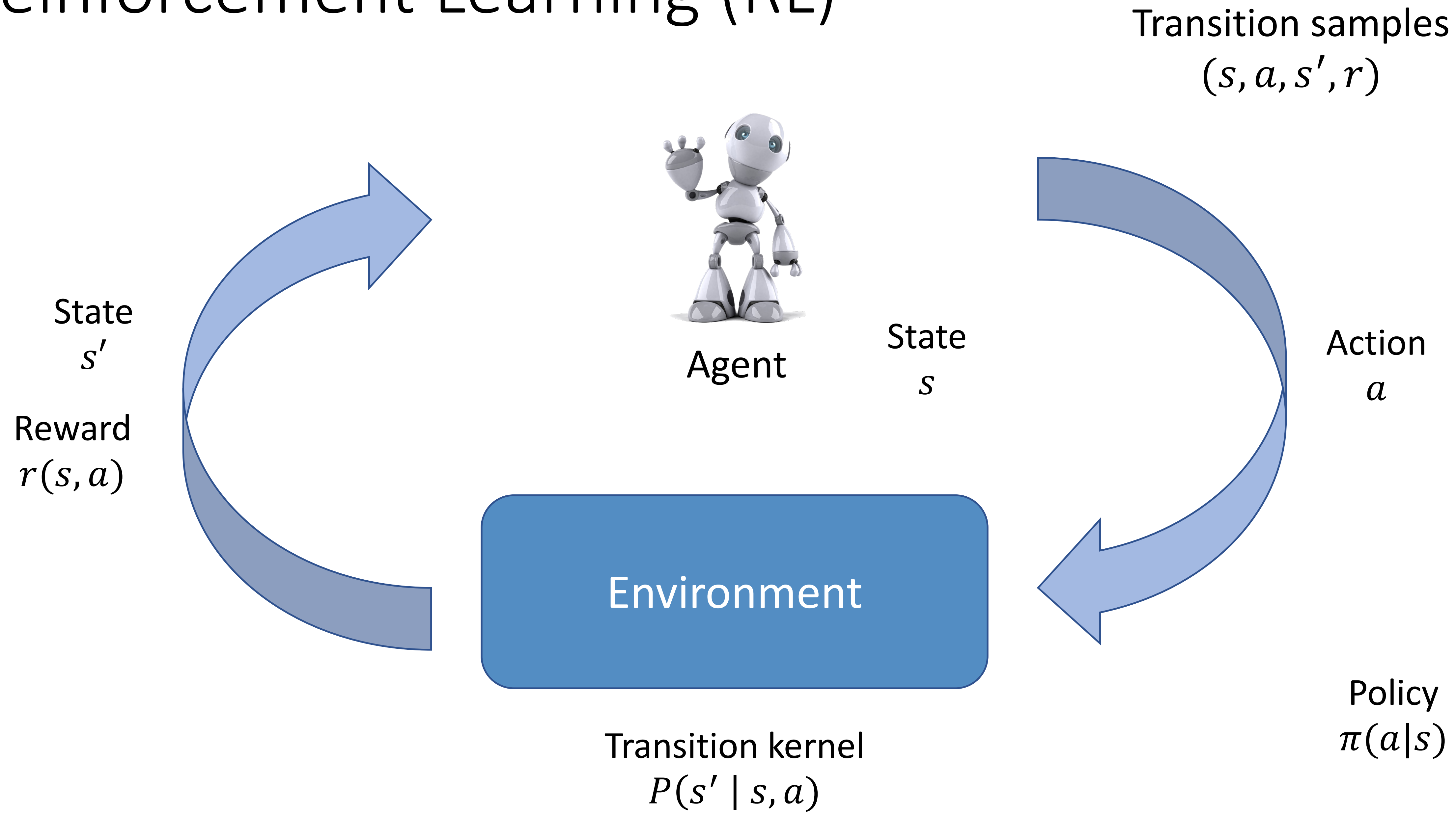
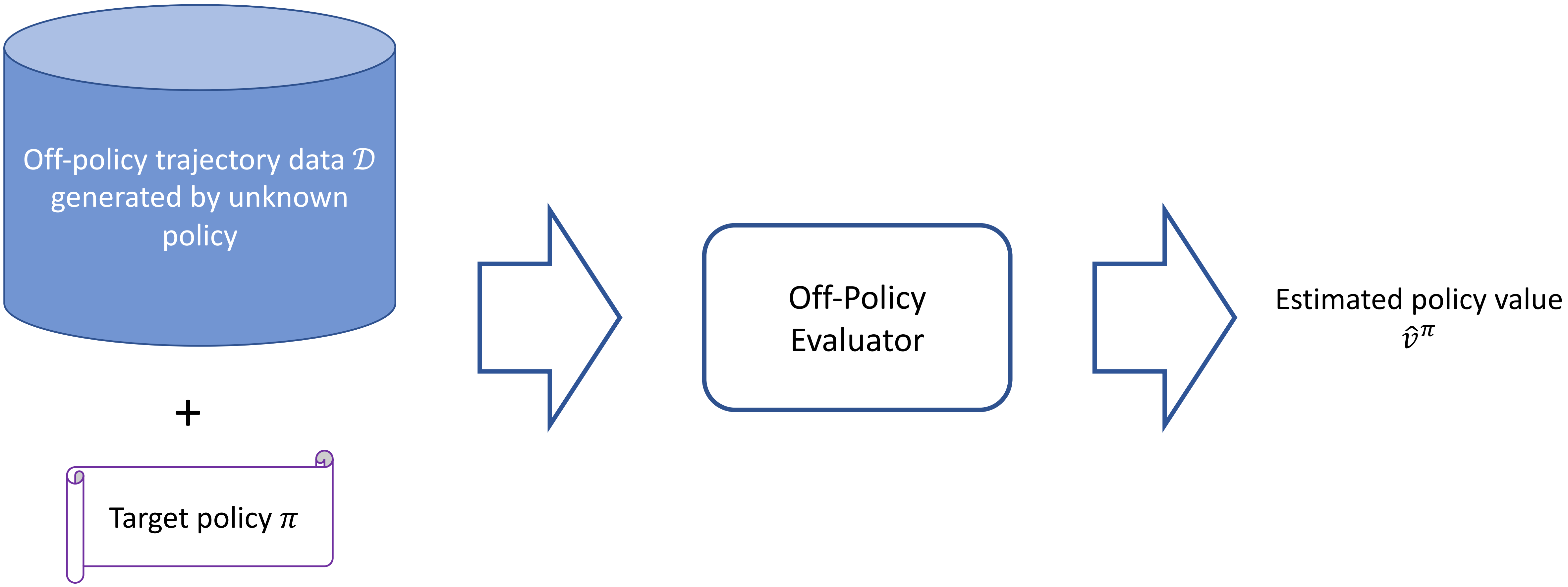
Sample Complexity of Nonparametric Off-Policy Evaluation on Low-Dimensional Manifolds Using Deep Networks

