Nash Q-Learning for General-Sum Stochastic Games

Hu, J. and Wellman, M.P.

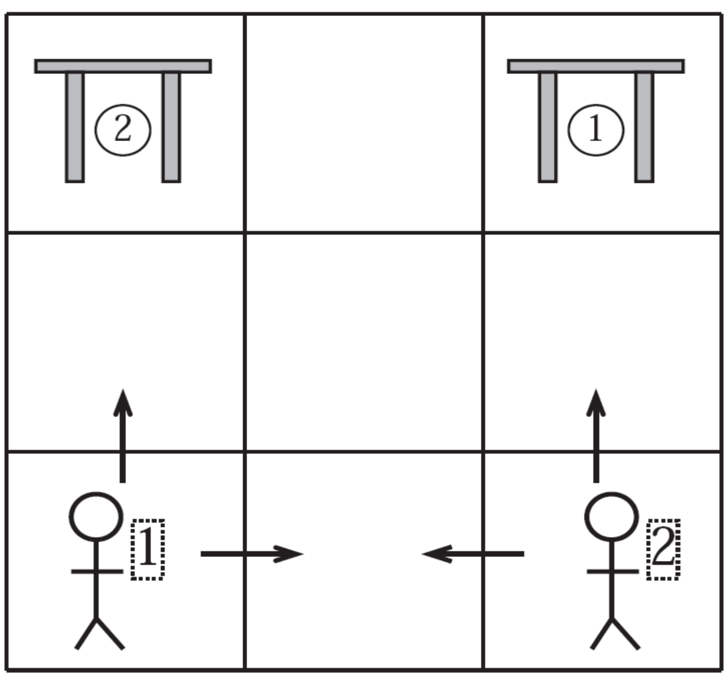
Lauri Lyly, llyly@cc.hut.fi

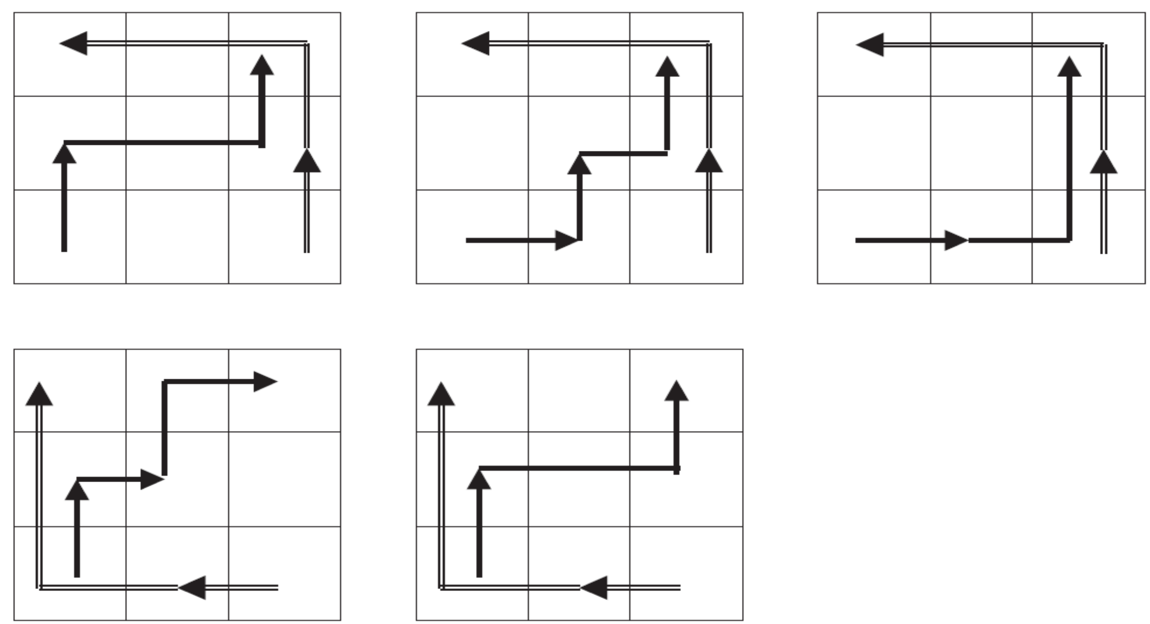
Stochastic general-sum games

- Stochasticity: Environment is in part formed by other agents
  - nondeterministic, noncooperative nature, no agreements
- Arbitrary relation between agents' rewards
  - Extends last time's topic zero-sum

