On the Foundation of **Distributionally Robust Reinforcement Learning**

Shengbo Wang (Stanford)

Nian Si (Chicago Booth)

Jose Blanchet (Stanford)

Zhengyuan Zhou (NYU Stern)

shengbo.wang@stanford.edu

Background: Foundation of RL & MDP

- Markov decision processes (MDPs) are foundational to Reinforcement learning (RL).
- *Dynamic Programming Principle* (DPP) is fundamental both in theory and practice, central to algorithm designs.

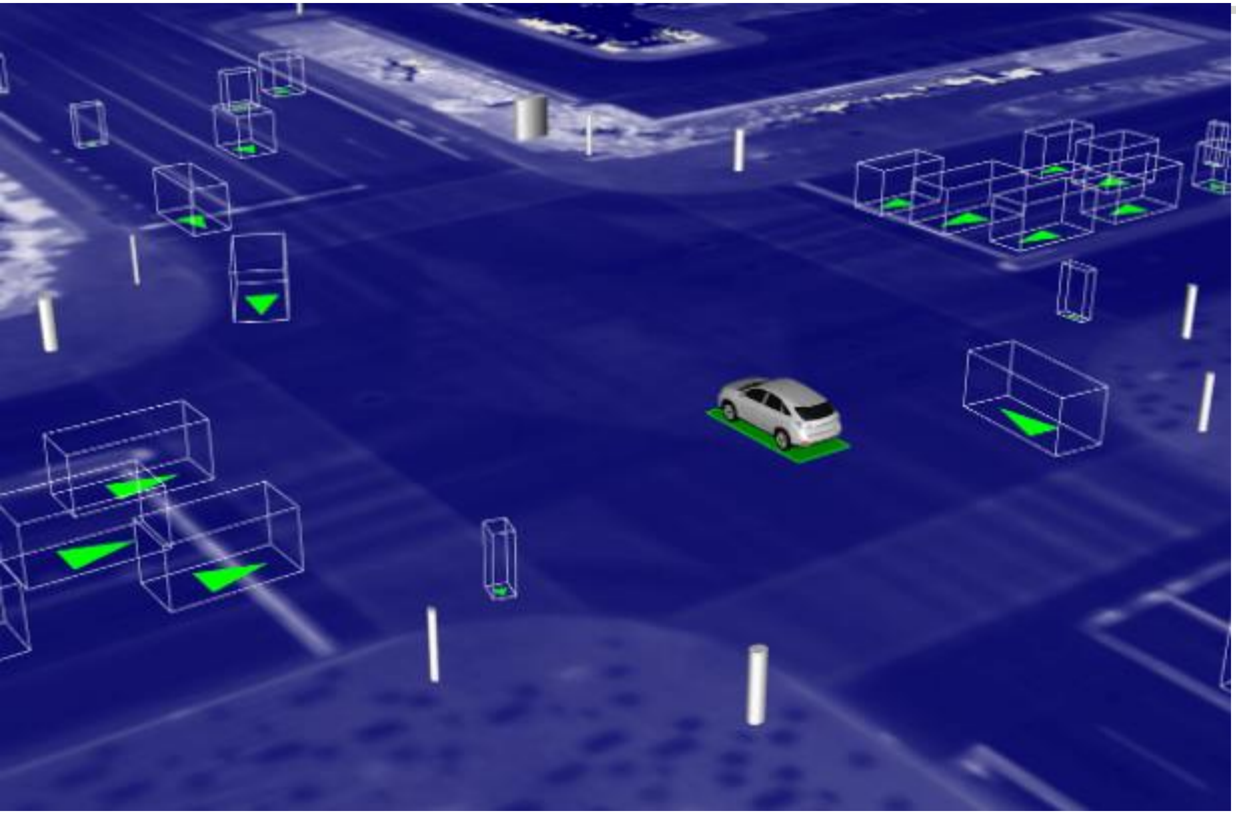
Background: Foundation of RL & MDP

- Markov decision processes (MDPs) are foundational to Reinforcement learning (RL).
- *Dynamic Programming Principle* (DPP) is fundamental both in theory and practice, central to algorithm designs.
- Key consequence of DPP:
  - Deterministic Markov policies are optimal.
  - In the infinite horizon discounted case, stationary policies are optimal.

