Stable Offline Value Function Learning with Bisimulation-based Representations




Brahma S. Pavse Yudong Chen Qiaomin Xie Josiah P. Hanna

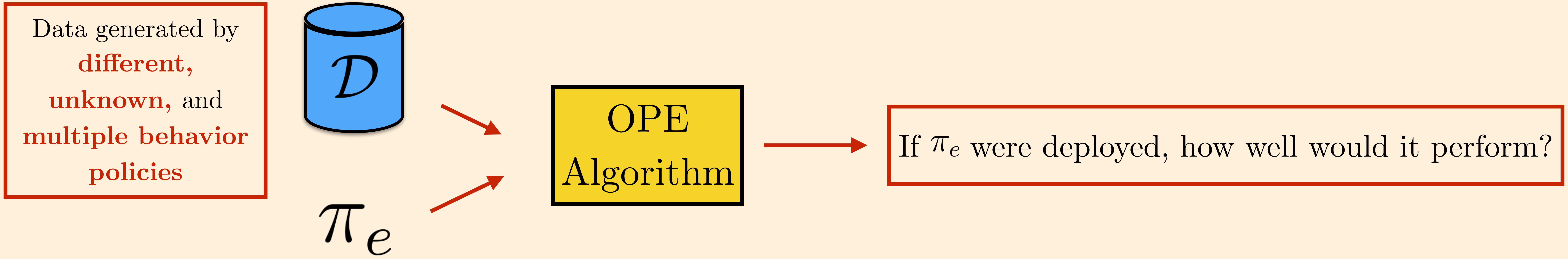
University of Wisconsin — Madison

pavse@wisc.edu

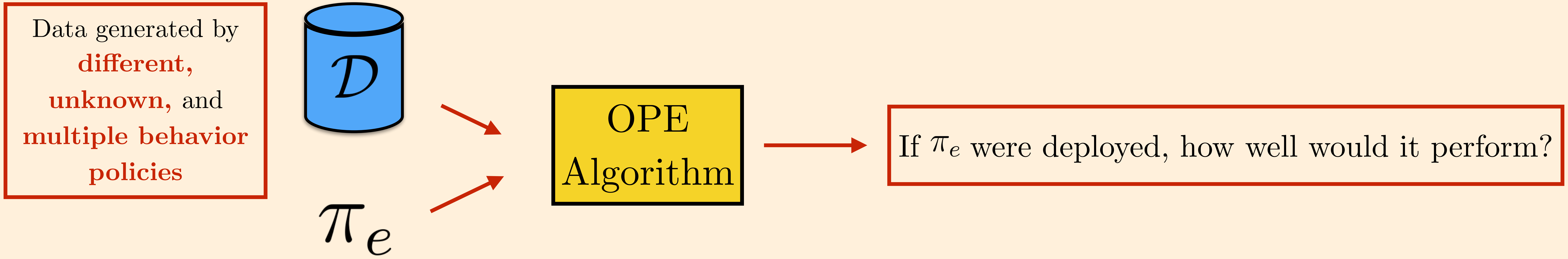
Paper:

Problem Setting: Offline Policy Evaluation (OPE)

Problem Setting: Offline Policy Evaluation (OPE)



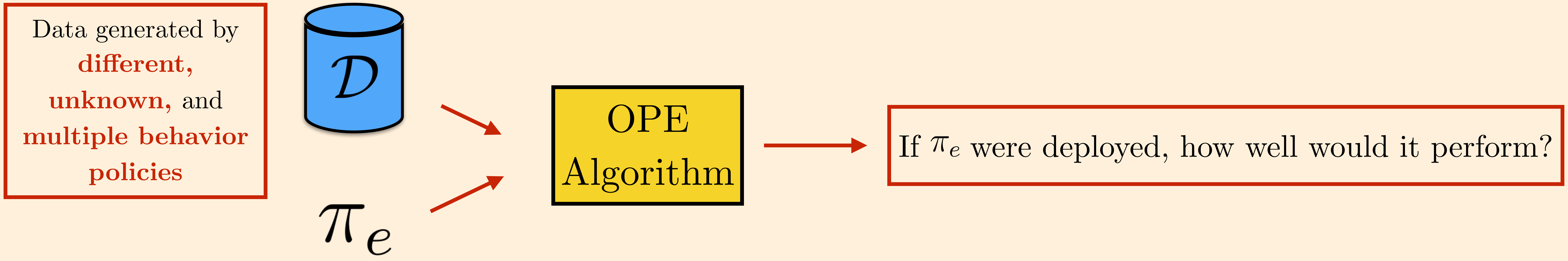
Problem Setting: Offline Policy Evaluation (OPE)



Desiderata of an OPE algorithm:

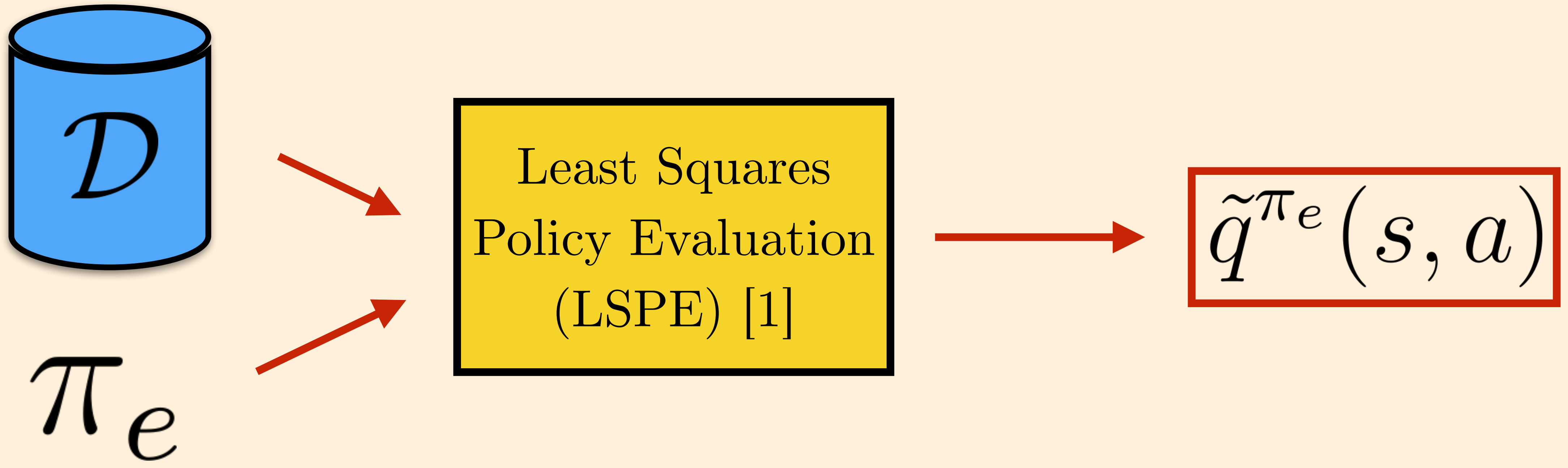
1) **accurate**
2) **convergent prediction**

Problem Setting: Offline Policy Evaluation (OPE)



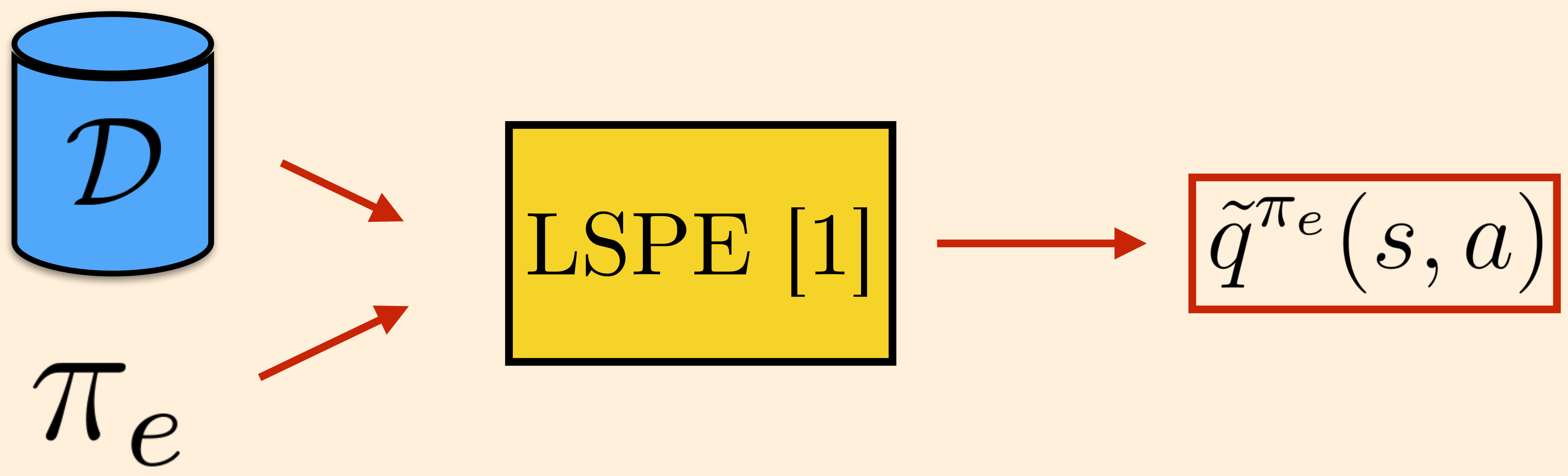
Desiderata of an OPE algorithm:

1) **accurate**
2) **convergent prediction**



1. Nedic, A. and Bertsekas, D. Least squares policy evaluation algorithms with linear function approximation. 2003.

Main Contribution

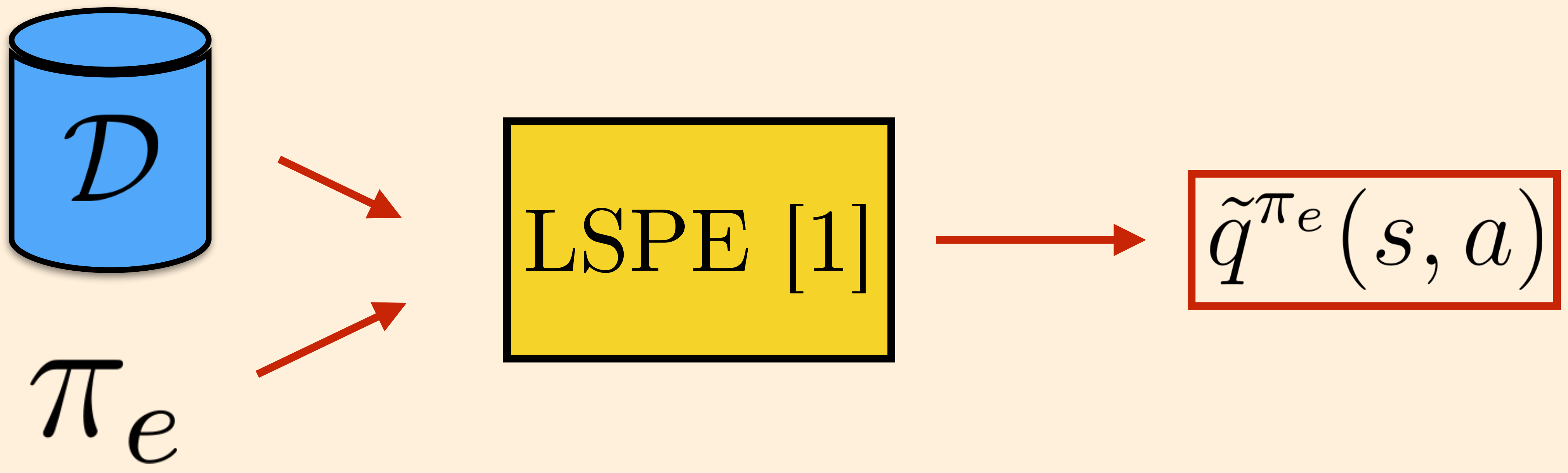


Desiderata of an OPE algorithm:

1) **accurate**
2) **convergent prediction**

1. Nedic, A. and Bertsekas, D. Least squares policy evaluation algorithms with linear function approximation. 2003.

Main Contribution



Desiderata of an OPE algorithm:

1) **accurate**
2) **convergent prediction**

Main Contribution: Bisimulation-based Representation learning for OPE

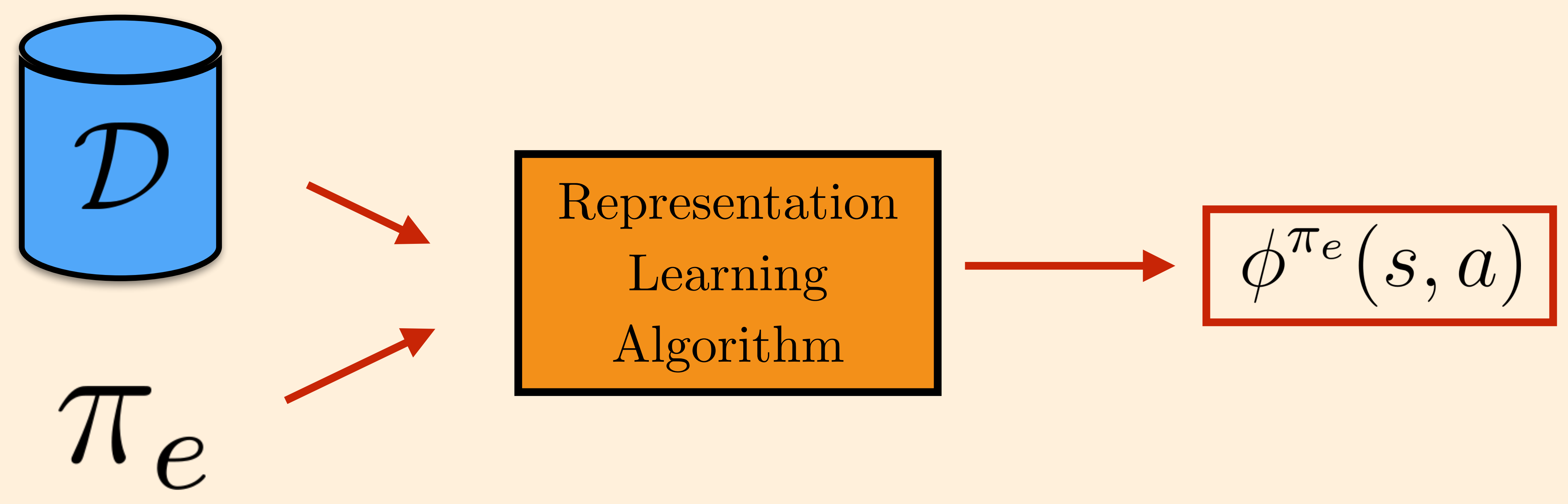
Shaping state-action features with **bisimulation-based** representation learning **before** feeding into LSPE can lead to **convergent** OPE predictions.

