2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Empirical Algorithms for General Stochastic Systems with Continuous States and Actions

Hiteshi Sharma, Rahul Jain and William Haskell

Abstract— **In this paper, we present Randomized Empirical Value Learning (RAEVL) algorithm for MDPs with continuous state and action spaces. This algorithm combines the ideas of random search over action space with randomized function approximation method to generalize the value functions over state space . Our theoretical analysis is done under a random operator framework combined with stochastic dominance argument. This provides finite-time analysis of the proposed algorithm as well as give the sample complexity.**

I. Introduction

In this paper, we consider a controlled Markov process as a model of a general stochastic system with finite memory. Bellman's dynamic programming principle can be used to find optimal control laws. A large variety of algorithms have been developed to find an exact solution to Bellman's optimality equation. Unfortunately, when the state space is continuous, designing algorithms that find exact solutions (even if asymptotically) is near impossible. State space discretization (or aggregation methods) may not work numerically and certainly do not scale [14]. Function approximation methods [12], [4], [3] can often have arbitrary (and unknown) gaps to optimality. Thus, in [6], we introduced an empirical value learning algorithm that used randomized function approximation in universal function approximation spaces (e.g., RKHS) to find arbitrarily accurate solutions with high probability. This combined "empirical dynamic programming" (EDP) ideas introduced in [5] with randomized function approximation ideas [13] to yield algorithms that work very well numerically, while still being able to provide probabilistic guarantees on non-asymptotic performance.

In many robotics problems, action space (and state space) is continuous with dimensions in low double digits [8]. And it is not possible to discretize state or action space without losing control performance (while also running into issues of scalability). Continuous action space problems are much harder. Current methods available for such settings include considering parametric families of policies, and then doing either following policy gradient [1], [10] or doing policy optimization [15], [16]. While for some of these algorithms, a proof of convergence is available, they can have arbitrary gaps to optimality since it is nearly impossible to know which parametric policy family is close enough to an optimal policy. Moreover, training such algorithms requires a large amount of data which may not be available. In the reinforcement learning (RL) literature, the closest useful techniques come from deep RL that rely on using deep neural networks for function approximation but can work with continuous action spaces. As with deep learning methods, such techniques suffer from long training times, no guarantees on optimality or performance, and the need to tune a large number of hyper-parameters before one can get reasonable performance on a given problem [9].

In this paper, we take a different approach. First, our goal is to develop algorithms that are universal (work on any problem), simple to implement, computationally tractable (and fairly easy), can provide arbitrarily good approximation to optimal performance and come with some performance guarantees, hopefully non-asymptotic, even if probabilistic. We build on our work in [6] for continuous state space but finite action space problems. The idea is very simple: Do function approximation in a universal function approximation space such as an RKHS. Do randomized function fitting by picking basis functions randomly in each iteration. Replace expectation in the Bellman operator with a sample average approximation by drawing samples of the next state. To optimize over the actions in the Bellman equation, sample a few actions, and just optimize over those. We call this algorithm Random Actions for Empirical Value Learning (or RAEVL).

Thus, the algorithm we introduce for continuous action and state space problems is pretty simple, and in fact performs quite well numerically as we show in Section V. But due to randomization and sampling of various kinds, its convergence analysis becomes intricate. This is addressed by viewing each iteration as operation of a random Bellman operator. In [5], probabilistic contraction analysis techniques were developed for analyzing iterated random operators. Convergence is proved by constructing a simple stochastically dominating Markov chain (which also yields the rate of convergence). We use the same framework but needed to modify some details for use in this paper. The second part of the analysis is related to concentration analysis of randomized function approximation [11]. The third part of the analysis is related to error analysis of the empirical optimum (obtained by taking samples of a function) with respect to the true optimum.

