A General Framework for Sample-Efficient Function Approximation in RL

Huizhuo Yuan
CS @ UCLA

Present at RL Theory Seminar
Sep. 26th, 2023

Joint work with Zixiang Chen*, Chris Junchi Li*, Quanquan Gu, Michael I. Jordan

Outline

01 **Background**
- Preliminaries
- Literature of RL models
- Model-based and Model-free RL
- General Function Approximation

02 **Our ABC Framework**
- Scope and sample complexity of our framework
- Admissible bellman characterization
- Functional eluder dimension
- Decomposable estimation function

03 **MDP Instances**
- Linear mixture MDP
- Low witness rank
- Kernelized nonlinear regulator (KNR)

04 **Algorithm and Main Results**
- Algorithm
- Proof sketch
- Main theorem

05 **Implications for specific MDPs**
- Linear mixture MDPs
- Kernelized nonlinear regulator (KNR)
- Low witness rank

Applications of RL

https://www.deepmind.com/research/highlighted-research/alphago https://openai.com/research/solving-rubiks-cube

https://ai.meta.com/blog/rebel-a-general-game-playing-ai-bot-that-excels-at-poker-and-more/ https://waymo.com/

Applications of RL

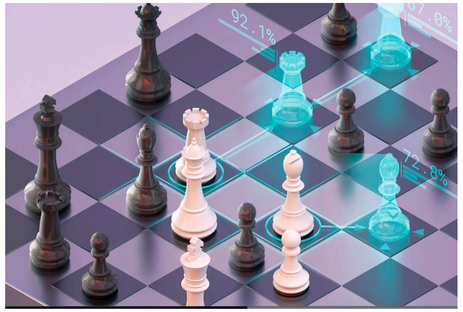
- Reinforcement Learning (RL):
  - Decision making process
  - Empirical successes:
  - Function approximation in RL:

https://www.deepmind.com/research/highlighted-research/alphago https://openai.com/research/solving-rubiks-cube

https://ai.meta.com/blog/rebel-a-general-game-playing-ai-bot-that-excels-at-poker-and-more/ https://waymo.com/

Applications of RL

- Reinforcement Learning (RL):
  - Decision making process
  - Empirical successes:
  - Function approximation in RL:

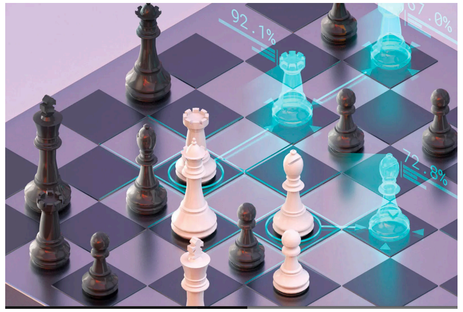


https://www.deepmind.com/research/highlighted-research/alphago https://openai.com/research/solving-rubiks-cube

https://ai.meta.com/blog/rebel-a-general-game-playing-ai-bot-that-excels-at-poker-and-more/ https://waymo.com/

Applications of RL

- Reinforcement Learning (RL):
  - Decision making process
  - Empirical successes:
  - Function approximation in RL:

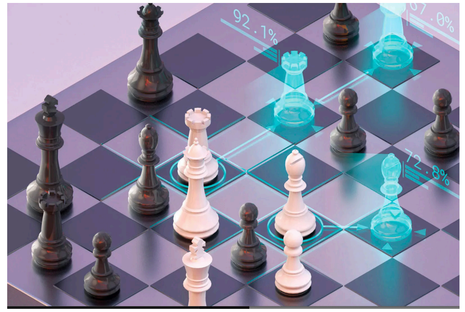


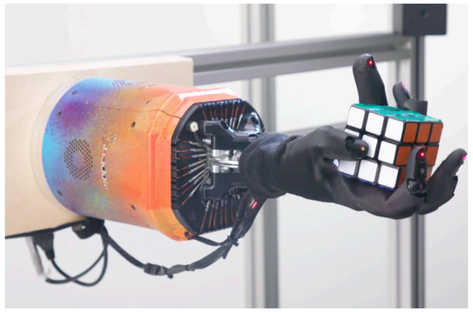
https://www.deepmind.com/research/highlighted-research/alphago https://openai.com/research/solving-rubiks-cube

https://ai.meta.com/blog/rebel-a-general-game-playing-ai-bot-that-excels-at-poker-and-more/ https://waymo.com/

Applications of RL

- Reinforcement Learning (RL):
  - Decision making process
  - Empirical successes:
  - Function approximation in RL:



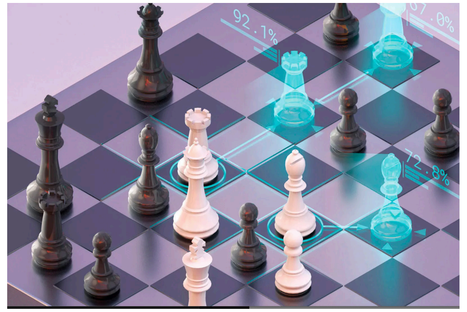


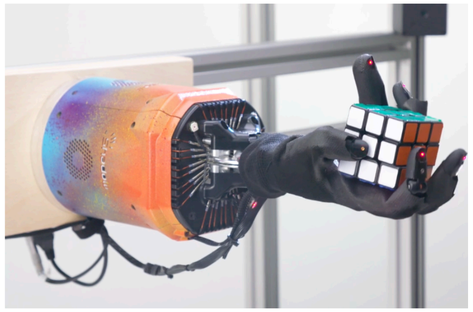
https://www.deepmind.com/research/highlighted-research/alphago https://openai.com/research/solving-rubiks-cube

https://ai.meta.com/blog/rebel-a-general-game-playing-ai-bot-that-excels-at-poker-and-more/ https://waymo.com/

Applications of RL

- Reinforcement Learning (RL):
  - Decision making process
  - Empirical successes:
  - Function approximation in RL:




https://www.deepmind.com/research/highlighted-research/alphago https://openai.com/research/solving-rubiks-cube

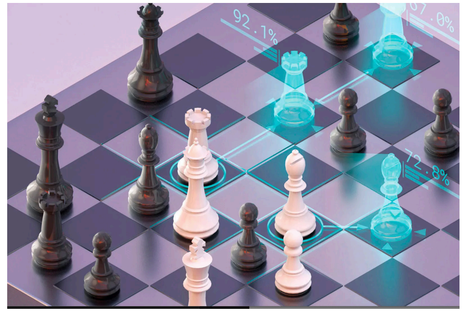
https://ai.meta.com/blog/rebel-a-general-game-playing-ai-bot-that-excels-at-poker-and-more/ https://waymo.com/

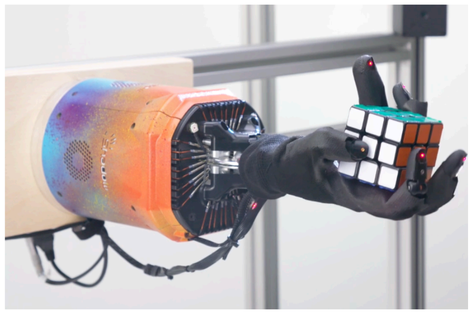
Applications of RL

- Reinforcement Learning (RL):
  - Decision making process
  - Empirical successes:
  - Function approximation in RL:

    Dealing with large state

    and action spaces




https://www.deepmind.com/research/highlighted-research/alphago https://openai.com/research/solving-rubiks-cube

https://ai.meta.com/blog/rebel-a-general-game-playing-ai-bot-that-excels-at-poker-and-more/ https://waymo.com/

Background

Markov Decision Processes (MDP)

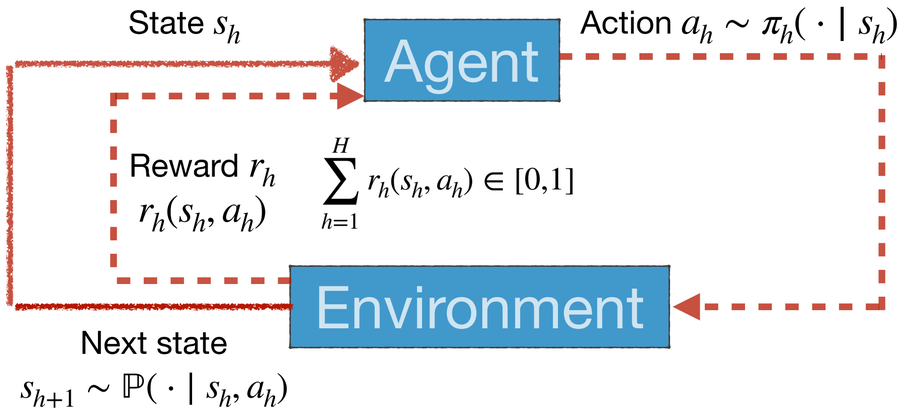
$\mathcal{S}$: space of feasible states

$\mathcal{A}$: action space

H: the horizon in each episode

$\mathbb{P} := \{\mathbb{P}_h\}_{h\in[H]}$: transition probability

$r_h(s,a) \geq 0$: the reward received



Background

Markov Decision Processes (MDP)

- Markov decision process (MDP): $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$

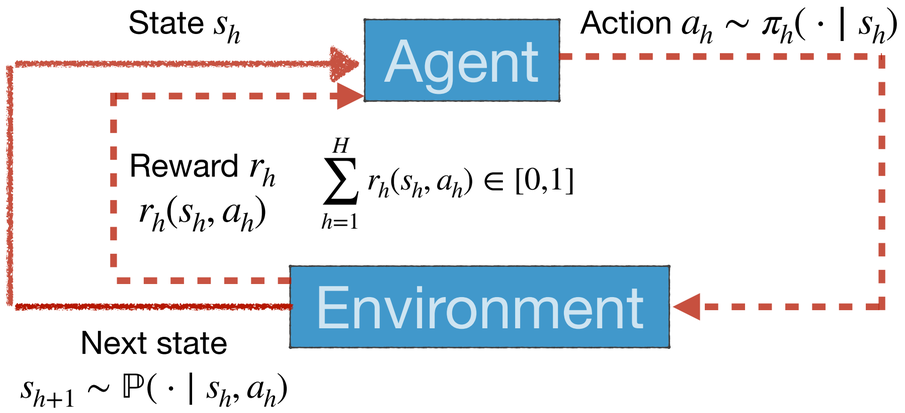
$\mathcal{S}$: space of feasible states

$\mathcal{A}$: action space

H: the horizon in each episode

$\mathbb{P} := \{\mathbb{P}_h\}_{h\in[H]}$: transition probability

$r_h(s,a) \geq 0$: the reward received

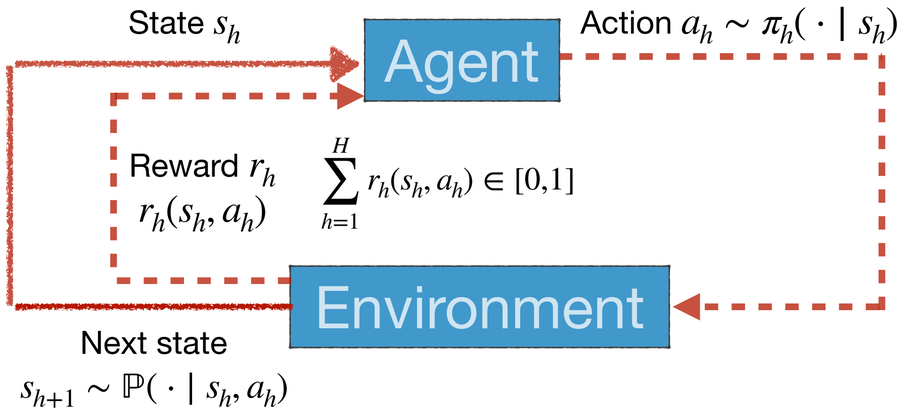


Background

Markov Decision Processes (MDP)

