Multi-Agent Markov Decision Processes with Controlled Observations

Jayadev Joy
Department of Electrical and Computer Engineering
New York University
Brooklyn, United States
joy.jayadev@nyu.edu

Quanyan Zhu
Department of Electrical and Computer Engineering
New York University
Brooklyn, United States
quanyan.zhu@nyu.edu

***Abstract*—This paper conducts a thorough and comprehensive analysis of a continuous-time discounted jump Markov decision process, seamlessly integrating controlled actions and observations. A distinguishing characteristic of this study lies in its restriction of observations exclusively to discrete time points, necessitating a meticulous approach to strategically determine the optimal timing for subsequent observations and craft a control trajectory between these defined observation instances. To further advance our understanding, we extend this Markov decision process framework, involving controlled actions and observations, to a more intricate system setting that encompasses multiple agents or players. Notably, our investigation maintains a specific focus on the nuanced dynamics inherent in a two-player system, providing a deep and detailed exploration of the complexities that arise within this unique multi-agent context. By deliberately limiting our examination to the two-player scenario, our goal is to offer a concentrated and thorough exploration of the challenges and strategic considerations involved in extending controlled actions and observations within the broader framework of a continuous-time discounted jump Markov decision process.**

***Index Terms*—Multi-Agent, Markov Jump Process, Markov Decision Process, Controlled Observation, Dynamic Programming, Value Iteration, Queueing Systems**

I. Introduction

Markov Decision Processes (MDPs) constitute a versatile and widely employed mathematical framework for modeling and solving complex decision-making problems under uncertainty. Rooted in the domain of stochastic processes, MDPs find applications across diverse disciplines, including robotics, finance, operations research, and artificial intelligence. At their conceptual core lies the Markov property, a key feature that enables succinct representation by asserting that the future state of a system depends solely on the present state and the action taken, rendering the decision history irrelevant. This characteristic simplifies the intricate task of modeling dynamic systems subject to probabilistic transitions. In an MDP, decision-makers aim to optimize a predefined objective function, navigating a dynamic environment by selecting actions at each time step. The decisions are made by considering both the immediate rewards associated with chosen actions and the long-term cumulative impact on the system's state. This dual consideration captures the essence of strategic decision-making in the face of uncertainty and evolving conditions, establishing MDPs as a fundamental and indispensable tool in addressing challenges arising in dynamic and uncertain domains.

MDPs have established themselves as indispensable tools across diverse domains, including queueing systems, communication networks, and robotics, offering versatile solutions to real-world challenges. However, certain traditional MDP methodologies hinge on the presupposition of uninterrupted and easily accessible observations. This presumption proves inadequate in contexts where observations are constrained or present a significant economic burden. Take, for example, cyber-physical systems where the physical segregation of controllers and sensors results in intermittent sensing, driven by limitations in sensor capabilities such as finite battery life or constrained processing capacity. A parallel scenario unfolds in extensive networks like the Internet of Things (IoT), where the continuous data acquisition process incurs not only substantial costs but also proves unnecessary for optimal system operation. As we delve into these intricacies, the imperative to develop MDP frameworks that navigate these challenges becomes increasingly evident, ensuring applicability across a broader spectrum of dynamic and resource-constrained environments.

In addressing these intricate challenges, our proposed solution involves introducing a MDP framework carefully designed to function within controlled and restricted observation parameters. Our investigation focuses on a continuous-time jump MDP that incorporates discounted cost criteria, explicitly considering the financial implications associated with observations. Within this refined framework, decision-makers encounter the constraint of not having continuous access to state observations. Instead, they face the task of determining the optimal timing for the subsequent observation following each observation point, coupled with crafting an action trajectory during the interim period between these points. Consequently, the decision-maker is compelled to make a joint determination involving both the control trajectory and observation timings. This intricate scenario unravels a pivotal trade-off between actions and observations: while a wealth of information and observations empowers decision-makers to formulate more effective actions, thereby enhancing overall system performance, this advantage is counterbalanced by escalated communication expenses and increased power consumption. The nuanced

nature of this trade-off underscores the complexity inherent in optimizing decision-making processes within environments characterized by restricted and controlled observations.

