# A Universal Empirical Dynamic Programming Algorithm for Continuous State MDPs

William B. Haskell, Rahul Jain, Hiteshi Sharma, and Pengqian Yu

***Abstract*—We propose universal randomized function approximation-based empirical value learning (EVL) algorithms for Markov decision processes. The "empirical" nature comes from each iteration being done empirically from samples available from simulations of the next state. This makes the Bellman operator a random operator. A parametric and a nonparametric method for function approximation using a parametric function space and a reproducing kernel Hilbert space respectively are then combined with EVL. Both function spaces have the universal function approximation property. Basis functions are picked randomly. Convergence analysis is performed using a random operator framework with techniques from the theory of stochastic dominance. Finite time sample complexity bounds are derived for both universal approximate dynamic programming algorithms. Numerical experiments support the versatility and computational tractability of this approach.**

***Index Terms*—Continuous state-space Markov decision processes (MDPs), dynamic programming (DP), reinforcement learning (RL).**

## I. Introduction

THERE exist a wide variety of approximate dynamic programming (DP) [2, Ch. 6], [3] and reinforcement learning (RL) algorithms [4] for finite state-space Markov decision processes (MDPs). But many real-world problems of interest have either a continuous state space, or very large state space that it is best approximated as one. Action space will be considered finite. Approximate DP (ADP) and RL algorithms do exist for continuous state-space MDPs but choosing which one to employ is an art form: different techniques (state-space aggregation and function approximation [5]) and algorithms work for different problems [6]–[8], and universally applicable algorithms are lacking. For example, fitted value iteration (FVI) [9] is very effective for some problems but requires the choice of an appropriate basis functions for good approximation. Most of the existing work on ADP requires domain knowledge of the problem at hand for effective implementation. Here, we are interested in ADP methods, which are effective without any previous problem knowledge.

In this paper, we propose ADP algorithms for continuous state-space MDPs with finite action space that are universal (approximating function space can provide arbitrarily good approximation for any problem), computationally tractable, simple to implement, and yet we have nonasymptotic sample complexity bounds. The first is accomplished by picking functions spaces for approximation that are dense in the space of continuous functions (i.e., for any continuous function $f$, and $\epsilon > 0$, there is an element of our approximating function space that is within $\epsilon$ of $f$ in the sup-norm.) The second goal is achieved by relying on randomized selection of basis functions for approximation and also by "empirical" DP [10]. The third is enabled because standard Python routines can be used for function fitting and the fourth is by analysis in a random operator framework, which provides nonasymptotic rate of convergence and sample complexity bounds.

There is a large body of well-known literature on RL and ADP for continuous state-space MDPs. We discuss the most directly related. In [11], a sampling-based state-space aggregation scheme combined with sample average approximation for the expectation in the Bellman operator was proposed. Under some regularity assumptions, the approximate value function can be computed at any state and an estimate of the expected error is given. But the algorithm seems to suffer from poor numerical performance. A linear programming-based constraint-sampling approach was introduced in [12]. Finite sample error guarantees, with respect to this constraint-sampling distribution, are provided but the method suffers from issues of feasibility. The closest paper to this study is the study by Munos and Szepesvári [9] that does function fitting with a given basis and does "empirical" value iteration in each step. Unfortunately, it is not a universal method as approximation quality depends on the function basis picked. Other papers worth noting are the study by Ormoneit and Sen[13] that discusses kernel-based value iteration and the bias-variance tradeoff, and the study by Grunewalder *et al.* [14] that proposed a kernel-based algorithm with random sampling of the state and action spaces, and proves asymptotic convergence. Other related works worth mentioning are [15], [16] (approximate value iteration), [17], [18] (the LP approach to ADP), and [19]–[21] (approximate policy iteration).

Manuscript received August 22, 2018; revised August 23, 2018 and March 1, 2019; accepted March 3, 2019. Date of publication April 1, 2019; date of current version December 27, 2019. This work was supported by the Singapore Ministry of Education Project MOE2015-T2-2-148. The work of Jain and Sharma was supported by an ONR Young Investigator Award #N000141210766 and by NSF Award CCF-1817212. A preliminary version of this paper appeared in CDC 2017 [1]. Recommended by Associate Editor E. Zhou. *(Corresponding author: Rahul Jain.)*

W. B. Haskell and P. Yu are with the Department of Industrial and Systems Engineering, National University of Singapore, Singapore 129792 (e-mail: isehwb@nus.edu.sg; yupengqian@u.nus.edu).

R. Jain and H. Sharma are with the EE Department, University of Southern California, Los Angeles, CA 90089 USA (e-mail: rahul.jain@usc.edu; hiteshis@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2019.2907414

Recent applications stress policy gradient methods [22], [23] and deep learning-based function approximation [24] for which theoretical performance guarantees for general problems are not available. The method presented in this paper may be seen as another alternative.

This paper is inspired by the "random function" approach that uses randomization to (nearly) solve otherwise intractable problems (see, e.g., [25] and[26]) and the "empirical" approach that reduces computational complexity of working with expectations [10]. We propose two new algorithms. For the first parametric approach, we pick a parametric function family. In each iteration, a number of functions are picked randomly for function fitting by sampling the parameters. A preliminary version of this for $l_2$ function fitting appeared in [1]. For the second nonparametric approach, we pick a reproducing kernel Hilbert space (RKHS) for approximation. Both function spaces are dense in the space of continuous functions. In each iteration, we sample a few states from the state space. Empirical value learning (EVL) is then performed on these states. Each step of EVL involves approximating the Bellman operator with an empirical (random) Bellman operator by plugging a sample average approximation from simulation for the expectation. This is akin to doing stochastic approximations with step size 1. We employ a probabilistic convergence analysis technique of iterated random operators based on stochastic dominance that we developed in [10]. This method is general in the sense that not only can we handle various norms, but also various random contractive operators.

The main contribution of this paper is the development of randomized function approximation-based (offline) DP algorithms that are universally applicable (i.e., do not require appropriate choice of basis functions for good approximation). A secondary contribution is further development of the random operator framework for convergence analysis in the $\mathcal{L}_p-$norm that also yields finite time sample complexity bounds.

The paper is organized as follows. Section II presents preliminaries including the continuous state-space MDP model and the empirical DP framework for finite state MDPs introduced in [10]. Section III presents two EVL algorithms—first, a randomized parametric function fitting method, and second, a nonparametric randomized function fitting in an RKHS space. We also provide statements of main theorems about nonasymptotic error guarantees. Section IV presents a unified analysis of the two algorithms in a random operator framework. Numerical results are reported in Section V. Supplemental proofs are relegated to the appendix.

## II. Preliminaries

Consider a discrete time discounted MDP given by the 5-tuple, $(\mathbb{S},\, \mathbb{A},\, Q,\, c,\, \gamma)$. The state space $\mathbb{S}$ is a compact subset of $\mathbb{R}^d$ with the Euclidean norm, with corresponding Borel $\sigma-$algebra $\mathcal{B}(\mathbb{S})$. Let $\mathcal{F}(\mathbb{S})$ be the space of all $\mathcal{B}(\mathbb{S})-$measurable bounded functions $f : \mathbb{S} \to \mathbb{R}$ in the supremum norm $\|f\|_\infty := \sup_{s\in\mathbb{S}} |f(s)|$. Moreover, let $\mathcal{M}(\mathbb{S})$ be the space of all probability distributions over $\mathbb{S}$ and define the $\mathcal{L}_p$ norm as $\|f\|_{p,\,\mu}^p := (\int_{\mathbb{S}} |f(s)|^p \mu(ds))$ for $p \in [1,\, \infty)$ and given $\mu \in \mathcal{M}(\mathbb{S})$. We assume that the action space $\mathbb{A}$ is finite. The transition law $Q$ governs the system evolution. For $B \in \mathcal{B}(\mathbb{S})$, $Q(B \,|\, s,\, a)$ is the probability of next visiting the set $B$ given that action $a \in \mathbb{A}$ is chosen in state $s \in \mathbb{S}$. The cost function $c : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ is a bounded measurable function that depends on state-action pairs. Finally, $\gamma \in (0,\, 1)$ is the discount factor.

We will denote by $\Pi$ the class of *stationary deterministic Markov policies*: mappings $\pi : \mathbb{S} \to \mathbb{A}$, which only depend on history through the current state. For a given state $s \in \mathbb{S}$, $\pi(s) \in \mathbb{A}$ is the action chosen in state $s$ under the policy $\pi$. The state and action at time $t$ are denoted $s_t$ and $a_t$, respectively. Any policy $\pi \in \Pi$ and initial state $s \in \mathbb{S}$ determine a probability measure $P_s^\pi$ and a stochastic process $\{(s_t,\, a_t),\, t \geq 0\}$ defined on the canonical measurable space of trajectories of state-action pairs. The expectation operator with respect to $P_s^\pi$ is denoted $\mathbb{E}_s^\pi[\cdot]$.

We will assume that the cost function $c$ satisfies $|c(s,\, a)| \leq c_{\max} < \infty$ for all $(s,\, a) \in \mathbb{S} \times \mathbb{A}$. Under this assumption, $\|v^\pi\|_\infty \leq v_{\max} := c_{\max}/(1-\gamma)$ where $v^\pi$ is the value function for policy $\pi \in \Pi$ defined as $v^\pi(s) = \mathbb{E}_s^\pi[\sum_{t=0}^\infty \gamma^t c(s_t,\, a_t)]$, $\forall s \in \mathbb{S}$. For later use, we define $\mathcal{F}(\mathbb{S};\, v_{\max})$ to be the space of all functions $f \in \mathcal{F}(\mathbb{S})$ such that $\|f\|_\infty \leq v_{\max}$.

The optimal value function is $v^*(s) := \inf_{\pi\in\Pi} \mathbb{E}_s^\pi[\sum_{t=0}^\infty \gamma^t c(s_t,\, a_t)]$, $\forall s \in \mathbb{S}$. To characterize the optimal value function, we define the Bellman operator $T : \mathcal{F}(\mathbb{S}) \to \mathcal{F}(\mathbb{S})$ via

$$[T\,v]\,(s) := \min_{a\in\mathbb{A}} \left\{c\,(s,\, a) + \gamma\, \mathbb{E}_{X\sim Q(\cdot\,|\,s,\,a)}\,[v(X)]\right\} \quad \forall s \in \mathbb{S}.$$

It is well known that the optimal value function $v^*$ is a fixed point of $T$, i.e., $T\,v^* = v^*$ [27, Th. 6.2.5]. Classical value iteration is based on iterating $T$ to obtain a fixed point, it produces a sequence $(v_k)_{k\geq 0} \subset \mathcal{F}(\mathbb{S})$ given by $v_{k+1} = T\,v_k$, $k \geq 0$. Also, we know that $(v_k)_{k\geq 0}$ converges to $v^*$ geometrically in $\|\cdot\|_\infty$.

We are interested in approximating the optimal value function $v^*$ within a tractable class of approximating functions $\mathcal{F} \subset \mathcal{F}(\mathbb{S})$. We have the following definitions, which we use to measure the approximation power of $\mathcal{F}$ with respect to $T$. We define

$$d_{p,\,\mu}\,(\mathcal{G},\mathcal{F}) := \sup_{g\in\mathcal{G}} \inf_{f\in\mathcal{F}} \|f-g\|_{p,\,\mu}$$

to be the distance between two function classes; then $d_{p,\,\mu}(T\,\mathcal{F},\, \mathcal{F})$ is the inherent $\mathcal{L}_p$ *Bellman error for the function class* $\mathcal{F}$. Similarly, defining

$$d_\infty\,(\mathcal{G},\mathcal{F}) := \sup_{g\in\mathcal{G}} \inf_{f\in\mathcal{F}} \|f-g\|_\infty$$

gives $d_\infty(T\mathcal{F},\mathcal{F})$ as the inherent $\mathcal{L}_\infty$ Bellman error for an approximating class $\mathcal{F}$.

We often compare $\mathcal{F}$ to the Lipschitz continuous functions $\mathrm{Lip}(L)$ defined as

$$\{f \in \mathcal{F}\,(\mathbb{S}) : |f(s) - f\,(s')\,| \leq L\,\|s-s'\| \quad \forall s,\, s' \in \mathbb{S}\}\,.$$

In our case, we say that an approximation class $\mathcal{F}$ is *universal* if $d_\infty(\mathrm{Lip}(L),\, \mathcal{F}) = 0$ for all $L \geq 0$. Note that on a compact state space $\mathbb{S}$, universality in the supremum norm implies universality in the $\mathcal{L}_1$ and $\mathcal{L}_2$ norms as well.

One of the difficulties of DP algorithms such as value iteration above is that each iteration of the Bellman operator involves

computation of an expectation, which may be expensive. Thus, in [10], Haskell *et al.* proposed replacing the Bellman operator with an empirical (or random) Bellman operator

$$\left[\hat{T}_n v\right](s) := \min_{a\in\mathbb{A}} \left\{ c(s,\, a) + \frac{\gamma}{n}\sum_{i=1}^{n}[v(X_i)] \right\}$$

where $X_i$ are samples of the next state from $Q(\cdot \,|\, s,\, a)$, which can be obtained from simulation. Now, we can iterate the empirical Bellman operator

$$v_{k+1} = \hat{T}_n v_k \quad \forall k \geq 0$$

an algorithm we called empirical value iteration (EVI). The sequence of iterates $\{v_k\}$ is a random process. Since $T$ is a contractive operator, its iterates converge to its fixed point $v^*$. The random operator $\hat{T}_n$ may be expected to inherit the contractive property in a probabilistic sense and its iterates converge to some sort of a probabilitic fixed point. We introduce $(\epsilon, \delta)$ versions of two such notions introduced in [10].

*Definition 1:* A function $\hat{v} : \mathbb{S} \to \mathbb{R}$ is an $(\epsilon, \delta)$-strong probabilistic fixed point for a sequence of random operators $\{\hat{T}_n\}$ if there exists an $N$ such that for all $n > N$

$$\mathbb{P}\left(||\hat{T}_n\hat{v} - \hat{v}|| > \epsilon\right) < \delta.$$

It is called a *strong probabilistic fixed point*, if the above is true for every positive $\epsilon$ and $\delta$.

*Definition 2:* A function $\hat{v} : \mathbb{S} \to \mathbb{R}$ is an $(\epsilon, \delta)$-weak probabilistic fixed point for a sequence of random operators $\{\hat{T}_n\}$ if there exist $N$ and $K$ such that for all $n > N$ and all $k > K$

$$\mathbb{P}\left(||\hat{T}_n^k v_0 - \hat{v}|| > \epsilon\right) < \delta \quad \forall v_0 \in \mathcal{F}(\mathbb{S}).$$

It is called a *weak probabilistic fixed point*, if the above is true for every positive $\epsilon$ and $\delta$. Note that the stochastic iterative algorithms such as EVL often find the weak probabilistic fixed point of $\{\hat{T}_n\}$ whereas what we are looking for is $v^*$, the fixed point of $T$. In [10], it was shown that asymptotically the weak probabilistic fixed point of $\{\hat{T}_n\}$ coincides with its strong probabilistic fixed points, which coincide with the fixed point of $T$ under certain fairly weak assumptions and a natural relationship between $T$ and $\{\hat{T}_n\}$

$$\lim_{n\to\infty} \mathbb{P}\left(||\hat{T}_n v - Tv|| > \epsilon\right) = 0 \quad \forall v \in \mathcal{F}(\mathbb{S}).$$

This implies that stochastic iterative algorithms such as EVL will find approximate fixed points of $T$ with high probability.

