Viscosity Solutions Approach to Finite-Horizon Continuous-Time Markov Decision Processes

Zhong-Wei Liao

Beijing Normal University (Zhuhai)

zhwliao@bnu.edu.cn

(work with Prof. Jinghai Shao)

Report on The 18th Workshop on Markov Processes and Related Topics
Tianjin University

2023-07-31

Outline

1. **Introduction**

– *Background: CTMDPs and Control Class*

– *Motivation - an Example*

Background: CTMDPs and Control Class

▷ Continuous-time Markov decision process:

CTMDPs have been studied intensively due to their rich application in queuing systems, population processes.

→ Monographs:

Guo & Hernández-Lerma (2009);

Bäuerle & Rieder (2011);

Ghosh & Saha (2012);

Prieto-Rumeau & Hernández-Lerma (2012).

Background: CTMDPs and Control Class

CTMDPs is defined on $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geqslant 0}, \mathbb{P})$ with usual conditions. The state space is denumerable, i.e., $\mathcal{S} = \{1, 2, \ldots\}$, the action space U is a compact subset of $\mathbb{R}^k$.

For each $\mu \in \mathscr{P}(U)$, $Q(\mu) = (q_{ij}(\mu))_{i,j\in\mathcal{S}}$ is a transition rate matrix over the state space $\mathcal{S}$. The jump process X_t is an $\mathscr{F}_t$-adapted process on $\mathcal{S}$ satisfying

$$\mathbb{P}\left(X_{t+\delta} = j \middle| X_t = i, \mu_t = \mu\right) = \begin{cases} q_{ij}(\mu)\delta + o(\delta), & \text{if } i \neq j\ , \\ 1 + q_{ii}(\mu)\delta + o(\delta), & \text{otherwise.} \end{cases}$$

Background: CTMDPs and Control Class

For each control policy $\pi = (X_\cdot, \mu_\cdot, s, i) \in \Pi_{s,i}$, the objective function with **finite horizon optimization criterion** is

$$J(s, i, \pi) = \mathbb{E}\left[\int_s^T f(t, X_t, \mu_t)\mathrm{d}t + g(T, X_T)\right],$$

where f is a running cost function and g is a terminal cost function. In correspondence with this objective function, the value function $V(s, i)$ is defined as

$$V(s, i) = \inf_{\pi \in \Pi_{s,i}} J(s, i, \pi), \quad (s, i) \in [0, T] \times \mathcal{S}.$$

The control class $\Pi_{s,i}$ is of all admissible **history-dependent control policies** with initial condition $(s, i) \in [0, T] \times \mathcal{S}$.

Background: CTMDPs and Control Class

- ▷ Finite horizon optimization criterion:

  Piunovskiy & Zhang (SICON, 2011); Guo et al. (AAP, 2015);
  Guo & Liao (SICON, 2019); Huang (DEDS, 2018).

- ▷ History-dependent control policies:

  Kushner (JOTA, 1975); Pliska (SPA, 1975); Yushkevich (TPA, 1978);
  Feinberg et al. (IEEE, 2013); Dufour & Miller (SICON, 2006);
  Guo et al. (SICON, 2012); Kumar & Chandan (SAA, 2015).

Motivation - an Example

The story starts from a well-known conclusion.

Theorem on history-dependent control (cf. Feinberg et al., IEEE, 2013)

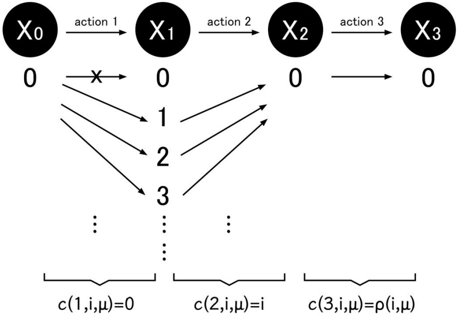
Let the initial state $x \in \mathcal{S}$ be fixed. For any history-dependent policy π, there exists a Markov policy φ such that for all $Z \in \mathscr{B}(\mathcal{S})$, $B \in \mathscr{B}(U)$ and $t \in \mathbb{R}_+$,

$$P_x^\pi(t; Z, B) = P_x^\varphi(t; Z, B), \quad \text{where } P_x^\pi(t; Z, B) := \int_\Omega \mathbb{1}_{\{X_t \in Z\}} \pi(B|\omega, t) \mathbb{P}_x(\mathrm{d}\omega).$$

This result shows the existence of a Markov policy with the same state-action distribution as that of an history-dependent policy. It implies that **it is sufficient to restrict the set of history-dependent policies to the set of Markov policies** for problems with multiple criteria and constraints. (Feinberg et al., IEEE, 2013, p.5731, line 4 from below).

