Taylor & Francis
Taylor & Francis Group

International Journal of Control

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tcon20

Viscosity solutions approach to finite-horizon continuous-time Markov decision process

Zhong-Wei Liao & Jinghai Shao

To cite this article: Zhong-Wei Liao & Jinghai Shao (09 Apr 2024): Viscosity solutions approach to finite-horizon continuous-time Markov decision process, International Journal of Control, DOI: 10.1080/00207179.2024.2338448

To link to this article: https://doi.org/10.1080/00207179.2024.2338448

Published online: 09 Apr 2024.

Internat. J. Control 98 (2025), no. 2, 369 – 379.

INTERNATIONAL JOURNAL OF CONTROL
https://doi.org/10.1080/00207179.2024.2338448

Taylor & Francis
Taylor & Francis Group

Viscosity solutions approach to finite-horizon continuous-time Markov decision process

Zhong-Wei Liao[a] and Jinghai Shao[b]

[a]Faculty of Arts and Sciences, Beijing Normal University, Zhuhai, People's Republic of China; [b]Center for Applied Mathematics, Tianjin University, Tianjin, People's Republic of China

ABSTRACT

This paper investigates the optimal control problems for the finite-horizon continuous-time Markov decision processes with delay-dependent control policies. We develop compactification methods in decision processes and show that the existence of optimal policies. Subsequently, through the dynamic programming principle of the delay-dependent control policies, the differential-difference Hamilton-Jacobi-Bellman (HJB) equation in the setting of discrete space is established. Under certain conditions, we give the comparison principle and further prove that the value function is the unique viscosity solution to this HJB equation. Based on this, we show that among the class of delay-dependent control policies, there is an optimal one which is Markovian.

ARTICLE HISTORY
Received 11 July 2023
Accepted 28 March 2024

KEYWORDS
Markov decision process; finite-horizon criterion; delay-dependent control; viscosity solution; HJB equation

2020 MATHEMATICS SUBJECT CLASSIFICATIONS
90C40; 93E20; 60J27

1. Introduction

Continuous-time Markov decision processes (CTMDPs) have been studied intensively due to their rich application in queuing systems, population processes, see, e.g. the monographs (Baüerle & Rieder, 2011; Ghosh & Saha, 2012; Guo & Hernández-Lerma, 2009; Prieto-Rumeau & Hernández-Lerma, 2012) and the extensive references therein. From the viewpoint of realistic applications, it is natural to investigate the optimal control problem with delay-dependent controls. The delay caused in the approach of observing the state of the system, making a decision based on this state, and then inputting this decision back into the studied system. However, the system maybe has changed its state at that time. More generally, this control policies are also known as history-dependent control policies, see, for example, Guo et al. (2015, 2012), Guo and Liao (2019), Huang (2018), Kumar and Chandan (2015), Piunovskiy and Zhang (2011), Prieto-Rumeau and Lorenzo (2010), and Zhang (2017). In this work we develop the viscosity solutions approach of CTMDPs, and it is worth noting that due to the consideration of delay-dependent controls, the controlled system is no longer a Markovian process. For this reason, relevant theoretical tools, such as compactification methods, comparison principle, differential-difference HJB equations and viscosity solutions approach, have also been discussed again.

It is a fundamental problem in the study of MDPs to distinguish the impact on the value function by taking account of all history-dependent policies or of merely Markovian policies. For the discrete-time MDPs in a finite state space, Derman and Strauch (1966, Theorem 2) established a basic result which implies that for any history-dependent policy there exists a randomised Markovian policy such that the associated controlled process admits the same marginal state-action distributions. This result also implies that with respect to the criteria of expected discounted, non-discounted costs and expected average costs, the optimisation problem over history-dependent policies and over Markovian policies will derive the same value function, see Derman and Strauch (1966) and Feinberg et al. (2013). For the situations of infinite state spaces or unbounded cost functions, more cautious research methods are needed. We constructed an explicit example (see Appendix) to illustrate that if there are no appropriate constraints on the transition probability matrix and cost function, the value function on the history-dependent policies is not equal to that on the Markovian policies. Therefore, discussing appropriate constraints to ensure consistency of the value function across different policy sets is also one of the topics of this article.

As is well known, the expected finite-horizon criterion is a widely used optimality criterion for CTMDPs optimisation problems, which has been studied by numerous works, see e.g. Baüerle and Rieder (2011), Ghosh and Saha (2012), Guo et al. (2015), Miller (1968), Pliska (1975), and Yushkevich (1978). For finite-horizon CTMDPs with finite state and action space, Miller (1968) gave a necessary and sufficient condition for the existence of a piecewise constant optimal policy. Subsequently, the state space of CTMDPs had been generalised to denumerable space (cf. Yushkevich, 1978) and Borel space (cf. Pliska, 1975), and the existence of an optimal Markov policy had been proven under the bounded hypothesis of transition rates and cost functions. Recently, Baüerle and Rieder (2011) studies

CONTACT Zhong-Wei Liao ✉ zhwliao@hotmail.com 🖃 Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, People's Republic of China

the finite-horizon CTMDPs with Markov polices by a method based on the equivalent transformation from finite-horizon CTMDPs to infinite-horizon discrete-time Markov decision processes. The corresponding optimality equation had been established according to the existing theory on discrete-time Markov decision processes. In addition, Ghosh and Saha (2012) considered the finite-horizon CTMDPs in Borel state space with bounded transition rates and Markov policies. The existence of a unique solution to the optimality equation is guaranteed by the Banach fixed point theorem, relatively, the existence of an optimal Markov policy is based on Itô-Dynkin's formula. The finite-horizon CTMDPs with unbounded transition rates are investigated in Guo et al. (2015).

The work (Guo et al., 2015) also studied the history-dependent control problem for jumping processes. The precise construction of such kind of controlled system is presented. However, via the main result (Guo et al., 2015, Theorem 4.1), the value function $V^*(t,i)$ $(t>0)$, defined in Guo et al. (2015, p.1069), associated with the optimal control problem over the set of randomised Markov policies can be characterised as a unique solution to a differential equation, and in such case the optimal Markov control policies are shown to exist. Nevertheless, if considering the control problem over the set of history-dependent control policies, there is no result in Guo et al. (2015) on the existence of the optimal control and on the characterisation of the associated value function. In the current work, we shall show the existence of the optimal controls over the set of delay-dependent controls and characterise the associated value function.

The approaches used in the aforementioned works in the study of CTMDPs rely on the characterisation of the Markov chains, and are not suitable to our current situation any longer since the controlled process is no longer a Markovian one caused by the delays. We develop the compactificion method used usually in the control problem for diffusion processes to the setting of jumping processes in order to show the existence of the optimal delay-dependent control policies. This is the starting point of this work. Precisely, the main contributions of the present paper are as follows:

(i) In comparison with Ghosh and Saha (2012) and Guo et al. (2015), our method used in the existence of an optimal delay-dependent control does not involve the solvability of the optimality equation, but is based on the compactification method, which is an effective method in the research of the optimal control problem of jump-diffusion processes, cf. Chow et al. (1985), Dufour and Miller (2006), Haussmann and Suo (1995a, 1995b). The basic idea is inspired by Kushner (1975), Haussmann and Suo (1995a, 1995b). Our approach is also suitable to other optimality criteria in the study of CTMDPs such as expected discounted, average and risk-sensitive.

(ii) According to the measurable selection theorem (cf. Stroock & Varadhan, 1979), the dynamic programming principle is established in Theorem 4.1, which deduces that the value function is a solution to a HJB equation provided the value function to be regular enough. Here the HJB equation is a differential-difference equation. We develop the viscosity solution approach to such equation, and especially we establish the comparison principle for such differential-difference HJB equation. In the second-order HJB equations, Jensen (1988) extended the classical Alexandrov Theorem (Alexandrov, 1939) to semi-convex functions, providing the results of comparison principle and establishing the uniqueness of viscosity solutions in this context. Later the uniqueness result was generalised by Ishii (1984, 1989) to the equations satisfying standard Lipschitz regularity assumptions. While the differential-difference HJB equation studied in our paper differs from second-order partial differential equations, the comparison principle and the uniqueness of viscosity solutions we present are inspired by the aforementioned results.

The rest of our paper is organised as follows. In Section 2, we state the concept of delay-dependent controls and the optimality problems of CTMDPs, and further introduced the main assumptions of this article. For the convenience, the optimality problem is reformulated on the canonical path space. In Section 3, by developing the compactification method within the framework of MDPs, we prove the existence of the optimal delay-dependent controls. In Section 4, we study the HJB equation derived from the dynamic programming principle through the viscosity solution approach. In order to prove the existence and uniqueness of viscosity solution, we also prove the comparison principle under the framework of MDPs. Invoking the corresponding results on the optimal control problem over Markovian control policies, we further show that there must exist an optimal Markovian control policy for the control problem over the class of delay-dependent control policies.

