Submitted to the Annals of Statistics

OPTIMAL POLICY EVALUATION USING KERNEL-BASED TEMPORAL DIFFERENCE METHODS

BY YAQI DUAN[1,a], MENGDI WANG[2,b] AND MARTIN J. WAINWRIGHT[3,c]

[1]*Leonard N. Stern School of Business, New York University,* [a]*yaqi.duan@stern.nyu.edu*

[2]*Department of ECE, Princeton University,* [b]*mengdiw@princeton.edu*

[3]*Departments of EECS and Mathematics, Massachusetts Institute of Technology,* [c]*wainwrigwork@gmail.com*

We study non-parametric methods for estimating the value function of an infinite-horizon discounted Markov reward process (MRP). We analyze the kernel-based least-squares temporal difference (LSTD) estimate, which can be understood either as a non-parametric instrumental variables method, or as a projected approximation to the Bellman fixed point equation. Our analysis imposes no assumptions on the transition operator of the Markov chain, but rather only conditions on the reward function and population-level kernel LSTD solutions. Using empirical process theory and concentration inequalities, we establish a non-asymptotic upper bound on the error with explicit dependence on the effective horizon $H = (1-\gamma)^{-1}$ of the Markov reward process, the eigenvalues of the associated kernel operator, as well as the instance-dependent variance of the Bellman residual error. In addition, we prove minimax lower bounds over sub-classes of MRPs, which shows that our guarantees are optimal in terms of the sample size n and the effective horizon H. Whereas existing worst-case theory predicts cubic scaling (H^3) in the effective horizon, our theory reveals a much wider range of scalings, depending on the kernel, the stationary distribution, and the variance of the Bellman residual error. Notably, it is only parametric and near-parametric problems that can ever achieve the worst-case cubic scaling.

1. Introduction. Markov decision processes provide a classical framework for modeling how to make optimal decisions in a sequential setting. They have been studied extensively in statistics and operations research [3, 8, 9], control theory [5, 40], and computer science [46]. Moreover, Markov decision processes have proven useful in a wide variety of applications (e.g., [10, 46]), including inventory management, allocation of fire-fighting resources, competitive game playing, and industrial process control. In much of the classical work, the model structure and parameters were assumed to be known, and principles of dynamic programming were used to characterize and compute optimal decision rules, also known as policies. By contrast, reinforcement learning (RL) refers to a class of statistical procedures suitable for settings in which the model structure and/or parameters are unknown. In this context, a central problem is to use samples to evaluate the quality of a given policy, as assessed via its value function. Indeed, the estimation of value functions serves as a fundamental building block for many RL algorithms [7, 46].

When a given policy is fixed, a Markov decision process reduces to a Markov reward process (MRP). The value of any given initial state in an MRP corresponds to the expected cumulative reward along a trajectory when starting from the given state; the collection of all such state values defines the value function. The problem of estimating this function is known

MSC2020 subject classifications: Primary 62G05; secondary 62M05.

Keywords and phrases: sequential decision-making, dynamic programming, reinforcement learning, Markov reward process, non-parametric estimation, policy evaluation, temporal difference learning, reproducing kernel Hilbert space.

as *policy evaluation*, or *value function estimation*, and we use these terms interchangeably. In practice, policy evaluation is challenging because the state space might be continuous, or even when discrete, it might involve a prohibitively large number of possible states. For this reason, practical methods for policy evaluation typically involve some form of function approximation.

The simplest and most well-studied approach is based on linear function approximation, in which the value function is approximated as a weighted combination of a fixed set of features. This particular choice leads to the least-squares policy evaluation estimator, also known as the least-squares temporal difference (LSTD) estimate, along with its online temporal difference variants (e.g., [11, 31, 46, 48]). The choice of linear functions is attractive in that the LSTD estimate is easy to compute, based on solving a linear system of equations. However, the expressivity of linear functions is limited, and so that it is natural to seek approximations in richer function classes.

In many types of statistical problems—among them regression, density estimation, dimension reduction, and clustering—methods based on reproducing kernel Hilbert spaces (RKHSs) have proven useful [4, 21, 42, 50]. As we discuss in Section 1.1, kernel methods have also proven useful in the specific context of reinforcement learning. Kernel methods allow for much richer representations of functions, by working—in an implicit way—over a possibly infinite set of features, as defined by the eigenfunctions of the associated kernel integral operator. However, at the same time, due to the classical representer theorem [27], a broad class of kernel-based estimators can be computed relatively easily by working directly with n-dimensional kernel matrices, where n is the sample size.

The main contribution of this paper is to provide a sharp and non-asymptotic characterization of the statistical properties of a family of kernel-based procedures for policy evaluation. So as to bring our specific contributions into sharp focus, we study the case of infinite-horizon γ-discounted Markov reward processes (MRPs), but much of our analysis and associated techniques also have consequences for kernel methods in the finite-horizon setting. In our analysis, we assume that we have access to the reward function and i.i.d. transition pairs drawn from the stationary distribution. We analyze a kernel-based temporal difference estimator, whose population limit corresponds to the fixed point of a projected Bellman operator. We measure the difference between the empirical and population estimators in $L^2(\mu)$ norm, with μ denoting the stationary distribution. We refer to this $L^2(\mu)$ error as the *estimation error*. At a high level, the main contribution of this paper is to provide a sharp and partially instance-dependent analysis of this estimation error.

1.1. *Related work and our contributions.*

We begin by discussing related work and then, with this context in place, provide a high-level overview of our contributions.

Related work. Here we provide a partial overview of past work, with an emphasis on those estimation-theoretic papers most relevant for putting our results in context. The utility of kernel methods in reinforcement learning is by now well-established, as attested to by the lengthy line of previous papers on the topic (e.g., [1, 2, 13, 17, 18, 20, 28, 47]). In the special case of a linear kernel function, the kernel-based method studied in this paper reduces to the classical least-squares temporal difference (LSTD) method [11, 45, 46].

In terms of papers that provide guarantees on statistical estimation error for policy evaluation using non-parametric methods, early work by Ormoneit and Sen [37] studied the use of local-averaging kernel methods for approximating value functions; they proved various types of asymptotic consistency results. Munos and Szepesvari [33] studied methods for fitted value iteration (FVI) under various types of ℓ_p-norms; under metric entropy conditions on the function space, they proved various types of consistency results, but without providing sharp or minimax-optimal guarantees. In later work, Farahmand et al. [16] studied a class

of regularized procedures for both policy evaluation and policy optimization. Their analysis is attractive in allowing for quite general function classes, with reproducing kernel Hilbert spaces being an important special case. They provided guarantees under bounds on the sup-norm metric entropy of the function classes at hand, and for certain function classes, they argued that their bounds achieved the optimal scaling in sample size n. Farahmand et al. also conjectured that it should be possible to prove similar guarantees using metric entropy conditions in the μ-norm, and indeed, in the special case of RKHS classes, one consequence of our results is to confirm this conjecture. A more recent line of work has studied variants of fitted Q-iteration (FQI) using neural network approximation, and provided statistical guarantees under different notions of smoothness. For example, Fan et al. [15] exploited the Hölder smoothness of the range of Bellman operator to derive bounds on estimation error; Nguyen-Tang et al. [35] approximated deep ReLU networks using Besov classes; and Long et al. [29] analyzed two-layer neural networks based on neural tangent kernels or Barron spaces. All these works contribute to the understanding of empirical success of deep reinforcement learning.

A notable feature of much past work is while it provides bounds on statistical error, it does not carefully track the dependence on the (effective) horizon and model dynamics, as well as the variance of the Bellman residual. As we argue in this paper, understanding how non-parametric procedures depend on the latter quantities is essential—in particular, they are the ingredients that actually distinguish the dynamic problem of value function estimation from a typical (static) prediction problem, with ordinary non-parametric regression being the archetypal example. In order to reveal this dependence, the analysis of this paper makes use of empirical process techniques [49, 50] that have proven successful for analyzing kernel ridge regression and related estimators (e.g., [41, 54, 59]). Essential for obtaining sharp rates is the local Rademacher complexity, which has an explicit expression in terms of the eigenvalues of the kernel integral operator [30]; see Chapters 12 and 13 in the book [50] for more details.

It is also worth noting that recent years have witnessed considerable progress in understanding policy evaluation in off-policy settings, and/or providing guarantees that have optimal instance-dependent rates. This work can be separated into work that is either asymptotic [23–25] and non-asymptotic [26, 38, 53, 55] in nature. In this non-asymptotic setting, much of this work is focused on either the tabular case, or the simpler setting of linear function approximation, as opposed to the non-parametric cases of interest here. We note that our results do depend on the problem instance, but this instance-dependence is not (yet) as sharp as that established in the simpler setting of tabular problems; see the papers [26, 38] for sharp results of this type.

This paper also makes connections to the large body of work on instrumental variable (IV) methods (e.g., [12, 34, 51, 52]). It is known that the least-squares temporal difference (LSTD) estimate can be viewed as a classical linear IV estimate [11]. More generally, the kernel-based procedures in this paper correspond to a non-parametric form of an instrumental variable method. While portions of our analysis are specific to reinforcement learning, we suspect that our techniques can be adapted so as to provide non-asymptotic and instance-dependent guarantees for other non-parametric IV estimates.

Our contributions. Consistency of any statistical estimator is certainly a desirable requirement. A more ambitious goal, and a centerpiece of high-dimensional statistics, is to give a more refined non-asymptotic characterization, one which tracks not only sample size but also other structural properties of the problem. In the context of policy evaluation for Markov reward processes with discount factor $\gamma \in (0, 1)$, such structural properties include: (a) the complexity of the population-level value function θ^* being estimated; (b) the "richness" of the function class used for approximation relative to the stationary measure of the Markov chain; (c) the effective horizon $H := (1 - \gamma)^{-1}$, which measures the typical scale over which

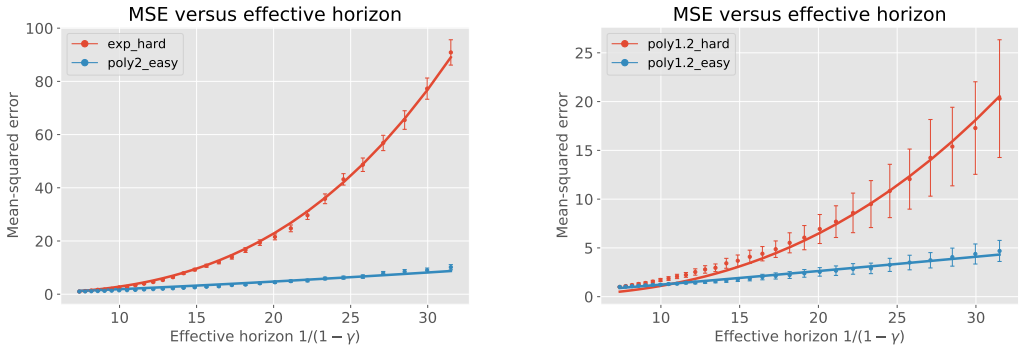
the discounted reward process evolves; and (d) the underlying noise function, given by the variance of the Bellman residual. The latter two properties are of particular interest, since they distinguish the dynamic nature of value function estimation from a standard problem of static non-parametric estimation.

The main contribution of this paper is to give a precise characterization, including both matching upper and lower bounds, on how a well-tuned version of the kernel-based LSTD estimate depends on all of these structural parameters. Notably, our characterization is instance-dependent, in that the bounds vary considerably depending on the structure of θ^* and the associated variance of the Bellman residual error, along with the eigenvalues of the kernel integral operator, which vary as a function of both the kernel function class, and the stationary measure of the Markov chain. En route to doing so, we provide specific guidance on how the regularization parameter, essential for non-parametric methods such as those based on RKHSs, should be chosen.

Theorem 3.1 provides two types of non-asymptotic bounds on the estimation error of a regularized kernel LSTD estimate: a "slow" rate and a "fast" rate. These two guarantees differ in the way that the inherent noise of the problem is measured. While the "slow" guarantee holds for any sample size n, the guarantee is based on a crude measure of the noise level, based on bounds on the sup-norm and Hilbert norm of the population-level value function. The second "fast" guarantee holds only once the sample size exceeds a certain threshold; at the same time, it depends on the variance of the Bellman residual error, which is a fundamental quantity for the problem. Indeed, in our second main result, stated as Theorem 3.4, we study the best performance of any procedure of two particular sub-classes of MRPs, and prove lower bounds that match the "fast" rates from Theorem 3.1 in terms of all relevant problem-dependent quantities. These matching upper and lower bounds establish the optimality of our procedure.

Our theory applies to a fairly general class of kernel functions in arbitrary dimension, with the rates specified in terms of the eigenvalues $\{\mu_j\}_{j=1}^{\infty}$ of the induced kernel operator. It is important to note that these eigenvalues depend not just on the kernel, but also on the stationary distribution of the Markov chain. One special case, of interest in its own right, are kernels and stationary distributions for which these eigenvalues decay at a polynomial rate, say $\mu_j \asymp (1/j)^{2\alpha}$ for some $\alpha > 1/2$. Figure 1 highlights some interesting predictions made by our theory regarding how the optimal $L^2(\mu)$-error should scale with the effective horizon $H = (1-\gamma)^{-1}$. As discussed in more detail in Section 3.3, we construct a "hard" ensemble of MRPs for which our theory—both upper and lower bounds—guarantees that for a fixed sample size, the squared $L^2(\mu)$-error should grow as $H^{\frac{2(3\alpha+1)}{2\alpha+1}}$. In the limit as $\alpha \to +\infty$, the kernel class becomes a parametric function class, and the horizon dependence becomes the familiar cubic one H^3 first elucidated by Azar et al. [19]. However, for genuinely non-parametric classes where α is relatively small, the dependence on the effective horizon is much milder—e.g., it scales as $H^{8/3}$ for a kernel with $\alpha = 1$. This reveals the interesting phenomenon that non-parametric forms of value estimation exhibit milder horizon dependence. Moreover, since our theory is instance-dependent via the variance of Bellman residual, we can show that global minimax predictions are often conservative. In particular, we also construct an "easy" ensemble for which the scaling in horizon is much milder, given by $H^{\frac{4\alpha}{2\alpha+1}}$. Again, as shown in Figure 1, these theoretical predictions capture the scaling of the error in the effective horizon with high accuracy.

1.2. *Paper organization and notation.* The remainder of the paper is structured as follows. We begin in Section 2 by introducing background on Markov reward processes and policy estimation, along with reproducing kernel Hilbert spaces and the kernel LSTD estimate analyzed in this paper. Section 3 is devoted to the statement of our main results, along with discussion of some of their consequences.



(a) (b)

Fig 1. Plots of the mean-squared error $\mathbb{E}\|\widehat{\theta} - \theta^*\|_\mu^2$ versus the effective horizon $H = \frac{1}{1-\gamma}$ for different ensembles of problems. For each point (on each curve in each plot), the MSE was approximated by taking a Monte Carlo average over $T = 1000$ trials; sample standard deviations are shown as error bars. In each case, our theory predicts that the MSE should grow as a function of the form H^η, where the exponent $\eta > 0$ is determined by α and the ensemble type: we have $\eta = \frac{2(3\alpha+1)}{2\alpha+1}$ for the hard ensemble, and $\eta = \frac{4\alpha}{2\alpha+1}$ for the easy ensemble. In each panel, solid curves correspond to our theoretical predictions for a sample size $n = 4000$ and a range of horizons H. (a) Plots comparing exponential decay kernel $\mathcal{K}_3$ under the "hard" ensemble to the 1-polynomial decay kernel $\mathcal{K}_2$ under the "easy" ensemble. Theory predicts that the MSE scales as H^3 and $H^{1.33}$ in these two cases respectively; as shown, these theoretical predictions agree well with the empirical results. (b) Plots comparing the behavior of the kernel $\mathcal{K}_1$ (with 0.6-polynomial decay) under the "hard" ensemble versus the "easy" ensemble. Theory predicts that the MSE should scale as $H^{2.55}$ and $H^{1.09}$ in these two cases.

Theorem 3.1 provides two finite-sample upper bounds and ranges of regularization to achieve them. Theorem 3.4 establishes matching minimax lower bounds over two MRP subclasses. Section 3.3 provides the results of numerical experiments with synthetic data that illustrate various qualitative features of our theoretical predictions. Section 4 contains the proofs of Theorems 3.1 and 3.4, and we conclude with a discussion in Section 5.

Notation. For any event $\mathcal{E}$, we use $\mathbb{1}\{\mathcal{E}\}$ to denote the $\{0,1\}$-valued indicator function. We use C, c, c_0 etc. to denote universal constants whose numerical values may vary from line to line. For a positive integer D, we adopt the shorthand $[D] := \{1, 2, \ldots, D\}$. Given a distribution μ, we define the $L^2(\mu)$-norm $\|f\|_\mu := \sqrt{\int f^2 \mu(dx)}$ for $f \in L^2(\mu)$. We also define the supremum norm $\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|$. For two measures p, q with p absolutely continuous with respect to q, we define the Kullback–Leibler (KL) divergence $D_{\mathrm{KL}}(p \,\|\, q) := \mathbb{E}_p\big[\log\big(\frac{dp}{dq}\big)\big]$, along with the χ^2-divergence $\chi^2(p \,\|\, q) := \mathbb{E}_q\big[\big(\frac{dp}{dq} - 1\big)^2\big]$.

2. Background and problem set-up.

In this section, we begin by formulating the value function estimation problem more precisely in Section 2.1. Section 2.2 is devoted to background on reproducing kernel Hilbert spaces (RKHSs), along with a description of the kernel *least squares temporal difference* (LSTD) estimator.

2.1. *Problem formulation.*

A discounted Markov reward process, denoted by $\mathscr{I}(\mathcal{P}, r, \gamma)$, consists of the combination of a Markov chain, a discount factor $\gamma \in (0,1)$, along with a reward function r. In the infinite-horizon discounted setting studied here, the Markov chain is homogeneous, defined on a state space $\mathcal{X}$ with a transition kernel $\mathcal{P} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The

reward function $r : \mathcal{X} \to \mathbb{R}$ models the reward associated with each given state, and for some specified discount factor $\gamma \in (0, 1)$, our goal is to estimate the expected discount sum of all future rewards. More precisely, we define the *value function* $V^* : \mathcal{X} \to \mathbb{R}$ via

$$V^*(x) := \mathbb{E}\big[\textstyle\sum_{h=0}^{\infty} \gamma^h \, r(X_h) \mid X_0 = x\big], \tag{1}$$

where the expectation is taken over a trajectory $(x, X_1, X_2, \ldots)$ from the Markov chain governed by the transition kernel $\mathcal{P}$. The existence of the value function V^* is guaranteed by mild assumptions such as the boundedness of reward r. For future reference, we note that the value function V^* is the solution to the Bellman fixed point equation

$$V^*(x) = r(x) + \gamma \, \mathbb{E}_{X'|x} \left[V^*(X') \mid X = x \right] \qquad \text{for any } x \in \mathcal{X}. \tag{2}$$

In this paper, we study the problem of estimating the value function V^* on the basis of samples from the Markov chain, when the reward function r and discount factor $\gamma \in (0, 1)$ are given.[1] Throughout our discussion, we consider the i.i.d. observation model, where the dataset consists of n i.i.d. sample pairs $\{(x_i, x_i')\}_{i=1}^n \subset \mathcal{X} \times \mathcal{X}$. We let μ be any stationary distribution of the Markov chain $\mathcal{P}$. The sample pair (x_i, x_i') is generated by

$$x_i \sim \mu, \qquad \text{and} \qquad x_i' \sim \mathcal{P}(\cdot \mid x_i). \tag{3}$$

The joint distribution induced by the pair $(\mu, \mathcal{P})$ corresponds to the stationary joint distribution over consecutive state pairs in the Markov chain.

Given an estimate $\widehat{\theta}$ of the value function, we measure its error in the squared-$L^2(\mu)$-norm

$$\|\widehat{\theta} - V^*\|_\mu^2 := \mathbb{E}_{X \sim \mu}\big[\big(\widehat{\theta}(X) - V^*(X)\big)^2\big], \tag{4}$$

where μ is the population distribution of samples $\{x_i\}_{i=1}^n$. In simple cases—such as the tabular setting, in which the state space $\mathcal{X}$ is a finite set—policy evaluation is a parametric problem, since the value function can be encoded as a vector with one entry per state.

Of interest to us in this paper are problems with "richer" state spaces, for which estimating the value function is more challenging, and often non-parametric in nature. In such settings, it is standard to seek approximate solutions of the Bellman operator, via the notion of a *projected fixed point* (e.g., [6, 32, 48, 58]). Given a convex class of functions $\mathbb{G}$ closed in $L^2(\mu)$, the projection operator $\Pi : L^2(\mu) \to \mathbb{G}$ is given by

$$\Pi(f) := \arg\min_{g \in \mathbb{G}} \|g - f\|_\mu \qquad \text{for any function } f \in L^2(\mu). \tag{5}$$

We then seek a solution to the projected fixed point equation

$$\theta^* = \Pi\big(\mathcal{T}(\theta^*)\big) \tag{6}$$

where $\mathcal{T}(\theta^*)(x) := r(x) + \gamma \, \mathbb{E}_{X'|x} \, \theta^*(X')$ is the Bellman operator. Since the Bellman operator is contractive[2] in the $L^2(\mu)$-norm and Π is non-expansive, this fixed point equation has a unique solution. We remark that the function θ^* defined in equation (6) minimizes the mean squared projected Bellman error $\mathrm{MSPBE}(\theta) := \|\theta - \Pi(\mathcal{T}(\theta))\|_\mu^2$.

When the approximating function class $\mathbb{G}$ is chosen to be the linear span of fixed features, then this approach leads to the least-squares temporal difference (LSTD) method. In this paper, our primary focus is more flexible function classes, as defined by reproducing kernel Hilbert spaces. Let us now describe this approach.

---

[1] As we discuss, our results can be easily extended to the setting with an unknown reward function r; so as to bring the essential challenges into clear focus, we take it as known for the bulk of our development.

[2] This fact is a consequence of the choice $\gamma \in (0, 1)$ and the non-expansiveness of the transition operator on $L^2(\mu)$, due to the stationarity of μ.

2.2. *Kernel least-squares temporal differences.* Reproducing kernel Hilbert spaces (RKHSs) provide a fertile ground for developing non-parametric estimators. In this paper, we analyze a standard RKHS-based estimate in reinforcement learning, known as the kernel least-squares estimate, which we now introduce. We begin with some basic background on reproducing kernel Hilbert spaces; see the books [4, 21, 50] for more details.

An RKHS is a particular type of Hilbert space of real-value functions f with domain $\mathcal{X}$. As a Hilbert space, the RKHS has an inner product $\langle f, g \rangle_{\mathbb{H}}$ along with the associated norm $\|f\|_{\mathbb{H}}$. The distinguishing property of an RKHS is the existence of a symmetric kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that acts as the representer of evaluation. In particular, for each $x \in \mathcal{X}$, the function $z \mapsto \mathcal{K}(z, x)$ belongs to the Hilbert space, and moreover we have

$$\langle \mathcal{K}(\cdot, x),\, f \rangle_{\mathbb{H}} = f(x) \quad \text{for all } f \in \mathbb{H}. \tag{7}$$

In order to simplify notation, in much of our development, we adopt the shorthand $\Phi_x := \mathcal{K}(\cdot, x)$ for this *representer of evaluation*.

The population-level kernel LSTD estimate θ^* is, by definition, equal to the projected fixed point (6) with $\mathbb{G}$ corresponding to the closure of $\mathbb{H}$. Since $\mathbb{H}$ is a reproducing kernel Hilbert space, this fixed point has a more explicit expression in terms of certain operators defined on the Hilbert space. In particular, the covariance and cross-covariance operators are defined as

$$\Sigma_{\text{cov}} := \mathbb{E}_{X \sim \mu}[\Phi_X \otimes \Phi_X] \quad \text{and} \quad \Sigma_{\text{cr}} := \mathbb{E}_{(X, X') \sim \mu \times \mathcal{P}}[\Phi_X \otimes \Phi_{X'}]\,. \tag{8}$$

By construction, the covariance operator $\Sigma_{\text{cov}}(f)$, when applied to some $f \in \mathbb{H}$, has the property that $\langle g, \Sigma_{\text{cov}}(f) \rangle_{\mathbb{H}} = \mathbb{E}_{X \sim \mu}[g(X) f(X)]$, with a similar property for the cross-covariance operator. In terms of these operators, the population-level kernel LSTD fixed point must satisfy[3] the fixed point relation

$$\Sigma_{\text{cov}}\, \theta^* = \Sigma_{\text{cov}}\, r + \gamma\, \Sigma_{\text{cr}}\, \theta^*\,. \tag{9}$$

This fixed point relation follows from the convex optimality conditions associated with the projected fixed point. In particular, the error function

$$e(x) := \theta^*(x) - \mathcal{T}(\theta^*)(x) = \theta^*(x) - r(x) - \gamma \mathbb{E}_{X'|x}[\theta^*(X')]$$

must be orthogonal (in $L^2(\mu)$) to any element of the Hilbert space. Enforcing orthogonality with respect to Φ_X and re-arranging leads to the condition (9). We also note that in the special case of linear kernel, equation (9) defines the population version of the least-squares temporal difference (LSTD) estimate; see the book by Sutton and Barto [46] for the derivation of the relation (9) in this special case.

Of more interest to us in this paper is the estimate defined by a richer class of kernel functions. The population-level estimate θ^* depends on the unknown operators Σ_{cov} and Σ_{cr}. In order to obtain an estimator, we need to replace these unknown quantities with data-dependent versions. In this paper, we analyze the *regularized kernel LSTD estimate* $\widehat{\theta}$ given by the solution to the equation

$$\big(\widehat{\Sigma}_{\text{cov}} + \lambda_n \mathcal{I}\big)\, \widehat{\theta} = \big(\widehat{\Sigma}_{\text{cov}} + \lambda_n \mathcal{I}\big)\, r + \gamma\, \widehat{\Sigma}_{\text{cr}}\, \widehat{\theta}. \tag{10}$$

where $\lambda_n > 0$ is a user-defined regularization parameter, $\mathcal{I}$ is the identity operator on the Hilbert space, and we have defined the empirical operators

$$\widehat{\Sigma}_{\text{cov}} := \frac{1}{n} \sum_{i=1}^{n} \Phi_{x_i} \otimes \Phi_{x_i}, \quad \text{and} \quad \widehat{\Sigma}_{\text{cr}} := \frac{1}{n} \sum_{i=1}^{n} \Phi_{x_i} \otimes \Phi_{x_i'}.$$

[3] In writing this equation, we have assumed that the reward function r belongs to the Hilbert space; if not, it should be replaced by the projection $\Pi(r)$.

Note that equation (10) is a fixed point equation in the (possibly infinite-dimensional) Hilbert space. However, as a consequence of the representer theorem [27], this fixed point relation can be formulated as an n-dimensional linear system involving kernel matrices. See Lemma E.1 in Appendix E.1 of the supplementary material for this computationally efficient representation, which we use in our experiments. When the RKHS $\mathbb{H}$ has a finite dimension d, we can also reduce equation (10) to a d-dimensional linear system and solve it efficiently.

Consider the empirical estimate $\widehat{\theta}$ as an estimate of the unknown value function V^*. The error $\|\widehat{\theta} - V^*\|_\mu$ can be upper bounded as

$$\|\widehat{\theta} - V^*\|_\mu \leq \underbrace{\|\widehat{\theta} - \theta^*\|_\mu}_{\text{Estimation error}} + \underbrace{\|\theta^* - V^*\|_\mu}_{\text{Approximation error}} . \tag{11}$$

The approximation error in this decomposition has been studied in past work, and there are various ways to bound it (e.g., [6, 48]); see the papers [32, 58] for some refined and optimal results.

In this paper, our main interest is to study the statistical estimation error $\|\widehat{\theta} - \theta^*\|_\mu$, and to characterize its behavior as a function of sample size and structural properties of the MRP and RKHS. The eigenvalues of the kernel integral operator play an important role here; in particular, under relatively mild conditions (required to satisfy Mercer's theorem, and assumed here), the kernel function admits a decomposition of the form

$$\mathcal{K}(x, z) = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(z), \tag{12}$$

where $\{\mu_j\}_{j=1}^{\infty}$ are a non-negative sequence of eigenvalues, and $\{\phi_j\}_{j=1}^{\infty}$ are the kernel eigenfunctions, orthonormal in $L^2(\mu)$. As we show, the statistical estimation error is controlled by a kernel complexity function that depends on the rate at which the eigenvalues decay.

3. Main results. We now turn to the statement of our main results, along with some discussion of their consequences. Section 3.1 is devoted to upper bounds on the $L^2(\mu)$-error of the kernel LSTD estimator, whereas Section 3.4 provides minimax lower bounds, applicable to any estimator.

3.1. *Non-asymptotic upper bounds on kernel LSTD.* Our first main result provides a non-asymptotic upper bound on the $L^2(\mu)$-error of the kernel LSTD estimator. We begin by stating the assumptions under which this upper bound holds. First, we assume that the kernel function is uniformly bounded, in the sense that

$$\sup_{x \in \mathcal{X}} \sqrt{\mathcal{K}(x, x)} \leq b \tag{13}$$

for some finite constant b. Note that any continuous kernel function over a compact domain $\mathcal{X}$ satisfies this condition; moreover, even on unbounded domains, various standard kernels (e.g., Gaussian, Laplacian etc.) satisfy this condition.

In addition, one of our results—namely, a so-called "fast rate"—requires a bound on the sup-norm of the kernel eigenfunctions $\{\phi_j\}_{j=1}^{\infty}$: that is, we assume that

$$\max_{j \geq 1} \|\phi_j\|_\infty \leq \kappa \qquad \text{for some finite quantity } \kappa. \tag{14}$$

For example, any convolutional kernel has eigenfunctions given by the Fourier basis, and so satisfies this condition. In the examples that follow the theorem, we provide additional examples of kernels that have bounded eigenfunctions.

Central to our analysis is a certain inequality, one that arises from a localized analysis of the empirical process defined by the kernel class. For static prediction problems—that is, problems that lack the dynamic evolution of the RL setting—the idea of localization is well-known to be necessary in order to obtain bounds on the estimation error (e.g., [49]); we also refer the reader to Chapters 13 and 14 in the book [50] for additional background, including specifics on kernel ridge regression (§13.4.2). Our use of localization here identifies very clearly how the structural properties of the Markov reward process determine the statistical accuracy of the estimate. In particular, the key ingredients in this analysis are the following:

Kernel and stationary distribution: The kernel function $\mathcal{K}$ interacts with the MRP's stationary distribution μ so as to determine the eigenvalues $\{\mu_j\}_{j=1}^{\infty}$ of the kernel-integral operator.

Effective horizon: The discount factor $\gamma \in (0,1)$ enters via the effective horizon $H := \frac{1}{1-\gamma}$.

Structural properties of fixed point: The structural properties of the projected fixed point θ^* are captured by a user-defined radius R such that

$$R \geq \max\left\{\|\theta^* - r\|_{\mathbb{H}}, \tfrac{2\|\theta^*\|_\infty}{b}\right\}. \tag{15}$$

Bellman residual variance: Playing the role of the noise level is the variance of the Bellman residual error, when evaluated at θ^*. It is given by

$$\sigma^2(\theta^*) := \mathbb{E}\Big[\big(\theta^*(X) - r(X) - \gamma\theta^*(X')\big)^2\Big], \tag{16}$$

where (X, X') are successive samples from the Markov chain, with the starting state X drawn according to the stationary distribution.

Our analysis differs from that of a static non-parametric regression problem (special case of $\gamma = 0$) in several important ways. First, the effective horizon $H = (1-\gamma)^{-1}$ plays no role in the static setting, and the Bellman residual variance is a more complex object than a typical observation noise variance, since it involves the dynamics defined by the fixed point θ^*.

3.1.1. *Kernel-based critical inequality.* We now turn to the critical inequality that determines the estimation error of the kernel LSTD estimate. It is an inequality that involves the kernel eigenvalues $\{\mu_j\}_{j=1}^{\infty}$, the radius R, and the discount γ via the effective horizon $H(\gamma) = (1-\gamma)^{-1}$. More precisely, we consider positive solutions $\delta > 0$ to the *ζ-based critical inequality*

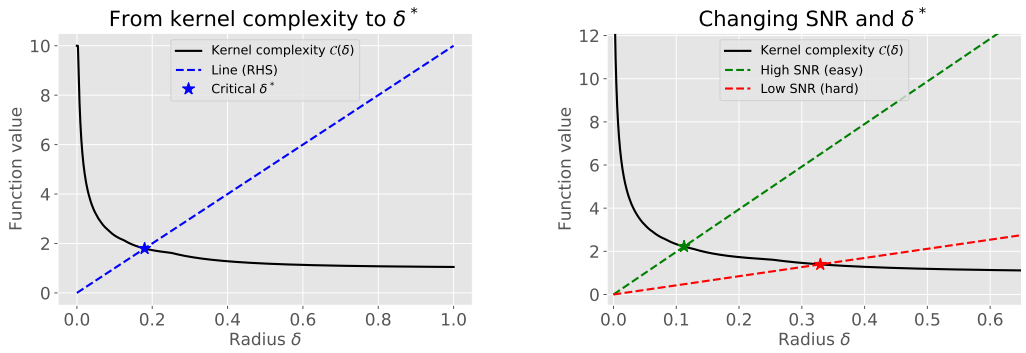
$$\mathcal{C}(\delta) := \sqrt{\sum_{j=1}^{\infty} \min\left\{\frac{\mu_j}{\delta^2}, 1\right\}} \leq \underbrace{\frac{\sqrt{n}\, R}{H(\gamma)\, \zeta}}_{\text{Slope (SNR)}} \delta\,, \tag{CI(ζ))}$$

where $\zeta > 0$ is a parameter to be specified. Note that the function on the left-hand side is decreasing in δ, whereas the right-hand side is linear in δ with the indicated slope. Consequently, inequality (CI(ζ)) has a unique smallest positive solution, which we denote by $\delta_n(\zeta)$. To be clear, in addition to depending on the sample size n and ζ, this smallest positive solution also depends on the eigenvalues as well as the pair (R, γ), but we suppress this dependence so as to simplify notation.

