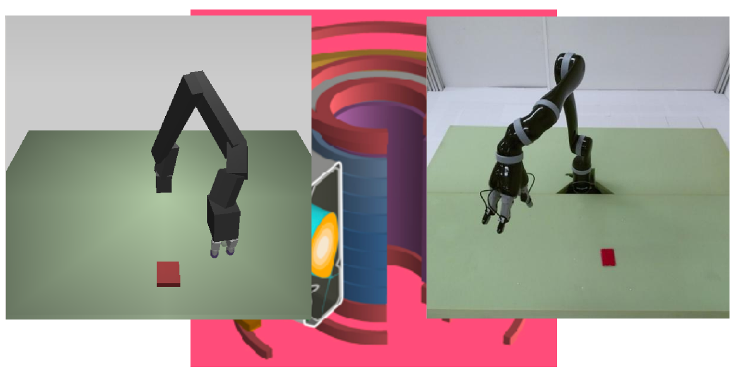
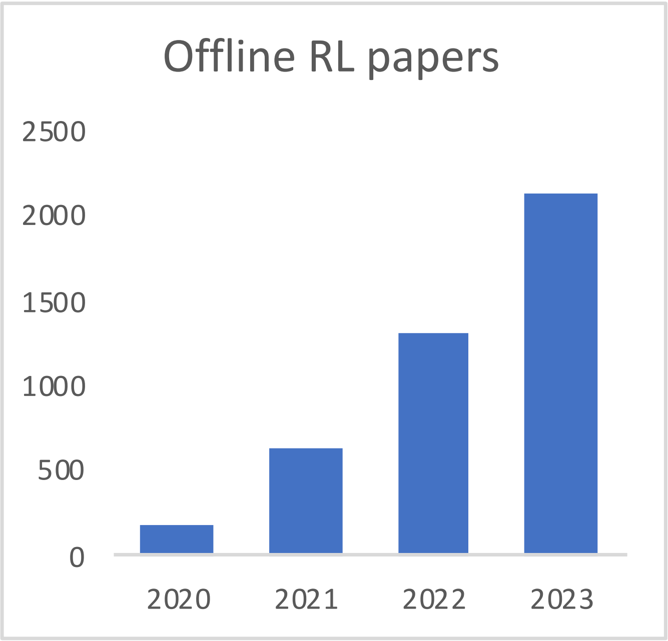
# Rethinking the theoretical foundation of reinforcement learning

Nan Jiang
University of Illinois at Urbana-Champaign
Nov 19, 2024
@MIT

- (Offline) RL in real life




Key ingredient: **simulator**

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

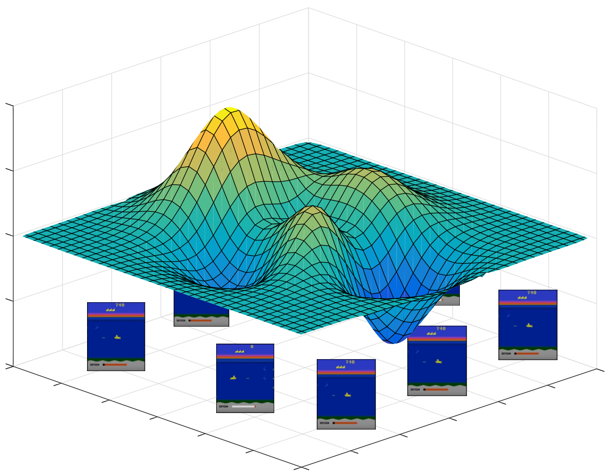
Why are we not seeing (offline) RL deployed everywhere already?

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

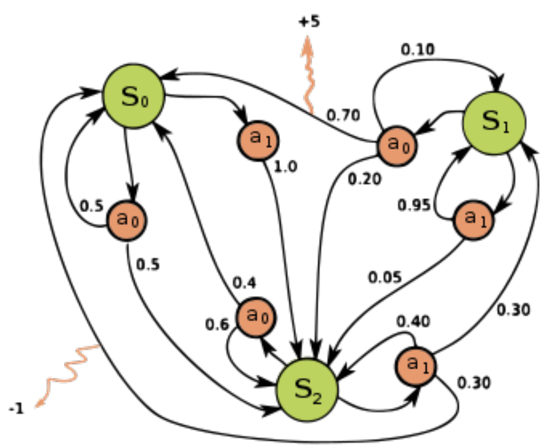
~2000

*

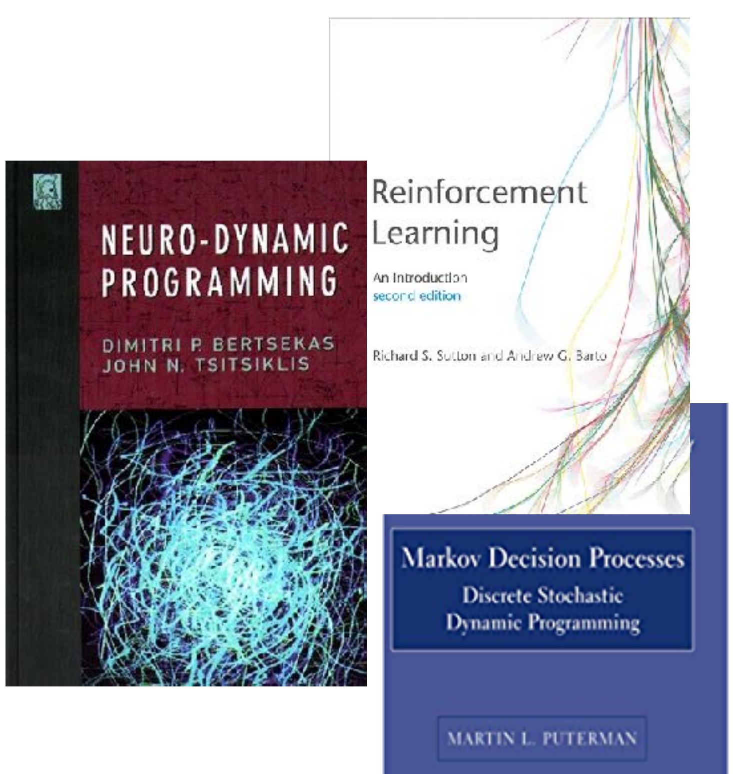
~2015

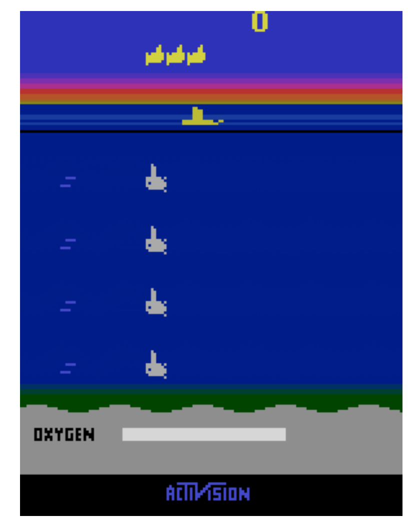


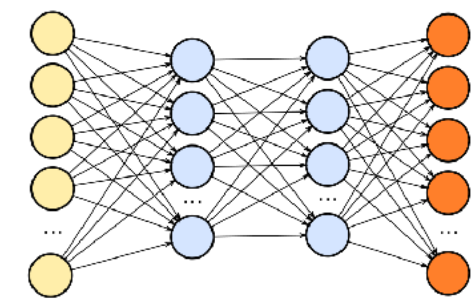
*Bellman rank, Eluder dimension, Concentrability, …*




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



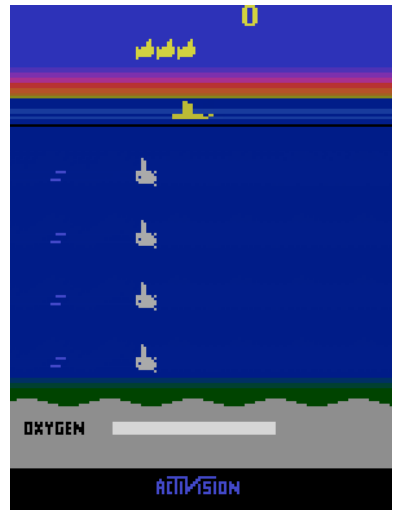



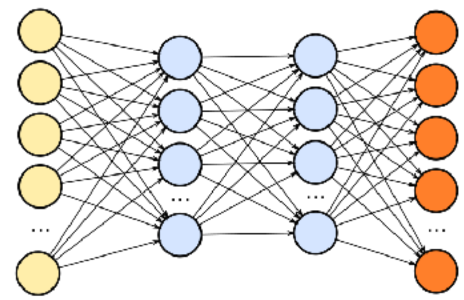



***Empirical:*** *Atari, Mujoco, OpenAI Gym, target network, architecture, …*

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

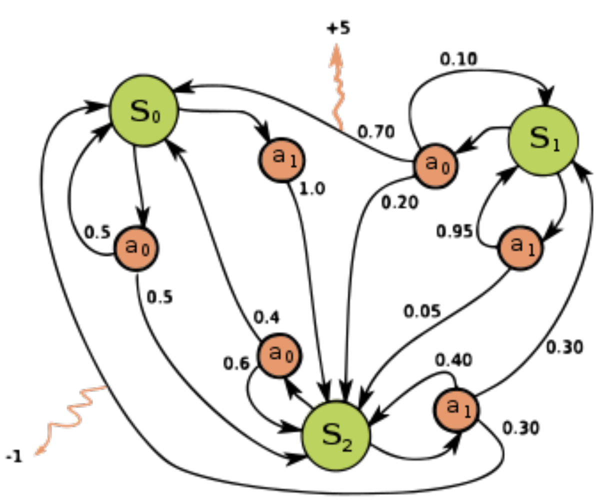
- (Offline) RL in real life
- Role of theory in modern RL
- Theoretical foundation



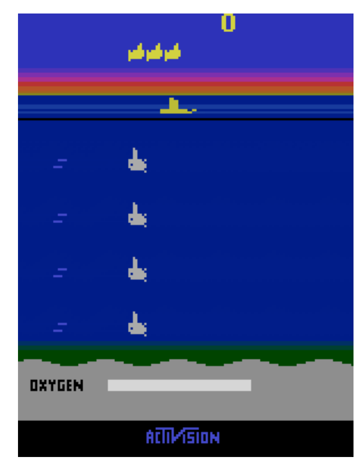
simplify

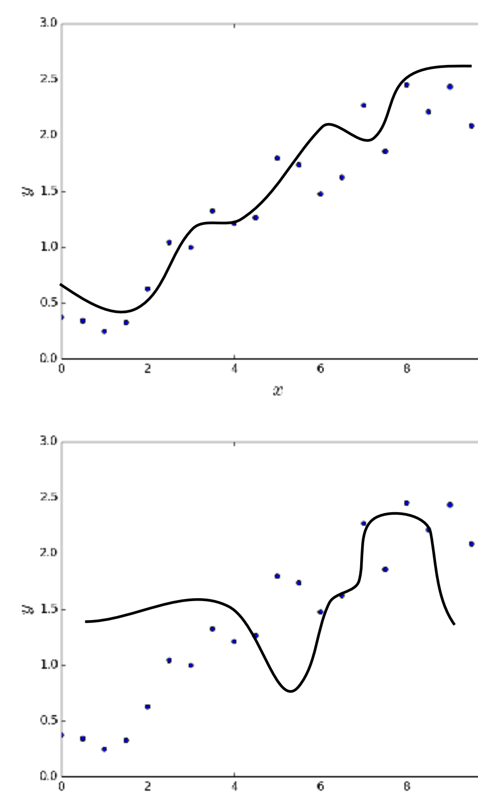
extend




| | $Q^*(s, a)$ |
|---|---|
| $(s_1, a_1)$ | … |
| $(s_1, a_2)$ | … |
| $(s_2, a_1)$ | … |

"tabular" RL



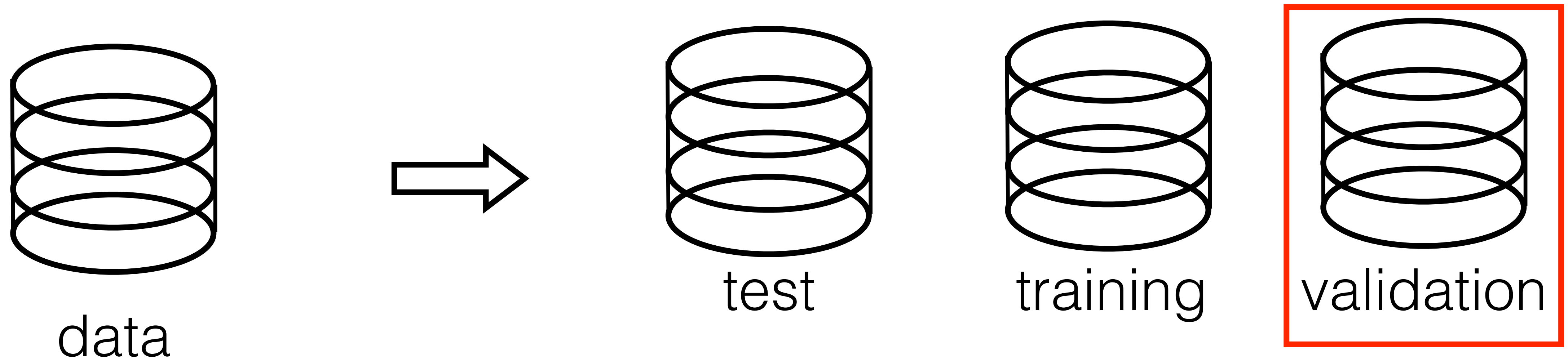



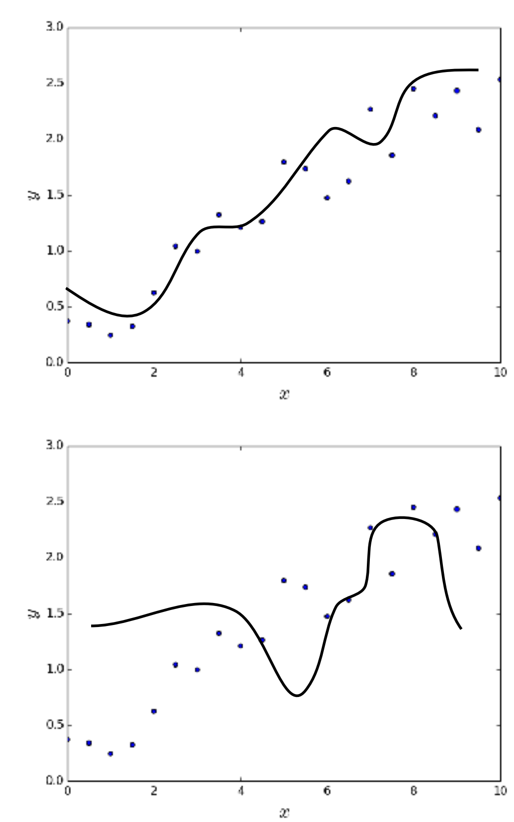
- (Offline) RL in real life
- Role of theory in modern RL
- Theoretical foundation

