Constrained continuous-time Markov decision processes on the finite horizon

Xianping Guo

Sun Yat-Sen University, Guangzhou
Email: mcsgxp@mail.sysu.edu.cn

16-19 July, 2018, Chengdou

Outline

- The optimal control problem
- Preliminary facts
- Occupation measures and their properties
- Characterization of constrained-optimal policies

1. The optimal control problem

- S: state space, a denumerable set;
- A: action space A, equipped with the Borel σ-algebra $\mathcal{B}(A)$;
- $A(t,i)(\in \mathcal{B}(A))$: sets of actions available to a controller when the system is in state $i \in S$ at time t;
- $q(j|t,i,a)$: Nonhomogeneous transition rates such that

$$q^*(i) := \sup_{t\geq 0, a\in A(t,i)} |q(t,i,a)| < \infty \quad \forall\, i \in S; \tag{1}$$

- $r(t,i,a)$ and $g(t,i)$: Reward and terminal reward,respectively;
- $c_k(t,i,a)$ and $g_k(t,i)$: Costs and terminal costs,$k = 1, \cdots, N$.

For each sample $\omega = (i_0, \theta_1, i_1, \ldots, \theta_n, i_n, \cdots)$, let

$$T_k(\omega) := \theta_1 + \theta_2 + \ldots + \theta_k, T_\infty(\omega) := \lim_{k\to\infty} T_k(\omega).$$

be the k-jump and explosion time,respectively, where θ_k denotes the holding time of state i_{k-1}.

Let $T_0(\omega) \equiv 0$, and define the state process $\{x_t, t \geq 0\}$ by

$$x_t := \sum_{k\geq 0} I_{\{T_k \leq t < T_{k+1}\}} i_k + \Delta I_{\{t \geq T_\infty\}}.$$

Here and below, I_E stands for the indicator function on E, and the Δ and a_Δ are cemetery state and action, respectively.

- Randomized history-dependent policies $\pi(da|\omega,t)$: is defined by the following expression with kernels $\pi^k(da|\cdot)$

$$\pi(da|\omega,t) = \sum_{k\geq 0} I_{\{T_k<t\leq T_{k+1}\}}\pi^k(da|i_0,\theta_1,\ldots,\theta_k,i_k,t-T_k) + I_{\{0\}}(t)\pi^0(da|i_0,0) + I_{\{t\geq T_\infty\}}\delta_{a_\Delta}(da).$$

- Π : The class of all randomized history-dependent policies.

- Π^m: The class of all Markov policies $\pi(da|t,i)$.

- f: a deterministic Markov policy f: A measurable map f on $[0,\infty)\times S$ with $f(t,i)\in A(t,i)$.

Given a initial distribution γ on S, each $\pi \in \Pi$ together with $q(j|t,i,a)$ ensures a unique probability space $(\Omega, \mathcal{F}, \mathbb{P}_\gamma^\pi)$.

Let $T \in (0, \infty)$ be the fixed finite (time) horizon.

For each policy $\pi \in \Pi$, we define

$$V(\pi, u, h) = \mathbb{E}_\gamma^\pi \left[\int_0^T \int_A u(t, x_t, a) \pi(da|\omega, t) dt + h(T, x_T) \right]$$

provided that the expectations are well defined.

Let d_k be the constrained constants, and then define

$$U := \{\pi \in \Pi : \; V(\pi, c_k, g_k) \leq d_k, \quad \text{for } k = 1, \ldots, N\}, \tag{2}$$

which denotes the set of policies satisfying the N constraints.

A policy $\pi \in \Pi$ is called feasible if it is in U. Throughout this article, to avoid trivial cases, we suppose that $U \neq \emptyset$, and this assumption will not be mentioned explicitly below.

Definition 1 A policy $\pi^* \in U$ is called constrained-optimal if

$$V(\pi^*, r, g) = \sup_{\pi \in U} V(\pi, r, g). \tag{3}$$

The main objective of this talk is to show the existence and structure of a constrained-optimal Markov policy.

2. Preliminary facts

In this section, we present some assumptions and preliminary facts that are used to prove our main results.

Assumption A. There exist a function $V_1 \geq 1$ on S and constants $c > 0$, $b \geq 0$, $M > 0$ such that

(i) $\sum_{j \in S} q(j|t,i,a)V(j) \leq cV(i) + b$, for all (t,i,a);

(ii) $q^*(i) \leq MV(i)$ for all $i \in S$, with $q^*(i)$ as in (1);

(iii) $|u(t,i,a)| \leq MV(i)$ for all $u \in \{r, g, c_k, g_k\}$ and (t,i,a).

(iv) $L := \sum_{i \in S} V(i)\gamma(i) < \infty$, where γ is the distribution.

Lemma 1. Under Assumption A, for each $\pi \in \Pi$, the following assertions hold.

