FLATIRON INSTITUTE
Center for Computational Mathematics

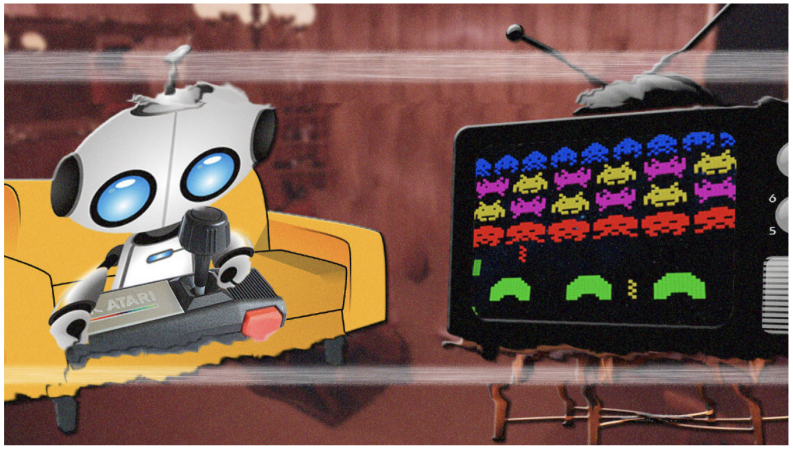
Perturbational Complexity by Distribution Mismatch: A Systematic Analysis of Reinforcement Learning in Reproducing Kernel Hilbert Space

Jiequn Han

Center for Computational Mathematics, Flatiron Institute

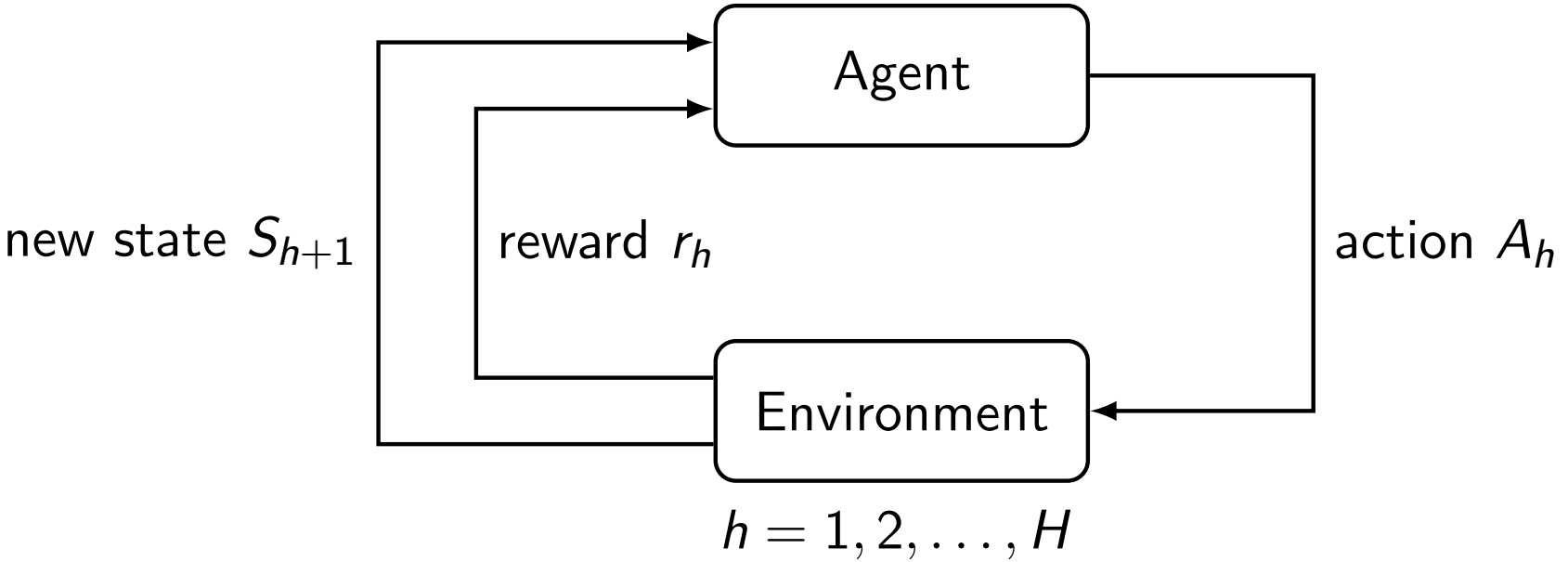
Joint work with Jihao Long, Princeton University

