
Multiagent Learning Sampler

CS 536: Machine Learning
Littman (Wu, TA)

Multiagent Learning

The MDP (and POMDP) models we've talked about assume that everything in the environment that's not part of the decision maker is controlled by stationary random distributions.

Even ignoring the issue of the world not being stationary, there is the issue of other agents...

How should we behave in an environment that has other, self-interested decision makers? (Show Michael Bowling's video.)

Recommended Reading

Material from a workshop and two tutorials I gave last year, with help from Gerry Tesauro and Michael Bowling.

- Fudenberg & Levine (1998), *The Theory of Learning in Games*.
- Littman (1994), "Markov games as a framework for multi-agent reinforcement learning" in *ICML-94*.
- Singh, Kearns & Mansour (2000), "Nash convergence of gradient dynamics in general-sum games" in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*.

Outline

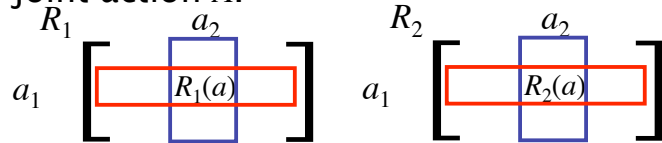
- A. Normal-form Games
- B. Repeated Games
- C. Stochastic Games

A. Normal-Form Games

n players, A_i action set for player i ,

A is joint action space: $A_1 \times \dots \times A_n$.

R_i reward to player i as a function of the joint action A .



If 2 players, sometimes called *bimatrix game*.

Strategies

Notation: σ is a joint strategy for all players.

$$R_i(\sigma) = \sum_{a \in A} \sigma(a) R_i(a)$$

- σ_{-i} is a joint strategy for all players except i .
- $\langle \sigma_i, \sigma_{-i} \rangle$ is a joint strategy where i uses strategy σ_i and everyone else σ_{-i} .

Types of games:

- zero-sum games: $R_1 + R_2 = 0$.
- team games: $R_1 = \dots = R_n$.
- general-sum games: (ex. Prisoner's dilemma)

$$R_1 \begin{array}{c} C \ D \\ \begin{bmatrix} 3 & 0 \\ 4 & 1 \end{bmatrix} \end{array} \quad R_2 \begin{array}{c} C \ D \\ \begin{bmatrix} 3 & 4 \\ 0 & 1 \end{bmatrix} \end{array}$$

Ex.: Rock-Paper-Scissors

Two players. Each independently and simultaneously chooses an action: rock, paper, or scissors.

Rewards: rock *beats* scissors
scissors *beats* paper
paper *beats* rock

Matrices: R_1 $\begin{array}{c} r \ p \ s \\ \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} \end{array}$ R_2 $\begin{array}{c} r \ p \ s \\ \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \end{array}$

Solutions: Minimax

Consider *matching pennies*:

$$R_1 \begin{array}{c} H \ T \\ \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{array} \quad R_2 \begin{array}{c} H \ T \\ \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \end{array}$$

Q: What do we do when the world is out to get us?

A: Make sure it can't.

Play strategy with the best worst-case outcome.

$$\operatorname{argmax}_{\sigma_i \in \Delta(A_i)} \min_{a_{-i} \in A_{-i}} R_i(\langle \sigma_i, \sigma_{-i} \rangle)$$

Minimax optimal strategy.

Minimax

Back to matching pennies:

$$R_1 \begin{matrix} H & T \\ H & \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \\ T & \end{matrix} \quad \sigma_1^* \begin{matrix} H \\ T \end{matrix} \begin{bmatrix} -1/2 \\ -1/2 \end{bmatrix}$$

$R_1(\langle \sigma_1^*, \sigma_{-i} \rangle) = 0$ for all σ_{-i}

Minimax optimal guarantees the **safety value**.

Generally, bad opponent choice can help (saddle point).

Nash Equilibria

Nash Equilibrium: joint strategy of best responses only.

If opponents don't change, no incentive to shift.

For all i , if there is no σ_i^* such that

$$R_i(\langle \sigma_i^*, \sigma_{-i} \rangle) > R_i(\langle \sigma_i, \sigma_{-i} \rangle)$$

then σ is a Nash equilibrium.

