13th Workshop on Markov Processes and Related Topics

A probability criterion for zero-sum stochastic games

Xianping Guo

Sun Yat-Sen University, Guangzhou

Email: mcsgxp@mail.sysu.edu.cn

17-21 July 2017, Wuhan

Outline

- Models of the games
- The game problem
- The main results
- An applied example

1. The model of stochastic games

$$\{S, (A(i) \subset A, B(i) \subset B, i \in S), Q(\cdot|i,a,b), r(i,a,b)\}$$

- S: State space, a denumerable set space S;
- $A(i)/B(i)$: Finite sets of actions available at $i \in S$ for player 1/player 2 respectively;
- $Q(j|i,a,b)$: Transition probability from i to j at under the pair of actions $(a,b) \in A(i) \times B(i)$;
- $r(i,a,b)$: Nonnegative reward/cost function for player 1 (i.e., the cost for player 2) under actions (a,b) at state i.

The evolution of the game: When the system state is i_0 at the initial decision epoch 0, there is a common level λ_0 in the two players' mind, that is, player 1 tries his or her best to get rewards more than λ_0, while player 2 will try to control the cost no more than λ_0). Then, the players independently of each other choose actions $a_0 \in A(i_0)$ and $b_0 \in B(i_0)$. Consequently, the system jumps to state $i_1 \in S$ with one-step transition probability $Q(i_1|i_0, a_0, b_0)$ at time 1, and a payoff $r(i_0, a_0, b_0)$ is generated, and thus there remains a level $\lambda_1 := \lambda_0 - r(i_0, a_0, b_0)$ for both players.

Based on the current state i_1, level λ_1, state i_0, and the previous level λ_0, the players chose their actions..... The game is developed in this way, and so we get an admissible history h_n of the game up to the nth decision epoch, i.e.,

$$h_n = (i_0, \lambda_0, a_0, b_0, \ldots, i_{n-1}, \lambda_{n-1}, a_{n-1}, b_{n-1}, i_n, \lambda_n),$$

where $(i_m, a_m, b_m) \in K := \{(i, a, b) | i \in S, a \in A(i), b \in B(i)\}$, $\lambda_0 \in \mathbb{R} := (-\infty, +\infty)$, and $\lambda_{m+1} := \lambda_m - r(i_m, a_m, b_m)$

For convenience, we denote by H_n the set of all admissible histories h_n of the system up to the nth decision epoch, and assume that H_n is endowed with a Borel σ-algebra.

Φ_1: Set of all stochastic kernels φ satisfying $\varphi(A(i)|i,\lambda)=1$

History-dependent policy (for player 1): $\pi^1=\{\pi_n^1, n\geq 0\}$ of stochastic kernels π_n^1 such that $\pi_n^1(A(i_n)|h_n)\equiv 1$

Markov policy: $\pi_n^1(\cdot|h_n)=\varphi_n(\cdot|x_n,\lambda_n)\in\Phi_1$ for all $n\geq 0$.

Stationary policy: $\pi_n^1(\cdot|h_n)\equiv\varphi(\cdot|x_n,\lambda_n)$ for some $\varphi\in\Phi_1$.

We write such a policy $\{\pi_n^1, n\geq 0\}$ as φ.

$\Pi_1/\Pi_1^m/\Pi_1^s$: the sets of all history-dependent/Markov/stationary policies for player 1, respectively.

Similarly, $\Pi_2/\Pi_2^m/\Pi_2^s$: the sets of all history-dependent/stationary policies for player 2, respectively.

Given $(i,\lambda) \in S \times \mathbb{R}$ and $(\pi^1,\pi^2) \in \Pi_1 \times \Pi_2$, the Tulcea's theorem gives the existence of a unique probability space $(\Omega, \mathcal{F}, P_{(i,\lambda)}^{\pi^1,\pi^2})$ and a process $\{i_n, \lambda_n, a_n, b_n\}$ such that

$$P_{(i,\lambda)}^{\pi^1,\pi^2}(a_n = a, b_n = b|h_n) = \pi_n^1(a|h_n)\pi_n^2(b|h_n),$$
$$P_{(i,\lambda)}^{\pi^1,\pi^2}(i_{n+1} = j|h_n, a_n, b_n) = Q(j|i_n, a_n, b_n),$$

for each $j \in S$, $a \in A(i)$, $b \in B(i)$ and $h_n \in H_n$ with $n \geq 0$.