We present summaries of two case studies that serve to illuminate the framework of controlled observations within the single-agent setting. The first case study explores a gated queueing system, where we derive an explicit dynamic programming equation. This allows us to characterize both the optimal observation and the optimal action explicitly. In the second case study, we delve into an inventory control problem featuring Poisson arrival and departure processes. Here, we employ value iterations to numerically determine the optimal observation points and corresponding optimal actions. The findings from these case studies unveil a discernible pattern: in states where the optimal action undergoes frequent adaptations, it proves optimal for the decision-maker to increase observation frequency. Conversely, in states where the optimal action experiences minimal changes, the decision-maker opts for less frequent observations. Additionally, we extend our exploration to a multi-agent setting, specifically examining the gated queuing system through two different approaches. It's worth noting that this investigation assumes that all servers have access to knowledge regarding the instantaneous server speeds of all servers.

Organization of the paper: In Section II, we review the single-agent MDP with controlled observations. We formulate, analyze, and develop a general theoretic framework of the problem of continuous-time jump MDP with controlled observation in a multi-agent system setting in Section III. Three case studies of a gated polling system (single server), inverntory control and gated polling system (multiple servers) are presented in Sections IV, V and VI respectively.

II. Single-Agent MDP with Controlled Observations [1]

Our initial exploration commences with a succinct review of a single-agent Markov Decision Process (MDP) characterized by controlled observations. This particular model finds relevance in scenarios where continuous updates of observations might be constrained or present a considerable cost. In this framework, the decision-maker assumes the crucial role of determining both the next observation time and the corresponding action (control trajectory) for that specific interval. This dual determination effectively leads to an intricate interplay between controlling trajectories and selecting observation points. It introduces a notable trade-off between actions and observations, with the overarching objective being the identification of optimal observation epochs and actions. The intricate dynamics inherent in this joint determination process underscore the complexity involved in striking an optimal balance between the quality of observations and the efficacy of actions within the context of a single-agent MDP with controlled observations.

A. Notations

In our examination, we delve into a continuous-time Markov Decision Process (MDP) denoted as $\{X(t) : X(t) \in S\}$, wherein the transitions are under the influence of control, and this control process is aptly represented by $A(t)$. Our focus extends to the consideration of stationary Markov policies, expressed in the form (with p chosen appropriately):

$$\pi(x) = \{(\alpha(x), \tau(x)) : \text{for any } x \in S\} \tag{1}$$

$$\alpha(x) = \{a(t) : a(.) \in L^p[t_x, t_x + T]\} \tag{2}$$

$$\tau(x) = \{T : T \in [\underline{T}, \infty]\} \tag{3}$$

Here, S denotes the state space, t_x represents the time at which the state was observed to be x, $\underline{T}$ signifies the minimum time gap between successive observations, and L^p denotes the space of p-integrable functions.

The transition probability, representing the likelihood that the continuous-time Markov Decision Process (MDP) resides in state $X(T) = x'$ after a duration T, assumes importance. This probability is conditioned on the initial state being $X(0) = x$ and the selection of the open-loop control $a(\cdot)$ for the specified time interval. Mathematically, it is expressed as:

$$\begin{aligned} &q(x, x'; a, T) \\ &= P(X(T) = x' | X(0) = x, A([0, T]) = a(.), T_1 = T) \end{aligned} \tag{4}$$

Given the restricted and controlled nature of observations, we introduce the concept of a consolidated reward, designed to encapsulate the time period between successive observations. Let $\bar{T}_k = \sum_{i<k} T_i$ represent the time at which the k^{th} observation is conducted, where T_k denotes the time interval between the k^{th} and $(k+1)^{th}$ observation epoch. The consolidated reward for this specific time period, spanning from the k^{th} to the $(k+1)^{th}$ observation epoch, is expressed as:

$$\begin{aligned} \bar{r}_k &= \bar{r}(X(\bar{T}_k), A(\bar{T}_k + .), T_k) \\ &= E\Big[\int_0^{T_k} \beta^t r(X(\bar{T}_k + t), A(\bar{T}_k + t) | X(\bar{T}_k), \\ &\qquad A(\bar{T}_k + .), T_k)\, dt \Big] \end{aligned} \tag{5}$$

B. Problem Formulation

Our objective is centered around optimizing the accumulated reward, meticulously constructed through the utilization of discounted values attributed to the consolidated rewards:

$$v(x) = \sup_{\pi(x)} J(\pi(x), x) \tag{6}$$

$$J(\pi(x), x) = \sum_{k=1}^{\infty} E\Big[\beta^{\bar{T}_k} \Big(\bar{r}_k - g(T_k) \Big) \Big] \tag{7}$$

