Primal-Dual Spectral Representation for Off-policy Evaluation

Anonymous CPAL submission

Off-policy evaluation (OPE) is one of the most fundamental problems in reinforcement learning (RL) to estimate the expected long-term payoff of a given target policy with *only* experiences from another behavior policy that is potentially unknown. The distribution correction estimation (DICE) family of estimators have advanced the state of the art in OPE by breaking the *curse of horizon*. However, the major bottleneck of applying DICE estimators lies in the difficulty of solving the saddle-point optimization involved, especially with neural network implementations. In this paper, we tackle this challenge by establishing a *linear representation* of value function and stationary distribution correction ratio, *i.e.*, primal and dual variables in the DICE framework, using the spectral decomposition of the transition operator. Such primal-dual representation not only bypasses the non-convex non-concave optimization in vanilla DICE, therefore enabling an computational efficient algorithm, but also paves the way for more efficient utilization of historical data. We highlight that our algorithm, SpectralDICE, is the first to leverage the linear representation of primal-dual variables that is both computation and sample efficient, the performance of which is supported by a rigorous theoretical sample complexity guarantee and a thorough empirical evaluation on various benchmarks.

1 Introduction

The past decade has witnessed the ubiquitous success of reinforcement learning (RL) across various domains. Despite the original rationale that RL agents should learn a reward-maximizing policy from continuous interactions with the environment, there also exist a wide range of applicational scenarios where *online* interaction with the environment may be expensive, inefficient, risky, unethical, and/or even infeasible, examples of which include robotics [1, 2], autonomous driving [3, 4], healthcare [5, 6], education [7, 8], dialogue systems [9, 10] and recommendation systems [11, 12]. These application scenarios motivate the study of *offline* RL, where the learning agent only has access to historical data collected by a separate behavior policy.

Off-policy evaluation (OPE) is one of the most fundamental problems in offline RL that aims at estimating the expected cumulative reward of a given target policy using only historical data collected by a different, potentially unknown behavior policy. In the past decade, various off-policy performance estimators have been proposed [13–16]. However, these estimators generally suffer from the *curse of horizon* [17]—step-wise variances accumulate in a multiplicative way, resulting in prohibitively high trajectory variances and thus unreliable estimators. The recently proposed Distribution Correction Estimation (DICE) family of estimators have advanced the state of the art in OPE, leveraging the primal-dual formulation of policy evaluation for a saddle-point optimization approach that directly estimates the stationary distribution correction ratio, and hence breaking the curse of horizon [18, 19].

Nevertheless, as systems scale up in terms of the size of state-action spaces, the saddle-point optimization in the formulation of DICE estimators become increasingly challenging to solve. Such *curse of dimensionality* is common for RL methods in general, and people have been working to alleviate the computational burden by exploiting function approximators. However, many known function approximators require additional assumptions to ensure computational and statistical properties [20–25], which may not be easily statisfiable in practice. Moreover, the induced optimiza-

Submitted to Second Conference on Parsimony and Learning (CPAL 2025). Do not distribute.

tion upon function approximators may be difficult to solve [26–28]. In particular, under a generic neural network parametrization, computing the DICE estimator [29] requires solving non-convex non-concave saddle-point optimizations, which is known to be NP-hard in theory and also yields unstable performance in practice, and is therefore regarded as intractable.

This dilemma brings up a very natural question:

*Can we design an OPE algorithm that is both **efficient** and **practical**?*

By "efficient" we mean its statistical complexity avoids an exponential dependence on both the length of history and the dimension of state-action spaces, *i.e.*, eliminating both *curse of horizon* and *curse of dimensionality*. By "practical" we mean the algorithm is free from unstable saddle-point optimizations and can be easily implemented and applied in practical settings.

In this paper, we provide an *affirmative* answer to this question by revealing a novel linear structure encapsulating both Q-functions and distribution correction ratios via a spectral representation of the transition operator, which has many nice properties to enable efficient representation learning and off-policy evaluation.

Contributions. Specifically, the contributions of this paper can be summarized as follows:

- We propose a novel *primal-dual spectral representation* of the state-action transition operator, which makes both the Q-function and the stationary distribution correction ratio (*i.e.*, the primal and dual variables in DICE) linearly representable in the primal/dual feature spaces, and thus enhances the tractability of the corresponding DICE estimator.
- We design SpectralDICE, an off-policy evaluation algorithm based on our primal-dual spectral representation, which bypasses the non-convex non-concave saddle-point optimization in vanilla DICE with generic neural network parameterization, and also makes efficient use of historical data. As far as we are concerned, our algorithm is the first to leverage the linear representation of both primal and dual variables that is *computation and sample efficient*.
- The performance of the SpectralDICE algorithm is justified both theoretically with a rigorous sample complexity guarantee and empirically by a thorough evaluation on various RL benchmarks.

1.1 Related Work

Off-Policy Evaluation (OPE). Off-policy evaluation has long been an active field of RL research. In the case where the behavior policy is known, various off-policy performance estimators have been proposed, including direct method (DM) estimators [30, 31], importance sampling (IS) estimators [13, 14], doubly-robust (DR) estimators [15, 16, 32] and other mixed-type estimators [24, 33, 34], which generally suffer from the *curse of dimension*. In an effort to settle this issue, there is also abundant literature on estimating the correction ratio of the stationary distribution [17, 35], among which the distribution correction estimation (DICE) family of estimators are the state of the art that leverage a novel primal-dual formulation of OPE to eliminate the curse of horizon, and in the meantime, allow unknown behavior policies [18, 19, 29, 36, 37]. However, as discussed above, the induced saddle-point optimization becomes unstable with neural networks, impeding the practical application of DICE estimators.

Spectral Representation in MDPs. Spectral decomposition of the transition kernel is known to induce a linear structure of Q-functions, which enables the design of provably efficient algorithms assuming known (primal) spectral feature maps [38–40]. These algorithms break the curse of dimensionality in the sense that their computation or sample complexity is independent of the size of the state-action space, but rather, only depends polynomially on the feature space dimension, the intrinsic dimension of the problem.

With the growing interest in spectral structures of MDPs, representation learning for RL has recently attracted much theory-oriented attention in the online setting [41, 42]. Practical representation-based online RL algorithms have been designed via kernel techniques [43, 44], latent variable mod-

els [45, 46], contrastive learning [47, 48], and diffusion score matching [49]. Recently, a unified representation learning framework is proposed from a novel viewpoint that leverages the spectral decomposition of the transition operator [40].

Spectral representations have also been exploited in the offline setting [50–52], where the temporal difference algorithm is applied in the linear space induced by the primal spectral feature for estimating Q-functions. The linear structure of the occupancy measure induced by the dual spectral feature is recently utilized in Huang et al. [53], which leads to an offline RL algorithm for stationary density ratio estimation. Although the algorithm is theoretically sound, the stationary density ratio breaks the linearity in occupancy, and hence the algorithm is not computationally efficient. As far as we know, there is no such offline RL algorithm that efficiently utilizes both primal and dual representations.

2 Preliminaries

Notations. Denote by $\|\cdot\|_p$ the p-norm of vectors or the L^p-norm of functionals, and by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y}$ the Euclidean inner product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Denote by $\widehat{\mathbb{E}}_{d^{\mathcal{D}}}[\cdot]$ the empirically approximated expectation using samples from dataset $\mathcal{D} \sim d^{\mathcal{D}}$. Denote by $\Delta(S)$ the set of distributions over set S, the element of which shall be regarded as densities whenever feasible. Denote the indicator function by $\mathbb{1}\{\cdot\}$. Write $[n] := \{1, \dots, n\}$ for $n \in \mathbb{Z}_+$. Regard $f(n) \lesssim g(n)$ as $f(n) = O(g(n))$.

Markov Decision Processes (MDPs). We consider an *infinite-horizon* discounted Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \mu_0, \gamma)$, where $\mathcal{S}$ is the (possibly infinite) state space, $\mathcal{A}$ is the (possibly infinite) action space; $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function; $\mu_0 \in \Delta(\mathcal{S})$ is the initial state distribution, and $\gamma \in (0, 1)$ is the reward discount factor, so that the discounted cumulative reward can be defined as $\sum_{t=0}^{\infty} \gamma^t r_t$. We consider *stationary Markovian policies* $\mathit{\Pi} := \{\pi : \mathcal{S} \to \Delta(\mathcal{A})\}$ that admit an action distribution depending on the current state only. Given any policy $\pi \in \mathit{\Pi}$, let $\mathbb{E}_{\pi,\mathbb{P}}[\cdot]$ denote the expectation over the trajectory governed by π and $\mathbb{P}$ (possibly under prescribed initial conditions). Let $d^\pi_{\mathbb{P}}(\cdot, \cdot) \in \Delta(\mathcal{S} \times \mathcal{A})$ denote the (*stationary*) *state-action occupancy measure* under policy π, *i.e.*, the normalized discounted probability of visiting (s, a) in a trajectory induced by policy π, defined by

$$d^\pi_{\mathbb{P}}(s, a) = (1 - \gamma)\mathbb{E}_{\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}\{s_t = s, a_t = a\}\right].$$

Similarly, let $d^\pi_{\mathbb{P}}(\cdot) \in \Delta(\mathcal{S})$ denote the *state occupancy measure* subject to the relation $d^\pi_{\mathbb{P}}(s, a) = d^\pi_{\mathbb{P}}(s)\pi(a|s)$. Further, define the state/state-action value functions (a.k.a. V- and Q-functions) as:

$$V^\pi_{\mathbb{P}}(s) := \mathbb{E}_{\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right],$$
$$Q^\pi_{\mathbb{P}}(s, a) := \mathbb{E}_{\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s,\ a_0 = a\right].$$

In this way, the value of policy π in $\mathcal{M}$ is defined by

$$\rho_{\mathbb{P}}(\pi) := (1 - \gamma)\mathbb{E}_{s\sim\mu_0}[V^\pi_{\mathbb{P}}(s)] = (1 - \gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}[Q^\pi_{\mathbb{P}}(s, a)],$$

where the factor $(1 - \gamma)$ is introduced for normalization. We omit the subscript $\mathbb{P}$ in clear context.

Remark 1. In order to better illustrate how the proposed method works in MDPs with continuous state-action spaces, we abuse the notation a bit to regard $\mathbb{P}$, π and d^π as *densities*. Parallel results for the discrete case can be analogously derived without difficulties.

The Primal-Dual Characterization of $\rho(\pi)$. Distribution Correction Estimation (DICE) [29] is a primal-dual-based method that evaluates the value of a given target policy π in the offline setting, using the linear programming (LP) formulation of policy values [54]. Specifically, it is known that

we can equivalently characterize $\rho(\pi)$ defined in (1) by the *primal LP*:

$$\begin{aligned} \min_{Q(\cdot,\cdot)} \quad & (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}[Q(s,a)], \\ \text{s.t.} \quad & Q(s,a) \geqslant r(s,a) + \gamma\mathbb{E}_{\substack{s'\sim\mathbb{P}(\cdot|s,a),\\ a'\sim\pi(\cdot|s')}}[Q(s',a')],\ \forall (s,a)\in\mathcal{S}\times\mathcal{A}. \end{aligned} \tag{1}$$

Further, it can be shown that strong duality holds in (1), with Lagrangian multipliers exactly the state-action occupancy measures $d^\pi(\cdot,\cdot)$. Then we characterize $\rho(\pi)$ by the following *primal-dual LP*:

$$\min_{Q(\cdot,\cdot)}\max_{d(\cdot,\cdot)} \quad (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}[Q(s,a)] + \mathbb{E}_{(s,a)\sim d^\pi(\cdot,\cdot)}\left[r(s,a) + \gamma\mathbb{E}_{\substack{s'\sim\mathbb{P}(\cdot|s,a),\\ a'\sim\pi(\cdot|s')}}[Q(s',a')] - Q(s,a)\right]. \tag{2}$$

We highlight that this primal-dual LP formulation is favored in the offline RL setting in that historical experiences can be utilized to empirically approximate the expectations in (2) after some simple change-of-variables. In particular, for any measurable function $f(s,a)$, the importance sampling (IS) estimator for the expected value of $f(s,a)$ is given by

$$\mathbb{E}_{(s,a)\sim d^\pi}[f(s,a)] = \mathbb{E}_{(s,a)\sim d^{\pi_\mathrm{b}}}\left[\frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}\cdot f(s,a)\right], \tag{3}$$

where $\zeta(s,a) := \frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}$ is known as the *stationary distribution correction ratio* for dataset $\mathcal{D}\sim d^\mathcal{D}$.

The DICE family estimators [18, 55, 56] is designed by plugging the IS expectation estimator (3) into (2), such that the stationary distribution correction ratio $\zeta(\cdot,\cdot)$ is parameterized along with the Q-function to formulate an optimization, with various regularization available [57]. It is evident that the DICE family estimators are applicable to the offline RL setting with unknown behavior policy.

Spectral Representation. We can always perform spectral decomposition of the dynamic operator to obtain a spectral representation of *any* MDP [40]. In particular, *low-rank MDPs* refer to such MDPs with intrinsic finite-rank spectral representation structures that enable scalable RL algorithms, and are thus of theoretical interest [38, 58]. Formally, $\mathcal{M}$ is said to be a *low-rank MDP* if there exists a *primal* feature map $\boldsymbol{\phi}: \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ and *dual* features $\tilde{\boldsymbol{\mu}}: \mathcal{S}\to\mathbb{R}^d$, $\boldsymbol{\theta}_r\in\mathbb{R}^d$, such that $\mathbb{P}(s'|s,a) = \langle\boldsymbol{\phi}(s,a), \tilde{\boldsymbol{\mu}}(s')\rangle$, $r(s,a) = \langle\boldsymbol{\phi}(s,a), \boldsymbol{\theta}_r\rangle$, for any $s,s'\in\mathcal{S}$, $a\in\mathcal{A}$. Here both the primal feature $\boldsymbol{\phi}$ and the dual features $\tilde{\boldsymbol{\mu}}$, $\boldsymbol{\theta}_r$ are assumed to be unknown, and thus must be learned from data [41, 42].

Unfortunately, it is revealed in Ren et al. [40], Zhang et al. [48] that learning the features of a low-rank MDP is difficult from the unnormalized density fitting point of view. To settle this tractability issue, the above papers propose a reparameterization of the dual feature as $\tilde{\boldsymbol{\mu}}(\cdot) = q(\cdot)\boldsymbol{\mu}(\cdot)$, where we introduce an auxiliary distribution $q(\cdot)\in\Delta(\mathcal{S})$ that will be specified later. Therefore, we will stick to the following spectral decomposition of the transition kernel in this paper:

$$\mathbb{P}(s'|s,a) = \langle\boldsymbol{\phi}(s,a), q(s')\boldsymbol{\mu}(s')\rangle,\ \forall s\in\mathcal{S}, a\in\mathcal{A}, s'\in\mathcal{S}. \tag{4}$$

Under such reparameterization, it is known that the spectral representaton can be learned efficiently.

Additionally, we also assume μ_0 to be linearly representable in the dual feature space.

Assumption 1 (initial representation)**.** There exists $\boldsymbol{\omega}_0\in\mathbb{R}^d$, such that $\mu_0(s) = q(s)\langle\boldsymbol{\mu}(s), \boldsymbol{\omega}_0\rangle$, $\forall s$.

Off-Policy Evaluation (OPE). Consider a setting where we are given $\mathcal{D} = \{(s_i, a_i, s_i') \mid i\in[N]\}$, an offline dataset of N historical transitions, sampled by a *behavior policy* π_b that could be unknown. The objective is to estimate the expected cumulative rewards $\rho(\pi)$ of a different *target policy* π.

For satisfactory performance, it is important that the behavior policy provides sufficient data coverage for the frequent transitions experienced by policy π. Specifically, we assume the occupancy ratio between π and π_b satisfies the following regularity assumption.

Assumption 2 (concentratability)**.** $\frac{d^\pi(s,a)}{d^{\pi_\mathrm{b}}(s,a)} \leqslant C^\pi_\infty, \forall s\in\mathcal{S}, a\in\mathcal{A}$.

We point out that the concentratability assumption is standard in offline RL literature [22, 59], and is also implicitly enforced in recent work like Huang et al. [53] (see Definition 1 therein). We are

aware that the coefficient C_∞^π can potentially be translated into different feature-related constants [42], which does not change the asymptotics of sample complexity, yet only adds to the technical complexity. For clarity, we will stick to the simple Assumption 2 in this paper.

3 SpectralDICE: OPE using Primal-Dual Spectral Representation

In this section, we first introduce a novel linear representation for the stationary distribution correction ratio using the *dual* spectral feature of transition kernel. We highlight that this linear structure, together with the known linear representation of Q-functions, helps to bypass the non-convex non-concave optimization required in the computation of DICE estimators, and also enables efficient utilization of historical data sampled by unknown behavior policies. Based on the above ideas, we present SpectralDICE, the proposed off-policy evaluation (OPE) algorithm using our primal-dual spectral representation.

