Convergent Q-learning in Discounted Markov Games

Yizhou Zhang

joint work with Prof. Eric Mazumdar

December 9, 2025

Caltech

Multi-Agent Learning Increasingly Deployed in Real Wrold

We want our agents to be robust and performant.

Caltech

Multi-Agent Learning Increasingly Deployed in Real Wrold

We want our agents to be robust and performant.

Stable against environmental or strategic deviations

Caltech

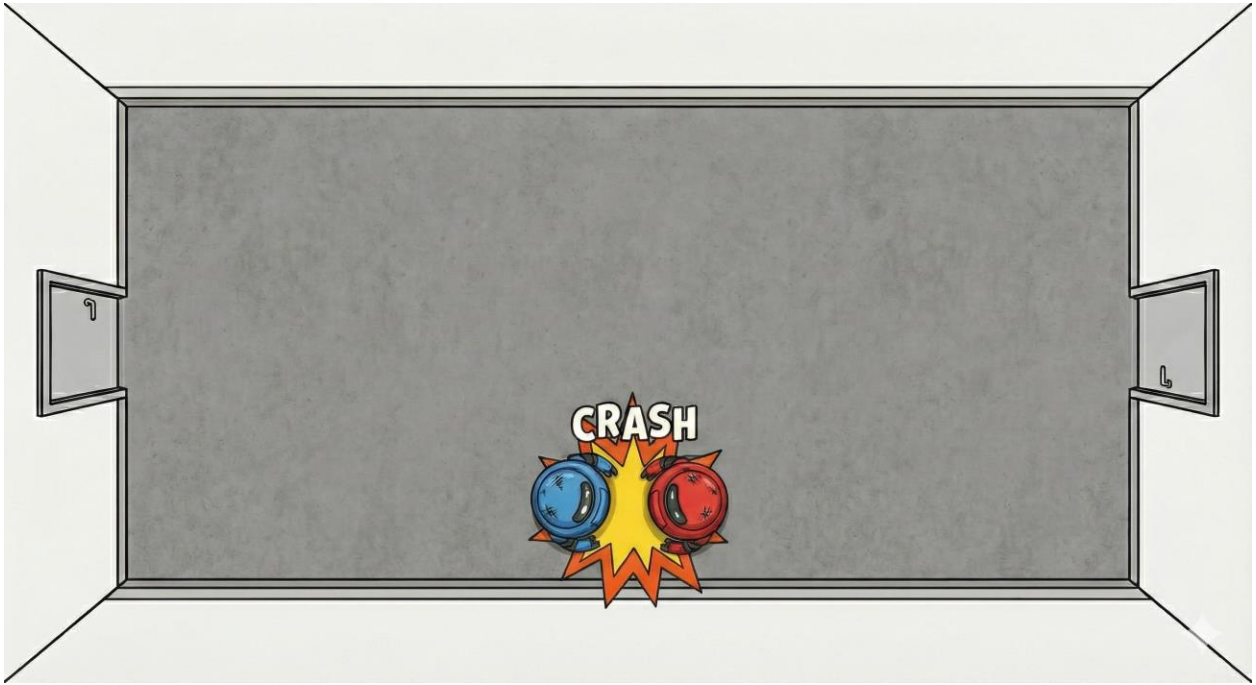
Multi-Agent Learning Increasingly Deployed in Real Wrold

We want our agents to be robust and performant.

Achieve “optimality” against others, a.k.a. “Equilibrium”

