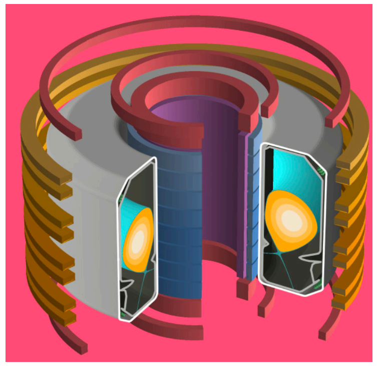
Rethinking the theoretical foundation of reinforcement learning

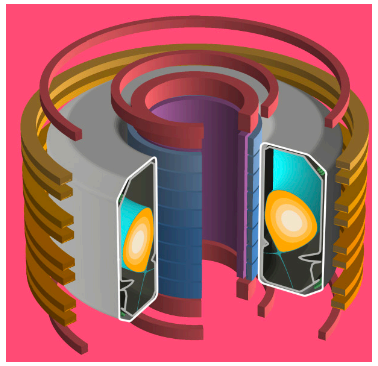
Nan Jiang
University of Illinois at Urbana-Champaign
Feb 24, 2024

@IISc Workshop
Reinforcement Learning: Recent Advances and Challenges Ahead


Google DeepMind
Challenge Match
LEE SEDOL
00:28:28


The international journal of science / 10 February 2022
nature
DRIVING FORCE
AI algorithm outcompetes human champions in Gran Turismo racing game



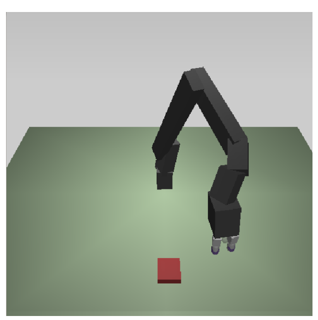
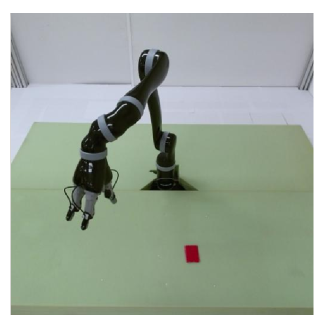


Key ingredient: **simulator**

- Unlimited data
- Decision w/o real consequences
- Can easily evaluate new strategy

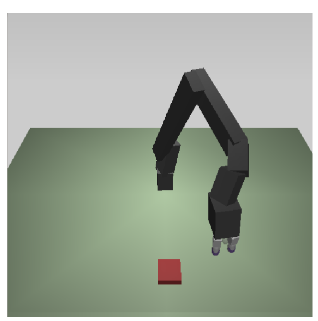
Key ingredient: **simulator**

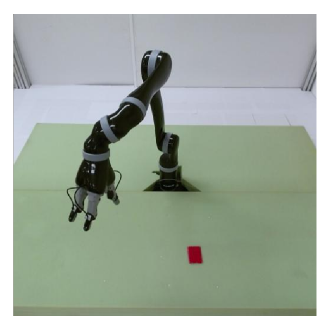
- Unlimited data X
- Decision w/o real consequences X
- Can easily evaluate new strategy X

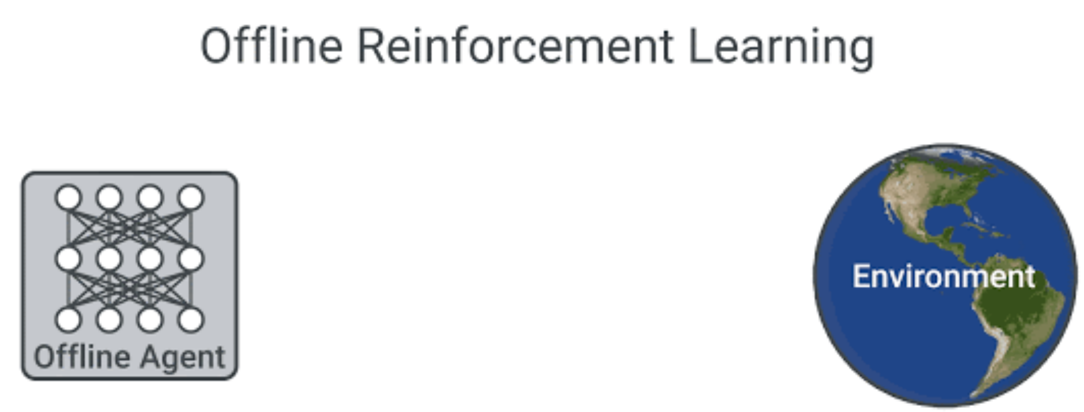




Key ingredient: **simulator**

- Unlimited data X
- Decision w/o real consequences X
- Can easily evaluate new strategy X



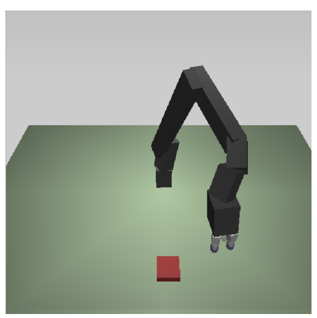


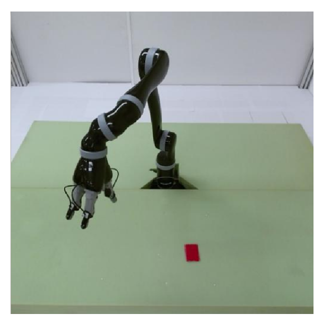


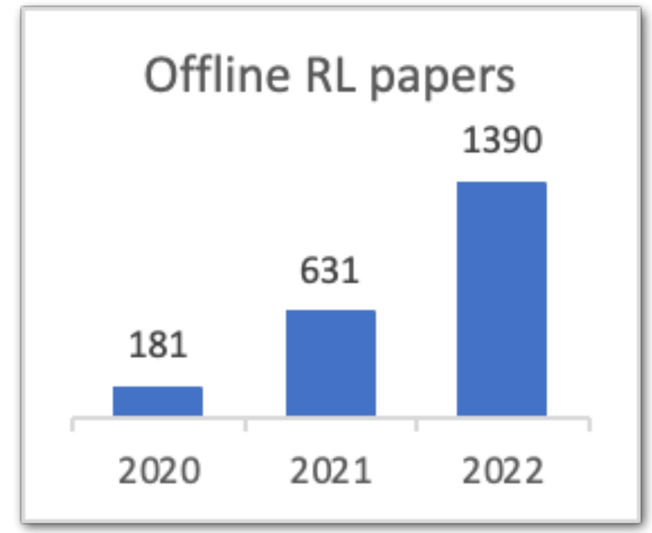
Key ingredient: **simulator**

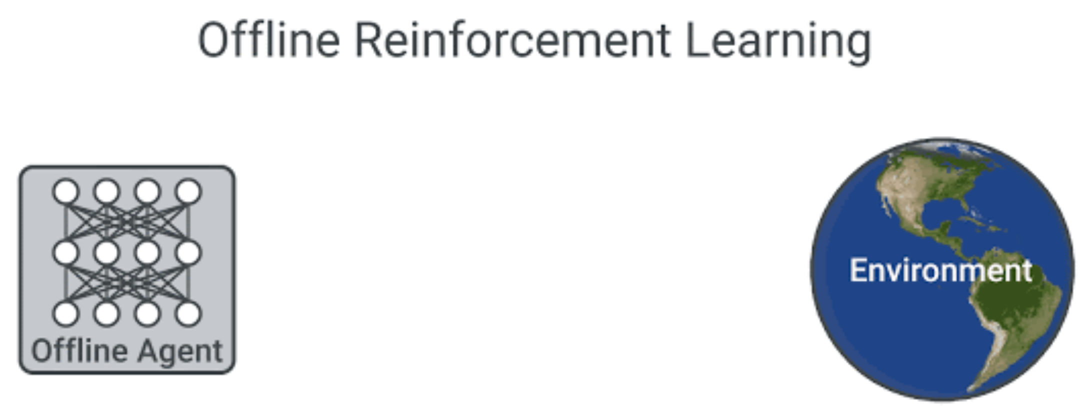
- Unlimited data X
- Decision w/o real consequences X
- Can easily evaluate new strategy X

- (Offline) RL in real life









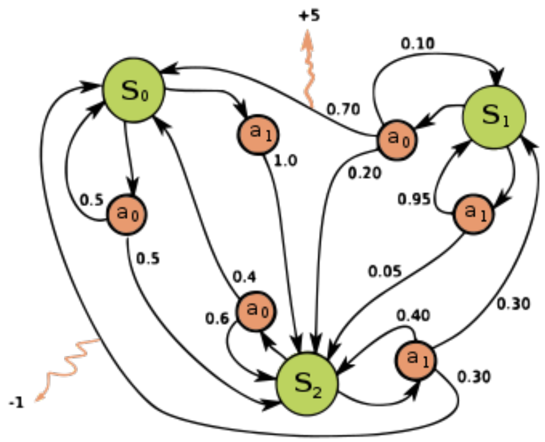
Key ingredient: **simulator**

- Unlimited data X
- Decision w/o real consequences X
- Can easily evaluate new strategy X

Why are we not seeing (offline) RL deployed everywhere already?

- (Offline) RL in real life
- Role of theory in modern RL

~2015



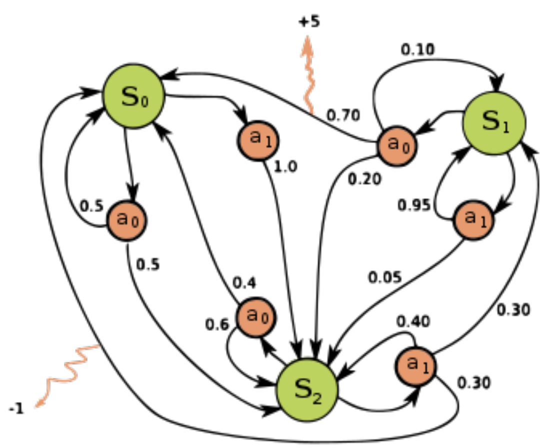
$\sqrt{HSAT}$ *regret,*
SAH^2/ϵ^2 *sample complexity, …*

~2000

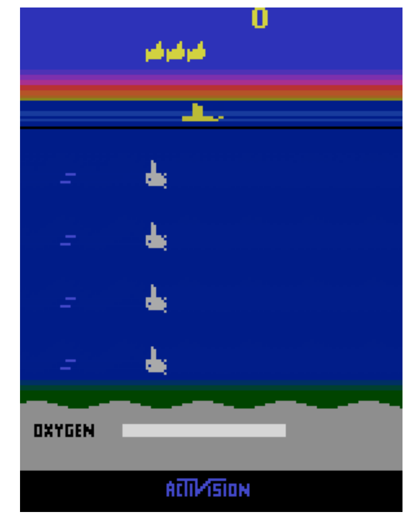
- (Offline) RL in real life
- Role of theory in modern RL

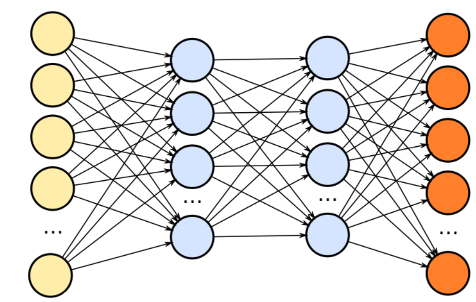
~2015

~2000



$\sqrt{HSAT}$ *regret,*
SAH^2/ϵ^2 *sample complexity, …*



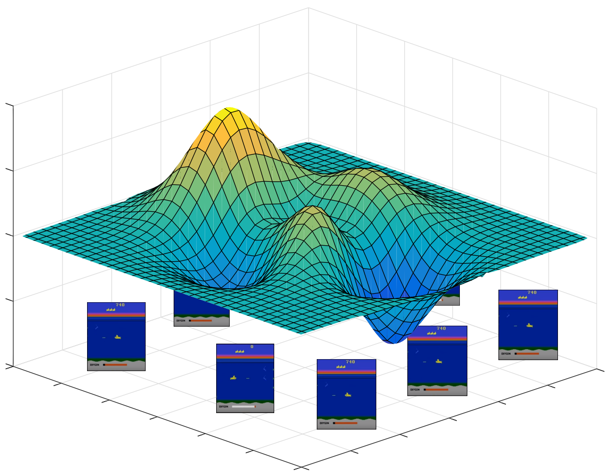


Empirical: *Atari, Mujoco, OpenAI Gym, target network, architecture, …*

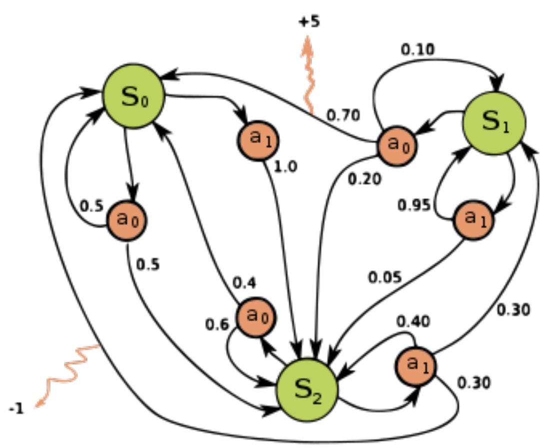
- (Offline) RL in real life
- Role of theory in modern RL

~2015

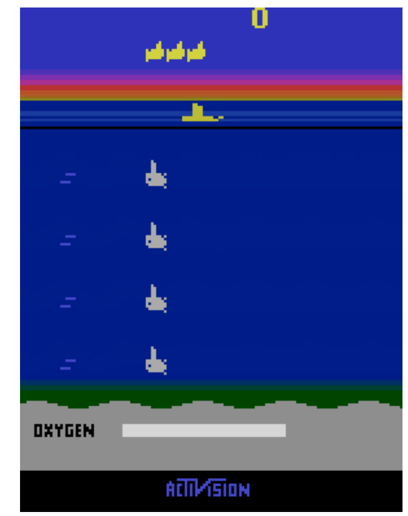
~2000

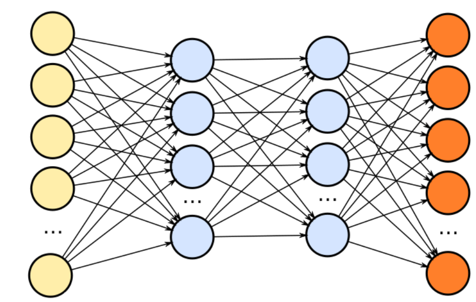


Bellman rank, Eluder dimension, Concentrability, …



$\sqrt{HSAT}$ *regret, SAH^2/ϵ^2 sample complexity, …*





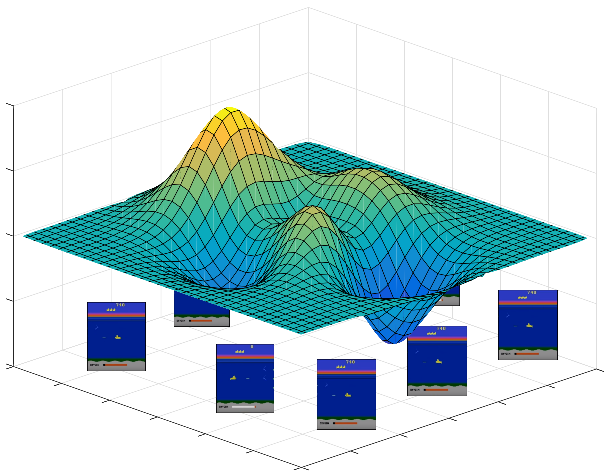
***Empirical:** Atari, Mujoco, OpenAI Gym, target network, architecture, …*

- (Offline) RL in real life
- Role of theory in modern RL

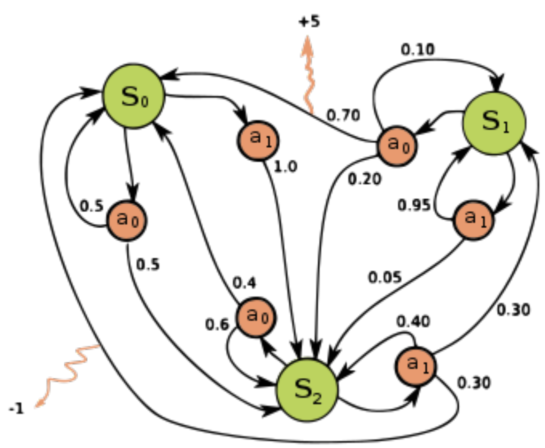
~2015

*

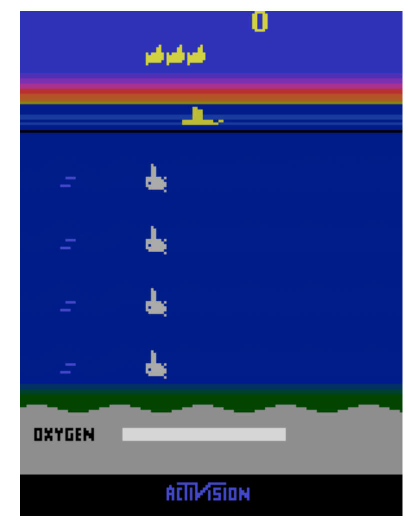
~2000

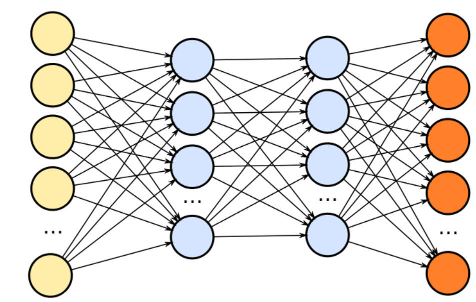


Bellman rank, Eluder dimension, Concentrability, …



$\sqrt{HSAT}$ *regret,* SAH^2/ϵ^2 *sample complexity, …*

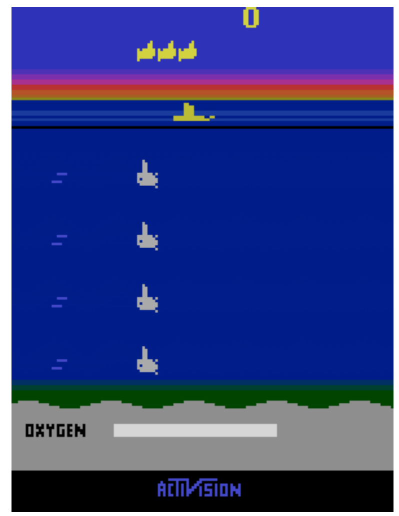


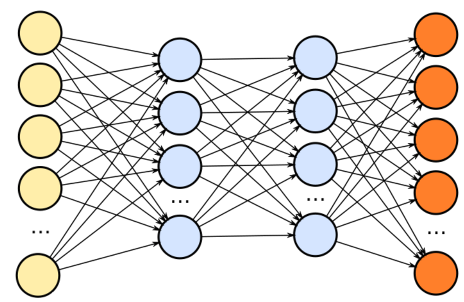


***Empirical:** Atari, Mujoco, OpenAI Gym, target network, architecture, …*

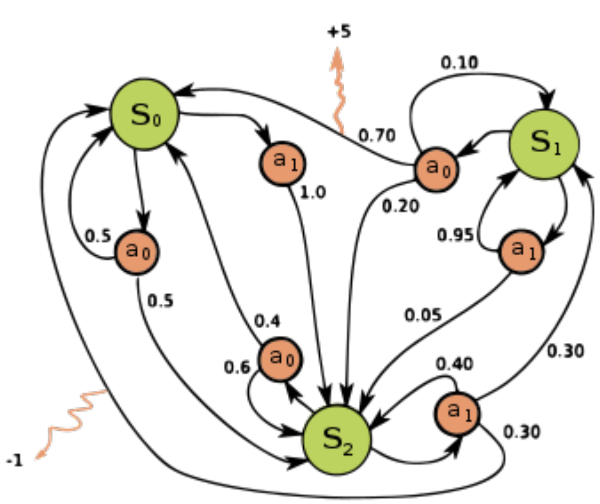
* finite-sample analysis of ADP & MCTS 00~10

- (Offline) RL in real life
- Role of theory in modern RL
- Theoretical foundation

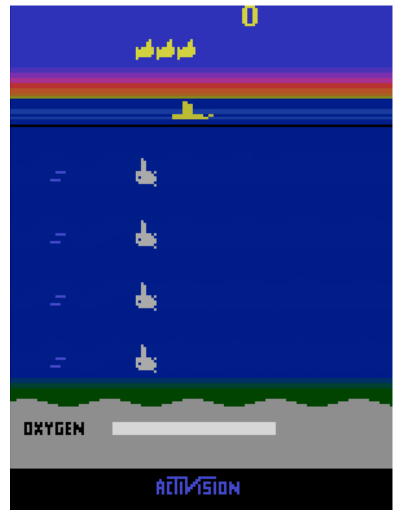


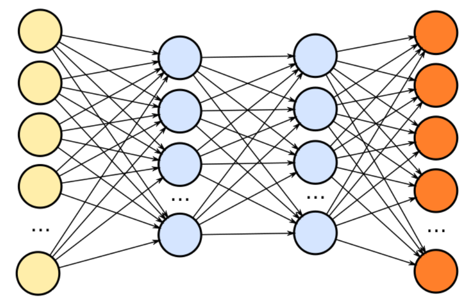


simplify



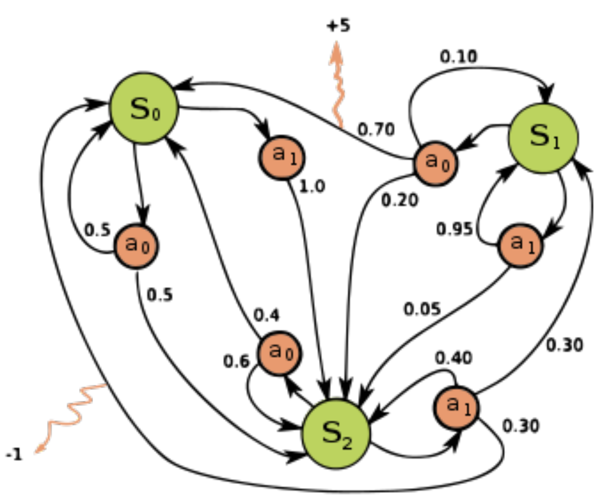
- (Offline) RL in real life
- Role of theory in modern RL
- Theoretical foundation





simplify

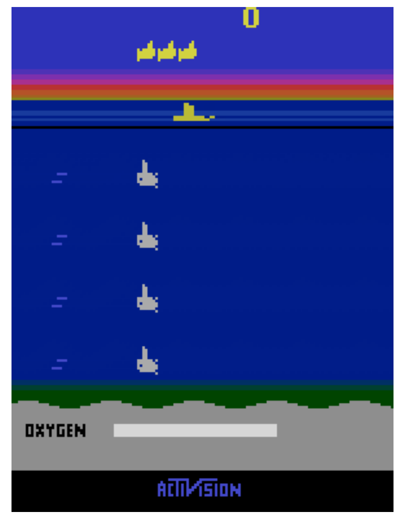
extend

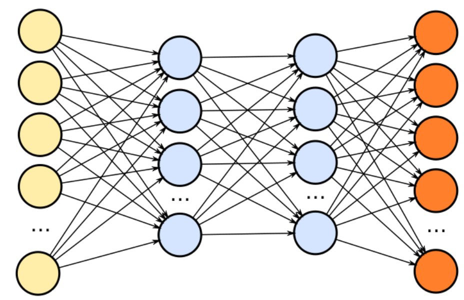


	$Q^*(s, a)$
(s_1, a_1)	$\dots$
(s_1, a_2)	$\dots$
(s_2, a_1)	$\dots$

"tabular" RL

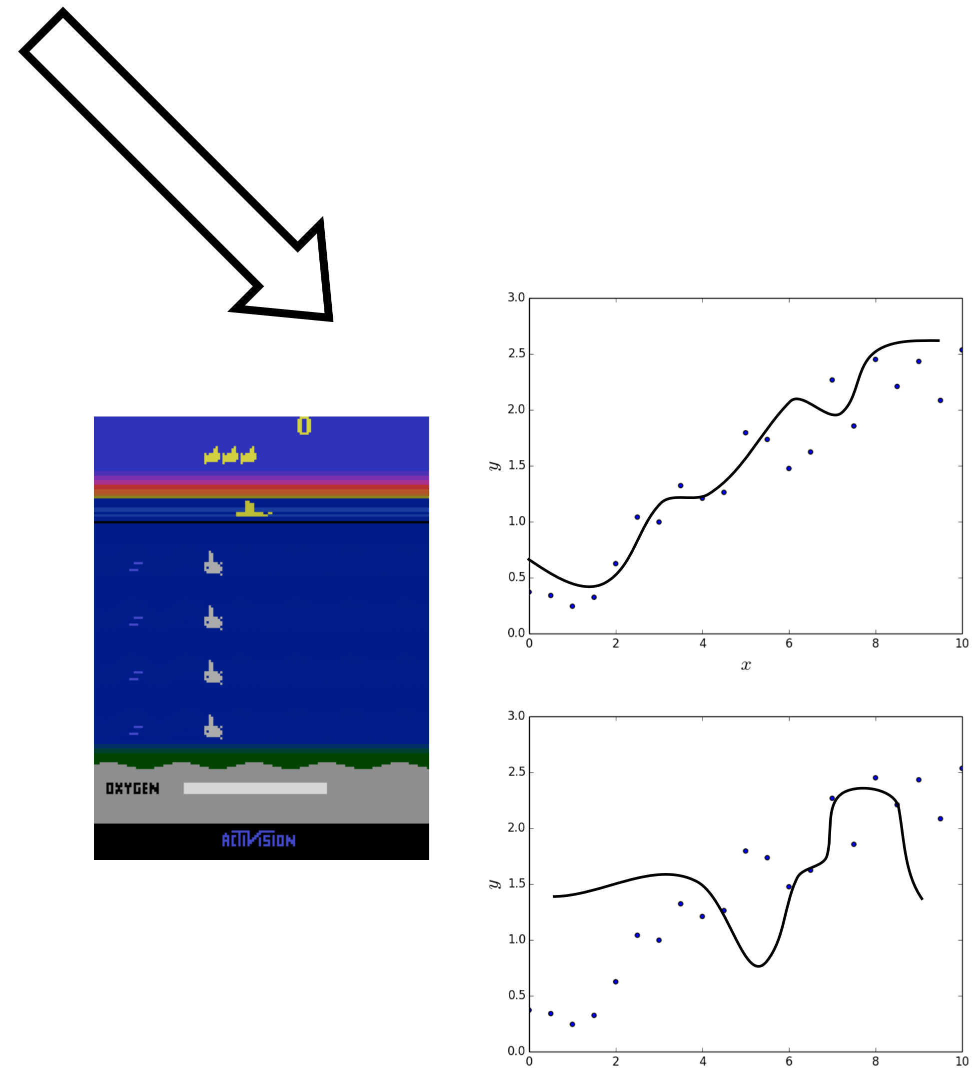
- (Offline) RL in real life
- Role of theory in modern RL
- Theoretical foundation

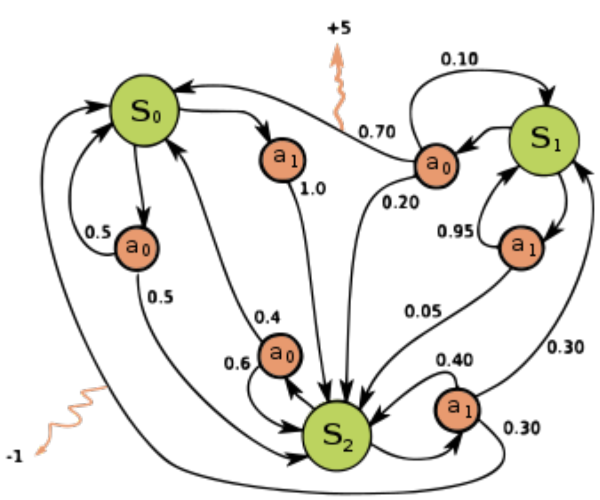




simplify

extend





	$Q^*(s, a)$
(s_1, a_1)	...
(s_1, a_2)	...
(s_2, a_1)	...

“tabular” RL

- (Offline) RL in real life
- Role of theory in modern RL
- Theoretical foundation

- (Offline) RL in real life
- Role of theory in modern RL
- Theoretical foundation

modern RL ⇒ real life

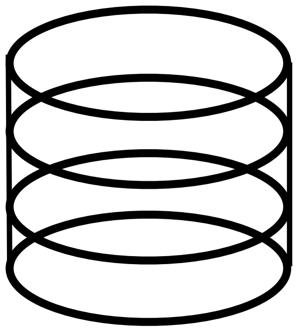
Rethinking
theoretical foundation

- (Offline) RL in real life
- Role of theory in modern RL
- Theoretical foundation

