, \theta_{k+1} \rangle) = R_k(s_k, x_k) + R^\theta(\theta_{k+1}) - R^\theta(\theta_k).
\]  

(3)

We illustrate the marginal reward function via Figure 2, where \(\theta_k = (3, 5)\), \(\theta_{k+1} = (7, 6, 5)\), and \(x_k = 2\). Letting \(R^\theta(\cdot)\) be the number of customers included in a route plan, \(R^\theta(\theta_k) = 2\) and \(R^\theta(\theta_{k+1}) = 3\). Then, noting \(R_k(s_k, x_k) = 1\), \(R^\Delta_k (\langle s_k, \theta_k \rangle, \langle x_k, \theta_{k+1} \rangle) = 1 + 3 - 2 = 2\).

While it is possible to include route plans in action selection but with no value by choosing \(R^\theta(\cdot) = 0\) (in which case the route-based MDP behaves as a conventional MDP), defining the reward function as a marginal difference highlights the long-term impact of route plan changes. In particular, as the above example demonstrates, the marginal reward function shifts rewards that may potentially be accumulated later in the horizon into an earlier period. Thus, connecting solution approaches to a route-based model may naturally encourage anticipatory decision-making, even when methods focus primarily on current-period marginal rewards.

### 4.2 Route-Based MDP Model for the VRPSSR

In this section, we present a route-based MDP model for the VRPSSR, formally demonstrating differences between our formulation and the conventional MDP. To fully characterize the model, we need only specify the structure of route plans and how route plans are valued.

We define a route plan as a sequence of customers, each of which has requested service but has not yet been visited. Let \(\mathcal{N}'(s_k, x_k) = \{n \in \mathcal{N} \setminus \{x_k\} : z_k(n) = 1\}\) be the set of status-1 customers
remaining in state \( s_k \) after selecting action \( x_k \). Then, given a current state \( s_k \), we define the space of routing plans \( \Theta_{k+1}(s_k) \) as the set of feasible ordered subsets of \( \mathcal{N}'(s_k, x_k) \). An ordered subset is feasible if the customers in the subset can be visited in order, beginning at location \( x_k \), and ending at the depot, without violating duration limit \( T \). We define the value of a plan \( \theta_k \) as the number of status-1 customers on the route: \( R^\theta(\theta_k) = |\theta_k| \), where \(|·|\) is the cardinality operator.

5 Model Equivalence

In this section, we make a formal connection between the conventional MDP model of §3 and the route-based MDP model of §4. Notably, for a given problem instance, we show the optimal value of a state for a suitably defined route-based MDP is related to the optimal value of a state in the conventional model, and that from an initial state the optimal values are the same.

To introduce the result, we first discuss the Bellman Equation. Often referred to as a value function or as the optimality equation, the Bellman Equation is a way of expressing the maximum reward that may be accumulated from a given state \( s_k \) onward to a terminal state. The optimality of the Bellman equation follows from MDP model assumptions (see Puterman (2005) for a detailed presentation, properties, and derivation). For a conventional MDP, the equation is

\[
V(s_k) = \max_{x \in \mathcal{X}(s_k)} \left\{ R_k(s_k, x) + \mathbb{E} \left[ V(s_{k+1} \mid s^x_k) \right] \right\},
\]

where the first term is the current-period reward and the second term, often referred to as the reward-to-go, is the value of the post-decision state. As is common, for all terminal post-decision states \( s^x_K \) we assume a null value for the reward-to-go:

\[
\mathbb{E} \left[ V(s_{K+1} \mid s^x_K) \right] = 0.
\]

Similarly, in the route-based case, we write the Bellman Equation as

\[
V^\theta(s_k, \theta_k) = \max_{(x_k, \theta_{k+1}) \in \mathcal{X}(s_k) \times \Theta_{k+1}(s_k)} \left\{ R^\Delta((s_k, \theta_k), (x_k, \theta_{k+1})) + \mathbb{E} \left[ V^\theta(s_{k+1}, \theta_{k+1} \mid s^x_k) \right] \right\},
\]

and assume for all \( s^x_K \):

\[
\mathbb{E} \left[ V^\theta(s_{K+1}, \theta_{K+1} \mid s^x_K) \right] = 0.
\]
While similar in form to the conventional Bellman Equation, the route-based Bellman Equation differs in two important ways. One, the first term of Equation (6) pulls reward into the current period that is captured in the second term of Equation (4). Two, the second term of Equation (6) is the value of the route-based post-decision state, but following the definition of $R^\Delta$, it represents the expected change in marginal rewards rather than an accumulation of rewards as in Equation (4).

