Towards incorporating safety in reinforcement learning

Maryam Kamgarpour
École Polytechnique Fédérale de Lausanne, Switzerland

Spring School on Control & Reinforcement Learning
CWI Amesterdam, The Netherlands

Mar 20,2025

Example application: autonomous cars

Multiagent autonomous car simulation environment [Zhou et al. 2021]

Autonomous car experiments in the lab

Example application: electricity markets

Multiagent systems

- N agents, with agent $i \in \{1, \ldots, N\}$
  - action a^i, joint action $\boldsymbol{a} = (a^i, a^{-i})$
  - objective $J^i(a^i, a^{-i})$

- Objectives: $\{J^i(.)\}_{i=1}^N$ may not be known

Learning in multiagent systems

Agent i does not know $J^i(.)$ but can query it

How do agents optimize their decisions?

Motivation

Learning Nash equilibria in static games
- Convex games
- Learning with bandit feedback

No-regret learning in dynamic games
- Challenges in Markov games
- No-regret algorithmic approach
- Model-based multiagent RL

Conclusions

Outline

Motivation

Learning Nash equilibria in static games
- Convex games
- Learning with bandit feedback

No-regret learning in dynamic games

Conclusions

Nash equilibrium

- Agent i's action set: A^i

- $\boldsymbol{a}^*$ is equilibrium: $\forall i, \quad J^i(a^{*i}, a^{*-i}) \leq J^i(a^i, a^{*-i}), \ \forall a^i \in A^i$
  - agent i has no reason to deviate from a^i

Nash equilibria may not exist

Pair up with your colleague and play Rock, Paper, Scissors for 10 rounds!

- What is the action set A^i of each player?
- Define cost as 1 for losing and -1 for winning (zero-sum game)
- Is there a Nash equilibrium?
  - $\boldsymbol{a}^*$ is equilibrium: $\forall i, \quad J^i({a^*}^i, {a^*}^{-i}) \leq J^i(a^i, {a^*}^{-i}), \ \forall a^i \in A^i$

Nash equilibria may not exist

Bertrand competition: Two producers setting prices for electricity

- $x^i \in \mathbb{R}_+$ for $i = 1, 2$: the price announced by each producer
- production marginal costs: $c \in \mathbb{R}_+$
- one unit demand
- demand chooses producer with lowest cost. If producers set the same price, then equal demand for each producer

Exercise: derive the Nash equilibrium for each of these two cases

- no capacity limit by each producer
- each producer can serve maximum $2/3$ of the demand

Convex games

- $J^i(a^i, a^{-i})$: convex in a^i, continuous in $\boldsymbol{a} = (a^i, a^{-i})$
- $a^i \in A^i \subset \mathbb{R}^d$: convex and compact
- Examples
  - mixed strategy extension of a finite action game
  - traffic networks, electricity market

Mixed strategy extension of rock-paper-scissors

For player i: $i = 1, 2$

- actions a_1^i, a_2^i, a_3^i: probability of playing rock, paper, and scissors, respectively
- What is the action set A^i?
- What is the cost of each player?
- Why is the game convex?

Nash equilibrium in convex games

Consider

- $J^i(a^i, a^{-i})$: convex in a^i, continuous in $\boldsymbol{a} = (a^i, a^{-i})$
- $a^i \in A^i \subset \mathbb{R}^d$: convex and compact

Nash equilibrium: $\boldsymbol{a}^* = (a^{*1}, a^{*2}, \ldots, a^{*N}) \in \boldsymbol{A}$ guaranteed to exist

Idea of proof:

- $\boldsymbol{a}^*$ is a Nash equilibrium $\iff$ $a^{*i} \in \arg\min_{a^i} J^i(a^i, a^*{-i})$
- Define best-response map:
  BR : $(a^1, \ldots, a^N) \mapsto \subset A^1 \times \cdots \times A^N$.
- $\boldsymbol{a}^*$ is a Nash equilibrium $\iff$ $\boldsymbol{a}^*$ is a fixed point of the best-response map
- Kaktuani's fixed point theorem: for convex games best-response map has a fixed point

