Model based MARL for general-sum Markov games

Aditya Mahajan

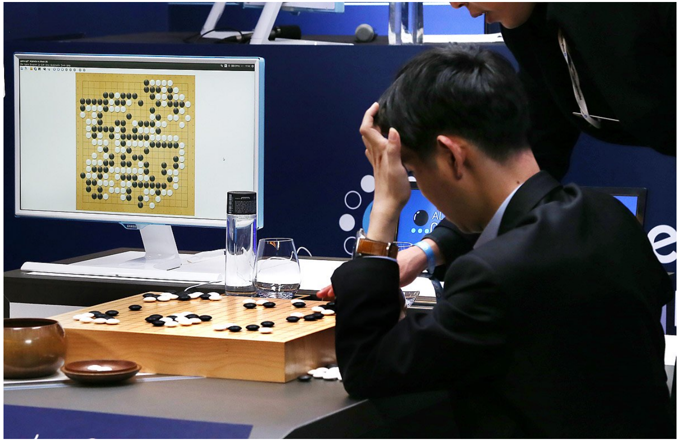
McGill University

CRM Workshop on Agents behaviour in combinatorial game theory
17th Nov 2021

- email: aditya.mahajan@mcgill.ca
- homepage: http://cim.mcgill.ca/~adityam

Recent successes of RL

Recent successes of RL

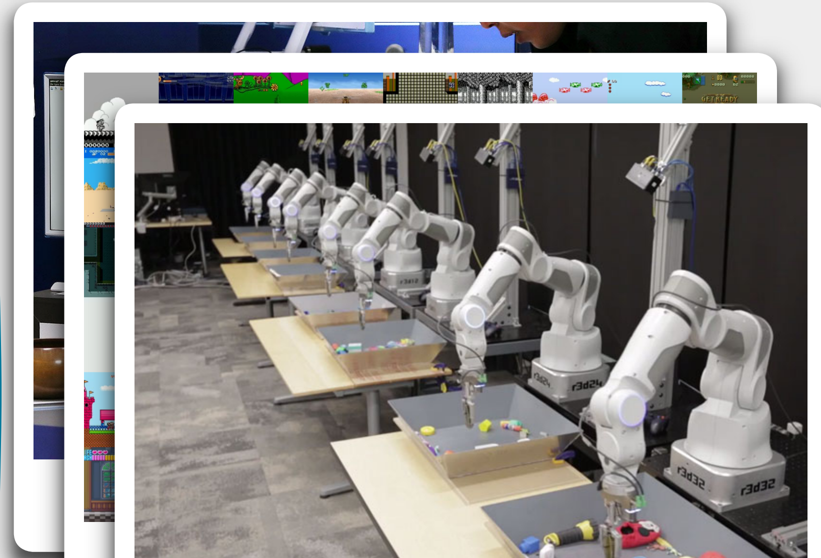


Alpha Go

Recent successes of RL

Arcade games

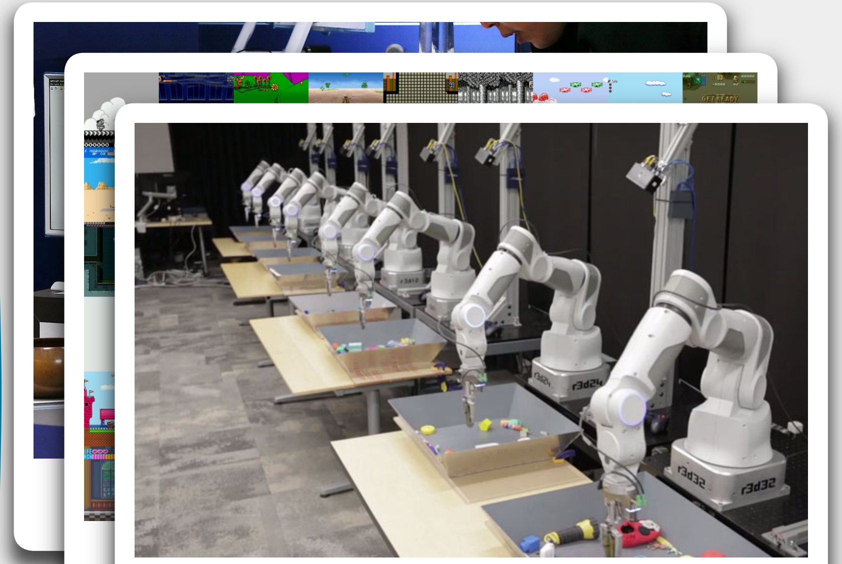
Recent successes of RL



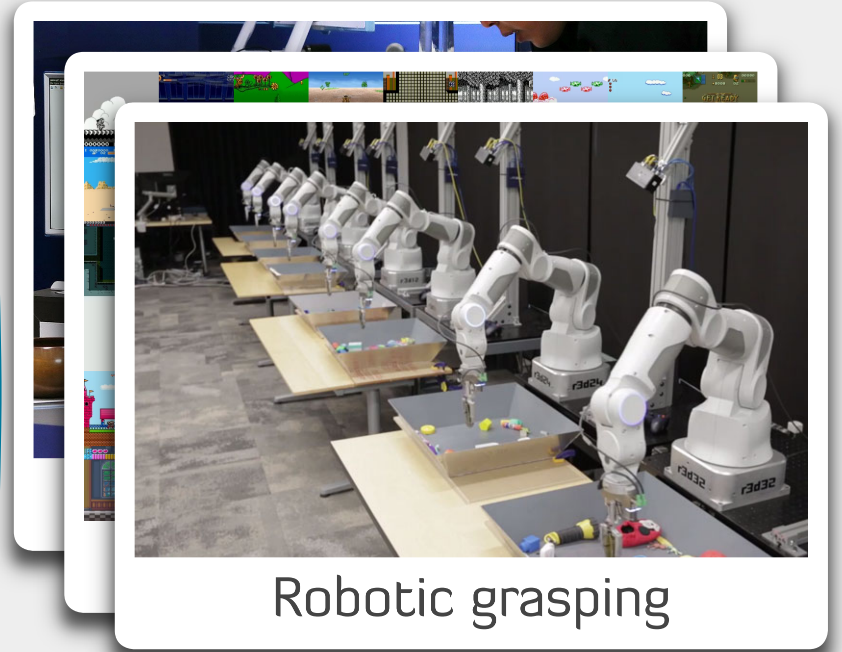
Robotic grasping

Recent successes of RL

