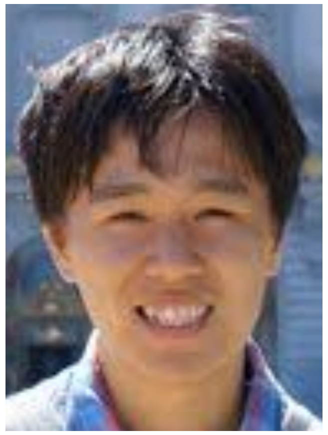
Model-based RL in Contextual Decision Processes:
PAC Bounds and Exponential Improvements over Model-free Approaches

Wen Sun

CMU -> MSR NYC

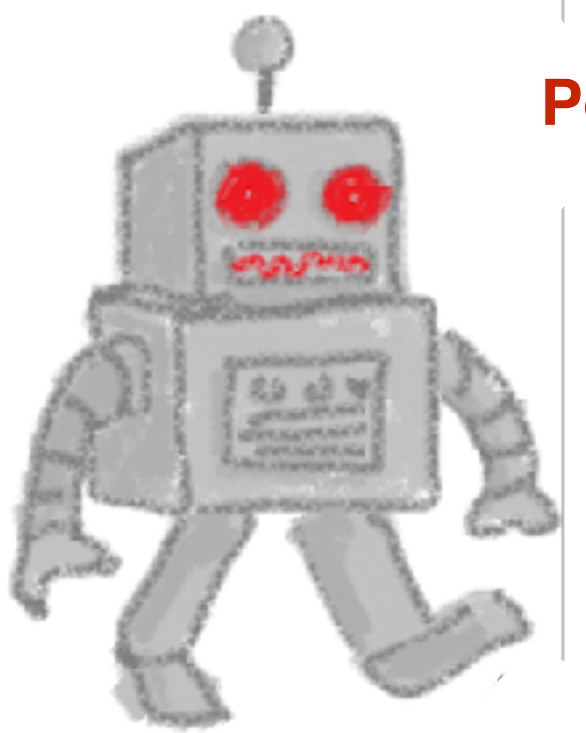
Joint work with Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford




Reinforcement Learning

Markov Decision Process

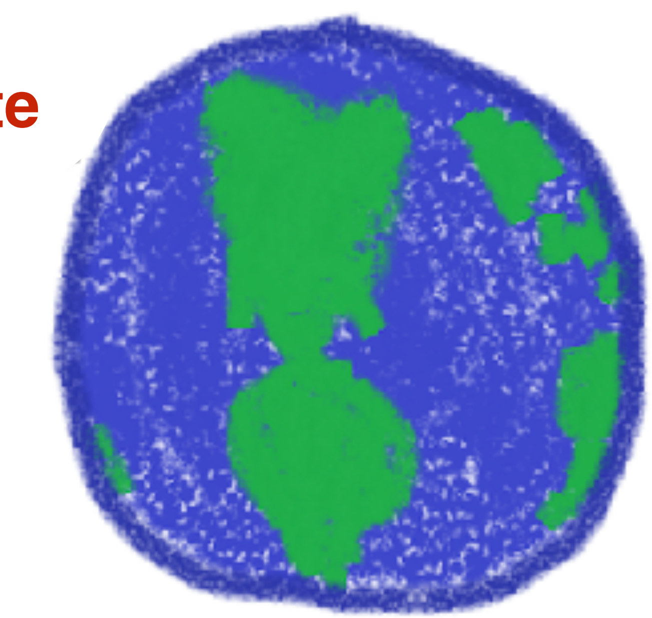
Agent



$$\pi(x) \rightarrow a$$

Policy: determine **action** based on **state**

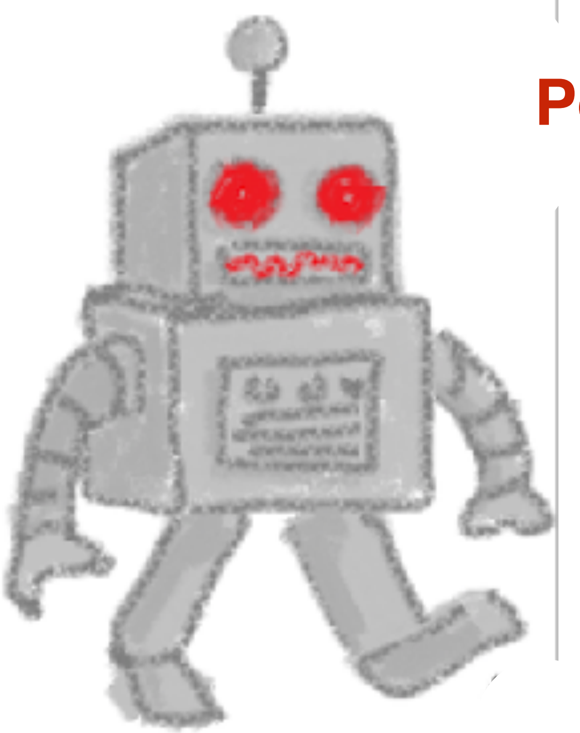
Environment



Reinforcement Learning

Markov Decision Process

Agent



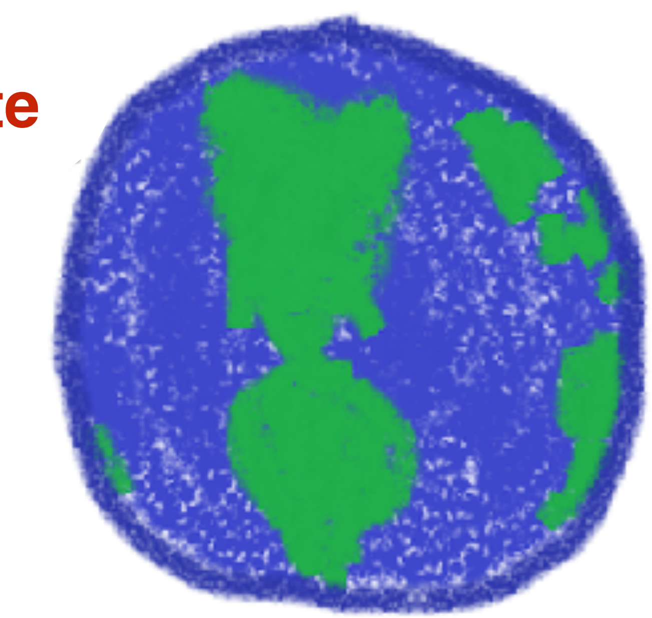
$$\pi(x) \rightarrow a$$

Policy: determine **action** based on **state**

Send **reward** and **next state** from a Markovian transition dynamics

$$r(x,a), \quad x' \sim P(\cdot|x,a)$$

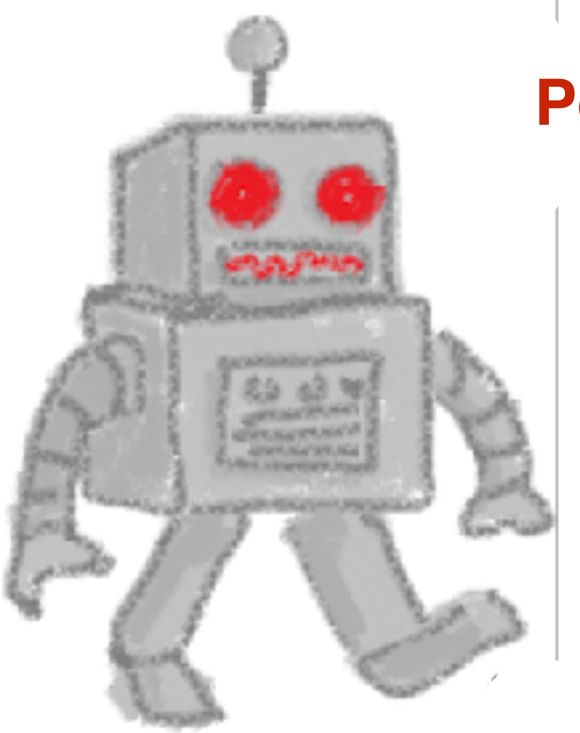
Environment



Reinforcement Learning

Markov Decision Process

Agent



$$\pi(x) \rightarrow a$$

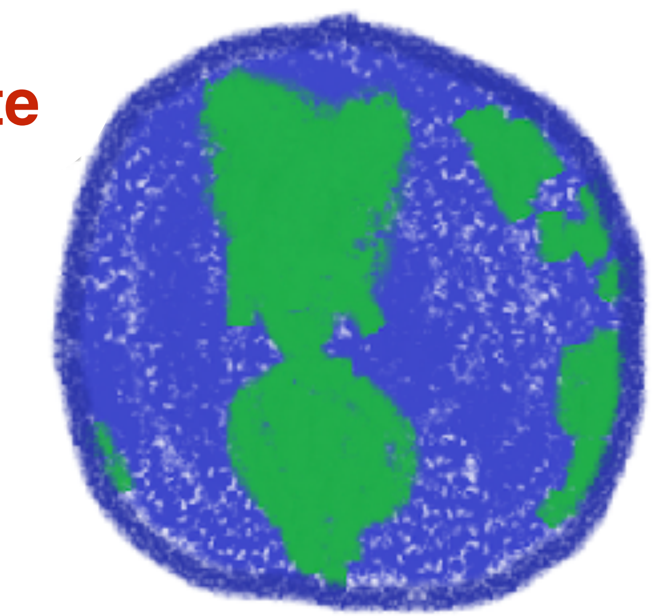
Policy: determine **action** based on **state**

H Steps

Send **reward** and **next state** from a Markovian transition dynamics

$$r(x, a), \quad x' \sim P(\cdot|x, a)$$

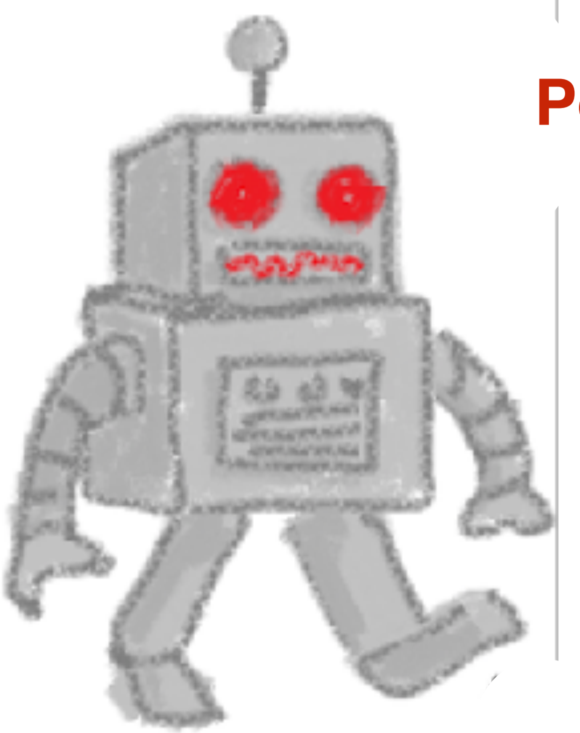
Environment



Reinforcement Learning

Markov Decision Process

Agent



$$\pi(x) \to a$$

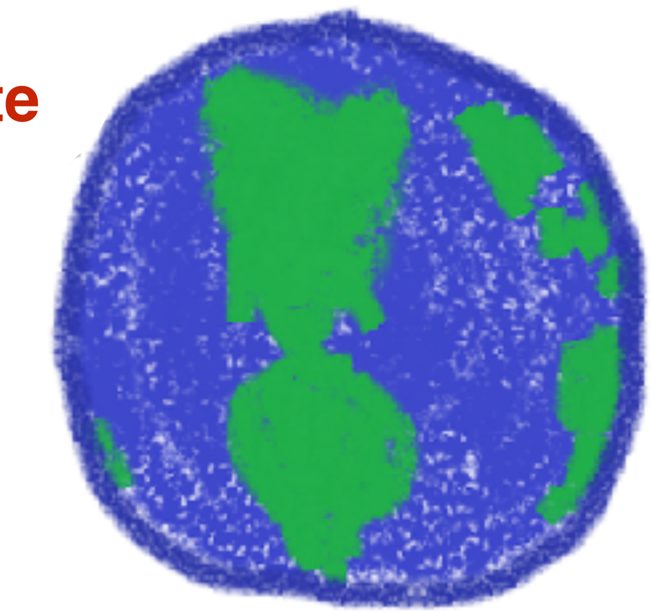
Policy: determine **action** based on **state**