1. Nedic, A. and Bertsekas, D. Least squares policy evaluation algorithms with linear function approximation. 2003.

Representation Learning + OPE Pipeline

1. Nedic, A. and Bertsekas, D. Least squares policy evaluation algorithms with linear function approximation. 2003.

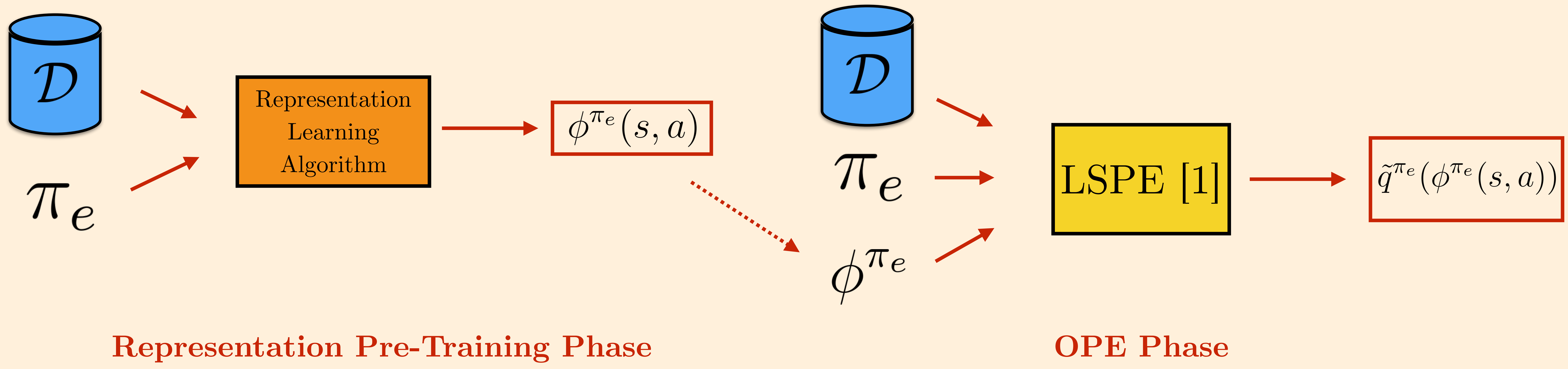
Representation Learning + OPE Pipeline



Representation Pre-Training Phase

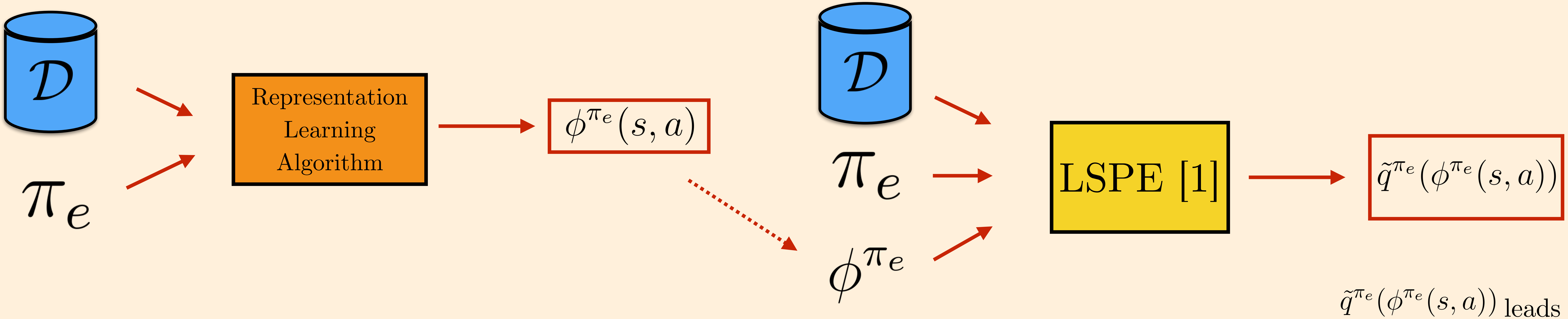
1. Nedic, A. and Bertsekas, D. Least squares policy evaluation algorithms with linear function approximation. 2003.

Representation Learning + OPE Pipeline



1. Nedic, A. and Bertsekas, D. Least squares policy evaluation algorithms with linear function approximation. 2003.

Representation Learning + OPE Pipeline



Representation Pre-Training Phase

OPE Phase

$\tilde{q}^{\pi_e}(\phi^{\pi_e}(s, a))$ leads to **more accurate** OPE estimates than $\tilde{q}^{\pi_e}(s, a)$

1. Nedic, A. and Bertsekas, D. Least squares policy evaluation algorithms with linear function approximation. 2003.

Our Work: Kernel Representations for OPE (KROPE)

Our Work: Kernel Representations for OPE (KROPE)

- Builds upon Kernel Similarity Metric (KSMe) [1].

1. Castro et al. 2023. A Kernel Perspective on Behavioural Metrics for Markov Decision Processes.

Our Work: Kernel Representations for OPE (KROPE)

- Builds upon Kernel Similarity Metric (KSMe) [1].
- KROPE similarity metric (short-term + long-term similarity):

$$k^{\pi_e}(s_1, a_1; s_2, a_2) := 1 - \frac{|r(s_1, a_1) - r(s_2, a_2)|}{|r_{\max} - r_{\min}|} + \gamma \mathbb{E}_{a_1' \sim \pi_e(s_1'), a_2' \sim \pi_e(s_2')}[k_e^{\pi}(s_1', a_1'; s_2', a_2')]$$

1. Castro et al. 2023. A Kernel Perspective on Behavioural Metrics for Markov Decision Processes.

Our Work: Kernel Representations for OPE (KROPE)

- Builds upon Kernel Similarity Metric (KSMe) [1].
- KROPE similarity metric (short-term + long-term similarity):

$$k^{\pi_e}(\underline{s_1, a_1}; \underline{s_2, a_2}) := 1 - \frac{|r(s_1, a_1) - r(s_2, a_2)|}{|r_{\max} - r_{\min}|} + \gamma \mathbb{E}_{a_1' \sim \pi_e(s_1'), a_2' \sim \pi_e(s_2')}[k_e^{\pi}(s_1', a_1'; s_2', a_2')]$$

1. Castro et al. 2023. A Kernel Perspective on Behavioural Metrics for Markov Decision Processes.

Our Work: Kernel Representations for OPE (KROPE)

- Builds upon Kernel Similarity Metric (KSMe) [1].
- KROPE similarity metric (short-term + long-term similarity):

$$k^{\pi_e}(s_1, a_1; s_2, a_2) := 1 - \frac{|r(s_1, a_1) - r(s_2, a_2)|}{|r_{\max} - r_{\min}|} + \gamma \mathbb{E}_{a_1' \sim \pi_e(s_1'), a_2' \sim \pi_e(s_2')}[k_e^{\pi}(s_1', a_1'; s_2', a_2')]$$

1. Castro et al. 2023. A Kernel Perspective on Behavioural Metrics for Markov Decision Processes.

Our Work: Kernel Representations for OPE (KROPE)

- Builds upon Kernel Similarity Metric (KSMe) [1].
- KROPE similarity metric (short-term + long-term similarity):

$$k^{\pi_e}(s_1, a_1; s_2, a_2) := 1 - \frac{|r(s_1, a_1) - r(s_2, a_2)|}{|r_{\max} - r_{\min}|} + \gamma \mathbb{E}_{a_1' \sim \pi_e(s_1'), a_2' \sim \pi_e(s_2')}[k_e^{\pi}(s_1', a_1'; s_2', a_2')]$$

- **State-action pairs that are similar under this metric have similar q^{π_e} values.**

