Real-time Bidding based Vehicle Sharing

Paper XXX

ABSTRACT

We consider the one-way vehicle sharing systems where customers can pick a car at one station and drop it off at another (e.g., Zipcar, Car2Go). We aim to optimize the distribution of cars, and quality of service, by pricing rentals appropriately. However, with highly uncertain demands and other uncertain parameters (e.g., pick-up and drop-off location, time, duration), pricing each individual rental becomes prohibitively difficult. As a first step towards overcoming this difficulty, we propose a bidding approach inspired from auctions, and reminiscent of Priceline or Hotwire. In contrast to current car-sharing systems, the operator does not set prices. Instead, customers submit bids and the operator decides to rent or not. The operator can even accept negative bids to motivate drivers to rebalance available cars in unpopular routes. We model the operator's sequential decision problem as a *constrained Markov decision problem* (CMDP), whose exact solution can be found by solving a sequence of stochastic shortest path problems in real-time. Furthermore, we propose an online approximate algorithm using the *actor-critic* method of reinforcement learning, for which this algorithm has a fast convergence rate and small variance in generalization error. We also show that its solution converges to the stationary (locally optimal) policy.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Planning and Scheduling — Planning under Uncertainty

General Terms

Algorithms, Theory, Management

Keywords

One-way vehicle sharing, Dynamic rebalancing, Constrained Markov decision problems (CMDPs), Actor-critic method

1. INTRODUCTION

One-way vehicle sharing system is an urban mobility on demand (MOD) platform which effectively utilizes usages of idle vehicles, reduces demands to parking spaces, alleviates traffic congestion during rush hours, and cuts down excessive carbon footprints due to personal transportation. The MOD vehicle sharing system consists of a network of parking stations and a fleet of vehicles. Customers arrive at particular stations can pick up a vehicle and drop it off at any other destination station. Existing vehicle sharing examples include Zipcar [17], Car2Go [32] and Autoshare [30] for one-way car sharing, and Velib [25] and City-bike [11] for one-way bike sharing. Figure 1 shows a typical Toyota i-Road one-way vehicle sharing system [19].

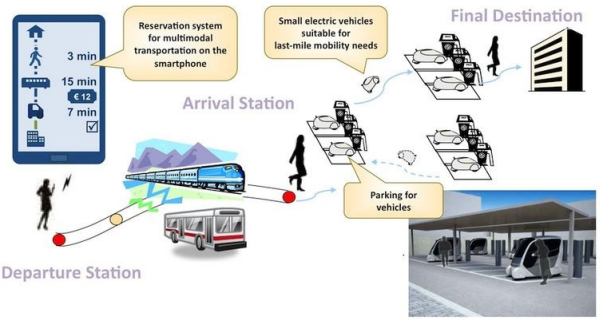

Figure 1: A Typical One-way Vehicle Sharing System that Allows Users to Pick-up and Drop-off Vehicles at Different Locations [19]

Traditional vehicle sharing system requires users to have the same drop-off and pick-up locations. This is known as the two-way vehicle sharing system. The challenges of operating two-way vehicle sharing systems are relatively small because by apriori vehicle scheduling, customers' demands can be easily fulfilled at each station. However, this service is less convenient for the users comparing to a one-way vehicle sharing system. Intuitively one-way vehicle sharing systems have a huge business potential as they allow more flexible trips than the the two-way vehicle sharing system.

Despite the apparent advantages of one-way vehicle sharing systems they do present significant operational problems. Due to the asymmetric travel patterns in a city, many stations will eventually experience imbalance of vehicle departures and customer arrivals. Stations with low customer demands (i.e., in suburbs) have excessive un-used vehicles and require many parking spaces, while stations with high demands (i.e., in city center) cannot fulfill most customers' requests during rush hours. To maintain the quality of service, many existing fleet management strategies empirically redistribute empty vehicles among stations with tow trucks or by hiring crew drivers. Still, this solution is ad-hoc and inefficient. In some cases, these scheduled re-balancing strategies may cause extra congestion to road networks as well.

In the next generation one-way vehicle sharing systems, demand-supply imbalance can be addressed by imposing incentive pricing to vehicle rentals. A typical incentive pricing mechanism can be found in [28] whose details are generalized in Figure 2. Here each

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May, 4–8, 2015, Istanbul, Turkey.*

station adjusts its rental price based on current inventory and customers' requests. Rather than passively balancing demand and supply by adjusting rental prices at each station, in this paper we study a bidding mechanism to vehicle rentals where at each station customers place bids based on their travel durations and destinations, and the company decides which bids to accept. Bidding systems on vehicle rentals have already been implemented in companies like Priceline and Hotwire [2] for several years. However, this bidding system only maximizes revenue for rental companies and it does not take into the account of balancing demand and supply. In our proposed bidding mechanism, similar to an auction marketplace, the rental company dynamically adjusts its favors to potential customers' requests based on inventory needs of origin and destination stations. Potential customers may bid prices for vehicle rentals based on realizing current inventory levels and competitions with other customers. Incentives will also be given to customers who rent vehicles in low demand stations and return in high demand stations. This causes some trips to be more expensive while other trips to be free of charges or even have finance rewards. Furthermore, the accepted bids of vehicle rentals among identical station pairs and rental durations will change in real-time as well.

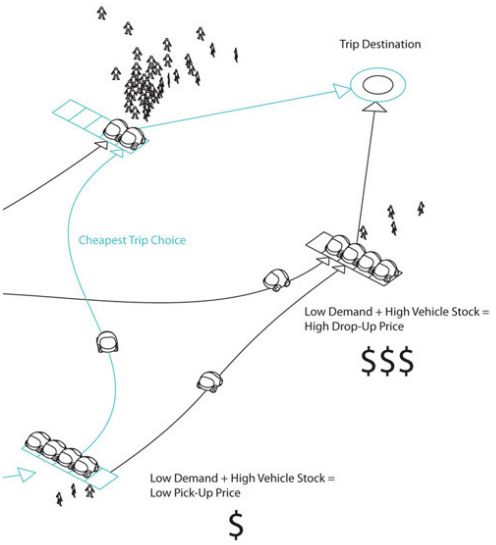

Figure 2: The Incentive Pricing Mechanism that Adjusts Rental Price Based on Inventories and Customers' Demands [28]

The design of this bidding mechanism is important for several reasons. First, accepted vehicle rental bids instantly reflect current demands and supplies in different stations. Second, by providing on-demand financial rewards for rebalancing vehicles, the rental company saves overhead costs in hiring crew drivers and renting extra parking spaces. Third, this pricing mechanism improves vehicle utilizations by encouraging extra vehicle rentals to less popular destinations and during non-rush hours. The efficiency of this bidding mechanism is scaled by the average rental duration and the size of system. In small sites such as a university campus, throughput performance can be instantly improved by providing rebalancing incentives, while there is a latency to reflect this improvement in large domains such as a metropolitan district.

1.1 Literature Review

There are several methods in literature to address demand-supply imbalance in one-way vehicle sharing system by relocating vehicles. The first suggested way is by periodic relocation of vehicles among stations by staff members. This method had been studied by [3], [18], [33] using discrete event simulations. [24] explored a stochastic mixed-integer programming (MIP) model with an objective of minimizing cost for vehicle relocation such that a probabilistic service level is satisfied. Experimental results showed that these systems improved efficiencies after re-balancing. Similar studies of static rebalancing in vehicle sharing can also be found in [15], [35], [22]. However with empirical re-balancing strategies, improvements in throughput performance are unstable, and this approach increases the sunk cost by hiring staff drivers.

Second, the user-based approach uses clients to relocate vehicles through various incentive mechanisms. Based on the distribution of parked vehicles, [36] have proposed a method to optimize vehicle assignment by trip splitting and trip joining. [23] and [10] proposed a dynamic pricing principle that enables shared vehicle drivers to trade-off between convenience and pricing. They concluded that significantly fewer vehicles were needed for the system to run efficiently. However, trip-joining policies may not be a viable solution in car-sharing due to safety and sociological concerns, and elasticity of price/location depends on fast real-time information updates, which may seem impractical in real applications.

Third, several authors have proposed trip selections for vehicle allocations. [13] formulated a multistage stochastic linear integer model for vehicle fleet management that maximizes profits of one-way car-sharing operators and account for demand variations. [9] developed several mathematical programming models to balance vehicles through choices of location, number and size of stations, and maximize the profit in a one-way car-sharing system. In both cases the car-rental company decides the number of reservations to accept and vehicles to relocate in order to maximize profit. However, both models do not provide guarantees to service levels and the proposed algorithms are not scalable in practical applications.

1.2 Contribution

The contribution of this paper is three-fold.

- In Section 2, we propose a novel mathematical model on vehicle sharing for which real time rental assignments are made based on customer arrivals and their proposed bids. The objective is to maximize the long term revenue collected from vehicle rentals at every station. There is also a constraint in each station to guarantee the long term average quality of service. In Section 3, this model is re-formulated into a CMDP for solution algorithm analysis.

- In Section 4, we rigorously derive an exact solution algorithm whose solution can be found by solving a sequence of unconstrained stochastic shortest path problems, instead of solving the large scale CMDP. This is the first main result in this paper.

- In Section 5, we also develop and analyze an iterative algorithm that effectively finds a near optimal vehicle-rental policy using reinforcement learning (the *actor-critic* method). This method incrementally updates the policy parameter at each iteration. It has a fast convergence rate and small variance in generalization error, compared to Monte-Carlo based policy gradient methods. Most actor-critic methods are proposed under the framework in average reward while our actor-critic method approximates the optimal policy of the stochastic shortest path problem. This is the second main result in this paper.

This work is only the first step in designing a market-based mechanism to tackle rebalancing issues in one-way vehicle sharing systems. We describe a wealth of open problems in Section 6.

2. MATHEMATICAL MODEL

2.1 Input from the Environment

Suppose the company has C vehicles, indexed from $1, \ldots, C$, and S stations, indexed from $1, \ldots, S$. The company's policy only allows each passenger to rent for a maximum of $\overline{T}$ time slots and the maximum fare for each rental period is $\overline{F}$.

In this paper, we consider a discrete time model $t = 0, 1, \ldots,$. At time $t \geq 0$, there is a multi-variate (four-dimensional) stationary probability distributions Φ with domain $\{1, \ldots, S\} \times \{1, \ldots, S\} \times [0, \overline{T}] \times [0, \overline{F}]$, representing the customers' origin station, destination, rental duration and proposed travel fare. We assume the multi-variate probability distribution Φ is known in advance. If the multi-variate distribution is unknown, it can easily be empirically estimated [12]. Since the vehicle sharing system can at most accept C requests, we generate C i.i.d. random variables from Φ:

$$((\mathbf{O}_t^1, \mathbf{G}_t^1, \mathbf{T}_t^1, \mathbf{F}_t^1), \ldots, (\mathbf{O}_t^C, \mathbf{G}_t^C, \mathbf{T}_t^C, \mathbf{F}_t^C))^1.$$

If $\mathbf{T}_t^k = 0$, it represents that there are no customers picking the k^{th} vehicle at time t. For $j \in \{1, \ldots, S\}$, denote by $\mathcal{A}_t^j$ the number of customers arriving at time t who wish to travel to station j. Based on the definition of random variable $\mathbf{T}_t^k$, one easily sees that this quantity can be expressed as

$$\mathcal{A}_t^j := \sum_{k=1}^{C} \mathbf{1}\{\mathbf{T}_t^k > 0, \mathbf{G}_t^k = j\}.$$

This model captures both concepts of renting and rebalancing. Notice that the random price offered by the customer k, i.e., $\mathbf{F}_t^k$ for $k \in \{1, \ldots, C\}$ can either be positive or negative. When this quantity is positive, it means that the customer is willing to paying $\mathbf{F}_t^k$ to rent a vehicle for $\mathbf{T}_t^k$ periods to travel from station $\mathbf{O}_t^k$ to $\mathbf{G}_t^k$. If this quantity is negative, it means that the company is paying $\mathbf{F}_t^k$ to the k^{th} customer, if a vehicle is needed to re-balance from station $\mathbf{O}_t^k$ to $\mathbf{G}_t^k$ in $\mathbf{T}_t^k$ periods.

In most cases, one observes periodic patterns of customer's arrivals (i.e., many customers during rush hours versus no customers during midnight), destination locations (i.e., most customers travel to city center to work in the mornings and return to residential area in the evenings), rental period (i.e., duration of travel to work) and maximum affordable fees. Therefore, it is reasonable to make the following assumption.

ASSUMPTION 1. *The state process* $\{\omega_t : t = 0, 1, \ldots,\}$ *where* $\omega_t := ((\mathbf{O}_t^1, \mathbf{G}_t^1, \mathbf{T}_t^1, \mathbf{F}_t^1), \ldots, (\mathbf{O}_t^C, \mathbf{G}_t^C, \mathbf{T}_t^C, \mathbf{F}_t^C))$ *follows the transition of a finite state ergodic Markov Chain.*

Notice that the sample space of ω_t is finite and it is denoted by Ω. Since ω_t is an ergodic Markov chain, it takes finite values and regularly returns to an initial state ω_0 after a random but finite period.

Since $(\mathbf{O}_t^1, \mathbf{G}_t^1, \mathbf{T}_t^1, \mathbf{F}_t^1), \ldots, (\mathbf{O}_t^C, \mathbf{G}_t^C, \mathbf{T}_t^C, \mathbf{F}_t^C)$ are i.i.d. random vectors, intuitively there is no difference in assigning any specific vehicles to corresponding potential customers if the customers' information is not known in advance. Rather, based on the vehicle biding mechanism in our problem formulation, the company obtains the stochastic customer information vector ω_t before deciding any actions on renting, parking or rebalancing. Therefore at each destination station, it has a pre-determined passenger ranking function to select "better customers", i.e., customers which maximize revenue (or minimize rebalancing cost) and minimize vehicle usage. We define f_{rank}^j as the customer ranking function for destination station $j \in \{1, \ldots, S\}$ based on the price-time ratio: $\mathbf{1}\{\mathbf{F} \geq 0\}\mathbf{F}/\mathbf{T} + \mathbf{1}\{\mathbf{F} \leq 0\}\mathbf{F}\mathbf{T}$ for $\mathbf{T} \neq 0$. Specifically, for any arbitrary customer information vector

$$\omega = ((\mathbf{O}^1, \mathbf{G}^1, \mathbf{T}^1, \mathbf{F}^1), \ldots, (\mathbf{O}^C, \mathbf{G}^C, \mathbf{T}^C, \mathbf{F}^C)),$$

the customer ranking function $f_{\text{rank}}^j(\omega)$ assigns score $-\infty$ to the elements with $\mathbf{T}^k = 0$ or $\mathbf{G}^k \neq j$, for $k \in \{1, \ldots, C\}$ in ω, and assigns score $\mathbf{1}\{\mathbf{F}^k \geq 0\}\mathbf{F}^k/\mathbf{T}^k + \mathbf{1}\{\mathbf{F}^k \leq 0\}\mathbf{F}^k\mathbf{T}^k$ to other elements whose destination station $\mathbf{G}^k = j$ for $k \in \{1, \ldots, C\}$.

REMARK 1. *The operator favors customers with high rental price and short travel time, i.e., for the customers who pay for rental (*$\mathbf{F}^k \geq 0$ *for* $k \in \{1, \ldots, i'\}$*):*

$$\frac{\mathbf{F}^k}{\mathbf{T}^k} \geq \frac{\mathbf{F}^{k+1}}{\mathbf{T}^{k+1}},$$

and favors drivers with low financial reward and short rebalancing time, i.e., for the customers who receive financial reward from rebalancing ($\mathbf{F}^k \leq 0$ *for* $k \in \{i' + 1, \ldots, \mathcal{A}^j\}$*):*

$$\mathbf{F}^k\mathbf{T}^k \geq \mathbf{F}^{k+1}\mathbf{T}^{k+1}.$$

If each vehicle speed is almost identical, similar analogy can also be applied to travel distance as well.