H Steps

Send **reward** and **next state** from a Markovian transition dynamics

$$r(x, a), \quad x' \sim P(\cdot | x, a)$$

Environment

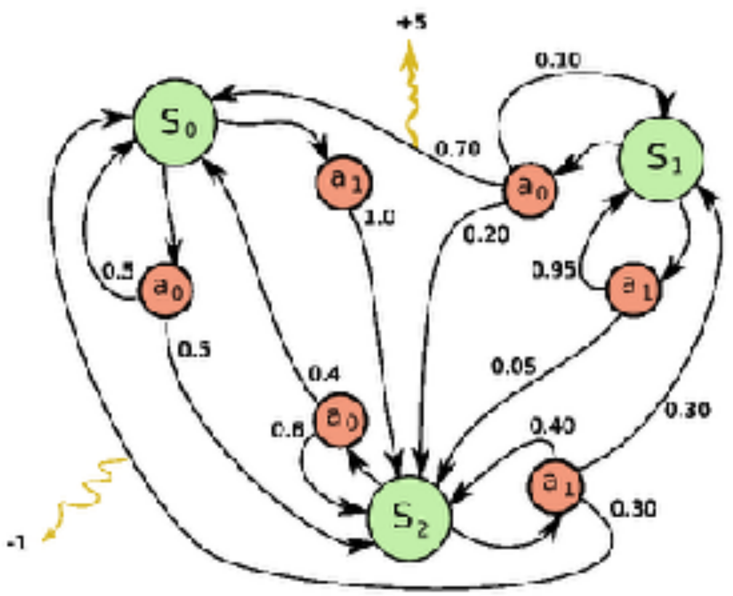


Maximize expected total reward:

$$J(\pi) = \mathbb{E}[r_1 + r_2 + \cdots + r_H | \pi]$$

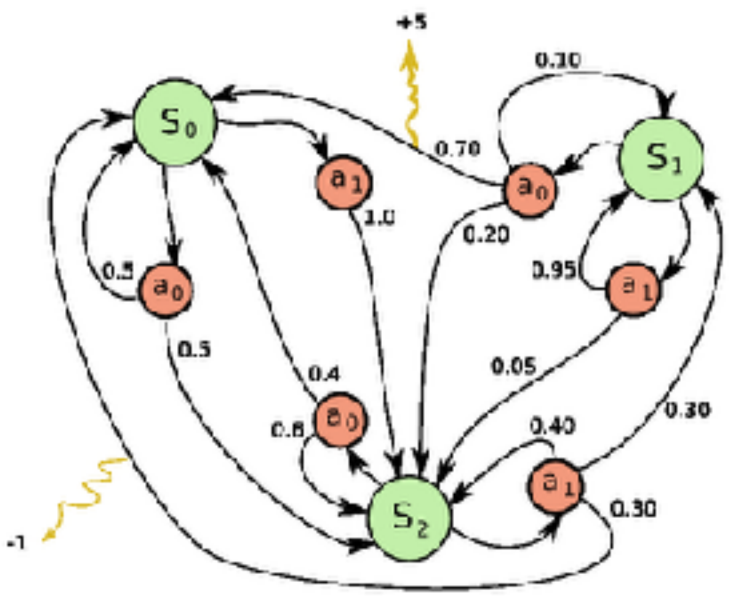
Progress of RL in Theory

Sample Efficiency in Small Discrete MDPs



Progress of RL in Theory

Sample Efficiency
in Small Discrete MDPs



Sample Complexity:

To achieve ϵ near-optimal policy,
need at most

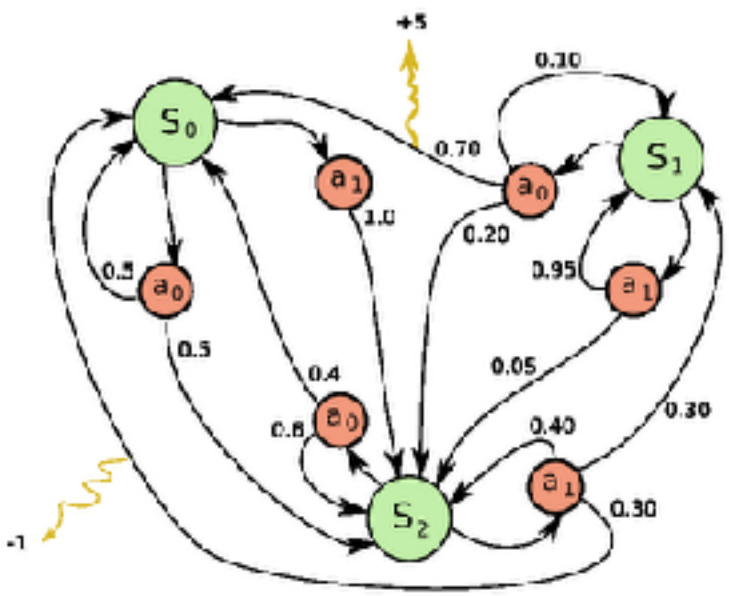
$$\text{poly}(\#\text{ of states}, \#\text{ of actions}, \text{Horizon}, 1/\epsilon)$$

many interactions

[e.g., Kearns & Singh, 02, Dann & Brunskill, 15, Azar et.al, 17]

Progress of RL in Theory

Sample Efficiency in Small Discrete MDPs



Sample Complexity:

To achieve ϵ near-optimal policy,
need at most

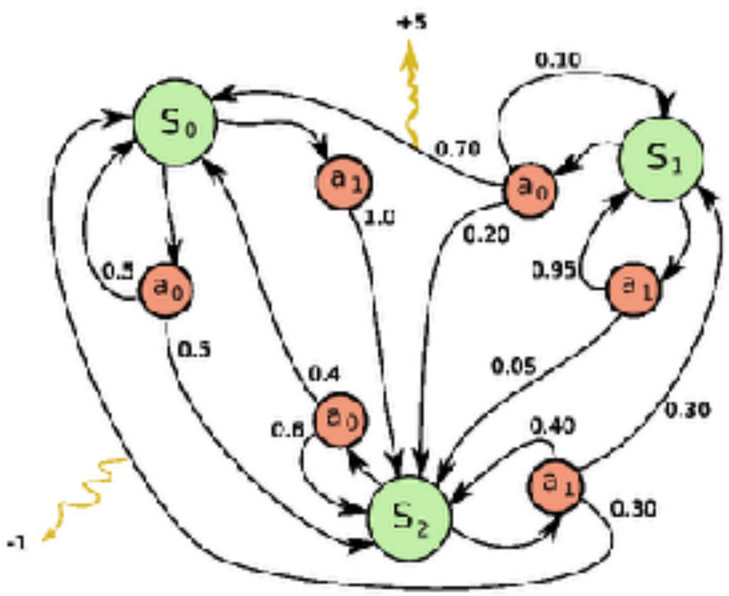
$$\boxed{\text{poly}(\#\text{ of states},}\ \#\text{ of actions}, \text{Horizon}, 1/\epsilon)$$

many interactions

[e.g., Kearns & Singh, 02, Dann & Brunskill, 15, Azar et.al, 17]

Progress of RL in Theory

Sample Efficiency in Small Discrete MDPs



Sample Complexity:

To achieve ϵ near-optimal policy, need at most

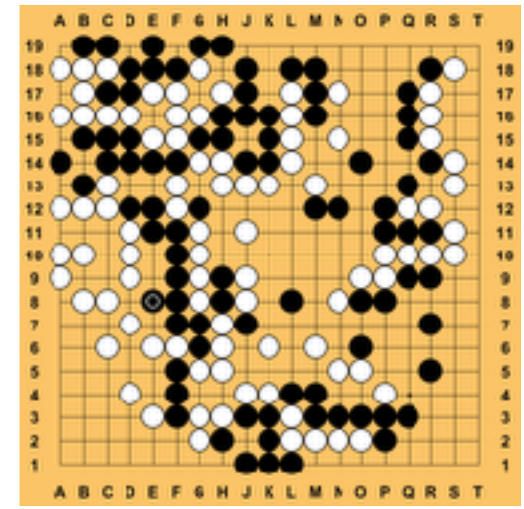
$$\boxed{\mathrm{poly}(\#\text{ of states},}\ \#\text{ of actions}, \text{Horizon}, 1/\epsilon)$$

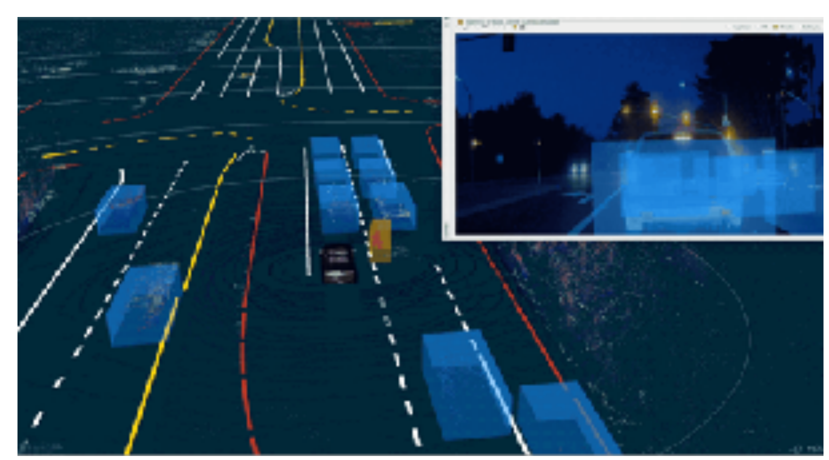
many interactions

[e.g., Kearns & Singh, 02, Dann & Brunskill, 15, Azar et.al, 17]