Example: Cournot competition

Consider two producers but this time, setting the quantity produced rather than prices

- quantity to produce denoted by $x^i \in \mathbb{R}_+$ for $i = 1, 2$
- production marginal cost of $c \in \mathbb{R}_+$
- market price $p : \mathbb{R}^2 \to \mathbb{R}$: $p(x^1, x^2) = a - b(x^1 + x^2)$.
- each producer has a capacity constraint of $k^i \in \mathbb{R}_+$

Show that the above game is convex. How would you find a Nash equilibrium?

Nash equilibrium and variational inequalities

Consider *additionally* J^i differentiable in $\boldsymbol{a} = (a^i, a^{-i})$

- Nash equilibrium: $\boldsymbol{a}^*$ characterized by <u>pseudo-gradient</u>: $\boldsymbol{M} : \mathbb{R}^{Nd} \to \mathbb{R}^{Nd}$

$$\boldsymbol{M}(\boldsymbol{a}) = [\nabla_i J^i(a^i, a^{-i})]_{i=1}^N$$

$\boldsymbol{a}^*$ Nash equilibrium $\iff$ $\boldsymbol{M}(\boldsymbol{a}^*)^T(\boldsymbol{a} - \boldsymbol{a}^*) \geq 0, \forall \boldsymbol{a} \in \boldsymbol{A}$

- Exercise: derive the pseudo-gradient of the rock-paper-scissors game in mixed strategies, and the Cournot game

Learning in convex games

Consider an independent learning approach:

$$a^i_{t+1} = \mathsf{Proj}_{A^i}(a^i_t - \eta_t \nabla_{a^i} \widehat{J^i(a^i_t, a^{-i}_t)})$$

Learning in convex games

Consider an independent learning approach:

$$a^i_{t+1} = \mathsf{Proj}_{A^i}(a^i_t - \eta_t \nabla_{a^i} \widehat{J^i(a^i_t, a^{-i}_t)})$$

Challenges compared to the single agent setting:

1. How can agent i estimate $\nabla_{a^i} J^i(\boldsymbol{a})$ without access to a^{-i}?

2. Is convexity of J^i in a^i sufficient for convergence?

Gradient estimation - bandit feedback setting

Player i does not know J^i but can query it

- Finite difference: $\widehat{\nabla_{a^i} J^i(\boldsymbol{a})} \approx \frac{J^i(a^i, a^{-i}) - J^i(a^i + \delta, a^{-i})}{\delta}$?
  - requires coordination
  - bandit feedback: $J^i(a^i, a^{-i}), J^i(a^i + \delta, a'^{-i})$

Gradient estimation using randomization

- approach: randomize query $\delta \sim \mathcal{N}(0, \sigma^2)$

- $\mathrm{E}\{\frac{\delta}{\sigma} J^i(a^i, a^{-i})\} \approx \nabla_{a^i} J^i(., a^{-i})$
  - bias: $O(\sigma)$, variance $O(\frac{1}{\sigma^2})$ [Nesterov, Spokoiny 2019]

The pseudo-gradient role in convergence

Game pseudo-gradient: $\boldsymbol{M}(\boldsymbol{a}) = [\nabla_i J^i(a^i, a^{-i})]_{i=1}^N$
Consider independent learning approach with known gradients, unconstrained:

$$\begin{bmatrix} a_{t+1}^1 \\ \vdots \\ a_{t+1}^N \end{bmatrix} = \begin{bmatrix} a_t^1 \\ \vdots \\ a_t^N \end{bmatrix} - \eta_t \underbrace{\begin{bmatrix} \nabla_{a^1} J^1(\boldsymbol{a}_t) \\ \vdots \\ \nabla_{a^N} J^N(\boldsymbol{a}_t) \end{bmatrix}}_{\boldsymbol{M}(\boldsymbol{a}_t) \neq \nabla_{\boldsymbol{a}} J(\boldsymbol{a}_t)}$$

