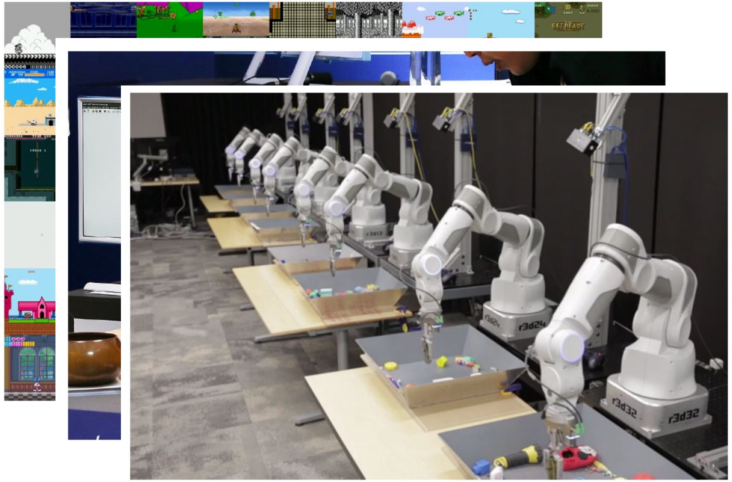
Robustness of Markov perfect equilibrium to model approximations in general-sum dynamic games

Jayakumar Subramanian, Senior Research Scientist, MDSR Lab, Adobe India
Amit Sinha, PhD Student, McGill University
Aditya Mahajan, Associate Professor, McGill University

Recent Successes of RL



Robotic grasping

- Algorithms based on comprehensive theory
- Theory restricted almost exclusively to single agent environments or models reduced to single agent environments
- Real world – strategic agents:
  - Industrial organization
  - Energy markets
  - …

How do we develop a theory for learning with strategic agents?

System Model: Markov/Stochastic/Dynamic games

- n players
- Action space: $\mathcal{A} = (\mathcal{A}^1 \times \ldots \times \mathcal{A}^n)$
- Action profile: $A_t = (A_t^1, \ldots, A_t^n) \in \mathcal{A}$
- Game state: $S_t \in \mathcal{S}$
- Game dynamics: $S_{t+1} \sim \mathcal{P}(\cdot \mid S_t, A_t)$
- Per-stage reward of player i: r^i: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Value (i.e. total reward of player i):
- $$V^i(s) = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^i(S_t, A_t) \middle| S_0 = s\right]$$

Solution Concept

Markov perfect equilibrium (MPE)

- Refinement of Nash equilibrium, where all players play (time-homogeneous) Markov policies
- Always exists for finite-state and finite-action games
- Exists under mild technical conditions, in general
- Various computational algorithms: non-linear programming, homotopy methods etc.

MPE of general-sum games is qualitatively different from zero-sum games and teams:

- A dynamic game can have multiple MPEs
- Different MPEs may have different payoff profiles

Problem Formulation

Learning MPE in games with unknown dynamics

- Suppose that the game dynamics are unknown
- … but we have access to a generative model (i.e. a system simulator) or historical data:
  - Can we learn an MPE or an approximate MPE?

Want to characterize

- <u>Robustness</u>: How robust is an MPE to model approximations?
- <u>Sample complexity</u>: How many samples do we need to learn an approximate MPE?
- <u>Regret</u>: How much better could we have done, had we known the model upfront?

Review: Markov perfect equilibrium and approximation

- (Time-homogeneous) Markov policy profile $\pi = (\pi^1, \dots, \pi^n)$, *where* $\pi^i \colon \mathcal{S} \to \Delta(\mathcal{A}^i)$
- Value of a Markov profile: $V_\pi^i(s) = (1-\gamma)\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r^i(S_t, A_t) \middle| S_0 = s\right]$

Markov perfect equilibrium (MPE)

A Markov policy profile π is a **Markov perfect equilibrium** if for all i and s:

$$V^i_{(\pi^i,\pi^{-i})}(s) \geq V^i_{(\tilde{\pi}^i\ \pi^{-i})}(s), \forall \tilde{\pi}^i \colon \mathcal{S} \to \Delta(\mathcal{A}^i)$$

Approximate MPE

Given $\alpha = (\alpha^1, \dots, \alpha^n)$, a Markov policy profile π is an α**-approximate Markov perfect equilibrium** if for all i and s:

$$V^i_{(\pi^i,\pi^{-i})}(s) \geq V^i_{(\tilde{\pi}^i\ \pi^{-i})}(s) - \alpha^i, \forall \tilde{\pi}^i \colon \mathcal{S} \to \Delta(\mathcal{A}^i)$$