Though Equations (4) and (6) have different interpretations, Proposition 1 provides a relation between the conventional and route-based value functions. The proposition requires the values of initial and terminal route plans be zero. Condition 1 formalizes the requirement:

**Condition 1.** The value of the initial route plan $\theta_0$ and the terminal route plan $\theta_{K+1}$ are zero:

$$R^\theta(\theta_0) = R^\theta(\theta_{K+1}) = 0.$$  

Proposition 1 states the value functions of the two formulations follow the simple relation $V^\theta(s_k, \theta_k) = V(s_k) - R^\theta(\theta_k)$. Thus, with Condition 1, an optimal policy for each problem achieves the same value from the initial state and the models are equivalent. The intuition behind the relation follows from recognizing the optimal value of a state in the route-based MDP results not only from current-period rewards but also from route plan values. Thus, relative to the conventional MDP, the route-based MDP accumulates value earlier in the horizon.

For an instance of the VRPSSR, Figure 3 depicts how rewards might accrue while servicing six customer requests in the conventional and route-based MDP models. In the conventional model, we observe unit increases at each decision epoch as a result of visiting a status-1 customer. The route-based model achieves the same value of six achieved by the conventional model, but in contrast we observe up-and-down value fluctuations across decision epochs. A rise in value indicates an increase in the marginal number of customers served by an updated routing plan whereas a decrease in value signifies a decline in the marginal number of planned visits.

We now formally state and prove the proposition.

**Proposition 1 (Value Function Relation).** If Condition 1 is satisfied, then $V^\theta(s_k, \theta_k) = V(s_k) - R^\theta(\theta_k)$ for all $s_k$ in $S$, for all $\theta_k$ in $\Theta(s_k)$, and for $k = 0, 1, \ldots, K$.

**Proof.** The proof is by induction. We first show the result for terminal period $K$:

$$V^\theta(s_K, \theta_K) = \max_{(x_K, \theta_{K+1}) \in X(s_K) \times \Theta(s_K)} \left\{ R^\Delta_K((s_K, \theta_K), (x_K, \theta_{K+1})) \right\}$$
Figure 3: Accumulated Reward Over Time for Route-Plan versus Conventional MDP

\[
V^\theta(s_{k}, \theta_{k}) = \max_{(x_{k}, \theta_{k+1}) \in \mathcal{X}(s_{k}) \times \Theta(s_{k})} \left\{ R_k((s_{k}, \theta_{k}), (x_{k}, \theta_{k+1})) + \mathbb{E}_k [V^\theta(s_{k+1}, \theta_{k+1}) | s_k] \right\}
\]

The first equality follows from the definition of the route-based Bellman Equation and the assumption of Equation (7). Equation (8) follows from the definition of the marginal reward function for the route-based MDP. Equation (9) follows from Condition 1 and the fact that \( R^\theta(\theta_{K}) \) is constant over the maximization. The equality in Equation (10) is valid because the selection of a plan \( \theta_{K+1} \) does not impact \( R_K(s_{K}, x_{K}) \). Equation (11) follows from Equations (4) and (5).

We assume the result holds for periods \( k + 1, k + 2, \ldots, K - 1 \). Then, for period \( k \):

\[
= \max_{(x_{k}, \theta_{k+1}) \in \mathcal{X}(s_{k}) \times \Theta(s_{k})} \left\{ R_K(s_{K}, x_{K}) + R^\theta(\theta_{K+1}) - R^\theta(\theta_{K}) \right\}
\]

\[
= R^\theta(\theta_{K}) + \max_{(x_{K}, \theta_{K+1}) \in \mathcal{X}(s_{K}) \times \Theta(s_{K})} \left\{ R_K(s_{K}, x_{K}) \right\}
\]

\[
= R^\theta(\theta_{K}) - R^\theta(\theta_{K})
\]