Xiang Ji, Minshuo Chen, Mengdi Wang, Tuo Zhao

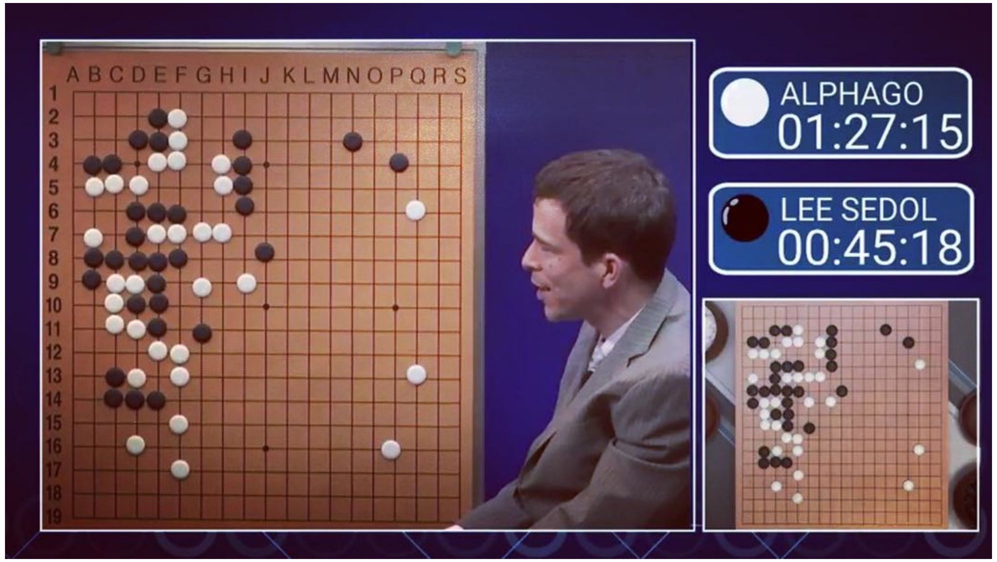
Reinforcement Learning (RL)



Off-Policy RL: Off-Policy Evaluation (OPE)



Deep RL



Neural network's success in supervised learning:

- Great universal function approximator
- Low sample complexity even with high data dimension

$$\text{Generalization error} \sim \sqrt{\frac{\mathrm{Comp}(\mathcal{F})}{n}} = \sqrt{\frac{f(D)}{n}}$$

Q1: Can these nice properties in supervised learning be preserved in RL?