## III. Algorithms and Main Results

When the state space $\mathbb{S}$ is very large, or even uncountable, exact DP methods are not practical, or even feasible. Instead, one must use a variety of approximation methods. In particular, function approximation (or fitting the value function with a fixed function basis) is a common technique. The idea is to sample a finite set of states from $\mathbb{S}$, approximate the Bellman update at these states, and then extend to the rest of $\mathbb{S}$ through function fitting similar to [9]. Furthermore, the expectation in the Bellman operator, for example, is also approximated by taking a number of samples of the next state. There are two main difficulties with this approach: First, the function fitting depends on the function basis chosen, making the results problem-dependent. Second, with a large basis (for good approximation), function fitting can be computationally expensive.

In this paper, we aim to address these issues by first picking universal approximating function spaces, and then using randomization to pick a smaller basis and thus reduce the computational burden of the function fitting step. We consider two functional families, one is a parametric family $\mathcal{F}(\Theta)$ parameterized over parameter space $\Theta$ and the other is a nonparametric regularized RKHS. By $\mu \in \mathcal{M}(\mathbb{S})$, we will denote a probability distribution from which to sample states in $\mathbb{S}$, and by a $\mathcal{F} \subset \mathcal{F}(\mathbb{S};\, v_{\max})$, we will denote a functional family in which to do value function approximation.

Let us denote by $(v_k)_{k\geq 0} \subset \mathcal{F}(\mathbb{S};\, v_{\max})$ the iterates of the value functions produced by an algorithm and a sample of size $N \geq 1$ from $\mathbb{S}$ is denoted $s^{1:N} = (s_1, \ldots,\, s_N)$. The empirical $p-$norm of $f$ is defined as $\|f\|_{p,\,\hat{\mu}}^p := \frac{1}{N}\sum_{n=1}^{N} |f(s_n)|^p$ for $p \in [1,\, \infty)$ and as $\|f\|_{\infty,\,\hat{\mu}} := \sup_{n=1,\ldots,N} |f(s_n)|$ for $p = \infty$, where $\hat{\mu}$ is the empirical measure corresponding to the samples $s^{1:N}$.

We will make the following technical assumptions for the rest of the paper similar to those made in [9].

*Assumption 1:*

1) For all $(s,\, a) \in \mathbb{S} \times \mathbb{A}$, $Q(\cdot \,|\, s,\, a)$ is absolutely continuous with respect to $\mu$ and
$$C_\mu := \sup_{(s,\,a)\in\mathbb{S}\times\mathbb{A}} \|dQ(\cdot \,|\, s,\, a)/d\mu\|_\infty < \infty.$$
2) Given any sequence of policies $\{\pi_m\}_{m\geq 1}$, the future state distribution $\rho Q^{\pi_1}\cdots Q^{\pi_m}$ is absolutely continuous with respect to $\mu$
$$c_{\rho,\,\mu}(m) := \sup_{\pi_1,\ldots,\pi_m} \|d(\rho Q^{\pi_1}\cdots Q^{\pi_m})/d\mu\|_\infty < \infty$$
and $C_{\rho,\,\mu} := \sum_{m\geq 0} \gamma^m c_{\rho,\,\mu}(m) < \infty$.

The above assumptions are conditions on transition probabilities, the first being a sufficient condition for the second. $\rho$ can be regarded as an "importance" distribution on $\mathbb{S}$, that is possibly different from the distribution $\mu$ on $\mathbb{S}$ that is used to sample states. Assumption 1 is essentially a regularity condition on the MDP: It ensures that the MDP cannot make arbitrary transitions with high probability with respect to the initial state distribution $\mu$. $C_{\rho,\mu}$ is called the discounted-average concentrability coefficient of the future-state distributions in [9]. Note that the assumption is satisfied when $\mu$ is the Lebesgue measure on $\mathbb{S}$ and the transition kernel has a bounded density with respect to $\mu$.

### A. Random Parametric Basis Function (RPBF) Approximation

We introduce an EVL algorithm with function approximation using random parametrized basis functions (EVL+RPBF). It requires a parametric family $\mathcal{F}$ built from a set of parameters $\Theta$ with probability distribution $\nu$ and a feature function $\phi : \mathbb{S} \times \Theta \to \mathbb{R}$ (that depends on both states and parameters) with the assumption that $\sup_{(s,\,\theta)\in\mathbb{S}\times\Theta} |\phi(s;\, \theta)| \leq 1$. This can

easily be met in practice by scaling $\phi$ whenever $\mathbb{S}$ and $\Theta$ are both compact and $\phi$ is continuous in $(s,\, \theta)$. Let $\alpha : \Theta \to \mathbb{R}$ be a weight function and define $\mathcal{F}(\Theta) :=$

$$\left\{ f(\cdot) = \int_{\Theta} \phi\left(\cdot;\, \theta\right) \alpha\left(\theta\right) d\theta \;:\; |\alpha\left(\theta\right)| \leq C\, \nu\left(\theta\right) \quad \forall \theta \in \Theta \right\}.$$

We note that the condition $|\alpha(\theta)| \leq C\,\nu(\theta)$ for all $\theta \in \Theta$ is equivalent to requiring that $\|\alpha\|_{\infty,\,\nu} := \sup_{\theta\in\Theta} |\alpha(\theta)/\nu(\theta)| \leq C$ where $\|\alpha\|_{\infty,\,\nu}$ is the $\nu-$weighted supremum norm of $\alpha$ and $C$ is a constant.

The function space $\mathcal{F}(\Theta)$ may be chosen to have the "universal" function approximation property in the sense that any Lipschitz continuous function can be approximated arbitrarily closely in this space as shown in [25]. By [25, Th. 2], many such choices of $\mathcal{F}(\Theta)$ are possible and are developed in [25, Sec. 5]. For example, $\mathcal{F}(\Theta)$ is universal in the following two cases.

1) $\phi(s;\, \theta) = \cos(\langle \omega,\, s\rangle + b)$ where $\theta = (\omega,\, b) \in \mathbb{R}^{d+1}$; and $\nu(\theta)$ is given by $\omega \sim \text{Normal}(0,\, 2\,\gamma\, I)$ and $b \sim \text{Uniform}\,[-\pi,\, \pi]$.
2) $\phi(s;\, \theta) = \text{sign}(s_k - t)$ where $\theta = (t,\, k) \in \mathbb{R} \times \{1, \ldots,\, d\}$; and $\nu(\theta)$ to be given by $k \sim \text{Uniform}\{1, \ldots,\, d\}$ and $t \sim \text{Uniform}[-a,\, a]$.

In this approach, we have a parametric function family $\mathcal{F}(\Theta)$ but instead of optimizing over parameters in $\Theta$, we randomly sample them first and then do function fitting, which involves optimizing over finite weighted combinations $\sum_{j=1}^{J} \alpha_j \phi(\cdot;\, \theta_j)$. Unfortunately, this leads to a nonconvex optimization problem. Hence, instead of optimizing over $\theta^{1:J} = (\theta_1, \ldots,\, \theta_J)$ and $\alpha^{1:J} = (\alpha_1, \ldots,\, \alpha_J)$ jointly, we first do randomization over $\theta^{1:J}$ and then optimization over $\alpha^{1:J}$, as in [26], to bypass the nonconvexity inherent in optimizing over $\theta^{1:J}$ and $\alpha^{1:J}$ simultaneously. This approach allows us to deploy rich parametric families without much additional computational cost. Once we draw a random sample $\{\theta_j\}_{j=1}^{J}$ from $\Theta$ according to $\nu$, we obtain a random function space: $\widehat{\mathcal{F}}(\theta^{1:J}) :=$

$$\left\{ f(\cdot) = \sum_{j=1}^{J} \alpha_j \phi\left(\cdot;\, \theta_j\right) \;:\; \|\left(\alpha_1, \ldots,\, \alpha_J\right)\|_{\infty} \leq C/J \right\}.$$

Step 1 of such an algorithm (Algorithm 1) involves sampling states $s^{1:N}$ over which to do value iteration and sampling parameters $\theta^{1:J}$ to pick basis functions $\phi(\cdot;\theta)$, which are used to do function fitting. Step 2 involves doing an EVI over states $s^{1:N}$ by sampling next states $(X_m^{s_n,\,a})_{m=1}^{M}$ according to the transition kernel $Q$, and using the current iterate of the value function $v_k$. Note that fresh (i.i.d.) samples of the next state are regenerated in each iteration. Step 3 involves finding the best fit to $\tilde{v}_k$, the iterate from Step 2, within $\widehat{\mathcal{F}}(\theta^{1:J})$ wherein randomly sampled parameters $\theta^{1:J}$ specify the basis functions for function fitting and weights $\alpha^{1:J}$ are optimized, which is a convex optimization problem.

We note that Step 3 of the algorithm can be replaced by another method for function fitting (as we do in the next section). The above algorithm differs from FVI algorithm of [9] in how it does function fitting. FVI does function fitting with a deterministic and given set of basis functions, which limits its universality, whereas we do function fitting in a much larger space, which has the universal function approximation property, but are able to reduce computational complexity by exploiting randomization.

**Algorithm 1:** EVL with Random Parameterized basis Functions (EVL+RPBF).

Input: probability distribution $\mu$ on $\mathbb{S}$ and $\nu$ on $\Theta$;
Sample sizes $N \geq 1$, $M \geq 1$, $J \geq 1$; initial seed $v_0$.
counter $k = 0$ and iterations $K \geq 1$.
For $k = 1, \ldots, K$
1) Sample $(s_n)_{n=1}^{N} \sim \mu^N$ and $(\theta_j)_{j=1}^{J} \sim \nu^J$.
2) Compute

$$\tilde{v}_k\left(s_n\right) = \min_{a\in\mathbb{A}} \left\{ c\left(s_n,\, a\right) + \frac{\gamma}{M} \sum_{m=1}^{M} v_k\left(X_m^{s_n,\,a}\right) \right\},$$

where $(X_m^{s_n,\,a}) \sim Q(\cdot \,|\, s_n,\, a)$, $m = 1, \cdots, M$ are i.i.d.
3) $\alpha^k = \arg\min_{\alpha}\ \frac{1}{N} \sum_{n=1}^{N} (\sum_{j=1}^{J} \alpha_j \phi(s_n;\, \theta_j) - \tilde{v}(s_n))^2$
s.t. $\|(\alpha_1, \ldots,\, \alpha_J)\|_{\infty} \leq C/J$.
$v_{k+1}(s) = \sum_{j=1}^{J} \alpha_j^k \phi(s; \theta_j)$.
4) Increment $k \leftarrow k+1$ and return to Step 1.

In [9, Sec. 7], it is shown that if the transition kernel and cost are smooth such that there exist $L_Q$ and $L_c$ for which

$$\|Q\left(\cdot \,|\, s,\, a\right) - Q\left(\cdot \,|\, s',\, a\right)\|_{TV} \leq L_Q \|s - s'\|_2 \tag{1}$$

and

$$|c\left(s,\, a\right) - c\left(s',\, a\right)| \leq L_c \|s - s'\|_2 \tag{2}$$

hold for all $s,\, s' \in \mathbb{S}$ and $a \in \mathbb{A}$, then the Bellman operator $T$ maps bounded functions to Lipschitz continuous functions. In particular, if $v$ is uniformly bounded by $v_{\max}$, then $T\,v$ is $(L_c + \gamma\, v_{\max} L_Q)-$Lipschitz continuous. Subsequently, the inherent $\mathcal{L}_\infty$ Bellman error satisfies $d_\infty(T\,\mathcal{F},\, \mathcal{F}) \leq d_\infty(\text{Lip}(L),\, \mathcal{F})$ since $T\,\mathcal{F} \subset \text{Lip}(L)$. So, it only remains to choose an $\mathcal{F}(\Theta)$ that is dense in $\text{Lip}(L)$ in the supremum norm, for which many examples exist.

We now provide nonasymptotic sample complexity bounds to establish that Algorithm 1 yields an approximately optimal value function with high probability. We provide guarantees for both the $\mathcal{L}_1$ and $\mathcal{L}_2$ metrics on the error.

Denote

$$N_2(\varepsilon, \delta') = \; 2^7 5^2 \bar{v}_{\max}^4 \log\left[ \frac{40\, e\, (J_2+1)}{\delta} \left(10\, e\, \bar{v}_{\max}^2\right)^J \right]$$

$$M_2(\varepsilon, \delta') = \; \left(\frac{\bar{v}_{\max}^2}{2}\right) \log\left[\frac{10\, N_2\, |\mathbb{A}|}{\delta'}\right]$$

$$J_2(\varepsilon, \delta') = \; \left( \frac{5C}{\varepsilon} \left( 1 + \sqrt{2 \log \frac{5}{\delta'}} \right) \right)^2, \text{ and}$$

$$K_2^* = \; 2 \left\lceil \frac{\ln\left(C_{\rho,\,\mu}^{1/2}\, \varepsilon\right) - \ln\left(2\, v_{\max}\right)}{\ln\, \gamma} \right\rceil$$

where $\bar{v}_{\max} = v_{\max}/\varepsilon$. Set $\delta' := 1 - (1-\delta/2)^{1/(K_2^*-1)}$. Then, we have the following sample complexity bound on Algorithm 1 with $\mathcal{L}_2$ error. We note that $\mathcal{L}_{2,\mu}(\mathbb{S})$ is a Hilbert space and that many powerful function approximation results exist for this setting because of the favorable properties of a Hilbert space.

*Theorem 1:* Given an $\varepsilon > 0$, and a $\delta \in (0,\,1)$, choose $J \geq J_2(\varepsilon, \delta')$, $N \geq N_2(\varepsilon, \delta')$, $M \geq M_2(\varepsilon, \delta')$, Then, for $K \geq \log(4/(\delta\,\mu^*(\delta; K_2^*)))$, we have

$$\|v_K - v^*\|_{2,\rho} \leq 2\tilde{\gamma}^{1/2} C_{\rho,\mu}^{1/2} \left(d_{2,\mu}\left(T\,\mathcal{F}\left(\Theta\right),\,\mathcal{F}\left(\Theta\right)\right) + 2\,\varepsilon\right)$$

with probability at least $1-\delta$.