1. Castro et al. 2023. A Kernel Perspective on Behavioural Metrics for Markov Decision Processes.

Our Work: Kernel Representations for OPE (KROPE)

- Builds upon Kernel Similarity Metric (KSMe) [1].
- KROPE similarity metric (short-term + long-term similarity):

$$k^{\pi_e}(s_1, a_1; s_2, a_2) := 1 - \frac{|r(s_1, a_1) - r(s_2, a_2)|}{|r_{\max} - r_{\min}|} + \gamma \mathbb{E}_{a_1' \sim \pi_e(s_1'), a_2' \sim \pi_e(s_2')}[k_e^{\pi}(s_1', a_1'; s_2', a_2')]$$

- **State-action pairs that are similar under this metric have similar q^{π_e} values.**
- Under function approximation, learn features: $k^{\pi_e}(s_1, a_1; s_2, a_2) = \phi(s_1, a_1)^\top \phi(s_2, a_2)$

1. Castro et al. 2023. A Kernel Perspective on Behavioural Metrics for Markov Decision Processes.

Theoretical Analysis Highlights

Theoretical Analysis Highlights

$$\mathbb{E}_{\mathcal{D}}[\Phi\Phi^\top] = \mathbb{E}_{\mathcal{D}}[K_1] + \gamma\mathbb{E}_{\mathcal{D},\pi_e}[P^{\pi_e}\Phi(P^{\pi_e}\Phi)^\top]$$

1. Castro et al. 2023. A Kernel Perspective on Behavioural Metrics for Markov Decision Processes.

Theoretical Analysis Highlights

$$\mathbb{E}_{\mathcal{D}}[\Phi\Phi^\top] = \mathbb{E}_{\mathcal{D}}[K_1] + \gamma\mathbb{E}_{\mathcal{D},\pi_e}[P^{\pi_e}\Phi(P^{\pi_e}\Phi)^\top]$$

Theorem 1: **LSPE will converge to its fixed point solution.**

1. Castro et al. 2023. A Kernel Perspective on Behavioural Metrics for Markov Decision Processes.

Theoretical Analysis Highlights

$$\mathbb{E}_{\mathcal{D}}[\Phi\Phi^\top] = \mathbb{E}_{\mathcal{D}}[K_1] + \gamma\mathbb{E}_{\mathcal{D},\pi_e}[P^{\pi_e}\Phi(P^{\pi_e}\Phi)^\top]$$

Theorem 1: **LSPE will converge to its fixed point solution.**

Theorem 2: KROPE state-action features are **Bellman Complete**.

1. Castro et al. 2023. A Kernel Perspective on Behavioural Metrics for Markov Decision Processes.

Empirical Analysis Highlights

Empirical Analysis Highlights

	Algorithm							
Dataset (DMC)	FQE	BCRL+EXP	BCRL	BEER	DR3	DBC	ROPE	KROPE (ours)
CartPoleSwingUp	Div.	2.0 ± 1.6	2.2 ± 0.8	Div.	0.9 ± 0.0	Div.	0.2 ± 0.1	$\mathbf{0.0 \pm 0.0}$
CheetahRun	$\mathbf{0.0 \pm 0.0}$	0.3 ± 0.2	0.8 ± 0.3	$\mathbf{0.0 \pm 0.0}$	0.4 ± 0.0	Div.	Div.	$\mathbf{0.0 \pm 0.0}$
FingerEasy	Div.	0.6 ± 0.1	0.8 ± 0.2	Div.	0.9 ± 0.0	Div.	$\mathbf{0.1 \pm 0.0}$	0.6 ± 0.0
WalkerStand	$\mathbf{0.0 \pm 0.0}$	0.2 ± 0.2	0.2 ± 0.1	1.9 ± 3.6	0.1 ± 0.0	Div.	0.2 ± 0.0	$\mathbf{0.0 \pm 0.0}$
Dataset (D4RL)	FQE	BCRL+EXP	BCRL	BEER	DR3	DBC	ROPE	KROPE (ours)
cheetah random	$\mathbf{0.9 \pm 0.0}$	Div.	Div.	$\mathbf{0.9 \pm 0.0}$	$\mathbf{0.9 \pm 0.0}$	$\mathbf{0.9 \pm 0.0}$	1.0 ± 0.0	1.0 ± 0.0
cheetah medium	Div.	Div.	0.2 ± 0.2	Div.	Div.	Div.	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$
cheetah med-expert	Div.	0.2 ± 0.1	0.3 ± 0.1	Div.	Div.	Div.	0.1 ± 0.0	$\mathbf{0.0 \pm 0.0}$
hopper random	Div.	Div.	Div.	Div.	0.8 ± 0.0	Div.	Div.	$\mathbf{0.1 \pm 0.0}$
hopper medium	Div.	Div.	Div.	Div.	Div.	Div.	Div.	Div.
hopper med-expert	Div.	Div.	Div.	Div.	0.6 ± 0.0	Div.	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$
walker random	Div.	Div.	Div.	Div.	1.0 ± 0.0	Div.	Div.	$\mathbf{0.5 \pm 0.1}$
walker medium	Div.	Div.	Div.	Div.	Div.	Div.	Div.	Div.
walker med-expert	Div.	1.3 ± 0.4	2.6 ± 2.1	Div.	6.6 ± 11.6	Div.	$\mathbf{0.1 \pm 0.0}$	Div.