(a) $\mathbb{E}_\gamma^\pi[V(x_t)] \leq e^{ct}[L + \frac{b}{c}]$ for each $t \geq 0$;

(b) $\mathbb{P}_\gamma^\pi(x_t = i) = \gamma(i) + \mathbb{E}_\gamma^\pi\big[\int_0^t \int_A q(i|s, x_{s-}, a)\pi(da|e, s)ds\big]$, for each $t \geq 0$ and $i \in S$;

(c) $\sum_{i \in S} \mathbb{P}_\gamma^\pi(x_t = i) = 1$, for each $t \geq 0$.

Lemma 1(b) gives the analog of the forward Kolmogorov equation, which will be used to derive the analog of the Ito-Dynkin formula for the process $\{x_t, t \geq 0\}$. To serve the analog, we introduce some additional conditions and notations.

Assumption B. There exist a function $V_1 \geq 1$ on S and constants $c_1 > 0$, $b_1 \geq 0$ and $M_1 > 0$ such that

(i) $\sum_{j\in S} V_1(j)q(j|t,i,a) \leq c_1 V_1(i) + b_1$, for all $(t,i,a) \in \mathbb{K}$;

(ii) $V(i)[1 + q^*(i)] \leq M_1 V_1(i)$, with $q^*(i)$ as in (1);

(iii) $L' := \sum_{i\in S} V_1(i)\gamma(i) < \infty$.

Let $I := [0, T]$. Given any function $\bar{w} \geq 1$ on S, a Borel measurable function φ on a Borel space $Z \times S$ is called $\bar{w}$-bounded if

$$\|\varphi\|_{\bar{w}} := \sup_{(z,i)\in Z\times S} \frac{|\varphi(z,i)|}{\bar{w}(i)} < \infty.$$

- $\mathbb{B}_{\bar{w}}(I \times S)$: the space of all $\bar{w}$-bounded functions on $I \times S$;
- $C_b(I \times S)$: the space of all bounded continuous functions.

If $\varphi(t,i)$ is absolutely continuous in $t \in I$, we denote by $\varphi_t(t,i)$: the derivative of $\varphi(t,i)$ with respect to t, and by $L_\varphi(i) \subseteq I$: the collection of points in I, when the $\varphi_t(t,i)$ is not defined.

With V and V_1 as in Assumption B, let

$$\mathbb{B}^{1,0}_{V,V_1}(I \times S) := \{\varphi \in \mathbb{B}_V(I \times S) : \varphi_t \in \mathbb{B}_{V+V_1}(I \times S).\}$$

On the other hand, for any Markov policy π and functions $u(\ldots,t,i,a)$, we use the following notation:

$$u(\ldots,t,i,\pi) := \int_A u(\ldots,t,a)\pi(da|t,i).$$

Lemma 2. Under Assumptions A and B, we have

(a) For each $\pi \in \Pi$ and $h \in B_1(I \times S)$,

$$\mathbb{E}_\gamma^\pi \left[\int_0^T \sum_{i \in S} \int_t^T \int_A h(s,i) q(i|t, x_t, a) \pi(da|\omega, t) ds dt \right]$$
$$= \mathbb{E}_\gamma^\pi \left[\int_0^T h(t, x_t) dt \right] - \sum_{i \in S} \left[\int_0^T h(t,i) dt \right] \gamma(i).$$

(b) For any $\pi \in \Pi^m$ and $1 \leq k \leq N$, $V(\pi, c_k, 0; t, i)$ is a the unique solution in $\mathbb{B}^{1,0}_{V,V_1}(I \times S)$ of the equation

$$\varphi_t(t,i) + c_k(t,i,\pi) + \sum_{j \in S} \varphi(t,j) q(j|t,i,\pi) = 0 \ \forall \ t \in L_\varphi^c(i)$$

with the condition $\varphi(T, i) = 0$ for each $i \in S$.

3. Occupation measures and properties

In this section, we introduce the occupation measure of a policy for the finite horizon CTMDP, and present some basic properties of the space of the occupation measures.

Definition 1. For each $\pi \in \Pi$, the occupation measure η^π of π on K, is defined by

$$\eta^\pi(dt, i, da) := \mathbb{E}_\gamma^\pi[I_{\{x_t = i\}} \pi(da|\omega, t)]dt, \qquad i \in S, \tag{4}$$

where

$$K := \{(t, i, a) : t \in [0, T],\ i \in S,\ a \in A(t, i)\}.$$

Let $c_0(t,i,a) := r(t,i,a)$, $g_0(t,i,a) := g(t,i,a)$, and

$$
\begin{aligned}
H_k(t,i,a) \;&:=\; c_k(t,i,a) + \sum_{j\in S} g_k(T,j)q(j|t,i,a) \\
&\quad + \frac{1}{T}\sum_{j\in S} g_k(T,j)\gamma(j), k = 0,1,\ldots,N.
\end{aligned}
$$

Lemma 3. Under Assumptions A and B, for each $\pi \in \Pi$ and $0 \le k \le N$, we have

$$
\begin{aligned}
V(\pi, H_k) \;&=\; \int_0^T \sum_{i\in S} \int_A H_k(t,i,a)\eta^\pi(dt,i,da) \\
&=: \; E^{\eta^\pi}[H_k]
\end{aligned}
$$

Then, our constrained optimality problem (3) can be rewrite as follows:

$$Maximize\ E^{\eta}[H_0] \text{ over } \eta \in \mathcal{D}_c,$$

where $\mathcal{D}_c := \{\eta^{\pi} | E^{\eta^{\pi}}[H_k] \leq d_k, k = 1, \ldots, N, \pi \in \Pi\}$.