Putting all of this together then gives us RAEVL, one of the first practical algorithms for finding near-optimal policies in general (Markovian) stochastic systems with continuous states and actions with probabilistic guarantees on non-asymptotic performance (and explicit sample complexity

Rahul Jain and Hiteshi Sharma are with the EE Department at the University of Southern California. They were supported by NSF Awards CCF-1817212 and ECCS-1810447. (rahul.jain,hiteshis)@usc.edu

William B. Haskell is with Industrial and Systems Engineering, National University of Singapore isehwb@nus.edu.sg

978-1-7281-1398-2/19/$31.00 ©2019 IEEE

bounds).

II. Problem Formulation

Consider a MDP $(\mathcal{X}, \mathcal{U}, P, r, \gamma)$ where $\mathcal{X}$ is the state space and $\mathcal{U}$ is the action space. The transition probability kernel is given by $P(\cdot|x, u)$, i.e., if action u is executed in state x, the probability that the next state is in a Borel-measurable set B is $P(x_{t+1} \in B | x_t = x, u_t = u)$ where x_t and u_t are the state and action at time t. The reward function is $r : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$. We are interested in maximizing the long-run expected discounted reward where the discount parameter is γ.

Let Π denote the class of stationary deterministic Markov policies mappings $\pi : \mathcal{X} \to \mathcal{U}$ which only depend on history through the current state. We only consider such policies since it is well known that there is an optimal MDP policy in this class. When the initial state is given, any policy π determines a probability measure P^{π}. Let the expectation with respect to this measure be $\mathbb{E}^{\pi}$. We focus on infinite horizon discounted reward criterion. The expected discounted reward or the value function for a policy π and initial state x is given as

$$v^{\pi}(x) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \, r(x_t, a_t) \middle| x_0 = x\right]$$

The optimal value function is given as

$$v^{*}(x) = \sup_{\pi \in \Pi} \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \, r(x_t, a_t) \middle| x_0 = x\right]$$

and the policy which maximizes the value function is the optimal policy, π^*. Now we make the following assumptions on the regularity of the MDP.

Assumption 1: (**Regularity of MDP**) The state space $\mathcal{X}$ and the action space $\mathcal{U}$ are compact subset of d_X and d_U dimensional Euclidean space respectively. The rewards are uniformly bounded by $r_{\max}$, i.e, $r(x, u) \leq r_{\max}$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. Furthermore $\mathcal{U}$ is convex.

The assumption above implies that for any policy π, $v^{\pi} \leq v_{\max} = r_{\max}/(1-\gamma)$. The next assumption is on Lipschitz continuity of MDP in action variable.

Assumption 2: (**Lipschitz continuity**) The reward and the transition kernel are Lipschitz continuous with respect to the action i.e., there exists constants L_r and L_p such that for all $(x, u, u') \in \mathcal{X} \times \mathcal{U} \times \mathcal{U}$ and a measurable set B of $\mathcal{X}$, the following holds

$$|r(x, u) - r(x, u')| \leq L_r \|u - u'\|$$
$$|P(B|x, u) - P(B|x, u')| \leq L_p \|u - u'\|$$

The compactness of action space combined with Lipschitz continuity implies that the greedy policies do exist. Let $B(\mathcal{X})$ be the set of functions on $\mathcal{X}$ such that $\|f\|_{\infty} \leq v_{\max}$. Let us now define the Bellman operator $T : B(\mathcal{X}) \to B(\mathcal{X})$ as follows

$$T\, v(x) = \max_{u} \left[r(x, u) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,u)} v(x') \right].$$

It is well known that the operator T is a contraction with respect to $\|\cdot\|_{\infty}$ norm and the contraction parameter is the discount factor, γ. Hence, the sequence of iterates $v_k = T\, v_{k-1}$ converge to v^* geometrically. Since, we will be analyzing the L_2 norm, we do not have contraction property with respect to this norm. Hence, we need bounded Radon-Nikodym derivative of transitions which we illustrate in the next assumption. Such an assumption has been used earlier for finite action space [12], [7] and for continuous action space in [2].

