Stochastic Nonlinear Control via Finite-dimensional Spectral Dynamics Embedding

Zhaolin Ren*,[1], Tongzheng Ren*,[2], Haitong Ma[1], Na Li[1], and Bo Dai[3]

***Abstract*— This paper presents an approach, Spectral Dynamics Embedding Control (SDEC), to optimal control for nonlinear stochastic systems. This method leverages an infinite-dimensional feature to linearly represent the state-action value function and exploits finite-dimensional truncation approximation for practical implementation. To characterize the effectiveness of these finite-dimensional approximations, we provide an in-depth theoretical analysis to characterize the *approximation error* induced by the finite-dimension truncation and *statistical error* induced by finite-sample approximation in both policy evaluation and policy optimization. Our analysis includes two prominent kernel approximation methods–truncations onto random features and Nystrom features. Empirically, our algorithm performs favorably against existing stochastic control algorithms on several benchmark problems.**

***Index Terms*— Stochastic nonlinear control, reinforcement learning, dynamic programming**

I. Introduction

Stochastic optimal nonlinear control—i.e. finding an optimal feedback policy to maximize cumulative rewards for a stochastic nonlinear system—has been a long-standing challenging problem in control literature [1], [2]. Various control techniques have been developed for nonlinear control, including gain scheduling [3], [4], feedback linearization [5], iterative linear-quadratic regulator [6], sliding model control [7], geometric control [8], back stepping [9], control Lyapunov functions [10], model-predictive control [11], and tools that leverage inequality approximation and optimization-based methods like sum-of-squares (SOS) programming [12]. Nonlinear control often focuses on the stability of closed-loop systems, while control optimality analysis is often heuristic or limited to special classes of systems. Moreover, these methods either lead to highly suboptimal solutions, can only be applied to a subclass of nonlinear systems that satisfy special conditions, or require a large amount of computation and thus cannot handle large-size problems.

To take advantage of the rich theory and tools developed for linear systems, kernel-based linearization has recently regained attraction. The representative approaches includes Koopman operator theory [13], [14], kernelized nonlinear regulator (KNR) [15], reproducing kernel Hilbert space (RKHS) dynamics embedding control [16], and optimal control with occupation kernel [17], among others. The Koopman operator lifts states into an infinite-dimensional space of known measurement functions, where the dynamics become linear in the new space. However, depending on the system itself, possible coupling nonlinear terms between state and control would still appear. Alternatively, KNR [15] studies the sample complexity of learning control for an unknown nonlinear system whose dynamics are assumed to lie in a pre-given RKHS whose dimension can be infinite, and hence are *linear* with respect to the feature maps for states and actions but neglect the computation challenges induced by the infinite dimension and the nonlinear features. [16] considers Markov decision processes (MDP) and represents the conditional transition probability of a MDP in a *pre-defined* RKHS so that calculations involved in solving the MDP could be done via inner products in the infinite-dimensional RKHS. [17] introduce the Liouville's equation in occupation kernel space to represent the trajectories, with which finding the optimal value can be reformulated as linear programming in the infinite-dimensional space under some strict assumptions.

Although kernelized linearization has brought a promising new perspective to nonlinear control, these representative approaches fall short in both computational and theoretical respects. *Computationally*, control in an infinite-dimensional space is intractable. Hence, a finite-dimensional approximation is necessary. Though data-driven computational procedures for kernel selection and RKHS reparametrization have been proposed in these existing methods, they are inefficient in the sense that **i)**, the dynamics or trajectories are *presumed* to be lying in some RKHS with a *pre-defined* finitely-approximated kernel, which is a very strict assumption; in fact, finding good kernel representations for the dynamics is a challenging task; and **ii)**, the dynamics information for kernelization is *only* exploited through samples, and other structure information in the dynamics are ignored even when the dynamics formula is known explicitly. Meanwhile, *theoretically*, the optimality of control with finite-dimensional approximations—*i.e.*, the policy value gap between the finite-dimensional approximation and optimal policy in an infinite-dimensional RKHS—has largely been ignored and not been rigorously analyzed.

Recently, [18] provided a novel kernel linearization method, spectral dynamics embedding, by establishing the connection between stochastic nonlinear control models and linear Markov Decision Processes (MDPs), which exploits the random noise property to factorize the transition probability operator, and induce an infinite-dimensional space for *linearly* representing the *state-action value function* for any *arbitrary*

*Equal contributions [1] Z. Ren, H. Ma and N. Li are with Harvard University, Email: zhaolinren@g.harvard.edu, haitongma@g.harvard.edu, nali@seas.harvard.edu; [1]T. Ren is with University of Texas, Austin, Email: tongzheng@utexas.edu; [3]B. Dai is with Google DeepMind and Georgia Tech, Email: bodai@{google.com, cc.gatech.edu}.

This work is done when T. Ren was a student researcher at Google DeepMind. Z. Ren, Ma, and Li are funded by NSF AI institute: 2112085, NSF ECCS: 2328241, NSF CNS: 2003111, and NIH R01LM014465

policy. Spectral dynamics embedding bypasses the drawbacks of the existing kernel linearization methods in the sense that **i)** the kernel is *automatically induced* by the system dynamics, which avoids the difficulty in deciding the kernel features, utilizes the knowledge of the dynamics and noise model and eliminates the modeling approximation induced by a predefined kernel; and **ii)** the kernel linearization and its finite feature approximation is *computational-friendly* with well-studied performance guarantees [19], [20]. The superiority of spectral dynamics embedding has been justified empirically in the reinforcement learning setting [18] where system dynamics are unknown. However, there still lacks any end-to-end control algorithm with theoretical guarantees that utilizes the finite-dimensional approximation of spectral dynamics embedding, motivating the present work.

Our contributions. In this paper, we close the gap by justifying the finite random and Nyström feature approximation of spectral dynamics embedding in nonlinear control rigorously, enabling the practical usage of the kernelized linear representation in control. Our contributions lie in the following folds.

We first formalize a computational tractable stochastic nonlinear control algorithm with finite-dimension truncation of spectral dynamics embedding representation, *Spectral Dynamics Embedding Control (SDEC)*. Specifically, we first extract the finite-dimensional spectral dynamics embedding through both Monte-Carlo random and Nyström approximations as explained in Section III-A, and then, we conduct the dynamic programming for value function estimation upon these representations through least square policy evaluation. The policy is improved by natural policy gradient for optimal control based on the obtained value functions. The concrete algorithm is derived in Section III-B. We note that the particular way of policy evaluation and policy update is actually compatible with cutting-edge deep reinforcement learning methods, such as Soft Actor-Critic [21], by using our proposed representation to approximate the critic function. In other words, one of the novelties of SDEC is that it exploits the known nonlinear dynamics to obtain a natural, inherent representation space that could be adopted by various dynamical programming or policy gradient-based methods for control.

To characterize the effectiveness of these finite-dimensional approximations, we provide an in-depth theoretical analysis to characterize the *approximation error* induced by the finite-dimension truncation and *statistical error* induced by finite-sample approximation in both policy evaluation and policy optimization. To the best of our knowledge, the analysis is the first time this has been done, due to the challenging complications in stochastic nonlinear control. Specifically, we show the gap between optimal policy and the SDEC induced policy is inversely proportional to a polynomial dependency w.r.t. number of features and the number of samples used in dynamic programming. Lastly, we conduct a numerical study on several robotic control problems to justify our theoretical analysis in Section V.

We also would like to note that the preliminary work has been published in [22]. This paper has several key differences with the conference version in [22]. First, this paper has more detailed proofs and more exposition to motivate the results. Second, this paper includes study of the use of Nyström features in the kernel approximation, which was not considered in [22]; moreover, we prove a novel high probability kernel approximation error bound for Nyström features (Lemma 11), which may be of independent interest. Third, we prove theoretical justification of the validity of two core technical assumptions key to our analysis (Assumptions 2 and 3), which is an important contribution over the conference version in [22], where these assumptions were merely assumed to be true. Finally, this paper has a significantly more fleshed out numerical section, with results on more challenging benchmarks such as 2D Drones, Pendubot balancing, and Cartpole, and comparisons to energy-based control as well as standard Soft-Actor-Critic [21], a state-of-the-art RL algorithm. Our performance metrics include quantitative as well as qualitative evaluation.

A. Related Work in Reinforcement Learning (RL)

Presumably, recently developed model-free deep RL methods for unknown systems, *e.g.*, [21], [23], could be also applied to optimal nonlinear control by treating the nonlinear dynamics as a generative or simulation model to generate data. The success of deep RL is largely attributed to the flexibility of deep neural networks for policy or value function modeling. However, such flexibility also brings difficulty in control by making the optimization extremely non-convex, and thus, may lead to sub-optimal policy, wasting the modeling power.

Though our proposed SDEC shares some similar aspects of actor-critic family of algorithms in RL, in the sense that SDEC also exploits policy improvement (actor) with state-action value function estimation (critic). However, we emphasize the major difference is that SDEC takes advantage of the spectral dynamics embedding for both state-value function and policy, instead of an arbitrary network parameterization. This not only reduces sample complexity by exploiting the known nonlinear dynamics structure but also bypasses the non-convexity difficulty in optimization, allowing us to provide rigorous theoretical guarantees with only mild assumptions. We note that while our work builds on the concept of linear MDPs [24], such linear MDPs merely refer to MDPs whose transition admits a linear factorization, which is distinct from the linear programming formulation of solving MDPs [25], [26].

II. Problem Setup and Preliminaries

In this section, we introduce the stochastic nonlinear control problem that will be studied in this paper and reformulate it as a MDP. We will also briefly introduce the background knowledge about reproducing kernel Hilbert space (RKHS) and random features.

A. Stochastic Nonlinear Control Problem in MDPs

We consider the standard discrete-time nonlinear control model with γ-discounted infinite horizon, defined by

$$s_{t+1} = f(s_t, a_t) + \epsilon_t, \quad \text{where} \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2 I_d), \tag{1}$$

such that $\gamma \in (0,1)$, $s \in \mathcal{S} = \mathbb{R}^d$ is the state, $a \in \mathcal{A}$ is the control action, and $\{\epsilon_t\}_{t=1}^{\infty}$ are independent Gaussian noises. The function $f(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ describes the general nonlinear dynamics, and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ gives the reward

function on the state and action pair. Without loss of generality, we assume there is a fixed initial state s_0. Given a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ with $\Delta(\mathcal{A})$ as the space of probability measures over $\mathcal{A}$, the accumulated reward over infinite horizon is given by

$$J^\pi = \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right], \tag{2}$$

where the expectation is w.r.t. the stochastic dynamics $P(s(t+1) \mid s(t), a(t)$ and the (possibly random) policy π which we use to choose the actions $a(t) \sim \pi(\cdot \mid s(t)$. In this paper, we study the optimal *control/planning* problem which is to seek a policy π^* that maximizes (2), given the dynamics f and the reward function r. Note that the nonlinearity of f and r, as well as the stochasticity from ϵ makes this optimal control problem difficult as reviewed in the introduction.

The above stochastic nonlinear optimal control problem can also be described via Markov Decision Process. Consider an episodic homogeneous-MDP, denoted by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$, where $P(\cdot|s,a) : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ describes the state transition distribution, where $\Delta(\mathcal{S})$ denotes the space of probability measures on the set $\mathcal{S}$. Then, the stochastic nonlinear control model (1) can be recast as an MDP with transition dynamics

$$P(s'|s,a) \propto \exp\left(-\frac{\|f(s,a) - s'\|_2^2}{2\sigma^2}\right). \tag{3}$$

Meanwhile, given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the corresponding Q^π-function is given by

$$Q^\pi(s,a) = \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s, a_0 = a\right]. \tag{4}$$

It is straightforward to show the Bellman recursion for Q^π or optimal Q^*-function, respectively,

$$Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_P\left[V^\pi(s')\right] \tag{5}$$

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_P\left[\max_{a' \in \mathcal{A}} Q^*(s', a')\right], \tag{6}$$

with $V^\pi(s) := \mathbb{E}_\pi\left[Q^\pi(s,a)\right]$. Equivalently, the accumulated reward J^π is $V^\pi(s_0)$. The goal can be reformulated as seeking the optimal policy $\pi^* = \operatorname{argmax}_\pi V^\pi(s_0)$.

For an MDP with finite states and actions, optimal control can be obtained by solving dynamic program for the Q-function via the Bellman relation (6). However, for continuous states and actions, representing the Q-function and conducting dynamic programming (DP) in function space becomes a major issue for obtaining the optimal policy. In this paper, we will introduce spectral dynamics embedding, a novel kernelized linearization, to linearly represent the Q-function, and develop a practical and provable method to learn and approximate the optimal control policy.

B. Positive definite kernels, its decompositions, and connection to RKHS

In this section, we will introduce the necessary background to understanding our proposed kernel embeddings. We first define what a positive definite (PD) kernel is, and introduce two decompositions of PD kernels.

Definition 1 ((Positive-Definite) Kernel [27]). *A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive-definite kernel on the non-empty set $\mathcal{X}$ if $\forall n \geqslant 1$, $\forall \{a_i\}_{i\in[n]} \subset \mathbb{R}$ and mutually distinct sets $\{x_i\}_{i\in[n]} \subset \mathcal{X}$, $\sum_{i\in[n]} \sum_{j\in[n]} a_i a_j k(x_i, x_j) \geqslant 0$, where the inequality is strict unless $\{a_i\}_{i\in[n]}$'s are all zero.*

A bounded, continuous PD kernel k admits two important decompositions, namely the Bochner decomposition [28] and the Mercer decomposition [29]. We begin with introducing the Bochner decomposition.

Lemma 1 (Bochner [28]). *If $k(x,x') = k(x-x')$ is a positive definite kernel, then there exists a set Ω, a measure $\rho(\omega)$ on Ω, and Fourier random feature $\psi_\omega(x) : \mathcal{X} \to \mathbb{C}$ such that*

$$k(x,x') = \mathbb{E}_{\rho(\omega)}\left[\psi_\omega(x)\, \psi_\omega(x')\right]. \tag{7}$$

It should be emphasized that the $\psi_\omega(\cdot)$ in Bochner decomposition may not be unique. The Bochner decomposition provides the random feature [30], which is applying Monte-Carlo approximation for (7), leading to a finite-dimension approximation for kernel methods. For example, for the Gaussian kernel, $k(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$, the corresponding $\rho(\omega)$ is a Gaussian proportional to $\exp\left(-\frac{\sigma^2\|\omega\|^2}{2}\right)$ with $\psi_\omega(x) = \exp\left(-\mathbf{i}\omega^\top x\right)$ where $\mathbf{i}$ is the imaginary unit; for the Laplace kernel, the $\rho(\omega)$ is a Cauchy distribution with the same $\psi_\omega(\cdot)$. Please refer to Table 1 in [31] for more examples.

An alternative but also useful decomposition for positive definite kernels is Mercer's decomposition, stated below[1].

Lemma 2 (Mercer [29], [32]). *Let $\mathcal{X} \subset \mathbb{R}^d$, μ be a strictly positive finite Borel measure on $\mathcal{X}$, and $k(x,x')$ a bounded continuous positive definite kernel. Then there exists a countable orthonormal basis $\{e_i\}_{i=1}^\infty$ of $\mathcal{L}_2(\mu)$ with corresponding eigenvalues $\{\sigma_i\}_{i=1}^\infty$,[2] such that*

$$k(x,x') = \sum_{i=1}^{\infty} \sigma_i e_i(x) e_i(x'), \tag{8}$$

where the convergence is absolute and uniform for all $(x,x') \in \mathcal{X} \times \mathcal{X}$. Without loss of generality, we assume $\sigma_1 \geqslant \sigma_2 \geqslant \cdots > 0$.

While analytical formulas of the Mercer decomposition exist for special kernels such as the Gaussian kernel [27], for general kernels they can be difficult to obtain analytically. Thus, it is common in the literature to approximate the Mercer features numerically, *e.g.*, Nyström method [33].

In addition, there is a one-to-one correspondence between PD kernels and Reproducing Kernel Hilbert Spaces (RKHSs). We first formally define the RKHS.

Definition 2 (Reproducing Kernel Hilbert Space [34]). *The Hilbert space $\mathcal{H}$ of an $\mathbb{R}$-valued function[3] defined on a non-empty set $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) if for all $x \in \mathcal{X}$, the (linear) evaluation functional given by $L_x : f \to f(x)$ for any $f \in \mathcal{H}$ is continuous.*

Next, we state the following classical result, which states a one-to-one correspondence between RKHS and PD kernels.

Lemma 3 (RKHS and PD kernels [34]). *Suppose $\mathcal{H}$ is a RKHS of $\{f : \mathcal{X} \to \mathbb{R}\}$. Then, there exists a unique positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $k(x,\cdot) \in \mathcal{H}$ for every $x \in \mathcal{X}$, that satisfies the* reproducing *property, i.e. for every $f \in \mathcal{H}$,*

[1] We note that under the conditions set forth in Lemma 2 (especially the finiteness of the measure μ), it can be shown using [32] that the Mercer expansion also holds for unbounded $\mathcal{X} \subset \mathbb{R}^d$.

[2] Given a probability measure μ defined on $\mathcal{X} \subseteq \mathbb{R}^d$, we denote $L_2(\mu)$ as the set of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\int_{\mathcal{X}} (f(x))^2 d\mu(x) < \infty$.

[3] We recall that a Hilbert space is an inner product space that is also a complete metric space with respect to the distance induced by its inner product $\langle \cdot, \rangle_{\mathcal{H}}$.

$f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}}$, where $\langle\cdot,\rangle_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$ (recall any Hilbert space has an inner product).

Conversely, for any positive definite kernel $k : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$, there exists a unique RKHS $\mathcal{H}_k$ of functions $f : \mathcal{X}\to\mathbb{R}$ for which k is the reproducing kernel. Moreover, given any $x, y \in \mathcal{X}$, $\langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}_k} = k(x,y)$.

We emphasize that any inner product, unless taking the form $\langle\cdot,\cdot\rangle_{H_k}$ for some RKHS H_k, does not refer to an RKHS inner product. Indeed, we only need the concept of RKHS when proving the approximation error for Nystrom features.

III. Control with Spectral Dynamics Embedding

In this section, we first introduce spectral dynamics embedding [18], a novel kernelized linearization, by which the Q-function for arbitrary policy can be represented *linearly*, therefore, the policy evaluation can be conducted within the linear space. This is significantly different from existing kernel linearization methods, which are designed for linearizing the dynamic model. We then propose the corresponding finite-dimensional approximated linear space for tractable optimal control. Specifically, we execute dynamic programming in the linear function space constructed by the spectral dynamics embedding for policy evaluation, upon which we improve the policy with natural policy gradient. This leads to *Spectral Dynamics Embedding Control (SDEC)* in Algorithm 1.

A. Spectral Dynamics Embedding

As we discussed in Section II, we have recast the stochastic nonlinear control model as an MDP. By recognizing the state transition operator (3) as a Gaussian kernel, we can further decompose the transition dynamics and reward in a *linear representation*, in two ways. The first utilizes the Bochner decomposition, as presented below.