2. Formulation and assumptions

The objective of this section is to describe briefly the controlled process and the associated optimal control criterion in this paper. Let $(\Omega,\mathscr{F},(\mathscr{F}_t)_{t\geqslant 0},\mathbb{P})$ be a filtered probability space satisfying the usual conditions, i.e. $(\Omega,\mathscr{F},\mathbb{P})$ is complete, the filtration $(\mathscr{F}_t)_{t\geqslant 0}$ is right-continuous and $\mathscr{F}_0$ contains all $\mathbb{P}$-null sets in $\mathscr{F}$. Let $\mathscr{S}=\{1,2,\dots\}$ be the countable state space, U be the action space which is a compact subset of $\mathbb{R}^k$ for some $k\in\mathbb{N}$. Denote by $\mathscr{P}(U)$ the collection of all probability measures over U, which is endowed with L_1-Wasserstein distance W_1 defined by:

$$W_1(\mu,\nu)=\inf\left\{\int_{U\times U}|x-y|\pi(\mathrm{d}x,\mathrm{d}y);\ \pi\in\mathscr{C}(\mu,\nu)\right\},$$

where $\mathscr{C}(\mu,\nu)$ stands for the set of all couplings of μ and ν in $\mathscr{P}(U)$. Since U is compact, $\mathscr{P}(U)$ becomes a compact Polish space under the metric W_1, and the weak convergence of probability measures in $\mathscr{P}(U)$ is equivalent to the convergence in the W_1 distance (cf. e.g. Ambrosio et al., 2005, Chapter 7). In this work we investigate the finite-horizon optimal control problem on $[0,T]$, where $T>0$ is fixed throughout this work.

For each $\mu\in\mathscr{P}(U)$, $(q_{ij}(\mu))$ is a transition rate matrix over the state space $\mathscr{S}$, which is assumed to be conservative, i.e.

$$\sum_{j\neq i}q_{ij}(\mu)=q_i(\mu)=-q_{ii}(\mu),\quad \forall\, i\in\mathscr{S},\ \mu\in\mathscr{P}(U).$$

The process (Λ_t) is an $\mathscr{F}_t$-adapted jump process on $\mathscr{S}$ satisfying

$$\mathbb{P}(\Lambda_{t+\delta}=j\,|\,\Lambda_t=i,\mu_t=\mu)\\ =\begin{cases} q_{ij}(\mu)\delta+o(\delta), & \text{if } i\neq j,\\ 1+q_{ii}(\mu)\delta+o(\delta), & \text{otherwise,}\end{cases} \tag{1}$$

provided $\delta>0$.

In order to introduce the delay-dependent control, we first introduce some notations. Given any metric space E, denote by $\mathscr{C}([0,T];E)$ the collection of continuous functions $x:[0,T]\to E$, and $\mathscr{D}([0,T];E)$ the collection of right-continuous functions with left limits $\lambda:[0,T]\to E$. For $r_0\in(0,T)$ and $s\in[0,T]$, define a shift operator $\theta_{s,r_0}:\mathscr{D}([0,T];\mathscr{S})\to\mathscr{D}([0,T];\mathscr{S})$ by

$$(\theta_{s,r_0}\lambda)(t)=\lambda((t-r_0)\vee s),\quad t\in[0,T]. \tag{2}$$

Moreover, $\theta^k_{s,r_0}\lambda(t):=\lambda((t-kr_0)\vee s)$ for $\lambda\in\mathscr{D}([0,T];\mathscr{S})$, $k\in\mathbb{Z}_+$. For more properties and discussions on shift operator, please refer to Billingsley (2013, Appendix M22, p. 258). Next, we introduce the concept of delay-dependent control.

Definition 2.1: Fix an arbitrary $m\in\mathbb{Z}_+$ and $r_0>0$. Given any $s\in[0,T)$ and $i\in\mathscr{S}$, a randomised delay-dependent control is a term $\alpha=(\Lambda_t,\mu_t,s,i)$ such that

(i) (Λ_t) is an $\mathscr{F}_t$-adapted jump process satisfying (1) with initial value $\Lambda_s=i$.
(ii) There exists a measurable map $h:[0,T]\times\mathscr{S}^{m+1}\to\mathscr{P}(U)$ such that

$$\mu_t=h(t,\theta^0_{s,r_0}\Lambda(t),\ldots,\theta^m_{s,r_0}\Lambda(t))\\ \text{for almost all } t\in[s,T]. \tag{3}$$

The parameter $r_0>0$ is used to characterise the time interval of delay of the controlled processes, and $m\in\mathbb{Z}_+$ for the number of delay. The collection of all delay-dependent control α with initial condition (s,i) is denoted by $\Pi_{s,i}$. When the starting time of the optimal control problem is s, as we have no further information on the controlled system before the initial time s, we use the state of the process (Λ_t) at time s to represent its states before time s, which is reflected by the definition of μ_t through Equation (3). Such treatment has been used in the study of optimal control problem over history-dependent policies; see, for instance, Guo et al. (2015, 2012).

Let $f:[0,T]\times\mathscr{S}\times\mathscr{P}(U)\to[0,\infty)$, $g:\mathscr{S}\to[0,\infty)$ be two lower semi-continuous functions. The expected cost for the delay-dependent control $\alpha\in\Pi_{s,i}$ is defined by

$$J(s,i,\alpha)=\mathbb{E}\left[\int_s^T f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+g(\Lambda_T)\right], \tag{4}$$

and the value function is defined by

$$V(s,i)=\inf_{\alpha\in\Pi_{s,i}}J(s,i,\alpha). \tag{5}$$

It immediately implies that the value function V satisfies $V(T,i)=g(i)$, $\forall\, i\in\mathscr{S}$. A delay-dependent control $\alpha^*\in\Pi_{s,i}$ is said to be *optimal*, if $V(s,i)=J(s,i,\alpha^*)$.

The set of delay-dependent controls introduced in Definition 2.1 contains many interesting control policies. Next, we present some examples below.

Example 2.2: We consider the optimal control problem with initial time $s=0$.

(1) $\mu_t=h(\Lambda_t)$ for some $h:\mathscr{S}\to\mathscr{P}(U)$. In this situation, α is corresponding to the stationary randomised Markov policy studied by many works; see, e.g. Guo and Hernández-Lerma (2009).
(2) $\mu_t=h(\Lambda_{(t-r_0)\vee 0})$ for some $h:\mathscr{S}\to\mathscr{P}(U)$. Now the control policies are purely determined by the jump process with a positive delay. This kind of controls is very natural to be used in the realistic application.
(3) $\mu_t=h(t,\Lambda_{(t-r_0)\vee 0},\Lambda_{(t-2r_0)\vee 0})$ for some $h:[0,T]\times\mathscr{S}\times\mathscr{S}\to\mathscr{P}(U)$.
(4) $\mu_t=h(t,\Lambda_{(t-r_0)\vee 0})$ for some $h(t,i)=\delta_{u_t(i)}$ for each $i\in\mathscr{S}$, where $t\mapsto u_t(i)$ is a curve in U and δ_x denote the Dirac measure in U.

In this paper we impose the following assumptions on the primitive Q-matrix of the continuous-time Markov decision process (Λ_t).

Assumption: (H1) *$\mu\mapsto q_{ij}(\mu)$ is continuous for every $i,j\in\mathscr{S}$, and*

$$M:=\sup_{i\in\mathscr{S}}\sup_{\mu\in\mathscr{P}(U)}q_i(\mu)<\infty.$$

(H2) *There exists a compact function $\Phi:\mathscr{S}\to[1,\infty)$, a compact set $B_0\subset\mathscr{S}$, constants $\lambda_0>0$ and $\kappa_0\geqslant 0$ such that*

$$Q_\mu\Phi(i):=\sum_{j\neq i}q_{ij}(\mu)\big(\Phi(j)-\Phi(i)\big)\\ \leqslant\lambda_0\Phi(i)+\kappa_0\mathbf{1}_{B_0}(i).$$

(H3) *There exists $K\in\mathbb{N}$ such that for every $i\in\mathscr{S}$ and $\mu\in\mathscr{P}(U)$, $q_{ij}(\mu)=0$, if $|j-i|>K$.*

Here if for every $c\in\mathbb{R}$, the set $\{i\in\mathscr{S};\Phi(i)\leqslant c\}$ is a compact set, then Φ is called a compact function. Condition (H3) is a technical condition, which is used when we consider to use the dominated convergence theorem in the argument of our main theorem.

In contrast to the well-studied continuous-time Markov decision process, the controlled system (Λ_t) studied in this work is no longer a Markov chain, and the delay-dependent control policy makes it more difficult to describe the evolution of (Λ_t). Following Shao (2020), we shall develop the classical compactness method to deal with the control problem with delay-dependent controls. The compactification method is usually used to cope with the optimal control problem for stochastic differential equations (cf. Haussmann & Suo, 1995a, 1995b; Kushner, 1975 and references therein). We extend this method to deal with stochastic processes in discrete space.