Distributionally Robust Reinforcement Learning

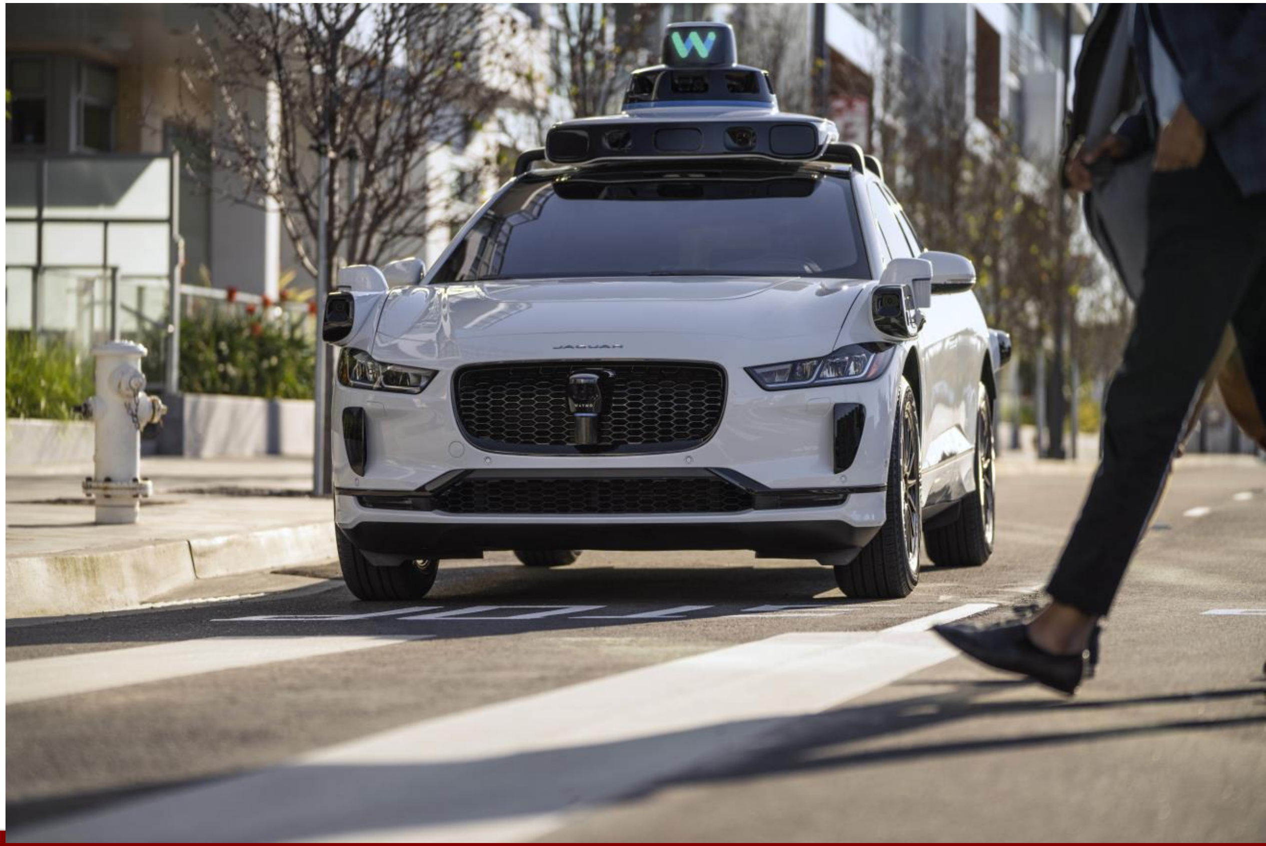
- *Distributionally robust RL (DRRL)* is an emerging area in RL.
- Motivation from classical control & RL:
  - Discrepancies between simulated training environment and deployment environment.
  - Unobserved confounders can disqualify Markov optimality, making the optimal policy of the MDP untrustworthy.
  - Full POMDP formulations may be too difficult to construct or solve, and usually lead to history-dependent controls.

Example Autonomous Driving…

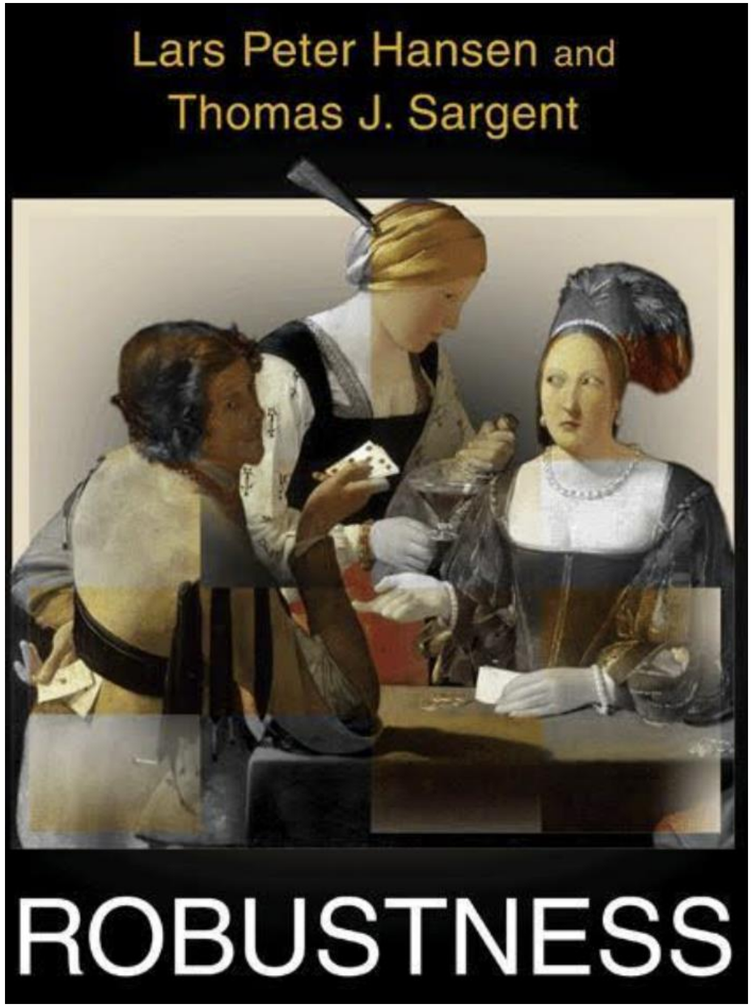


Simulator

Real environment



Model Misspecification: Adversarial Approach



- Adversarial distributional shifts.

- Static case is well studied (although still important questions remain)…

- Our focus is on the ***dynamic*** case.
- We adopt a dynamic "game" formulation.
  Two players: **Controller** v.s. **Adversary**

DRRL & DRMDP

DRMDP: Foundation of DRRL.

Expressiveness:

- Can be used to model a rich family of dynamic learning problems.

Effectiveness:

- Adversarial robustness leads to conservatism.
- Need to restrict the power of the adversary.