Lemma 4 (Bochner linear representation, cf. Proposition 2 in [22])**.** *Consider any $\alpha\in[0,1)$. Denote $\theta_r = [0,0,1]^\top$,*

$$\phi_{\omega,b}(s,a) = \left[\psi_{\omega,b}(s,a), r(s,a)\right], \text{ where} \tag{9}$$

$$\psi_{\omega,b}(s,a) = \frac{g_\alpha(f(s,a))}{\alpha^d}\cos\left(\frac{\omega^\top f(s,a)}{\sqrt{1-\alpha^2}}+b\right), \tag{10}$$

$$\mu_{\omega,b}(s') = p_\alpha(s')[\cos(\sqrt{1-\alpha^2}\omega^\top s'+b), 0]^\top,$$

where $g_\alpha(f(s,a)) := \exp\left(\frac{\alpha^2\|f(s,a)\|^2}{2(1-\alpha^2)\sigma^2}\right)$, $\omega\sim\mathcal{N}(0,\frac{1}{\sigma^2}I_d)$, $b\sim U([0,2\pi]) := \mathrm{Unif}([0,2\pi])$ and $p_\alpha(s') = \frac{\alpha^d}{(2\pi\sigma^2)^{d/2}}\exp\left(-\frac{\|\alpha s'\|^2}{2\sigma^2}\right)$ is a Gaussian distribution for s' with standard deviation $\frac{\sigma}{\alpha}$. Then,

$$P(s'|s,a) = \mathbb{E}_{\omega\sim\mathcal{N}\left(0,\frac{1}{\sigma^2}I_d\right),b\sim U([0,2\pi])}\left[\phi_{\omega,b}(s,a)^\top\mu_{\omega,b}(s')\right] := \left\langle\phi_{\omega,b}(s,a),\mu_{\omega,b}(s')\right\rangle_{\mathcal{N}\left(0,\frac{1}{\sigma^2}I_d\right)\times U([0,2\pi])}, \tag{11a}$$

$$r(s,a) = \mathbb{E}_{\omega\sim\mathcal{N}\left(0,\frac{1}{\sigma^2}I_d\right),b\sim U([0,2\pi])}\left[\phi_{\omega,b}(s,a)^\top\theta_r\right] := \left\langle\phi_{\omega,b}(s,a),\theta_r\right\rangle_{\mathcal{N}\left(0,\frac{1}{\sigma^2}I_d\right)\times U([0,2\pi])}, \tag{11b}$$

Proof. The representation of reward $r(s,a)$ through $\phi_{\omega,b}(s,a)$ is straightforward since we explicitly include $r(\cdot)$ in $\phi_{\omega,b}(\cdot)$. For $P(s'|s,a)$, we first notice that $\forall\alpha\in(0,1)$, we have

$$P(s'|s,a)\propto\exp\left(-\frac{\left\|s'-f(s,a)\right\|^2}{2\sigma^2}\right)$$

$$= \exp\left(-\frac{\|\alpha s'\|^2}{2\sigma^2}\right)\exp\left(-\frac{\left\|(1-\alpha^2)s'-f(s,a)\right\|^2}{2\sigma^2(1-\alpha^2)}\right)\exp\left(\frac{\alpha^2\|f(s,a)\|^2}{2(1-\alpha^2)\sigma^2}\right). \tag{12}$$

The factorization of the transition $P(s'|s,a)$ in (11a) can thus be derived from the property of the Gaussian kernel by applying Lemma 1 to the second term in (12), with the $\phi_{\omega,b}(\cdot)$ and $\mu_{\omega,b}(\cdot)$ as the random Fourier features of Gaussian kernel. □

The $\phi_{\omega,b}(\cdot)$ is named as the *Bochner Spectral Dynamics Embedding*. We emphasize the decomposition in [18] is a special case of in (11) with $\alpha = 0$. The tunable α provides benefits for the analysis and may also be used to improve empirical performance, as seen in our empirical simulations. We note that another linear representation $\phi_M(s,a)\in\ell_2$,[4] which we will name as the *Mercer Spectral Dynamics Embedding*, is possible via the Mercer decomposition of the kernel, as described below.

Lemma 5 (Mercer linear representation)**.** *Let μ denote a strictly positive Borel measure on $\mathcal{S}$, and denote $k_\alpha(x,x') = \exp\left(-\frac{(1-\alpha^2)\|x-x'\|^2}{2\sigma^2}\right)$ for any $0\leqslant\alpha<1$. By Mercer's theorem, k_α admits a decomposition for any $(x,x')\in\mathcal{X}$:*

$$k_\alpha(x,x') = \sum_{i=1}^{\infty}\sigma_i e_i(x)e_i(x'),\ \{e_i\} \text{ a basis for } \mathcal{L}_2(\mu). \tag{13}$$

We denote $k_\alpha(x,x') = \langle\tilde{e}(x),\tilde{e}(x')\rangle_{\ell_2}$ where $\tilde{e}_i(x) = \sqrt{\sigma_i}e_i(x)$ for each positive integer i. Denote $\theta_{r,M} = [1,0,0,\dots]^\top\in\ell_2$ (and the same $g_\alpha(f(s,a))$ and $p_\alpha(s')$ as in Lemma 4), and

$$\phi_M(s,a) = [r(s,a), \psi_M(s,a)], \text{ where} \tag{14}$$

$$\psi_M(s,a) = \frac{g_\alpha(f(s,a))}{\alpha^d}\tilde{e}\left(\frac{f(s,a)}{1-\alpha^2}\right)^\top, \tag{15}$$

$$\mu_M(s') = p_\alpha(s')\left[0,\tilde{e}(s')\right]^\top \tag{16}$$

Then,

$$P(s'|s,a) = \langle\phi_M(s,a),\mu_M(s')\rangle_{\ell_2} \tag{17a}$$

$$r(s,a) = \langle\phi_M(s,a),\theta_{r,M}\rangle_{\ell_2} \tag{17b}$$

Proof. The proof is analogous to the proof of Lemma 4, where here we apply Mercer's theorem to decompose the middle term in terms of the kernel k_α

$$k_\alpha\left(s',\frac{f(s,a)}{1-\alpha^2}\right) = \exp\left(-\frac{(1-\alpha^2)\left\|s'-\frac{f(s,a)}{1-\alpha^2}\right\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\left\|(1-\alpha^2)s'-f(s,a)\right\|^2}{2\sigma^2(1-\alpha^2)}\right).$$

□

Remark 1. *In the proof of our kernel approximation result for the Nyström method later (Lemma 11), for technical reasons it is easier to work with k_α (as defined in the statement of Lemma 5) rather than k. Since the Nyström method is an approximation of the Mercer decomposition, we also decompose k_α in Lemma 5.*

The MDP with the factorization either in (11) or in (17) shows that we can represent the stochastic nonlinear control

[4] The ℓ_2 space is the set of square-summable infinite-dimensional vectors v indexed by the positive integers, such that $\|v\|_{\ell_2}^2 = \sum_{i=1}^{\infty}v_i^2 < \infty$.

problem as an instantiation of the *linear MDP* [24], [35], where the state transition and reward satisfy

$$P(s'|s,a) = \langle \phi(s,a), \mu(s') \rangle, \; r(s,a) = \langle \phi(s,a), \theta \rangle$$

with the feature functions $\phi(s,a)$ be either $\{\phi_{\omega,b}(s,a)\}$ with $\omega \sim \mathcal{N}\left(0, \frac{1}{\sigma^2} I_d\right), b \sim U([0,2\pi])$ or $\phi_M(s,a) \in \ell_2$. The most significant benefit of linear MDP is the property that it induces a function space composed by $\phi(s,a)$ where the Q-function for arbitrary policy can be linearly represented.

Lemma 6 (cf. Proposition 2.3 in [24])**.** *For any policy, there exist weights* $\{\theta^\pi_{\omega,b}\}$ *(where* $\omega \sim \mathcal{N}(0, \frac{1}{\sigma^2} I_d), b \sim U([0,2\pi])$*) and* $\theta^\pi_M \in \ell_2$ *such that the corresponding state-action value function satisfies*[5]

$$\begin{aligned} Q^\pi(s,a) &= \left\langle \phi_{\omega,b}(s,a), \theta^\pi_{\omega,b} \right\rangle_{\mathcal{N}(0,\frac{1}{\sigma^2} I_d) \times U([0,2\pi])} \\ &= \left\langle \phi_M(s,a), \theta^\pi_M \right\rangle_{\ell_2}. \end{aligned}$$

Proof. Our claim can be verified easily by applying the decomposition in (11) to Bellman recursion:

$$\begin{aligned} &Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_P \left[V^\pi(s') \right] \qquad (18) \\ &= \Big\langle \phi_{\omega,b}(s,a), \underbrace{\theta_r + \gamma \int_{\mathcal{S}} \mu_{\omega,b}(s') V^\pi(s') dp_\alpha(s')}_{\theta^\pi_{\omega,b}} \Big\rangle_{\mathcal{N}(0,\frac{1}{\sigma^2} I_d) \times U([0,2\pi])}, \\ &= \Big\langle \phi_M(s,a), \underbrace{\theta_{r,M} + \gamma \int_{\mathcal{S}} \mu_M(s') V^\pi(s') dp_\alpha(s')}_{\theta^\pi_M} \Big\rangle_{\ell_2}, \end{aligned}$$

where the second equation comes from the random feature decomposition and the third equation comes from the Mercer decomposition. □

Immediately, for the stochastic nonlinear control model (1) with an arbitrary dynamics $f(s,a)$, the spectral dynamics embeddings $\phi_{\omega,b}(\cdot)$ and $\phi_M(\cdot)$ provides a linear space, in which we can conceptually conduct dynamic programming for policy evaluation and optimal control with global optimal guarantee. However, compared to the existing literature of *linear MDP* ([24], [35]) that assumes a given linear factorization with finite dimension ϕ, both the Bochner and Mercer Spectral Dynamics embedding decomposes general stochastic nonlinear model explicitly through an infinite-dimensional kernel view, motivating us to investigate the finite-dimensional approximation.

B. Control with Finite-dimensional Approximation

Although the spectral dynamics embeddings in Lemma 4 and Lemma 5 provide linear spaces to represent the family of Q-function, there is still a major challenge to be overcome for practical implementation. Specifically, for the Bochner random feature embedding, the dimension of $\phi_{\omega,b}(\cdot)$ is *infinite* with $\omega \sim \mathcal{N}\left(0, \frac{1}{\sigma^2} I_d\right)$ and $b \sim U([0,2\pi])$, which is computationally intractable. Similarly, for the Mercer embedding, the dimension of $\phi_M(\cdot) \in \ell_2$ is also infinite. For the random feature embedding, [18] suggested the finite-dimensional random feature, which is the Monte-Carlo approximation for the Bochner decomposition [30], and demonstrated strong empirical performances for reinforcement learning. For the Mercer embedding, a well-known method to derive a finite-dimensional approximation of the Mercer expansion is the Nyström method [33], which seeks to approximate the subspace spanned by the top eigenfunctions in the Mercer expansion by using the eigendecomposition of an empirical Gram matrix. Compared to random features, it is known that the Nyström method can yield tighter approximation of the kernel when the kernel eigenvalues satisfy certain decay assumptions [36], motivating us to study and compare both finite-dimensional truncations.

[5] We note that these inner products refer to the inner product defined in Equation 11 and the standard ℓ_2 space inner product respectively, and not any RKHS inner products.

Algorithm 1: Spectral Dynamics Embedding Control (SDEC)

Data: Transition Model $s' = f(s,a) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$, Reward Function $r(s,a)$, Number of Random/Nyström Feature m, Number of Nyström Samples $n_{\text{Nys}} \geqslant m$, Nyström Sampling Distribution μ_{Nys}, Number of Sample n, Factorization Scale α, Learning Rate η, Finite-Truncation-Method $\in$ {'Random features', 'Nystrom'}

Result: π

`Spectral Dynamics Embedding Generation`

1 Generate $\phi(s,a)$ using Algorithm 2

`Policy evaluation and update`

2 Initialize $\theta_0 = 0$ and $\pi_0(a|s) \propto \exp(\phi(s,a)^\top \theta_0)$.

3 **for** $k = 0, 1, \cdots, K$ **do**

`Least Square Policy Evaluation`

4 Sample *i.i.d.* $\{(s_i, a_i, s'_i), a'_i\}_{i \in [n]}$ where $(s_i, a_i) \sim \nu_{\pi_k}$, $s'_i = f(s_i, a_i) + \varepsilon$ where ν_{π_k} is the stationary distribution of π_k, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$, $a'_i \sim \pi_k(s'_i)$.

5 Initialize $\hat{w}_{k,0} = 0$.

6 **for** $t = 0, 1, \cdots, T-1$ **do**

7 Solve

$$\hat{w}_{k,t+1} = \arg\min_w \left\{ \sum_{i \in [n]} \left(\phi(s_i,a_i)^\top w - r(s_i,a_i) - \gamma \phi(s'_i, a'_i)^\top \hat{w}_{k,t} \right)^2 \right\} \qquad (19)$$

8

9 **end**

`Natural Policy Gradient for Control`

10 Update $\theta_{k+1} = \theta_k + \eta \hat{w}_{k,T}$ and

$$\pi_{k+1}(a|s) \propto \exp(\phi(s,a)^\top \theta_{k+1}). \qquad (20)$$

11 **end**

In this section, we first formalize the Spectral Dynamics Embedding Control (SDEC) algorithm, implementing the dynamic programming efficiently in a principled way as shown in Algorithm 1, whose theoretical property will be analyzed in the next section. In Spectral Dynamics Embedding Control (SDEC), there are three main components,

1) *Generating spectral dynamics embedding (Line 1 in Algorithm 1, which calls Algorithm 2 as a subroutine.)* To generate the spectral dynamics embedding, we consider either using random features as a finite-dimensional truncation of

Algorithm 2: Spectral Dynamics Embedding Generation

Data: Same data as Algorithm 1
Result: $\phi(\cdot,\cdot)$
if *Finite-Truncation-Method = 'Random features'* **then**
  Sample *i.i.d.* $\{\omega_i\}_{i\in[m]}$ and $\{b_i\}_{i\in[m]}$ where $\omega_i \sim \mathcal{N}(0, \frac{1}{\sigma^2} I_d)$, $b_i \sim U([0, 2\pi])$ and construct the feature

$$\psi_{\mathrm{rf}}(s,a)=\frac{g_\alpha(f(s,a))}{\alpha^d}\left[\cos\left(\omega_i^\top \frac{f(s,a)}{\sqrt{1-\alpha^2}}+b_i\right)\right]_{i\in[m]} \tag{21}$$

  Set $\phi(s,a) = [\psi_{\mathrm{rf}}(s,a), r(s,a)]$
else if *Finite-Truncation-Method = 'Nystrom'* **then**
  Sample n_{nys} random samples $\{x_1, \ldots, x_{\mathrm{nys}}\}$ independently from $\mathcal{S}$ following the distribution μ_{Nys}.
  Construct n_{nys}-by-n_{nys} Gram matrix given by $K_{i,j}^{(n_{\mathrm{Nys}})} = k_\alpha(x_i, x_j)$
  Compute the Eigendecomposition $K^{(n_{\mathrm{nys}})}U = \Lambda U$, with $\lambda_1 \geqslant \ldots \geqslant \lambda_{n_{\mathrm{nys}}}$ denoting the corresponding eigenvalues.
  Construct the feature

$$\psi_{\mathrm{nys}}(s,a)=\left[\frac{g_\alpha(f(s,a))}{\alpha^d\sqrt{\lambda_i}}\sum_{\ell=1}^{n_{\mathrm{nys}}} U_{i,\ell}k_\alpha\left(x_\ell, \frac{f(s,a)}{1-\alpha^2}\right)\right]_{i\in[m]} \tag{22}$$

  Set $\phi(s,a) = [\psi_{\mathrm{nys}}(s,a), r(s,a)]$
end

the Bochner random features, or the Nyström method as a finite-dimensional truncation of the Mercer features.

- **Random features.** Following Lemma 4, we construct finite-dimensional ϕ by Monte-Carlo approximation, building $\psi_{\mathrm{rf}} = [\psi_{\omega_i,b_i}]_{i=1}^m$, where $\omega_i \sim \mathcal{N}\left(0, \frac{1}{\sigma^2} I\right)$ and $b_i \sim U([0, 2\pi])$. We then concatenate ψ_{rf} with r to form ϕ.
- **Nyström features.** At a high level, we use the Nyström method [33] to obtain the Nyström feature decomposition of $k_\alpha(\cdot,\cdot) \approx \varphi_{\mathrm{nys}}(\cdot)^\top \varphi_{\mathrm{nys}}(\cdot)$ for some $\varphi_{\mathrm{nys}}(\cdot) \in \mathbb{R}^m$ by drawing n_{nys} random samples iid from a sampling distribution μ_{nys} over the state space $\mathcal{S}$, where the sampling distribution can be specified by the user, and in our analysis is chosen to be p_α (note we always choose $m \leqslant n_{\mathrm{nys}}$) and then working with the eigendecomposition of the corresponding Gram matrix; we derive and motivate the Nyström method in greater detail in Appendix VI-B of our supplement [37]. Then, noting that

$$k_\alpha\left(s', \frac{f(s,a)}{1-\alpha^2}\right) \approx \varphi_{\mathrm{nys}}\left(\frac{f(s,a)}{1-\alpha^2}\right)^\top \varphi_{\mathrm{nys}}(s'),$$

  echoing the Mercer decomposition in Lemma 5, we set

$$\psi_{\mathrm{nys}}(s,a) = \frac{g_\alpha(f(s,a))}{\alpha^d}\varphi_{\mathrm{nys}}\left(\frac{f(s,a)}{1-\alpha^2}\right),$$

  which can be viewed as a finite-dimensional truncation of the Mercer feature in (15). We specify the exact form of $\varphi_{\mathrm{nys}}(\cdot)$ in (50) of Appendix VI-B, which results in the feature ψ_{nys} in (22) of Algorithm 2.

2) *Policy evaluation (Line 4 to 9 in Algorithm 1).* We conduct least square policy evaluation for estimating the state-action value function of current policy upon the generated finite-dimensional truncation features. Specifically, we sample[6] from the stationary distribution $\nu_\pi(s,a)$ from dynamics under current policy π and solve a series of least square regression (19) to learn a $Q^\pi(s,a) = \phi(s,a)^\top w^\pi$, with $w^\pi \in \mathbb{R}^{m\times 1}$ by minimizing a Bellman recursion type loss.
3) *Policy update (Line 10 in Algorithm 1).* Once we have the state-value function for current policy estimated well in step 2), we will update the policy by natural policy gradient or mirror descent [38]. Specifically, we have the functional gradient of accumulated reward w.r.t. a direct parameterization policy π as $\nabla_\pi J^\pi = \frac{1}{1-\gamma}\nu^\pi(s) Q^\pi(s,a)$, where $\nu^\pi(s,a)$ denotes the stationary distribution of π. At t-iteration, we estimated the $Q^{\pi_t}(s,a)$ with $\phi(s,a)^\top w^{\pi_t}$, we have the mirror descent update [39] for π as

$$\pi^{t+1} = \operatorname*{argmax}_{\pi(\cdot|s)\in\Delta(\mathcal{A})} \left\langle \pi, \nu^{\pi_t}(s)\phi(s,a)^\top w^{\pi_t}\right\rangle + \frac{1}{\eta} KL(\pi||\pi_t)$$

leading to the following closed-form solution,

$$\pi^{t+1}(a|s) \propto \pi_t(a|s)\exp\left(\phi(s,a)^\top \eta w^{\pi_t}\right)$$
$$= \exp\left(\phi(s,a)^\top \theta_{t+1}\right)$$

where $\theta_{t+1} = \sum_{i=0}^t \eta w^{\pi_i}$, giving the update rule in (20).

Once the finite-dimensional spectral dynamics embedding has been generated in step 1), the algorithm alternates between step 2) and step 3) to improve the policy.

