Stochastic Games (SG)

This tutorial includes:

- An introduction on games theory
- An intuitive and formal explanation on what is stochastic game and how it is related to our control course.
- Some examples and results

- This presentation relay on the work of:
  - L. S. Shapley
  - Michael Kearns
  - Cachon and Zipkin
  - Netessine and Rudi

Example: The Prisoner's Dilemma

- Two suspects in a crime are interrogated in separate rooms
- Each has two choices: confess or deny
- With no confessions, enough evidence to convict on lesser charge; one confession enough to establish guilt
- Police officer plea bargains for confessing
- Encode strategic conflict as a payoff matrix:

payoffs	confess	deny
confess	-3,-3	0,-4
deny	-4,0	-1,-1

- What should happen?

Example: Hawks and Doves

- Two players compete for a valuable resource
- Each has a confrontational strategy ("hawk") and a conciliatory strategy ("dove")
- Value of resource is V ; cost of losing a confrontation is C
- Suppose C > V (think nuclear first strike)
- Encode strategic conflict as a payoff matrix:

payoffs	hawk	dove
hawk	(V-C)/2, (V-C)/2	V,0
dove	0,V	V/2,V/2

- What should happen?

Assumptions

- Players optimize their payoffs
- Players are selfish and play their best response

A Formal Definition of a Game

- Set of players $i = 1,\ldots,n$ (assume $n = 2$ for now)
- Each player has a set of m basic actions or pure strategies (such as “hawk” or “dove”)
- Notation: a_i will denote the strategy chosen by player i
- Joint action: $\vec{a}$
- Payoff to player i given by matrix or table $M_i(\vec{a})$
- Goal of players: maximize their own payoff

Game Strategy

A strategy could be pure (=deterministic) or mixed (=randomized)

Mixed strategy

- Each player i has an independent distribution p_i over their pure strategies
- Use $\vec{p} = (p_1; \ldots ; p_n)$ to denote the product distribution induced over joint action $\vec{a}$
- Use $\vec{a} \sim \vec{p}$ to indicate a distributed according to $\vec{p}$

The Concept of Equilibrium

- An equilibrium among the players is a strategic standoff
- No player can improve on their current strategy
- Different types of equilibrium assume different models of communication, coordination, and collusion among the players; Nash equilibrium assumes no communication or bargaining.

Nash Equilibrium

- Expected return to player i over mixed strategy $\vec{a}$ is $E_{\vec{a}\sim\vec{p}}[M_i(\vec{a})]$
- A Nash equilibrium is a situation where no player has a unilateral incentive to deviate

Formally:

- Let $\vec{p}[i : p_i']$ denote $\vec{p}$ with p_i replaced by p_i'
- Thus: $\vec{p}$ is a Nash equilibrium (NE) if for every player i, and every mixed strategy p_i':

  $E_{\vec{a}\sim\vec{p}}[M_i(\vec{a})] \geq E_{\vec{a}\sim\vec{p}[i:p_i']}[M_i(\vec{a})]$

Nash 1951: NE always exist in mixed strategies

NE of the Prisoner's Dilemma

- The payoff matrix:

payoffs	confess	deny
Confess	-3,-3	0,-4
Deny	-4,0	-1,-1

- One (pure) NE: (confess,confess)

NE of Hawks and Doves

- The payoff matrix (C>V):

payoffs	hawk	dove
hawk	(V-C)/2, (V-C)/2	V,0
dove	0,V	V/2,V/2

- Three NE:
  - pure: (hawk,dove)
  - pure: (dove,hawk)
  - mixed: (Pr[hawk] = V/C,Pr[hawk] = V/C)

Game Value

- <u>Suggestion</u>: Can we define the game value by the utility that each player get at a Nash Equilibrium?
- <u>Problem</u>: A Game can have few values

Value of the Prisoner's Dilemma

- The payoff matrix:

payoffs	confess	deny
Confess	-3,-3	0,-4
Deny	-4,0	-1,-1

- One (pure) NE: (confess,confess)
- val[M] = (-3,-3)

Value of Hawks and Doves

- The payoff matrix (C>V):

payoffs	hawk	dove
hawk	(V-C)/2, (V-C)/2	V,0
dove	0,V	V/2,V/2

- Three NE payoffs:
  - pure: (hawk,dove) -> $val_1[M] = (V,0)$
  - pure: (dove,hawk) -> $val_2[M] = (0,V)$
  - mixed: (Pr[hawk] = V/C,Pr[hawk] = V/C) ->

$$val_3[M] = \left(\left(1-\frac{V}{C}\right)\frac{V}{2}, \left(1-\frac{V}{C}\right)\frac{V}{2}\right)$$

Security level = 0

Game Value

- Security level: the payoff that player can ensure for themselves regardless of their opponent's behavior; $s[M_1] = \max_\alpha \min_\beta (M_1(\alpha,\beta))$
- A zero-sum game have only one value which is it's security level, in a general-sum game security level is lower bound for the value

Different Types of Games

- Cooperative and Non-cooperative game
- Zero-Sum Vs. General-Sum games
- Repeated games
- Example of strategy that changes over time...