Nash Equilibrium

- Best-response joint strategy
- Study limited to stationary strategies (policies)
- Rewards of others are perceived, strategies are not





Nash Q-Values

$$Q_*^i(s, a^1, \ldots, a^n) = r^i(s, a^1, \ldots, a^n) + \beta \sum_{s' \in S} p(s' | s, a^1, \ldots, a^n) v^i(s', \pi_*^1, \ldots, \pi_*^n)$$

	Multiagent	**Single-Agent**
Q-function	$Q(s, a^1, \ldots, a^n)$	$Q(s, a)$
"Optimal" Q-value	Current reward + Future rewards when all agents play specified Nash equilibrium strategies from the next period onward	Current reward + Future rewards by playing the optimal strategy from the next period onward

Definitions of Q-values

Stage game

- One-period game as opposed to stochastic

Let σ^{-k} be the product of strategies of all agents other than k, $\sigma^{-k} \equiv \sigma^1 \cdots \sigma^{k-1} \cdot \sigma^{k+1} \cdots \sigma^n$.

- Mainly used in convergence proof
- Nash equilibrium for the stage game. M is a “payoff function”:

$$\sigma^k \sigma^{-k} M^k \geq \hat{\sigma}^k \sigma^{-k} M^k \quad \text{for all } \sigma^k \in \hat{\sigma}(A^k).$$

Update rule

- Same update rule for agent itself and its conjecture on other agent's Q-functions
- Q-functions can be initialized for example to 0
- Asynchronous updating: only entries pertaining to current state are updated

$$Q_{t+1}^{i}(s, a^{1}, \ldots, a^{n}) = (1-\alpha_{t})\,Q_{t}^{i}(s, a^{1}, \ldots, a^{n}) + \alpha_{t}\left[r_{t}^{i} + \beta NashQ_{t}^{i}(s')\right]$$

$$NashQ_{t}^{i}(s') = \pi^{1}(s') \cdots \pi^{n}(s') \cdot Q_{t}^{i}(s')$$

The Nash Q-learning algorithm

Initialize:

Let $t = 0$, get the initial state s_0.

Let the learning agent be indexed by i.

For all $s \in S$ and $a^j \in A^j$, $j = 1, \ldots, n$, let $Q_t^j(s, a^1, \ldots, a^n) = 0$.

Loop

Choose action a_t^i.

Observe $r_t^1, \ldots, r_t^n; a_t^1, \ldots, a_t^n$, and $s_{t+1} = s'$

Update Q_t^j for $j = 1, \ldots, n$

$$Q_{t+1}^j(s, a^1, \ldots, a^n) = (1 - \alpha_t) Q_t^j(s, a^1, \ldots, a^n) + \alpha_t [r_t^j + \beta NashQ_t^j(s')]$$

where $\alpha_t \in (0, 1)$ is the learning rate, and $NashQ_t^k(s')$ is defined in (7)

Let $t := t + 1$.

Convergence proof requirements

- **Assumption 1**: Every state-action tuple is visited infinitely often
- **Assumption 2**: Learning rate alpha(t) satisfies:
  - Sum from goes towards infinity
  - Squared sum does not
  - alpha = 0 if the element being updated doesn't correspond to current state-action tuple (asynchronous updating)

Proof basis and result

- Q-learning process updated by pseudo-contraction operator using the usual form:
  Q=(1-alpha(t))Q(t)+alpha(t)[P(t)Q(t)]
  Contraction: Values approach optimal Q
- Link between stage games and stochastic games
- Goal: Show that NashQ is a pseudo-contraction operator
- Actually a real contraction operator in restricted conditions
  - Game with special types of Nash equilibrium points