Tractability:

- Dynamic programming principle (DPP) for DRMDP.
- In the literature, such DPP is ***assumed*** or applied under the assumption that may not guarantee a DPP.

shengbo.wang@stanford.edu

Goal:

Study if DPP holds or not for DRMDPs under a wide range of *attributes* of the controller and the adversary.

shengbo.wang@stanford.edu

The Infinite Horizon Discount Case

- Controlled Markov chain on finite state action spaces S, A: $\{X_k, A_k : k \geq 0\}$
- Transition probabilities: $\{p(s'|s,a) : s, s' \in S, a \in A\}$
- Reward function: $r(s,a)$

shengbo.wang@stanford.edu

The Infinite Horizon Discount Case

- Controlled Markov chain on finite state action spaces S, A: $\{X_k, A_k : k \geq 0\}$
- Transition probabilities: $\{p(s'|s,a) : s, s' \in S, a \in A\}$
- Reward function: $r(s,a)$
- Standard v.s. DRMDP

$$\sup_{\pi \in \Pi} E_s^{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r(X_t, A_t)\right] \quad \text{v.s.} \quad \sup_{\pi \in \Pi} \inf_{\kappa \in \mathrm{K}} E_s^{\pi,\kappa}\left[\sum_{k=0}^{\infty} \gamma^k r(X_t, A_t)\right]$$

DRMDP

$$\sup_{\pi\in\Pi}\inf_{\kappa\in K} E_s^{\pi,\kappa}\left[\sum_{k=0}^{\infty}\gamma^k r(X_t, A_t)\right]$$

- Π: the set of admissible controls (to be discussed = tbd).
- K: the (constrained) set of adversarial policies (tbd).
- The law of $\{X_k, A_k : k \geq 0\}$ is determined by (π, κ).

Related Literature

- Stochastic games: [Shapley 1953] [Solan and vieille 2015] [Hansen-Sargent 2008]. Often focus on both adversary and controller history-dependent.
- Robust control formulation: [Gonzalez-Trejo et al. 2002], [Huang et al. 2017], [Shapiro 2021]. Both the adversary and the controller are history dependent. The adversary sees the action realized by the controller.
- DRMDP: [Nilim and El Ghaoui 2005], [Iyengar 2005], [Xu and Mannor 2010], [Wiesemann et al. 2013]. Markov or time-homogeneous adversary vs history dependent controller.
- [Xu and Mannor 2010], [Wiesemann et al. 2013] ***further constraints the adversary cannot see the controller's realized action at current time***. <u>Convex action set for the adversary.</u>

Attributes

Attributes of Controller & Adversary

$$\sup_{\pi \in \Pi} \inf_{\kappa \in \mathrm{K}} E_s^{\pi,\kappa} \left[\sum_{k=0}^{\infty} \gamma^k r(X_t, A_t) \right]$$

- Information strucrure: History-dependent v.s. Markov v.s. Time-homogeneous (for both the controller and the adversary).
- Randomized v.s. deterministic controller policies.
- Adversary admissible set.
- Does the adversary see the *realized* action of the controller?

Information Asymmetries

$$\pi = (\pi_t : t \geq 0) \in \Pi$$

- History-dependent:

  $$\pi_t(a_t | x_0, a_0, \ldots, x_t)$$

- Markov:

  $$\pi_t(a_t | x_t)$$

- Time-homogeneous:

  $$\pi(a_t | x_t)$$

$$\kappa = (\kappa_t : t \geq 0) \in \mathrm{K}$$

- History-dependent:

  $$\kappa_t(x_{t+1} | x_0, a_0, \ldots, x_t, a_t)$$

- Markov:

  $$\kappa_t(x_{t+1} | x_t, a_t)$$

- Time-homogeneous:

  $$\kappa(x_{t+1} | x_t, a_t)$$

Probability on the Path Space

$$\sup_{\pi \in \Pi} \inf_{\kappa \in \mathrm{K}} E_s^{\pi,\kappa}\left[\sum_{k=0}^{\infty} \gamma^k r(X_t, A_t)\right]$$

$$\pi = (\pi_t : t \geq 0) \in \Pi \qquad \kappa = (\kappa_t : t \geq 0) \in \mathrm{K}$$

- Path until time t:

$$B = \{A_0 = a_0, X_1 = x_1 \ldots X_t = x_t, A_t = a_t\}$$

- Probability of path

$$P_s^{\pi,\kappa}(B) := \pi_0(a_0|s)\kappa_0(x_1|s, a_0) \ldots \kappa_{t-1}(x_t|g_{t-1})\pi_t(a_t|h_t)$$

where $g_t = (x_0, a_0, \ldots, x_t, a_t), \quad h_t = (x_0, a_0, \ldots, x_t).$

- Same for asymmetric information structures.

Constrained Controller

- History-dependent: $\pi_t(a_t | x_0, a_0, \ldots, x_t)$
- Constraint: $\pi_t(\cdot | x_0, a_0, \ldots, x_t) \in \mathcal{Q} \subset \mathcal{P}(A)$
  - e.g. deterministic $\mathcal{Q} = \{\delta_a : a \in A\}$, **non-convex**
  - or fully randomized $\mathcal{Q} = \mathcal{P}(A)$ policies, **convex**.

Adversary: Rectangularity

- A crucial way to constraint the adversary.
- S- v.s. SA-rectangular adversary.

S-Rectangularity: Motivating Example

- Inventory control: $X_{t+1} = (X_t + A_t - K_t)_+$
  - $\{K_t : t \geq 0\}$ is the demand process and A_t is the ordered inventory.
  - Adversary changes K_t.

S-Rectangularity: Motivating Example

- Inventory control: $X_{t+1} = (X_t + A_t - K_t)_+$
  - $\{K_t : t \geq 0\}$ is the demand process and A_t is the ordered inventory.
  - Adversary changes K_t.
- SA-rectangularity: Observe X_t, A_t and then choose K_t.

S-Rectangularity: Motivating Example

- Inventory control: $X_{t+1} = (X_t + A_t - K_t)_+$
  - $\{K_t: t \geq 0\}$ is the demand process and A_t is the ordered inventory.
  - Adversary changes K_t.
- SA-rectangularity: Observe X_t, A_t and then choose K_t.
- More natural to assume that, the adversary can only observe X_t but not the controller's realized action A_t.
- S-rectangularity.