modern RL  real life

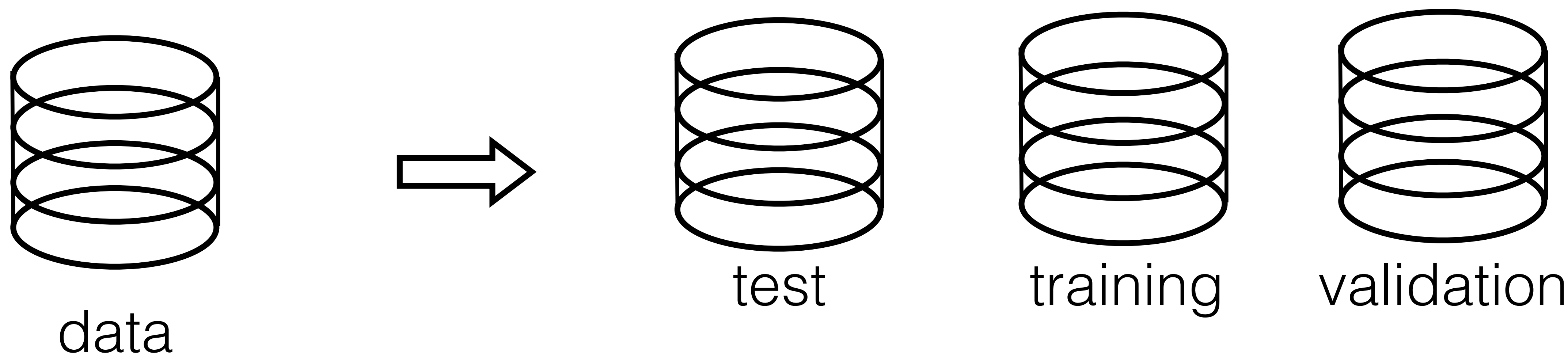
Rethinking
theoretical foundation

Supervised learning pipeline

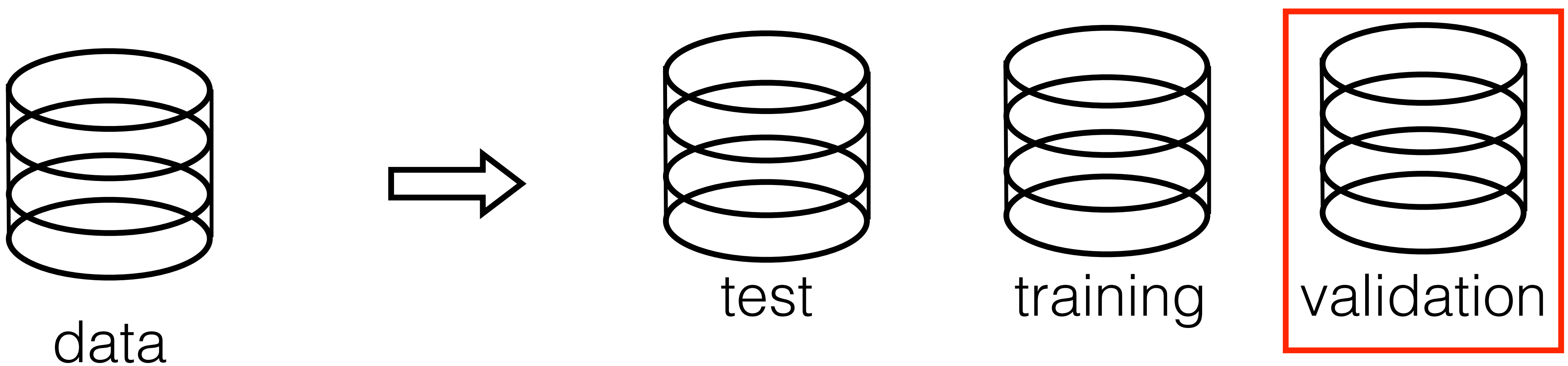


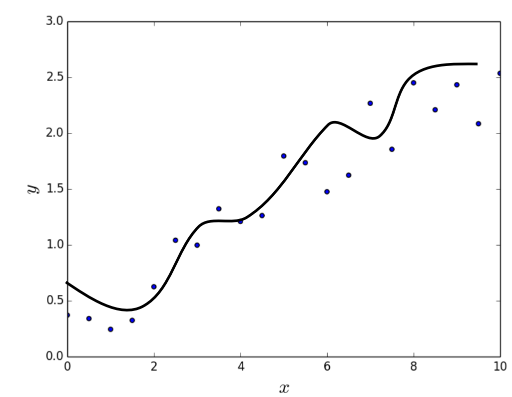
data

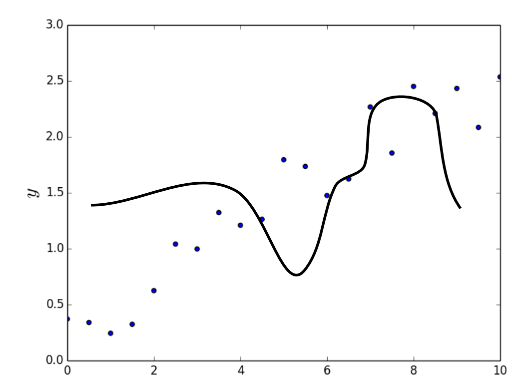
Supervised learning pipeline



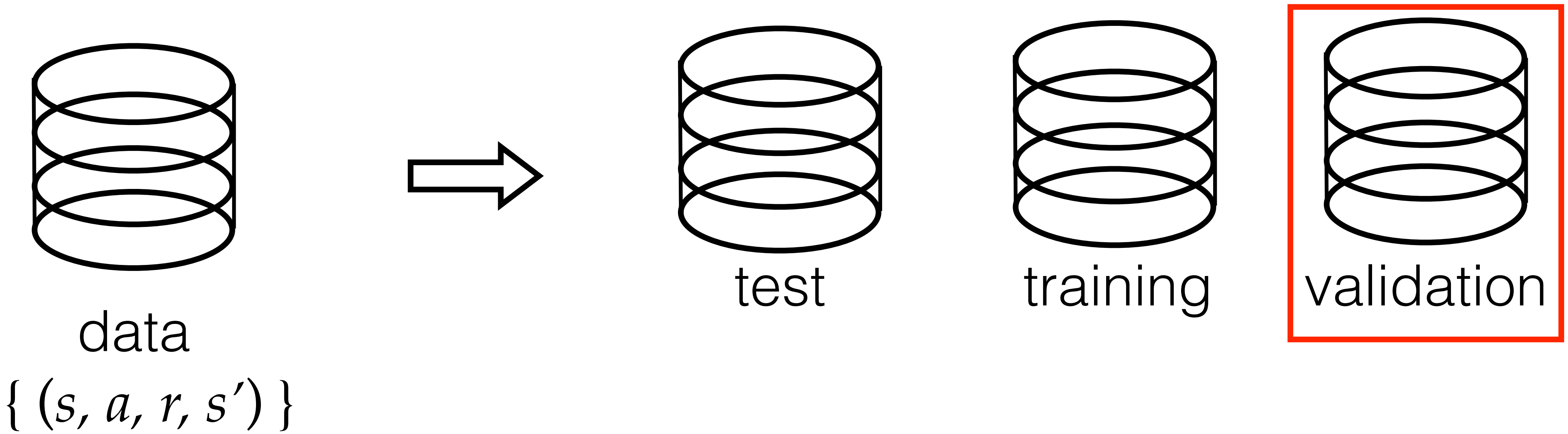
Supervised learning pipeline

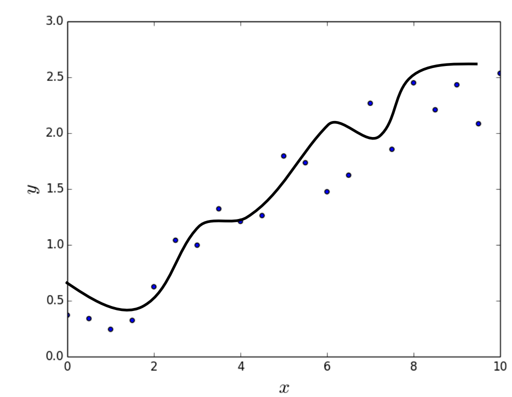


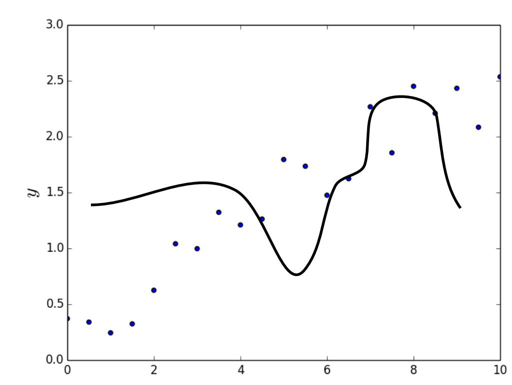




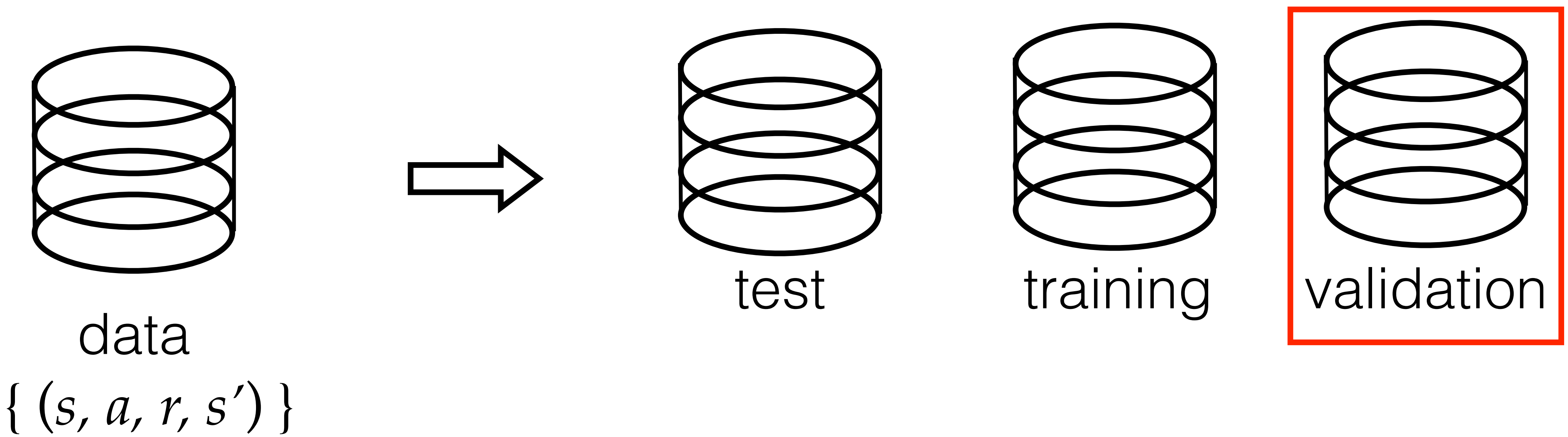
Offline RL pipeline



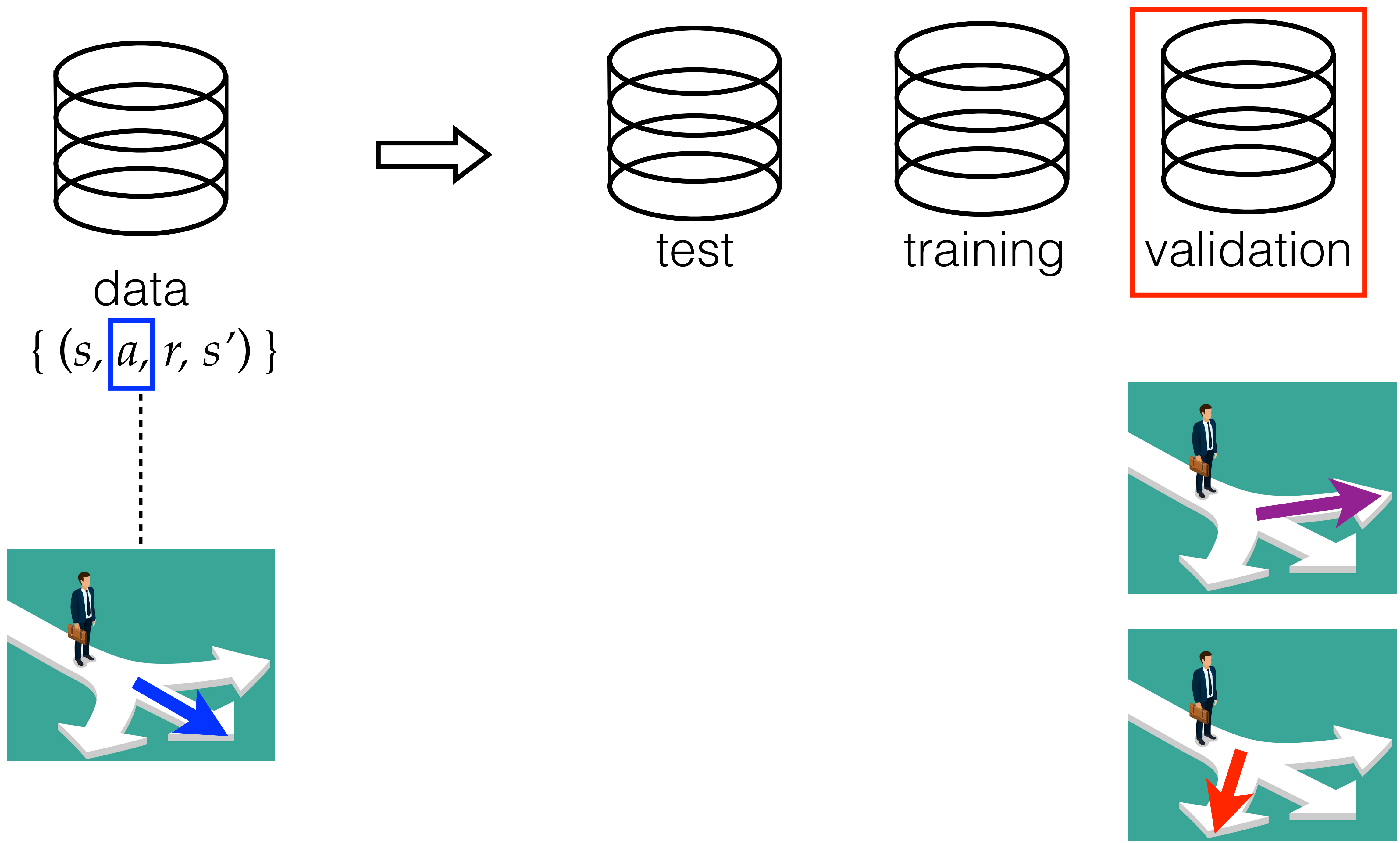




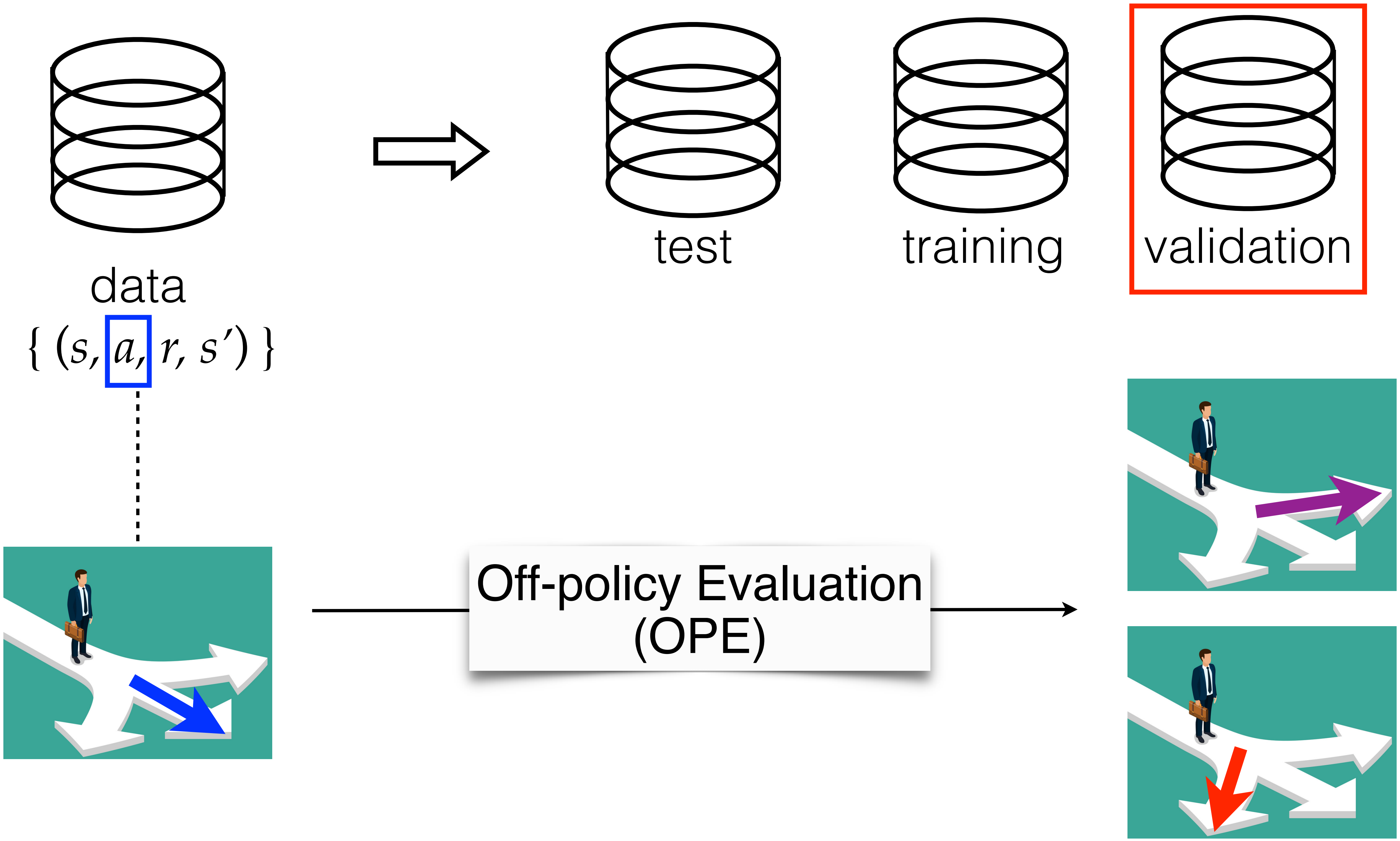
Offline RL pipeline



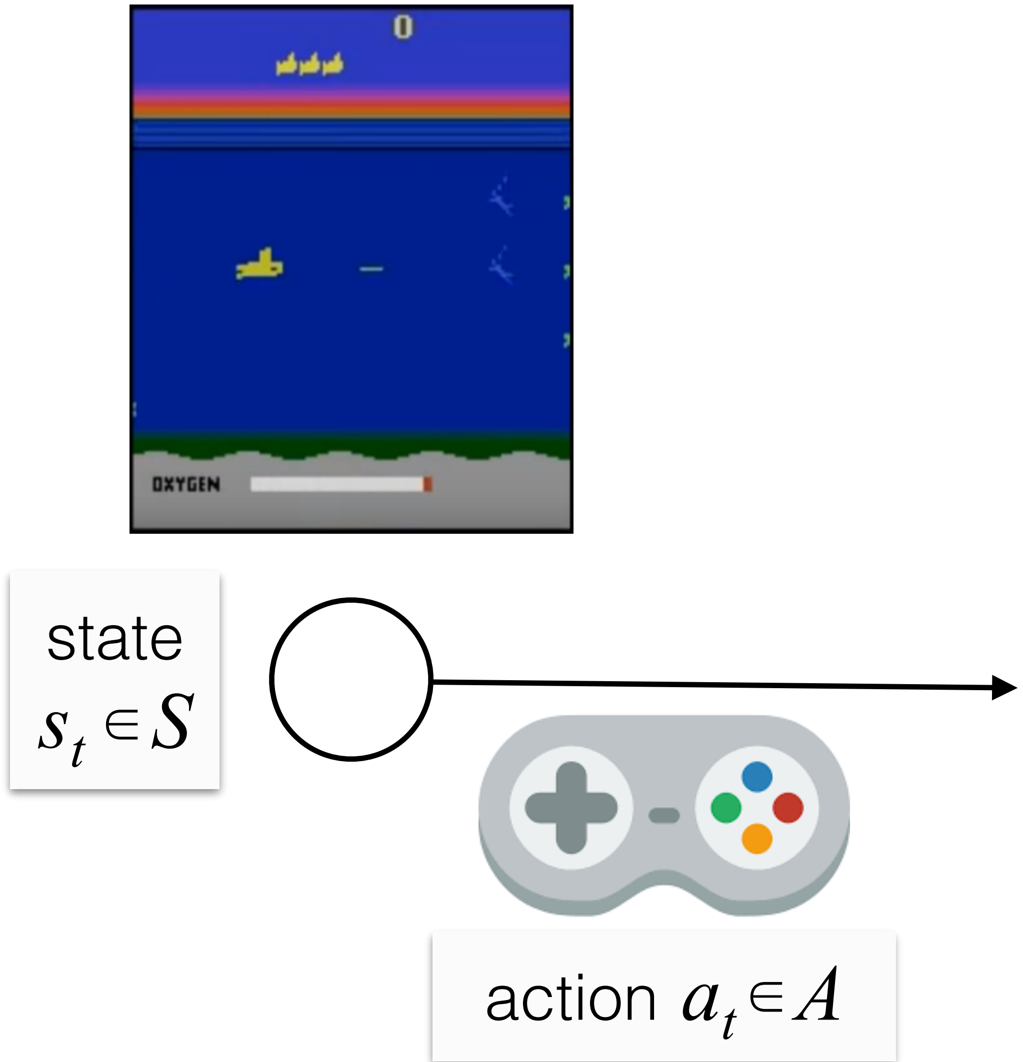
Offline RL pipeline

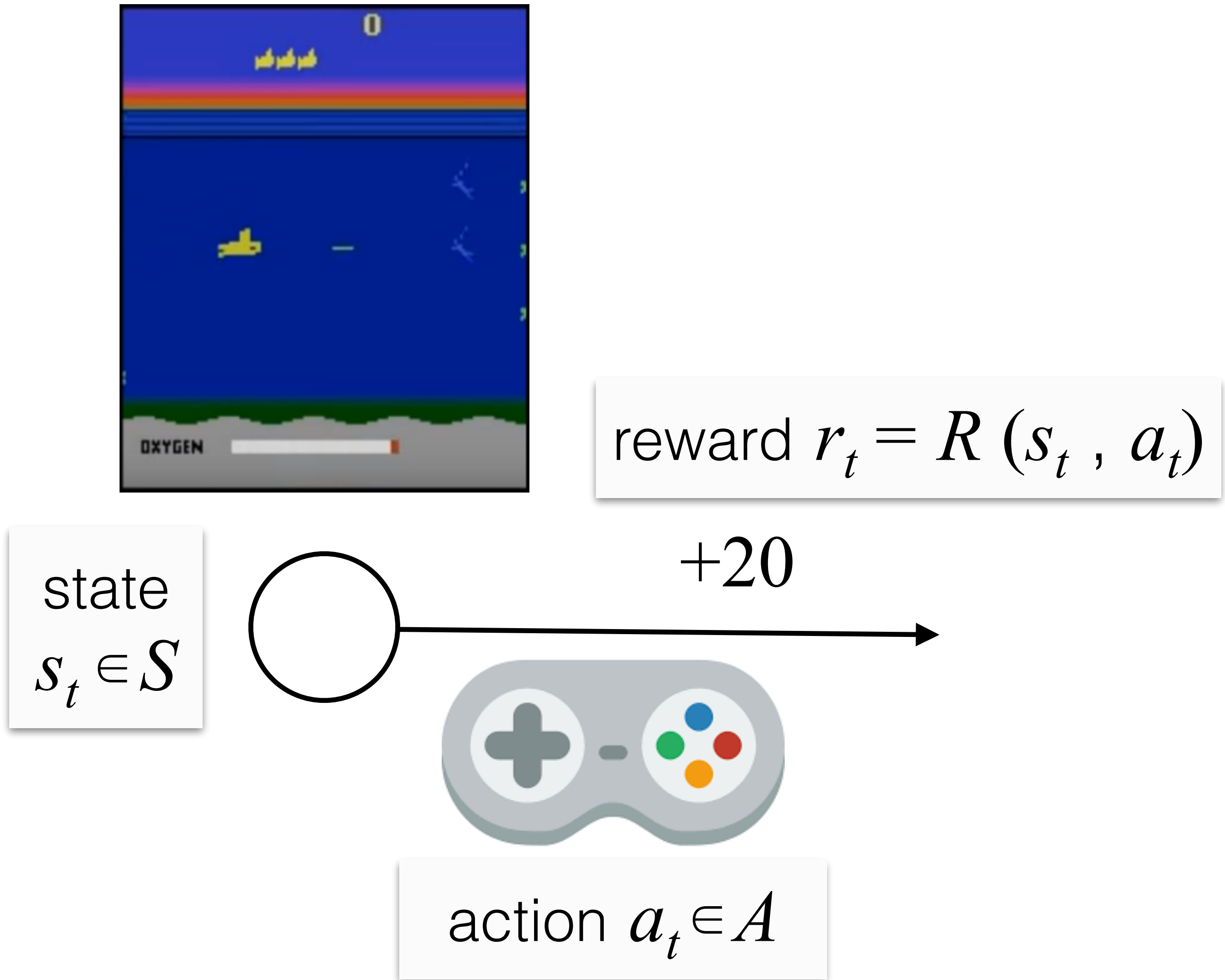


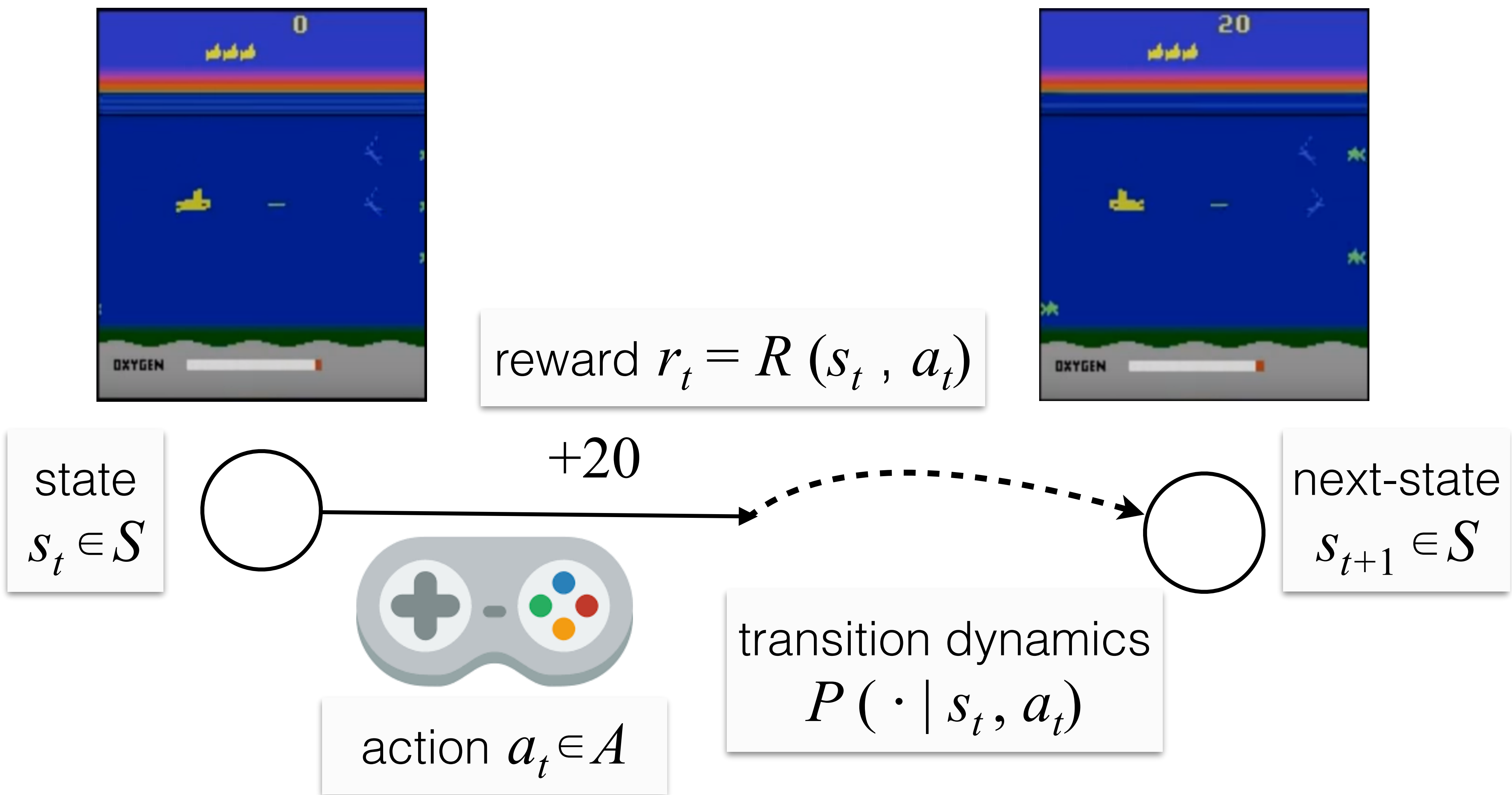
Offline RL pipeline

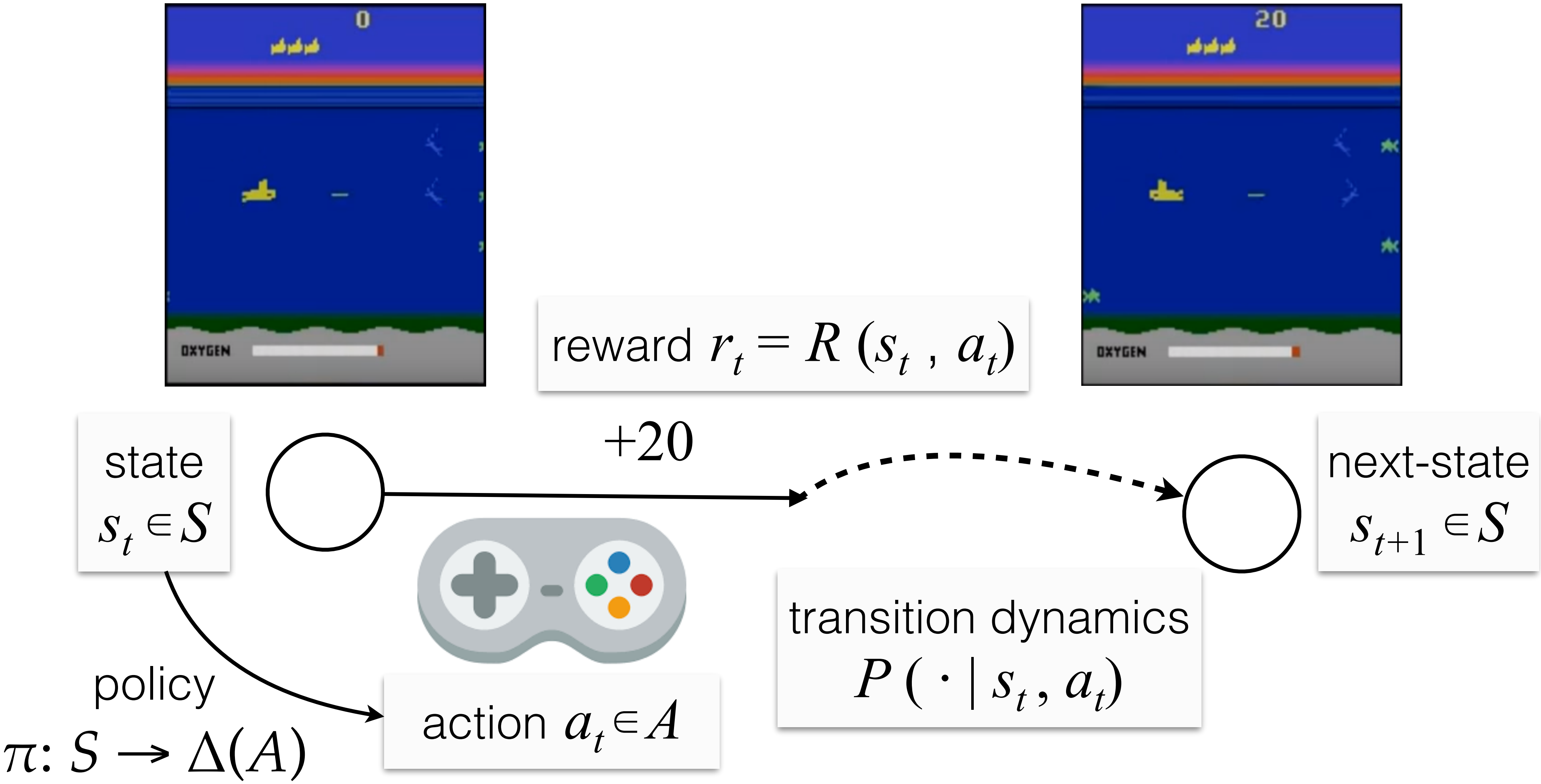


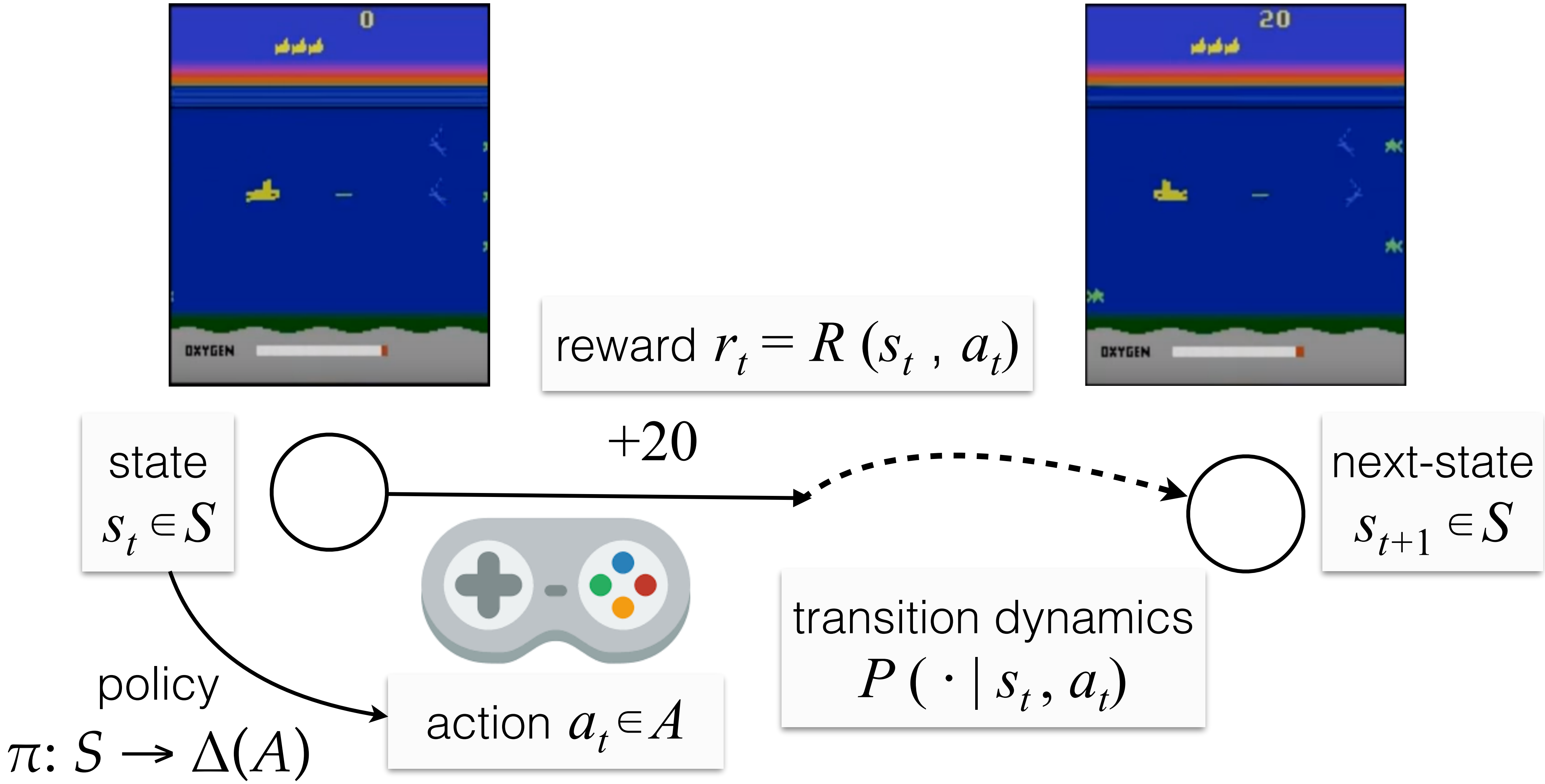
state
$s_t \in S$


OXYGEN
state
$s_t \in S$
action $a_t \in A$

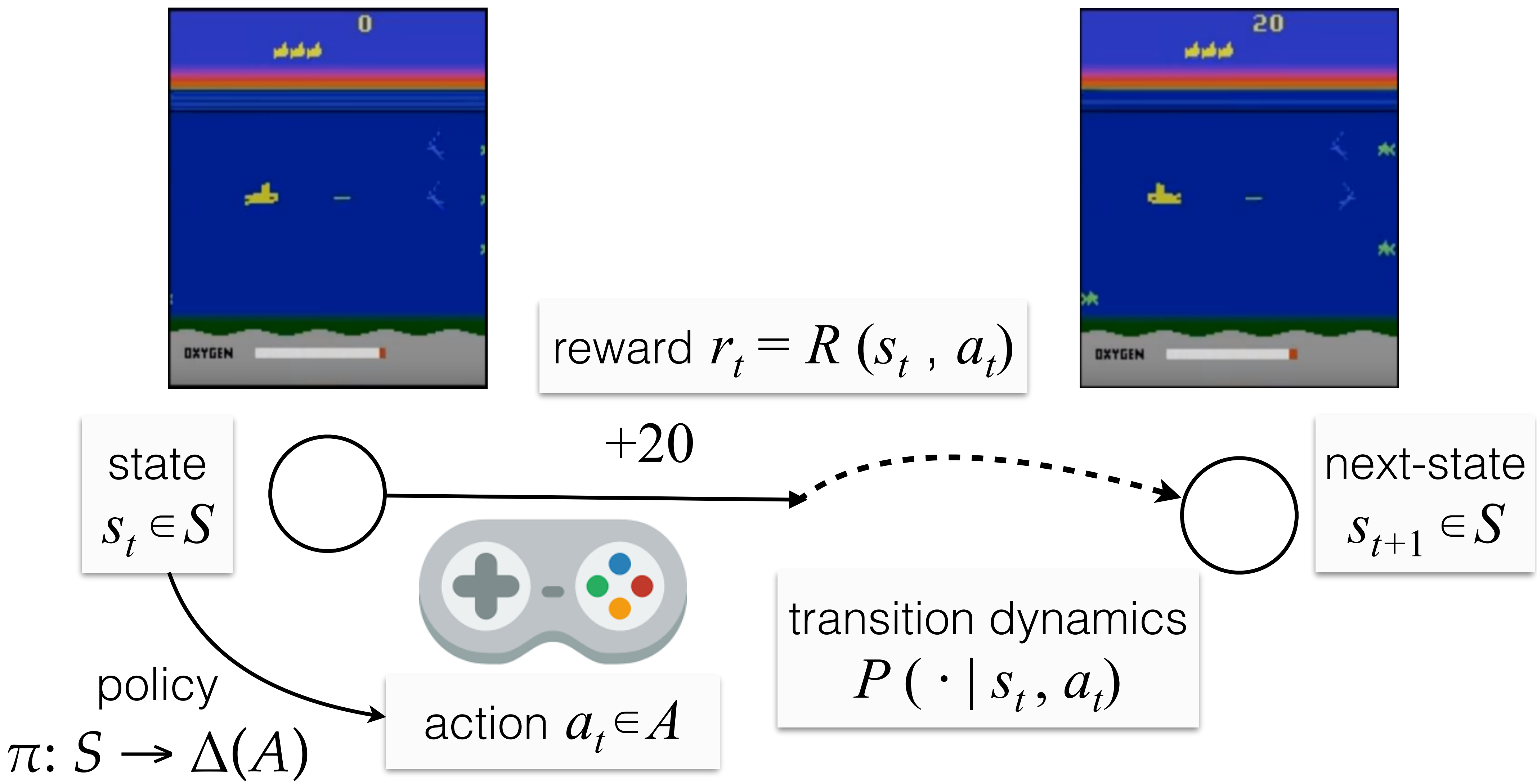

reward $r_t = R(s_t, a_t)$
state
$s_t \in S$
+20
action $a_t \in A$


0
OXYGEN
20
OXYGEN
reward $r_t = R(s_t, a_t)$
+20
state
$s_t \in S$
action $a_t \in A$
transition dynamics
$P(\cdot \mid s_t, a_t)$
next-state
$s_{t+1} \in S$

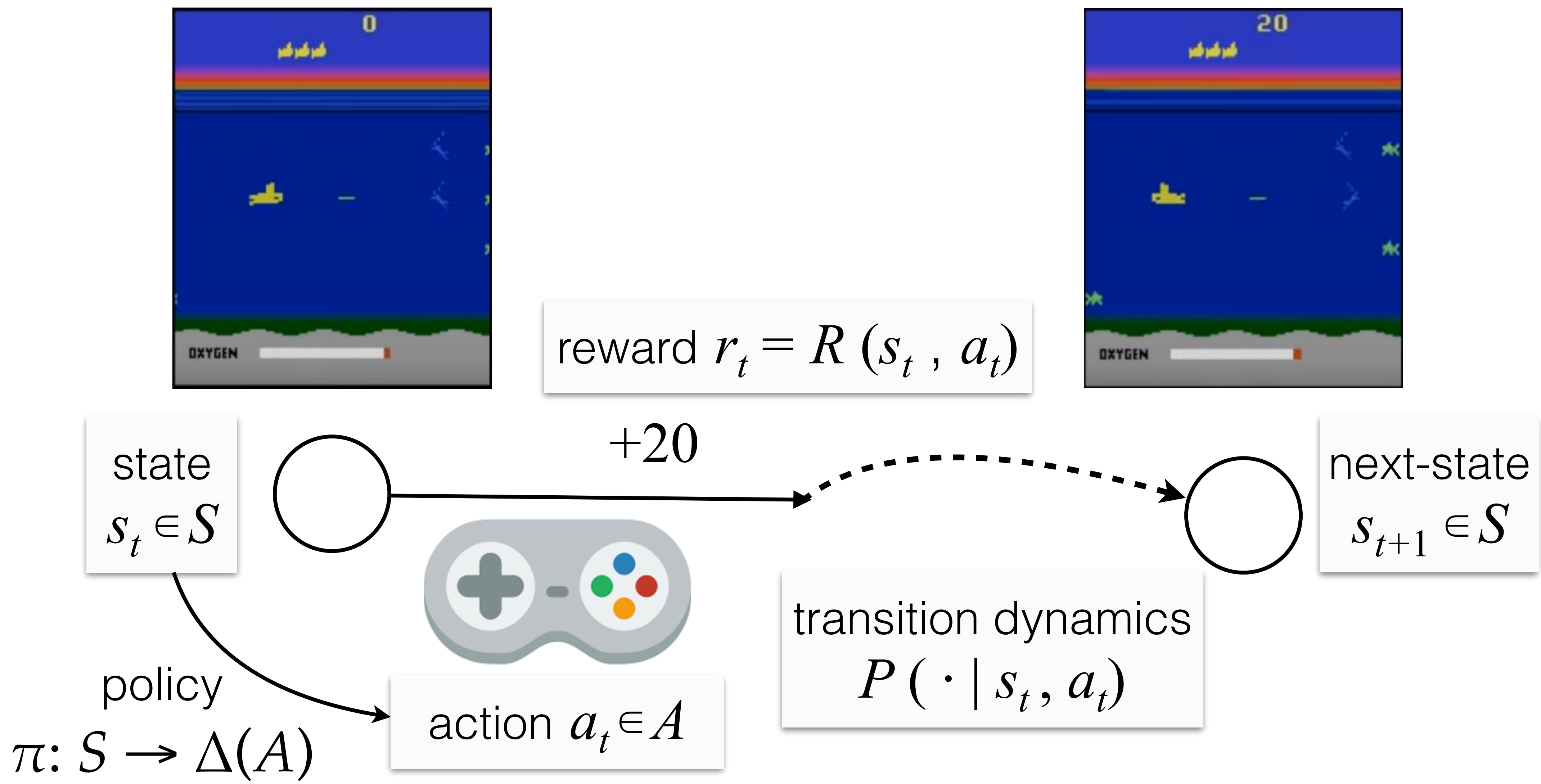

reward $r_t = R(s_t, a_t)$
+20
state
$s_t \in S$
next-state
$s_{t+1} \in S$
transition dynamics
$P(\cdot \mid s_t, a_t)$
policy
$\pi: S \to \Delta(A)$
action $a_t \in A$



$$J(\pi) := \mathbb{E}_\pi [\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0]$$

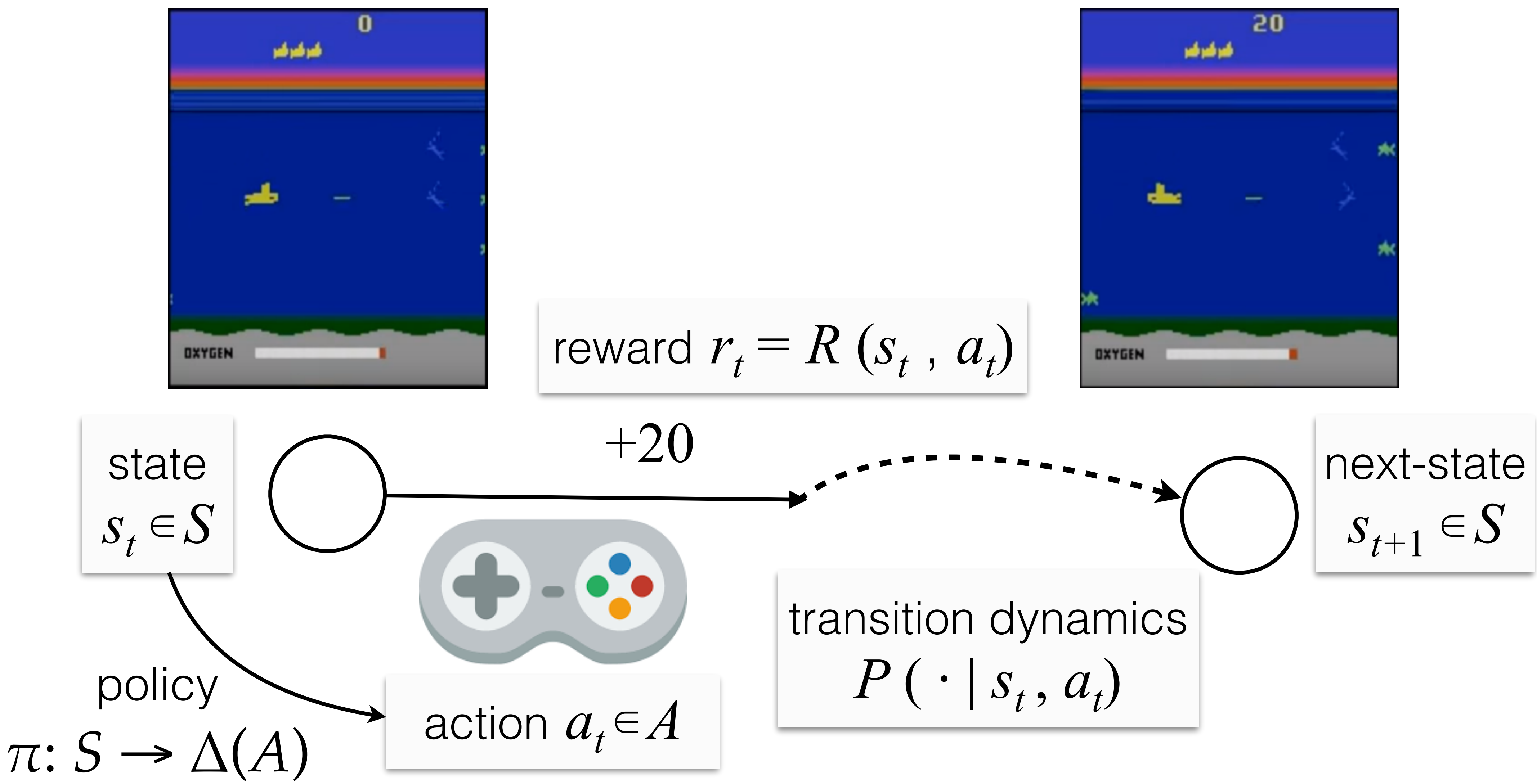


Policy evaluation: estimate $J(\pi) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0]$ given π



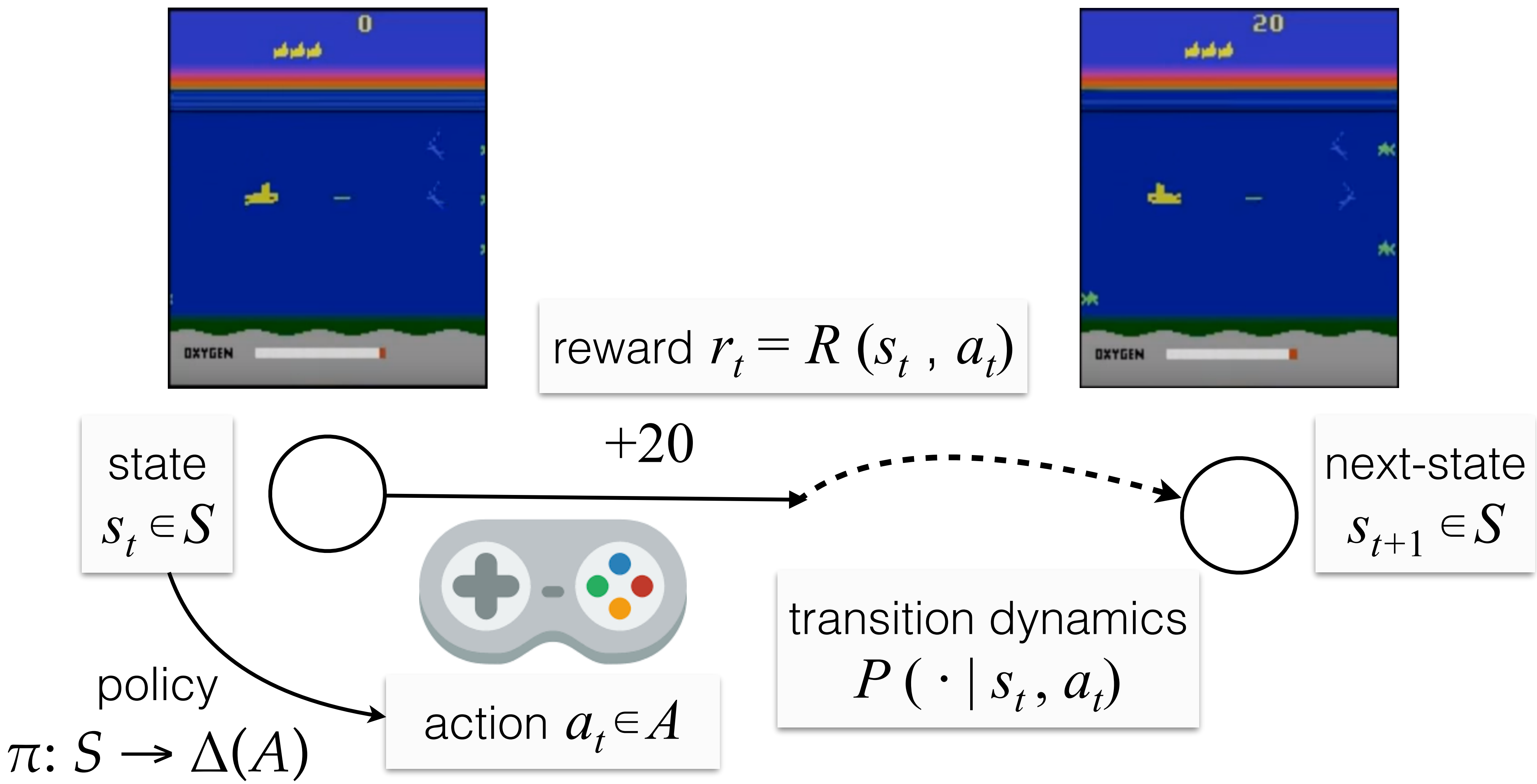
Policy evaluation: estimate $J(\pi) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0]$ given π

Policy optimization: $\max_\pi J(\pi)$



Policy evaluation: estimate $J(\pi) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0]$ given π

$$= Q^\pi(s_0, \pi)$$

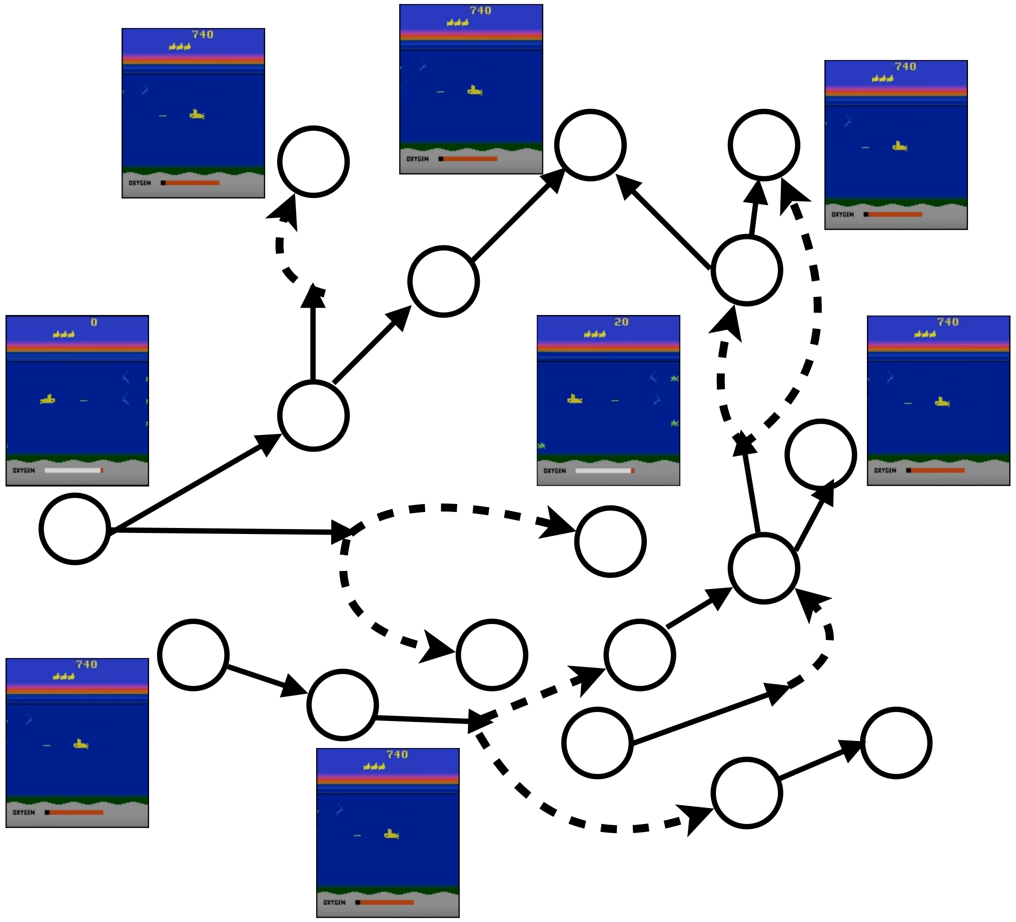


Policy evaluation: estimate $J(\pi) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0]$ given π

$$= Q^\pi(s_0, \pi)$$

How to find Q^π?

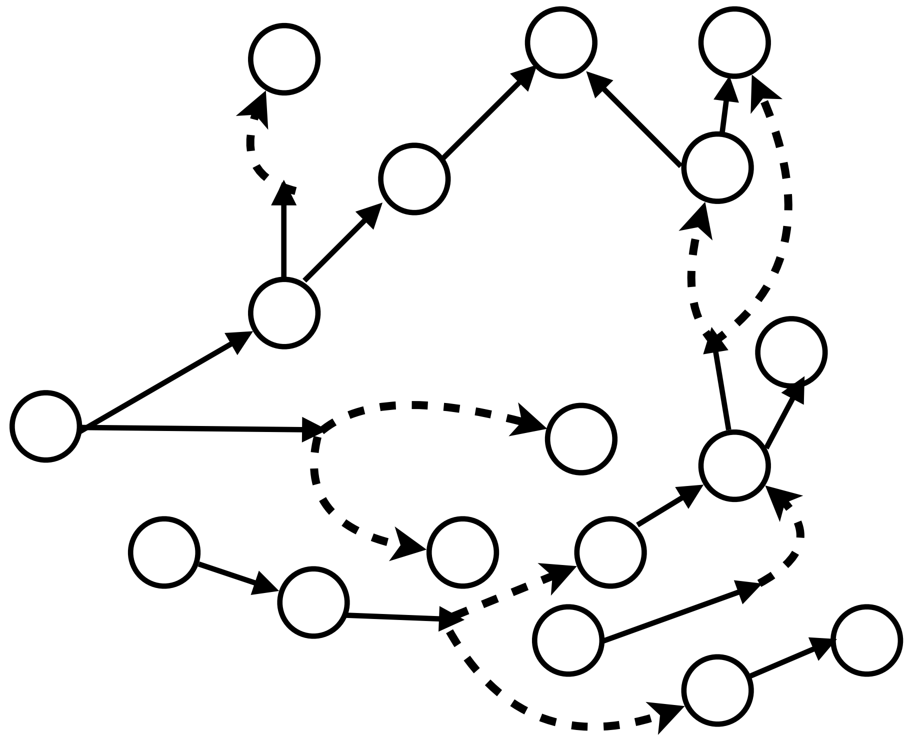
$Q^\pi = \mathcal{T}^\pi Q^\pi \quad \rightarrow \quad |SxA|$ equations

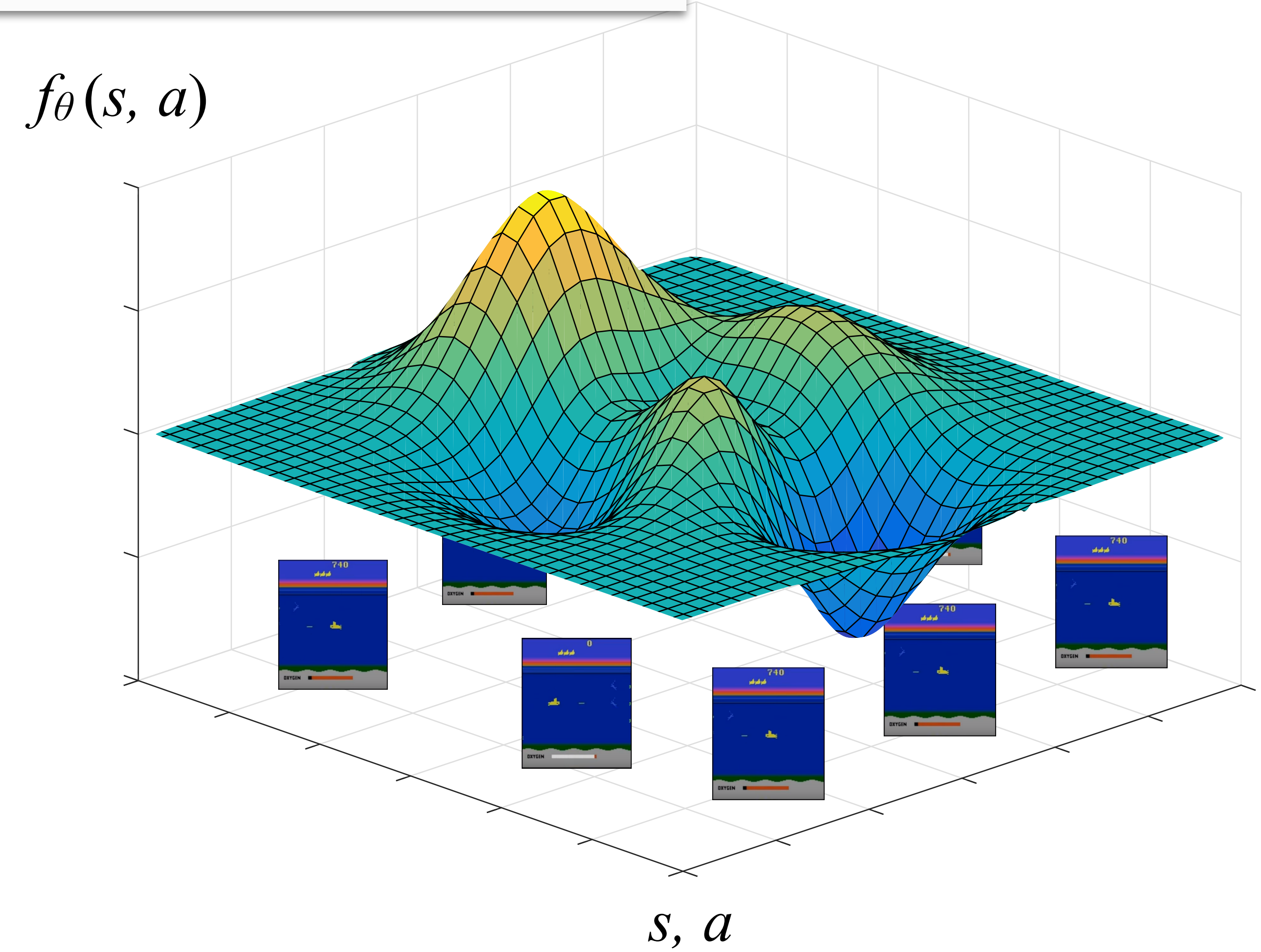


How to find Q^π?

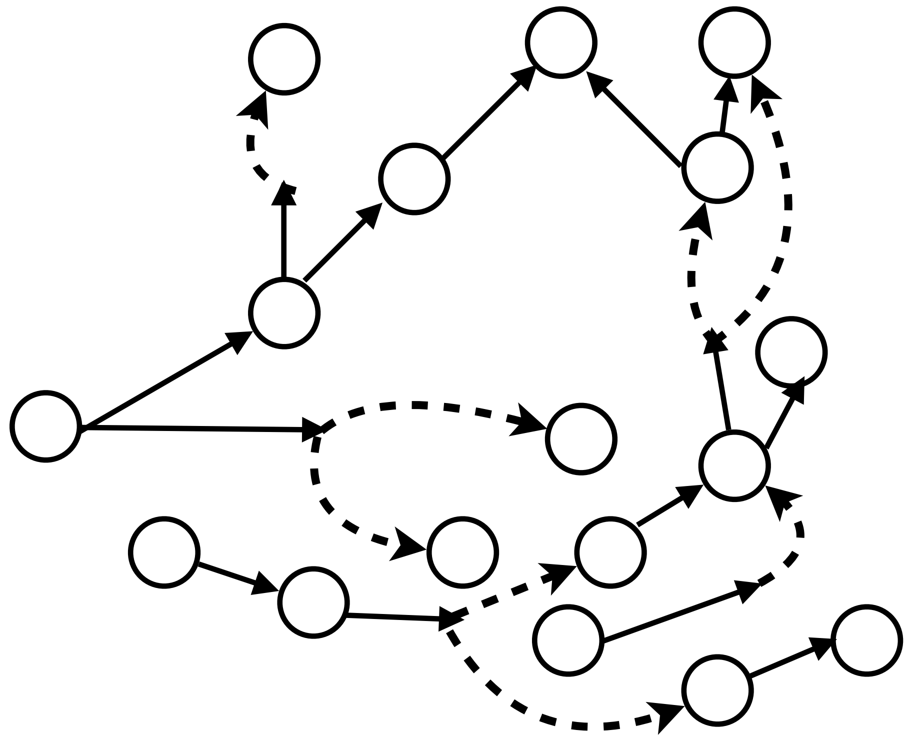
$Q^\pi = \mathcal{T}^\pi Q^\pi \;\; \rightarrow \;\; |S\text{x}A|$ equations **X**

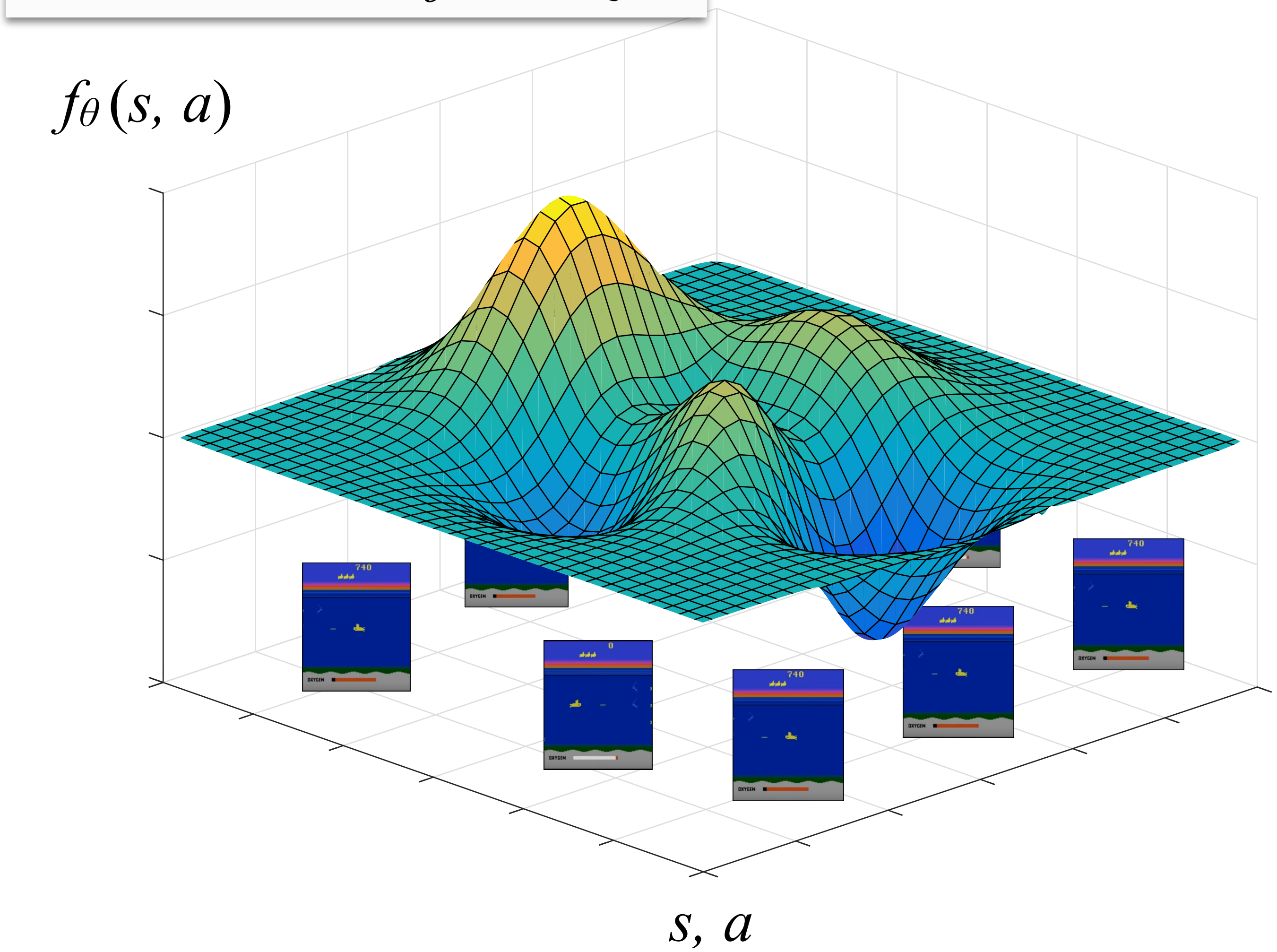
Find θ s.t. $f_\theta \approx Q^\pi$





Find θ s.t. $f_\theta \approx Q^\pi$

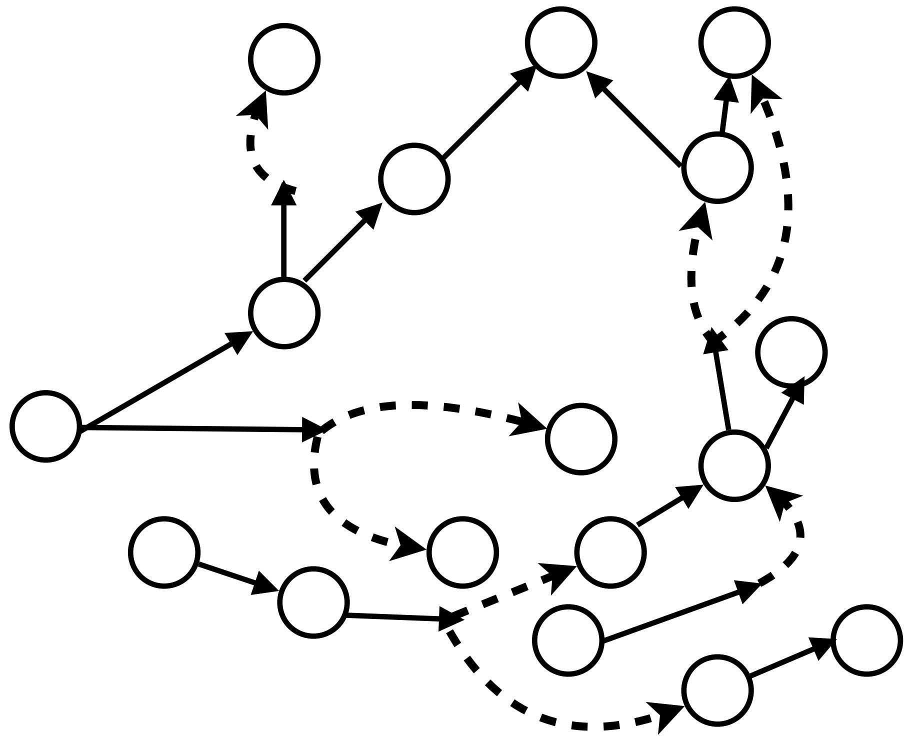




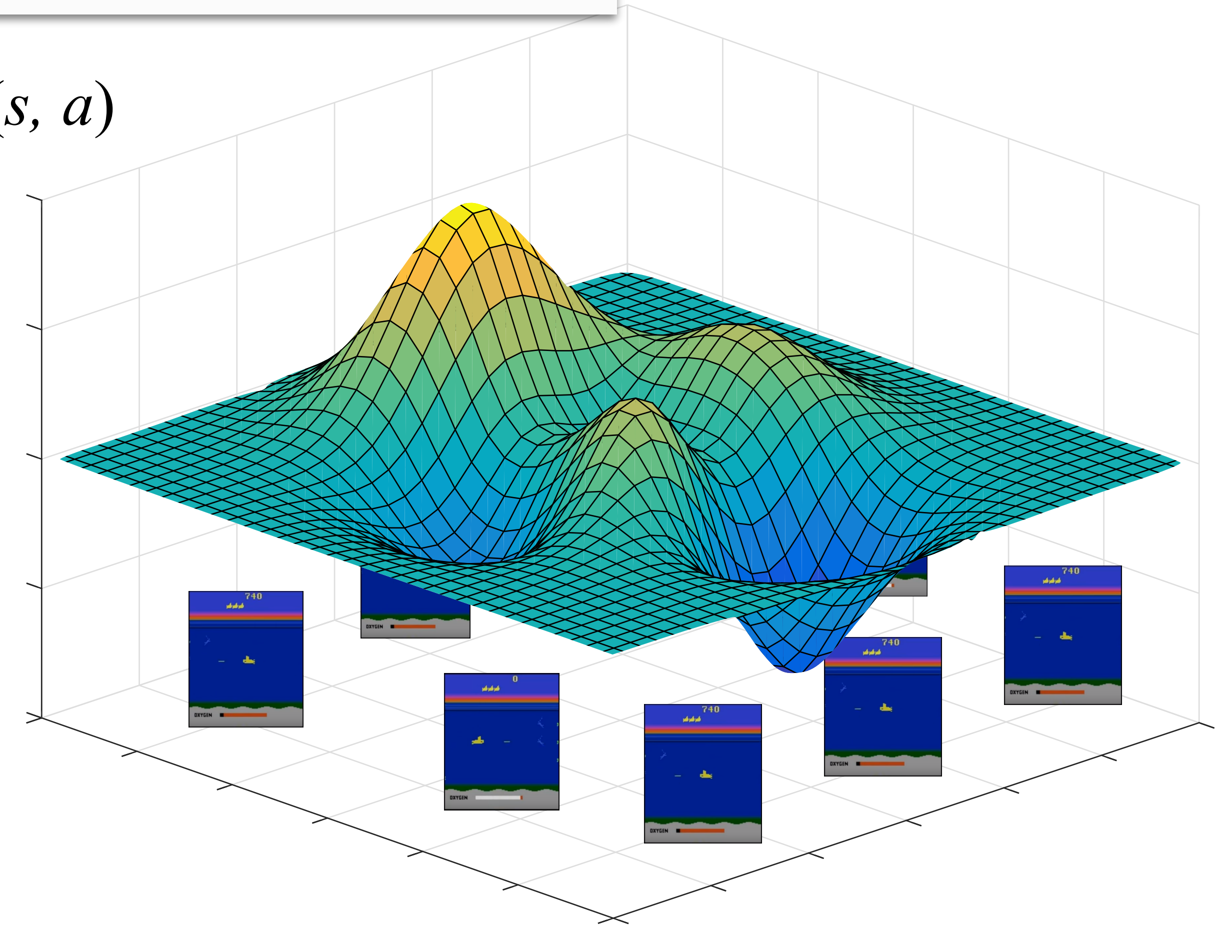
Validation:

(FQE: learn Q^π)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma f_{k-1}(s', \pi) \right)^2 \right]$$



Find θ s.t. $f_\theta \approx Q^\pi$

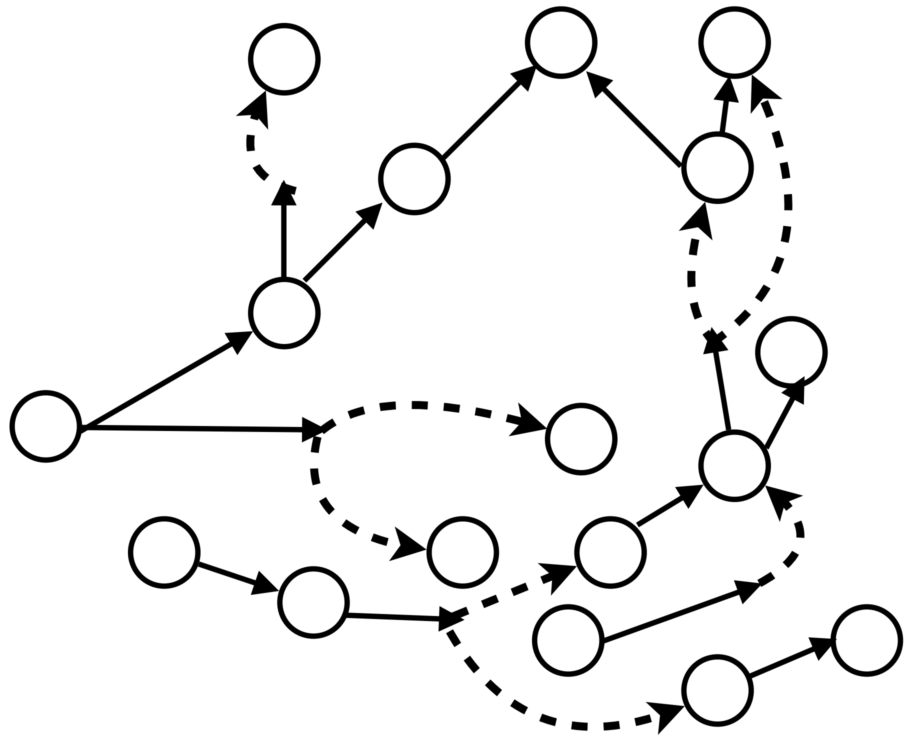


Validation:

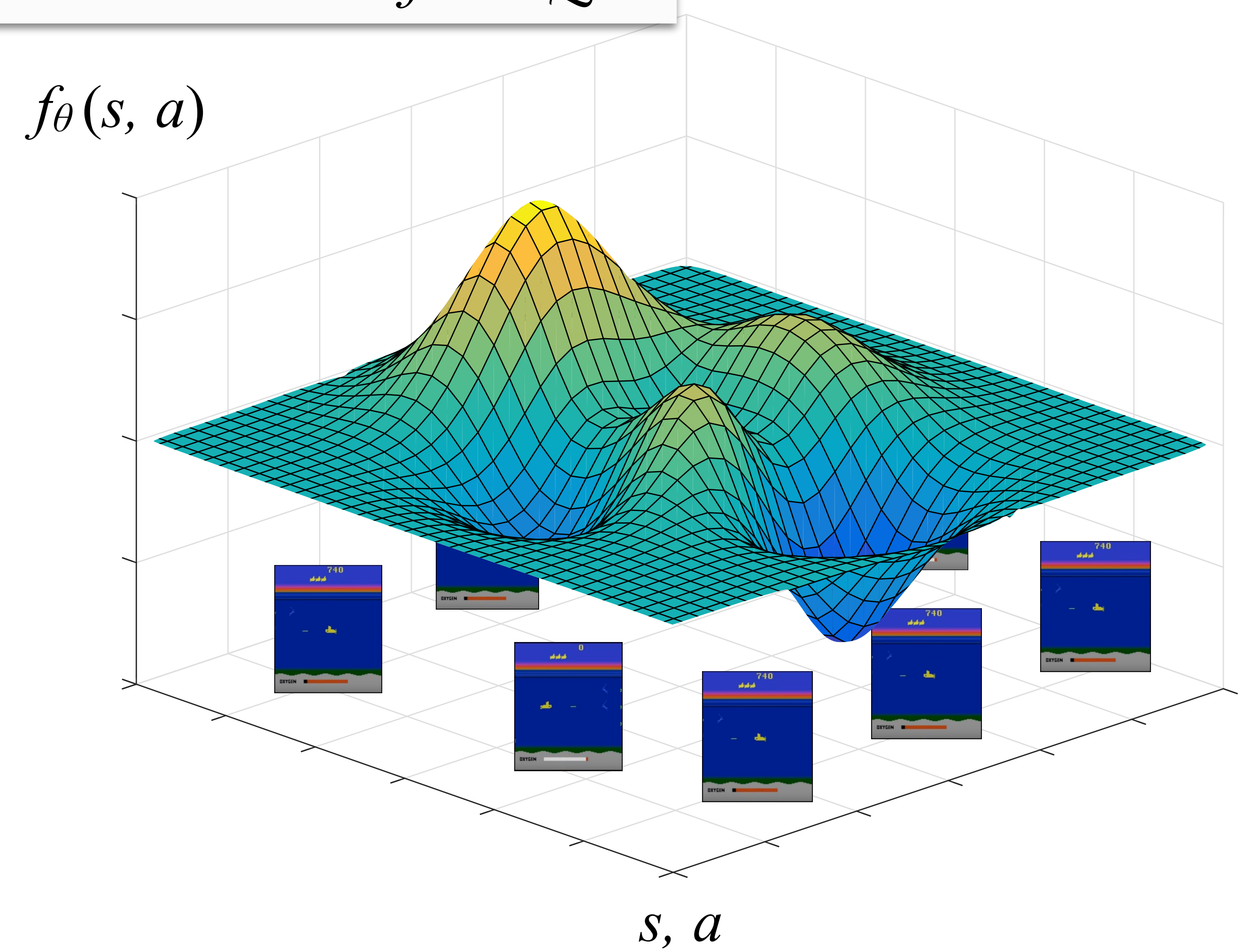
(FQE: learn Q^π)

iterative

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma f_{k-1}(s', \pi) \right)^2 \right]$$



Find θ s.t. $f_\theta \approx Q^\pi$



$(s, a, r, s') \sim D$

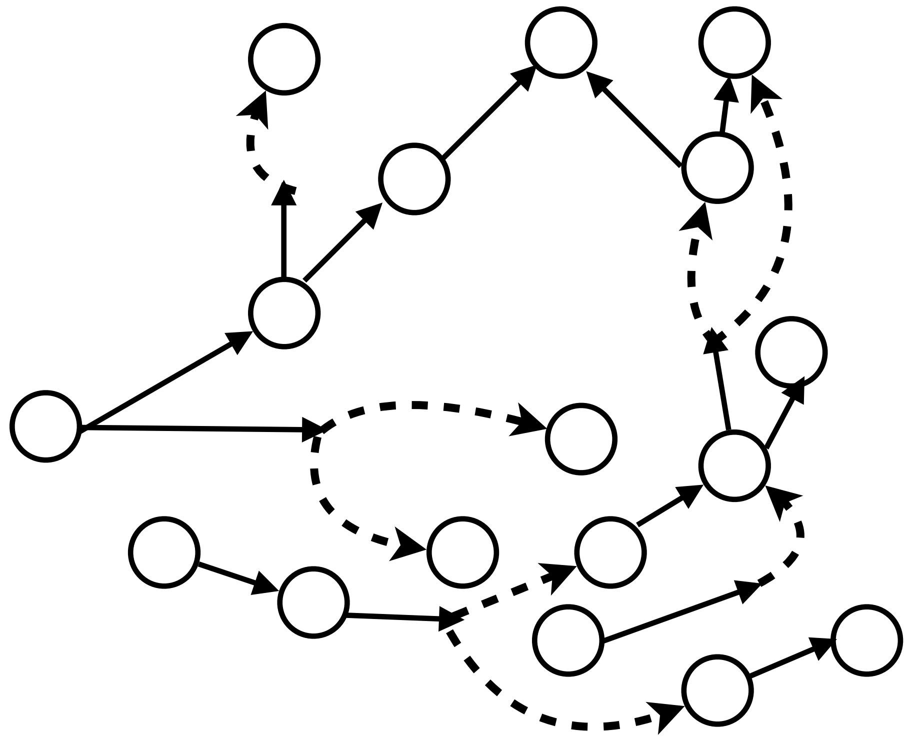
Validation:

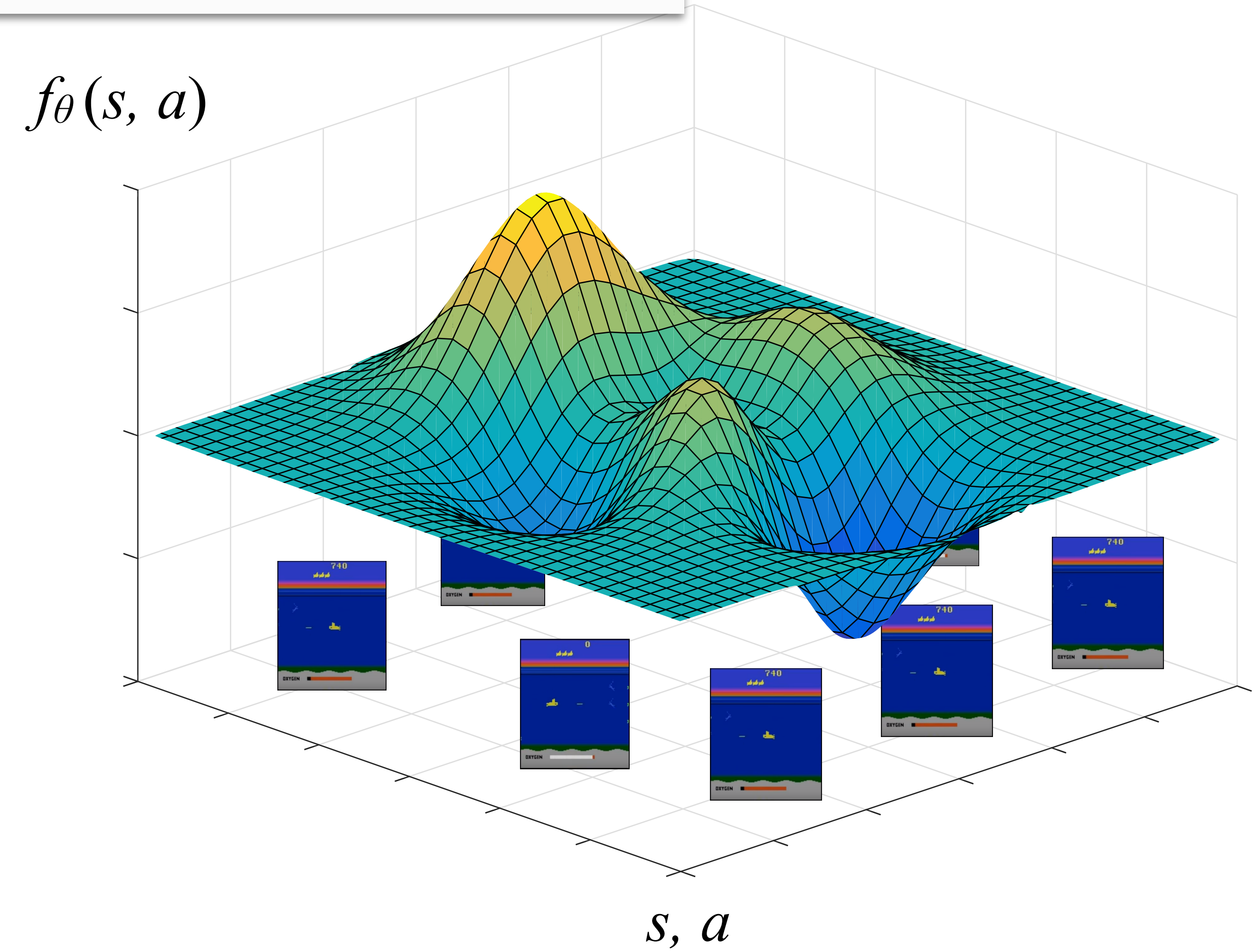
(FQE: learn Q^π)

iterative

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s, a) - r - \gamma f_{k-1}(s', \pi) \right)^2 \right]$$

Find θ s.t. $f_\theta \approx Q^\pi$





$$s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, s_3, \ldots$$

$(s, a, r, s') \sim D$

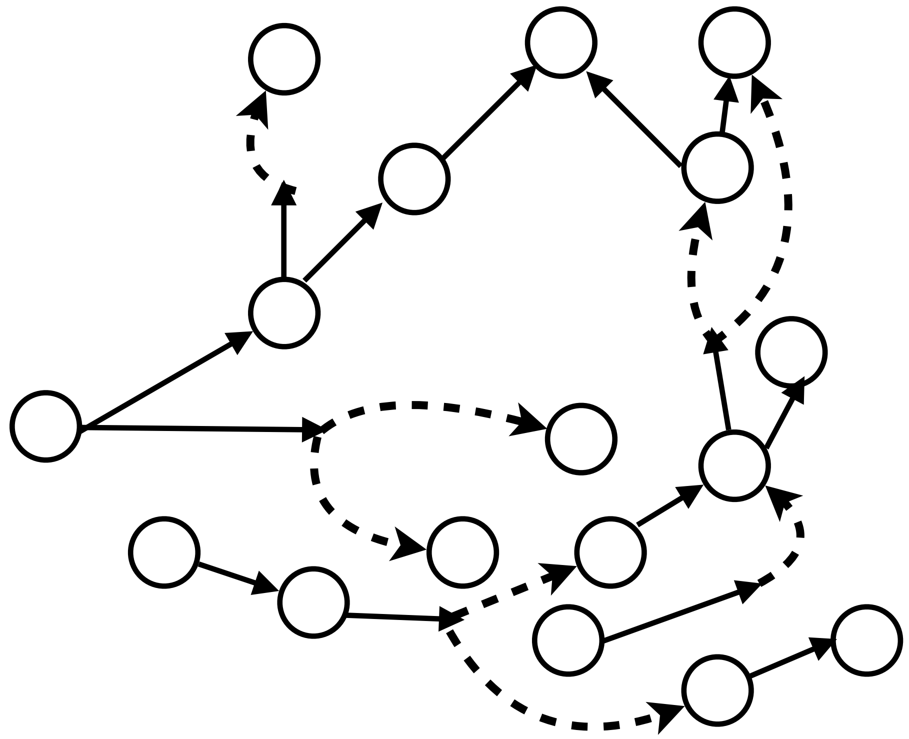
Validation:

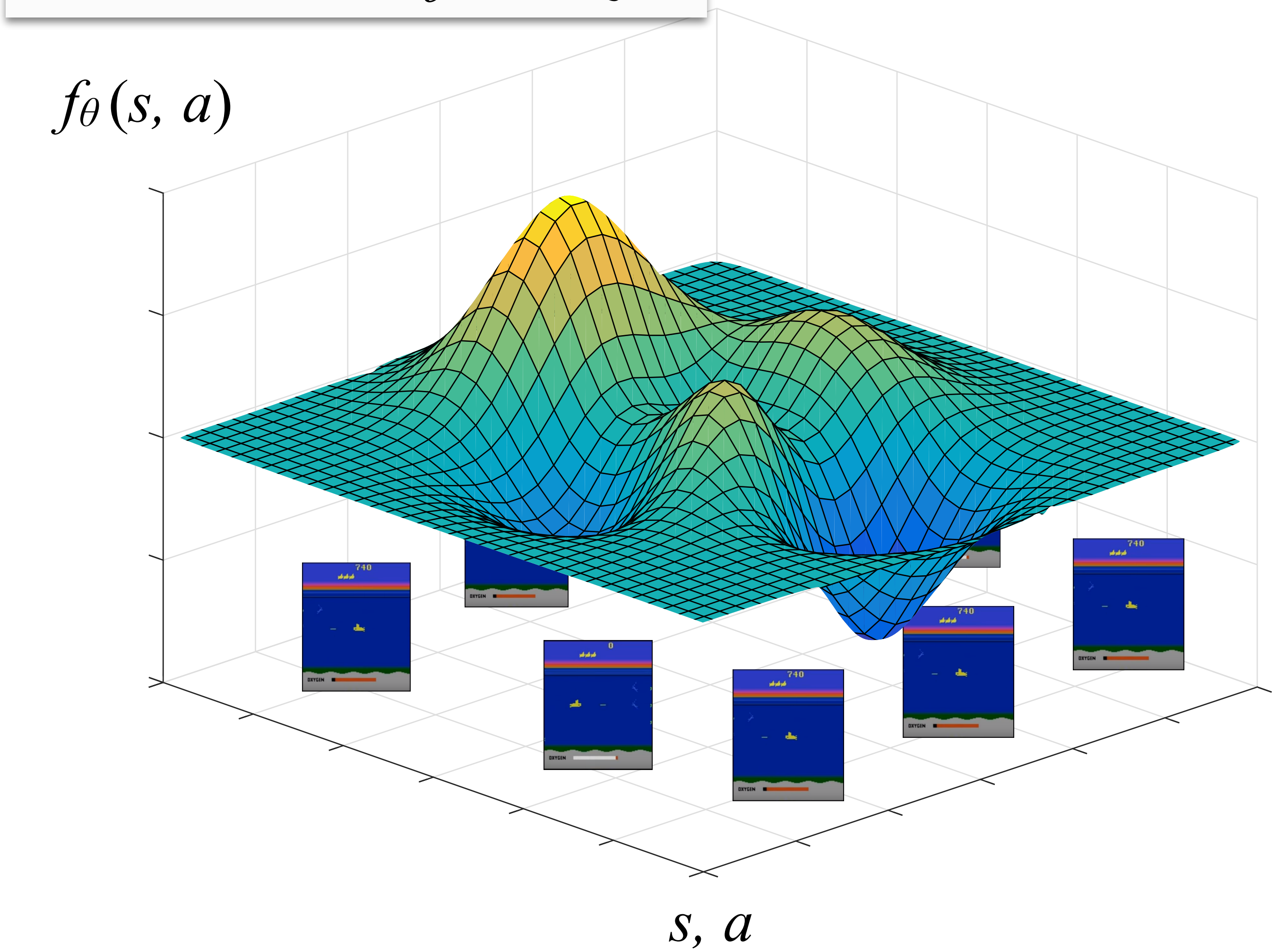
(FQE: learn Q^π)

iterative

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D [(f_\theta(s,a) - r - \gamma f_{k-1}(s', \pi))^2]$$

Find θ s.t. $f_\theta \approx Q^\pi$





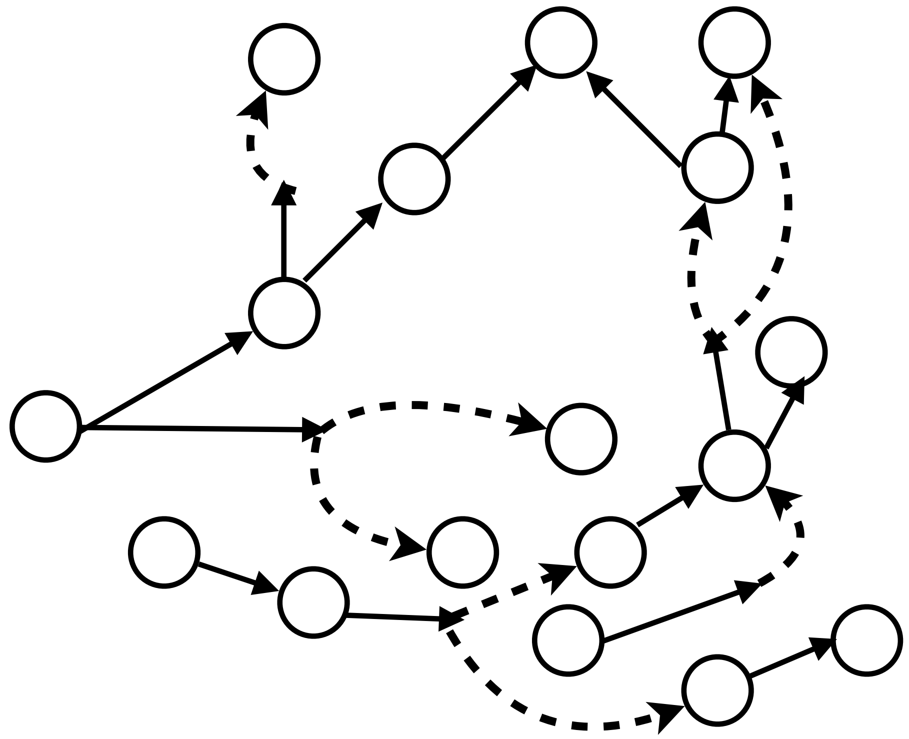
Validation:

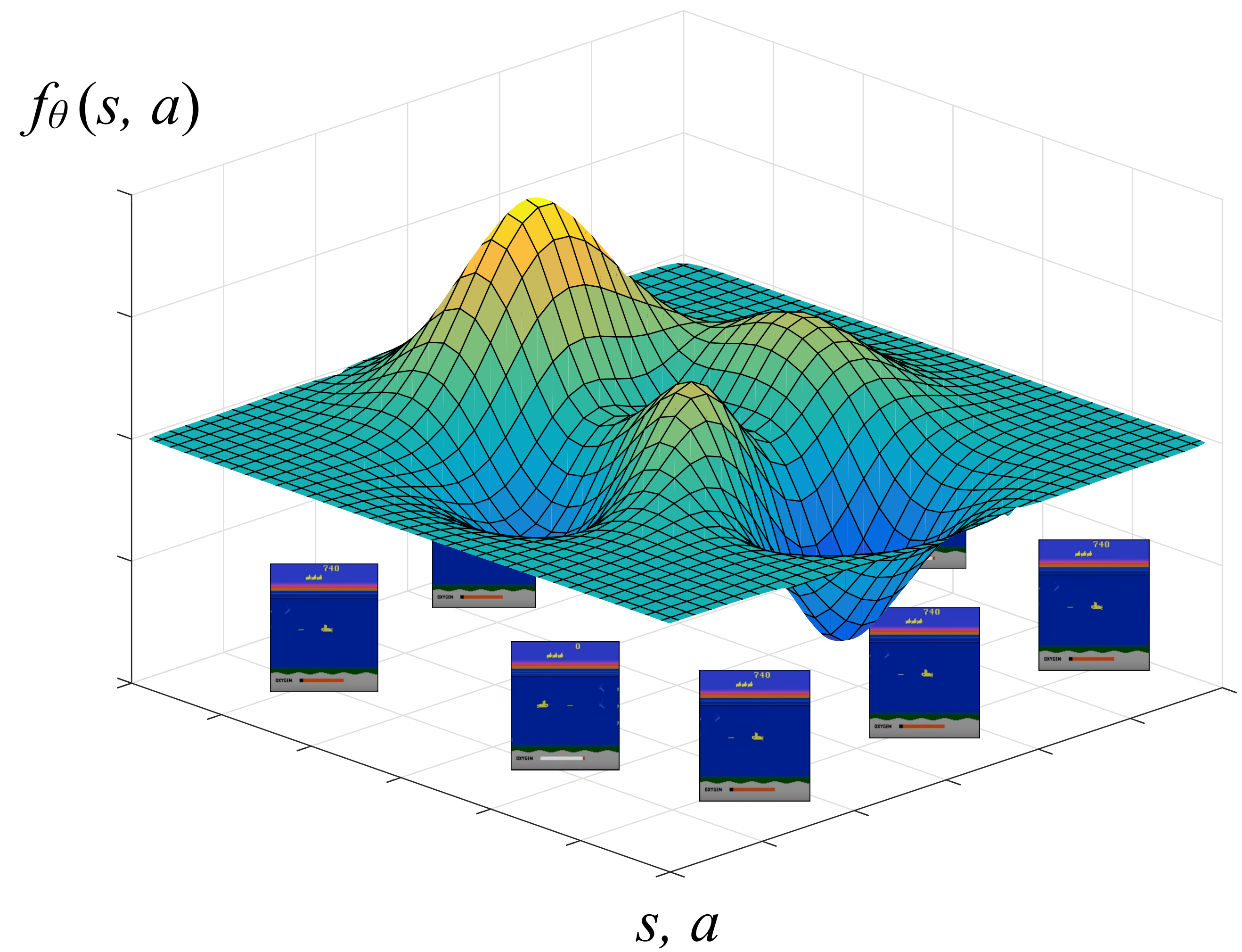
(FQE: learn Q^π)

iterative

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D\left[\left(f_\theta(s,a) - r - \gamma f_{k-1}(s', \pi)\right)^2\right]$$

$\approx \mathcal{T}^\pi f_{k-1}$ ← $\mathrm{E}[\ \cdot \ | s, a]$





Training: $\hat{f} = f_k$ where

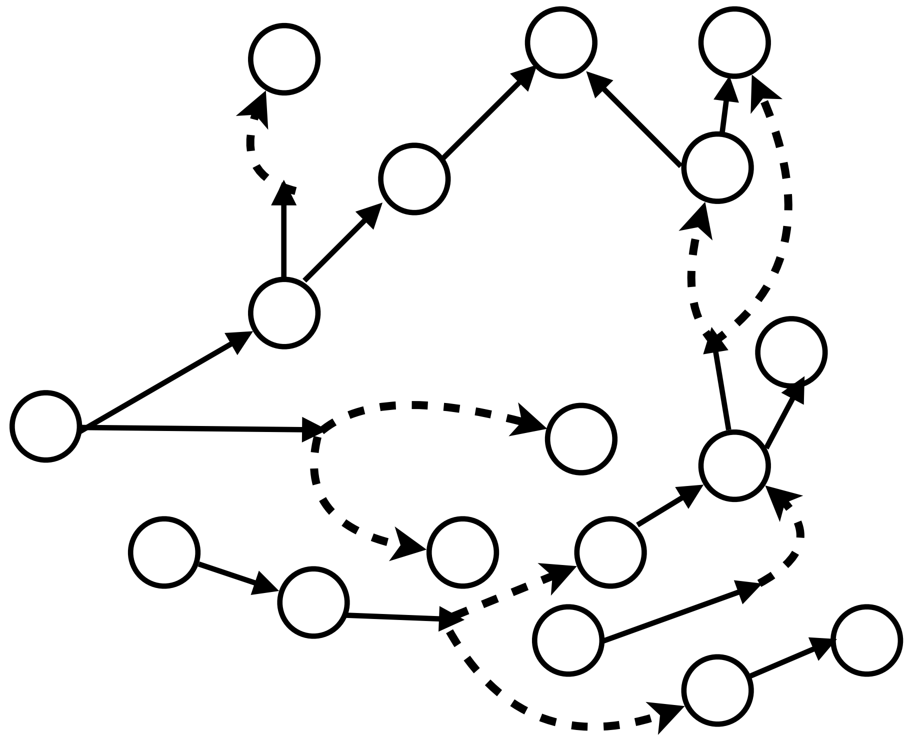
(FQI: learn Q^*)

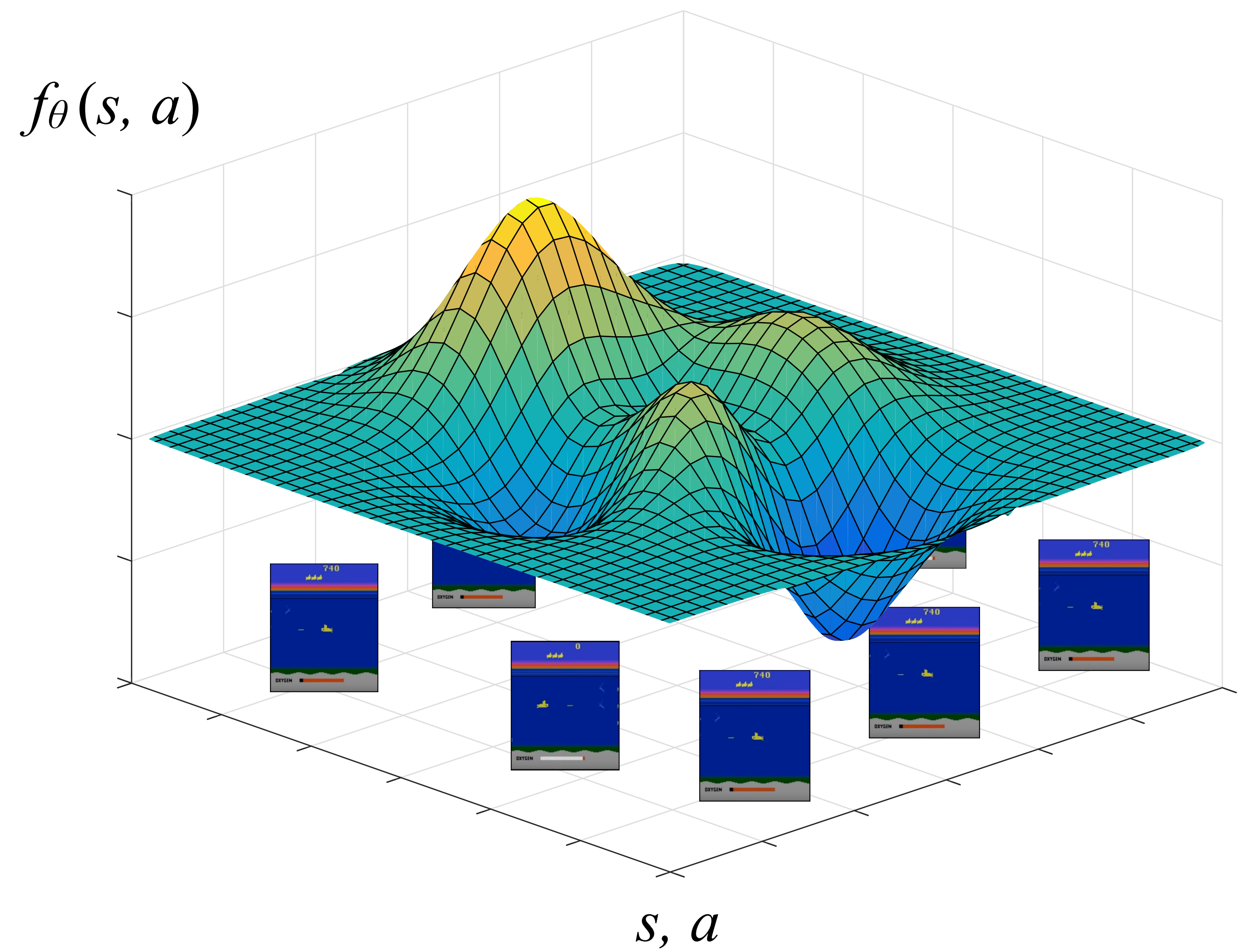
$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s', a') \right)^2 \right]$$

Validation:

(FQE: learn Q^π)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma f_{k-1}(s', \pi) \right)^2 \right]$$





Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

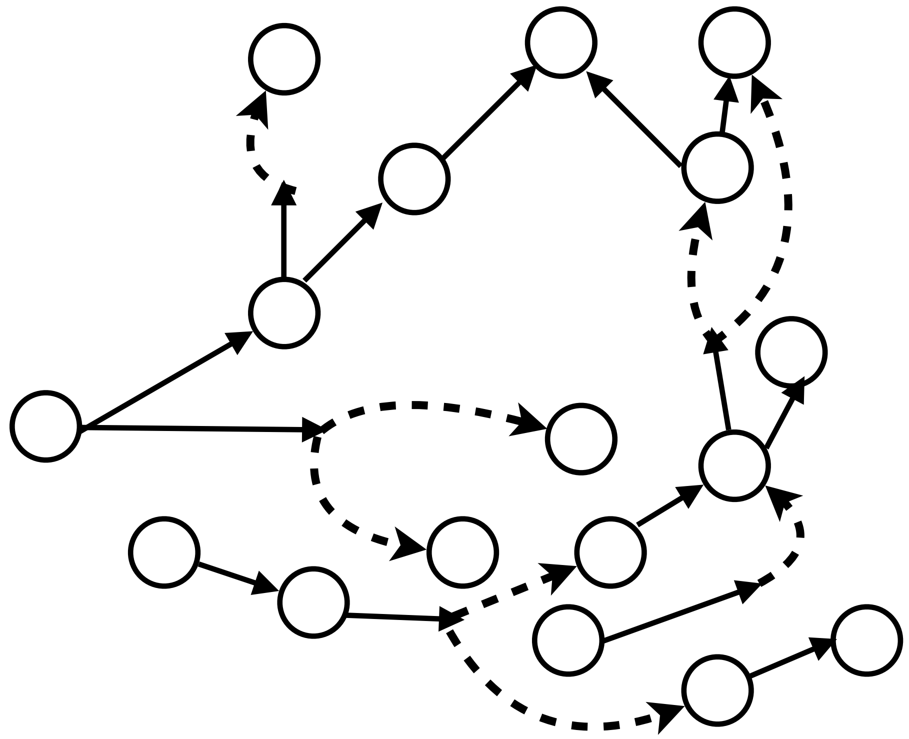
$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s', a') \right)^2 \right]$$

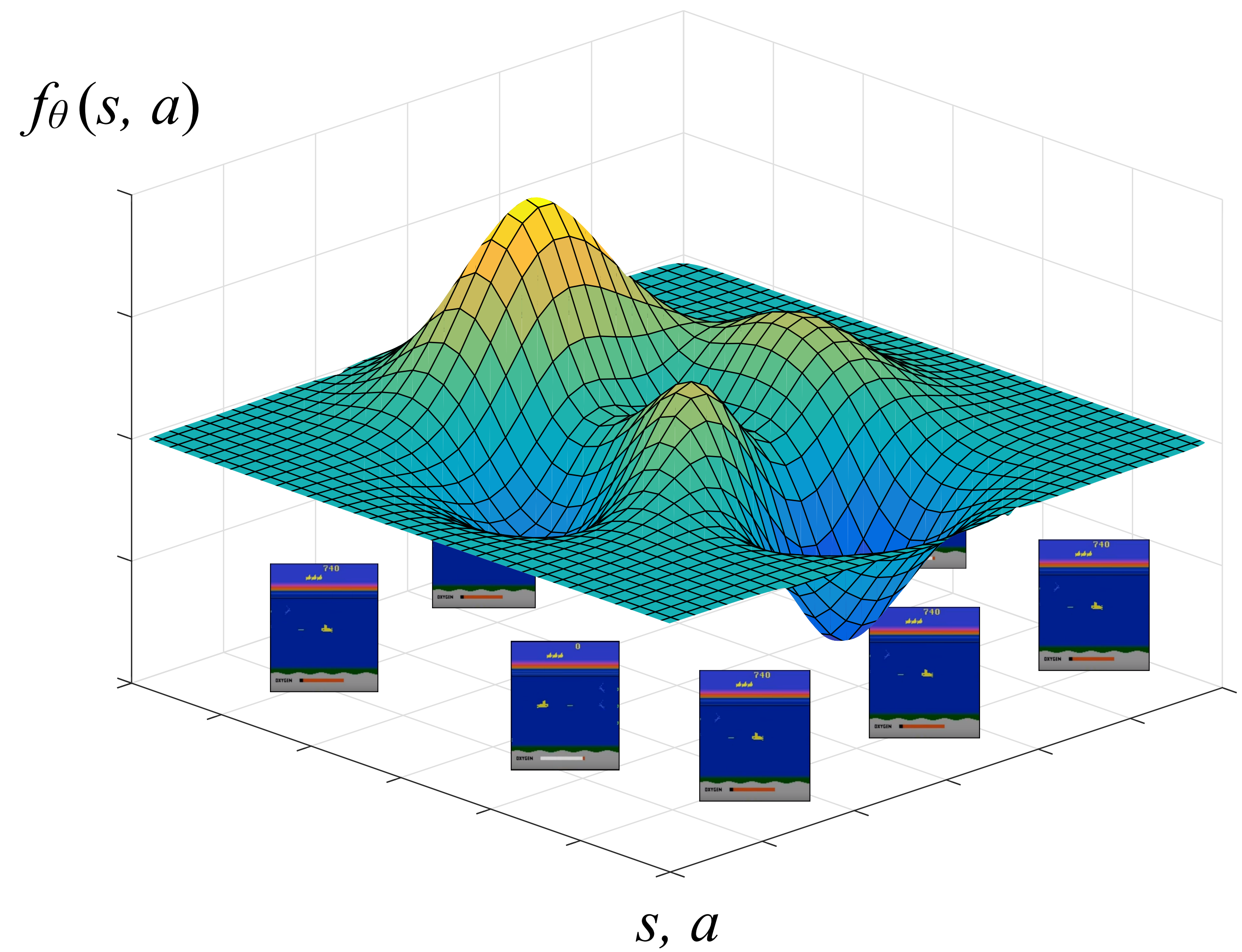
$$\Downarrow \; \pi = \text{greedy w.r.t. } \hat{f}$$

Validation:

(FQE: learn Q^π)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma f_{k-1}(s', \pi) \right)^2 \right]$$

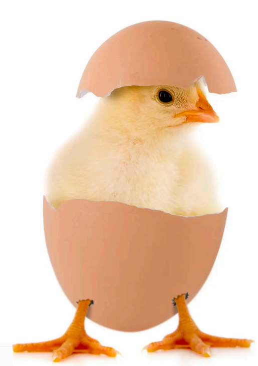




Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s', a') \right)^2 \right]$$



$$\Downarrow \; \pi = \text{greedy w.r.t. } \hat{f}$$

Validation:

(FQE: learn Q^π)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma f_{k-1}(s', \pi) \right)^2 \right]$$

Hyperparameter-free methods?

Importance sampling [Precup'00]

- Hyperparameter-free ✓
- No Markovianity required ✓

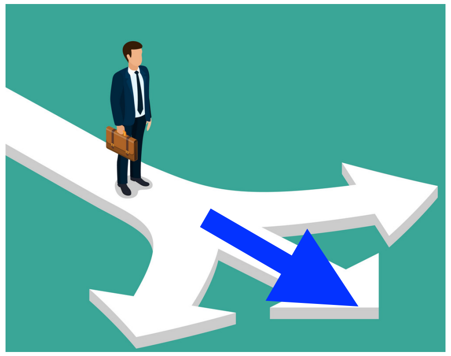
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Hyperparameter-free methods?

Importance sampling [Precup'00]

- Hyperparameter-free ✓
- No Markovianity required ✓

Data



Candidate

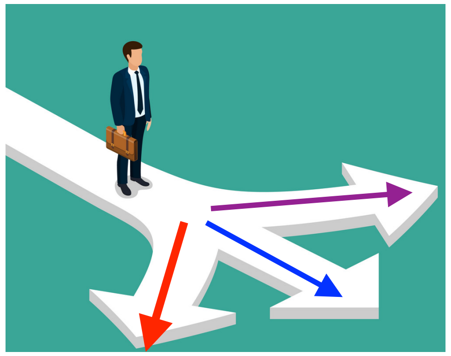
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Hyperparameter-free methods?

Importance sampling [Precup'00]

- Hyperparameter-free ✓
- No Markovianity required ✓

Data



Candidate

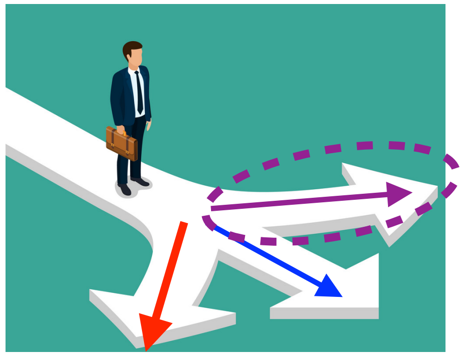
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Hyperparameter-free methods?

Importance sampling [Precup'00]

- Hyperparameter-free ✓
- No Markovianity required ✓

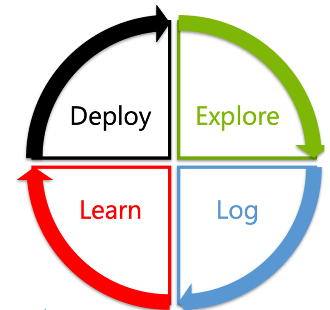
Data



Candidate

Precup. 2000. Eligibility traces for off-policy policy evaluation.

Hyperparameter-free methods?

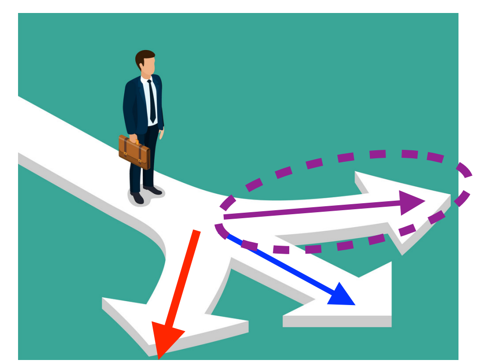


Azure

Importance sampling [Precup'00]

- Hyperparameter-free ✓
- No Markovianity required ✓
- Industry deployment (ctx. bandit, horizon=1)

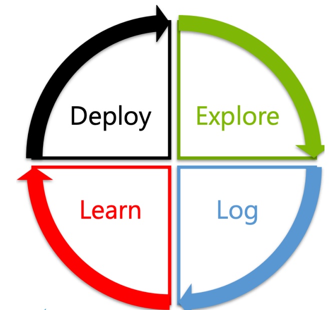
Data



Candidate

Precup. 2000. Eligibility traces for off-policy policy evaluation.

Hyperparameter-free methods?



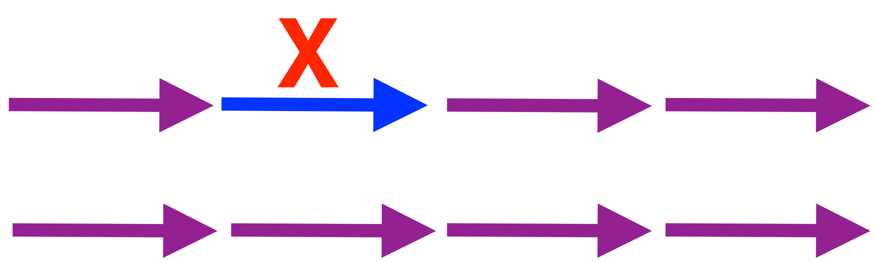
Azure

Importance sampling [Precup'00]

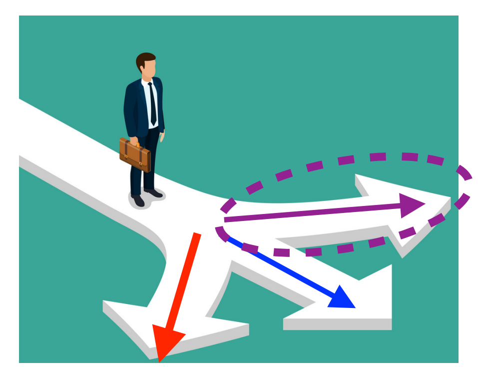
- Hyperparameter-free ✓
- No Markovianity required ✓
- Industry deployment (ctx. bandit, horizon=1)
- Exponential-in-horizon variance!

Data

Candidate



Data



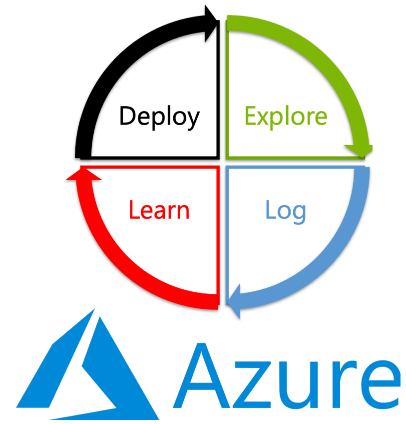
Candidate

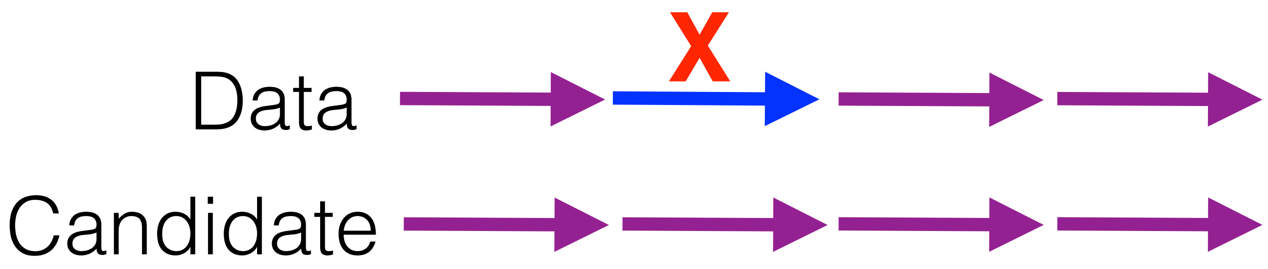
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Hyperparameter-free methods?

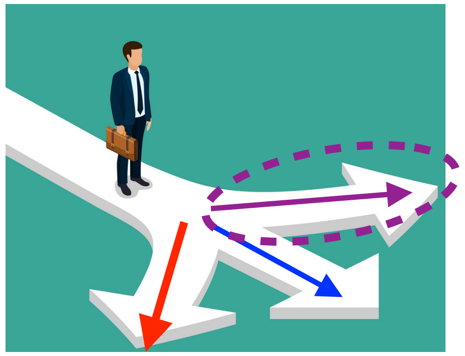
Importance sampling [Precup'00]

- Hyperparameter-free ✓
- No Markovianity required ✓
- Industry deployment (ctx. bandit, horizon=1)
- Exponential-in-horizon variance!
- Variance reduction?





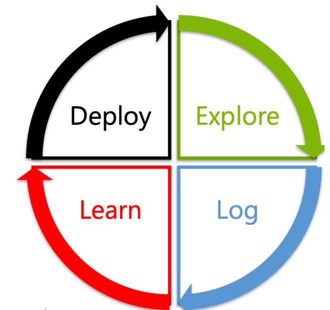
Data



Candidate

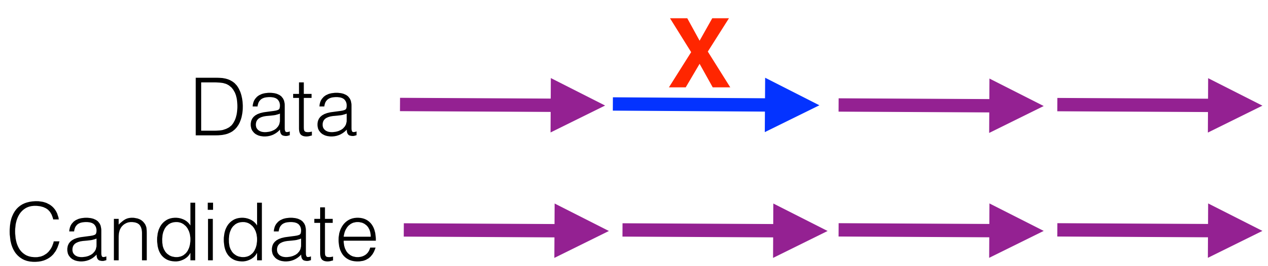
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Hyperparameter-free methods?

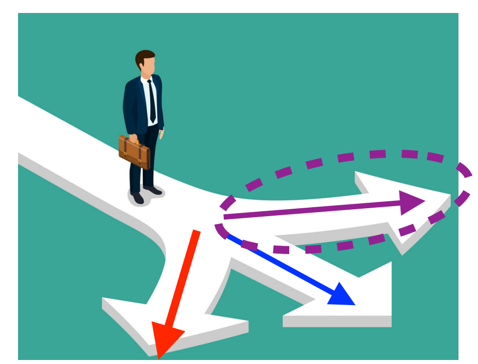


Importance sampling [Precup'00]

- Hyperparameter-free ✓
- No Markovianity required ✓
- Industry deployment (ctx. bandit, horizon=1)
- Exponential-in-horizon variance!
- Variance reduction?



Data

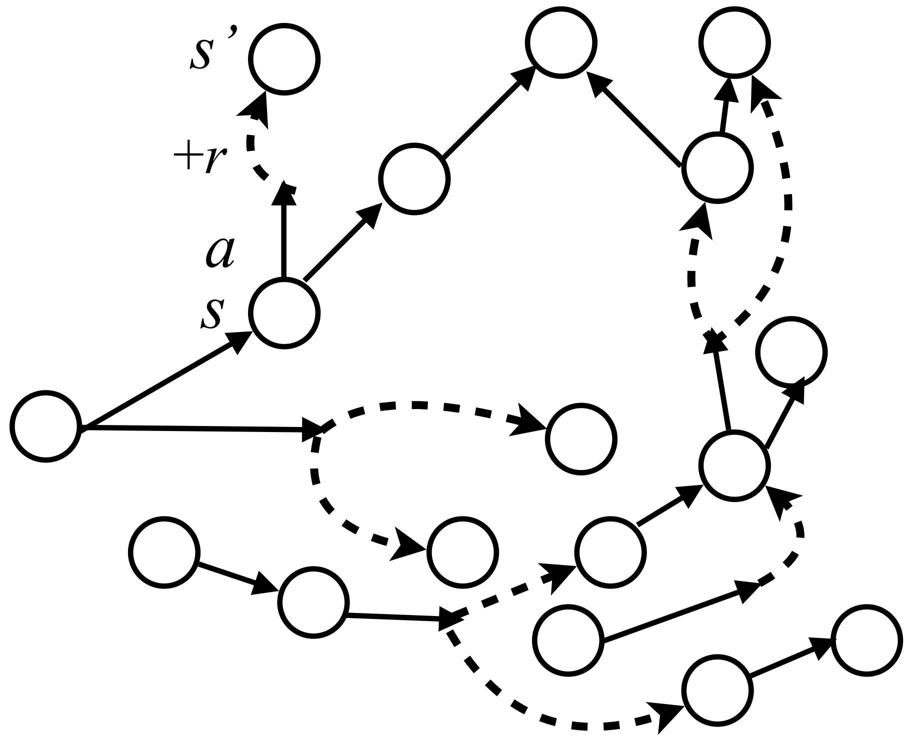


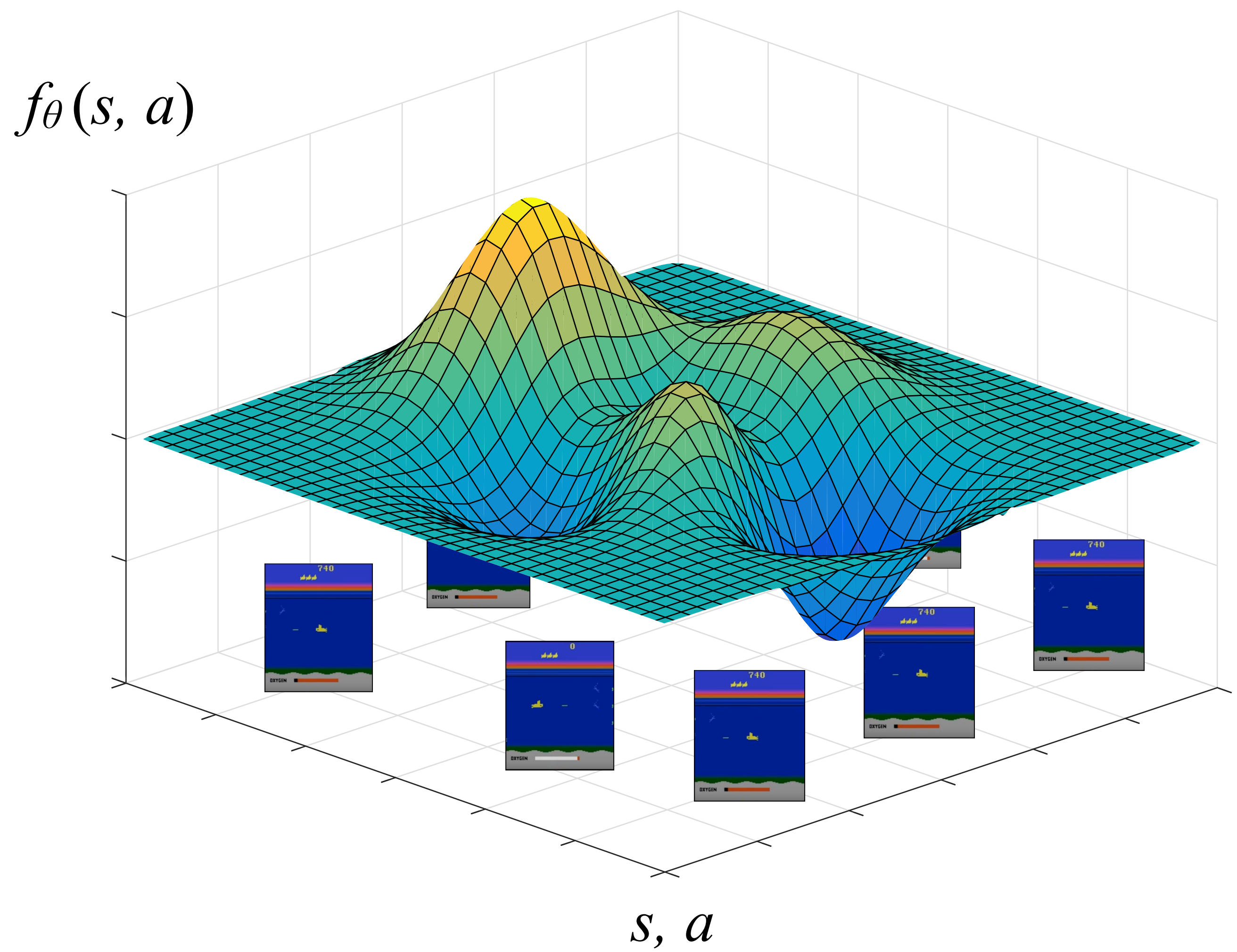
Candidate

Doubly robust [JL'16]

- Even perfect control variate cannot eliminate exponential variance!

Precup. 2000. Eligibility traces for off-policy policy evaluation.
Nan Jiang, Lihong Li. ICML-16. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning.



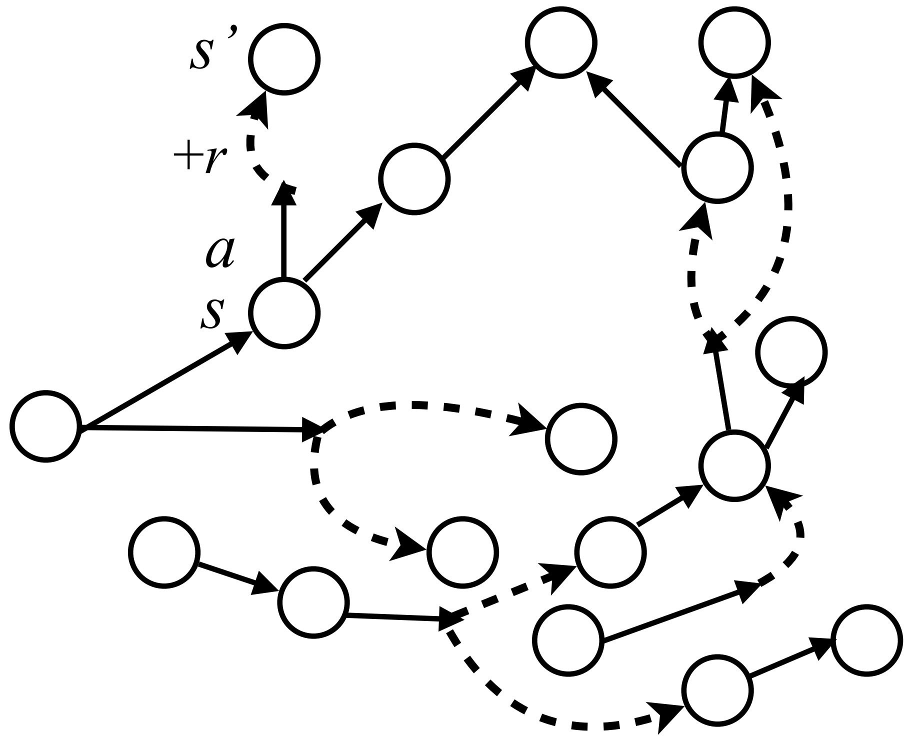


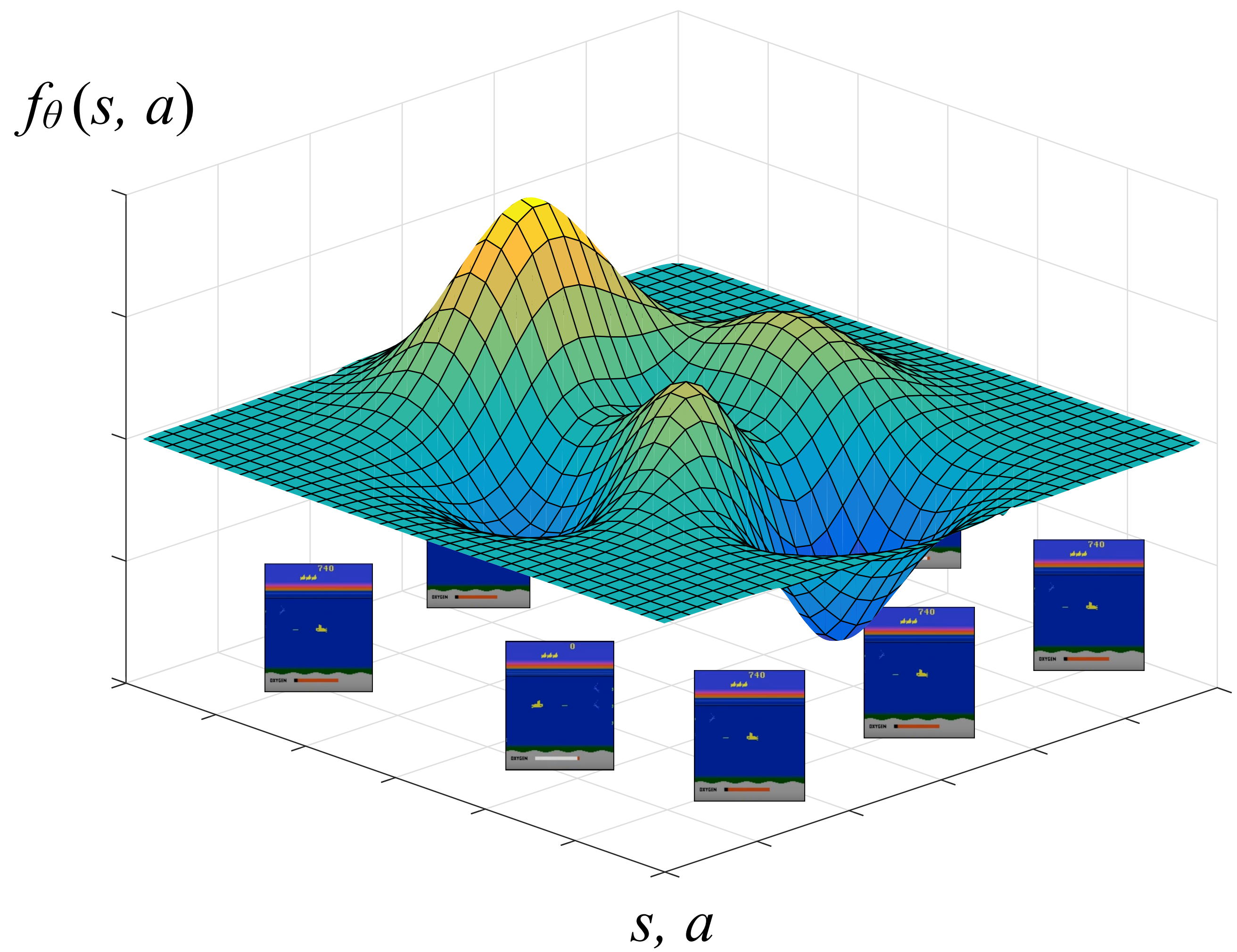
Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s', a'))^2 \right]$$

$$\boxed{\pi} = \text{greedy w.r.t. } \hat{f}$$





Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D\left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s', a')\right)^2\right]$$

$$\pi = \text{greedy w.r.t. } \boxed{\hat{f}}$$

Reformulation: Value-function Selection

Simple(?) Problem

- Run different training algorithms
- Get candidate value functions f_1, f_2, …
- Holdout data $\{(s, a, r, s')\}$

Reformulation: Value-function Selection

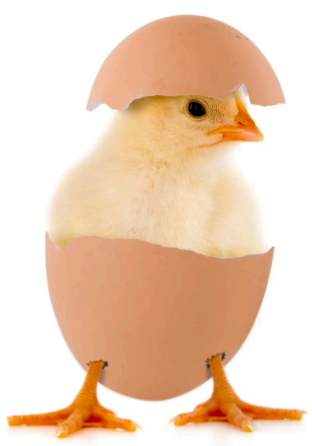
Simple(?) Problem

- Run different training algorithms
- Get candidate value functions f_1, f_2, …
- Holdout data $\{(s, a, r, s')\}$
- Select a good approx of Q^* w/ a "small" holdout dataset?
    - "small" = no |S| or exponential-in-horizon

Reformulation: Value-function Selection

Simple(?) Problem

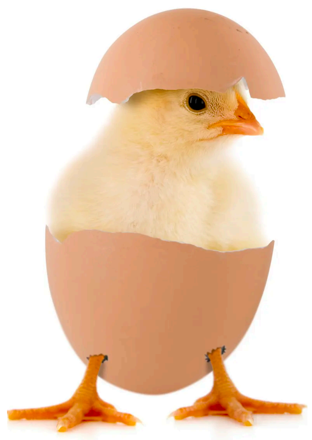
- Run different training algorithms
- Get candidate value functions f_1, f_2, …
- Holdout data $\{(s, a, r, s')\}$
- Select a good approx of Q^* w/ a "small" holdout dataset?
  - "small" = no |S| or exponential-in-horizon
  - & no further function approximation!



Reformulation: Value-function Selection

Simple(?) Problem

- Run different training algorithms
- Get candidate value functions f_1, f_2, …
- Holdout data $\{(s, a, r, s')\}$
- Select a good approx of Q^* w/ a "small" holdout dataset?
    - "small" = no |S| or exponential-in-horizon
    - & no further function approximation!
- Simpler: identify Q^* out of f_1, f_2



The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$

 Baird. 1995. Residual algorithms: Reinforcement learning with function approximation. * original paper concerns V^π

The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f\in\mathcal{F}} L(f)$

$\mathcal{F} = \{f_1, f_2, \ldots\}$

 Baird. 1995. Residual algorithms: Reinforcement learning with function approximation. * original paper concerns V^π

The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$
- RL doesn't work like that!

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

* exception: on-policy MC methods

The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$
- RL doesn't work like that!

Training: $\hat{f} = f_k$ where

optimization

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

* exception: on-policy MC methods

The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$
- RL doesn't work like that!

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

optimization

$\approx \mathcal{T} f_{k-1}$

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

iterative

* exception: on-policy MC methods

The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$
- RL doesn't work like that!

Training: $\hat{f} = f_k$ where

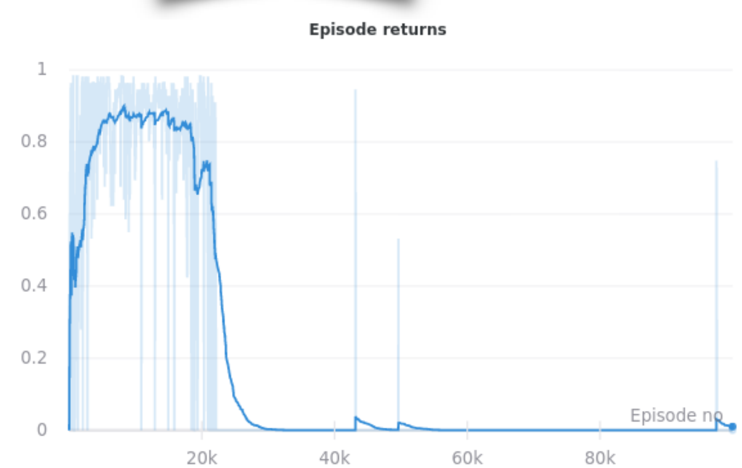
(FQI: learn Q^*)

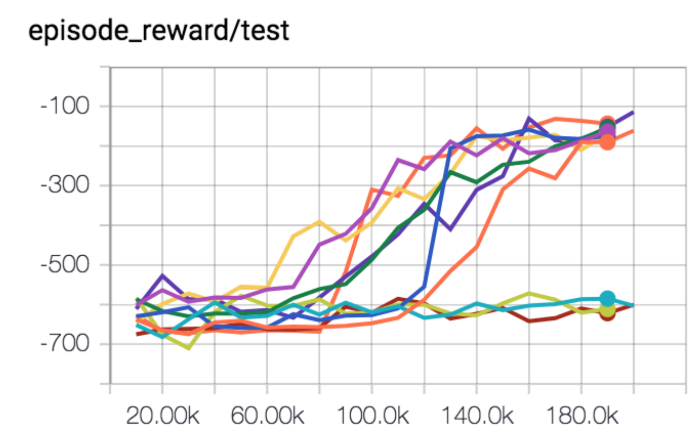
optimization

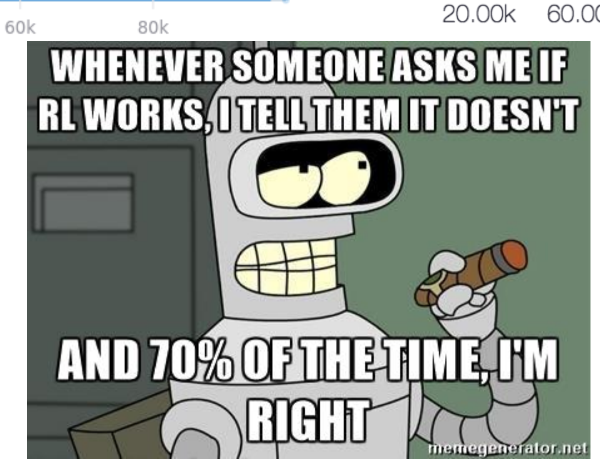
$\approx \mathcal{T} f_{k-1}$

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a') \right)^2 \right]$$

iterative







https://www.alexirpan.com/2018/02/14/rl-hard.html

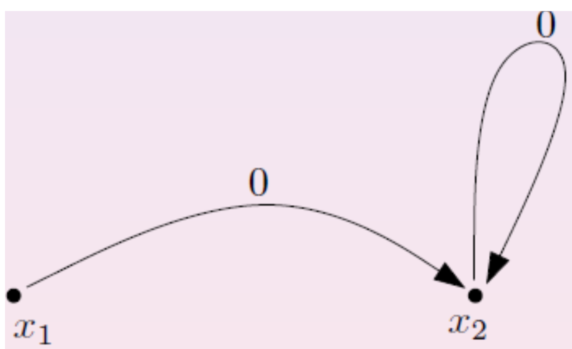
The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f\in\mathcal{F}} L(f)$
- RL doesn't work like that!

