Koopman Q-learning: Offline Reinforcement learning Via Symmetries of Dynamics

Matthias Weissenbacher [1*] **Samarth Sinha** [2] **Animesh Garg** [2]

Yoshinobu Kawahara [1,3]

[1] RIKEN Center for Advanced Intelligence Project, [2] University of Toronto, Vector Institute, [3] Institute of Mathematics for Industry, Kyushu University

Overview and Motivation

Offline RL uses static training data

Further environment exploration not possible

Overview and Motivation

Offline RL uses static training data

Further environment exploration not possible

+

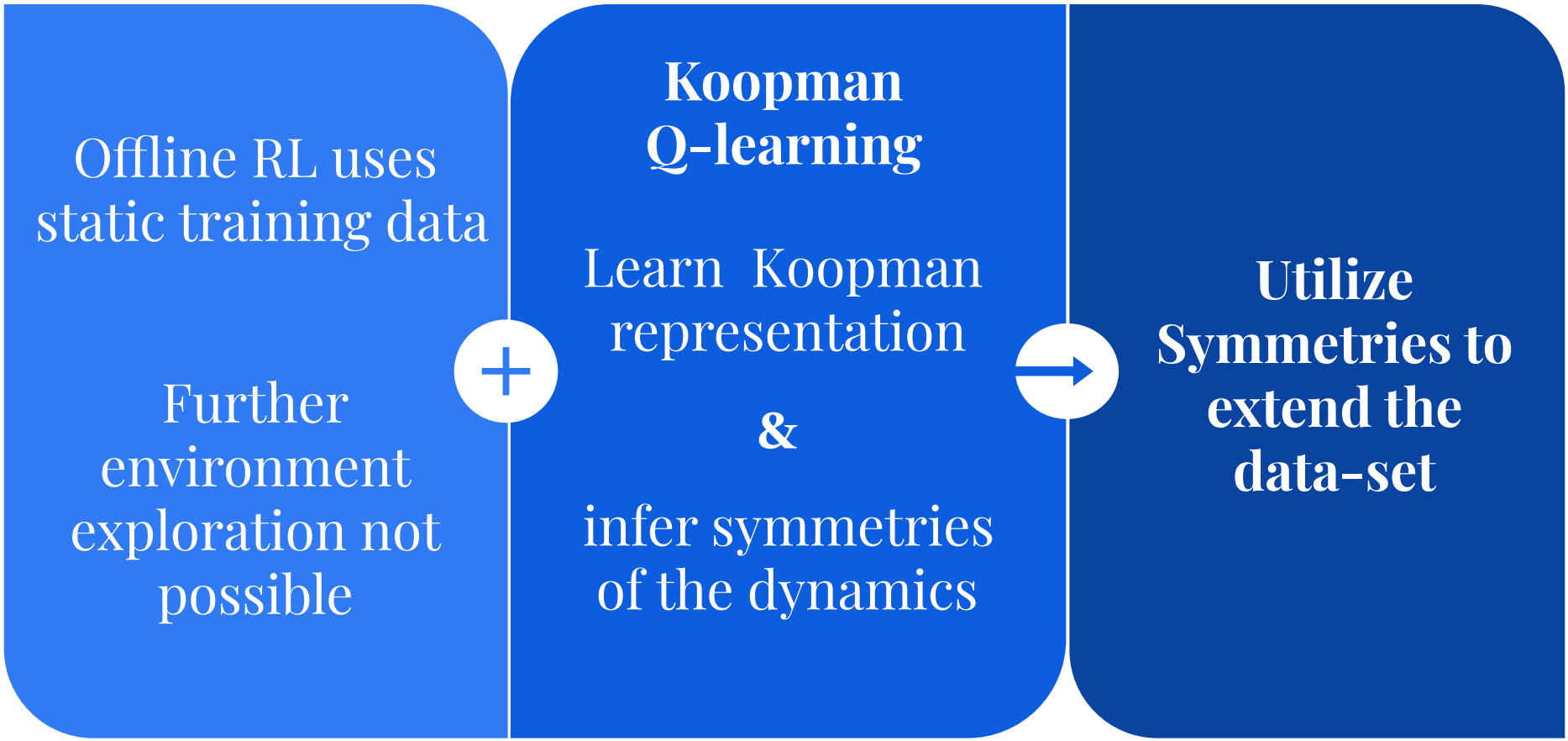
Koopman Q-learning

Learn Koopman representation

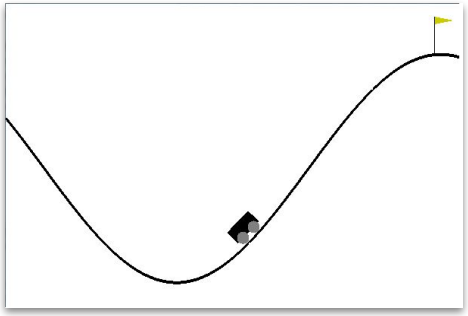
&

infer symmetries of the dynamics

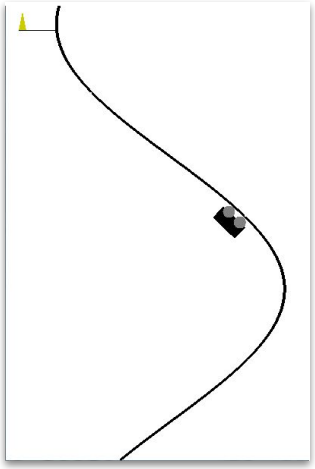
Overview and Motivation



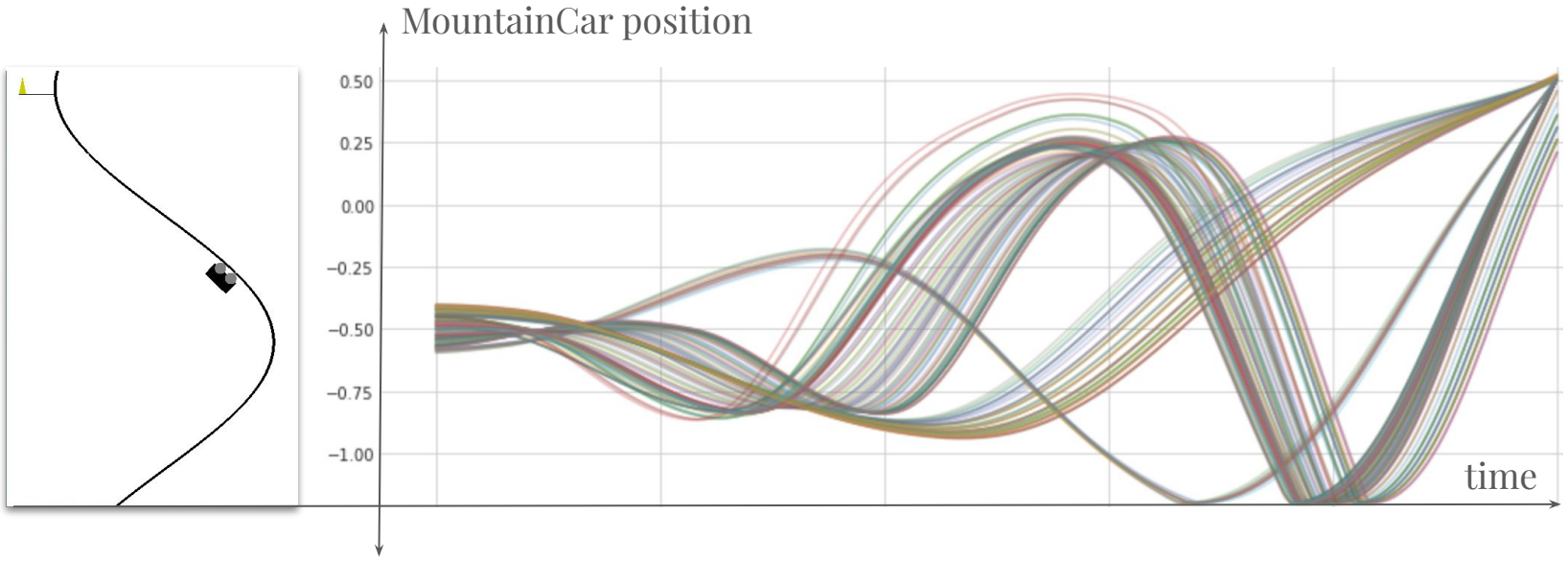
In Koopman space the Dynamics are linear