\[
= V(s_{K}) - R^\theta(\theta_{K}).
\]
\[
\begin{align*}
&= \max_{(x_k, \theta_{k+1}) \in \mathcal{X}(s_k) \times \Theta(s_k)} \left\{ R_k(s_k, x_k) + R^\theta(\theta_{k+1}) - R^\theta(\theta_k) \\
&\quad + \mathbb{E} \left[ V(s_{k+1}) - R^\theta(\theta_{k+1})|s_k^{x_k} \right] \right\} \\
&= \max_{(x_k, \theta_{k+1}) \in \mathcal{X}(s_k) \times \Theta(s_k)} \left\{ R_k(s_k, x_k) + R^\theta(\theta_{k+1}) - R^\theta(\theta_k) \\
&\quad - R^\theta(\theta_{k+1}) + \mathbb{E} \left[ V(s_{k+1})|s_k^{x_k} \right] \right\} \\
&= -R^\theta(\theta_k) + \max_{x_k \in \mathcal{X}(s_k)} \left\{ R_k(s_k, x_k) + \mathbb{E} \left[ V(s_{k+1})|s_k^{x_k} \right] \right\} \\
&= -R^\theta(\theta_k) + \max_{x_k \in \mathcal{X}(s_k)} \left\{ R_k(s_k, x_k) + \mathbb{E} \left[ V(s_{k+1})|s_k^{x_k} \right] \right\}
\end{align*}
\]

The first equality follows from the definition of the Bellman Equation for the route-based MDP. Equation (12) follows from the definition of the marginal reward for the route-based MDP and from the induction hypothesis. Equation (13) acknowledges the value of route plan \( \theta_{k+1} \) can be calculated separately from the post-decision value of \( s_{k+1} \). Equation (14) recognizes the value of \( \theta_k \) depends on neither \( x_k \) nor \( \theta_{k+1} \). The equality in Equation (15) recognizes that neither \( R_k(s_k, x_k) \) nor \( \mathbb{E} \left[ V(s_{k+1})|s_k^{x_k} \right] \) are affected by \( \theta_{k+1} \). Finally, Equation (16) follows from the definition of the period-\( k \) value function in the conventional MDP.

\section{A Route-based Model for a Dynamic Dial-A-Ride Problem}

Again drawing an analogy to deterministic optimization, just as various mixed integer programming formulations can connect models with solution approaches, the presentation of a DRP and accompanying solution method can benefit from an appropriate model. In this section, as a proof of concept, we formulate a route-based MDP model for the \textit{dynamic dial-a-ride problem} (DDARP), showing how route plans and reward functions can be structured to align a model with the solution approach of Schilde et al. (2014).

Similar to the VRPSSR, the DDARP seeks to service customer requests arriving randomly over a given time horizon, but with two added complexities often encountered in ride-sharing ventures. One, each request must be picked up and dropped off within a certain time frame. Two, travel times are stochastic, becoming known only after a vehicle enters a path between two locations. Drawing
on a single-vehicle version of the DDARP presented in Schilde et al. (2014), we associate with a stochastic service request by customer \( n \) a pickup location, a drop-off destination, and a time window \([e_n, l_n]\) where \( e_n \) is the earliest a customer needs pickup and \( l_n \) is the latest time by which the customer desires to arrive at the destination. The objective is to minimize the expected sum of tardiness (drop-off time for a customer \( n \) exceeds \( l_n \)), earliness (pickup time for a customer \( n \) occurs before \( e_n \)), and ride-time violations (the length of time a customer spends in the vehicle exceeds 40 minutes) accrued over service of all customer requests.

Though we build on the route-based MDP for the VRPSSR, our choice of route plan for the DDARP requires more information than a sequence of pickups and drop-offs. In particular, because our choice of route evaluation function estimates the total tardiness, earliness, and ride-time violations for a given route, we carry in a route plan additional information about planned arrival times, planned pickup times, and time windows necessary to make the estimate. These choices allow for an unambiguous mapping from route plans to rewards and facilitate a route-based solution method. In the sections that follow, we present a route-based MDP model for the DDARP and discuss how the model connects with the solution method of Schilde et al. (2014).

### 6.1 Route-Based MDP

We begin a route-based MDP model by defining conventional MDP model elements. A decision epoch occurs when the vehicle arrives at a location. The current state \( s_k \) at decision epoch \( k \) includes the time of arrival, the vehicle’s current location, passengers on the vehicle, the times at which the passengers were picked up, the pickup and delivery locations of these in-process requests as well as those of any outstanding requests, and the associated time windows. An action \( x_k \) provides service at the vehicle’s current location and moves the vehicle to a location in the set \( X(s_k) \) of origins and destinations of outstanding requests as well as destinations of in-process requests. Cost \( R(s_k, x_k) \) is the total tardiness, earliness, and ride-time violation incurred in the current state. Transition to a subsequent state begins with the observation of \( \omega_{k+1} \), the travel time required to reach the next location followed by the realization of new customer requests upon arrival to the next location, the event marking the start of decision epoch \( k + 1 \).

