Krylov–Bellman boosting: Super-linear policy evaluation in general state spaces

Eric Xia
University of California, Berkeley

Martin J. Wainwright
Massachusetts Institute of Technology

Abstract

We present and analyze the Krylov–Bellman Boosting algorithm for policy evaluation in general state spaces. It alternates between fitting the Bellman residual using non-parametric regression (as in boosting), and estimating the value function via the least-squares temporal difference (LSTD) procedure applied with a feature set that grows adaptively over time. By exploiting the connection to Krylov methods, we equip this method with two attractive theoretical guarantees. First, we provide a general convergence bound that allows for separate statistical estimation errors in residual fitting and LSTD computation. Consistent with our numerical experiments, this bound shows that convergence rates depend on the restricted spectral structure, and are typically super-linear. Second, by combining this meta-result with sample-size dependent guarantees for residual fitting and LTSD computation, we obtain concrete statistical guarantees that depend on the sample size along with the complexity of the function class used to fit the residuals. We illustrate the behavior of the KBB algorithm for various types of policy evaluation problems, and typically find large reductions in sample complexity relative to the standard approach of fitted value iteration.

We focus on policy evaluation for Markov decision processes defined over general state spaces. Policy evaluation occupies a central role in reinforcement learning. It is of interest both in its own right and as a key subroutine in many algorithms. For instance, classical algorithms for computing optimal policies, such as policy iteration (Howard, 1960), and iterative schemes like policy gradient (Williams, 1992; Silver et al., 2014; Sutton et al., 1999; Kakade, 2001)

methods and actor-critic schemes (Konda and Tsitsiklis, 2001; Mnih et al., 2016; Bhatnagar et al., 2009).

Policy evaluation in discrete state spaces, i.e., the tabular setting, is very well-understood (e.g., (Bertsekas, 2017b; Bellman, 1957; Puterman and Brumelle, 1979)); far more challenging are Markov decision processes involving general state spaces. In this setting, there are at least two broad classes of approaches. One approach is based on using a suitably "rich" non-parametric function class to repeatedly refit an estimate of the value function; procedures of this type are known collectively as fitted value iteration (Ernst et al., 2005; Antos et al., 2007; Munos and Szepesvári, 2008; Scherrer et al., 2015; Chen and Jiang, 2019), or FVI for short. Another classical approach to approximate policy evaluation is via the *least-squares temporal difference* method, or LSTD for short (Bradtke and Barto, 1996; Sutton and Barto, 2018; Boyan, 2002). Given a finite collection of features or basis functions, it computes the value function within the associated linear span that provides the "best fit", as formalized in terms of a projected fixed point (e.g., (Bertsekas and Yu, 2009; Yu and Bertsekas, 2010)). In practice, the quality of the LSTD fit depends strongly on the given basis, and finding a good basis is often a difficult task. There is a line of past work (Keller et al., 2006; Menache et al., 2003) on automated procedures for feature construction.

Our contributions: In this paper, we propose and analyze a procedure for approximate policy evaluation that combines the LSTD framework with the boosting approach of fitting residuals. The method estimates the value function by computing a sequence of LSTD solutions, where the basis used at each round is augmented with a non-parametric estimate of the Bellman residual from the previous round. The LSTD steps can be interpreted as computing a sequence of Krylov subspaces, and accordingly, given these foundations, we refer to our method as the *Krylov–Bellman boosting* algorithm, or KBB for short. In the tabular setting, an idealized form of the KBB algorithm with exact Bellman residuals has been studied in past work (Parr et al., 2007, 2008), called the method of Bellman error basis functions (BEBF). However, this algorithm has been entirely limited to the setting of finite-dimension state spaces,

and relied on exact computation with full knowledge of the transition dynamics, which are both unrealistic restrictions. Instead, we use non-parametric procedures to (approximately) fit the residual in general state spaces. Moreover, any theoretical analysis of the procedure needs to account for the associated statistical errors (as well those in approximate LSTD calculation).

Within this broader context, the primary contributions of this paper take the form of two theorems. Our first main result, stated as Theorem 3, is a general convergence guarantee for the KBB algorithm that allows for errors in both the residual fitting and LSTD computations. Our convergence guarantee involves a sequence of so-called restricted spectral values (cf. equation (15)) that track how the effective conditioning of the residual problem improves as the algorithm proceeds. Our second set of results, stated as Theorem 4 and Corollary 1, provides guarantees in the RL setting, and in particular, for independent samples of state, next-state and reward triples. Under this set-up, Theorem 4 provides an upper bound on the regression error for a general function class, whereas Corollary 1 combines this guarantee with Theorem 3 to provide an "end-to-end" guarantee for a particular instantiation of the KBB algorithm. While we focus on the i.i.d. generative setting for concreteness, we note that our framework for analysis is general: by using tail bounds for dependent processes, one could also provide concrete guarantees for trajectory-based sampling models.

1 BACKGROUND

Here we describe some background necessary to explain and motivate our results. In Section 1.1 we describe the basic setup of the problem of interest, solving for the value function of a Markov reward process, as well as an associated sampling model. Section 1.2 then describes an approximate method for solving for the value function as linear combination of given basis functions. Finally, in Section 1.3 we describe Krylov subspace methods for solving linear equations.

1.1 Markov decision and reward processes

A Markov decision process consists of a state space $\mathcal{X}$, an action space $\mathcal{U}$, a reward function R, along with a collection of probability transition functions. In this paper, we focus exclusively on the policy evaluation procedure, in which case a given policy is fixed up front. For a fixed policy, a Markov decision process reduces to a *Markov reward process* (MRP), which can be characterized more simply in terms of the state space $\mathcal{X}$, along with a probability transition kernel $\mathbf{\mathcal{P}}$, and a reward function.

In a Markov reward process, the states evolve over time according to a collection of transition kernels $\mathbf{\mathcal{P}}$. At some

time $t = 1, 2, \ldots$, when in state x_t, the next state is generated by sampling $X_{t+1} \sim \mathbf{\mathcal{P}}(\cdot \mid x_t)$, where for each $x \in \mathcal{X}$, the function $\mathbf{\mathcal{P}}(\cdot \mid x)$ defines a probability measure over $\mathcal{X}$. Under standard regularity assumptions, the operator $\mathbf{\mathcal{P}}$ has a stationary measure μ satisfying the consistency condition

$$\mu(x') = \mathbb{E}_{X \sim \mu}[\mathbf{\mathcal{P}}(x' \mid X)].$$

The MRP also is equipped with a reward function $R : \mathcal{X} \to \mathbb{R}$, which we assume to be uniformly bounded and μ-measurable. At time $t = 1, 2, \ldots$, when in state x_t, we receive a reward $R(x_t)$.

The problem of policy evaluation corresponds to computing the *value function*

$$V^{\star}(x) \coloneqq \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t \cdot R(X_t) \mid x_0 = x \big], \tag{1}$$

where the conditional expectation is taken over a trajectory $(X_1, X_2, \ldots)$ from the underlying Markov chain. Since the reward function R is measurable and uniformly bounded, the value function exists and is uniquely defined. Moreover, it is measurable, uniformly bounded by $\frac{\|R\|_\infty}{1-\gamma}$, and therefore $V^{\star} \in L^2(\mathcal{X}, \mu)$.

Given a value function $V \in L^2(\mathcal{X}, \mu)$ we define the Bellman operator $\mathcal{T} : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ via

$$\mathcal{T}(V)(x) \coloneqq R(x) + \gamma \mathbb{E}_{x' \sim \mathbf{\mathcal{P}}(\cdot \mid x)}[V(x')] \quad \forall x \in \mathcal{X}. \tag{2}$$

By classical results on dynamic programming, the value function is a fixed point of the Bellman operator—that is, $V^{\star} = \mathcal{T}(V^{\star})$. Moreover, the Bellman operator $\mathcal{T}$ is γ-contractive in the norm $\| \cdot \|_\mu$; see the standard references (Puterman, 2014; Sutton and Barto, 2018; Bertsekas, 2009) for further background.

In order to clarify the connection to solving linear operator equations, define the transition operator $\mathcal{P}(V)(x) \coloneqq \mathbb{E}_{x' \sim \mathbf{\mathcal{P}}(\cdot \mid x)}\big[V(x')\big]$. Observing that $\mathcal{P}$ is a linear operator on $L^2(\mathcal{X}, \mu)$, we can rewrite the Bellman fixed point relation as

$$V^{\star} = R + \gamma \mathcal{P} V^{\star} \iff (\mathcal{I} - \gamma \mathcal{P}) V^{\star} = R, \tag{3}$$

where $\mathcal{I}$ denotes the identity operator on $L^2(\mathcal{X}, \mu)$. Thus, computing the value function is equivalent to solving the linear operator equation (3).

We conclude by describing the sampling model studied in this paper. The operator $\mathcal{P}$ and the reward function R are unknown to us, but we have access to sample transition pairs (x, r, x') generated as follows

$$x \sim \mu, \quad r = R(x), \quad \text{and} \quad x' \sim \mathbf{\mathcal{P}}(\cdot \mid x). \tag{4}$$

To be concrete, intermediate stages of our procedure make use of a dataset $\mathcal{D} = \{(x_t, r_t, x'_t)\}_{t=1}^n$ of n samples.

1.2 Linear function approximation and LSTD

When the state space $\mathcal{X}$ is sufficiently complex—either finite but with a very large number of states, or continuous in nature—any efficient scheme for policy evaluation requires some form of function approximation. In loose terms, we choose some function space, and then seek to find the "best" approximation to the value function within this space.

A classical and widely-used approach is based on *linear function approximation*. Given a collection of functions $\phi_j \in L^2(\mathcal{X}, \mu)$ for $j = 1, \ldots, J$, known either as features or basis functions, we use the subspace $\mathcal{S} = \mathrm{span}\{\phi_1, \phi_2, \ldots, \phi_J\}$ to approximate the value function. An attractive feature is that any function in the subspace $\mathcal{S}$ can be written as $V_\theta = \sum_{j=1}^{J} \theta_j \phi_j$ for some vector $\theta \in \mathbb{R}^J$, so that computations can be reduced to linear-algebraic operations over $\mathbb{R}^J$. One way to define the best approximation is via the notion of a projected fixed point (e.g., (Bertsekas and Yu, 2009; Yu and Bertsekas, 2010)). Defining the projection onto the subspace $\mathcal{S}$ via $\Pi(f) := \arg\min_{V \in \mathcal{S}} \|f - V\|_\mu$, we seek a function $V^{\mathrm{LSTD}} \in \mathcal{S}$ that satisfies the projected fixed point condition $V^{\mathrm{LSTD}} = \Pi(\mathcal{T}(V^{\mathrm{LSTD}}))$. Since Π is non-expansive and $\mathcal{T}$ is a contraction, both with respect to the $\|\cdot\|_\mu$-norm, there is a unique solution V^{LSTD}, known as the *population LSTD solution*. For future reference, we note that an abstract characterization of V^{LSTD} is in terms of the orthogonality conditions

$$\langle \phi_j, \mathcal{B}(V^{\mathrm{LSTD}})\rangle_\mu = 0 \qquad \text{for all } j = 1, \ldots, J, \quad (5)$$

where $\mathcal{B}(V^{\mathrm{LSTD}}) := V^{\mathrm{LSTD}} - \mathcal{T}(V^{\mathrm{LSTD}})$ is the *Bellman residual*.

Less abstractly, the computation of the LSTD estimate V^{LSTD} can be reduced to solving a linear system over $\mathbb{R}^J$. Let $\Phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_t(x)) \in \mathbb{R}^t$ be the vector obtained by evaluating each basis function at the state x. By standard LSTD theory (Bradtke and Barto, 1996), we can compute the LSTD solution V^{LSTD} by solving for $\theta^{\mathrm{LSTD}} \in \mathbb{R}^t$ in the linear system

$$\mathbb{E}_{(X,X')}\big[\Phi(X)\big(\Phi(X) - \gamma\Phi(X')\big)^T\big]\theta^{\mathrm{LSTD}} = \\ \mathbb{E}_{X \sim \mu}\big[R(X)\Phi(X)\big], \quad (6)$$

and then setting $V^{\mathrm{LSTD}} = V_{\theta^{\mathrm{LSTD}}}$.

In most practical settings, we cannot compute these expectations exactly, but instead have access to samples $\mathcal{D} = \{(x_i, r_i, x_i')\}_{i=1}^{n}$ generated according to the model (4). We can use these samples to form empirical estimates as follows. The *plug-in estimate* $\widehat{\theta}^{\mathrm{LSTD}}$ is obtained by solving the

linear system

$$\big(\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)(\Phi(x_i) - \gamma\Phi(x_i'))^T\big)\widehat{\theta}^{\mathrm{LSTD}} \\ = \frac{1}{n}\sum_{i=1}^{n}r_i\Phi(x_i), \quad (7)$$

which leads to the *empirical LSTD estimate* $\widehat{V}^{\mathrm{LSTD}} = V_{\widehat{\theta}^{\mathrm{LSTD}}}$.

1.3 Krylov subspace methods

Portions of our development rely on connections to Krylov subspace methods for solving systems of linear equations, which we describe briefly here. See the book by Saad (Saad, 2003) for more details. Suppose we are interested in solving a linear system $Ax = b$, where $A \in \mathbb{R}^{d \times d}$ is a matrix, and $b \in \mathbb{R}^n$ is a vector. Krylov subspace methods are based on approximating the solution to such a linear system via an expanding sequence of subspaces. For each positive integer $j = 1, \ldots, d$, we define the jth-order Krylov subspace as

$$\mathcal{K}_j(A, b) := \mathrm{span}\big\{b, Ab, A^2 b, \ldots, A^j b\big\}. \quad (8)$$

Using the subspace $\mathcal{K}_j(A, b)$, we can define an approximate solution $\widehat{x} \in \mathcal{K}_j(A, b)$

$$\widehat{x} \in \mathcal{K}_j(A, b) \qquad \text{such that} \qquad b - A\widehat{x} \perp \mathcal{K}_j(A, b). \quad (9)$$

Many algorithms for computing approximations to x involve forming the Krylov subspace $\mathcal{K}_j(A, b)$ and then solving the system (9); the conjugate gradient method is a notable instance. In the setting of tabular reinforcement learning, the problem of policy evaluation is equivalent to solving a linear system, so that Krylov methods are applicable, as has been noted in past work (Petrik, 2007). Our method, to be described in the next section, exploits the connection to Krylov theory at the more abstract level of operators, and allows for statistical errors in the updates.

2 KRYLOV–BELLMAN BOOSTING AND ITS GUARANTEES

In this section, we first describe the Krylov–Bellman boosting (KBB) procedure (Section 2.1), and then state some theoretical guarantees on its performance (Section 2.2). The KBB algorithm involves a sequence of LSTD solutions, along with a sequence of fits to the Bellman residual. Our first result (Theorem 3) is a general guarantee, one that allows for arbitrary errors in these intermediate computations. In Section 2.3, we analyze a general family of nonparametric procedures for fitting the residual, and provide an end-to-end result that bounds the behavior of the KBB procedure with statistical error.

2.1 Krylov–Bellman Boosting

The KBB procedure can be viewed as a procedure for adaptively choosing features or basis function; each feature is a function belonging to $L^2(\mathcal{X}, \mu)$. In the Krylov–Bellman boosting procedure, we assume access to a mechanism that generates samples of the form (x_i, r_i, x_i') generated according to the sampling model (4), with which we estimate the desired quantities. As described in Section 1.2, a dataset of the form $\mathcal{D}^{LSTD} = \{(x_j, r_j, x_j')\}_{j=1}^m$ can be used to compute the empirical LSTD estimate (7). Moreover, given our current estimate V of the value function, consider the problem of estimating the Bellman residual $\mathcal{B}(V)$ using a dataset $\mathcal{D}^{REG} = \{(x_i, r_i, x_i')\}_{i=1}^n$. In the specific instantiation of the KBB algorithm given here, we implement and analyze a *non-parametric least-squares estimate* of this residual. More precisely, given a suitably chosen function class $\mathcal{F}$ and the dataset $\mathcal{D}^{REG}$, we compute the approximation

$$\mathcal{B}(V) \approx \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \left[V(x_i) - \big(r_i + \gamma V(x_i')\big) - f(x_i) \right]^2 \right\}$$
(10)

The Krylov–Bellman boosting (KBB) procedure alternates between LSTD fitting and fitting of the residual, as stated formally as Algorithm 1.

An important choice is the function class $\mathcal{F}$ used to compute the regression estimates (11). There are various families that are commonly used in machine learning and statistics, including splines and other linear smoothing methods (Wahba, 1990; Hastie et al., 2001), reproducing kernel Hilbert spaces (Hofmann et al., 2008), random forests and regression trees (Breiman et al., 1984; Breiman, 2001), boosting procedures (Freund and Schapire, 1997), and neural networks (Schmidhuber, 2015). For our numerical experiments presented in Section A, we used a regression tree boosting procedure (Chen and Guestrin, 2016).

2.2 Convergence guarantees

We now provide some theoretical guarantees for the Krylov–Bellman Boosting procedure. Our first result (Theorem 3) is of a general nature: it allows arbitrary procedures to be used in computing the LSTD solutions and residual fits at intermediate stages of the algorithm. In fact, while we have described the KBB residual fits in terms of non-parametric regression, Theorem 3 actually applies to other procedures that might be used for this task. (For instance, if the model dynamics were known, one could use numerical procedures to approximate the Bellman update.) Our second main result (Theorem 4) is more specific in nature, in that it provides bounds on the error in the LSTD and residual fitting steps induced by using empirical samples, as used in our introduction of the procedure.

For ease of notation, define the discount operator $\mathcal{Q} := \mathcal{I} - \gamma\mathcal{P}$. In the analysis given here, we assume that the Markov chain is reversible.[1] This assumption implies that the transition operator is a self-adjoint operator on $L^2(\mathcal{X}, \mu)$, so that $\langle \mathcal{P}f, g \rangle_\mu = \langle f, \mathcal{P}g \rangle_\mu$ for all $f, g \in L^2(\mathcal{X}, \mu)$. This self-adjoint operator $\mathcal{Q}$ defines the inner product $\langle f, g \rangle_\mathcal{Q} := \langle f, \mathcal{Q}g \rangle_\mu$, along with the induced norm $\|f\|_\mathcal{Q} := \sqrt{\langle f, \mathcal{Q}f \rangle_\mu}$. Under certain regularity conditions, it can be shown (cf. the supplement) that the operator $\mathcal{Q} := \mathcal{I} - \gamma\mathcal{P}$ is self-adjoint, invertible, and satisfies the relation

$$(1 - \gamma)\|f\|_\mu^2 \leqslant \langle f, \mathcal{Q}f \rangle_\mu \leqslant (1 + \gamma)\|f\|_\mu^2 \qquad (12)$$

for all $f \in L^2(\mathcal{X}, \mu)$. In addition, the operator $\mathcal{Q}$ is bounded and (under mild regularity conditions) has a discrete spectrum.