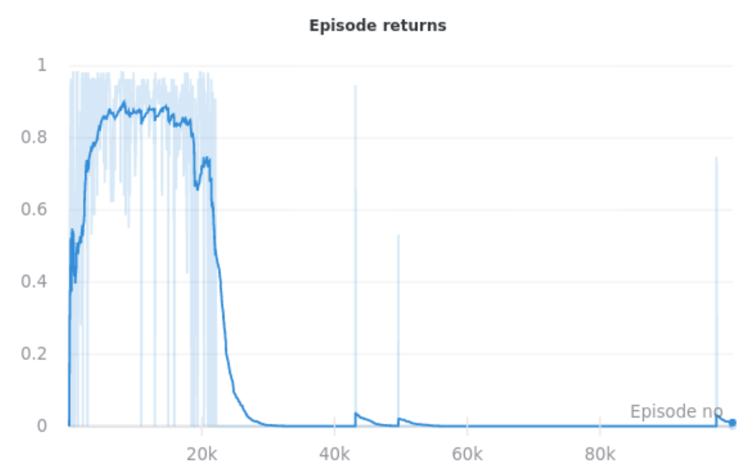
Training: $\hat{f} = f_k$ where $\approx \mathcal{T} f_{k-1}$

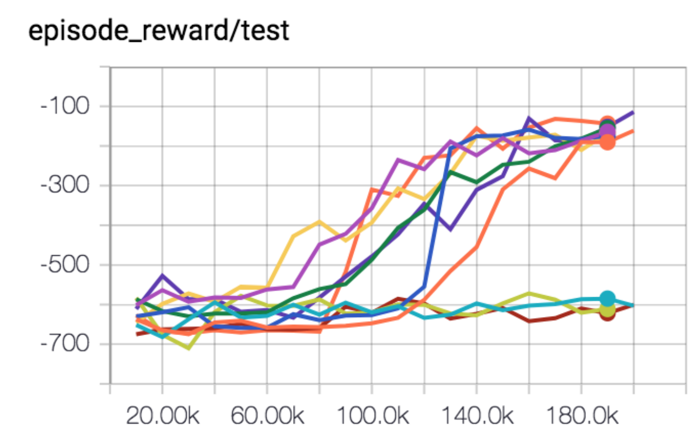
(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D\left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2\right]$$

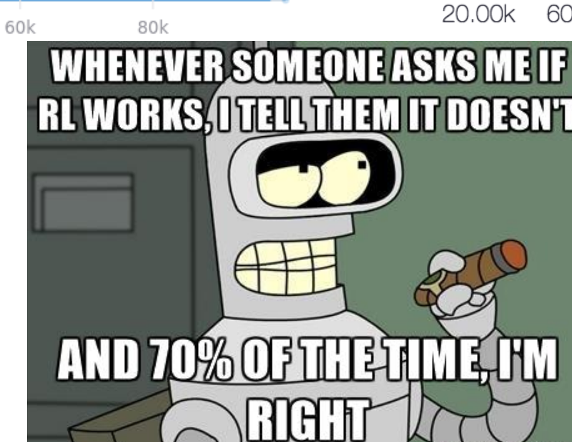


Divergence under 1-d linear
[TvR'96]





"Deadly triad"



https://www.alexirpan.com/2018/02/14/rl-hard.html

The training perspective

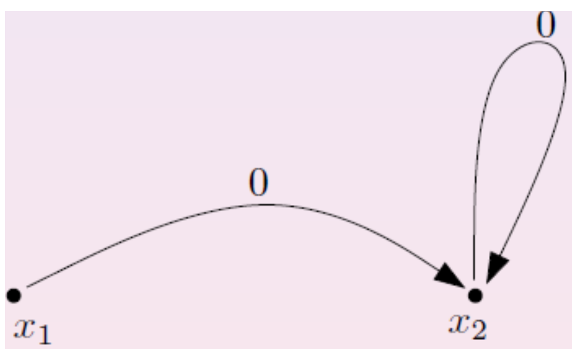
- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f\in\mathcal{F}} L(f)$
- RL doesn't work like that!

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

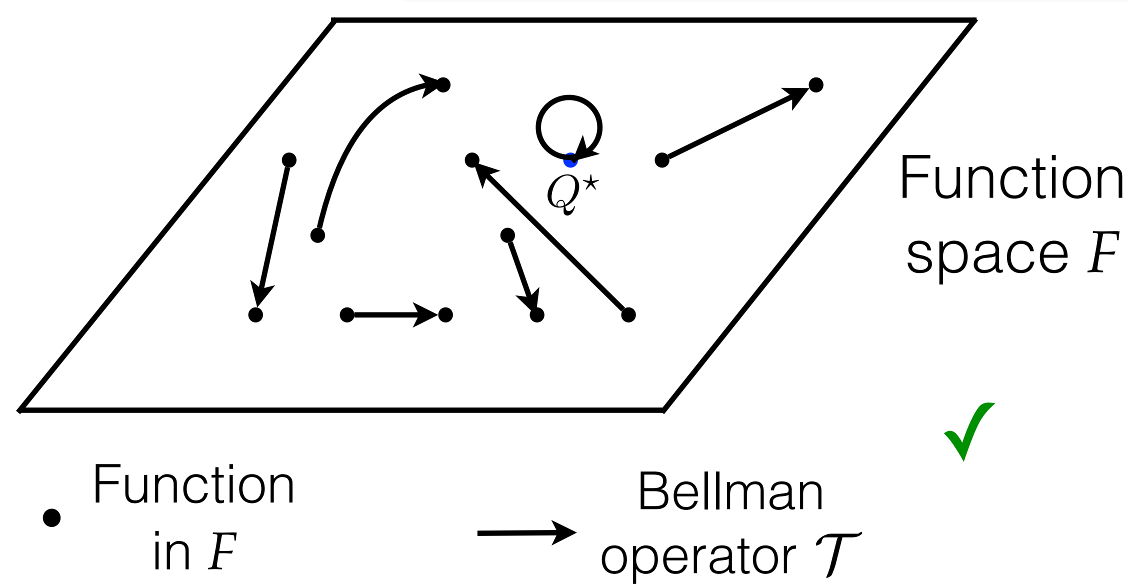
$\approx \mathcal{T} f_{k-1}$

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D\left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2\right]$$



Divergence under 1-d linear
[TvR'96]

"Bellman-completeness"
$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathcal{F}$



✓

The training perspective

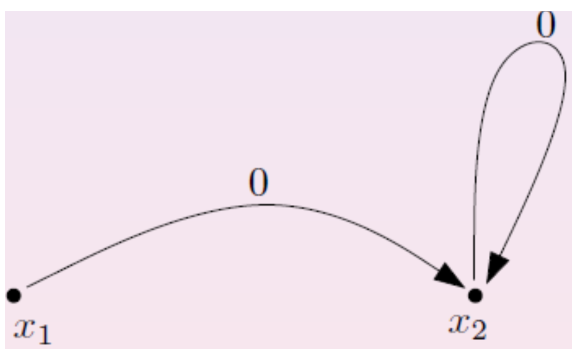
- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$
- RL doesn't work like that!

Training: $\hat{f} = f_k$ where

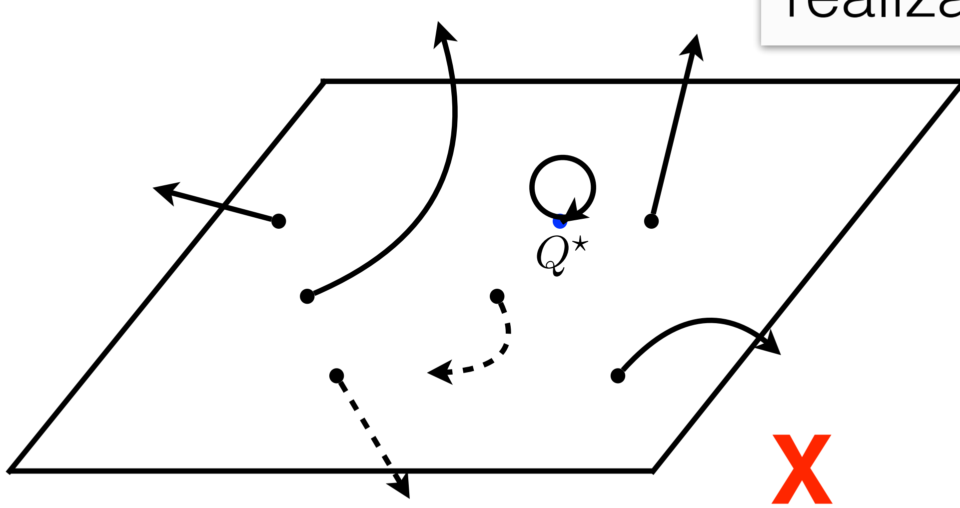
(FQI: learn Q^*)

$\approx \mathcal{T} f_{k-1}$

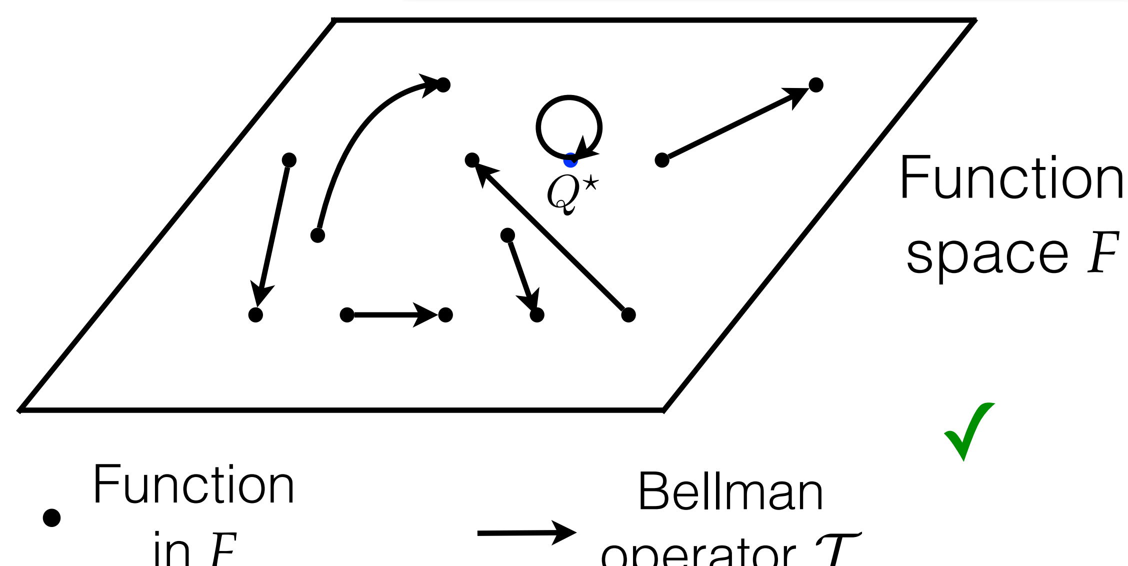
$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a') \right)^2 \right]$$



realizability (of Q^*)



"Bellman-completeness"
$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathcal{F}$



The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f\in\mathcal{F}} L(f)$

The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$
- $f = Q^\star \Leftrightarrow f = \mathcal{T} f$, so how about

$$f - \mathcal{T} f$$

The training perspective

- Baird'95: design L s.t. $Q^\star \stackrel{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$
- $f = Q^\star \Leftrightarrow f = \mathcal{T} f$, so how about

$$L = \mathbb{E}_D\left[(f - \mathcal{T} f)^2\right]$$

The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$
- $f = Q^\star \Leftrightarrow f = \mathcal{T} f$, so how about

$$L = \mathbb{E}_D \left[(f - \mathcal{T} f)^2 \right]$$

$$= \mathbb{E}_D \left[(f(s,a) - \mathbb{E}[r + \gamma \max_{a'} f(s', a') | s, a])^2 \right]$$

The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f \in \mathcal{F}} L(f)$
- $f = Q^\star \Leftrightarrow f = \mathcal{T} f$, so how about

$$L = \mathbb{E}_D \left[(f - \mathcal{T} f)^2\right]$$

$$= \mathbb{E}_D \left[(f(s,a) - \mathbb{E}[r + \gamma \max_{a'} f(s', a') | s, a])^2\right]$$

$$\neq$$

$$\mathbb{E}_D \left[(f(s,a) - (r + \gamma \max_{a'} f(s', a')))^2\right]$$

- Naive "1-sample" estimator is biased

* over-estimate by a Bayes-error-like term: $\mathbb{E}_{d^D}\left[\mathbb{V}_{s'|s,a}[r + \gamma \max_{a'} f(s', a')]\right]$

The training perspective

- Baird'95: design L s.t. $Q^\star \overset{?}{=} \arg\min_{f\in\mathcal{F}} L(f)$
- $f = Q^\star \Leftrightarrow f = \mathcal{T}f$, so how about

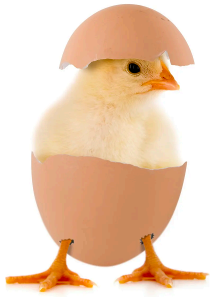
$$L = \mathbb{E}_D\left[(f - \mathcal{T}f)^2\right]$$

$$= \mathbb{E}_D\left[(f(s,a) - \mathbb{E}[r + \gamma \max_{a'} f(s',a') | s,a])^2\right]$$

$$\neq$$

$$\mathbb{E}_D\left[(f(s,a) - (r + \gamma \max_{a'} f(s',a')))^2\right]$$

- Naive "1-sample" estimator is biased
  - debasing requires simulator ("*double sampling*" [Baird'95])
  - or, helper class $\mathcal{F}' \ni \mathcal{T}f$ [ASM'08, FS'10]



* over-estimate by a Bayes-error-like term: $\mathbb{E}_{d^D}\left[\mathbb{V}_{s'|s,a}[r + \gamma \max_{a'} f(s',a')]\right]$

Basis of resolution

Our goal

To select b/t f_1, f_2

Gordan. 1995. Stable Function Approximation in Dynamic Programming.

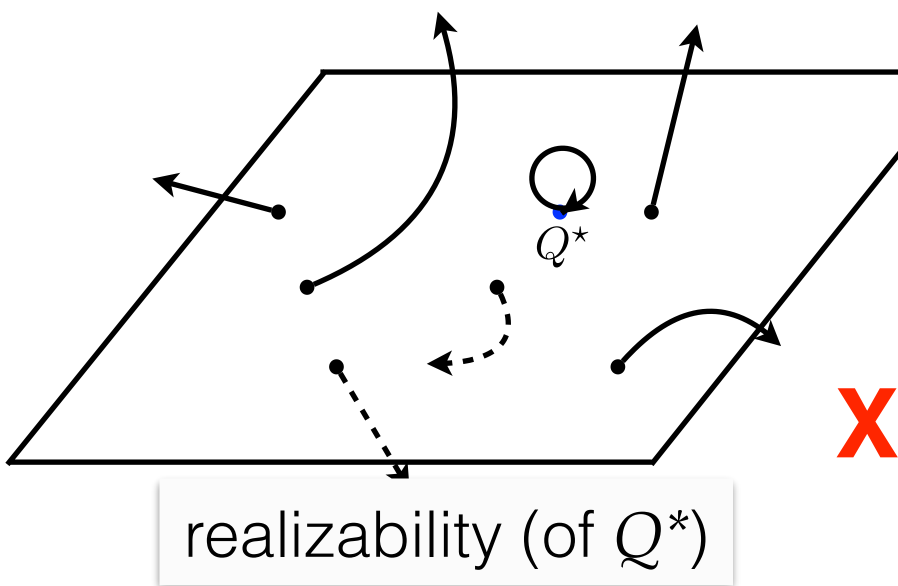
Basis of resolution

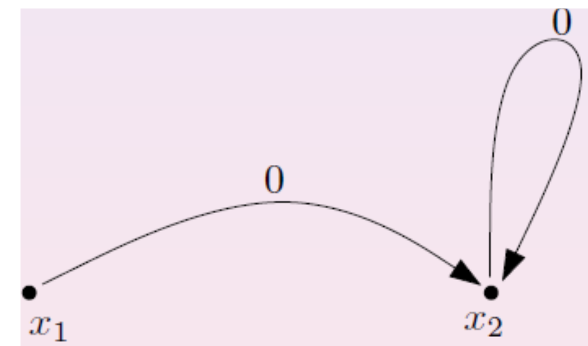
Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta \in \mathcal{F}} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

iterative





Divergence under 1-d linear $\mathcal{F}$

[TvR'96]

To select b/t f_1, f_2

Gordan. 1995. Stable Function Approximation in Dynamic Programming.

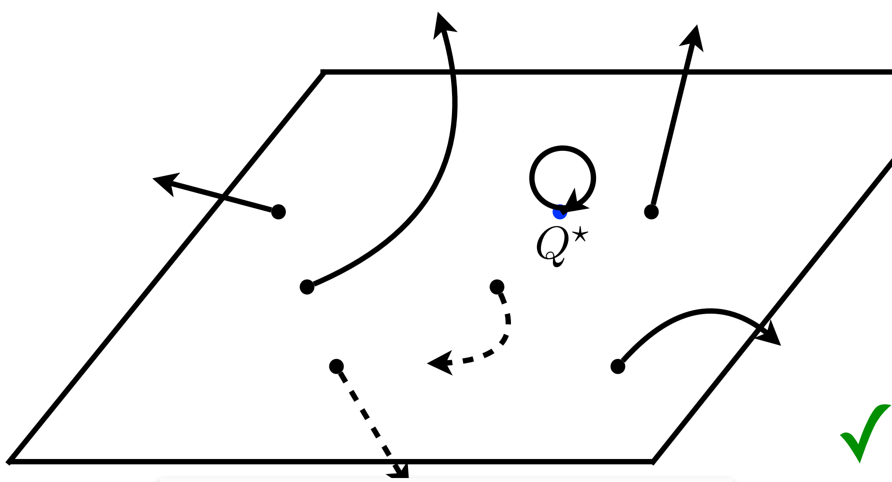
Basis of resolution

Training: $\hat{f} = f_k$ where

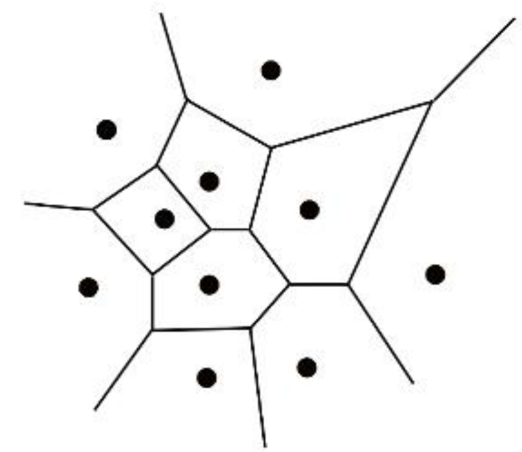
(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta \in \mathcal{F}} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a') \right)^2 \right]$$

iterative



realizability (of Q^*)



Convergence under piecewise constant $\mathcal{F}$! [Gordon'95]

To select b/t f_1, f_2

Gordan. 1995. Stable Function Approximation in Dynamic Programming.

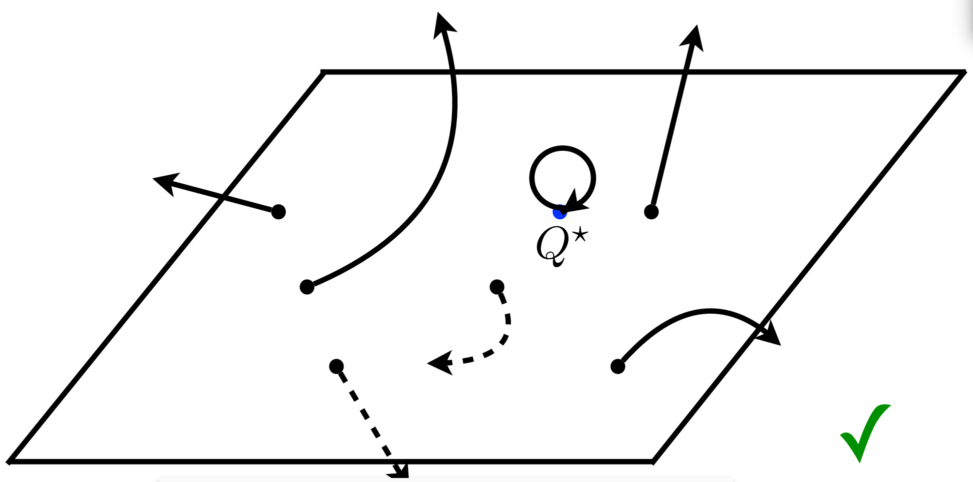
Basis of resolution

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

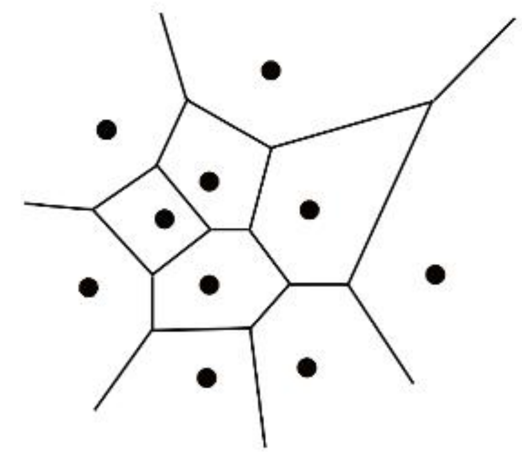
$$f_k \leftarrow \arg\min_{f_\theta \in \mathcal{F}} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

iterative



realizability (of Q^*)

✓



Convergence under piecewise constant $\mathcal{F}$! [Gordon'95]

same

To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant
- can express Q^*
- small # partitions (bounded complexity)

Gordan. 1995. Stable Function Approximation in Dynamic Programming.

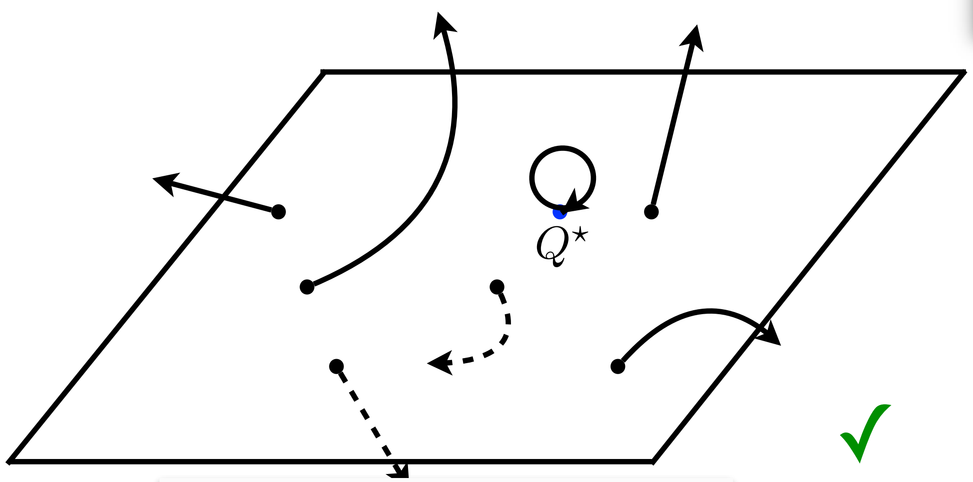
Basis of resolution

Training: $\hat{f} = f_k$ where

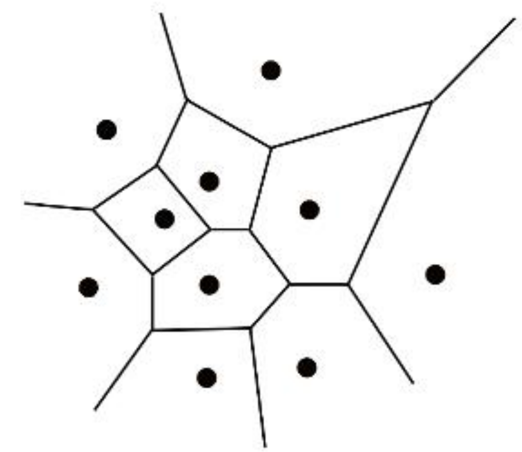
(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta \in \mathcal{F}} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

iterative



realizability (of Q^*) ✓



Convergence under piecewise constant $\mathcal{F}$! [Gordon'95]

same

To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant
- can express Q^*
- small # partitions (bounded complexity)

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T}f)\|_{2,D}$

Gordan. 1995. Stable Function Approximation in Dynamic Programming.

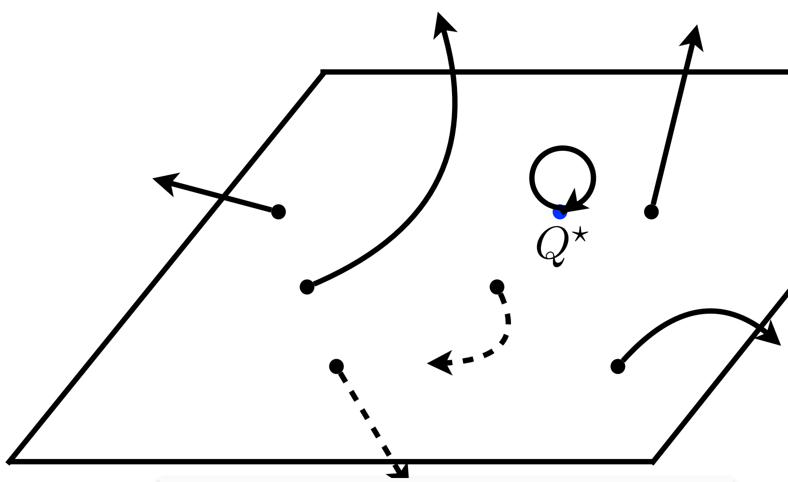
Basis of resolution

Training: $\hat{f} = f_k$ where

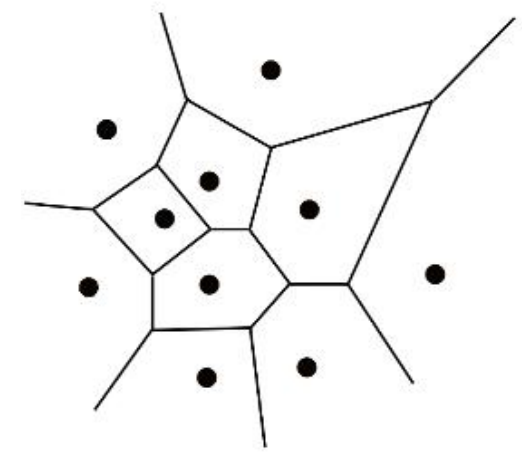
(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta \in \mathcal{F}} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

iterative

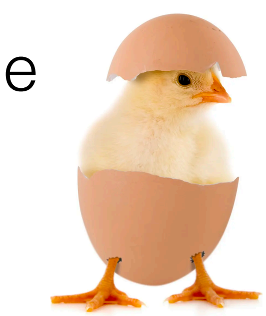


realizability (of Q^*) ✓



Convergence under piecewise constant $\mathcal{F}$! [Gordon'95]

same



To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant
- can express Q^*
- small # partitions (bounded complexity)

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T} f)\|_{2,D}$

Gordan. 1995. Stable Function Approximation in Dynamic Programming.

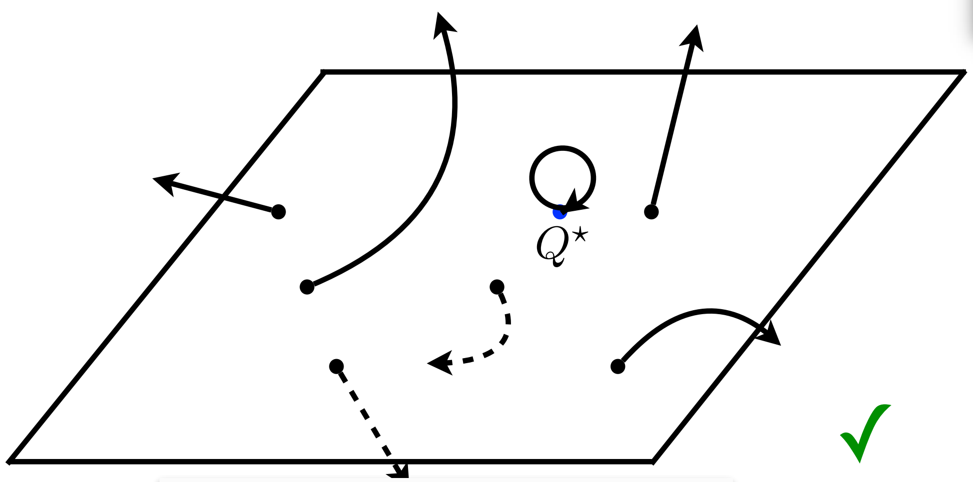
Basis of resolution

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

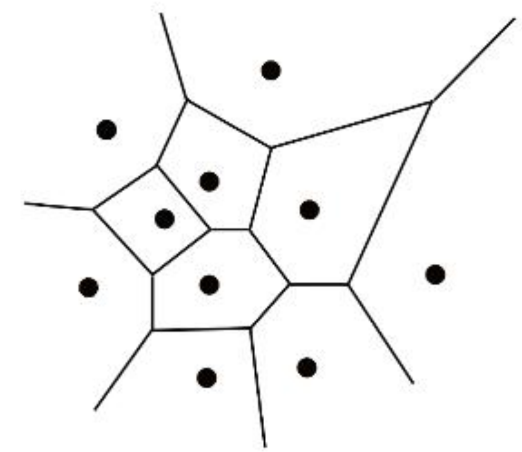
$$f_k \leftarrow \arg\min_{f_\theta \in \mathcal{F}} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a') \right)^2 \right]$$

iterative

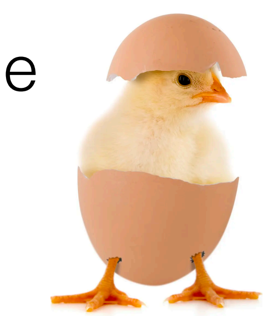


realizability (of Q^*)

✓



Convergence under piecewise constant $\mathcal{F}$! [Gordon'95]



same

To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant
- can express Q^*
- small # partitions (bounded complexity)

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T} f)\|_{2,D}$

Our method: create such a magical G "out of nothing"!

Gordan. 1995. Stable Function Approximation in Dynamic Programming.

Does a magical G always exist?

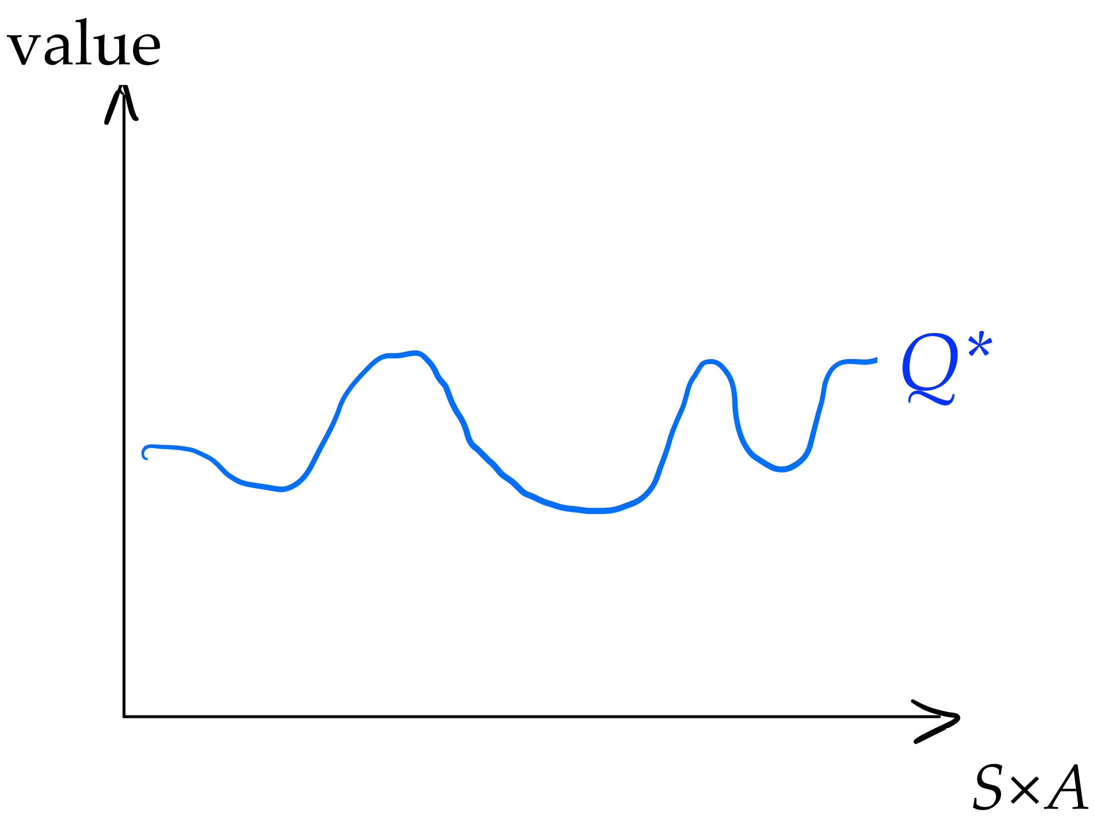
To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant
- can express Q^*
- small # partitions (bounded complexity)

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T} f)\|_{2,D}$

Does a magical G always exist?

- YES! Just partition SxA according to **output** of Q^*



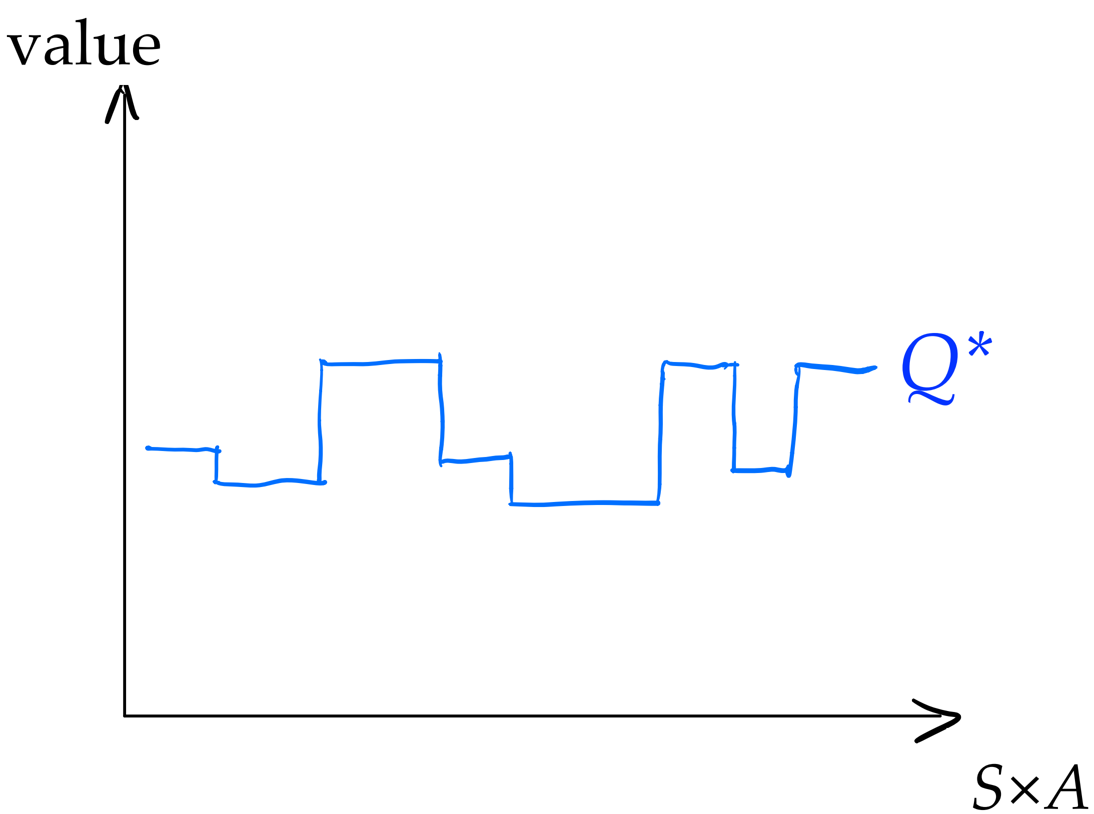
To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant
- can express Q^*
- small # partitions (bounded complexity)

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T}f)\|_{2,D}$

Does a magical G always exist?

- YES! Just partition SxA according to **output** of Q^*



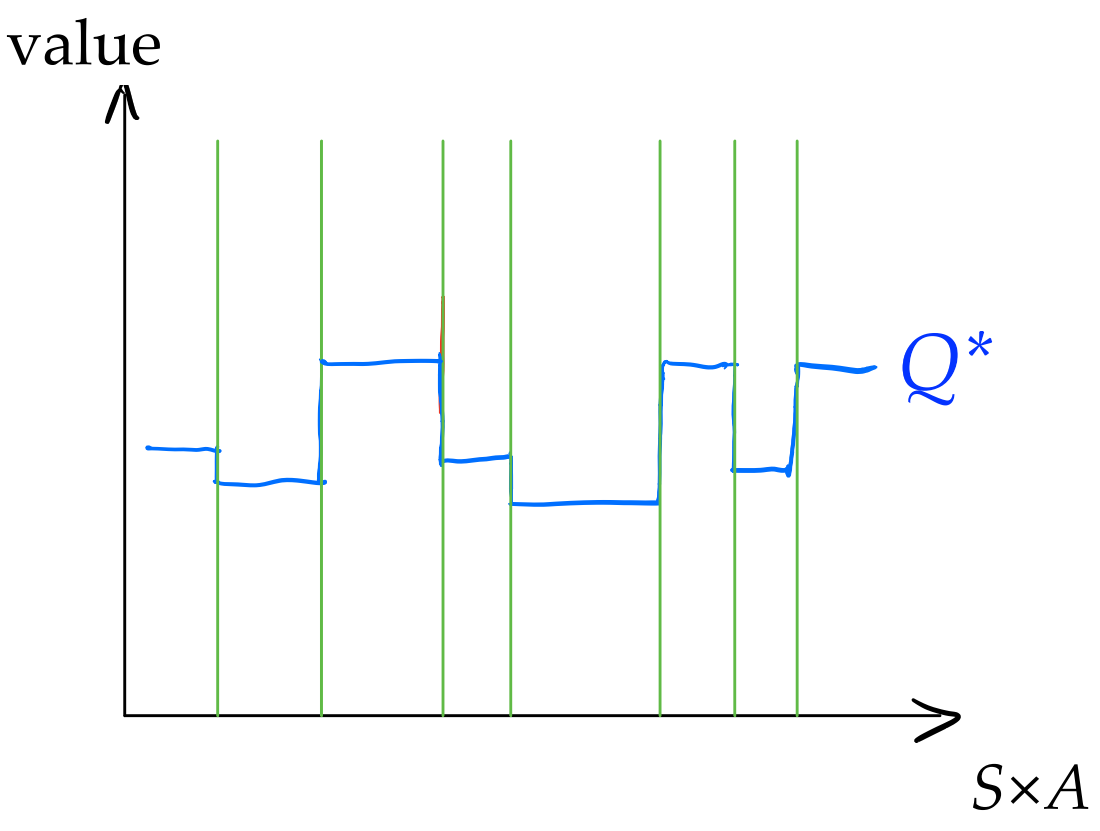
To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant
- can express Q^*
- small # partitions (bounded complexity)

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T}f)\|_{2,D}$

Does a magical G always exist?

- YES! Just partition SxA according to **output** of Q^*



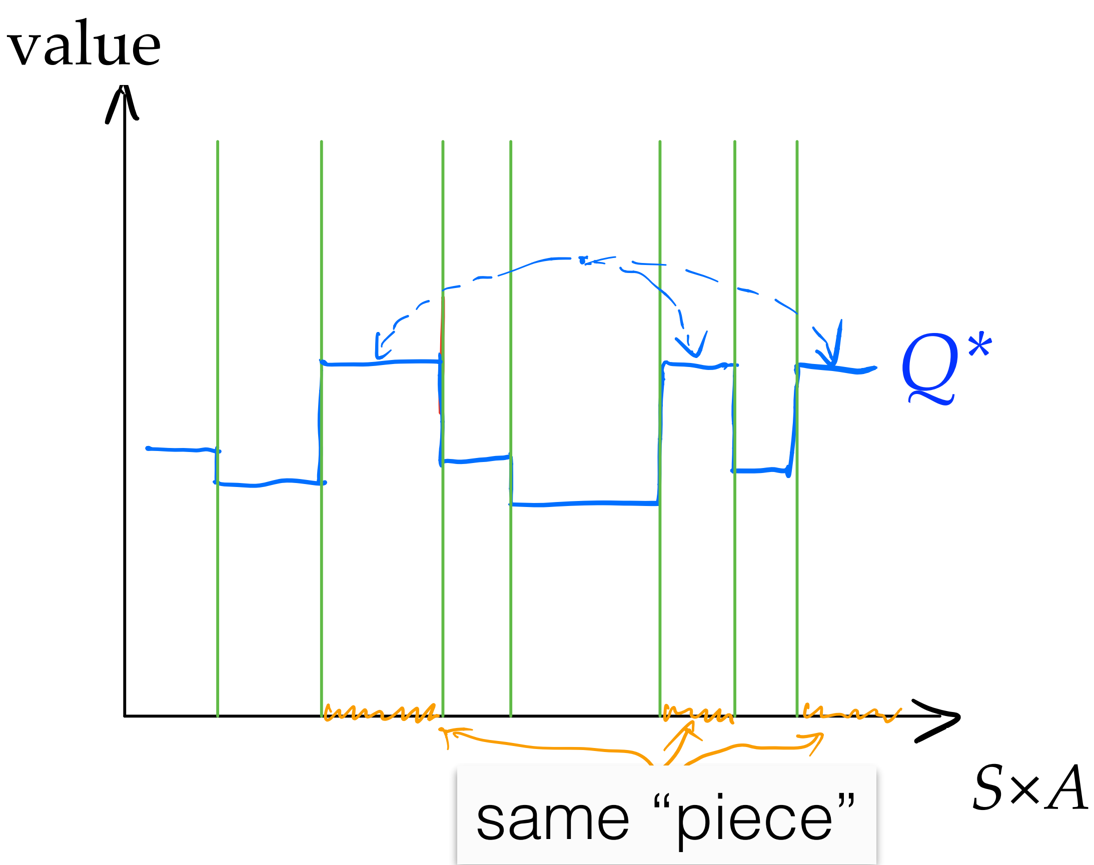
To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant ✓
- can express Q^* ✓
- small # partitions (bounded complexity)

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T} f)\|_{2,D}$

Does a magical G always exist?

- YES! Just partition SxA according to **output** of Q^*



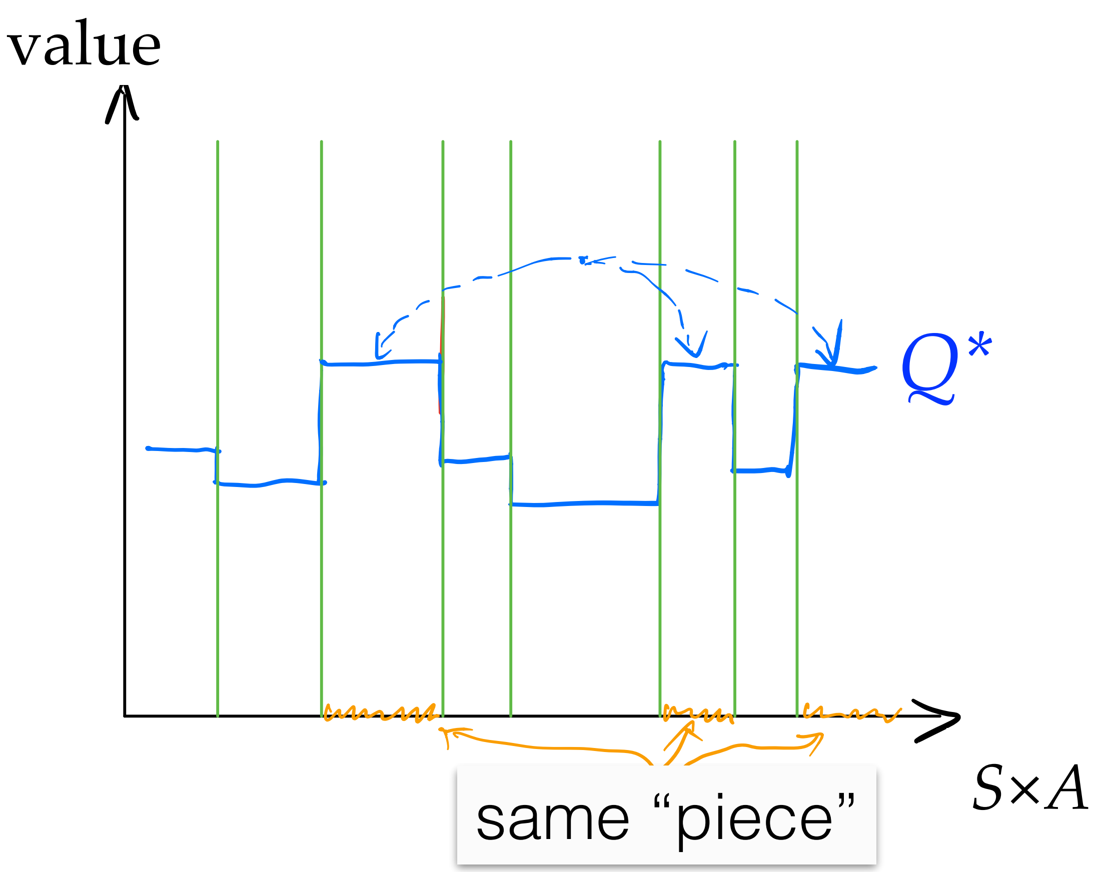
To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant ✓
- can express Q^* ✓
- small # partitions (bounded complexity)

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T} f)\|_{2,D}$

Does a magical G always exist?

- YES! Just partition SxA according to **output** of Q^*



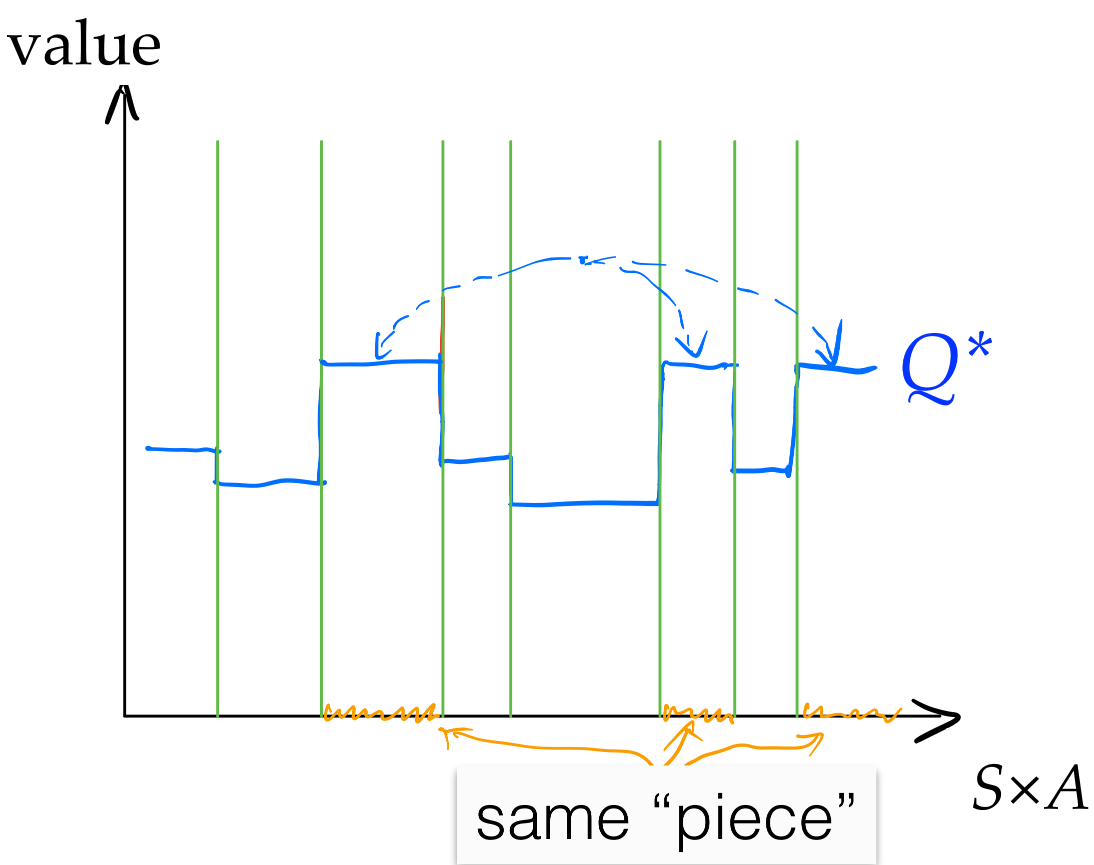
To select b/t f_1, f_2, suffices to have class G s.t.

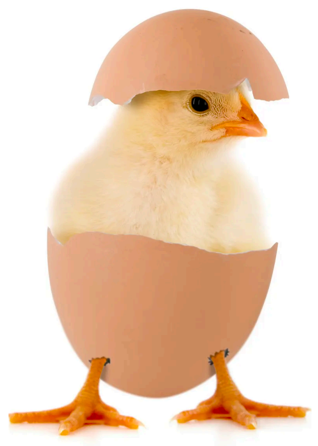
- piecewise constant ✓
- can express Q^* ✓
- **$O(1/\varepsilon)$** partitions (bounded complexity) ✓

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T} f)\|_{2,D}$

Does a magical G always exist?

- YES! Just partition SxA according to **output** of Q^*



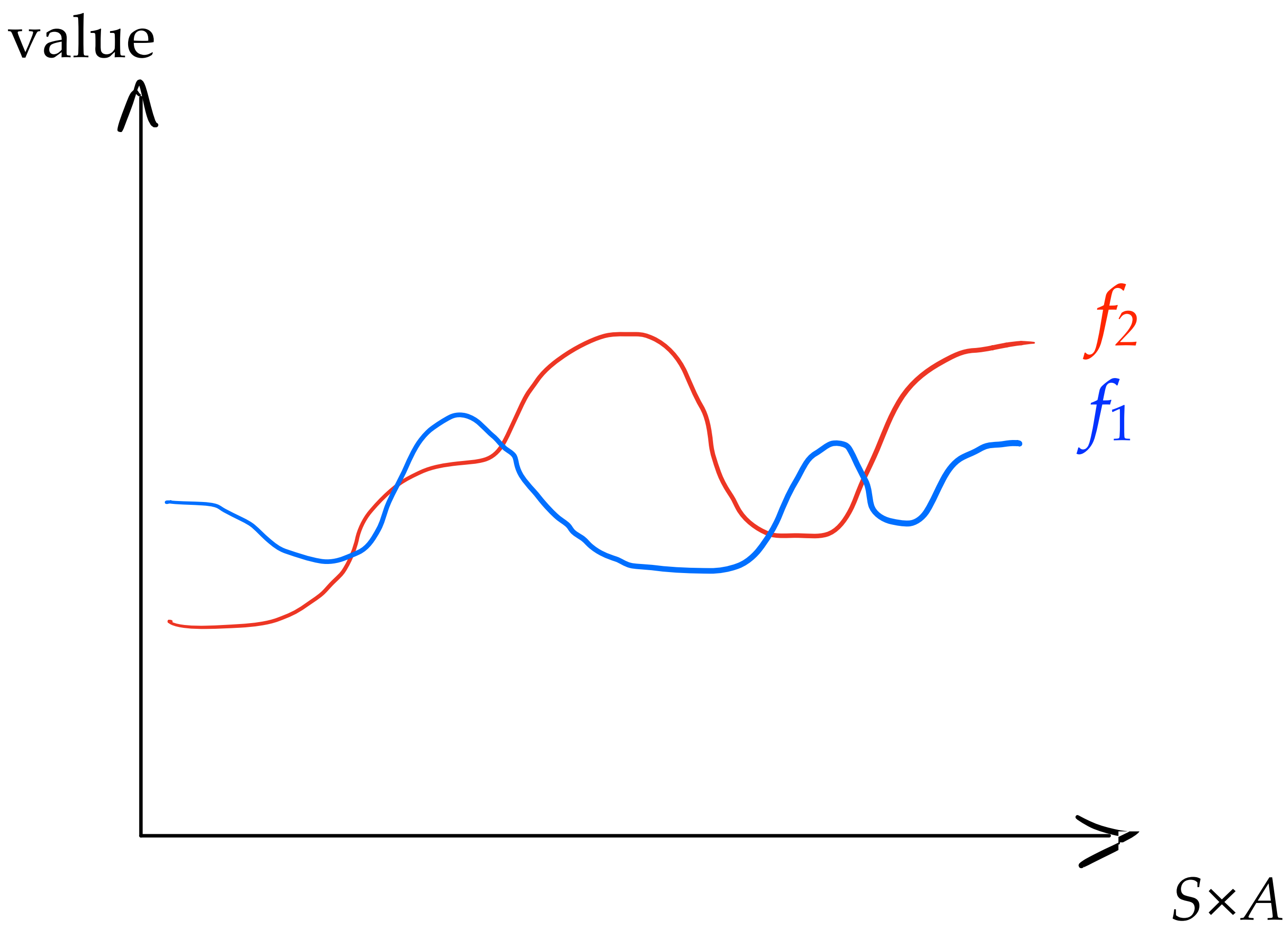


To select b/t f_1, f_2, suffices to have class G s.t.

