Hybrid RL:
Efficient RL Using Both Offline and Online Data

Yuda Song
Carnegie Mellon University

Based on work w/ Yifei Zhou, Ayush Sekhari,

Drew Bagnell, Akshay Krishnamurthy, Wen Sun

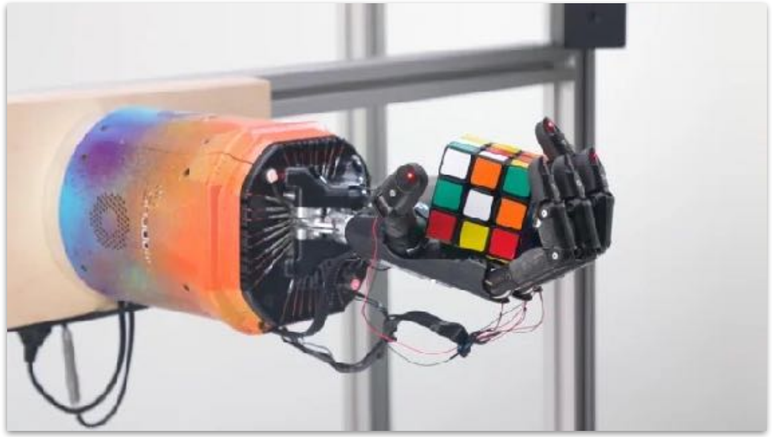
Can we design provably efficient algorithms for

Rich Function Approximation + Model-free RL ?

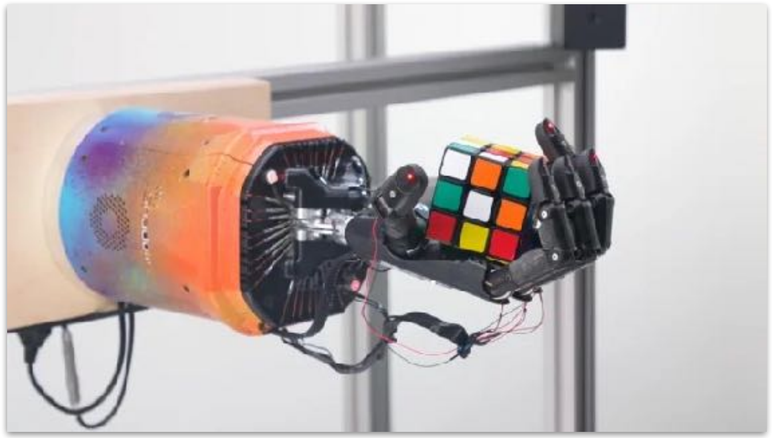
Efficient Model-free RL w/ Rich FA?

Let us just do **online RL.**

RL with naive exploration is hard!



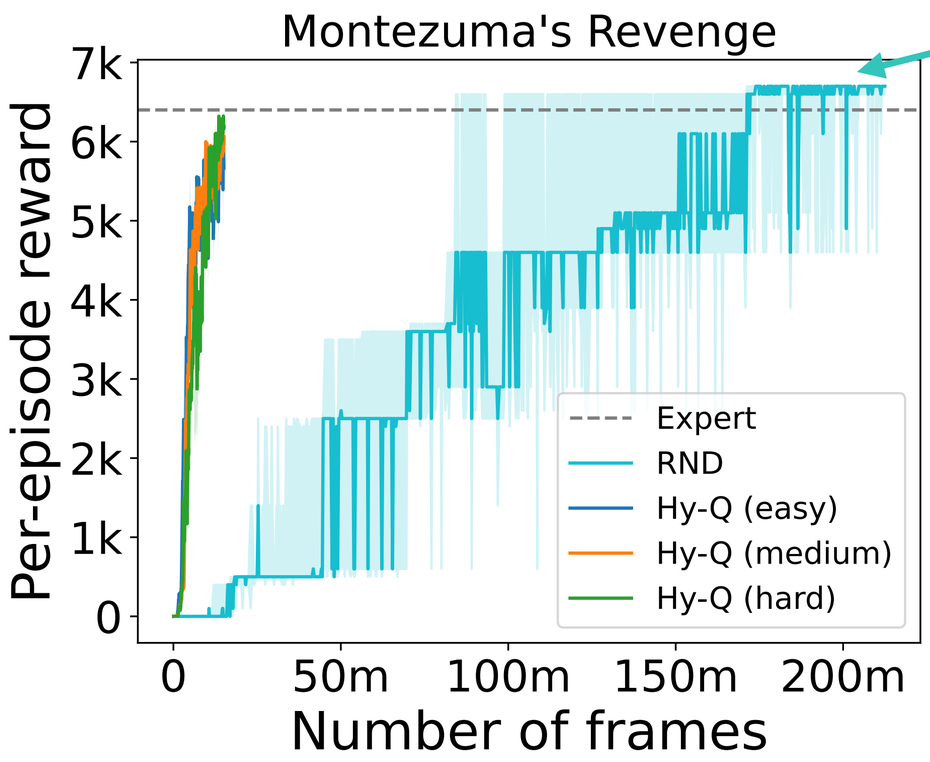
RL with naive exploration is hard!



Challenges:

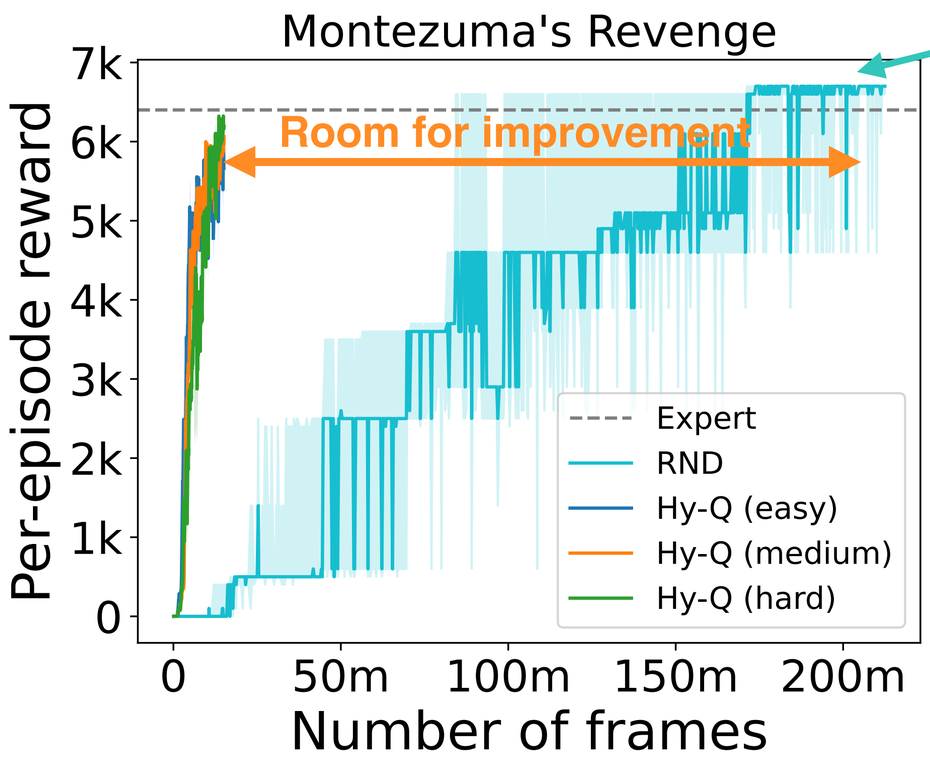
- Naive exploration without utilizing the structure of the problem requires a prohibitive number of samples.
- Some problems may allow a large number of samples (particularly in simulation), others do not.

RL with naive exploration is hard!



Online RL with heuristic exploration

RL with naive exploration is hard!



Online RL with heuristic exploration

Efficient Model-free RL w/ Rich FA?

Let us just do **online RL.**

Requires **a lot of data** without **informed exploration**.

Efficient Model-free RL w/ Rich FA?

Let us just do **online RL.**

Requires **a lot of data** without **informed exploration**.

Efficient Model-free RL w/ Rich FA?

Let us just do **online RL.**

Requires **a lot of data** without **informed exploration**.

Can we leverage **existing data**?

Leverage existing data: offline RL

Image credit: Ego4D, RT-X

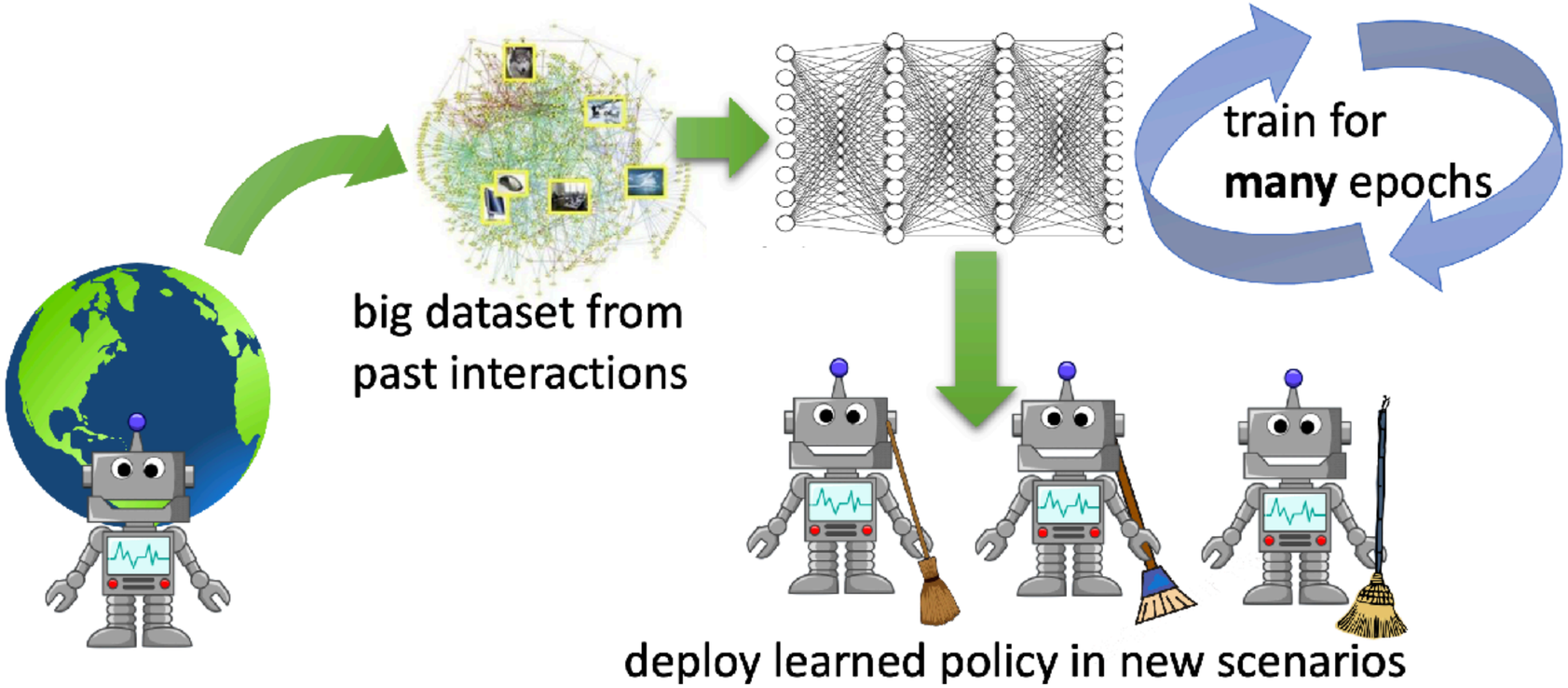
Leverage existing data: offline RL

- Pros: Usually very large data size.
- Cons: Mixed quality: expert & low-quality data.

Image credit: Ego4D, RT-X

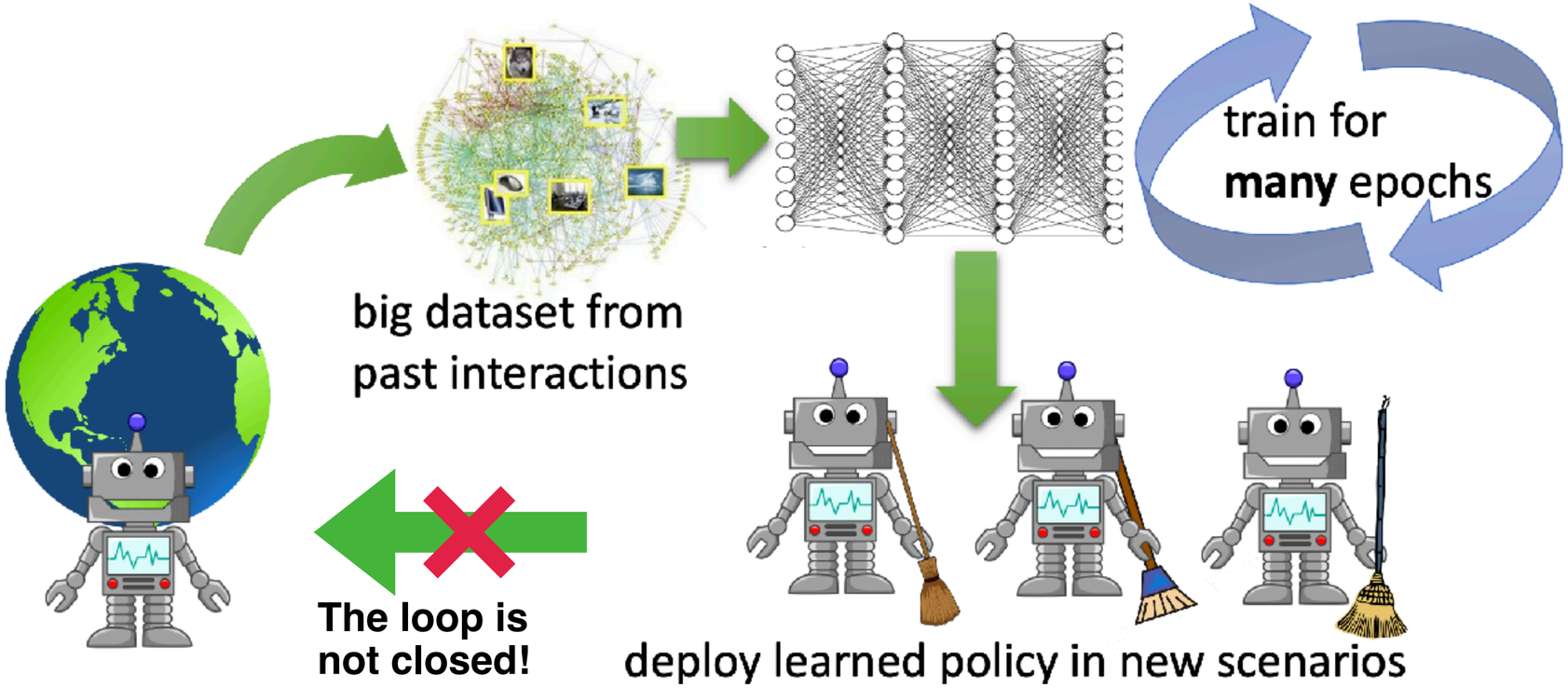
Offline RL

RL with pre-collected dataset



Offline RL

RL with pre-collected dataset



Offline RL

The reality: Making offline RL work reliably is hard…

Jin, Ying, Zhuoran Yang, and Zhaoran Wang. "Is pessimism provably efficient for offline rl?." *International Conference on Machine Learning*. PMLR, 2021.
Jin, Xue-Kun, et al. "Hybrid Value Estimation for Off-policy Evaluation and Offline Reinforcement Learning." *arXiv preprint arXiv:2206.02000* (2022).

Offline RL

The reality: Making offline RL work reliably is hard…

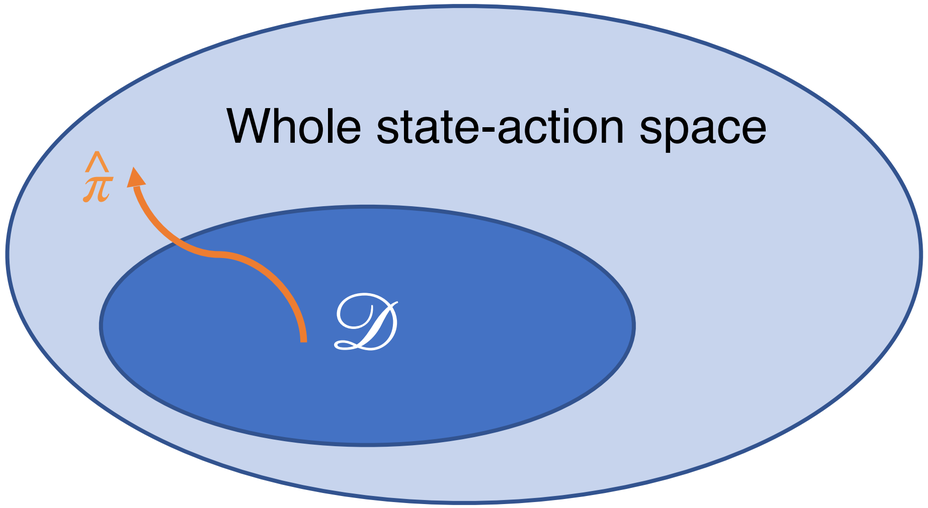
Issue 1: **Distribution shift**

Jin, Ying, Zhuoran Yang, and Zhaoran Wang. "Is pessimism provably efficient for offline rl?." *International Conference on Machine Learning*. PMLR, 2021.
Jin, Xue-Kun, et al. "Hybrid Value Estimation for Off-policy Evaluation and Offline Reinforcement Learning." *arXiv preprint arXiv:2206.02000* (2022).

Offline RL

The reality: Making offline RL work reliably is hard…

Issue 1: **Distribution shift**

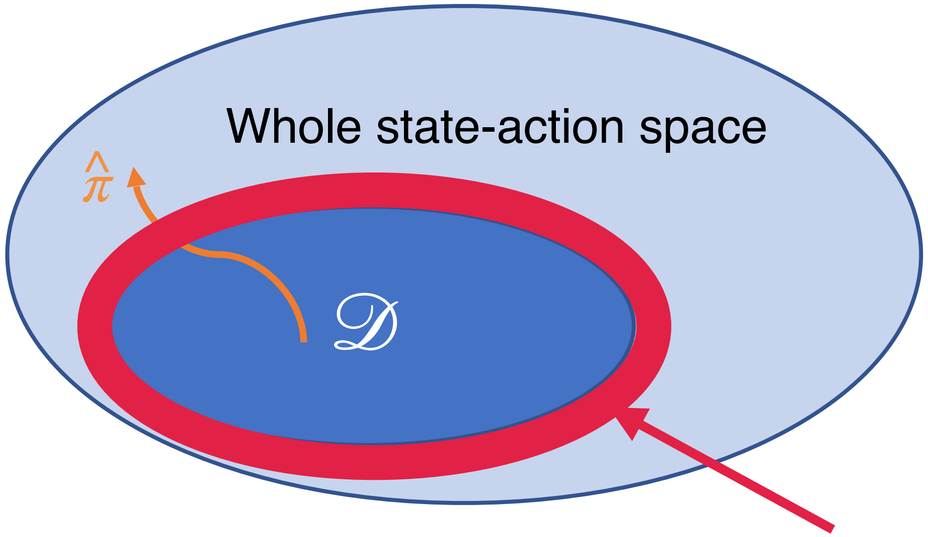


Jin, Ying, Zhuoran Yang, and Zhaoran Wang. "Is pessimism provably efficient for offline rl?." *International Conference on Machine Learning*. PMLR, 2021.
Jin, Xue-Kun, et al. "Hybrid Value Estimation for Off-policy Evaluation and Offline Reinforcement Learning." *arXiv preprint arXiv:2206.02000* (2022).

Offline RL

The reality: Making offline RL work reliably is hard…

Issue 1: **Distribution shift**



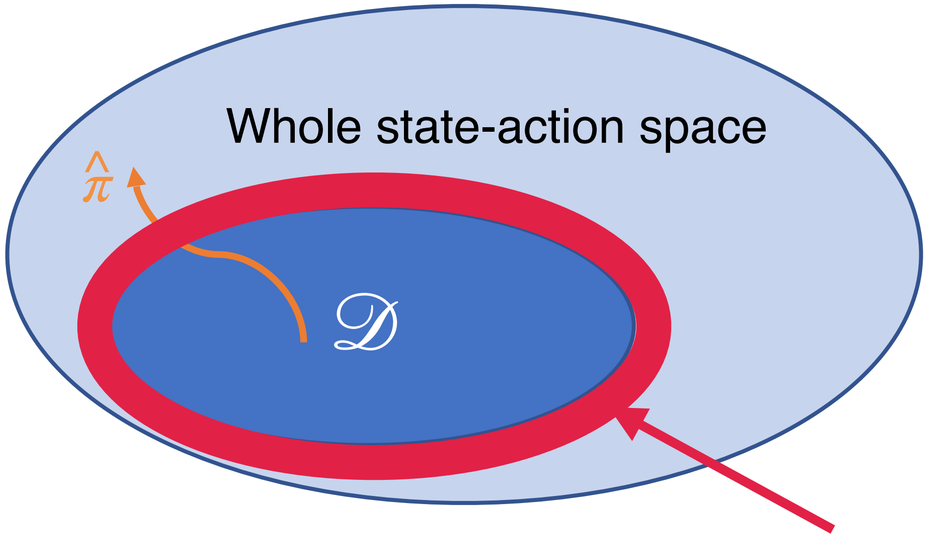
- Carefully design algorithms that restrict search space inside the coverage.
- E.g., Pessimism in the face of uncertainty [Jin et al., 2021].
- Implementing Pessimism under general setting can be challenging.

Jin, Ying, Zhuoran Yang, and Zhaoran Wang. "Is pessimism provably efficient for offline rl?." *International Conference on Machine Learning*. PMLR, 2021.
Jin, Xue-Kun, et al. "Hybrid Value Estimation for Off-policy Evaluation and Offline Reinforcement Learning." *arXiv preprint arXiv:2206.02000* (2022).

Offline RL

The reality: Making offline RL work reliably is hard…

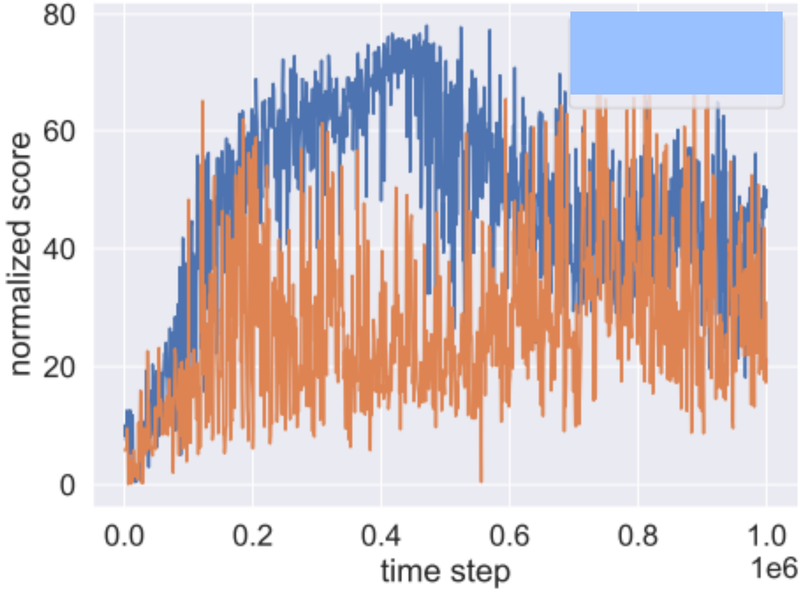
Issue 1: **Distribution shift**



- Carefully design algorithms that restrict search space inside the coverage.
- E.g., Pessimism in the face of uncertainty [Jin et al., 2021].
- Implementing Pessimism under general setting can be challenging.

Issue 2: **The need to verify via online**

Curves of **online performance** of two popular offline deep RL baselines under a standard D4RL benchmark

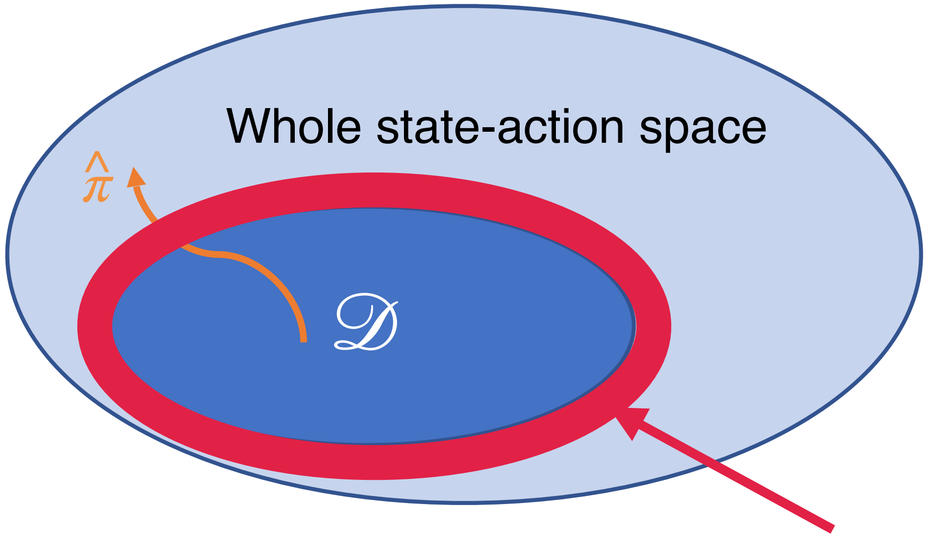


Jin, Ying, Zhuoran Yang, and Zhaoran Wang. "Is pessimism provably efficient for offline rl?." *International Conference on Machine Learning*. PMLR, 2021.
Jin, Xue-Kun, et al. "Hybrid Value Estimation for Off-policy Evaluation and Offline Reinforcement Learning." *arXiv preprint arXiv:2206.02000* (2022).

Offline RL

The reality: Making offline RL work reliably is hard…

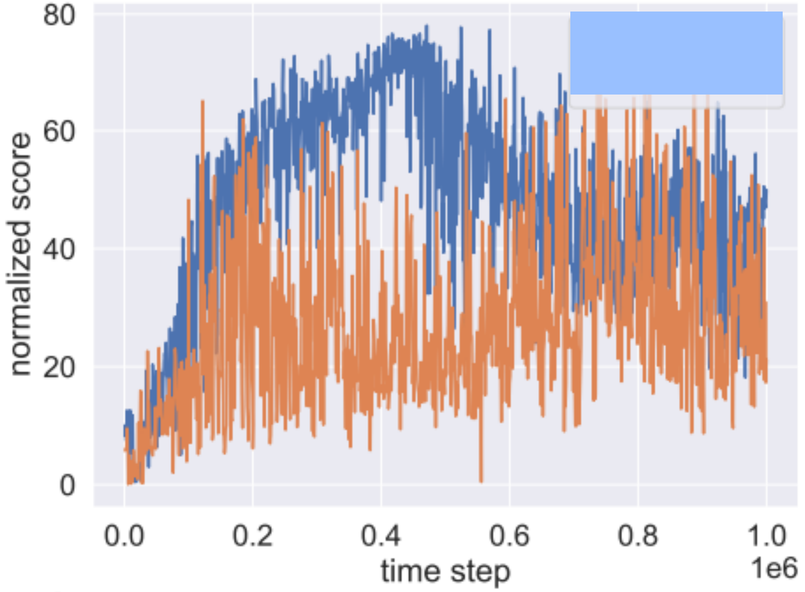
Issue 1: **Distribution shift**



- Carefully design algorithms that restrict search space inside the coverage.
- E.g., Pessimism in the face of uncertainty [Jin et al., 2021].
- Implementing Pessimism under general setting can be challenging.

Issue 2: **The need to verify via online**

Curves of **online performance** of two popular offline deep RL baselines under a standard D4RL benchmark



Q: How do we tune hyperparameters and when do we stop the training?