Let $W_t := \phi_t - \mathcal{B}(V_{t-1})$ denote the error in the regression procedure at round t. Our first assumption provides control on the regression accuracy:

Assumption 1 (Reg-Err). *For each iteration $t = 1, 2, \ldots$, the regression procedure is δ_t-accurate:*

$$\mathbb{E}\big[\|W_t\|_\mathcal{Q}^2\big] \leqslant \delta_t^2. \qquad (13)$$

Our second assumption provides control on the accuracy of the LSTD computation at each step:

Assumption 2 (LSTD-Err). *At each iteration t, the approximate LSTD solution V_t is ϵ^{LSTD}-accurate:*

$$\|V_t - V^\star\|_\mathcal{Q}^2 \overset{(a)}{\lesssim} \|V_t^{\text{LSTD}} - V^\star\|_\mathcal{Q}^2 + (\epsilon^{\text{LSTD}})^2, \quad and$$

$$\|\mathcal{B}(V_t) - \mathcal{B}(V_t^{\text{LSTD}})\|_\mathcal{Q} \overset{(b)}{\lesssim} \epsilon^{\text{LSTD}}. \qquad (14)$$

We provide a brief discussion of these assumptions. Assumption 1 assumes that the approximate residual fit (10) is accurate up to some degree. To validate this assumption, we provide a guarantee in Theorem 4 on the regression step. Assumption 2 is an assumption on the accuracy of the standard LSTD estimate; see the paper (Dalal et al., 2018) for a concrete guarantee on the given LSTD estimator.

Our theory involves the behavior of the operator $\mathcal{Q}$ when restricted to certain subspaces of the function space. Recall that at each iteration, the algorithm performs LSTD over the linear subspace $\mathcal{S}_t = \text{span}\{\phi_1, \ldots, \phi_t\}$. Letting $\mathcal{S}_t^\perp$ denote the orthogonal complement, we define the constraint set $\mathcal{C}_t := \big\{z \in \mathcal{S}_t^\perp \mid \langle z, \mathcal{Q}^{-1}z \rangle_\mu = 1\big\}$, and the $\mathcal{C}_t$-*restricted spectral values*

$$m_t := \inf_{z \in \mathcal{C}_t^\perp} \|z\|_\mu^2 \quad \text{and} \quad M_t := \sup_{z \in \mathcal{C}_t^\perp} \|z\|_\mu^2. \qquad (15)$$

---

[1] Although our theory exploits reversibility, it does not seem essential to the practical behavior of the algorithm itself, as shown by the simulations to follow in the sequel.

Algorithm 1 Krylov–Bellman Boosting

1: `Inputs:` (i) Datasets $\mathcal{D}_t^{LSTD}$ and $\mathcal{D}_t^{REG}$, $t = 0, 1, 2, \ldots$, each containing transition tuples generated according to the model (4). (ii) Function class $\mathcal{F}$ over which to perform regression.

2: Initialize $V_0 = 0$, $S_0 = \varnothing$.

3: **for** $t = 0, \ldots$ **do**

4:   Use dataset $\mathcal{D}_t^{REG} = \{(x_i, r_i, x_i')\}_{i=1}^{n_t}$ to estimate $\mathcal{B}(V_t)$ by solving the non-parametric least-squares problem:

$$\phi_{t+1} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} \left[V_t(x_i) - \left(r_i + \gamma \cdot V_t(x_i') \right) - f(x_i) \right]^2 \right\}. \tag{11}$$

5:   Update basis set $S_{t+1} = S_t \cup \{\phi_{t+1}\}$.
     Define function $x \mapsto \Phi_{t+1}(x) := \left(\phi_1(x), \phi_2(x), \ldots, \phi_{t+1}(x) \right) \in \mathbb{R}^{t+1}$.

6:   Use dataset $\mathcal{D}_t^{LSTD} = \{(x_i, r_i, x_i')\}_{i=1}^{m_t}$ to estimate LSTD coefficient vector $\theta^{t+1} \in \mathbb{R}^{t+1}$ via

$$\theta^{t+1} = \left[\frac{1}{m_t} \sum_{i=1}^{m_t} \Phi_{t+1}(x_i)(\Phi_{t+1}(x_i) - \gamma \Phi_{t+1}(x_i'))^T \right]^{-1} \left\{ \frac{1}{m_t} \sum_{i=1}^{m_t} r_i \cdot \Phi_{t+1}(x_i) \right\}.$$

7:   Set $V_{t+1} = \sum_{j=1}^{t+1} \theta_j^{t+1} \phi_j$.

8: **end for**

Note we have the sandwich relation $1 - \gamma \leqslant m_t \leqslant M_t \leqslant 1 + \gamma$ at each iteration t. We let $\mathbb{E}_{t+1}$ denote the expectation over the noise at iterate $t + 1$.

Theorem 1 (General KBB bound). *Suppose that the Krylov–Bellman procedure is run with a δ_t-accurate regression procedure (cf.24), and an ϵ^{LSTD}-accurate LSTD implementation (cf. 25). Then at each step $t = 1, 2, \ldots$, the error satisfies the $\mathcal{Q}$-norm bound*

$$\mathbb{E}_{t+1} \|V_{t+1} - V^\star\|_{\mathcal{Q}}^2 \leqslant \left(1 - \tfrac{m_t^2}{8M_t} \right) \|V_t - V^\star\|_{\mathcal{Q}}^2$$
$$+ \tfrac{10}{M_t} \cdot \delta_t^2 + \tfrac{8M_t}{m_t^2} \cdot (\epsilon_t^{\text{LSTD}})^2.$$

See the appendix for the complete proof.

Note that the function $t \mapsto \frac{m_t^2}{8M_t}$ is non-decreasing in t by definition, so convergence is at least geometric, and can be faster if the ratio $\frac{m_t^2}{8M_t}$ is growing quickly. Using the fact that $m_t \geqslant 1 - \gamma$ and $M_t \leqslant 1 + \gamma$, we have the contraction factor is at least $1 - \frac{m_t^2}{8M_t} \leqslant 1 - \frac{(1-\gamma)^2}{8(1+\gamma)}$ at every iteration. However in the next section we illustrate that the convergence can be much faster than this worst case behavior. In addition, we can control the error terms δ_t^2 and $(\epsilon_t^{\text{LSTD}})^2$ by the number of samples, n_t and m_t, used in the regression procedure and LSTD estimate, respectively.

2.3   Regression bounds and consequences for KBB

Here we provide some results on the regression procedure in Algorithm 1 for completeness. Given data points $\mathcal{D}^{REG} = \{(x_i, r_i, x_i')\}_{i=1}^n$, we can write the regression problem in the generative form as $y_i = f^*(x_i) + w_i$

for $i = 1, 2, \ldots, n$, where

$$f^*(x) = V(x) - \mathcal{T}V(x), \quad \text{and} \tag{16}$$
$$w_i = \gamma \left(\mathcal{P}V(x_i) - V(x_i') \right). \tag{17}$$

Here we have used the fact that the rewards are deterministic. It should be noted that the "noise" variables w_i are actually dependent on the states x_i, so that some care is required in the analysis.

We can write our regression procedure as

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left(y_i - f(x_i) \right)^2.$$

Note that it need *not* be the case that $f^* \in \mathcal{F}$. In order to allow for the possibility of such mis-specification, we define the $\| \cdot \|_\mu$-projection $\Pi_{\mathcal{F}}$ from $L^2(\mathcal{X}, \mu)$ onto $\mathcal{F}$, which is well-defined when $\mathcal{F}$ is closed and convex.

Our results involve a non-asymptotic upper bound on $\|\widehat{f} - \mathcal{B}(V)\|_\mu$; it is easy to convert it to a $\| \cdot \|_{\mathcal{Q}}$-guarantee at the cost of a $(1 + \gamma)$- factor (see the supplement for details). Additionally, it involves a *statistical estimation error* β_n that depends on the function class $\mathcal{F}$. This quantity decreases to zero as the sample size n increases, but the rate of decrease depends on the complexity of the function class $\mathcal{F}$; see Appendix D.1 for its precise definition, and associated details.

Theorem 2 (Regression bound). *For a given function $V \in L^2(\mathcal{X}, \mu)$, suppose that we compute an estimate $\widehat{f}$ of the Bellman residual $\mathcal{B}(V)$ via the regression procedure (10) over a b-bounded and convex function class $\mathcal{F}$. Then we*

have

$$\mathbb{E}\|\widehat{f} - \mathcal{B}(V)\|_{\mu}^2 \leqslant c\Big\{\beta_n^2 + \frac{1}{n}\left(b^2 + \|V\|_{\infty}^2\right) \tag{18}$$
$$+ \|\Pi_{\mathcal{F}}(\mathcal{B}(V)) - \mathcal{B}(V)\|_{\mu}^2\Big\},$$

where the statistical estimation error β_n^2 depends on the complexity of $\mathcal{F}$.

See the supplement for the proof. Note that our regression procedures depends on β_n^2, as well as the approximation error $\|\Pi_{\mathcal{F}}(\mathcal{B}(V)) - \mathcal{B}(V)\|_{\mu}^2$; such dependence is to be expected since we are fitting $\widehat{f}$ in some function class $\mathcal{F}$ that does not necessarily contain $\mathcal{B}(V)$.

Combining Theorems 3 and 4 yields the following "end-to-end" guarantee on the KBB algorithm:

Corollary 1 (End-to-end guarantee for KBB). *Under the assumptions of Theorems 3 and 4, there are universal constants c_1 and c_2 such that the KBB update satisfies*

$$\mathbb{E}_{t+1}\|V_{t+1} - V^{\star}\|_{\mathcal{Q}}^2 \leqslant \left(1 - \frac{m_t^2}{8M_t}\right)\|V_t - V^{\star}\|_{\mathcal{Q}}^2$$
$$+ \frac{c_2 M_t}{m_t^2} \cdot (\epsilon^{\text{LSTD}})^2$$
$$+ \frac{c_1}{M_t}\Big\{\beta_n^2 + \frac{1}{n}\left(b^2 + \|V_t\|_{\infty}^2\right)$$
$$+ \|\Pi_{\mathcal{F}}(\mathcal{B}V_t) - \mathcal{B}V_t\|_{\mu}^2\Big\},$$

where $n = n_t = |\mathcal{D}_t^{REG}|$.

2.4 Estimation error via localized complexities

The statistical estimation error can be decomposed as the sum

$$\beta_n^2 = \omega_n^2 + \varepsilon_n^2, \tag{19}$$

where each of the two quantities $(\omega_n, \varepsilon_n)$ arise from localized complexity measures, in particular sub-Gaussian or Rademacher complexities, that play a central role in empirical process theory (van de Geer, 2000; Bartlett et al., 2005; Wainwright, 2019). When fitting the residual $f^* = \mathcal{B}(V)$, the relevant set is the *shifted function class* $\widetilde{\mathcal{F}} = \{f - \widetilde{f} \mid f \in \mathcal{F}\}$, where $\widetilde{f} \coloneqq \Pi_{\mathcal{F}}(f^*)$ denotes the $\|\cdot\|_{\mu}$-projection of f^* onto $\mathcal{F}$.

In the non-parametric fit of the Bellman residual, there are two sources of "noise". The first is the Bellman noise, and it leads to the *sub-Gaussian complexity* at scale $\delta > 0$ defined as

$$\mathcal{G}_n(\delta; \widetilde{\mathcal{F}}) \coloneqq \mathbb{E}\left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_{\mu} \leqslant \delta}\left|\frac{1}{n}\sum_{i=1}^n w_i g(x_i)\right|\right] \tag{20}$$

Here the x_i is generated i.i.d. from the stationary distribution μ, whereas the noise w_i is generated according to the model (16).

The second form of "noise" has to do with the discrepancy between the empirical norm $\|f\|_n \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(x_i)}$ and the population norm $\|f\|_{\mu} = \sqrt{\mathbb{E}[f^2(X)]}$. In order to prove a uniform bound on this discrepancy, we make use of the *local Rademacher complexity* of $\widetilde{\mathcal{F}}$ at scale $\delta > 0$ as

$$\mathcal{R}_n(\delta; \widetilde{\mathcal{F}}) \coloneqq \mathbb{E}\left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_{\mu} \leqslant \delta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)\right|\right]. \tag{21}$$

Here $\{x_i\}_{i=1}^n$ are i.i.d. samples from the stationary distribution μ, and $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables taking values in $\{-1, +1\}$ with equal probabilities, independent of $\{x_i\}_{i=1}^n$.

The two quantities ε_n and ω_n in the decomposition (19) are both quantities known as *critical radii* in empirical process theory. They are defined, respectively as the smallest positive solutions to the inequalities

$$\frac{\mathcal{G}_n(\varepsilon_n, \widetilde{\mathcal{F}})}{\varepsilon_n} \leqslant \frac{\varepsilon_n}{2}, \quad \text{and} \quad \frac{\mathcal{R}_n(\omega_n; \widetilde{\mathcal{F}})}{\omega_n} \leqslant \frac{\omega_n}{b}. \tag{22}$$

As long as the function class $\widetilde{\mathcal{F}}$ is convex (as assumed here), it can be shown that both of the functions $\delta \mapsto \frac{\mathcal{G}_n(\delta)}{\delta}$ and $\delta \mapsto \frac{\mathcal{R}_n(\delta)}{\delta}$ are non-increasing on the positive real line, and so the critical radii always exist. We refer the reader to Chapters 13 and 14 of the book (Wainwright, 2019) for more details on critical radii, and the arguments used to establish claims of this type. See the appendix for the critical radii computed for specific function classes.

3 NUMERICAL RESULTS

We illustrate the behavior of the KBB algorithm for a range of problems, including both idealized and semi-realistic settings. In performing these simulations, goal is to reveal some qualitative differences when compared with value iteration and fitted value iteration, as well as to provide empirical confirmation of the accuracy of our theoretical predictions. Recall that the Bellman operator $\mathcal{T}$ is γ-contractive with respect to $\|\cdot\|_{\mu}$, so we can iteratively apply $\mathcal{T}$ and ensures convergence to the true value function $V^{\star}$; this is the value iteration algorithm. The fitted value iteration (FVI) algorithm applies when the Bellman update *cannot* be computed $\mathcal{T}(V)$ exactly, and must be approximated. Given a dataset $\mathcal{D} = \{(x_i, r_i, x_i')\}_{i=1}^n$ and a function class $\mathcal{F}$, we compute an approximation via

$$\widehat{\mathcal{T}(V)} \in \arg\min_{f \in \mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n\left[r_i + \gamma V(x_i') - f(x_i)\right]^2\right\}.$$

In fitted value iteration, we iterate repeating this regression procedure for every step (potentially with a different dataset). Algorithm 1, in each iteration, computes the Bellman residual $\mathcal{B}(V_t) = V_t - \mathcal{T}V_t$ which involves comput-

ing the Bellman operator $\mathcal{T}V_t$, so we compare each iteration to one iteration of value iteration and fitted value iteration, using the same regression procedure and number of samples. In our implementations of Algorithm 1, we use $\mathcal{D}_t^{LSTD} = \mathcal{D}_t^{REG}$—that is, the same samples used in fitting the regression procedure are re-used for the LSTD fit. When running the algorithm, we have observed that ensuring the first regression is very accurate results in much better convergence. In the first step, Krylov-Bellman is estimating the reward function R. Having accurate estimates of the reward function is likely essential as the algorithm is effectively solving the linear system $\mathcal{Q}V = R$.

3.1 Circular random walk

For a more structured model, we create a tabular MDP with $|\mathcal{X}| = 200$ states as follows: indexing the states modulo 200 (i.e. state -2 is actually state 198), for a state x, it has $\frac{1}{3}$ probability of transitioning to itself, and then a $\frac{1}{6}$ probability of transitioning to states $x-2, x-1, x+1$, and $x+2$. The dynamics here model a random walk on the circle, and have a more interesting eigenstructure than the purely random transition models from the previous setting. As shown in Figure 4, the qualitative behavior of the algorithms is similar. The value iteration algorithms, either exact or fitted, still exhibits the linear behavior that is indicated by the γ-contractive nature of the Bellman operator $\mathcal{T}$. Krylov–Bellman Boosting is again much faster than either of the algorithms, especially in the case of $\gamma = 0.99$.

3.2 A non-linear example

Consider the nonlinear system in the 3-dimensional state variable $x_t = \begin{bmatrix} x_{t,1} & x_{t,2} & x_{t,3} \end{bmatrix} \in \mathbb{R}^3$ and control vector $u_t \in \mathbb{R}^3$ given by

$$\begin{bmatrix} x_{t+1,1} - x_{t+1,2}^2 \\ x_{t+1,2} \\ x_{t+1,3} - x_{t+1,1}^2 \end{bmatrix} = \mathbf{A} \begin{bmatrix} x_{t,1} - x_{t,2}^2 \\ x_{t,2} \\ x_{t,3} - x_{t,1}^2 \end{bmatrix} + \mathbf{B}u_t + w_t$$

with control and cost function given by

$$u_t = \mathbf{K} \begin{bmatrix} x_{t,1} - x_{t,2}^2 \\ x_{t,2} \\ x_{t,3} - x_{t,1}^2 \end{bmatrix}, \quad \text{and}$$

$$c_t(x_t, u_t) = \begin{bmatrix} x_{t,1} - x_{t,2}^2 \\ x_{t,2} \\ x_{t,3} - x_{t,1}^2 \end{bmatrix}^T \mathbf{Q} \begin{bmatrix} x_{t,1} - x_{t,2}^2 \\ x_{t,2} \\ x_{t,3} - x_{t,1}^2 \end{bmatrix} + u^T \mathbf{R} u_t.$$

Here $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{K}$ are all three-dimensional matrices, with each of $\mathbf{Q}$ and $\mathbf{R}$ positive semidefinite, whereas the random vector $w_t \in \mathbb{R}^3$ follows a zero-mean multivariate Gaussian random distribution. We choose a nonlinear control system of this form for ease of computation: for a system of this general type, despite being nonlinear, it is straightforward to perform exact value iteration via a change of coordinates

(see the book (Isidori, 1995) for further details). We generate the associated matrices $\mathbf{A}, \mathbf{B}, \mathbf{Q}, \mathbf{K}$, and $\mathbf{R}$ randomly.