Phenomenal Success of Reinforcement Learning (RL)



Deep RL: powerful function approximation

Episodic Markov Decision Process (MDP)



Markov Decision Process (MDP): $(\mathcal{S}, \mathcal{A}, H, P, r, \mu)$

- $\mathcal{S}$: state space / $\mathcal{A}$: action space (both in Euclidean spaces)
- μ: initial state distribution, $S_1 \sim \mu$
- H: episode length
- $A_h \sim \pi_h(\cdot \mid S_h)$: policy (action selection rule)
- $S_{h+1} \sim P(\cdot \mid h, S_h, A_h)$: transition probability at step h (unknown)
- r_h: observed reward at step h, $r(h, S_h, A_h)$ (unknown)

Optimal Total Reward and Optimal Policy

Expected total reward

$$J(M,\pi) = \mathbb{E}_\pi \left[\sum_{h=1}^{H} r(h, S_h, A_h) \,|\, S_1 \sim \mu \right]$$
$$= \sum_{h=1}^{H} \int_{\mathcal{S}\times\mathcal{A}} r(h,s,a)\, \mathrm{d}\rho_{h,P,\pi,\mu}(s,a)$$

where $\rho_{h,P,\pi,\mu}$ denotes the distribution of (S_h, A_h) under initial distribution μ and policy π

Optimal policy π^* maximizes the total reward, and optimal total reward is

$$J^*(M) = \sup_\pi J(M,\pi) = J(M,\pi^*)$$

Goal: find near-optimal total reward/policy through finite interactions (S_h, A_h, r_h, S_{h+1}) with the environment

From Table to Function Approximation

Key components in RL algorithms: policy function $\pi_h(s, a)$, value function $V_h(s), Q_h(s, a)$

Tabular MDP: Both $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite.

Function approximation:

- Linear model.
- Kernel function: given a positive definite kernel k on $\mathcal{S} \times \mathcal{A}$, there exists a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ s.t. $\forall z \in \mathcal{S} \times \mathcal{A}$ and $f \in \mathcal{H}_k$, $f(z) = \langle f, k(z, \cdot) \rangle_k$

When can a reinforcement learning problem be solved efficiently using kernel function approximation?

Mainly focus on sample complexity, high dimensions.

Existing Works and Our Contributions

- Lower bound: quite few results beyond the tabular setting
  - [NYW19] proves an optimal lower bound for Lipschitz function approximation
  - [BKWY20] shows hard examples depending on the horizon exponentially in a linear setting
  - [CJ19] shows hard examples even when the set of candidate approximating functions is finite and includes the optimal Q-value function
  - Our results: provide lower bound for a general class of RL problems
- Upper bound: two types of assumptions for RKHS
  - [DMPKV20, YJWWJ20a, YJWWJ20b] assume fast eigenvalue decay of the kernel
  - [FMS10, LH21] assume finite concentration coefficients
  - Our results: provide upper bound valid under either of the two

Both results build on the same complexity measure

Learning in RKHS

Supervised Learning: for any target distribution f lying in the unit ball of an RKHS $\mathcal{H}$ and a fixed probability distribution ν, one can efficiently obtain an estimation $\hat{f}$ in the unit ball such that

$$\|f - \hat{f}\|_{L^2(\nu)} \leq \epsilon = O(n^{-\alpha}) \quad \text{(no curse of dimensionality)}$$

target function in supervised learning

$\downarrow$

reward function in reinforcement learning

If the reward function lies in an RKHS, what is the condition of the RKHS and transition dynamics to ensure that the reinforcement learning problem can be solved efficiently?