C. Problem Analysis

Equations (6) and (7) seamlessly translate into the representation of a discounted cost discrete-time Markov Decision Process (MDP). This formulation incorporates a discount factor denoted as $\underline{\beta} = \beta^{\underline{T}}$, where β represents the discount factor and $\underline{T}$ corresponds to the time duration. The Markovian state and the action taken are respectively represented as:

$$Z_k = (X(\bar{T}_k), \tilde{T}_k) \tag{8}$$

$$A_k = (a_k(.), T_k) \tag{9}$$

Here, $\tilde{T}_k = \bar{T}_k - (k-1)\underline{T}$, and the running reward is expressed as:

$$R(Z_k, A_k) = \beta^{\tilde{T}_k}\Big(\bar{r}(X(\bar{T}_k), a(.), T_k) - g(T_k)\Big) \tag{10}$$

Adopting the previous notation changes, the optimization problem (7) can now be succinctly restated as:

$$J(\pi(x), x) = \sum_{k=1}^{\infty} \underline{\beta}^{k-1} R(Z_k, A_k) \tag{11}$$

With the groundwork laid in the current Markov Decision Process (MDP) literature, we are well-equipped to apply relevant results to the optimization problem outlined above. A notable observation is that the value function, as defined by (6), adheres to the following dynamic programming equation:

$$v(x) = \sup_{a,T}\Big(\bar{r}(x, a(.), T) + \beta^T \sum_{x' \in S} q(x, x'; a, T) v(x') - g(T)\Big) \tag{12}$$

If the optimal policy $\pi^*(x) = \{(\alpha^*(x), \tau^*(x)) : x \in S\} = \{(a^*(\cdot), T^*)\}$ exists, it must satisfy (13) for all $x \in S$.

$$v(x) = \bar{r}(x, a^*(.), T^*) + \beta^{T^*} \sum_{x' \in S} q(x, x'; a^*, T^*) v(x') - g(T^*) \tag{13}$$

The dynamic programming equations inherent in our model can be effectively tackled through the application of the value iteration method. Commencing with the appropriate initialization of $v_0(x)$ and armed with the k^{th} estimate of the value function, denoted as $\{v_k(x)\}_{x \in S}$, equation (13) becomes instrumental in computing the subsequent $(k+1)^{th}$ estimate, $\{v_{k+1}(x)\}_{x \in S}$. This iterative process is recurrently applied until the error falls below a predefined threshold, ensuring the achievement of proper convergence.

III. Multi-Agent MDP with Controlled Observations

Our examination now shifts towards extending the aforementioned model into a multi-agent setting, introducing the presence of multiple decision-makers or players within the system. This extension inherently adds another layer of complexity as the dynamics of the system are now intricately shaped by the actions of multiple players. Each player, in this context, assumes the responsibility of jointly determining both the control trajectory and the observation points, amplifying the intricacy of decision-making interactions within the system. This intricate web of decisions reflects a collaborative effort among the players to optimize both the trajectory of controls and the instances of observation. It's crucial to note that such multi-agent scenarios introduce a dynamic interplay among the decision-makers, each contributing to the overall evolution of the system. Our exploration is intentionally confined to the nuanced dynamics of a two-player system, aiming to delve deeply into the complexities inherent in extending controlled actions and observations within the broader context of a multi-agent environment.

A. Notations

In the upcoming discussions, we will leverage superscripts within parenthesis, (1) and (2), as explicit references to Players 1 and 2, respectively. This choice is deliberate, as subscripts traditionally carry the role of denoting observation indices. To ensure the equations remain both concise and transparent, we embrace a notational approach wherein $i \in \{1, 2\}$. This consistent framework allows us to systematically define and articulate the various components constituting our Markov Decision Process (MDP). Specifically, we center our exploration around stationary Markov policies, adopting a formal expression in the form (with p appropriately chosen):

$$\pi^{(i)}(x) = \{(\alpha^{(i)}(x), \tau^{(i)}(x)) : \text{for any } x \in S\} \tag{14}$$

$$\alpha^{(i)}(x) = \{a^{(i)}(t) : a^{(i)}(.) \in L^p[t_x^{(i)}, t_x^{(i)} + T^{(i)}\} \tag{15}$$

$$\tau^{(i)}(x) = \{T^{(i)} : T^{(i)} \in [\underline{T}^{(i)}, \infty]\} \tag{16}$$

Here, S denotes the state space, $t_x^{(i)}$ represents the time at which player i observed the state to be x, $\underline{T}^{(i)}$ signifies the minimum time gap between successive observations for the respective player, and L^p denotes the space of p-integrable functions.