2.2 State Variables

The operator makes decisions based on the stochastic inputs generated from the environment and the current system observations of each vehicle in the fleet. These observations are represented by the state variables as follows:

- For $i \in \{1, \ldots, C\}$ and $t \geq 0$, $q_t^i \in \{1, \ldots, S\}$ is the destination station at time t of the i^{th} vehicle. Also define $q_t = (q_t^1, \ldots, q_t^C)$ as the stochastic state vector of $\{q_t^i\}$.
- For $i \in \{1, \ldots, C\}$ and $t \geq 0$, $\tau_t^i \in \{0, 1, 2, \ldots, \overline{T}\}$ is the current travel time remaining to destination on the i^{th} vehicle. Also define $\tau_t = (\tau_t^1, \ldots, \tau_t^C)$ as the state vector of $\{\tau_t^i\}$.

2.3 Decision Variables

At any time slot t, in order to maximize the expected revenue and satisfy the service level agreement constraints, the company makes a decision to park, re-balance or to rent vehicle to any potential passengers. The company's decision is a function mapping from the realizations of the current states and the current stochastic inputs to the action space. More information on the control policy will be later given in Section 3.2.

Specifically, at each time slot t, we have the following set of decision variables:

- For each station $j \in \{1, \ldots, S\}$, $\mathbf{u}_t^j \in \{0, 1, \ldots, C\}$ is a decision variable that represents the number of vehicles to dedicate to destination station j at time t. Also define the decision $\mathbf{u}_t = (\mathbf{u}_t^1, \ldots, \mathbf{u}_t^S)$ as the operator's decision vector of $\{\mathbf{u}_t^j\}_{j=1,\ldots,S}$.

These decision variables have the following constraint to upper bound the decision variable at time $t \geq 0$:

$$\mathbf{u}_t^j \leq \mathcal{A}_t^j, \forall j \in \{1, \ldots, S\}. \quad (1)$$

[1]We will later see that for any $k \in \{1, \ldots, C\}$, the state process $(\mathbf{O}_t^k, \mathbf{G}_t^k, \mathbf{T}_t^k, \mathbf{F}_t^k)$ is a finite state, ergodic Markov chain. Therefore, we assume the random travel time $\mathbf{T}_t^k$ and fare $\mathbf{F}_t^k$ are rounded to the nearest integer greater than or equal to this element, i.e., $\mathbf{T}_t^k \leftarrow \lceil \mathbf{T}_t^k \rceil$ and $\mathbf{F}_t^k \leftarrow \lceil \mathbf{F}_t^k \rceil$.

Furthermore, the total number of vehicle assignment is equal to C, i.e.,

$$\sum_{j=1}^{S} \mathbf{u}_t^j = C, \forall j \in \{1, \ldots, S\}. \quad (2)$$

2.4 State Dynamics

Before stating the state dynamics of (q_t, τ_t), we start by constructing a destination allocation function for each vehicle. Define the quota index $\mathbf{Q} = (\mathbf{Q}^1, \ldots, \mathbf{Q}^S)$ whose domain lies in $\{0, 1, \ldots, C\}^S$. For each $k \in \{1, \ldots, S\}$, $\mathbf{Q}^k$ is a quota index that counts the number of vehicle assignments to destination station k. Recall the arbitrary information vector ω from Section 2.1. At any origin $j \in \{1, \ldots, S\}$, construct an allocation function $\mathcal{G}(\omega, \mathbf{Q}, j) : \Omega \times \{0, 1, \ldots, C\}^S \times \{1, \ldots, S\} \to \{1, \ldots, S\} \times \{1, \ldots, S\} \times [0, \overline{T}] \times [0, \overline{F}]$ for which this function examines the current origin station of each request and outputs the corresponding information based on the available quota and maximum score. Specifically, let $\omega^j = \{(\mathbf{O}, \mathbf{G}, \mathbf{T}, \mathbf{F}) : (\mathbf{O}, \mathbf{G}, \mathbf{T}, \mathbf{F}) \in \omega,\ \mathbf{O} = j\}$ be a sub-vector of ω whose elements have origins at $j \in \{1, \ldots, S\}$. Then, define $\text{Assign}(f_{\text{rank}}^{j'}(\omega^j)) = (\mathbf{O}, \mathbf{G}, \mathbf{T}, \mathbf{F})$ as a function that finds an element in ω^j with maximum score corresponding to destination station j', where $\{v^{j'}\}_{j' \in \{1,\ldots,S\}}$ is a shorthand notation for vector $(v^1, \ldots, v^S)$. If there exists a destination station $j' \in \{1, \ldots, S\}$ with $\mathbf{Q}^{j'} > 0$ and $\max f_{\text{rank}}^{j'}(\omega^j) \neq -\infty$, then

$$\mathcal{G}(\omega, \mathbf{Q}, j) = \underset{j' \in \{1,\ldots,S\}: \mathbf{Q}^{j'} > 0}{\arg\max} \left\{ \text{Assign}(f_{\text{rank}}^{j'}(\omega^j)) \right\}_{j' \in \{1,\ldots,S\}}.$$

Otherwise,

$$\mathcal{G}(\omega, \mathbf{Q}, j) = (\text{NIL}, \text{NIL}, \text{NIL}, \text{NIL}).$$

Then, we have the following algorithm that assigns state updates $(q_{t+1}^i, \tau_{t+1}^i)$ for each vehicle.

Algorithm 1 State Updates at Time t

Input: Customer information vector ω_t and Decision variable $\mathbf{u}_t^1, \ldots, \mathbf{u}_t^S$
Initialize quota index $\mathbf{Q} = (\mathbf{Q}^1, \ldots, \mathbf{Q}^S)$ such that $\mathbf{Q}^j = \mathbf{u}_t^j$ at each station $j \in \{1, \ldots, S\}$, available customer information $\omega = \omega_t$ and stage-wise revenue function $\mathbf{R}(q_t, \tau_t, \omega_t, \mathbf{u}_t) = 0$
for $i = 1, 2, \ldots, C$ **do**
  for $j = 1, 2, \ldots, S$ **do**
    Compute $(j, j^*, \mathcal{T}_t^i, \mathcal{F}_t^i) = \mathcal{G}(\omega, \mathbf{Q}, j)$
    if $q_t^i = j$ and $\tau_t^i = 0$ and $j^* \neq$ NIL **then**
      Set $(q_{t+1}^i, \tau_{t+1}^i) = (j^*, \mathcal{T}_t^i)$, $\mathbf{R}(q_t, \tau_t, \omega_t, \mathbf{u}_t) = \mathbf{R}(q_t, \tau_t, \omega_t, \mathbf{u}_t) + \mathcal{F}_t^i$,
      Update $\mathbf{Q}^{j^*} \leftarrow \mathbf{Q}^{j^*} - 1$ in $\mathbf{Q}$, replace the corresponding element $(j, j^*, \mathcal{T}_t^i, \mathcal{F}_t^i)$ in ω with $(j, j^*, 0, \mathcal{F}_t^i)$ and **break**
    else
      Set $(q_{t+1}^i, \tau_{t+1}^i) = (q_t^i, \max(\tau_t^i - 1, 0))$
    end if
  end for
end for
return State updates: (q_{t+1}, τ_{t+1})

2.5 Revenue and Constraint Cost Functions

Recall the stage-wise revenue function from Algorithm 1, the total average revenue generated is given by

$$\lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbf{R}(q_t, \tau_t, \omega_t, \mathbf{u}_t) \right].^2$$

[2]It is an easy extension to add a penalty function to address the limits in parking spaces. Since this addition does not constitute to any major changes in our model, we omit this term in our paper for the sake of brevity.

We also impose the following set of service level agreement constraints that upper bounds the average number of customers at each station $j \in \{1, \ldots, S\}$ for rental purposes, i.e.,

$$\lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \left(\sum_{i=1}^{C} \mathbf{1}\{\mathbf{G}_t^i = j, \mathbf{T}_t^i > 0, \mathbf{F}_t^i > 0\} - \mathbf{u}_t^j \right) \right] \leq \mathbf{d}^j,$$

where $\{\mathbf{d}^j\}_{j=1}^S$ is the vector of quality-of-service thresholds, prespecified by the system operator.

Our objective for this problem is to maximize the expected revenue collected by renting vehicles while satisfying the customer service level agreement constraints at each station. The mathematical problem formulation will be introduced in the next section.

3. CMDP FORMULATION

In this section, we formalize the vehicle rebalancing problem using a CMDP. Before getting into the details, we rigorously state the assumptions for the stochastic input state ω_t.

3.1 Ergodic Markov Chain Assumption on ω_t

Recall from Section 2.1 that ω_t is a finite state ergodic Markov Chain supported on Ω. Let $\omega_0 \in \Omega$ be the initial state of ω_t. Since ω_t is an ergodic Markov chain, there exists a sequence of finite random return time T_r, for $r \in \mathbb{N}$, such that ω_{T_r} revisits ω_0 for the r-th time at time T_r. Without loss of generality, we assume the first renewal time frame starts at $t = 0$, i.e., $T_0 = 0$. Define N_t as the number of visits of ω_0 at time t. Specifically,

$$N_t = \max\{r : T_r \leq t\}. \quad (3)$$

From this sequence of return times, we define the r^{th} epoch (the interval that starts and revisits ω_0) as $[T_r, T_r + \Delta T_r]$ and the length of this epoch is defined as $\Delta T_r = T_{r+1} - T_r$. Since ω_t is an ergodic Markov chain, the sequence of $\{\Delta T_r\}_{r \in \mathbb{N}}$ is i.i.d. [31]. Let ΔT be a random variable that is equal in distribution to ΔT_r, $\forall r$. The positive recurrence assumption implies that $\lambda < \infty$. We assume that the second moment of ΔT is bounded: $\mathbb{E}\left[\Delta T^2\right] < \infty$ and define the mean return rate of state w_0 as $\lambda = 1/E(\Delta T)$.

3.2 The CMDP

A finite MDP is a quintuple $\mathbf{X} \times \Omega, \mathbf{U}, \mathbf{R}, \mathbf{P}, \mathbf{x}_0$ where:

1. The state space is defined as $\mathbf{X} \times \Omega$ where $\mathbf{X} = \{1, \ldots, S\}^C \times \{0, 1, 2, \ldots, \overline{T}\}^C$. The state at time t is given by $\mathbf{x}_t^\omega = (\mathbf{x}_t, \omega_t)$ where $\mathbf{x}_t = (q_t, \tau_t)$.
2. $\mathbf{U} = \{0, 1, \ldots, C\}^S$ is the control space and $\mathbf{u}_t$ is the action taken at time t. Also define the set of admissible controls in x^ω as $\mathbf{U}(x^\omega) \subseteq \mathbf{U}$, such that $\mathbf{U}(x^\omega) = \{\mathbf{u} \in \mathbf{U} : \mathbf{u}^j \leq \mathcal{A}^j,\ \forall j \in \{1, \ldots, S\}\}$.
3. $\mathbf{R} : \mathbf{X} \times \Omega \times \mathbf{U} \to \mathbb{R}$ is the *immediate* reward defined in Algorithm 1.
4. $\mathbf{P}_{x^\omega, y^{\omega,\prime}}^u$ is the transition probability from state x^ω to state $y^{\omega,\prime}$ when action u is applied, i.e., $\mathbf{P}_{x^\omega, y^{\omega,\prime}}^u = \mathbb{P}[\mathbf{x}_{t+1} = y | \mathbf{x}_t^\omega = x^\omega, \mathbf{u}_t = u] \mathbb{P}[\omega_{t+1} = \omega' | \omega_t = \omega]$.
5. $\mathbf{x}_0 = (q_0, \tau_0)$ and ω_0 are the initial states of $\mathbf{x}_t$ and ω_t. q_0 is the initial destination vector, which equals to the initial location vector. τ_0 is the vectors of initial travel times, which is a zero vector. ω_0 is the initial (renewal) state.

Since the reward $\mathbf{R}(\mathbf{x}_t^\omega, \mathbf{u}_t)$ and transition probabilities do not depend on time, the above model is *stationary* . Based on the above definitions, the sequence of states and actions over time constitutes a stochastic process that we will denote as $(\mathbf{x}_t^\omega, \mathbf{u}_t)$. Without loss of generality (since the model is stationary), we assume that the evolution starts at $t = 0$.

REMARK 2. *Transition probability $\mathbb{P}[\mathbf{x}_{t+1} = y|\mathbf{x}_t^\omega = x^\omega, \mathbf{u}_t = u]$ follows from the evolution of (q_t, τ_t) in Algorithm 1. In general the explicit formulation of $\mathbb{P}[\mathbf{x}_{t+1} = y|\mathbf{x}_t^\omega = x^\omega, \mathbf{u}_t = u]$ is very complicated. This partly motivates us to propose an episodic sampling algorithm for finding a near-optimal vehicle rental policy.*

REMARK 3. *The dimension of the state space are $|\Omega|(S(1 + \overline{T}))^C$ and C^S respectively. When the numbers of vehicles and stations are moderately large, the state and action spaces and the computational power of solving the CMDP grows exponentially large as well. This is known as the "curse of dimensionality". Since solving for an exact solution is impractical in many applications, we will later propose an episodic sampling algorithm for finding a near-optimal vehicle rental policy.*

Next, a CMDP extends the Markov decision problem (MDP) by introducing additional constraints. A CMDP is defined by the following elements: $\mathbf{X} \times \Omega, \mathbf{U}, \mathbf{R}, \mathbf{P}, \mathbf{x}_0, \{\mathbf{D}^j\}_{j=1}^S, \{\mathbf{d}^j\}_{j=1}^S$ where $\mathbf{X} \times \Omega, \mathbf{U}, \mathbf{R}, \mathbf{P}, \mathbf{x}_0$ are the same as above. For $j \in \{1, \ldots, S\}$,

1. $\mathbf{D}^j : \mathbf{X} \times \Omega \times \mathbf{U} \to \mathbb{R}$, is a constraint cost expressed as $\mathbf{D}^j(\mathbf{x}_t^\omega, \mathbf{u}_t) = \sum_{i=1}^C \mathbf{1}\{\mathbf{G}_t^i = j, \mathbf{T}_t^i > 0, \mathbf{F}_t^i > 0\} - \mathbf{u}_t^j$.

The optimal control of an CMDP entails the determination of a closed-loop stationary policy μ defining which action should be applied at time t in order to minimize an aggregate (sum) objective function of the immediate costs, while ensuring that the total average constraint costs defined are (in expectation) bounded by the vector of quality-of-service thresholds $\{\mathbf{d}^j\}_{j=1}^S$. This notion can be formalized as follows. A policy μ induces a stationary mass distribution[3] over the realizations of the stochastic process $(\mathbf{x}_t^\omega, \mathbf{u}_t)$. Let $\mathbf{MS}$ be the set of closed-loop, Markovian, stationary, policies $\mu : \mathbf{X} \times \Omega \to \mathbb{P}(\mathbf{U})$. It is well known that for CMDPs there is no loss of optimality in restricting the attention on policies in $\mathbf{MS}$ (instead, e.g., of also considering history-dependent or randomized policies). For more details about the existence of dominating policies, please see Proposition 4.1 in [1].

