# Sample Complexity of Kernel-Based Q-Learning

Sing-Yuan Yeh

Advised by Dr. Sattar Vakili (MRUK) & Fu-Chieh Chang (MRTW)

Mediatek Research & NTU

June 14, 2024

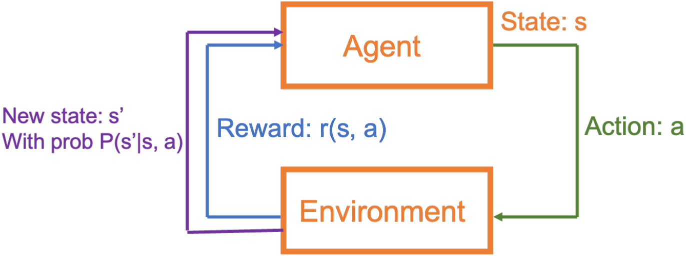
# Table of Contents

# Introduction (LR)

1. Sample complexity of spectral embedding: [Rudi & Canas, NeuralPS, (2014)]
    - Diffusion maps converge in $L^\infty$ [Dusun, Wu & Wu, ACHA, (2019)]
    - Vector diffusion maps converge in $L^2$ [Singer & Wu, Inf. Inference, (2015)]
2. Sample complexity of Q-learning
    - RL with finite $\mathcal{S} \times \mathcal{A}$ and linear feature [Jin et al., PMLR, (2020)]
    - Bandit with infinite $\mathcal{S} \times \mathcal{A}$ and kernel-, network-based [Yang et al., NeuralPS, (2020)]

# Outline

# Notation

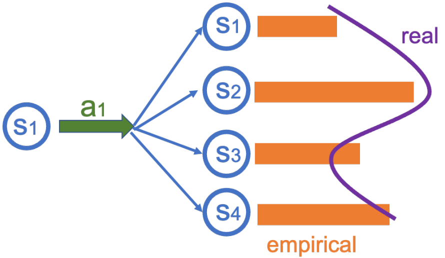



- A set of states, $\mathcal{S}$ and a set of actions, $\mathcal{A}$.
- $\mathcal{S}$ and $\mathcal{A}$ might be infinite, *i.e.* $S = |\mathcal{S}| \leq \infty$, $A = |\mathcal{A}| \leq \infty$.
- An agent will decide to play $a$ at $s$ by policy $\pi$.
- After playing $a$ at $s$, get reward $r(s, a) \in [0, 1]$.
- After playing $a$ at $s$, transition to $s'$ with probability $P(s'|s, a)$.

# Goal [Jin et al. (2020)]

## Goal

Given a model of the environment, how many transitions do we need to observe for finding an “near” optimal policy with high probability.




| | $s_1$ | $s_2$ | $\cdots$ | $s_S$ |
|---|---|---|---|---|
| $(s_1, a_1)$ | | | | |
| $(s_1, a_2)$ | | | | |
| $\vdots$ | | | | |
| $(s_S, a_A)$ | | | | |

## Question

However, if $S$ and $A$ is infinite?

# Value function

- Consider discounted Markov decision process with discounted factor $\gamma$.
- For a policy $\pi : \mathcal{S} \to \mathcal{A}$, the value function is defined as

$$v^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} r\left(s^{t}, \pi\left(s^{t}\right)\right) \mid s^{0} = s\right]$$

- A policy $\pi$ is said to be $\epsilon$-optimal if $\|v^{\pi} - v^{*}\|_{\infty} \leq \epsilon$. That is,

$$v^{\pi}(s) \geq v^{*}(s) - \epsilon, \quad \text{for all } s \in \mathcal{S}\,.$$

  where $\pi^{*}$ attains the maximal value.
- A Q-function of policy $\pi$ is defined by

$$Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v^{\pi}(s')\,.$$

- Select the actions based on proxy Q-function $\hat{Q}(s,a) = \hat{Q}(z)$.