Motivation - an Example

Let $\mathcal{S} = U = \{0, 1, \ldots\}$, decision time $[0, T] = \{0, 1, 2, 3\}$.



$$\mathbb{P}(X_0 = 0) = 1;$$

$$P_1(j|i, a) = \begin{cases} \frac{1}{Kj^2}, & j \neq 0, \\ 0, & j = 0; \end{cases}$$

$$P_2(0|i, a) = P_3(0|i, a) = 1, \quad \forall i \in \mathcal{S}.$$

where $K := \sum_{j=1}^{\infty} \frac{1}{j^2}$. Then,

$$\mathbb{P}(X_0 = 0) = 1, \ \mathbb{P}(X_1 \geqslant 1) = 1,$$

$$\mathbb{P}(X_2 = 0) = \mathbb{P}(X_3 = 0) = 1.$$

For $\mu \in \mathscr{P}(U)$, let $M_i(\mu) = \sum_{k \geqslant 0} k^i \mu(k)$ and $\mathrm{Var}(\mu) = M_2(\mu) - M_1^2(\mu)$. Define

$$\rho(i, \mu) = \left[-2M_1(\mu) + \infty \cdot \mathrm{Var}(\mu) \right] \mathbb{1}_{\{M_1(\mu) < \infty\}}, \quad \forall i \in \mathcal{S}.$$

Motivation - an Example

$$\rho(i,\mu) = \left[-2M_1(\mu) + \infty\cdot \mathrm{Var}(\mu)\right]\mathbb{1}_{\{M_1(\mu)<\infty\}}, \quad \forall i \in \mathcal{S}.$$

Here, $\rho(i,\mu)$ takes value in $(-\infty,+\infty]$. If $M_1(\mu) < +\infty$ and $\mathrm{Var}(\mu) > 0$, then $\rho(i,\mu) = +\infty$. If $\rho(i,\mu) < +\infty$, μ must be a Dirac measure in the form $\mu = \delta_{i_0}$ for some $i_0 \in U$.

For any control policy π, consider the objective function on $(0,0) \in [0,T] \times \mathcal{S}$

$$J(0,0,\pi) := \mathbb{E}\left[\sum_{t=1}^{3} c(t, X_{t-}, \pi_t)\middle| X_0 = 0\right], \text{ for any } \pi \in \Pi \text{ or } \Pi^M,$$

where Π (resp. Π^M) is the set of history-dependent (resp. Markov) control policies, and $\Pi^M \subset \Pi$. We will show that

$$V(0,0) = \inf_{\pi\in\Pi} J(0,0,\pi) = -\infty \quad \text{but} \quad V^M(0,0) = \inf_{\pi\in\Pi^M} J(0,0,\pi) = +\infty.$$

Motivation - an Example

By the definition,

$$J(0,0,\pi) = \mathbb{E}\big[c(2, X_1, \pi_2) + c(3, X_2, \pi_3)\big] = \mathbb{E}\big[X_1 + \rho(X_2, \pi_3)\big]$$
$$= \mathbb{E}\big[X_1 + \rho(0, \pi_3)\big] \quad (X_2 = 0 \text{ a.s. and } \mathbb{E}(X_1) = \sum_i \frac{1}{Ki} = +\infty).$$

For every **Markov control policy** $\pi^M \in \Pi^M$,

(i) if $\rho(0, \pi_3^M) = +\infty$ (i.e., $M_1(\pi_3^M) < \infty$ and $\mathrm{Var}(\pi_3^M) > 0$), then $J(0,0,\pi^M) = +\infty$.