Let

$$\mathscr{U} = \{\mu : [0, T] \to \mathscr{P}(U) \text{ is measurable}\}. \tag{6}$$

$\mathscr{U}$ can be viewed as a subspace of $\mathscr{P}([0, T] \times U)$ through the map

$$(\mu_t)_{t\in[0,T]} \mapsto \bar{\mu},$$

where $\bar{\mu}$ is determined by

$$\bar{\mu}(A \times B) = \frac{1}{T}\int_A \mu_t(B)\,\mathrm{d}t.$$

Endow $\mathscr{U}$ with the induced weak convergence topology from $\mathscr{P}([0, T] \times U)$. This topology is equivalent to the topology induced by the following Wasserstein distance on $\mathscr{P}([0, T] \times U)$:

$$W_1(\bar{\mu}, \bar{\nu}) = \inf_{\Gamma\in\mathscr{C}(\bar{\mu},\bar{\nu})} \int_{([0,T]\times U)^2} \big(|s-t| + |x-y|\big)\,\mathrm{d}\Gamma((s,x),(t,y)),$$

where $\mathscr{C}(\bar{\mu}, \bar{\nu})$ stands for the collection of couplings of $\bar{\mu}$ and $\bar{\nu}$ over $([0, T] \times U)^2$. The canonical path space for our problem is defined as

$$\hat{\Omega} = \mathscr{D}([0, T]; \mathscr{S}) \times \mathscr{U}$$

endowed with the product topology, which is a metrizable and separable space (cf. Haussmann & Suo, 1995a). Denote by $\tilde{\mathscr{D}}^1$ (resp. $\tilde{\mathscr{D}}^2$) the Borel σ-algebra of $\mathscr{D}([0, T]; \mathscr{S})$ (resp. $\mathscr{U}$), and $\tilde{\mathscr{D}}^1_t$ (resp. $\tilde{\mathscr{D}}^2_t$) the σ-algebra up to time t. Define the σ-algebra of $\hat{\Omega}$ as

$$\hat{\mathscr{F}} := \tilde{\mathscr{D}}^1 \times \tilde{\mathscr{D}}^2, \quad \text{and} \quad \hat{\mathscr{F}}_t = \tilde{\mathscr{D}}^1_t \times \tilde{\mathscr{D}}^2_t.$$

For each delay control $\alpha = (\Lambda_t, \mu_t, s, i) \in \Pi_{s,i}$, we define a measurable map $\Phi_\alpha : \Omega \to \hat{\Omega}$ as

$$\Phi_\alpha(\omega) = (\Lambda_t(\omega), \mu_t(\omega))_{t\in[0,T]},$$
$$\Lambda_r(\omega) \equiv i,\ \mu_r(\omega) \equiv \mu_s,\ 0 \leqslant r \leqslant s.$$

Then, there exists a corresponding probability on $(\hat{\Omega}, \hat{\mathscr{F}})$ defined by $R = \mathbb{P} \circ \Phi_\alpha^{-1}$. We denote by $\hat{\Pi}_{s,i}$ the space of probabilities induced by the delay-dependent control set $\Pi_{s,i}$ with initial condition (s, i). By the definition of value function, we have

$$V(s,i) = \inf_{\alpha\in\Pi_{s,i}} J(s,i,\alpha) = \inf_{R\in\hat{\Pi}_{s,i}} \mathbb{E}_R\left[\int_s^T f(t,\Lambda_t,\mu_t)\,\mathrm{d}t + g(\Lambda_T)\right].$$

The topology and properties of the canonical path space have been well studied, see, for instance Haussmann and Suo (1995a), Meyer and Zheng (1984), and Stroock and Varadhan (1979) and the references therein.

3. Existence of optimal delay-dependent controls

By developing the compactification method presented, for instance, in Haussmann and Suo (1995a), Kushner (1975), and Shao (2020) investigated the optimal control problem for the regime-switching processes. There, the control on the transition rate matrix of the jumping process (Λ_t) has been studied. In this paper we shall apply the result (Shao, 2020, Theorem 2.3) to the current situation to obtain the existence of optimal delay-dependent controls of our continuous-time Markov decision processes under the mild conditions (H1)–(H3).

Theorem 3.1: *Assume* (H1)–(H3) *hold. Then for every* $s \in [0, T)$, $i \in \mathscr{S}$, *there exists an optimal delay-dependent control* $\alpha^* \in \Pi_{s,i}$.

Proof: This theorem is proved by using the idea of Shao (2020, Theorem 2.3). The proof is a little long. In order to save space, here we only sketch the idea and point out the different points compared with that of Shao (2020, Theorem 2.3).

We only need to consider the nontrivial case $V(s, i) < \infty$. For simplicity of notation, we consider the case $s = 0$, and separate the proof into three steps.

Step 1. According to the definition of $V(0, i)$, there exists a sequence of delay-dependent controls $\alpha_n = (\Lambda^{(n)}_t, \mu^{(n)}_t, 0, i) \in \Pi_{0,i}$ such that

$$\lim_{n\to\infty} J(0, i, \alpha_n) = V(0, i). \tag{7}$$

Denote by R_n the probability measures on $(\hat{\Omega}, \hat{\mathscr{F}})$ corresponding to α_n. Let $\mathscr{L}^n_\mu$ (resp. $\mathscr{L}^n_\Lambda$) be the marginal distribution of R_n with respect to $(\mu^{(n)}_t)_{t\in[0,T]}$ (resp. $(\Lambda^{(n)}_t)_{t\in[0,T]}$) in $\mathscr{U}$ (resp. $\mathscr{D}([0, T]; \mathscr{S})$). Since $\mathscr{P}([0, T] \times U)$ is compact and further $\mathscr{U}$ is compact as a closed subset, we have $(\mathscr{L}^n_\mu)_{n\geqslant1}$ is tight.

We proceed to prove that $(\mathscr{L}^n_\Lambda)_{n\geqslant1}$ is tight. For each $n \geqslant 1$, due to (H1), the process $\Lambda^{(n)}_t$ is non-explosive, and the number of jumps within a finite time is finite. Therefore, for the function Φ given in (H2), we have $\mathbb{E}\Phi(\Lambda^{(n)}_t) < \infty$. Using (H2) and Itô-Dynkin's formula (cf. Guo et al., 2015, Theorem 3.1), we have

$$\begin{aligned}\mathbb{E}\Phi(\Lambda^{(n)}_t) &= \Phi(i) + \mathbb{E}\int_0^t Q_{\mu_s}\Phi(\Lambda^{(n)}_s)\,\mathrm{d}s \\ &\leqslant \Phi(i) + \mathbb{E}\int_0^t \big(\lambda_0\Phi(\Lambda^{(n)}_s) + \kappa_0\big)\,\mathrm{d}s \\ &\leqslant (\Phi(i) + \kappa_0 T) + \int_0^t \lambda_0\mathbb{E}\Phi(\Lambda^{(n)}_s)\,\mathrm{d}s.\end{aligned}$$

The above inequality is obtained by interchanging the order of expectation and integration based on Fubini's theorem. Furthermore, applying Gronwall's inequality, it holds that

$$\mathbb{E}\Phi(\Lambda^{(n)}_t) \leqslant \big(\Phi(i) + \kappa_0 T\big)\mathrm{e}^{\lambda_0 t}, \quad t \in [0, T],\ n \geqslant 1. \tag{8}$$

For any $\varepsilon > 0$, we can find $N_\varepsilon > 0$ such that

$$\sup_n \mathbb{P}(\Lambda^{(n)}_t \in K^c_\varepsilon) \leqslant \sup_n \frac{\mathbb{E}\Phi(\Lambda^{(n)}_t)}{N_\varepsilon} \leqslant \frac{(\Phi(i) + \kappa_0 T)\mathrm{e}^{\lambda_0 T}}{N_\varepsilon} < \varepsilon, \tag{9}$$

where $K_\varepsilon = \{j \in \mathscr{S}; \Phi(j) \leqslant N_\varepsilon\}$. Since Φ is a compact function, K_ε is a compact set. Moreover, for every $0 \leqslant u \leqslant \delta$, due to (H1),

$$\begin{aligned}\mathbb{E}\big[\mathbf{1}_{\Lambda^{(n)}_{t+u}\neq\Lambda^{(n)}_t}\big] &\leqslant 1-\mathbb{P}(\Lambda^{(n)}_s=\Lambda^{(n)}_t,\forall\, s\in[t,t+u])\\ &\leqslant 1-\mathrm{e}^{-Mu}\leqslant 1-\mathrm{e}^{-M\delta}=:\gamma_n(\delta). \qquad (10)\end{aligned}$$

To apply (Ethier & Kurtz, 1986, Theorem 8.6, p.138), by taking $q(i,j)=\mathbf{1}_{i\neq j}$, $\beta=1$, $\gamma_n(\delta)$ given in (10) and invoking (9), we obtain the tightness of $(\mathscr{L}^n_\Lambda)_{n\geqslant 1}$.

Step 2. Since the marginal distributions $(\mathscr{L}^n_\Lambda)_{n\geqslant 1}$ and $(\mathscr{L}^n_\mu)_{n\geqslant 1}$ are both tight, $(R_n)_{n\geqslant 1}$ is tight as well. Hence, there exists a subsequence n_k, $k\geqslant 1$, such that R_{n_k} weakly converges to some probability measure R_0 on $(\hat{\Omega},\hat{\mathscr{F}})$ as $k\to\infty$. By virtue of Skorokhod's representation theorem (cf. e.g. Ethier & Kurtz, 1986, Chapter 3), there exists a probability space $(\Omega',\mathscr{F}',\mathbb{P}')$ on which is defined a sequence of $\hat{\Omega}$-valued random variables $Y_{n_k}=(\Lambda^{(n_k)}_t,\mu^{(n_k)}_t)_{t\in[0,T]}$ with distribution R_{n_k}, $k\geqslant 1$, and $Y_0=(\Lambda^{(0)}_t,\mu^{(0)}_t)_{t\in[0,T]}$ with distribution R_0 such that

$$\lim_{k\to\infty} Y_{n_k}=Y_0,\quad \mathbb{P}'-\text{a.s.} \qquad (11)$$

Analogous to the Step 2 in the argument of Shao (2020, Theorem 2.3), we can show that $\alpha^*:=(\Lambda^{(0)}_t,\mu^{(0)}_t,0,i)$ is a delay-dependent control in $\Pi_{0,i}$. During this procedure, we need to replace the sigma fields $\mathscr{F}^{X,\Lambda}_{-n,t}$ by the following

$$\mathscr{F}^{\Lambda}_{-n,t}:=\overline{\sigma\{(\Lambda^{(k)}_t,\dots,\Lambda^{(k)}_{t-mr_0});k\geqslant n\}}.$$

Step 3. Invoking (7) and the lower semi-continuity of f and g, we obtain

$$\begin{aligned}V(0,i)&=\lim_{k\to\infty}\mathbb{E}\left[\int_0^T f(t,\Lambda^{(n_k)}_t,\mu^{(n_k)}_t)\,\mathrm{d}t+g(\Lambda^{(n_k)}_T)\right]\\ &\geqslant \mathbb{E}\left[\int_0^T f(t,\Lambda^{(0)}_t,\mu^{(0)}_t)\,\mathrm{d}t+g(\Lambda^{(0)}_T)\right]\\ &\geqslant V(0,i).\end{aligned}$$

By taking $\alpha^*=(\Lambda^{(0)}_t,\mu^{(0)}_t,0,i)\in\Pi_{0,i}$, the previous inequalities imply that α^* is an optimal delay-dependent control of the continuous-time Markov jump process. The proof of this theorem is complete. ■

4. Dynamic programming principle and viscosity solution

In the rest of the paper, we introduce the dynamic programming principle for the controlled processes with delay-dependent control and the differential equation satisfied by the value function. To do so, we introduce some notations. Assume that τ is an $\hat{\mathscr{F}}_t$-stopping time satisfying $0\leqslant\tau\leqslant T$, $\hat{\mathscr{F}}_\tau$ is denoted by the collection of sets A such that $A\cap\{\tau\leqslant t\}\in\hat{\mathscr{F}}_t$, $\forall\, t\in[0,T]$.

