# The existence of optimal control for continuous-time Markov decision processes in random environments

Jinghai Shao

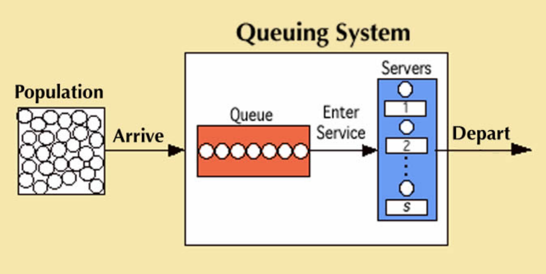
Center for Applied Mathematics, Tianjin University

July 12, 2019

# Contents

# Contents

- CTMDPs have been extensively studied and widely applied in various application fields such as telecommunication, queueing systems, population processes, epidemiology, and so on.
- As an illustrative example, consider the controlled queueing systems:



# Control Model

Consider the state space $\mathcal{S} = \{1, 2, \ldots\}$, on which there exists a continuous-time Markov chain $(\Lambda_t)$ with

$$(q_{ij}(a)) \quad \text{for } a \in U, \text{ action space.}$$

Assume

$$U \subset \mathbb{R}^k, \quad \text{compact;} \sum_{j \in \mathcal{S}} q_{ij}(a) = 0, \quad \forall i \in \mathcal{S},\ a \in U;$$

$$\sup_{a \in U} \sup_{i \in \mathcal{S}} q_i(a) < \infty.$$

For example, choose appropriate control policy to minimize the cost

- `finite-horizon expected cost:`
  $$V_T(i,\pi) := \mathbb{E}\Big[\int_0^T c(\Lambda_t,\pi_t)\mathrm{d}t\Big],$$ where $T>0$.
- `infinite-horizon expected discounted cost:`
  $$V(i,\pi) := \mathbb{E}\Big[\int_0^\infty \mathrm{e}^{-\lambda t} c(\Lambda_t,\pi_t)\mathrm{d}t\Big],$$ where $\lambda>0$, discount factor.

Randomized Markov policies: A randomized Markov policy is a real-valued function $\pi_t(C|i)$ that satisfies the following conditions:

(i) For all $i\in\mathcal{S}$ and $C\in\mathscr{B}(U)$, the mapping $t\mapsto\pi_t(C|i)$ is measurable;

(ii) For all $i\in\mathcal{S}$, $t\geq 0$, $C\mapsto\pi_t(C|i)$ is a probability measure on $\mathscr{B}(U)$.

*stationary* : if $\pi_t(C|i)\equiv\pi(C|i)$.

*deterministic* : if $\pi_t(C|i)=\delta_{u_t}(C|i)$, Dirac measure.

♣ $\Pi$ : the set of all randomized Markov policies.

* X.P. Guo, Hernandez-Lerma, Springer-Verlag, Berlin, 2009.
* X.P. Guo, X. Huang, Y. Huang, *Finite-horizon optimality for CTMDPs with unbounded transition rates*, Adv. Appl. Prob. 2015.
* X.P. Guo, U. Rieder, *Average optimality for CTMDPs in Polish spaces*, Ann. Appl. Probab. 2006.
* A. Piunovskiy, Y. Zhang, *Discounted CTMDPs with unbounded rates: the convex analytic approach*, SIAM J. Control Optim. 2011.

# An existing method

Consider

$$J_\lambda(i,\pi) := \mathbb{E}\Big[\int_0^\infty \mathrm{e}^{-\lambda t} c(\Lambda_t,\pi_t)\mathrm{d}t\Big],$$

and the corresponding value function

$$J_\lambda^*(i) := \inf_{\pi\in\Pi} J_\lambda(i,\pi).$$

Key point: The function $J_\lambda^*$ satisfies the HJB equation

$$J_\lambda^*(i) = \inf_{a\in U}\Big\{\frac{c(i,a)}{\lambda+q_i(a)} + \frac{1}{\lambda+q_i(a)}\sum_{j\neq i} J_\lambda^*(j)q_{ij}(a)\Big\},\quad i\in\mathcal{S}.$$

