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

Shengbo Wang (Stanford)

Nian Si (Chicago Booth)

Jose Blanchet (Stanford)

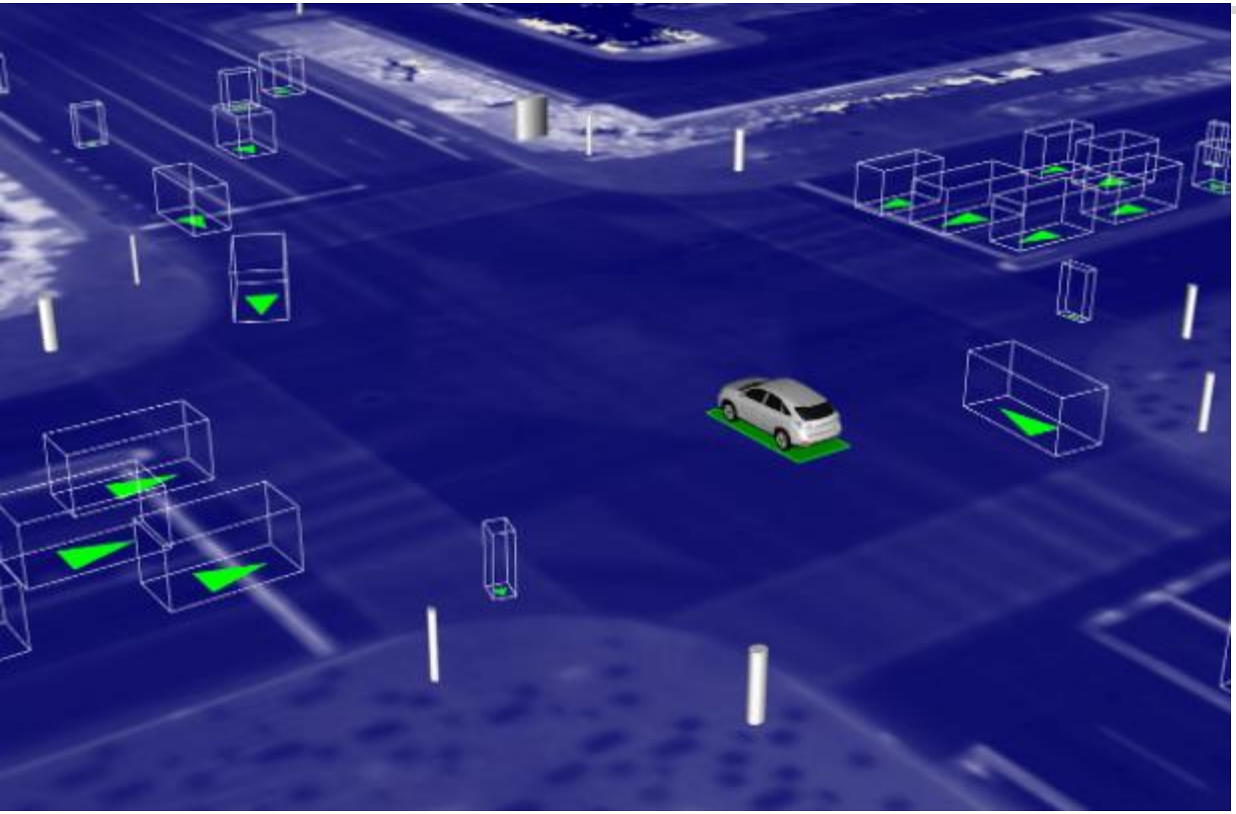
Zhengyuan Zhou (NYU Stern)