Caltech

Issue 1:
Instability in Learning



Caltech

Issue 1:
Instability in Learning

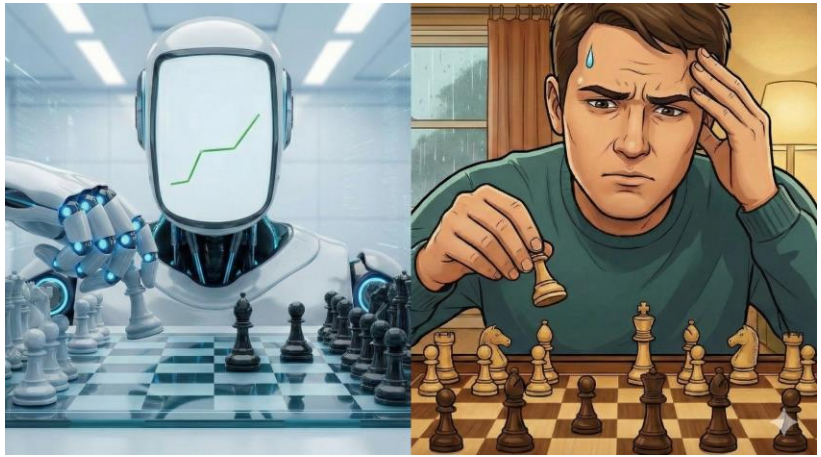
Instability: Fundamental limitation.
Nash equilibria are intractable to compute!