The transition probability, indicative of the likelihood that the Markov Decision Process (MDP) resides in state $X(T) = x'$ after a duration T, is of paramount significance. This probability is conditioned on the initial state being $X(0) = x$, and it takes into account the selection of open-loop controls $a^{(1)}(\cdot)$ and $a^{(2)}(\cdot)$ for the specified time interval. Mathematically, it is expressed as:

$$q(x, x'; a^{(1)}, a^{(2)}, T) = P(X(T) = x' | X(0) = x, \\ A^{(1)}([0, T]) = a^{(1)}(.), A^{(2)}([0, T]) = a^{(2)}(.), T_1 = T) \tag{17}$$

The consolidated reward for the time period between the k^{th} and $(k+1)^{th}$ observation epoch for each player is given by:

$$\begin{aligned} \bar{r}_k^{(i)} &= \bar{r}^{(i)}(X(\bar{T}_k^{(i)}), A^{(1)}(\bar{T}_k^{(i)} + .), A^{(2)}(\bar{T}_k^{(i)} + .), T_k^{(i)}) \\ &= E\Big[\int_0^{T_k^{(i)}} (\beta^{(i)})^t r(X(\bar{T}_k^{(i)} + t), A^{(1)}(\bar{T}_k^{(i)} + t), \\ &\quad A^{(2)}(\bar{T}_k^{(i)} + t) | X(\bar{T}_k^{(i)}), a^{(i)}(\bar{T}_k^{(i)} + .), T_k^{(i)})\, dt\Big] \end{aligned} \tag{18}$$

Here, $\bar{T}_k^{(i)} = \sum_{j<k} T_j^{(i)}$ denotes the time at which the k^{th} observation is made by the respective player.

B. Problem Formulation

Our goal is to optimize the accumulated reward, meticulously constructed through the utilization of discounted values attributed to the consolidated rewards:

$$v^{(i)}(x) = \sup_{\pi^{(i)}(x)} J^{(i)}(\pi^{(1)}(x), \pi^{(2)}(x), x) \tag{19}$$

$$J^{(i)}(\pi^{(1)}(x), \pi^{(2)}(x), x) = \sum_{k=1}^{\infty} E\Big[(\beta^{(i)})^{\bar{T}_k^{(i)}} \Big(\bar{r}_k^{(i)} - g^{(i)}(T_k^{(i)})\Big)\Big] \tag{20}$$

C. Problem Analysis

Equations (19) and (20) lend themselves to representation as discrete-time Markov Decision Processes (MDPs) for each player, incorporating a discounted cost framework. The discount factor is denoted as $\underline{\beta}^{(i)} = (\beta^{(i)})^{\underline{T}^{(i)}}$, where $\beta^{(i)}$ represents the discount factor and $\underline{T}^{(i)}$ corresponds to the time duration. The Markovian state and the action taken are represented respectively as:

$$Z_k^{(i)} = (X(\bar{T}_k^{(i)}), \tilde{T}_k^{(i)}) \tag{21}$$

$$W_k^{(i)} = (a_k^{(1)}(.), a_k^{(2)}(.), T_k^{(i)}) \tag{22}$$

Here, $\tilde{T}_k^{(i)} = \bar{T}_k^{(i)} - (k-1)\underline{T}^{(i)}$, and the running reward is expressed as:

$$R^{(i)}(Z_k^{(i)}, W_k^{(i)}) = (\beta^{(i)})^{\tilde{T}_k^{(i)}} \Big(\bar{r}^{(i)}(X(\bar{T}_k^{(i)}), a^{(1)}(.), a^{(2)}(.), T_k^{(i)}) - g^{(i)}(T_k^{(i)})\Big) \tag{23}$$

Adopting the previous notation changes, the optimization problem (7) can now be succinctly restated as:

$$J^{(i)}(\pi^{(1)}(x), \pi^{(2)}(x), x) = \sum_{k=1}^{\infty} (\underline{\beta}^{(i)})^{k-1} R^{(i)}(Z_k^{(i)}, W_k^{(i)}) \tag{24}$$