Theorem 4.1: *Assume* (H1)–(H3) *hold. For each $\hat{\mathscr{F}}_t$-stopping time τ satisfying $s\leqslant\tau\leqslant T$, then*

$$V(s,i)=\inf\left\{\mathbb{E}_R\left[\int_s^\tau f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+V(\tau,\Lambda_\tau)\right];R\in\hat{\Pi}_{s,i}\right\}.$$

Proof: Define a subset of $\hat{\Pi}_{s,i}$ as

$$\begin{aligned}\hat{\Pi}^0_{s,i}=\Bigg\{R\in\hat{\Pi}_{s,i}:V(s,i)&\\ =\mathbb{E}_R\left[\int_s^T f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+g(\Lambda_T)\right]\Bigg\}.&\end{aligned}$$

By Theorem 3.1, $\hat{\Pi}^0_{s,i}\neq\emptyset$ for any $s\in[0,T]$ and $i\in\mathscr{S}$. According to measurable choices theorem presented by Stroock and Varadhan (1979), there exists a Borel-measurable map $H:[0,t]\times\mathscr{S}\to\mathscr{P}(U)$, which is called measurable selector, satisfying for each $(s,i)\in[0,t]\times\mathscr{S}$, $H(s,i)\in\hat{\Pi}^0_{s,i}$. Refer to Haussmann and Suo (1995a, Lemma 3.9) for more details of the existence of the measurable selector. Hence, for any $\hat{\omega}\in\hat{\Omega}$, $H(\tau(\hat{\omega}),\Lambda_{\tau(\hat{\omega})})$ is a probability measure on $(\hat{\Omega},\hat{\mathscr{F}})$ and satisfies

$$\begin{aligned}&V(\tau(\hat{\omega}),\Lambda_{\tau(\hat{\omega})})\\ &=\mathbb{E}_{H(\tau(\hat{\omega}),\Lambda_{\tau(\hat{\omega})})}\left[\int_{\tau(\hat{\omega})}^T f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+g(\Lambda_T)\right]. \qquad (12)\end{aligned}$$

Note that the topology on $\hat{\Omega}$ is separable, then $\hat{\mathscr{F}}_t$ is countably generated, and then for every probability measure $\mathbb{P}$ on $(\hat{\Omega},\hat{\mathscr{F}})$, the regular conditional probability distribution of $\mathbb{P}$ for given $\hat{\mathscr{F}}_\tau$ exists, cf. Haussmann and Suo (1995a, 1995b). According to Haussmann and Suo (1995b, Lemma 3.3), for each $R\in\hat{\Pi}_{s,i}$, there exists a unique probability measure, denoted by R^H, such that $R^H(A)=R(A),\forall\, A\in\mathscr{F}_\tau$ and the regular conditional probability distribution of R^H for given $\mathscr{F}_\tau$ is $H(\tau(\cdot),\Lambda_{\tau(\cdot)})$. Moreover, by Haussmann and Suo (1995b, Proposition 3.8), it holds that $R^H\in\hat{\Pi}_{s,i}$. Hence, we have

$$V(\tau(\hat{\omega}),\Lambda_{\tau(\hat{\omega})})=\mathbb{E}_{R^H}\left[\int_{\tau(\hat{\omega})}^T f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+g(\Lambda_T)\Big|\mathscr{F}_\tau\right].$$

Due to the definition of value function $V(s,i)$, we have

$$\begin{aligned}V(s,i)&\leqslant\mathbb{E}_{R^H}\Bigg[\int_s^\tau f(t,\Lambda_t,\mu_t)\,\mathrm{d}t\\ &\qquad+\int_\tau^T f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+g(\Lambda_T)\Bigg]\\ &=\mathbb{E}_{R^H}\Bigg[\int_s^\tau f(t,\Lambda_t,\mu_t)\,\mathrm{d}t\\ &\qquad+\mathbb{E}_{R^H}\left[\int_\tau^T f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+g(\Lambda_T)\Big|\mathscr{F}_\tau\right]\Bigg]\\ &=\mathbb{E}_R\left[\int_s^\tau f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+V(\tau,\Lambda_\tau)\right],\end{aligned}$$

where the last equation is based on the relationship between R and R^H. The arbitrariness of $R\in\hat{\Pi}_{s,i}$ implies that

$$\begin{aligned}&V(s,i)\\ &\leqslant\inf\left\{\mathbb{E}_R\left[\int_s^\tau f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+V(\tau,\Lambda_\tau)\right];R\in\hat{\Pi}_{s,i}\right\}.\end{aligned}$$

Conversely, by Theorem 3.1, there exists an optimal delay-dependent control $\alpha^*\in\Pi_{s,i}$ and then denote by $R^*\in\hat{\Pi}_{s,i}$ the corresponding probability measure on $(\hat{\Omega},\hat{\mathscr{F}})$. Then we have

$$\begin{aligned}&V(s,i)\\&=\mathbb{E}_{R^*}\Big[\int_s^\tau f(t,\Lambda_t,\mu_t)\,\mathrm{d}t\\&\quad+\int_\tau^T f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+g(\Lambda_T)\Big]\\&\geqslant\mathbb{E}_{R^*}\Big[\int_s^\tau f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+V(\tau,\Lambda_\tau)\Big]\\&\geqslant\inf\Big\{\mathbb{E}_R\Big[\int_s^\tau f(t,\Lambda_t,\mu_t)\,\mathrm{d}t+V(\tau,\Lambda_\tau)\Big];R\in\hat{\Pi}_{s,i}\Big\}.\end{aligned}$$

The dynamic programming principle is thus proved. ■

The next result is about the continuity of value function. Since $\mathscr{S}$ is a countable state space equipped with discrete topology, we only need to consider the continuity of $V(s,i)$ in the time variable s.

Proposition 4.2: *Assume* (H1)–(H3) *hold. Suppose that f, g are bounded and f satisfies the following condition,*

$$|f(t,i,\mu)-f(s,i,\mu)|\leqslant C_0|t-s|,\quad 0\leqslant s,t\leqslant T,\tag{13}$$

uniformly for $i\in\mathscr{S}$ and $\mu\in\mathscr{P}(U)$. Then, the value function $V(s,i)$ is Lipschitz continuous with respect to the time variable s. In fact, there exists a constant $C>0$ such that for any $i\in\mathscr{S}$

$$\big|V(s,i)-V(s',i)\big|\leqslant C|s-s'|,\quad 0\leqslant s,s'\leqslant T.$$

Proof: For convenience, denote by C_1 and C_2 the constants such that

$$\sup_{(t,i,\mu)\in[0,T]\times\mathscr{S}\times\mathscr{P}(U)}|f(t,i,\mu)|\leqslant C_1\quad\text{and}\quad\sup_{i\in\mathscr{S}}|g(i)|\leqslant C_2.$$

Fix any $i\in\mathscr{S}$ and assume $0\leqslant s\leqslant s'\leqslant T$. According to Theorem 3.1, there exists an optimal delay-dependent control $\alpha^*=(\Lambda_t,\mu_t,s,i)\in\Pi_{s,i}$ such that $V(s,i)=J(s,i,\alpha^*)$. By time shift, we can define a couple of processes with initial point (s',i) as following

$$\Lambda'_t=\Lambda_{t-\Delta s},\quad \mu'_t=\mu_{t-\Delta s},\quad\forall\, t\in[s',T],$$

where $\Delta s:=s'-s$. It is easy to verify that (1) and (3) hold for (Λ'_t,μ'_t), which means that $\alpha':=(\Lambda'_t,\mu'_t,s',i)$ is a delay-dependent control in $\Pi_{s',i}$. Using (H1) and (1), we have

$$\mathbb{E}\Big[\mathbf{1}_{\Lambda'_t\neq\Lambda_t}\Big]=\mathbb{P}\left(\Lambda_{t-\Delta s}\neq\Lambda_t\right)\leqslant M\Delta s+o(\Delta s).$$