- $\mathcal{D} := \{\eta^{\pi} : \ \pi \in \Pi\}$: the set of all occupation measures;
- $P(K)$: the collection of measures η on K with $\eta(K) = T$;
- $\bar{\eta}(dt, i)$: the marginal of η on $I \times S$;
- $P_{\bar{\omega}}(K) := \{\eta \in P(K) | \ \sum_{i \in S} \bar{\omega}(i) \bar{\eta}(I \times \{i\}) < \infty\}$.

The following theorem characterizes occupation measures.

Theorem 1. Under Assumptions A and B, we have

(a) For each $\eta \in P_V(K)$, it holds that $\eta \in \mathcal{D}$ if and only if

$$\int_0^T \sum_{i \in S} \int_{A(t,i)} \sum_{j \in S} \left(\int_t^T q(j|t,i,a) h(s,j) ds \right) \eta(dt,i,da)$$

$$= \int_0^T \sum_{j \in S} h(s,j) \bar{\eta}(ds,j) - \int_0^T h(s,\gamma) ds \; \forall \; h \in C_b(I \times S)$$

(b) For each $\pi \in \Pi$, there exists a Markov policy ϕ such that

$$\eta^\pi = \eta^\phi$$

(c) $\mathcal{D}$ is convex.

Definition 2. For each $\bar{w} \geq 1$ on S, the $\bar{\omega}$-weak topology on $P_{\bar{w}}(K)$ is defined as the weakest topology with respect to which, $\int_0^T \sum_{i\in S} \int_A u(t,i,a)\eta(dt,i,a)$ is continuous in $\eta \in P_{\bar{w}}(K)$ for each continuous function u on K such that $\sup_{(t,i,a)\in K} \frac{|u(t,i,a)|}{\bar{\omega}(i)} < \infty$.

Here and below, $P_{V+V_1}(K)$ and $P_V(K)$ are endowed with the $(V+V_1)$ and V -weak topologies, respectively.

Lemma 4. Under Assumptions A and B, if $q(j|t,i,a)$ is continuous in $(t,i,a) \in K$ for each fixed $j \in S$, then $\mathcal{D}$ is closed in $P_{V+V_1}(K)$ and in $P_V(K)$.

For the compactness of the set of occupation measures, we introduce the following condition.

Assumption C. Let V and V_1 be as in Assumption B.

(i) $q(j|t,i,a)$ are continuous in $(t,i,a) \in K$ (for fixed $j \in S$).

(ii) There exist compact subsets K_m of K satisfying $\bigcup_m K_m = K$ and $\lim_{m\to\infty} \inf_{(t,i,a)\in K\setminus K_m} \frac{V_1(i)}{V(i)} = \infty$, where $\inf \emptyset := \infty$.

Assumption C implies that each $A(t,i)$ is compact.

Theorem 2. Suppose that Assumptions A, B, and C hold. Then, $\mathcal{D}$ is compact in $P_V(K)$.

4. Characterization of optimal policies

This part establishes the existence and structure of a constrained-optimal policy.

Assumption D.

(a) $r(t,i,a), c_k(t,i,a)$ and $\sum_{j\in S} V(j)q(j|t,i,a)$ are continuous on K.

(b) Either $q^*(i)$ or $g_k(T,i)$ are bounded on S;

Theorem 3. Under Assumptions A, B, C and D, there exists a Markov constrained-optimal policy.

Under the assumptions, we define the space of performance vectors for the model with the criteria:

$$\mathcal{U} := \{(V(\pi, r, g), V(\pi, c_1, g_1), \ldots, V(\pi, c_N, g_N)) \mid \pi \in \Pi\}.$$

Definition 3. A policy $\pi \in \Pi$ is said to be a mixture of $N+1$ deterministic Markov policies $f_k, k = 0, 1, 2, \ldots, N$, if

$$\eta^{\pi}(dt, i, da) = \sum_{k=0}^{N} p_k \eta^{f_k}(dt, i, da),$$

where $p_k \geq 0$ for all $0 \leq k \leq N$, and $p_0 + p_1 + \cdots + p_N = 1$.

We next give our main statement.

Theorem 4. Under Assumptions A–D, the following assertions hold:

(a) The space of performance vectors, $\mathcal{U}$, is nonempty, compact and convex.

(b) Any extreme point of $\mathcal{U}$ (there exists at least one), say v^{ex}, is generated by a deterministic Markov policy, say f, i.e., $v^{ex} = (V(f, r, g), V(f, c_1, g_1), \ldots, V(f, c_N, g_N))$.

(c) There exists a constrained-optimal policy, which is a $(N+1)$-mixture of deterministic Markov policies.

Many Thanks !!!