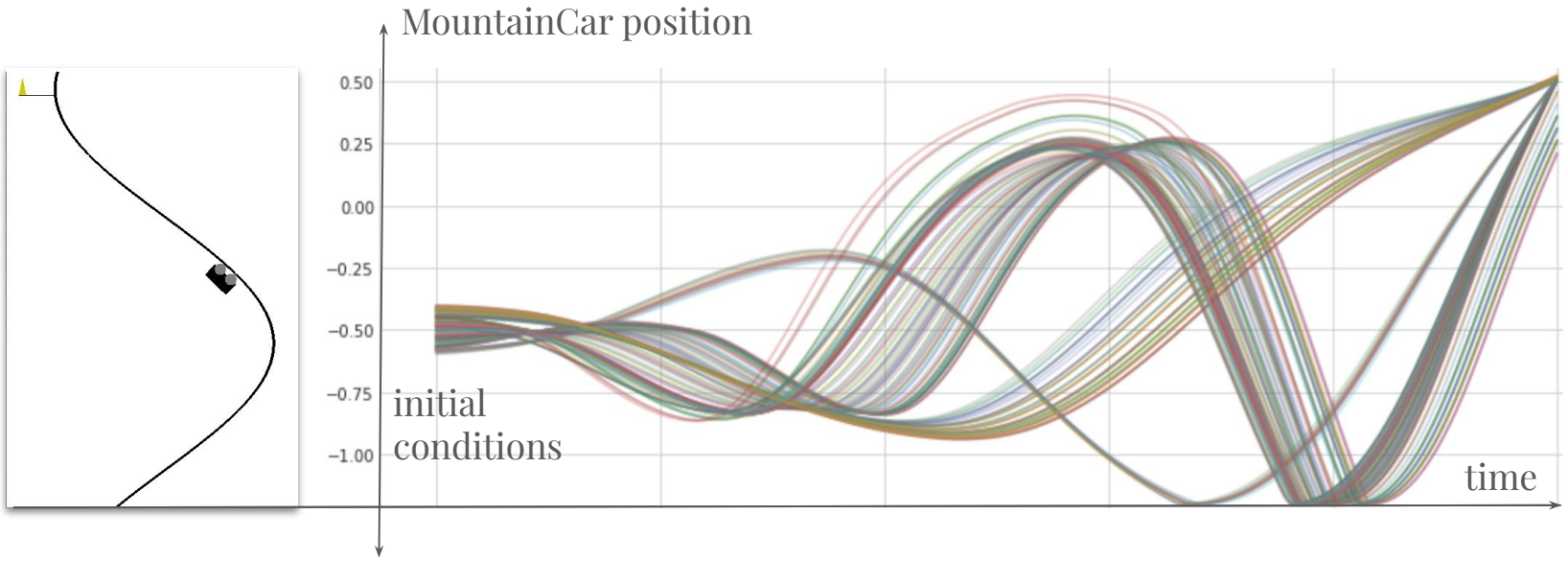
In Koopman space the Dynamics are linear



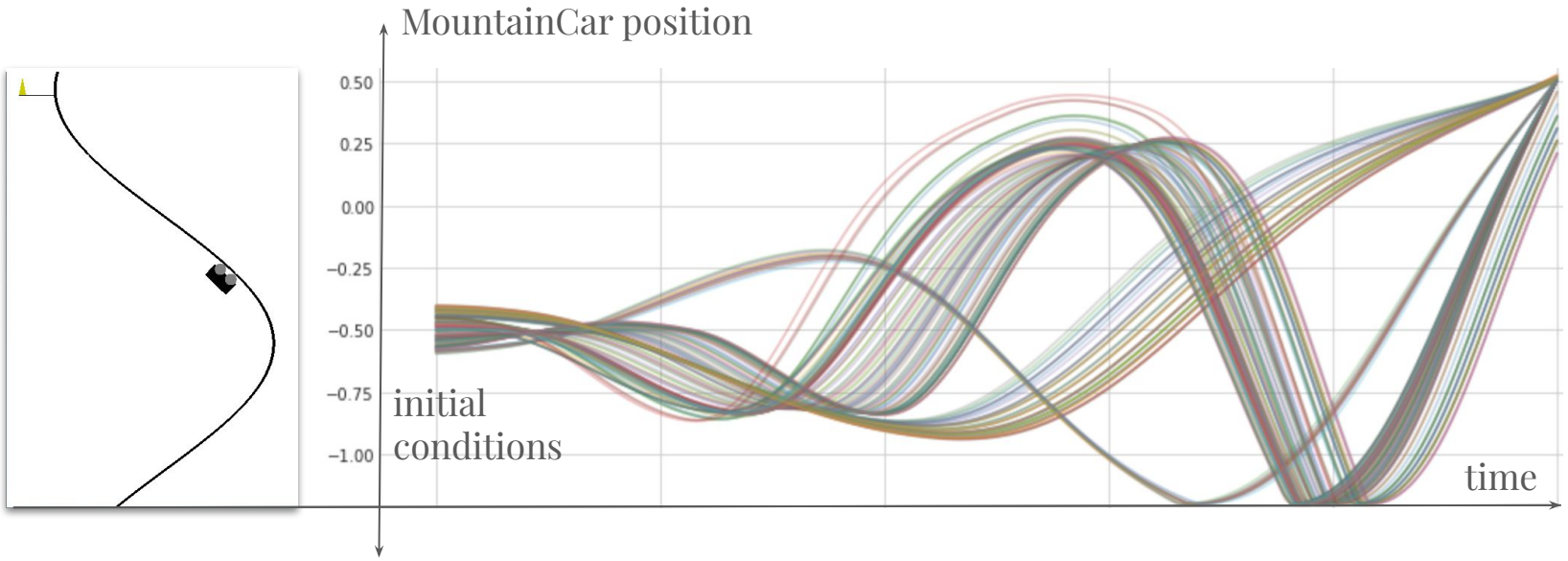
In Koopman space the Dynamics are linear



