Two-person zero-sum risk-sensitive stochastic games with incomplete reward information on one side

Fang Chen

co-operate with: Xianping Guo

July 2023

Outline

Risk-sensitive criterion

- The risk preferences of players are taken into consideration by the expectation of the exponential utility of the total reward.
- References on discrete-time stochastic games (DTSGs):
  1) Basu, A. and Ghosh, M. K. (2014) Zero-sum risk-sensitive stochastic games on a countable state space. *Stochastic Process. Appl.*
  2) Bäuerle, N. and Rieder, U. (2017) Zero-sum risk-sensitive stochastic games. *Stochastic Process. Appl.*
  3) Ghosh, M. K., Golui, S., Pal, C. and Pradhan, S. (2023) Discrete-time zero-sum games for Markov chains with risk-sensitive average cost criterion. *Stochastic Process. Appl.*

complete information VS incomplete information

- The existing literature on risk-sensitive DTSGs considers complete information games.
- Complete information game: players do not have private information, which is known only to themselves and not to other players.
- Incomplete information games: players may have private information.

Game Model

A risk-sensitive stochastic game with incomplete reward information:

$$G(\theta, r, p) := \{\theta, E, A, B, q(y|x, a, b), K, p, r(k, x, a, b)\}$$

- $\theta \in (0, \infty)$: risk-sensitive parameter;
- E: finite state space;
- A/B: finite action space for player 1/player 2;
- $q(y|x, a, b)$: the transition probability to the state y from the state x under the action pair (a, b);
- K: finite reward type set used to describe the reward information;
- $p = \{p_k, k \in K\} \in \mathcal{P}(K)$: law of the reward types;
- $r(k, x, a, b)$: reward function; assume that r is nonnegative.

The evolution of the game $G(\theta, r, p)$

- Initially, k is chosen on K with the probability p_k. It is informed only to player 1.
- Both players observe the initial state x_0. Player 1 chooses a_0 according to the information k and x_0, whereas player 2 chooses b_0 only according to x_0.
- The system jumps to state x_1 with probability $q(x_1|x_0, a_0, b_0)$.
- At the stage n, player 1 chooses a_n according to k and the history h_n, whereas player 2 chooses b_n according only to h_n.
- Finally, given a discount factor $\beta \in (0, 1)$, player 1 receives the reward $\sum_{n=0}^{\infty} \beta^n r(k, x_n, a_n, b_n)$, which is paid by player 2.

♠ incomplete information on one side: only player 1 has private information

Policy

Definition 1

(a) *A randomized policy for player 1 is a sequence* $\pi = \{\pi_n^{(k)}, k \in K, n \geq 0\}$ *of stochastic kernels* $\pi_n^{(k)}$ *on* A *given* H_n, *where* $H_n := E \times (A \times B \times E)^n$.

(b) *A randomized policy for player 2 is a sequence* $\sigma = \{\sigma_n, n \geq 0\}$ *of stochastic kernels* σ_n *on* B *given* H_n.

(c) *Denote by* Π_i *the set of all randomized policies a for player* i $(i = 1, 2)$

For any $\pi = \{\pi_n^{(k)}, k \in K, n \geq 0\}$ and $h_n \in H_n$, denote

$$\pi_n(\cdot|h_n) := \{\pi_n^{(k)}(\cdot|h_n), k \in K\}.$$

Clearly, $\pi_n(\cdot|h_n) \in \mathcal{P}(A|K)$.

Risk-sensitive reward

Given $(\theta, r, p, x) \in (0, \infty) \times \mathcal{C} \times \mathcal{P}(K) \times E$, the expected risk-sensitive reward for player 1 under the policy pair $(\pi, \sigma) \in \Pi_1 \times \Pi_2$ is

$$V(\theta, r, p, x, \pi, \sigma) := \mathbb{E}_{p,x}^{\pi,\sigma}\left[e^{\theta \sum_{n=0}^{\infty} \beta^n r(\Lambda, X_n, A_n, B_n)}\right] \quad (1)$$

- $\mathcal{C}$: the family of all non-negative functions on $K \times E \times A \times B$
- $\mathbb{E}_{p,x}^{\pi,\sigma}$ is the expectation with respect to $\mathbb{P}_{p,x}^{\pi,\sigma}$ on
$$(\Omega, \mathcal{F}) := (K \times (E \times A \times B)^\infty, \mathcal{B}(K \times (E \times A \times B)^\infty));$$
- Λ, X_n, A_n, and B_n are random variables on $(\Omega, \mathcal{F})$ defined by
$$\Lambda(\omega) := k, \quad X_n(\omega) := x_n, \quad A_n(\omega) := a_n, \quad B_n(\omega) := b_n,$$
for each $n \geq 0$ and $\omega = (k, x_0, a_0, b_0, \ldots, x_n, a_n, b_n, \ldots) \in \Omega$.