Q2: What does it take? Are standard RL assumptions enough?

Off-Policy Evaluation (OPE)

Episodic MDP: state space S, action space A, horizon H
time-inhomogeneous transition kernel $\{P_h\}_{h=1}^{H}$
time-inhomogeneous reward $\{r_h\}_{h=1}^{H}$

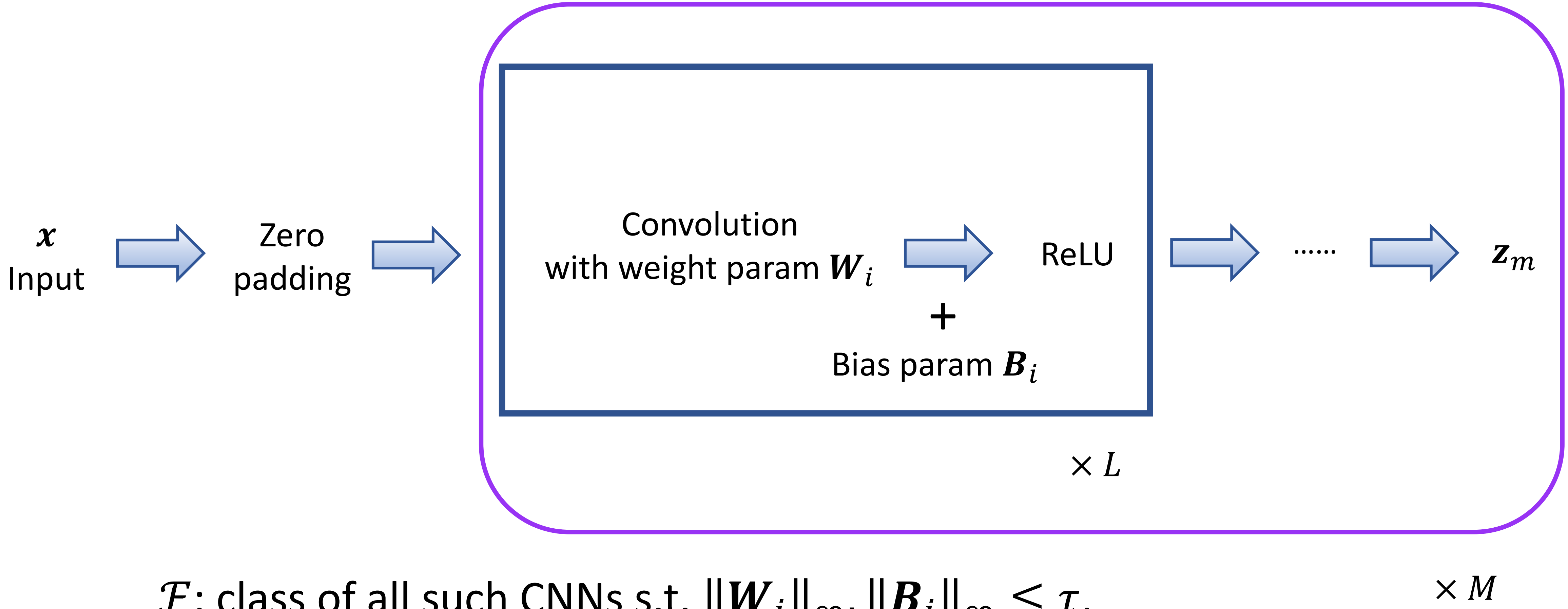
Off-policy dataset: $\mathcal{D} = \left\{\left\{\left(s_{h,k}, a_{h,k}, s'_{h,k}, r_{h,k}\right)\right\}_{h=1}^{H}\right\}_{k=1}^{K}$

For every step h, $\left\{s_{h,k}\right\}_{k=1}^{K}$ are i.i.d.
$\left\{a_{h,k}\right\}_{k=1}^{K}$ are generated from unknown behavior policy π_0.