Structured similar to Figures 1 and 2, Figure 4 illustrates conventional and route-based MDP model elements for the DDARP in a single graphic. The left pane illustrates the current state. At
time $t_k = 20$ the vehicle arrives to drop-off location D1 for Customer 1, picked up at time 0 and with a late time window of 15. In-process requests include Customer 2, currently a passenger on the vehicle, with drop-off location D2 and a closing time window of 50. Outstanding requests include customer 3 and the just-made request from Customer 4. Respectively, the outstanding requests have pickup and delivery locations P3, D3, P4, and D4 denoted on the grid and time windows of $[20, 80]$ and $[40, 90]$. Available actions include drop-off of Customer 1 plus movement to drop-off Customer 2 or to pickup Customers 3 or 4. The cost associated with any of these actions is $R(s_k, \cdot) = 20 - 15 = 5$ time units of tardiness dropping off Customer 1. The center pane illustrates the action servicing Customer D1 and moving the vehicle to P3. The right pane shows a travel time realization of 10 time units from D1 to P3 and does not indicate any new customer requests. We use the remaining portions of Figure 4 to illustrate route plans.

The space of routing plans $\Theta_k(s_{k-1})$ in state $s_{k-1}$ is the set of feasible ordered subsets of destinations associated with in-process requests (excluding the vehicle’s current location) as well as pickup and drop-off locations accompanying outstanding requests. We represent such a subset by $(\theta_{k1}, \ldots, \theta_{km})$ and label it feasible if pickups are sequenced ahead of deliveries for each customer and if the first element of the sequence is selected action $x_{k-1}$. Additionally, to facilitate route
plan evaluation, we associate with a sequence of pickups and deliveries four pieces of information. First, denote by $a(\theta_{k_i})$ the planned arrival time to location $\theta_{k_i}$ (a number that might be a calculated expectation, an estimate via simulation, or a quantity obtained via some other method). Second, denote by $p(\theta_{k_i})$ the pickup time for the service request associated with location $\theta_{k_i}$. If $\theta_{k_i}$ is a pickup location, then $p(\theta_{k_i})$ is a planned quantity (again obtained via some measure). If $\theta_{k_i}$ is a drop-off location, then $p(\theta_{k_i})$ is set to the pickup time of the customer’s origin. Third, denote by $e(\theta_{k_i})$ the earliest time service may begin at location $\theta_{k_i}$, set to early time window $e_{\theta_{k_i}}$ if $\theta_{k_i}$ is a pickup location and to $-\infty$ if $\theta_{k_i}$ is a delivery location. Fourth, denote by $l(\theta_{k_i})$ the latest time at which service is desired at location $\theta_{k_i}$, set to late time window $l_{\theta_{k_i}}$ if $\theta_{k_i}$ is a destination and to $\infty$ otherwise. Then, a route plan is the sequence of five-tuples

$$\theta_k = \left( (\theta_{k_1}, a(\theta_{k_1}), p(\theta_{k_1}), e(\theta_{k_1}), l(\theta_{k_1})), \ldots, (\theta_{k_m}, a(\theta_{k_m}), p(\theta_{k_m}), e(\theta_{k_m}), l(\theta_{k_m})) \right). \quad (17)$$

We calculate the reward $R^\theta(\theta_k)$ associated with a route plan $\theta_k$ as the sum of projected earliness (I), tardiness (II), and ride time violations (III):

$$R^\theta(\theta_k) = \sum_{i=1}^{m} \left( \max (e(\theta_{k_i}) - a(\theta_{k_i}), 0) + \max (a(\theta_{k_i}) - l(\theta_{k_i}), 0) \right)$$

$$\left( I \right) + \max (a(\theta_{k_i}) - p(\theta_{k_i}) - 40, 0) \right). \quad (III)$$

Defining initial and terminal routing plans $\theta_0$ and $\theta_{K+1}$ as empty, and noting $R^\theta(\emptyset) = 0$, our modeling choices satisfy Condition 1 and thus Proposition 1 holds, thereby establishing a formal relationship between the conventional DDARP model and our route-based MDP model.