- Algorithms based on comprehensive theory



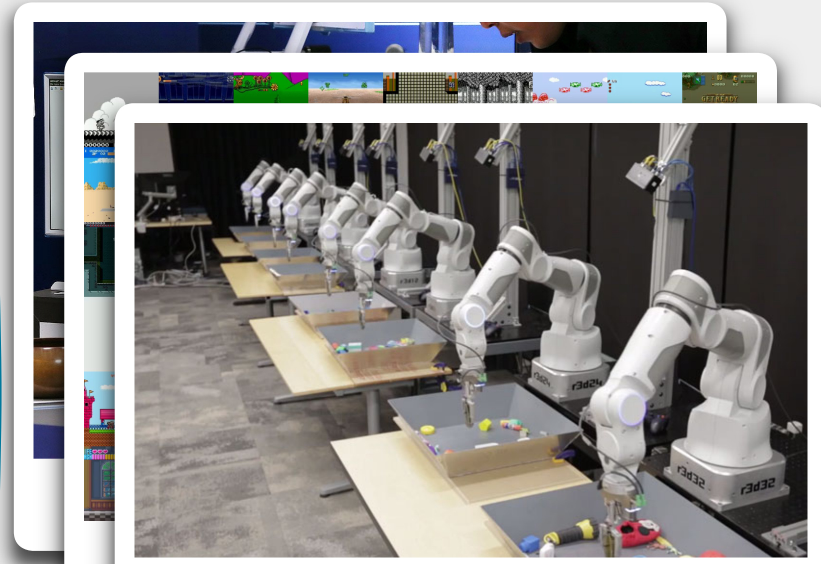
Robotic grasping



Robotic grasping

Recent successes of RL

- Algorithms based on comprehensive theory
- The theory is restricted almost exclusively to single agent envs or models which can be reduced to single agent envs.



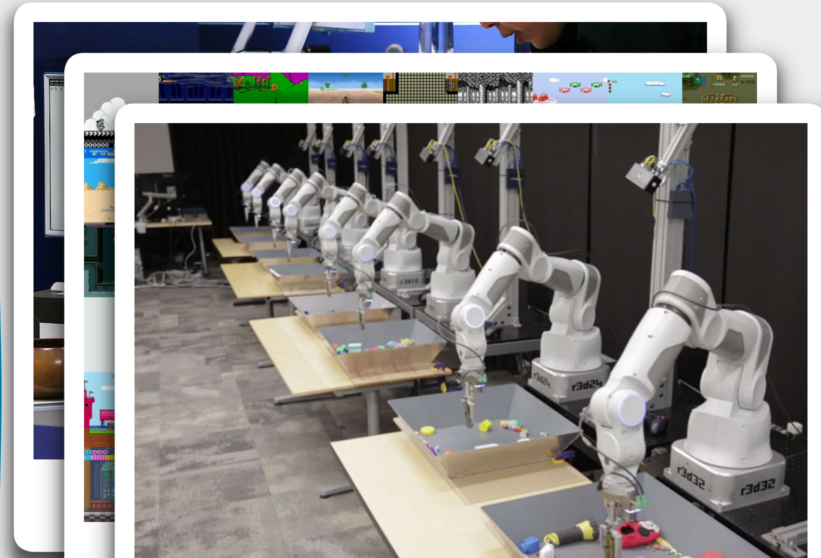
Robotic grasping

Recent successes of RL

- Algorithms based on comprehensive theory
- The theory is restricted almost exclusively to single agent envs or models which can be reduced to single agent envs.

