$12^{ème}$ Biennale de la SMAI

Chance-constrained zero-sum discounted stochastic games

Lucas Osmani[1], Abdel Lisser[2], and Vikas Vikram Singh[3]

Laboratoire des signaux et systemes[1,2] and Department of Mathematics IIT Delhi [3]

June 2025

Table of Contents

Table of Contents

Topic of the talk

- We consider zero-sum stochastic games with probabilistic rewards.
- We assume that the distribution of the rewards is known to both players.
- The aim of each player is to get the maximum payoff he can guarantee with a given probability $p \in (0, 1)$, against the worst possible move from his opponent.
- The problem is formulated as a pair of chance-constrained optimization programs.
- This work is to be published in the Annals of operations' research (ANOR)

Table of Contents

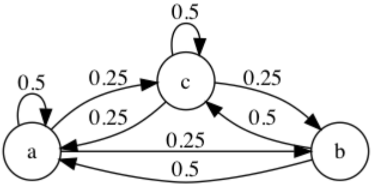
The model

Finite stochastic games

A two-players zero-sum stochastic game is defined by a tuple $\langle X, (A^1(x))_{x \in X}, (A^2(x))_{x \in X}, r, p \rangle$,

- X is a finite state space, and A^1, A^2, are finite action spaces.
- r is a reward function: when the game is in state x, and actions a^1 and a^2 are chosen, player 1 earns $r(x, a^1, a^2)$ while player 2 earns $-r(x, a^1, a^2)$.
- $p(y|x, a^1, a^2)$ denotes a probability that game moves to state y from x when player 1 and player 2 choose actions a^1 and a^2, respectively.

The model



Controlled Markov chains

The game starts at time $t = 0$ from an initial state x_0 which is selected according to an initial distribution m, i.e., x_0 is selected with probability $m(x_0)$. Player 1 and player 2 choose actions a_0^1 and a_0^2, respectively, and player 1 receives $r(x_0, a_0^1, a_0^2)$ and player 2 receives $-r(x_0, a_0^1, a_0^2)$. The game moves to state x_1 at time $t = 1$ with probability $p(x_1|x_0, a_0^1, a_0^2)$, and the same process repeats infinitely.

The model

Strategies

The strategy of a player represents a sequence of decision rules according to which actions are taken during the entire play:

- General strategies are history-dependent (they depend on the previous states and actions)
- A stationary strategy of player 1 is defined by a vector $f = (f(x))_{x \in X}$ where $f(x) \in \wp(A^1(x))$: whenever game is at state x, player 1 chooses action a^1 with probability $f(x, a^1)$.
- A stationary strategy g of player 2 is similarly defined.
- We denote the set of stationary strategies of player 1 and player 2 by F_S and G_S

The model

The discounted overall reward

Let X_t, A_t^1 and A_t^2 denote state and actions of player 1 and player 2 at time t, respectively. Future stage rewards are discounted by a factor $\alpha \in [0, 1)$. The objective of the game is:

$$V(m, f, g) = \sum_{t=0}^{\infty} \alpha^t \mathbb{E}_{f,g}^m \left(r(X_t, A_t^1, A_t^2) \right) . \quad (1)$$

- Player 1 wants to maximize V, and player 2 wants to minimize V.
- When rewards are deterministic, there exists a saddle point of V in $F_S \times G_S$, as proved by L.S. Shapley (1953).

The probabilistic reward

We consider a random reward function
$\tilde{r}(\omega) = (\tilde{r}(x, a^1, a^2, \omega))_{x \in X, a^1 \in A^1(x), a^2 \in A^2(x)}$

The random overall reward

$$\tilde{V}(m, f, g, \omega) = \sum_{t=0}^{\infty} \alpha^t \mathbb{E}^m_{f,g} \left(\tilde{r}(X_t, A^1_t, A^2_t, \omega) \right). \tag{2}$$

The aim of each player is to get the maximum payoff, that can be guaranteed with at least a given probability $p \in (0, 1)$, against the worst possible move from the opponent.