By the definition of the value function, we have

$$\begin{aligned}&|V(s',i)-V(s,i)|\\&\leqslant\mathbb{E}\Big[\int_{s'}^T\big|f(t,\Lambda'_t,\mu'_t)-f(t,\Lambda_t,\mu_t)\big|\,\mathrm{d}t\Big]\\&\quad+\mathbb{E}\big[|g(\Lambda'_T)-g(\Lambda_T)|\big]+\mathbb{E}\Big[\int_s^{s'}|f(t,\Lambda_t,\mu_t)|\,\mathrm{d}t\Big].\end{aligned}\tag{14}$$

According to the boundedness of f and g, we obtain

$$\begin{aligned}&\mathbb{E}\Big[\int_s^{s'}|f(t,\Lambda_t,\mu_t)|\,\mathrm{d}t\Big]\leqslant C_1\Delta s,\quad\text{and}\\&\mathbb{E}\big[|g(\Lambda'_T)-g(\Lambda_T)|\big]\leqslant 2C_2\mathbb{E}\Big[\mathbf{1}_{\Lambda'_T\neq\Lambda_T}\Big]\\&\leqslant 2MC_2\Delta s+o(\Delta s).\end{aligned}$$

To estimate the first term of (14), we combine the boundedness and (13),

$$\begin{aligned}&\mathbb{E}\Big[\int_{s'}^T|f(t,\Lambda'_t,\mu'_t)-f(t,\Lambda_t,\mu_t)|\,\mathrm{d}t\Big]\\&=\mathbb{E}\Big[\int_s^{T-\Delta s}|f(t+\Delta s,\Lambda_t,\mu_t)\,\mathrm{d}t-f(t,\Lambda_t,\mu_t)|\,\mathrm{d}t\Big]\\&\quad+\mathbb{E}\Big[\int_s^{s'}|f(t,\Lambda_t,\mu_t)|\,\mathrm{d}t\Big]+\mathbb{E}\Big[\int_{T-\Delta s}^T|f(t,\Lambda_t,\mu_t)|\,\mathrm{d}t\Big]\\&\leqslant TC_0\Delta s+2C_1\Delta s.\end{aligned}$$

Hence,

$$|V(s,i)-V(s',i)|\leqslant(3C_1+2MC_2+TC_0)\Delta s+o(\Delta s).$$

By the symmetric position of s and s', we have $|V(s,i)-V(s',i)|\leqslant C|s-s'|$. ■

According to Proposition 4.2 and Rademacher's theorem, we know that for each $i\in\mathscr{S}$, $t\to V(t,i)$ is almost everywhere differentiable in $[0,T]$ with respect to Lebesgue measure. In some practical applications, the property of almost everywhere differentiable is not enough, especially when $\mathscr{S}$ is a general state space rather than a countable space. But, it is not easy to justify whether $V(t,i)$ is differentiable every where in $[0,T]$. In such situation, it is useful to introduce the concept of viscosity solution to further characterise $V(t,i)$. Consider the following equation

$$\begin{aligned}&-\frac{\partial v}{\partial t}-\inf_{\mu\in\mathscr{P}(U)}\Big\{\sum_{j\neq i}q_{ij}(\mu)\big(v(t,j)-v(t,i)\big)+f(t,i,\mu)\Big\}=0,\\&v(T,i)=g(i).\end{aligned}\tag{15}$$

Definition 4.3: Let $v:[0,T)\times\mathscr{S}\to\mathbb{R}$ be a continuous function.

(i) v is called a viscosity supersolution of (15) if $v(T,i)=g(i)$,

$$\begin{aligned}&-\frac{\partial\phi}{\partial t}(t_0,i_0)-\inf_{\mu\in\mathscr{P}(U)}\Big\{\sum_{j\neq i_0}q_{i_0j}(\mu)\big(\phi(t_0,j)-\phi(t_0,i_0)\big)\\&+f(t_0,i_0,\mu)\Big\}\geqslant 0\end{aligned}$$

for all $(t_0,i_0)\in[0,T)\times\mathscr{S}$ and for all $\phi\in C^1([0,T)\times\mathscr{S})$ such that (t_0,i_0) is a minimum point of $v-\phi$.

(ii) v is called a viscosity subsolution of (15) if $v(T,i)=g(i)$,

$$-\frac{\partial\phi}{\partial t}(t_0,i_0)-\inf_{\mu\in\mathscr{P}(U)}\Bigg\{\sum_{j\neq i_0}q_{i_0j}(\mu)\big(\phi(t_0,j)-\phi(t_0,i_0)\big)$$
$$+f(t_0,i_0,\mu)\Big\}\leqslant 0$$

for all $(t_0,i_0)\in[0,T)\times\mathscr{S}$ and for all $\phi\in C^1([0,T)\times\mathscr{S})$ such that (t_0,i_0) is a maximum point of $v-\phi$.

(iii) v is called a viscosity solution to (15) if it is both a viscosity subsolution and a viscosity supersolution of (15).

The next result says that the value function is a solution to the HJB Equation (15) in the viscosity sense.

Theorem 4.4: *Under the conditions of Proposition* 4.2, *the value function* $V(t,i)$ *is a viscosity solution to* (15).

Proof: We first consider the viscosity subsolution property. Let $(t_0,i_0)\in[0,T)\times\mathscr{S}$ and $\phi\in C^1([0,T)\times\mathscr{S})$ be a test function such that

$$0=(V-\phi)(t_0,i_0)=\max\{(V-\phi)(t,i);(t,i)\in[0,T)\times\mathscr{S}\}. \tag{16}$$

Take an arbitrary point $\tilde\mu\in\mathscr{P}(U)$ and let

$$\mu_t=\tilde\mu,\quad \forall\, t\in[s,T],$$

which is a constant control policy and obviously satisfies the conditions of Definition 2.1. According to the dynamic programming principle (Theorem 4.1), we have

$$V(t_0,i_0)\leqslant\mathbb{E}\left[\int_{t_0}^{t}f(r,\Lambda_r,\tilde\mu)\,\mathrm{d}r+V(t,\Lambda_t)\right].$$

Due to (16), it holds $V\leqslant\phi$, and hence

$$\phi(t_0,i_0)\leqslant\mathbb{E}\left[\int_{t_0}^{t}f(r,\Lambda_r,\tilde\mu)\,\mathrm{d}r+\phi(t,\Lambda_t)\right]. \tag{17}$$

Applying Itô-Dynkin's formula to the function ϕ (cf. Guo et al., 2015, Theorem 3.1), we get

$$\mathbb{E}\phi(t,\Lambda_t)$$
$$=\phi(t_0,i_0)+\mathbb{E}\left[\int_{t_0}^{t}\left(\frac{\partial\phi}{\partial r}(r,\Lambda_r)+Q(\tilde\mu)\phi(r,\Lambda_r)\right)\mathrm{d}r\right]. \tag{18}$$

Inserting (18) into (17) leads to

$$-\mathbb{E}\left[\int_{t_0}^{t}\left(\frac{\partial\phi}{\partial r}(r,\Lambda_r)+Q(\tilde\mu)\phi(r,\Lambda_r)+f(r,\Lambda_r,\tilde\mu)\right)\mathrm{d}r\right]\leqslant 0. \tag{19}$$

Dividing both sides of (19) by $t-t_0$ and letting $t\downarrow t_0$, we get from the almost sure right-continuity of the trajectories of (Λ_t) that

$$-\frac{\partial\phi}{\partial t}(t_0,i_0)-\sum_{j\neq i_0}q_{i_0j}(\tilde\mu)(\phi(t_0,j)$$
$$-\phi(t_0,i_0))+f(t_0,i_0,\tilde\mu)\leqslant 0. \tag{20}$$

Then, by the arbitrariness of $\tilde\mu\in\mathscr{P}(U)$, $V(t,i)$ is a viscosity subsolution of (15).

Next, we proceed to the viscosity supersolution property. Let $(t_0,i_0)\in[0,T)\times\mathscr{S}$ and $\phi\in C^1([0,T)\times\mathscr{S})$ be a test function such that

$$0=(V-\phi)(t_0,i_0)=\min\{(V-\phi)(t,i);(t,i)\in[0,T)\times\mathscr{S}\}. \tag{21}$$

The desired result will be shown by contradiction. Assume

$$-\frac{\partial\phi}{\partial t}(t_0,i_0)-\inf_{\mu\in\mathscr{P}(U)}\big\{Q(\mu)\phi(t_0,i_0)+f(t_0,i_0,\mu)\big\}<0. \tag{22}$$

By (H1), the compactness of $\mathscr{P}(U)$ and the continuity of f, we obtain from (22) that there exist $\varepsilon,\eta>0$ such that for any $0\leqslant t-t_0\leqslant\eta$, it holds

$$-\frac{\partial\phi}{\partial t}(t,i_0)-\inf_{\mu\in\mathscr{P}(U)}\big\{Q(\mu)\phi(t,i_0)+f(t,i_0,\mu)\big\}\leqslant-\varepsilon. \tag{23}$$

Let $(t_k)_{k\geqslant1}$ be a sequence satisfying $t_k>t_0$ for any $k\geqslant1$ and $\lim_{k\to\infty}t_k=t_0$. Using the dynamic programming principle (Theorem 4.1) again, for each $k\geqslant1$, there exists $\alpha^{(k)}=(\Lambda_t^{(k)},\mu_t^{(k)},t_0,i_0)\in\Pi_{t_0,i_0}$ such that

$$V(t_0,i_0)\geqslant\mathbb{E}\left[\int_{t_0}^{\beta_k}f(r,\Lambda_r^{(k)},\mu_r^{(k)})\,\mathrm{d}r+V(\beta_k,\Lambda_{\beta_k}^{(k)})\right]$$
$$-\frac{\varepsilon}{2}(t_k-t_0),$$

where $\beta_k=t_k\wedge\tau_k$, and τ_k is defined by

$$\tau_k=\inf\{t\in[t_0,T];\Lambda_t^{(k)}\neq\Lambda_{t_0}^{(k)}\}\wedge(t_0+\eta). \tag{24}$$