- Markov decision process (MDP): $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$

$\mathcal{S}$: space of feasible states
$\mathcal{A}$: action space
H: the horizon in each episode
$\mathbb{P} := \{\mathbb{P}_h\}_{h\in[H]}$: transition probability
$r_h(s,a) \geq 0$: the reward received



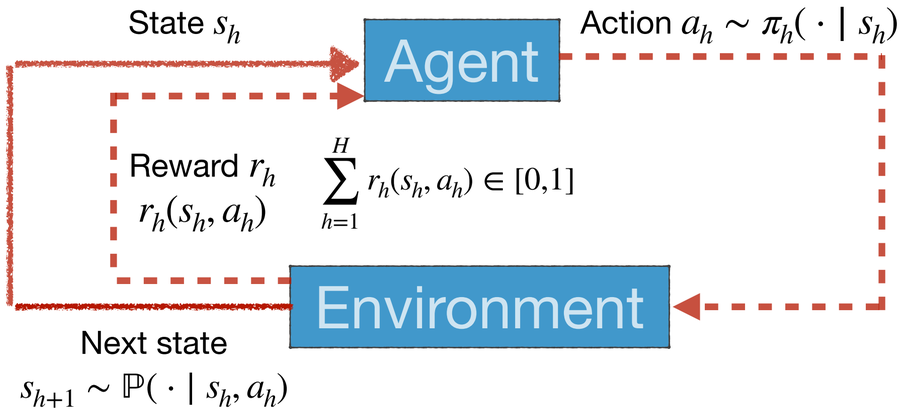
- Deterministic policy: $\left\{\pi_h : \mathcal{S} \mapsto \mathcal{A}\right\}_{h\in[H]}$

Background

Markov Decision Processes (MDP)

- Markov decision process (MDP): $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$

$\mathcal{S}$: space of feasible states
$\mathcal{A}$: action space
H: the horizon in each episode
$\mathbb{P} := \{\mathbb{P}_h\}_{h\in[H]}$: transition probability
$r_h(s,a) \geq 0$: the reward received



- Deterministic policy: $\left\{\pi_h : \mathcal{S} \mapsto \mathcal{A}\right\}_{h\in[H]}$

- $Q_h^\pi(s,a) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s, a_h = a \right]$, $V_h^\pi(s) = Q_h^\pi(s, \pi(s))$

Background

Bellman Operator, Policy and Regret

Background

Bellman Operator, Policy and Regret

- Bellman Operator: $\left(\mathcal{T}_h Q_{h+1}\right)(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q_{h+1}(s', a')$

Background

Bellman Operator, Policy and Regret

- Bellman Operator: $\left(\mathcal{T}_h Q_{h+1}\right)(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q_{h+1}(s', a')$
- Optimal policy: $\pi^* := \arg\max_{\pi} V_h^{\pi}(s), \forall s \in \mathcal{S}$

Background

Bellman Operator, Policy and Regret

- Bellman Operator: $\left(\mathcal{T}_h Q_{h+1}\right)(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q_{h+1}(s', a')$
- Optimal policy: $\pi^* := \arg\max_{\pi} V_h^{\pi}(s), \forall s \in \mathcal{S}$

$$V_h^* := V_h^{\pi^*}, Q_h^* := Q_h^{\pi^*}$$

Background

Bellman Operator, Policy and Regret

- Bellman Operator: $\left(\mathcal{T}_h Q_{h+1}\right)(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q_{h+1}(s', a')$
- Optimal policy: $\pi^* := \arg\max_{\pi} V_h^{\pi}(s), \forall s \in \mathcal{S}$

$$V_h^* := V_h^{\pi^*}, Q_h^* := Q_h^{\pi^*}$$

- Objective: ε-optimal policy: $V_1^{\pi}(s_1) \geq V_1^*(s_1) - \epsilon$

Background

Bellman Operator, Policy and Regret

- Bellman Operator: $\left(\mathcal{T}_h Q_{h+1}\right)(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q_{h+1}(s', a')$
- Optimal policy: $\pi^* := \arg\max_{\pi} V_h^{\pi}(s), \forall s \in \mathcal{S}$

$$V_h^* := V_h^{\pi^*}, Q_h^* := Q_h^{\pi^*}$$

- Objective: ε-optimal policy: $V_1^{\pi}(s_1) \geq V_1^*(s_1) - \epsilon$
- Cumulative regret: $\text{Regret}(T) := \sum_{t=1}^{T} \left[V_1^*(s_1) - V_1^{\pi^t}(s_1) \right]$

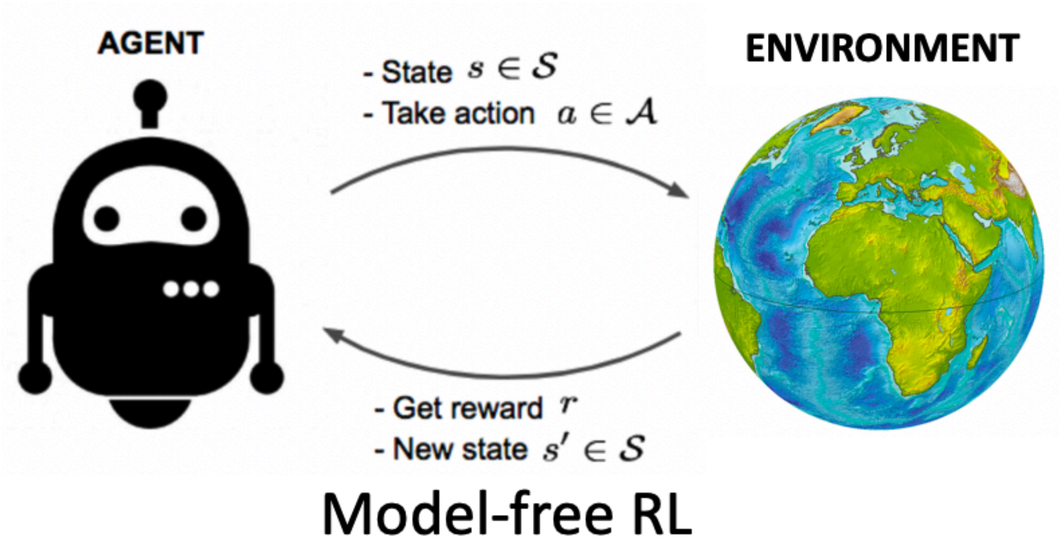
Rich Literature of RL Models

- Tabular MDP
- Factored MDP [Kearns & Koller '99]
- Predictive State Rrepresentations [Littman et al. '01]
- State Aggregation [Li '09, Dong et al. '20]
- Block MDP [Jiang et al. '17]
- Linear Quadratic Regulator (LQR) [Dean et al. '19]
- Linear MDP [Jin et al. '20]
- Linear Mixture MDP [Modi et al. '20, Ayoub et al. '20, Zhou et al.'20]
- Bellman Complete [Munos '05, Zanette et al '20]
- Kernelized Nonlinear Regulator [Kakade et al. '20]
- Low Occupancy Complexity [Du et al. '21]
- Linear Q*/V* [Du et al. '21]
- Kernel Reactive POMDPs [Krishnamurthy et al. '16]
- Flambe/Feature Selection [Agarwal et al. '20]
- Generalized Linear Bellman Complete [Wang et al. '19]

Model-based vs Model-free RL

Model-based vs Model-free RL

Diagram of model-free reinforcement learning



Model-based vs Model-free RL

Diagram of model-free reinforcement learning

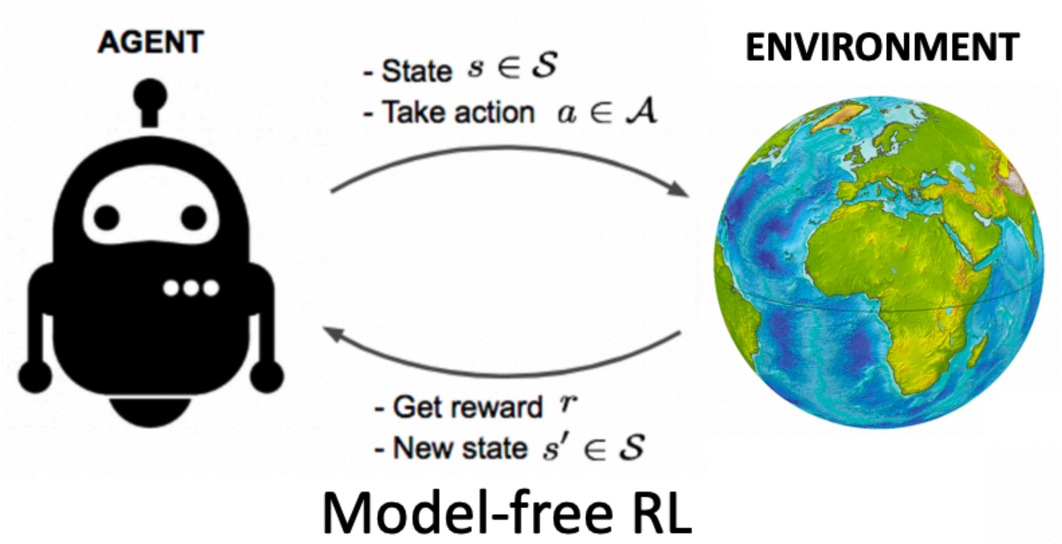
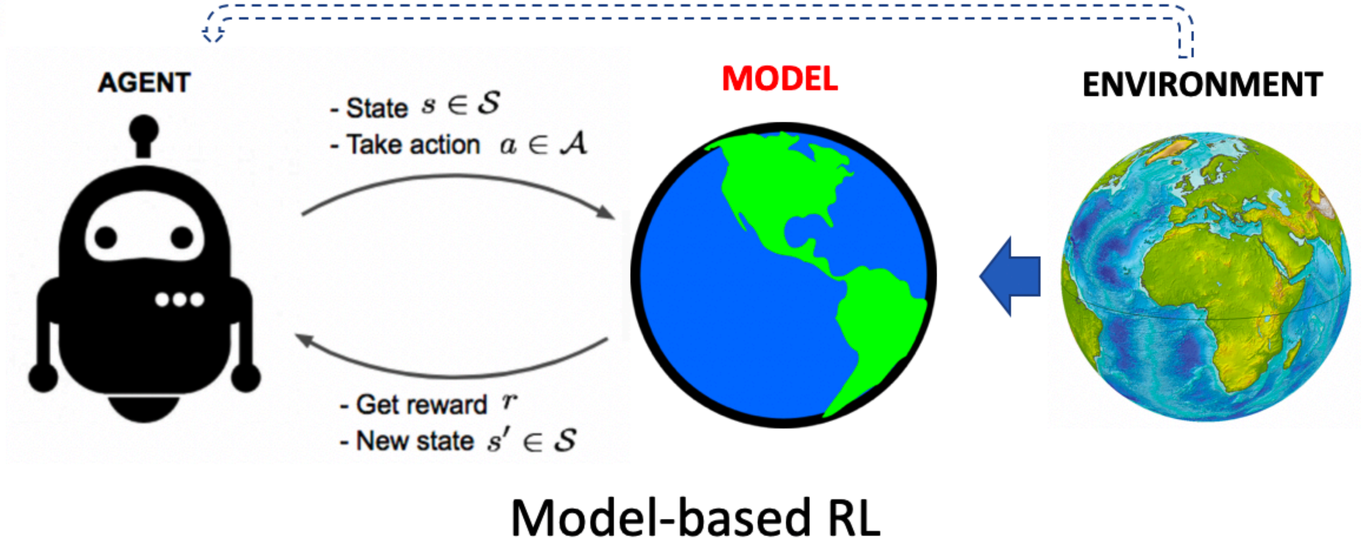


Diagram of model-based reinforcement learning



General Function Approximation

General Function Approximation

- Bellman Rank [Jiang et al. ’17]
- Eluder Dimension [Wang et al. ’20]
- Bellman-Eluder Dimension [Jin et al. ’21]