# Least square method

Let's focus on $PV(\bar{s},\bar{a}) = \sum_{s'\in\mathcal{S}} P(s'|\bar{s},\bar{a})V(s')$. Consider the function class $\mathcal{F}$. Approximate $PV(\bar{s},\bar{a})$ by least square problem,

$$PV(\bar{s},\bar{a}) \leftarrow \min_{f\in\mathcal{F}} \left\{ \sum_{(\bar{s},\bar{a})\in\mathcal{U}} \left[\widehat{P}V(\bar{s},\bar{a}) - f(\bar{s},\bar{a})\right]^2 + \mathrm{pen}(f) \right\}.$$

where $\mathrm{pen}(f)$ is regularization term.

## Goal

- Pick representative set $\mathcal{U}$
- Compute the empirical transition probability

## Question

What is $\mathcal{F}$? In linear case, $\mathcal{F} = \{\phi(s,a)^\top w : \phi(\cdot,\cdot), w \in \mathbb{R}^D\}$ [Jin et al.].

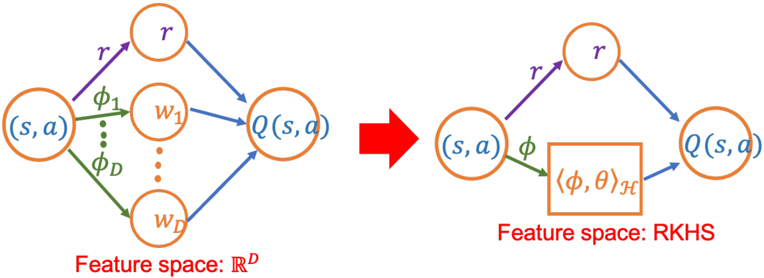
# Kernel ridge regression (KRR)

- Consider an unknown function $f \in \mathcal{H}_K$, and a set $\mathcal{U}_J = \{z_j\}_{j=1}^{J} \subset \mathcal{Z}$ of $J$ inputs.
- Assume $J$ noisy observations $\{Y(z_j) = f(z_j) + \epsilon_j\}_{j=1}^{J}$ are provided, where $\epsilon_j$ are i.i.d. zero mean sub-Gaussian noise terms.
- Define $Y_{\mathcal{U}_J} = [Y(z_1), \ldots, Y(z_J)]^\top \in \mathbb{R}^{J \times 1}, k_{\mathcal{U}_J}(z) = [K(z, z_1), \ldots, K(z, z_J)]^\top \in \mathbb{R}^{J \times 1}$ and $K_{\mathcal{U}J} = [K(z_i, z_j)]_{i,j=1}^{J} \in \mathbb{R}^{J \times J}$.

# Kernel ridge regression

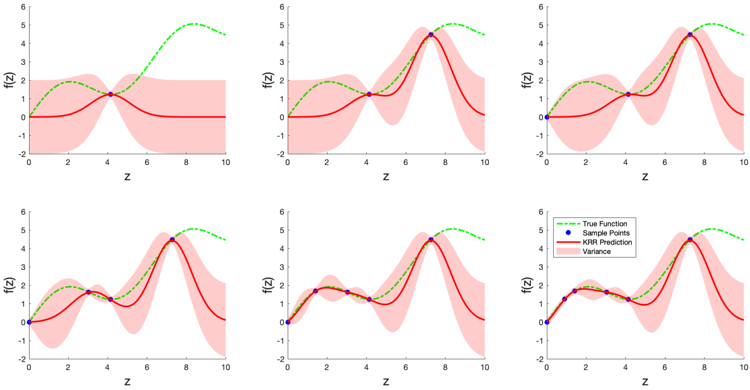
- Given $\lambda > 0$, kernel ridge regression provides the following regressor and uncertainty estimate, respectively.