Jin, Ying, Zhuoran Yang, and Zhaoran Wang. "Is pessimism provably efficient for offline rl?." *International Conference on Machine Learning*. PMLR, 2021.
Jin, Xue-Kun, et al. "Hybrid Value Estimation for Off-policy Evaluation and Offline Reinforcement Learning." *arXiv preprint arXiv:2206.02000* (2022).

Efficient Model-free RL w/ Rich FA?

Let us just do **online RL.**

Requires **a lot of data** without **informed exploration**.

Leverage **existing data -> offline RL.**

Computation & **Verification**.

Efficient Model-free RL w/ Rich FA?

Let us just do **online RL.**

Requires **a lot of data** without **informed exploration**.

Leverage **existing data -> offline RL.**

Computation & **Verification**.

Efficient Model-free RL w/ Rich FA?

Let us just do **online RL.**

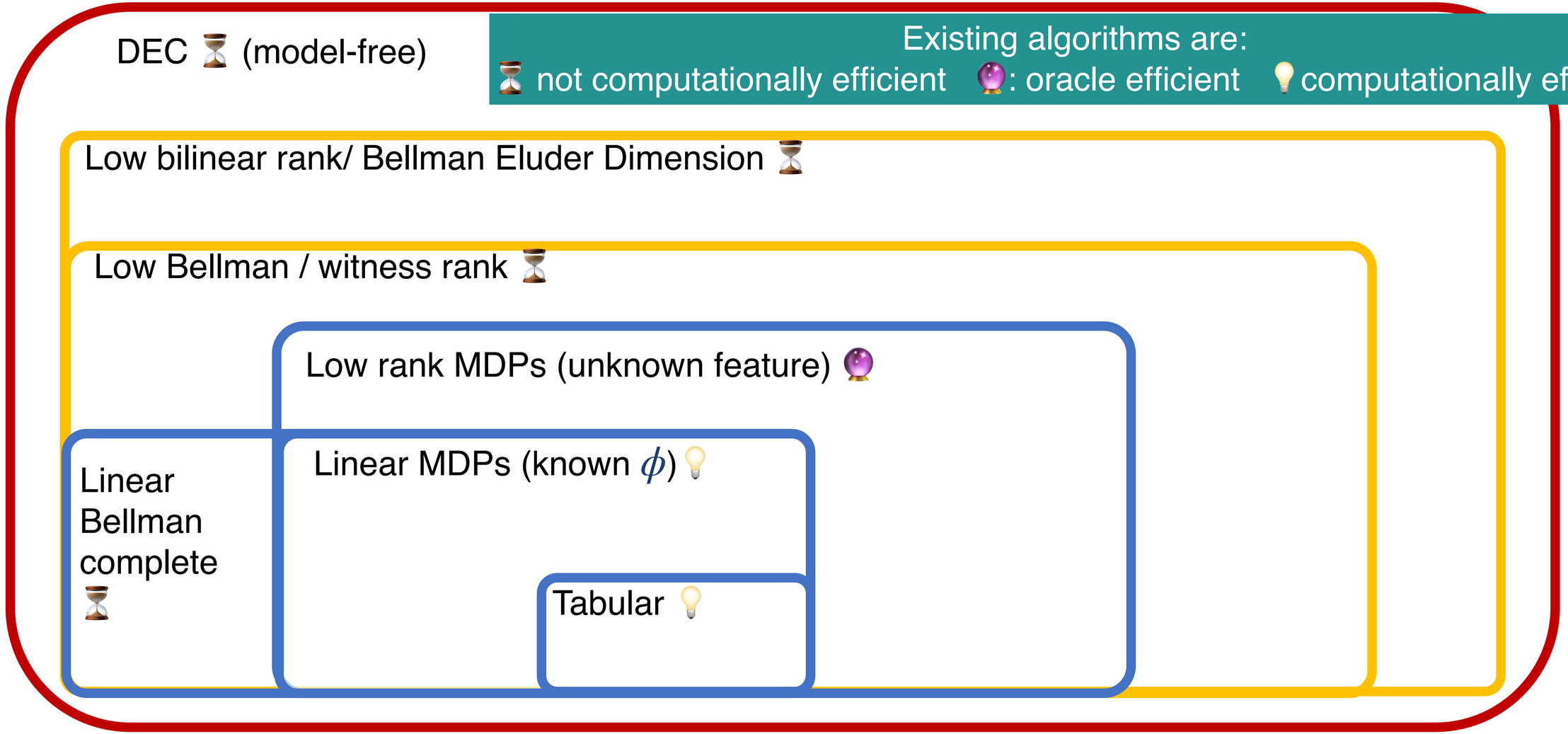
Requires **a lot of data** without **informed exploration**.

Leverage **existing data -> offline RL.**

Computation & **Verification**.

Leverage **structure** of the problem.

Online RL with structural assumptions



Slide credit:

Akshay Krishnamurthy

Kane, Daniel, et al. "Exponential hardness of reinforcement learning with linear function approximation." *arXiv preprint arXiv:2302.12940* (2023).

Efficient Model-free RL w/ Rich FA?

Let us just do **online RL.**

Requires **a lot of data** without **informed exploration**.

Leverage **existing data -> offline RL.**

Computation & Verification.

Leverage **structure** of the problem.

Again, Computation.

Our toolkits

Online RL

Offline RL

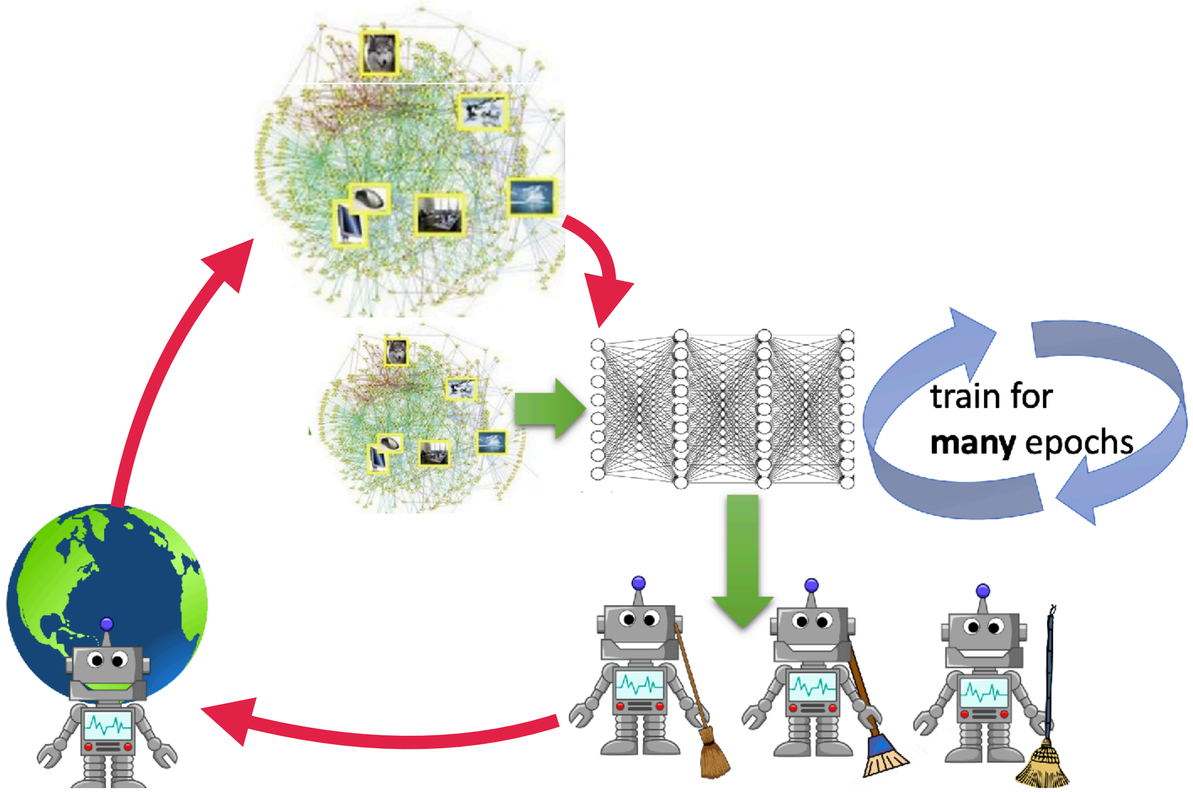
Structure

Our toolkits

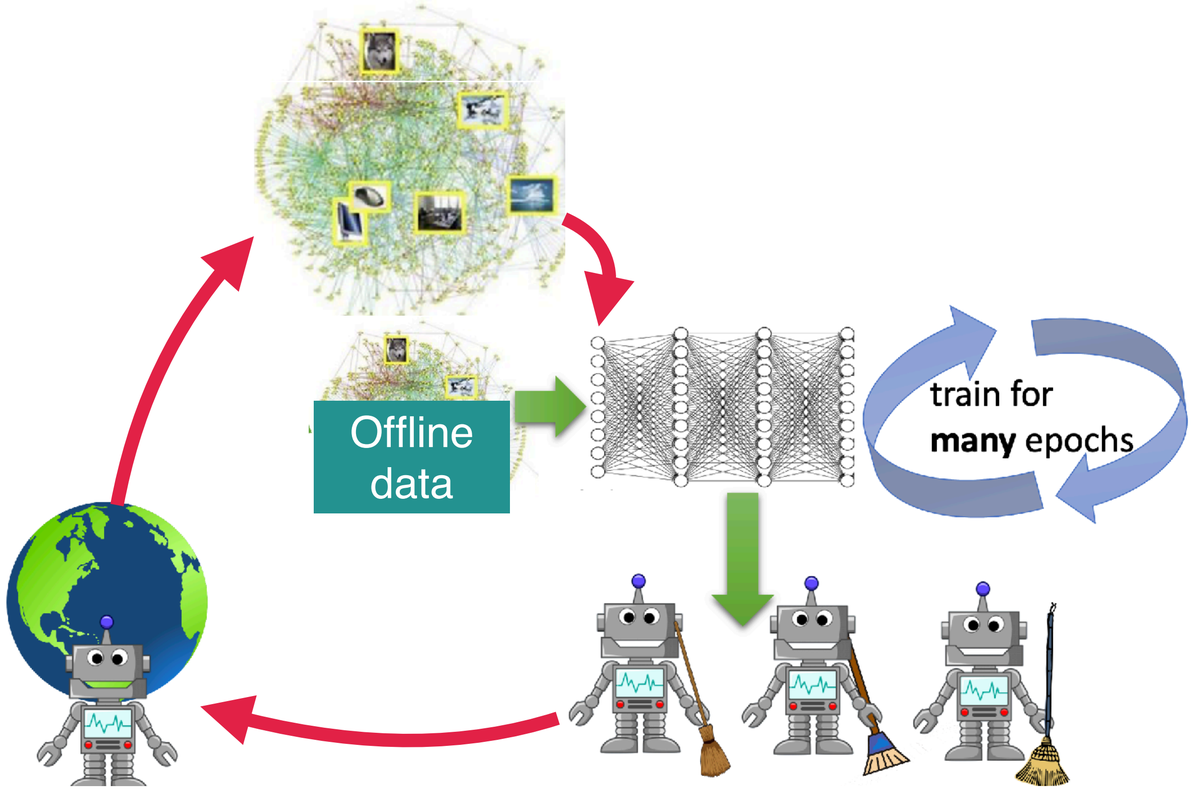
Online RL
Offline RL

Structure

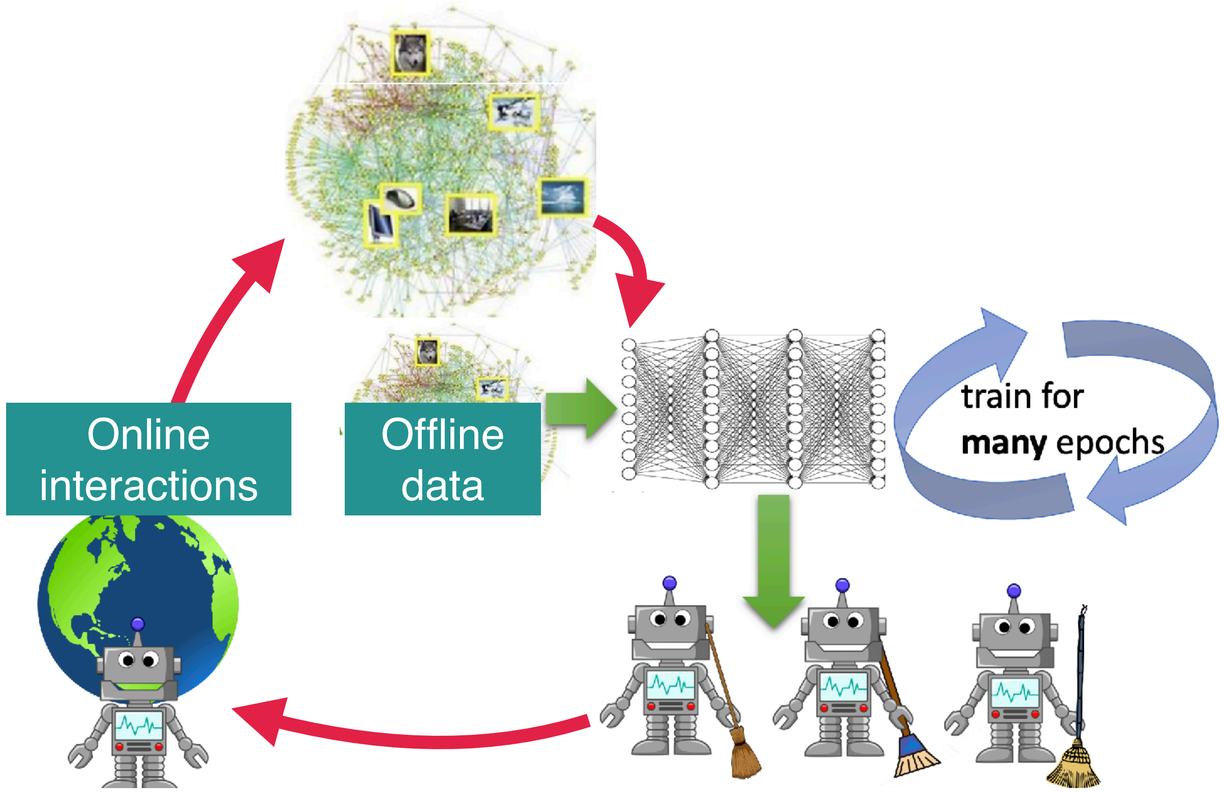
To the rescue: Offline data + Online interaction



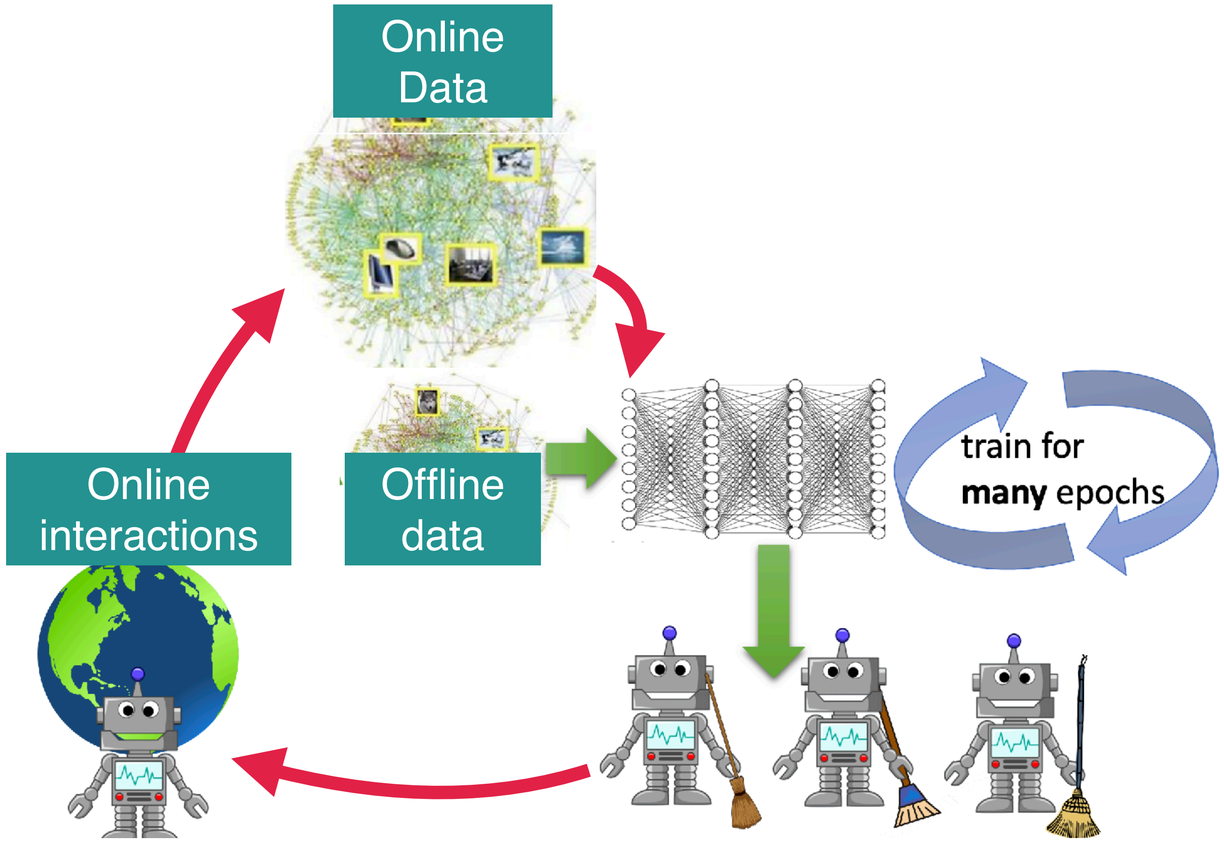
To the rescue: Offline data + Online interaction



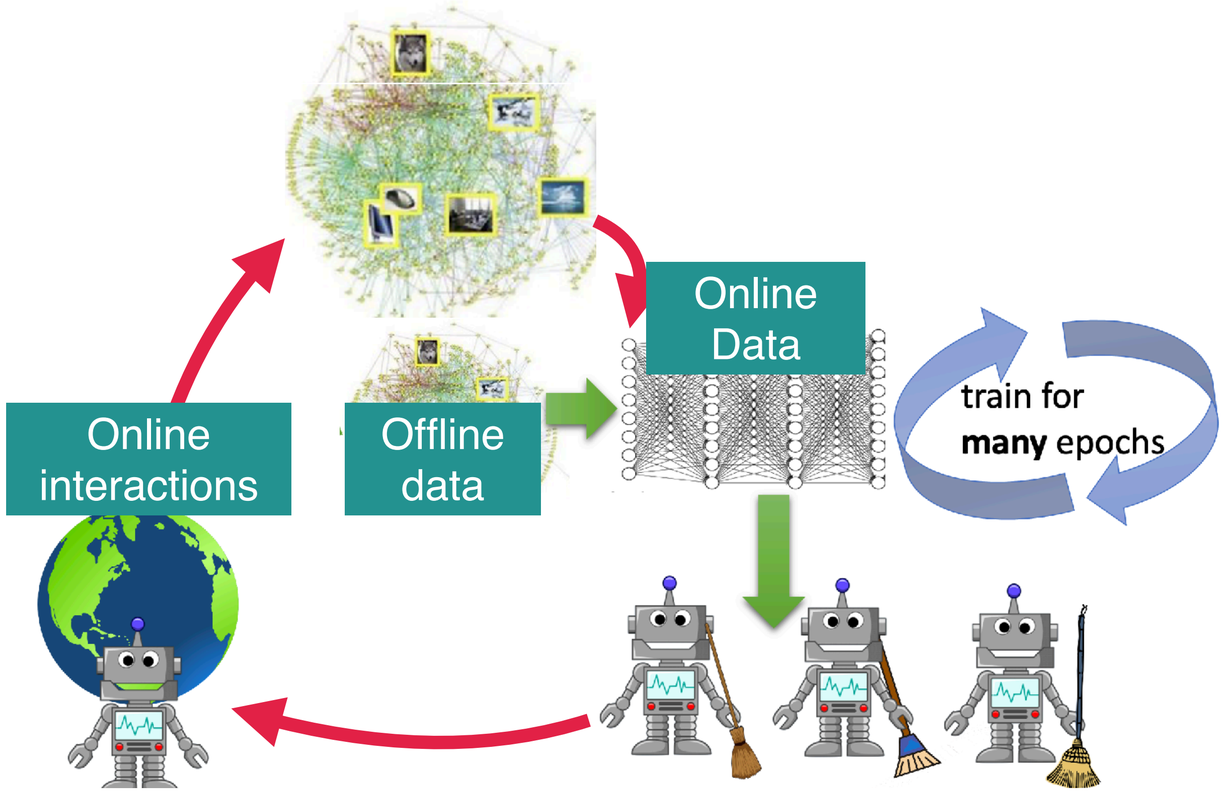
To the rescue: Offline data + Online interaction



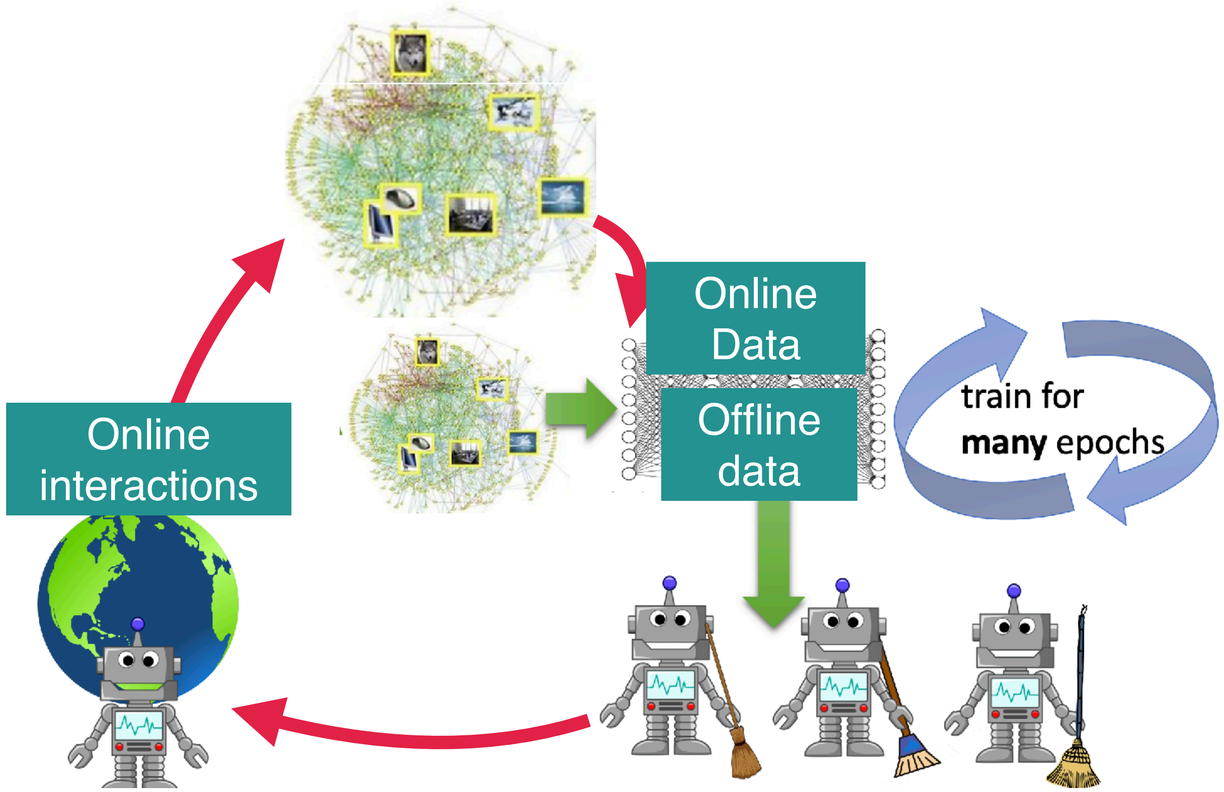
To the rescue: Offline data + Online interaction



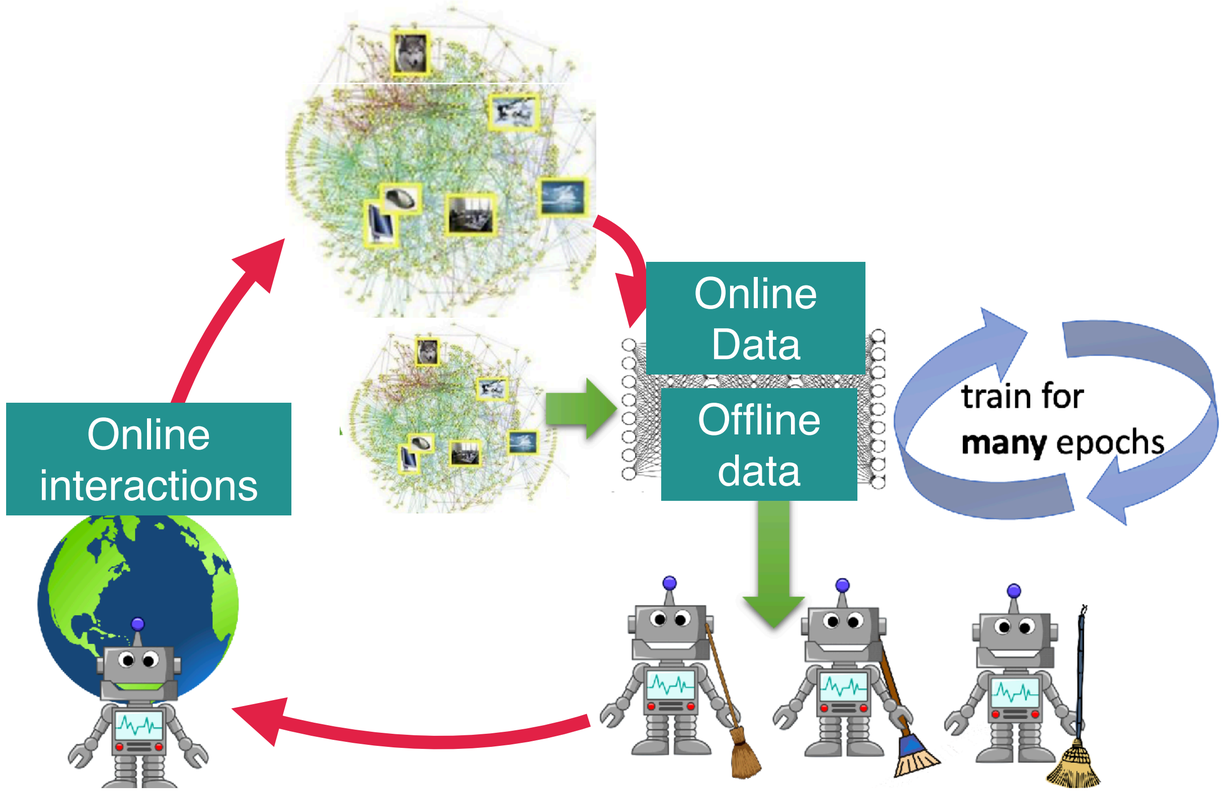
To the rescue: Offline data + Online interaction



