Neural Network Approximation for Pessimistic Offline Reinforcement Learning

Di Wu
Joint work with Yuling Jiao, Xiliang Lu, Li Shen, Haizhao Yang

March Seminar, School of Mathematics and Statistics, Wuhan University
2024.3

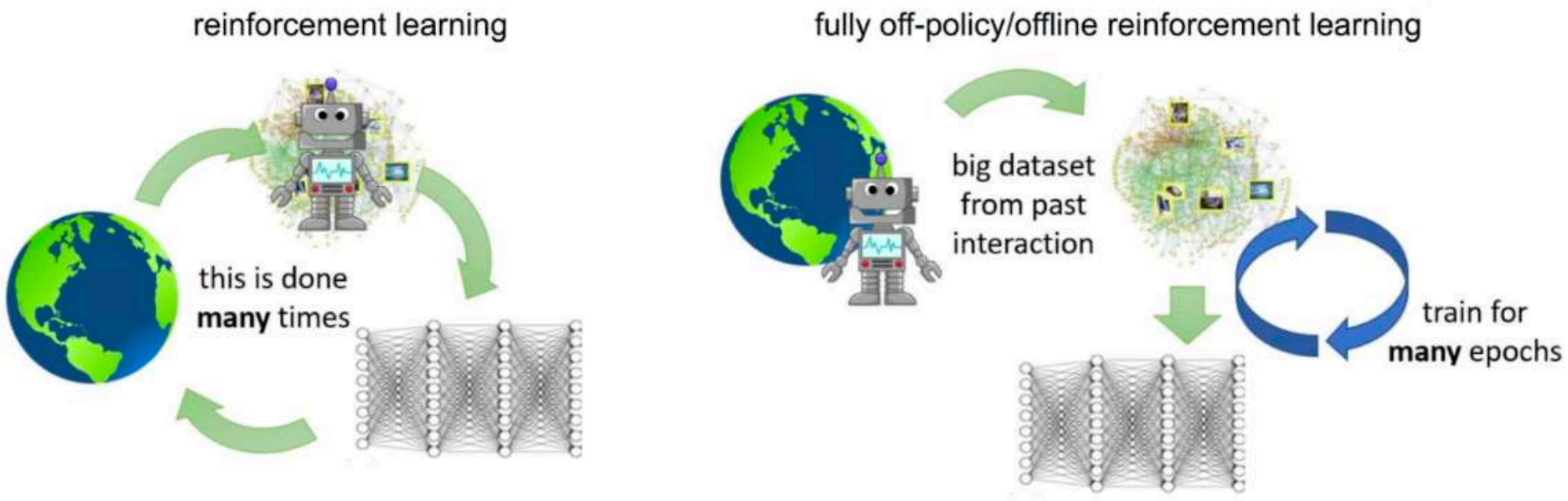
Offline RL?