- ex: $J^1(\boldsymbol{a}) = a^1 a^2 = -J^2(\boldsymbol{a})$, $\begin{bmatrix} \nabla_{a^1} J^1(\boldsymbol{a}) \\ \nabla_{a^2} J^2(\boldsymbol{a}) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}$
- single agent analysis don't generally work

Game pseudo-gradient monotonicity

$$\boldsymbol{M}(\boldsymbol{a}) = [\nabla_i J^i(a^i, a^{-i})]_{i=1}^N$$

$$\big(M(\boldsymbol{a}) - M(\boldsymbol{a}')\big)^T(\boldsymbol{a} - \boldsymbol{a}') \geq \nu\|\boldsymbol{a} - \boldsymbol{a}'\|^2,\ \forall \boldsymbol{a}, \boldsymbol{a}' \in \boldsymbol{A}$$

- monotone: $\nu = 0$
- strictly monotone: $\nu = 0$ and inequality strict for $\boldsymbol{a} \neq \boldsymbol{a}'$
- strongly monotone: $\nu > 0$

For differentiable $\boldsymbol{M}$, monotonicity conditions

- based on positive (semi)definiteness of $\nabla \boldsymbol{M}(.) \in \mathbb{R}^{Nd \times Nd}$
- symmetric $\nabla \boldsymbol{M}$: <u>potential game</u> with (strong/strict) convex objective

Exercises - monotone games

- Show that a mixed strategy extension of a finite-action zero-sum game has monotone pseudo-gradient.
  Hint: for player 1,2 having m, n actions, player 1's cost: ${a^1}^T C a^2$ and player 2's cost is $-{a^1}^T C a^2$, $C \in \mathbb{R}^{m \times n}$.

- Show that the Cournot game has strongly monotone pseudo-gradient

Convergence condition of independent learning approach

Agent i: $a^i_{t+1} = \mathrm{Proj}_{A^i}\big(a^i_t - \eta_t \widehat{\nabla_{a^i} J^i(\boldsymbol{a})}\big),\ i = 1, \ldots, N,$

$$\hat{\boldsymbol{M}}(\boldsymbol{a}) := \begin{bmatrix} \widehat{\nabla_{a^1} J^1(\boldsymbol{a})} \\ \vdots \\ \widehat{\nabla_{a^N} J^N(\boldsymbol{a})} \end{bmatrix}$$

Theorem

Let $\boldsymbol{M}$ Lipschitz and strictly monotone. For $\sum_t \eta_t = \infty$, $\sum_t \frac{\eta_t^2}{\sigma_t^2} < \infty$, $\boldsymbol{a}_t$ converges to $\boldsymbol{a}^*$. *[Tatarenko, **MK**, IEEE TAC 2018]*

Proof sketch:

$$\mathrm{E}\{\|\boldsymbol{a}_{t+1} - \boldsymbol{a}^*\|^2\} \quad \leq \|\boldsymbol{a}_t - \boldsymbol{a}^*\|^2 + \underbrace{\xi_t}_{O(\eta_t \sigma_t + \frac{\eta_t^2}{\sigma_t^2})} \quad -\eta_t \underbrace{\boldsymbol{M}(\boldsymbol{a}_t)^T(\boldsymbol{a}_t - \boldsymbol{a}^*)}_{>0}$$

[Robbins, Siegmund 1985]

Advances in bandit learning for convex games

Class of games for which payoff-based learning approach converges

- Merely monotone $\boldsymbol{M}(\boldsymbol{a})$

  extra-gradient, optimistic gradient descent-ascent, Tikhonov regularization, ...
- Games with coupling constraints
- Games that have a variationally stable equilibrium

[Tatarenko, **MK**, IEEE TCNS 2024], [Bravo et al., 2018], [Gao, Pavel, 2022], [Zou, Lygeros, 2023], ...