♠ Λ is the reward information variable.

Value function

- Upper value function:

$$\overline{V}(\theta, r, p, x) := \inf_{\sigma \in \Pi_2} \sup_{\pi \in \Pi_1} V(\theta, r, p, x, \pi, \sigma)$$

- Lower value function:

$$\underline{V}(\theta, r, p, x) := \sup_{\pi \in \Pi_1} \inf_{\sigma \in \Pi_2} V(\theta, r, p, x, \pi, \sigma)$$

- Value function: for $(\theta, r, p) \in (0, \infty) \times \mathcal{C} \times \mathcal{P}(K)$, if

$$\underline{V}(\theta, r, p, x) = \overline{V}(\theta, r, p, x) \quad \forall x \in E,$$

the common function is called the value function of $G(\theta, r, p)$ and is denoted by $V^*(\theta, r, p, x)$.

Optimal policies

Definition 2

(a) *A policy* $\pi^* \in \Pi_1$ *for player 1 is called optimal in* $G(\theta, r, p)$ *if*

$$\inf_{\sigma \in \Pi_2} V(\theta, r, p, x, \pi^*, \sigma) = \underline{V}(\theta, r, p, x) \quad \forall x \in E.$$

(b) *Symmetrically, a policy* $\sigma^* \in \Pi_2$ *for player 2 is called optimal in* $G(\theta, r, p)$ *if*

$$\sup_{\pi \in \Pi_1} V(\theta, r, p, x, \pi, \sigma^*) = \overline{V}(\theta, r, p, x) \quad \forall x \in E.$$

♠ Our goals: proving the existence of the value function and constructing optimal policies for players.

In complete information games

- Complete information games: the existence of the value function and optimal policies are proved at the same time by the Shapley equation

$$u(\theta, x) = \sup_{\mu \in \mathcal{P}(A)} \inf_{\nu \in \mathcal{P}(B)} \sum_{a,b,y} \mu(a)\nu(b)q(y|x,a,b)e^{\theta r(x,a,b)} u(\theta\beta, y); \quad (2)$$

- the property (P1) is key :

$$(P1): \quad \mathbb{E}[e^{\theta \sum_{n=0}^{\infty} \beta^n r(X_n,A_n,B_n)} | X_0, A_0, B_0] = e^{\theta r(X_0,A_0,B_0)} \mathbb{E}[e^{\theta \sum_{n=1}^{\infty} \beta^n r(X_n,A_n,B_n)} | X_0, A_0, B_0].$$

- (P1) does not hold in incomplete information case; (2) is not suitable:

$$\mathbb{E}[e^{\theta \sum_{n=0}^{\infty} \beta^n r(\Lambda,X_n,A_n,B_n)} | X_0, A_0, B_0] \neq e^{\theta r(\Lambda,X_0,A_0,B_0)} \mathbb{E}[e^{\theta \sum_{n=1}^{\infty} \beta^n r(\Lambda,X_n,A_n,B_n)} | X_0, A_0, B_0];$$

Scheme for solving incomplete information games

- Establish the existence of the value function
- Derive a new Shapley equation by introducing a functional of rewards
- Show that the value function solves the Shapley equation
- Construct an optimal policy for player 1
- Construct an optimal policy for player 2

The value function

Theorem 1 (The existence of the value function)

(a) *For each* $N \geq 0$, *and* $(\theta, r, p, x) \in (0, \infty) \times \mathcal{C} \times \mathcal{P}(K) \times E$.

$$\begin{aligned} & \inf_{\sigma \in \Pi_2} \sup_{\pi \in \Pi_1} \mathbb{E}_{p,x}^{\pi,\sigma} \left[e^{\theta \sum_{n=0}^{N} \beta^n r(\Lambda, X_n, A_n, B_n)} \right] \\ = & \sup_{\pi \in \Pi_1} \inf_{\sigma \in \Pi_2} \mathbb{E}_{p,x}^{\pi,\sigma} \left[e^{\theta \sum_{n=0}^{N} \beta^n r(\Lambda, X_n, A_n, B_n)} \right] \\ =: & V_N^*(\theta, r, p, x). \end{aligned}$$

(b) *The value function* V^* *exists and satisfies*

$$V^* = \lim_{N \to \infty} V_N^*.$$

Key points of proof:

- The finiteness assumption ensures that the N-horizon game can be transformed into a static game where an action for player 1 is

$$d_1 : K \times \cup_{n=0}^{N} H_n \to A,$$

and an action for player 2 is $d_2 : \cup_{n=0}^{N} H_n \to B$. Both action spaces in the static game are finite, thus (a) holds.