Remark 2 (Policy evaluation and update). *We note that although SDEC uses the least square policy evaluation and natural policy gradient method for the policy update, the proposed dynamic spectral embedding is compatible with other policy evaluation and policy update methods, including cutting-edge deep reinforcement learning methods, such as Soft Actor-Critic [21], by using our proposed representation to approximate the critic function. In other words, one of the novelties of SDEC, is that it exploits the known nonlinear dynamics to obtain a natural, inherent representation space that could be adopted by various dynamical programming or policy gradient-based methods.*

Remark 3 (Beyond Gaussian noise). *Although SDEC mainly focuses the stochastic nonlinear dynamics with Gaussian noise, the Nyström feature is clearly applicable for any transition operator. Moreover, we can easily extend the random feature for more flexible noise models. Specifically, we consider the exponential family condition distribution for the transition [40],* i.e.,

$$P(s'|s,a) \propto p(s')\exp\left(f(s,a)^\top \zeta(s')\right), \tag{23}$$

The transition (23) *generalizes the noise model in the transition in two-folds:* **i)**, *the noise model is generalized beyond Gaussian to exponential family, which covers most common noise models,* e.g., *exponential, gamma, Poisson, Wishart, and so on; and* **ii)**, *the nonlinear transformation over $\zeta(s')$ also generalize the dynamics. The random features for* (23) *are derived in Appendix VI-A.*

IV. Theoretical Analysis

The major difficulty in analyzing the optimality of the policy obtained by SDEC is the fact that after finite-dimensional truncation, the transition operator constructed

[6] Our analysis also assumes that we sample exactly from $\nu_\pi(s,a)$, but in practice, we can only approximately sample from this distribution, which may incur another source of error.

by the random or Nyström features is no longer a valid distribution, *i.e.*, it lacks non-negativity and normalization, which induces a *pseudo-MDP* [41] as the approximation. For instance, in general, for the random features, the term $\widehat{P}(s'|s,a) := \frac{1}{m}\sum_{i=1}^{m} \phi_{\omega_i,b_i}(f(s,a))\,\mu_{\omega_i,b_i}(s')$ with $\{\omega_i, b_i\}_{i=1}^{m} \sim \mathcal{N}\left(0, \sigma^{-2} I_d\right) \times U([0, 2\pi])$ may be negative for some values of s' and (s,a). As a consequence, the value function for the pseudo-MDP is not bounded, as the normalization condition of the transition operator no longer holds. Then, the vanilla proof strategy used in majority of the literature since [42], *i.e.*, analyzing the optimality gap between policies through simulation lemma, is no longer applicable.

In this section, we bypass the difficulty from pseudo-MDP, and provide rigorous investigation of the impact of the approximation error for policy evaluation and optimization in SDEC, filling the long-standing gap. We first specify the assumptions, under which we derive our theoretical results below. These assumptions are commonly used in the literature [24], [38], [43], [44].

Assumption 1 (Regularity Condition for Dynamics and Reward)**.** *$\|f(s,a)\| \leqslant c_f$, and $r(s,a) \leqslant c_r$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. For the ease of presentation, we omit the polynomial dependency on c_f and c_r and focus on the dependency of other terms of interest.*

Assumption 2 (Regularity Condition for Feature)**.** *The features are linearly independent.*

Assumption 3 (Regularity Condition for Stationary Distribution [44])**.** *Any policy π has full support over $\mathcal{A}$. In addition, for any policy π, the discounted stationary state distribution d^π has full support over any region in $\mathcal{S}$ on which an optimal policy π^* is supported, and the discounted stationary state-action distribution ν^π satisfies the following conditions:*

$$\sigma_{\min}\left(\mathbb{E}_{\nu_\pi}\left[\phi(s,a)\phi(s,a)^\top\right]\right) \geqslant \Upsilon_1, \tag{24}$$

$$\sigma_{\min}\left(\mathbb{E}_{\nu_\pi}\left[\phi(s,a)\left(\phi(s,a) - \gamma \mathbb{E}_{\nu_\pi}\phi(s',a')\right)^\top\right]\right) \geqslant \Upsilon_2, \tag{25}$$

where $\Upsilon_1, \Upsilon_2 > 0$, and $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix.

Assumption 1 is an assumption that holds whenever the state space is compact. In the case of unbounded state spaces, given the ability to design rewards, the boundedness of $\|f(s,a)\|$ is more difficult to ensure. However, so long as we are in the fairly reasonable scenario where the value function is 0 outside of a bounded set, this assumption on the boundedness of $\|f(s,a)\|$ can be seen to no longer be necessary. To see why, suppose that the value function $V^\pi(\cdot)$ function for any policy π is 0 outside of a bounded set $\mathcal{S}_0 \subset \mathcal{S}$; this is a fairly representative scenario since for many problems, assuming we have the ability to do reward design, there exists a region of interest which we would like to give high rewards to encourage the agent to stay inside. Outside of this region, we can simply set the rewards to be 0. With this in mind, observe that we need the bound on $\sup_{s,a} f(s,a)$ because we need the term

$$\sup_{((s,a),\omega)} \left|\phi_\omega(s,a)^\top \beta_\omega^\pi\right| \leqslant \sup_{((s,a),\omega)} |\psi_\omega(s,a)|_1 \, \|\beta_\omega^\pi\|_\infty = O\left(\frac{\sup_{(s,a)} g_\alpha(f(s,a))}{\alpha^d(1-\gamma)}\right),$$

which appears in the proof of Proposition 1 to be bounded (the analysis is similar for Nystrom features in Proposition 2). When we set $\alpha = 0$, the term $g_\alpha(f(s,a)) = \exp(\frac{\alpha^2\|f(s,a)\|^2}{2\sigma^2(1-\alpha^2)})$ is simply equal to 1. In this case, it is clear that term $\psi_{\omega,b}(s,a) = \frac{g_\alpha(f(s,a))}{\alpha^d}\cos\left(\frac{\omega^\top f(s,a)}{\sqrt{1-\alpha^2}} + b\right)$ is always bounded. Meanwhile, the term $\|\beta_{\omega,b}^\pi\| = \int_{\mathcal{S}} p_\alpha(s')\mu_{\omega,b}(s')V^\pi(s')ds'$ is bounded even though $p_\alpha(s') = 1$ for $\alpha = 0$, since $V^\pi(s') = 0$ outside of a bounded set $\mathcal{S}_0 \subset \mathcal{S}$. Thus, even when the state and action spaces are infinite, so long as we are in the fairly general scenario where the value function $V^\pi(\cdot)$ function for any policy π is 0 outside of a bounded set $\mathcal{S}_0 \subset \mathcal{S}$, our analysis should go through without the need for bounding $\|f(s,a)\|$ on the entirety of the state-action space..

Assumptions 2 and 3 are also typically challenging to verify. Nonetheless, for our proposed features $\phi(s,a) = [\psi(s,a), r(s,a)]$, where $\psi(s,a)$ can either denote $\psi_{rf}(s,a)$ or $\psi_{nys}(s,a)$, we show that the transition features $\psi(s,a)$ are linearly independent under very mild conditions, which then implies that Assumption 2 holds when the reward function $r(s,a)$ does not lie exactly in the span of the transition feature functions $\psi(s,a)$; this again is a mild condition that is for instance satisfied for any (locally) quadratic reward function when random features are used. When this happens, we then show that the two key conditions (24) and (25) in Assumption 3 holds for all but finitely many discount factors λ.

Lemma 7. *The random features $\psi_{rf}(s,a)$ are linearly independent almost surely.*

Lemma 8. *Consider any feature dimension $0 < m \leqslant n_{Nys}$. Suppose the Nystrom Gram matrix $K^{(n_{Nys})}$ has rank at least m. Then, the Nystrom features $\psi_{nys}(s,a)$ are linearly independent.*

Lemma 9. *Suppose the features $\phi(s,a)$ are linearly independent over the interior of the support of ν^π for any policy π. Then, for all but finitely many $0 \leqslant \lambda < 1$, the two conditions (24) and (25) in Assumption 3 holds.*

Proofs of Lemmas 7,8 and 9 are deferred to our supplement [37].

Throughout the analysis, for notational simplicity it will be helpful to define the term

$$\tilde{g}_\alpha := \sup_{s,a}\left(g_\alpha(f(s,a))\right)\alpha^{-d} \tag{26}$$

We provide a brief outline of our overall proof strategy. First, in Section IV-A we study the policy evaluation error. To do so, for any policy π_k encountered during the algorithm, let $\hat{Q}^{\pi_k}$ denote the learned Q-value at the end of the policy evaluation step in Line 9 of Algorithm 1. Then, by the triangle inequality, the approximation error satisfies

$$\|Q^{\pi_k} - \hat{Q}^{\pi_k}\|_{\nu_{\pi_k}} \leqslant \underbrace{\|Q^{\pi_k} - \tilde{Q}^{\pi_k}\|_{\nu_{\pi_k}}}_{\textit{approximation error}} + \underbrace{\|\tilde{Q}^{\pi_k} - \hat{Q}^{\pi_k}\|_{\nu_{\pi_k}}}_{\textit{statistical error}},$$

where $\tilde{Q}^{\pi_k}$ denotes the solution to a projected Bellman equation where the projection is onto the either the finite-dimensional random features ((Line 2 in Algorithm 2)) or Nyström features (Line 8 in Algorithm 2); we explain this in greater detail in Section IV-A. Essentially, the approximation error corresponds to the truncation error incurred by using finite-dimensional features, while the statistical error is the error incurred by using finitely many samples to approximate $\tilde{Q}^{\pi_k}$. Bounding these two sources of errors gives us a total error for policy evaluation bound in Theorem 1.

Second, in Section IV-B, we study the policy optimization error of the natural policy gradient in Line 10 of Algorithm

1. We note that the policy optimization error includes a component inherited from the policy evaluation error. This culminates in Theorem 2, which provides the overall performance guarantee for the control optimality of SDEC.

A. Error Analysis for Policy Evaluation

Our analysis starts from the error for policy evaluation (Line 4 to 9 in Algorithm 1). We decompose the error into two parts, one is the approximation error due to the limitation of our basis (*i.e.*, finite m in Line 2 and Line 8 of Algorithm 2), and one is the statistical error due to the finite number of samples we use (*i.e.*, finite n in Line 4 of Algorithm 1). For notational simplicity, we omit π and use ν to denote the stationary distribution corresponding to π in this section.

a) Approximation Error: We first provide the bound on the approximation error in terms of representing the Q-function for arbitrary policy due to using an imperfect finite-dimensional basis. When restricted to using a finite-dimensional basis, the best possible $\tilde{Q}$ approximation is the solution to a projected Bellman equation (cf. [43]), defined as follows: given any (possibly finite-dimensional) feature map $\Phi := \{\phi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, we define the approximation $\tilde{Q}^\pi_\Phi$ as follows

$$\tilde{Q}^\pi_\Phi = \Pi_{\nu,\Phi}(r + P^\pi \tilde{Q}^\pi_\Phi), \tag{27}$$

where we recall ν is the stationary distribution under π, the operator P^π is defined as

$$(P^\pi h)(s,a) = \mathbb{E}_{(s',a')\sim P(s,a)\times\pi} h(s',a'). \tag{28}$$

and $\Pi_{\nu,\Phi}$ is the projection operator as defined as

$$\Pi_{\nu,\Phi} Q = \underset{h\in\mathrm{span}(\Phi)}{\arg\min}\ \mathbb{E}_\nu \left(Q(s,a) - h(s,a)\right)^2. \tag{29}$$

Our interest in $\tilde{Q}^\pi_\Phi$ stems from the fact that the least-squares policy evaluation step in Algorithm 1 (see equation (19)) recovers $\tilde{Q}^\pi_\Phi$ if the number of samples, n, goes to infinity. We will address the statistical error from using a finite n later.

We are now ready to introduce our bound for the approximation error in using finite-dimensional random and Nyström features. We begin with the result for random features. To do so, we first need the following technical result, which is a restatement of Lemma 1 from [19].

Lemma 10 (cf. Lemma 1 from [19])**.** *Let p be a distribution on a space Ω, and consider a mapping $\phi(x;\omega)\in\mathbb{R}$. Suppose $f^*(x) = \int_\Omega p(\omega)\beta(\omega)\phi(x;\omega)d\omega$ for some scalar $\beta(\omega)\in\mathbb{R}$ where $\sup_{x,\omega}|\beta(\omega)\phi(x;\omega)| \leqslant C$ for some $C>0$. Consider $\{\omega_i\}_{i=1}^k$ drawn iid from p, and denote the sample average of f^* as $\hat{f}(x) = \frac{1}{K}\sum_{k=1}^K \beta(\omega_k)\phi(x;\omega_k)$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the random draws of $\{\omega_i\}_{i=1}^k$,*

$$\sqrt{\int_{\mathcal{X}} \left(\hat{f}(x) - f^*(x)\right)^2 d\mu(x)} \leqslant \frac{C}{\sqrt{K}}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right).$$

Proposition 1 (Q-Approximation error with random features)**.** *We define the feature map Φ_{rf} for random features as follows*

$$\Phi_{\mathrm{rf}} = \{[\psi_{\mathrm{rf}}(s,a), r(s,a)]\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}, \tag{30}$$

where $\psi_{\mathrm{rf}}(s,a)$ is defined in (21) in Algorithm 1. Then, for any $\delta>0$, with probability at least $1-\delta$, we have that

$$\left\|Q^\pi - \tilde{Q}^\pi_{\Phi_{\mathrm{rf}}}\right\|_\nu = \tilde{O}\left(\frac{\gamma\tilde{g}_\alpha}{(1-\gamma)^2\sqrt{m}}\right), \tag{31}$$

where $\|\cdot\|_\nu$ is the L_2 norm defined as $\|f\|_\nu = \int f^2 d\nu$, and $\tilde{Q}^\pi_{\Phi_{\mathrm{rf}}}$ is defined in (27).

Proof. With the contraction property and results from [43], we have

$$\left\|Q^\pi - \tilde{Q}^\pi_{\Phi_{\mathrm{rf}}}\right\|_\nu \leqslant \frac{1}{1-\gamma}\left\|Q^\pi - \Pi_{\nu,\Phi_{\mathrm{rf}}} Q^\pi\right\|_\nu, \tag{32}$$

where $\Pi_{\nu,\Phi_{\mathrm{rf}}}$ is defined as in (29). By definition $\Pi_{\nu,\Phi_{\mathrm{rf}}}$ is contractive under $\|\cdot\|_\nu$. Note that

$$Q^\pi(s,a) = r(s,a) + \gamma\mathbb{E}_{\omega\sim\mathcal{N}(0,\frac{1}{\sigma^2}I_d),b\sim U([0,2\pi])}\left[\phi_{\omega,b}(s,a)^\top \int_{\mathcal{S}} p_\alpha(s')\mu_\omega(s')V^\pi(s')ds'\right].$$

With Hölder's inequality, as well as $\|\mu_{\omega,b}(s')\| = O(1)$ by definition, $\|V^\pi\|_\infty = O((1-\gamma)^{-1})$, and using the notation $\beta^\pi_{\omega,b} := \int_{\mathcal{S}} p_\alpha(s')\mu_{\omega,b}(s')V^\pi(s')ds'$, we have for every (ω,b) that $\left\|\beta^\pi_{\omega,b}\right\| = O((1-\gamma)^{-1})$. In addition, recalling that $\phi_{\omega,b}(s,a) = [\psi_{\omega,b}(s,a), r(s,a)]$, we have

$$\sup_{((s,a),(\omega,b))} |\phi_{\omega,b}(s,a)\beta^\pi_{\omega,b}| \leqslant \sup_{((s,a),(\omega,b))} |\psi_{\omega,b}(s,a)|\,|\beta^\pi_{\omega,b}|$$

since the coordinate in $\beta^\pi_{\omega,b}$ corresponding to the reward coordinate of $\phi_{\omega,b}(s,a)$ is 0. Since

$$\sup_{((s,a),(\omega,b))} |\psi_{\omega,b}(s,a)| \leqslant 2\sup_{(s,a)} \frac{g_\alpha(f(s,a))}{\alpha^d},$$

this implies that

$$\sup_{((s,a),(\omega,b))} |\phi_{\omega,b}(s,a)\beta^\pi_{\omega,b}| = O\left(\frac{\sup_{(s,a)} g_\alpha(f(s,a))}{\alpha^d(1-\gamma)}\right)$$

Thus, applying Lemma 10, we have that

$$\left\|Q^\pi - \Pi_{\nu,\Phi_{\mathrm{rf}}} Q^\pi\right\|_\nu = \tilde{O}\left(\frac{\gamma\sup_{s,a} g_\alpha(f(s,a))}{\alpha^d(1-\gamma)\sqrt{m}}\right). \tag{33}$$

Substituting (33) into (32) finishes the proof. □

We see above that with high probability, the approximation error for random features shrinks at a rate of $O\left(m^{-1/2}\right)$ as we increase the number of random features m. We proceed to show our approximation error result for Nyström features. An essential step to proving our approximation error for Nyström features is the following proposition, which is a general high-probability approximation error bound when approximating a smooth kernel using Nyström features.

Lemma 11 (Kernel approximation error using Nyström features)**.** *Consider the following Mercer decomposition (on $\mathcal{S}$) of $k_\alpha(\cdot,\cdot)$:*

$$k_\alpha(x,x') = \sum_{i=1}^\infty \sigma_i e_i(x)e_i(x'), \tag{34}$$

where $\{e_i\}_{i=1}^\infty$ forms a countable orthonormal basis for $L_2(\mu_{\mathrm{nys}})$ with corresponding eigenvalues $\{\sigma_i\}_{i=1}^\infty$. Let $X^{n_{\mathrm{nys}}} = \{x_i\}_{i=1}^{n_{\mathrm{nys}}}$ be an i.i.d n_{nys}-point sample from μ_{nys}. In addition, let $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_{n_{\mathrm{nys}}}$ denote the eigenvalues of the (unnormalized) Gram matrix $K^{(n_{\mathrm{nys}})}$ in its eigendecomposition $K^{(n_{\mathrm{nys}})}U = \Lambda U$ where $U^\top U = UU^\top = I$. Suppose that $\sigma_j, \lambda_j/n_{\mathrm{nys}} \lesssim \exp(-\beta j^{1/h})$ for some $\beta>0$ and $h>0$. Suppose that $n_{\mathrm{nys}} \geqslant 3$ and that $\lfloor(2\log n_{\mathrm{nys}})^h/\beta^h\rfloor \leqslant m \leqslant n_{\mathrm{nys}}$.

Consider the rank-m kernel approximation $\hat{k}^{n_{\mathrm{nys}}}_{\alpha,m}$ constructed using Nyström features, defined as follows:

$$\hat{k}^{n_{\mathrm{nys}}}_{\alpha,m}(s,t) = \varphi_{\mathrm{nys}}(s)^\top \varphi_{\mathrm{nys}}(t), \tag{35}$$

where $\varphi_{\text{nys}}(\cdot) \in \mathbb{R}^m$ *and is defined as*

$$(\varphi_{\text{nys}})_i(\cdot) := \frac{1}{\sqrt{\lambda_i}} u_i^\top k_\alpha(X^{n_{\text{nys}}}, \cdot), \quad \forall i \in [m], \tag{36}$$

where u_i *denotes the* i*-th column of* U*, and* $k_\alpha(X^{n_{\text{nys}}}, \cdot)$ *denotes a* n_{nys}*-dimensional vector where* $(k_\alpha(X^{n_{\text{nys}}}, \cdot))_\ell = k_\alpha(x_\ell, \cdot)$*; for details on deriving* $(\varphi_{\text{nys}})_i$*, see Appendix VI-B in our supplement [37]. Then,* $\forall \delta > 0$*, with probability at least* $1-\delta$*,*

$$\int_{\mathcal{S}} \sqrt{\left(k_\alpha - \hat{k}^{n_{\text{nys}}}_{\alpha,m}\right)(x,x)} d\mu_{\text{nys}}(x) = \tilde{O}\left(\frac{1}{n_{\text{nys}}}\right). \tag{37}$$

We defer the proof of Lemma 11 to Appendix VI-C. We note here that the $\tilde{O}(1/n_{\text{nys}})$ high probability error decay rate is a novel result, outperforming the $\tilde{O}(1/\sqrt{n_{\text{nys}}})$ rate in [20]. We achieve this improvement by combining Bernstein's inequality and results in local Rademacher complexity theory [45]. Using Lemma 11, we can then demonstrate the following approximation error for the Nyström features.