$\neq$

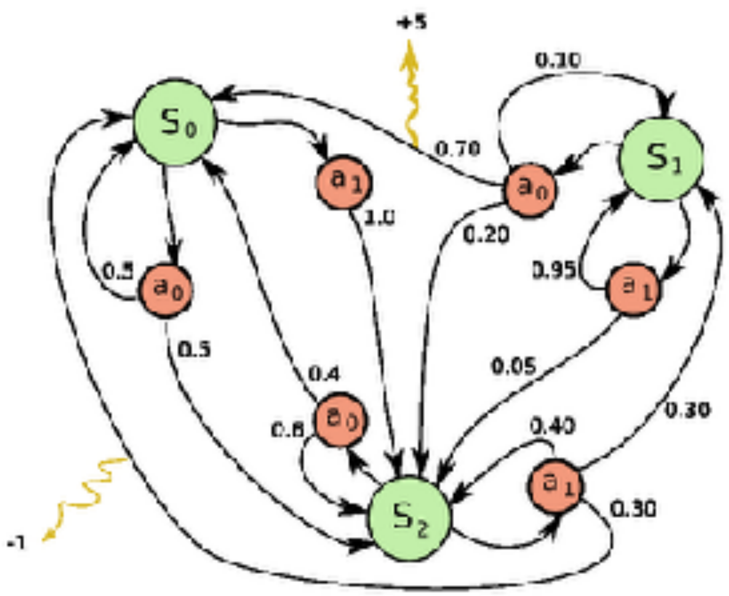
Large-Scale Decision Making Problems





Progress of RL in Theory

Sample Efficiency
in Small Discrete MDPs



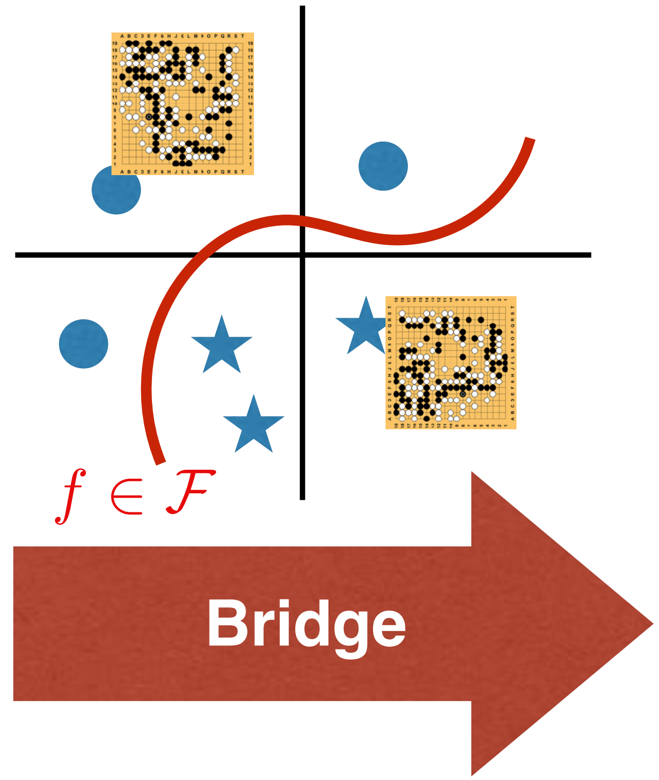
Sample Complexity:

To achieve ϵ near-optimal policy,
need at most

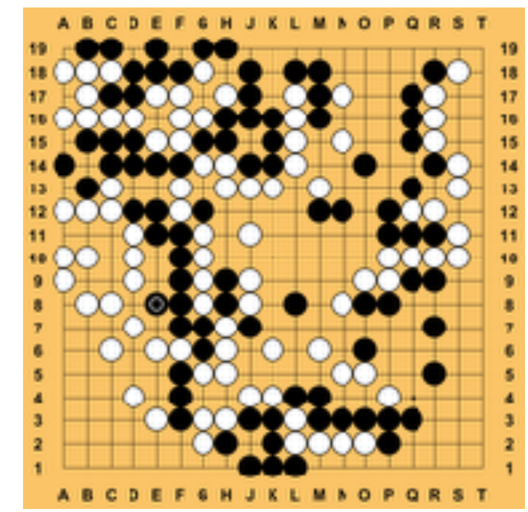
$$\boxed{\text{poly}(\#\text{ of states},}\ \#\text{ of actions, Horizon}, 1/\epsilon)$$

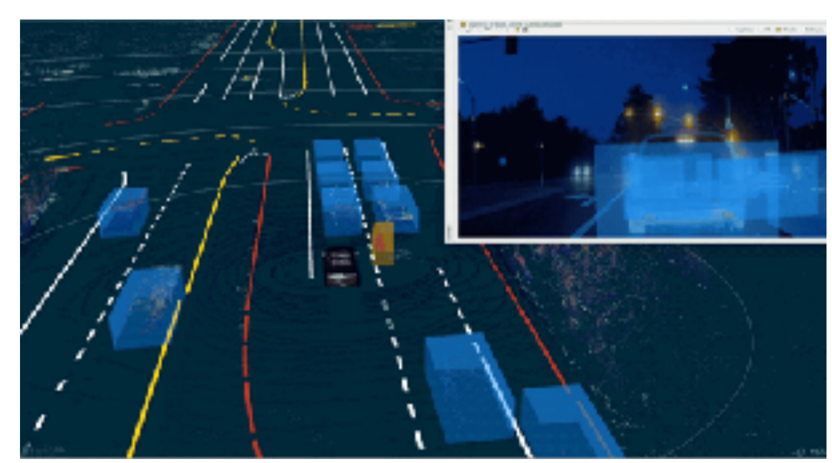
many interactions

[e.g., Kearns & Singh, 02, Dann & Brunskill, 15, Azar et.al, 17]



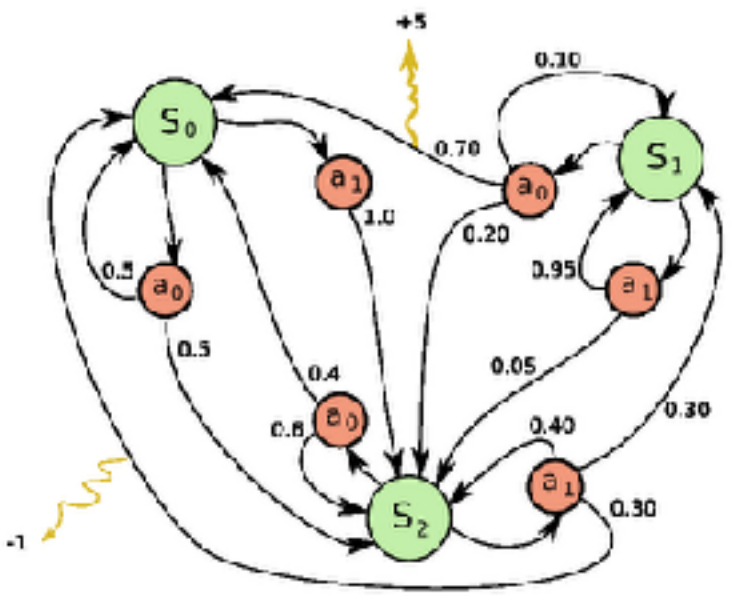
Large-Scale Decision Making Problems





Progress of RL in Theory

Sample Efficiency in Small Discrete MDPs



Sample Complexity:

To achieve ϵ near-optimal policy,

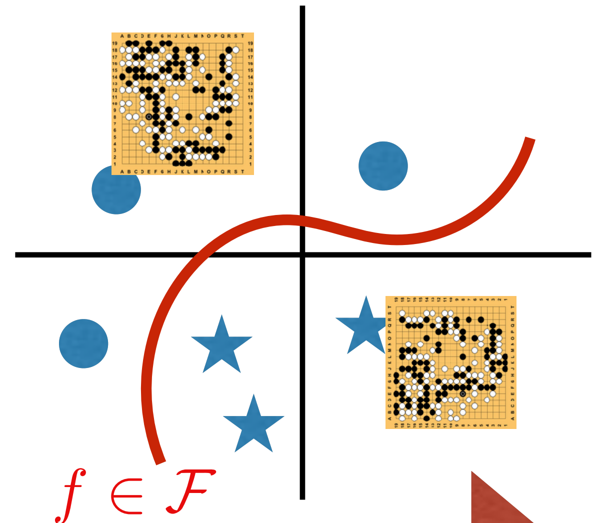
need at most

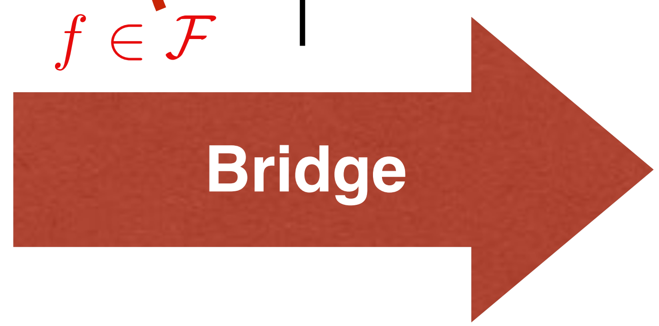
$$\boxed{\text{poly}(\#\text{ of states},}\ \#\text{ of actions, Horizon}, 1/\epsilon)$$

many interactions

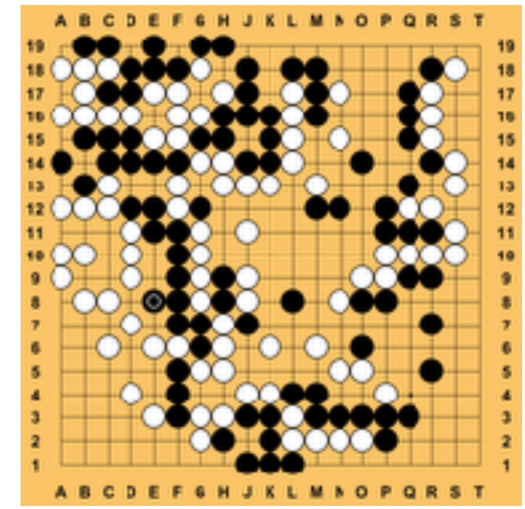
[e.g., Kearns & Singh, 02, Dann & Brunskill, 15, Azar et.al, 17]

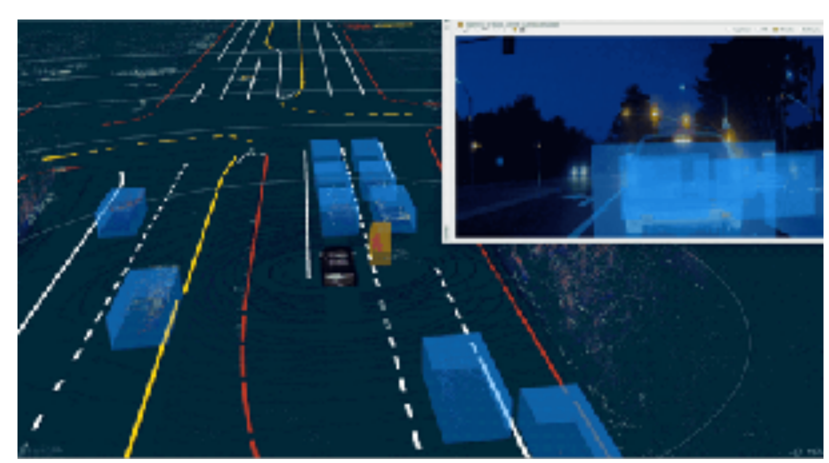
VC-dim





Large-Scale Decision Making Problems





Previous Works on PAC RL w/ General Function Approximation

Previous Works on PAC RL w/ General Function Approximation

Contextual Bandits (horizon=1)

(e.g., Auer et al., 02, Langford & Zhang, 07)

Previous Works on PAC RL w/ General Function Approximation

Contextual Bandits (horizon=1)