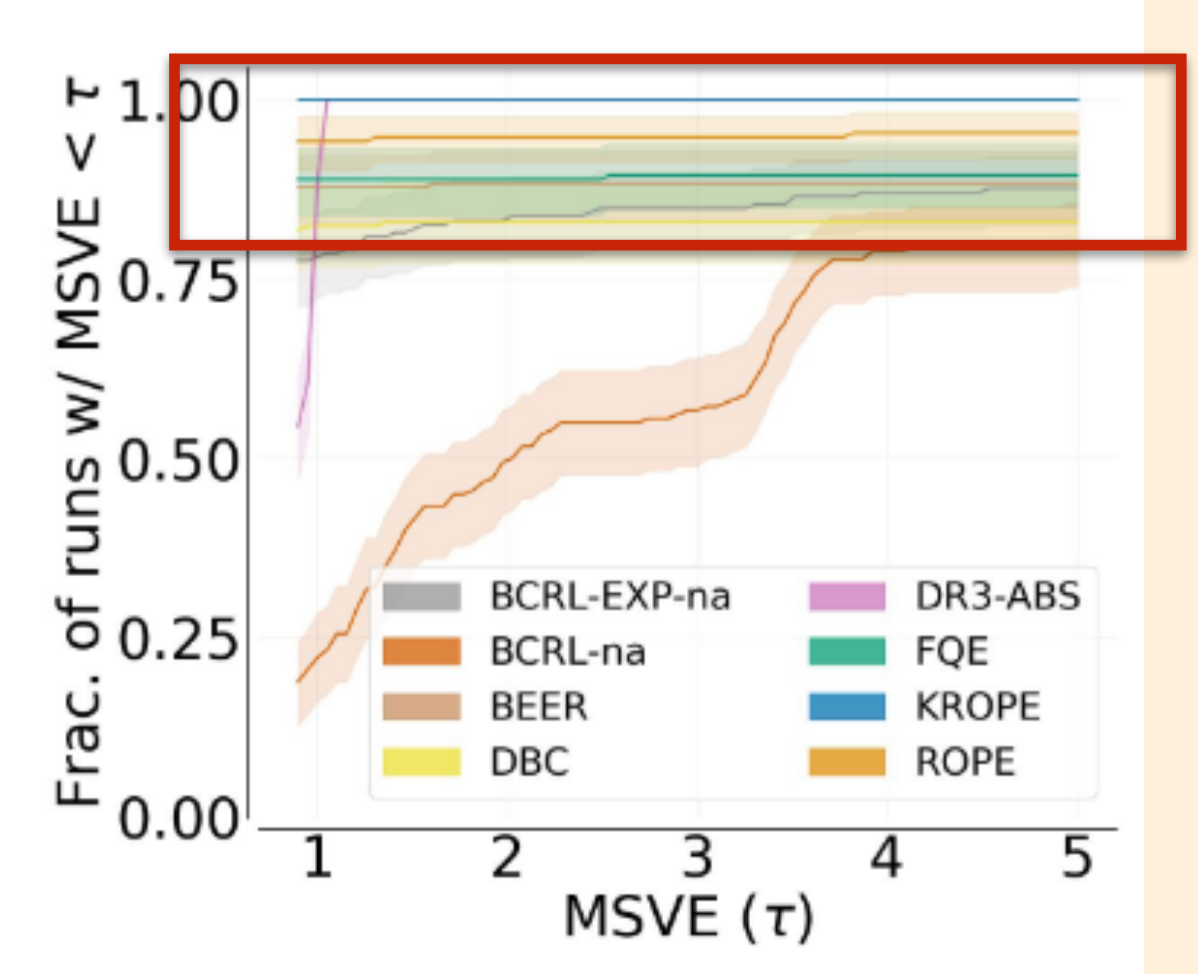
Lower OPE error than 1) other bisimulation, 2) model-based, and 3) co-adaptation based methods

Empirical Analysis Highlights

Dataset (DMC)	Algorithm: FQE	BCRL+EXP	BCRL	BEER	DR3	DBC	ROPE	KROPE (ours)
CartPoleSwingUp	Div.	2.0 ± 1.6	2.2 ± 0.8	Div.	0.9 ± 0.0	Div.	0.2 ± 0.1	$\mathbf{0.0 \pm 0.0}$
CheetahRun	$\mathbf{0.0 \pm 0.0}$	0.3 ± 0.2	0.8 ± 0.3	$\mathbf{0.0 \pm 0.0}$	0.4 ± 0.0	Div.	Div.	$\mathbf{0.0 \pm 0.0}$
FingerEasy	Div.	0.6 ± 0.1	0.8 ± 0.2	Div.	0.9 ± 0.0	Div.	$\mathbf{0.1 \pm 0.0}$	0.6 ± 0.0
WalkerStand	$\mathbf{0.0 \pm 0.0}$	0.2 ± 0.2	0.2 ± 0.1	1.9 ± 3.6	0.1 ± 0.0	Div.	0.2 ± 0.0	$\mathbf{0.0 \pm 0.0}$
Dataset (D4RL)	**FQE**	**BCRL+EXP**	**BCRL**	**BEER**	**DR3**	**DBC**	**ROPE**	**KROPE (ours)**
cheetah random	$\mathbf{0.9 \pm 0.0}$	Div.	Div.	$\mathbf{0.9 \pm 0.0}$	$\mathbf{0.9 \pm 0.0}$	$\mathbf{0.9 \pm 0.0}$	1.0 ± 0.0	1.0 ± 0.0
cheetah medium	Div.	Div.	0.2 ± 0.2	Div.	Div.	Div.	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$
cheetah med-expert	Div.	0.2 ± 0.1	0.3 ± 0.1	Div.	Div.	Div.	0.1 ± 0.0	$\mathbf{0.0 \pm 0.0}$
hopper random	Div.	Div.	Div.	Div.	0.8 ± 0.0	Div.	Div.	$\mathbf{0.1 \pm 0.0}$
hopper medium	Div.	Div.	Div.	Div.	Div.	Div.	Div.	Div.
hopper med-expert	Div.	Div.	Div.	Div.	0.6 ± 0.0	Div.	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$
walker random	Div.	Div.	Div.	Div.	1.0 ± 0.0	Div.	Div.	$\mathbf{0.5 \pm 0.1}$
walker medium	Div.	Div.	Div.	Div.	Div.	Div.	Div.	Div.
walker med-expert	Div.	1.3 ± 0.4	2.6 ± 2.1	Div.	6.6 ± 11.6	Div.	$\mathbf{0.1 \pm 0.0}$	Div.