General Function Approximation

- Bellman Rank [Jiang et al. '17]
- Eluder Dimension [Wang et al. '20]
- Bellman-Eluder Dimension [Jin et al. '21]

$$\mathbb{E}_{\pi_{f'}}\left[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)\right]$$

General Function Approximation

- Bellman Rank [Jiang et al. ’17]
- Eluder Dimension [Wang et al. ’20]
- Bellman-Eluder Dimension [Jin et al. ’21]

$$\mathbb{E}_{\pi_{f'}}\left[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)\right]$$

Model-free

General Function Approximation

- Bellman Rank [Jiang et al. ’17]
- Eluder Dimension [Wang et al. ’20]
- Bellman-Eluder Dimension [Jin et al. ’21]

$$\mathbb{E}_{\pi_{f'}}\left[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)\right]$$

Model-free

- Witness Rank [Sun et al. ’19]
- Bilinear Classes [Du et al. ’21]

General Function Approximation

- Bellman Rank [Jiang et al. ’17]
- Eluder Dimension [Wang et al. ’20]
- Bellman-Eluder Dimension [Jin et al. ’21]

$$\mathbb{E}_{\pi_{f'}}\left[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)\right]$$

Model-free

- Witness Rank [Sun et al. ’19]

Model-based

- Bilinear Classes [Du et al. ’21]

General Function Approximation

- Bellman Rank [Jiang et al. ’17]
- Eluder Dimension [Wang et al. ’20]
- Bellman-Eluder Dimension [Jin et al. ’21]

$$\mathbb{E}_{\pi_{f'}}\left[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)\right]$$

Model-free

- Witness Rank [Sun et al. ’19]

Model-based

- Bilinear Classes [Du et al. ’21]

Covers both model-based and model-free

General Function Approximation

- Bellman Rank [Jiang et al. ’17]
- Eluder Dimension [Wang et al. ’20]
- Bellman-Eluder Dimension [Jin et al. ’21]

$$\mathbb{E}_{\pi_{f'}}\left[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)\right]$$

Model-free

- Witness Rank [Sun et al. ’19]

Model-based

- Bilinear Classes [Du et al. ’21]

Covers both model-based and model-free

- DEC [Foster et al. ’21]:

General Function Approximation

- Bellman Rank [Jiang et al. ’17]
- Eluder Dimension [Wang et al. ’20]
- Bellman-Eluder Dimension [Jin et al. ’21]

$$\mathbb{E}_{\pi_{f'}}\left[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)\right]$$

Model-free

- Witness Rank [Sun et al. ’19]

Model-based

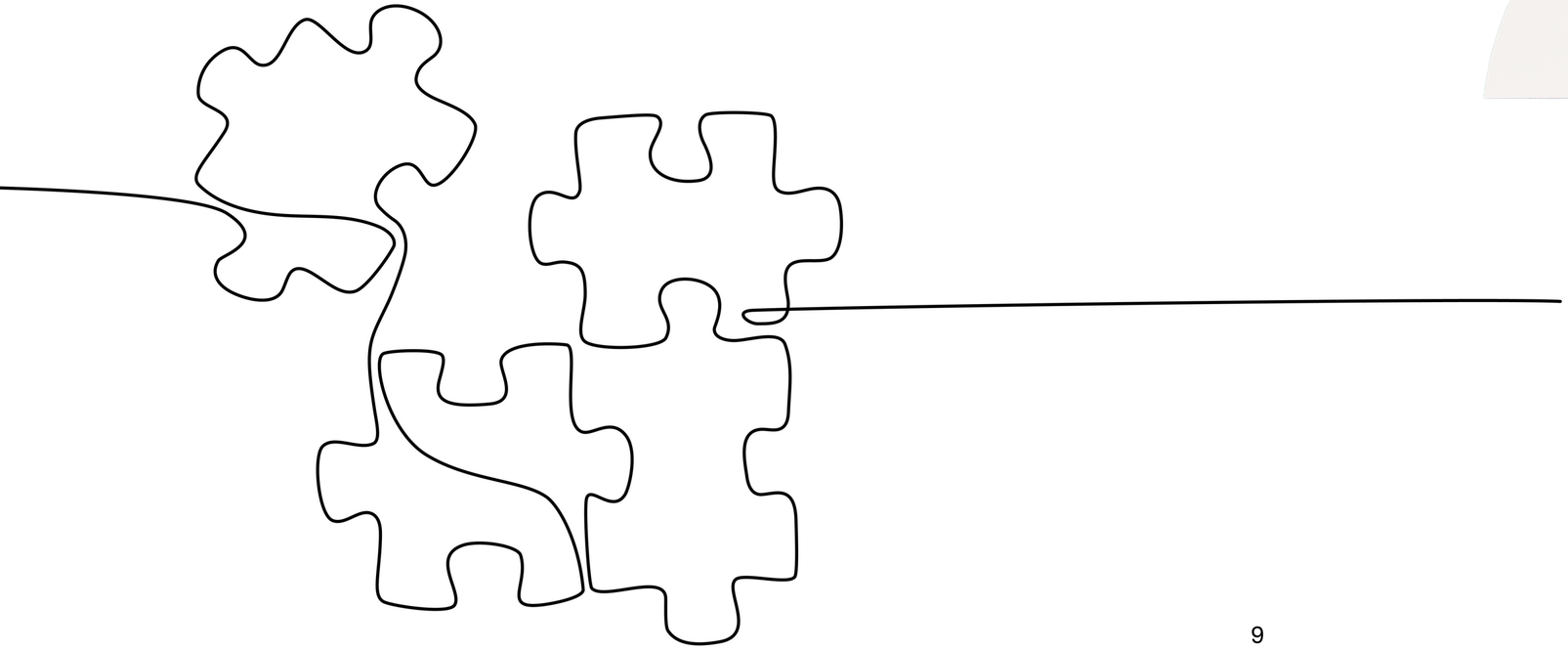
- Bilinear Classes [Du et al. ’21]

Covers both model-based and model-free

- DEC [Foster et al. ’21]:

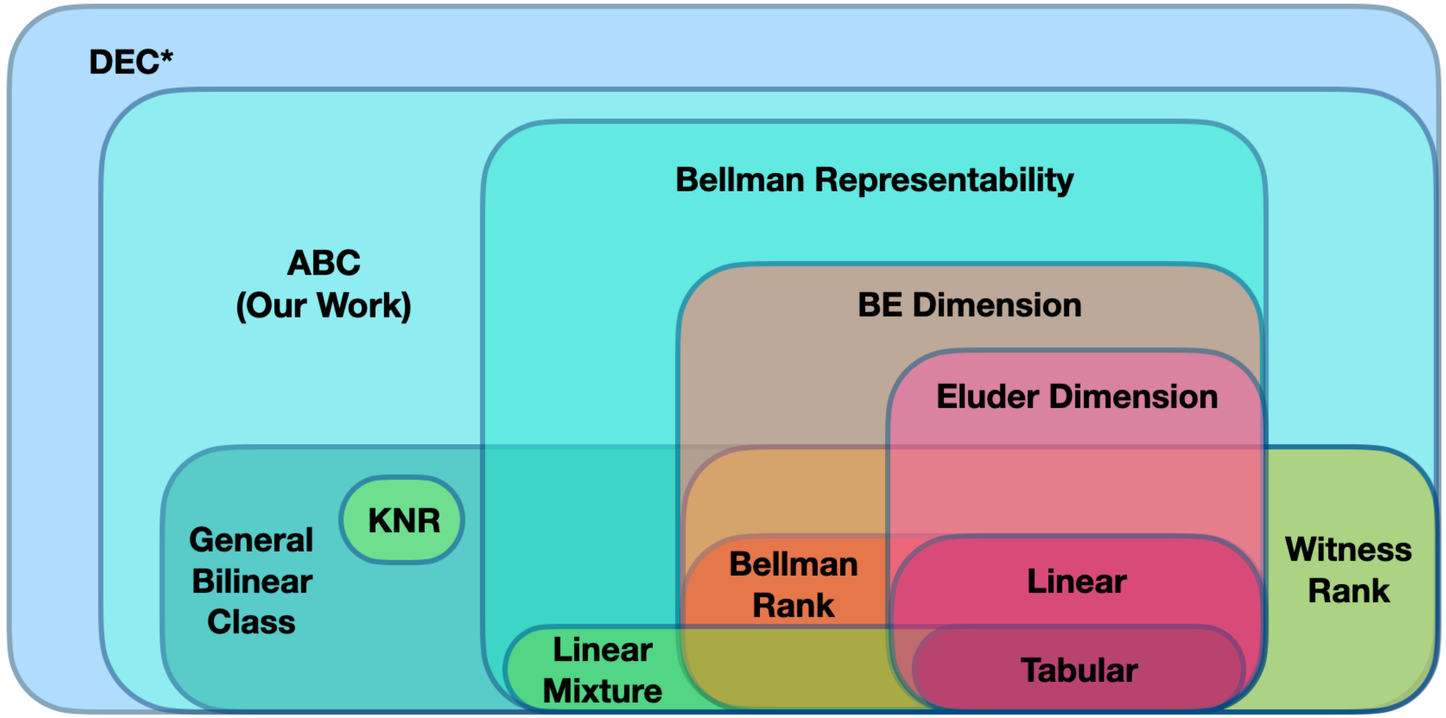
A necessary condition

Is there a unified framework that includes all model-free and model-based RL classes while maintaining sharp sample efficiency



Our ABC Framework

Is there a unified framework that includes all model-free and model-based RL classes while maintaining sharp sample efficiency?



Venn-Diagram Visualization of Prevailing Sample-Efficient RL Classes

Comparison of Sample Complexity

Comparison of sample complexity for different MDP models under different RL frameworks

Comparison of Sample Complexity

	Bilinear Class	Low BE Dimensio	DEC and Bellman Representability	ABC Class (with Low FE
Linear MDPs [Yang & Wang ’19; Jin et				
Linear Mixture MDPs				
Bellman Rank [Jiang et al. ’17]				
Eluder Dimension [Wang et al. ’20]				
Witness Rank [Sun et al. ’19]				
Low Occupancy Complexity				
Kernelized Nonlinear Regulator [Kakade et al. ’20]				
Linear Q*/V* [Du et al. ’21]				

Comparison of sample complexity for different MDP models under different RL frameworks

Comparison of Sample Complexity

	Bilinear Class	Low BE Dimensio	DEC and Bellman Representability	ABC Class (with Low FE
Linear MDPs [Yang & Wang '19; Jin et	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Linear Mixture MDPs				
Bellman Rank [Jiang et al. '17]				
Eluder Dimension [Wang et al. '20]				
Witness Rank [Sun et al. '19]				
Low Occupancy Complexity	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Kernelized Nonlinear Regulator [Kakade et al. '20]				
Linear Q*/V* [Du et al. '21]	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2

Comparison of sample complexity for different MDP models under different RL frameworks

Comparison of Sample Complexity

	Bilinear Class	Low BE Dimensio	DEC and Bellman Representability	ABC Class (with Low FE
Linear MDPs [Yang & Wang '19; Jin et	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Linear Mixture MDPs	d^3H^4/ε^2	✘	d^3H^3/ε^2	d^2H^2/ε^2
Bellman Rank [Jiang et al. '17]				
Eluder Dimension [Wang et al. '20]				
Witness Rank [Sun et al. '19]				
Low Occupancy Complexity	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Kernelized Nonlinear Regulator [Kakade et al. '20]				
Linear Q*/V* [Du et al. '21]	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2

Comparison of sample complexity for different MDP models under different RL frameworks