To be clear, the relevance of the kernel complexity function $\mathcal{C}$ on the left-hand side (CI(ζ)) is well-known from past work on kernel ridge regression; in particular, it arises from an analysis of the local Rademacher complexity of a kernel class (e.g., [30, 50]). Equally important

understanding kernel-based LSTD methods—and what is novel in our analysis—is the dependence on the structural parameters on the right-hand of the critical inequality; as our results show, these choices capture precisely how the statistical estimation error of kernel LSTD methods depends on various aspects of the problem structure, including the effective horizon and the Bellman residual variance.

Since the critical inequality (CI(ζ)) plays a central role in our analysis, it is worth gaining intuition for how different components of the MRP affect the solution $\delta_n(\zeta)$. Panel (a) in Figure 2 illustrates the basic geometry of the critical inequality. One instance of the kernel



(a) (b)

Fig 2. Illustrations of the structure of the critical inequality (CI(ζ)). (a) Plots of the kernel complexity $\delta \to \mathcal{C}(\delta)$ on the left-hand side, along with the linear function on the right-hand side. The critical $\delta^* = \delta_n(\zeta)$ is found at the intersection of this curve and line as marked in a blue star. (b) Effects of changing the slope of the right-hand side line, which corresponds to a type of signal-to-noise ratio (SNR). As the SNR decreases, leading to a harder problem, the critical δ^* shifts rightwards to larger values.

complexity function $\delta \mapsto \mathcal{C}(\delta)$, obtained from a kernel with 1-polynomial decaying eigenvalues (see equation (22) in the sequel), is plotted in black. Note that this function is monotonically decreasing in δ. The dotted blue line corresponds to the right-hand side, obtained for a particular value of the slope parameter. The critical radius $\delta^* \equiv \delta_n(\zeta)$, obtained at the intersection of the kernel complexity of this line, is marked with a star.

The slope on the right-hand side of the inequality (CI(ζ)) corresponds to a type of signal-to-noise ratio (SNR). Panel (b) in Figure 2 shows the effect of changing this SNR parameter. As the SNR decreases—so that the slope decreases—the fixed point δ^* shifts rightward to larger values. One consequence of our analysis is that we are able to show precisely the rate at which these leftward and rightward shifts in the statistical estimation error occur, as a function of the MRP's structural parameters (in addition to the sample size n).

3.1.2. *Non-asymptotic upper bounds.* With this set-up and intuition in place, let us now turn to the statement of our non-asymptotic upper bounds on the quality of the kernel LSTD estimate. We provide two guarantees, both of which involve solutions to the the critical inequality CI(ζ) but with different choices of ζ. In each case, the tightest bound is afforded by $\delta_n(\zeta)$. We make two different choices of ζ. First, we establish a bound, one that holds for all sample sizes, with the choice $\zeta = bR$. We then prove a sharper result, one that holds for a finite sample size that is suitably lower bounded, and involves setting $\zeta = \kappa\sigma(\theta^*)$, where $\sigma^2(\theta^*)$ is the variance of the Bellman residual error (16).

Both parts of our theorem guarantee that the kernel LSTD estimator satisfies a bound of the form

$$\|\widehat{\theta} - \theta^*\|_\mu^2 \leq c_1 R^2 \left\{ \delta^2 + \frac{\lambda_n}{1-\gamma} \right\} \tag{17}$$

with probability at least $1 - 2\exp\big(-\frac{c_2 n \delta^2 (1-\gamma)^2}{b^2}\big)$, where (c_1, c_2) are universal constants. The two parts differ in the allowable settings of δ and λ_n for which the bound (17) holds.

THEOREM 3.1 (Non-asymptotic upper bounds). *There is a universal constant c_0 such that:*

(a) *Slow rate: Under the kernel boundedness condition (13), the bound (17) holds for any solution $\delta = \delta(n, R, \gamma, b)$ to the critical inequality $\mathrm{CI}(bR)$ and any $\lambda_n \geq c_0 \delta^2 (1-\gamma)$.*

(b) *Fast rate: Suppose in addition that the kernel eigenfunctions are uniformly bounded (14). Let $\delta_n(\kappa\sigma(\theta^*))$ be the smallest solution to the critical inequality $\mathrm{CI}(\kappa\sigma(\theta^*))$, and suppose that n is large enough to ensure that*

$$R^2 \delta_n^2(\kappa\sigma(\theta^*)) \leq c \, \frac{\kappa \, \sigma^2(\theta^*)}{(1-\gamma)\sqrt{n}}. \tag{18}$$

Then the bound (17) holds for any solution $\delta = \delta(n, R, \gamma, \sigma(\theta^*))$ to the critical inequality $\mathrm{CI}(\kappa\sigma(\theta^*))$ and any $\lambda_n \geq c_0 \delta^2 (1-\gamma)$.

The proof of Theorem 3.1, given in Section 4.1, involves first proving a "basic inequality" that is satisfied by the error $\widehat{\theta} - \theta^*$. We then use empirical process theory and concentration inequalities to establish high probability bounds on the terms in this basic inequality.

It is also interesting to compare the regimes in which the fast and slow rates from our theory apply. See Appendix B in the supplementary material for an in-depth discussion of these regimes.

3.2. *A simpler bound and some corollaries.* It should be noted that the bounds in Theorem 3.1 hold if we set $\delta = \delta_n$, corresponding to the smallest positive solution to the critical inequality CI(ζ), along with $\lambda_n = c_0(1-\gamma)\delta_n^2$. We are then guaranteed to have

$$\|\widehat{\theta} - \theta^*\|_\mu^2 \leq \underbrace{c_1(1+c_0)}_{:=c'} R^2 \delta_n^2 \tag{19}$$

with probability at least $1 - 2\exp\big(-\frac{c_2 n \delta_n^2 (1-\gamma)^2}{b^2}\big)$. Let us consider some examples of this simpler upper bound to illustrate.

3.2.1. *Linear kernels and standard LSTD.* We begin by considering the special case of a linear kernel, in which case the kernel LSTD estimate reduces to the classical linear LSTD estimate. Given a d-dimensional feature map of the form $\varphi : \mathcal{X} \to \mathbb{R}^d$, let us consider linear value functions $\theta(x) := \langle \theta, \varphi(x) \rangle = \sum_{j=1}^d \theta_j \varphi_j(x)$. Here we have overloaded the notation in letting $\theta \in \mathbb{R}^d$ denote a parameter vector. Similarly, we write the reward function as $r(x) = \langle r, \varphi(x) \rangle$ for some vector $r \in \mathbb{R}^d$.

In this case, the Hilbert space can be identified with $\mathbb{R}^d$ equipped with the Euclidean inner product as the Hilbert inner product, and the vector $\varphi(x) \in \mathbb{R}^d$ plays the role of the representer of evaluation. Note that we have $\|\theta^* - r\|_{\mathbb{H}} = \|\theta^* - r\|_2$, and since $\mathcal{K}(x, y) = \langle \varphi(x), \varphi(y) \rangle$, the covariance operator takes the form $\Sigma_{\mathrm{cov}} = \mathbb{E}\big[\varphi(X)\varphi(X)^\top\big] =$

$\sum_{j=1}^{d} \mu_j v_j v_j^T$, a d-dimensional symmetric positive semidefinite matrix with eigenvalues $\{\mu_j\}_{j=1}^{d}$, and eigenvectors $\{v_j\}_{j=1}^{d}$. [4] We have $\mathcal{K}(x,y) = \langle \varphi(x), \varphi(y) \rangle$, and so

$$b = \sup_{x \in \mathcal{X}} \sqrt{\mathcal{K}(x,x)} = \max_{x} \|\varphi(x)\|_2 \quad \text{and} \quad \kappa = \sup_{x \in \mathcal{X}} \max_{j \geq 1} |\langle v_j, \varphi(x) \rangle| / \sqrt{\mu_j}.$$

We now study the structure of the critical inequality CI(ζ), and derive two bounds for the standard LSTD estimate. Both bounds are of the form

$$\|\widehat{\theta} - \theta^*\|_{\mu}^2 = \mathbb{E}\big[\langle \widehat{\theta} - \theta^*, \varphi(X) \rangle^2\big] \leq \varepsilon^2(\zeta) := c' \frac{\zeta^2}{(1-\gamma)^2} \frac{d}{n} \tag{20}$$

for different choices of ζ, and hold with probability at least $1 - 2\exp\big(-\frac{c_2 n \varepsilon^2(\zeta)(1-\gamma)^2}{b^2 R^2}\big)$. We summarize as follows:

COROLLARY 3.2 (Linear kernels and standard LSTD). For the linear kernel and associated standard LSTD estimate:

(a) For any sample size n, the bound (20) holds with

$$\varepsilon^2(bR) = c' \frac{b^2 R^2}{(1-\gamma)^2} \frac{d}{n}. \tag{21a}$$

(b) For a sample size lower bounded as $\sqrt{n} \geq \frac{200 \kappa d}{1-\gamma}$, the bound (20) holds with

$$\varepsilon^2\big(\kappa \sigma(\theta^*)\big) = c' \frac{\kappa^2 \sigma^2(\theta^*)}{(1-\gamma)^2} \frac{d}{n}. \tag{21b}$$

PROOF. For any $\delta > 0$, we have $\sum_{j=1}^{d} \min\big\{\frac{\mu_j}{\delta^2}, 1\big\} \leq d$. Consequently, the critical inequality CI(ζ) is satisfied as long as $\sqrt{d} \leq \frac{\sqrt{n} R (1-\gamma)}{\zeta} \delta$. The smallest $\delta = \delta(\zeta)$ is given by

$$R^2 \delta^2(\zeta) = \frac{\zeta^2}{(1-\gamma)^2} \frac{d}{n}.$$

Setting $\zeta = bR$ yields the claim (21a).

As for the faster rate claimed in the bound (21b), we need to check when the requirement of Theorem 3.1(b)—in particular the sample size condition (18)—is satisfied. In this case, we have $R^2 \delta_n^2(\kappa \sigma(\theta^*)) \leq \frac{\kappa^2 \sigma^2(\theta^*)}{(1-\gamma)^2} \frac{d}{n}$, so that in order to satisfy the bound (18), it suffices to have $\sqrt{n} \geq \frac{200 \kappa d}{1-\gamma}$. The claim (21b) then follows. □

3.2.2. *Kernels with α-polynomial decay.* Let us now consider a "richer" class of kernel functions, for which the kernel estimator is truly non-parametric. In particular, let us consider the class of kernels that satisfy the *α-polynomial decay condition*

$$\mu_j \leq c\, j^{-2\alpha} \quad \text{for some exponent } \alpha > \tfrac{1}{2}. \tag{22}$$

There are many examples of kernels used in practice that satisfy a decay condition of this form, including the Laplacian kernel $\mathcal{K}(x,x') = \exp(-\|x - x'\|_1)$, as well as various types of Sobolev and spline kernels that are used in non-parametric regression and density estimation. See Chapters 12 and 13 in the book [50] for more details on such kernels.

[4] The j-th eigenfunction of kernel $\mathcal{K}$ takes the form $\phi_j(x) := \langle v_j, \varphi(x) \rangle / \sqrt{\mu_j}$, which satisfies $\|\phi_j\|_{\mu} = 1$.

Let us study the structure of the critical inequality CI(ζ) for kernels whose eigenvalues satisfy the α-polynomial decay condition (22). We derive two bounds, both of which are of the form

$$\|\widehat{\theta} - \theta^*\|_\mu^2 \leq \varepsilon^2(\zeta) := R^2 \underbrace{c'\Big(\frac{\zeta^2}{R^2(1-\gamma)^2}\frac{1}{n}\Big)^{\frac{2\alpha}{2\alpha+1}}}_{\delta^2(\zeta)} = c'\, R^{\frac{2}{2\alpha+1}}\Big(\frac{\zeta^2}{(1-\gamma)^2}\frac{1}{n}\Big)^{\frac{2\alpha}{2\alpha+1}}, \tag{23}$$

for different choices of ζ, and hold with probability at least $1 - 2\exp\big(-\frac{c_2 n\delta^2(\zeta)(1-\gamma)^2}{b^2}\big)$.

COROLLARY 3.3. (a) For any sample size n, the bound (23) holds with

$$\varepsilon^2(bR) = c' R^2\Big(\frac{b^2}{(1-\gamma)^2}\frac{1}{n}\Big)^{\frac{2\alpha}{2\alpha+1}}. \tag{24a}$$

(b) Suppose that the sample size n is large enough to ensure that $R^2\delta_n^2(\sigma(\theta^*)) \leq c\,\frac{\kappa\,\sigma^2(\theta^*)}{(1-\gamma)\sqrt{n}}$. Then the bound (23) holds with

$$\varepsilon^2\big(\kappa\sigma(\theta^*)\big) = c' R^2\Big(\frac{\kappa^2\sigma^2(\theta^*)}{R^2(1-\gamma)^2}\frac{1}{n}\Big)^{\frac{2\alpha}{2\alpha+1}}. \tag{24b}$$

PROOF. Let us find a solution to the critical inequality CI(ζ) for a kernel satisfying the α-polynomial decay condition (22). Let k be the largest positive integer such that $\delta^2 \leq ck^{-2\alpha}$. With this choice, we have

$$\sqrt{\sum_{j=1}^{\infty}\min\Big\{\frac{\mu_j}{\delta^2}, 1\Big\}} \leq \sqrt{k + \frac{c}{\delta^2}\sum_{j=k+1}^{\infty} j^{-2\alpha}}.$$

Now we have

$$\sum_{j=k+1}^{\infty} j^{-2\alpha} \leq \int_k^{\infty} t^{-2\alpha}dt \;\leq\; \frac{1}{2\alpha-1}(1/k)^{2\alpha-1}.$$

Consequently, there is a universal constant c', depending only on α, such that the critical inequality CI(ζ) will be satisfied for a $\delta > 0$ such that $c'\delta^{-\frac{1}{2\alpha}} \leq \sqrt{n}\,\frac{R(1-\gamma)}{\zeta}\delta$. Solving this inequality yields that

$$\delta^2 \asymp \Big(\frac{\zeta^2}{R^2(1-\gamma)^2}\frac{1}{n}\Big)^{\frac{2\alpha}{2\alpha+1}} \tag{25}$$

satisfies the critical inequality CI(ζ).

Putting together the pieces, we conclude that there is a universal constant c such that

$$\|\widehat{\theta} - \theta^*\|_\mu^2 \leq cR^2\Big(\frac{\zeta^2}{R^2(1-\gamma)^2}\frac{1}{n}\Big)^{\frac{2\alpha}{2\alpha+1}} = c\,R^{\frac{2}{2\alpha+1}}\Big(\frac{\zeta^2}{(1-\gamma)^2}\frac{1}{n}\Big)^{\frac{2\alpha}{2\alpha+1}} \tag{26}$$

with high probability. This bound holds with $\zeta = bR$ for all sample sizes, and it holds with $\zeta = \kappa\sigma(\theta^*)$ once the sample size is sufficiently large to ensure that

$$R^2\delta_n^2(\kappa\sigma(\theta^*)) \asymp R^{\frac{2}{2\alpha+1}}\Big(\frac{\kappa^2\sigma^2(\theta^*)}{(1-\gamma)^2}\frac{1}{n}\Big)^{\frac{2\alpha}{2\alpha+1}} \lesssim \frac{\kappa\,\sigma^2(\theta^*)}{(1-\gamma)\sqrt{n}}.$$

Since $\frac{2\alpha}{2\alpha+1} > \frac{1}{2}$, this bound will hold once n exceeds a finite threshold. □

Connection to effective dimension: In order to gain intuition for our theoretical results, it can be helpful to consider a notion of effective dimension. Given a solution δ_n to the critical inequality, let us define

$$d_n \equiv d_n(\delta_n) := \max\left\{j \mid \mu_j \geq \delta_n^2\right\}. \tag{27}$$

This notion of effective dimension corresponds to the number of eigenvalues above the squared error level δ_n^2 returned by our critical inequality. As a concrete illustration, suppose that the kernel has eigenvalues decaying $\mu_j \asymp j^{-2\alpha}$. In ths special case, the calculations from equations (25) and (26) imply that

$$d_n \asymp \left(\frac{\sqrt{n}\, R(1-\gamma)}{\zeta}\right)^{\frac{2}{2\alpha+1}} \qquad \text{and} \qquad \|\widehat{\theta} - \theta^*\|_{\mu}^2 \lesssim \frac{\zeta^2}{(1-\gamma)^2}\frac{d_n}{n}. \tag{28}$$

Thus, when reformulated in terms of effective dimension, our bounds on the estimation error $\|\widehat{\theta} - \theta^*\|_{\mu}^2$ now take a form similar to inequality (20) in the linear kernel example, except that we replace the dimension d of linear kernel with the effective dimension d_n from equation (28).

It should be observed that—via its dependence on the critical error δ_n—the effective dimension d_n is affected by several structural properties of the problem. In particular, we observe:

- For non-parametric problems, the effective dimension d_n grows as the sample size n increases, so the squared error $\|\widehat{\theta} - \theta^*\|_{\mu}^2$ decays more slowly than the classical parametric rate n^{-1}.
- For sufficiently regular problems, the term $R(1-\gamma)/\zeta$ gets smaller as we increase the effective horizon $H = (1-\gamma)^{-1}$. It shows that the effective dimension d_n shrinks as horizon H grows, so as to maintain a simple model and stabilize the estimation. In this way, for a non-parametric problem, the overall dependence of error $\|\widehat{\theta} - \theta^*\|_{\mu}^2$ on the effective horizon H is always milder than that of the linear kernel, for which it grows cubically (as H^3) [19]. This highlights an interesting distinction between parametric and non-parametric procedures in policy evaluation that does not seem to have been appreciated to date.

3.3. *Some illustrative simulations.* Some simulations are useful in illustrating the predictions of our theory, and most concretely the sharpness of Corollary 3.3. In particular, from the bound (23), the error depends on the eigenvalue exponent α from equation (22) in two distinct ways. On one hand, the dependence on the effective horizon $H = \frac{1}{1-\gamma}$ worsens as the exponent α increases. On the other hand, the dependence on the inverse sample size $(1/n)$—corresponding to how quickly the estimation error vanishes—improves as α increases. Corollary 3.3 makes very explicit predictions about these dependencies, and the sharpness of these predictions can be verified empirically.

In order to do so, we constructed three different kernels $\mathcal{K}_i$, $i = 1, 2, 3$ with eigenvalues $\{\mu_j\}_{j=1}^{\infty}$ decaying as

$$\mu_j(\mathcal{K}_i) = \begin{cases} j^{-6/5} & \text{for } i = 1 \\ j^{-2} & \text{for } i = 2 \\ \exp\left(-(j-1)^2\right) & \text{for } i = 3. \end{cases} \tag{29}$$

Note that $\mathcal{K}_1$ has α-polynomial decay (22) with $\alpha = 3/5$, $\mathcal{K}_2$ with $\alpha = 1$, and the exponential decay of $\mathcal{K}_3$ can be viewed as a limiting case $\alpha = +\infty$.

In parallel, we constructed two different probability transition functions that allowed us to vary the dependence of the radius R and the Bellman residual variance $\sigma^2(\theta^*)$ on the effective horizon.

"Hard" ensemble : The transition function underlying our hard ensemble is constructed so that

$$R \asymp \sigma^2(\theta^*) \asymp \frac{1}{1-\gamma},$$

where the notation $\asymp$ means bounded above and below by constants independent of γ.

"Easy" ensemble: For our easy ensemble, we construct the probability transition matrix and rewards so that both R and $\sigma^2(\theta^*)$ remain of constant order as γ is varied.

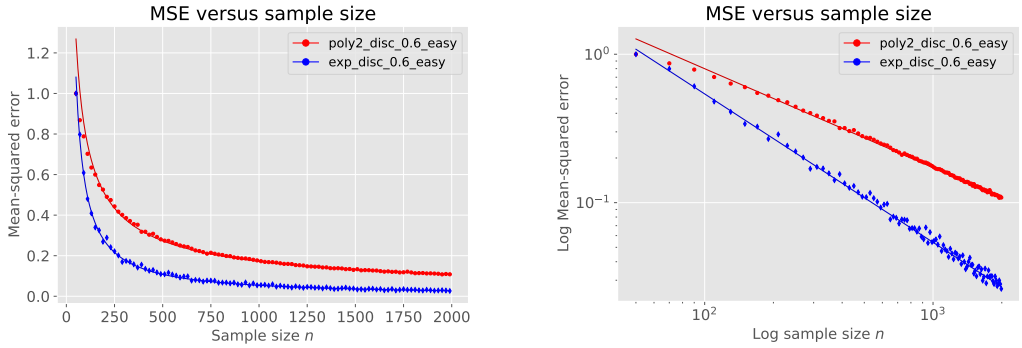
See Appendix A of the supplementary material for more details on these constructions. In all cases, we implemented the kernel LSTD estimate using the regularization parameter $\lambda_n = c(1-\gamma)\delta_n^2$ for a fixed constant $c = 0.01$.

3.3.1. *Dependence on sample size.* We begin by commenting the dependence of the kernel LSTD estimator on the sample size, ensuring that our guarantees are consistent with known results. For any kernel with α-polynomial decay (22), Corollary 3.3 predicts that the mean-squared error should decay as

$$\|\widehat{\theta} - \theta^*\|_{\mu}^2 \asymp \left(\frac{1}{n}\right)^{\frac{2\alpha}{2\alpha+1}}, \tag{30}$$

where, for this particular comparison, we disregard other terms that are independent of the sample size n. This decay rate is a standard one in the context of non-parametric regression [44, 50], so to be expected here as well.

The novelty in our analysis lies in the characterization of the other structural parameters, to which we now turn.



(a) (b)

Fig 3. Plots of the mean-square error $\mathbb{E}\|\widehat{\theta} - \theta^*\|_{\mu}^2$ versus the sample size n for two different kernels. For each point (for each curve on each plot), the MSE was approximated by taking a Monte Carlo average over $T = 2000$ trials with the sample standard deviations shown as error bars. Our theory predicts that the mean-squared error should drop off as $\left(\frac{1}{n}\right)^{\nu}$ for an exponent $\nu > 0$ determined by the kernel. Solid curves correspond to these theoretical predictions. We fixed discount factor $\gamma = 0.6$. (a) MSE versus sample size on ordinary scale. Our theory predicts that (disregarding logarithmic factors), the MSE should scale as $\left(\frac{1}{n}\right)$ for the exponential kernel $\mathcal{K}_3$, and as $\left(\frac{1}{n}\right)^{2/3}$ for the 1-polynomial decaying kernel $\mathcal{K}_2$. Note that these theoretical predictions align very well with the empirical behavior. (b) Plots of the same data on a log-log scale, showing the expected linear relationship between log MSE and log sample size.

3.3.2. *Dependence on effective horizon.* In our second simulation study, we examine the behavior of the $L^2(\mu)$-error as a function of the effective horizon $H := \frac{1}{1-\gamma}$. For kernels with eigenvalues that exhibit α-polynomial decay, our theory—in particular via the bound (24b) from Corollary 3.3—gives specific predictions about this dependence as well.

- With the probability transitions from the "hard" ensemble, it can be shown that $R \asymp \sigma^2(\theta^*) \asymp H = \frac{1}{1-\gamma}$. As a consequence, our theory predicts that for a fixed sample size n, we should observe the following scaling

$$\|\widehat{\theta} - \theta^*\|_\mu^2 \asymp H^{\frac{2(3\alpha+1)}{2\alpha+1}}. \tag{31a}$$

- With the probability transitions from our "easy" ensemble, for which $R \asymp \sigma^2(\theta^*) \asymp 1$, the predicted slope of this linear scaling is

$$\|\widehat{\theta} - \theta^*\|_\mu^2 \asymp H^{\frac{4\alpha}{2\alpha+1}}. \tag{31b}$$

See Appendix A of the supplementary material for the calculations of both of these theoretical predictions. Note that predictions for the kernel $\mathcal{K}_3$, with its exponentially decaying values, can be obtained as a limiting case with $\alpha \to +\infty$.

3.4. *Minimax lower bounds.* Thus far, we have established some upper bounds on the performance of a specific estimator. To what extent are these bounds improvable? In order to answer this question, it is natural to investigate the fundamental (statistical) limitations of kernel-based value function estimation. In this section, we do so by deriving some minimax lower bounds on the behavior of any procedures for estimating the value function.

Minimax lower bounds are obtained by assessing the performance of any estimator in a uniform sense over a particular class of problems. In particular, for classes of MRPs $\mathfrak{M}$ to be defined, we prove lower bounds of the following type. For a given MRP instance $\mathscr{I}$, we assume that we observe a dataset $\{(x_i, x_i')\}_{i=1}^n$ of n i.i.d. samples generated from the given MRP. An estimator $\widehat{\theta}$ of the value function is any measurable function of the data mapping into $\mathbb{R}^{\mathcal{X}}$. For suitable classes $\mathfrak{M}$ indexed by pairs of parameters $(\bar{R}, \bar{\sigma})$, we prove that the squared-$L^2(\mu)$ error of any estimator, when measured in a uniform sense over the family, is lower bounded as $c_1 \bar{R}^2 \delta_n^2$. Here $c_1 > 0$ is a universal constant, and the error parameter δ_n critical inequality (CI(ζ)) that specifies our upper bounds; see equation (33) for the precise definition.

3.4.1. *Families of MRPs and regular kernels.* We begin by describing the families of MRPs over which we prove minimax lower bounds. In all of our constructions, both the reward function r and the optimal value function θ^* are members of a Hilbert space with a set of eigenfunctions $\{\phi_j\}_{j=1}^\infty$, and a sequence of eigenvalues $\{\mu_j\}_{j=1}^\infty$ that vary as part of the construction. In all cases, our construction ensures that the eigenfunction bound (14) holds with $\kappa = 2$, along with the kernel being trace class. In particular, we have

$$\max_{j \geq 1} \|\phi_j\|_\infty \leq \kappa = 2, \quad \text{and} \quad \sum_{j=1}^\infty \mu_j \leq \tfrac{b^2}{4}. \tag{32a}$$

Note that these conditions imply that

$$\sup_{x \in \mathcal{X}} \sqrt{\mathcal{K}(x,x)} = \sup_{x \in \mathcal{X}} \Big(\sum_{j=1}^\infty \mu_j \phi_j^2(x) \Big)^{1/2} \leq b,$$

so that the b-boundedness condition (13) from our upper bound holds. In addition, our families of MRPs are also defined by the constraints

$$\max\left\{\|\theta^* - r\|_{\mathbb{H}}, \tfrac{2\|\theta^*\|_\infty}{b}\right\} \leq \bar{R}, \quad \text{and} \quad \sigma(\theta^*) \leq \bar{\sigma}. \tag{32b}$$

We say that a family $\mathfrak{M}$ of MRPs is $(\bar{R}, \bar{\sigma})$-*valid* if its members satisfy the bound (32b), along with the conditions (32a).

So as to match our upper bounds, we prove lower bounds that involve an error term δ_n defined as the smallest positive solution to the inequality

$$\sqrt{\sum_{j=1}^{\infty} \min\left\{\frac{\mu_j}{\delta^2}, 1\right\}} \leq \sqrt{n}\, \frac{\bar{R}(1-\gamma)}{2\bar{\sigma}}\, \delta. \tag{33}$$

From past work on kernel ridge regression [54], it is known that such lower bounds cannot hold for kernels with eigenvalues that decay in pathological ways. The notion of a regular kernel, which we define here, precludes such pathology. For a given δ_n, the associated statistical dimension $d_n \equiv d_n(\delta_n)$ is given by $d_n(\delta_n) := \max\left\{j \mid \mu_j \geq \delta_n^2\right\}$. The kernel is regular if there is a universal constant c such that

$$\left\{\frac{2\bar{\sigma}}{\bar{R}(1-\gamma)}\right\}^2 d_n \geq c\, n\, \delta_n^2. \tag{34}$$

Standard kernels, including the linear kernel and more general kernels with eigenvalues that decay at a polynomial or exponential rate, are all regular.

3.4.2. *Statement of bounds.* With this set-up, we are now ready to state our minimax lower bounds. For a given $(\bar{R}, \bar{\sigma})$-*valid* family of MRPs, we say that the lower bound $\mathrm{LB}(\bar{R}, \bar{\sigma}, \delta_n)$ holds if

$$\inf_{\widehat{\theta}} \sup_{\mathscr{J} \in \mathfrak{M}(\bar{R}, \bar{\sigma})} \mathbb{P}_{\mathscr{J}}\left(\|\widehat{\theta} - \theta^*\|_{\mu}^2 \geq c_1\, \bar{R}^2 \delta_n^2\right) \geq c_2. \tag{$\mathrm{LB}(\bar{R}, \bar{\sigma}, \delta_n)$}$$

In this statement, the quantities (c_1, c_2) are universal constants.

We prove minimax lower bounds in two regimes of parameters $(\bar{R}, \bar{\sigma})$, depending on how these parameters scale with the effective horizon $\frac{1}{1-\gamma}$. In Regime A, this scaling is linear in the effective horizon—namely

$$\bar{R} \geq \tfrac{1}{6(1-\gamma)} \max\left\{\tfrac{\gamma}{\sqrt{\mu_1}}, \tfrac{2}{b}\right\}, \quad \text{and} \quad \bar{\sigma}^2 \in \left[\tfrac{1+\gamma}{5(1-\gamma)}, \tfrac{1+\gamma}{1-\gamma}\right]. \tag{35a}$$

In Regime B, by contrast, both of these quantities can be order one with the effective horizon—viz.

$$\bar{R} \geq \max\left\{\tfrac{1}{2\sqrt{\mu_1}}, \tfrac{2}{\gamma b}\right\}, \quad \text{and} \quad \bar{\sigma}^2 \in \left(\tfrac{1}{8}, 1\right]. \tag{35b}$$

We discuss the motivation for considering these two regimes following the statement of our bounds.

THEOREM 3.4 (Minimax lower bounds). (a) *For any pair* $(\bar{R}, \bar{\sigma})$ *in Regime A* (35a)*, there is a* $(\bar{R}, \bar{\sigma})$*-valid family of MRPs such that the lower bound* $\mathrm{LB}(\bar{R}, \bar{\sigma}, \delta_n)$ *holds for any sample size* n *such that*

$$\bar{R}^2 \delta_n^2 \leq \frac{2\,\kappa \bar{\sigma}^2}{(1-\gamma)^{\frac{3}{2}}\sqrt{n}}. \tag{36a}$$

(b) *Consider any pair* $(\bar{\sigma}, \bar{R})$ *in Regime B* (35b)*, and suppose that the eigensequence satisfies* $\min_{3 \leq j \leq d_n} \{\sqrt{\mu_{j-1}} - \sqrt{\mu_j}\} \geq \frac{\delta_n}{2d_n}$*. Then there is a* $(\bar{R}, \bar{\sigma})$*-valid family of MRPs such that the lower bound* $\mathrm{LB}(\bar{R}, \bar{\sigma}, \delta_n)$ *holds for a sample size* n *large enough such that*

(36b)
$$\bar{R}^2 \delta_n^2 \leq \frac{12\, \kappa \bar{\sigma}^2}{(1-\gamma)\sqrt{n}} \qquad \text{and} \qquad \bar{R}\delta_n \leq 10 \kappa \bar{\sigma} \left(1 - \tfrac{\mu_2}{\mu_1}\right) \min\left\{ \frac{\kappa\bar{\sigma}/(\sqrt{\mu_1}\bar{R})}{(1-\gamma)^2 \log n}, \frac{\sqrt{\mu_1}}{b} \right\}.$$

See Section 4.2 for the proof of Theorem 3.4.

The main take-away from this result is the following: by comparing the bounds in Theorem 3.4 with the achievable rate from Theorem 3.1(b), we see that the kernel LSTD estimator is an optimal procedure. More precisely, it achieves the minimax-optimal scaling $\bar{R}^2\delta_n^2$ of the squared-$L^2(\mu)$ norm. As we discuss below, there are some differences in the minimum sample size required for the bounds to be valid, with the lower bound requirements being less stringent than our upper bounds from Theorem 3.1.

A few high-level comments on the proof: it makes use of the Fano method, a well-known approach for proving minimax lower bounds (e.g., [50, 57]). This method involves constructing a family of MRP instances that are "well-separated", and arguing that any method with relatively low estimation error is capable of solving a multi-way testing problem defined over this family. For static estimation problems, including non-parametric regression and density estimation, it is well understood how to construct such families (e.g., [22, 44]). The novelty and technical challenges in our analysis lie in the construction of such difficult families for MRPs; doing so requires negotiating the interplay between the kernel structure and the Markovian dynamics. In our construction, we begin by setting up a "difficult" ensemble of MRPs with a (discrete) 2-point state space. We then use a tensorization approach, along with the Walsh basis, to embed many copies of this 2-state MRP into a non-parametric problem with state space $\mathcal{X} = [0, 1)$. By also defining the kernels using the Walsh basis, we retain complete control over the eigenvalues, so that we can establish the claimed lower bounds.

Some differences. Our upper and lower bounds differ in terms of their required lower bounds on sample size; as we discuss in Appendix D.1 of the supplementary material, the requirements of the lower bounds in Theorem 3.4 are milder than our corresponding condition for the kernel LSTD estimate. Apart from the sample size conditions, Theorem 3.4 also requires the kernel regularity condition (34), along with the eigensequence condition in part (b). As we discuss in more detail in Appendix D.1 of the supplementary material, these conditions are relatively mild, and satisfied by various kernels used in practice (including any kernel with eigenvalues that exhibit α-polynomial decay (22)).

Regimes of $(\bar{R}, \bar{\sigma})$. Let us now discuss the two regimes of parameters.