- The existence of the value function directly follows from (a) and

$$0 \leq V(\theta, p, r, x, \pi, \sigma) - V_N(\theta, r, p, x, \pi, \sigma) \leq e^{\frac{\theta||r||}{1-\beta}}(e^{\frac{\theta||r||\beta^{N+1}}{1-\beta}} - 1).$$

Shapley equation

Q1: The property (P1) does not hold:

$$\mathbb{E}_{p,x}^{\pi,\sigma}[e^{\theta\sum_{n=0}^{\infty}\beta^n r(\Lambda,X_n,A_n,B_n)}|X_0,A_0,B_0]$$
$$\neq e^{\theta r(\Lambda,X_0,A_0,B_0)}\mathbb{E}_{p,x}^{\pi,\sigma}[e^{\theta\sum_{n=1}^{\infty}\beta^n r(\Lambda,X_n,A_n,B_n)}|X_0,A_0,B_0];$$

- Given any $(x,a,b)\in E\times A\times B$, define $\mathbb{H}^{x,a,b}$ from $\mathcal{C}$ to $\mathcal{C}$ as

$$\mathbb{H}^{x,a,b}(r)(\hat{k},\hat{x},\hat{a},\hat{b}) := r(k,\hat{x},\hat{a},\hat{b}) + \beta^{-1}(1-\beta)r(\hat{k},x,a,b), \tag{3}$$

for all $(\hat{k},\hat{x},\hat{a},\hat{b})\in K\times E\times A\times B$. $\mathbb{H}^{x,a,b}(r)\in\mathcal{C}$

Q2: How do players update the probability distribution p over K?

- A mapping $Q:\mathcal{P}(A|K)\times A\times\mathcal{P}(K)\to\mathcal{P}(K)$ is defined as

$$Q_k(\mu,a,p) := \frac{\mu(a|k)p_k}{\sum_{l\in K}\mu(a|l)p_l} \tag{4}$$

for all $\mu\in\mathcal{P}(A|K), a\in A, p\in\mathcal{P}(K), k\in K$. $Q(\mu,a,p)\in\mathcal{P}(K)$

Shapley equation

- For $(\mu,\nu) \in \mathcal{P}(A|K) \times \mathcal{P}(B)$, define an operator from $\mathbb{M}$ to $\mathbb{M}$ as

$$T^{\mu,\nu}u(\theta,r,p,x) := \sum_{k,a,b,y} p_k \mu(a|k)\nu(b)q(y|x,a,b) \\ \times u(\theta\beta, \mathbb{H}^{x,a,b}(r), Q(\mu,a,p), y)$$

  where $\mathbb{M}$ denotes the set of all nonnegative real-valued functions u on $(0,\infty) \times \mathcal{C} \times \mathcal{P}(K) \times E$.

- Define an operator T from $\mathbb{M}$ to $\mathbb{M}$ by

$$Tu := \sup_{\mu\in\mathcal{P}(A|K)} \inf_{\nu\in\mathcal{P}(B)} T^{\mu,\nu}u, \quad u \in \mathbb{M}. \tag{5}$$

- Then, we derive a new Shapley equation $u = Tu$.

The Shapley equation

Theorem 2

The value function V^* solves the Shapley equation, i.e.,

$$V^*(\theta, r, p, x) = \sup_{\mu \in \mathcal{P}(A|K)} \inf_{\nu \in \mathcal{P}(B)} T^{\mu,\nu} V^*(\theta, r, p, x)$$

for all $(\theta, r, p, x) \in (0, \infty) \times \mathcal{C} \times \mathcal{P}(K) \times E$.

Remark: the case of $K = \{k\}$.

Since $\mathcal{P}(K) = \{1\}$, we skip the third component of the value and obtain $V^*(\theta\beta, \mathbb{H}^{x,a,b}(r), y) = e^{\theta r(k,x,a,b)} V^*(\theta\beta, r, y)$. Hence, Theorem 2 gives

$$V^*(\theta, r, x) = \sup_{\mu \in \mathcal{P}(A)} \inf_{\nu \in \mathcal{P}(B)} \sum_{a,b,y} \mu(a)\nu(b)q(y|x, a, b)e^{\theta r(k,x,a,b)} V^*(\theta\beta, r, y),$$

which is consistent with the case of complete information.