- piecewise constant ✓
- can express Q^* ✓
- **$O(1/\varepsilon)$** partitions (bounded complexity) ✓

Then: minimize $\|f - \mathrm{Proj}_{\mathcal{G}}(\mathcal{T}f)\|_{2,D}$

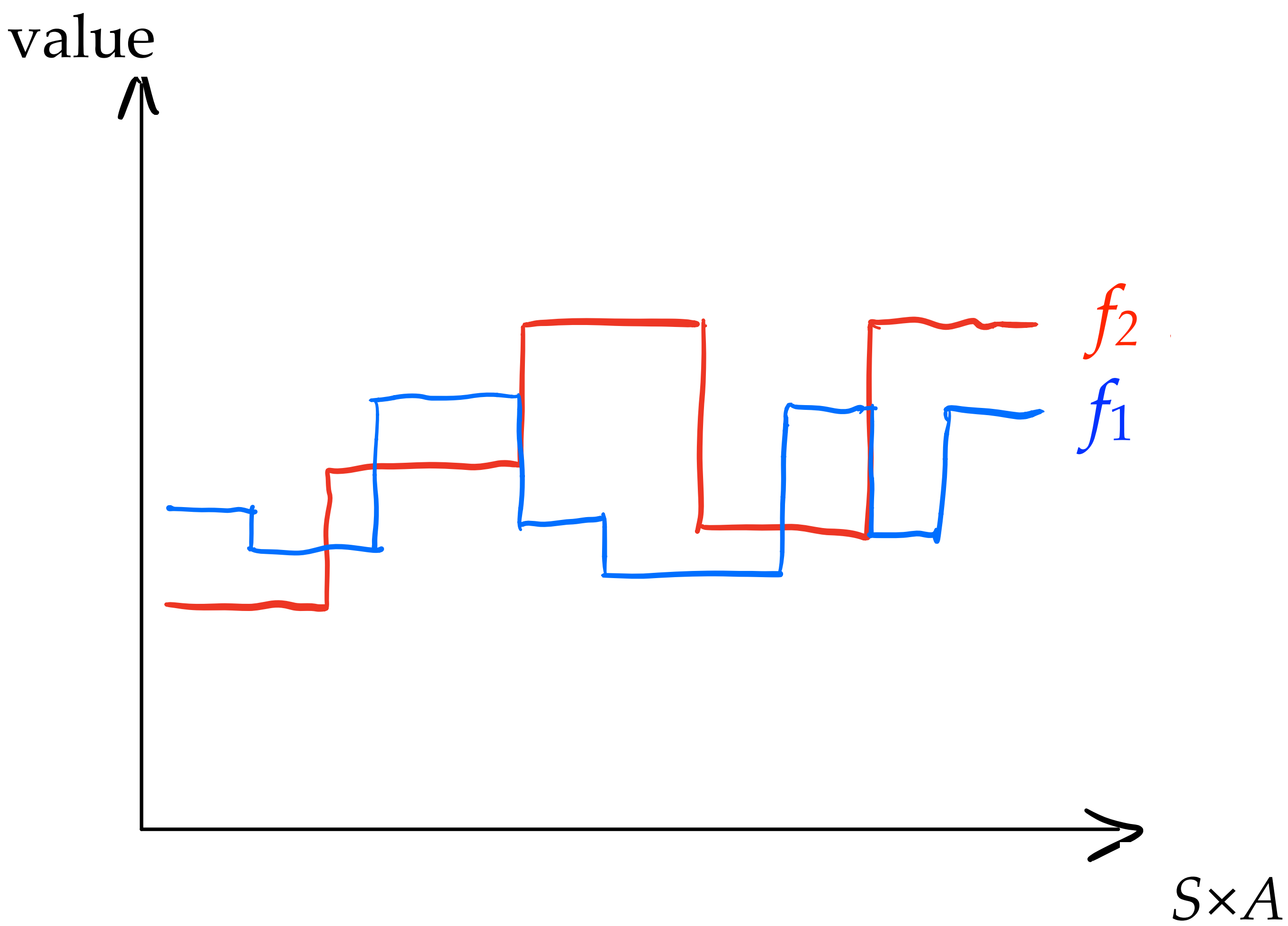
Batch Value-Function Tournament [XJ, ICML-21]



- (Simplified) problem: identify Q^* out of $F = \{ f_1, f_2 \}$

Xie & **Jiang**. ICML-21. Batch Value-function Approximation with Only Realizability.

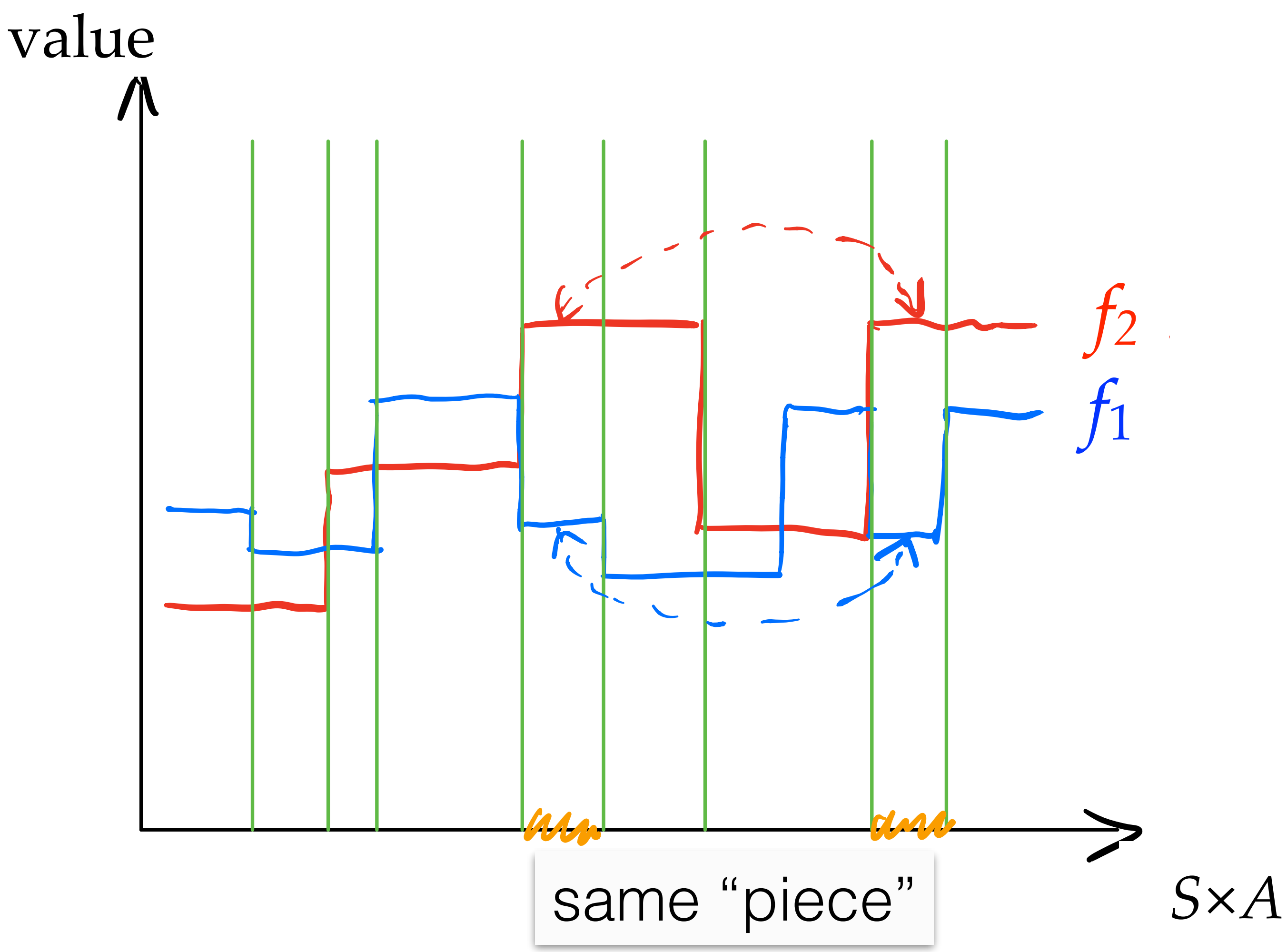
Batch Value-Function Tournament [XJ, ICML-21]



- (Simplified) problem: identify Q^* out of $F = \{ f_1, f_2 \}$
- Partition SxA according to both functions simultaneously!

Xie & **Jiang**. ICML-21. Batch Value-function Approximation with Only Realizability.

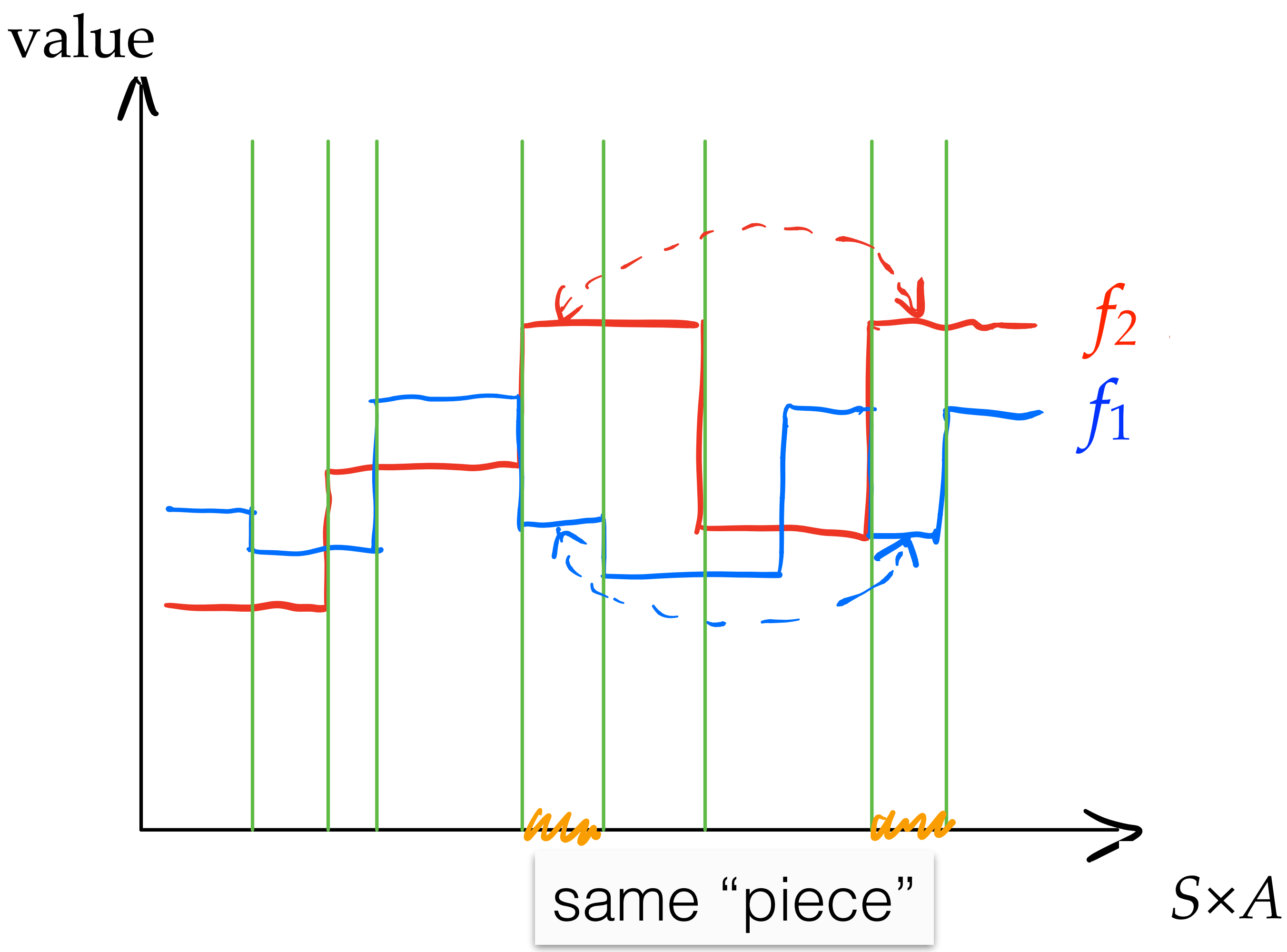
Batch Value-Function Tournament [XJ, ICML-21]



- (Simplified) problem: identify Q^* out of $F = \{ f_1, f_2 \}$
- Partition SxA according to both functions simultaneously!

Xie & **Jiang**. ICML-21. Batch Value-function Approximation with Only Realizability.

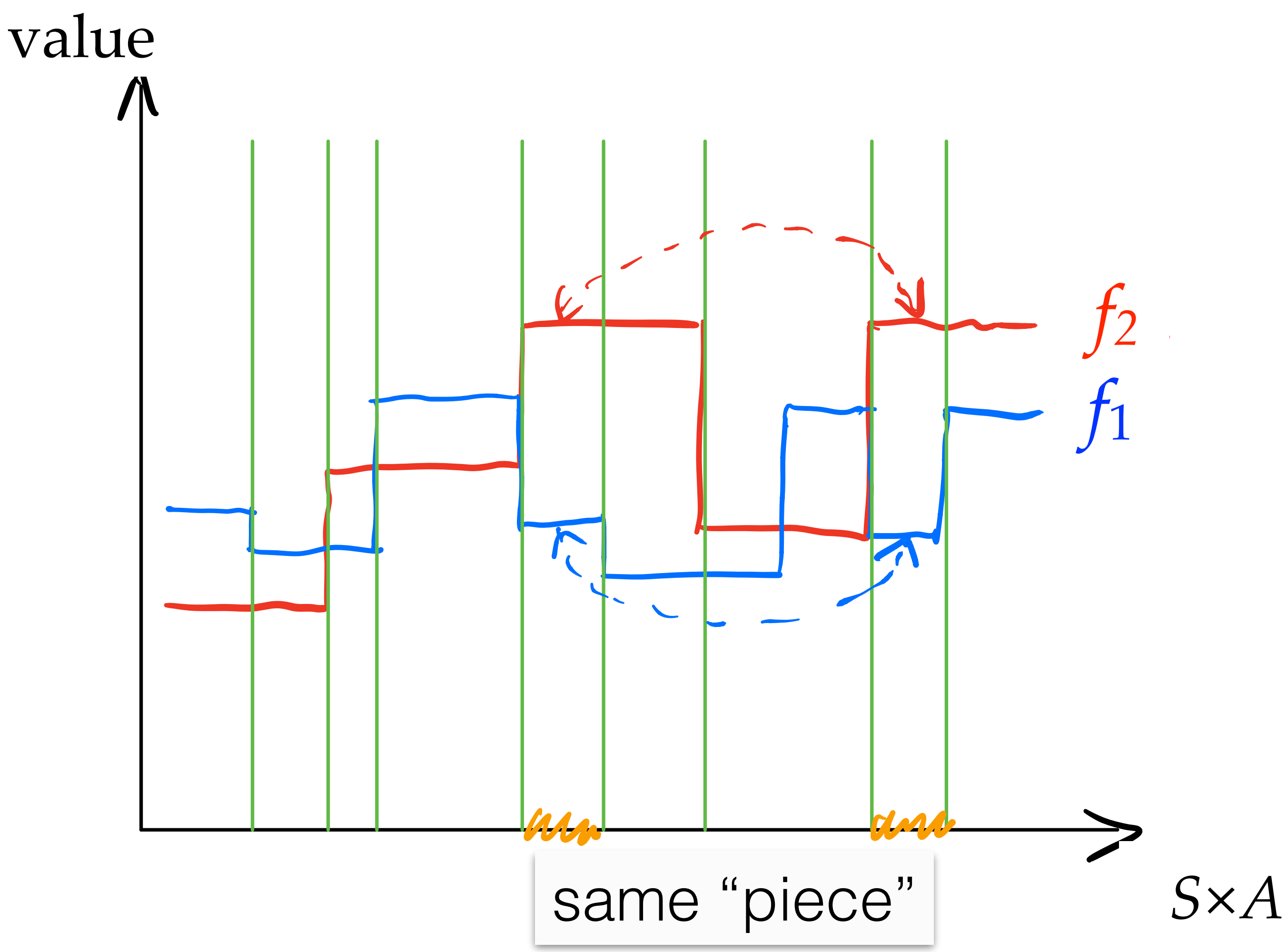
Batch Value-Function Tournament [XJ, ICML-21]



- (Simplified) problem: identify Q^* out of $F = \{ f_1, f_2 \}$
- Partition SxA according to both functions simultaneously!
  - Pw-const class $G_{1,2}$ w/ size **$O(1/\varepsilon^2)$** **!!**

Xie & **Jiang**. ICML-21. Batch Value-function Approximation with Only Realizability.

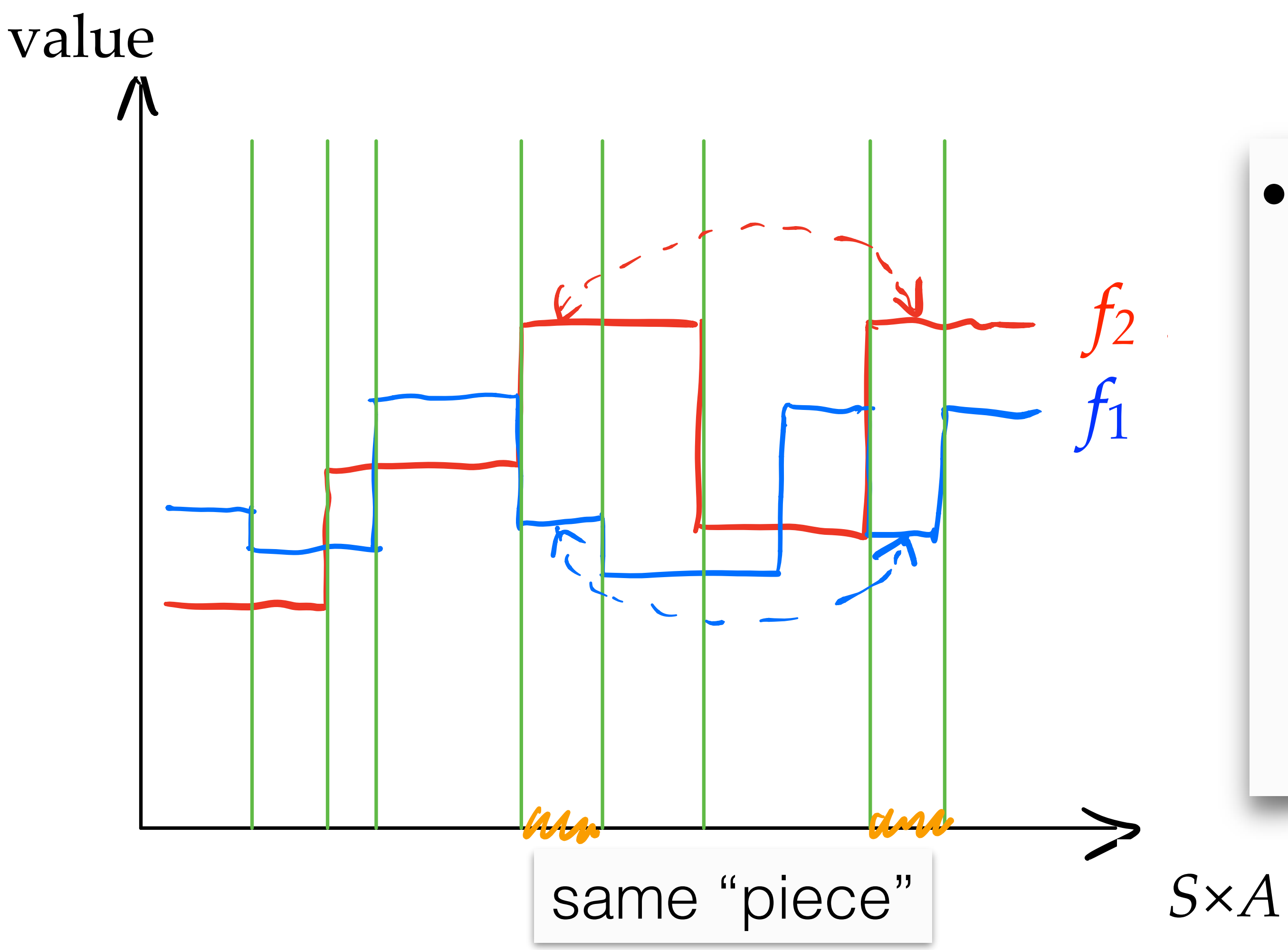
Batch Value-Function Tournament [XJ, ICML-21]



- (Simplified) problem: identify Q^* out of $F = \{ f_1, f_2 \}$
- Partition SxA according to both functions simultaneously!
  - Pw-const class $G_{1,2}$ w/ size **$O(1/\varepsilon^2)$** **!!**
- Naive extension to >2 functions in F: $O(1/\varepsilon^{|F|})$

Xie & **Jiang**. ICML-21. Batch Value-function Approximation with Only Realizability.

Batch Value-Function Tournament [XJ, ICML-21]



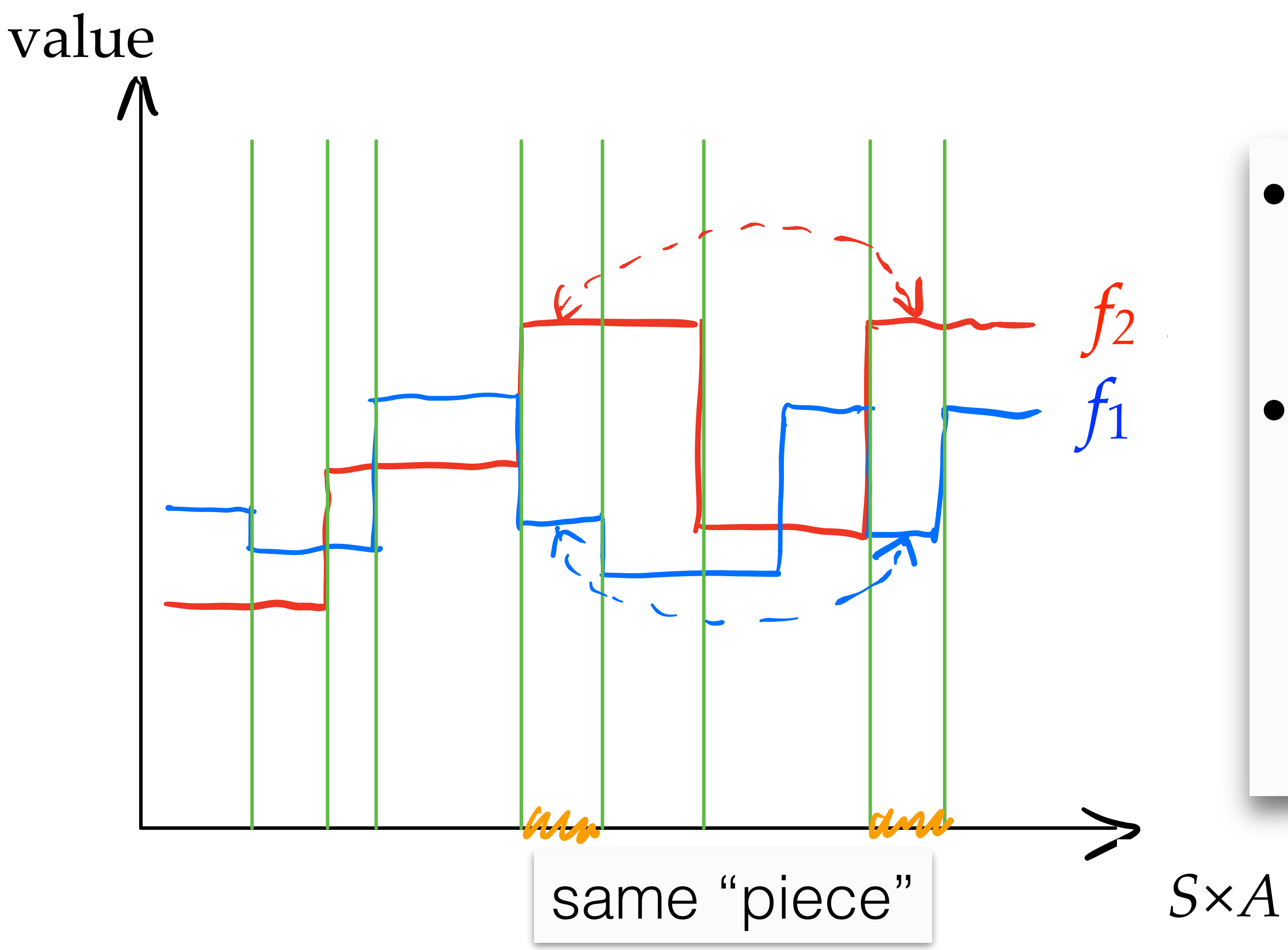
- Algorithm: **BVFT**

$$\arg\min_i \max_j \|f_i - \mathrm{Proj}_{\mathcal{G}_{i,j}}(\mathcal{T} f_i)\|_{2,D}$$

- (Simplified) problem: identify Q^* out of $F = \{ f_1, f_2 \}$
- Partition SxA according to both functions simultaneously!
  - Pw-const class $G_{1,2}$ w/ size **$O(1/\varepsilon^2)$** **!!**
- Naive extension to >2 functions in F: $O(1/\varepsilon^{|F|})$
  - Pairwise comparison + tournament

Xie & **Jiang**. ICML-21. Batch Value-function Approximation with Only Realizability.

Batch Value-Function Tournament [XJ, ICML-21]



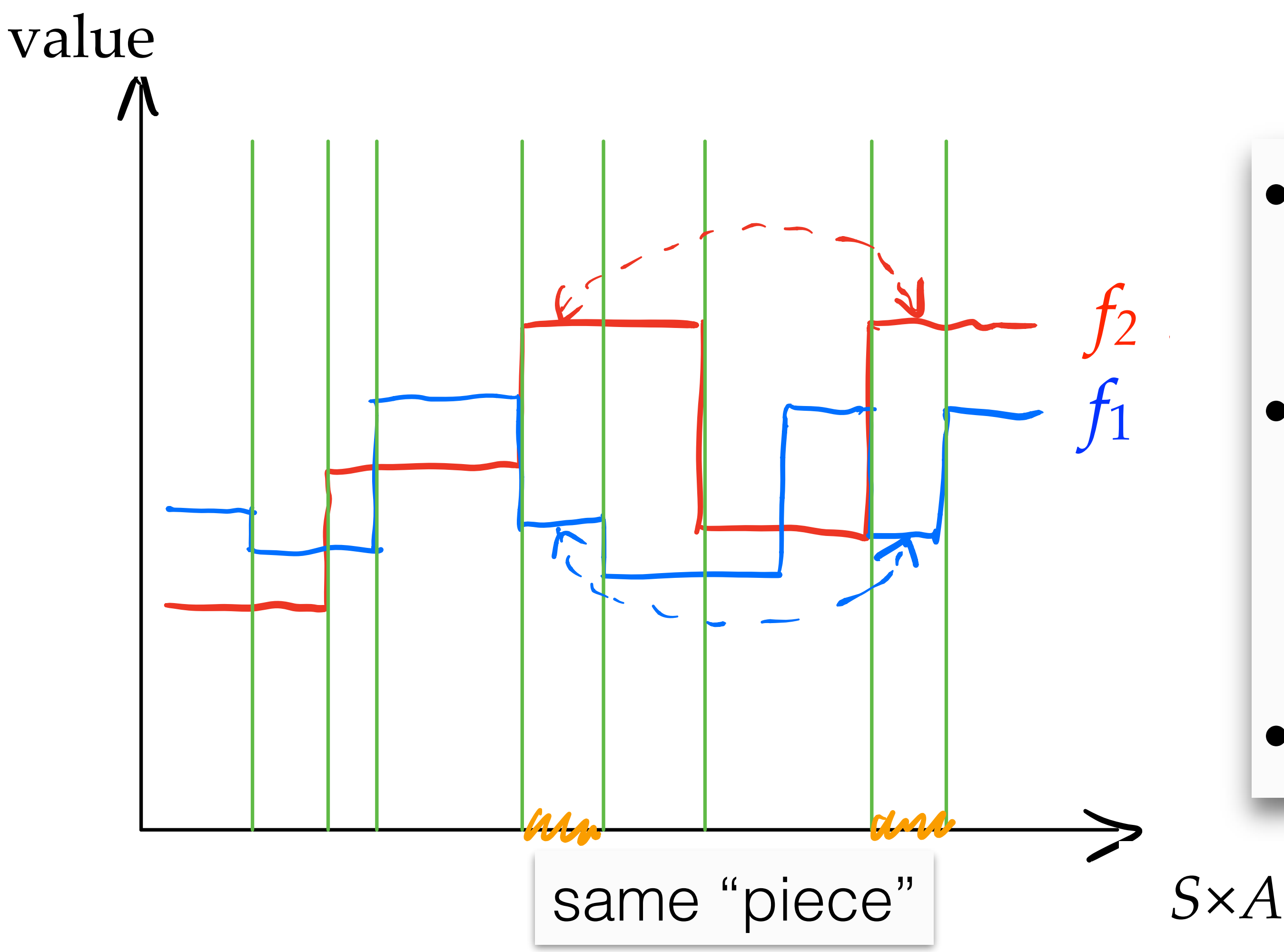
- Algorithm: **BVFT** $\arg\min_i \max_j \|f_i - \text{Proj}_{\mathcal{G}_{i,j}}(\mathcal{T} f_i)\|_{2,D}$
- Sample complexity poly in horizon, $1/\varepsilon$, log(#candidates), and C (data coverage)

- (Simplified) problem: identify Q^* out of $F = \{ f_1, f_2 \}$
- Partition SxA according to both functions simultaneously!
  - Pw-const class $G_{1,2}$ w/ size $\mathbf{O(1/\varepsilon^2)}$ **!!**
- Naive extension to >2 functions in F: $O(1/\varepsilon^{|F|})$
  - Pairwise comparison + tournament

Formal guarantee in backup slide

Xie & **Jiang**. ICML-21. Batch Value-function Approximation with Only Realizability.

Batch Value-Function Tournament [XJ, ICML-21]

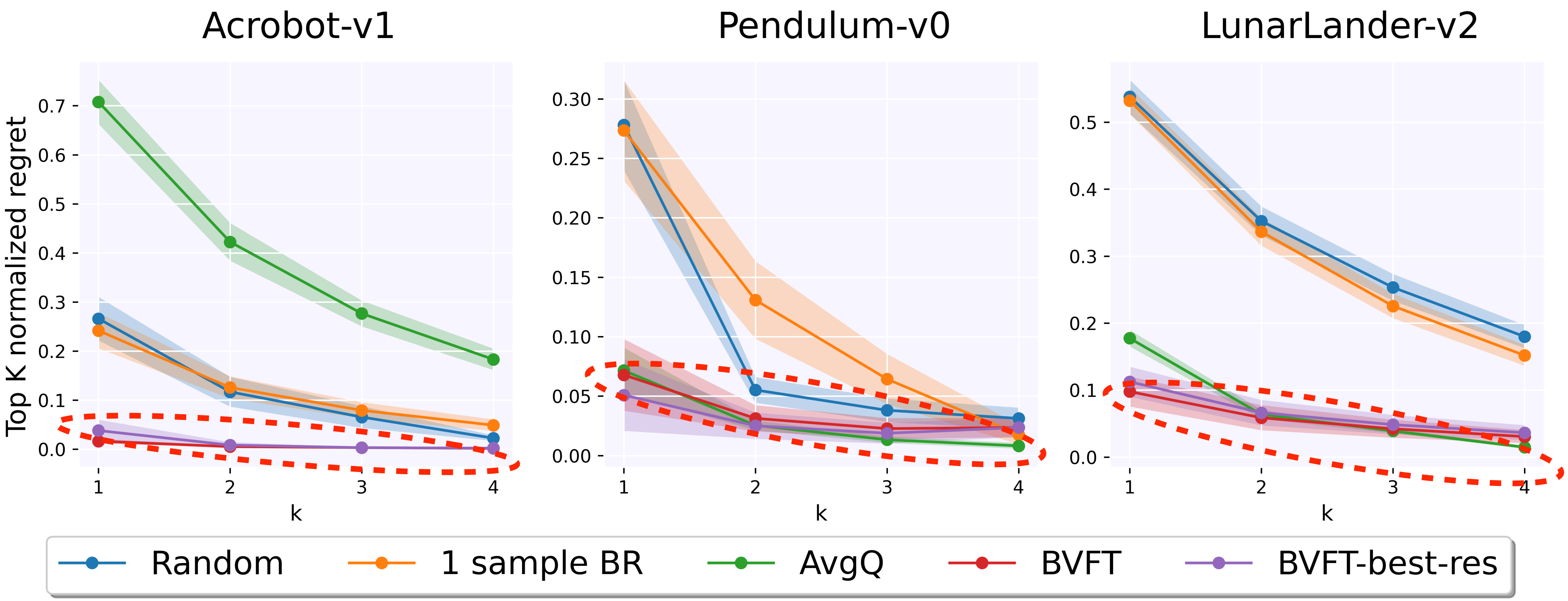


- Algorithm: **BVFT**
  $\arg\min_i \max_j \|f_i - \mathrm{Proj}_{\mathcal{G}_{i,j}}(\mathcal{T} f_i)\|_{2,D}$
- Sample complexity poly in horizon, $1/\varepsilon$, log(#candidates), and C (data coverage)
- Computation: #data points * $|F|^2$

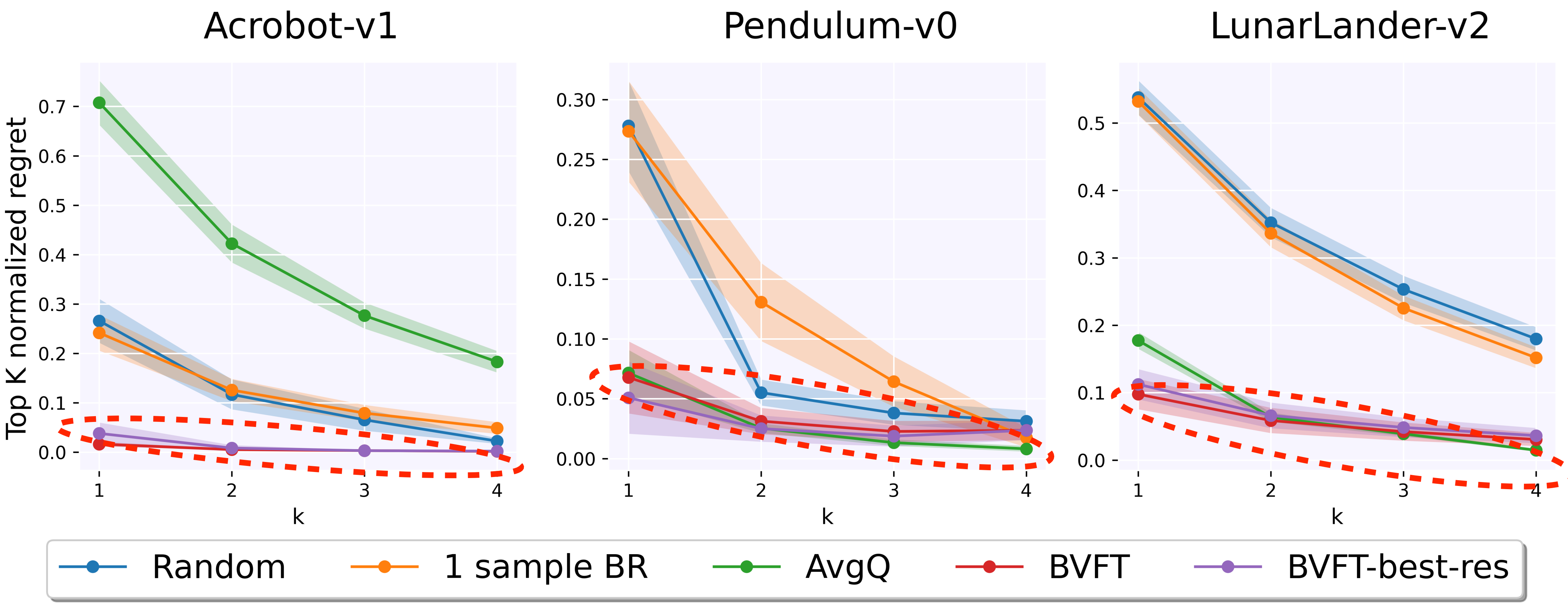
- (Simplified) problem: identify Q^* out of $F = \{ f_1, f_2 \}$
- Partition SxA according to both functions simultaneously!
  - Pw-const class $G_{1,2}$ w/ size $O(1/\varepsilon^2)$!!
- Naive extension to >2 functions in F: $O(1/\varepsilon^{|F|})$
  - Pairwise comparison + tournament

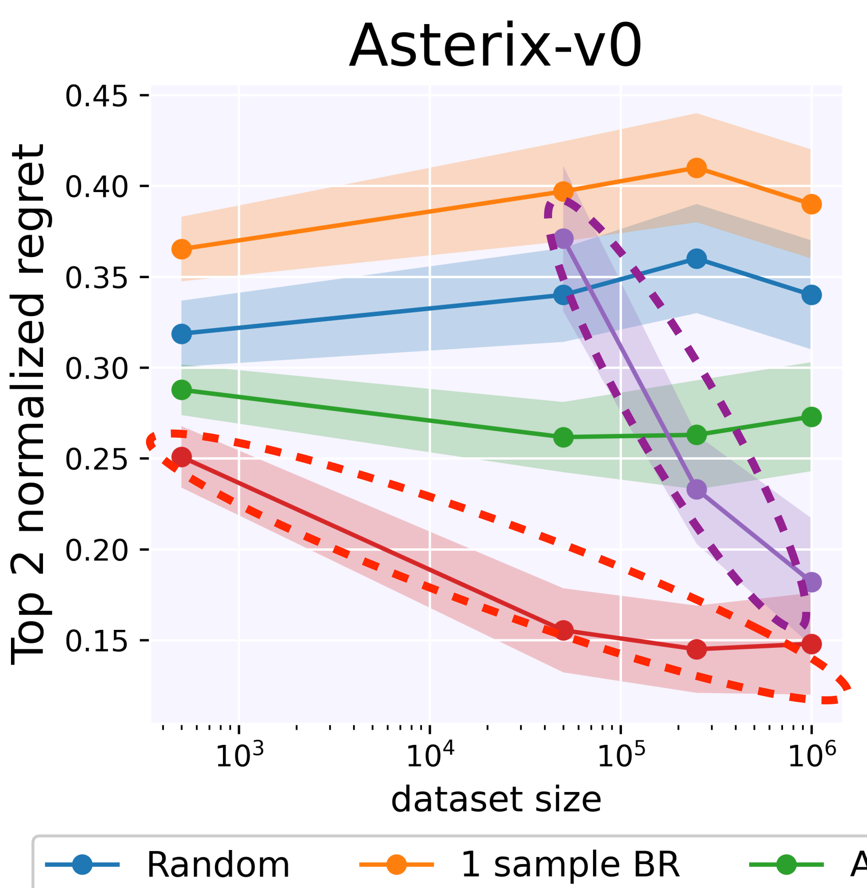
Formal guarantee in backup slide

Xie & Jiang. ICML-21. Batch Value-function Approximation with Only Realizability.



Zhang & **Jiang**. NeurIPS-21. Towards Hyperparameter-free Policy Selection for Offline Reinforcement Learning.





$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma f_{k-1}(s', \pi) \right)^2 \right]$$

Neural architecture designed by "cheating"

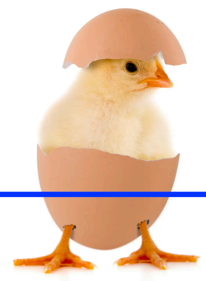
Zhang & **Jiang**. NeurIPS-21. Towards Hyperparameter-free Policy Selection for Offline Reinforcement Learning.

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

$\Downarrow$ π = greedy w.r.t. $\hat{f}$



Validation:

(FQE: learn Q^π)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma f_{k-1}(s',\pi))^2 \right]$$

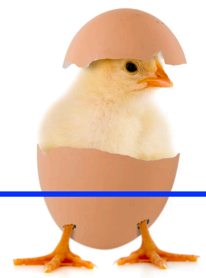
- BVFT: H-P free solution for value-function selection

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

$\Downarrow$ π = greedy w.r.t. $\hat{f}$



Validation:

(FQE: learn Q^π)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma f_{k-1}(s',\pi))^2 \right]$$

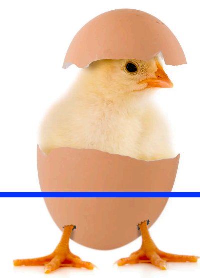
- BVFT: H-P free solution for value-function selection
- Many open problems in validation
  - Data coverage issues (see lower bound [FKSX'22])

Foster et al. COLT-22. Offline reinforcement learning: Fundamental barriers for value function approximation.

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*) $f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2]$

$\Downarrow$ π = greedy w.r.t. $\hat{f}$



Validation:

(FQE: learn Q^π) $f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D[(f_\theta(s,a) - r - \gamma f_{k-1}(s',\pi))^2]$

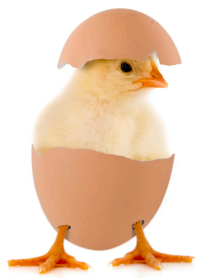
- BVFT: H-P free solution for value-function selection
- Many open problems in validation
  - Data coverage issues (see lower bound [FKSX'22])
  - Combine with different OPE methods
    e.g., marginalized importance sampling
    [LLTZ'18, NCDL'19, UHJ'20, JH'20, VJY'21, HJ'22]
  - Practical toolkit (cf. for OPE [VLJY'20])

Foster et al. COLT-22. Offline reinforcement learning: Fundamental barriers for value function approximation.
Voloshin et al. NeurIPS-21. Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning.

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*) $$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2\right]$$

$$\Downarrow \quad \pi = \text{greedy w.r.t. } \hat{f}$$



Validation:

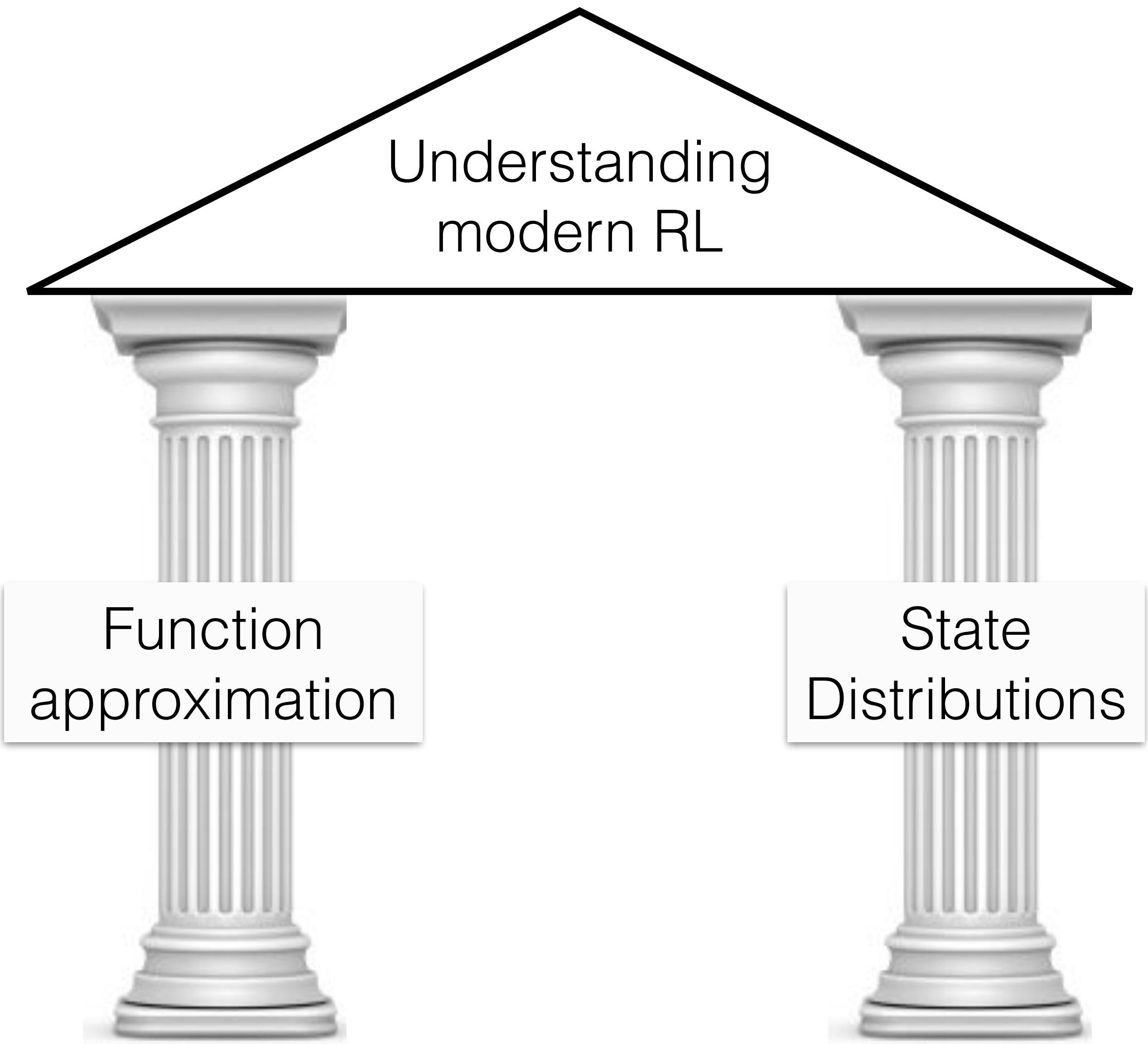
(FQE: learn Q^π) $$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma f_{k-1}(s',\pi))^2\right]$$

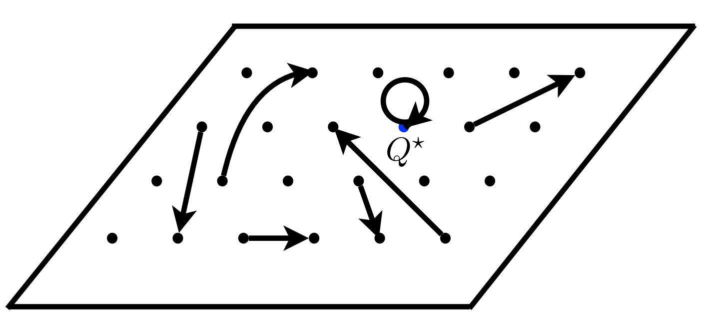
- BVFT: H-P free solution for value-function selection
- Many open problems in validation
  - Data coverage issues (see lower bound [FKSX'22])
  - Combine with different OPE methods
    e.g., marginalized importance sampling
    [LLTZ'18, NCDL'19, UHJ'20, JH'20, VJY'21, HJ'22]
  - Practical toolkit (cf. for OPE [VLJY'20])

Foster et al. COLT-22. Offline reinforcement learning: Fundamental barriers for value function approximation.
Voloshin et al. NeurIPS-21. Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning.

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*) $$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s', a'))^2 \right]$$

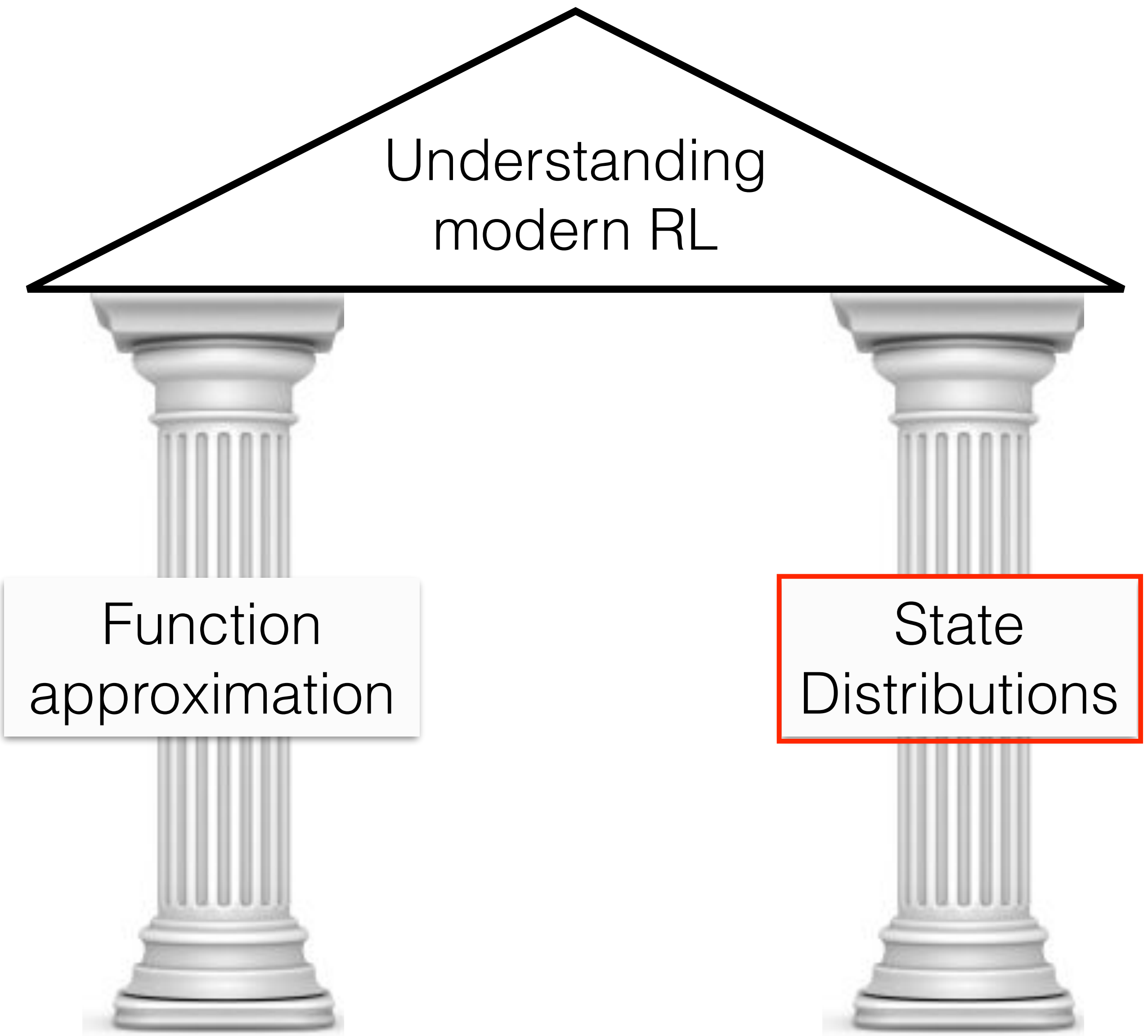


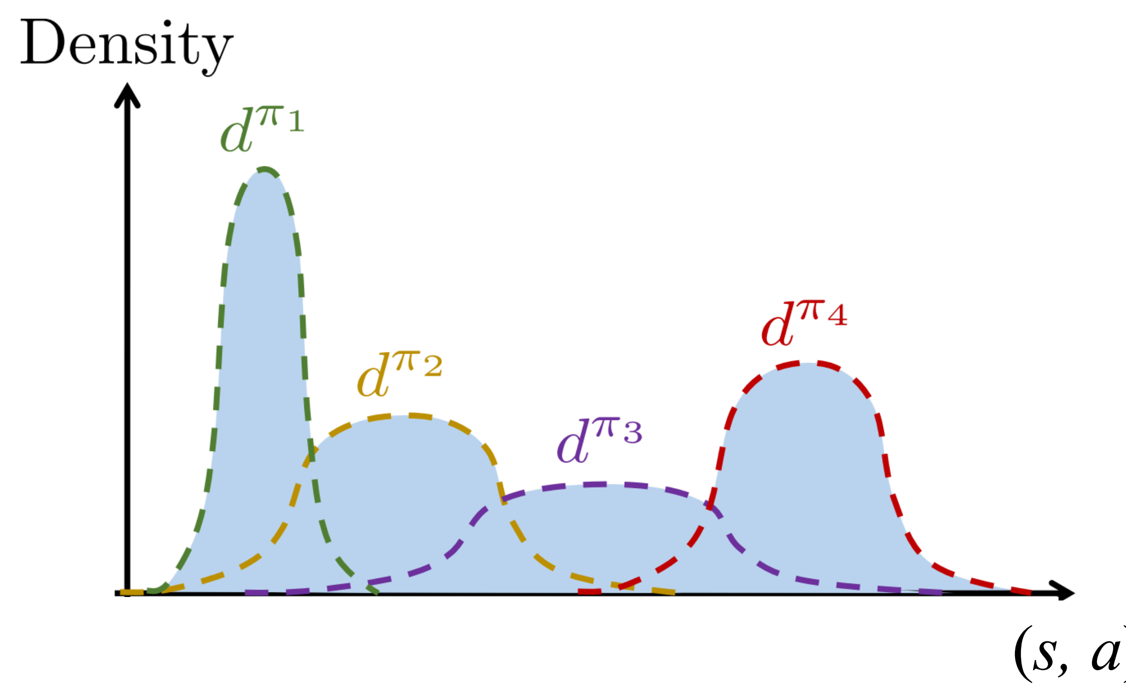


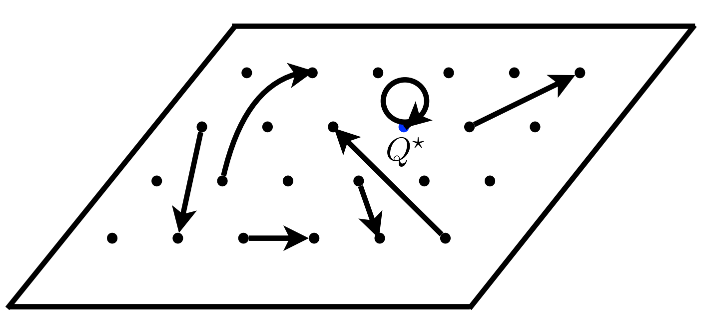
Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a') \right)^2 \right]$$