- The scalings in Regime A (35a) arise naturally when we assume only that the reward function is uniformly bounded—say $\|r\|_\infty \leq 1$. In this case, by the law of total variance [43], we have the bound $\sigma^2(\theta^*) \leq \frac{2}{1-\gamma}$, and there exist MRPs for which this $(1-\gamma)^{-1}$ is achieved, consistent with the first inclusion in condition (35a). In terms of the choice of $\bar{R}$, we can construct MRPs with bounded reward functions such that $\|\theta^* - r\|_\mu \lesssim \frac{1}{1-\gamma}$ and $\|\theta^*\|_\infty \lesssim \frac{1}{1-\gamma}$. With these scalings, the constraint on $\bar{R}$ in condition (35a) is satisfied.
- Turning to Regime B (35b), it corresponds to a class of problems for which estimation is much easier. Instances with this scaling arise when we impose a constraint of the form $\|\theta^*\|_\mu \leq 1$. This constraint ensures that $\sigma^2(\theta^*) \leq 1$ because the variance is

dominated by the second moment. As for the parameter $\bar{R}$, the RKHS norm is connected with the $L^2(\mu)$-norm via inequality $\|\theta^* - r\|_{\mathbb{H}} \geq \frac{1}{\sqrt{\mu_1}}\|\theta^* - r\|_\mu$. Therefore, we can ensure that the constraint $\bar{R} \geq \max\{\|\theta^* - r\|_{\mathbb{H}}, \frac{2\|\theta^*\|_\infty}{b}\}$ holds by constructing MRPs with $\|\theta^* - r\|_\mu \lesssim 1$ and $\|\theta^*\|_\infty \lesssim \frac{1}{\gamma}$. With these choices, we can ensure that $\bar{R} \gtrsim \max\{\frac{1}{\sqrt{\mu_1}}, \frac{1}{\gamma b}\}$, as required in the definition (35b).

4. Proofs. We now turn to the proofs of our main results. Section 4.1 is devoted to overviews of the proofs of the upper bounds stated in Theorem 3.1, whereas Section 4.2 contains the proofs of the lower bounds stated in Theorem 3.4. In this main body, given space constraints, we focus on providing the higher level road map to the proof structure, and defer the technically more challenging aspects of the proofs to the supplementary material.

4.1. *Proof overview for Theorem 3.1.* In this section, we provide an overview of the proof of the finite-sample upper bounds stated in Theorem 3.1. Our proof consists of three main steps. First, we use the definition of the estimator to derive a basic inequality to give an upper bound on the the squared $L^2(\mu)$ error. Then we use techniques from empirical process theory and concentration of measure to upper bound the terms on the right-hand side of this inequality. Finally, we exploit this analysis to choose the regularization parameter λ_n in a manner that yields an optimal trade-off between the bias and variance terms.

4.1.1. *The building blocks.* Recall that $\widehat{\theta}$ denotes our estimate, whereas θ^* denotes the population-level kernel LSTD solution. We begin our analysis by deriving an inequality that must be satisfied by the the error $\widehat{\Delta} = \widehat{\theta} - \theta^*$. We state our results in terms of the functional

$$\rho(f) := \Big(\mathbb{E}[f^2(X) - \gamma f(X)f(X')]\Big)^{1/2}, \tag{37}$$

where (X, X') are successive states sampled from the Markov chain, with X drawn according to the stationary distribution. As shown in the proof of Theorem 4.1 below, it follows from the Cauchy-Schwarz inequality that we always have the lower bound

$$\mathbb{E}[f^2(X) - \gamma f(X)f(X')] \geq (1-\gamma)\|f\|_\mu^2 \geq 0. \tag{38}$$

so that our definition of ρ is meaningful.

A basic inequality on the error. We begin by stating an inequality that must be satisfied by the error. It lies at the foundation of our analysis:

LEMMA 4.1 (Basic inequality). *The error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ satisfies the inequality*

$$(1-\gamma)\|\widehat{\Delta}\|_\mu^2 \overset{(i)}{\leq} \rho^2(\widehat{\Delta}) \overset{(ii)}{=} \Big\{\sum_{j=1}^{3} T_j\Big\} - \lambda_n\|\widehat{\Delta}\|_{\mathbb{H}}^2, \tag{39}$$

where

$$T_1 = \big\langle \widehat{\Delta},\, \widehat{\Sigma}_{\mathrm{cov}}(r - \theta^*) + \gamma\widehat{\Sigma}_{\mathrm{cr}}\theta^* \big\rangle_{\mathbb{H}}, \tag{40a}$$

$$T_2 = \lambda_n \big\langle \widehat{\Delta},\, r - \theta^* \big\rangle_{\mathbb{H}}, \tag{40b}$$

$$T_3 = \big\langle \widehat{\Delta},\, (\Gamma - \widehat{\Gamma})\widehat{\Delta} \big\rangle_{\mathbb{H}}, \tag{40c}$$

where $\Gamma = \Sigma_{\mathrm{cov}} - \gamma\Sigma_{\mathrm{cr}}$ *and* $\widehat{\Gamma} = \widehat{\Sigma}_{\mathrm{cov}} - \gamma\widehat{\Sigma}_{\mathrm{cr}}$.

See Appendix C.1 of the supplementary material for the proof of this claim. This proof is relatively straightforward, based on exploiting the conditions that define the population and empirical projected fixed points.

Controlling the terms. Our next step is to derive upper bounds on the three terms on the right-hand side of our basic inequality (39). Recall that $\|\theta^* - r\|_{\mathbb{H}} \leq R$ by assumption.

The quantity T_2 is easily handled: we have

$$T_2 \overset{(i)}{\leq} \lambda_n \|\widehat{\Delta}\|_{\mathbb{H}} \|r - \theta^*\|_{\mathbb{H}} \overset{(ii)}{\leq} \frac{\lambda_n}{2} \left\{ \|\widehat{\Delta}\|_{\mathbb{H}}^2 + R^2 \right\}, \tag{41}$$

where step (i) follows from the Cauchy-Schwarz inequality, and step (ii) follows from the Fenchel-Young inequality.

As for the terms T_1 and T_3, we state some auxiliary lemmas that bound them with high probability.

LEMMA 4.2. *Let $\delta_n = \delta_n(\zeta)$ for either $\zeta = bR$ or $\zeta = \kappa\sigma(\theta^*)$. There are are universal constants (c, c') such that*

$$T_1 \leq c\,(1-\gamma)\,\delta_n^2 \left\{ \|\widehat{\Delta}\|_{\mathbb{H}}^2 + R^2 \right\} + c\,R\,(1-\gamma)\,\delta_n \|\widehat{\Delta}\|_{\mu} \tag{42}$$

with probability at least $1 - \exp\big(-c' \frac{n\delta_n^2(1-\gamma)^2}{b^2}\big)$.

See Appendix C.2 of the supplementary material for the proof of this claim. This proof involves more technical effort: a reformulation in terms of the supremum of a certain empirical process, followed by a localization step so as to obtain the sharp rates given here.

LEMMA 4.3. *(a) With the choice $\delta_n = \delta_n(bR)$ there are universal constants (c, c') such that*

$$T_3 \leq c\,(1-\gamma)\,\delta_n^2 \left\{ \|\widehat{\Delta}\|_{\mathbb{H}}^2 + R^2 \right\} + \frac{\rho^2(\widehat{\Delta})}{2}, \tag{43}$$

with probability at least $1 - \exp\big(-c' \frac{n\delta_n^2(1-\gamma)}{b^2}\big)$.
(b) If, in addition, the sample size condition (18) holds, then the same bound holds with $\delta_n = \delta_n(\kappa\sigma(\theta^*))$.

The proof of this claim is given in Appendix C.3 of the supplementary material. It is the most technically challenging of the three, making use of localization involving both functional $\rho(f)$ from equation (37) and the Hilbert norm. As shown in this proof, establishing the bounds in Lemma 4.3 amounts to proving a non-asymptotic bound on the supremum of an empirical process involving functions of the form $g(x, x') = f^2(x) - \gamma f(x) f(x')$, uniformly over a suitable class of functions f.

4.1.2. *Putting together the pieces.* We now put together the pieces in order to complete the proof of Theorem 3.1. In particular, we use Theorems 4.2 and 4.3 to bound the terms $\{T_j\}_{j=1}^3$ on the right hand side of the bound (39) from Theorem 4.1. Applying all of these bounds and combining all the terms, we find that with probability at least $1 - 2\exp\big(-c' \frac{n\delta_n^2(1-\gamma)^2}{b^2}\big)$, we have

$$\rho^2(\widehat{\Delta}) \leq \underbrace{c(1-\gamma)\delta_n^2 \Big\{ \|\widehat{\Delta}\|_{\mathbb{H}}^2 + R^2 \Big\} + cR(1-\gamma)\,\delta_n \|\widehat{\Delta}\|_{\mu}}_{\text{Bound on } T_1} + \underbrace{\frac{\lambda_n}{2} \Big\{ \|\widehat{\Delta}\|_{\mathbb{H}}^2 + R^2 \Big\}}_{\text{Bound on } T_2}$$
$$+ \underbrace{c(1-\gamma)\,\delta_n^2 \Big\{ \|\widehat{\Delta}\|_{\mathbb{H}}^2 + R^2 \Big\} + \frac{\rho^2(\widehat{\Delta})}{2}}_{\text{Bound on } T_3} - \lambda_n \|\widehat{\Delta}\|_{\mathbb{H}}^2.$$

Re-arranging terms yields

$$\tfrac{1}{2}\rho^2(\widehat{\Delta}) \leq cR\,(1-\gamma)\,\delta_n\|\widehat{\Delta}\|_{\mu} + \|\widehat{\Delta}\|_{\mathbb{H}}^2\Big\{2c(1-\gamma)\delta_n^2 - \frac{1}{2}\lambda_n\Big\} + R^2\Big\{2c(1-\gamma)\delta_n^2 + \frac{1}{2}\lambda_n\Big\}.$$

Setting $\lambda_n \geq 4c(1-\gamma)\delta_n^2$ ensures that the second term is negative. Combining with the lower bound $\rho^2(\widehat{\Delta}) \geq (1-\gamma)\|\widehat{\Delta}\|_{\mu}^2$, we find that

$$\frac{1-\gamma}{2}\|\widehat{\Delta}\|_{\mu}^2 \leq cR(1-\gamma)\,\delta_n\|\widehat{\Delta}\|_{\mu} + \lambda_n R^2.$$

By the Fenchel-Young inequality, we have

$$cR\,(1-\gamma)\,\delta_n\|\widehat{\Delta}\|_{\mu} + \lambda_n R^2 \leq \tfrac{1-\gamma}{4}\|\widehat{\Delta}\|_{\mu}^2 + c^2R^2\,(1-\gamma)\,\delta_n^2 + \lambda_n R^2.$$

Putting together the pieces, we conclude that there is a universal constant $\bar{c}$ such that

$$\|\widehat{\Delta}\|_{\mu}^2 \leq \bar{c}R^2\left\{\delta_n^2 + \frac{\lambda_n}{1-\gamma}\right\},$$

as claimed. This concludes the proof of Theorem 3.1 for $\delta = \delta_n$.

We note that all of the same steps actually hold for any $\delta \geq \delta_n$, so that the bound given in the theorem is also valid.

4.2. *Proof overview for Theorem 3.4.* We now turn to an overview of the proof of the minimax lower bounds stated in Theorem 3.4. In Section 4.2.1, we explicitly define the MRP families $\mathfrak{M}_A$ and $\mathfrak{M}_B$. Section 4.2.2 then presents a high-level overview of the proof structure that works for both Regimes A and B. The bulk of our technical analysis is deferred to Appendix D of the supplementary material, where we provide the constructions of RKHSs $\mathbb{H}_A$ and $\mathbb{H}_B$ and MRP model families $\mathfrak{M}_A$ and $\mathfrak{M}_B$, and verify the conditions required by the proof framework outlined in Section 4.2.2.

4.2.1. *Full specification of the minimax lower bound.* We begin by providing a more rigorous statement of the minimax lower bound $\mathrm{LB}(\bar{R}, \bar{\sigma}, \delta_n)$. In either Regime A or B, we fix an RKHS $\mathbb{H} = \mathbb{H}_A$ or $\mathbb{H}_B$ and a reward function $r = r_A$ or r_B such that $r \in \mathbb{H}$, and then consider MRPs of the form $\mathscr{I}(\mathcal{P}, r, \gamma)$, of which the value function $\theta^* \in \mathbb{H}$. Throughout our construction, we let $\mu(\mathcal{P})$ be the stationary distribution associated with transition kernel $\mathcal{P}$ and always use notation μ to denote the Lebesgue measure (on state space $\mathcal{X} = [0,1)$). The stationary distribution $\mu(\mathcal{P})$ plays multiple roles. First, the observation pairs $\{(x_i, x_i')\}_{i=1}^n$ are generated by drawing x_i from distribution $\mu(\mathcal{P})$ and then a successor state x_i' from the probability transition $\mathcal{P}$. Note moreover that metric $\|\cdot\|_{\mu}$ in equation ($\mathrm{LB}(\bar{R}, \bar{\sigma}, \delta_n)$) is an abbreviation of the $L^2\big(\mu(\mathcal{P})\big)$-norm, i.e. we measure the estimation error by

$$\|\widehat{\theta} - \theta^*\|_{\mu(\mathcal{P})}^2 := \mathbb{E}_{\mu(\mathcal{P})}\big[\big(\widehat{\theta}(X) - \theta^*(X)\big)^2\big].$$

Finally, the covariance operator $\Sigma_{\mathrm{cov}}(\mathcal{P})$ is induced by distribution $\mu(\mathcal{P})$ (recall equation (8))—i.e., $\Sigma_{\mathrm{cov}}(\mathcal{P}) = \mathbb{E}_{X\sim\mu(\mathcal{P})}[\Phi_X \otimes \Phi_X]$—with Φ_X denoting the representer of evaluation. Below, we denote by $\big\{\big(\mu_j(\mathcal{P}), \phi_j(\mathcal{P})\big)\big\}_{j=1}^{\infty}$ the eigenpairs associated with operator $\Sigma_{\mathrm{cov}}(\mathcal{P})$.

As alluded to above, the lower bounds require precise definitions of the MRP families over which they hold. We define two collections of problem instances $\mathfrak{M}_A(\bar{R}, \bar{\sigma})$ and $\mathfrak{M}_B(\bar{R}, \bar{\sigma})$ that are considered in Regimes A and B respectively. In Regime A, we suppose the reward function r_A is uniformly bounded, i.e. $\|r_A\|_{\infty} \leq 1$, and define a $(\bar{R}, \bar{\sigma})$-valid MRP family

$$\begin{aligned}(44a)\quad \mathfrak{M}_A \equiv \mathfrak{M}_A(\bar{R}, \bar{\sigma}) &\equiv \mathfrak{M}_A\big(\bar{R}, \bar{\sigma}, \{\mu_j\}_{j=1}^{\infty}; r_A, \gamma, \mathbb{H}_A\big)\\ &:= \big\{\text{MRP } \mathscr{I}(\mathcal{P}, r_A, \gamma) \mid \text{(i) The value function } \theta^* \in \mathbb{H}_A \text{ and inequalities (32b) hold.}\\ &\quad \text{(ii) The eigenpairs satisfy } \mu_j(\mathcal{P}) = \mu_j \text{ for } j = 1, 2, \ldots \text{ and } \sup_{j\in\mathbb{Z}_+}\|\phi_j(\mathcal{P})\|_{\infty} \leq 2.\big\}.\end{aligned}$$

In parallel, the $(\bar{R}, \bar{\sigma})$-valid MRP family $\mathfrak{M}_B$ is set as

$$\text{(44b)}\quad \mathfrak{M}_B \equiv \mathfrak{M}_B(\bar{R}, \bar{\sigma}) \equiv \mathfrak{M}_B\big(\bar{R}, \bar{\sigma}, \{\mu_j\}_{j=1}^{\infty}; r_B, \gamma, \mathbb{H}_B\big)$$
$$:= \big\{\text{MRP } \mathscr{I}(\mathcal{P}, r_B, \gamma) \mid \text{(i) } \gamma\|\theta^*\|_{\mu(\mathcal{P})} \le 1.\ \text{(ii) } \theta^* \in \mathbb{H}_B \text{ and inequalities (32b) hold.}$$
$$\text{(iii) The eigenpairs satisfy } \mu_j(\mathcal{P}) \le \mu_j \text{ for any } j \ge 2 \text{ and } \sup_{j\in\mathbb{Z}_+}\|\phi_j(\mathcal{P})\|_\infty \le 2.\big\}.$$

The major differences between MRP families $\mathfrak{M}_A$ and $\mathfrak{M}_B$ are the regularity conditions and the eigenvalue constraints. In Regime A, the reward function r_A is properly normalized so that $\|r_A\|_\infty \le 1$, whereas we impose an upper bound on the value function norm $\gamma\|\theta^*\|_{\mu(\mathcal{P})}$ in Regime B. Additionally, in family $\mathfrak{M}_A$ of Regime A, the eigenvalues are exactly equal to the pre-specified parameters $\{\mu_j\}_{j=1}^{\infty}$. In contrast, we only require the eigenvalues are upper bounded by parameters $\{\mu_j\}_{j=1}^{\infty}$ in family $\mathfrak{M}_B$ of Regime B, so they may be different.

As a point of clarification, the critical radius δ_n in lower bound $\mathrm{LB}(\bar{R}, \bar{\sigma}, \delta_n)$ is defined by the pre-specified constants $\{\mu_j\}_{j=1}^{\infty}$, not the eigenvalues $\{\mu_j(\mathcal{P})\}_{j=1}^{\infty}$. In Regime B in particular, parameter δ_n might be different from $\delta_n(\mathcal{P})$, the radius obtained by plugging eigenvalues $\{\mu_j(\mathcal{P})\}_{j=1}^{\infty}$ into critical inequality (33). However, we still have relation $\delta_n \ge \delta_n(\mathcal{P})$ due to our condition on eigenvalues in definition (44b). See Appendix D.1 of the supplementary material for a proof of this claim. In this case, lower bound $\mathrm{LB}(\bar{R}, \bar{\sigma}, \delta_n)$ further implies $\|\widehat{\theta} - \theta^*\|_{\mu(\mathcal{P})}^2 \ge c_1 \bar{R}^2 \delta_n^2(\mathcal{P})$.

4.2.2. *High-level overview.* We provide a high-level overview of the proof structure that is shared by Regimes A and B. The main argument is based on Fano's method, with the key (and technically challenging) step being the construction of an ensemble of value estimation problems that are "well-separated". While it is well-known how to do so for classical (static) non-parametric problems, doing so for Markov reward processes requires some new ideas.

Here we give the high-level overview, deferring the technical details of the construction itself to the supplementary material. Our approach is to construct a collection $\{\mathscr{I}_m\}_{m=1}^{M}$ of MRP instances, all of which share the same state space $\mathcal{X} = [0, 1)$ and reward function r. Let μ denote the Lebesgue measure over $\mathcal{X}$, and let $\mathcal{P}_m$, θ_m^* and $\mu_m \equiv \mu(\mathcal{P}_m)$ denote (respectively) the transition kernel, value function and stationary distribution associated with $\mathscr{I}_m$. Let $\mathcal{P}_m^{1:n}$ be the distribution of data $\{(x_i, x_i')\}_{i=1}^n$ when the ground-truth model is $\mathscr{I}_m$.

Suppose that an index J is uniformly distributed over $[M]$ and observations $\{(x_i, x_i')\}_{i=1}^n$ are generated i.i.d. from $\mathscr{I}_J$. Given this set-up, an application of Fano's method (cf. §15.3.2 in the book [50] for details) yields the lower bound

$$\inf_{\widehat{\theta}} \max_{m^\dagger \in [M]} \mathbb{P}_{m^\dagger}\left[\big\|\widehat{\theta} - \theta^*_{m^\dagger}\big\|_\mu \ge \frac{1}{2}\min_{m\neq m'}\big\|\theta^*_m - \theta^*_{m'}\big\|_\mu\right]$$
$$\ge 1 - \frac{\log 2 + \max_{m,m'\in[M]} D_{\mathrm{KL}}\big(\mathcal{P}_m^{1:n} \,\big\|\, \mathcal{P}_{m'}^{1:n}\big)}{\log M}.$$

Moreover, suppose that we can also ensure that

$$\frac{d\mu_m}{d\mu}(x) \ge \frac{1}{2} \qquad \text{for all } x \in \mathcal{X} \text{ and } m \in [M]. \tag{45}$$

In this way, we can connect the $L^2(\mu)$ error with the $L^2(\mu_m)$ error via the inequality $\|\widehat{\theta} - \theta^*_m\|_{\mu_m} \ge \frac{1}{\sqrt{2}}\|\widehat{\theta} - \theta^*\|_\mu$. It then follows that

$$\inf_{\widehat{\theta}} \max_{m^\dagger \in [M]} \mathbb{P}_{m^\dagger}\left[\big\|\widehat{\theta} - \theta^*_{m^\dagger}\big\|_{\mu_{m^\dagger}} \ge \frac{1}{2\sqrt{2}}\min_{m\neq m'}\big\|\theta^*_m - \theta^*_{m'}\big\|_\mu\right] \tag{46}$$

$$\geq 1 - \frac{\log 2 + \max_{m,m' \in [M]} D_{\mathrm{KL}}\big(\mathcal{P}_m^{1:n} \,\big\|\, \mathcal{P}_{m'}^{1:n}\big)}{\log M}.$$

Exploiting the Fano inequality so as to obtain a "good" lower bound involves constructing a suitable family of models. Recalling the statistical dimension d_n. In our proof of either Regime A or B, we establish the existence of a family $\{\mathscr{I}_m\}_{m=1}^M$ with log cardinality $\log M \geq \frac{d_n}{10}$, and such that

$$\max_{m,m' \in [M]} D_{\mathrm{KL}}\big(\mathcal{P}_m^{1:n} \,\big\|\, \mathcal{P}_{m'}^{1:n}\big) \leq \frac{d_n}{40} \qquad \text{and} \tag{47a}$$

$$\min_{m \neq m'} \big\|\theta_m^* - \theta_{m'}^*\big\|_{\mu} \geq c_1' \sqrt{c}\, R \delta_n \tag{47b}$$

where c_1' is a universal constant and c is given in condition (34). See Lemmas D.2 and D.3 at the end of Appendix D.2.4 of the supplementary material for the precise statement of these claims.

Given these claims, we can combine the pieces to prove Theorem 3.4. Given the condition (47a) and the bound $\log M \geq \frac{d_n}{10}$, we have $\frac{1}{\log M} \max_{m,m' \in [M]} D_{\mathrm{KL}}\big(\mathcal{P}_m^{1:n} \,\big\|\, \mathcal{P}_{m'}^{1:n}\big) \leq \frac{1}{4}$. Additionally, given that $d_n \geq 10$, it holds that $\log M \geq \frac{d_n}{10} \geq 1$ and therefore $\frac{\log 2}{\log M} \leq \log 2$. Combining these inequalities, we find that the right hand side of inequality (46) is larger than a positive constant $\big\{1 - \frac{1}{4} - \log 2\big\}$. We then substitute the minimum value function distance $\min_{m \neq m'} \big\|\theta_m^* - \theta_{m'}^*\big\|_{\mu}$ in the left hand side of inequality (46) by its lower bound in the inequality (47b). This completes the high-level overview of the proof of Theorem 3.4.

With this perspective in place, the remaining steps—and the technically challenging portion of the argument—should be clear. In particular, the remainder of our argument involves:

- constructing two reproducing kernel Hilbert spaces, denoted by $\mathbb{H}_A$ and $\mathbb{H}_B$, along with two subsets $\{\mathscr{I}_m\}_{m=1}^M$ belonging to either $\mathfrak{M}_A$ or $\mathfrak{M}_B$.
- verifying that both groups of the MRP instances satisfy the claimed properties (45), (47a) and (47b).

We defer the constructions and the proofs of the claims to Appendix D in the supplementary material.

5. Discussion. In this paper, we have analyzed the performance of a regularized kernel-based least-squares temporal difference (LSTD) estimator for policy evaluation. Our main contribution was to prove non-asymptotic upper bounds on the statistical estimation error, along with guidance for the choices of the regularization parameter required to achieve such bounds. Notably, our upper bounds depend on the problem structure via the sample size, the effective horizon, the eigenvalues of the kernel operator, and the variance of the Bellman residual. As we show, the bounds show a wide range of behavior as these different structural components are altered. Moreover, we prove a matching minimax lower bounds over distinct subclasses of problems that demonstrate the sharpness of our upper bounds.

Our study leaves open a number of intriguing questions; let us mention a few of them here to conclude. First, although our bounds are instance-dependent, this dependence is not as refined as recent results in the simpler tabular and linear function settings [26, 32, 56]. In particular, our current results do not explicitly track the mixing properties of the transition kernel, which should enter in any such refined analysis. Second, the analysis of this paper was carried out under i.i.d. assumptions on transition sampling model. However, in practice, the data may be collected from Markov chain trajectories or adaptive experiments and the transition pairs are no longer independent. It is interesting to understand how dependence in data affects sample complexity of policy evaluation, and since the first posting of this paper,

a follow-up paper [14] due to subset of the current authors provides some insight into this question. Third, this paper assumes that samples are drawn from the stationary distribution of the Markov chain; in practice, such data may not be available, so that it is interesting to consider extensions of this kernel LSTD estimator suitable for the off-policy setting. In Appendix F of the supplementary material, we present some extensions of our current results to the off-policy setting. However, it remains open as to whether the optimality and sharpness of our analysis still hold in the off-policy setting, and further research is needed to address this question. Last, the results in this paper use the $L^2(\mu)$-norm to quantify the error. In applications of policy evaluation, other error metrics may be of interest, including pointwise errors ($\left|\widehat{\theta}(x) - \theta^*(x)\right|$ for a fixed state $x \in \mathcal{X}$), or sup-norm guarantees ($\|\widehat{\theta} - \theta^*\|_\infty$). These are interesting directions for future study.

Funding. This work was partially supported by NSF-DMS grant 2015454, NSF-IIS grant 1909365, NSF-FODSI grant 202350, and DOD-ONR Office of Naval Research N00014-21-1-2842 to MJW.

REFERENCES

[1] BAGNELL, J. A. and SCHNEIDER, J. (2003). Policy search in kernel Hilbert space Technical Report, Carnegie Mellon University.
[2] BAREETO, A. M. S., PRECUP, D. and PINEAU, J. (2016). Practical kernel-based reinforcement learning. *Journal of Machine Learning Research* **17** 1–70.
[3] BELLMAN, R. and DREFUS, S. (1962). *Applied dynamic programming*. Princeton University Press, Princeton, NJ.
[4] BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, Norwell, MA.
[5] BERTSEKAS, D. P. (1995). *Dynamic programming and optimal control* **1**. Athena scientific Belmont, MA.
[6] BERTSEKAS, D. P. (2011). Dynamic programming and optimal control 3rd edition, volume II. *Belmont, MA: Athena Scientific*.
[7] BERTSEKAS, D. P. and TSITSIKLIS, J. N. (1996). *Neuro-Dynamic Programming*, 1st ed. Athena Scientific.
[8] BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Stat.* **33** 719–726.
[9] BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Stat.* **36** 226–235.
[10] BOUCHERIE, R. J. and VAN DIJK, N. M. (2017). *Markov decision processes in practice*. Springer, New York.
[11] BRADTKE, S. J. and BARTO, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine learning* **22** 33–57.
[12] CHEN, X. and REISS, M. (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory* **27** 497–521.
[13] DAI, B., HE, N., PAN, Y., BOOTS, B. and SONG, L. (2017). Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics* 1458–1467. PMLR.
[14] DUAN, Y. and WAINWRIGHT, M. J. (2022). Policy evaluation from a single path: Multi-step methods, mixing and mis-specification Technical Report, MIT. arXiv:2211.03899.
[15] FAN, J., WANG, Z., XIE, Y. and YANG, Z. (2020). A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control* 486–489. PMLR.
[16] FARAHMAND, A.-M., GHAVAMZADEH, M., SZEPESVÁRI, C. and MANNOR, S. (2016). Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research* **17** 4809–4874.
[17] FENG, Y., LI, L. and LIU, Q. (2019). A kernel loss for solving the Bellman equation. In *Advances in Neural Information Processing Systems* 15456–15467.
[18] FENG, Y., REN, T., TANG, Z. and LIU, Q. (2020). Accountable off-policy evaluation with kernel Bellman statistics. In *International Conference on Machine Learning* 3102–3111. PMLR.
[19] GHESHLAGHI AZAR, M., MUNOS, R. and KAPPEN, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning* **91** 325–349.
[20] GRUNEWALDER, S., LEVER, G., BALDASSARRE, L., PONTIL, M. and GRETTON, A. (2012). Modelling transition dynamics in MDPs with RKHS embeddings Technical Report, UCL.
[21] GU, C. (2002). *Smoothing spline ANOVA models*. *Springer Series in Statistics*. Springer, New York, NY.

[22] HASMINSKII, R. Z. (1978). A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Prob. Appl.* **23** 794–798.
[23] JIANG, N. and LI, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning* 652–661. PMLR.
[24] KALLUS, N. and UEHARA, M. (2019). Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *stat* **1050** 12.
[25] KALLUS, N. and UEHARA, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research* **21** 1–63.
[26] KHAMARU, K., PANANJADY, A., RUAN, F., WAINWRIGHT, M. J. and JORDAN, M. I. (2021). Is Temporal Difference Learning Optimal? An Instance-Dependent Analysis. *SIAM J. Math. Data Science* **3** 1013–1040.
[27] KIMELDORF, G. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *Jour. Math. Anal. Appl.* **33** 82–95.
[28] KOPPEL, A., WARNELL, G., STUMP, E., STONE, P. and RIBEIRO, A. (2020). Policy evaluation in continuous MDPs with efficient kernelized gradient temporal difference. *IEEE Transactions on Automatic Control*.
[29] LONG, J., HAN, J. and E, W. (2021). An L^2 Analysis of Reinforcement Learning in High Dimensions with Kernel and Neural Network Approximation. *arXiv preprint arXiv:2104.07794*.
[30] MENDELSON, S. (2002). Geometric Parameters of Kernel Machines. In *Computational Learning Theory* 29–43. Springer Berlin Heidelberg.
[31] MOU, W., LI, C. J., WAINWRIGHT, M. J., BARTLETT, P. L. and JORDAN, M. I. (2020). On Linear Stochastic Approximation: Fine-grained Polyak-Ruppert and Non-Asymptotic Concentration. In *Conference on Learning Theory (COLT)* **125** 2947–2997.
[32] MOU, W., PANANJADY, A. and WAINWRIGHT, M. J. (2023). Optimal oracle inequalities for solving projected fixed-point equations. *Mathematics of Operations Research* **48** 2308–2336.
[33] MUNOS, R. and SZEPESVARI, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research* **1** 815–857.
[34] NEWEY, W. K. and POWELL, J. L. (2003). Instrumental variable estimation of non-parametric models. *Econometrica* **71** 1565–1578.
[35] NGUYEN-TANG, T., GUPTA, S., TRAN-THE, H., VENKATESH, S. et al. (2021). Sample Complexity of Offline Reinforcement Learning with Deep ReLU Networks. *arXiv preprint arXiv:2103.06671*.
[36] O'LEARY, D. and STEWART, G. (1990). Computing the eigenvalues and eigenvectors of symmetric arrowhead matrices. *Journal of Computational Physics* **90** 497–505.
[37] ORMONEIT, D. and SEN, Ś. (2002). Kernel-based reinforcement learning. *Machine learning* **49** 161–178.
[38] PANANJADY, A. and WAINWRIGHT, M. J. (2020). Instance-Dependent ℓ_∞-Bounds for Policy Evaluation in Tabular Reinforcement Learning. *IEEE Transactions on Information Theory* **67** 566–585.
[39] PERDOMO, J. C., KRISHNAMURTHY, A., BARTLETT, P. and KAKADE, S. (2022). A sharp characterization of linear estimators for offline policy evaluation. *arXiv preprint arXiv:2203.04236*.
[40] PUTERMAN, M. L. (2005). *Markov decision processes: Discrete stochastic dynamic programming*. Wiley.
[41] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research* **12** 389–427.
[42] SHAWE-TAYLOR, J., CRISTIANINI, N. et al. (2004). *Kernel methods for pattern analysis*. Cambridge university press.
[43] SOBEL, M. J. (1982). The variance of discounted Markov decision processes. *Journal of Applied Probability* **19** 794–802.
[44] STONE, C. J. (1982). Optimal global rates of convergence for non-parametric regression. *Annals of Statistics* **10** 1040–1053.
[45] SUTTON, R. S. (1988). Learning to predict via the methods of temporal differences. *Machine Learning* **3** 9–44.
[46] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
[47] TAYLOR, G. and PARR, R. (2009). Kernelized value function approximation for reinforcement learning. In *Proceedings of the 26th annual international conference on machine learning* 1017–1024.
[48] TSITSIKLIS, J. N. and VAN ROY, B. (1997). Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems* 1075–1081.
[49] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.
[50] WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge University Press.
[51] WHITE, H. (1982). Instrumental variables regression with independent observations. *Econometrica* **50** 483–499.

[52] WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.

[53] XIE, T., MA, Y. and WANG, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems* 9668–9678.

[54] YANG, Y., PILANCI, M., WAINWRIGHT, M. J. et al. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics* **45** 991–1023.

[55] YIN, M. and WANG, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. *arXiv preprint arXiv:2001.10742*.

[56] YIN, M. and WANG, Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems* **34** 4065–4078.

[57] YU, B. (1997). Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer-Verlag, Berlin.

[58] YU, H. and BERTSEKAS, D. P. (2010). Error bounds for approximations from projected linear equations. *Mathematics of Operations Research* **35** 306–329.

[59] ZHANG, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Computation* **17** 2077–2098.