(ii) if $\rho(0, \pi_3^M) < +\infty$, then π_3^M must be a Dirac measure. π_3^M is a Markov policy, which is a functional of X_2 (note $X_2 \equiv 0$), there exists a function $f : \mathcal{S} \to U$ such that $\pi_3^M(\mathrm{d}x) = \delta_{f(0)}(\mathrm{d}x)$. Hence, $J(0,0,\pi^M) = \mathbb{E}[X_1 - 2f(0)] = +\infty$.

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

Motivation - an Example

For the **history-dependent control policy**, we choose a special one π^H given by

$$\pi_1^H(\mathrm{d}x) = \pi_2^H(\mathrm{d}x) = \delta_0(\mathrm{d}x), \quad \pi_3^H(\mathrm{d}x) = \delta_{X_1}(\mathrm{d}x).$$

Note that π_3^H depends on X_1, is not Markov, i.e. $\pi^H \notin \Pi^M$. Then,

$$J(0,0,\pi^H) = \mathbb{E}\big[X_1 - 2X_1\big] = -\mathbb{E}[X_1] = -\infty.$$

Therefore, $V(0,0) = \inf_{\pi \in \Pi} J(0,0,\pi) = -\infty$.

Compare with Feinberg et al. (IEEE, 2013)

The value functions in two control classes are fundamentally different. Therefore, it **can not restrict the set of history-dependent policies to the set of Markov policies** without providing additional conditions.

Motivation - an Example

▷ **Motivation 1**:

How much **improvement can the value function be achieved** by expanding control classes, such as deterministic policy $\to$ stochastic Markov policy $\to$ history-dependent policy?

Motivation - an Example

▷ **Motivation 2**:

There are still some **unresolved issues** regarding Markov decision processes with history-dependent policies. For example:

Definition in Guo et al. (AAP, 2015, p. 1069)

The expected T-horizon criterion $V_\pi(0, i)$ is defined by

$$V_\pi(t, i) = \mathbb{E}\left[\int_t^T r(s, X_s, \pi_s)\mathrm{d}s + g(T, X_T)\Big| X_t = i\right], \quad \forall (t, i) \in [0, T] \times \mathcal{S}.$$

The T-horizon value function of the CTMDPs is define as

$$V^*(0, i) = \sup_{\pi \in \Pi} V_\pi(0, i),$$

$$V^*(t, i) = \sup_{\pi \in \Pi^M} V_\pi(t, i), \quad \text{for } (t, i) \in (0, T] \times \mathcal{S},$$

where Π (resp. Π^M) is the set of history-dependent (resp. Markov) control policies.

Motivation - an Example

▷ **Motivation 3**:

For the Hamilton-Jacobi-Bellman (HJB) equations induced by the value function of CT-MDPs, the higher order smoothness of the solution is considered. Therefore, we introduce the concept of **viscosity solution** in CTMDPs.

2. Canonical Path Space and History-Dependent Control

- *History-Dependent Control and Skorokhod's Representation*
- *Canonical Path Space and Compactification Method*
- *Mind Mapping*

History-Dependent Control and Skorokhod's Representation

CTMDPs is defined on $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geqslant 0}, \mathbb{P})$ with usual conditions. The state space is denumerable $\mathcal{S} = \{1, 2, \ldots\}$ and the action space U is a compact subset of $\mathbb{R}^k$. For each $\mu \in \mathscr{P}(U)$, $Q(\mu) = \left(q_{ij}(\mu)\right)_{i,j\in\mathcal{S}}$ is a irreducible and conservative rate matrix, and the jump process X_t satisfying

$$\mathbb{P}\left(X_{t+\delta} = j \middle| X_t = i, \mu_t = \mu\right) = \begin{cases} q_{ij}(\mu)\delta + o(\delta), & \text{if } i \neq j\,, \\ 1 + q_{ii}(\mu)\delta + o(\delta), & \text{otherwise.} \end{cases} \qquad (\triangle)$$

History-Dependent Control and Skorokhod's Representation

Definition: admissible history-dependent control

Policy $\pi = (X_\cdot, \mu_\cdot, s, i)$ is called admissible history-dependent control policy if:

(i) X_t is an $\mathscr{F}_t$-adapted jump process satisfying ($\triangle$) with initial value $X_s = i$.