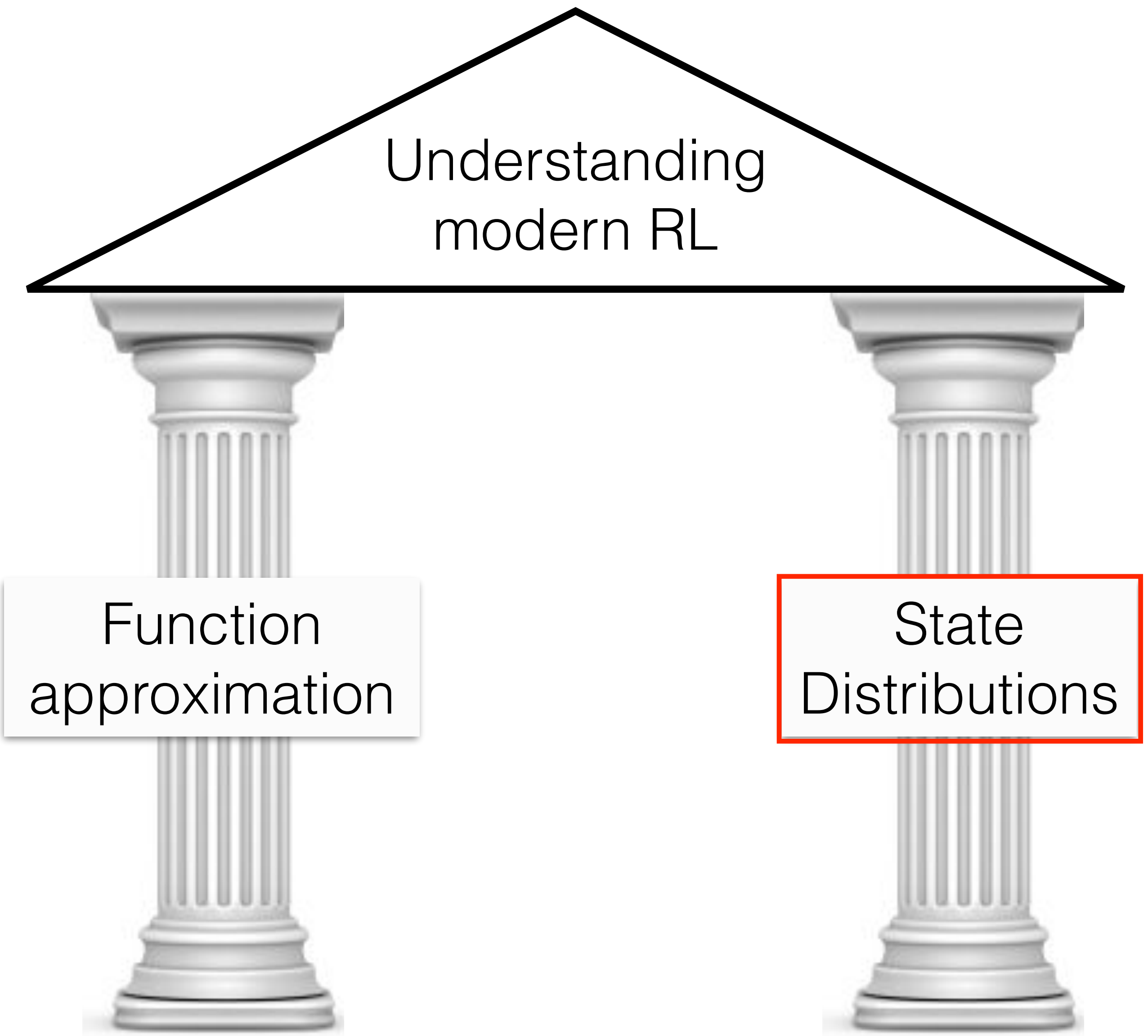


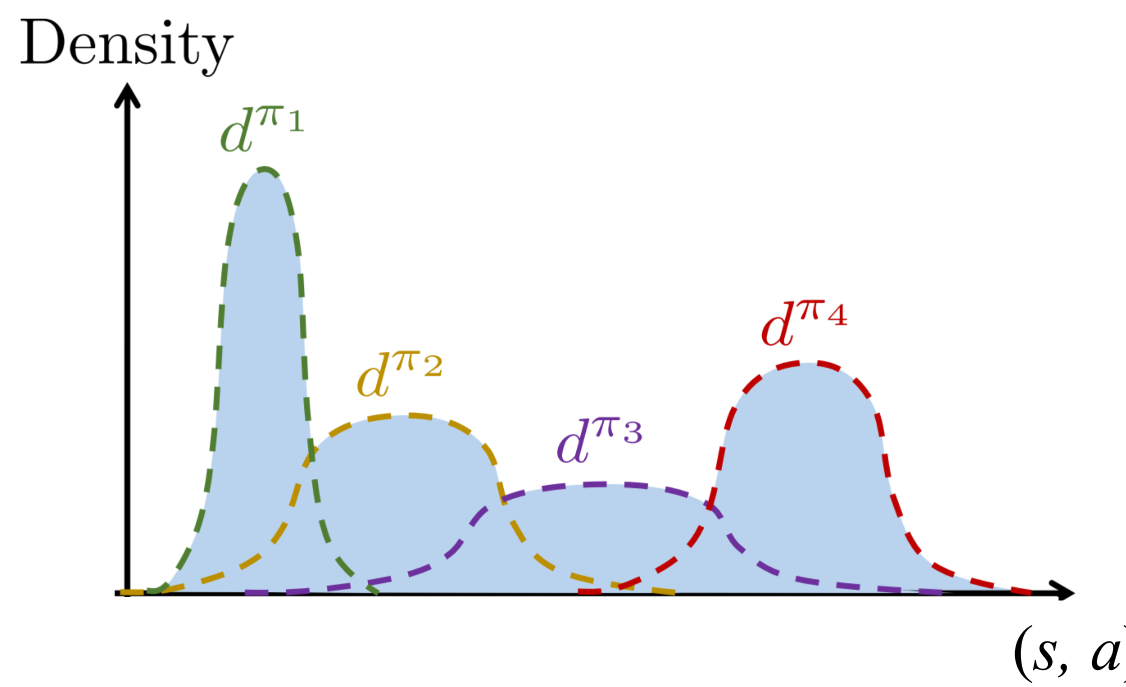


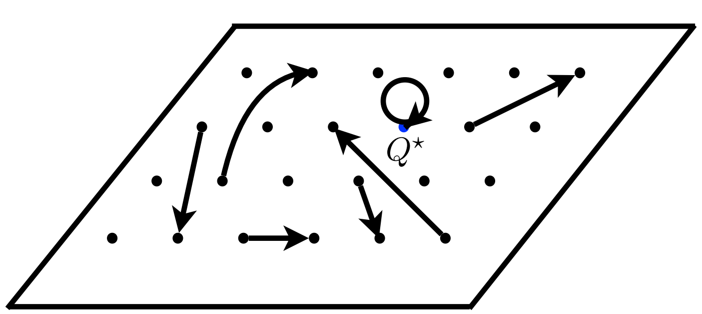
Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[\left(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s', a') \right)^2 \right]$$







ICML | 2022

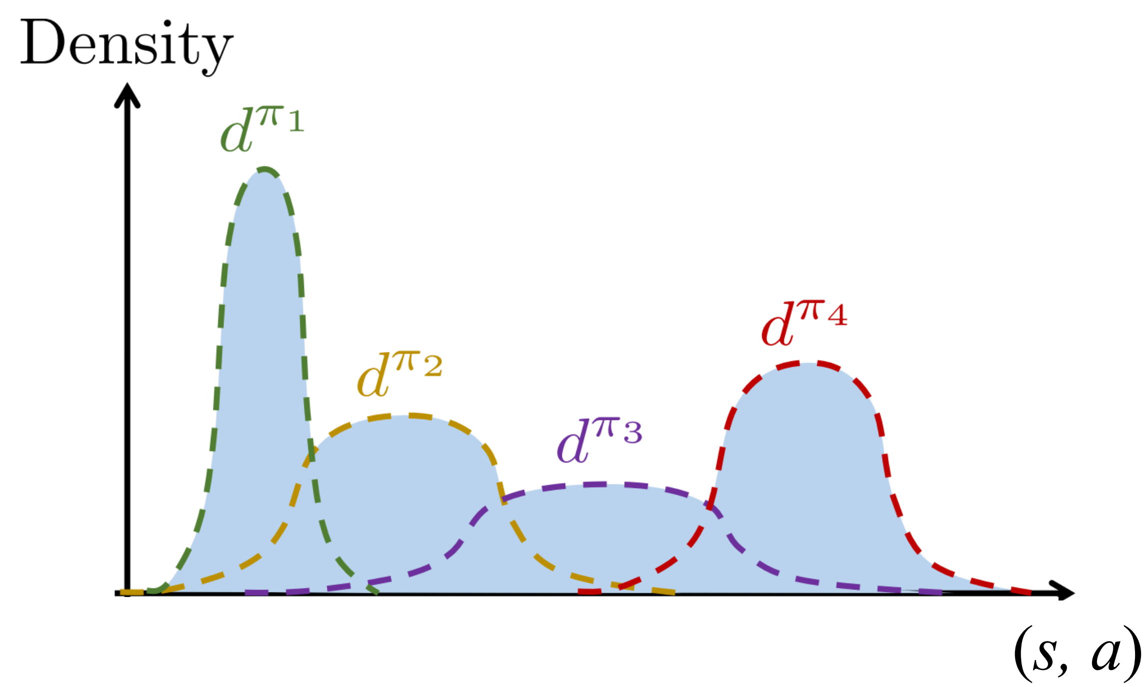
Outstanding Paper Runner Up

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*) $f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D\left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2\right]$

Standard assumption

- $\max_\pi \|d^\pi / D\|_\infty \leq C$
- All policies covered by data

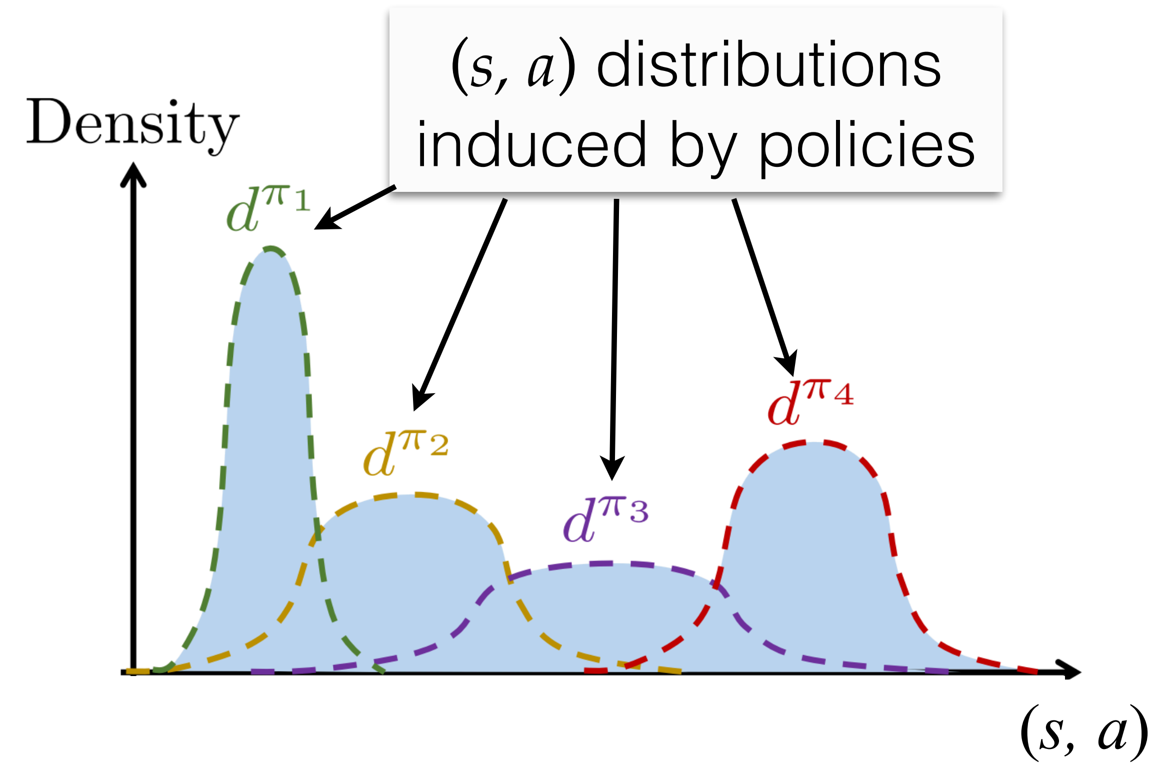


Training: $\hat{f} = f_k$ where

(FQI: learn Q^*) $$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

Standard assumption

- $\max_\pi \|d^\pi / D\|_\infty \leq C$
- All policies covered by data



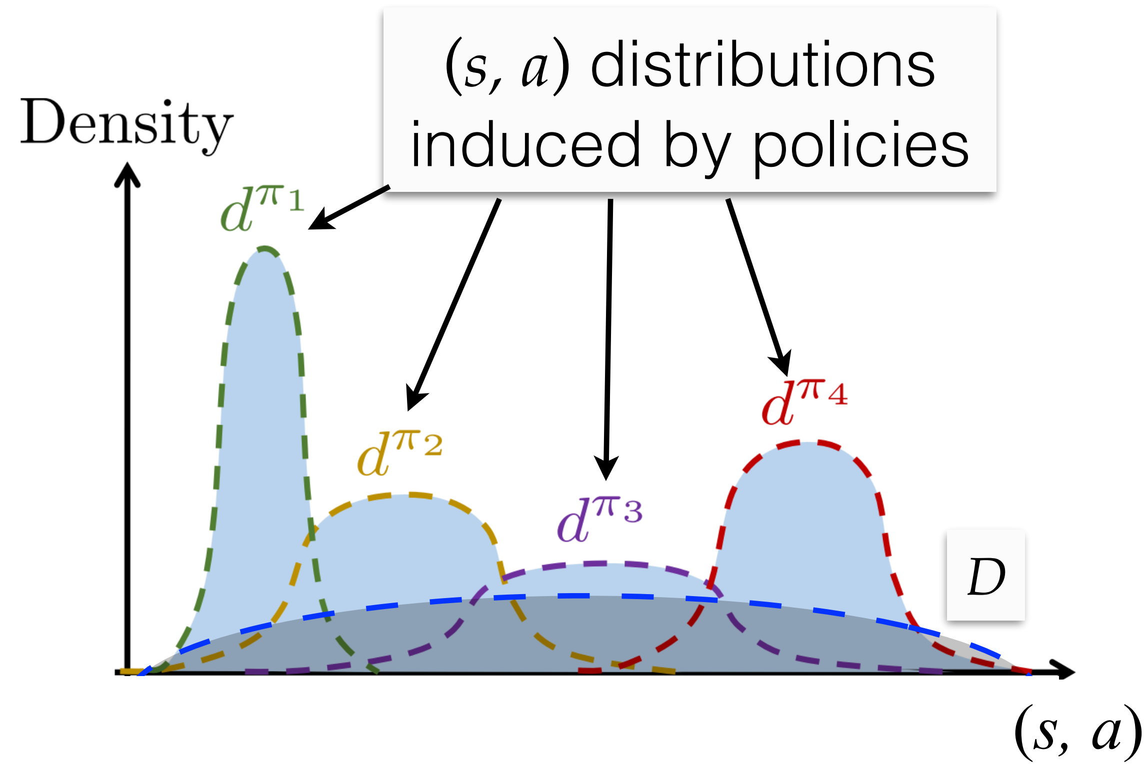
Training: $\hat{f} = f_k$ where

(FQI: learn Q^*)

$$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$$

Standard assumption

- $\max_\pi \| d^\pi / D \|_\infty \le C$
- All policies covered by data

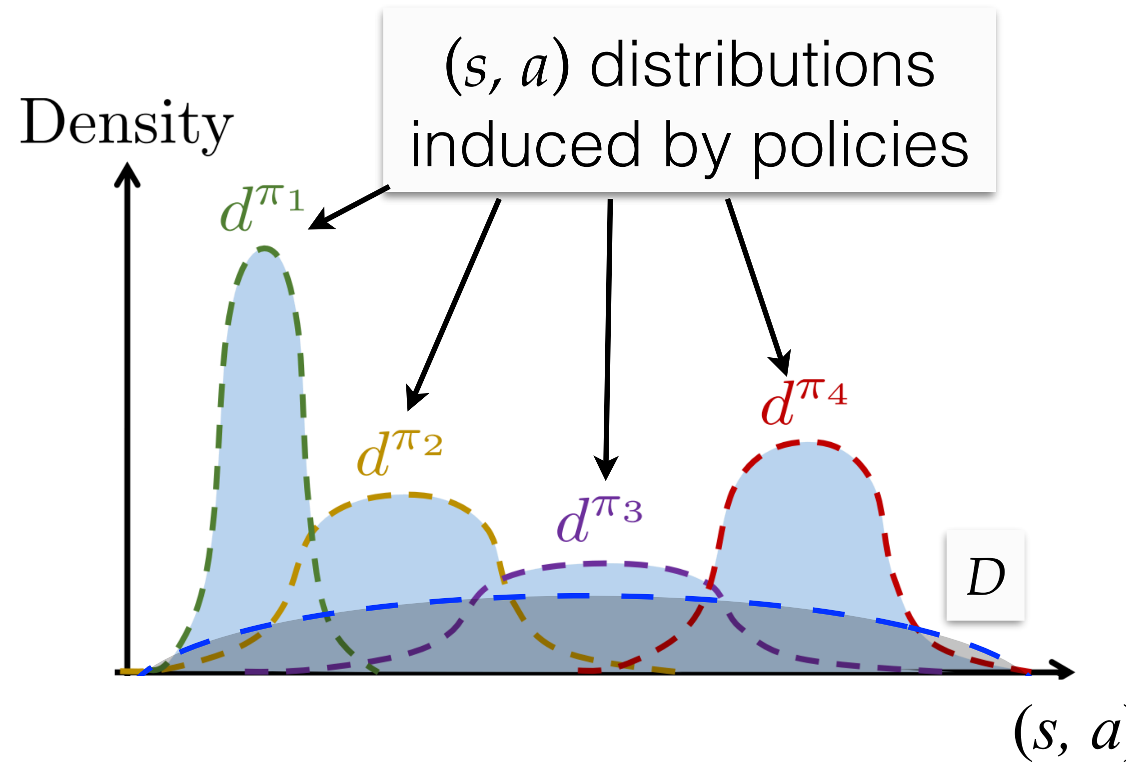


Training: $\hat{f} = f_k$ where

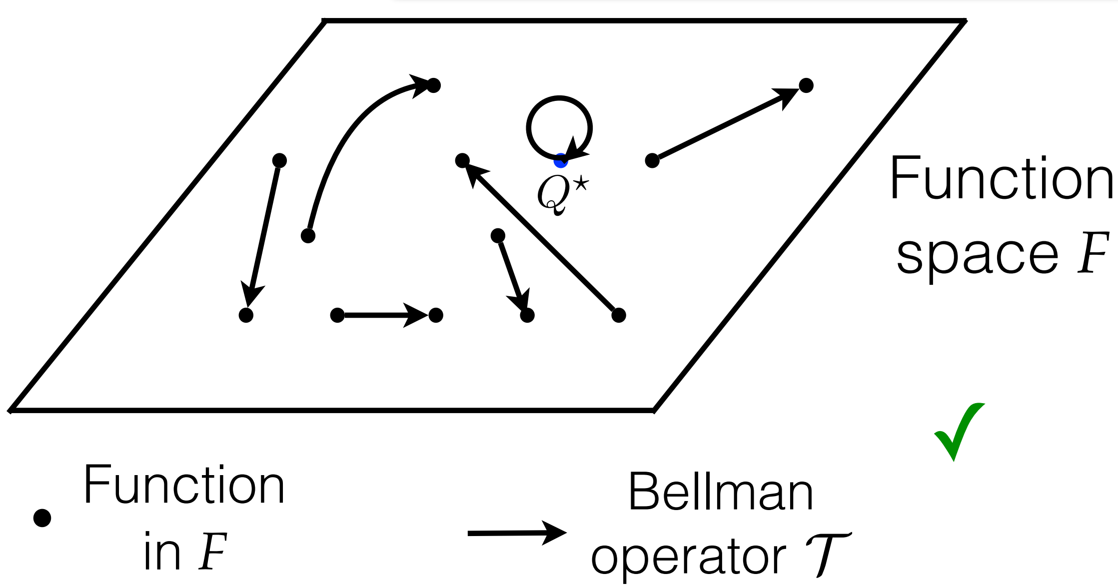
(FQI: learn Q^*) $f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D\left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2\right]$

Standard assumption

- $\max_\pi \|d^\pi / D\|_\infty \leq C$
- All policies covered by data
- compete with optimal policy



"Bellman-completeness"
$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathcal{F}$



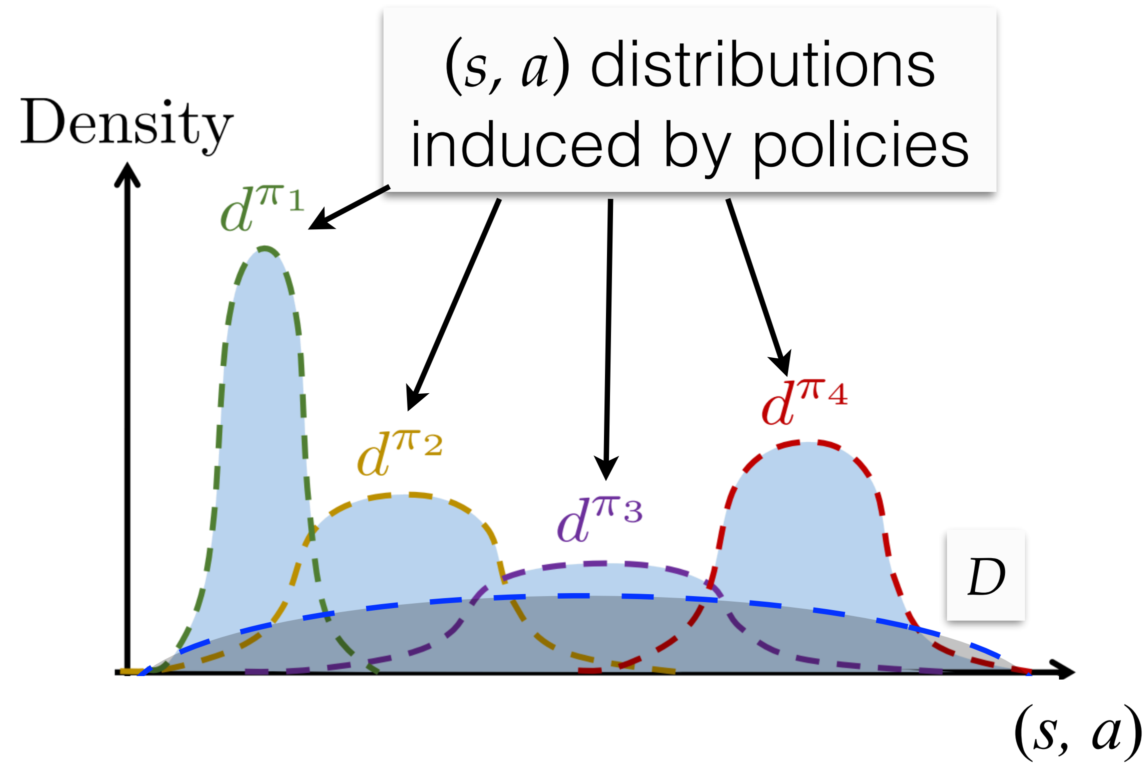
✓

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*) $$f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D\left[(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2\right]$$

Standard assumption

- $\max_\pi \|d^\pi / D\|_\infty \leq C$
- All policies covered by data
- compete with optimal policy



Challenge: real-world data lacks exploration!

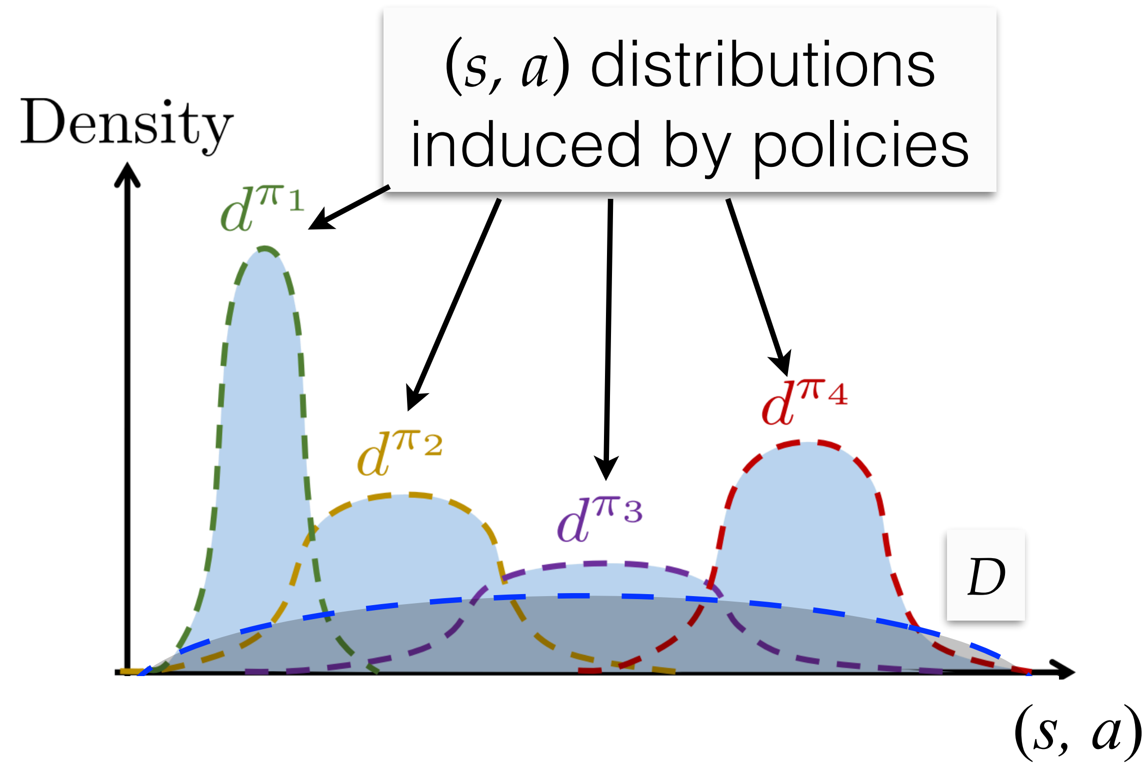
- Data may not contain all bad behaviors

Training: $\hat{f} = f_k$ where

(FQI: learn Q^*) $f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D [(f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2]$

Standard assumption

- $\max_\pi \|d^\pi / D\|_\infty \leq C$
- All policies covered by data
- compete with optimal policy

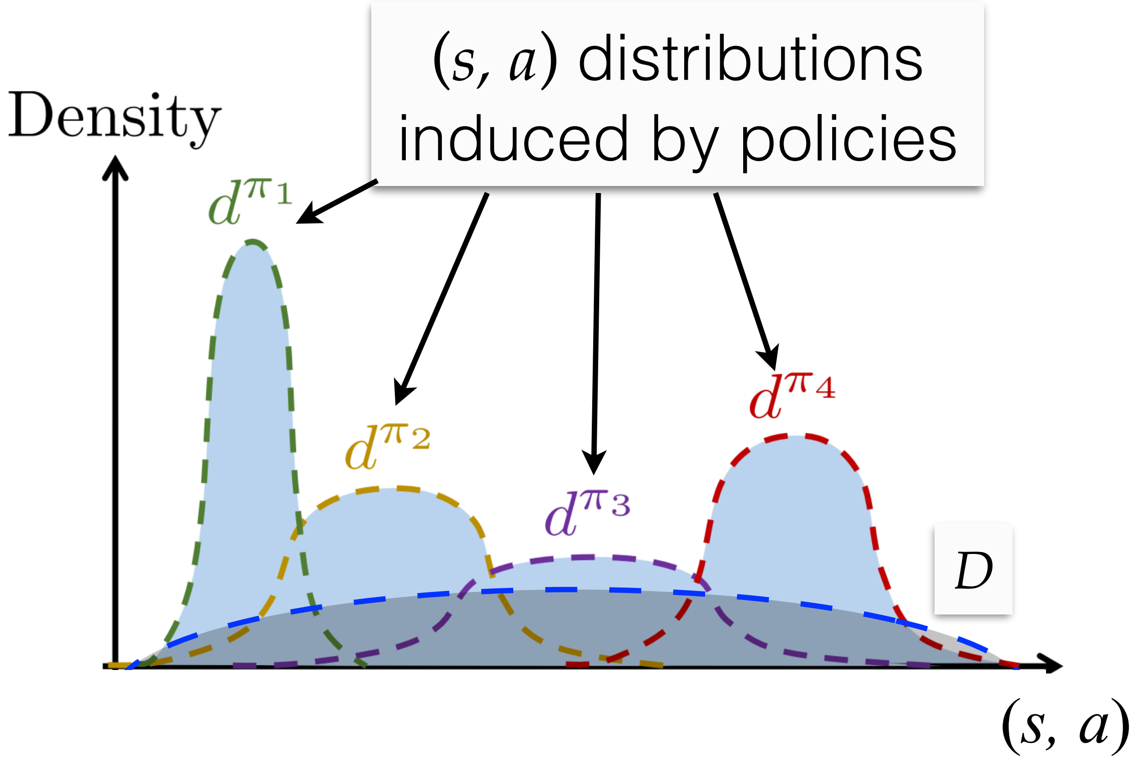


Challenge: real-world data lacks exploration!

- Data may not contain all bad behaviors
- Alg may **over-estimate** their performance

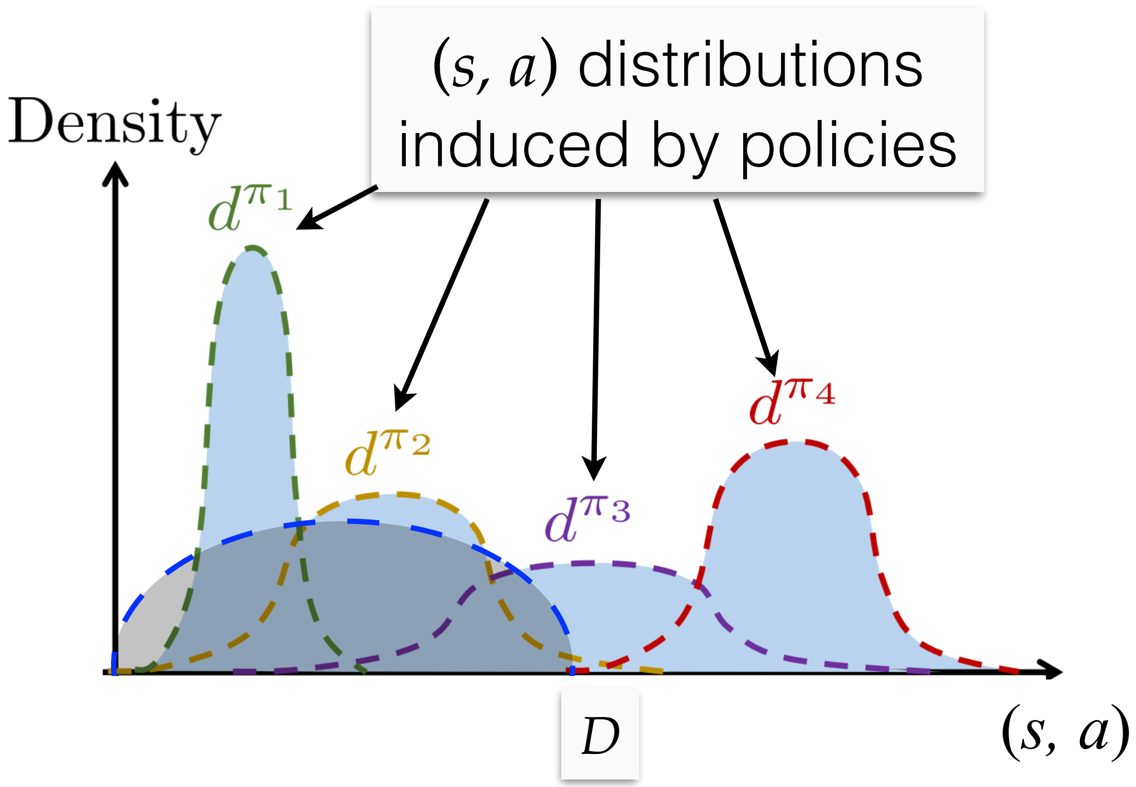
Desirable assumption

- $\max_{\pi} \|d^{\pi}/D\|_{\infty} \leq C$
- All policies covered by data



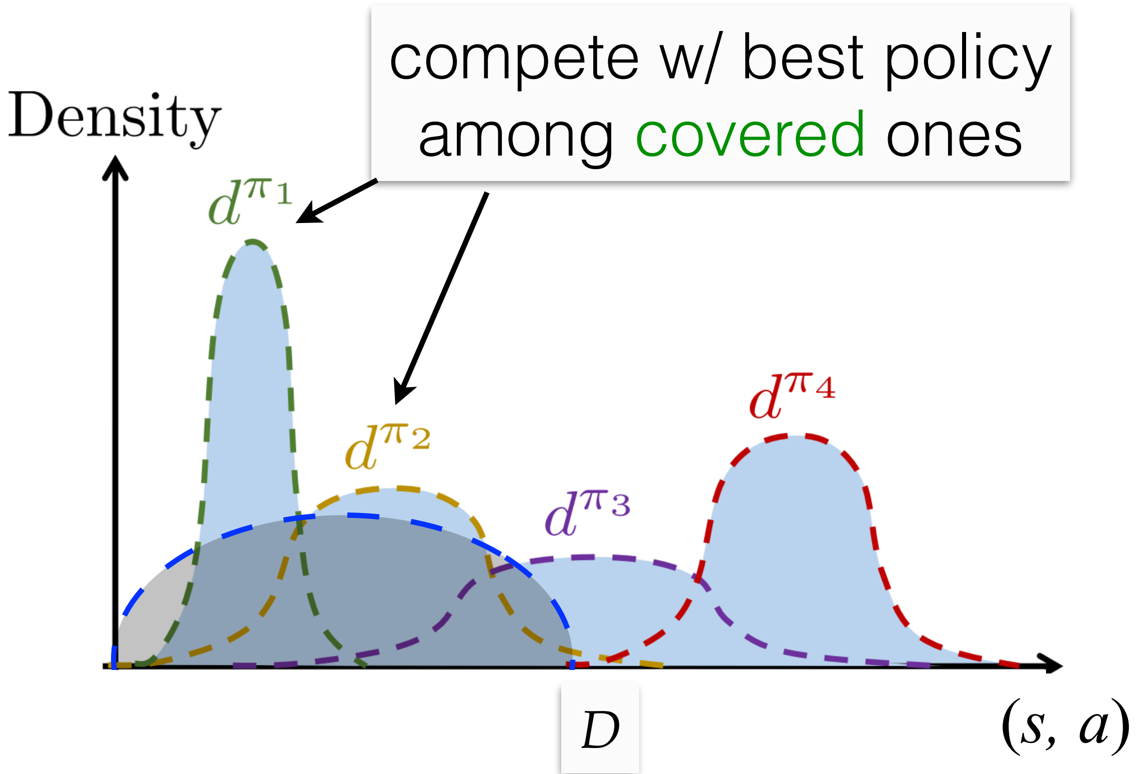
Desirable assumption

- ~~$\max_\pi \|d^\pi / D\|_\infty \leq C$~~
- ~~All policies covered by data~~



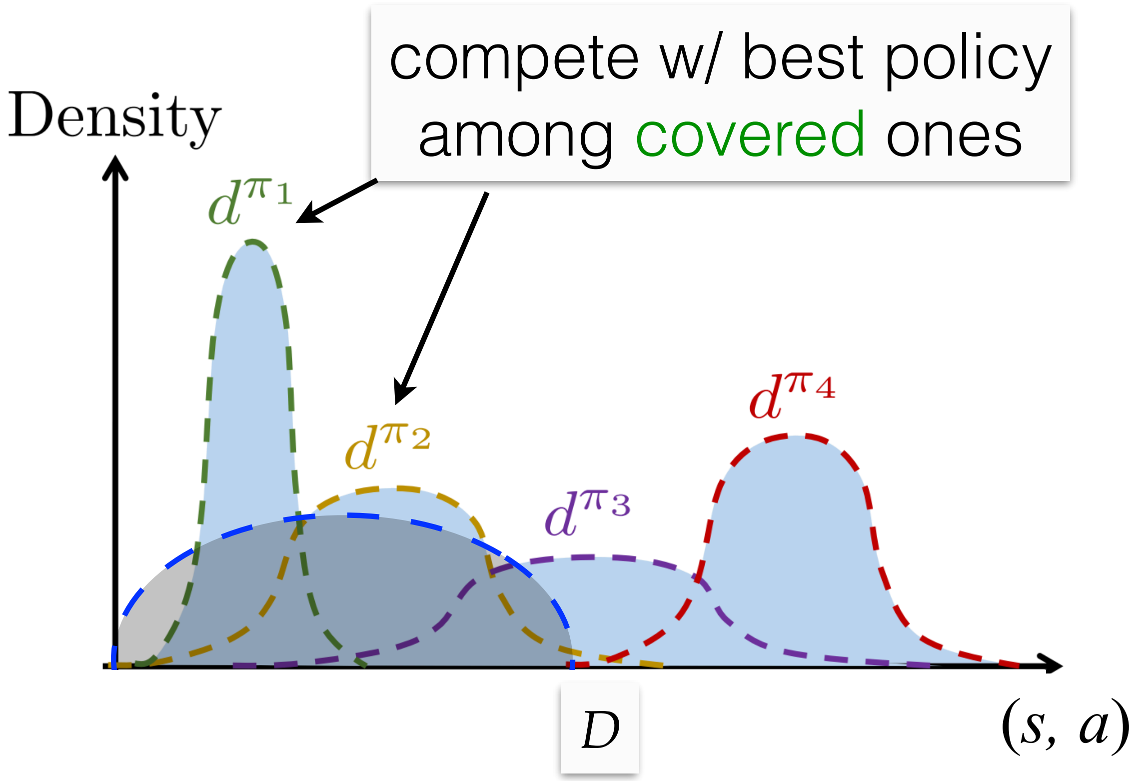
Desirable assumption

- ~~$\max_\pi \|d^\pi / D\|_\infty \le C$~~
- ~~All policies covered by data~~
- (implicit) ***a*** good policy is covered



Desirable assumption

- ~~$\max_\pi \|d^\pi / D\|_\infty \leq C$~~
- ~~All policies covered by data~~
- (implicit) ***a*** good policy is covered

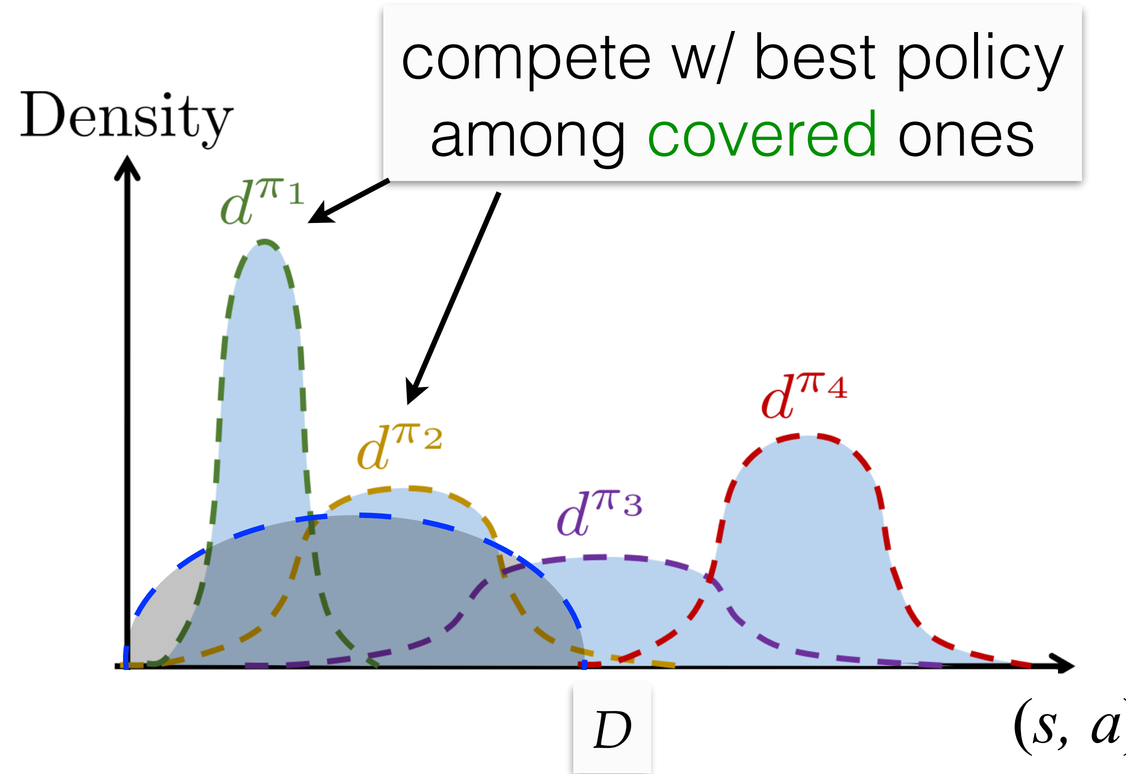


Offline RL (exploitation)

Goal: stay within data distribution

Desirable assumption

- ~~$\max_\pi \|d^\pi / D\|_\infty \leq C$~~
- ~~All policies covered by data~~
- (implicit) ***a*** good policy is covered



Offline RL (exploitation)

Goal: stay within data distribution

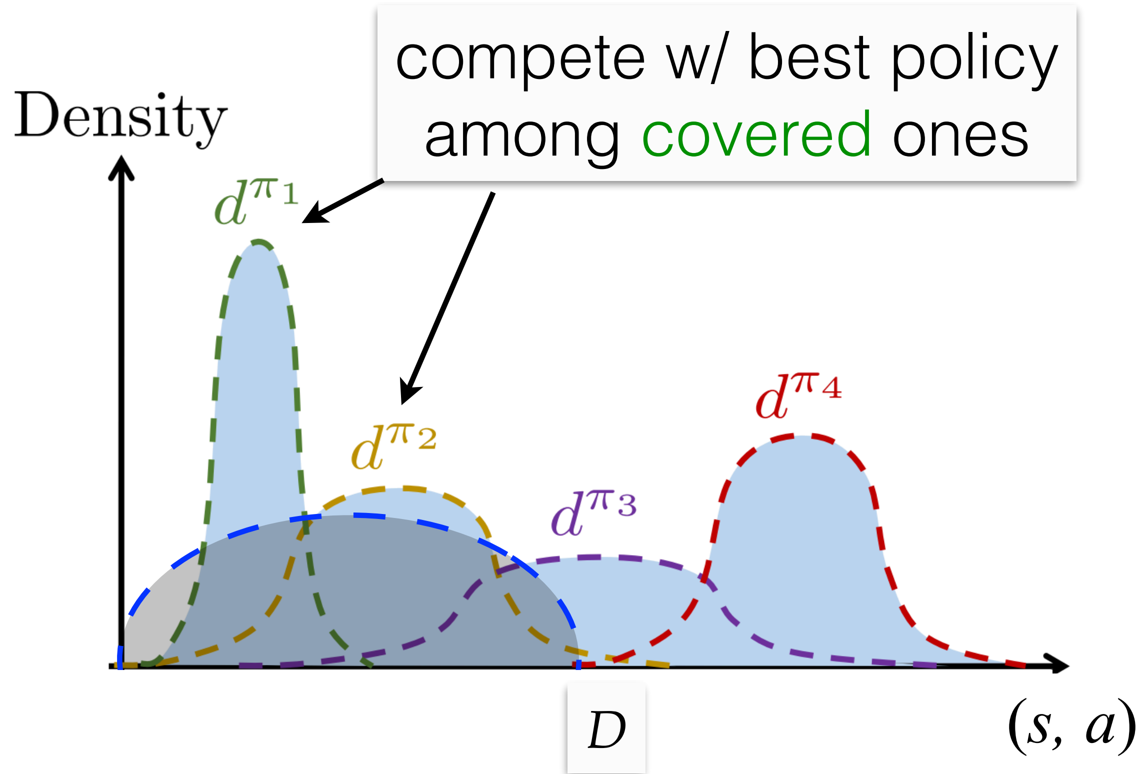
Online RL (exploration)

Goal: leave current data distribution

Principle: *optimism*

Desirable assumption

- ~~$\max_\pi \|d^\pi / D\|_\infty \leq C$~~
- ~~All policies covered by data~~
- (implicit) ***a*** good policy is covered



Offline RL (exploitation)

Goal: stay within data distribution

Principle: *pessimism*

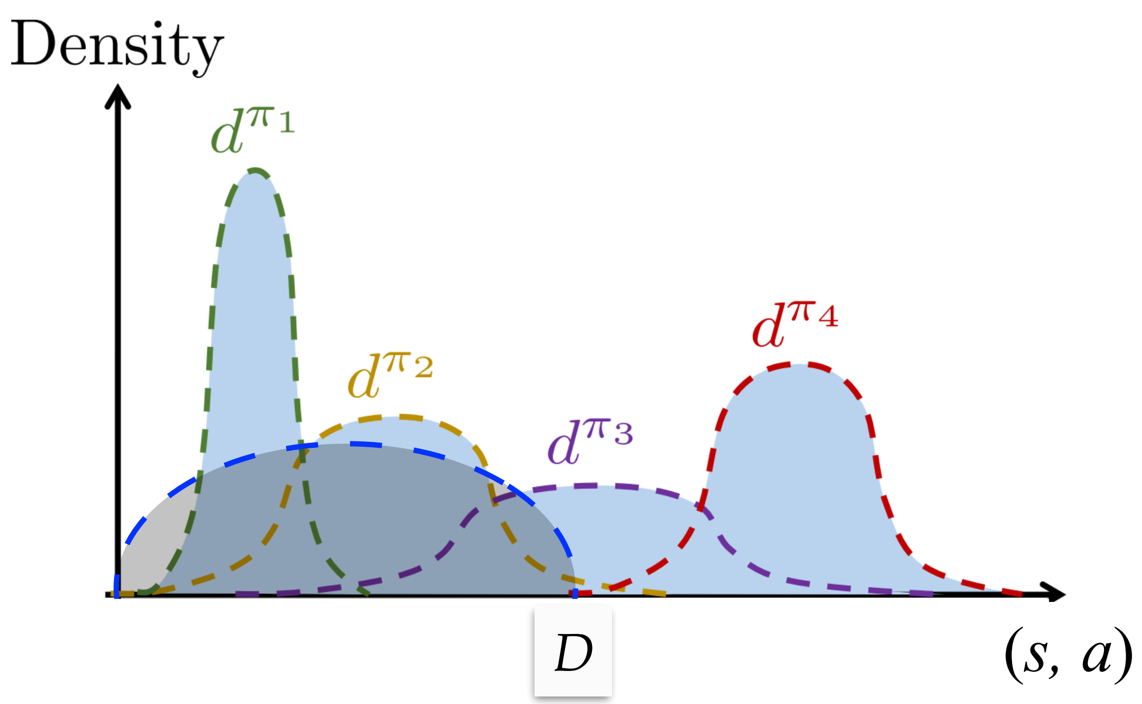
Online RL (exploration)

Goal: leave current data distribution

Principle: *optimism*

Pessimism in face of uncertainty

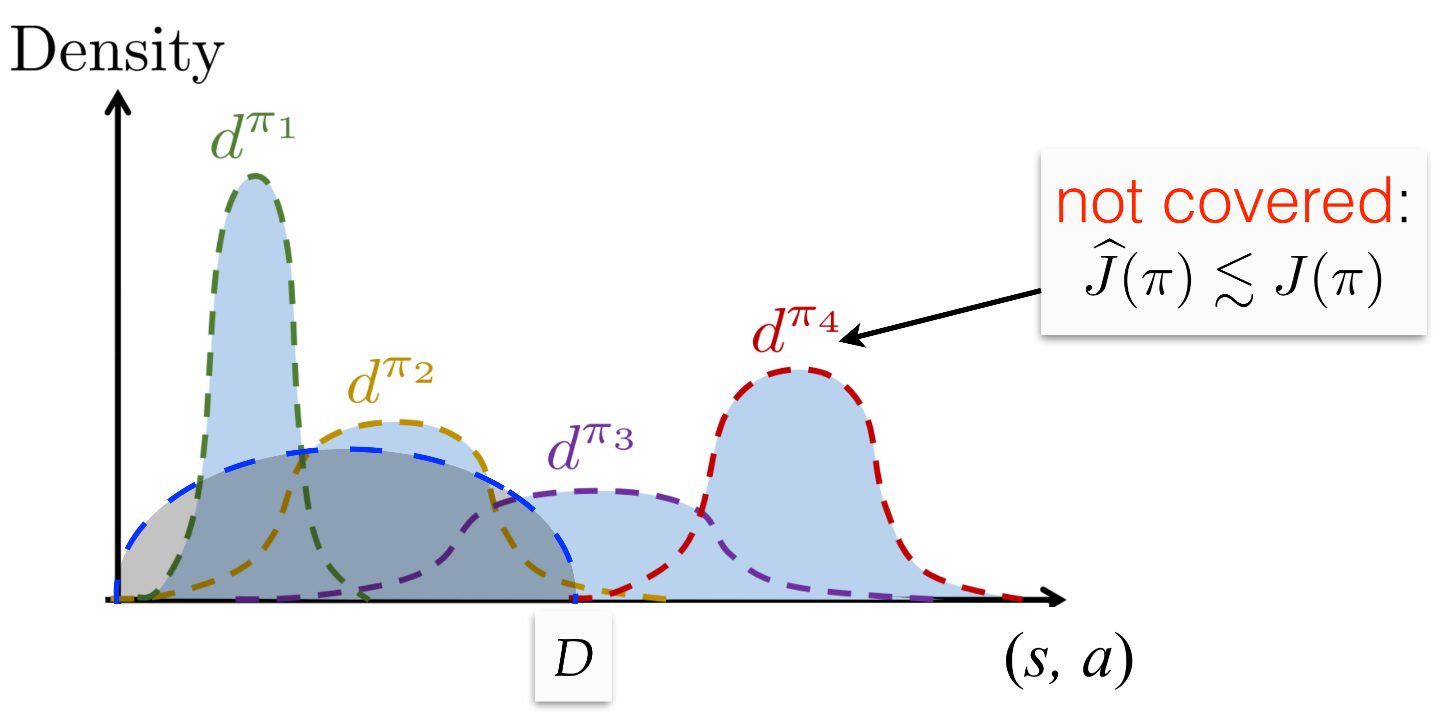
- Policy optimization: $\arg\max_{\pi \in \Pi} J(\pi) := Q^{\pi}(s_0, \pi)$
  - Π: policy class



Pessimism in face of uncertainty

- Policy optimization: $\arg\max_{\pi \in \Pi} J(\pi) := Q^{\pi}(s_0, \pi)$
  - Π: policy class

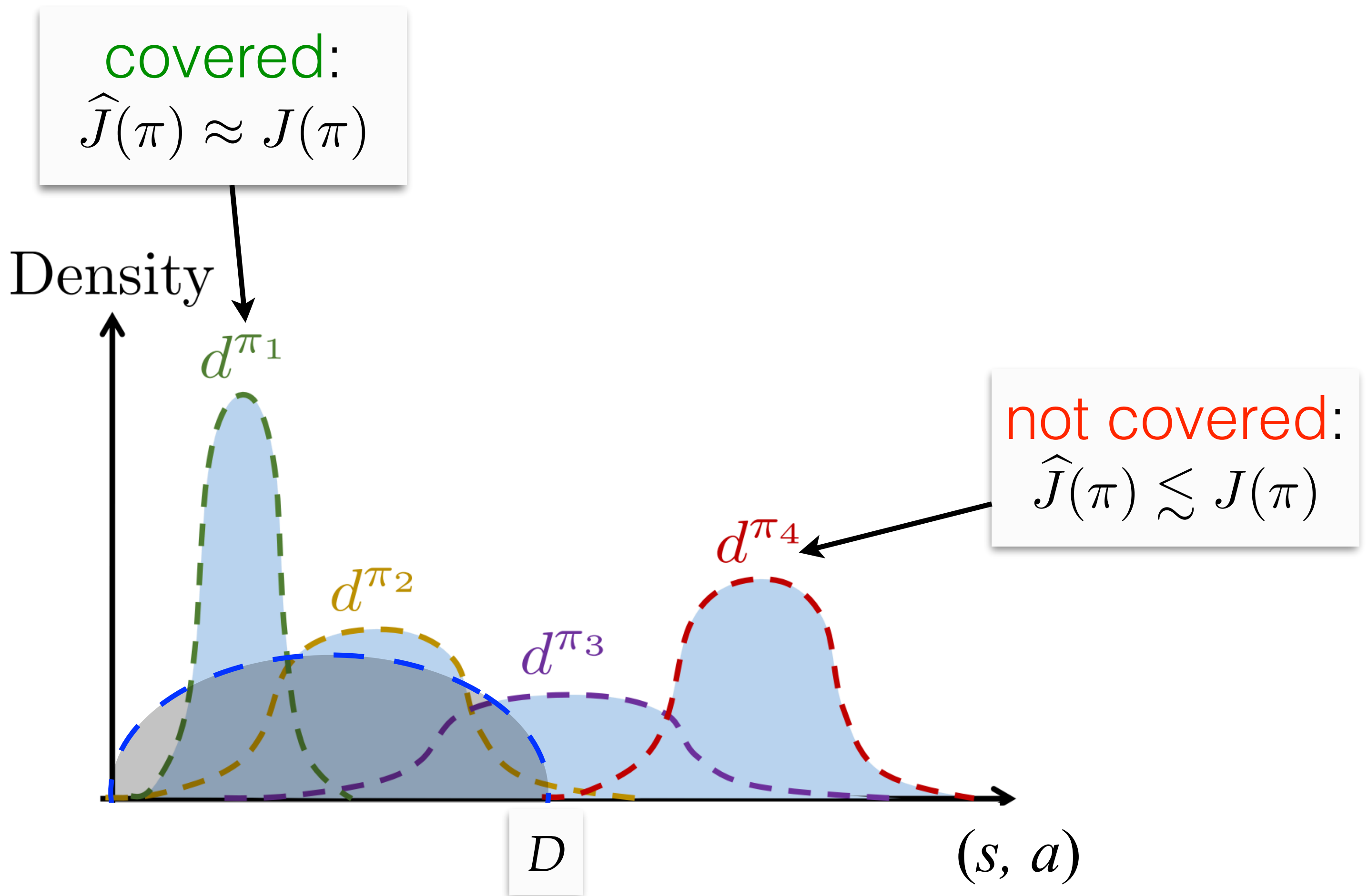
$$\arg\max_{\pi \in \Pi} \widehat{J}(\pi) = \qquad \text{lower bound of } J(\pi)$$



Pessimism in face of uncertainty

- Policy optimization: $\arg\max_{\pi \in \Pi} J(\pi) := Q^{\pi}(s_0, \pi)$
  - Π: policy class

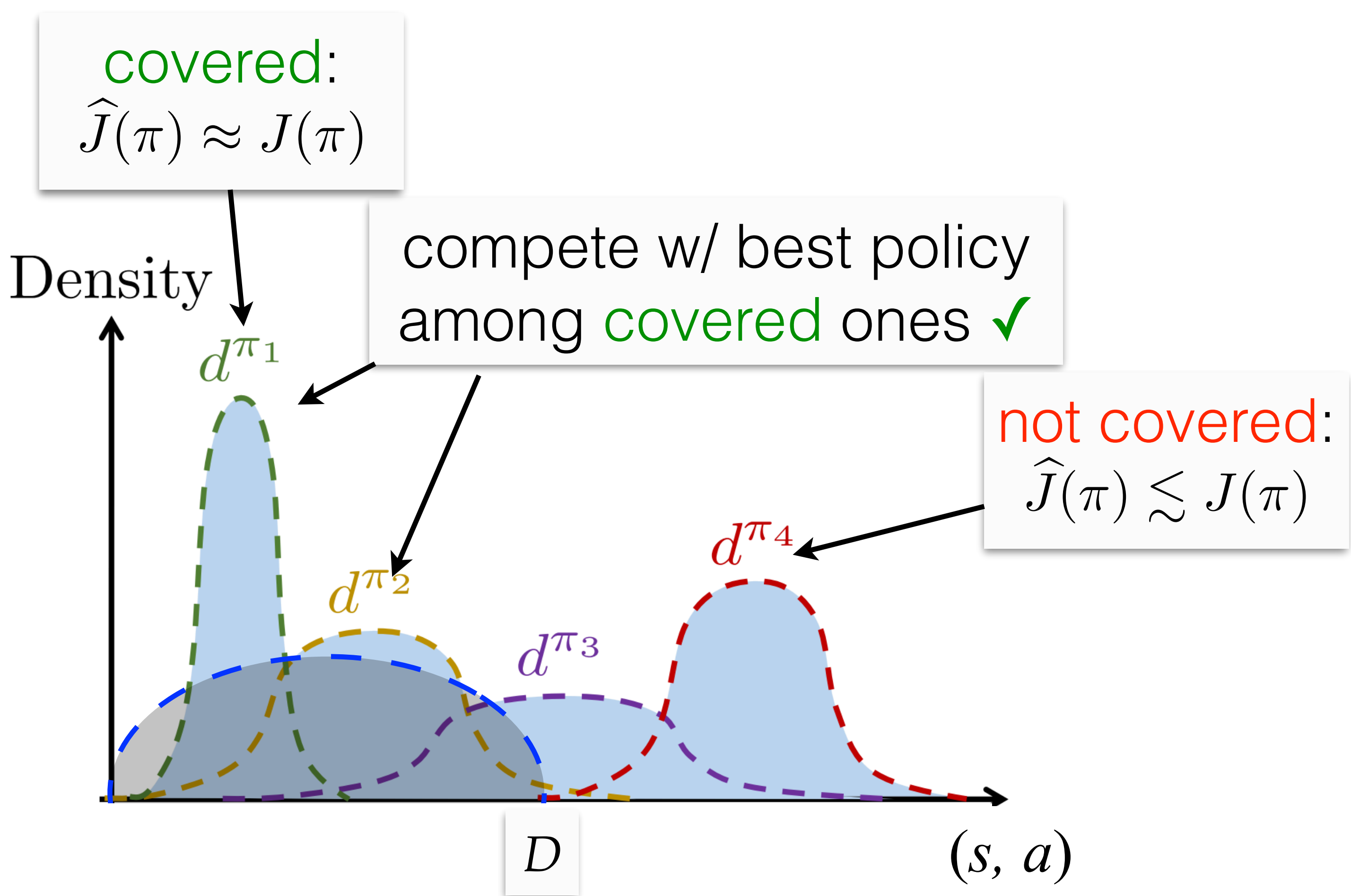
$$\arg\max_{\pi \in \Pi} \widehat{J}(\pi) = \text{tight lower bound of } J(\pi)$$



Pessimism in face of uncertainty

- Policy optimization: $\arg\max_{\pi \in \Pi} J(\pi) := Q^{\pi}(s_0, \pi)$
  - Π: policy class

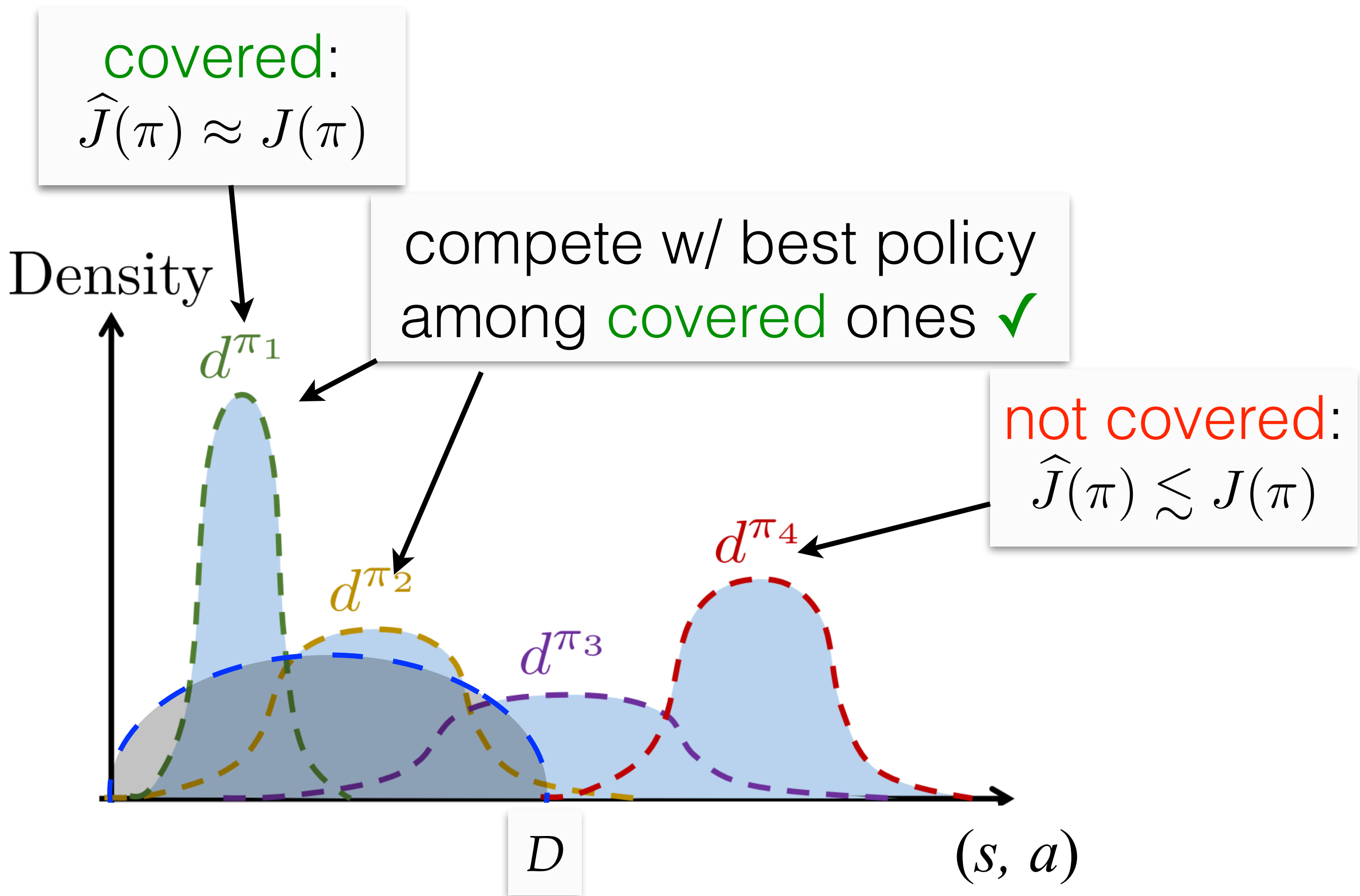
$$\arg\max_{\pi \in \Pi} \widehat{J}(\pi) = \text{tight lower bound of } J(\pi)$$



Pessimism in face of uncertainty

- Policy optimization: $\arg\max_{\pi\in\Pi} J(\pi) := Q^\pi(s_0, \pi)$
  - Π: policy class

$$\arg\max_{\pi\in\Pi} \widehat{J}(\pi) = \text{tight lower bound of } J(\pi)$$



 Nilim and Ghaoui. OR-05. Robust control of Markov decision processes with uncertain transition matrices.