SUPPLEMENTARY TO "OPTIMAL POLICY EVALUATION USING KERNEL-BASED TEMPORAL DIFFERENCE METHODS"

BY YAQI DUAN[1,a], MENGDI WANG[2,b] AND MARTIN J. WAINWRIGHT[3,c]

[1]*Leonard N. Stern School of Business, New York University,* [a]*yaqi.duan@stern.nyu.edu*

[2]*Department of ECE, Princeton University,* [b]*mengdiw@princeton.edu*

[3]*Departments of EECS and Mathematics, Massachusetts Institute of Technology,* [c]*wainwrigwork@gmail.com*

APPENDIX A: DETAILS OF SIMULATIONS

In this appendix, we provide the details of the families of MRPs used for the simulation results in Section 3.3.

A.1. Families of MRPs. We constructed families of MRPs all with state space $\mathcal{X} = [0, 1)$, and the reward function

$$r(x) := \mathbb{1}\{x \in [0, \tfrac{1}{2})\} - \mathbb{1}\{x \in [\tfrac{1}{2}, 1)\}. \tag{48a}$$

The transition operator is given by

$$\mathcal{P}(x' \mid x) := \begin{cases} 2(1-p), & \text{if } x, x' \in [0, \frac{1}{2}) \text{ or } x, x' \in [\frac{1}{2}, 1), \\ 2p, & \text{if } \begin{cases} x \in [0, \frac{1}{2}) \\ x' \in [\frac{1}{2}, 1) \end{cases} \text{ or } \begin{cases} x \in [\frac{1}{2}, 1) \\ x' \in [0, \frac{1}{2}). \end{cases} \end{cases} \tag{48b}$$

See Figure 4 for an illustration of the structure of this transition function. By construction,

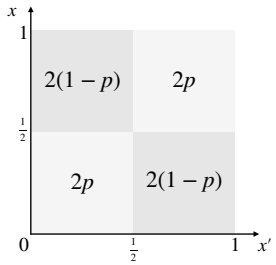


Fig 4: The density of data $\{(x_i, x_i')\}_{i=1}^n$.

the uniform distribution μ is stationary. The samples $\{(x_i, x_i')\}_{i=1}^n$ were i.i.d. drawn from the pair $(\mu, \mathcal{P})$. The two ensembles of probability transitions (Ensembles A and B) used in our simulations are distinguished by the choice $p \in \{\frac{1}{4}, \frac{1-\gamma}{\gamma}\}$.

In addition to the two ensembles of transition functions, our experiments involve comparisons between three different kernels, all of which were constructed based on the Walsh system. Let $W_j : [0, 1) \to \{-1, 1\}$ be the j-th Walsh function. For each $i = 1, 2, 3$, we define a kernel $\mathcal{K}_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as

$$\mathcal{K}_i(x, y) := \sum_{j=1}^{\infty} \mu_j(\mathcal{K}_i)\, W_{j-1}(x) W_{j-1}(y). \tag{49a}$$

This choice ensures that each $\mathcal{K}_i$ has $\{W_{j-1}\}_{j=1}^{\infty}$ as its eigenfunctions. We choose the associated kernel eigenvalues as

$$\mu_j(\mathcal{K}_i) = \begin{cases} j^{-6/5} & \text{for } i = 1 \\ j^{-2} & \text{for } i = 2 \\ \exp\big(-(j-1)^2\big) & \text{for } i = 3. \end{cases} \tag{49b}$$

Let $\mathbb{H}_i$ be the RKHS associated with kernel $\mathcal{K}_i$.

Calculations of predicted slopes. Let us now calculate the theoretically predicted slopes given in equations (31a) and (31b). Note that Corollary 3.3(b)—see in particular the bound (26)—predicts that the $L^2(\mu)$ error should scale as

$$R^{\frac{1}{2\alpha+1}} \Big(\frac{\kappa^2 \sigma^2(\theta^*)}{(1-\gamma)^2} \frac{1}{n} \Big)^{\frac{\alpha}{2\alpha+1}}. \tag{50}$$

Since $\kappa = 1$ for our construction, in order to understand the scaling with the effective horizon, we need to calculate the quantities $\sigma^2(\theta^*)$ and R. Some calculations show that the value function θ^* is given by

$$\theta^*(x) = \tfrac{1}{1-\gamma+2\gamma p}\big(\mathbb{1}\{x \in [0, \tfrac{1}{2})\} - \mathbb{1}\{x \in [\tfrac{1}{2}, 1)\}\big). \tag{51}$$

Consequently, we can see that $r, \theta^* \in \mathbb{H}_i$ for $i = 1, 2, 3$.

Moreover, we find that variance term $\sigma^2(\theta^*)$ takes the form

$$\sigma^2(\theta^*) = \frac{4\gamma^2 p(1-p)}{(1-\gamma+2\gamma p)^2}. \tag{52a}$$

For each RKHS $\mathbb{H}_i$, the radius $R(\mathcal{K}_i)$ is given by

$$R(\mathcal{K}_i) := \max\big\{\|\theta^* - r\|_{\mathbb{H}_i}, \tfrac{2\|\theta^*\|_\infty}{b(\mathcal{K}_i)}\big\} = \tfrac{1}{1-\gamma+2\gamma p} \max\Big\{\tfrac{\gamma(1-2p)}{\sqrt{\mu_1(\mathcal{K}_i)}}, \tfrac{2}{b(\mathcal{K}_i)}\Big\} \tag{52b}$$

with $b(\mathcal{K}_i) = \sqrt{\sum_j \mu_j(\mathcal{K}_i)}$.

For the choice $p = \frac{1-\gamma}{\gamma}$, it can be seen that both $\sigma^2(\theta^*)$ and $R(\mathcal{K}_i)$ scale as $\frac{1}{1-\gamma}$. Substituting these scalings into equation (50) (and retaining only the dependence on the effective horizon) yields

$$\Big(\frac{1}{1-\gamma}\Big)^{\frac{1}{2\alpha+1}} \Big(\frac{1}{(1-\gamma)^3}\Big)^{\frac{\alpha}{2\alpha+1}} = \Big(\frac{1}{1-\gamma}\Big)^{\frac{3\alpha+1}{2\alpha+1}},$$

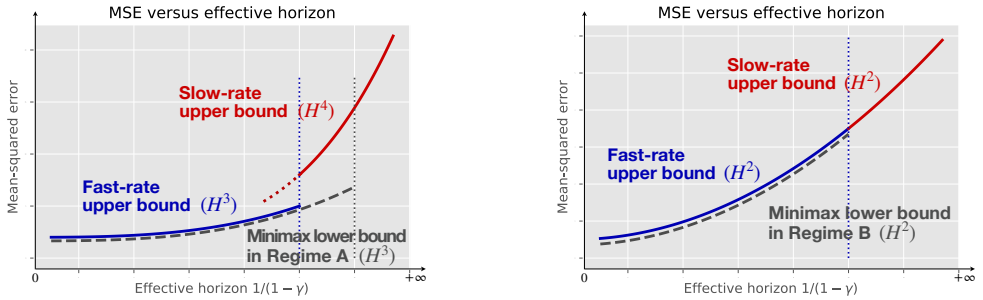
as claimed in equation (31a).

For the choice $p = \frac{1}{4}$, both $\sigma^2(\theta^*)$ and $R(\mathcal{K}_i)$ remain bounded as the effective horizon grows, so that the corresponding slope is $\Big(\frac{1}{(1-\gamma)^2}\Big)^{\frac{\alpha}{2\alpha+1}} = \Big(\frac{1}{1-\gamma}\Big)^{\frac{2\alpha}{2\alpha+1}}$, as claimed in equation (31b).

APPENDIX B: COMPARISON OF THE FAST VERSUS SLOW RATES

In this appendix, we compare how, how the different bounds in our paper depend on the effective horizon $H = (1-\gamma)^{-1}$. Specifically, we compare the slow and fast-rate upper bounds from Theorem 3.1 with the minimax lower bound from Theorem 3.4 in the context of parametric function approximation (cf. Section 3.2.1).

We can see that the bounds exhibit distinct behaviors for different regimes characterized by the scalings of R and $\sigma(\theta^*)$. In general, when certain sample size conditions are satisfied, the fast-rate upper bound and the minimax lower bound align with each other. The slow-rate upper bound, although generally looser than the fast-rate bound, provides coverage across the entire range of parameters, making it applicable for large planning horizons H or small sample sizes n.



(a) Regime A (b) Regime B

Fig 5. Illustration of the bounds on the estimation error $\|\widehat{\theta}-\theta^*\|_{\mu}^2$ v.s. horizon $H=(1-\gamma)^{-1}$. In both regimes, the red curves represent slow-rate upper bounds, while the blue curves represent fast-rate upper bounds. The grey curves correspond to minimax lower bounds. The dashed vertical lines indicate conditions (18) and (36). (a) Plot for Regime A with $R \asymp \sigma^2(\theta^*) \asymp (1-\gamma)^{-1}$. The fast-rate bound is better than the slow-rate bound. However, the slow-rate bound remains feasible for the entire range of planning horizon H. The minimax lower bound matches the fast-rate upper bound. (b) Plot for Regime B with $R \asymp \sigma^2(\theta^*) \asymp 1$. Both the fast and slow rate upper bounds exhibit the same rate and match the minimax lower bound.

APPENDIX C: TECHNICAL RESULTS FOR THEOREM 3.1

In this appendix, we prove the technical lemmas that underlie the proof of Theorem 3.1. Appendix C.1 contains the proof of Lemma 4.1, which provides the basic inequality on the error. Appendices C.2 and C.3 are devoted, respectively, to the proof of Lemmas 4.2 and 4.3 that are used in the proof of Theorem 3.1. Recall that these two lemmas provide high-probability upper bounds on the quantities T_1 and T_3, respectively, as defined in Lemma 4.1.

C.1. Proof of Lemma 4.1. Recall that $\widehat{\theta}$ is defined by the estimating equation (10), whereas the actual population-level estimate θ^* satisfies equation (9). Subtracting these two equations yields

$$\big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I}\big)\,\widehat{\theta} - \Sigma_{\mathrm{cov}}\theta^* = \big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I} - \Sigma_{\mathrm{cov}}\big) r + \gamma\big(\widehat{\Sigma}_{\mathrm{cr}}\widehat{\theta} - \Sigma_{\mathrm{cr}}\theta^*\big).$$

Recalling that $\widehat{\Delta} = \widehat{\theta} - \theta^*$ is the error, we substitute $\widehat{\theta} = \theta^* + \widehat{\Delta}$ to find that

$$\big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I} - \gamma\widehat{\Sigma}_{\mathrm{cr}}\big)\widehat{\Delta} = \big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I} - \Sigma_{\mathrm{cov}}\big)(r-\theta^*) + \gamma\big(\widehat{\Sigma}_{\mathrm{cr}} - \Sigma_{\mathrm{cr}}\big)\theta^*.$$

Again making use of equation (9), we have $\Sigma_{\mathrm{cov}}(r-\theta^*) + \gamma\Sigma_{\mathrm{cr}}\theta^* = 0$, which implies that

$$\big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I} - \gamma\widehat{\Sigma}_{\mathrm{cr}}\big)\widehat{\Delta} = \big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I}\big)(r-\theta^*) + \gamma\widehat{\Sigma}_{\mathrm{cr}}\theta^*.$$

Taking the Hilbert inner product of both sides with $\widehat{\Delta}$ then yields

$$\big\langle \widehat{\Delta},\, \big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I} - \gamma\widehat{\Sigma}_{\mathrm{cr}}\big)\widehat{\Delta}\big\rangle_{\mathbb{H}} = \big\langle \widehat{\Delta},\, \big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I}\big)(r-\theta^*) + \gamma\widehat{\Sigma}_{\mathrm{cr}}\theta^*\big\rangle_{\mathbb{H}}. \tag{53}$$

The left hand side of equation (53) can then be written as

$$\big\langle \widehat{\Delta},\, \big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I} - \gamma\widehat{\Sigma}_{\mathrm{cr}}\big)\widehat{\Delta}\big\rangle_{\mathbb{H}} = \big\langle \widehat{\Delta},\, (\Sigma_{\mathrm{cov}} - \gamma\Sigma_{\mathrm{cr}})\widehat{\Delta}\big\rangle_{\mathbb{H}} + \lambda_n\|\widehat{\Delta}\|_{\mathbb{H}}^2 + \big\langle \widehat{\Delta},\, (\widehat{\Gamma} - \Gamma)\widehat{\Delta}\big\rangle_{\mathbb{H}},$$

where $\Gamma := \Sigma_{\mathrm{cov}} - \gamma\Sigma_{\mathrm{cr}}$, and $\widehat{\Gamma} := \widehat{\Sigma}_{\mathrm{cov}} - \gamma\widehat{\Sigma}_{\mathrm{cr}}$. The right hand side of equation (53) satisfies

$$\big\langle \widehat{\Delta},\, \big(\widehat{\Sigma}_{\mathrm{cov}} + \lambda_n \mathcal{I}\big)(r-\theta^*) + \gamma\widehat{\Sigma}_{\mathrm{cr}}\theta^*\big\rangle_{\mathbb{H}} = \big\langle \widehat{\Delta},\, \widehat{\Sigma}_{\mathrm{cov}}(r-\theta^*) + \gamma\widehat{\Sigma}_{\mathrm{cr}}\theta^*\big\rangle_{\mathbb{H}} + \lambda_n\big\langle \widehat{\Delta},\, r-\theta^*\big\rangle_{\mathbb{H}}.$$

In this way, we reduce equation (53) to

$$\rho^2(\widehat{\Delta}) \;=\; \big\langle \widehat{\Delta},\, (\Sigma_{\mathrm{cov}} - \gamma\Sigma_{\mathrm{cr}})\widehat{\Delta}\big\rangle_{\mathbb{H}} = \big\langle \widehat{\Delta},\, \widehat{\Sigma}_{\mathrm{cov}}(r-\theta^*) + \gamma\widehat{\Sigma}_{\mathrm{cr}}\theta^*\big\rangle_{\mathbb{H}} + \lambda_n\big\langle \widehat{\Delta},\, r-\theta^*\big\rangle_{\mathbb{H}} + \big\langle \widehat{\Delta},\, (\Gamma-\widehat{\Gamma})\widehat{\Delta}\big\rangle_{\mathbb{H}} - \lambda_n\|\widehat{\Delta}\|_{\mathbb{H}}^2. \tag{54}$$

We have thus established equality (ii) in equation (39) from the lemma statement.

It remains to prove the lower bound (i) in equation (39). Letting $X \sim \mu$ and $X' \sim \mathcal{P}(\cdot \mid X)$, we can write

$$\mathbb{E}[f(X)f(X')] = \big\langle f,\, \Sigma_{\mathrm{cr}} f\big\rangle_{\mathbb{H}} \quad \text{and} \quad \mathbb{E}[f^2(X)] = \big\langle f,\, \Sigma_{\mathrm{cov}} f\big\rangle_{\mathbb{H}}.$$

Consequently, by applying Young's inequality, we find that

$$\underbrace{\mathbb{E}[f(X)f(X')]}_{\langle f, \Sigma_{\mathrm{cr}} f\rangle_{\mathbb{H}}} \leq \frac{1}{2}\Big\{\mathbb{E}[f^2(X)] + \mathbb{E}[f^2(X')]\Big\} \;=\; \underbrace{\mathbb{E}[f^2(X)]}_{\langle f, \Sigma_{\mathrm{cov}} f\rangle_{\mathbb{H}}}, \tag{55}$$

where the equality follows since X and X' have the same marginal distributions, due to the stationarity of μ. This completes the proof of Lemma 4.1.

C.2. Proof of Lemma 4.2. Define the i.i.d. random variables $\nu_i = r(x_i) - \theta^*(x_i) + \gamma\theta^*(x_i')$. Since $\Sigma_{\mathrm{cov}}\theta^* = \Sigma_{\mathrm{cov}} r + \gamma\Sigma_{\mathrm{cr}}\theta^*$, we can write T_1 as

$$T_1 = \frac{1}{n}\sum_{i=1}^n \Big(\widehat{\Delta}(x_i)\nu_i - \mathbb{E}[\widehat{\Delta}(x_i)\nu_i]\Big).$$

For scalars $t > 0$, we define the family of random variables

$$Z_n(t) := \sup_{\substack{\|f\|_\mu \leq t \\ \|f\|_{\mathbb{H}} \leq R}} \Big|\frac{1}{n}\sum_{i=1}^n \big(f(x_i)\nu_i - \mathbb{E}[f(x_i)\nu_i]\big)\Big|, \tag{56}$$

and let $t_n > 0$ be the smallest positive solution to the inequality

$$\mathbb{E}[Z_n(t)] \leq (1-\gamma)\,\frac{t^2}{4}.$$

Our first step is to relate t_n to the critical radius δ_n involved in Theorem 3.1.

LEMMA C.1. *There is a universal constant c_0 such that, for any $\zeta \in \{bR, \kappa\sigma(\theta^*)\}$, we have*

$$t_n \leq u_n(\zeta) := c_0\, R\, \delta_n(\zeta). \tag{57}$$

The remainder of the proof applies to both $u_n(bR)$ or $u_n(\kappa\sigma(\theta^*))$ without any differences, so we adopt the generic notation u_n for either. Our next step is to use u_n to define an event that allows us to establish the claim of Lemma 4.2. For a given $f \in \mathbb{H}$, we say that inequality $I(f)$ holds when

$$(I(f)) \quad \Big|\frac{1}{n}\sum_{i=1}^n \big(f(x_i)\nu_i - \mathbb{E}[f(x_i)\nu_i]\big)\Big| \;\geq\; (1-\gamma)\,u_n^2 \max\Big\{1, \tfrac{\|f\|_{\mathbb{H}}}{R}\Big\} + (1-\gamma)\,u_n\,\|f\|_\mu\,.$$

Here $c > 0$ is a universal constant to be specified as part of the proof. Now consider the event $\mathcal{A} := \big\{\exists\, f \in \mathbb{H} \text{ s.t } I(f) \text{ holds}\big\}$. Note that conditioned on $\mathcal{A}^c$, we have the bound

$$\begin{aligned} T_1 &\leq (1-\gamma)\,u_n^2 \max\Big\{1, \tfrac{\|\widehat{\Delta}\|_{\mathbb{H}}}{R}\Big\} + (1-\gamma)\,u_n\,\|\widehat{\Delta}\|_\mu \\ &\leq c_0^2\,(1-\gamma)\,\delta_n^2\,\Big\{\|\widehat{\Delta}\|_{\mathbb{H}}^2 + R^2\Big\} + c_0\,(1-\gamma)\,R\,\|\widehat{\Delta}\|_\mu \delta_n, \end{aligned}$$

as desired.

Consequently, the remainder of our proof is directed at bounding $\mathbb{P}[\mathcal{A}]$. We do so by relating the event $\mathcal{A}$ to a tail event associated with the random variable $Z_n(u_n)$. In particular, we make the following claim:

LEMMA C.2. *We have the upper bound*

$$\mathbb{P}[\mathcal{A}] \leq \mathbb{P}\big[Z_n(u_n) \geq (1-\gamma)\, u_n^2\big]. \tag{58}$$

Our final lemma provides control on the upper tail of $Z_n(u_n)$.

LEMMA C.3. *There is a universal constant c_1 such that*

$$\mathbb{P}\big[Z_n(u_n) \geq (1-\gamma)\, u_n^2\big] \leq \exp\big(-c_1\, n\, \tfrac{u_n^2(1-\gamma)^2}{b^2R^2}\big) = \exp\big(-c_1 c_0^2\, \tfrac{n\delta_n^2(1-\gamma)^2}{b^2}\big). \tag{59}$$

Combining Lemmas C.2 and C.3 yields the conclusion of Lemma 4.2 with $c' = c_1 c_0^2$.

It remains to prove our three auxiliary lemmas, and we prove Lemmas C.1, C.2, and C.3 in Appendices C.2.1 to C.2.3, respectively.

C.2.1. *Proof of Lemma C.1.* By definition of t_n, we have $(1-\gamma)\,\frac{t_n^2}{4} = \mathbb{E}[Z_n(t_n)]$. Consequently, we can prove the claim by upper bounding the expectation. Let $\{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Rademacher variables, independent of $\{(x_i, x_i')\}_{i=1}^n$. From a standard symmetrization argument, we have

$$\mathbb{E}[Z_n(t_n)] \leq 2\,\mathbb{E}\left[\sup_{\substack{\|f\|_{\mu} \leq t_n \\ \|f\|_{\mathbb{H}} \leq R}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\nu_i\Big|\right].$$

Proof for $\delta_n(bR)$. We begin by proving the claim when $\zeta = bR$. Note that for any (x_i, x_i'), we have

$$|\nu_i| = \big|r(x_i) - \theta^*(x_i) + \gamma\theta^*(x_i')\big| \leq b\|\theta^* - r\|_{\mathbb{H}} + \|\theta^*\|_\infty \leq 2bR,$$

using the definition of R. Consequently, by the Ledoux-Talagrand contraction, we have

$$\begin{aligned}
(1-\gamma)\,\frac{t_n^2}{4} = \mathbb{E}[Z_n(t_n)] &\leq 4bR\,\mathbb{E}\left[\sup_{\substack{\|f\|_{\mu} \leq t_n \\ \|f\|_{\mathbb{H}} \leq R}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\right] \\
&\overset{(i)}{=} 4bR^2\,\mathbb{E}\left[\sup_{\substack{\|g\|_{\mu} \leq t_n/R \\ \|g\|_{\mathbb{H}} \leq 1}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)\Big|\right] \\
&\overset{(ii)}{\leq} 4bR^2\,\frac{R(1-\gamma)}{bR}\left\{\delta_n^2 + \frac{t_n\delta_n}{R}\right\} \\
&= 4R^2\,(1-\gamma)\left\{\delta_n^2 + \frac{t_n\delta_n}{R}\right\}
\end{aligned}$$

where equality (i) follows by reparameterizing the supremum in terms of the rescaled functions $g = f/R$, and inequality (ii) follows from the definition of $\delta_n(bR)$. This implies that there is a universal constant c_0 such that $t_n \leq c_0\, R\, \delta_n$, as claimed.

Proof for $\delta_n(\kappa\sigma(\theta^*))$. In this case, we begin by observing that

$$\mathbb{E}[Z_n(t_n)] \leq 2\sigma(\theta^*)\,\mathbb{E}\left[\sup_{\substack{\|f\|_{\mu}\leq t_n \\ \|f\|_{\mathbb{H}}\leq R}} \Big|\frac{1}{n}\sum_{i=1}^n f(x_i)\xi_i\Big|\right]$$

where the variables $\xi_i = \frac{\varepsilon_i \nu(x_i,x_i')}{\sigma(\theta^*)}$ have zero mean and unit variance.

We now reparameterize the supremum in terms of the rescaled functions $g = f/R$, so that $\|g\|_{\mathbb{H}} \leq 1$ and $\|g\|_{\mu} \leq t_n/R$. In this way, we find that

$$(1-\gamma)\,\frac{t_n^2}{4} \;=\; \mathbb{E}[Z_n(t_n)] \;\leq\; \big(2\sigma(\theta^*)R\big)\,\mathbb{E}\left[\sup_{\substack{\|g\|_{\mu}\leq t_n/R \\ \|g\|_{\mathbb{H}}\leq 1}} \Big|\frac{1}{n}\sum_{i=1}^n g(x_i)\xi_i\Big|\right].$$

Recall that $g \in \mathbb{H}$ can be written in the form $g = \sum_{j=1}^{\infty} g_j\phi_j$ for some coefficients $\{g_j\}_{j=1}^{\infty}$ such that $\|g\|_{\mu}^2 = \sum_{j=1}^{\infty} g_j^2$, and $\|g\|_{\mathbb{H}}^2 = \sum_{j=1}^{\infty} \frac{g_j^2}{\mu_j}$. Consequently, for any g involved in the supremum, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{\substack{\|g\|_{\mu}\leq t_n/R \\ \|g\|_{\mathbb{H}}\leq 1}} \Big|\frac{1}{n}\sum_{i=1}^n g(x_i)\xi_i\Big|\right] &= \mathbb{E}\left[\sup_{\substack{\|g\|_{\mu}\leq t_n/R \\ \|g\|_{\mathbb{H}}\leq 1}} \Big|\sum_{j=1}^{\infty} g_j\Big(\frac{1}{n}\sum_{i=1}^n \xi_i\phi_j(x_i)\Big)\Big|\right] \\
&\leq \mathbb{E}\left[\Big\{2\sum_{j=1}^{\infty}\min\big\{\tfrac{t_n^2}{R^2},\mu_j\big\}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i\phi_j(x_i)\Big)^2\Big\}^{1/2}\right] \\
&\leq \sqrt{\frac{2}{n}\sum_{j=1}^{\infty}\min\big\{\tfrac{t_n^2}{R^2},\mu_j\big\}\mathbb{E}\big[\xi^2\phi_j^2(X)\big]}\,.
\end{aligned}$$

Since $\mathbb{E}[\xi^2] = 1$ and $\phi_j^2(X) \leq \kappa^2$ by assumption, we have $\mathbb{E}\big[\xi^2\phi_j^2(X)\big] \leq \kappa^2$. Thus, we have established that

$$\mathbb{E}\left[\sup_{\substack{\|g\|_{\mu}\leq t_n/R \\ \|g\|_{\mathbb{H}}\leq 1}} \Big|\frac{1}{n}\sum_{i=1}^n g(x_i)\xi_i\Big|\right] \leq \kappa\sqrt{\frac{2}{n}\sum_{j=1}^{\infty}\min\big\{\tfrac{t_n^2}{R^2},\mu_j\big\}} \;\leq\; \sqrt{2}\kappa\,\frac{R(1-\gamma)}{\kappa\sigma(\theta^*)}\Big\{\delta_n^2 + \frac{\delta_n t_n}{R}\Big\},$$

where the final inequality follows from the definition of $\delta_n = \delta_n(\kappa\sigma(\theta^*))$.

Putting together all the pieces, we have

$$\begin{aligned}
(1-\gamma)\,\frac{t_n^2}{4} \leq \mathbb{E}[Z_n(t_n)] &\leq \big(2\sigma(\theta^*)R\big)\,\frac{\sqrt{2}\,R\,(1-\gamma)}{\sigma(\theta^*)}\Big\{\delta_n^2 + \frac{\delta_n t_n}{R}\Big\} \\
&= 2\sqrt{2}\,R^2\,(1-\gamma)\Big\{\delta_n^2 + \frac{\delta_n t_n}{R}\Big\}.
\end{aligned}$$

This implies that there is a universal constant c_0 such that $t_n \leq c_0\,R\,\delta_n$, as claimed.

C.2.2. *Proof of Lemma C.2.* First, we claim that if $I(f)$ holds for any function, then we can find a function g with $\|g\|_{\mathbb{H}} \leq R$ such that

$$\Big|\frac{1}{n}\sum_{i=1}^n g(x_i)\nu_i\Big| \;\geq\; (1-\gamma)\big\{u_n^2 + u_n\,\|g\|_{\mu}\big\}. \tag{60}$$

Indeed, if $\|f\|_{\mathbb{H}} \leq R$, then we are done. Otherwise, we define the rescaled function $g = \frac{R}{\|f\|_{\mathbb{H}}} f$, and note that it also belongs to the Hilbert space, and satisfies $\|g\|_{\mathbb{H}} = R$. Moreover, since f satisfies $I(f)$, we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i)\nu_i\right| &= \frac{R}{\|f\|_{\mathbb{H}}}\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i)\nu_i\right| \\ &\geq \frac{R}{\|f\|_{\mathbb{H}}}\left\{(1-\gamma)\,u_n^2 \max\left\{1, \tfrac{\|f\|_{\mathbb{H}}}{R}\right\} + (1-\gamma)\,u_n\,\|f\|_{\mu}\right\} \\ &= (1-\gamma)\left\{u_n^2 + u_n\,\|g\|_{\mu}\right\}.\end{aligned}$$

Next, we claim that we can also find a function h such that, in addition, satisfies the bound $\|h\|_{\mu} \leq u_n$ and

$$\left|\frac{1}{n}\sum_{i=1}^{n} h(x_i)\nu_i\right| \geq (1-\gamma)\,u_n^2. \tag{61}$$

If the function g constructed above satisfies $\|g\|_{\mu} \leq u_n$, then we are done. Otherwise, we set $h = \frac{u_n}{\|g\|_{\mu}} g$. Note that $h \in \mathbb{H}$ satisfies $\|h\|_{\mathbb{H}} \leq \|g\|_{\mathbb{H}} = R$ and $\|h\|_{\mu} = u_n$. Moreover, since g satisfies inequality (60), we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} h(x_i)\nu_i\right| = \frac{u_n}{\|g\|_{\mu}}\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i)\nu_i\right| &\geq \frac{u_n}{\|g\|_{\mu}}(1-\gamma)\max\left\{u_n^2 + u_n\,\|g\|_{\mu}\right\} \\ &\geq (1-\gamma)\,u_n^2.\end{aligned}$$

Consequently, we have shown that if the event $\mathcal{A}$ holds, then we can find a function h such that $\|h\|_{\mathbb{H}} \leq R$ and $\|h\|_{\mu} \leq u_n$, and such that the lower bound (61) holds. The existence of this h implies that $Z_n(u_n) \geq (1-\gamma)\,u_n^2$, which shows that $\mathcal{A} \subset \{Z_n(u_n) \geq (1-\gamma)\,u_n^2\}$, as claimed.

C.2.3. *Proof of Lemma C.3.* By definition of u_n from Lemma C.1, we have $u_n \geq t_n$, and hence

$$\mathbb{E}[Z_n(u_n)] \leq (1-\gamma)\,u_n\,\frac{t_n}{4} \leq (1-\gamma)\,\frac{u_n^2}{4},$$

using the definition of t_n. Our next step is to prove that there is a universal constant c_1 such that

$$\mathbb{P}\Big[Z_n(u_n) \geq 2\,\mathbb{E}[Z_n(u_n)] + (1-\gamma)\,\tfrac{u_n^2}{2}\Big] \leq \exp\big(-c_1 n\,\tfrac{u_n^2(1-\gamma)^2}{b^2R^2}\big). \tag{62}$$

The statement given in the lemma follows by combining these two claims.

It remains to prove the tail bound (62). By definition, the random variable $Z_n(t_n)$ corresponds to the supremum of an empirical process in terms of functions of the form

$$g(x_i, x_i') = f(x_i)\underbrace{\Big\{\big(r(x_i) - \theta^*(x_i)\big) + \gamma\theta^*(x_i')\Big\}}_{\nu(x_i, x_i')}$$

where f varies, while satisfying the constraints $\|f\|_{\mu} \leq u_n$ and $\|f\|_{\mathbb{H}} \leq R$. In order to establish concentration for this supremum, we can apply Talagrand's theorem (cf. Theorem 3.27 in the book [50]). Doing so requires us to bound $\|g\|_{\infty}$, as well as $\mathbb{E}[g^2]$, uniformly over the relevant function class.

Recall that our definition of b ensures that $\|h\|_\infty \le b\,\|h\|_{\mathbb{H}}$ for any $h \in \mathbb{H}$. Consequently, we have

$$\sup_{x,x'} |\nu(x,x')| \le b\|\theta^* - r\|_{\mathbb{H}} + \|\theta^*\|_\infty = bR, \quad \text{and}$$
$$\|g\|_\infty = \sup_{x,x'} \big|f(x)\nu(x,x')\big| \le \|f\|_\infty bR \le b^2R^2,$$

where we have used the fact that $\|f\|_\infty \le b\|f\|_{\mathbb{H}} \le bR$. On the other hand, we have

$$\|g\|_\mu^2 = \mathbb{E}\big[f^2(X)\nu^2(X,X')\big] \le b^2R^2\mathbb{E}[f^2(X)] \le b^2R^2u_n^2.$$

By Talagrand's theorem (cf. equation (3.86) in the book [50]), there are universal constants c_2, c_3 such that

$$\mathbb{P}\Big[Z_n(u_n) \ge 2\,\mathbb{E}[Z_n(u_n)] + c_2bRu_n\sqrt{s} + c_3b^2R^2s\Big] \le \exp(-ns).$$

Setting $s = c_1\frac{u_n^2(1-\gamma)^2}{b^2R^2}$ for a sufficiently small constant c_1 yields the claim in equation (62).

C.3. Proof of Lemma 4.3. Recall our definitions of the operator $\Gamma = \Sigma_{\text{cov}} - \gamma\Sigma_{\text{cr}}$, as well as its empirical version $\widehat{\Gamma} = \widehat{\Sigma}_{\text{cov}} - \gamma\widehat{\Sigma}_{\text{cr}}$, as given in Lemma 4.1. Recall the functional $\rho^2(f) = \mathbb{E}[f^2(X) - \gamma f(X)f(X')]$, as previously defined in equation (37). For each $t > 0$, define the random variable

$$\widetilde{Z}_n(t) := \sup_{\substack{\rho(f)\le t \\ \|f\|_{\mathbb{H}}\le R}} \big|\langle f, (\widehat{\Gamma} - \Gamma)f\rangle_{\mathbb{H}}\big|, \tag{63}$$

and let $t_n > 0$ be the smallest positive solution to the inequality

$$\mathbb{E}[\widetilde{Z}_n(t)] \le \frac{t^2}{8}. \tag{64}$$

We begin by relating this critical radius t_n to our original radius δ_n:

LEMMA C.4. *There is a universal constant c_0 such that*

$$t_n \le u_n := c_0\,R\sqrt{1-\gamma}\,\delta_n(bR). \tag{65}$$

If, in addition, the sample size condition (18) holds, then the same bound holds with $\delta_n(\kappa\sigma(\theta^*))$.

See Appendix C.3.1 for the proof of this claim.