To the rescue: Offline data + Online interaction

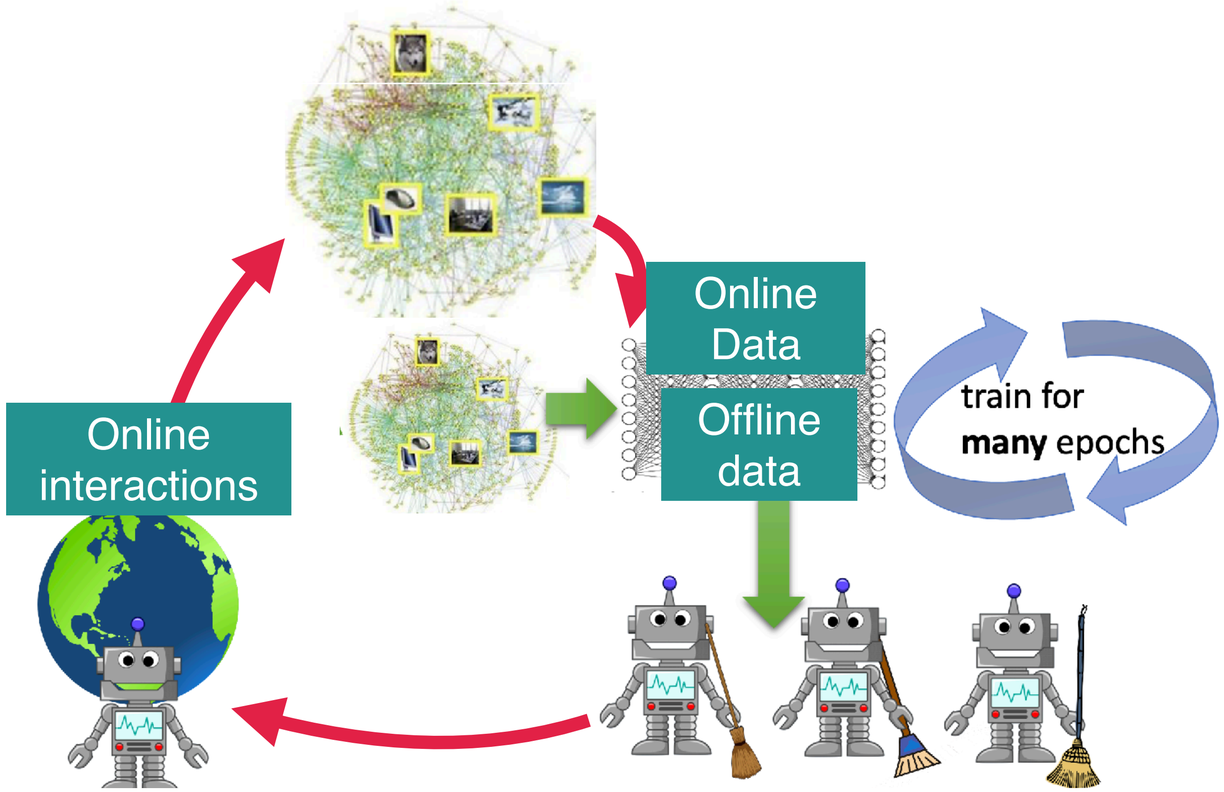


To the rescue: Offline data + Online interaction



Previous works:

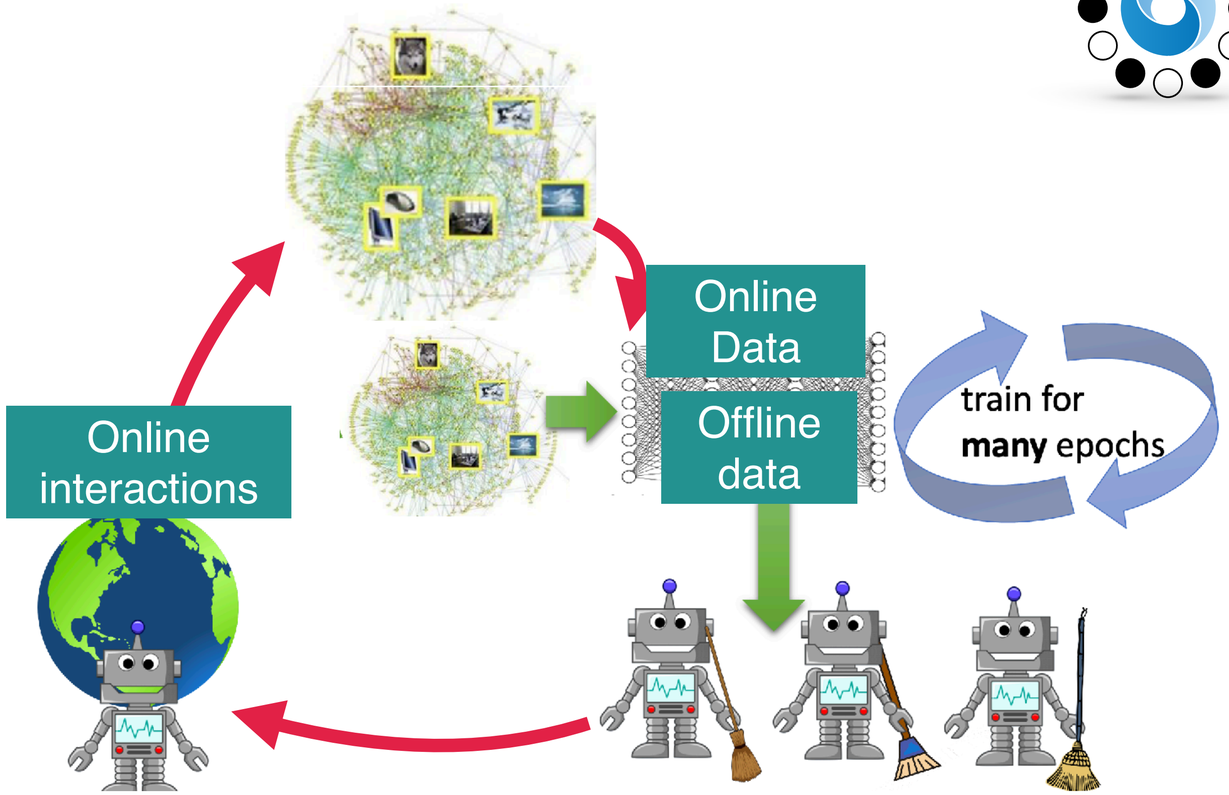
To the rescue: Offline data + Online interaction



Previous works:

- Lots of applications:

To the rescue: Offline data + Online interaction

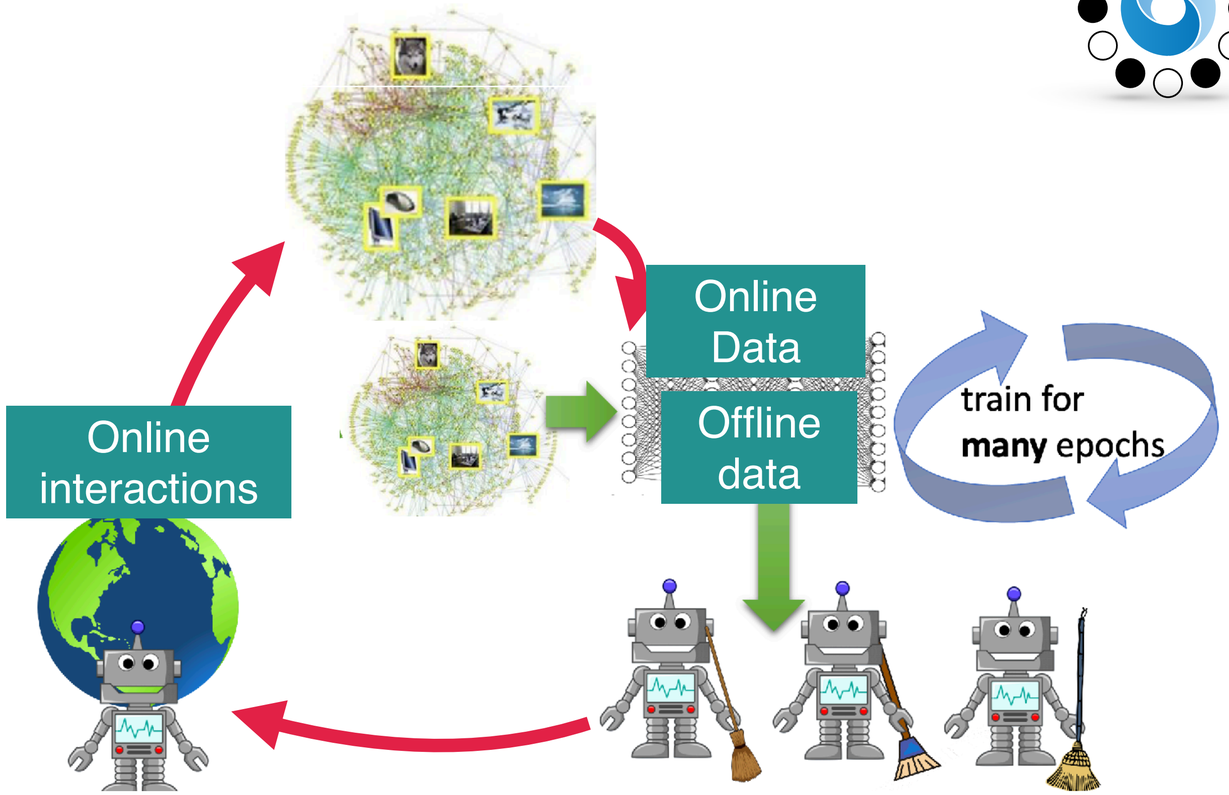


AlphaGo

Previous works:

- Lots of applications:

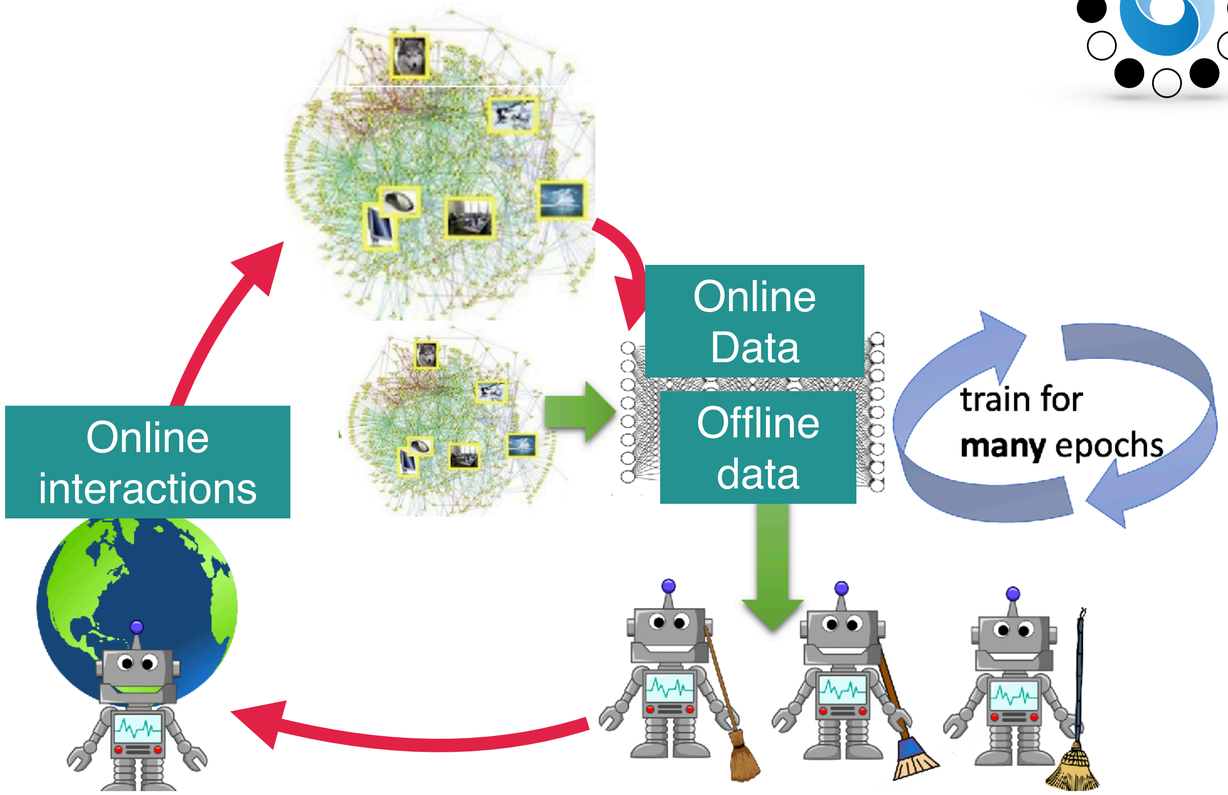
To the rescue: Offline data + Online interaction



Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]

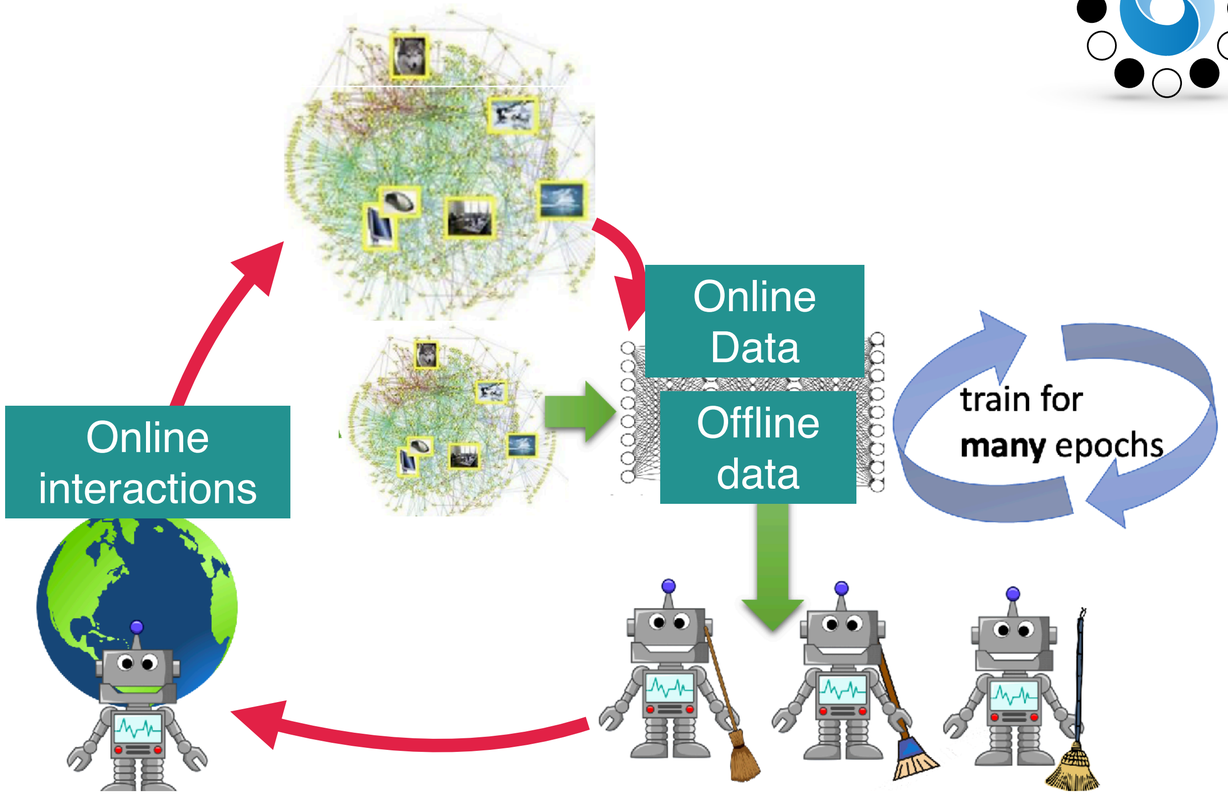
To the rescue: Offline data + Online interaction



Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]

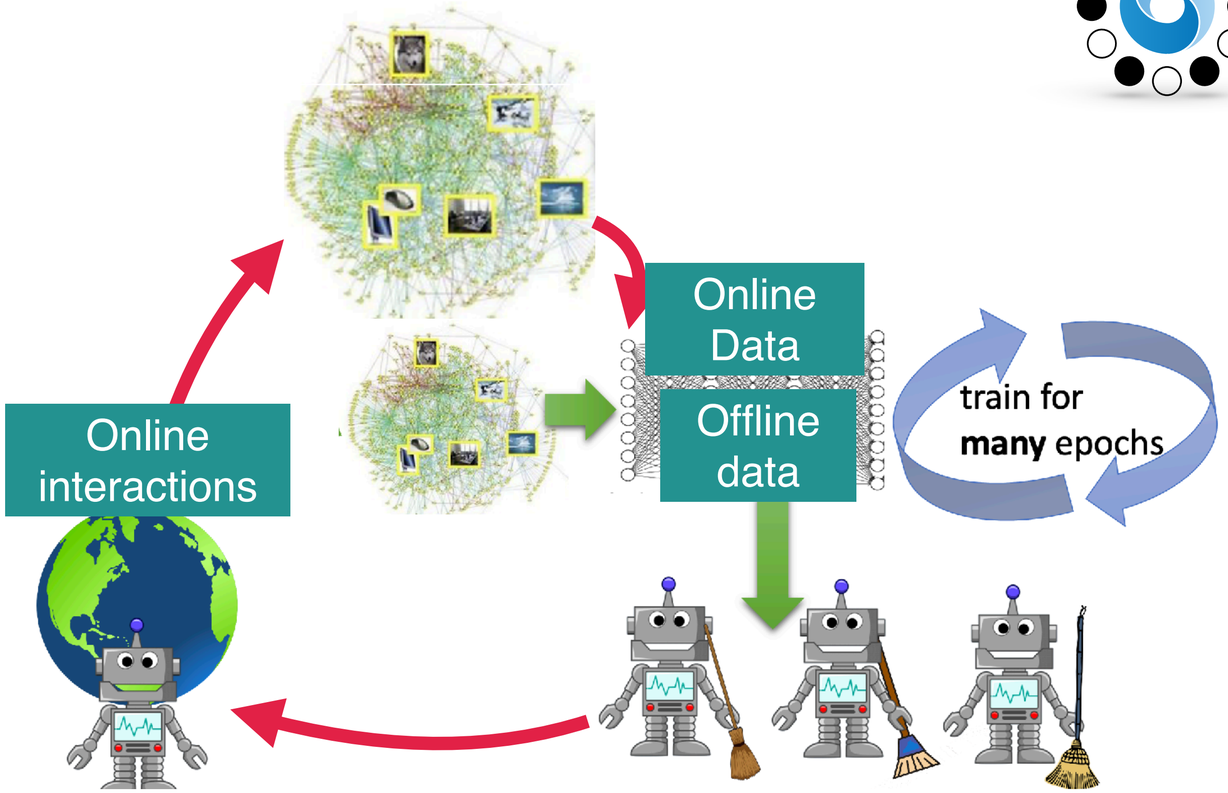
To the rescue: Offline data + Online interaction



Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]
  - **Robotics**: [Vecerik et al., 2017]

To the rescue: Offline data + Online interaction

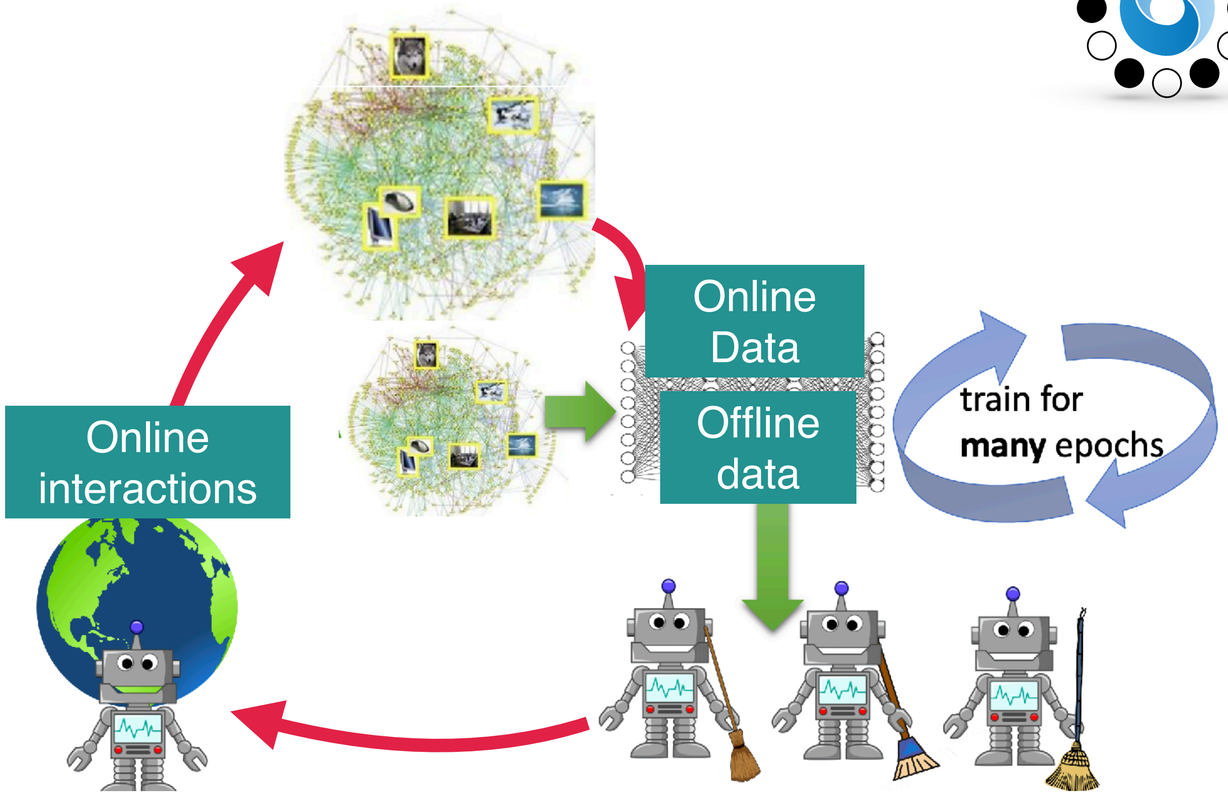


AlphaGo

Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]
  - **Robotics**: [Vecerik et al., 2017]

To the rescue: Offline data + Online interaction

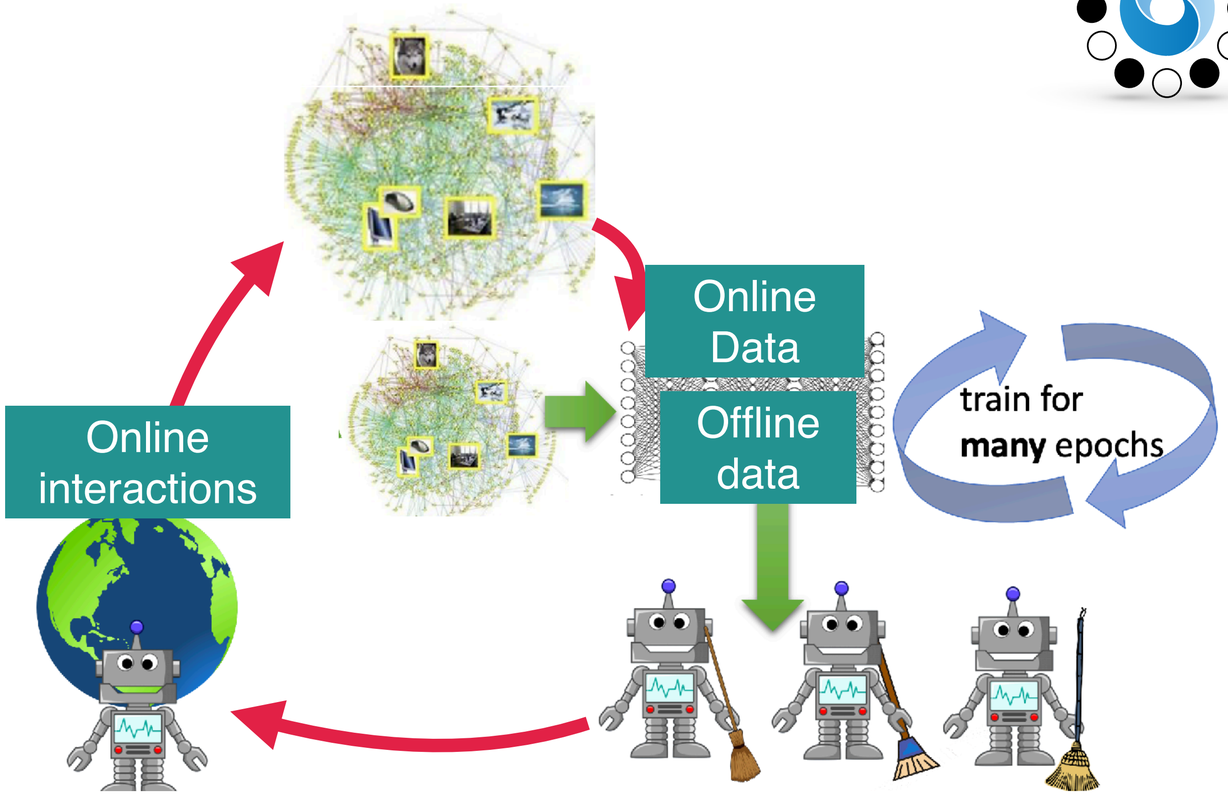


AlphaGo

Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]
  - **Robotics**: [Vecerik et al., 2017]
  - **NLP**: ChatGPT.

To the rescue: Offline data + Online interaction

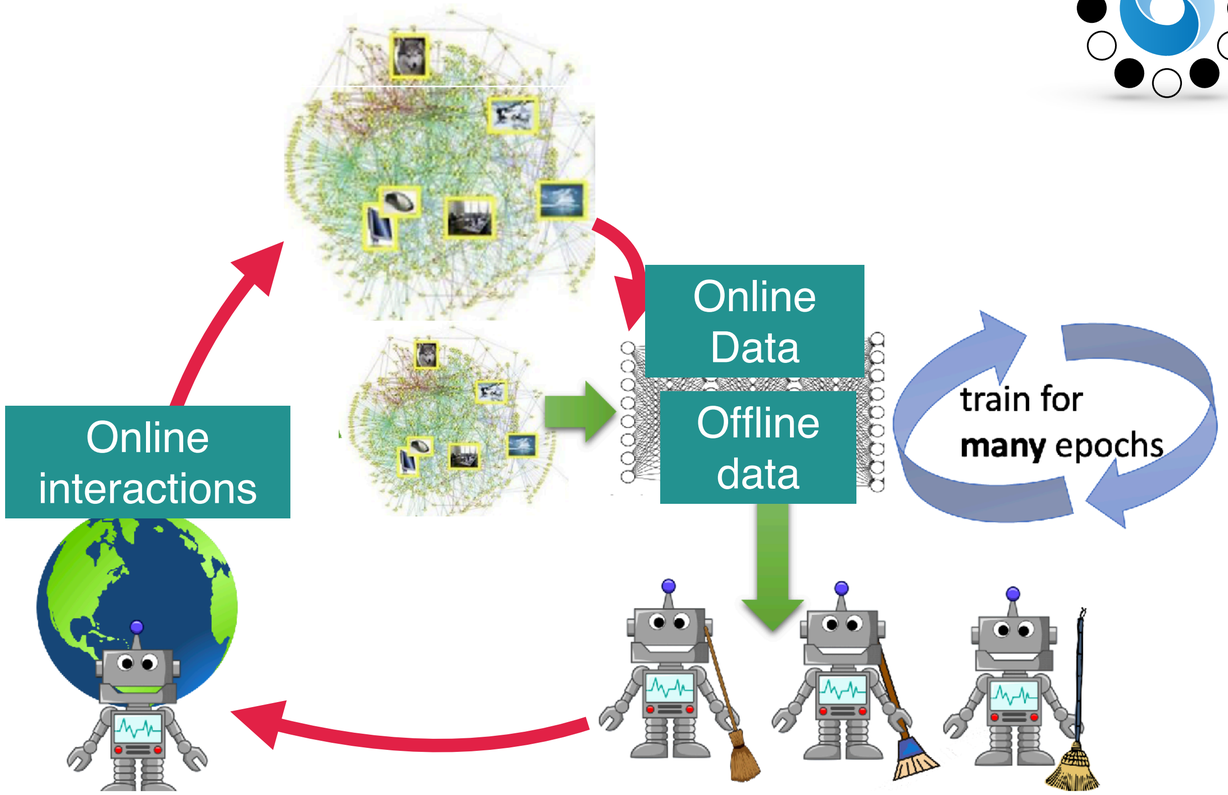


AlphaGo

Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]
  - **Robotics**: [Vecerik et al., 2017]
  - **NLP**: ChatGPT.
- On the theory side:

To the rescue: Offline data + Online interaction

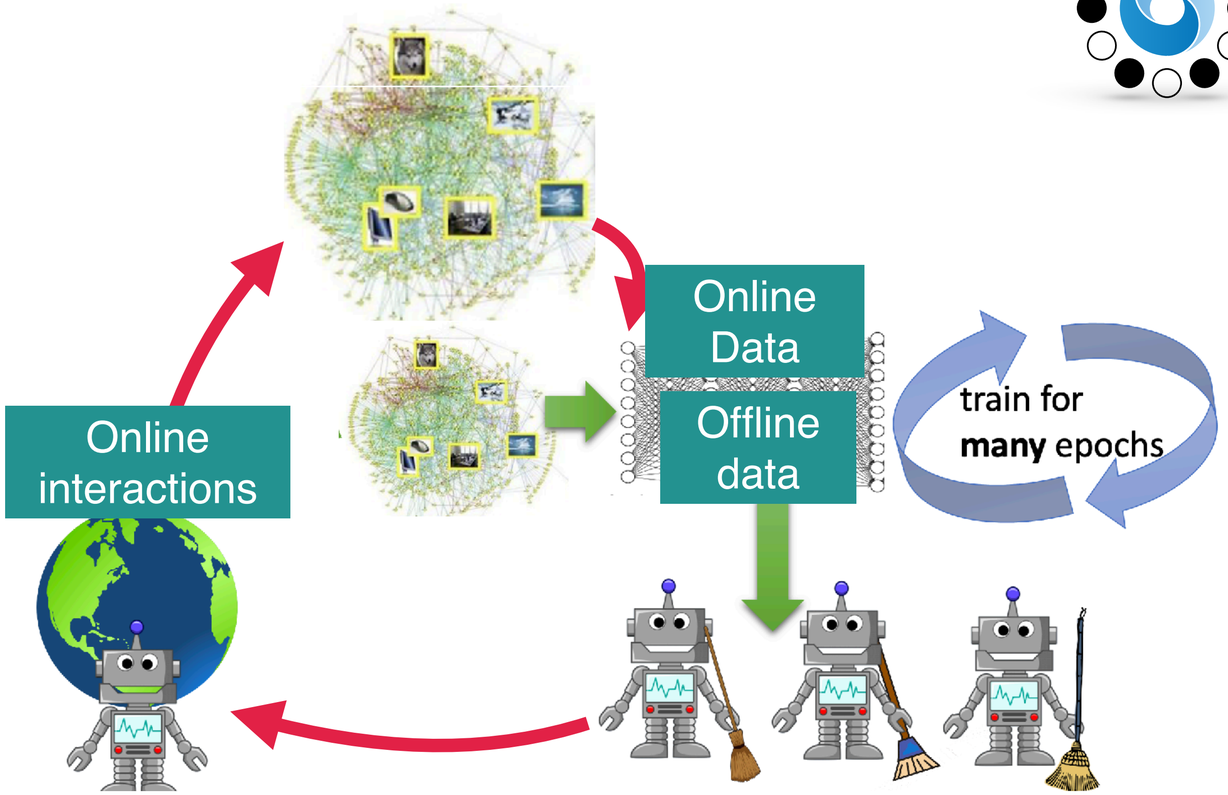


AlphaGo

Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]
  - **Robotics**: [Vecerik et al., 2017]
  - **NLP**: ChatGPT.
- On the theory side:
  - CPI [Kakade & Langford, 2002], PSDP [Bagnell et al., 2003]: leveraging a reset model.