Let

$$\varphi_{ij}^{(n)}(a) := \begin{cases} \frac{\delta_{ij}}{\lambda+q_i(f)} & n=1, \\ \frac{1}{\lambda+q_i(f)}\left[\delta_{ij} + \sum_{k\neq i} q_{ik}(f)\varphi_{kj}^{(n-1)}(f)\right] & n=2. \end{cases}$$

Then

$$\begin{aligned} J_\lambda(i,f) &= \sum_{j\in\mathcal{S}} \int_0^\infty \mathrm{e}^{-\lambda t} c(j,f) P_f(0,i,t,j)\mathrm{d}t \\ &= \sum_{j\in\mathcal{S}} c(j,f)\left[\lim_{n\to\infty} \varphi_{ij}^{(n)}(f)\right]. \end{aligned}$$

# Framework

Let us consider further a diffusion process satisfying SDE:

$$\mathrm{d}X_t = b(X_t, \Lambda_t)\mathrm{d}t + \sigma(X_t, \Lambda_t)\mathrm{d}B_t,$$

where $(B_t)$ is a $d$-dimension B.M., $b : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}^d$, and $\sigma : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}^{d\times d}$.
The optimal control problem:

$$\inf_{\Pi} \mathbb{E}\Big[\int_0^T f(t, X_t, \Lambda_t, \mu_t)\mathrm{d}t + g(X_T, \Lambda_T)\Big],$$

where $\Pi$ is the set of admissible control policies which will be given later.

# Some notations

1. Let $\psi : [0,T] \to [0,\infty)$ be an increasing function such that
$$\lim_{r\to 0} \psi(r) = 0 \qquad \forall\, r \in [0,T].$$
2. $\mathscr{P}(U)$: all the probab. measures over $U$, endowed with the $L^1$-Wasserstein distance, becoming a Polish space.
3. $\mathcal{D}([0,T];\mathscr{P}(U))$: measurable maps $[0,T] \mapsto (\mathscr{P}(U), W_1)$, càdlàg.
4. Endow $\mathcal{D}([0,T];\mathscr{P}(U))$ with the pseudopath topology, which makes it being a Polish space.
5. For $\mu : [0,T] \to \mathscr{P}(U)$ in $\mathcal{D}([0,T];\mathscr{P}(U))$, put
$$w_\mu([a,b)) = \sup\{W_1(\mu_t,\mu_s);\ s,t \in [a,b)\}, \quad a,b \in [0,T], a<b;$$
$$w''_\mu(\delta) = \sup \min\left\{W_1(\mu_t,\mu_{t_1}), W_1(\mu_t,\mu_{t_2})\right\},$$
where the supremum is taken over $t_1$, $t$, and $t_2$ satisfying
$$t_1 \le t \le t_2, \qquad t_2 - t_1 \le \delta.$$

The process $(X_t)$ is determined by the following SDE:

$$\mathrm{d}X_t = b(X_t, \Lambda_t)\mathrm{d}t + \sigma(X_t, \Lambda_t)\mathrm{d}B_t, \tag{1}$$

where $(B_t)$ is a Brownian motion; $(\Lambda_t)$ is a continuous-time Markov process on $\mathcal{S}$ associated with the $q$-pair $(q_i(u))$, $q_{ij}(u))$ satisfying

$$\mathbb{P}(\Lambda_{t+\delta} = j|\Lambda_t = i, \mu_t = \mu) = \begin{cases} q_{ij}(\mu)\delta + o(\delta) & i \neq j, \\ 1 - q_i(\mu)\delta + o(\delta), & i = j, \end{cases} \tag{2}$$

provided $\delta > 0$. The decision-maker still tries to minimize the cost through controlling the transition rates of the Markov chain $(\Lambda_t)$, but now the cost function may depend on the diffusion process $(X_t)$.

