

Equilibrium Selection in Multi-agent Reinforcement Learning

Bianchi Ludovica, Calabretta Silvia

University of Trieste

September 28, 2025

Contents

- 1 Basic definitions
- 2 Game Theory Setting
- 3 Unified Framework
- 4 Results and applications
- 5 Conclusion

Problem setting

- **Problem:** how can we extend successful equilibrium selection methods from normal-form games to stochastic games?
- **Our Approach:** a unified framework that uses the learning rules from normal-form games as foundational components for solving stochastic games.
- **Main goal:** establish theoretical guarantees and demonstrate convergence across different settings.

Outline

- 1 Basic definitions
- 2 Game Theory Setting
- 3 Unified Framework
- 4 Results and applications
- 5 Conclusion

Definitions

Definition

A **Stochastic Game (SG)** is described through the model

$\mathcal{M} = \{\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, P, r, \rho, H\}$, where:

- n : number of agents;
- H : horizon of the Markov game;
- $s_h \in \mathcal{S}$: state at horizon h ;
- $a_{i,h} \in \mathcal{A}_i$: local action of agent i , $a_h = \{a_{1,h}, \dots, a_{n,h}\}$, and thus $\mathcal{A} = \mathcal{A}_1 \times \dots \times \mathcal{A}_n$;
- P : state transition probabilities, $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$;
- $r = (r_{i,h})$: rewards s.t. $r_{i,h} : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$;
- ρ : initial local state distribution.

A **stochastic policy** $\pi = \{\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})\}_{h=1}^H$ specifies a strategy in which agents choose their actions based on the current state.

Definitions

Given a policy π , we define the V function and the Q function as:

$$V_{i,h}^{\pi}(s) := \mathbb{E}^{\pi} \left[\sum_{h'=h}^H r_{i,h'}(s_{h'}, a_{h'}) \mid s_h = s \right]$$

$$Q_{i,h}^{\pi}(s, a) := \mathbb{E}^{\pi} \left[\sum_{h'=h}^H r_{i,h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a \right].$$

Definition

A deterministic policy π^* is called a **Markov perfect (pure Nash) equilibrium (MPE)** if and only if for all $i \in [n], h \in [H], s \in \mathcal{S}$, the following inequality holds:

$$Q_{i,h}^{\pi^*}(s, a^*) > Q_{i,h}^{\pi}(s, a_i, a_{-i}^*), \quad \forall a_i \neq a_i^*, \forall \pi,$$

where a^* is the action such that $\pi^*(a^* \mid s) = 1$.

Pareto Optimal Policies

Definition

A policy π^* is a **Pareto optimal policy** if it maximizes the social utility, i.e.

$$\sum_{i=1}^n V_{i,h}^{\pi^*}(s) \geq \sum_{i=1}^n V_{i,h}^{\pi}(s) \quad \forall \pi, \forall h \in [H], \forall s \in \mathcal{S}.$$

Definition

A policy π^* is a **Pareto optimal MPE** if maximizes the social utility among all MPEs, i.e.

$$\sum_{i=1}^n V_{i,h}^{\pi^*}(s) \geq \sum_{i=1}^n V_{i,h}^{\pi}(s) \quad \forall \pi \in \Pi_{\text{MPE}}, \forall h \in [H], \forall s \in \mathcal{S}.$$

Outline

- 1 Basic definitions
- 2 Game Theory Setting
- 3 Unified Framework
- 4 Results and applications
- 5 Conclusion

Normal-form Games

A *normal-form game* is captured by the reward functions $\{r_i : \mathcal{A} \rightarrow \mathbb{R}\}_{i=1}^n$ of each agent i .

For equilibrium selection, the group of agents respond to the reward outcome from the previous action following certain *iterative learning rules*, according to a transition kernel K^ϵ :

$$(a^{(t+1)}, \xi^{(t+1)}) \sim K^\epsilon(\cdot, \cdot \mid a^{(t)}, \xi^{(t)}),$$

where $\xi \in \mathcal{E}$ are auxiliary variables, and $\epsilon \in (0, 1)$ represents the rate of mistakes.