Key points of proof:

- For $\pi = \{\pi_n^{(k)}, k \in K, n \geq 0\}$ and $\sigma = \{\sigma_n, n \geq 0\}$

$$V(\theta, r, p, x, \pi, \sigma) = \sum_{k \in K} \sum_{a \in A} \sum_{b \in B} p_k \pi_0^{(k)}(a|x)\sigma_0(b|x) \sum_{y \in E} q(y|x, a, b) \\ \cdot V(\theta\beta, \mathbb{H}^{x,a,b}(r), Q(\pi_0(\cdot|x), a, p), y, {}^{(x,a,b)}\pi, {}^{(x,a,b)}\sigma). \quad (6)$$

where ${}^{(x,a,b)}\pi = \{{}^{(x,a,b)}\pi_n^{(k)}, k \in K, n \geq 0\}$ is defined by

$${}^{(x,a,b)}\pi_n^{(k)}(\cdot|h_n) = \pi_{n+1}^{(k)}(\cdot|x, a, b, h_n), \quad k \in K, n \geq 0, h_n \in H_n.$$

${}^{(x,a,b)}\sigma$, one-shift policy of σ, is similarly defined.

- At 1-th decision epoch, if the history $h_1 = (x, a, b, y)$ is observed and the action a is chosen according to $\pi_0(\cdot|x)$, players can consider the problem as a new game $G(\theta\beta, \mathbb{H}^{x,a,b}(r), Q(\pi_0(\cdot|x), a, p))$ with the initial state y.

Construct an optimal policy for player 1:

Theorem 2

$$V^*(\theta, r, p, x) = \sup_{\mu \in \mathcal{P}(A|K)} \inf_{\nu \in \mathcal{P}(B)} T^{\mu,\nu} V^*(\theta, r, p, x)$$
$$= \sup_{\mu \in \mathcal{P}(A|K)} \inf_{\nu \in \mathcal{P}(B)} \sum_{k,a,b,y} p_k \mu(a|k)\nu(b) q(y|x,a,b)$$
$$\times V^*(\theta\beta, \mathbb{H}^{x,a,b}(r), Q(\mu, a, p), y)$$

- An optimal policy $\pi^* = \{\pi_n^{*(k)}, k \in K, n \geq 0\}$ for player 1 in $G(\theta, r, p)$ should satisfy the following two characterizations:

(C1): $V^*(\theta, r, p, x) = \inf_{\nu \in \mathcal{P}(B)} T^{\pi_0^*(\cdot|x),\nu} V^*(\theta, r, p, x)$ for all $x \in E$;

(C2): The one-shift policy ${}^{(x_0,a_0,b_0)}\pi^*$ is optimal in the new game

$$G(\theta\beta, \mathbb{H}^{x_0,a_0,b_0}(r), Q(\pi_0^*(\cdot|x_0), a_0, p)).$$

Construct an optimal policy for player 1:

- For a given triple $(\theta, r, p) \in (0,\infty) \times \mathcal{C} \times \mathcal{P}(K)$, we recursively define

$$\{(p^*[h_n], \pi_n^*(\cdot|h_n)) \in \mathcal{P}(K) \times \mathcal{P}(A|K), n \geq 0, h_n \in H_n\}.$$

- For each $h_0 = x_0 \in H_0$ let $p^*[h_0] = p$, $r[h_0] := r$ and

$$\pi_0^*(\cdot|h_0) = \arg\max_{\mu \in \mathcal{P}(A|K)} \{ \inf_{\nu \in \mathcal{P}(B)} T^{\mu,\nu} V^*(\theta, r[h_0], p^*[h_0], x_0)\}.$$

- For $h_{n+1} = (x_0, a_0, b_0, \ldots, x_n, a_n, b_n, x_{n+1}) = (h_n, a_n, b_n, x_{n+1}) \in H_{n+1}$ let

$$p^*[h_{n+1}] = Q(\pi_n^*(\cdot|h_n), a_n, p^*[h_n]), \quad r[h_{n+1}] = \mathbb{H}^{x_n,a_n,b_n}(r[h_n])$$
$$\pi_{n+1}^*(\cdot|h_{n+1}) = \arg\max_{\mu \in \mathcal{P}(A|K)} \{ \inf_{\nu \in \mathcal{P}(B)} T^{\mu,\nu} V^*(\theta\beta^{n+1}, r[h_{n+1}], p^*[h_{n+1}], x_{n+1})\}.$$