Challenge to Analysis: Q-value Function as An Example

Optimal Q-value function:

$$Q_h^*(s,a) = \sup_\pi \mathbb{E}_{P,\pi}[\sum_{h'=h}^{H} r(h', S_{h'}, A_{h'}) \,|\, S_h = s, A_h = a].$$

Optimal policy can be derived as the greedy policy of Q_h^*

$$\mathrm{supp}(\pi_h^*(\cdot \,|\, s)) \subset \{a \in \mathcal{A} : Q_h^*(s,a) = \max_{a' \in \mathcal{A}} Q_h^*(s,a')\}$$

If we have an estimation $\hat{Q}_h$ close to Q_h^* in the sense of $L^2(\nu)$, how to evaluate the performance of $\hat{\pi}$, the greedy policy of $\hat{Q}_h$?

Performance difference lemma: need the distribution under $\hat{\pi}$, unknown!

$$\begin{aligned} & J^*(M) - J(M, \hat{\pi}) \\ =& \sum_{h=1}^{H} \int_{\mathcal{S}\times\mathcal{A}} \sum_{a' \in \mathcal{A}} Q_h^*(s,a')[\pi_h^*(a' \,|\, s) - \hat{\pi}_h(a' \,|\, s)] \,\mathrm{d}\rho_{h,P,\hat{\pi},\mu}(s,a) \end{aligned}$$

Distribution Mismatch

$$J^*(M) - J(M, \hat{\pi})$$
$$= \sum_{h=1}^{H} \int_{\mathcal{S}\times\mathcal{A}} \sum_{a' \in \mathcal{A}} Q_h^*(s, a')[\pi_h^*(a' \mid s) - \hat{\pi}_h(a' \mid s)] \, \mathrm{d}\rho_{h,P,\hat{\pi},\mu}(s, a)$$

Distribution mismatch: mismatch between the distribution ν for estimation and the distribution for evaluation that is unknown a priori

"This lemma elucidates a fundamental measure mismatch. $\cdots$ Thus even if the optimal policy advantage is small with respect to π and μ, the advantages may not be small with respect to π^* and μ" (Kakade and Langford, 2002)

Perturbation Response by Distribution Mismatch

Definition 1

1. For any set Π of probability distribution on $\mathcal{S} \times \mathcal{A}$, we define a semi-norm Π-*norm* $\|\cdot\|_\Pi$ on $C(\mathcal{S} \times \mathcal{A})$ such that

$$\|f\|_\Pi := \sup_{\rho \in \Pi} \left| \int_{\mathcal{S} \times \mathcal{A}} f(s,a)\, \mathrm{d}\rho(s,a) \right|.$$

2. Given a Banach space $\mathcal{B}$, a positive constant $\epsilon > 0$ and a probability distribution $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, we define $\mathcal{B}_{\epsilon,\nu}$, a *$\nu$-perturbation space with scale ϵ*, as follows:

$$\mathcal{B}_{\epsilon,\nu} := \{f \in \mathcal{B}^1, \|f\|_{L^2(\nu)} \leq \epsilon\}.$$

3. The *perturbation response by distribution mismatch* is defined as the radius of $\mathcal{B}_{\epsilon,\nu}$ under Π-norm,

$$\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu) := \sup_{f \in \mathcal{B}_{\epsilon,\nu}} \|f\|_\Pi.$$

Remarks on Perturbation Response

$$\|f\|_{\Pi} \coloneqq \sup_{\rho \in \Pi} \left| \int_{\mathcal{S}\times\mathcal{A}} f(s,a)\, \mathrm{d}\rho(s,a) \right|$$

$$\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu) \coloneqq \sup_{f \in \mathcal{B}^1, \|f\|_{L^2(\nu)} \leq \epsilon} \|f\|_{\Pi}$$

- If $\Pi = \{\nu\}$, then $\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu) \leq \epsilon$
- If $\Pi = \mathcal{P}(\mathcal{S} \times \mathcal{A})$, then $\|f\|_{\Pi} = \|f\|_{\infty}$
  - used to handle the distribution mismatch in the tabular and linear RL
  - but may suffer from the curse of dimensionality in the RKHSs

The scale of $\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu)$ measures the discrepancy between ν and Π and reflects the error due to the fact that we do not know the state-action distribution under the policy of interest

Setting (Known Unknowns)

Solve among a family of MDPs $\mathcal{M} = \{M_\theta = (\mathcal{S}, \mathcal{A}, P_\theta, r_\theta, H, \mu) \colon \theta \in \Theta\}$