To the rescue: Offline data + Online interaction

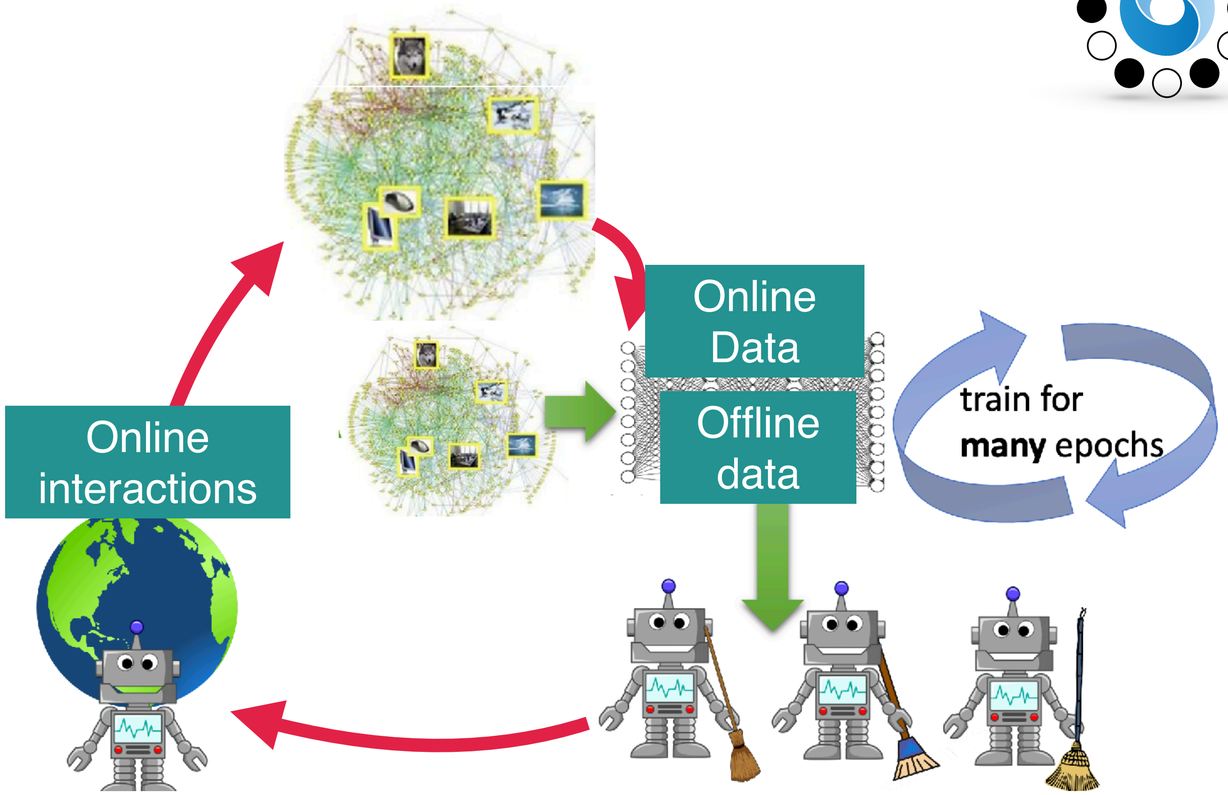


AlphaGo

Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]
  - **Robotics**: [Vecerik et al., 2017]
  - **NLP**: ChatGPT.
- On the theory side:
  - CPI [Kakade & Langford, 2002], PSDP [Bagnell et al., 2003]: leveraging a reset model.
  - Agnostic SysID [Ross & Bagnell., 2012]: model-based method.

To the rescue: Offline data + Online interaction

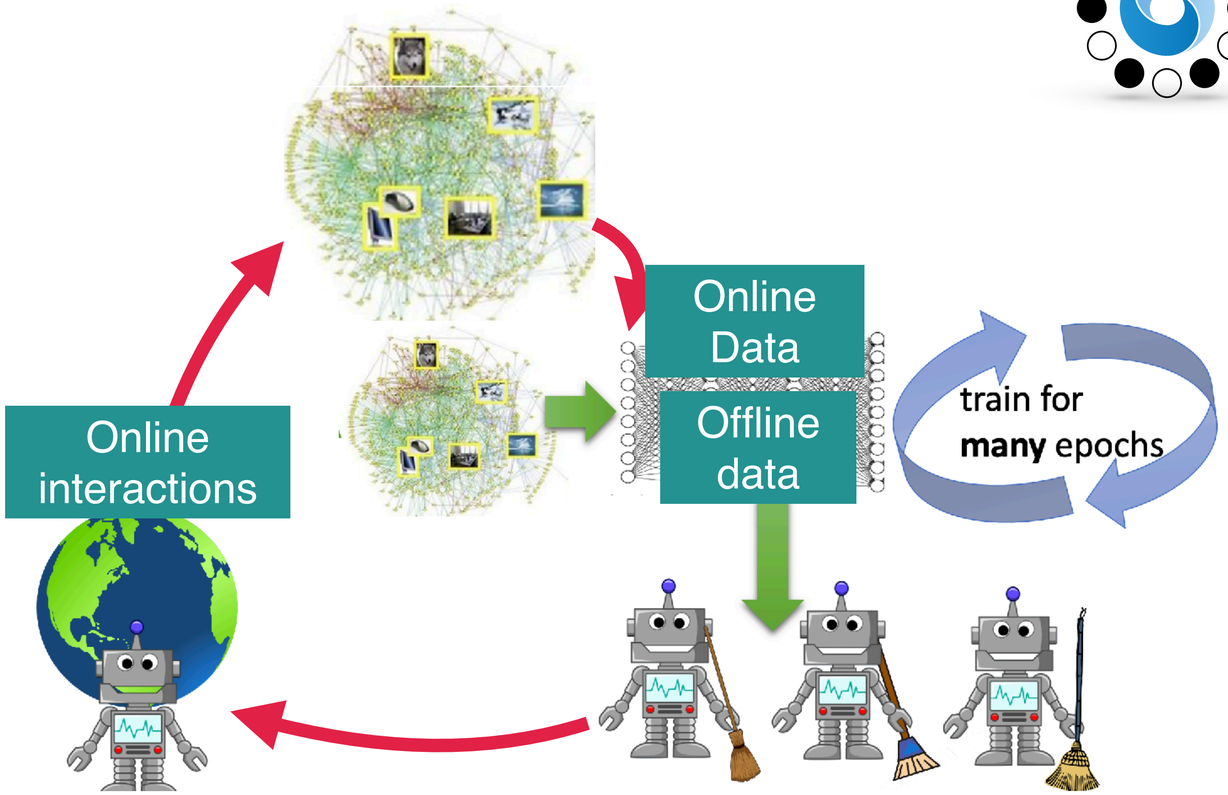


AlphaGo

Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]
  - **Robotics**: [Vecerik et al., 2017]
  - **NLP**: ChatGPT.
- On the theory side:
  - CPI [Kakade & Langford, 2002], PSDP [Bagnell et al., 2003]: leveraging a reset model.
  - Agnostic SysID [Ross & Bagnell., 2012]: model-based method.
  - Policy-finetuning [Xie et al., 2021]: tabular MDPs.

To the rescue: Offline data + Online interaction



Previous works:

- Lots of applications:
  - **Games**: AlphaGo [Silver et al., 2016]
  - **Robotics**: [Vecerik et al., 2017]
  - **NLP**: ChatGPT.
- On the theory side:
  - CPI [Kakade & Langford, 2002], PSDP [Bagnell et al., 2003]: leveraging a reset model.
  - Agnostic SysID [Ross & Bagnell., 2012]: model-based method.
  - Policy-finetuning [Xie et al., 2021]: tabular MDPs.

Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." *nature* 529.7587 (2016): 484-489
Vecerik, Mel, et al. "Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards." (2017).
Kakade, Sham, and John Langford. "Approximately optimal approximate reinforcement learning." *Proceedings of the Nineteenth International Conference on Machine Learning*. 2002.
Bagnell, James, et al. "Policy search by dynamic programming." *Advances in neural information processing systems* 16 (2003).
Ross, Stephane, and J. Andrew Bagnell. "Agnostic system identification for model-based reinforcement learning." *International Conference on Machine Learning*. 2012.
Xie, Tengyang, et al. "Policy finetuning: Bridging sample-efficient offline and online reinforcement learning." *Advances in neural information processing systems* 34 (2021): 27395-27407.

Hybrid RL: towards Efficient Model-free RL with Rich FA

Hybrid RL

Efficient if: reset/ model-based/ demonstration.

Structure

Hybrid RL: towards Efficient Model-free RL with Rich FA

Hybrid RL

Structure

Efficient if: reset/ model-based/ demonstration.

Hybrid RL: towards Efficient Model-free RL with Rich FA

Hybrid RL

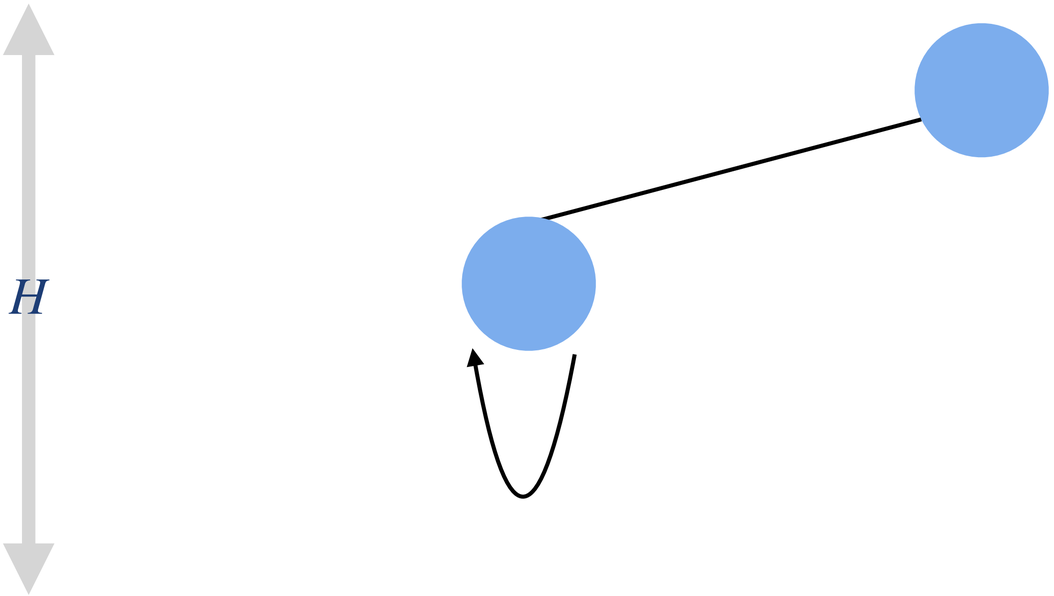
Structure

Efficient model-free RL with Rich FA.

The ability to recover

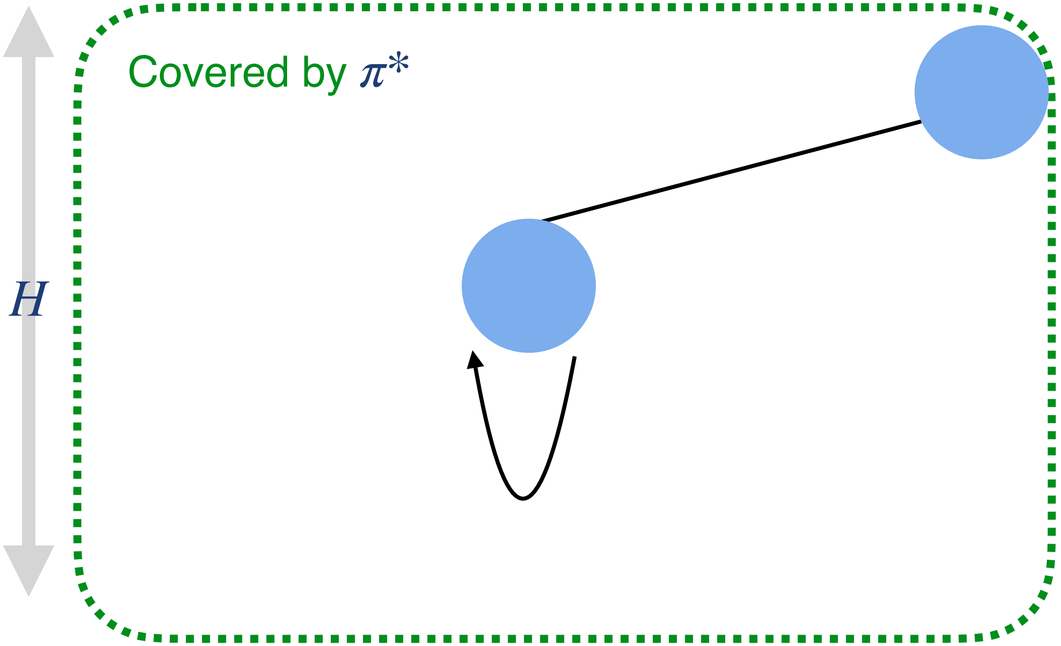
Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

The ability to recover



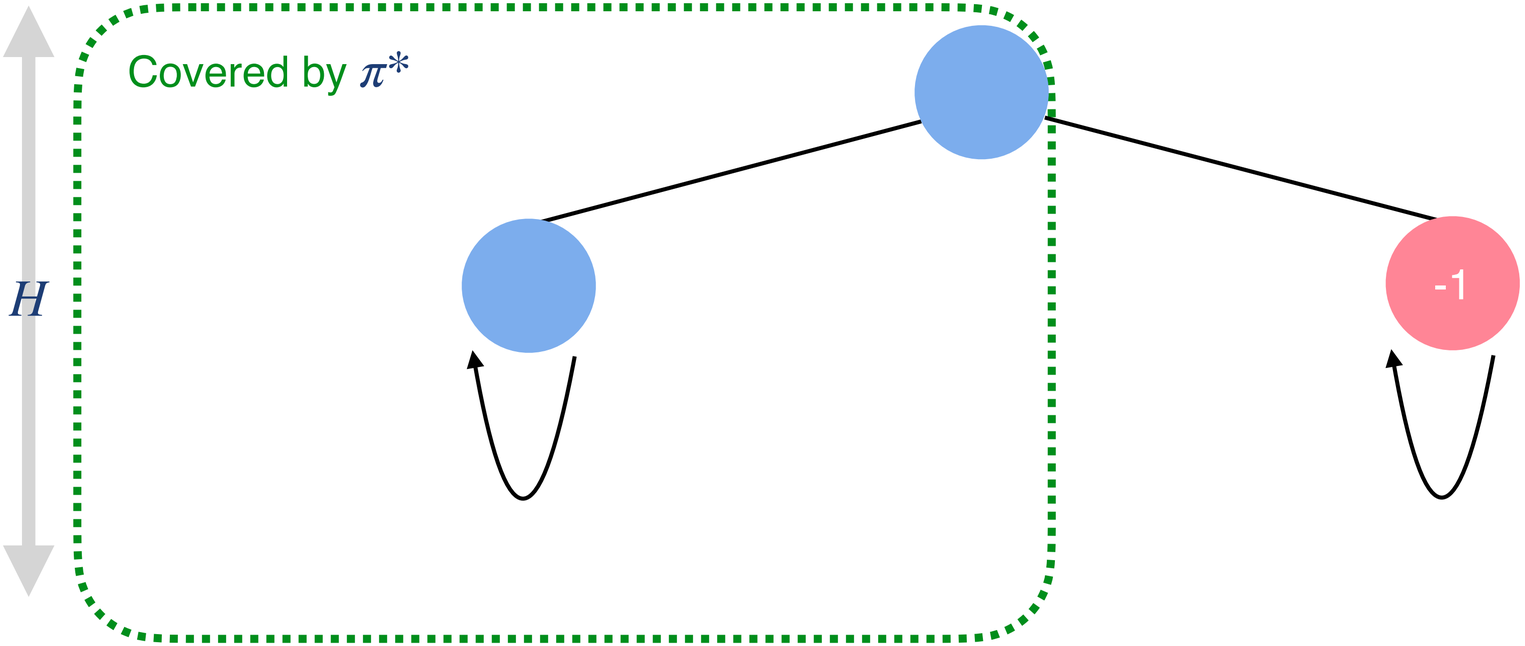
Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

The ability to recover



Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

The ability to recover



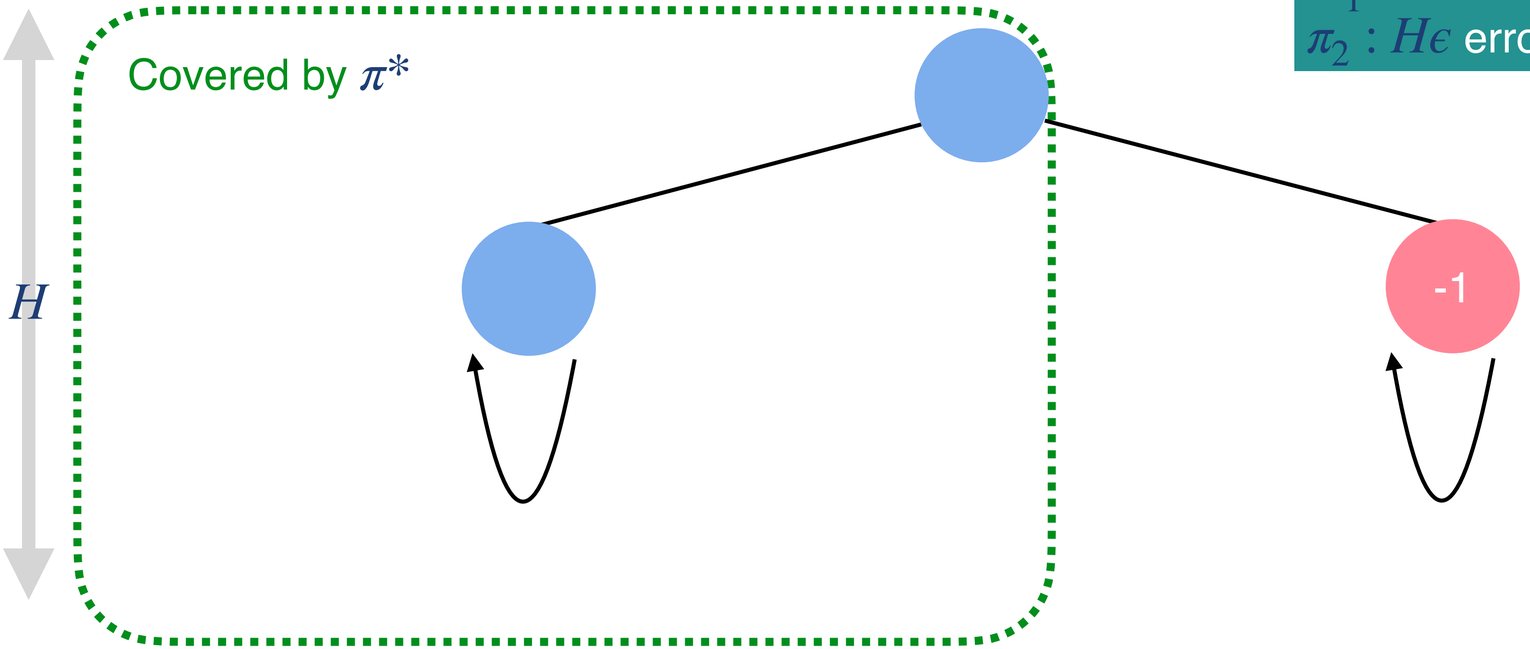
Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

The ability to recover

Policy class with average error ϵ

π_1 : $H\epsilon$ error at $h = 1$, takes left action at $h > 1$

π_2 : $H\epsilon$ error at $h = 1$, takes right action at $h > 1$



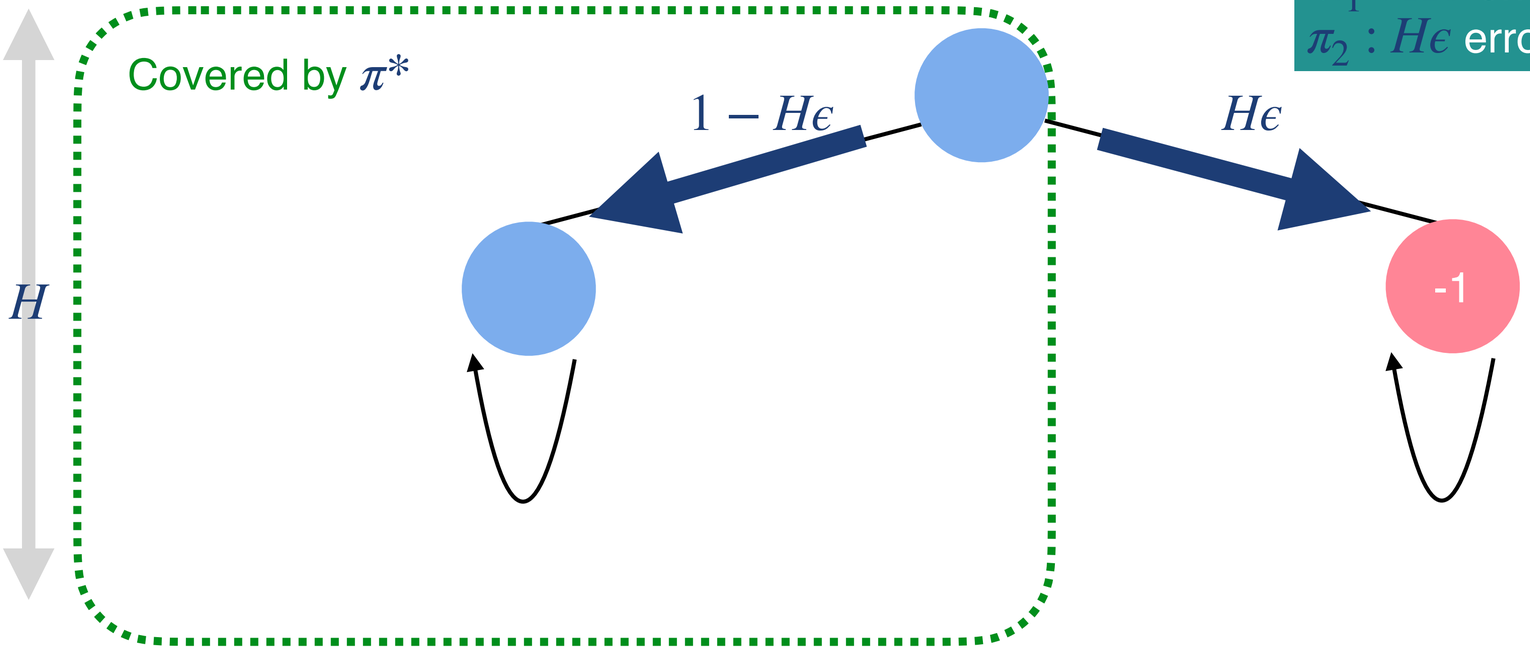
Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

The ability to recover

Policy class with average error ϵ

π_1 : $H\epsilon$ error at $h = 1$, takes left action at $h > 1$

π_2 : $H\epsilon$ error at $h = 1$, takes right action at $h > 1$



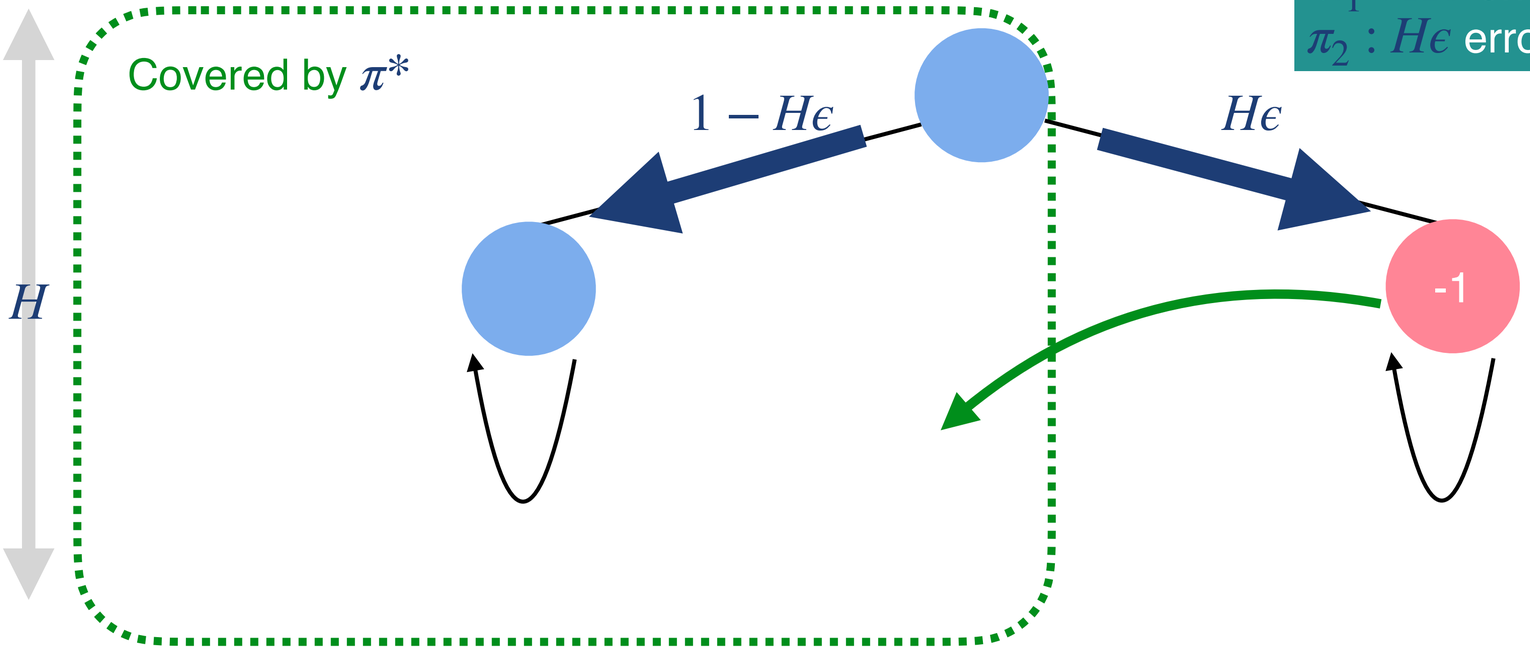
Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