*Remarks:*

1) That is, if we choose enough samples $N_2$ of the states, enough samples $M_2$ of the next state, and enough random samples $J_2$ of the parameter $\theta$, and then for large enough number of iterations $K_2$, the $\mathcal{L}_2$ error in the value function is determined by the inherent Bellman error of the function class $\mathcal{F}(\Theta)$.
2) For the function families $\mathcal{F}(\Theta)$ discussed earlier (RPBF), the inherent Bellman error, $d_{2,\mu}(T\,\mathcal{F}(\Theta),\,\mathcal{F}(\Theta)) = 0$ indeed, and so the value function will have small $\mathcal{L}_2$ error with high probability.
3) Note that the sample complexity bounds are independent of the state-space dimension though the computational complexity of sampling from the state space does indeed depend on that dimension.

Next we give a similar guarantee for $\mathcal{L}_1$ error for Algorithm 1 by considering approximation in $\mathcal{L}_{1,\mu}(\mathbb{S})$.

Denote

$$\begin{aligned}
N_1(\varepsilon,\delta') &= 2^7 5^2 \bar{v}_{\max}^2 \log\left[\frac{40\,e\,(J_1+1)}{\delta}\left(10\,e\,\bar{v}_{\max}\right)^J\right]\\
M_1(\varepsilon,\delta') &= \left(\frac{\bar{v}_{\max}^2}{2}\right)\log\left[\frac{10\,N_1\,|\mathbb{A}|}{\delta'}\right]\\
J_1(\varepsilon,\delta') &= \left(\frac{5C}{\varepsilon}\left(1+\sqrt{2\log\frac{5}{\delta'}}\right)\right)^2\\
K_1^* &= \left\lceil\frac{\ln\left(C_{\rho,\mu}\varepsilon\right)-\ln\left(2\,v_{\max}\right)}{\ln\,\gamma}\right\rceil,\ \text{and}\\
\mu^*\left(p;\,K^*\right) &= (1-p)\,p^{(K^*-1)}
\end{aligned}$$

where $C$ is the same constant that appears in the definition of $\mathcal{F}(\Theta)$ (see [26]) and $\bar{v}_{\max} = v_{\max}/\varepsilon$. Set $\delta' := 1-(1-\delta/2)^{1/(K_1^*-1)}$.

*Theorem 2:* Given an $\varepsilon > 0$, and a $\delta \in (0,\,1)$, choose $J \geq J_1(\varepsilon,\delta')$, $N \geq N_1(\varepsilon,\delta')$, and $M \geq M_1(\varepsilon,\delta')$, Then, for $K \geq \log(4/(\delta\,\mu^*(\delta; K_1^*)))$, we have

$$\|v_K - v^*\|_{1,\rho} \leq 2\,C_{\rho,\mu}\left(d_{1,\mu}\left(T\,\mathcal{F}\left(\Theta\right),\,\mathcal{F}\left(\Theta\right)\right) + 2\,\varepsilon\right)$$

with probability at least $1-\delta$.

*Remarks:*

1) Again, note that the above result implies that the RBPF function family $\mathcal{F}(\Theta)$ has inherent Bellman error $d_{1,\mu}(T\,\mathcal{F}(\Theta),\,\mathcal{F}(\Theta)) = 0$, so that for enough samples $N_1$ of the states, enough samples $M_1$ of the next state, and enough random samples $J_1$ of the parameter $\theta$, and then for large enough number of iterations $K_1$, the value function will have small $\mathcal{L}_1$ error with high probability.
2) As above, note that there is no dependence on state-space dimension in the sample complexity bounds though computational complexity of sampling states from the state space indeed depends on it.

### B. Nonparametric Function Approximation in RKHS

We now consider nonparametric function approximation combined with EVL. We employ a RKHS for function approximation since for suitably chosen kernels, it is dense in the space of continuous functions and hence has a "universal" function approximation property. In the RKHS setting, we can obtain guarantees directly with respect to the supremum norm.

We will consider a regularized RKHS setting with a continuous, symmetric, and positive semidefinite kernel $K : \mathbb{S}\times\mathbb{S} \to \mathbb{R}$ and a regularization constant $\lambda > 0$. The RKHS space, $\mathcal{H}_K$, is defined to be the closure of the linear span of $\{K(s,\cdot)\}_{s\in\mathbb{S}}$ endowed with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$. The inner product $\langle\cdot,\,\cdot\rangle_{\mathcal{H}_K}$ for $\mathcal{H}_K$ is defined such that $\langle K(x,\,\cdot),\,K(y,\,\cdot)\rangle_{\mathcal{H}_K} = K(x,\,y)$ for all $x,\,y\in\mathbb{S}$, i.e., $\langle\sum_i \alpha_i K(x_i,\,\cdot),\,\sum_j \beta_j K(y_j,\,\cdot)\rangle_{\mathcal{H}_K} = \sum_{i,\,j}\alpha_i\beta_j K(x_i,\,y_j)$. Subsequently, the inner product satisfies the reproducing property: $\langle K(s,\,\cdot),\,f\rangle_{\mathcal{H}_K} = f(s)$ for all $s\in\mathbb{S}$ and $f\in\mathcal{H}_K$. The corresponding RKHS norm is defined in terms of the inner product $\|f\|_{\mathcal{H}_K} := \sqrt{\langle f,\,f\rangle_{\mathcal{H}_K}}$. We assume that our kernel $K$ is bounded so that $\kappa := \sup_{s\in\mathbb{S}}\sqrt{K(s,\,s)} < \infty$.

To find the best fit $f\in\mathcal{H}_K$ to a function with data $\{(s_n, \tilde{v}(s_n))\}_{n=1}^N$, we solve the regularized least squares problem

$$\min_{f\in\mathcal{H}_K}\left\{\frac{1}{N}\sum_{n=1}^N\left(f\left(s_n\right)-\tilde{v}\left(s_n\right)\right)^2 + \lambda\,\|f\|_{\mathcal{H}_K}^2\right\}. \tag{3}$$

This is a convex optimization problem (the norm squared is convex), and has a closed-form solution by the Representer Theorem. In particular, the optimal solution is of the form $\hat{f}(s) = \sum_{n=1}^N \alpha_n K(s_n,\,s)$ where the weights $\alpha^{1:N} = (\alpha_1,\ldots,\,\alpha_N)$ are the solution to the linear system

$$\left(\left[K\left(s_i,\,s_j\right)\right]_{i,\,j=1}^N + \lambda\,N\,I\right)\left(\alpha_n\right)_{n=1}^N = \left(\tilde{v}\left(s_n\right)\right)_{n=1}^N. \tag{4}$$

This yields EVL algorithm with randomized function fitting in a regularized RKHS (EVL+RKHS) displayed as Algorithm 2.

Note that the optimization problem in Step 3 in Algorithm 2 is analogous to the optimization problem in Step 3 of Algorithm 1, which finds an approximate best fit within the finite-dimensional space $\widehat{\mathcal{F}}(\theta^{1:J})$, rather than the entire space $\mathcal{F}(\Theta)$, while Problem (3) in Algorithm 2 optimizes over the entire space $\mathcal{H}_K$. This difference can be reconciled by the Representer Theorem, since it states that optimization over $\mathcal{H}_K$ in Problem (3) is equivalent to optimization over the finite-dimensional space spanned by $\{K(s_n,\,\cdot) : n = 1,\ldots,\,N\}$. Note that the regularization $\lambda\,\|f\|_{\mathcal{H}_K}^2$ is a requirement of the Representer Theorem.

We define the *regression function* $f_M : \mathbb{S}\to\mathbb{R}$ via

$$f_M(s) \triangleq \mathbb{E}\left[\min_{a\in\mathbb{A}}\left\{c\left(s,\,a\right) + \frac{\gamma}{M}\sum_{m=1}^M v\left(X_m^{s,\,a}\right)\right\}\right] \qquad \forall s\in\mathbb{S}$$

**Algorithm 2:** EVL with Regularized RKHS (EVL+RKHS).

Input: probability distribution $\mu$ on $\mathbb{S}$;
sample sizes $N \geq 1$, $M \geq 1$; penalty $\lambda$;
initial seed $v_0$; counter $k = 0$.
For $k = 1, \ldots, K$
1) Sample $\{s_n\}_{n=1}^N \sim \mu$.
2) Compute

$$\tilde{v}_k(s_n) = \min_{a \in \mathbb{A}} \left\{ c(s_n, a) + \frac{\gamma}{M} \sum_{m=1}^{M} v_k\left(X_m^{s_n, a}\right) \right\},$$

where $\{X_m^{s_n, a}\}_{m=1}^M \sim Q(\cdot \mid s_n, a)$ are i.i.d.
3) $v_{k+1}(\cdot)$ is given by

$$\arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{n=1}^{N} (f(s_n) - \tilde{v}(s_n))^2 + \lambda ||f||_{\mathcal{H}_K} \right\}.$$

4) Increment $k \leftarrow k + 1$ and return to Step 1.

it is the expected value of our empirical estimator of $T v$. As expected, $f_M \to T v$ as $M \to \infty$. We note that $f_M$ is not necessarily equal to $T v$ by Jensen's inequality. We require the following assumption on $f_M$ to continue.

*Assumption 2:* For every $M \geq 1$, $f_M(s) = \int_{\mathbb{S}} K(s, y)\alpha(y)\mu(dy)$ for some $\alpha \in \mathcal{L}_{2,\mu}(\mathbb{S})$.

Regression functions play a key role in the statistical learning theory, Assumption 2 states that the regression function lies in the span of the kernel $K$. It is satisfied whenever $K$ is a universal kernel. Some examples of universal kernels follow. Additionally, when $\mathcal{H}_K$ is dense in the space of Lipschitz functions, then the inherent Bellman error is zero. For example, $K(s, s') = \exp(-\gamma \|s - s'\|_2)$, $K(s, s') = 1 - \frac{1}{a}\|s - s'\|_1$, and $K(s, s') = \exp(\gamma \|s - s'\|_1)$ are all universal kernels.

Denote

$$N_\infty(\varepsilon, \delta') = \left(\frac{4 C_K \kappa}{\varepsilon(1-\gamma)}\right)^6 \log\left(\frac{4}{\delta'}\right)^2$$

$$M_\infty(\varepsilon) = \frac{160\, v_{\max}^2}{(\varepsilon(1-\gamma))^2} \log\left(\frac{2|\mathbb{A}|\gamma(8 v_{\max} - \varepsilon(1-\gamma))}{\varepsilon(1-\gamma)(2-\gamma)}\right)$$

$$K_\infty^* = \left\lceil \frac{\ln(\varepsilon) - \ln(4 v_{\max})}{\ln \gamma} \right\rceil$$

where $C_K$ is a constant independent of the dimension of $\mathbb{S}$ (see [28] for the details on how $C_K$ depends on the kernel $K$) and set $\delta' = 1 - (1 - \delta/2)^{1/(K_\infty^* - 1)}$.

*Theorem 3:* Suppose Assumption 2 holds. Given any $\varepsilon > 0$ and $\delta \in (0, 1)$, choose an $N \geq N_\infty(\varepsilon, \delta')$ and an $M \geq M_\infty(\varepsilon)$. Then, for any $K \geq \log(4/(\delta\, \mu^*(\delta; K_\infty^*)))$

$$\|v_K - v^*\|_\infty \leq \varepsilon$$

with probability at least $1 - \delta$.

Note that we provide guarantees on $\mathcal{L}_1$ and $\mathcal{L}_2$ error (can be generalized to $\mathcal{L}_p$) with the RPBF method and for $\mathcal{L}_\infty$ error with the RKHS-based randomized function fitting method. Getting guarantees for the $\mathcal{L}_p$ error with the RKHS method has proved quite difficult, as has bounds on the $\mathcal{L}_\infty$ error with the RBPF method.

## IV. Analysis in a Random Operator Framework

We will analyze Algorithms 1 and 2 in terms of random operators since this framework is general enough to encompass many such algorithms. The reader can see that Step 2 of both algorithms involves iteration of the empirical Bellman operator, whereas Step 3 involves a randomized function fitting step, which is done differently and in different spaces in both algorithms. We use random operator notation to write these algorithms in a compact way, and then derive a clean and to a large-extent unified convergence analysis. The key idea is to use the notion of stochastic dominance to bound the error process with an easy to analyze "dominating" Markov chain. Then, we can infer the solution quality of our algorithms via the probability distribution of the dominating Markov chain. This analysis idea refines (and in fact, simplifies) the idea we introduced in [10] for MDPs with finite state and action spaces (where there is no function fitting) in the supremum norm. In this paper, we develop the technique further, give a stronger convergence rate, account for randomized function approximation, and also generalize the technique to $\mathcal{L}_p$ norms.

We introduce a probability space $(\Omega, \mathcal{B}(\Omega), P)$ on which to define random operators, where $\Omega$ is a sample space with elements denoted $\omega \in \Omega$, $\mathcal{B}(\Omega)$ is the Borel $\sigma-$algebra on $\Omega$, and $P$ is a probability distribution on $(\Omega, \mathcal{B}(\Omega))$. A random operator is an operator-valued random variable on $(\Omega, \mathcal{B}(\Omega), P)$. We define the first random operator on $\mathcal{F}(\mathbb{S})$ as $\widehat{T}(v) = (s_n, \tilde{v}(s_n))_{n=1}^N$ where $(s_n)_{n=1}^N$ is chosen from $\mathbb{S}$ according to a distribution $\mu \in \mathcal{M}(\mathbb{S})$ and

$$\tilde{v}(s_n) = \min_{a \in \mathbb{A}} \left\{ c(s_n, a) + \frac{\gamma}{M} \sum_{m=1}^{M} v\left(X_m^{s_n, a}\right) \right\}$$

$n = 1, \ldots, N$ is an approximation of $[T v](s_n)$ for all $n = 1, \ldots, N$. In other words, $\widehat{T}$ maps from $v \in \mathcal{F}(\mathbb{S}; v_{\max})$ to a randomly generated sample of $N$ input–output pairs $(s_n, \tilde{v}(s_n))_{n=1}^N$ of the function $T v$. Note that $\widehat{T}$ depends on sample sizes $N$ and $M$. Next, we have the function reconstruction operator $\widehat{\Pi}_{\mathcal{F}}$, which maps the data $(s_n, \tilde{v}(s_n))_{n=1}^N$ to an element in $\mathcal{F}$. Note that $\widehat{\Pi}_{\mathcal{F}}$ is not necessarily deterministic since Algorithms 1 and 2 use randomized function fitting. We can now write both algorithms succinctly as

$$v_{k+1} = \widehat{G}\, v_k := \widehat{\Pi}_{\mathcal{F}} \widehat{T}\, v_k \tag{5}$$

which can be further written in terms of residual error $\varepsilon_k = \widehat{G}\, v_k - T\, v_k$ as

$$v_{k+1} = \widehat{G}\, v_k = T\, v_k + \varepsilon_k. \tag{6}$$

Iteration of these operators corresponds to repeated samples from $(\Omega, \mathcal{B}(\Omega), P)$, so we define the space of sequences $(\Omega^\infty, \mathcal{B}(\Omega^\infty), \mathcal{P})$ where $\Omega^\infty = \times_{k=0}^\infty \Omega$ with elements denoted $\boldsymbol{\omega} = (\omega_k)_{k \geq 0}$, $\mathcal{B}(\Omega^\infty) = \times_{k=0}^\infty \mathcal{B}(\Omega)$, and $\mathcal{P}$ is the probability measure on $(\Omega^\infty, \mathcal{B}(\Omega^\infty))$ guaranteed by the Kolmogorov extension theorem applied to $\mathcal{P}$.