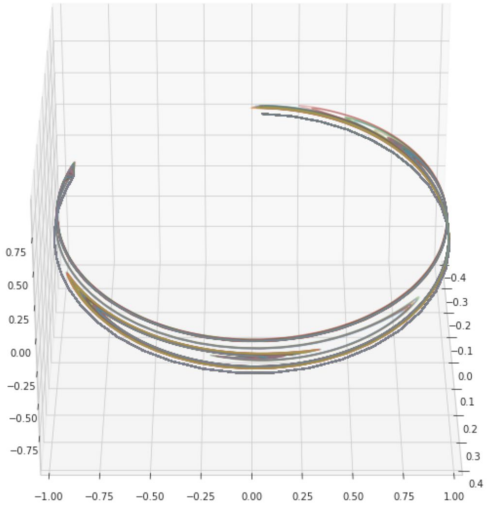
In Koopman space the Dynamics are linear

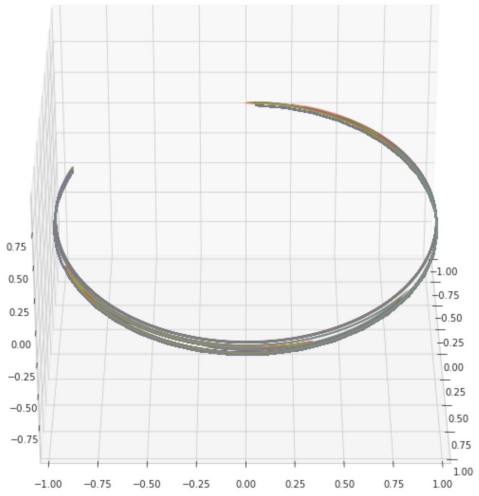


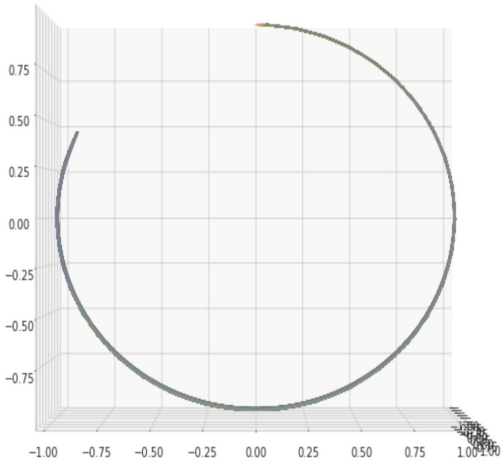
In Koopman space the Dynamics are linear



