Lecture 24: Stochastic Game with Nash Equilibrium Concept

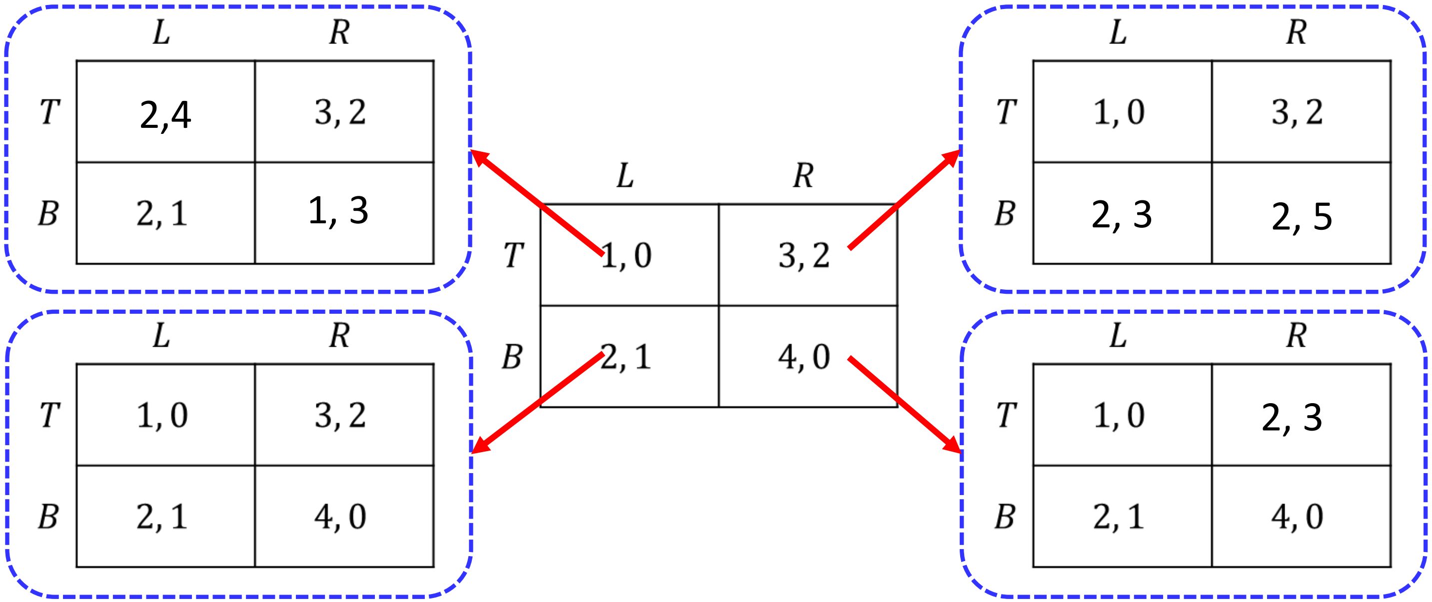
Motivations

What if we didn't always repeat back to the same stage game?

- A stochastic game is a generalization of repeated games
  - agents repeatedly play games from a set of normal-form games
  - the game played at any iteration depends on the previous game played and on the actions taken by all agents in that game

What if there are multiple decision makers in Markov Decision Process?

- A stochastic game is a generalized Markov decision process
  - there are multiple players one reward function for each agent
  - the state transition function and reward functions depend on the action choices of both players



- Stochastic game is a moral general setting where learning is taking place
  - The game transits to another game depending on the joint actions by agents
  - Same players and same actions sets are used through games

- Most of the techniques discussed in the context of repeated games are applicable more generally to stochastic games
  - ✓ specific results obtained for repeated games do not always generalize.

Formal Definition

Definition (Stochastic game)

A stochastic game is a tuple (N, S, A, R, T), where

- N is a finite set of n players
- S is a finite set of states (stage games),
- $A = A_1 \times \cdots \times A_n$, where A_i is a finite set of actions available to player i,
- $T : S \times A \times S \longmapsto [0,1]$ is the transition probability function; $T(s, a, s')$ is the probability of transitioning from state s to state s' after joint action a,
- $R = r_1 \dots, r_n$, where $r_i : S \times A \longmapsto \mathbb{R}$ is a real-valued payoff function for player i

Transition model

- All agents $(1, \dots, n)$ share the joint state s
- The transition equation is similar to the Markov Decision Process decision transition:

MDP : $\sum_{s'} T(s, a, s') = \sum_{s'} p(s'|a, s) = 1, \forall s \in S, \forall a \in A$

SG: $\sum_{s'} T(s, a_1, \dots, a_i, \dots, a_n, s') = \sum_{s'} p(s'|a_1, \dots, a_i, \dots, a_n, s) = 1$

$\forall s \in S, \forall a_i \in A_i, i = (1, \dots, n)$

Reward function

- Reward function r_i for agent i depends on the current joint state s, the joint action $a = (a_1, \dots, a_n)$, and the next joint future state s'

MDP : $r(s, a, s')$

SG: $r_i(s, a_1, \dots, a_i, \dots, a_n, s')$

Policy

	L	R
T	1, 0	3, 2
B	2, 1	4, 0

State 1 : s^1

	L	R
T	2,4	3, 2
B	2, 1	1, 3

State 2 : s^2

	L	R
T	1, 0	3, 2
B	2, 1	4, 0

State 3 : s^3

	L	R
T	1, 0	3, 2
B	2, 3	2, 5

State 4 : s^4

	L	R
T	1, 0	2, 3
B	2, 1	4, 0

- Policy π_1 will give the action that will be taken by player 1 at a given state (stage game):

$$a_1 = \pi_1(s),\ a_1 \in \{T, B\}$$

Value function

- As we did in MDP, we can define value function
- Let π_i be the policy of player $i \in N$. For a given initial state s, the value of state s for player i is defined as

$$V_i(s, \pi_1, \dots, \pi_i, \dots, \pi_n) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1, \dots, \pi_i, \dots, \pi_n, s_0 = s\right]$$

  - The accumulated rewards depends on the policies of other agents
  - The immediate reward is expressed as expected value, because some policy π_i can be stochastic

- In a *discounted stochastic game*, the objective of each player is to maximize the discounted sum of rewards, with discount factor $\gamma \in [0,1)$.

Equilibrium strategy

Definition (Nash equilibrium policy in Stochastic game)

In a stochastic game $\Gamma = (N, S, A, R, T)$, a Nash equilibrium policy is a tuple of n policies $\pi^* = (\pi_1^*, \ldots, \pi_n^*)$ such that for all $s \in S$ and $i = 1, \ldots n$,

$$V_i(s, \pi_1^*, \ldots, \pi_i^*, \ldots, \pi_n^*) \geq V_i(s, \pi_1^*, \ldots, \pi_i, \ldots, \pi_n^*) \text{ for all } \pi_i \in \Pi_i$$

- A Nash equilibrium is a joint policy where each agent's policy is a best response to the others
- For a stochastic game, each agent's policy is defined over the entire time horizon of the game
- **A Nash equilibrium state value** $V_i(s, \pi_1^*, \ldots, \pi_n^*)$ is defined as the sum of discounted rewards when all agents following the Nash equilibrium policies $\pi^* = (\pi_1^*, \ldots, \pi_n^*)$
  - Notations: $V_i^*(s) = V_i^{\pi^*}(s) = V_i(s, \pi_1^*, \ldots, \pi_n^*)$

Equilibrium policy

Theorem (Fink 1964)

Every n −player discounted stochastic game processes at least one Nash equilibrium policy in stationary policies

- Action selection rule for non-stationary policy is different depending on time
  - $\pi_t(s) \neq \pi_{t+1}(s)$

- There are generally a great multiplicity of non-stationary equilibria, whose fact is partially demonstrated by Folk Theorems

Optimal Q-values → Nash Q-Values

Single agent

Q-values

$Q^{\pi}(s,a)$: The expected utility of taking action a from state s, and then following policy π

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty}\gamma^{k} r_{t+k} \;\middle|\; S_t = s, A_t = a\right)$$

Optimal Q-values

$$Q^{*}(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

$$= \max_{\pi} \mathbb{E}[r(s,a,s') + \gamma V^{\pi}(s') | s_t = s, a_t = a]$$

$$= \mathbb{E}\left[r(s,a,s') + \gamma \max_{\pi} V^{\pi}(s') \,|\, s_t = s, a_t = a\right]$$

$$= \mathbb{E}[r(s,a,s') + \gamma V^{*}(s') | s_t = s, a_t = a] \qquad \because V^{*}(s') \equiv \max_{\pi} V^{\pi}(s')$$

$$= \mathbb{E}\left[r(s,a,s') + \gamma \max_{a'} Q^{*}(s',a') \,|\, s_t = s, a_t = a\right] \qquad \because V^{*}(s') \equiv \max_{a'} Q^{*}(s',a')$$

Optimization over policy becomes greedy optimization over action!