The results of the algorithm are plotted in Figure 6. We still observe the linear behavior of value iteration in both settings, however this algorithm requires us to know the underlying nonlinear system exactly as well as its transition dynamics. In general, Algorithm 1 performs much better than fitted value iteration. When $\gamma = 0.9$, initially the algorithm is much faster than value iteration, but eventually it catches up and overtakes it due to the noise in the algorithm. In the setting of $\gamma = 0.99$, we see that Krylov–Bellman Boosting is much faster than either value iteration or fitted value iteration. However in this case, we are running Krylov–Bellman Boosting on the state space given by x_t, which has nonlinear transition dynamics and cost functions.

4 PROOF-SKETCH

In this section, we provide proof sketch for our first main result, Theorem 3. Due to space constraints, the full proof as well as the proof of the other results can be found in the appendix.

Roughly speaking, the idea underlying the proof is that one round of the KBB algorithm is at least as good as a gradient-type update that is *restricted to the subspace* $\mathcal{C}_t^\perp$. For this reason, the algorithm depends on the restricted spectral values (15) over this subspace, which determine the conditioning of the effective problem at each round.

In more detail, suppose both the LSTD and regression procedures are exact, and consider the problem of minimizing the functional $\mathcal{F}(V) \coloneqq \langle V - V^\star, \mathcal{Q}(V - V^\star) \rangle_\mu$. It has (Frechet) derivative $\nabla \mathcal{F}(V) = \mathcal{Q}(V - V^\star)$, and moreover, since the operator $\mathcal{Q}$ is positive and self-adjoint, this is a convex optimization problem, achieving its (unique) minimum at the true value function $V^\star$. Using these properties along with conditions defining LSTD optimality, we can prove that, for any choice of stepsize $\alpha > 0$, we have

$$\|V_{t+1} - V^\star\|_{\mathcal{Q}}^2 \leqslant \|V_t - \alpha \mathcal{Q}(V_t - V^\star) - V^\star\|_{\mathcal{Q}}^2$$

Using the fact that $\mathcal{Q}(V_t - V^\star) = \mathcal{B}(V_t)$ and then expanding the square yields

$$\begin{aligned} &\|V_t - \alpha\mathcal{B}(V_t) - V^\star\|_{\mathcal{Q}}^2 \\ &= \|V_t - V^\star\|_{\mathcal{Q}}^2 - 2\alpha\langle V_t - V^\star, \mathcal{B}(V_t)\rangle_{\mathcal{Q}} \\ &\quad + \alpha^2 \|\mathcal{B}(V_t)\|_{\mathcal{Q}} \\ &\overset{(i)}{=} \|V_t - V^\star\|_{\mathcal{Q}}^2 - 2\alpha\|\mathcal{B}(V_t)\|_\mu^2 + \alpha^2\|\mathcal{B}(V_t)\|_{\mathcal{Q}}^2, \end{aligned}$$

where step (i) makes use of the equivalence $\mathcal{Q}(V_t - V^\star) = \mathcal{B}(V_t)$. Since V_t is the LSTD fit over $\mathcal{S}_t$, the optimality conditions for the projected fixed point equation (1.2) ensure that $\mathcal{B}(V_t) \perp \mathcal{S}_t$, or equivalently $\mathcal{B}(V_t) \in \mathcal{S}_t^\perp$. As an important consequence, we can control $\|\mathcal{B}(V_t)\|_\mu$ and $\|\mathcal{B}(V_t)\|_{\mathcal{Q}}$
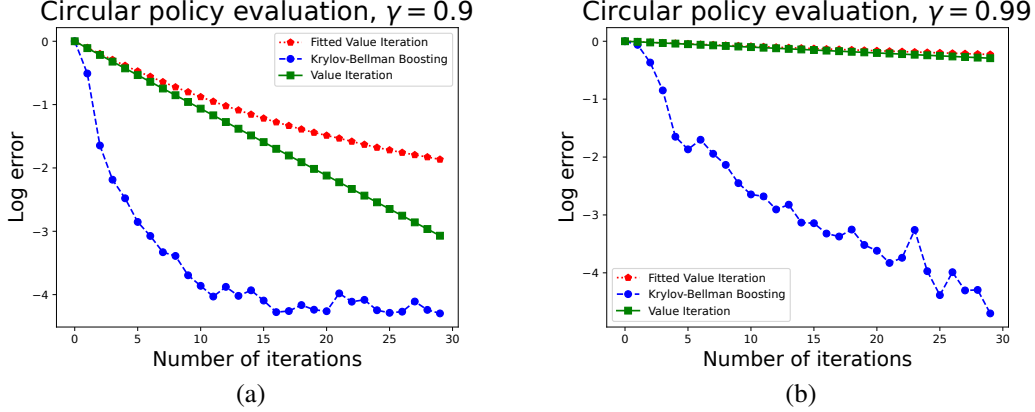
Figure 1. Illustration of the behavior of three different algorithms, value iteration (red), fitted value iteration (green), and Krylov-Bellman Boosting (blue). We plot the log error in the $\|\cdot\|_\mu$-norm (vertical axis) against the number of iterations used (horizonal axis). (a) Circular MRP with discount factor $\gamma = 0.9$. (b) Circular MRP with discount factor $\gamma = 0.99$.
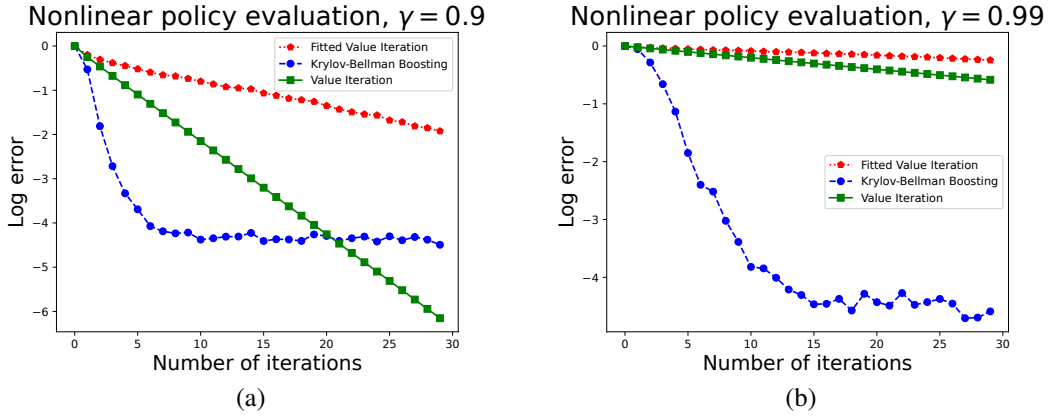

Figure 2. Illustration of the behavior of three different algorithms, value iteration (red), fitted value iteration (green), and Krylov-Bellman Boosting (blue) on the nonlinear policy evaluation problem. We plot the log error in the $\|\cdot\|_\mu$-norm (vertical axis) against the number of iterations used (horizontal axis). Algorithm 1 is still much faster than fitted-value iteration, but in the case with $\gamma = 0.9$ the Krylov–Bellman Boosting is faster than value iteration initially but because of the noise, value iteration catches up and overtakes it. (a) Discount factor $\gamma = 0.9$. (b) Discount factor $\gamma = 0.99$.

in terms of the restricted spectral values (15). We claim that the Bellman residual $\mathcal{B}(V^{\text{LSTD}})$ satisfies the bounds

$$\|\mathcal{B}(V_t)\|_\mathcal{Q}^2 \leqslant 2M_t\|V^{\text{LSTD}} - V^\star\|_\mathcal{Q}^2, \quad \text{and} \tag{23}$$
$$\|\mathcal{B}(V_t)\|_\mu^2 \geqslant m_t\|V^{\text{LSTD}} - V^\star\|_\mathcal{Q}^2.$$

Indeed, by combining these two inequalities we obtain

$$\|V_{t+1} - V^\star\|_\mathcal{Q} \leqslant \left(1 - 2\alpha m_t + 2\alpha^2 M_t\right)\|V_t - V^\star\|_\mathcal{Q}^2.$$

Setting $\alpha = \frac{m_t}{2M_t}$ yields

$$\|V_{t+1} - V^\star\|_\mathcal{Q} \leqslant \left(1 - \frac{m_t^2}{4M_t}\right)\|V_t - V^\star\|_\mathcal{Q}.$$

This provides on outline of the proof; the complete proof is more involved as it needs to address what arises when there are errors in computing the LSTD fit V_t and the Bellman residual $\mathcal{B}(V_t)$.

5 DISCUSSION

In this paper, we presented and analyzed the Krylov–Bellman boosting (KBB) procedure. It is an efficient algorithm for policy evaluation based on samples for general state spaces, that is built on the machinery of Krylov subspace methods for solving linear operator equations. As opposed to an idealized Krylov method, our approach allows for errors in both the LSTD steps and the fitting of the Bellman residual, and we provide general convergence guarantees that track the effect of such errors. On the empirical front, the KBB algorithm typically exhibits a superlinear rate of convergence, which is significantly faster than the linear or geometric rate obtained by fitted value iteration. Consistent with these empirical observations, our theory also reveals superlinear convergence. In particular, we show how the KBB convergence rate is determined by the spectral properties of the Bellman residual opera-

tor when restricted to a shrinking sequence of subspaces. Finally, we analyzed the use of non-parametric regression routines for fitting the Bellman residual. We proved a general theorem that bounds the residual fitting error in terms of a combination of approximation error, and statistical estimation error. There exist many interesting avenues for future work. While the current theory applies to reversible Markov chains, the method itself exhibits rapid convergence for general Markov reward processes. It would be interesting to close this gap between theory and practice by developing theory applicable to non-reversible Markov chains. Finally, given the promising nature of our sandbox simulations, it would be interesting to implement the KBB procedure on larger scale and more realistic MDP problems.

References

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted Q-iteration in continuous action-space MDPs. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957.

Dimitri P. Bertsekas. *Neuro-Dynamic Programming*. Springer US, Boston, MA, 2009. ISBN 978-0-387-74759-0.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, Belmont, MA, USA, 4th edition, 2017a.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmont, MA, USA, 4th edition, 2017b.

Dimitri P Bertsekas and Huizhen Yu. Projected equation methods for approximate solution of large linear systems. *Journal of Computational and Applied Mathematics*, 227 (1):27–50, 2009.

Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009. ISSN 0005-1098.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986. ISSN 0304-4076.

Justin A Boyan. Technical update: Least-squares temporal difference learning. *Machine learning*, 49(2):233–246, 2002.

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 1996.

Leo Breiman. Random forests. *Machine Learning*, 45(1): 5–32, 2001. ISSN 0885-6125.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 1984.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1042–1051. PMLR, 09–15 Jun 2019.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2.

Gal Dalal, Balazs Szorenyi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td(0) with function approximation. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982. ISSN 00129682, 14680262.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(18):503–556, 2005. URL http://jmlr.org/papers/v6/ernst05a.html.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55 (1):119–139, 1997. ISSN 0022-0000.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, 36(3):1171–1220, 2008.

R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

Alberto Isidori. *Nonlinear Control Systems, Third Edition*. Communications and Control Engineering. Springer, 1995. ISBN 978-1-4471-3909-6. doi: 10.1007/978-1-84628-615-5.

Sham Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Rudolf Kalman. Contributions to the theory of optimal control. *Boletin de la Sociedad Matematica Mexicana*, 5: 102–119, 1960.

Philipp W. Keller, Shie Mannor, and Doina Precup. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 449–456, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143901.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Society for Industrial and Applied Mathematics*, 42, 04 2001.

Erwin Kreyszig. *Introductory functional analysis with applications*, volume 17. John Wiley & Sons, 1991.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Ishai Menache, Shie Mannor, and Nahum Shimkin. Basis function adaptation in temporal difference reinforcement learning. *Annals of Operations Research*, 134, 05 2003. doi: 10.1007/s10479-005-5732-z.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008.

Ronald Parr, Christopher Painter-Wakefield, Lihong Li, and Michael Littman. Analyzing feature generation for value-function approximation. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 737–744, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933.

Ronald E. Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L. Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *ICML '08*, 2008.

Marek Petrik. An analysis of Laplacian methods for value function approximation in MDPs. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2574–2579, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

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

Martin L. Puterman and Shelby L. Brumelle. On the convergence of policy iteration in stationary dynamic programming. *Mathematics of Operations Research*, 4(1): 60–69, 1979.

Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003. ISBN 978-0-89871-534-7. doi: 10.1137/1.9780898718003.

Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, 16(49): 1629–1676, 2015.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. ISSN 0893-6080.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. *31st International Conference on Machine Learning, ICML 2014*, 1, 06 2014.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Michel Talagrand. A new look at independence. *The Annals of Probability*, 24(1):1 – 34, 1996.

Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer, 2014.

Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990. doi: 10.1137/1.9781611970128.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, (8):229 – 256, 1992.

Huizhen Yu and Dimitri P. Bertsekas. Error bounds for approximations from projected linear equations. *Mathematics of Operations Research*, 35(2):306–329, 2010. ISSN 0364765X, 15265471.

A Complete numerical results

We illustrate the behavior of our algorithm in a mixture of idealized and semi-realistic settings. Appendix A.1 highlights the behavior of our algorithm in the tabular setting, where $\mathcal{X}$ is finite, and Appendix A.2 illustrates the performance in the linear-quadratic setting, where $\mathcal{X} = \mathbb{R}^d$, but the transitions and costs have simple representations. Appendix A.3 demonstrates the behavior of our algorithm for a nonlinear system with more complicated dynamics.

We compare our algorithm to value iteration and fitted value iteration. Recall that the Bellman operator $\mathcal{T}$ is γ-contractive with respect to $\|\cdot\|_\mu$, so we can iteratively apply $\mathcal{T}$ and ensures convergence to the true value function $V^\star$; this is the value iteration algorithm. The fitted value iteration (FVI) algorithm applies when the Bellman update *cannot* be computed $\mathcal{T}(V)$ exactly, and must be approximated. Given a dataset $\mathcal{D} = \{(x_i, r_i, x_i')\}_{i=1}^n$ and a function class $\mathcal{F}$, we compute an approximation via

$$\widehat{\mathcal{T}(V)} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \left[r_i + \gamma V(x_i') - f(x_i) \right]^2 \right\}.$$

In fitted value iteration, we iterate repeating this regression procedure for every step (potentially with a different dataset). Algorithm 1, in each iteration, computes the Bellman residual $\mathcal{B}(V_t) = V_t - \mathcal{T}V_t$ which involves computing the Bellman operator $\mathcal{T}V_t$, so we compare each iteration to one iteration of value iteration and fitted value iteration, using the same regression procedure and number of samples. In our implementations of Algorithm 1, we use $\mathcal{D}_t^{LSTD} = \mathcal{D}_t^{REG}$—that is, the same samples used in fitting the regression procedure are re-used for the LSTD fit. When running the algorithm, we have observed that ensuring the first regression is very accurate results in much better convergence. In the first step, Krylov-Bellman is estimating the reward function R. Having accurate estimates of the reward function is likely essential as the algorithm is effectively solving the linear system $\mathcal{Q}V = R$.

A.1 Tabular MRPs

In a tabular MRP, the state space $\mathcal{X}$ is finite; for ease of notation we represent it as $\mathcal{X} = \{1, 2, \ldots, d\}$ where $d = |\mathcal{X}|$. We can then represent all quantities of interest as either a matrix or a vector: the transition operator $\mathcal{P}$ can be written as a matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ and the reward function R can be identified with a vector in $R \in \mathbb{R}^{|\mathcal{X}|}$ via

$$R_x \coloneqq R(x) \qquad \text{and} \qquad \mathbf{P}_{x,x'} = \mathcal{P}(x' \mid x) \qquad \text{for all } x \in \mathcal{X},$$

where $\mathcal{P}(x' \mid x)$ is the probability of transition to x' from x. Similarly we can represent a value function V as an equivalent vector in $\mathbb{R}^{|\mathcal{X}|}$. The Bellman operator can then be written in the matrix form as $\mathcal{T}V = R + \gamma \mathbf{P}V$. In this setting we perform the regression procedure over the function class $\mathcal{F} = \mathbb{R}^{|\mathcal{X}|}$, and the regression is equivalent to computing the sample average for every state.

Random Transitions: In our first example of a tabular MRP, we create a $|\mathcal{X}| = 300$ dimensional MRP by generating a transition matrix $\mathbf{P} \in \mathbb{R}^{300 \times 300}$ with i.i.d. $\text{Unif}(0,1)$ entries, with each row rescaled so as to ensure row-stochasticity. Similarly, we generate a reward vector $R \in \mathbb{R}^{300}$ with i.i.d $\text{Unif}(0,1)$ entries. We generate such independent samples of such models, and set the discount parameter as $\gamma = 0.9$ in one case, and $\gamma = 0.99$ in the other. The results are plotted in Figure 3. The plots clearly illustrate the advantages of Algorithm 1 over fitted value iteration and value iteration. As the KBB error becomes quite small, we see the error floor (due to statistical error in evaluating the residual) start to dominate, as should be expected.

Circular Random Walk: For a more structured model, we create a tabular MDP with $|\mathcal{X}| = 200$ states as follows: indexing the states modulo 200 (i.e. state -2 is actually state 198), for a state x, it has $\frac{1}{3}$ probability of transitioning to itself, and then a $\frac{1}{6}$ probability of transitioning to states $x - 2, x - 1, x + 1$, and $x + 2$. The dynamics here model a random walk on the circle, and have a more interesting eigenstructure than the purely random transition models from the previous setting. As shown in Figure 4, the qualitative behavior of the algorithms is similar. The value iteration algorithms, either exact or fitted, still exhibits the linear behavior that is indicated by the γ-contractive nature of the Bellman operator $\mathcal{T}$. Krylov–Bellman Boosting is again much faster than either of the algorithms, especially in the case of $\gamma = 0.99$.