Challenge: many games, e.g. Markov games, do not in general satisfy convergence conditions above

Outline

Motivation

Learning Nash equilibria in static games

No-regret learning in dynamic games
- Challenges in Markov games
- No-regret algorithmic approach
- Model-based multiagent RL

Conclusions

Markov games

- Dynamics: $s_{h+1} \sim P(.|s_h, a_h^1, \ldots, a_h^N)$, $s_0 \sim \rho$
- Policy $\pi^i : S \to \Delta(A^i)$
- $J^i(\pi^i, \pi^{-i}) = \mathbb{E} \sum_{h=0}^{\infty} \alpha^t c^i(s_h, \pi^1(s_h), \ldots, \pi^N(s_h))$
- Nash equilibrium policy $(\pi^{*1}, \ldots, \pi^{*N})$:

$$J^i(\pi^{*i}, \pi^{*-i}) \leq J^i(\pi^i, \pi^{*-i}), \ \forall \pi^i, \ \forall i$$

  - existence [Shapley 1953, Fink 1964, Takahashi 1964]

Bandit learning in Markov games

~~Given $s_{h+1} \sim P(.|s_h, a_h^1, \ldots, a_h^N)$~~

- Parametrize a policy $a_t^i \sim \pi_{\theta^i}(.|s_h)$, $\theta^i \in \mathbb{R}^d$
- Find equilibrium $\boldsymbol{\theta}^* = (\theta^1, \ldots, \theta^N)$ through interactions

- Reinforcement learning approach

Policy gradient class of algorithms

Single agent RL: $J(\theta) = \mathbb{E}\sum_{h=0}^{\infty} \alpha^t c(s_h, \pi_\theta(s_h))$

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta J(\theta_t)$$

- convergence conditions [Agarwal et al., 2021],[Hu et al. 2023], [Bhandari et al. 2024], ...

Multiagent RL: $J^i(\theta^i, \theta^{-i}) = \mathbb{E}\sum_{h=0}^{\infty} \alpha^t c^i(s_h, \pi_{\theta^1}(s_h), \ldots, \pi_{\theta^N}(s_h))$

$$\theta^i_{t+1} = \theta^i_t - \eta_t \nabla_{\theta^i} J^i(\theta^i_t, \theta^{-i}_t)$$

- generally non-convergent

Challenging even in linear quadratic setting

single agent

$$J_\rho(\theta) = \mathbb{E}_{s_0}[\sum_{h=0}^{\infty} s_h^T Q s_h + a_h^T R a_h]$$

$$s_{h+1} = A s_h + B a_h$$

$$a_h = \theta^T s_h, \; s_0 \sim \rho$$

Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator

Maryam Fazel [*1] Rong Ge [*2] Sham M. Kakade [*1] Mehran Mesbahi [*1]

Abstract

Direct policy gradient methods for reinforcement learning and continuous control problems are a popular approach for a variety of reasons: 1) they

2016) and Atari game playing (Mnih et al., 2015). Deep reinforcement learning (DeepRL) is becoming increasingly popular for tackling such challenging sequential decision making problems.

multiagent

$$J_\rho^i(\boldsymbol{\theta}) = \mathbb{E}_{s_0}[\sum_{h=0}^{\infty} s_h^T Q^i s_h + (a^i)_h^T R^i a_h^i]$$

$$s_{h+1} = A s_h + \sum_{i=1}^{N} B a_h^i$$

$$a_h^i = (\theta^i)^T s_h, \; s_0 \sim \rho$$

Policy-Gradient Algorithms Have No Guarantees of Convergence in Linear Quadratic Games

Eric Mazumdar
University of California, Berkeley
Berkeley, CA
mazumdar@berkeley.edu

Lillian J. Ratliff
University of Washington
Seattle, WA
ratliffl@uw.edu

Michael I. Jordan
University of California, Berkeley
Berkeley, CA
jordan@cs.berkeley.edu