Constrained SA- and S-Rectangular Adversary

- SA-rectangularity: Adversary observes both state and action (s,a) and selects $p(\cdot|s,a) \in \mathcal{P}_{s,a}$ from $\mathcal{P}_{s,a} \subset \mathcal{P}(S)$
- S-rectangularity: Adversary observes only state s and selects $p(\cdot|s,\cdot) \in \mathcal{P}_s$ from $\mathcal{P}_s \subset \{A \to \mathcal{P}(S)\}$

Constrained SA- and S-Rectangular Adversary

- SA-rectangularity: Adversary observes both state and action (s,a) and selects $p(\cdot|s,a) \in \mathcal{P}_{s,a}$ from $\mathcal{P}_{s,a} \subset \mathcal{P}(S)$
- S-rectangularity: Adversary observes only state s and selects $p(\cdot|s,\cdot) \in \mathcal{P}_s$ from $\mathcal{P}_s \subset \{A \to \mathcal{P}(S)\}$
- Here $\mathcal{P}_{s,a} \subset \mathcal{P}(S)$, $\mathcal{P}_s \subset \{A \to \mathcal{P}(S)\}$ are prescribed (designed by the modeler) sets.
  - E.g. $\mathcal{P}_{s,a}(\delta) = \{\mu \in \mathcal{P}(S) : d(\mu, \mu_{0,s,a}) \leq \delta\}$

SA- and S-Rectangularity

- History-dependent SA-Rectangular adversary chooses:

$$\kappa_t(\cdot|x_0, \ldots, x_t, a_t) \in \mathcal{P}_{x_t,a_t} \subset \mathcal{P}(S)$$

SA- and S-Rectangularity

- History-dependent SA-Rectangular adversary chooses:

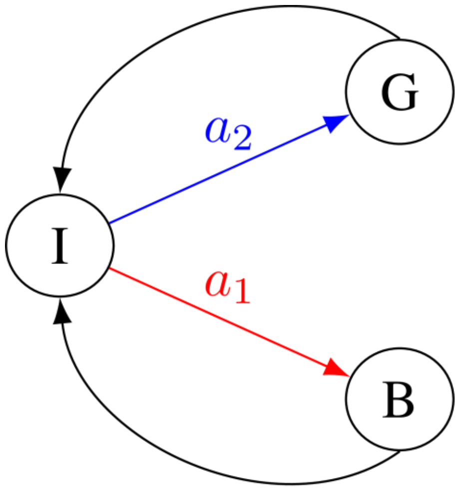
$$\kappa_t(\cdot|x_0, \ldots, x_t, a_t) \in \mathcal{P}_{x_t, a_t} \subset \mathcal{P}(S)$$

- History-dependent S-Rectangular adversary chooses

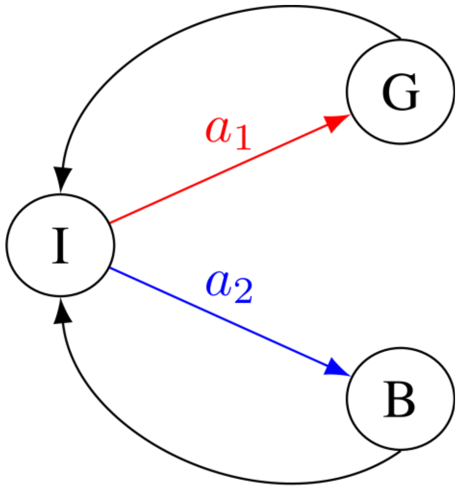
$$\kappa_t(\cdot|x_0, \ldots, x_t, \cdot) \in \mathcal{P}_{x_t} \subset \{A \to \mathcal{P}(S)\}$$

- Note: it turns out that SA-rectangular adversary is equivalent to a special S-rectangular adversary.

S-Rectangularity: Illustrative Example



(a) $p^{(1)}$



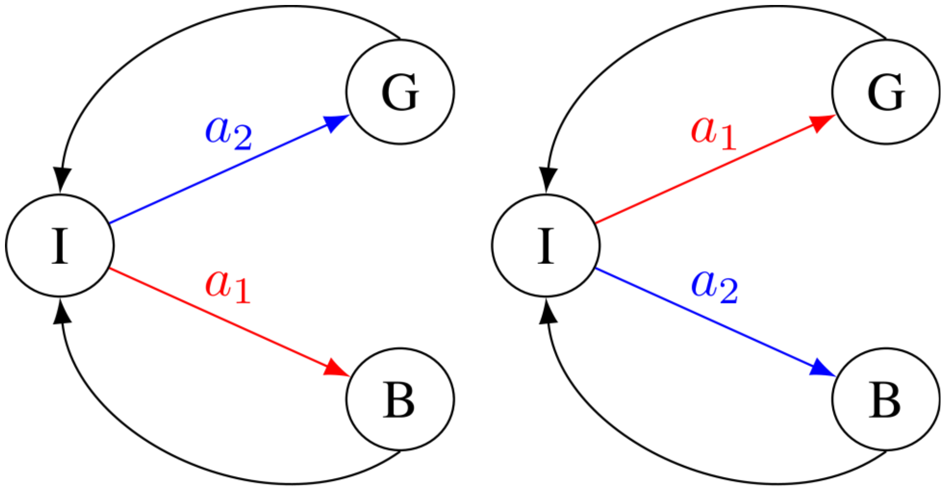
(b) $p^{(2)}$

$$p^{(1)}_{I,a_1}(B) = 1 \text{ and } p^{(1)}_{I,a_2}(G) = 1,$$
$$p^{(2)}_{I,a_1}(G) = 1 \text{ and } p^{(2)}_{I,a_2}(B) = 1.$$

Assume **deterministic** adversary:

- The adversary can choose transition diagram either (a) or (b). Controller choose uniform at random.