Proposition 2 (Q-Approximation error with Nyström features)**.** *Suppose all the assumptions in Lemma 11 hold. Suppose also that we pick the sampling distribution* μ_{nys} *such that* $\mu_{\text{nys}}(x) = p_\alpha(x)$*. We define the feature map* Φ_{nys} *for the Nyström features as follows:*

$$\Phi_{\text{nys}} = \{[\psi_{\text{nys}}(s,a), r(s,a)]\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}, \tag{38}$$

where $\psi_{\text{nys}}(s,a) \in \mathbb{R}^m$ *is defined in* (38) *in Algorithm 1. Then, for any* $\delta > 0$*, with probability at least* $1-\delta$*,*

$$\left\|Q^\pi - \tilde{Q}^\pi_{\text{nys}}\right\|_\nu \leqslant \tilde{O}\left(\frac{\gamma \tilde{g}_\alpha}{(1-\gamma)^2 n_{\text{nys}}}\right) \leqslant \tilde{O}\left(\frac{\gamma \tilde{g}_\alpha}{(1-\gamma)^2 m}\right),$$

where $\|\cdot\|_\nu$ *is the* L_2 *norm defined as* $\|f\|_\nu = \int f^2 d\nu$*,* $\tilde{Q}^\pi_{\Phi_{\text{nys}}}$ *is defined in* (27)*, and* $\tilde{g}_\alpha := \sup_{s,a} \frac{g_\alpha(f(s,a))}{\alpha^d}$*.*

Proof. Similar to the analysis at the start of the proof of Proposition 1, with the contraction property and results from [43], we have that

$$\left\|Q^\pi - \tilde{Q}^\pi_{\Phi_{\text{nys}}}\right\|_\nu \leqslant \frac{1}{1-\gamma}\left\|Q^\pi - \Pi_{\nu,\Phi_{\text{nys}}} Q^\pi\right\|_\nu, \tag{39}$$

where $\Pi_{\nu,\Phi_{\text{nys}}}$ is defined as in (29). It remains for us to bound $\left\|Q^\pi - \Pi_{\nu,\Phi_{\text{nys}}} Q^\pi\right\|_\nu$. First, by (35), recalling the definition of $\varphi_{\text{nys}}(\cdot)$ in (36), we have

$$\hat{k}^{n_{\text{nys}}}_m\left(s', \frac{f(s,a)}{1-\alpha^2}\right) = \varphi_{\text{nys}}(s')^\top \varphi_{\text{nys}}\left(\frac{f(s,a)}{(1-\alpha^2)}\right), \tag{40}$$

as our Nyström approximation of $k_\alpha\left(s', \frac{f(s,a)}{1-\alpha^2}\right)$. Since we have $\psi_{\text{nys}}(s,a) = g_\alpha(f(s,a))\varphi_{\text{nys}}\left(\frac{f(s,a)}{1-\alpha^2}\right)$, where $\psi_{\text{nys}}(\cdot)$ is defined as in (22), this motivates us to consider the following Q-value approximation:

$$\hat{Q}^\pi_{\text{nys}}(s,a) := r(s,a) + \gamma \psi_{\text{nys}}(s,a)^\top \left(\frac{\int_{\mathcal{S}} \varphi_{\text{nys}}(s') V^\pi(s') p_\alpha(s') ds'}{\alpha^d}\right),$$

which we note this is a feasible solution to the objective

$$\underset{f\in \text{span}(\Phi_{\text{nys}})}{\arg\min} \ \mathbb{E}_\nu \left(Q^\pi(s,a) - f(s,a)\right)^2.$$

Thus, we have

$$\begin{aligned}
&\left\|\Pi_{\nu,\Phi_{\text{nys}}}(Q^\pi) - Q^\pi\right\|_\nu \leqslant \left\|\hat{Q}^\pi_{\text{nys}} - Q^\pi\right\|_\nu \\
&= \gamma \left\| \psi_{\text{nys}}(s,a)^\top \left(\frac{\int_{\mathcal{S}} \varphi_{\text{nys}}(\sqrt{1-\alpha^2}s') V^\pi(s') p_\alpha(s') ds'}{\alpha^d}\right) \right. \\
&\qquad \left. - \int_{\mathcal{S}} P(s' \mid s,a) V^\pi(s') ds' \right\|_\nu \\
&= \gamma \left\| \left(\frac{g_\alpha(f(s,a))}{\alpha^d}\right) \left(\int_{\mathcal{S}} \varphi_{\text{nys}}\left(\frac{f(s,a)}{1-\alpha^2}\right)^\top \varphi_{\text{nys}}(s') V^\pi(s') dp_\alpha(s') \right.\right. \\
&\qquad \left.\left. - \int_{\mathcal{S}} k_\alpha(s', \tfrac{f(s,a)}{1-\alpha^2}) V^\pi(s') dp_\alpha(s') \right) \right\|_\nu \\
&\leqslant \gamma \left(\frac{V_{\max} g_{\alpha,\sup}}{\alpha^d}\right) \times \\
&\qquad \underbrace{\left\| \int_{\mathcal{S}} \left| \hat{k}^{n_{\text{nys}}}_m\left(s', \tfrac{f(s,a)}{1-\alpha^2}\right) - k_\alpha\left(s', \tfrac{f(s,a)}{1-\alpha^2}\right) \right| p_\alpha(s') ds' \right\|_\nu}_{T_1},
\end{aligned} \tag{41}$$

where $g_{\alpha,\sup} := \sup_{s,a} g_\alpha(f(s,a))$. By our choice of $\mu_{\text{nys}} = p_\alpha(x)$, continuing from (41), we have

$$\begin{aligned}
T_1 &= \left\| \int_{\mathcal{S}} \left| \hat{k}^{n_{\text{nys}}}_m\left(s', \tfrac{f(s,a)}{1-\alpha^2}\right) - k_\alpha\left(s', \tfrac{f(s,a)}{1-\alpha^2}\right) \right| p_\alpha(s') ds' \right\|_\nu \\
&= \left\| \int_{\mathcal{S}} \left| \left\langle k_\alpha\left(\tfrac{f(s,a)}{1-\alpha^2}, \cdot\right), \hat{k}^{n_{\text{nys}}}_m(s', \cdot) - k_\alpha(s', \cdot) \right\rangle_{\mathcal{H}_{k_\alpha}} \right| dp_\alpha(s') \right\|_\nu \\
&\leqslant \left\| \int_{\mathcal{S}} \sqrt{k_\alpha\left(\tfrac{f(s,a)}{1-\alpha^2}, \tfrac{f(s,a)}{1-\alpha^2}\right)} \sqrt{(k_\alpha - \hat{k}^{n_{\text{nys}}}_m)(s',s')} dp_\alpha(s') \right\|_\nu \\
&\leqslant \left\| \int_{\mathcal{S}} \sqrt{(k_\alpha - \hat{k}^{n_{\text{nys}}}_m)(s',s')} dp_\alpha(s') \right\|_\nu
\end{aligned}$$

Note to move from the second last line to the last line, we used the fact that $k_\alpha(\cdot,\cdot) \leqslant 1$. Using Proposition 11, by utilizing the decay assumption on the eigenvalues of both the Mercer expansion and empirical gram matrix, we thus have that for any $\delta > 0$, with probability at least $1-\delta$,

$$T_1 \leqslant \left(\int_{\mathcal{S}} \sqrt{(k_\alpha - \hat{k}^{n_{\text{nys}}}_m)(s',s')} \mu(s') ds' \right) = \tilde{O}\left(\frac{1}{n_{\text{nys}}}\right).$$

We conclude by combining (41) and the bound on T_1. □

Remark 4. *Proposition 2 shows that using the Nyström method improves the approximation error to* $O\left(\left(n^{-1}_{\text{nys}}\right)\right)$*, where* n_{nys} *is the number of samples we sample to construct the Gram matrix used to build the Nyström features. We emphasize that an important condition for this hold is the assumption in the above proposition that the number of features* m *satisfy* $m \geqslant \frac{(2\log n_{nys})^h}{\beta^h}$*. The rate* $O\left(\left(n^{-1}_{\text{nys}}\right)\right)$ *is always an improvement over* $O(1/m)$ *where* m *is the number of features, since by design SDEC chooses* $m \leqslant n_{\text{nys}}$*. The only assumption required for this improvement is that the eigenvalues of the empirical Gram matrix and Mercer expansion satisfy certain decay assumptions, which is a natural assumption to make in the kernel literature (cf. Theorem 2 in [46]).*

b) Statistical Error: We now provide the bound on the statistical error due to using a finite number of samples n. We note that this is a general result that applies to any features $\phi(s,a)$, inclding both Nyström and random features, that satisfies Assumptions 1 to 3.[7]

Proposition 3. *For each policy* π *encountered in the algorithm, let* $\hat{Q}^\pi_{\Phi,T}$ *denote the policy given by* $\hat{Q}^\pi_{\Phi,T}(s,a) = \phi(s,a)^\top \hat{w}_T$*, where* $\hat{w}_T$ *is defined as in* (19)*, and* Φ *can either be the reward concatenated with the Nyström or random features, i.e. either* Φ_{nys} *or* Φ_{rf}*. Let* $\phi_{\sup} := \sup_{s,a} \|\phi(s,a)\|$*. Then, for sufficiently large* n*, there exists an universal constant* $C > 0$ *independent of* m*,* n*,* T *and* $(1-\gamma)^{-1}$*, such that with probability at least* $1-\delta$*, we have*

$$\left\|\tilde{Q}^\pi_\Phi - \hat{Q}^\pi_{\Phi,T}\right\|_\nu \leqslant \gamma^T \left\|\tilde{Q}^\pi_\Phi\right\|_\nu + \frac{C\phi^6_{\sup} \text{polylog}(m, T/\delta)}{(1-\gamma)\Upsilon_1^2 \Upsilon_2 \sqrt{n}}, \tag{42}$$

where we recall $\tilde{Q}^\pi_\Phi$ *is defined as in* (27)*.*

[7] Similar analysis of statistical error of policy evaluation with linear function approximation also exist in the literature [44], [47].

Proof. We define $\tilde{w}$ which satisfies the condition

$$\tilde{w}=\left(\mathbb{E}_\nu\left[\phi(s,a)\phi(s,a)^\top\right]\right)^{-1} \left(\mathbb{E}_\nu\left[\phi(s,a)\left(r(s,a)+\gamma\mathbb{E}_{(s',a')\sim P(s,a)\times\pi}\left[\phi(s',a')^\top\tilde{w}\right]\right)\right]\right), \tag{43}$$

and let $\tilde{Q}^\pi_\Phi(s,a)=\phi(s,a)^\top\tilde{w}$. It is straightforward to see that $\tilde{w}$ is the fixed point of the population (*i.e.*, $n\to\infty$) projected least square update (19). For notational convenience, we drop the Φ in the subscript of $\Pi_{\nu,\Phi}$. With the update (19), we have that

$$\begin{aligned}\|\Phi(\tilde{w}-\hat{w}_{t+1})\|_\nu \leqslant&\gamma\|\Phi(\tilde{w}-\hat{w}_t)\|_\nu+\left\|\left(\Pi_\nu-\hat{\Pi}_\nu\right)r\right\|_\nu\\&+\gamma\left\|\left(\Pi_\nu P^\pi-\hat{\Pi}_\nu\hat{P}^\pi\right)\Phi\hat{w}_t\right\|_\nu,\end{aligned}$$

where we use the contractivity under $\|\cdot\|_\nu$. Telescoping over t leads to

$$\begin{aligned}\|\Phi(\tilde{w}-\hat{w}_T)\|_\nu \leqslant&\gamma^T\|\Phi(\tilde{w}-\hat{w}_0)\|_\nu+\frac{1}{1-\gamma}\left\|\left(\Pi_\nu-\hat{\Pi}_\nu\right)r\right\|_\nu\\&+\frac{\gamma}{1-\gamma}\max_{t\in[T]}\left\|\left(\Pi_\nu P^\pi-\hat{\Pi}_\nu\hat{P}^\pi\right)\Phi\hat{w}_t\right\|_\nu.\end{aligned}$$

With the concentration of the second and third terms as shown in Appendix VI-D, we conclude the proof. □

Proposition 3 demonstrates that, the statistical error of the linear parts can be decomposed into two parts. The first part $\gamma^T\left\|\tilde{Q}^\pi\right\|_\nu$ is due to the fact that we start from $\hat{w}_0=0$ and shrinks as the number of least square policy evaluation iterations T increase. The second part $\frac{C\tilde{g}_\alpha^6 m^3\mathrm{polylog}(m,T/\delta)}{(1-\gamma)\Upsilon_1^2\Upsilon_2\sqrt{n}}$ denotes the statistical error due to the finite sample and shrinks as the number of samples n increases. We can balance the two parts with $T=\Theta(\log n)$ and obtain an estimation error that shrinks as $O(n^{-1/2})$ with high probability.

c) Total Error for Policy Evaluation: Combine the approximation error in Proposition 1 and Proposition 2 and the statistical error in Proposition 3, we have the following bound on the error for least square policy evaluation:

Theorem 1. *Let $T=\Theta(\log n)$. With probability at least $1-\delta$, we have that for the random features Φ_{rf},*

$$\left\|Q^\pi-\hat{Q}^\pi_{\Phi_{\mathrm{rf}},T}\right\|_\nu=\tilde{O}\left(\underbrace{\frac{\tilde{g}_\alpha}{(1-\gamma)^2\sqrt{m}}}_{\text{approx. error}}+\underbrace{\frac{\tilde{g}_\alpha^6 m^3}{(1-\gamma)\Upsilon_1^2\Upsilon_2\sqrt{n}}}_{\text{stat. error}}\right), \tag{44}$$

and for the Nyström features

$$\left\|Q^\pi-\hat{Q}^\pi_{\Phi_{\mathrm{nys}},T}\right\|_\nu=\tilde{O}\left(\underbrace{\frac{\tilde{g}_\alpha}{(1-\gamma)^2 n_{\mathrm{nys}}}}_{\text{approx. error}}+\underbrace{\frac{\tilde{g}_\alpha^6 n_{nys}^3/\lambda_m^3}{(1-\gamma)\Upsilon_1^2\Upsilon_2\sqrt{n}}}_{\text{stat. error}}\right). \tag{45}$$

Proof. This directly comes from the triangle inequality, *i.e.*,

$$\left\|Q^\pi-\hat{Q}^\pi_{\Phi,T}\right\|_\nu\leqslant\left\|Q^\pi-\tilde{Q}^\pi_\Phi\right\|_\nu+\left\|\tilde{Q}^\pi_\Phi-\hat{Q}^\pi_{\Phi,T}\right\|_\nu.$$

The first approximation error term is bound using Proposition 1 and Proposition 2 for random and Nyström features respectively, while the second statistical error term can be bound using Proposition 3, and applying the result from Appendix G in our supplement [37] that for random features, $\sup_{s,a}\|\phi(s,a)\|=O(\tilde{g}_\alpha\sqrt{m})$, and that for Nystrom features, $\sup_{s,a}\|\phi(s,a)\|=O(\tilde{g}_\alpha\sqrt{\frac{n_{nys}}{\lambda_m}})$, where λ_m is the m-th largest eigenvalue of the Gram matrix $K^{(n_{nys})}$. □

Theorem 1 provides the estimation error of Q with least square policy evaluation under the $\|\cdot\|_\nu$ norm for both the random and Nyström features. It can be used to provide an estimation error of J^π shown in the following corollary. Meanwhile, it also performs as a cornerstone of control optimality analysis, as we will show in the next section. The bound also reveals a fundamental tradeoff between the approximation error and statistical error: for random features, a larger number m of features will make the finite kernel truncation capable of approximating the original infinite-dimensional function space better (cf. the $\tilde{O}\left(\frac{1}{\sqrt{m}}\right)$ approximation error term in the RHS of (44)) but it also requires a larger number of learning samples n in the policy evaluation step in order to train the weights well (cf. the $\tilde{O}\left(\frac{m^3}{\sqrt{n}}\right)$ statistical error term in the RHS of (44)). While the statistical error term in Nyström features remains the same, the approximation error term here (cf. (45)) improves to $\tilde{O}(1/n_{\mathrm{nys}})\leqslant\tilde{O}(1/m)$ (which holds since the number of features we pick, m, is always less than n_{nys} which is the number of samples used to obtain the Nyström features), smaller than the $\tilde{O}(1/\sqrt{m})$ approximation error term in (44).

B. Error Analysis of Natural Policy Gradient for Control

We now provide the error analysis for policy optimization. We start with the following "regret lemma" that is adapted from Lemma 6.2 in [38]. We remark that in the following proof we assume the action space $\mathcal{A}$ is finite for simplicity, however, the proposed SDEC is applicable for infinite action space without any modification.

Lemma 12 (cf. Lemma 6.2 in [38]). *Let $\phi_{\sup}:=\sup_{s,a}\|\phi(s,a)\|$. For a fixed comparison policy $\tilde{\pi}$, we have that*

$$\min_{k<K}\left\{V^{\tilde{\pi}}-V^{\pi_k}\right\}=O\left(\frac{1}{1-\gamma}\left(\frac{\log|\mathcal{A}|}{\eta K}+\frac{\eta\phi_{\sup}^4}{\Upsilon_2^2}+\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{err}_k\right)\right), \tag{46}$$

where err_k is defined as

$$\mathrm{err}_k:=\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[A^{\pi_k}(s,a)-\hat{w}_{k,T}^\top\left(\phi(s,a)-\mathbb{E}_{a'\sim\pi_k(s)}\left[\phi(s,a')\right]\right)\right],$$

where $A^{\pi_k}(s,a)=Q^{\pi_k}(s,a)-V^{\pi_k}(s)$, and $\hat{w}_{k,T}$ is as defined in (19).