Comparison of Sample Complexity

	Bilinear Class	Low BE Dimensio	DEC and Bellman Representability	ABC Class (with Low FE
Linear MDPs [Yang & Wang '19; Jin et	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Linear Mixture MDPs	d^3H^4/ε^2	✘	d^3H^3/ε^2	d^2H^2/ε^2
Bellman Rank [Jiang et al. '17]	$d^2H^5\|\mathcal{A}\|/\varepsilon^2$	$dH^2\|\mathcal{A}\|/\varepsilon^2$	$d^2H^3\|\mathcal{A}\|/\varepsilon^2$	$dH^2\|\mathcal{A}\|/\varepsilon^2$
Eluder Dimension [Wang et al. '20]				
Witness Rank [Sun et al. '19]				
Low Occupancy Complexity	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Kernelized Nonlinear Regulator [Kakade et al. '20]				
Linear Q*/V* [Du et al. '21]	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2

Comparison of sample complexity for different MDP models under different RL frameworks

Comparison of Sample Complexity

	Bilinear Class	Low BE Dimensio	DEC and Bellman Representability	ABC Class (with Low FE
Linear MDPs [Yang & Wang '19; Jin et	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Linear Mixture MDPs	d^3H^4/ε^2	✘	d^3H^3/ε^2	d^2H^2/ε^2
Bellman Rank [Jiang et al. '17]	$d^2H^5\lvert\mathcal{A}\rvert/\varepsilon^2$	$dH^2\lvert\mathcal{A}\rvert/\varepsilon^2$	$d^2H^3\lvert\mathcal{A}\rvert/\varepsilon^2$	$dH^2\lvert\mathcal{A}\rvert/\varepsilon^2$
Eluder Dimension [Wang et al. '20]	✘	$\dim_E H^2/\varepsilon^2$	$\dim_E^2 H^3/\varepsilon^2$	$\dim_E H^2/\varepsilon^2$
Witness Rank [Sun et al. '19]				
Low Occupancy Complexity	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Kernelized Nonlinear Regulator [Kakade et al. '20]				
Linear Q*/V* [Du et al. '21]	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2

Comparison of sample complexity for different MDP models under different RL frameworks

Comparison of Sample Complexity

	Bilinear Class	Low BE Dimensio	DEC and Bellman Representability	ABC Class (with Low FE
Linear MDPs [Yang & Wang '19; Jin et	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Linear Mixture MDPs	d^3H^4/ε^2	✘	d^3H^3/ε^2	d^2H^2/ε^2
Bellman Rank [Jiang et al. '17]	$d^2H^5\|\mathcal{A}\|/\varepsilon^2$	$dH^2\|\mathcal{A}\|/\varepsilon^2$	$d^2H^3\|\mathcal{A}\|/\varepsilon^2$	$dH^2\|\mathcal{A}\|/\varepsilon^2$
Eluder Dimension [Wang et al. '20]	✘	$\dim_E H^2/\varepsilon^2$	$\dim_E^2 H^3/\varepsilon^2$	$\dim_E H^2/\varepsilon^2$
Witness Rank [Sun et al. '19]				$W_\kappa H^2\|\mathcal{A}\|/\varepsilon^2$
Low Occupancy Complexity	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Kernelized Nonlinear Regulator [Kakade et al. '20]				
Linear Q*/V* [Du et al. '21]	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2

Comparison of sample complexity for different MDP models under different RL frameworks

Comparison of Sample Complexity

	Bilinear Class	Low BE Dimensio	DEC and Bellman Representability	ABC Class (with Low FE
Linear MDPs [Yang & Wang '19; Jin et	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Linear Mixture MDPs	d^3H^4/ε^2	✘	d^3H^3/ε^2	d^2H^2/ε^2
Bellman Rank [Jiang et al. '17]	$d^2H^5\|\mathcal{A}\|/\varepsilon^2$	$dH^2\|\mathcal{A}\|/\varepsilon^2$	$d^2H^3\|\mathcal{A}\|/\varepsilon^2$	$dH^2\|\mathcal{A}\|/\varepsilon^2$
Eluder Dimension [Wang et al. '20]	✘	$\dim_E H^2/\varepsilon^2$	$\dim_E^2 H^3/\varepsilon^2$	$\dim_E H^2/\varepsilon^2$
Witness Rank [Sun et al. '19]				$W_\kappa H^2\|\mathcal{A}\|/\varepsilon^2$
Low Occupancy Complexity	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Kernelized Nonlinear Regulator [Kakade et al. '20]				$d_\phi^2 d_s H^4/\varepsilon^2$
Linear Q*/V* [Du et al. '21]	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2

Comparison of sample complexity for different MDP models under different RL frameworks

Comparison of Sample Complexity

	Bilinear Class	Low BE Dimensio	DEC and Bellman Representability	ABC Class (with Low FE
Linear MDPs [Yang & Wang '19; Jin et	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Linear Mixture MDPs	d^3H^4/ε^2	✘	d^3H^3/ε^2	d^2H^2/ε^2
Bellman Rank [Jiang et al. '17]	$d^2H^5\|\mathcal{A}\|/\varepsilon^2$	$dH^2\|\mathcal{A}\|/\varepsilon^2$	$d^2H^3\|\mathcal{A}\|/\varepsilon^2$	$dH^2\|\mathcal{A}\|/\varepsilon^2$
Eluder Dimension [Wang et al. '20]	✘	$\dim_E H^2/\varepsilon^2$	$\dim_E^2 H^3/\varepsilon^2$	$\dim_E H^2/\varepsilon^2$
Witness Rank [Sun et al. '19]	—	✘	—	$W_\kappa H^2\|\mathcal{A}\|/\varepsilon^2$
Low Occupancy Complexity	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2
Kernelized Nonlinear Regulator [Kakade et al. '20]	—	✘	—	$d_\phi^2 d_s H^4/\varepsilon^2$
Linear Q*/V* [Du et al. '21]	d^3H^4/ε^2	d^2H^2/ε^2	d^3H^3/ε^2	d^2H^2/ε^2

Comparison of sample complexity for different MDP models under different RL frameworks

Outline

01 **Background**
- Preliminaries
- Literature of RL models
- Model-based and Model-free RL
- General Function Approximation

02 **Our ABC Framework**
- Scope and sample complexity of our framework
- Admissible bellman characterization
- Functional eluder dimension
- Decomposable estimation function

03 **MDP Instances**
- Linear mixture MDP
- Low witness rank
- Kernelized nonlinear regulator (KNR)

04 **Algorithm and Main Results**
- Algorithm
- Proof sketch
- Main theorem

05 **Implications for specific MDPs**
- Linear mixture MDPs
- Kernelized nonlinear regulator (KNR)
- Low witness rank

Admissible Bellman Characterization

Admissible Bellman Characterization

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$

Admissible Bellman Characterization

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$
- Hypothesis classes $\mathcal{F}$ (and $\mathcal{G}$) [Du et al. '21]: identified by a pair $\{Q_f, V_f\}_{f \in \mathcal{F}}$

Admissible Bellman Characterization

Model-free

$\{Q_f\}, V_{h,f}(s) = \max_a Q_{h,f}(s,a)$

Model-based

$Q_{h,f} = Q^*_{h,M_f}, V_{h,f} = V^*_{h,M_f}$

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$
- Hypothesis classes $\mathcal{F}$ (and $\mathcal{G}$) [Du et al. '21]: identified by a pair $\{Q_f, V_f\}_{f \in \mathcal{F}}$

Admissible Bellman Characterization

Model-free

$\{Q_f\}, V_{h,f}(s) = \max_a Q_{h,f}(s,a)$

Model-based

$Q_{h,f} = Q^*_{h,M_f}, V_{h,f} = V^*_{h,M_f}$

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$
- Hypothesis classes $\mathcal{F}$ (and $\mathcal{G}$) [Du et al. '21]: identified by a pair $\{Q_f, V_f\}_{f \in \mathcal{F}}$
- Per step observed transition: $o_h := (s_h, a_h, s_{h+1})$

Admissible Bellman Characterization

Model-free

$\{Q_f\}, V_{h,f}(s) = \max_a Q_{h,f}(s,a)$

Model-based

$Q_{h,f} = Q^*_{h,M_f}, V_{h,f} = V^*_{h,M_f}$

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$
- Hypothesis classes $\mathcal{F}$ (and $\mathcal{G}$) [Du et al. '21]: identified by a pair $\{Q_f, V_f\}_{f \in \mathcal{F}}$
- Per step observed transition: $o_h := (s_h, a_h, s_{h+1})$
- Discriminator function class: $\mathcal{V}$, as in integral probability metrics (IPM) and witness rank

Admissible Bellman Characterization

Model-free

$\{Q_f\}, V_{h,f}(s) = \max_a Q_{h,f}(s,a)$

Model-based

$Q_{h,f} = Q^*_{h,M_f}, V_{h,f} = V^*_{h,M_f}$

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$
- Hypothesis classes $\mathcal{F}$ (and $\mathcal{G}$) [Du et al. '21]: identified by a pair $\{Q_f, V_f\}_{f \in \mathcal{F}}$
- Per step observed transition: $o_h := (s_h, a_h, s_{h+1})$
- Discriminator function class: $\mathcal{V}$, as in integral probability metrics (IPM) and witness rank
- Coupling function: $G_{h,f^*}(f,g) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^d$

Admissible Bellman Characterization

Model-free

$\{Q_f\}, V_{h,f}(s) = \max_a Q_{h,f}(s,a)$

Model-based

$Q_{h,f} = Q^*_{h,M_f}, V_{h,f} = V^*_{h,M_f}$

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$
- Hypothesis classes $\mathcal{F}$ (and $\mathcal{G}$) [Du et al. '21]: identified by a pair $\{Q_f, V_f\}_{f \in \mathcal{F}}$
- Per step observed transition: $o_h := (s_h, a_h, s_{h+1})$
- Discriminator function class: $\mathcal{V}$, as in integral probability metrics (IPM) and witness rank
- Coupling function: $G_{h,f^*}(f,g) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^d$

f^*: true model

Admissible Bellman Characterization

Model-free

$\{Q_f\}, V_{h,f}(s) = \max_a Q_{h,f}(s,a)$

Model-based

$Q_{h,f} = Q^*_{h,M_f}, V_{h,f} = V^*_{h,M_f}$

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$
- Hypothesis classes $\mathcal{F}$ (and $\mathcal{G}$) [Du et al. '21]: identified by a pair $\{Q_f, V_f\}_{f \in \mathcal{F}}$
- Per step observed transition: $o_h := (s_h, a_h, s_{h+1})$
- Discriminator function class: $\mathcal{V}$, as in integral probability metrics (IPM) and witness rank
- Coupling function: $G_{h,f^*}(f,g) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^d$

f^*: true model

- Estimation function ℓ: surrogate loss function of the Bellman error.

Admissible Bellman Characterization

Model-free

$\{Q_f\}, V_{h,f}(s) = \max_a Q_{h,f}(s,a)$

Model-based

$Q_{h,f} = Q^*_{h,M_f}, V_{h,f} = V^*_{h,M_f}$

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$
- Hypothesis classes $\mathcal{F}$ (and $\mathcal{G}$) [Du et al. '21]: identified by a pair $\{Q_f, V_f\}_{f \in \mathcal{F}}$
- Per step observed transition: $o_h := (s_h, a_h, s_{h+1})$
- Discriminator function class: $\mathcal{V}$, as in integral probability metrics (IPM) and witness rank
- Coupling function: $G_{h,f^*}(f,g) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^d$ f^*: true model
- Estimation function ℓ: surrogate loss function of the Bellman error.

$\ell_{h,f}(o_h, f_{h+1}, g_h, v) : (\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \times \mathcal{F} \times \mathcal{G} \times \mathcal{V} \to \mathbb{R}^{d_s}$

Admissible Bellman Characterization

Model-free

$\{Q_f\}, V_{h,f}(s) = \max_a Q_{h,f}(s,a)$

Model-based

$Q_{h,f} = Q^*_{h,M_f}, V_{h,f} = V^*_{h,M_f}$

- MDP M, a trajectory of states and actions $s_1, a_1, \ldots, s_H$
- Hypothesis classes $\mathcal{F}$ (and $\mathcal{G}$) [Du et al. '21]: identified by a pair $\{Q_f, V_f\}_{f \in \mathcal{F}}$
- Per step observed transition: $o_h := (s_h, a_h, s_{h+1})$
- Discriminator function class: $\mathcal{V}$, as in integral probability metrics (IPM) and witness rank
- Coupling function: $G_{h,f^*}(f,g) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^d$

f^*: true model

- Estimation function ℓ: surrogate loss function of the Bellman error.