shengbo.wang@stanford.edu

Background: Foundation of RL & MDP

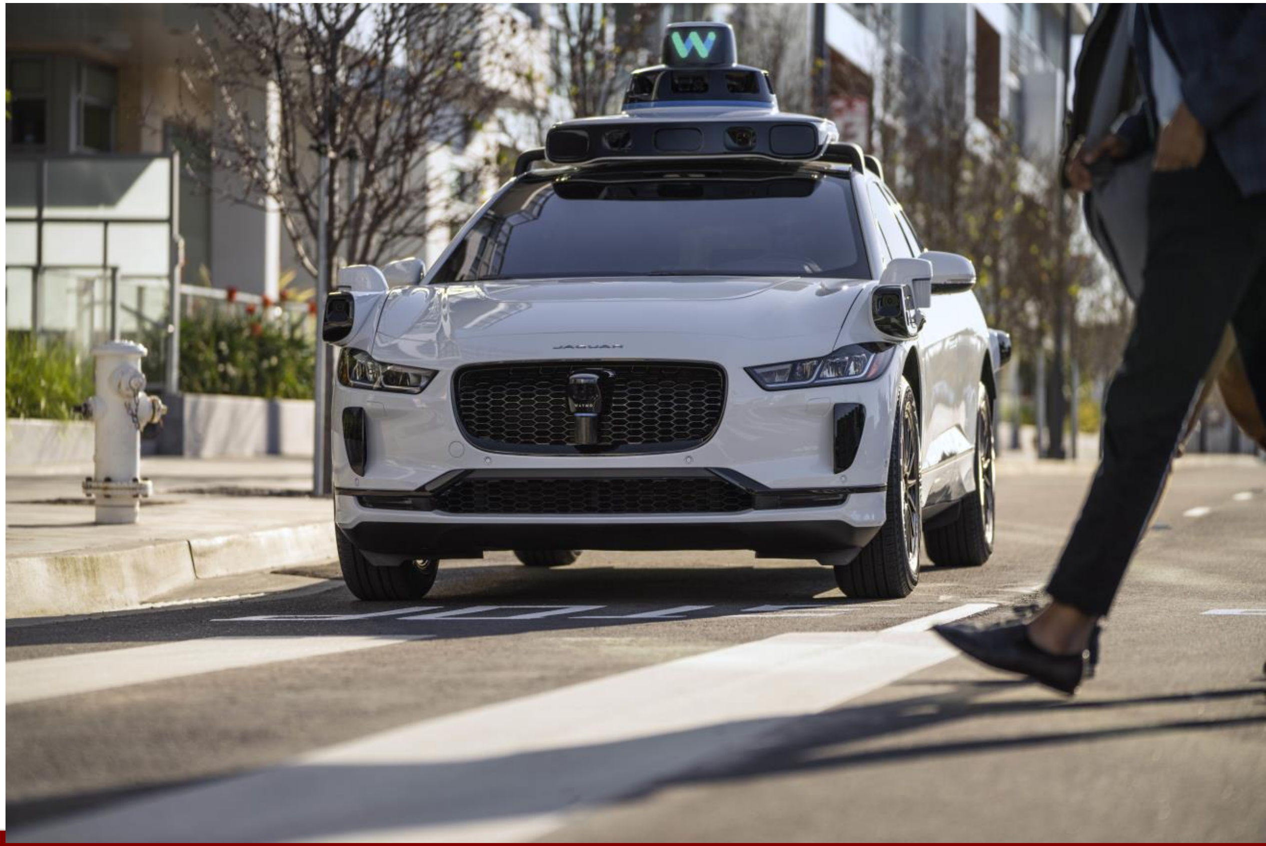
- Markov decision processes (MDPs) are foundational to Reinforcement learning (RL).
- *Dynamic Programming Principle* (DPP) is fundamental both in theory and practice, central to algorithm designs.
- Key consequence of DPP: Markovian policy is optimal.
  - In the infinite horizon discounted case, stationary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

Example Autonomous Driving…



Simulator

Real environment



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.

DRMDP & DRRL

DRMDP: Foundation of DRRL.

Expressiveness:

- Adversarial robustness might lead to over conservative policies.
- Need to restrict the power of the adversary.

Tractability:

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

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

DRMDP: 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

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

DRMDP

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

- 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 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

shengbo.wang@stanford.edu

Attributes of Controller & Adversary

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

- Time-homogeneous v.s. Markov v.s. History-dependent (for both the controller and the adversary).
- Randomized v.s. non-randomized controller.
- Convex or non-convex adversary.
- Does the adversary see the *realized* action of the controller?