The random sequences $(v_k)_{k\geq 0}$ in Algorithms 1 and 2 given by

$$\begin{aligned} v_{k+1} &= \widehat{\Pi}_{\mathcal{F}} \widehat{T}(\omega_k)\, v_k \\ &= \widehat{\Pi}_{\mathcal{F}} \widehat{T}(\omega_k)\, \widehat{\Pi}_{\mathcal{F}} \widehat{T}(\omega_{k-1}) \cdots \widehat{\Pi}_{\mathcal{F}} \widehat{T}(\omega_0)\, v_0 \end{aligned}$$

for all $k \geq 0$ is a stochastic process defined on $(\Omega^\infty, \mathcal{B}(\Omega^\infty), \mathcal{P})$. We now analyze error propagation over the iterations.

Let us now bound how the Bellman residual at each iteration of EVL is changing. There have already been some results that address the error propagation both in $\mathcal{L}_\infty$ and $\mathcal{L}_p$ $(p \geq 1)$ norms [16]. After adapting [9, Lemma 3], we obtain the following $p$-norm error bounds on $v_K - v^*$ in terms of the errors $\{\varepsilon_k\}_{k\geq 0}$.

*Lemma 4:* For any $K \geq 1$, and $\varepsilon > 0$, suppose $\|\varepsilon_k\|_{p,\mu} \leq \varepsilon$ for all $k = 0, 1, \ldots, K-1$, then

$$\|v_K - v^*\|_{p,\rho} \leq 2 \left( \frac{1-\gamma^{K+1}}{1-\gamma} \right)^{\frac{p-1}{p}} \left[ C_{\rho,\mu}^{1/p} \varepsilon + \gamma^{K/p} (2\, v_{\max}) \right] \tag{7}$$

where $C_{\rho,\mu}$ is as defined in Assumption 3. Note that Lemma 4 assumes that $\|\varepsilon_k\|_{p,\mu} \leq \varepsilon$, which we will show subsequently that it is true with high probability.

The second inequality is for the supremum norm.

*Lemma 5:* For any $K \geq 1$ and $\varepsilon > 0$, suppose $\|\varepsilon_k\|_\infty \leq \varepsilon$ for all $k = 0, 1, \ldots, K-1$, then

$$\|v_K - v^*\|_\infty \leq \varepsilon / (1-\gamma) + \gamma^K (2\, v_{\max}). \tag{8}$$

Inequalities (7) and (8) are the key to analyzing iteration of (6).

### A. Convergence Analysis Using Stochastic Dominance

We now provide a (unified) convergence analysis for iteration of a sequence of random operators given by (5) and (6). Later, we will show how it can be applied to Algorithms 1 and 2. We will use $\|\cdot\|$ to denote a general norm in the following discussion, since our idea applies to all instances of $p \in [1, \infty)$ and $p = \infty$ simultaneously. The magnitude of the error in iteration $k \geq 0$ is then $\|\varepsilon_k\|$. We make the following key assumption for a general EVL algorithm.

*Assumption 3:* For $\varepsilon > 0$, there is a $q \in (0, 1)$ such that $\Pr\{\|\varepsilon_k\| \leq \varepsilon\} \geq q$ for all $k \geq 0$.

Assumption 3 states that we can find a lower bound on the probability of the event $\{\|\varepsilon_k\| \leq \varepsilon\}$ that is independent of $k$ and $(v_k)_{k\geq 0}$ (but does depend on $\varepsilon$). Equivalently, we are giving a lower bound on the probability of the event $\{\|T\, v_k - \widehat{G}\, v_k\| \leq \varepsilon\}$. This is possible for all of the algorithms that we proposed earlier. In particular, we can control $q$ in Assumption 3 through the sample sizes in each iteration of EVL. Naturally, for a given $\varepsilon$, $q$ increases as the number of samples grows.

We first choose $\varepsilon > 0$ and the number of iterations $K^*$ for our EVL algorithms to reach a desired accuracy [this choice of $K^*$ comes from the inequalities (7) and (8)]. We call iteration $k$ "good" if the error $\|\varepsilon_k\|$ is within our desired tolerance $\varepsilon$ and "bad" when the error is greater than our desired tolerance. We then construct a stochastic process $(X_k)_{k\geq 0}$ on $(\Omega^\infty, \mathcal{B}(\Omega^\infty), \mathcal{P})$ with state space $\mathcal{K} := \{1, 2, \ldots, K^*\}$ such that

$$X_{k+1} = \begin{cases} \max\{X_k - 1,\, 1\}, & \text{if iteration } k \text{ is "good"} \\ K^*, & \text{otherwise.} \end{cases}$$

The stochastic process $(X_k)_{k\geq 0}$ is easier to analyze than $(v_k)_{k\geq 0}$ because it is defined on a finite state space, however $(X_k)_{k\geq 0}$ is not necessarily a Markov chain.

We next construct a "dominating" Markov chain $(Y_k)_{k\geq 0}$ to help us analyze the behavior of $(X_k)_{k\geq 0}$. We construct $(Y_k)_{k\geq 0}$ on $(\mathcal{K}^\infty, \mathcal{B})$, the canonical measurable space of trajectories on $\mathcal{K}$, so $Y_k : \mathcal{K}^\infty \to \mathbb{R}$, and we let $\mathcal{Q}$ denote the probability measure of $(Y_k)_{k\geq 0}$ on $(\mathbb{R}^\infty, \mathcal{B})$. Since $(Y_k)_{k\geq 0}$ will be a Markov chain by construction, the probability measure $\mathcal{Q}$ is completely determined by an initial distribution on $\mathbb{R}$ and a transition kernel for $(Y_k)_{k\geq 0}$. We always initialize $Y_0 = K^*$, and then construct the transition kernel as follows:

$$Y_{k+1} = \begin{cases} \max\{Y_k - 1,\, 1\}, & \text{w.p. } q \\ K^*, & \text{w.p. } 1-q \end{cases}$$

where $q$ is the probability of a "good" iteration with respect to the corresponding norm. Note that $(Y_k)_{k\geq 0}$, we introduce here is different and has much smaller state space than the one we introduced in [10] leading to stronger convergence guarantees.

We now describe a stochastic dominance relationship between the two stochastic processes $(X_k)_{k\geq 0}$ and $(Y_k)_{k\geq 0}$. We will establish that $(Y_k)_{k\geq 0}$ is "larger" than $(X_k)_{k\geq 0}$ in a stochastic sense.

*Definition 3:* Let $X$ and $Y$ be two real-valued random variables, then $X$ is *stochastically dominated* by $Y$, written $X \leq_{st} Y$, when $\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$ for all increasing functions $f : \mathbb{R} \to \mathbb{R}$. Equivalently, $X \leq_{st} Y$ when $\Pr\{X \geq \theta\} \leq \Pr\{Y \geq \theta\}$ for all $\theta$ in the support of $Y$.

Let $\{\mathcal{F}_k\}_{k\geq 0}$ be the filtration on $(\Omega^\infty, \mathcal{B}(\Omega^\infty), \mathcal{P})$ corresponding to the evolution of information about $(X_k)_{k\geq 0}$, and let $[X_{k+1} \mid \mathcal{F}_k]$ denote the conditional distribution of $X_{k+1}$ given the information $\mathcal{F}_k$. We have the following initial results on the relationship between $(X_k)_{k\geq 0}$ and $(Y_k)_{k\geq 0}$.

The following theorem, our main result for our random operator analysis, establishes the relationship between the stochastic process $\{X_k\}_{k\geq 0}$ and the Markov chain $\{Y_k\}_{k\geq 0}$. Under Assumption 3, this result allows us to bound the stochastic process $\{X_k\}_{k\geq 0}$, which keeps track of the error in EVL with the dominating Markov chain $\{Y_k\}_{k\geq 0}$.

*Theorem 6:* Under Assumption 3

1) $X_k \leq_{st} Y_k$ for all $k \geq 0$.
2) $\Pr\{Y_k \leq \eta\} \leq \Pr\{X_k \leq \eta\}$ for any $\eta \in \mathbb{R}$ and all $k \geq 0$.

The proof is relegated to Appendix C. By Theorem 6, if $X_K \leq_{st} Y_K$ and we can make $\Pr\{Y_K \leq \eta\}$ large, then we will also obtain a meaningful bound on $\Pr\{X_K \leq \eta\}$. Following this observation, the next two corollaries are the main mechanisms for our general sample complexity results for EVL.

The following corollary follows from bounding the mixing time of the dominating Markov chain $\{Y_k\}_{k\geq 0}$ and employing our general $p-$norm error bound Lemma 4.

*Corollary 7:* For a given $p \in [1, \infty)$, and any $\varepsilon > 0$, and $\delta \in (0, 1)$, suppose Assumption 3 holds for this $\varepsilon$, and choose

any $K^* \geq 1$. Then, for $q \geq (1/2+\delta/2)^{1/(K^*-1)}$ and $K \geq \log(4/((1/2-\delta/2)(1-q)q^{K^*-1}))$, we have

$$\|v_K - v^*\|_{p,\rho} \leq 2\left(\frac{1-\gamma^{K^*+1}}{1-\gamma}\right)^{\frac{p-1}{p}} \left[C_{\rho,\mu}^{1/p}\varepsilon + \gamma^{K^*/p}(2\,v_{\max})\right]$$

with probability at least $\delta$.

The proof is relegated to Appendix C.

The next Corollary uses the same reasoning for the supremum norm case. It follows from bounding the mixing time of the dominating Markov chain $\{Y_k\}_{k\geq 0}$ and employing our general $\infty-$norm error bound Lemma 5.

*Corollary 8:* Given any $\varepsilon > 0$ and $\delta \in (0,\,1)$, suppose Assumption 3 holds for this $\varepsilon$, and choose any $K^* \geq 1$. For $q \geq (1/2+\delta/2)^{1/(K^*-1)}$ and $K \geq \log(4/((1/2-\delta/2)(1-q)q^{K^*-1}))$, we have

$$\Pr\left\{\|v_K - v^*\|_\infty \leq \varepsilon/(1-\gamma) + \gamma^{K^*}(2\,v_{\max})\right\} \geq \delta.$$

The sample complexity results for both EVL algorithms from Section III follow from Corollaries 7 and 8. This is shown next.

### B. Proofs of Theorems 1, 2, and 3

We now apply our random operator framework to both EVL algorithms. We will see that it is easy to check the conditions of Corollaries 7 and 8, from which we obtain specific sample complexity results. We will use Theorems 17, 16, and 19, which are all "one-step" results that bound the error in a single step of Algorithm 1 (in the 1- and 2-norm) and Algorithm 2 (in the $\infty$-norm) compared to the true Bellman operator.

We first give the proof of Theorem 1. We let $p(N,\,M,\,J,\,\varepsilon)$ denote the a lower bound on the probability of the event $\{\|\widehat{T}\,v - T\,v\|_{2,\mu} \leq \varepsilon\}$.

*Proof of Theorem 1:* Starting with inequality (7) for $p=2$ and using the statement of Theorem 17 in Appendix C, we have $\|v_K - v^*\|_{2,\rho}$

$$\leq\; 2\left(\frac{1}{1-\gamma}\right)^{1/2} C_{\rho,\mu}^{1/2}\left(d_{2,\mu}\left(T\,\mathcal{F}(\Theta),\,\mathcal{F}(\Theta)\right)+\varepsilon\right) + 4\left(\frac{1}{1-\gamma}\right)^{1/2} v_{\max}\gamma^{K/2}$$

when $\|\varepsilon_k\|_{2,\mu} \leq d_{2,\mu}(T\,\mathcal{F}(\Theta),\,\mathcal{F}(\Theta)) + \varepsilon$ for all $k=0,\,1,\,\ldots,K-1$. We choose $K^* \geq 1$ to satisfy

$$4\left(\frac{1}{1-\gamma}\right)^{1/2} v_{\max}\gamma^{K^*/2} \leq 2\left(\frac{1}{1-\gamma}\right)^{1/2} C_{\rho,\mu}^{1/2}\varepsilon$$

which implies $K^* = 2\lceil \frac{\ln(C_{\rho,\mu}^{1/2}\varepsilon)-\ln(2\,v_{\max})}{\ln\gamma}\rceil$. On the basis of Corollary 7, we just need to choose $N$, $M$, $J$ such that $p(N,\,M,\,J,\,\varepsilon) \geq (1-\delta/2)^{1/(K^*-1)}$. We then apply the statement of Theorem 16 with $p = 1-(1-\delta/2)^{1/(K^*-1)}$. ■

We now give the proof of Theorem 2 along the same lines as for Theorem 2. Let $p(N,\,M,\,J,\,\varepsilon)$ denote the lower bound on the probability of the event $\{\|\widehat{T}\,v - T\,v\|_{1,\mu} \leq \varepsilon\}$ for $\varepsilon > 0$. We also note that $d_{1,\mu}(T\,v,\,\mathcal{F}(\Theta)) \leq d_{1,\mu}(T\,\mathcal{F}(\Theta),\,\mathcal{F}(\Theta))$ for all $v \in \mathcal{F}(\Theta)$.

*Proof of Theorem 2:* Starting with inequality (7) for $p=1$ and using the statement of Theorem 16 in Appendix D, we have

$$\|v_K - v^*\|_{1,\rho} \leq 2\,C_{\rho,\mu}\left(d_{1,\mu}\left(T\,\mathcal{F}(\Theta),\,\mathcal{F}(\Theta)\right)+\varepsilon\right) + 4\,v_{\max}\gamma^K$$

when $\|\varepsilon_k\|_{1,\mu} \leq d_{1,\mu}(T\,\mathcal{F}(\Theta),\,\mathcal{F}(\Theta)) + \varepsilon$ for all $k=0,\,1,\,\ldots,K-1$. Choose $K^*$ such that

$$4\,v_{\max}\gamma^K \leq 2\,C_{\rho,\mu}\varepsilon \Rightarrow K^* = \left\lceil \frac{\ln(C_{\rho,\mu}\varepsilon)-\ln(2\,v_{\max})}{\ln\gamma}\right\rceil.$$

On the basis of Corollary 7, we just need to choose $N$, $M$, $J$ such that $p(N,\,M,\,J,\,\varepsilon) \geq (1-\delta/2)^{1/(K^*-1)}$. We then apply the statement of Theorem 16 with probability $1-(1-\delta/2)^{1/(K^*-1)}$. ■

We now provide proof of $\mathcal{L}_\infty$ function fitting in RKHS based on Theorem 19 in Appendix C. For this proof, we let $p(N,\,M,\,\varepsilon)$ denote a lower bound on the probability of the event $\{\|\widehat{T}\,v - T\,v\|_\infty \leq \varepsilon\}$.