modern RL ⟹ real life

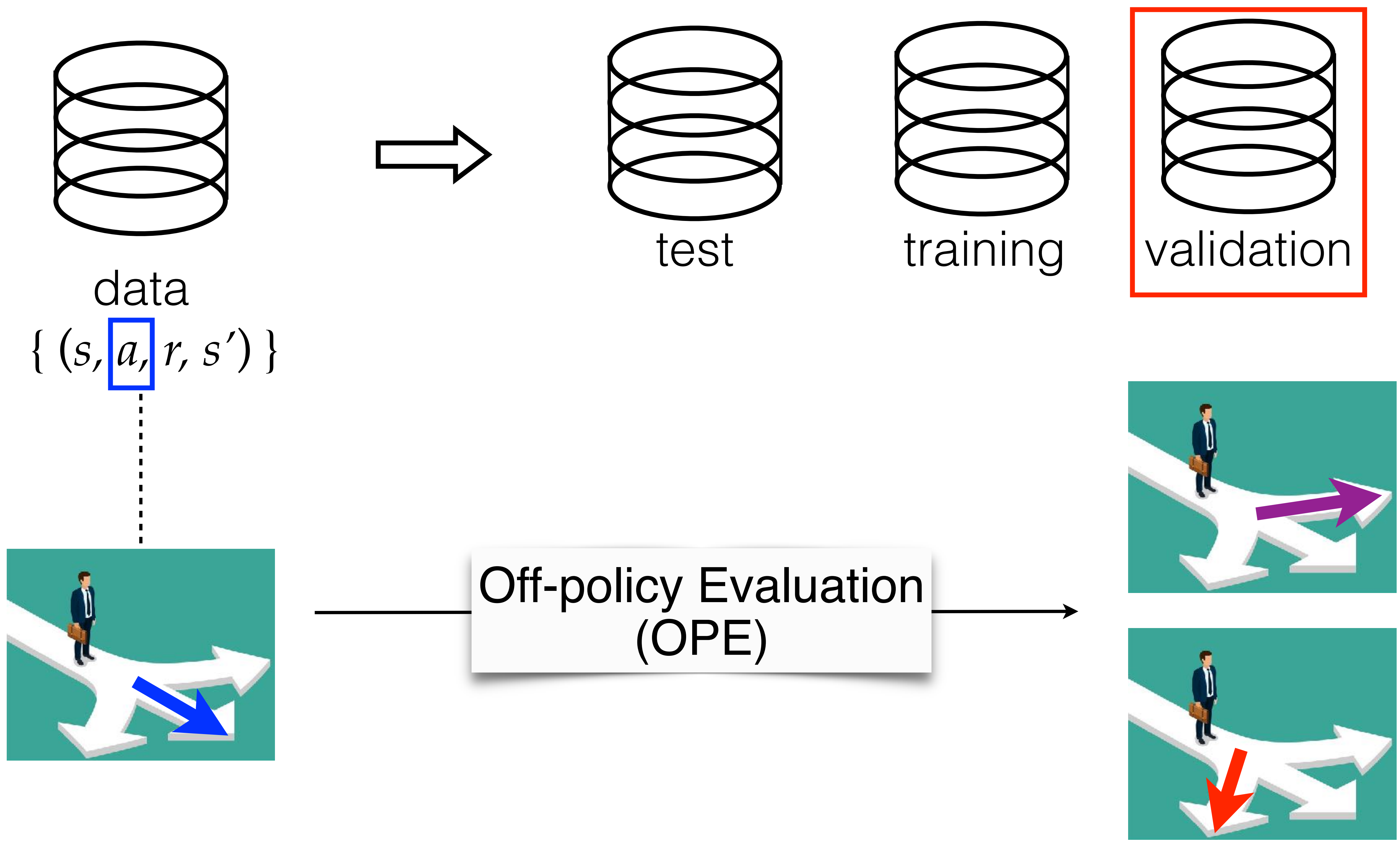
Rethinking
theoretical foundation

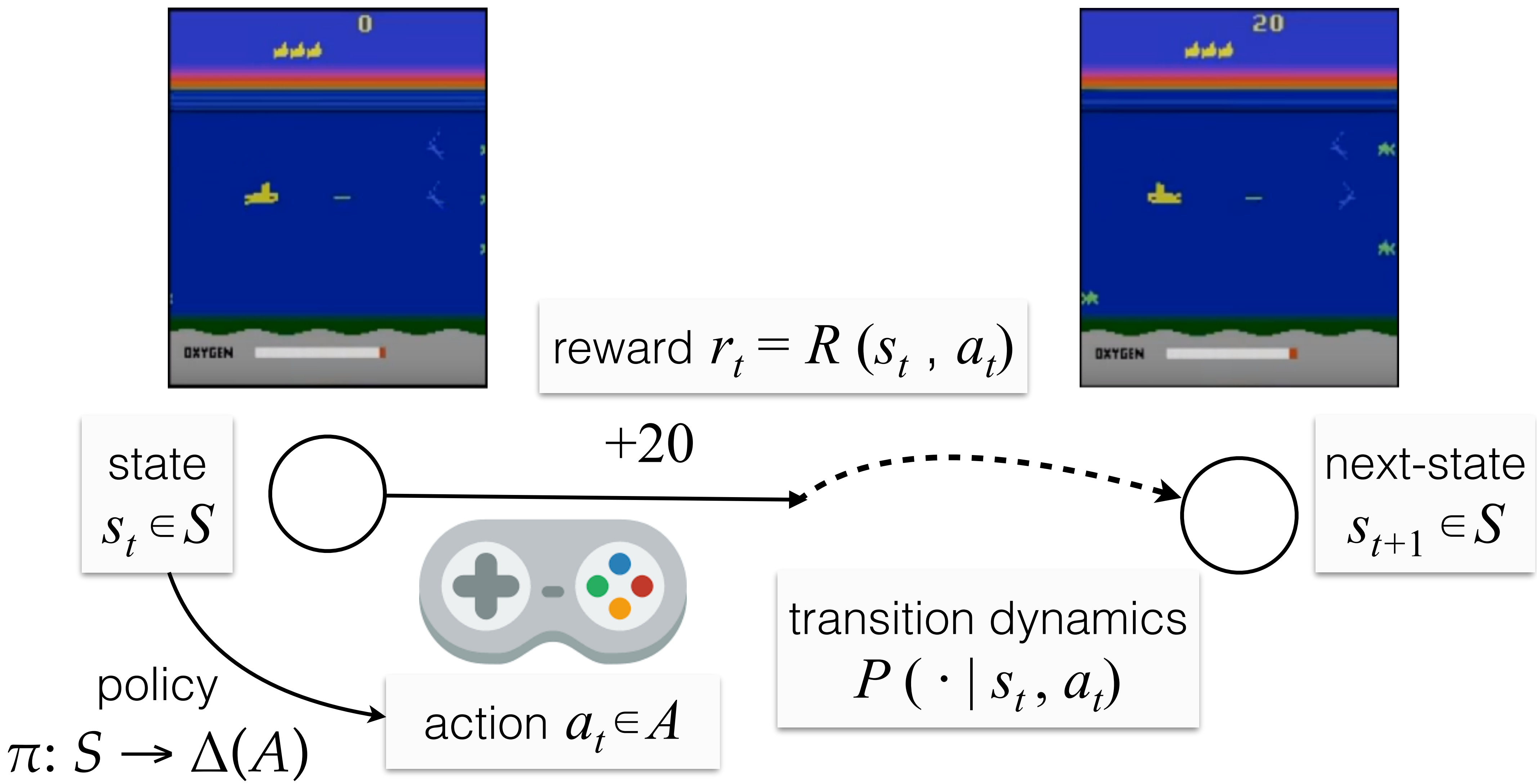
# Supervised learning pipeline



# Offline RL pipeline



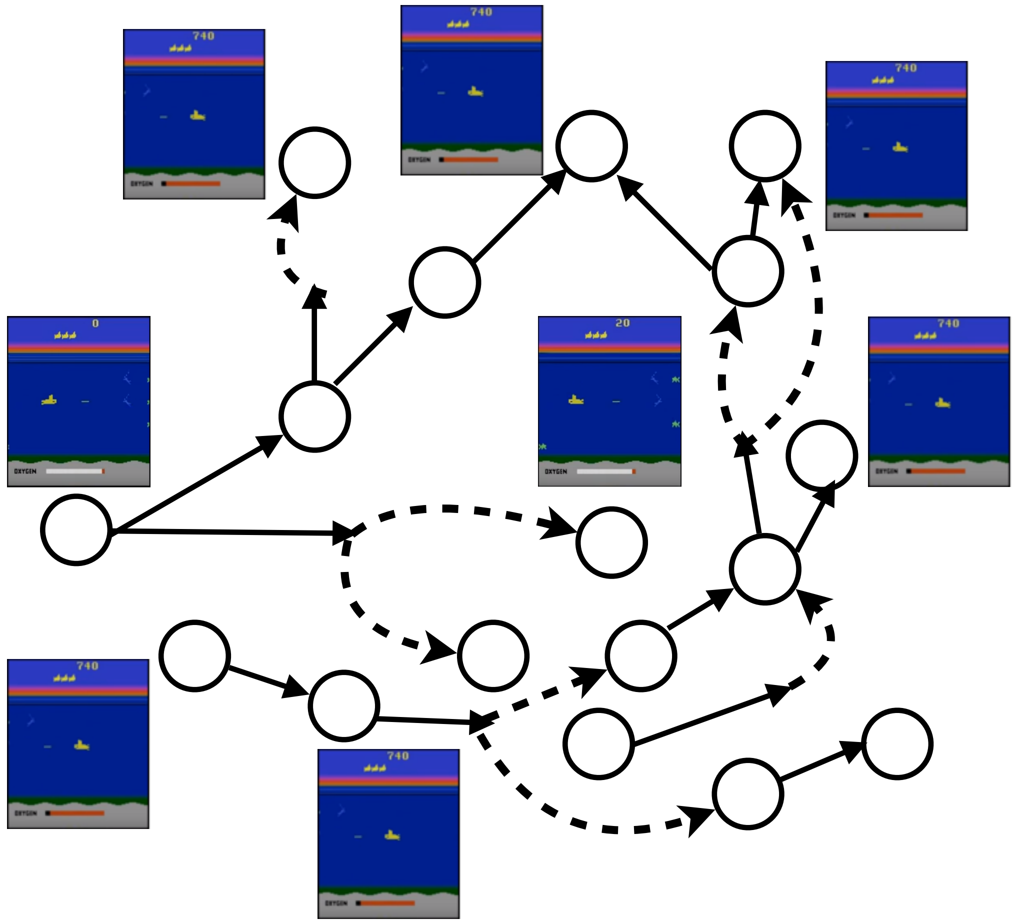



Policy evaluation: estimate $J(\pi) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0]$ given $\pi$

Policy optimization: $\max_\pi J(\pi) \qquad = Q^\pi(s_0, \pi)$

How to find $Q^\pi$?

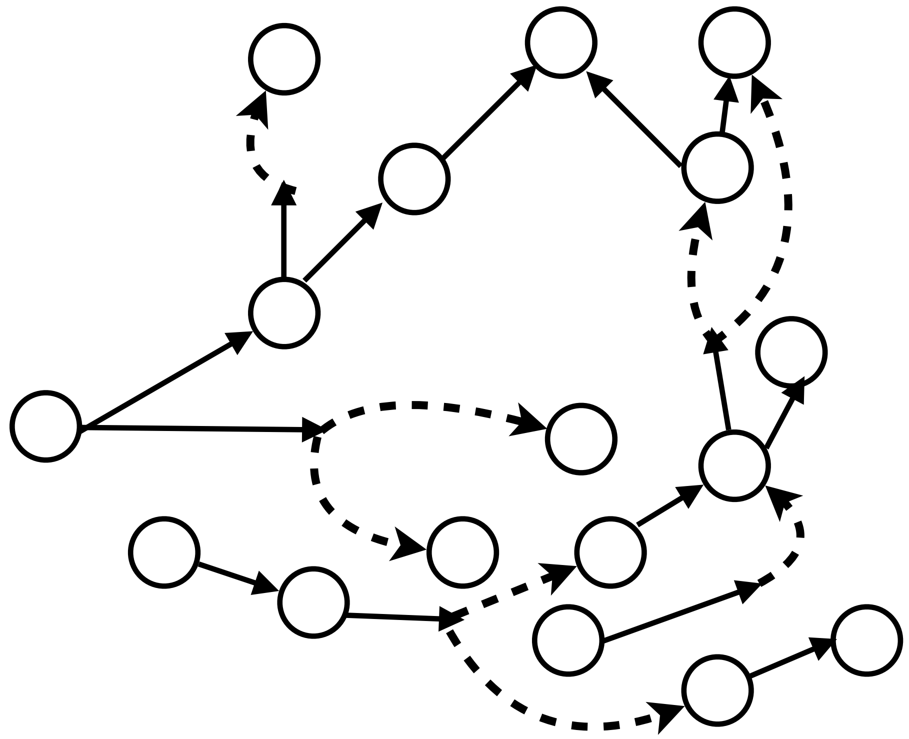
$$Q^\pi = \mathcal{T}^\pi Q^\pi \quad \rightarrow \quad |SxA| \text{ equations}$$

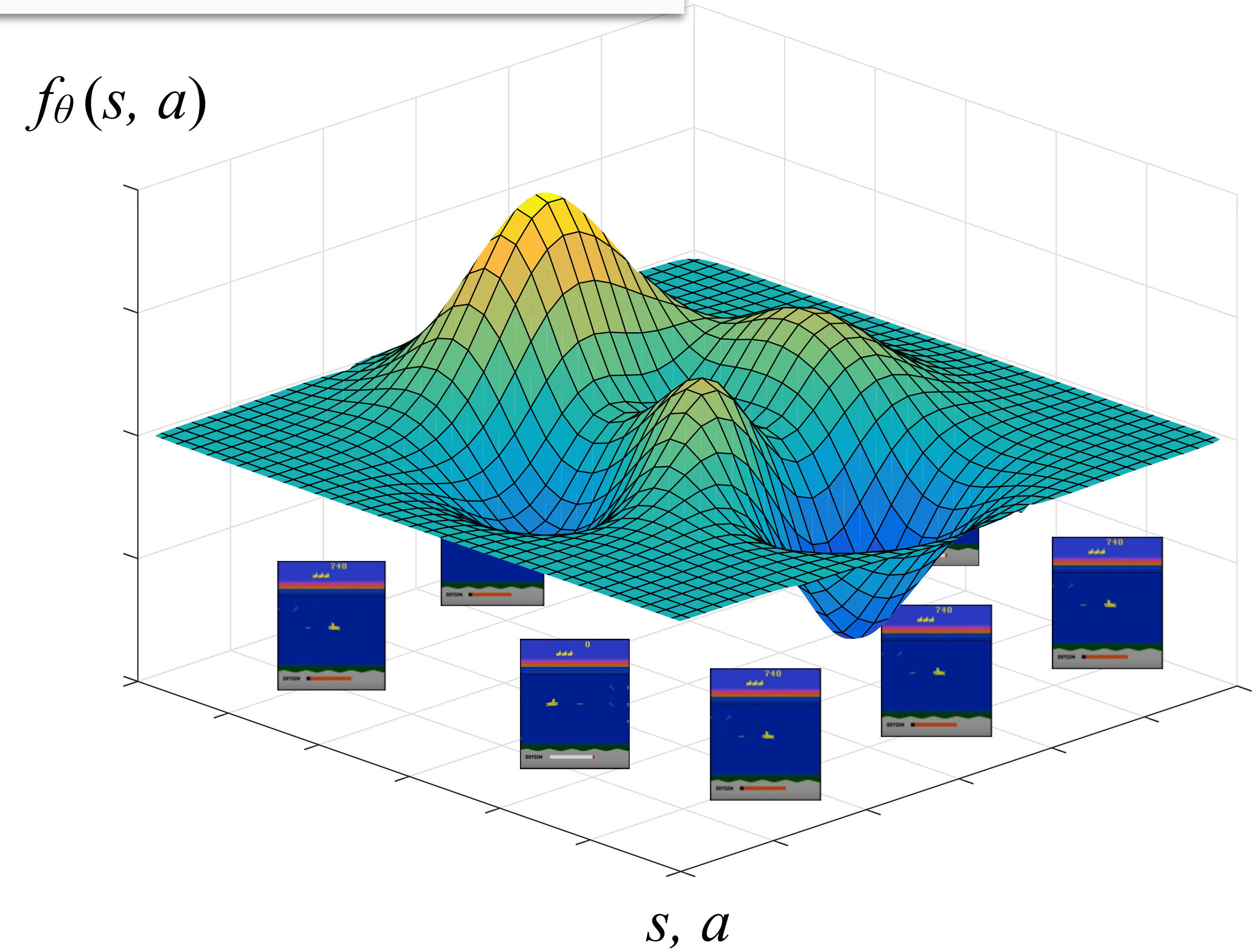


How to find $Q^\pi$?