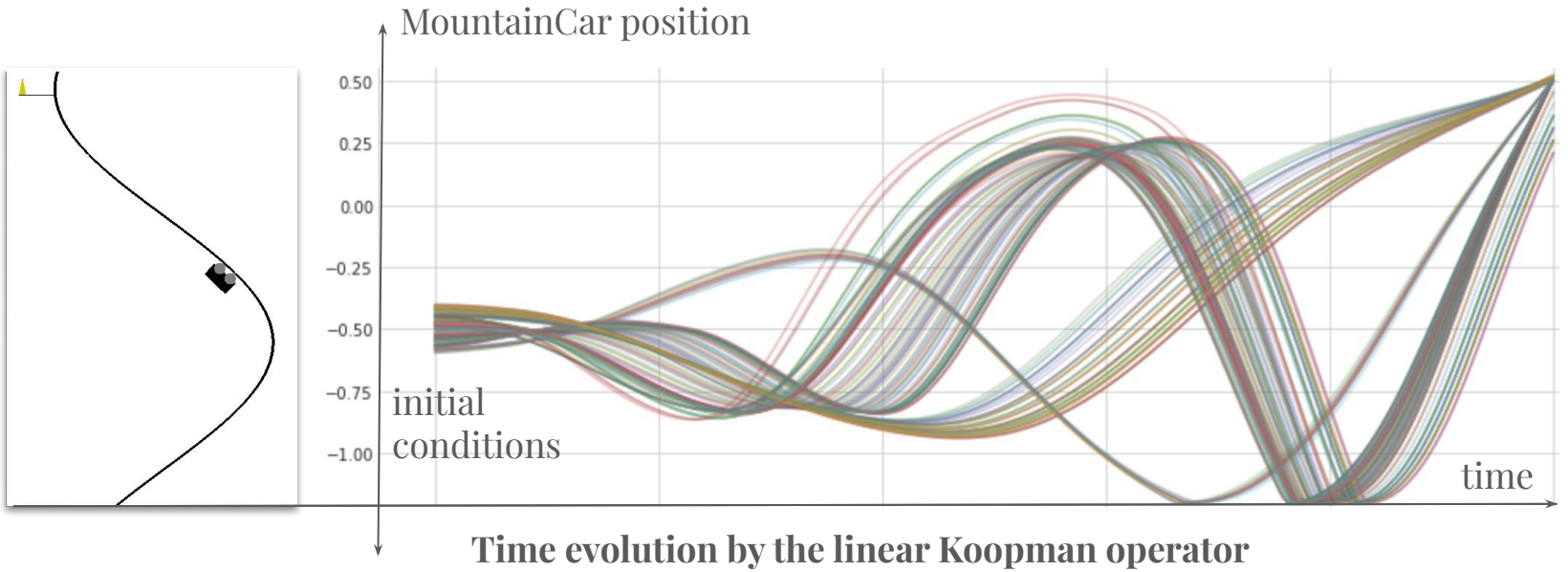
Koopman
observables





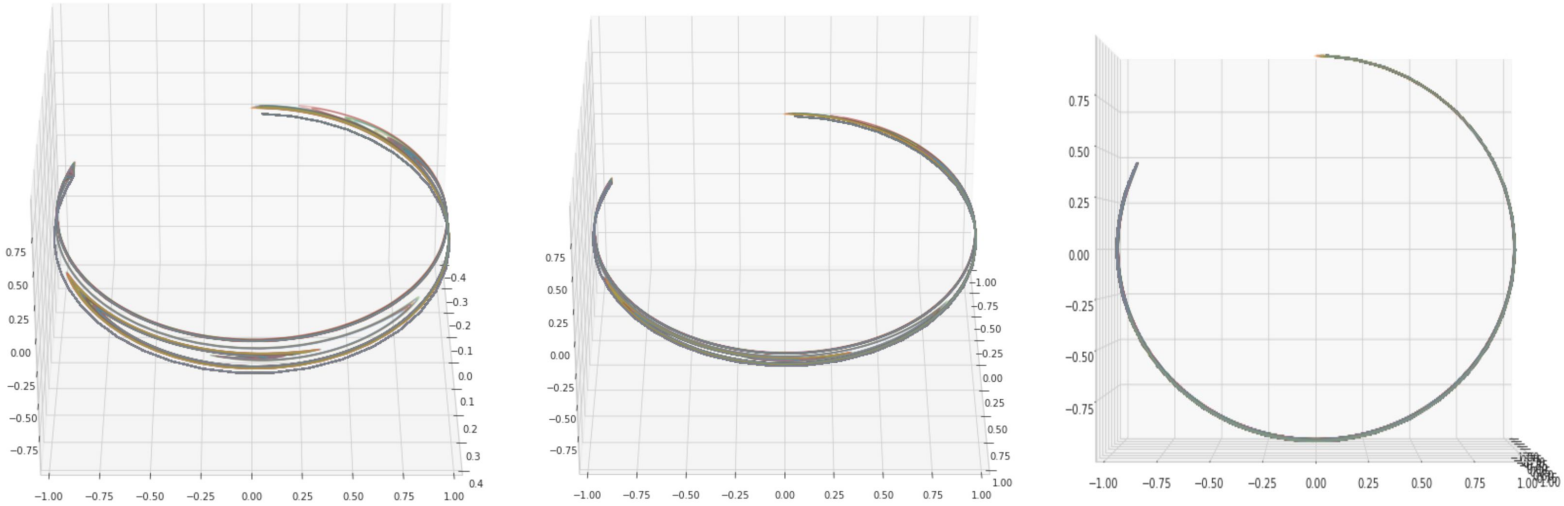


In Koopman space the Dynamics are linear



Time evolution by the linear Koopman operator

Koopman
observables



Symmetries

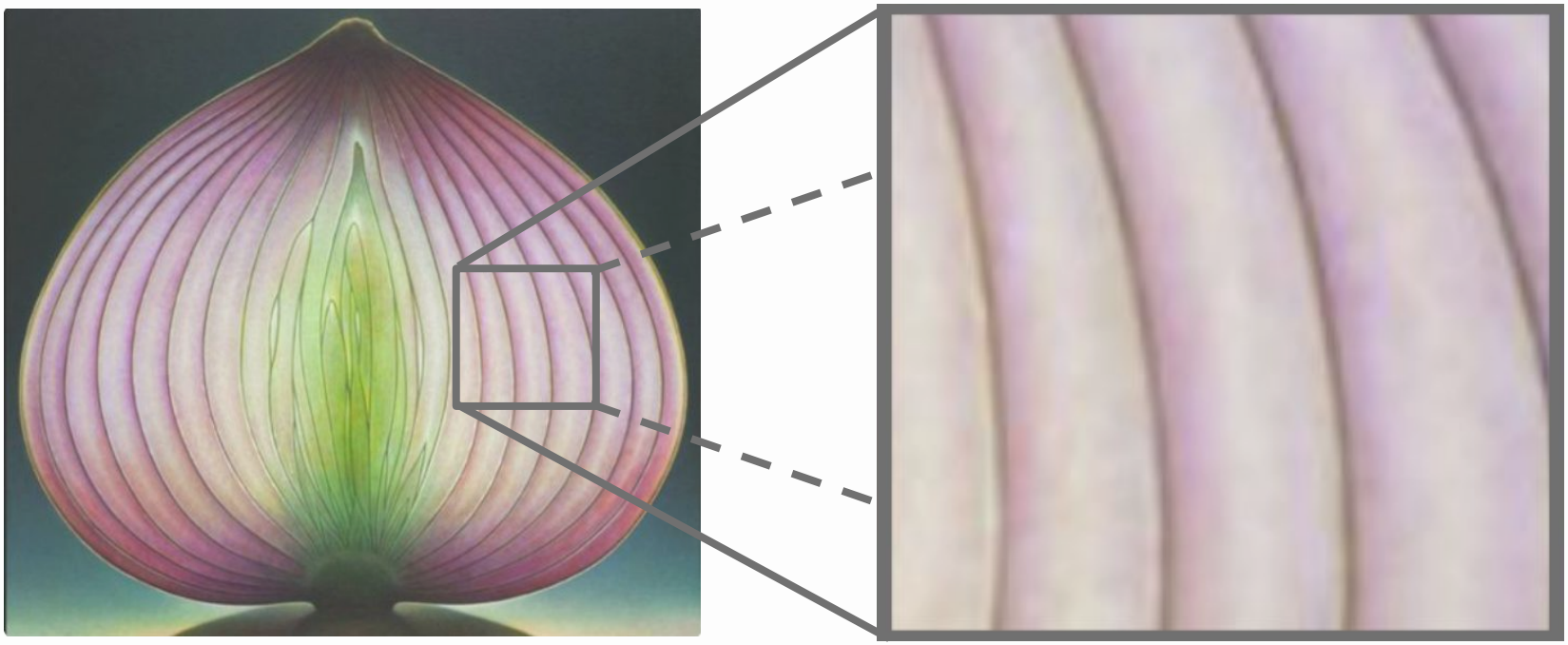