With the groundwork laid in the existing Markov Decision Process (MDP) literature, we are poised to apply pertinent findings to the optimization problem at hand. It becomes evident that the value function, as defined in (6), adheres to the following dynamic programming equation:

$$v^{(i)}(x) = \sup_{a^{(i)}, T^{(i)}} \Big(\bar{r}^{(i)}(x, a^{(1)}(.), a^{(2)}(.), T^{(i)}) + (\beta^{(i)})^{T^{(i)}} \sum_{x' \in S} q(x, x'; a^{(1)}, a^{(2)}, T^{(i)}) v^{(i)}(x') - g^{(i)}(T^{(i)})\Big) \tag{25}$$

If the optimal policy $\pi^{(i)*}(x) = \{(\alpha^{(i)*}(x), \tau^{(i)*}(x)) : x \in S\} = \{(a^{(i)*}(\cdot), T^{(i)*})\}$ exists, it must satisfy (26) for all $x \in S$.

$$v^{(i)}(x) = \bar{r}^{(i)}(x, a^{(i)*}(.), a^{-(i)}(.), T^{(i)*}) + (\beta^{(i)})^{T^{(i)*}} \sum_{x' \in S} q(x, x'; a^{(i)*}(.), a^{-(i)}(.), T^{(i)*}) v^{(i)}(x') - g^{(i)}(T^{(i)*}) \tag{26}$$

The dynamic programming equations inherent in our model find a resolution through the application of the value iteration method. Commencing with the appropriate initialization of $v_0^{(i)}(x)$ and armed with the k^{th} estimate of the value function, denoted as $\{v_k^{(i)}(x)\}_{x \in S}$, equation (26) is instrumental in computing the subsequent $(k+1)^{th}$ estimate $\{v_{k+1}^{(i)}(x)\}_{x \in S}$. This iterative process continues until the error falls below a predefined threshold, ensuring the attainment of proper convergence.

IV. Case Study I: Gated Queuing Systems (Single Server)

In examining a queuing system designed as a dynamic gated polling system, our focus extends to the control of gate openings and the dynamic adjustment of server speed in response to the number of customers. The observation epochs in this context align with instances of gate openings. Conceptually, this system can be envisioned as having two waiting rooms: upon gate opening, all customers from the outer room enter the inner room, and the gate promptly closes. A designated server speed is then allocated to handle the workload in the inner room. Simultaneously, the determination of the next gate opening instance takes place, while arriving customers accumulate in the outer room until the gate reopens.

The influx of arrivals into this system is modeled through a Poisson process characterized by a constant rate, denoted as . Service times for customers are assumed to be independent and identically distributed, provided they are served at the same speed. The objective of the agent is to optimize a discounted cost related to the waiting times of the customers. We see that the expected waiting time of all the customers outside the gate during the k^{th} observation period is:

$$\frac{\lambda T_k^2}{2} \tag{27}$$

If the server speed is a_k and if X_{k-1} is the number of customers entering the room at $(k-1)^{th}$ gate open instance, we have:

$$E[W_k | X_{k-1}, T_k, a_k] = \frac{\lambda T_k^2}{2} + \frac{X_{k-1}^2 + X_{k-1}}{2a_k} \tag{28}$$

Taking the cost for frequent observations and the cost for server speed into account, we would like to optimize the following:

$$\sum_{k=1}^{\infty} E\left[\beta^{\bar{T}_k}\left(\frac{\lambda T_k^2}{2} + \frac{X_{k-1}^2 + X_{k-1}}{2A_k} + g(T_k) + \eta(A_k)\right)\right] \tag{29}$$

This is an example of separable utilities with:

$$\bar{r}_a(x, a) = \frac{x^2 + x}{2a} + \eta(a) \tag{30}$$

$$\bar{r}_T(x, T) = \frac{\lambda T^2}{2} \tag{31}$$

Further, the transition probabilities are independent of the action A_k:

$$q(x, x'; a, T) = e^{-\lambda T} \frac{(\lambda T)^{x'}}{(x')!} \tag{32}$$

Assuming a linear server speed cost $\eta(a) = \eta a$, we immediately arrive at:

$$r_x^*(x) = \sqrt{2\eta x(x+1)} \tag{33}$$

$$a_x^*(x) = \sqrt{\frac{x(x+1)}{2\eta}} \tag{34}$$

Solving the DP equations for this example numerically, we arrive at the optimal policy to be:

$$A_x^* = \sqrt{\frac{x(x+1)}{2\eta}} \tag{35}$$

$$T_x^* = T^* \tag{36}$$

for some constant T^*. It is interesting to observe that the server speed depends upon the number of customers who have entered the inner room, while the optimal observation epochs are independent of the state of the system.