3.1 Primal-Dual Spectral Representation

We start by specifying the primal-dual spectral representation used in SpectralDICE. At first glance, it may seem natural to directly learn the spectral representation of $\mathbb{P}$ as defined in (4). However, it turns out that this naive approach includes the target policy π in the linear representation of $d^\pi(\cdot,\cdot)$, which in turn induces a complicated representation for the stationary distribution correction ratio $\zeta(\cdot,\cdot)$ [53], and thus, leads to an intractable optimization (2) for the computation of the DICE estimator.

The above challenge inspires us to properly reparameterize the spectral decomposition (4). Specifically, since we only work with a fixed target policy π for off-policy evaluation, we shall consider the following alternative representation of the state-action transition kernel $\mathbb{P}^\pi(s',a'|s,a) := \mathbb{P}(s'|s,a)\pi(a'|s')$:

$$\mathbb{P}^\pi(s',a'|s,a) = \Big\langle \boldsymbol{\phi}(s,a), q(s')\pi_{\mathrm{b}}(a'|s') \underbrace{\tfrac{\pi(a'|s')}{\pi_{\mathrm{b}}(a'|s')}\boldsymbol{\mu}(s')}_{\boldsymbol{\mu}^\pi(s',a')} \Big\rangle. \tag{5}$$

Note that Assumption 2 guarantees a non-zero denominator when the nominator is non-zero. We refer to (5) as the *primal-dual spectral representation* of the state-action) transition kernel $\mathbb{P}^\pi$, where $\boldsymbol{\phi}(\cdot,\cdot)$ and $\boldsymbol{\mu}^\pi(\cdot,\cdot)$ are still called *primal* and *dual* spectral features, respectively. The superscript π of the dual spectral feature emphasizes its dependence on the target policy.

The primal-dual spectral representation has several nice properties. In particular, we can show that the Q-function $Q^\pi(s,a)$, the state-action occupancy measure $d^\pi(s,a)$, and the stationary distribution correction ratio $\zeta(s,a)$ can all be represented in linear forms using the primal/dual features, as summarized below.

Lemma 1. *With primal-dual spectral representation (5), the Q-function $Q^\pi(\cdot,\cdot)$ is linearly representable in the primal feature space with cofactor $\boldsymbol{\theta}_Q^\pi \in \mathbb{R}^d$:*

$$Q^\pi(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_Q^\pi \rangle, \ \forall s \in S, a \in \mathcal{A}. \tag{6}$$

Further, under Assumption 1, the state-action occupancy measure $d^\pi(\cdot,\cdot)$ is also linearly representable in the dual feature space with cofactor $\boldsymbol{\omega}_d^\pi \in \mathbb{R}^d$:

$$d^\pi(s,a) = q(s)\pi_{\mathrm{b}}(a|s)\langle \boldsymbol{\mu}^\pi(s,a), \boldsymbol{\omega}_d^\pi \rangle, \ \forall s \in S, a \in \mathcal{A}.$$

Specifically, when the auxiliary distribution $q(\cdot)$ is selected as the state-occupancy measure $d^{\pi_{\mathrm{b}}}(\cdot)$ of the behavior policy π_{b}, the stationary distribution correction ratio can also be linearly represented as:

$$\zeta(s,a) = \frac{d^\pi(s,a)}{q(s)\pi_{\mathrm{b}}(a|s)} = \langle \boldsymbol{\mu}^\pi(s,a), \boldsymbol{\omega}_d^\pi \rangle. \tag{7}$$

Proof. Note that the original dual feature in (4) can be restored by $\boldsymbol{\mu}(s') = \frac{\pi_{\mathrm{b}}(a'|s')}{\pi(a'|s')}\boldsymbol{\mu}^\pi(s', a')$ for any $a' \in \mathcal{A}$. Then by Bellman recursive equation we have:

$$\begin{aligned} Q^\pi(s,a) &= \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_r \rangle + \gamma \int V^\pi(s') \langle \boldsymbol{\phi}(s,a), q(s')\boldsymbol{\mu}(s') \rangle \mathrm{d}s' \\ &= \Big\langle \boldsymbol{\phi}(s,a), \underbrace{\boldsymbol{\theta}_r + \gamma \int V^\pi(s') q(s') \boldsymbol{\mu}(s') \mathrm{d}s'}_{\boldsymbol{\theta}_Q^\pi} \Big\rangle. \end{aligned}$$

Similarly, by the recursive property of d^π we have:

$$\begin{aligned} d^\pi(s,a) &= (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \int d^\pi(\tilde{s},\tilde{a}) \mathbb{P}^\pi(s,a|\tilde{s},\tilde{a}) \mathrm{d}\tilde{s}\mathrm{d}\tilde{a} \\ &= (1-\gamma) q(s) \big\langle \pi_{\mathrm{b}}(a|s)\boldsymbol{\mu}^\pi(s,a), \boldsymbol{\omega}_0 \big\rangle + \gamma \Big\langle q(s)\pi_{\mathrm{b}}(a|s)\boldsymbol{\mu}^\pi(s,a), \int d^\pi(\tilde{s},\tilde{a})\boldsymbol{\phi}(\tilde{s},\tilde{a})\mathrm{d}\tilde{s}\mathrm{d}\tilde{a} \Big\rangle \\ &= \Big\langle q(s)\pi_{\mathrm{b}}(a|s)\boldsymbol{\mu}^\pi(s,a), \underbrace{(1-\gamma)\boldsymbol{\omega}_0 + \gamma \int d^\pi(\tilde{s},\tilde{a})\boldsymbol{\phi}(\tilde{s},\tilde{a})\mathrm{d}\tilde{s}\mathrm{d}\tilde{a}}_{\boldsymbol{\omega}_d^\pi} \Big\rangle, \end{aligned}$$

where we use the initial representation (Assumption 1) and the fact that $\pi(a|s)\boldsymbol{\mu}(s) = \pi_{\mathrm{b}}(a|s)\boldsymbol{\mu}^\pi(s,a)$. The representation of $\zeta(\cdot,\cdot)$ is hence a direct corollary since $q(s)\pi_{\mathrm{b}}(a|s) = d^{\pi_{\mathrm{b}}}(s,a)$ when $q(\cdot) = d^{\pi_{\mathrm{b}}}(\cdot)$. □

Then, using the linear spectral representations of Q and ζ in (6) and (7), we shall equivalently formulate the DICE estimator as follows.

Corollary 2. *With primal-dual spectral representation (5) where $q(\cdot) \equiv d^{\pi_{\mathrm{b}}}(\cdot)$, under Assumption 1,*

$$\begin{aligned} \rho_{\mathbb{P}}(\pi) = \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \Big\{ & (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0, \\ a\sim\pi(\cdot|s)}} \big[\boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_Q\big] \\ & + \mathbb{E}_{\substack{s\sim d^{\pi_{\mathrm{b}}}(\cdot),\ a\sim\pi_{\mathrm{b}}(a|s), \\ s'\sim\mathbb{P}(\cdot|s,a),\ a'\sim\pi(\cdot|s')}} \Big[\big(\boldsymbol{\mu}^\pi(s,a)^\top \boldsymbol{\omega}_d\big)\big(r(s,a) + \gamma \boldsymbol{\phi}(s',a')^\top \boldsymbol{\theta}_Q - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_Q\big) \Big] \Big\}. \end{aligned} \tag{8}$$

The proof of Corollary 2 is deferred to Appendix B.1 due to limited space. We highlight that our new DICE formulation (8) bears several benefits:

- **Offline data compatible.** The estimator is favorable for OPE since the expectation over the (s,a,s') transition pair can be effectively approximated by samples from the offline dataset $\mathcal{D}$, as long as the auxiliary distribution $q(\cdot)$ is selected as the state occupancy measure $d^{\pi_{\mathrm{b}}}$ of the behavior policy π_{b} such that $\Pr[(s,a,s') \in \mathcal{D}] = q(s)\pi_{\mathrm{b}}(a|s)\mathbb{P}(s'|s,a)$.
- **Optimization tractable.** Given (learned) $\boldsymbol{\phi}(s,a)$ and $\boldsymbol{\mu}^\pi(s,a)$, the saddle-point optimization in (8) is convex-concave with respect to both θ_Q and ω_d, which perfectly bypasses the optimization difficulty in vanilla DICE estimators with neural-network-parameterized $Q^\pi(\cdot,\cdot)$ and $\zeta(\cdot,\cdot)$. Meanwhile, compared to the counterpart obtained by directly applying the naive spectral representation (4) (details of which can be found in Appendix B.2), the proposed estimator (8) is tractable in that it is free of the policy ratio $\frac{\pi(a|s)}{\pi_{\mathrm{b}}(a|s)}$ that is unknown.

From now on, we will always regard $q(\cdot) \equiv d^{\pi_{\mathrm{b}}}(\cdot)$ for the aforementioned nice properties to hold.

3.2 Spectral Representation Learning

In the last section, we have elaborated on how to perform OPE using off-policy data given a primal-dual spectral representation. Now it only suffices to specify how to learn such a representation, which we regard as an abstract subroutine $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\mu}}^\pi) \leftarrow \textsc{RepLearn}(\mathcal{F}, \mathcal{D}, \pi)$. Here $\mathcal{F}$ denotes the collection of candidate representations. We highlight that our algorithm works with any representation learning method that has a bounded learning error, without any further requirements on the

Algorithm 1 SpectralDICE: **DI**stribution **C**orrection **E**stimation with **Spectral** Representation

Require: Target policy π, off-policy dataset $\mathcal{D}$, function family $\mathcal{F}$.

1: Learn a spectral representation $(\hat{\phi}, \hat{\boldsymbol{\mu}}^\pi) \leftarrow \textsc{RepLearn}(\mathcal{F}, \mathcal{D}, \pi)$.

2: Plug in the spectral representation $(\hat{\phi}, \hat{\boldsymbol{\mu}}^\pi)$ to compute the following DICE estimator:

$$\hat{\rho}(\pi) = \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \Bigg\{ (1-\gamma) \widehat{\mathbb{E}}_{\substack{s\sim\mu_0, \\ a\sim\pi(\cdot|s)}} \Big[\hat{\phi}(s,a)^\top \boldsymbol{\theta}_Q \Big] + \widehat{\mathbb{E}}_{\substack{(s,a,s')\sim\mathcal{D}, \\ a'\sim\pi(\cdot|s')}} \Big[\big(\hat{\boldsymbol{\mu}}^\pi(s,a)^\top \boldsymbol{\omega}_d\big) \big(r(s,a) + \gamma \hat{\phi}(s',a')^\top \boldsymbol{\theta}_Q - \hat{\phi}(s,a)^\top \boldsymbol{\theta}_Q\big) \Big] \Bigg\}. \tag{9}$$

3: **return** $\hat{\rho}(\pi)$

learning mechanism. Given a range of spectral representation learning methods available in literature [40, 45, 48, 49], for the sake of clarity we only consider a few candidates here, while other methods may also be applicable:

1. **Ordinary Least Squares (OLS).** Inspired by Ren et al. [40], an OLS objective can be constructed as follows. Denote by $\mathbb{Q}^\pi(s',a',s,a) := d^{\pi_\mathrm{b}}(s,a)\mathbb{P}^\pi(s',a'|s,a)$ the joint distribution of state-action transitions under behavior policy π_b, based on which we plug in (5) to obtain

$$\frac{\mathbb{Q}^\pi(s',a',s,a)}{\sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}} = \sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}\boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}^\pi(s',a'),$$

which further induces the following OLS objective:

$$\min_{(\hat{\phi},\hat{\boldsymbol{\mu}}^\pi)\in\mathcal{F}} \int \left(\frac{\mathbb{Q}^\pi(s',a',s,a)}{\sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}} - \sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}\hat{\boldsymbol{\phi}}(s,a)^\top \hat{\boldsymbol{\mu}}^\pi(s',a') \right)^2 \mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a'$$

Therefore, $(\hat{\phi}, \hat{\boldsymbol{\mu}}^\pi)$ can be learned by solving [40, 60]:

$$\min_{(\hat{\phi},\hat{\boldsymbol{\mu}}^\pi)\in\mathcal{F}} \Big\{ \widehat{\mathbb{E}}_{(s,a)\sim d^{\pi_\mathrm{b}},(\tilde{s}',\tilde{a}')\sim d^{\pi_\mathrm{b}}} \Big[\big(\hat{\boldsymbol{\phi}}(s,a)^\top \hat{\boldsymbol{\mu}}^\pi(\tilde{s}',\tilde{a}')\big)^2 \Big] - 2\widehat{\mathbb{E}}_{(s,a)\sim d^{\pi_\mathrm{b}},(s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)} \Big[\hat{\boldsymbol{\phi}}(s,a)^\top \hat{\boldsymbol{\mu}}^\pi(s',a') \Big] \Big\},$$

where the last term becomes a constant after expansion and is thus omitted. For practical implementation, we can use stochastic gradient descent to solve the above stochastic optimization problem.

2. **Noise-Contrastive Estimation (NCE).** NCE is a widely used method for contrastive representation learning in RL [47, 48]. To learn $(\hat{\phi}, \hat{\boldsymbol{\mu}}^\pi)$, we consider a binary contrastive learning objective [47]:

$$\min_{(\hat{\phi},\hat{\boldsymbol{\mu}}^\pi)\in\mathcal{F}} \widehat{E}_{(s,a)\sim d^{\pi_\mathrm{b}}} \Big[\widehat{\mathbb{E}}_{(s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)} \Big[\log\Big(1 + \frac{1}{\hat{\boldsymbol{\phi}}(s,a)^\top \hat{\boldsymbol{\mu}}^\pi(s',a')}\Big) \Big] + \widehat{\mathbb{E}}_{(s',a')\sim P_\mathrm{neg}} \Big[\log\Big(1 + \hat{\boldsymbol{\phi}}(s,a)^\top \hat{\boldsymbol{\mu}}^\pi(s',a')\Big) \Big] \Big],$$

where P_neg is a negative sampling distribution.

Details of these representation learning methods along with their learning errors can be found in Appendix C.

3.3 SpectralDICE

With the two key components specified above, now we are ready to state SpectralDICE, the proposed offline policy evaluation (OPE) algorithm using spectral representations, as in Algorithm 1.

Specifically, given a policy π, assuming access to an offline dataset $(s,a,s') \sim \mathcal{D}$ sampled by the behavior policy π_b, we follow a two-step algorithm to evaluate the target policy π in an off-policy manner:

1. **Representation learning.** We may choose any representation learning method that comes with a bounded learning error as the RepLearn subroutine, and the overall sample complexity will depend on this choice (see Section 4).

2. **DICE-based policy evaluation.** With the learned representation $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\mu}}^\pi)$, we use the primal-dual DICE estimator (9) to estimate the value of the target policy π. Note that the data distribution $d^{\mathcal{D}}(s,a,s') = d^{\pi_{\mathrm{b}}}(s)\pi_{\mathrm{b}}(a|s)\mathbb{P}(s'|s,a)$ is exactly compatible with the formulation in (8).

Remark 2 (Numerical considerations). It is known that directly solving (9) leads to potential numerical instability issues due to the objective's linearity in $\boldsymbol{\theta}_Q$ and $\boldsymbol{\omega}_d$ [19]. Fortunately, it is shown in Yang et al. [57] that certain regularization leads to strictly concave inner maximization while keeping the optimal *solution* unbiased (see Appendix B.3 for details). In our implementation, we append a regularizer $-\lambda\widehat{\mathbb{E}}_{(s,a)\sim\mathcal{D}}\big[f(\hat{\boldsymbol{\mu}}^\pi(s,a)^\top\boldsymbol{\omega}_d)\big]$ to the objective in (9), where f is a differentiable function with closed and convex Fenchel conjugate f_* (see Appendix E.1), and λ is a tunable constant. Furthermore, since $\boldsymbol{\mu}^\pi(s,a)^\top\boldsymbol{\omega}_d = \zeta(s,a) \leqslant C_\infty^\pi$ and $\boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_Q = Q(s,a) \leqslant \frac{1}{1-\gamma}$, we also restrict $\boldsymbol{\theta}_Q$ and $\boldsymbol{\omega}_d$ in regions $\Theta(\hat{\boldsymbol{\phi}}) = \{\boldsymbol{\theta}_Q \mid 0 \leqslant \hat{\boldsymbol{\phi}}(s,a)^\top\boldsymbol{\theta}_Q \leqslant \frac{1}{1-\gamma}\}$ and $\Omega(\hat{\boldsymbol{\mu}}^\pi) = \{\boldsymbol{\omega}_d \mid \hat{\boldsymbol{\mu}}^\pi(s,a)^\top\boldsymbol{\omega}_d \leqslant C_\infty^\pi\}$, respectively.