With this set-up, the remainder of proof has a structure similar to that of Lemma 4.2. We say that a function $f \in \mathbb{H}$ satisfies inequality $J(f)$ if

$$\big|\langle f, (\widehat{\Gamma} - \Gamma)f\rangle_{\mathbb{H}}\big| \ge u_n^2 \max\Big\{1, \tfrac{\|f\|_{\mathbb{H}}^2}{R^2}\Big\} + \frac{\rho^2(f)}{2}. \tag{$J(f)$}$$

Now consider the event $\mathcal{B} := \big\{\exists\, f \in \mathbb{H} \text{ s.t } J(f) \text{ holds}\big\}$. Note that conditioned on $\mathcal{B}^c$, we have the bound

$$\begin{aligned}\big|\langle\widehat{\Delta}, (\widehat{\Gamma} - \Gamma)\widehat{\Delta}\rangle_{\mathbb{H}}\big| &\le u_n^2 \max\Big\{1, \tfrac{\|\widehat{\Delta}\|_{\mathbb{H}}^2}{R^2}\Big\} + \frac{\rho^2(\widehat{\Delta})}{2} \\ &\le c_0^2\,\delta_n^2\,(1-\gamma)\max\Big\{R^2, \|\widehat{\Delta}\|_{\mathbb{H}}^2\Big\} + \frac{\rho^2(\widehat{\Delta})}{2}\end{aligned}$$

where the second inequality follows from the definition of u_n given in equation (64). This is the bound claimed in the statement of Lemma 4.3. Consequently, it suffices to bound the probability $\mathbb{P}[\mathcal{B}]$.

We begin by upper bounding the probability of $\mathbb{P}[\mathcal{B}]$ in terms of the tail behavior of the random variable $\widetilde{Z}_n(u_n)$ as follows:

LEMMA C.5. *We have the upper bound*

$$\mathbb{P}[\mathcal{B}] \leq \mathbb{P}\Big[\widetilde{Z}_n(u_n) \geq \tfrac{u_n^2}{2}\Big]. \tag{66}$$

See Appendix C.3.2 for the proof.

Our second lemma provides control on the upper tail of $\widetilde{Z}_n(u_n)$.

LEMMA C.6. *There is a universal constant c_1 such that*

$$\mathbb{P}\Big[\widetilde{Z}_n(u_n) \geq \tfrac{u_n^2}{2}\Big] \leq \exp\big(-c_1 n \tfrac{u_n^2}{b^2R^2}\big) = \exp\big(-c_1 c_0^2 \tfrac{n\delta_n^2(1-\gamma)}{b^2}\big). \tag{67}$$

See Appendix C.3.3 for the proof of this claim.

C.3.1. *Proof of Lemma C.4.* Define the random variables $y_i = (x_i, x_i')$ along with the function $g(y_i) := f^2(x_i) - \gamma f(x_i)f(x_i')$, and note that $\widetilde{Z}_n(t_n)$ is a supremum of the empirical process $\big\{\frac{1}{n}\sum_{i=1}^n (g(y_i) - \mathbb{E}[g(y_i)])\big\}$ as g varies as a function of f, and f satisfies the constraints $\|f\|_{\mathbb{H}} \leq R$ and $\rho(f) \leq t_n$.

By a standard symmetrization argument, we have

$$\mathbb{E}[\widetilde{Z}_n(t_n)] = \mathbb{E}\left[\sup_{\substack{\rho(f)\leq t_n \\ \|f\|_{\mathbb{H}}\leq R}} \Big|\frac{1}{n}\sum_{i=1}^n \big(g(y_i) - \mathbb{E}[g(Y)]\big)\Big|\right] \leq 2\mathbb{E}\left[\sup_{\substack{\rho(f)\leq t_n \\ \|f\|_{\mathbb{H}}\leq R}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(y_i)\Big|\right],$$

where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d. sequence of Rademacher variables.

Now by the lower bound (38), the constraint $\rho(f) \leq t_n$ implies that $\|f\|_\mu \leq \frac{t_n}{\sqrt{1-\gamma}}$. Introducing the shorthand $\mathcal{E} = \big\{f \in \mathbb{H} \mid \|f\|_\mu \leq \frac{t_n}{\sqrt{1-\gamma}},\ \|f\|_{\mathbb{H}} \leq R\big\}$, we have

$$\begin{aligned}
\tfrac{1}{2}\mathbb{E}[\widetilde{Z}_n(t_n)] &\leq \mathbb{E}\left[\sup_{f\in\mathcal{E}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f^2(x_i) - \gamma f(x_i)f(x_i')\big)\Big|\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{E}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f^2(x_i)\Big|\right] + \mathbb{E}\left[\sup_{f\in\mathcal{E}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)f(x_i')\Big|\right].
\end{aligned}$$

From this point, our proof diverges, depending on the two choices of δ_n.

Proof for $\delta_n(bR)$. In this case, we use the fact that $\|f\|_\infty \leq bR$. Combined with the Ledoux-Talagrand contraction, we find that

$$\begin{aligned}
\tfrac{1}{2}\mathbb{E}[\widetilde{Z}_n(t_n)] &\leq 2bR\,\mathbb{E}\left[\sup_{f\in\mathcal{E}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\right] + 4bR\,\mathbb{E}\left[\sup_{f\in\mathcal{E}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\right] \\
&= 6bR\,\mathbb{E}\left[\sup_{f\in\mathcal{E}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\right].
\end{aligned}$$

Define the rescaled ellipse $\tilde{\mathcal{E}} := \frac{1}{R}\mathcal{E} = \big\{h \in \mathbb{H} \mid \|h\|_{\mu} \leq \frac{t_n}{R\sqrt{1-\gamma}},\ \|h\|_{\mathbb{H}} \leq 1\big\}$. By construction, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{E}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\Big|\right] = R\,\mathbb{E}\left[\sup_{h\in\tilde{\mathcal{E}}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(x_i)\Big|\right].$$

Finally, by definition of $\delta_n(bR)$, we are guaranteed that

$$\mathbb{E}\left[\sup_{h\in\tilde{\mathcal{E}}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(x_i)\Big|\right] \leq \frac{R(1-\gamma)}{bR}\max\Big\{\delta_n^2,\ \delta_n\tfrac{t_n}{R\sqrt{1-\gamma}}\Big\}.$$

Putting together the pieces, using the definition of t_n, we have shown that

$$\begin{aligned}\frac{t_n^2}{8} \;=\; \mathbb{E}[\widetilde{Z}_n(t_n)] &\leq \big(6bR^2\big)\,\frac{R(1-\gamma)}{bR}\,\max\Big\{\delta_n^2,\ \delta_n\tfrac{t_n}{R\sqrt{1-\gamma}}\Big\}\\ &= 6R^2\,(1-\gamma)\,\max\Big\{\delta_n^2,\ \delta_n\tfrac{t_n}{R\sqrt{1-\gamma}}\Big\}.\end{aligned}$$

This implies that there is a universal constant c_0 such that $t_n^2 \leq c_0^2\,R^2\,(1-\gamma)\,\delta_n^2$, as claimed in the statement of the lemma.

Proof for $\delta_n(\kappa\sigma(\theta^*))$. Recall the ellipse $\mathcal{E} = \big\{f \in \mathbb{H} \mid \|f\|_{\mu} \leq \frac{t_n}{\sqrt{1-\gamma}},\ \|f\|_{\mathbb{H}} \leq R\big\}$. We claim that it suffices to show that under the assumed bound (18) on the sample size, we have

$$\sup_{f\in\mathcal{E}}\|f\|_{\infty} \leq \beta := \frac{\kappa\sigma(\theta^*)}{128}\Big\{1 + \frac{t_n}{\delta_n R\sqrt{1-\gamma}}\Big\}. \tag{68}$$

Indeed, if this bound holds, then we can perform the Ledoux-Talagrand contraction with the constraint $\|f\|_{\infty} \leq \beta$, so as to conclude that

$$\tfrac{1}{2}\mathbb{E}[\widetilde{Z}_n(t_n)] \leq 6\beta\,\mathbb{E}\left[\sup_{f\in\mathcal{E}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\Big|\right].$$

Proceeding as before, we find that

$$\begin{aligned}\frac{t_n^2}{16} \;=\; \tfrac{1}{2}\mathbb{E}[\widetilde{Z}_n(t_n)] &\leq 6\beta\,\frac{R^2(1-\gamma)}{\kappa\sigma(\theta^*)}\Big\{\delta_n^2 + \frac{\delta_n t_n}{R\sqrt{1-\gamma}}\Big\}\\ &= \frac{R^2(1-\gamma)}{32}\Big\{1 + \frac{t_n}{\delta_n R\sqrt{1-\gamma}}\Big\}\Big\{\delta_n^2 + \frac{\delta_n t_n}{R\sqrt{1-\gamma}}\Big\}\\ &= \frac{R^2(1-\gamma)}{32}\Big\{\delta_n^2 + 2\frac{\delta_n t_n}{R\sqrt{1-\gamma}}\Big\} + \frac{t_n^2}{32}.\end{aligned}$$

This bound implies that $t_n \leq c_0\,R\,\sqrt{1-\gamma}\,\delta_n$, as claimed.

Accordingly, let us prove the bound (68). Any $f \in \mathcal{E}$ has the expansion $f = \sum_{j\geq 1} f_j\phi_j$ for some coefficients such that $\sum_{j=1}^{\infty} f_j^2 \leq t_n^2/(1-\gamma)$ and $\sum_{j=1}^{\infty} f_j^2/\mu_j \leq R^2$. Consequently, we have

$$\begin{aligned}\|f\|_{\infty} \;= \sup_x\Big|\sum_{j=1}^{\infty} f_j\phi_j(x)\Big| &\leq R\,\sup_x\Big\{2\sum_{j=1}^{\infty}\min\big\{\tfrac{t_n^2}{R^2(1-\gamma)},\mu_j\big\}\phi_j^2(x)\Big\}^{1/2}\\ &\overset{(i)}{\leq} R\kappa\Big\{2\sum_{j=1}^{\infty}\min\big\{\tfrac{t_n^2}{R^2(1-\gamma)},\mu_j\big\}\Big\}^{1/2}\end{aligned}$$

$$\stackrel{(ii)}{\leq} R\kappa\sqrt{2n}\,\frac{R(1-\gamma)}{\kappa\sigma(\theta^*)}\left\{\delta_n^2+\frac{\delta_n t_n}{R\sqrt{1-\gamma}}\right\}$$
$$=\left\{\frac{\sqrt{2n}R^2(1-\gamma)\delta_n^2}{\kappa\sigma^2(\theta^*)}\right\}\kappa\sigma(\theta^*)\left\{1+\frac{t_n}{\delta_n R\sqrt{1-\gamma}}\right\}$$
$$\stackrel{(iii)}{\leq}\frac{\kappa\sigma(\theta^*)}{128}\left\{1+\frac{t_n}{\delta_n R\sqrt{1-\gamma}}\right\},$$

where step (i) uses the fact that $\|\phi_j\|_\infty \leq \kappa$ by assumption; step (ii) uses the definition of $\delta_n = \delta_n(\kappa\sigma(\theta^*))$; and step (iii) follows from the assumed bound (18).

C.3.2. *Proof of Lemma C.5.* We first claim that if there is some $f \in \mathbb{H}$ such that $J(f)$ holds, then we can construct a function $g \in \mathbb{H}$ such that $\|g\|_{\mathbb{H}} \leq R$, and

$$\left|\langle g, (\widehat{\Gamma}-\Gamma)g\rangle_{\mathbb{H}}\right| \geq u_n^2 + \frac{\rho^2(g)}{2}. \tag{69}$$

Indeed, if the given function f satisfies $\|f\|_{\mathbb{H}} \leq R$, then we are done. Otherwise, we define the rescaled function $g := \frac{Rf}{\|f\|_{\mathbb{H}}} \in \mathbb{H}$, which satisfies $\|g\|_{\mathbb{H}} = R$. Now observe that

$$\left|\langle g, (\widehat{\Gamma}-\Gamma)g\rangle_{\mathbb{H}}\right| = \frac{R^2}{\|f\|_{\mathbb{H}}^2}\left|\langle f, (\widehat{\Gamma}-\Gamma)f\rangle_{\mathbb{H}}\right| \geq \frac{R^2}{\|f\|_{\mathbb{H}}^2}\left\{u_n^2\max\left\{1,\tfrac{\|f\|_{\mathbb{H}}^2}{R^2}\right\}+\frac{\rho^2(f)}{2}\right\}$$
$$\geq u_n^2+\frac{\rho^2(g)}{2},$$

as claimed.

We now claim that that there must exist some function h with $\|h\|_{\mathbb{H}} \leq R$ and $\rho(h) \leq u_n$ such that $\left|\langle h, (\widehat{\Gamma}-\Gamma)h\rangle_{\mathbb{H}}\right| \geq \frac{u_n^2}{2}$. Indeed, if the g constructed above satisfies $\rho(g) \leq u_n$, then this function has the desired property. Otherwise, we may assume that $\rho(g) > u_n$, and define $h = \frac{u_n}{\rho(g)}g \in \mathbb{H}$. Observe that $\|h\|_{\mathbb{H}} \leq \|g\|_{\mathbb{H}} = R$, and $\rho(h) = u_n$ by construction. Moreover, since g satisfies the lower bound (69), we have

$$\left|\langle h, (\widehat{\Gamma}-\Gamma)h\rangle_{\mathbb{H}}\right| = \frac{u_n^2}{\rho^2(g)}\left|\langle g, (\widehat{\Gamma}-\Gamma)g\rangle_{\mathbb{H}}\right| \geq \frac{u_n^2}{\rho^2(g)}\left\{u_n^2+\frac{\rho^2(g)}{2}\right\}$$
$$\geq \frac{u_n^2}{2}.$$

Putting together the pieces, we have established that the event $\mathcal{B}$ is contained within the event $\left\{\widetilde{Z}_n(u_n) \geq \frac{u_n^2}{2}\right\}$, as claimed.

C.3.3. *Proof of Lemma C.6.* As usual, we proceed by first bounding the mean $\mathbb{E}[\widetilde{Z}_n(u_n)]$, and then establishing concentration around this mean. Since $u_n \geq t_n$, by standard properties of Rademacher complexities, we have

$$\mathbb{E}[\widetilde{Z}_n(u_n)] \leq u_n\,\frac{t_n}{8} \leq \frac{u_n^2}{8}. \tag{70}$$

Consequently, in order to complete the proof, it suffices to show that

$$\mathbb{P}\left[\widetilde{Z}_n(u_n) \geq 2\,\mathbb{E}[\widetilde{Z}_n(u_n)] + \tfrac{u_n^2}{4}\right] \leq \exp\left(-c_1 n\,\tfrac{u_n^2}{b^2R^2}\right). \tag{71}$$

Recall from the proof of Lemma C.4 that the random variable $\widetilde{Z}_n(u_n)$ is the supremum of an empirical process defined by the random variables $y = (x, x')$ and functions of the form

$g(y) = f^2(x) - \gamma f(x) f(x')$. In order to apply Talagrand's concentration inequality, we need to bound $\|g\|_\infty$ and $\mathbb{E}[g^2(Y)]$ uniformly over the class. We have

$$\|g\|_\infty \leq (1+\gamma)\,\|f\|_\infty^2 \leq 2b^2R^2$$

where the final inequality uses the facts that $\gamma \leq 1$, and $\|f\|_\infty \leq bR$ for any function with $\|f\|_{\mathbb{H}} \leq R$. On the other hand, again using the fact that $\|f\|_\infty \leq bR$, we have

$$\begin{aligned}
\mathbb{E}[g^2(Y)] &= \mathbb{E}\Big[f^2(X)\big(f(X) - \gamma f(X')\big)^2\Big] \\
&\leq b^2R^2\,\mathbb{E}\Big[f^2(X) + \gamma^2 f^2(X') - 2\gamma f(X) f(X')\Big] \\
&\overset{(i)}{\leq} 2b^2R^2 \underbrace{\mathbb{E}\Big[f^2(X) - \gamma f(X) f(X')\Big]}_{\rho^2(f)} \overset{(ii)}{\leq} 2b^2R^2u_n^2,
\end{aligned}$$

where inequality (i) uses the fact that $\mathbb{E}[f^2(X')] = \mathbb{E}[f^2(X)]$ and $\gamma \leq 1$; and inequality (ii) uses the fact that $\rho^2(f) \leq u_n^2$ for all functions f in the relevant class.

Consequently, by applying Talagrand's theorem (cf. equation (3.86) in the book [50]), there are universal constants c_2, c_3 such that

$$\mathbb{P}\Big[\widetilde{Z}_n(u_n) \geq 2\,\mathbb{E}[\widetilde{Z}_n(u_n)] + c_2 bRu_n\sqrt{s} + c_3 b^2R^2 s\Big] \leq \exp(-ns).$$

Setting $s = c_1 \frac{u_n^2}{b^2R^2}$ for a sufficiently small constant c_1 yields the claim.

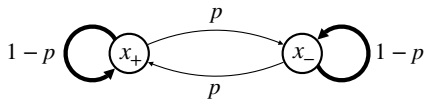
APPENDIX D: AUXILIARY RESULTS FOR THEOREM 3.4

This appendix is devoted to various auxiliary results associated with Theorem 3.4. In Appendix D.1, we discuss the sample size requirements of the theorem, along with the conditions on the kernel eigenvalues. Appendix D.2 presents constructions of RKHSs $\mathbb{H}_A$ and $\mathbb{H}_B$ as well as MRP families $\mathfrak{M}_A$ and $\mathfrak{M}_B$. In the remaining subsections, we provide various technical results used in the construction. Appendix D.3 is devoted to the analysis of MRP class $\mathfrak{M}_A$; whereas Appendix D.4 concerns family $\mathfrak{M}_B$.

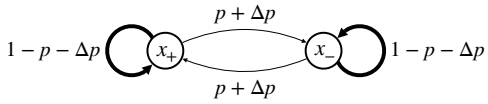
D.1. Conditions of Theorem 3.4. In this appendix, we discuss the requirements needed for our lower bounds to be valid. First, we claim that the sample size conditions required for the lower bounds in Theorem 3.4, parts (a) and (b), are are weaker than the requirement (18) for the upper bounds in Theorem 3.1. It is clear that the constraint in part (a) and the first inequality in constraint (36b) are looser than bound (18). Additionally, the second inequality in condition (36b) is easy to satisfy if the eigengap $(1 - \frac{\mu_2}{\mu_1})$ has a constant order and the uniform bound b and the radius $\bar{R}$ are not too large. For these reasons, the lower bounds require even milder conditions on the sample size n.

We also note that the eigengap condition $\min_{3\leq j\leq d_n}\{\sqrt{\mu_{j-1}} - \sqrt{\mu_j}\} \geq \frac{\delta_n}{2d_n}$ in Theorem 3.4 is rather mild. For instance, if we consider a kernel whose eigenvalues exhibit α-polynomial decay (22)—that is, say $\mu_j = cj^{-2\alpha}$ for some constant $c > 0$ and exponent $\alpha > \frac{1}{2}$. In this case, we have

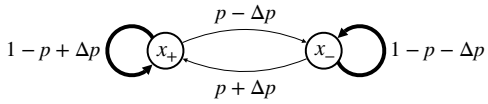
$$\begin{aligned}
\sqrt{\mu_{j-1}} - \sqrt{\mu_j} &= \sqrt{c}\,\big\{(j-1)^{-\alpha} - j^{-\alpha}\big\} \\
&= \sqrt{c}\,j^{-\alpha}\big\{(1 - \frac{1}{j})^{-\alpha} - 1\big\} \geq \sqrt{c}\,j^{-\alpha}\frac{\alpha}{j} = \sqrt{\mu_j}\,\frac{\alpha}{j} \geq \frac{\sqrt{\mu_j}}{2j}.
\end{aligned}$$



(a) Base Markov chain $\mathbf{P}_0(p)$.



(b) Construction of $\mathbf{P}_A(p, \Delta p)$.



(c) Construction of $\mathbf{P}_B(p, \Delta p)$.

Fig 6. Two-state MRP instances $\mathbf{P}_0$, $\mathbf{P}_A$ and $\mathbf{P}_B$. The MRP instances are parameterized by scalars $p \in \left[0, \frac{1}{2}\right]$ and $\Delta p \in [-p, p]$. The parameters are chosen so as to ensure that all edges are labeled with valid probabilities.

If $j \le d_n$, then $\sqrt{\mu_j} \ge \delta_n$. Therefore, $\sqrt{\mu_{j-1}} - \sqrt{\mu_j} \ge \frac{\delta_n}{2d_n}$ for any $j \le d_n$ and the assumption is satisfied.

Finally, we comment on the critical inequality (33), and show that in Regime B, by replacing $\{\mu_j\}_{j=1}^{\infty}$ with $\{\mu_j(\mathcal{P})\}_{j=1}^{\infty}$, we would get a smaller critical radius. We recall from definition (44b) of $\mathfrak{M}_B$ that any $\mathscr{I}(\mathcal{P}, r_B, \gamma) \in \mathfrak{M}_B$ satisfies $\mu_j(\mathcal{P}) \le \mu_j$ for any $j \ge 2$. Since the statistical dimension $d_n = \max\left\{j \mid \mu_j \ge \delta_n^2\right\}$, we have $\min\left\{\mu_j(\mathcal{P}), \delta_n^2\right\} \le \min\left\{\mu_j, \delta_n^2\right\}$ for any $j \in \mathbb{Z}_+$ as long as $d_n \ge 2$. Recall that δ_n is the smallest positive solution to inequality (33), therefore,

$$\sqrt{\sum_{j=1}^{\infty} \min\left\{\frac{\mu_j(\mathcal{P})}{\delta_n^2}, 1\right\}} \le \sqrt{\sum_{j=1}^{\infty} \min\left\{\frac{\mu_j}{\delta_n^2}, 1\right\}} \le \sqrt{n}\, \frac{\bar{R}\,(1-\gamma)}{2\bar{\sigma}}\, \delta_n\,.$$

In other words, δ_n satisfies the critical inequality defined by $\{\mu_j(\mathcal{P})\}_{j=1}^{\infty}$. Hence, $\delta_n \ge \delta_n(\mathcal{P})$, where $\delta_n(\mathcal{P})$ is the critical radius induced by $\{\mu_j(\mathcal{P})\}_{j=1}^{\infty}$. In this way, $\mathrm{LB}(\bar{R}, \bar{\sigma}, \delta_n)$ further implies another lower bound $\|\widehat{\theta} - \theta^*\|_{\mu(\mathcal{P})}^2 \ge c_1\, \bar{R}^2 \delta_n^2(\mathcal{P})$, as claimed in Section 4.2.1.

D.2. Constructions of MRP instances. In this part, we present the constructions of RKHSs and MRP families used in the lower bound proof, and verify the claims (45), (47a) and (47b) in Section 4.2.2. In particular, Appendices D.2.1 and D.2.2 introduce the general structure of the MRP instances we devised. Appendix D.2.3 is devoted to the design of two RKHSs. In Appendix D.2.4, we define two model families $\mathfrak{M}_A$ and $\mathfrak{M}_B$ by specifying the parameters in our MRP construction in Appendix D.2.2.

D.2.1. *Construction of simple two-state MRPs.* As a warm-up, we first construct a simple two-state Markov chain and two perturbed variants of it. Each variant is the basic building block that underlies our full-scale "hard" instances in $\mathfrak{M}_A$ or $\mathfrak{M}_B$. Denote the states by x_+ and x_-. Given a scalar $p \in \left[0, \frac{1}{2}\right]$, the base Markov chain is defined by the 2×2 transition matrix

$$\mathbf{P}_0 \equiv \mathbf{P}_0(p) := \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}. \tag{72a}$$

See Figure 6a for an illustration of the transition dynamics of the base model.

We further define the perturbed variants $\mathbf{P}_A$ and $\mathbf{P}_B$ of model $\mathbf{P}_0$. In addition to the parameter p, we introduce another scalar $\Delta p \in [-p, p]$. The Markov chains are constructed as follows:

$$\mathbf{P}_A \equiv \mathbf{P}_A(p, \Delta p) := \begin{pmatrix} 1-p-\Delta p & p+\Delta p \\ p+\Delta p & 1-p-\Delta p \end{pmatrix}, \tag{72b}$$
$$\mathbf{P}_B \equiv \mathbf{P}_B(p, \Delta p) := \begin{pmatrix} 1-p+\Delta p & p-\Delta p \\ p+\Delta p & 1-p-\Delta p \end{pmatrix}.$$

Panels (b) and (c) in Figure 6 represent these two processes respectively.

Consider a reward function $\boldsymbol{r}$ given by $\boldsymbol{r}(x_+) := r$ and $\boldsymbol{r}(x_-) := -r$ where $r \in \mathbb{R}$ is a scalar. Let $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ be the value functions associated with transition kernels $\mathbf{P}_0$, $\mathbf{P}_A$ and $\mathbf{P}_B$. Then $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ can be viewed as perturbations of $\boldsymbol{\theta}_0$ in two different directions. Specifically, we perform calculations and find that $\boldsymbol{\theta}_0 = (1-\gamma+2\gamma p)^{-1}\boldsymbol{r}$ and the differences $\Delta\boldsymbol{\theta}_A = \boldsymbol{\theta}_A - \boldsymbol{\theta}_0$ and $\Delta\boldsymbol{\theta}_B = \boldsymbol{\theta}_B - \boldsymbol{\theta}_0$ satisfy the relations

$$\Delta\boldsymbol{\theta}_A(x_+) = -\Delta\boldsymbol{\theta}_A(x_-) \qquad \text{and} \qquad \Delta\boldsymbol{\theta}_B(x_+) = \Delta\boldsymbol{\theta}_B(x_-). \tag{73}$$

In the sequel, we construct full-scale MRP instances using the Markov chains $\mathbf{P}_A$ and $\mathbf{P}_B$.

D.2.2. *Construction of MRPs over state space $\mathcal{X} = [0, 1)$.* In this part, we assemble K different two-state Markov chains $\{\mathbf{P}^{(k)}\}_{k=1}^K$ into a full-scale model $\mathcal{P}$ over state space $\mathcal{X} = [0, 1)$. In our constructions, matrix $\mathbf{P}^{(k)}$ takes the form of $\mathbf{P}_A\big(p, \Delta p^{(k)}\big)$ in Regime A and $\mathbf{P}_B\big(p, \Delta p^{(k)}\big)$ in Regime B, where the parameters K, p and $\{\Delta p^{(k)}\}_{k=1}^K$ will be specified later.

We evenly partition the state space $\mathcal{X} = [0, 1)$ into $2K$ intervals

$$\Delta_+^{(k)} := \big[\tfrac{k-1}{2K}, \tfrac{k}{2K}\big) \quad \text{and} \quad \Delta_-^{(k)} := \big[\tfrac{1}{2} + \tfrac{k-1}{2K}, \tfrac{1}{2} + \tfrac{k}{2K}\big) \qquad \text{for } k = 1, 2, \ldots, K. \tag{74}$$

For each index $k \in [K]$, the dynamics of $\mathcal{P}$ on intervals $\Delta_+^{(k)}$ and $\Delta_-^{(k)}$ follow the local model $\mathbf{P}^{(k)}$. With slight abuse of notation, we denote the two states of Markov chain $\mathbf{P}^{(k)}$ by x_+ and x_- for any $k \in [K]$. The transition kernel $\mathcal{P}$ is then defined as [1]

$$\mathcal{P}(x' \mid x) := \begin{cases} 2K\, \mathbf{P}^{(k)}(x_+ \mid x_+) & \text{if } x, x' \in \Delta_+^{(k)}, \\ 2K\, \mathbf{P}^{(k)}(x_- \mid x_+) & \text{if } x \in \Delta_+^{(k)}, x' \in \Delta_-^{(k)}, \\ 2K\, \mathbf{P}^{(k)}(x_- \mid x_-) & \text{if } x, x' \in \Delta_-^{(k)}, \\ 2K\, \mathbf{P}^{(k)}(x_+ \mid x_-) & \text{if } x \in \Delta_-^{(k)}, x' \in \Delta_+^{(k)}, \\ 0 & \text{otherwise.} \end{cases} \tag{75}$$

Figure 7 illustrates our construction of model $\mathcal{P}$.

In the full-scale MRP $\mathscr{I}(\mathcal{P}, r, \gamma)$, we take a reward function

$$r(x) := r\left(\mathbb{1}\big\{x \in [0, \tfrac{1}{2})\big\} - \mathbb{1}\big\{x \in [\tfrac{1}{2}, 1)\big\}\right) \tag{76}$$

[1] A technical side-comment: note that the Markov chain (75) is not ergodic. However, this issue can be remedied with a slight modification of the transition kernel $\mathcal{P}$. Let $\widetilde{\mu}$ be a stationary distribution of $\mathcal{P}$. We fix a number $\epsilon \in (0, 1)$. At each time step, let the Markov chain follow $\mathcal{P}$ with probability $1-\epsilon$, and transit to a next state according to $\widetilde{\mu}$ with probability ϵ. This procedure defines a new transition kernel $\widetilde{\mathcal{P}}(\cdot \mid x) := \epsilon\widetilde{\mu}(\cdot) + (1-\epsilon)\mathcal{P}(\cdot \mid x)$, which induces a new Markov chain that is ergodic, and has a unique stationary distribution. Since $\epsilon > 0$ can be chosen arbitrarily close to zero, we can recover statements about the original model in this way. The ϵ-modification would induce unnecessary clutter, so that we focus on model (75) in the following discussion.

with a scalar $r \in \mathbb{R}$. By our construction, the transition kernel $\mathcal{P}$ and reward function r produce a value function θ^* that is piecewise constant over intervals $\Delta_+^{(k)}$ and $\Delta_-^{(k)}$. Moreover, we have

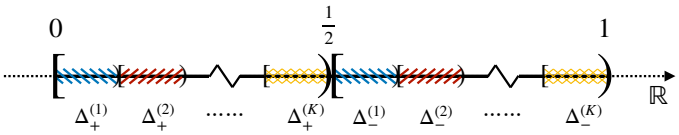
$$\theta^*(x) = \begin{cases} \boldsymbol{\theta}^{(k)}(x_+) & \text{if } x \in \Delta_+^{(k)}, \\ \boldsymbol{\theta}^{(k)}(x_-) & \text{if } x \in \Delta_-^{(k)}, \end{cases} \tag{77}$$

where $\boldsymbol{\theta}^{(k)} := (\mathbf{I} - \gamma \mathbf{P}^{(k)})^{-1} \boldsymbol{r} \in \mathbb{R}^2$ is the value vector given by transition matrix $\mathbf{P}^{(k)}$ and reward vector $\boldsymbol{r} = [r, -r]^\top$.

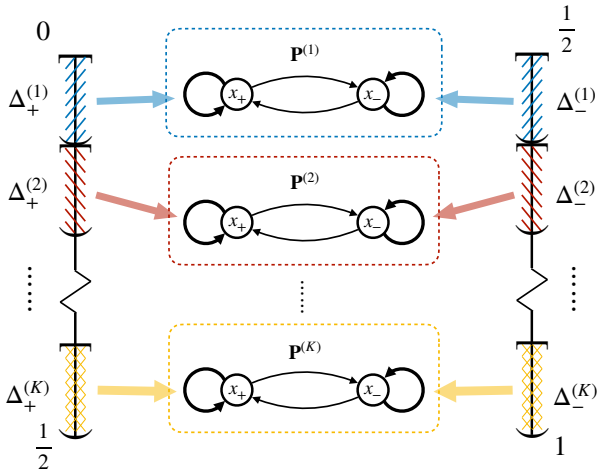
We consider the form of the full-scale value function θ^* when taking $\mathbf{P}^{(k)} = \mathbf{P}_0(p)$, $\mathbf{P}_A\big(p, \Delta p^{(k)}\big)$ or $\mathbf{P}_B\big(p, \Delta p^{(k)}\big)$. If we set $\mathbf{P}^{(k)} = \mathbf{P}_0(p)$, then the value function is given by

$$\theta^*(x) = \theta_0^*(x) := (1 - \gamma + 2\gamma p)^{-1} r(x). \tag{78}$$

We refer to θ_0^* as the base value function. If $\mathbf{P}^{(k)} = \mathbf{P}_A\big(p, \Delta p^{(k)}\big)$, then due to equation (73), we have $\theta^* = \theta_0^* + \Delta\theta^*$ with function $\Delta\theta^*$ satisfying $\Delta\theta^*(x) = -\Delta\theta^*(x + \frac{1}{2})$ for any $x \in [0, \frac{1}{2})$. When $\mathbf{P}^{(k)} = \mathbf{P}_B\big(p, \Delta p^{(k)}\big)$, the value function θ^* admits a decomposition $\theta^* = \theta_0^* + \Delta\theta^*$ with $\Delta\theta^*(x) = \Delta\theta^*(x + \frac{1}{2})$ for any $x \in [0, \frac{1}{2})$. In the following, we construct function spaces $\mathbb{H}_A$ and $\mathbb{H}_B$ of which the elements possess these properties.



(a) Partition of state space $\mathcal{X} = [0, 1)$.



(b) Construction of transition kernel $\mathcal{P}$ over state space $\mathcal{X} = [0, 1)$ using 2-state Markov chains $\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots, \mathbf{P}^{(K)}$.

Fig 7. Embedding of two-state Markov chains $\{\mathbf{P}^{(k)}\}_{k=1}^K$ into state space $\mathcal{X} = [0, 1)$. Up: partition of state space $\mathcal{X}$ into intervals $\{\Delta_+^{(k)}, \Delta_-^{(k)}\}_{k=1}^K$. Bottom: the transitions on intervals $\Delta_+^{(k)}$ and $\Delta_-^{(k)}$ follow a local Markov chain $\mathbf{P}^{(k)}$.