$$\hat{f}_{U_j}(z) = k_{\mathcal{U}_j}^\top(z)\left(K_{\mathcal{U}_j} + \lambda^2 I\right)^{-1} Y_{\mathcal{U}_J},$$
$$\Sigma_{\mathcal{U}_j}^2(z) = K(z,z) - k_{\mathcal{U}_j}^\top(z)\left(K_{\mathcal{U}_j} + \lambda^2 I\right)^{-1} k_{\mathcal{U}_j}(z).$$



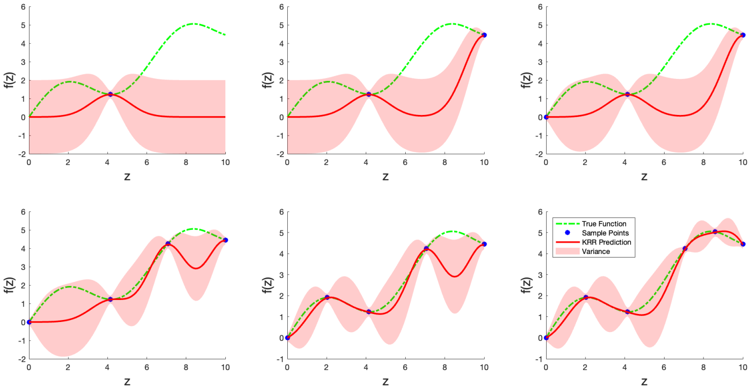
# Without maximal variance reduction (Random choose)

Pick representative set $\mathcal{U}$ randomly.



# With maximal variance reduction [Vakali et al. (2021)]

Pick $z \leftarrow \arg\max_{z \in \mathcal{Z}} \sum^2_{\mathcal{U}_{j-1}}(z)$ and collect $\mathcal{U}_j \leftarrow \mathcal{U}_{j-1} \cup \{z\}$.



# Confidence interval of $f(z)$

**Theorem in [Vakaki et al. (2021, 2022)]**

The noise are sub-Gaussian with parameter $R$ and $\|f\|_{\mathcal{H}_K} \leq C_K$. Then, the following each hold uniformly in $z \in \mathcal{Z}$, with probability $1-\delta$,

$$f(z) \geq \hat{f}_{\mathcal{U}_J}(z) - \beta(\delta)\Sigma_{\mathcal{U}_J}(z), \text{ and } f(z) \leq \hat{f}_{\mathcal{U}_J}(z) + \beta(\delta)\Sigma_{\mathcal{U}_j}(z)$$

where $\beta(\delta) = \mathcal{O}\left(C_K + \frac{R}{\lambda}\sqrt{d \log\left(\frac{JC_K}{\delta}\right)}\right)$

Next, we need to estimate $\Sigma_{\mathcal{U}_j}(z)$

# Maximal information gain

## Theorem in Srinivas et al. (2010)

For any set $\mathcal{U}_J \subset \mathcal{Z}$, we have

$$\sum_{j=1}^{J} \Sigma^2_{\mathcal{U}_{j-1}}\left(z_j\right) \leq \frac{2}{\log\left(1+1/\lambda^2\right)} \Gamma_{K,\lambda}(J).$$

where $\Gamma_{K,\lambda}(J)$ complexity term of a kernel $K$.

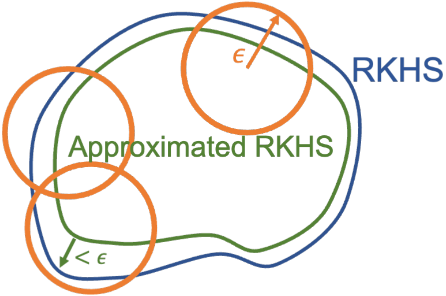
By Mercer Theorem, a kernel function $K$ can be represented by

$$K\left(z, z'\right) = \sum_{m=1}^{\infty} \sigma_m \psi_m(z) \psi_m\left(z'\right) .$$

- if $\sigma_m \leq C_p m^{-\beta_p}$, then $\Gamma_{K,\lambda}(J) = \mathcal{O}\left(J^{\frac{1}{\beta_p}} \log^{1-\frac{1}{\beta_p}}(J)\right)$.
- if $\sigma_m \leq C_{e,1} \exp\left(-C_{e,2} m^{\beta_e}\right)$, then $\Gamma_{K,\lambda}(J) = \mathcal{O}\left(\log^{1+\frac{1}{\beta_e}}(J)\right)$.