4 Theoretical Guarantee

In this section, we provide a rigorously theoretical analysis regarding the sample complexity of the proposed SpectralDICE algorithm. For the sake of technical conciseness, we make the following assumption on the candidate family $\mathcal{F}$. We argue that this is not a restrictive assumption, but rather, only helps to highlight the key contributions with simplified analysis.

Assumption 3 (realizability)**.** Assume a finite family $\mathcal{F}$, such that $\langle\hat{\boldsymbol{\phi}}(s,a), d^{\pi_{\mathrm{b}}}(s',a')\hat{\boldsymbol{\mu}}^\pi(s',a')\rangle$ is a valid state-action transition kernel for any $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\mu}}^\pi) \in \mathcal{F}$, and the ground-truth representation $(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^{\pi,\star}) \in \mathcal{F}$.

Representation Learning Error. The key to subsequent analyses is to first bound the error of representation learning, which is of some theoretical interest by itself. Generally speaking, we expect *probably approximately correct* (*PAC*) bounds for representation learning in the following format.

Claim 3. *With probability at least* $1-\delta$*, we have*

$$\mathbb{E}_{(s,a)\sim d_{\mathbb{P}}^{\pi_{\mathrm{b}}}}\left[\left\|\hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a) - \mathbb{P}^\pi(\cdot,\cdot|s,a)\right\|_1\right] \leqslant \xi(|\mathcal{F}|, N, \delta),$$

where $\hat{\mathbb{P}}^\pi(s',a'|s,a) := d_{\mathbb{P}}^{\pi_{\mathrm{b}}}(s',a')\hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(s')$*,* N *is the number of samples in* $\mathcal{D}$*, and the upper bound* ξ *only depends on* $|\mathcal{F}|$*,* N *and* δ*.*

We point out that, under certain regularity assumptions, the above claim can be proven for many spectral representation learning algorithms. Specifically, when RepLearn is implemented by OLS or NCE, we can show that $\xi(|\mathcal{F}|, N, \delta) = \Theta\Big(\sqrt{\frac{1}{N}\log\frac{|\mathcal{F}|}{\delta}}\Big)$.

Policy Evaluation Error. The performance of the proposed SpectralDICE algorithm is evaluated by the *policy evaluation error* $\mathcal{E} := \hat{\rho}(\pi) - \rho_{\mathbb{P}}(\pi)$, which can be bounded by the following theorem.

Theorem 4 (Main Theorem)**.** *Suppose Claim 3 holds for the* RepLearn *subroutine. Then under Assumptions 1 to 3, with probability at least* $1-\delta$*, we have*

$$\mathcal{E} \lesssim \frac{1}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{N}} + \frac{1}{(1-\gamma)^2}\cdot\xi(|\mathcal{F}|, N, \delta/2).$$

Proof sketch. We first split $\mathcal{E}$ into the following terms:

$$\mathcal{E} \;=\; \underbrace{\hat{\rho}(\pi) - \bar{\rho}(\pi)}_{\text{statistical}} \;+\; \underbrace{\bar{\rho}(\pi) - \rho_{\hat{\mathbb{P}}}(\pi)}_{\text{dataset}} \;+\; \underbrace{\rho_{\hat{\mathbb{P}}}(\pi) - \rho_{\mathbb{P}}(\pi)}_{\text{representation}},$$

where we introduce an auxiliary problem:

$$\bar{\rho}(\pi) = \min_{\boldsymbol{\theta}_Q}\max_{\boldsymbol{\omega}_d}\Big\{(1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}\Big[\hat{\boldsymbol{\phi}}(s,a)^\top\boldsymbol{\theta}_Q\Big]$$

$$+ \mathbb{E}_{\substack{s\sim d^{\pi_b}(\cdot),\ a\sim\pi_b(a|s),\\ (s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)}} \Big[\big(\hat{\boldsymbol{\mu}}^\pi(s,a)^\top \boldsymbol{\omega}_d\big) \cdot \big(r(s,a) + \gamma \hat{\boldsymbol{\phi}}(s',a')^\top \boldsymbol{\theta}_Q - \hat{\boldsymbol{\phi}}(s,a)^\top \boldsymbol{\theta}_Q\big) \Big] \Big\}.$$

Note that (9) is the empirical estimation of $\bar{\rho}(\pi)$, and that $\bar{\rho}(\pi)$ is (subtly) inequivalent to $\rho_{\hat{\mathbb{P}}}(\pi)$—the expectation is still taken over $(s',a') \sim \mathbb{P}^\pi(\cdot,\cdot|s,a)$ rather than $\hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a) = \langle \hat{\boldsymbol{\phi}}(s,a), \hat{\boldsymbol{\mu}}^\pi(\cdot,\cdot)\rangle$.

Intuitively, the latter two terms are directly related to the representation learning error established in Claim 3, which can actually be bounded as follows:

$$\rho_{\hat{\mathbb{P}}}(\pi) - \rho_{\mathbb{P}}(\pi) \lesssim \frac{\gamma}{(1-\gamma)^2} \cdot \xi(|\mathcal{F}|, N, \delta/2),$$
$$\bar{\rho}(\pi) - \rho_{\hat{\mathbb{P}}}(\pi) \lesssim \frac{1}{1-\gamma} \cdot \xi(|\mathcal{F}|, N, \delta/2).$$

On the other hand, the first term is only caused by replacing the expectations with their empirical estimators, which can be bounded by concentration inequalities as:

$$\hat{\rho}(\pi) - \bar{\rho}(\pi) \lesssim \frac{1}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{N}}.$$

Plugging these terms back completes the proof. □

Finally, we conclude that the sample complexity of SpectralDICE equipped with either OLS or NCE RepLearn subroutine is $\tilde{O}(N^{-1/2})$ (under mild regularity assumptions). Details are deferred to Appendix D.

5 Experiments

In this section, we present experimental results in both continuous and discrete environments to demonstrate the strength of the proposed SpectralDICE algorithm. We also study the impact of hyperparameters, data coverage and the choice of behavior policy on the OPE performance, and illustrate the efficacy of the proposed representation learning method.

The empirical results show that our method outperforms BestDICE, the state-of-the-art DICE implementation without representation learning, in terms of both the convergence rate and the final prediction error. In comparison to other baselines, SpectralDICE achieves comparable performance with higher efficiency in simple environments, and performs significantly better than others in the most challenging environment.

5.1 Continuous Environments

Setting. We start by comparing SpectralDICE with various baseline OPE methods in literature, including BestDICE [57], Fitted Q Evaluation (FQE) [61], Model-Based (MB) method [62], Importance Sampling (IS) method [13] and Doubly-Robust (DR) method [32]. We follow the experiment protocol in Yang et al. [57] to evaluate and compare the OPE performances of these algorithms in three continuous MuJoCo environments, namely `Cartpole`, `Reacher` and `Half-Cheetah`, in an increasing order of difficulty. In our implementation, for representation learning, we parameterize each of $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\mu}}^\pi$ with a 2-layer feed-forward neural network. For the OPE step, regularizer is appended to (9), and the estimated policy value is retrieved by $\hat{\rho}(\pi) = \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\hat{\boldsymbol{\mu}}^\pi(s,a)^\top \boldsymbol{\omega}_d \cdot r(s,a)]$ (see Remark 2). Both steps are regarded as stochastic optimization problems, and are solved by stochastic gradient descent and stochastic gradient descent-ascent, respectively. Optimization hyperparameters are selected via grid search. Performance is quantified by *OPE error* $|\hat{\rho}(\pi) - \rho(\pi)|$.

Results. The OPE performances of different methods in three environments are shown in Figure 1. It is observed that SpectralDICE achieves comparable performance in fewer optimization steps as compared to all the other baselines, and further, outperforms them in terms of both convergence rate and final estimation error in the most challenging `Half-Cheetah` environment. Further, although

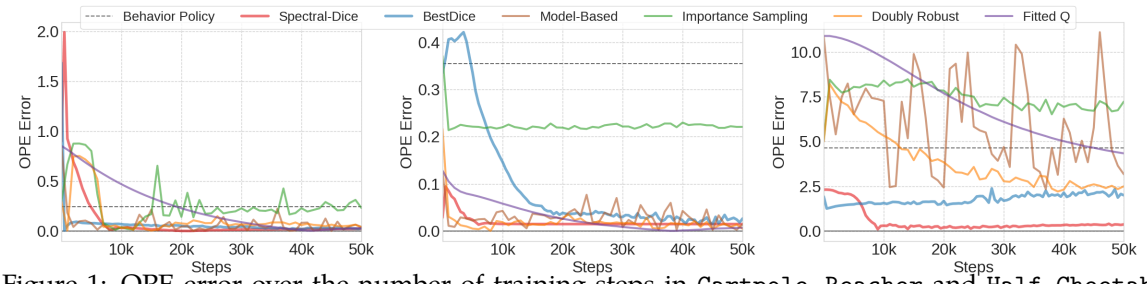

Figure 1: OPE error over the number of training steps in `Cartpole`, `Reacher` and `Half-Cheetah` environments (from left to right). Due to the use of convex-concave formulation, we can see that SpectralDICE converges faster and more stably to the target policy with a smaller OPE error in all three environments.

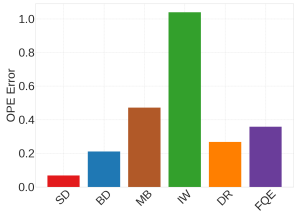

Figure 2: Averaged relative OPE errors over three environments.

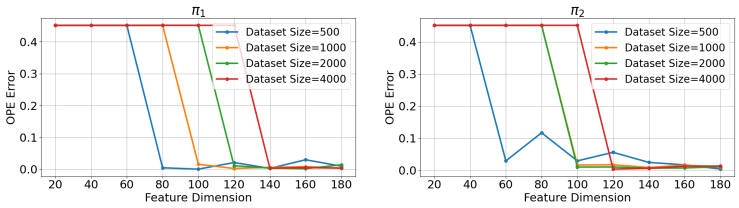

Figure 3: OPE error of SpectralDICE in `Four Rooms` with varying behavior policies ("far-away" policy π_1 vs. "similar" policy π_2), dataset sizes and feature dimensions.

FQE achieves an error close to SpectralDICE in simpler environments, its performance significantly degrades when the transition dynamics becomes more complex, demonstrating the importance and power of spectral representation.

Here we also highlight the comparison between two DICE-based methods—SpectralDICE (ours) and BestDICE. All settings showcase the advantage of our primal-dual spectral representation over the generic neural network representation, which justify the argument that, compared to the non-convex non-concave optimization in vanilla DICE, our convex-concave optimization leads to faster convergence and enhanced stability within a wider range of environments.

For a clearer comparison, we further present the averaged relative OPE error across these three environments in Figure 2. Here the *relative OPE error* is defined by $\frac{|\hat{\rho}(\pi)-\rho(\pi)|}{|\bar{\rho}(\pi_b)-\rho(\pi)|}$, *i.e.*, OPE error normalized by the value difference between the target and behavior policies. Under this metric, it becomes more evident that our method outperforms all the baselines in terms of estimation accuracy by a large margin.

5.2 Discrete Environment

Setting. We proceed to test our method in `Four Rooms` [63], a classical discrete environment featuring convenient visualization, to study the algorithm's sensitivity for hyperparameters and illustrate the efficacy of representation learning. For representation learning in this tabular MDP, we perform singular value decomposition (SVD) of the matrix $\left[\frac{\mathbb{P}^\pi(s',a'|s,a)}{d^\pi(s',a')}\right]$ (indexed by (s,a) and (s',a')) and select the top d singular vectors as $\hat{\boldsymbol{\phi}}(s,a)$ and $\hat{\boldsymbol{\mu}}^\pi(s',a')$.

Sensitivity Study. We study the algorithm's sensitivity with respect to behavior policy π_b, dataset size N and spectral feature dimension d by examining their impact on the OPE performance. For π_b, we vary between two behavior policies π_1 and π_2, where π_1 has a larger ℓ_1-distance from the target policy than π_2. The results are shown in Figure 3. It can be observed that the proposed algorithm is always able to achieve low OPE errors with sufficiently large feature dimensions, showcasing its wide applicability under different behavior policies, data availability and hyperparameters.

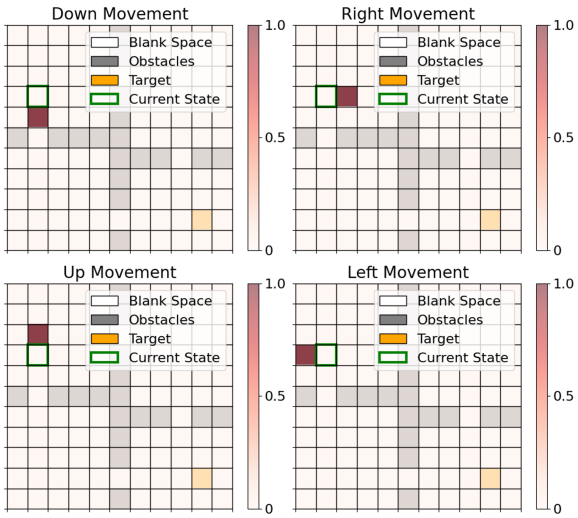

Figure 4: Visualization of the learned transition kernel for a fixed state and all the four actions.

Efficacy of Representation Learning. To give a hint of the efficacy of our representation learning scheme RepLearn, we visualize in Figure 4 the learned transition kernel $\hat{\mathbb{P}}$ for a fixed state and all the four actions, where $\hat{\mathbb{P}}$ is restored from the spectral representation by (4). As shown in the heat map (where darker color indicates higher probability), the RepLearn algorithm successfully learns a set of primal-dual features that accurately encode the correct transition dynamics.

More experimental details are deferred to Appendix A.

6 Conclusion

In this paper, to relieve the intrinsic tension between breaking the curse of horizon and overcoming the curse of dimensionality via DICE estimators, we propose a novel primal-dual spectral representation method that establishes linear spectral representations for both the primal variable (*i.e.*, Q-function) and the dual variable (*i.e.*, stationary distribution correction ratio), which leads to SpectralDICE, an efficient and practical OPE algorithm that eliminates the non-convex non-concave saddle-point optimization in DICE and makes efficient use of historical data. The performance of SpectralDICE is justified by a theoretical sample complexity guarantee and the empirical outperformance. Future directions include taking one step further to design offline policy optimization methods using primal-dual spectral representations, and applying the algorithm for efficient imitation learning.

References

[1] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018.

[2] Gregory Kahn, Adam Villaflor, Pieter Abbeel, and Sergey Levine. Composable action-conditioned predictors: Flexible off-policy learning for robot navigation. In *Conference on robot learning*, pages 806–816. PMLR, 2018.

[3] Tianyu Shi, Dong Chen, Kaian Chen, and Zhaojian Li. Offline reinforcement learning for autonomous driving with safety and exploration enhancement. *arXiv preprint arXiv:2110.07067*, 2021.

[4] Xing Fang, Qichao Zhang, Yinfeng Gao, and Dongbin Zhao. Offline reinforcement learning for autonomous driving with real world driving data. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 3417–3422. IEEE, 2022.

[5] Abhyuday Jagannatha, Philip Thomas, and Hong Yu. Towards high confidence off-policy reinforcement learning for clinical applications. In *CausalML Workshop, ICML*, 2018.

[6] Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

[7] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.

[8] Ahmad Slim, Husain Al Yusuf, Nadine Abbas, Chaouki T Abdallah, Gregory L Heileman, and Ameer Slim. A Markov decision processes modeling for curricular analytics. In *2021 20th IEEE international conference on machine learning and applications (ICMLA)*, pages 415–421. IEEE, 2021.

[9] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

[10] Haoming Jiang, Bo Dai, Mengjiao Yang, Tuo Zhao, and Wei Wei. Towards automatic evaluation of dialog systems: A model-free off-policy evaluation approach. *arXiv preprint arXiv:2102.10242*, 2021.

[11] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

[12] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019.

[13] Josiah Hanna, Scott Niekum, and Peter Stone. Importance sampling policy evaluation with an estimated behavior policy. In *International Conference on Machine Learning*, pages 2605–2613. PMLR, 2019.

[14] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in neural information processing systems*, 32, 2019.

[15] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.

[16] Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

[17] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

[18] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019.

[19] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. AlgaeDICE: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

[20] Geoffrey J Gordon. Stable function approximation in dynamic programming. In *Machine learning proceedings 1995*, pages 261–268. Elsevier, 1995.

[21] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are PAC-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

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

[23] Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

[24] Pulkit Katdare, Nan Jiang, and Katherine Rose Driggs-Campbell. Marginalized importance sampling for off-environment policy evaluation. In *Conference on Robot Learning*, pages 3778–3788. PMLR, 2023.