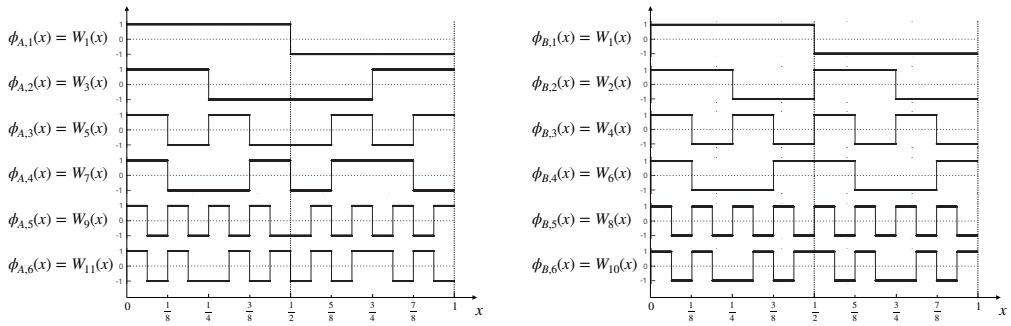


Fig 8. Construction of functions $\{\phi_{A,j}\}_{j=1}^{\infty}$ and $\{\phi_{B,j}\}_{j=1}^{\infty}$. Left: Basis functions $\phi_{A,1}, \phi_{A,2}, \ldots, \phi_{A,6}$. Right: Basis functions $\phi_{B,1}, \phi_{B,2}, \ldots, \phi_{B,6}$.

D.2.3. *Constructing the Hilbert spaces* $\mathbb{H}_A$ *and* $\mathbb{H}_B$. Given a sequence $\{\mu_j\}_{j=1}^{\infty}$ of non-negative numbers, we construct two RKHSs $\mathbb{H}_A$ and $\mathbb{H}_B$ of functions with domain $\mathcal{X} = [0,1)$, such that both the associated kernels have eigenvalues $\{\mu_j\}_{j=1}^{\infty}$ under the Lebesgue measure μ. Our construction is designed to produce kernels that are especially amenable to analysis, and can easily connect to the MRPs defined in equation (75). In particular, we leverage the Walsh system, an orthonormal basis of $L^2(\mu)$ that can represent discrete functions conveniently. For any $j \in \mathbb{N}$, the j-th Walsh function is given by

$$W_j(x) := (-1)^{\sum_{i=0}^{\infty} k_i x_{i+1}} \text{ for } j = \sum_{i=0}^{\infty} k_i 2^i,\ x = x_0 + \sum_{i=1}^{\infty} x_i 2^{-i}$$

with $k_i, x_i \in \{0,1\}$ and $x_0 \in \mathbb{Z}$. Specifically, the first Walsh function takes the form $W_1(x) = \mathbb{1}\{x \in [0, \frac{1}{2})\} - \mathbb{1}\{x \in [\frac{1}{2}, 1)\}$.

Below we construct two groups of functions $\{\phi_{A,j}\}_{j=1}^{\infty}$ and $\{\phi_{B,j}\}_{j=1}^{\infty}$ that are bases of $\mathbb{H}_A$ and $\mathbb{H}_B$ respectively:

$$\text{(79a)} \qquad \phi_{A,j}(x) := W_{2j-1}(x) = W_{j-1}(2x)\, W_1(x) \qquad \text{for } j = 1, 2, 3, \ldots \qquad \text{and}$$

$$\text{(79b)} \qquad \phi_{B,j}(x) := \begin{cases} W_1(x) & \text{if } j = 1, \\ W_{2(j-1)}(x) = W_{j-1}(2x) & \text{if } j = 2, 3, \ldots. \end{cases}$$

See Figure 8 for an illustration of the top 6 basis functions in each group. Based on $\{\phi_{A,j}\}_{j=1}^{\infty}$ and $\{\phi_{B,j}\}_{j=1}^{\infty}$, we define the kernel functions $\mathcal{K}_A$ and $\mathcal{K}_B$ as

$$\text{(80)} \qquad \mathcal{K}_\iota(x,y) := \sum_{j=1}^{\infty} \mu_j\, \phi_{\iota,j}(x)\, \phi_{\iota,j}(y) \qquad \text{for } \iota = A \text{ or } B$$

and let $\mathbb{H}_A$ and $\mathbb{H}_B$ be the RKHSs induced by $\mathcal{K}_A$ and $\mathcal{K}_B$.

The function classes $\{\phi_{A,j}\}_{j=1}^{\infty}$ and $\{\phi_{B,j}\}_{j=1}^{\infty}$ above are both orthonormal in $L^2(\mu)$. Indeed, we have $\int_{\mathcal{X}} \phi_{\iota,i}(x)\phi_{\iota,j}(x)\mu(dx) = \mathbb{1}\{i = j\}$ for $\iota = A$ or B and any $i, j \in \mathbb{Z}_+$. Hence, the kernels $\mathcal{K}_A$ and $\mathcal{K}_B$ have eigenpairs $\{(\mu_j, \phi_{A,j})\}_{j=1}^{\infty}$ and $\{(\mu_j, \phi_{B,j})\}_{j=1}^{\infty}$ associated with the Lebesgue measure μ.

Our choice of bases $\{\phi_{A,j}\}_{j=1}^{\infty}$ and $\{\phi_{B,j}\}_{j=1}^{\infty}$ is especially tailored to the MRP construction in Appendix D.2.2. Suppose K is a power of 2 and let $\{\Delta_+^{(k)}, \Delta_-^{(k)}\}_{k=1}^{K}$ be a partition of state space $\mathcal{X} = [0,1)$ given in equation (74). Then for any function f that is piecewise constant with respect to the partition and satisfies $f(x) = -f(x + \frac{1}{2})$ for any $x \in [0, \frac{1}{2})$, it

can always be linearly expressed by functions $\{\phi_{A,1}, \ldots, \phi_{A,K}\}$. Similarly, the function set $\{\phi_{B,2}, \ldots, \phi_{B,K}\}$ is capable of representing any discrete function f that is adapted to the partition and satisfies $f(x) = f(x + \frac{1}{2})$ for any $x \in [0, \frac{1}{2})$.

D.2.4. *Two families of MRPs.* We now construct a family $\{\mathscr{I}_m\}_{m=1}^M$ of MRP instances using the transition kernel and reward function defined in equations (75) and (76), with value functions belonging to either $\mathbb{H}_A$ or $\mathbb{H}_B$. Recall our definition $d_n = \max\{j \mid \mu_j \geq \delta_n^2\}$ of the effective dimension (at sample size n) of the underlying kernel class. Consider the Boolean hypercube $\{0,1\}^{d_n - 1}$, and let $\{\boldsymbol{\alpha}_m\}_{m=1}^M$ be a $\frac{1}{4}$-(maximal) packing of it with respect to the (rescaled) Hamming metric

$$\rho_H(\boldsymbol{\alpha}, \boldsymbol{\alpha}') := \frac{1}{d_n - 1} \sum_{k=1}^{d_n - 1} \mathbb{1}\{\alpha_k \neq \alpha'_k\}. \tag{81}$$

It is known from standard results on metric entropy (e.g., see Example 5.3 in the book [50]) that there exists such a set with log cardinality lower bounded as $\log M \geq \frac{d_n}{10}$. Using this packing of the Boolean hypercube, we now show how to construct the MRP instance $\mathscr{I}_m$ based on the binary vector $\boldsymbol{\alpha}_m$.

In either Regime A or B, the MRP instances $\{\mathscr{I}_m\}_{m=1}^M$ share the same reward function $r(x) = r_A(x)$ or $r_B(x)$. Each model $\mathscr{I}_m$ has a transition kernel $\mathcal{P}_m$ that lies within a neighborhood of a base Markov chain $\mathcal{P}_0$. The difference between $\mathcal{P}_m$ and $\mathcal{P}_0$ is encoded by vector $\boldsymbol{\alpha}_m$. Specifically, we pick a transition kernel $\mathcal{P}_m$ such that the difference in value functions $\theta_m^* - \theta_0^*$ is a linear combination of functions $\{\phi_{A,j}\}_{j=2}^{d_n}$ or $\{\phi_{B,j}\}_{j=2}^{d_n}$, with vector $\boldsymbol{\alpha}_m$ determining the linear coefficients. Here, θ_0^* is the base value function given by equation (78).

In our constructions below, we take $K := 2^{\lceil \log_2 d_n \rceil}$. It is ensured that the functions $\{\phi_{A,j}\}_{j=1}^{d_n}$ and $\{\phi_{B,j}\}_{j=1}^{d_n}$ are piecewise constant with respect to the partition $\{\Delta_+^{(k)}, \Delta_-^{(k)}\}_{k=1}^K$. Recall from definition (75) that transition kernel $\mathcal{P}_m$ is determined by local models $\{\mathbf{P}_m^{(k)}\}_{k=1}^K$. In the sequel, we specify the choices of $\{\mathbf{P}_m^{(k)}\}_{k=1}^K$ so that the value function θ_m^* has the desired form.

Regime A. We first construct MRP instances $\{\mathscr{I}_m\}_{m=1}^M$ that belong to the model class $\mathfrak{M}_A$. In order that the regularity condition $\|r_A\|_\infty \leq 1$ holds, we simply set parameter $r := 1$ in equation (76) so that the reward function $r_A(x) = W_1(x)$.

In our design of the transition kernel $\mathcal{P}_m$, the local Markov chains are $\mathbf{P}_m^{(k)} := \mathbf{P}_A\big(p, \Delta p_m^{(k)}\big)$ where $\mathbf{P}_A$ is given in equation (72b) and the parameter p is chosen as $p := \frac{3(1-\gamma)}{\gamma}$. We remark that the uniform distribution μ is stationary under model $\mathcal{P}_m$, so we pick $\mu_m = \mu(\mathcal{P}_m) = \mu$. We take parameters $\{\Delta p_m^{(k)}\}_{k=1}^K$ such that the value function θ_m^* of MRP $\mathscr{I}_m$ satisfies

$$\theta_m^* = \theta_0^* - \frac{2\gamma}{(1 - \gamma + 2\gamma p)^2} f_m, \tag{82}$$

where $\theta_0^* = (1 - \gamma + 2\gamma p)^{-1} W_1$ and

$$f_m(x) := \sqrt{\frac{p(1-p)}{120\, n}} \sum_{j=2}^{d_n} \alpha_m^{(j-1)}\, \phi_{A,j}(x). \tag{83}$$

In order to do so, we set

$$\Delta p_m^{(k)} := \frac{1 - \gamma + 2\gamma p}{1 - \gamma + 2\gamma p - 2\gamma f_m(x_k)}\, f_m(x_k) \qquad \text{for } m \in [M] \text{ and } k \in [K] \tag{84}$$

in the local Markov chain $\mathbf{P}_m^{(k)} = \mathbf{P}_A\big(p, \Delta p_m^{(k)}\big)$. Recall from equation (77) that $\theta_m^*(x) = \boldsymbol{\theta}_m^{(k)}(x_+) = -\boldsymbol{\theta}_m^{(k)}(x_-)$ for any $x \in \Delta_+^{(k)}$, where $\boldsymbol{\theta}_m^{(k)} \in \mathbb{R}^2$ is the value vector induced by model $\mathbf{P}_m^{(k)}$ and reward vector $\boldsymbol{r} = [1, -1]^\top$. Under our choice of $\Delta p_m^{(k)}$ in equation (84), the value function θ_m^* has the desired form as in equation (82).

Regime B. We now construct MRP instances $\{\mathscr{I}_m\}_{m=1}^M$ in family $\mathfrak{M}_B$. In this scenario, we take the parameter $p := \frac{1}{8}$ in local models $\mathbf{P}_m^{(k)} := \mathbf{P}_B\big(p, \Delta p_m^{(k)}\big)$ and $r := p + \frac{1-\gamma}{2\gamma}$ in the definition (76) of reward function so that $r_B(x) = \big(p + \frac{1-\gamma}{2\gamma}\big) W_1(x)$. Moreover, we set $\Delta p_m^{(k)} := f_m(x_k)$ in model $\mathbf{P}_m^{(k)} = \mathbf{P}_B\big(p, \Delta p_m^{(k)}\big)$, where

$$f_m(x) := \frac{p}{25\sqrt{n}} \sum_{j=2}^{d_n} \alpha_m^{(j-1)}\, \phi_{B,j}(x) \tag{85}$$

and x_k is any point in interval Δ_k^+ or Δ_k^-. The value function θ_m^* then satisfies

$$\theta_m^* = \theta_0^* + \frac{1}{1-\gamma}\, f_m\,. \tag{86}$$

We observe that in this case, the transition kernel $\mathcal{P}_m$ has a stationary distribution

$$\mu_m(x) := \begin{cases} 1 + \frac{f_m(x)}{p} & \text{if } x \in \Delta_k^+, \\ 1 - \frac{f_m(x)}{p} & \text{if } x \in \Delta_k^-. \end{cases} \tag{87}$$

The measure μ_m is not the uniform distribution μ; however, our construction ensures that $\frac{d\mu_m}{d\mu}(x) \geq \frac{1}{2}$. See Appendix D.4.1 for the proof of this claim.

We claim that both of our constructions yield MRPs that belong to the desired classes:

LEMMA D.1. *The previously described constructions yield MRP instances $\mathscr{I}_m$ such that $\{\mathscr{I}_m\}_{m=1}^M \subset \mathfrak{M}_A$ in Regime A, and $\{\mathscr{I}_m\}_{m=1}^M \subset \mathfrak{M}_B$ in Regime B.*

We prove the Regime A claim in Appendix D.3.1, and the Regime B claim in Appendix D.4.2.

We now need to establish upper bounds on the pairwise KL divergences, and lower bounds on the pairwise $L^2(\mu)$-distances, as stated informally in equations (47). The precise statements are as follows:

LEMMA D.2. *For either of the two classes ($\mathfrak{M}_A$ in Regime A, or $\mathfrak{M}_B$ in Regime B), our construction ensures that*

$$D_{KL}\big(\mathcal{P}_m^{1:n} \,\big\|\, \mathcal{P}_{m'}^{1:n}\big) \leq \frac{d_n}{40} \qquad \textit{for any } m, m' \in [M]. \tag{88}$$

See Appendices D.3.2 and D.4.3, respectively, for the proofs corresponding to the classes $\mathfrak{M}_A$ and $\mathfrak{M}_B$.

LEMMA D.3. *Our construction ensures that there exists a universal constant c_1' such that*

$$\min_{m \neq m'} \big\|\theta_m^* - \theta_{m'}^*\big\|_\mu \geq c_1' \sqrt{c}\, \bar{R} \delta_n. \tag{89}$$

The claim holds for both $\{\mathscr{I}_m\}_{m=1}^M \subset \mathfrak{M}_A$ and $\{\mathscr{I}_m\}_{m=1}^M \subset \mathfrak{M}_B$.

This claim is proved in Appendices D.3.3 and D.4.4 for $\mathfrak{M}_A$ and $\mathfrak{M}_B$ respectively.

D.3. Proofs of auxiliary results in Regime A. In this part, Appendix D.3.1 presents the proof of Lemma D.1, which shows the well-definedness of our MRP instances and verifies that they belong to the model family $\mathfrak{M}_A$. Appendix D.3.2 is devoted to the proof of Lemma D.2, which provides an upper bound on the pairwise KL distances in our construction. On the other hand, Appendix D.3.3 provides the proof of Lemma D.3, which lower bounds the pairwise distances between the value functions in our model family.

D.3.1. *Proof of Lemma D.1.* We verify the conditions in the definition (44a) of family $\mathfrak{M}_A$. In order that our constructed MRP instances $\mathscr{I}_m \in \mathfrak{M}_A$ for any $m \in [M]$, we check the constraints in equation (44a) one by one. We first note that $\theta_m^* \in \mathbb{H}_A$ by our construction. As for condition (ii) in definition (44a), we recall that all models $\{\mathscr{I}_m\}_{m=1}^M \subset \mathfrak{M}_A$ have Lebesgue measure μ as the common stationary distribution, thus the covariance operator Σ_{cov} has eigenpairs $\{(\mu_j, \phi_{A,j})\}_{j=1}^\infty$, where $\{\mu_j\}_{j=1}^\infty$ are the pre-specified parameters and $\{\phi_{A,j}\}_{j=1}^\infty$ are the bases of $\mathbb{H}_A$ defined in equation (79a). Since $\sup_{j\in\mathbb{Z}_+}\|\phi_{A,j}\|_\infty = 1 \le \kappa$, condition (ii) is satisfied. In the sequel, we only need to verify inequalities (32b). Specifically, we will prove that for each $\mathscr{I}_m$, the following properties hold:

- The Bellman residual variance satisfies $\sigma^2(\theta_m^*) \le \bar{\sigma}^2$, and;
- The norms satisfy $\max\big\{\|\theta_m^* - r_A\|_{\mathbb{H}_A}, \frac{2\|\theta_m^*\|_\infty}{b}\big\} \le \bar{R}$.

Before proving the two claims above, we first develop upper bounds on $\|f_m\|_\infty$ and $\big|\Delta p_m^{(k)}\big|$, which are crucial in our estimations below. We claim that

$$\|f_m\|_\infty \le \frac{p}{9} \qquad \text{and} \qquad \big|\Delta p_m^{(k)}\big| \le \frac{p}{8} \qquad \text{for any } k \in [K] \text{ and } m \in [M]\,. \tag{90}$$

In fact, by using the definition of f_m in equation (82) and the fact that $\|\phi_{A,j}\|_\infty \le \kappa$, we find that

$$\|f_m\|_\infty \le \sqrt{\frac{p\,(1-p)}{120\,n}} \sum_{j=2}^{d_n} \|\phi_{A,j}\|_\infty \le \kappa d_n \sqrt{\frac{p}{120\,n}}\,.$$

The critical inequality (33) ensures $d_n \le n\left\{\frac{\bar{R}\delta_n(1-\gamma)}{\kappa\bar{\sigma}}\right\}^2$, and therefore

$$\|f_m\|_\infty \le \kappa n \left\{\frac{\bar{R}\delta_n(1-\gamma)}{\kappa\bar{\sigma}}\right\}^2 \sqrt{\frac{p}{120\,n}} \overset{(i)}{\le} \sqrt{\tfrac{1}{30}\,p\,(1-\gamma)} \le \frac{p}{9}\,,$$

where we have used condition (36a) in the step (i). We plug the inequality $\|f_m\|_\infty \le \frac{p}{9}$ into the definition of $\Delta p_m^{(k)}$ in equation (84). It follows that

$$\big|\Delta p_m^{(k)}\big| = \frac{1-\gamma+2\gamma p}{1-\gamma+2\gamma p - 2\gamma f_m(x_k)}\,|f_m(x_k)| \le \frac{1-\gamma+2\gamma p}{1-\gamma+2\gamma p - 2\gamma(p/9)}\,(p/9) \le \frac{p}{8}\,.$$

Upper bound on $\sigma^2(\theta_m^*)$. We consider the condition $\sigma(\theta_m^*) \le \bar{\sigma}$. Recall that the MRP $\mathscr{I}_m$ consists of K local models, each is determined by the transition matrix $\mathbf{P}_m^{(k)} = \mathbf{P}_A\big(p, \Delta p_m^{(k)}\big)$ and reward vector $\boldsymbol{r} = [1, -1]^\top$. The Bellman residual variance $\sigma^2(\theta_m^*)$ of the full-scale MRP $\mathscr{I}_m$ is the average of those of the small local MRPs. Let $\boldsymbol{\theta}_m^{(k)}$ be the value function associated with the k-th local MRP. We use some algebra and find that

$$\sigma^2\big(\boldsymbol{\theta}_m^{(k)}\big) = \frac{4\gamma^2\big(p + \Delta p_m^{(k)}\big)\big(1 - p - \Delta p_m^{(k)}\big)}{\big(1-\gamma+2\gamma p + 2\gamma\Delta p_m^{(k)}\big)^2}\,.$$

Since $\left|\Delta p_m^{(k)}\right| \leq p/8$ and $p = \frac{3(1-\gamma)}{\gamma}$, we have

$$\sigma^2\big(\boldsymbol{\theta}_m^{(k)}\big) \leq \frac{(4\gamma^2)\,(p+p/8)\,(1-p-p/8)}{(1-\gamma+2\gamma p-2\gamma(p/8))^2} \leq \frac{1+\gamma}{5(1-\gamma)} \leq \bar{\sigma}^2.$$

By taking the average of $\sigma^2\big(\boldsymbol{\theta}_m^{(k)}\big)$ over indices $k \in [K]$, we conclude that

$$\sigma^2(\theta_m^*) = \frac{1}{K}\sum_{k=1}^{K} \sigma^2\big(\boldsymbol{\theta}_m^{(k)}\big) \leq \bar{\sigma}^2.$$

Upper bounds on $\|\theta_m^* - r_A\|_{\mathbb{H}_A}$ *and* $\|\theta_m^*\|_\infty$. We first consider the RKHS norm $\|\theta_m^* - r_A\|_{\mathbb{H}_A}$. Recall from equations (76) and (82) that the reward and value functions r_A and θ_m^* are

$$r_A = \phi_{A,1} \qquad \text{and}$$

$$\theta_m^* = \frac{1}{1-\gamma+2\gamma p}\,\phi_{A,1} - \frac{2\gamma}{(1-\gamma+2\gamma p)^2}\sqrt{\frac{p\,(1-p)}{120\,n}}\,\sum_{j=2}^{d_n} \alpha_m^{(j-1)}\,\phi_{A,j}\,.$$

We take a shorthand $\eta := \frac{2\gamma}{(1-\gamma+2\gamma p)^2}\sqrt{\frac{p\,(1-p)}{120}}$. Note that $\{\sqrt{\mu_j}\,\phi_{A,j}\}_{j=1}^{\infty}$ is an orthonormal basis in RKHS $\mathbb{H}_A$ and $\delta_n^2 \leq \mu_j$ for any $j \in [d_n]$; as a consequence, we have

(91)
$$\|\theta_m^* - r_A\|_{\mathbb{H}_A}^2 = \Big\{\frac{\gamma\,(1-2p)}{1-\gamma+2\gamma p}\Big\}^2 \frac{1}{\mu_1} + \frac{\eta^2}{n}\sum_{j=2}^{d_n}\frac{(\alpha_m^{(j-1)})^2}{\mu_j} \leq \Big\{\frac{\gamma\,(1-2p)}{1-\gamma+2\gamma p}\Big\}^2 \frac{1}{\mu_1} + \frac{\eta^2\,d_n}{n\delta_n^2}\,.$$

Since $p = \frac{3(1-\gamma)}{\gamma}$, the first term in the upper bound above satisfies

(92a)
$$\Big\{\frac{\gamma\,(1-2p)}{1-\gamma+2\gamma p}\Big\}^2 \frac{1}{\mu_1} = \Big\{\frac{7\gamma-6}{7(1-\gamma)}\Big\}^2 \frac{1}{\mu_1} \leq \Big\{\frac{\gamma}{7(1-\gamma)\sqrt{\mu_1}}\Big\}^2 \leq \Big\{\frac{6\bar{R}}{7}\Big\}^2,$$

where we have used the relation $\bar{R} \geq \frac{\gamma}{6(1-\gamma)\sqrt{\mu_1}}$. As for the second term in the right hand side of inequality (91), we recall that the critical inequality (33) ensures $\frac{d_n}{n\delta_n^2} \leq \big\{\frac{\bar{R}(1-\gamma)}{\bar{\sigma}}\big\}^2$, therefore,

$$\frac{\eta^2\,d_n}{n\delta_n^2} \leq \bar{R}^2\,\big\{\eta(1-\gamma)/\bar{\sigma}\big\}^2.$$

Combining the definition of η, the equality $p = \frac{3(1-\gamma)}{\gamma}$ and the relation $\bar{\sigma}^2 \geq \frac{1+\gamma}{5(1-\gamma)}$, we find that $\eta(1-\gamma)/\bar{\sigma} \leq \frac{1}{98}$. It follows that

$$\frac{\eta^2\,d_n}{n\delta_n^2} \leq \Big\{\frac{\bar{R}}{98}\Big\}^2.$$

Plugging inequalities (92) into (91) yields $\|\theta_m^* - r_A\|_{\mathbb{H}_A} \leq \bar{R}$.

We now estimate the sup-norm $\|\theta_m^*\|_\infty$. We use the inequality $\|f_m\|_\infty \leq \frac{p}{9}$ and find that

$$\begin{aligned}\|\theta_m^*\|_\infty &\leq \frac{1}{1-\gamma+2\gamma p}\,\|\phi_{A,1}\|_\infty + \frac{2\gamma}{(1-\gamma+2\gamma p)^2}\,\|f_m\|_\infty \\ &\leq \frac{1}{1-\gamma+2\gamma p} + \frac{2\gamma}{(1-\gamma+2\gamma p)^2}\,(p/9) \leq \frac{1}{6(1-\gamma)}\,.\end{aligned}$$

Since $\bar{R} \geq \frac{2}{3b(1-\gamma)}$, we have $\frac{2\|\theta_m^*\|_\infty}{b} \leq \bar{R}$.

Integrating the two parts, we conclude that $\max\big\{\|\theta_m^* - r_A\|_{\mathbb{H}_A}, \frac{2\|\theta_m^*\|_\infty}{b}\big\} \leq \bar{R}$.

D.3.2. *Proof of Lemma D.2.* Since the n samples $\{(x_i, x_i')\}_{i=1}^n$ are i.i.d., we have

$$D_{\mathrm{KL}}\big(\mathcal{P}_{m'}^{1:n} \,\big\|\, \mathcal{P}_m^{1:n}\big) = n\, D_{\mathrm{KL}}(\mathcal{P}_{m'} \,\|\, \mathcal{P}_m)\,.$$

Thus, the remainder of our proof focuses on bounding $D_{\mathrm{KL}}(\mathcal{P}_{m'} \,\|\, \mathcal{P}_m)$, for an arbitrary pair $m, m' \in [M]$.

From Jensen's inequality and the concavity of the logarithm, the KL divergence can be upper bounded by the χ^2-divergence—that is

$$D_{\mathrm{KL}}(\mathcal{P}_{m'} \,\|\, \mathcal{P}_m) \le \chi^2(\mathcal{P}_{m'} \,\|\, \mathcal{P}_m) = \int_{\mathcal{X}^2} \mu(x)\, \frac{\big(\mathcal{P}_{m'}(x' \mid x) - \mathcal{P}_m(x' \mid x)\big)^2}{\mathcal{P}_m(x' \mid x)}\, dx\, dx'.$$

Recall our shorthand notation $K = 2^{\lceil \log_2 d_n \rceil}$, where the kernel dimension d_n was previously defined as $d_n = \max\big\{j \mid \mu_j \ge \delta_n^2\big\}$. Since our construction of transition model $\mathcal{P}_m$ is an ensemble of K blocks $\{\mathbf{P}_m^{(k)}\}_{k=1}^K$, each involving two states, the χ^2-divergence can be written as the sum

$$\chi^2(\mathcal{P}_{m'} \,\|\, \mathcal{P}_m) = \frac{1}{K} \sum_{k=1}^{K} \chi^2\big(\mathbf{P}_{m'}^{(k)} \,\big\|\, \mathbf{P}_m^{(k)}\big)\,. \tag{93}$$

The local χ^2-divergence is defined as

$$\chi^2\big(\mathbf{P}_{m'}^{(k)} \,\big\|\, \mathbf{P}_m^{(k)}\big) := \sum_{x,x' \in \{x_+, x_-\}} \boldsymbol{\mu}(x)\, \frac{\big(\mathbf{P}_{m'}^{(k)}(x' \mid x) - \mathbf{P}_m^{(k)}(x' \mid x)\big)^2}{\mathbf{P}_m^{(k)}(x' \mid x)}$$

where $\boldsymbol{\mu} := \big[\frac{1}{2}, \frac{1}{2}\big]$ is the stationary distribution. We recall from equation (72b) the expression of local model $\mathbf{P}_m^{(k)} = \mathbf{P}_A\big(p, \Delta p_m^{(k)}\big)$ and derive that

$$\chi^2\big(\mathbf{P}_{m'}^{(k)} \,\big\|\, \mathbf{P}_m^{(k)}\big) \;=\; \frac{\big(\Delta p_{m'}^{(k)} - \Delta p_m^{(k)}\big)^2}{\big(p + \Delta p_m^{(k)}\big)\big(1 - p - \Delta p_m^{(k)}\big)}\,. \tag{94}$$

We develop upper and lower bounds on the numerator and denominator separately.

We first consider the numerator $\big(\Delta p_{m'}^{(k)} - \Delta p_m^{(k)}\big)^2$. Following some algebra, we find that

$$\Delta p_{m'}^{(k)} - \Delta p_m^{(k)} = \frac{f_m(x_k) - f_{m'}(x_k)}{\big(1 - \frac{2\gamma\, f_{m'}(x_k)}{1-\gamma+2\gamma p}\big)\big(1 - \frac{2\gamma\, f_m(x_k)}{1-\gamma+2\gamma p}\big)}\,,$$

where x_k is any point in interval $\Delta_+^{(k)}$. It was shown in the bound (90) that $\|f_m\|_\infty \le p/9$, therefore, $\min\big\{1 - \frac{2\gamma\, f_{m'}(x_k)}{1-\gamma+2\gamma p}, 1 - \frac{2\gamma\, f_m(x_k)}{1-\gamma+2\gamma p}\big\} \ge \frac{19}{21}$. It follows that

$$\big(\Delta p_{m'}^{(k)} - \Delta p_m^{(k)}\big)^2 \le 2\,\big(f_m(x_k) - f_{m'}(x_k)\big)^2. \tag{95a}$$

As for the numerator $\big(p + \Delta p_m^{(k)}\big)\big(1 - p - \Delta_m^{(k)}\big)$ in the right hand side of equality (94), we have proved in the bound (90) that $\big|\Delta p_m^{(k)}\big| \le p/8$, so that

$$\big(p + \Delta p_m^{(k)}\big)\big(1 - p - \Delta_m^{(k)}\big) \ge \tfrac{7}{8}\, p(1-p)\,. \tag{95b}$$

Combining inequalities (95) with equation (94) yields

$$\chi^2\big(\mathbf{P}_{m'}^{(k)} \,\big\|\, \mathbf{P}_m^{(k)}\big) \le \frac{3\,\big(f_{m'}(x_k) - f_m(x_k)\big)^2}{p(1-p)}\,. \tag{96}$$

We plug the bound (96) into equation (93) and find that

$$\chi^2(\mathcal{P}_{m'} \,\|\, \mathcal{P}_m) \le \frac{3}{p(1-p)} \left\{ \frac{1}{K} \sum_{k=1}^{K} \big(f_{m'}(x_k) - f_m(x_k)\big)^2 \right\}$$
$$\overset{(i)}{=} \frac{3}{p(1-p)} \int_{\mathcal{X}} \big(f_{m'}(x) - f_m(x)\big)^2 dx = \frac{3}{p(1-p)} \, \|f_{m'} - f_m\|_{\mu}^2 \,.$$

Here step (i) is due to the property that $f_m(x) = f_m(x_k)$ for any $x \in \Delta_k^+$ and $f_m(x) = -f_m(x_k)$ for any $x \in \Delta_k^-$. Regarding the $L^2(\mu)$-distance $\|f_m - f_{m'}\|_{\mu}$, we leverage the orthonormality of basis functions $\{\phi_{A,j}\}_{j=1}^{d_n}$ in $L^2(\mu)$ and find that

$$\|f_{m'} - f_m\|_{\mu}^2 = \frac{p(1-p)}{120n} \sum_{j=2}^{d_n} \big(\alpha_m^{(j-1)} - \alpha_{m'}^{(j-1)}\big)^2 \le p(1-p) \, \frac{d_n}{120n} \,.$$

Therefore, we have

$$\chi^2(\mathcal{P}_{m'} \,\|\, \mathcal{P}_m) \le \frac{d_n}{40n}.$$

Putting together the pieces yields

$$D_{\mathrm{KL}}\big(\mathcal{P}_{m'}^{1:n} \,\big\|\, \mathcal{P}_m^{1:n}\big) \le n \, \chi^2(\mathcal{P}_{m'} \,\|\, \mathcal{P}_m) \le \frac{d_n}{40},$$

as claimed in the lemma statement.

D.3.3. *Proof of Lemma D.3.* We now lower bound the $L^2(\mu)$-norm between the value functions of different models in our family. Recall the expression of value function θ_m^* in equation (82). We find that

$$\theta_m^* = \theta_0^* + \frac{\eta}{\sqrt{n}} \sum_{j=2}^{d_n} \alpha_m^{(j-1)} \, \phi_{A,j}$$

where $\eta = \frac{2\gamma}{(1-\gamma+2\gamma p)^2} \sqrt{\frac{p\,(1-p)}{120}}$. Since $\{\phi_{A,j}\}_{j=1}^{\infty}$ is an orthonormal basis in $L^2(\mu)$, we can write

$$\|\theta_{m'}^* - \theta_m^*\|_{\mu}^2 = \frac{\eta^2}{n} \sum_{j=2}^{d_n} \big(\alpha_{m'}^{(j-1)} - \alpha_m^{(j-1)}\big)^2.$$

By our construction, $\{\boldsymbol{\alpha}_m\}_{m=1}^{M}$ is a $\frac{1}{4}$-packing of the Boolean hypercube $\{0,1\}^{d_n-1}$ with respect to the rescaled Hamming distance, therefore,

$$\sum_{j=2}^{d_n} \big(\alpha_{m'}^{(j-1)} - \alpha_m^{(j-1)}\big)^2 \ge \frac{d_n - 1}{4} \,.$$

We use the conditions $p = \frac{3(1-\gamma)}{\gamma}$, $\gamma \in [0.9, 1)$ and $\bar{\sigma}^2 \le \frac{1+\gamma}{1-\gamma}$, and find by some algebra that

$$\eta \ge \frac{c_2'}{1-\gamma} \sqrt{\frac{1+\gamma}{1-\gamma}} \ge \frac{c_2' \, \bar{\sigma}}{1-\gamma}$$

where $c_2' > 0$ is a universal constant. Combining the inequalities, we obtain

$$\|\theta_{m'}^* - \theta_m^*\|_{\mu} \ge \frac{c_1' \, \bar{\sigma}}{1-\gamma} \sqrt{\frac{d_n}{n}}$$

for another universal constant $c_1' > 0$. By further using the regularity condition (34), we can derive inequality (89) in the lemma statement.

