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

Jihao Long [*][1] and Jiequn Han [†][2]

[1]Program of Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA.
[2]Center for Computational Mathematics, Flatiron Institute, New York, NY 10010, USA.

Abstract. Most existing theoretical analysis of reinforcement learning (RL) is limited to the tabular setting or linear models due to the difficulty in dealing with function approximation in high dimensional space with an uncertain environment. This work offers a fresh perspective into this challenge by analyzing RL in a general reproducing kernel Hilbert space (RKHS). We consider a family of Markov decision processes $\mathcal{M}$ of which the reward functions lie in the unit ball of an RKHS and transition probabilities lie in a given arbitrary set. We define a quantity called *perturbational complexity by distribution mismatch* $\Delta_{\mathcal{M}}(\epsilon)$ to characterize the complexity of the admissible state-action distribution space in response to a perturbation in the RKHS with scale ϵ. We show that $\Delta_{\mathcal{M}}(\epsilon)$ gives both the lower bound of the error of all possible algorithms and the upper bound of two specific algorithms (fitted reward and fitted Q-iteration) for the RL problem. Hence, the decay of $\Delta_{\mathcal{M}}(\epsilon)$ with respect to ϵ measures the difficulty of the RL problem on $\mathcal{M}$. We further provide some concrete examples and discuss whether $\Delta_{\mathcal{M}}(\epsilon)$ decays fast or not in these examples. As a byproduct, we show that when the reward functions lie in a high dimensional RKHS, even if the transition probability is known and the action space is finite, it is still possible for RL problems to suffer from the curse of dimensionality.

1 Introduction

Modern reinforcement learning (RL) algorithms in practice often utilize function approximation tools to deal with problems involving an enormous amount of states in high dimensions. However, the majority of existing theoretical analysis of RL is only applicable to the tabular setting (see, e.g., [1–6]), in which both the state and action spaces are discrete and finite, and no function approximation is involved. Relatively simple function approximation methods, such as the linear model in [7, 8] or generalized linear model in [9, 10], have been recently studied in the context of RL with various statistical estimates. Yet, these results are not sufficient to explain the practical success of RL algorithms in high dimensions. In the tabular setting, the number of samples required by an RL algorithm is proportional to the size of state-action pairs, which is enormous in practice. For linear

---

model or generalized linear model, the assumption therein is pretty restrictive for practice. The kernel function is a class of models that can approximate more general functions than the tabular setting or linear model, and it is widely used in practice. Moreover, the kernel function approximation is closely related to neural network approximation, as established in the theory of neural tangent kernel in [11] and Barron space in [12]. RL with kernel function approximation has been recently studied in [13–17]. Still, results therein either suffer from the curse of dimensionality or require stringent assumptions on the kernel or the dynamics. As pointed out in [18] and [17, Section 5], L^∞-estimation, as a widely used technique in classical RL analysis of the tabular setting, may give rise to a curse of dimensionality in kernel method, which signifies new difficulties in RL algorithms with kernel function approximation. In this paper, we aim at a systematic study of RL with kernel function approximation and consider the following question:

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

Note that in this paper "efficiency" is considered in terms of sample complexity, i.e., the number of data an algorithm needs to collect to achieve a specified performance criterion. Our analysis will focus on the sample complexity of the RL problem, while other complexities such as computational complexity will not be covered. We are particularly interested in high dimensional state-action spaces and want to identify RL problems that can be solved efficiently with kernel function approximation even in high dimensions. Similar questions have been studied in the realm of supervised learning. There the answer is quite clear: once the target function lies in a reproducing kernel Hilbert space (RKHS), no matter how large the dimension is, the corresponding supervised learning problem can be solved efficiently (see, e.g., [19]). In RL, the reward function plays a similar role with the target function in supervised learning. This analogy motivates us to study a more concrete question:

If the unknown 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?

Below we give some intuition of the main challenge to answer this question and the key concept we introduce in this paper. Given an RKHS $\mathcal{H}$ and a probability distribution ν, existing results in supervised learning have shown that for any target distribution g lying in the unit ball of $\mathcal{H}$, one can efficiently obtain an estimation $\hat{g}$ such that

$$\|g - \hat{g}\|_{\mathcal{H}} \leq 2, \|g - \hat{g}\|_{L^2(\nu)} \leq \epsilon$$

for any $\epsilon > 0$ (see, e.g., [19] or Lemma 5.1). We can then view $\hat{g}$ as a ν-perturbation of g and define the ν-perturbation space with scale ϵ as

$$\mathcal{H}_{\epsilon,\nu} = \{g \in \mathcal{H} \colon \|g\|_{\mathcal{H}} \leq 1, \|g\|_{L^2(\nu)} \leq \epsilon\}.$$

While the distribution ν is given in supervised learning, in the theoretical analysis of RL, one needs to control the difference of the expectation between the target function and es-

timation with a probability distribution unknown a priori. That probability distribution is unknown because that is the state distribution or state-action distribution induced by a particular policy, which is unknown a priori. Take the estimation of the optimal Q-value function Q_h^* (see (5.8) below for the detailed definition) for example. The optimal policy can be derived from the optimal Q-value function through the greedy policy. In practice, given any probability distribution ν, under certain conditions, one can estimate the optimal Q-value function Q_h^* in the sense of $L^2(\nu)$ using Q-learning algorithm in [20] or fitted Q-iteration algorithm (see Algorithm 3 or [21] and [17]). In other words, one can obtain $\hat{Q}_h^*$, a ν-perturbation of Q_h^*. However, when evaluating the performance of $\hat{\pi}$, the greedy policy derived from $\hat{Q}_h^*$, one needs to control the error between Q_h^* and $\hat{Q}_h^*$ under the state-action distribution induced by the policy $\hat{\pi}$ (see the performance difference lemma in [22] or (5.16)), which is unknown before one obtains $\hat{Q}_h^*$. We call this phenomenon *distribution mismatch*: mismatch between the distribution ν for estimation and the distribution for evaluation that is unknown a priori. This phenomenon is ubiquitous in the analysis of RL (see, e.g., [22, Section 6]). Although not detailed above, when estimating the optimal Q-value function in the sense of $L^2(\nu)$, one needs to deal with the error propagation between steps, and distribution mismatch also brings difficulty.

To quantify the error brought by distribution mismatch, we define a semi-norm

$$\|g\|_\Pi = \sup_{\rho \in \Pi} \left| \int g \, \mathrm{d}\rho \right|,$$

where Π is a set of probability distributions and introduce the *perturbation response by distribution mismatch*:

$$\mathcal{R}(\Pi, \mathcal{H}, \epsilon, \nu) = \sup_{g \in \mathcal{H}_{\epsilon,\nu}} \|g\|_\Pi.$$

One shall notice that if $\Pi = \{\nu\}$, then $\mathcal{R}(\Pi, \mathcal{H}, \epsilon, \nu)$ cannot be greater than ϵ. However, in analysis, we usually can only choose Π as the possible state-action distributions under a class of policies. The scale of perturbation response by distribution mismatch measures the discrepancy between ν and Π and reflects the error brought by the fact that we do not know the state-action distribution under the policy of interest. If Π consists of all probability distributions, then the above semi-norm is just the L^∞-norm, which is used to handle the distribution mismatch in the tabular and linear RL problems. However, for many common RKHSs, the L^∞-estimation may suffer from the curse of dimensionality; see [18] and [17, Section 5] for a detailed discussion. The challenge of L^∞-estimation in high dimensional space reveals the difficulty of RL problems in the RKHS compared to the tabular setting or linear function approximation. In this sense, the introduced Π-norm can be understood as a generalization of the L^∞-norm to overcome this difficulty. This concept takes into account the distribution structure of the RL problem and allows us to do a more delicate error analysis. Following this idea, we introduce the *perturbational complexity by distribution mismatch* $\Delta_\mathcal{M}(\epsilon)$ for a large class of families of Markov decision processes (MDPs) and prove that once the perturbational complexity decreases fast with respect to ϵ, the RL problem can be solved efficiently. On the other hand, by considering the RL problem in which one only knows the reward function lies in the unit ball of a

general RKHS and transition probability lies in a given arbitrary set, we show that the perturbational complexity $\Delta_{\mathcal{M}}(\epsilon)$ must decay fast with respect to ϵ if this RL problem can be solved efficiently.

Combining the above two types of results together, we show that the perturbational complexity $\Delta_{\mathcal{M}}(\epsilon)$ measures the intrinsic difficulty of an RL problem. Note that most of our results still hold if we replace RKHS with a Banach space in which we can efficiently obtain an L^2-estimation, e.g., linear space or Barron space in [12]. Furthermore, our results shed some light on studying practical RL algorithms. First, the structure of Π has been used in the previous analysis of RL in various settings; see e.g., [13, 23–25]. While previous works mainly focus on the so-called concentration coefficients of Π and use related assumptions to prove upper bounds for RL problems, our work shows the necessity of additional assumptions on Π in order to ensure that the RL problem in the RKHS can be efficiently solved. As indicated by Proposition 6.3, if the eigenvalue decay of the kernel is slow and Π consists of all probability distributions, then $\Delta_{\mathcal{M}}(\epsilon)$ also decays slowly. Therefore, to design efficient RL algorithms, one needs to better understand the set Π, particularly when the eigenvalue decay of the kernel is slow. Second, when the unknown reward function lies in the unit ball of an RKHS and the transition probability is known, Theorems 4.1 and 5.1 show that solving the RL problem is equivalent to using finite values of a target function g to obtain a function estimate $\hat{g}$ that is accurate with respect to the Π-norm; see Remark 5.3 for detailed discussions. Theorems 4.2 and 5.2 also establish a partial connection between these two problems in the case of unknown transition probability. Therefore, it is helpful to study this supervised learning problem as a prototype of the RL problem.

Our contributions:

1. We define the perturbational complexity by distribution mismatch $\Delta_{\mathcal{M}}(\epsilon)$ for the families of MDPs $\mathcal{M}$ of which the reward functions lie in the unit ball of an RKHS and transition probabilities lie in a given arbitrary set. We then show that $\Delta_{\mathcal{M}}(\epsilon)$ gives a lower bound for the error of every algorithm on the corresponding RL problem (Theorems 4.1 and 4.2).

2. In the case of known transition (all transition probabilities in the families of MDPs are the same), we show that $\Delta_{\mathcal{M}}(\epsilon)$ also gives an upper bound of the error of the fitted reward algorithm (Algorithm 2) without any further assumption (Theorem 5.1).

3. In the case of unknown transition (general case), with an additional assumption on Bellman operators (5.11), we show that $\Delta_{\mathcal{M}}(\epsilon)$ gives an upper bound for the error of the fitted Q-iteration algorithm (Algorithm 3 and Theorem 5.2).

4. We give a concrete form of the perturbation response by distribution mismatch (Lemma 6.1) and show that when the assumptions on concentration coefficients in the existing literature (see e.g., [13, 17, 21, 24]) are satisfied or the eigenvalue decay of the kernel is fast, $\Delta_{\mathcal{M}}(\epsilon)$ decays fast with respect to ϵ (Proposition 6.2 and Proposition 6.3).

5. 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 (Proposition 6.4).

Related literature. While the optimal lower bound of the error of RL algorithms in the tabular setting has been established in [2, 4], there are much fewer results about lower bounds of RL with function approximation. [26] proves an optimal lower bound for Lipschitz function approximation. [27] shows that even when the value function, policy function, reward function, and transition probability can be approximated by a linear function, it is still possible that solving the RL problem requires samples exponentially depending on the horizon. [24] shows that even when the set of candidate approximating functions is finite and includes the optimal Q-value function, there does not exist an algorithm whose sample size is a polynomial function of the logarithm of the size of the candidate function set, the size of action space, horizon, and the reciprocal of accuracy. In other words, the previous works either consider function spaces (Lipschitz function space) that are too large to derive meaningful upper bound or only give lower bounds on special cases. Instead, we consider a fairly general class of RL problems associated with the RKHS and give both lower bound and upper bound through the perturbational complexity by distribution mismatch.

Previous works establish several upper bounds for RL algorithms with kernel function approximation. Based on the type of used assumptions, these works can be divided into two categories. The first category of upper bounds in [14–16] depends on the eigenvalue decay of kernel while the second category in [17, 23] requires accessibility to reference distributions that can uniformly bound all possible state-action distributions under admissible policies (called assumption on concentration coefficients). In this work, we show that the perturbational complexity $\Delta_{\mathcal{M}}(\epsilon)$ decays fast in both situations and establish an upper bound for the fitted reward algorithm (see Algorithm 2 in Section 5.1) and the fitted Q-iteration algorithm (see Algorithm 3 in Section 5.2) under the assumption that $\Delta_{\mathcal{M}}(\epsilon)$ decays fast. In this sense, our work generalizes both categories of the previous work.

Besides the error bounds of the RL algorithms, there is recent work in [28] studying policy evaluation in RKHS as a component of the RL algorithm and analyzing its optimal convergence rate.

Notation: Let $\mathcal{X}$ be an arbitrary subset of a Euclidean space, we use $C(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$ to denote the bounded continuous function space and probability distribution space on $\mathcal{X}$, respectively. We use $\|\cdot\|_{C(\mathcal{X})}$ to denote the uniform norm on $C(\mathcal{X})$:

$$\|g\|_{C(\mathcal{X})} = \sup_{x \in \mathcal{X}} |g(x)|.$$

Given a probability distribution ν on $\mathcal{X}$, we use $\|\cdot\|_{L^2(\nu)}$ and $\|\cdot\|_\infty$ to denote L^2-norm and L^∞-norm, respectively. Given two probability distributions μ and ν in $\mathcal{P}(\mathcal{X})$, define the total variation distance:

$$\|\mu - \nu\|_{TV} = \sup\{|\mu(A) - \nu(A)| \colon A \text{ is a measurable subset of } \mathcal{X}\}.$$

When μ is absolute continuous with respect to ν, define the Radon-Nikodym derivative

$\frac{d\mu}{d\nu}$ and the Kullback-Leibler divergence:

$$\mathrm{KL}(\mu \,||\, \nu) = \int_{\mathcal{X}} \log\left(\frac{d\mu}{d\nu}\right) d\mu.$$

For any random variable, we use $\mathcal{L}(\cdot)$ to denote the law of the random variable. Given a positive integer H, $[H]$ denotes the set $\{1, \cdots, H\}$. $\mathbb{N}^+$ denotes the set of all positive integers. $\mathbb{S}^{d-1}$ denotes the unit sphere of $\mathbb{R}^d$: $\{x \in \mathbb{R}^d, \|x\|_2 = 1\}$. Given a Banach space $\mathcal{B}$, we use $\|\cdot\|_{\mathcal{B}}$ to denote the norm of $\mathcal{B}$. We say $f(n) = \Theta(g(n))$, if there exist two constants $c, C > 0$ independent of n such that $cg(n) \leq f(n) \leq Cg(n)$.[1]

2 Preliminary

2.1 Markov Decision Process

We consider an episodic MDP $(\mathcal{S}, \mathcal{A}, H, P, r, \mu)$ as the mathematical model for the RL problem. Here H is a positive constant integer indicating the length of each episode. $\mathcal{S}$ and $\mathcal{A}$ denote the set of all the states and actions, respectively. We assume $\mathcal{S}$ is a subset of a Euclidean space and $\mathcal{A}$ is a compact subset of a Euclidean space. $P: [H] \times \mathcal{S} \times \mathcal{A} \mapsto \mathcal{P}(\mathcal{S})$ is the state transition probability. For each $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $P(\cdot \,|\, h, s, a)$ denotes the transition probability for the next state at step h if the current state is s and action a is taken. $r: [H] \times \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function, denoting the expected reward at step h if we choose action a at the state s. We assume each observed reward is the sum of the expected reward and an independent standard Gaussian noise. $\mu \in \mathcal{P}(\mathcal{S})$ is the initial distribution.