V. Case Study II: Inventory Control

In this specific case study, our examination delves into the intricacies of managing inventory in the context of energy harvesting. The primary goal is to carefully maintain an inventory level proximate to θ. Notably, both arrivals and departures unfold as random events, with each arrival or departure corresponding to a singular unit. Departures transpire in adherence to a Poisson process characterized by a fixed rate μ. On the other hand, arrivals follow an inhomogeneous Poisson process, introducing a controlled time-varying rate denoted as $a()$. The nuanced dynamics inherent in these processes add a layer of complexity to the inventory control challenge, prompting an exploration of strategic approaches to sustain the desired inventory level amidst the stochastic nature of these events.

At each observation epoch, we define a fixed time interval $[0, T]$. Let X_k represent the inventory amount at the conclusion of the k^{th} observation. Specifically, with each arrival, one unit is incrementally added to the inventory, while each demand departure results in a reduction of one unit. Consequently, the inventory level at any given time before the $(k+1)^{th}$ epoch can be expressed as:

$$X_k(t) = X_k + \mathcal{A}(t; a) - \mathcal{D}(t) = X_k + \sum_{i=1}^{\mathcal{N}(t;a)} \xi_i \tag{37}$$

where $\mathcal{A}(t; a)$ and $\mathcal{D}(t)$ are the number of arrivals and departures by time t, respectively, since the last observation and $\mathcal{N}(t; a)$ is the total number of arrivals and departures with $\xi_i = 1$ if it is arrival and -1 if it is departure. The overall utility depends on the cost spent on the acceleration process and the deviation from the targeted inventory θ which is:

$$\int_0^\infty \beta^t \{E[(X_k(t) - \theta)^2] + \nu a(t)\} dt \tag{38}$$

The consolidated utility for this example turns out to be:

$$r(x, a(t), t) = (x - \theta + \bar{a}(t) - \mu t)^2 + \bar{a}(t) + \mu t + a(t)\nu \tag{39}$$

The transition probability given the control $a(.)$ and time period T would be:

$$q(x, x`; a, T) = \begin{cases} \sum_{k=x`-x}^{\infty} \frac{e^{-\bar{a}T-\mu T}(\bar{a}T)^k(\mu T)^{k-x`+x}}{k!(k-x`+x)!}, & \text{if } x` > x \\ \sum_{k=x-x`}^{\infty} \frac{e^{-\bar{a}T-\mu T}(\bar{a}T)^{k-x+x`}(\mu T)^k}{k!(k-x+x`)!}, & \text{else} \end{cases} \tag{40}$$

In each iteration k, for a given x, we need to solve the following optimal control problem:

$$v_{k+1}(x) = \inf_{a \in L^\infty[0,T]} \int_0^T \beta^t r(x, a(t), t) dt + h(\bar{a}(T), T) \tag{41}$$

where the terminal cost is

$$h(\bar{a}(T), T) = \beta^T \sum_{x`} q(x, x`; a, T) v(x`) + g(T) \tag{42}$$

At iteration k, we need to solve the following optimal control problem for any given x:

$$\inf_{a \in L^\infty[0,T]} \int_0^T \beta^t \{(x - \theta + y(t) - \mu t)^2 + y(t) + \mu t + a(t)\nu\} dt + h(y(T), T) \quad s.t. \quad \dot{y}(t) = a(t), \quad y(0) = 0 \tag{43}$$

The Hamiltonian of the above problem is given by

$$H(t, a, y, \lambda) = \beta^t \{(x - \theta + y - \mu t)^2 + y + \mu t + a\nu\} + \lambda a \tag{44}$$

where λ is the costate. With the application of minimum principle, the optimal solutions a^* and the corresponding state y^* need to satisfy the following conditions:

$$\dot{y}^*(t) = a^*(t) \tag{45}$$

$$\dot{\lambda} = -\beta^t \{2(x - \theta + y^*(t) - \mu t) + 1\} \tag{46}$$

$$\lambda(T) = \frac{\partial h}{\partial y}(y(T), T) \tag{47}$$

$$a^*(t) = \begin{cases} 0, & \text{if } \beta^t \nu + \lambda > 0 \\ \bar{\bar{a}}, & \text{otherwise} \end{cases} \tag{48}$$