Due to (21), we have $V\geqslant\phi$ and

$$\phi(t_0,i_0)\geqslant\mathbb{E}\left[\int_{t_0}^{\beta_k}f(r,\Lambda_r^{(k)},\mu_r^{(k)})\,\mathrm{d}r+\phi(\beta_k,\Lambda_{\beta_k}^{(k)})\right]$$
$$-\frac{\varepsilon}{2}(t_k-t_0). \tag{25}$$

Using Itô-Dynkin's formula to the function ϕ, we have

$$\mathbb{E}\left[\int_{t_0}^{\beta_k}f(r,\Lambda_r^{(k)},\mu_r^{(k)})+\left(\frac{\partial\phi}{\partial r}+Q(\mu_r^{(k)})\phi\right)(r,\Lambda_r^{(k)})\,\mathrm{d}r\right]$$
$$\leqslant\frac{\varepsilon}{2}(t_k-t_0).$$

Then (23) and the definition of β_k implies that

$$\frac{\mathbb{E}[\beta_k-t_0]}{t_k-t_0}\leqslant\frac{1}{2},\quad k\geqslant1. \tag{26}$$

On the other hand, by (H1), we have

$$\mathbb{P}(\beta_k-t_0\leqslant t_k-t_0)\leqslant\mathbb{P}\big(\sup_{s\in[t_0,t_k]}|\Lambda_s^{(k)}-\Lambda_{t_0}^{(k)}|>0\big)$$
$$\leqslant 1-\mathrm{e}^{-M(t_k-t_0)},$$

Therefore,

$$\lim_{k\to\infty}\mathbb{P}(\beta_k-t_0\geqslant t_k-t_0)=1.$$

Since

$$\mathbb{P}(\beta_k-t_0\geqslant t_k-t_0)\leqslant\frac{\mathbb{E}[\beta_k-t_0]}{t_k-t_0}\leqslant 1,$$

we get finally that

$$\lim_{k\to\infty}\frac{\mathbb{E}[\beta_k-t_0]}{t_k-t_0}=1, \tag{27}$$

which contradicts (26). Consequently, we have

$$-\frac{\partial\phi}{\partial t}(t_0,i_0)-\inf_{\mu\in\mathscr{P}(U)}\Big\{\sum_{j\neq i_0}\big(\phi(t_0,j)-\phi(t_0,i_0)\big)+f(t_0,i_0,\mu)\Big\}\geqslant 0. \tag{28}$$

This means that $V(t,i)$ is a viscosity supersolution of (15). We conclude the proof of this theorem by the definition of viscosity solution to (15). ■

In the end let us discuss the uniqueness of the viscosity solution to (15). For this purpose it is sufficient to establish the following comparison principle for (15). We shall develop the method used to establish the comparison principle for HJB equations associated with diffusion processes to the equations associated with purely jumping processes.

Theorem 4.5: *Assume the conditions of Proposition 4.2 hold. Let V_1 (resp. V_2) be a bounded viscosity supersolution (resp. viscosity subsolution) of (15) in $[0,T)\times\mathscr{S}$. Then*

$$\sup_{[0,T]\times\mathscr{S}}[V_2-V_1]=\sup_{\{T\}\times\mathscr{S}}[V_2-V_1]=0.$$

Proof: Obviously, we just need to show that

$$\sup_{[0,T]\times\mathscr{S}}[V_2-V_1]\leqslant\sup_{\{T\}\times\mathscr{S}}[V_2-V_1]=0. \tag{29}$$

By the condition boundedness of V_1 and V_2, there exists a constant $K_0>0$ such that

$$K_0\geqslant\sup_{t\in[0,T]}\sup_{j\in\mathscr{S}}\big\{|V_1(t,j)|\vee|V_2(t,j)|\big\}. \tag{30}$$

There exists a sequence of $C^2(\mathbb{R})$ functions $\lambda_n(x)$ such that $\lambda_n(x)=0$ for $x\leqslant 0$, $0<\lambda_n'(x)<1$, $\lambda_n(x)\uparrow\max\{x,0\}$ as $n\to\infty$. Let

$$\eta_n(s,t)=t+\lambda_n(s-t),\quad s,t\in[0,T].$$

Then $\eta_n(s,t)\uparrow\max\{s,t\}$ as $n\to\infty$. Define a function on $[0,T]\times[0,T]$ as

$$\Psi_{i_0}^n(t,s)=V_2(t,i_0)-V_1(s,i_0)-\frac{1}{2\delta}(t-s)^2+\frac{\beta}{\delta}(\eta_n(s,t)-T),$$

where $\delta,\beta>0$ are two parameters. Again, the continuity of V_1 and V_2 implies that $\Psi_{i_0}^n$ achieves the maximum on $[0,T]\times[0,T]$. Denote by $(\bar t,\bar s)\in[0,T]\times[0,T]$ an arbitrary one of the maximum points, and note that $(\bar t,\bar s)$ may depend on the parameters δ, β and n.

We first give an estimate of the distance between $\bar s$ and $\bar t$. For any $\rho\geqslant 0$, let

$$D_\rho=\big\{(t,s)\in[0,T]\times[0,T]:|t-s|^2\leqslant\rho\big\},$$
$$m_{i_0}^{(1)}(\rho)=2\sup\big\{|V_1(t,i_0)-V_1(s,i_0)|:(t,s)\in D_\rho\big\},$$
$$m_{i_0}^{(2)}(\rho)=2\sup\big\{|V_2(t,i_0)-V_2(s,i_0)|:(t,s)\in D_\rho\big\}.$$

Then $m_{i_0}^{(1)}$ and $m_{i_0}^{(2)}$ are increasing functions satisfying $m_{i_0}^{(1)}(0)=m_{i_0}^{(2)}(0)=0$. Moreover, it follows from the continuity of V_1 and V_2 and the compactness of $[0,T]\times[0,T]$ that $m_{i_0}^{(1)},m_{i_0}^{(2)}$ are continuous. Since $V_1(\cdot,i_0)$ and $V_2(\cdot,i_0)$ are bounded, $m_{i_0}^{(1)}$ and $m_{i_0}^{(2)}$ are bounded as well and bounded by $M_{i_0}:=\sup\{m_{i_0}^{(1)}(\rho)+m_{i_0}^{(2)}(\rho):\rho\geqslant 0\}<\infty$. We obtain from the fact $\Psi_{i_0}^n(\bar t\vee\bar s,\bar t\vee\bar s)\leqslant\Psi_{i_0}^n(\bar t,\bar s)$ that

$$\frac{1}{\delta}(\bar t-\bar s)^2\leqslant 2\big(V_2(\bar t,i_0)-V_2(\bar t\vee\bar s,i_0)+V_1(\bar t\vee\bar s,i_0)-V_1(\bar s,i_0)\big)\leqslant M_{i_0}.$$

Hence,

$$|\bar t-\bar s|\leqslant\sqrt{\delta M_{i_0}},\quad\text{and hence}\quad\bar t-\bar s\to 0,\quad\text{as }\delta\to 0. \tag{31}$$

Next, we shall show by contradiction that $\bar t$ equals to T. Assume that $\bar t\in[0,T)$. Define an auxiliary function on $[0,T]\times\mathscr{S}$ as

$$\psi_{i_0}^{(1)}(s,j)=-\frac{1}{2\delta}(\bar t-s)^2-2K_0\big(1-\mathbf{1}_{i_0}(j)\big)+\frac{\beta}{\delta}(\eta_n(s,\bar t)-T).$$

For each $s\in[0,T]$, since $\Psi_{i_0}^n(\bar t,s)\leqslant\Psi_{i_0}^n(\bar t,\bar s)$, it holds that

$$V_1(\bar s,i_0)+\frac{1}{2\delta}(\bar t-\bar s)^2-\frac{\beta}{\delta}(\eta_n(\bar s,\bar t)-T)\leqslant V_1(s,i_0)+\frac{1}{2\delta}(\bar t-s)^2-\frac{\beta}{\delta}(\eta_n(s,\bar t)-T),$$

and further for each $j\in\mathscr{S}, j\neq i_0$,

$$\begin{aligned}2K_0&\geqslant V_1(s,i_0)-V_1(s,j)\\&\geqslant V_1(\bar s,i_0)-V_1(s,j)+\frac{1}{2\delta}(\bar t-\bar s)^2-\frac{1}{2\delta}(\bar t-s)^2-\frac{\beta}{\delta}(\eta_n(\bar s,\bar t)-\eta_n(s,\bar t)).\end{aligned}$$

Hence, $(\bar s,i_0)$ is the minimum point of the function $(s,j)\mapsto V_1(s,j)-\psi_{i_0}^{(1)}(s,j)$. Since V_1 is the viscosity supersolution of (15), we have

$$-\frac{1}{\delta}(\bar t-\bar s)-\frac{\beta}{\delta}\lambda_n'(\bar s-\bar t)-\inf_{\mu\in\mathscr{P}(U)}\big\{-2K_0q_{i_0}(\mu)+f(\bar s,i_0,\mu)\big\}\geqslant 0. \tag{32}$$

Similarly, consider the test function on $[0, T] \times \mathscr{S}$ as

$$\psi^{(2)}_{i_0}(t,j) = \frac{1}{2\delta}(t-\bar{s})^2 + 2K_0\big(1-\mathbf{1}_{i_0}(j)\big) - \frac{\beta}{\delta}(\eta_n(\bar{s},t)-T).$$