We denote a policy by $\pi = \{\pi_h\}_{h=1}^H \in \mathcal{P}(\mathcal{A} \,|\, \mathcal{S}, H)$, where

$$\mathcal{P}(\mathcal{A} \,|\, \mathcal{S}, H) = \left\{ \{\pi_h(\cdot \,|\, \cdot)\}_{h=1}^H \colon \pi_h(\cdot \,|\, s) \in \mathcal{P}(\mathcal{A}) \text{ for any } s \in \mathcal{S} \text{ and } h \in [H] \right\}.$$

Given a time step h, a transition probability P, a policy π and an initial distribution μ, we use $\rho_{h,P,\pi,\mu}$ to denote the distribution of (S_h, A_h) where $S_1 \sim \mu$, A_h follows the policy $\pi_h(\cdot \,|\, S_h)$ and S_{h+1} is distributed according to the transition probability $P(\cdot \,|\, h, S_h, A_h)$. Moreover, we use $\Pi(h, P, \mu)$ to denote the set of all the possible distributions of $\rho_{h,P,\pi,\mu}$ as follows

$$\Pi(h, P, \mu) = \{\rho_{h,P,\pi,\mu} \colon \pi \in \mathcal{P}(\mathcal{A} \,|\, \mathcal{S}, H)\},$$

and let

$$\Pi(P, \mu) = \bigcup_{h \in [H]} \Pi(h, P, \mu).$$

Given an MDP M and a policy π, we define the total reward as follows:

$$J(M, \pi) = \sum_{h=1}^H \int_{\mathcal{S} \times \mathcal{A}} r(h, s, a) \, d\rho_{h,P,\pi,\mu}(s, a).$$

---

[1] Later we also use Θ to denote an index set associated with a family of MDPs. The specific meaning should be always clear from context.

The optimal total reward is defined as $J^*(M) = \sup_{\pi \in \mathcal{P}(\mathcal{A} \mid \mathcal{S}, H)} J(M, \pi)$. We assume there exists at least one optimal policy π^* such that $J(M, \pi^*) = J^*(M)$.

2.2   Reproducing Kernel Hilbert Space (RKHS)

Suppose $k : (\mathcal{S} \times \mathcal{A}) \times (\mathcal{S} \times \mathcal{A}) \mapsto \mathbb{R}$ is a continuous positive definite kernel that satisfies

1. $k(z, z') = k(z', z), \forall z, z' \in \mathcal{S} \times \mathcal{A}$;

2. $\forall m \geq 1, z_1, \cdots, z_m \in \mathcal{S} \times \mathcal{A}$ and $c_1, \cdots, c_m \in \mathbb{R}$, we have:

$$\sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j k(z_i, z_j) \geq 0.$$

Then, there exists a Hilbert space $\mathcal{H}_k \subset C(\mathcal{S} \times \mathcal{A})$ such that

1. $\forall z \in \mathcal{S} \times \mathcal{A}, k(z, \cdot) \in \mathcal{H}_k$;

2. $\forall z \in \mathcal{S} \times \mathcal{A}$ and $g \in \mathcal{H}_k, g(z) = \langle g, k(z, \cdot) \rangle_k$,

and k is called the reproducing kernel of $\mathcal{H}_k$ in [29] and we use $\| \cdot \|_k$ and $\langle \cdot, \cdot \rangle_k$ to denote the norm and inner product in the Hilbert space $\mathcal{H}_k$, respectively.

Given a probability distribution ν on $\mathcal{S} \times \mathcal{A}$, we will use $\{\Lambda_i^\nu\}_{i \in \mathbb{N}^+}$ and $\{\psi_i^\nu\}_{i \in \mathbb{N}^+}$ to denote the eigenvalues and eigenfunctions of the operator

$$(\mathcal{K}_\nu g)(z) := \int_{\mathcal{S} \times \mathcal{A}} k(z, z') g(z') \, \mathrm{d}\nu(z')$$

from $L^2(\nu)$ to $L^2(\nu)$. We further require that $\{\Lambda_i^\nu\}_{i \in \mathbb{N}^+}$ is nonincreasing and $\{\psi_i^\nu\}_{i \in \mathbb{N}^+}$ is orthonormal in $L^2(\nu)$. The famous Mercer decomposition states that

$$k(z, z') = \sum_{i=1}^{+\infty} \Lambda_i^\nu \psi_i^\nu(z) \psi_i^\nu(z'). \tag{2.1}$$

Moreover, for any $g \in \mathcal{H}_k$

$$\|g\|_k^2 = \sum_{i=1}^{+\infty} \frac{1}{\Lambda_i^\nu} |\langle g, \psi_i^\nu \rangle_{L^2(\nu)}|^2. \tag{2.2}$$

See, e.g., [30, Section 2.1].

Given any two probability distributions ρ and ρ' on $\mathcal{S} \times \mathcal{A}$, the maximum mean discrepancy (MMD) is defined as follows (see e.g. [31]):

$$\mathrm{MMD}_k(\rho, \rho') = \sup_{\|g\|_k \leq 1} \left| \int_{\mathcal{S} \times \mathcal{A}} g(z) \, \mathrm{d}\rho(z) - \int_{\mathcal{S} \times \mathcal{A}} g(z) \, \mathrm{d}\rho'(z) \right|.$$

An equivalent but more concrete expression of MMD is

$$\mathrm{MMD}_k(\rho, \rho') = \sqrt{\int_{\mathcal{S} \times \mathcal{A}} \int_{\mathcal{S} \times \mathcal{A}} k(z, z') \, \mathrm{d}(\rho - \rho')(z) \, \mathrm{d}(\rho - \rho')(z')}.$$

3   Problem setup

We first specify our prior knowledge of the RL problem. We want to solve an RL problem whose underlying MDP belongs to a family of MDPs

$$\mathcal{M} = \{M_\theta = (\mathcal{S}, \mathcal{A}, P_\theta, r_\theta, H, \mu)\colon \theta \in \Theta\},$$

where $\mathcal{S}$, $\mathcal{A}$, H and μ are common state space, action space, length of each episode and initial distribution. The possible transition probability P_θ and reward function r_θ is indexed by θ, and Θ is an index set. We do not know the exact value of θ but can access a generative simulator. In other words, for any step $h \in [H]$ and state-action pair (s, a), we can observe a state $x \sim P_\theta(\,\cdot\,|\,h, s, a)$ and a noisy reward $y \sim \mathcal{N}(r_\theta(h, s, a), 1)$, which is called one sample or one access to the generative simulator. So far we need to assume the noise of the reward is Gaussian to prove the lower bounds, but the noise can be relaxed to be sub-Gaussian in the upper bounds. Another popular form of the simulator is the so-called episodic simulator, through which one can only choose the initial state and a policy to observe the whole path and corresponding rewards. Our lower bound is still valid if we only have an episodic simulator but might be loose. How to obtain a tight lower bound in those settings is left to future work.

We assume $\theta = (\theta_P, \theta_r)$ and the index set is a Cartesian product

$$\Theta = \{(\theta_P, \theta_r)\colon \theta_P \in \Theta_P,\ \theta_r \in \Theta_r\},$$

where θ_P and θ_r are the actual indexes of the transition probability and reward function, i.e., $P_\theta = P_{\theta_P}$, $r_\theta = r_{\theta_r}$. We also assume

$$\{r_{\theta_r}\colon \theta_r \in \Theta_r\} = \{r\colon \|r(h, \cdot, \cdot)\|_\mathcal{B} \leq 1,\ \forall h \in [H]\},$$

where $\mathcal{B}$ is a Banach space such that $\mathcal{B}$ is a subset of $C(\mathcal{S} \times \mathcal{A})$ and $\|\cdot\|_{C(\mathcal{S} \times \mathcal{A})} \leq B\|\cdot\|_\mathcal{B}$ with a positive constant B.

Following the intuition introduced in Section 1, we give the following definitions in preparation for the analysis.

Definition 3.1.

(i) *For any set Π consisting of probability distributions on $\mathcal{S} \times \mathcal{A}$, we define a semi-norm $\|\cdot\|_\Pi$ on $C(\mathcal{S} \times \mathcal{A})$:*

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

We call this semi-norm Π-norm.

(ii) *Given a Banach space, 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 ν-perturbation space with scale ϵ, as follows:*

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

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

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

3.1 Properties of perturbation response by distribution mismatch

We first state two propositions later used to give readers more understanding of the properties of perturbation response by distribution mismatch $\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu)$. The proofs of these two propositions are postponed to Section 6. The first proposition gives a more concrete formula of $\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu)$. Specifically, when $\mathcal{B}$ is an RKHS, $\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu)$ can be determined by a maximin problem related to MMD; see (3.1).

Proposition 6.1. We have

$$\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu) = \sup_{\rho \in \Pi} \inf_{g \in L^2(\nu)} [\|\rho - g \circ \nu\|_{\mathcal{B}^*} + \epsilon \|g\|_{L^2(\nu)}],$$

where $g \circ \nu$ is a signed measure such that

$$\frac{\mathrm{d}g \circ \nu}{\mathrm{d}\nu} = g,$$

$\mathcal{B}^*$ is the dual space of $\mathcal{B}$ and $\|\rho\|_{\mathcal{B}^*}$ is the dual norm of linear functional

$$\rho(g) := \int_{\mathcal{S} \times \mathcal{A}} g(z) \, \mathrm{d}\rho(z), \quad \forall g \in \mathcal{B},$$

for any signed measure ρ on $\mathcal{S} \times \mathcal{A}$ (we slightly abuse the notation that ρ are both the signed measure and linear functional in $\mathcal{B}$). If $\mathcal{B}$ is an RKHS with kernel k, then

$$\mathcal{R}(\Pi, \mathcal{H}_k, \epsilon, \nu) = \sup_{\rho \in \Pi} \inf_{g \in L^2(\nu)} [\mathrm{MMD}_k(\rho, g \circ \nu) + \epsilon \|g\|_{L^2(\nu)}]. \tag{3.1}$$

When $\mathcal{B}$ is an RKHS, the kernel's eigenvalues encode much information. The following proposition shows that the perturbation response of $\mathcal{P}(\mathcal{S} \times \mathcal{A})$, the set of all probability distributions on $\mathcal{S} \times \mathcal{A}$, is closely related to the kernel's eigenvalues. Later in Section 6 we will discuss how this proposition gives us the implication in the efficiency of RL algorithms.

Proposition 6.3. Assume that

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

For any $\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, define

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

We have

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

and, by $n(\nu) \leq n$,

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

Moreover, if there exists a distribution $\hat{v} \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ such that

$$\sup_{i \in \mathbb{N}^+} \|\psi_i^{\hat{v}}\|_\infty < +\infty,$$

then

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

3.2 General algorithm

Now we state in Algorithm 1 the general RL algorithm for estimating the optimal value $J^*(M_\theta)$ with n samples. In Algorithm 1, the superscript θ indicates that the collected data depends on the underlying MDP M_θ. The superscript ξ denotes the collection $\{f_1, \cdots, f_n, F\}$, where f_i are measurable mappings: $([H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{R})^{\otimes i-1} \times \mathbb{R} \mapsto [H] \times \mathcal{S} \times \mathcal{A}$, F is a measurable mapping: $([H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{R})^{\otimes n} \times \mathbb{R} \mapsto \mathbb{R}$. ξ can be viewed as an RL algorithm, which adaptively chooses the step-state-action tuple (h, s, a) at the i-th step based on all received data $\mathcal{D}_{i-1}^{\theta,\xi}$ according to function f_i and receives a subsequent state and reward through the generative simulator. After collecting n samples, the algorithm outputs an estimate of the optimal value based on all data according to function F. The randomness of the whole process in Algorithm 1 is related to i.i.d. standard normal random variables $\{\epsilon_i\}_{1 \le i \le n}$, $\{u_i\}_{1 \le i \le n}$, and $\bar{u}$, which all live in a common probability space $(\Omega, \mathbb{P})$. ϵ_i denotes the noise in the observed reward. u_i denotes the randomness of the transition, for which we assume that, by the isomorhism theorem [32, Section 41], $p_\theta : [H] \times \mathcal{S} \times \mathcal{A} \times \mathbb{R} \mapsto \mathcal{S}$ is a measurable function satisfying $p_\theta(h, s, a, u_i) \sim P_\theta(\cdot \mid h, s, a)$ for any $\theta \in \Theta$ and $h \in [H]$. Again by the isomorphism theorem, we use $\bar{u}$ to denote all the randomness of the algorithm ξ itself besides the randomness within the simulator. Mathematically, Algorithm 1 can also be summarized as follows:

$$\begin{cases} \mathcal{D}_0^{\theta,\xi} = \varnothing, \quad \mathcal{D}_i^{\theta,\xi} = \mathcal{D}_{i-1}^{\theta,\xi} \cup \{(h_i^{\theta,\xi}, s_i^{\theta,\xi}, a_i^{\theta,\xi}, x_i^{\theta,\xi}, y_i^{\theta,\xi})\}, \quad 1 \le i \le n, \\ J_n^{\theta,\xi} = F(\mathcal{D}_n^{\theta,\xi}, \bar{u}), \quad (h_i^{\theta,\xi}, s_i^{\theta,\xi}, a_i^{\theta,\xi}) = f_i(\mathcal{D}_{i-1}^{\theta,\xi}, \bar{u}), \\ x_i^{\theta,\xi} = p_\theta(h_i^{\theta,\xi}, s_i^{\theta,\xi}, a_i^{\theta,\xi}, u_i), \quad y_i^{\theta,\xi} = r_\theta(h_i^{\theta,\xi}, s_i^{\theta,\xi}, a_i^{\theta,\xi}) + \epsilon_i. \end{cases} \quad (3.2)$$

We use Ξ_n to denote the set of all possible choices of ξ. So Ξ_n is the set of all possible RL algorithms which only access the generative simulator n times. Our goal is to find the best ξ, or the best RL algorithm, to minimize the worst-case error of the optimal total reward given n opportunities to access the simulator:

$$\inf_{\xi \in \Xi_n} \sup_{\theta \in \Theta} \mathbb{E}|J_n^{\theta,\xi} - J^*(M_\theta)|.$$

In Sections 4 and 5 below, we give lower and upper bounds for the worst-case error, respectively. In both sections, we first consider the special case where the transition probability is known and then generalize our results to the case where the transition probability is unknown. In practice, it is often of interest to estimate the optimal policy as well. In

---

Algorithm 1 General Reinforcement Learning Algorithm for Estimating the Optimal Value

---

Input: Number of samples n

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

for $i = 1, \cdots, n$ **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 $x_i^{\theta,\xi} = p_\theta(h_i^{\theta,\xi}, s_i^{\theta,\xi}, a_i^{\theta,\xi}, u_i)$ 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 $J^*(M_\theta)$

---

the upper bound part, we also provide algorithms to obtain the optimal policy that gives the estimated optimal total reward. Nevertheless, in the lower bound part, we abstractly estimate the optimal total reward without estimating the optimal policy. Note that we can always use the Monte-Carlo method to estimate the optimal total reward accurately given an optimal or near-optimal policy. So our lower bound result still serves as a valid difficulty measure of the RL problem aiming at finding the optimal policy.