When the Poisson arrival process is homogeneous, the problem becomes a finite dimensional optimization problem with

$$v_{k+1}(x) = \inf_{a \in [0, \bar{\bar{a}}], T} \int_0^T \beta^t r(x, at, t) dt + h(aT, T) \tag{49}$$

Hence, we have

$$v_{k+1}(x) = \min_{a \in [0, \bar{\bar{a}}], T} \int_0^T \beta^t \{(x - \theta + at - \mu)^2 + at + \mu t\} dt + h(aT, T) \tag{50}$$

VI. Case Study III: Gated Queuing Systems (Multiple Servers)

In this case study, we delve into the intricacies of a gated queuing system featuring dynamic gate control independently managed by two players. The distinctive aspect of this setup lies in the shared utilization of a common sensor by both players. This sensor serves the crucial function of observing the system's state independently when each player makes decisions. Notably, the cost associated with these observations is not determined in isolation; rather, it is jointly influenced by the collaborative usage of the sensor by both players. Within this intricate problem context, the quest for the optimal policy for both players takes on the metaphorical role of the sensor allocating a server speed to each server during observation epochs, precisely when customers are introduced into the system. Given the cooperative alignment of both players toward a shared objective, an assumption is made that they possess mutual awareness of each other's actions on the system. This awareness extends to each server being cognizant of the instantaneous server speed at which the other operates, fostering a collaborative and well-informed decision-making process.

The influx of arrivals into this system is modeled through a Poisson process characterized by a constant rate, denoted as . Service times for customers are assumed to be independent and identically distributed, provided they are served at the same speed. The objective of both the players is to optimize a discounted cost related to the waiting times of the customers. We see that the expected waiting time of all the customers outside the gate during the k^{th} observation period is:

$$\frac{\lambda T_k^2}{2} \tag{51}$$

Let us assume the server speed to be a_k and X_{k-1} to be the number of customers entering the room at $(k-1)^{th}$ gate open instance.

A. Method I

In the initial approach, our strategy revolves around the assignment of server speeds for each server, utilizing the variable X_{k-1}. The collective cost in this context is encapsulated by the discounted expected waiting time, represented as $\sum_{k=1}^{\infty} \beta^{\bar{T}_k} E[W_k]$. This consolidated cost provides a comprehensive measure, capturing the cumulative impact of discounted waiting times across multiple epochs in the system. It serves as a key metric in evaluating the overall efficiency and performance of the system under consideration. We have:

$$E[W_k|X_{k-1}, T_k, a_k^{(1)}, a_k^{(2)}] = \frac{\lambda T_k^2}{2} + \frac{X_{k-1}^2 + 2X_{k-1}}{2(a_k^{(1)} + a_k^{(2)})} \tag{52}$$

Taking the cost for frequent observations and the cost for server speed into account, we would like to optimize the following:

$$J(\pi(x), x) = \sum_{k=1}^{\infty} E\left[\beta^{\bar{T}_k}\left(\frac{\lambda T_k^2}{2} + \frac{X_{k-1}^2 + 2X_{k-1}}{2(a_k^{(1)} + a_k^{(2)})} + g(T_k) + \eta^{(1)}(A_k^{(1)}) + \eta^{(2)}(A_k^{(2)})\right)\right] \tag{53}$$

Further, the transition probabilities are independent of the action A_k:

$$q(x, x'; a^{(1)}, a^{(2)}, T) = e^{-\lambda T}\frac{(\lambda T)^{x'}}{(x')!} \tag{54}$$

Assuming a linear server speed cost $\eta(a) = \eta a$, the optimal policy for the above problem is to always assign the entire load to the server with the lower server speed cost and the corresponding speed and observation epochs are given by:

$$A_x^* = \sqrt{\frac{x(x+1)}{2\eta}} \tag{55}$$

$$T_x^* = T^* \tag{56}$$

for some constant T^*. It is interesting to observe that the server speed depends upon the number of customers who have entered the inner room, while the optimal observation epochs are independent of the state of the system. We also notice that in this approach, the optimal policy resembles that of a single server setting, with the role of the active server taken by the server with lower server speed cost.