Assumptions

Assumption

For any given $\epsilon > 0$ and any set of reward $\{r_i : \mathcal{A} \rightarrow \mathbb{R}\}_{i=1}^n$, the Markov chain induced by the transition kernel $K^\epsilon(a^{(t+1)}, \xi^{(t+1)} \mid a^{(t)}, \xi^{(t)}, \{r_i\}_{i=1}^n)$ is **ergodic**.

Definition

Given the assumption above, we denote as $\pi^\epsilon(a, \xi)$ the unique **stationary distribution** the Markov chain K^ϵ induces on $\mathcal{A} \times \mathcal{E}$.

$\pi^\epsilon(a)$ denotes the marginal stationary distribution on \mathcal{A} , i.e.

$$\pi^\epsilon(a) = \sum_{\xi \in \mathcal{E}} \pi^\epsilon(a, \xi).$$

Definition

The **Stochastically stable equilibrium (SSE)** of a learning rule K^ϵ are actions a^* such that $\lim_{\epsilon \rightarrow 0} \pi^\epsilon(a^*) > 0$.

Assumption

For any pair of (a, ξ) and (a', ξ') , there exists a constant $R((a, \xi) \rightarrow (a', \xi'))$ and $C_1, C_2 > 0$, such that

$$C_1 \epsilon^{R((a, \xi) \rightarrow (a', \xi'))} < K^\epsilon(a', \xi' | a, \xi; \{r_i\}_{i=1}^n) < C_2 \epsilon^{R((a, \xi) \rightarrow (a', \xi'))}$$

Definition

The constant $R((a, \xi) \rightarrow (a', \xi'))$ is called the **resistance** of the transition $(a, \xi) \rightarrow (a', \xi')$ under the transition kernel $K^\epsilon(\cdot | \cdot; \{r_i\}_{i=1}^n)$.

Notice, from the assumption:

- If $\lim_{\epsilon \rightarrow 0} K^\epsilon(a', \xi' | a, \xi; \{r_i\}_{i=1}^n) > 0$, then $R((a, \xi) \rightarrow (a', \xi')) = 0$.
- If $K^\epsilon(a', \xi' | a, \xi; \{r_i\}_{i=1}^n) = 0 \quad \forall \epsilon$, then $R((a, \xi) \rightarrow (a', \xi')) = +\infty$.

We can model the transitions as a *directed graph* where each edge $(a, \xi) \rightarrow (a', \xi')$ is weighted by its resistance $R((a, \xi) \rightarrow (a', \xi'))$.

The **total resistance** of a set of edges T is the sum of their individual resistances:

$$R(T) := \sum_{(a, \xi) \rightarrow (a', \xi') \in T} R((a, \xi) \rightarrow (a', \xi'))$$

We denote as $\mathcal{T}(a, \xi)$ the set of *spanning trees* rooted in (a, ξ) , each consisting of $|\mathcal{A}||\mathcal{E}| - 1$ directed edges, such that from every other vertex (a', ξ') there is a unique directed path to the root (a, ξ) .

Definition

The **stochastic potential** $\gamma(a, \xi)$ of an action a and a hidden variable ξ is defined as:

$$\gamma(a, \xi) := \min_{T \in \mathcal{T}(a, \xi)} R(T)$$

Notation: $\gamma(a, \xi; \{r_i\}_{i=1}^n)$ specifies that γ refers to the game $\{r_i\}_{i=1}^n$.