[25] Fengdi Che, Chenjun Xiao, Jincheng Mei, Bo Dai, Ramki Gummadi, Oscar A Ramirez, Christopher K Harris, A Rupam Mahmood, and Dale Schuurmans. Target networks and over-parameterization stabilize off-policy bootstrapping with function approximation. *arXiv preprint arXiv:2405.21043*, 2024.

[26] Justin Boyan and Andrew Moore. Generalization in reinforcement learning: Safely approximating the value function. *Advances in neural information processing systems*, 7, 1994.

[27] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine learning proceedings 1995*, pages 30–37. Elsevier, 1995.

[28] John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

[29] Ofir Nachum and Bo Dai. Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

[30] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

[31] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[32] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[33] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

[34] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

[35] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

[36] Mengjiao Yang, Bo Dai, Ofir Nachum, George Tucker, and Dale Schuurmans. Offline policy selection under uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 4376–4396. PMLR, 2022.

[37] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020.

[38] 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. PMLR, 2020.

[39] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[40] Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. *arXiv preprint arXiv:2208.09515*, 2022.

[41] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. *Advances in neural information processing systems*, 33:20095–20107, 2020.

[42] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. *arXiv preprint arXiv:2110.04652*, 2021.

[43] Tongzheng Ren, Tianjun Zhang, Csaba Szepesvári, and Bo Dai. A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence*, pages 1686–1696. PMLR, 2022.

[44] Tongzheng Ren, Zhaolin Ren, Na Li, and Bo Dai. Stochastic nonlinear control via finite-dimensional spectral dynamic embedding. *arXiv preprint arXiv:2304.03907*, 2023.

[45] Tongzheng Ren, Chenjun Xiao, Tianjun Zhang, Na Li, Zhaoran Wang, Sujay Sanghavi, Dale Schuurmans, and Bo Dai. Latent variable representation for reinforcement learning. *arXiv preprint arXiv:2212.08765*, 2022.

[46] Hongming Zhang, Tongzheng Ren, Chenjun Xiao, Dale Schuurmans, and Bo Dai. Provable representation with efficient planning for partially observable reinforcement learning. *arXiv preprint arXiv:2311.12244*, 2023.

[47] Shuang Qiu, Lingxiao Wang, Chenjia Bai, Zhuoran Yang, and Zhaoran Wang. Contrastive UCB: Provably efficient contrastive self-supervised learning in online reinforcement learning. In *International Conference on Machine Learning*, pages 18168–18210. PMLR, 2022.

[48] Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making linear MDPs practical via contrastive representation learning. In *International Conference on Machine Learning*, pages 26447–26466. PMLR, 2022.

[49] Dmitry Shribak, Chen-Xiao Gao, Yitong Li, Chenjun Xiao, and Bo Dai. Diffusion spectral representation for reinforcement learning. *arXiv preprint arXiv:2406.16121*, 2024.

[50] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

[51] Chengzhuo Ni, Anru R Zhang, Yaqi Duan, and Mengdi Wang. Learning good state and action representations via tensor decomposition. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1682–1687. IEEE, 2021.

[52] Jonathan Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning Bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pages 2938–2971. PMLR, 2022.

[53] Audrey Huang, Jinglin Chen, and Nan Jiang. Reinforcement learning in low-rank MDPs with density features. In *International Conference on Machine Learning*, pages 13710–13752. PMLR, 2023.

[54] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[55] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block MDPs: A model-free representation learning approach. In *International Conference on Machine Learning*, pages 26517–26547. PMLR, 2022.

[56] Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. CoinDICE: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33:9398–9411, 2020.

[57] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized Lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561, 2020.

[58] Hengshuai Yao, Csaba Szepesvári, Bernardo Avila Pires, and Xinhua Zhang. Pseudo-MDPs and factored linear action models. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 1–9. IEEE, 2014.

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

[60] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

[61] Ilya Kostrikov and Ofir Nachum. Statistical bootstrapping for uncertainty estimation in off-policy evaluation. *arXiv preprint arXiv:2007.13609*, 2020.

[62] Michael R Zhang, Tom Le Paine, Ofir Nachum, Cosmin Paduraru, George Tucker, Ziyu Wang, and Mohammad Norouzi. Autoregressive dynamics models for offline policy evaluation and optimization. *arXiv preprint arXiv:2104.13877*, 2021.

[63] Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211, 1999.

[64] Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

[65] Michel Broniatowski and Amor Keziou. Minimization of ϕ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442, 2006.

[66] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[67] Sergei Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Appendix

A More Experimental Results

Additional Experiments. We evaluate the OPE performance of the proposed SpectralDICE algorithm and the aforementioned baselines (see Section 5.1) in three additional environments, namely `Walker2d`, `Hopper` and `Ant`, the results of which are shown in Figure 5. These additional experiments further justify that our algorithm outperforms all the other baselines in a consistent and robust way, enjoying both a faster convergence rate and a smaller OPE error. These additional experimental results further confirm the superiority of SpectralDICE.

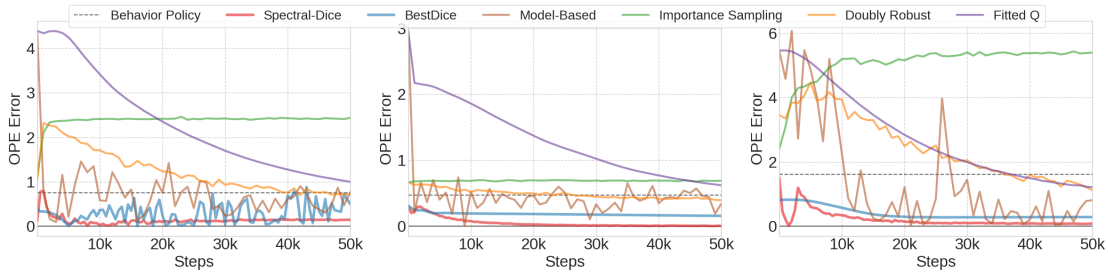

Figure 5: OPE error over the number of training steps in `Walker2d`, `Hopper` and `Ant` (from left to right).

Implementation Details. For the baseline algorithms, we follow the implementation of BestDICE in [57] and the implementations of FQE, MB, IS, DR in [64]. The optimization hyperparameters including learning rate, optimizer parameter, network architecture, batch size, *etc.*, are selected via grid search. All the experiments were conducted using V100 GPUs on a multi-node cluster.

For the continuous environments, the target policy is obtained using deep reinforcement learning agents (Deep Q-Network (DQN) agent for `Cartpole`, and Soft Actor-Critic (SAC) agents for all the other environments). The behavior policy is then obtained by sampling from a Gaussian distribution centered at the mean action of the target policy, where the variance of the Gaussian distribution can be adjusted to get behavior policies at different distances from the target policy. To build the offline dataset, we collect 400 trajectories using the behavior policy, where each trajectory is truncated to 250 steps.

The source code is available at `https://anonymous.4open.science/r/spectral_dice-720A`.

B Primal-Dual Spectral Representation

In this appendix, we present the key properties of the proposed primal-dual spectral representation with proofs, as well as a brief discussion on why the spectral representation of that specific form is preferable.

B.1 DICE Estimator with Primal-Dual Spectral Representation

We first present the proof of Corollary 2 that is already stated in the main text.

Corollary 2. *With primal-dual spectral representation (5) where $q(\cdot) \equiv d^{\pi_b}(\cdot)$, under Assumption 1,*

$$\rho(\pi) = \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \Big\{ (1-\gamma) \mathbb{E}_{\substack{s\sim\mu_0, \\ a\sim\pi(\cdot|s)}} \left[\boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_Q \right] + \mathbb{E}_{\substack{s\sim d^{\pi_b}(\cdot),\ a\sim\pi_b(a|s), \\ s'\sim\mathbb{P}(\cdot|s,a),\ a'\sim\pi(\cdot|s')}} \left[\left(\boldsymbol{\mu}^\pi(s,a)^\top \boldsymbol{\omega}_d \right) \left(r(s,a) + \gamma \boldsymbol{\phi}(s',a')^\top \boldsymbol{\theta}_Q - \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_Q \right) \right] \Big\}.$$

Proof of Corollary 2. Recall the primal-dual LP formulation of policy evaluation stated in (2), which can be equivalently rewritten using the primal-dual spectral representation (5) as follows:

$$\rho(\pi) = \min_{Q(\cdot,\cdot)} \max_{d^\pi(\cdot,\cdot)} \left\{ (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}[Q(s,a)] + \int d^\pi(s,a)\left[r(s,a) + \gamma\mathbb{E}_{\substack{s'\sim\mathbb{P}(\cdot|s,a),\\ a'\sim\pi(\cdot|s')}}[Q(s',a')] - Q(s,a)\right]\mathrm{d}s\mathrm{d}a \right\} \tag{10a}$$

$$= \min_{Q(\cdot,\cdot)} \max_{d^\pi(\cdot,\cdot)} \left\{ (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}[Q(s,a)] + \int q(s)\pi_{\mathrm{b}}(a|s)\cdot\frac{d^\pi(s,a)}{q(s)\pi_{\mathrm{b}}(a|s)}\left[r(s,a) + \gamma\mathbb{E}_{\substack{s'\sim\mathbb{P}(\cdot|s,a),\\ a'\sim\pi(\cdot|s')}}[Q(s',a')] - Q(s,a)\right]\mathrm{d}s\mathrm{d}a \right\} \tag{10b}$$

$$= \min_{Q(\cdot,\cdot)} \max_{d^\pi(\cdot,\cdot)} \left\{ (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}[Q(s,a)] + \mathbb{E}_{\substack{s\sim q(\cdot),\ a\sim\pi_{\mathrm{b}}(a|s),\\ s'\sim\mathbb{P}(\cdot|s,a),\ a'\sim\pi(\cdot|s')}}\left[\frac{d^\pi(s,a)}{q(s)\pi_{\mathrm{b}}(\cdot|s)}\big(r(s,a)+\gamma Q(s',a') - Q(s,a)\big)\right] \right\} \tag{10c}$$

$$= \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \left\{ (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}\left[\boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_Q\right] + \mathbb{E}_{\substack{s\sim d^{\pi_{\mathrm{b}}}(\cdot),\ a\sim\pi_{\mathrm{b}}(\cdot|s),\\ s'\sim\mathbb{P}(\cdot|s,a),\ a'\sim\pi(\cdot|s')}}\left[\boldsymbol{\mu}^\pi(s,a)^\top\boldsymbol{\omega}_d\big(r(s,a)+\gamma\boldsymbol{\phi}(s',a')^\top\boldsymbol{\theta}_Q - \boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_Q\big)\right] \right\}, \tag{10d}$$

where in (10b) we perform the IS-style change-of-variable used in DICE estimators (see (3)); in (10d) we plug in the primal-dual spectral representation of Q^π and d^π stated in (5), as well as the fact that $q(\cdot) \equiv d^{\pi_{\mathrm{b}}}(\cdot)$. □

B.2 Failure of the Naive Spectral Representation

In Section 3.1, it is mentioned that directly applying the naive spectral representation (4) proposed in Ren et al. [40] induces a complicated representation for $\zeta(\cdot,\cdot)$, which in turn leads to an intractable optimization (2) for the computation of the DICE estimator. The above point is further elaborated here in a formal way.

Note that, in Lemma 1, the linear representation of Q^π only builds upon the low-rank MDP assumption, and therefore it still holds with the naive spectral representation (4). Meanwhile, it can be checked that

$$d^\pi(s,a) = \Big\langle q(s)\pi(a|s)\boldsymbol{\mu}(s), \underbrace{(1-\gamma)\boldsymbol{\omega}_0 + \gamma\int d^\pi(\tilde{s},\tilde{a})\boldsymbol{\phi}(\tilde{s},\tilde{a})\mathrm{d}\tilde{s}\mathrm{d}\tilde{a}}_{\boldsymbol{\omega}_d^\pi}\Big\rangle, \tag{11}$$

which can be obtained by plugging the relation $\pi(a|s)\boldsymbol{\mu}(s) = \pi_{\mathrm{b}}(a|s)\boldsymbol{\mu}^\pi(s,a)$ into the linear representation of $d^\pi(\cdot,\cdot)$ to eliminate $\boldsymbol{\mu}^\pi$ from the representation. Consequently, the LP formulation (10) becomes

$$\rho(\pi) = \min_{Q(\cdot,\cdot)} \max_{d^\pi(\cdot)} \left\{ (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}[Q(s,a)] + \int d^\pi(s,a)\left[r(s,a) + \gamma\mathbb{E}_{\substack{s'\sim\mathbb{P}(\cdot|s,a),\\ a'\sim\pi(\cdot|s')}}[Q(s',a')] - Q(s,a)\right]\mathrm{d}s\mathrm{d}a \right\}$$

$$= \min_{Q(\cdot,\cdot)} \max_{d^\pi(\cdot)} \left\{ (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}[Q(s,a)] + \int q(s)\pi_{\mathrm{b}}(a|s)\cdot\frac{\pi(a|s)}{\pi_{\mathrm{b}}(a|s)}\frac{d^\pi(s,a)}{q(s)\pi(a|s)}\left[r(s,a) + \gamma\mathbb{E}_{\substack{s'\sim\mathbb{P}(\cdot|s,a),\\ a'\sim\pi(\cdot|s')}}[Q(s',a')] - Q(s,a)\right]\mathrm{d}s\mathrm{d}a \right\}$$

$$= \min_{Q(\cdot,\cdot)} \max_{d^\pi(\cdot)} \left\{ (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}[Q(s,a)] + \mathbb{E}_{\substack{s\sim q(\cdot),\ a\sim\pi_{\mathrm{b}}(\cdot|s),\\ s'\sim\mathbb{P}(\cdot|s,a),\ a'\sim\pi(\cdot|s')}}\left[\frac{\pi(a|s)}{\pi_{\mathrm{b}}(a|s)}\frac{d^\pi(s,a)}{q(s)\pi(a|s)}\big(r(s,a)+\gamma Q(s',a') - Q(s,a)\big)\right] \right\}$$

$$= \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \left\{ (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}\left[\boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_Q\right] + \mathbb{E}_{\substack{s\sim q(\cdot),\ a\sim\pi_{\mathrm{b}}(\cdot|s),\\ s'\sim\mathbb{P}(\cdot|s,a),\ a'\sim\pi(\cdot|s')}}\left[\frac{\pi(a|s)}{\pi_{\mathrm{b}}(a|s)}\big(\boldsymbol{\mu}(s)^\top\boldsymbol{\omega}_d\big)\big(r(s,a)+\gamma\boldsymbol{\phi}(s',a')^\top\boldsymbol{\theta}_Q - \boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_Q\big)\right] \right\},$$

which involves an unknown ratio $\frac{\pi(a|s)}{\pi_{\mathrm{b}}(a|s)}$ when the behavior policy π_{b} is unknown, and is thus intractable.

The above failed attempt implies that the policy ratio should be "absorbed" into the representation to be implicitly learned during representation learning, which exactly inspires the primal-dual spectral representation (5).

B.3 Solving the Minimax Problem via Regularization

It is known that directly solving (9) leads to potential numerical instability issues due to the objective's linearity in $\boldsymbol{\theta}_Q$ and $\boldsymbol{\omega}_d$ [19]. Fortunately, it is shown in Yang et al. [57] that certain regularization leads to strictly concave inner maximization while keeping the optimal *solution* $\boldsymbol{\omega}_d^\star$ unbiased. Specifically, in practice we may append the following regularizer to the objective in (8):

$$\rho_{\mathrm{reg}}(\pi) = \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \Big\{ (1-\gamma)\mathbb{E}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}\big[\boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_Q\big] + \mathbb{E}_{\substack{s\sim d^{\pi_{\mathrm{b}}}(\cdot),\ a\sim\pi_{\mathrm{b}}(a|s),\\ s'\sim\mathbb{P}(\cdot|s,a),\ a'\sim\pi(\cdot|s')}}\Big[\big(\boldsymbol{\mu}^\pi(s,a)^\top\boldsymbol{\omega}_d\big)\cdot \big(r(s,a) + \gamma\boldsymbol{\phi}(s',a')^\top\boldsymbol{\theta}_Q - \boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_Q\big)\Big] - \lambda\mathbb{E}_{(s,a)\sim\mathcal{D}}\big[f(\hat{\boldsymbol{\mu}}^\pi(s,a)^\top\boldsymbol{\omega}_d)\big] \Big\}. \tag{12}$$

Here f is a differentiable convex function with closed and convex Fenchel conjugate f_* (see Appendix E.1), and $\lambda > 0$ is a tunable constant that controls the magnitude of regularization. It is evident that the regularized objective is concave in $\boldsymbol{\omega}_d$, which facilitates the inner maximization. What's more, it has also be shown that such regularization does not alter the optimal solution $\boldsymbol{\omega}_d^\star$, as summarized below.