(ii) exists a measurable map F_t such that $\mu_t = F_t(X_{s,t})$, where $X_{s,t} = \left\{ X_r \middle| r \in [s, t] \right\}$.

For $\pi = (X_\cdot, \mu_\cdot, s, i) \in \Pi_{s,i}$, the objective function is:

$$J(s, i, \pi) = \mathbb{E}\left[\int_s^T f(t, X_t, \mu_t)\mathrm{d}t + g(T, X_T)\right],$$

The value function $V(s, i)$ is defined as:

$$V(s, i) = \inf_{\pi \in \Pi_{s,i}} J(s, i, \pi), \quad (s, i) \in [0, T] \times \mathcal{S}.$$

If $V(s, i) = J(s, i, \pi^*)$ for some $\pi^* \in \Pi_{s,i}$, π^* is called optimal.

History-Dependent Control and Skorokhod's Representation

For any $\mu \in \mathscr{P}(U)$, construct a family of intervals $\{\Gamma_{ij}(\mu) : i, j \in \mathcal{S}\}$ on $[0, \infty)$ as

$$\Gamma_{ij}(\mu) = \left[\sum_{k=1}^{i-1} q_k(\mu) + \sum_{l=1}^{j-1} q_{il}(\mu),\ \sum_{k=1}^{i-1} q_k(\mu) + \sum_{l=1}^{j} q_{il}(\mu) \right),$$

In particular, let $\Gamma_{ij}(\mu) = \emptyset$ when $j = i$ or $q_{ij}(\mu) = 0$.

Therefore, $\{\Gamma_{ij}(\mu) : i, j \in \mathcal{S}\}$ is pairwise disjoint and the length of interval $\Gamma_{ij}(\mu)$ equals $q_{ij}(\mu)$. Next, define $\vartheta(i, \mu, z) = \sum_{j\in\mathcal{S}} (j - i)\mathbb{1}_{\Gamma_{ij}(\mu)}(z)$. Then, **Skorokhod's representation** said X_t can be represented by the solution of SDE:

$$\mathrm{d}X_t = \int_{\mathbb{R}_+} \vartheta(X_{s-}, \mu_{s-}, z) N(\mathrm{d}s, \mathrm{d}z), \qquad X_0 = i.$$

Canonical Path Space and Compactification Method

$$\textit{Process}: \ \mathscr{D}([0,T];\mathcal{S}) := \left\{ f \,\middle|\, f: [0,T] \to \mathcal{S}, \text{ right-continuous with left limits} \right\}.$$

$$\textit{Control}: \ \mathscr{U} := \left\{ \mu_\cdot \,\middle|\, \mu_\cdot : [0,T] \to \mathscr{P}(U), \text{ Borel measurable} \right\}.$$

The canonical path space of our optimization problem is

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

which is endowed with the product topology. We can construct L_1-Wasserstein distance $W_{1,\mathscr{X}}$ on the space $\mathscr{P}(\mathscr{X})$ and make it a Polish space:

$$W_{1,\mathscr{X}}(R_1,R_2) = \inf_{\zeta \in \mathscr{C}(R_1,R_2)} \left\{ \iint_{\mathscr{X}\times\mathscr{X}} \rho\big((X_\cdot,\mu_\cdot),(X'_\cdot,\mu'_\cdot)\big) \mathrm{d}\zeta \right\}, \quad R_1,R_2 \in \mathscr{P}(\mathscr{X}).$$

where $\rho\big((X_\cdot,\mu_\cdot),(X'_\cdot,\mu'_\cdot)\big) = \|X_\cdot - X'_\cdot\|_\infty + W_1(\mu,\mu')$.