Many real-world applications have strategic agents

- Industrial organization
- Energy markets
- Communication networks
- Cyber-security
- . . .



Robotic grasping

Recent successes of RL

- Algorithms based on comprehensive theory
- The theory is restricted almost exclusively to **single agent envs** or models which **can be reduced** to single agent envs.

Many real-world applications have **strategic agents**

- Industrial organization
- Energy markets
- Communication networks
- Cyber-security

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

Outline

System Model

- Markov/Stochastic/Dynamic game
  - Markov-perfect equilibrium
  - Approximate MPE
  - Characterization via Bellman operators

Outline

System Model

- Markov/Stochastic/Dynamic game
  - Markov-perfect equilibrium
  - Approximate MPE
  - Characterization via Bellman operators

RL in games

- Why is RL in games hard?

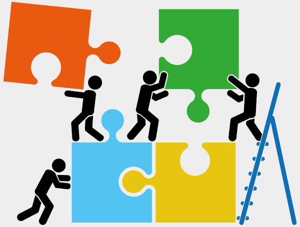
Outline

System Model

- Markov/Stochastic/Dynamic game
  - Markov-perfect equilibrium
  - Approximate MPE
  - Characterization via Bellman operators

RL in games

- Why is RL in games hard?



Model-based RL

- Robustness of MPE to model approx.
- Sample complexity bounds

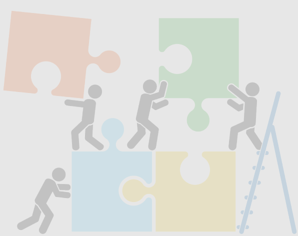
Outline

System Model

- Markov/Stochastic/Dynamic game
  - Markov-perfect equilibrium
  - Approximate MPE
  - Characterization via Bellman operators

RL in games

- Why is RL in games hard?



Model-based RL

- Robustness of MPE to model approx.
- Sample complexity bounds

System Model

Markov/Stochastic/Dynamic games

- n players.
- Action space $\mathcal{A} = (\mathcal{A}^1 \times \cdots \times \mathcal{A}^n)$.
- Action profile $A_t = (A_t^1, \dots, A_t^n) \in \mathcal{A}$.
- Game state $S_t \in \mathcal{S}$.
- Game dynamics $S_{t+1} \sim P(\cdot | S_t, A_t)$.
- Per-stage reward of player i: $r^i \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Value (i.e., total reward) of player i):

$$V^i(s) = (1-\gamma)\,\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^i(S_t, A_t) \,\middle|\, S_0 = s\right].$$

System Model

Markov/Stochastic/Dynamic games

- n players.
- Action space $\mathcal{A} = (\mathcal{A}^1 \times \cdots \times \mathcal{A}^n)$.
- Action profile $A_t = (A_t^1, \dots, A_t^n) \in \mathcal{A}$.
- Game state $S_t \in \mathcal{S}$.
- Game dynamics $S_{t+1} \sim P(\cdot | S_t, A_t)$.
- Per-stage reward of player i: $r^i \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Value (i.e., total reward) of player i):

$$V^i(s) = (1-\gamma)\,\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^i(S_t, A_t) \,\middle|\, S_0 = s\right].$$

Special cases

- **Finite horizon games:**
  Take time as part of the state space.
  Go to an absorbing state at end of horizon.
- **Zero-sum games:**
  $n = 2;\ r^1(s,a) + r^2(s,a) = 0.$
- **Teams or common-interest games**
  $r^1(s,a) = \cdots = r^n(s,a).$
- **MDPs:** $n = 1$.

Solution concept

Markov perfect equilibrium (MPE)

- Refinement of NE, 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.

Solution concept

Markov perfect equilibrium (MPE)

- Refinement of NE, 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 ZSG and teams:

- A 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:

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?

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:

- Sample complexity: How many samples do we need to learn an approximate MPE?
- Regret: How much better could we have done, had we known the model upfront?