Lemma 5 (Nachum et al. [19], Yang et al. [57])**.** *The solution* $(\boldsymbol{\theta}_Q^{\mathrm{reg},\star}, \boldsymbol{\omega}_d^{\mathrm{reg},\star})$ *to* (12) *satisfies:*

$$\begin{aligned}
\boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_Q^{\mathrm{reg},\star} &= \boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_Q^\star - \lambda(\mathcal{I}-\mathcal{P}^\pi)^{-1} f'\Big(\tfrac{d^\pi(s,a)}{d^{\pi_{\mathrm{b}}}(s,a)}\Big),\\
\boldsymbol{\mu}^\pi(s,a)^\top\boldsymbol{\omega}_d^\star &= \boldsymbol{\mu}^\pi(s,a)^\top\boldsymbol{\omega}_d^{\mathrm{reg},\star},\\
\rho_{\mathrm{reg}}(\pi) &= \rho(\pi) - \lambda\mathsf{D}_f(d^\pi\|d^{\pi_{\mathrm{b}}}),
\end{aligned}$$

where $(\boldsymbol{\theta}_Q^\star, \boldsymbol{\omega}_d^\star)$ *is the solution to* (8).

We emphasize that the regularized problem is unbiased only in the sense that $\boldsymbol{\omega}_d^{\mathrm{reg},\star} = \boldsymbol{\omega}_d^\star$. Therefore, in general we need to plug $\boldsymbol{\omega}_d^{\mathrm{reg},\star}$ back into (8) and solve the outer minimization again to recover $\boldsymbol{\theta}_Q^\star$. Nevertheless, when λ is sufficiently small, we shall regard $\boldsymbol{\theta}_Q^{\mathrm{reg},\star} \approx \boldsymbol{\theta}_Q^\star$ to relieve the computational burden.

In practice, we can only solve the empirical version of (12), *i.e.*,

$$\rho_{\mathrm{reg}}(\pi) = \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \Big\{ (1-\gamma)\widehat{\mathbb{E}}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}\big[\boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_Q\big] + \widehat{\mathbb{E}}_{\substack{s\sim\mu_0,\\ a\sim\pi(\cdot|s)}}\Big[\big(\boldsymbol{\mu}^\pi(s,a)^\top\boldsymbol{\omega}_d\big)\cdot \big(r(s,a) + \gamma\boldsymbol{\phi}(s',a')^\top\boldsymbol{\theta}_Q - \boldsymbol{\phi}(s,a)^\top\boldsymbol{\theta}_Q\big)\Big] - \lambda\widehat{\mathbb{E}}_{(s,a)\sim\mathcal{D}}\big[f(\hat{\boldsymbol{\mu}}^\pi(s,a)^\top\boldsymbol{\omega}_d)\big] \Big\}.$$

C Representation Learning Methods and Their Error Bounds

In this appendix, we introduce two candidate methods—*ordinary least squares* (*OLS*) and *noise-contrastive estimation* (*NCE*)—for the RepLearn subroutine. Further, we also provide their representation learning error bounds in the form of Claim 3, which is restated here for readers' convenience:

Claim 3. *With probability at least* $1-\delta$*, the representation learning error of* $\textsc{RepLearn}(\mathcal{F}, \mathcal{D}, \pi)$ *is bounded by*

$$\mathbb{E}_{(s,a)\sim d_{\mathbb{P}}^{\pi_{\mathrm{b}}}}\Big[\big\|\hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a) - \mathbb{P}^\pi(\cdot,\cdot|s,a)\big\|_1\Big] \leqslant \xi(|\mathcal{F}|, N, \delta),$$

where $\hat{\mathbb{P}}^\pi(s',a'|s,a) := d_{\mathbb{P}}^{\pi_{\mathrm{b}}}(s',a')\hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(s')$*, and* N *is the number of samples in* $\mathcal{D}$*.*

It should be emphasized that the two methods discussed here are not the only candidates for RepLearn. Rather, any representation learning method that comes with a learning error bound in the required form is applicable, without any further requirements on the learning mechanism.

C.1 Ordinary Least Sqaures (OLS)

Method. Inspired by Ren et al. [40], the objective of OLS can be constructed as follows. Denote by $\mathbb{Q}^\pi(s',a',s,a) := d^{\pi_\mathrm{b}}(s,a)\mathbb{P}^\pi(s',a'|s,a)$ the joint distribution of state-action transitions under behavior policy π_b. Then we plug $\mathbb{Q}^\pi$ into (5) and rearrange the terms to obtain

$$\frac{\mathbb{Q}^\pi(s',a',s,a)}{\sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}} = \sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}\boldsymbol{\phi}(s,a)^\top\boldsymbol{\mu}^\pi(s',a').$$

Therefore, we propose to optimize over the following OLS objective:

$$\begin{aligned}
&\min_{(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\mu}}^\pi)\in\mathcal{F}} \int \left(\frac{\mathbb{Q}^\pi(s',a',s,a)}{\sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}} - \sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}\hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(s',a') \right)^2 \mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a' \\
= &\min_{(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\mu}}^\pi)\in\mathcal{F}} \Bigg\{ \int \frac{\mathbb{Q}^\pi(s',a',s,a)^2}{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}\mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a' - 2\mathbb{E}_{(s,a)\sim d^{\pi_\mathrm{b}},(s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)}\left[\hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(s',a')\right] \\
&\qquad + \mathbb{E}_{(s,a)\sim d^{\pi_\mathrm{b}},(s',a')\sim d^{\pi_\mathrm{b}}}\left[\left(\hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(s',a')\right)^2\right] \Bigg\},
\end{aligned}$$

Note that the first term $\int \frac{\mathbb{Q}^\pi(s',a',s,a)^2}{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}\mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a'$ is a constant that can be omitted in optimization, while the second and third terms can be effectively approximated by sampling from the dataset $\mathcal{D}$ and the target policy π. Therefore, in practice we learn $(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\mu}}^\pi)$ by solving the following optimization:

$$\min_{(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\mu}}^\pi)\in\mathcal{F}} \left\{ \widehat{\mathbb{E}}_{(s,a)\sim d^{\pi_\mathrm{b}},(\tilde{s}',\tilde{a}')\sim d^{\pi_\mathrm{b}}}\left[\left(\hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(\tilde{s}',\tilde{a}')\right)^2\right] - 2\widehat{\mathbb{E}}_{(s,a)\sim d^{\pi_\mathrm{b}},(s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)}\left[\hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(s',a')\right] \right\}, \tag{13}$$

where the expectations are replaced by their empirical estimations using data sampled from $\mathcal{D}$.

Error Bound. We proceed to show the representation learning error bound for the OLS method, which requires the following regularity assumption on the transition kernel $\mathbb{P}^\pi$ and the occupancy measure d^{π_b}.

Assumption 4 (regularity for OLS)**.** (1) lower-bounded transition kernel: $\mathbb{P}^\pi(s',a'|s,a) \geqslant \frac{1}{C_\mathbb{P}} > 0$, $\forall s,a,s',a'$; (2) effective behavior policy coverage: $\frac{d^{\pi_\mathrm{b}}(s,a)}{d^{\pi_\mathrm{b}}(s',a')} \leqslant C_\mathrm{cov}$, $\forall s,a,s',a'$.

We point out that the major rationale behind these mild assumptions is to rule out the cases where certain transitions are scarcely sampled due to the singularity in transition kernel or behavior policy.

Theorem 6 (OLS learning error)**.** *Under Assumptions 1 to 3 and the additional Assumption 4 for regularity, let $(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\mu}}^\pi)$ be the solution to (13), and set $\hat{\mathbb{P}}^\pi(s',a'|s,a) := d^{\pi_\mathrm{b}}(s',a')\hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(s')$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$, we have*

$$\mathbb{E}_{(s,a)\sim d_\mathbb{P}^{\pi_\mathrm{b}}}\left[\left\|\mathbb{P}^\pi(\cdot,\cdot|s,a) - \hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a)\right\|_1\right] \leqslant \sqrt{C_\mathbb{P}C_\mathrm{reg}}\cdot\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{N}},$$

where $C_\mathrm{reg} = \frac{4}{3}\sqrt{C_\mathrm{cov}} + 8C_\mathrm{cov}$ is a universal constant determined by the PAC bound for OLS..

Proof. We would like to apply the fast-rate PAC bound for OLS regression (Lemma 18). For the sake of clarity, we explicitly define the family of candidate regression functions as

$$\widetilde{\mathcal{F}} := \left\{ f : (s,a,s',a') \mapsto \sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}\boldsymbol{\phi}(s,a)^\top\boldsymbol{\mu}^\pi(s',a') \;\middle|\; (\boldsymbol{\phi},\boldsymbol{\mu}^\pi)\in\mathcal{F} \right\}.$$

It is evident that any $f\in\widetilde{\mathcal{F}}$ is bounded as follows:

$$0 \leqslant f(s,a,s',a') = \sqrt{\frac{d^{\pi_\mathrm{b}}(s,a)}{d^{\pi_\mathrm{b}}(s',a')}}\tilde{\mathbb{P}}^\pi(s',a'|s,a) \leqslant \sqrt{C_\mathrm{cov}},$$

where we use the fact that $\langle\hat{\boldsymbol{\phi}}(s,a), d^{\pi_\mathrm{b}}(s',a')\hat{\boldsymbol{\mu}}^\pi(s',a')\rangle$ is always some valid transition kernel $\tilde{\mathbb{P}}^\pi$ (by Assumption 3), and the additional regularity assumption (Assumption 4). Further, since the

family $\widetilde{\mathcal{F}}$ is realizable (by Assumption 3), there exists an optimal $f^\star \in \widetilde{\mathcal{F}}$ such that

$$f^\star(s,a,s',a') = \frac{\mathbb{Q}^\pi(s',a',s,a)}{\sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}}.$$

As $f(s,a,s',a'), f^\star(s,a,s',a') \in [0, \sqrt{C_{\mathrm{cov}}}]$, we deduce from Lemma 18 that, with probability at least $1-\delta$,

$$\int \Big(f^\star(s,a,s',a') - \hat{f}(s,a,s',a')\Big)^2 \mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a' \leqslant C_{\mathrm{reg}} \cdot \frac{\log(|\mathcal{F}|/\delta)}{N}, \tag{14}$$

where $C_{\mathrm{reg}} := \frac{4}{3}\sqrt{C_{\mathrm{cov}}} + 8C_{\mathrm{cov}}$, and $\hat{f}(s,a,s',a') := \sqrt{d^{\pi_\mathrm{b}}(s,a)d^{\pi_\mathrm{b}}(s',a')}\hat{\boldsymbol{\phi}}(s,a)^\top \hat{\boldsymbol{\mu}}^\pi(s',a')$. Consequently,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d_{\mathbb{P}}^{\pi_\mathrm{b}}}\Big[\big\|\mathbb{P}^\pi(\cdot,\cdot|s,a) - \hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a)\big\|_1\Big] \\
&= \int d_{\mathbb{P}}^{\pi_\mathrm{b}}(s,a)\Big|\mathbb{P}^\pi(s',a'|s,a) - \hat{\mathbb{P}}^\pi(s',a'|s,a)\Big|\mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a' && \text{(15a)} \\
&= \int \Big|\mathbb{Q}^\pi(s',a',s,a) - \hat{\mathbb{Q}}^\pi(s',a',s,a)\Big|\mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a' && \text{(15b)} \\
&\leqslant \sqrt{\int \left(\sqrt{\mathbb{Q}^\pi(s',a',s,a)} - \frac{\hat{\mathbb{Q}}^\pi(s',a',s,a)}{\sqrt{\mathbb{Q}^\pi(s',a',s,a)}}\right)^2 \mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a' \cdot \int \mathbb{Q}^\pi(s',a',s',a')\mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a'} && \text{(15c)} \\
&= \sqrt{\int \frac{d^{\pi_\mathrm{b}}(s',a')}{\mathbb{P}^\pi(s',a'|s,a)}\Big(f^\star(s,a,s',a') - \hat{f}(s,a,s',a')\Big)^2 \mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a'} && \text{(15d)} \\
&\leqslant \sqrt{\max_{s,a,s',a'}\left\{\frac{d^{\pi_\mathrm{b}}(s',a')}{\mathbb{P}^\pi(s',a'|s,a)}\right\}} \cdot \sqrt{C_{\mathrm{reg}} \cdot \frac{\log(|\mathcal{F}|/\delta)}{N}} && \text{(15e)} \\
&\leqslant \sqrt{C_{\mathbb{P}} C_{\mathrm{reg}}} \cdot \sqrt{\frac{\log(|\mathcal{F}|/\delta)}{N}}, && \text{(15f)}
\end{aligned}$$

where in (15b) we use the definition of $\mathbb{Q}^\pi$, and define $\hat{\mathbb{Q}}^\pi := d_{\mathbb{P}}^{\pi_\mathrm{b}}(s,a)\hat{\mathbb{P}}^\pi(s',a'|s,a)$; in (15c) we use Cauchy-Schwartz inequality; in (15d) we use the definition of $\hat{f}$ and $f^\star$; in (15e) we plug in the PAC bound (14); in (15f) we use Assumption 4 to bound the coefficient. This completes the proof. □

C.2 Noise-Constrastive Learning (NCE)

Method. NCE is a widely used method for contrastive representation learning in RL [47, 48]. To learn $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\mu}}^\pi)$, we consider a binary contrastive learning objective [47]:

$$\min_{(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\mu}}^\pi)\in\mathcal{F}} \widehat{\mathbb{E}}_{(s,a)\sim d^{\pi_\mathrm{b}}}\left[\widehat{\mathbb{E}}_{(s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)}\left[\log\left(1 + \frac{1}{\hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(s',a')}\right)\right] + \widehat{\mathbb{E}}_{(s',a')\sim P_{\mathrm{neg}}}\left[\log\left(1 + \hat{\boldsymbol{\phi}}(s,a)^\top\hat{\boldsymbol{\mu}}^\pi(s',a')\right)\right]\right], \tag{16}$$

where P_{neg} is a negative sampling distribution that will be specified with justification later. We highlight that the above objective implicitly guarantees an equal number of positive and negative samples.

The following derivations follow a similar pathway as those in [47]. For notational consistency that facilitates the application of known results, we introduce the following auxiliary notations. Define

$$\widetilde{\mathcal{F}} := \big\{f : (s,a,s',a') \mapsto \boldsymbol{\phi}(s,a)^\top\boldsymbol{\mu}^\pi(s',a') \mid (\boldsymbol{\phi},\boldsymbol{\mu}^\pi) \in \mathcal{F}\big\}.$$

For clarity, we augment the sampled transitions to include a label y indicating whether the sample is positive ($y=1$) or negative ($y=0$). Formally, given a dataset $\mathcal{D} = \{(s_i,a_i,s_i',a_i') \mid i \in [N]\}$ of positive transitions, we randomly sample N negative transitions $(\tilde{s}_i,\tilde{a}_i) \sim P_{\mathrm{neg}}$ ($i \in [N]$, i.i.d.), and define the augmented dataset

$$\widetilde{\mathcal{D}} := \big\{(s_i,a_i,s_i',a_i',1), (s_i,a_i,\tilde{s}_i,\tilde{a}_i,0) \;\big|\; i \in [N]\big\}.$$

In this way, the NCE objective (16) can be equivalently rewritten (in MLE format) as

$$\max_{f\in\widetilde{\mathcal{F}}} \widehat{\mathbb{E}}_{(s,a,s',a',y)\sim d^{\widetilde{\mathcal{D}}}}[\log \psi_f(s,a,s',a',y)], \tag{17}$$

where the likelihood function ψ_f is defined by

$$\psi_f(s,a,s',a',y) := \left(\frac{f(s,a,s',a')}{1+f(s,a,s',a')}\right)^y \cdot \left(\frac{1}{1+f(s,a,s',a')}\right)^{1-y}.$$

We point out that $\psi_f(s,a,s',a',\cdot) \in \Delta(\mathcal{Y})$ for any (s,a,s',a'), where $\mathcal{Y} := \{0,1\}$. In fact, given $f^\star$ that optimizes the unconstrained non-empirical version of (17), $\psi_{f^\star}$ can be interpreted as the probability of obtaining label y given (s,a,s',a'), as summarized in the following lemma that is similar to Lemma C.1 in Qiu et al. [47].