Game Theory Result

Theorem

An action $a^* \in \mathcal{A}$ is the **Stochastically Stable Equilibrium (SSE)** of a normal-form game $\{r_i\}_{i=1}^n$ and a learning rule K^ϵ if and only if there exists a hidden variable $\xi^* \in \mathcal{E}$ such that (a^*, ξ^*) minimizes the stochastic potential, i.e., $\gamma(a^*, \xi^*) = \min_{a, \xi} \gamma(a, \xi)$. Furthermore, for any pair of states (a, ξ) and (a^*, ξ^*) , there exists a constant $C > 0$ such that

$$\frac{\pi^\epsilon(a, \xi)}{\pi^\epsilon(a^*, \xi^*)} < C e^{\gamma(a, \xi) - \gamma(a^*, \xi^*)}.$$

Learning Rules

Log-linear Learning: $\mathcal{E} = \emptyset$,

$$K^\epsilon \left(a^{(t+1)} = (a_{-i}^{(t)}, a_i^{(t+1)}) \mid a^{(t)}; \{r_i\}_{i=1}^n \right) = \frac{1}{n} \frac{\epsilon^{-r_i(a^{(t+1)})}}{\sum_{a'_i} \epsilon^{-r_i(a'_i, a_{-i}^{(t)})}}$$

where

- $\frac{1}{n}$: probability of choosing a random agent to update his action;
- $\epsilon^{-r_i(a^{(t+1)})}$: probability that increases for actions with higher rewards.

Learning Rules

Marden Mood Learning: $\xi = (\xi_1, \dots, \xi_n)$, where $\xi_i \in \{C, D\}$,

Action dynamics:

$$\begin{cases} \text{if } \xi_i^{(t)} = D \rightarrow a_i^{(t+1)} \sim \text{Unif}(\mathcal{A}_i) \\ \text{if } \xi_i^{(t)} = C \rightarrow a_i^{(t+1)} \begin{cases} = a_i^{(t)} & \text{with prob } 1 - \epsilon^C \\ \sim \text{Unif}(\mathcal{A}_i \setminus \{a_i^{(t)}\}) & \text{with prob } \epsilon^C \end{cases} \end{cases}$$

Mood dynamics:

$$\begin{cases} \text{if } \xi_i^{(t)} = C \text{ and } a^{(t+1)} = a^{(t)} \rightarrow \xi_i^{(t+1)} = C \\ \text{else} \rightarrow \xi_i^{(t+1)} = \begin{cases} C & \text{with prob } \epsilon^{1-r_i(a^{(t+1)})} \\ D & \text{otherwise} \end{cases} \end{cases}$$

Outline

- 1 Basic definitions
- 2 Game Theory Setting
- 3 Unified Framework**
- 4 Results and applications
- 5 Conclusion

The Algorithm

Core idea: to apply the learning rule K^ϵ independently at each stage (h) and state (s).

The algorithm has two main steps:

- 1 **Actor:** applies the learning rule K^ϵ , substituting the normal-form game's rewards with the Q values.
- 2 **Critic:** updates the values $Q_{i,h}^t(s, a)$ using a Bellman-like iteration.

Notation

- $\pi_h^{(t)}(\cdot, \cdot | s)$: the joint probability distribution of the action $a_h^{(t)}(s)$ and the hidden variable $\xi_h^{(t)}(s)$.
- $\pi_h^{(t)}(a | s)$: the marginal distribution of the action $a_h^{(t)}(s)$, obtained by summing over all hidden variables, i.e., $\pi_h^{(t)}(a | s) = \sum_{\xi \in E} \pi_h^{(t)}(a, \xi | s)$.
- $\pi_h^\epsilon(\cdot, \cdot | s)$ and $\pi_h^\epsilon(a | s)$: the stationary distributions as $t \rightarrow +\infty$.

Algorithm 1

Initialization: $\forall s \in \mathcal{S}$

- $Q_{i,h}^{(0)}(s, a) = r_{i,h}(s, a), \quad \forall h \in [H], i \in [N], a \in \mathcal{A}.$
- $V_{i,H+1}^{(t)}(s) = 0, \quad \forall i \in [N], t \geq 0.$
- $a_h^{(0)}(s) \in \mathcal{A}, \xi_h^{(0)}(s) \in \mathcal{E}$ randomly.

for $t = 0, 1, \dots$ **do**

for $h = H, H-1, \dots, 1$ **do**

Actor: $(a_h^{(t+1)}(s), \xi_h^{(t+1)}(s)) \sim K^\epsilon(\cdot | a_h^{(t)}(s), \xi_h^{(t)}(s); \{Q_{i,h}^{(t)}(s, \cdot)\}_{i=1}^n)$