The ability to recover

Policy class with average error ϵ

π_1 : $H\epsilon$ error at $h = 1$, takes left action at $h > 1$

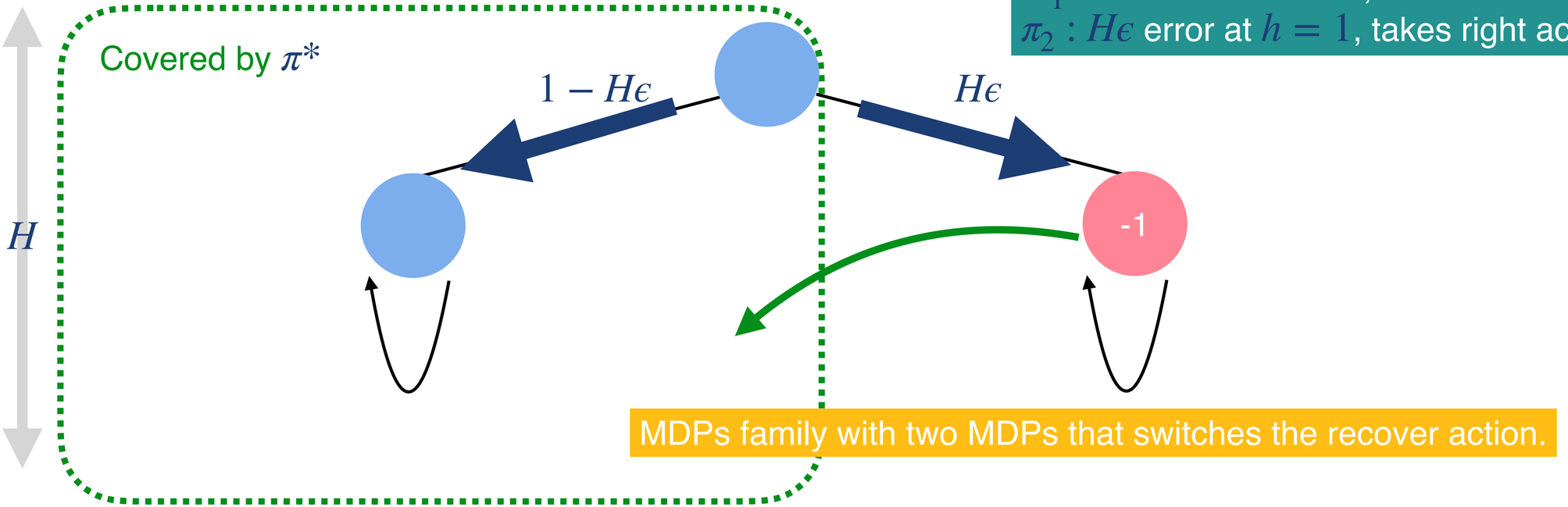
π_2 : $H\epsilon$ error at $h = 1$, takes right action at $h > 1$



Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

The ability to recover

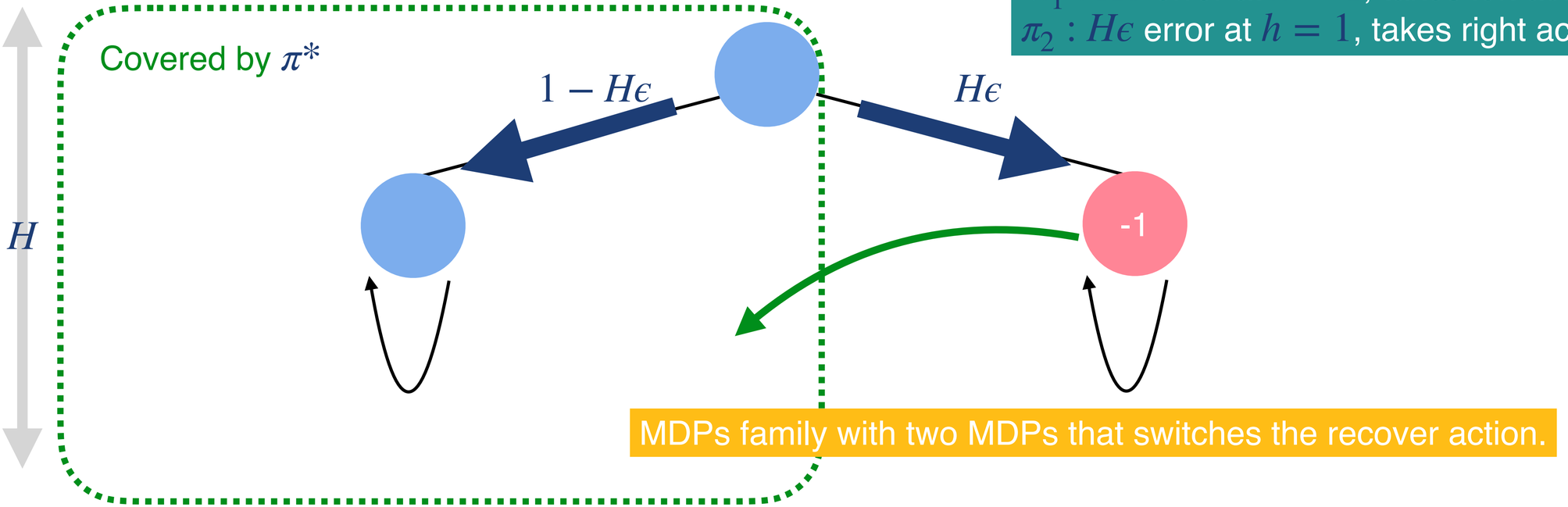
Policy class with average error ϵ
π_1 : $H\epsilon$ error at $h = 1$, takes left action at $h > 1$
π_2 : $H\epsilon$ error at $h = 1$, takes right action at $h > 1$



Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

The ability to recover

Policy class with average error ϵ
π_1 : $H\epsilon$ error at $h = 1$, takes left action at $h > 1$
π_2 : $H\epsilon$ error at $h = 1$, takes right action at $h > 1$



- With misspecification level ϵ, offline RL can not achieve sub-optimality better than $H^2\epsilon$ with probability 1/2.
- With recoverability, hybrid RL results in sub-optimality $H\epsilon$.
- Resembles DAgger [Ross et al., 2011] vs. Behavior Cloning.

Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

Setting

Episodic finite horizon MDPs:

- $\mathscr{M} = \{\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^{H}, R, d_0\}$
- Can only reset from the initial state distribution d_0.
- $V^{\pi} := \mathbb{E}\left[r(s_0, a_0) + r(s_1, a_1) + r(s_2, a_2) + \ldots + r(s_{H-1}, a_{H-1}) \,|\, a \sim \pi, P\right]$

Setting

Episodic finite horizon MDPs:

- $\mathscr{M} = \{\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^{H}, R, d_0\}$
- Can only reset from the initial state distribution d_0.
- $V^{\pi} := \mathbb{E}\left[r(s_0, a_0) + r(s_1, a_1) + r(s_2, a_2) + \ldots + r(s_{H-1}, a_{H-1}) \,|\, a \sim \pi, P\right]$

Access to an offline distribution:

- Offline data is sampled from offline distributions $\nu_0, \ldots, \nu_{H-1}$.
- $\mathscr{D}_{off;h} = \{s, a, r, s'\}_{i=1}^{m_{off}}$, where $s, a \sim \nu_h, s' \sim P(s, a)$.
- **We assume offline distributions "cover" some high quality policy's traces.**

Setting

Episodic finite horizon MDPs:

- $\mathscr{M} = \{\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^{H}, R, d_0\}$
- Can only reset from the initial state distribution d_0.
- $V^{\pi} := \mathbb{E}\left[r(s_0, a_0) + r(s_1, a_1) + r(s_2, a_2) + \ldots + r(s_{H-1}, a_{H-1}) \,|\, a \sim \pi, P\right]$

Access to an offline distribution:

- Offline data is sampled from offline distributions $\nu_0, \ldots, \nu_{H-1}$.
- $\mathscr{D}_{off;h} = \{s, a, r, s'\}_{i=1}^{m_{off}}$, where $s, a \sim \nu_h, s' \sim P(s, a)$.
- **We assume offline distributions "cover" some high quality policy's traces.**

Function approximation:

- Function class $\mathscr{F}_h : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, $\forall h$
- Realizability: $Q_h^{\star} \in \mathscr{F}_h$

Setting

Episodic finite horizon MDPs:

- $\mathscr{M} = \{\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^{H}, R, d_0\}$
- Can only reset from the initial state distribution d_0.
- $V^{\pi} := \mathbb{E}\left[r(s_0, a_0) + r(s_1, a_1) + r(s_2, a_2) + \ldots + r(s_{H-1}, a_{H-1}) \,|\, a \sim \pi, P\right]$

Access to an offline distribution:

- Offline data is sampled from offline distributions $\nu_0, \ldots, \nu_{H-1}$.
- $\mathscr{D}_{off;h} = \{s, a, r, s'\}_{i=1}^{m_{off}}$, where $s, a \sim \nu_h, s' \sim P(s, a)$.
- **We assume offline distributions "cover" some high quality policy's traces.**

Function approximation:

- Function class $\mathscr{F}_h : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, $\forall h$
- Realizability: $Q_h^{\star} \in \mathscr{F}_h$

Learning Goal:

Finding a policy that is ***at least as good as the best one*** covered by offline data.

Algorithm: Hybrid Q-Iteration (HyQ)

In high level, HyQ iteratively runs FQI [Gordon, 1999; Munos & Szepesvari, 2008] on **combination of offline and online data**

Gordon, Geoffrey J. "Stable fitted reinforcement learning." *Advances in neural information processing systems* 8 (1995).
Munos, Rémi, and Csaba Szepesvári. "Finite-Time Bounds for Fitted Value Iteration." *Journal of Machine Learning Research* 9.5 (2008).

Algorithm: Hybrid Q-Iteration (HyQ)

In high level, HyQ iteratively runs FQI [Gordon, 1999; Munos & Szepesvari, 2008] on **combination of offline and online data**

At the t-th iteration:

1. For every h, collect **online** data w/ π^t:

$$s_h, a_h \sim \pi^t, r_h, s_{h+1} \sim P(\,.\,|\,s_h, a_h)$$

2. Online data aggregation:

$$\mathscr{D}_{on;h} = \mathscr{D}_{on;h} + \{s_h, a_h, r_h, s_{h+1}\}, \forall h$$

3. Run Fitted Q-Iteration on the combined **online+offline** dataset:

$$\pi^{t+1} \leftarrow \text{FQI}\left(\{\mathscr{D}_{\text{off;h}} + \mathscr{D}_{\text{on;h}}\}_{\forall \text{h}}\right)$$

Gordon, Geoffrey J. "Stable fitted reinforcement learning." *Advances in neural information processing systems* 8 (1995).
Munos, Rémi, and Csaba Szepesvári. "Finite-Time Bounds for Fitted Value Iteration." *Journal of Machine Learning Research* 9.5 (2008).

Algorithm: Hybrid Q-Iteration (HyQ)

In high level, HyQ iteratively runs FQI [Gordon, 1999; Munos & Szepesvari, 2008] on **combination of offline and online data**

At the t-th iteration:

1. For every h, collect **online** data w/ π^t:

$$s_h, a_h \sim \pi^t, r_h, s_{h+1} \sim P(\,.\,|\,s_h, a_h)$$

2. Online data aggregation:

$$\mathcal{D}_{on;h} = \mathcal{D}_{on;h} + \{s_h, a_h, r_h, s_{h+1}\}, \forall h$$

3. Run Fitted Q-Iteration on the combined **online+offline** dataset:

$$\pi^{t+1} \leftarrow \text{FQI}\left(\{\mathcal{D}_{\text{off;h}} + \mathcal{D}_{\text{on;h}}\}_{\forall \text{h}}\right)$$

Gordon, Geoffrey J. "Stable fitted reinforcement learning." *Advances in neural information processing systems* 8 (1995).
Munos, Rémi, and Csaba Szepesvári. "Finite-Time Bounds for Fitted Value Iteration." *Journal of Machine Learning Research* 9.5 (2008).

Algorithm: Hybrid Q-Iteration (HyQ)

In high level, HyQ iteratively runs FQI [Gordon, 1999; Munos & Szepesvari, 2008] on **combination of offline and online data**

At the t-th iteration:

1. For every h, collect **online** data w/ π^t:

$$s_h, a_h \sim \pi^t, r_h, s_{h+1} \sim P(\,.\,|\,s_h, a_h)$$

2. Online data aggregation:

$$\mathscr{D}_{on;h} = \mathscr{D}_{on;h} + \{s_h, a_h, r_h, s_{h+1}\}, \forall h$$

3. Run Fitted Q-Iteration on the combined **online+offline** dataset:

$$\pi^{t+1} \leftarrow \text{FQI}\left(\{\mathscr{D}_{\text{off;h}} + \mathscr{D}_{\text{on;h}}\}_{\forall \text{h}}\right)$$

FQI:

("DP" using regression on the given data)

1. At last step $H-1$:

$$\hat{f}_{H-1} := \arg\min_{f \in \mathscr{F}_{H-1}} \mathbb{E}_{\mathscr{D}_{H-1}} (f(s,a) - r)^2$$

2. For every h from $H-2$ to 0:

$$\hat{f}_h := \arg\min_{f \in \mathscr{F}_h} \mathbb{E}_{\mathscr{D}_h} \left(f(s,a) - (r + \max_{a'} \hat{f}_{h+1}(s', a')) \right)^2$$

3. Extract policy:

$$\pi_h(s) = \arg\max_a \hat{f}_h(s,a)$$

Gordon, Geoffrey J. "Stable fitted reinforcement learning." *Advances in neural information processing systems* 8 (1995).
Munos, Rémi, and Csaba Szepesvári. "Finite-Time Bounds for Fitted Value Iteration." *Journal of Machine Learning Research* 9.5 (2008).

Theory results

Assumptions

1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a') \in \mathcal{F}_h$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right| = \left| \langle X_h(g), W_h(f) \rangle \right|$$

Theory results

Assumptions

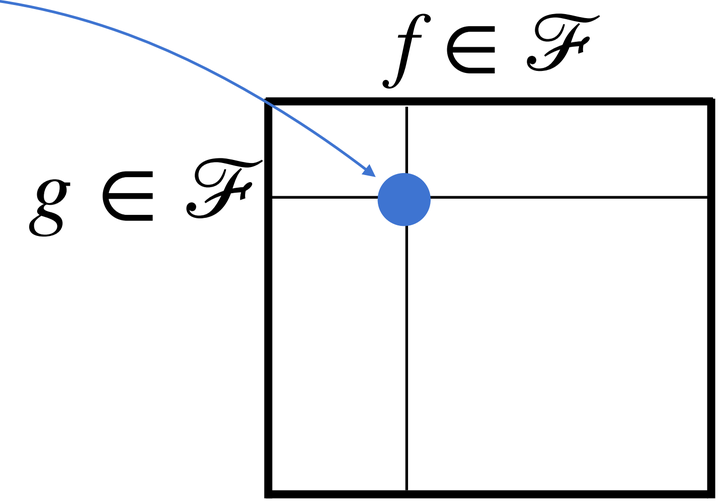
1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a') \in \mathcal{F}_h$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right| = \left| \langle X_h(g), W_h(f) \rangle \right|$$

$$\left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right|$$

($\pi_g(s) = \arg\max_a g(s)$ is the greedy policy of g)



Theory results

Assumptions

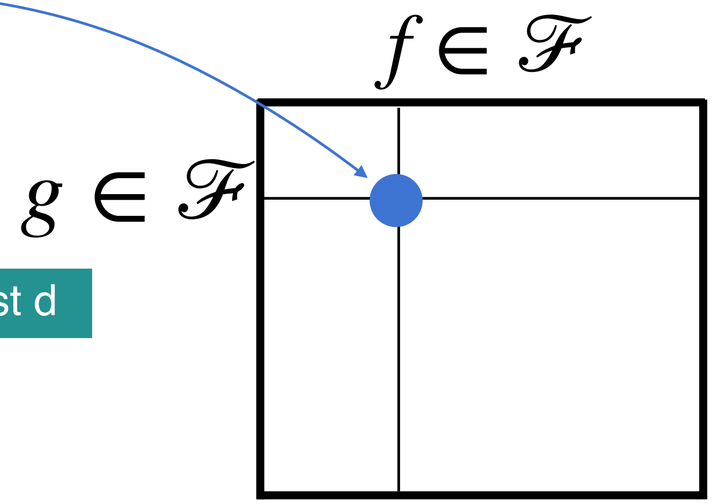
1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a') \in \mathcal{F}_h$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right| = \left| \langle X_h(g), W_h(f) \rangle \right|$$

$$\left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right|$$

($\pi_g(s) = \arg\max_a g(s)$ is the greedy policy of g)



Bilinear rank d means this matrix has rank at most d

Theory results

Assumptions

1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a') \in \mathcal{F}_h$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right| = \left| \langle X_h(g), W_h(f) \rangle \right|$$

Theory results

Assumptions

1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a') \in \mathcal{F}_h$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right| = \left| \langle X_h(g), W_h(f) \rangle \right|$$

Coverage of offline distribution

$$C_\pi = \max \left\{ 0, \max_{f \in \mathcal{F}} \frac{\sum_h \mathbb{E}_{s_h, a_h \sim \pi} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)}{\sqrt{\sum_h \mathbb{E}_{s_h, a_h \sim \nu_h} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)^2}} \right\}$$

Theory results

Assumptions

1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a') \in \mathcal{F}_h$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right| = \left| \langle X_h(g), W_h(f) \rangle \right|$$

Coverage of offline distribution

Average signed Bellman-error of f under π

$$C_\pi = \max \left\{ 0, \max_{f \in \mathcal{F}} \frac{\sum_h \mathbb{E}_{s_h, a_h \sim \pi} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)}{\sqrt{\sum_h \mathbb{E}_{s_h, a_h \sim \nu_h} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)^2}} \right\}$$

Theory results

Assumptions

1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a') \in \mathcal{F}_h$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right| = \left| \langle X_h(g), W_h(f) \rangle \right|$$

Coverage of offline distribution

Average signed Bellman-error of f under π

$$C_\pi = \max \left\{ 0, \max_{f \in \mathcal{F}} \frac{\sum_h \mathbb{E}_{s_h, a_h \sim \pi} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)}{\sqrt{\sum_h \mathbb{E}_{s_h, a_h \sim \nu_h} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)^2}} \right\}$$

Average squared Bellman-error of f under offline distribution

Theory results

Assumptions

1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a') \in \mathcal{F}_h$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right| = \left| \langle X_h(g), W_h(f) \rangle \right|$$

Coverage of offline distribution

$$C_\pi = \max \left\{ 0, \max_{f \in \mathcal{F}} \frac{\sum_h \mathbb{E}_{s_h, a_h \sim \pi} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)}{\sqrt{\sum_h \mathbb{E}_{s_h, a_h \sim \nu_h} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)^2}} \right\}$$

Theory results

Assumptions

1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a') \in \mathcal{F}_h$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h, a_h \sim \pi_g} (f_h(s_h, a_h) - \mathcal{T} f_{h+1}(s_h, a_h)) \right| = \left| \langle X_h(g), W_h(f) \rangle \right|$$

Coverage of offline distribution

$$C_\pi = \max \left\{ 0, \max_{f \in \mathcal{F}} \frac{\sum_h \mathbb{E}_{s_h, a_h \sim \pi} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)}{\sqrt{\sum_h \mathbb{E}_{s_h, a_h \sim \nu_h} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h, a_h) \right)^2}} \right\}$$

Smaller than prior definition of coverage:

Tabular: $C_\pi \leq \max_{s,a,h} \frac{d_h^\pi(s,a)}{\nu_h(s,a)}$

Linear: $C_\pi \leq \max_{x,h} \frac{x^\top \mathbb{E}_{s,a \sim \pi} \phi(s,a) \phi(s,a)^\top x}{x^\top \mathbb{E}_{s,a \sim \nu_h} \phi(s,a) \phi(s,a)^\top x}$

https://twitter.com/nanjiang_cs/status/1666994607073230848

Theory results

Assumptions

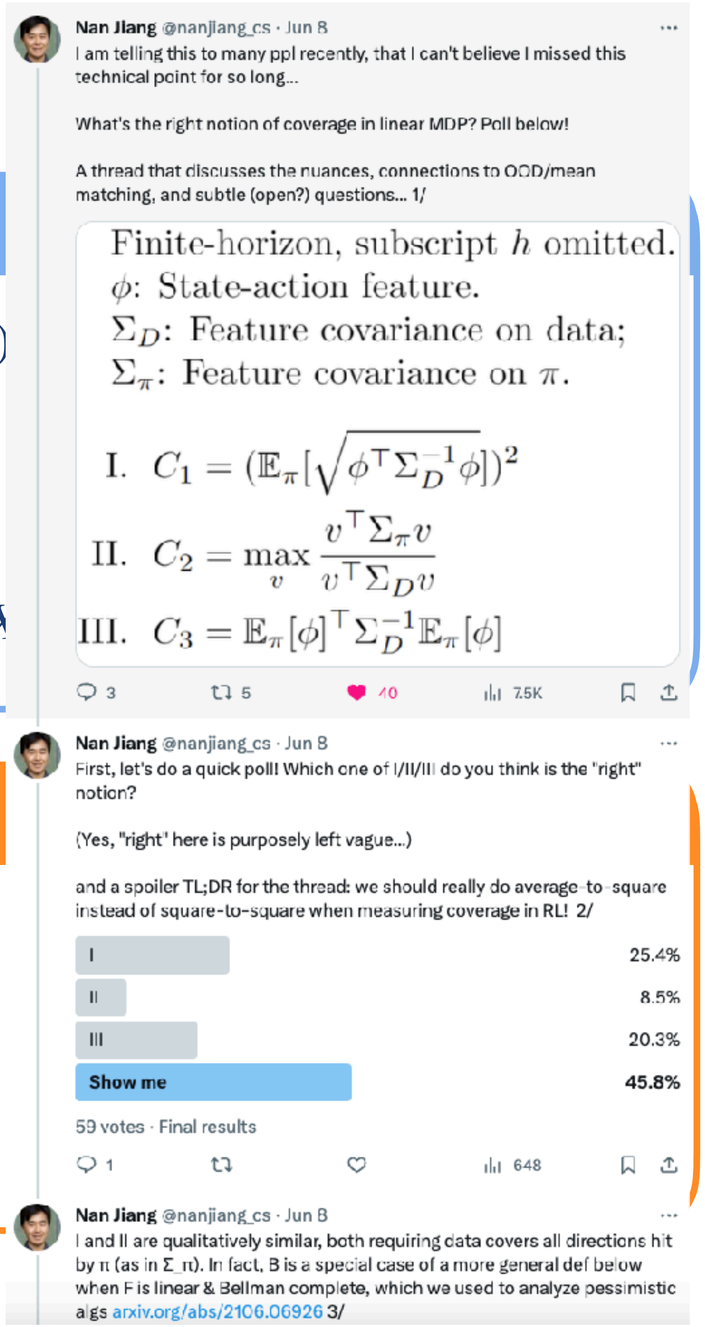
1. **Bellman Completeness**: $\forall f \in \mathcal{F}_{h+1} : r(s,a) + \mathbb{E}_{s' \sim P(s,a)} \max_{a'} f(s',a')$

2. **Low Bilinear Rank**: there exist $X_h, W_h : \mathcal{F} \to \mathbb{R}^d, \forall h$ such that:

$$\forall f, g \in \mathcal{F} : \left| \mathbb{E}_{s_h,a_h \sim \pi_g} (f_h(s_h,a_h) - \mathcal{T} f_{h+1}(s_h,a_h)) \right| = \left| \langle X$$

Coverage of offline distribution

$$C_\pi = \max \left\{ 0, \max_{f \in \mathcal{F}} \frac{\sum_h \mathbb{E}_{s_h,a_h \sim \pi} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h,a_h) \right)}{\sqrt{\sum_h \mathbb{E}_{s_h,a_h \sim \nu_h} \left(\mathcal{T} f_{h+1}(s,a) - f_h(s_h,a_h) \right)^2}} \right\}$$



Nan Jiang @nanjiang_cs · Jun 8

I am telling this to many ppl recently, that I can't believe I missed this technical point for so long...

What's the right notion of coverage in linear MDP? Poll below!

A thread that discusses the nuances, connections to OOD/mean matching, and subtle (open?) questions... 1/

Finite-horizon, subscript h omitted.
ϕ: State-action feature.
Σ_D: Feature covariance on data;
Σ_π: Feature covariance on π.

I. $C_1 = (\mathbb{E}_\pi[\sqrt{\phi^\top \Sigma_D^{-1} \phi}])^2$

II. $C_2 = \max_v \frac{v^\top \Sigma_\pi v}{v^\top \Sigma_D v}$

III. $C_3 = \mathbb{E}_\pi[\phi]^\top \Sigma_D^{-1} \mathbb{E}_\pi[\phi]$

3 · 5 · 40 · 7.5K

Nan Jiang @nanjiang_cs · Jun 8

First, let's do a quick poll! Which one of I/II/III do you think is the "right" notion?

(Yes, "right" here is purposely left vague...)

and a spoiler TL;DR for the thread: we should really do average-to-square instead of square-to-square when measuring coverage in RL! 2/

I 25.4%
II 8.5%
III 20.3%
Show me 45.8%

59 votes · Final results

1 · 648

Nan Jiang @nanjiang_cs · Jun 8

I and II are qualitatively similar, both requiring data covers all directions hit by π (as in Σ_π). In fact, B is a special case of a more general def below when F is linear & Bellman complete, which we used to analyze pessimistic algs arxiv.org/abs/2106.06926 3/

Theory results

Theorem

Run HyQ for T iterations, with probability $1-\delta$, we have that for any comparator π^e,

$$\sum_{t=1}^{T}(V^{\pi^e}-V^{\pi^t}) \leq \widetilde{O}\left(\max\{1,C_{\pi^e}\}H^2\sqrt{dT}\cdot\sqrt{\ln(|\mathcal{F}|/\delta)}\right)$$

Theory results

Theorem

Run HyQ for T iterations, with probability $1-\delta$, we have that for any comparator π^e,

$$\sum_{t=1}^{T}(V^{\pi^e}-V^{\pi^t}) \leq \widetilde{O}\left(\max\{1,C_{\pi^e}\}H^2\sqrt{dT}\cdot\sqrt{\ln(|\mathcal{F}|/\delta)}\right)$$

Corollary (informal):

If $\pi^\star$ is covered, i.e., $C_{\pi^\star}<\infty$, then we learn to compete to $\pi^\star$ in poly sample complexity, with **poly number of calls to least square regression oracles.**

Theory results

Theorem

Run HyQ for T iterations, with probability $1-\delta$, we have that for any comparator π^e,

$$\sum_{t=1}^{T}(V^{\pi^e}-V^{\pi^t}) \leq \widetilde{O}\left(\max\{1,C_{\pi^e}\}H^2\sqrt{dT}\cdot\sqrt{\ln(|\mathcal{F}|/\delta)}\right)$$

If offline distribution covers high quality policies, a large family of

RL settings can be solved via Supervised learning w/ least square regression oracles.

