Cornell Bowers C·IS
College of Computing
and Information Science

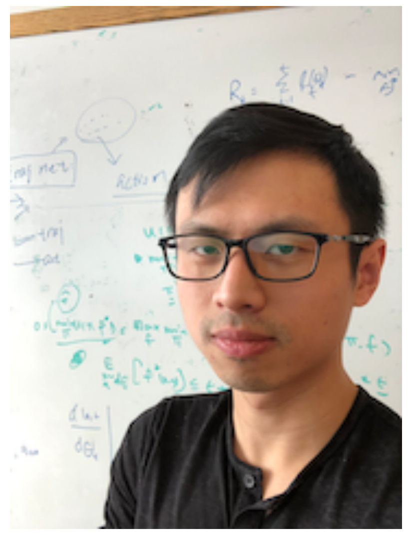
Cornell Engineering
Operations Research
and Information Engineering

Learning Bellman Complete Representations for Off-Policy Evaluation

Jonathan D. Chang and Kaiwen Wang

Cornell University

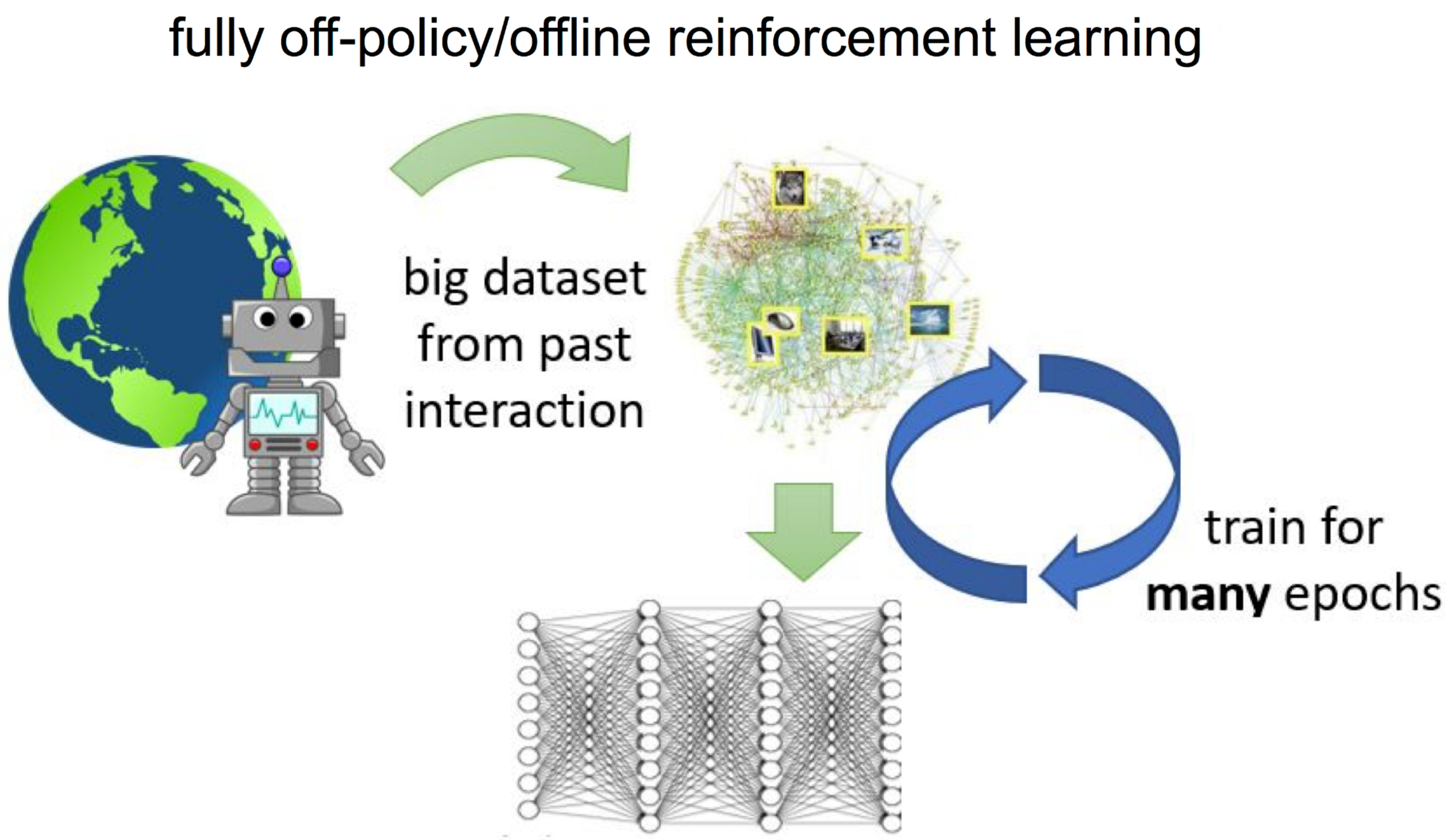
Joint work with Nathan Kallus and Wen Sun

Off-Policy Evaluation (OPE)

Evaluate target policy π^e using data from behavior policy π_0, in a **high-dimensional, complex environment with**

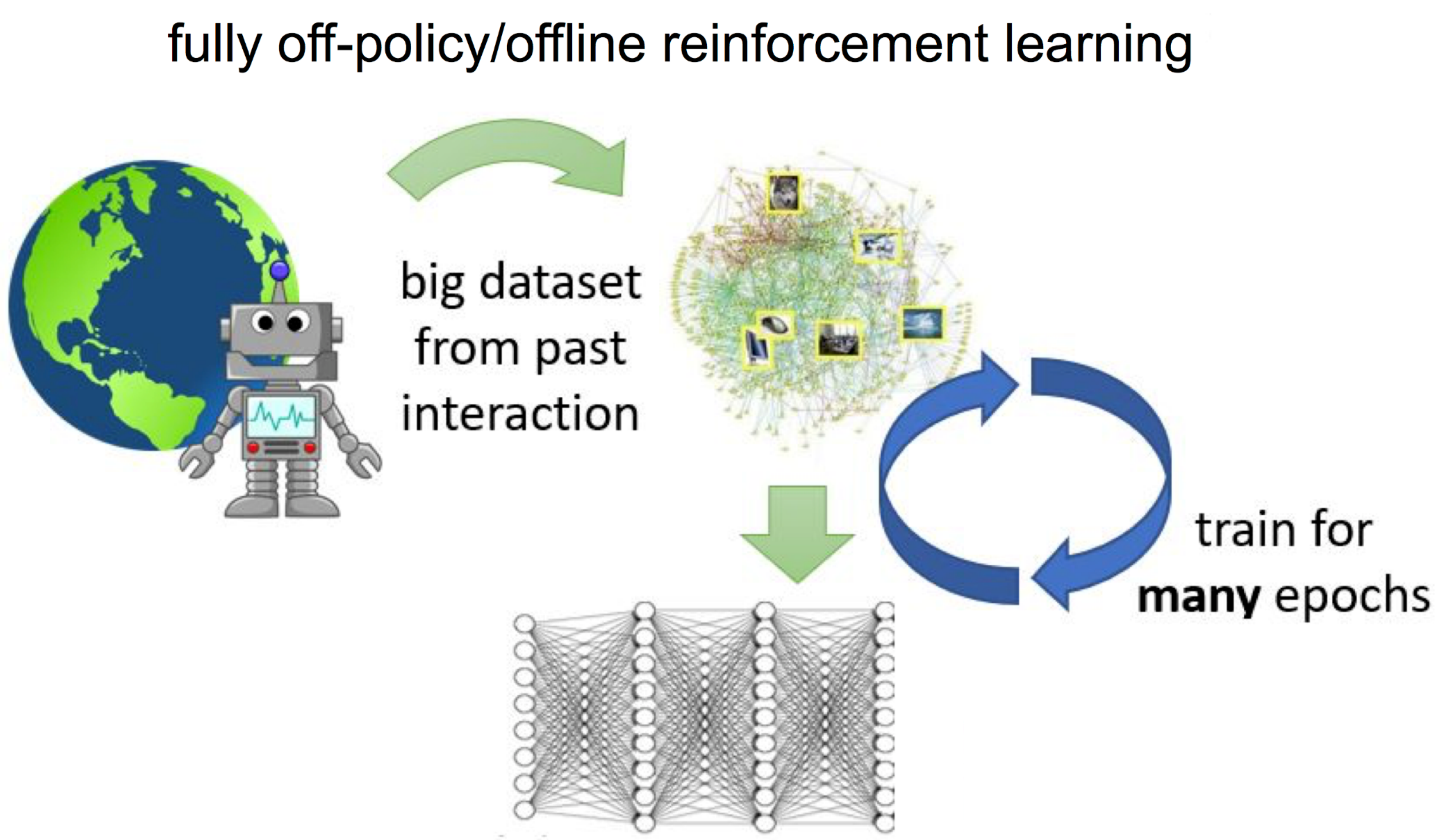
- Image observations,
- Continuous actions.

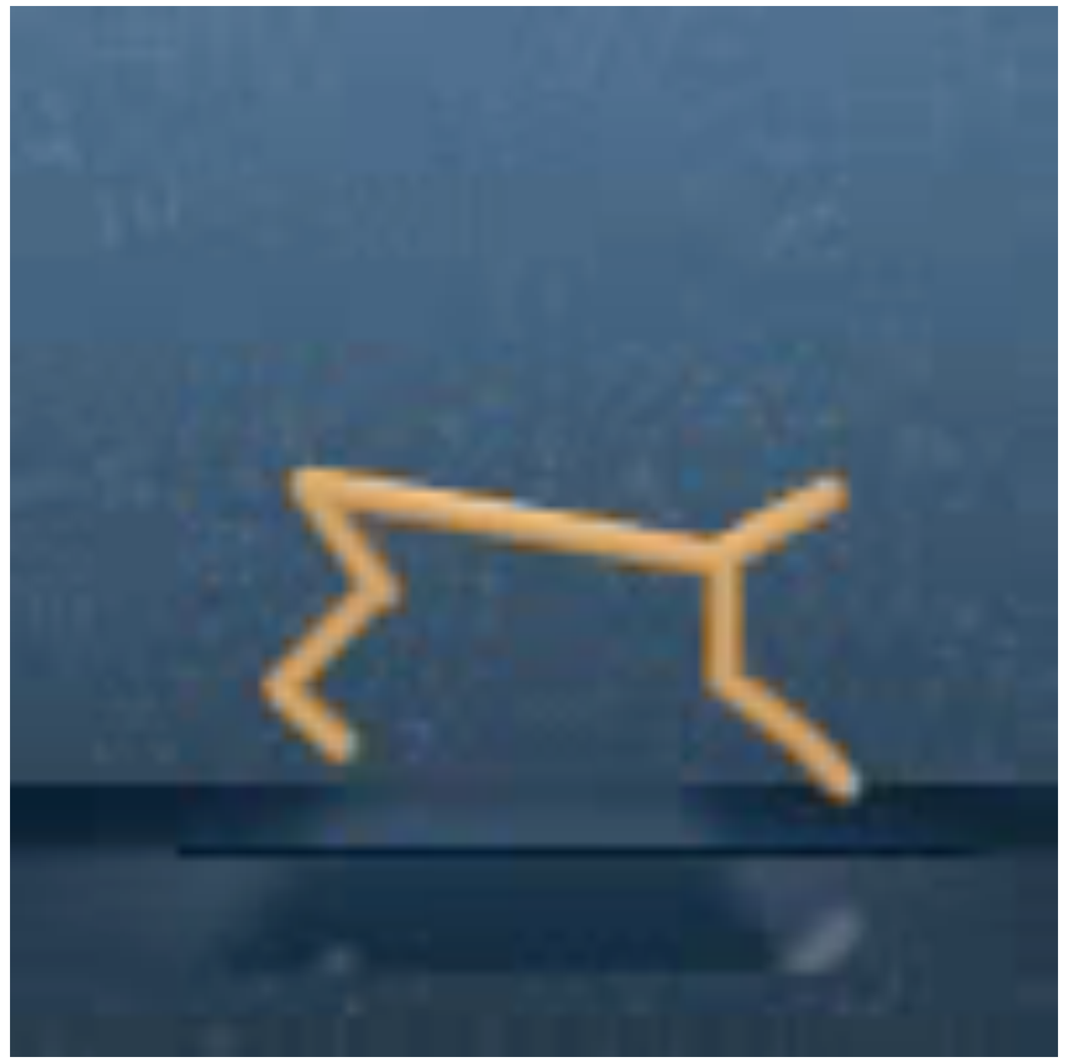


Off-Policy Evaluation (OPE)

Evaluate target policy π^e using data from behavior policy π_0, in a **high-dimensional, complex environment with**

- Image observations,
- Continuous actions.

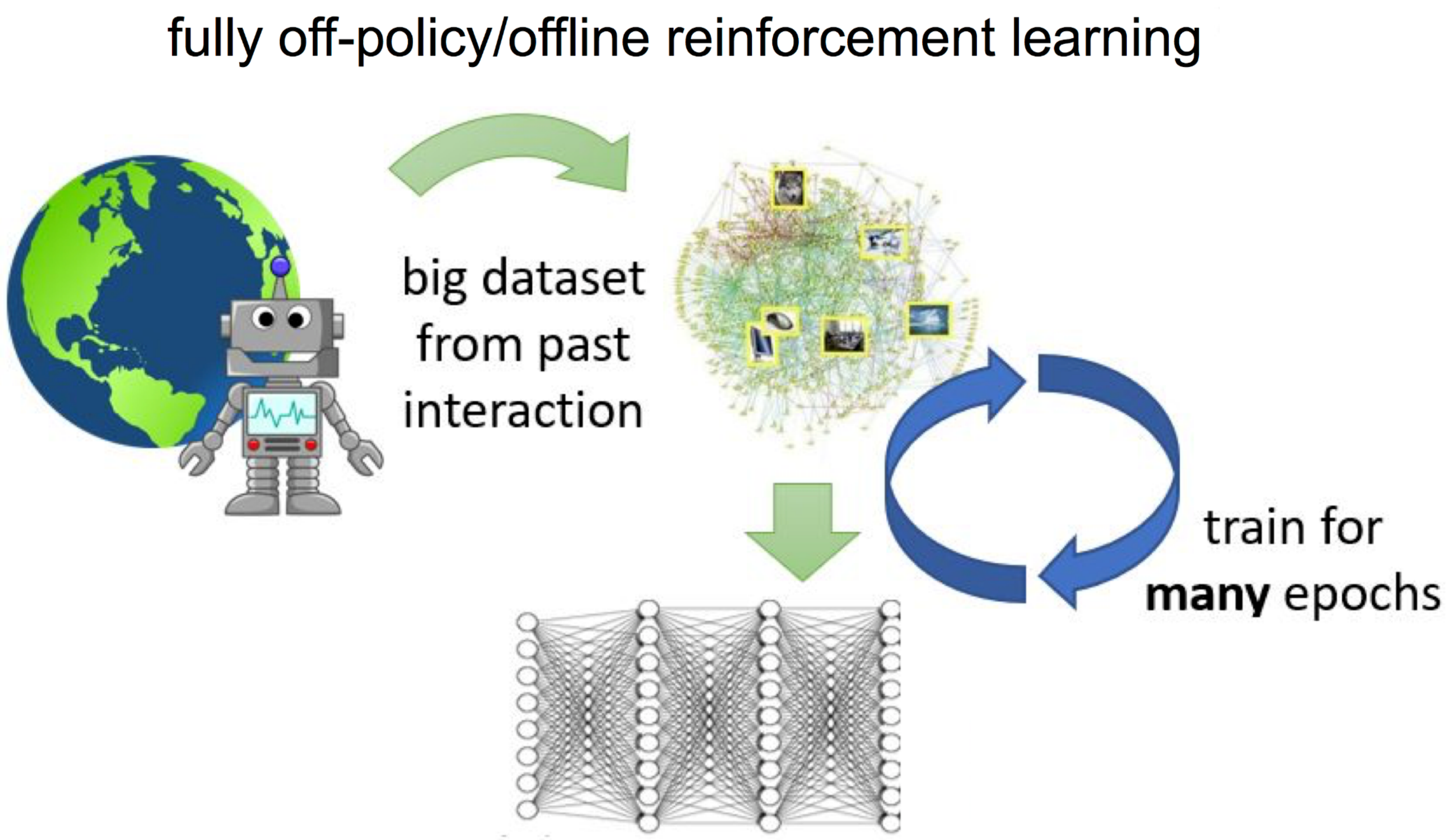


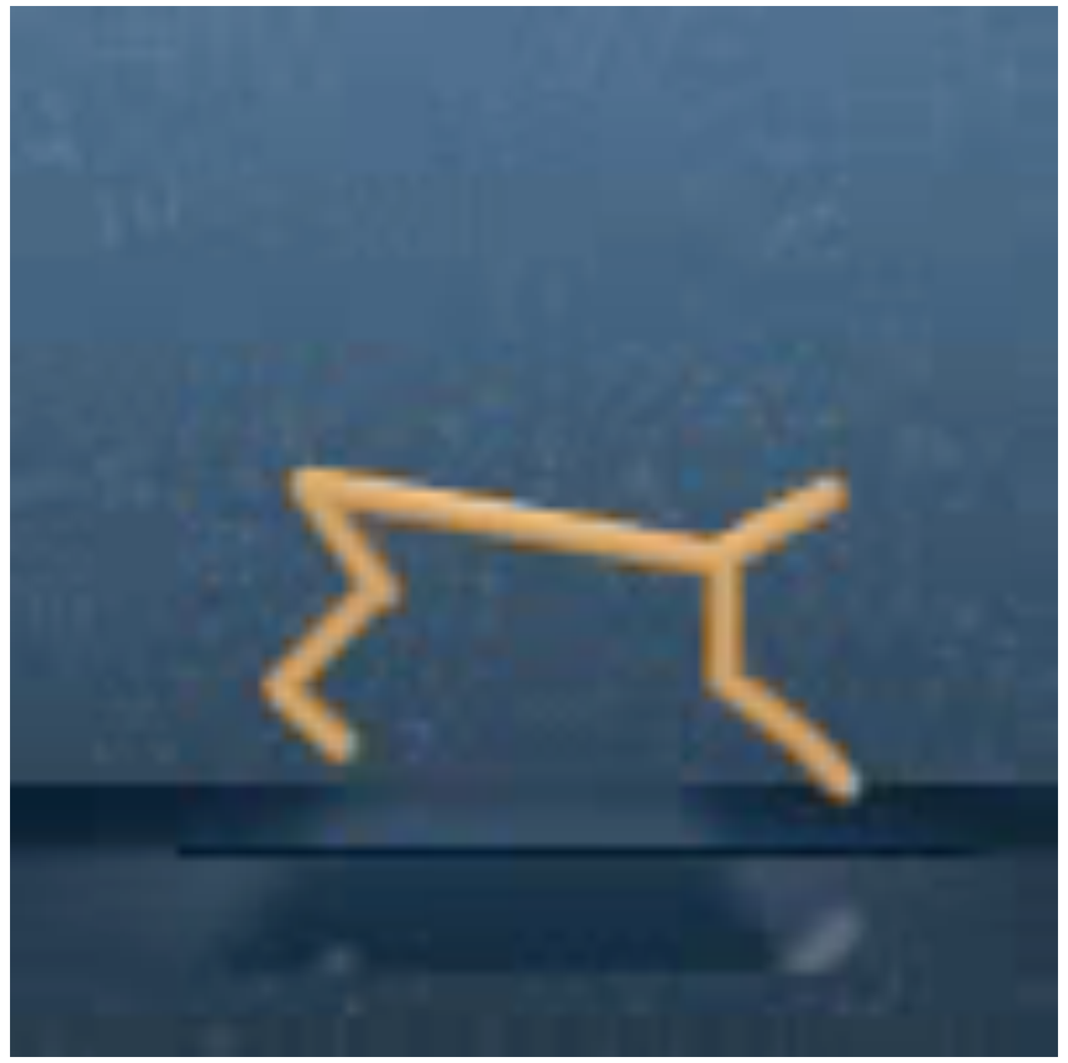


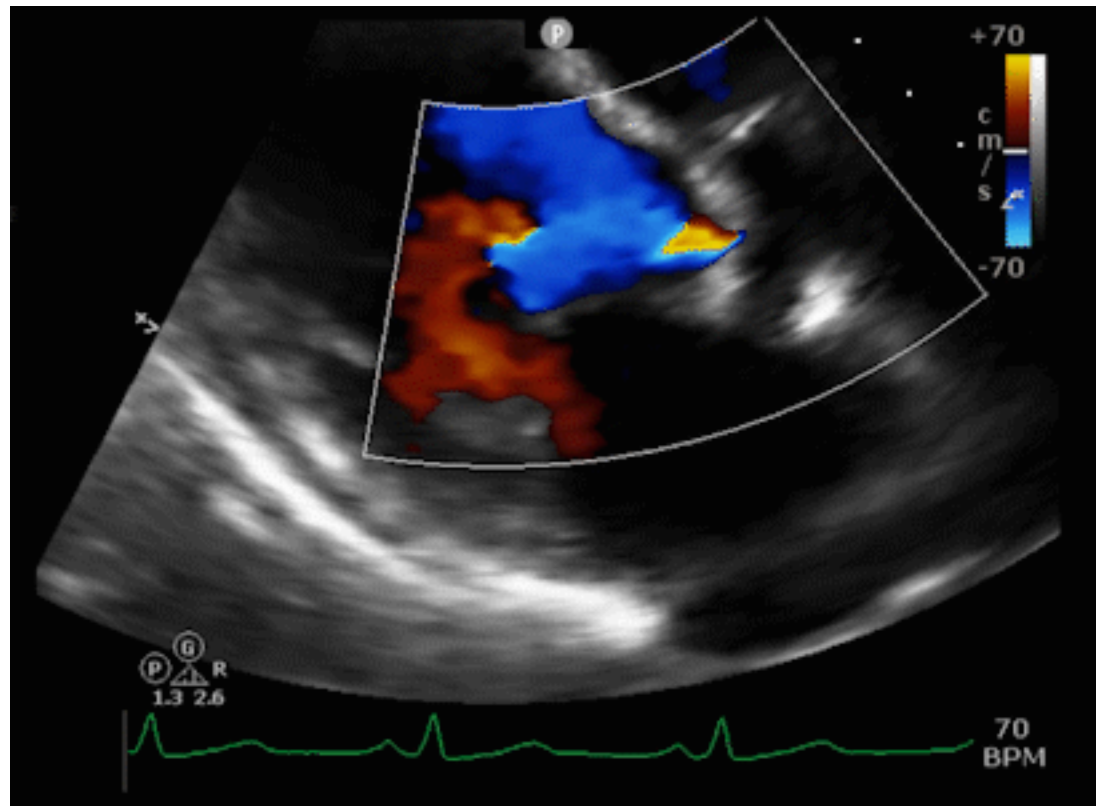
Off-Policy Evaluation (OPE)

Evaluate target policy π^e using data from behavior policy π_0, in a **high-dimensional, complex environment with**

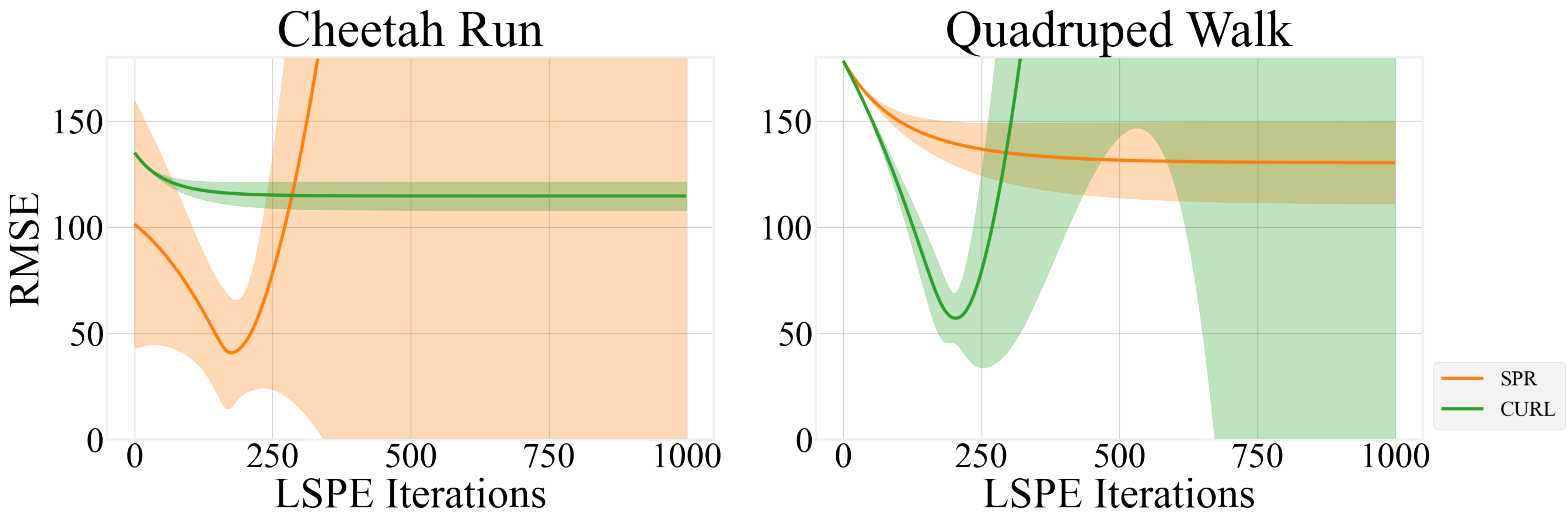
- Image observations,
- Continuous actions.