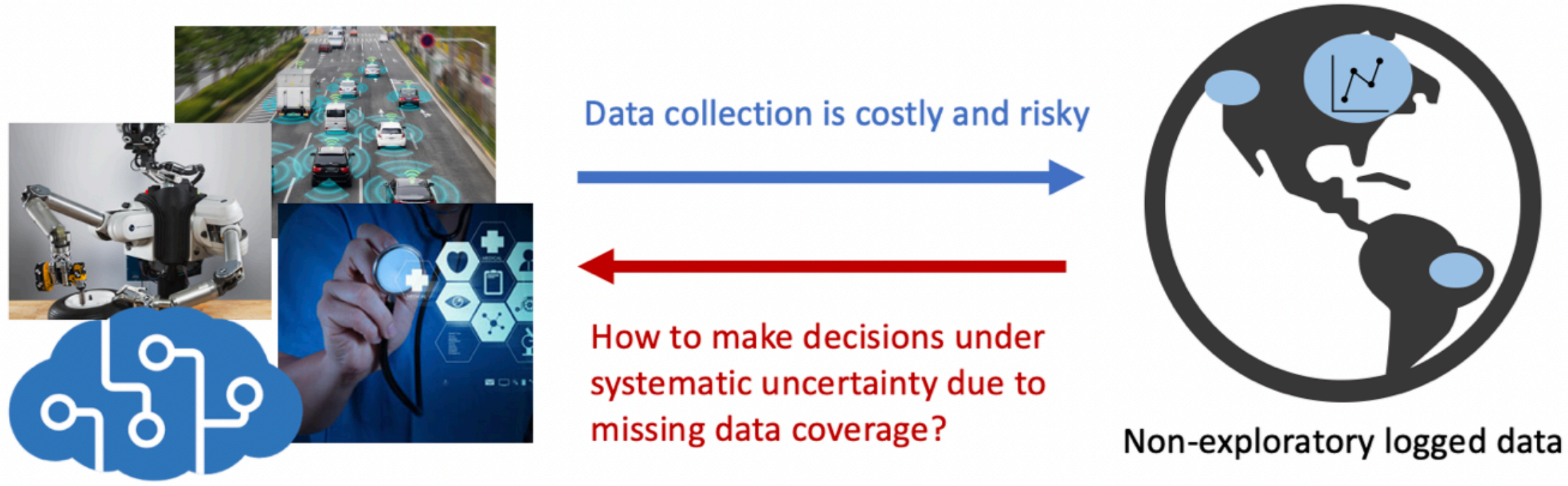
Characteristics and Challenge

- What?
  Learn completely from a previously collected dataset.

- Why?
  Safety (Healthcare) & Cost (Robot)

- Hard?
  Yes!
  Data distribution shift &
  Out-of-distribution values





Offline RL Methods

- To Alleviate A Main Challenge:
- DISTRIBUTION SHIFT

$$\boxed{\begin{aligned}
&\zeta_k(s,a) = |Q_k(s,a) - Q^*(s,a)| \\
&\delta_k(s,a) = |Q_k(s,a) - \mathcal{T}Q_{k-1}(s,a)| \\
&\zeta_k(s,a) \leq \delta_k(s,a) + \gamma \max_{a'} \mathbb{E}_{s'}[\zeta_{k-1}(s',a')]
\end{aligned}}$$

Kumar et al. '19 "Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction"

- The second term could be high on out-of-distribution (s, a) pairs since never directly minimized while training.

- **Policy-Based Regularization**
  (Kumar et al. '19 "Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction")
- Drawbacks:
  (1) may be overly conservative, similar to behavior cloning
  (2) estimating the behavior policy can be challenging
- **Pessimistic Value-Based Methods**
  (Kumar et al. '20 "Conservative Q-Learning for Offline Reinforcement Learning")
- Benefits:
  (1) Learning optimality
  (2) Robust Policy Improvement

Related Works

Theoretical Insight

- Early works:
  assume data to be fully covered,
  which is unrealistic
- More recent studies:
  partial coverage, tabular and linear
  function approximations
- General function approximations:
  finiteness and convexity

Existing Works	Assumption	Coverage
Szepesvári and Munos (2005); Munos (2007) Antos, Szepesvári, and Munos (2007, 2008) Farahmand, Szepesvári, and Munos (2010) Scherrer (2014); Liu et al. (2019); Chen and Jiang (2019) Jiang (2019); Wang et al. (2019); Feng, Li, and Liu (2019) Liao et al. (2022); Zhang et al. (2020) Uehara, Huang, and Jiang (2020); Xie and Jiang (2021)	\	Full
Nguyen-Tang et al. (2022a)	Neural network	
Rashidinejad et al. (2021); Yin, Bai, and Wang (2021) Shi et al. (2022); Li et al. (2022)	Tabular MDP	Partial
Jin, Yang, and Wang (2021); Chang et al. (2021) Zhang et al. (2022); Nguyen-Tang et al. (2022b); Bai et al. (2022)	Linear MDP	
Jiang and Huang (2020)	Compact	
Zhan et al. (2022)	Strongly convex	
Uehara, Huang, and Jiang (2020); Rashidinejad et al. (2022) Zanette and Wainwright (2022); Xie et al. (2021); Cheng et al. (2022)	Finite	
Ji et al. (2023) Our work	Neural network	

Table 1: A comparison of existing works concerning assumptions related to data coverage, and approximation.