Different Nash equilibria

- Global optimal point using stage game notation

$$\sigma M^k \geq \hat{\sigma} M^k \quad for\ all\ \hat{\sigma} \in \sigma(A).$$

- Saddle point

$$\begin{aligned} \sigma^k \sigma^{-k} M^k &\geq \hat{\sigma}^k \sigma^{-k} M^k \quad for\ all\ \hat{\sigma}^k \in \sigma(A^k), \\ \sigma^k \sigma^{-k} M^k &\leq \sigma^k \hat{\sigma}^{-k} M^k \quad for\ all\ \hat{\sigma}^{-k} \in \sigma(A^{-k}) \end{aligned}$$

- All equilibria chosen for update must be same type

Stage games compared

Global optimal point (*Up*, *Left*)

(Q_t^1, Q_t^2)	*Left*	*Right*
Up	10, 9	0, 3
Down	3, 0	-1, 2

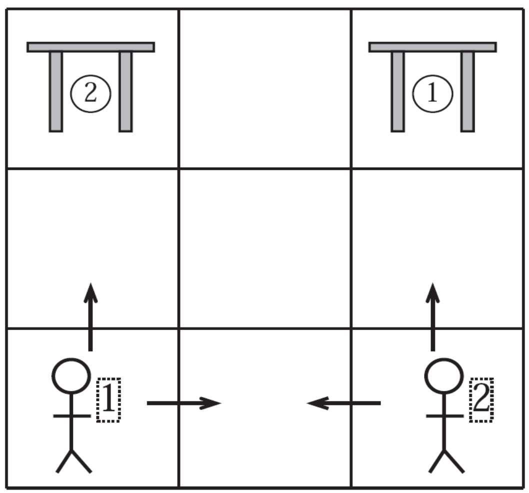
Saddle point (*Down*,*Right*)

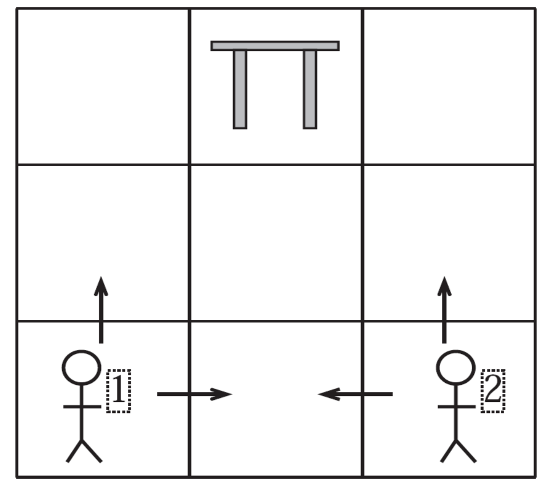
(Q_*^1, Q_*^2)	*Left*	*Right*
Up	5, 5	0, 6
Down	6, 0	2, 2