Corollary (informal):

If $\pi^\star$ is covered, i.e., $C_{\pi^\star}<\infty$, then we learn to compete to $\pi^\star$ in poly sample complexity, with **poly number of calls to least square regression oracles.**

One-page Proof Sketch

Consider iteration t, and recall π^t is the greedy policy of f^t from FQI, we have

$$V^{\pi^e} - V^{\pi^t} \leq \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim \pi^e} (\mathcal{T} f_{h+1}^t(s,a) - f_h^t(s,a)) + \sum_{h=0}^{H-1} \left| \mathbb{E}_{s_h, a_h \sim \pi^t} (\mathcal{T} f_{h+1}^t(s,a) - f_h^t(s,a)) \right|$$

One-page Proof Sketch

Consider iteration t, and recall π^t is the greedy policy of f^t from FQI, we have

$$V^{\pi^e} - V^{\pi^t} \leq \text{Signed Bellman error of } f^t \text{ under } \pi^e + \sum_{h=0}^{H-1} \left| \mathbb{E}_{s_h, a_h \sim \pi^t} (\mathcal{T} f_{h+1}^t(s,a) - f_h^t(s,a)) \right|$$

One-page Proof Sketch

Consider iteration t, and recall π^t is the greedy policy of f^t from FQI, we have

$$V^{\pi^e} - V^{\pi^t} \leq \text{Signed Bellman error of } f^t \text{ under } \pi^e + \sum_{h=0}^{H-1} \left| \mathbb{E}_{s_h, a_h \sim \pi^t} (\mathcal{T} f_{h+1}^t(s,a) - f_h^t(s,a)) \right|$$

(a) FQI always minimizes Bellman error under offline distribution ν

(b) ν covers $\pi^e \Rightarrow$ this term always is small

One-page Proof Sketch

Consider iteration t, and recall π^t is the greedy policy of f^t from FQI, we have

$$V^{\pi^e} - V^{\pi^t} \leq \text{Signed Bellman error of } f^t \text{ under } \pi^e + \text{Bellman error of } f^t \text{ under } \pi^t$$

(a) FQI always minimizes Bellman error under offline distribution ν

(b) ν covers $\pi^e \Rightarrow$ this term always is small

One-page Proof Sketch

Consider iteration t, and recall π^t is the greedy policy of f^t from FQI, we have

$$V^{\pi^e} - V^{\pi^t} \leq \text{Signed Bellman error of } f^t \text{ under } \pi^e + \text{Bellman error of } f^t \text{ under } \pi^t$$

(a) FQI always minimizes Bellman error under offline distribution ν

(b) ν covers $\pi^e \Rightarrow$ this term always is small

(a) Fact: f^t has small Bellman error under mixture of $\pi^1, \ldots, \pi^{t-1}$

(b) If π^t distribution is covered by $\pi^1, \ldots, \pi^{t-1}$, then we are done!

(c) Otherwise, we just explored; but this cannot happen too many times due to low Bilinear rank.

One-page Proof Sketch

Consider iteration t, and recall π^t is the greedy policy of f^t from FQI, we have

$$V^{\pi^e} - V^{\pi^t} \leq \text{Signed Bellman error of } f^t \text{ under } \pi^e + \text{Bellman error of } f^t \text{ under } \pi^t$$

Has the Explore-or-Terminate flavor but without explicit exploration scheme!

(a) FQI always minimizes Bellman error under offline distribution ν

(b) ν covers $\pi^e \Rightarrow$ this term always is small

(a) Fact: f^t has small Bellman error under mixture of $\pi^1, \ldots, \pi^{t-1}$

(b) If π^t distribution is covered by $\pi^1, \ldots, \pi^{t-1}$, then we are done!

(c) Otherwise, we just explored; but this cannot happen too many times due to low Bilinear rank.

An analogy of the computational benefit

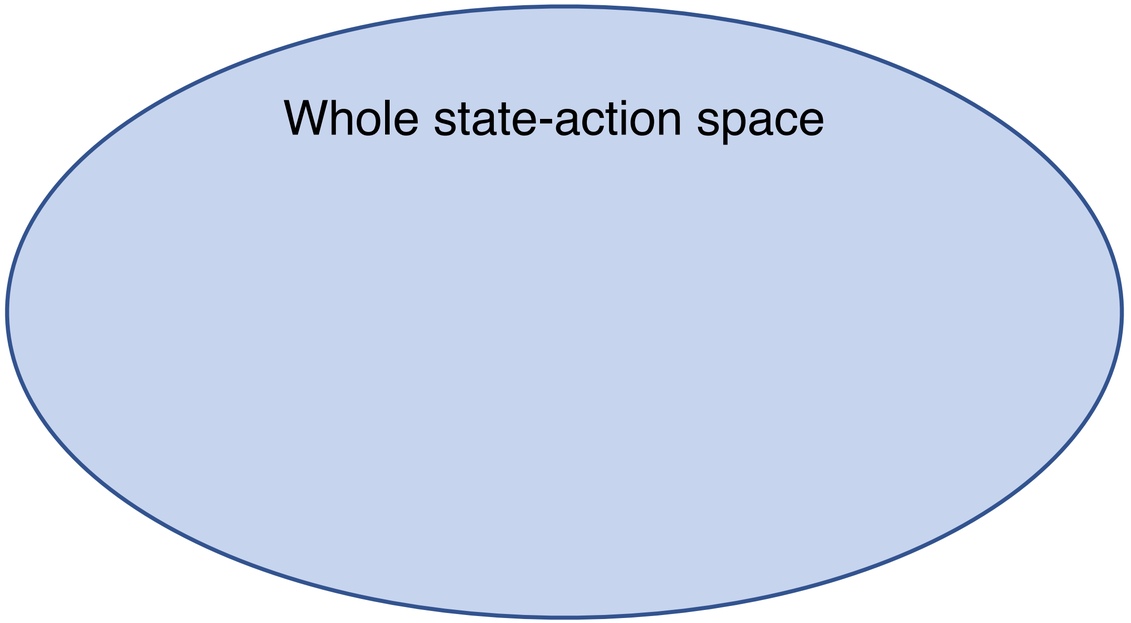
	Offline RL	Hybrid RL
Model-based	**Version space algorithms as min-max games [Uehara & Sun, 2021]:** $\pi = \arg\max_{\pi\in\Pi} \min_{P\in\mathscr{M}} V_P^\pi$ s.t., $\mathbb{E}[\mathrm{MLE}(\mathrm{P}(\cdot \mid \mathrm{s}, \mathrm{a}))] \le \xi$	**Maximum likelihood estimation on the combined offline + online data [Ross & Bagnell, 2012]:** $\mathscr{D} := \mathscr{D}_{off} + \mathscr{D}_{on}$ $\hat{P} = \mathrm{MLE}(\mathscr{D}), \quad \hat{\pi} = \arg\max_{\pi\in\Pi} \mathrm{V}_{\hat{\mathrm{P}}}^\pi$
Model-free	**Version space algorithms as min-max games [Xie et al., 2021]:** $\pi = \arg\max_{\pi\in\Pi} \min_{f\in\mathscr{F}} f(s_0, \pi)$ s.t., $\mathbb{E}[(f(s,a) - \mathscr{T}f(s,a))^2] \le \xi$	**Minimizing Bellman error on the combined offline + online data [This work]:** $\mathscr{D} := \mathscr{D}_{off} + \mathscr{D}_{on}$ $\hat{f} := \arg\min_{f\in\mathscr{F}} \mathbb{E}_{\mathscr{D}} \left(f(s,a) - \mathscr{T}f(s,a)\right)^2,$ $\hat{\pi} = \arg\max_{\pi\in\Pi} \hat{f}(\pi)$

Uehara, Masatoshi, and Wen Sun. "Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage." *International Conference on Learning Representations*. 2021.
Ross, Stephane, and J. Andrew Bagnell. "Agnostic system identification for model-based reinforcement learning." *International Conference on Machine Learning*. 2012.
Xie, Tengyang, et al. "Bellman-consistent pessimism for offline reinforcement learning." *Advances in neural information processing systems* 34 (2021): 6683-6694.

Can we do better?

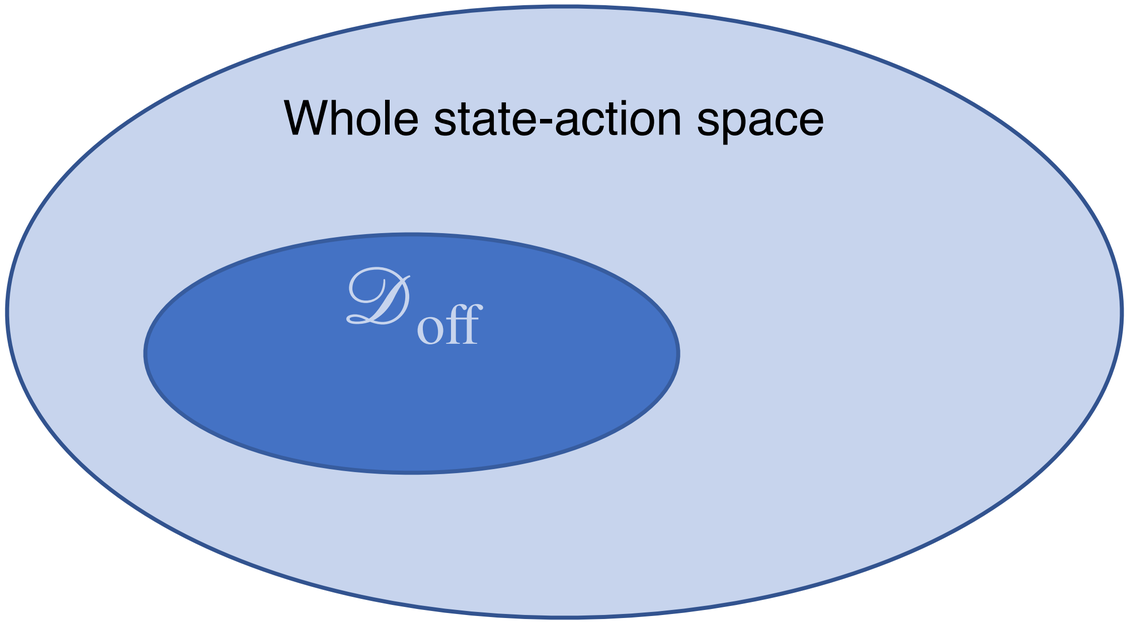
Dann, Chris, et al. "Guarantees for epsilon-greedy reinforcement learning with function approximation." *International conference on machine learning*. PMLR, 2022.

Can we do better?



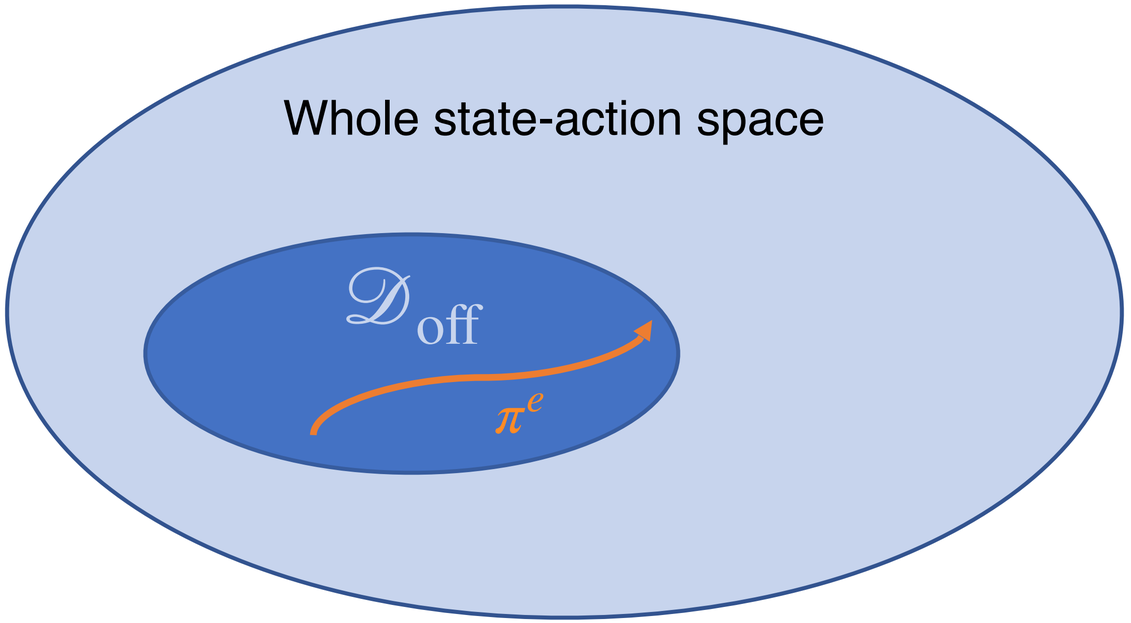
Dann, Chris, et al. "Guarantees for epsilon-greedy reinforcement learning with function approximation." *International conference on machine learning*. PMLR, 2022.

Can we do better?



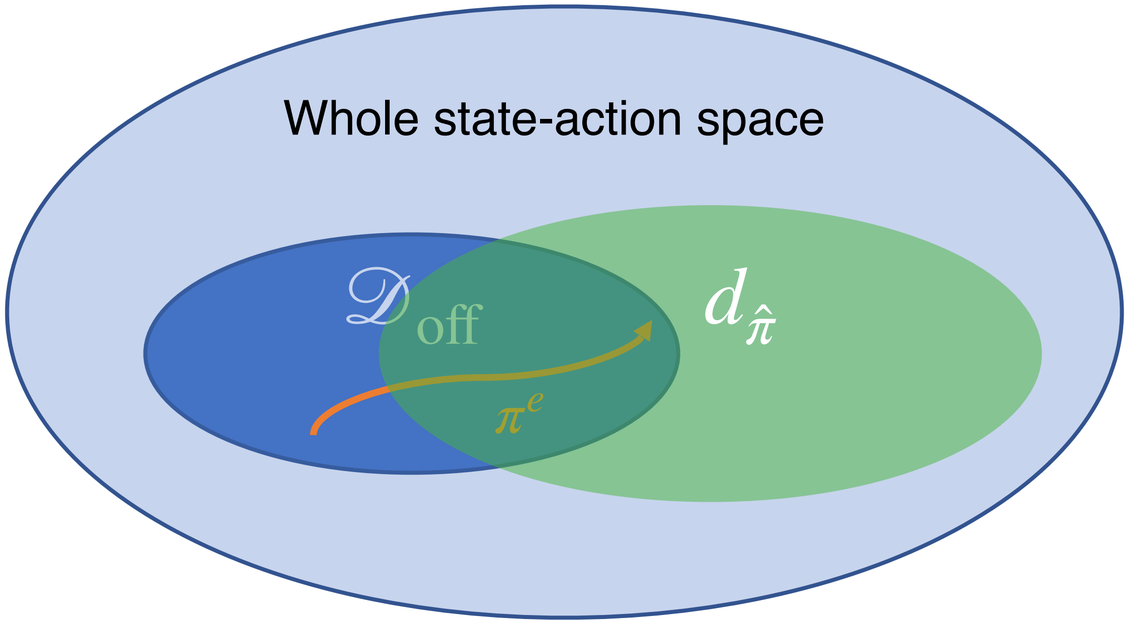
Dann, Chris, et al. "Guarantees for epsilon-greedy reinforcement learning with function approximation." *International conference on machine learning*. PMLR, 2022.

Can we do better?



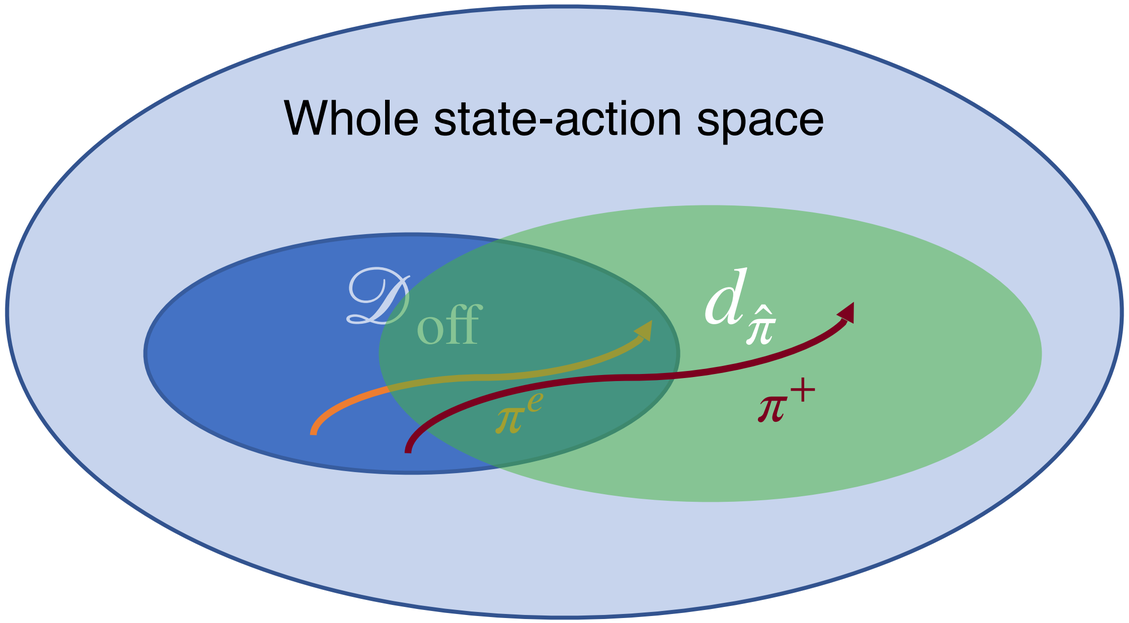
Dann, Chris, et al. "Guarantees for epsilon-greedy reinforcement learning with function approximation." *International conference on machine learning*. PMLR, 2022.

Can we do better?



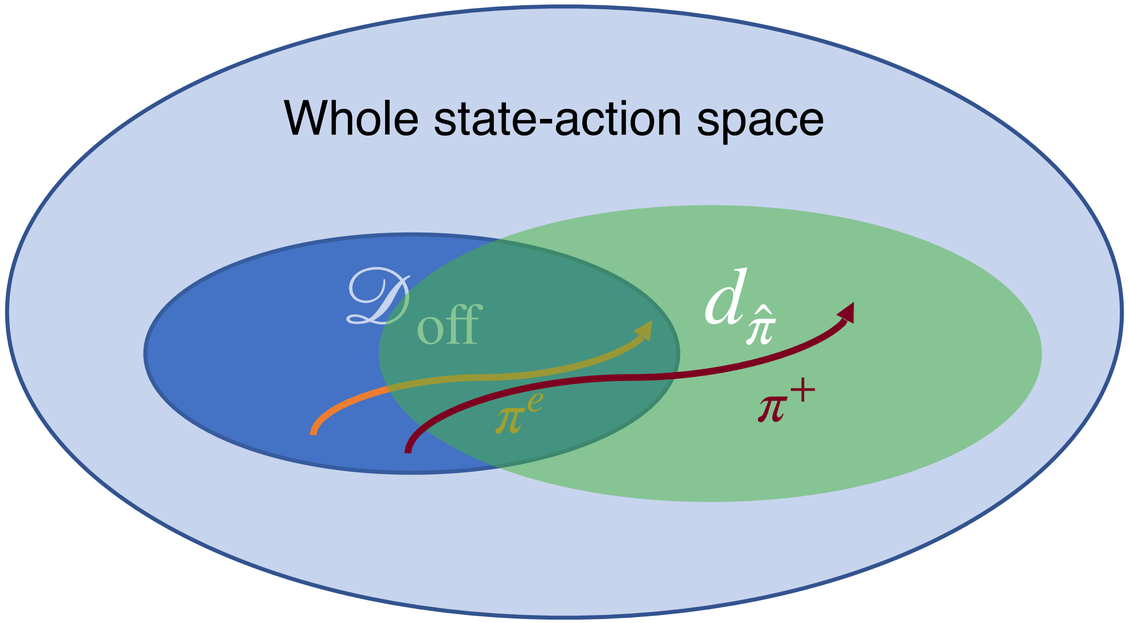
Dann, Chris, et al. "Guarantees for epsilon-greedy reinforcement learning with function approximation." *International conference on machine learning*. PMLR, 2022.

Can we do better?



Dann, Chris, et al. "Guarantees for epsilon-greedy reinforcement learning with function approximation." *International conference on machine learning*. PMLR, 2022.

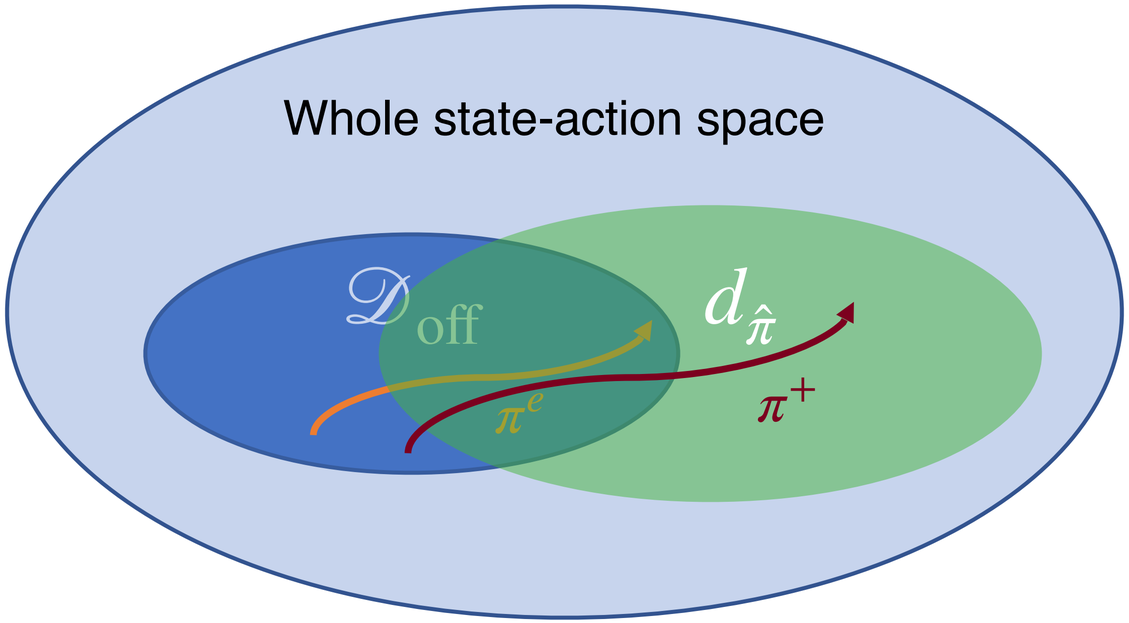
Can we do better?



$$C_\pi^+ = \max\left\{0,\ \max_{f\in\mathcal{F}} \frac{\sum_h \mathbb{E}_{s_h,a_h\sim\pi}\left(\mathcal{T}f(s,a) - f(s_h,a_h)\right)}{\sqrt{\sum_h \mathbb{E}_{s_h,a_h\sim(\frac{1}{2}\nu_h+\frac{1}{2}\bar{d}_{\hat{\pi}})}\left(f(s,a)-\mathcal{T}f(s_h,a_h)\right)^2}}\right\}$$

Dann, Chris, et al. "Guarantees for epsilon-greedy reinforcement learning with function approximation." *International conference on machine learning*. PMLR, 2022.

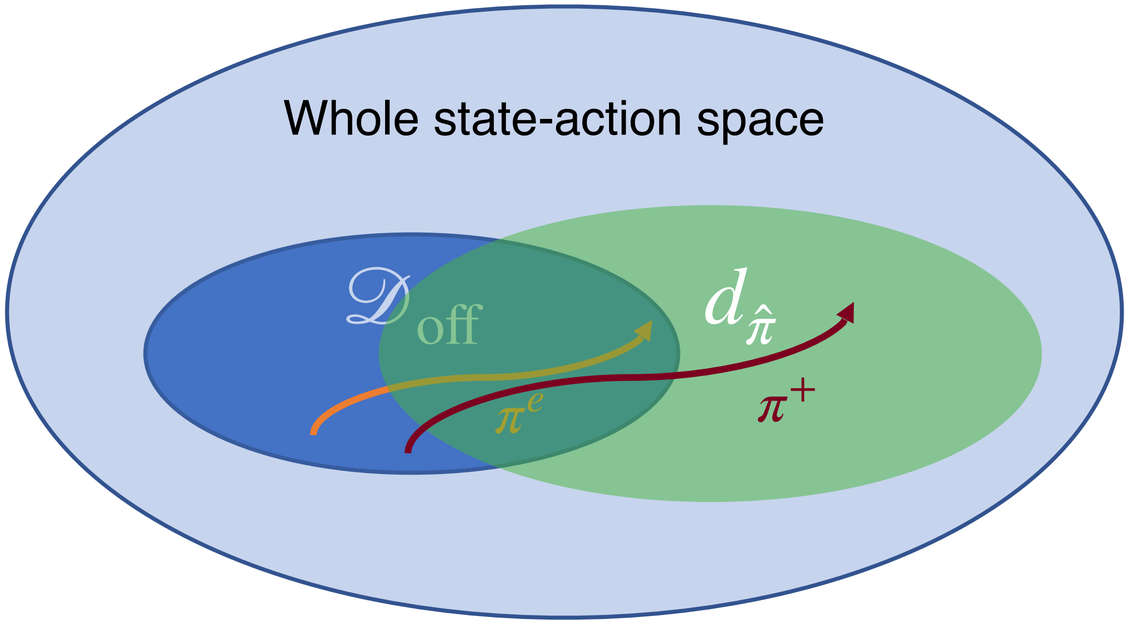
Can we do better?



$$C_\pi^+ = \max\left\{0,\ \max_{f\in\mathcal{F}} \frac{\sum_h \mathbb{E}_{s_h,a_h\sim\pi}\left(\mathcal{T}f(s,a) - f(s_h,a_h)\right)}{\sqrt{\sum_h \mathbb{E}_{s_h,a_h\sim(\frac{1}{2}\nu_h+\frac{1}{2}\bar{d}_{\hat{\pi}})}\left(f(s,a)-\mathcal{T}f(s_h,a_h)\right)^2}}\right\}$$

Dann, Chris, et al. "Guarantees for epsilon-greedy reinforcement learning with function approximation." *International conference on machine learning*. PMLR, 2022.

Can we do better?



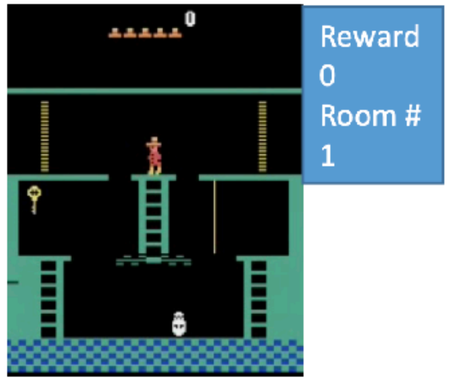
$$C_\pi^+ = \max\left\{0,\ \max_{f\in\mathcal{F}} \frac{\sum_h \mathbb{E}_{s_h,a_h\sim\pi}\left(\mathcal{T}f(s,a) - f(s_h,a_h)\right)}{\sqrt{\sum_h \mathbb{E}_{s_h,a_h\sim\left(\frac{1}{2}\nu_h+\frac{1}{2}\bar{d}_{\hat{\pi}}\right)}\left(f(s,a)-\mathcal{T}f(s_h,a_h)\right)^2}}\right\}$$

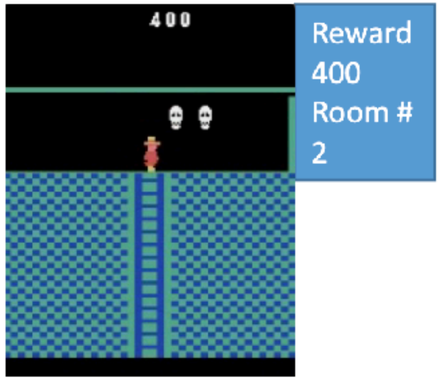
- We can now compare with the best policy covered by **{offline distribution + online distribution induced by the learned policies}**.
- Beneficial in:
  - **Local improvement**: the neighborhood of the offline distribution contains very good policies.
  - **Global improvement**: epsilon-greedy is effective in certain scenarios. See [Dann et al., 2022] for a great characterization.
  - Hint back at the ability to recover benefit.