Critic: **for** $i = 1, \dots, n, h = 1, \dots, H$

$$V_{i,h}^{(t+1)}(s) = \frac{t}{t+1} V_{i,h}^{(t)}(s) + \frac{1}{t+1} Q_{i,h}^{(t)}(s, a_h^{(t)}(s))$$

$$Q_{i,h}^{(t+1)}(s, a) = r_{i,h}(s, a) + \sum_{s'} P_h(s' | s, a) V_{i,h+1}^{(t+1)}(s')$$

end for

end for

end for

Main Result

Theorem

Given the previous assumptions, the learning rule K^ϵ has a stationary distribution π^ϵ , and when Algorithm 1 is executed under K^ϵ , its long-run behavior converges to it.

Further, $\lim_{\epsilon \rightarrow 0} \pi^\epsilon = \pi^$, and $\pi_h^*(\cdot \mid s)$ has support on actions for which there exists $\xi \in \mathcal{E}$ such that the pair (a, ξ) minimizes the stochastic potential, i.e.,*

$$\gamma(a, \xi; \{Q_{i,h}^{\pi_{h+1:H}^*}(s, \cdot)\}_{i=1}^n) = \min_{a', \xi'} \gamma(a', \xi'; \{Q_{i,h}^{\pi_{h+1:H}^*}(s, \cdot)\}_{i=1}^n)$$

Sketch of the Proof

- **Technical challenge:** to apply the learning rules to $\{Q_{i,h}^{(t)}(s, \cdot)\}_{i=1}^n$ instead of $\{r_i\}_{i=1}^n$. The difficulty is that $Q_{i,h}^{(t)}(s, \cdot)$ is a random variable varying with each iteration t , and potentially correlated with outputs in the past iterations.
- **Two-stage game:** we prove the theorem for the stochastic game with only two stages, i.e., $H = 2$.
- **Particular cases:** we choose the log-linear learning rule and identical rewards for every agent i at every stage $h = 1, 2$. Thus, the Q-functions are the same and simply denoted as $Q_h^{(t)}$.

Sketch of the Proof

Last stage: $H=2$. At each state s , it holds $Q_H^{(t)}(s, \cdot) = r_H(s, \cdot)$ thus the dynamic is the same as applying log-linear learning on a normal-form game, with the reward matrix given as $r_H(s, \cdot)$.

- From normal-form game setting, $\pi_H^\epsilon(\cdot|s)$ exists.
- From concentration lemmas of the MC:

$$V_H^{(t+1)}(s) = \frac{\sum_{\tau=1}^{t+1} r_H(s, a_h^{(\tau)}(s))}{t+1} \xrightarrow{\text{a.s.}} V_H^{\pi_H^\epsilon}(s).$$

- This gives for $H-1=1$:

$$\begin{aligned} Q_{h=1}^{(t)}(s, a) &= r_{h=1}(s, a) + \sum_{s'} P_{h=1}(s'|s, a) V_H^{(t)}(s') \\ &\xrightarrow{\text{a.s.}} r_{h=1}(s, a) + \sum_{s'} P_{h=1}(s'|s, a) V_H^{\pi_H^\epsilon}(s') = Q_{h=1}^{\pi_H^\epsilon}(s, a). \end{aligned}$$

Sketch of the Proof

Stage: $h=1$. We first need:

Lemma

If the random variables $Q_{h=1}^{(t)}(s, \cdot)$ almost surely converge to $Q_{h=1}^{\pi_H^\epsilon}(s, \cdot)$, then the two random processes:

$$a_{h=1}^{(t+1)}(s) \sim K^\epsilon(\cdot | a_{h=1}^{(t)}(s); Q_{h=1}^{(t)}(s, \cdot))$$

$$a'_{h=1}^{(t+1)}(s) \sim K^\epsilon(\cdot | a'_{h=1}^{(t)}(s); Q_{h=1}^{\pi_H^\epsilon}(s, \cdot))$$

share the same stationary distributions and the same concentration properties.