Realizability assumption:

$f^* \in \mathcal{F}, Q^*_h(s,a) = Q_{h,f^*}(s,a)$

Generalized Completeness:

$\mathcal{T}_h \mathcal{F}_{h+1} \subseteq \mathcal{G}_h$

$\ell_{h,f}(o_h, f_{h+1}, g_h, v) : (\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \times \mathcal{F} \times \mathcal{G} \times \mathcal{V} \to \mathbb{R}^{d_s}$

Admissible Bellman Characterization

Admissible Bellman Characterization

- Dominating average estimation function

Admissible Bellman Characterization

- Dominating average estimation function

$$\max_{v \in \mathscr{V}} \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f, g) \right)^2$$

Admissible Bellman Characterization

- Dominating average estimation function

$$\max_{v \in \mathscr{V}} \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f, g) \right)^2$$

- Bellman Dominance

Admissible Bellman Characterization

- Dominating average estimation function

$$\max_{v \in \mathscr{V}} \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f, g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f, f) \right|$$

Admissible Bellman Characterization

- Dominating average estimation function

$$\max_{v \in \mathscr{V}} \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f, g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f, f) \right|$$

When taking

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) = Q_{h,g}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1})$$

One possible choice of G is

$$G_{h,f^*}(f, g) = \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{op}} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right]$$

Reduces to low BE dimension

Admissible Bellman Characterization

- Dominating average estimation function

$$\max_{v \in \mathscr{V}} \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f, g) \right)^2$$

When taking

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = Q_{h,g}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1})$$

One possible choice of G is

$$G_{h,f^*}(f, g) = \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{op}} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right]$$

Reduces to low BE dimension

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f, f) \right|$$

Further, if $G_{h,f^*}(f, g)$ is $\left\langle W_h(f), X_h(g) \right\rangle$, reduces to low Bellman rank

Functional Eluder Dimension

Functional Eluder Dimension

- The length of the existing longest sequence $f_1, \ldots, f_n \in \mathscr{F}$ satisfying for some $\epsilon' \geq \epsilon$ and any $2 \leq t \leq n$, there exists $g \in \mathscr{F}$ such that $\sqrt{\sum_{i=1}^{t-1} \big(G(g, f_i)\big)^2} \leq \epsilon'$ while $|G(g, f_t)| > \epsilon'$

Functional Eluder Dimension

- The length of the existing longest sequence $f_1, \ldots, f_n \in \mathcal{F}$ satisfying for some $\epsilon' \geq \epsilon$ and any $2 \leq t \leq n$, there exists $g \in \mathcal{F}$ such that $\sqrt{\sum_{i=1}^{t-1} \left(G(g, f_i)\right)^2} \leq \epsilon'$ while $|G(g, f_t)| > \epsilon'$

- When $G_h(f, g) := \mathbb{E}_{\pi_{h,g}}(Q_{h,f} - \mathcal{T}_h Q_{f,h+1})$, $\dim_{\mathsf{FE}}(\mathcal{F}, G, \epsilon) = \dim_{\mathsf{BE}}(\mathcal{F}, G, \epsilon)$

Decomposable Estimation Function

Decomposable Estimation Function

- Decomposability (inherited from the Bellman error)

Decomposable Estimation Function

- Decomposability (inherited from the Bellman error)

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) - \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h\right] = \ell_{h,f'}(o_h, f_{h+1}, \mathcal{T}(f)_h, v)$$

Decomposable Estimation Function

- Decomposability (inherited from the Bellman error)

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) - \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h\right] = \ell_{h,f'}(o_h, f_{h+1}, \mathcal{T}(f)_h, v)$$

- Global Discriminator Optimality: there exists a global maximum $v_h^*(f) \in \mathcal{V}$

Decomposable Estimation Function

- Decomposability (inherited from the Bellman error)

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) - \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h\right] = \ell_{h,f'}(o_h, f_{h+1}, \mathcal{T}(f)_h, v)$$

- Global Discriminator Optimality: there exists a global maximum $v_h^*(f) \in \mathcal{V}$

$$\left\| \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v_h^*(f)) \mid s_h, a_h\right] \right\| \geq \left\| \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h\right] \right\|$$

Decomposable Estimation Function

- Decomposability (inherited from the Bellman error)

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) - \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h\right] = \ell_{h,f'}(o_h, f_{h+1}, \mathcal{T}(f)_h, v)$$

- Global Discriminator Optimality: there exists a global maximum $v_h^*(f) \in \mathcal{V}$

$$\left\| \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v_h^*(f)) \mid s_h, a_h\right] \right\| \geq \left\| \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h\right] \right\|$$

- When $\ell_{h,f'}(o_h, f_{h+1}, g_h, v) := Q_{h,g}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})$

Decomposable Estimation Function

- Decomposability (inherited from the Bellman error)

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) - \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h\right] = \ell_{h,f'}(o_h, f_{h+1}, \mathcal{T}(f)_h, v)$$

- Global Discriminator Optimality: there exists a global maximum $v_h^*(f) \in \mathcal{V}$

$$\left\| \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v_h^*(f)) \mid s_h, a_h\right] \right\| \geq \left\| \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h\right] \right\|$$

- When $\ell_{h,f'}(o_h, f_{h+1}, g_h, v) := Q_{h,g}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})$

$$\left[Q_{h,g}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})\right] - \left[Q_{h,g}(s_h, a_h) - (\mathcal{T}_h V_{h+1})(s_h, a_h)\right] = (\mathcal{T}_h V_{h+1})(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})$$

Decomposable Estimation Function

- Decomposability (inherited from the Bellman error)

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) - \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h\right] = \ell_{h,f'}(o_h, f_{h+1}, \mathcal{T}(f)_h, v)$$

- Global Discriminator Optimality: there exists a global maximum $v_h^*(f) \in \mathcal{V}$

$$\left\| \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v_h^*(f)) \mid s_h, a_h\right] \right\| \geq \left\| \mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h\right] \right\|$$

- When $\ell_{h,f'}(o_h, f_{h+1}, g_h, v) := Q_{h,g}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})$

$$\left[Q_{h,g}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})\right] - \left[Q_{h,g}(s_h, a_h) - (\mathcal{T}_h V_{h+1})(s_h, a_h)\right] = (\mathcal{T}_h V_{h+1})(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1})$$

- Global discriminator optimality requires a rich enough $\mathcal{V}$, which is generally satisfied by existing MDP models

Outline

01 **Background**
- Preliminaries
- Literature of RL models
- Model-based and Model-free RL
- General Function Approximation

02 **Our ABC Framework**
- Scope and sample complexity of our framework
- Admissible bellman characterization
- Functional eluder dimension
- Decomposable estimation function

03 **MDP Instances**
- Linear mixture MDP
- Low witness rank
- Kernelized nonlinear regulator (KNR)

04 **Algorithm and Main Results**
- Algorithm
- Proof sketch
- Main theorem

05 **Implications for specific MDPs**
- Linear mixture MDPs
- Kernelized nonlinear regulator (KNR)
- Low witness rank

Linear Mixture MDP

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learn-ing for linear mixture markov decision processes. In Conference on Learning Theory, pages 4532–4576. PMLR, 2021a

Linear Mixture MDP

- Linear in feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learn-ing for linear mixture markov decision processes. In Conference on Learning Theory, pages 4532–4576. PMLR, 2021a

Linear Mixture MDP

- Linear in feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

$$\mathbb{P}_h(s' \mid s, a) = \left\langle \phi(s, a, s'), \theta_h^* \right\rangle, \quad r_h(s, a) = \left\langle \psi(s, a), \theta_h^* \right\rangle$$

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learn-ing for linear mixture markov decision processes. In Conference on Learning Theory, pages 4532–4576. PMLR, 2021a

Linear Mixture MDP

- Linear in feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

$$\mathbb{P}_h(s' \mid s, a) = \left\langle \phi(s, a, s'), \theta_h^* \right\rangle, \quad r_h(s, a) = \left\langle \psi(s, a), \theta_h^* \right\rangle$$

- The transition and reward is uniquely given by θ_h

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learn-ing for linear mixture markov decision processes. In Conference on Learning Theory, pages 4532–4576. PMLR, 2021a

Linear Mixture MDP

- Linear in feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

$$\mathbb{P}_h(s' \mid s, a) = \left\langle \phi(s, a, s'), \theta_h^* \right\rangle, \quad r_h(s, a) = \left\langle \psi(s, a), \theta_h^* \right\rangle$$

- The transition and reward is uniquely given by θ_h

$$\sum_{s'} \mathbb{P}_h^{\theta}(s' \mid s, a) V(s') = \left\langle \sum_{s'} \phi(s, a, s') V(s'), \theta_h \right\rangle, \quad r_h^{\theta}(s, a) = \left\langle \psi(s, a), \theta_h \right\rangle$$

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learn-ing for linear mixture markov decision processes. In Conference on Learning Theory, pages 4532–4576. PMLR, 2021a

Linear Mixture MDP

- Linear in feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

$$\mathbb{P}_h(s' \mid s, a) = \left\langle \phi(s, a, s'), \theta_h^* \right\rangle, \quad r_h(s, a) = \left\langle \psi(s, a), \theta_h^* \right\rangle$$

- The transition and reward is uniquely given by θ_h

$$\sum_{s'} \mathbb{P}_h^{\theta}(s' \mid s, a) V(s') = \left\langle \sum_{s'} \phi(s, a, s') V(s'), \theta_h \right\rangle, \quad r_h^{\theta}(s, a) = \left\langle \psi(s, a), \theta_h \right\rangle$$

- The hypothesis class is defined as: $\mathcal{F}_h = \{\theta_h \in \mathcal{H}\}$

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learn-ing for linear mixture markov decision processes. In Conference on Learning Theory, pages 4532–4576. PMLR, 2021a

Linear Mixture MDP

- Linear in feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

$$\mathbb{P}_h(s' \mid s, a) = \left\langle \phi(s, a, s'), \theta_h^* \right\rangle, \quad r_h(s, a) = \left\langle \psi(s, a), \theta_h^* \right\rangle$$

- The transition and reward is uniquely given by θ_h

$$\sum_{s'} \mathbb{P}_h^{\theta}(s' \mid s, a) V(s') = \left\langle \sum_{s'} \phi(s, a, s') V(s'), \theta_h \right\rangle, \quad r_h^{\theta}(s, a) = \left\langle \psi(s, a), \theta_h \right\rangle$$

- The hypothesis class is defined as: $\mathcal{F}_h = \{\theta_h \in \mathcal{H}\}$

- We choose the following DEF, where $g, f' \in \mathcal{F}$ are two models

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learn-ing for linear mixture markov decision processes. In Conference on Learning Theory, pages 4532–4576. PMLR, 2021a

Linear Mixture MDP

- Linear in feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

$$\mathbb{P}_h(s' \mid s, a) = \left\langle \phi(s, a, s'), \theta_h^* \right\rangle, \quad r_h(s, a) = \left\langle \psi(s, a), \theta_h^* \right\rangle$$

- The transition and reward is uniquely given by θ_h

$$\sum_{s'} \mathbb{P}_h^{\theta}(s' \mid s, a) V(s') = \left\langle \sum_{s'} \phi(s, a, s') V(s'), \theta_h \right\rangle, \quad r_h^{\theta}(s, a) = \left\langle \psi(s, a), \theta_h \right\rangle$$