4 Lower bound

4.1 The case of known transition

We first consider the case that the transition probability P_θ is known, assuming $\Theta_P = \{0\}$ is a single-point set. In this case, we have the following definition of the perturbational complexity by distribution mismatch.

Definition 4.1. *The* perturbational complexity by distribution mismatch *in the case of known transition is*

$$\Delta_{\mathcal{M}}(\epsilon) := \inf_{\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})} \mathcal{R}(\Pi(P_0, \mu), \mathcal{B}, \epsilon, \nu). \tag{4.1}$$

The following theorem shows that this quantity can give a lower bound of the worst-case error.

Theorem 4.1. *Assume $\Theta_P = \{0\}$, i.e., 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,

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

Proof. With fixed $g \in \mathcal{B}$ such that $\|g\|_{\mathcal{B}} \leq 1$, $\xi \in \Xi_n$ and $h^* \in [H]$, we first estimate

$$\|\mathcal{L}(J_n^{\theta_1,\xi}) - \mathcal{L}(J_n^{\theta_2,\xi})\|_{TV},$$

where $\theta_1, \theta_2 \in \Theta$ satisfying

$$P_{\theta_1} = P_{\theta_2} = P_0, \quad r_{\theta_1} = 0, \quad r_{\theta_2}(h,s,a) = \begin{cases} 0, & \text{when } h \neq h^*, \\ g(s,a), & \text{when } h = h^*. \end{cases}$$

By the definition (3.2) and Pinsker's inequality (see e.g. [33,34]), we have

$$\|\mathcal{L}(J_n^{\theta_1,\xi}) - \mathcal{L}(J_n^{\theta_2,\xi})\|_{TV}^2 \leq \|\mathcal{L}(\mathcal{D}_n^{\theta_1,\xi}, \bar{u}) - \mathcal{L}(\mathcal{D}_n^{\theta_2,\xi}, \bar{u})\|_{TV}^2$$
$$\leq \frac{1}{2}\mathrm{KL}(\mathcal{L}(\mathcal{D}_n^{\theta_1,\xi}, \bar{u})\|\mathcal{L}(\mathcal{D}_n^{\theta_2,\xi}, \bar{u})).$$

Calculation gives that

$$\mathrm{KL}(\mathcal{L}(\mathcal{D}_n^{\theta_1,\xi}, \bar{u})\|\mathcal{L}(\mathcal{D}_n^{\theta_2,\xi}, \bar{u}))$$
$$= \mathbb{E}\log\left(\prod_{i=1}^n \exp\left(-\frac{(y_i^{\theta_1,\xi})^2}{2} + \frac{y_i^{\theta_1,\xi} - r_{\theta_2}(h_i^{\theta_1,\xi}, s_i^{\theta_1,\xi}, a_i^{\theta_1,\xi})^2}{2}\right)\right)$$
$$= \frac{1}{2}\sum_{i=1}^n \mathbb{E}[|g(s_i^{\theta_1,\xi}, a_i^{\theta_1,\xi})|^2 - 2g(s_i^{\theta_1,\xi}, a_i^{\theta_1,\xi})\epsilon_i]\mathbb{1}_{h_i^{\theta_1,\xi}=h^*}$$
$$\leq \frac{1}{2}\sum_{i=1}^n \mathbb{E}[|g(s_i^{\theta_1,\xi}, a_i^{\theta_1,\xi})|^2].$$

Combining the last two inequalities, we know that

$$\|\mathcal{L}(J_n^{\theta_1,\xi}) - \mathcal{L}(J_n^{\theta_2,\xi})\|_{TV}^2 \leq \frac{1}{4}\sum_{i=1}^n \mathbb{E}[g(s_i^{\theta_1,\xi}, a_i^{\theta_1,\xi})]^2 = \frac{n}{4}\int_{\mathcal{S}\times\mathcal{A}} |g(s,a)|^2 \, d\nu^\xi(s,a),$$

where

$$\nu^\xi = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(s_i^{\theta_1,\xi}, a_i^{\theta_1,\xi}).$$

By definition, we have

$$\Delta_{\mathcal{M}}(n^{-\frac{1}{2}}) \leq \sup_{h\in[H]} \sup_{\|g\|_{\mathcal{B}}\leq 1, \sqrt{n}\|g\|_{L^2(\nu^\xi)}\leq 1} \sup_{\rho\in\Pi(h,P_0,\mu)} \int_{\mathcal{S}\times\mathcal{A}} g(s,a) \, d\rho(s,a).$$

So we can choose $g \in \mathcal{B}$ with $\|g\|_{\mathcal{B}} \leq 1$ and $h^* \in [H]$ to define a reward function r_{θ_2} such that

$$\int_{\mathcal{S}\times\mathcal{A}} |g(s,a)|^2 \, d\nu^\xi(s,a) \leq \frac{1}{n},$$
$$J^*(M_{\theta_2}) \geq \frac{2}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}}),$$

which means that

$$\|\mathcal{L}(J_n^{\theta_1,\xi}) - \mathcal{L}(J_n^{\theta_2,\xi})\|_{TV} \leq \frac{1}{2},$$

$$\left\{x \in \mathbb{R}\colon |x - J^*(M_{\theta_2})| < \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})\right\} \cap \left\{x \in \mathbb{R}\colon |x| < \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})\right\} = \varnothing.$$

Noticing that $J^*(M_{\theta_1}) = 0$, we can use the second condition above to have

$$\mathbb{P}(|J_n^{\theta_1,\xi} - J^*(M_{\theta_1})| \geq \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})$$

$$= \ \mathbb{P}(|J_n^{\theta_1,\xi}| \geq \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})$$

$$\geq \ \mathbb{P}(|J_n^{\theta_1,\xi} - J^*(M_{\theta_2})| < \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})).$$

Therefore,

$$\frac{1}{2} \geq \mathbb{P}(|J_n^{\theta_2,\xi} - J^*(M_{\theta_2})| < \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})) - \mathbb{P}(|J_n^{\theta_1,\xi} - J^*(M_{\theta_2})| < \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}}))$$

$$\geq 1 - \mathbb{P}(|J_n^{\theta_2,\xi} - J^*(M_{\theta_2})| \geq \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})) - \mathbb{P}(|J_n^{\theta_1,\xi} - J^*(M_{\theta_1})| \geq \frac{1}{3}\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})).$$

Rearranging the above inequality, we can then conclude that for any $\xi \in \Xi_n$,

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

This completes the proof. $\qquad\square$

4.2 The case of unknown transition

In this section, we deal with the general case when there are multiple possible transition probabilities. Following the idea of Theorem 4.1,

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

can provide a lower bound of the worst-case error, where $\mathcal{M}_\theta$ is the subset of $\mathcal{M}$ whose transition probability is P_θ. However, we can have an even better lower bound. The critical observation is that we do not know the exact value of P_θ, but can only sample from P_θ with finite observed data. So the optimal distribution ν for estimation corresponding to $\Delta_{\mathcal{M}_\theta}(n^{-\frac{1}{2}})$ is generally inaccessible, and we can leverage this fact to better characterize the perturbational complexity to improve the lower bound. To this end, we assume a general sampling algorithm as follows to characterize the set of distribution ν:

$$\begin{cases} \mathcal{D}_0^{\theta,\bar{\xi}} = \varnothing, \quad \mathcal{D}_i^{\theta,\bar{\xi}} = \mathcal{D}_{i-1}^{\theta,\bar{\xi}} \cup \{(h_i^{\theta,\bar{\xi}}, s_i^{\theta,\bar{\xi}}, a_i^{\theta,\bar{\xi}}, x_i^{\theta,\bar{\xi}})\}, \quad 1 \leq i \leq n, \\ (h_i^{\theta,\bar{\xi}}, s_i^{\theta,\bar{\xi}}, a_i^{\theta,\bar{\xi}}) = \bar{f}_i(\mathcal{D}_{i-1}^{\theta,\bar{\xi}}, \bar{u}), \quad x_i^{\theta,\bar{\xi}} = p_\theta(h_i^{\theta,\bar{\xi}}, s_i^{\theta,\bar{\xi}}, a_i^{\theta,\bar{\xi}}, u_i), \\ \nu^{\theta,\bar{\xi}} = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(s_i^{\theta,\bar{\xi}}, a_i^{\theta,\bar{\xi}}). \end{cases} \quad (4.2)$$

Here $\bar{\xi} = (\bar{f}_1, \cdots, \bar{f}_n)$ and $\bar{f}_i$ are measurable mappings $([H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S})^{\otimes(i-1)} \times \mathbb{R} \mapsto [H] \times \mathcal{S} \times \mathcal{A}$. Eq. (4.2) shares a similar spirit with (3.2), but only focuses on sampling the next-step states from the transition probability. Similar to ξ in (3.2), $\bar{\xi}$ can be viewed as a sampling algorithm which adaptively chooses n step-state-action tuples (h, s, a) and obtains samples from $P_\theta(\cdot \mid h, s, a)$. We let $\bar{\Xi}_n$ denote the set of all possible choices of $\bar{\xi}$. Now we define the perturbational complexity by distribution mismatch in the case of unknown transition.

Definition 4.2. *The* perturbational complexity by distribution mismatch *in the case of unknown transition is*

$$\Delta_{\mathcal{M}}(\epsilon) := \inf_{\bar{\xi} \in \bar{\Xi}_{[1/\epsilon^2]}} \sup_{\theta \in \Theta} \mathcal{R}(\Pi(P_\theta, \mu), \mathcal{B}, \epsilon, \nu^{\theta, \bar{\xi}}). \tag{4.3}$$

Here we choose the number of samples to be $[1/\epsilon^2]$ so that the result is consistent with the case of known transition. The following theorem shows that perturbational complexity gives a lower bound of the worst-case error in the case of unknown transition.

Theorem 4.2. *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,

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

Proof. Following the proof of Theorem 4.1, we know that for any $\xi \in \Xi_n$,

$$\sup_{\theta \in \Theta} \mathbb{P}\left(|J_n^{\theta, \xi} - J^*(M_\theta)| \geq \frac{1}{3} \sup_{\theta \in \Theta} \mathcal{R}(\Pi(P_\theta, \mu), \mathcal{B}, n^{-\frac{1}{2}}, \nu^{\theta, \xi})\right) \geq \frac{1}{4}.$$

Here

$$\nu^{\theta, \xi} = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(s_i^{\theta_0(\theta), \xi}, a_i^{\theta_0(\theta), \xi}),$$

and $\{s_i^{\theta_0(\theta), \xi}, a_i^{\theta_0(\theta), \xi}\}_{1 \leq i \leq n}$ is generated by the sampling path (3.2) with $P_{\theta_0(\theta)} = P_\theta$ and $r_{\theta_0(\theta)} = 0$. Hence, $y_i^{\theta_0(\theta), \xi} = \epsilon_i$ for any $1 \leq i \leq n$. Using the isomorphism theorem [32, Section 41], we can find the measurable mappings $T_i : \mathbb{R} \mapsto \mathbb{R}$ for $1 \leq i \leq n+1$ such that

$$(T_1(\bar{u}), \cdots, T_n(\bar{u}), t_{n+1}(\bar{u})) \text{ has the same distribution with } (\epsilon_1, \cdots, \epsilon_n, \bar{u}).$$

Therefore, for any $\xi \in \Xi_n$, there exists $\bar{\xi} \in \bar{\Xi}_n$ such that for all $\theta \in \Theta$,

$$\nu^{\theta, \bar{\xi}} = \nu^{\theta, \xi}.$$

Therefore,

$$\sup_{\theta \in \Theta} \mathcal{R}(\Pi(P_\theta, \mu), \mathcal{B}, n^{-\frac{1}{2}}, \nu^{\theta, \xi}) \geq \Delta_{\mathcal{M}}(n^{-\frac{1}{2}}),$$

which concludes our proof. $\square$

5   Upper bound

In this section, we discuss how to use $\Delta_{\mathcal{M}}(\epsilon)$ in Definitions 4.1, 4.2 to design sample-efficient RL algorithms. We will use C to denote a universal positive constant, which may vary from line to line. As motivated in the introduction, one important reason for us to consider the perturbation response is to study those high-dimensional spaces in which L^∞ estimation can not be obtained efficiently through finite samples. Many common RKHSs are such examples. So in this section we focus on the case that $\mathcal{B}$ is an RKHS with kernel k. We remark that most of our results still hold for general Banach spaces whenever an L^2 estimation can be obtained efficiently through finite samples, such as linear space and Barron space [12].