- Optimal Q-value for a single-agent is the sum of the current reward and future discounted rewards when playing the optimal strategy from the next period onward

Multi agents

Q-values for agent i

$Q_i^{\pi}(s, a_1, \dots, a_n)$: The expected utility of taking joint action $(a_1, \dots, a_n)$ from state s, and then following policy π

$$Q_i^{\pi}(s, a_1, \dots, a_n) = \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{i,t+k} \;\middle|\; S_t = s, A = (a_1, \dots, a_n)\right)$$

Optimal Q-values for agent i

$$Q_i^{*}(s, a_1, \dots, a_n) = \max_{\pi_1, \dots, \pi_n} Q_i^{\pi}(s, a_1, \dots, a_n)$$

$$= \max_{\pi_1, \dots, \pi_n} \mathbb{E}[r_i(s, a_1, \dots, a_n, s') + \gamma V_i(s', \pi_1, \dots, \pi_n) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= \mathbb{E}\left[r_i(s, a_1, \dots, a_n, s') + \gamma \max_{\pi_1, \dots, \pi_n} V_i(s', \pi_1, \dots, \pi_n) | s_t = s, a_t = (a_1, \dots, a_n)\right]$$

$$= \mathbb{E}[r_i(s, a_1, \dots, a_n, s') + \gamma V_i(s', \pi_1^*, \dots, \pi_n^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= \mathbb{E}\left[r_i(s, a_1, \dots, a_n, s') + \gamma \max_{a_1, \dots, a_n} Q_i^*(s', a_1, \dots, a_n) \,| s_t = s, a_t = (a_1, \dots, a_n)\right]$$

- Optimal Q-value for agent i occurs when all agents are jointly coordinating to maximize agent i's accumulated reward
  - ➢ Rarely occurs! : Optimal Q-values for all agents are not achieved simultaneously

Multi agents

Q-values for agent i

$Q_i^{\pi}(s, a_1, \dots, a_n)$: The expected utility of taking joint action $(a_1, \dots, a_n)$ from state s, and then following policy π

$$Q_i^{\pi}(s, a_1, \dots, a_n) = \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{i,t+k} \mid S_t = s, A = (a_1, \dots, a_n)\right)$$

Optimal Q-values for agent i

$$Q_i^{*}(s, a_1, \dots, a_n) = \max_{\pi_1,\dots,\pi_n} Q_i^{\pi}(s, a_1, \dots, a_n)$$

$$= \max_{\pi_1,\dots,\pi_n} \mathbb{E}[r_i(s, a_1, \dots, a_n, s') + \gamma V_i(s', \pi_1, \dots, \pi_n) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= \mathbb{E}\left[r_i(s, a_1, \dots, a_n, s') + \gamma \max_{\pi_1,\dots,\pi_n} V_i(s', \pi_1, \dots, \pi_n) | s_t = s, a_t = (a_1, \dots, a_n)\right]$$

$$= \mathbb{E}[r_i(s, a_1, \dots, a_n, s') + \gamma V_i(s', \pi_1^*, \dots, \pi_n^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= \mathbb{E}\left[r_i(s, a_1, \dots, a_n, s') + \gamma \max_{a_1,\dots,a_n} Q_i^{*}(s', a_1, \dots, a_n) | s_t = s, a_t = (a_1, \dots, a_n)\right]$$

- Optimal Q-value for agent i occurs when all agents are jointly coordinating to maximize agent i's accumulated reward
  - Rarely occurs! : Optimal Q-values for all agents are not achieved simultaneously

Multi agents

Q-values for agent i

$Q_i^{\pi}(s, a_1, \dots, a_n)$: The expected utility of taking joint action $(a_1, \dots, a_n)$ from state s, and then following policy π

$$Q_i^{\pi}(s, a_1, \dots, a_n) = \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{i,t+k} \mid S_t = s, A = (a_1, \dots, a_n)\right)$$

Nash Q-values for agent i

$$\begin{aligned}
Q_i^{*}(s, a_1, \dots, a_n) &= \underset{\pi_1,\dots,\pi_n}{\mathrm{Nash}}\ Q_i^{\pi}(s, a_1, \dots, a_n) \\
&= \underset{\pi_1,\dots,\pi_n}{\mathrm{Nash}}\ \mathbb{E}[r_i(s, a_1, \dots, a_n, s') + \gamma V_i(s', \pi_1, \dots, \pi_n) | s_t = s, a_t = (a_1, \dots, a_n)] \\
&= \mathbb{E}\left[r_i(s, a_1, \dots, a_n, s') + \gamma \underset{\pi_1,\dots,\pi_n}{\mathrm{Nash}}\ V_i(s', \pi_1, \dots, \pi_n) | s_t = s, a_t = (a_1, \dots, a_n)\right] \\
&= \mathbb{E}[r_i(s, a_1, \dots, a_n, s') + \gamma V_i(s', \pi_1^*, \dots, \pi_n^*) | s_t = s, a_t = (a_1, \dots, a_n)] \\
&= \mathbb{E}\left[r_i(s, a_1, \dots, a_n, s') + \gamma \underset{a_1,\dots,a_n}{\mathrm{Nash}}\ Q_i^{*}(s', a_1, \dots, a_n) \mid s_t = s, a_t = (a_1, \dots, a_n)\right]
\end{aligned}$$

Equilibrium over policies becomes stage game equilibrium over action!

- A **Nash Q value** $Q_i^{*}(s, a_1, \dots, a_n)$ is the expected sum of discounted rewards when all agents take the joint action $a = (a_1, \dots, a_n)$ at given state s and follow a Nash equilibrium strategy $\pi^* = (\pi_1^*, \dots, \pi_n^*)$

Nash Bellman equation

For single agent:

$$V^*(s') = \max_a Q^*(s', a)$$

$$Q^*(s, a) = \mathbb{E}[r(s, a, s') + \gamma V^*(s') | s_t = s, a_t = a]$$

$$= \mathbb{E}\left[r(s, a, s') + \gamma \max_{a'} Q^*(s', a') \, | s_t = s, a_t = a\right]$$

For multiple agents:

$$V_i(s', \pi_1^*, \dots, \pi_n^*) = \underset{a_1, \dots, a_n}{\text{Nash}} \; Q_i^*(s', a_1, \dots, a_n)$$

$$Q_i^*(s, a_1, \dots, a_n) = \mathbb{E}[r(s, a, s') + \gamma V_i(s', \pi_1^*, \dots, \pi_n^*) | s_t = s, a_t = a]$$

$$= \mathbb{E}\left[r(s, a, s') + \gamma \underset{a_1, \dots, a_n}{\text{Nash}} \; Q_i^*(s', a_1, \dots, a_n) \, | s_t = s, a_t = (a_1, \dots, a_n)\right]$$

How to compute A Nash (equilibrium) state value $V_i(s, \pi_1^*, ..., \pi_n^*)$

For multiple agents:

$$V_i(s', \pi_1^*, \dots, \pi_n^*) = \underset{a_1,\dots,a_n}{\text{Nash}}\ Q_i^*(s', a_1, \dots, a_n)$$

$$Q_i^*(s, a_1, \dots, a_n) = \mathbb{E}[r(s,a,s') + \gamma V_i(s', \pi_1^*, \dots, \pi_n^*) | s_t = s, a_t = a]$$

$$= \mathbb{E}\left[r(s,a,s') + \gamma \underset{a_1,\dots,a_n}{\text{Nash}}\ Q_i^*(s', a_1, \dots, a_n) \,|s_t = s, a_t = (a_1, \dots, a_n)\right]$$

- Nash **equilibrium** Q value $\underset{a_1,\dots,a_n}{\text{Nash}}\ Q_i^*(s', a_1, \dots, a_n)$ can be computed by computing player ith Nash equilibrium value for the stage game $[Q_i^*(s', a_1, \dots, a_n), \dots, Q_n^*(s', a_1, \dots, a_n)]$

  - for example when $i = 1,2$

	a_2^1	a_2^2
a_1^1	$Q_1^*(s', a_1^1, a_2^1), Q_2(s', a_1^1, a_2^1)$	$Q_1^*(s', a_1^1, a_2^2), Q_2(s', a_1^1, a_2^2)$
a_1^2	$Q_1^*(s', a_1^2, a_2^1), Q_2(s', a_1^2, a_2^1)$	$Q_1^*(s', a_1^2, a_2^2), Q_2(s', a_1^2, a_2^2)$

Nash equilibrium

Simplifying Notation

For multiple agents:

$$r_i(s, a_1, \dots, a_n, s') \to r_i(s, \vec{a}, s')$$

$$V_i(s, \pi_1^*, \dots, \pi_n^*) \to V_i^*(s)$$

$$Q_i^*(s, a_1, \dots, a_n) \to Q_i^*(s', \vec{a}\ )$$

$$Q_i^*(s, a_1, \dots, a_n) = \mathbb{E}[r_i(s, a_1, \dots, a_n, s') + \gamma V_i(s', \pi_1^*, \dots, \pi_n^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= \mathbb{E}\left[r_i(s, a_1, \dots, a_n, s') + \gamma \underset{a_1,\dots,a_n}{\text{Nash}}\ Q_i^*(s', a_1, \dots, a_n)\ | s_t = s, a_t = (a_1, \dots, a_n)\right]$$

$$Q_i^*(s', \vec{a}\ ) = \mathbb{E}[r_i(s, \vec{a}, s') + \gamma V_i^*(s') | s_t = s, a_t = \vec{a}]$$

$$= \mathbb{E}[r_i(s, \vec{a}, s') + \gamma\ \text{Nash}\ Q_i^*(s')\ | s_t = s, a_t = \vec{a}]$$

$$\underset{a_1,\dots,a_n}{\text{Nash}}\ Q_i^*(s', a_1, \dots, a_n) = Q_i^*(s', \vec{a}_{NE}) = \text{Nash}\ Q_i^*(s')$$

Computing Nash Q-values analytically

- If we know **Nash equilibrium policy** $\pi^* = (\pi_1^*, \dots, \pi_n^*)$, we can compute the Nash equilibrium state values $V_i(s, \pi_1^*, \dots, \pi_n^*)$ (i.e., policy evaluation)

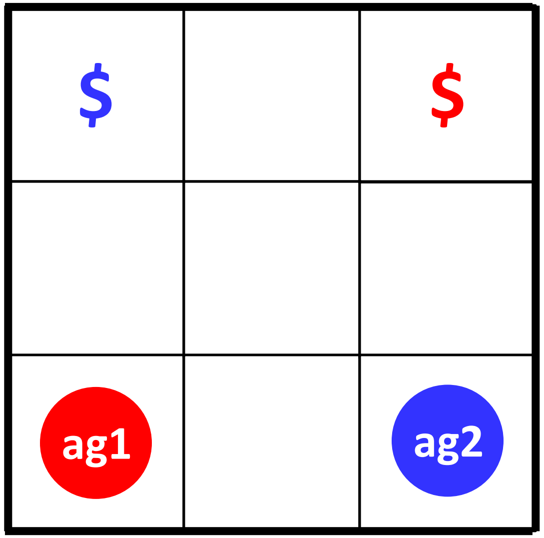
$$V_i(s, \pi_1^*, \dots, \pi_n^*) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1^*, \dots, \pi_n^*, s_0 = s\right]$$

- If we know **Nash equilibrium state value** $V_i(s, \pi_1^*, \dots, \pi_n^*)$ and transition models $p(s'|s, a_1, \dots, a_n)$, we can compute **Nash Q-values (i.e., Nash Q-function)** using backward induction (analytical approach)

$$Q_i^*(s, a_1, \dots, a_n) = \mathbb{E}[r_i(s, a_1, \dots, a_n, s') + \gamma V_i(s', \pi_1^*, \dots, \pi_n^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$