- The hypothesis class is defined as: $\mathcal{F}_h = \{\theta_h \in \mathcal{H}\}$

- We choose the following DEF, where $g, f' \in \mathcal{F}$ are two models

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^{\top} \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right) - r_h - V_{h+1,f'}(s_{h+1})$$

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learn-ing for linear mixture markov decision processes. In Conference on Learning Theory, pages 4532–4576. PMLR, 2021a

Linear Mixture MDP

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^{\top} \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right) - r_h - V_{h+1,f}(s_{h+1})$$

Linear Mixture MDP

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^{\top} \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right) - r_h - V_{h+1,f}(s_{h+1})$$

- The expectation of DEF equals to:

Linear Mixture MDP

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right) - r_h - V_{h+1,f'}(s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \left(\theta_{h,g} - \theta_h^* \right)^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right)$$

Linear Mixture MDP

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^{\top}\left(\psi(s_h, a_h) + \sum_{s'}\phi(s_h, a_h, s')V_{h+1,f'}(s')\right) - r_h - V_{h+1,f'}(s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}}\left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h\right] = \left(\theta_{h,g} - \theta_h^*\right)^{\top}\left(\psi(s_h, a_h) + \sum_{s'}\phi(s_h, a_h, s')V_{h+1,f'}(s')\right)$$

- The above measures the discrepancy in optimality between g and f^*

Linear Mixture MDP

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right) - r_h - V_{h+1,f'}(s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \left(\theta_{h,g} - \theta_h^* \right)^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right)$$

- The above measures the discrepancy in optimality between g and f^*
- On the other hand, the expectation of Bellman residual is

Linear Mixture MDP

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right) - r_h - V_{h+1,f}(s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_{h,f}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \left(\theta_{h,g} - \theta_h^* \right)^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right)$$

- The above measures the discrepancy in optimality between g and f^*
- On the other hand, the expectation of Bellman residual is

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] = \left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h, a_h \sim \pi_f} \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right)$$

Linear Mixture MDP

$$\ell_{h,f'}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right) - r_h - V_{h+1,f'}(s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_{h,f'}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \left(\theta_{h,g} - \theta_h^* \right)^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f'}(s') \right)$$

- The above measures the discrepancy in optimality between g and f^*
- On the other hand, the expectation of Bellman residual is

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] = \left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h, a_h \sim \pi_f} \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right)$$

- Belongs to ABC class with

Linear Mixture MDP

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right) - r_h - V_{h+1,f}(s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_{h,f}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \left(\theta_{h,g} - \theta_h^* \right)^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right)$$

- The above measures the discrepancy in optimality between g and f^*
- On the other hand, the expectation of Bellman residual is

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] = \left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h, a_h \sim \pi_f} \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right)$$

- Belongs to ABC class with

$$G_{h,f^*}(f, g) := \left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h, a_h \sim \pi_g} \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,g}(s') \right)$$

Linear Mixture MDP

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = \theta_{h,g}^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right) - r_h - V_{h+1,f}(s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_{h,f}(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \left(\theta_{h,g} - \theta_h^* \right)^\top \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right)$$

- The above measures the discrepancy in optimality between g and f^*
- On the other hand, the expectation of Bellman residual is

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] = \left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h, a_h \sim \pi_f} \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,f}(s') \right)$$

- Belongs to ABC class with

$$G_{h,f^*}(f, g) := \left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h, a_h \sim \pi_g} \left(\psi(s_h, a_h) + \sum_{s'} \phi(s_h, a_h, s') V_{h+1,g}(s') \right)$$

- The FE dimension of $\mathscr{F}$ w.r.t G is d

Low Witness Rank

Sun, Wen, et al. "Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches." *Conference on learning theory*. PMLR, 2019.

Low Witness Rank

- Model-based assumption, Factored MDP, low Bellman rank
  $\mathscr{V} = \{\mathscr{V}_h\}_{h\in[H]}, \mathscr{V}_h \subset \mathscr{S} \times \mathscr{A} \times \mathscr{S} \to \mathbb{R}$, hypothesis class $\mathscr{F}$, mappings
  $X_h : \mathscr{F} \to \mathbb{R}^d, W_h : \mathscr{F} \to \mathbb{R}^d$

Sun, Wen, et al. "Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches." *Conference on learning theory*. PMLR, 2019.

Low Witness Rank

- Model-based assumption, Factored MDP, low Bellman rank
  $\mathscr{V} = \{\mathscr{V}_h\}_{h\in[H]}, \mathscr{V}_h \subset \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, hypothesis class $\mathscr{F}$, mappings $X_h : \mathscr{F} \to \mathbb{R}^d, W_h : \mathscr{F} \to \mathbb{R}^d$

$$\max_{v\in\mathscr{V}_h} \mathbb{E}_{s_h\sim\pi_f, a_h\sim\pi_g} \left| \mathbb{E}_{\tilde{s}\sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

Sun, Wen, et al. "Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches." *Conference on learning theory*. PMLR, 2019.

Low Witness Rank

- Model-based assumption, Factored MDP, low Bellman rank
  $\mathscr{V} = \{\mathscr{V}_h\}_{h\in[H]}, \mathscr{V}_h \subset \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, hypothesis class $\mathscr{F}$, mappings $X_h : \mathscr{F} \to \mathbb{R}^d, W_h : \mathscr{F} \to \mathbb{R}^d$

$$\max_{v\in\mathscr{V}_h} \mathbb{E}_{s_h\sim\pi_f, a_h\sim\pi_g} \left| \mathbb{E}_{\tilde{s}\sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h\sim\pi_f, a_h\sim\pi_g} \left| \mathbb{E}_{\tilde{s}\sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

Sun, Wen, et al. "Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches." *Conference on learning theory*. PMLR, 2019.

Low Witness Rank

- Model-based assumption, Factored MDP, low Bellman rank
  $\mathscr{V} = \{\mathscr{V}_h\}_{h\in[H]}, \mathscr{V}_h \subset \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, hypothesis class $\mathscr{F}$, mappings $X_h : \mathscr{F} \to \mathbb{R}^d, W_h : \mathscr{F} \to \mathbb{R}^d$

$$\max_{v\in\mathscr{V}_h} \mathbb{E}_{s_h\sim\pi_f, a_h\sim\pi_g} \left| \mathbb{E}_{\tilde{s}\sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h\sim\pi_f, a_h\sim\pi_g} \left| \mathbb{E}_{\tilde{s}\sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

- The hypothesis class is defined as: $\mathscr{F}_h = \mathscr{M}$

Sun, Wen, et al. "Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches." *Conference on learning theory*. PMLR, 2019.

Low Witness Rank

- Model-based assumption, Factored MDP, low Bellman rank
  $\mathscr{V} = \{\mathscr{V}_h\}_{h\in[H]}, \mathscr{V}_h \subset \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, hypothesis class $\mathscr{F}$, mappings
  $X_h : \mathscr{F} \to \mathbb{R}^d, W_h : \mathscr{F} \to \mathbb{R}^d$

$$\max_{v\in\mathscr{V}_h} \mathbb{E}_{s_h\sim\pi_f, a_h\sim\pi_g} \left| \mathbb{E}_{\tilde{s}\sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h\sim\pi_f, a_h\sim\pi_g} \left| \mathbb{E}_{\tilde{s}\sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

- The hypothesis class is defined as: $\mathscr{F}_h = \mathscr{M}$
- We choose the following DEF, where $g \in \mathscr{F}$

Sun, Wen, et al. "Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches." *Conference on learning theory*. PMLR, 2019.

Low Witness Rank

- Model-based assumption, Factored MDP, low Bellman rank
  $\mathscr{V} = \{\mathscr{V}_h\}_{h\in[H]}, \mathscr{V}_h \subset \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, hypothesis class $\mathscr{F}$, mappings $X_h : \mathscr{F} \to \mathbb{R}^d, W_h : \mathscr{F} \to \mathbb{R}^d$

$$\max_{v\in\mathscr{V}_h} \mathbb{E}_{s_h\sim\pi_f, a_h\sim\pi_g} \left| \mathbb{E}_{\tilde{s}\sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h\sim\pi_f, a_h\sim\pi_g} \left| \mathbb{E}_{\tilde{s}\sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s}\sim\mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

- The hypothesis class is defined as: $\mathscr{F}_h = \mathscr{M}$

- We choose the following DEF, where $g \in \mathscr{F}$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s}\sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

Sun, Wen, et al. "Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches." *Conference on learning theory*. PMLR, 2019.

Low Witness Rank

Low Witness Rank

$$\max_{v \in \mathscr{V}_h} \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left[\mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right] \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

Low Witness Rank

- The expectation of DEF equals to:

$$\max_{v \in \mathscr{V}_h} \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left[\mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right] \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

Low Witness Rank

$$\max_{v \in \mathscr{V}_h} \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s})$$

Low Witness Rank

$$\max_{v \in \mathscr{V}_h} \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s})$$

- The above measures the discrepancy in optimality between g and the true model

Low Witness Rank

$$\max_{v \in \mathscr{V}_h} \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s})$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual is by definition

Low Witness Rank

$$\max_{v \in \mathscr{V}_h} \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s})$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual is by definition

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

Low Witness Rank

$$\max_{v \in \mathscr{V}_h} \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s})$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual is by definition

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

- Belongs to ABC class with

Low Witness Rank

$$\max_{v \in \mathscr{V}_h} \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s})$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual is by definition

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

- Belongs to ABC class with

$$G_{h,f^*}(f, g) := \left\langle W_h(f), X_h(g) \right\rangle$$

Low Witness Rank

$$\max_{v \in \mathscr{V}_h} \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s}) \right| \geq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

$$\ell_h(o_h, f_{h+1}, g_h, v) = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - v(s_h, a_h, s_{h+1})$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = \mathbb{E}_{\tilde{s} \sim g_h} v(s_h, a_h, \tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} v(s_h, a_h, \tilde{s})$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual is by definition

$$\kappa \cdot \mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_g} \left| \mathbb{E}_{\tilde{s} \sim g_h} V_{h+1,g}(\tilde{s}) - \mathbb{E}_{\tilde{s} \sim \mathbb{P}_h} V_{h+1,g}(\tilde{s}) \right| \leq \left\langle W_h(g), X_h(f) \right\rangle$$

- Belongs to ABC class with

$$G_{h,f^*}(f, g) := \left\langle W_h(f), X_h(g) \right\rangle$$

- The FE dimension of $\mathscr{F}$ w.r.t G is d

KNR

Mania, Horia, Michael I. Jordan, and Benjamin Recht. "Active learning for nonlinear system identification with guarantees." *The Journal of Machine Learning Research* 23.1 (2022): 1433-1462.

Kakade, Sham, et al. "Information theoretic regret bounds for online nonlinear control." Advances in Neural Information Processing Systems 33 (2020): 15312-15325.

KNR

- Models a nonlinear control dynamics on an RKHS $\mathscr{H}$ of finite or countably infinite dimension, feature mapping $\phi(s_h, a_h) : \mathcal{S} \times \mathcal{A} \to H$, uniformly bounded $\|\phi(s, a)\| \leq B$

Mania, Horia, Michael I. Jordan, and Benjamin Recht. "Active learning for nonlinear system identification with guarantees." *The Journal of Machine Learning Research* 23.1 (2022): 1433-1462.

Kakade, Sham, et al. "Information theoretic regret bounds for online nonlinear control." Advances in Neural Information Processing Systems 33 (2020): 15312-15325.

KNR

- Models a nonlinear control dynamics on an RKHS $\mathscr{H}$ of finite or countably infinite dimension, feature mapping $\phi(s_h, a_h) : \mathcal{S} \times \mathcal{A} \to H$, uniformly bounded $\|\phi(s, a)\| \leq B$

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

Mania, Horia, Michael I. Jordan, and Benjamin Recht. "Active learning for nonlinear system identification with guarantees." *The Journal of Machine Learning Research* 23.1 (2022): 1433-1462.