Formal Formulation

Population Level

$$\widehat{\pi}^* \in \arg\max_{\pi\in\Pi} \min_{f\in\mathcal{F}_\mu^{\pi,\epsilon}} \mathcal{L}_\mu(\pi, f),$$

$$\begin{aligned}\mathcal{L}_\mu(\pi, f) &:= \mathbb{E}_\mu[f(s,\pi) - f(s,a)]\\ \mathcal{F}_\mu^{\pi,\epsilon} &:= \{f \in \mathcal{F} \mid \mathcal{E}_\mu(\pi, f) \le \epsilon\}\\ \mathcal{E}_\mu(\pi, f) &:= \|f - \mathcal{T}^\pi f\|_{2,\mu}^2\end{aligned}$$

Empirical Level

$$\widehat{\pi} = \arg\max_{\pi\in\Pi_\theta} \min_{f\in\mathcal{N}\mathcal{N}_2\cap\mathcal{F}_\mathcal{D}^{\pi,\epsilon}} \mathcal{L}_\mathcal{D}(\pi, f)$$

$$\begin{aligned}\mathcal{E}_D(\pi, f) :=&\mathbb{E}_\mathcal{D}\left[(f(s,a) - r - \gamma f(s',\pi))^2\right]\\ &- \min_{f'\in\mathcal{F}} \mathbb{E}_\mathcal{D}\left[(f'(s,a) - r - \gamma f(s',\pi))\right],\end{aligned}$$

$$\begin{aligned}\mathcal{L}_\mathcal{D}(\pi, f) &:= \mathbb{E}_\mathcal{D}[f(s,\pi) - f(s,a)]\\ \mathcal{F}_\mathcal{D}^{\pi,\epsilon} &:= \{f \in \mathcal{F} \mid \mathcal{E}_\mathcal{D}(\pi, f) \le \epsilon\}\end{aligned}$$

Function Class

$$\begin{aligned}&f_0(x) = x,\\ &f_\ell(x) = \sigma\left(W_\ell f_{\ell-1}(x) + b_\ell\right), \quad \ell = 1, \ldots, L-1,\\ &f(x) = f_L(x) = W_L f_{L-1}(x) + b_L.\end{aligned}$$

$$\mathcal{H}^\zeta = \left\{f : [0,1]^d \to \mathbb{R} \middle| \max_{\|\alpha\|_1\le s} \|\partial^\alpha f\|_\infty \le B, \right.\\ \left. \max_{\|\alpha\|_1 = s} \sup_{x\ne y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x-y\|_\infty^r} \le B\right\}.$$

Main Result

Loss Consistency

Hölder Continuous $\qquad$ $\forall \pi \in \Pi_\theta, f \in \mathcal{NN}_2$, *we have* $\mathcal{T}^\pi f \in \mathcal{F}$. $\qquad$ *C-mixing*

Theorem 4.1 *Under Assumptions about smoothness, completeness and mixing data, for* $\mathcal{NN}_1$, $\mathcal{NN}_2$ *with width* $\mathcal{W} = \mathcal{O}(d^{s+1}|\mathcal{D}|^{\frac{d}{2d+4\zeta^*}})$ *and depth* $\mathcal{L} = \mathcal{O}(\log(|\mathcal{D}|))$, *the following non-asymptotic error bound holds*

$$\mathbb{E}[\widetilde{\mathcal{R}}_\mu(\widehat{\pi}, \epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*, \epsilon)] \leq C_1 R_{\max} d^{s+(\zeta \vee 1)/2} |\mathcal{D}|^{\frac{-\zeta^*}{d+2\zeta^*}} \log(|\mathcal{D}|)^{2+\frac{1}{\eta}} + C_2\sqrt{\epsilon},$$

where $\zeta^* = \zeta(1 \wedge \zeta)$, C_1 *is a constant depending on* $s, B, \mathcal{C}(\widehat{\pi}; \mu), \mathcal{C}(\widehat{\pi}^*_\delta; \mu)$ *and* C_2 *is a constant depending on* $\mathcal{C}(\widehat{\pi}; \mu), \mathcal{C}(\widehat{\pi}^*_\delta; \mu)$.