For risk-neutral optimization in CMDPs, the goal is to find an optimal policy μ^* for the following problem:

$$J^{\text{OPT}} = \text{maximize}_{\mu \in \mathbf{MS}} \quad \overline{\mathbf{R}}(\mu) \tag{4}$$

$$\text{subject to} \quad \overline{\mathbf{D}}^j(\mu) \leq \mathbf{d}^j, \forall j, \tag{5}$$

where the multi-stage reward and constraint cost functions for all $j \in \{1, \ldots, S\}$ are given by

$$\overline{\mathbf{R}}(\mu) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \mathbf{R}(\mathbf{x}_t^\omega, \mathbf{u}_t) \mid \mathbf{x}_0^\omega \ \mathbf{u}_t \sim \mu\right],$$

$$\overline{\mathbf{D}}^j(\mu) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum\nolimits_{t=0}^{T-1} \mathbf{D}^j(\mathbf{x}_t^\omega, \mathbf{u}_t) \mid \mathbf{x}_0^\omega \ \mathbf{u}_t \sim \mu\right].$$

Previously, we stated that for the average reward CMDP problem (4) to (5), if this problem is feasible, Theorem 4.1 of [1] implies there exists an optimal stationary Markovian policy μ^*. Because this system experiences regular renewals, the performance of any stationary policy can be characterized by ratios of expectations over one renewal frame. Define

$$\widehat{\mathbf{R}}_r(\mu) = \sum_{h=T_r}^{T_{r+1}-1} \mathbf{R}(\mathbf{x}_h^\omega, \mathbf{u}_h) \mid \mathbf{x}_{T_r}^\omega, \mathbf{u}_h \sim \mu$$

$$\widehat{\mathbf{D}}_r^j(\mu) = \sum_{h=T_r}^{T_{r+1}-1} \mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h) \mid \mathbf{x}_{T_r}^\omega, \mathbf{u}_h \sim \mu$$

as the revenue function and constraint cost function at the r^{th} renewal time interval induced by policy μ. By recalling the renewal time-frame r and the renewal time T_r for the renewal input process ω_t and the feasibility assumption of the CMDP, the following expression holds:

$$\mathbb{E}\left[\widehat{\mathbf{R}}_r(\mu^*)\right] = \frac{J_{\text{OPT}}}{\lambda}, \quad \mathbb{E}\left[\widehat{\mathbf{D}}_r^j(\mu^*)\right] \leq \frac{\mathbf{d}^j}{\lambda}, \forall j \in \{1, \ldots, S\}$$

This is because, by the Renewal Cost Theorem (Theorem 3.6.1, [31]), $\widehat{\mathbf{R}}_r(\mu)$ is an i.i.d. process, $\forall r$, and one obtains

$$\lambda\mathbb{E}\left[\widehat{\mathbf{R}}_r(\mu^*)\right] = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\sum_{r=0}^{N_t-1} \widehat{\mathbf{R}}_r(\mu^*)\right] = \overline{\mathbf{R}}(\mu^*).$$

Analogous arguments can also be applied to show that for the constraint cost functions,

$$\lambda\mathbb{E}\left[\widehat{\mathbf{D}}_r^j(\mu^*)\right] = \overline{\mathbf{D}}^j(\mu^*).$$

In cases where the CMDP is stationary and has finite state and action spaces, one can solve for the optimal control policies using the convex analytic approach and finite dimensional linear programming (see Theorem 4.3 in [1] for further details). However, when the state and action spaces are exponentially large (especially when the size of C and S are large), any direct applications of CMDP methods from [1] are numerically and computationally intractable. In the next section, we will introduce an approximation algorithm to solve the optimization problem (4) to (5) using unconstrained stochastic control methods.

4. EXACT SOLUTION TO CMDP

Since the vehicle sharing problem aims at maximizing revenue subjected to service level constraints in each station, we have just shown that this problem can be modeled as a CMDP with average objective and constraint cost functions. However, one of the biggest challenges to solving CMDPs with convex analytic methods from [1] is handling large state and action spaces, because the size of these spaces and computational effort grow exponentially with the number of the dimensions. On the other hand, approximating a CMDP with reinforcement learning makes use of its Lagrangian formulation [6], [29], for which the convergence analysis is more complicated than its unconstrained MDP counterpart.

Inspired by the Lyapunov optimization [26], a technique that stabilizes a queueing network and minimizes the time average of a network penalty function, we propose a sequential algorithm for solving the CMDP in (4) to (5). By constructing an augmented state, so called the "virtual queue", we will later show that an optimal policy to the above CMDP can be iteratively generated by solving an unconstrained stochastic shortest path problem.

4.1 State Augmentation and Stability Conditions

[3]Such mass distribution not only exists, but can be explicitly computed.

In order to simplify the following analysis, define the following short-hand notation:

$$\Delta\mathbf{D}^j(\mathbf{x}^\omega, \mathbf{u}) = \mathbf{D}^j(\mathbf{x}^\omega, \mathbf{u}) - \mathbf{d}^j.$$

For any time slot $t \geq 0$, define the augmented state variables $\mathbf{z}_t^j$, for $j \in \{1, \ldots, S\}$ as follows:

$$\mathbf{z}_{t+1}^j = \max\left(\mathbf{z}_t^j + \Delta\mathbf{D}^j(\mathbf{x}_t^\omega, \mathbf{u}_t), 0\right), \tag{6}$$

where $\mathbf{z}_0^j$ is a pre-specified initial condition for the virtual queue. We also define $\mathbf{z}_t = (\mathbf{z}_t^1, \ldots, \mathbf{z}_t^S)$ as a vector of augmented state variables at time t. Induced by an arbitrary policy μ, $\mathbf{z}_t^j$, for $j \in \{1, \ldots, S\}$, become stochastic processes. Modified from Definition 2.1 of [26], we have the following definition of mean rate stability.

DEFINITION 2. *Recall T_r as the initial time of the r^{th} renewal time frame. A discrete time process Λ_t, induced by policy μ, is mean rate stable if*

$$\lim_{r\to\infty} \frac{1}{r}\mathbb{E}[\Lambda_{T_r} \mid \mu] = 0.$$

It can be easily shown that if $\mathbf{z}_t^j$ is mean rate stable, then it implies the constraint in (5) is satisfied (feasibility).

To see this, from equation (6), one can easily see that $\mathbf{z}_{t+1}^j \geq \mathbf{z}_t^j + \Delta\mathbf{D}^j(\mathbf{x}_t^\omega, \mathbf{u}_t)$ for all $j \in \{1, \ldots, S\}$. Now we generate the state trajectory of $\mathbf{z}^j$ based on its dynamics in (6), induced by policy μ. Recall $E(\Delta T) = 1/\lambda < \infty$ and $N_t = \max\{r : T_r \leq t\}$. By a telescopic sum over $t \in \{T_r, \ldots, T_{r+1} - 1\}$ and $r \in \{0, 1, \ldots, N_t - 1\}$, this implies for any $j \in \{1, \ldots, S\}$,

$$\mathbf{z}_{T_{N_t}}^j - \mathbf{z}_0^j = \sum_{r=0}^{N_t-1} \mathbf{z}_{j,T_{r+1}} - \mathbf{z}_{T_r}^j \geq \sum_{r=0}^{N_t-1}\sum_{h=T_r}^{T_{r+1}-1} \Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h).$$

By taking expectation with respect to the state trajectory of $\mathbf{z}^j$ and policy μ, dividing by N_t/λ and letting $t \to \infty$ on both sides, one obtains for each j,

$$\lim_{t\to\infty} \frac{\mathbb{E}[\mathbf{z}_{T_{N_t}}^j \mid \mu] - \mathbf{z}_0^j}{N_t/\lambda} \geq \lim_{t\to\infty} \frac{1}{N_t/\lambda}\mathbb{E}\left[\sum_{r=0}^{N_t-1}\left(\widehat{\mathbf{D}}_r^j(\mu) - \frac{\mathbf{d}^j}{\lambda}\right)|\mathbf{x}_0^\omega\right].$$

Recall $T_{N_t} \to \infty$ when both N_t and t tend to infinity. Since $\mathbf{z}_0^j$ is a bounded initial condition of $\mathbf{z}_t^j$ for all $j \in \{1, \ldots, S\}$, if the stochastic process $\mathbf{z}_t^j$ is mean rate stable, the left side of the above inequality becomes zero and the above expression becomes

$$\lim_{t\to\infty} \frac{1}{N_t/\lambda}\mathbb{E}\left[\sum_{r=0}^{N_t-1}\left(\widehat{\mathbf{D}}_r^j(\mu) - \frac{\mathbf{d}^j}{\lambda}\right)|\mathbf{x}_0^\omega\right] \leq 0, \ \forall j.$$

The Elementary Renewal Theory (Theorem 3.3.4, [**?**]) implies that

$$\lim_{t\to\infty} \frac{t}{N_t/\lambda} = 1 \text{ almost surely.}$$

For $T_{N_t} + 1 \leq t \leq T_{N_t+1}$ and $|\Delta\mathbf{D}^j(\mathbf{x}_t^\omega, \mathbf{u}_t)| \leq M$, one obtains at $j \in \{1, \ldots, S\}$

$$\begin{aligned} 0 &\leq \lim_{t\to\infty} \frac{1}{N_t/\lambda}\left|\mathbb{E}\left[\sum_{h=T_{N_t}+1}^{t} \Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h)|\mathbf{x}_0^\omega, \mathbf{u}_h \sim \mu\right]\right| \\ &\leq \lim_{t\to\infty} \frac{M\lambda}{N_t} = 0. \end{aligned}$$

4.2 Algorithm $\mathcal{OPT}^{\mathrm{OL}}$

In this section, we provide the exact solution algorithm. Before getting to the main result, define the r^{th} epoch (the interval that starts and revisits ω_0) as $[T_r, T_r + \Delta T_r]$ and the length of this epoch is defined as $\Delta T_r = T_{r+1} - T_r$. Since ω_t is an ergodic Markov chain, the sequence of $\{\Delta T_r\}_{r\in\mathbb{N}}$ is i.i.d. [31]. Let ΔT be a random variable that is equal in distribution to ΔT_r, $\forall r$. The positive recurrence assumption implies that $\lambda < \infty$. We assume that the second moment of ΔT is bounded: $\mathbb{E}\left[\Delta T^2\right] < \infty$ and define the mean return rate of state w_0 as $\lambda = 1/\lambda$.

Now, we state the following algorithm that performs a policy update at the beginning of each renewal time frame.

- Initialize: Set $r \leftarrow 0$. For $r \in \{0, 1, \ldots\}$, construct the regularization function W_r based on the following set of rules:

$$\begin{aligned} &0 < W_0, \ W_r \leq W_{r+1}, \ \forall r \geq 0, \\ &\lim_{\mathfrak{R}\to\infty} \frac{W_{\mathfrak{R}}}{\mathfrak{R}} = 0, \ \lim_{\mathfrak{R}\to\infty} \sum_{r=0}^{\mathfrak{R}-1} \frac{1}{\mathfrak{R}W_r} = 0. \end{aligned} \tag{7}$$

- Step 1: At the beginning of the r^{th} renewal time frame, solve the following stochastic shortest path problem:

  MDP problem $\mathcal{SP}$ —At time $t = T_r$, given initial states $\mathbf{x}_{T_r}^\omega \in \mathbf{X} \times \Omega$ and $\mathbf{z}_{T_r}$, solve the following stochastic shortest path problem:

  $$\mathcal{U}_r^* \in \arg\max_\mu \mathbb{E}\left[\sum_{h=T_r}^{T_{r+1}-1} \mathbf{R}_r^{\mathrm{OL}}(\mathbf{x}_h^\omega, \mathbf{u}_h)|\mathbf{x}_{T_r}^\omega, \mathbf{u}_h \sim \mu\right]$$

  where

  $$\mathbf{R}_r^{\mathrm{OL}}(\mathbf{x}_h^\omega, \mathbf{u}_h) = \sum_{i=1}^{C} \mathbf{R}(\mathbf{x}_h^\omega, \mathbf{u}_h) - \sum_{j=1}^{S} \frac{\mathbf{z}_{T_r}^j}{W_r} \Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h).$$

  Obtain the realization of the next renewal time T_{r+1} and set the subsequence of online policy μ^{OL}, from $t = T_r$ to $t = T_{r+1}$, as follows,

  $$\{\mu_{T_r}^{\mathrm{OL}}, \ldots, \mu_{T_{r+1}-1}^{\mathrm{OL}}\} = \{\mathcal{U}_r^*, \ldots, \mathcal{U}_r^*\}.$$

- Step 2: During the course of the frame, update virtual queues $\mathbf{z}_t^j$ for $j \in \{1, \ldots, S\}$ at every time slot by (6) and update state $\mathbf{x}_t^\omega$ of the MDP. At the end of the frame, go back to step 1 and set $r \leftarrow r + 1$.

Notice that the expectation operator in problem $\mathcal{SP}$ is taken over the state process $\mathbf{x}_h^\omega$, induced by policy μ and the i.i.d. random variable $\Delta T = T_{r+1} - T_r$ (renewal interval). From the ergodic Markov chain assumption and renewal process theories [31], one can calculate the distribution of ΔT when transition probability $\mathbf{P}_{x^\omega, y^{\omega,\prime}}^u$ and policy μ are known. While calculating the distribution of ΔT may not be straightforward, optimal policy of problem $\mathcal{SP}$ can also be found using dynamic programming by defining a stopping set corresponding to the next renewal time T_{r+1}. More details will be discussed in Section 4.5 and Section 5.

Comparing the optimization problem $\mathcal{SP}$ to the CMDP in (4) to (5), one notices that instead of directly optimizing the reward at the current renewal time frame, the online algorithm optimizes a weighted combination of the stage-wise reward and a *Lyapunov function derived regularization term*. We will later show that by solving problem $\mathcal{SP}$ and following the update rules of the regularization term W_r at each episode, μ^{OL} is an optimal policy for the CMDP in (4) to (5).

REMARK 4. *The analysis of algorithm $\mathcal{OPT}^{\mathrm{OL}}$ is similar to the drift-plus-penalty method, a technique in in Lyapunov optimization which is mainly used in wireless network optimization [37], [14] and routing [38] problems.*

4.3 Optimality

In this section, we analyze the performance of μ^{OL} and show that this policy induces a multi-stage reward that equals to the optimal solution of CMDP in (4) to (5). In other words, by assuming $\mathbf{z}_t^j$ is mean rate stable in this section, μ^{OL} is an optimal solution to the CMDP in (4) to (5).

THEOREM 3 (PERFORMANCE). *The control policy μ^{OL} is optimal, i.e., $J(\mu^{\mathrm{OL}}) = J^{\mathrm{OPT}}$ almost surely.*

Proof. **C**onsider the weighted quadratic Lyapunov function

$$L_r(\mathbf{z}_t) = \sum_{j=1}^{S} \frac{(\mathbf{z}_t^j)^2}{2W_r}.$$

At $t \in \{T_r, \ldots, T_{r+1}-1\}$, let the Lyapunov drift be

$$\Delta_r(\mathbf{z}_t) = \mathbb{E}\left[L_r(\mathbf{z}_{t+1}) - L_r(\mathbf{z}_t)|\mathbf{z}_t\right].$$

Recall the dynamics of $\mathbf{z}_t^j$ in (6) for $i \in \{1, \ldots, C\}$. By expanding the Lyapunov drift term, one obtains

$$\begin{aligned}
\Delta_r(\mathbf{z}_t) &= \mathbb{E}\left[\sum_{j=1}^{S} \frac{(\mathbf{z}_{t+1}^j)^2}{2W_r} - \frac{(\mathbf{z}_t^j)^2}{2W_r}|\mathbf{z}_t\right] \\
&\leq \mathbb{E}\left[\sum_{j=1}^{S} \frac{1}{2W_r}\left(\Delta\mathbf{D}^j(\mathbf{x}_t^\omega, \mathbf{u}_t)\right)^2 + \sum_{j=1}^{S} \frac{\mathbf{z}_t^j}{W_r}\Delta\mathbf{D}^j(\mathbf{x}_t^\omega, \mathbf{u}_t) \mid \mathbf{x}_t^\omega, \mathbf{z}_t\right].
\end{aligned}$$

Consider the drift-plus-penalty term

$$\Delta_r(\mathbf{z}_t) - \mathbb{E}[\mathbf{R}(\mathbf{x}_t^\omega, \mathbf{u}_t) \mid \mathbf{z}_t].$$

Recall from expression (7) that $W_{r+1} \geq W_r$ and $r \geq 0$. By substituting $t = T_r$, taking a telescoping sum from $h \in \{T_r, \ldots, T_{r+1}-1\}$ and conditional expectation with respect to $\mathbf{z}_{T_r}$, it follows that with arbitrary admissible control actions $(\mathbf{u}_{T_r}, \ldots, \mathbf{u}_{T_{r+1}-1})$,

$$\begin{aligned}
&\mathbb{E}\left[L_{r+1}(\mathbf{z}_{T_{r+1}}) - L_r(\mathbf{z}_{T_r}) - \sum_{h=T_r}^{T_{r+1}-1} \mathbf{R}(\mathbf{x}_h^\omega, \mathbf{u}_h)|\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}\right] \\
\leq &\mathbb{E}\left[L_r(\mathbf{z}_{T_{r+1}}) - L_r(\mathbf{z}_{T_r}) - \sum_{h=T_r}^{T_{r+1}-1} \mathbf{R}(\mathbf{x}_h^\omega, \mathbf{u}_h)|\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}\right] \\
\leq &\mathbb{E}\Bigg[\sum_{h=T_r}^{T_{r+1}-1} -\mathbf{R}(\mathbf{x}_h^\omega, \mathbf{u}_h) + \sum_{j=1}^{S} \frac{\mathbf{z}_h^j}{W_r}\Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h) \\
&\qquad + \sum_{j=1}^{S} \frac{1}{2W_r}\left(\Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h)\right)^2 |\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}\Bigg]. \qquad (8)
\end{aligned}$$