S-Rectangularity: Illustrative Example



(a) $p^{(1)}$ (b) $p^{(2)}$

$$p^{(1)}_{I,a_1}(B) = 1 \text{ and } p^{(1)}_{I,a_2}(G) = 1,$$
$$p^{(2)}_{I,a_1}(G) = 1 \text{ and } p^{(2)}_{I,a_2}(B) = 1.$$

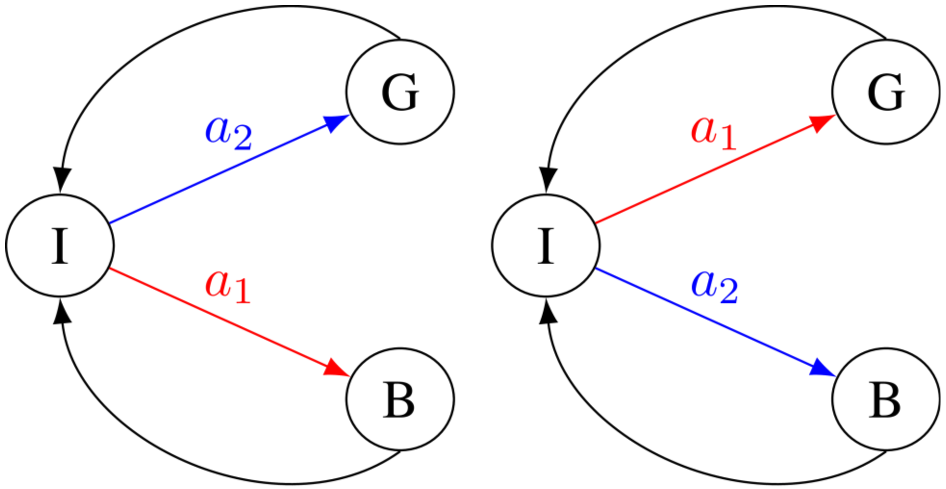
Assume **deterministic** adversary:

- The adversary can choose transition diagram either (a) or (b). Controller choose uniform at random.

Based on the “seeing/not seeing action” intuition:

- S-Rectangular: 50-50.
- SA-Rectangular: Starting from state I, the adversary can make the next state B regardless of the controller’s action.

S-Rectangularity: Illustrative Example



(a) $p^{(1)}$ (b) $p^{(2)}$

$$p^{(1)}_{I,a_1}(B) = 1 \text{ and } p^{(1)}_{I,a_2}(G) = 1,$$
$$p^{(2)}_{I,a_1}(G) = 1 \text{ and } p^{(2)}_{I,a_2}(B) = 1.$$

Assume **deterministic** adversary:

- The SA-rectangular adversary is much more powerful.
- Might lead to conservative policies.
- S-rectangularity further constrains the adversary.

Summary of *Attributes*

- History-dependent controller policy class Π induced by the sets of admissible policies $\mathcal{Q} \subset \mathcal{P}(A)$ is:

$$\Pi := \{\pi = (\pi_t : t \geq 0) : \pi_t(\cdot|x_0, \ldots, x_t) \in \mathcal{Q}\}$$

- History-dependent S-Rectangular adversary policy class K induced by the sets $\{\mathcal{P}_s : s \in S\}$ where $\mathcal{P}_s \subset \{A \to \mathcal{P}(S)\}$ is:

$$\mathrm{K} := \{\kappa = (\kappa_t : t \geq 0) : \kappa_t(\cdot|x_0, \ldots, x_t, \cdot) \in \mathcal{P}_{x_t}\}$$

- Similarly defined for Markov, time-homogeneous players.

DPP

Postulated DPP

$$v^*(s) = \sup_{\pi \in \Pi} \inf_{\kappa \in \mathrm{K}} E_s^{\pi,\kappa}\left[\sum_{k=0}^{\infty} \gamma^k r(X_t, A_t)\right]$$

$$u^*(s) = \sup_{d \in \mathcal{Q}} \inf_{p(\cdot|s,\cdot) \in \mathcal{P}_s} E_{A\sim d}[r(s,A)] + \gamma E_{A\sim d}\left[\sum_{y \in S} p(y|s,A) u^*(y)\right].$$

where $\mathcal{Q}$: controllers policy set, and $\mathcal{P}_s$ S-rectangular adversary.

Π, K are derived from $\mathcal{Q}, \mathcal{P}_s$

Postulated DPP

$$v^*(s) = \sup_{\pi\in\Pi} \inf_{\kappa\in K} E_s^{\pi,\kappa}\left[\sum_{k=0}^{\infty}\gamma^k r(X_t, A_t)\right]$$

$$? \; ||$$

$$u^*(s) = \sup_{d\in\mathcal{Q}} \inf_{p(\cdot|s,\cdot)\in\mathcal{P}_s} E_{A\sim d}[r(s,A)] + \gamma E_{A\sim d}\left[\sum_{y\in S} p(y|s,A)u^*(y)\right].$$

where $\mathcal{Q}$: controllers policy set, and $\mathcal{P}_s$ S-rectangular adversary.
Π, K are derived from $\mathcal{Q}, \mathcal{P}_s$

Whether the DPP (a.k.a Bellman equation) holds for symmetric and asymmetric information (HD, Markov, TH)?

SA-Rectangular

		Adversary History-dependent	Adversary Markov	Adversary Time-homogeneous
Controller	History-dependent	✔ González-Trejo et al. [2002]	✔ Iyengar [2005]	✔
	Markov	✔	✔ Nilim and El Ghaoui [2005]	✔ Nilim and El Ghaoui [2005]
	Time-homogeneous	✔	✔ Iyengar [2005] Nilim and El Ghaoui [2005]	✔ Nilim and El Ghaoui [2005]

S-Rectangular with Convex Ambiguity Sets

Randomized Controller		Convex Adversary	
	History-dependent	Markov	Time-homogeneous
History-dependent	✔	✔ Xu and Mannor [2010]	✔ Wiesemann et al. [2013] Xu and Mannor [2010]
Markov	✔	✔ Li and Shapiro [2023]	✔
Time-homogeneous	✔	✔	✔ Le Tallec [2007]

Not the same table for deterministic (non-convex) controller!