A.2 Linear Quadratic Regulator

The linear quadratic regulator (LQR) is a canonical model in optimal control; it plays a central role in many applications, including aeronautics, robotics and industrial process control (Bertsekas, 2017a; Kalman, 1960). Here we consider a
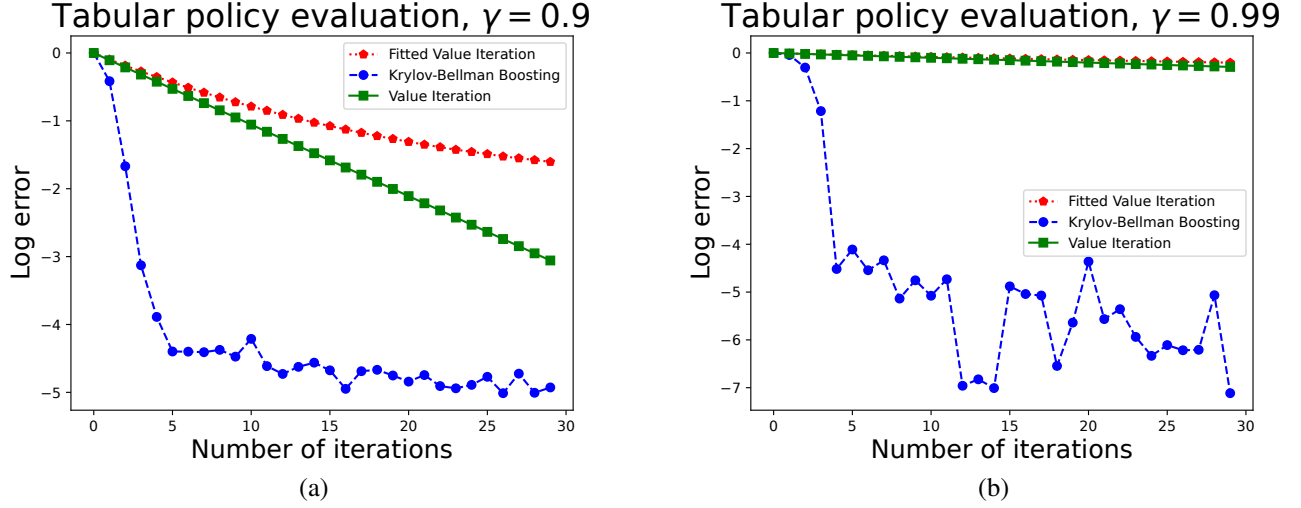
Figure 3. Illustration of the behavior of three different algorithms, value iteration (red), fitted value iteration (green), and Krylov Bellman Boosting (blue) on the random tabular MRP. We plot the log error in the $\|\cdot\|_\mu$-norm (vertical axis) against the number of iterations used (horizontal axis). (a) MRP with discount factor $\gamma = 0.9$. Value iteration excepts the characteristic linear plot that reflects the γ-contractive nature of value iteration. Fitted value iteration approximates value iteration, and performs worse as a result of being an inexact procedure. Krlov-Bellman Boosting performs much better than either of these algorithms, although we see that its convergence slows down drastically and becomes somewhat unstable. (b) MRP with discount factor $\gamma = 0.99$. Like previously, value-iteration exhibits linear plot that illustrates the contractivity of the Bellman operator $\mathcal{T}$, albeit at a much slower rate. Fitted value iteration is also still slower than value iteration, and Krylov–Bellman Boosting is much faster than both of them, but exhibits much greater instability in this setting.

particular variant of the LQR problem known as the linear quadratic Gaussian (LQG) problem. The state space $\mathcal{X} = \mathbb{R}^d$ and action space $\mathcal{U} = \mathbb{R}^v$ are continuous, and the transition dynamics starting from state $x_t \in \mathbb{R}^d$ and taking action $u \in \mathbb{R}^v$ is given by

$$x_{t+1} = \mathbf{A}x_t + \mathbf{B}u_t + W_t,$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is the transition matrix, $\mathbf{B} \in \mathbb{R}^{d \times v}$ is the action matrix, and $w_t \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. We formulate this problem in terms of a cost function, rather than a reward function (although for the purposes for policy evaluation it is a meaningless distinction). Given symmetric positive semidefinite matrices $\mathbf{Q} \in \mathbb{R}^{d \times d}$ and $\mathbf{R} \in \mathbb{R}^{v \times v}$, the cost function at each stage is

$$\psi(x, u) = x^T \mathbf{Q} x + u^T \mathbf{R} u.$$

We restrict our attention to policies of the form $u_t = \mathbf{K} x_t$ where $\mathbf{K} \in \mathbb{R}^{v \times d}$. We then have a Markov reward process with dynamics and costs given by

$$x_{t+1} = (\mathbf{A} + \mathbf{B}\mathbf{K})x_t + w_t \quad \text{and} \quad \psi(x_t) = x_t^T (\mathbf{Q} + \mathbf{K}^T \mathbf{R} \mathbf{K}) x_t.$$

Assume that the eigenvalues of $\mathbf{A} + \mathbf{B}\mathbf{K}$ are contained within the unit circle on the complex plane. The goal is to solve for the value function $V^\star$ of this MRP. The value function can be written as

$$V^\star(x) = x^T \mathbf{P}^* x + \frac{\gamma}{1-\gamma} \cdot \text{trace}(\mathbf{P}^* \boldsymbol{\Sigma}),$$

where the matrix $\mathbf{P}^* \in \mathbb{R}^{d \times d}$ is the solution to the Lyapunov equation

$$\mathbf{P}^* = \mathbf{Q} + \mathbf{K}^T \mathbf{R} \mathbf{K} + \gamma (\mathbf{A} + \mathbf{B}\mathbf{K})^T \mathbf{P}^* (\mathbf{A} + \mathbf{B}\mathbf{K}).$$

Value iteration involves computing the update

$$\mathbf{P}_{t+1} = \mathbf{Q} + \mathbf{K}^T \mathbf{R} \mathbf{K} + \gamma (\mathbf{A} + \mathbf{B}\mathbf{K})^T \mathbf{P}_t (\mathbf{A} + \mathbf{B}\mathbf{K}) \quad \text{and} \quad c_{t+1} = \gamma c_t + \gamma \text{trace}(\mathbf{P}_t \boldsymbol{\Sigma}).$$

The pair $(\mathbf{P}_y, c_t)$ define the value function via the value function via $V_t(x) = x^T \mathbf{P}_t x + c_t$. When $\mathbf{A} + \mathbf{B}\mathbf{K}$ has eigenvalues contained within the unit circle, this update is guaranteed to be a contraction.
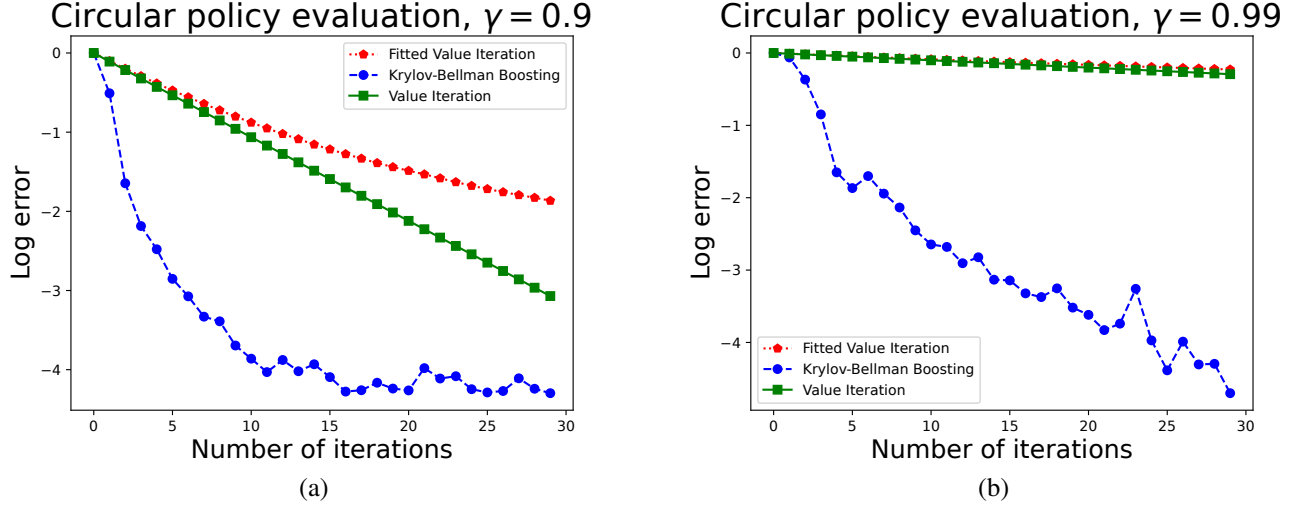
Figure 4. Illustration of the behavior of three different algorithms, value iteration (red), fitted value iteration (green), and Krylov Bellman Boosting (blue) on the random tabular MRP. We plot the log error in the $\|\cdot\|_\mu$-norm (vertical axis) against the number of iterations used (horizontal axis). The algorithms in this setting perform quite similarly to those in Figure 3, although the Krylov–Bellman Boosting algorithm seems to be more stable in this setting than in the previous one. (a) Circular MRP with discount factor $\gamma = 0.9$. (b) Circular MRP with discount factor $\gamma = 0.99$.

In the numerical simulations, we choose $d = 5$ and $v = 3$. The matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{K}$ are chosen to have i.i.d. Unif$(0, 1)$ entries, and then rescaled to ensure that the eigenvalues are contained within the unit circle. The symmetric matrices $\mathbf{Q}$ and $\mathbf{R}$ are computed by generating matrices with i.i.d. Unif$(0, 1)$ random variables and then left-multiplied by its transpose to ensure symmetry. The regression steps in both fitted value iteration and Krylov–Bellman Boosting use the XGBoost routine (Chen and Guestrin, 2016) to perform the fitting. Result are given in Figure 5. As before, we observe a geometric (or linear) convergence rate for value iteration, consistent with the γ-contractive nature of the update. We also see approximately linear convergence of fitted value iteration, albeit at a slower rate compared to value iteration. Krylov–Bellman Boosting is again more rapidly convergent than these two procedures. (We note that KBB hits an error floor due to the statistical sampling involved, whereas exact value iteration, which is based on unrealistic knowledge of the true dynamics, does not involve any such error.) For a larger discount factor $\gamma = 0.99$, the gains afforded by the KBB procedure become more significant.

A.3 A non-linear example

Consider the nonlinear system in the 3-dimensional state variable $x_t = \begin{bmatrix} x_{t,1} & x_{t,2} & x_{t,3} \end{bmatrix} \in \mathbb{R}^3$ and control vector $u_t \in \mathbb{R}^3$ given by

$$\begin{bmatrix} x_{t+1,1} - x_{t+1,2}^2 \\ x_{t+1,2} \\ x_{t+1,3} - x_{t+1,1}^2 \end{bmatrix} = \mathbf{A} \begin{bmatrix} x_{t,1} - x_{t,2}^2 \\ x_{t,2} \\ x_{t,3} - x_{t,1}^2 \end{bmatrix} + \mathbf{B}u_t + w_t$$

with control and cost function given by

$$u_t = \mathbf{K} \begin{bmatrix} x_{t,1} - x_{t,2}^2 \\ x_{t,2} \\ x_{t,3} - x_{t,1}^2 \end{bmatrix}, \quad \text{and} \quad c_t(x_t, u_t) = \begin{bmatrix} x_{t,1} - x_{t,2}^2 \\ x_{t,2} \\ x_{t,3} - x_{t,1}^2 \end{bmatrix}^T \mathbf{Q} \begin{bmatrix} x_{t,1} - x_{t,2}^2 \\ x_{t,2} \\ x_{t,3} - x_{t,1}^2 \end{bmatrix} + u^T \mathbf{R}u_t.$$

Here $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{K}$ are all three-dimensional matrices, with each of $\mathbf{Q}$ and $\mathbf{R}$ positive semidefinite, whereas the random vector $w_t \in \mathbb{R}^3$ follows a zero-mean multivariate Gaussian random distribution. We choose a nonlinear control system of this form for ease of computation: for a system of this general type, despite being nonlinear, it is straightforward to perform exact value iteration via a change of coordinates (see the book (Isidori, 1995) for further details). In the simple case where
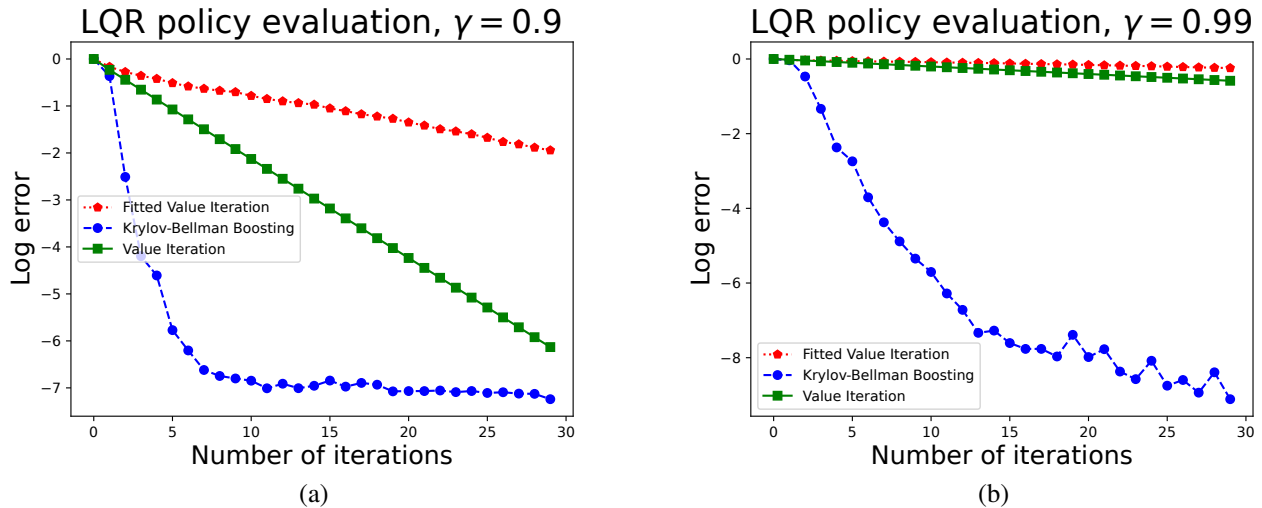
Figure 5. Illustration of the behavior of three different algorithms, value iteration (red), fitted value iteration (green), and Krylov Bellman Boosting (blue) on the LQR policy evaluation problem. We plot the log error in the $\|\cdot\|_\mu$-norm (vertical axis) against the number of iterations used (horizontal axis). The algorithms in this setting perform quite similarly to those in Figure 3, although the Krylov–Bellman Boosting algorithm seems to be more stable in this setting than in the previous one. (a) Linear Quadratic Regulator with discount factor $\gamma = 0.9$. (b) Linear Quadratic Regulator with discount factor $\gamma = 0.99$.

$w_t = 0$ and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

we can verify that the system has the following state space representation

$$x_{t+1,1} = x_{t,2} + \left[x_{t,3} - x_{t,1}^2 + u_{t,1} \right]^2,$$
$$x_{t+1,2} = x_{t,3} - x_{t,1}^2 + u_{t,1},$$
$$x_{t+1,3} = x_{t,1} + 2x_{t,2} \left[x_{t,3} - x_{t,1}^2 + u_{t,1} \right]^2 + \left[x_{t,3} - x_{t,1}^2 + u_{t,1} \right]^4.$$

We generate the associated matrices $\mathbf{A}, \mathbf{B}, \mathbf{Q}, \mathbf{K}$, and $\mathbf{R}$ in the same manner as Appendix A.2.

The results of the algorithm are plotted in Figure 6. We still observe the linear behavior of value iteration in both settings, however this algorithm requires us to know the underlying linear system exactly as well as its transition dynamics. In general, Algorithm 1 performs much better than fitted value iteration. When $\gamma = 0.9$, initially the algorithm is much faster than value iteration, but eventually it catches up and overtakes it due to the noise in the algorithm. In the setting of $\gamma = 0.99$, we see that Krylov–Bellman Boosting is much faster than either value iteration or fitted value iteration. However in this case, we are running Krylov–Bellman Boosting on the state space given by x_t, which has nonlinear transition dynamics and cost functions.

A.4   ARCH

For the final example, we consider the autoregressive conditional heteroskedasticity model (ARCH) (Engle, 1982; Bollerslev, 1986). The state vectors X_t evolve according to the non-linear dynamics

$$X_{t+1} = \mathbf{A}X_t + \sqrt{q + X_t \mathbf{Q} X_t} \cdot W_t$$

where $q \geqslant 0$ is a scalar, the matrix $\mathbf{Q} \in \mathbb{R}^{d \times d}$ is positive semidefinite, and $W_t \sim \mathcal{N}(0, \mathbf{\Sigma})$ is system noise. We turn this into a Markov reward process by assigning a cost function $\psi(x) = x^T \mathbf{R} x$ and adding discounting. Then the value function $V^\star$ is the solution to the Bellman relation

$$V(x) = \mathcal{T}(V)(x) := x^T \mathbf{R} x + \gamma \mathbb{E}_W \left[V(\mathbf{A}x + \sqrt{q + x^T \mathbf{Q} x} \cdot W) \right].$$
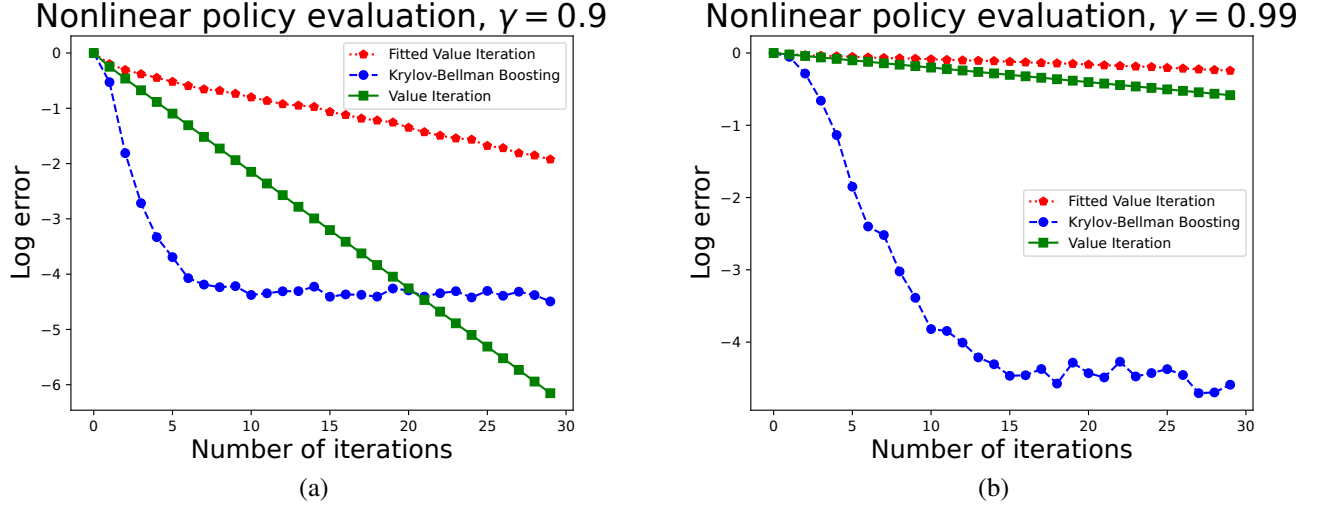
Figure 6. Illustration of the behavior of three different algorithms, value iteration (red), fitted value iteration (green), and Krylov Bellman Boosting (blue) on the nonlinear policy evaluation problem. We plot the log error in the $\|\cdot\|_\mu$-norm (vertical axis) against the number of iterations used (horizontal axis). Algorithm 1 is still much faster than fitted-value iteration, but in the case with $\gamma = 0.9$ the Krylov–Bellman Boosting is faster than value iteration initially but because of the noise, value iteration catches up and overtakes it. (a) Discount factor $\gamma = 0.9$. (b) Discount factor $\gamma = 0.99$.