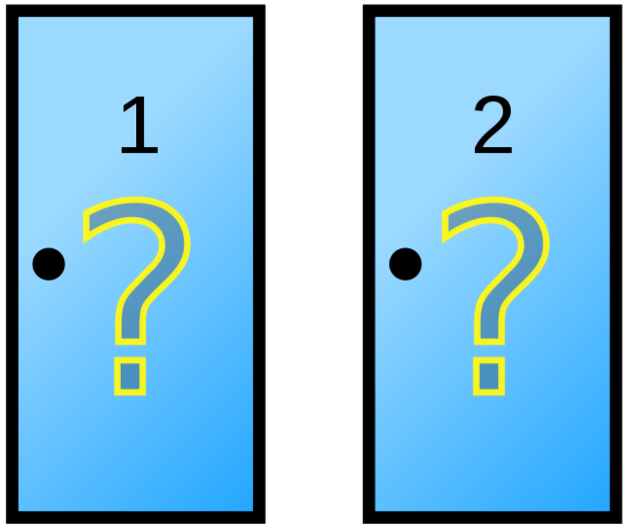
Prior work: Point-wise pessimism [JZW'21]

- Learn $\widehat{Q}^\pi \leq Q^\pi \Longrightarrow \widehat{J}(\pi) \leq J(\pi)$
- Overly-conservative by imagining impossible scenarios
    - *Ex.* In linear MDPs, true Q^π linear, but $\widehat{Q}^\pi$ quadratic

Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

Prior work: Point-wise pessimism [JZW'21]

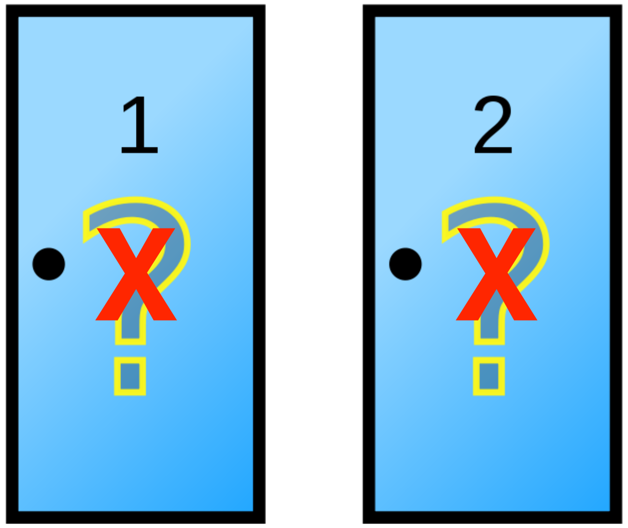
- Learn $\widehat{Q}^\pi \leq Q^\pi \Longrightarrow \widehat{J}(\pi) \leq J(\pi)$
- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true Q^π linear, but $\widehat{Q}^\pi$ quadratic



Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

Prior work: Point-wise pessimism [JZW'21]

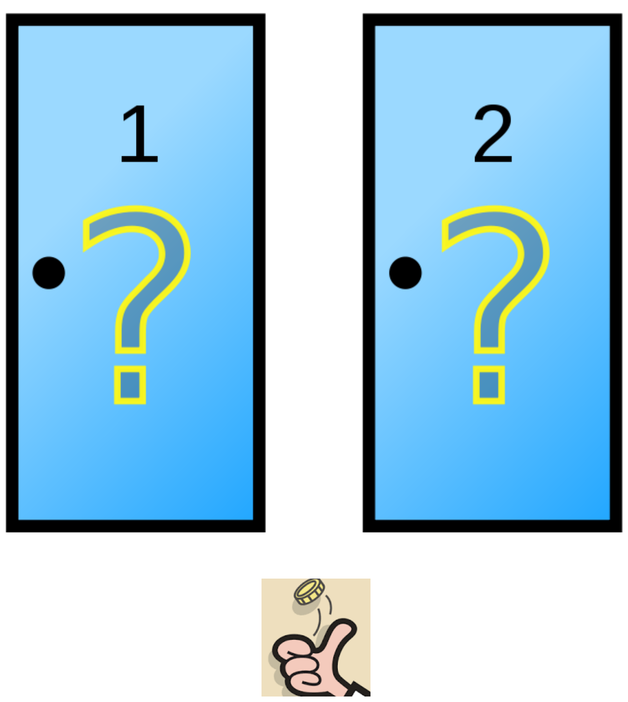
- Learn $\widehat{Q}^\pi \leq Q^\pi \Longrightarrow \widehat{J}(\pi) \leq J(\pi)$
- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true Q^π linear, but $\widehat{Q}^\pi$ quadratic



Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

Prior work: Point-wise pessimism [JZW'21]

- Learn $\widehat{Q}^\pi \leq Q^\pi \Longrightarrow \widehat{J}(\pi) \leq J(\pi)$
- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true Q^π linear, but $\widehat{Q}^\pi$ quadratic

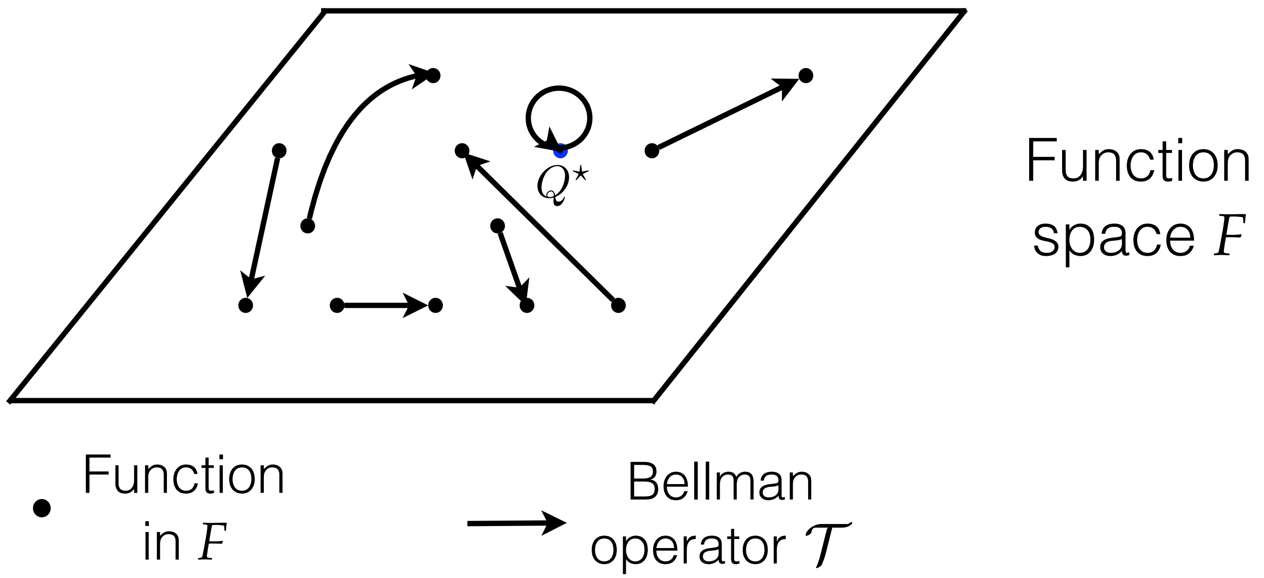


Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

Prior work: Point-wise pessimism [JZW'21]

- Learn $\widehat{Q}^\pi \leq Q^\pi \Longrightarrow \widehat{J}(\pi) \leq J(\pi)$
- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true Q^π linear, but $\widehat{Q}^\pi$ quadratic
- Strong assumptions for point-wise uncertainty

"Bellman-completeness"
$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathcal{F}$

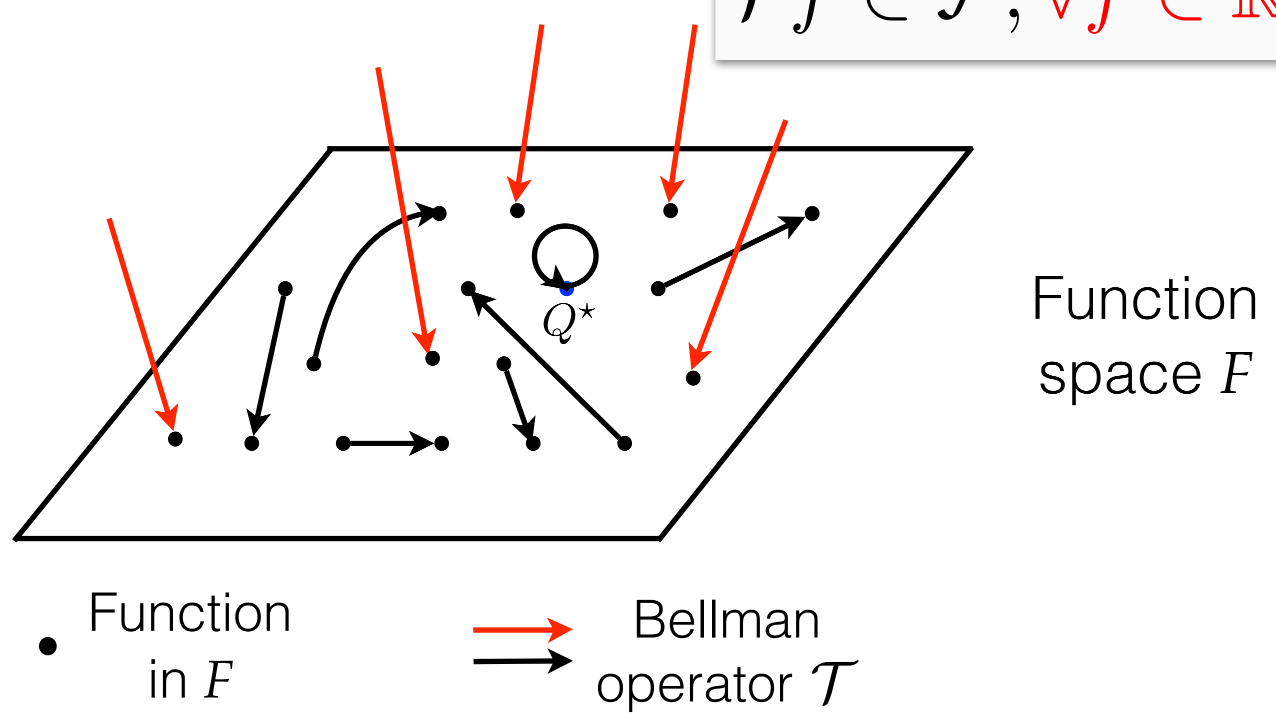


Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

Prior work: Point-wise pessimism [JZW'21]

- Learn $\widehat{Q}^\pi \leq Q^\pi \Longrightarrow \widehat{J}(\pi) \leq J(\pi)$
- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true Q^π linear, but $\widehat{Q}^\pi$ quadratic
- Strong assumptions for point-wise uncertainty

$$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$$

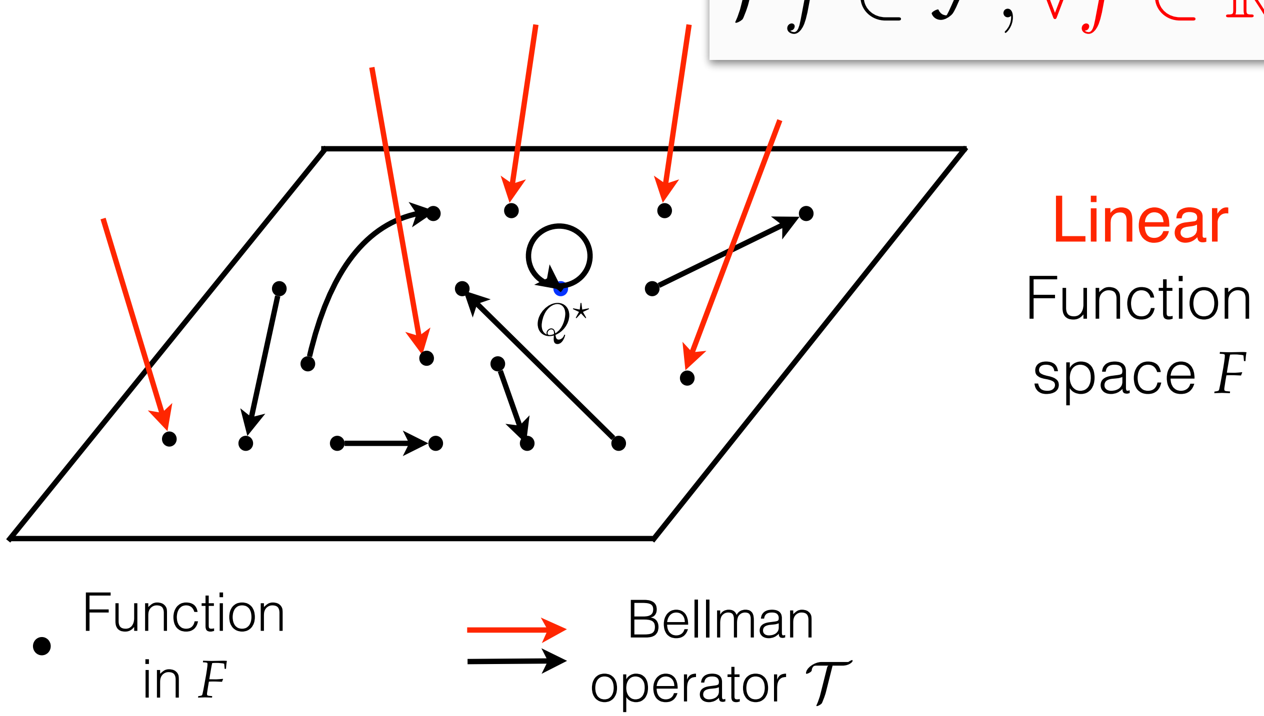


Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

Prior work: Point-wise pessimism [JZW'21]

- Learn $\widehat{Q}^\pi \le Q^\pi \Longrightarrow \widehat{J}(\pi) \le J(\pi)$
- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true Q^π linear, but $\widehat{Q}^\pi$ quadratic
- Strong assumptions for point-wise uncertainty

$$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$$

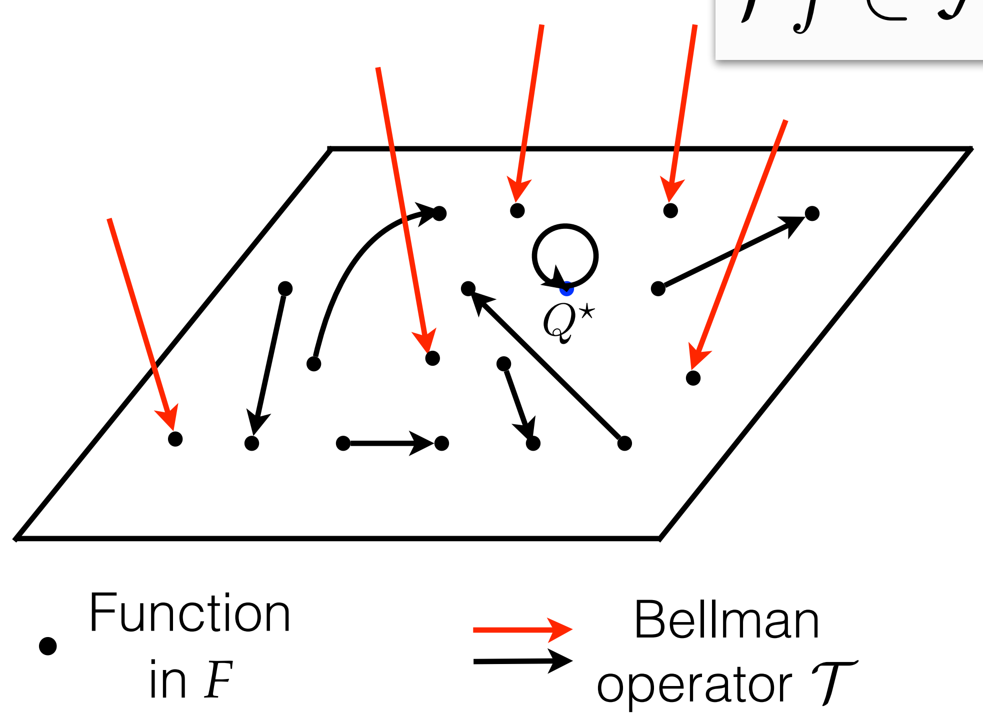


Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

Prior work: Point-wise pessimism [JZW'21]

- Learn $\widehat{Q}^\pi \le Q^\pi \Longrightarrow \widehat{J}(\pi) \le J(\pi)$
- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true Q^π linear, but $\widehat{Q}^\pi$ quadratic
- Strong assumptions for point-wise uncertainty

$$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$$



Linear Function space F

$\Big\} \Longleftrightarrow$ Linear MDP [JYWJ'20]

Restrict MDP dynamics

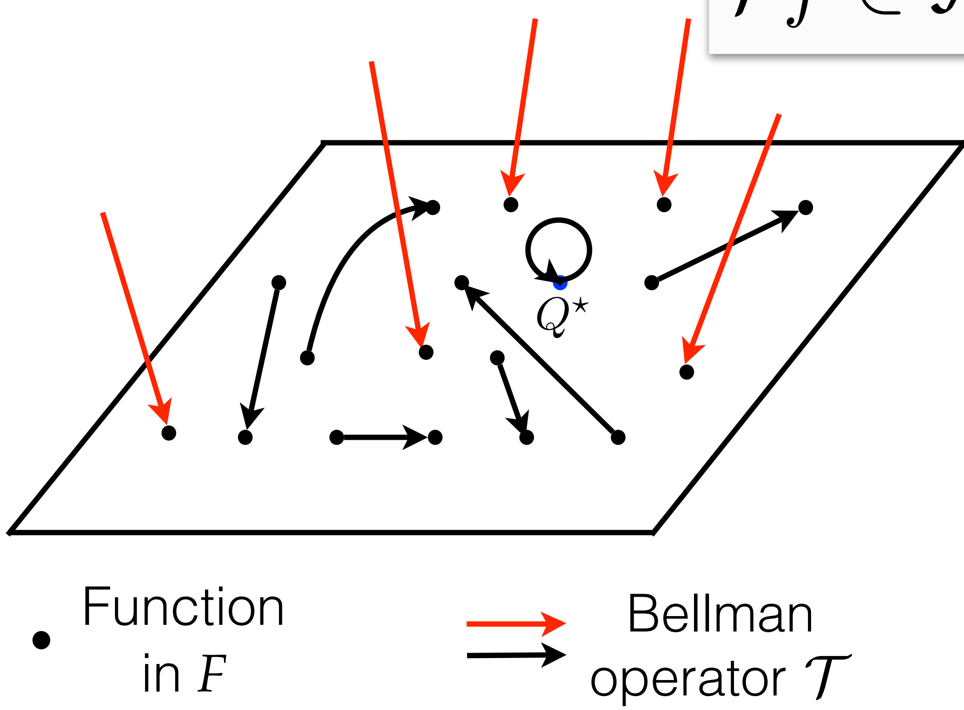
P(s'|s,a) = ✓ x ?

Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

Prior work: Point-wise pessimism [JZW'21]

- Learn $\widehat{Q}^{\pi} \leq Q^{\pi} \Longrightarrow \widehat{J}(\pi) \leq J(\pi)$
- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true Q^{π} linear, but $\widehat{Q}^{\pi}$ quadratic
- Strong assumptions for point-wise uncertainty

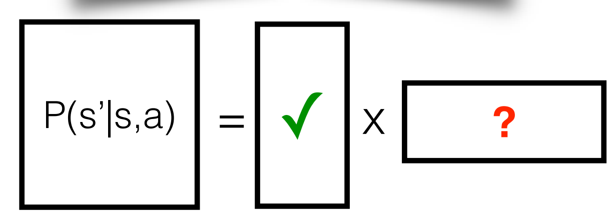
$$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$$



Linear Function space F

$\Leftrightarrow$ Linear MDP [JYWJ'20]

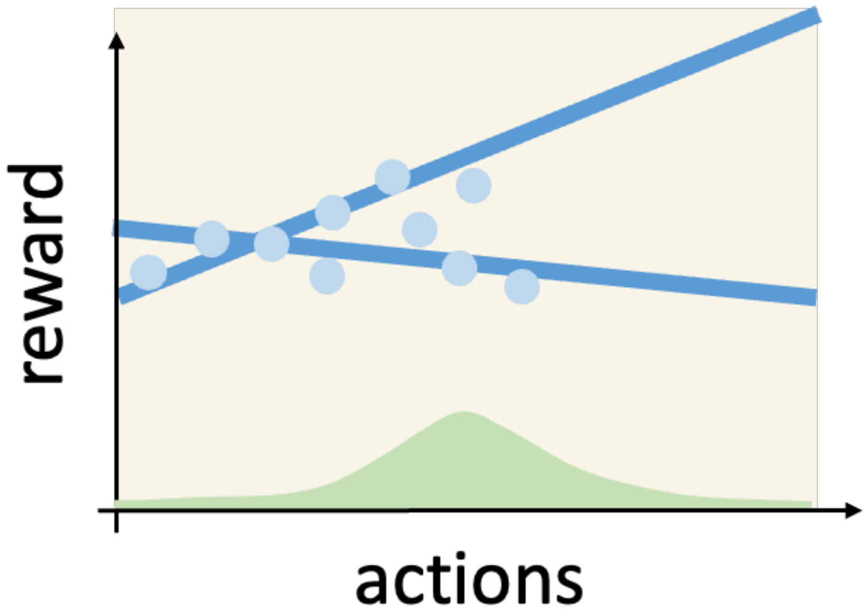
Restrict MDP dynamics



Beyond linear MDPs?
- No restriction on MDP
- Allow non-linear F

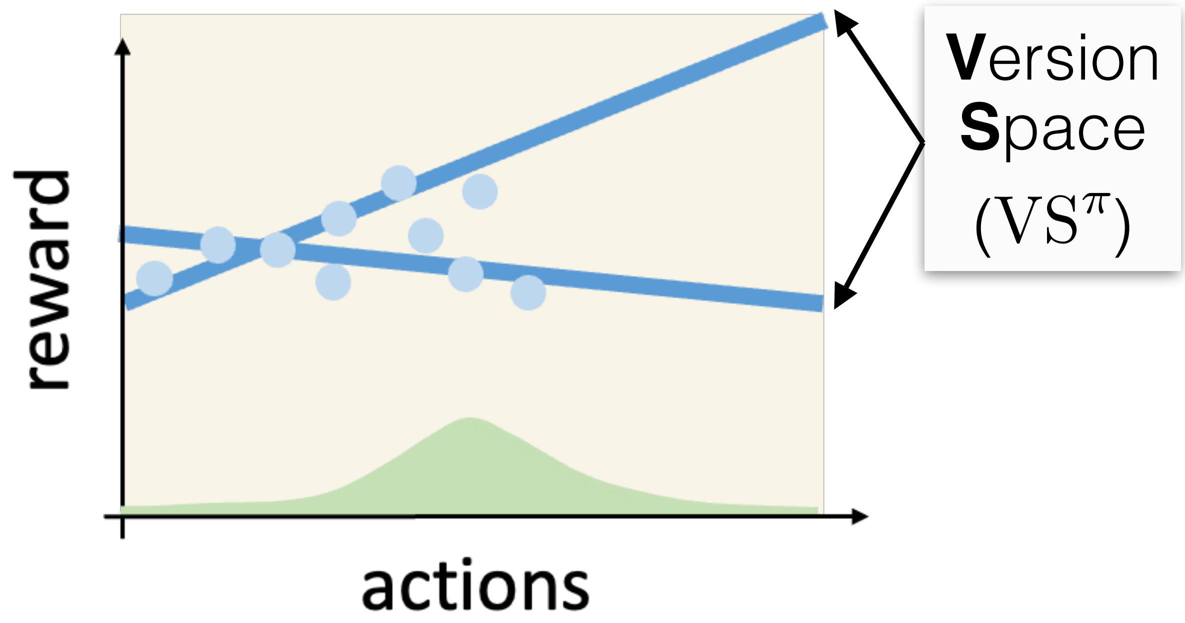
Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

Bellman-consistent pessimism [XCJMA’21]



Antos et al. ML’08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.
Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Bellman-consistent pessimism [XCJMA'21]

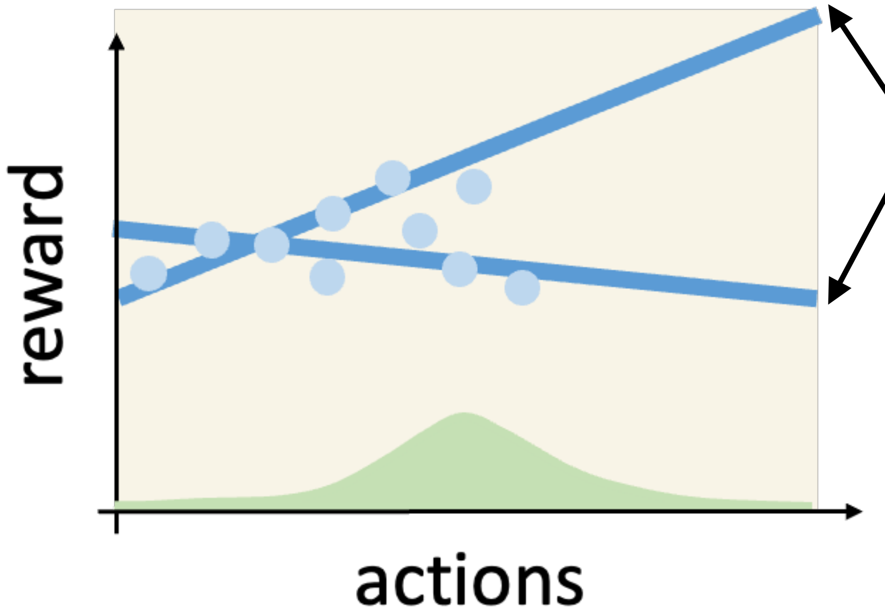


$$\widehat{J}(\pi) = \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

Antos et al. ML'08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.

Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Bellman-consistent pessimism [XCJMA'21]



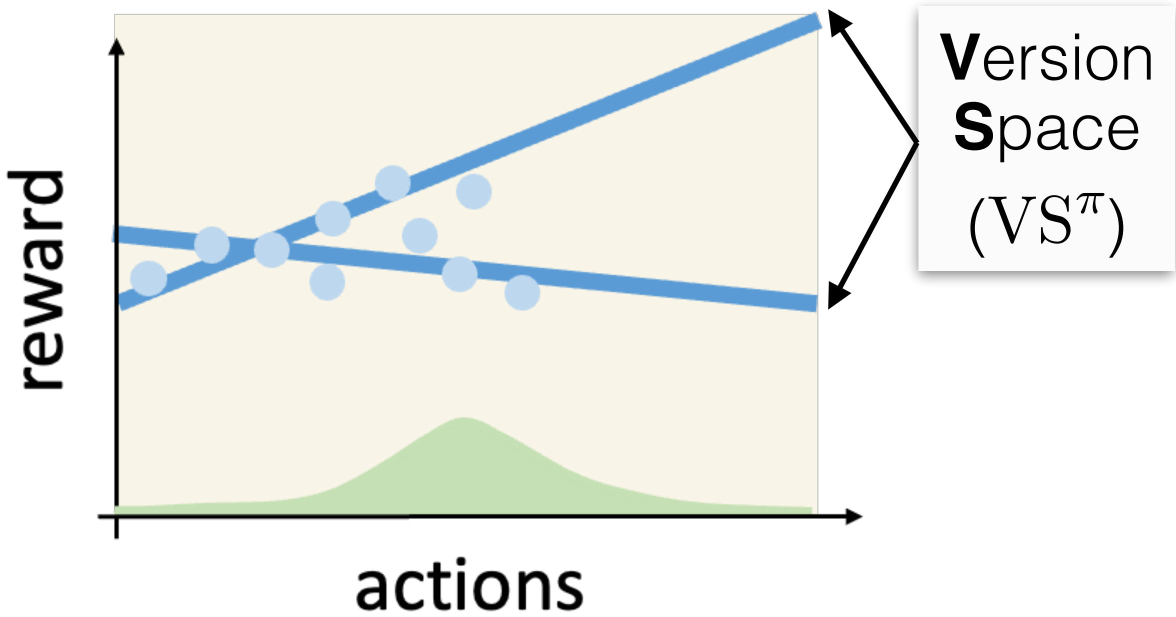
Version **S**pace (VS^π)

$$\mathrm{VS}^\pi := \{f : \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2] \approx 0\}$$

$$\widehat{J}(\pi) = \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

Antos et al. ML'08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.
Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Bellman-consistent pessimism [XCJMA'21]

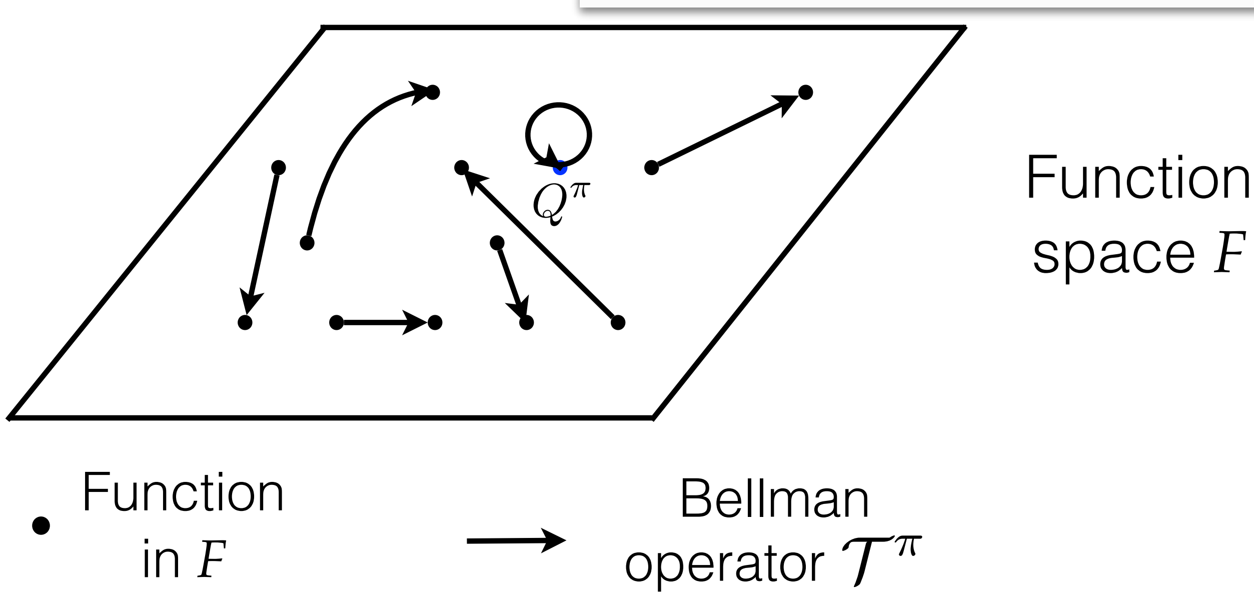


$$\mathrm{VS}^\pi := \{f : \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2] \approx 0\}$$

$$\widehat{J}(\pi) = \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

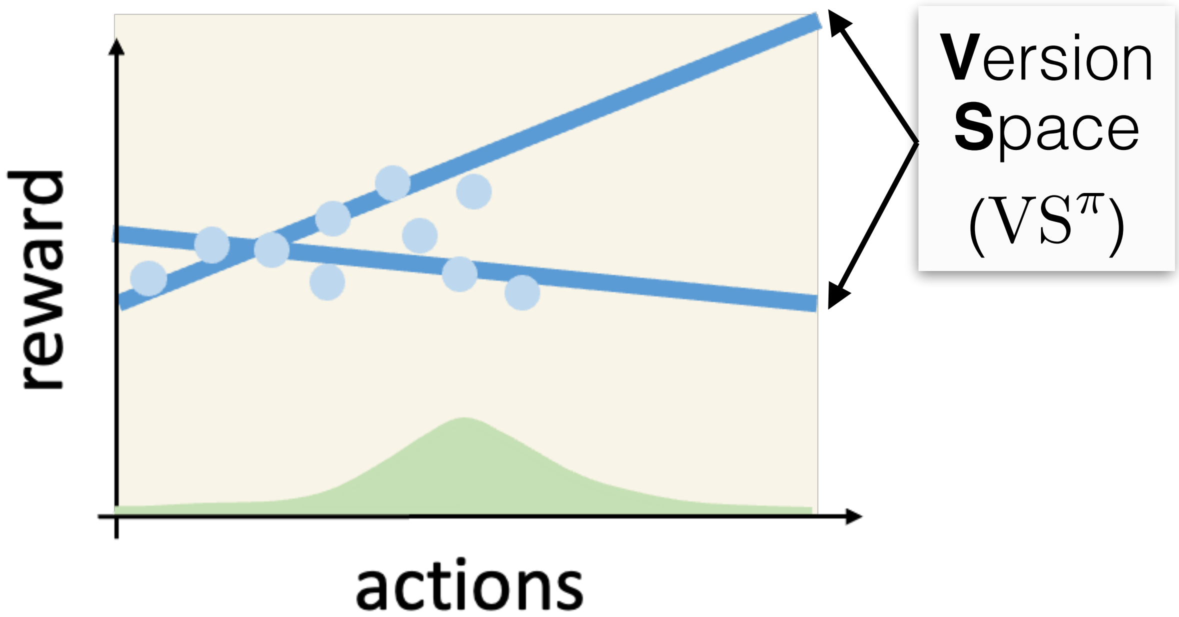
"Bellman-completeness" [AMS'08]

$\mathcal{T}^\pi f \in \mathcal{F}, \forall f \in \mathcal{F}, \pi \in \Pi$



Antos et al. ML'08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.
Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Bellman-consistent pessimism [XCJMA'21]



$$\mathrm{VS}^\pi := \{f : \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2] \approx 0\}$$

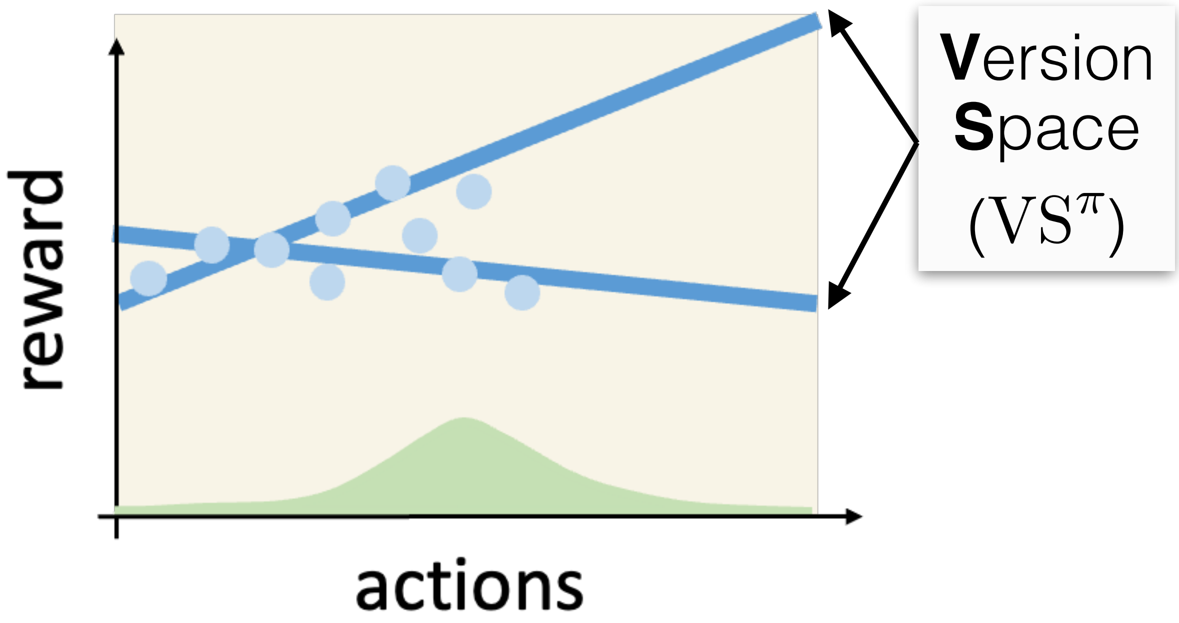
$$\widehat{J}(\pi) = \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true Q^π linear, but $\widehat{Q}^\pi$ quadratic

- Strong assumptions for point-wise uncertainty
  - Restrict MDP dynamics

Antos et al. ML'08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.
Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Bellman-consistent pessimism [XCJMA'21]



$$\mathrm{VS}^\pi := \{f : \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2] \approx 0\}$$

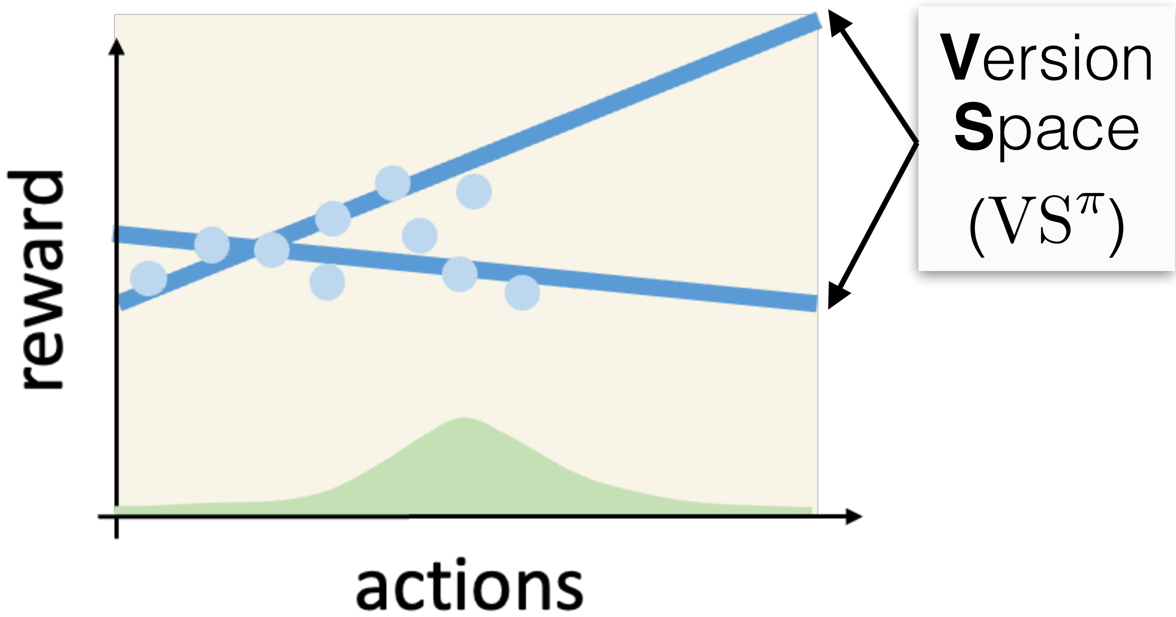
$$\widehat{J}(\pi) = \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

- Less conservative: only plausible scenarios
  - *Ex.* In linear MDPs, argmin f is linear

- Strong assumptions for point-wise uncertainty
  - Restrict MDP dynamics

Antos et al. ML'08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.
Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Bellman-consistent pessimism [XCJMA'21]



$$\mathrm{VS}^\pi := \{f : \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2] \approx 0\}$$

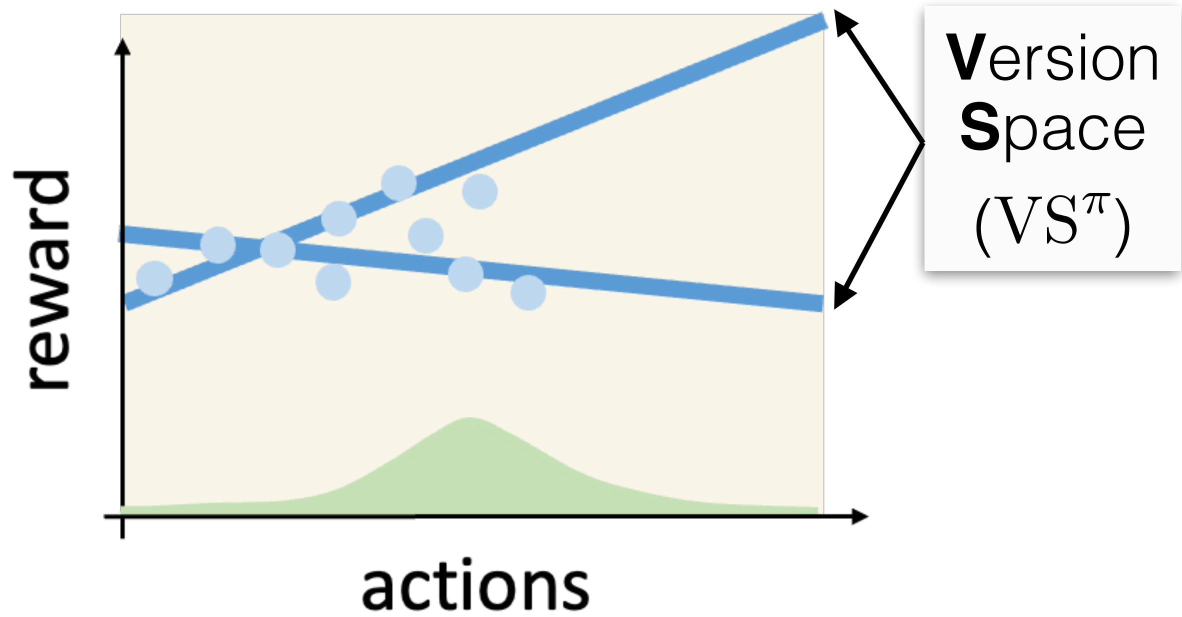
$$\widehat{J}(\pi) = \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

- Less conservative: only plausible scenarios
  - *Ex.* In linear MDPs, argmin f is linear

- Standard representation assumptions [AMS'08]
  - No restriction on MDP & allow non-linear F

Antos et al. ML'08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.
Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Bellman-consistent pessimism [XCJMA'21]



$$\mathrm{VS}^\pi := \{f : \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2] \approx 0\}$$

$$\widehat{J}(\pi) = \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

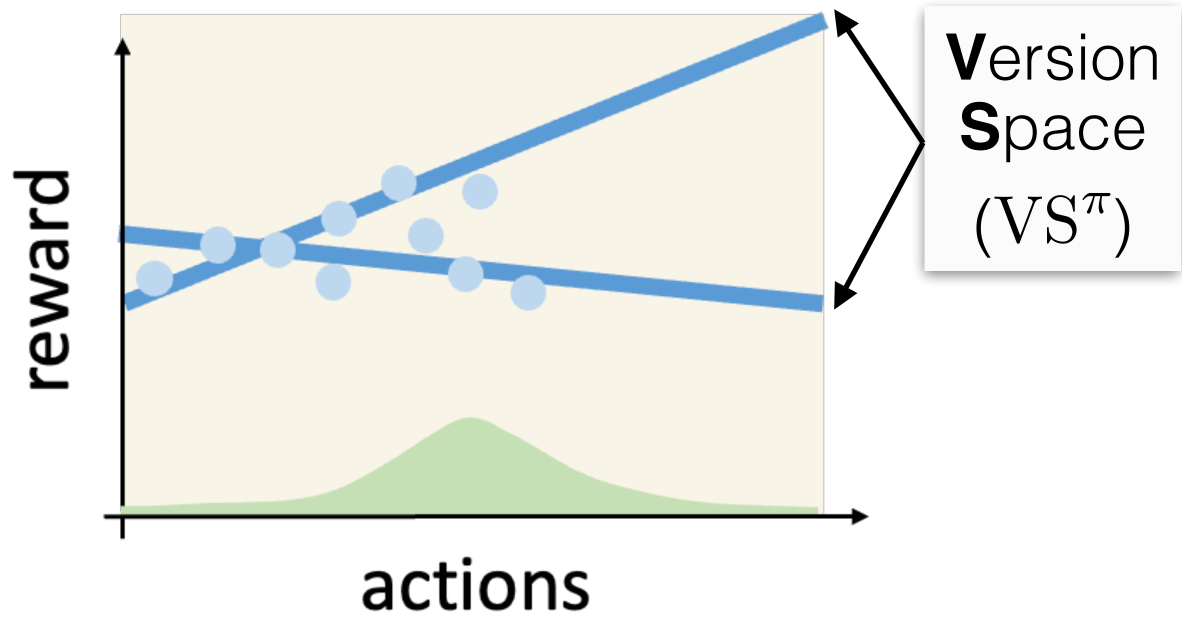
- Less conservative: only plausible scenarios
  - *Ex.* In linear MDPs, argmin f is linear
  - Rate improved over [JZW'21] in linear MDPs

- Standard representation assumptions [AMS'08]
  - No restriction on MDP & allow non-linear F

Formal guarantee in backup slide

Antos et al. ML'08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.
Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Bellman-consistent pessimism [XCJMA'21]



$$\mathrm{VS}^\pi := \{f : \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2] \approx 0\}$$

$$\arg\max_{\pi \in \Pi} \widehat{J}(\pi) = \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

Computational efficiency?

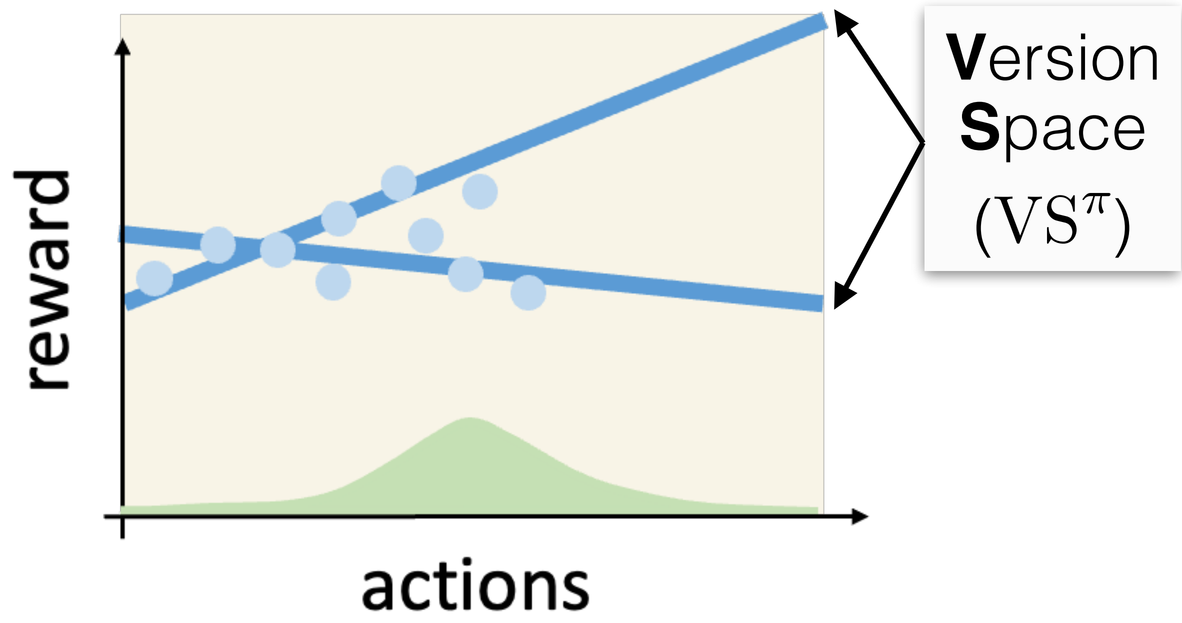
- Less conservative: only plausible scenarios
  - *Ex.* In linear MDPs, argmin f is linear
  - Rate **improved** over [JZW'21] in linear MDPs

- Standard representation assumptions [AMS'08]
  - No restriction on MDP & allow non-linear F

Formal guarantee in backup slide

Antos et al. ML'08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.
Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Bellman-consistent pessimism [XCJMA'21]



$$\mathrm{VS}^\pi := \{f : \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2] \approx 0\}$$

$$\arg\max_{\pi \in \Pi} \widehat{J}(\pi) = \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

Computational efficiency?

$$\arg\max_{\pi\in\Pi} \min_{f\in\mathcal{F}} \max_{f'\in\mathcal{F}} f(s_0,\pi) + \lambda \mathbb{E}_D[(f(s,a) - r - \gamma f(s',\pi))^2] - \mathbb{E}_D[(f'(s,a) - r - \gamma f(s',\pi))^2]$$

- Less conservative: only plausible scenarios
  - *Ex.* In linear MDPs, argmin f is linear
  - Rate improved over [JZW'21] in linear MDPs

- Standard representation assumptions [AMS'08]
  - No restriction on MDP & allow non-linear F

Formal guarantee in backup slide

Antos et al. ML'08. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.

Xie et al. NeurIPS-21. Bellman-consistent Pessimism for Offline Reinforcement Learning.