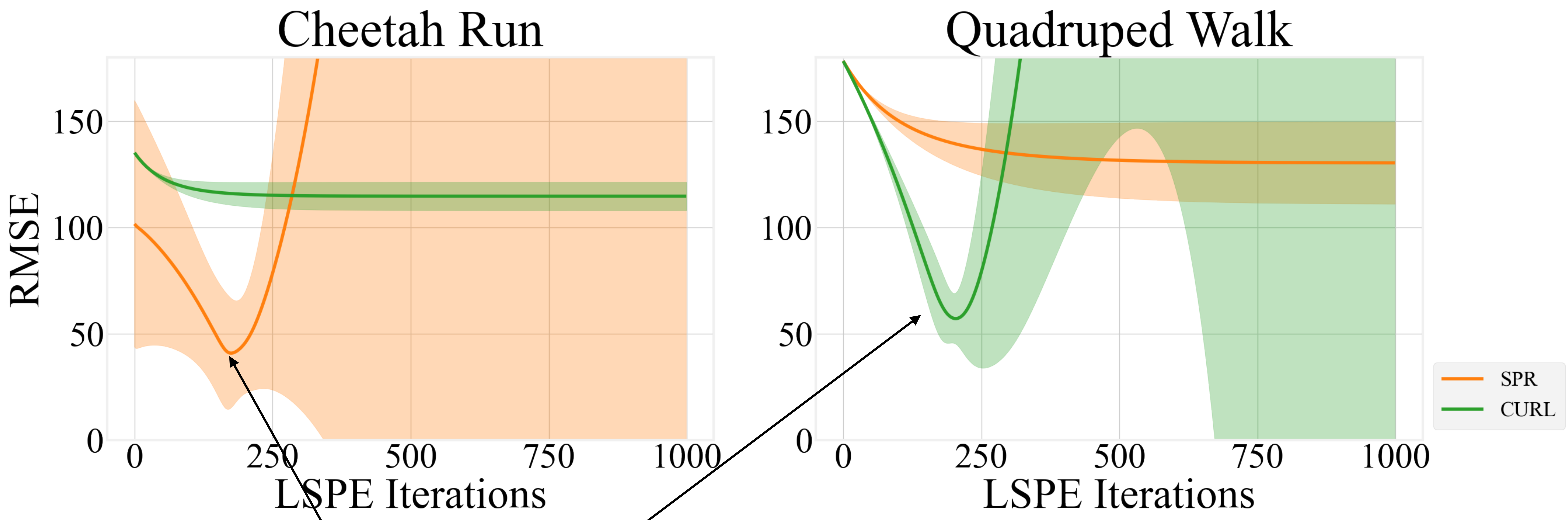




Online representation learning objectives fail in offline RL!



Online representation learning objectives fail in offline RL!



Exponential Error Amplification!

Can we provably perform good OPE on high-dimensional tasks through *Representation Learning*?

One-Step OPE

Ordinary Least Squares on
$$R(s,a) := \mathbb{E}\left[R \mid S = s, A = a\right].$$

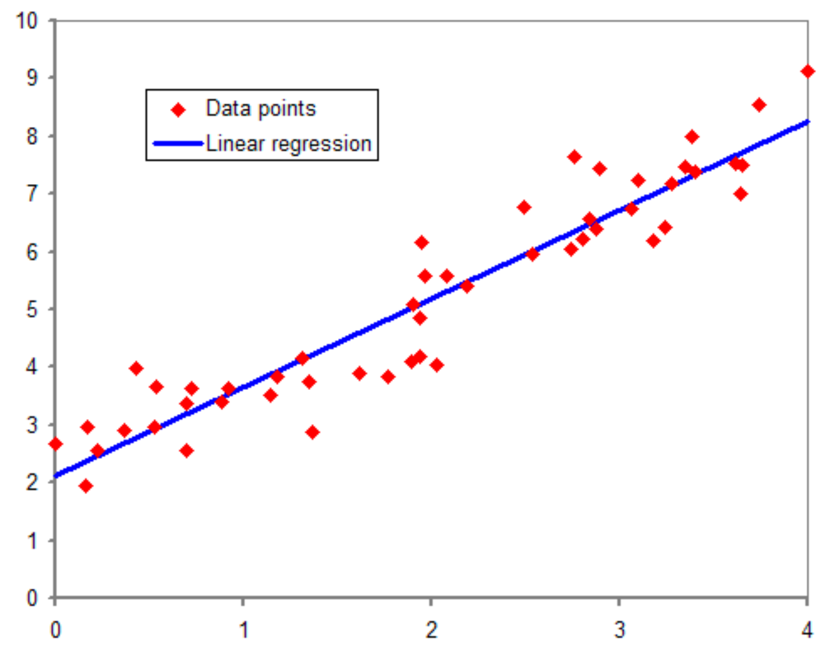
Realizability
$$R(s,a) = \phi(s,a)^T w$$

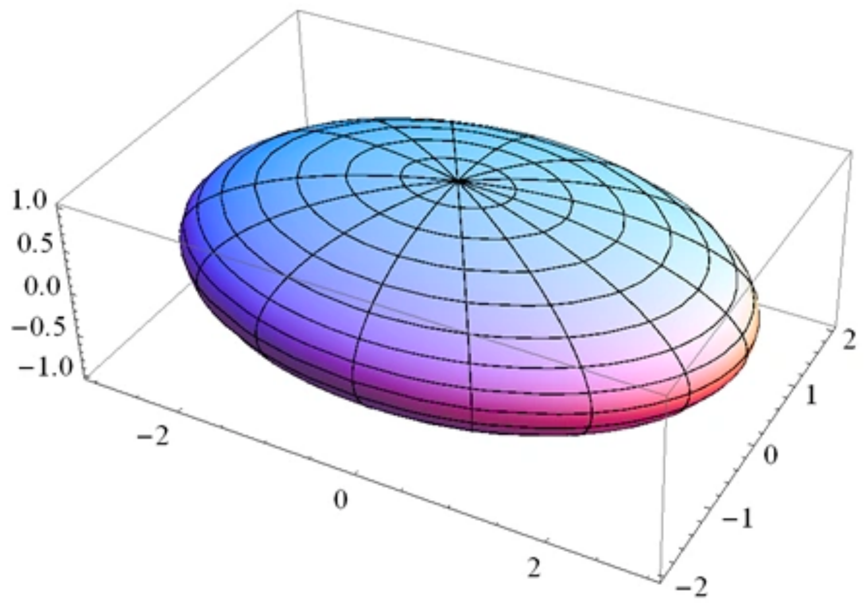
+

Coverage
$$\mathbb{E}_\nu[\phi(s,a)\phi(s,a)^T]$$

=

Sample Efficiency!



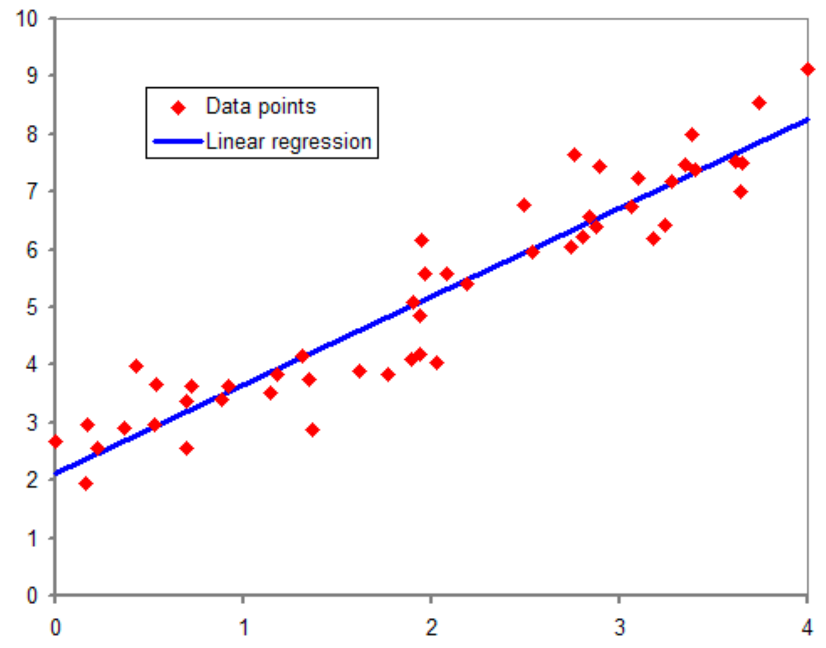


Multi-Step (RL) OPE

Estimate $Q^{\pi}(s,a) := \mathbb{E}_{\pi}\left[\sum_{h=0}^{\infty} \gamma^t r_h \mid s_0 = s, a_0 = a\right].$

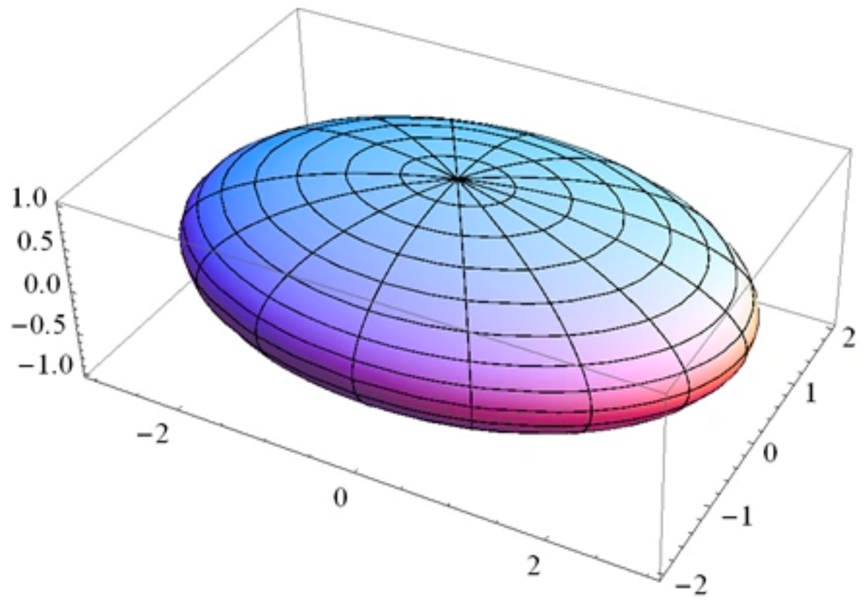
Curse of horizon!

Realizability
$Q^{\pi}(s,a) = \phi(s,a)^T w^{\pi}$



+

Coverage
$\mathbb{E}_{\nu}[\phi(s,a)\phi(s,a)^T]$



=

Information Theoretic
Lower Bound of $\Omega((d/2)^H)$

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline RL with linear function approximation? *ICLR 2021*.

Multi-Step (RL) OPE

Estimate

$$Q^\pi(s,a) := \mathbb{E}_\pi\left[\sum_{h=0}^{\infty} \gamma^t r_h \mid s_0 = s, a_0 = a\right].$$

Bellman Completeness

$f \in \mathcal{F} \implies \mathcal{T}^\pi(f) \in \mathcal{F}$

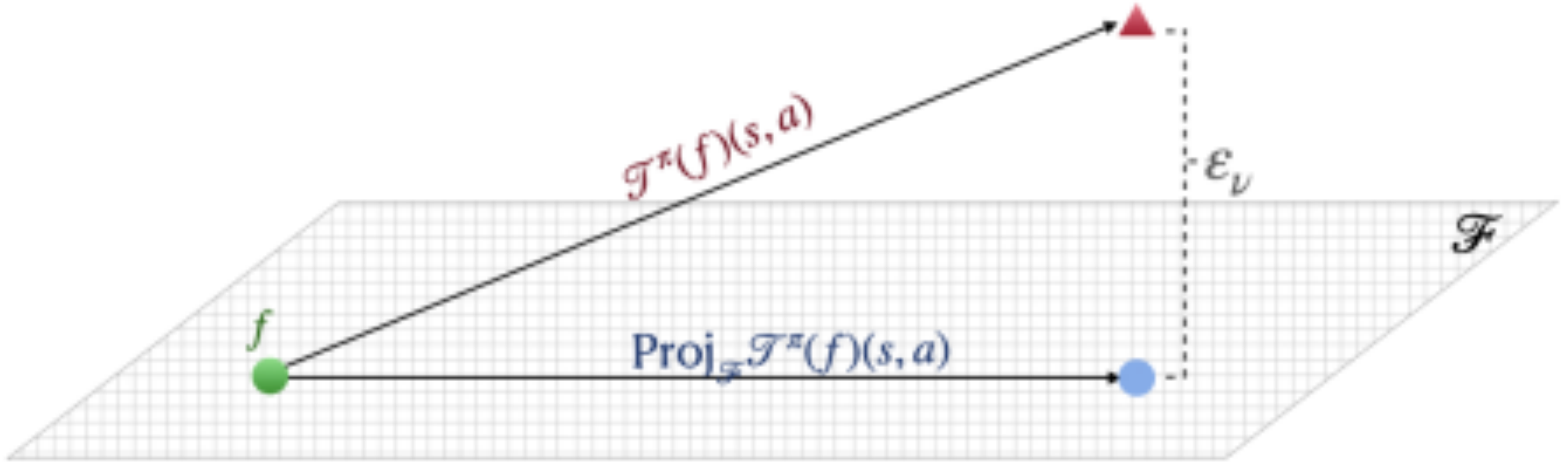
+

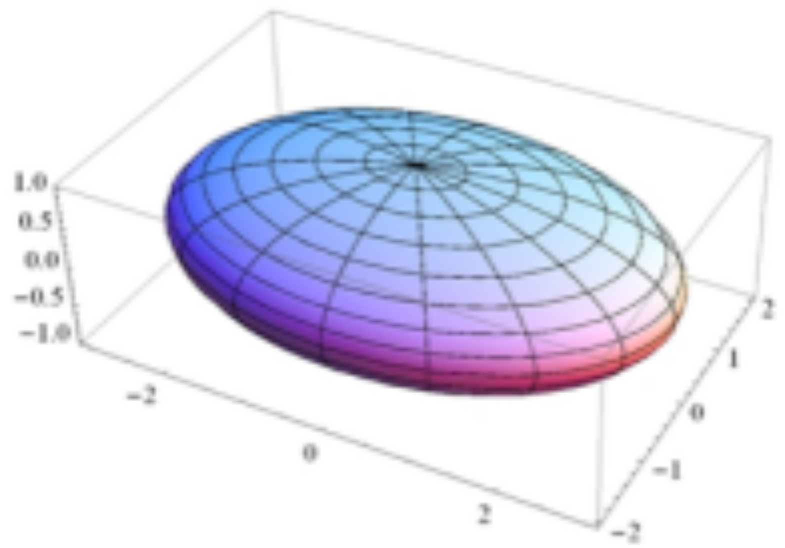
Coverage

$\mathbb{E}_\nu[\phi(s,a)\phi(s,a)^T]$

=

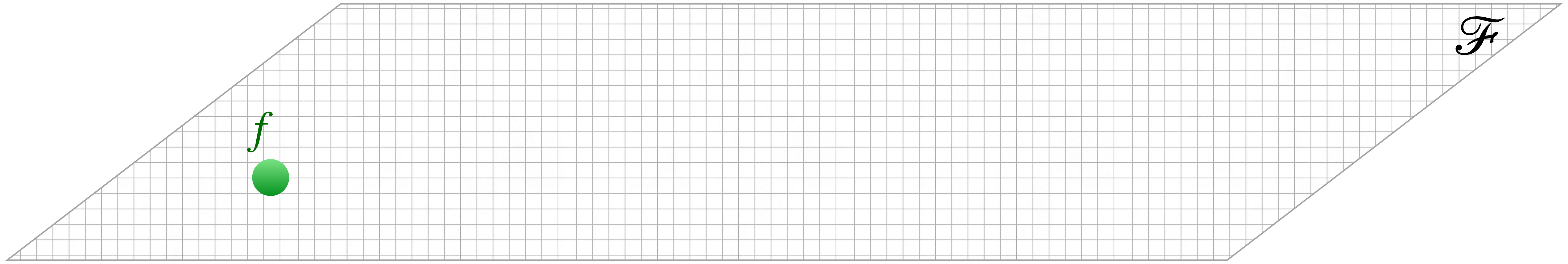
Sample Efficiency!





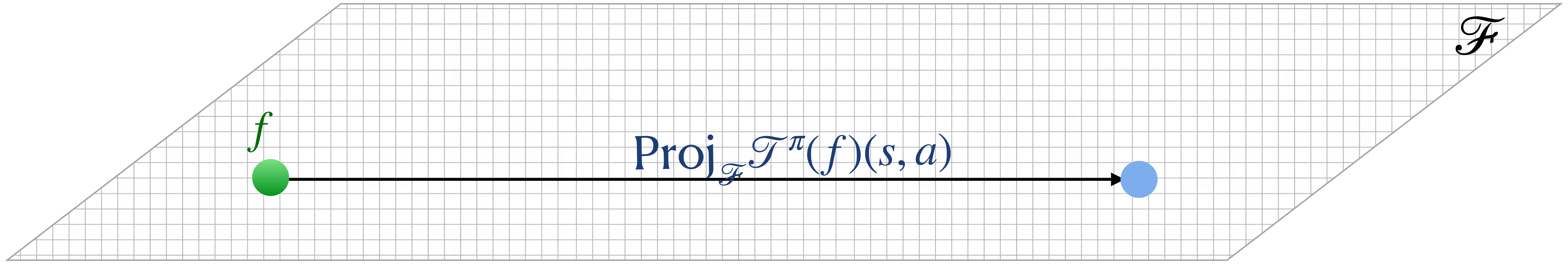
Bellman Completeness

Bellman Operator $\mathcal{T}^{\pi}(f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[f(s',\pi)]$



Bellman Completeness

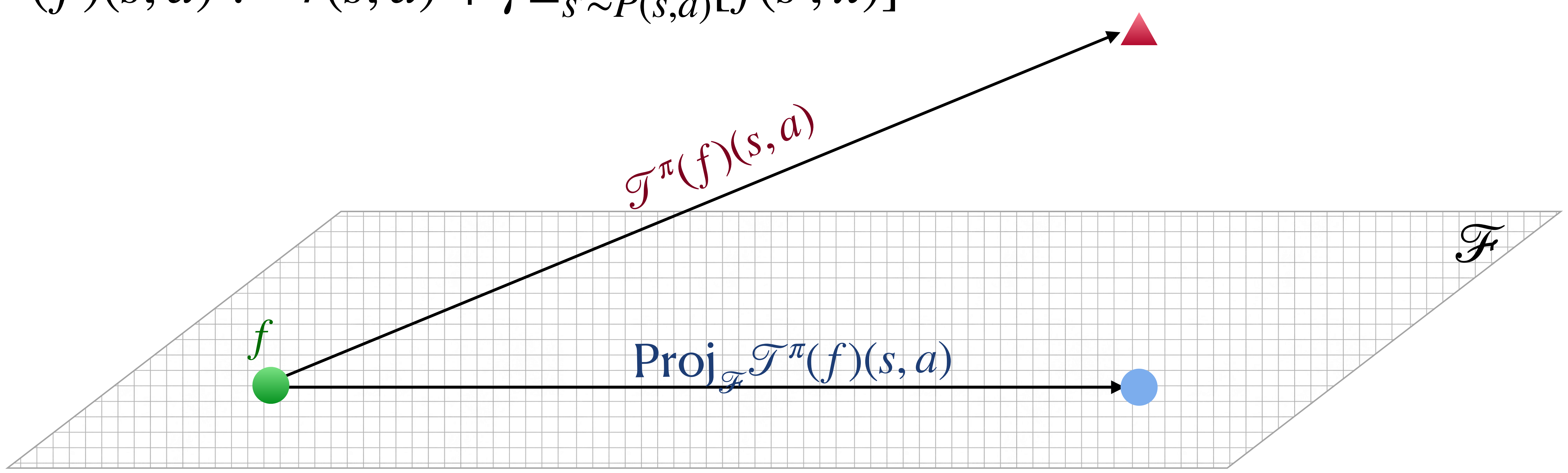
Bellman Operator $\mathcal{T}^{\pi}(f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[f(s', \pi)]$



 - If f is Bellman Complete

Bellman Completeness

Bellman Operator $\mathcal{T}^{\pi}(f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[f(s',\pi)]$