Lower OPE error than 1) other bisimulation, 2) model-based, and 3) co-adaptation based methods

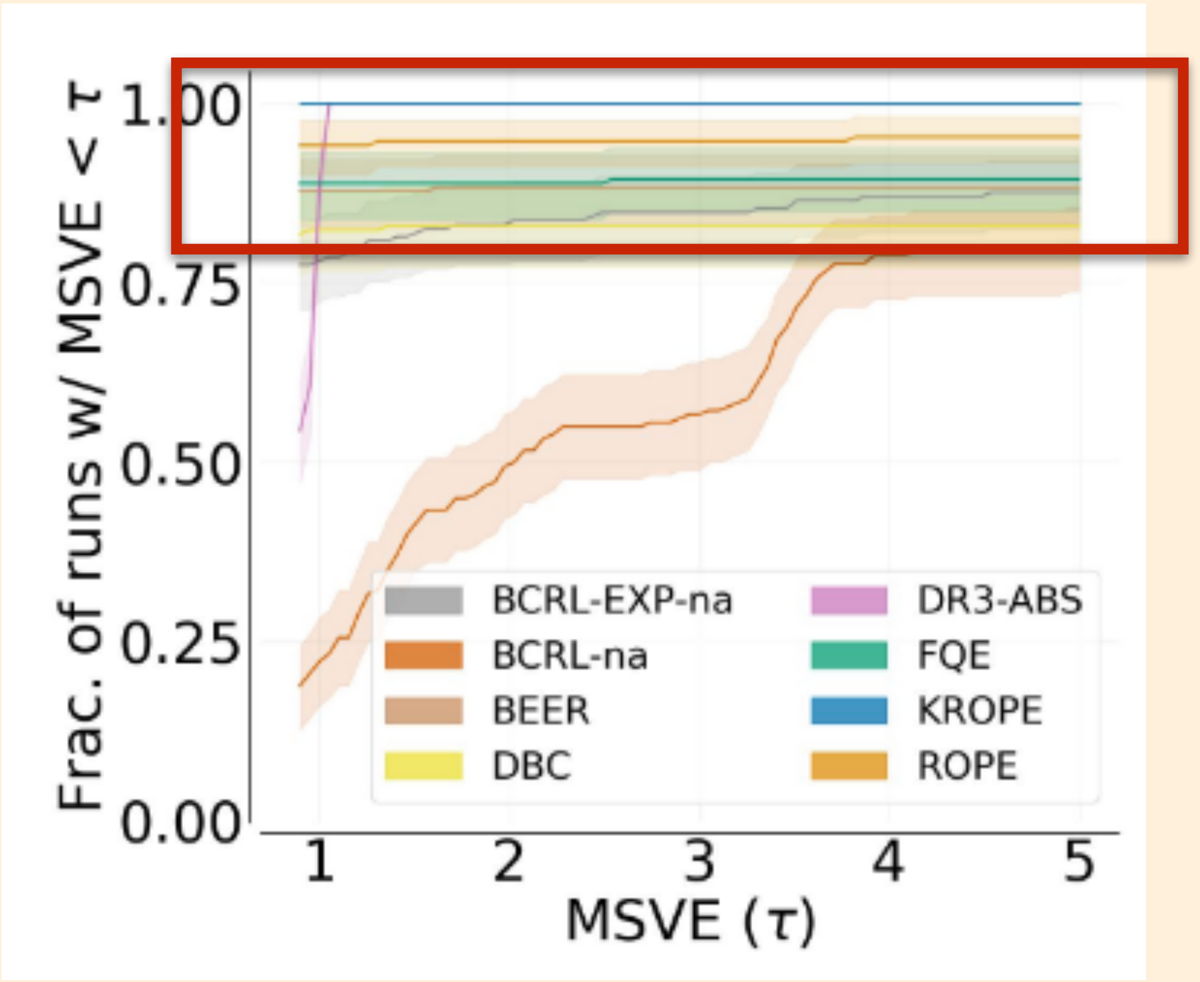
Robust across hyperparameters



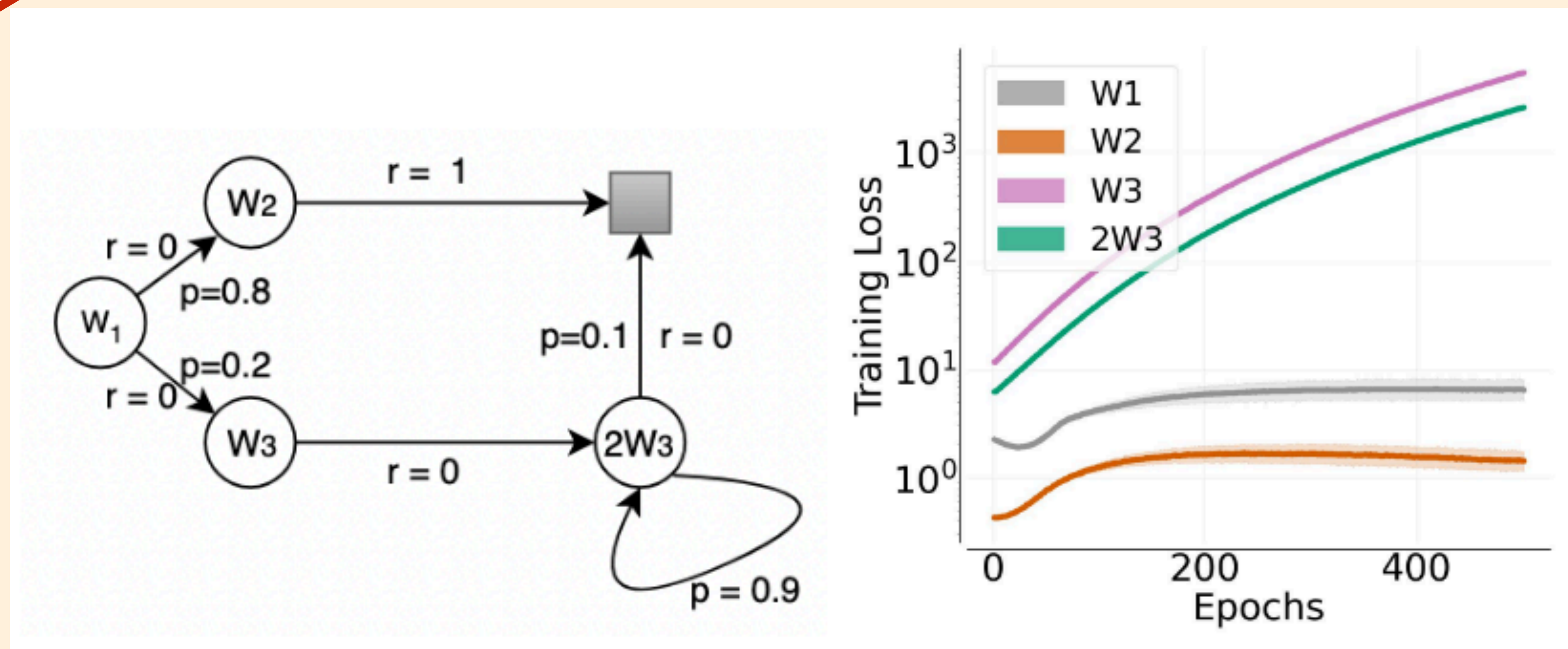
Empirical Analysis Highlights

	Algorithm							
Dataset (DMC)	FQE	BCRL+EXP	BCRL	BEER	DR3	DBC	ROPE	KROPE (ours)
CartPoleSwingUp	Div.	2.0 ± 1.6	2.2 ± 0.8	Div.	0.9 ± 0.0	Div.	0.2 ± 0.1	**0.0 ± 0.0**
CheetahRun	**0.0 ± 0.0**	0.3 ± 0.2	0.8 ± 0.3	**0.0 ± 0.0**	0.4 ± 0.0	Div.	Div.	**0.0 ± 0.0**
FingerEasy	Div.	0.6 ± 0.1	0.8 ± 0.2	Div.	0.9 ± 0.0	Div.	**$0.1 \pm 0.$**	0.6 ± 0.0
WalkerStand	**0.0 ± 0.0**	0.2 ± 0.2	0.2 ± 0.1	1.9 ± 3.6	0.1 ± 0.0	Div.	0.2 ± 0.0	**0.0 ± 0.0**
Dataset (D4RL)	FQE	BCRL+EXP	BCRL	BEER	DR3	DBC	ROPE	KROPE (ours)
cheetah random	**0.9 ± 0.0**	Div.	Div.	**0.9 ± 0.0**	**0.9 ± 0.0**	**0.9 ± 0.0**	1.0 ± 0.0	1.0 ± 0.0
cheetah medium	Div.	Div.	0.2 ± 0.2	Div.	Div.	Div.	**0.0 ± 0.0**	**0.0 ± 0.0**
cheetah med-expert	Div.	0.2 ± 0.1	0.3 ± 0.1	Div.	Div.	Div.	0.1 ± 0.0	**0.0 ± 0.0**
hopper random	Div.	Div.	Div.	Div.	0.8 ± 0.0	Div.	Div.	**0.1 ± 0.0**
hopper medium	Div.	Div.	Div.	Div.	Div.	Div.	Div.	Div.
hopper med-expert	Div.	Div.	Div.	Div.	0.6 ± 0.0	Div.	**$0.0 \pm 0.$**	**0.0 ± 0.0**
walker random	Div.	Div.	Div.	Div.	1.0 ± 0.0	Div.	Div.	**0.5 ± 0.1**