Proof. We provide a brief sketch here. Note that $\|\phi(s,a)\|_2=O(\phi_{\sup})$. Thus, by Remark 6.7 in [38], we know that $\log\pi(a|s)$ is smooth with the smoothness parameter $\beta=O(\phi_{\sup}^2)$, with which we have

$$\begin{aligned}&\mathbb{E}_{s\sim d^{\tilde{\pi}}}\left[\mathrm{KL}(\tilde{\pi}(\cdot|s)\|\pi_k(\cdot|s))-\mathrm{KL}(\tilde{\pi}(\cdot|s)\|\pi_{k+1}(\cdot|s))\right]\\=&\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[\log\frac{\pi_{k+1}(a|s)}{\pi_k(a|s)}\right]\\\geqslant&\eta\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[\left(\nabla_\theta\log\pi_k(a|s)\right)^\top\hat{w}_k^T\right]-\frac{\eta^2\beta}{2}\left\|\hat{w}_k^T\right\|_2^2\\=&\eta\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[A^{\pi_k}(s,a)\right]-\frac{\eta^2\beta}{2}\left\|\hat{w}_k^T\right\|_2^2\\&+\eta\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[\left(\nabla_\theta\log\pi_k(a|s)\right)^\top\hat{w}_k^T-A^{\pi_k}(s,a)\right]\\=&(1-\gamma)\eta\left(V^{\tilde{\pi}}-V^{\pi_k}\right)-\frac{\eta^2\beta}{2}\left\|\hat{w}_k^T\right\|_2^2-\eta\mathrm{err}_k,\end{aligned}$$

where second inequality comes from smoothness condition and the last step comes from the performance difference lemma (Lemma 3.2 in [38]). We can then complete the proof by a telescoping sum, and also use the fact that for sufficiently large number of learning samples n, the learnt $\hat{w}_{k,T}\approx\tilde{w}$

(where $\tilde{w}$ is defined in (43)) and that $\|\tilde{w}\| = O(\phi_{\sup}\Upsilon_2^{-1})$. More details are provided in Appendix VI-E. □

Given an arbitrary policy $\tilde{\pi}$, for K steps of the policy update step in SDEC (yielding $\{V^{\pi_k}\}_{k=1}^K$), this lemma bounds a difference term where the largest of the $\{V^{\pi_k}\}_{k=1}^K$'s is subtracted from $V^{\tilde{\pi}}$. We note that the first term of the bound in (46) decreases at the rate of $O\left(K^{-1}\right)$, with the second term in the bound in (46) coming from the smoothness parameter (see Appendix VI-E for details) and the third term in the bound in (46) arising from the estimation error in the policy evaluation step. We can eventually balance the first and second terms to select the optimal stepsize η, while the third term is irreducible and inherited from the policy evaluation step. Now, we can provide the performance guarantee for the control optimality with natural policy gradient.

Theorem 2. *Let* $\eta = \Theta\left(\Upsilon_2\tilde{g}_\alpha^{-2}m^{-1}\sqrt{\log|\mathcal{A}|}\right)$ *for random features and* $\eta = \Theta\left(\Upsilon_2\tilde{g}_\alpha^{-2}n_{nys}^{-1}\lambda_m\sqrt{\log|\mathcal{A}|}\right)$ *for Nystrom features. With high probability, for the* $V_{\Phi_{\rm rf}}^{\pi_k}$ *learnt using random features* $\Phi_{\rm rf}$*, we have*

$$\mathbb{E}\left[\min_{k<K}\left\{V^{\pi^*}-V_{\Phi_{\rm rf}}^{\pi_k}\right\}\right] = \tag{47}$$
$$\tilde{O}\left(\frac{\tilde{g}_\alpha^2 m}{\Upsilon_2}\sqrt{\frac{\log|\mathcal{A}|}{K}}+\frac{1}{1-\gamma}\sqrt{\max_{s,a,\pi,k}\frac{d^{\pi^*}(s)\pi(a|s)}{\nu_{\pi_k}(s,a)}}\left(\frac{\tilde{g}_\alpha}{(1-\gamma)^2\sqrt{m}}+\frac{\tilde{g}_\alpha^6 m^3}{(1-\gamma)\Upsilon_1^2\Upsilon_2\sqrt{n}}\right)\right).$$

Meanwhile for the $V_{\Phi_{\rm nys}}^{\pi_k}$ *learnt using Nyström features* $\Phi_{\rm nys}$*, we have*

$$\mathbb{E}\left[\min_{k<K}\left\{V^{\pi^*}-V_{\Phi_{\rm nys}}^{\pi_k}\right\}\right] = \tag{48}$$
$$\tilde{O}\left(\frac{n_{nys}\lambda_m^{-1}\tilde{g}_\alpha^2}{\Upsilon_2}\sqrt{\frac{\log|\mathcal{A}|}{K}}+\frac{1}{1-\gamma}\sqrt{\max_{s,a,\pi,k}\frac{d^{\pi^*}(s)\pi(a|s)}{\nu_{\pi_k}(s,a)}}\left(\frac{\tilde{g}_\alpha}{(1-\gamma)^2 n_{\rm nys}}+\frac{\tilde{g}_\alpha^6 n_{nys}^3/\lambda_m^3)}{(1-\gamma)\Upsilon_1^2\Upsilon_2\sqrt{n}}\right)\right)$$

Proof. This is proved by setting $\tilde{\pi} := \pi^*$ in Lemma 12, selecting the appropriate η as our choice in the lemma, and characterizing the term err_k as in Appendix VI-F. Finiteness of the policy distribution ratio is ensured by the coverage conditions in Assumption 3. □

We emphasize that Theorem 2 characterizes the gap between optimal policy and the solution provided by SDEC, taking in account of finite-step in policy optimization (K), finite-dimension (m) in the feature space, finite number of samples used to generate Nyström features ($n_{\rm nys}$) and finite-sample (n) approximation in policy evaluation, respectively, which, to the best of our knowledge, is established for the first time. As we can see, for the random features, when m increases, we can reduce the approximation error, but the optimization and sample complexity will also increase accordingly. For Nyström features, this tradeoff may be better especially since the approximation term scales as $O(1/n_{\rm nys})$), and depending on the spectrum eigendecay, m may be chosen to be significantly smaller than $n_{\rm nys}$. In either case, we can further balance the terms for the optimal dimension of features. Finally, we note that the final bound has a polynomial dependence on $\tilde{g}_\alpha$, which at first glance scales exponentially with c_f. However, α can be picked to be $\tilde{O}(1/c_f)$ to cancel out this dependence.

V. Simulations

A. Details of SDEC implementation

In this section, we will empirically implement the SDEC and compare it with other methods. As we discussed in Remark 2, the advantage of the proposed spectral dynamics embedding is that it is compatible with a wide range of dynamical programming/reinforcement learning methods, allowing us to take advantage of the computational tools recently developed in deep reinforcement learning. Though in SDEC we stick to a particular type of policy evaluation and policy update for the purpose of theory development, in our empirical implementation of SDEC, we combine spectral dynamicsal embedding with Soft Actor-Critic (SAC) [21], a cutting-edge deep reinforcement learning method that has demonstrated strong empirical performance. Specifically, we use random features or Nyström to parameterize the critic, and use this as the critic in SAC. In SAC, it is necessary to maintain a parameterized function $Q(s,a)$ which estimates the soft Q-value (which includes not just the reward but also an entropy term encouraging exploration). For a given policy π, the soft Q-value satisfies the relationship $Q^\pi(s_t,a_t) = r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}\sim p}[V(s_{t+1})]$, where $V(s_t) = \mathbb{E}_{a_t\sim\pi}[Q(s_t,a_t) - \beta\log\pi(a_t \mid s_t)]$. Based on the spectral dynamics embedding proposed in our paper, we parameterize the Q^π-function as $Q^\pi(s,a) = r(s,a) + \tilde{\phi}_\omega(s,a)^\top\tilde{\theta}^\pi$ where $\tilde{\phi}_\omega(s,a)$ is our feature approximations. We use two types of feature approximations, random features and Nyström features. The random features are composed by $\tilde{\phi}_\omega(s,a) = \left[\cos(\omega_1^\top s'+b_1),\cdots,\cos(\omega_m^\top s'+b_m)\right]^\top$, where $s' = f(s,a)$, with $\{\omega_i\}_{i=1}^m \sim \mathcal{N}(0,I_d)$ and $\{b_i\}$ drawn iid from $\mathrm{Unif}([0,2\pi])$. The $\tilde{\theta}^\pi \in \mathbb{R}^m$ is updated using (mini-batch) gradient descent for (19). For the Nyström feature, we first sample m data $s_1,\ldots,s_m$ uniformly in the state space to construct the corresponding Gram matrix $K \in \mathbb{R}^{m\times m}$. Then we do spectral decomposition of K to get orthogonal eigenvectors U and diagonal matrix Λ. The Nyström features are composed by $\tilde{\phi}_\omega(s,a) = [k(s',s_1)\ldots,k(s',s_m)]U\Lambda^{-1/2}$. We then use this Q-function for the actor update in SAC.

B. Experimental Setup

We implemented the proposed SDEC algorithm on four nonlinear control tasks, Pendulum swingup, pendubot balance, cartpole swingup, and 2D drone hovering. Pendulum and Cartpole are two classic nonlinear control tasks, the tasks are swinging up the pendulum to the upright position. Pendubot is another classic underactuated system where two rods are linked together with an un-acturated joint [48]. The task is to balance the two rods in the upright position. The 2D drone task simplifies drones to a 2D plane with two motors. The task is hovering the drone at a fixed position.

C. Performance versus other nonlinear controllers

We compare SDEC with iterative LQR (iLQR) [6] and Koopman-based control [14], Energy-based control [49], [50], and soft actor-critic (SAC) [21]. We note that iLQR, the energy-based controller and SDEC all make use of the knowledge of the environment dynamics. DKUC and SAC do not make use of the environment dynamics. The first two are well-known alternatives for nonlinear control, and the energy-based control is a Lyapunov-based method to design stabilizing controllers. For iLQR, we used the implementation in [51], where we added a log barrier function to account for the actuator constraint. Meanwhile, for the Koopman-based control, we adapted an implementation called Deep KoopmanU with Control (DKUC) proposed in the paper [52],

where we combine the dynamics learned from DKUC with MPC to enforce the input constraint. We modified the energy-based swing-up controller from [50]. For SDEC and DKUC, we trained using 4 different random seeds, and both algorithms have access to the same amount of environment interaction.

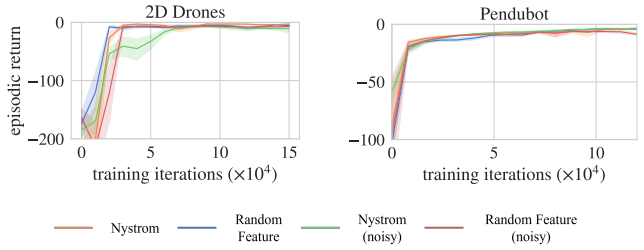

Fig. 1. SDEC performances with random features and Nyström on noisy and non-noisy environments.

f

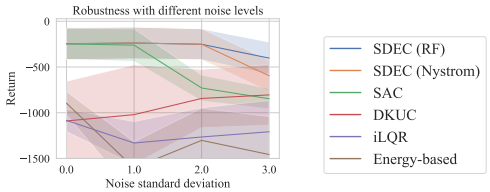

Fig. 2. Performances of all algorithms on Pendulum with varying noise levels $\sigma \in \{0, 1, 2, 3\}$.

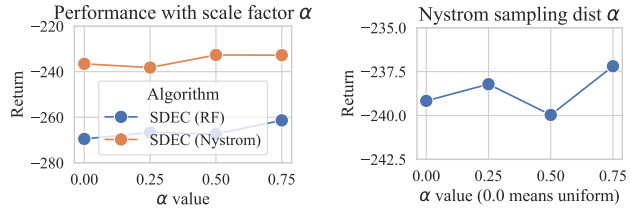

Fig. 3. SDEC performance on the Pendulum with different α in scaling $f(s, a)$ (left) and in Nystrom sampling distribution (right).

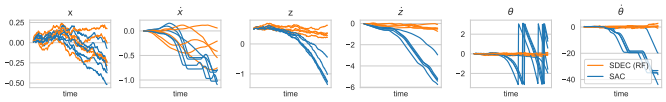

Fig. 4. 5 trajectories of 2D quadrotor controlled by SDEC and SAC.

The performance of the various algorithms in both the noiseless and noisy setting (with $\sigma = 1$) can be found in Table I. The mean and standard deviation (in brackets) over 4 random seeds are shown. The number of random and Nyström features used is 512 for pendulum, 4096 for pendubot and 2D Drones and 8192 for Cartpole. We observe that the SDEC strongly outperforms other nonlinear controllers in the noisy setting (higher reward is better; in this case the negative sign is due to the rewards representing negative cost). For the noiseless setting, SDEC are comparable to the best nonlinear baseline controllers. For comparison with SAC, we observe different results depending on whether the system is fully actuated. For the pendulum and 2D drones where the systems are fully actuated, we observe significant advantages of the proposed SDEC over baselines, especially in noisy environments. For the Cartpole and pendubot that are underacutated, the proposed SDEC can still outperform other nonlinear control baselines but achieve comparable performance with SAC. We further plot the performance of SDEC during the course of its learning in Figure 1 and 2. The evaluation is performed every 25 learning episodes on a fresh evaluation set of 100 episodes, and the y-axis represents the average episodic reward on the evaluation set. The shaded regions represent a 1 standard deviation confidence interval (across 4 random seeds). Figure 1 compares performance with different features and noise levels. In both the noiseless and noisy settings, the performance of SDEC continuously improves until it saturated after a few episodes. Figure 2 shows the evaluation performance on the pendulum swingup task during training with noisier environments. The performance still continuously improves and then saturates. Performance with different finite approximation dimensions are compared in Figure 5. As dimension increases, the performance improves. Nyström features are constantly better than random features and benefit more from dimension increase, which aligns with our theoretical results.

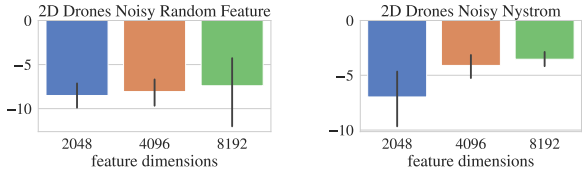

Fig. 5. Performance after training with different feature dimensions on noisy 2D drones with random and Nyström features.

D. Robustness and Stability

To evaluate the robustness of the proposed algorithm, We show the performance under different noise levels in Figure 2. We can see that both two SDEC algorithms show good robustness against increasing noise levels, while SAC has significantly worse performance as the noise level increases. As for all the nonlinear control baselines, they are not able to obtain good performance with noise dynamics. We can study stability using the results in Table I. The reward functions are designed as the squared error with respect to the equilibrium state. We see the proposed SDEC shows good stability, especially with noise. We also show 5 trajectories of noisy 2D drones controlled by SDEC and SAC in Figure 4. The goal is to stabilize the drone at the initial position $[0.0, 0.5, 0.0]$. The initial state is the origin and the noise level is $\sigma = 1.0$. The six plots show six drone states including the 2D position and the roll angle $[x, z, \theta]$, and their time derivatives. The trajectories show that SDEC can stabilize the drone even with the presence of noise. However, with SAC, the drones quickly become unstable and flip over due to the noise.

VI. Conclusion

There has been a long-standing gap between the theoretical understanding and empirical success of the kernelized

TABLE I
PERFORMANCE COMPARISON. BOLD NUMBERS ARE THE BEST AVERAGE PERFORMANCE OVER ALL ALGORITHMS.

	Pendulum	Pendulum (Noisy)	Pendubot	Pendubot (Noisy)	2D Drones	2D Drones (Noisy)	Cartpole	Cartpole (Noisy)
SDEC(RF)	-279.0±175	-240±173	-0.793±0.190	-3.883±0.283	-5.074±2.023	-7.423±2.549	-212±116	-211±93
SDEC(Nyström)	**-245.0±174**	**-225±156**	-2.635±1.072	-3.880±1.192	-2.315±0.323	**-4.424±0.924**	-216±118	**-207±90**
iLQR	-1084±544	-1191±415	-429.6±113.8	-1339±605	**-0.973±0.841**	753.7±119.7	-430±87.3	-705±59
DKUC	-1090±436	-1021±546	-446.0±54.2	-330.8±140.1	-750.8±427.7	-867.0±94.9	-486±189	-774±309
Energy-based	-895.8±128	-1604±288	-0.797±0.588	-16.63±5.62	-322.9±14.6	-339.8±35.3	-779±247	-694±91
SAC	-250±173	-261±179	**-0.295±0.132**	**-0.751±0.458**	-10.87±12.15	-20.06±14.82	**-206±100**	-214±96

linearization control with spectral dynamics embedding, *i.e.*, the error induced by finite-dimensional approximation has not be clearly analyzed. In this work, we close this gap by studying the policy evaluation and optimization error for finite-dimensional random and Nystrom features, theoretically validating the practical usage of SDEC. Along the way, we have also developed new analysis tools and results that may be of independent interest, e.g. the novel $O(1/n_{\text{nys}})$ high-probability convergence rate for Nyström kernel approximation.

REFERENCES

[1] J.-J. E. Slotine, W. Li *et al.*, *Applied nonlinear control*. Prentice hall Englewood Cliffs, NJ, 1991, vol. 199, no. 1.

[2] H. Khalil, *Nonlinear Control*. Pearson Education Limited, 2015.

[3] D. A. Lawrence and W. J. Rugh, "Gain scheduling dynamic linear controllers for a nonlinear plant," *Automatica*, vol. 31, no. 3, 1995.

[4] W. J. Rugh and J. S. Shamma, "Research on gain scheduling," *Automatica*, vol. 36, no. 10, pp. 1401–1425, 2000.

[5] B. Charlet, J. Lévine, and R. Marino, "On dynamic feedback linearization," *Systems & Control Letters*, vol. 13, no. 2, pp. 143–151, 1989.

[6] A. Sideris and J. E. Bobrow, "An efficient sequential linear quadratic algorithm for solving nonlinear optimal control problems," in *Proceedings of American Control Conference.* IEEE, 2005, pp. 2275–2280.

[7] C. Edwards and S. Spurgeon, *Sliding mode control: theory and applications*. Crc Press, 1998.

[8] R. Brockett, "The early days of geometric nonlinear control," *Automatica*, vol. 50, no. 9, pp. 2203–2224, 2014.

[9] P. V. Kokotovic, "The joy of feedback: nonlinear and adaptive," *IEEE Control Systems Magazine*, vol. 12, no. 3, pp. 7–17, 1992.

[10] J. A. Primbs, V. Nevistić, and J. C. Doyle, "Nonlinear optimal control: A control lyapunov function and receding horizon perspective," *Asian Journal of Control*, vol. 1, no. 1, pp. 14–24, 1999.

[11] J. B. Rawlings, D. Q. Mayne, and M. Diehl, *Model predictive control: theory, computation, and design*. Nob Hill Publishing Madison, WI, 2017, vol. 2.

[12] S. Prajna, A. Papachristodoulou, P. Seiler, and P. A. Parrilo, "Sostools and its control applications," in *Positive polynomials in control*. Springer, 2005, pp. 273–292.

[13] B. O. Koopman, "Hamiltonian systems and transformation in hilbert space," *Proceedings of the national academy of sciences of the united states of america*, vol. 17, no. 5, p. 315, 1931.

[14] M. Korda and I. Mezić, "Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control," *Automatica*, vol. 93, pp. 149–160, 2018.

[15] S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi, and W. Sun, "Information theoretic regret bounds for online nonlinear control," *Adv. in Neural Info. Processing Systems*, vol. 33, pp. 15 312–15 325, 2020.

[16] S. Grunewalder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton, "Modelling transition dynamics in mdps with rkhs embeddings," *arXiv preprint arXiv:1206.4655*, 2012.

[17] R. Kamalapurkar and J. A. Rosenfeld, "An occupation kernel approach to optimal control," *arXiv preprint arXiv:2106.00663*, 2021.

[18] T. Ren, T. Zhang, C. Szepesvári, and B. Dai, "A free lunch from the noise: Provable and practical exploration for representation learning," in *Uncertainty in Artificial Intelligence*. PMLR, 2022, pp. 1686–1696.

[19] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," *Advances in neural information processing systems*, vol. 21, 2008.