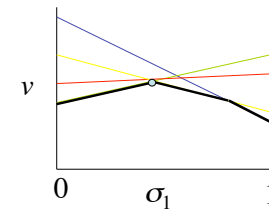
Unique value in a zero-sum game (minimax!).

Not unique value in general.

Prisoner's dilemma has a unique Nash equilibrium. (What is it?)

Minimax: Linear Programming

- Finding minimax optimal *equivalent* to LP (so in P).
- Payoff vs. opponent action linear function of strategy.
- Since opponent minimizes, choices on lower surface.
- Choose best point: $\max_{v, \sigma_1} v$ s.t. $R_1(\sigma_1, b) \geq v$ for all i



Computing a Nash

Two players: with factoring, the most significant open problem on the boundary of P today. (Papadimitriou 01)

On the one hand:

- nondeterministic algorithm: given the right support, in P.
- in $F(NP \cap co-NP)$; unlikely to be NP-hard.

On the other hand:

- "Is there a Nash?": Easy, always "yes".
- But NP-hard to $\max(v_1+v_2)$, $\max(\min(v_1, v_2))$, $\max(v_i)$, Pareto-optimal, more than one equilibrium, specific action in/out of support. (Gilboa & Zemel 89; Conitzer & Sandholm 03)

Simplex-like algorithm (Lemke-Howson) does well in practice.

Three players: Equilibrium can be irrational. (Nash 51)

Joint Distributions

Bach/Stravinsky game:

$$\begin{array}{c}
 R_1 \quad B \quad S \\
 B \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \\
 S \\
 \end{array}
 \quad
 \begin{array}{c}
 R_2 \quad B \quad S \\
 B \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \\
 S
 \end{array}$$

Pure equilibria: always B: (2, 1), always S: (1, 2)

$$\begin{array}{c}
 R_1 \\
 B \begin{bmatrix} 2/3 \\ 2/3 \end{bmatrix} \\
 S \\
 \end{array}
 \quad
 \begin{array}{c}
 R_2 \quad B \quad S \\
 \begin{bmatrix} 2/3 & 2/3 \end{bmatrix}
 \end{array}$$

Mixed equilibrium: 1/3 preferred: (2/3, 2/3)

Joint Distributions

Bach/Stravinsky game:

$$\begin{array}{c}
 R_1 \quad B \quad S \\
 B \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \\
 S \\
 \end{array}
 \quad
 \begin{array}{c}
 R_2 \quad B \quad S \\
 B \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \\
 S
 \end{array}$$

Not achievable
as product of
distributions

Distributions on joint actions

$$\begin{array}{c}
 B \quad S \\
 B \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\
 S \\
 \end{array}
 \quad
 \begin{array}{c}
 B \quad S \\
 B \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\
 S
 \end{array}$$

$$\begin{array}{c}
 B \quad S \\
 B \begin{bmatrix} 2/3 & 1/9 \\ 4/9 & 2/9 \end{bmatrix} \\
 S \\
 \end{array}
 \quad
 \begin{array}{c}
 B \quad S \\
 B \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \\
 S
 \end{array}$$

Correlated Equilibria

Different equilibrium concept

- Let σ be distribution over *joint actions*.
- Joint action, sampled, each player shown its part.
- σ is a **correlated equilibrium** if no player can gain by deviating from its prescribed action.

Comparison of Equilibria

Nash equilibrium

- Players told distribution over joint actions of other players
- Players given prescribed distribution over its own actions
- Player flips weighted coin to select its action
- Should have no incentive to deviate unilaterally.

Correlated equilibrium

- Players told distribution over joint actions of all players (including their own)
- Each player's prescription factored into the joint distribution
- Player told which action to play from joint action
- Should have no incentive to deviate unilaterally

Computing Correlated Equilibria

Because joint distribution not constrained to be product, can set up as LP (for n players).

In P: $\max(v_1+v_2)$, $\max(\min(v_1, v_2))$, $\max(v_1)$

Arguably more natural than Nash. Much less attention.

All Nash and their mixtures are correlated equilibria.

Issue: Where does the correlator come from?

Learn to use a random bits as a correlating device?