Canonical Path Space and Compactification Method

For any $\pi = (X_t, \mu_t, s, i) \in \Pi_{s,i}$, define a measureable map $\Phi_\pi : \Omega \to \mathscr{X}$ as

$$\Phi_\pi(\omega) = (X_t(\omega), \mu_t(\omega))_{t\in[0,T]}.$$

Hence, Φ_π induces a probability measure on the canonical path space $\mathscr{X}$, which is

$$R_\pi := \mathbb{P} \circ \Phi_\pi^{-1}.$$

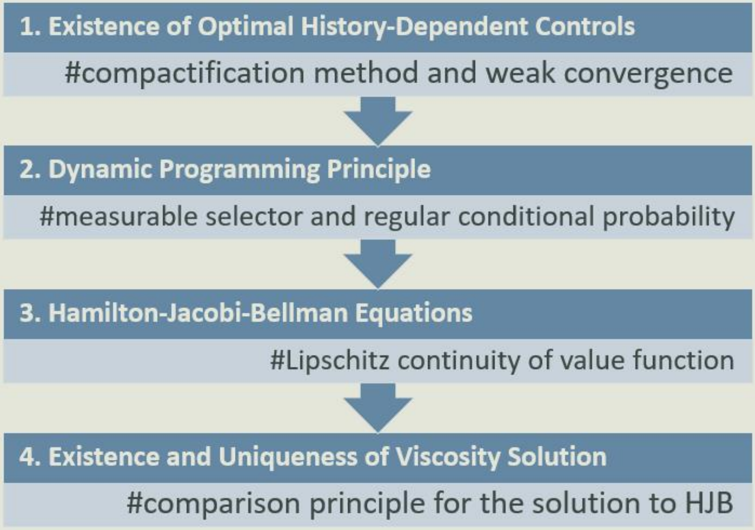
Introduce the probability measure set on $\mathscr{X}$ induced by controls as

$$\mathscr{R}_{s,i} := \left\{ R \in \mathscr{P}(\mathscr{X}) \,\middle|\, \text{exists } \pi \in \Pi_{s,i} \text{ such that } R = \mathbb{P} \circ \Phi_\pi^{-1} \right\}.$$

By the definition of value function, we have

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

Mind Mapping



3. **Main Results**

– *Assumptions*

– *Existence of Optimal Controls & Dynamic Programming Principle*

– *HJB Equations & Existence and Uniqueness of Viscosity Solution*

Assumptions

To prove the existence of optimal history-dependent control, we introduce the following assumptions:

Assumptions

$H1:$ $\mu \to q_{ij}(\mu)$ is continuous, and there exists $M := \sup_{i\in\mathcal{S}} \sup_{\mu\in\mathscr{P}(U)} q_i(\mu) < \infty$.

$H2:$ exist compact function $\Phi : \mathcal{S} \to [1,\infty)$, $B_0 \in \mathscr{B}(\mathcal{S})$ and constants $\lambda > 0$, $\kappa < \infty$ such that

$$Q(\mu)\Phi(u) := \sum_{j\neq u} q_{uj}(\mu)\big(\Phi(j) - \Phi(u)\big) \leqslant \lambda\Phi(u) + \kappa\mathbb{1}_{B_0}(u), \quad u \in \mathcal{S}, \mu \in \mathscr{P}(U).$$

$H3:$ exists a constant $K \in \mathbb{N}$ such that $\forall u \in \mathscr{P}(U)$, $q_{ij}(\mu) = 0$ when $|j - i| > K$.

Existence of Optimal Controls & Dynamic Programming Principle

Thm1: Existence of Optimal Controls

Assuming that (H1) - (H3) hold, then for all $(s, i) \in [0, T] \times \mathcal{S}$, there exists an optimal history-dependent control policy $\pi^* \in \Pi_{s,i}$, such that $V(s, i) = J(s, i, \pi^*)$.

Thm2: Dynamic Programming Principle

Assume (H1) - (H3) hold. For any $(s, i) \in [0, T] \times \mathcal{S}$ and $\mathscr{X}_t$-stopping time τ satisfying $s \leqslant \tau \leqslant T$, the following holds

$$V(s, i) = \inf \left\{ \mathbb{E}_R \left[\int_s^\tau f(t, X_t, \mu_t) \mathrm{d}t + V(\tau, X_\tau) \right] \,\middle|\, R \in \mathscr{R}_{s,i} \right\}.$$

HJB Equations & Existence and Uniqueness of Viscosity Solution

Easy to show that $V(s, i)$ is Lipschitz continuous with respect to the time variable s. Hence, $V(t, i)$ is differentiable in $[0, T]$ a.s..

$$-\frac{\partial v}{\partial t} - \inf_{\mu \in \mathscr{P}(U)} \left\{ \sum_{j \neq i} q_{ij}(\mu) \left[v(t, j) - v(t, i) \right] + f(t, i, \mu) \right\} = 0, \quad v(T, i) = g(T, i). \tag{1}$$

In some practical applications, the property of almost everywhere differentiable is not enough, especially when $\mathcal{S}$ is a general state space rather than a countable space. It is useful to introduce the concept of viscosity solution to further characterize $V(t, i)$.