*Proof of Theorem 3:* By inequality (8), we choose $\varepsilon$ and $K^* \geq 1$ such that $\varepsilon/(1-\gamma) \leq \epsilon/2$ and $\gamma^{K^*}(2\,v_{\max}) \leq \epsilon/2$ by setting

$$K^* \geq \left\lceil \frac{\ln(\epsilon)-\ln(4\,v_{\max})}{\ln(\gamma)}\right\rceil.$$

On the basis of Corollary 7, we next choose $N$ and $M$ such that $p(N,\,M,\,\varepsilon) \geq (1-\delta/2)^{1/(K^*-1)}$. We then apply the statement of Theorem 19 with error $\epsilon(1-\gamma)/2$ and probability $1-(1-\delta/2)^{1/(K^*-1)}$. ■

## V. Numerical Experiments

We now present numerical performance of our algorithm by testing it on the benchmark optimal replacement problem [9], [11]. The setting is that a product (such as a car) becomes more costly to maintain with time/miles, and must be replaced at some point. Here, the state $s_t \in \mathbb{R}_+$ represents the accumulated utilization of the product. Thus, $s_t = 0$ denotes a brand new durable good. Here, $\mathbb{A} = \{0,1\}$, so at each time step, $t$, we can either replace the product $(a_t = 0)$ or keep it $(a_t = 1)$. Replacement incurs a cost $C$, whereas keeping the product has a maintenance cost, $c(s_t)$, associated with it. The transition probabilities are as follows:

$$q(s_{t+1}|s_t,a_t) = \begin{cases} \lambda e^{-\lambda(s_{t+1}-s_t)}, & \text{if } s_{t+1} \geq s_t \text{ and } a_t = 1,\\ \lambda e^{-\lambda s_{t+1}}, & \text{if } s_{t+1} \geq 0 \text{ and } a_t = 0, \text{and}\\ 0, & \text{otherwise}\end{cases}$$

and the reward function is given by

$$r(s_t,a_t) = \begin{cases} -c(s_t), & \text{if } a_t = 1, \text{and}\\ -C - c(0), & \text{if } a_t = 0.\end{cases}$$

For our computation, we use $\gamma = 0.6, \lambda = 0.5, C = 30$, and $c(s) = 4s$. The optimal value function and the optimal policy can be computed analytically for this problem. For EVL+RPBF, we use $J$ random parameterized Fourier functions $\{\phi(s,\theta_j) = \cos(\theta_j^T s + b)\}_{j=1}^J$ with $\theta_j \sim \mathcal{N}(0, 0.01)$ and $b \sim \text{Unif}[-\pi,\pi]$. We fix J = 5. For EVL+RKHS, we use Gaussian kernel defined as $k(x,y) = \exp(||x-y||^2/(2\sigma^2))$ with $1/\sigma^2 = 0.01$ and $\mathcal{L}_2$ regularization. We fix the regularization coefficient to be $10^{-2}$. The underlying function space for FVI is polynomials of degree 4. The results are plotted after 20 iterations.

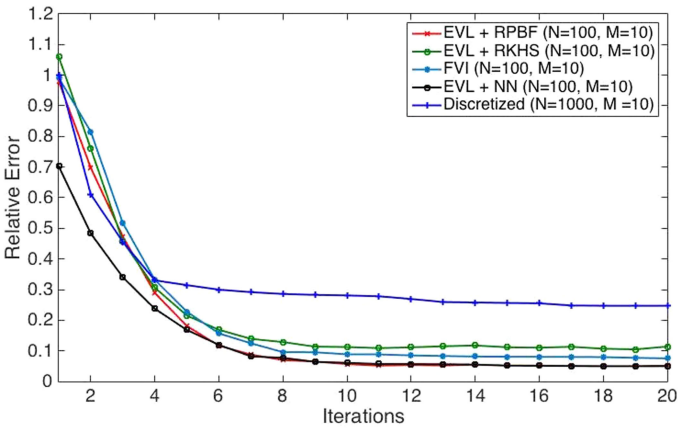



Fig. 1. Relative error with iterations for various algorithms.

The error in each iteration for different algorithms with $N = 100$ states and $M = 5$ is shown in Fig. 1. On *Y*-axis, it shows the relative error computed as $\sup_{s\in\mathbb{S}} |v^*(s) - v^{\pi_k}(s)/v^*(s)|$ with iterations $k$ on the *X*-axis. It shows that EVL+RPBF has relative error below 10% after 20 iterations. FVI is close to it but EVL+RKHS has larger relative error though it may improve with a higher $M$ or by using other kernels. This is also reflected in the actual runtime performance: EVL+RPBF takes 8705 s, FVI 8654 s, and EVL+RKHS takes 42 173 s to get within 0.1 relative error. The computational complexity of kernel methods increases quadratically with number of samples and needs a matrix inversion resulting in a slower performance.

Note that performance of FVI depends on being able to choose suitable basis functions, which for the optimal replacement problem is easy. For other problems, we may expect both EVL algorithms to perform better. So, we tested the algorithms on the cart-pole balancing problem, another benchmark problem but for which the optimal value function is unknown. We formulate it as a continuous four-dimensional state space with 2 action MDP. The state comprises of the position of the cart, $x$, velocity of the cart, $\dot{x}$, angle of the pole in radians, $\theta$, and the angular velocity of the pole, $\dot{\theta}$ . The actions are to add a force of $-10N$ or $+10N$ to the cart, pushing it left or right. We add $\pm 50\%$ noise to these actions. For system dynamics, let $m_c$ and $m_p$ be the mass of cart and pole, respectively. Let $l$ be the length of the pole. If $F_t$ is the force applied to the cart at time $t$, then acceleration of pole is

$$\ddot{\theta}_t = \frac{g\sin\theta_t + \cos\theta_t \left( \frac{-F_t - m_p l \dot{\theta}_t^{\,2} \sin\theta_t}{m_c + m_p} \right)}{l\left( \frac{4}{3} - \frac{m_p \cos^2\theta_t}{m_c + m_p} \right)}$$

and acceleration of cart is

$$\ddot{x}_t = \frac{F_t + m_p l \left( \dot{\theta}_t^{\,2} \sin\theta_t - \ddot{\theta}_t \cos\theta_t \right)}{m_c + m_p}.$$

TABLE I
RUNTIME PERFORMANCE OF VARIOUS ALGORITHMS ON THE CART-POLE PROBLEM (M=MINUTES)

| Goal | EVL+RPBF | FVI | EVL+RKHS |
|---|---|---|---|
| 50 | 5.4 m | 4.8 m | 8.7 m |
| 100 | 18.3 m | 23.7 m | 32.1 m |
| 150 | 36.7 m | 41.5 m | 54.3 m |

Now let $\tau$ be the time step for Euler's method, we have the following state transition equations:

$$x_{t+1} = x_t + \tau \dot{x}_t$$
$$\dot{x}_{t+1} = \dot{x}_t + \tau \ddot{x}_t$$
$$\theta_{t+1} = x_t + \tau \dot{\theta}_t$$
$$\dot{\theta}_{t+1} = \dot{\theta}_t + \tau \ddot{\theta}_t.$$

Rewards are zero except for failure state (if the position of cart reaches beyond $\pm 2.4$, or the pole exceeds an angle of $\pm 12$ degrees), it is $-1$. For our experiments, we choose $N = 100$ and $M = 1$. In case of RPBF, we consider parameterized Fourier basis of the form $\cos(\mathbf{w}^T\mathbf{s} + b)$ where $\mathbf{w} = [w_1, w_2]$, $w_1, w_2 \sim \mathcal{N}(0, 1)$ and $b \sim \text{Unif}[-\pi, \pi]$. We fix $J = 10$ for our EVL+RPBF. For RKHS, we consider Gaussian kernel, $K(s_1, s_2) = \exp(-\sigma||s_1 - s_2||^2/2)$ with $\sigma = 0.01$. We limit each episode to 1000 time steps. We compute the average length of the episode for which we are able to balance the pole without hitting the failure state. This is the goal in Table I. The other columns show run-time needed for the algorithms to learn to achieve such a goal.

From the table, we can see that EVL+RPBF outperforms FVI and EVL+RKHS. Note that guarantees for FVI are only available for $\mathcal{L}_2$-error and for EVL-RPBF for $\mathcal{L}_p$-error. EVL-RKHS is the only algorithm that can provide guarantees on the sup-norm error. Also note that when for problems for which the value functions are not so regular, and good basis functions are difficult to guess, the EVL+RKHS method is likely to perform better but as of now we do not have a numerical example to demonstrate this.

We also tested our algorithms on the Acrobot problem, a 2-link pendulum with only the second joint actuated. The objective is to swing the end-effector to a height, which is at least the length of one link above the base starting with both links pointing downwards. The state here is six dimensional, which are $\sin(\cdot)$ and $\cos(\cdot)$ of the two rotational joint angles and the joint angular velocities. There are three actions available: +1, 0 or $-1$, corresponding to the torque on the joint between the two pendulum links. We modify the environment available from `OpenAI` by injecting a uniform noise in the actions so that the transitions are not deterministic. The reward is 1 if the goal state is reached, else 0. We choose $N = 2000$, $M = 1$, $J = 100$. Fig. 2 represents the reward for both of the proposed algorithms. Not only does EVL+RPBF perform better, it is also faster than EVL+RKHS by an average of 3.67 min per iteration. The reason for this is that the EVL+RKHS algorithm is designed to provide guarantees on sup-error, a much more stringent requirement than the $L_p$-error that EVL+RPBF algorithm provides guarantees on.

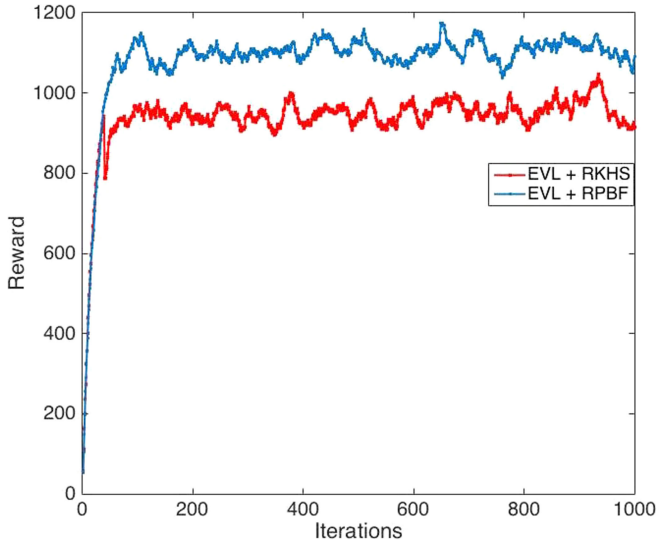



Fig. 2. Performance on the Acrobot problem.

## VI. Conclusion

In this paper, we have introduced universally applicable ADP algorithms for continuous state-space MDPs with finite action spaces. The algorithms introduced are based on using randomization to improve computational tractability and reduce the "curse of dimensionality" via the synthesis of the "random function approximation" and "empirical" approaches. Our first algorithm is based on a random parametric function fitting by sampling parameters in each iteration. The second is based on sampling states, which then yield a set of basis functions in an RKHS from the kernel. Both function fitting steps involve convex optimization problems and can be implemented with standard packages. Both algorithms can be viewed as iteration of a type of random Bellman operator followed by a random projection operator. Iterated random operators in general are difficult to analyze. Nevertheless, we can construct Markov chains that stochastically dominate the error sequences, which simplify the analysis [10]. In fact, the introduced method may be viewed as a "probabilistic contraction analysis" method in contrast to stochastic Lyapunov techniques and other methods for analyzing stochastic iterative algorithms. They yield convergence but also nonasymptotic sample complexity bounds. Numerical experiments on the cart-pole balancing and the Acrobat problems suggests good performance in practice. More rigorous numerical analysis will be conducted as part of future work.

## Appendix

### A. Supplement for Section III

The following computation shows that $T$ maps bounded functions to Lipschitz continuous functions when $Q$ and $c$ are both Lipschitz continuous in the sense of (1) and (2). Suppose $\|v\|_\infty \leq v_{\max}$, then $T\,v$ is Lipschitz continuous with constant $L_c + \gamma\, v_{\max} L_Q$. We have

$$\begin{aligned}
&|\,[T\,v]\,(s) - [T\,v]\,(s')\,| \\
&\leq \max_{a\in\mathbb{A}} |c\,(s,\,a) - c\,(s',\,a)\,| \\
&\quad + \gamma \max_{a\in\mathbb{A}} \left|\int v\,(y)\,Q\,(dy\,|\,s,\,a) - \int v\,(y)\,Q\,(dy\,|\,s',\,a)\,\right| \\
&\leq L_c \|s - s'\|_2 + \gamma\, v_{\max} \max_{a\in\mathbb{A}} \int |Q\,(dy\,|s,a) - Q\,(dy\,|\,s',a)| \\
&\leq (L_c + \gamma\, v_{\max} L_Q)\,\|s - s'\|_2 .
\end{aligned}$$

### B. Supplement for Section IV

First, we need to adapt [9, Lemma 3] to obtain point-wise error bounds on $v_K - v^*$ in terms of the errors $\{\varepsilon_k\}_{k\geq 0}$. These bounds are especially useful when analyzing the performance of EVL with respect to other norms besides the supremum norm, since $T$ does not have a contractive property with respect to any other norm.

For any $\pi \in \Pi$, we define the operator $Q^\pi : \mathcal{F}(\mathbb{S}) \to \mathcal{F}(\mathbb{S})$ (which gives the transition mapping as a function of $\pi$) via

$$(Q^\pi v)\,(s) \triangleq \int_{\mathbb{S}} v\,(y)\,Q\,(dy\,|\,s,\,\pi(s)) \qquad \forall s \in \mathbb{S}.$$

Then, we define the operator $T^\pi : \mathcal{F}(\mathbb{S}) \to \mathcal{F}(\mathbb{S})$ via

$$[T^\pi v]\,(s) \triangleq c\,(s,\,\pi(s)) + \gamma \int_{\mathbb{S}} v\,(x)\,Q\,(dx\,|\,s,\,\pi(s)) \quad \forall s \in \mathbb{S}.$$

For later use, we let $\pi^* \in \Pi$ be an optimal policy satisfying

$$\pi^*(s) \in \arg\min_{a\in\mathbb{A}} \left\{ c\,(s,\,a) + \gamma \int_{\mathbb{S}} v^*\,(x)\,Q\,(dx\,|\,s,\,a) \right\}$$

$\forall s \in \mathbb{S}$,, i.e., it is greedy with respect to $v^*$. More generally, a policy $\pi \in \Pi$ is greedy with respect to $v \in \mathcal{F}(\mathbb{S})$ if $T^\pi v = T\,v$.