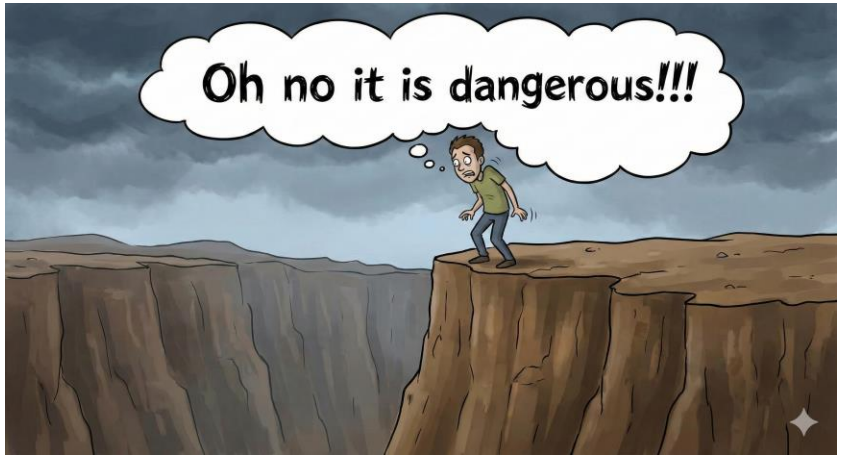
Caltech

Issue 2:
Ignoring Human Traits

Caltech

Issue 2:
Ignoring Human Traits

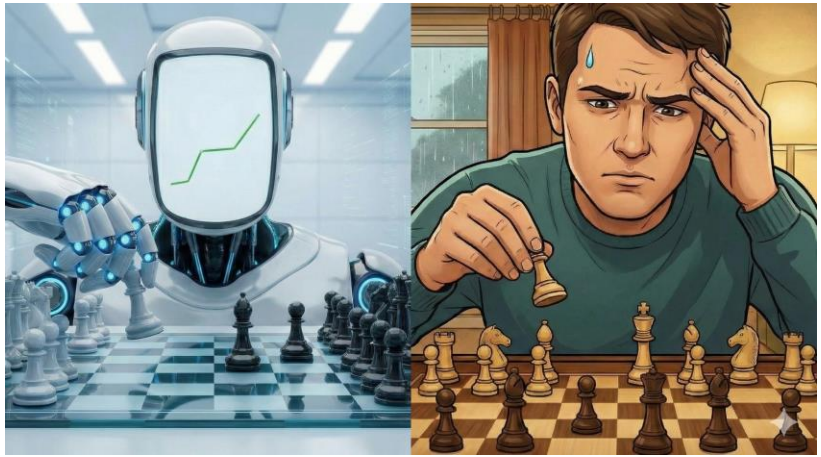




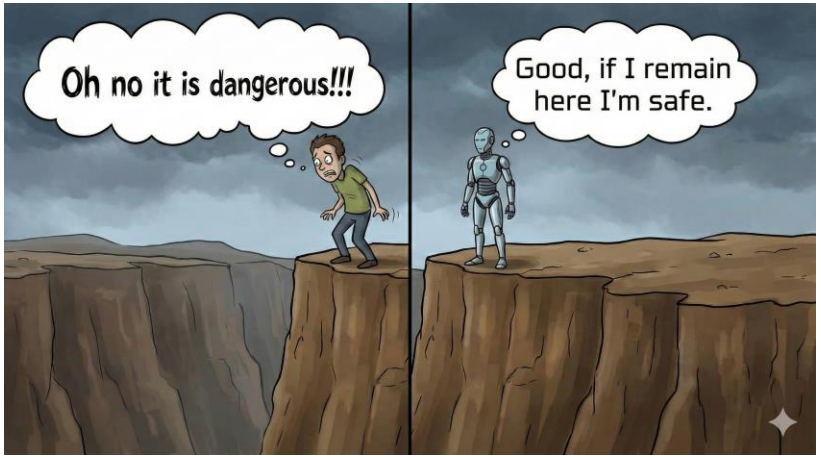
Humans are boundedly rational.

Caltech

Issue 2:
Ignoring Human Traits



Humans are boundedly rational.



Humans are risk-averse.

Caltech

Issue 2:
Ignoring Human Traits

Two aspects Nash equilibrium fails to capture:
Risk-aversion and bounded rationality.

Caltech

Question:

Is there an equilibrium notion that:
Is tractable to compute, and incorporates
risk-aversion and bounded rationality?

Caltech

Objective Construction: Matrix Games

Player i's objective with risk-aversion and bounded-rationality:

Minimize Cost + risk-aversion + bounded rationality

Caltech

Objective Construction: Matrix Games

Player i's objective with risk-aversion and bounded-rationality:

$$\min_{\pi_i} \max_{p_i} -\pi_i^T R_i p_i - D_i(p_i, \pi_{-i})/\tau_i + \epsilon_i \nu_i(\pi_i)$$

Example: reverse KL negative entropy

Caltech

Solution Concept: RQE
Risk-averse Quantal-response Equilibrium

Player i's objective with risk-aversion and bounded-rationality:

$$\min_{\pi_i} \max_{p_i} -\pi_i^T R_i p_i - D_i(p_i, \pi_{-i})/\tau_i + \epsilon_i \nu_i(\pi_i)$$

Example: reverse KL negative entropy

Definition (RQE for matrix games): An RQE is a joint strategy $\pi^* = (\pi_1^*, \pi_2^*, \ldots, \pi_n^*)$ such that each π_i^* maximizes the objective given fixed π_{-i}^*.

Caltech

RQE in Matrix Games: Uniqueness & Lipschitz Continuity

Theorem (Informal): If $\epsilon_1\epsilon_2\tau_1\tau_2 > 1$, the RQE of the game satisfies the following properties for all payoff structures:

1. Uniqueness;
2. Lipschitz continuity w.r.t. payoff matrix R;
3. Tractable to compute.

Proof relies on monotonicity of the expanded game between original players and adversaries