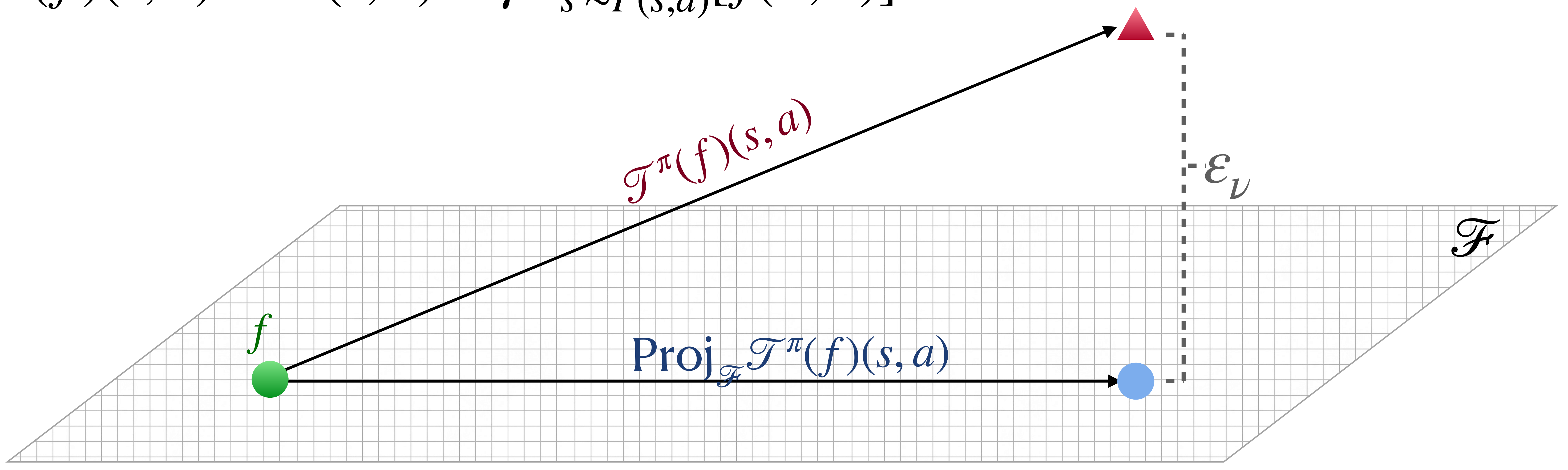


- If f is Bellman Complete

- If f is *not* Bellman Complete

Bellman Completeness

Bellman Operator $\mathcal{T}^{\pi}(f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[f(s', \pi)]$



- If f is Bellman Complete

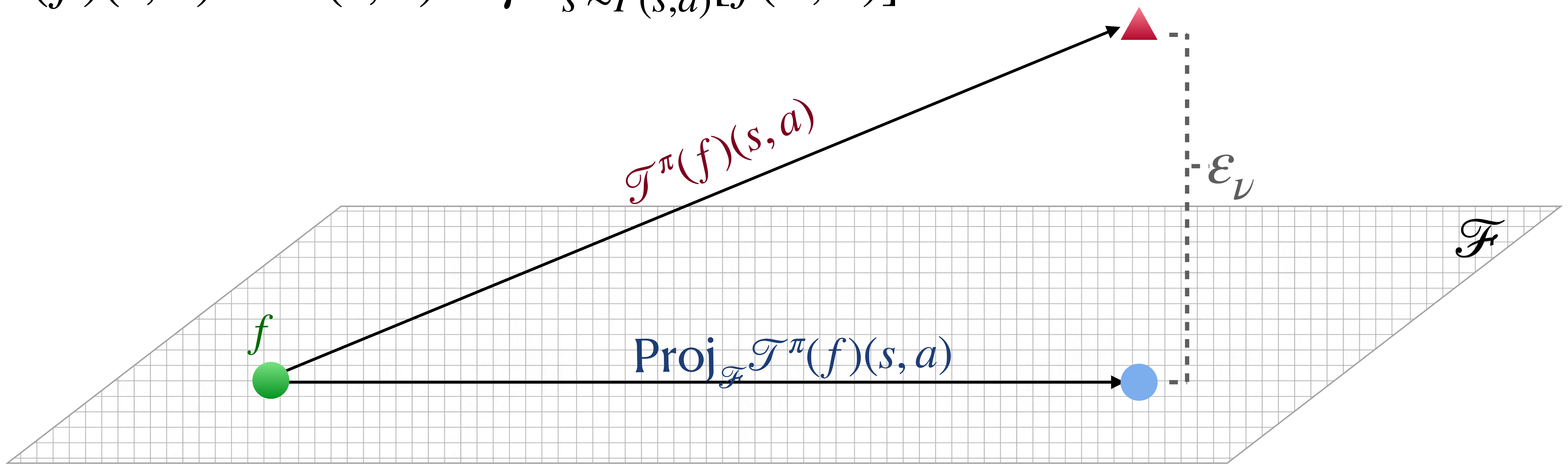
- If f is *not* Bellman Complete

Note

- If *exactly* bellman complete, $\varepsilon_{\nu} = 0$

Bellman Completeness

Bellman Operator $\mathcal{T}^{\pi}(f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[f(s',\pi)]$



- If f is Bellman Complete

- If f is *not* Bellman Complete

Note

- If *exactly* bellman complete, $\varepsilon_{\nu} = 0$

We say a representation ϕ is Linear Bellman Complete if $\mathcal{F} = \left\{\phi^T w : w \in \mathbb{R}^d\right\}$ is Bellman Complete.

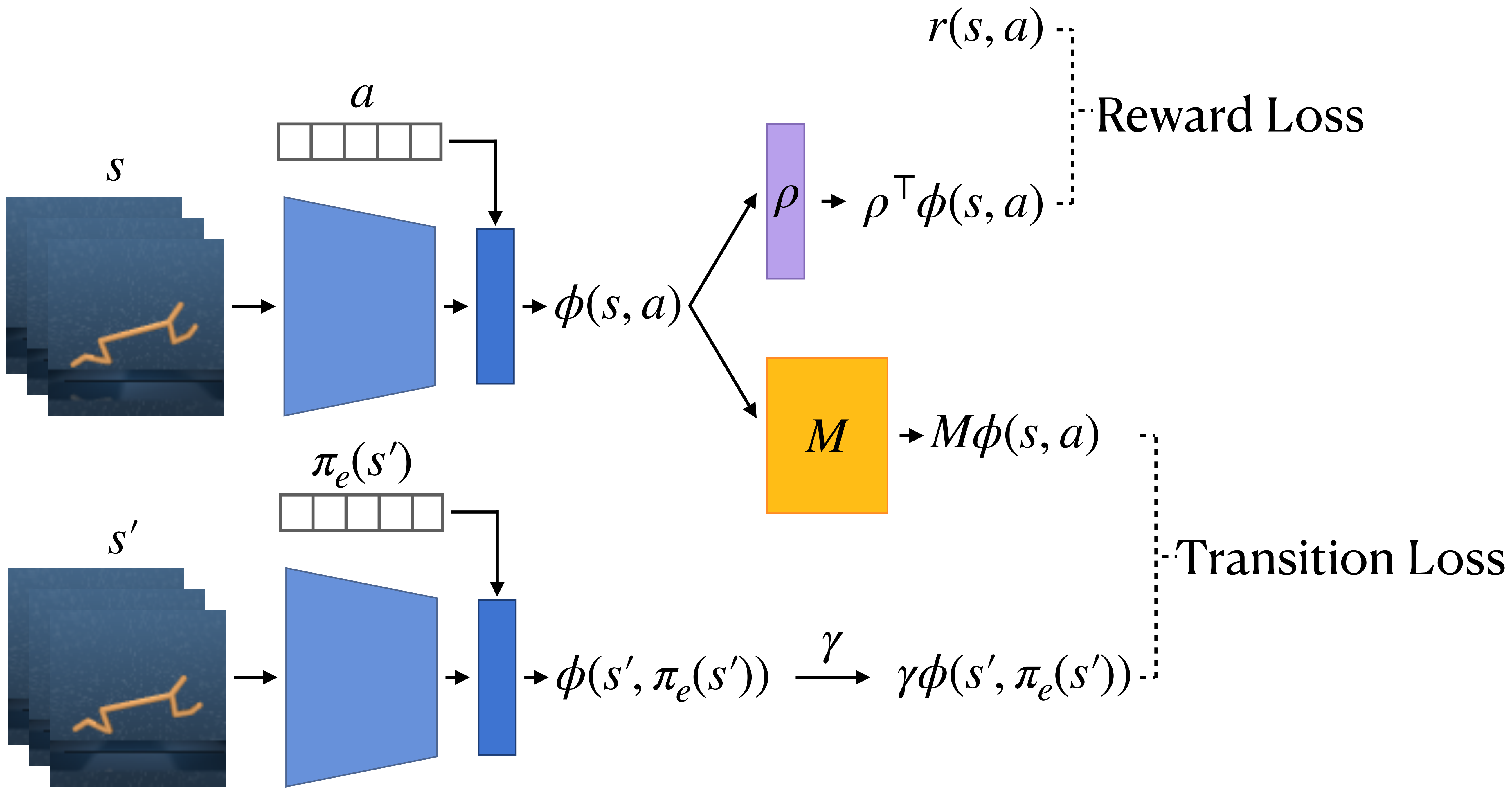
Equivalent Characterization

Linear BC is *equivalent* to, there exist $(\rho, M) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}$ so that

$$\mathbb{E}_\nu \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathbb{E}_{s' \sim P(s,a)} \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0.$$

* formal result with norm constraints on ρ, M in paper.

Equivalent Characterization

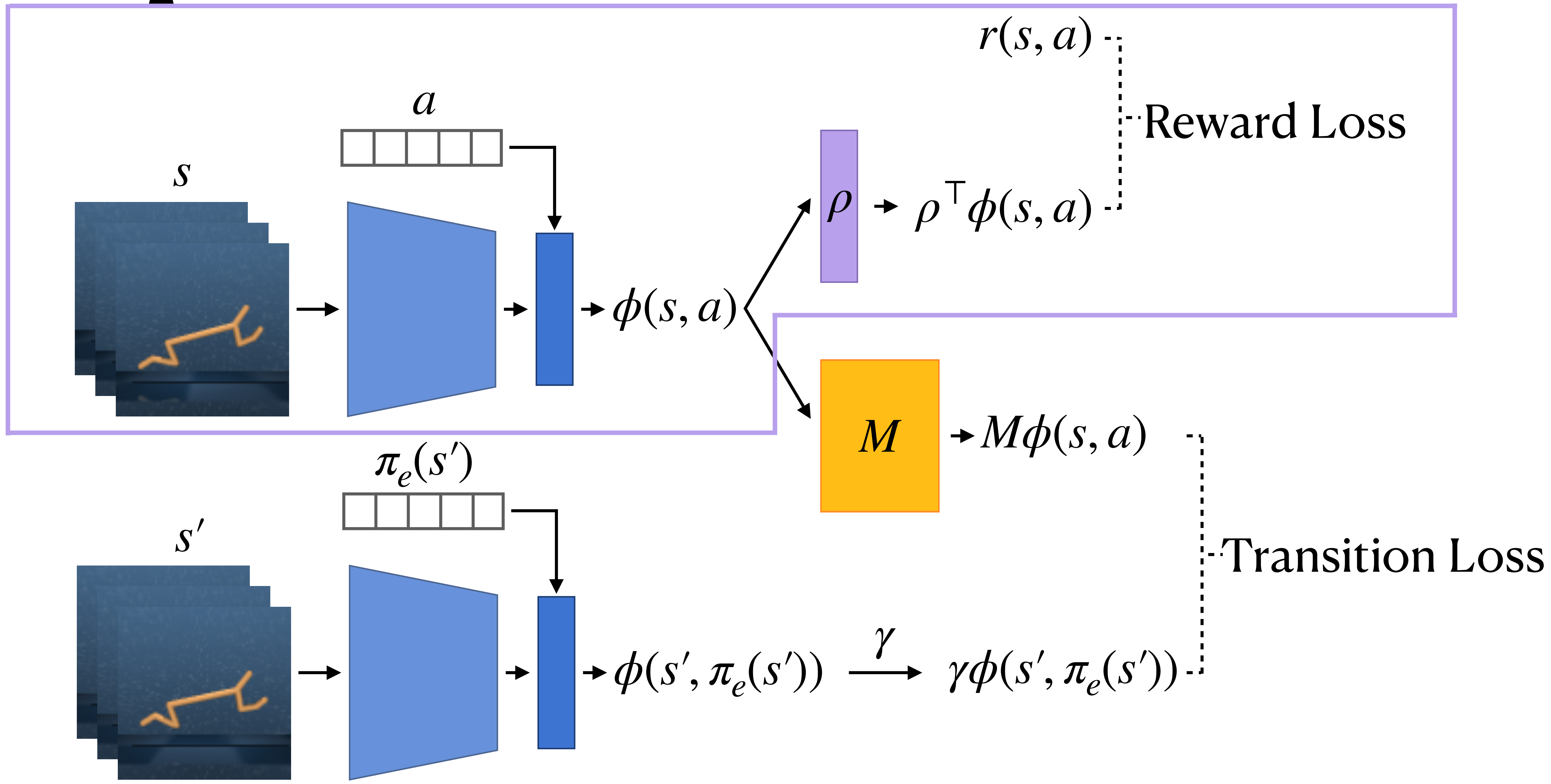


Linear BC is *equivalent* to, there exist $(\rho, M) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$ so that

$$\mathbb{E}_\nu \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathbb{E}_{s'\sim P(s,a)} \phi(s',\pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0.$$

* formal result with norm constraints on ρ, M in paper.

Equivalent Characterization

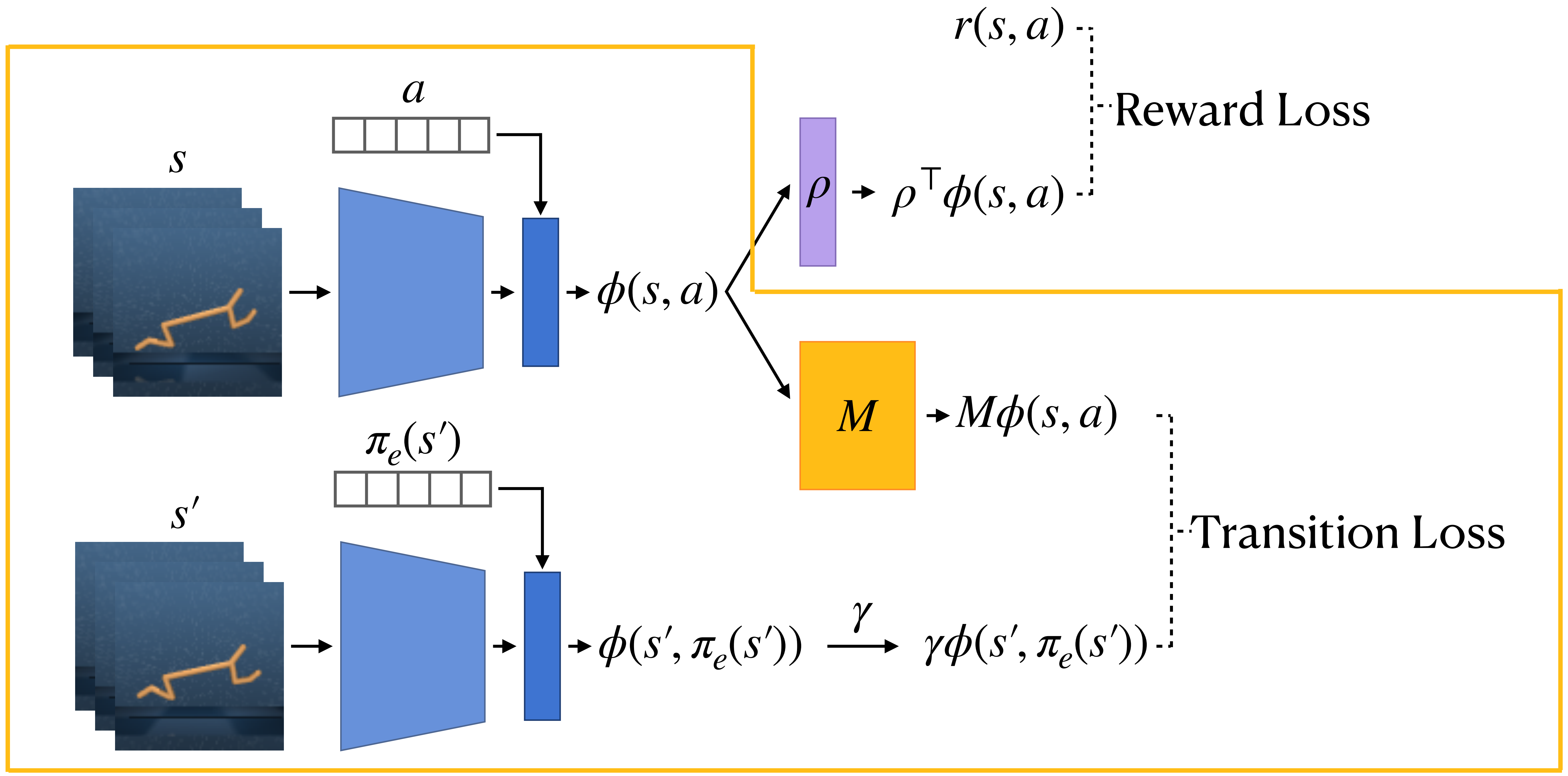


Linear BC is *equivalent* to, there exist $(\rho, M) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}$ so that

$$\mathbb{E}_\nu \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathbb{E}_{s' \sim P(s,a)} \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0.$$

* formal result with norm constraints on ρ, M in paper.

Equivalent Characterization



Linear BC is *equivalent* to, there exist $(\rho, M) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}$ so that

$$\mathbb{E}_\nu \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathbb{E}_{s' \sim P(s,a)} \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0.$$

* formal result with norm constraints on ρ, M in paper.

Learning Bellman Complete Features

with coverage

- Suppose the representation class Φ contains a Linear BC feature $\phi^\star$, with **coverage** $\lambda_{min}(\mathbb{E}_\nu[\phi^\star(s,a)\phi^\star(s,a)^T]) \geq \beta$.
- Self-supervised objective:

$$\widehat{\phi} \in \arg\min_{\phi \in \Phi} \left[\min_{(\rho, M) \in \Theta} \mathbb{E}_{\mathscr{D}} \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 \right],$$

$$\text{s.t. } \lambda_{min}\left(\mathbb{E}_{\mathscr{D}}[\phi(s,a)\phi(s,a)^T]\right) \geq \beta/2.$$

* formal result with stochastic transitions in paper.

BCRL

BCRL

1. **Learn $\widehat{\phi}$ by minimizing self-supervised Bellman Completeness loss.**

BCRL

1. Learn $\widehat{\phi}$ by minimizing self-supervised Bellman Completeness loss.
2. Run LSPE with the learned $\widehat{\phi}$.

Theory: Representation Learning

- Theorem: For any δ and large enough dataset of size N, with probability at least $1 - \delta$, we have that the ERM $\widehat{\phi}$ satisfies,

1. Approximately Linear BC, with $\varepsilon_\nu = \tilde{\mathcal{O}}\left(\dfrac{d \cdot \mathrm{comp}(\Phi)}{\sqrt{N}}\right)$,

2. Coverage, with $\lambda_{min}\left(\mathbb{E}_\nu[\,\widehat{\phi}\,(s,a)\,\widehat{\phi}\,(s,a)^T]\right) \geq \beta/4$.

* formal result with stochastic transitions in paper.

Theory: End-to-End OPE Guarantee

- Theorem: For any δ and large enough dataset of size N, with probability at least $1-\delta$, BCRL with K iterates of LSPE evaluates well **for any distribution p_0**,

$$\left| V_{p_0}^{\pi_e} - \mathbb{E}_{s\sim p_0}[\hat{f}_K(s,\pi_e)] \right| \in \tilde{\mathcal{O}}\left(\frac{\gamma^{K/2}}{(1-\gamma)} + \frac{\sqrt{\left\| \frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu} \right\|_\infty}}{(1-\gamma)^2} \cdot \varepsilon_\nu + \frac{d}{(1-\gamma)^2\sqrt{\beta N}} \right).$$

* formal result with stochastic transitions in paper.

Theory: End-to-End OPE Guarantee

- Theorem: For any δ and large enough dataset of size N, with probability at least $1-\delta$, BCRL with K iterates of LSPE evaluates well **for any distribution p_0**,

$$\left| V_{p_0}^{\pi_e} - \mathbb{E}_{s\sim p_0}[\hat{f}_K(s,\pi_e)] \right| \in \tilde{\mathcal{O}}\left(\frac{\gamma^{K/2}}{(1-\gamma)} + \frac{\sqrt{\left\| \frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu} \right\|_\infty}}{(1-\gamma)^2} \cdot \varepsilon_\nu + \frac{d}{(1-\gamma)^2\sqrt{\beta N}} \right).$$

Exponentially decaying in num. LSPE iterations K

* formal result with stochastic transitions in paper.

Theory: End-to-End OPE Guarantee

- Theorem: For any δ and large enough dataset of size N, with probability at least $1-\delta$, BCRL with K iterates of LSPE evaluates well **for any distribution p_0**,

$$\left| V^{\pi_e}_{p_0} - \mathbb{E}_{s\sim p_0}[\hat{f}_K(s,\pi_e)] \right| \in \tilde{\mathcal{O}}\left(\frac{\gamma^{K/2}}{(1-\gamma)} + \frac{\sqrt{\left\| \frac{\mathrm{d}d^{\pi_e}_{p_0}}{\mathrm{d}\nu} \right\|_\infty}}{(1-\gamma)^2} \cdot \varepsilon_\nu + \frac{d}{(1-\gamma)^2\sqrt{\beta N}} \right).$$

Exponentially decaying in num. LSPE iterations K

Non-Linear BC part bounded by density ratio

* formal result with stochastic transitions in paper.