Goal: Given off-policy dataset $\mathcal{D}$, **estimate** the value of target policy π from a fixed initial state distribution ξ, given by

$$v^{\pi} := \mathbb{E}^{\pi}\left[\sum_{h=1}^{H} r_h(s_h, a_h) \mid s_1 \sim \xi\right]$$

Convolutional Neural Networks (CNN)



$\mathcal{F}$: class of all such CNNs s.t. $\|\boldsymbol{W}_i\|_\infty, \|\boldsymbol{B}_i\|_\infty \leq \tau$,
filter size bounded by I,
number of channels bounded by J,
$\|f\|_\infty \leq V$

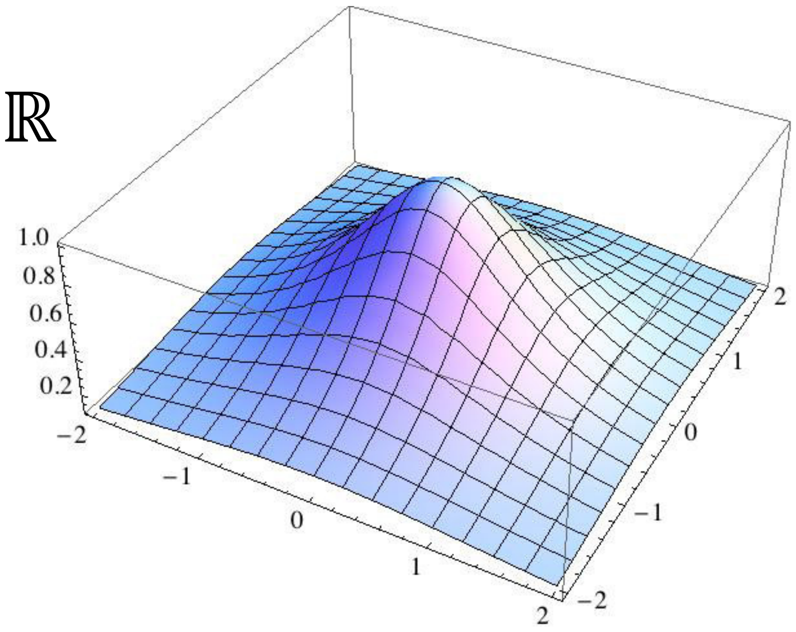
Learning on a low-dimensional manifold

- Data represented in $\mathbb{R}^D$, actually on a d-dim manifold ($d \ll D$)

  RL: State-action space $\mathcal{X} \coloneqq \mathcal{S} \times \mathcal{A}$

- Ground truth: Q-functions $\{Q_h^\pi\}_{h=1}^H$

  Nonparametric Besov functions from $\mathcal{X}$ to $\mathbb{R}$



Algorithm

Neural Fitted Q-Evaluation (Neural FQE)

For h = H, … , 1:

Sample $\mathcal{D}_h = \{(s_{h,k}, a_{h,k}, s'_{h,k}, r_{h,k})\}_{k=1}^{K}$

$$\hat{Q}_h^\pi \leftarrow \mathrm{argmin}_{f\in\mathcal{F}} \sum_{k=1}^{K} \left(f(s_{h,k}, a_{h,k}) - r_{h,k} - \int_{\mathcal{A}} \hat{Q}_{h+1}^\pi(s'_{h,k}, a)\pi_h(a|s'_{h,k})\mathrm{d}a \right)^2$$

$$\hat{v}^\pi \coloneqq \int_{\mathcal{S}\times\mathcal{A}} \hat{Q}_1^\pi(s,a)\,\pi(a|s)\xi(s)\mathrm{d}(s,a)$$

Assumptions

__Assumption 1__ $\mathcal{X} := \mathcal{S} \times \mathcal{A}$ is d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$. $\forall x \in \mathcal{X}, \|x\|_\infty \le B$.

__Assumption 2__ (Bellman completeness) Under target policy π, $\forall h$, $\forall f \in \mathcal{F}$, $\mathcal{T}_h^\pi f \in \mathcal{B}_{p,q}^\alpha(\mathcal{X})$ and there exists c_0 s.t. $\|\mathcal{T}_h^\pi f\|_{\mathcal{B}_{p,q}^\alpha(\mathcal{X})} \le c_0$.

[Ruosong Wang et al. 20]: Even with linear realizability & good data coverage,

$\exists$ an MDP s.t. all algorithms need $\Omega\left((D/2)^H\right)$ samples for OPE up to constant error w.p. 0.9.

Easily satisfied if MDP has "smooth" dynamics.

[Ruosong Wang et al. 20] Wang, Ruosong, Dean P. Foster, and Sham M. Kakade. "What are the Statistical Limits of Offline RL with Linear Function Approximation?." *arXiv preprint arXiv:2010.11895* (2020).