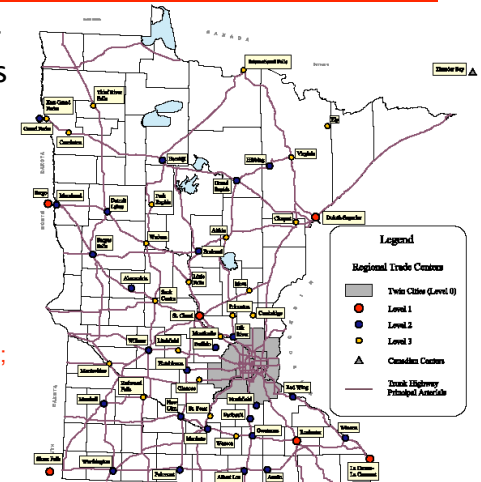
Representing Many-Player Games

Takes $O(n^m)$ space for n players, m actions

Interactions often very sparse; rewards depend only on neighbors.

Represent, solve games graphically

(Kearns et al. 01; Koller & Milch 01; Ortiz & Kearns 03; Vickrey & Koller 02; etc.).



Learning in Games: Mismatches

Supervised learning:

- fixed distribution on inputs to a fixed target function
- or slow "drift"

Clustering or unsupervised learning:

- static data distribution

Reinforcement learning:

- Markovian state-based models

Assumptions severely violated when learning in games:

- dynamics continually changing due to the others' choices
- not in accordance with any statistical model

B. Repeated Games

- You can't learn playing the game once.

Candy Game!

- Repeatedly playing a game raises new questions.

- How many times? Is this common knowledge?

Finite horizon Infinite horizon

- How evaluate overall performance?
- Trade off present and future reward?

- Average Reward: $\lim_{T \rightarrow \infty} 1/T \sum_{t=1}^T r_t$
- Discounted: $\sum_{t=1}^T \gamma^t r_t$

Fictitious Play

Pretend to play a repeated game to find an equilibrium.

1. Initialize $C_i(a_i)$ to count number of plays of a_i by i .
2. Repeat
 1. Choose $a_i = \operatorname{argmax}_{a_i} R_i(a_i, C_{-i})$.
 2. Increment $C_i(a_i)$.

Play best responses to experience (Brown 49; Robinson 51).

If C_i converges, it must be a Nash (Fudenberg & Levine 98).

Converges in 2-player 2-action games, zero-sum, “dominance solvable”, but not in general.

Best Response Learners

Learner seeks max reward assuming fixed environment.

Q learning (Sen et al. 94; Claus & Boutilier 98; Tan 93; Crites & Sandholm 95; Bowling 00; Tesauro 95; Uther 97).

- Learns values, selects actions that seem successful.

Gradient Ascent (Williams 93; Sutton et al., 00; Baxter & Bartlett 00; Singh et al. 00; Bowling & Veloso 02, 03; Zinkevich 03).

- Explicit policy, adjusted based on reward feedback.

Learn coordination policies (Claus & Boutilier 98; Wong & Sandholm 02; Young 93; Brafman & Tannenholtz 02, 03).

- Players need to move together.

Let's Try It

Matching Pennies (start with 1.5, 2)

	H	T
H	1,-1	-1,1
T	-1,1	1,-1

Shapley Game

0,0	1,0	0,1
0,1	0,0	1,0
1,0	0,1	0,0

Gradient Dynamics

Two player, two action game.

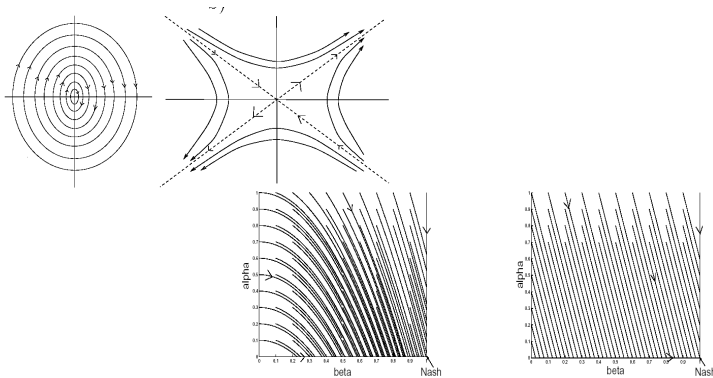
Players keep probability distribution over its two actions.