Online RL

Offline RL

$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

$$\arg\max_{\pi \in \Pi} \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

Bellman-consistent
(version space)

Point-wise

	Online RL	Offline RL
Bellman-consistent (version space)	$\arg\max_{\pi\in\Pi} \max_{f\in \mathrm{VS}^\pi} f(s_0, \pi)$	$\arg\max_{\pi\in\Pi} \min_{f\in \mathrm{VS}^\pi} f(s_0, \pi)$
Point-wise		

	Online RL	Offline RL
Bellman-consistent (version space)	$\arg\max_{\pi\in\Pi} \max_{f\in\mathrm{VS}^\pi} f(s_0,\pi)$ • Statistical guarantee in very general settings [JKALS'17] ✓	$\arg\max_{\pi\in\Pi} \min_{f\in\mathrm{VS}^\pi} f(s_0,\pi)$
Point-wise		

Online RL | Offline RL

Bellman-consistent (version space)

$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17] ✓
- NP-hardness under strong oracles [DJKALS'18] X

$$\arg\max_{\pi \in \Pi} \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

Point-wise

Online RL

Offline RL

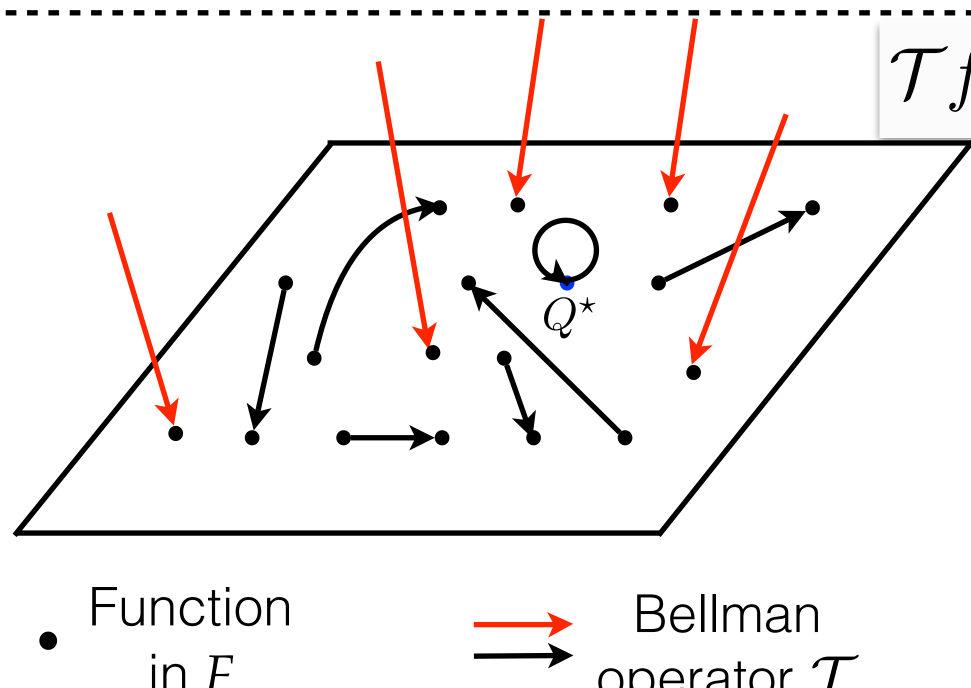
Bellman-consistent (version space)

$$\arg\max_{\pi\in\Pi} \max_{f\in \mathrm{VS}^\pi} f(s_0,\pi)$$

- Statistical guarantee in very general settings [JKALS'17] ✓
- NP-hardness under strong oracles [DJKALS'18] X

$$\arg\max_{\pi\in\Pi} \min_{f\in \mathrm{VS}^\pi} f(s_0,\pi)$$

Point-wise



$\mathcal{T}f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$

Linear Function space F

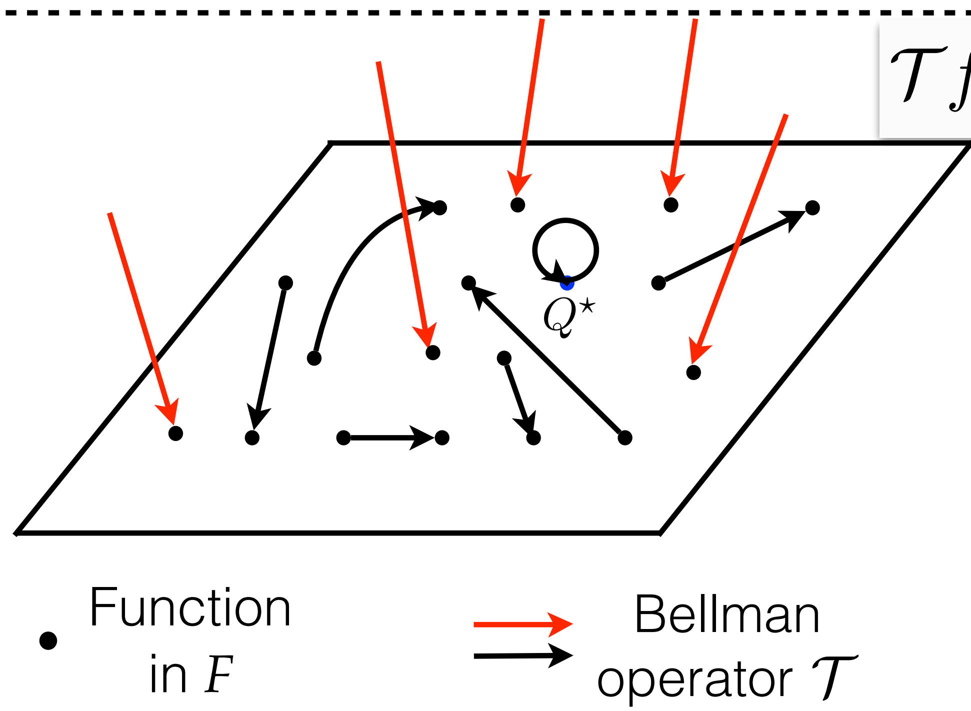
$\Leftrightarrow$ Linear MDP [JYWJ'20]

Online RL	Offline RL
$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$	$\arg\max_{\pi \in \Pi} \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$

Bellman-consistent (version space)

- Statistical guarantee in very general settings [JKALS'17] ✓
- NP-hardness under strong oracles [DJKALS'18] X

Point-wise



Linear Function space F

$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ } $\Leftrightarrow$ Linear MDP [JYWJ'20]

Computationally efficient ✓

Online RL

Offline RL

Bellman-consistent (version space)

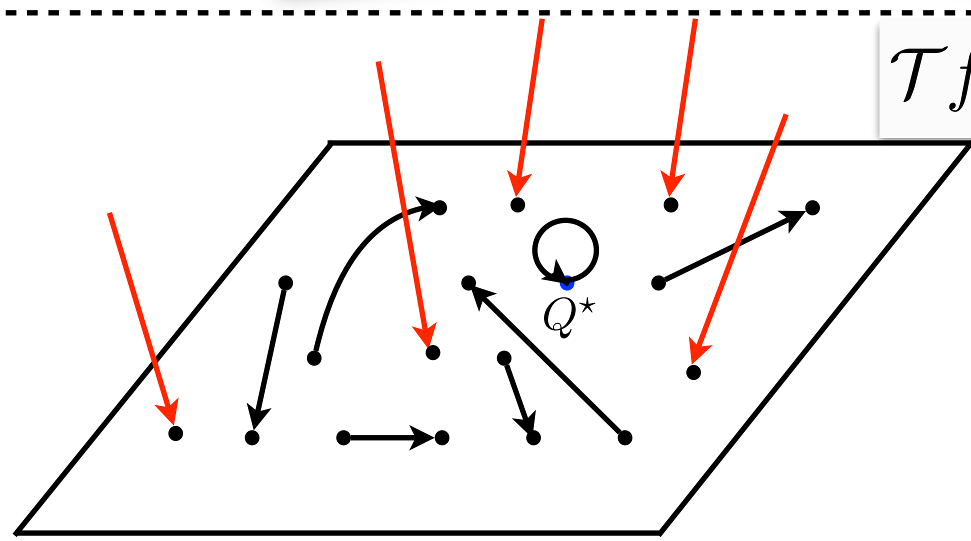
$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17] ✓
- NP-hardness under strong oracles [DJKALS'18] X

$$\arg\max_{\pi \in \Pi} \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

Statistical generality
vs
Computational tractability?

Point-wise



$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ } ⇔ Linear MDP [JYWJ'20]

Linear Function space F

Computationally efficient ✓

• Function in F

Bellman operator $\mathcal{T}$

Online RL

Offline RL

Bellman-consistent (version space)

$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17] ✓
- NP-hardness under strong oracles [DJKALS'18] X

$$\arg\max_{\pi \in \Pi} \min_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

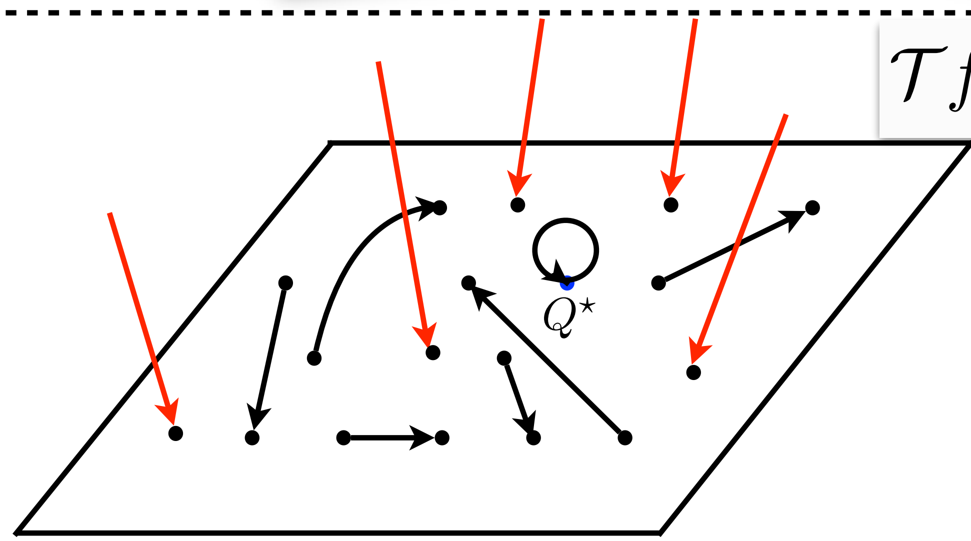
$$f_k \leftarrow \arg\min_{f \in \mathrm{VS}^{\pi_k}} f(s_0, \pi_k)$$

$$\pi_{k+1} \leftarrow \mathrm{NoRegret}(\pi_k, f_k)$$

- Oracle-efficient ! ✓

Statistical generality
vs
Computational tractability?

Point-wise



$$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$$

Linear Function space F

$\Leftrightarrow$ Linear MDP [JYWJ'20]

Computationally efficient ✓

• Function in F

→ Bellman operator $\mathcal{T}$

Freund and Schapire.1999. Adaptive game playing using multiplicative weights.

Bellman-consistent (version space)

Online RL

$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^{\pi}} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17] ✓
- NP-hardness under strong oracles [DJKALS'18] ✗

Statistical generality
vs
Computational tractability?

Offline RL

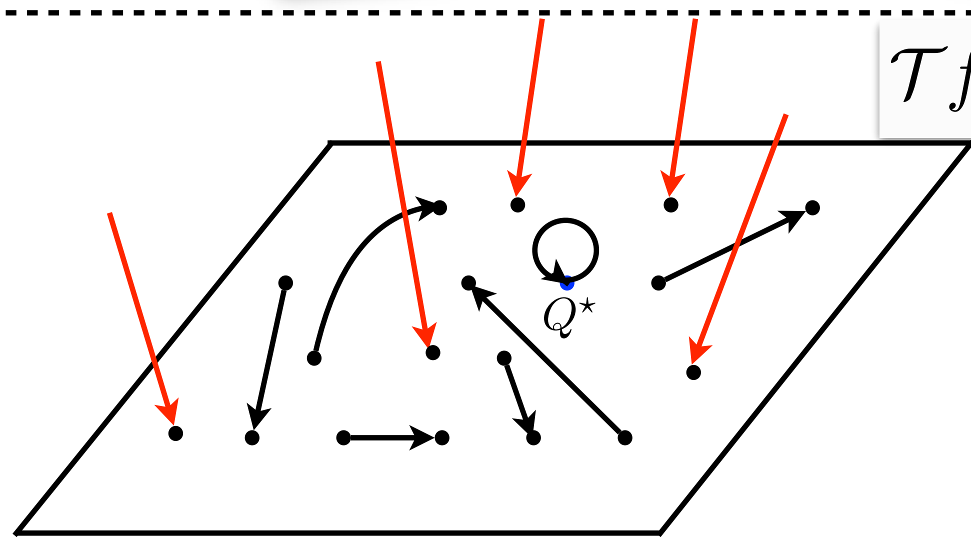
$$\arg\max_{\pi \in \Pi} \min_{f \in \mathrm{VS}^{\pi}} f(s_0, \pi)$$

$$f_k \leftarrow \arg\min_{f \in \mathrm{VS}^{\pi_k}} f(s_0, \pi_k)$$

$$\pi_{k+1} \leftarrow \mathrm{NoRegret}(\pi_k, f_k)$$

- Oracle-efficient ! ✓
- Oracle itself is efficient in the linear setting (pessimistic LSTD)

Point-wise



$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$

Linear Function space F

$\Leftrightarrow$ Linear MDP [JYWJ'20]

Computationally efficient ✓

• Function in F

Bellman operator $\mathcal{T}$

Freund and Schapire.1999. Adaptive game playing using multiplicative weights.

Robustness of offline RL

Example: network control

- Status quo: time-tested heuristics

Robustness of offline RL

Example: network control

- Status quo: time-tested heuristics
- RL: training instability
- Amplified by the difficulty of h-p tuning

Robustness of offline RL

Example: network control

- Status quo: time-tested heuristics
- RL: training instability
- Amplified by the difficulty of h-p tuning

Imitation Learning

WHEN SHOULD WE PREFER OFFLINE REINFORCEMENT LEARNING OVER BEHAVIORAL CLONING?

Aviral Kumar[*,1,2], Joey Hong[*,1], Anikait Singh[1], Sergey Levine[1,2]

Robustness of offline RL

Example: network control

- Status quo: time-tested heuristics
- RL: training instability
- Amplified by the difficulty of h-p tuning

Imitation Learning

- Reliably learn the data policy

Offline RL

- Can be worse than data policy

Robustness of offline RL

Example: network control

- Status quo: time-tested heuristics
- RL: training instability
- Amplified by the difficulty of h-p tuning

Imitation Learning

- Reliably learn the data policy
- Performance ceiling

Offline RL

- Can be worse than data policy
- Potential for optimality

Robustness of offline RL

Example: network control

- Status quo: time-tested heuristics
- RL: training instability
- Amplified by the difficulty of h-p tuning

Imitation Learning

- Reliably learn the data policy
- Performance ceiling

Offline RL

- Can be worse than data policy
- Potential for optimality

Robustness of offline RL

Example: network control

- Status quo: time-tested heuristics
- RL: training instability
- Amplified by the difficulty of h-p tuning

Imitation Learning

- Reliably learn the data policy
- Performance ceiling

Offline RL

- Can be worse than data policy
- Potential for optimality

Best of both worlds?

ATAC: *Relative* Pessimism [CXJA'22]

$$\arg\max_{\pi \in \Pi} \text{tight lower bound of } J(\pi)$$

Cheng et al. ICML-22. Adversarially Trained Actor Critic for Offline Reinforcement Learning.
Swamy et al. ICML-21. Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap.

ATAC: *Relative* Pessimism [CXJA'22]

data policy

$$\arg\max_{\pi \in \Pi} \text{tight lower bound of } J(\pi) - J(\pi_D)$$

Cheng et al. ICML-22. Adversarially Trained Actor Critic for Offline Reinforcement Learning.
Swamy et al. ICML-21. Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap.

ATAC: *Relative* Pessimism [CXJA'22]

data policy

$$\arg\max_{\pi\in\Pi} \text{ tight lower bound of } J(\pi) - J(\pi_D)$$

"Performance-diff Lemma" [Langford & Kakade'02]

$$J(\pi) - J(\pi_D) \propto \mathbb{E}_{(s,a)\sim D}[Q^\pi(s,\pi) - Q^\pi(s,a)]$$

$$\arg\max_{\pi\in\Pi} \mathbb{E}_{(s,a)\sim D}[\widehat{Q}^\pi(s,\pi) - \widehat{Q}^\pi(s,a)]$$

$$\text{where } \widehat{Q}^\pi \in \arg\min_{f\in\mathcal{F}} \mathbb{E}_{(s,a)\sim D}[f(s,\pi) - f(s,a)] + \lambda\, \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2]$$

Cheng et al. ICML-22. Adversarially Trained Actor Critic for Offline Reinforcement Learning.
Swamy et al. ICML-21. Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap.

ATAC: *Relative* Pessimism [CX**J**A'22]

$$\arg\max_{\pi\in\Pi} \text{tight lower bound of } J(\pi) - J(\pi_D)$$

"Performance-diff Lemma" [Langford & Kakade'02]

$$J(\pi) - J(\pi_D) \propto \mathbb{E}_{(s,a)\sim D}[Q^\pi(s,\pi) - Q^\pi(s,a)]$$

$$\arg\max_{\pi\in\Pi} \mathbb{E}_{(s,a)\sim D}[\widehat{Q}^\pi(s,\pi) - \widehat{Q}^\pi(s,a)]$$

$$\text{where} \quad \widehat{Q}^\pi \in \arg\min_{f\in\mathcal{F}} \mathbb{E}_{(s,a)\sim D}[f(s,\pi) - f(s,a)] + \lambda\, \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2]$$

- λ small (≈ 0): (adversarial) **Imitation Learning**!
  - strong discriminator (π must imitate π_D)
  - IL requires weaker assumptions ($\pi_D \in \Pi + Q^\pi \in \mathcal{F}, \forall \pi \in \Pi$)

Cheng et al. ICML-22. Adversarially Trained Actor Critic for Offline Reinforcement Learning.
Swamy et al. ICML-21. Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap.

ATAC: *Relative* Pessimism [CXJA'22]

$$\arg\max_{\pi \in \Pi} \text{ tight lower bound of } J(\pi) - J(\pi_D)$$

"Performance-diff Lemma" [Langford & Kakade'02]

$$J(\pi) - J(\pi_D) \propto \mathbb{E}_{(s,a)\sim D}[Q^\pi(s,\pi) - Q^\pi(s,a)]$$

$$\arg\max_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim D}[\widehat{Q}^\pi(s,\pi) - \widehat{Q}^\pi(s,a)]$$

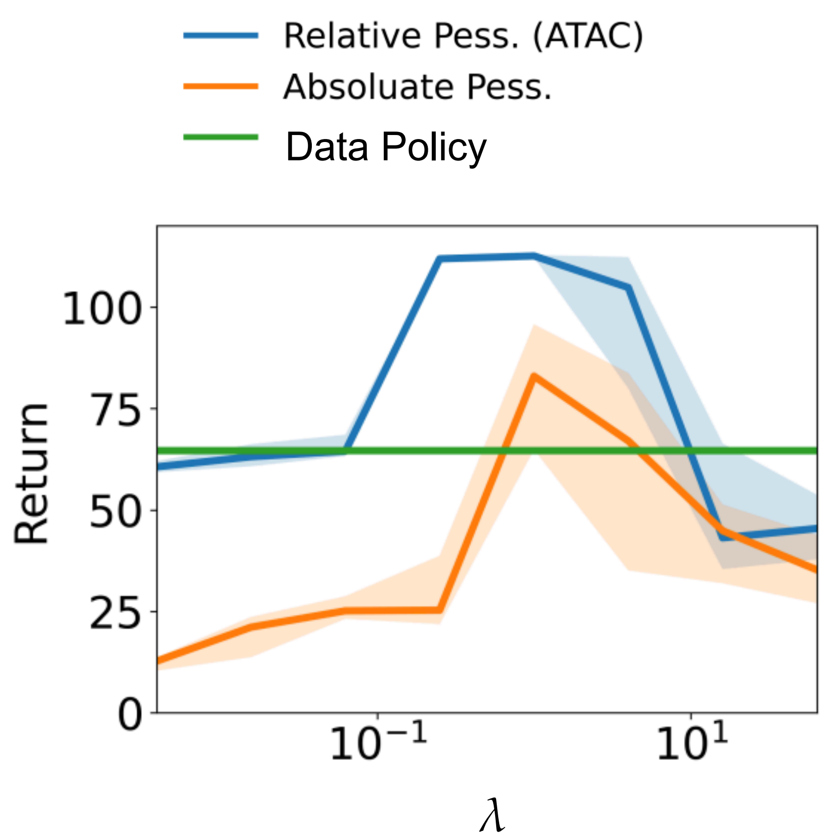
$$\text{where } \widehat{Q}^\pi \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(s,a)\sim D}[f(s,\pi) - f(s,a)] + \lambda\, \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2]$$

Bellman-error regularization

- λ small (≈ 0): (adversarial) **Imitation Learning**!
  - strong discriminator (π must imitate π_D)
  - IL requires weaker assumptions ($\pi_D \in \Pi + Q^\pi \in \mathcal{F}, \forall \pi \in \Pi$)
- Well-specified λ: offline RL
  - weakens discriminator, allowing π to further improve

Cheng et al. ICML-22. Adversarially Trained Actor Critic for Offline Reinforcement Learning.
Swamy et al. ICML-21. Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap.

Empirical evaluation



(d) hopper-medium-expert

	Behavior	ATAC*	CQL	COMBO	TD3BC	IQL	BC
halfcheetah-rand	-0.1	4.8	35.4	**38.8**	10.2	-	2.1
walker2d-rand	0.0	**8.0**	7.0	7.0	1.4	-	1.6
hopper-rand	1.2	**31.8**	10.8	17.9	11.0	-	9.8
halfcheetah-med	40.6	**54.3**	44.4	**54.2**	42.8	47.4	36.1
walker2d-med	62.0	**91.0**	74.5	75.5	79.7	78.3	6.6
hopper-med	44.2	**102.8**	86.6	94.9	**99.5**	66.3	29.0
halfcheetah-med-replay	27.1	49.5	46.2	**55.1**	43.3	44.2	38.4
walker2d-med-replay	14.8	**94.1**	32.6	56.0	25.2	73.9	11.3
hopper-med-replay	14.9	**102.8**	48.6	73.1	31.4	94.7	11.8
halfcheetah-med-exp	64.3	**95.5**	62.4	90.0	**97.9**	86.7	35.8
walker2d-med-exp	82.6	**116.3**	98.7	96.1	101.1	**109.6**	6.4
hopper-med-exp	64.7	**112.6**	**111.0**	**111.1**	**112.2**	91.5	**111.9**
pen-human	207.8	79.3	37.5	-	-	71.5	34.4
hammer-human	25.4	**6.7**	**4.4**	-	-	1.4	1.5
door-human	28.6	8.7	**9.9**	-	-	4.3	0.5
relocate-human	86.1	**0.3**	**0.2**	-	-	**0.1**	**0.0**
pen-cloned	107.7	73.9	39.2	-	-	37.3	56.9
hammer-cloned	8.1	2.3	2.1	-	-	2.1	0.8
door-cloned	12.1	**8.2**	0.4	-	-	1.6	-0.1
relocate-cloned	28.7	**0.8**	**-0.1**	-	-	**-0.2**	**-0.1**
pen-exp	105.7	**159.5**	107.0	-	-	-	85.1
hammer-exp	96.3	**128.4**	86.7	-	-	-	**125.6**
door-exp	100.5	**105.5**	**101.5**	-	-	-	34.9
relocate-exp	101.6	**106.5**	95.0	-	-	-	**101.3**

New perspective that bridges IL and offline RL

- IL ($\lambda \approx 0$): strong discriminator (π must imitate π_D)
- RL weakens discriminator, allowing π to further improve

Cheng et al. ICML-22. Adversarially Trained Actor Critic for Offline Reinforcement Learning.

Online RL

Offline RL

Bellman-consistent (version space)

Online RL:

$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^{\pi}} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17] ✓
- NP-hardness under strong oracles [DJKALS'18] ✗

Offline RL:

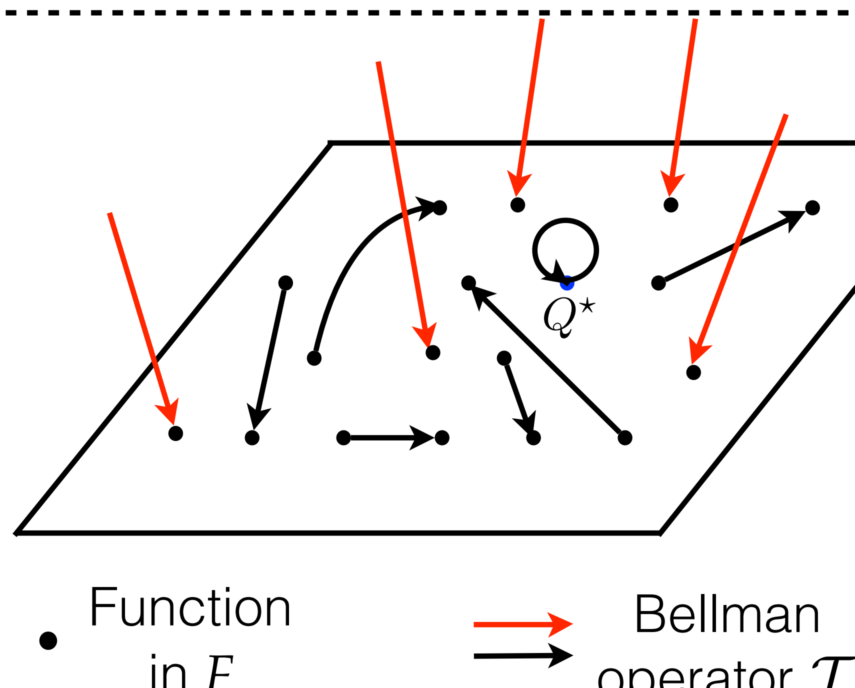
$$\arg\max_{\pi \in \Pi} \min_{f \in \mathrm{VS}^{\pi}} f(s_0, \pi)$$

$$f_k \leftarrow \arg\min_{f \in \mathrm{VS}^{\pi_k}} f(s_0, \pi_k)$$

$$\pi_{k+1} \leftarrow \mathrm{NoRegret}(\pi_k, f_k)$$

- Oracle-efficient ! ✓
- Oracle itself is efficient in the linear setting (pessimistic LSTD)

Point-wise



Linear Function space F

$$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$$

$\Leftrightarrow$ Linear MDP [JYWJ'20]

Computationally efficient ✓

Online RL

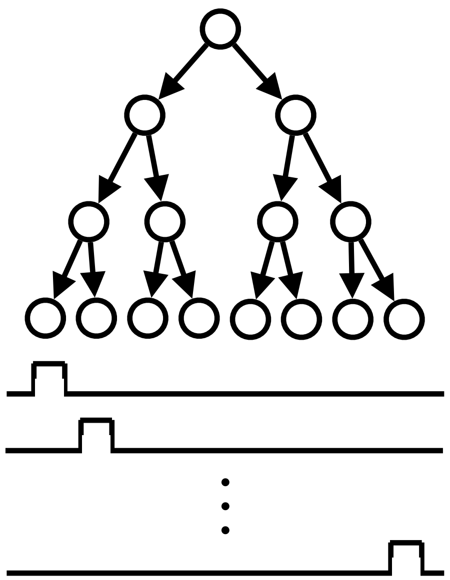
$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17]

Online RL

$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^{\pi}} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17]



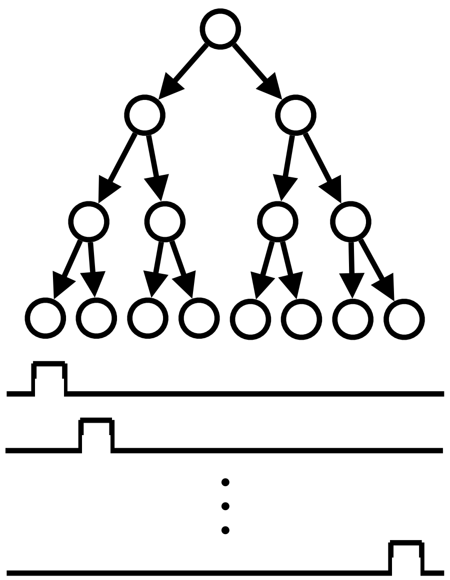
Bounded statistical complexities (e.g., VC-type dim) are insufficient! [KAL'16, JKALS'17]

Online RL

$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^\pi} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17]

Structural assumptions



Bounded statistical complexities (e.g., VC-type dim) are insufficient! [KAL'16, JKALS'17]

Online RL

$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^{\pi}} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17]

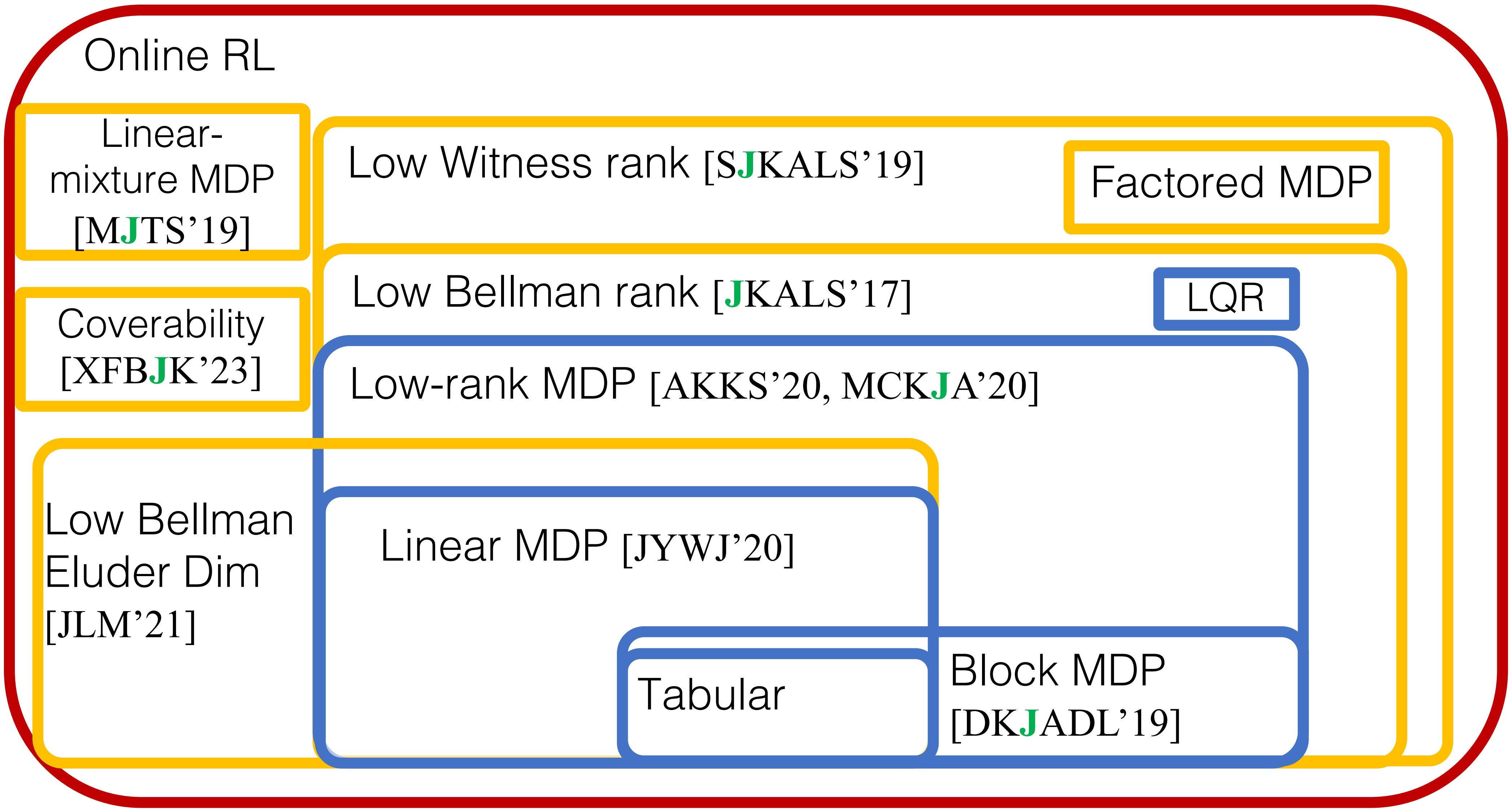
Structural assumptions

Low Bellman rank [JKALS'17]

Online RL

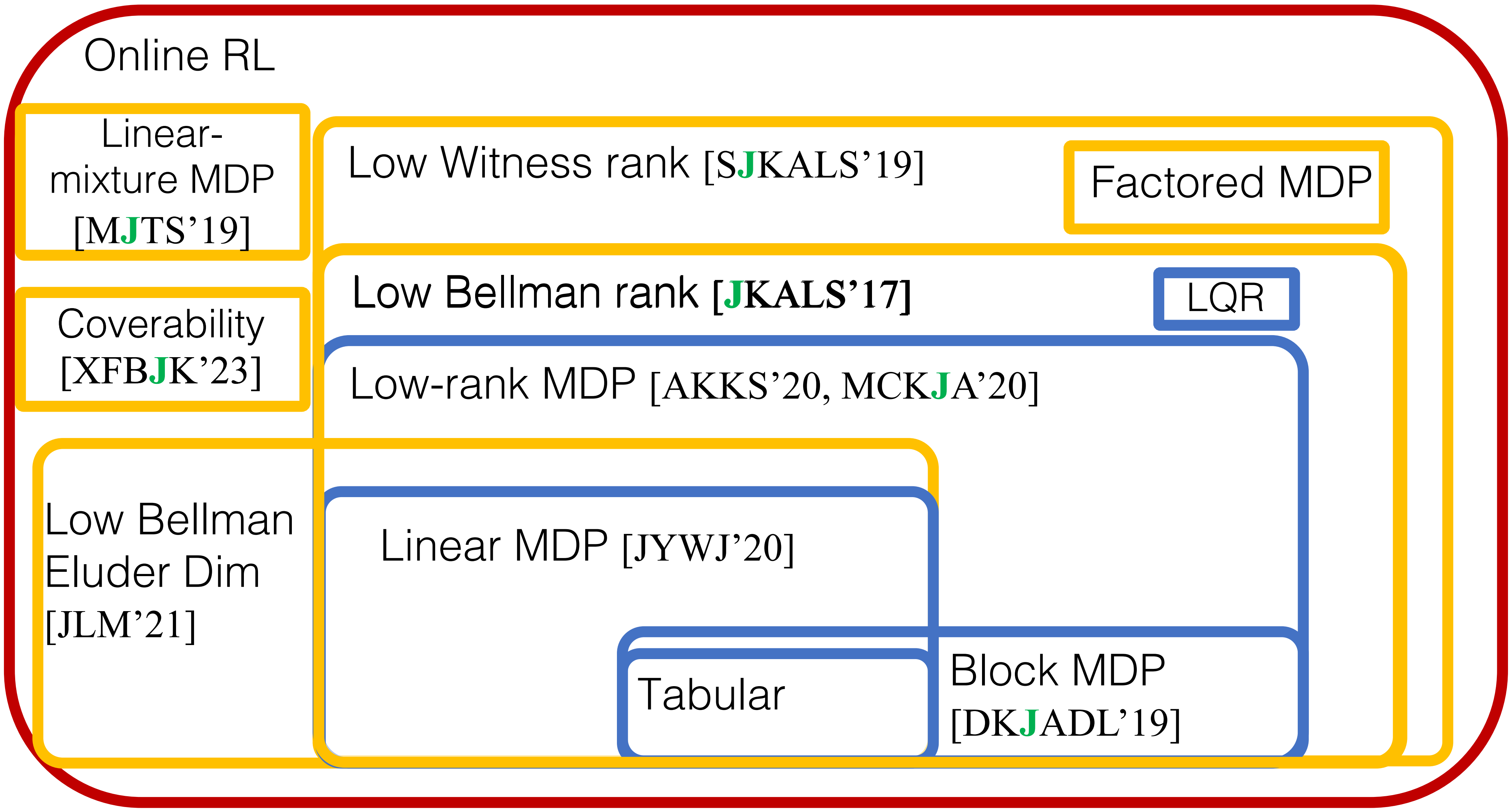
$$\arg\max_{\pi \in \Pi} \max_{f \in \mathrm{VS}^{\pi}} f(s_0, \pi)$$

- Statistical guarantee in very general settings [JKALS'17]



- Adapted from FOCS'20 Tutorial by Agarwal, Krishnamurthy, and Langford
- Also related: bilinear classes [DKLLMSW'21], DEC [FKQR'21]
- Bellman-eluder [JLM'21] generalizes deterministic version of [RvR'13]

- All consider modeling value functions

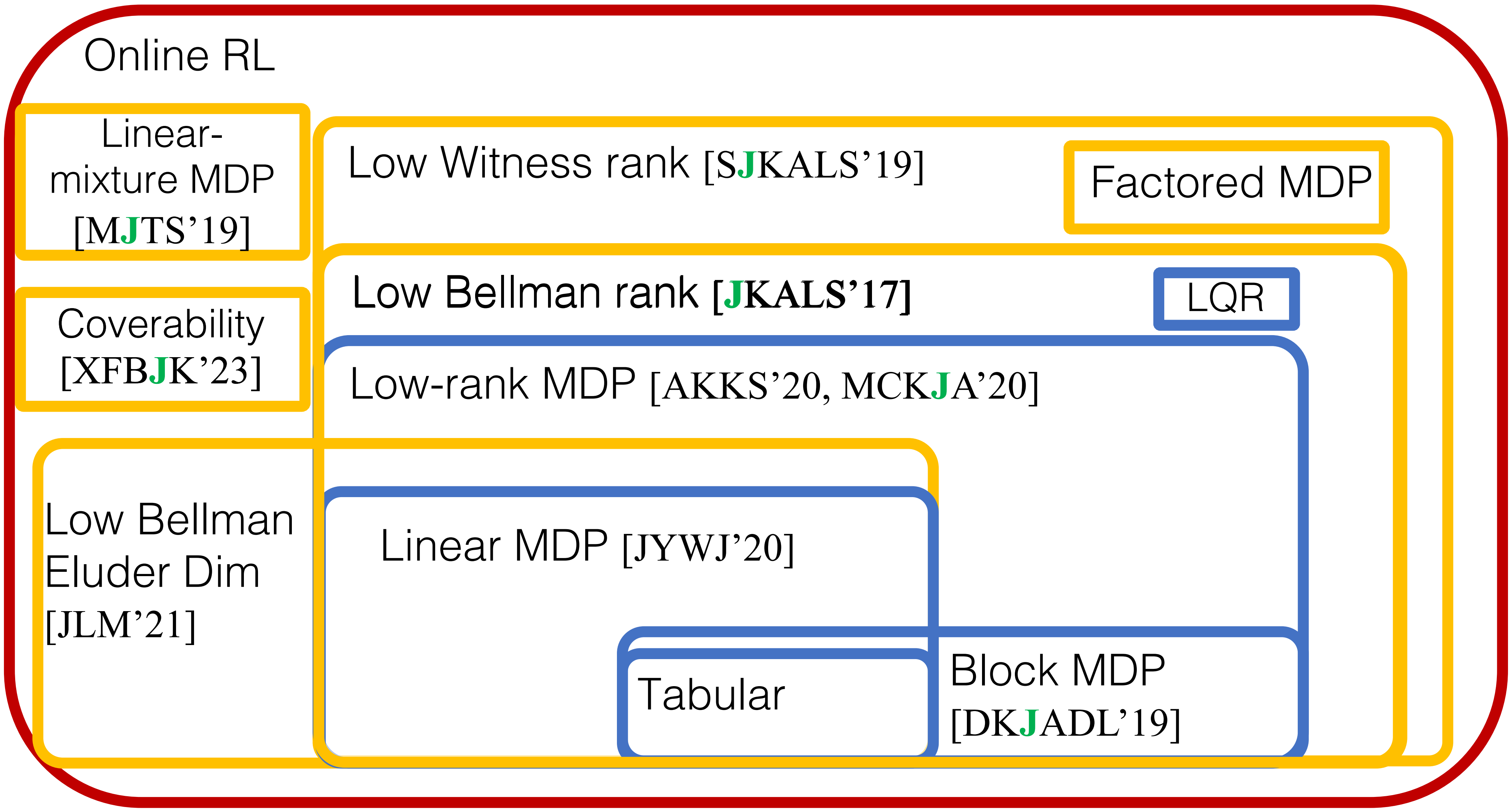


- Adapted from FOCS'20 Tutorial by Agarwal, Krishnamurthy, and Langford
- Also related: bilinear classes [DKLLMSW'21], DEC [FKQR'21]
- Bellman-eluder [JLM'21] generalizes deterministic version of [RvR'13]

- All consider modeling value functions

Discriminative learning

Generative learning?

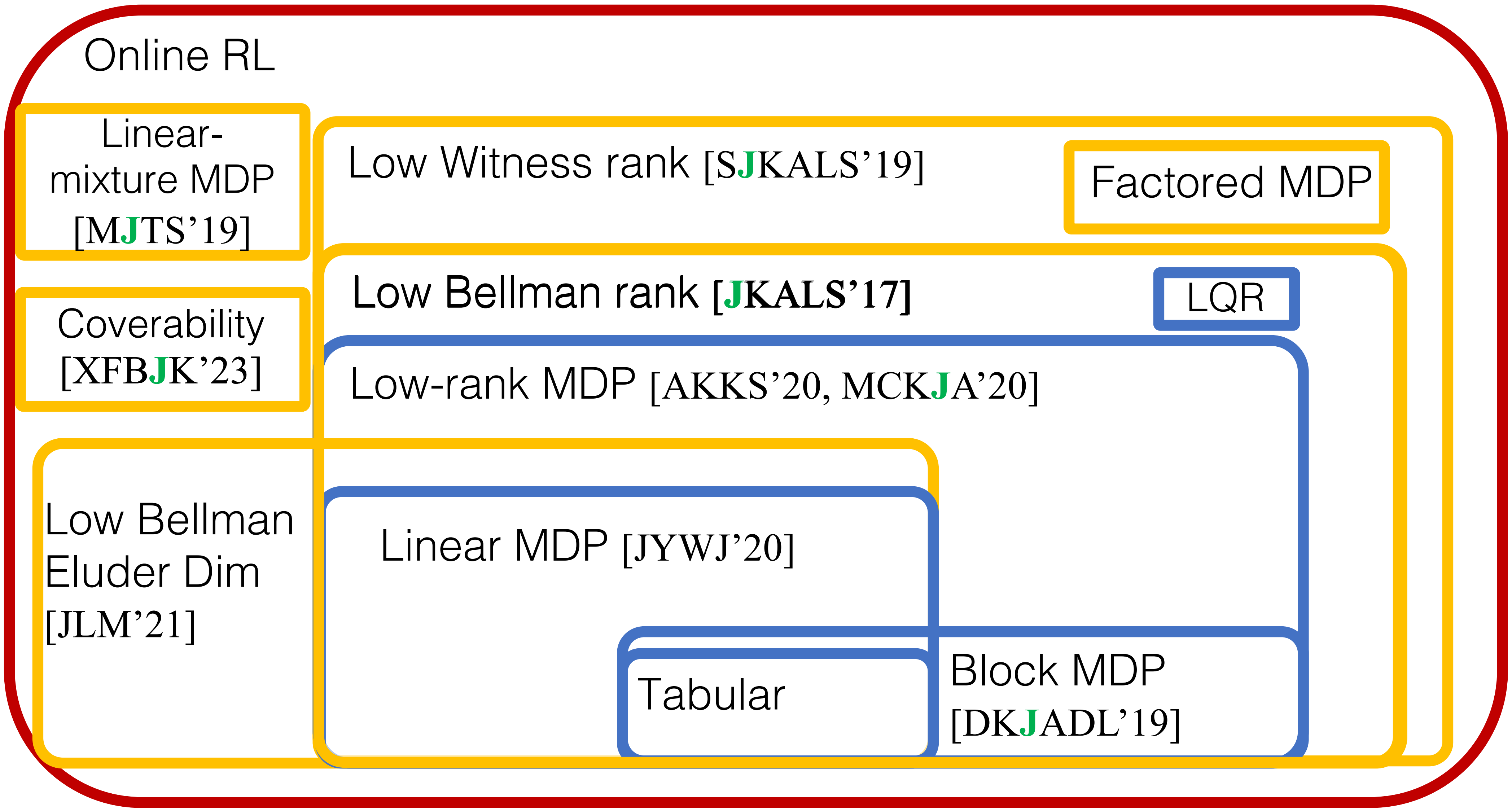


- Adapted from FOCS'20 Tutorial by Agarwal, Krishnamurthy, and Langford
- Also related: bilinear classes [DKLLMSW'21], DEC [FKQR'21]
- Bellman-eluder [JLM'21] generalizes deterministic version of [RvR'13]

- All consider modeling value functions
  - Learning occupancies with density estimation [HCJ'23]

Discriminative learning

Generative learning?



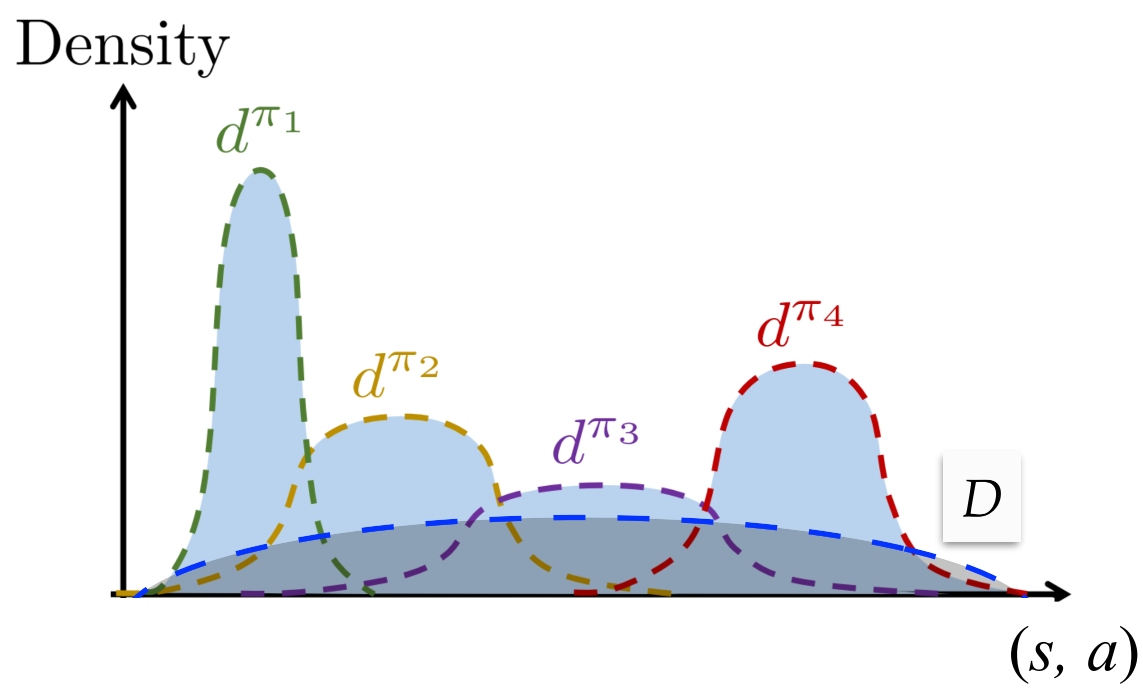
- Adapted from FOCS'20 Tutorial by Agarwal, Krishnamurthy, and Langford
- Also related: bilinear classes [DKLLMSW'21], DEC [FKQR'21]
- Bellman-eluder [JLM'21] generalizes deterministic version of [RvR'13]

- All consider modeling value functions

Discriminative learning

  - Learning occupancies with density estimation [HCJ'23]

Generative learning?



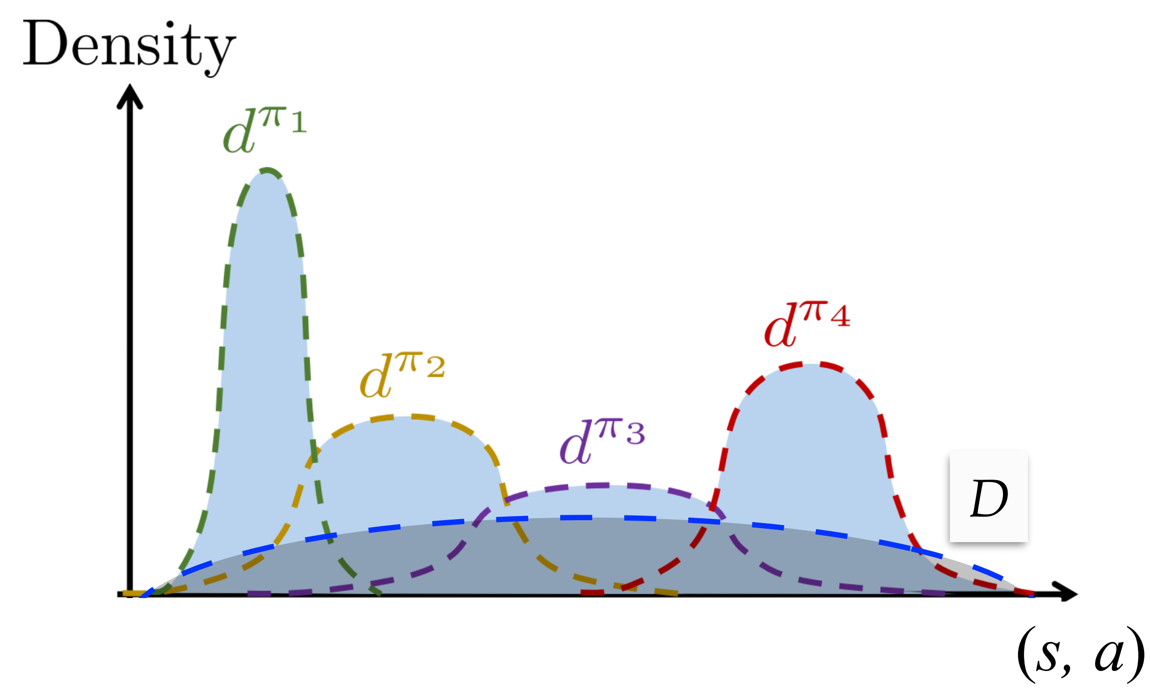
Coverability [XFBJK'23]

- Online vs. Offline: Unification

- All consider modeling value functions
  - Learning occupancies with density estimation [HCJ'23]

Discriminative learning

Generative learning?



Coverability [XFBJK'23]

- Online vs. Offline: Unification
  - Not all MDPs admit such D
  - Those who do can be explored efficiently

Longterm directions

- RL (theory) so far: mostly single-agent & Markovian

Longterm directions

- RL (theory) so far: mostly single-agent & Markovian
- Significant challenges in real-world systems

AI EDGE Institute

Longterm directions

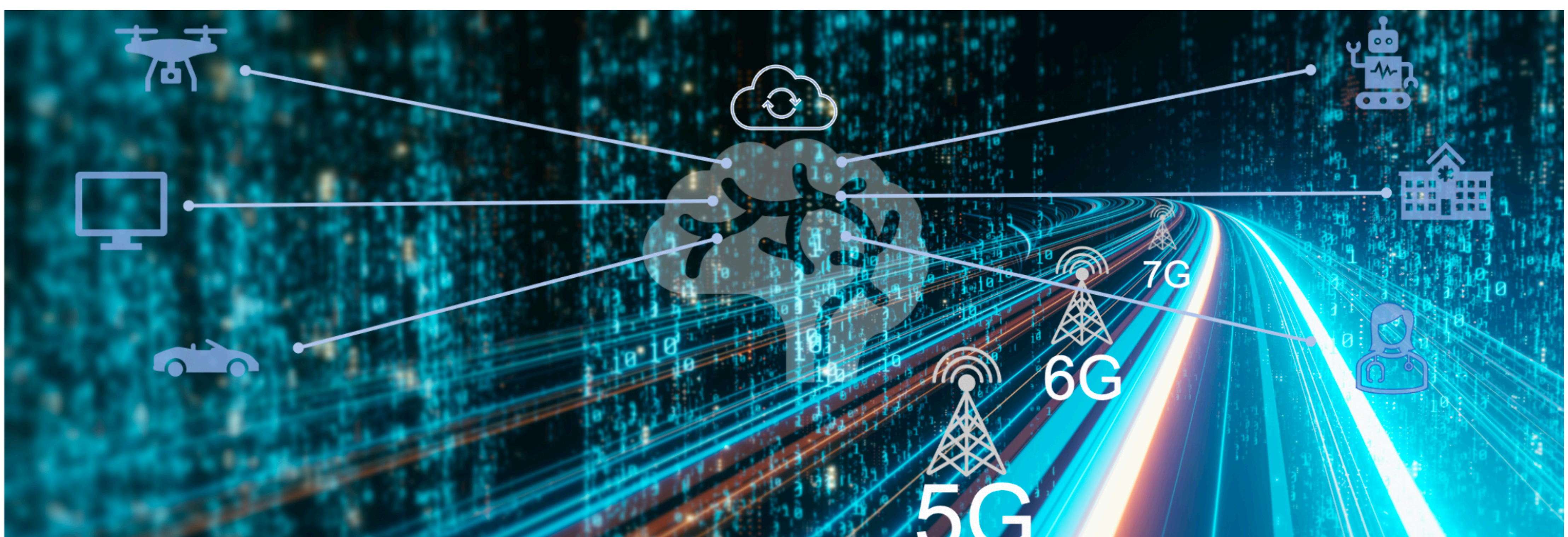
- RL (theory) so far: mostly single-agent & Markovian
- Significant challenges in real-world systems
  - Multi-agent (possibly w/ strategic interactions)

AI EDGE Institute

Longterm directions

- RL (theory) so far: mostly single-agent & Markovian
- Significant challenges in real-world systems
  - Multi-agent (possibly w/ strategic interactions)
  - Partial observability

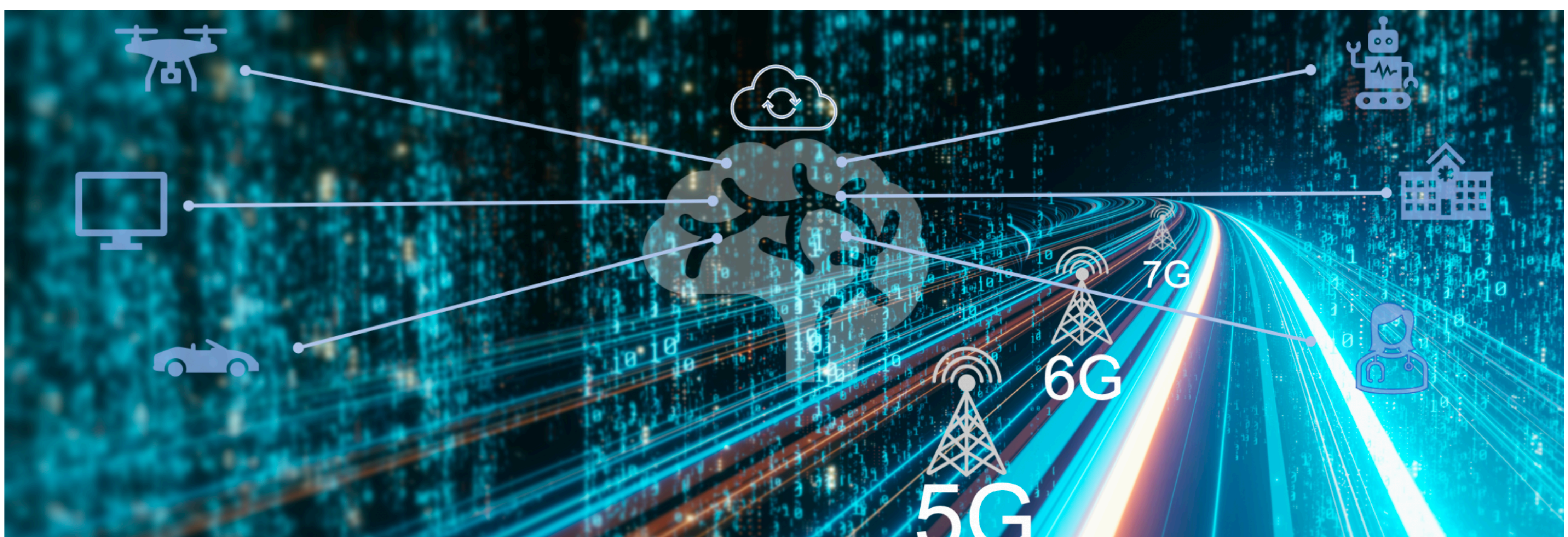
AI EDGE Institute



Longterm directions

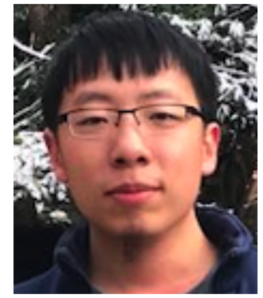
- RL (theory) so far: mostly single-agent & Markovian
- Significant challenges in real-world systems
  - Multi-agent (possibly w/ strategic interactions) [ZBJ’23]
  - Partial observability [KJS’15a’15b, JKS’16’18] [UKBCJKSS’23, ZJ’24]

AI EDGE Institute



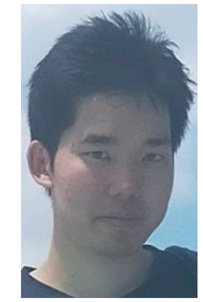
PhD Alumni

Tengyang Xie (Assistant Professor, Wisconsin-Madison, 24)



Jingling Chen (Research Engineer, TikTok)

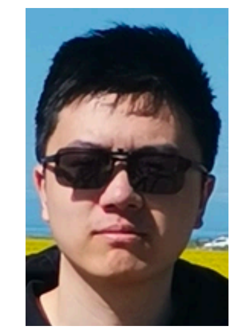
Alumini (visiting student & MS)



Masatoshi Uehara (Assistant Professor, Wisconsin-Madison, 25)

Siyuan Zhang

Priyank Agarwal (Columbia PhD)



Jiawei Huang (ETH Zurich PhD)

Current Students

Philip Amortila

Audrey Huang

Yuheng Zhang

Wei Xiong

Collaborators (Reverse Chronological): Dylan Foster, Ching-An Cheng, Akshay Krishnamurthy, Yu Bai, Alekh Agarwal, Sham Kakade, Chengchun Shi, Wen Sun, Aditya Modi, Paul Mineiro, John Langford, Jason Lee, Cameron Voloshin, Hoang M. Le, Yisong Yue, Gellert Weisz, Csaba Szepesvári, Nathan Kallus, Yu-Xiang Wang, Simon Du, Miroslav Dudík, Christoph Dann, Robert Schapire, Alex Kulesza, Satinder Singh, Ambuj Tewari, Hal Daumé III

Thank you! Questions?

NSF CAREER & AI-Edge

Adobe Data Science Award

DEVCOM ARMY RESEARCH LABORATORY IoBT