Kakade, Sham, et al. "Information theoretic regret bounds for online nonlinear control." Advances in Neural Information Processing Systems 33 (2020): 15312-15325.

KNR

- Models a nonlinear control dynamics on an RKHS $\mathscr{H}$ of finite or countably infinite dimension, feature mapping $\phi(s_h, a_h) : \mathcal{S} \times \mathcal{A} \to H$, uniformly bounded $\|\phi(s, a)\| \leq B$

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

- Hypothesis class: $\left\{ U_h^* \in \mathbb{R}^{d_s} \times \mathscr{H} \right\}_{h \in [H]}$

Mania, Horia, Michael I. Jordan, and Benjamin Recht. "Active learning for nonlinear system identification with guarantees." *The Journal of Machine Learning Research* 23.1 (2022): 1433-1462.

Kakade, Sham, et al. "Information theoretic regret bounds for online nonlinear control." Advances in Neural Information Processing Systems 33 (2020): 15312-15325.

KNR

- Models a nonlinear control dynamics on an RKHS $\mathscr{H}$ of finite or countably infinite dimension, feature mapping $\phi(s_h, a_h) : \mathcal{S} \times \mathcal{A} \to H$, uniformly bounded $\|\phi(s, a)\| \leq B$

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

- Hypothesis class: $\left\{ U_h^* \in \mathbb{R}^{d_s} \times \mathscr{H} \right\}_{h \in [H]}$
- DEF:

Mania, Horia, Michael I. Jordan, and Benjamin Recht. "Active learning for nonlinear system identification with guarantees." *The Journal of Machine Learning Research* 23.1 (2022): 1433-1462.

Kakade, Sham, et al. "Information theoretic regret bounds for online nonlinear control." Advances in Neural Information Processing Systems 33 (2020): 15312-15325.

KNR

- Models a nonlinear control dynamics on an RKHS $\mathscr{H}$ of finite or countably infinite dimension, feature mapping $\phi(s_h, a_h) : \mathcal{S} \times \mathcal{A} \to H$, uniformly bounded $\|\phi(s, a)\| \leq B$

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

- Hypothesis class: $\left\{ U_h^* \in \mathbb{R}^{d_s} \times \mathscr{H} \right\}_{h \in [H]}$

- DEF:

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

Mania, Horia, Michael I. Jordan, and Benjamin Recht. "Active learning for nonlinear system identification with guarantees." *The Journal of Machine Learning Research* 23.1 (2022): 1433-1462.

Kakade, Sham, et al. "Information theoretic regret bounds for online nonlinear control." Advances in Neural Information Processing Systems 33 (2020): 15312-15325.

KNR

KNR

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

KNR

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

- The expectation of DEF equals to:

KNR

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = (U_{h,g} - U_h^*) \phi(s_h, a_h)$$

KNR

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = (U_{h,g} - U_h^*) \phi(s_h, a_h)$$

- The above measures the discrepancy in optimality between g and the true model

KNR

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = (U_{h,g} - U_h^*) \phi(s_h, a_h)$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual i

KNR

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = (U_{h,g} - U_h^*) \phi(s_h, a_h)$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual i

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] \leq \frac{2H}{\sigma} \mathbb{E}_{s_h, a_h \sim \pi_f} \left\| (U_{h,f} - U_h^*) \phi(s_h, a_h) \right\|_2$$

KNR

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = (U_{h,g} - U_h^*) \phi(s_h, a_h)$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual i

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] \leq \frac{2H}{\sigma} \mathbb{E}_{s_h, a_h \sim \pi_f} \left\| (U_{h,f} - U_h^*) \phi(s_h, a_h) \right\|_2$$

- Belongs to ABC class with

KNR

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = (U_{h,g} - U_h^*) \phi(s_h, a_h)$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual i

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] \leq \frac{2H}{\sigma} \mathbb{E}_{s_h, a_h \sim \pi_f} \left\| (U_{h,f} - U_h^*) \phi(s_h, a_h) \right\|_2$$

- Belongs to ABC class with

$$G_{h,f^*}(f, g) := \sqrt{\mathbb{E}_{s_h, a_h \sim \pi_g} \left\| (U_{h,f} - U_h^*) \phi(s_h, a_h) \right\|^2}$$

KNR

$$s_{h+1} = U_h^* \phi(s_h, a_h) + \epsilon_{h+1}, \epsilon_{h+1} \sim \mathcal{N}(0, \sigma^2 I)$$

$$\ell_{h,f}(o_h, f_{h+1}, g_h, v) = U_{h,g} \phi(s_h, a_h) - s_{h+1}$$

- The expectation of DEF equals to:

$$\mathbb{E}_{s_{h+1}} \left[\ell_h(o_h, f_{h+1}, g_h, v) \mid s_h, a_h \right] = (U_{h,g} - U_h^*) \phi(s_h, a_h)$$

- The above measures the discrepancy in optimality between g and the true model
- On the other hand, the expectation of Bellman residual i

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] \leq \frac{2H}{\sigma} \mathbb{E}_{s_h, a_h \sim \pi_f} \left\| (U_{h,f} - U_h^*) \phi(s_h, a_h) \right\|_2$$

- Belongs to ABC class with

$$G_{h,f^*}(f, g) := \sqrt{\mathbb{E}_{s_h, a_h \sim \pi_g} \left\| (U_{h,f} - U_h^*) \phi(s_h, a_h) \right\|^2}$$

- The FE dimension of $\mathcal{F}$ w.r.t G is d_ϕ

Properties of ABC

Properties of ABC

- Dominating average estimation function
- Bellman Dominance

Properties of ABC

- Dominating average estimation function

$$\max_{v \in \mathscr{V}} \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f, g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f, f) \right|$$

Properties of ABC

- Coupling function G can be nonlinear

- Dominating average estimation function

$$\max_{v \in \mathscr{V}} \mathbb{E}_{s_h \sim \pi_g, a_h \sim \pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f, g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_{h,f}(s_h, a_h) - r(s_h, a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f, f) \right|$$

Properties of ABC

- Coupling function G can be nonlinear

KNR: $\sqrt{\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|^2}$

- Dominating average estimation function

$$\max_{v\in\mathscr{V}} \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f,g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f,f) \right|$$

Properties of ABC

- Coupling function G can be nonlinear

KNR: $\sqrt{\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|^2}$

Generalized linear Bellman complete: $\sqrt{\langle W_h(f), X_h(f) \rangle}$

- Dominating average estimation function

$$\max_{v\in\mathscr{V}} \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f,g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f,f) \right|$$

Properties of ABC

- Coupling function G can be nonlinear

KNR: $\sqrt{\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|^2}$

Generalized linear Bellman complete: $\sqrt{\langle W_h(f), X_h(f) \rangle}$

- Dominating average estimation function

$$\max_{v\in\mathscr{V}} \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f,g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f,f) \right|$$

- ℓ can be different from the expected Bellman error (any model-based)

Properties of ABC

- Coupling function G can be nonlinear

KNR: $\sqrt{\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|^2}$

Generalized linear Bellman complete: $\sqrt{\langle W_h(f), X_h(f) \rangle}$

- Dominating average estimation function

$$\max_{v\in\mathscr{V}} \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f,g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f,f) \right|$$

- ℓ can be different from the expected Bellman error (any model-based)
- Dominating average estimation function ℓ can be larger:

Properties of ABC

- Coupling function G can be nonlinear

KNR: $\sqrt{\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|^2}$

Generalized linear Bellman complete: $\sqrt{\langle W_h(f), X_h(f) \rangle}$

- Dominating average estimation function

$$\max_{v\in\mathscr{V}} \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f,g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f,f) \right|$$

- ℓ can be different from the expected Bellman error (any model-based)

- Dominating average estimation function ℓ can be larger:

Linear mixture MDP:

$$\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| \mathbb{E} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(\left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h,a_h\sim\pi_g} \left[\psi(s_h,a_h) + \sum_{s'} \phi(s_h,a_h,s') V_{h+1,g}(s') \right] \right)^2$$

Properties of ABC

- Coupling function G can be nonlinear

KNR: $\sqrt{\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|^2}$

Generalized linear Bellman complete: $\sqrt{\langle W_h(f), X_h(f) \rangle}$

- Dominating average estimation function

$$\max_{v\in\mathscr{V}} \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f,g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f,f) \right|$$

- ℓ can be different from the expected Bellman error (any model-based)
- Dominating average estimation function ℓ can be larger:

Linear mixture MDP:

$$\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| \mathbb{E} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(\left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h,a_h\sim\pi_g} \left[\psi(s_h,a_h) + \sum_{s'} \phi(s_h,a_h,s') V_{h+1,g}(s') \right] \right)^2$$

- Bellman dominance can be smaller:

Properties of ABC

- Coupling function G can be nonlinear

KNR: $\sqrt{\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|^2}$

Generalized linear Bellman complete: $\sqrt{\langle W_h(f), X_h(f) \rangle}$

- Dominating average estimation function

$$\max_{v\in\mathscr{V}} \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f,g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f,f) \right|$$

- ℓ can be different from the expected Bellman error (any model-based)
- Dominating average estimation function ℓ can be larger:

Linear mixture MDP:

$$\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| \mathbb{E} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(\left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h,a_h\sim\pi_g} \left[\psi(s_h,a_h) + \sum_{s'} \phi(s_h,a_h,s') V_{h+1,g}(s') \right] \right)^2$$

- Bellman dominance can be smaller:

KNR:

$$\mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] \leq \frac{2H}{\sigma} \mathbb{E}_{s_h,a_h\sim\pi_f} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|_2$$

Properties of ABC

- Coupling function G can be nonlinear

KNR: $\sqrt{\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|^2}$

Generalized linear Bellman complete: $\sqrt{\langle W_h(f), X_h(f) \rangle}$

- Dominating average estimation function

$$\max_{v\in\mathscr{V}} \mathbb{E}_{s_h\sim\pi_g,a_h\sim\pi_{op}} \left\| \mathbb{E}_{s_{h+1}} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(G_{h,f^*}(f,g) \right)^2$$

- Bellman Dominance

$$\kappa \cdot \left| \mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r(s_h,a_h) - V_{h+1,f}(s_{h+1}) \right] \right| \leq \left| G_{h,f^*}(f,f) \right|$$

- ℓ can be different from the expected Bellman error (any model-based)
- Dominating average estimation function ℓ can be larger:

Linear mixture MDP:

$$\mathbb{E}_{s_h,a_h\sim\pi_g} \left\| \mathbb{E} \left[\ell_{h,g}(o_h, f_{h+1}, f_h, v) \mid s_h, a_h \right] \right\|^2 \geq \left(\left(\theta_{h,f} - \theta_h^* \right)^\top \mathbb{E}_{s_h,a_h\sim\pi_g} \left[\psi(s_h,a_h) + \sum_{s'} \phi(s_h,a_h,s') V_{h+1,g}(s') \right] \right)^2$$

- Bellman dominance can be smaller:

KNR:

$$\mathbb{E}_{s_h,a_h\sim\pi_f} \left[Q_{h,f}(s_h,a_h) - r_h - V_{h+1,f}(s_{h+1}) \right] \leq \frac{2H}{\sigma} \mathbb{E}_{s_h,a_h\sim\pi_f} \left\| (U_{h,f} - U_h^*)\phi(s_h,a_h) \right\|_2$$

- ℓ can be vector valued

Outline

01 **Background**
- Preliminaries
- Literature of RL models
- Model-based and Model-free RL
- General Function Approximation

02 **Our ABC Framework**
- Scope and sample complexity of our framework
- Admissible bellman characterization
- Functional eluder dimension
- Decomposable estimation function

03 **MDP Instances**
- Linear mixture MDP
- Low witness rank
- Kernelized nonlinear regulator (KNR)

04 **Algorithm and Main Results**
- Algorithm
- Proof sketch
- Main theorem

05 **Implications for specific MDPs**
- Linear mixture MDPs
- Kernelized nonlinear regulator (KNR)
- Low witness rank