Lemma 7 (non-empirical solution to NCE)**.** *The optimal solution to the unconstrained non-empirical version of* (17), i.e., $f^\star := \max_f \mathbb{E}_{(s,a,s',a',y)\sim d^{\widetilde{\mathcal{D}}}}[\log \psi_f(s,a,s',a',y)]$, *is characterized by*

$$f^\star(s,a,s',a') = \frac{\mathbb{P}^\pi(s',a'|s,a)}{P_{\text{neg}}(s',a')}.$$

Proof. Note that the objective can be rewritten as

$$\begin{aligned}
&\mathbb{E}_{(s,a,s',a',y)\sim d^{\widetilde{\mathcal{D}}}}[\log \psi_f(s,a,s',a',y)] \\
=& \int d^{\widetilde{\mathcal{D}}}(s,a,s',a') \left(\sum_{y\in\mathcal{Y}} \Pr(y|s,a,s',a') \log \psi_f(s,a,s',a',y)\right) \mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a' \\
=& -\int d^{\widetilde{\mathcal{D}}}(s,a,s',a') \cdot \mathsf{H}\big(\Pr(y|s,a,s',a'); \psi_f(s,a,s',a',y)\big) \mathrm{d}s\mathrm{d}a\mathrm{d}s'\mathrm{d}a'.
\end{aligned}$$

Here $\mathsf{H}(\cdot;\cdot)$ is the cross entropy between distributions, which, by Gibbs' inequality, is minimized only when

$$\Pr(y|s,a,s',a') = \psi_{f^\star}(s,a,s',a',y) = \left(\frac{f^\star(s,a,s',a')}{1+f^\star(s,a,s',a')}\right)^y \cdot \left(\frac{1}{1+f^\star(s,a,s',a')}\right)^{1-y}. \tag{18}$$

On the other hand, Bayes' rule states that (note that $\Pr(y|s,a) = \frac{1}{2}, \forall y \in \mathcal{Y}$):

$$\Pr(y=1|s,a,s',a') = \frac{\Pr(s',a'|s,a,y=1)\Pr(y=1|s,a)}{\sum_{y\in\mathcal{Y}} \Pr(s',a'|s,a,y)\Pr(y|s,a)} = \frac{\mathbb{P}^\pi(s',a'|s,a)}{P_{\text{neg}}(s',a') + \mathbb{P}^\pi(s',a'|s,a)}. \tag{19}$$

Comparing (18) and (19) gives

$$\frac{f^\star(s,a,s',a')}{1+f^\star(s,a,s',a')} = \frac{\mathbb{P}^\pi(s',a'|s,a)}{P_{\text{neg}}(s',a') + \mathbb{P}^\pi(s',a'|s,a)} \quad\Longrightarrow\quad f^\star(s,a,s',a') = \frac{\mathbb{P}^\pi(s',a'|s,a)}{P_{\text{neg}}(s',a')}.$$

This completes the proof. □

Remark 3. For conciseness, here we slightly abuse the notation $\Pr(\cdot)$ to denote the distribution (density or mass) of joint and conditional distributions involving random variables $(s,a,s',a',y) \sim d^{\widetilde{\mathcal{D}}}$. Specifically, we write $\Pr(\cdots, x, \cdots)$ to indicate an arbitrary value x taken by the random variable, and we also write $\Pr(\cdots, x = x_0, \cdots)$ to emphasize the value x_0 taken by that random variable.

Lemma 7 is important in that it echoes the form of primal-dual spectral representation in (5). Specifically, we shall take $P_{\text{neg}}(\cdot,\cdot) \equiv d^{\pi_{\text{b}}}(\cdot,\cdot)$ for an exact match, which is also implementable using offline data since $d^{\pi_{\text{b}}}$ can be effectively approximated by sampling the trajectories. We will stick to this choice of P_{neg} from now on.

Error Bound. We proceed to show the representation learning error bound for the NCE method, which requires the following regularity assumption on the negative sampling distribution P_{neg}, or equivalently, as per the choice above, the state-action occupancy measure $d^{\pi_{\text{b}}}(\cdot,\cdot)$ for the behavior policy π_{b}.

Assumption 5 (regularity for NCE). $d^{\pi_b}_{\mathbb{P}}(s,a) \geqslant \frac{1}{C_d} > 0, \forall s, a.$

We point out that Assumption 5 is a standard assumption for the negative sampling distribution [47], aiming at eliminating the cases where certain transitions are scarcely drawn as negative samples and thus obstruct efficient representation learning for those cases. The assumption is also slightly stronger than the effective behavior policy coverage assumption required by the OLS method (see Assumption 4).

Theorem 8 (NCE learning error). *Under Assumptions 1 to 3 and the additional Assumption 5 for regularity, let* $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\mu}}^\pi)$ *be the solution to* (16) *with* $P_{\mathrm{neg}}(\cdot,\cdot) \equiv d^{\pi_b}(\cdot,\cdot)$, *and set* $\hat{\mathbb{P}}^\pi(s',a'|s,a) := d^{\pi_b}(s',a')\hat{\boldsymbol{\phi}}(s,a)^\top \hat{\boldsymbol{\mu}}^\pi(s')$. *Then, for any* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *we have*

$$\mathbb{E}_{(s,a)\sim d^{\pi_b}_{\mathbb{P}}}\left[\left\|\mathbb{P}^\pi(\cdot,\cdot|s,a) - \hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a)\right\|_1\right] \leqslant 2\sqrt{2}(1+C_d)\cdot\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{N}}.$$

The proof of Theorem 8 largely follows the same pathway and techniques established in Qiu et al. [47]. Nevertheless, our proof is less technically involved since the offline non-episodic setting significantly weakens the correlation between samples. For the sake of completeness, we restate the proof below.

Proof. We start by observing $\Pr(y,s',a'|s,a) := \Pr(y|s,a,s',a')\Pr(s',a'|s,a)$, where $\Pr(s',a'|s,a)$ can in turn be calculated using Bayes' rule as follows:

$$\begin{aligned}\Pr(s',a'|s,a) &= \Pr(s',a'|s,a,y=0)\Pr(y=0|s,a) + \Pr(s',a'|s,a,y=1)\Pr(y=1|s,a)\\ &= \tfrac{1}{2}\big(\mathbb{P}^\pi(s',a'|s,a) + P_{\mathrm{neg}}(s',a')\big).\end{aligned} \tag{20}$$

Here we use the fact that the data distribution $d^{\widetilde{\mathcal{D}}}$ implicitly assigns an equal number of labels as $y=0$ and $y=1$ by the design of NCE objective (16). Since $\Pr(s',a'|s,a)$ is a constant that is independent from f, we can further rewrite the NCE objective to be

$$\arg\max_{f\in\widetilde{\mathcal{F}}}\left\{\widehat{\mathbb{E}}_{(s,a,s',a',y)\in\widetilde{\mathcal{D}}}\left[\log \Pr\nolimits_f(y|s,a,s',a')\right]\right\} = \arg\max_{f\in\widetilde{\mathcal{F}}}\left\{\widehat{\mathbb{E}}_{(s,a,s',a',y)\in\widetilde{\mathcal{D}}}\left[\log \Pr\nolimits_f(y,s',a'|s,a)\right]\right\}, \tag{21}$$

where we define the shorthand notations

$$\begin{aligned}\Pr\nolimits_f(y|s,a,s',a') &:= \left(\frac{f(s,a,s',a')}{1+f(s,a,s',a')}\right)^y \cdot \left(\frac{1}{1+f(s,a,s',a')}\right)^{1-y},\\ \Pr\nolimits_f(y,s',a'|s,a) &:= \left(\frac{f(s,a,s',a')\Pr(s',a'|s,a)}{1+f(s,a,s',a')}\right)^y \cdot \left(\frac{\Pr(s',a'|s,a)}{1+f(s,a,s',a')}\right)^{1-y}\end{aligned}$$

for any $f \in \widetilde{\mathcal{F}}$. Note that the right-hand side of (21) is in the desired MLE form, with ground-truth conditional density $\Pr_{f^\star}(y,s',a'|s,a)$ specified by some $f^\star \in \widetilde{\mathcal{F}}$, thanks to the realizability assumption (Assumption 3). Now, using the PAC bound for MLE shown in Agarwal et al. [41] (see Lemma 19), we have

$$\sum_{i=1}^N \mathbb{E}_{(s_i,a_i)\sim d^{\pi_b}_{\mathbb{P}}}\left[\left\|\Pr\nolimits_{\hat{f}}(\cdot,\cdot,\cdot|s_i,a_i) - \Pr\nolimits_{f^\star}(\cdot,\cdot,\cdot|s_i,a_i)\right\|_1^2\right] \leqslant 8\log(|\mathcal{F}|/\delta)$$

Since all (s_i,a_i) pairs are sampled i.i.d. from the same distribution d^{π_b}, we shall further conclude that

$$\mathbb{E}_{(s,a)\sim d^{\pi_b}_{\mathbb{P}}}\left[\left\|\Pr\nolimits_{\hat{f}}(\cdot,\cdot,\cdot|s,a) - \Pr\nolimits_{f^\star}(\cdot,\cdot,\cdot|s,a)\right\|_1^2\right] \leqslant \frac{8\log(|\mathcal{F}|/\delta)}{N}. \tag{22}$$

We proceed to further relate (22) with the desired format. For this purpose, note that

$$\begin{aligned}&\left\|\Pr\nolimits_{\hat{f}}(\cdot,\cdot,\cdot|s,a) - \Pr\nolimits_{f^\star}(\cdot,\cdot,\cdot|s,a)\right\|_1\\ =\ &\left\|\Pr\nolimits_{\hat{f}}(y=1,\cdot,\cdot|s,a) - \Pr\nolimits_{f^\star}(y=1,\cdot,\cdot|s,a)\right\|_1 + \left\|\Pr\nolimits_{\hat{f}}(y=0,\cdot,\cdot|s,a) - \Pr\nolimits_{f^\star}(y=0,\cdot,\cdot|s,a)\right\|_1\end{aligned} \tag{23a}$$

$$= 2\left\| \frac{\Pr(\cdot,\cdot|s,a)}{1+\hat{f}(s,a,\cdot,\cdot)} - \frac{\Pr(\cdot,\cdot|s,a)}{1+f^\star(s,a,\cdot,\cdot)} \right\|_1 \tag{23b}$$

$$= 2\int \frac{\left|\hat{f}(s,a,s',a') - f^\star(s,a,s',a')\right| \cdot \Pr(s',a'|s,a)}{\big(1+\hat{f}(s,a,s',a')\big)\big(1+f^\star(s,a,s',a')\big)} \mathrm{d}s'\mathrm{d}a', \tag{23c}$$

where in (23a) we use the definition of L^1-norm; in (23b) we use the fact that

$$\Pr_{\hat{f}}(y,\cdot,\cdot|s,a) - \Pr_{f^\star}(y,\cdot,\cdot|s,a) = (-1)^y \left(\frac{\Pr(\cdot,\cdot|s,a)}{1+\hat{f}(s,a,\cdot,\cdot)} - \frac{\Pr(\cdot,\cdot|s,a)}{1+f^\star(s,a,\cdot,\cdot)} \right).$$

Now, plugging Lemma 7 and (20) into the integrand in (23c), we have

$$\frac{\left|\hat{f}(s,a,s',a') - f^\star(s,a,s',a')\right| \cdot \Pr(s',a'|s,a)}{\big(1+\hat{f}(s,a,s',a')\big)\big(1+f^\star(s,a,s',a')\big)}$$

$$= \frac{\left|\mathbb{P}^\pi(s',a'|s,a)/P_{\mathrm{neg}}(s',a') - \hat{f}(s,a,s',a')\right| \cdot \frac{1}{2}\big(\mathbb{P}^\pi(s',a'|s,a) + P_{\mathrm{neg}}(s',a')\big)}{\big(1+\hat{f}(s,a,s',a')\big)\big(1+\mathbb{P}^\pi(s',a'|s,a)/P_{\mathrm{neg}}(s',a')\big)} \tag{24a}$$

$$= \frac{\left|\mathbb{P}^\pi(s',a'|s,a) - P_{\mathrm{neg}}(s',a')\hat{f}(s,a,s',a')\right|}{2\big(1+\hat{f}(s,a,s',a')\big)} \tag{24b}$$

$$\geqslant \frac{\left|\mathbb{P}^\pi(s',a'|s,a) - P_{\mathrm{neg}}(s',a')\hat{f}(s,a,s',a')\right|}{2(1+C_d)}, \tag{24c}$$

where we use the upper bound $\hat{f}(s,a,s',a') = \hat{\mathbb{P}}^\pi(s',a'|s,a)/d^{\pi_\mathrm{b}}(s',a') \leqslant C_d$ in (24c). Hence,

$$\left\|\mathbb{P}^\pi(\cdot,\cdot|s,a) - \hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a)\right\|_1$$

$$= \int \left|\mathbb{P}^\pi(s',a'|s,a) - P_{\mathrm{neg}}(s',a')\hat{f}(s,a,s',a')\right| \mathrm{d}s'\mathrm{d}a' \tag{25a}$$

$$\leqslant 2(1+C_d)\int \frac{\left|\hat{f}(s,a,s',a') - f^\star(s,a,s',a')\right| \cdot \Pr(s',a'|s,a)}{\big(1+\hat{f}(s,a,s',a')\big)\big(1+f^\star(s,a,s',a')\big)} \mathrm{d}s'\mathrm{d}a' \tag{25b}$$

$$= (1+C_d)\left\|\Pr_{\hat{f}}(\cdot,\cdot,\cdot|s,a) - \Pr_{f^\star}(\cdot,\cdot,\cdot|s,a)\right\|_1, \tag{25c}$$

where we use $\hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a) = P_{\mathrm{neg}}(\cdot,\cdot)\hat{f}(s,a,\cdot,\cdot)$ in (25a), (24) in (25b), and (23) in (25c). Finally,

$$\mathbb{E}_{(s,a)\sim d_{\mathbb{P}}^{\pi_\mathrm{b}}}\left[\left\|\mathbb{P}^\pi(\cdot,\cdot|s,a) - \hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a)\right\|_1\right]$$

$$\leqslant \sqrt{\mathbb{E}_{(s,a)\sim d_{\mathbb{P}}^{\pi_\mathrm{b}}}\left[\left\|\mathbb{P}^\pi(\cdot,\cdot|s,a) - \hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a)\right\|_1^2\right]} \tag{26a}$$

$$\leqslant \sqrt{(1+C_d)^2\mathbb{E}_{(s,a)\sim d_{\mathbb{P}}^{\pi_\mathrm{b}}}\left[\left\|\Pr_{\hat{f}}(\cdot,\cdot,\cdot|s,a) - \Pr_{f^\star}(\cdot,\cdot,\cdot|s,a)\right\|_1^2\right]} \tag{26b}$$

$$\leqslant 2\sqrt{2}(1+C_d)\cdot\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{N}}, \tag{26c}$$

where we use Cauchy-Schwartz inequality in (26a), (25) in (26b), and (22) in (26c). □

D Sample Complexity Guarantee

In this appendix, we derive the sample complexity guarantee for the proposed SpectralDICE algorithm, assuming a known bound on the representation learning error induced by the RepLearn subroutine (see Claim 3). As discussed in the main text, the objective is to bound the estimation error $\mathcal{E} := \hat{\rho}(\pi) - \rho(\pi)$, which can be intuitively split into the following three terms that are easier to bound:

$$\mathcal{E} \quad = \quad \underbrace{\hat{\rho}(\pi) - \bar{\rho}(\pi)}_{\text{statistical error}} + \underbrace{\bar{\rho}(\pi) - \rho_{\hat{\mathbb{P}}}(\pi)}_{\text{dataset error}} + \underbrace{\rho_{\hat{\mathbb{P}}}(\pi) - \rho_{\mathbb{P}}(\pi)}_{\text{representation error}}.$$

We point out that the statistical error results from replacing the expectation with empirical estimates, the dataset error comes from the offline dataset that samples transitions from the true transition kernel $\mathbb{P}^\pi$ instead of the learned kernel $\hat{\mathbb{P}}^\pi$, and the representation error accounts for the error induced by plugging in the learned representation $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\mu}}^\pi)$ instead of the ground truth $(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^{\pi,\star})$ into the DICE estimator.

As described in the proof sketch, for the rest of this appendix, we provide an upper bound for each of these three terms, and eventually conclude with an overall sample complexity guarantee.

Representation Error. We start by bounding the representation error term, which by intuition should be a direct consequence of the representation learning error shown in Claim 3.