From previous observation and from the Lemma we derive: the original random process $K^\epsilon(\cdot | \cdot; Q_{h=1}^{(t)}(s, \cdot))$ is equivalent to the auxiliary process $K^\epsilon(\cdot | \cdot; Q_{h=1}^{\pi_H^\epsilon}(s, \cdot))$ where $Q_{h=1}^{\pi_H^\epsilon}(s, \cdot)$ is no longer iteration varying.

Sketch of the Proof

Applying a classical result in equilibrium selection we obtain:

Lemma

For a given state s and horizon h , define $a^ \in \arg \max_a Q_h^{\pi^\epsilon}(s, a)$. There exists a uniform constant C (with respect to ϵ), such that for any ϵ*

$$\pi_{h=1}^\epsilon(a \mid s) < C e^{-\frac{1}{\epsilon}(Q_h^{\pi^\epsilon}(s, a^*) - Q_h^{\pi^\epsilon}(s, a))}, \quad \forall a \in \mathcal{A}$$

$$\pi_{h=2}^\epsilon(a \mid s) < C e^{-\frac{1}{\epsilon}(r_H(s, a^*) - r_H(s, a))}, \quad \forall a \in \mathcal{A}$$

- For $h = H = 2$: if $\epsilon \rightarrow 0$, $\pi_H^\epsilon \rightarrow \pi_H^*$, where

$$\pi_H^*(\cdot \mid s) \in \arg \max_{\pi_H(\cdot \mid s)} \sum_a \pi_H(a \mid s) r_H(s, a)$$

- For $h = 1$: if $\epsilon \rightarrow 0$, $Q_h^{\pi^\epsilon}(s, a) \rightarrow Q_h^{\pi^*}(s, a)$, thus $\pi_h^\epsilon \rightarrow \pi_h^*$, where

$$\pi_h^*(\cdot \mid s) \in \arg \max_{\pi_h(\cdot \mid s)} \sum_a \pi_h(a \mid s) Q_h^{\pi^*}(s, a)$$

Sketch of the Proof

The above definitions for π_h^* satisfy the Bellman optimality condition, thus π^* is the optimal policy that maximizes the cumulative reward.

This shows the following particular case:

Theorem

In an identical interest 2-stage game with log-linear learning algorithm, π^ϵ exists. Further, as $\epsilon \rightarrow 0$, $\pi^\epsilon \rightarrow \pi^$, being π^* the global optimal policy.*

Outline

- 1 Basic definitions
- 2 Game Theory Setting
- 3 Unified Framework
- 4 Results and applications**
- 5 Conclusion

Potential Games

Definition

A normal-form game $\{r_i\}_{i=1}^n$ is called a **potential game** if there exists a potential function $\phi : \mathcal{A} \rightarrow \mathbb{R}$ such that for any agent i :

$$\phi(a_i, a_{-i}) - \phi(a'_i, a_{-i}) = r_i(a_i, a_{-i}) - r_i(a'_i, a_{-i}) \quad \forall a_i, a'_i \in \mathcal{A}_i, \quad a_{-i} \in \mathcal{A}_{-i}$$

Definition

A stochastic game is a **Markov Potential Game (MPG)** if there exist functions $\{\phi_h : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}\}_{h=1}^H$ such that

$$\Phi_h^\pi(s, a) := \mathbb{E}^\pi \left[\sum_{h'=h}^H \phi_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a \right]$$

satisfies $\forall \pi, \forall i, \forall a$

$$\Phi_h^\pi(s, a_i, a'_{-i}) - \Phi_h^\pi(s, a_i, a_{-i}) = Q_{i,h}^\pi(s, a_i, a'_{-i}) - Q_{i,h}^\pi(s, a_i, a_{-i}).$$

We call Φ_h^π the total potential function and ϕ the stage potential.