Review: Markov perfect equilibrium

Review: Markov perfect equilibrium

- (Time-homogeneous) Markov policy profile:

$$\pi = (\pi^1, \dots, \pi^n), \quad \text{where } \pi^i \colon \mathcal{S} \to \Delta(\mathcal{A}^i).$$

Review: Markov perfect equilibrium

- (Time-homogeneous) Markov policy profile:

$$\pi = (\pi^1, \dots, \pi^n), \quad \text{where } \pi^i \colon \mathcal{S} \to \Delta(\mathcal{A}^i).$$

- Value of a Markov policy profile:

$$V^i_\pi(s) = (1-\gamma)\, \mathbb{E}_\pi \Bigg[\sum_{t=0}^{\infty} \gamma^t r^i(S_t, A_t) \,\Bigg|\, S_0 = s \Bigg]$$

Review: Markov perfect equilibrium

- (Time-homogeneous) Markov policy profile:

$$\pi = (\pi^1, \dots, \pi^n), \quad \text{where } \pi^i \colon \mathcal{S} \to \Delta(\mathcal{A}^i).$$

- Value of a Markov policy profile:

$$V^i_\pi(s) = (1-\gamma)\, \mathbb{E}_\pi \Big[\sum_{t=0}^{\infty} \gamma^t r^i(S_t, A_t) \,\Big|\, S_0 = s \Big]$$

Markov perfect equilibrium (MPE)

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

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

Review: Markov perfect equilibrium

- (Time-homogeneous) Markov policy profile:

$$\pi = (\pi^1, \dots, \pi^n), \quad \text{where } \pi^i \colon \mathcal{S} \to \Delta(\mathcal{A}^i).$$

- Value of a Markov policy profile:

$$V^i_\pi(s) = (1-\gamma)\, \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t r^i(S_t, A_t) \,\Big|\, S_0 = s \Big]$$

Markov perfect equilibrium (MPE)

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

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

Approximate MPE

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

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

Alternative characterization: Bellman operators

Bellman operators

- Given Markov policy profile π, define $\mathcal{B}^i_\pi \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_\pi v](s) = \sum_{a \in \mathcal{A}} \pi(a|s) \Big[(1-\gamma) r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \Big]$$

Alternative characterization: Bellman operators

Bellman operators

- Given Markov policy profile π, define $\mathcal{B}^i_\pi \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_\pi v](s) = \sum_{a \in \mathcal{A}} \pi(a|s) \Big[(1-\gamma) r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \Big]$$

- Given Markov policy profile π, define $\mathcal{B}^i_{(*,\pi^{-i})} \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_{(*,\pi^{-i})} v](s) = \max_{a^i \in \mathcal{A}^i} \sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \Big[(1-\gamma) r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \Big]$$

Alternative characterization: Bellman operators

Bellman operators

- Given Markov policy profile π, define $\mathcal{B}^i_\pi \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_\pi v](s) = \sum_{a \in \mathcal{A}} \pi(a|s) \Big[(1-\gamma) r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \Big]$$

- Given Markov policy profile π, define $\mathcal{B}^i_{(*,\pi^{-i})} \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_{(*,\pi^{-i})} v](s) = \max_{a^i \in \mathcal{A}^i} \sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \Big[(1-\gamma) r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \Big]$$

Fixed-point

$$\boxed{V^i_\pi = \mathcal{B}^i_\pi V^i_\pi}$$

Alternative characterization: Bellman operators

Bellman operators

- Given Markov policy profile π, define $\mathcal{B}^i_\pi \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_\pi v](s) = \sum_{a \in \mathcal{A}} \pi(a|s) \Big[(1-\gamma) r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \Big]$$

Fixed-point

$$\boxed{V^i_\pi = \mathcal{B}^i_\pi V^i_\pi}$$

- Given Markov policy profile π, define $\mathcal{B}^i_{(*,\pi^{-i})} \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_{(*,\pi^{-i})} v](s) = \max_{a^i \in \mathcal{A}^i} \sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \Big[(1-\gamma) r^i(s,a) +$$

Fixed-point

$$\boxed{V^i_{(*,\pi^{-i})} = \mathcal{B}^i_{(*,\pi^{-i})} V^i_{(*,\pi^{-i})}}$$

Alternative characterization: Bellman operators

Bellman operators

- Given Markov policy profile π, define $\mathcal{B}^i_\pi \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_\pi v](s) = \sum_{a \in \mathcal{A}} \pi(a|s) \Big[(1-\gamma) r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \Big]$$

Fixed-point

$$\boxed{V^i_\pi = \mathcal{B}^i_\pi V^i_\pi}$$

- Given Markov policy profile π, define $\mathcal{B}^i_{(*,\pi^{-i})} \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_{(*,\pi^{-i})} v](s) = \max_{a^i \in \mathcal{A}^i} \sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \Big[(1-\gamma) r^i(s,a) +$$