D.4. Proofs of auxiliary results in Regime B. This section contains proofs of auxiliary results that underlie the minimax lower bound over model family $\mathfrak{M}_B$. Specifically, Appendix D.4.1 proves the density ratio condition (45), that is, $\frac{d\mu_m}{d\mu}(x) \geq \frac{1}{2}$. Appendix D.4.2 is devoted to the proof of Lemma D.1, which shows that our constructed models $\{\mathscr{I}_m\}_{m=1}^M$ belong to the family $\mathfrak{M}_B$. Appendix D.4.3 proves Lemma D.2, which upper bounds the pairwise KL-divergence. Appendix D.4.4 presents the proof of Lemma D.3, which estimates the pairwise distance in value functions.

D.4.1. *Proof of density ratio condition.* We prove the density ratio condition (45), i.e. $\frac{d\mu_m}{d\mu}(x) \geq \frac{1}{2}$. Recall our definition of f_m in equation (85). Since $\sup_{j\geq 1}\|\phi_{B,j}\|_\infty \leq \kappa$, we have

$$\|f_m\|_\infty \leq \frac{\kappa\, p\, d_n}{25\sqrt{n}} \overset{(i)}{\leq} n\left\{\frac{\bar{R}\delta_n(1-\gamma)}{\kappa\bar{\sigma}}\right\}^2 \frac{\kappa\, p}{25\sqrt{n}} \overset{(ii)}{\leq} \frac{p\,(1-\gamma)}{2}. \tag{97}$$

Here step (i) is due to the critical inequality (33) and in step (ii) we have used the inequality $\bar{R}^2\delta_n^2 \leq \frac{12\,\kappa\bar{\sigma}^2}{(1-\gamma)\sqrt{n}}$ in condition (36b). We plug inequality (97) into the expression of stationary distribution μ_m in equation (87). It follows that $\frac{d\mu_m}{d\mu}(x) \geq \frac{1}{2}$, as claimed.

D.4.2. *Proof of Lemma D.1.* By our construction, condition $\theta_m^* \in \mathbb{H}_B$ in equation (44b) naturally holds. In the sequel, we verify the remaining constraints in the definition of $\mathfrak{M}_B$, including

- the regularity condition $\gamma\|\theta_m^*\|_{\mu_m} \leq 1$ and the Bellman residual variance bound $\sigma^2(\theta_m^*) \leq \bar{\sigma}^2$,
- the norm condition $\max\left\{\|\theta_m^* - r\|_{\mathbb{H}_B}, \frac{2\|\theta_m^*\|_\infty}{b}\right\} \leq \bar{R}$,
- the property that the covariance operator $\Sigma_{\mathrm{cov}}(\mathcal{P}_m)$ has eigenpairs $\left\{\left(\mu_j(\mathcal{P}_m), \phi_j(\mathcal{P}_m)\right)\right\}_{j=1}^\infty$ with $\mu_j(\mathcal{P}_m) \leq \mu_j$ for $j \geq 2$ and $\sup_j \left\|\phi_j(\mathcal{P}_m)\right\|_\infty \leq 2 = \kappa$.

Upper bounds on $\gamma\|\theta_m^*\|_{\mu_m}$ and $\sigma^2(\theta_m^*)$. The MRP $\mathscr{I}_m$ consists of K blocks, each is a small local MRP determined by the transition matrix $\mathbf{P}_m^{(k)} = \mathbf{P}_B\big(p, \Delta p_m^{(k)}\big)$ and a reward vector $\boldsymbol{r} = \left\{p + \frac{1-\gamma}{2\gamma}\right\}[1,-1]^\top$. The stationary distribution of $\mathbf{P}_m^{(k)}$ takes the form

$$\boldsymbol{\mu}_m^{(k)} = \left[\tfrac{1}{2} + \tfrac{\Delta p_m^{(k)}}{2p}, \tfrac{1}{2} - \tfrac{\Delta p_m^{(k)}}{2p}\right]. \tag{98}$$

The $\boldsymbol{\mu}_m^{(k)}$-weighted norm of value function $\boldsymbol{\theta}_m^{(k)}$ and the variance term $\sigma^2(\boldsymbol{\theta}_m^{(k)})$ satisfy

$$\gamma^2\left\|\boldsymbol{\theta}_m^{(k)}\right\|_{\boldsymbol{\mu}_m^{(k)}}^2 = \frac{1}{4} + \frac{\gamma\,(1-\gamma+\gamma p)}{p\,(1-\gamma)^2}\left(\Delta p_m^{(k)}\right)^2 \quad \text{and}$$

$$\sigma^2(\boldsymbol{\theta}_m^{(k)}) = \frac{1-p}{p}\left\{p^2 - \left(\Delta p_m^{(k)}\right)^2\right\} \leq p\,(1-p).$$

The squared $L^2(\mu_m)$-norm of the full-scale value function θ_m^* is the average of $\left\|\boldsymbol{\theta}_m^{(k)}\right\|_{\boldsymbol{\mu}_m^{(k)}}^2$ over indices $k \in [K]$. We use the relation $\Delta p_m^{(k)} = f_m(x_k)$ and find that

$$\gamma^2\|\theta_m^*\|_{\mu_m}^2 = \frac{1}{4} + \frac{1}{K}\sum_{k=1}^K \frac{\gamma\,(1-\gamma+\gamma p)}{p\,(1-\gamma)^2} f_m^2(x_k) = \frac{1}{4} + \frac{\gamma\,(1-\gamma+\gamma p)}{p\,(1-\gamma)^2}\|f_m\|_\mu^2. \tag{99a}$$

Due to the orthonormality of bases $\{\phi_{B,j}\}_{j=2}^{d_n}$ in $L^2(\mu)$, we have

$$\|f_m\|_\mu = \frac{p}{25\sqrt{n}}\|\boldsymbol{\alpha}_m\|_2 \leq \frac{p}{25}\sqrt{\frac{d_n}{n}}.$$

According to the critical inequality (33), it holds that

$$\|f_m\|_{\mu} \leq \frac{p}{25}\sqrt{\frac{d_n}{n}} \leq \frac{p}{25}\left\{\frac{\bar{R}\delta_n(1-\gamma)}{\kappa\bar{\sigma}}\right\} \overset{(i)}{\leq} \frac{2}{5}\,p\,(1-\gamma)\,. \tag{99b}$$

In step (i), we have used the inequality $\bar{R}\delta_n \leq 10\,\kappa\bar{\sigma}$, which is implied by condition (36b). We plug inequality (99b) into equation (99a) and conclude that $\gamma\|\theta_m^*\|_{\mu_m} \leq 1$. Therefore, the regularity condition in Regime B is satisfied.

Similarly, we calculate the variance term $\sigma^2(\theta_m^*)$ by taking the average of $\{\sigma^2(\boldsymbol{\theta}_m^{(k)})\}_{k=1}^K$. It follows that $\sigma^2(\theta_m^*) \leq p\,(1-p)$. Since $\bar{\sigma}^2 \geq \frac{1}{8} = p$, we have $\sigma(\theta_m^*) \leq \bar{\sigma}$, as required by equation (32b).

Upper bounds on $\|\theta_m^* - r_B\|_{\mathbb{H}_B}$ and $\|\theta_m^*\|_\infty$. We first consider the RKHS norm $\|\theta_m^* - r_B\|_{\mathbb{H}_B}$. Recall that the reward and value functions r_B and θ_m^* take the form

$$r_B = \left\{p + \frac{1-\gamma}{2\gamma}\right\}\phi_{B,1}\,,$$

$$\theta_m^* = \frac{1}{2\gamma}\,\phi_{B,1} + \frac{1}{1-\gamma}\,f_m = \frac{1}{2\gamma}\,\phi_{B,1} + \frac{p}{25(1-\gamma)\sqrt{n}}\sum_{j=2}^{d_n}\alpha_m^{(j-1)}\,\phi_{B,j}\,.$$

Since $\{\sqrt{\mu_j}\,\phi_{B,j}\}_{j=1}^\infty$ is an orthonormal basis of $\mathbb{H}_B$, we use the property that $\mu_j \geq \delta_n^2$ for any $j \leq d_n$ and find that

$$\|\theta_m^* - r_B\|_{\mathbb{H}_B}^2 = \frac{(\frac{1}{2}-p)^2}{\mu_1} + \frac{p^2}{25^2\,(1-\gamma)^2\,n}\sum_{j=2}^{d_n}\frac{(\alpha_m^{(j-1)})^2}{\mu_j} \leq \frac{(\frac{1}{2}-p)^2}{\mu_1} + \frac{p^2\,d_n}{25^2\,(1-\gamma)^2\,n\delta_n^2}\,.$$

The critical inequality (33) ensures $\frac{d_n}{n\delta_n^2} \leq \left\{\frac{\bar{R}(1-\gamma)}{\kappa\bar{\sigma}}\right\}^2$ and implies

$$\|\theta_m^* - r_B\|_{\mathbb{H}_B}^2 \leq \frac{(\frac{1}{2}-p)^2}{\mu_1} + \frac{p^2}{25^2\kappa^2\bar{\sigma}^2}\bar{R}^2 \leq \frac{9}{64\mu_1} + \frac{\bar{R}^2}{25^2} \leq \bar{R}^2\,, \tag{100a}$$

where we have used the properties $\frac{1}{8} = p \leq \bar{\sigma}^2 \leq 1$, $\kappa \geq 1$ and $\frac{1}{\sqrt{\mu_1}} \leq 2\bar{R}$.

As for the upper bound on sup-norm $\|\theta_m^*\|_\infty$, we apply the estimation of $\|f_m\|_\infty$ in inequality (97) and find that

$$\|\theta_m^*\|_\infty \leq \tfrac{1}{2\gamma}\,\|\phi_{B,1}\|_\infty + \tfrac{1}{1-\gamma}\,\|f_m\|_\infty \leq \tfrac{1}{2\gamma} + \tfrac{p}{2} \leq \tfrac{1}{\gamma}\,. \tag{100b}$$

Therefore, it holds that $\frac{2\|\theta_m^*\|_\infty}{b} \leq \bar{R}$.

Combining inequalities (100a) and (100b), we conclude that

$$\max\left\{\|\theta_m^* - r_B\|_{\mathbb{H}_B}, \tfrac{2\|\theta_m^*\|_\infty}{b}\right\} \leq \bar{R}.$$

Analysis of eigenpairs $\left\{\left(\mu_j(\mathcal{P}_m), \phi_j(\mathcal{P}_m)\right)\right\}_{j=1}^\infty$. Recall that $\Sigma_{\text{cov}}(\mathcal{P}_m)$ is the covariance operator of kernel $\mathcal{K}_B$ associated with distribution $\mu_m = \mu(\mathcal{P}_m)$, $\{\mu_j(\mathcal{P}_m)\}_{j=1}^\infty$ are the eigenvalues of $\Sigma_{\text{cov}}(\mathcal{P}_m)$ arranged in non-increasing order, and $\phi_j(\mathcal{P}_m)$ is the eigenfunction corresponding to $\mu_j(\mathcal{P}_m)$. In the following Lemma D.4, we develop upper bounds on the eigenvalues and the sup-norms of the eigenfunctions.

LEMMA D.4. *Under our construction of kernel $\mathcal{K}_B$ and MRP instances $\{\mathscr{I}_m\}_{m=1}^M \subset \mathfrak{M}_B$ in Regime B, for any $m \in [M]$, the eigenpairs $\left\{\left(\mu_j(\mathcal{P}_m), \phi_j(\mathcal{P}_m)\right)\right\}_{j=1}^\infty$ satisfy the claims below:*

(a) *It holds that $\mu_j(\mathcal{P}_m) \leq \mu_j$ for any $j \geq 2$.*

(b) *Suppose* $\min_{3\le j\le d_n}\{\sqrt{\mu_{j-1}}-\sqrt{\mu_j}\}\ge\frac{\delta_n}{2d_n}$ *and the sample size* n *is sufficiently large such that condition* (36b) *holds. Then the eigenfunctions* $\{\phi_j(\mathcal{P}_m)\}_{j=1}^{\infty}$ *satisfy* $\sup_{j\in\mathbb{Z}_+}\|\phi_j(\mathcal{P}_m)\|_\infty\le 2$.

We establish the proof of Lemma D.4 by first connecting the eigenpairs $\{(\mu_j(\mathcal{P}_m),\phi_j(\mathcal{P}_m))\}_{j=1}^{\infty}$ to the spectrum of an arrowhead matrix, and then developing the desired bounds based on properties of the matrix. See Appendix E.2 for the details.

D.4.3. *Proof of Lemma D.2.* Similar to the proof in Appendix D.3.2, we also upper bound the KL-divergence $D_{\mathrm{KL}}(\mathcal{P}_{m'}^{1:n}\,\|\,\mathcal{P}_m^{1:n})$ by the average of χ^2-divergences between local models $\mathbf{P}_{m'}^{(k)}$ and $\mathbf{P}_m^{(k)}$. The calculation of local χ^2-divergence in the MRPs $\{\mathscr{I}_m\}_{m=1}^M\subset\mathfrak{M}_B$ is different from that in Appendix D.3.2, since the stationary distributions $\boldsymbol{\mu}_m^{(k)}$ and $\boldsymbol{\mu}_{m'}^{(k)}$ (given in equation (98)) are unequal. In particular, the local χ^2-divergence takes the form

$$\chi^2(\mathbf{F}_{m'}^{(k)}\,\|\,\mathbf{F}_m^{(k)})=\sum_{x,x'\in\{x_+,x_-\}}\frac{\big(\mathbf{F}_{m'}^{(k)}(x'\,|\,x)-\mathbf{F}_m^{(k)}(x'\,|\,x)\big)^2}{\mathbf{F}_m^{(k)}(x'\,|\,x)}\tag{101}$$

where the matrix $\mathbf{F}_\iota^{(k)}:=\big[\mathrm{diag}\,(\boldsymbol{\mu}_\iota^{(k)})\big]\,\mathbf{P}_\iota^{(k)}\in\mathbb{R}^{2\times 2}$ for $\iota=m$ or m' .

We learn from inequality (97) that $\|f_m\|_\infty\le\frac{p}{2}$, therefore, $|\Delta p_m^{(k)}|\le\frac{p}{2}$ in local Markov chain $\mathbf{P}_m^{(k)}=\mathbf{P}_B(p,\Delta p_m^{(k)})$. It follows that $\boldsymbol{\mu}_m^{(k)}(x)\ge\frac14$ and $\mathbf{P}_m^{(k)}(x'\,|\,x)\ge\frac12\mathbf{P}_0(x'\,|\,x)$ for any $x,x'\in\{x_+,x_-\}$. Here, $\mathbf{P}_0$ is the base Markov chain defined in equation (72a). These lower bounds imply that

$$\mathbf{F}_m^{(k)}(x'\,|\,x)\ge\frac18\,\mathbf{P}_0(x'\,|\,x)\qquad\text{for } x,x'\in\{x_+,x_-\}.$$

Substituting the above inequality into equation (101) yields

$$\chi^2(\mathbf{F}_{m'}^{(k)}\,\|\,\mathbf{F}_m^{(k)})\le 8\sum_{x,x'\in\{x_+,x_-\}}\frac{\big(\mathbf{F}_{m'}^{(k)}(x'\,|\,x)-\mathbf{F}_m^{(k)}(x'\,|\,x)\big)^2}{\mathbf{P}_0(x'\,|\,x)}.$$

We use some algebra and derive that

$$\chi^2(\mathbf{F}_{m'}^{(k)}\,\|\,\mathbf{F}_m^{(k)})\le\frac{8\,(\Delta p_{m'}^{(k)}-\Delta p_m^{(k)})^2}{p^3(1-p)}\big\{p+(\Delta p_{m'}^{(k)}+\Delta p_m^{(k)})^2\big\}\overset{(i)}{\le}\frac{16\,(\Delta p_{m'}^{(k)}-\Delta p_m^{(k)})^2}{p^2(1-p)}.$$

In step (i) above, we have used the relation $\max\{|\Delta p_m^{(k)}|,|\Delta p_{m'}^{(k)}|\}\le\frac{p}{2}$ once again.

Collecting all the local χ^2-divergences yields

$$D_{\mathrm{KL}}(\mathcal{P}_{m'}^{1:n}\,\|\,\mathcal{P}_m^{1:n})\le\frac{n}{K}\sum_{k=1}^K\chi^2(\mathbf{F}_{m'}^{(k)}\,\|\,\mathbf{F}_m^{(k)})\le\frac{16\,n}{p^2(1-p)}\bigg\{\frac1K\sum_{k=1}^K(\Delta p_{m'}^{(k)}-\Delta p_m^{(k)})^2\bigg\}.$$

Recall that by our construction, $f_\iota(x)=\Delta p_\iota^{(k)}$ for any $x\in\Delta_+^{(k)}\cup\Delta_-^{(k)}$ and $\iota=m$ or m', therefore, it holds that

$$\frac1K\sum_{k=1}^K(\Delta p_{m'}^{(k)}-\Delta p_m^{(k)})^2=\int_{\mathcal{X}}\big(f_{m'}(x)-f_m(x)\big)^2\,dx=\|f_{m'}-f_m\|_{\boldsymbol{\mu}}^2.$$

Due to the orthogonality of basis $\{\phi_{B,j}\}_{j=1}^{\infty}$ in $L^2(\mu)$, the definitions of $f_{m'}$ and f_m in equation (85) imply

$$\|f_{m'} - f_m\|_\mu^2 = \frac{p^2}{625\, n} \sum_{j=2}^{d_n} \left(\alpha_{m'}^{(j-1)} - \alpha_m^{(j-1)}\right)^2 \leq \frac{p^2\, d_n}{625\, n}. \tag{102}$$

Putting together the pieces, we prove that $D_{\mathrm{KL}}\left(\mathcal{P}_{m'}^{1:n} \,\middle\|\, \mathcal{P}_m^{1:n}\right) \leq \frac{d_n}{40}$, as claimed.

D.4.4. *Proof of Lemma D.3.* Due to the definitions of θ_m^* and $\theta_{m'}^*$ in equation (85), we find that

$$\|\theta_{m'}^* - \theta_m^*\|_\mu = \frac{1}{1-\gamma} \|f_{m'} - f_m\|_\mu .$$

We recall from equation (102) that the $L^2(\mu)$-difference $\|f_{m'} - f_m\|_\mu$ can be expressed by vectors $\boldsymbol{\alpha}_m$ and $\boldsymbol{\alpha}_{m'}$. Using the property that $\left\{\boldsymbol{\alpha}_m\right\}_{m=1}^M$ is a $\frac{1}{4}$-packing of the Boolean hypercube $\{0,1\}^{d_n - 1}$, we find that

$$\|f_{m'} - f_m\|_\mu^2 = \frac{p^2}{625\, n} \sum_{j=2}^{d_n} \left(\alpha_{m'}^{(j-1)} - \alpha_m^{(j-1)}\right)^2 \geq \frac{p^2}{25^2\, n} \frac{d_n - 1}{4},$$

Plugging the lower bound on $\|f_{m'} - f_m\|_\mu$ into the expression of $\|\theta_{m'}^* - \theta_m^*\|_\mu$, we have

$$\|\theta_{m'}^* - \theta_m^*\|_\mu \geq \frac{p}{50\,(1-\gamma)} \sqrt{\frac{d_n - 1}{n}}.$$

It follows from the conditions $\bar{\sigma} \leq 1$, $p = \frac{1}{8}$ and $\kappa = 2$ that

$$\|\theta_{m'}^* - \theta_m^*\|_\mu \geq \frac{c_1' \kappa \bar{\sigma}}{1-\gamma} \sqrt{\frac{d_n}{n}}$$

for some universal constant $c_1' > 0$. Under the regularity condition (34), the above lower bound further implies inequality (89) in the lemma statement.

APPENDIX E: PROOF OF TECHNICAL LEMMAS

In this appendix, we collect together various technical lemmas.

E.1. A kernel-based computation. Here we provide an explicit expression for the kernel LSTD estimate in terms of kernel matrices. Define the kernel covariance matrix $\mathbf{K}_{\mathrm{cov}} \in \mathbb{R}^{n \times n}$ and cross-covariance matrix $\mathbf{K}_{\mathrm{cr}} \in \mathbb{R}^{n \times n}$ with entries

$$\mathbf{K}_{\mathrm{cov}}(i,j) = \mathcal{K}(x_i, x_j)/n\,, \qquad \text{and} \qquad \mathbf{K}_{\mathrm{cr}}(i,j) = \mathcal{K}(x_i, x_j')/n \quad \text{for } i,j = 1,\ldots,n\,. \tag{103}$$

The following lemma yields an explicit linear-algebraic expression for the solution:

LEMMA E.1 (Kernel-based computation). *The LSTD estimator $\widehat{\theta}$ takes the form*

$$\widehat{\theta} = r + \frac{\gamma}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i\, \mathcal{K}(\cdot, x_i)\,, \tag{104}$$

where the coefficient vector $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^n$ is the solution to the linear system

$$\left(\mathbf{K}_{\mathrm{cov}} + \lambda_n \mathbf{I}_n - \gamma \mathbf{K}_{\mathrm{cr}}^\top\right) \widehat{\boldsymbol{\alpha}} \;=\; \boldsymbol{y}\,. \tag{105}$$

Here $\boldsymbol{y} \in \mathbb{R}^n$ has entries $y_i = r(x_i')/\sqrt{n}$.

PROOF. We first show that function $(\widehat{\theta} - r)$ can be linearly expressed by the representers of evaluation $\{\Phi_{x_i}\}_{i=1}^n$ as in equation (104). Take a linear subspace $\widehat{\mathbb{H}}$ of $\mathbb{H}$ that is spanned by representer functions $\{\Phi_{x_i}\}_{i=1}^n$. By denoting $\widetilde{\Sigma}_{\text{cov}} := \widehat{\Sigma}_{\text{cov}} + \lambda_n \mathcal{I}$, we recast equation (10) into

$$\widetilde{\Sigma}_{\text{cov}}(\widehat{\theta} - r) = \widehat{\Sigma}_{\text{cr}}\widehat{\theta}\,. \tag{106}$$

The right hand side satisfies $\widehat{\Sigma}_{\text{cr}}\widehat{\theta} \in \widehat{\mathbb{H}}$ by definition. As long as we can show that

$$\widetilde{\Sigma}_{\text{cov}}^{-1}\widehat{\mathbb{H}} \subset \widehat{\mathbb{H}}\,, \tag{107}$$

it follows from equation (106) that $\widehat{\theta} - r = \widetilde{\Sigma}_{\text{cov}}^{-1}\big(\widehat{\Sigma}_{\text{cr}}\widehat{\theta}\big) \in \widehat{\mathbb{H}}$, which then implies the existence of a coefficient vector $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^n$ such that

$$\widehat{\theta} = r + \frac{\gamma}{\sqrt{n}}\sum_{i=1}^n \widehat{\alpha}_i \Phi_{x_i} = r + \frac{\gamma}{\sqrt{n}}\sum_{i=1}^n \widehat{\alpha}_i \mathcal{K}(\cdot, x_i)\,.$$

We now prove the relation (107) by contradiction. In fact, if there exists a function $f \in \widehat{\mathbb{H}}$ such that $g = \widetilde{\Sigma}_{\text{cov}}^{-1} f \notin \widehat{\mathbb{H}}$, then $\widetilde{\Sigma}_{\text{cov}} g = \frac{1}{n}\sum_{i=1}^n \Phi_{x_i} g(x_i) + \lambda_n g \notin \widehat{\mathbb{H}}$, which contradicts the condition $f \in \widehat{\mathbb{H}}$.

Below we derive the explicit form of vector $\widehat{\boldsymbol{\alpha}}$. We take a shorthand $\widehat{f} := \gamma^{-1}(\widehat{\theta} - r) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\alpha}_i \mathcal{K}(\cdot, x_i)$. It follows from equation (106) that

$$\big(\widehat{\Sigma}_{\text{cov}} - \gamma\widehat{\Sigma}_{\text{cr}}\big)\widehat{f} + \lambda_n \widehat{f} = \widehat{\Sigma}_{\text{cr}} r\,. \tag{108}$$

Plugging the definitions of $\widehat{\Sigma}_{\text{cov}}$ and $\widehat{\Sigma}_{\text{cr}}$ into equation (108), we find that the left hand side equals

$$\begin{aligned}
&\frac{1}{n\sqrt{n}}\sum_{i=1}^n \Phi_{x_i}\sum_{j=1}^n \widehat{\alpha}_j\big(\mathcal{K}(x_i, x_j) - \gamma\mathcal{K}(x_i', x_j)\big) + \frac{\lambda_n}{\sqrt{n}}\sum_{i=1}^n \widehat{\alpha}_i \Phi_{x_i} \\
&= \frac{1}{\sqrt{n}}\big[\Phi_{x_1}, \Phi_{x_2}, \ldots, \Phi_{x_n}\big]\big(\mathbf{K}_{\text{cov}} - \gamma\mathbf{K}_{\text{cr}}^\top + \lambda_n \mathbf{I}_n\big)\widehat{\boldsymbol{\alpha}}\,.
\end{aligned}$$

The right hand side of equation (108) takes the form

$$\frac{1}{n}\sum_{i=1}^n \Phi_{x_i} r(x_i') = \frac{1}{\sqrt{n}}\big[\Phi_{x_1}, \Phi_{x_2}, \ldots, \Phi_{x_n}\big]\boldsymbol{y}\,.$$

Comparing both sides, we have shown that the coefficient vector $\widehat{\boldsymbol{\alpha}}$ satisfies the linear system (105), thereby completing the proof. □

E.2. Proof of Lemma D.4. We observe that the distribution μ_m is relatively close to the uniform measure μ over $[0, 1)$. Therefore, we expect that the eigenspectra of $\Sigma_{\text{cov}}(\mathcal{P}_m)$ and Σ_{cov} should be similar, where $\Sigma_{\text{cov}}(\mathcal{P}_m)$ and Σ_{cov} are the covariance operators associated with distributions μ_m and μ respectively. Recall that by our construction of the kernel $\mathcal{K}_B$ in equation (80), Σ_{cov} has eigenpairs $\{(\mu_j, \phi_{B,j})\}_{j=1}^\infty$. In the following, we expand $\Sigma_{\text{cov}}(\mathcal{P}_m)$ using the basis functions $\{\phi_{B,j}\}_{j=1}^\infty$, which yields an arrowhead matrix $\boldsymbol{\Sigma}$. We take shorthands $\tilde{\mu}_j \equiv \mu_j(\mathcal{P}_m)$ and $\tilde{\phi}_j \equiv \phi_j(\mathcal{P}_m)$, and connect the eigenpairs $\{(\tilde{\mu}_j, \tilde{\phi}_j)\}_{j=1}^\infty$ of $\Sigma_{\text{cov}}(\mathcal{P}_m)$ with the spectrum of $\boldsymbol{\Sigma}$ in Appendix E.2.1. The bounds on eigenvalues and the norms of eigenfunctions are developed in Appendices E.2.2 and E.2.3 respectively.

E.2.1. *Explicit forms of the eigenvalues and eigenfunctions.* We calculate the pairwise inner products of functions $\{\phi_{B,j}\}_{j=1}^{\infty}$ under the distribution μ_m. By definition of $\{\phi_{B,j}\}_{j=1}^{\infty}$ in equation (79b), we have $\phi_{B,j}^2(x) = 1$, therefore, $\int_{\mathcal{X}} \phi_{B,j}^2(x)\ \mu_m(dx) = 1$ for any $j = 1, 2, \ldots$. We then consider $\int_{\mathcal{X}} \phi_{B,i}(x)\phi_{B,j}(x)\ \mu_m(dx)$ with $i \neq j$. Suppose that $i, j \geq 2$. Recall that by our construction, $\phi_{B,j}(x) = \phi_{B,j}(x + \frac{1}{2})$ for any $x \in \big[0, \frac{1}{2}\big)$ and $j \geq 2$. Moreover, we have $\mu_m(x) + \mu_m(x + \frac{1}{2}) = 2$ by equation (87). Based on these observations, we derive that

$$\begin{aligned}\int_{\mathcal{X}} \phi_{B,i}(x)\phi_{B,j}(x)\mu_m(dx) &= \int_0^{\frac{1}{2}} \phi_{B,i}(x)\phi_{B,j}(x)\big\{\mu_m(dx) + \mu_m\big(d(x+\tfrac{1}{2})\big)\big\} \\ &= 2\int_0^{\frac{1}{2}} \phi_{B,i}(x)\phi_{B,j}(x)dx = \int_{\mathcal{X}} \phi_{B,i}(x)\phi_{B,j}(x)dx = 0\,,\end{aligned}$$

where the last equality is because $\phi_{B,i}$ and $\phi_{B,j}$ are orthogonal in $L^2(\mu)$ for any $i \neq j$. As for the cases where $i = 1$ and $j \geq 2$, we find that

$$\begin{aligned}&\int_{\mathcal{X}} \phi_{B,i}(x)\phi_{B,j}(x)\ \mu_m(dx) \overset{(i)}{=} \int_0^{\frac{1}{2}} \phi_{\theta,j}(x)\ \big\{\mu_m(dx) - \mu_m\big(d(x+\tfrac{1}{2})\big)\big\} \\ &\overset{(ii)}{=} \frac{2}{p}\int_0^{\frac{1}{2}} \phi_{B,j}(x) f_m(x)\ dx \overset{(iii)}{=} \frac{1}{p}\int_{\mathcal{X}} \phi_{B,j}(x) f_m(x) dx \overset{(iv)}{=} \begin{cases} \frac{\alpha_m^{(j)}}{25\sqrt{n}} & \text{if } j \leq d_n\,, \\ 0 & \text{otherwise}\,. \end{cases}\end{aligned}$$

Here step (i) follows from the fact that $\phi_{B,1}(x) = \mathbb{1}\big\{x \in [0, \frac{1}{2})\big\} - \mathbb{1}\big\{x \in [\frac{1}{2}, 1)\big\}$; step (ii) follows from the equality $\mu_m(x) - \mu_m(x + \frac{1}{2}) = (2/p)\, f_m(x)$ by equation (87); step (iii) is because $f_m(x) = f_m(x + \frac{1}{2})$ for any $x \in [0, \frac{1}{2})$; and step (iv) results from our choice of f_m in equation (85).

Based on the calculations above, we are now ready to explicitly express the eigenvalues and eigenfunctions of operator $\Sigma_{\text{cov}}(\mathcal{P}_m)$. Define a d_n-by-d_n matrix

$$\boldsymbol{\Sigma} := \begin{pmatrix} \mu_1 & \boldsymbol{x}^\top \\ \boldsymbol{x} & \mathbf{D} \end{pmatrix} \tag{109}$$

where $\mathbf{D}$ is a diagonal matrix given by $\mathbf{D} := \operatorname{diag}\{\mu_2, \mu_3, \ldots, \mu_{d_n}\} \in \mathbb{R}^{(d_n-1)\times(d_n-1)}$ and the vector $\boldsymbol{x}$ satisfies $\boldsymbol{x} := \frac{1}{25\sqrt{n}}\sqrt{\mu_1}\mathbf{D}\,\boldsymbol{\alpha}_m \in \mathbb{R}^{d_n-1}$. Recall that the binary vector $\boldsymbol{\alpha}_m$ is a component in the packing of Boolean hypercube $\{0,1\}^{d_n-1}$.

Let $\{\tilde{\mu}_j\}_{j=1}^{d_n}$ be the eigenvalues of matrix $\boldsymbol{\Sigma}$ in non-increasing order and define $\tilde{\mu}_j := \mu_j$ for $j \geq d_n + 1$. Then $\{\tilde{\mu}_j\}_{j=1}^{\infty}$ are the eigenvalues of covariance operator $\Sigma_{\text{cov}}(\mathcal{P}_m)$. For any index $j \geq d_n + 1$, the basis function $\phi_{B,j}$ is the eigenfunction associated with eigenvalue $\tilde{\mu}_j = \mu_j$, i.e. $\tilde{\phi}_j = \phi_{B,j}$. When $j \in [d_n]$, let $\boldsymbol{v}_j \in \mathbb{R}^{d_n}$ be the j-th eigenvector of the arrowhead matrix $\boldsymbol{\Sigma}$ defined in equation (109). The function $\tilde{\phi}_j := (\phi_{B,1}, \phi_{B,2}, \ldots, \phi_{B,d_n})\,\boldsymbol{v}_j$ is the eigenfunction associated with eigenvalue $\tilde{\mu}_j$.

In the sequel, we leverage the properties of the arrowhead matrix $\boldsymbol{\Sigma}$ to analyze the eigenpairs $\{(\tilde{\mu}_j, \tilde{\phi}_j)\}_{j=1}^{\infty}$.

E.2.2. *Bounds on eigenvalues.* We learn from Cauchy interlacing theorem that

$$\tilde{\mu}_1 \geq \mu_2 \geq \tilde{\mu}_2 \geq \ldots \geq \mu_{d_n} \geq \tilde{\mu}_{d_n}. \tag{110}$$

Therefore, $\tilde{\mu}_j \leq \mu_j$ for $j \geq 2$.

E.2.3. *Bonds on the norms of eigenfunctions.* By our construction, we have $\|\phi_{B,j}\|_\infty = 1$ for all $j = 1, 2, \ldots$. It follows that $\|\tilde{\phi}_j\|_\infty = \|\phi_{B,j}\|_\infty \le \kappa$ for $j \ge d_n + 1$. As for indices $j \in [d_n]$, it holds that $\|\tilde{\phi}_j\|_\infty \le \|\boldsymbol{v}_j\|_1 \sup_{i\in[d_n]} \|\phi_{B,i}\|_\infty = \|\boldsymbol{v}_j\|_1$. In what follows, we verify that $\|\boldsymbol{v}_j\|_1 \le \kappa$ for any $j \in [d_n]$.