(e.g., Auer et al., 02, Langford & Zhang, 07)

Contextual Decision Process

(Krishnamurthy et al.,16, Jiang et al., 17, Dann et al, 18)

Previous Works on PAC RL w/ General Function Approximation

Contextual Bandits (horizon=1)

(e.g., Auer et al., 02, Langford & Zhang, 07)

Contextual Decision Process

(Krishnamurthy et al.,16, Jiang et al., 17, Dann et al, 18)

Model-based vs Model-free

Previous Works on PAC RL w/ General Function Approximation

Contextual Bandits (horizon=1)

(e.g., Auer et al., 02, Langford & Zhang, 07)

Contextual Decision Process

(Krishnamurthy et al.,16, Jiang et al., 17, Dann et al, 18)

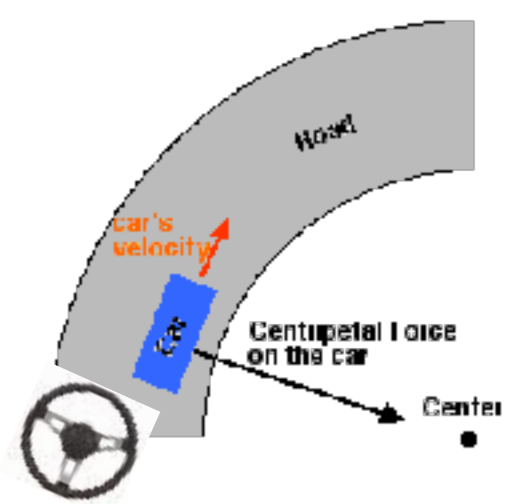
Model-based vs Model-free

A PAC model-based Algorithm

Setup of Model-Based RL

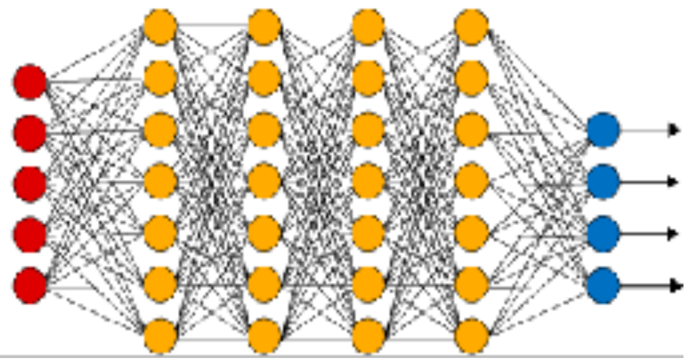
Setup of Model-Based RL

$$x' \sim P^\star(\cdot|x, a)$$

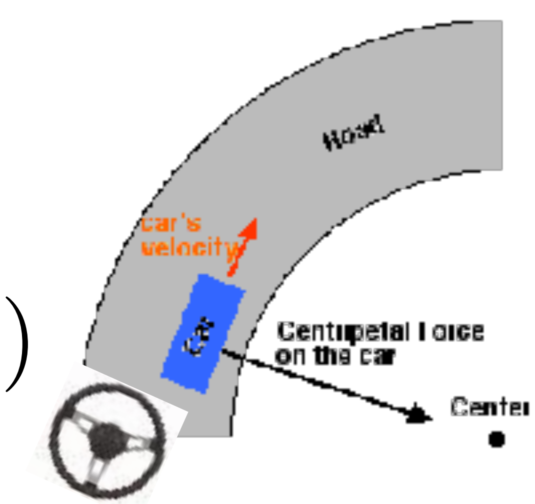


Setup of Model-Based RL

Function Approximators

(x, a)  $P(\cdot|x, a)$

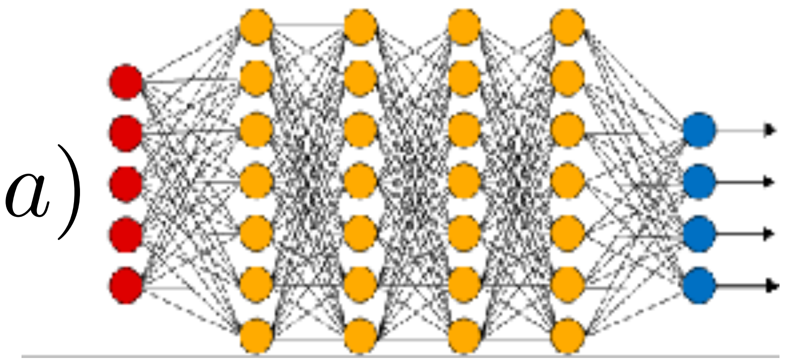
$$\mathcal{P} = \{P : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})\}$$

$x' \sim P^\star(\cdot|x, a)$ 

Setup of Model-Based RL

Function Approximators

(x, a)

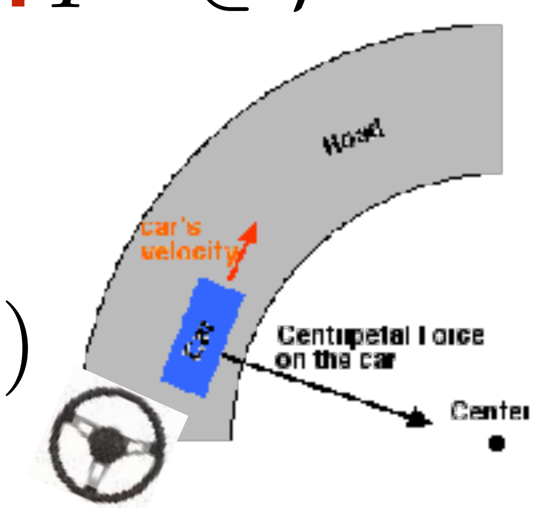


$P(\cdot|x, a)$

$$\mathcal{P} = \{P : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})\}$$

1. Realizability: $P^\star \in \mathcal{P}$

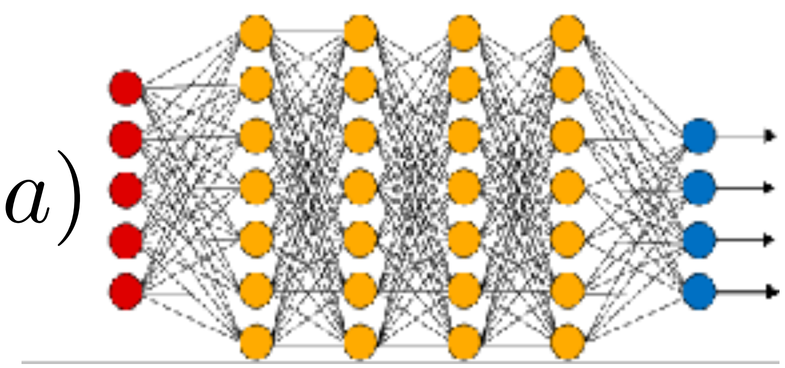
$x' \sim P^\star(\cdot|x, a)$



Setup of Model-Based RL

Function Approximators

(x, a)



$P(\cdot|x, a)$

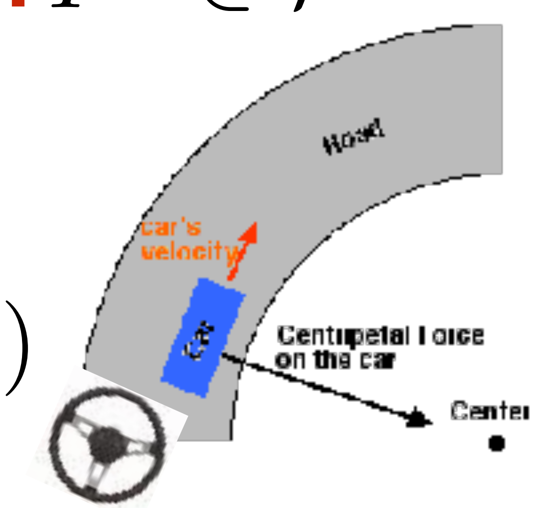
$$\mathcal{P} = \{P : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})\}$$

2. Access to
Optimal Planner (OP)

$$OP(P, r) \Rightarrow \pi_P$$

1. Realizability: $P^\star \in \mathcal{P}$

$$x' \sim P^\star(\cdot|x, a)$$

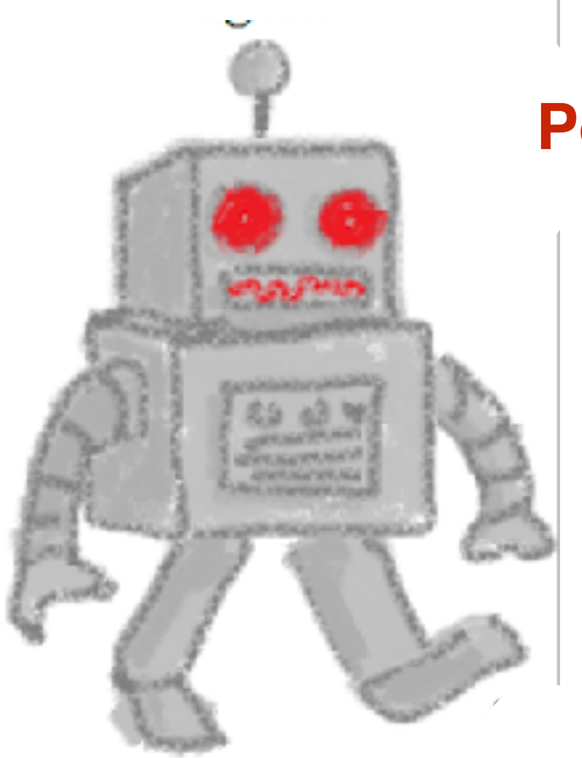


Definition of Model-free RL

Input: $\mathcal{Q} \triangleq \{Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$

Definition of Model-free RL

Input: $\mathcal{Q} \triangleq \{Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$



Policy: determine **action** based on **state**

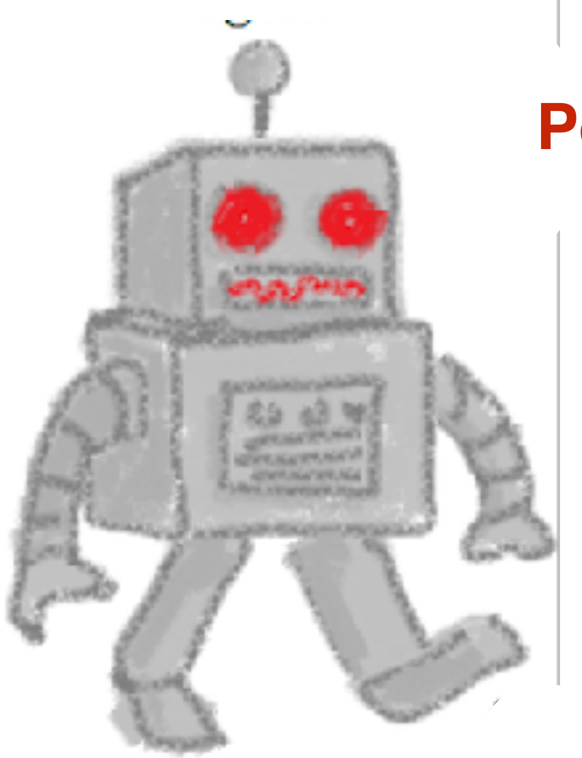
H Steps

Env: reveal next state x

Definition of Model-free RL

Input: $\mathcal{Q} \triangleq \{Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$