## Definition

A $\psi$-relaxed control is a term $\alpha = (\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P}, B_t, X_t, \Lambda_t, \mu_t, s, x, i)$ such that

(1) $(s, x, i) \in [0, T] \times \mathbb{R}^d \times \mathcal{S}$;

(2) $(\Omega, \mathscr{F}, \mathbb{P})$ is a probability space with the filtration $\{\mathscr{F}_t\}_{t\in[0,T]}$;

(3) $(B_t)$ is a $d$-dim B.M. on $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P})$, and $(X_t, \Lambda_t)$ is a stochastic process on $\mathbb{R}^d \times \mathcal{S}$ satisfying (1) and (2) with $X_s = x$, $\Lambda_s = i$;

(4) $\mu_t \in \mathscr{P}(U)$ is adapted to the $\sigma$-field generated by $\Lambda_t$, $t \mapsto \mu_t$ is in $\mathcal{D}([0, T]; \mathscr{P}(U))$ almost surely, and for every $i' \in \mathcal{S}$ the curve $t \mapsto \nu_t(\,\cdot, i'\,) := \mu_t(\,\cdot\,|\Lambda_t = i')$ satisfies

$$w''_\nu(\delta) \leq \psi(\delta), \quad \delta \in (0, T];$$

– The collection of all $\psi$-relaxed control with initial value $(s, x, i)$ is denoted by $\widetilde{\Pi}_{s,x,i}$.

Actually, the randomized policy can be viewed as a Markov feedback control.

$$\begin{aligned}\mu_t(C) &= \sum_{i\in\mathcal{S}} \pi_t(C|i)\mathbf{1}_{\Lambda_t=i}\\ &= \pi_t(C|\Lambda_t), \qquad t\geq 0.\end{aligned}$$

# Contents

# Assumptions

(H1) $U \subset \mathbb{R}^k$ is a compact set for some $k \in \mathbb{N}$.

(H2) $\forall\, u \in U$, $(q_{ij}(u))$ is conservative. $M := \sup_{u\in U} \sup_{i\in\mathcal{S}} q_i(u) < \infty$.

(H3) $\forall\, i, j \in \mathcal{S}$, $u \mapsto q_{ij}(u)$ is continuous on $U$.

(H4) $\exists$ a compact function $\Phi : \mathcal{S} \to [1, \infty)$, a compact set $B_0 \in \mathscr{B}(\mathcal{S})$, constants $\lambda > 0$ and $\kappa_0 < \infty$ such that

$$Q_u\Phi(i) := \sum q_{ij}(u)\Phi(j) \leq \lambda\Phi(i) + \kappa_0 \mathbf{1}_{B_0}(i), \qquad i \in \mathcal{S},\ u \in U.$$

(H5) $\exists$ a constant $C_1 > 0$ such that

$$|b(x,i)-b(y,i)|^2+\|\sigma(x,i)-\sigma(y,i)\|^2 \leq C_1|x-y|^2, \quad x, y \in \mathbb{R}^d,\ i \in \mathcal{S},$$

where $|x|^2 = \sum_{k=1}^d x_k^2$, $\|\sigma\|^2 = \mathrm{tr}(\sigma\sigma')$.

(H6) $\exists\, C_2 > 0$ such that $|b(x,i)|^2+\|\sigma(x,i)\|^2 \leq C_2(1+|x|^2)$, $x \in \mathbb{R}^d, i \in \mathcal{S}$.

## Theorem 1

Assume that (H1)-(H6) hold, and $f$, $g$ are lower semi-continuous functions bounded from below. Then for every $s \in [0, T)$, $x \in \mathbb{R}^d$, $i \in \mathcal{S}$, there exists an optimal $\psi$-relaxed control $\alpha^* \in \widetilde{\Pi}_{s,x,i}$, i.e.

$$\begin{aligned} V(s,x,i) &= J(s,x,i,\alpha^*) \\ &= \inf_{\alpha \in \widetilde{\Pi}_{s,x,i}} \mathbb{E}\Big[ \int_s^T f(t, X_t, \Lambda_t, \mu_t) \mathrm{d}t + g(X_T, \Lambda_T) \Big]. \end{aligned}$$

## Theorem 2

Suppose (H1)-(H6) hold. Assume that $f$ and $g$ are continuous functions and there exists a positive constant $C_3$ such that

$$|f(t,x,i,u)-f(t,x',i,u)| + |g(x,i) - g(x',i)| \leq C_3|x-x'|,$$
$$|f(t,x,i,u)| + |g(x,i)| \leq C_3,$$

for every $t \in [0,T]$, $x, x' \in \mathbb{R}^d$, $i \in \mathcal{S}$ and $u \in U$. Then $V(s,x,i)$ is continuous on $[0,T] \times \mathbb{R}^d \times \mathcal{S}$.