By defining the following short-hand notation:

$$\Delta L_r = L_{r+1}(\mathbf{z}_{T_{r+1}}) - L_r(\mathbf{z}_{T_r}),$$

the above expression implies

$$\begin{aligned}
&\mathbb{E}\left[\Delta L_r - \widehat{\mathbf{R}}_r(\mu^{\mathrm{OL}})|\mathbf{z}_{T_r}\right] \\
\leq &\mathbb{E}\Bigg[-\widehat{\mathbf{R}}_r(\mu^{\mathrm{OL}}) + \sum_{h=T_r}^{T_{r+1}-1} \sum_{j=1}^{S} \frac{\mathbf{z}_h^j}{W_r}\Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h) \\
&\quad + \sum_{h=T_r}^{T_{r+1}-1} \sum_{j=1}^{S} \frac{1}{2W_r}\left(\Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h)\right)^2 |\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}, \mu^{\mathrm{OL}}\Bigg]. \qquad (9)
\end{aligned}$$

Now, by expanding the stage-wise reward function $\mathbf{r}_{\mathrm{OL},r}$, we exploit the structure in the first part of the right side in inequality (9), i.e.,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=T_r}^{T_{r+1}-1} \sum_{j=1}^{S} \frac{\mathbf{z}_h^j}{W_r}\Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h)|\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}, \mu^{\mathrm{OL}}\right] \\
= &\underbrace{\mathbb{E}\left[\sum_{h=T_r}^{T_{r+1}-1} \sum_{j=1}^{S} \frac{(\mathbf{z}_h^j - \mathbf{z}_{T_r}^j)}{W_r}\Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h)|\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}, \mu^{\mathrm{OL}}\right]}_{\mathcal{B}_1} \\
&+ \underbrace{\sum_{j=1}^{S} \frac{\mathbf{z}_{T_r}^j}{W_r}\mathbb{E}\left[\widehat{\mathbf{D}}_r^j(\mu^{\mathrm{OL}}) - \mathbf{d}^j\right]}_{\mathcal{B}_2}.
\end{aligned}$$

For expression $\mathcal{B}_1$, by triangular inequality one obtains

$$\begin{aligned}
\mathcal{B}_1 \leq &\mathbb{E}\left[\sum_{h=T_r}^{T_{r+1}-1} \sum_{j=1}^{S} \frac{|\mathbf{z}_h^j - \mathbf{z}_{T_r}^j|}{W^j}\left|\Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h)\right| |\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}, \mu^{\mathrm{OL}}\right] \\
\leq &M^2 \sum_{j=1}^{S} \frac{1}{W_r}\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-1} \sum_{\tau'=1}^{t-T_r} 1|\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}, \mu^{\mathrm{OL}}\right] \\
\leq &\frac{M^2}{2}\frac{S}{W_r}\mathbb{E}[\Delta T(\Delta T - 1)], \qquad (10)
\end{aligned}$$

where the second inequality is due to the fact that

$$|\mathbf{z}_{t_2}^j - \mathbf{z}_{t_1}^j| \leq \sum_{h=1}^{t_2-t_1} M, \ \forall t_1, t_2 > 0, \qquad (11)$$

and the last inequality is from the fact that the inter-arrival time ΔT is an i.i.d random variable at each renewal time frame r. The inequality in (11) holds because the largest magnitude change in $\mathbf{z}_t^j$ per time slot is upper bounded by $M \geq |\Delta\mathbf{D}^j(\mathbf{x}_t^\omega, \mathbf{u}_t)|, j\{1, \ldots, S\}$, $t \geq 0$.

On the other hand, based on the definitions of the constraint cost function and the stage-wise cost upper bound M, the second part of the right side in inequality (9) can be written as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=T_r}^{T_{r+1}-1} \sum_{j=1}^{S} \frac{1}{2W_r}\left(\Delta\mathbf{D}^j(\mathbf{x}_h^\omega, \mathbf{u}_h)\right)^2 |\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}, \mu^{\mathrm{OL}}\right] \\
\leq &\mathbb{E}\left[\sum_{h=T_r}^{T_{r+1}-1} \frac{M^2}{2} \sum_{j=1}^{S} \frac{1}{W_r}|\mathbf{x}_{T_r}^\omega, \mathbf{z}_{T_r}, \mu^{\mathrm{OL}}\right] \leq \frac{M^2}{2}\frac{S}{\lambda W_r} \qquad (12)
\end{aligned}$$

where the last inequality is also from the fact that the inter-arrival time ΔT is an i.i.d random variable at each renewal time frame r.

Inserting the results of (10) and (12) into expression (9), one

obtains

$$\mathbb{E}\left[\Delta L_r - \widehat{\mathbf{R}}_r(\mu^{\mathrm{OL}}) \mid \mathbf{z}_{T_r}\right] \leq \frac{M^2}{2}\frac{S}{W_r}\mathbb{E}\left[(\Delta T)^2\right] - \mathbb{E}\left[\widehat{\mathbf{r}}_{\mathrm{OL},r}(\mu^{\mathrm{OL}})\right]$$
$$\leq \frac{M^2}{2}\frac{S}{W_r}\mathbb{E}\left[(\Delta T)^2\right] - \mathbb{E}\left[\widehat{\mathbf{r}}_{\mathrm{OL},r}(\mu^*)\right] \tag{13}$$

For the second inequality, it is clear that minimizing the right hand side of the above expression over $\mathbf{u}_h$, $h \in \{T_r, \ldots, T_{r+1}-1\}$, is equivalent to maximizing the objective of $\mathcal{OPT}_{\mathrm{OL}}$, and given that μ^*, the stationary optimal policy of problem (4) to (5) is feasible for $\mathcal{OPT}_{\mathrm{OL}}$. Now consider the following expression

$$\mathcal{B}_2 := \sum_{j=1}^{S}\frac{\mathbf{z}^j_{T_r}}{W_r}\mathbb{E}\left[\widehat{\mathbf{D}}^j_r(\mu^*) - \mathbf{d}^j\right].$$

By the definitions of $W_r > 0$ and $\mathbf{z}^j_t \geq 0$, feasibility of μ^* implies

$$\mathcal{B}_2 \leq 0. \tag{14}$$

Therefore, expression (13) implies

$$\mathbb{E}\left[\Delta L_r - \widehat{\mathbf{R}}_r(\mu^{\mathrm{OL}})|\mathbf{z}_{T_r}\right] \leq \frac{M^2}{2}\frac{S}{W_r}\mathbb{E}\left[(\Delta T)^2\right] - \mathbb{E}[\widehat{\mathbf{R}}_r(\mu^*)|\mathbf{z}_{T_r}].$$

By taking expectation in the above expression with respect to $\mathbf{z}_{T_r}$ and using the optimality condition of μ^*, this expression becomes

$$\mathbb{E}\left[\Delta L_r - \widehat{\mathbf{R}}_r(\mu^{\mathrm{OL}}) \mid \mathbf{z}_{T_r}\right] \leq \frac{M^2}{2}\frac{S}{W_r}\mathbb{E}\left[(\Delta T)^2\right] - \mathbb{E}[\widehat{\mathbf{R}}_r(\mu^*) \mid \mathbf{z}_{T_r}]$$
$$= \frac{M^2}{2}\frac{S}{W_r}\mathbb{E}\left[(\Delta T)^2\right] - \frac{J_{\mathrm{OPT}}}{\lambda}.$$

Recall the definition $N_t = \max\{r : T_r \leq t\}$ from (3). By a telescoping sum over $r = 0, \ldots, N_t - 1$ and dividing both sides by N_t/λ, the above expression becomes

$$\frac{1}{N_t/\lambda}\mathbb{E}\left[L_{N_t}(\mathbf{z}(T_{N_t})) - L_0(\mathbf{z}_0) - \sum_{r=0}^{N_t-1}\widehat{\mathbf{R}}_r(\mu^{\mathrm{OL}})|\mathbf{x}^\omega_0, \mathbf{z}_0\right]$$
$$\leq \frac{M^2}{2}\frac{1}{N_t}\sum_{r=0}^{N_t-1}\frac{S\lambda}{W_r}\mathbb{E}\left[(\Delta T)^2\right] - J_{\mathrm{OPT}}. \tag{15}$$

Recall $T_r \to \infty$ as $r \to \infty$ and $N_t = \max\{r : T_r \leq t\}$. Then, as $t \to \infty$, we get $N_t \to \infty$, this implies that

$$\lim_{t\to\infty}\frac{\mathbb{E}\left[L_0(\mathbf{z}_0) \mid |\mathbf{x}^\omega_0, \mathbf{z}_0, \mu^{\mathrm{OL}}\right]}{N_t/\lambda} = 0.$$

By taking the limit on $t \to \infty$ and noticing that $L_{N_t}(\mathbf{z}(T_{N_t})) \geq 0$, expression (15) implies

$$\lim_{t\to\infty}\frac{\mathbb{E}\left[\sum_{h=0}^{t}\mathbf{R}(\mathbf{x}^\omega_h, \mathbf{u}_h) - \sum_{h=T_{N_t}+1}^{t}\mathbf{R}(\mathbf{x}^\omega_h, \mathbf{u}_h) \mid \mathbf{x}^\omega_0, \mu^{\mathrm{OL}}\right]}{N_t/\lambda}$$
$$\geq \lim_{t\to\infty}\frac{\mathbb{E}\left[\sum_{h=0}^{t}\mathbf{R}(\mathbf{x}^\omega_h, \mathbf{u}_h) - \sum_{h=T_{N_t}+1}^{t}(\mathbf{R}(\mathbf{x}^\omega_h, \mathbf{u}_h))^+ \mid \mathbf{x}^\omega_0, \mu^{\mathrm{OL}}\right]}{N_t/\lambda}$$
$$\geq J_{\mathrm{OPT}} - \frac{\lambda M^2}{2}\sum_{j=1}^{S}\frac{1}{W^j}\mathbb{E}\left[(\Delta T)^2\right]. \tag{16}$$

[4]Now, since $T_{N_t} + 1 \leq t \leq T_{N_t+1}$, and $T_{N_t+1} - T_{N_t} = \Delta T_{N_t}$ where ΔT_{N_t} equals to ΔT in distribution, by noting that $\mathbb{E}[\Delta T]$ and $\mathbb{E}\left[\Delta T^2\right] < \infty$ for each time slot t, one easily obtains

$$0 \leq \lim_{t\to\infty}\frac{1}{N_t/\lambda}\mathbb{E}\left[\sum_{h=T_{N_t}+1}^{t}(\mathbf{R}(\mathbf{x}^\omega_h, \mathbf{u}_h))^+ \mid \mathbf{x}^\omega_0, \mu^{\mathrm{OL}}\right]$$
$$\leq \lim_{t\to\infty}\frac{\lambda\overline{\mathcal{F}}C}{N_t} = 0,$$

where $\overline{\mathcal{F}}$ is the maximum fare collected from signing a rental contract (see Section 2.4). Next, recall from the Elementary Renewal Theory (Theorem 3.3.4, [**?**]) that $\lim_{t\to\infty} t/(N_t/\lambda) = 1$ almost surely. By combining all previous arguments, one further obtains the following expression:

$$\overline{\mathbf{R}}(\mu^{\mathrm{OL}}) \geq J_{\mathrm{OPT}} - \mathbb{E}\left[(\Delta T)^2\right]\frac{M^2}{2}\lambda S\left(\lim_{t\to\infty}\frac{1}{N_t}\sum_{r=0}^{N_t-1}\frac{1}{W_r}\right) \tag{17}$$

almost surely. By the properties in (7), one obtains

$$\lim_{t\to\infty}\sum_{r=0}^{N_t-1}1/(N_t W_r) = 0.$$

Therefore, the above expression implies that $J(\mu^{\mathrm{OL}}) \geq J_{\mathrm{OPT}}$. On the other hand, we will later show that by the mean rate stability property of the augmented state $\mathbf{z}^j_t$ in (6), for $j \in \{1, \ldots, S\}$, μ^{OL} is a feasible control policy to problem (4) to (5), which further implies $J(\mu^{\mathrm{OL}}) \leq J_{\mathrm{OPT}}$. Therefore one concludes that μ^{OL} is an optimal policy by combining both arguments. ■

4.4 Feasibility

In order to complete the proof on the optimality of μ^{OL}, in this section we will show that $\mathbf{z}^j_t$ is mean rate stable and thus μ^{OL} is a feasible policy to the CMDP in (4) to (5).

THEOREM 4 (FEASIBILITY). *The augmented state $\mathbf{z}^j_t$ in* (6)*, for $j \in \{1, \ldots, S\}$ is mean rate stable. This further implies when control policy μ^{OL} is executed, constraint* (5) *is satisfied.*

Proof. **R**ecall the drift-plus-penalty inequality in (8) for any admissible control actions $(\mathbf{u}_{T_r}, \ldots, \mathbf{u}_{T_r+\Delta T-1})$. Similar to the proof in Theorem 3, it is clear that minimizing the right hand side of the above expression over $\mathbf{u}_h$, $h \in \{T_r, \ldots, T_r + \Delta T - 1\}$, is equivalent to minimizing the objective of $\mathcal{OPT}_{\mathrm{OL}}$. Also recall that μ^* is feasible for $\mathcal{OPT}_{\mathrm{OL}}$. By recalling ΔL_r as the short-hand for $L_{r+1}(\mathbf{z}_{T_{r+1}}) - L_r(\mathbf{z}_{T_r})$, this implies

$$\mathbb{E}\left[\Delta L_r - \widehat{\mathbf{R}}_r(\mu_{\mathrm{OL}}) \mid \mathbf{z}_{T_r}\right] \leq \frac{M^2}{2}\frac{S}{W_r}\mathbb{E}\left[(\Delta T)^2\right] - \mathbb{E}\left[\widehat{\mathbf{R}}_r(\mu^*)\right]$$

Recall from Section 2.4 and 2.5 that $\overline{\mathcal{F}}$ is the maximum fare collected from signing a rental contract and $\mathbf{c}_{\text{penalty},j}(C)$ is the maximum penalty incurred from parking violation at the j^{th} station. Since the inter-arrival time ΔT_r is an i.i.d. random variable for each $r \in \{0, 1, 2, \ldots\}$ and $\sum_{j=1}^{S}\mathbf{c}_{\text{penalty},j}(C) \leq \mathbf{R}(\mathbf{x}^\omega_h, \mathbf{u}_h) \leq \overline{\mathcal{F}}C$ surely, one can write

$$\mathbb{E}\left[\widehat{\mathbf{R}}_r(\mu^*)\right] - \mathbb{E}\left[\widehat{\mathbf{R}}_r(\mu_{\mathrm{OL}})\right] \leq \mathbb{E}[\Delta T]\left(\overline{\mathcal{F}}C + \sum_{j=1}^{S}\mathbf{c}_{\text{penalty},j}(C)\right).$$

Define $\overline{\mathbf{R}} = (\overline{\mathcal{F}}C + \sum_{j=1}^{S}\mathbf{c}_{\text{penalty},j}(C))$ as the upper bound of $\max_{x^\omega, x^{\omega,\prime}, u, u'}|\mathbf{R}(x^\omega, u) - \mathbf{R}(x^{\omega,\prime}, u')|$, and combining with previous results, one obtains

$$\mathbb{E}\left[\Delta L_r|\mathbf{x}^\omega_{T_r}, \mathbf{z}_{T_r}, \mu_{\mathrm{OL}}\right] \leq \frac{M^2}{2}\frac{S}{W_r}\mathbb{E}\left[(\Delta T)^2\right] + \frac{\overline{\mathbf{R}}}{\lambda}.$$

[4]Notice that $(f)^+ = \max(f, 0)$ represents the positive part of f.