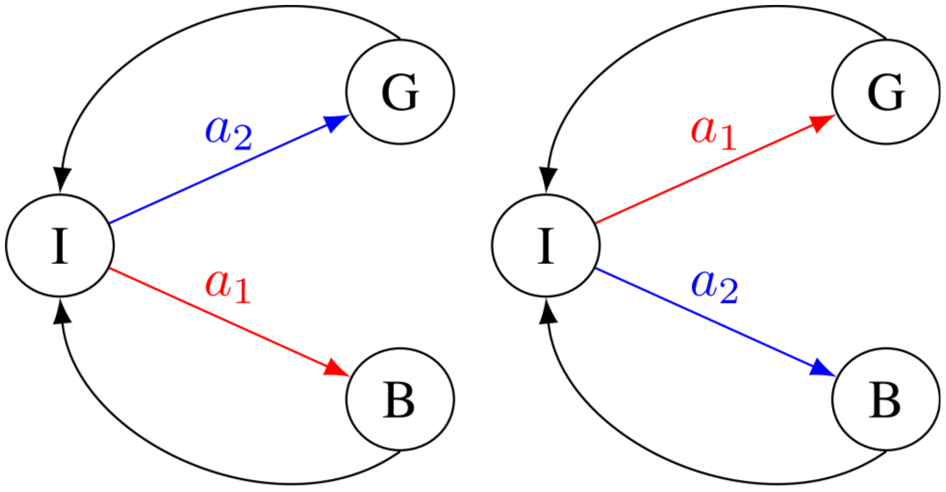
S-Rectangularity: Motivating Example

- Inventory control: $X_{t+1} = (X_t + A_t - K_t)_+$
  - $\{K_t : t \geq 0\}$ i.i.d. is the demand process and A_t is the ordered inventory.
- Natural to assume that at each time t, the adversary can only change K_t dependent on X_t but not dependent on the controller's realized action A_t.
- S-rectangularity.
- SA-rectangularity: Observe X_t, A_t and then choose K_t.

SA- and S-Rectangularity

- SA-rectangularity: Adversary observes both state and action (s,a) and selects an action $p(\cdot|s,a) \in \mathcal{P}_{s,a}$
- S-rectangularity: Adversary observes only state s and selects an action $p(\cdot|s,\cdot) \in \mathcal{P}_s$
- Here $\mathcal{P}_{s,a} \subset \mathcal{P}(S)$, $\mathcal{P}_s \subset \{A \to \mathcal{P}(S)\}$ are prescribed action sets (designed by the modeler).
- In both cases, this selection can be dependent on the history or restricted to be Markov or time-homogeneous.

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:

- One example of S-rectangular: The transition diagram can be either (a) or (b)
- SA-Rectangular: Starting from state I, the adversary can make the next state B regardless of the controller's action.

DPP

Postulated DPP

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

$$? \quad \|$$

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

where $\mathcal{Q}$: controllers policy set; e.g. deterministic $\mathcal{Q} = \{\delta_a : a \in A\}$ or fully randomized $\mathcal{Q} = \mathcal{P}(A)$ policies.

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

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

SA-Rectangularity

		Adversary		
		History-dependent	Markov	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]

Same table for deterministic & randomized controller.

shengbo.wang@stanford.edu

S-Rectangularity with Convex Ambiguity Sets

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 controller!

The Master Theorem

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

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

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

Then, regardless of the information asymmetry, we have

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

Master Theorem: Implications

Lemma 1 (SA-rectangularity). *Let u^* be the solution of the Distributionally robust Bellman equation and define the q-function*

$$q^*(s,a) = r(s,a) + \gamma \inf_{p_{s,a} \in \mathcal{P}_{s,a}} p_{s,a}[u^*].$$

Then, $u^*(\cdot) = \max_{a \in A} q^*(\cdot, a)$ and it also solves the inf-sup equation.

Lemma 2 (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.*

shengbo.wang@stanford.edu

S-Rectangularity with Non-Convex Ambiguity Sets

		Non-Convex Adversary		
		History-dependent	Markov	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** controller!

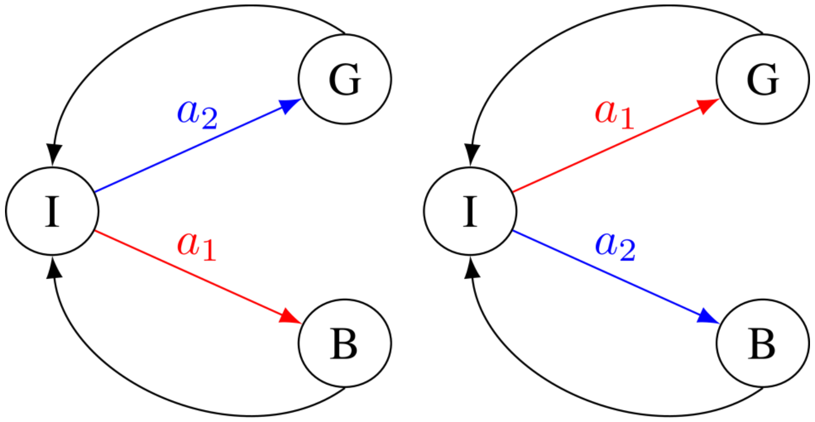
S-Rectangularity with Non-Convex Ambiguity Sets

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

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

Counterexample 1:

History-dependent controller v.s. Non-Convex Time-Homogeneous 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!

Conclusions

- DRRL is an emerging area that heavily relies on DPP (Bellman equation).
- Attributes such as information constraints and rectangularity can usually be imposed to reduce over conservativeness, 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.

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

Thanks for listening!

Questions?

shengbo.wang@stanford.edu