Theory: End-to-End OPE Guarantee

- Theorem: For any δ and large enough dataset of size N, with probability at least $1-\delta$, BCRL with K iterates of LSPE evaluates well **for any distribution p_0**,

$$\left| V_{p_0}^{\pi_e} - \mathbb{E}_{s\sim p_0}[\hat{f}_K(s,\pi_e)] \right| \in \tilde{\mathcal{O}}\left(\frac{\gamma^{K/2}}{(1-\gamma)} + \frac{\sqrt{\left\| \frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu} \right\|_\infty}}{(1-\gamma)^2} \cdot \varepsilon_\nu + \frac{d}{(1-\gamma)^2\sqrt{\beta N}} \right).$$

Exponentially decaying in num. LSPE iterations K

Non-Linear BC part bounded by density ratio

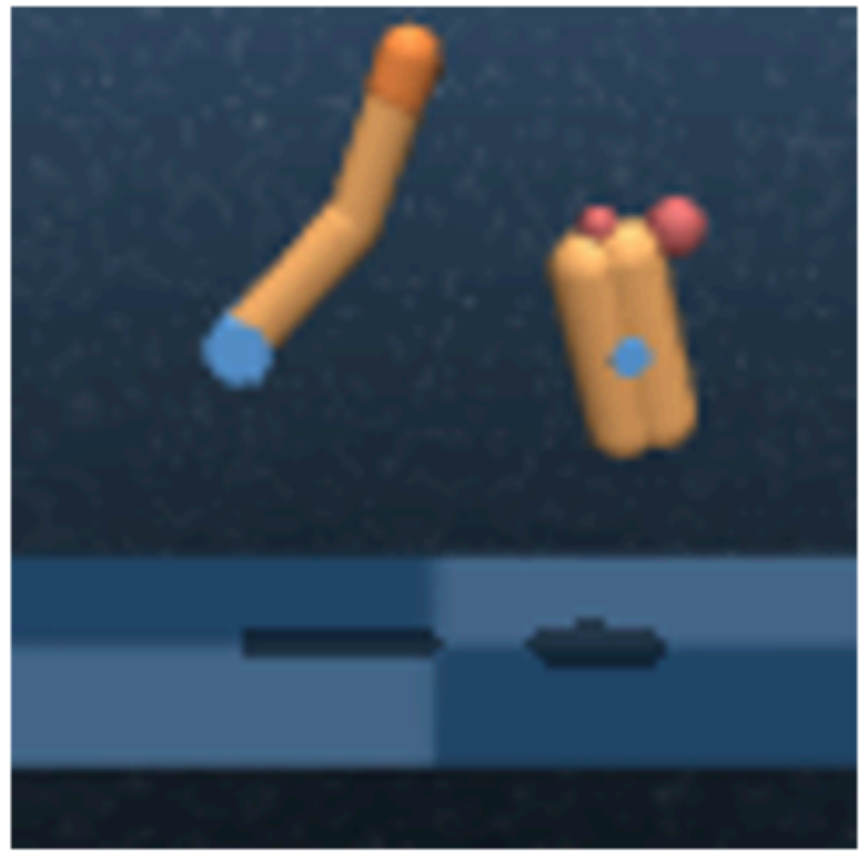
Statistical error from evaluation, converging to zero as N grows

* formal result with stochastic transitions in paper.

Experiments

Setup

DeepMind Control Suite

4 **Image Based** Continuous Control Tasks

Experiments

Setup

DeepMind Control Suite

4 **Image Based** Continuous Control Tasks

Offline Datasets

Task	Target performance	Behavior Performance
Finger Turn Hard	927	226 (24%)
Cheetah Run	758	192 (25%)
Quadruped Walk	873	236 (27%)
Humanoid Stand	827	277 (33%)

Experiments

Setup

DeepMind Control Suite

4 **Image Based** Continuous Control Tasks

Offline Datasets

Task	Target performance	Behavior Performance
Finger Turn Hard	927	226 (24%)
Cheetah Run	758	192 (25%)
Quadruped Walk	873	236 (27%)
Humanoid Stand	827	277 (33%)

Offline DB: 100K (~200 Trajectories)

Example Trajectories: Cheetah Run

Behavior Policy

Train on this ...

Example Trajectories: Cheetah Run

Behavior Policy

Train on this …

Example Trajectories: Cheetah Run

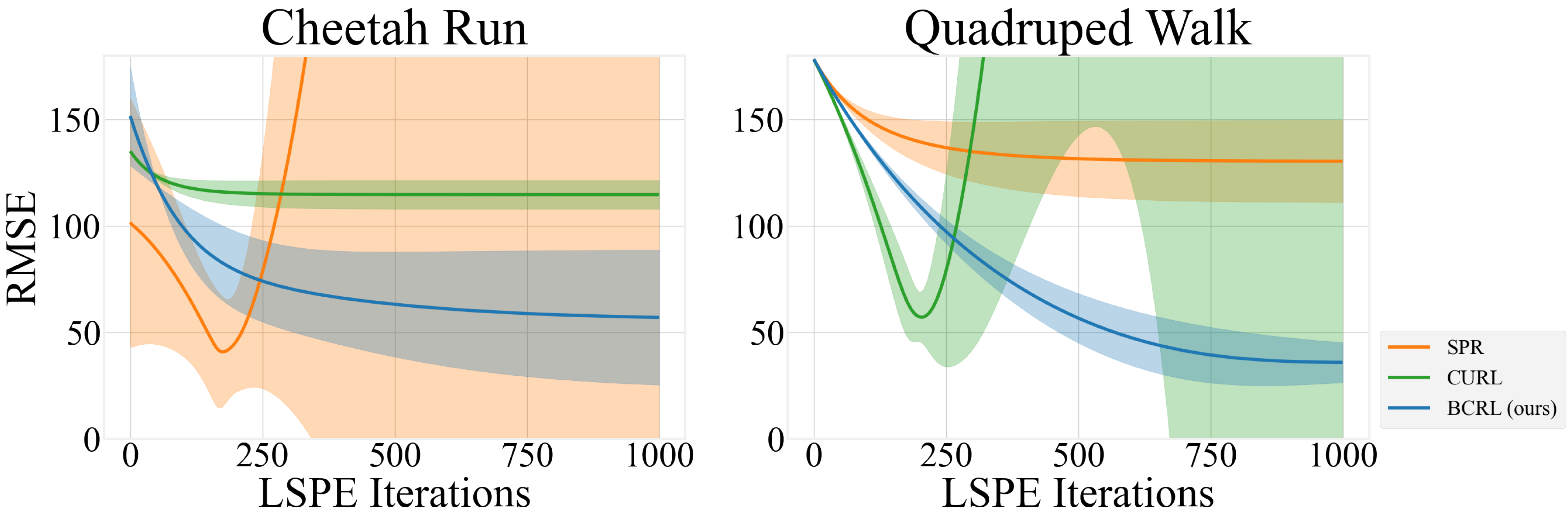
Behavior Policy

Train on this ...

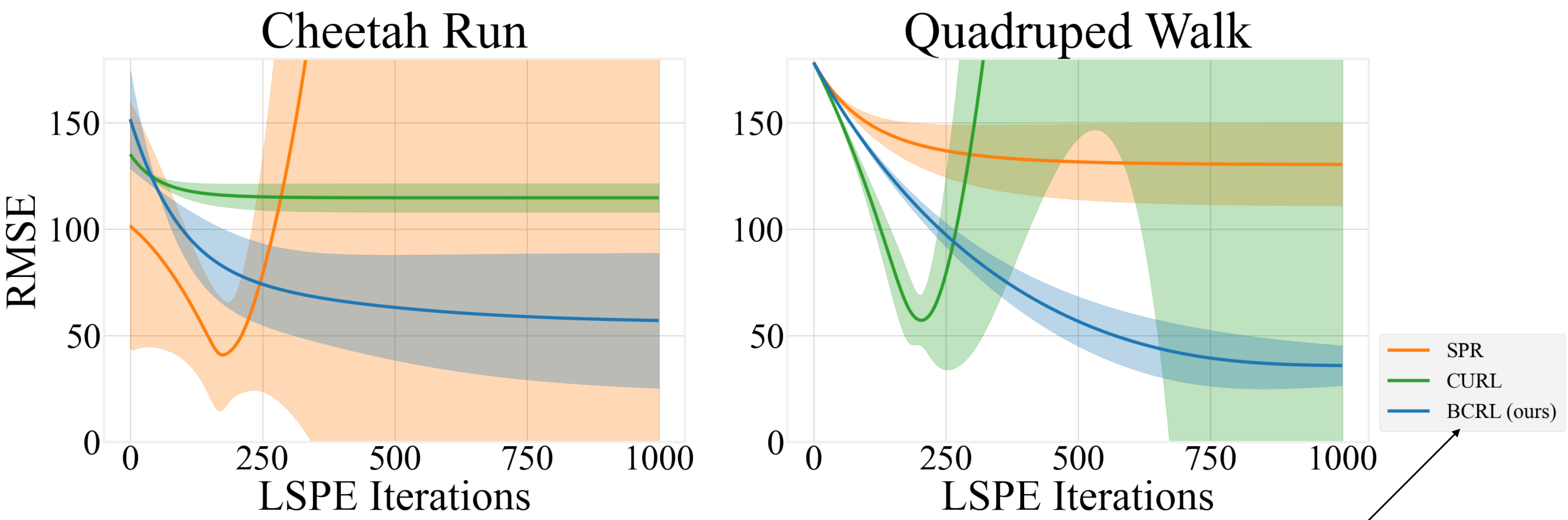
Target Policy

... to evaluate this

How is BCRL as a Representation?

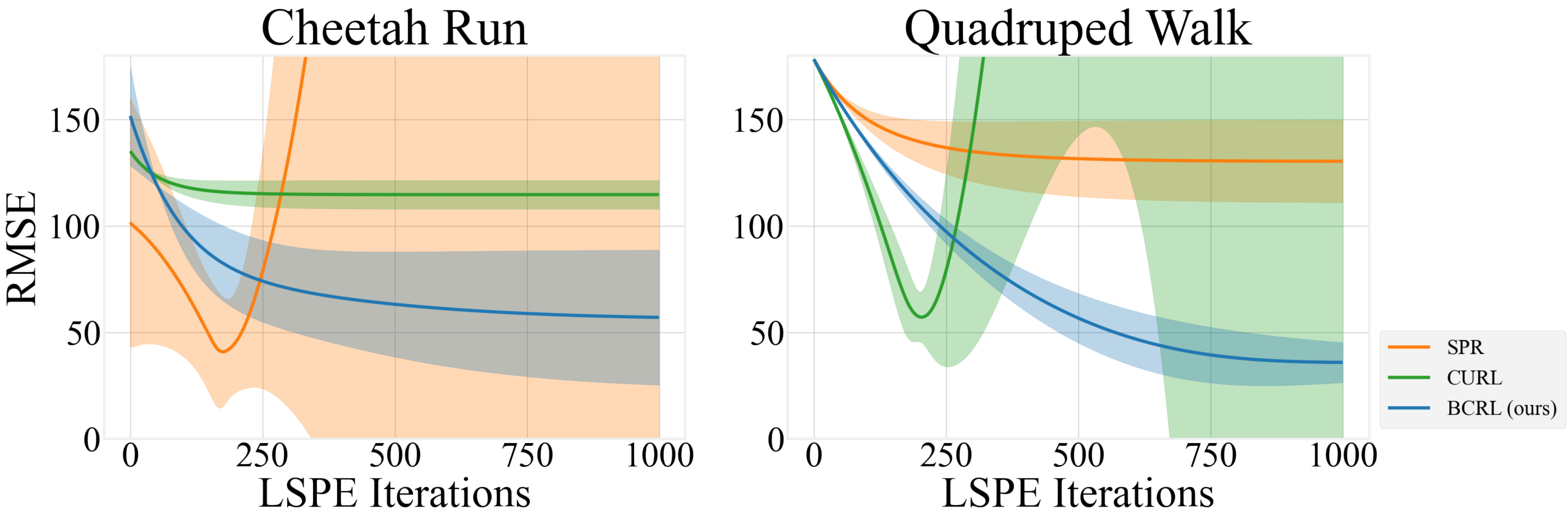


How is BCRL as a Representation?



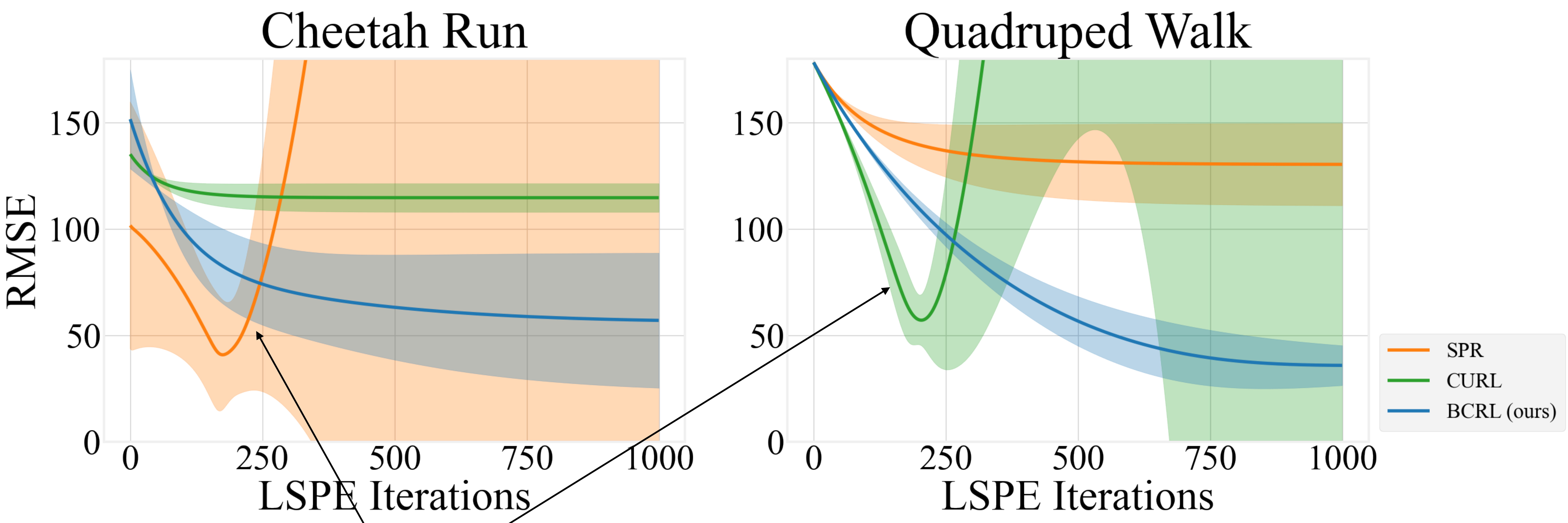
Representation Learning Comparison:
CURL and SPR

How is BCRL as a Representation?



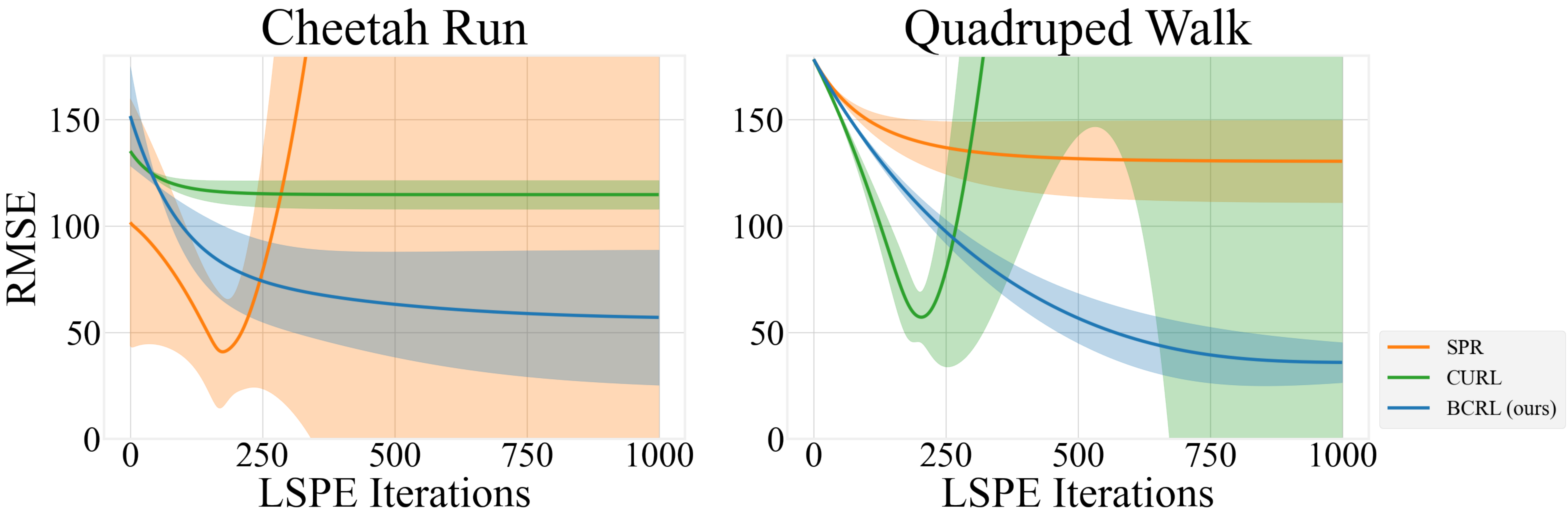
LSPE Iterations with Fixed Trained Representations

How is BCRL as a Representation?



Exponential Error Amplification!

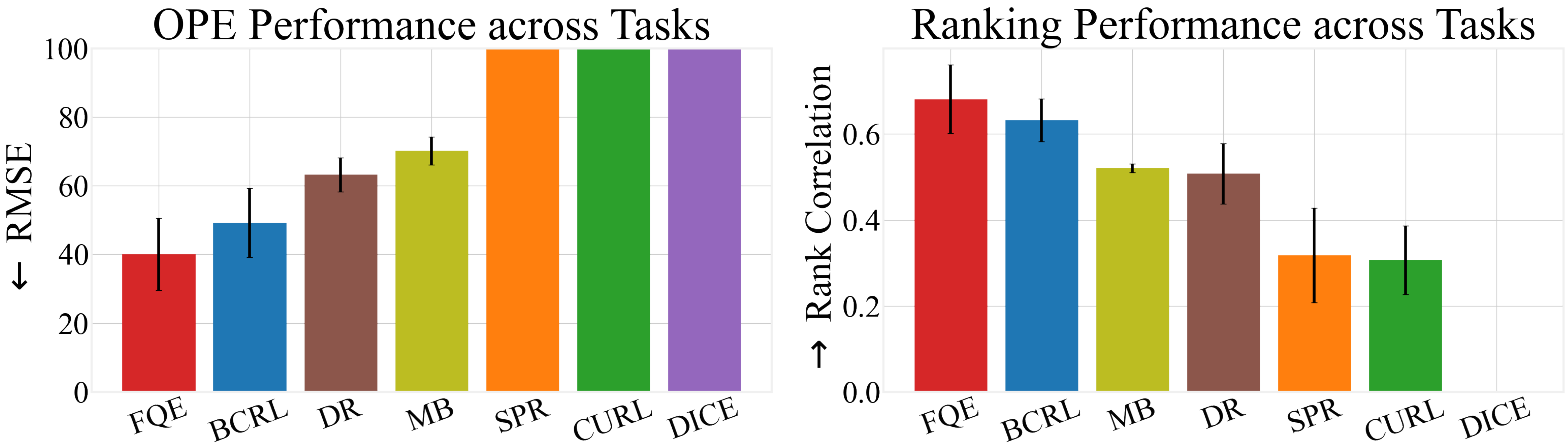
How is BCRL as a Representation?



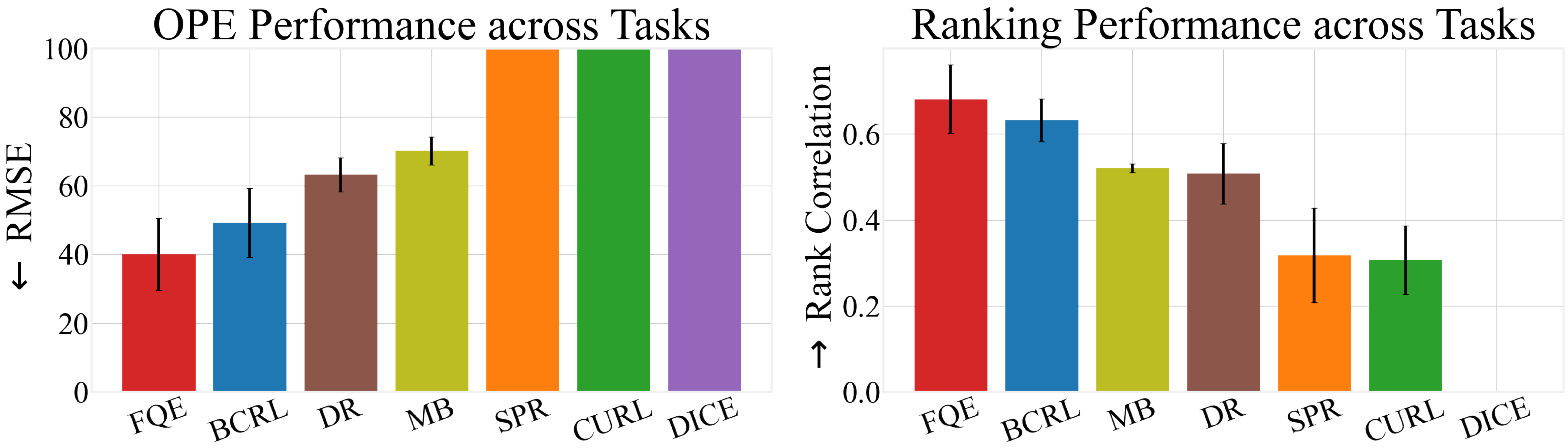
Takeaway

- Representations learned by BCRL outperforms baselines
- BCRL does not exhibit exponential error amplification in *any* task