By taking expectation with respect to $\mathbf{z}_{T_r}$ for which the trajectories are induced by policy $\mu_{\rm OL}$, and using a telescoping sum over $r = 0, \ldots, \mathfrak{R}-1$, the above expression becomes

$$\mathbb{E}\left[L_{\mathfrak{R}}(\mathbf{z}(T_{\mathfrak{R}})) - L_0(\mathbf{z}_0) \,|\, \mu_{\rm OL}\right] \leq \sum_{r=0}^{\mathfrak{R}-1} \frac{M^2}{2} \frac{S\mathbb{E}\left[(\Delta T)^2\right]}{W_r} + \mathfrak{R}\frac{\overline{\mathbf{R}}}{\lambda}. \tag{18}$$

On the other hand, by Cauchy-Schwarz inequality and using the fact that every elements in $\mathbf{z}(T_{\mathfrak{R}})$ is non-negative, one obtains

$$\mathbb{E}\left[L_{\mathfrak{R}}(\mathbf{z}(T_{\mathfrak{R}})) \mid \mu_{\rm OL}\right] \geq \sum_{j=1}^{S} \frac{1}{2} \frac{\left(\mathbb{E}\left[\mathbf{z}^j(T_{\mathfrak{R}})\right) \mid \mu_{\rm OL}\right])^2}{W_{\mathfrak{R}}}.$$

For each $j \in \{1, \ldots, S\}$, by substituting this inequality to expression (18), dividing by $\mathfrak{R}^2/W_{\mathfrak{R}}$ and taking square-root on both sides, we have the following set of inequalities

$$\frac{\mathbb{E}\left[\mathbf{z}^j(T_{\mathfrak{R}}) \mid \mu_{\rm OL}\right]}{\mathfrak{R}} \leq \sqrt{\frac{W_{\mathfrak{R}}}{\mathfrak{R}^2} \sum_{r=0}^{\mathfrak{R}-1} \frac{S}{W_r} M^2 \mathbb{E}\left[(\Delta T)^2\right] + \frac{W_{\mathfrak{R}} L_0(\mathbf{z}_0)}{\mathfrak{R}^2} + \frac{2W_{\mathfrak{R}}\overline{\mathbf{R}}}{\lambda\mathfrak{R}}},$$

for every $j \in \{1, \ldots, S\}$. Then the properties of the regularization function in (7) imply $\lim_{\mathfrak{R}\to\infty} \frac{W_{\mathfrak{R}}}{\mathfrak{R}} = 0$, $\lim_{\mathfrak{R}\to\infty} \sum_{r=0}^{\mathfrak{R}-1} \frac{1}{\mathfrak{R}W_r} = 0$, and the above expressions further imply that as $\mathfrak{R} \to \infty$, the augmented state $\mathbf{z}_t^j$ in (6), for $j \in \{1, \ldots, S\}$ is mean rate stable. ■

4.5 Bellman Equation to Problem $\mathcal{SP}$

Without loss of generality, we analyze problem $\mathcal{SP}$ at the zeroth renewal frame, i.e., $r = 0$ with start time $t = T_0 = 0$. Generalizations to cases with $r > 0$ is straight-forward and is omitted for the sake of brevity. The virtual queue backlogs $\mathbf{z}_0^j$, for $j \in \{1, \ldots, S\}$ and the initial state $\mathbf{x}_0^\omega = (\mathbf{x}_0, \omega_0) \in \mathbf{X} \times \Omega$ are given. Recall the random renewal interval size as ΔT and next renewal time $T_1 = T_0 + \Delta T = \Delta T$. Based on the renewal process assumption, we know that with probability one, the stochastic system states ω_t is going to re-visit ω_0 in finite time. Therefore, the renewal interval size is finite, i.e., $\Delta T < \infty$ almost surely. For problem $\mathcal{SP}$, define the state space as $(\mathbf{X} \cup \{x^T\}) \times \Omega$ and the action space as $\mathbf{U}$. Here the set of transient states is $\mathbf{X} \times \Omega$ and (x^T, ω_0) is the terminal state $\mathbf{x}_{\Delta T}^\omega = (\mathbf{x}_{\Delta T}, \omega_{\Delta T})$. From the above arguments, we immediately have the following property for stochastic shortest path problems, i.e., every policy $\mu \in \mathbf{MS}$ satisfies the following condition:

$$\sum_{h=0}^{\infty} \mathbb{P}[\mathbf{x}_h^\omega \neq (x^T, \omega_0) | \mathbf{x}_0^\omega, \mu] < \infty, \, \forall \mathbf{x}_0^\omega \in \mathbf{X} \times \Omega. \tag{19}$$

For $x^\omega, y^{\omega,\prime} \in (\mathbf{X} \cup \{x^T\}) \times \Omega$ and $u \in \mathbf{U}$, define the reward function as

$$\mathbf{R}_{\mathcal{SP}}(x^\omega, u) = \begin{cases} \mathbf{R}_r^{\rm OL}(x^\omega, u) & \text{if } x^\omega = (x, \omega) \neq (x^T, \omega_0) \\ 0 & \text{otherwise} \end{cases},$$

and the transition probability as

$$\mathbf{P}^u_{\mathcal{SP}, x^\omega, y^{\omega,\prime}} = \begin{cases} \mathbf{P}^u_{x^\omega, y^{\omega,\prime}} & \text{if } x^\omega \neq (x^T, \omega_0) \\ \mathbf{1}\{y^\omega = (x^T, \omega_0)\} & \text{otherwise} \end{cases}.$$

By reformulating problem $\mathcal{SP}$ as a MDP, we define the Bellman operator with respect to the policy μ for any real-valued function $V : \mathbf{X} \times \Omega \to \mathbb{R}$, at any given state x^ω:

$$F_\mu[V](x^\omega) := \sum_{u \in \mathbf{U}(x^\omega)} \mu(u|x^\omega) \left(\mathbf{R}_{\mathcal{SP}}(x^\omega, u) + \sum_{y^{\omega,\prime} \in \mathbf{X}\times\Omega} \mathbf{P}^u_{x^\omega, y^{\omega,\prime}} V(y^{\omega,\prime}) \right).$$

First, the Bellman operator satisfies the following properties in [4].

PROPOSITION 5. *The Bellman operator $F_\mu[V]$ has the following properties:*

- *(Monotonicity) If $V_1(x^\omega) \geq V_2(x^\omega)$, for any $x^\omega \in \mathbf{X} \times \Omega$, then $F_\mu[V_1](x^\omega) \geq F_\mu[V_2](x^\omega)$.*
- *(Constant shift) For any $K \in \mathbb{R}$, $F_\mu[V](x^\omega) - |K| \leq F_\mu[V + K](x^\omega) \leq F_\mu[V](x^\omega) + |K|$ for any $x^\omega \in \mathbf{X} \times \Omega$.*
- *(Contraction) There exists $\kappa \in (0, 1)$ such that*

  $$\|F_\mu[V_1] - F_\mu[V_2]\|_\infty \leq \kappa \|V_1 - V_2\|_\infty,$$

  where $\|f\|_\infty = \max_{x^\omega \in \mathbf{X}\times\Omega} |f(x^\omega)|$.

We also have the following standard results from the Bellman equation of stochastic shortest path problems [4].

THEOREM 6 (BELLMAN EQUALITY). *For any policies $\mu(\cdot|\cdot) \in \mathbf{MS}$, the associated reward function*

$$V_\mu(x^\omega) = \mathbb{E}\left[\sum_{h=0}^{\Delta T - 1} \mathbf{R}_{\mathcal{SP}}(\mathbf{x}_h^\omega, \mathbf{u}_h) \mid \mathbf{x}_0^\omega = x^\omega, \mu\right],$$

at any $x^\omega \in \mathbf{X} \times \Omega$ satisfies

$$\lim_{N\to\infty} F_\mu^N[V](x^\omega) = V_\mu(x^\omega), \, \forall x^\omega \in \mathbf{X} \times \Omega$$

for any initial value function $V : \mathbf{X} \times \Omega \to \mathbb{R}$. Furthermore, the function $\{V_\mu(x^\omega)\}_{x^\omega \in \mathbf{X}\times\Omega}$ is a unique solution to fixed point equation $F_\mu[V](x^\omega) = V(x^\omega)$, for any $x^\omega \in \mathbf{X} \times \Omega$.

Based on methods in dynamic programming such as policy iteration, one can solve for the optimal policy $\mu^{\rm OL}$ based on Bellman equality (see Ch.3 in volume 2 of [4] for details). However, this is still challenging because the state and action spaces in problem $\mathcal{SP}$ are large and the computational effort grow exponentially with the number of the dimensions. In the next section, we will provide methods to approximate "good" policy for this problem.

5. APPROXIMATION TO PROBLEM $\mathcal{SP}$

A natural and venerable way of approximating problem $\mathcal{SP}$ when the state and action spaces are large is to approximate the value function and policy parametrically using reinforcement learning such as policy gradient [34], [16] and actor critic [20], [7]. In these methods, the policy is taken to be an arbitrary differentiable function of a parameter vector, namely θ, and we would like to update the policy parameter in the descent direction with respect to the gradient of the objective function. Since the exact gradient is unknown, in policy gradient, one constructs stochastic unbiased estimates of the actual gradient by sampling trajectories, but this may result in slow learning due to high variance for gradient estimates. On the other hand, actor-critic simultaneously performs online estimations of the value function approximation (actor) and policy parameters (critic). This can be viewed as a bootstrapping method to policy gradient which accelerates learning by trading bias for variance. Here we developed an online approximation algorithm for problem $\mathcal{SP}$ based on *actor critic* and show that it asymptotically converges to the local optimal solution[5].

[5]Actor critic methods converge to the local optimal solution of problem $\mathcal{SP}$ with respect to pre-defined parametrized classes of value functions and policies. More details will be provided in subsequent analysis.

Recall that a *stationary policy* $\mu(\cdot|x^\omega)$ is a probability distribution over actions, conditioned on the current state $x^\omega = (x, \omega)$. In policy gradient methods, we define a class of parameterized stochastic policies $\left\{\mu(\cdot|x^\omega;\theta), x^\omega = (x,\omega), \theta \in \Theta \subseteq \mathbb{R}^{\kappa_1}\right\}$. Since in this setting a policy μ is represented by its κ_1-dimensional parameter vector θ, policy dependent functions can be written as a function of θ in place of μ. So, we use μ and θ interchangeably in this section.

5.1 Value Function Approximation

Consider the $v-$dependent linear value function approximation of $V_\theta(x^\omega)$, in the form of $\phi^\top(x^\omega)v$, where $\phi(x^\omega) \in \mathbb{R}^{\kappa_2}$ represents the state-dependent feature. The feature vectors can also be dependent on θ as well. But for notational convenience, we drop the indices corresponding to θ. The low dimensional subspace is therefore $S_V = \{\Phi v | v \in \mathbb{R}^{\kappa_2}\}$ where $\phi : \mathbf{X} \times \Omega \to \mathbb{R}^{\kappa_2}$ is a function mapping such that $\Phi(x^\omega) = \phi^\top(x^\omega)$. We also make the standard assumption on the rank of matrix ϕ [4].

ASSUMPTION 7. *The basis functions $\left\{\phi^{(i)}\right\}_{i=1}^{\kappa_2}$ are linearly independent. In particular, $\kappa_2 \ll n$ and Φ is full rank.*

Let $v \in \mathbb{R}^{\kappa_2}$ be the best approximation parameter vector. Then $\tilde{V}_\theta^v(x^\omega) = v^\top\phi(x^\omega)$ is the best linear approximation of $V_\theta(x^\omega)$.

Since our goal is to approximate the value function of a stochastic shortest path problem with stopping time ΔT, we define the feature vector as follows:

$$\phi(x^\omega) = \mathbf{0}, \text{ if } x^\omega = (\mathbf{x}^T, \omega_0).$$

To estimate v from simulated trajectories of the stochastic shortest path MDP, it is reasonable to consider the projections from $\mathbb{R}$ onto S_V with respect to a norm that is weighted according to the occupation measure

$$d_\theta(y^{\omega,\prime}|x^\omega) = \sum_{h=0}^{\infty} \mathbb{P}(\mathbf{x}_h^\omega = y^{\omega,\prime} \mid \mathbf{x}_0^\omega = x^\omega, \mu), \forall x^\omega, y^{\omega,\prime} \in \mathbf{X}\times\Omega,$$

where $\mathbf{x}_0^\omega = x^\omega$ is the initial condition. We also make the following standard assumption for the state-action pair visiting probability.

ASSUMPTION 8. *For all $\theta \in \Theta$, each state-action pair has a positive probability of being visited, i.e., $d_\theta(y^{\omega,\prime}|\mathbf{x}_0^\omega)\mu(u|y^{\omega,\prime};\theta) > 0$ for any $u \in \mathbf{U}$ and $y^\omega \in \mathbf{X}\times\Omega$.*

For a function $f : \mathbf{X}\times\Omega \to \mathbb{R}$, we introduce the weighted norm: $\|f\|_d = \sqrt{\sum_{y^{\omega,\prime}} d(y^{\omega,\prime}|x^\omega)(f(y^{\omega,\prime}))^2}$ where d is the occupation measure (with non-negative elements).

We also denote by Π the projection from $\mathbf{X}\times\Omega$ to S_V. We are now ready to describe the approximation scheme. Consider the following projected fixed point equation

$$V(x^\omega) = \Pi F_\theta[V](x^\omega)$$

where F_θ is the Bellman operator with respect to policy μ and let $\tilde{V}_\theta^v$ denote the solution of the above equation. The existence of this unique fixed point is guaranteed by the following contraction property of the projected Bellman operator: ΠF_θ, whose proof is given in Proposition 7.1.1 in [4].

LEMMA 9. *There exists $\kappa \in (0,1)$ such that*

$$\|\Pi F_\theta[V_1] - \Pi F_\theta[V_2]\|_d \le \kappa\|V_1 - V_2\|_d.$$

Therefore, by Banach fixed point theorem, a unique fixed point solution exists for equation: $\Pi F_\theta[V](x^\omega) = V(x^\omega)$ for any $x^\omega = (x,\omega)$. Denote by $\tilde{V}_\theta^v$ the fixed point solution and v the corresponding weight, which is unique by the full rank assumption. From Lemma 9, one obtains a unique value function estimates from the following projected Bellman equation:

$$\Pi F_\theta[\tilde{V}_\theta^v](x^\omega) = \tilde{V}_\theta^v(x^\omega),\ \tilde{V}_\theta^v(x^\omega) = (v)^\top\phi(x^\omega). \tag{20}$$

Note that we can re-write the projected Bellman equation in explicit form as follows:

$$\Pi F_\theta[\Phi v] = \Phi v \iff \Pi\Bigg[\Bigg\{\sum_{u\in\mathbf{U}}\mu(u|x^\omega;\theta)\Big(\mathbf{r}_{\mathcal{S}}(x^\omega,u)+ \sum_{y^{\omega,\prime}\in\mathbf{X}\times\Omega}\mathbf{P}^u_{x^\omega,y^{\omega,\prime}}(v)^\top\phi\left(y^{\omega,\prime}\right)\Big)\Bigg\}_{x^\omega\in\mathbf{X}\times\Omega}\Bigg] = \Phi v.$$