Fixed-point

$$\boxed{V^i_{(*,\pi^{-i})} = \mathcal{B}^i_{(*,\pi^{-i})} V^i_{(*,\pi^{-i})}}$$

MPE

A policy π is an MPE if for all i

$$V^i_\pi = V^i_{(*,\pi^{-i})}$$

Alternative characterization: Bellman operators

Bellman operators

- Given Markov policy profile π, define $\mathcal{B}^i_\pi \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_\pi v](s) = \sum_{a \in \mathcal{A}} \pi(a|s) \Big[(1-\gamma) r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \Big]$$

Fixed-point

$$\boxed{V^i_\pi = \mathcal{B}^i_\pi V^i_\pi}$$

- Given Markov policy profile π, define $\mathcal{B}^i_{(*,\pi^{-i})} \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as:

$$[\mathcal{B}^i_{(*,\pi^{-i})} v](s) = \max_{a^i \in \mathcal{A}^i} \sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \Big[(1-\gamma) r^i(s,a) +$$

Fixed-point

$$\boxed{V^i_{(*,\pi^{-i})} = \mathcal{B}^i_{(*,\pi^{-i})} V^i_{(*,\pi^{-i})}}$$

MPE

A policy π is an MPE if for all i

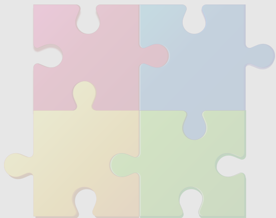
$$V^i_\pi = V^i_{(*,\pi^{-i})}$$

α-MPE

A policy π is an α-MPE if for all i

$$V^i_\pi = V^i_{(*,\pi^{-i})} - \alpha^i$$

Outline

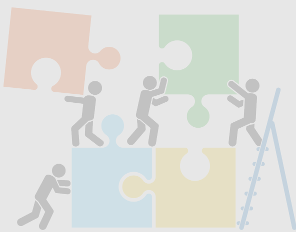


System Model

- Markov/Stochastic/Dynamic game
  - Markov-perfect equilibrium
  - Approximate MPE
  - Characterization via Bellman operators

RL in games

- Why is RL in games hard?



Model-based RL

- Robustness of MPE to model approx.
- Sample complexity bounds

Review: How does RL (Q-learning) work in MDPs?

Expand the Bellman operator

$$V(s) = \max_{a \in \mathcal{A}} Q(s, a)$$

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

Review: How does RL (Q-learning) work in MDPs?

Expand the Bellman operator

$$V(s) = \max_{a \in \mathcal{A}} Q(s, a)$$

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

Approximate via stochastic approximation

$$Q(s, a) \leftarrow Q(s, a) + \alpha \Big[r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q(s_+, a') - Q(s, a) \Big]$$

Review: How does RL (Q-learning) work in MDPs?

Expand the Bellman operator

$$V(s) = \max_{a \in \mathcal{A}} Q(s, a)$$

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

Approximate via stochastic approximation

$$Q(s, a) \leftarrow Q(s, a) + \alpha \Big[r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q(s_+, a') - Q(s, a) \Big]$$

unbiased sample

Review: How does RL (Q-learning) work in MDPs?

Expand the Bellman operator

$$V(s) = \max_{a \in \mathcal{A}} Q(s, a)$$

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

Approximate via stochastic approximation

$$Q(s, a) \leftarrow Q(s, a) + \alpha \Big[r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q(s_+, a') - Q(s, a) \Big]$$

unbiased sample

Why does Q-learning converge?

- Under approprpriate technical conditions, SA tracks an ODE (Borkar 1997).
- Since the Bellman operator is a contraction, the ODE has a unique equilibrium point which is globally asymptotically stable (Borkar and Soumyanatha, 1997).

Review: How does RL (Q-learning) work in zero-sum games?

Expand the Bellman operator

$$V(s) = \max_{a^1 \in \mathcal{A}^1} \min_{a^2 \in \mathcal{A}^2} Q(s, (a^1, a^2))$$

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

Review: How does RL (Q-learning) work in zero-sum games?

Expand the Bellman operator

$$V(s) = \max_{a^1 \in \mathcal{A}^1} \min_{a^2 \in \mathcal{A}^2} Q(s, (a^1, a^2))$$

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

Approximate via stochastic approximation

Use $r(s, a) + \gamma \max_{a^1 \in \mathcal{A}^1} \min_{a^2 \in \mathcal{A}^2} Q(s_+, (a^1, a^2))$

unbiased sample

Review: How does RL (Q-learning) work in zero-sum games?

Expand the Bellman operator

$$V(s) = \max_{a^1 \in \mathcal{A}^1} \min_{a^2 \in \mathcal{A}^2} Q(s, (a^1, a^2))$$

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

Approximate via stochastic approximation

Use $r(s, a) + \gamma \max_{a^1 \in \mathcal{A}^1} \min_{a^2 \in \mathcal{A}^2} Q(s_+, (a^1, a^2))$

unbiased sample

Minimax Q-learning (Littman 1994)

Review: How does RL (Q-learning) work in zero-sum games?