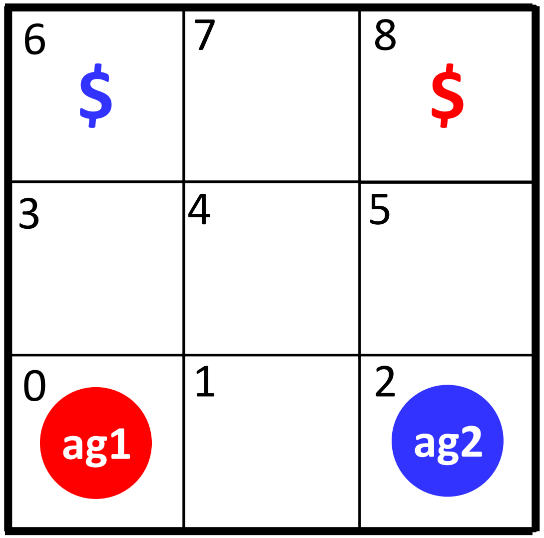
$$= r_i(s, a_1, \dots, a_n, s') + \sum_{s'} p(s'|s, a_1, \dots, a_n) V_i(s', \pi_1^*, \dots, \pi_n^*)$$

Grid Game 1



- Grid game has deterministic moves
- Two agents start from respective lower corners, trying to reach their goal cells in the top row
- Agent can move only one cell a time, and in four possible directions: Left, Right, Up, Down
- If two agents attempt to move into the same cell (excluding a goal cell), they are bounced back to their previous cells
- The game ends as soon as an agent reaches its goal
    - The objective of an agent in this game is therefore to reach its goal with a minimum No. of steps
- Agents do not know
    - the locations of their goals at the beginning of the learning period
    - their own and the other agents' payoff functions
- Agent choose their action simultaneously and observe
    - the previous actions of both agents and the current joint state
    - the immediate rewards after both agents choose their actions

Grid Game 1 represented as stochastic game

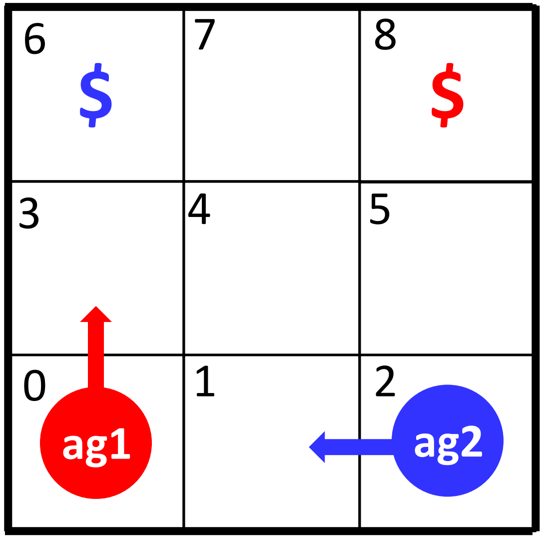


- The action space of agent i, $i = 1{,}2$, is $A_i = \{Left, Right, Down, Up\}$
- The sate space is $S = \{(0{,}1), (0{,}2), \dots, (8{,}7)\}$
  - $s = (l_1, l_2)$ represents the agents' joint location
  - $l_i \in \{0, 2, \dots, 8\}$ is the indexed location
- The reward function is, for $i = 1, 2$,

$$r_i = \begin{cases} 100 & \text{if } L(l_i, a_i) = Goal_i \\ -1 & \text{if } L(l_1, a_1) = L(l_2, a_2) \text{ and } L(l_i, a_i) \neq Goal_i \text{ for } i = 1{,}2 \\ 0 & \text{otherwise} \end{cases}$$

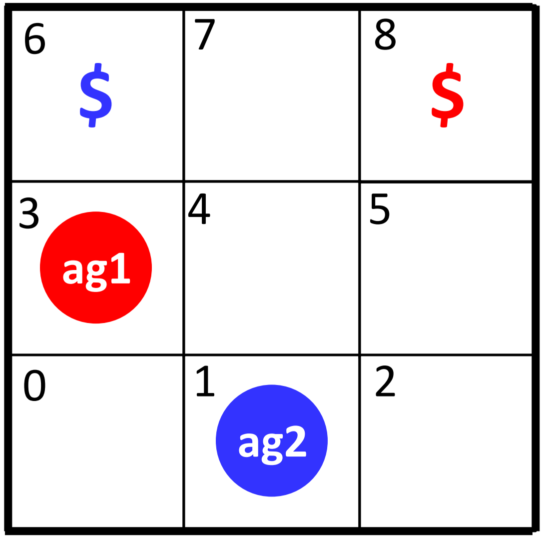
$l_i' = L(l_i, a_i)$ is the next location when executing a_i at l_i

Grid Game 1 represented as stochastic game



- $s = (l_1, l_2) = (0,2)$
- $a = (a_1, a_2) = (Up, Left)$

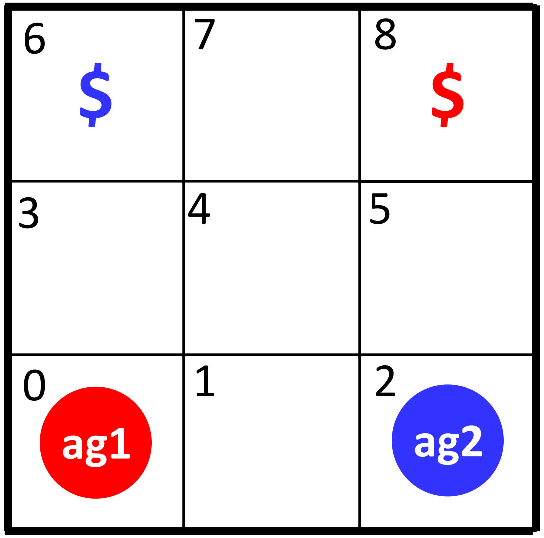
Grid Game 1 represented as stochastic game



- $s = (l_1, l_2) = (0,2)$
- $a = (a_1, a_2) = (Up, Left)$

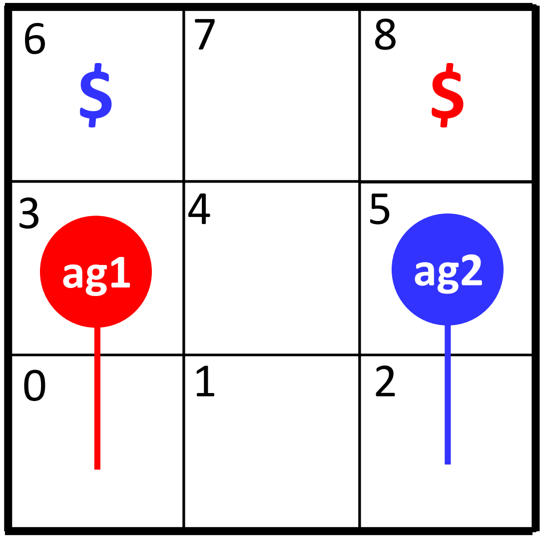
- $s' = \big(L(l_1, a_1), L(l_2, a_2)\big) = (3, 1)$
- $r_1 = 0$
- $r_2 = 0$