5.1   The case of known transition

Again we first consider the case where the transition probability is known, i.e., $\Theta_P = \{0\}$. We let
$$\hat{v} = \underset{v \in \mathcal{P}(\mathcal{S} \times \mathcal{A})}{\arg\min} \ \mathcal{R}(\Pi(P_0, \mu), \mathcal{H}_k, n^{-\frac{1}{2}}, v).$$

Here $\hat{v}$ is an intrinsic property of the MDP family $\mathcal{M}$ without any dependence of sampling. In practice, the minimizer $\hat{v}$ is probably hard to obtain but similar argument holds if we can obtain a probability distribution $\hat{v}'$ such that

$$\mathcal{R}(\Pi(P_0, \mu), \mathcal{H}_k, n^{-\frac{1}{2}}, \hat{v}')$$

is small. Once we know the distribution $\hat{v}$, we can sample n^2 i.i.d. samples $z_1, \cdots, z_{n^2}$ from $\hat{v}$ and perform the following fitted reward algorithm to estimate the optimal policy. Notice that, in order to solve (5.1) below, we only need to solve a convex optimization problem with the same objective function but in the finite-dimensional set

$$\left\{ \|r\|_k \leq 1 \colon \ r = \sum_{i=1}^{n^2} c_i k(\cdot, z_i) \right\}.$$

The reason is that for any $r \in \mathcal{H}_k$, we can always find coefficients $\{c_i\}_{1 \leq i \leq n^2}$ to construct a function $\sum_{i=1}^{n^2} c_i k(\cdot, z_i)$ such that

$$r(z_j) = \sum_{i=1}^{n^2} c_i k(z_j, z_i), \ 1 \leq j \leq n^2, \quad \text{and} \quad \|r\|_k \geq \left\| \sum_{i=1}^{n^2} c_i k(\cdot, z_i) \right\|_k.$$

See, e.g., [35, Proposition 4.2].

   The algorithm described above based on the fitted reward is summarized in Algorithm 2, and we have the following convergence result.

Theorem 5.1. *Assume* $\Theta_P = \{0\}$, *i.e., there is only one possible transition probability, and*

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

Algorithm 2 Fitted Reward Algorithm

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

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

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

   Compute $\hat{r}_\theta(h, \cdot)$ as the minimizer of the optimization problem

$$\min_{\|r\|_k \leq 1} \sum_{i=1}^{n^2} [r(z_i) - y_i^{\theta,h}]^2 \tag{5.1}$$

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 $\hat{M}_\theta$.

For any $\theta \in \Theta$ and $p \in (0, 1)$, with probability at least $1 - p$, we have

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

Remark 5.1. As we need Hn^2 samples to achieve the upper bound $\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})$, the convergence rate with respect to n in this theorem does not match the lower bound offered in Theorem 4.1. Still, $\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})$ does give an indication whether an RL problem can be solved efficiently or not. For example, if $\Delta_{\mathcal{M}}(n^{-\frac{1}{2}}) = \Theta(n^{-\frac{1}{2}})$, then the convergence rate with respect to n is between $n^{-\frac{1}{4}}$ and $n^{-\frac{1}{2}}$. We can hence know that the corresponding RL problem can be solved efficiently. On the other hand, if and only if $\Delta_{\mathcal{M}_d}(n^{-\frac{1}{2}}) = \Theta(n^{-\frac{1}{d}})$, where $(\mathcal{M}_d)_{d \in \mathbb{N}^+}$ are families of MDPs and d denotes the dimension of state-action space of $\mathcal{M}_d$, we know that the RL problems suffer from the curse of dimensionality. Overall, whether we can establish a dimension-free convergence rate for $\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})$ determines whether we can construct a dimension-free RL algorithm. Similar arguments hold for the case of unknown transition; see Theorems 4.2 and 5.2.

    Before proving Theorem 5.1, we first prove the following lemma concerning the L^2-distance between r_θ and $\hat{r}$ solved in Algorithm 2.

Lemma 5.1. *Assume that*

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

Let $(z_{1,1}, \cdots, z_{1,n}), \cdots, (z_{n,1}, \cdots, z_{n,n})$ be i.i.d. drawn from a distribution $v \in (\mathcal{S} \times \mathcal{A})^{\otimes n}$ (the n-ary Cartesian power of $\mathcal{S} \times \mathcal{A}$) and

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i,$$

where $\{\nu_i\}_{1 \leq i \leq n}$ are marginal distributions of ν. Random variables $\{\epsilon_{i,j}\}_{1 \leq i \leq n, 1 \leq j \leq n}$, conditional on $\{z_{i,j}\}_{1 \leq i \leq n, 1 \leq j \leq n}$, are independent, 1-subguassian and mean zero. Given $r^* \in \mathcal{H}_k$ with $\|r^*\|_k \leq 1$, we let

$$\hat{r} = \arg\min_{\|r\|_k \leq 1} \sum_{i=1}^{n} \sum_{j=1}^{n} [r(z_{i,j}) - r^*(z_{i,j}) - \epsilon_{i,j}]^2. \tag{5.2}$$

Then, for any $p \in (0,1)$, with probability at least $1 - p$, we have

$$\|\hat{r} - r^*\|_{L^2(\bar{\nu})} \leq C \sqrt{\frac{1}{n} \left[1 + \log\left(\frac{n}{p}\right) \right]}.$$

Remark 5.2. Note that the setting of Lemma 5.1 is more general than Theorem 5.1 since we do not require $z_{i,1}, \cdots, z_{i,n}$ to be i.i.d. random variables and we know the error is Gaussian instead of sub-Gaussian. We will use this general setting later to deal with the case of unknown transition; see the proof of Theorem 5.2.

Proof. Following (5.2), we know that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} [\hat{r}(z_{i,j}) - r^*(z_{i,j}) - \epsilon_{i,j}]^2 \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_{i,j}^2.$$

Therefore,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} [\hat{r}(z_{i,j}) - r^*(z_{i,j})]^2 \leq 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_{i,j} [\hat{r}(z_{i,j}) - r^*(z_{i,j})] \leq 4 \sup_{\|r\|_k \leq 1} \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_{i,j} r(z_{i,j}).$$

Notice that

$$\sup_{\|r\|_k \leq 1} \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_{i,j} r(z_{i,j}) = \sup_{\|r\|_k \leq 1} \langle r, \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_{i,j} k(\cdot, z_{i,j}) \rangle_k$$

$$= \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{i'=1}^{n} \sum_{j'=1}^{n} \epsilon_{i,j} \epsilon_{i',j'} k(z_{i,j}, z_{i',j'})}.$$

In other words, there exists a positive semi-definite matrix $K \in \mathbb{R}^{n^2 \times n^2}$ whose diagonal entries are no larger than 1 and a random vector $\tilde{\epsilon} \in \mathbb{R}^{n^2}$ whose entries are $\{\epsilon_{i,j}\}_{1 \leq i \leq n, 1 \leq j \leq n}$ such that

$$\sup_{\|r\|_k \leq 1} \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_{i,j} r(z_{i,j}) = \sqrt{(\tilde{\epsilon}^{\mathrm{T}}) K \tilde{\epsilon}}.$$

Let $K = J^{\mathrm{T}} J$, then we have $\|J\|_S \leq \|J\|_F = \sqrt{\mathrm{trace}(K)} \leq n$, where $\|\cdot\|_S$ and $\|\cdot\|_F$ denote the spectral norm and Frobenious norm, respectively, and

$$\sup_{\|r\|_k \leq 1} \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_{i,j} r(z_{i,j}) = \|J\tilde{\epsilon}\|_2.$$

By the concentration of anisotropic random vectors (see, e.g., [36, Theorem 6.3.2]), we can conclude that $\sup_{\|r\|_k \le 1} \sum_{i=1}^n \sum_{j=1}^n \epsilon_{i,j} r(z_{i,j})$, conditional on $\{z_{i,j}\}_{1 \le i \le n, 1 \le j \le n}$, is Cn-subgaussian. Hence, with probability at least $1 - p$, we have

$$\sup_{\|r\|_k \le 1} \sum_{i=1}^n \sum_{j=1}^n \epsilon_{i,j} r(z_{i,j}) \le Cn\sqrt{1 + \log(1/p)}.$$

Therefore, with probability at least $1 - p$,

$$\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n [\hat{r}(z_{i,j}) - r^*(z_{i,j})]^2 \le \frac{C}{n}\sqrt{1 + \log(1/p)}. \tag{5.3}$$

Noticing $\|\hat{r} - r^*\|_k \le \|\hat{r}\|_k + \|r^*\|_k \le 2$, we next define $\bar{\nu}_n = \frac{1}{n^2} \sum_{i=1}^n \delta_{z_{i,j}}$ and estimate

$$\sup_{\|r\|_k \le 2, \|r\|_{L^2(\bar{\nu}_n)} \le C\sqrt{\frac{1}{n}[1+\log(1/p)]}} \|r\|_{L^2(\bar{\nu})}$$

to prove the result. Given $\epsilon > 0$, we have

$$\sup_{\|r\|_k \le 1, \|r\|_{L^2(\bar{\nu}_n)} \le \epsilon} \|r\|_{L^2(\bar{\nu})}$$

$$= \sup_{\|r\|_k \le 1, \|r\|_{L^2(\bar{\nu}_n)} \le \epsilon} \sup_{\|g\|_{L^2(\bar{\nu})} \le 1} \int_{\mathcal{S} \times \mathcal{A}} g(z) r(z) \, \mathrm{d}\bar{\nu}(z)$$

$$= \sup_{\|g\|_{L^2(\bar{\nu})} \le 1} \sup_{\|r\|_k \le 1, \|r\|_{L^2(\bar{\nu}_n)} \le \epsilon} \int_{\mathcal{S} \times \mathcal{A}} g(z) r(z) \, \mathrm{d}\bar{\nu}(z)$$

$$= \sup_{\|g\|_{L^2(\bar{\nu})} \le 1} \inf_{g' \in L^2(\bar{\nu}_n)} \left[\mathrm{MMD}_k(g \circ \bar{\nu}, g' \circ \bar{\nu}_n) + \epsilon \|g'\|_{L^2(\bar{\nu}_n)} \right]$$

$$= \sup_{\|g\|_{L^2(\bar{\nu})} \le 1} \inf_{c_{1,1}, \cdots, c_{n,n}} \left[\mathrm{MMD}_k\left(g \circ \bar{\nu}, \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n c_{i,j} \delta_{z_{i,j}}\right) + \epsilon \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n c_{i,j}^2} \right]$$

$$\le \sup_{\|g\|_{L^2(\bar{\nu})} \le 1} \inf_{c_{1,1}, \cdots, c_{n,n}} \left[\frac{1}{n} \sum_{j=1}^n \mathrm{MMD}_k\left(g \circ \nu_j, \frac{1}{n} \sum_{i=1}^n c_{i,j} \delta_{z_{i,j}}\right) + \epsilon \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n c_{i,j}^2} \right]$$

$$\le 2\left\{ \sup_{\|g\|_{L^2(\bar{\nu})} \le 1} \inf_{c_{1,1}, \cdots, c_{n,n}} \left[\frac{1}{n} \sum_{j=1}^n \mathrm{MMD}_k^2\left(g \circ \nu_j, \frac{1}{n} \sum_{i=1}^n c_{i,j} \delta_{z_{i,j}}\right) + \frac{\epsilon^2}{n^2} \sum_{i=1}^n \sum_{j=1}^n c_{i,j}^2 \right] \right\}^{\frac{1}{2}}$$

$$= 2\left\{ \sup_{\sum_{j=1}^n m_j^2 \le n} \frac{1}{n} \sum_{j=1}^n \sup_{\|g\|_{L^2(\nu_j)} \le m_j} \inf_{c_1, \cdots, c_n} \left[\mathrm{MMD}_k^2\left(g \circ \nu_j, \frac{1}{n} \sum_{i=1}^n c_i \delta_{z_{i,j}}\right) + \frac{\epsilon^2}{n} \sum_{i=1}^n c_i^2 \right] \right\}^{\frac{1}{2}}. \tag{5.4}$$

In the third equality above, we have used Eq. (6.2) (when $\mathcal{B}$ is an RKHS with kernel k) that

is proved in Lemma 6.1 below. For $1 \le j \le n$, we then estimate

$$\sup_{\|g\|_{L^2(\nu_j)} \le m_j} \inf_{c_1, \cdots, c_n} \left[\mathrm{MMD}_k^2 \left(g \circ \nu_j, \frac{1}{n} \sum_{i=1}^n c_i \delta_{z_{i,j}} \right) + \frac{\epsilon^2}{n} \sum_{i=1}^n c_i^2 \right].$$

We first find a probability distribution λ on $\mathbb{N}^+$ and a measurable mapping $\phi : \mathbb{N}^+ \times (\mathcal{S} \times \mathcal{A})$ such that

$$k(z, z') = \mathbb{E}_{\omega \sim \lambda} \phi(\omega, z) \phi(\omega, z'),$$

which can be achieved using Mercer decomposition (see, e.g., [30, Section 2.3]). Then,

$$\mathrm{MMD}_k^2 \left(g \circ \nu_j, \frac{1}{n} \sum_{i=1}^n c_i \delta_{z_{i,j}} \right) = \mathbb{E}_{\omega \sim \lambda} \left[\int_{\mathcal{S} \times \mathcal{A}} g(z) \phi(\omega, z) \, \mathrm{d}\nu_j(z) - \frac{1}{n} \sum_{i=1}^n c_i \phi(\omega, z_{i,j}) \right]^2.$$

Recalling that $z_{1,j}, \cdots, z_{n,j}$ are i.i.d. drawn from ν_j, we can use Proposition 1 in [30] to obtain that with probability at least $1 - p$

$$\sup_{\|g\|_{L^2(\nu_j)} \le 1} \inf_{\sum_{i=1}^n c_i^2 \le 4n} \mathbb{E}_{\omega \sim \lambda} \left[\int_{\mathcal{S} \times \mathcal{A}} g(z) \phi(\omega, z) \, \mathrm{d}\nu_j(z) - \frac{1}{n} \sum_{i=1}^n c_i \phi(\omega, z_{i,j}) \right]^2 \le 4t,$$

where t satisfies that

$$5d(t) \log \left(\frac{16 d(t)}{p} \right) = n, \tag{5.5}$$

and

$$\begin{aligned} d(t) &= \sup_{z \in \mathcal{S} \times \mathcal{A}} \langle \phi(\cdot, z), (\Sigma + tI)^{-1} \phi(\cdot, z) \rangle_{L^2(\lambda)} \\ &\le t^{-1} \sup_{z \in \mathcal{S} \times \mathcal{A}} \langle \phi(\cdot, z), \phi(\cdot, z) \rangle_{L^2(\lambda)} \\ &= t^{-1} \sup_{z \in \mathcal{S} \times \mathcal{A}} k(z, z) \le t^{-1}, \end{aligned} \tag{5.6}$$

in which Σ is a self-adjoint, positive semi-definite operator on $L^2(\lambda)$ (see the detailed definition in [30, Section 2.1]). Therefore, from (5.5), (5.6), we have

$$\sup_{\|g\|_{L^2(\nu_j)} \le 1} \inf_{\sum_{i=1}^n c_i^2 \le 4n} \mathbb{E}_{\omega \sim \lambda} \left[\int_{\mathcal{S} \times \mathcal{A}} g(z) \phi(\omega, z) \, \mathrm{d}\nu_j(z) - \frac{1}{n} \sum_{i=1}^n c_i \phi(\omega, z_{i,j}) \right]^2$$

$$\le \frac{4}{d(t)} \le \frac{C}{n} \left[1 + \log \left(\frac{n}{p} \right) \right],$$

which means that

$$\sup_{\|g\|_{L^2(\nu_j)} \le m_j} \inf_{c_1, \cdots, c_n} \left[\mathrm{MMD}_k^2 \left(g \circ \nu_j, \frac{1}{n} \sum_{i=1}^n c_i \delta_{z_{i,j}} \right) + \frac{\epsilon^2}{n} \sum_{i=1}^n c_i^2 \right] \le C m_j^2 \frac{1}{n} \left[1 + \log \left(\frac{n}{p} \right) \right] + C m_j^2 \epsilon^2.$$