Chance-constrained formulation

Objective for player 1

$$\delta^*(p_1) := \max_{f \in F_S, \delta \in \mathbb{R}} \delta$$
$$\text{s.t.} \quad \min_{g \in G_S} \mathbb{P}(\tilde{V}(m, f, g) \geq \delta) \geq p_1. \tag{P1}$$

Objective for player 2

$$\eta^*(p_2) := \min_{g \in G_S, \eta \in \mathbb{R}} \eta$$
$$\text{s.t.} \quad \min_{f \in F_S} \mathbb{P}(\tilde{V}(m, f, g) \leq \eta) \geq p_2. \tag{P2}$$

Table of Contents

Reward distribution

Let $n = \sum_{x \in X} |A^1(x)||A^2(x)|$

Elliptical rewards

$\tilde{r} \sim Ellip_n(\mu, \Theta, \psi)$ where μ is a mean vector, Θ is a positive definite covariance matrix, and ψ is a characteristic generator, such that $\tilde{r}$ admits a strictly positive density.

Let $F^{-1}(\cdot)$ be a quantile function of $\tilde{r}$.

Occupation measures

The state-actions occupation measures

$$\rho_m^{f,g}(x, a^1, a^2) = \sum_{t=0}^{\infty} \alpha^t \mathbb{P}_{f,g}^m(X_t = x, A_t^1 = a^1, A_t^2 = a^2)$$

The value function has the following representation:

$$\tilde{V}(m, f, g, \omega) = \sum_{x \in X, a^1 \in A^1(x), a^2 \in A^2(x)} \tilde{r}(x, a^1, a^2, \omega) \rho_m^{f,g}(x, a^1, a^2) \quad (3)$$

Deterministic equivalent reformulation

Theorem

The problem is reformulated as follows:

$$\delta^*(p_1) = \max_{f \in F_S} \min_{g \in G_S} \left(\mu^\top \rho_m^{f,g} + F^{-1}(1 - p_1) \| \Theta^{\frac{1}{2}} \rho_m^{f,g} \|_2 \right), \tag{4}$$

$$\eta^*(p_2) = \min_{g \in G_S} \max_{f \in F_S} \left(\mu^\top \rho_m^{f,g} + F^{-1}(p_2) \| \Theta^{\frac{1}{2}} \rho_m^{f,g} \|_2 \right), \tag{5}$$

Deterministic equivalent reformulation

Proof.

We have $\tilde{V}(m,f,g) = \tilde{r}^\top \rho_m^{f,g}$. Define a standard normal random variable $Z = \frac{\tilde{r}^\top \rho_m^{f,g} - \mu^\top \rho_m^{f,g}}{\|\Theta^{\frac{1}{2}} \rho_m^{f,g}\|_2}$. Then, the chance constraint of (P1) can be reformulated as follows

$$\begin{aligned}
&\mathbb{P}(\tilde{V}(m,f,g) \geq \delta) \geq p_1, \ \forall \ g \in G_S, \\
\iff &\mathbb{P}\left(Z \geq \frac{\delta - \mu^\top \rho_m^{f,g}}{\|\Theta^{\frac{1}{2}} \rho_m^{f,g}\|_2} \right) \geq p_1, \ \forall \ g \in G_S, \\
\iff &\delta \leq \min_{g \in G_S} \mu^\top \rho_m^{f,g} + F^{-1}(1-p_1)\|\Theta^{\frac{1}{2}} \rho_m^{f,g}\|_2.
\end{aligned}$$

This implies that the optimal value $\delta^*(p_1)$ of player 1 satisfies (4). Similarly, the optimal cost $\eta^*(p_2)$ satisfies (5). □