Log-linear Convergence

Definition

For a MPG with total potential function Φ_h^π , the **potential-maximizing policy** π^* is the policy such that $\Phi_h^{\pi^*}(s, a) \geq \Phi_h^\pi(s, a) \quad \forall s \in \mathcal{S}, a \in \mathcal{A}$ for all policy π .

Corollary

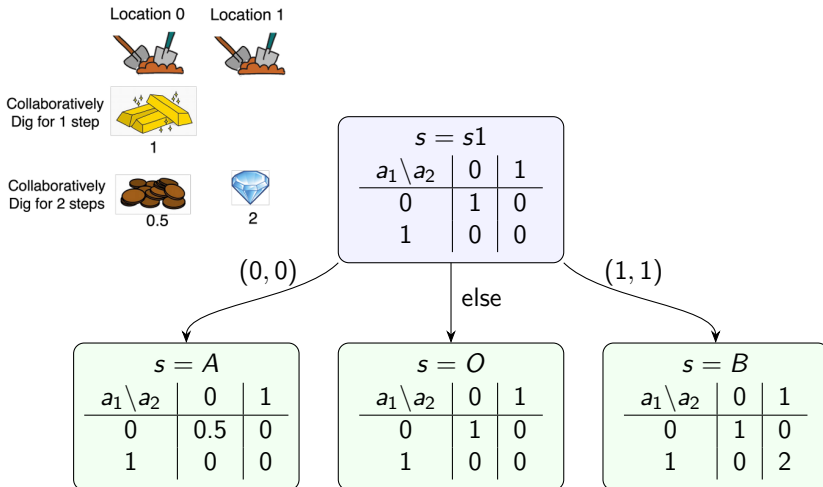
For a PG, under the log-linear learning rule, the SSEs are the potential maximizing actions a^* such that $a^* \in \arg \max_a \phi(a)$. Further, the stationary distribution π^ϵ satisfies:

$$\frac{\pi^\epsilon(a)}{\pi^\epsilon(a^*)} < C \epsilon^{\phi(a^*) - \phi(a)}$$

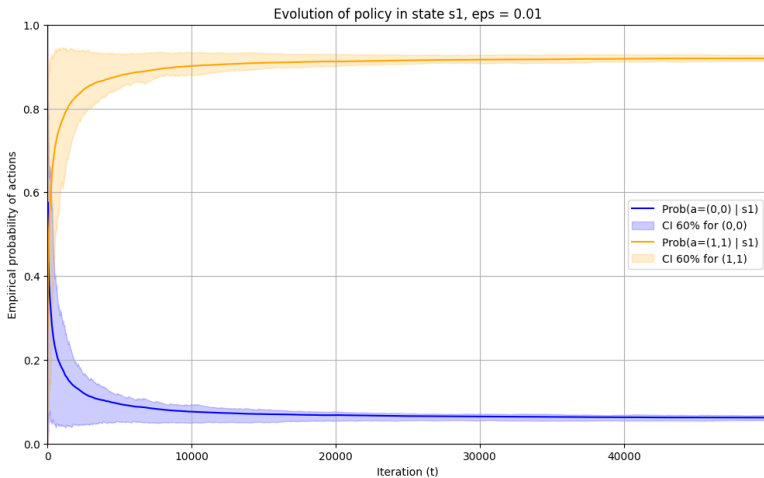
Corollary

For a MPG, the policy π^* to which the Algorithm 1 converges under the log-linear learning rule is a potential-maximizing policy.

Treasure Game



Treasure Game



Marden Mood Convergence

Definition

- A normal-form game $\{r_i\}_{i=1}^n$ is **interdependent** if $\forall a \in \mathcal{A}$ and every proper subset $J \subset [n]$, there exist $i \notin J$ and $a'_J \in \prod_{j \in J} \mathcal{A}_j$ such that

$$r_i(a'_J, a_{-J}) \neq r_i(a_J, a_{-J}).$$

- A stochastic game is **interdependent** if, for all $\pi, h \in [H], s \in \mathcal{S}$, the stage game induced by $\{Q_{i,h}^\pi(s, \cdot)\}_{i=1}^n$ is interdependent.