$$Q^\pi = \mathcal{T}^\pi Q^\pi \;\rightarrow\;$$ $|SxA|$ equations **X**

Find $\theta$ s.t. $f_\theta \approx Q^\pi$






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

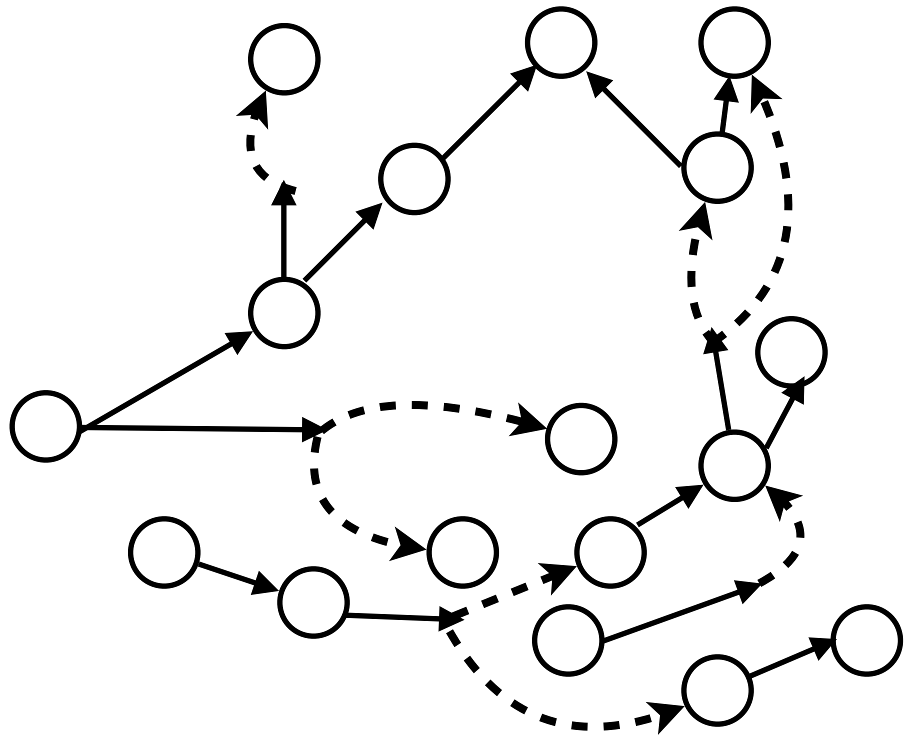
$(s, a, r, s') \sim D$

**Validation**:

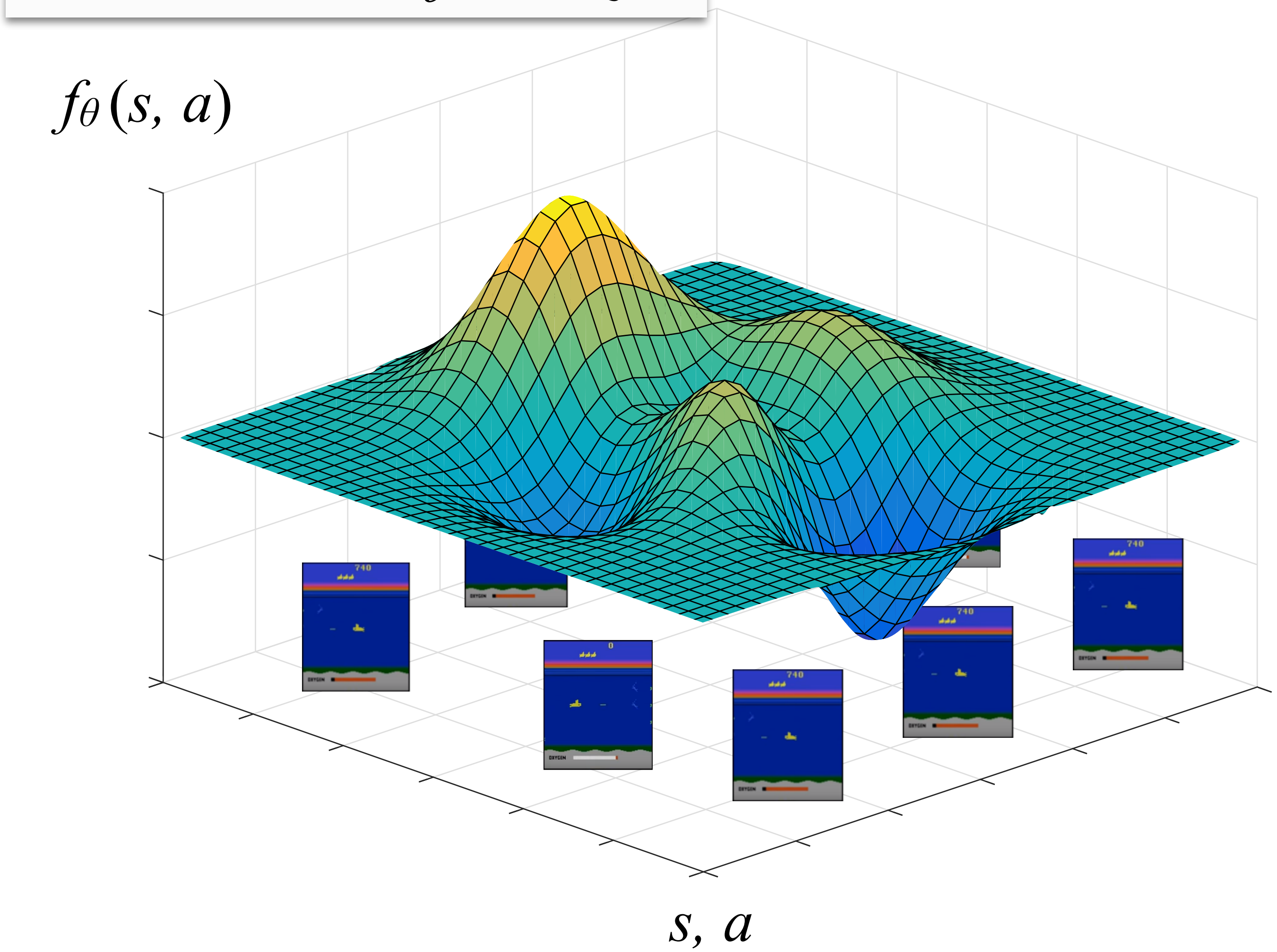
(FQE: learn $Q^\pi$)

iterative

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



Find $\theta$ s.t. $f_\theta \approx Q^\pi$



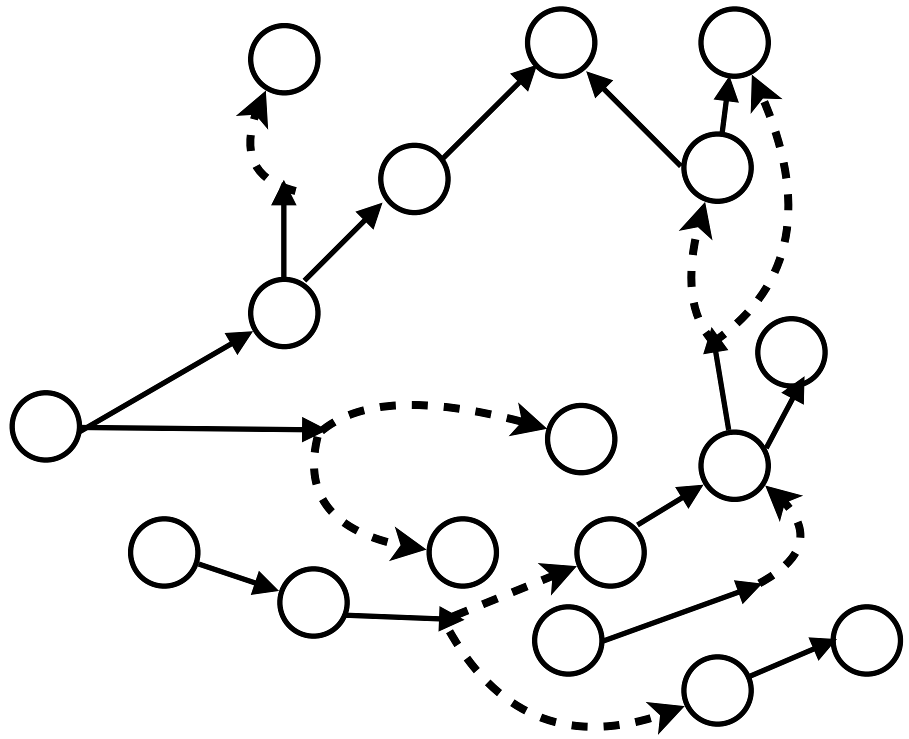

**Validation**:

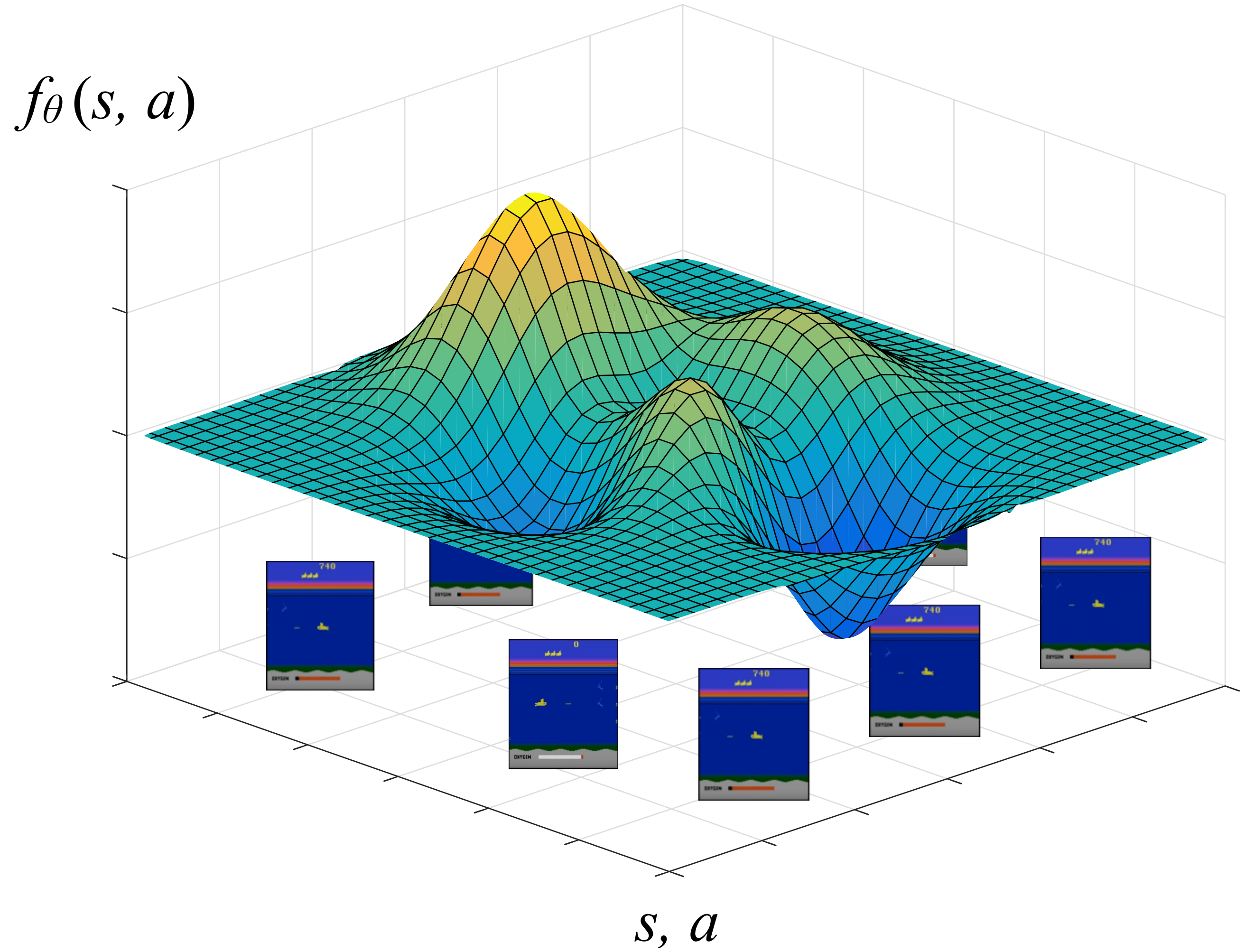
(FQE: learn $Q^\pi$)

iterative

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

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






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

(FQI: learn $Q^*$)

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

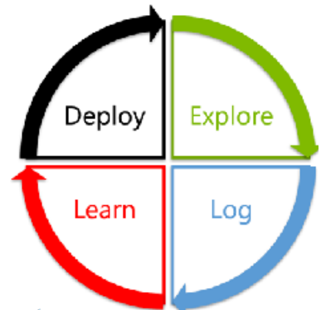
$$\Downarrow \quad \pi = \text{greedy w.r.t. } \hat{f}$$

Validation:

(FQE: learn $Q^\pi$)

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

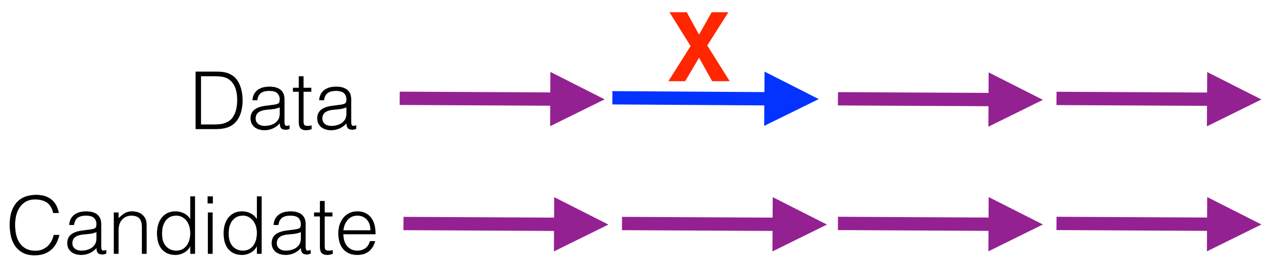
# Hyperparameter-free methods?