Rotation

Mirror flip

Global and Local Symmetries

Global
Mirror flip

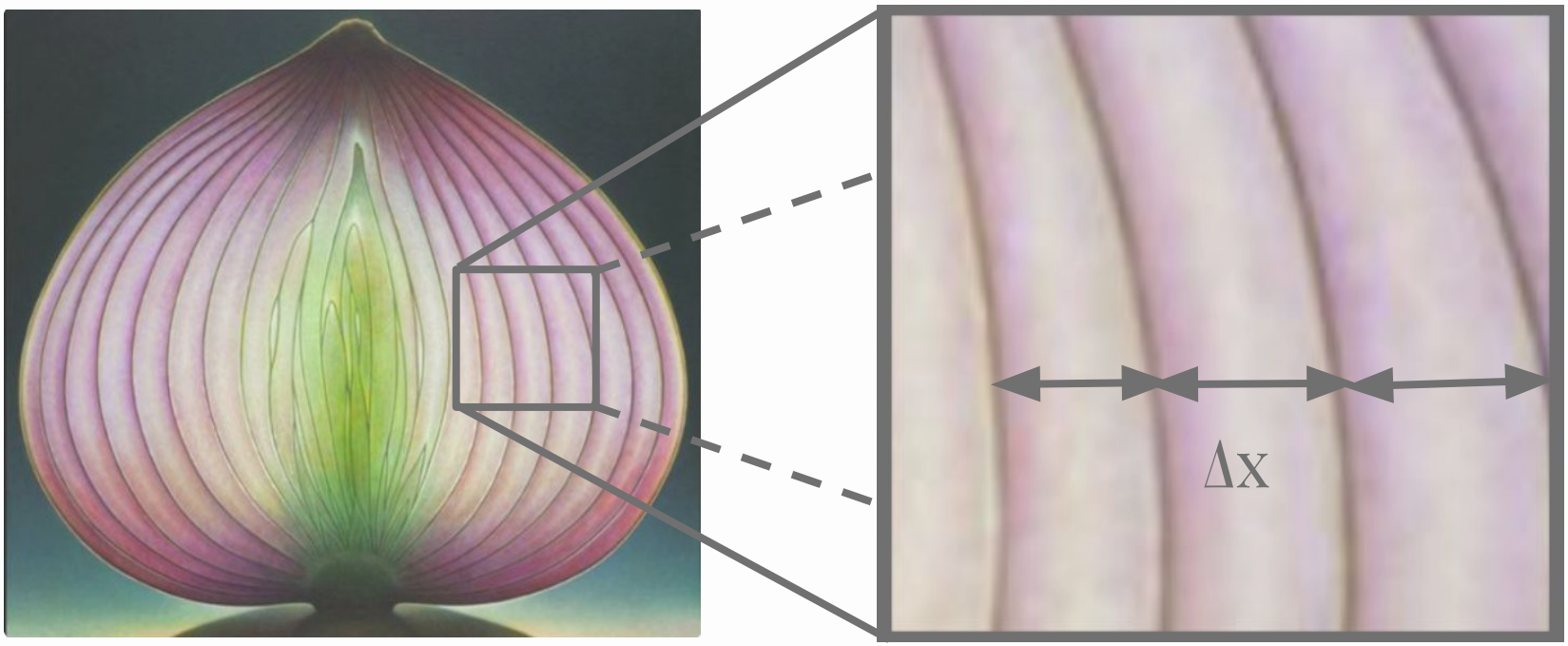
Global and Local Symmetries



Global
Mirror flip

Local

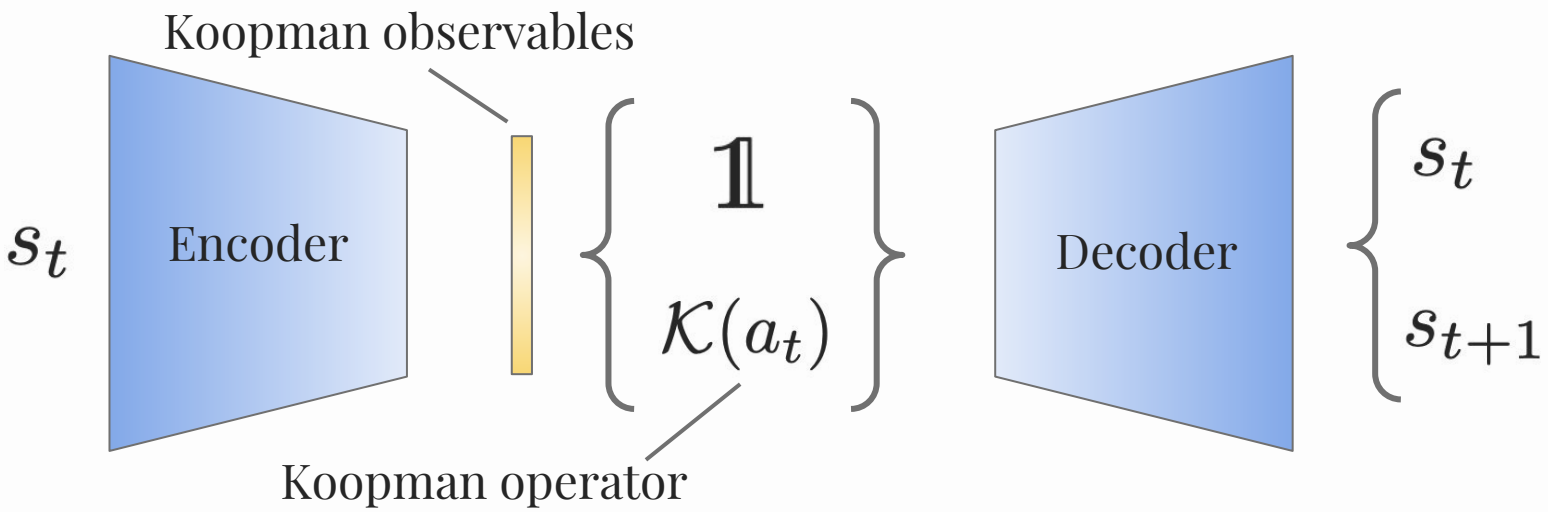
Global and Local Symmetries



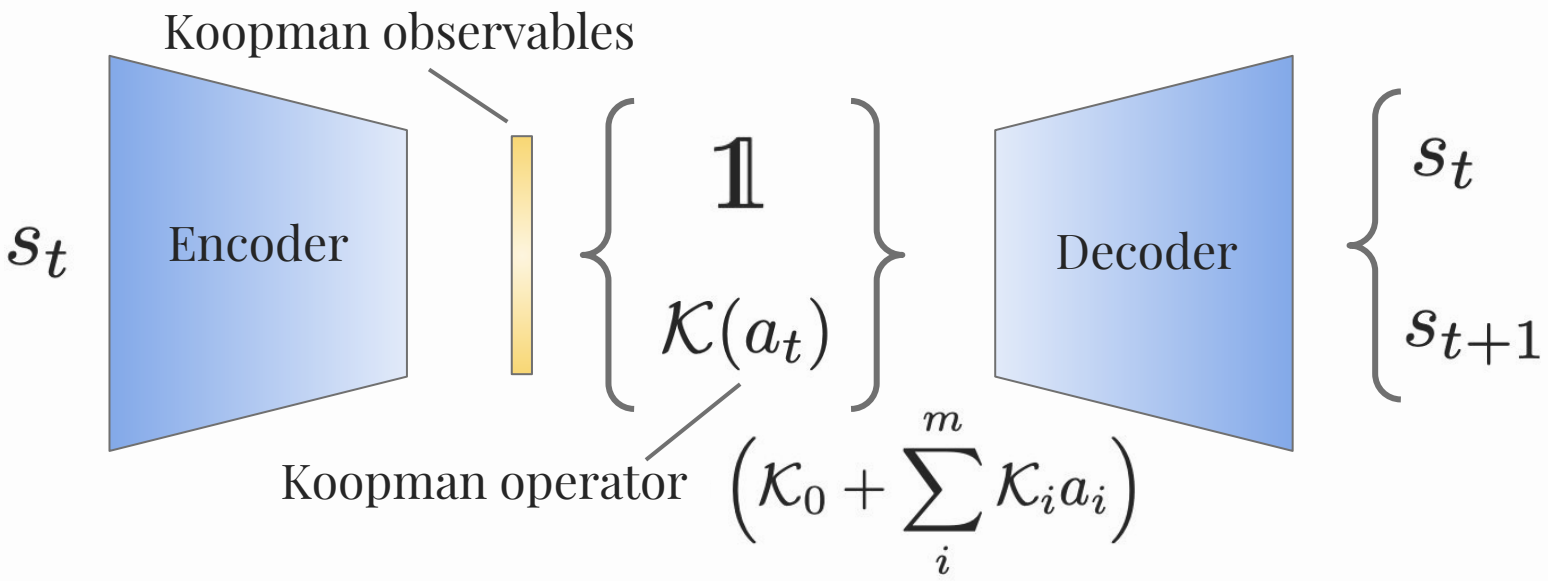