**Theorem 3 (Dynamic programming principle)**

Assume all the conditions of Theorem 2 are still valid. Then for $s < t < T$,

$$V(s,x,i) = \inf\Big\{\mathbb{E}_\alpha\Big[\int_s^t f(r,X_r,\Lambda_r,\mu_r)\mathrm{d}r + V(t,X_t,\Lambda_t)\Big];\ \alpha\in\widetilde{\Pi}_{s,x,i}\Big\}.$$

# Key steps to prove Theorem 1

To simplify the proof, transform the relaxed controls into the canonical path space. Let

$$\mathcal{U} = \{\nu \in \mathcal{D}([0,T];\mathscr{P}(U));\ w''_\nu(\delta) \le \psi(\delta)\}$$
$$\mathcal{Y} = C([0,T];\mathbb{R}^d) \times \mathcal{D}([0,T];\mathcal{S}) \times \mathcal{U}.$$

Denote by $\widetilde{\mathcal{D}}$, $\widetilde{\mathcal{U}}$ the Borel measurable sets, and $\widetilde{\mathcal{D}}_t$, $\widetilde{\mathcal{U}}_t$ the $\sigma$-fields up to time $t$. Each $\psi$-relaxed control $\alpha = (\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P}, B_t, X_t, \Lambda_t, \mu_t, s, x, i)$ can be transformed into $\mathcal{Y}$ via the map

$$\Psi(\omega) = (X_t(\omega), \Lambda_t(\omega), \mu_t(\omega))_{t\in[0,T]},$$

with $X_r := x, \Lambda_r := i, \mu_r := \mu_s,\ \forall r \in [0,s]$.
We can use $R := \mathbb{P} \circ \Psi^{-1}$ to represent the control $\alpha$ in canonical space $\mathcal{Y}$.

# Key steps to prove Theorem 1

Consider the nontrivial case $V(0,x,i)<\infty$, and $\exists$ $(R_n)_{n\geq 1}$ such that

$$\lim_{n\to\infty} J(0,x,i,R_n)=V(0,x,i). \tag{e1}$$

1. Prove the tightness of the distributions of $(X_t)_{t\in[0,T]}$, $(\Lambda_t)_{t\in[0,T]}$ and $(\mu_t)_{t\in[0,T]}$ under the sequence of probab. measures $R_n$, $n\geq 1$.
   - Taking a subsequence if necessary, using Skorokhod's representation theorem, $\exists$ a probab. space $(\Omega',\mathscr{F}',\mathbb{P}')$ and $(X_t^{(n)},\Lambda_t^{(n)},\mu_t^{(n)})_{t\in[0,T]}$ taking values in $\mathcal{Y}$ with the distribution $R_n$, such that

$$(X_t^{(n)},\Lambda_t^{(n)},\mu_t^{(n)})_{t\in[0,T]}\longrightarrow(X_t^{(0)},\Lambda_t^{(0)},\mu_t^{(0)})_{t\in[0,T]},\ \text{a.s.}\quad n\to\infty.$$

# Key steps to prove Theorem 1

2. Prove that $(X_t^{(0)}, \Lambda_t^{(0)}, \mu_t^{(0)})$ satisfies
   - $X_t^{(0)} = x + \int_0^t b(X_s^{(0)}, \Lambda_s^{(0)})\mathrm{d}s + \int_0^t \sigma(X_s^{(0)}, \Lambda_s^{(0)})\mathrm{d}B_s.$
   - $$\mathbb{P}(\Lambda_{t+\delta}^{(0)} = j | \Lambda_t^{(0)} = i', \mu_t^{(0)} = \mu) = \begin{cases} q_{i'j}(\mu)\delta + o(\delta) & i' \neq j, \\ 1 - q_{i'}(\mu)\delta + o(\delta), & i' = j. \end{cases}$$
   - $\mu_t^{(0)}$ is adapted to $\sigma(\Lambda_t^{(0)})$.
3. the control $\alpha^*$ associated with $(X_t^{(0)}, \Lambda_t^{(0)}, \mu_t^{(0)})$ is an optimal $\psi$-relaxed control.