An optimal policy for player 1

Theorem 3 (Optimal policy for player 1)

Let $\pi^* = \{\pi_n^{*(k)}, k \in K, n \geq 0\}$. *The policy* π^* *is an optimal policy for player 1 in the game* $G(\theta, r, p)$, *i.e.,*

$$\inf_{\sigma \in \Pi_2} V(\theta, r, p, x, \pi^*, \sigma) = V^*(\theta, r, p, x) \quad \forall x \in E.$$

An optimal policy for player 2

- A natural idea for constructing an optimal policy σ^* for player 2 in $G(\theta, r, p)$ is via the Shapley equation.
- Corresponding to (C2), $\sigma^* = \{\sigma_n^*, n \geq 0\}$ should have the characteristic: the one-shift policy ${}^{(x_0,a_0,b_0)}\sigma^*$ of σ^* is optimal in

$$G(\theta\beta, \mathbb{H}^{x_0,a_0,b_0}(r), Q(\pi_0^*(\cdot|x_0), a_0, p)).$$

  This implies that σ^* must depend on π^*.
- Any optimal policy for player 2 cannot depend on anyone for player 1. We cannot obtain any optimal policy for player 2 by this idea.

♠ Construct an optimal policy for player 2 by introducing dual games.

Dual risk-sensitive games

A dual risk-sensitive game with incomplete reward information is defined as

$$G^{\#}(\theta, r, z) := \{\theta, K, E, A, B, z, q(y|x, a, b), r(k, x, a, b)\},$$

- where θ, K, E, A, B, q and r are the same as $G(\theta, r, p)$.
- The difference is that $p \in \mathcal{P}(K)$ in $G(\theta, r, p)$ is replaced by $z = \{z_k, k \in K\} \in \mathbb{R}_+^{|K|}$, which is used to modify the expected discounted risk-sensitive reward.

- Initially, both players observe an initial state x_0. According to the initial state x_0, player 1 chooses a reward type $k \in K$, which dose not change and is hidden from player 2 in the subsequent evolution.
- The subsequent evolution of the dual game $G^{\#}(\theta, r, z)$ is the same as that of $G(\theta, r, p)$.
- Finally, player 1 receives the reward $\sum_{n=0}^{\infty} \beta^n r(k, x_n, a_n, b_n) - z_k$.

Definition 3

A policy $\pi^{\#}$ for player 1 in the dual game is given by a two-tuple (ξ, π) with $\xi \in \mathcal{P}(K|E)$ and $\pi \in \Pi_1$. Denote by $\Pi_1^{\#}$ the set of all policies for player 1 in the dual game.

The dual games

- For $\pi^{\#} = (\xi, \pi) \in \Pi_1^{\#}$, $\sigma \in \Pi_2$, and $x \in E$, the expected risk-sensitive reward for player 1 in $G^{\#}(\theta, r, z)$ is defined as

$$U(\theta, r, z, x, \pi^{\#}, \sigma) = \mathbb{E}_x^{\pi^{\#},\sigma}\left[e^{\theta \sum_{n=0}^{\infty} \beta^n r(\Lambda, X_n, A_n, B_n)} - z_\Lambda\right]. \quad (7)$$

- $\underline{U}(\theta, r, z, x) := \sup_{\pi^{\#} \in \Pi_1^{\#}} \inf_{\sigma \in \Pi_2} U(\theta, r, z, x, \pi^{\#}, \sigma)$
- $\overline{U}(\theta, r, z, x) := \inf_{\sigma \in \Pi_2} \sup_{\pi^{\#} \in \Pi_1^{\#}} U(\theta, r, z, x, \pi^{\#}, \sigma)$
- Given (θ, r, z), if $\underline{U}(\theta, r, z, x) = \overline{U}(\theta, r, z, x)$ holds for all $x \in E$, the common function is called the value function of $G^{\#}(\theta, r, z)$ and is denoted by $U^*(\theta, r, z, x)$.

Definition 4

For $(\theta, r, z) \in (0, \infty) \times \mathcal{C} \times R_+^{|K|}$, *a policy* $\sigma^* \in \Pi_2$ *is called optimal for player 2 in the dual game* $G^\#(\theta, r, z)$ *if*

$$\sup_{\pi^\# \in \Pi_1^\#} U(\theta, r, z, x, \pi^\#, \sigma^*) = \overline{U}(\theta, r, z, x) \quad \forall x \in E.$$

Two questions:

♠ Is there an optimal policy for player 2 in the dual game?