# Covering (Entropy bound)

The measure quantities of complexity of $\widehat{V}_w$ is number of $\epsilon$-covering of RKHS.




Hence, by Yang et al (2020), we can find the finite dimensional subspace to approximate infinite dimensional RKHS.

# Algorithm

1. Pick $(s_j, a_j) \leftarrow \arg\max_{(s,a)\in\mathcal{Z}} \; \Sigma^2_{\mathcal{U}_{j-1}}(s,a)$.
2. Collect $\mathcal{U}_j \leftarrow \mathcal{U}_{j-1} \cup \{(s_j, a_j)\}$.
3. Repeat Step 1 & 2 $J$ times.
4. Declare a vector $Y^{(\ell)}_{\mathcal{U}_J} = \mathbf{0}_J$.
5. Obtain a sample transition state $s' \sim P(\cdot|s_j, a_j)$.
6. Update $Y^{(\ell)}(s_j, a_j) \leftarrow$
   $\Pi_{[0,\frac{1}{1-\gamma}]} \max_{a\in\mathcal{A}} \left\{ r(s',a) + \gamma k^\top_{\mathcal{U}_J}(s',a) \left(K_{\mathcal{U}_J} + \lambda I_J\right)^{-1} Y^{(\ell-1)}_{\mathcal{U}_J} \right\}$.
7. Repeat Step 5, 6 $L$ times, i.e. draw $L$ sample from $(s_j, a_j)$.
8. Repeat Step 5, 6, 7 $J$ times, i.e. go over all $(s_j, a_j) \in \mathcal{U}$.
9. Output proxy $Q$ function
   $\widehat{Q}^{(L)}(\cdot) = r(\cdot) + \gamma k^\top_{\mathcal{U}_J}(\cdot) \left(K_{\mathcal{U}J} + \lambda^2 I_J\right)^{-1} Y^{(L)}_{\mathcal{U}_J}$.

# Estimation of value function

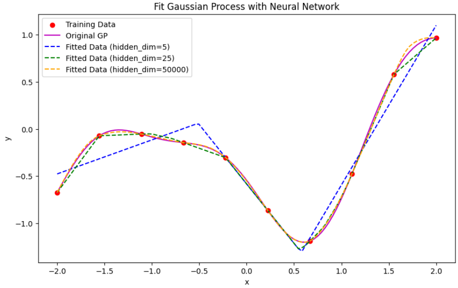
**Main Theorem (Yeh, Chang, Yeuh, Wu, Bernacchia & Vakali)**

With probability at least $1 - \delta$,

$$\|V^{\pi} - V^{*}\|_{\infty} \leq \underbrace{2\beta(\delta)\left(\frac{\gamma}{1-\gamma}\right)^{2}\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{J}}}_{\text{information gain}} + \underbrace{2\gamma^{L-1}\left(\frac{1}{1-\gamma}\right)^{2}}_{\text{Bellman operator}}$$

# Networks-based Q-learning

- Gaussian process has been pointed out as a shallow but infinitely wide neural network (NN) with Gaussian weights. [Neal (1996); Matthews et al. (2018); Lee et al. (2018)]
- The dynamic of training overparametrized NN process can be captured by the frame work of neural tangent kernel (NTK). [Jacot et al. (2018)]



# Outline

# Reference


[1] S. Vakili, N. Bouziani, J. Jalali, A. Bernacchia, and D.-S. Shiu, *Optimal order simple regret for gaussian process bandits*, NeuralPS, (2021).

[2] Z. Yang, C. Jin, Z. Wang, M. Wang and M. Jordan, *On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces*, NeuralPS, (2020).

Thank You for Your Attention!