OPE Performance



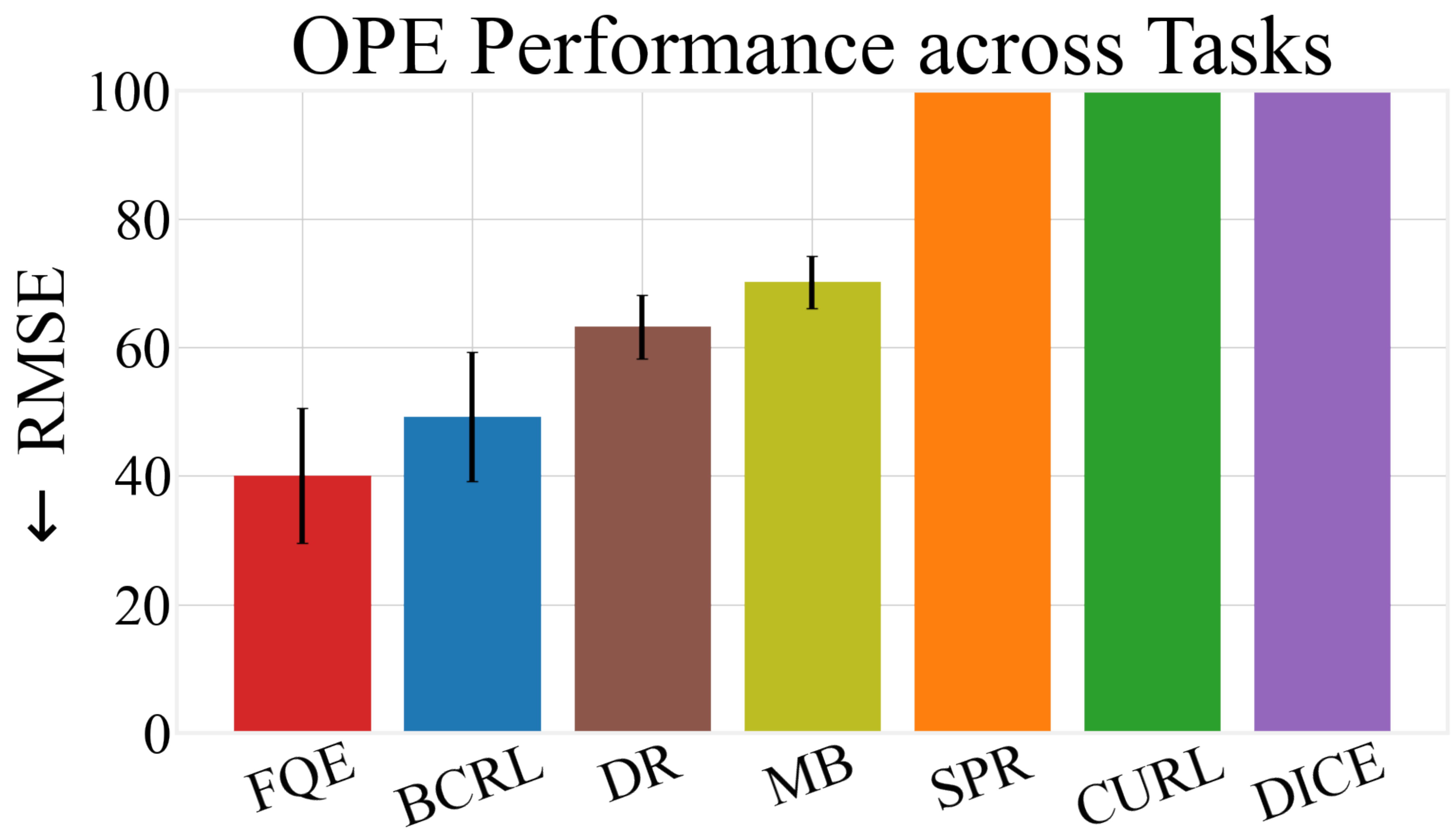
OPE Performance

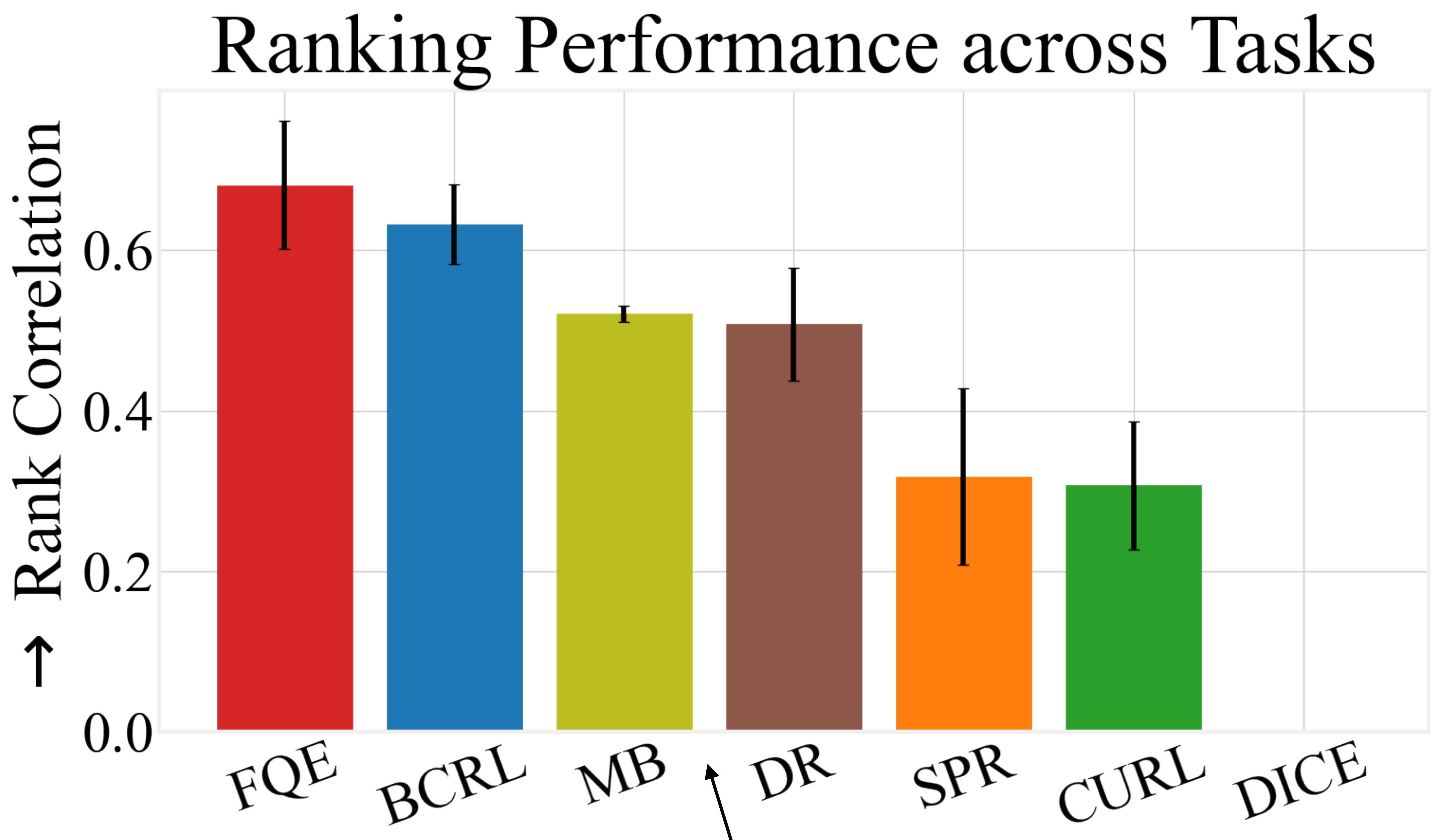


Additional Baselines

- Fitted Q-Evaluation (FQE)
- Doubly Robust Estimator (DR)
- Dreamer-v2 (Model-Based, MB)
- Distribution Correction Estimator (DICE)

OPE Performance



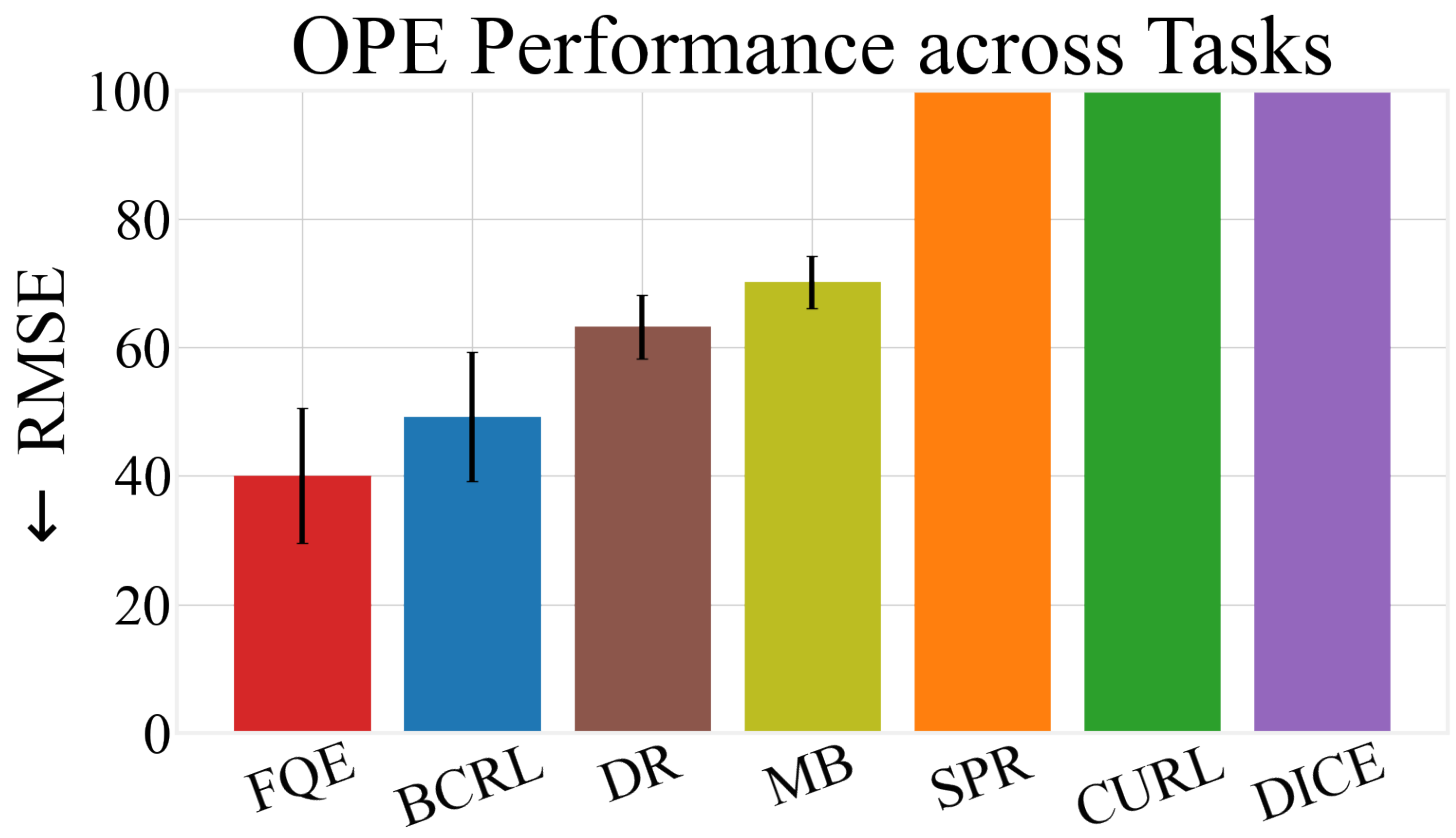


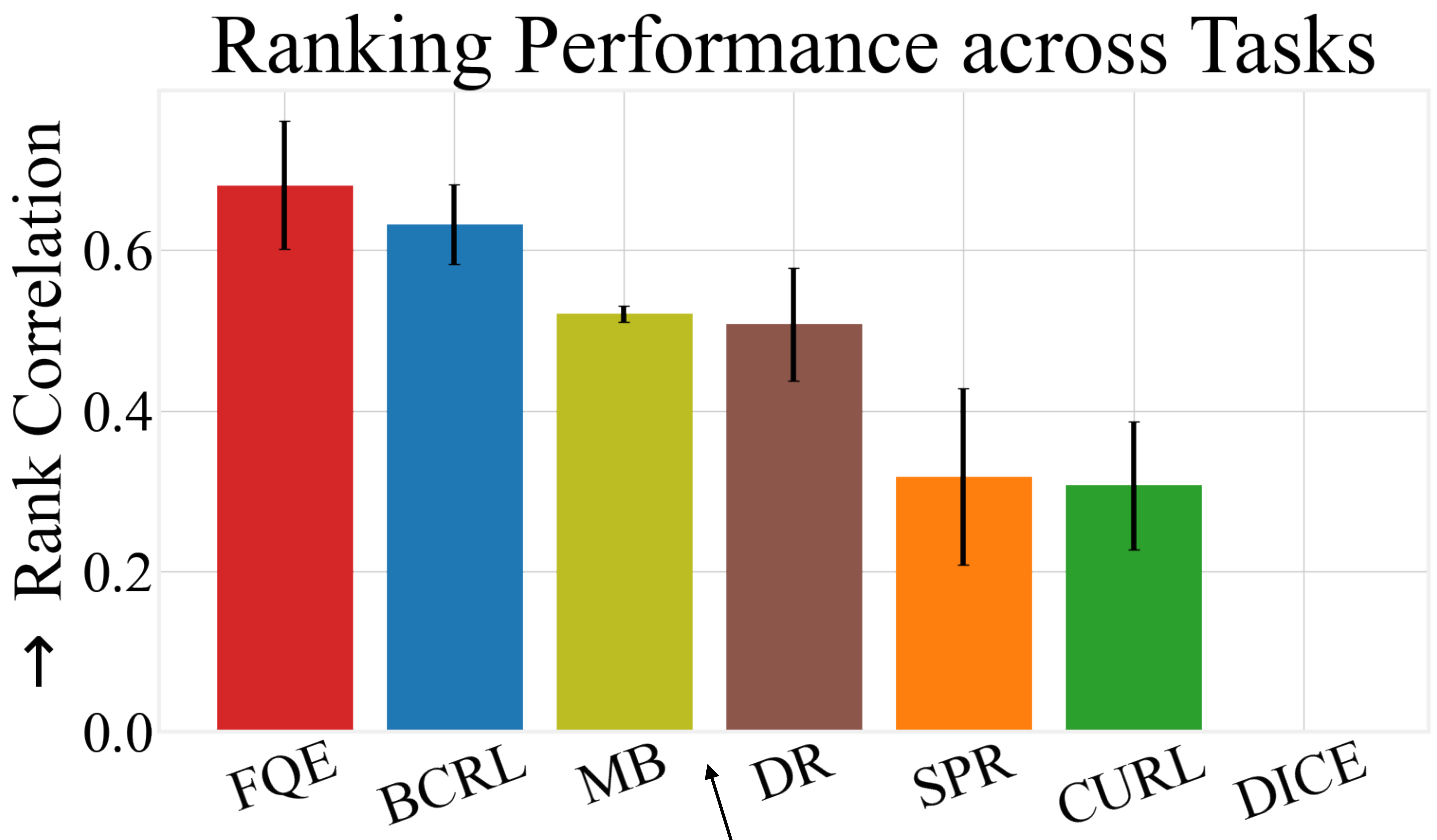
Additional Baselines

- Fitted Q-Evaluation (FQE)
- Doubly Robust Estimator (DR)
- Dreamer-v2 (Model-Based, MB)
- Distribution Correction Estimator (DICE)

Spearman ranking correlation on
4 different (~10%, ~50%, ~75%, 100%) target policies

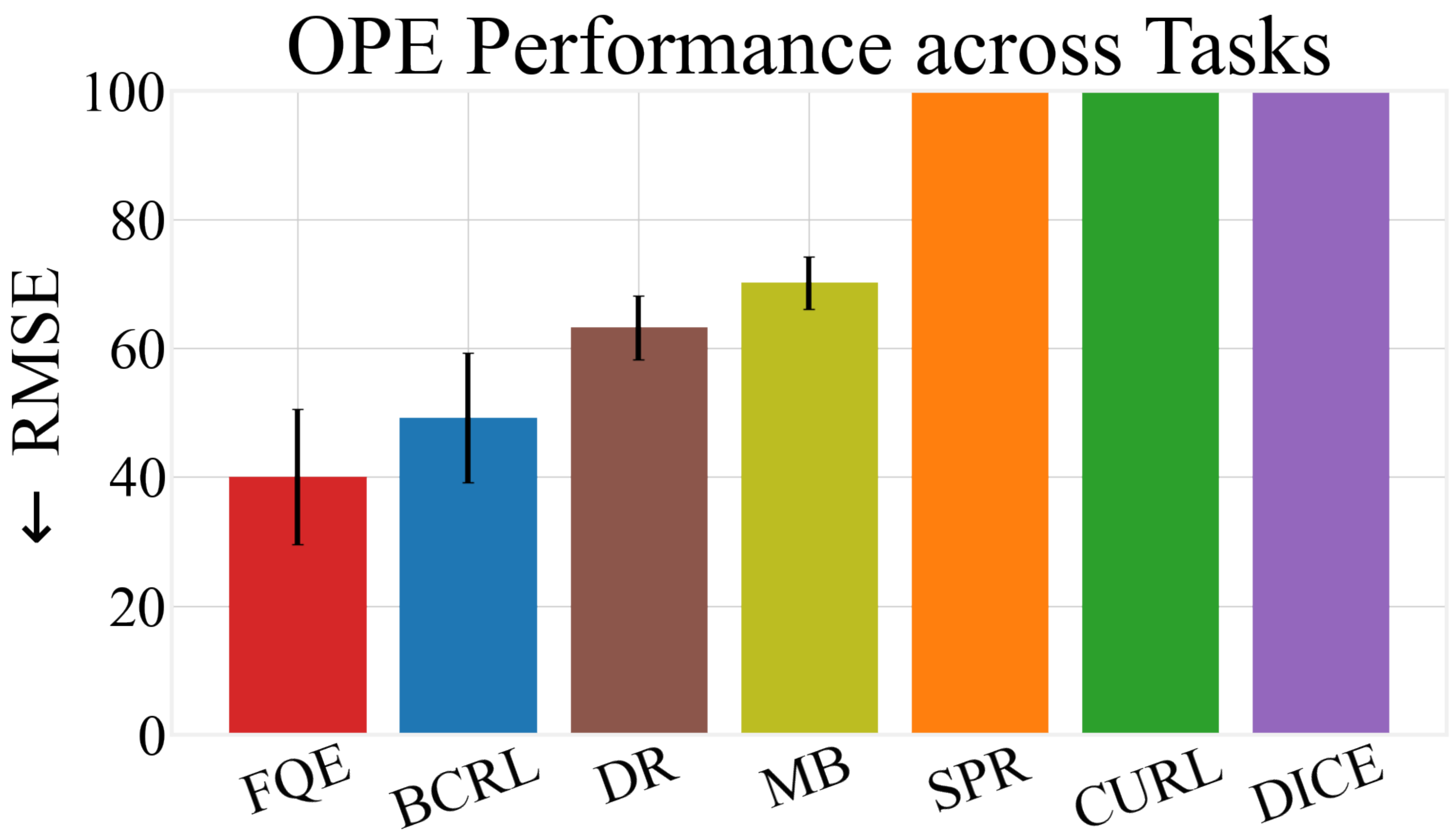
OPE Performance

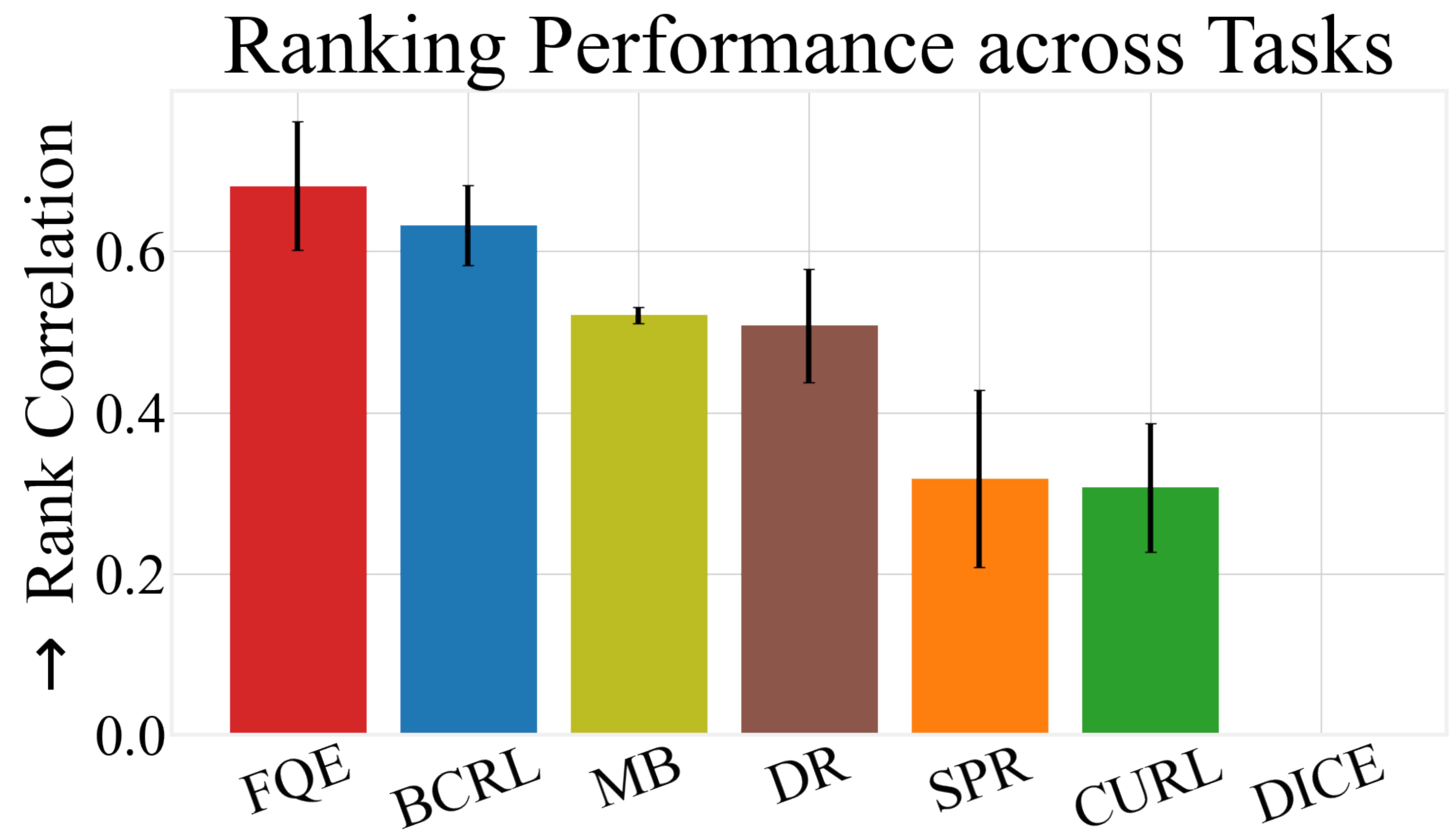




Spearman ranking correlation on
4 different (~10%, ~50%, ~75%, 100%) target policies