By the definition of projection, the unique solution $v \in \mathbb{R}^\ell$ satisfies

$$v \in \arg\min_v \|F_\theta[\Phi v] - \Phi v\|_{d_\theta}^2 \iff v \in \arg\min_v \sum_{y^{\omega,\prime}\in\mathbf{X}\times\Omega} d_\theta(y^{\omega,\prime}|x^\omega)\left(\sum_{u'\in\mathbf{U}}\mu(u'|y^{\omega,\prime};\theta)\tilde{Q}_\theta^v(y^{\omega,\prime},u') - \phi^\top(y^{\omega,\prime})v\right)^2.$$

where for any $x^\omega \in \mathbf{X}\times\Omega$ and $u \in \mathbf{U}$,

$$\tilde{Q}_\theta^v(x^\omega,u) = \sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega}\mathbf{P}^u_{x^\omega,x^{\omega,\prime}}v^\top\phi(x^{\omega,\prime}) + \mathbf{r}_{\mathcal{S}}(x^\omega,u)$$

is the approximate $Q-$function using linear unction approximation. By the projection theorem on Hilbert space, the orthogonality condition for v becomes:

$$\begin{aligned}&\sum_{y^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}}\pi_\theta(y^{\omega,\prime},u'|x^\omega)\phi(y^{\omega,\prime})(v)^\top\phi(y^{\omega,\prime})\\ =&\sum_{y^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}}\Big\{\pi_\theta(y^{\omega,\prime},u'|x^\omega)\phi(y^{\omega,\prime})\mathbf{r}_{\mathcal{S}}(y^{\omega,\prime},u')+ \sum_{z^{\omega,\prime\prime}\in\mathbf{X}\times\Omega}\pi_\theta(y^{\omega,\prime},u'|x^\omega)\mathbf{P}^{u'}_{y^{\omega,\prime},z^{\omega,\prime\prime}}\phi(y^{\omega,\prime})\phi^\top\left(z^{\omega,\prime\prime}\right)v\Big\}\end{aligned}$$

where for any $x^\omega, y^{\omega,\prime} \in \mathbf{X}\times\Omega$ and $u' \in \mathbf{U}$

$$\begin{aligned}\pi_\theta(y^{\omega,\prime},u'|x^\omega) =& d_\theta(y^{\omega,\prime}|x^\omega)\mu(u'|x,\omega;\theta)\\ =&\sum_{h=0}^{\infty}\mathbb{P}(\mathbf{x}_h^\omega = y^{\omega,\prime},\mathbf{u}_h = u' \mid \mathbf{x}_0^\omega = x^\omega,\mu)\end{aligned}$$

is the state-action occupation measure. This condition can be written as $Av = b$ where

$$A = \sum_{y^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}}\pi_\theta(y^{\omega,\prime},u'|x^\omega)\phi(y^{\omega,\prime})\cdot \left(\phi^\top(y^{\omega,\prime}) - \sum_{z^{\omega,\prime\prime}\in\mathbf{X}\times\Omega}\mathbf{P}^u_{y^{\omega,\prime},z^{\omega,\prime\prime}}\phi^\top\left(z^{\omega,\prime\prime}\right)\right) \tag{21}$$

is a finite dimensional matrix in $\mathbb{R}^{\kappa_2\times\kappa_2}$ and

$$b = \sum_{y^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}}\pi_\theta(y^{\omega,\prime},u'|x^\omega)\phi(y^{\omega,\prime})\mathbf{r}_{\mathcal{S}}(y^{\omega,\prime},u') \tag{22}$$

is a finite dimensional vector in $\mathbb{R}^{\kappa_2}$. The matrix A is invertible since Lemma 9 guarantees that (20) has a unique solution v. Note that the projected equation $Av = b$ can be re-written as

$$v = v - \xi(Av - b)$$

for any positive scaler $\xi \geq 0$. By expanding the structure of the occupation measures, this further implies

$$A = \mathbb{E}\left[\sum_{h=0}^{\Delta T-1} \phi(\mathbf{x}_h^\omega)\left(\phi^\top(\mathbf{x}_h^\omega) - \phi^\top(\mathbf{x}_{h+1}^\omega)\right) \mid \mathbf{x}_0^\omega = x^\omega, \mu\right],$$

$$b = \mathbb{E}\left[\sum_{h=0}^{\Delta T-1} \phi(\mathbf{x}_h^\omega)\mathbf{r}_{\mathcal{S}}(\mathbf{x}_h^\omega, \mathbf{u}_h) \mid \mathbf{x}_0^\omega = x^\omega, \mu\right].$$

5.2 The Actor-Critic Algorithm

In this section, we propose an actor-critic algorithm that use linear approximation in the gradient estimates and update the parameters episodically (after the states reach the stopping region). This algorithm is based on the gradient estimate of θ and temporal difference update. Algorithm 2 contains the pseudo-code of this algorithm. The projection operator Γ_Θ is defined as $\arg\min_{\hat{\theta}\in\Theta} \frac{1}{2}\|\theta - \hat{\theta}\|_2^2$ and is necessary to ensure the convergence of the algorithm. The step-size schedules satisfy the standard conditions for stochastic approximation the algorithm, i.e.,

$$\sum_k \zeta_k = \sum_k \zeta_k' = \infty,\ \sum_k (\zeta_k)^2, \sum_k (\zeta_k')^2 < \infty,\ \zeta_k' = o\big(\zeta_k\big). \tag{23}$$

It ensures that the critic update is on the fast time-scale $\{\zeta_k\}$ and the policy parameter updates are on the slow time-scale $\{\zeta_k'\}$. This results in a two time-scale stochastic approximation algorithm.

Algorithm 2 Actor-Critic Algorithm for Problem $\mathcal{SP}$

Input: Parameterized policy $\mu(\cdot|\cdot;\theta)$ and value function feature vector $\phi(\cdot)$
Initialization: policy parameter $\theta = \theta_0$; value function weight vector $v = v_0$
for $k = 0, 1, 2, \ldots$ **do**
  Set $\mathbf{x}_0^\omega = (\mathbf{x}_0, \omega_0)$ and $h = 0$;
  while $\omega_h \neq (x^T, \omega_0)$ **do**
    Draw action $\mathbf{u}_h \sim \mu(\cdot|\mathbf{x}_h^\omega;\theta_k)$ and observe reward $\mathbf{R}_{\mathcal{SP}}(\mathbf{x}_h^\omega, \mathbf{u}_h)$; Assign state updates $\mathbf{x}_{h+1}$ using Algorithm 1; Observe next stochastic state $\omega_{h+1} \sim \mathbb{P}(\omega_{h+1} = \cdot \mid \omega_h)$; Set the next aggregate state as $\mathbf{x}_{h+1}^\omega = (\mathbf{x}_{h+1}, \omega_{h+1})$; Perform the following update

$$\textbf{TD Error:}\quad \delta_h(v_k) = -v_k^\top \phi(\mathbf{x}_h^\omega) + v_k^\top \phi\left(\mathbf{x}_{h+1}^\omega\right) + \mathbf{R}_{\mathcal{SP}}(\mathbf{x}_h^\omega, \mathbf{u}_h) \tag{24}$$

    Update $h \leftarrow h + 1$
  end while
  Set $\Delta T_k = h$ and perform the following updates

$$\textbf{Critic Update:}\quad v_{k+1} = v_k + \zeta_k \sum_{h=0}^{\Delta T_k - 1} \phi(\mathbf{x}_h^\omega)\delta_h(v_k) \tag{25}$$

$$\textbf{Actor Update:}\quad \theta_{k+1} = \Gamma_\Theta\left(\theta_k - \zeta_k'\left(\sum_{h=0}^{\Delta T_k - 1} \nabla_\theta \log \mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)|_{\theta=\theta_k}\delta_h(v_k)\right)\right) \tag{26}$$

end for
return policy and value function parameters θ, v

5.2.1 TD(0) Critic Update

In this section, we want to show that the TD(0) critic update $\{v_k\}$ converges to the "best" linear function approximation v with respect to the policy parameter θ, i.e., $v \in \arg\min_v \|F_\theta[\Phi v] - \Phi v\|_{d_\theta}^2$. Recall the TD(0) update v_k in equation (25), where the scaler δ_h in equation (24) is known as the temporal difference (TD). Before getting into the convergence analysis, we have the following technical lemma whose proof is given in Theorem 5.1 in [5].

LEMMA 10. *Every eigenvalues of matrix A has positive real part.*

We now have the following theorem showing the convergence of the critic updates.

THEOREM 11. *The TD(0) iterates converges to the unique fixed point v almost surely, at $k \to \infty$.*

Proof. **R**ecall the TD(0) update v_k in equation (25), where the scaler δ_h in equation (24) is known as the temporal difference (TD). Based on the definitions of matrices A and b in equation (21) to (22), it is easy to see that the TD(0) critic update v_k in equation (25) can be re-written as the following stochastic approximation scheme:

$$v_{k+1} = v_k + \zeta_k(b - Av_k + \delta A_{k+1}) \tag{27}$$

where the noise term δA_{k+1} satisfies the Martingale difference equation, i.e, $\mathbb{E}[\delta A_{k+1} \mid \mathcal{F}_k] = 0$ and $\mathcal{F}_k$ is the filtration generated by different independent trajectories. By writing

$$\delta A_{k+1} = -(b - Av_k) + \sum_{h=0}^{\Delta T_k - 1} \phi(\mathbf{x}_h^\omega)\delta_h(v_k)$$

and noting

$$\mathbb{E}\left[\sum_{h=0}^{\Delta T-1} \phi(\mathbf{x}_h^\omega)\delta_h(v_k) \mid \mathcal{F}_k\right] = -Av_k + b,$$

one can easily check that the above stochastic approximation scheme is equivalent to the TD(0) iterates in (25) and δA_{k+1} is a Martingale difference, i.e., $\mathbb{E}[\delta A_{k+1} \mid \mathcal{F}_k] = 0$. Let

$$h(v) = -Av + b.$$

Note that $H(v)$ is Lipschitz, the step size satisfies the properties in (23), the noise term δA_{k+1} satisfies the Martingale difference equation, $H_c(v) := H(cv)/c$ for $c \geq 1$ converges uniformly to a continuous function $H_\infty(v)$ for any v in a compact set, i.e., $H_c(v) \to H_\infty(v)$ as $c \to \infty$, and the ordinary differential equation (ODE) $\dot{v} = H_\infty(v)$ has the origin as its unique globally asymptotically stable equilibrium. By Theorem 3.1 in [8], the TD iterates $\{v_k\}$ is bounded almost surely, i.e., $\sup_k \|v_k\| < \infty$ almost surely.

Finally, from the standard stochastic approximation result and the above conditions, the convergence of the TD(0) iterates in (25) can be related to the asymptotic behavior of the ODE

$$\dot{v} = H(v) = b - Av. \tag{28}$$

By Theorem 2 in Chapter 2 of [8], when the properties hold, then $v_k \to v$ with probability 1 where the limit v is the unique solution satisfying $H(v) = 0$, i.e., $Av = b$. Therefore, the TD(0) iterates converges to the unique fixed point v almost surely, at $k \to \infty$. ■

5.2.2 Actor Update

We turn to show that the policy gradient update of θ in (26) attains a stationary (locally optimal) point. Before getting to the main result, we first to compute the gradient of V_θ with respect to θ at $\mathbf{x}_0^\omega = x^\omega$. Define the $Q-$value function as

$$Q_\theta(x^\omega, u) = \sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega} \mathbf{P}_{x^\omega, x^{\omega,\prime}}^u V_\theta(x^{\omega,\prime}) + \mathbf{r}_{\mathcal{SP}}(x^\omega, u),$$

where one can check that

$$\sum_{u\in\mathbf{U}} \mu(u|x^\omega;\theta)Q_\theta(x^\omega,u) = V_\theta(x^\omega).$$

One obtains the following expression by direct gradient evaluation:

$$\begin{aligned}
&\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\\
=&\sum_{u\in\mathbf{U}} \nabla_\theta\mu(u|\mathbf{x}_0^\omega;\theta)Q_\theta(\mathbf{x}_0^\omega,u) + \mu(u|\mathbf{x}_0^\omega;\theta)\nabla_\theta Q_\theta(\mathbf{x}_0^\omega,u)\\
=&\sum_{u\in\mathbf{U}} \nabla_\theta\mu(u|\mathbf{x}_0^\omega;\theta)Q_\theta(\mathbf{x}_0^\omega,u) + \mu(u|\mathbf{x}_0^\omega;\theta)\sum_{\mathbf{x}_1^\omega\in\mathbf{X}\times\Omega}\mathbf{P}^u_{\mathbf{x}_0^\omega,\mathbf{x}_1^\omega}\nabla_\theta V_\theta\left(\mathbf{x}_1^\omega\right)\\
=&h_\theta(\mathbf{x}_0^\omega) + \sum_{\mathbf{x}_1^\omega\in\mathbf{X}\times\Omega,\mathbf{u}_0\in\mathbf{U}} \mu(\mathbf{u}_0|\mathbf{x}_0^\omega;\theta)\mathbf{P}^{\mathbf{u}_0}_{\mathbf{x}_0^\omega,\mathbf{x}_1^\omega}\nabla_\theta V_\theta\left(\mathbf{x}_1^\omega\right)
\end{aligned}$$

where

$$h_\theta(\mathbf{x}_0^\omega) = \sum_{u\in\mathbf{U}} \nabla_\theta\mu(u|\mathbf{x}_0^\omega;\theta)Q_\theta(\mathbf{x}_0^\omega,u).$$

Since the above expression is a recursion, one further obtains

$$\nabla_\theta V_\theta(\mathbf{x}_0^\omega) = h_\theta(\mathbf{x}_0^\omega) + \sum_{\mathbf{x}_1^\omega\in\mathbf{X}\times\Omega,\mathbf{u}_0\in\mathbf{U}} \mu(\mathbf{u}_0|\mathbf{x}_0^\omega;\theta)\mathbf{P}^{\mathbf{u}_0}_{\mathbf{x}_0^\omega,\mathbf{x}_1^\omega}$$

$$\left[h_\theta(\mathbf{x}_1^\omega) + \sum_{\mathbf{x}_2^\omega\in\mathbf{X}\times\Omega,\mathbf{u}_1\in\mathbf{U}} \mu(\mathbf{u}_1|\mathbf{x}_1^\omega;\theta)\mathbf{P}^{\mathbf{u}_1}_{\mathbf{x}_1^\omega,\mathbf{x}_2^\omega}\nabla_\theta V_\theta\left(\mathbf{x}_2^\omega\right)\right].$$

By the definition of occupation measures and note that from (19),

$$\lim_{h\to\infty}\mathbb{P}[\mathbf{x}_h^\omega \neq (x^T,\omega_0)|\mathbf{x}_0^\omega,\theta] = 0, \text{ for any } \mathbf{x}_0^\omega\in\mathbf{X}\times\Omega,$$

the above expression becomes

$$\begin{aligned}
&\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\\
=&\sum_{h=0}^{\infty}\sum_{x^\omega\in\mathbf{X}\times\Omega}\mathbb{P}(\mathbf{x}_h^\omega = x^\omega|\mathbf{x}_0^\omega,\mu)h_\theta(x^\omega) = \sum_{x^\omega\in\mathbf{X}\times\Omega} d_\theta(x^\omega|\mathbf{x}_0^\omega)h_\theta(x^\omega)\\
=&\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega} d_\theta(x^{\omega,\prime}|\mathbf{x}_0^\omega)\sum_{u'\in\mathbf{U}}\nabla_\theta\mu(u'|x^{\omega,\prime};\theta)Q_\theta(x^{\omega,\prime},u')\\
=&\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}} \pi_\theta(x^{\omega,\prime},u'|\mathbf{x}_0^\omega)\nabla_\theta\log\mu(u'|x^{\omega,\prime};\theta)A_\theta(x^{\omega,\prime},u')
\end{aligned}$$

where $A_\theta(x^\omega,u) = Q_\theta(x^\omega,u) - V_\theta(x^\omega)$ is the advantage function. The last equality is due to $\mu(u|x;\theta)\nabla_\theta\log\mu(u|x^\omega;\theta) = \nabla_\theta\mu(u|x^\omega;\theta)$ and $\sum_u\nabla_\theta\mu(u|x^\omega;\theta) = \nabla_\theta\sum_u\mu(u|x^\omega;\theta) = \nabla_\theta(1) = 0$.

Furthermore, we have the following technical lemma on the Lipchitz continuity of the gradient of the objective function $\nabla_\theta V_\theta(\mathbf{x}_0^\omega)$.