Assumption 3: (**Stochastic Transitions**) For all $(x,\, u) \in \mathcal{X} \times \mathcal{U}$, $P(\cdot \,|\, x,\, u)$ is absolutely continuous with respect to μ and $C_{\mu} \triangleq \sup_{(x,\, u) \in \mathcal{X} \times \mathcal{U}} \left\| \frac{dP(\cdot\,|\,x,\,u)}{d\mu} \right\|_{\infty} < \infty$.

Since we have a sampling based algorithm, we need a function space to approximate value function. In this paper, we focus on randomized function approximation via random features. Let Θ be a set of parameters and let $\phi : \mathcal{X} \times \Theta \to \mathbb{R}$ be a feature function. The feature functions need to satisfy $\sup_{(x,\, \theta) \in \mathcal{X} \times \Theta} |\phi\,(x;\, \theta)\,| \leq 1$, for e.g., Fourier features. Let $\mathcal{F}\,(\Theta)$ be defined as

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

But we are interested in finding the best fit within finite sums of the form $\sum_{j=1}^{J} \alpha_j \phi\,(x;\, \theta_j)$. Doing classical function fitting with $\sum_{j=1}^{J} \alpha_j \phi\,(x;\, \theta_j)$ leads to nonconvex optimization problems because of the joint dependence in α and θ. Instead, we fix a density ν on Θ and draw a random sample θ_j from Θ for $j = 1, 2, \ldots J$. Once these $(\theta_j)_{j=1}^{J}$ are fixed, we consider the space of functions $\widehat{\mathcal{F}}\left(\theta^{1:J}\right) \triangleq$

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

Now, it remains to calculate weights α by minimizing a convex loss. Furthermore, let us define the $L_{2,\mu}$ norm of a function for a given a probability distribution μ on $\mathcal{X}$ as $\|f\|_{2,\,\mu}^2 = \left(\int_{\mathcal{X}} |f\,(x)\,|^2 \mu\,(dx)\right)$. The empirical norm at given samples $(x_1, x_2, \ldots x_N)$ is defined as $\|f\|_{2,\,\hat{\mu}}^2 = \frac{1}{N} \sum_{i=1}^{N} |f(x_i)|^2$. Recall two distance measures for function spaces:

- $d_{2,\,\mu}\,(T\,f,\, \mathcal{F}) \triangleq \inf_{f' \in \mathcal{F}} \|f' - T\,f\|_{2,\,\mu}$ is the approximation error for a specific f;
- $d_{2,\,\mu}\,(T\,\mathcal{F},\, \mathcal{F}) \triangleq \sup_{f \in \mathcal{F}} d_{2,\,\mu}\,(T\,f,\, \mathcal{F})$ is the inherent Bellman error for the entire class $\mathcal{F}$.

III. The Algorithm

We now present our RAndomized Empirical Value Learning (RAEVL) algorithm. It is an empirical variant of value iteration for continuous state and action space. It samples both states and actions. Note that the Bellman operator computes expectation with respect to the next state and then optimize over action space for each state. This is not feasible for continuous state and action space. Instead, we replace the expectation with an empirical average and use empirical optimization in the original Bellman operator. This means for a given state $x \in \mathcal{X}$ and sample sizes M and L, we first sample L actions uniformly and then generate M samples of next state (for each sampled action). This leads us to define an empirical Bellman operator $\widehat{T}_{M,L} : B(\mathcal{X}) \to B(\mathcal{X})$ as

$$\widehat{T}_{M,L}\, v(x) = \max_{u_1, u_2, \ldots u_L} \left[r(x, u_l) + \frac{\gamma}{M} \sum_{m=1}^{M} v(x'_{lm}) \right]$$

where $x'_{lm} \sim P(\cdot|x, u_l)$ for all $l = 1, 2 \ldots L$ and $m = 1, 2, \ldots M$. Instead of evaluating this operator at each state x, we sample $\{x_n\}_{n=1}^{N}$ from the state space $\mathcal{X}$ according to distribution μ. Then we compute $\widehat{v}(x_n) = [\widehat{T}_{M,L}\, v](x_n)$.