- $\mathcal{S}$, $\mathcal{A}$, H and μ are common state space, action space, episode length and initial distribution
- The possible transition probability P_θ and reward function r_θ is indexed by $\theta = (\theta_P, \theta_r)$, and $\Theta = \Theta_P \times \Theta_r$ is an index set as a Cartesian product.
- Reward functions lie in a unit ball of space $\mathcal{S}$

$$\{r_{\theta_r} \colon \theta_r \in \Theta_r\} = \{r \colon r(h, \cdot, \cdot) \in \mathcal{S}^1, \forall h \in [H]\}$$

  - $\mathcal{S} = \mathcal{B}$, a general Banach space for the lower bound
  - $\mathcal{S} = \mathcal{H}_k$, an RKHS with kernel k for the upper bound
- Θ_P is a given arbitrary set.
- Assume a generative simulator: for any h and state-action pair (s, a), we can observe a state $x \sim P_\theta(\cdot \mid h, s, a)$ and a noisy reward $y \sim \mathcal{N}(r_\theta(h, s, a), 1)$, called one access to the simulator

Worst-Case Error

Algorithm 1 General RL Algorithm for Estimating the Optimal Value

Input: Number of samples n

Initialize: $\mathcal{D}_0^{\theta,\xi} = \emptyset$.

for $i = 1, \ldots, n$ **do**

Obtain i-th step-state-action tuple through $(h_i^{\theta,\xi}, s_i^{\theta,\xi}, a_i^{\theta,\xi}) = f_i(\mathcal{D}_{i-1}^{\theta,\xi}, \bar{u})$

Collect the subsequent state $x_i^{\theta,\xi} \sim P_\theta(\cdot \mid h_i^{\theta,\xi}, s_i^{\theta,\xi}, a_i^{\theta,\xi})$ and the noisy reward $y_i^{\theta,\xi} = r_\theta(h_i^{\theta,\xi}, s_i^{\theta,\xi}, a_i^{\theta,\xi}) + \epsilon_i$ from the simulator

Set $\mathcal{D}_i^{\theta,\xi} = \mathcal{D}_{i-1}^{\theta,\xi} \cup \{(h_i^{\theta,\xi}, s_i^{\theta,\xi}, a_i^{\theta,\xi}, x_i^{\theta,\xi}, y_i^{\theta,\xi})\}$

end

Output: $J_n^{\theta,\xi} = F(\mathcal{D}_n^{\theta,\xi}, \bar{u})$ as an estimate of the optimal value $J^*(M_\theta)$

Find the best ξ to minimize worst-case error $\inf_{\xi \in \Xi_n} \sup_{\theta \in \Theta} \mathbb{E}|J_n^{\theta,\xi} - J^*(M_\theta)|$.

Two Cases and Some Notations

Two cases

- Known transition: $\Theta_P = \{0\}$
- Unknown transition: Θ_P is a given arbitrary set

Notations of distribution

- $\rho_{h,P,\pi,\mu}$: the distribution of (S_h, A_h) under initial distribution μ and policy π
- $\Pi(h, P, \mu) = \{\rho_{h,P,\pi,\mu} : \pi \text{ is an admissible policy}\}$
- $\Pi(P, \mu) = \bigcup_{h\in[H]} \Pi(h, P, \mu)$

Lower Bound (Known Transition)

Definition 2

The *perturbational complexity by distribution mismatch* in the case of known transition is

$$\Delta_{\mathcal{M}}(\epsilon) := \inf_{\nu \in \mathcal{P}(\mathcal{S}\times\mathcal{A})} \mathcal{R}(\Pi(P_0, \mu), \mathcal{B}, \epsilon, \nu).$$

Theorem 1 (Long and Han'21)

If there is only one possible transition probability, then

$$\inf_{\xi \in \Xi_n} \sup_{\theta \in \Theta} \mathbb{P}(|J_n^{\theta,\xi} - J^*(M_\theta)| \geq \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})) \geq \frac{1}{4}.$$

Therefore,

$$\inf_{\xi \in \Xi_n} \sup_{\theta \in \Theta} \mathbb{E}|J_n^{\theta,\xi} - J^*(M_\theta)| \geq \frac{1}{12}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}}).$$

Remark: we should care about how $\Delta_{\mathcal{M}}(\epsilon)$ decays with ϵ!

