XVIIth conference on stochastic programming

Discounted zero-sum stochastic games with random rewards

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]

July 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.

Table of Contents

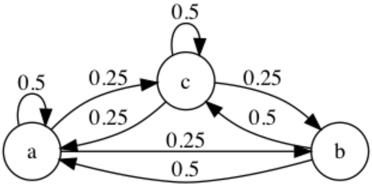
The model

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 $\tilde{r}(\omega) = (\tilde{r}(x, a^1, a^2, \omega))_{x \in X, a^1 \in A^1(x), a^2 \in A^2(x)}$ defined in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$

The random overall reward

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

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

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

Objective for player 2

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

Table of Contents

Reward distribution

We assume that the random reward vector $\tilde{r}$ follows an elliptical 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

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

The value function has the following representation:

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

Deterministic equivalent reformulation

Problems (P1) and (P2) are reformulated, respectively as (4) and (5)

Theorem

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

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

Where $\gamma_m^{f,g}$ is the state-actions occupation measure.

Deterministic equivalent reformulation

Proof.

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

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

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

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

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

Results when $p_2 \leq 0.5$

We prove that the optimum in (P2) can be computed as the minimum of parameterized stochastic games.

Theorem

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

We prove the following two results:

1. $H(\cdot)$ is continuously differentiable almost everywhere.
2. the minimum of $H(\cdot)$ lies on the sphere

Algorithm for $p_2 \leq 0.5$

We use a Riemannian gradient sampling algorithm proposed by S.Hosseini and A.Uschmajew (2017), to find the minimum of $H(.)$ on the unit sphere.

Algorithm 1

1. Compute spherical gradients of H at random points $(\lambda_n^i) \in B(\lambda_n, \epsilon_n) \cap S(0,1)$, where $\lambda_n \in S(0,1)$ is the current iterate.
2. Transport the gradients to the tangent space at λ_n.
3. Compute the least norm vector in the convex hull of the transported gradients, denoted by g_n.
4. Perform a line search, and then the update $\lambda_n \leftarrow \frac{\lambda_n - tg_n}{\|\lambda_n - tg_n\|_2}$
5. Set $\epsilon_n \leftarrow \epsilon_n \times \theta$ where $\theta \in (0,1)$

Algorithm for $p_2 \leq 0.5$

We prove that Algorithm 1 converges to a stationary point of $H(.)$. Given an optimal λ^*, the optimal strategy of player 2 is obtained by solving a linear program.

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

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

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

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

Then problem (6) is equivalently written as:

Nonlinear program

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

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 for risk averse player

Algorithm 2

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

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

Bilinear reformulation for risk averse player

- 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.

Bilinear reformulation for risk averse player

$$\max_{y,\rho_i} \quad y \tag{9}$$

$$\text{s.t.} \quad (i)\ y \leq \tilde{\mu}_i^\top \rho_i + F^{-1}(1-p_1)\|\tilde{\Sigma}_i \rho_i\|_2, \quad i \in I$$

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

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

Where K^{g_i} is the occupation measure polytope, when g_i a fixed pure strategy, and $(\tilde{\mu}_i, \tilde{\Sigma}_i)$ are obtained by removing the entries of $(\mu, \Theta^{\frac{1}{2}})$ corresponding to an action which is not chosen by g_i.

Table of Contents

Restriction to stationary strategies

We assume that strong duality holds for $(P1)$,

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

Then,

$$\delta^*(p_1) = \max_{f \in F} \min_{g \in G} \left(\mu^\top \gamma_m^{f,g} + F^{-1}(1-p_1) \|\Theta^{\frac{1}{2}} \gamma_m^{f,g}\|_2 \right) \tag{10}$$

Where F and G are the sets of history-dependent strategies.

Table of Contents

Numerical results

We assume the reward vector is normally distributed 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 solutions of risk-averse and risk-seeking problems

p	Risk-averse problem			Risk-seeking problem	
	Algorithm 2		Gurobi	Algorithm 1	
	$\delta^*(p)$	Optimal strategy	Objective	$\eta^*(1-p)$	Optimal strategy
		$f^*(x_1) = \begin{pmatrix} 0.03 \\ 0.97 \\ 0 \end{pmatrix}$			$g^*(x_1) = \begin{pmatrix} 0.26 \\ 0 \\ 0.74 \end{pmatrix}$
0.55	-0.748	$f^*(x_2) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$	-0.925	-0.043	$g^*(x_2) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$
		$f^*(x_3) = \begin{pmatrix} 0.27 \\ 0.72 \\ 0 \end{pmatrix}$			$g^*(x_3) = \begin{pmatrix} 0.65 \\ 0 \\ 0.35 \end{pmatrix}$
		$f^*(x_1) = \begin{pmatrix} 0.17 \\ 0.83 \\ 0 \end{pmatrix}$			$g^*(x_1) = \begin{pmatrix} 0 \\ 0.15 \\ 0.85 \end{pmatrix}$
0.6	-3.106	$f^*(x_2) = \begin{pmatrix} 0.06 \\ 0.94 \\ 0 \end{pmatrix}$	-3.632	-2.031	$g^*(x_2) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$
		$f^*(x_3) = \begin{pmatrix} 0.34 \\ 0.66 \\ 0 \end{pmatrix}$			$g^*(x_3) = \begin{pmatrix} 0.69 \\ 0 \\ 0.31 \end{pmatrix}$
		$f^*(x_1) = \begin{pmatrix} 0.20 \\ 0.74 \\ 0.05 \end{pmatrix}$			$g^*(x_1) = \begin{pmatrix} 0 \\ 0.22 \\ 0.78 \end{pmatrix}$
0.7	-7.878	$f^*(x_2) = \begin{pmatrix} 0.24 \\ 0.72 \\ 0.04 \end{pmatrix}$	-9.046	-7.553	$g^*(x_2) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$
		$f^*(x_3) = \begin{pmatrix} 0.36 \\ 0.58 \\ 0.06 \end{pmatrix}$			$g^*(x_3) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

Table of Contents

Conclusion and remarks

We considered a stochastic game with random rewards, and formulated a chance-constrained optimization program for each player. Under the assumption of elliptical rewards, we proved equivalence of the chance-constrained progams to a minimax.
We studied risk-averse and risk-seeking players separately, and proposed algorithms for each case.

Some references

- Discounted zero-sum stochastic games with random rewards (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.
- S.Hosseini and A.Uschmajew (2017) propose a Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds

Thank you for your attention.