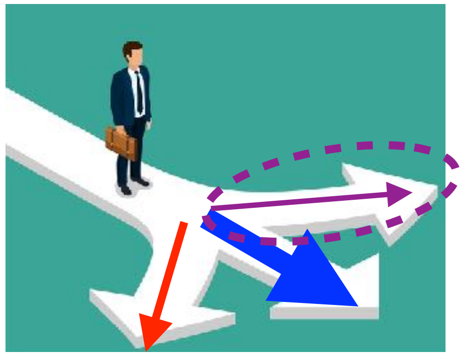

Azure


## Importance sampling [Precup'00]

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




Data

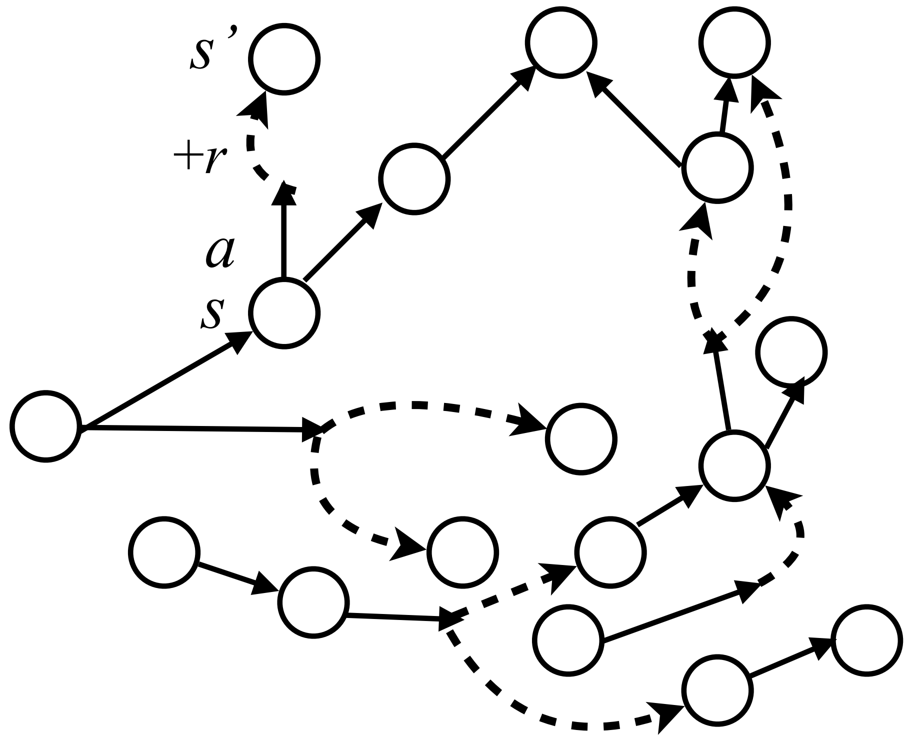


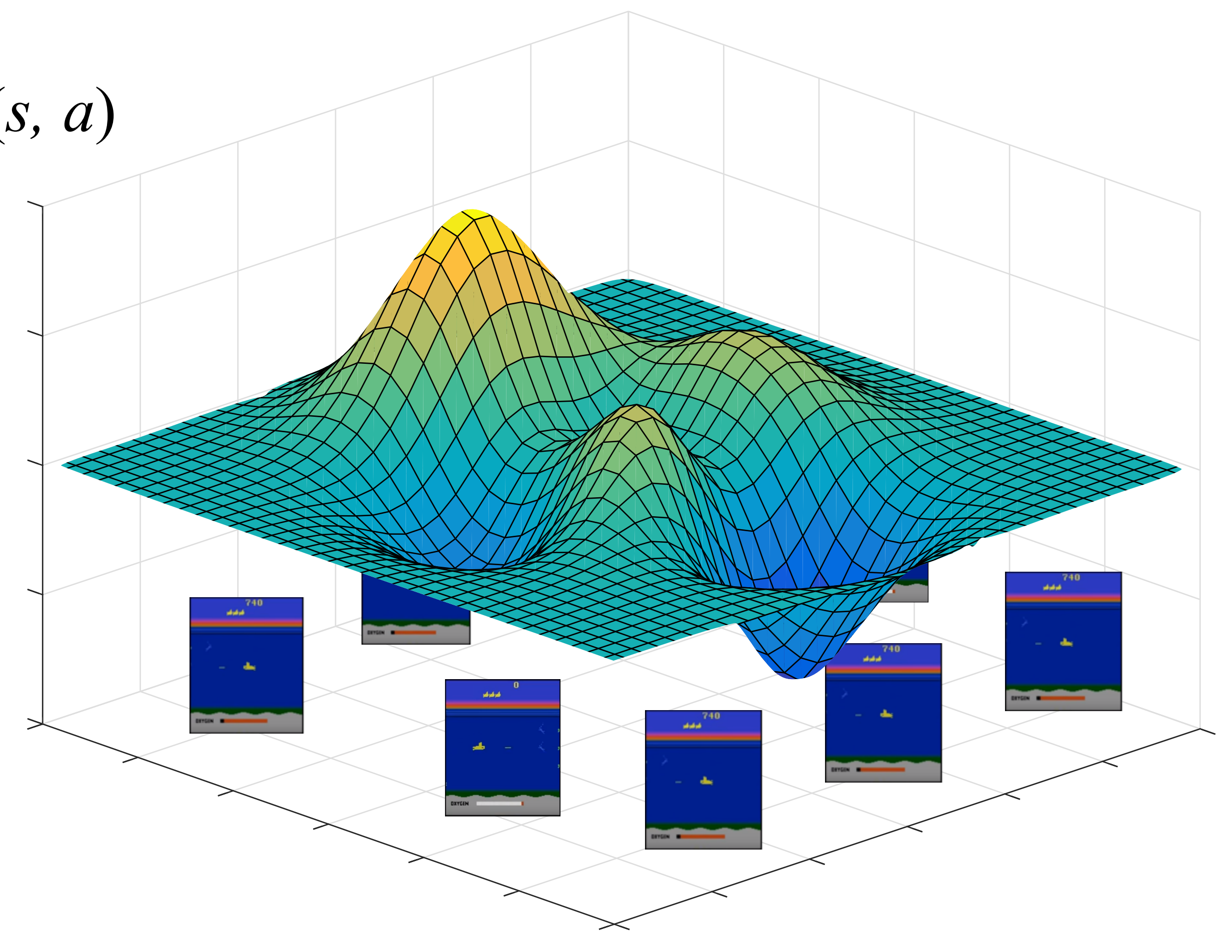
Candidate

## 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 $Q^*$)

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

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

# Reformulation: Value-function Selection

## Simple(?) Problem

- Run different training algorithms
- Get candidate value functions $f_1$, $f_2$, …
- Holdout data $\{(s, a, r, s')\}$
- 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. $Q^\star \overset{?}{=} \arg\min_{f\in\mathcal{F}} L(f)$ $\quad \mathcal{F} = \{f_1, f_2, \ldots\}$
- RL doesn't work like that!

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

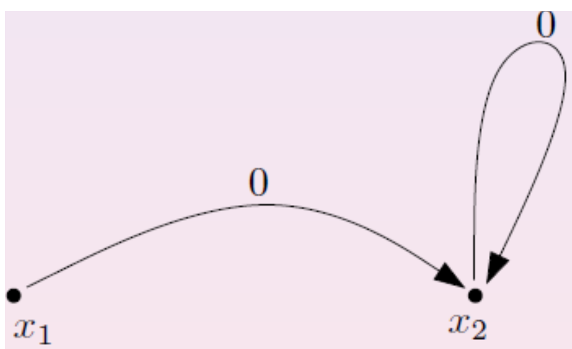
(FQI: learn $Q^*$)

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

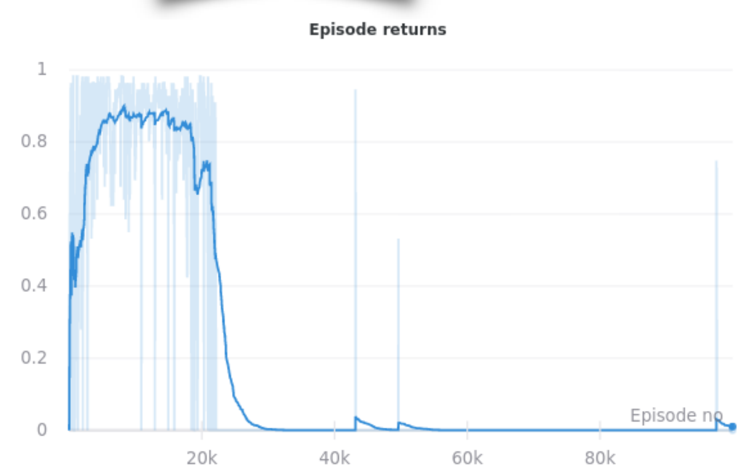
optimization

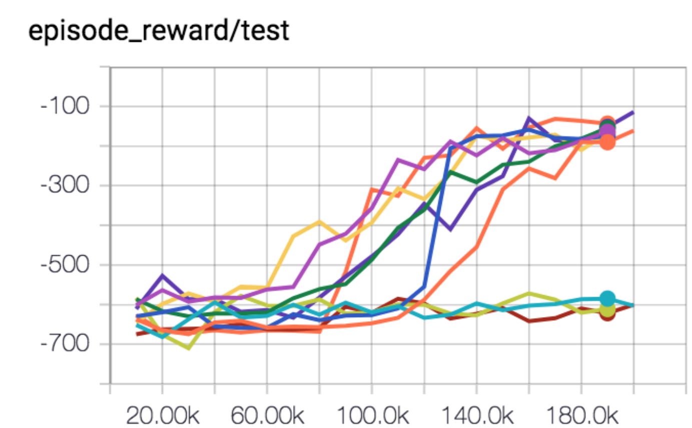
$\approx \mathcal{T} f_{k-1}$

iterative

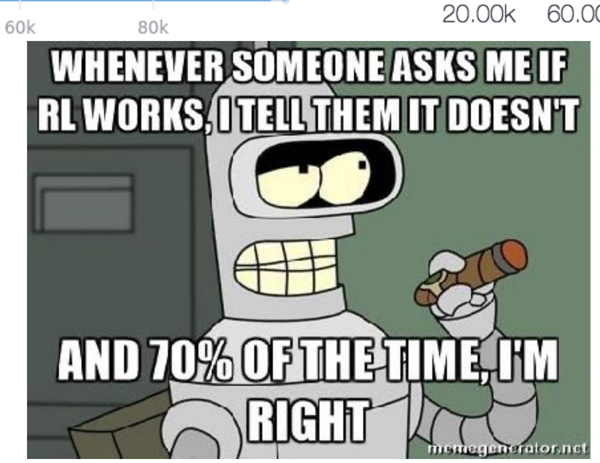



Divergence under 1-d linear
[TvR'96]




"Deadly triad"




 Baird. 1995. Residual algorithms: Reinforcement learning with function approximation
https://www.ale[illegible]

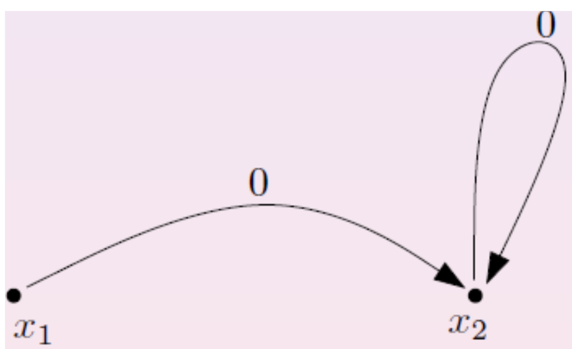
# The training perspective

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

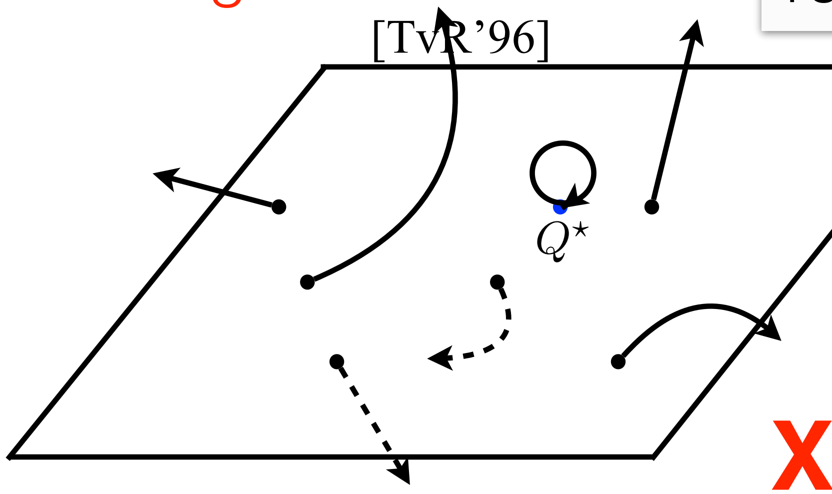
(FQI: learn $Q^*$)

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

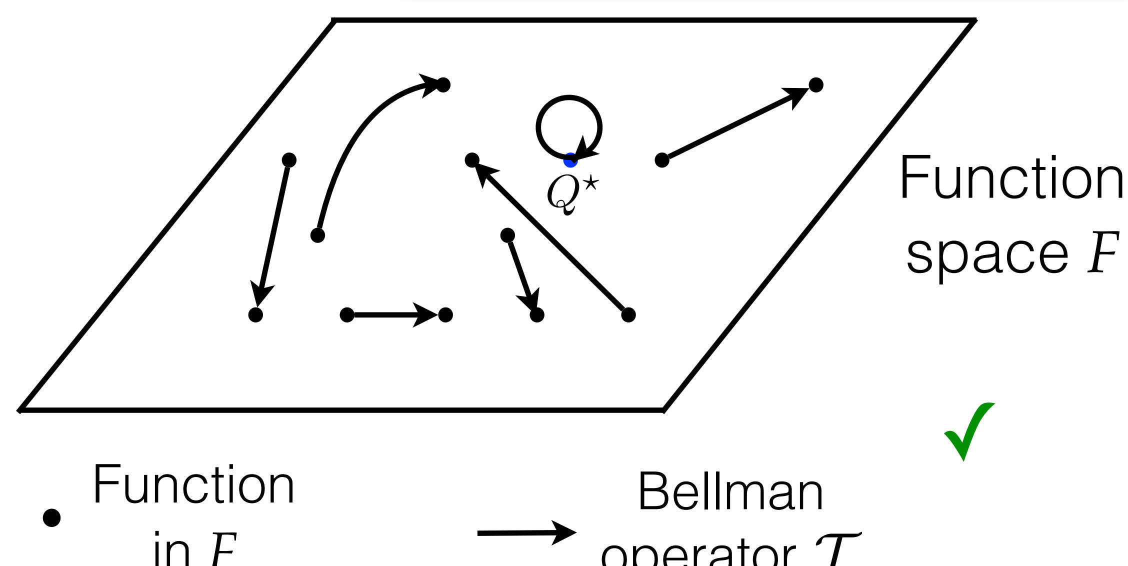



Divergence under 1-d lin realizability (of $Q^*$)

[TvR'96]




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



# The training perspective

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

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

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

$$\neq$$

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

- 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: $\mathbb{E}_{d^D}\left[\mathbb{V}_{s'|s,a}[r + \gamma \max_{a'} f(s',a')]\right]$

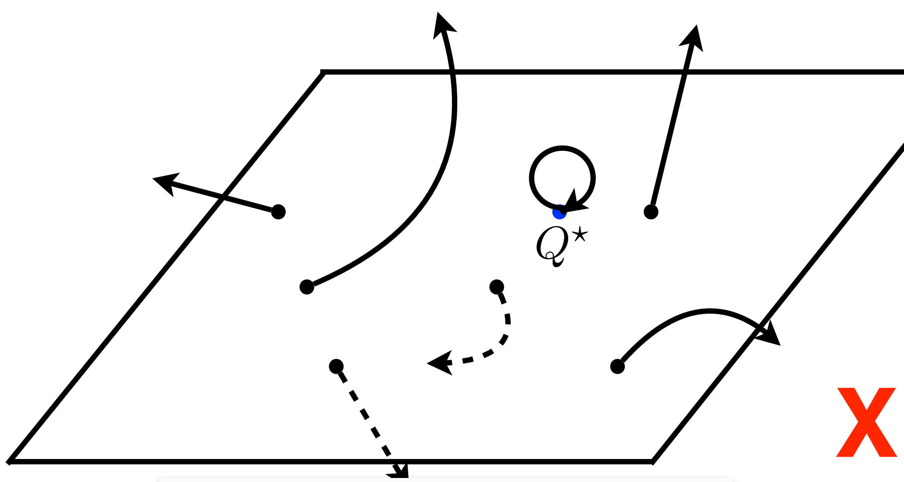
# Basis of resolution

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

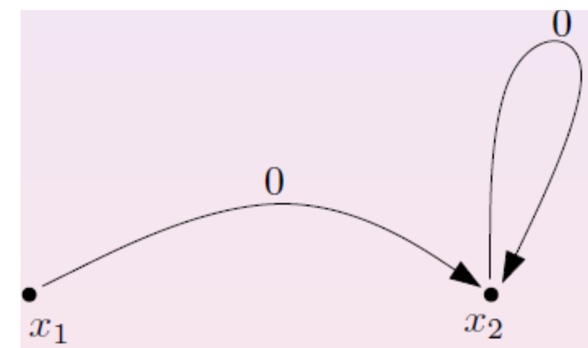
(FQI: learn $Q^*$)

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

iterative




realizability (of $Q^*$)




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.