Grid Game 1 represented as stochastic game



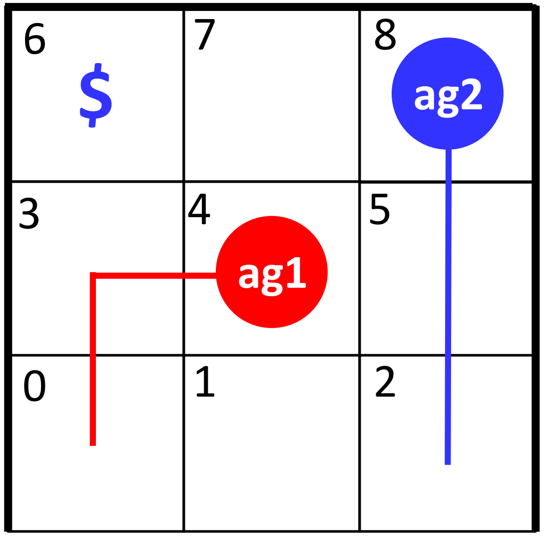
Nash Equilibrium strategies

Grid Game 1 represented as stochastic game



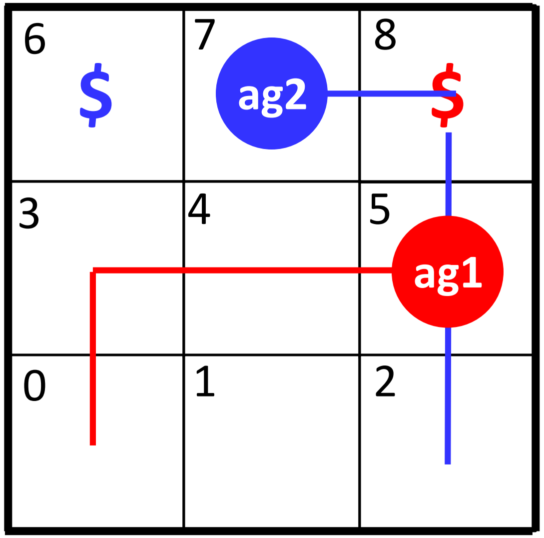
Nash Equilibrium strategies

Grid Game 1 represented as stochastic game



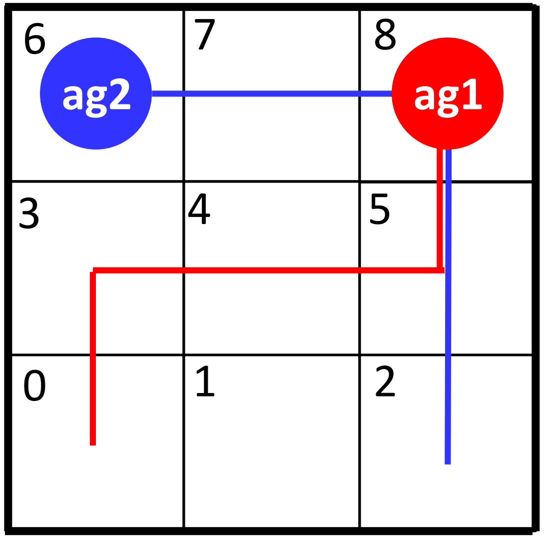
Nash Equilibrium policies

Grid Game 1 represented as stochastic game



Nash Equilibrium policies

Grid Game 1 represented as stochastic game



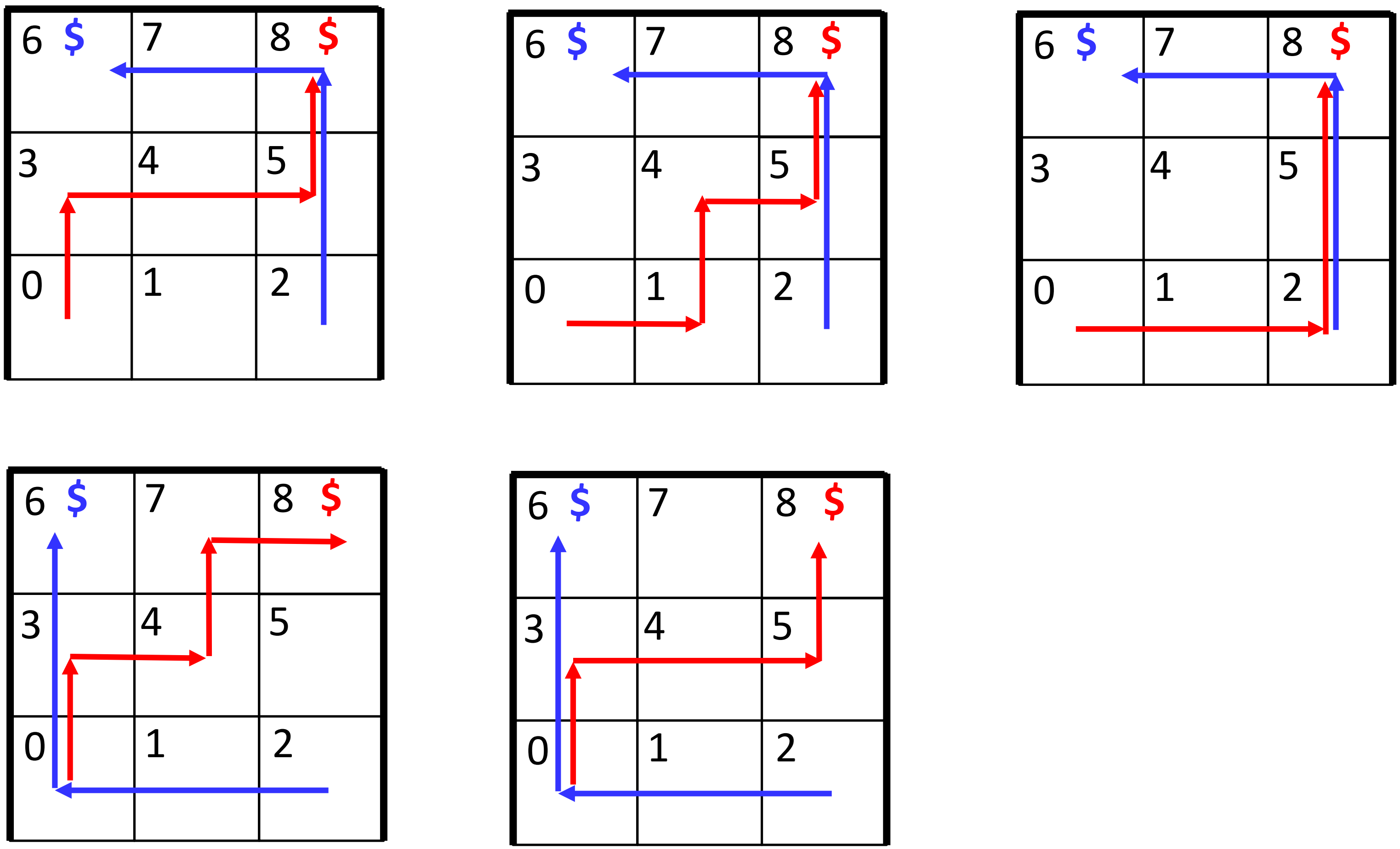
Nash Equilibrium policies

State s	$\pi_1(s)$
$(0, any)$	U
$(3, any)$	$Right$
$(4, any)$	$Right$
$(5, any)$	Up

Nash strategy for agent 1

Grid Game 1 represented as stochastic game

All Nash Equilibrium policies



Grid Game 1 represented as stochastic game

Nash Q values for the initial state $s_0 = (0,2)$

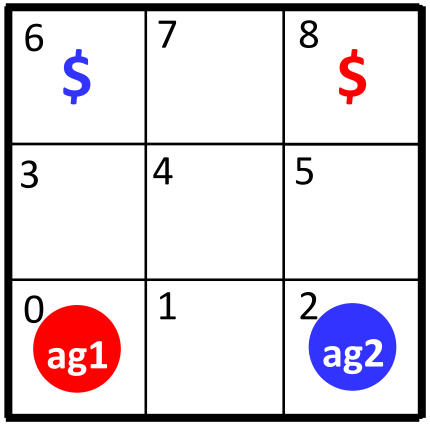
- The value of the game for agent 1 is defined as its accumulated reward when both agents follow their Nash equilibrium polices $\pi^* = (\pi_1^*, \dots, \pi_n^*,)$:

$$V_1(s_0, \pi_1^*, \pi_2^*) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1^*, \dots, \pi_n^*, s_0 = s\right]$$

- In Grid game 1 and initial state $s_0 = (0,2)$, this becomes, given $\gamma = 0.99$,

$$V_1(s_0, \pi_1^*, \pi_2^*) = 0 + 0.99 \times 0 + 0.99^2 \times 0 + 0.99^3 \times 100$$
$$= 97.0$$

$s_0 = (0,2)$



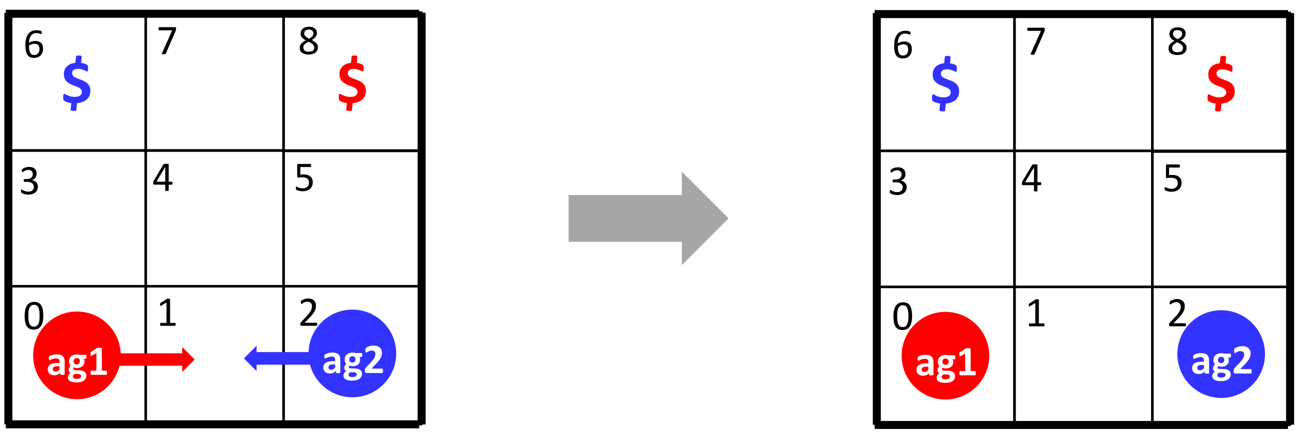
Grid Game 1 represented as stochastic game

Nash Q values for the initial state $s_0 = (0,2)$

$$
\begin{aligned}
Q_1^*(s_0, a_1, a_2) &= \mathbb{E}[r_1(s_0, a_1, a_2) + \gamma V_1(s', \pi_1^*, \pi_2^*) | s_t = s, a_t = (a_1, \dots, a_n)] \\
&= r_1(s_0, a_1, a_2) + \gamma \sum_{s'} p(s'|s_0, a_1, a_2) V_1(s', \pi_1^*, \pi_2^*)
\end{aligned}
$$

$$
\begin{aligned}
Q_1^*(s_0 = (0,2), Right, Left\,) &= -1 + 0.99 \times V_1(s' = (0,2), \pi_1^*, \pi_2^*) \\
&= -1 + 0.99 \times 97 = 95.1
\end{aligned}
$$

$s_0 = (0,2)$ $\qquad\qquad$ $s' = s_0 = (0,2)$

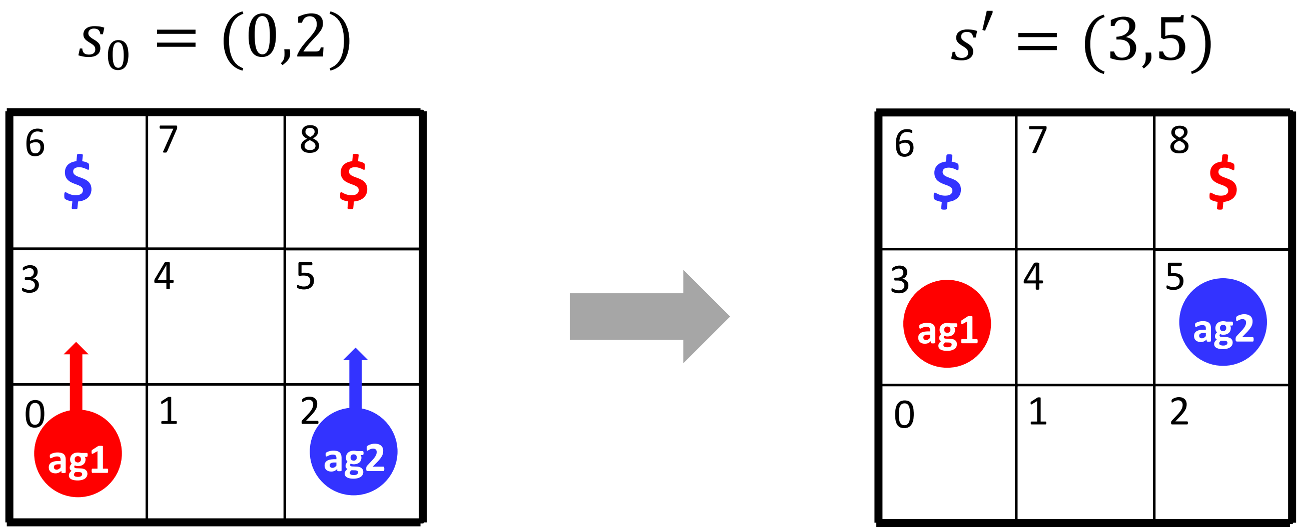


Grid Game 1 represented as stochastic game

Nash Q values for the initial state $s_0 = (0,2)$

$$Q_1^*(s_0, a_1, a_2) = \mathbb{E}[r_1(s_0, a_1, a_2) + \gamma V_1(s', \pi_1^*, \pi_2^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$
$$= r_1(s_0, a_1, a_2) + \gamma \sum_{s'} p(s' | s_0, a_1, a_2) V_1(s', \pi_1^*, \pi_2^*)$$

$$Q_1^*(s_0 = (0,2), Up, Up) = 0 + 0.99 \times V_1(s' = (3,5), \pi_1^*, \pi_2^*)$$
$$= 0 + 0.99 \times \{0 + 0.99 \times 0 + 0.99^2 \times 100\} = 97.0$$