Master Theorem

Theorem 1. *Let u^* solve the following two equations simultaneously*

$$u(s) = \sup_{d \in \mathcal{Q}} \inf_{p(\cdot|s,\cdot) \in \mathcal{P}_s} E_{A\sim d}[r(s,A)] + \gamma E_{A\sim d} \sum_{s' \in S} p(s'|s,A)u(s'),$$

$$u(s) = \inf_{p(\cdot|s,\cdot) \in \mathcal{P}_s} \sup_{d \in \mathcal{Q}} E_{A\sim d}[r(s,A)] + \gamma E_{A\sim d} \sum_{s' \in S} p(s'|s,A)u(s').$$

Then, regardless of the information asymmetry, we have

$$u^*(s) = v^*(s) = \sup_{\pi \in \Pi} \inf_{\kappa \in \mathrm{K}} E_s^{\pi,\kappa} \sum_{k=0}^{\infty} \gamma^k r(X_k, A_k).$$

Proof Sketch

$$u^*(s) = \sup_{d\in\mathcal{Q}} \inf_{p(\cdot|s,\cdot)\in\mathcal{P}_s} E_{A\sim d}[r(s,A)] + \gamma E_{A\sim d}\sum_{s'\in S} p(s'|s,A)u^*(s')$$

$$u^*(s) = \inf_{p(\cdot|s,\cdot)\in\mathcal{P}_s} \sup_{d\in\mathcal{Q}} E_{A\sim d}[r(s,A)] + \gamma E_{A\sim d}\sum_{s'\in S} p(s'|s,A)u^*(s')$$

- Note: fix d, the kernel achieving the inner inf depends on d.
- If interchangeable:
  - Let d_s^* achieve the first line outer sup.
  - Also let $p^*(\cdot|s,\cdot)$ achieve the second line outer inf.
  - Then d_s^* is optimal for $p^*(\cdot|s,\cdot)$ and $p^*(\cdot|s,\cdot)$ is the worst case under d_s^*.

Proof Sketch

$$\sup_{\pi \in \Pi} \inf_{\kappa \in \mathrm{K}} E_s^{\pi,\kappa} \sum_{k=0}^{\infty} \gamma^k r(X_t, A_t)$$

- This implies
  - HD controller v.s. TH adversary v^*_{big}. Let adversary use $p^*(\cdot|s,\cdot)$
    The controller has to response with $\pi_t(a|h_t) = d^*_s(a)$ and hence value $v^*_{big} \le u^*$

  - TH controller v.s. HD adversary v^*_{small}. Fix control $\pi(a|s) = d^*_s(a)$
    Then by "backward induction", it is optimal for the adversary to choose

$$\kappa_t(\cdot|x_0, \ldots, x_t, \cdot) = p^*(\cdot|s,\cdot)$$

resulting in value $v^*_{small} \ge u^*$

Proof Sketch

$$\sup_{\pi \in \Pi} \inf_{\kappa \in K} E_s^{\pi,\kappa} \sum_{k=0}^{\infty} \gamma^k r(X_t, A_t)$$

- This implies
  - HD controller v.s. TH adversary v^*_{big}. Let adversary use $p^*(\cdot|s,\cdot)$
    The controller has to response with $\pi_t(a|h_t) = d^*_s(a)$ and hence value $v^*_{big} \leq u^*$

  - TH controller v.s. HD adversary v^*_{small}. Fix control $\pi(a|s) = d^*_s(a)$
    Then by "backward induction", it is optimal for the adversary to choose
    $$\kappa_t(\cdot|x_0, \ldots, x_t, \cdot) = p^*(\cdot|s, \cdot)$$
    resulting in value $v^*_{small} \geq u^*$
- But $v^*_{big} \geq v^*_{small}$, so $v^*_{big} = v^*_{small} = u^*$
- Extreme cases $v^* = u^*$, DPP holds under all information structures.

Interesting Fact

- Note: The proof also implies that

$$v^*(s) = \sup_{\pi \in \Pi} \inf_{\kappa \in \mathrm{K}} E_s^{\pi,\kappa} \sum_{k=0}^{\infty} \gamma^k r(X_t, A_t)$$

$$= \inf_{\kappa \in \mathrm{K}} \sup_{\pi \in \Pi} E_s^{\pi,\kappa} \sum_{k=0}^{\infty} \gamma^k r(X_t, A_t).$$

Implications: SA-Rectangular

Lemma (SA-rectangularity). *Let u^* be the solution of the Bellman equation and define the q-function*

$$q^*(s,a) = r(s,a) + \gamma \inf_{p_{s,a}\in\mathcal{P}_{s,a}} \sum_{s'\in S} p(s'|s,a)u^*(s').$$

If $\{\delta_a : a \in A\} \subset \mathcal{Q}$, then $u^*(\cdot) = \max_{a\in A} q^*(\cdot, a)$ and it also ***solves the inf-sup equation.***

Implications: SA-Rectangular

Lemma (SA-rectangularity). *Let u^* be the solution of the Bellman equation and define the q-function*

$$q^*(s,a) = r(s,a) + \gamma \inf_{p_{s,a} \in \mathcal{P}_{s,a}} \sum_{s' \in S} p(s'|s,a) u^*(s').$$

If $\{\delta_a : a \in A\} \subset \mathcal{Q}$, then $u^*(\cdot) = \max_{a \in A} q^*(\cdot, a)$ and it also solves the inf-sup equation.

- We can define another fixed point equation:

$$q(s,a) = r(s,a) + \gamma \inf_{p_{s,a} \in \mathcal{P}_{s,a}} \sum_{s'} p(s'|s,a) \max_{b \in A} q(s',b).$$

- q^* is its unique solution if $\{\delta_a : a \in A\} \subset \mathcal{Q}$
- Then $u^*(\cdot) = \max_{a \in A} q^*(\cdot, a)$

Implications: SA-Rectangular

Lemma (SA-rectangularity). *Let u^* be the solution of the Bellman equation and define the q-function*

$$q^*(s,a) = r(s,a) + \gamma \inf_{p_{s,a}\in\mathcal{P}_{s,a}} \sum_{s'\in S} p(s'|s,a)u^*(s').$$

If $\{\delta_a : a \in A\} \subset \mathcal{Q}$, *then* $u^*(\cdot) = \max_{a\in A} q^*(\cdot, a)$ *and it also solves the inf-sup equation.*