Policy: determine **action** based on **state**



H Steps

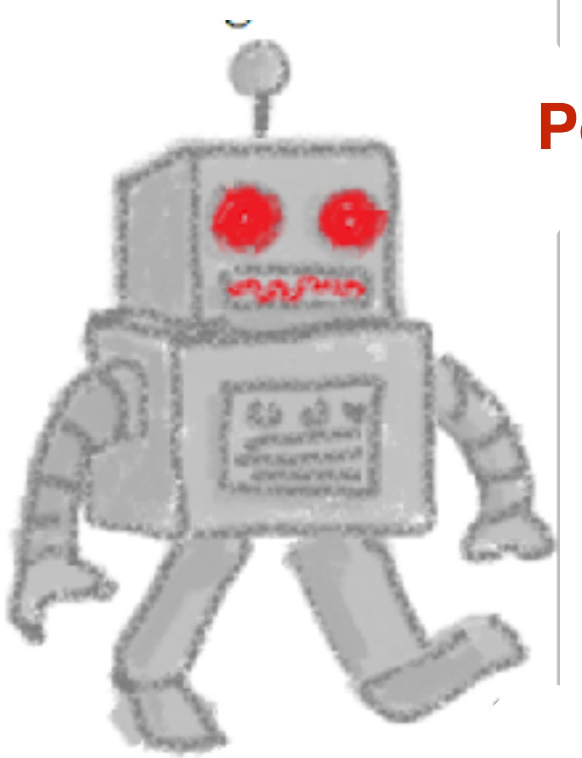
Q-profile:

$$[Q(x,a)]_{Q \in \mathcal{Q}, a \in \mathcal{A}}$$

all possible Q values evaluated at state x

Definition of Model-free RL

Input: $\mathcal{Q} \triangleq \{Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$



Policy: determine **action** based on **state**

H Steps

Q-profile:

$$[Q(x,a)]_{Q \in \mathcal{Q}, a \in \mathcal{A}}$$

all possible Q values evaluated at state x

Efficient Q-learning (Jin et.al, 18)
Fitted Q-Iteration (Ernst et.al., 05)
OLIVE (Jiang et.al, 17)
Policy Gradient (Williams 92)

An Exponential Improvement over Model-free RL

An Exponential Improvement over Model-free RL

There exists MDPs (e.g., Factored MDPs), s.t., to learn near optimal policy,

Model-Based RL:

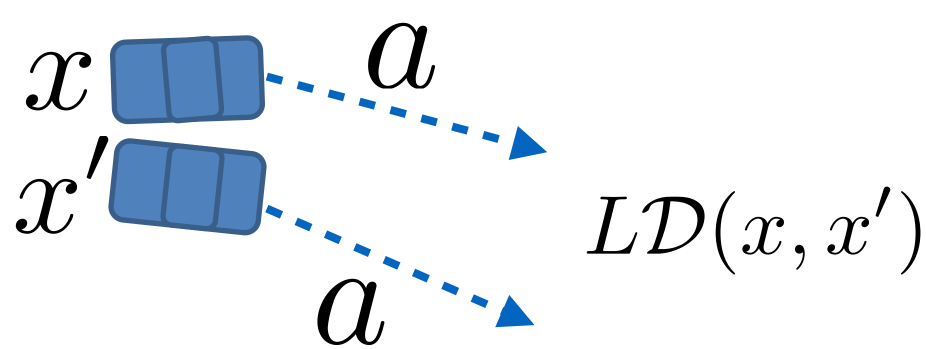
Polynomial Sample Complexity

VS

ANY Model-Free RL:

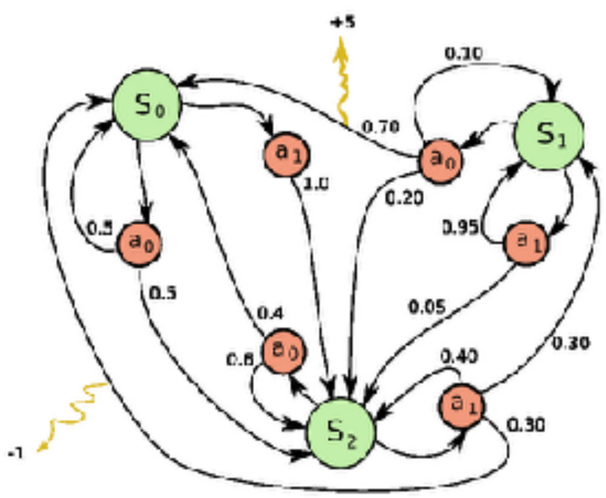
$$\Omega(\exp(H))$$

We have been studying model-based RL, BUT…



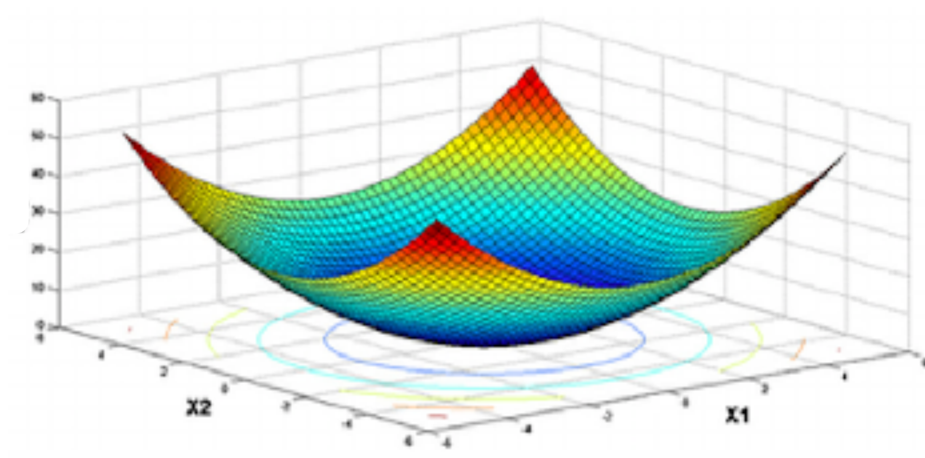
Lipschitz MDPs

[Kearn, Langford, Kakade, 03]



Small Tabular MDP

[Kearn & Singh, 02]



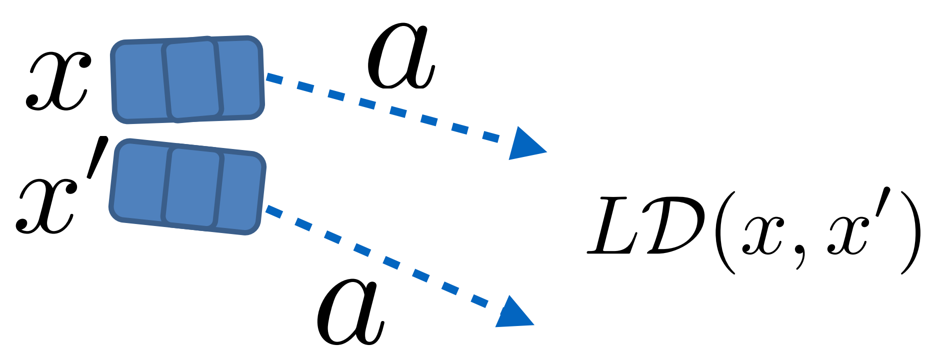
Linear Quadratic Regulator (LQR)

[Dean et.al, 18]

Factored MDPs

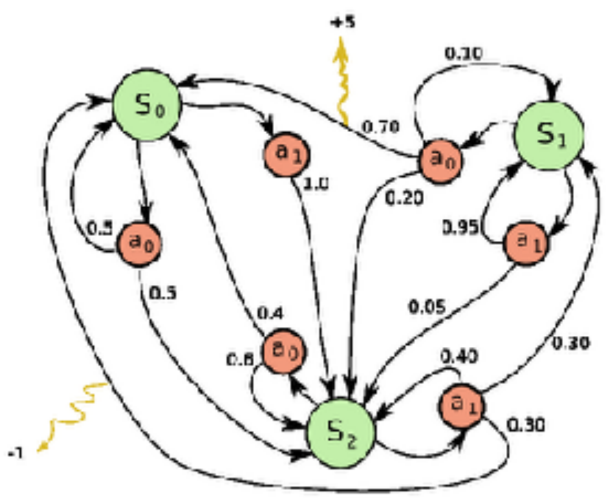
[Guestrin et.al, 03; Osband & Van Roy,13]

We have been studying model-based RL, BUT…



Lipschitz MDPs

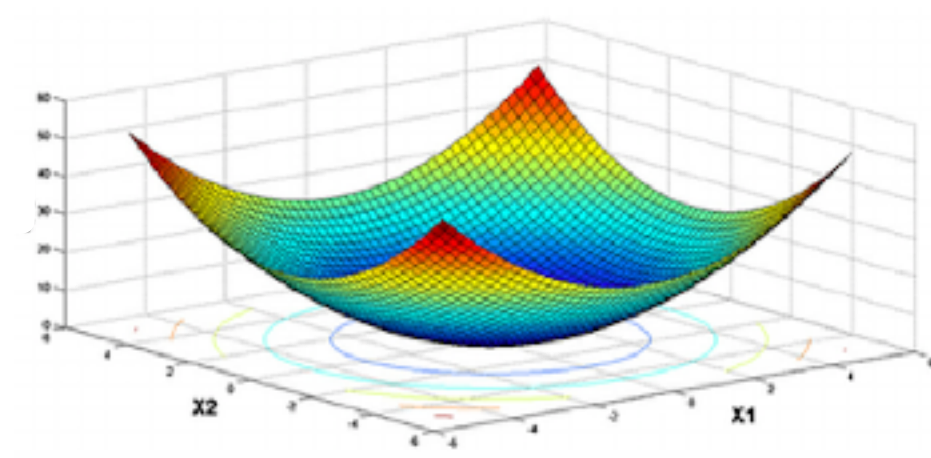
[Kearn, Langford, Kakade, 03]



Small Tabular MDP

[Kearn & Singh, 02]

A Unified Algorithm?