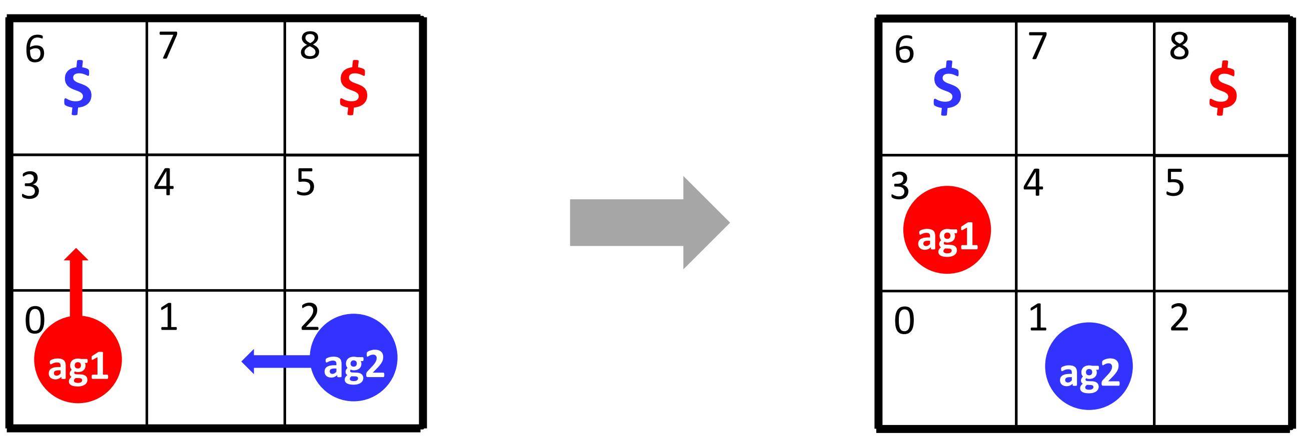
Grid Game 1 represented as stochastic game

Nash Q values for the initial state $s_0 = (0,2)$

$$
\begin{aligned}
Q_1^*(s_0, a_1, a_2) &= \mathbb{E}[r_1(s_0, a_1, a_2) + \gamma V_1(s', \pi_1^*, \pi_2^*) | s_t = s, a_t = (a_1, \dots, a_n)] \\
&= r_1(s_0, a_1, a_2) + \gamma \sum_{s'} p(s' | s_0, a_1, a_2) V_1(s', \pi_1^*, \pi_2^*)
\end{aligned}
$$

$$
\begin{aligned}
Q_1^*(s_0 = (0,2), Up, Left\ ) &= 0 + 0.99 \times V_1(s' = (3,1), \pi_1^*, \pi_2^*) \\
&= 0 + 0.99 \times \{0 + 0.99 \times 0 + 0.99^2 \times 100\} = 97.0
\end{aligned}
$$

$s_0 = (0,2)$ $\qquad\qquad$ $s' = (3,1)$



Grid Game 1 represented as stochastic game

Nash Q values for the initial state $s_0 = (0,2)$

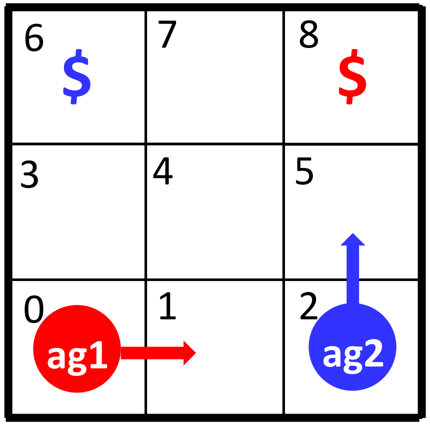
$$Q_1^*(s_0, a_1, a_2) = \mathbb{E}[r_1(s_0, a_1, a_2) + \gamma V_1(s', \pi_1^*, \pi_2^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= r_1(s_0, a_1, a_2) + \gamma \sum_{s'} p(s' | s_0, a_1, a_2) V_1(s', \pi_1^*, \pi_2^*)$$

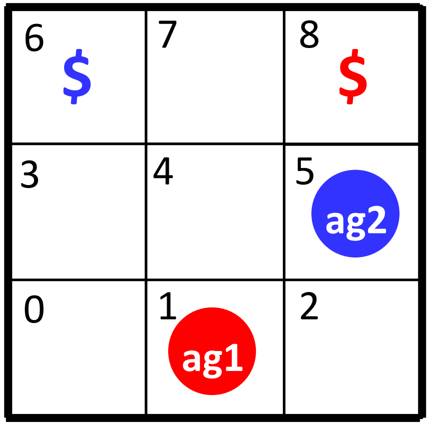
$$Q_1^*(s_0 = (0,2), Right, Up) = 0 + 0.99 \times V_1(s' = (1,5), \pi_1^*, \pi_2^*)$$

$$= 0 + 0.99 \times \{0 + 0.99 \times 0 + 0.99^2 \times 100\} = 97.0$$

$s_0 = (0,2)$



$s' = (1,5)$



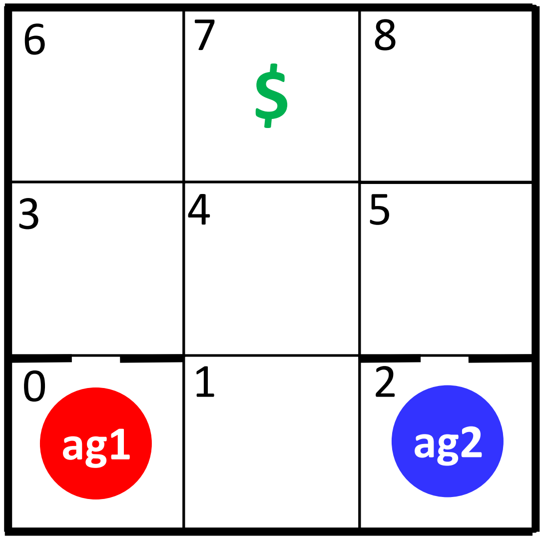
Grid Game 1 represented as stochastic game

Nash Q values for the initial state $s_0 = (0,2)$

	$a_2 = Left$	$a_2 = Up$
$a_1 = Right$	$Q_1^*(s_0, R, L), Q_2^*(s_0, R, L)$	$Q_1^*(s_0, R, U), Q_2^*(s_0, R, U)$
$a_2 = Up$	$Q_1^*(s_0, U, L), Q_2^*(s_0, U, L)$	$Q_1^*(s_0, U, U), Q_2^*(s_0, U, U)$

	$a_2 = Left$	$a_2 = Up$
$a_1 = Right$	$95.1, 95.1$	$97.0, 97.0$
$a_2 = Up$	$97.0, 97.0$	$97.0, 97.0$