# Contents

Let $\psi : [0,T] \to [0,\infty)$ be increasing, $\lim_{r\to 0}\psi(r) = 0$.
For $r_0 \in (0,T)$, define a shift operator $\theta_{r_0} : \mathcal{D}([0,T];\mathcal{S}) \to \mathcal{D}([0,T];\mathcal{S})$ by

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

Moreover, $\theta_{r_0}^k\lambda(t) := \lambda_{(t-kr_0)\vee 0}$ for $\lambda \in \mathcal{D}([0,T];\mathcal{S})$, $k \geq 0$.
$m \geq 1$ is a fixed integer. A functional $h : [0,T] \times \mathcal{S}^{m+1} \to \mathscr{P}(U)$ is said to be in the class $\Upsilon_\psi$ if for every $i_0,\ldots,i_m \in \mathcal{S}$, $t \mapsto \tilde{\mu}(t) := h(t,i_0,\ldots,i_m)$ satisfies

$$w_{\tilde{\mu}}([t_1,t_2)) \leq \psi(|t_2-t_1|), \quad t_1,\, t_2 \in [0,T].$$

**Definition:** For $s \in [0,T)$ and $i \in \mathcal{S}$, a *history-dependent control* is a term $\alpha = (\Lambda_t, \mu_t)$ such that

(1) $(\Lambda_t)$ is an $\mathscr{F}_t$-adapted jumping process satisfying

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

with initial value $\Lambda_s = i$ for $s \in [0,T)$ and $i \in \mathcal{S}$.

(2) There exists $h \in \Upsilon_\psi$ such that

$$\mu_t = h(t, \theta_{r_0}^0 \Lambda(t), \ldots, \theta_{r_0}^m \Lambda(t)).$$

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

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

and the value function is defined by

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

A history-dependent control $\alpha^* \in \Pi_{s,i}$ is said to be *optimal*, if

$$V(s, i) = J(s, i, \alpha^*).$$

1. $\mu_t = h(\Lambda_t)$ for some $h : \mathcal{S} \to \mathscr{P}(U)$. In this situation, $\alpha$ is corresponding to the stationary randomized Markov policy studied by many works.
2. $\mu_t = h(\Lambda_{(t-r_0)\vee 0})$ for some $h : \mathcal{S} \to \mathscr{P}(U)$. Now the control policies are purely determined by the jumping process with a positive delay. This kind of controls is very natural to be used in the realistic application.

**Assumptions**:

(A1) $\mu \mapsto q_{ij}(\mu)$ is continuous $\forall\, i,\, j \in \mathcal{S}$, and $M := \sup_{i\in\mathcal{S}} \sup_{\mu\in\mathscr{P}(U)} q_i(\mu) < \infty$.

(A2) There exists a compact function $\Phi : \mathcal{S} \to [1, \infty)$, a compact set $B_0 \subset \mathcal{S}$, contants $\lambda_0 > 0$ and $\kappa_0 \geq 0$ such that

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

(A3) There exists a $K \in \mathbb{N}$ such that for every $i \in \mathcal{S}$ and $\mu \in \mathscr{P}(U)$, $q_{ij}(\mu) = 0$, if $|j - i| > K$.

**Theorem 4**

Assume (A1)-(A3) hold. Then for every $s \in [0, T)$, $i \in \mathcal{S}$, there exists an optimal control $\alpha^* \in \Pi_{s,i}$.

- X.P. Guo, X.X. Huang, Y.H. Huang, Adv. Appl. Prob. 2015.

# Further work

Consider the following SDE:

$$\mathrm{d}X_t = b(X_t, \Lambda_t, \mu_t)\mathrm{d}t + \sigma(X_t, \Lambda_t, \mu_t)\mathrm{d}B_t,$$

where $b : \mathbb{R}^d \times \mathcal{S} \times \mathscr{P}(U) \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \times \mathcal{S} \times \mathscr{P}(U) \to \mathbb{R}^{d\times d}$, and $(B_t)_{t\geq 0}$ is a $d$-dimensional $\mathscr{F}_t$-Brownian motion. Here $(\Lambda_t)_{t\geq 0}$ is a continuous-time jumping process on $\mathcal{S}$ satisfying

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

provided $\delta > 0$ for every $x \in \mathbb{R}^d$, $\nu \in \mathscr{P}(U)$, $i,\ j \in \mathcal{S}$.

# Thank You For Your Attention !

EMAIL:shaojh@tju.edu.cn