Given the data $\{(x_n, \widehat{v}(x_n))\}_{n=1}^{N}$, we generalize the value function over the state space by computing a best fit within $\widehat{\mathcal{F}}\left(\theta^{1:J}\right)$ by solving

$$\begin{aligned} \min_{\alpha} \; & \frac{1}{N} \sum_{n=1}^{N} |\sum_{j=1}^{J} \alpha_j \phi(x_n; \theta_j) - \widehat{v}(x_n)|^2 \\ \text{s.t.} \; & \|(\alpha_1, \ldots, \alpha_J)\|_\infty \le C/J. \end{aligned}$$

This optimization problem only optimizes over weights $\alpha^{1:J}$ since parameters $\theta^{1:J}$ have already been randomly sampled. Let $\Pi_{\widehat{\mathcal{F}}}(J, N)$ denote this optimization problem which we denote as $\Pi_{\widehat{\mathcal{F}}}$ for compact notation. We are now ready to present our algorithm.

Algorithm 1 RAEVL

```
Input: probability distribution $\mu$ on $\mathcal{X}$; sample sizes $N, M, J, L \ge 1$; counter $k = 0$, initial seed $v_0$

For k=1, 2, ...
  1) Sample $(x_n)_{n=1}^{N}$ from $\mathcal{X}$ according to $\mu$.
  2) For each $n$, sample $(u_l)_{l=1}^{L}$ uniformly for the action space
  3) For each $n$ and $l$, sample i.i.d. next states $y_{lm} \sim P(\cdot|x_n, u_l)$
  4) Empirical value iteration: $\widehat{v}_k(x_n) = \widehat{T}_{M,L}\, v_{k-1}(x_n)$
  5) Function approximation: $v_{k+1} = \Pi_{\widehat{\mathcal{F}}}\, \widehat{v}_k(x_n)$
  6) Increment $k \leftarrow k + 1$ and return to Step 1.
```

RAEVL can be seen as an iteration of a composition of two random operators. Let us define $\widehat{G}(N, M, L, J) = \Pi_{\widehat{\mathcal{F}}}(N, J) \circ \widehat{T}_{M,L}$ where $\circ$ denotes the composition. Let $\widehat{G}$ be a compact notation for this operator. Hence, in our algorithm $v_{k+1} = \widehat{G}\, v_k$. We will use the random operator framework to analyze our algorithm. We now present our main theorem for which we define

$$J_0(\epsilon, \delta) = \left[\frac{5\, C}{\epsilon} \left(1 + \sqrt{2 \log \frac{5}{\delta}} \right) \right]^2,$$

$$N_0(\epsilon, \delta) = \left(\frac{512\, v_{\max}^2}{(\epsilon/7)^4} \right) \log \left[\frac{56\, e\, (J+1)}{\delta} \left(\frac{2\, e\, v_{\max}}{(\epsilon/7)^2} \right)^J \right]$$

$$L_0(\epsilon, \delta) = \left(\frac{(L_r + \gamma\, v_{\max}\, L_p)\, \mathrm{diam}(\mathcal{U})}{\epsilon/7} \right)^{d_U} \log \frac{7\, N}{\delta}$$

and $M_0(\epsilon, \delta) = \left(\frac{2\, v_{\max}^2}{(\epsilon/7)^2} \right) \log \left[\frac{14\, N\, L}{\delta} \right]$.