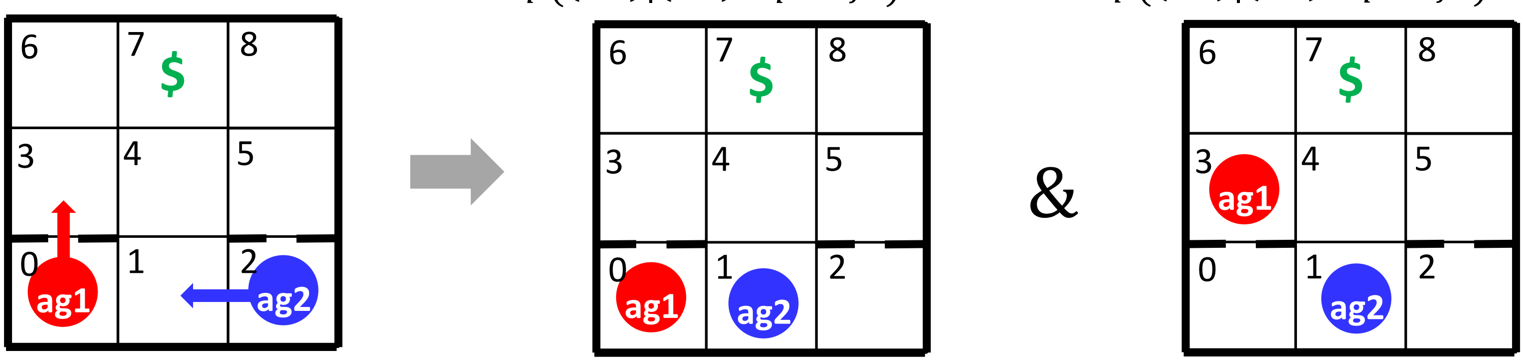
Grid Game 2



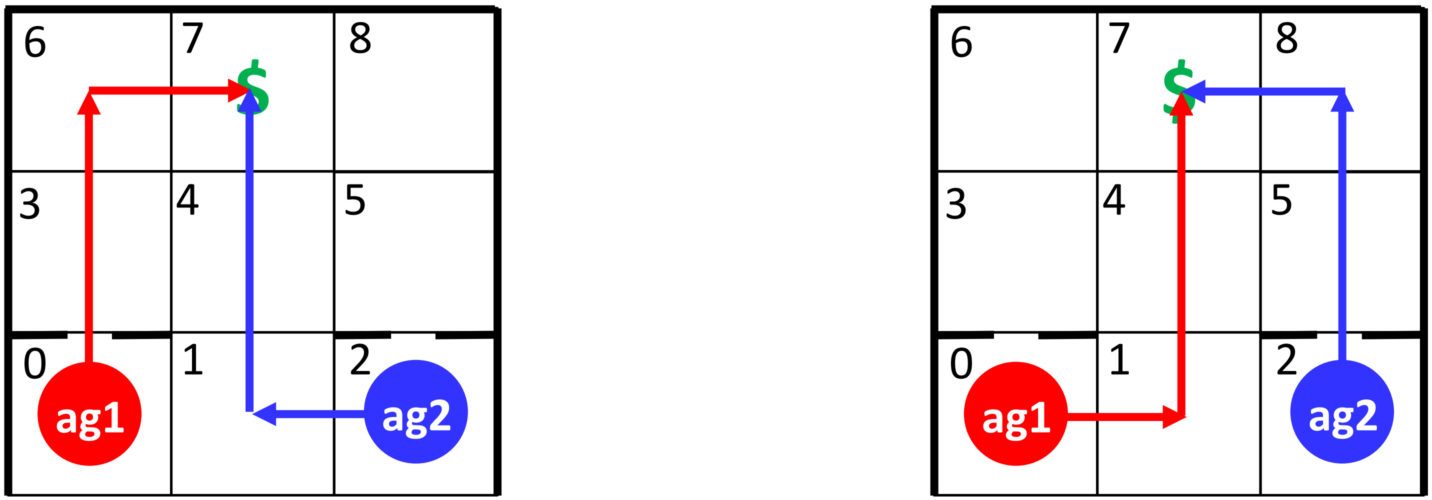
- First to reach goal gets $100
- If both reaches the money at the same time, both win
- Semi wall (50% go through)
- Cannot occupy the same grid

- Grid game has both **stochastic** and **deterministic** moves
- If agent choses Up from position 0 or 2, it moves up with probability 0.5 and remains in its previous position with probability 0.5

$p\big((0{,}1)|(0{,}2), Up, Left\big) = 0.5$ $\quad$ $p\big((3{,}1)|(0{,}2), Up, Left\big) = 0.5$



Grid Game 2



- There are two Nash equilibrium paths
- The value of the game for agent 1 is defined as its accumulated reward when both agents follow their Nash equilibrium strategies $\pi^* = (\pi_1^*, \dots, \pi_n^*)$:

$$V_1(s, \pi_1^*, \pi_2^*) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1^*, \dots, \pi_n^*, s_0 = s\right]$$

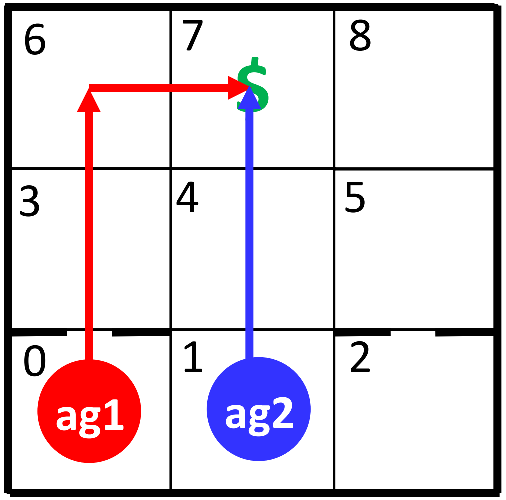
  - $V_1((0,1), \pi_1^*, \pi_2^*) = 0 + 0.99 \times 0 + 0.99^2 \times 0 = 0$
  - $V_1((0,x), \pi_1^*, \pi_2^*) = 0 \text{ for } x = 3, .., 8$
  - $V_1((1,2), \pi_1^*, \pi_2^*) = 0 + 0.99 \times 100 = 99$
  - $V_1((1,3), \pi_1^*, \pi_2^*) = 0 + 0.99 \times 0 + 0.99^2 \times 0 = 0$
  - $V_1((1,x), \pi_1^*, \pi_2^*) = 0 + 0.99 \times 0 + 0.99^2 \times 0 = 0$

Grid Game 2

- The value of the game for agent 1 is defined as its accumulated reward when both agents follow their Nash equilibrium strategies $\pi^* = (\pi_1^*, \ldots, \pi_{n,}^*)$:

$$V_1(s, \pi_1^*, \pi_2^*) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1^*, \ldots, \pi_n^*, s_0 = s\right]$$

  - $V_1((0,1), \pi_1^*, \pi_2^*) = 0 + 0.99 \times 0 + 0.99^2 \times 0 = 0$

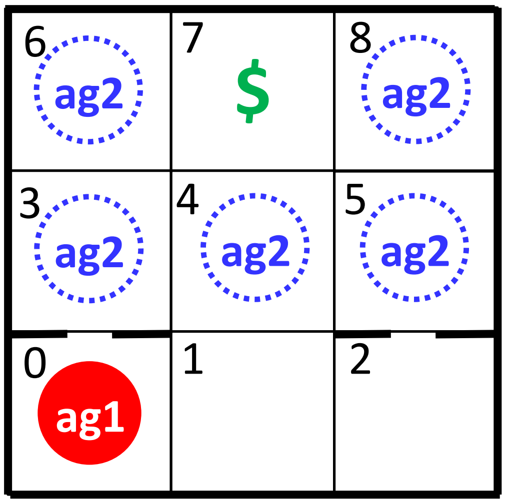


Grid Game 2

- The value of the game for agent 1 is defined as its accumulated reward when both agents follow their Nash equilibrium strategies $\pi^* = (\pi_1^*, \ldots, \pi_n^*,)$:

$$V_1(s, \pi_1^*, \pi_2^*) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1^*, \ldots, \pi_n^*, s_0 = s\right]$$

  - $V_1((0, x), \pi_1^*, \pi_2^*) = 0 \quad \text{for} \;\; x = 3, .., 8$

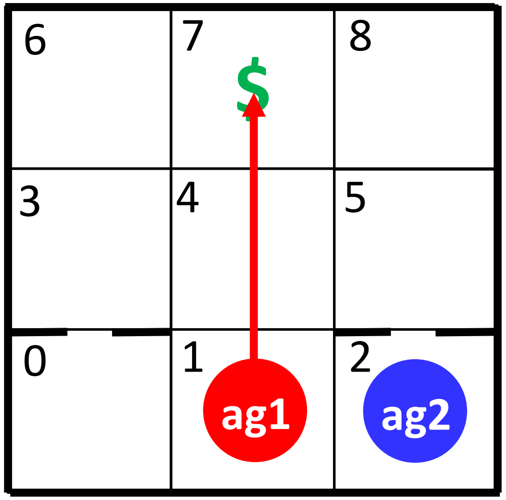


Grid Game 2

- The value of the game for agent 1 is defined as its accumulated reward when both agents follow their Nash equilibrium strategies $\pi^* = (\pi_1^*, \ldots, \pi_{n,}^*)$:

$$V_1(s, \pi_1^*, \pi_2^*) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1^*, \ldots, \pi_n^*, s_0 = s\right]$$

  - $V_1((1,2), \pi_1^*, \pi_2^*) = 0 + 0.99 \times 100 = 99$

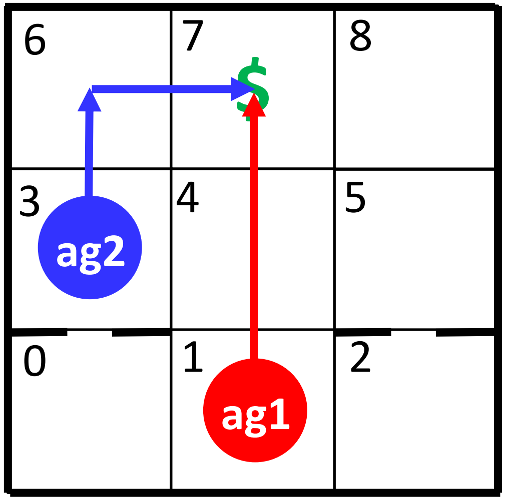


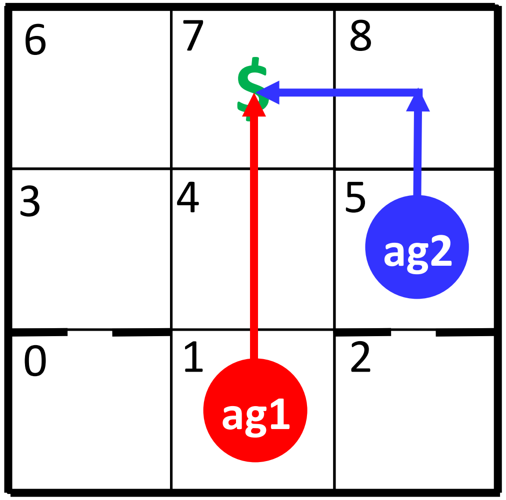
Grid Game 2

- The value of the game for agent 1 is defined as its accumulated reward when both agents follow their Nash equilibrium strategies $\pi^* = (\pi_1^*, \dots, \pi_{n,}^*)$:

$$V_1(s, \pi_1^*, \pi_2^*) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1^*, \dots, \pi_n^*, s_0 = s\right]$$

  - $V_1((1,3), \pi_1^*, \pi_2^*) = 0 + 0.99 \times 100 = 99 = V_1((1,5), \pi_1^*, \pi_2^*)$



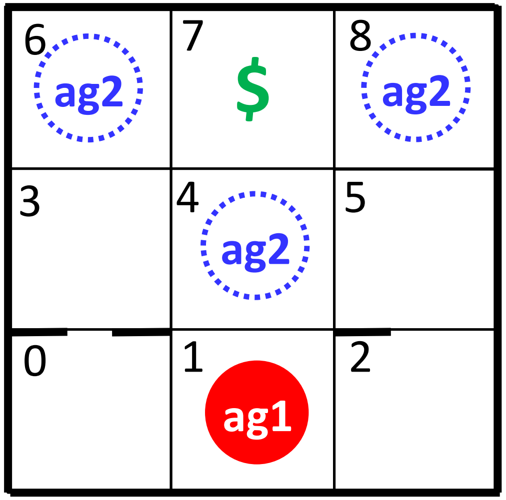


Grid Game 2

- The value of the game for agent 1 is defined as its accumulated reward when both agents follow their Nash equilibrium strategies $\pi^* = (\pi_1^*, \ldots, \pi_{n,}^*)$:

$$V_1(s, \pi_1^*, \pi_2^*) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1^*, \ldots, \pi_n^*, s_0 = s\right]$$