[20] S. Hayakawa, H. Oberhauser, and T. Lyons, "Sampling-based nystr\" om approximation and kernel quadrature," *arXiv preprint arXiv:2301.09517*, 2023.

[21] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, 2018.

[22] T. Ren, Z. Ren, N. Li, and B. Dai, "Stochastic nonlinear control via finite-dimensional spectral dynamic embedding," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 795–800.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[24] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Conference on Learning Theory*. PMLR, 2020, pp. 2137–2143.

[25] O. Hernández-Lerma and J. B. Lasserre, "Approximation schemes for infinite linear programs," *SIAM Journal on Optimization*, vol. 8, no. 4, pp. 973–988, 1998.

[26] ——, *Discrete-time Markov control processes: basic optimality criteria*. Springer Science & Business Media, 2012, vol. 30.

[27] G. E. Fasshauer, "Positive definite kernels: past, present and future," *Dolomites Research Notes on Approximation*, vol. 4, pp. 21–63, 2011.

[28] A. Devinatz, "Integral representations of positive definite functions," *Transactions of the American Mathematical Society*, vol. 74, no. 1, pp. 56–77, 1953.

[29] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philos. Trinsdictions Rogyal Soc*, vol. 209, pp. 415–446, 1909.

[30] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Adv. in Neural Info. Processing Systems*, vol. 20, 2007.

[31] B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M.-F. Balcan, and L. Song, "Scalable kernel methods via doubly stochastic gradients," *Adv. in Neural Info. Processing Systems*, vol. 27, pp. 3041–3049, 2014.

[32] H. Sun, "Mercer theorem for rkhs on noncompact sets," *Journal of Complexity*, vol. 21, no. 3, pp. 337–349, 2005.

[33] C. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," *Advances in Neural Information Processing Systems*, vol. 13, pp. 682–688, 2000.

[34] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, vol. 68, no. 3, pp. 337–404, 1950.

[35] L. Yang and M. Wang, "Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 746–10 756.

[36] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou, "Nyström method vs random fourier features: A theoretical and empirical comparison," *Advances in neural information processing systems*, vol. 25, 2012.

[37] Z. Ren, T. Ren, H. Ma, N. Li, and B. Dai, "Stochastic nonlinear control via finite-dimensional spectral dynamic embedding," https://www.zhaolinren.com/files/SDECsupplement.pdf, 2025.

[38] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *J Mach Learn Res*, vol. 22, no. 98, pp. 1–76, 2021.

[39] A. S. Nemirovskij and D. B. Yudin, "Problem complexity and method efficiency in optimization," 1983.

[40] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[41] H. Yao, C. Szepesvári, B. A. Pires, and X. Zhang, "Pseudo-mdps and factored linear action models," in *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, 2014, pp. 1–9.

[42] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," *Machine learning*, vol. 49, no. 2, pp. 209–232, 2002.

[43] H. Yu and D. P. Bertsekas, "New error bounds for approximations from projected linear equations," in *European Workshop on Reinforcement Learning*. Springer, 2008, pp. 253–267.

[44] Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, C. Szepesvari, and G. Weisz, "Politex: Regret bounds for policy iteration using expert prediction," in *International Conference on Machine Learning*, 2019.

[45] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local rademacher complexities," 2005.

[46] M. Belkin, "Approximation beats concentration? an approximation view on inference with smooth radial kernels," in *Conference On Learning Theory*. PMLR, 2018, pp. 1348–1361.

[47] Y. Duan, Z. Jia, and M. Wang, "Minimax-optimal off-policy evaluation with linear function approximation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2701–2709.

[48] M. W. Spong and D. J. Block, "The pendubot: A mechatronic system for control research and education," in *Proceedings of 1995 34th IEEE conference on decision and control*, vol. 1. IEEE, 1995, pp. 555–556.

[49] I. Fantoni, R. Lozano, and M. W. Spong, "Energy based control of the pendubot," *IEEE transactions on automatic control*, vol. 45, no. 4, pp. 725–729, 2000.

[50] K. J. Åström and K. Furuta, "Swinging up a pendulum by energy control," *Automatica*, vol. 36, no. 2, pp. 287–295, 2000.

[51] V. Roulet, S. Srinivasa, M. Fazel, and Z. Harchaoui, "Iterative linear quadratic optimization for nonlinear control: Differentiable programming algorithmic templates," *arXiv preprint arXiv:2207.06362*, 2022.

[52] H. Shi and M. Q.-H. Meng, "Deep koopman operator with control for nonlinear systems," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7700–7707, 2022.

[53] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[54] J. A. Tropp *et al.*, "An introduction to matrix concentration inequalities," *Foundations and Trends® in Machine Learning*, vol. 8, no. 1-2, pp. 1–230, 2015.

APPENDIX

A. Random Feature for Exponential Family Transition

With the transition operator in (23), we have

$$
\begin{aligned}
&P(s'|s,a) = \exp\left(A\left(s,a\right)\right)\exp\left(f\left(s,a\right)^{\top}\zeta\left(s'\right)\right)p(s')\\
&=\underbrace{\exp\left(A\left(s,a\right)\right)\exp\left(\frac{\|f(s,a)\|^2}{2}\right)}_{h(s,a)}\\
&\cdot\exp\left(-\frac{\|f\left(s,a\right)-\zeta\left(s'\right)\|^2}{2}\right)\underbrace{\exp\left(\frac{\|\zeta\left(s'\right)\|^2}{2}\right)p(s')}_{g(s')}\\
&=\mathbb{E}_{\omega}\left[\left\langle h\left(s,a\right)\psi_{\omega}\left(f(s,a)\right),\mu_{\omega}\left(s'\right)g(s')\right\rangle\right],
\end{aligned}
$$

where $\psi_{\omega,b}\left(f(s,a)\right) = \cos\left(\omega^{\top}f\left(s,a\right)+b\right)$ and $\mu_{\omega,b} = \cos\left(\omega^{\top}\zeta\left(s'\right)+b\right)$ with $\omega\sim\mathcal{N}\left(0,\mathbf{I}\right)$ and $b\sim U([0,2\pi])$.

B. Nyström features

We refer to Appendix VI-B in our supplement [37] for a more detailed derivation and justification of Nyström features. For simplicity, in the sections relating to the Nyström method, let $n := n_{\text{nys}}$ denote the number of samples used to learn the Nyström features. In brief, for any $m \leqslant n$, we approximate k_α by the rank-m approximation $\hat{k}^n_{\alpha,m}$, where

$$
\hat{k}^n_{\alpha,m}(x,y) = k_\alpha(x,X^n)\left(\sum_{i=1}^{m}\frac{1}{\lambda_i}u_iu_i^{\top}\right)k_\alpha(X^n,x), \tag{49}
$$

and X^n is n i.i.d samples drawn from μ_{Nys}. This is the Nyström kernel approximation that we will consider in the paper. Since it can be shown that $\hat{k}^n_{\alpha,m}(x,y) = \varphi_{\text{nys}}(x)^{\top}\varphi_{\text{nys}}(y)$, where

$$
(\varphi_{\text{nys}})_i(\cdot) = \frac{1}{\sqrt{\lambda_i}}u_i^{\top}k_\alpha(X^n,\cdot), \quad \forall i\in[m], \tag{50}
$$

we may view $\varphi_{\text{nys}}(\cdot)\in\mathbb{R}^m$ as the (rank-m) Nyström features associated with the Nyström kernel approximation.

C. Detailed Proof for Proposition 11 (Kernel approximation error of Nyström features)

By Lemma 3, in the RKHS corresponding to k_α, which we denote as $\mathcal{H}_{k_\alpha}$, given any $x,y\in\mathcal{S}$, $\langle k_\alpha(x,\cdot),k_\alpha(y,\cdot)\rangle_{\mathcal{H}_{k_\alpha}} = k_\alpha(x,y)$. Let $\phi(\cdot)\in\mathbb{R}^m$ be as defined in (50). Note that for each $i\in[m]$, we can view $(\varphi_{\text{nys}})_i(\cdot)$ as an element in $\mathcal{H}_{k_\alpha}$. This and several other useful properties of the features $\{(\varphi_{\text{nys}})_i(\cdot)\}_{i=1}^m$ is shown in this next result (cf. [20]).

Lemma 13. *Let $(\varphi_{\text{nys}})_i(\cdot) := \frac{1}{\sqrt{\lambda_i}}u_i^{\top}k_\alpha(X^n,\cdot)$ for each $i\in[m]$. Then, the following statements hold.*

1) *The set $\{(\varphi_{\text{nys}})_i(\cdot)\}_{i=1}^m$ is orthonormal in $\mathcal{H}_{k_\alpha}$.*
2) *Define an orthogonal projection map $P_m^{X^n}:\mathcal{H}_{k_\alpha}\to\mathcal{H}_{k_\alpha}$ onto the span of $\{(\varphi_{\text{nys}})_i(\cdot)\}_{i=1}^m$, such that*

$$
P_m^{X^n}(h(\cdot)) = \sum_{i=1}^{m}\langle h(\cdot),(\varphi_{\text{nys}})_i(\cdot)\rangle_{\mathcal{H}_{k_\alpha}}(\varphi_{\text{nys}})_i(\cdot).
$$

Then, $P_m^{X^n}k_\alpha(x,\cdot) = \hat{k}^n_{\alpha,m}(x,\cdot)$ for every $x\in\mathcal{S}$. Thus, for every $(x,y)\in\mathcal{S}$, recalling the definition of $\hat{k}^n_{\alpha,m}$ in (49), we have

$$
\begin{aligned}
\hat{k}^n_{\alpha,m}(x,y) &= \langle k_\alpha(x,\cdot),\hat{k}^n_{\alpha,m}(\cdot,y)\rangle_{\mathcal{H}_{k_\alpha}}\\
&= \langle P_m^{X^n}k_\alpha(x,\cdot),P_m^{X^n}k_\alpha(y,\cdot)\rangle_{\mathcal{H}_{k_\alpha}}
\end{aligned}
$$

3) *Let $P^{X^n}_{m,\perp}$ denote the orthogonal projection onto the complement of $\{(\varphi_{\text{nys}})_i(\cdot)\}_{i=1}^m$ in $\mathcal{H}_{k_\alpha}$. Then,*

$$
\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}} = \sqrt{(k_\alpha-\hat{k}^n_{\alpha,m})(x,x)} \tag{51}
$$

We defer the proof to our supplement [37], and note that the lemma follows from results in [20]. We next bound the approximation error of the Nyström method, i.e. find a bound for $\int_{\mathcal{S}}\sqrt{(k_\alpha-\hat{k}^n_{\alpha,m})(x,x)}d\mu(x)$, which by (51) is equal to $\int_{\mathcal{S}}\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_K}d\mu(x)$. To do so, we need a few preliminary results. We first introduce some additional notation. For any distance d and class of functions $\mathcal{F}$, we define

$$
\mathcal{N}(\epsilon,\mathcal{F},d) := \min_{G\in\mathcal{G}}|G|, \quad \mathcal{G} := \{G:\forall f\in\mathcal{F},\exists g\in G \text{ s.t. } d(f,g)\leqslant\epsilon\},
$$

and $|G|$ denotes the cardinality of the set G. In addition, for any function $f\in\mathcal{L}_2(\mu)$, we denote $\mu(f) := \int_{\mathcal{S}}f(x)d\mu(x)$, and $\mu_{X^n}(f) = \frac{1}{n}\sum_{i=1}^n f(x_i)$. Given any $f,g\in\mathcal{L}_2(\mu)$, we also define $d_{L_2(\mu_{X^n})}(f,g) := \sqrt{\frac{1}{n}\sum_{i=1}^n(f(x_i)-g(x_i))^2}$. The first result below, which utillizies local Rademacher complexity [45], is based on Corollary 3.7 in [45], but applies to classes of real-valued function (as oppposed to binary functions as stated in Corollary 3.7 in [45]).

Lemma 14. *Let $\mathcal{F}$ be a class of $[0,1]$-valued functions such that for any $\epsilon>0$*

$$
\log\mathcal{N}(\epsilon,\mathcal{F},d_{L_2(\mu_{X^n})}) \leqslant S\log(1+4/\epsilon).
$$

Then, there exists a constant $c>0$ such that for any $\delta>0$ and any $C>0$, with probability at least $1-\delta$, for any $f\in\mathcal{F}$,

$$
\mu(f) \leqslant \frac{C}{C-1}\mu_{X^n}f + cC\left(\frac{S\log(n/S)}{n}+\frac{\log(1/\delta)}{n}\right).
$$

Proof. The result follows naturally by using the log covering number bound in the proof of Corollary 3.7 in [45]. □

We note that local Rademacher complexity is critical in deriving the $O(1/n)$ convergence rate above. Standard Rademacher complexity bounds [40] can only ensure a $O(1/\sqrt{n})$ convergence rate, which is strictly worse. The next result we need is this lemma, which builds on Lemma 14.

Lemma 15. *Suppose Π_r is an arbitrary deterministic r-dimensional orthogonal projection in $\mathcal{H}_{k_\alpha}$. Then, there exists*

a constant $c > 0$ such that for any constant $C > 1$ and orthogonal projection P in $\mathcal{H}_{k_\alpha}$ possibly depending on X^n, we have that for any $\delta > 0$, with probability at least $1-\delta$,

$$\mu(\|P\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) \leqslant \frac{C}{C-1}\mu_{X^n}(\|P\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) + cC\left(\frac{r^2\max\{\log(n/r^2),1\}}{n} + \frac{\log(1/\delta)}{n}\right).$$

Proof. First, given any orthogonal projection P in $\mathcal{H}_{k_\alpha}$, let $\{u_i\}_{i\in I}$ denote an orthonormal basis for $P\mathcal{H}_{k_\alpha}$, and let $P_\perp$ denotes the orthogonal projection onto its complement. In addition, let $\{e_1,\dots,e_r\}$ denote a basis for $\Pi_r\mathcal{H}_{k_\alpha}$. We note that for any projection P in $\mathcal{H}_{k_\alpha}$, straightforward calculations show that (see details in online report)

$$\|P\Pi_r k_\alpha(x,\cdot)\|^2_{\mathcal{H}_{k_\alpha}} = v(x)^\top A_{P,\Pi_r} v(x),$$

where $v(x) = (e_1(x), e_2(x), \dots, e_r(x)) \in \mathbb{R}^r$, and $A_{P,\Pi_r} \in \mathbb{R}^{r\times r}$ is a r by r dimensional positive semidefinite matrix where $(A_{P,\Pi_r})_{\ell,s} = \langle Pe_\ell, Pe_s\rangle_{\mathcal{H}_{k_\alpha}}$. We can similarly define a $r\times r$ positive semidefinite matrix $A_{P_\perp,\Pi_r} \in \mathbb{R}^{r\times r}$ where $(A_{P_\perp,\Pi_r})_{\ell,s} = \langle P_\perp e_\ell, P_\perp e_s\rangle_{\mathcal{H}_{k_\alpha}}$. Since

$$(A_{P,\Pi_r})_{\ell,s} + (A_{P_\perp,\Pi_r})_{\ell,s} = \langle Pe_\ell, Pe_s\rangle_{\mathcal{H}_{k_\alpha}} + \langle P_\perp e_\ell, P_\perp e_s\rangle_{\mathcal{H}_{k_\alpha}} = \langle e_\ell, e_s\rangle_{\mathcal{H}_{k_\alpha}} = \delta_{\ell,s},$$

it follows that $A_{P,\Pi_r} \preceq I_{r\times r}$, and hence can be decomposed as $A_{P,\Pi_r} = U^\top U$ for some matrix $U \in \mathbb{R}^{r\times r}$ such that $\|U\|_2 \leqslant 1$. Thus, we can reexpress $\|P\Pi_r\mathcal{H}_{k_\alpha}\|_{\mathcal{H}_{k_\alpha}}$ as

$$\|P\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}} = \|Uv(x)\|_2$$

Since this holds for any random orthogonal projection P, the class of functions $P\Pi_r k_\alpha(x,\cdot)$ is a subset of the class of functions $F_U := \{f_U(x) = \|Uv(x)\|_2 : U \in \mathbb{R}^{r\times r}\}$. We next seek to find a covering number bound for the class F_U. We note that straightforward calculations show that if we have an $\epsilon/\sqrt{2}$ covering set $\mathcal{U}_\epsilon$ of $\mathcal{U} := \{U : U \in \mathbb{R}^{r\times r}, \|U\|_2 \leqslant 1\}$ under the distance $\|\cdot\|_2$, then we have an ϵ-covering F_ϵ of F_U under the distance $d_{L_2(\mu_{X^n})}$. Thus, since $\log\mathcal{N}(\epsilon/\sqrt{2}, \mathcal{U}, \|\cdot\|_2) \leqslant r^2\log(1+4/\epsilon)$ (as $\mathcal{U}$ is the unit ball of $\mathcal{R}^{r^2}$ under the spectral norm; see Example 5.8 in [53]), it follows that $\log\mathcal{N}(\epsilon, F_U, d_{L_2(\mu_{X^n})}) \leqslant r^2\log(1+4/\epsilon)$ as well. Our result then follows by applying Lemma 14. □

We note Lemma 15 is a new result, and it is key to obtaining the fast rate of $O\left(n_{\text{nys}}^{-1}\right)$ for Nyström kernel approximation in Lemma 11. Now we prove Lemma 11.