Table of Contents

Results when $p_2 \leq 0.5$

We focus on player 2, when $p_2 \leq 0.5$,

Parameterized stochastic games

$$H(\lambda) = \min_{g \in G_S} \max_{f \in F_S} \sum_{t=0}^{\infty} \alpha^t \mathbb{E}^m_{f,g}(\tilde{u}(X_t, A^1_t, A^2_t))$$
$$= \max_{f \in F_S} \min_{g \in G_S} \sum_{t=0}^{\infty} \alpha^t \mathbb{E}^m_{f,g}(\tilde{u}(X_t, A^1_t, A^2_t)),$$

$\tilde{u}$ is given by $\tilde{u}(x, a^1, a^2) = \mu(x, a^1, a^2) + F^{-1}(p_2)(\Theta^{\frac{1}{2}}\lambda)_{x,a^1,a^2}$, and $\lambda \in \mathbb{R}^n$

Theorem

$$\eta^*(p_2) = \min_{\|\lambda\|_2 \leq 1} H(\lambda).$$

Results when $p_2 \leq 0.5$

1. $H(\cdot)$ is differentiable almost everywhere, and it admits directional derivatives
2. the minimum of $H(\cdot)$ lies on the sphere

Let $X^*(\lambda)$ and $Y^*(\lambda)$ denote the set of saddle points of the stochastic game $H(\lambda)$ for players 1 and 2 respectively.

Theorem

Let λ^* be a local minimum of $H(\cdot)$ on the unit sphere, then for every $g \in Y^*(\lambda^*)$, there exists $f \in X^*(\lambda^*)$, such that $\lambda^* = \frac{\Theta^{\frac{1}{2}} \rho_m^{f,g}}{\|\Theta^{\frac{1}{2}} \rho_m^{f,g}\|_2}$.

Algorithm for $p_2 \leq 0.5$

Algorithm 1

1. The inner loop: solve the stochastic game $H(\lambda_n)$ with vector $\lambda_n \in S$
2. Update $\lambda_{n+1} = \Gamma(\lambda_n)$

Where $\Gamma(\lambda) = \frac{\Theta^{\frac{1}{2}} \rho_m^{f,g}}{\|\Theta^{\frac{1}{2}} \rho_m^{f,g}\|_2}$, $f \in X^*(\lambda)$ and $g \in Y^*(\lambda)$ (We assume a unique saddle point) .

We show how to obtain optimal strategy for player 2 given an optimal λ^*. The convergence of this procedure is shown numerically.

Table of Contents

Results when $p_1 \geq 0.5$

We focus on player 1, when $p_1 \geq 0.5$. When for every $x \in X$, there exists an action $a^1 \in A^1(x)$ such that $f(x, a^1) = 1$ and $f(x, b) = 0$ for all $b \in A^1(x)$ such that $b \neq a^1$, we call f a pure stationary strategy. Similarly we can define a pure stationary strategy of player 2. We denote the set of pure stationary strategies of player 1 and player 2 by F_{PS} and G_{PS}, respectively

Theorem

$$\delta^*(p_1) = \max_{f \in F_S} \min_{g \in G_{PS}} \left\{ \langle \mu, \rho_m^{f,g} \rangle + F^{-1}(1 - p_1) \| \Theta^{\frac{1}{2}} \rho_m^{f,g} \|_2 \right\} \tag{6}$$

Since G_{PS} is finite, we obtain a discrete minimax formulation.

Nonlinear programming formulation

Let I be the index set for stationary deterministic strategies of player 2 and $(g_i)_{i \in I}$ denote their complete enumeration. For each $i \in I$, define a function

$$\phi_i(f) = \langle \mu, \rho_m^{f,g_i} \rangle + F^{-1}(1 - p_1) \| \Theta^{\frac{1}{2}} \rho_m^{f,g_i} \|_2.$$

The problem is equivalently written as:

Nonlinear program

$$\begin{aligned} & \delta^*(p_1) := \max \ y && (7) \\ \text{s.t.} \quad & (i) \ \phi_i(f) \geq y, \ \forall \ i \in I, \\ & (ii) \sum_{a^1 \in A^1(x)} f(x, a^1) = 1, \ \forall \ x \in X, \\ & (iii) \ f(x, a^1) \geq 0, \ \forall \ x \in X, a^1 \in A^1(x). \end{aligned}$$