For use throughout this section, we let $\pi_k$ be a greedy policy with respect to $v_k$ so that $T^{\pi_k} v_k = T\,v_k$ for all $k \geq 0$. Then, for fixed $K \geq 1$, we define the operators

$$\begin{aligned}
A_K &\triangleq \frac{1}{2}\left[\left(Q^{\pi^*}\right)^K + Q^{\pi_{K-1}} Q^{\pi_{K-2}} \cdots Q^{\pi_0}\right] \\
A_k &\triangleq \frac{1}{2}\left[\left(Q^{\pi^*}\right)^{K-k-1} + Q^{\pi_{K-1}} Q^{\pi_{K-2}} \cdots Q^{\pi_{k+1}}\right]
\end{aligned}$$

for $k = 0, \ldots, K-1$, formed by composition of transition kernels. We let $\vec{1}$ be the constant function equal to one on $\mathbb{S}$, and we define the constant $\tilde{\gamma} = \frac{2(1-\gamma^{K+1})}{1-\gamma}$ for use shortly. We note that $\{A_k\}_{k=0}^K$ are all linear operators and $A_k \vec{1} = \vec{1}$ for all $k = 0, \ldots, K$.

*Lemma 9:* For any $K \geq 1$,

1) $v_K - v^* \leq \sum_{k=0}^{K-1} \gamma^{K-k-1} (Q^{\pi^*})^{K-k-1} \varepsilon_k + \gamma^K (Q^{\pi^*})^K (v_0 - v^*)$;
2) $v_K - v^* \geq \sum_{k=0}^{K-1} \gamma^{K-k-1} (Q^{\pi_{K-1}} Q^{\pi_{K-2}} \cdots Q^{\pi_{k+1}}) \varepsilon_k + \gamma^K (Q^{\pi_{K-1}} Q^{\pi_{K-2}} \cdots Q^{\pi_0})(v_0 - v^*)$;
3) $|v_K - v^*| \leq 2[\sum_{k=0}^{K-1} \gamma^{K-k-1} A_k\,|\varepsilon_k| + \gamma^K A_K\,(2\,v_{\max})]$.

*Proof:*

1) For any $k \geq 1$, we have $T\,v_k \leq T^{\pi^*} v_k$ and $T^{\pi^*} v_k - T^{\pi^*} v^* = \gamma\, Q^{\pi^*}(v_k - v^*)$, so $v_{k+1} - v^* =$

$$\begin{aligned}
&T\,v_k + \varepsilon_k - T^{\pi^*} v_k + T^{\pi^*} v_k - T^{\pi^*} v^* \\
&\leq \gamma\, Q^{\pi^*}\,(v_k - v^*) + \varepsilon_k .
\end{aligned}$$

The result then follows by induction.

2) Similarly, for any $k \geq 1$, we have $T\, v^* \leq T^{\pi_k} v^*$ and $T\, v_k - T^{\pi_k} v^* = T^{\pi_k} v_k - T^{\pi_k} v^* = \gamma\, Q^{\pi_k}(v_k - v^*)$, so $v_{k+1} - v^* =$

$$T\, v_k + \varepsilon_k - T^{\pi_k} v^* + T^{\pi_k} v^* - T\, v^* \geq \gamma\, Q^{\pi_k}\,(v_k - v^*) + \varepsilon_k.$$

Again, the result follows by induction.

3) If $f \leq g \leq h$ in $\mathcal{F}(\mathbb{S})$, then $|g| \leq |f| + |h|$, so combining parts 1) and 2) gives

$$|v_K - v^*| \leq 2 \sum_{k=0}^{K-1} \gamma^{K-k-1} A_k |\varepsilon_k| + 2\,\gamma^K A_K |v_0 - v^*|.$$

Then, we note that $|v_0 - v^*| \leq 2\, v_{\max}$. ■

Now we use Lemma 9 to derive $p-$norm bounds.

*Proof of Lemma 4:* Using $\sum_{k=0}^{K} \gamma^k = (1-\gamma^{K+1})/(1-\gamma)$, we define the constants

$$\alpha_k = \frac{(1-\gamma)\,\gamma^{K-k-1}}{1-\gamma^{K+1}} \quad \forall k = 0, \ldots, K-1$$

$$\alpha_K = \frac{(1-\gamma)\,\gamma^K}{1-\gamma^{K+1}}$$

and we note that $\sum_{k=0}^{K} \alpha_k = 1$. Then, we obtain $|v_K - v^*|$

$$\leq \tilde{\gamma} \left[ \sum_{k=0}^{K-1} \alpha_k A_k\, |\varepsilon_k| + \alpha_K A_K\, (2\, v_{\max}) \right]$$

from Lemma 9 3). Next, we compute $\|v_K - v^*\|_{p,\,\rho}^p =$

$$\int_{\mathbb{S}} |v_K(s) - v^*(s)|^p \rho\,(ds)$$

$$\leq \tilde{\gamma}^p \int_{\mathbb{S}} \left[ \sum_{k=0}^{K-1} \alpha_k A_k\, |\varepsilon_k| + \alpha_K A_K\, (2\, v_{\max})\, \vec{1} \right]^p (s) \rho\,(ds)$$

$$\leq \tilde{\gamma}^p \int_{\mathbb{S}} \left[ \sum_{k=0}^{K-1} \alpha_k A_k\, |\varepsilon_k|^p + \alpha_K A_K\, (2\, v_{\max})^p\, \vec{1} \right] (s) \rho\,(ds)$$

using Jensen's inequality and convexity of $x \to |x|^p$. Now, we have $\rho\, A_k \leq c_{\rho,\,\mu}(K-k-1)\mu$ for $k = 0, \ldots, K-1$ by Assumption 1 2) and so for all $k = 0, \ldots, K-1$

$$\int_{\mathbb{S}} [A_k\, |\varepsilon_k|^p]\,(s) \rho\,(ds) \leq c_{\rho,\,\mu}\,(K-k-1)\, \|\varepsilon_k\|_{p,\,\mu}^p.$$

We arrive at $\|v_K - v^*\|_{p,\,\rho}^p$

$$\leq\ \tilde{\gamma}^p \left[ \sum_{k=0}^{K-1} \alpha_k c_{\rho,\,\mu}\,(K-k-1)\, \|\varepsilon_k\|_{p,\,\mu}^p + \alpha_K\, (2\, v_{\max})^p \right]$$

$$= 2^p \tilde{\gamma}^{p-1} \left[ \sum_{k=0}^{K-1} \gamma^{K-k-1} c_{\rho,\,\mu}\,(K-k-1)\, \|\varepsilon_k\|_{p,\,\mu}^p + \gamma^K\, (2\, v_{\max})^p \right]$$

where we use $|v_0 - v^*|^p \leq (2\, v_{\max})^p$. Now, by subadditivity of $x \to |x|^t$ for $t = 1/p \in (0,\, 1]$ with $p \in [1,\, \infty)$, assumption that $\|\varepsilon_k\|_{p,\,\mu} \leq \varepsilon$ for all $k = 0,\, 1, \ldots, K-1$, and since $\sum_{k=0}^{K-1} \gamma^{K-k-1} c_{\rho,\,\mu}(K-k-1) \leq C_{\rho,\,\mu}$ by Assumption 1 2), we see

$$\|v_K - v^*\|_{p,\,\rho} \leq 2 \left( \frac{1-\gamma^{K+1}}{1-\gamma} \right)^{\frac{p-1}{p}} \left[ C_{\rho,\,\mu}^{1/p} \varepsilon + \gamma^{K/p}\, (2\, v_{\max}) \right]$$

which gives the desired result. ■

Supremum norm error bounds follow more easily from Lemma 9.

*Proof of Lemma 5:* We have

$$\|v_K - v^*\|_\infty \leq \max \left\{ \left\| \sum_{k=0}^{K-1} \gamma^{K-k-1} \left(Q^{\pi^*}\right)^{K-k-1} \varepsilon_k + \gamma^K \left(Q^{\pi^*}\right)^K (v_0 - v^*) \right\|_\infty \right.$$

$$\left. \left\| \sum_{k=0}^{K-1} \gamma^{K-k-1} \left(Q^{\pi_{K-1}} Q^{\pi_{K-2}} \cdots Q^{\pi_{K+1}}\right) \varepsilon_k + \gamma^K \left(Q^{\pi_{K-1}} Q^{\pi_{K-2}} \cdots Q^{\pi_0}\right) (v_0 - v^*) \right\|_\infty \right\}$$

by Lemma 9. Now

$$\left\| \sum_{k=0}^{K-1} \gamma^{K-k-1} \left(Q^{\pi^*}\right)^{K-k-1} \varepsilon_k + \gamma^K \left(Q^{\pi^*}\right)^K (v_0 - v^*) \right\|_\infty \leq \sum_{k=0}^{K-1} \gamma^{K-k-1} \|\varepsilon_k\|_\infty + \gamma^K \|v_0 - v^*\|_\infty$$

and

$$\left\| \sum_{k=0}^{K-1} \gamma^{K-k-1} \left(Q^{\pi_{K-1}} Q^{\pi_{K-2}} \cdots Q^{\pi_{K+1}}\right) \varepsilon_k + \gamma^K \left(Q^{\pi_{K-1}} Q^{\pi_{K-2}} \cdots Q^{\pi_0}\right) (v_0 - v^*) \right\|_\infty \leq \sum_{k=0}^{K-1} \gamma^{K-k-1} \|\varepsilon_k\|_\infty + \gamma^K \|v_0 - v^*\|_\infty$$

where we use the triangle inequality, the fact that $|(Q\, f)(s)| \leq \int_{\mathbb{S}} |f(y)| Q(dy \,|\, s) \leq \|f\|_\infty$ for any transition kernel $Q$ on $\mathbb{S}$ and $f \in \mathcal{F}(\mathbb{S})$, and $|v_0 - v^*| \leq 2\, v_{\max}$. For any $K \geq 1$

$$\|v_K - v^*\|_\infty \leq \sum_{k=0}^{K-1} \gamma^{K-k-1} \|\varepsilon_k\|_\infty + \gamma^K\, (2\, v_{\max}) \quad (9)$$

follows immediately since $\sum_{k=0}^{K-1} \gamma^{K-k-1} \varepsilon \leq \varepsilon/(1-\gamma)$ for all $K \geq 1$. ■

We emphasize that Lemma 5 does not require any assumptions on the transition probabilities, in contrast to Lemma 4, which requires Assumption 1 2).

### C. Supplement for Section IV: Function Approximation

We record several pertinent results here on the type of function reconstruction used in our EVL algorithms. The first lemma is illustrative of approximation results in Hilbert spaces, it gives

an $O(1/\sqrt{J})$ convergence rate on the error from using $\widehat{\mathcal{F}}(\theta^{1:J})$ compared to $\mathcal{F}(\Theta)$ in $\mathcal{L}_{2,\mu}(\mathbb{S})$ in probability.

*Lemma 10 (see[26, Lemma 1]):* Fix $f^* \in \mathcal{F}(\Theta)$, for any $\delta \in (0,\,1)$, there exists a function $\hat{f} \in \widehat{\mathcal{F}}(\theta^{1:J})$ such that

$$\|f^* - \hat{f}\|_{2,\mu} \leq \frac{C}{\sqrt{J}}\left(1 + \sqrt{2\,\log\frac{1}{\delta}}\right)$$

with probability at least $1-\delta$.

The next result is an easy consequence of [26, Lemma 1] and bounds the error from using $\widehat{\mathcal{F}}(\theta^{1:J})$ compared to $\mathcal{F}(\Theta)$ in $\mathcal{L}_{1,\mu}(\mathbb{S})$.

*Lemma 11:* Fix $f^* \in \mathcal{F}(\Theta)$, for any $\delta \in (0,\,1)$, there exists a function $\hat{f} \in \widehat{\mathcal{F}}(\theta^{1:J})$ such that

$$\|\hat{f} - f^*\|_{1,\mu} \leq \frac{C}{\sqrt{J}}\left(1 + \sqrt{2\,\log\frac{1}{\delta}}\right)$$

with probability at least $1-\delta$.

*Proof:* Choose $f,\,g \in \mathcal{F}(\mathbb{S})$, then by Jensen's inequality, we have $\|f-g\|_{1,\mu} = \mathbb{E}_\mu[|f(S)-g(S)|]$

$$= \mathbb{E}_\mu\left[\left(\left(f(s)-g(s)\right)^2\right)^{1/2}\right] \leq \sqrt{\mathbb{E}_\mu\left[\left(f(s)-g(s)\right)^2\right]}.$$

The desired result then follows by [26, Lemma 1]. ■

Now we consider function approximation in the supremum norm. Recall the definition of the regression function

$$f_M(s) = \mathbb{E}\left[\min_{a\in\mathbb{A}}\left\{c(s,\,a) + \frac{\gamma}{M}\sum_{m=1}^{M} v\left(X_m^{s,\,a}\right)\right\}\right]$$

$\forall s \in \mathbb{S}$. Then, we have the following approximation result, for which we recall the constant $\kappa := \sup_{s\in\mathbb{S}}\sqrt{K(s,\,s)}$.

*Corollary 12 (see[28, Corollary 5]):* For any $\delta \in (0,\,1)$

$$\|f_{z,\lambda} - f_M\|_{\mathcal{H}_K}$$
$$\leq \tilde{C}\,\kappa\left(\frac{\log(4/\delta)^2}{N}\right)^{1/6} \text{ for } \lambda = \left(\frac{\log(4/\delta)^2}{N}\right)^{1/3}$$

with probability at least $1-\delta$.

*Proof:* Uses the fact that for any $f \in \mathcal{H}_K$, $\|f\|_\infty \leq \kappa\,\|f\|_{\mathcal{H}_K}$. For any $s \in \mathbb{S}$, we have $|f(s)| = |\langle K(s,\,\cdot),\,f(\cdot)\rangle_{\mathcal{H}_K}|$ and subsequently

$$\begin{aligned}
|\langle K(s,\,\cdot),\,f(\cdot)\rangle_{\mathcal{H}_K}| &\leq \|K(s,\,\cdot)\|_{\mathcal{H}_K}\|f\|_{\mathcal{H}_K}\\
&= \sqrt{\langle K(s,\,\cdot),\,K(s,\,\cdot)\rangle_{\mathcal{H}_K}}\|f\|_{\mathcal{H}_K}\\
&= \sqrt{K(s,\,s)}\|f\|_{\mathcal{H}_K}\\
&\leq \sup_{s\in\mathbb{S}}\sqrt{K(s,\,s)}\|f\|_{\mathcal{H}_K}
\end{aligned}$$

where the first inequality is by Cauchy–Schwartz and the second is by assumption that $K$ is a bounded kernel. ■

The preceding result is about the error when approximating the regression function $f_M$, but $f_M$ generally is not equal to $T\,v$. We bound the error between $f_M$ and $T\,v$ as well in the next section.