Proof of Lemma 11. Recall that for any function $f \in \mathcal{L}_2(\mu)$, we denote $\mu(f) := \int_{\mathcal{S}} f(x)d\mu(x)$, and $\mu_{X^n}(f) = \frac{1}{n}\sum_{i=1}^n f(x_i)$. Observe that by (51), we have

$$\sqrt{(k_\alpha - \hat{k}^n_{\alpha,m})(x,x)} = \|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}.$$

Thus it suffices for us to bound $\mu(\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) = \int_{\mathcal{S}}\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}d\mu(x)$. We follow the idea in the proof of Proposition 9 in [20], and use the observation that for any $C > 0$ and $f, f_- \in \mathcal{L}_2(\mu)$ where $f_- \leqslant f$ pointwise, we have

$$\begin{aligned}&\mu(f) - C\mu_{X^n}(f)\\ &= (\mu(f) - \mu(f_-)) + (\mu(f_-) - C\mu_{X^n}(f_-)) + C(\mu_{X^n}(f_-) - \mu_{X^n}(f))\\ &\leqslant (\mu(f) - \mu(f_-)) + (\mu(f_-) - C\mu_{X^n}(f_-)),\end{aligned} \tag{52}$$

where we used the fact that $f_- \leqslant f$ pointwise to derive the inequality above. Now, consider setting

$$f(x) = \|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}},$$
$$f_-(x) = \|P^{X^n}_{m,\perp}\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}} - \|P^{X^n}_{m,\perp}\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}},$$

where Π_r denotes the orthogonal projection onto $\{e_1,e_2,\dots,e_r\} \subset \mathcal{H}_{k_\alpha}$ which are the first r eigenfunctions in the Mercer expansion of $k_\alpha(x,\cdot)$ (so each e_i satisfies $\|e_i\|_{\mathcal{H}_{k_\alpha}} = 1$), and $\Pi_{r,\perp}$ is the orthogonal projection onto the complement of $\{e_1,\dots,e_r\}$ in $\mathcal{H}_{k_\alpha}$. We first note that $f_-(x) \leqslant f(x)$ holds pointwise by the triangle inequality. Continuing from (52), picking $C = 2$, we have via appropriate decompositions (see supplement [37] for details)

$$\begin{aligned}&\mu(\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) - 2\mu_{X^n}(\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})\\ &\leqslant 2\mu\left(\|P^{X^n}_{m,\perp}\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}\right)\\ &\quad+\mu\left(\|P^{X^n}_{m,\perp}\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}\right) - 2\mu_{X^n}\left(\|P^{X^n}_{m,\perp}\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}\right)\\ &\quad+2\mu_{X^n}\left(\|P^{X^n}_{m,\perp}\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}\right)\end{aligned} \tag{53}$$

From (53), we see that we need to bound three terms. The first term we need to bound is $2\mu(\|P^{X^n}_{m,\perp}\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})$. By nonexpansiveness of $P^{X^n}_{m,\perp}$,

$$\begin{aligned}&2\mu(\|P^{X^n}_{m,\perp}\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) \leqslant 2\mu(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})\\ &\leqslant 2\sqrt{\mu(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|^2_{\mathcal{H}_{k_\alpha}})} = 2\sqrt{\sum_{j>r}e_j(x)^2} = 2\sqrt{\sum_{j>r}\sigma_j}.\end{aligned} \tag{54}$$

Next we bound

$$\left(\mu(\|P^{X^n}_{m,\perp}\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) - 2\mu_{X^n}(\|P^{X^n}_{m,\perp}\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})\right).$$

To do so, we apply Lemma 15 (by picking $C = 2$ in that proposition) , which shows that there exists $c > 0$ such that for any $\delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned}&\left(\mu(\|P^{X^n}_{m,\perp}\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) - 2\mu_{X^n}(\|P^{X^n}_{m,\perp}\Pi_r k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})\right)\\ &\leqslant c\left(\frac{r^2\max\{\log(n/r^2),1\}}{n} + \frac{\log(1/\delta)}{n}\right).\end{aligned} \tag{55}$$

Finally, to bound $\mu_{X^n}(\|P^{X^n}_{m,\perp}\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})$, we use Bernstein's inequality [53]. It suffices to bound $\mu_{X^n}(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})$ since $\|P^{X^n}_{m,\perp}\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}} \leqslant \|\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}$. Next, observe that

$$\text{Var}(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) \leqslant \mu\left(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|^2_{\mathcal{H}_{k_\alpha}}\right) = \sum_{j>r}\sigma_j.$$

Then, since $0 \leqslant \sqrt{(k_\alpha - \hat{k}^n_{\alpha,m})(x,x)} \leqslant \sqrt{k_\alpha(x,x)} \leqslant 1$for any $x \in \mathcal{X}$, by Bernstein's inequality, there exists a constant $c > 0$ such that for any $\delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned}&\mu_{X^n}(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) - \mu(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})\\ &\leqslant c\log(1/\delta)\left(\sqrt{\text{Var}(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})}/\sqrt{n} + 1/n\right)\\ &\leqslant c\log(1/\delta)\left(\sqrt{\sum_{j>r}\sigma_j}/\sqrt{n} + 1/n\right)\end{aligned}$$

Since $\mu(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) \leqslant \sqrt{\sum_{j>r}\sigma_j}$ (this follows from (54)), we thus have (with probability at least $1-\delta$),

$$\begin{aligned}&\mu_{X^n}(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})\\ &\leqslant \mu(\|\Pi_{r,\perp}k_\alpha(x,\cdot)\}\|_{\mathcal{H}_{k_\alpha}}) + c\log(1/\delta)\left(\sqrt{\sum_{j>r}\sigma_j}/\sqrt{n} + 1/n\right)\\ &\leqslant \sqrt{\sum_{j>r}\sigma_j} + c\log(1/\delta)\left(\sqrt{\sum_{j>r}\sigma_j}/\sqrt{n} + 1/n\right).\end{aligned} \tag{56}$$

Thus, by plugging (54), (55) and (56) into (53), we find that

$$\begin{aligned}&\mu(\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}) - 2\mu_{X^n}(\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})\\ &\leqslant c\log(1/\delta)\left(\sqrt{\sum_{j>r}\sigma_j} + \frac{r^2\max\{\log(n/r^2),1\}}{n}\right).\end{aligned} \tag{57}$$

By picking $r \geqslant \lfloor(2\log n)^h/\beta^h\rfloor$ and the argument in Remark 1 in [20], using the decay assumption on $\{\sigma_j\}$, we have that

$$\sqrt{\sum_{j>r}\sigma_j} + \frac{r^2\max\{\log(n/r^2),1\}}{n} = O\left(\frac{(\log n)^{2h+1}}{n}\right) \tag{58}$$

Thus, to bound $\mu(\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}})$, all it remains is to

bound $2\mu_{X^n}(\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|_{\mathcal{H}_{k_\alpha}}$. To this end, observe that

$$\mu_{X^n}(\|P^{X^n}_{m,\perp}k_\alpha(x,\cdot)\|) = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(k_\alpha - \hat{k}^n_{\alpha,m})(x_i,x_i)}$$
$$\leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sqrt{(k_\alpha - \hat{k}^n_{\alpha,m})(x_i,x_i)}^2} = \sqrt{\frac{1}{n}\sum_{i>m}\lambda_i}, \tag{59}$$

where the first inequality follows from Jensen's, and the second follows from Lemma 2 in [20]. The result then follows by plugging in (58) and (59) into (57), and using the assumption that $m \geqslant \lfloor (2\log n)^h/\beta^h \rfloor$ as well as the argument in Remark 1 in [20], which by the decay assumption on the empirical eigenvalues gives us $\sqrt{\sum_{i>m} n^{-1}\lambda_i} = O\left(n^{-1}(\log n)^{h/2}\right)$. □

Remark 5. *Our result significantly strengthens the high-probability result in Corollary 1 of [20], since we attain an $O(1/n)$ high-probability rate rather than their $O(1/\sqrt{n})$ rate.*

Remark 6. *We note that by Theorem 2 in [46], both the (normalized) empirical and Mercer eigenvalues ($\frac{1}{n}\lambda_j$ and σ_j respectively) satisfy*

$$\lambda_i/n \leqslant C'\exp(-Ci^{1/d}), \quad \sigma_i \leqslant C'\exp(-Ci^{1/d})$$

for some measure-independent constants $C, C' > 0$, where d is the dimension of $\mathcal{S}$. This justifies our assumption on eigenvalue decay in the statement of Lemma 11. To model more situations where the decay exponent may be better than expected, we use a general spectral decay exponent h rather than d.

D. Detailed Proof for Proposition 3

For notational ease, we define Φ as the concatenation of $\phi(s,a)$ over all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and define the operator P^π as

$$(P^\pi f)(s,a) = \mathbb{E}_{(s',a')\sim P(s,a)\times\pi} f(s',a').$$

We additionally define $\tilde{w}$ which satisfies the condition

$$\tilde{w} = \left(\mathbb{E}_\nu\left[\phi(s,a)\phi(s,a)^\top\right]\right)^{-1} \left(\mathbb{E}_\nu\left[\phi(s,a)\left(r(s,a) + \gamma\mathbb{E}_{(s',a')\sim P(s,a)\times\pi}\left[\phi(s',a')^\top\tilde{w}\right]\right)\right]\right),$$

and let $\tilde{Q}(s,a) = \phi(s,a)^\top\tilde{w}$. It is straightforward to see that $\tilde{w}$ is the fixed point of the population (*i.e.*, $n\to\infty$) projected least square update (19). Furthermore, note that

$$\tilde{w} = \left(\mathbb{E}_\nu\left[\phi(s,a)\left(\phi(s,a) - \gamma\mathbb{E}_{(s',a')\sim P(s,a)\times\pi}\phi(s',a')\right)^\top\right]\right)^{-1} \mathbb{E}_\nu\left[\phi(s,a)r(s,a)\right].$$

With this characterization, letting $\phi_{sup} := \sup_{s,a}\|\phi(s,a)\|$, we have $\|\tilde{w}\| = O\left(\phi_{\sup}\Upsilon_2^{-1}\right)$. We define the operator D_ν as

$$\|f\|_\nu^2 = \langle f, D_\nu f\rangle, \tag{60}$$

and we omit the subscript when ν is the Lebesgue measure, and we also use $\hat{\Pi}_\nu$ and $\hat{P}^\pi$ to define the empirical counterpart of Π_ν and P^π. With the update (19), we have that

$$\Phi\hat{w}_{t+1} = \hat{\Pi}_\nu(r + \gamma\hat{P}^\pi\Phi\hat{w}_t), \quad \Phi\tilde{w} = \Pi_\nu(r + \gamma P^\pi\Phi\tilde{w}),$$

which leads to

$$\Phi(\tilde{w} - \hat{w}_{t+1}) = (\Pi_\nu - \hat{\Pi}_\nu)r + \gamma(\Pi_\nu P^\pi)\Phi(\tilde{w} - \hat{w}_t) + \gamma(\Pi_\nu P^\pi - \hat{\Pi}_\nu\hat{P}^\pi)\Phi\hat{w}_t.$$

With the triangle inequality, we have that

$$\|\Phi(\tilde{w} - \hat{w}_{t+1})\|_\nu \leqslant \gamma\|\Phi(\tilde{w} - \hat{w}_t)\|_\nu + \left\|\left(\Pi_\nu - \hat{\Pi}_\nu\right)r\right\|_\nu + \gamma\left\|\left(\Pi_\nu P^\pi - \hat{\Pi}_\nu\hat{P}^\pi\right)\Phi\hat{w}_t\right\|_\nu,$$

where we use the contractivity under $\|\cdot\|_\nu$. Telescoping over t leads to

$$\|\Phi(\tilde{w} - \hat{w}_T)\|_\nu \leqslant \gamma^T\|\Phi(\tilde{w} - \hat{w}_0)\|_\nu + \frac{1}{1-\gamma}\left\|\left(\Pi_\nu - \hat{\Pi}_\nu\right)r\right\|_\nu + \frac{\gamma}{1-\gamma}\max_{t\in[T]}\left\|\left(\Pi_\nu P^\pi - \hat{\Pi}_\nu\hat{P}^\pi\right)\Phi\hat{w}_t\right\|_\nu.$$

Note that $\|\phi(s,a)\| = O(\phi_{\sup})$, and $\|\tilde{w}\| = O(\phi_{\sup}\Upsilon_2^{-1})$. We follow the proof of Theorem 5.1 from [44] to bound the second term and the third term. In particular, as the subsequent analysis makes clear, the third term dominates the second term, so we focus on bounding the third term. For notational simplicity, we denote $w := \hat{w}_t$ in this section. Note that

$$\left\|\left(\Pi_\nu P^\pi - \hat{\Pi}_\nu\hat{P}^\pi\right)\Phi w\right\|_\nu = \left\|D_\nu^{1/2}\left(\Pi_\nu P^\pi - \hat{\Pi}_\nu\hat{P}^\pi\right)\Phi w\right\|$$
$$\leqslant \left\|D_\nu^{1/2}\Pi_\nu\left(P^\pi - \hat{P}^\pi\right)\Phi w\right\| + \left\|D_\nu^{1/2}\left(\Pi_\nu - \hat{\Pi}_\nu\right)\hat{P}^\pi\Phi w\right\|$$

For the first term, as Π_ν can be written as $\Pi_\nu = \Phi\left(\Phi^\top D_\nu\Phi\right)^{-1}\Phi D_\nu$, we have

$$\left\|D_\nu^{1/2}\Pi_\nu\left(P^\pi - \hat{P}^\pi\right)\Phi w\right\| \leqslant \left\|D_\nu^{1/2}\Phi\right\|\left\|\left(\Phi^\top D_\nu\Phi\right)^{-1}\right\|\left\|\Phi^\top D_\nu\left(P^\pi - \hat{P}^\pi\right)\Phi w\right\| = O\left(\phi_{\sup}\Upsilon_1^{-1}\left\|\Phi^\top D_\nu\left(P^\pi - \hat{P}^\pi\right)\Phi\right\|\|w\|\right).$$

For the second term, we have

$$\left\|D_\nu^{1/2}\left(\Pi_\nu - \hat{\Pi}_\nu\right)\hat{P}^\pi\Phi w\right\| = \left\|D_\nu^{1/2}\Phi\left(\left(\Phi^\top D_\nu\Phi\right)^{-1}\Phi D_\nu - \left(\Phi^\top\hat{D}_\nu\Phi\right)^{-1}\Phi\hat{D}_\nu\right)\hat{P}^\pi\Phi w\right\|$$
$$\leqslant \left\|D_\nu^{1/2}\Phi\left(\Phi^\top D_\nu\Phi\right)^{-1}\Phi\left(D_\nu - \hat{D}_\nu\right)\hat{P}^\pi\Phi w\right\| + \left\|D_\nu^{1/2}\Phi\left(\left(\Phi^\top D_\nu\Phi\right)^{-1} - \left(\Phi^\top\hat{D}_\nu\Phi\right)^{-1}\right)\Phi\hat{D}_\nu\hat{P}^\pi\Phi w\right\|.$$

For the former term, we have that

$$\left\|D_\nu^{1/2}\Phi\left(\Phi^\top D_\nu\Phi\right)^{-1}\Phi\left(D_\nu - \hat{D}_\nu\right)\hat{P}^\pi\Phi w\right\| \leqslant \left\|D_\nu^{1/2}\Phi\right\|\left\|\left(\Phi^\top D_\nu\Phi\right)^{-1}\right\|\left\|\Phi(D_\nu - \hat{D}_\nu)\hat{P}^\pi\Phi w\right\| = O\left(\phi_{\sup}\Upsilon_1^{-1}\left\|\Phi(D_\nu - \hat{D}_\nu)\hat{P}^\pi\Phi\right\|\|w\|\right).$$

For the latter term, note that

$$\left(\Phi^\top D_\nu\Phi\right)^{-1} - \left(\Phi^\top\hat{D}_\nu\Phi\right)^{-1} = \left(\Phi^\top D_\nu\Phi\right)^{-1}\left(\Phi^\top\left(\hat{D}_\nu - D_\nu\right)\Phi\right)\left(\Phi^\top D_\nu\Phi\right)^{-1}.$$

Hence, we have

$$\left\|D_\nu^{1/2}\Phi\left(\left(\Phi^\top D_\nu\Phi\right)^{-1} - \left(\Phi^\top\hat{D}_\nu\Phi\right)^{-1}\right)\Phi\hat{D}_\nu\hat{P}^\pi\Phi w\right\| = O\left(\phi_{\sup}^3\Upsilon_1^{-1}\left\|\Phi^\top\left(\hat{D}_\nu - D_\nu\right)\Phi\right\|\left\|\left(\Phi^\top\hat{D}_\nu\Phi\right)^{-1}\right\|\|w\|\right)$$

Now we bound the remaining terms $\left\|\Phi^\top D_\nu\left(P^\pi - \hat{P}^\pi\right)\Phi\right\|$, $\left\|\Phi(D_\nu - \hat{D}_\nu)\hat{P}^\pi\Phi\right\|$ and $\left\|\Phi^\top\left(\hat{D}_\nu - D_\nu\right)\Phi\right\|$ with standard operator concentration inequalities. We start with the term $\left\|\Phi^\top\left(\hat{D}_\nu - D_\nu\right)\Phi\right\|$. With the matrix Azuma inequality [54], we have that

$$\left\|\Phi^\top\left(\hat{D}_\nu - D_\nu\right)\Phi\right\| = \tilde{O}\left(\phi_{\sup}^2 n^{-1/2}\right).$$

Hence, for sufficient large n, we have that $\left\|\left(\Phi^\top\hat{D}_\nu\Phi\right)^{-1}\right\| = O(\Upsilon_1^{-1})$. Furthermore, as $\hat{P}^\pi$ is a stochastic operator, we also have

$$\left\|\Phi(D_\nu - \hat{D}_\nu)\hat{P}^\pi\Phi\right\| = \tilde{O}\left(\phi_{\sup}n^{-1/2}\right).$$

Now we turn to the term $\left\|\Phi^\top D_\nu\left(P^\pi - \hat{P}^\pi\right)\Phi\right\|$. Note that

$$\left\|\Phi^\top D_\nu\left(P^\pi - \hat{P}^\pi\right)\Phi\right\| \leqslant \left\|\Phi(D_\nu - \hat{D}_\nu)\hat{P}^\pi\Phi\right\| + \left\|\Phi\left(\hat{D}_\nu\hat{P}^\pi - D_\nu P^\pi\right)\Phi\right\|.$$

We already have the bound for the first term. For the second term, we still apply the matrix Azuma's inequality and obtain

$$\left\|\Phi\left(\hat{D}_\nu\hat{P}^\pi - D_\nu P^\pi\right)\Phi\right\| = \tilde{O}\left(\phi_{\sup}^2 n^{-1/2}\right).$$

Thus for large enough n, we have $\|w\| := \|\hat{w}_t\| = \Theta(\|\tilde{w}\|)$. Combining the previous results completes the proof. □

E. Detailed Proof for Lemma 12

Note that $\|\phi(s,a)\|_2 = O(\phi_{\sup})$ and $\|\tilde{w}\|_2 = O(\phi_{\sup}\Upsilon_2^{-1})$. By Remark 6.7 in [38], we know that $\log\pi(a|s)$ is smooth with the smoothness parameter $\beta = O(\phi_{\sup}^2)$. As a result, we

have that

$$\log \frac{\pi_{k+1}(a|s)}{\pi_k(a|s)} \geqslant \eta \left(\nabla_\theta \log \pi_k(a|s)\right)^\top \hat{w}_{k,T} - \frac{\eta^2\beta}{2}\left\|\hat{w}_{k,T}\right\|_2^2.$$

With this inequality, we have

$$\begin{aligned}
&\mathbb{E}_{s\sim d^{\tilde{\pi}}}\left[\mathrm{KL}(\tilde{\pi}(\cdot|s)\|\pi_k(\cdot|s)) - \mathrm{KL}(\tilde{\pi}(\cdot|s)\|\pi_{k+1}(\cdot|s))\right]\\
=&\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[\log\frac{\pi_{k+1}(a|s)}{\pi_k(a|s)}\right]\\
\geqslant&\eta\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[\left(\nabla_\theta\log\pi_k(a|s)\right)^\top\hat{w}_{k,T}\right] - \frac{\eta^2\beta}{2}\left\|\hat{w}_{k,T}\right\|_2^2\\
=&\eta\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[A^{\pi_k}(s,a)\right] - \frac{\eta^2\beta}{2}\left\|\hat{w}_{k,T}\right\|_2^2\\
&+\eta\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[\left(\nabla_\theta\log\pi_k(a|s)\right)^\top\hat{w}_{k,T} - A^{\pi_k}(s,a)\right]\\
=&(1-\gamma)\eta\left(V^{\tilde{\pi}} - V^{\pi_k}\right) - \frac{\eta^2\beta}{2}\left\|\hat{w}_{k,T}\right\|_2^2 - \eta\mathrm{err}_k,
\end{aligned}$$

where in the last step we use the performance difference lemma (Lemma 3.2 in [38]). Telescoping over k gives

$$\begin{aligned}
\sum_{k=0}^{K-1}\left\{V^{\tilde{\pi}} - V^{\pi_k}\right\} \leqslant& \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\left[\mathrm{KL}(\tilde{\pi}(\cdot|s)\|\pi_0(\cdot|s))\right]}{1-\gamma}\\
&+\frac{\eta^2\beta}{2}\sum_{k=0}^{K-1}\left\|\hat{w}_{k,T}\right\|_2^2 + \frac{1}{1-\gamma}\sum_{k=0}^{K-1}\mathrm{err}_k.
\end{aligned}$$

Note that $\mathrm{KL}(\tilde{\pi}(\cdot|s)\|\pi_0(\cdot|s)) \leqslant \log|\mathcal{A}|$ and $\|\hat{w}_{k,T}\| = O\left(\|\tilde{w}\|\right) = O(\phi_{\sup}\Upsilon_2^{-1})$, we finish the proof. □

F. Detailed Proof for Theorem 2

With Lemma 12 and our choice of η, we only need to bound the term err_k. We observe that

$$\mathrm{err}_k = \mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[A^{\pi_k}(s,a) - \hat{w}_{k,T}{}^\top\left(\phi(s,a) - \mathbb{E}_{a'\sim\pi_k(s)}\left[\phi(s,a')\right]\right)\right]$$

Note that

$$\begin{aligned}
&\mathbb{E}_{d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[A^{\pi_k}(s,a) - \hat{w}_{k,T}{}^\top\left(\phi(s,a) - \mathbb{E}_{a'\sim\pi_k(s)}\left[\phi(s,a')\right]\right)\right]\\
=&\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[Q^{\pi_k}(s,a) - \hat{Q}_T{}^{\pi_k}(s,a)\right]\\
&-\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\pi_k(\cdot|s)}\left[Q^{\pi_k}(s,a) - \hat{Q}_T{}^{\pi_k}(s,a)\right]\\
\leqslant&\sqrt{\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\tilde{\pi}(\cdot|s)}\left[\left(Q^{\pi_k}(s,a) - \hat{Q}_T{}^{\pi_k}(s,a)\right)^2\right]}\\
&+\sqrt{\mathbb{E}_{s\sim d^{\tilde{\pi}},a\sim\pi_k(\cdot|s)}\left[\left(Q^{\pi_k}(s,a) - \hat{Q}_T{}^{\pi_k}(s,a)\right)^2\right]}.
\end{aligned}$$

Plugging in π^* as $\tilde{\pi}$, combining the results in the previous subsection, we have concluded the proof. □

G. Bounding the norms of $\psi_{\mathsf{rf}}(\mathsf{s},\mathsf{a})$ and $\psi_{\mathsf{nys}}(\mathsf{s},\mathsf{a})$

We have the following lemma which bounds the norm of $\|\psi_{rf}(s,a)\|$ and $\|\psi_{nys}(s,a)\|$.