Expand the Bellman operator

$$V(s) = \max_{a^1 \in \mathcal{A}^1} \min_{a^2 \in \mathcal{A}^2} Q(s, (a^1, a^2))$$

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

Approximate via stochastic approximation

Use $r(s, a) + \gamma \max_{a^1 \in \mathcal{A}^1} \min_{a^2 \in \mathcal{A}^2} Q(s_+, (a^1, a^2))$

unbiased sample

Minimax Q-learning (Littman 1994)

Why does Minimax Q-learning converge?

- Exactly same reason as before.
- The important part is that the minimax Bellman operator is a contraction

Review: How does RL (Q-learning) work in general-sum games?

Expand the Bellman operator

$$V(s) = \underset{a \in \mathcal{A}}{\mathrm{Nash}}\, Q(s, a)$$

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

Review: How does RL (Q-learning) work in general-sum games?

Expand the Bellman operator

$$V(s) = \underset{a \in \mathcal{A}}{\text{Nash}}\, Q(s, a)$$

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

Approximate via stochastic approximation

Use $r(s, a) + \gamma \underset{a \in \mathcal{A}}{\text{Nash}}\, Q(s_+, a)$

unbiased sample

Review: How does RL (Q-learning) work in general-sum games?

Expand the Bellman operator

$$V(s) = \operatorname*{Nash}_{a \in \mathcal{A}} Q(s, a)$$

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

Approximate via stochastic approximation

Use $r(s, a) + \gamma \operatorname*{Nash}_{a \in \mathcal{A}} Q(s_+, a)$

unbiased sample

Nash Q-learning (Hu Wellman 2003)

Review: How does RL (Q-learning) work in general-sum games?

Expand the Bellman operator

$$V(s) = \underset{a \in \mathcal{A}}{\text{Nash}} Q(s, a)$$

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

Approximate via stochastic approximation

Use $r(s, a) + \gamma \underset{a \in \mathcal{A}}{\text{Nash}} Q(s_+, a)$

unbiased sample

Nash Q-learning (Hu Wellman 2003)

How to guanratee convergence?

- The Nash operator is not a contraction. Need to assume that all Q-functions encountered during learning 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.
- Few known examples other than zero-sum games or common interest games.

Other challenges with RL in general-sum games

Policy evaluation Bellman equaitons

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) Q_\pi(s, a)$$

$$Q_\pi(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V_\pi(s')$$

Other challenges with RL in general-sum games

Policy evaluation Bellman equaitons

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) Q_\pi(s, a)$$

$$Q_\pi(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V_\pi(s')$$

NoSDE games (Zinkevich, Greenwald, Littman 2006)

- A specific family of general-sum games with the following properties:
  - The game has a unique MPE in mixed strategies.
  - For any game $\mathcal{G} = \langle \mathcal{S}, \mathcal{A}, P, r \rangle$ with unique MPE strategy π, there exists another NoSDE game $\mathcal{G}' = \langle \mathcal{S}, \mathcal{A}, P, r' \rangle$ with unique MPE strategy π' such that

$$\pi \neq \pi' \text{ and } V_\pi^{\mathcal{G}} \neq V_{\pi'}^{\mathcal{G}'} \quad \text{but} \quad Q_\pi^{\mathcal{G}} = Q_{\pi'}^{\mathcal{G}'}$$

Other challenges with RL in general-sum games

Policy evaluation Bellman equaitons

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) Q_\pi(s, a)$$

$$Q_\pi(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V_\pi(s')$$

Implications

- Value-based (critic only) algorithms cannot work!
- Lot of the follow-up literature focuses on other solution concepts: cyclic equilibrium, correlated equilibrium, etc.

NoSDE games (Zinkevich, Greenwald, Littman 2006)

- A specific family of general-sum games with the following properties:
  - The game has a unique MPE in mixed strategies.
  - For any game $\mathcal{G} = \langle \mathcal{S}, \mathcal{A}, P, r \rangle$ with unique MPE strategy π, there exists another NoSDE game $\mathcal{G}' = \langle \mathcal{S}, \mathcal{A}, P, r' \rangle$ with unique MPE strategy π' such that

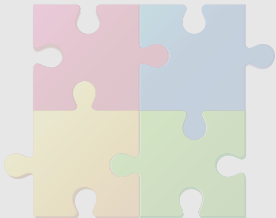
$$\pi \neq \pi' \text{ and } V_\pi^{\mathcal{G}} \neq V_{\pi'}^{\mathcal{G}'} \quad \text{but} \quad Q_\pi^{\mathcal{G}} = Q_{\pi'}^{\mathcal{G}'}$$

Simple observation: Model-based approaches side-step all such challenges.