What is a stochastic game?

Shapley 1953:

"In a stochastic game the play proceeds by steps from position to position, according to transition probabilities controlled jointly by the two players"

A formal notation

- N - number of players
- S - set of states (finite/countable)
- At each state $s \in S$ the compact sets of admissible actions A_{js} are available to player j
- $P(A_{js})$ – set of all probability distributions on A_{js}

A formal notation

- a - vector of actions of the N players, where a_j is a randomized (mixed) action on $P(A_{js})$
- $M(j,s,a)$ – immediate reward earned by player j at this stage if the players act according to a
- $q(s'|s,a)$ – transition probability of the system to a new state s'
- $\pi_j(s)$- policy of player j at state s

Game value or SG Dynamic Programming

- β- discount factor
- $v_j(s,\pi)$- Expected stationary policy of player j over infinite horizon:

$$V_j(s,\pi)=E_s^{\pi}\sum_{t=1}^{\infty}\beta^{t-1}M_j(s_t,\alpha_t) \qquad \beta<1$$

- The Expected stationary policy of player j over finite horizon T:

$$V_j^T(s,\pi)=E_s^{\pi}\sum_{t=1}^{T}\beta^{t-1}M_j(s_t,\alpha_t) \qquad \beta\le 1$$

Nash Equilibrium in SG

- we say that $(\pi_1; \pi_2)$ is a Nash Eq. (for two players) if for any start state s_0 and any π_0',
  $V_1(s_0; \pi_0'; \pi_2) \leq V_1(s_0; \pi_1; \pi_2)$,
  and for any start state s_0 and any π_0',
  $V_2(s_0; \pi_1; \pi_0') \leq V_2(s_0; \pi_1; \pi_2)$

Example 1 – Pollution Tax Model

- Two firms contribute to the emission of certain pollutant. The government can detect only the combined emissions, and only if it is high.
- The Profit Matrix:

Profit	Clean	Dirty
Clean	(4,5)	(3,8)
Dirty	(7,4)	(6,7)

- What is the Nash Equilibrium?

Example 1 – Pollution Tax Model

(state 1: no tax)

Profit / trans. pr.	Clean	Dirty
Clean	(4,5) / (1,0)	(3,8) / (0,1)
Dirty	(7,4) / (0,1)	(6,7) / (0,1)

(state 2: tax = 3)

Profit / trans. pr.	Clean	Dirty
Clean	(1,2) / (1,0)	(0,5) / (0,1)
Dirty	(4,1) / (0,1)	(3,4) / (0,1)

Example 2 – Strike Negotiation Model

- Management and union negotiate about salary level
- At day t-1 the Management offered an increase of $x_1(t-1)$ and union demanded $x_2(t-1)$ (of course $x_1(t-1) < x_2(t-1)$)
- At time t: $x_k \in [x_1(t-1)\ ,\ x_2(t-1)]$
- If $x_1(t) < x_2(t)$ strike continue

Example 2 – Strike Negotiation Model

- Strike cost $L(t)$ to management $S(t)$ for union
- If $x_1(t) \geq x_2(t)$ strike stop and agree on a new salary level $x_a = 0.5(x_1(t) + x_2(t))$
- Future Utility: $f_1(x_a,t)$ cost of Management and $f_2(x_a,t)$ profit to union
- The decision moment is t_a

Example 2 – Strike Negotiation Model

- Management try to minimize

$$(1-\beta)\sum_{\tau=0}^{t_a-1}\beta^{\tau} l(\tau)+(1-\beta)\beta_a^{t} f_1(x_a, t_a)$$

- Union try to maximize

$$(1-\beta)\beta_a^{t} f_2(x_a, t_a)-(1-\beta)\sum_{\tau=0}^{t_a-1}\beta^{\tau} s(\tau)$$

Stochastic games and MDP

- The analogy: MDP is a stochastic game where all other players have only one choice
- We look for an Equilibrium, i.e. a strategy under which if each player plays in order to maximize it's utility, this strategy will be "stable"... does such policy exist? If yes, can we find such policy? Is stationary policies suffice?
- In what way will my strategy change if I consider other player strategy? Can one affect on finding optimal equilibrium?