Under some regularity conditions on the triple $(q, \mathbf{A}, \mathbf{R})$, the Bellman operator $\mathcal{T}$ is γ-contractive under the norm $\|\cdot\|_\mu$. Consequently, we can solve for the value function via value iteration in the same manner as previous settings. A straightforward calculation yields

$$V^\star(x) = x^T \mathbf{P}^* x + \frac{\gamma q}{1 - \gamma} \cdot \text{trace}(\mathbf{P}^* \mathbf{\Sigma})$$

where the matrix $\mathbf{P}^* \in \mathbb{R}^{d \times d}$ solves the linear system

$$\mathbf{P}^* = \mathbf{R} + \gamma \left\{ \mathbf{A}^T \mathbf{P}^* \mathbf{A} + \mathbf{Q} \cdot \text{trace}(\mathbf{P}^* \mathbf{\Sigma}) \right\}.$$

In our numerical simulations we choose $d = 5$, and $q = 0.5$. The matrix $\mathbf{A}$ is chosen to have i.i.d. Unif$(0, 1)$ entries, and then rescaled to ensure that the system is stable. The matrix $\mathbf{Q}$ is computed by generating a matrix with i.i.d. Unif$(0, 1)$ random variables, and then left-multiplying by its transpose to ensure symmetry. The results are plotted in Figure 7. The behavior in this setting seems to be the same as what we observed in the past few settings. Initially, Krylov–Bellman Boosting is much faster than both fitted value iteration and value iteration. In the setting of $\gamma = 0.9$, Krylov Bellman Boosting hits an error floor relatively early, and it takes value iteration and fitted value iteration quite a while to catch up. For $\gamma = 0.99$, we observe a similar behavior, but both value iteration and fitted value iteration take significantly longer to match Krylov–Bellman Boosting in accuracy.

B Proofs

In this section, we present a complete proof of our results. For the convenience of our reader, we restate the major results and assumptions.

KBB General Guarantee

We have the following guarantees on the steps in the KBB procedure: Let $W_t := \phi_t - \mathcal{B}(V_{t-1})$ denote the error in the regression procedure at round t. Our first assumption provides control on the regression accuracy:

Assumption 3 (Reg-Err). *For each iteration $t = 1, 2, \ldots$, the regression procedure is δ_t-accurate:*

$$\mathbb{E}\big[\|W_t\|_\mathcal{Q}^2\big] \leqslant \delta_t^2. \tag{24}$$
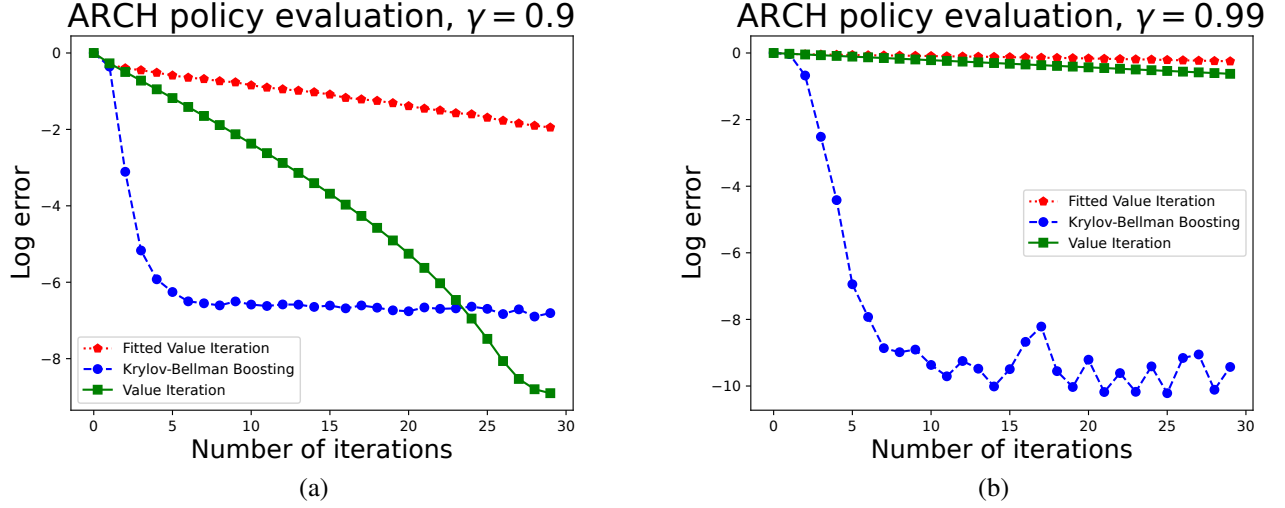
Figure 7. Illustration of the behavior of three different algorithms, value iteration (red), fitted value iteration (green), and Krylov Bellman Boosting (blue) on the ARCH policy evaluation problem. We plot the log error in the $\|\cdot\|_\mu$-norm (vertical axis) against the number of iterations used (horizonal axis). Algorithm 1 is still much faster than fitted-value iteration, but in the case with $\gamma = 0.9$ the Krylov–Bellman Boosting is faster than value iteration initially but because of the noise, value iteration catches up and overtakes it. (a) ARCH with discount factor $\gamma = 0.9$. (b) ARCH with discount factor $\gamma = 0.99$

Our second assumption provides control on the accuracy of the LSTD computation at each step:

Assumption 4 (LSTD-Err). *At each iteration t, the approximate LSTD solution V_t is ϵ^{LSTD}-accurate:*

$$\|V_t - V^\star\|_{\mathcal{Q}}^2 \overset{(a)}{\leqslant} \|V_t^{\mathrm{LSTD}} - V^\star\|_{\mathcal{Q}}^2 + (\epsilon^{\mathrm{LSTD}})^2, \quad and$$

$$\|\mathcal{B}(V_t) - \mathcal{B}(V_t^{\mathrm{LSTD}})\|_{\mathcal{Q}} \overset{(b)}{\leqslant} \epsilon^{\mathrm{LSTD}}.$$

(25)

Theorem 3 (General KBB bound). *Suppose that the Krylov–Bellman procedure is run with a δ_t-accurate regression procedure (cf.24), and an ϵ^{LSTD}-accurate LSTD implementation (cf. 25). Then at each step $t = 1, 2, \ldots$, the error satisfies the $\mathcal{Q}$-norm bound*

$$\mathbb{E}_{t+1}\|V_{t+1} - V^\star\|_{\mathcal{Q}}^2 \leqslant \left(1 - \frac{m_t^2}{8M_t}\right)\|V_t - V^\star\|_{\mathcal{Q}}^2 + \frac{34}{M_t} \cdot \delta_t^2 + \frac{8M_t}{m_t^2} \cdot (\epsilon_t^{\mathrm{LSTD}})^2.$$

Regression bound

We prove the following guarantee on the regression procedure over some function class $\mathcal{F}$.

Theorem 4 (Regression bound). *For a given function $V \in L^2(\mathcal{X}, \mu)$, suppose that we compute an estimate $\widehat{f}$ of the Bellman residual $\mathcal{B}(V)$ via the regression over a b-bounded and convex function class $\mathcal{F}$. Then we have*

$$\mathbb{E}\|\widehat{f} - \mathcal{B}(V)\|_\mu^2 \leqslant c\left\{\omega_n^2 + \varepsilon_n^2 + \frac{1}{n}\left(b^2 + \|V\|_\infty^2\right) + \|\Pi_{\mathcal{F}}(\mathcal{B}(V)) - \mathcal{B}(V)\|_\mu^2\right\},$$

(26)

where the statistical estimation error ω_n^2 and ε_n^2 depends on the complexity of $\mathcal{F}$.

We begin by proving Theorem 1 in Appendix B.1. This proof relies on a result of possible interest—stated as Proposition 1 and proved in Appendix B.2—that controls the one-step of the algorithm when the LSTD calculations are exact.

B.1 Proof of Theorem 3

Our proof of this result consists of two parts:

(a) First, we analyze an idealized version of the algorithm in which there is no error in the LSTD calculation. We summarize our conclusions in Proposition 1 stated below.

(b) Second, we leverage our understanding of the "exact-LSTD" setting in order to analyze the general algorithm. Our strategy is to reduce the general update to perturbed form of the "exact-LSTD" update.

We begin by bounding the behavior of an update that is not actually used in the algorithm itself, but is a useful auxiliary quantity. In particular, suppose that we are given an LSTD solution V^{LSTD}, computed from a subspace $\mathcal{S}$ with spectral parameters (m, M), as previously defined (15). For a stepsize $\alpha > 0$, define the update

$$\mathcal{F}_\alpha(V^{\text{LSTD}}) = V^{\text{LSTD}} - \alpha\{\mathcal{B}(V^{\text{LSTD}}) + W\},$$

where W represents noise in the computation of the Bellman residual. We let $b = \mathbb{E}[W]$ and $\sigma^2 = \mathbb{E}\|W - b\|_\mathcal{Q}^2$ denote the bias and variance, respectively, of this noise. Recall that by our regression-noise assumption 24, the mean-squared error satisfies the upper bound $\|b\|_\mathcal{Q}^2 + \sigma^2 \leq \delta^2$—we say that the regression procedure is δ-accurate for short.

Proposition 1. *Given a δ-accurate regression procedure and a stepsize $\alpha \in [0, \frac{1}{2}]$, for any LSTD solution V^{LSTD}, the update $\mathcal{F}_\alpha(V^{\text{LSTD}})$ satisfies the bound*

$$\mathbb{E}\|\mathcal{F}_\alpha(V^{\text{LSTD}}) - V^\star\|_\mathcal{Q}^2 \leq \left(1 - \tfrac{3\alpha m}{2} + 2\alpha^2 M\right)\|V^{\text{LSTD}} - V^\star\|_\mathcal{Q}^2 + \left(\tfrac{8\alpha}{m} + \alpha^2\right)\|b\|_\mathcal{Q}^2 + \alpha^2 \cdot \sigma^2. \tag{27a}$$

where $\mathbb{E}$ denotes expectation over the regression noise W. In particular, for the choice $\alpha = \frac{m}{2M}$, we have

$$\mathbb{E}\|\mathcal{F}_\alpha(V^{\text{LSTD}}) - V^\star\|_\mathcal{Q}^2 \leq \left(1 - \tfrac{m^2}{4M}\right)\|V^{\text{LSTD}} - V^\star\|_\mathcal{Q}^2 + \tfrac{5}{M}\|b\|_\mathcal{Q}^2 + \tfrac{1}{2M} \cdot \sigma^2. \tag{27b}$$

See Appendix B.2 for the proof of this proposition.

In addition, we require another auxiliary result, which characterizes the LSTD solution as a projection under the $\mathcal{Q}$-norm:

Lemma 1. *For a reversible Markov chain, the LSTD estimate V^{LSTD} is the projection of $V^\star$ onto $\mathcal{S}$ under the $\|\cdot\|_\mathcal{Q}$-norm— viz.*

$$V^{\text{LSTD}} = \arg\min_{V \in \mathcal{S}} \|V - V^\star\|_\mathcal{Q}^2. \tag{28}$$

See Appendix B.3 for the proof.

Equipped with these two auxiliary results, we are now ready to prove Theorem 3, which applies to the algorithm that includes errors in both the regression and LSTD phase. Focusing on the update from V_t to V_{t+1}, we need to bound the error $\|V_{t+1} - V^\star\|_\mathcal{Q}$ in terms of the error $\|V_t - V^\star\|_\mathcal{Q}$. In order to do so, our analysis also involves the exact LSTD solutions V_{t+1}^{LSTD} and V_t^{LSTD}; let us emphasize that these quantities are *not* actually computed in the algorithm itself.

Condition (a) in the definition 25 of LSTD accuracy allows us to relate V_{t+1} to V_{t+1}^{LSTD} via the inequality

$$\|V_{t+1} - V^\star\|_\mathcal{Q}^2 \leq \|V_{t+1}^{\text{LSTD}} - V^\star\|_\mathcal{Q}^2 + (\epsilon_t^{\text{LSTD}})^2. \tag{29a}$$

Next, we observe that for any scalar α, the function $V_t^{\text{LSTD}} - \alpha\big(\mathcal{B}(V_t) + W_{t+1}\big)$ is an element of $\mathcal{S}_{t+1}$, since $\phi_{t+1} = \mathcal{B}(V_t) + W_{t+1}$ is the basis function added at round t. Consequently, applying Lemma 1 with V_{t+1}^{LSTD} and the subspace $\mathcal{S}_{t+1}$ guarantees that

$$\|V_{t+1}^{\text{LSTD}} - V^\star\|_\mathcal{Q}^2 \leq \|V_t^{\text{LSTD}} - \alpha\big(\mathcal{B}(V_t) + W_{t+1}\big) - V^\star\|_\mathcal{Q}^2. \tag{29b}$$

Now observe that, by definition of the one-step update $\mathcal{F}_\alpha$, we have the equivalence

$$V_t^{\text{LSTD}} - \alpha\big(\mathcal{B}(V_t) + W_{t+1}\big) = \mathcal{F}_\alpha(V_t^{\text{LSTD}}) + \alpha\{\mathcal{B}(V_t^{\text{LSTD}}) - \mathcal{B}(V_t)\}.$$

Using this decomposition and applying the Fenchel-Young inequality with a parameter $\tau > 0$ to be chosen, we find that

$$\|V_{t+1} - V^\star\|_\mathcal{Q}^2 \leq (1 + \tau)\|\mathcal{F}_\alpha(V_t^{\text{LSTD}}) - V^\star\|_\mathcal{Q}^2 + \left(1 + \tfrac{1}{\tau}\right)\|\mathcal{B}(V_t) - \mathcal{B}(V_t^{\text{LSTD}})\|_\mathcal{Q}^2 + (\epsilon_t^{\text{LSTD}})^2$$

$$\overset{(i)}{\leq} (1 + \tau)\|\mathcal{F}_\alpha(V_t^{\text{LSTD}}) - V^\star\|_\mathcal{Q}^2 + \left(2 + \tfrac{1}{\tau}\right)(\epsilon_t^{\text{LSTD}})^2,$$

where step (i) follows inequality (b) in the definition of ϵ^{LSTD}-accuracy for the LSTD computation.

We can apply Proposition 1 with the choice $\alpha = \frac{m_t}{2M_t}$ to upper bound the first term, thereby obtaining

$$\mathbb{E}_{t+1}\big[\|\mathcal{F}_\alpha(V_t^{\mathrm{LSTD}}) - V^\star\|_{\mathcal{Q}}^2\big] \leqslant \kappa_t\|V_t^{\mathrm{LSTD}} - V^\star\|_{\mathcal{Q}}^2 + \frac{17}{M_t}\|b_{t+1}\|_{\mathcal{Q}}^2 + \frac{1}{2M_t}\cdot\sigma_t^2.$$

where we have introduced the shorthand $\kappa_t \coloneqq 1 - \frac{m_t^2}{4M_t} \in (0,1)$. Putting together the pieces, we have

$$\mathbb{E}_{t+1}\|V_{t+1} - V^\star\|_{\mathcal{Q}}^2 \leqslant (1+\tau)\,\kappa_t\|V_t^{\mathrm{LSTD}} - V^\star\|_{\mathcal{Q}}^2 + (1+\tau)\left(\frac{17}{M_t}\|b_{t+1}\|_{\mathcal{Q}}^2 + \frac{1}{2M_t}\cdot\sigma_t^2\right) + \left(2+\frac{1}{\tau}\right)(\epsilon^{\mathrm{LSTD}})^2.$$

Setting the Fenchel-Young parameter as $\tau = \frac{1-\kappa_t}{2\kappa_t}$ yields

$$(1+\tau)\,\kappa_t = \frac{1+\kappa_t}{2} = 1 - \frac{m_t^2}{8M_t}.$$

We also have $\tau = \frac{1-\kappa_t}{2\kappa_t} = \frac{m_t^2}{8M_t-2m_t^2} \leqslant \frac{1}{2}$, using the fact that $m_t^2 \leqslant 2M_t$ in the inequality. Moreover, we observe that

$$2 + \frac{1}{\tau} = 2 + \frac{2\kappa_t}{1-\kappa_t} = \frac{2}{1-\kappa_t} = \frac{8M_t}{m_t^2}.$$

Putting together the pieces, we conclude

$$\mathbb{E}_{t+1}\|V_{t+1} - V^\star\|_{\mathcal{Q}}^2 \leqslant \left(1 - \frac{m_t^2}{8M_t}\right)\|V_t^{\mathrm{LSTD}} - V^\star\|_{\mathcal{Q}}^2 + \frac{34}{M_t}\cdot\|b_{t+1}\|_{\mathcal{Q}}^2 + \frac{1}{M_t}\cdot\sigma_t^2 + \frac{8M_t}{m_t^2}\cdot(\epsilon^{\mathrm{LSTD}})^2. \tag{30}$$

Finally, since both V_t^{LSTD} and V_t belong to the subspace $\mathcal{S}_t$, Lemma 1 guarantees that

$$\|V_t^{\mathrm{LSTD}} - V^\star\|_{\mathcal{Q}}^2 \leqslant \|V_t - V^\star\|_{\mathcal{Q}}^2.$$

Substituting this upper bound into the inequality (30) and using the fact that $\|b_{t+1}\|_{\mathcal{Q}}^2 + \sigma_t^2 \leqslant \delta_t^2$ yields the claim.