Lower Bound (Unknown Transition)

Direct application of Theorem 1 gives

$$\sup_{\theta \in \Theta} \Delta_{\mathcal{M}_\theta}(n^{-\frac{1}{2}})$$

as a lower bound

But the lower bound can be tighten by considering the following general sampling algorithm

$$\begin{cases} \mathcal{D}_0^{\theta,\bar{\xi}} = \emptyset, \mathcal{D}_i^{\theta,\bar{\xi}} = \mathcal{D}_{i-1}^{\theta,\bar{\xi}} \cup \{(h_i^{\theta,\bar{\xi}}, s_i^{\theta,\bar{\xi}}, a_i^{\theta,\bar{\xi}}, x_i^{\theta,\bar{\xi}})\}, \ 1 \leq i \leq n-1, \\ (h_i^{\theta,\bar{\xi}}, s_i^{\theta,\bar{\xi}}, a_i^{\theta,\bar{\xi}}) = g_i(\mathcal{D}_{i-1}^{\theta,\bar{\xi}}, \bar{u}), x_i^{\theta,\bar{\xi}} \sim P_\theta(\cdot \mid h_i^{\theta,\bar{\xi}}, s_i^{\theta,\bar{\xi}}, a_i^{\theta,\bar{\xi}}), \\ \nu^{\theta,\bar{\xi}} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \mathcal{L}(s_i^{\theta,\bar{\xi}}, a_i^{\theta,\bar{\xi}}). \end{cases}$$

$\bar{\Xi}_n$ denotes the set of all the possible sampling algorithms $\bar{\xi}$.

Lower Bound (Unknown Transition, Cont.)

Definition 3

The *perturbational complexity by distribution mismatch* in the case of unknown transition is

$$\Delta_{\mathcal{M}}(\epsilon) := \inf_{\bar{\xi}\in\bar{\Xi}_{[1/\epsilon^2]}} \sup_{\theta\in\Theta} \mathcal{R}(\Pi(P_\theta,\mu),\mathcal{B},\epsilon,\nu^{\theta,\bar{\xi}}).$$

Theorem 2 (Long and Han'21)

We have

$$\inf_{\xi\in\Xi_n} \sup_{\theta\in\Theta} \mathbb{P}(|J_n^{\theta,\xi} - J^*(M_\theta)| \geq \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})) \geq \frac{1}{4}.$$

Therefore,

$$\inf_{\xi\in\Xi_n} \sup_{\theta\in\Theta} \mathbb{E}|J_n^{\theta,\xi} - J^*(M_\theta)| \geq \frac{1}{12}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}}).$$

Upper Bound (Known Transition, Fitted Reward)

$$\hat{\nu} = \underset{\nu \in \mathcal{P}(\mathcal{S}\times\mathcal{A})}{\arg\min}\ \mathcal{R}(\Pi(P_0, \mu), \mathcal{H}_k, n^{-\frac{1}{2}}, \nu).$$

Algorithm 2 Fitted Reward Algorithm

Input: n^2 i.i.d. samples $z_1, \ldots, z_{n^2}$ from distribution $\hat{\nu}$

for $h = 1, 2, \ldots, H$ **do**

Sample $y_1^{\theta,h}, \ldots, y_{n^2}^{\theta,h}$ from $\mathcal{N}(r_\theta(h, z_1), 1), \ldots, \mathcal{N}(r_\theta(h, z_{n^2}), 1)$

Compute $\hat{r}_\theta(h, \cdot)$ as the minimizer of the optimization problem $\min_{\|r\|_k \le 1} \sum_{i=1}^{n^2} [r(z_i) - y_i^{\theta,h}]^2$

end

Collect the fitted reward function to form the MDP $(\mathcal{S}, \mathcal{A}, H, P_0, \hat{r}_\theta, \mu)$, of which both reward function and transition are known. Denote it as $\hat{M}_\theta$.

Output: $\hat{\pi}_\theta$ as the optimal policy of $\hat{M}_\theta$.