- To interchange, use the inf in fixed point equation to find the worst adversary

$$q(s,a) = r(s,a) + \gamma \inf_{p_{s,a}\in\mathcal{P}_{s,a}} \sum_{s'} p(s'|s,a) \max_{b\in A} q(s',b).$$

$$u^*(s) = \inf_{p(\cdot|s,\cdot)\in\mathcal{P}_s} \sup_{d\in\mathcal{Q}} E_{A\sim d}[r(s,A)] + \gamma E_{A\sim d} \sum_{s'\in S} p(s'|s,A)u^*(s')$$

Implications: SA-Rectangular

No need for convexity

Controller	Adversary: History-dependent	Adversary: Markov	Adversary: Time-homogeneous
History-dependent	✔ González-Trejo et al. [2002]	✔ Iyengar [2005]	✔
Markov	✔	✔ Nilim and El Ghaoui [2005]	✔ Nilim and El Ghaoui [2005]
Time-homogeneous	✔	✔ Iyengar [2005] Nilim and El Ghaoui [2005]	✔ Nilim and El Ghaoui [2005]

Same table for deterministic & randomized controller.

Implications: Convex S-Rectangular

Lemma (S-rectangularity with convex ambiguity sets). *With $\mathcal{Q} = \mathcal{P}(A)$ convex and compact, and convex $\mathcal{P}_s$ for all $s \in S$, by the Sion's minimax theorem, we have that the sup and inf in the Master theorem interchanges.*

$$u^*(s) = \sup_{d \in \mathcal{Q}} \inf_{p(\cdot|s,\cdot) \in \mathcal{P}_s} \boxed{E_{A\sim d}[r(s,A)] + \gamma E_{A\sim d} \sum_{s' \in S} p(s'|s,A) u^*(s')}$$

Affine in either *d* or *p*!

Implications: Convex S-Rectangular

Randomized Controller		Convex Adversary: History-dependent	Convex Adversary: Markov	Convex Adversary: Time-homogeneous
	History-dependent	✔	✔ Xu and Mannor [2010]	✔ Wiesemann et al. [2013] Xu and Mannor [2010]
	Markov	✔	✔ Li and Shapiro [2023]	✔
	Time-homogeneous	✔	✔	✔ Le Tallec [2007]

Not the same table for **deterministic (non-convex)** controller!

Non-Convex Ambiguity Sets

S-Rectangularity with Non-Convex Ambiguity Sets

		Non-Convex Adversary: History-dependent	Non-Convex Adversary: Markov	Non-Convex Adversary: Time-homogeneous
Randomized Controller	History-dependent	✔	✔	✗ Wiesemann et al. [2013]
	Markov	✔	✔ Li and Shapiro [2023]	✗
	Time-homogeneous	✔	✔	✔ Le Tallec [2007]

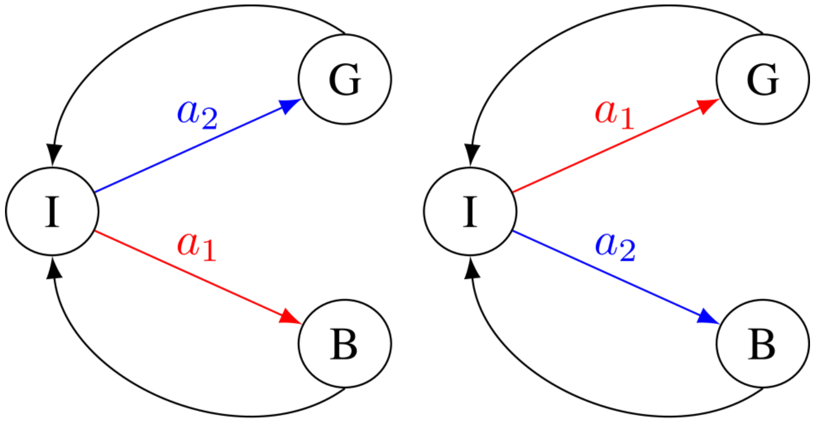
Not the same table for **deterministic (non-convex)** controller (on the next page)!

S-Rectangularity with Non-Convex Ambiguity Sets

		Convex/Non-Convex Adversary		
		History-dependent	Markov	Time-homogeneous
Deterministic Controller	History-dependent	✔	✗	✗
	Markov	✔	✔	✗
	Time-homogeneous	✔	✔	✔

Not the same table for **randomized (covex)** controller (on the previous slide)!

Counterexample 1:
History-dependent convex controller v.s. Non-Convex Time-Homogeneous S-Rectangular Adversary



(a) $p^{(1)}$ (b) $p^{(2)}$

$$p^{(1)}_{I,a_1}(B) = 1 \text{ and } p^{(1)}_{I,a_2}(G) = 1,$$

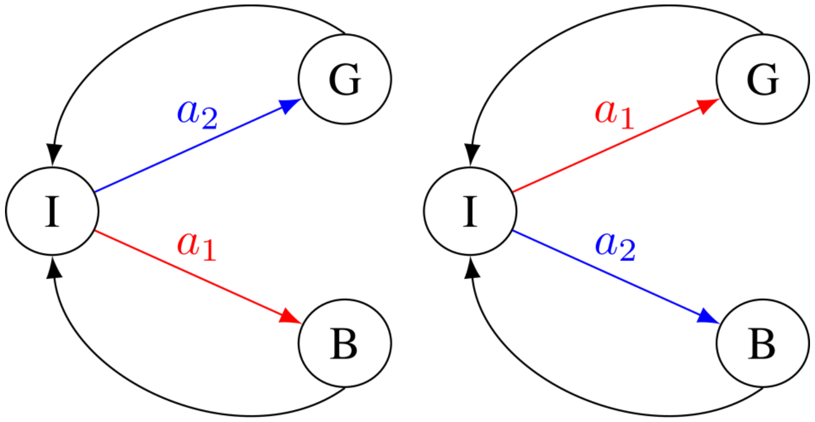
$$p^{(2)}_{I,a_1}(G) = 1 \text{ and } p^{(2)}_{I,a_2}(B) = 1.$$

$$r(I) = 0, r(G) = 1, r(B) = -1.$$

The solution to the DR-DPP:

$$u^*(I) = 0, u^*(G) = 1, u^*(B) = -1.$$

Counterexample 1:
History-dependent convex controller v.s. Non-Convex **Time-Homogeneous** S-Rectangular Adversary