Results

Main theorem:

Under Assumption 1 and 2, let $\mathcal{F}$ be the class of CNNs with magnitude $\mathrm{V} = H$, number of channels $\mathrm{J} = O(D)$, $\mathrm{M} = O\left(K^{\frac{d}{2\alpha+d}}\right)$ layers each with $L = O(\log K + D)$ filters. Neural FQE achieves

$$\mathbb{E}|\hat{v}^{\pi} - v^{\pi}| \leq CH^2 \kappa K^{-\frac{\alpha}{2\alpha+d}} \log^{2.5} K\,, \quad (1)$$

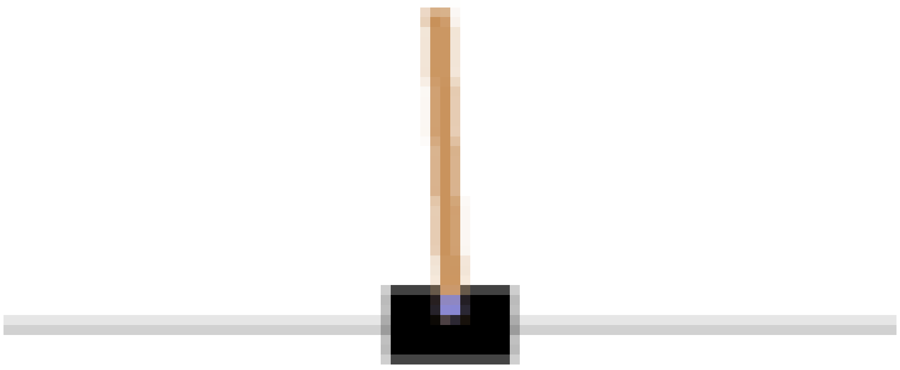
where

$$\kappa := \frac{1}{H} \sum_{h=1}^{H} \sqrt{\sup_{f \in \mathcal{G}} \frac{\mathbb{E}_{q_h^{\pi}}[f(x)]^2}{\mathbb{E}_{q_h^{\pi_0}}[f(x)^2]}}$$

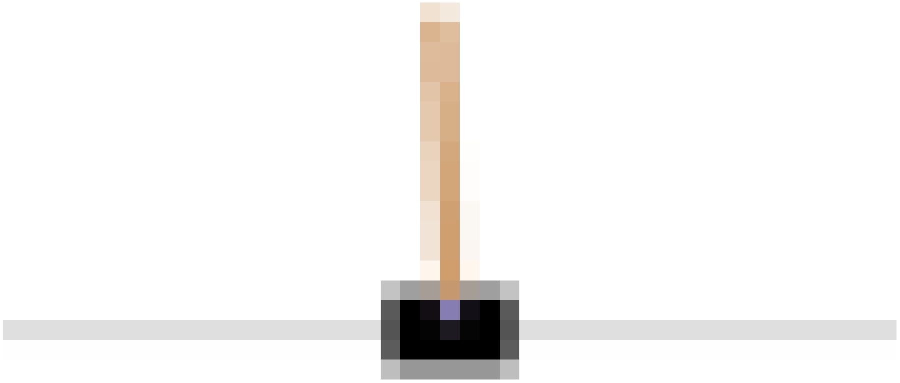
q_h^{π} is the state-action occupancy measure of target policy π at step h;
$q_h^{\pi_0}$ is the sampling distribution for step h.
$\mathcal{G}$ is the Minkowski sum of $\mathcal{B}_{p,q}^{\alpha}(\mathcal{X})$ and $\mathcal{F}$, i.e. $\mathcal{G} = \left\{g_1 + g_2 : g_1 \in \mathcal{B}_{p,q}^{\alpha}(\mathcal{X}), g_2 \in \mathcal{F}\right\}$.

- Estimation error depends mostly on d
- Better distributional mismatch characterization than absolute continuity $\left\|\frac{q_h^{\pi}}{q_h^{\pi_0}}\right\|_{\infty}$.

Experiments



CartPole in high resolution



CartPole in low resolution

Ground truth: 65.2

(A) No distribution shift		(B) Off-policy	
High res	Low res	High res	Low res
64.6 ± 2.0	63.5 ± 1.9	60.4 ± 2.8	60.0 ± 3.3
66.0 ± 1.3	66.5 ± 1.7	67.0 ± 1.8	68.0 ± 2.3
65.1 ± 1.0	65.1 ± 1.2	65.0 ± 1.6	65.1 ± 2.0

Neural FQE performs similarly on both datasets $\Rightarrow$ estimation error mostly independent from data representation dimension

Thank you!