Lemma 9. *Conditioned on the event that the inequality in Claim 3 holds, under Assumptions 1 to 3,*

$$\rho_{\hat{\mathbb{P}}}(\pi) - \rho_{\mathbb{P}}(\pi) \leqslant \frac{\gamma C^\pi_\infty}{(1-\gamma)^2} \cdot \xi(|\mathcal{F}|, N, \delta).$$

Proof. By the well-known Simulation Lemma (see Lemma 20), we have

$$\begin{aligned}
&\rho_{\hat{\mathbb{P}}}(\pi) - \rho_{\mathbb{P}}(\pi) \\
&= \frac{\gamma}{1-\gamma} \mathbb{E}_{(s,a)\sim d^\pi_{\mathbb{P}}} \left[\mathbb{E}_{s'\sim\hat{\mathbb{P}}(\cdot|s,a)} \left[V^\pi_{\hat{\mathbb{P}}}(s') \right] - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[V^\pi_{\hat{\mathbb{P}}}(s') \right] \right] && \text{(27a)} \\
&= \frac{\gamma}{1-\gamma} \mathbb{E}_{(s,a)\sim d^\pi_{\mathbb{P}}} \left[\mathbb{E}_{(s',a')\sim\hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a)} \left[Q^\pi_{\hat{\mathbb{P}}}(s',a') \right] - \mathbb{E}_{(s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)} \left[Q^\pi_{\hat{\mathbb{P}}}(s',a') \right] \right] && \text{(27b)} \\
&= \frac{\gamma}{1-\gamma} \int d^\pi_{\mathbb{P}}(s,a) \mathrm{d}s\mathrm{d}a \int Q^\pi_{\hat{\mathbb{P}}}(s',a') \Big(\hat{\mathbb{P}}^\pi(s',a'|s,a) - \mathbb{P}^\pi(s',a'|s,a) \Big) \mathrm{d}s'\mathrm{d}a' && \text{(27c)} \\
&\leqslant \frac{\gamma}{(1-\gamma)^2} \int C^\pi_\infty d^{\pi_{\mathrm{b}}}_{\mathbb{P}}(s,a) \mathrm{d}s\mathrm{d}a \int \left| \mathbb{P}^\pi(s',a'|s,a) - \hat{\mathbb{P}}^\pi(s',a'|s,a) \right| \mathrm{d}s'\mathrm{d}a' && \text{(27d)} \\
&= \frac{\gamma C^\pi_\infty}{(1-\gamma)^2} \mathbb{E}_{(s,a)\sim d^{\pi_{\mathrm{b}}}_{\mathbb{P}}} \left[\left\| \mathbb{P}^\pi(s',a'|s,a) - \hat{\mathbb{P}}^\pi(s',a'|s,a) \right\|_1 \right] && \text{(27e)} \\
&\leqslant \frac{\gamma C^\pi_\infty}{(1-\gamma)^2} \cdot \xi(|\mathcal{F}|, N, \delta), && \text{(27f)}
\end{aligned}$$

where in (27a) we use the Simulation Lemma; in (27b) we use the relationship between value functions; in (27d) we plug in $d^\pi_{\mathbb{P}}(s,a) \leqslant C^\pi_\infty d^{\pi_{\mathrm{b}}}_{\mathbb{P}}(s,a)$ (Assumption 2) and the fact that $Q^\pi_{\hat{\mathbb{P}}}(\cdot,\cdot) \leqslant \frac{1}{1-\gamma}$; in (27f) we use Claim 3. □

Dataset Error. The dataset error can be accounted for by a bounded difference in the objective function, which turns out to be another consequence of the representation learning error. For this purpose, we first show the following technical lemma that formalizes the above intuition.

Lemma 10. $\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} F_1(\boldsymbol{x},\boldsymbol{y}) - \min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} F_2(\boldsymbol{x},\boldsymbol{y}) \leqslant \max_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{y}\in\mathcal{Y}} \left| F_1(\boldsymbol{x},\boldsymbol{y}) - F_2(\boldsymbol{x},\boldsymbol{y}) \right|.$

Proof. Let $\varepsilon := \max_{\boldsymbol{x},\boldsymbol{y}} \left| F_1(\boldsymbol{x},\boldsymbol{y}) - F_2(\boldsymbol{x},\boldsymbol{y}) \right|$. Then we have

$$\begin{aligned}
\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} F_1(\boldsymbol{x},\boldsymbol{y}) &\leqslant \min_{\boldsymbol{x}} \left\{ \max_{\boldsymbol{y}} F_2(\boldsymbol{x},\boldsymbol{y}) + \max_{\boldsymbol{y}} \{ F_1(\boldsymbol{x},\boldsymbol{y}) - F_2(\boldsymbol{x},\boldsymbol{y}) \} \right\} && \text{(28a)} \\
&\leqslant \min_{\boldsymbol{x}} \left\{ \max_{\boldsymbol{y}} F_2(\boldsymbol{x},\boldsymbol{y}) + \varepsilon \right\} && \text{(28b)} \\
&= \min_{\boldsymbol{x}} \max_{\boldsymbol{y}} F_2(\boldsymbol{x},\boldsymbol{y}) + \varepsilon, && \text{(28c)}
\end{aligned}$$

where in (28a) we use the fact that $\max_{\boldsymbol{y}} \{ f(\boldsymbol{y}) + g(\boldsymbol{y}) \} \leqslant \max_{\boldsymbol{y}} f(\boldsymbol{y}) + \max_{\boldsymbol{y}} g(\boldsymbol{y})$. □

Now we are ready to show the following lemma regarding dataset error.

Lemma 11. *Conditioned on the event that the inequality in Claim 3 holds, under Assumptions 1 to 3,*

$$\bar{\rho}(\pi) - \rho_{\hat{\mathbb{P}}}(\pi) \leqslant \frac{C^\pi_\infty}{1-\gamma} \cdot \xi(|\mathcal{F}|, N, \delta).$$

Proof. For the sake of clarity, denote the optimization objectives of $\bar{\rho}(\pi)$ and $\rho_{\hat{\mathbb{P}}}(\pi)$ as follows:

$$\bar{\rho}(\pi) = \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \bar{F}(\boldsymbol{\theta}_Q, \boldsymbol{\omega}_d), \qquad \rho_{\hat{\mathbb{P}}}(\pi) = \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \hat{F}(\boldsymbol{\theta}_Q, \boldsymbol{\omega}_d).$$

Then we can show that

$$\begin{aligned}
|\bar{\rho}(\pi) - \rho_{\hat{\mathbb{P}}}(\pi)| &= \left| \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \bar{F}(\boldsymbol{\theta}_Q, \boldsymbol{\omega}_d) - \min_{\boldsymbol{\theta}_Q} \max_{\boldsymbol{\omega}_d} \hat{F}(\boldsymbol{\theta}_Q, \boldsymbol{\omega}_d) \right| \leqslant \left| \bar{F}(\boldsymbol{\theta}_Q, \boldsymbol{\omega}_d) - \hat{F}(\boldsymbol{\theta}_Q, \boldsymbol{\omega}_d) \right| && \text{(29a)} \\
&= \left| \int d_{\mathbb{P}}^{\pi_{\mathrm{b}}}(s,a) \Big(\mathbb{P}^{\pi}(s',a'|s,a) - \hat{\mathbb{P}}^{\pi}(s',a'|s,a) \Big) \big(\hat{\boldsymbol{\mu}}^{\pi}(s,a)^{\top} \boldsymbol{\omega}_d \big) \cdot \right. \\
&\qquad \left. \Big(r(s,a) + \gamma \hat{\boldsymbol{\phi}}(s',a')^{\top} \boldsymbol{\theta}_Q - \hat{\boldsymbol{\phi}}(s,a)^{\top} \boldsymbol{\theta}_Q \Big) \mathrm{d}s \mathrm{d}a \mathrm{d}s' \mathrm{d}a' \right| && \text{(29b)} \\
&\leqslant \int d_{\mathbb{P}}^{\pi_{\mathrm{b}}}(s,a) \big| \mathbb{P}^{\pi}(s',a'|s,a) - \hat{\mathbb{P}}^{\pi}(s',a'|s,a) \big| \cdot \big| \hat{\boldsymbol{\mu}}^{\pi}(s,a)^{\top} \boldsymbol{\omega}_d \big| \cdot \\
&\qquad \big| r(s,a) + \gamma \hat{\boldsymbol{\phi}}(s',a')^{\top} \boldsymbol{\theta}_Q - \hat{\boldsymbol{\phi}}(s,a)^{\top} \boldsymbol{\theta}_Q \big| \mathrm{d}s \mathrm{d}a \mathrm{d}s' \mathrm{d}a' && \text{(29c)} \\
&\leqslant \mathbb{E}_{(s,a) \sim d_{\mathbb{P}}^{\pi_{\mathrm{b}}}} \left[\big\| \mathbb{P}^{\pi}(\cdot,\cdot|s,a) - \hat{\mathbb{P}}^{\pi}(\cdot,\cdot|s,a) \big\|_1 \cdot C_{\infty}^{\pi} \cdot \tfrac{1}{1-\gamma} \right] && \text{(29d)} \\
&= \frac{C_{\infty}^{\pi}}{1-\gamma} \cdot \xi(|\mathcal{F}|, N, \delta), && \text{(29e)}
\end{aligned}$$

where in (29a) we use Lemma 10; in (29c) we use the integral triangle inequality; in (29d) we plug in $\big|\hat{\boldsymbol{\mu}}^{\pi}(s,a)^{\top} \boldsymbol{\omega}_d\big| \leqslant C_{\infty}^{\pi}$ and $\big| r(s,a) + \gamma \hat{\boldsymbol{\phi}}(s',a')^{\top} \boldsymbol{\theta}_Q - \hat{\boldsymbol{\phi}}(s,a)^{\top} \boldsymbol{\theta}_Q \big| \leqslant \frac{1}{1-\gamma}$ (see Remark 2); in (29e) we use Claim 3. □

Statistical Error. Finally, the statistical error is caused by replacing the expectations with their empirical estimations, which can be bounded by Hoeffding's concentration inequality (see Lemma 16).

Lemma 12. *Under Assumptions 1 to 3, with probability at least $1-\delta$, we have*

$$\hat{\rho}(\pi) - \bar{\rho}(\pi) \leqslant \frac{C_{\infty}^{\pi}}{1-\gamma} \sqrt{\frac{\log(1/2\delta)}{2N}}.$$

Proof. For clarity, label the samples in $\mathcal{D}$ as $\mathcal{D} = \{(s_i, a_i, s_i', a_i') \mid i \in [N]\}$, and define

$$F(s,a,s',a') := \big(\hat{\boldsymbol{\mu}}^{\pi}(s,a)^{\top} \boldsymbol{\omega}_d \big) \big(r(s,a) + \gamma \hat{\boldsymbol{\phi}}(s',a')^{\top} \boldsymbol{\theta}_Q - \hat{\boldsymbol{\phi}}(s,a)^{\top} \boldsymbol{\theta}_Q \big).$$

Note that $\big|\hat{\boldsymbol{\mu}}^{\pi}(s,a)^{\top} \boldsymbol{\omega}_d\big| \leqslant C_{\infty}^{\pi}$ and $\big| r(s,a) + \gamma \hat{\boldsymbol{\phi}}(s',a')^{\top} \boldsymbol{\theta}_Q - \hat{\boldsymbol{\phi}}(s,a)^{\top} \boldsymbol{\theta}_Q \big| \leqslant \frac{1}{1-\gamma}$ (see Remark 2), we have

$$\big| F(s,a,s',a') \big| \leqslant \frac{C_{\infty}^{\pi}}{1-\gamma}, \ \forall s,a,s',a'.$$

Therefore, by Hoeffding's inequality (see Lemma 16), we conclude that

$$\Pr\left[\left| \frac{1}{N} \sum_{i=1}^{N} F(s_i,a_i,s_i',a_i') - \mathbb{E}_{\substack{s \sim d^{\pi_{\mathrm{b}}}(\cdot),\ a \sim \pi_{\mathrm{b}}(a|s), \\ s' \sim \mathbb{P}(\cdot|s,a),\ a' \sim \pi(\cdot|s')}} [F(s,a,s',a')] \right| > t \right] \leqslant 2 \exp\left(- \frac{2Nt^2}{4(C_{\infty}^{\pi})^2/(1-\gamma)^2} \right).$$

Or equivalently, with probability at least $1-\delta$, we have

$$\left| \widehat{\mathbb{E}}_{\substack{(s,a,s') \sim \mathcal{D}, \\ a' \sim \pi(\cdot|s')}} [F(s,a,s',a')] - \mathbb{E}_{\substack{s \sim d^{\pi_{\mathrm{b}}}(\cdot),\ a \sim \pi_{\mathrm{b}}(a|s), \\ s' \sim \mathbb{P}(\cdot|s,a),\ a' \sim \pi(\cdot|s')}} [F(s,a,s',a')] \right| \leqslant \frac{C_{\infty}^{\pi}}{1-\gamma} \sqrt{\frac{\log(1/2\delta)}{2N}}.$$

Finally, the conclusion follows from Lemma 10 using the same argument as above. □

Conclusion. Now we are ready to prove the Main Theorem.

Theorem 4. *Suppose Claim 3 holds for the* RepLearn$(\mathcal{F}, \mathcal{D}, \pi)$ *subroutine. Then under Assumptions 1 to 3, with probability at least $1-\delta$, we have*

$$\mathcal{E} \leqslant \frac{C_{\infty}^{\pi}}{1-\gamma} \sqrt{\frac{\log(1/\delta)}{2N}} + \frac{C_{\infty}^{\pi}}{(1-\gamma)^2} \cdot \xi(|\mathcal{F}|, N, \delta/2).$$

Proof. Consider the following high-probability events:

$$\mathcal{C}_1 : \mathbb{E}_{(s,a)\sim d_{\hat{\mathbb{P}}}^{\pi_b}}\left[\left\|\hat{\mathbb{P}}^{\pi}(\cdot,\cdot|s,a) - \mathbb{P}^{\pi}(\cdot,\cdot|s,a)\right\|_1\right] \leqslant \xi(|\mathcal{F}|, N, \delta/2),$$

$$\mathcal{C}_2 : \left|\widehat{\mathbb{E}}_{\substack{(s,a,s')\sim\mathcal{D},\\ a'\sim\pi(\cdot|s')}}[F(s,a,s',a')] - \mathbb{E}_{\substack{s\sim d^{\pi_b}(\cdot),\ a\sim\pi_b(a|s),\\ s'\sim\mathbb{P}(\cdot|s,a),\ a'\sim\pi(\cdot|s')}}[F(s,a,s',a')]\right| \leqslant \frac{C_\infty^\pi}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{2N}}.$$

As per Claim 3 and Lemma 12, we know $\Pr[\mathcal{C}_i] \geqslant 1-\delta/2$ $(i=1,2)$. Hence by Union Bound,

$$\Pr[\mathcal{C}_1 \cap \mathcal{C}_2] \geqslant 1-\delta.$$

On the other hand, conditioned on $\mathcal{C}_1 \cap \mathcal{C}_2$, Lemma 9, Lemma 11 and Lemma 12 in combination guarantee that

$$\begin{aligned}
\mathcal{E} &= \hat{\rho}(\pi) - \bar{\rho}(\pi) + \bar{\rho}(\pi) - \rho_{\hat{\mathbb{P}}}(\pi) + \rho_{\hat{\mathbb{P}}}(\pi) - \rho_{\mathbb{P}}(\pi) \\
&\leqslant \frac{C_\infty^\pi}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{2N}} + \frac{C_\infty^\pi}{1-\gamma}\cdot\xi(|\mathcal{F}|,N,\delta/2) + \frac{\gamma C_\infty^\pi}{(1-\gamma)^2}\xi(|\mathcal{F}|,N,\delta/2) \\
&= \frac{C_\infty^\pi}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{2N}} + \frac{C_\infty^\pi}{(1-\gamma)^2}\cdot\xi(|\mathcal{F}|,N,\delta/2)
\end{aligned}$$

This completes the proof. □

For completeness, we also include the corollaries of the Main Theorem that characterize the overall sample complexity of our SpectralDICE algorithm using OLS and NCE representation learning methods.

Corollary 13 (sample complexity of OLS-based SpectralDICE)**.** *Under Assumptions 1 to 3 and the additional Assumption 4 for regularity, let $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\mu}}^\pi)$ be the solution to the OLS problem (13). Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\mathcal{E} \leqslant \frac{C_\infty^\pi}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{2N}} + \frac{C_\infty^\pi\sqrt{C_{\mathbb{P}}C_{\mathrm{reg}}}}{(1-\gamma)^2}\cdot\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{N}} \lesssim \frac{1}{(1-\gamma)^2}\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{N}},$$

where $C_{\mathrm{reg}} = \frac{4}{3}\sqrt{C_{\mathrm{cov}}} + 8C_{\mathrm{cov}}$ is a universal constant determined by the PAC bound for OLS..