♠ How to construct an optimal policy for player 2 in the primal game by that in the dual game?

Lemma 1

(a) *The value function* U^* *of* $G^{\#}(\theta, r, z)$ *exists.*

(b) $U^*(\theta, r, z, x) = \max_{p \in \mathcal{P}(K)}\{V^*(\theta, r, p, x) - \langle p, z\rangle\}$.

(c) $V^*(\theta, r, p, x) = \min_{z \in \mathbb{B}_r^\theta}\{U^*(\theta, r, z, x) + \langle p, z\rangle\}$ *where*

$$\mathbb{B}_r^\theta = \{z = (z_k)_{k \in K} \in \mathbb{R}_+^{|K|} | z_k \leq e^{\frac{\theta||r||}{1-\beta}}, k \in K\}$$

is a compact subset of $\mathbb{R}_+^{|K|}$.

♠ Lemma 1 shows the relationship between primal games and dual games.

Proposition 1

Given any $(\theta, r, z) \in (0, \infty) \times \mathcal{C} \times R_+^{|K|}$, *the value function* U^* *of the dual game* $G^*(\theta, r, z)$ *satisfies that for each* $x \in E$

$$U^*(\theta, r, z, x) = \min_{\nu \in \mathcal{P}(B)} \min_{f \in \mathcal{L}_r^\theta} \max_{p \in \mathcal{P}(K)} \max_{\mu \in \mathcal{P}(A|K)} \Gamma_{f,p}^{\mu,\nu} U^*(\theta, r, z, x), \tag{8}$$

where

$$\mathcal{L}_r^\theta := \left\{ f : A \times B \times E \to \mathbb{R}_+^{|K|} \,\middle|\, f(a, b, y) \in \mathbb{B}_r^\theta \;\; \forall (a, b, y) \in A \times B \times E \right\}$$

and

$$\Gamma_{f,p}^{\mu,\nu} U^*(\theta, r, z, x) := -\langle p, z \rangle + \sum_{k \in K} \sum_{a \in A} \sum_{b \in B} p_k \mu(a|k) \nu(b) \sum_{y \in E} q(y|x, a, b) \Big(U^*\left(\theta\beta, \mathbb{H}^{x,a,b}(r), f(a, b, y), y\right) + \left\langle Q(\mu, a, p), f(a, b, y) \right\rangle \Big).$$

Remark

From Proposition 1, given $h_1 = (x, a, b, y)$, player 2 can view choosing an action at $(n+1)$-th decision epoch in the game $G^{\#}(\theta, r, z)$ as choosing an action at n-th decision epoch in a new game

$$G^{\#}(\theta\beta, \mathbb{H}^{x,a,b}(r), f^*(a, b, y))$$

with an initial state y, where

$$f^* = \arg\min_{f \in \mathcal{L}_r^\theta} \Big\{ \min_{\nu \in \mathcal{P}(B)} \max_{p \in \mathcal{P}(K)} \max_{\mu \in \mathcal{P}(A|K)} \Gamma_{f,p}^{\mu,\nu} U^*(\theta, r, z, x) \Big\}.$$

Note that

$$\mathcal{L}_r^\theta := \Big\{ f : A \times B \times E \to \mathbb{R}_+^{|K|} \ \Big| f(a, b, y) \in \mathbb{B}_r^\theta \ \ \forall (a, b, y) \in A \times B \times E \Big\}$$

Given $(\theta, r, z) \in (0, \infty) \times \mathcal{C} \times R_+^{|K|}$, we define $\sigma_z^* = \{\sigma_{z,n}^*, n \geq 0\} \in \Pi_2$ (depending on (θ, r, z)) for player 2 as follows. For $h_0 = x_0 \in H_0$ let

$$\sigma_{z,0}^*(\cdot|h_0) := \arg\min_{\nu \in \mathcal{P}(B)} \Big\{ \min_{f \in \mathcal{L}_r^\theta} \max_{p \in \mathcal{P}(K)} \max_{\mu \in \mathcal{P}(A|K)} \Gamma_{f,p}^{\mu,\nu} U^*(\theta, r, z, x_0) \Big\}, \qquad (9)$$