S. Shankar Sastry
University of California, Berkeley
Berkeley, CA
sastry@coe.berkeley.edu

ABSTRACT

We show by counterexample that policy-gradient algorithms have no guarantees of even local convergence to Nash equilibria in continuous action and state space multi-agent settings. To do so, we analyze gradient-play in N–player general-sum linear quadratic games, a classic game setting which is recently emerging as a bench-

of multi-agent reinforcement learning have made use of policy optimization algorithms such as multi-agent actor-critic [13, 17, 30], multi-agent proximal policy optimization [2], and even simple multi-agent policy-gradients [15] in problems where the various agents have high-dimensional continuous state and action spaces like StarCraft II [32].

Sources of challenge

Consider scalar dynamics, 2-player infinite horizon LQ games

- Nash equilibria $\subset$ solution set of a system of 3rd degree polynomials in 2 variables
  - at least 1 and at most 3 equilibria [Salizzoni, Ouhamma, MK, 2024]
- hard to derive conditions for monotonicity, uniqueness of equilibria or the existence of a potential function

Number of equilibria as a function of system parameters

Tractability of learning Nash equilibria in Markov games

Guarantees exist for subclasses of Markov games

- Finite-horizon linear quadratic games [Hambly et al 2021]
- Finite state and action potential games [Leonardos et al. 2022], [R. Zhang et al. 2021], [Ding et al. 2022]
- Zero-sum Markov games [Daskalakis et al. 2020], [Wei et al. 2021], [Cen et al. 2021], [K. Zhang et al. 2023], [Ouhamma, **MK**, 2023]...

For more general classes, computation/learning of Nash equilibria is intractable [Daskalakis et al. 2023]

Relaxing the equilibrium notion

A *probability distribution* $\mathcal{P}^*$ on $\boldsymbol{A}$ is coarse correlated equilibrium (CCE) if

$$\forall i \quad \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{P}^*}[J^i(\boldsymbol{\theta})] \leq \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{P}^*}[J^i(\tilde{\theta}^i, \theta^{-i})], \ \forall \tilde{\theta}^i$$

- Decoupled no-regret algorithm converges to CCE
- Our goal: scalable no-regret learning algorithm

The no-regret benchmark

- Consider player i facing an unknown sequence of costs J_t^i

- Regret: $R^i(T) = \underbrace{\sum_{t=0}^{T} J_t^i(\theta_t)}_{\text{incurred cost}} - \underbrace{\min_{\theta} \sum_{t=0}^{T} J_t^i(\theta)}_{\text{best cost}}$
- Algorithm is no-regret: $R^i(T)/T \to 0$

The no-regret benchmark

- Consider player i facing an unknown sequence of costs J_t^i

- Regret: $R^i(T) = \underbrace{\sum_{t=0}^{T} J_t^i(\theta_t)}_{\text{incurred cost}} - \underbrace{\min_{\theta} \sum_{t=0}^{T} J_t^i(\theta)}_{\text{best cost}}$
- Algorithm is no-regret: $R^i(T)/T \to 0$

In games: $J_t^i(.) := J^i(., \theta_t^{-i})$ for player i

No-regret learning and equilibria

Let each player adopt a no-regret algorithm

- empirical distribution of actions $\rightarrow$ CCE
  - zero-sum games: $\rightarrow$ Nash equilibria
- regret rate $\implies$ rate of convergence to an approximate CCE

Optimal regret rates based on player's feedback

n: number of actions for player, T: number of iterations

- Bandit feedback [Auer et al. 2003]

Optimal regret rates based on player's feedback

n: number of actions for player, T: number of iterations

- Bandit feedback [Auer et al. 2003]

  - Regret $R^i(T)$ grows as $\sqrt{Tn\log n}$

Optimal regret rates based on player's feedback

n: number of actions for player, T: number of iterations

- Bandit feedback [Auer et al. 2003]

  - Regret $R^i(T)$ grows as $\sqrt{Tn\log n}$

- Full feedback [Freund et al.1997]

  - Regret $R^i(T)$ grows as $\sqrt{T\log n}$

Optimal regret rates based on player's feedback

n: number of actions for player, T: number of iterations

- Bandit feedback [Auer et al. 2003]

  - Regret $R^i(T)$ grows as $\sqrt{Tn\log n}$

- Full feedback [Freund et al.1997]

  - Regret $R^i(T)$ grows as $\sqrt{T\log n}$

Can we improve the dependence on n in a game?