Combining the last inequality with (5.4), we can obtain that

$$\sup_{\|r\|_k \leq 1, \|r\|_{L^2(\bar{v}_n)} \leq \epsilon} \|r\|_{L^2(\bar{v})} \leq C\sqrt{\frac{1}{n}\left[1 + \log\left(\frac{n}{p}\right)\right]} + C\epsilon. \tag{5.7}$$

Recalling inequality (5.3) and choosing $\epsilon = C\sqrt{\frac{1}{n}[1 + \log(1/p)]}$ in (5.7), we know that with probability at least $1 - p$

$$\sup_{\|r\|_k \leq 2, \|r\|_{L^2(\bar{v}_n)} \leq C\sqrt{\frac{1}{n}[1+\log(1/p)]}} \|r\|_{L^2(\bar{v})} \leq C\sqrt{\frac{1}{n}\left[1 + \log\left(\frac{n}{p}\right)\right]},$$

which concludes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ □

Proof of Theorem 5.1. Using Lemma 5.1 and the union bound, we know that for any $\theta \in \Theta$, with probability at least $1 - p$,

$$\|\hat{r}_\theta(h, \cdot) - r_\theta(h, \cdot)\|_{L^2(\hat{v})} \leq C\sqrt{\frac{1}{n}\left[1 + \log\left(\frac{nH}{p}\right)\right]}, \quad \forall h \in [H].$$

Recalling the definition of $\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})$ in Definition 4.1 in the case of known transition

$$\Delta_{\mathcal{M}}(n^{-\frac{1}{2}}) = \sup_{h \in [H]} \sup_{\|r\|_k \leq 1, \sqrt{n}\|r\|_{L^2(\hat{v})} \leq 1} \|r\|_{\Pi(h, P_0, \mu)},$$

we know that for any $\pi \in \mathcal{P}(\mathcal{A} \,|\, \mathcal{S}, H)$ and $h \in [H]$,

$$\left| \int_{\mathcal{S} \times \mathcal{A}} r_\theta(h, s, a) \, \mathrm{d}\rho_{h, P_0, \pi, \mu}(s, a) - \int_{\mathcal{S} \times \mathcal{A}} \hat{r}_\theta(s, a) \, \mathrm{d}\rho_{h, P_0, \pi, \mu}(h, s, a) \right|$$

$$\leq C\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})\sqrt{1 + \log\left(\frac{nH}{p}\right)},$$

which means that for any $\pi \in \mathcal{P}(\mathcal{A} \,|\, \mathcal{S}, H)$

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

Therefore,

$$0 \leq J^*(M_\theta) - J(M_\theta, \hat{\pi}_\theta) = J^*(M_\theta) - J^*(\hat{M}_\theta) + J^*(\hat{M}_\theta) - J(M_\theta, \hat{\pi}_\theta)$$

$$= \sup_{\pi \in \mathcal{P}(\mathcal{A} \,|\, \mathcal{S})} J(M_\theta, \pi) - \sup_{\pi \in \mathcal{P}(\mathcal{A} \,|\, \mathcal{S})} J(\hat{M}_\theta, \pi) + J(\hat{M}_\theta, \hat{\pi}_\theta) - J(M_\theta, \hat{\pi}_\theta)$$

$$\leq CH\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})\sqrt{1 + \log\left(\frac{nH}{p}\right)}.$$

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ □

Remark 5.3. We remark that equivalence exists between the RL problem with known transition and the supervised learning problem with respect to the Π-norm in the RKHS. On the one hand, the above proof shows that, for any function g in the unit ball of the RKHS and any $h \in [H]$, once we can use finite samples to obtain an estimation $\hat{g}$ which is accurate with respect to $\| \cdot \|_{\Pi(h,P_0,\mu)}$, then the corresponding RL problem can be solved efficiently. On the other hand, Lemma 5.1 shows that one can always use finite samples to efficiently obtain an accurate estimation, in the L^2 sense, of a target function lying in the unit ball of the RKHS, and Theorem 4.1 tells that once the RL problem can be solved efficiently, $\Delta_{\mathcal{M}}(\epsilon)$ must decay fast with respect to ϵ. Therefore, we conclude that if the RL problem can be solved efficiently, for any $h \in [H]$, one can always obtain an estimation $\hat{g}$ of a target function g in the unit ball of the RKHS which is accurate with respect to $\| \cdot \|_{\Pi(h,P_0,\mu)}$.

5.2 The case of unknown transition

Similar to the case of known transition in the previous subsection, we consider

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

an intrinsic property of the MDP family $\mathcal{M}$ in the case of unknown transition. Again, similar argument also holds if $\hat{\zeta}$ is not the minimizer but

$$\sup_{\theta \in \Theta} \mathcal{R}(\Pi(P_\theta,\mu), \mathcal{H}_k, n^{-\frac{1}{2}}, \nu^{\theta,\hat{\zeta}})$$

is small. Given $\hat{\zeta}$, we sample

$$(\hat{z}_{1,1}^\theta, \cdots, \hat{z}_{1,n}^\theta), \cdots, (\hat{z}_{n,1}^\theta, \cdots, \hat{z}_{n,n}^\theta)$$

as i.i.d. copies of $((s_1^{\theta,\hat{\zeta}}, a_1^{\theta,\hat{\zeta}}), \cdots, (s_n^{\theta,\hat{\zeta}}, a_n^{\theta,\hat{\zeta}}))$ defined in (4.2). We then use $\{\hat{z}_{i,j}\}_{1 \le i \le n, 1 \le j \le n}$ to perform the following variant of fitted Q-iteration algorithm (see, e.g., [17, 21, 24]), as summarized in Algorithm 3. To understand the idea of Algorithm 3/fitted Q-iteration, we define the optimal action-value function (Q-value function) $Q_h^* : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ as the optimal expected cumulative reward of the MDP starting from step h:

$$Q_h^{\theta,*}(s,a) = \sup_{\pi \in \mathcal{P}(\mathcal{A} \,|\, \mathcal{S}, H)} \mathbb{E}_{P_\theta,\pi}\left[\sum_{h'=h}^{H} r_\theta(h', S_{h'}, A_{h'}) \,|\, S_h = s, \, A_h = a \right]. \tag{5.8}$$

Here the expectation is taken among the MDP paths generated by transition probability P_θ and policy π. Let $\pi^{\theta,*} = (\pi_h^{\theta,*})_{h \in [H]}$ be the greedy policy with respect to $Q^{\theta,*} = (Q_h^{\theta,*})_{h \in [H]}$:

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

for any $s \in \mathcal{S}$ and $h \in [H]$. Here $\text{supp}(\pi_h^{\theta,*}(\,\cdot\,|\,s))$ denotes the support of $\pi_h^{\theta,*}(\,\cdot\,|\,s)$. We can then conclude that $\pi^{\theta,*}$ is the optimal policy of M_θ (see, e.g. [37, Theorem 4.5.1]):

$$J(M_\theta, \pi^{\theta,*}) = \sup_{\pi \in \mathcal{P}(\mathcal{A}\,|\,\mathcal{S},H)} J(M_\theta, \pi).$$

We define the Bellman optimal operator $\mathcal{T}_h^\theta : \mathcal{H}_k \mapsto \mathcal{H}_k$ as follows:

$$(\mathcal{T}_h^\theta g)(s,a) = r_\theta(h,s,a) + \mathbb{E}_{s' \sim P_\theta(\,\cdot\,|\,h,s,a)}[\max_{a' \in \mathcal{A}} g(s',a')]. \tag{5.9}$$

Then the famous Bellman equations gives

$$Q_h^{\theta,*} = \mathcal{T}_h^\theta Q_{h+1}^{\theta,*}, \quad \forall h \in [H], \quad Q_{H+1}^{\theta,*} = 0. \tag{5.10}$$

We assume that for any $g \in \mathcal{H}_k$ and $\theta \in \Theta$,

$$\|\mathcal{T}_h^\theta g\|_k \le \|g\|_k + 1, \tag{5.11}$$

which implies that

$$\|Q_h^{\theta,*}\|_k \le H - h + 1. \tag{5.12}$$

Finally Eqs. (5.10), (5.12) motivate us to solve the optimization problem (5.13) in Algorithm 3. Note that (5.13) can be solved as a finite-dimensional convex optimization problem in the same way as (5.1); see the comment before Algorithm 2.

Algorithm 3 Fitted Q-Iteration Algorithm

Input: n^2 samples $(\hat{z}_{1,1}^\theta, \cdots, \hat{z}_{1,n}^\theta), \cdots, (\hat{z}_{n,1}^\theta, \cdots, \hat{z}_{n,n}^\theta)$ as i.i.d. copies of $((s_1^{\theta,\hat{\xi}}, a_1^{\theta,\hat{\xi}}), \cdots, (s_n^{\theta,\hat{\xi}}, a_n^{\theta,\hat{\xi}}))$ defined in (4.2).
Initialize: $Q_{H+1}^\theta(s,a) = 0$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.
for $h = H, H-1, \cdots, 1$ **do**
    for $i = 1, \cdots, n$ and $j = 1, \cdots, n$ **do**
        Sample $r_{i,j}^\theta \sim \mathcal{N}(r_\theta(h, \hat{z}_{i,j}^\theta), 1)$ and $s_{i,j}^{\theta,\prime} \sim P_\theta(\,\cdot\,|\,h, \hat{z}_{i,j}^\theta)$
        Compute $y_{i,j}^\theta = r_{i,j}^\theta + \max_{a' \in \mathcal{A}} Q_{h+1}(s_{i,j}^{\theta,\prime}, a')$
    end
    Compute Q_h^θ as the minimizer of the optimization problem

$$\min_{\|g\|_k \le H-h+1} \sum_{i=1}^n \sum_{j=1}^n [g(\hat{z}_{i,j}^\theta) - y_{i,j}^\theta]^2 \tag{5.13}$$

end
Output: $\hat{\pi}^\theta$ as the greedy policies with respect to $(Q_h^\theta)_{h \in [H]}$.

We have the following convergence result regarding Algorithm 3.

Theorem 5.2. *Assume that* (5.11) *holds and*

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

Then, for any $\theta \in \Theta$, with probability at least $1 - p$,

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

Remark 5.4. Assumption (5.11) is used to control the approximation error in fitted Q-iteration. Similar assumptions are made in [15–17, 21, 24, 38]. By choosing $g = 0$ in (5.11), we can see that (5.11) is stronger than the setting in Theorem 4.2, which only requires $\|r_\theta(h, \cdot)\|_k \leq 1$ for any $\theta \in \Theta$ and $h \in [H]$. How to fill out this gap is left for future work.

Proof. For each time step $h \in [H]$, by Eq. (5.13), we have $\|Q_h^\theta\|_k \leq H - h + 1$. With the optimal Bellman operator (5.9), we have

$$y_{i,j}^\theta = r_\theta(h, \hat{z}_{i,j}^\theta) + \epsilon_{i,j} + \max_{a' \in \mathcal{A}} Q_{h+1}^\theta(s_{i,j}^{\theta,'}, a')$$

$$= (\mathcal{T}_h^\theta Q_{h+1}^\theta)(\hat{z}_{i,j}^\theta) + (\max_{a' \in \mathcal{A}} Q_{h+1}^\theta(s_{i,j}^{\theta,'}, a') - \mathbb{E}_{s' \sim P_\theta(\cdot \mid h, \hat{z}_{i,j}^\theta)}[\max_{a' \in \mathcal{A}} Q_{h+1}^\theta(s_{i,j}^{\theta,'}, a')]) + \epsilon_{i,j},$$

where $\epsilon_{i,j}$ are standard normal random variables. With the boundedness of Q

$$\sup_{z \in \mathcal{S} \times \mathcal{A}} |Q_{h+1}^\theta(z)| = \sup_{z \in \mathcal{S} \times \mathcal{A}} |\langle Q_{h+1}^\theta, k(\cdot, z) \rangle_k|$$

$$\leq (H - h) \sup_{z \in \mathcal{S} \times \mathcal{A}} \|k(\cdot, z)\|_k = (H - h) \sup_{z \in \mathcal{S} \times \mathcal{A}} \sqrt{k(z, z)} \leq H - h,$$

we know that $y_{i,j}^\theta - (\mathcal{T}_h^\theta Q_{h+1}^\theta)(\hat{z}_{i,j}^\theta)$ is CH-subgaussian; see, e.g., [36, Example 2.5.8]. With the assumption (5.11), we know $\|\mathcal{T}_h^\theta Q_{h+1}^\theta\|_k \leq H - h + 1$. Hence, we can use Lemma 5.1 and the union bound to obtain that with probability at least $1 - p$, for any $h \in [H]$,

$$\|Q_h^\theta - \mathcal{T}_h^\theta Q_{h+1}^\theta\|_{L^2(\nu^{\theta, \hat{\xi}})} \leq CH \sqrt{\frac{1}{n}\left[1 + \log\left(\frac{nH}{p}\right)\right]}.$$