Upper Bound (Known Transition)

Theorem 3 (Long and Han'21)

If there is only one possible transition probability, and

$$\sup_{z \in \mathcal{S} \times \mathcal{A}} k(z,z) \leq 1.$$

For any $\theta \in \Theta$ and $p \in (0,1)$, with probability at least $1-p$, we can access the simulator $n^2 H$ times to have

$$|J(M_\theta, \hat{\pi}_\theta) - J^*(M_\theta)| \leq CH\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})\sqrt{1+\log(\frac{nH}{p})}.$$

Remark 1: again, we should care about how $\Delta_{\mathcal{M}}(\epsilon)$ decays with ϵ!

Remark 2: in the case of known transition, a low complexity RL problem is equivalent to that $\|f - \hat{f}\|_{\Pi(h,P_0,\mu)}$ can be small with finite samples for any $h \in [H]$

Upper Bound (Unknown Transition, Fitted Q-iteration)

$$\hat{\xi} = \arg\min_{\bar{\xi}\in\bar{\Xi}_n} \sup_{\theta\in\Theta} \mathcal{R}(\Pi(P_\theta,\mu),\mathcal{H}_k,n^{-\frac{1}{2}},\nu^{\theta,\bar{\xi}}),$$

Algorithm 3 Fitted Q-Iteration Algorithm

Input: n^2 samples $(\hat{z}^\theta_{1,1},\ldots,\hat{z}^\theta_{1,n}),\ldots,(\hat{z}^\theta_{n,1},\ldots,\hat{z}^\theta_{n,n})$ as i.i.d. copies of $((s_1^{\theta,\hat{\xi}},a_1^{\theta,\hat{\xi}}),\ldots,(s_n^{\theta,\hat{\xi}},a_n^{\theta,\hat{\xi}}))$

Initialize: $Q^\theta_{H+1}(s,a)=0$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$.

for $h=H,H-1,\ldots,1$ **do**

  for $i=1,\ldots,n$ *and* $j=1,\ldots,n$ **do**

    Sample $r^\theta_{i,j}\sim\mathcal{N}(r_\theta(h,\hat{z}^\theta_{i,j}),1)$ and $s^{\theta,\prime}_{i,j}\sim P_\theta(\cdot\,|\,h,\hat{z}^\theta_{i,j})$

    Compute $y^\theta_{i,j}=r^\theta_{i,j}+\max_{a'\in\mathcal{A}}Q_{h+1}(s^{\theta,\prime}_{i,j},a')$

  end

  Compute Q^θ_h as the minimizer of the optimization problem $\min_{\|f\|_k\le H-h+1}\sum_{i=1}^n\sum_{j=1}^n[f(\hat{z}^\theta_{i,j})-y^\theta_{i,j}]^2$

end

Output: $\hat{\pi}_\theta$ as the greedy policies with respect to $\{Q^\theta_h\}_{h=1}^H$.

Upper Bound (Unknown Transition)

Theorem 4 (Long and Han'21)

Assume $\|\mathcal{T}_h^\theta f\|_k \leq \|f\|_k + 1$ *where* $\mathcal{T}_h^\theta$ *is the Bellman optimal operator and*

$$\sup_{z \in \mathcal{S} \times \mathcal{A}} k(z,z) \leq 1.$$

Then, for any $\theta \in \Theta$, *with probability at least* $1-p$, *we can access the simulator* $n^2 H$ *times to have*

$$|J(M_\theta, \hat{\pi}_\theta) - J^*(M_\theta)| \leq C H^3 \Delta_{\mathcal{M}}(n^{-\frac{1}{2}}) \sqrt{1 + \log(\frac{nH}{p})}.$$

Connection with Concentratability

Proposition 5

Assume that there exists $1 < p \leq 2$ *and*

$$M = \sup_{\rho \in \Pi} \| \frac{\mathrm{d}\rho}{\mathrm{d}\nu} \|_{L^p(\nu)} < +\infty.$$

Then,

$$\mathcal{R}(\Pi, \mathcal{H}_k, n^{-\frac{1}{2}}, \nu) \leq 2Mn^{\frac{1}{p}-1}.$$