Modified on each step to maximize reward (gradient) modified to remain a probability.

$$\frac{\partial V_r(\alpha, \beta)}{\partial \alpha} = \beta u - (r_{22} - r_{12})$$

$$\frac{\partial V_c(\alpha, \beta)}{\partial \beta} = \alpha u' - (c_{22} - c_{12})$$

Finite Set of Possibilities



Results

If both players follow the IGA rule, then both player's average payoffs will converge to the expected payoff of some Nash equilibrium

If the strategy pair trajectory converges at all, then it converges to a Nash pair.

Only for 2 actions, recently generalized by Zinkovich.

Nash Equilibria in Repeated PD

Repeating a one-shot equilibrium is an equilibrium.

But, including history-dependent strategies adds more...

R_1	C	D		R_2	C	D
C	3	0]	C	3	4
D	4	1]	D	0	1

PD, average reward. Tit-for-tat (Axelrod 84). Echo choice.

- Always D: 1
- Alternate C-D-C-D..., or D-C-D-C...: 2.
- Always C or TFT: 3. TFT best response: Nash!

Folk Theorem

For any repeated game with average reward, every *feasible* and *enforceable* vector of payoffs for the players can be achieved by some Nash equilibrium strategy. (Osborne & Rubinstein, 94)

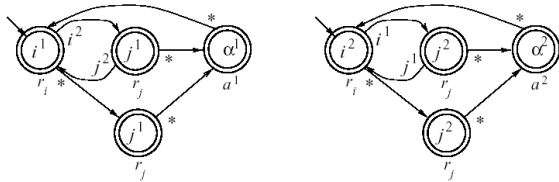
- A payoff vector is *feasible* if it is a linear combination of individual action payoffs.
- A payoff vector is *enforceable* if all players get at least their minimax value.

Enforcement comes from threat of punishment, as TFT.

Algorithmic Application

For any two-player repeated game under the average-reward criterion, a Nash equilibrium pair of controllers can be synthesized in poly time. (Littman & Stone 03)

Defines feasible payoff. If enforceable, creates FSMs. If not, can use modification of minimax solution.



Related Problems

Can be done efficiently for more than 2 players?

Take advantage of graphical structure?

How solve distributed two-player game? (Need to coordinate threats in the Nash setting.)

C. Stochastic Games

Sequential decision making

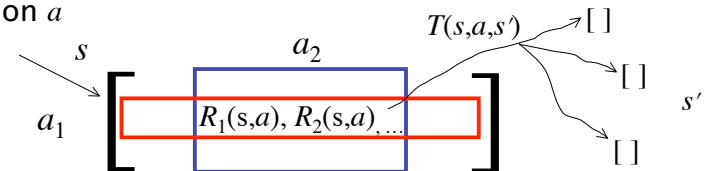
- Repeated games
 - multiple decision makers
 - one state
- Markov decision processes (Bellman 57; Puterman 94)
 - one decision maker
 - multiple states
- Stochastic games (Markov games) (Shapley 53)
 - multiple decision makers
 - multiple states

Notation

S : Set of states

$R_i(s,a)$: Reward to player i in state s under joint action a

$T(s,a,s')$: Probability of transition from s' to state s on a



From dynamic programming approach:

$Q_i(s,a)$: Long-run payoff to i from s on a then equilibrium

Computing Equilibria

Q functions exist (at least for discounted).

(Fink 64)

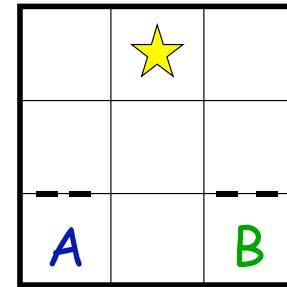
At least as hard as normal-form games.

Values can be irrational (Littman 96), even if

- inputs rational
- transitions deterministic
- two-player zero sum payoffs

Value iteration converges for zero-sum games.

Example: Grid Game 3



(Hu & Wellman 01)

U, D, R, L, X

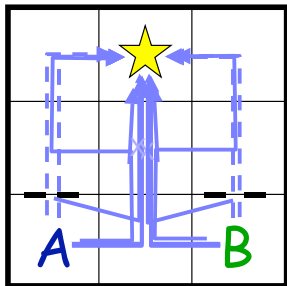
No move on collision

Semiwalls (50%)

-1 for step, -10 for collision, +100 for goal, 0 if back to initial config.