  - $V_1((1, x), \pi_1^*, \pi_2^*) = 0 \quad \text{for} \quad x = 4, 6, 8$

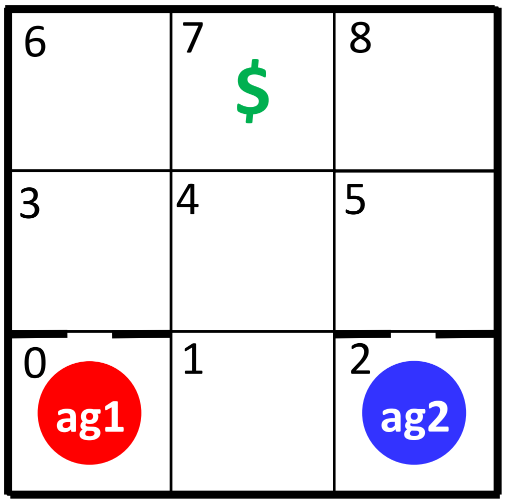


Grid Game 2

- The value of the game for agent 1 is defined as its accumulated reward when both agents follow their Nash equilibrium strategies $\pi^* = (\pi_1^*, \dots, \pi_n^*)$:

$$V_1(s, \pi_1^*, \pi_2^*) = \sum_{t=0}^{\infty} \gamma^t E\left[r_{i,t} | \pi_1^*, \dots, \pi_n^*, s_0 = s\right]$$

  - $V_1((0,2), \pi_1^*, \pi_2^*) = V_1(s_0, \pi_1^*, \pi_2^*)$ can be computed only in expectation
  - We solve $V_1(s_0, \pi_1^*, \pi_2^*)$ from the state game $\left(Q_1^*(s_0, a_1, a_2), Q_2^*(s_0, a_1, a_2)\right)$



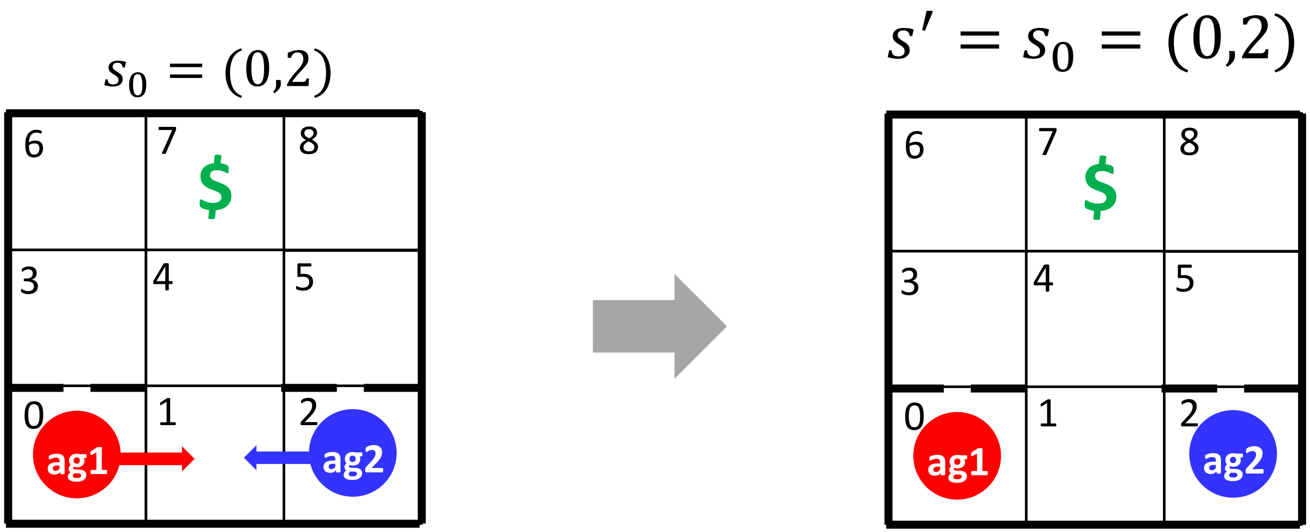
Grid Game 2

Nash Q values for the initial state $s_0 = (0,2)$

$$Q_1^*(s_0, a_1, a_2) = \mathbb{E}[r_1(s_0, a_1, a_2) + \gamma V_1(s', \pi_1^*, \pi_2^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= r_1(s_0, a_1, a_2) + \gamma \sum_{s'} p(s' | s_0, a_1, a_2) V_1(s', \pi_1^*, \pi_2^*)$$

$$Q_1^*(s_0 = (0,2), Right, Left\ ) = -1 + 0.99 \times V_1(s_0, \pi_1^*, \pi_2^*)$$



Grid Game 2

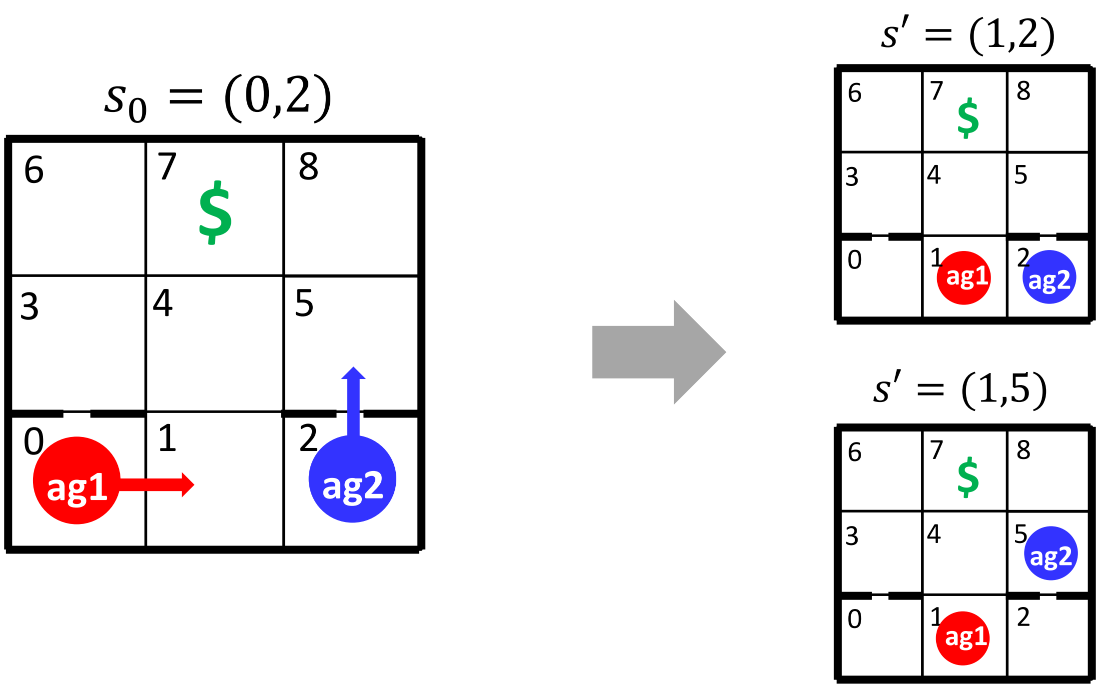
Nash Q values for the initial state $s_0 = (0,2)$

$$Q_1^*(s_0, a_1, a_2) = \mathbb{E}[r_1(s_0, a_1, a_2) + \gamma V_1(s', \pi_1^*, \pi_2^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= r_1(s_0, a_1, a_2) + \gamma \sum_{s'} p(s' | s_0, a_1, a_2) V_1(s', \pi_1^*, \pi_2^*)$$

$$Q_1^*(s_0 = (0,2), Right, Up) = 0 + 0.99 \times \left\{ \frac{1}{2} V_1\left((1,2), \pi_1^*, \pi_2^*\right) + \frac{1}{2} V_1\left((1,5), \pi_1^*, \pi_2^*\right) \right\}$$

$$= 0 + 0.99 \times (0.5 \times 99 + 0.5 \times 99) = 98$$



Grid Game 2

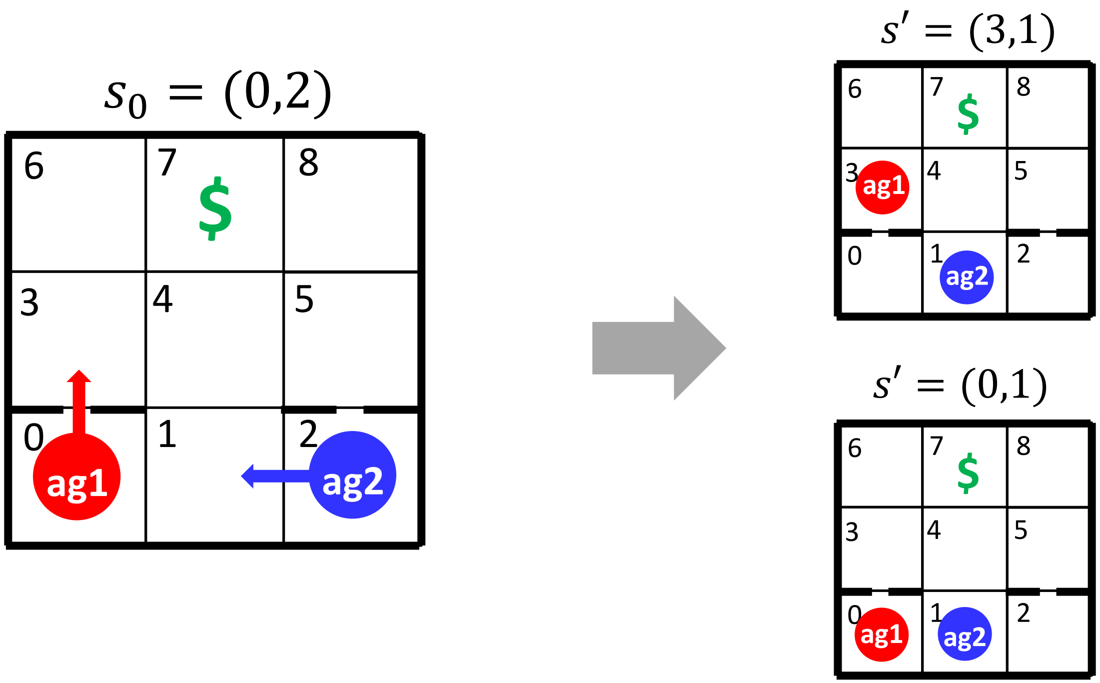
Nash Q values for the initial state $s_0 = (0,2)$