Connection with Kernel Decomposition

Given a probability distribution ν on $\mathcal{S} \times \mathcal{A}$, consider the operator $(\mathcal{K}_\nu)f(z) := \int_{\mathcal{S}\times\mathcal{A}} k(z, z')f(z')\,\mathrm{d}\nu(z')$ and use $\{\Lambda_i^\nu\}_{i\in\mathbb{N}^+}$ (nonincreasing) and $\{\psi_i^\nu\}_{i\in\mathbb{N}^+}$ (orthonormal) to denote its eigenvalues and eigenfunctions

Proposition 6

Assume $\sup_{z\in\mathcal{S}\times\mathcal{A}} k(z, z) \leq 1$, *then*

$$\inf_{\nu\in\mathcal{P}(\mathcal{S}\times\mathcal{A})} \mathcal{R}(\mathcal{P}(\mathcal{S}\times\mathcal{A}), \mathcal{H}_k, n^{-\frac{1}{2}}, \nu) \geq \frac{1}{2}\Big(\sup_{\rho\in\mathcal{P}(\mathcal{S}\times\mathcal{A})} \sum_{i=n+1}^{+\infty} \Lambda_i^\rho\Big)^{\frac{1}{2}}.$$

Moreover, if there exists a distribution $\hat{\nu} \in \mathcal{P}(\mathcal{S}\times\mathcal{A})$ *such that* $\sup_{i\in\mathbb{N}^+} \|\psi_i^{\hat{\nu}}\|_\infty < +\infty$, *then*

$$\mathcal{R}(\mathcal{P}(\mathcal{S}\times\mathcal{A}), \mathcal{H}_k, n^{-\frac{1}{2}}, \hat{\nu}) \leq 2\sqrt{\Big[\frac{n(\hat{\nu})}{n} + \sum_{i=n(\hat{\nu})+1}^{\infty} \Lambda_i^{\hat{\nu}}\Big]} \sup_{i\in\mathbb{N}^+} \|\psi_i^{\hat{\nu}}\|_\infty,$$

where $n(\hat{\nu}) = \max\{i \in \mathbb{N}^+ : n\Lambda_i^{\hat{\nu}} \geq 1\}$.

Decay of Eigenvalues

Remark 1. When there is ν so that the eigenvalue decay of $\mathcal{K}_\nu$ is fast, we can expect good convergence of RL algorithms.

Remark 2. When the eigenvalue decay is slow, like that in Laplace kernel and neural tangent kernel on sphere $\mathbb{S}^{d-1}$, the lower bound decays slow ($\sim n^{-\frac{1}{d-1}}$). The knowledge of Π plays a vital role.

Remark 3 (challenge in high dimensional action space). Assume

$$\mathcal{S} = \{s_0\}, \quad \mathcal{A} = \mathbb{S}^{d-1}, \quad H = 1.$$

Then the RL problem is essentially to find the maximum value of the reward function lying in the unit ball of $\mathcal{H}_k$ based on the values of n points. We need to assume the decay of eigenvalue is fast enough to break the curse of dimensionality.

Conclusions

- The perturbational complexity by distribution mismatch $\Delta_{\mathcal{M}}(\epsilon)$, gives a lower bound for the error of every algorithm on the considered RL problem.
- In the case of known transition, $\Delta_{\mathcal{M}}(\epsilon)$ also gives an upper bound of the error of the fitted reward algorithm.
- In the case of unknown transition (general case), with an additional assumption on Bellman operators, $\Delta_{\mathcal{M}}(\epsilon)$ gives an upper bound for the error of the fitted Q-iteration algorithm.
- $\Delta_{\mathcal{M}}(\epsilon)$ generalize existing results for fast convergence based on the assumptions of the finite concentration coefficients or fast eigenvalue decay of the kernel.
- We give a concrete example in which the reward functions lie in a high dimensional RKHS, the transition probability is known, and the action space is finite, but the corresponding RL problem can not be solved without the curse of dimensionality.

Open Problems

- Close the upper bound and lower bound
- Relax the assumption of Bellman operator or prove its necessity
- Bound for known reward but unknown transition
- Bound for episodic simulator
- Use perturbational complexity to guide the design of RL algorithms

Thank you for your attention