OPE Performance



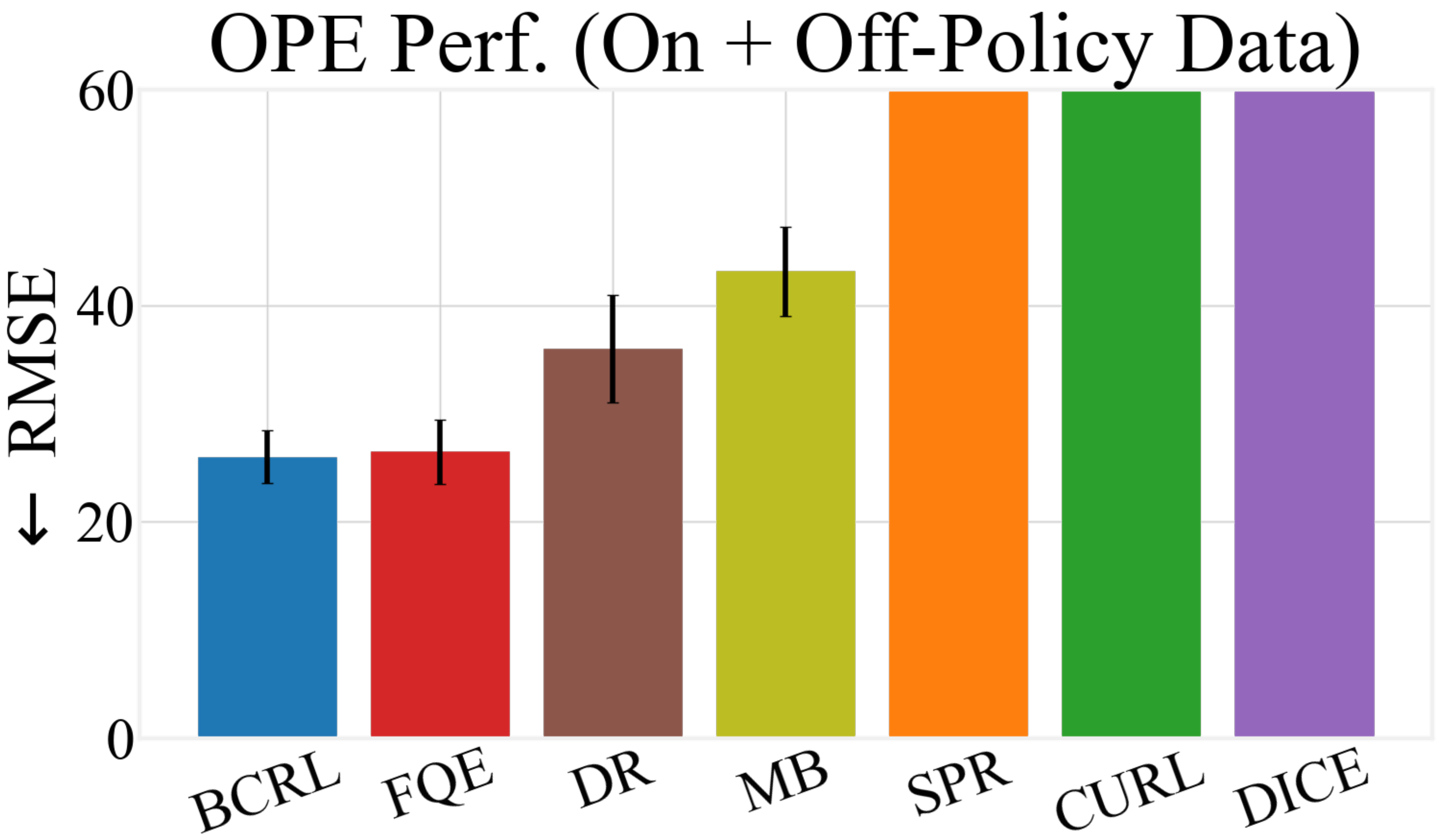


Takeaway

- BCRL is competitive with FQE and outperforms other OPE baselines

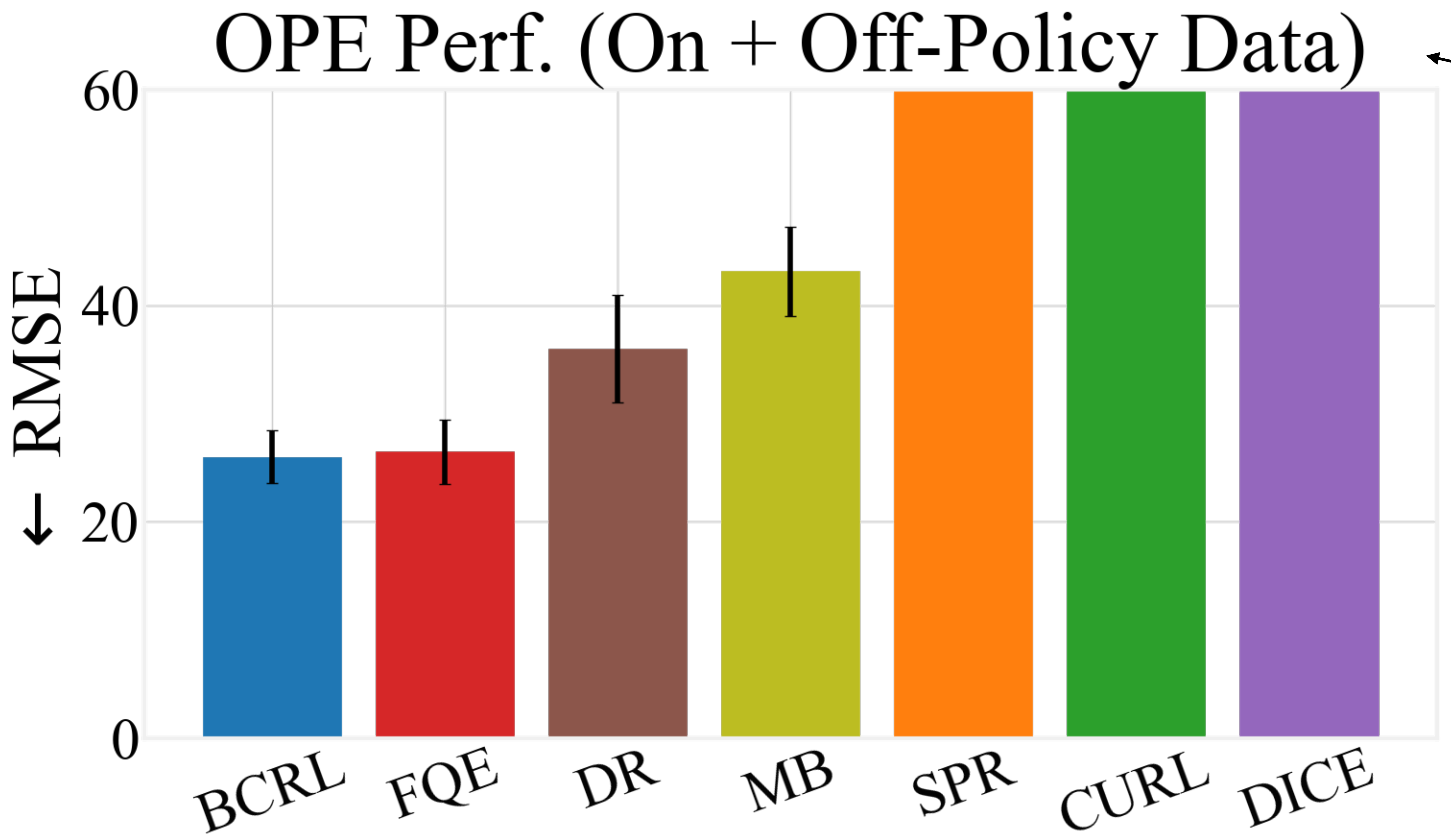
OPE Performance with On + Off policy Data

Testing Bellman Completeness



OPE Performance with On + Off policy Data

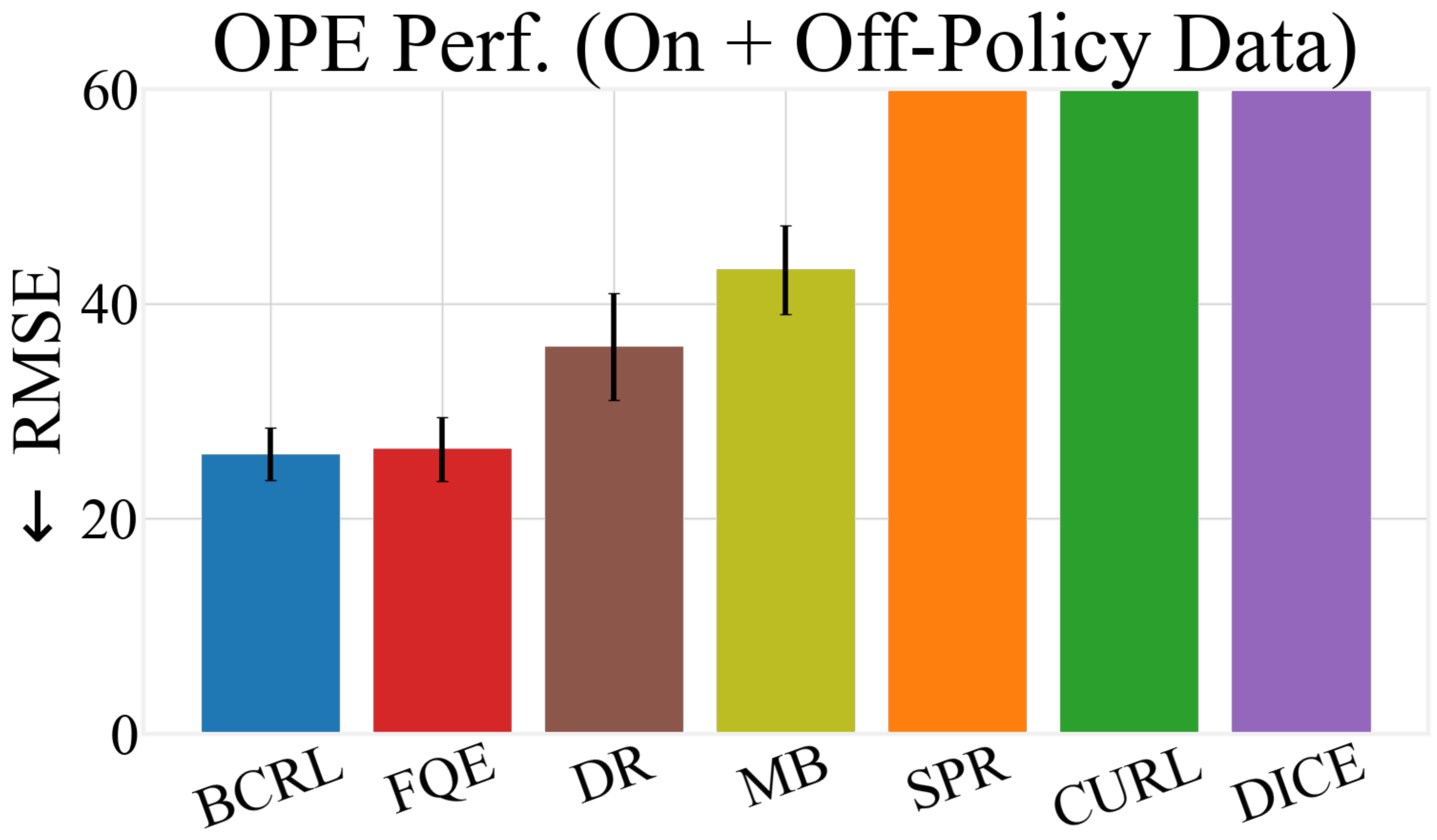
Testing Bellman Completeness



100K off-policy + 100K on-policy samples

OPE Performance with On + Off policy Data

Testing Bellman Completeness



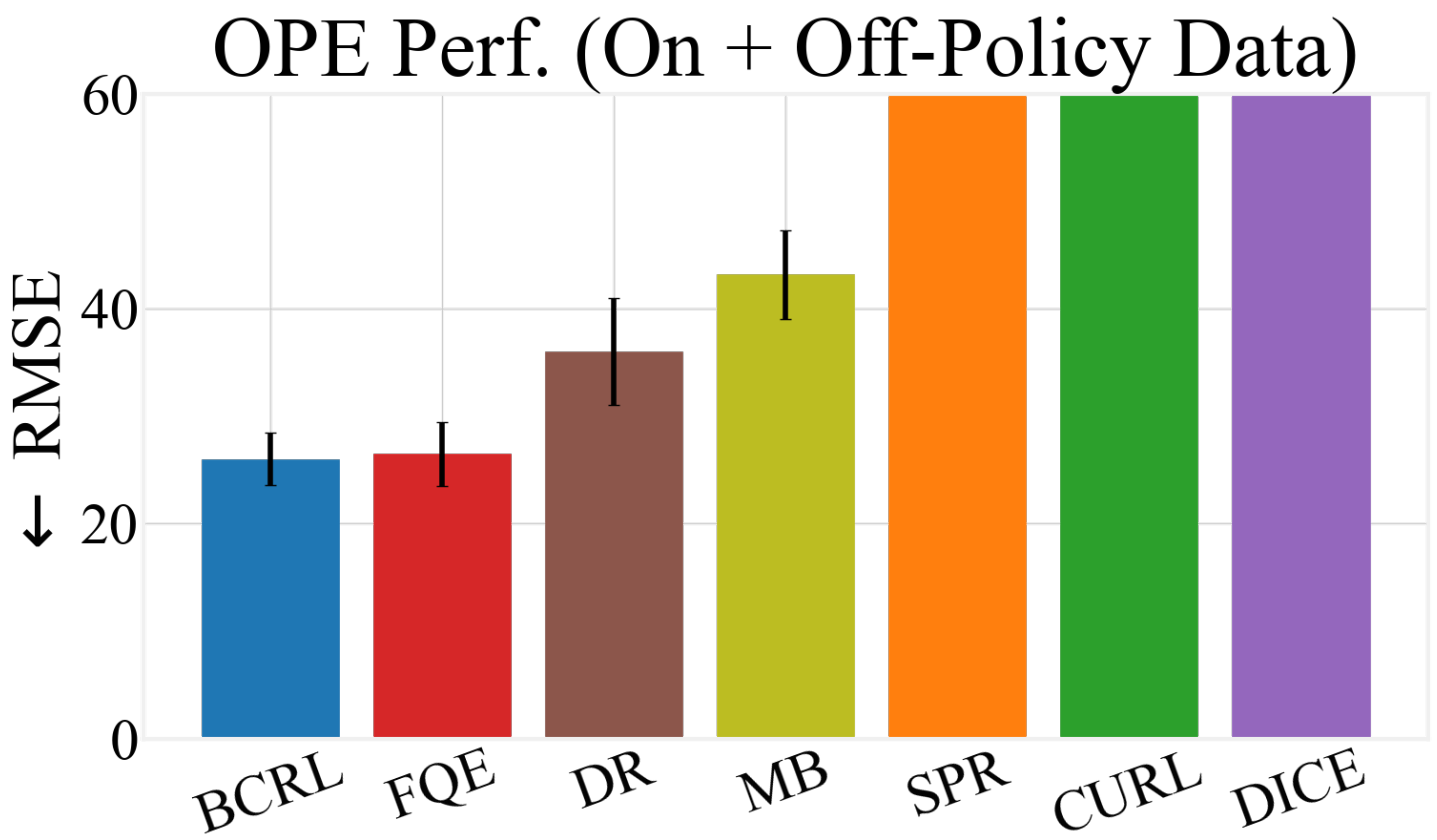
NOTE:
Adding on-policy ensures offline data **coverage over target policy** for all baselines

↓

Demonstrate the **unique benefit** of learning Bellman complete representations!