Then, $\Psi^n_{i_0}(t,\bar{s}) \leqslant \Psi^n_{i_0}(\bar{t},\bar{s})$ implies that for each $t \in [0, T]$,

$$\begin{aligned} &V_2(t,i_0) - \frac{1}{2\delta}(t-\bar{s})^2 + \frac{\beta}{\delta}(\eta_n(\bar{s},t)-T) \\ &\leqslant V_2(\bar{t},i_0) - \frac{1}{2\delta}(\bar{t}-\bar{s})^2 + \frac{\beta}{\delta}(\eta_n(\bar{s},\bar{t})-T), \end{aligned}$$

and for each $j \in \mathscr{S}$ with $j \neq i_0$,

$$\begin{aligned} 2K_0 \geqslant{} & V_2(t,j) - V_2(\bar{t},i_0) + \frac{1}{2\delta}(\bar{t}-\bar{s})^2 - \frac{1}{2\delta}(t-\bar{s})^2 \\ &+ \frac{\beta}{\delta}(\eta_n(\bar{s},t)-\eta_n(\bar{s},\bar{t})). \end{aligned}$$

This means that $(\bar{t}, i_0)$ is a maximum point of $(t,j) \mapsto V_2(t,j) - \psi^{(2)}_{i_0}(t,j)$. Since V_2 is the viscosity subsolution of (15), we have

$$\begin{aligned} &\frac{\beta}{\delta}\big(1-\lambda_n'(\bar{s}-\bar{t})\big) - \frac{1}{\delta}(\bar{t}-\bar{s}) \\ &- \inf_{\mu\in\mathscr{P}(U)}\big\{2K_0q_{i_0}(\mu)+f(\bar{t},i_0,\mu)\big\} \leqslant 0. \end{aligned} \tag{33}$$

Combining the inequalities (31)–(33) and (13), we arrive at

$$\begin{aligned} \frac{\beta}{\delta} &\leqslant \inf_{\mu\in\mathscr{P}(U)}\big\{2K_0q_{i_0}(\mu)+f(\bar{t},i_0,\mu)\big\} \\ &\quad - \inf_{\mu\in\mathscr{P}(U)}\big\{-2K_0q_{i_0}(\mu)+f(\bar{s},i_0,\mu)\big\} \\ &= \sup_{\mu\in\mathscr{P}(U)}\big\{2K_0q_{i_0}(\mu)-f(\bar{s},i_0,\mu)\big\} \\ &\quad - \sup_{\mu\in\mathscr{P}(U)}\big\{-2K_0q_{i_0}(\mu)-f(\bar{t},i_0,\mu)\big\} \\ &\leqslant \sup_{\mu\in\mathscr{P}(U)}\big\{4K_0q_{i_0}(\mu)+f(\bar{t},i_0,\mu)-f(\bar{s},i_0,\mu)\big\} \\ &\leqslant 4K_0M + C_0|\bar{t}-\bar{s}|. \end{aligned} \tag{34}$$

Invoking the estimate (31), this yields that

$$\beta \leqslant 4K_0M\delta + C_0\delta^{3/2}\sqrt{M_{i_0}}.$$

Thus, letting $\delta \to 0$, we get that $\beta \leqslant 0$, which contradicts the assumption that $\beta > 0$. Consequently, it must hold

$$\bar{t} = T. \tag{35}$$

By the choice of $(\bar{t},\bar{s})$, it holds that for every $t \in [0, T)$,

$$\begin{aligned} &V_2(t,i_0) - V_1(t,i_0) + \frac{\beta}{\delta}(t-T) \\ &= \Psi^n_{i_0}(t,t) \leqslant \Psi^n_{i_0}(\bar{t},\bar{s}) \\ &= V_2(T,i_0) - V_1(\bar{s},i_0) - \frac{1}{2\delta}(T-\bar{s})^2 \\ &\leqslant V_2(T,i_0) - V_1(\bar{s},i_0). \end{aligned} \tag{36}$$

Thus, letting first $\beta \to 0$ and then $\delta \to 0$, noting $\lim_{\delta\to 0}|\bar{t}-\bar{s}| = 0$ due to (31), we obtain that

$$V_2(t,i_0) - V_1(t,i_0) \leqslant V_2(T,i_0) - V_1(T,i_0).$$

The desired conclusion (29) follows from the arbitrariness of $i_0 \in \mathscr{S}$. ■

The following uniqueness result is an immediate result of Theorems 4.4 and 4.5.

Corollary 4.6: *Under the conditions of Proposition* 4.2*, the value function* $V(t,i)$ *is the unique viscosity solution to the Equation* (15).

Next, noticing that the Equation (15) does not rely on the delay-dependent control policies, we shall take advantage of this property to show the existence of an optimal Markovian control policy over the class of delay-dependent controls.

Theorem 4.7: *Under the conditions of Proposition* 4.2*, for every* $t \in [0,T], i \in \mathscr{S}$*, there exists an optimal control* α^* *for* $V(t,i)$*, which depends only on the current state of the process* (Λ_t)*, i.e. a Markovian control policy.*

Proof: Introduce a sub-class $\Pi^m_{s,i}$ of $\Pi_{s,i}$ by

$$\begin{aligned} \Pi^m_{s,i} = \big\{&\alpha \in (\Lambda_t,\mu_t,s,i) \in \Pi_{s,i};\ \exists\, h : \mathscr{S} \to \mathscr{P}(U) \\ &\text{such that } \mu_t = h(\Lambda_t)\big\}, \end{aligned}$$

which is the class of stationary randomised Markov policy. Let

$$\widetilde{V}(s,i) = \inf_{\alpha\in\Pi^m_{s,i}} J(s,i,\alpha), \tag{37}$$

which is consistent with the value function studied in Guo et al. (2015, p.1069). According to Guo et al. (2015, Theorem 4.1) and Proposition 4.2, the HJB equation corresponding to the value function $\widetilde{V}(s,i)$ is consistent with Equation (15). Additionally, Guo et al. (2015, Theorem 4.1) further demonstrates that for any i, the value function $\widetilde{V}(\cdot,i)$ is an almost everywhere differentiable solution to the HJB Equation (15). By utilising the proof method of Theorem 4.4, it can be proved that $\widetilde{V}$ is also the viscosity solution of the HJB Equation (15). Hence, the uniqueness of viscosity solution given in Theorem 4.6 means that $\widetilde{V}(s,i) = V(s,i)$. Using Guo et al. (2015, Theorem 4.1) again or along the procedure of Theorem 3.1, there exists an $\tilde{\alpha} \in \Pi^m_{s,i}$ such that $\widetilde{V}(s,i) = J(s,i,\tilde{\alpha})$. Therefore,

$$V(s,i) = \widetilde{V}(s,i) = J(s,i,\tilde{\alpha}), \tag{38}$$

which means that $\tilde{\alpha} \in \Pi^m_{s,i} \subset \Pi_{s,i}$ is the desired optimal control policy in $\Pi_{s,i}$ associated with $V(s,i)$. ■

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Supported in part by National Key R&D Program of China (No. 2022YFA1006000), National Natural Science Foundation of China (No. 12271397, 11831014), Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515010222) and Project of Guangdong Provincial Department of Education (No. 2023GXJK671, 2023KTSCX223).

References

Alexandrov, A. D. (1939). Almost everywhere existence of the second differential of a convex function and some properties of convex functions. *Leningrad State University Annals (Math Ser)*, *37*, 3–35.

Ambrosio, L., Gigli, N., & Savaré, G. (2005). *Gradient flows in metric spaces and in the space of probability measures*. Lectures in mathematics ETH Zrich. Birkhüser Verlag.

Baüerle, N., & Rieder, U. (2011). *Markov decision processes with applications to finance*. Heidelberg.

Billingsley, P. (2013). *Convergence of probability measures* (2nd ed.). Wiley.

Chow, P. L., Menaldi, J., & Robin, M. (1985). Additive control of stochastic linear system with finite time horizon. *SIAM Journal on Control and Optimization*, *23*(6), 858–899. https://doi.org/10.1137/0323051

Derman, C., & Strauch, R. E. (1966). A note on memoryless rules for controlling sequential control processes. *The Annals of Mathematical Statistics*, *37*(1), 276–278. https://doi.org/10.1214/aoms/1177699618

Dufour, F., & Miller, B. (2006). Maximum principle for stochastic control problems. *SIAM Journal on Control and Optimization*, *45*(2), 668–698. https://doi.org/10.1137/040612403

Ethier, S., & Kurtz, T. (1986). *Markov processes characterization and convergence*. Wiley.

Feinberg, E., Mandava, M., & Shiryaev, A. N. (2013, December). Sufficiency of Markov policies for continuous-time Markov decision processes and solutions to Kolmogorov's forward equation for jump Markov processes. In *Proceedings of 2013 IEEE 52nd Annual Conference on Decision and Control*, Florence, Italy (pp. 5728–5732).

Ghosh, M. K., & Saha, S. (2012). *Continuous-Time controlled jump markov processes on the finite horizon*. Birkhäuser.

Guo, X. P., & Hernández-Lerma, O. (2009). *Continuous-time Markov decision processes. Theory and applications*. Springer-Verlag.