We characterize sample-complexity bounds

- co-author: Jayakumar Subramanian and Amit Sinha
- paper: https://arxiv.org/abs/2110.02355

Outline

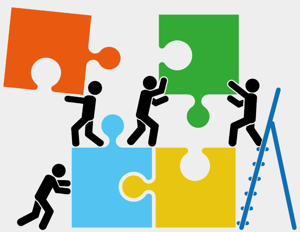


System Model

- Markov/Stochastic/Dynamic game
  - Markov-perfect equilibrium
  - Approximate MPE
  - Characterization via Bellman operators

RL in games

- Why is RL in games hard?



Model-based RL

- Robustness of MPE to model approx.
- Sample complexity bounds

Quantifying an approximate model

True model

$$(P, r)$$

Approx. model

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

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

Quantifying an approximate model

True model

(P, r)

Approx. model

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

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

(ε, δ)-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):

$$\left| r(s, a) - \hat{r}(s, a) \right| \leqslant \varepsilon \quad \text{and} \quad d_{\mathfrak{F}}\big(P(\cdot|s, a), \hat{P}(\cdot|s, a)\big) \leqslant \delta$$

Quantifying an approximate model

True model

$$(P, r)$$

Approx. model

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

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

(ε, δ)-approximation of a game

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

$$\big| r(s,a) - \widehat{r}(s,a) \big| \le \varepsilon \quad \text{and} \quad \boxed{d_{\mathfrak{F}}}\big(P(\cdot|s,a), \widehat{P}(\cdot|s,a)\big) \le \delta$$

Definition depend on the choice of metric on probability spaces

Robustness of MPE to model approximation

$$\text{If} \left\{ \begin{array}{c} \widehat{\mathcal{G}} \text{ is an } (\varepsilon, \delta)\text{-approximation of } \mathcal{G} \\ \text{and} \\ \hat{\pi} \text{ is an MPE of } \widehat{\mathcal{G}} \end{array} \right\} \text{ then } \hat{\pi} \text{ is an } \alpha\text{-MPE of } \mathcal{G}$$

Robustness of MPE to model approximation

$$\text{If} \left\{ \begin{array}{c} \hat{\mathcal{G}} \text{ is an } (\varepsilon, \delta)\text{-approximation of } \mathcal{G} \\ \text{and} \\ \hat{\pi} \text{ is an MPE of } \hat{\mathcal{G}} \end{array} \right\} \text{ then } \hat{\pi} \text{ is an } \alpha\text{-MPE of } \mathcal{G}$$

Instance dependent approximation bounds

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

Robustness of MPE to model approximation

$$\text{If} \left\{ \begin{array}{c} \hat{\mathcal{G}} \text{ is an } (\varepsilon, \delta)\text{-approximation of } \mathcal{G} \\ \text{and} \\ \hat{\pi} \text{ is an MPE of } \hat{\mathcal{G}} \end{array} \right\} \text{ then } \hat{\pi} \text{ is an } \alpha\text{-MPE of } \mathcal{G}$$

Instance dependent approximation bounds

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

Succintly, $\Delta^i_{\hat{\pi}} = \left\| P\,\hat{V}^i_{\hat{\pi}} - \hat{P}\,\hat{V}^i_{\hat{\pi}} \right\|_\infty$

Robustness of MPE to model approximation

If $\left\{ \begin{array}{c} \hat{\mathcal{G}} \text{ is an } (\varepsilon, \delta)\text{-approximation of } \mathcal{G} \\ \text{and} \\ \hat{\pi} \text{ is an MPE of } \hat{\mathcal{G}} \end{array} \right\}$ then $\hat{\pi}$ is an α-MPE of $\mathcal{G}$

Instance **dependent** approximation bounds

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

Instance **independent** approximation bounds

Robustness of MPE to model approximation

$$\text{If } \left\{ \begin{array}{c} \hat{\mathcal{G}} \text{ is an } (\varepsilon, \delta)\text{-approximation of } \mathcal{G} \\ \text{and} \\ \hat{\pi} \text{ is an MPE of } \hat{\mathcal{G}} \end{array} \right\} \text{ then } \hat{\pi} \text{ is an } \alpha\text{-MPE of } \mathcal{G}$$

Instance dependent approximation bounds

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

Instance independent approximation bounds

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

Robustness of MPE to model approximation

$$\text{If} \left\{ \begin{array}{c} \hat{\mathcal{G}} \text{ is an } (\varepsilon, \delta)\text{-approximation of } \mathcal{G} \\ \text{and} \\ \hat{\pi} \text{ is an MPE of } \hat{\mathcal{G}} \end{array} \right\} \text{ then } \hat{\pi} \text{ is an } \alpha\text{-MPE of } \mathcal{G}$$