Global
Mirror flip

Local
shift symmetry by Δx

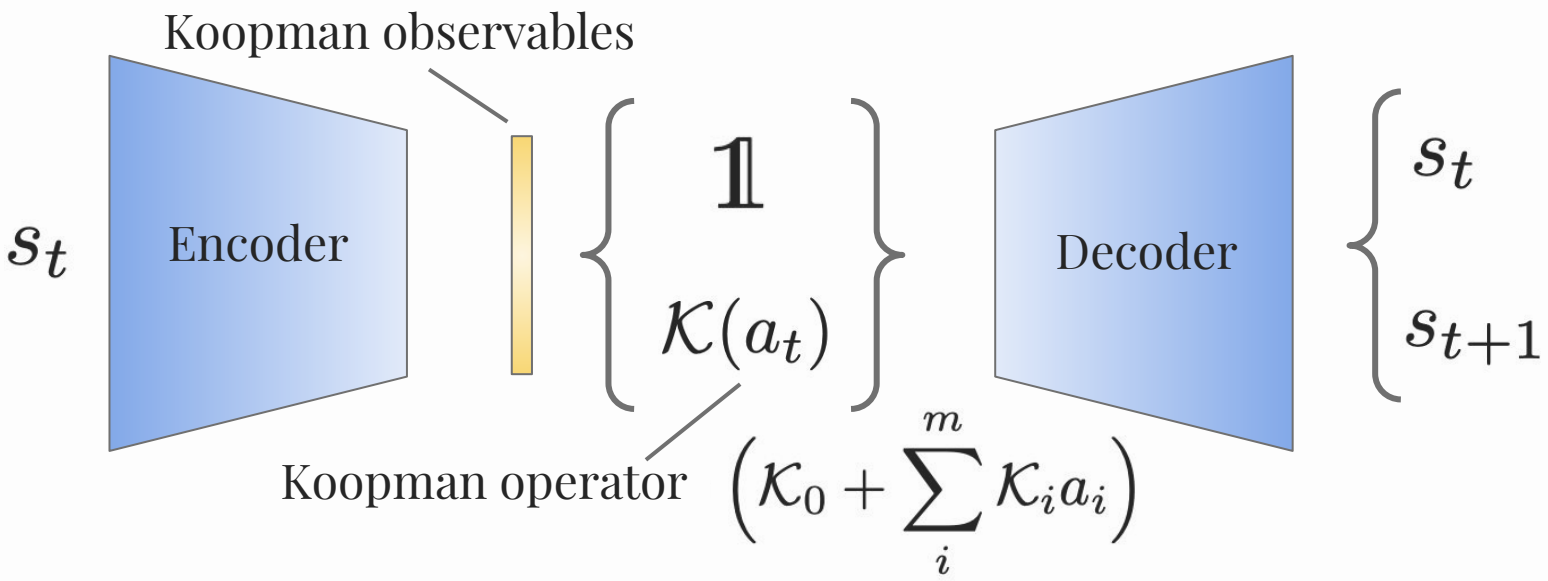
Learn Koopman Latent Space Representation



Learn Koopman Latent Space Representation



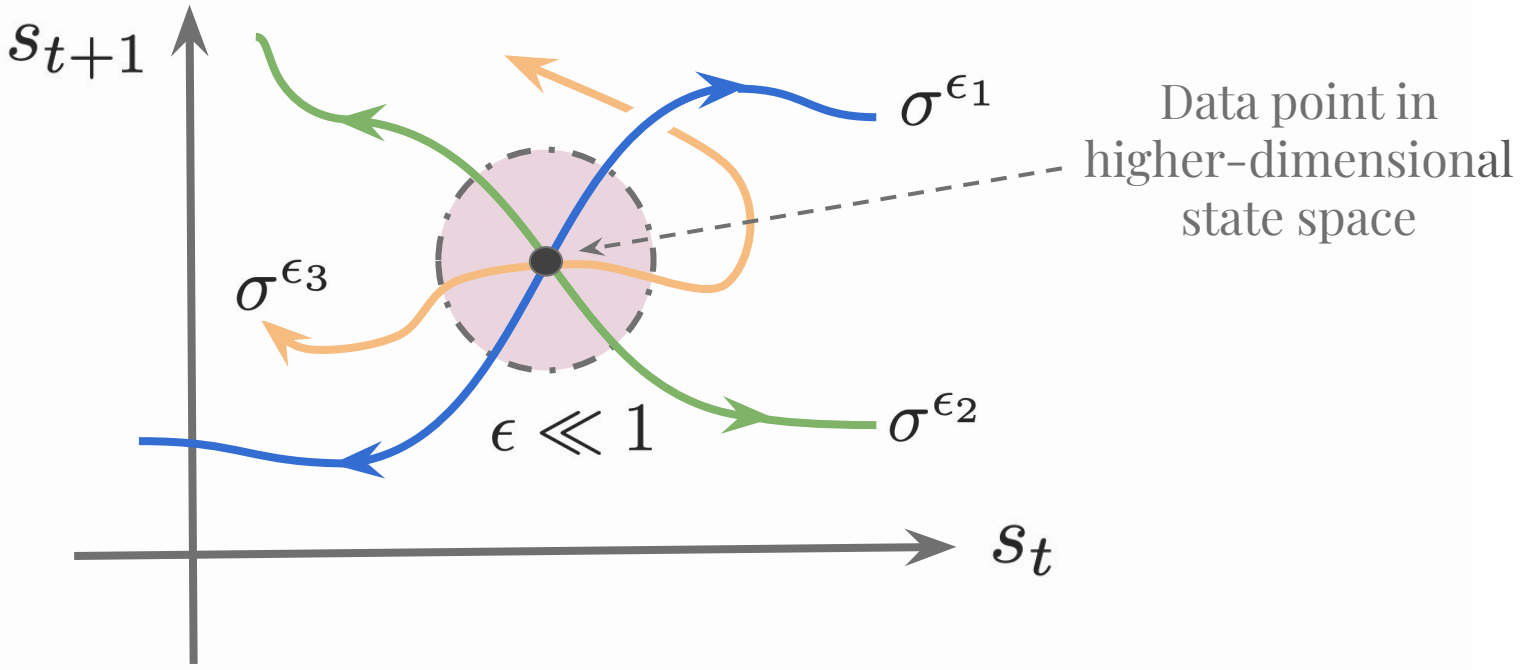
Learn Koopman Latent Space Representation