PROPOSITION 12. *Assume for any state-action pair (x^ω,u), $\mu(u|x^\omega;\theta)$ is continuously differentiable in θ and $\nabla_\theta\mu(u|x^\omega;\theta)$ is Lipschitz in θ for every $u\in\mathbf{U}$ and $x^\omega = (x,\omega)\in\mathbf{X}\times\Omega$. Then $\nabla_\theta V_\theta(\mathbf{x}_0^\omega)$ is Lipschitz in θ.*

Proof. **R**ecall that for $\mathcal{H} = \{\mathbf{x}_0^\omega,\mathbf{u}_0,\ldots,\mathbf{x}_{\Delta T-1}^\omega,\mathbf{u}_{\Delta T-1}\}$ being an arbitrary transient state-action trajectory, one obtains

$$\nabla_\theta V_\theta(\mathbf{x}_0^\omega) = \mathbb{E}\left[\sum_{\mathcal{H}}\mathbb{P}_\theta(\mathcal{H})\cdot\nabla_\theta\log\mathbb{P}_\theta(\mathcal{H})\left(\sum_{h=0}^{\Delta T-1}\mathbf{r}_{\mathcal{S}}(\mathbf{x}_h^\omega,\mathbf{u}_h)\right)\right]$$

where the expectation is taken over the random stopping time ΔT, and $\nabla_\theta\log\mathbb{P}_\theta(\mathcal{H}) = \sum_{h=0}^{\Delta T-1}\nabla_\theta\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)/\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)$ whenever $\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)\in(0,1]$. Now the assumption of this proposition implies that $\nabla_\theta\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)$ is a Lipschitz function in θ for $h\in\{0,\ldots,\Delta T-1\}$ and $\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)$ is differentiable in θ. Therefore, by recalling that

$$\mathbb{P}_\theta(\mathcal{H}) = \prod_{h=0}^{\Delta T-1}\mathbf{P}^{\mathbf{u}_h}_{\mathbf{x}_h^\omega,\mathbf{x}_{h+1}^\omega}\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta),$$

combining these arguments and noting that the sum of products of Lipschitz functions is Lipschitz continuous, one concludes that $\nabla_\theta V_\theta(\mathbf{x}_0^\omega)$ is Lipschitz continuous in θ. ■

For any linear function approximation vector $v\in\mathbb{R}^{\kappa_1}$, define the $v-$dependent approximated advantage function

$$\tilde{A}_\theta^v(x^\omega,u) = \tilde{Q}_\theta^v(x^\omega,u) - v^\top\phi(x^\omega),\ x^\omega\in\mathbf{X}\times\Omega,\ u\in\mathbf{U}$$

The following Lemma first shows that $\delta_k(v)$ is an unbiased estimator of $\tilde{A}_\theta^v$.

LEMMA 13. *For any given policy μ and $v\in\mathbb{R}^{\kappa_2}$, we have*

$$\tilde{A}_\theta^v(x^\omega,u) = \mathbb{E}[\delta_k(v)\mid\mathbf{x}_h^\omega = x^\omega,\mathbf{u}_h = u],\ x^\omega\in\mathbf{X}\times\Omega,\ u\in\mathbf{U}.$$

Proof. **N**ote that for any $v\in\mathbb{R}^{\kappa_2}$,

$$\mathbb{E}[\delta_k(v)\,|\,\mathbf{x}_h^\omega,\mathbf{u}_h] = \mathbf{r}_{\mathcal{S}}(\mathbf{x}_h^\omega,\mathbf{u}_h) - v^\top\phi(x^\omega) + \mathbb{E}\left[v^\top\phi(\mathbf{x}_{h+1}^\omega)\,|\,\mathbf{x}_h^\omega,\mathbf{u}_h\right],$$

where

$$\mathbb{E}\left[v^\top\phi(\mathbf{x}_{h+1}^\omega)\mid\mathbf{x}_h^\omega = x^\omega,\mathbf{u}_h = u\right] = \sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega}\mathbf{P}^u_{x^\omega,x^{\omega,\prime}}v^\top\phi(x^{\omega,\prime}).$$

By recalling the definition of $\tilde{Q}_\theta^v(x^\omega,u)$, the proof is completed. ■

Recall $\tilde{V}_\theta^v(\mathbf{x}_0^\omega) = v^\top\phi(\mathbf{x}_0^\omega)$ as the linear function approximation of $V_\theta(\mathbf{x}_0^\omega)$, where the approximation vector v depends on the policy parameter θ. Define $\nabla_\theta\tilde{V}_\theta^v(\mathbf{x}_0^\omega):\Theta\to\mathbb{R}$ as the linear function approximation of $\nabla_\theta V_\theta(\mathbf{x}_0^\omega)$ as follows:

$$\nabla_\theta\tilde{V}_\theta^v(\mathbf{x}_0^\omega) := \sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}}\pi_\theta(x^{\omega,\prime},u'|\mathbf{x}_0^\omega)\nabla_\theta\log\mu(u'|x^{\omega,\prime};\theta)\tilde{A}^v(x^{\omega,\prime},u').$$

Similar to Proposition 12, we have the following technical Lemma on $\nabla_\theta\tilde{V}_\theta^v(\mathbf{x}_0^\omega)$.

PROPOSITION 14. *Assume for any state-action pair (x^ω,u), $\mu(u|x^\omega;\theta)$ is continuously differentiable in θ and $\nabla_\theta\mu(u|x^\omega;\theta)$ is Lipschitz in θ for every $u\in\mathbf{U}$ and $x^\omega = (x,\omega)\in\mathbf{X}\times\Omega$. Then the function $\nabla_\theta\tilde{V}_\theta^v(\mathbf{x}_0^\omega)$ is Lipschitz in θ.*

Proof. **F**irst consider the approximation vector v. Recall that this vector satisfies the linear equation $Av = b$ where A and b are functions of θ found from the Hilbert space projection of Bellman operator. It has been shown in Lemma 1 of [**?**] that, by exploiting the inverse of A using Cramer's rule, one can show that v is continuously differentiable of θ. Next, consider the occupation measure π_θ. From an application of Theorem 2 of [**?**] (or Theorem 3.1 of [**?**]), it can be seen that the stationary distribution π_θ of the process $\mathbf{x}_h^\omega$ is continuously differentiable in θ. Recall from Assumption (B1) that $\nabla_\theta\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)$ is a Lipschitz function in θ for any $a\in\mathcal{A}$ and $h\in\{0,\ldots,\Delta T-1\}$ and $\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)$ is differentiable in θ. Therefore, by combining these arguments and noting that the sum of products of Lipschitz functions is Lipschitz, one concludes that $\nabla_\theta\tilde{V}_\theta^v(\mathbf{x}_0^\omega)$ is Lipschitz in θ. ■

Now we turn to the main result on proving the convergence of $\theta-$update. Since v converges in a faster scale than θ, one can also replace v with the limit $v^*(\theta)$ in the convergence analysis. Then the $\theta-$update in (26) can be re-written as follows:

$$\theta_{k+1} = \Gamma_\Theta\left(\theta_k - \zeta_k'\sum_{h=0}^{\Delta T_k-1}\nabla_\theta\log\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)|_{\theta=\theta_k}\delta_h(v^*(\theta_k))\right). \tag{29}$$

For any policy parameter $\theta \in \Theta$, define

$$\epsilon_\theta(v_k) = \|F_\theta[\Phi v_k] - \Phi v_k\|_\infty^2$$

as the residual of the value function approximation at step k, induced by policy $\mu(\cdot|\cdot,\cdot;\theta)$. By triangular inequality and fixed point theorem $F_\theta[V_\theta] = V_\theta$, it can be easily seen that $\|V_\theta - \Phi v_k\|_\infty^2 \leq \epsilon_\theta(v_k) + \|F_\theta[\Phi v_k] - F_\theta[V_\theta]\|_\infty^2 \leq \epsilon_\theta(v_k) + \kappa\|\Phi v_k - V_\theta\|_\infty^2$ for $\kappa \in (0,1)$ given in Proposition 5. The last inequality follows from the contraction mapping with $\kappa \in (0,1)$. Thus, one concludes that $\|V_\theta - \Phi v_k\|_\infty^2 \leq \epsilon_\theta(v_k)/(1-\kappa)$.

Before stating the main result, define

$$\Upsilon_\theta[K(\theta)] := \lim_{0<\eta\to 0} \frac{\Gamma_\Theta\big(\theta + \eta K(\theta)\big) - \Gamma_\Theta(\theta)}{\eta}$$

as the left directional derivative of the function $\Gamma_\Theta(\theta)$ in the direction of $K(\theta)$. By the left directional derivative $\Upsilon_\theta\left[-\nabla_\theta \tilde{V}_\theta^v(\mathbf{x}_0^\omega)\right]$ in the gradient descent algorithm for θ, the gradient will point at the descent direction along the boundary of Θ whenever the $\theta-$update hits its boundary. The following theorem provides a convergence result to the policy parameter.

THEOREM 15. *Assume for any state-action pair (x^ω, u), the control policy $\mu(u|x^\omega;\theta)$ is continuously differentiable in θ and its gradient $\nabla_\theta\mu(u|x^\omega;\theta)$ is Lipschitz in θ for every $u \in \mathbf{U}$ and $x^\omega = (x,\omega) \in \mathbf{X}\times\Omega$. Suppose $\widehat{\theta}^*$ is the equilibrium point of the continuous system θ satisfying*

$$\Upsilon_\theta\left[-\nabla_\theta\left(\tilde{V}_\theta^v(\mathbf{x}_0^\omega)\mid_{v=v^*(\theta)}\right)\right] = 0. \tag{30}$$

Then the sequence of $\theta-$updates in (26) *converges to $\widehat{\theta}^*$ almost surely. Furthermore, suppose $\theta^* \in \operatorname{argmin}_{\theta\in\Theta} V_\theta(\mathbf{x}_0^\omega)$ is a local minimum point. If $\epsilon_\theta(v_k)\to 0$ as $v_k \to v^*$, then this sequence of $\theta-$updates converges to θ^* almost surely.*

Proof. We will mainly focus on deriving the convergence of $\theta_k \to \theta^*$ (second part of the theorem). Since we just showed in Proposition 14 that $\nabla_\theta\tilde{V}_\theta^v(\mathbf{x}_0^\omega)$ is Lipchitz in θ, the convergence proof of $\theta_k \to \widehat{\theta}^*$ (first part of the theorem) follows from identical arguments.

First, the $\theta-$update from (29) can be re-written as follows:

$$\theta_{k+1} = \Gamma_\Theta\left(\theta_k + \zeta_k'\left(-\nabla_\theta V_\theta(\mathbf{x}_0^\omega)|_{\theta=\theta_k} + \delta\theta_{k+1} + \delta\theta_\epsilon\right)\right)$$

where

$$\begin{aligned}\delta\theta_{k+1} = &\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega, u'\in\mathbf{U}} \pi_{\theta_k}(x^{\omega,\prime},u'|\mathbf{x}_0^\omega)\nabla_\theta\log\mu(u'|x^{\omega,\prime};\theta)|_{\theta=\theta_k}\cdot\\ &\tilde{A}_{\theta_k}^v(x^{\omega,\prime},u') - \sum_{h=0}^{\Delta T_k-1}\nabla_\theta\log\mu(\mathbf{u}_h|\mathbf{x}_h^\omega;\theta)|_{\theta=\theta_k}\delta_h(v^*(\theta_k)).\end{aligned} \tag{31}$$

Since ΔT is an i.i.d. random variable with bounded first and second moments, one can show that $\delta\theta_{k+1}$ is square integrable, i.e.,

$$\begin{aligned}&\mathbb{E}[\|\delta\theta_{k+1}\|^2 \mid \mathcal{F}_{\theta,k}] \leq 2\mathbb{E}(\Delta T)^2]\|\nabla_\theta\log\mu(u'|x^{\omega,\prime};\theta)|_{\theta=\theta_k}\|_\infty^2\cdot\\ &\left(\|\tilde{A}_{\theta_k}^v(x^{\omega,\prime},u')\|_\infty^2 + \max_h|\delta_h(v^*(\theta_k))|^2\right) \leq 64K_1^2\mathbb{E}[(\Delta T)^2]\cdot\\ &\left(\max_{x^\omega,u}|\mathbf{r}_{\mathcal{SP}}(x^\omega,u)|^2 + 2\max_{x^\omega}\|\phi(x^\omega)\|^2\sup_k\|v_k\|^2\right)(1+\|\theta_k\|^2).\end{aligned}$$

The Lipschitz upper bound $\|\nabla_\theta\log\mu(u'|x^{\omega,\prime};\theta)|_{\theta=\theta_k}\| \leq K_1(1+\|\theta_k\|)$ is based on the assumption of this theorem and $\sup_k\|v_k\| < \infty$ is based on the Lyapunov analysis in the Critic update.

Second, notice that for $v = v^*(\theta_k)$,

$$\begin{aligned}\delta\theta_\epsilon = &\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega, u'\in\mathbf{U}}\pi_{\theta_k}(x^{\omega,\prime},u'|\mathbf{x}_0^\omega)\nabla_\theta\log\mu(u'|x^{\omega,\prime};\theta)|_{\theta=\theta_k}\cdot\\ &(A_{\theta_k}(x^{\omega,\prime},u') - \tilde{A}_{\theta_k}^v(x^{\omega,\prime},u')) \leq 2\mathbb{E}[\Delta T]\sqrt{\frac{\epsilon_\theta(v)}{1-\kappa}}\|\psi_{\theta_k}\|_\infty.\end{aligned}$$

where $\psi_\theta(x^\omega,u) = \nabla_\theta\log\mu(u|x^\omega;\theta)$ is the "compatible feature". For the last inequality, recall π_θ is the state-action occupation measure and define

$$\Delta V_\theta^v(x^\omega) = V_\theta(x^\omega) - \phi^\top(x^\omega)v$$

as difference between the value function and its approximation at $x^\omega \in \mathbf{X}\times\Omega$, convexity of quadratic functions implies

$$\begin{aligned}&\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}}\pi_\theta(x^{\omega,\prime},u'|\mathbf{x}_0^\omega)(A_\theta(x^{\omega,\prime},u') - \tilde{A}_\theta^v(x^{\omega,\prime},u'))\\
=&\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}}\pi_\theta(x^{\omega,\prime},u'|\mathbf{x}_0^\omega)(Q_\theta(x^{\omega,\prime},u') - \tilde{Q}_\theta^v(x^{\omega,\prime},u'))\\
&+\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega}d_\theta(x^{\omega,\prime}|\mathbf{x}_0^\omega)\Delta V_\theta^v(x^{\omega,\prime})\\
\leq&\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}}\pi_\theta(x^{\omega,\prime},u'|\mathbf{x}_0^\omega)\sum_{x^{\omega,\prime\prime}\in\mathbf{X}\times\Omega}\mathbf{P}_{x^{\omega,\prime},x^{\omega,\prime\prime}}^{u'}\Delta V_\theta^v(x^{\omega,\prime\prime})\\
&+\mathbb{E}[\Delta T]\sqrt{\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega}\frac{d_\theta(x^{\omega,\prime}|\mathbf{x}_0^\omega)}{\mathbb{E}[\Delta T]}(\Delta V_\theta^v(x^{\omega,\prime}))^2}\\
\leq&\left(\sqrt{\sum_{x^{\omega,\prime}\in\mathbf{X}\times\Omega,u'\in\mathbf{U}}\frac{\pi_\theta(x^{\omega,\prime},u'|\mathbf{x}_0^\omega)}{\mathbb{E}[\Delta T]}\sum_{x^{\omega,\prime\prime}\in\mathbf{X}\times\Omega}\mathbf{P}_{x^{\omega,\prime},x^{\omega,\prime\prime}}^{u'}(\Delta V_\theta^v(x^{\omega,\prime\prime}))^2}\right.\\
&\left.+\sqrt{\frac{\epsilon_\theta(v)}{1-\kappa}}\right)\times\mathbb{E}[\Delta T]\\
\leq&\mathbb{E}[\Delta T]\sqrt{\sum_{x^{\omega,\prime\prime}\in\mathbf{X}\times\Omega}\frac{d_\theta(x^{\omega,\prime\prime}|\mathbf{x}_0^\omega) - 1\{\mathbf{x}_0^\omega = x^{\omega,\prime\prime}\}}{\mathbb{E}[\Delta T]}(\Delta V_\theta^v(x^{\omega,\prime\prime}))^2}\\
&+\mathbb{E}[\Delta T]\sqrt{\frac{\epsilon_\theta(v)}{1-\kappa}}\\
\leq&2\mathbb{E}[\Delta T]\sqrt{\frac{\epsilon_\theta(v)}{1-\kappa}}.\end{aligned}$$

By Lemma 13, one obtains $\mathbb{E}\left[\delta\theta_{k+1} \mid \mathcal{F}_{\theta,k}\right] = 0$, where $\mathcal{F}_{\theta,k} = \sigma(\theta_m,\, \delta\theta_m,\, m \leq k)$ is the filtration generated by different independent trajectories. On the other hand, $|\delta\theta_\epsilon| \to 0$ almost surely as $\epsilon_{\theta_k}(v^*(\theta_k)) \to 0$. Therefore, the $\theta-$update in (29) is a stochastic approximation of the continuous system $\theta_{(t)}$ which satisfies the following ODE

$$\dot{\theta} = \Upsilon_\theta\left[-\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\right] \tag{32}$$

with an error term that is a sum of a vanishing bias and a Martingale difference. Now consider the continuous time system $\theta \in \Theta$ in (32). We may write

$$\frac{dV_\theta(\mathbf{x}_0^\omega)}{dt} = \big(\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\big)^\top\ \Upsilon_\theta\big[-\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\big]. \tag{33}$$

We have the following cases:

Case 1: When $\theta \in \Theta^\circ$.