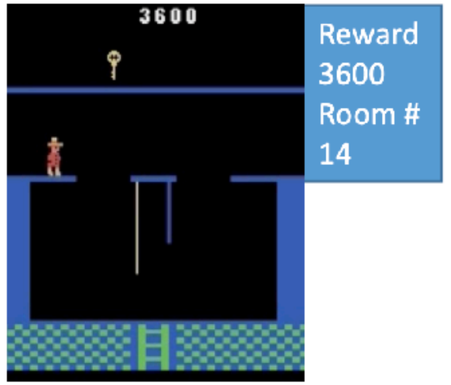
Dann, Chris, et al. "Guarantees for epsilon-greedy reinforcement learning with function approximation." *International conference on machine learning*. PMLR, 2022.

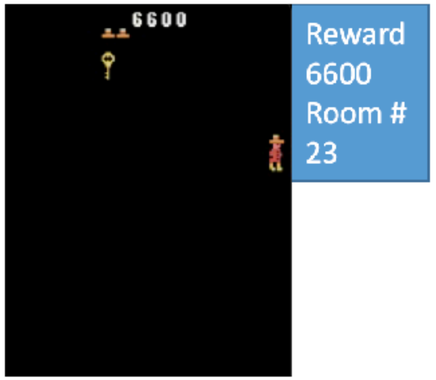
Experiments: comparing with online methods

Montezuma's Revenge





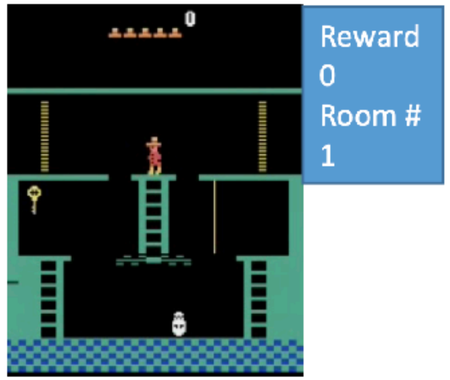


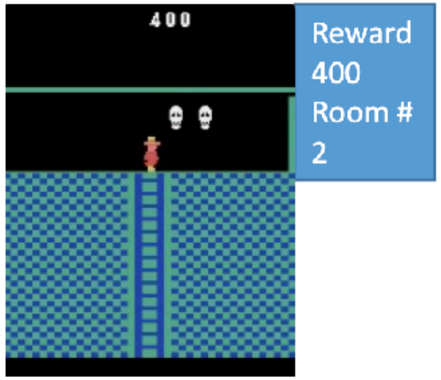


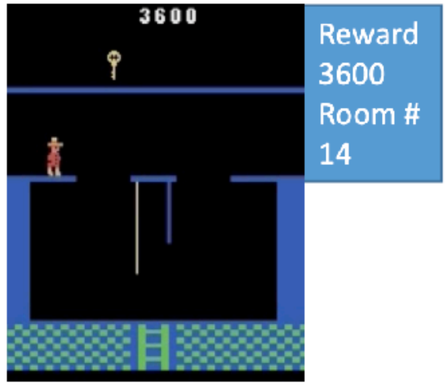
Experiments: comparing with online methods

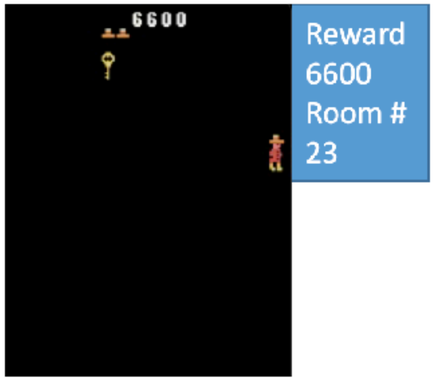
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal





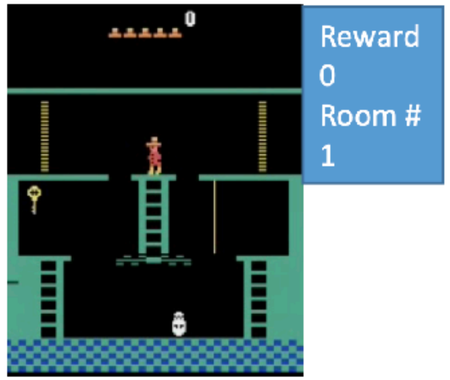


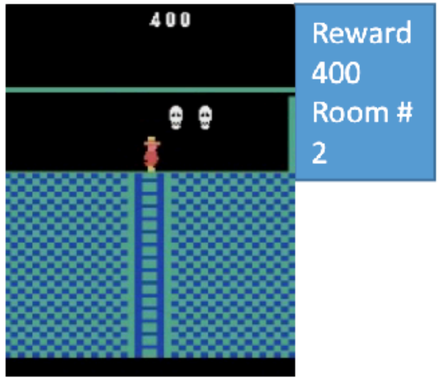


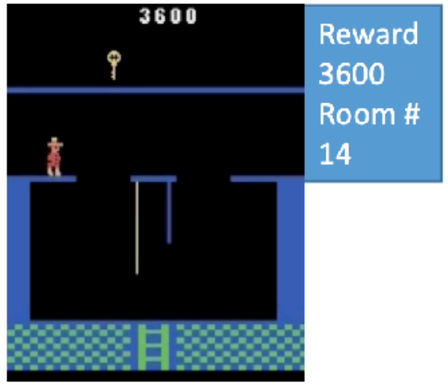
Experiments: comparing with online methods

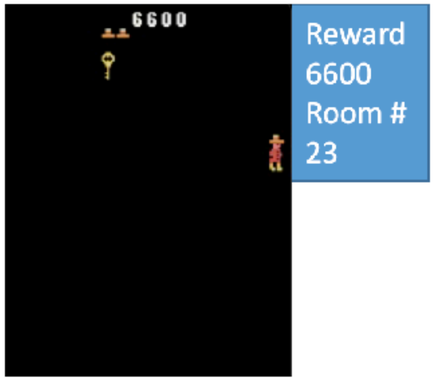
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal









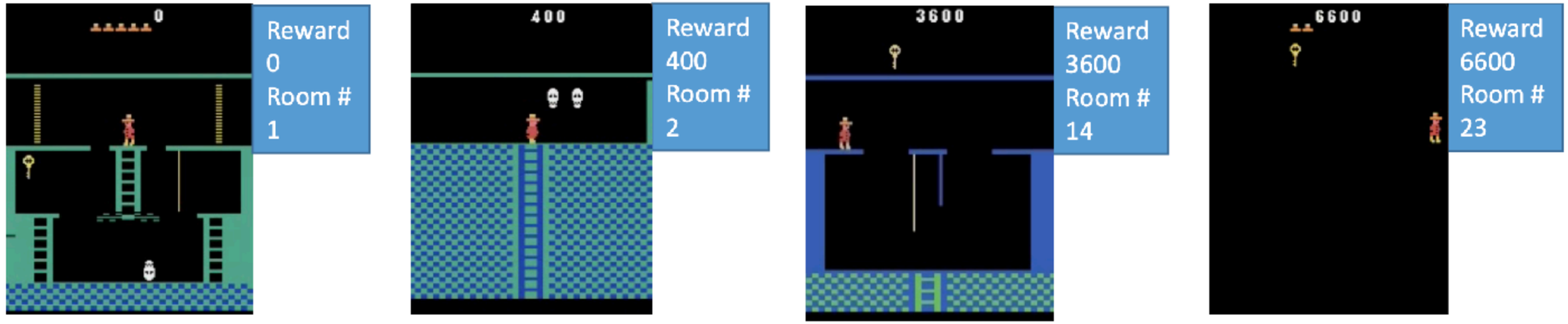
Offline dataset:

- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total

Experiments: comparing with online methods

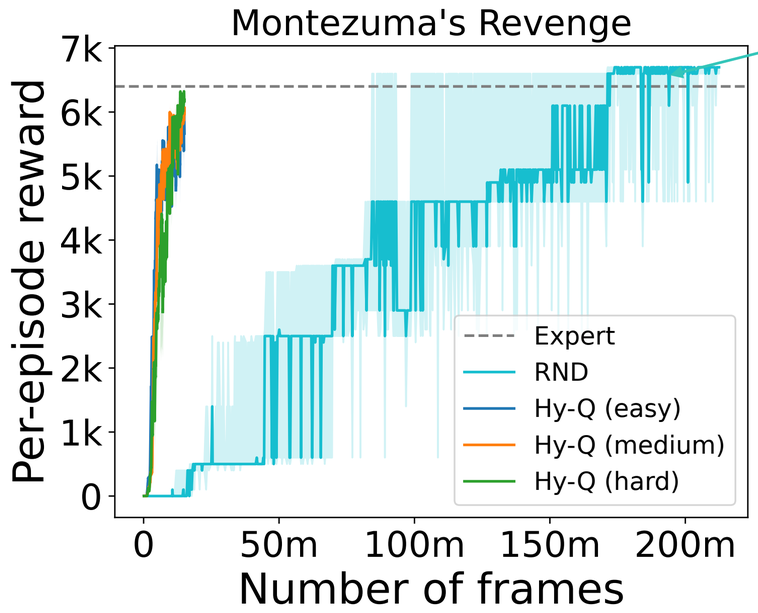
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal



Offline dataset:

- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total

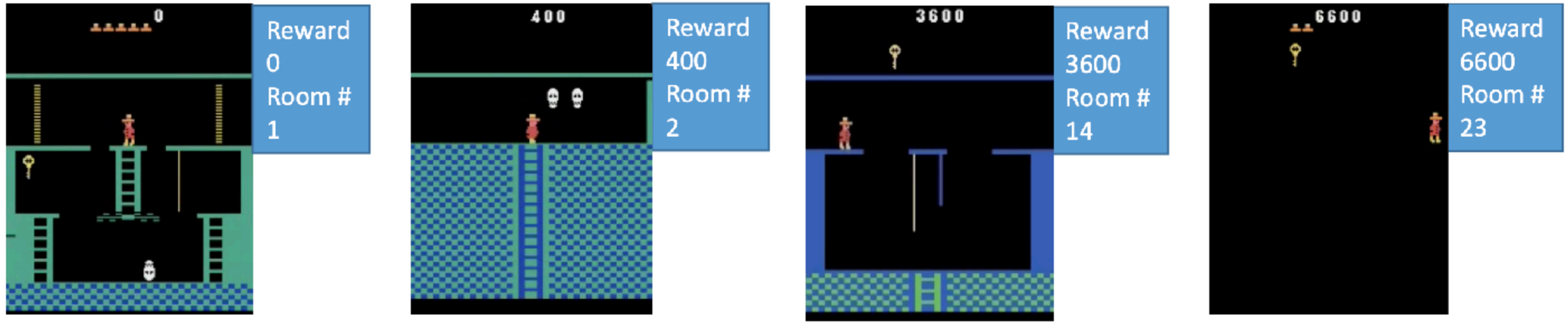


RND [Burda et al., 2018]: a method designed for M-revenge

Experiments: comparing with online methods

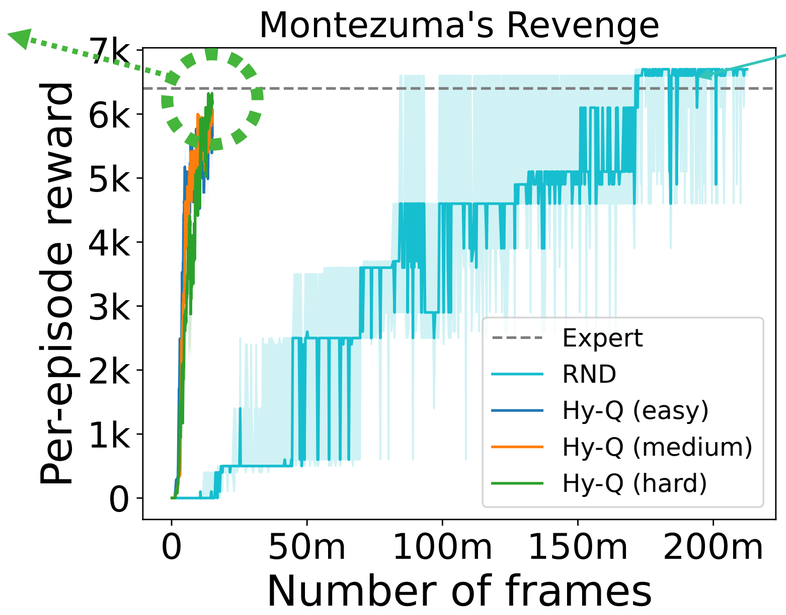
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal



Offline dataset:

- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total

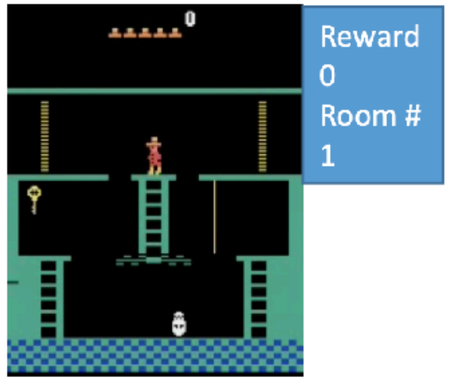


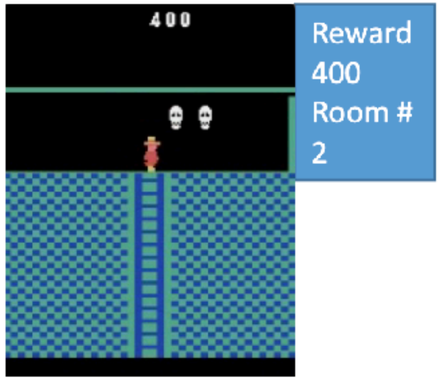
RND [Burda et al., 2018]: a method designed for M-revenge

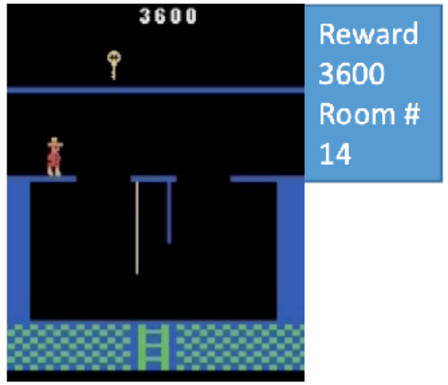
Experiments: comparing with online methods

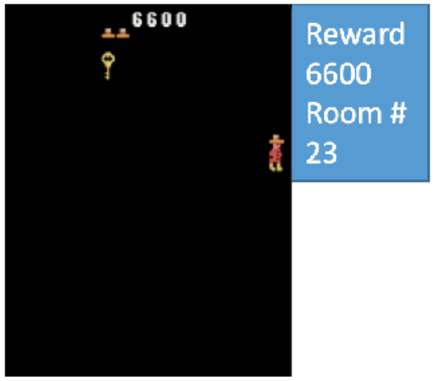
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal





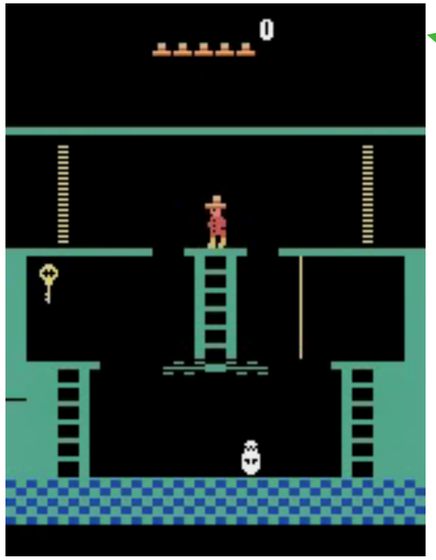


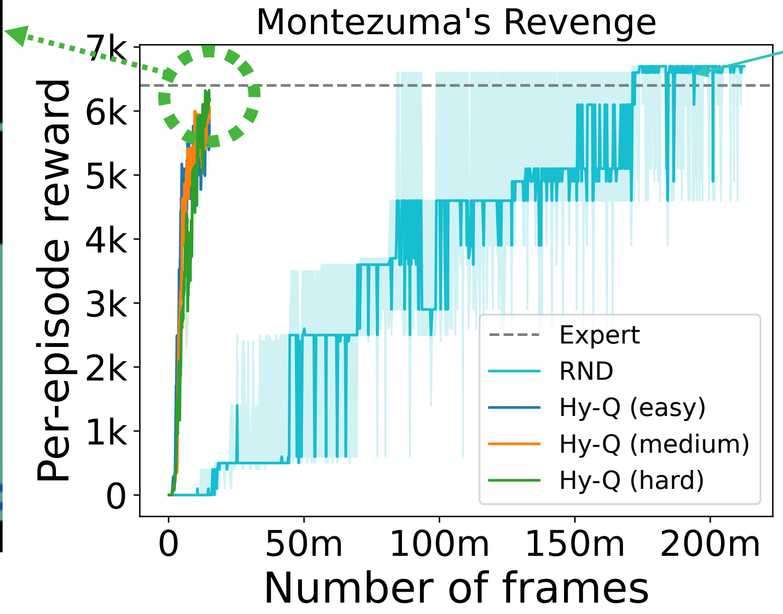


RND [Burda et al., 2018]: a method designed for M-revenge

Offline dataset:

- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total

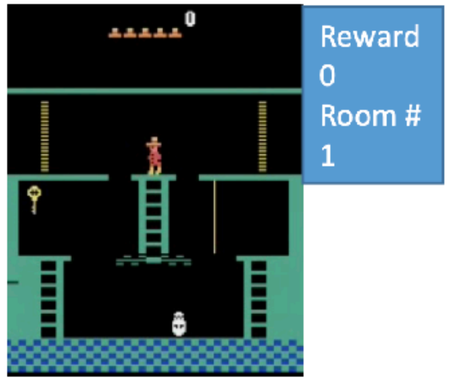


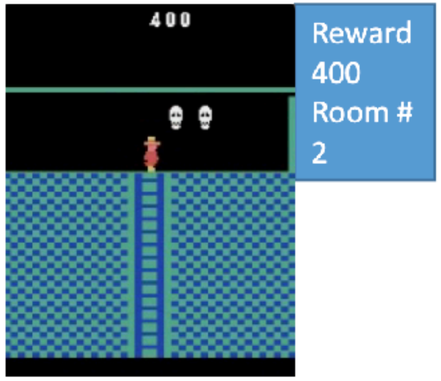


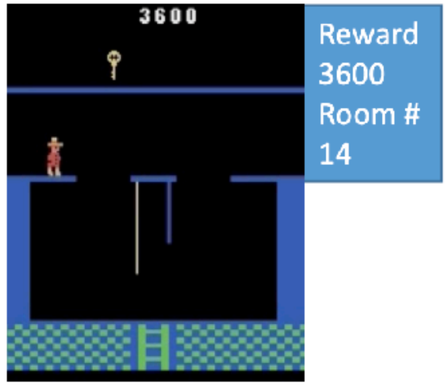
Experiments: comparing with online methods

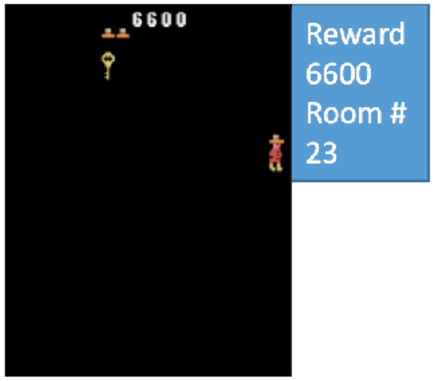
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal





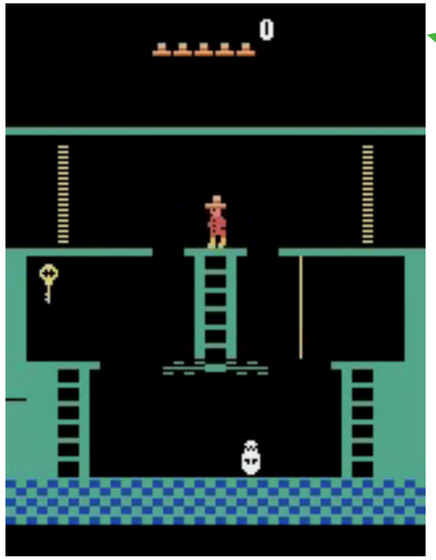


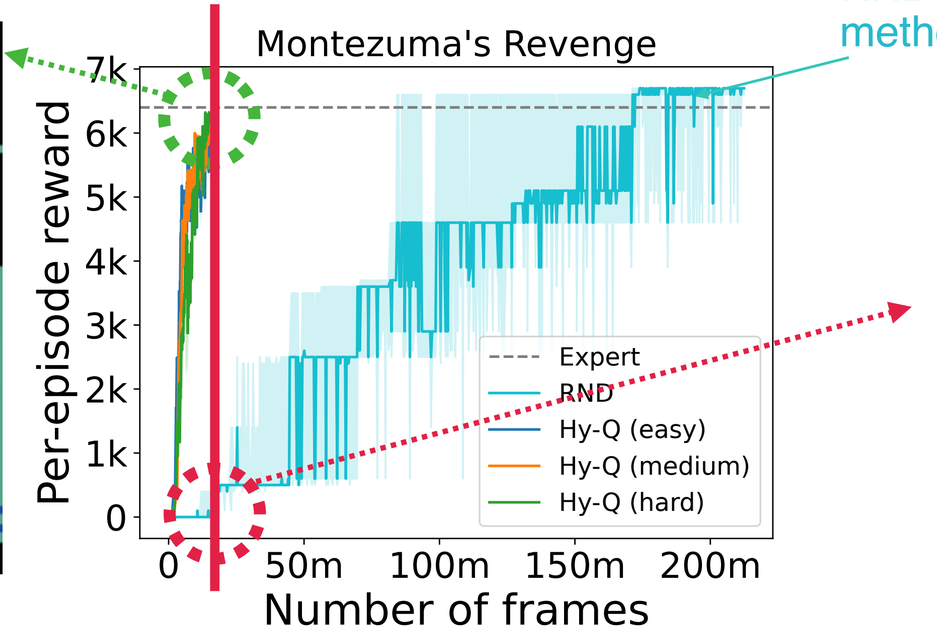


RND [Burda et al., 2018]: a method designed for M-revenge

Offline dataset:

- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total

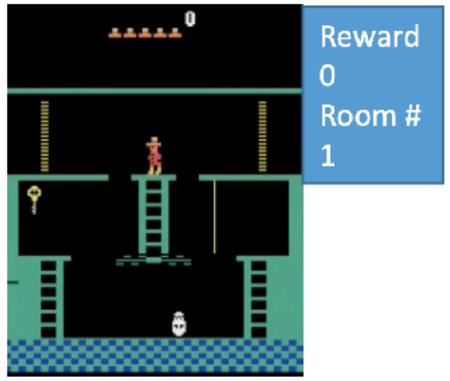


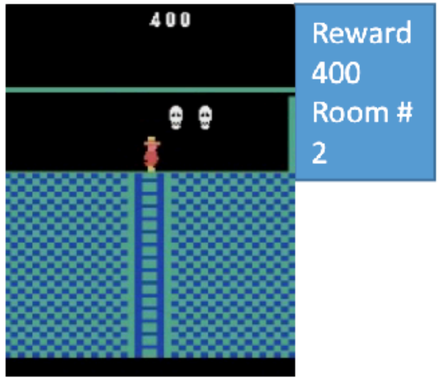


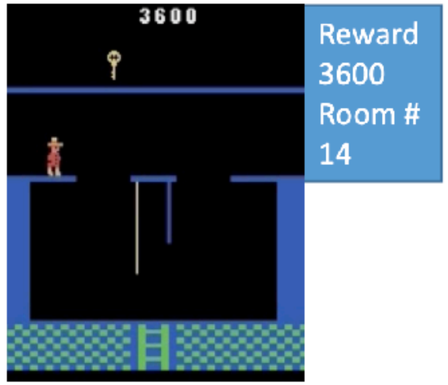
Experiments: comparing with online methods

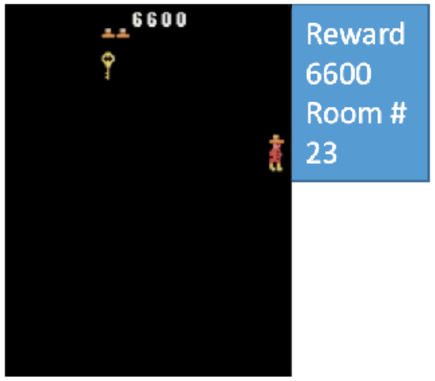
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal





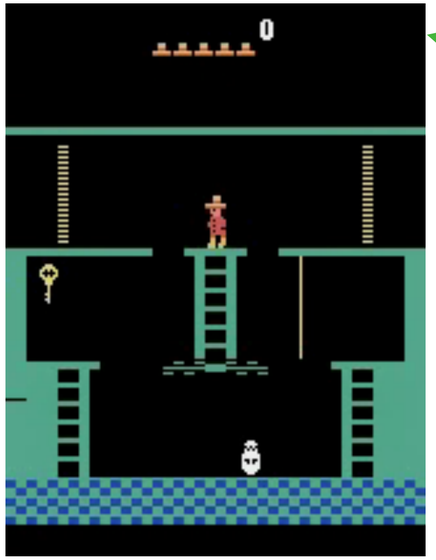


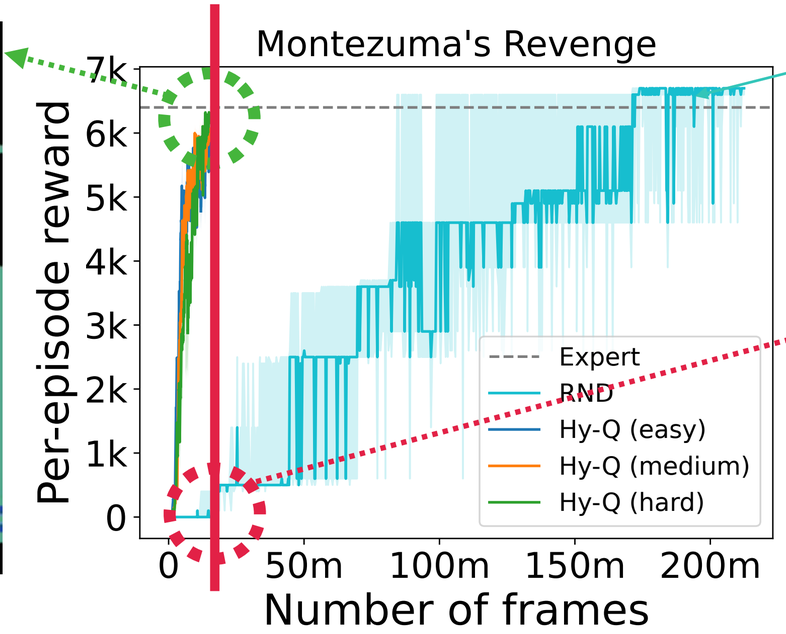


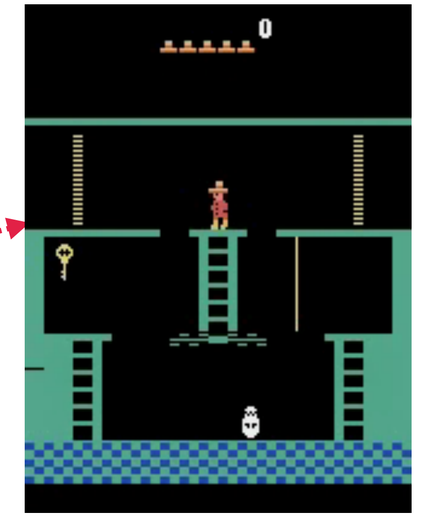
RND [Burda et al., 2018]: a method designed for M-revenge