Model-based no-regret learning

- Dynamics $s_{h+1} = f(s_h, a_h^1, a_h^2 \ldots, a_h^N) + \omega_h$
  - $s_h \in S \subset \mathbb{R}^p$, $a_h^i \in A^i \subset \mathbb{R}^q$
- Objective $J^i(\pi^i, \pi^{-i}) = \mathrm{E}[\sum_{h=0}^{H-1} r^i(s_h, \pi^i(s_h), \pi^{-i}(s_h))]$

f has a bounded norm in a reproducing Kernel space $\implies$
f can be modeled by a Gaussian process

Modeling class for the dynamics

- $f(x) \sim \mathcal{GP}\big(\mu(x), k(x, x')\big)$
- μ: mean, k: covariance (kernel)

$$k_{poly}(x, x') = \left(l + x^\top x'\right)^d, \; k_{SE}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{l^2}\right)$$

- mean μ_t, variance Σ_t regression based on past data

prior

posterior

Model-based learning of CCE in Markov games

- Initialize $\mathcal{P}_0 = \mathcal{P}(\boldsymbol{\theta}_0)$. For $t = 0, 1, \ldots, T$
  - sample $(\pi_t^1, \ldots, \pi_t^N) \sim \mathcal{P}_t$

  - estimate dynamics, compute corresponding CCE: $\mathcal{P}_{t+1}$

Theorem

Under Lipschitz continuity of f, $\{r^i, \pi^i\}_{i=1}^N$

$$R^i(T) = \mathcal{O}\big(\sqrt{T\mathcal{I}_T}\big)$$

- *$\mathcal{I}_T$: information gain, sublinear in p, q, N, H*

[Sessa, **MK**, Krause, ICML 2022]

Example: Multi-agent RL in autonomous driving

SMARTS autonomous car simulation environment [Zhou et al. 2021]

- testing multi-agent RL algorithms for autonomous driving
- realistic traffic data and car dynamics

Multiagent reinforcement learning for autonomous driving

- Objective: progress towards the goal, avoid collision
- Dynamics: $P(.|s_h, a_h^1, a_h^2)$
  - s: positions and velocities of cars
  - a^i: heading and speed, $i = 1, 2$
  - $\pi_{\theta^i}(s)$: parametrized by neural networks, $i = 1, 2$

The autonomous cars can coordinate and overtake the human-driven car

Implementation on multiagent autonomous car simulation environment [Zhou et al. 2021]

Average rewards for the agents

Single-agent optima

Multi-agent equilibrium

- Coordination $\implies$ less breaking, more successful merges

Figure: Green: proposed approach, black: independent Q-learning

- Model-based approach much more sample efficient

Outline

Motivation

Learning Nash equilibria in static games

No-regret learning in dynamic games

Conclusions

Summary

- Bandit learning of Nash equilibria
  - decoupled algorithm based only on function evaluations
  - challenging to extend to Markov games

- No-regret learning
  - computationally tractable and converges to CCEs
  - can improve learning rate with a model-based approach

Outlook

- Learning equilibria in Markov games under safety and reachability constraints
- Provable algorithms under partial and asymmetric information
- Learning of "good" equilibria, mechanism design
- Applications: power markets, robotics, autonomous driving

Acknowledgements

- This talk: P. Giuseppe Sessa, Salizzoni, T. Tatarenko, A. Krause

`https://www.epfl.ch/labs/sycamore/`

- Funding: ERC, Swiss National Fund, NCCR Automation