Corollary

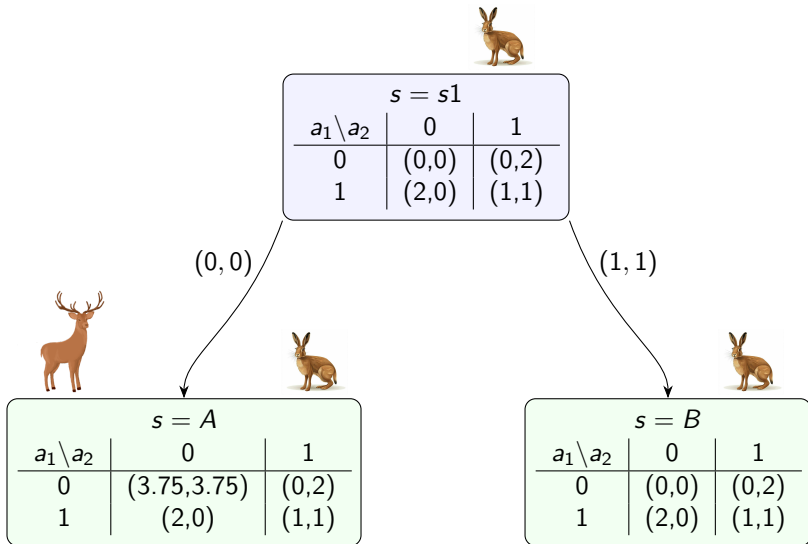
In general-sum normal-form games, under interdependence and the Marden Mood learning rule, the SSEs are Pareto-optimal.



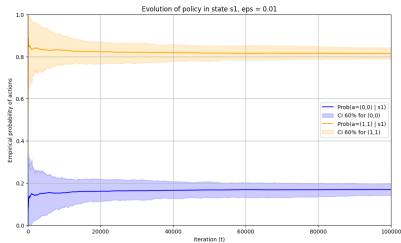
Corollary

In general-sum stochastic games, the policy π^* to which Algorithm 1 converges under the Marden Mood learning rule is Pareto-optimal.

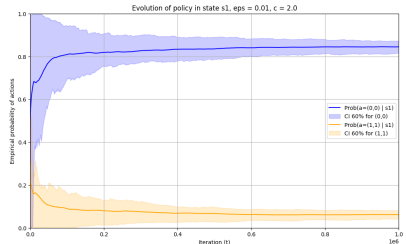
Stag Hunt Game



Stug Hunt Game



Log-linear learning convergence



Marden mood learning convergence

Outline

- 1 Basic definitions
- 2 Game Theory Setting
- 3 Unified Framework
- 4 Results and applications
- 5 Conclusion

Final

Our work demonstrates how classical game theory can be successfully applied within modern reinforcement learning.

- The framework is modular and adapts to different games and behaviors, depending on the chosen learning rule.
- Its robustness relies on theoretical convergence guarantees.
- The approach can be naturally extended to richer scenarios and more complex dynamics.

- [1] Runyu Zhang, Jeff Shamma, and Na Li. *Equilibrium Selection for Multi-agent Reinforcement Learning: A Unified Framework*. 2024. arXiv: 2406.08844 [cs.GT].
- [2] Lawrence E. Blume. “The Statistical Mechanics of Strategic Interaction”. In: *Games and Economic Behavior* 5.3 (1993), pp. 387–424. ISSN: 0899-8256.
- [3] Jason R. Marden, H. Peyton Young, and Lucy Y. Pao. “Achieving pareto optimality through distributed learning”. In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. 2012, pp. 7419–7424. DOI: 10.1109/CDC.2012.6426834.
- [4] H. Peyton Young. “The Evolution of Conventions”. In: *Econometrica* 61.1 (1993), pp. 57–84. ISSN: 00129682, 14680262. (Visited on 09/19/2025).
- [5] Jason R. Marden and Jeff S. Shamma. “Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation”. In: *Games and Economic Behavior* 75.2 (2012), pp. 788–808. ISSN: 0899-8256.