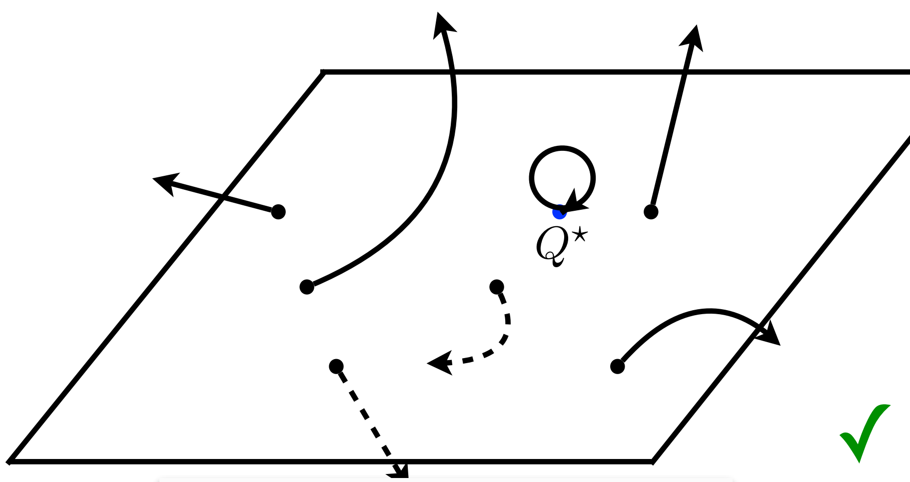
# Basis of resolution

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

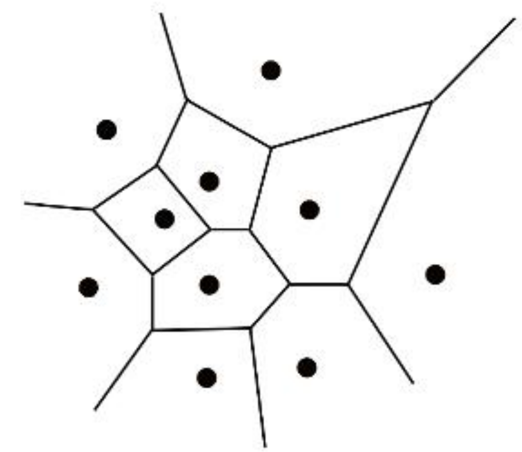
(FQI: learn $Q^*$)

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

iterative

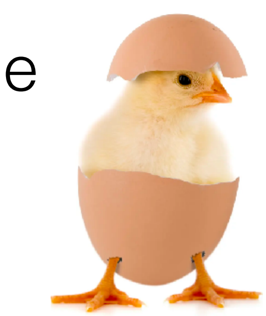



realizability (of $Q^*$) ✓



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

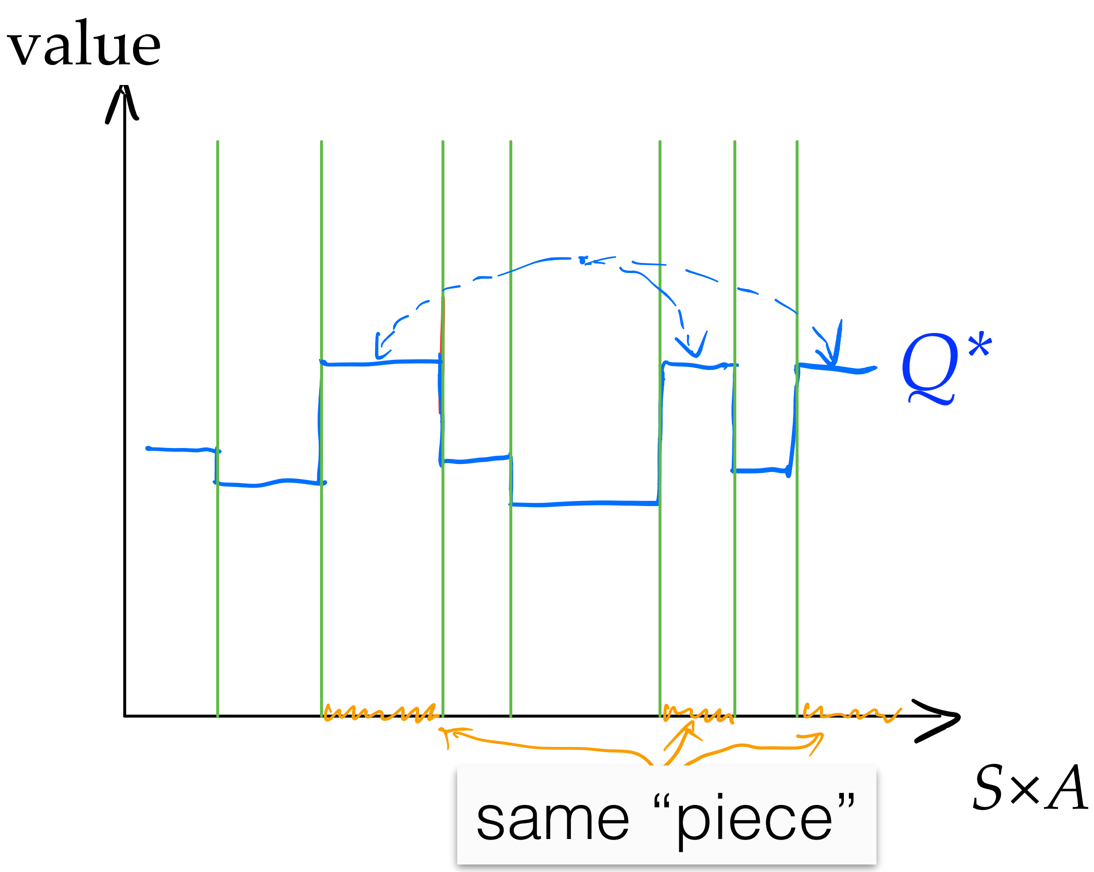
Our method: create such a magical $G$ "out of nothing"!



Gordan. 1995. Stable Function Approximation in Dynamic Programming.

# Does a magical $G$ always exist?

- YES! Just partition $S$x$A$ according to **output** of $Q^*$

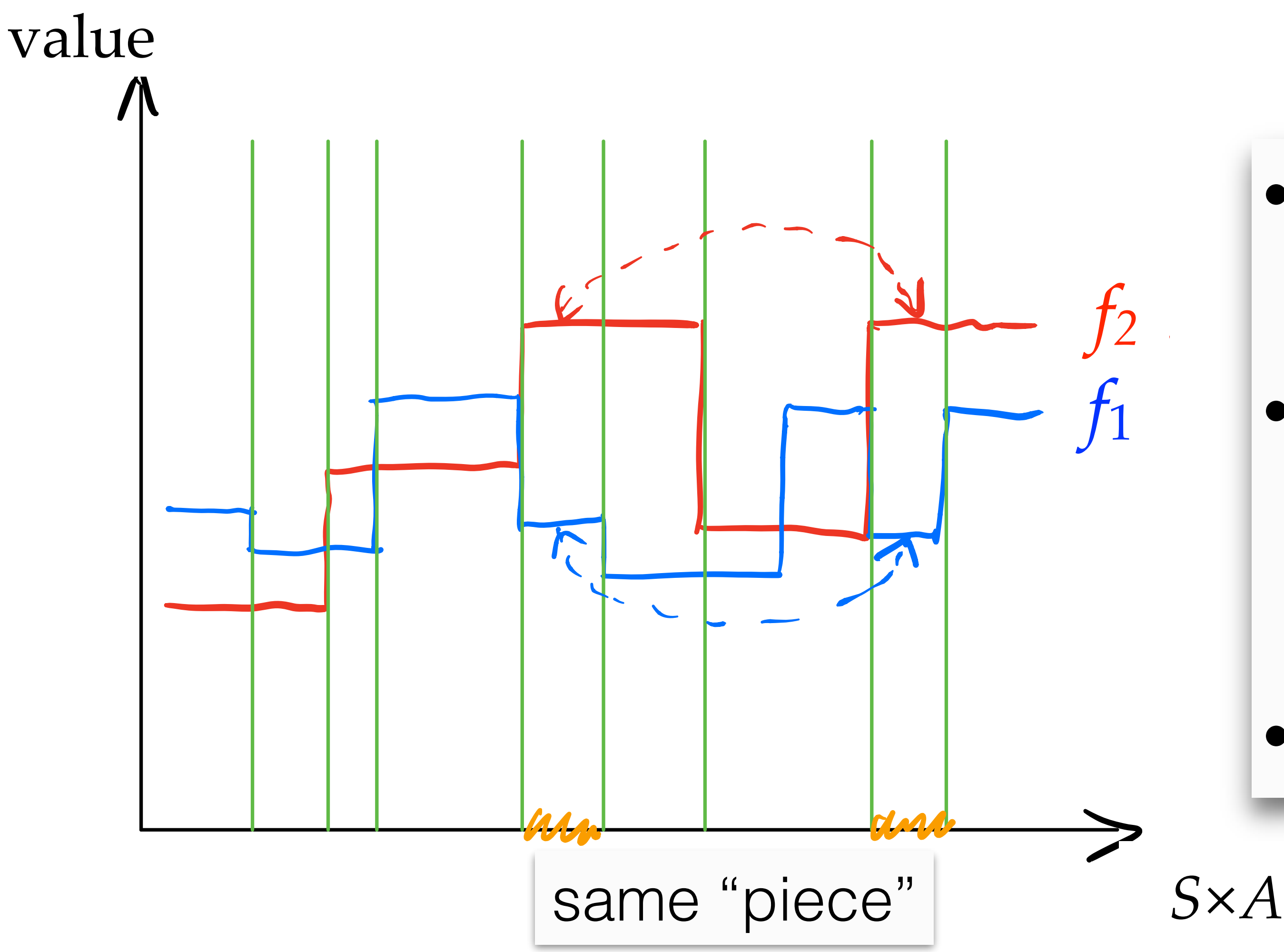



To select b/t $f_1$, $f_2$, suffices to have class $G$ s.t.

- 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]



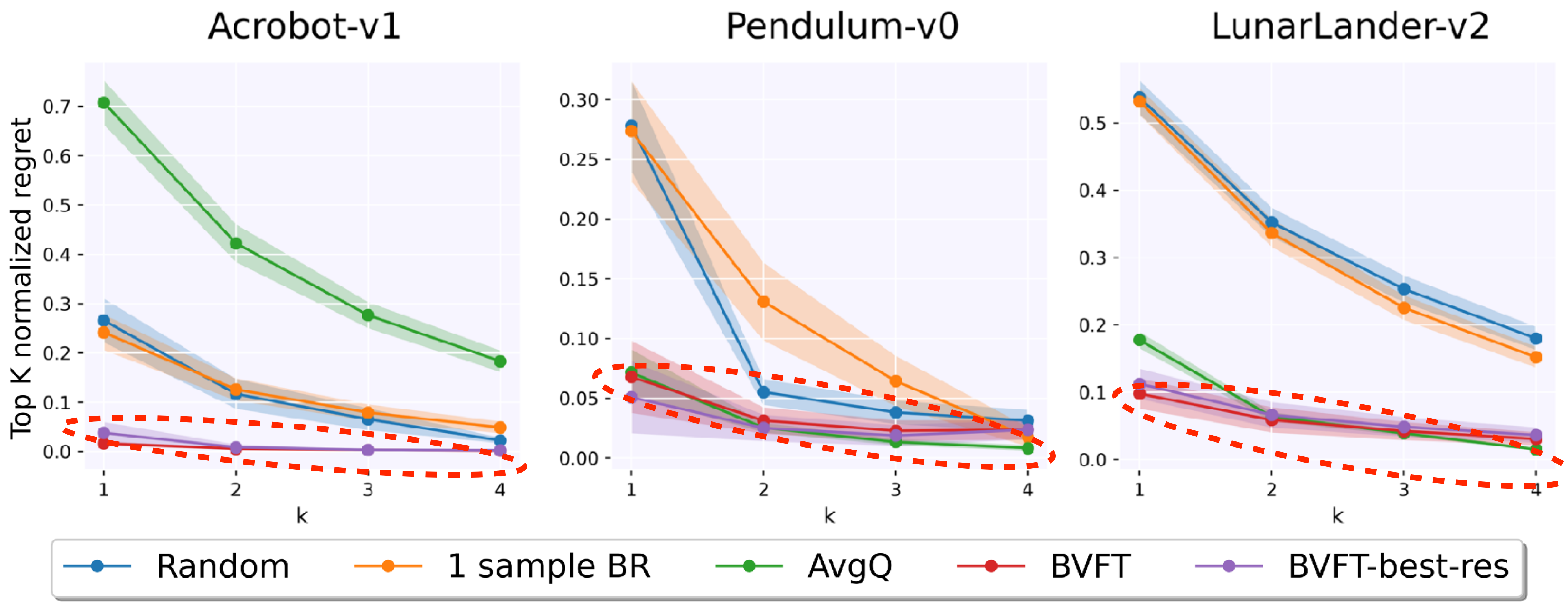

- Algorithm: **BVFT**
  $\arg\min_i \max_j \|f_i - \mathrm{Proj}_{\mathcal{G}_{i,j}}(\mathcal{T} f_i)\|_{2,D}$
- Sample complexity poly in horizon, $1/\varepsilon$, log(#candidates), and $C$ (data coverage)
- Computation: #data points * $|F|^2$

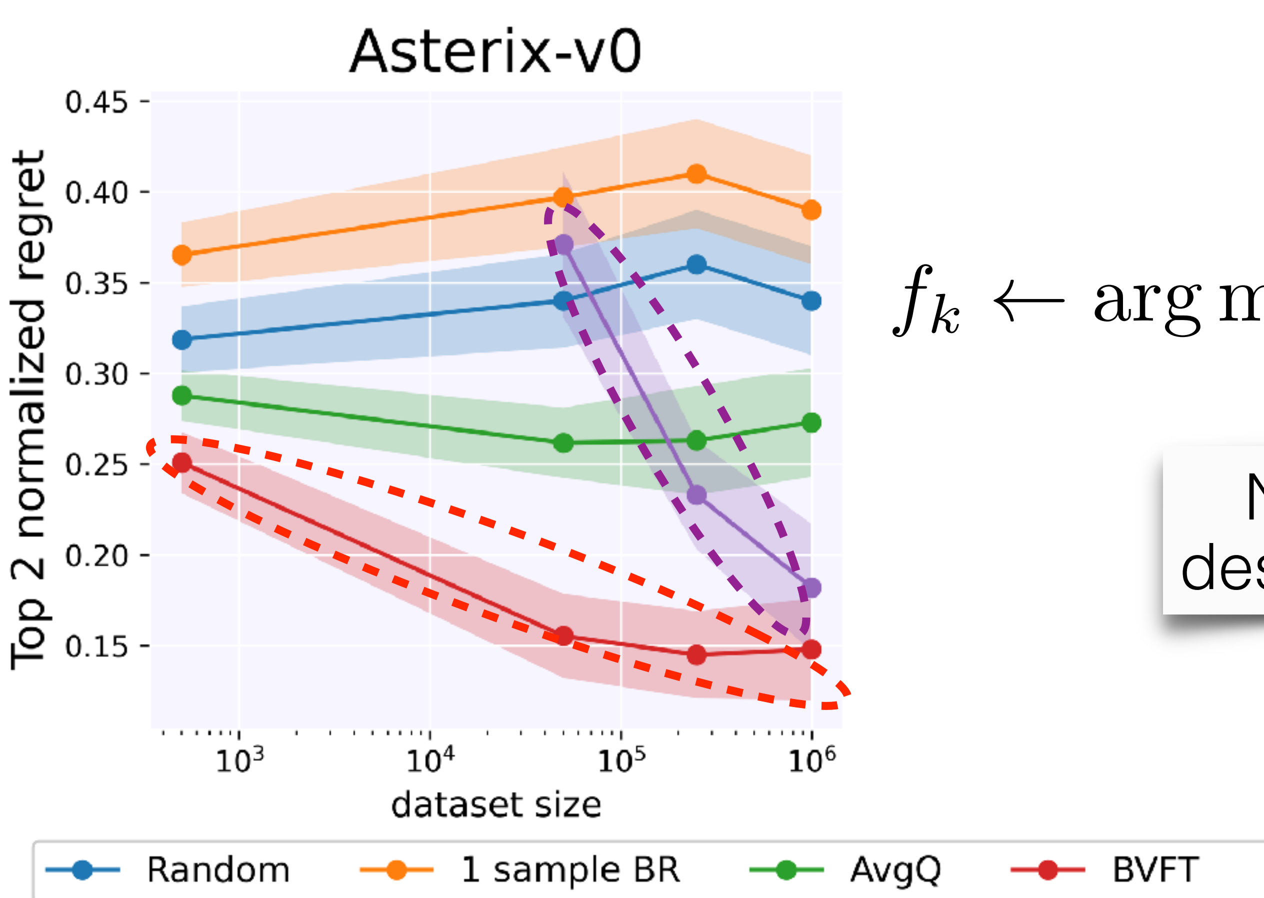
- (Simplified) problem: identify $Q^*$ out of $F = \{ f_1, f_2 \}$
- Partition $S$x$A$ according to both functions simultaneously!
  - Pw-const class $G_{1,2}$ w/ size **$O(1/\varepsilon^2)$** **!!**
- Naive extension to >2 functions in $F$: $O(1/\varepsilon^{|F|})$
  - Pairwise comparison + tournament

Formal guarantee in backup slide



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




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

Neural architecture designed by "cheating"



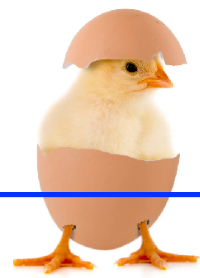
Zhang & **Jiang**. NeurIPS-21. Towards Hyperparameter-free Policy Selection for Offline Reinforcement Learning.