*Proof:* (Theorem 6) First we note that, by [10, Lemma A.1], $[Y_{k+1}\,|\,Y_k = \eta]$ is stochastically increasing in $\eta$ for all $k \geq 0$, i.e., $[Y_{k+1}\,|\,Y_k = \eta] \leq_{st} [Y_{k+1}\,|\,Y_k = \eta']$ for all $\eta \leq \eta'$. Then, by [10, Lemma A.2], $[X_{k+1}\,|\,X_k = \eta,\,\mathcal{F}_k] \leq_{st} [Y_{k+1}\,|\,Y_k = \eta]$ for all $\eta \in f$ and $\mathcal{F}_k$ for all $k \geq 0$.

1) Trivially, $X_0 \leq_{st} Y_0$ since $X_0 \leq_{as} Y_0$. Next, we see that $X_1 \leq_{st} Y_1$ by [10, Lemma A.1]. We prove the general case by induction. Suppose $X_k \leq_{st} Y_k$ for $k \geq 1$, and for this proof define the random variable

$$\mathcal{Y}(\theta) = \begin{cases}\max\{\theta-1,\,1\}, & \text{w.p. } q\\ K^*, & \text{w.p. } 1-q\end{cases}$$

to be the conditional distribution of $Y_k$ conditional on $\theta$, as a function of $\theta$. We see that $Y_{k+1}$ has the same distribution as $[\mathcal{Y}(\theta)\,|\,\theta = Y_k]$ by definition. Since $\mathcal{Y}(\theta)$ are stochastically increasing by Lemma [10, Lemma A.1], we see that $[\mathcal{Y}(\theta)\,|\,\theta = Y_k] \geq_{st} [\mathcal{Y}(\theta)\,|\,\theta = X_k]$ by [29, Th. 1.A.6] and our induction hypothesis. Now, $[\mathcal{Y}(\theta)\,|\,\theta = X_k] \geq_{st} [X_{k+1}\,|\,X_k,\,\mathcal{F}_k]$ by [29, Th. 1.A.3(d)] and Lemma [10, Lemma A.2] for all histories $\mathcal{F}_k$. It follows that $Y_{k+1} \geq_{st} X_{k+1}$ by transitivity of $\geq_{st}$.

2) Follows from part 1) by the definition of $\leq_{st}$. ■

*Proof:* (Corollary 7) Since $(Y_k)_{k\geq 0}$ is an irreducible Markov chain on a finite state space, its steady-state distribution $\mu = (\mu(i))_{i=1}^{K^*}$ on $\mathcal{K}$ exists. By [10, Lemma 4.3], the steady-state distribution of $(Y_k)_{k\geq 0}$ is $\mu = (\mu(i))_{i=1}^{K^*}$ given by

$$\begin{aligned}
\mu(1) &= q^{K^*-1}\\
\mu(i) &= (1-q)\,q^{K^*-i}, \qquad \forall i = 2,\ldots,K^*-1\\
\mu(K^*) &= 1-q.
\end{aligned}$$

The constant

$$\mu_{\min}(q;\,K^*) := \min\left\{q^{K^*-1},\,(1-q)\,q^{(K^*-2)},\,(1-q)\right\}$$

for all $q \in (0,\,1)$ and $K^* \geq 1$, which is the minimum of the steady-state probabilities appears shortly in the Markov chain mixing time bound for $(Y_k)_{k\geq 0}$. We note that $\mu^*(q;\,K^*) = (1-q)q^{K^*-1} \leq \mu_{\min}(q;\,K^*)$ is a simple lower bound for $\mu_{\min}(q;\,K^*)$ (we defined $\mu^*(q;\,K^*) = (1-q)q^{K^*-1}$ earlier).

Now, recall that $\|\mu-\nu\|_{TV} = \frac{1}{2}\sum_{\eta=1}^{K^*}|\mu(\eta)-\nu(\eta)|$ is the total variation distance for probability distributions on $\mathcal{K}$. Let $Q^k$ be the marginal distribution of $Y_k$ for $k \geq 0$. By a Markov chain mixing time argument, e.g., [30, Th. 12.3], we have that

$$\begin{aligned}
t_{\text{mix}}(\delta') &:= \min\left\{k \geq 0 : \|Q^k - \mu\|_{TV} \leq \delta'\right\}\\
&\leq \log\left(\frac{1}{\delta'\mu_{\min}(q;\,K^*)}\right)\\
&\leq \log\left(\frac{1}{\delta'(1-q)\,q^{K^*-1}}\right)
\end{aligned}$$

for any $\delta' \in (0,\,1)$. So, for $K \geq \log(1/(\delta'(1-q)q^{K^*-1}))$, we have $|\Pr\{Y_K = 1\} - \mu(1)| =$

$$|\Pr\{Y_K = 1\} - q^{K^*-1}| \leq 2\,\|Q^K - \mu\|_{TV} \leq 2\,\delta'$$

where we use $\mu(1) = q^{K^*-1}$. By Theorem 6, $\Pr\{X_K = 1\} \geq \Pr\{Y_K = 1\}$ and so

$$\Pr\{X_K = 1\} \geq q^{K^*-1} - 2\,\delta'.$$

Choose $q$ and $\delta'$ to satisfy $q^{K^*-1} = 1/2 + \delta/2$ and $2\,\delta' = q^{K^*-1} - \delta = 1/2 - \delta/2$ to get $q^{K^*-1} - 2\,\delta' \geq \delta$, and the desired result follows. ■

### D. Bellman Error

The layout of this section is modeled after the arguments in [9], but with the added consideration of randomized function fitting. We use the following easy-to-establish fact.

*Fact 13:* Let $X$ be a given set, and $f_1 : X \to \mathbb{R}$ and $f_2 : X \to \mathbb{R}$ be two real-valued functions on $X$. Then,

1) $|\inf_{x\in X} f_1(x) - \inf_{x\in X} f_2(x)| \leq \sup_{x\in X} |f_1(x) - f_2(x)|$, and
2) $|\sup_{x\in X} f_1(x) - \sup_{x\in X} f_2(x)| \leq \sup_{x\in X} |f_1(x) - f_2(x)|$.

For example, Fact 13 can be used to show that $T$ is contractive in the supremum norm.

The next result is about $\widehat{T}$, it uses Hoeffding's inequality to bound the estimation error between $\{\tilde{v}(s_n)\}_{n=1}^N$ and $\{[T\,v](s_n)\}_{n=1}^N$ in probability.

*Lemma 14:* For any $p \in [1, \infty]$, $f, v \in \mathcal{F}(\mathbb{S};\, v_{\max})$, and $\varepsilon > 0$

$$\Pr\{|\|f - T\,v\|_{p,\,\hat{\mu}} - \|f - \tilde{v}\|_{p,\,\hat{\mu}}| > \varepsilon\} \leq 2\,N\,|\mathbb{A}|\,\exp\left(\frac{-2\,M\,\varepsilon^2}{v_{\max}^2}\right).$$

*Proof:* First we have $|\|f - T\,v\|_{p,\,\hat{\mu}} - \|f - \tilde{v}\|_{p,\,\hat{\mu}}| \leq \|T\,v - \tilde{v}\|_{p,\,\hat{\mu}}$ by the reverse triangle inequality. Then, for any $s \in \mathbb{S}$, we have $|[T\,v](s) - \tilde{v}(s)| =$

$$\begin{aligned}&\max_{a\in\mathbb{A}} \left| \left\{ c\,(s,\,a) + \gamma \int_{\mathbb{S}} v\,(x)\,Q\,(dx \,|\, s,\,a) \right\} \right. \\ &\qquad \left. - \left\{ c\,(s,\,a) + \frac{\gamma}{M} \sum_{m=1}^{M} v\,(X_m^{s,\,a}) \right\} \right| \\ \leq\ &\gamma \max_{a\in\mathbb{A}} \left| \int_{\mathbb{S}} v\,(x)\,Q\,(dx \,|\, s,\,a) - \frac{1}{M} \sum_{m=1}^{M} v\,(X_m^{s,\,a}) \right|\end{aligned}$$

by Fact 13. We may also take $v(s) \in [0,\, v_{\max}]$ for all $s \in \mathbb{S}$ by assumption on the cost function, so by the Hoeffding inequality and the union bound, we obtain

$$\Pr\left\{ \max_{n=1,\ldots,N} |[T\,v]\,(s_n) - \tilde{v}\,(s_n)| \geq \varepsilon \right\} \leq 2\,N\,|\mathbb{A}|\,\exp\left(\frac{-2\,M\,\varepsilon^2}{v_{\max^2}}\right)$$

and thus

$$\begin{aligned}&\Pr\{\|T\,v - \tilde{v}\|_{p,\,\hat{\mu}} \geq \varepsilon\} \\ &= \Pr\left\{ \left( \frac{1}{N} \sum_{n=1}^{N} |[T\,v]\,(s_n) - \tilde{v}\,(s_n)|^p \right)^{1/p} \geq \varepsilon \right\} \\ &\leq \Pr\left\{ \max_{n=1,\ldots,N} |[T\,v]\,(s_n) - \tilde{v}\,(s_n)| \geq \varepsilon \right\}\end{aligned}$$

which gives the desired result. ■

To continue, we introduce the following additional notation corresponding to a set of functions $\mathcal{F} \subset \mathcal{F}(\mathbb{S})$:

1) $\mathcal{F}(s^{1:N}) \triangleq \{(f(s_1), \ldots, f(s_N)) : f \in \mathcal{F}\}$
2) $\mathcal{N}(\varepsilon,\, \mathcal{F}(s^{1:N}))$ is the $\varepsilon$-covering number of $\mathcal{F}(s^{1:N})$ with respect to the 1-norm on $\mathbb{R}^N$.

The next lemma uniformly bounds the estimation error between the true expectation and the empirical expectation over the set $\widehat{\mathcal{F}}(\theta^{1:J})$ (in the following statement, $e$ is Euler's number).

*Lemma 15:* For any $\varepsilon > 0$ and $N \geq 1$

$$\Pr\left\{ \sup_{f\in\widehat{\mathcal{F}}(\theta^{1:J})} \left| \frac{1}{N} \sum_{n=1}^{N} f\,(S_n) - \mathbb{E}_\mu\,[f(s)] \right| > \varepsilon \right\} \leq 8\,e\,(J+1) \left( \frac{2\,e\,v_{\max}}{\varepsilon} \right)^J \exp\left( \frac{-N\,\varepsilon^2}{128\,v_{\max}^2} \right).$$

*Proof:* For any $\mathcal{F} \subset \mathcal{F}(\mathbb{S};\, v_{\max})$, $\varepsilon > 0$, and $N \geq 1$, we have

$$\Pr\left\{ \sup_{f\in\widehat{\mathcal{F}}(\theta^{1:J})} \left| \frac{1}{N} \sum_{n=1}^{N} f\,(S_n) - \mathbb{E}_\mu\,[f(s)] \right| > \varepsilon \right\} \leq 8\,\mathbb{E}\left[ \mathcal{N}\left( \varepsilon/8,\, \widehat{\mathcal{F}}\left(\theta^{1:J}\right)\left(s^{1:N}\right) \right) \right] \exp\left( \frac{-N\,\varepsilon^2}{128\,v_{\max}^2} \right).$$

It remains to bound $\mathbb{E}[\mathcal{N}(\varepsilon/8,\, \widehat{\mathcal{F}}(\theta^{1:J})(s^{1:N}))]$. We note that $\widehat{\mathcal{F}}(\theta^{1:J})$ is a subset of

$$\left\{ f(\cdot) = \sum_{j=1}^{J} \alpha_j \phi\,(\cdot;\, \theta_j) \,:\, (\alpha_1, \ldots, \alpha_J) \in \mathbb{R}^J \right\}$$

which is a vector space with dimension $J$. By [31, Corollary 11.5], the pseudo-dimension of $\widehat{\mathcal{F}}(\theta^{1:J})$ is bounded above by $J$. Furthermore

$$\mathcal{N}\left( \varepsilon,\, \widehat{\mathcal{F}}\left(\theta^{1:J}\right)\left(s^{1:N}\right) \right) \leq e\,(J+1) \left( \frac{2\,e\,v_{\max}}{\varepsilon} \right)^J$$

by [32, Corollary 3], which gives the desired result. ■

To continue, we let $v' = v'(v,\, N,\, M,\, J,\, \mu,\, \nu)$ denote the (random) output of one iteration of EVL applied to $v \in \mathcal{F}(\mathbb{S})$ as a function of the parameters $N$, $M$, $J \geq 1$ and the probability distributions $\mu$ and $\nu$. The next theorem bounds the error between $T\,v$ and $v'$ in one iteration of EVL with respect to $\mathcal{L}_{1,\,\mu}(\mathbb{S})$, it is a direct adaptation of [9, Lemma 1] modified to account for the randomized function fitting and the effective function space being $\mathcal{F}(\theta^{1:J})$.

*Theorem 16:* Choose $v \in \mathcal{F}(\mathbb{S};\, v_{\max})$, $\varepsilon > 0$, and $\delta \in (0, 1)$. Also choose $J \geq [\frac{5C}{\varepsilon}(1 + \sqrt{2\,\log\frac{5}{\delta}})]^2$, $N \geq 2^7 5^2 \bar{v}_{\max}^2 \log$

$[\frac{40\,e(J+1)}{\delta}(10\,e\,\bar{v}_{\max})^J]$, and $M \geq (\frac{v_{\max}^2}{2\,\varepsilon^2})\log[\frac{10\,N\,|\mathbb{A}|}{\delta}]$. Then, for $v' = v'(v,\, N,\, M,\, J,\, \mu,\, \nu)$, we have $\|v' - T\,v\|_{1,\,\mu} \leq d_{1,\,\mu}(T\,v,\, \mathcal{F}(\Theta)) + \varepsilon$ with probability at least $1-\delta$.