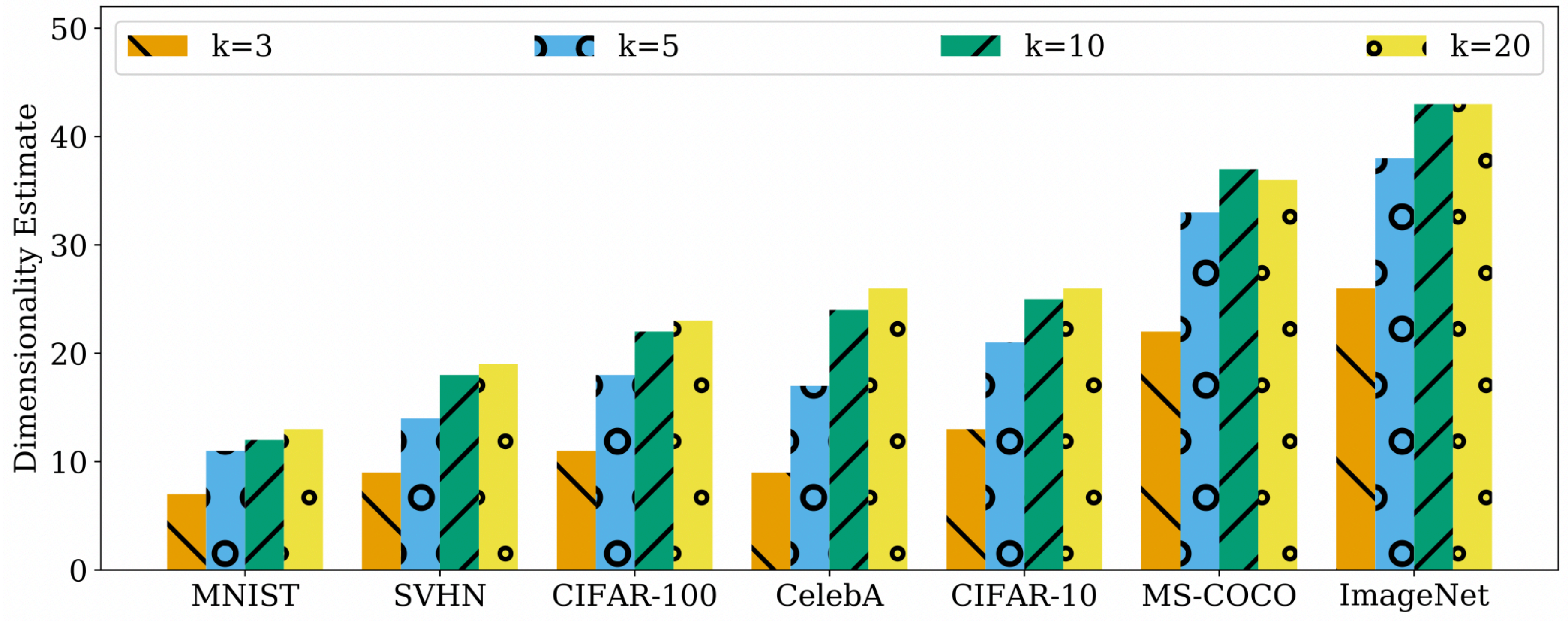
- Effective Controllability
- Optimal Rate
- Bellman Constraint Hyperparameter

- Curse of Dimensionality!

Curse of Dimensionality?

Two Scenarios

- Low-dimensional mainifold assumption



Credit: Phillip Pope et al. (2021).

e.g., MNIST, CIFAR, ImageNet

- Low-complexity assumption

$$f = G^k \circ G^{k-1} \circ \cdots \circ G^1,$$

where $G^i : \mathbb{R}^{l_{i-1}} \to \mathbb{R}^{l_i}$ is defined by $G^i(x) = [g_1^i(W_1^i x), \ldots, g_{l_i}^i(W_{l_i}^i x)]^\top$, with $W_j^i \in \mathbb{R}^{d_i \times l_{i-1}}$ being a matrix and $g_j^i : \mathbb{R}^{d_i} \to \mathbb{R}$ being a function.