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

(FQI: learn $Q^*$)

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

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



Validation:

(FQE: learn $Q^\pi$)

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

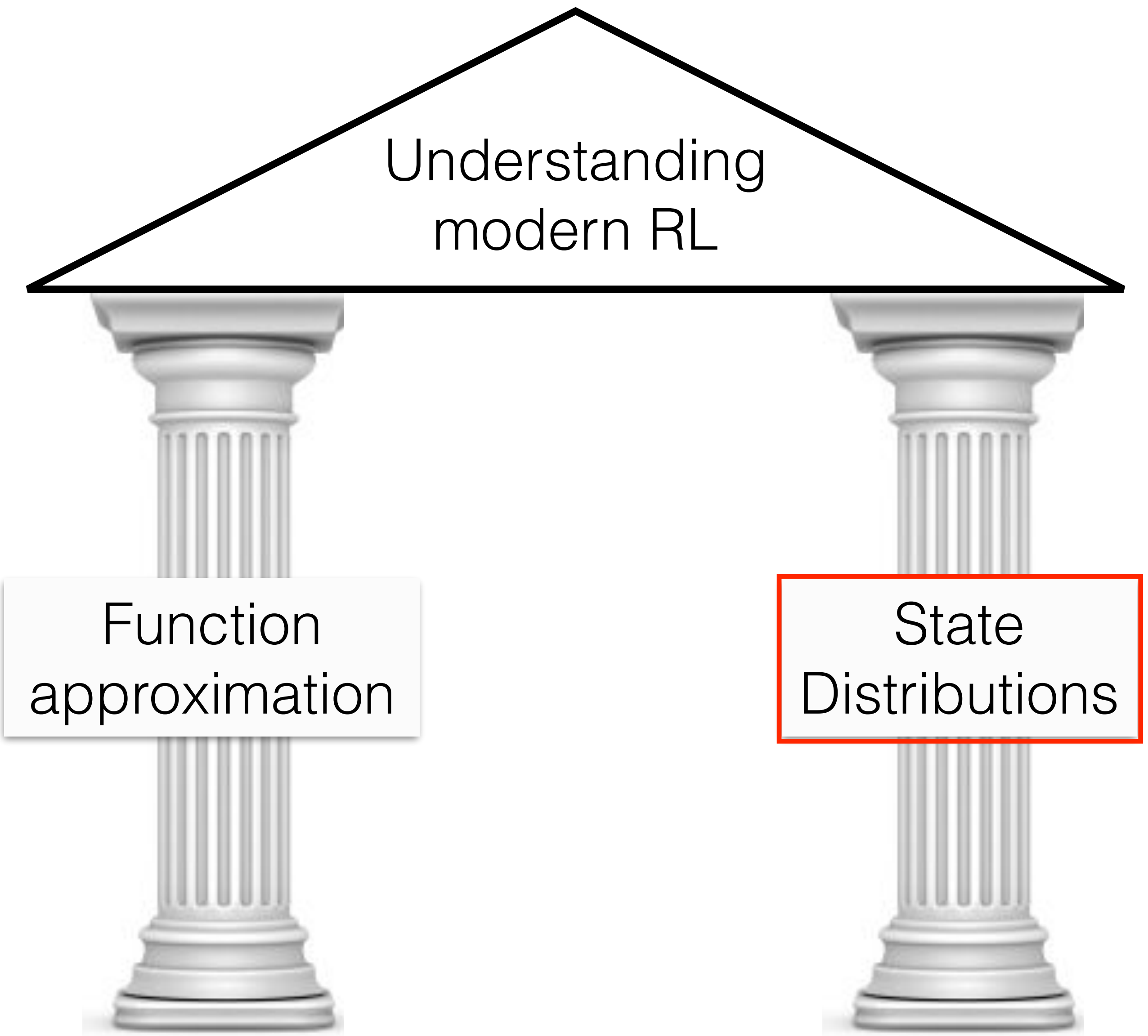
- BVFT: H-P free solution for value-function selection
- Many open problems in validation
    - Data coverage issues (see lower bound [FKSX'22])
    - Combine with different OPE methods e.g., marginalized importance sampling [LLTZ'18, NCDL'19, UHJ'20, JH'20, VJY'21, HJ'22]
    - Practical toolkit (cf.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

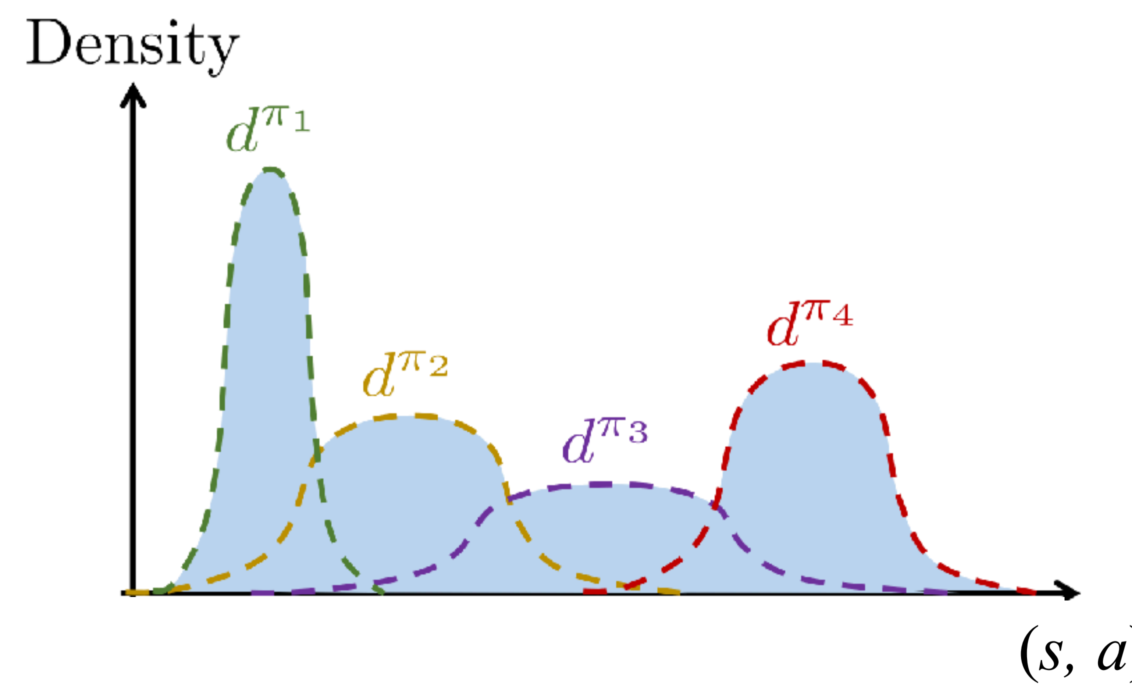
Foster et al. COLT-22. Offline reinforcement learning: Fundamental barriers for value function approximation.
Voloshin et al. NeurIPS-21. Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning.

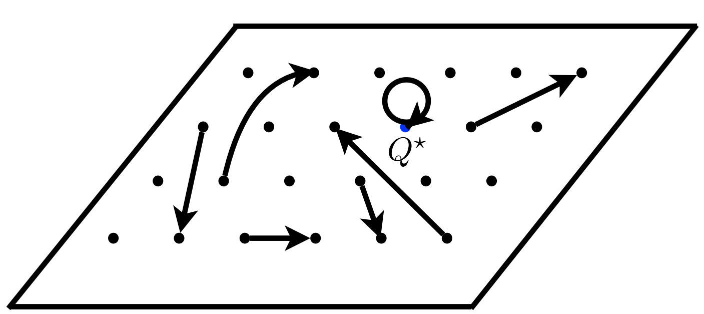
Training: $\hat{f} = f_k$ where

(FQI: learn $Q^*$)

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




ICML | 2022

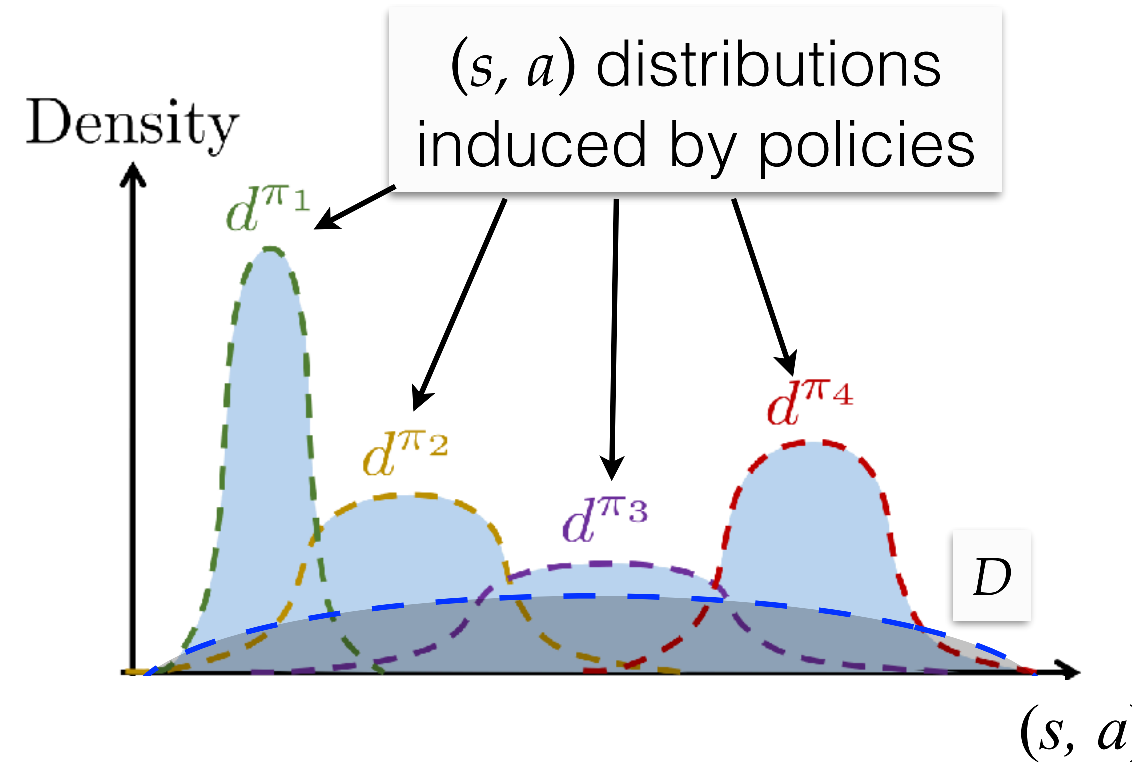
Outstanding Paper Runner Up

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

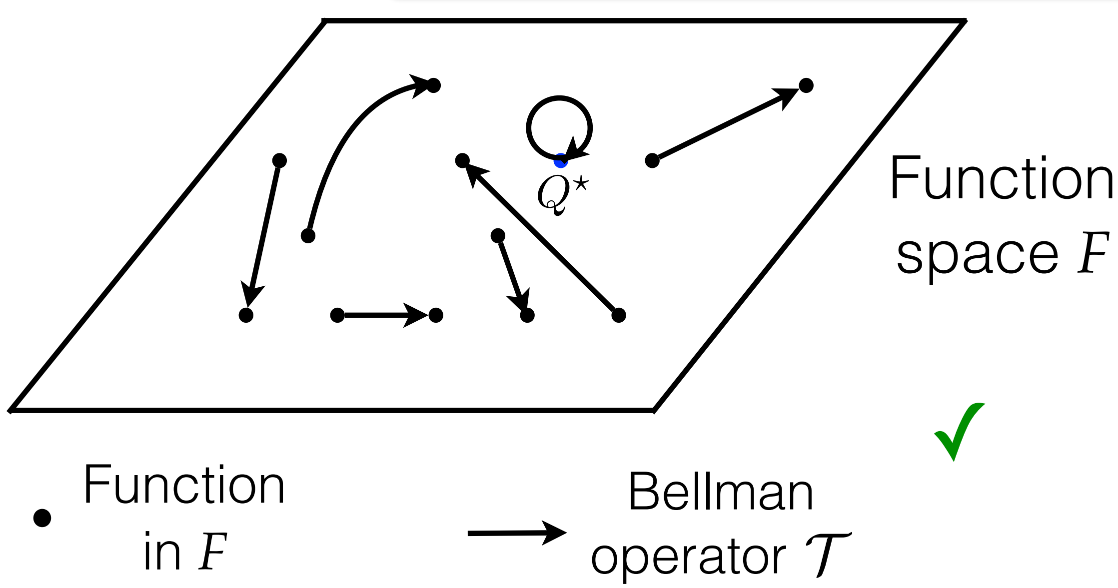
(FQI: learn $Q^*$) $f_k \leftarrow \arg\min_{f_\theta} \mathbb{E}_D \left[ (f_\theta(s,a) - r - \gamma \max_{a'} f_{k-1}(s',a'))^2 \right]$