Noticing that $\|Q_h^\theta - \mathcal{T}_h^\theta Q_{h+1}^\theta\|_k \leq 2H$ and the definition of $\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})$ in (4.3), we have that, with probability at least $1 - p$, for any $h \in [H]$,

$$\sup_{\pi \in \mathcal{P}(\mathcal{A} \mid \mathcal{S}, H)} \left| \int_{\mathcal{S} \times \mathcal{A}} [Q_h^\theta(s, a) - (\mathcal{T}_h^\theta Q_{h+1}^\theta)(s, a)] \, d\rho_{h, P_\theta, \pi, \mu}(s, a) \right|$$

$$\leq CH \Delta_{\mathcal{M}}(n^{-\frac{1}{2}}) \sqrt{1 + \log\left(\frac{nH}{p}\right)}. \tag{5.14}$$

Using the optimal Bellman equation (5.10), we have:

$$\left| \int_{\mathcal{S}\times\mathcal{A}} [Q_h^\theta - Q_h^{\theta,*}](s,a)\, \mathrm{d}\rho_{h,P_\theta,\pi,\mu}(s,a) \right|$$

$$\leq \left| \int_{\mathcal{S}\times\mathcal{A}} [Q_h^\theta - (\mathcal{T}_h^\theta Q_{h+1}^\theta)](s,a)\, \mathrm{d}\rho_{h,P_\theta,\pi,\mu}(s,a) \right|$$

$$+ \left| \int_{\mathcal{S}\times\mathcal{A}} [(\mathcal{T}_h^\theta Q_{h+1}^{\theta,*}) - (\mathcal{T}_h^\theta Q_{h+1}^\theta)](s,a)\, \mathrm{d}\rho_{h,P_\theta,\pi,\mu}(s,a) \right|. \qquad (5.15)$$

Notice that

$$\left| \int_{\mathcal{S}\times\mathcal{A}} [(\mathcal{T}_h^\theta Q_{h+1}^{\theta,*}) - (\mathcal{T}_h^\theta Q_{h+1}^\theta)](s,a)\, \mathrm{d}\rho_{h,P_\theta,\pi,\mu}(s,a) \right|$$

$$\leq \left| \int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}_{s'\sim P_\theta(\,\cdot\,|\,h,s,a)} [\max_{a'\in\mathcal{A}} Q_{h+1}^{\theta,*} - \max_{a'\in\mathcal{A}} Q_{h+1}^\theta](s',a')\, \mathrm{d}\rho_{h,P_\theta,\pi,\mu}(s,a) \right|$$

$$\leq \max \left\{ \left| \int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}_{s'\sim P_\theta(\,\cdot\,|\,h,s,a)} \max_{a'\in\mathcal{A}} [Q_{h+1}^{\theta,*} - Q_{h+1}^\theta](s',a')\, \mathrm{d}\rho_{h,P_\theta,\pi,\mu}(s,a) \right|, \right.$$

$$\left. \left| \int_{\mathcal{S}\times\mathcal{A}} \mathbb{E}_{s'\sim P_\theta(\,\cdot\,|\,h,s,a)} \max_{a'\in\mathcal{A}} [Q_{h+1}^\theta - Q_{h+1}^{\theta,*}](s',a')\, \mathrm{d}\rho_{h,P_\theta,\pi,\mu}(s,a) \right| \right\}$$

$$\leq \sup_{\pi\in\mathcal{P}(\mathcal{A}\,|\,\mathcal{S},H)} \left| \int_{\mathcal{S}\times\mathcal{A}} [Q_{h+1}^\theta - Q_{h+1}^{\theta,*}](s,a)\, \mathrm{d}\rho_{h+1,P_\theta,\pi,\mu}(s,a) \right|,$$

where the last inequality holds because we can choose policies π^1 and π^2 such that $\pi_{h'}^1 = \pi_{h'}^2 = \pi$ for $h' \neq h+1$ and π_{h+1}^1 and π_{h+1}^2 are the greedy policies of $[Q_{h+1}^{\theta,*} - Q_{h+1}^\theta](s,a)$ and $[Q_{h+1}^\theta - Q_{h+1}^{\theta,*}](s,a)$, respectively. Combining the last inequality, inequalities (5.14) and (5.15), we have that with probability $1-p$, $\forall h \in [H]$

$$\sup_{\pi\in\mathcal{P}(\mathcal{A}|\mathcal{S},H)} \left| \int_{\mathcal{S}\times\mathcal{A}} [Q_h^\theta - Q_h^{\theta,*}](s,a)\, \mathrm{d}\rho_{h,P_\theta,\pi,\mu}(s,a) \right|$$

$$\leq CH\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})\sqrt{1+\log\left(\frac{nH}{p}\right)} + \sup_{\pi\in\mathcal{P}(\mathcal{A}\,|\,\mathcal{S},H)} \left| \int_{\mathcal{S}\times\mathcal{A}} [Q_{h+1}^\theta - Q_{h+1}^{\theta,*}](s,a)\, \mathrm{d}\rho_{h+1,P_\theta,\pi,\mu}(s,a) \right|.$$

With the recursive relationship above, we have that, with probability at least $1-p$,

$$\sup_{h\in[H]} \sup_{\pi\in\mathcal{P}(\mathcal{A}\,|\,\mathcal{S},H)} \left| \int_{\mathcal{S}\times\mathcal{A}} [Q_h^\theta - Q_h^{\theta,*}](s,a)\, \mathrm{d}\rho_{h,P_\theta,\pi,\mu}(s,a) \right|$$

$$\leq CH^2\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})\sqrt{1+\log\left(\frac{nH}{p}\right)}.$$

We can then use the famous performance difference lemma (see, e.g., [39, Lemma 3.2]

or [22, Lemma 6.1]) to obtain that

$$0 \leq J^*(M_\theta) - J(M_\theta, \hat{\pi}^\theta)$$

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

$$= \sum_{h=1}^{H} \int_{\mathcal{S} \times \mathcal{A}} \sum_{a' \in \mathcal{A}} \{ [Q_h^{\theta,*} - Q_h^\theta](s, a') \pi_h^{\theta,*}(a' \mid s) + Q_h^\theta(s, a')[\pi_h^{\theta,*} - \hat{\pi}_h^\theta](a' \mid s)$$

$$+ [Q_h^\theta - Q_h^{\theta,*}](s, a') \hat{\pi}_h^\theta(a' \mid s) \} \, \mathrm{d}\rho_{h, P_\theta, \hat{\pi}^\theta, \mu}(s, a). \tag{5.16}$$

Noticing that $\sum_{a' \in \mathcal{A}} Q_h^\theta(s, a')[\pi_h^{\theta,*} - \hat{\pi}_h^\theta](a'|s) \leq 0$ since $\hat{\pi}^\theta$ is the greedy policy with respect to $Q_{h'}^\theta$ we can conclude that, with probability at least $1 - p$,

$$|J^*(M_\theta) - J(M_\theta, \hat{\pi}^\theta)| \leq CH \sup_{h \in [H]} \sup_{\pi \in \mathcal{P}(\mathcal{A} \mid \mathcal{S}, H)} \left| \int_{\mathcal{S} \times \mathcal{A}} [Q_h^\theta - Q_h^{\theta,*}](s, a) \, \mathrm{d}\rho_{h, P_\theta, \pi, \mu}(s, a) \right|$$

$$\leq CH^3 \Delta_{\mathcal{M}}(n^{-\frac{1}{2}}) \sqrt{1 + \log\left(\frac{nH}{p}\right)}.$$

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

6 Discussion on perturbational complexity by distribution mismatch

In this section, we discuss in more details the perturbation response $\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu)$ and the perturbational complexity $\Delta_{\mathcal{M}}(\epsilon)$. We first give a more concrete expression of $\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu)$.

Proposition 6.1. *We have*

$$\mathcal{R}(\Pi, \mathcal{B}, \epsilon, \nu) = \sup_{\rho \in \Pi} \inf_{g \in L^2(\nu)} [\|\rho - g \circ \nu\|_{\mathcal{B}^*} + \epsilon \|g\|_{L^2(\nu)}],$$

where $g \circ \nu$ is a signed measure such that

$$\frac{\mathrm{d}g \circ \nu}{\mathrm{d}\nu} = g,$$

$\mathcal{B}^*$ is the dual space of $\mathcal{B}$ and $\|\rho\|_{\mathcal{B}^*}$ is the dual norm of linear functional

$$\rho(g) := \int_{\mathcal{S} \times \mathcal{A}} g(z) \, \mathrm{d}\rho(z), \quad \forall g \in \mathcal{B},$$

for any signed measure ρ on $\mathcal{S} \times \mathcal{A}$ (we slightly abuse the notation that ρ are both the signed measure and linear functional in $\mathcal{B}$). If $\mathcal{B}$ is an RKHS with kernel k, then

$$\mathcal{R}(\Pi, \mathcal{H}_k, \epsilon, \nu) = \sup_{\rho \in \Pi} \inf_{g \in L^2(\nu)} [\mathrm{MMD}_k(\rho, g \circ \nu) + \epsilon \|g\|_{L^2(\nu)}]. \tag{6.1}$$

Proof. It is sufficient to prove that

$$\sup_{\|g\|_{\mathcal{B}}\leq 1, \|g\|_{L^2(\nu)}\leq \epsilon} \int_{\mathcal{S}\times\mathcal{A}} g(z)\,\mathrm{d}\rho(z) = \inf_{g\in L^2(\nu)} [\|\rho - g\circ\nu\|_{\mathcal{B}^*} + \epsilon\|g\|_{L^2(\nu)}]. \tag{6.2}$$

Eq. (6.1) can be derived using the definition of maximum mean discrepancy.

Define $F_1, F_2\colon \mathcal{B}\mapsto(-\infty,+\infty]$:

$$F_1(g) = \begin{cases} \int_{\mathcal{S}\times\mathcal{A}} g(z)\,\mathrm{d}\rho(z) & \text{if } \|g\|_{\mathcal{B}}\leq 1, \\ +\infty & \text{if } \|g\|_{\mathcal{B}}> 1, \end{cases} \quad F_2(g) = \begin{cases} 0 & \text{if } \|g\|_{L^2(\nu)}\leq\epsilon, \\ +\infty & \text{if } \|g\|_{L^2(\nu)}>\epsilon. \end{cases}$$

Then F_1 and F_2 are both convex functions in $\mathcal{B}$. We can define the conjugate functions F_1^* and $F_2^*\colon \mathcal{B}^*\mapsto(-\infty,+\infty]$:

$$F_1^*(l) = \sup_{g\in\mathcal{B}}[l(g) - F_1(g)] = \sup_{\|g\|_{\mathcal{B}}\leq 1}[l(g) - \rho(g)] = \|l-\rho\|_{\mathcal{B}^*},$$

$$F_2^*(l) = \sup_{g\in\mathcal{B}}[l(g) - F_2(g)] = \sup_{g\in\mathcal{B},\|g\|_{L^2(\nu)}\leq\epsilon} l(g) = \epsilon \sup_{g\in\mathcal{B},\|g\|_{L^2(\nu)}\leq 1} l(g).$$

We then compute

$$\sup_{g\in\mathcal{B},\|g\|_{L^2(\nu)}\leq 1} l(g).$$

Let $l\in\mathcal{B}^*$ such that

$$M_l = \sup_{g\in\mathcal{B},\|g\|_{L^2(\nu)}\leq 1} l(g) < +\infty.$$

Notice that l is a linear mapping in $\mathcal{B}$ such that

$$|l(g)| \leq M_l\|g\|_{L^2(\nu)}, \quad \forall g\in\mathcal{B}.$$

Using Hahn-Banach theorem [40, Corollary 1.2] and Riesz representation theorem in $L^2(\nu)$ [40, Theorem 4.11], we know that there exists $R_l\in L^2(\nu)$ such that

$$l(g) = \int_{\mathcal{S}\times\mathcal{A}} R_l(z)g(z)\,\mathrm{d}\nu(z), \quad \forall g\in\mathcal{B}. \tag{6.3}$$

Hence,

$$\sup_{g\in\mathcal{B},\|g\|_{L^2(\nu)}\leq 1} l(g) = \|P_{\mathcal{B}}R_l\|_{L^2(\nu)},$$

where $P_{\mathcal{B}}$ is the orthogonal projection from $L^2(\nu)$ to $\mathcal{B}$. Consequently,

$$F_2^*(l) = \begin{cases} \epsilon\|P_{\mathcal{B}}R_l\|_{L^2(\nu)} & \text{if exists } R_l\in L^2(\nu) \text{ s.t. Eq. (6.3) holds,} \\ +\infty & \text{otherwise.} \end{cases}$$

Noticing that F_1 is continuous at 0, we can use Fenchel-Rockafellar Theorem [40, Theorem 1.12] to obtain that

$$\inf_{r\in\mathcal{B}}[F_1(r) + F_2(r)] = -\inf_{l\in\mathcal{B}^*}[F_1^*(l) + F_2^*(-l)],$$

which means that

$$\sup_{\|g\|_{\mathcal{B}}\leq 1,\|g\|_{L^2(\nu)}\leq\epsilon}\int_{\mathcal{S}\times\mathcal{A}}g(z)\,\mathrm{d}\rho(z)=\inf_{g\in L^2(\nu)}[\|\rho-g\circ\nu\|_{\mathcal{B}^*}+\epsilon\|P_{\mathcal{B}}g\|_{L^2(\nu)}].$$

Finally, noticing that $\|\rho-g\circ\nu\|_{\mathcal{B}^*}=\|\rho-(P_{\mathcal{B}}g)\circ\nu\|_{\mathcal{B}^*}$, we obtain

$$
\begin{aligned}
\inf_{g\in L^2(\nu)}[\|\rho-g\circ\nu\|_{\mathcal{B}^*}+\epsilon\|P_{\mathcal{B}}g\|_{L^2(\nu)}]&=\inf_{g\in L^2(\nu)}[\|\rho-(P_{\mathcal{B}}g)\circ\nu\|_{\mathcal{B}^*}+\epsilon\|P_{\mathcal{B}}g\|_{L^2(\nu)}]\\
&=\inf_{g\in L^2(\nu)}[\|\rho-g\circ\nu\|_{\mathcal{B}^*}+\epsilon\|g\|_{L^2(\nu)}],
\end{aligned}
$$

which completes the proof. $\qquad\square$

In the following, we again only consider the case that $\mathcal{B}$ is an RKHS with kernel k. We first show that the finite concentration coefficients, considered in [13, 17, 21, 23–25, 41, 42], implies that the perturbation response must decay fast. So our results generalize the previous works based on the assumption of concentratability. Note that the original assumption on concentration coefficients is only stated for the case $p=2$, and the corresponding M is called concentration coefficients in previous works.