Figure 4 illustrates route plan evolution and evaluation. The left pane of Figure 4 shows a current route plan $\theta_k$ with a sequence of locations $(D_1, P_3, D_2, D_3)$, each location tagged with the four-tuple of associated data. Because the planned arrival time of 55 to D2 is five time units beyond the closing time window of 50, and because the projected ride time of Customer 2 is five time units greater than 40, $R^\theta(\theta_k) = 5 + 5 = 10$. Following action selection of $x_k = P_3$, the center pane of Figure 4 depicts updated route plan $\theta_{k+1}$ with a sequence of locations $(P_3, P_4, D_2, D_3, D_4)$ and associated data. The updated route plan adds Customer 4 to the location sequence and adjusts
planned arrival and pickup times to account for the five time units of tardiness at location D1. Because the planned arrival time of 60 to D2 is 10 time units past the latest drop-off time of 50 and because the projected ride time of Customer 2 is 10 time units beyond the limit, $R^{\theta}(\theta_{k+1}) = 10 + 10 = 20$. Thus, the marginal reward is $R^\Delta_k((s_k, \theta_k), (x_k, \theta_{k+1})) = 5 + 20 - 10 = 15$. The right pane of Figure 4 shows the arrival of the vehicle to location P3 at time 30, five time units before the planned arrival time of 35. Though an earliness penalty will be incurred, the arrival time to P3 allows a subsequent route plan update to move projected arrival times ahead by five time units, an update that would yield a negative marginal reward in decision epoch $k + 1$.

### 6.2 Solution Method

The route-based MDP model we propose for the DDARP aligns with the solution methodology of Schilde et al. (2014). In particular, rather than explicitly considering conventional MDP actions at each epoch, Schilde et al. (2014) work instead with routes that address not only current-period decisions, but which seek to anticipate potential vehicle movement in future scenarios.

At decision epoch $k$ in state $s_k$, the method of Schilde et al. (2014) begins by inserting new requests realized in $\omega_k$ into a route plan $\theta_k$. As in equation (17), each element of the plan is a customer location with an associated four-tuple of data. Then, Schilde et al. (2014) seek to improve the route plan via a variable neighborhood search (VNS) scheme. The VNS evaluates candidate route plans via reward function $R^\theta$ and by sampling travel times, returning both an action for the current period as well as an updated route plan. Thus, the VNS concurrently searches the space of feasible actions $\mathcal{X}(s_k)$ and the space of route plans $\Theta(s_k)$. Because Schilde et al. (2014) seek to maximize cost savings achieved by $x_k$ and $\theta_{k+1}$ relative to $\theta_k$, the VNS procedure may be viewed as a method to minimize marginal reward $R^\Delta((s_k, \theta_k), (x_k, \theta_{k+1}))$.

Connecting the solution approach of Schilde et al. (2014) to the route-based DRP via the marginal reward, as well as operating on routes, are strong links between model and method. Similar connections to a route-based MDP model might be made with the works of Bent and Van Hentenryck (2004), Coelho et al. (2014), Gendreau et al. (1999), Ghiani et al. (2012), Klapp et al. (to appear), Powell et al. (2000), Thomas (2007), and Voccia et al. (to appear). In each of these papers, the DRP solution methods operate on routes, at each epoch simultaneously considering current-period actions and updated routing plans.
7 Conclusion

We present a route-based MDP model for DRPs, demonstrating equivalence to conventional MDPs under an easily satisfied condition. Importantly, our modeling approach provides a direct connection between a dynamic and stochastic optimization model and the route-based solution methods found in much of the DRP literature. Moreover, the model-method connection we establish provides routing researchers with a framework to clearly describe their problems and related solution approaches.

As a model for DRPs, our work opens the door to a more rigorous study of DRPs. One direction for future research is to examine various components of route-based MDP models, exploring procedures to formulate, generate, and evaluate route plans. Further, route-based MDPs may inspire new DRP solution methods. For instance, though Ulmer et al. (2016) do not explicitly present a route-based MDP model in their work on dynamic same-day delivery, they operate on a set of route plans and evaluate the route plans by means of value-function approximation. More importantly, Ulmer et al. (2016) include the additional information embedded in their route plan to improve the performance of a value-function approximation. There are likely other problem classes in which information derived from a route plan might enhance the quality of a solution approach.

Another interesting avenue for future research is to further generalize the route-based model. For example, instead of a sequence of customer visits, a route plan might be expanded to an a priori policy (see Campbell and Thomas (2008) and Gendreau et al. (2016) for overviews of a priori route policies). A priori policies are characterized by routes or sets of routes, but also include recourse rules that allow the routes to respond to realizations of stochastic information.

Finally, though we present the route-based MDP model specifically for DRPs, the idea of a plan-based MDP can be extended to other domains where sequential decision-making might benefit from operating on plans rather than on only conventional MDP model elements. We anticipate potential applications in scheduling, project management, and treatment pathway identification in healthcare.
Acknowledgements

Justin Goodson wishes to express appreciation for the support of Saint Louis University’s Center for Supply Chain Management.

References


25