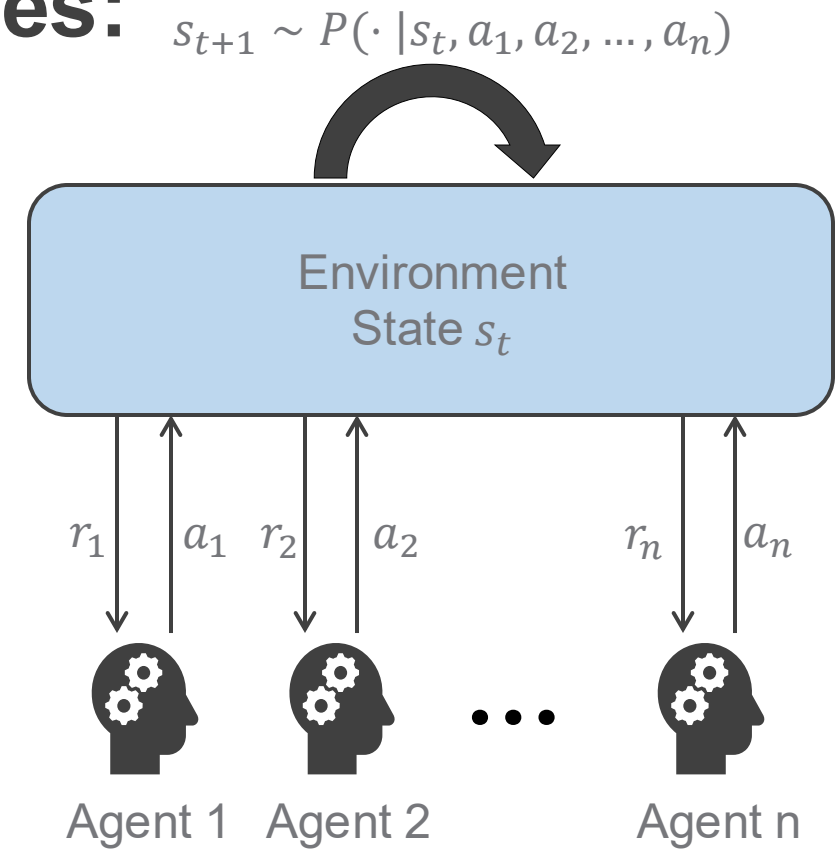
Caltech

RQE in Markov Games: Definition

Markov Games (discounted):

- n agents
- shared state s_t
- individual action $a_{i,t}$
- rewards $r_i(s, a_1, a_2, \dots, a_n)$
- agent i objective (risk-neutral):

$$\max \sum_t \gamma^t r_i(s_t, a_t)$$



Caltech

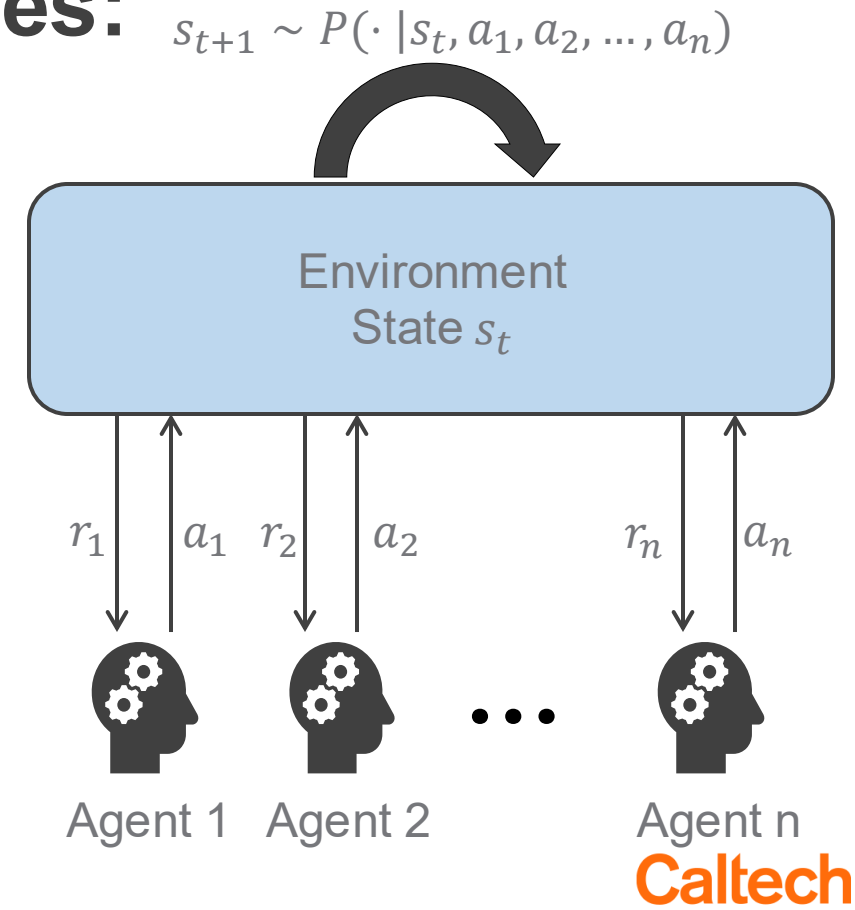
RQE in Markov Games: Definition

Markov Games (discounted):