Since Θ° is the interior of the set Θ and Θ is a convex compact set, there exists a sufficiently small $\eta_0 > 0$ such that $\theta - \eta_0\nabla_\theta V_\theta(\mathbf{x}_0^\omega) \in \Theta$ and

$$\Gamma_\Theta\big(\theta - \eta_0\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\big) - \theta = -\eta_0\nabla_\theta V_\theta(\mathbf{x}_0^\omega).$$

Therefore, the definition of $\Upsilon_\theta\big[-\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\big]$ implies

$$\frac{dV_\theta(\mathbf{x}_0^\omega)}{dt} = -\left\|\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\right\|^2 \leq 0. \tag{34}$$

At the same time, $dV_\theta(\mathbf{x}_0^\omega)/dt < 0$ whenever $\|\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\| \neq 0$.
Case 2: When $\theta \in \partial\Theta$ and $\theta - \eta\nabla_\theta V_\theta(\mathbf{x}_0^\omega) \in \Theta$ for any $\eta \in (0, \eta_0]$ and some $\eta_0 > 0$.
The condition $\theta - \eta\nabla_\theta V_\theta(\mathbf{x}_0^\omega) \in \Theta$ implies that

$$\Upsilon_\theta\big[-\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\big] = -\nabla_\theta V_\theta(\mathbf{x}_0^\omega).$$

Then we obtain

$$\frac{dV_\theta(\mathbf{x}_0^\omega)}{dt} = -\left\|\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\right\|^2 \leq 0. \tag{35}$$

Furthermore, $dV_\theta(\mathbf{x}_0^\omega)/dt < 0$ when $\|\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\| \neq 0$.
Case 3: When $\theta \in \partial\Theta$ and $\theta - \eta\nabla_\theta V_\theta(\mathbf{x}_0^\omega) \notin \Theta$ for some $\eta \in (0, \eta_0]$ and any $\eta_0 > 0$.
For any $\eta > 0$, define $\theta_\eta := \theta - \eta\nabla_\theta V_\theta(\mathbf{x}_0^\omega)$. The above condition implies that when $0 < \eta \to 0$, $\Gamma_\Theta\big[\theta_\eta\big]$ is the projection of θ_η to the tangent space of Θ. For any elements $\hat{\theta} \in \Theta$, since the following set $\{\theta \in \Theta : \|\theta - \theta_\eta\|_2 \leq \|\hat{\theta} - \theta_\eta\|_2\}$ is compact, the projection of θ_η on Θ exists. Furthermore, since $f(\theta) := \frac{1}{2}\|\theta - \theta_\eta\|_2^2$ is a strongly convex function and $\nabla f(\theta) = \theta - \theta_\eta$, by first order optimality condition, one obtains

$$\nabla f(\theta_\eta^*)^\top(\theta - \theta_\eta^*) = (\theta_\eta^* - \theta_\eta)^\top(\theta - \theta_\eta^*) \geq 0, \quad \forall\theta \in \Theta$$

where θ_η^* is an unique projection of θ_η (the projection is unique because $f(\theta)$ is strongly convex and Θ is a convex compact set). Since the projection (minimizer) is unique, the above equality holds if and only if $\theta = \theta_\eta^*$.

Therefore, for any $\theta \in \Theta$ and $\eta > 0$,

$$\begin{aligned}
&\big(\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\big)^\top\ \Upsilon_\theta\big[-\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\big] \\
=&\big(\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\big)^\top \left(\lim_{0<\eta\to 0}\frac{\theta_\eta^* - \theta}{\eta}\right) \\
=&\left(\lim_{0<\eta\to 0}\frac{\theta - \theta_\eta}{\eta}\right)^\top \left(\lim_{0<\eta\to 0}\frac{\theta_\eta^* - \theta}{\eta}\right) \\
=&\lim_{0<\eta\to 0}\frac{-\|\theta_\eta^* - \theta\|^2}{\eta^2} + \lim_{0<\eta\to 0}\big(\theta_\eta^* - \theta_\eta\big)^\top\left(\frac{\theta_\eta^* - \theta}{\eta^2}\right) \leq 0.
\end{aligned}$$

From these arguments, one concludes that $dV_\theta(\mathbf{x}_0^\omega)/dt \leq 0$ and this quantity is non-zero whenever $\|\Upsilon_\theta\left[-\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\right]\| \neq 0$.

Now define the following Lyapunov function

$$\mathcal{V}_\theta(\mathbf{x}_0^\omega) = V_\theta(\mathbf{x}_0^\omega) - V_{\theta^*}(\mathbf{x}_0^\omega)$$

where θ^* is a local minimum point. Then there exists a ball centered at θ^* with radius r such that for any $\theta \in B_{\theta^*}(r)$, $\mathcal{V}_\theta(\mathbf{x}_0^\omega)$ is a locally positive definite function, i.e., $\mathcal{V}_\theta(\mathbf{x}_0^\omega) \geq 0$. On the other hand, by the definition of a local minimum point, one obtains $\Upsilon_\theta[-\nabla_\theta V_\theta(\mathbf{x}_0^\omega)]|_{\theta=\theta^*} = 0$ which means that θ^* is also a stationary point, i.e., $\theta^* \in \Theta_c$.

Note that $d\mathcal{V}_{\theta_{(t)}}(\mathbf{x}_0^\omega)/dt = dV_{\theta_{(t)}}(\mathbf{x}_0^\omega)/dt \leq 0$ and the time-derivative is non-zero whenever $\|\Upsilon_\theta\left[-\nabla_\theta V_\theta(\mathbf{x}_0^\omega)\right]\| \neq 0$. Therefore, by Lyapunov theory for asymptotically stable systems [**?**], the above arguments imply that with any initial condition $\theta_{(0)} \in B_{\theta^*}(r)$, the state trajectory $\theta_{(t)}$ of (32) converges to θ^*, i.e.,

$$V_{\theta^*}(\mathbf{x}_0^\omega) \leq V_{\theta_{(t)}}(\mathbf{x}_0^\omega) \leq V_{\theta_{(0)}}(\mathbf{x}_0^\omega)$$

for any $t \geq 0$.

Based on the above properties and noting that 1) $\nabla_\theta V_\theta(\mathbf{x}_0^\omega)$ is a Lipschitz function in θ, 2) the step-size rule follows from (23), 3) $\delta\theta_{k+1}$ is a square integrable Martingale difference, and 4) $\theta_k \in \Theta$, $\forall k$ implies that $\sup_k \|\theta_k\| < \infty$ almost surely, one can invoke Theorem 2 in Chapter 6 of [8] (multi-time scale stochastic approximation theory) to show that sequence $\{\theta_k\}$, $\theta_k \in \Theta$ converges almost surely to the solution of ODE (32) which further converges almost surely to $\theta^* \in \Theta$. ■

The above theorem shows that by setting the appropriate step-sizes ζ_k and ζ_k' using the rules in (23), the actor critic method converges to a stationary point. Furthermore, if the value function approximation error between $V_\theta(\mathbf{x}_0^\omega)$ and $\tilde{V}_\theta^{v^*}(\mathbf{x}_0^\omega)$ goes to zero, the converged policy parameter θ^* induces a locally optimal policy $\mu(\cdot|\cdot;\theta^*)$ for the stochastic shortest path problem $\mathcal{SP}$.

REMARK 5. *In the actor-critic method, the control policy space is parameterized using the Boltzmann family*

$$\mu(\mathbf{u}|x^\omega;\theta) = \frac{\exp(\gamma\theta^\top\phi_{\mathbf{U}}(x^\omega,\mathbf{u}))}{\sum_{u^j\in[0,\mathcal{A}^j],\forall j,\sum_{j=1}^S u^j=C}\exp(\gamma\theta^\top\phi_{\mathbf{U}}(x^\omega,u))}$$

where $\phi_{\mathbf{U}} \in \mathbb{R}^{\kappa_3}$ is the basis function of the policy space and $\gamma > 0$ is the temperature parameter of this policy family that controls the rate of exploration versus exploitation. Since the admissible control set $\mathbf{U}(x^\omega)$ may still be large (in the order of C^S), direct summation of the denominator in $\mu(\mathbf{u}|x^\omega;\theta)$ has a computational complexity in the order of C^S. While this brute-force computation may be intractable, one can form another family of parameterized policies by approximating the action space $\mathbf{U} = \{0,\ldots,C\}^S$ by the compact real set $[0,C]^S$, setting the policy as

$$\widehat{\mu}(\mathbf{u}'|x^\omega;\theta) = \frac{\exp(\gamma\theta^\top\phi_{\mathbf{U}}(x^\omega,\mathbf{u}'))}{\int_{u^j\in[0,\mathcal{A}^j],\forall j,\sum_{j=1}^S u^j=C}\exp(\gamma\theta^\top\phi_{\mathbf{U}}(x^\omega,u'))du'}$$

and obtaining the control action $\mathbf{u}$ by rounding $\mathbf{u}' \sim \widehat{\mu}(\mathbf{u}'|x^\omega;\theta)$ to its nearest integer vector.

REMARK 6. *In the above theorem, we established asymptotic limits for Algorithm 2 using the ODE approach. Comparing to existing reinforcement learning methods like SARSA and Q-learning [5], the major advantage of using Algorithm 2 is due to its nature of on-policy incremental updates on both value function approximation and policy parameters [20], [7]. While the critic attempts to update the value function approximation, the actor updates the policy parameters at the same time. This two-step interactive procedure provides more useful reinforcement feedbacks to the actor/critic counterparts, and it potentially leads to a better solution. However, to the best of our knowledge, there are no known convergence rate results available for actor-critic algorithms. It would be an interesting direction for future research to obtain finite-time convergence bounds for this algorithm.*

6. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel mathematical model on one-way vehicle sharing whose real time rental assignment is based on incentive bidding. By rigorously formulating this problem as a CMDP, we derive an exact solution algorithm whose solution can be found by solving a sequence of unconstrained stochastic shortest path problems (problem $\mathcal{SP}$). Furthermore, we also develop and analyze an iterative algorithm that effectively finds a near optimal vehicle-rental policy using the *actor-critic* method. This episodic approximation algorithm is important to the decision-maker, especially if number of stations and vehicles in the CMDP is large. Future work includes: **1)** Providing convergence proofs for our actor critic algorithm; **2)** Extending the current bidding mechanism using auction mechanisms [21] and algorithmic game theory [27]; and **3)** Evaluating our algorithms in a real life vehicle sharing platform.

REFERENCES

[1] E. Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

[2] C. Anderson. Setting prices on Priceline. 2009.

[3] M. Barth and M. Todd. Simulation model performance analysis of a multiple station shared vehicle system. *Transportation Research Part C: Emerging Technologies*, 7(4):237–259, 1999.

[4] D. Bertsekas. *Dynamic programming and optimal control*, volume 1 & 2. Athena Scientific, 1995.

[5] D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[6] S. Bhatnagar. An actor–critic algorithm with function approximation for discounted cost constrained Markov decision processes. *Systems & Control Letters*, 59(12):760–766, 2010.

[7] S. Bhatnagar et al. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

[8] V. Borkar. Stochastic approximation. *Cambridge Books*, 2008.

[9] G. Correia and A. Antunes. Optimization approach to depot location and trip selection in one-way carsharing systems. *Transportation Research Part E: Logistics and Transportation Review*, 48(1):233–247, 2012.

[10] A. Di Febbraro et al. One-way carsharing: Solving the relocation problem. *Transportation research record*, (2319):113–120, 2012.

[11] M. DiDonato. *City-bike maintenance and availability*. PhD thesis, Worcester Polytechnic Institute, 2002.

[12] V. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

[13] W. Fan et al. Carsharing: Dynamic decision-making problem for vehicle allocation. *Transportation Research Record: Journal of the Transportation Research Board*, 2063(1):97–104, 2008.

[14] L. Georgiadis et al. *Resource allocation and cross-layer control in wireless networks*. Now Publishers Inc, 2006.

[15] R. Gunther et al. Balancing bicycle sharing systems: Improving a VNS by efficiently determining optimal loading operations. volume 7919, pages 130–143, 2013.

[16] S. Kakade. A natural policy gradient. In *NIPS*, volume 14, pages 1531–1538, 2001.

[17] P. Keegan. Zipcar-The best new idea in business. *CNNMoney. com*, 2009.

[18] A. Kek et al. Relocation simulation model for multiple-station shared-use vehicle systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1986(1):81–88, 2006.

[19] J. Kendall. Toyota takes the i-road. *Automotive Engineering International*, 121(3), 2013.

[20] V. Konda and V. Borkar. Actor-critic–type learning algorithms for Markov decision processes. *SIAM Journal on control and Optimization*, 38(1):94–123, 1999.

[21] V. Krishna. *Auction theory*. Academic press, 2009.

[22] D. Mauro et al. The bike sharing rebalancing problem: Mathematical formulations and benchmark instances. *Omega*, 45:7–19, 2014.

[23] W. Mitchell. *Reinventing the automobile: Personal urban mobility for the 21st century*. MIT press, 2010.

[24] R. Nair. Fleet management for vehicle sharing operations. *Transportation Science*, 45(4):524–540, 2011.

[25] R. Nair et al. Large-scale vehicle sharing systems: Analysis of vélib'. *International Journal of Sustainable Transportation*, 7(1):85–106, 2013.

[26] M. Neely. Stochastic network optimization with application to communication and queueing systems. *Synthesis Lectures on Communication Networks*, 3(1):1–211, 2010.

[27] N. Nisan et al. *Algorithmic game theory*. Cambridge University Press, 2007.

[28] D. Papanikolaou et al. *The market economy of trips*. PhD thesis, Massachusetts Institute of Technology, 2011.

[29] L. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in Neural Information Processing Systems*, pages 252–260, 2013.

[30] E. Reynolds and K. McLaughlin. Autoshare: The smart alternative to owning a car. *Autoshare, Toronto, Ontario, Canada*, 2001.

[31] S. Ross. *Stochastic processes*, volume 2. John Wiley & Sons New York, 1996.

[32] A. Schmauss. Car2go in Ulm, Germany, as an advanced form of car-sharing. *European Local Transport Information Service (ELTIS)*, 2009.

[33] J. Shu et al. Bicycle-sharing system: deployment, utilization and the value of re-distribution. *National University of Singapore-NUS Business School, Singapore*, 2010.

[34] R. Sutton et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, pages 1057–1063. Citeseer, 1999.

[35] R. Tal et al. Static repositioning in a bike-sharing system: models and solution approaches. *European Journal of Transportation and Logistics*, 2:187–229, 2013.

[36] K. Uesugi et al. Optimization of vehicle assignment for car sharing system. In *Knowledge-based intelligent information and engineering systems*, pages 1105–1111. Springer, 2007.

[37] R. Urgaonkar and M. Neely. Opportunistic scheduling with reliability guarantees in cognitive radio networks. *Mobile Computing, IEEE Transactions on*, 8(6):766–777, 2009.

[38] L. Ying et al. On combining shortest-path and back-pressure routing over multihop wireless networks. *IEEE/ACM Transactions on Networking (TON)*, 19(3):841–854, 2011.