Ascent directions

An ascent direction $d \in \mathbb{R}^N$ at a stationary policy $f \in F_S$ can be obtained from an optimal solution of the following quadratic program:

Quadratic program

$$
\begin{aligned}
\max_{y,d} \quad & y - \frac{1}{2}\|d\|^2 \\
\text{s.t.} \quad & y \leq \phi_i(f) + \nabla \phi_i(f)^\top d, \quad \forall\, i \in I_\epsilon(f), \\
& f(x, a^1) + d(x, a^1) \geq 0, \quad \forall\, x \in X,\; a^1 \in A^1(x), \\
& \sum_{a \in A^1(x)} d(x, a) = 0, \quad \forall\, x \in X.
\end{aligned}
\tag{8}
$$

Where $I_\epsilon(f) = \{j \in I \mid \phi_j(f) \leq \min_{i \in I} \phi_i(f) + \epsilon\}$

Algorithm

Algorithm 2

1. Find an ascent direction for the function to maximize, this is the result of the quadratic program (8).
2. Perform a line search.
3. Update the current strategy.

This algorithm converges to a KKT point of the nonlinear program (7).

Bilinear reformulation

- Alternatively, the problem can be formulated using a standard optimization program, including linear, bilinear, and SOCP constraints.
- This approach relies on several change of variables, into the space of discounted occupation measures.
- In practice, this problem is solved using a Gurobi solver.

Table of Contents

Numerical results

We make two assumptions:

- The rewards do not depend on players' actions
- The reward vector is normally distributed

In the first experiment, we consider a simple example where $|X| = 3$ and for every $x \in X$, $|A^1(x)| = |A^2(x)| = 3$. Let $X = \{x_1, x_2, x_3\}$

Numerical results

Table: Optimal solution of risk-seeking problems

p	$\delta^*(p)$	Algorithm 2 Optimal strategy	Dual vector	$\eta^*(p)$	Algorithm 1 Optimal strategy	Dual vector
0.45	-1.04958	$f^*(x_1) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ $f^*(x_2) = \begin{pmatrix} 0.81 \\ 0 \\ 0.19 \end{pmatrix}$ $f^*(x_3) = \begin{pmatrix} 0.67 \\ 0.08 \\ 0.25 \end{pmatrix}$	$\lambda^* = \begin{pmatrix} 0.71 \\ 0.50 \\ 0.49 \end{pmatrix}$	-2.40466	$g^*(x_1) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ $g^*(x_2) = \begin{pmatrix} 0.47 \\ 0 \\ 0.53 \end{pmatrix}$ $g^*(x_3) = \begin{pmatrix} 0 \\ 0.44 \\ 0.56 \end{pmatrix}$	$\lambda^* = \begin{pmatrix} 0.68 \\ 0.52 \\ 0.51 \end{pmatrix}$
0.4	-0.35613	$f^*(x_1) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ $f^*(x_2) = \begin{pmatrix} 0.81 \\ 0 \\ 0.19 \end{pmatrix}$ $f^*(x_3) = \begin{pmatrix} 0.61 \\ 0.09 \\ 0.29 \end{pmatrix}$	$\lambda^* = \begin{pmatrix} 0.71 \\ 0.50 \\ 0.50 \end{pmatrix}$	-3.08954	$g^*(x_1) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ $g^*(x_2) = \begin{pmatrix} 0.58 \\ 0.42 \\ 0 \end{pmatrix}$ $g^*(x_3) = \begin{pmatrix} 0 \\ 0.45 \\ 0.55 \end{pmatrix}$	$\lambda^* = \begin{pmatrix} 0.69 \\ 0.51 \\ 0.51 \end{pmatrix}$
0.3	1.15072	$f^*(x_1) = \begin{pmatrix} 0 \\ 0.22 \\ 0.78 \end{pmatrix}$ $f^*(x_2) = \begin{pmatrix} 0.78 \\ 0 \\ 0.22 \end{pmatrix}$ $f^*(x_3) = \begin{pmatrix} 0.49 \\ 0.11 \\ 0.40 \end{pmatrix}$	$\lambda^* = \begin{pmatrix} 0.76 \\ 0.48 \\ 0.43 \end{pmatrix}$	-4.54933	$g^*(x_1) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ $g^*(x_2) = \begin{pmatrix} 0.60 \\ 0.39 \\ 0 \end{pmatrix}$ $g^*(x_3) = \begin{pmatrix} 0 \\ 0.47 \\ 0.53 \end{pmatrix}$	$\lambda^* = \begin{pmatrix} 0.69 \\ 0.51 \\ 0.51 \end{pmatrix}$