- n agents
- shared state s_t
- individual action $a_{i,t}$
- rewards $r_i(s, a_1, a_2, \dots, a_n)$
- agent i objective (RQE):

$$\max \sum_t \gamma^t r_i(s_t, a_t)$$

+risk-aversion+bounded rationality



Caltech

RQE in Markov Games: Bellman Operator

Single-agent

Q function:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} r(s_t, a_t) | s_0 = s, a_0 = a\right]$$

Caltech

RQE in Markov Games: Bellman Operator

Single-agent

Q function:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\right]$$

Multi-agent

Q function:

$$Q_i^{\pi}(s,\mathbf{a}) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, \mathbf{a}_t) | s_0 = s, \mathbf{a}_0 = \mathbf{a}\right]$$

Caltech

RQE in Markov Games: Bellman Operator

Single-agent

Q function:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r(s_t,a_t)|s_0=s,a_0=a\right]$$

Bellman operator:

$$(\mathcal{T}Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q(s',a')\right]$$

Contraction → Q-learning converges

Multi-agent

Q function:

$$Q_i^{\pi}(s,\mathbf{a}) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r_i(s_t,\mathbf{a}_t)|s_0=s,\mathbf{a}_0=\mathbf{a}\right]$$

Caltech

RQE in Markov Games: Bellman Operator

Single-agent

Q function:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r(s_t,a_t)|s_0=s,a_0=a\right]$$

Bellman operator:

$$(\mathcal{T}Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q(s',a')\right]$$

Multi-agent

Q function:

$$Q_i^{\pi}(s,\mathbf{a}) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r_i(s_t,\mathbf{a}_t)|s_0=s,\mathbf{a}_0=\mathbf{a}\right]$$

Bellman operator (with Nash):

$$(\mathcal{T}Q)_i(s,\mathbf{a}) = r(s,\mathbf{a}) + \gamma \mathbb{E}_{s'}\left[\texttt{Nash}_i(\mathbf{Q}(s',\cdot))\right]$$

Contraction → Q-learning converges

Caltech

RQE in Markov Games: Bellman Operator

Single-agent

Q function:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r(s_t,a_t)|s_0=s,a_0=a\right]$$

Bellman operator:

$$(\mathcal{T}Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q(s',a')\right]$$

Contraction → Q-learning converges

Multi-agent

Q function:

$$Q_i^{\pi}(s,\mathbf{a}) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r_i(s_t,\mathbf{a}_t)|s_0=s,\mathbf{a}_0=\mathbf{a}\right]$$

Bellman operator (with Nash):

$$(\mathcal{T}Q)_i(s,\mathbf{a}) = r(s,\mathbf{a}) + \gamma \mathbb{E}_{s'}\left[\texttt{Nash}_i(\mathbf{Q}(s',\cdot))\right]$$

Not a contraction!

Additionally, Nash equilibria intractable to compute…

Caltech

RQE in Markov Games: Bellman Operator

Single-agent

Q function:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r(s_t,a_t) | s_0 = s, a_0 = a\right]$$

Bellman operator:

$$(\mathcal{T}Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q(s',a')\right]$$

Contraction → Q-learning converges

Multi-agent

Q function:

$$Q_i^{\pi}(s,\mathbf{a}) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r_i(s_t,\mathbf{a}_t) | s_0 = s, \mathbf{a}_0 = \mathbf{a}\right]$$

Bellman operator (with RQE):

$$(\mathcal{T}Q)_i(s,\mathbf{a}) = r(s,\mathbf{a}) + \gamma \mathbb{E}_{s'}\left[\mathtt{RQE}_i(\mathbf{Q}(s',\cdot))\right]$$

Contraction?

Replace Nash value with RQE value → tractable!

Caltech

RQE in Markov Games: Contraction of Bellman Operator

Theorem (Informal): If $\epsilon_1\epsilon_2\tau_1\tau_2 > c(\gamma)$, the Bellman operator defined with RQE satisfy the contraction property for some $\gamma_0 < 1$:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma_0\|Q - Q'\|_\infty$$

Corollary: Q-learning provably converges to RQE!

Caltech

Takeaways

- RQE incorporates risk-aversion and bounded rationality.

- RQE is tractable to compute in matrix/Markov games.

More technical details will be covered in my talk in Session FrA04, 10:45am-11:00am

Caltech