$E_{(i,\lambda)}^{\pi^1,\pi^2}$: the expectation operator associated with $P_{(i,\lambda)}^{\pi^1,\pi^2}$.

For the target set D, let

$$\tau_D := \begin{cases} \inf\{n \geq 0 : i_n \in D\} & \text{if } \{n \geq 0 : i_n \in D\} \neq \emptyset, \\ +\infty & \text{otherwise.} \end{cases}$$

2. The game problem

Probability criterion: For each $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$,

$$G(i, \lambda, \pi^1, \pi^2) := P_{(i,\lambda)}^{\pi^1,\pi^2} \left(\sum_{n=0}^{\tau_D - 1} r(i_n, a_n, b_n) > \lambda \right),$$

which gives capacity for player 1 to reach the profit level λ, and also measures the risk of player 2 to control the cost level λ.

The corresponding standard expectation criterion:

$$V(i, \cdot, \pi^1, \pi^2) := E_{(i,\cdot)}^{\pi^1,\pi^2} \left(\sum_{n=0}^{\tau_D - 1} r(i_n, a_n, b_n) \right).$$

The functions

$$L(i,\lambda) := \sup_{\pi^1\in\Pi_1}\inf_{\pi^2\in\Pi_2} G(i,\lambda,\pi^1,\pi^2),$$
$$U(i,\lambda) := \inf_{\pi^2\in\Pi_2}\sup_{\pi^1\in\Pi_1} G(i,\lambda,\pi^1,\pi^2)$$

are called the lower value and the upper value of the game, respectively. Clearly,

$$L(i,\lambda) \le U(i,\lambda)$$

Definition 1: If $L(i,\lambda) \equiv U(i,\lambda)$, then we call the common function the value of the game, which is denoted by V.

Definition 2: Suppose that the value of the game V exists. A policy $\pi^{*1} \in \Pi_1$ is said to be optimal for player 1 if

$$\inf_{\pi^2 \in \Pi_2} G(i, \lambda, \pi^{*1}, \pi^2) = V(i, \lambda)$$

Similarly, $\pi^{*2} \in \Pi_2$ is called optimal for player 2 if

$$\sup_{\pi^1 \in \Pi_1} G(i, \lambda, \pi^1, \pi^{*2}) = V(i, \lambda)$$

If $\pi^{*k} \in \Pi_k$ is optimal for player $k (k = 1, 2)$, then (π^{*1}, π^{*2}) is called a pair of optimal policies (also known as a saddle point).

Main goals: The existence and computation of a saddle point.

3. The main results

$\mathcal{P}(U)$: The set of all probability measures on the set U, endowed with the weak topology.

$\mathcal{F}_m$: The set of functions $h : D^c \times \mathbb{R} \to [0, 1]$, such that $h(i, \cdot)$ is Borel-measurable on $\mathbb{R}$ for each $i \in D^c$ and $h(i, \lambda) = 1$ if $\lambda < 0$ for each $i \in D^c$.

$T^{\varphi,\phi}$, T, T^{π^1,π^2}: The operators on $\mathcal{F}_m$ are defined as follows:

For any $h \in \mathcal{F}_m$, $i \in D^c$, $\varphi \in \mathcal{P}(A(i))$, $\phi \in \mathcal{P}(B(i))$ and $(\pi^1, \pi^2) \in \Pi_1^s \times \Pi_2^s$, if $\lambda \geq 0$,

$$T^{\varphi,\phi}h(i,\lambda) := \sum_{a\in A(i)} \sum_{b\in B(i)} \varphi(a)\phi(b)[I_{\{\lambda<r(i,a,b)\}}Q(D|i,a,b) + \sum_{j\in D^c} h(j,\lambda - r(i,a,b))Q(j|i,a,b)],$$

$$Th(i,\lambda) := \sup_{\varphi\in\mathcal{P}(A(i))} \inf_{\phi\in\mathcal{P}(B(i))} T^{\varphi,\phi}h(i,\lambda), \tag{1}$$

$$T^{\pi^1,\pi^2}h(i,\lambda) := T^{\pi^1(\cdot|i,\lambda),\pi^2(\cdot|i,\lambda)}h(i,\lambda),$$

with $T^{\varphi,\phi}h(i,\lambda) = Th(i,\lambda) = T^{\pi^1,\pi^2}h(i,\lambda) := 1$ for $\lambda < 0$.