Numerical results

Table: Optimal solution of risk-aversion problems

p	$\delta^*(p)$	Algorithm 3 Optimal strategy	CPU(s)	$\eta^*(p)$	Algorithm 4 Optimal strategy	CPU(s)
0.55	-2.41212	$f^*(x_1) = \begin{pmatrix} 0.93 \\ 0.06 \\ 0 \end{pmatrix}$ $f^*(x_2) = \begin{pmatrix} 0.83 \\ 0 \\ 0.17 \end{pmatrix}$ $f^*(x_3) = \begin{pmatrix} 0.19 \\ 0.21 \\ 0.601 \end{pmatrix}$	0.71	-1.04635	$g^*(x_1) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ $g^*(x_2) = \begin{pmatrix} 0.50 \\ 0.14 \\ 0.36 \end{pmatrix}$ $g^*(x_3) = \begin{pmatrix} 0.04 \\ 0.40 \\ 0.56 \end{pmatrix}$	0.74
0.6	-3.09056	$f^*(x_1) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ $f^*(x_2) = \begin{pmatrix} 0.83 \\ 0 \\ 0.17 \end{pmatrix}$ $f^*(x_3) = \begin{pmatrix} 0.02 \\ 0.26 \\ 0.71 \end{pmatrix}$	0.7	-0.35373	$g^*(x_1) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ $g^*(x_2) = \begin{pmatrix} 0.49 \\ 0.06 \\ 0.44 \end{pmatrix}$ $g^*(x_3) = \begin{pmatrix} 0.08 \\ 0.38 \\ 0.54 \end{pmatrix}$	0.72
0.7	-4.54958	$f^*(x_1) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ $f^*(x_2) = \begin{pmatrix} 0.84 \\ 0 \\ 0.16 \end{pmatrix}$ $f^*(x_3) = \begin{pmatrix} 0 \\ 0.29 \\ 0.71 \end{pmatrix}$	0.67	1.15688	$g^*(x_1) = \begin{pmatrix} 0.05 \\ 0 \\ 0.95 \end{pmatrix}$ $g^*(x_2) = \begin{pmatrix} 0.49 \\ 0 \\ 0.51 \end{pmatrix}$ $g^*(x_3) = \begin{pmatrix} 0.18 \\ 0.29 \\ 0.53 \end{pmatrix}$	0.8

Duality between risk-averse and risk-seeking players

Consider the following stochastic game:
$X = \{1, 2\}$, $A^1(1) = A^2(1) = \{1, 2\}$, $A^1(2) = A^2(2) = \{1\}$, $\mu(1) = 1$, $\mu(2) = 0$. $\Theta = I$. $m = \delta_1$, and the transition probabilities given by:

$$p(1 \to 1, a^1 = 1, a^2 = 1) = 1$$
$$p(1 \to 2, a^1 = 1, a^2 = 1) = 0$$
$$p(1 \to 1, a^1 = 2, a^2 = 2) = 1$$
$$p(1 \to 2, a^1 = 2, a^2 = 2) = 0$$
$$p(1 \to 1, a^1 = 2, a^2 = 1) = 0$$
$$p(1 \to 2, a^1 = 2, a^2 = 1) = 1$$
$$p(1 \to 1, a^1 = 1, a^2 = 2) = 0$$
$$p(1 \to 2, a^1 = 1, a^2 = 2) = 1$$