Figure 2: Two stage games with different types of Nash equilibria

Experimentation framework

- Two grid-world games





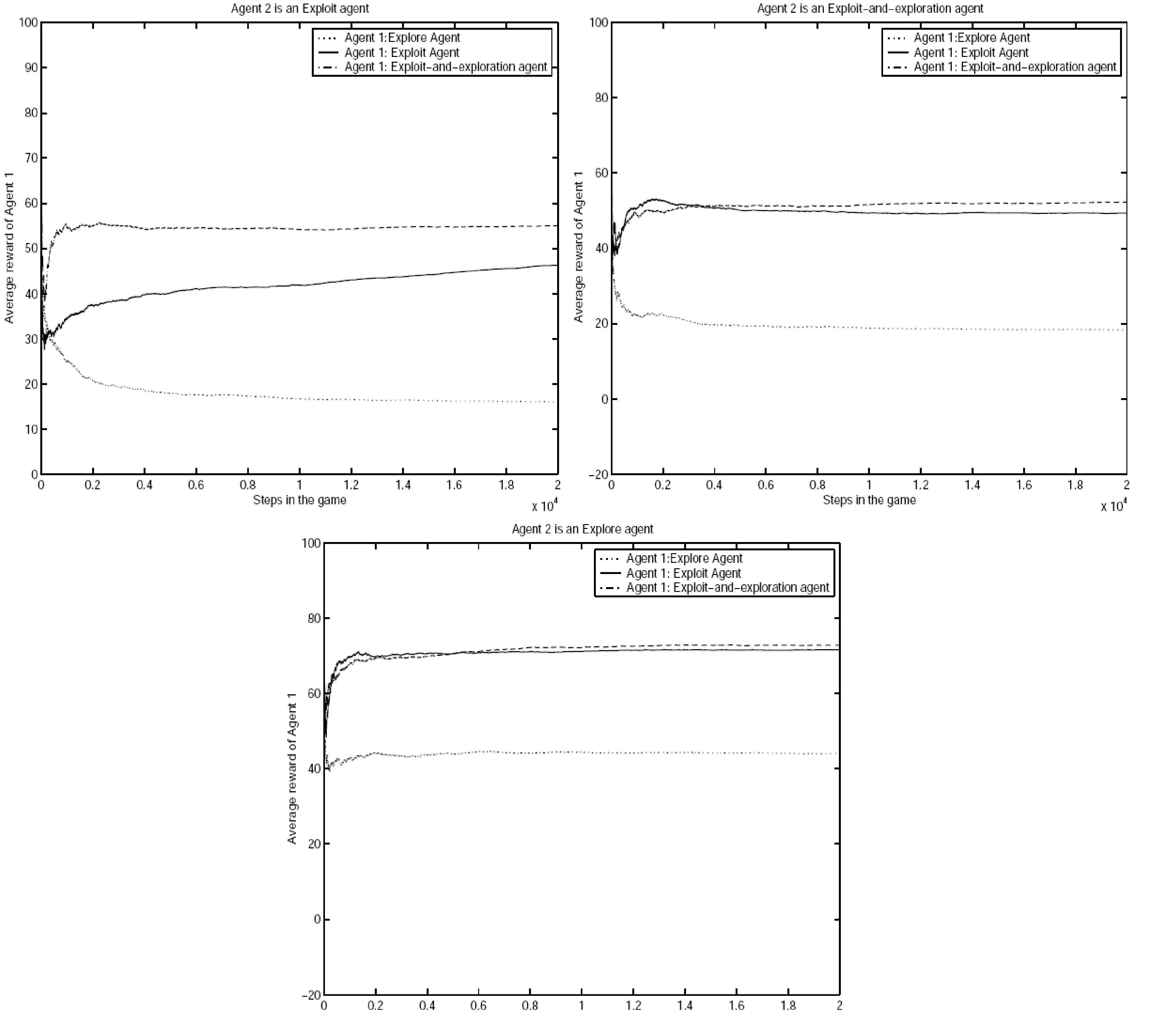
- Motivation for grid games: state-specific actions, qualitative transitions, immediate and long-term rewards

Learning process

- Violates assumption 3 of monotonic selection of global optima or saddle points.
- Still converges in most cases regardless of selection
- Offline and online learning rated separately

Learning strategy		Results of learning
Agent 1	Agent 2	Percent that reach a Nash equilibrium
Single	Single	20%
Single	First Nash	60%
	Second Nash	50%
	Best Expected Nash	76%
First Nash	Second Nash	60%
	Best Expected Nash	76%
Second Nash	Best Expected Nash	84%
Best expected Nash	Best Expected Nash	100%
First Nash	First Nash	100%
Second Nash	Second Nash	100%

Table 11: Learning performance in Grid Game 1



Agent 2 is an Exploit agent
Agent 1:Explore Agent
Agent 1: Exploit Agent
Agent 1: Exploit-and-exploration agent
Average reward of Agent 1
Steps in the game
x 10^4
Agent 2 is an Exploit-and-exploration agent
Agent 1:Explore Agent
Agent 1: Exploit Agent
Agent 1: Exploit-and-exploration agent
Average reward of Agent 1
Steps in the game
x 10^4
Agent 2 is an Explore agent
Agent 1:Explore Agent
Agent 1: Exploit Agent
Agent 1: Exploit-and-exploration agent
Average reward of Agent 1

Conclusions

- No current method provides performance guarantees for general-sum stochastic games
- Works as a starting point
  - Other promising variants
  - Nash equilibrium itself can be refined

References

- [7] Hu, J. and Wellman, M.P. (2003). Nash Q-Learning for General-Sum Stochastic Games. Journal of Machine Learning Research 4:1039–1069.