In order to calculate $G(i,\lambda,\pi^1,\pi^2)$, we rewrite

$$\begin{aligned} G(i,\lambda,\pi^1,\pi^2) &= P_{(i,\lambda)}^{\pi^1,\pi^2}\left(\sum_{n=0}^{\tau_D-1} r(i_n,a_n,b_n) > \lambda\right) \\ &= P_{(i,\lambda)}^{\pi^1,\pi^2}\left(\sum_{n=0}^{\infty} I_{\{\bigcap_{k=0}^{n}\{i_k\in D^c\}\}} r(i_n,a_n,b_n) > \lambda\right) \\ &= \lim_{n\to\infty} G_n(i,\lambda,\pi^1,\pi^2) \end{aligned}$$

where

$$G_n(i,\lambda,\pi^1,\pi^2) := P_{(i,\lambda)}^{\pi^1,\pi^2}\left(\sum_{m=0}^{n} I_{\{\bigcap_{k=0}^{m}\{i_k\in D^c\}\}} r(i_m,a_m,b_m) > \lambda\right)$$

Obviously, $G_n(\cdot,\cdot,\pi^1,\pi^2) \le G_{n+1}(\cdot,\cdot,\pi^1,\pi^2)$ (by $r \ge 0$).

Lemma 1: Given any $\pi^1 = \{\varphi_n, n \geq 0\} \in \Pi_1^m$, $\pi^2 = \{\phi_n, n \geq 0\} \in \Pi_2^m$, define

$${}^{(1)}\pi^1 := \{\varphi_n, n \geq 1\} \in \Pi_1^m, \ {}^{(1)}\pi^2 := \{\phi_n, , n \geq 1\} \in \Pi_2^m.$$

Then, for each $n \geq 0$, we have

(a) $G_n(\cdot, \cdot, \pi^1, \pi^2) \in \mathcal{F}_m$ and $G(\cdot, \cdot, \pi^1, \pi^2) \in \mathcal{F}_m$;

(b) $G_{n+1}(\cdot, \cdot, \pi^1, \pi^2) = T^{\varphi_0, \phi_0} G_n(\cdot, \cdot, {}^{(1)}\pi^1, {}^{(1)}\pi^2)$;

(c) $G(\cdot, \cdot, \pi^1, \pi^2) = T^{\varphi_0, \phi_0} G(\cdot, \cdot, {}^{(1)}\pi^1, {}^{(1)}\pi^2)$;

(d) $G(\cdot, \cdot, \varphi, \phi) = T^{\varphi, \phi} G(\cdot, \cdot, \varphi, \phi)$ for every $(\varphi, \phi) \in \Pi_1^s \times \Pi_2^s$.

To further show the uniqueness of the solution to the equation $h = T^{\varphi,\phi}h$, we need the following assumption.

Assumption 1. $P^{\pi^1,\pi^2}_{(i,\lambda)}(\tau_D < \infty) \equiv 1$

Assumption 1 indicates that, no matter what the initial state is, what the level is, and what the pair of randomized Markov policies is, the system will fail within finite time.

To verify Assumption 1, it is desired to give a sufficient condition imposed on the *primitive* data of the game model.

Lemma 2. If $\inf_{(i,a,b)\in D^c\times A(i)\times B(i)} Q(D|i,a,b) > 0$, then Assumption 1 holds.

Lemma 3. Under Assumption 1, for any function u in $\mathcal{F}_m$ the following statements hold.

(a) If $u(i,\lambda) \leq T^{\pi^1,\phi_k}u(i,\lambda)$ for all $k \geq 0$, $\pi^1 \in \Pi_1^s$, and $\overline{\pi}^2 = \{\phi_k, k \geq 0\} \in \Pi_2^m$, then $u(i,\lambda) \leq G(i,\lambda,\pi^1,\overline{\pi}^2)$.