Duality between risk-averse and risk-seeking players

We assume that player 1 is risk-averse, and denote $C = -F^{-1}(1-p_1) \geq 0$. Player 2 is risk-seeking, and $p_2 = 1 - p_1$. The strategies of player 1 and 2 only depend on one number, $p = f(1,1) \in [0,1]$ and $q = g(1,1) \in [0,1]$. Notice that the game is symetric (same actions for each player), and zero-sum. Occupation measures can be computed explicitly:

$$\gamma_1^{f,g}(1) = \sum_{t=0}^{+\infty} \alpha^t (pq + (1-p)(1-q))^t = \frac{1}{1 - \alpha(pq + (1-p)(1-q))}$$

We have $\gamma_1^{f,g}(2) = \frac{1}{1-\alpha} - \gamma_1^{f,g}(1)$. Take $\alpha = 0.5$, and the objective is now explicit:

$$\delta^* = \max_{p\in[0,1]} \min_{q\in[0,1]} \frac{1 - C\sqrt{1 + (p+q-pq)^2}}{1 - \frac{1}{2}(pq + (1-p)(1-q))} \tag{9}$$

Duality between risk-averse and risk-seeking players

The bivariate function that appears in (9) does not necessarily have a saddle point in $[0,1] \times [0,1]$, depending on the value of the parameter C. In consequence, strong duality does not always hold. We can prove that strong duality is equivalent to the optimality of the stationary strategy class.

Bilinear reformulation

$$\max_{y,\rho_i} \quad y \tag{10}$$

$$\text{s.t.} \quad (i)\ y \le \mu^\top \hat{\rho}_i + F^{-1}(1-p_1)\|\Theta^{\frac{1}{2}}\hat{\rho}_i\|_2, \quad i \in I$$

$$(ii)\ \rho_i \in K^{g_i},\ i \in I,$$

$$(iii)\ \rho_i(x,a^1)\sum_{a\in A^1(x)}\rho_1(x,a) = \rho_1(x,a^1)\sum_{a\in A^1(x)}\rho_i(x,a),\ \forall\ i \in I \setminus\ \{1\}$$

Where K^{g_i} is the occupation measure polytope, when g_i a fixed pure strategy. We compare Gurobi solver with Algorithm 2.

Numerical experiments

Table: Comparison between Algorithm 2 and Gurobi

Example	—X—	A	Algorithm 3		Gurobi		Duality gap (upper bound)
			Objective value	CPU(s)	Objective value	CPU(s)	
1	3	2	2.9797	0.2	2.9798	0.05	4.10^{-2}
2	3	3	-14.6838	0.9	-14.6838	10	2.10^{-4}
3	4	2	-6.69755	0.6	-6.69754	0.07	3.10^{-4}
4	4	3	4.74672	6.6	4.74676	7	0.1
5	4	4	2.60343	8.7	2.57851	200	5.10^{-2}
6	5	2	-2.34345	1.7	-2.34345	0.61	1.10^{-3}
7	5	3	-1.848352	380	-1.84834	4	1.10^{-2}
8	5	4	-6.21586	74.6	-	-	-
9	6	4	-0.16407	141.7	-	-	-

Table of Contents

Conclusion and remarks

The proposed approach can be generalized, for the study of other reward distributions, in particular $\alpha -$ *stable* ones. The continuous-time version of this problem could be formulated in the future.

Some references

- Chance-constrained zero-sum discounted stochastic games (2025)
- Stochastic games were first studied by L.S. Shapley (1953).
- E. Delage and S. Mannor (2010) studied Markov decision processes with random rewards.
- R. Blau (1974) studied zero-sum games with a random payoff matrix, using a chance-constrained formulation that we draw inspiration from.
- V.V. Singh and A. Lisser (2018) studied existence of Nash equilibria in a class of games with random payoffs.
- N. Bäuerle and U. Rieder (2016) studied risk-sensitive stochastic games.

Thank you for your attention.