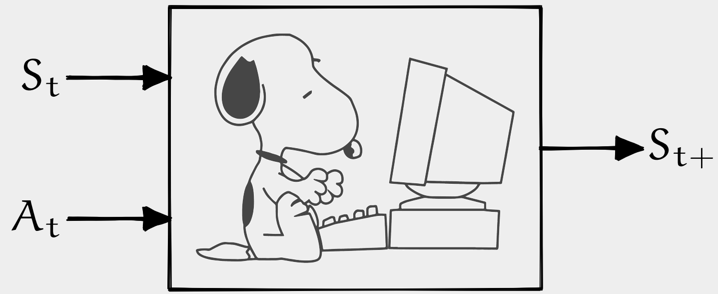
Instance dependent approximation bounds

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

Instance independent approximation bounds

- When $d_{\mathfrak{F}}$ is total-variation metric: $\alpha^i \leqslant 2\left(\varepsilon + \dfrac{\gamma \delta \operatorname{span}(\hat{r}^i)}{(1-\gamma)}\right)$
- When $d_{\mathfrak{F}}$ is Wasserstein metric: $\alpha^i \leqslant 2\left(\varepsilon + \dfrac{\gamma \delta L_r}{(1-\gamma L_P)}\right)$, where $\left\{ \begin{array}{l} L_r : \text{Lip. constant of } r \\ L_P : \text{Lip. constant of } P \end{array} \right.$

Learning with a generative model

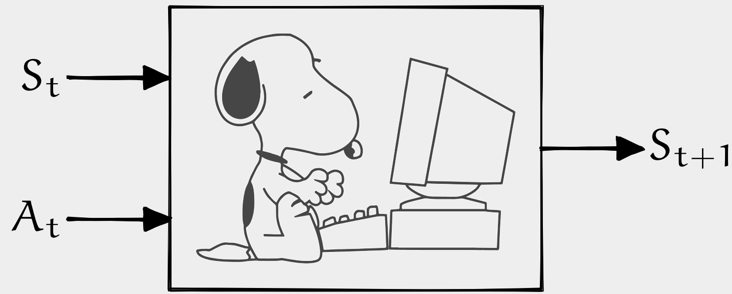


$\widehat{P}$ estimated from generated samples

$\widehat{P}(s'|s, a) = \#N(s', s, a)/\#N(s, a)$

Learning with a generative model

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

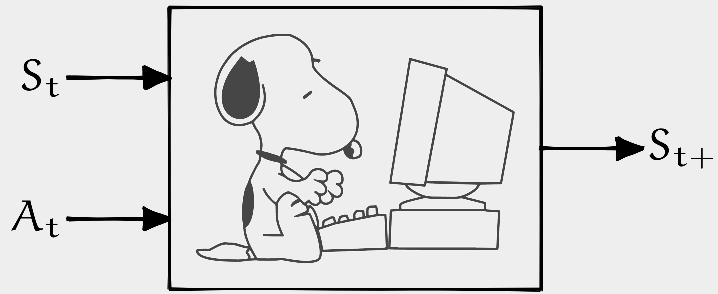


$\hat{P}$ estimated from generated samples

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

Learning with a generative model

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



$\hat{P}$ estimated from generated samples

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

Main Result

For any $\alpha > 0$ and $p > 0$, if we generate

$$m \geqslant \left\lceil \left(\frac{\gamma}{1-\gamma}\right)^2 \frac{2\log(2|\mathcal{S}|(\prod_{i=1}^{n} |\mathcal{A}^i|)n)/p}{\alpha^2} \right\rceil$$

samples for each state action pair, then the MPE of the generated model is an α-MPE of the true model with probability $1-p$.

Some remarks

Proof sketch

- In the robustness result, 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 inequality.

Some remarks

Proof sketch

- In the robustness result, 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 inequality.

Tightness of the bounds

- For MDPs ($n = 1$), the bound is loose by a factor of $1/(1-\gamma)$.

Some remarks

Proof sketch

- In the robustness result, 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 inequality.

Tightness of the bounds

- For MDPs ($n = 1$), the bound is loose by a factor of $1/(1-\gamma)$.
- Tighter bounds for MDPs rely on Bernstein inequality to bound $\text{var}(\hat{V}_{\hat{\pi}_m})$ (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

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

Conclusion

Takeaway message: 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.

Conclusion

Takeaway message: 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 the sample complexity bounds?
- How do we characterize regret?
- . . . What do we even mean by regret when there are multiple equilibria?

- **email**: aditya.mahajan@mcgill.ca
- **web**: http://cim.mcgill.ca/~adityam

Thank you

Funding

- NSERC Discovery
- DND IDEaS Network

References

- **Approx for POMDPs**: https://arxiv.org/abs/2010.08843
- **Approx for Games**: https://arxiv.org/abs/2009.12367