*Proof:* Let $\varepsilon' > 0$ be arbitrary and choose $f^* \in \mathcal{F}(\Theta)$ such that $\|f^* - T\,v\|_{1,\,\mu} \leq \inf_{f\in\mathcal{F}(\Theta)} \|f - T\,v\|_{1,\,\mu} + \varepsilon'$. Then, choose $\hat{f} \in \widehat{\mathcal{F}}(\theta^{1:J})$ such that $\|\hat{f} - T\,v\|_{1,\,\mu} \leq \|f^* - T\,v\|_{1,\,\mu} + \varepsilon/5$ with probability at least $1 - \delta/5$ by Lemma 11 by choosing $J \geq 1$ to satisfy

$$\frac{C}{\sqrt{J}}\left(1 + \sqrt{2\,\log\frac{1}{(\delta/5)}}\right) \leq \frac{\varepsilon}{5}$$

$$\Rightarrow J \geq \left[\left(\frac{5\,C}{\varepsilon}\right)\left(1 + \sqrt{2\,\log\frac{5}{\delta}}\right)\right]^2.$$

Now consider the inequalities

$$\begin{aligned}
\|v' - T\,v\|_{1,\,\mu} &\leq \|v' - T\,v\|_{1,\,\hat{\mu}} + \varepsilon/5 && (10)\\
&\leq \|v' - \tilde{v}\|_{1,\,\hat{\mu}} + 2\,\varepsilon/5 && (11)\\
&\leq \|\hat{f} - \tilde{v}\|_{1,\,\hat{\mu}} + 2\,\varepsilon/5 && (12)\\
&\leq \|\hat{f} - T\,v\|_{1,\,\hat{\mu}} + 3\,\varepsilon/5 && (13)\\
&\leq \|\hat{f} - T\,v\|_{1,\,\mu} + 4\,\varepsilon/5 && (14)\\
&\leq \|f^* - T\,v\|_{1,\,\mu} + \varepsilon && (15)\\
&\leq d_{1,\,\mu}\left(T\,v,\, \mathcal{F}\left(\Theta\right)\right) + \varepsilon + \varepsilon'. && (16)
\end{aligned}$$

First, note that inequality (12) is immediate since $\|v' - \tilde{v}\|_{1,\,\hat{\mu}} \leq \|f - \tilde{v}\|_{1,\,\hat{\mu}}$ for all $f \in \widehat{\mathcal{F}}(\theta^{1:J})$ by the choice of $v'$ as the minimizer in Step 3 of Algorithm 1. Second, inequalities (10) and (14) follow from Lemma 15 by choosing $N \geq 1$ to satisfy

$$8\,e\,(J+1)\left(\frac{2\,e\,v_{\max}}{\varepsilon/5}\right)^J \exp\left(\frac{-N\,(\varepsilon/5)^2}{128\,v_{\max}^2}\right) \leq \frac{\delta}{5}$$

$$\Rightarrow N \geq 2^7 5^2 \bar{v}_{\max}^2 \log\left[\frac{40\,e\,(J+1)}{\delta}\,(10\,e\,\bar{v}_{\max})^J\right].$$

Third, inequality (16) follows from the choice of $f^* \in \mathcal{F}$. Finally, inequalities (11) and (13) follow from Lemma 14 by choosing $M \geq 1$ to satisfy

$$2\,N\,|\mathbb{A}|\,\exp\left(\frac{-2\,M\,\varepsilon^2}{{v_{\max}}^2}\right) \leq \frac{\delta}{5}$$

$$\Rightarrow M \geq \left(\frac{v_{\max}^2}{2\,\varepsilon^2}\right)\log\left[\frac{10\,N\,|\mathbb{A}|}{\delta}\right].$$

Since $\varepsilon'$ was arbitrary, the desired result then follows by the union bound. ■

Using similar steps as Theorem 16, the next theorem bounds the error in one iteration of EVL with respect to $\mathcal{L}_{2,\,\mu}(\mathbb{S})$.

*Theorem 17:* Choose $v \in \mathcal{F}(\mathbb{S}; v_{\max})$, $\varepsilon > 0$, and $\delta \in (0,1)$. Also choose $J \geq [\frac{5\,C}{\varepsilon}(1 + \sqrt{2\,\log\frac{5}{\delta}})]^2$, $N \geq 2^7 5^2 \bar{v}_{\max}^4 \log[\frac{40\,e(J+1)}{\delta}(10\,e\,\bar{v}_{\max})^J]$, and $M \geq (\frac{v_{\max}^2}{2\,\varepsilon^2})\log[\frac{10\,N\,|\mathbb{A}|}{\delta}]$. Then, for $v' = v'(v,\, N,\, M,\, J,\, \mu,\, \nu)$, we have $\|v' - T\,v\|_{2,\,\mu} \leq d_{2,\,\mu}(T\,v,\, \mathcal{F}(\Theta)) + \varepsilon$ with probability at least $1-\delta$.

In the next lemma, we show that we can make the bias between the regression function $f_M$ and the Bellman update $T\,v$ arbitrarily small uniformly over $s \in \mathbb{S}$ through the choice of $M \geq 1$.

*Lemma 18:* For any $\varepsilon > 0$ and $M \geq 1$

$$\|f_M - T\,v\|_\infty \leq \gamma\left[\varepsilon + 2\,|\mathbb{A}|\,\exp\left(\frac{-2\,M\,\varepsilon^2}{{v_{\max}}^2}\right)(v_{\max} - \varepsilon)\right].$$

*Proof:* For any $s \in \mathbb{S}$, we compute

$$\begin{aligned}
&|f_M(s) - [T\,v](s)|\\
&\leq \mathbb{E}\left[\left|\min_{a\in\mathbb{A}}\left\{c(s,\,a) + \frac{\gamma}{M}\sum_{m=1}^{M} v\left(X_m^{s,\,a}\right)\right\}\right.\right.\\
&\quad\left.\left. - \min_{a\in\mathbb{A}}\left\{c(s,\,a) + \gamma\,\mathbb{E}_{X\sim Q(\cdot\,|\,s,\,a)}\left[v(X)\right]\right\}\right|\right]\\
&\leq \gamma\,\mathbb{E}\left[\max_{a\in\mathbb{A}}\left|\frac{1}{M}\sum_{m=1}^{M} v\left(X_m^{s,\,a}\right) - \mathbb{E}_{X\sim Q(\cdot\,|\,s,\,a)}\left[v(X)\right]\right|\right]\\
&\leq \gamma\left[\varepsilon + 2\,|\mathbb{A}|\,\exp\left(\frac{-2\,M\,\varepsilon^2}{{v_{\max}}^2}\right)(v_{\max} - \varepsilon)\right]
\end{aligned}$$

where the second inequality follows from Fact 13 and the third is by the Hoeffding inequality. ■

We make use of the following RKHS function fitting result for the one step Bellman error in the supremum norm.

*Theorem 19:* Fix $v \in \mathcal{F}(\mathbb{S}; v_{\max})$, $\varepsilon > 0$, and $\delta \in (0,1)$. Also choose $N \geq (\frac{2\,C_K\,\kappa}{\varepsilon})^6 \log(4/\delta)^2$ and $M \geq \frac{v_{\max}^2}{2(\varepsilon/4)^2}\log(\frac{4\,|\mathbb{A}|\,\gamma(v_{\max}-\varepsilon/4)}{(4-2\,\gamma)\varepsilon})$, where $C_K$ is a constant independent of the dimension of $\mathbb{S}$. Then, for

$$\hat{f}_\lambda \triangleq \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{N}\sum_{n=1}^{N}\left(f\left(S_n\right) - Y_n\right)^2 + \lambda\,\|f\|_{\mathcal{H}_K}^2\right\}$$

we have $\|\hat{f}_\lambda - T\,v\|_\infty \leq \varepsilon$ with probability at least $1-\delta$.

*Proof:* By the triangle inequality, $\|\hat{f}_\lambda - T\,v\|_\infty \leq \|\hat{f}_\lambda - f_M\|_\infty + \|f_M - T\,v\|_\infty$. We choose $N \geq 1$ to satisfy

$$C_K\,\kappa\left(\frac{\log\left(4/\delta\right)^2}{N}\right)^{1/6} \leq \frac{\varepsilon}{2} \Rightarrow N \geq \left(\frac{2\,C_K\,\kappa}{\varepsilon}\right)^6 \log\left(4/\delta\right)^2$$

so that $\|\hat{f}_\lambda - f_M\|_\infty \leq \varepsilon/2$ with probability at least $1-\delta$ by [28, Corollary 5] and the fact that $\|f\|_\infty \leq \kappa\,\|f\|_{\mathcal{H}_K}$. Then, we choose $M \geq 1$ to satisfy

$$\gamma\left[\varepsilon/4 + 2\,|\mathbb{A}|\,\exp\left(\frac{-2\,M\,(\varepsilon/4)^2}{v_{\max}^2}\right)(v_{\max} - \varepsilon/4)\right] \leq \frac{\varepsilon}{2}$$

$$\Rightarrow M \geq \frac{v_{\max}^2}{2\,(\varepsilon/4)^2}\log\left(\frac{4\,|\mathbb{A}|\,\gamma\,(v_{\max} - \varepsilon/4)}{(4-2\,\gamma)\,\varepsilon}\right)$$

so that $\|f_M - T\,v\|_\infty \leq \varepsilon/2$ by Lemma 18. ■

## REFERENCES


[1] W. B. Haskell, P. Yu, H. Sharma, and R. Jain, "Randomized function fitting-based empirical value iteration," in *Proc. IEEE 56th Annu. Conf. Decision Control*, 2017, pp. 2467–2472.

[2] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed, vol. 2. Belmont, MA, USA: Athena Sci., 2011.

[3] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, vol. 703. New York, NY, USA: Wiley, 2007.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[5] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.
[6] D. P. Bertsekas, "Dynamic programming and suboptimal control: A survey from adp to mpc," 2005. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.68.4541
[7] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2010.
[8] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics)*. New York, NY, USA: Wiley-Interscience, 2007.
[9] R. Munos and C. Szepesvári, "Finite-time bounds for fitted value iteration," *J. Mach. Learn. Res.*, vol. 9, pp. 815–857, 2008.
[10] W. B. Haskell, R. Jain, and D. Kalathil, "Empirical dynamic programming," *Math. Operations Res.*, vol. 41, no. 2, pp. 402–429, 2016.
[11] J. Rust, "Using randomization to break the curse of dimensionality," *Econometrica: J. Econometric Soc.*, vol. 65, no. 3, pp. 487–516, 1997.
[12] D. P. De Farias and B. Van Roy, "On constraint sampling in the linear programming approach to approximate dynamic programming," *Math. Oper. Res.*, vol. 29, no. 3, pp. 462–478, 2004.
[13] D. Ormoneit and Ś. Sen, "Kernel-based reinforcement learning," *Mach. Learn.*, vol. 49, nos. 2/3, pp. 161–178, 2002.
[14] S. Grünewälder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton, "Modelling transition dynamics in MDPS with RKHS embeddings," *ICML*, Madison, WI, USA: Omnipress, pp. 535–542, 2012,
[15] C. Szepesvári, "Efficient approximate planning in continuous space markovian decision problems," *AI Commun.*, vol. 14, no. 3, pp. 163–176, 2001.
[16] R. Munos, "Performance bounds in l_p-norm for approximate value iteration," *SIAM J. Control Optim.*, vol. 46, no. 2, pp. 541–561, 2007.
[17] D. P. De Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Oper. Res.*, vol. 51, no. 6, pp. 850–865, 2003.
[18] N. Bhat, V. Farias, and C. C. Moallemi, "Non-parametric approximate dynamic programming via the kernel method," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 386–394.
[19] R. Munos, "Error bounds for approximate policy iteration," in *Proc. Int. Conf. Mach. Learn.*, 2003, vol. 3, pp. 560–567.
[20] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Proc. Advances Neural Inf. Process. Syst.*, 2000, pp. 1008–1014.
[21] R. Jain and P. Varaiya, "Simulation-based optimization of Markov decision processes: An empirical process theory approach," *Automatica*, vol. 46, no. 8, pp. 1297–1304, 2010.
[22] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1057–1063.
[23] J. Peters and J. A. Bagnell, "Policy gradient methods," in *Encyclopedia of Machine Learning and Data Mining*. New York, NY, USA: Springer, 2016, pp. 1–4.
[24] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[25] A. Rahimi and B. Recht, "Uniform approximation of functions with random bases," in *Proc. 46th Annu. Allerton Conf. Commun., Control, Comput.*, 2008, pp. 555–561.
[26] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1313–1320.
[27] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 2014.
[28] S. Smale and D.-X. Zhou, "Shannon sampling ii: Connections to learning theory," *Appl. Comput. Harmon. Anal.*, vol. 19, no. 3, pp. 285–302, 2005.
[29] M. Shaked and J. G. Shanthikumar, *Stochastic Orders*. New York, NY, USA: Springer, 2007.
[30] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*. Providence, RI, USA: Am. Math. Soc., 2008.
[31] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[32] D. Haussler, "Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik–Chervonenkis dimension," *J. Combinatorial Theory, Ser. A*, vol. 69, no. 2, pp. 217–232, 1995.


**William B. Haskell** received the B.S. degree in mathematics and the M.S. degree in econometrics from the University of Massachusetts Amherst, Amherst, MA, USA, in 2005 and 2006, respectively. He then received the M.S. degree in operations research, the M.A. degree in mathematics, and the Ph.D. degree in operations research from the University of California Berkeley, Berkeley, CA, USA, in 2007, 2010, and 2011, respectively.

He is currently an Assistant Professor with the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore. His research interests include algorithms for convex optimization and dynamic programming, with an emphasis on risk-aware decision-making.

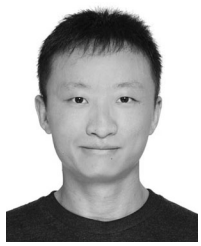
**Rahul Jain** received the B.Tech. degree from IIT Kanpur, Kanpur, India, the M.A. degree in statistics and the Ph.D. degree in EECS from the University of California, Berkeley, Berkeley, CA, USA.

He is the K.C. Dahlberg Early Career Chair and Associate Professor of ECE, CS* and ISE* (*by courtesy) with the University of Southern California, Los Angeles, CA, USA. His research interests include reinforcement learning, stochastic control, statistical learning, stochastic networks, and game theory, and power systems, transportation, and healthcare on the applications side.

Dr. Jain has received numerous awards including the NSF CAREER award, the ONR Young Investigator award, an IBM Faculty award, the James H. Zumberge Faculty Research and Innovation Award, and is currently a US Fulbright Scholar.

**Hiteshi Sharma** received the M.Tech. degree in electrical engineering from IIT Bombay, Mumbai, India, where she worked on dynamic spectrum sharing in cognitive radio networks. She is currently working toward the Ph.D. degree with the University of Southern California (USC), Los Angeles, CA, USA, working with Prof. R. Jain.

Her research interests include approximate dynamic programming, reinforcement learning, and online learning.



**Pengqian Yu** received the B.S. degree in mechanical design, manufacturing, and automation from the College of Mechanical Engineering, Chongqing University, Chongqing, China, in 2012, and the Ph.D. degree from the Department of Mechanical Engineering, National University of Singapore (NUS), Singapore, in 2016.

Since 2017, he has been a Postdoctoral Research Fellow with the Department of Industrial Systems Engineering and Management, NUS. His research interests include sequential decision-making under uncertainty, machine learning, approximate dynamic programming, and reinforcement learning.