Proposition 6.2. *Assume that there exists $1<p\leq 2$ and a distribution ν such that*

$$M=\sup_{\rho\in\Pi}\Big\|\frac{\mathrm{d}\rho}{\mathrm{d}\nu}\Big\|_{L^p(\nu)}<+\infty,\tag{6.4}$$

and

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

Then,

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

Proof. For any $\rho\in\Pi$, define

$$g_\rho=\frac{\mathrm{d}\rho}{\mathrm{d}\nu},$$

and choose $K>0$, then

$$
\begin{aligned}
\inf_{g\in L^2(\nu)}&\left[\mathrm{MMD}_k(\rho,g\circ\nu)+\frac{\|g\|_{L^2(\nu)}}{\sqrt{n}}\right]\\
&\leq\mathrm{MMD}_k(g_\rho\circ\nu,(g_\rho 1_{|g_\rho|\leq K})\circ\nu))+\frac{\|g_\rho 1_{|g_\rho|\leq K}\|_{L^2(\nu)}}{\sqrt{n}}\\
&\leq\|g_\rho 1_{|g_\rho|>K}\|_{L^1(\nu)}+\frac{\|g_\rho 1_{|g_\rho|\leq K}\|_{L^2(\nu)}}{\sqrt{n}}\\
&=\|g_\rho^p\cdot g_\rho^{1-p}1_{|g_\rho|>K}\|_{L^1(\nu)}+\frac{\sqrt{\|g_\rho^p\cdot g_\rho^{2-p}1_{|g_\rho|\leq K}\|_{L^1(\nu)}}}{\sqrt{n}}
\end{aligned}
$$

$$\leq M^p K^{1-p} + \frac{M^{\frac{p}{2}} K^{1-\frac{p}{2}}}{\sqrt{n}}.$$

By choosing $K = Mn^{\frac{1}{p}}$, we know that

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

This completes the proof. □

Next, we give a sufficient condition such that $\Delta_{\mathcal{M}}(n^{-\frac{1}{2}})$ decays fast with respect to n in the case of unknown transition. The original idea is from [17, Section 3.3]. The condition is that there exists a distribution λ on $[H] \times \mathcal{S} \times \mathcal{A}$ such that the induced state-action distribution

$$\bar{\nu}^\theta(S \times A) = \int_{[H] \times \mathcal{S} \times \mathcal{A}} P_\theta(S \mid h, s, a) \, \mathrm{d}\lambda(h, s, a) \times \mathrm{Uniform}_{\mathcal{A}}(A),$$

$$\forall \text{ measurable set } S \subset \mathcal{S}, A \subset \mathcal{A}$$

satisfying that

$$\sup_{\theta \in \Theta} \mathcal{R}(\Pi(P_\theta, \mu), \mathcal{B}, n^{-\frac{1}{2}}, \bar{\nu}^\theta)$$

decays fast with respect n. This condition holds when

$$\sup_{\theta \in \Theta} \sup_{\rho \in \Pi(P_\theta, \mu)} \left\| \frac{\mathrm{d}\rho}{\mathrm{d}\bar{\nu}^\theta} \right\|_{L^2(\bar{\nu}^\theta)} < +\infty,$$

or when the eigenvalue decay of the kernel is fast; see the discussion below. Given this condition, we can choose a sampling algorithm $\bar{\xi} = (\tilde{f}_1, \cdots, \tilde{f}_{2n})$ in (4.2) satisfying that

$$\tilde{f}_i(\mathcal{D}_{i-1}^{\theta, \bar{\xi}}, \bar{u}) = \begin{cases} f_i(\bar{u}) & \text{when } 1 \leq i \leq n, \\ (1, x_{i-n}^{\theta, \bar{\xi}}, f_i(\bar{u})) & \text{when } n+1 \leq i \leq 2n, \end{cases}$$

where $f_1(\bar{u}), \cdots f_n(\bar{u})$ are i.i.d. random variables with distribution λ and $f_{n+1}(\bar{u}), \cdots, f_{2n}(\bar{u})$ are i.i.d. random variables with distribution $\mathrm{Uniform}_{\mathcal{A}}$. By construction, $\frac{1}{n} \sum_{i=n+1}^{2n} \mathcal{L}(z_i^{\theta, \bar{\xi}}) = \bar{\nu}^\theta$, and thus we know that

$$\Delta_{\mathcal{M}}((2n)^{-\frac{1}{2}}) \leq \sup_{\theta \in \Theta} \mathcal{R}(\Pi(P_\theta, \mu), \mathcal{H}_k, (2n)^{-\frac{1}{2}}, \nu^{\theta, \bar{\xi}})$$

$$\leq 2 \sup_{\theta \in \Theta} \mathcal{R}(\Pi(P_\theta, \mu), \mathcal{H}_k, n^{-\frac{1}{2}}, \bar{\nu}^\theta)$$

must decay fast with respect to n.

In the following, we establish the connection between the kernel's eigenvalues and perturbation response. The following proposition is the core of this connection.

Proposition 6.3. *Assume that*

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

For any $\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, define

$$n(\rho) = \max\{i \in \mathbb{N}^+ : n\Lambda_i^\rho \ge 1\}. \tag{6.5}$$

We have

$$\mathcal{R}(\mathcal{P}(\mathcal{S} \times \mathcal{A}), \mathcal{H}_k, n^{-\frac{1}{2}}, v) \ge \frac{1}{2} \sqrt{\sup_{\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A})} \sum_{i=n(v)+1}^{+\infty} \Lambda_i^\rho}, \tag{6.6}$$

and, by $n(v) \le n$,

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

Moreover, if there exists a distribution $\hat{v} \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ such that

$$\sup_{i \in \mathbb{N}^+} \|\psi_i^{\hat{v}}\|_\infty < +\infty, \tag{6.7}$$

then

$$\mathcal{R}(\mathcal{P}(\mathcal{S} \times \mathcal{A}), \mathcal{H}_k, n^{-\frac{1}{2}}, \hat{v}) \le 2\sqrt{\frac{n(\hat{v})}{n} + \sum_{i=n(\hat{v})+1}^{\infty} \Lambda_i^{\hat{v}} \sup_{i \in \mathbb{N}^+} \|\psi_i^{\hat{v}}\|_\infty}. \tag{6.8}$$

Remark 6.1. Assumption (6.7) is widely used in the literature; see, e.g., [16, 30, 43, 44]. However, note that even for C^∞-kernel, (6.7) does not always hold; see, e.g., [45]. To see a concrete example of lower bound (6.6) and upper bound (6.8), assume $\Lambda_i^{\hat{v}} = \Theta(i^{-\alpha})$ for $\alpha > 1$, and we have

$$\mathcal{R}(\mathcal{P}(\mathcal{S} \times \mathcal{A}), \mathcal{H}_k, n^{-\frac{1}{2}}, \hat{v}) = \Theta(n^{-\frac{\alpha-1}{2\alpha}}).$$

Proof. We first prove that $n(\rho) \le n$ for any $\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$. When $i \ge n+1$, using the Mercer decomposition (2.1) and the fact that $\{\psi_i^v\}_{i \in \mathbb{N}^+}$ is orthonormal in $L^2(v)$, we have,

$$n\Lambda_i^v \le \sum_{j=1}^n \Lambda_j^v \le \sum_{j=1}^{+\infty} \Lambda_j^v = \int_{\mathcal{S} \times \mathcal{A}} k(z,z)\,dv(z) \le 1.$$

By the definition (6.5), we know that $n(\rho) \le n$.

We will use the expression (6.1)

$$\mathcal{R}(\mathcal{P}(\mathcal{S} \times \mathcal{A}), \mathcal{H}_k, \epsilon, v) = \sup_{\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A})} \inf_{g \in L^2(v)} [\mathrm{MMD}_k(\rho, g \circ v) + \epsilon \|g\|_{L^2(v)}]$$

to prove our result. First, given $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ and $\rho_0 \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, we can use Mercer's decomposition (2.1) to obtain

$$\inf_{g \in L^2(\nu)} \left(\mathrm{MMD}(\rho_0, g \circ \nu) + \frac{\|g\|_{L^2(\nu)}}{\sqrt{n}} \right)^2$$

$$\geq \inf_{g \in L^2(\nu)} \left[\mathrm{MMD}^2(\rho_0, g \circ \nu) + \frac{\|g\|_{L^2(\nu)}^2}{n} \right]$$

$$= \inf_{g \in L^2(\nu)} \left[\int_{\mathcal{S} \times \mathcal{A}} \int_{\mathcal{S} \times \mathcal{A}} k(z, z') \, \mathrm{d}(\rho_0 - g \circ \nu)(z) \, \mathrm{d}(\rho_0 - g \circ \nu)(z') + \frac{\|g\|_{L^2(\nu)}^2}{n} \right]$$

$$= \inf_{g \in L^2(\nu)} \sum_{i=1}^{+\infty} \left[\Lambda_i^\nu \left(\int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) \, \mathrm{d}\rho_0(z) - \int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) g(z) \, \mathrm{d}\nu(z) \right)^2 \right.$$

$$\left. + \frac{1}{n} \left(\int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) g(z) \, \mathrm{d}\nu(z) \right)^2 \right]$$

$$= \inf_{\{g_i\}_{i \in \mathbb{N}^+}} \sum_{i=1}^{+\infty} \left[\Lambda_i^\nu \left(\int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) \, \mathrm{d}\rho_0(z) - g_i \right)^2 + \frac{1}{n} |g_i|^2 \right]$$

$$= \sum_{i=1}^{+\infty} \frac{\Lambda_i^\nu \left(\int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) \, \mathrm{d}\rho_0(z) \right)^2}{n \Lambda_i^\nu + 1}.$$

Similarly, we have

$$\inf_{g \in L^2(\hat{\nu})} \left(\mathrm{MMD}(\rho_0, g \circ \hat{\nu}) + \frac{\|g\|_{L^2(\hat{\nu})}}{\sqrt{n}} \right)^2 \leq 2 \sum_{i=1}^{+\infty} \frac{\Lambda_i^{\hat{\nu}} (\int_{\mathcal{S} \times \mathcal{A}} \psi_i^{\hat{\nu}}(z) \, \mathrm{d}\rho_0(z))^2}{n \Lambda_i^{\hat{\nu}} + 1}$$

$$\leq 2 \sum_{i=1}^{+\infty} \frac{\Lambda_i^{\hat{\nu}}}{n \Lambda_i^{\hat{\nu}} + 1} \sup_{i \in \mathbb{N}^+} \|\psi_i^{\hat{\nu}}\|_\infty^2.$$

Noticing that

$$\sum_{i=1}^{+\infty} \frac{\Lambda_i^{\hat{\nu}}}{n \Lambda_i^{\hat{\nu}} + 1} \leq 2 \left[\frac{n(\hat{\nu})}{n} + \sum_{i=n(\hat{\nu})+1}^{+\infty} \Lambda_i^{\hat{\nu}} \right],$$

we obtain inequality (6.8).

We then prove inequality (6.6). Given any $z_0 \in \mathcal{S} \times \mathcal{A}$, picking $\rho_0 = \delta_{z_0}$, we have

$$\inf_{g \in L^2(\nu)} \left[\mathrm{MMD}(\delta_{z_0}, g \circ \nu) + \frac{\|g\|_{L^2(\nu)}}{\sqrt{n}} \right]^2 \geq \sum_{i=1}^{+\infty} \frac{\Lambda_i^\nu (\psi_i^\nu(z_0))^2}{n \Lambda_i^\nu + 1}.$$

Therefore, for any $\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$, by taking average with respect to z_0, we have

$$\sup_{\nu' \in \mathcal{P}(\mathcal{S} \times \mathcal{A})} \inf_{g \in L^2(\nu)} \left[\mathrm{MMD}(\nu', g \circ \nu) + \frac{\|g\|_{L^2(\nu)}}{\sqrt{n}} \right]^2 \geq \sum_{i=1}^{+\infty} \frac{\Lambda_i^\nu \int |\psi_i^\nu(z)|^2 \, \mathrm{d}\rho(z)}{n \Lambda_i^\nu + 1}.$$

Noticing that $n\Lambda_i^\nu \leq 1$ when $i \geq n(\nu) + 1$ by definition 6.5, we have

$$\sup_{\nu' \in \mathcal{P}(\mathcal{S} \times \mathcal{A})} \inf_{g \in L^2(\nu)} \left[\text{MMD}(\nu', g \circ \nu) + \frac{\|g\|_{L^2(\nu)}}{\sqrt{n}} \right]^2 \geq \frac{1}{2} \sum_{i=n(\nu)+1}^{+\infty} \Lambda_i^\nu \int_{\mathcal{S} \times \mathcal{A}} |\psi_i^\nu(z)|^2 \, d\rho(z).$$

$$(6.9)$$

Again, using Mercer decomposition (2.1) and the eigenvalues $\{\Lambda_i^\rho\}_{i \in \mathbb{N}^+}$ corresponding to the operator $\mathcal{K}_\rho$, we have

$$\sum_{i=1}^{+\infty} \Lambda_i^\nu \int_{\mathcal{S} \times \mathcal{A}} |\psi_i^\nu(z)|^2 \, d\rho(z) = \int_{\mathcal{S} \times \mathcal{A}} k(z, z) \, d\rho(z) = \sum_{i=1}^{+\infty} \Lambda_i^\rho. \qquad (6.10)$$

Checking (6.9), (6.10) together, we need to have an upper bound of $\sum_{i=1}^{n(\nu)} \Lambda_i^\nu \int_{\mathcal{S} \times \mathcal{A}} |\psi_i^\nu(z)|^2 \, d\rho(z)$. Let

$$c_j = \sum_{i=1}^{n(\nu)} \Lambda_i^\nu \left(\int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) \psi_j^\rho(z) \, d\rho(z) \right)^2.$$

By the Parserval's equality in $L^2(\rho)$, we have

$$\sum_{i=1}^{n(\nu)} \Lambda_i^\nu \int_{\mathcal{S} \times \mathcal{A}} |\psi_i^\nu(z)|^2 \, d\rho(z) = \sum_{i=1}^{n(\nu)} \sum_{j=1}^{+\infty} \Lambda_i^\nu \left(\int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) \psi_j^\rho(z) \, d\rho(z) \right)^2 = \sum_{j=1}^{\infty} c_j. \quad (6.11)$$