Derive Symmetries of Dynamics

σ^ϵ is symmetry $\Longleftrightarrow$ commutes with $\mathcal{K}(a_t)$

Apply Symmetry Shifts to Offline RL Data



D4RL

Fu et. al

S4RL - random shift

Sinha et. al

Ours - symmetry shift

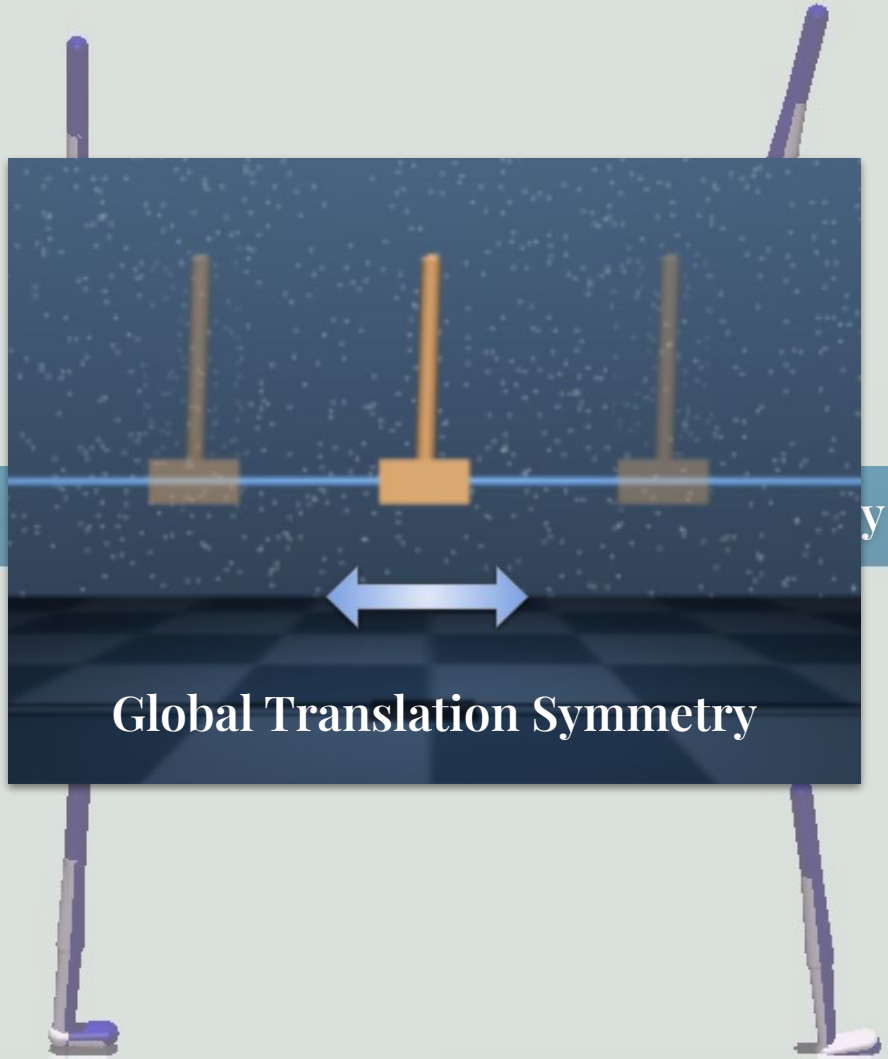
Ours - symmetry shift large

D4RL

S4RL - random shift

Ours -

y shift large



Policy Optimization step Q-learning (CQL)

Kumar et. al

Bellman error

$$\hat{Q}_{i+1} \leftarrow \arg\min_{Q} \left(\mathop{\mathbb{E}}_{s_t, a_t, s_{t+1} \sim \mathcal{D}} \left[\left(r_t + \gamma \hat{Q}_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_{t+1} | s_{t+1} \right), a_{t+1} \right) - Q_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_t | s_t \right), a_t \right) \right)^2 \right] \right.$$
$$\left. + \tilde{\alpha} \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}} \left[\log \sum_{a} \exp \left(Q_i(s_t, a) \right) - \mathop{\mathbb{E}}_{a \sim \pi(s_t)} \left[Q_i(s_t, a) \right] \right] \right)$$

$$\tilde{s}_t = D\Big(\left(\mathbb{1} + \epsilon \sigma_{a_t} \right) \hat{*} E(s_t) \Big)$$

Policy Optimization step Q-learning (CQL)

Kumar et. al

Bellman error

$$\hat{Q}_{i+1} \leftarrow \arg\min_{Q} \left(\mathop{\mathbb{E}}_{s_t, a_t, s_{t+1} \sim \mathcal{D}} \left[\left(r_t + \gamma \hat{Q}_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_{t+1} | s_{t+1} \right), a_{t+1} \right) - Q_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_t | s_t \right), a_t \right) \right)^2 \right] \right.$$
$$\left. + \tilde{\alpha} \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}} \left[\log \sum_{a} \exp \left(Q_i(s_t, a) \right) - \mathop{\mathbb{E}}_{a \sim \pi(s_t)} \left[Q_i(s_t, a) \right] \right] \right)$$

$$\tilde{s}_t = D\Big(\left(\mathbb{1} + \epsilon \sigma_{a_t} \right) \hat{*} E(s_t) \Big)$$

Koopman space embedding

Policy Optimization step Q-learning (CQL)

Kumar et. al

Bellman error

$$\hat{Q}_{i+1} \leftarrow \arg\min_{Q} \left(\mathop{\mathbb{E}}_{s_t, a_t, s_{t+1} \sim \mathcal{D}} \left[\left(r_t + \gamma \hat{Q}_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_{t+1} | s_{t+1} \right), a_{t+1} \right) - Q_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_t | s_t \right), a_t \right) \right)^2 \right] \right.$$
$$\left. + \tilde{\alpha} \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}} \left[\log \sum_{a} \exp \left(Q_i(s_t, a) \right) - \mathop{\mathbb{E}}_{a \sim \pi(s_t)} \left[Q_i(s_t, a) \right] \right] \right)$$

$$\tilde{s}_t = D\Big(\big(\mathbb{1} + \epsilon \sigma_{a_t} \big) \hat{*} E(s_t) \Big)$$

Symmetry shift

Koopman space embedding

Policy Optimization step Q-learning (CQL)

Kumar et. al

Bellman error

$$\hat{Q}_{i+1} \leftarrow \arg\min_{Q} \left(\mathop{\mathbb{E}}_{s_t, a_t, s_{t+1} \sim \mathcal{D}} \left[\left(r_t + \gamma \hat{Q}_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_{t+1} | s_{t+1} \right), a_{t+1} \right) - Q_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_t | s_t \right), a_t \right) \right)^2 \right] \right.$$
$$\left. + \tilde{\alpha} \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}} \left[\log \sum_{a} \exp \left(Q_i(s_t, a) \right) - \mathop{\mathbb{E}}_{a \sim \pi(s_t)} \left[Q_i(s_t, a) \right] \right] \right)$$

$$\tilde{s}_t = D\Big(\underbrace{(\mathbb{1} + \epsilon \sigma_{a_t})}_{\text{Symmetry shift}} \hat{*} \underbrace{E(s_t)}_{\text{Koopman space embedding}} \Big)$$

inverse embedding

Policy Optimization step Q-learning (CQL)

Kumar et. al

Bellman error

$$\hat{Q}_{i+1} \leftarrow \arg\min_{Q} \left(\mathop{\mathbb{E}}_{s_t, a_t, s_{t+1} \sim \mathcal{D}} \left[\left(r_t + \gamma \hat{Q}_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_{t+1} | s_{t+1} \right), a_{t+1} \right) - Q_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_t | s_t \right), a_t \right) \right)^2 \right] \right.$$
$$\left. + \tilde{\alpha} \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}} \left[\log \sum_{a} \exp \left(Q_i(s_t, a) \right) - \mathop{\mathbb{E}}_{a \sim \pi(s_t)} \left[Q_i(s_t, a) \right] \right] \right)$$

$$\tilde{s}_t = D\Big(\underbrace{(\mathbb{1} + \epsilon \sigma_{a_t})}_{\text{Symmetry shift}} \hat{*} \underbrace{E(s_t)}_{\text{Koopman space embedding}} \Big)$$

inverse embedding

Normally distributed random variable

Policy Optimization step Q-learning (CQL)

Kumar et. al

Bellman error

$$\hat{Q}_{i+1} \leftarrow \arg\min_{Q} \left(\mathbb{E}_{s_t, a_t, s_{t+1} \sim \mathcal{D}} \left[\left(r_t + \gamma \hat{Q}_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_{t+1} | s_{t+1} \right), a_{t+1} \right) - Q_i \left(\sigma^{\epsilon}_{a_t} \left(\tilde{s}_t | s_t \right), a_t \right) \right)^2 \right] + \tilde{\alpha} \, \mathbb{E}_{s_t \sim \mathcal{D}} \left[\log \sum_a \exp \left(Q_i(s_t, a) \right) - \mathbb{E}_{a \sim \pi(s_t)} \left[Q_i(s_t, a) \right] \right] \right)$$