Standard assumption

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




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

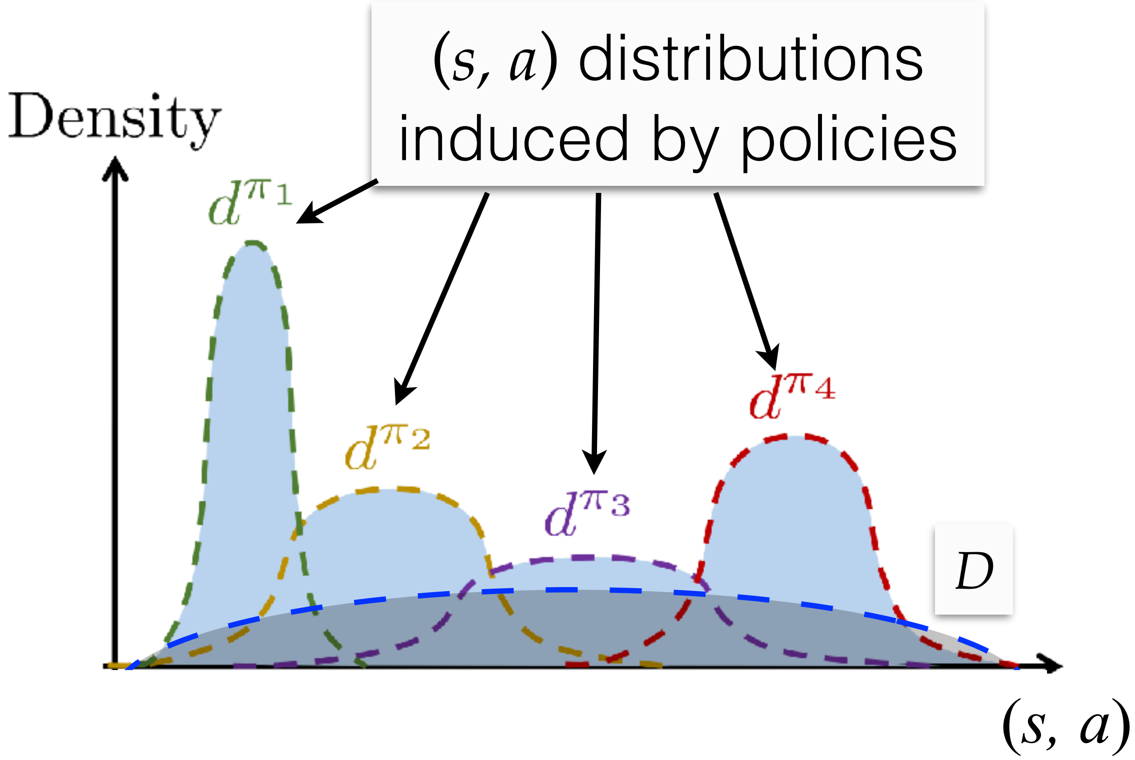


**Training**: $\hat{f} = f_k$ where

(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

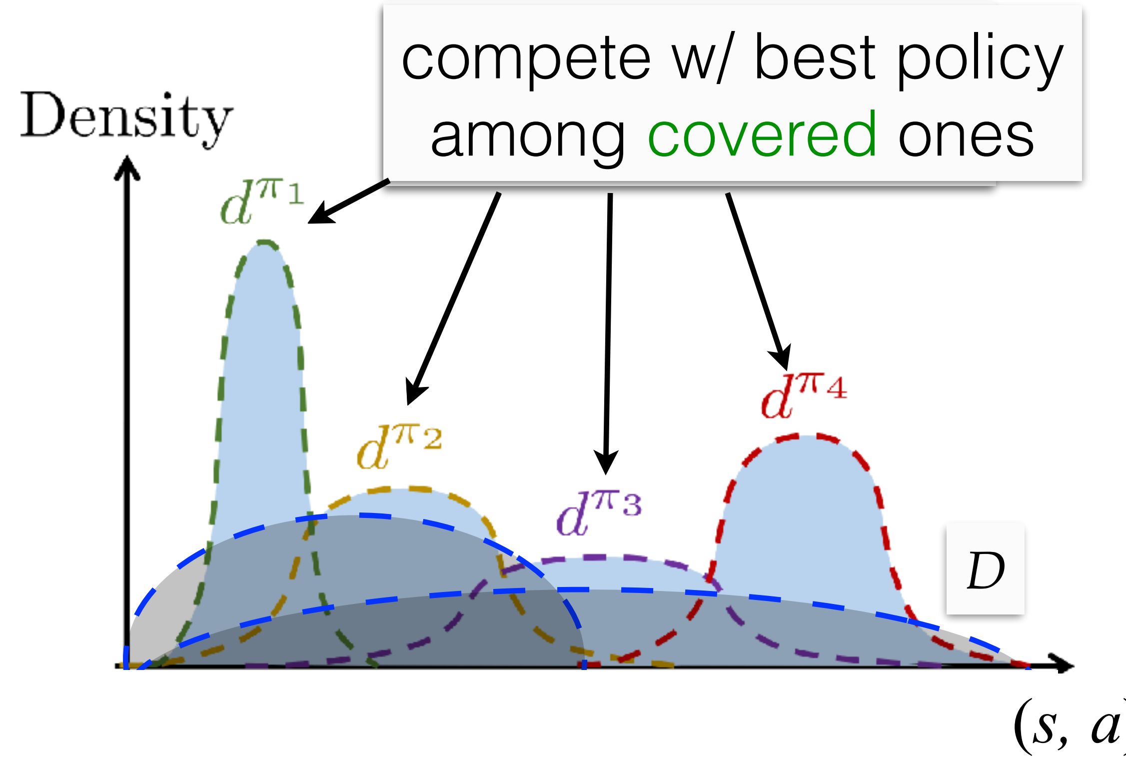



**Challenge:** real-world data lacks exploration!

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

Desirable assumption

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




Offline RL (exploitation)

Goal: stay within data distribution

Principle: *pessimism*

Online RL (exploration)

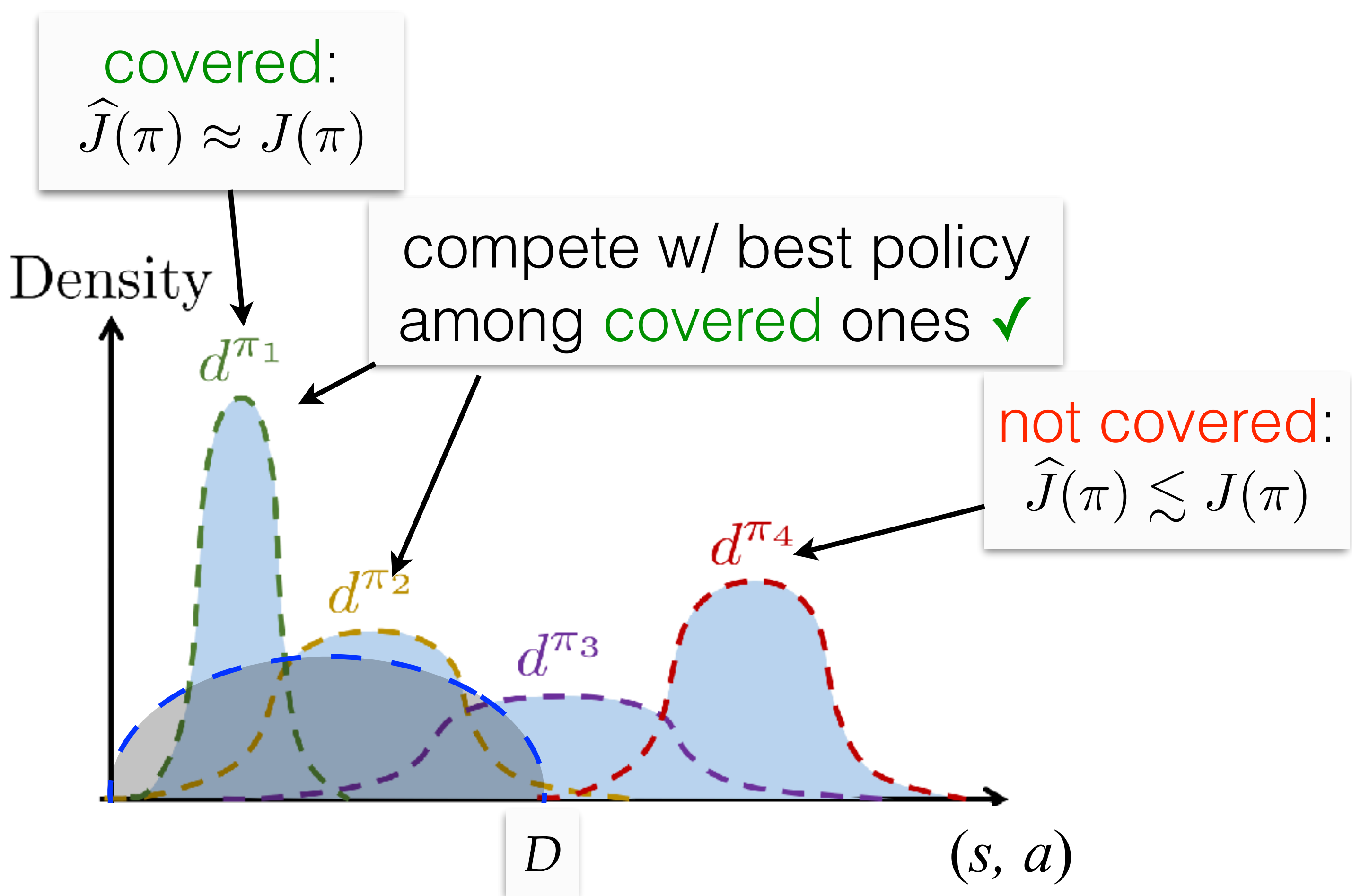
Goal: leave current data distribution

Principle: *optimism*

# Pessimism in face of uncertainty

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

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

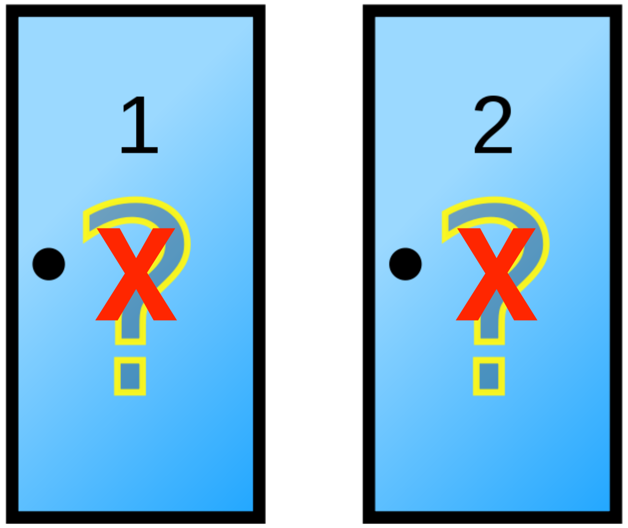





Nilim and Ghaoui. OR-05. Robust control of Markov decision processes with uncertain transition matrices.

# 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 $Q^\pi$ linear, but $\widehat{Q}^\pi$ quadratic

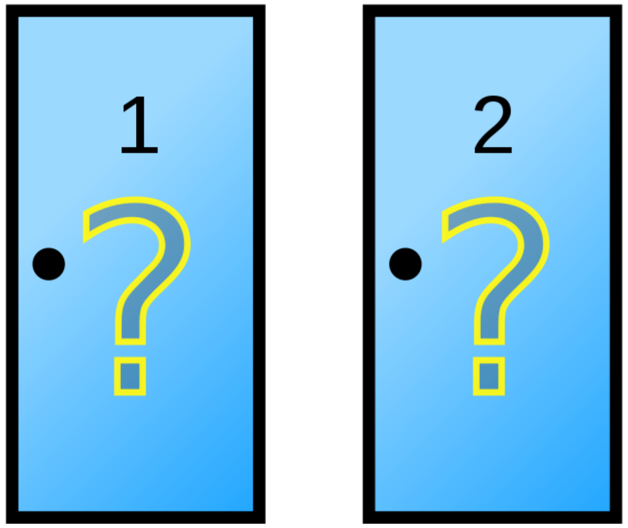





Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

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

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



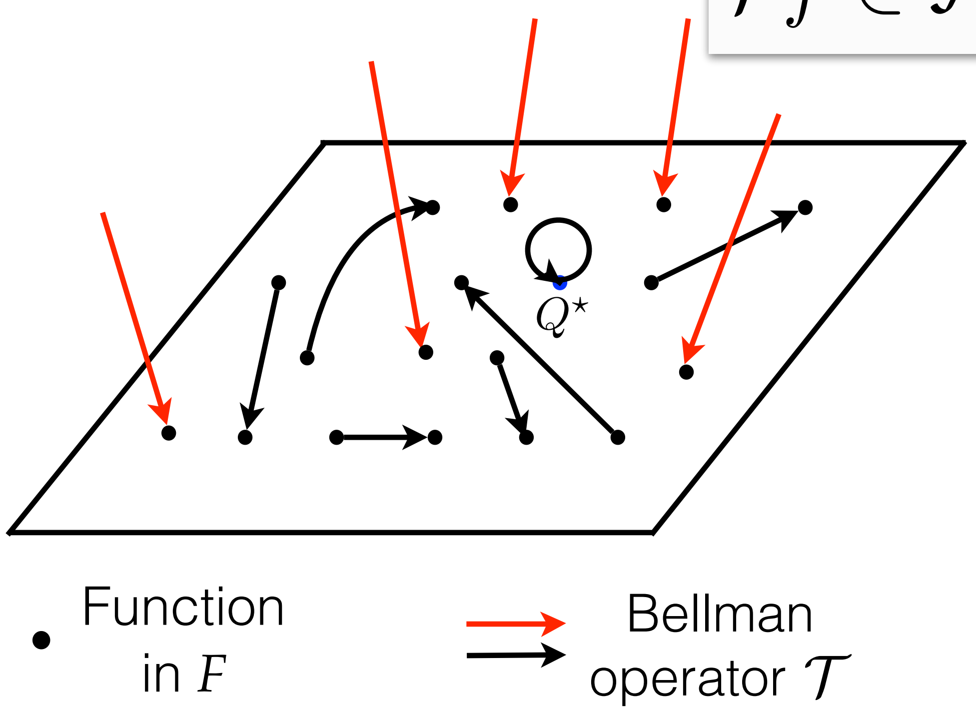



Jin et al. ICML-21. Is pessimism provably efficient for offline RL?

# 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 $Q^{\pi}$ 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

P(s'|s,a) = ✓ x ?

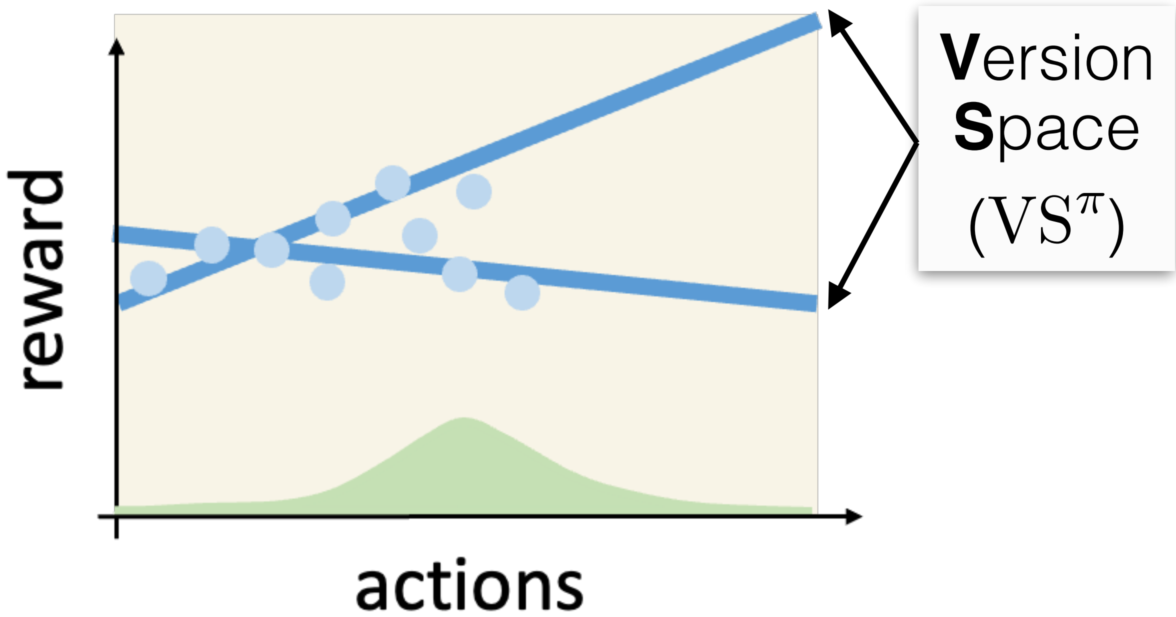
Beyond linear MDPs?