B. Method II

In this particular approach, our focus lies in the allocation of the number of customers to each server, derived from the total number of customers in the system. Subsequently, the optimal policy for each server is determined based on the assigned number of customers. The associated cost in this scenario is defined as the discounted expected waiting time, expressed as $\sum_{k=1}^{\infty} \beta^{\bar{T}_k} E[W_k]$. Given specific values for $X_{k-1}^{(1)}$, $X_{k-1}^{(2)}$, and T_k, and assuming a linear server speed cost denoted as $\eta(a) = \eta a$, we have:

$$E[W_k|X_{k-1}^{(1)}, X_{k-1}^{(1)}, T_k] = \frac{\lambda T_k^2}{2} + \sqrt{2\eta^{(1)} X_{k-1}^{(1)}(X_{k-1}^{(1)} + 1)} + \sqrt{2\eta^{(2)} X_{k-1}^{(2)}(X_{k-1}^{(2)} + 1)} \tag{57}$$

Taking the cost for frequent observations and the cost for server speed into account, we would like to optimize the following:

$$J(\pi(x), x) = \sum_{k=1}^{\infty} E\left[\beta^{\bar{T}_k}\left(\frac{\lambda T_k^2}{2} + \sqrt{2\eta^{(1)} X_{k-1}^{(1)}(X_{k-1}^{(1)} + 1)} + \sqrt{2\eta^{(2)} X_{k-1}^{(2)}(X_{k-1}^{(2)} + 1)} + g(T_k)\right)\right] \tag{58}$$

Further, the transition probabilities are independent of the actions $A_k^{(1)}$ and $A_k^{(2)}$:

$$q(x, x'; a^{(1)}, a^{(2)}, T) = e^{-\lambda T} \frac{(\lambda T)^{x'}}{(x')!} \tag{59}$$

The solution to the provided numerical expression defines the optimal policy for the aforementioned problem:

$$\sqrt{\frac{\eta^{(1)}}{\eta^{(2)}}} = \frac{2x^{(2)} + 1}{2x^{(1)} + 1} \sqrt{\frac{x^{(1)}(x^{(1)} + 1)}{x^{(2)}(x^{(2)} + 1)}} \tag{60}$$

$$T_x^* = T^* \tag{61}$$

The optimal server speeds for the servers follows directly from $x^{(1)}$ and $x^{(2)}$:

$$A_k^{(1)*} = \sqrt{\frac{x^{(1)}(x^{(1)} + 1)}{2\eta^{(1)}}} \tag{62}$$

$$A_k^{(2)*} = \sqrt{\frac{x^{(2)}(x^{(2)} + 1)}{2\eta^{(2)}}} \tag{63}$$

for some constant T^*. It is interesting to observe that the server speed depends upon the number of customers allotted to each server, while the optimal observation epochs are independent of the state of the system. We also notice that in this approach, the optimal policy resembles that of a single server setting, with the role of the number of customers in the system replaced by the number of customers allotted to each server.

VII. Conclusion

In this manuscript, we have provided a comprehensive summary and examination of continuous-time jump Markov decision processes, focusing on the joint control of actions and observation epochs within the context of a single-agent system. Furthermore, we have expanded this framework to encompass a multi-agent system setting. The continuous-time jump MDP model has been reformulated into a conventional MDP problem through the creation of consolidated utilities between two observation epochs. We derive dynamic programming equations that facilitate the application of value iterations, enabling the characterization of optimal times for the next observation and optimal control trajectories. Our exploration encompasses two case studies within the single-agent system paradigm. The first case study delves into the theoretical characterization of optimal observation and action in a gated queueing system. The second case study addresses an inventory control problem featuring a Poisson arrival process, necessitating numerical computations for optimal observation epochs and actions. Additionally, we extend our investigation to a multi-agent setting, specifically examining the gated queuing system using two distinct approaches. It is noteworthy that this inquiry assumes universal access to knowledge regarding the instantaneous server speeds among all servers. Future endeavors will delve into the analysis of a multiple-server gated queuing system, where centralized knowledge of instantaneous server speeds is not universally accessible. This analysis will necessitate the integration of a game-theoretic model into the MDP framework, adding a layer of complexity to the study.

References

[1] Yunhan Huang, Veeraruna Kavitha and Quanyan Zhu, "Continuous-Time Markov Decision Processes with Controlled Observations," 2019

[2] M. L. Puterman, Markov Decision Processes.: Discrete Stochastic Dynamic Programming. John Wiley Sons, 2014.

[3] R. Durrett, Probability: theory and examples. Cambridge university press, 2019, vol. 49.