Both can get goal.

Choices in Grid Game



see: Hawks/Doves, Traffic, "chicken"

Average reward:

- (32.3, 16.0) , C, S
- (16.0, 32.3) , S, C
- (-1.0, -1.0) , C, C
- (15.8, 15.8) , S, S
- (15.9, 15.9) , mix
- (25.7, 25.8) , L, F
- (25.8, 25.7) , F, L

"Equilibrium" Learners

Extensions of Q-learning (Watkins 89) to joint actions

$$Q(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q(s', a')$$

Minimax-Q (Littman 94): Converges, zero-sum equil. (Littman & Szepesvári 96)

$$Q_i(s, a) = R_i(s, a) + \sum_{s'} T(s, a, s') \min_{a'_i} Q_i(s', a')$$

Nash-Q (Hu & Wellman 98): Applies to general-sum scenarios; heuristic.

$$Q_i(s, a) = R_i(s, a) + \sum_{s'} T(s, a, s') \text{Nash}_{a', Q_{-i,i}} Q_i(s', a')$$

Friend-or-Foe-Q (Littman 01): Opponents friends (use max) or foes (use minimax); converges, equilibria if saddlepoint/global optima.

CE-Q (Hall & Greenwald 03): Use correlated equilibria; empirically equil.

$$Q_i(s, a) = R_i(s, a) + \sum_{s'} T(s, a, s') \text{CE}_{a', Q_{-i,i}} Q_i(s', a')$$

SG Analogies to MDPs

In the zero-sum case, results analogous to MDPs:

- optimal value function, policy, Q function
- can be found via simulation, search, DP (not LP!)
- can define convergent Q-learning-like algorithm

Failed analogies for general-sum games:

- equilibrium value function need not be unique
- Q-learning-like algorithms don't generally converge
- values not sufficient to specify policy
- no efficient algorithm known

Active area of research. What's the right thing to do?

The Q-value "Program"

These algorithms seek Q functions.

In general sum games, are Q functions enough?

If so, equilibrium policies when only one player moves in a state would necessarily be deterministic.

Are they?

No.

Q values are not sufficient; we need another approach.

R-MAX (Brafman & Tennenholtz 02)

At the beginning, mark all states as "unlearned".

Compute and execute optimal T-step policy.

Use model in which "unlearned" states have maximum reward. ("Optimism under uncertainty.")

Learn the model as you go. Once

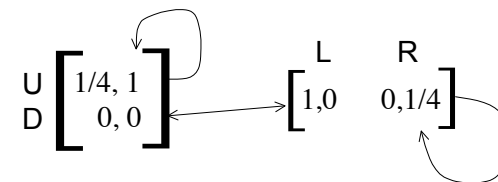
$$K_1 = \max((4NTR_{max})/\epsilon^3, -6 \ln^3(\delta/6Nk^2))+1$$

visits accrue, mark the state as "learned".

Theorem: With prob. $1-\delta$, R-MAX achieves reward within 2ϵ of optimal after polynomially many timesteps.

For zero-sum Markov games, repeated games, MDPs.

Marty's Game



Row U: Column chooses L (~ 1) over R ($\sim 1/4$).

Column L: Row chooses D (~ 1) over U ($\sim 1/4$).

Row D: Column chooses R ($\sim 1/4$) over L (0).

Column R: Row chooses U ($\sim 1/4$) over D (0).

No deterministic equilibrium policy.

Emerging Connections

Behavioral game theory (Camerer 03)

- Is equilibrium behavior “rational”?
- Depends on what others do. (Not Nash!)
- Yet, fairly systematic; need to understand this.
- Nobel prize in Economics '02 (Smith, Kahneman)

Neuroeconomics

- Evidence that cooperation is rewarding. (Angier 02, NYT)

Social network theory

- *Small worlds* interactions. (Kleinberg 00; Watts & Strogatz 98)

Current Research

Shoham: “If Nash-Q is the answer, what is the question?”

What is research in multiagent learning really trying to do?

- compute equilibria?
- win against general opponents?
- design agents that get along with each other?
- understand how people learn?