and for $h_n = (h_{n-1}, a_{n-1}, b_{n-1}, x_n) \in H_n$ $(n \geq 1)$ let

$$\sigma_{z,n}^*(\cdot|h_n) := \arg\min_{\nu \in \mathcal{P}(B)} \Big\{ \min_{f \in \mathcal{L}_{r[h_n]}^{\theta\beta^n}} \max_{p \in \mathcal{P}(K)} \max_{\mu \in \mathcal{P}(A|K)} \Gamma_{f,p}^{\mu,\nu} U^*\big(\theta\beta^n, r[h_n], f^*[h_{n-1}](a_{n-1}, b_{n-1}, x_n), x_n\big) \Big\}, \qquad (10)$$

where $f^*[h_0] := \arg\min_{f \in \mathcal{L}_r^\theta} \Big\{ \min_{\nu \in \mathcal{P}(B)} \max_{p \in \mathcal{P}(K)} \max_{\mu \in \mathcal{P}(A|K)} \Gamma_{f,p}^{\mu,\nu} U^*(\theta, r, z, x_0) \Big\}$, and

$$f^*[h_n] := \arg\min_{f \in \mathcal{L}_{r[h_n]}^{\theta\beta^n}} \Big\{ \min_{\nu \in \mathcal{P}(B)} \max_{p \in \mathcal{P}(K)} \max_{\mu \in \mathcal{P}(A|K)} \Gamma_{f,p}^{\mu,\nu} U^*\big(\theta\beta^n, r[h_n], f^*[h_{n-1}](a_{n-1}, b_{n-1}, x_n), x_n\big) \Big\}. \qquad (11)$$

Theorem 4 (Optimal policy for player 2 in the dual game)

Given any $(\theta, r, z) \in (0, \infty) \times \mathcal{C} \times R_+^{|K|}$, the policy σ_z^* is an optimal policy for player 2 in the dual game $G^\#(\theta, r, z)$.

- For $(\theta, r, p) \in (0, \infty) \times \mathcal{C} \times \mathcal{P}(K)$, let

$$z^x := \arg\min_{z \in \mathbb{B}_r^\theta} \{U^*(\theta, r, z, x) + \langle p, z \rangle\}, \quad x \in E.$$

Denote by $\sigma_{z^x}^* = \{\sigma_{z^x,n}^*, n \geq 0\}$ the optimal policy for player 2 in the dual game $G^\#(\theta, r, z^x)$. Then, define $\sigma_p^* = \{\sigma_{p,n}^*, n \geq 0\}$ as

$$\sigma_{p,n}^*(\cdot|h_n) := \sigma_{z^{x_0},n}^*(\cdot|h_n), \quad h_n = (x_0, a_0, b_0, \ldots, x_n) \in H_n.$$

Theorem 5 (Optimal policy for player 2 in the primal game)

The policy σ_p^* is an optimal policy for player 2 in $G(\theta, r, p)$.

An example

- $K = \{1, 2\}$, $E = \{x_1, x_2, x_3\}$, $A = \{a_1, a_2\}$, $B = \{b_1, b_2\}$;
- $q(y|x_3, a_2, b) := \delta_{x_2}(y), \quad q(y|x_3, a_1, b) := \delta_{x_1}(y) \quad \forall b \in B, y \in E$;
- $q(y|x_2, a, b) = q(y|x_1, a, b) := \delta_{x_1}(y) \quad \forall a \in A, b \in B, y \in E$;
- $r(k, x_1, a, b) = 1$ for all $k \in K, a \in A, b \in B$;
- for $x \in \{x_2, x_3\}$,

$$
\begin{aligned}
r(1, x, a_1, b_1) = 4,\ & r(1, x, a_1, b_2) = 0, \\
r(1, x, a_2, b_1) = 2,\ & r(1, x, a_2, b_2) = 2, \\
r(2, x, a_1, b_1) = 0,\ & r(2, x, a_1, b_2) = 4, \\
r(2, x, a_2, b_1) = 2,\ & r(2, x, a_2, b_2) = 2.
\end{aligned}
$$

The value

For any $p = (p_1, p_2) \in \mathcal{P}(K)$: assume that $(e^{4\theta} + 1)e^{\theta\beta} - e^{2\theta}(e^{4\theta\beta} + 1) \geq 0$