Corollary 14 (sample complexity of NCE-based SpectralDICE)**.** *Under Assumptions 1 to 3 and the additional Assumption 5 for regularity, let $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\mu}}^\pi)$ be the solution to the NCE problem (16) with $P_{\mathrm{neg}}(\cdot,\cdot) \equiv d^{\pi_b}(\cdot,\cdot)$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\mathcal{E} \leqslant \frac{C_\infty^\pi}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{2N}} + \frac{2\sqrt{2}C_\infty^\pi(1+C_d)}{(1-\gamma)^2}\cdot\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{N}} \lesssim \frac{1}{(1-\gamma)^2}\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{N}}.$$

Remark 4 (Sampling the dataset). Throughout this paper, we have been slightly abusing the notation $(s,a,s') \sim \mathcal{D}$, which is a little subtle in practice since only trajectories (rather than transitions) are collected. To ensure the correct data distribution $d^{\mathcal{D}}(s,a) = d^{\pi_b}(s,a)$, we shall first randomly sample the trajectories, within which we sample each transition $(s_t, a_t, s_{t+1}, a_{t+1})$ with probability $(1-\gamma)\gamma^t$.

E Technical Lemmas

In this final appendix, we include all the technical lemmas used in the previous sections.

E.1 f-Divergence

Definition 1 (f-divergence)**.** Let $\mathbb{P}$ and $\mathbb{Q}$ be two probabilities distribution over a sample space $\mathcal{X}$, such that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. Given a convex function $f : \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ such that $f(1) = 0$ and $f(0) := \lim_{t\to 0^+} f(t)$. Then the *$f$-divergence* of $\mathbb{P}$ with respect to $\mathbb{Q}$ is defined as

$$\mathsf{D}_f(\mathbb{P}\|\mathbb{Q}) := \int_{\mathcal{X}} f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{Q}.$$

The following *variational representation* of f-divergences is well-known in literature.

Lemma 15 (variational representation using Fenchel conjugate)**.** *Let $\mathcal{F}$ denote the class of measurable real valued functions on $\mathcal{X}$ that is absolutely integratable with respect to $\mathbb{Q}$. Then*

$$\mathsf{D}_f(\mathbb{P}\|\mathbb{Q}) = \sup_{g\in\mathcal{F}}\Big\{\mathbb{E}_{x\sim\mathbb{P}}[g(x)] - \mathbb{E}_{x\sim\mathbb{Q}}[f_*(g(x))]\Big\},$$

where f_* is the Fenchel conjugate of f. Further, if f is differentiable, then the optimal dual variable is given by

$$g^\star(x) = f'\Big(\tfrac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big) \quad\Longrightarrow\quad \mathsf{D}_f(\mathbb{P}\|\mathbb{Q}) = \mathbb{E}_{x\sim\mathbb{P}}\Big[f'\Big(\tfrac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big)\Big] - \mathbb{E}_{x\sim\mathbb{Q}}\Big[f_*\Big(f'\Big(\tfrac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big)\Big)\Big]$$

Proof. See Theorem 4.4 in Broniatowski and Keziou [65]. □

E.2 Concentration Inequalities

Lemma 16 (Hoeffding's inequality, Hoeffding [66])**.** *Let $X_1, X_2, \cdots, X_N$ be i.i.d. random variables with mean μ and taking values in $[a, b]$ almost surely. Then for any $\varepsilon > 0$ we have*

$$\Pr\left[\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| > \varepsilon\right] \leqslant 2\exp\left(-\frac{2N\varepsilon^2}{(b-a)^2}\right).$$

In other words, with probability at least $1-\delta$, we have

$$\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \leqslant (b-a)\sqrt{\frac{\log(1/2\delta)}{2N}}.$$

Lemma 17 (Bernstein's inequality, Bernstein [67])**.** *Let $X_1, X_2, \cdots, X_N$ be i.i.d. random variables with mean μ, variance σ^2, and bounded range $|X_i - \mu| \leqslant B$ almost surely. Then with probability at least $1-\delta$, we have*

$$\pm\left(\frac{1}{N}\sum_{i=1}^N X_i - \mu\right) \leqslant \sigma\sqrt{\frac{2\log(1/\delta)}{N}} + \frac{B\log(1/\delta)}{3N}.$$

E.3 Statistical Learning: PAC Bounds

In this section, we present the standard PAC bounds for OLS and MLE. Although these are both classic results, we fail to trace back to the original literature of the former, and thus provide a short proof here for completeness.

Lemma 18 (PAC bound for OLS, fast rate)**.** *Consider a regression problem over a finite family $\mathcal{F} = \{f : \mathcal{X} \to [a, b]\}$ of bounded functions with data distribution $(X, Y) \sim \mathcal{M}$, where the objective is to solve for*

$$\arg\min_{f\in\mathcal{F}} \mathcal{L}(f), \quad \textit{where } \mathcal{L}(f) := \mathbb{E}_{(X,Y)\sim\mathcal{M}}\big[(f(X) - Y)^2\big].$$

Suppose the regression function $f^\star(x) := \mathbb{E}[Y \mid X = x] \in \mathcal{F}$ (realizability), and we have access to i.i.d. sample $(x_i, y_i) \sim \mathcal{M}, \forall i \in [N]$. Let the Empirical Risk Minimization (ERM) estimator be

$$\hat{f} := \arg\min_{f\in\mathcal{F}} \hat{\mathcal{L}}(f), \quad \textit{where } \hat{\mathcal{L}}(f) := \frac{1}{N}\sum_{i=1}^N (f(x_i) - y_i)^2.$$

Then, with probability at least $1-\delta$, the ERM estimator induces a regret that is at most

$$\mathcal{L}(\hat{f}) \leqslant \mathcal{L}(f^\star) + C_{\mathrm{reg}}\frac{\log(|\mathcal{F}|/\delta)}{N}.$$

Suppose further that the ground truth is deterministic such that $y = f^\star(x)$ for some $f^\star \in \mathcal{F}$, in which case we have

$$\mathcal{L}(\hat{f}) \leqslant C_{\mathrm{reg}}\frac{\log(|\mathcal{F}|/\delta)}{N}.$$

Here $C_{\mathrm{reg}} = 8(b-a)^2 + \frac{4}{3}(b-a)$ is a universal constant depending only on the range $[a, b]$.

Proof. Define a random variable $Z_i := (f(X_i) - Y_i)^2 - (f^\star(X_i) - Y_i)^2$, such that

$$\mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}[Z_i(f)] = \mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}\left[(f(X_i) - Y_i)^2 - (f^\star(X_i) - Y_i)^2\right] \tag{30a}$$

$$= \mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}\left[\left((f(X_i) - f^\star(X_i)) + (f^\star(X_i) - Y_i)\right)^2 - (f^\star(X_i) - Y_i)^2\right] \tag{30b}$$

$$= \mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}\left[(f(X_i) - f^\star(X_i))^2\right] + 2\mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}[(f(X_i) - f^\star(X_i))(f^\star(X_i) - Y_i)] \tag{30c}$$

$$= \mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}\left[(f(X_i) - f^\star(X_i))^2\right] =: \mathcal{E}(f), \tag{30d}$$

where in (30c) we use the following fact $\mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}[(f(X_i) - f^\star(X_i))(f^\star(X_i) - Y_i)] = \mathbb{E}_{X_i}[(f(X_i) - f^\star(X_i)) \cdot \mathbb{E}_{Y_i\sim\mathcal{M}(\cdot|X_i)}[f^\star(X_i) - Y_i]] = 0$ that directly follows from the definition of $f^\star$. Similarly, for any $t \in [T]$,

$$\mathrm{Var}_{(X_i,Y_i)\sim\mathcal{M}}[Z_i(f)] = \mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}\left[Z_i(f)^2\right] - \left(\mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}[Z_i(f)]\right)^2 \tag{31a}$$

$$\leqslant \mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}\left[\left((f(X_i) - Y_i)^2 - (f^\star(X_i) - Y_i)^2\right)^2\right] \tag{31b}$$

$$= \mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}\left[(f(X_i) - f^\star(X_i))^2(f(X_i) + f^\star(X_i) - 2Y_i)^2\right] \tag{31c}$$

$$\leqslant 4(b-a)^2\mathbb{E}_{(X_i,Y_i)\sim\mathcal{M}}\left[(f(X_i) - f^\star(X_i))^2\right] = 4(b-a)^2\mathcal{E}(f). \tag{31d}$$

where in (31a) we simply drop the second term, and in (31c) we use the fact $f(X_i) + f^\star(X_i) - 2Y_i \in [-2(b-a), 2(b-a)]$ as $f(X_i), f^\star(X_i), Y_i \in [a, b]$. Further, for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $f \in \mathcal{F}$, we have $f(x) - y \in [-(b-a), b-a]$, implying $Z_i(f) \in [-(b-a), b-a]$ and $\mathbb{E}[Z_i(f)] \in [-(b-a), (b-a)]$. Therefore, $|Z_i(f) - \mathbb{E}[Z_i(f)]| \leqslant 2(b-a)$. Then by Bernstein's inequality (Lemma 17), we conclude that, with probability at least $1-\delta$,

$$\mathbb{E}[Z_i(f)] - \frac{1}{N}\sum_{i=1}^N Z_i(f) \leqslant \sqrt{\mathrm{Var}[Z_i(f)]}\sqrt{\frac{2\log(1/\delta)}{N}} + \frac{2(b-a)\log(1/\delta)}{3N}. \tag{32}$$

To proceed, plug (30) and (31) into (32), and we have

$$\mathcal{E}(f) - (\hat{\mathcal{L}}(f) - \hat{\mathcal{L}}(f^\star)) \leqslant 2(b-a)\sqrt{\mathcal{E}(f)}\sqrt{\frac{2\log(1/\delta)}{N}} + \frac{2(b-a)\log(1/\delta)}{3N} \tag{33a}$$

$$\leqslant \left(\frac{1}{2}\mathcal{E}(f) + \frac{4(b-a)^2\log(1/\delta)}{N}\right) + \frac{2(b-a)\log(1/\delta)}{3N}, \tag{33b}$$

where in (33a) we apply the AM-GM inequality. Finally, we rearrange the terms to obtain

$$\mathcal{E}(f) \leqslant 2(\hat{\mathcal{L}}(f) - \hat{\mathcal{L}}(f^\star)) + \frac{C_{\mathrm{reg}}\log(1/\delta)}{N} \tag{34}$$

for any fixed $f \in \mathcal{F}$, with probability at least $1-\delta$. Finally, we take the union bound with respect to all $f \in \mathcal{F}$, such that with probability at least $1-\delta$, we have

$$\mathcal{E}(f) \leqslant 2(\hat{\mathcal{L}}(f) - \hat{\mathcal{L}}(f^\star)) + \frac{C_{\mathrm{reg}}\log(|\mathcal{F}|/\delta)}{N},\ \forall f \in \mathcal{F}. \tag{35}$$

In particular, (35) also applies to the ERM estimator $\hat{f}$, which gives

$$\mathcal{E}(\hat{f}) \leqslant 2(\hat{\mathcal{L}}(\hat{f}) - \hat{\mathcal{L}}(f^\star)) + \frac{C_{\mathrm{reg}}\log(|\mathcal{F}|/\delta)}{N} \leqslant \frac{C_{\mathrm{reg}}\log(|\mathcal{F}|/\delta)}{N}. \tag{36}$$

Here we use the inequality $\hat{\mathcal{L}}(\hat{f}) \leqslant \hat{\mathcal{L}}(f^\star)$, as $\hat{f}$ minimizes $\hat{\mathcal{L}}(\cdot)$ within $\mathcal{F}$. This completes the proof. □

Lemma 19 (PAC bound for MLE, Agarwal et al. [41])**.** *Consider a conditional probability estimation problem over a finite family $\mathcal{F} = \{f : (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}\}$, where the objective is to estimate $f^\star(x, y) := \mathbb{P}(y|x)$. Suppose the ground truth $f^\star \in \mathcal{F}$ (realizability), and we have access to (potentially correlated) samples $\{(x_i, y_i) \mid i \in [N]\}$ such that $x_i \sim \mathcal{D}_i$ ($\mathcal{D}_i$ is allowed to depend on $(x_{1:i-1}, y_{1:i-1})$, forming a martigale process) and $y_i \sim \mathbb{P}(\cdot|x_i)$. Let the Maximum Likelihood Estimator (MLE) be*

$$\hat{f} := \arg\max_{f\in\mathcal{F}} \sum_{i=1}^N \log f(x_i, y_i).$$

Then, with probability at least $1-\delta$, the error of the MLE estimator is bounded as follows:

$$\sum_{i=1}^{N} \mathbb{E}_{x\sim\mathcal{D}_i}\left[\left\|\hat{f}(x,\cdot) - f^\star(x,\cdot)\right\|_1^2\right] \leqslant 8\log(|\mathcal{F}|/\delta).$$

Specifically, when $\{(x_i, y_i) \mid i \in [N]\}$ are i.i.d. samples from a dataset $\mathcal{D}$, we have

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\left\|\hat{f}(x,\cdot) - f^\star(x,\cdot)\right\|_1^2\right] \leqslant \frac{8\log(|\mathcal{F}|/\delta)}{N}.$$

E.4 Simulation Lemma in MDPs

The following Simulation Lemma is a simplified version of Lemma 21 in Uehara et al. [42].

Lemma 20 (Simulation Lemma)**.** *Given two MDPs $(\mathbb{P}, r)$ and $(\hat{\mathbb{P}}, r)$, for any policy $\pi \in \Pi$, we have*

$$\rho_{\hat{\mathbb{P}}}(\pi) - \rho_{\mathbb{P}}(\pi) = \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi_{\mathbb{P}}}\left[\mathbb{E}_{s'\sim\hat{\mathbb{P}}(\cdot|s,a)}\left[V^\pi_{\hat{\mathbb{P}}}(s')\right] - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[V^\pi_{\hat{\mathbb{P}}}(s')\right]\right].$$

Proof. Note that, for any uniformly bounded function $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{E}_{s\sim\mu_0, a\sim\pi(\cdot|s)}[f(s,a)] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty}\left(\gamma^t f(s_t,a_t) - \gamma^{t+1} f(s_{t+1},a_{t+1})\right) \,\middle|\, s_0\sim\mu_0, a_0\sim\pi(\cdot|s_0)\right] \\
&= \frac{1}{1-\gamma}\sum_{s,a} f(s,a)\cdot\mathbb{E}_{\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty}\left(\gamma^t \mathbb{1}\{s_t=s, a_t=a\} - \gamma^{t+1}\mathbb{1}\{s_{t+1}=s, a_{t+1}=a\}\right) \,\middle|\, s_0\sim\mu_0, a_0\sim\pi(\cdot|s_0)\right] \\
&= \frac{1}{1-\gamma}\sum_{s,a} f(s,a)\cdot\left(d^\pi_{\mathbb{P}}(s,a) - \gamma\sum_{\tilde{s},\tilde{a}} d^\pi_{\mathbb{P}}(\tilde{s},\tilde{a})\mathbb{P}^\pi(s,a|\tilde{s},\tilde{a})\right) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi_{\mathbb{P}}}\left[f(s,a) - \gamma\mathbb{E}_{(s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)}[f(s',a')]\right].
\end{aligned}$$

Therefore, since $\rho_{\mathbb{P}}(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi_{\mathbb{P}}}[r(s,a)]$ and $\rho_{\hat{\mathbb{P}}}(\pi) = \mathbb{E}_{s\sim\mu_0,a\sim\pi(\cdot|s)}[Q^\pi_{\hat{\mathbb{P}}}(s,a)]$, we have

$$\begin{aligned}
\rho_{\hat{\mathbb{P}}}(\pi) - \rho_{\mathbb{P}}(\pi) &= \mathbb{E}_{s\sim\mu_0,a\sim\pi(\cdot|s)}[Q^\pi_{\hat{\mathbb{P}}}(s,a)] - \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi_{\mathbb{P}}}[r(s,a)] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi_{\mathbb{P}}}\left[Q^\pi_{\hat{\mathbb{P}}}(s,a) - \gamma\mathbb{E}_{(s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)}\left[Q^\pi_{\hat{\mathbb{P}}}(s',a')\right] - r(s,a)\right] \\
&= \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi_{\mathbb{P}}}\left[\mathbb{E}_{(s',a')\sim\hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a)}\left[Q^\pi_{\hat{\mathbb{P}}}(s',a')\right] - \mathbb{E}_{(s',a')\sim\mathbb{P}^\pi(\cdot,\cdot|s,a)}\left[Q^\pi_{\hat{\mathbb{P}}}(s',a')\right]\right],
\end{aligned}$$

where in the last equality we plug in the Bellman equation

$$Q^\pi_{\hat{\mathbb{P}}}(s,a) = r(s,a) + \gamma\mathbb{E}_{(s',a')\sim\hat{\mathbb{P}}^\pi(\cdot,\cdot|s,a)}[Q^\pi_{\hat{\mathbb{P}}}(s',a')].$$

Finally, we leverage the relationship between Q- and V-functions to complete the proof. □