- 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]



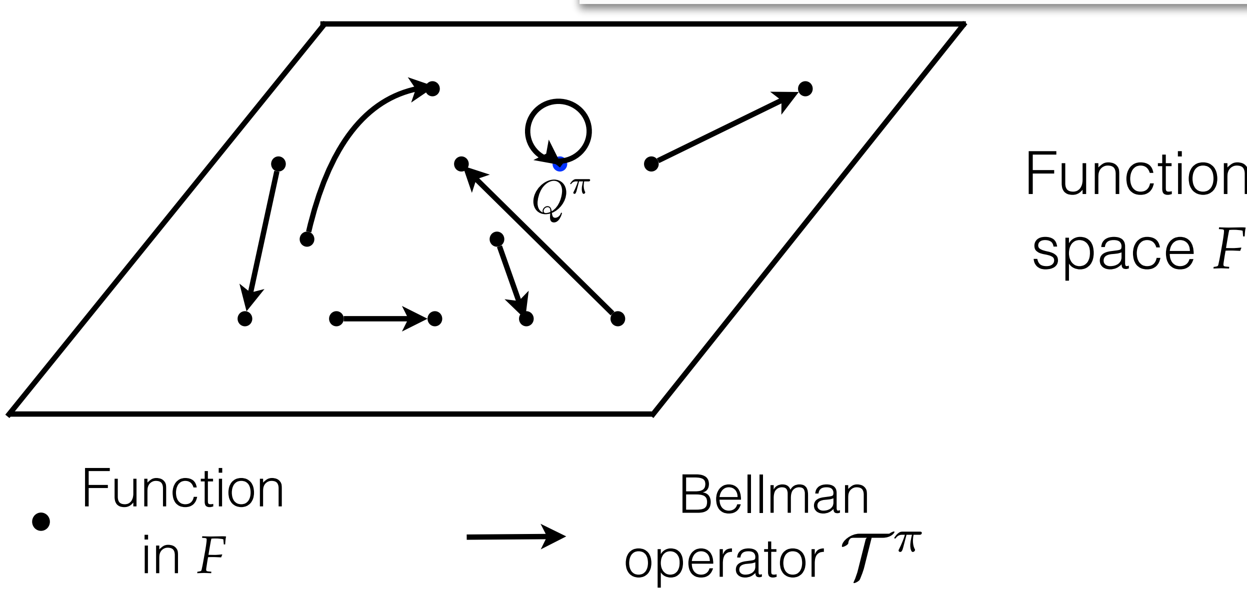

$$\mathrm{VS}^\pi := \{f : \mathbb{E}_D[(f - \mathcal{T}^\pi f)^2] \approx 0\}$$

✓

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

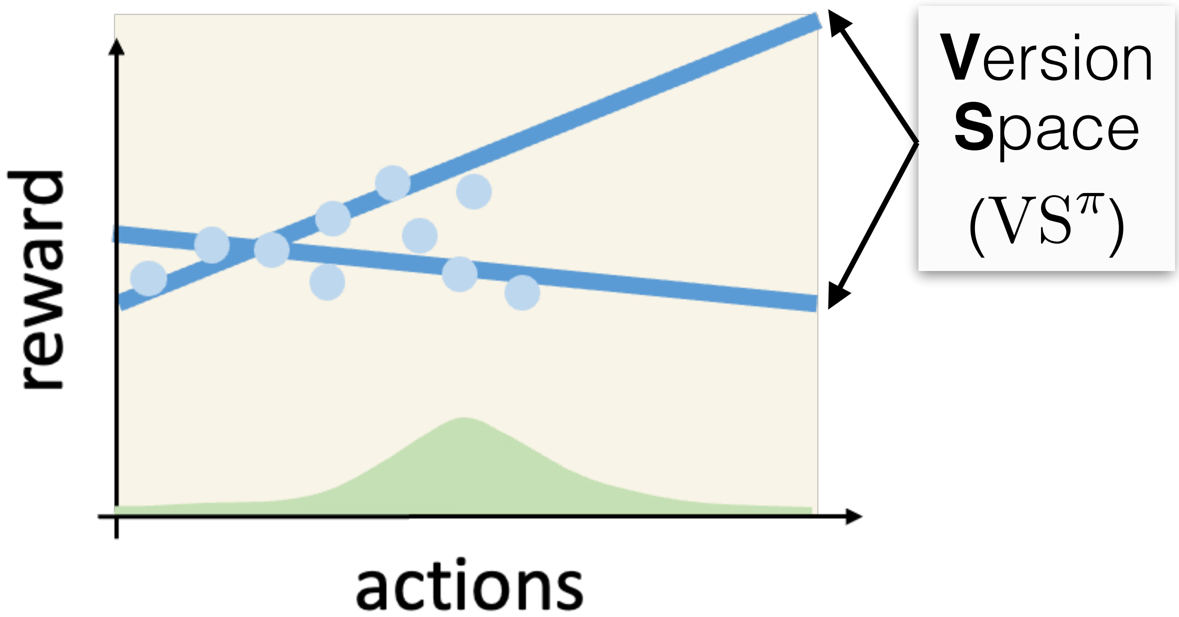
"Bellman-completeness" [AMS'08]

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




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]




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

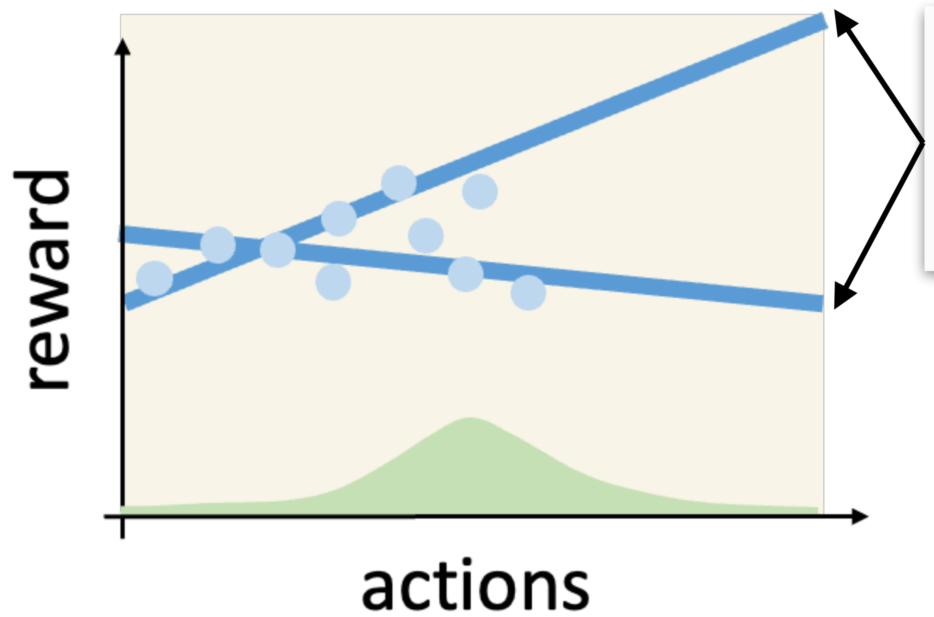
$$\widehat{J}(\pi) = \min_{f \in VS^\pi} f(s_0, \pi)$$

- Overly-conservative by imagining impossible scenarios
  - *Ex.* In linear MDPs, true $Q^\pi$ 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]




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

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

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.

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] X

Statistical generality
vs
Computational tractability?

## Offline RL

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

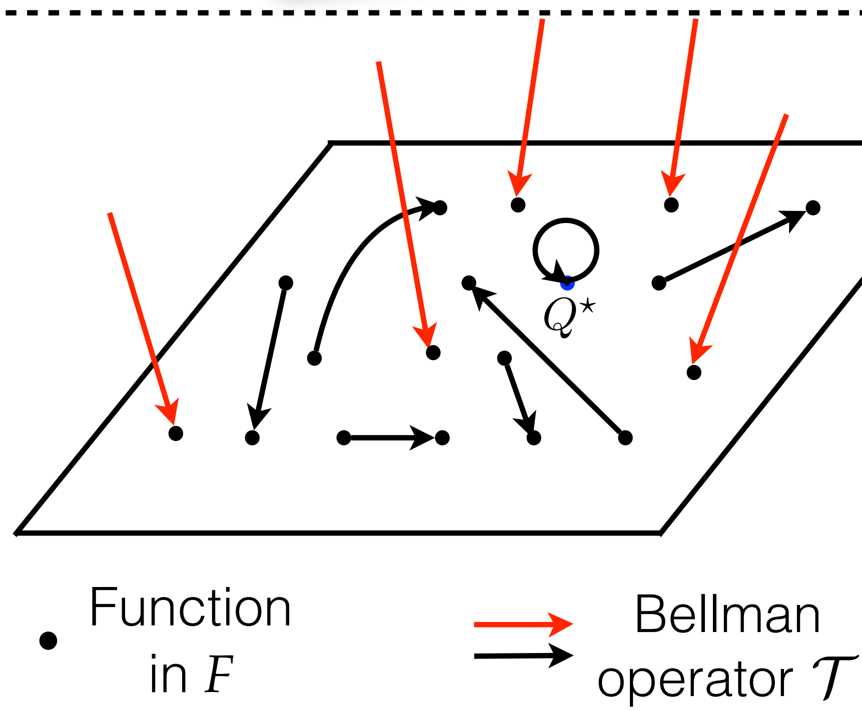
$$f_k \leftarrow \arg\min_{f\in \mathrm{VS}^{\pi_k}} f(s_0,\pi_k)$$

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

- 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 ✓



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

# 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
- 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 [CX**J**A'21]

data policy

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

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

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

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

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

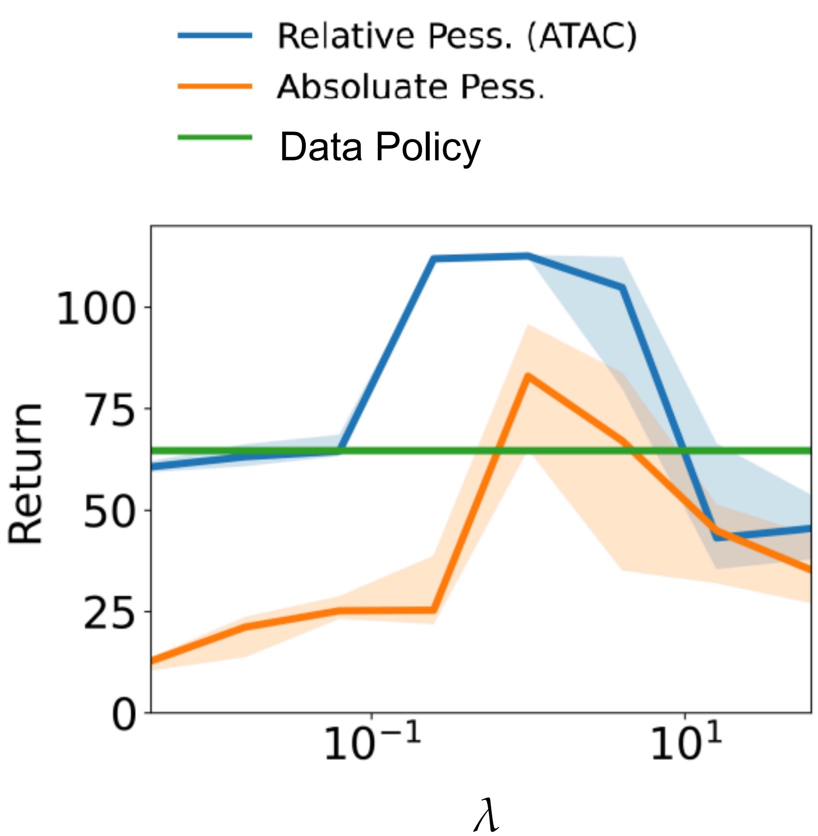
Bellman-error regularization

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



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

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] X

## Offline RL

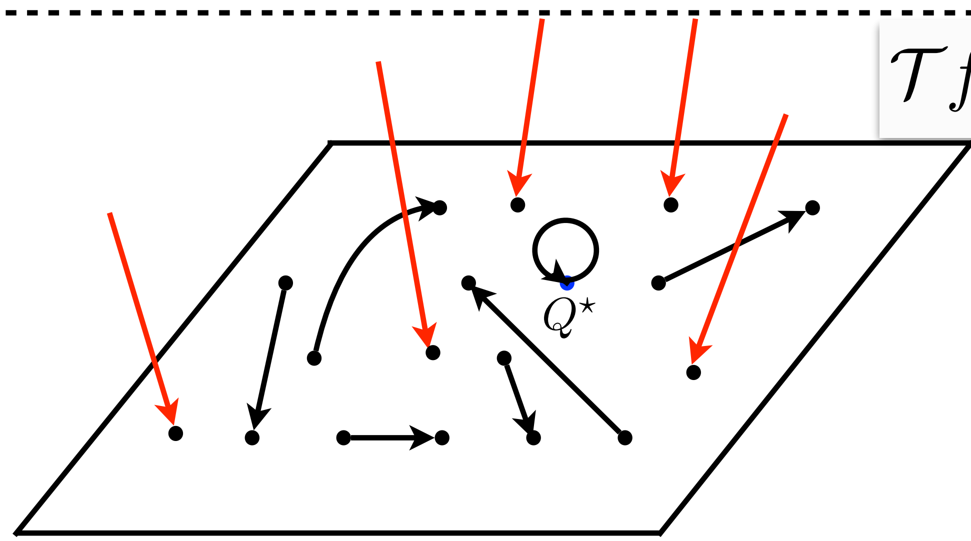
$$\arg\max_{\pi \in \Pi} \min_{f \in \mathrm{VS}^{\pi}} f(s_0, \pi)$$

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

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

- 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 $F$

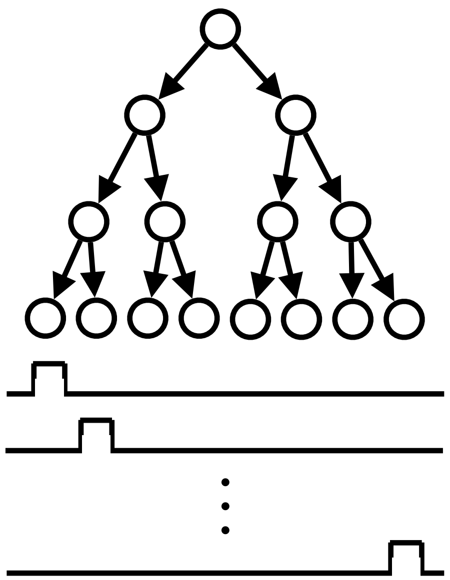
→ Bellman operator $\mathcal{T}$

# 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

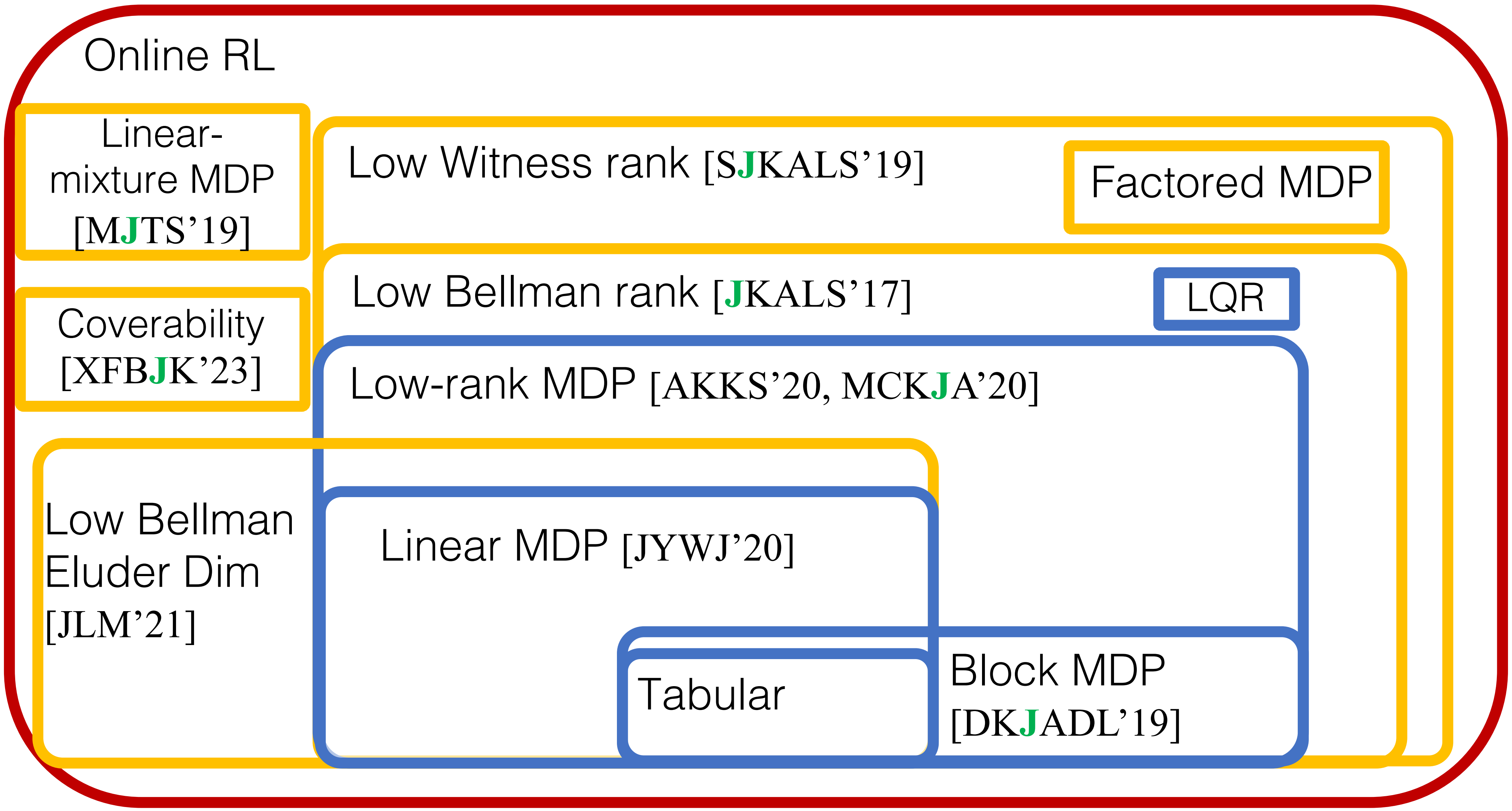


Bounded statistical complexities (e.g., VC-type dim) are insufficient! [KAL'16, JKALS'17]

# 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]




- 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]

# 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]

EDGE
Institute