- $V^*(\theta, r, p, x_1) = e^{\frac{\theta}{1-\beta}}$;
- $V^*(\theta, r, p, x_2) = \begin{cases} (p_1(e^{4\theta} + 1) + (1 - 2p_1)e^{2\theta})e^{\frac{\theta\beta}{1-\beta}}, & \text{if } p_1 \leq \frac{1}{2}, \\ (p_2(e^{4\theta} + 1) + (1 - 2p_2)e^{2\theta})e^{\frac{\theta\beta}{1-\beta}}, & \text{if } p_1 \geq \frac{1}{2}; \end{cases}$
- $V^*(\theta, r, p, x_3) = \begin{cases} p_1(e^{4\theta} + 1)e^{\frac{\theta\beta}{1-\beta}} + (1 - 2p_1)e^{2\theta + 2\theta\beta + \frac{\theta\beta^2}{1-\beta}}, & \text{if } p_1 \leq \frac{1}{2}, \\ p_2(e^{4\theta} + 1)e^{\frac{\theta\beta}{1-\beta}} + (1 - 2p_2)e^{2\theta + 2\theta\beta + \frac{\theta\beta^2}{1-\beta}}, & \text{if } p_1 \geq \frac{1}{2}; \end{cases}$

Optimal policy for player 1

- $\pi^{*(k)}_{p,0}(a|x_1) = \delta_{a_2}(a)$, $\pi^{*(k)}_{p,0}(a|x_3) = \pi^{*(k)}_{p,0}(a|x_2)$, where

$$\pi^{*(k)}_{p,0}(a|x_2) = \begin{cases} \delta_1(k)\delta_{a_1}(a) + \frac{p_1}{p_2}\delta_2(k)\delta_{a_1}(a) + (1 - \frac{p_1}{p_2})\delta_2(k)\delta_{a_2}(a), & \text{if } p_1 \le \frac{1}{2}, \\ \frac{p_2}{p_1}\delta_1(k)\delta_{a_1}(a) + (1 - \frac{p_2}{p_1})\delta_1(k)\delta_{a_2}(a) + \delta_2(k)\delta_{a_1}(a), & \text{if } p_1 \ge \frac{1}{2}. \end{cases}$$

- $\pi^{*(k)}_{p,1}(\cdot|x_3, a_2, b_2, x_2) = \pi^{*(k)}_{p,1}(\cdot|x_3, a_2, b_1, x_2)$, where

$$\pi^{*(k)}_{p,1}(\cdot|x_3, a_2, b_1, x_2) = \begin{cases} \delta_1(k)\delta_{a_1}(\cdot) + \delta_2(k)\delta_{a_2}(\cdot), & \text{if } p_1 \le \frac{1}{2}, \\ \delta_1(k)\delta_{a_2}(\cdot) + \delta_2(k)\delta_{a_1}(\cdot), & \text{if } p_1 \ge \frac{1}{2}. \end{cases}$$

- $h_1 \in H_1 \setminus \{(x_3, a_2, b_1, x_2), (x_3, a_2, b_2, x_2)\}$, $h_n \in H_n$ $(n \ge 2)$,

$$\pi^{*(k)}_{p,1}(a|h_1) = \pi^{*(k)}_{p,n}(a|h_n) := \delta_{a_2}(a);$$

Optimal policy for player 2 in $G(\theta, r, p)$ with $p_1 \geq \frac{1}{2}$

$\sigma^* = \{\sigma_n, n \geq 0\}$ as follows: for any $b \in B$

$$\sigma_0^*(b|x_2) = \delta_{b_1}(b)\frac{e^{2\theta} - 1}{e^{4\theta} - 1} + \delta_{b_2}(b)\frac{e^{4\theta} - e^{2\theta}}{e^{4\theta} - 1},$$

$$\sigma_0^*(b|x_3) = \delta_{b_1}(b)\frac{(e^{2\theta+\theta\beta} - 1)}{e^{4\theta} - 1} + \delta_{b_2}(b)\frac{(e^{4\theta} - e^{2\theta+\theta\beta})}{e^{4\theta} - 1},$$

$$\sigma_1^*(b|x_3, a_2, b_1, x_2) = \sigma_1^*(b|x_3, a_2, b_2, x_2) = \delta_{b_1}(b)\frac{e^{2\theta\beta} - 1}{e^{4\theta\beta} - 1} + \delta_{b_2}(b)\frac{e^{4\theta\beta} - e^{2\theta\beta}}{e^{4\theta\beta} - 1},$$

and for $h_1 \in H_1 \setminus \{(x_3, a_2, b_1, x_2), (x_3, a_2, b_2, x_2)\}$ and $h_n \in H_n$ $(n \geq 2)$,

$$\sigma_0^*(b|x_1) = \sigma_1^*(b|h_1) = \sigma_n^*(b|h_n) := \delta_{b_2}(b).$$

Combining the data of this example, Theorems 4 and 5, we have that σ^* is optimal for player 2 in the game $G(\theta, r, p)$ with $p_1 \geq \frac{1}{2}$.

Thanks!