OPE Performance with On + Off policy Data

Testing Bellman Completeness

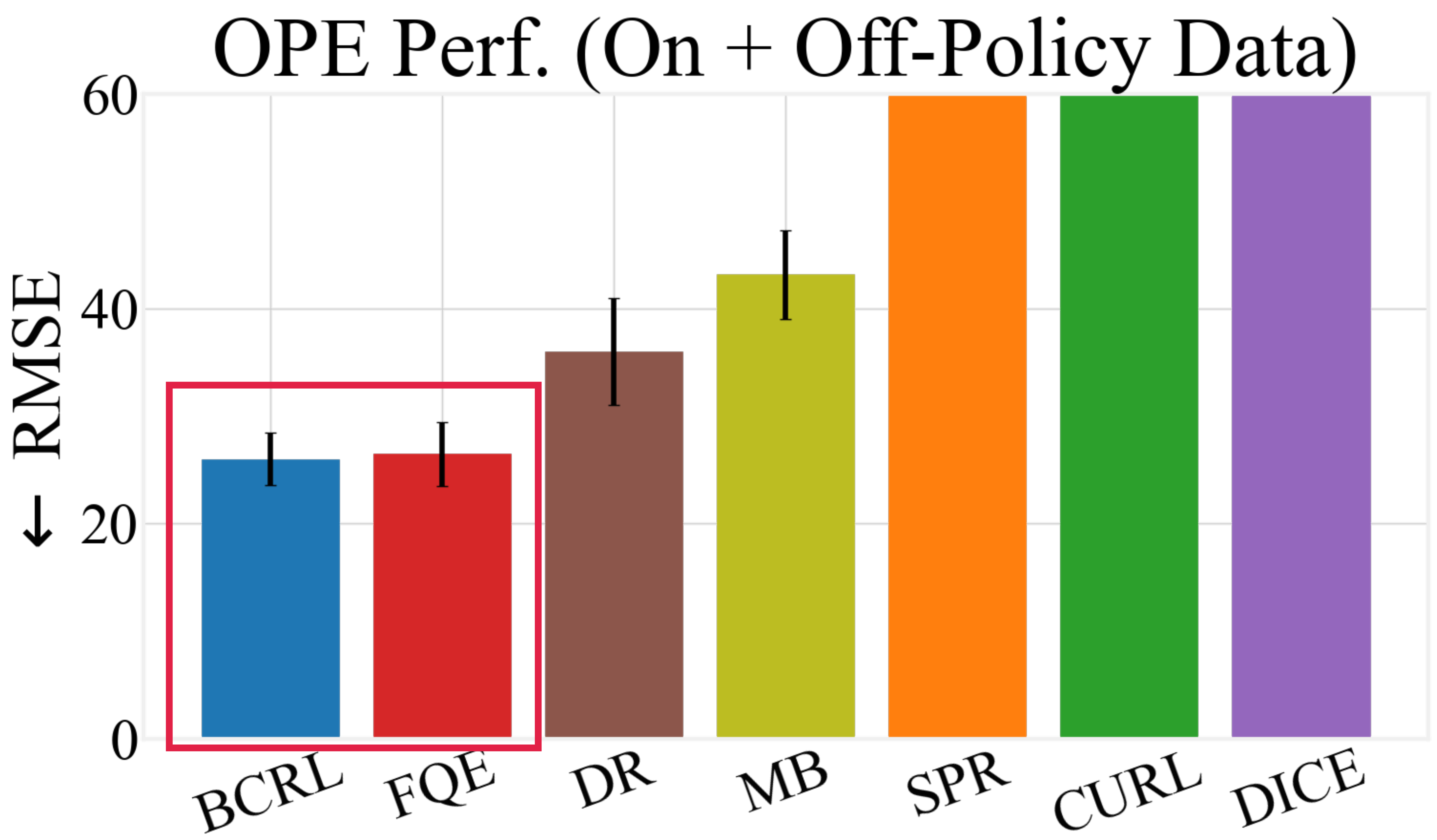


Takeaway

- Learning Bellman complete representations **improves OPE**

OPE Performance with On + Off policy Data

Testing Bellman Completeness

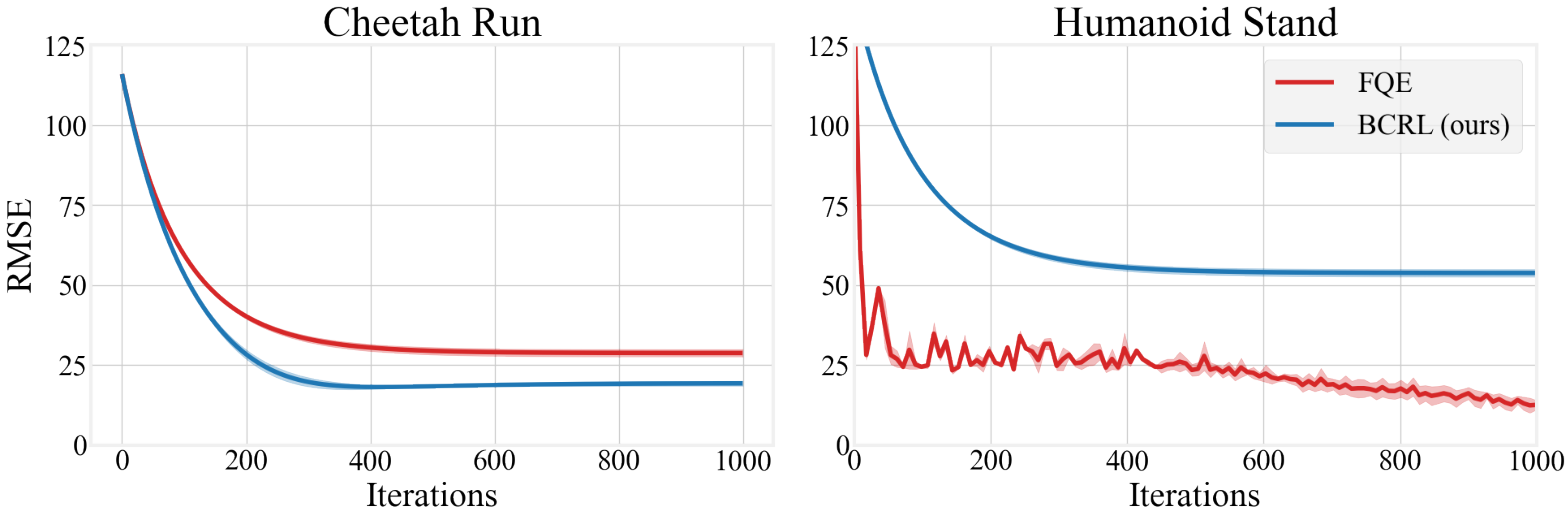


Takeaway

- Learning Bellman complete representations **improves OPE**

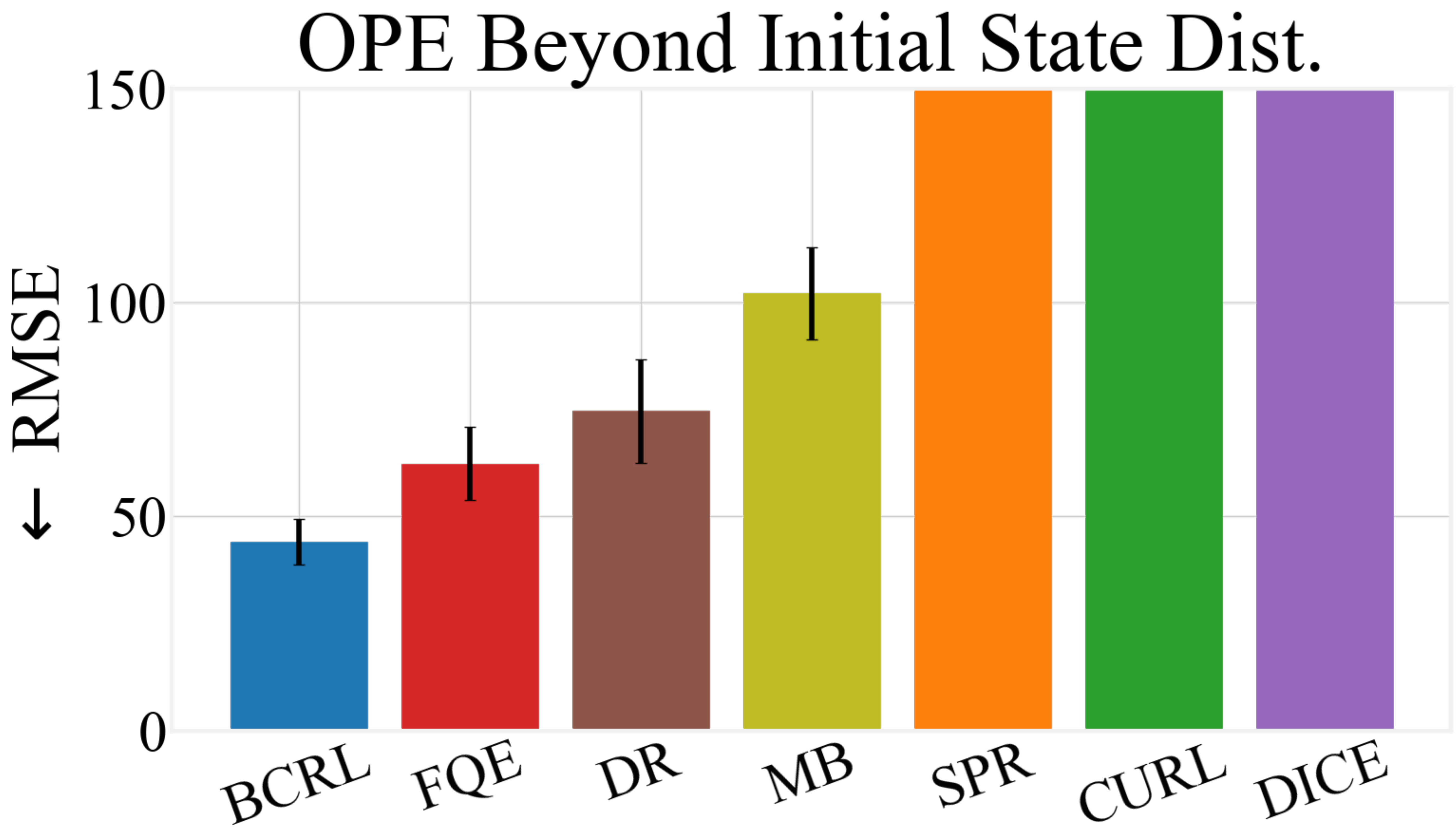
Detailed Look

Closer look at Cheetah and Humanoid



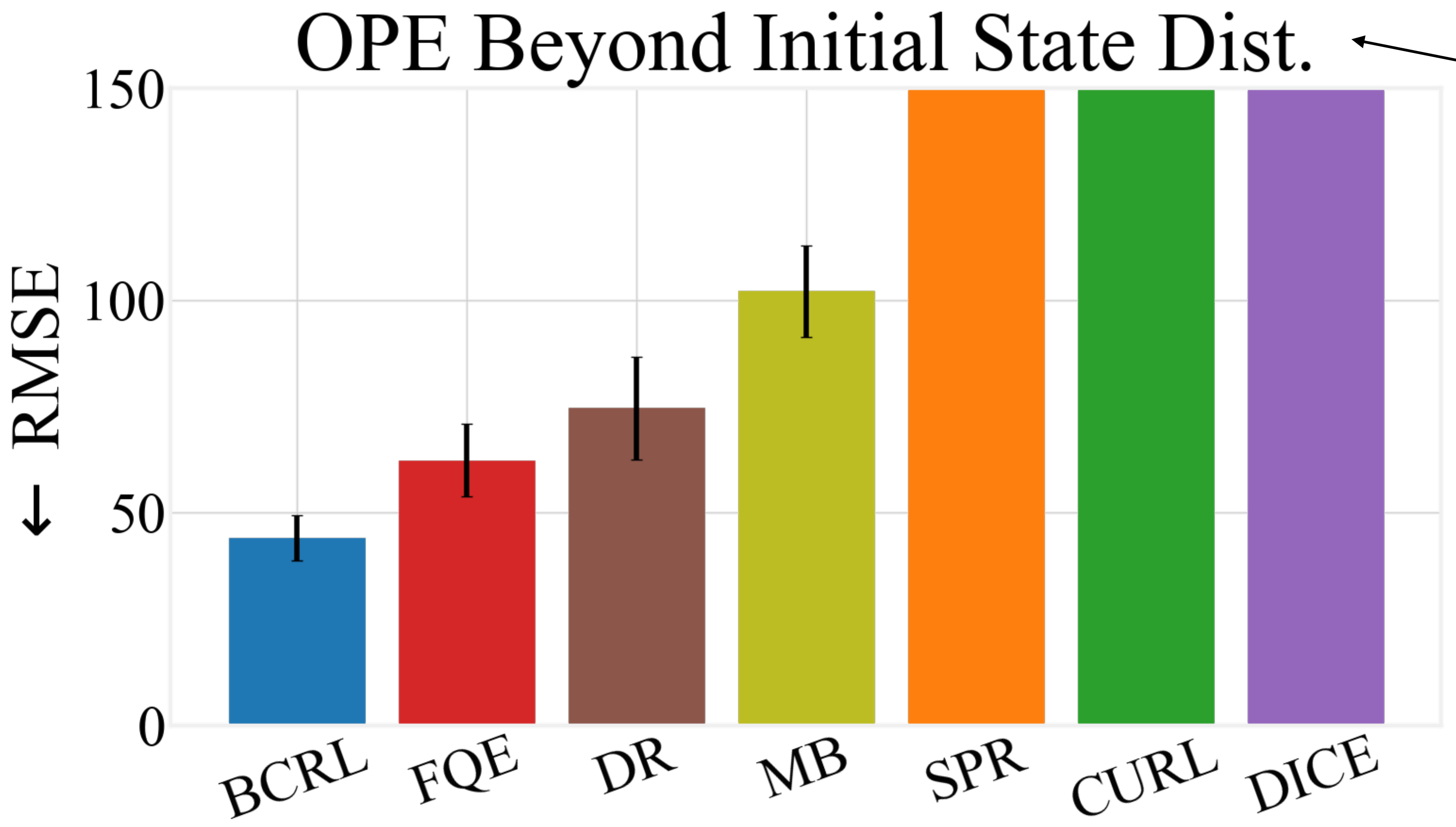
OPE Performance Beyond Init. State Distribution

Testing Coverage



OPE Performance Beyond Init. State Distribution

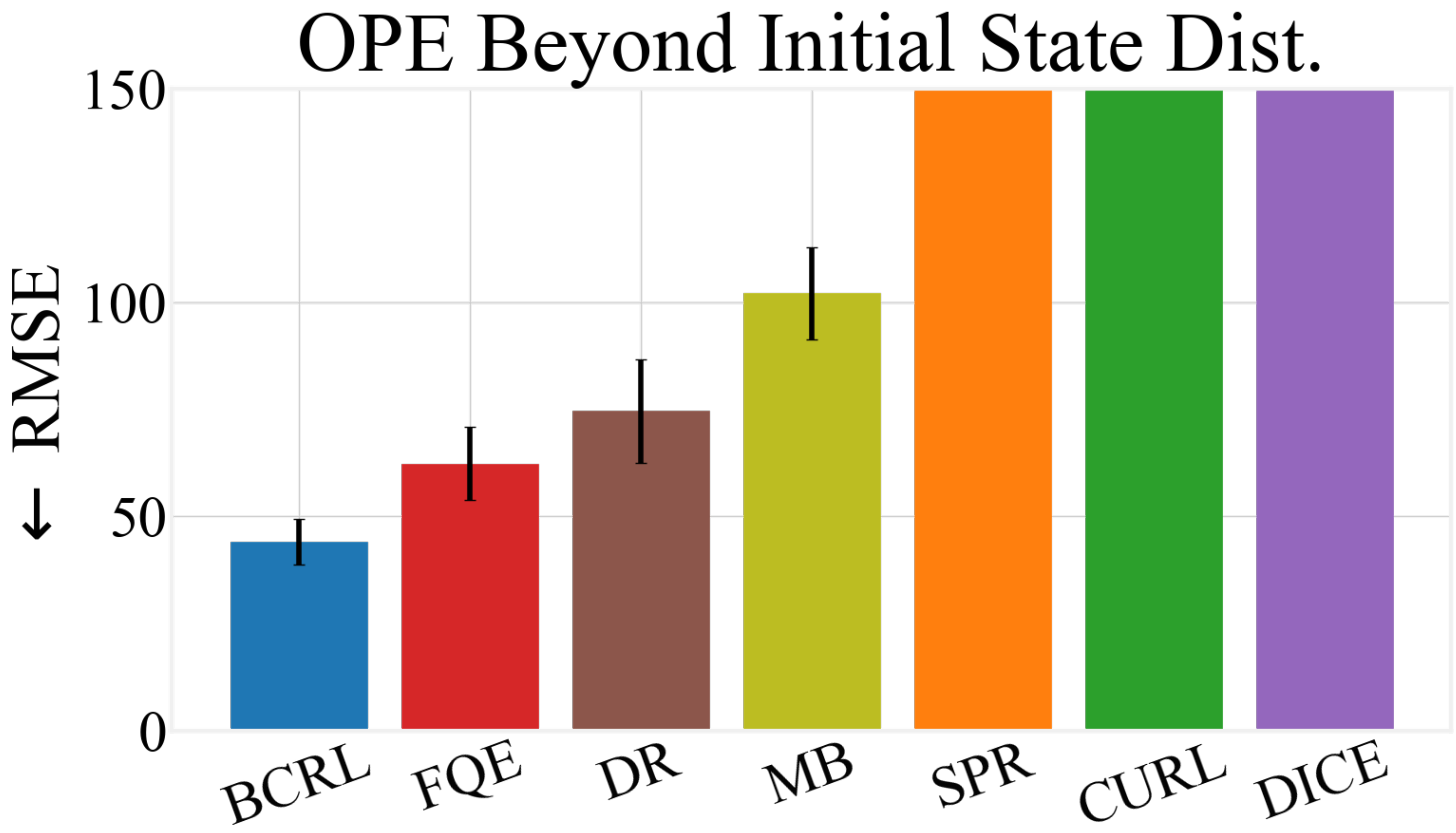
Testing Coverage



Evaluate at **all time steps**
of target policy rollout

OPE Performance Beyond Init. State Distribution

Testing Coverage



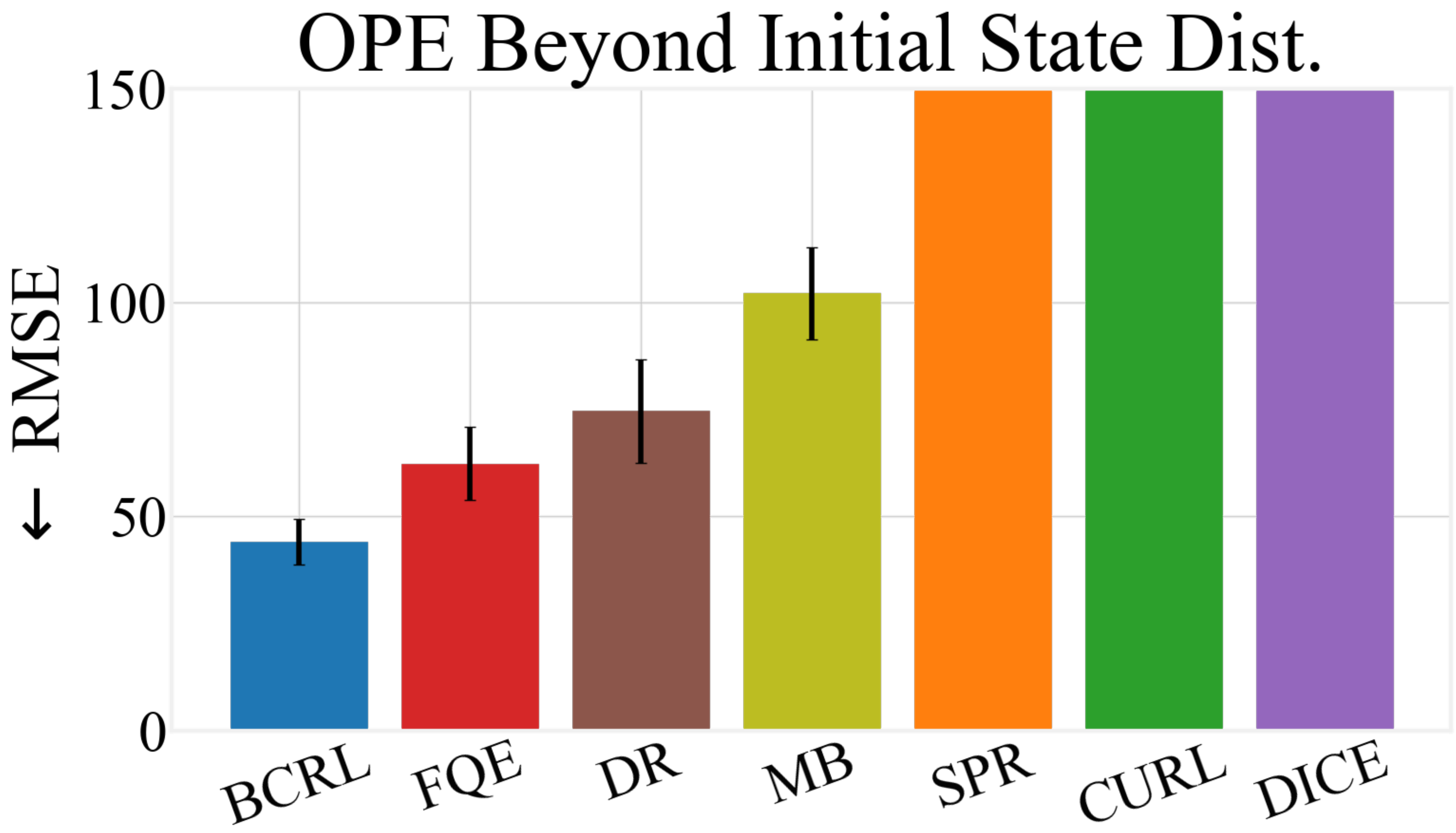
NOTE:
If representations are exactly Bellman Complete and has well-conditioned feature covariance matrix

↓

Should be able to evaluate well at **any state**

OPE Performance Beyond Init. State Distribution

Testing Coverage



Takeaway

- BCRL more robustly evaluates out-of-distribution

Takeaways

Takeaways

1. We can do provably good Offline Policy Evaluation with representations that are **bellman complete** and have **good coverage** over the offline data.

Takeaways

1. We can do provably good Offline Policy Evaluation with representations that are **bellman complete** and have **good coverage** over the offline data.
2. BCRL is able to both **scale** to complex image-based tasks and be a competitive policy evaluator.

Takeaways

1. We can do provably good Offline Policy Evaluation with representations that are **bellman complete** and have **good coverage** over the offline data.
2. BCRL is able to both **scale** to complex image-based tasks and be a competitive policy evaluator.
3. Although BCRL generally performs well, there is still room for improvement as seen in Humanoid Stand.

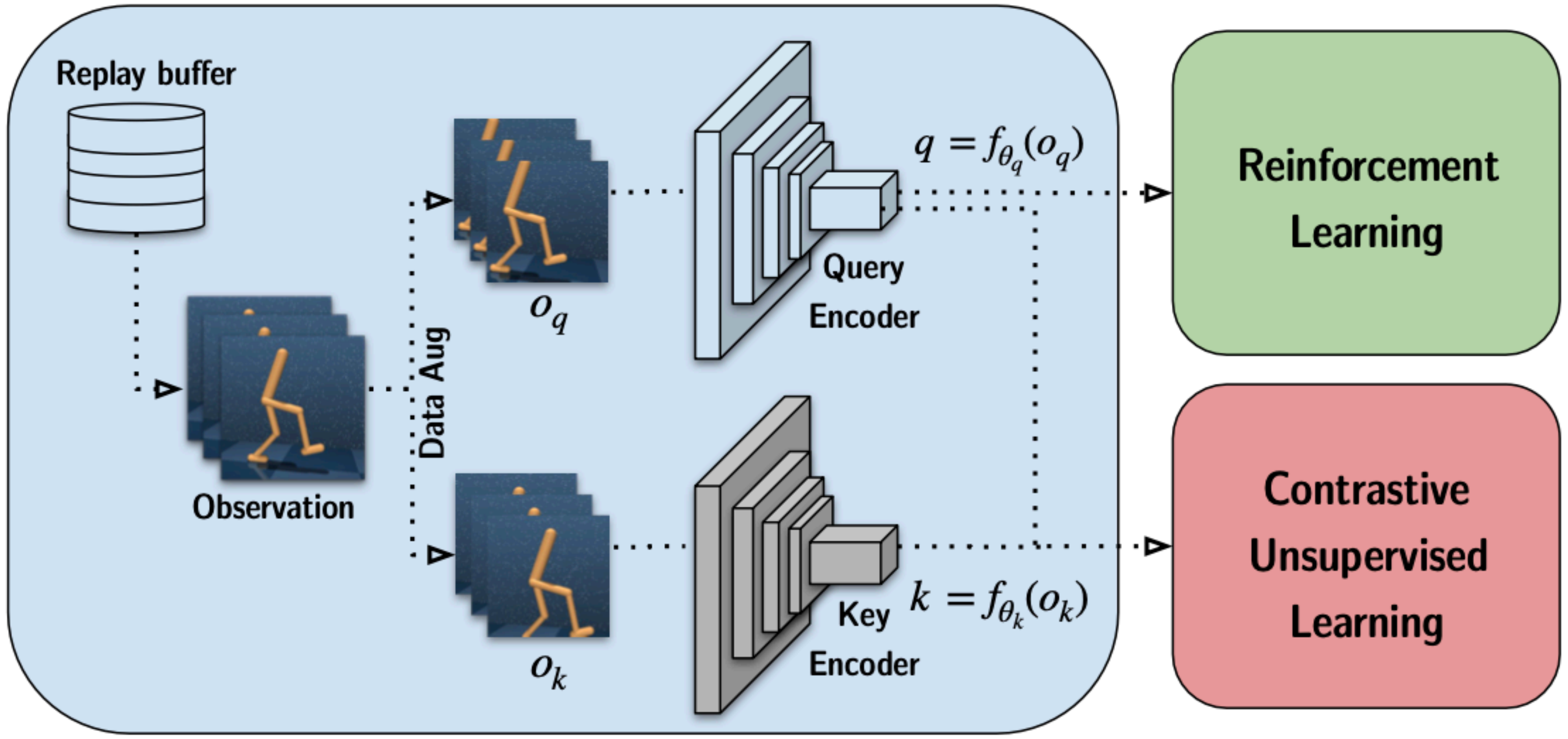
Thank you!

Github Repository: https://github.com/CausalML/bcrl

Appendices

CURL

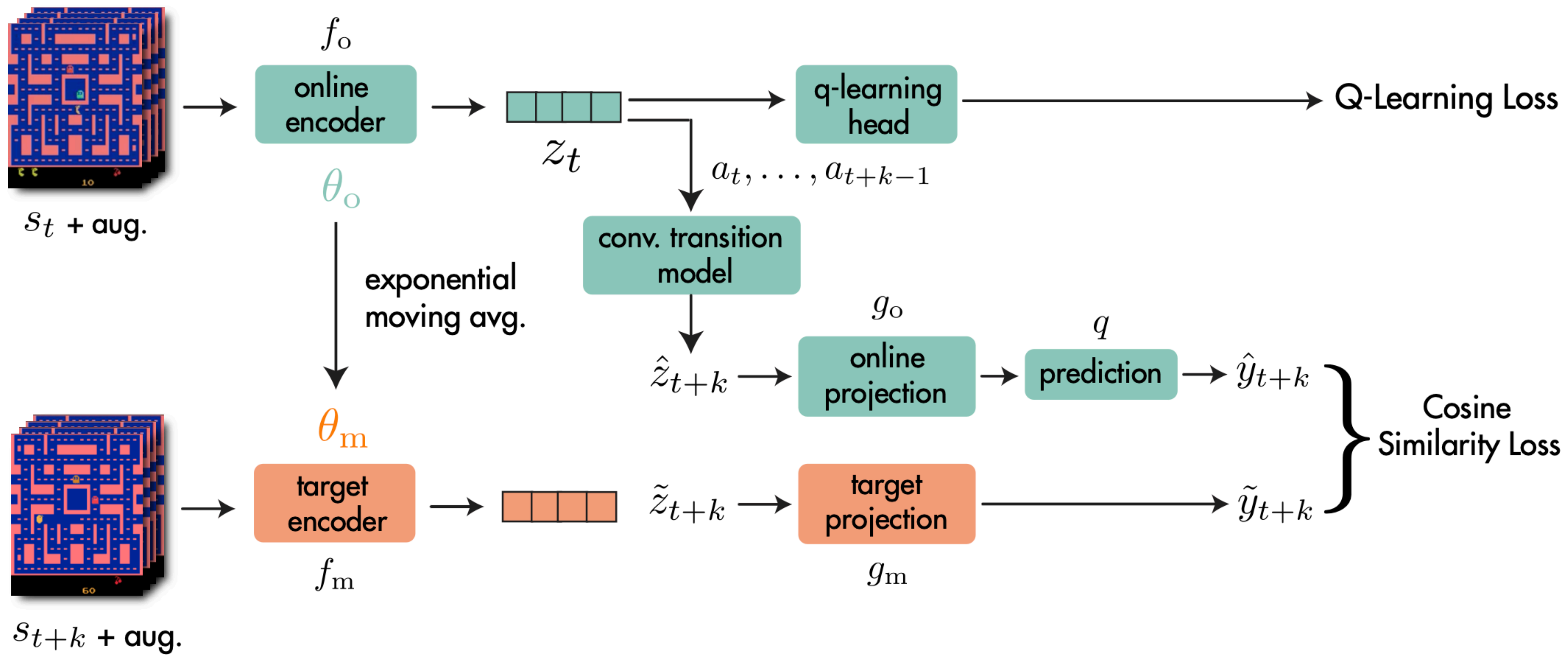
- Contrastive loss pushes different cropped frames to have different representations.



Laskin, Michael, Aravind Srinivas, and Pieter Abbeel. "Curl: Contrastive unsupervised representations for reinforcement learning." *International Conference on Machine Learning*. PMLR, 2020.

SPR

- Bootstrapping from latent representations by predicting into the future.



Schwarzer, M., Anand, A., Goel, R., Hjelm, R. D., Courville, A. C., and Bachman, P. Data-efficient reinforcement learning with momentum predictive representations. *ICLR*, 2021.

Ablations


Design Reg. Ablation
RMSE
150
100
50
0
0
250
500
750
1000
LSPE Iterations
with Reg.
w/o Reg.
Singular Values
10^{-3}
10^{-16}
10^{-29}
Bellman Completeness Ablation
RMSE
150
100
50
0
0
250
500
750
1000
LSPE Iterations
with BC
w/o BC

Equivalent Characterization

ϕ is Linear BC, meaning $\max_{w_1 \in B_W} \min_{w_2 \in B_W} \|w_2^T \phi - \mathcal{T}^\pi (w_1^T \phi)\|_\nu = 0.$

$$\Updownarrow$$

There exist $(\rho, M) \in B_W \times \mathbb{R}^{d \times d}$ with $\|M\|_2 < 1$ so that

$$\mathbb{E}_\nu \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathbb{E}_{s' \sim P(s,a)} \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0.$$

* formal result with norm constraints on ρ, M in paper.

Equivalent Characterization

ϕ is Linear BC, meaning $\max_{w_1 \in B_W} \min_{w_2 \in B_W} \|w_2^T \phi - \mathcal{T}^\pi(w_1^T \phi)\|_\nu = 0.$

Backward Direction:
For any w_1 set $w_2 = \rho + M^T w_1$.

$$\Updownarrow$$

There exist $(\rho, M) \in B_W \times \mathbb{R}^{d \times d}$ with $\|M\|_2 < 1$ so that

$$\mathbb{E}_\nu \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathbb{E}_{s' \sim P(s,a)} \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0.$$

* formal result with norm constraints on ρ, M in paper.