Guo, X. P., Huang, X. X., & Huang, Y. H. (2015). Finite-horizon optimality for continuous-time Markov decision processes with unbounded transition rates. *Advances in Applied Probability*, *47*(4), 1064–1087. https://doi.org/10.1239/aap/1449859800

Guo, X. P., Huang, Y. H., & Song, X. (2012). Linear programming and constrained average optimality for general continuous-time Markov decision processes in history-dependent policies. *SIAM Journal on Control and Optimization*, *50*(1), 23–47. https://doi.org/10.1137/100805169

Guo, X. P., & Liao, Z. W. (2019). Risk-sensitive discounted continuous-time Markov decision processes with unbounded rates. *SIAM Journal on Control and Optimization*, *56*(6), 3857–3883. https://doi.org/10.1137/18M1222016

Haussmann, U., & Suo, W. (1995a). Singular optimal stochastic controls I: Existence. *SIAM Journal on Control and Optimization*, *33*(3), 916–936. https://doi.org/10.1137/S0363012993250256

Haussmann, U., & Suo, W. (1995b). Singular optimal stochastic controls II: Dynamic programming. *SIAM Journal on Control and Optimization*, *33*(3), 937–959. https://doi.org/10.1137/S0363012993250529

Huang, Y. H. (2018). Finite horizon continuous-time Markov decision processes with mean and variance criteria. *Discrete Event Dynamic Systems*, *28*(4), 539–564. https://doi.org/10.1007/s10626-018-0273-1

Ishii, H. (1984). Uniqueness of unbounded viscosity solution of Hamilton-Jacobi equations. *Indiana University Mathematics Journal*, *33*(5), 721–748. https://doi.org/10.1512/iumj.1984.33.33038

Ishii, H. (1989). On uniqueness and existence of viscosity solutions of fully nonlinear second order elliptic PDE's. *Communications on Pure and Applied Mathematics*, *42*(1), 15–45. https://doi.org/10.1002/cpa.v42:1

Jensen, R. (1988). The maximum principle for viscosity solutions of second order fully nonlinear partial differential equations. *Archive for Rational Mechanics and Analysis*, *101*(1), 1–27. https://doi.org/10.1007/BF00281780

Kumar, K., & Chandan, P. (2015). Risk-sensitive control of continuous-time Markov processes with denumerable state space. *Stochastic Analysis and Applications*, *33*(5), 863–881. https://doi.org/10.1080/07362994.2015.1050674

Kushner, H. J. (1975). Existence results for optimal stochastic controls. *Journal of Optimization Theory and Applications*, *15*(4), 347–359. https://doi.org/10.1007/BF00933203

Meyer, P. A., & Zheng, W. A. (1984). Tightness criteria for laws of semimartingales. *Annales De L'IHP Probabilités Et Statistiques*, *20*, 353–372.

Miller, B. L. (1968). Finite state continuous time Markov decision processes with a finite planning horizon. *SIAM Journal on Control and Optimization*, *6*(2), 266–280. https://doi.org/10.1137/0306020

Piunovskiy, A., & Zhang, Y. (2011). Discounted continuous-time Markov decision processes with unbounded rates: The convex analytic approach. *SIAM Journal on Control and Optimization*, *49*(5), 2032–2061. https://doi.org/10.1137/10081366X

Pliska, S. R. (1975). Controlled jump processes. *Stochastic Processes and Their Applications*, *3*(3), 259–282. https://doi.org/10.1016/0304-4149(75)90025-3

Prieto-Rumeau, T., & Hernández-Lerma, O. (2012). *Selected topics on continuous-time controlled Markov chains and Markov games*. Imperial College Press.

Prieto-Rumeau, T., & J. M. Lorenzo (2010). Approximating ergodic average reward continuous-time controlled Markov chains. *IEEE Transactions on Automatic Control*, *55*(1), 201–207. https://doi.org/10.1109/TAC.2009.2033848

Shao, J. H. (2020). The existence of optimal feedback controls for stochastic dynamical systems with regime-switching. preprint, https://arxiv.org/abs/2003.13982.

Stroock, D. W., & Varadhan, S. R. S. (1979). *Multidimensional diffusion processes*. Springer-Verlag.

Yushkevich, A. (1978). Controlled Markov models with countable state space and continuous time. *Theory of Probability and Its Applications*, *22*(2), 215–235. https://doi.org/10.1137/1122029

Zhang, Y. (2017). Continuous-time Markov decision processes with exponential utility. *SIAM Journal on Control and Optimization*, *55*(4), 2636–2660. https://doi.org/10.1137/16M1086261

Appendix

In the section, we construct an example to illustrate that for discrete-time decision processes in an infinite state space, the optimisation problem may have essential difference between the control mechanism over history-dependent control policies and over Markovian policies. In this example, the value function corresponding to taking infimum over Markovian policies equals $+\infty$, while the one over history-dependent policies equals $-\infty$. Therefore, when analysing the influence of control policy class on the value function, more attentions should be paid.

Let the state space $X = \mathbb{Z}_+ = \{0, 1, 2, \dots\}$ and the action space $A = \mathbb{Z}_+$. Denote $\mathscr{P}(A)$ the set of probability measures on A, and let

$$\mathscr{P}_2(A) = \left\{ \mu \in \mathscr{P}(A) : \sum_{i \in A} i^2 \mu_i < \infty \right\}.$$

All the randomised policies considered in this example are assumed to take values in $\mathscr{P}_2(A)$. Consider the transition probability matrices given by

$$P_0(0) = 1, \quad P_1(j \mid i, a) = \begin{cases} \dfrac{1}{Kj^2}, & j \neq 0, \\ 0, & j = 0. \end{cases}$$

$$P_2(0 \mid i, a) = P_3(0 \mid i, a) = 1, \quad \forall\, i \in X,\ a \in A.$$

Here $K = \sum_{j=1}^{\infty} \frac{1}{j^2}$ is a constant. Let $\{\xi_k; k = 0, 1, 2, 3\}$ denote the controlled process. By the definition of $P_t(\cdot \mid i, a)$ above, it holds that

$$\mathbb{P}(\xi_0 = 0) = 1, \quad \mathbb{P}(\xi_1 \geqslant 1) = 1, \quad \mathbb{P}(\xi_2 = 0) = \mathbb{P}(\xi_3 = 0) = 1.$$

For a probability measure $\mu \in \mathscr{P}_2(A)$, denote by

$$m_1(\mu) = \sum_{i \geqslant 0} i \mu(i), \quad m_2(\mu) = \sum_{i \geqslant 0} i^2 \mu(i),$$

$$\mathrm{var}(\mu) = m_2(\mu) - (m_1(\mu))^2.$$

Let

$$\rho(x, \mu) = -2m_1(\mu) + \mathrm{var}(\mu) \tag{A1}$$

for $x \in X$ and $\mu \in \mathscr{P}_2(A)$. It is clear that $\rho(x, \mu)$ takes values in $(-\infty, +\infty)$, and is an unbounded function.

Define the cost function $c_t(\cdot, \cdot)$ by

$$c_1(i, \mu) = 0, \quad c_2(i, \mu) = i, \quad c_3(i, \mu) = \rho(i, \mu)$$
$$\text{for } i \in X \text{ and } \mu \in \mathscr{P}(A). \tag{A2}$$

For the control policy π,

$$V^{\pi}(i) := \mathbb{E}_i\left[\sum_{t=1}^{3} c_t(\xi_{t-1}, \pi_t)\right]. \tag{A3}$$

Let Π be the set of all history-dependent control policies, and Π^M the set of all Markov control policies. Clearly, $\Pi^M \subset \Pi$. The corresponding value functions are given by

$$V(i) = \inf_{\pi \in \Pi} V^{\pi}(i), \quad V^M(i) = \inf_{\pi \in \Pi^M} V^{\pi}(i), \ i \in X.$$

We shall show that

$$V(0) = -\infty, \quad \text{but } V^M(0) = +\infty. \tag{A4}$$

Indeed, according to (A2),

$$V^{\pi}(0) = \mathbb{E}\big[c_2(\xi_1, \pi_2) + c_3(\xi_2, \pi_3)\big] = \mathbb{E}[\xi_1 + \rho(\xi_2, \pi_3)\big]$$
$$= \mathbb{E}\big[\xi_1 + \rho(0, \pi_3)\big] \quad (\text{as } \xi_2 = 0 \text{ a.s.}).$$

Note that $\mathbb{E}[\xi_1] = \sum_{i=1}^{\infty} \frac{1}{Ki} = +\infty$.

For every Markov control policy π, π_3 is in $\mathscr{P}_2(A)$ and hence $\rho(0, \pi_3) < +\infty$. This further yields that $V^{\pi}(0) = +\infty$. Hence,

$$V^M(0) = \inf_{\pi \in \Pi^M} V^{\pi}(0) = +\infty.$$

For the set of history-dependent control policies, we choose a special one $\tilde{\pi}$ given by

$$\tilde{\pi}_1(\mathrm{d}x) = \delta_0(\mathrm{d}x), \quad \tilde{\pi}_2(\mathrm{d}x) = \delta_0(\mathrm{d}x), \quad \tilde{\pi}_3(\mathrm{d}x) = \delta_{\xi_1}(\mathrm{d}x).$$

Note that $\tilde{\pi}_3$ depends on ξ_1, not on ξ_2, so $\tilde{\pi}$ is not a Markov control policy. Also, it is clear that $\tilde{\pi}_3 \in \mathscr{P}_2(A)$. Then

$$V^{\tilde{\pi}}(0) = \mathbb{E}[\xi_1 - 2\xi_1] = -\mathbb{E}[\xi_1] = -\infty.$$

Hence,

$$V(0) = \inf_{\pi \in \Pi} V^{\pi}(0) \leqslant V^{\tilde{\pi}}(0) = -\infty.$$

Consequently, we have proved the desired result (A4).