e.g., additive model, single index model, projection pursuit model, PDEs

Improved Results

with same assumptions

Low Dimensionality

$$\mathbb{E}[\widetilde{\mathcal{R}}_\mu(\widehat{\pi},\epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*,\epsilon)]$$
$$\leq \frac{C_1 R_{\max}}{(1-\lambda)^{\zeta/2}}\sqrt{d}d_K^{s+(\zeta\vee 1+1)/2}|\mathcal{D}|^{\frac{-\zeta^*}{d_K+2\zeta^*}}\log(|\mathcal{D}|)^{2+\frac{1}{\eta}} + C_2\sqrt{\epsilon},$$

where $0<\lambda<1$, $d_K = \mathcal{O}(\dim_{\mathcal{M}}(K)/\lambda^2)$, $\zeta^* = \zeta(1\wedge\zeta)$, C_1 *is a constant depending on* $s, B, \mathcal{C}(\widehat{\pi};\mu), \mathcal{C}(\widehat{\pi}^*_\delta;\mu)$ *and* C_2 *is a constant depending on* $\mathcal{C}(\widehat{\pi};\mu), \mathcal{C}(\widehat{\pi}^*_\delta;\mu)$.

Low Complexity

$$\mathbb{E}[\widetilde{\mathcal{R}}_\mu(\widehat{\pi},\epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*,\epsilon)]$$
$$\leq C_1 R_{\max} d_*^{s+(\zeta\vee 1)/2}|\mathcal{D}|^{\frac{-\zeta^*}{d_*+2\zeta^*}}\log(|\mathcal{D}|)^{2+\frac{1}{\eta}} + C_2\sqrt{\epsilon},$$

where $\zeta^* = \min_i(\zeta_i\prod_{l=i+1}^{k}(\zeta^l\wedge 1))(1\wedge\zeta)$, $d_* = \max_i d_i$, C_1 *is a constant depending on* $B, \zeta, s, k, \mathcal{C}(\widehat{\pi};\mu), \mathcal{C}(\widehat{\pi}^*_\delta;\mu)$ *and* C_2 *is a constant depending on* $\mathcal{C}(\widehat{\pi};\mu), \mathcal{C}(\widehat{\pi}^*_\delta;\mu)$.

- Substitute d with low Minkowski dimension
- Substitute d with the minimum of each component
- Significant Improvement!

Proof Sketch

Oracle Inequality

$$\widetilde{\mathcal{R}}_\mu(\widehat{\pi},\epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*,\epsilon) \leq \underbrace{2\sup_{\phi\in\Pi_\theta}|\widetilde{\mathcal{R}}_{\mathcal{D}}(\phi,\epsilon) - \widetilde{\mathcal{R}}_\mu(\phi,\epsilon)|}_{(A)} + \underbrace{2\sup_{\phi\in\Pi_\theta}|\widehat{\mathcal{R}}_\mu(\phi,\epsilon) - \widetilde{\mathcal{R}}_{\mathcal{D}}(\phi,\epsilon)|}_{(B)}$$