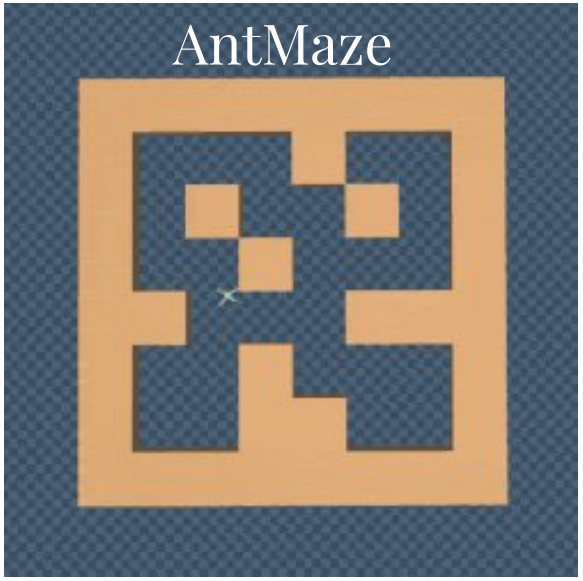
Koopman Forward Conservative Q-learning (KFC)

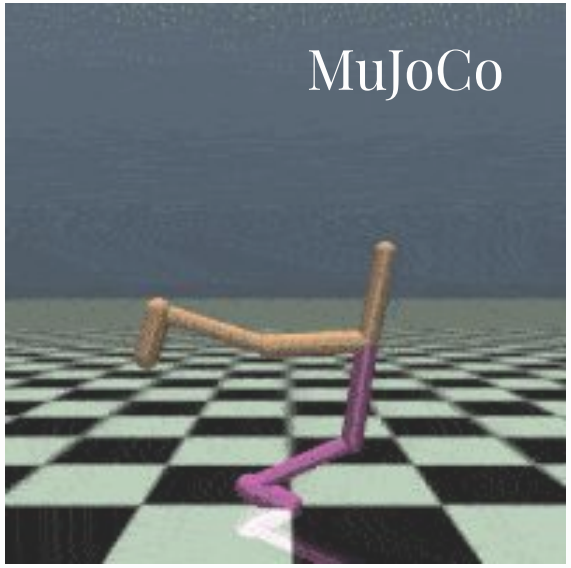
$$\tilde{s}_t = D\Big(\underbrace{(\mathbb{1} + \epsilon \, \sigma_{a_t})}_{\text{Symmetry shift}} \hat{*} \underbrace{E(s_t)}_{\text{Koopman space embedding}} \Big)$$

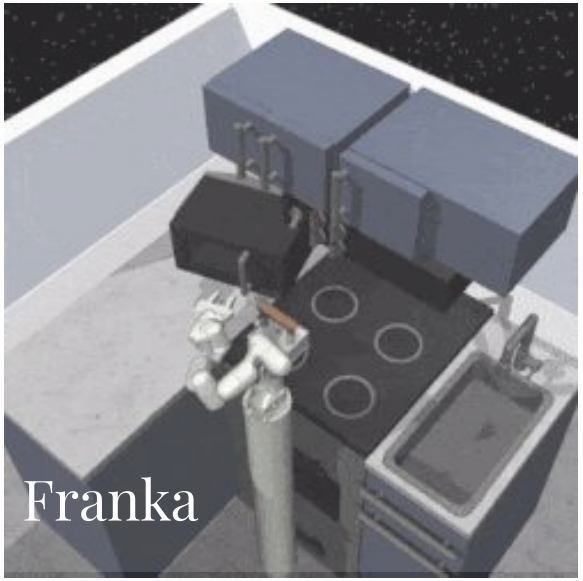
inverse embedding

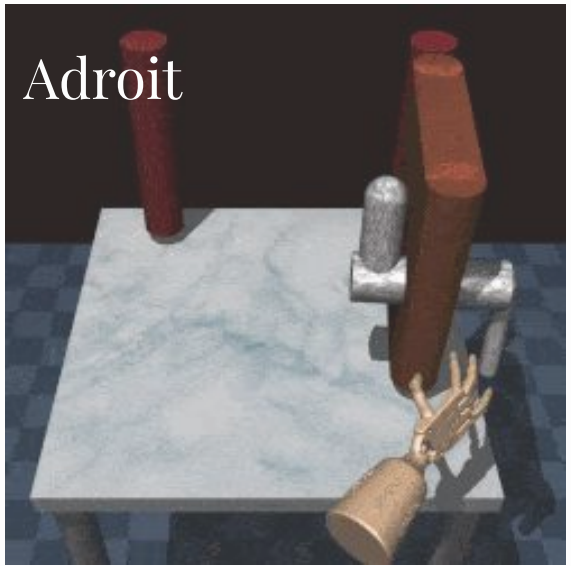
Normally distributed random variable

Empirical Tests - D4RL







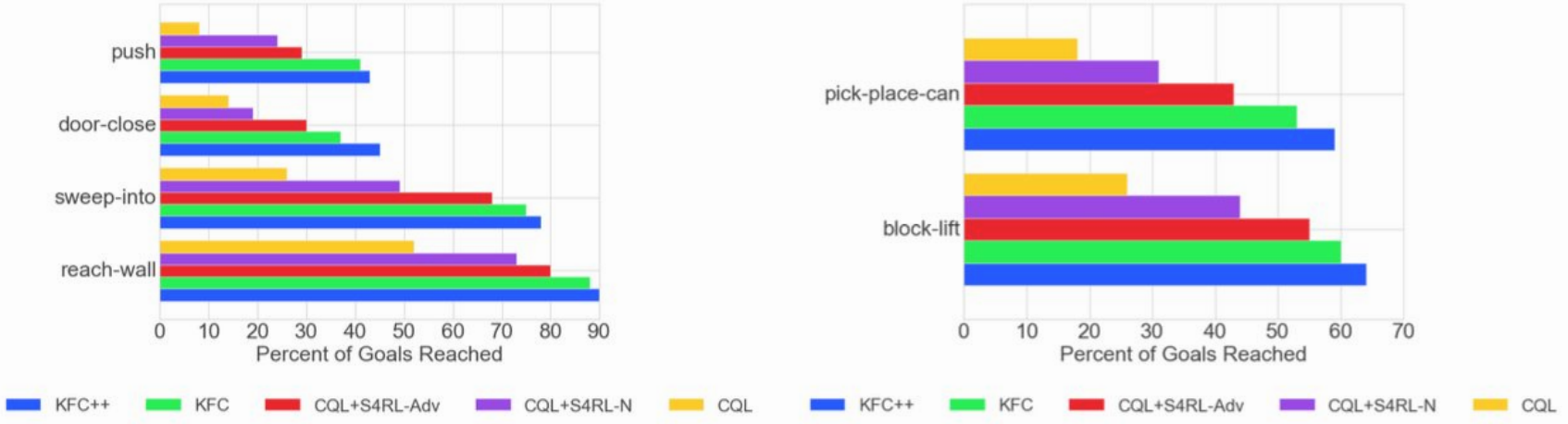


Empirical Results of KFC / KFC++

Task Name	CQL	S4RL	KFC	KFC++
Gym	41.6	61.6	66.4	**67.8**
Antmaze	50.6	62.9	65.5	**67.4**
Adroit	11.7	19.4	23.8	**24.3**
Franka	46.8	85.9	93.2	**95.4**

Empirical Results of KFC / KFC++

Task Name	CQL	S4RL	KFC	KFC++
Gym	41.6	61.6	66.4	**67.8**
Antmaze	50.6	62.9	65.5	**67.4**
Adroit	11.7	19.4	23.8	**24.3**
Franka	46.8	85.9	93.2	**95.4**



(a) MetaWorld Environments

(b) RoboSuite Environments