(a) $p^{(1)}$ (b) $p^{(2)}$

$$p^{(1)}_{I,a_1}(B) = 1 \text{ and } p^{(1)}_{I,a_2}(G) = 1,$$
$$p^{(2)}_{I,a_1}(G) = 1 \text{ and } p^{(2)}_{I,a_2}(B) = 1.$$
$$r(I) = 0, r(G) = 1, r(B) = -1.$$

The solution to the DR-DPP:

$$u^*(I) = 0, u^*(G) = 1, u^*(B) = -1.$$

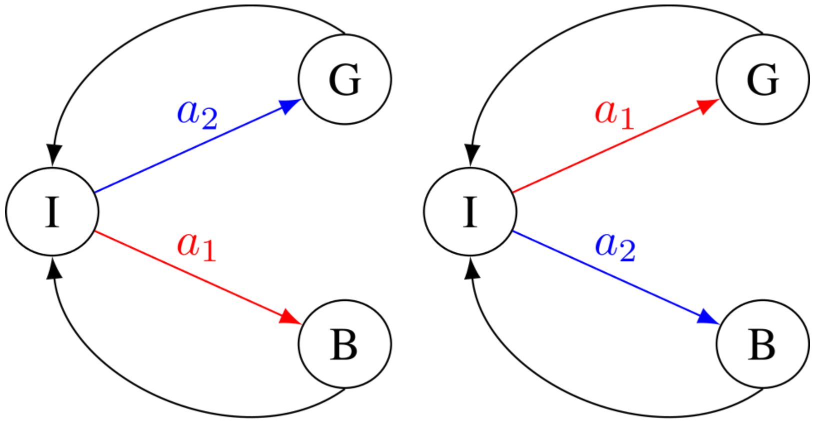
Starting History-dependent policy:

- At time 0, uniformly random an action at state I.
- If jump to G, choose same action for the following time steps.
- If jump to B, choose alternative action for the following time steps.
- For any Markov time-homogeneous adversary κ,
$$v(I, \pi, \kappa) = \gamma^3/(1-\gamma^2) > 0 = u^*(I).$$

Intuition: *Bandit learning* by the controller!

Counterexample 1:

History-dependent convex controller v.s. Non-Convex **Time-Homogeneous** S-Rectangular Adversary



(a) $p^{(1)}$ (b) $p^{(2)}$

$$p^{(1)}_{I,a_1}(B) = 1 \text{ and } p^{(1)}_{I,a_2}(G) = 1,$$

$$p^{(2)}_{I,a_1}(G) = 1 \text{ and } p^{(2)}_{I,a_2}(B) = 1.$$

$$r(I) = 0, r(G) = 1, r(B) = -1.$$

Thoughts:

- It is actually quite remarkable that we do have DPP in asymmetric case where the adversary is TH.

Conclusion

- DRRL is an emerging area that heavily relies on DPP (Bellman equation).
- Attributes such as information constraints and rectangularity can usually be imposed improve realism of the model, without losing tractability in terms of a DPP.
- Despite information *asymmetry* and the absence of convexity, DPP typically holds.
- DPP doesn’t hold in general: especially for the time-homogeneous adversary case.
- Equivalent DR stochastic control formulations exist.

Extension

- DR stochastic control formulation equivalent to DRMDPs.

shengbo.wang@stanford.edu

Extension: State Recursion Formulation

$$X_{t+1} = f(X_t, A_t, K_t) \qquad X_{t+1} = (X_t + A_t - K_t)_+$$

- Many OR related settings, e.g. inventory control, queuing, system engineering, state recursion formulation is convenient.
- Adversary cannot perturb f.
- Adversary can induce shifts in the distribution of K_t.

Extension: State Recursion Formulation

$$X_{t+1} = f(X_t, A_t, K_t) \qquad X_{t+1} = (X_t + A_t - K_t)_+$$

- Many OR related settings, e.g. inventory control, queuing, system engineering, state recursion formulation is convenient.
- Adversary cannot perturb f.
- Adversary can induce shifts in the distribution of K_t.

How do our theories translate?

Extension: State Recursion Formulation

$$X_{t+1} = f(X_t, A_t, K_t)$$

- SA-rectangular: Can choose *different* distribution of K for *different action.*
- S-rectangular: the *same* distributional choice of K are made *across all actions.*

Same intuition

Extension: State Recursion Formulation

$$X_{t+1} = f(X_t, A_t, K_t)$$

- SA-rectangular: Can choose *different* distribution of K for *different action.*
- S-rectangular: the *same* distributional choice of K are made *across all actions.*

- Action-aware.
- Action-agnostic.

Extension: State Recursion Formulation

- Ambiguity set $\mathcal{P} \subset \mathcal{P}(\mathbf{K})$.
- Bellman equation for the Action-aware (SA) case:

$$u^*(s) = \sup_{d \in \mathcal{Q}} E_{A \sim d} \left[r(s, A) + \gamma \inf_{\psi \in \mathcal{P}} E_{K \sim \psi} u^*(f(s, A, K)) \right]$$

- Bellman equation for the Action-agnostic (S) case:

$$u^*(s) = \sup_{d \in \mathcal{Q}} \inf_{\psi \in \mathcal{P}} E_{A \sim d, K \sim \psi} \left[r(s, A) + \gamma u^*(f(s, A, K)) \right]$$

- DPP: equivalent to the previous tables.

Paper: https://arxiv.org/abs/2311.09018
Slides: https://shengbo-wang.github.io/talks/

Thanks for listening!

Questions?

shengbo.wang@stanford.edu