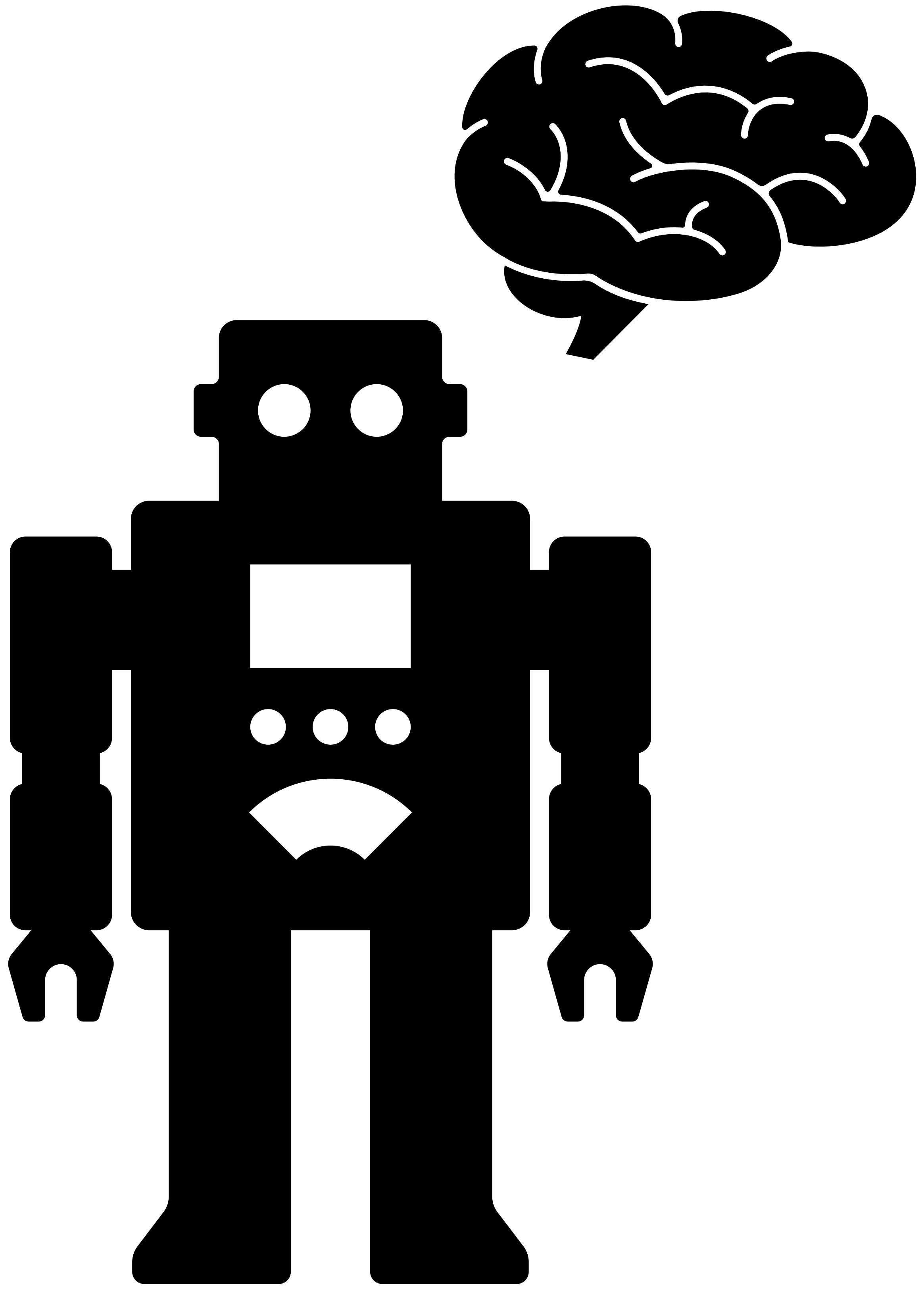
$$+ \underbrace{\inf_{\phi\in\Pi_\theta}\Big(\Big(\widehat{\mathcal{R}}_\mu(\widehat{\pi},\epsilon) - \widehat{\mathcal{R}}_{\mathcal{D}}(\widehat{\pi},\epsilon)\Big) + \Big(\widehat{\mathcal{R}}_{\mathcal{D}}(\phi,\epsilon) - \widehat{\mathcal{R}}_\mu(\phi,\epsilon)\Big) + \Big(\widetilde{\mathcal{R}}_\mu(\phi,\epsilon) - \widetilde{\mathcal{R}}_\mu(\widehat{\pi}^*,\epsilon)\Big)\Big)}_{(C)}$$

$$\widetilde{\mathcal{R}}_{\mathcal{D}}(\pi,\epsilon) = \max{}_{f\in\mathcal{NN}_2\cap\mathcal{F}_\mu^{\pi,\epsilon}}\mathcal{R}_\mu(\pi,f), \widehat{\mathcal{R}}_\mu(\pi,\epsilon) = \max{}_{f\in\mathcal{NN}_2\cap\mathcal{F}_\mu^{\pi,\epsilon}}\mathcal{R}_{\mathcal{D}}(\pi,f), \widehat{\mathcal{R}}_{\mathcal{D}}(\pi,\epsilon) = \max{}_{f\in\mathcal{NN}_2\cap\mathcal{F}_{\mathcal{D}}^{\pi,\epsilon}}\mathcal{R}_{\mathcal{D}}(\pi,f).$$

- (A): Approximation Error
- (B): Generalization Error
- (C): Bellman Estimation Error

Summary

- Deep Adversarial Offline RL Framework
  deep neural networks, sequential data, partial coverage
- Non-Asymptotic Rate for the Estimation Error
  mild assumptions, explicit illustration
- Mitigate the Curse of Dimensionality
  low-dimensional data structures or low-complexity target functions



Thanks