(b) If $u(i,\lambda) \geq T^{\varphi_k,\pi^2}u(i,\lambda)$ for all $k \geq 0$, policies $\pi^2 \in \Pi_2^s$, and $\overline{\pi}^1 = \{\varphi_k, k \geq 0\} \in \Pi_1^m$, then $u(i,\lambda) \geq G(i,\lambda,\overline{\pi}^1,\pi^2)$;

(c) For every $(\pi^1,\pi^2) \in \Pi_1^s \times \Pi_2^s$, $G(\cdot,\cdot,\pi^1,\pi^2)$ is the unique solution in $\mathcal{F}_m$ to the equation $h = T^{\pi^1,\pi^2}h$.

Let $u_{-1}(i,\lambda) := I_{(-\infty,0)}(\lambda)$, and for $n \geq 0$, define u_n by

$$u_n(i,\lambda) := Tu_{n-1}(i,\lambda)$$

$$= \sup_{\varphi \in \mathcal{P}(A(i))} \inf_{\phi \in \mathcal{P}(B(i))} \{ \sum_{a \in A(i)} \sum_{b \in B(i)} \varphi(a)\phi(b)[I_{\{\lambda < r(i,a,b)\}} Q(D|i,a,b)$$

$$+ \sum_{j \in D^c} u^*(j, \lambda - r(i,a,b)) Q(j|i,a,b)] \}$$

for $(i,\lambda) \in D^c \times \mathbb{R}$.

Now, we state the main result on the existence of a pair of optimal policies.

Theorem 1. Under Assumption 1, we have the assertions:

(a) The $\lim_{n\to\infty} u_n(i,\lambda) =: u^*(i,\lambda)$ exists and belongs to $\mathcal{F}_m$;

(b) u^* satisfies the Shapley's equation $u^*(i,\lambda) = Tu^*(i,\lambda)$;

(c) There exists a pair of stationary policies $(\pi_1^*, \pi_2^*) \in \Pi_1^s \times \Pi_2^s$ such that, for all $(i,\lambda) \in D^c \times \mathbb{R}$,

$$T^{\pi_1^*,\pi_2^*}u^*(i,\lambda) = \max_{\varphi\in\mathcal{P}(A(i))} T^{\varphi,\pi_2^*}u^*(i,\lambda) = \min_{\phi\in\mathcal{P}(B(i))} T^{\pi_1^*,\phi}u^*(i,\lambda)$$

(d) $u^*(i,\lambda)$ is the value of the game, and $u^*(i,\lambda) = G(i,\lambda,\pi_1^*,\pi_2^*)$;

(e) (π_1^*, π_2^*) in (c) above is a pair of optimal stationary policies.

4. An example

Example 1 (An inventory system with capacity M):

i_n: stock amount at the beginning of period $n = 0, 1, 2, \ldots$

a_n: order amount from an finite set $A(i_n)$,

b_n: supply amount from an finite set $B(i_n)$,

z_n: amount of the product's demand during the period n,

$r(i_n, a_n, b_n)$: a reward function

Thus, the stock level evolves as follows

$$i_{n+1} := \min\{(i_n + \min\{a_n, b_n\} - z_n)^+, M\}, \ n = 0, 1, \ldots$$

$\{z_n\}$ is re assumed to be i.i.d. with a distribution $P(z_n = k) =: p_k$ and independent of the stock level.

Then, the state space is $S := \{0, \cdots, M\}$, and the transition law is the following: for any $j \in S$,

$$Q(j|i,a,b) := \sum_{k=0}^{+\infty} I_{\{j\}}[\min\{(i + \min\{a,b\} - k)^+, M)\}]p_k$$

Let $D := \{0\}$, this means that the game plays only there is at least one stock amount.

To ensure the existence of a pair of optimal policies for this inventory system, we impose the following hypothesis:

(C_1) For some k_0 such that $k_0 \geq M + \min\{\|a\|, \|b\|\}$, $p_{k_0} > 0$,

where $\|a\| := \max_{a \in A(i)} a$ and $\|b\| := \max_{b \in B(i)} b$.

Proposition 1. Under the hypothesis C_1, there exists a pair of optimal policies for the inventory system above.

Many thanks for your attention !