Using the properties of arrowhead matrix $\boldsymbol{\Sigma}$ [36], we find that $\boldsymbol{v}_j$ can be explicitly written as

$$\boldsymbol{v}_j = \frac{\boldsymbol{u}_j}{\|\boldsymbol{u}_j\|_2} \qquad \text{with } \boldsymbol{u}_j = \begin{pmatrix} 1 \\ (\tilde{\mu}_i \mathbf{I} - \mathbf{D})^{-1}\boldsymbol{x} \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{\sqrt{\mu_1}}{25\sqrt{n}} (\tilde{\mu}_j \mathbf{I} - \mathbf{D})^{-1} \sqrt{\mathbf{D}}\, \boldsymbol{\alpha}_m \end{pmatrix} \tag{111}$$

for any $j \in [d_n]$. The eigenvalues $\{\tilde{\mu}_j\}_{j=1}^{d_n}$ are zeros to the characteristic function

$$\chi(\mu) := \mu_1 - \mu + \boldsymbol{x}^\top (\mu \mathbf{I} - \mathbf{D})^{-1} \boldsymbol{x}\,. \tag{$\chi(\mu)$}$$

Estimation of $\|\boldsymbol{v}_1\|_1$. We first consider $\|\boldsymbol{v}_1\|_1$, the ℓ_1-norm of the first eigenvector. Since $\|\boldsymbol{u}_j\|_2 \ge 1$, we use the expression of $\boldsymbol{v}_1$ in equation (111) and find that

$$\|\boldsymbol{v}_1\|_1 \le \|\boldsymbol{u}_1\|_1 = 1 + \frac{\sqrt{\mu_1}}{25\sqrt{n}} \sum_{i=2}^{d_n} \frac{\sqrt{\mu_i}}{\tilde{\mu}_1 - \mu_i}\,. \tag{112}$$

According to the characteristic equation $\chi(\tilde{\mu}_1) = 0$, it holds $\tilde{\mu}_1 - \mu_1 = \boldsymbol{x}^\top (\tilde{\mu}_1 \mathbf{I} - \mathbf{D})^{-1} \boldsymbol{x}$. Inequality (110) ensures that $\tilde{\mu}_1 \ge \mu_2 \ge \ldots \ge \mu_{d_n}$, therefore, $\boldsymbol{x}^\top (\tilde{\mu}_1 \mathbf{I} - \mathbf{D})^{-1} \boldsymbol{x} \ge 0$. It further implies $\tilde{\mu}_1 \ge \mu_1$. We plug it into inequality (112) and obtain that

$$\begin{aligned} \|\boldsymbol{v}_1\|_1 \le 1 + \frac{\sqrt{\mu_1}}{25\sqrt{n}} \sum_{i=2}^{d_n} \frac{\sqrt{\mu_i}}{\mu_1 - \mu_2} &\overset{(i)}{\le} 1 + \frac{\sqrt{\mu_1}\sqrt{d_n \sum_{i=1}^{\infty} \mu_i}}{25\sqrt{n}(\mu_1 - \mu_2)} \\ &\overset{(ii)}{\le} 1 + \frac{b\sqrt{\mu_1}}{25(\mu_1 - \mu_2)} \frac{1-\gamma}{\kappa\bar{\sigma}} \bar{R}\delta_n \overset{(iii)}{\le} 2 = \kappa\,. \end{aligned}$$

Here, step (i) is due to the Cauchy-Schwarz inequality; step (ii) is by inequality $\sum_{j=1}^{\infty} \mu_j \le \frac{b^2}{4}$ in condition (32a) and the critical inequality (33); and step (iii) is due to inequality $\bar{R}\delta_n \le 10\kappa\bar{\sigma}\big(1 - \frac{\mu_2}{\mu_1}\big)\frac{\sqrt{\mu_1}}{b}$ in condition (36b). We then conclude that $\|\phi_1\|_\infty \le \|\boldsymbol{v}_1\|_1 \le 2 = \kappa$, as claimed in the lemma statement.

Estimation of $\|\boldsymbol{v}_j\|_1$ *for* $j = 2, \ldots, d_n$. We next consider the ℓ_1-norms of eigenvectors $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_{d_n}$. Intuitively, when the sample size n is sufficiently large, vector $\boldsymbol{x}$ in matrix $\boldsymbol{\Sigma}$ is small and $\boldsymbol{\Sigma}$ is approximately diagonal. In this case, we expect that the eigenvector $\boldsymbol{v}_j$ is close to the j-th canonical basis $\mathbf{e}_j$ so that $\|\tilde{\phi}_j\|_\infty = \|\boldsymbol{v}_j\|_1 \approx \|\mathbf{e}_j\|_1 = 1$.

In order to prove this claim, we will show that the j-th entry of vector $\boldsymbol{u}_j$ (denoted by $\boldsymbol{u}_j(j)$) in equation (111) is noticeably larger than the other entries in $\boldsymbol{u}_j$. This is because the eigenvalue difference $|\tilde{\mu}_j - \mu_j|$ is rather small compared with eigengaps $|\tilde{\mu}_j - \mu_i|$ with $i \ne j$. Indeed, we will prove that it roughly holds $\mu_j - \tilde{\mu}_j \lesssim \frac{\mu_j}{n}$, thus $\boldsymbol{u}_j(j)$ has order $\Omega(\sqrt{n})$. Under our eigengap condition $\min_{3\le j\le d_n} \big\{\sqrt{\mu_{j-1}} - \sqrt{\mu_j}\big\} \ge \frac{\delta_n}{2d_n}$, the gaps $|\tilde{\mu}_j - \mu_i|$ with $i \ne j$ are relatively large so that the sum of entries $\big\{|\boldsymbol{u}_j(i)| \mid i \ne j\big\}$ is at most $\widetilde{\mathcal{O}}(\sqrt{d_n})$ [2]. Here, $\boldsymbol{u}_j(i)$ denotes the i-th entry of vector $\boldsymbol{u}_j$. To this end, rescaling $\boldsymbol{u}_j$ yields a vector $\boldsymbol{v}_j$ that approximates $\mathbf{e}_j$.

[2] $\widetilde{\mathcal{O}}$ stands for the big O notation, omitting logarithmic factors.

Let us now prove the arguments that were sketched above. For notational simplicity, we only consider $\boldsymbol{v}_j$ with $2 \le j \le d_n - 1$. The analysis of $\boldsymbol{v}_{d_n}$ is very similar. We first partition the entries of $\boldsymbol{u}_j$ into three groups and decompose the norm $\|\boldsymbol{v}_j\|_1$ accordingly. Specifically, we have $\|\boldsymbol{v}_j\|_1 = A_1 + A_2 + A_3$ where

$$A_1 := \frac{1}{\|\boldsymbol{u}_j\|_2}\big\{|\boldsymbol{u}_j(1)| + |\boldsymbol{u}_j(j)| + |\boldsymbol{u}_j(j+1)|\big\},$$

$$A_2 := \frac{1}{\|\boldsymbol{u}_j\|_2}\sum_{i=2}^{j-1}|\boldsymbol{u}_j(i)| \qquad \text{and} \qquad A_3 := \frac{1}{\|\boldsymbol{u}_j\|_2}\sum_{i=j+2}^{d_n}|\boldsymbol{u}_j(i)|.$$

By the Cauchy-Schwarz inequality, the term A_1 satisfies

$$A_1 \le \frac{|\boldsymbol{u}_j(1)| + |\boldsymbol{u}_j(j)| + |\boldsymbol{u}_j(j+1)|}{\sqrt{(\boldsymbol{u}_j(1))^2 + (\boldsymbol{u}_j(j))^2 + (\boldsymbol{u}_j(j+1))^2}} \le \sqrt{3}\,. \tag{113}$$

We take shorthands $\tilde{u}_{j,i} := \frac{25\sqrt{n}}{\sqrt{\mu_1}}\,\boldsymbol{u}_j(i) = \frac{\sqrt{\mu_i}}{\tilde{\mu}_j - \mu_i}\,\boldsymbol{\alpha}_m(i)$ for $i = 2, 3, \ldots, d_n$. Since $\boldsymbol{u}_j(j)$ dominates the other entries in $\boldsymbol{u}_j$, we approximate A_2 and A_3 by

$$A_2 \le \frac{1}{|\boldsymbol{u}_j(j)|}\sum_{i=2}^{j-1}|\boldsymbol{u}_j(i)| = \frac{1}{|\tilde{u}_{j,j}|}\sum_{i=2}^{j-1}|\tilde{u}_{j,i}| =: \widetilde{A}_2, \tag{114a}$$

$$A_3 \le \frac{1}{|\boldsymbol{u}_j(j)|}\sum_{i=j+2}^{d_n}|\boldsymbol{u}_j(i)| = \frac{1}{|\tilde{u}_{j,j}|}\sum_{i=j+2}^{d_n}|\tilde{u}_{j,i}| =: \widetilde{A}_3. \tag{114b}$$

In the following, we estimate upper bounds $\widetilde{A}_2$ and $\widetilde{A}_3$ in inequalities (114).

Under the eigengap condition $\min_{3\le i\le d_n}\big\{\sqrt{\mu_{i-1}} - \sqrt{\mu_i}\big\} \ge \frac{\delta_n}{2d_n}$, we can show that

$$\sum_{i=2}^{j-1}\frac{\sqrt{\mu_i}}{\mu_i - \mu_j} \le \frac{2d_n}{\delta_n}(1 + \log n), \qquad \sum_{i=j+2}^{d_n}\frac{\sqrt{\mu_i}}{\mu_{j+1} - \mu_i} \le \frac{2d_n}{\delta_n}(1 + \log n). \tag{115}$$

We assume the claim (115) to hold at this point and prove that both $\widetilde{A}_2$ and $\widetilde{A}_3$ are constant order.

In terms of the numerators of terms $\widetilde{A}_2$ and $\widetilde{A}_3$, the interlacing inequality (110) and the claim (115) imply that

$$\sum_{i=2}^{j-1}|\tilde{u}_{j,i}| = \sum_{i=2}^{j-1}\frac{\sqrt{\mu_i}}{\mu_i - \tilde{\mu}_j} \le \sum_{i=2}^{j-1}\frac{\sqrt{\mu_i}}{\mu_i - \mu_j} \le \frac{2d_n}{\delta_n}(1 + \log n), \tag{116a}$$

$$\sum_{i=j+2}^{d_n}|\tilde{u}_{j,i}| = \sum_{i=j+2}^{d_n}\frac{\sqrt{\mu_i}}{\tilde{\mu}_j - \mu_i} \le \sum_{i=2}^{j-1}\frac{\sqrt{\mu_i}}{\mu_{j+1} - \mu_i} \le \frac{2d_n}{\delta_n}(1 + \log n). \tag{116b}$$

Consider the common denominator $|\tilde{u}_{j,j}| = \frac{\sqrt{\mu_j}}{\mu_j - \tilde{\mu}_j}$ of $\widetilde{A}_2$ and $\widetilde{A}_3$. A key step in our analysis is to estimate the perturbation term $\mu_j - \tilde{\mu}_j$. Recall that $\tilde{\mu}_j$ satisfies the characteristic equation $\chi(\tilde{\mu}_j) = 0$, which translates into

$$\frac{\mu_1\mu_j}{25^2 n(\mu_j - \tilde{\mu}_j)} = \mu_1 - \tilde{\mu}_j + \frac{\mu_1}{25^2 n}\sum_{\substack{2\le i\le d_n,\\ i\ne j}}\frac{\mu_i}{\tilde{\mu}_j - \mu_i}.$$

We use the interlacing inequality (110) and obtain that

$$\frac{\mu_1\mu_j}{25^2 n(\mu_j-\tilde{\mu}_j)} \geq \mu_1-\mu_j-\frac{\mu_1}{25^2 n}\sum_{i=2}^{j-1}\frac{\mu_i}{\mu_i-\mu_j} \geq \mu_1-\mu_j-\frac{\mu_1^{\frac{3}{2}}}{25^2 n}\sum_{i=2}^{j-1}\frac{\sqrt{\mu_i}}{\mu_i-\mu_j}.$$

When the bounds (115) hold, we have

$$\begin{aligned}(117)\quad \frac{\mu_1^{\frac{3}{2}}}{25^2 n}\sum_{i=2}^{j-1}\frac{\sqrt{\mu_i}}{\mu_i-\mu_j} &\leq \frac{2\mu_1^{\frac{3}{2}} d_n}{25^2 n\delta_n}(1+\log n)\\ &\overset{(i)}{\leq} \frac{2\mu_1^{\frac{3}{2}}}{25^2}\left\{\frac{\bar{R}(1-\gamma)}{\kappa\bar{\sigma}}\right\}^2\delta_n\,(1+\log n) \overset{(ii)}{\leq} \frac{8}{125}(\mu_1-\mu_j),\end{aligned}$$

where step (i) is due to inequality (33); and in step (ii) we use inequality

$$\bar{R}\delta_n \leq 10\kappa\bar{\sigma}\big(1-\tfrac{\mu_2}{\mu_1}\big)\frac{\kappa\bar{\sigma}/(\sqrt{\mu_1}\bar{R})}{(1-\gamma)^2\log n}$$

in condition (36b). We integrate the pieces and derive that

$$\frac{\mu_1\mu_j}{25^2 n(\mu_j-\tilde{\mu}_j)} \geq \frac{1}{2}\,(\mu_1-\mu_j).$$

It further implies

$$(118)\qquad \frac{1}{|\tilde{u}_{j,j}|} = \frac{\mu_j-\tilde{\mu}_j}{\sqrt{\mu_j}} \leq \frac{2\,\mu_1\sqrt{\mu_j}}{25^2\,n(\mu_1-\mu_j)}.$$

Combining inequalities (116) and (118), we find that the terms $\widetilde{A}_2$ and $\widetilde{A}_3$ in bounds (114) satisfy

$$\begin{aligned}\max\{\widetilde{A}_2,\widetilde{A}_3\} &\leq \frac{2\mu_1\sqrt{\mu_j}}{25^2 n(\mu_1-\mu_j)}\left\{\frac{2d_n}{\delta_n}(1+\log n)\right\}\\ &\leq \frac{2}{\mu_1-\mu_j}\left\{\frac{2\mu_1^{\frac{3}{2}} d_n}{25^2 n\delta_n}(1+\log n)\right\} \overset{(i)}{\leq} \frac{16}{125},\end{aligned}$$

where step (i) follows from inequality (117). We plug inequalities (113) and (114) into the decomposition $\|\boldsymbol{v}_j\|_1 = A_1+A_2+A_3$ and derive that $\|\tilde{\phi}_j\|_\infty \leq \|\boldsymbol{v}_j\|_1 \leq 2 = \kappa$, as claimed in the lemma statement.

It only remains to prove the claim (115). We use some algebra and obtain that

$$\frac{\sqrt{\mu_i}}{\mu_i-\mu_j} \leq \frac{1}{\sqrt{\mu_i}-\sqrt{\mu_j}} \quad \text{for } i \leq j-1 \qquad \text{and}$$

$$\frac{\sqrt{\mu_i}}{\mu_{j+1}-\mu_i} \leq \frac{1}{\sqrt{\mu_{j+1}}-\sqrt{\mu_i}} \quad \text{for } i \geq j+2.$$

Under the eigengap condition $\min_{3\leq i\leq d_n}\{\sqrt{\mu_{i-1}}-\sqrt{\mu_i}\} \geq \frac{\delta_n}{2d_n}$, we have

$$\sqrt{\mu_{i_1}}-\sqrt{\mu_{i_2}} \geq (i_2-i_1)\,\frac{\delta_n}{2d_n} \qquad \text{for any } 2 \leq i_1 < i_2 \leq d_n.$$

It then follows that

$$\sum_{i=2}^{j-1} \frac{\sqrt{\mu_i}}{\mu_i - \mu_j} \leq \sum_{i=2}^{j-1} \frac{1}{\sqrt{\mu_i} - \sqrt{\mu_j}} \leq \frac{2d_n}{\delta_n} \sum_{i=2}^{j-1} \frac{1}{j-i}$$
$$\leq \frac{2d_n}{\delta_n} \left\{1 + \log(j-2)\right\} \leq \frac{2d_n}{\delta_n}(1 + \log n)\,.$$

The second bound in equation (115) can be proved in a similar way.

APPENDIX F: EXTENSION TO OFF-POLICY EVALUATION

Here we describe how the results in the main body can be extended to the off-policy setting, in which the the data used to collect the data differs from the stationary measure induced by the policy of interest. We begin in Appendix F.1 by setting up the problem of off-policy evaluation; in Appendix F.2, we state an extension of the non-asymptotic upper bounds from Theorem 3.1 to this off-policy setting. Appendix F.3 is devoted to the proofs.

F.1. Problem set-up. The off-policy setting is more general than the on-policy setting, as we describe here. Instead of a Markov reward process, we consider a Markov decision process (MDP) $\mathscr{J}(\mathcal{P}, r, \gamma)$ defined over a state space $\mathcal{S}$ and an action space $\mathcal{A}$, and with transition kernel $\mathcal{P}$. Note that a Markov reward process can be obtained from an MDP by fixing some policy; an MDP can be viewed as a collection of Markov reward processes indexed by policies. In an MDP, the transition kernel specifies the distribution of a future state conditioned on a state-action pair. The function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the reward defined over the state and action space. Our goal is to use data to evaluate the quality of a *target policy* π. In particular, we seek to estimate the state-action value function (Q-function) $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ given by

$$Q^\pi(s,a) := \mathbb{E}\big[\textstyle\sum_{h=0}^{\infty} \gamma^h\, r(S_h, A_h) \mid S_0 = s, A_0 = a\big],$$

where the trajectory $(S_0 = s, A_0 = a, S_1, A_1, S_2, A_2, \ldots)$ is generated by policy π via $A_h \sim \pi(\cdot \mid S_h)$, and $S_{h+1} \sim \mathcal{P}(\cdot \mid S_h, A_h)$.

In the off-policy setting, we do not observe data that has been generated by the target policy. Instead, the dataset consists of i.i.d. tuples $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1}^n$, where the samples $\{(s_i, a_i)\}_{i=1}^n$ are independently drawn from a distribution $\mu_{\mathcal{D}}$ over the space $\mathcal{S} \times \mathcal{A}$ and s_i' is the next state transiting from (s_i, a_i) according to transition kernel $\mathcal{P}$. The key distinction from the on-policy setting is that the distribution $\mu_{\mathcal{D}}$ over state-action pairs might be different from the stationary distribution μ_π associated with implementing the target policy π in steady state; for this reason, this setting is referred to as *off-policy evaluation*.

In order to apply an RKHS-method in the the off-policy setting, we consider a kernel function $\mathcal{K}$ defined over the state-action space[3] $\mathcal{S} \times \mathcal{A}$ and take the representer of evaluation as $\Phi_{s,a} := \mathcal{K}(\cdot, (s,a))$. The kernel LSTD estimate $\widehat{\theta}$ is again defined by the fixed-point equation (10) except that the (cross-)covariance operators are replaced by

(119)
$$\widehat{\Sigma}_{\text{cov}} := \frac{1}{n}\sum_{i=1}^n \Phi_{s_i,a_i} \otimes \Phi_{s_i,a_i} \quad \text{and} \quad \widehat{\Sigma}_{\text{cr}} = \frac{1}{n}\sum_{i=1}^n \Phi_{s_i,a_i} \otimes \Big\{ \int_{\mathcal{A}} \Phi_{(s_i',a)}\, \pi(da \mid s_i') \Big\}.$$

Again, we define θ^* as the population-level counterpart of $\widehat{\theta}$. Our goal is to derive non-asymptotic upper bounds on the estimation error $\|\widehat{\theta} - \theta^*\|^2_{\mu_{\mathcal{D}}}$.

[3]Recall that in the on-policy setting, our kernel was defined only over the state space.

F.2. Non-asymptotic upper bounds. In order to state and prove some non-asymptotic upper bounds on this estimator, we need to quantify the amount of distribution shift. We do so via coefficients $C_{\mathsf{shift}} \in \big(0, (1/\gamma)\big)$ and $C'_{\mathsf{shift}} \in [1, +\infty)$ given by

$$\sup_{f \in \mathbb{H}: \|f\|_{\mu_{\mathcal{D}}} > 0} \frac{\mathbb{E}[f(S,A) f(S',A')]}{\|f\|^2_{\mu_{\mathcal{D}}}} \leq C_{\mathsf{shift}} < (1/\gamma) \qquad \text{and} \tag{120a}$$

$$\sup_{f \in \mathbb{H}: \|f\|_{\mu_{\mathcal{D}}} > 0} \frac{\mathbb{E}\big[\mathbb{E}[f(S',A') \mid S']^2\big]}{\|f\|^2_{\mu_{\mathcal{D}}}} \leq C'_{\mathsf{shift}} < +\infty\,. \tag{120b}$$

The expectations are taken over quadruples (S, A, S', A') with $(S, A) \sim \mu_{\mathcal{D}}$, $S' \sim \mathcal{P}(\cdot \mid S, A)$ and $A' \sim \pi(\cdot \mid S')$. In the special case of a linear kernel, the assumption that $C_{\mathsf{shift}} < (1/\gamma)$ is closely related to a stability condition from the paper [39], shown to be necessary for a stable estimation of value function. In general, it always holds that $C_{\mathsf{shift}} \leq (1 + C'_{\mathsf{shift}})/2$.

In the special case of on-policy evaluation, for which $\mu_{\mathcal{D}}$ is equal the stationary distribution μ_π under the target policy π, we can take $C_{\mathsf{shift}} = C'_{\mathsf{shift}} = 1$. For a genuinely off-policy problem, the two distributions $\mu_{\mathcal{D}}$ and μ_π are different, so that the coefficients C_{shift} and C'_{shift} may become large. Let us relate these coefficients to the so-called concentrability coefficients from the RL literature. The concentrability coefficients $(c_{\mathsf{con}}, C_{\mathsf{con}})$ provide uniform bounds on the likelihood ratio $\frac{d\mu_\pi}{d\mu_{\mathcal{D}}}$ as follows:

$$c_{\mathsf{con}} \leq \frac{d\mu_\pi}{d\mu_{\mathcal{D}}}(s,a) \leq C_{\mathsf{con}} \qquad \text{for any state-action pair } (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Given bounds of this type, we can take $C_{\mathsf{shift}} = \frac{(1 + \frac{C_{\mathsf{con}}}{c_{\mathsf{con}}})}{2}$ in equation (120a) and $C'_{\mathsf{shift}} = \frac{C_{\mathsf{con}}}{c_{\mathsf{con}}}$ in equation (120b). See Appendix F.3 for the proofs of these claims.

Under conditions (120a) and (120b), we can prove a non-asymptotic upper bound on the kernel-based estimator, with the following modifications:

- (*Eigenvalues*) The eigenvalues $\{\mu_j\}_{j=1}^{\infty}$ are induced by measure $\mu_{\mathcal{D}}$.
- (*Effective horizon*) The effective horizon is defined by $H(\gamma) := \frac{1}{1 - \gamma\, C_{\mathsf{shift}}}$.
- (*Bellman residual variance*) The variance term is defined as

$$\sigma^2(\theta^*) := \mathbb{E}_{(S,A) \sim \mu_{\mathcal{D}}}\Big[\big(\theta^*(S,A) - r(S,A) - \gamma\, \mathbb{E}_{A' \sim \pi(\cdot|S')}[\theta^*(S',A') \mid S']\big)^2\Big]. \tag{121}$$

By using the parameters above, we can define the critical inequality and radius in the same way as in inequality (CI(ζ)).

As a corollary of Theorem 3.1, we can show that the regularized kernel LSTD estimate $\widehat{\theta}$ in the off-policy setting has upper bounds of the form

$$\|\widehat{\theta} - \theta^*\|^2_{\mu_{\mathcal{D}}} \;\leq\; c_1\, R^2 \Big\{\delta^2 + \frac{\lambda_n}{1 - \gamma\, C_{\mathsf{shift}}}\Big\} \tag{122}$$

with probability at least $1 - 2\exp\big(-\frac{c_2 n \delta^2 (1 - \gamma\, C_{\mathsf{shift}})^2}{b^2}\big)$ for any $\lambda_n \geq c_0 \delta^2 (1 - \gamma\, C_{\mathsf{shift}})$.

COROLLARY F.1 (Non-asymptotic upper bounds for off-policy evaluation). Under the distribution shift assumptions (120a) and (120b), the bound (122) holds for any solution δ to the critical inequality $\mathrm{CI}(\kappa\sigma(\theta^*))$ once the sample size n is large enough to ensure that

$$R^2 \delta_n^2(\kappa\sigma(\theta^*)) \leq c\, \frac{\kappa\, \sigma^2(\theta^*)}{\sqrt{C'_{\mathsf{shift}}}(1 - \gamma\, C_{\mathsf{shift}})\sqrt{n}}\,. \tag{123}$$

To be clear, unlike our results in the main body, we cannot guarantee the sharpness and optimality of these off-policy bounds when the coefficient C_{shift} is much larger than 1. Resolving this issue of optimality is an important direction for future research.

F.3. Proofs. The proof of Corollary F.1 closely resembles that of Theorem 3.1, with the main differences lying in the lower bound (i) in inequality (39) of Lemma 4.1 and the proof of Lemma 4.3 regarding the upper bound on term T_3. For the off-policy setting, we define the function ρ in the following modified way:

$$\rho(f) := \Big(\mathbb{E}\big[f^2(S,A) - \gamma\, f(S,A) f(S',A')\big]\Big)^{1/2},$$

where the state-action pair (S,A) is generated from the distribution $\mu_{\mathcal{D}}$, and (S',A') represents the succeeding state and action drawn from the MDP $\mathcal{P}$ and target policy π. Under the distribution shift condition (120a), the decomposition (39) takes the following modified form

$$(1-\gamma\, C_{\mathsf{shift}})\|\widehat{\Delta}\|_{\mu_{\mathcal{D}}}^2 \overset{(i)}{\leq} \rho^2(\widehat{\Delta}) \overset{(ii)}{=} \Big\{\sum_{j=1}^{3} T_j\Big\} - \lambda_n \|\widehat{\Delta}\|_{\mathbb{H}}^2, \tag{124}$$

where the terms $\{T_i\}_{i=1}^3$ adhere to the same definitions as in equations (40). By following a similar approach to deriving Lemmas 4.2 and 4.3, we can show that

$$T_1 \leq c\,(1-\gamma\, C_{\mathsf{shift}})\,\delta_n^2 \Big\{\|\widehat{\Delta}\|_{\mathbb{H}} + R^2\Big\} + c\,R\,(1-\gamma\, C_{\mathsf{shift}})\,\delta_n\,\|\widehat{\Delta}\|_{\mu_{\mathcal{D}}} \qquad \text{and} \tag{125a}$$

$$T_3 \leq c\,(1-\gamma\, C_{\mathsf{shift}})\,\delta_n^2 \Big\{\|\widehat{\Delta}\|_{\mathbb{H}} + R^2\Big\} + \frac{\rho^2(\widehat{\Delta})}{2}. \tag{125b}$$

Each bound holds with probability at least $1 - \exp\big(-c'\frac{n\delta_n^2(1-\gamma C_{\mathsf{shift}})^2}{b^2}\big)$ with $\delta_n = \delta_n(\kappa\sigma(\theta^*))$. By combining the aforementioned pieces in the same manner as in the proof of Theorem 3.1, we establish the bounds as stated in Corollary F.1.

It is worth noting that the proof of inequality (125b) involves estimating a Rademacher complexity of the form

$$\mathcal{R}_n(t) := \mathbb{E}\left[\sup_{\substack{\rho(f)\leq t \\ \|f\|_{\mathbb{H}}\leq R}} \Big|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\, \mathbb{E}_\pi\big[f(s_i',A) \mid s_i'\big]\Big|\right], \tag{126}$$

which is a novel component in the analysis. This complexity arises in the proof of Lemma C.4 (the off-policy version), which bounds the solution t_n to inequality $\mathbb{E}[\widetilde{Z}_n(t)] \leq t^2/8$ for $\widetilde{Z}_n(t) = \sup_{\rho(t)\leq t, \|f\|_{\mathbb{H}}\leq R} \big|\langle f, (\widehat{\Sigma}_{\mathrm{cov}} - \gamma\widehat{\Sigma}_{\mathrm{cr}})f\rangle_{\mathbb{H}}\big|$. To control $\mathcal{R}_n(t)$, we leverage the distribution shift assumption (120b). In Lemma F.2 below, we provide a statement of the upper bound.

LEMMA F.2. *Let δ_n denote the smallest solution to critical inequality (CI(ζ)) defined with the modified parameters. Then*

$$\mathcal{R}_n(t) \;\leq\; c\,\sqrt{C'_{\mathsf{shift}}}\,\frac{R^2(1-\gamma\, C_{\mathsf{shift}})}{\zeta}\,\max\Big\{\delta_n^2,\; \delta_n \tfrac{t}{R\sqrt{1-\gamma\, C_{\mathsf{shift}}}}\Big\}. \tag{127}$$

Proof of Lemma F.2. We introduce a norm $\|f\|_+ := \mathbb{E}\big[\mathbb{E}[f(S',A') \mid S']^2\big]^{\frac{1}{2}}$ for any function $f \in \mathbb{H}$. Let $\{\mu_j^+\}_{j=1}^\infty$ be the eigenvalues of the covariance operator

$$\Sigma_+ := \mathbb{E}\big[\mathbb{E}[\Phi_{S',A'} \mid S'] \otimes \mathbb{E}[\Phi_{S',A'} \mid S']\big].$$

Due to the Courant minimax principle, condition (120b) implies the relation $\mu_j^+ \leq C'_{\mathsf{shift}}\,\mu_j$, where $\{\mu_j\}_{j=1}^\infty$ denote the eigenvalues associated with the population-level data distribution $\mu_{\mathcal{D}}$.

We introduce another Rademacher complexity $\widetilde{\mathcal{R}}_n(\delta)$ closely related to $\mathcal{R}_n(t)$:

$$\widetilde{\mathcal{R}}_n(\delta) := \mathbb{E}\left[\sup_{\substack{\|f\|_+\le\delta\\ \|f\|_{\mathrm{H}}\le 1}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, \mathbb{E}_\pi\big[f(s_i', A)\mid s_i'\big]\right|\right].$$

It is evident that

$$\widetilde{\mathcal{R}}_n(\delta) \le \sqrt{\frac{2}{n}\sum_{j=1}^\infty \min\{\delta^2, \mu_j^+\}} \le \sqrt{\frac{2}{n}\sum_{j=1}^\infty \min\{\delta^2, C'_{\mathsf{shift}}\,\mu_j\}}$$

$$\lesssim \frac{R(1-\gamma\, C_{\mathsf{shift}})}{\zeta}\,\max\left\{\sqrt{C'_{\mathsf{shift}}}\,\delta_n^2,\ \delta_n\delta\right\}, \tag{128}$$

where the last inequality follows from the property of critical inequality (CI(ζ)).

Moreover, assumptions (120a) and (120b) ensure that any function f with $\rho(f)\le t$ satisfies

$$\|f\|_+^2 \le C'_{\mathsf{shift}}\|f\|_{\mu_{\mathcal{D}}}^2 \le C'_{\mathsf{shift}}(1-\gamma\, C_{\mathsf{shift}})^{-1}\rho^2(f) \le C'_{\mathsf{shift}}(1-\gamma\, C_{\mathsf{shift}})^{-1}t^2\,.$$

This allows us to establish the relation

$$\mathcal{R}_n(t) = R\,\mathbb{E}\left[\sup_{\substack{\rho(f)\le t/R\\ \|f\|_{\mathrm{H}}\le 1}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, \mathbb{E}_\pi\big[f(s_i', A)\mid s_i'\big]\right|\right]$$

$$\le R\,\mathbb{E}\left[\sup_{\substack{\|f\|_+\le\sqrt{C'_{\mathsf{shift}}}\cdot\frac{t}{R\sqrt{1-\gamma\, C_{\mathsf{shift}}}}\\ \|f\|_{\mathrm{H}}\le 1}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, \mathbb{E}_\pi\big[f(s_i', A)\mid s_i'\big]\right|\right]$$

$$= R\cdot\widetilde{\mathcal{R}}_n\left(\sqrt{C'_{\mathsf{shift}}}\cdot\frac{t}{R\sqrt{1-\gamma\, C_{\mathsf{shift}}}}\right). \tag{129}$$

Combining inequalities (128) and (129) completes the proof of Lemma F.2.

Relation between coefficients. In the following, we establish the relationships between the coefficients C_{shift}, C'_{shift} and $(c_{\mathsf{con}}, C_{\mathsf{con}})$. Using the concentrability coefficients, we observe that

$$c_{\mathsf{con}}\,\mathbb{E}_{(S,A)\sim\mu_{\mathcal{D}}}[f^2(S',A')] \le \mathbb{E}_{(S,A)\sim\mu_\pi}[f^2(S',A')] \overset{(*)}{=} \mathbb{E}_{(S,A)\sim\mu_\pi}[f^2(S,A)]$$
$$\le C_{\mathsf{con}}\,\mathbb{E}_{(S,A)\sim\mu_{\mathcal{D}}}[f^2(S,A)] = C_{\mathsf{con}}\,\|f\|_{\mu_{\mathcal{D}}}^2\,,$$

where the equality $(*)$ arises from the stationarity of distribution μ_π. It then follows from Young's inequality that

$$\mathbb{E}\big[f(S,A)f(S',A')\big] \le \frac{1}{2}\Big\{\mathbb{E}[f^2(S,A)] + \mathbb{E}[f^2(S',A')]\Big\} \le \frac{1+c_{\mathsf{con}}^{-1}C_{\mathsf{con}}}{2}\,\|f\|_{\mu_{\mathcal{D}}}^2\,.$$

Consequently, condition (120a) holds with parameter $C_{\mathsf{shift}} = (1+c_{\mathsf{con}}^{-1}C_{\mathsf{con}})/2$.

Similarly, we can establish inequality (120b) with $C'_{\mathsf{shift}} = C_{\mathsf{con}}/c_{\mathsf{con}}$ by noting that $\mathbb{E}\big[\mathbb{E}[f(S',A')\mid S']^2\big] \le \mathbb{E}[f^2(S',A')]$.