B.2 Proof of Proposition 1

Let V^{LSTD} be an LSTD solution defined by the subspace $\mathcal{S}$, and let (m, M) be the associated spectral quantities. Recall that our goal is to control the behavior of the operator $\mathcal{F}_\alpha(V^{\mathrm{LSTD}}) = V^{\mathrm{LSTD}} - \alpha\big\{\mathcal{B}(V^{\mathrm{LSTD}}) + W\big\}$, where W is the noise in the evaluation of the Bellman residual.

We first claim that the following decomposition holds

$$\mathbb{E}\|\mathcal{F}_\alpha(V^{\mathrm{LSTD}}) - V^\star\|_{\mathcal{Q}}^2 = T_1 + T_2 + \alpha^2\|b\|_{\mathcal{Q}}^2 + \alpha^2\sigma^2, \tag{31}$$

where the two terms are defined as

$$T_1 \coloneqq \|V^{\mathrm{LSTD}} - V^\star\|_{\mathcal{Q}}^2 - 2\alpha\|\mathcal{B}(V^{\mathrm{LSTD}})\|_\mu^2 + \alpha^2\|\mathcal{B}(V^{\mathrm{LSTD}})\|_{\mathcal{Q}}^2, \quad \text{and} \tag{32a}$$

$$T_2 \coloneqq 2\alpha^2\langle V^{\mathrm{LSTD}} - V^\star,\ \mathcal{Q}b\rangle_{\mathcal{Q}} - 2\alpha\langle V^{\mathrm{LSTD}} - V^\star,\ b\rangle_{\mathcal{Q}}. \tag{32b}$$

In order to prove the claim (31), we begin by observing that for any stepsize $\alpha > 0$, expanding the square yields

$$\|\mathcal{F}_\alpha(V^{\mathrm{LSTD}}) - V^\star\|_{\mathcal{Q}}^2 = \|V^{\mathrm{LSTD}} - V^\star\|_{\mathcal{Q}}^2 - 2\alpha\langle V^{\mathrm{LSTD}} - V^\star,\ \mathcal{B}(V^{\mathrm{LSTD}}) + W\rangle_{\mathcal{Q}} + \alpha^2\|\mathcal{B}(V^{\mathrm{LSTD}}) + W\|_{\mathcal{Q}}^2,$$

Taking expectation over the regression noise, we find that

$$\mathbb{E}\langle V^{\mathrm{LSTD}} - V^\star,\ \mathcal{B}(V^{\mathrm{LSTD}}) + W\rangle_{\mathcal{Q}} = \langle V^{\mathrm{LSTD}} - V^\star,\ \mathcal{B}(V^{\mathrm{LSTD}}) + b\rangle_{\mathcal{Q}}, \quad \text{and}$$

$$\mathbb{E}\|\mathcal{B}(V^{\mathrm{LSTD}}) + W\|_{\mathcal{Q}}^2 = \|\mathcal{B}(V^{\mathrm{LSTD}}) + b\|_{\mathcal{Q}}^2 + \sigma^2,$$

where the second equation follows from a bias-variance decomposition. Thus we have

$$\mathbb{E}\|\mathcal{F}_\alpha(V^{\mathrm{LSTD}}) - V^\star\|_{\mathcal{Q}} = \|V^{\mathrm{LSTD}} - V^\star\|_{\mathcal{Q}}^2 - 2\alpha\,\langle V^{\mathrm{LSTD}} - V^\star,\ \mathcal{B}(V^{\mathrm{LSTD}}) + b\rangle_{\mathcal{Q}}$$
$$+ \alpha^2\|\mathcal{B}(V^{\mathrm{LSTD}}) + b\|_{\mathcal{Q}}^2 + \alpha^2\sigma^2.$$

We have

$$\langle V^{\text{LSTD}} - V^{\star}, \mathcal{B}(V^{\text{LSTD}}) + b\rangle_{\mathcal{Q}} = \langle V^{\text{LSTD}} - V^{\star}, \mathcal{B}(V^{\text{LSTD}})\rangle_{\mathcal{Q}} + \langle V^{\text{LSTD}} - V^{\star}, b\rangle_{\mathcal{Q}}$$
$$= \|\mathcal{B}(V^{\text{LSTD}})\|_{\mu}^2 + \langle V^{\text{LSTD}} - V^{\star}, b\rangle_{\mathcal{Q}}, \tag{33}$$

and

$$\|\mathcal{B}(V^{\text{LSTD}}) + b\|_{\mathcal{Q}}^2 = \|\mathcal{B}(V^{\text{LSTD}})\|_{\mathcal{Q}}^2 + 2\langle \mathcal{B}(V^{\text{LSTD}}), b\rangle_{\mathcal{Q}} + \|b\|_{\mathcal{Q}}^2$$
$$= \|\mathcal{B}(V^{\text{LSTD}})\|_{\mathcal{Q}}^2 + 2\langle V^{\text{LSTD}} - V^{\star}, \mathcal{Q}b\rangle_{\mathcal{Q}} + \|b\|_{\mathcal{Q}}^2, \tag{34}$$

using the self-adjointness of $\mathcal{Q}$. Combining equations (33) and (34) yields

$$\mathbb{E}\|\mathcal{F}_{\alpha}(V^{\text{LSTD}}) - V^{\star}\|_{\mathcal{Q}}^2 = \|V^{\text{LSTD}} - V^{\star}\|_{\mathcal{Q}}^2 - 2\alpha\|\mathcal{B}(V^{\text{LSTD}})\|_{\mu}^2 + \alpha^2\|\mathcal{B}(V^{\text{LSTD}})\|_{\mathcal{Q}}^2$$
$$+ 2\alpha^2\langle V^{\text{LSTD}} - V^{\star}, \mathcal{Q}b\rangle_{\mathcal{Q}} - 2\alpha\langle V^{\text{LSTD}} - V^{\star}, b\rangle_{\mathcal{Q}} + \alpha^2(\|b\|_{\mathcal{Q}}^2 + \sigma^2)$$
$$= T_1 + T_2 + \alpha^2(\|b\|_{\mathcal{Q}}^2 + \sigma^2),$$

which establishes the claimed decomposition (31).

Our next step is to bound the terms T_1 and T_2 that were previously defined (32).

Bounding T_1: It suffices to show that the Bellman residual $\mathcal{B}(V^{\text{LSTD}})$ satisfies the bounds

$$\|\mathcal{B}(V^{\text{LSTD}})\|_{\mathcal{Q}}^2 \leqslant 2M\|V^{\text{LSTD}} - V^{\star}\|_{\mathcal{Q}}^2, \quad \text{and} \quad \|\mathcal{B}(V^{\text{LSTD}})\|_{\mu}^2 \geqslant m\|V^{\text{LSTD}} - V^{\star}\|_{\mathcal{Q}}^2. \tag{35}$$

Indeed, by combining these two inequalities, we find that

$$T_1 \leqslant \left(1 - 2\alpha m + 2\alpha^2 M\right)\|V^{\text{LSTD}} - V^{\star}\|_{\mathcal{Q}}^2,$$

In order to prove the claims (35), we first observe that the lower bound follows immediately from the definition of m. As for the upper bound, we have

$$\|z\|_{\mathcal{Q}}^2 = \langle \mathcal{Q}^{-1/2}z, \mathcal{Q}^2\mathcal{Q}^{-1/2}z\rangle_{\mu} \overset{(i)}{\leqslant} 2\langle \mathcal{Q}^{-1/2}z, \mathcal{Q}\mathcal{Q}^{-1/2}z\rangle_{\mu}$$
$$= 2\|z\|_{\mu}^2$$
$$\leqslant 2M\|V^{\text{LSTD}} - V^{\star}\|_{\mathcal{Q}}^2,$$

where step (i) follows from the inequality $\mathcal{Q}^2 \preceq 2\mathcal{Q}$.

Bounding T_2: We apply the Fenchel-Young inequality with a parameter $\tau > 0$ to be chosen, thereby obtaining the bound

$$|T_2| = 2\alpha|\langle V^{\text{LSTD}} - V^{\star}, (\alpha\mathcal{Q} - \mathcal{I})b\rangle_{\mathcal{Q}}| \leqslant 2\alpha\left(\frac{\tau}{2}\|V^{\text{LSTD}} - V^{\star}\|_{\mathcal{Q}} + \frac{1}{2\tau}\|(\alpha\mathcal{Q} - \mathcal{I})b\|_{\mathcal{Q}}^2\right)$$
$$= \alpha\tau\|V^{\text{LSTD}} - V^{\star}\|_{\mathcal{Q}}^2 + \frac{\alpha}{\tau}\|(\alpha\mathcal{Q} - \mathcal{I})b\|_{\mathcal{Q}}^2.$$

In order to bound the second term on the right-hand side of this inequality, we again apply the Fenchel-Young inequality so as to obtain

$$\|(\alpha\mathcal{Q} - \mathcal{I})b\|_{\mathcal{Q}}^2 \leqslant 2\alpha^2\|\mathcal{Q}b\|_{\mathcal{Q}}^2 + 2\|b\|_{\mathcal{Q}}^2.$$

Since $\mathcal{Q}^3 \preceq 2\mathcal{Q}^2 \preceq 4\mathcal{Q}$, we have $\|\mathcal{Q}b\|_{\mathcal{Q}}^2 = \langle b, \mathcal{Q}^3 b\rangle_{\mu} \leqslant 4\langle b, \mathcal{Q}b\rangle_{\mu} = 4\|b\|_{\mathcal{Q}}^2$, whence

$$T_2 \leqslant \alpha\tau\|V^{\text{LSTD}} - V^{\star}\|_{\mathcal{Q}}^2 + \frac{2\alpha}{\tau}\left(4\alpha^2 + 1\right)\|b\|_{\mathcal{Q}}^2.$$

We now set the Fenchel-Young parameter as $\tau = \frac{m}{2}$. With this choice, and recalling that $\alpha \in (0, \frac{1}{2}]$, we find that

$$T_2 \leqslant \frac{\alpha m}{2}\|V^{\text{LSTD}} - V^{\star}\|_{\mathcal{Q}} + \frac{8\alpha}{m}\|b\|_{\mathcal{Q}}^2.$$

Putting together the pieces, we conclude that

$$\mathbb{E}\|\mathcal{F}_\alpha(V^{\text{LSTD}}) - V^\star\|_{\mathcal{Q}}^2 \leqslant \left(1 - \frac{3\alpha m}{2} + 2\alpha^2 M\right)\|V^{\text{LSTD}} - V^\star\|_{\mathcal{Q}}^2 + \left(\frac{8\alpha}{m} + \alpha^2\right)\|b\|_{\mathcal{Q}}^2 + \alpha^2 \cdot \sigma^2,$$

as claimed in equation (27a).

Finally, setting $\alpha = \frac{m}{2M}$ and using the inequality $m^2 \leqslant 2M$ yields the claim (27b).

B.3 Proof of Lemma 1

Recall that V^{LSTD} is the LSTD solution defined by the subspace $\mathcal{S}$. For any $V \in \mathcal{S}$, we have

$$\begin{aligned}
\|V^{\text{LSTD}} - V - V^\star\|_{\mathcal{Q}}^2 &= \|V^{\text{LSTD}} - V^\star\|_{\mathcal{Q}}^2 + 2\langle V, V^{\text{LSTD}} - V^\star\rangle_{\mathcal{Q}} + \|V\|_{\mathcal{Q}}^2 \\
&= \|V^{\text{LSTD}} - V^\star\|_{\mathcal{Q}}^2 + 2\langle V, \mathcal{Q}(V^{\text{LSTD}} - V^\star)\rangle_\mu + \|V\|_{\mathcal{Q}}^2 \\
&\stackrel{(i)}{=} \|V^{\text{LSTD}} - V^\star\|_{\mathcal{Q}}^2 + \|V\|_{\mathcal{Q}}^2,
\end{aligned}$$

where the equality (i) follows from the inclusion $\mathcal{Q}(V^{\text{LSTD}} - V^\star) \in \mathcal{S}^\perp$, as guaranteed by definition of the LSTD fixed point. Consequently, we see that this quadratic form is minimized by setting $V = 0$, yielding the claim.

B.4 Proof of Theorem 4

Key to this proof is the notion of the *empirical local sub-Gaussian complexity* of $\widetilde{\mathcal{F}}$, defined at scale $\delta > 0$ as

$$\widehat{\mathcal{G}}_n(\delta; \widetilde{\mathcal{F}}, \{x_i\}_{i=1}^n) := \mathbb{E}\left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_n \leqslant \delta} \left|\frac{1}{n}\sum_{i=1}^n w_i g(x_i)\right| \;\middle|\; \{x_i\}_{i=1}^n\right].$$

Here the samples $\{x_i\}_{i=1}^n$ correspond to those in regression data set $\mathcal{D}^{REG}$, upon which we are conditioning. For convenience, we drop the dependence on the observation data $\{x_i\}_{i=1}^n$ and directly write $\widehat{\mathcal{G}}_n(\delta; \widetilde{\mathcal{F}})$. We use the shorthand $\mathbb{E}_w := \mathbb{E}[\cdot \mid \{x_i\}_{i=1}^n]$ in the paper occasionally for convenience.

Relative to the population-level sub-Gaussian complexity defined previously (20) (in which we take expectations over state variables), the complexity function $\widehat{\mathcal{G}}$ treats the data $\{x_i\}_{i=1}^n$ as fixed where $\mathcal{G}$ does not. In analogy to the critical radii (22) defined at the population level, we also define a critical radius at using $\widehat{\mathcal{G}}$: more specifically, let $\widehat{\varepsilon}_n > 0$ be the smallest positive solution to the inequality $\frac{\widehat{\mathcal{G}}(\delta;\widetilde{\mathcal{F}})}{\delta} \leqslant \frac{\delta}{2}$. Finally, recall that the dataset $\{x_i\}_{i=1}^n$ defines the empirical norm $\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f^2(x_i)$.

With this set-up, our proof is broken into three main steps:

(i) We control the empirical error $\|\widehat{f} - \widetilde{f}\|_n^2$ in terms of empirical sub-Gaussian critical radius $\widehat{\varepsilon}_n$ and the approximation error $\|\widetilde{f} - f^*\|_n^2$, as stated in Lemma 2.

(ii) We establish that $\widehat{\varepsilon}_n$ is controlled by the population critical radii ε_n and ω_n, given by Lemma 3.

(iii) We use the critical radius ω_n to establish uniform control and connect $\|\cdot\|_n$ with $\|\cdot\|_\mu$ (cf. Lemma 4).

We begin by stating several technical lemmas. Recall that $\omega_n > 0$ satisfies the critical inequality $\mathcal{R}(\omega_n; \widetilde{\mathcal{F}}) \leqslant \frac{\omega_n^2}{b}$. The statement of this lemma treats the observations $\{x_i\}_{i=1}^n$ as fixed, and provides control of $\|\widehat{f} - \widetilde{f}\|_n$ over the noise in w_i conditional on $\{x_i\}_{i=1}^n$.

Lemma 2. *For any $t \geqslant \widehat{\varepsilon}_n$, we have*

$$\|\widehat{f} - \widetilde{f}\|_n^2 \leqslant c_1 t\widehat{\varepsilon}_n + c_2\|\widetilde{f} - f^*\|_n^2$$

with probability exceeding $1 - e^{-c_3 n \cdot \frac{t\widehat{\varepsilon}_n}{\|V\|_\infty^2}}$.

See Appendix B.5.1 for the proof.

In Lemma 2, we have a guarantee in terms of the empirical critical radius $\widehat{\varepsilon}_n$, which is a random variable dependent on $\{x_i\}_{i=1}^n$ and $\widetilde{\mathcal{F}}$. The following lemma controls the expectation of $\widehat{\varepsilon}_n^2$.

Lemma 3. *The empirical sub-Gaussian complexity $\widehat{\varepsilon}_n$ has expectation bounded as*

$$\mathbb{E}\left[\widehat{\varepsilon}_n^2\right] \leqslant c_1\varepsilon_n^2 + c_2\,\omega_n^2 + \frac{c_3}{n}\left(b^2 + \|V\|_\infty^2\right)$$

for universal constants c_1, c_2, c_3.

See Appendix B.5.2 for a proof of this lemma.

The next lemma allows us to convert our empirical norm guarantee to one involving the stationary norm $\|\cdot\|_\mu$. It provides control on $\|f - \tilde{f}\|_n$ in terms of $\|f - \tilde{f}\|_\mu$ uniformly for all $f \in \widetilde{\mathcal{F}}$; we require a uniform version since $\widehat{f}$ is a random quantity.

Lemma 4. *For any $t \geqslant \omega_n$, we have*

$$\left|\|f - \tilde{f}\|_n^2 - \|f - \tilde{f}\|_\mu^2\right| \leqslant \frac{1}{2}\|f - \tilde{f}\|_\mu^2 + \frac{t^2}{2} \qquad \text{for all } f \in \widetilde{\mathcal{F}} \tag{36}$$

with probability exceeding $1 - c_1 e^{-c_2 \frac{nt^2}{b^2}}$.

See Theorem 14.1 in the reference (Wainwright, 2019) for the proof.

Note that for an arbitrary real-valued random variable Z, the inequality $Z \leqslant \max\{Z, 0\}$ allows to write $\mathbb{E}[Z] \leqslant \mathbb{E}[\max\{Z, 0\}] = \int_0^\infty \mathbb{P}(Z \geqslant u)\,du$, so that we can upper bound $\mathbb{E}[Z]$ by integrating an upper tail bound.