$$Q_1^*(s_0, a_1, a_2) = \mathbb{E}[r_1(s_0, a_1, a_2) + \gamma V_1(s', \pi_1^*, \pi_2^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= r_1(s_0, a_1, a_2) + \gamma \sum_{s'} p(s' | s_0, a_1, a_2) V_1(s', \pi_1^*, \pi_2^*)$$

$$Q_1^*(s_0 = (0,2), Up, Left\ ) = 0 + 0.99 \times \left\{ \frac{1}{2} V_1\big((3,1), \pi_1^*, \pi_2^*\big) + \frac{1}{2} V_1\big((0,1), \pi_1^*, \pi_2^*\big) \right\}$$

$$= 0 + 0.99 \times (0.5 \times 99 + 0.5 \times 0) = 49$$



Grid Game 2

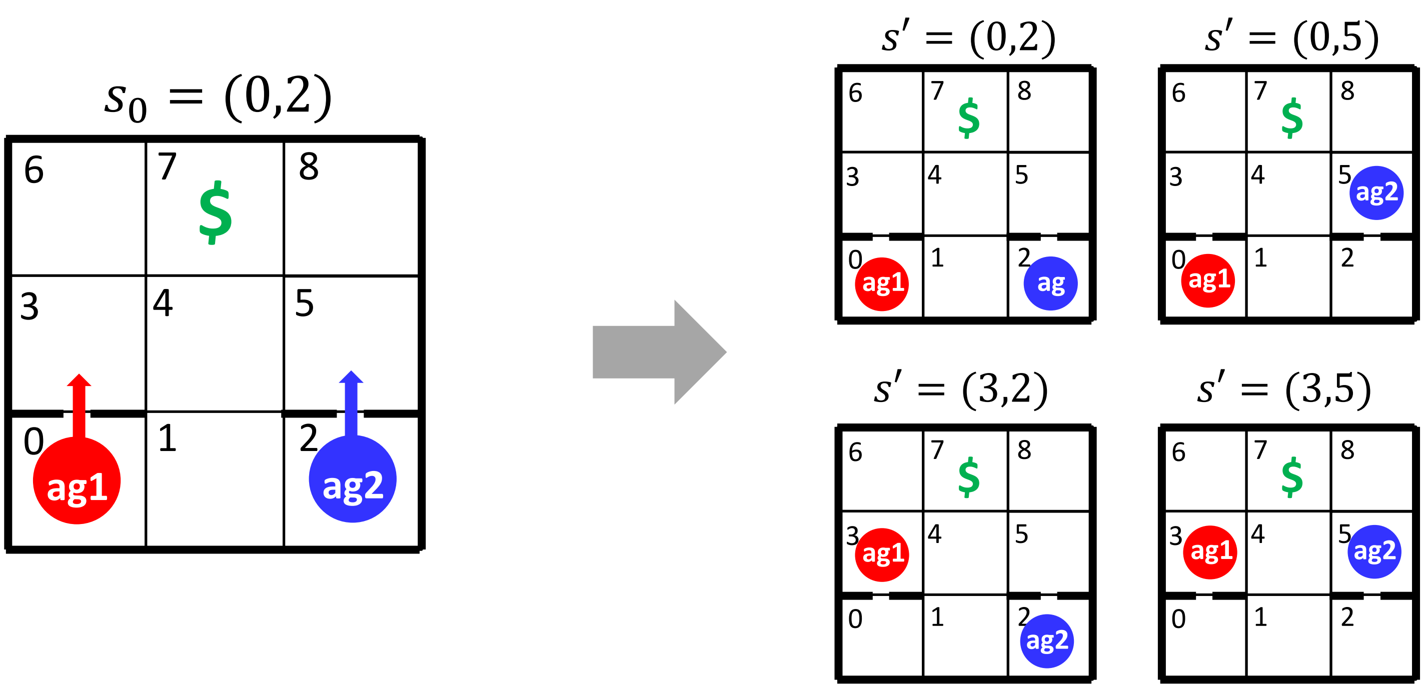
Nash Q values for the initial state $s_0 = (0,2)$

$$Q_1^*(s_0, a_1, a_2) = \mathbb{E}[r_1(s_0, a_1, a_2) + \gamma V_1(s', \pi_1^*, \pi_2^*) | s_t = s, a_t = (a_1, \dots, a_n)]$$

$$= r_1(s_0, a_1, a_2) + \gamma \sum_{s'} p(s' | s_0, a_1, a_2) V_1(s', \pi_1^*, \pi_2^*)$$

$$Q_1^*(s_0 = (0,2), Up, Up) = 0 + 0.99 \times \left\{ \frac{1}{4} V_1^*((0,2)) + \frac{1}{4} V_1^*((0,5)) + \frac{1}{4} V_1^*((3,2)) + \frac{1}{4} V_1^*((3,5)) \right\}$$

$$= 0 + 0.99 \times \left\{ \frac{1}{4} V_1^*(s_0) + \frac{1}{4} \times 0 + \frac{1}{4} \times 99 + \frac{1}{4} \times 99 \right\} = 0.99 \times \frac{1}{4} V_1^*(s_0) + 49$$



Grid Game 2

Nash Q values for the initial state $s_0 = (0,2)$

	$a_2 = Left$	$a_2 = Up$
$a_1 = Right$	$Q_1^*(s_0, R, L), Q_2^*(s_0, R, L)$	$Q_1^*(s_0, R, U), Q_2^*(s_0, R, U)$
$a_2 = Up$	$Q_1^*(s_0, U, L), Q_2^*(s_0, U, L)$	$Q_1^*(s_0, U, U), Q_2^*(s_0, U, U)$

	$a_2 = Left$	$a_2 = Up$
$a_1 = Right$	$-1 + 0.99V_1^*(s_0), -1 + 0.99V_2^*(s_0)$	$98, 49$
$a_2 = Up$	$49, 98$	$49 + \frac{0.99}{4}V_1^*(s_0), 49 + \frac{0.99}{4}V_2^*(s_0)$

Grid Game 2

Nash Q values for the initial state $s_0 = (0,2)$

	$a_2 = Left$	$a_2 = Up$
$a_1 = Right$	$-1 + 0.99V_1^*(s_0), -1 + 0.99V_2^*(s_0)$	$98, 49$
$a_2 = Up$	$\boxed{49, 98}$	$49 + \frac{0.99}{4}V_1^*(s_0), 49 + \frac{0.99}{4}V_2^*(s_0)$

$$V_1^*(s_0) = \text{Nash}\{Q_1^*(s_0, a_1, a_2), Q_2^*(s_0, a_1, a_2)\}$$

Case 1: $V_1^*(s_0) = 49$

	$Left$	Up
$Right$	$47, 96$	$98, 49$
Up	$49, 98$	$61, 73$

Grid Game 2

Nash Q values for the initial state $s_0 = (0,2)$

	$a_2 = Left$	$a_2 = Up$
$a_1 = Right$	$-1 + 0.99V_1^*(s_0), -1 + 0.99V_2^*(s_0)$	$\boxed{98, 49}$
$a_2 = Up$	$49, 98$	$49 + \frac{0.99}{4}V_1^*(s_0), 49 + \frac{0.99}{4}V_2^*(s_0)$

$$V_1^*(s_0) = \text{Nash}\{Q_1^*(s_0, a_1, a_2), Q_2^*(s_0, a_1, a_2)\}$$

Case 2: $V_1^*(s_0) = 98$

	$Left$	Up
$Right$	$96, 47$	$98, 49$
Up	$49, 98$	$73, 61$

Nash Q values for the initial state $s_0 = (0,2)$

	$a_2 = Left$	$a_2 = Up$
$a_1 = Right$	$-1 + 0.99V_1^*(s_0), -1 + 0.99V_2^*(s_0)$	$98, 49$
$a_2 = Up$	$49, 98$	$49 + \frac{0.99}{4}V_1^*(s_0), 49 + \frac{0.99}{4}V_2^*(s_0)$

$$V_1^*(s_0) = \text{Nash}\{Q_1^*(s_0, a_1, a_2), Q_2^*(s_0, a_1, a_2)\}$$

Case 3: $\{\pi_1(s_0), \pi_2(s_0)\} = (\{p(R) = 0.97, p(U) = 0.03\}, \{p(L) = 0.97, p(U) = 0.03\})$

	$Left$	Up
$Right$	$47.48, 47.48$	$98, 49$
Up	$49, 98$	$61.2, 61.2$

Optimal Q-function v.s. Nash Q-function

Definition (Optimal Q-function)

Optimal Q function is defined as

$$Q^*(s,a) = r(s,a,s') + \gamma \sum_{s' \in S} p(s'|s,a\,) V^*(s')$$

- $V^*(s'\,) = \max_a Q^*(s',a)$
- With **optimum** policy $\pi^*(s) = \operatorname{argmax}_a Q^*(s,a)$

Definition (Nash Q-function)

Nash-Q function is defined as

$$Q_i^*(s,\vec{a}) = r_i(s,\vec{a},s') + \gamma \sum_{s' \in S} p(s'|s,\vec{a}) \underbrace{V_i^*(s')}_{Nash\,Q_i(s')}$$

- $V_i^*(s') = \text{Nash}\,Q_i^*(s')$ is Nash equilibrium value that can be computed by solving the following state game

$$(Q_1^*(s',\vec{a})\,, \ldots, Q_n^*(s',\vec{a})\,)$$

Definition (Nash equilibrium policy in Stochastic game)

Compute the Nash equilibrium policies $\pi^* = (\pi_1^*, \pi_2^*)$ such that for all $s \in S$ and $i = 1, \dots 2$,

$$V_i(s, \pi_i^*, \pi_{-i}^*) \geq V_i(s, \pi_i^*, \pi_{-i}^*) \quad \text{for all } \pi_i \in \Pi_i$$