Algorithm (OPERA)

Set $\pi^t := \pi_{f^t}$ where f^t is taken as $\arg\max_{f \in \mathcal{F}} Q_{1,f}(s_1, \pi_f(s_1))$, subject to

$$\max_{v \in \mathcal{V}} \left\{ \sum_{i=1}^{t-1} \left\| \ell_{h,f^i}(o_h^i, f_{h+1}, f_h, v) \right\|^2 - \inf_{g_h \in \mathcal{G}_h} \sum_{i=1}^{t-1} \left\| \ell_{h,f^i}(o_h^i, f_{h+1}, g_h, v) \right\|^2 \right\} \leq \beta$$

Proof Sketch (Optimization based exploration)

$$\text{Regret}(T) := \sum_{t=1}^{T} \left[V_1^*(s_1) - V_1^{\pi^t}(s_1) \right]$$

- Take $\text{argmax}_{f \in \mathcal{F}} Q_{1,f}(s_1, \pi_f(s_1))$ in a confidence interval $\mathcal{C}^t(\beta)$.
- Our choice of β satisfies $f^* \in \mathcal{C}^t$.
- $V_1^*(s_1) - V_1^{\pi^t}(s_1) \leq V_{1,f^t}(s_1) - V_1^{\pi^t}(s_1)$
- $= \sum_{h=1}^{H} \mathbb{E}_{s_h, a_h \sim \pi^t} \left[Q_{h,f^t}(s_h, a_h) - r_h - V_{h+1,f^t}(s_{h+1}) \right]$

$$\sum_{t=1}^{T} \sum_{h=1}^{H} \left| \mathbb{E}_{s_h, a_h \sim \pi^t} \left[Q_{h,f^t}(s_h, a_h) - r_h - V_{h+1,f^t}(s_{h+1}) \right] \right|$$

Proof Sketch

Proof Sketch

$$\sum_{t=1}^{T} V_1^*(s_1) - V_1^{\pi^t}(s_1) \leq \frac{1}{\kappa} \sum_{t=1}^{T} \sum_{h=1}^{H} |\, G_{h,f^*}(f^t, f^t) \,|$$

Proof Sketch

$$\sum_{t=1}^{T} V_1^*(s_1) - V_1^{\pi^t}(s_1) \leq \frac{1}{\kappa} \sum_{t=1}^{T} \sum_{h=1}^{H} |G_{h,f^*}(f^t, f^t)|$$

Martingale concentration argument (Freedman's Inequality):

$$\sum_{i=1}^{t-1} (G_{h,f^*}(f^t, f^i))^2 \leq \mathcal{O}(\beta)$$

holds with probability at least $1 - \delta$, where
$\beta = c(\log(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta) + T\rho)$

Proof Sketch

$$\sum_{t=1}^{T} V_1^*(s_1) - V_1^{\pi^t}(s_1) \leq \frac{1}{\kappa} \sum_{t=1}^{T} \sum_{h=1}^{H} |\, G_{h,f^*}(f^t, f^t)\,|$$

Martingale concentration argument (Freedman's Inequality):

$$\sum_{i=1}^{t-1} (G_{h,f^*}(f^t, f^i))^2 \leq \mathcal{O}(\beta)$$

holds with probability at least $1-\delta$, where
$\beta = c(\log(TH\mathcal{N}_{\mathcal{L}}(\rho)/\delta) + T\rho)$

$$\sum_{i=1}^{t} |\, G(f_i, g_i)\,| \leq \mathcal{O}\left(\sqrt{\dim_{\mathsf{FE}}(\mathcal{F}, G, \omega)\beta t} + C \cdot \min\{t, \dim_{\mathsf{FE}}(\mathcal{F}, G, \omega)\} + t\omega\right)$$

Main Theorem

Theorem (Regret of OPERA)

For an MDP $\mathcal{M}$ with hypothesis classes $\mathcal{F}$, $\mathcal{G}$ under our structural assumption. We choose $\beta = \mathcal{O}(\log(THN_{\mathcal{L}}(1/T)/\delta))$ in OPERA. We have with probability at least $1-\delta$ the regret can be bounded as

$$\mathit{Regret}(T) \leq \mathcal{O}\left(\frac{H}{\kappa}\sqrt{T \cdot \dim_{FE}\left(\mathcal{F}, G, \sqrt{1/T}\right) \cdot \beta}\right).$$

- Dependent on both the functional eluder dimension $\dim_{\mathsf{FE}}$ and $N_{\mathscr{L}}(\sqrt{1/T})$
- For cases with linear structure of dimension d, i.e. Linear MDPs, Linear Mixture MDPs, Low Occupancy Complexity, Linear Q*/V*, the regret bound $\approx \mathcal{O}\left(\frac{dH}{\kappa}\sqrt{T}\right)$

Outline

01 **Background**
- Preliminaries
- Literature of RL models
- Model-based and Model-free RL
- General Function Approximation

02 **Our ABC Framework**
- Scope and sample complexity of our framework
- Admissible bellman characterization
- Functional eluder dimension
- Decomposable estimation function

03 **MDP Instances**
- Linear mixture MDP
- Low witness rank
- Kernelized nonlinear regulator (KNR)

04 **Algorithm and Main Results**
- Algorithm
- Proof sketch
- Main theorem

05 **Implications for specific MDPs**
- Linear mixture MDPs
- Kernelized nonlinear regulator (KNR)
- Low witness rank

Linear Mixture MDPs (Algorithm)

Set $\pi^t := \pi_{f^t}$ where f^t is taken as $\arg\max_{f \in \mathcal{F}} Q_{1,f}(s_1, \pi_f(s_1))$, subject to

$$\max_{v \in \mathcal{V}} \left\{ \sum_{i=1}^{t-1} \left\| \ell_{h,f^t}(o_h^i, f_{h+1}, f_h, v) \right\|^2 - \inf_{g_h \in \mathcal{G}_h} \sum_{i=1}^{t-1} \left\| \ell_{h,f^t}(o_h^i, f_{h+1}, g_h, v) \right\|^2 \right\} \leq \beta$$

Set $\pi^t := \pi_{f^t}$ where f^t is taken as $\arg\max_{f \in \mathcal{F}} Q_{1,f}(s_1, \pi_f(s_1))$, subject to

$$\hat{\theta}_{h,t} = \left(\Phi_h^{t-1} \left(\Phi_h^{t-1} \right)^\top \right)^{-1} \Phi_h^{t-1} \left(y_h^{t-1} \right)^\top, \qquad \left\| \theta_{h,t} - \hat{\theta}_{h,t} \right\|_{\Sigma_h^{t-1}}^2 \leq \beta \quad \forall h \in [H]$$

Where $\hat{\theta}_{h,t}$ is the solution to the value-target regression as in UCRL-VTR [Ayoub et al. '20]

- Global optimization version of UCRL-VTR on linear mixture MDPs
- Exhibits a $dH\sqrt{T}$ regret bound and d^2H^2/ϵ^2 sample complexity result
- Compared with the d^3H^4/ϵ^2 sample complexity (the best-known results on general frameworks that subsumes linear mixture MDPs)

KNR

Set $\pi^t := \pi_{f^t}$ where f^t is taken as $\arg\max_{f \in \mathcal{F}} Q_{1,f}(s_1, \pi_f(s_1))$, subject to

$$\max_{v \in \mathcal{V}} \left\{ \sum_{i=1}^{t-1} \left\| \ell_{h,f^t}(o_h^i, f_{h+1}, f_h, v) \right\|^2 - \inf_{g_h \in \mathcal{G}_h} \sum_{i=1}^{t-1} \left\| \ell_{h,f^t}(o_h^i, f_{h+1}, g_h, v) \right\|^2 \right\} \le \beta$$

Set $\pi^t := \pi_{f^t}$ where f^t is taken as $\arg\max_{f \in \mathcal{F}} Q_{1,f}(s_1, \pi_f(s_1))$, subject to

$$\sum_{i=1}^{t-1} \left\| U_{h,f}\phi(s_h^i, a_h^i) - s_{h+1}^i \right\|^2 - \inf_{g_h \in \mathcal{G}_h} \sum_{i=1}^{t-1} \left\| U_{h,g}\phi(s_h^i, a_h^i) - s_{h+1}^i \right\|^2 \le \beta$$

- The confidence set is equivalent to $\left\| (U_{h,f} - \hat{U}_{h,f})(\Sigma_h^{t-1})^{1/2} \right\|^2 \le \beta$
- $\Sigma_h^{t-1} := \Phi_h^{t-1}(\Phi_h^{t-1})^\top$ and $\hat{U}_{h,f}$ is the optimal solution to the least square problem$\arg\min_U \sum_{i=1}^{t-1} \|U\phi(s_h^i, a_h^i) - s_{h+1}^i\|^2$
- Inhomogeneous version of LC^3 [Kakade et al. '20] for KNR.
- Yields a regret bound of$\tilde{\mathcal{O}}\left(\sqrt{d_\phi^2 d_s H^4 T}\right)$. Matching the SOTA regret bound.

Low Witness Rank

Set $\pi^t := \pi_{f^t}$ where f^t is taken as $\arg\max_{f \in \mathcal{F}} Q_{1,f}(s_1, \pi_f(s_1))$, subject to

$$\max_{v \in \mathcal{V}} \left\{ \sum_{i=1}^{t-1} \left(\mathbb{E}_{\tilde{s} \sim f_h} v(s_h^i, a_h^i, \tilde{s}) - v(s_h^i, a_h^i, s_{h+1}^i) \right)^2 - \inf_{g_h \in \mathcal{G}_h} \sum_{i=1}^{t-1} \left(\mathbb{E}_{\tilde{s} \sim g_h} v(s_h^i, a_h^i, \tilde{s}) - v(s_h^i, a_h^i, s_{h+1}^i) \right)^2 \right\} \leq \beta$$

Set $\pi^t := \pi_{f^t}$ where f^t is taken as $\arg\max_{f \in \mathcal{F}} Q_{1,f}(s_1, \pi_f(s_1))$, subject to

$\max_{v \in \mathcal{V}} \left\{ \sum_{i=1}^{t-1} \left\| \ell_{h,f^t}(o_h^i, f_{h+1}, f_h, v) \right\|^2 - \inf_{g_h \in \mathcal{G}_h} \sum_{i=1}^{t-1} \left\| \ell_{h,f^t}(o_h^i, f_{h+1}, g_h, v) \right\|^2 \right\} \leq \beta$

- Outputs an ε-optimal policy π_{out} within $T = \tilde{\mathcal{O}}\left(H^2 |\mathcal{A}| W_\kappa \beta / (\kappa^2 \epsilon^2)\right)$ trajectories.
- Compared with previous best-known sample complexity result of $\widetilde{O}\left(H^3 W_\kappa^2 |\mathcal{A}| \log(T|\mathcal{M}||\mathcal{V}|/\delta)/(\kappa^2\epsilon^2)\right)$, our sample complexity is superior by a factor of dH up to a polylogarithmic prefactor in model parameters

Takeaway

- Unified framework that subsumes nearly all solvable MDP models, including model-based & model-free.
- A new type of estimation function for optimization-based exploration that can be vector.
- Propose Functional eluder dimension, with a sample-efficient algorithm named OPERA

A GENERAL FRAMEWORK FOR SAMPLE-EFFICIENT FUNCTION APPROXIMATION IN REINFORCEMENT LEARNING

Zixiang Chen[‡,*] **Chris Junchi Li**[◇,*] **Angela Yuan**[‡,*] **Quanquan Gu**[‡] **Michael I. Jordan**[◇,†]
[‡] Department of Computer Science, University of California, Los Angles
{chenzx19, hzyuan, qgu}@cs.ucla.edu
[◇] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
junchili@berkeley.edu, jordan@cs.berkeley.edu
[†] Department of Statistics, University of California, Berkeley

Thank you for Listening!