Let us apply this line of reasoning to the random variable

$$Z := \left|\|\widehat{f} - \tilde{f}\|_n^2 - \|\widehat{f} - \tilde{f}\|_\mu^2\right| - \frac{1}{2}\|\widehat{f} - \tilde{f}\|_\mu^2,$$

using the tail bound established in Lemma 4. With a change of variables, we have

$$\begin{aligned}
\mathbb{E}[Z] &\leqslant \int_0^\infty \mathbb{P}(Z \geqslant u)\,du \leqslant \frac{1}{2}\omega_n^2 + \int_{\omega_n^2/2}^\infty \mathbb{P}(Z \geqslant u)\,du \\
&\leqslant \frac{1}{2}\omega_n^2 + \int_0^\infty c_1 e^{-2c_2 \frac{nu}{b^2}}\,du \\
&= \frac{1}{2}\omega_n^2 + \frac{c_1}{2c_2} \cdot \frac{b^2}{n}.
\end{aligned} \tag{37}$$

Additionally, observe that Lemma 2 treats $\{x_i\}_{i=1}^n$ as fixed and the high probability is over the randomness with respect to $\{w_i\}$ conditional on $\{x_i\}$. Thus, integrating the tail bound again yields

$$\mathbb{E}_w\left[\|\widehat{f} - \tilde{f}\|_n^2\right] \leqslant c'\left\{\widehat{\varepsilon}_n^2 + \frac{\|V\|_\infty^2}{n} + \|\tilde{f} - f^*\|_n^2\right\}$$

for some universal constant c'. Taking expectations over $\{x_i\}_{i=1}^n$ then yields

$$\mathbb{E}\|\widehat{f} - \tilde{f}\|_\mu^2 \leqslant c'\left\{\mathbb{E}\left[\widehat{\varepsilon}_n^2\right] + \frac{\|V\|_\infty^2}{n} + \|\tilde{f} - f^*\|_\mu^2\right\}. \tag{38}$$

Finally, combining equations (37) and (38), we conclude that

$$\mathbb{E}\|\widehat{f} - \tilde{f}\|_\mu^2 \leqslant c''\left\{\mathbb{E}\left[\widehat{\varepsilon}_n^2\right] + \omega_n^2 + \frac{1}{n}\left(b^2 + \|V\|_\infty^2\right) + \|\tilde{f} - f^*\|_\mu^2\right\}.$$

The theorem then follows from applying Lemma 3 and rearranging.

B.5 Proof of auxiliary lemmas for Theorem 4

We collect here the proofs of the auxiliary lemmas.

B.5.1 Proof of Lemma 2

We begin with some straightforward algebraic manipulations. By expanding the square, we have

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - f(x_i))^2 + \frac{2}{n}\sum_{i=1}^{n}\left(f^*(x_i) - \tilde{f}(x_i)\right)(\tilde{y}_i - f(x_i)) + \frac{1}{n}\sum_{i=1}^{n}\left(f^*(x_i) - \tilde{f}(x_i)\right)^2,$$

where we define $\tilde{y}_i := \tilde{f}(x_i) + w_i$. Since the rightmost term does not involve f, we can write

$$\hat{f} \in \arg\min_{f \in \mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - f(x_i))^2 + \frac{2}{n}\sum_{i=1}^{n}\left(f^*(x_i) - \tilde{f}(x_i)\right)(\tilde{y} - f(x_i))\right\}.$$

By optimality of $\hat{f}$ and feasibility of $\tilde{f}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{y}_i - \hat{f}(x_i)\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left(f^*(x_i) - \tilde{f}(x_i)\right)(\tilde{y}_i - \hat{f}(x_i))$$

$$\leqslant \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{y}_i - \tilde{f}(x_i)\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left(f^*(x_i) - \tilde{f}(x_i)\right)(\tilde{y}_i - \tilde{f}(x_i)).$$

Rearranging terms yields the *basic inequality*

$$\frac{1}{2n}\sum_{i=1}^{n}\left(\hat{f}(x_i) - \tilde{f}(x_i)\right)^2 \leqslant T_1 + T_2, \tag{39}$$

where we define

$$T_1 := \frac{1}{n}\sum_{i=1}^{n}w_i\left(\hat{f}(x_i) - \tilde{f}(x_i)\right), \quad \text{and} \quad T_2 := \frac{1}{n}\sum_{i=1}^{n}\left(f^*(x_i) - \tilde{f}(x_i)\right)\left(\hat{f}(x_i) - \tilde{f}(x_i)\right).$$

We bound each of T_1 and T_2 in turn.

Bounding T_2:   By applying the Fenchel-Young inequality, we find that

$$T_2 \leqslant \frac{1}{n}\sum_{i=1}^{n}\left(f^*(x_i) - \tilde{f}(x_i)\right)^2 + \frac{1}{4n}\sum_{i=1}^{n}\left(\hat{f}(x_i) - \tilde{f}(x_i)\right)^2$$

$$= \|f^* - \tilde{f}\|_n^2 + \frac{1}{4}\|\hat{f} - \tilde{f}\|_n^2.$$

Bounding T_1:   We make use of techniques for controlling localized empirical processes; see Chapter 13 in the book (Wainwright, 2019) for relevant background. Recall that by definition, the scalar $\hat{\varepsilon}_n > 0$ satisfies the critical inequality $\frac{\hat{\mathcal{G}}(\delta;\widetilde{\mathcal{F}})}{\delta} \leqslant \frac{\delta}{2}$. Recall that $\widetilde{\mathcal{F}} = \{f - \tilde{f} : f \in \mathcal{F}\}$ is the shifted function class around $\tilde{f}$. For a scalar $u \geqslant \hat{\varepsilon}_n$, define the event

$$\mathcal{E}(u) = \left\{\exists\, g \in \widetilde{\mathcal{F}} \text{ with } \|g\|_n \geqslant u \text{ such that } \left|\frac{1}{n}\sum_{i=1}^{n}w_i\, g(x_i)\right| \geqslant 2\|g\|_n u\right\}. \tag{40}$$

The following result controls the probability of this event:

Lemma 5. *For all $u \geqslant \hat{\varepsilon}_n$, we have*

$$\mathbb{P}(\mathcal{E}(u)) \leqslant e^{-cn \cdot \frac{u^2}{\|V\|_\infty^2}}$$

for some universal constant $c > 0$.

See Appendix C.2 for the proof of this lemma.

By Lemma 5, since $\widehat{f} - \widetilde{f} \in \widetilde{\mathcal{F}}$, we have if $\|\widehat{f} - \widetilde{f}\|_n > \sqrt{t\widehat{\varepsilon}_n}$ for some $t \geqslant \widehat{\varepsilon}_n$, then

$$|T_1| = \left| \frac{1}{n} \sum_{i=1}^n w_i(\widehat{f}(x_i) - \widetilde{f}(x_i)) \right| \leqslant 2\|\widehat{f} - \widetilde{f}\|_n \sqrt{t\widehat{\varepsilon}_n}$$

with probability exceeding $1 - e^{-cn \cdot \frac{t\widehat{\varepsilon}_n}{\|V\|_\infty^2}}$.

Putting together the pieces, we conclude that

$$\frac{1}{2}\|\widehat{f} - \widetilde{f}\|_n^2 \leqslant \frac{1}{2}t\widehat{\varepsilon}_n + 2\|\widehat{f} - \widetilde{f}\|_n \sqrt{t\widehat{\varepsilon}_n} + \|f^* - \widetilde{f}\|_n^2 + \frac{1}{4}\|\widehat{f} - \widetilde{f}\|_n^2,$$

which implies that

$$\|\widehat{f} - \widetilde{f}\|_n^2 \leqslant 2t\widehat{\varepsilon}_n + 8\sqrt{t\widehat{\varepsilon}_n}\|\widehat{f} - \widetilde{f}\|_n + 4\|f^* - \widetilde{f}\|_n^2.$$

This inequality implies that there are universal constants c_1, c_2, c_3 such that

$$\|\widehat{f} - \widetilde{f}\|_n^2 \leqslant c_1 t\widehat{\varepsilon}_n + c_2\|\widetilde{f} - f^*\|_n^2 \tag{41}$$

with probability exceeding $1 - e^{-c_3 n \cdot \frac{t\widehat{\varepsilon}_n}{\|V\|_\infty^2}}$.

B.5.2 Proof of Lemma 3

It suffices to show that for any $u \geqslant \max\{\varepsilon_n, \frac{\omega_n^2}{\varepsilon_n}\}$, we have the bound

$$\widehat{\varepsilon}_n^2 \leqslant 4u\varepsilon_n \text{ with probability at least } 1 - c_1 e^{-c_2 n \frac{u\varepsilon_n}{(\|V\|_\infty + b)^2}}. \tag{42}$$

Indeed, the statement of the lemma can be obtained by integrating this tail bound. Accordingly, the remainder of our proof is devoted to establishing inequality (42) for some fixed but arbitrary u.

Recall that by the definition (22), the critical radii satisfy the inequalities

$$\frac{\widehat{\mathcal{G}}_n(\widehat{\varepsilon}_n, \widetilde{\mathcal{F}})}{\widehat{\varepsilon}_n} \leqslant \frac{\widehat{\varepsilon}_n}{2}, \qquad \frac{\mathcal{G}_n(\varepsilon_n, \widetilde{\mathcal{F}})}{\varepsilon_n} \leqslant \frac{\varepsilon_n}{2}, \quad \text{and} \quad \frac{\mathcal{R}_n(\omega_n; \widetilde{\mathcal{F}})}{\omega_n} \leqslant \frac{\omega_n}{b}.$$

For each $t \geqslant$, define the random variable

$$\overline{Z}_n(t) := \mathbb{E}_w \left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_\mu \leqslant t} \left| \frac{1}{n} \sum_{i=1}^n w_i g(x_i) \right| \right], \tag{43}$$

and observe that $\mathbb{E}[\overline{Z}_n(t)] = \mathcal{R}_n(t; \widetilde{\mathcal{F}})$ by construction. We also define the events

$$\mathcal{E}_0(t) := \left\{ \overline{Z}_n(t) \leqslant \mathcal{R}_n(t; \widetilde{\mathcal{F}}) + \frac{t\sqrt{u\varepsilon_n}}{8} \right\}, \quad \text{and} \quad \mathcal{E}_1(t) := \left\{ \sup_{f \in \widetilde{\mathcal{F}}} \frac{|\|f\|_n^2 - \|f\|_\mu^2|}{\|f\|_\mu^2 + t^2} \leqslant \frac{1}{2} \right\}. \tag{44}$$

Conditional on $\mathcal{E}_1(t)$, we have for all $f \in \widetilde{\mathcal{F}}$

$$\|f\|_n \leqslant \sqrt{\tfrac{3}{2}\|f\|_\mu^2 + \tfrac{1}{2}t^2} \leqslant 2\|f\|_\mu + t, \qquad \text{and}$$

$$\|f\|_\mu \leqslant \sqrt{2\|f\|_n^2 + t^2} \leqslant 2\|f\|_n + t.$$

Therefore conditioned on $\mathcal{E}_1(\sqrt{u\varepsilon_n})$,

$$\widehat{\mathcal{G}}_n(s) = \mathbb{E}_w \left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_n \leqslant s} \left| \frac{1}{n} \sum_{i=1}^n w_i g(x_i) \right| \right] \leqslant \mathbb{E}_w \left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_\mu \leqslant 2s + \sqrt{u\varepsilon_n}} \left| \frac{1}{n} \sum_{i=1}^n w_i g(x_i) \right| \right] = \overline{Z}_n(2s + \sqrt{u\varepsilon_n}). \tag{45}$$

Conditioning on both $\mathcal{E}_0\left(9\sqrt{u\varepsilon_n}\right)$ and $\mathcal{E}_1(\sqrt{u\varepsilon_n})$, we have

$$\widehat{\mathcal{G}}(4\sqrt{u\varepsilon_n}) \overset{(i)}{\leqslant} \overline{Z}_n(9\sqrt{u\varepsilon_n}) \overset{(ii)}{\leqslant} \mathcal{R}_n(9\sqrt{u\varepsilon_n}) + \frac{9}{8}u\varepsilon_n,$$

where step (i) follows from the bound (45), and step (ii) follows from the event $\mathcal{E}_0(9\sqrt{u\varepsilon_n})$. Since the function $\delta \mapsto \frac{\mathcal{R}_n(\delta)}{\delta}$ is non-increasing, and $9\sqrt{u\varepsilon_n} \geqslant \varepsilon_n$,

$$\frac{\mathcal{R}_n(9\sqrt{u\varepsilon_n})}{9\sqrt{u\varepsilon_n}} \leqslant \frac{\mathcal{R}_n(\varepsilon_n)}{\varepsilon_n} \leqslant \frac{\varepsilon_n}{2} \leqslant \frac{\sqrt{u\varepsilon_n}}{2} \implies \mathcal{R}_n(9\sqrt{u\varepsilon_n}) \leqslant \frac{9}{2}u\varepsilon_n,$$

using the definition of ε_n and the critical inequality. We then have

$$\widehat{\mathcal{G}}(4\sqrt{u\varepsilon_n}) \leqslant \frac{9}{2}u\varepsilon_n + \frac{9}{8}u\varepsilon_n = \frac{45}{8}u\varepsilon_n,$$

which implies

$$\frac{\widehat{\mathcal{G}}(4\sqrt{u\varepsilon_n})}{4\sqrt{u\varepsilon_n}} \leqslant \frac{45}{32}\sqrt{u\varepsilon_n} \leqslant \frac{4\sqrt{u\varepsilon_n}}{2}.$$

Thus $4\sqrt{u\varepsilon_n}$ satisfies the critical inequality defined by $\widehat{\mathcal{G}}$, implying that $\widehat{\varepsilon}_n \leqslant 4\sqrt{u\varepsilon_n}$, as desired.

The following lemma provides control on the events $\mathcal{E}_0(9\sqrt{u\varepsilon_n})^c$ and $\mathcal{E}_1(\sqrt{u\varepsilon_n})^c$.

Lemma 6. *For any* $u \geqslant \max\{\varepsilon_n, \frac{\omega_n^2}{\varepsilon_n}\}$, *we have*

$$\mathbb{P}(\mathcal{E}_0(9\sqrt{u\varepsilon_n})^c) \leqslant 2\exp\left(-c_1 n \cdot \frac{u\varepsilon_n}{\|V\|_\infty^2 + b\|V\|_\infty}\right), \quad and \tag{46}$$

$$\mathbb{P}(\mathcal{E}_1(\sqrt{u\varepsilon_n})^c) \leqslant c_2\exp\left(-c_3 n \cdot \frac{u\varepsilon_n}{b^2}\right). \tag{47}$$

See Appendix C.3 for the proof of this lemma.

C Further Proofs

Here we present the proof of various lemmas used in the main text of the paper.

C.1 Proof of equation (12)

Let us state the result to be proven more formally:

Lemma 7. *For any separable Hilbert space* $L^2(\mathcal{X}, \mu)$, *the operator* $(\mathcal{I} - \gamma\mathcal{P})$ *is self-adjoint, invertible, and satisfies the bounds*

$$(1-\gamma)\|f\|_\mu^2 \leqslant \langle f, (\mathcal{I} - \gamma\mathcal{P})f\rangle_\mu \leqslant (1+\gamma)\|f\|_\mu^2 \quad for\ all\ f \in L^2(\mathcal{X}, \mu). \tag{48}$$

The requirements for $L^2(\mathcal{X}, \mu)$ to be separable is quite mild; in the case where $\mathcal{X} = \mathbb{R}^d$ for some dimension d, one sufficient condition is that μ is a continuous distribution.

Proof. By definition of the stationary measure, the operator $\mathcal{P}$ satisfies the bound $\|\mathcal{P}V\|_\mu \leqslant \|V\|_\mu$. Thus, by classical functional analysis (e.g., see Chap. 8 in the reference (Kreyszig, 1991)), the inverse $(\mathcal{I} - \gamma\mathcal{P})^{-1}$ exists, and can be represented via the Neumann series expansion

$$(\mathcal{I} - \gamma\mathcal{P})^{-1} = \sum_{j=0}^\infty \gamma^j \mathcal{P}^j.$$

The other statements follow from using linearity of the inner product, and then applying Cauchy-Schwarz in conjunction with the operator norm bound of $\mathcal{P}$.

Since $\mathcal{Q} := \mathcal{I} - \gamma\mathcal{P}$ is a positive operator, it has a unique square root $(\mathcal{I} - \gamma\mathcal{P})^{1/2}$. In addition, since $\mathcal{Q}$ is invertible, so is its square root (which is also self-adjoint). $\square$

C.2   Proof of Lemma 5

This argument is inspired by the proof of Lemma 13.12 in the book (Wainwright, 2019), along with some new ingredients. We treat the observations $\{x_i\}_{i=1}^n$ as fixed here. Define the empirical process

$$Z_n(u) := \sup_{\substack{g \in \widetilde{\mathcal{F}}, \\ \|g\|_n \leqslant u}} \left| \frac{1}{n} \sum_{i=1}^n w_i\, g(x_i) \right|.$$

We claim that

$$\mathbb{P}\left(\mathcal{E}(u)\right) \leqslant \mathbb{P}(Z_n(u) \geqslant 2u^2).$$

In order to establish this claim, suppose that there exists $g \in \widetilde{\mathcal{F}}$ with $\|g\|_n \geqslant u$ such that

$$\left| \frac{1}{n} \sum_{i=1}^n w_i\, g(x_i) \right| \geqslant 2\|g\|_n u.$$

Define the rescaled function $\tilde{g} = \frac{u}{\|g\|_n} g$; observe $\tilde{g} \in \widetilde{\mathcal{F}}$ by convexity. We then have

$$\left| \frac{1}{n} \sum_{i=1}^n w_i\, \tilde{g}(x_i) \right| = \frac{u}{\|g\|_n} \left| \frac{1}{n} \sum_{i=1}^n w_i\, g(x_i) \right| \geqslant 2u^2,$$

as desired.

It remains to control the probability $\mathbb{P}(Z_n(u) \geqslant 2u^2)$. If we view $Z_n(u)$ as a function of $(w_1, \ldots, w_n)$, it can then be seen that it has Lipschitz constant of $\frac{u}{\sqrt{n}}$ in the Euclidean norm and is separately convex. Since $|w_i| \leqslant 2\|V\|_\infty$, we have by Theorem 3.4 in the book (Wainwright, 2019) we have

$$\mathbb{P}(Z_n(u) \geqslant \mathbb{E}[Z_n(u)] + s) \leqslant e^{-cn \cdot \frac{s^2}{u^2 \|V\|_\infty^2}}.$$

Observe by definition that $\mathbb{E}[Z_n(u)] = \widehat{\mathcal{G}}(u)$, so that

$$\frac{\widehat{\mathcal{G}}(u)}{u} \overset{(i)}{\leqslant} \frac{\widehat{\mathcal{G}}(\widehat{\varepsilon}_n)}{\widehat{\varepsilon}_n} \overset{(ii)}{\leqslant} \frac{\widehat{\varepsilon}_n}{2} \leqslant \widehat{\varepsilon}_n,$$

where step (i) uses Lemma 13.6 in the book (Wainwright, 2019) and the fact that $u \geqslant \widehat{\varepsilon}_n$, and step (ii) uses the critical inequality (22). Thus, we conclude that

$$\mathbb{P}(Z_n(u) \geqslant 2u^2) \leqslant \mathbb{P}(Z_n(u) \geqslant u\widehat{\varepsilon}_n + u^2) \leqslant e^{-cn \cdot \frac{u^2}{\|V\|_\infty^2}},$$

as desired.