HJB Equations & Existence and Uniqueness of Viscosity Solution

Def: Viscosity Solution

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

(i) v is called a viscosity supersolution, if (t_0, i_0) is a minimum point of $v - \phi$ ($\phi \in C^\infty([0, T] \times \mathcal{S})$), it holds that

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

(ii) v is called a viscosity subsolution, if (t_0, i_0) is a maximum point of $v - \phi$ ($\phi \in C^\infty([0, T] \times \mathcal{S})$), it holds that

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

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

HJB Equations & Existence and Uniqueness of Viscosity Solution

Thm3: Existence of Viscosity Solution

Assuming that (H1) - (H3) hold, the value function $V(t, i)$ is a viscosity solution to HJB equation.

Thm4: Comparison Principle for HJB

Assume (H1) - (H3) hold. Let V_1 (resp. V_2) be a viscosity supersolution (resp. viscosity subsolution) of HJB equation. Then

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

Cor: Uniqueness of Viscosity Solution

Assume (H1) - (H3) hold. The value function $V(t, i)$ is the unique viscosity solution to the HJB equation.

4. Some Proofs

– Existence of Optimal History-Dependent Controls

Proof. Existence of Optimal Controls

The case $V(s, i) = \infty$ is trivial. Focus on $0 \leqslant V(s, i) < \infty$, the proof is completed in three steps.

Step 1. Let $\pi_n = (X_\cdot^{(n)}, \mu_\cdot^{(n)}, s, i) \in \Pi_{s,i}$, such that $\lim_{n\to\infty} J(s, i, \pi_n) = V(s, i) < \infty$. Let R_n be the probability $\mathscr{X}$ associated with π_n.

Denote by $\mathscr{L}_X^n$ and $\mathscr{L}_\mu^n$ by the marginal distribution of X_t and μ_t, respectively.

i The tightness of $\mathscr{L}_X^n$ and $\mathscr{L}_\mu^n$.

ii The tightness of R_n.

Proof. Existence of Optimal Controls

Step 2. Since R_n is tight, there exists a subsequence R_{n_k} converges weakly to some probability distribution R^*. According to the Skorokhod representation theorem, there exists a probability space $(\Omega', \mathscr{F}', \mathbb{P}')$, and random variables Y^* and Y_{n_k}

$$\lim_{k\to\infty} Y_{n_k} = Y^*, \quad \mathbb{P}' - \text{a.s.},$$

$$Y_{n_k} = \left(X_t^{(n_k)}, \mu_t^{(n_k)}\right)_{t\in[s,T]} \sim R_{n_k} \quad \text{and} \quad Y^* = (X_t^*, \mu_t^*)_{t\in[s,T]} \sim R^*.$$

Next, We show that $\pi^* = (X_t^*, \mu_t^*, s, i)$ is an admissible history-dependent control policy on probability space $(\Omega', \mathscr{F}', \mathscr{F}_t', \mathbb{P}')$.

Proof. Existence of Optimal Controls

Step 3. Finally, all that remains is to prove that π^* is optimal:

$$V(s,i) = \lim_{k\to\infty} J(s,i,\pi_{n_k}) = \lim_{k\to\infty} \mathbb{E}\left[\int_s^T f(t, X_t^{(n_k)}, \mu_t^{(n_k)})\mathrm{d}t + g(T, X_T^{(n_k)})\right]$$
$$\geqslant \mathbb{E}\left[\int_s^T f(t, X_t^*, \mu_t^*)\mathrm{d}t + g(T, X_T^*)\right] = J(s,i,\pi^*) \geqslant V(s,i).$$

Thank you for your attention!