Challenges of RL in general-sum dynamic games

- The Bellman operator (for single agent RL) and the minimax Bellman operator (for zero-sum games) are contractions – thereby providing convergence guarantees for learning algorithms
- However, the Nash operator is not a contraction *(Hu, Wellman 2003)*. Hence stricter conditions for convergence: All Q functions encountered in learning must satisfy one of the following very strong assumptions*(Bowling 2000)*:
  - has a NE where each player receives its maximum payoff
  - has a NE where no player benefits from the deviation of any player.
- Value-based (critic only) algorithms cannot work! Shown by (Zinkevich, Greenwald, Littman 2006)

Model-based approaches side-step all such challenges

Quantifying an Approximate Model

True model

$$(P, r)$$

Approximate model

$$(\hat{P}, \hat{r})$$

Is an MPE of the approximate model an approximate MPE of the true model?

$(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$-approximation of a game

A game $\hat{\mathcal{G}} = (\hat{P}, \hat{r})$ is an (ε, δ) - approximation of game $\mathcal{G} = (P, r)$ if for all (s, a):

$|r(s,a) - \hat{r}(s,a)| \leq \varepsilon$ and $d_{\mathfrak{F}}(P(\cdot\,|s,a), \hat{P}(\cdot|s,a)) \leq \delta$

Definition depends on the choice of metric on probability spaces

Robustness of MPE to model approximation

If $\hat{\mathcal{G}} = (\hat{P}, \hat{r})$ is an **(ε, δ)-approximation** of game $\mathcal{G} = (P, r)$ and **$\hat{\pi}$ is an MPE of $\hat{\mathcal{G}}$** then **$\hat{\pi}$ is an α-MPE of $\mathcal{G}$**

Instance dependent approximation bounds

$$\alpha^i \leq 2\left(\varepsilon + \frac{\gamma \Delta_{\hat{\pi}}^i}{(1-\gamma)}\right) \text{ where } \Delta_{\hat{\pi}}^i = max_{s \in \mathcal{S}, a \in \mathcal{A}} \left| \sum_{s \in \mathcal{S}} \left[P(s'|s,a) \hat{V}_{\hat{\pi}}^i - \hat{P}(s'|s,a) \hat{V}_{\hat{\pi}}^i \right] \right|$$

Instance independent approximation bounds

When $d_{\mathfrak{F}}$ is the total-variation metric: $\alpha^i \leq 2\left(\varepsilon + \frac{\gamma\delta\, span(\hat{r}^i)}{(1-\gamma)}\right)$

When $d_{\mathfrak{F}}$ is the Wasserstein metric: $\alpha^i \leq 2\left(\varepsilon + \frac{\gamma\delta\, L_r)}{(1-\gamma) L_P}\right)$, where L_r, L_P : Lip. constants of r, P

Learning with a generative model

How many samples do we need from the generative model to ensure that the MPE of the generated game is an α-MPE of the true game.

$$\hat{P}(s'|s,a) = \frac{\#N(s',s,a)}{\#N(s,a)}$$

Main result

For any $\alpha > 0, p > 0$, if we generate $m \geq \left\lceil \left(\frac{\gamma}{1-\gamma}\right)^2 \frac{2\log(2|\mathcal{S}|(\prod_{i=1}^{n}|\mathcal{A}^i|)n)/p}{\alpha^2} \right\rceil$ samples,
then the MPE of the generated model is an α-MPE of the true model with probability $1 - p$

Some remarks

Proof sketch

- Robustness proofs: use approximation in MDPs
- Sample complexity: bound $\Delta^i_{\hat{\pi}_m} = \left\| P\hat{V}_{\hat{\pi}_m} - \hat{P}_m\hat{V}_{\hat{\pi}_m} \right\|_\infty$ using Hoeffding's inequality

Tightness of the bounds

- For MDPs, the bound is loose by a factor of $1/(1-\gamma)$
- Tighter bounds for MDPs rely on Bernstein inequality to bound $var(\hat{V}_{\hat{\pi}_m})$ (Agarwal et al. 2020, Li et al. 2020)
- Similar bounds were adapted to zero-sum games (Zhang et al 2020) but the proof relies on the uniqueness of the minmax value.

Open question: How to establish tighter sample complexity bounds for general-sum games?

Conclusion

Model based methods side-step many of the conceptual challenges of learning in games

- **Key technical result**: Novel and general characterization of **robustness of MPE** to model approximations
- **Future directions**:
  - How to tighten sample complexity bounds?
  - How do we characterize regret?
  - What do we even mean by regret when there are multiple equilibria?

Thanks