Offline dataset:

- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total



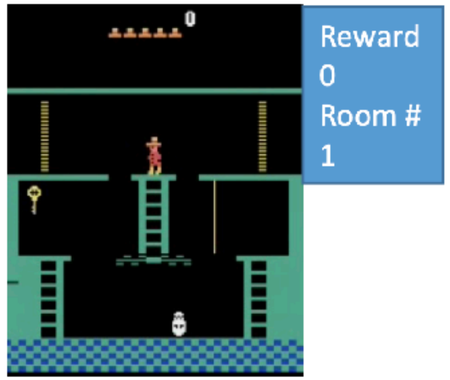


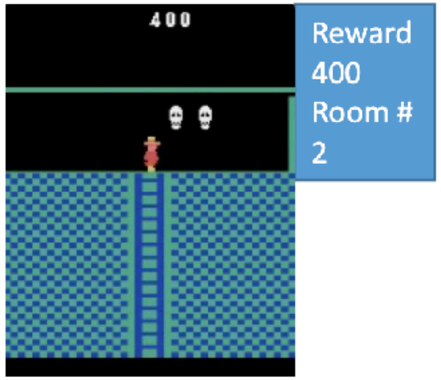


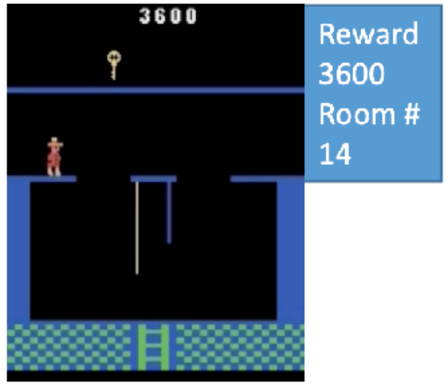
Experiments: comparing with online methods

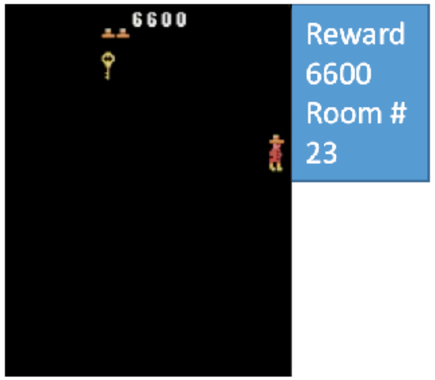
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal





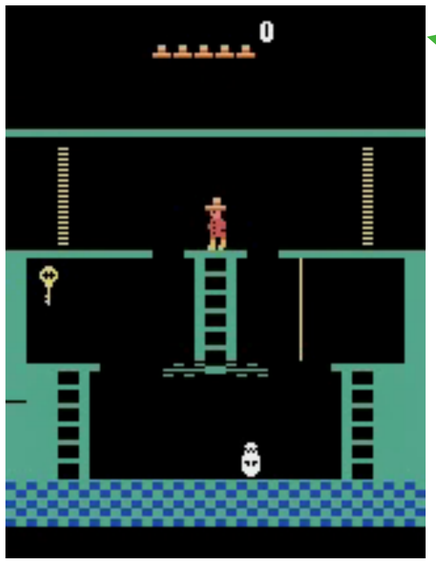


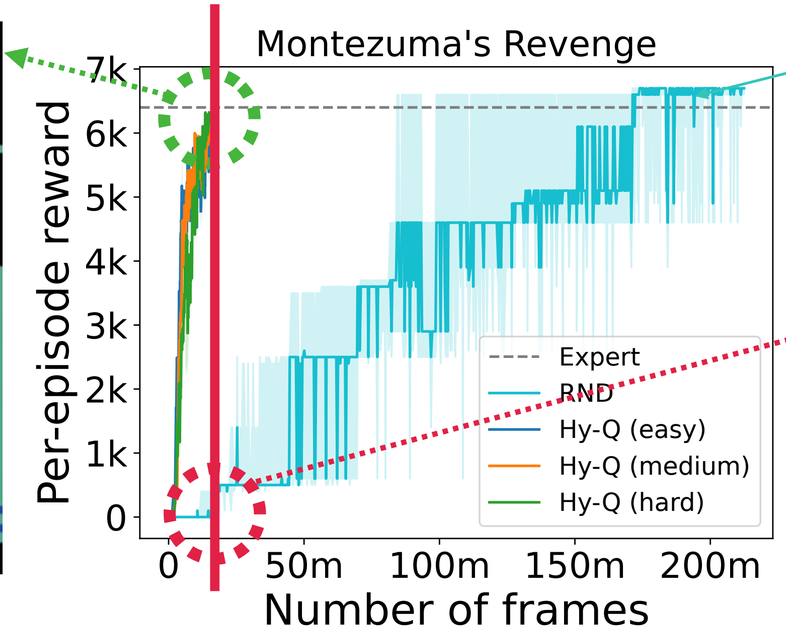


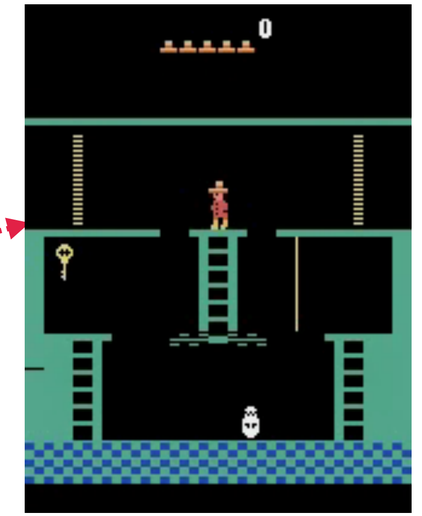
RND [Burda et al., 2018]: a method designed for M-revenge

Offline dataset:

- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total





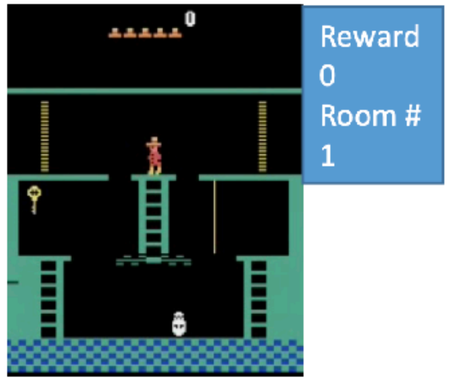


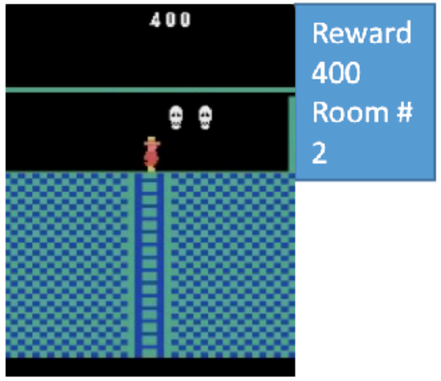
Burda, Yuri, et al. "Exploration by random network distillation." *International Conference on Learning Representations*. 2018.

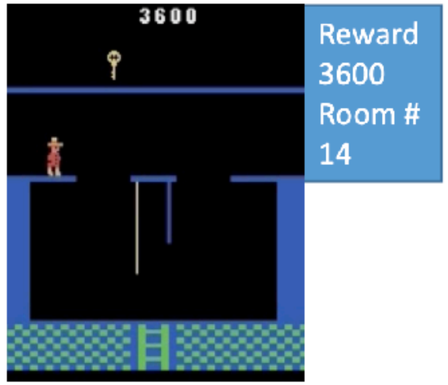
Experiments: comparing with offline RL & IL

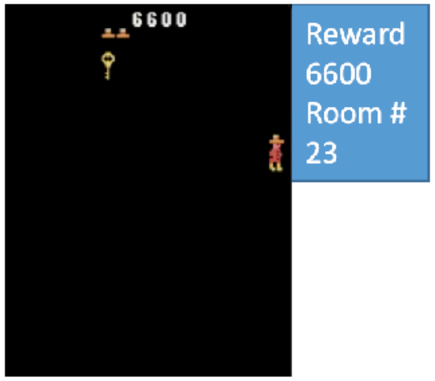
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal



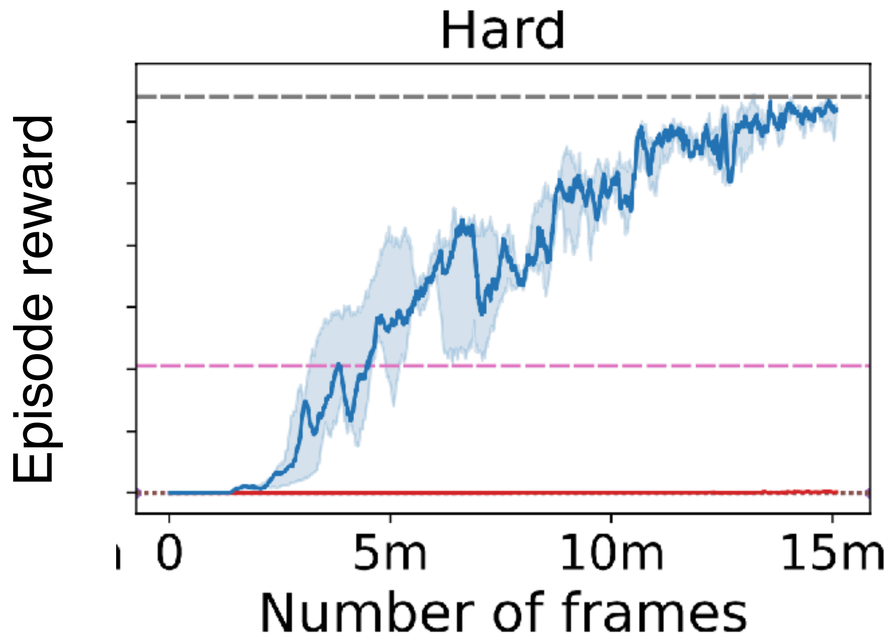






Offline dataset:

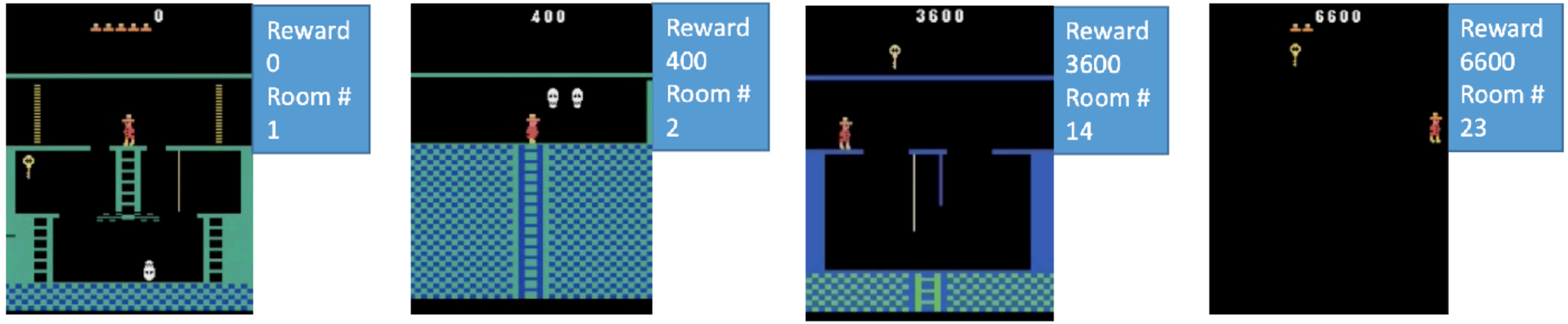
- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total



Experiments: comparing with offline RL & IL

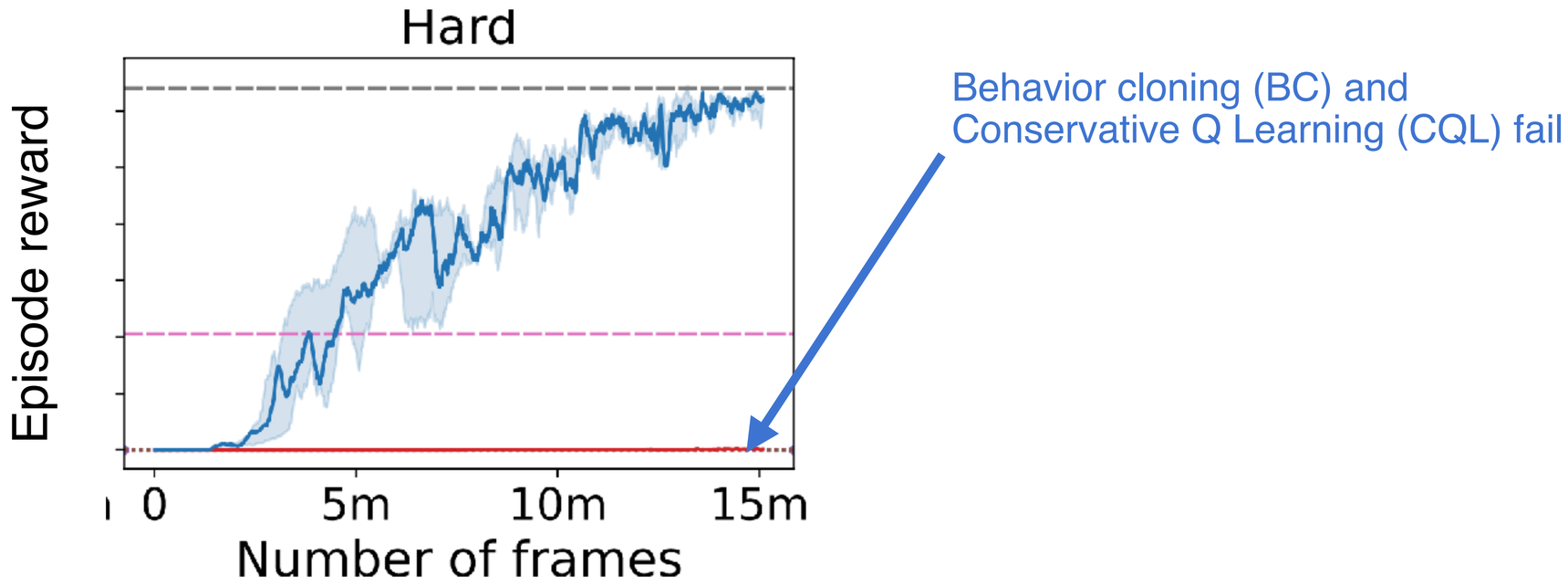
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal



Offline dataset:

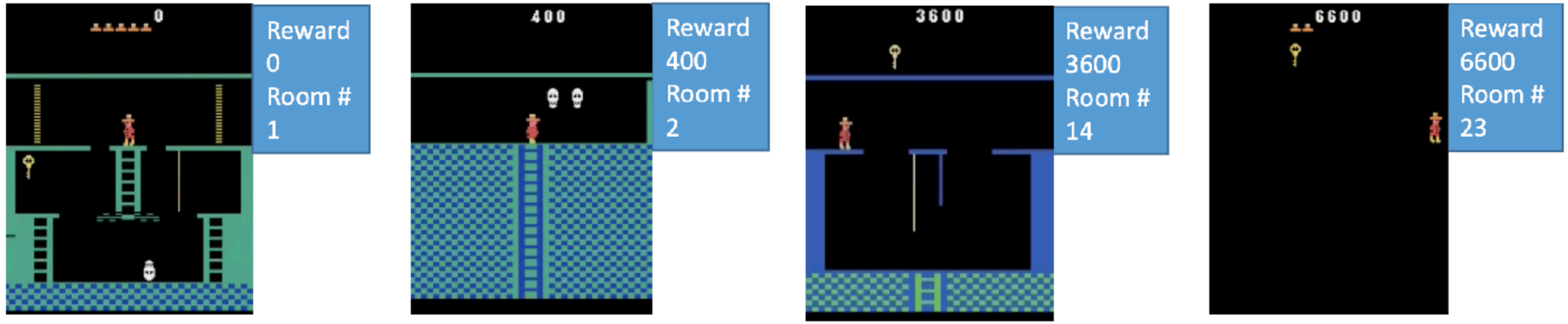
- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total



Experiments: comparing with offline RL & IL

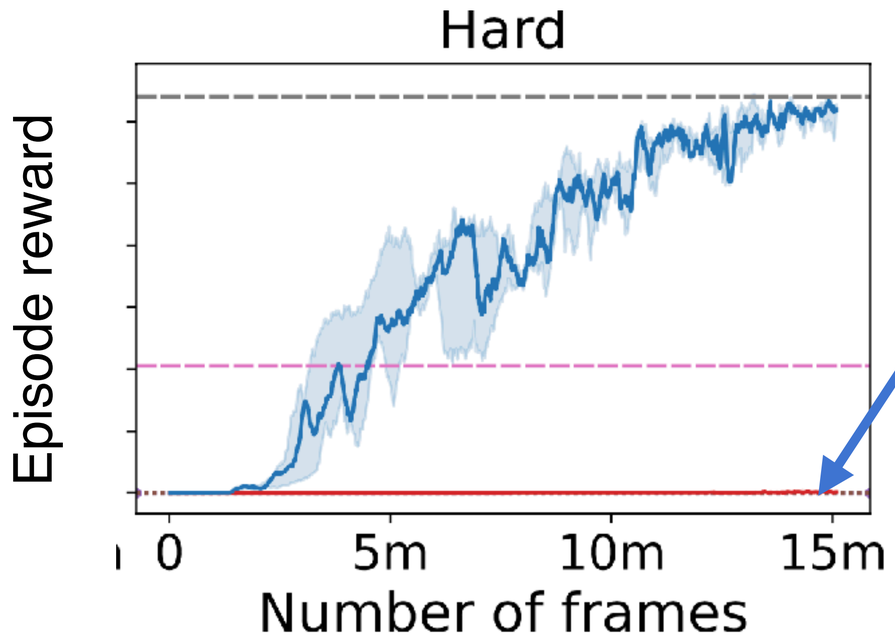
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal



Offline dataset:

- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total



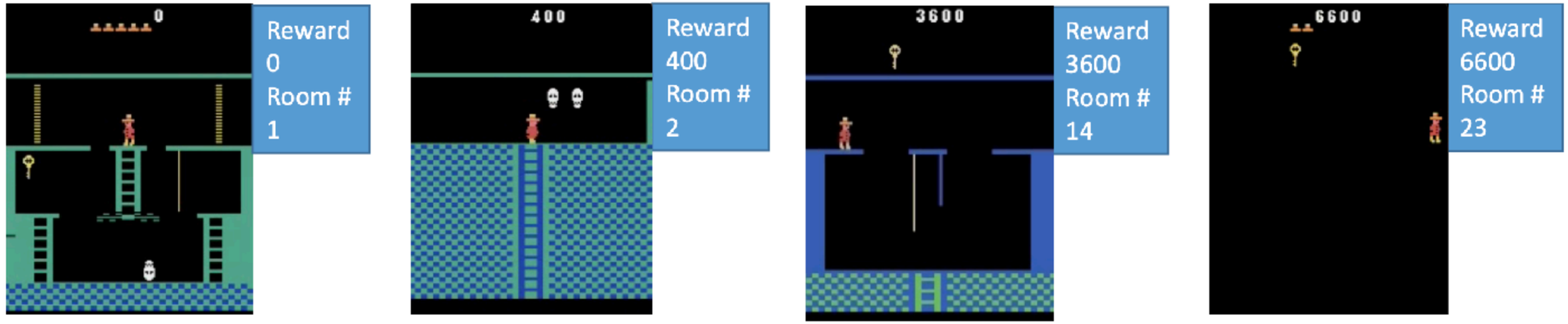
Behavior cloning (BC) and Conservative Q Learning (CQL) fail

- BC fails since offline data contains data from low quality policies
- CQL [Kumar et al, 2020] does work for a few Atari games; most other offline deep RL baselines only work for simpler low-dim tasks

Experiments: comparing with offline RL & IL

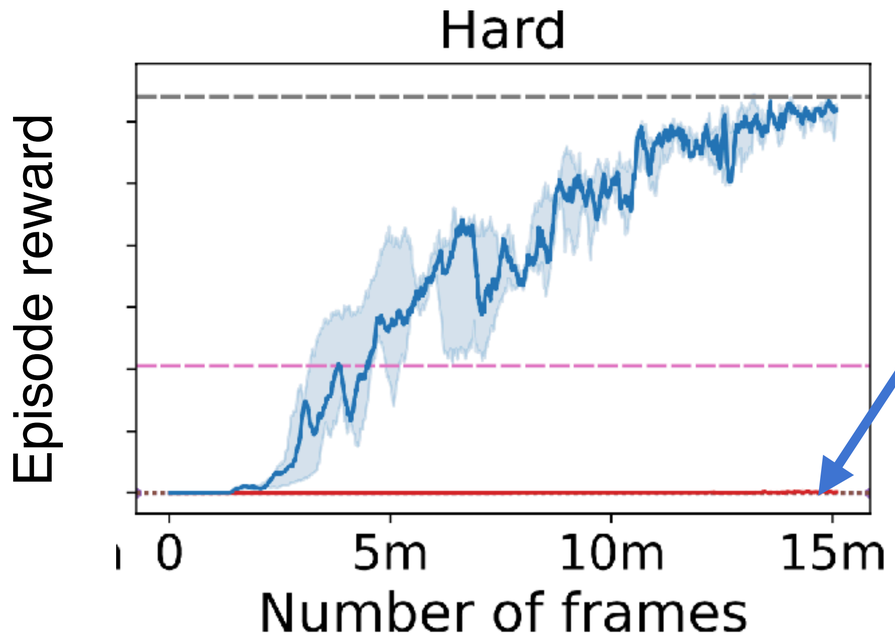
Montezuma's Revenge

- Image states
- 17 actions
- Extremely sparse reward signal



Offline dataset:

- Mixing data from an expert policy (50%) and a random policy (50%).
- 0.1 m samples in total



Behavior cloning (BC) and Conservative Q Learning (CQL) fail

- BC fails since offline data contains data from low quality policies
- CQL [Kumar et al, 2020] does work for a few Atari games; most other offline deep RL baselines only work for simpler low-dim tasks

Kumar, Aviral, et al. "Conservative q-learning for offline reinforcement learning." *Advances in Neural Information Processing Systems* 33 (2020): 1179-1191.

Summary

Hybrid RL

Online RL + Offline RL + Structure

Summary

Hybrid RL

Online RL + Offline RL + Structure

Comparing with online RL

- Statistical gain vs. online RL without structure.
- Computational gain vs. online RL with structure.

Summary

Hybrid RL

Online RL + Offline RL + Structure

Comparing with online RL

- Statistical gain vs. online RL without structure.
- Computational gain vs. online RL with structure.

Comparing with offline RL

- Computational gain vs. global pessimism.
- Online verification.
- Ability to recover.

Summary

Hybrid RL

Online RL + Offline RL + Structure

Comparing with online RL

- Statistical gain vs. online RL without structure.
- Computational gain vs. online RL with structure.

Comparing with offline RL

- Computational gain vs. global pessimism.
- Online verification.
- Ability to recover.

Benefit of hybrid RL

- **Solves a large family of RL just via Supervised Learning. (Just combine your offline and online data in a balanced way.)**
- Easy to implement, agnostic to specific structure, works well in practice.

Future directions

Li, Gen, et al. "Reward-agnostic Fine-tuning: Provable Statistical Benefits of Hybrid Reinforcement Learning.”
Wagenmaker, Andrew, and Pacchiano, Aldo. "Leveraging offline data in online reinforcement learning.”
Zhou, Yifei, et al. “Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees.”

Future directions

Hybrid RL has many exciting problems that are yet to solve:

Li, Gen, et al. "Reward-agnostic Fine-tuning: Provable Statistical Benefits of Hybrid Reinforcement Learning.”
Wagenmaker, Andrew, and Pacchiano, Aldo. "Leveraging offline data in online reinforcement learning.”
Zhou, Yifei, et al. “Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees.”

Future directions

Hybrid RL has many exciting problems that are yet to solve:

1. Hy-Q can only compare with the best policy covered by the offline dataset.

Li, Gen, et al. "Reward-agnostic Fine-tuning: Provable Statistical Benefits of Hybrid Reinforcement Learning."
Wagenmaker, Andrew, and Pacchiano, Aldo. "Leveraging offline data in online reinforcement learning."
Zhou, Yifei, et al. "Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees."

Future directions

Hybrid RL has many exciting problems that are yet to solve:

1. Hy-Q can only compare with the best policy covered by the offline dataset.

   - Comparing with **optimal policy**: tabular MDPs [Li et al., 2023], linear MDPs [Wagenmaker & Pacchiano, 2022]

Li, Gen, et al. "Reward-agnostic Fine-tuning: Provable Statistical Benefits of Hybrid Reinforcement Learning."
Wagenmaker, Andrew, and Pacchiano, Aldo. "Leveraging offline data in online reinforcement learning."
Zhou, Yifei, et al. "Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees."

Future directions

Hybrid RL has many exciting problems that are yet to solve:

1. Hy-Q can only compare with the best policy covered by the offline dataset.

   - Comparing with **optimal policy**: tabular MDPs [Li et al., 2023], linear MDPs [Wagenmaker & Pacchiano, 2022]

   - What happens beyond linear MDPs?

Li, Gen, et al. "Reward-agnostic Fine-tuning: Provable Statistical Benefits of Hybrid Reinforcement Learning."
Wagenmaker, Andrew, and Pacchiano, Aldo. "Leveraging offline data in online reinforcement learning."
Zhou, Yifei, et al. "Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees."

Future directions

Hybrid RL has many exciting problems that are yet to solve:

1. Hy-Q can only compare with the best policy covered by the offline dataset.
   - Comparing with **optimal policy**: tabular MDPs [Li et al., 2023], linear MDPs [Wagenmaker & Pacchiano, 2022]
   - What happens beyond linear MDPs?
2. Hy-Q requires Bellman completeness assumption.

Li, Gen, et al. "Reward-agnostic Fine-tuning: Provable Statistical Benefits of Hybrid Reinforcement Learning."
Wagenmaker, Andrew, and Pacchiano, Aldo. "Leveraging offline data in online reinforcement learning."
Zhou, Yifei, et al. "Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees."

Future directions

Hybrid RL has many exciting problems that are yet to solve:

1. Hy-Q can only compare with the best policy covered by the offline dataset.
   - Comparing with **optimal policy**: tabular MDPs [Li et al., 2023], linear MDPs [Wagenmaker & Pacchiano, 2022]
   - What happens beyond linear MDPs?
2. Hy-Q requires Bellman completeness assumption.
   - Actor-critic style hybrid RL algorithm: **best-of-both-world** guarantee: only requires either Bellman completeness or standard online PG assumption [Zhou et al., 2023].

Li, Gen, et al. "Reward-agnostic Fine-tuning: Provable Statistical Benefits of Hybrid Reinforcement Learning."
Wagenmaker, Andrew, and Pacchiano, Aldo. "Leveraging offline data in online reinforcement learning."
Zhou, Yifei, et al. "Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees."

Future directions

Hybrid RL has many exciting problems that are yet to solve:

1. Hy-Q can only compare with the best policy covered by the offline dataset.
   - Comparing with **optimal policy**: tabular MDPs [Li et al., 2023], linear MDPs [Wagenmaker & Pacchiano, 2022]
   - What happens beyond linear MDPs?
2. Hy-Q requires Bellman completeness assumption.
   - Actor-critic style hybrid RL algorithm: **best-of-both-world** guarantee: only requires either Bellman completeness or standard online PG assumption [Zhou et al., 2023].
3. Are the structural assumptions necessary?

Li, Gen, et al. "Reward-agnostic Fine-tuning: Provable Statistical Benefits of Hybrid Reinforcement Learning."
Wagenmaker, Andrew, and Pacchiano, Aldo. "Leveraging offline data in online reinforcement learning."
Zhou, Yifei, et al. "Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees."

Hybrid RL:
Using Both Offline and Online Data can Make RL Efficient

Yuda Song*, Yifei Zhou*, Ayush Sekhari,
Drew Bagnell, Akshay Krishnamurthy, Wen Sun