C.3   Proof of Lemma 6

To begin, we observe that the claim (47) follows by applying Lemma 4, along with the fact that $u \geqslant \frac{\omega_n^2}{\varepsilon_n}$ implies that $\sqrt{u\varepsilon_n} \geqslant \omega_n$. In order to prove the claim (46), assume that $u \geqslant \varepsilon_n$, and define the random variable

$$Z_n(t) := \sup_{g \in \widetilde{\mathcal{F}}, \|g\|_\mu \leqslant t} \left| \frac{1}{n} \sum_{i=1}^n w_i\, g(x_i) \right|.$$

By the definition (43) of $\bar{Z}_n(t)$, we have the equivalence

$$\bar{Z}_n(t) = \mathbb{E}_w[Z_n(t)] = \mathbb{E}[Z_n(t) \mid \{x_i\}_{i=1}^n],$$

and $\mathbb{E}[\bar{Z}_n(t)] = \mathbb{E}[Z_n(t)] = \mathcal{G}_n(t)$. The moment generating function of $\bar{Z}_n(t) - \mathcal{G}_n(t)$ can be bounded as

$$\mathbb{E}\left[e^{\lambda(\bar{Z}_n(t) - \mathcal{G}_n(t))} \right] = \mathbb{E}\left[e^{\lambda(\mathbb{E}_w[Z_n(t)] - \mathcal{G}_n(t))} \right] \leqslant \mathbb{E}\left\{ \mathbb{E}_w\left[e^{\lambda(Z_n(t) - \mathcal{G}_n(t))} \right] \right\} = \mathbb{E}\left[e^{\lambda(Z_n(t) - \mathcal{G}_n(t))} \right],$$

where the inequality follows from a (conditional) Jensen's inequality. Consequently, we can control $\bar{Z}_n(t) - \mathcal{G}_n(t)$ by controlling $Z_n(t) - \mathcal{G}_n(t)$ using a Chernoff-type argument. We prove an upper tail bound on $Z_n(t)$ via Talagrand concentration (Talagrand, 1996) which uses a Chernoff bound in its proof; see Theorem 3.27 in the book (Wainwright, 2019) for further details.

Define the random variable $\Sigma^2 = \sup_{f \in \widetilde{\mathcal{F}}, \|f\|_\mu \leqslant t} \frac{1}{n} \sum_{i=1}^n w_i^2 f(x_i)^2$. We then have by Talagrand concentration

$$\mathbb{P}(Z_n(t) \geqslant \mathcal{G}_n(t) + s) \leqslant 2 \exp\left(-cn \cdot \frac{s^2}{\mathbb{E}[\Sigma^2] + bs} \right)$$

where c is a universal constant. By Equation (3.84) from the book (Wainwright, 2019), we have

$$\mathbb{E}[\Sigma^2] \leqslant \sup_{f \in \widetilde{\mathcal{F}}, \|f\|_\mu \leqslant t} \mathbb{E}[w^2 f^2(x)] + 4b\|V\|_\infty \cdot \mathcal{G}_n(t)$$
$$\leqslant 4\|V\|_\infty^2 t^2 + 4b\|V\|_\infty \cdot \mathcal{G}_n(t),$$

where we have used the fact $|w| \leqslant 2\|V\|_\infty$ a.s. Thus, we have

$$\mathbb{P}\left(\bar{Z}_n(9\sqrt{u\varepsilon_n}) \geqslant \mathcal{G}_n(9\sqrt{u\varepsilon_n}) + \frac{9u\varepsilon_n}{8} \right) \leqslant 2 \exp\left(-cn \cdot \frac{u^2 \varepsilon_n^2}{\|V\|_\infty^2 u\varepsilon_n + b\|V\|_\infty \cdot \mathcal{G}_n(9\sqrt{u\varepsilon_n}) + b\|V\|_\infty u\varepsilon_n} \right)$$
$$\overset{(i)}{\leqslant} 2 \exp\left(-c'n \cdot \frac{u\varepsilon_n}{\|V\|_\infty^2 + b\|V\|_\infty} \right),$$

where step (i) follows from the fact that $\mathcal{G}(9\sqrt{u\varepsilon_n}) \leqslant \frac{9}{2} u\varepsilon_n$ for $u \geqslant \varepsilon_n$.

D   Critical radii calculations

In this section, we describe the form of the critical radii for several different function classes.

D.1   Examples of statistical error

For simplicity, we consider classes with $\|V\|_\infty = b = 1$. Here we summarize the form of the critical radii; see Appendix D.2 for the computations that underlie these conclusions.

Polynomials of degree d: As our first example, suppose that $\widetilde{\mathcal{F}}$ is the family of d-degree polynomials defined on $\mathcal{X} = [-1, 1]$ uniformly bounded by 1. For a given $\theta \in \mathbb{R}^{d+1}$, define $f_\theta(x) := \theta_0 + \theta_1 x + \ldots + \theta_{d+1} x^d$. Then we have

$$\widetilde{\mathcal{F}} := \{ f_\theta : [-1, 1] \to \mathbb{R} \text{ for some } \theta \in \mathbb{R}^{d+1} \text{ such that } \|f_\theta\|_\infty \leqslant 1 \}.$$

Some straightforward calculations yield $\omega_n^2 \lesssim \frac{d+1}{n}$, and $\varepsilon_n^2 \lesssim \frac{d+1}{n}$, so that we can conclude

$$\beta_n^2 \lesssim \frac{d+1}{n}.$$

Radial basis kernels: Consider $\mathcal{X} = [-1, 1]$ and the reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ given by the Gaussian kernel $\mathcal{K}(x, z) = e^{-\frac{1}{2}(x-z)^2}$ defined on the Cartesian product space $[-1, 1] \times [-1, 1]$. Consider the unit ball in this RKHS, given by

$$\widetilde{\mathcal{F}} = \{ f \in \mathbb{H} \mid \|f\|_{\mathbb{H}} \leqslant 1, \|f\|_\infty \leqslant 1 \}.$$

For this example, it can be shown that $\omega_n^2 \lesssim \frac{\log(n+1)}{n}$ and $\varepsilon_n^2 \lesssim \frac{\log(n+1)}{n}$, and hence

$$\beta_n^2 \lesssim \frac{\log(n+1)}{n}.$$

This is a simple example of a non-parametric class, since the effective degrees of freedom ($\log(n+1)$ in this case) grows as a function of the sample size.

Convex Lipschitz functions: We let $\mathcal{X} = [0, 1]$ and $\widetilde{\mathcal{F}}$ be the class of convex 1-Lipschitz functions, i.e.

$$\widetilde{\mathcal{F}} := \{f : [0, 1] \to \mathbb{R} \mid f(0) = 0, \text{ and } f \text{ is convex and 1-Lipschitz}\}.$$

We can show $\omega_n^2 \lesssim \frac{1}{n^{4/5}}$ and $\varepsilon_n^2 \lesssim \frac{1}{n^{4/5}}$, giving

$$\beta_n^2 \lesssim \frac{1}{n^{4/5}}.$$

D.2   Calculations for Appendix D.1

In this section we provide the details behind the calculations in Appendix D.1. Before proceeding to the computations, we have the following result relating the sub-Gaussian complexity $\mathcal{G}(\delta; \widetilde{\mathcal{F}})$ and the Rademacher complexity $\mathcal{R}(\delta; \widetilde{\mathcal{F}})$

Lemma 8. *We have for some universal constant c,*

$$\mathcal{G}(\delta; \widetilde{\mathcal{F}}) \leqslant c\|V\|_\infty \cdot \mathcal{R}(\delta; \widetilde{\mathcal{F}}).$$

This lemma allows us to instantly obtain a bound for the sub-Gaussian complexity after computing the Rademacher complexity of the function class $\widetilde{\mathcal{F}}$. In Appendix D.1, we assume $\|V\|_\infty = 1$; consequently $\omega_n^2 \lesssim \varepsilon_n^2$.

Proof. We have by the tower property,

$$\mathcal{G}(\delta; \widetilde{\mathcal{F}}) = \mathbb{E}\left[\mathbb{E}_w\left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_\mu \leqslant \delta} \left|\frac{1}{n}\sum_{i=1}^n w_i g(x_i)\right|\right]\right]$$

We can then bound, letting ε_i be i.i.d. Rademacher random variables independent of $\{x_i\}_{i=1}^n$,

$$\mathbb{E}_w\left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_\mu \leqslant \delta} \left|\frac{1}{n}\sum_{i=1}^n w_i g(x_i)\right|\right] \overset{(a)}{\leqslant} 2\mathbb{E}_{w,\varepsilon}\left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_\mu \leqslant \delta} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i w_i g(x_i)\right|\right]$$

$$= 4\|V\|_\infty \cdot \mathbb{E}_{w,\varepsilon}\left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_\mu \leqslant \delta} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \cdot \frac{w_i g(x_i)}{2\|V\|_\infty}\right|\right]$$

$$\overset{(b)}{\leqslant} 8\|V\|_\infty \cdot \mathbb{E}_\varepsilon\left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_\mu \leqslant \delta} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)\right|\right],$$

where step (a) follows from symmetrization (cf. Chapter 4 in the reference (Wainwright, 2019)) and step (b) follows from the bound $|w_i| \leqslant 2\|V\|_\infty$ and the Ledoux-Talagrand contraction inequality (Ledoux and Talagrand, 1991). $\quad\square$

D.3   Polynomials of degree d

To simplify the calculations, let $\phi_0, \ldots, \phi_d$ be an orthonormal basis of $\widetilde{\mathcal{F}}$ with respect to $\langle \cdot, \cdot \rangle_\mu$. This basis exists via applying Gram-Schmidt on the basis $\{1, x, \ldots, x^d\}$. We can then write any polynomial function in $\widetilde{\mathcal{F}}$ as $f_b(x) = b_0\phi_0(x) + b_1\phi_1(x) + \ldots + b_d\phi_d(x)$, which satisfies $\|f_b\|_\mu = \|b\|_2$. Define the matrix $\mathbf{M} \in \mathbb{R}^{n \times (d+1)}$ with $\mathbf{M}_{ij} = \phi_j(x_i)$. We then have, letting (with mild abuse of notation) $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$,

$$\mathcal{G}_n(\delta; \widetilde{\mathcal{F}}) = \mathbb{E}\left[\sup_{f_b \in \widetilde{\mathcal{F}}, \|f_b\|_\mu \leqslant \delta} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_b(x_i)\right|\right]$$

$$\leqslant \mathbb{E}\left[\sup_{\|b\|_2 \leqslant \delta} \left|\frac{1}{n}\varepsilon^T \mathbf{M} b\right|\right] \overset{(a)}{=} \frac{\delta}{n}\mathbb{E}\left[\|\varepsilon^T\mathbf{M}\|_2\right] \overset{(b)}{\leqslant} \sqrt{\mathbb{E}[\|\varepsilon^T\mathbf{M}\|_2^2]},$$

where step (a) follows from the fact that $\|\cdot\|_2$ is a dual norm to itself, and step (b) follows from Jensen's inequality.

Manipulating further, we have

$$\mathbb{E}\left[\|\varepsilon^T\mathbf{M}\|_2^2\right] = \mathbb{E}\left[\mathbb{E}_\varepsilon[\text{trace}(\mathbf{M}^T \varepsilon\varepsilon^T\mathbf{M})]\right] = \mathbb{E}\left[\text{trace}(\mathbb{E}_\varepsilon[\varepsilon\varepsilon^T]\mathbf{M}\mathbf{M}^T)\right],$$

where we have used the fact that $\mathbb{E}_\varepsilon[\cdot] := \mathbb{E}[\cdot \mid \{x_i\}_{i=1}^n]$ and $\mathbf{M}$ is a function of $\{x_i\}_{i=1}^n$. We then have

$$\mathbb{E}\|\varepsilon^T \mathbf{M}\|_2^2 \overset{(a)}{=} \mathbb{E}[\text{trace}(\mathbf{M}\mathbf{M}^T)] \overset{(b)}{=} (d+1)n,$$

where the equality (a) follows from the fact that $\mathbb{E}[\varepsilon\varepsilon^T] = \mathbf{I}$, and the equality (b) follows from the fact that $\{\phi_0, \dots, \phi_d\}$ is orthonormal. To identify the critical radius, we compute using the definition (22),

$$\frac{4\delta\sqrt{d+1}}{\delta\sqrt{n}} \leqslant \delta \implies \omega_n^2 \lesssim \frac{d+1}{n}.$$

By Lemma 8, we have $\varepsilon_n^2 \lesssim \frac{d+1}{n}$.

D.4 Radial basis kernels

By Mercer's theorem (see Chapter 12 in the book (Wainwright, 2019)), we have the kernel integral operator $T_{\mathcal{K}}(f)(x) := \int_{\mathcal{X}} \mathcal{K}(x, z) f(z)\, d\mu(z)$ has a set of eigenvalues $\{\mu_j\}_{j=1}^\infty$. Then we have (cf. Corollary 14.5 (Wainwright, 2019))

$$\mathcal{R}_n(\delta; \widetilde{\mathcal{F}}) \leqslant \sqrt{\frac{2}{n}} \cdot \sqrt{\sum_{j=1}^\infty \min\{\mu_j, \delta^2\}}.$$

Since a reversible Markov chain has a uniform stationary distribution, the eigenvalues of $T_{\mathcal{K}}$ satisfies

$$\mu_j \leqslant c_0 e^{-c_1 j \log j}, \qquad \text{for } j = 1, 2, \dots$$

for some universal constants c_0, c_1 (see Chapter 12 in the reference (Wainwright, 2019)).

Let k be the smallest positive integer such that $c_0 e^{-c_1 k \log k} \leqslant \delta^2$. We have

$$\begin{aligned}
\sqrt{\frac{2}{n}} \cdot \sqrt{\sum_{j=1}^\infty \min\{\mu_j, \delta^2\}} &\leqslant \sqrt{\sum_{j=1}^\infty \min\{\delta^2, c_0 e^{-c_1 j \log j}\}} \\
&\leqslant \sqrt{k\delta^2 + c_0 \sum_{j=k+1}^\infty e^{-c_1 j \log j}} \\
&\leqslant \sqrt{k\delta^2 + c_0 e^{-c_1 k \log k}} \leqslant c_0 \sqrt{k\delta^2},
\end{aligned}$$

where c_0, c_1 denote universal constants whose values may change from line to line. Some algebra then shows $\omega_n^2 \lesssim \frac{\log(n+1)}{n}$ and $\varepsilon_n^2 \lesssim \frac{\log(n+1)}{n}$, as claimed.

D.5 Convex Lipschitz functions

We use results involving chaining and empirical process theory to compute the critical radii in this setting, see the references (Wainwright, 2019; van de Geer, 2000; Talagrand, 2014) for a full exposition. Recall the definition of the event $\mathcal{E}_1(t)$ from Equation (44), and let ω_n denote the critical radius. By an argument similar to that of the proof of Lemma 3, we have for $k \geqslant 1$,

$$\mathbb{E}_\varepsilon\left[\sup_{g\in\widetilde{\mathcal{F}}, \|g\|_\mu \leqslant \omega_n} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, g(x_i)\right|\right] \leqslant \mathbb{E}_\varepsilon\left[\sup_{g\in\widetilde{\mathcal{F}}, \|g\|_n \leqslant (k+2)\omega_n} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, g(x_i)\right|\right],$$

with probability exceeding $1 - c_0 e^{-c_1 n \frac{k^2 \omega_n^2}{b^2}}$. We bound the right-hand side via chaining: we know from existing results (see Chapter 14 in the book (Wainwright, 2019)) that for Lipschitz function class $\widetilde{\mathcal{F}}$,

$$\mathbb{E}_\varepsilon\left[\sup_{g\in\widetilde{\mathcal{F}}, \|g\|_n \leqslant (k+2)\omega_n} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, g(x_i)\right|\right] \leqslant \frac{1}{\sqrt{n}}\left((k+2)\,\omega_n\right)^{3/4}.$$

Defining

$$Z := \mathbb{E}_\varepsilon \left[\sup_{g \in \widetilde{\mathcal{F}}, \|g\|_n \leqslant (k+2)\omega_n} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \, g(x_i) \right| \right]$$

We can integrate the tail bound to get

$$\begin{aligned}
\mathbb{E}[Z] &= \int_0^\infty \mathbb{P}(Z \geqslant u) \, du \\
&\leqslant \frac{1}{\sqrt{n}} (3\,\omega_n)^{3/4} + \frac{1}{\sqrt{n}} \int_{(3\omega_n)^{3/4}}^\infty \mathbb{P}(Z \geqslant u) \, du \\
&\overset{(a)}{\leqslant} \frac{1}{\sqrt{n}} (3\,\omega_n)^{3/4} + \frac{3c_0 \omega_n^{3/4}}{4} \int_0^\infty \exp\left(-c_1 n k^2 \omega_n^2\right) \cdot (k+2)^{-1/4} \, dk \\
&\leqslant (3\,\omega_n)^{3/4} + \frac{3c_0 \omega_n^{3/4}}{4\sqrt{n}} \int_0^\infty \exp\left(-c_1 n k^2 \omega_n^2\right) \, dk \\
&\overset{(b)}{=} \frac{1}{\sqrt{n}} (3\,\omega_n)^{3/4} + \frac{3c_0 \omega_n^{3/4}}{8\sqrt{n}} \cdot \sqrt{\frac{\pi}{n\omega_n^2}},
\end{aligned}$$

where step (a) follows from the change of variables $u = ((k+2)\,\omega_n)^{3/4}$ and step (b) follows from the integral of a Gaussian density. By definition (22), we have $\mathcal{R}_n(\omega_n; \widetilde{\mathcal{F}}) = \omega_n^2$, so we have for universal constants c_1 and c_2,

$$\omega_n^2 \leqslant c_1 \frac{\omega_n^{3/4}}{\sqrt{n}} + \frac{c_2}{\omega_n^{1/4} n}.$$

Thus we conclude $\omega_n^2 \lesssim \frac{1}{n^{4/5}}$, as claimed.