walker medium	Div.	Div.	Div.	Div.	Div.	Div.	Div.	Div.
walker med-expert	Div.	1.3 ± 0.4	2.6 ± 2.1	Div.	6.6 ± 11.6	Div.	**$0.1 \pm 0.$**	Div.

Robust across hyperparameters



Lower OPE error than 1) other bisimulation, 2) model-based, and 3) co-adaptation based methods



Divergence analysis: representation learning vs. direct value function learning?

Summary of Contributions

Summary of Contributions

- A theoretical understanding of the benefits of **bisimulation-based representations** for **stable offline policy evaluation**.

Summary of Contributions

- A theoretical understanding of the benefits of **bisimulation-based representations** for **stable offline policy evaluation**.
- An empirical analysis showing **improved OPE accuracy** and **hyperparameter robustness**.

Summary of Contributions

- A theoretical understanding of the benefits of **bisimulation-based representations** for **stable offline policy evaluation**.
- An empirical analysis showing **improved OPE accuracy** and **hyperparameter robustness**.
- A better understanding of when bootstrapping-based representation learning may **converge** in settings where value function-based bootstrapping may diverge.

Thank you!

Brahma S. Pavse Yudong Chen Qiaomin Xie Josiah P. Hanna

Paper:

University of Wisconsin — Madison

pavse@wisc.edu