Lemma 16. *The random features and Nystrom features constructed in Algorithm 2 satisfy the bounds*

$$\|\psi_{rf}(s,a)\| \leqslant \sqrt{m}\tilde{g}_\alpha$$

$$\|\psi_{nys}(s,a)\| \leqslant \sqrt{\frac{n_{nys}\tilde{g}_\alpha^2}{\lambda_m}}$$

respectively, where λ_m denotes the m-th largest eigenvalue of $K^{(n_{nys})}$.

Proof. Recall that

$$\psi_{rf}(s,a) := \{\frac{g_\alpha(f(s,a))}{\alpha^d}\cos(\omega_i^\top f(s,a) + b_i)\}_{i=1}^m$$

$$\psi_{nys}(s,a) := \{\frac{g_\alpha(f(s,a))}{\alpha^d\sqrt{\lambda_i}}\sum_{\ell=1}^{n_{nys}} U_{\ell,i}k_\alpha\left(x_\ell, \frac{f(s,a)}{1-\alpha^2}\right)\}_{i=1}^m,$$

where we recall that $U\in\mathbb{R}^{n_{nys}\times n_{nys}}$ is the U that appears in the eigendecomposition of the empirical Nystrom Gram matrix $K^{(nys)} = U\Lambda U^T$.

For random features, it is not hard to see that $\|\psi_{rf}(s,a)\| = O(\sqrt{m}\tilde{g}_\alpha)$, since each random feature takes the form of a randomly shifted cosine $\frac{g_\alpha(f(s,a))}{\alpha^d}$, which in our paper is upper bounded by the expression $\tilde{g}_\alpha$.

For Nystrom features, we note that since $0\leqslant k_\alpha(\cdot,\cdot)\leqslant 1$ and U is an orthonormal matrix, we see that

$$\begin{aligned}
&\|\psi_{nys}(s,a)\|^2\\
\leqslant&\tilde{g}_\alpha^2\left(\begin{bmatrix}k_\alpha(x_1,f(s,a)) & \dots & k_\alpha(x_{n_{nys}},f(s,a))\end{bmatrix}^\top U_{:,:m}\Lambda_m^{-1}U_{:,:m}^\top\begin{bmatrix}k_\alpha(x_1,f(s,a))\\ \dots\\ k_\alpha(x_{n_{nys}},f(s,a))\end{bmatrix}\right)\\
\leqslant&\tilde{g}_\alpha^2\cdot\frac{1}{\lambda_m}\left(\begin{bmatrix}k_\alpha(x_1,f(s,a)) & \dots & k_\alpha(x_{n_{nys}},f(s,a))\end{bmatrix}^\top U_{:,:m}U_{:,:m}^\top\begin{bmatrix}k_\alpha(x_1,f(s,a))\\ \dots\\ k_\alpha(x_{n_{nys}},f(s,a))\end{bmatrix}\right)\\
\overset{(a)}{\leqslant}&\frac{1}{\lambda_m}\tilde{g}_\alpha^2\left(\left\|\begin{bmatrix}k_\alpha(x_1,f(s,a))\\ \dots\\ k_\alpha(x_{n_{nys}},f(s,a))\end{bmatrix}\right\|^2\right)\\
\leqslant&n_{nys}\frac{1}{\lambda_m}\tilde{g}_\alpha^2
\end{aligned}$$

where we note that $U_{:,:m}$ denotes the first m columns of U, Λ_m denoted the top m by m block of Λ,and we used the fact that the columns of U are independent to derive (a). Thus, for the Nystrom features, we have the bound $\|\psi_{nys}(s,a)\|\leqslant\sqrt{\frac{n_{nys}\tilde{g}_\alpha^2}{\lambda_m}}$. □

H. Proofs on Lemmas 7, 8,9

Lemma 17 (Restatement of Lemma 7)**.** *The random features $\psi_{rf}(s,a)$ are linearly independent almost surely, i.e. the probability of randomly drawing $\{\omega_i,b_i\}_{i=1}^m\sim N(0,\frac{1}{\sigma^2}I_d)\times U([0,2\pi])$ is 0.*

Proof. Without loss of generality, we consider the case that $\alpha = 0$, such that the random features take the form $\psi_{rf}(s,a) = \{\cos(\omega_i^\top f(s,a) + b_i)\}_{i=1}^m$, where $\omega_i\sim N(0,\sigma^{-2}I_d)$ and $b_i\sim\mathrm{Unif}([0,2\pi])$, i.e. the random features are randomly shifted cosines; we note that the general case when $\alpha>0$ can be easily reduced to this case since the factor $g_\alpha(f(s,a))/\alpha^d$ shared by all the features is strictly positive. Suppose the random features are not linearly independent over the space $f(\mathcal{S}\times\mathcal{A})$. Then there exist scalars $\{a_i\}_{i=1}^m$, not all of which are 0, such that

$$S(x) := \sum_{i=1}^m a_i\cos(\omega_i^\top x + b_i) = 0,\quad \forall x\in f(\mathcal{S}\times\mathcal{A}).$$

Pick some x^0 in the interior of $f(\mathcal{S}\times\mathcal{A})$, and denote $y_i := \cos(\omega_i^\top x^0 + b_i)$; the fact that x^0 lies in the interior of $f(\mathcal{S}\times\mathcal{A})$ means that all derivatives of $S(x)$ vanish at $x = x^0$. Now, for any natural number $n\leqslant m-1$, differentiate $S(x)$ with respect to the first coordinate x_1 at the point x^0, giving us

$$\sum_{i=1}^m a_i(\omega_i)_1^{2n}y_i = 0.$$

This then yields the following relation

$$Va = 0,\quad \text{where } V := \begin{pmatrix}1 & \dots & 1\\ (\omega_1)_1^2 & \dots & (\omega_m)_1^2\\ \vdots & & \vdots\\ ((\omega_1)_1^2)^{m-1} & \dots & ((\omega_m)_1^2)^{m-1}\end{pmatrix},\ a := \begin{pmatrix}a_1\\ \vdots\\ a_m\end{pmatrix}.$$

We note that since $V\in\mathbb{R}^{m\times m}$ is a Vandermonde matrix, its determinant is non-zero so long as the $(\omega_i)_1$'s are all distinct, which happens with probability 1 if the ω_i's are drawn iid

from a Gaussian distribution. This implies then that the scalars $\{a_i\}_{i=1}^m$ must all be 0, which is a contradiction. Thus, the random features must be linearly independent over the space $f(\mathcal{S} \times \mathcal{A})$. □

Lemma 18 (Restatement of Lemma 8)**.** *Consider any feature dimension $0 < m \leqslant n_{Nys}$. Suppose the Nystrom Gram matrix $K^{(n_{Nys})}$ has rank at least m. Then, the Nystrom features $\psi_{nys}(s,a)$ are linearly independent.*

Proof. Without loss of generality, for simplicity we consider the case when $\alpha = 0$, in which case the Nystrom features are constructed as follows. First, let m denote the feature dimension, and let $n_{nys} \geqslant m$ denote the number of random samples $\{x_1, \ldots, x_{n_{nys}}\}$ drawn from $\mathcal{S}$ following the distribution μ_{Nys}. Recall that the (n_{nys} by n_{nys}) Gram matrix K is constructed as follows: $K_{i,j} = k(x_i, x_j) := \exp\left(\frac{-\|x_i, x_j\|^2}{2\sigma^2}\right)$. Compute the eigendecomposition $KU = \Lambda U$, where $UU^\top = U^\top U = I$, and Λ denoting a diagonal matrix with eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_{n_{nys}} \geqslant 0$. Suppose that m-th largest eigenvalue $\lambda_m > 0$. Then, for any (s,a) pair,the m Nystrom feature functions are given as

$$(\psi_{nys})_i(s,a) := \frac{1}{\sqrt{\lambda_i}} \sum_{\ell=1}^{n_{nys}} U_{\ell,i} k(x_\ell, f(s,a)), \qquad i \in [m].$$

Note that by construction, the m eigenvectors $U_{:,\ell}$ are linearly independent (so long as $\lambda_m > 0$). Suppose that the m Nystrom feature functions are not linearly independent. That implies then that there exists m scalars $\{a_i\}_{i=1}^m \subset \mathbb{R}$ (at least some of which are nonzero) such that for any (s,a),

$$\begin{aligned}
&\sum_{i=1}^m a_i (\psi_{nys})_i(s,a) = 0 \\
\iff &\sum_{i=1}^m a_i \left(\frac{1}{\sqrt{\lambda_i}} \sum_{\ell=1}^{n_{nys}} U_{\ell,i} k(x_\ell, f(s,a)) \right) = 0 \\
\iff &\sum_{\ell=1}^{n_{nys}} k(x_\ell, f(s,a)) \left(\sum_{i=1}^m a_i \left(\frac{1}{\sqrt{\lambda_i}} \sum_{\ell=1}^{n_{nys}} U_{\ell,i} \right) \right) = 0.
\end{aligned}$$

Note that since $k(\cdot,\cdot) > 0$, the above derivation implies that we must have for every $\ell \in [n_{nys}]$:

$$\begin{aligned}
&\sum_{i=1}^m a_i \left(\frac{1}{\sqrt{\lambda_i}} \sum_{\ell=1}^{n_{nys}} U_{\ell,i} \right) = 0 \\
\implies &U_{:,:m} \Lambda_m^{-1/2} a = 0,
\end{aligned}$$

where $U_{:,:m} \in \mathbb{R}^{n_{nys} \times m}$ denotes the first m columns of U, Λ_m denotes the top m by m block of Λ, and $a := \mathrm{vec}(a_1, \ldots, a_m) \in \mathbb{R}^m$. Since the columns in U are linearly independent by design, and $\Lambda_m \succ 0$ (by our assumption that $\lambda_m > 0$), it follows that the vector $a \in \mathbb{R}^m$ has to be 0, which contradicts the assumption that the m Nystrom features are linearly dependent. Thus, whenever $\lambda_m > 0$, it follows that the Nystrom features are linearly independent.

□

Lemma 19 (Restatement of Lemma 9)**.** *Suppose the features $\phi(s,a)$ are linearly independent over the interior of the set $\mathrm{supp}(\nu^\pi)$ for any policy π. Then, for all but finitely many $0 \leqslant \lambda < 1$, Assumption 3 holds.*

Proof. Fix some policy π. Suppose that the matrix $\mathbb{E}_{\nu^\pi}\left[\psi(s,a)\psi(s,a)^\top\right]$ (which must be positive-semidefinite) is not positive definite. Then, there exists some $0 \neq v \in \mathbb{R}^m$ such that

$$v^\top \mathbb{E}_{\nu^\pi}\left[\psi(s,a)\psi(s,a)^\top\right] v = 0.$$

In particular, this implies that $v^\top \psi(s,a) = 0$ for any $(s,a) \in \mathrm{int}(\mathrm{supp}(\nu^\pi))$. However, by linear independence of the features $\{\psi_i(s,a)\}_{i=1}^m$ over $\mathrm{int}(\mathrm{supp}(\nu^\pi))$, it follows that v must be 0. This forms a contradiction, and thus the matrix $\mathbb{E}_{\nu^\pi}\left[\psi(s,a)\psi(s,a)^\top\right]$ must be positive definite.

Next, we show that

$$\sigma_{\min}\left(\mathbb{E}_{\nu^\pi}\left[\psi(s,a)\left(\psi(s,a) - \gamma \mathbb{E}_{\nu^\pi}[\psi(s',a')]\right)^\top\right]\right) > 0$$

for all but finitely many $0 \leqslant \lambda < 1$. Denote the matrix $M(\lambda) := \mathbb{E}_{\nu^\pi}\left[\psi(s,a)\left(\psi(s,a) - \gamma \mathbb{E}_{\nu^\pi}[\psi(s',a')]\right)^\top\right]$. Observe that the determinant of $M(\lambda)$ is a polynomial in λ with degree at most m. Since $M(0) \neq 0$ (as we just showed that the matrix $\mathbb{E}_{\nu^\pi}\left[\psi(s,a)\psi(s,a)^\top\right]$ is positive definite), it follows that $\det(M(\lambda)$ is not the zero polynomial. Thus, it can have at most m roots, which shows that for all but finitely many λ, $M(\lambda)$ must be invertible, i.e. its smallest singular value is bounded away from 0. □

I. Necessity of local Rademacher complexity

A natural approach to bounding the approximation error without using local Rademacher complexity is as follows. For any function $f \in L_2(\mu)$, let $\mu_{X^n}(f) = \frac{1}{n}\sum_{i=1}^n f(x_i)$, where each x_i denotes the i-th sample from X^n, drawn iid according to the distribution μ. The key is that by construction of the Nystrom features,

$$\begin{aligned}
\mu_{X^n}(\left\|P_{m,\perp}^{X^n} k(x,\cdot)\right\|) &= \frac{1}{n}\sum_{i=1}^n \sqrt{(k - \hat{k}_m^n)(x_i,x_i)} \\
&\leqslant \sqrt{\frac{1}{n}\sum_{i=1}^n (k - \hat{k}_m^n)(x_i,x_i)} \\
&= \sqrt{\frac{1}{n}\sum_{i>m} \lambda_i},
\end{aligned}$$

where the λ_i's denote the eigenvalues of the empirical Gram matrix $K \in \mathbb{R}^{n \times n}$, where $K_{i,j} = k(x_i, x_j)$. By appropriate assumptions on the decay speed of the empirical eigenvalues, this term can be shown to be of the order $\tilde{O}(1/n)$ for m sufficiently large. Hence, if the true approximation error $\mu(\left\|P_{m,\perp}^{X^n} k(x,\cdot)\right\|)$ concentrates around the empirical approximation error $\mu_{X^n}(\left\|P_{m,\perp}^{X^n} k(x,\cdot)\right\|)$ also at a rate of $O(1/n)$, we will get an overall convergence speed of $\tilde{O}(1/n)$.

However, assuming for simplicity that $\left\|P_{m,\perp}^{X^n} k(x,\cdot)\right\|$ belongs to a function class $\mathcal{F}$, the catch is that standard Rademacher complexity arguments [53] can only show that

$$\begin{aligned}
&\mu(\left\|P_{m,\perp}^{X^n} k(x,\cdot)\right\|) - \mu_{X^n}(\left\|P_{m,\perp}^{X^n} k(x,\cdot)\right\|) \\
&\leqslant O(Rad(\mathcal{F})) + O(\sqrt{\frac{\log(1/\delta)}{n}}),
\end{aligned}$$

with probability at least $1 - O(\delta)$. Note that there is a term $O(1/\sqrt{n})$ is strictly worse than the $O(1/n)$ rate we hope to

achieve.

However, by leveraging local Rademacher complexity results, it is possible to show for any function class $\mathcal{F}$ with appropriately bounded covering number (see Lemma 11 in our original submission for the detailed conditions), for any $f \in \mathcal{F}$, with probability at least $1-\delta$, we can show that

$$\mu(f) \leqslant O(\mu_{X^n}(f)) + \tilde{O}(\frac{\log(1/\delta)}{n}),$$

where the $\tilde{O}$ hides dependencies on terms related to the size of the covering number of $\mathcal{F}$. Note that this improves the scaling with n from $1/\sqrt{n}$ in the generic bound to $1/n$.

If $\left\|P^{X^n}_{m,\perp} k(x,\cdot)\right\|$ belonged to a function class $\mathcal{F}$ that satisfied appropriate bounded conditions on its covering number, then this argument above based on local Rademacher complexity sufficies for us to improve the approximation error bound to depend on $1/n$ rather than on $1/\sqrt{n}$ as in the generic bound. However, $\left\|P^{X^n}_{m,\perp} k(x,\cdot)\right\|$ may not belong to a function class $\mathcal{F}$ that satisfies the regularity conditions in Lemma 11, so we cannot directly apply this result. Nonetheless, via a careful analysis, we are able to eventually rearrange the error bound such that it can be written as

$$\begin{aligned}
&\mu(\left\|P^{X^n}_{m,\perp} k(x,\cdot)\right\|) \\
&\leqslant \mu\left(\left\|P^{X_n}_{m,\perp} \Pi_r k_\alpha(x,\cdot)\right\|_{\mathcal{H}_{k_\alpha}}\right) - \\
&\qquad 2\mu_{X_n}\left(\left\|P^{X_n}_{m,\perp} \Pi_r k_\alpha(x,\cdot)\right\|_{\mathcal{H}_{k_\alpha}}\right) + \text{remainder terms},
\end{aligned}$$

where remainder terms can be shown to scale as the decay rate of either the eigenvalues of the Mercer expansion or Gram matrix plus a $O(1/n)$ term that comes from application of a Bernstein inequality, leaving a key term $\mu\left(\left\|P^{X_n}_{m,\perp} \Pi_r k_\alpha(x,\cdot)\right\|_{\mathcal{H}_{k_\alpha}}\right) - 2\mu_{X_n}\left(\left\|P^{X_n}_{m,\perp} \Pi_r k_\alpha(x,\cdot)\right\|_{\mathcal{H}_{k_\alpha}}\right)$. Using the fact that $\left\|P^{X_n}_{m,\perp} \Pi_r k_\alpha(x,\cdot)\right\|$ can be shown to be part of a function class with bounded covering number, we can then apply the local Rademacher complexity result in Lemma 11 to bound this term such that it is also of the form $\tilde{O}(1/n)$, leaving an overall rate of $\tilde{O}(1/n)$.