Equivalent Characterization

ϕ is Linear BC, meaning $\max_{w_1 \in B_W} \min_{w_2 \in B_W} \|w_2^T \phi - \mathcal{T}^\pi(w_1^T \phi)\|_\nu = 0.$

Forward Direction:
To get ρ: set $w_1 = 0$ and use w_2.
To get ith row of M:
set $w_1 = e_i$ and use $w_2 - \rho$.

$$\Updownarrow$$

Backward Direction:
For any w_1 set $w_2 = \rho + M^T w_1$.

There exist $(\rho, M) \in B_W \times \mathbb{R}^{d \times d}$ with $\|M\|_2 < 1$ so that

$$\mathbb{E}_\nu \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathbb{E}_{s' \sim P(s,a)} \phi(s', \pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 = 0.$$

* formal result with norm constraints on ρ, M in paper.

Theory: LSPE OPE Guarantee

- Theorem: Let ϕ be a ε_ν-approximate Linear BC feature.
  For any δ and large enough dataset of size N, with probability at least $1-\delta$, K iterates of LSPE evaluates well **for any distribution p_0**,

$$\left| V_{p_0}^{\pi_e} - \mathbb{E}_{s\sim p_0}[\hat{f}_K(s,\pi_e)] \right| \in \tilde{\mathcal{O}}\left(\frac{\gamma^{K/2}}{(1-\gamma)} + \frac{\sqrt{\left\| \frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu} \right\|_\infty}}{(1-\gamma)^2} \cdot \varepsilon_\nu + \sqrt{\kappa(p_0)} \frac{d}{(1-\gamma)^2\sqrt{N}} \right).$$

Theory: LSPE OPE Guarantee

- Theorem: Let ϕ be a ε_ν-approximate Linear BC feature.
  For any δ and large enough dataset of size N, with probability at least $1-\delta$, K iterates of LSPE evaluates well **for any distribution p_0**,

$$\left| V_{p_0}^{\pi_e} - \mathbb{E}_{s\sim p_0}[\hat{f}_K(s,\pi_e)] \right| \in \tilde{\mathcal{O}}\left(\frac{\gamma^{K/2}}{(1-\gamma)} + \frac{\sqrt{\left\| \frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu} \right\|_\infty}}{(1-\gamma)^2} \cdot \varepsilon_\nu + \sqrt{\kappa(p_0)} \frac{d}{(1-\gamma)^2\sqrt{N}} \right).$$

Exponentially decaying in
num. LSPE iterations K

Theory: LSPE OPE Guarantee

- Theorem: Let ϕ be a ε_ν-approximate Linear BC feature.
  For any δ and large enough dataset of size N, with probability at least $1-\delta$, K iterates of LSPE evaluates well **for any distribution p_0**,

$$\left| V^{\pi_e}_{p_0} - \mathbb{E}_{s\sim p_0}[\hat{f}_K(s,\pi_e)] \right| \in \tilde{\mathcal{O}}\left(\underbrace{\frac{\gamma^{K/2}}{(1-\gamma)}}_{\text{Exponentially decaying in num. LSPE iterations } K} + \underbrace{\frac{\sqrt{\left\| \frac{\mathrm{d}d^{\pi_e}_{p_0}}{\mathrm{d}\nu} \right\|_\infty}}{(1-\gamma)^2} \cdot \varepsilon_\nu}_{\text{Non-Linear BC part bounded by density ratio}} + \sqrt{\kappa(p_0)} \frac{d}{(1-\gamma)^2\sqrt{N}} \right).$$

Theory: LSPE OPE Guarantee

- Theorem: Let ϕ be a ε_ν-approximate Linear BC feature.
  For any δ and large enough dataset of size N, with probability at least $1-\delta$, K iterates of LSPE evaluates well **for any distribution p_0**,

$$\left| V_{p_0}^{\pi_e} - \mathbb{E}_{s\sim p_0}[\hat{f}_K(s,\pi_e)] \right| \in \tilde{\mathcal{O}}\left(\frac{\gamma^{K/2}}{(1-\gamma)} + \frac{\sqrt{\left\| \frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu} \right\|_\infty}}{(1-\gamma)^2} \cdot \varepsilon_\nu + \sqrt{\kappa(p_0)} \frac{d}{(1-\gamma)^2\sqrt{N}} \right).$$

Exponentially decaying in num. LSPE iterations K

Non-Linear BC part bounded by density ratio

Statistical error from evaluation, converging to zero as N grows

Least Squares Policy Evaluation

Algorithm 1 Least Squares Policy Evaluation (LSPE)

Input: Target policy π_e, features ϕ, dataset $\mathcal{D}$

Initialize $\widehat{\theta}_0 = \mathbf{0} \in B_W$.

for $k = 1, 2, ..., K$ **do**

  Set $\widehat{f}_{k-1}(s,a) = \widehat{\theta}_{k-1}^{\intercal}\phi(s,a)$,

  $\widehat{V}_{k-1}(s) = \mathbb{E}_{a\sim\pi_e(s)}[\widehat{f}_{k-1}(s,a)]$

  Perform linear regression:

$$\widehat{\theta}_k \in \arg\min_{\theta \in B_W} \frac{1}{N}\sum_{i=1}^{N}\left(\theta^{\intercal}\phi(s_i,a_i) - r_i - \gamma\widehat{V}_{k-1}(s_i')\right)^2$$

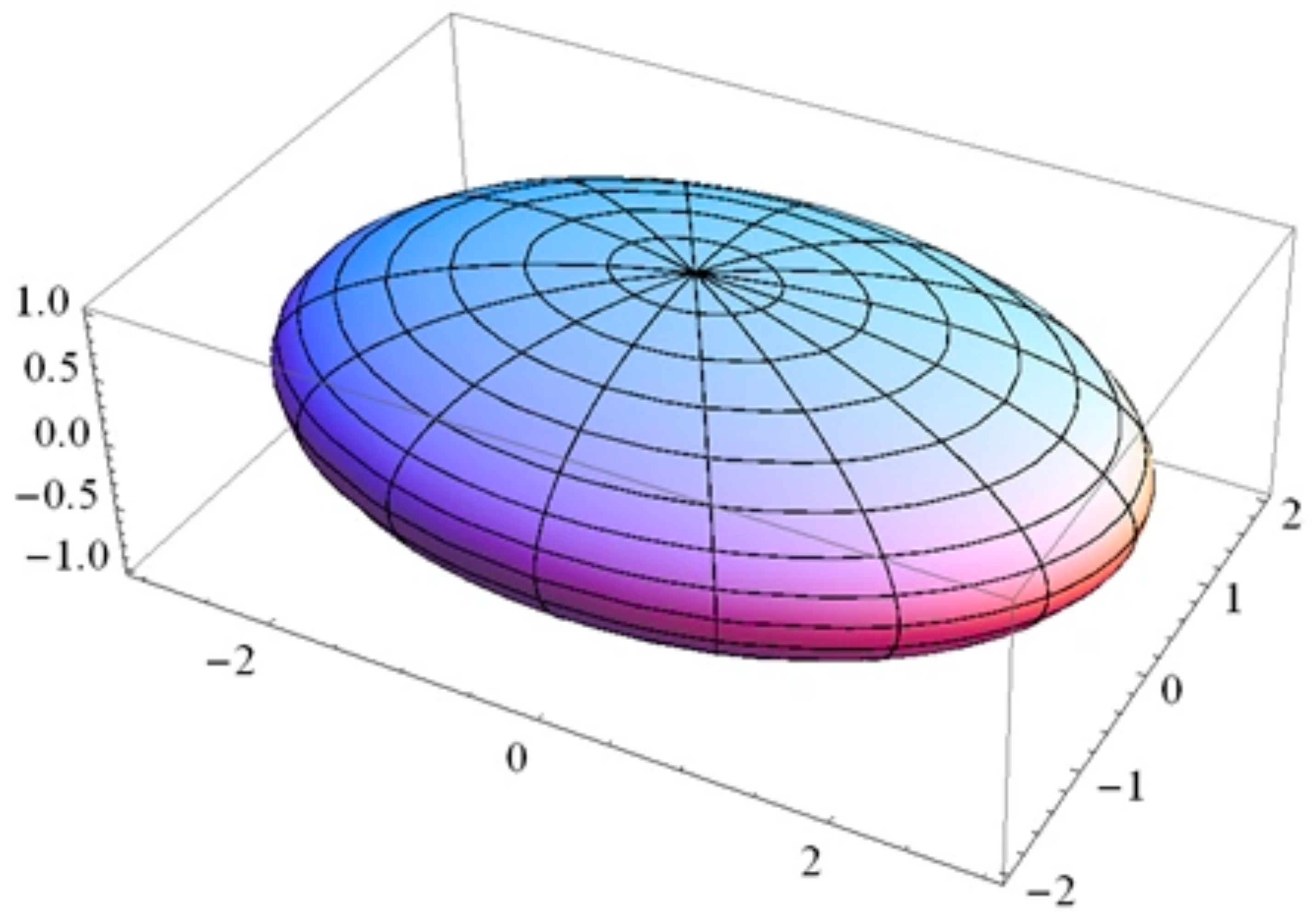
end for

Return $\widehat{f}_K$.

Relative Coverage

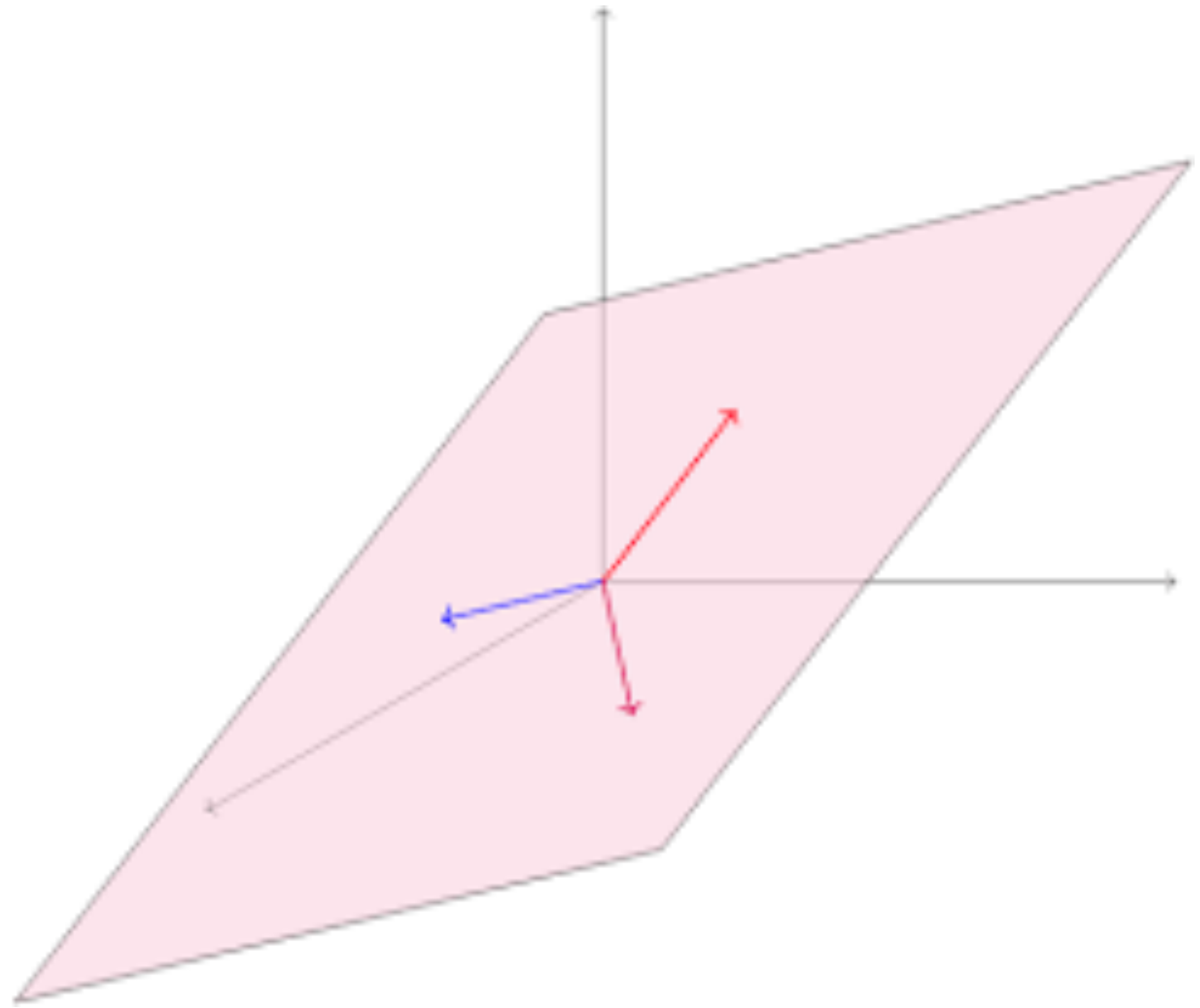
$$\kappa(p_0) := \sup_{x \in \mathbb{R}^d} \frac{x^T \mathbb{E}_{d_{p_0}^{\pi_e}}[\phi(s,a)\phi(s,a)^T]x}{x^T \Sigma(\phi) x}$$

where $\Sigma(\phi) = \mathbb{E}_{\nu}[\phi(s,a)\phi(s,a)^T]$.



Can be bounded when, e.g.

- $\Sigma(\phi)$ is invertible and well-conditioned.
- $\nu = \frac{1}{2} d_{p_0}^{\pi_e} + \frac{1}{2}\mu$, i.e. density ratio is upper bounded.
- $\lambda_{max}\left(\Sigma^{-1}\mathbb{E}_{d_{p_0}^{\pi}}[\phi(s,a)\phi(s,a)^T]\right)$.



Proof Breakdown

- First, a "value difference lemma":
  $$V_\mu^\pi - \mathbb{E}_\mu[f(s,\pi)] = \frac{1}{1-\gamma}\mathbb{E}_{d_\mu^\pi}[\mathcal{T}^\pi f(s,a) - f(s,\pi)].$$
- Then, $\|f_K - \mathcal{T}^\pi f_K\|_{d_{p_0}^\pi} \le \dfrac{4}{1-\gamma} \max_{k\in[K]} \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d_{p_0}^\pi} + \gamma^{K/2}$.

Proof Breakdown

- First, a "value difference lemma":

$$V_\mu^\pi - \mathbb{E}_\mu[f(s,\pi)] = \frac{1}{1-\gamma}\mathbb{E}_{d_\mu^\pi}[\mathcal{T}^\pi f(s,a) - f(s,\pi)].$$

- Then, $\|f_K - \mathcal{T}^\pi f_K\|_{d_{p_0}^\pi} \le \dfrac{4}{1-\gamma}\max_{k\in[K]}\|f_k - \mathcal{T}^\pi f_{k-1}\|_{d_{p_0}^\pi} + \gamma^{K/2}$.

Maximum per-iteration error from LSPE

$$\le \sup_{\vartheta\in B_W}\|\widehat{\theta}_\vartheta^T\phi - \mathcal{T}^\pi(\vartheta^T\phi)\|_{d_{p_0}^\pi}, \text{ where}$$

$$\widehat{\theta}_\vartheta := \arg\min_{\theta\in B_W}\widehat{\ell}(\theta,\vartheta)$$

$$\widehat{\ell}(\theta,\vartheta) := \mathbb{E}_{\mathcal{D}}\left[\left(r(s,a) + \gamma\vartheta^T\phi(s',\pi) - \theta^T\phi(s,a)\right)^2\right]$$

Proof Breakdown

- First, a "value difference lemma":

$$V_\mu^\pi - \mathbb{E}_\mu[f(s,\pi)] = \frac{1}{1-\gamma}\mathbb{E}_{d_\mu^\pi}[\mathcal{T}^\pi f(s,a) - f(s,\pi)].$$

- Then, $\|f_K - \mathcal{T}^\pi f_K\|_{d_{p_0}^\pi} \le \dfrac{4}{1-\gamma} \max\limits_{k\in[K]} \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d_{p_0}^\pi} + \gamma^{K/2}.$

Maximum per-iteration error from LSPE

$$\le \sup_{\vartheta\in B_W} \| \widehat{\theta}_\vartheta^T \phi - \mathcal{T}^\pi(\vartheta^T\phi)\|_{d_{p_0}^\pi}, \text{ where}$$

$$\widehat{\theta}_\vartheta := \arg\min_{\theta\in B_W} \widehat{\ell}(\theta,\vartheta)$$

$$\widehat{\ell}(\theta,\vartheta) := \mathbb{E}_{\mathcal{D}}\left[\left(r(s,a) + \gamma\vartheta^T\phi(s',\pi) - \theta^T\phi(s,a)\right)^2\right]$$

$$\le \sqrt{\kappa(p_0)} \sup_{\vartheta\in B_W} \|\widehat{\theta}_\vartheta - \theta_\vartheta\|_\Sigma + \sup_{\vartheta\in B_W} \|\theta_\vartheta^T\phi - \mathcal{T}^\pi(\vartheta^T\phi)\|_{d_{p_0}^\pi}$$

Proof Breakdown

- First, a "value difference lemma":

$$V_\mu^\pi - \mathbb{E}_\mu[f(s,\pi)] = \frac{1}{1-\gamma}\mathbb{E}_{d_\mu^\pi}[\mathcal{T}^\pi f(s,a) - f(s,\pi)].$$

- Then, $\|f_K - \mathcal{T}^\pi f_K\|_{d_{p_0}^\pi} \le \frac{4}{1-\gamma} \max_{k\in[K]} \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d_{p_0}^\pi} + \gamma^{K/2}$.

Maximum per-iteration error from LSPE

$$\le \sup_{\vartheta\in B_W} \|\widehat{\theta}_\vartheta^T\phi - \mathcal{T}^\pi(\vartheta^T\phi)\|_{d_{p_0}^\pi}, \text{ where}$$

$$\widehat{\theta}_\vartheta := \arg\min_{\theta\in B_W} \widehat{\ell}(\theta,\vartheta)$$

$$\widehat{\ell}(\theta,\vartheta) := \mathbb{E}_{\mathcal{D}}\left[\left(r(s,a) + \gamma\vartheta^T\phi(s',\pi) - \theta^T\phi(s,a)\right)^2\right]$$

$$\le \sqrt{\kappa(p_0)} \sup_{\vartheta\in B_W} \|\widehat{\theta}_\vartheta - \theta_\vartheta\|_\Sigma + \sup_{\vartheta\in B_W} \|\theta_\vartheta^T\phi - \mathcal{T}^\pi(\vartheta^T\phi)\|_{d_{p_0}^\pi}$$

$$\widetilde{\mathcal{O}}\left(\frac{dW}{\sqrt{N}}\right)$$

Proof Breakdown

- First, a "value difference lemma":

$$V_\mu^\pi - \mathbb{E}_\mu[f(s,\pi)] = \frac{1}{1-\gamma}\mathbb{E}_{d_\mu^\pi}[\mathcal{T}^\pi f(s,a) - f(s,\pi)].$$

- Then, $\|f_K - \mathcal{T}^\pi f_K\|_{d_{p_0}^\pi} \le \dfrac{4}{1-\gamma}\max_{k\in[K]}\|f_k - \mathcal{T}^\pi f_{k-1}\|_{d_{p_0}^\pi} + \gamma^{K/2}.$

Maximum per-iteration error from LSPE

$$\le \sup_{\vartheta\in B_W}\|\widehat{\theta}_\vartheta^T\phi - \mathcal{T}^\pi(\vartheta^T\phi)\|_{d_{p_0}^\pi}, \text{ where}$$

$$\widehat{\theta}_\vartheta := \arg\min_{\theta\in B_W}\widehat{\ell}(\theta,\vartheta)$$

$$\widehat{\ell}(\theta,\vartheta) := \mathbb{E}_{\mathcal{D}}\left[\left(r(s,a)+\gamma\vartheta^T\phi(s',\pi)-\theta^T\phi(s,a)\right)^2\right]$$

$$\le \sqrt{\kappa(p_0)}\sup_{\vartheta\in B_W}\|\widehat{\theta}_\vartheta - \theta_\vartheta\|_\Sigma + \sup_{\vartheta\in B_W}\|\theta_\vartheta^T\phi - \mathcal{T}^\pi(\vartheta^T\phi)\|_{d_{p_0}^\pi}$$

$$\widetilde{\mathcal{O}}\left(\frac{dW}{\sqrt{N}}\right)$$

$$\le \sqrt{\left\|\frac{\mathrm{d}d_{p_0}^{\pi_e}}{\mathrm{d}\nu}\right\|_\infty}\cdot\varepsilon_\nu$$

Stochastic BCRL

- Recall the Bellman Complete loss is,

$$\min_{(\rho,M)\in\Theta} \mathbb{E}_{\mathcal{D}} \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma \mathbb{E}_{s'\sim P(s,a)}[\phi(s',\pi_e)] \\ r(s,a) \end{bmatrix} \right\|_2^2 .$$

- When task is stochastic, double sampling issue.
- Fix by subtracting the overestimation bias.
  (which is the variance, and can be estimated!)

Stochastic BCRL

- $\mathbb{E}_{\nu \circ P} \left\| M\phi(s,a) - \gamma\phi(s',\pi_e) \right\|_2^2 - \mathbb{E}_{\nu} \left\| M\phi(s,a) - \gamma\mathbb{E}_{s' \sim P(s,a)}[\phi(s',\pi_e)] \right\|_2^2$
  $= \inf_{g} \mathbb{E}_{\nu \circ P} \left\| \gamma\phi(s',\pi_e) - \gamma\mathbb{E}_{s' \sim P(s,a)}[\phi(s',\pi_e)] \right\|_2^2$
- So, when MDP is stochastic, BCRL is:

$$\widehat{\phi} \in \arg\min_{\phi \in \Phi} \left[\min_{(\rho,M) \in \Theta} \mathbb{E}_{\mathcal{D}} \left\| \begin{bmatrix} M \\ \rho^T \end{bmatrix} \phi(s,a) - \begin{bmatrix} \gamma\phi(s',\pi_e) \\ r(s,a) \end{bmatrix} \right\|_2^2 - \min_{g \in \mathcal{G}} \mathbb{E}_{\mathcal{D}} \left\| g(s,a) - \gamma\phi(s',\pi_e) \right\|_2^2 \right]$$

$$\text{s.t. } \lambda_{min} \left(\mathbb{E}_{\mathcal{D}}[\phi(s,a)\phi(s,a)^T] \right) \geq \beta/2.$$

Theory: Representation Learning

- Theorem: Assume realizability of $\mathcal{G}$.

  For any δ and large enough dataset of size N, with probability at least $1-\delta$, we have that the ERM $\widehat{\phi}$ satisfies,

  1. Approximately Linear BC, with
  $$\varepsilon_\nu = \widetilde{\mathcal{O}}\left(\frac{d \cdot \mathrm{comp}(\Phi)}{\sqrt{N}} + \frac{\gamma \cdot \mathrm{comp}(\mathcal{G})}{\sqrt{N}}\right),$$
  2. Coverage, with $\lambda_{min}\left(\mathbb{E}_\nu[\widehat{\phi}(s,a)\widehat{\phi}(s,a)^T]\right) \geq \beta/4$.