Some Results

- Shapley 1953: In finite horizon, zero-sum stochastic game for 2 players, with positive stopping times, there exist an optimal (mixed) strategy which leads to a unique value of the game

proof:

  - Uniqueness - throw contraction operators
  - Existence - by setting a lower bound on the payments of each player

The Optimality Function in Zero-Sum Games

- Given a matrix game M, let val[M] denote its min-max value to the first player, and a, b the sets of optimal mixed strategies for the first and second players, respectively.
- For finite horizon:

$$v^0(s)=0$$

$$v^{t+1}(s)=val_{a,b}[M(s;a,b)+\beta\sum_{s'\in S}q(s'|s;a,b)v^t(s')]$$

- For infinite horizon:

$$t=0,1,2\ldots$$

$$v(s)=val_{a,b}[M(s;a,b)+\beta\sum_{s'\in S}q(s'|s;a,b)v(s')]$$

Some Results

- Ininfinite discounted case, a Nash pair (NE) always exists in the space of stationary policies

How to find the EP?

- LP applicable for some games
- Value Iteration
- A Modified Newton's Method

A Modified Newton's Method

- Define:

$$R(s, v_\beta) = \left[M(s; a, b) + \beta \sum_{s' \in S} q(s' | s; a, b) v_\beta(s') \right]_{a,b}$$

- Shapley's theorem proved that

$$L(v)(s) \stackrel{\text{def}}{=} val[R(s, v)] \quad \forall v \in \mathbb{R}^N, s \in S$$

is construction operator with unique fixed point $L(v) = v$

- This is equivalent to finding zero of:

$\psi(v) \stackrel{\text{def}}{=} L(v) - v$ or solving: $min\, J(v) = \frac{1}{2}[\psi(v)^T \psi(v)]$

$$s.t. \quad v \in \mathbb{R}^N$$

A Modified Newton's Method

- The general algorithm:
  In iteration k - v^k is the current solution
  - Search direction: $d^k \stackrel{\text{def}}{=} -[\psi'(v^k)]^{-1}\psi(v^k)$
  - Step size $\omega \in (0,1]$ set in order to insure convergence
  - New estimated solution:

$$v^{k+1} = v^k - \omega^k [\psi'(v^k)]^{-1}\psi(v^k)$$

  - If $J(v^k)=0$ stop and $v_\beta = v^k$.

How to find the Equilibrium?

- Value-Iteration Algorithm for Finite Horizon

Algorithm FiniteVI(T)**:**

Initialization:

For all $s \in S$, $k \in 1, 2$:

$Q_k[s, 0] \leftarrow M_k[s]$;

$\pi_k(s, 0) \leftarrow f_k(M_1[s], M_2[s])$;

Iteration $t = 1 \ldots T$**:**

For all $s \in S$, $k \in \{1, 2\}$:

For all pure strategies i and j:

$Q_k[s, t](i, j) \leftarrow M_k[s](i, j) + \sum_{s'} P(s'|s, i, j) v_f^k(Q_1[s', t-1], Q_2[s', t-1])$;

$\pi_k(s, t) \leftarrow f_k(Q_1[s, t], Q_2[s, t])$;

Return the policy pair (π_1, π_2);

How to find the Equilibrium?

- Value-Iteration Algorithm for Infinite Horizon

Algorithm InfiniteVI(T)**:**
Initialization: for all $s \in S$, $k \in 1, 2$:
$Q_k[s, 0] \leftarrow M_k[s]$;
$\pi_k(s) \leftarrow f_k(M_1[s], M_2[s])$;
Iteration $t = 1, \ldots, T$**:** for all $s \in S$, $k \in \{1, 2\}$:
For all pure strategies i and j:
$Q_k[s, t](i, j) \leftarrow M_k[s](i, j) + \gamma \sum_{s'} P(s'|s, i, j) v_f^k(Q_1[s', t-1], Q_2[s', t-1])$;
$\pi_k(s) \leftarrow f_k(Q_1[s, t], Q_2[s, t])$;
Return the policy pair (π_1, π_2);

Will my strategy change? Inventory models

- In two competitors inventory model - Netessine et al. (2005) showed that the order-up-to policy is a NE
- In two-stage supply chain Cachon and Zipkin (1999) showed that games have different optimal solution then MDP, though the same structure. Thus, NE policies (under competition) reduce efficiency