Linear Quadratic Regulator (LQR)

[Dean et.al, 18]

Factored MDPs

[Guestrin et.al, 03; Osband & Van Roy,13]

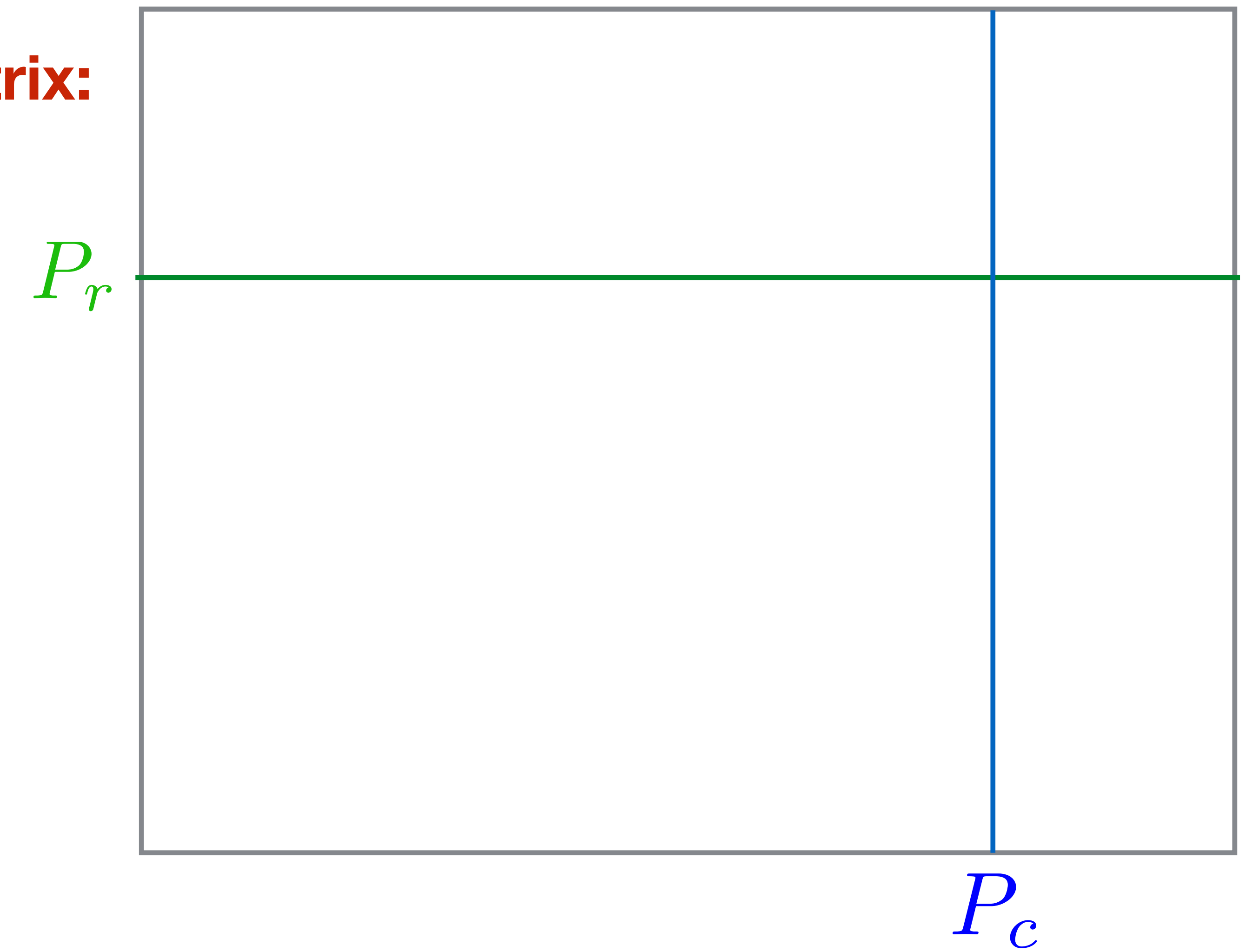
A Unified Measure—*Witness Rank*

Misfit Matrix:

$$\in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{P}|}$$

A Unified Measure—*Witness Rank*

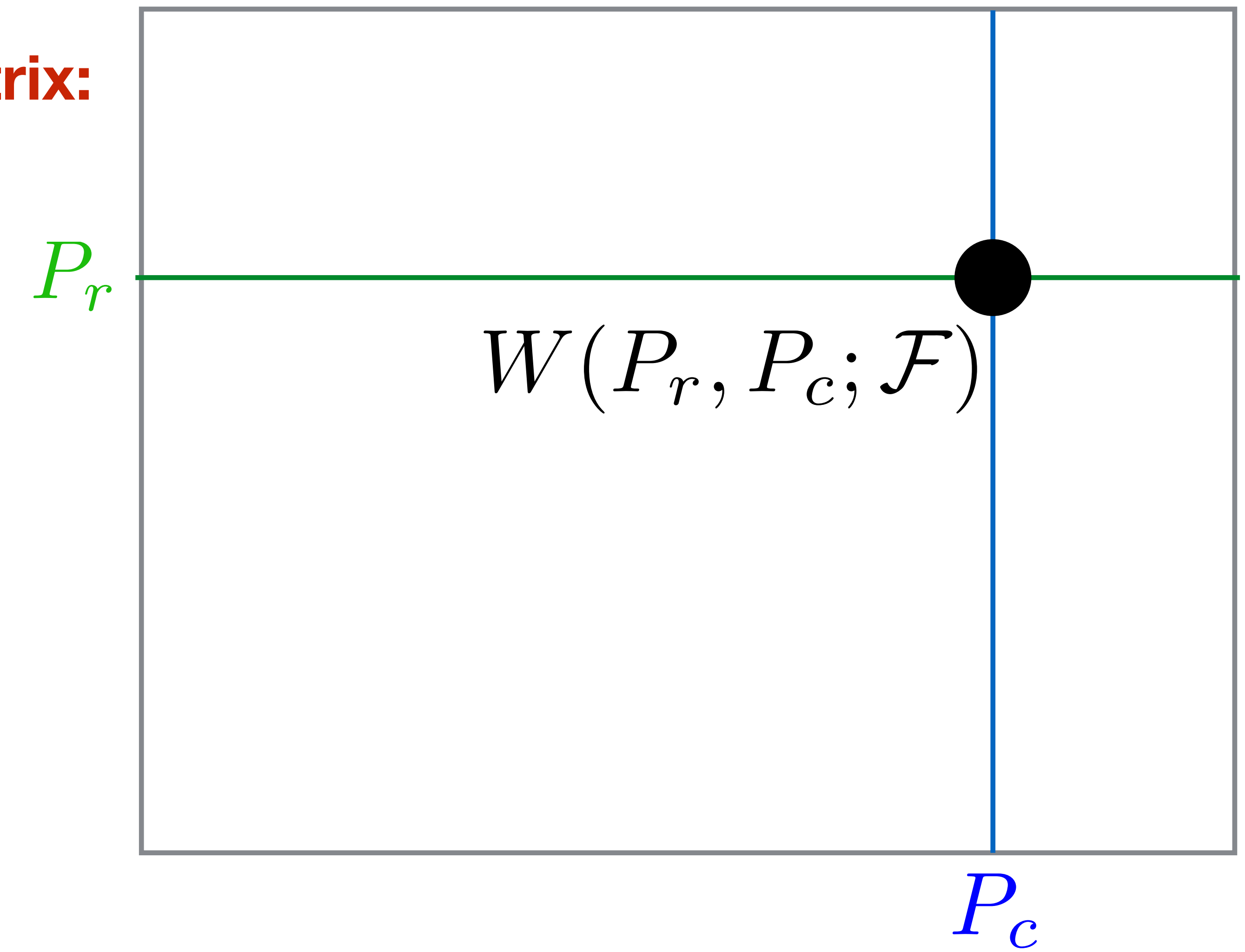
Misfit Matrix:



$$\in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{P}|}$$

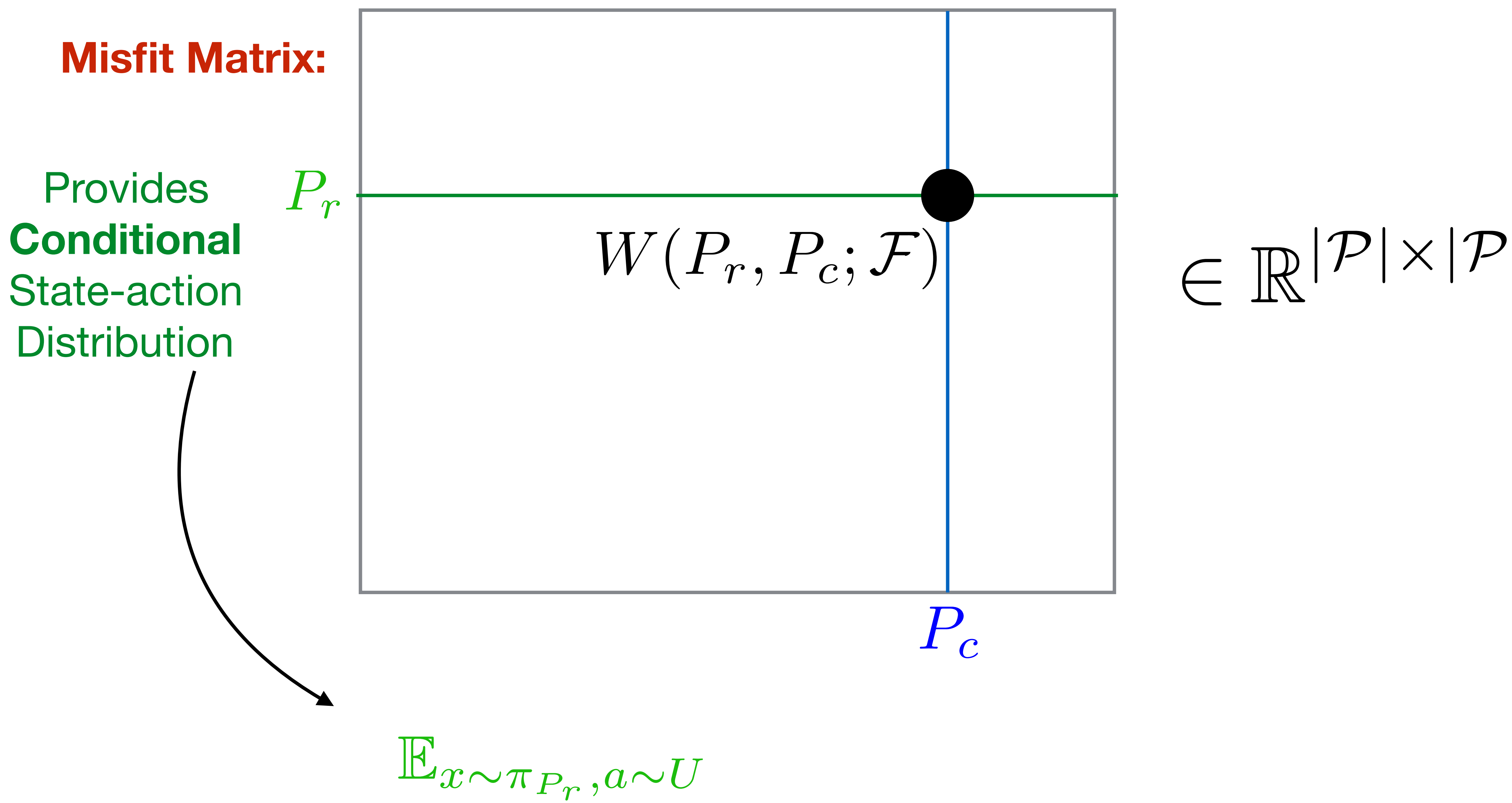
A Unified Measure—*Witness Rank*

Misfit Matrix:



$$\in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{P}|}$$

A Unified Measure—*Witness Rank*

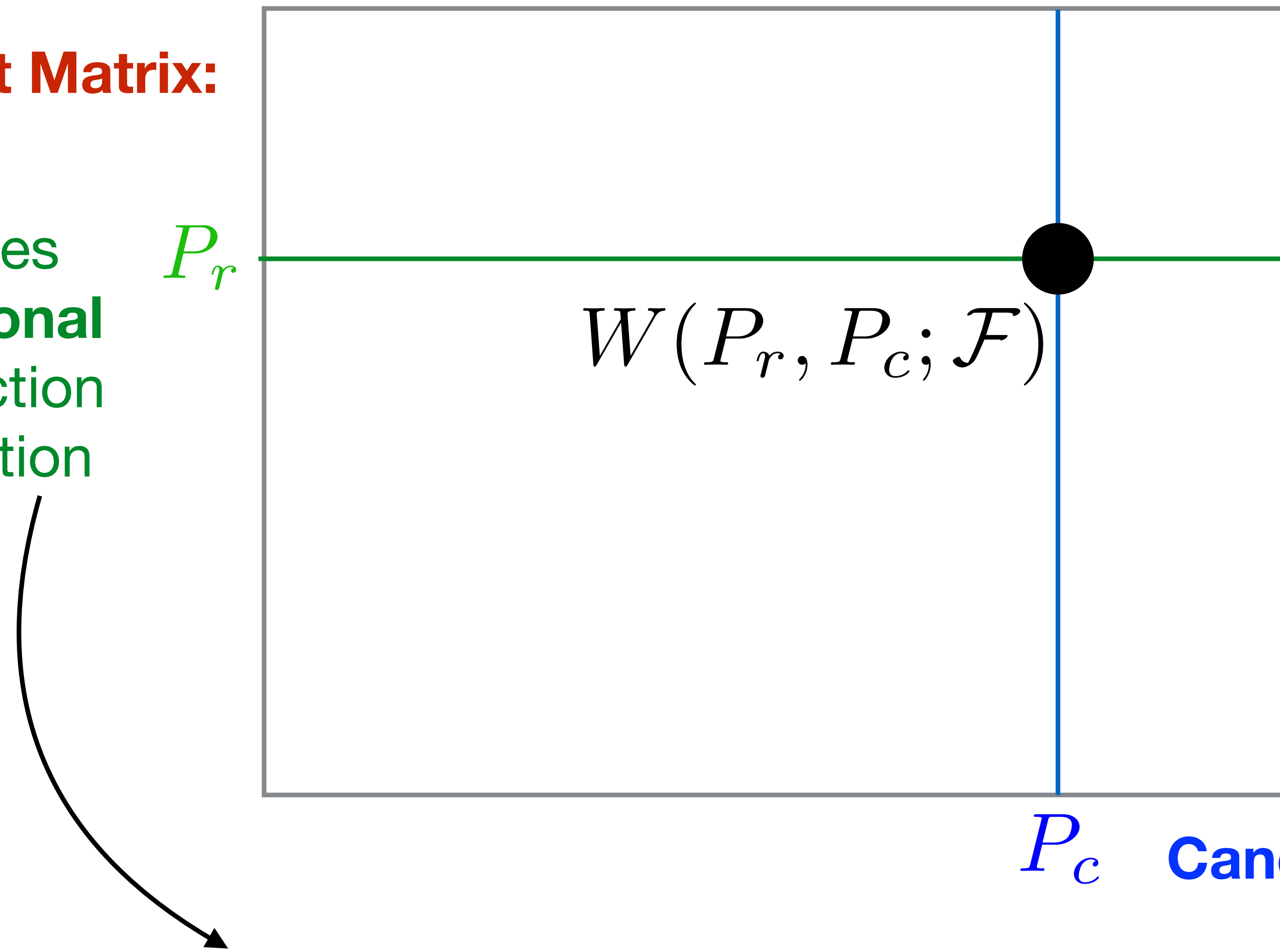


A Unified Measure—*Witness Rank*

Misfit Matrix:

Provides **Conditional** State-action Distribution

P_r



$\in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{P}|}$

P_c **Candidate**

$$\mathbb{E}_{x\sim\pi_{P_r}, a\sim U}\left[\mathbb{E}_{x'\sim P_c} f(x,a,x') - \mathbb{E}_{x'\sim P^\star} f(x,a,x')\right]$$

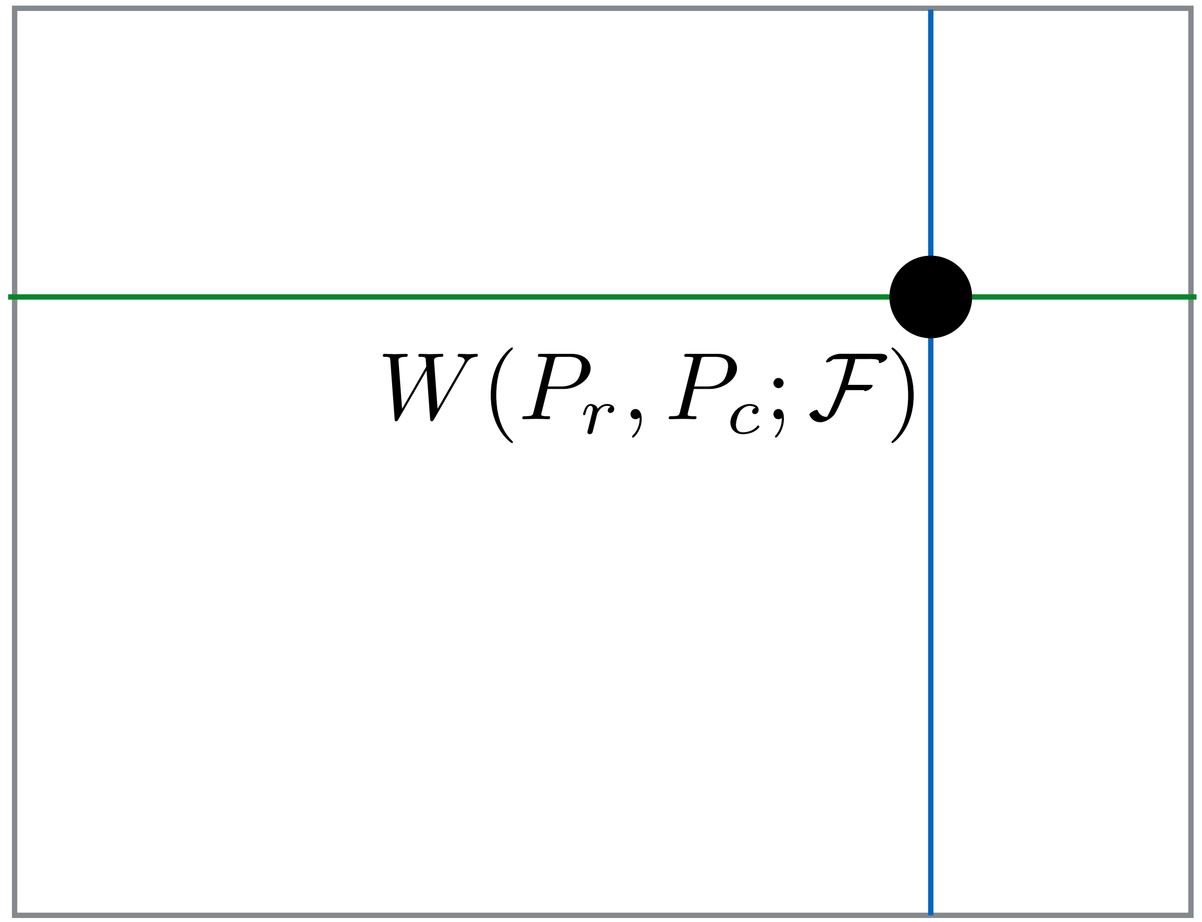
Imaginary Real

A Unified Measure—*Witness Rank*

Misfit Matrix:

Provides **Conditional** State-action Distribution

P_r



$\in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{P}|}$

P_c **Candidate**

$$W(P_r, P_c; \mathcal{F}) = \max_{f \in \mathcal{F}} \mathbb{E}_{x \sim \pi_{P_r}, a \sim U} \left[\mathbb{E}_{x' \sim P_c} f(x, a, x') - \mathbb{E}_{x' \sim P^\star} f(x, a, x') \right]$$

Witness functions (aka discriminators)

Imaginary

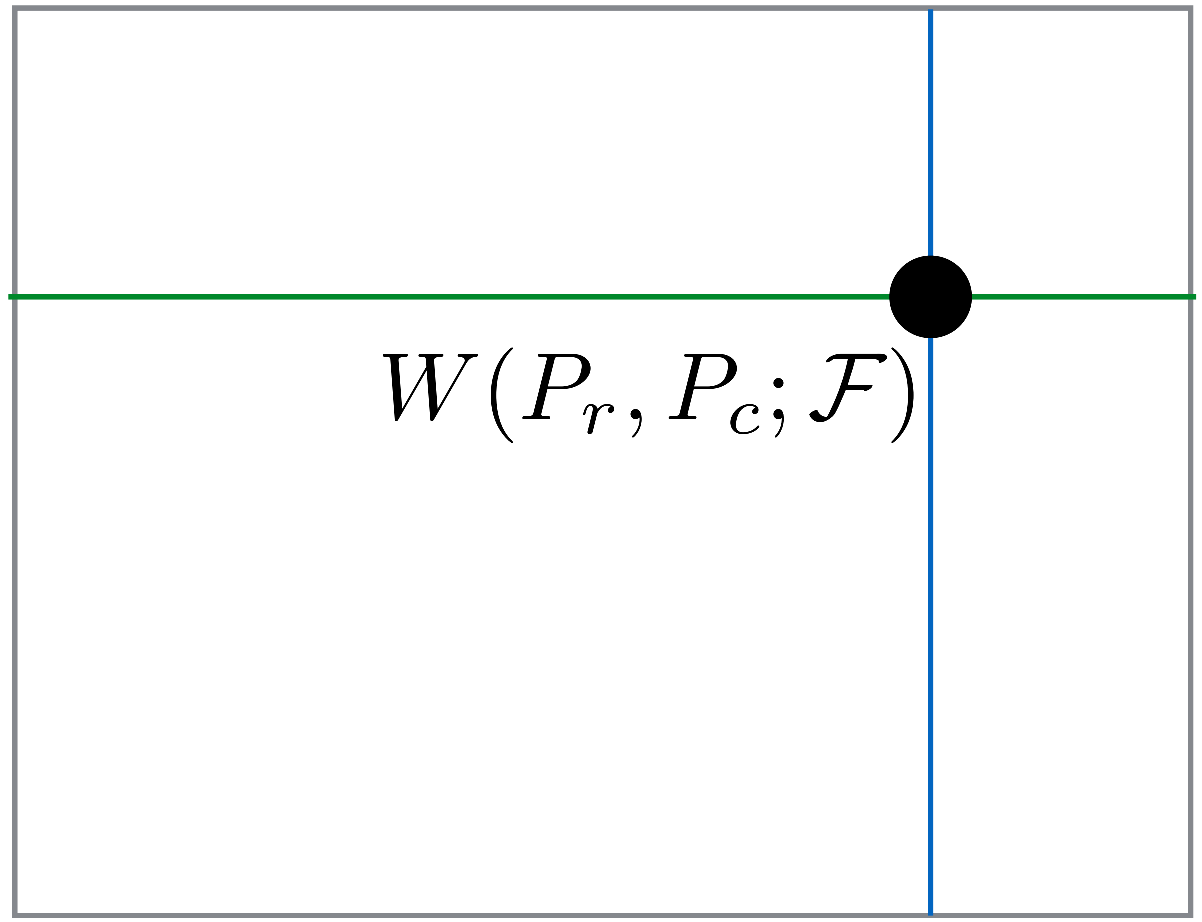
Real

A Unified Measure—*Witness Rank*

Misfit Matrix:

Provides
Conditional
State-action
Distribution

P_r



$\in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{P}|}$

P_c **Candidate**

$$W(P_r, P_c; \mathcal{F}) = \max_{f \in \mathcal{F}} \mathbb{E}_{x \sim \pi_{P_r}, a \sim U} \left[\mathbb{E}_{x' \sim P_c} f(x, a, x') - \mathbb{E}_{x' \sim P^\star} f(x, a, x') \right]$$

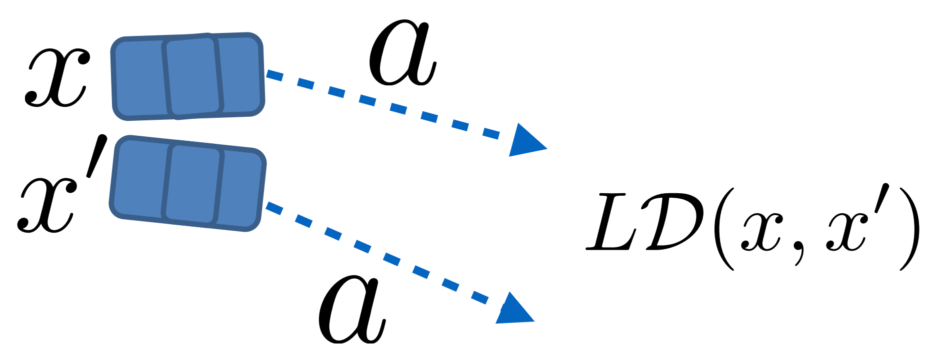
Witness functions
(aka discriminators)

Imaginary

Real

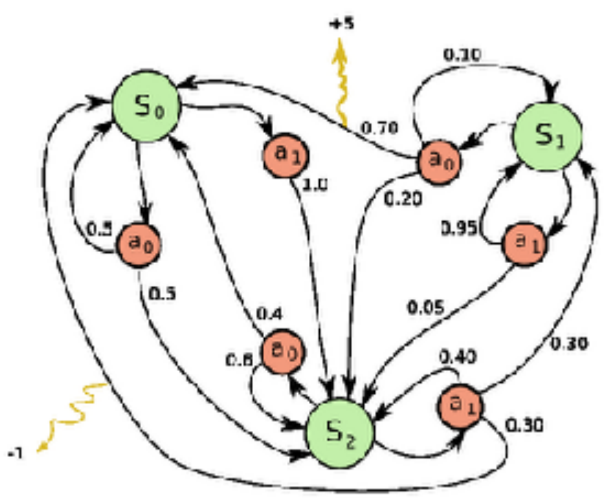
Witness Rank $\triangleq$ rank of this misfit matrix

Capture Complexities of Existing RL problems



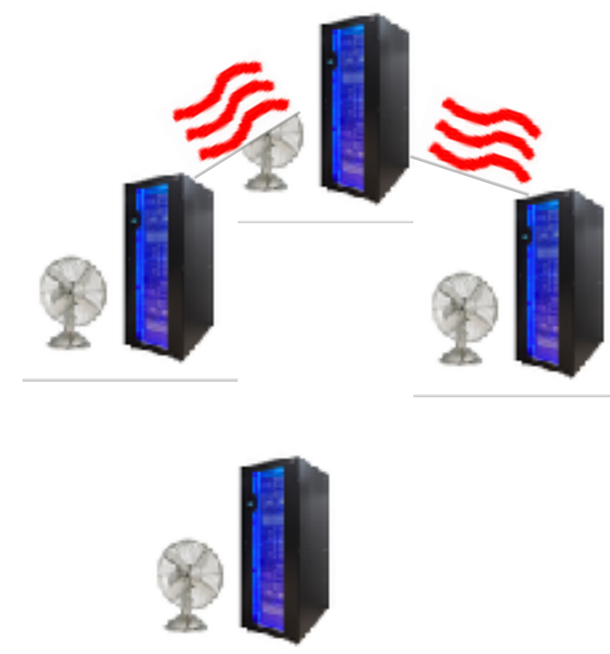
Lipschitz Continuous MDPs
[Kearn, Langford, Kakade, 03]

Rank <= Covering number of state space



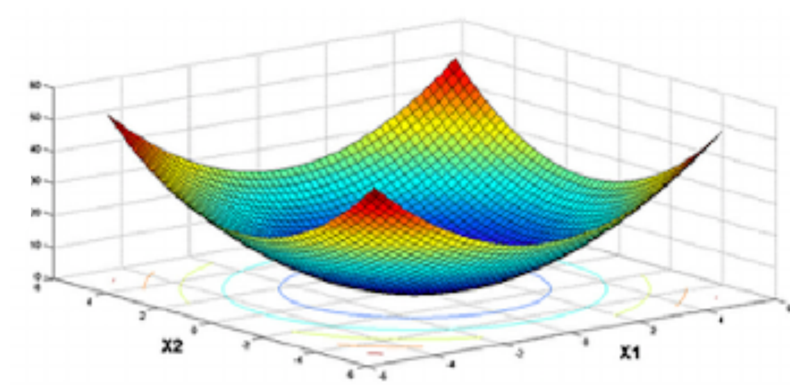
Small Discrete MDP

Rank <= # of state



Factored MDPs
[Guestrin et.al, 03; Osband & Van Roy,13]

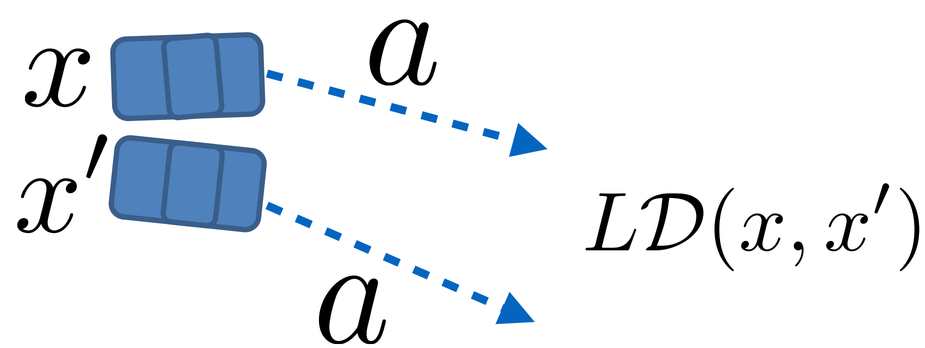
Rank <= exp(in-degree)



LQR

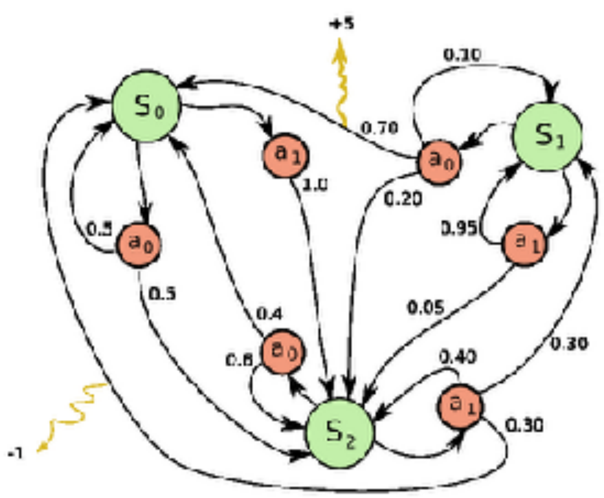
Rank <= O(d^2)

Capture Complexities of Existing RL problems



Lipschitz Continuous MDPs
[Kearn, Langford, Kakade, 03]

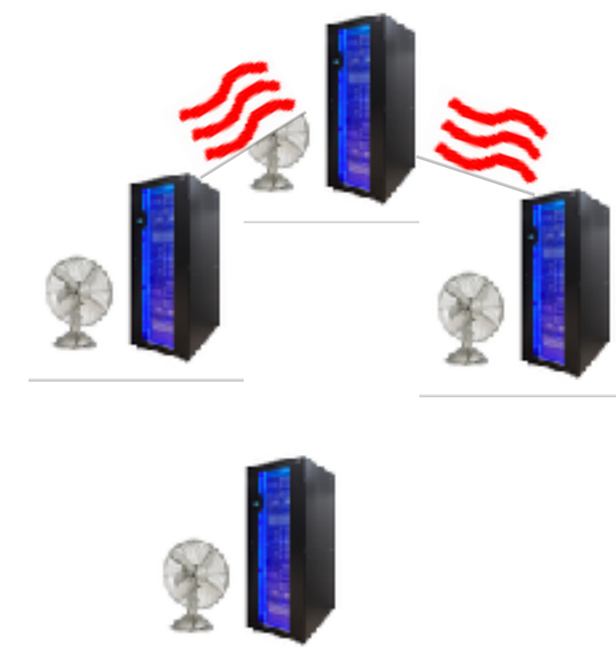
Rank <= Covering number of state space



Small Discrete MDP

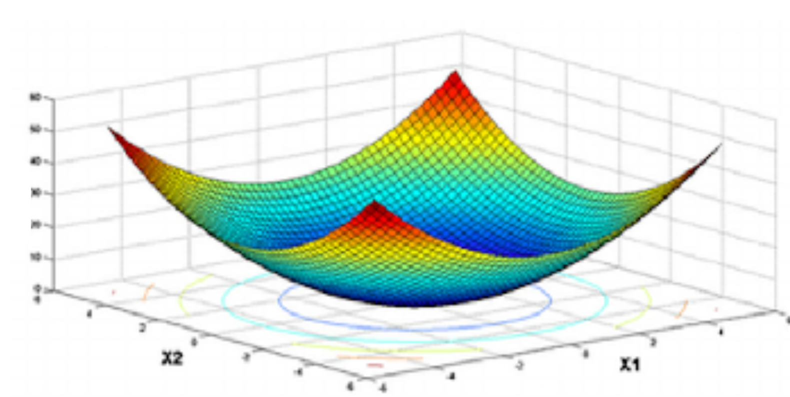
Rank <= # of state

A Unified Algorithm!



Factored MDPs
[Guestrin et.al, 03; Osband & Van Roy,13]

Rank <= exp(in-degree)



LQR

Rank <= O(d^2)

Sample Complexity

To achieve ϵ near-optimal policy (w/ high probability)

Sample Complexity

To achieve ϵ near-optimal policy (w/ high probability)

$$\tilde{O}\Big(\frac{H^3 R^2 |\mathcal{A}|}{\epsilon^2} \log\Big(\frac{|\mathcal{F}||\mathcal{P}|}{\delta}\Big)\Big)$$

Sample Complexity

To achieve ϵ near-optimal policy (w/ high probability)

Witness Rank

$$\tilde{O}\left(\frac{H^3 R^2 |\mathcal{A}|}{\epsilon^2} \log\left(\frac{|\mathcal{F}||\mathcal{P}|}{\delta}\right)\right)$$

Sample Complexity

To achieve ϵ near-optimal policy (w/ high probability)

Witness Rank

$$\tilde{O}\left(\frac{H^3 R^2 |\mathcal{A}|}{\epsilon^2} \log\left(\frac{|\mathcal{F}||\mathcal{P}|}{\delta}\right)\right)$$

Complexities of
model class & discriminator class

Poly Dependency on # of states

Take-home Messages

Take-home Messages

Model-based RL could be exponentially more sample efficient than model-free ones

Take-home Messages

Model-based RL could be exponentially more sample efficient than model-free ones

Sample efficiency is possible when Witness Rank is small