Note that

$$\begin{aligned}
c_j &= \sum_{i=1}^{n(\nu)} \Lambda_i^\nu \left(\int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) \psi_j^\rho(z) \, d\rho(z) \right)^2 \\
&\leq \sum_{i=1}^{+\infty} \Lambda_i^\nu \left(\int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) \psi_j^\rho(z) \, d\rho(z) \right)^2 \\
&= \int_{\mathcal{S} \times \mathcal{A}} \int_{\mathcal{S} \times \mathcal{A}} \psi_j^\rho(z) \psi_j^\rho(z') \sum_{i=1}^{+\infty} \Lambda_i^\nu \psi_i^\nu(z) \psi_i^\nu(z') \, d\rho(z) \, d\rho(z') \\
&= \int_{\mathcal{S} \times \mathcal{A}} \int_{\mathcal{S} \times \mathcal{A}} \psi_j^\rho(z) \psi_j^\rho(z') k(z, z') \, d\rho(z) \, d\rho(z') = \Lambda_j^\rho,
\end{aligned}$$

and, by Eq. (2.2),

$$n(\nu) = \sum_{i=1}^{n(\nu)} \Lambda_i^\nu \|\psi_i^\nu\|_{\mathcal{H}_k}^2 = \sum_{i=1}^{n(\nu)} \sum_{j=1}^{+\infty} \frac{\Lambda_i^\nu \left(\int_{\mathcal{S} \times \mathcal{A}} \psi_i^\nu(z) \psi_j^\rho(z) \, d\rho(z) \right)^2}{\Lambda_j^\rho} = \sum_{j=1}^{+\infty} \frac{c_j}{\Lambda_j^\rho}.$$

Combining the last two inequality, we have

$$0 \leq c_j \leq \Lambda_j^\rho, \quad \sum_{j=1}^{+\infty} \frac{c_j}{\Lambda_j^\rho} = n(\nu). \qquad (6.12)$$

Under the constraint (6.12) and the condition $\{\Lambda_j^\rho\}_{j\in\mathbb{N}^+}$ being nonincreasing, we know $\sum_{j=1}^{+\infty} c_j$ achieves the maximum value when $c_j = \Lambda_j^\rho$ for $j = 1, \cdots, n(\nu)$ and $c_j = 0$ for $j \geq n(\nu) + 1$. Therefore, by equation (6.11), we know that

$$\sum_{i=1}^{n(\nu)} \Lambda_i^\nu \int_{\mathcal{S}\times\mathcal{A}} |\psi_i^\nu(z)|^2 \, \mathrm{d}\rho(z) = \sum_{j=1}^{+\infty} c_j \leq \sum_{j=1}^{n(\nu)} \Lambda_j^\rho.$$

Combining the last inequality with (6.9), (6.10), we have

$$\sup_{\nu' \in \mathcal{P}(\mathcal{S}\times\mathcal{A})} \inf_{g \in L^2(\nu)} \left[\mathrm{MMD}(\nu', g \circ \nu) + \frac{\|g\|_{L^2(\nu)}}{\sqrt{n}} \right]^2 \geq \frac{1}{2} \sum_{i=n(\nu)+1}^{+\infty} \Lambda_j^\rho.$$

We finish our proof by noticing that both ν and ρ are arbitrary probability distributions. $\qquad\square$

The above proposition shows that the perturbation response $\mathcal{R}(\mathcal{P}(\mathcal{S}\times\mathcal{A}), \mathcal{H}_k, n^{-\frac{1}{2}}, \nu)$ is bounded by the eigenvalue decay. On the one hand, if the eigenvalue decay of $\mathcal{K}_\nu$ is fast, even when Π includes all probability distributions on $\mathcal{S}\times\mathcal{A}$, the decay of $\mathcal{R}(\Pi, \mathcal{H}_k, n^{-\frac{1}{2}}, \nu)$ with respect to n is fast, which explains the positive results for RL algorithms in RKHS established in [15,16]. On the other hand, we note that the kernel function can be defined through Mercer decomposition, with the only requirement of $\{\Lambda_i^\rho\}_{i\in\mathbb{N}^+}$ being $\sum_{i=1}^{+\infty} \Lambda_i^\rho < +\infty$. Therefore, without further assumption on the kernel k, we only know that the right-hand side of (6.6) converges to zero when n goes to infinity, but its speed can be arbitrary slow. In fact, for many popular RKHSs like the RKHSs corresponding to Laplace kernel and neural tangent kernel on sphere $\mathbb{S}^{d-1}$, the right-hand side of (6.6) can be bounded below by $n^{-\frac{1}{d-1}}$; see [17, Section 5] for discussion. For those RKHSs, the knowledge of Π, such as condition (6.4), plays a vital role in designing efficient RL algorithms. Without such knowledge of Π, there exist many RL problems whose sample complexity suffers from the curse of dimensionality. Below we give two such examples.

Single state, high dimensional action space. We first consider a problem in which the state space $\mathcal{S}$ consists of a single point s_0 while the action space $\mathcal{A}$ is $\mathbb{S}^{d-1}$ and $H = 1$. In this setting, 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. Based on Proposition 6.3, the convergence rate can be bounded below by the eigenvalue decay. Therefore, if we consider the RKHS corresponding to the Laplacian kernel or neural tangent kernel, the convergence rate suffers from the curse of dimensionality. We can then conclude that if we want to solve RL problems with high dimensional action space, we need to assume the decay of eigenvalue is fast enough to break the curse of dimensionality.

High dimensional state, finite action space. Even when the action space is finite, there still exist MDPs that cannot be solved without the curse of dimensionality. For any dimension $d \geq 2$, length of each episode $H \in \mathbb{N}^+$ and positive constant $\delta > 0$, we define an

MDP family $\mathcal{M}_{d,H,\delta}$ as follows:

$$
\mathcal{S} = \mathsf{S}^{d-1}, \quad \mathcal{A} = \{0,1\}, \quad H = H, \quad \Theta_P = \{0\}, \quad \mu = \text{Uniform}_{\mathsf{S}^{d-1}},
$$
$$
\{r_{\theta_r} : \theta_r \in \Theta_r\} = \{r : \|r(h,\cdot)\|_{\mathcal{H}_k} \leq 1, \, \forall h \in [H]\},
$$
$$
k((s,a),(s',a')) = \exp(-\|s - s'\|), \quad P_0(\cdot \,|\, h,s,a) = \delta_{T_{a,h}s}(\cdot),
$$
$$
T_{a,h}s = \begin{cases} (\phi_1, \cdots, \phi_{h_d} + \delta, \cdots, \phi_d), & \text{when } a = 0, \\ (\phi_1, \cdots, \phi_{h_d} - \delta, \cdots, \phi_d), & \text{when } a = 1, \end{cases}
$$

where $h_d = h \bmod d$ and we use the spherical coordinates $(\phi_1, \cdots, \phi_d)$ to denote the points on S^{d-1}. Notice that for every d, H, δ, the transition probability in the RL problem $\mathcal{M}_{d,H,\delta}$ is known since Θ_P is a single-point set. By construction, the transition probability P_0 indicates that the agent can take an action at each step to move on the sphere surface along one spherical coordinate (depending on h) with size δ. When δ is small and H is large, we know that for any delta function on S^{d-1}, we can find a policy such that the corresponding state distribution at step H is close to that delta function. In other words, the set $\Pi(H, P_0, \mu)$ is very large. Noticing that the kernel is a Laplacian kernel whose eigenvalue decay is slow, we expect the RL problem to be difficult to solve. The following theorem shows that there exists no dimension-free bound on $\Delta_{\mathcal{M}_{d,H,\delta}}(n^{-\frac{1}{2}})$ for all n, d, H, and δ. Therefore, the above RL problem can not be solved without the curse of dimensionality.

Proposition 6.4. *There exist no universal constants $\alpha, \beta > 0$ and constant $C_d > 0$ only depending on d such that*

$$
\sup_{\delta > 0} \Delta_{\mathcal{M}_{d,H,\delta}}(n^{-\frac{1}{2}}) \leq C_d H^\alpha \left(\frac{1}{n}\right)^\beta
$$

holds for all $n, H \in \mathbb{N}^+$ and $d \geq 2$.

Proof. For any $s \in \mathsf{S}^{d-1}$, we define the following policy π^s

$\pi^s_h(\cdot \,|\, s') :$ choose the policy a such that $T_{a,h}s'$ is closest to s among all possible choices.

Then, in the above MDP with policy π^s, when $h \geq \frac{C'_d}{\delta}$, we have

$$
\|S_h - s\| \leq C'_d \delta \text{ with probability 1.} \tag{6.13}
$$

Here $C'_d > 0$ is a constant only depending on d, which may vary from line to line in the following proof. Using Proposition 6.3 and the eigenvalue decay of the Laplacian kernel (see, e.g., [17, Section 5]), we know that for any $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$,

$$
\sup_{z \in \mathcal{S} \times \mathcal{A}} \inf_{g \in L^2(\nu)} \left[\text{MMD}_k(\delta_z, g \circ \nu) + \frac{\|g\|_{L^2(\nu)}}{\sqrt{n}} \right]
$$
$$
= \sup_{\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A})} \inf_{g \in L^2(\nu)} \left[\text{MMD}_k(\rho, g \circ \nu) + \frac{\|g\|_{L^2(\nu)}}{\sqrt{n}} \right]
$$
$$
\geq C'_d n^{-\frac{1}{2(d-1)}}.
$$

Therefore, there exists $(s^*, a^*) \in \mathcal{S} \times \mathcal{A}$ such that

$$\inf_{g \in L^2(\nu)} \left[\text{MMD}_k(\delta_{(s^*, a^*)}, g \circ \nu) + \frac{\|g\|_{L^2(\nu)}}{\sqrt{n}} \right] \geq C'_d n^{-\frac{1}{2(d-1)}}.$$

Combining the last equation with Lemma 6.1, if $H = [\frac{C'_d}{\delta}]$, we have

$$\mathcal{R}(\Pi(H, P_0, \mu), \mathcal{H}_k, n^{-\frac{1}{2}}, \nu) \geq \inf_{g \in L^2(\nu)} \left[\text{MMD}_k(\rho_{H, P_0, \pi^{s^*}, \mu}, g \circ \nu) + \frac{\|g\|_{L^2(\nu)}}{\sqrt{n}} \right]$$

$$\geq C'_d n^{-\frac{1}{2(d-1)}} - \text{MMD}_k(\delta_{(s^*, a^*)}, \rho_{H, P_0, \pi^{s^*}, \mu}).$$

By inequality (6.13), we have,

$$\text{MMD}_k(\delta_{(s^*, a^*)}, \rho_{H, P_0, \pi^{s^*}, \mu}) \leq C'_d \delta.$$

Therefore, we can choose

$$\delta = C'_d n^{-\frac{1}{2(d-1)}},$$

such that

$$\mathcal{R}(\Pi(H, P_0, \mu), \mathcal{H}_k, n^{-\frac{1}{2}}, \nu) \geq C'_d n^{-\frac{1}{2(d-1)}}$$

and

$$H = [C'_d n^{\frac{1}{2(d-1)}}].$$

Therefore, combining the last two equations and the definition of $\Delta_{\mathcal{M}_{d,H,\delta}}(n^{-\frac{1}{2}})$ in the case of known transition (4.1), if the constants α, β and C_d exist, we must have

$$C'_d n^{-\frac{1}{2(d-1)}} \leq C_d (C'_d)^\alpha n^{\frac{\alpha}{2(d-1)} - \beta}$$

holds for all $n \in \mathbb{N}^+$ and $d \geq 2$. Therefore,

$$\frac{\alpha + 1}{2(d-1)} \geq \beta$$

holds for all $d \geq 2$, which is a contradiction. □

7 Conclusions and future works

In this paper, we define the perturbational complexity by distribution mismatch $\Delta_{\mathcal{M}}(\epsilon)$ when the reward functions lie in the unit ball of an RKHS and the transition probabilities lie in a given arbitrary set. We show that $\Delta_{\mathcal{M}}(\epsilon)$ is an informative indicator of whether the RL problems can be solved efficiently or not. Some concrete properties of $\Delta_{\mathcal{M}}(\epsilon)$ are studied in several cases. There are still quite a few unsolved problems related to this topic. First, the upper bound and lower bound in the current work are not matched. Second, in

the case of unknown transition, we still need assumption (5.11) to prove the convergence of fitted Q-iteration algorithms. How to relax this assumption or show the necessity of this assumption remains unclear yet. Third, our lower bound mainly utilizes the uncertainty of reward functions. If the reward function is known or the reward function is deterministic such that there is no noise in reward, our lower bound can not be applied. It is of interest to study the lower bound and corresponding upper bound in these situations. Fourth, our setup of RL requires a generative simulator. In some RL problems, however, one can only access an episodic simulator. How to establish similar results in this case is still open. Finally, we wish to use information related to the perturbational complexity to guide the design of efficient RL algorithms in practice.

Acknowledgments

References

[1] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[2] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. On the sample complexity of reinforcement learning with a generative model. In *International Conference on Machine Learning*, page 17071714, 2012.

[3] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

[4] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[5] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 5713–5723, 2017.

[6] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, pages 4863–4873, 2018.

[7] Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[8] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

[9] Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2021.

[10] Gen Li, Yuxin Chen, Yuejie Chi, Yuantao Gu, and Yuting Wei. Sample-efficient reinforcement learning is feasible for linearly realizable mdps with limited revisiting. *arXiv preprint arXiv:2105.08024*, 2021.

[11] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[12] Weinan E, Chao Ma, and Lei Wu. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019.

[13] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

[14] Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Regret bounds for kernel-based reinforcement learning. *arXiv preprint arXiv:2004.05599*, 2020.

[15] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems*, 33, 2020.

[16] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*, 2020.

[17] Jihao Long, Jiequn Han, and Weinan E. An L^2 analysis of reinforcement learning in high dimensions with kernel and neural network approximation. *arXiv preprint arXiv:2104.07794, to appear in CSIAM Transactions on Applied Mathematics*, 2021.

[18] Frances Y Kuo, Grzegorz W Wasilkowski, and Henryk Woźniakowski. Multivariate L_∞ approximation in the worst case setting over reproducing kernel Hilbert spaces. *Journal of Approximation Theory*, 152(2):135–160, 2008.

[19] Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[20] Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020.

[22] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[23] Amir Massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

[24] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

[25] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[26] Chengzhuo Ni, Lin F Yang, and Mengdi Wang. Learning to control in metric space with optimal regret. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 726–733. IEEE, 2019.

[27] Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

[28] Yaqi Duan, Mengdi Wang, and Martin J Wainwright. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*, 2021.

[29] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[30] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[31] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[32] Paul R Halmos. *Measure Theory*, volume 18. Springer, 2013.

[33] Mark S Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, 1964.

[34] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[35] Vern I Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152. Cambridge University Press, 2016.

[36] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

[37] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[38] Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, 2020.

[39] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[40] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer Science & Business Media, 2010.

[41] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[42] Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of tetris. *The Journal of Machine Learning Research*, 16:1629–1676, 2015.

[43] Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Conference on Learning Theory*, pages 79–93, 2009.

[44] Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

[45] Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.