Theorem 1: Suppose Assumptions 1, 2 and 3 hold. Choose an $\epsilon > 0$ and $\delta \in (0, 1)$. Set $\delta' = 1 - (1/2 + \delta/2)^{1/(K^*-1)}$ and

$$K^* = \left\lceil \frac{\log\left(C_\mu^{1/2} \epsilon\right) - \log(2\, v_{\max})}{\log \gamma} \right\rceil.$$

Then if $N \ge N_0(\epsilon, \delta')$, $M \ge M_0(\epsilon, \delta')$, $J \ge J_0(\epsilon, \delta')$, $L \ge L_0(\epsilon, \delta')$ and

$$K \ge \log\left(4 / \left((1/2 - \delta/2)\, (1 - q)\, q^{K^*-1} \right) \right),$$

the following holds with probability at least $1 - \delta$,

$$\|v_K - v^*\|_{2,\mu} \le \widetilde{C} \left[d_{2,\mu}\left(T\, \mathcal{F}(\Theta), \mathcal{F}(\Theta) \right) + 2\epsilon \right] \quad (1)$$

where $\widetilde{C} = 2 \left((1 - \gamma^{K+1})/(1 - \gamma) \right)^{1/2} C_\mu^{1/2}$.

IV. Proof of Theorem 1

There are three sources of error in RAEVL: empirical optimization, sample average and function approximation. We will get a handle on each of these approximation errors to give us a bound on the error in one iteration. We then use a stochastic dominance argument to analyze the error process. As mentioned before, we view RAEVL as an iteration of the random operator $\widehat{G}$. Let ϵ_k be the gap between this random operator and the exact Bellman operator at iteration k, i.e,

$$v_{k+1} = \widehat{G}\, v_k = T\, v_k + \epsilon_k$$

Let us also define a (random) operator, $\widetilde{T}_L$ as follows

$$\widetilde{T}_L\, v(x) = \max_{u_1, u_2, \ldots u_L} \left[r(x, u_l) + \gamma \mathbb{E}_{x' \sim P(\cdot|x, u_l)} v(x') \right]$$

where $u_l \sim \mathrm{Unif}(\mathcal{U})$ for $l = 1, 2, \ldots L$. Let the Q-value function be $Q(x, u) = r(x, u) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,u)} v(x')$ for all $v \in B(\mathcal{X})$. We now argue that the Q-value function is L_U-Lipschitz continuous in action variable where $L_U = L_r + \gamma\, v_{\max} L_p$. For all $(x, u, u') \in \mathcal{X} \times \mathcal{U} \times \mathcal{U}$ and $v \in B(\mathcal{X})$,

$$\begin{aligned} & |Q(x, u) - Q(x, u')| \\ & \le |r(x, u) - r(x, u')| + \gamma \int_{\mathcal{X}} |(P(dy|x, u) - P(dy|x, u'))\, v(y)| \\ & \le L_r |u - u'| + \gamma\, v_{\max} \int_{\mathcal{X}} |P(dy|x, u) - P(dy|x, u')| \\ & \le (L_r + \gamma\, v_{\max} L_p)\, \|u - u'\| \end{aligned}$$

where the last inequalities follow from Assumption 2. The next lemma bounds the error due to sampling for finding the best action. The proof is given in the appendix.

Lemma 2: Choose $\epsilon > 0$ and $\delta \in (0, 1)$. Let $\mathrm{diam}(\mathcal{U})$ be the diameter of the action space $\mathcal{U}$. Then for all $v \in B(\mathcal{X})$, if

$$L \ge \left(\frac{L_U\, \mathrm{diam}(\mathcal{U})}{\epsilon} \right)^{d_U} \log \frac{1}{\delta} \quad (2)$$

then

$$\mathbb{P}\left(|T\, v(x) - \widetilde{T}_L\, v(x)| > \epsilon \right) < \delta$$

for all $x \in \mathcal{X}$.

Next, we bound the error due to sample average. This will give us a sample complexity bound for next state samples. The proof is a simple application of Hoeffding's inequality followed by an union bound.

Lemma 3: Choose $\epsilon > 0$, $\delta \in (0, 1)$ and $L \geq 1$. Then for all $v \in B(\mathcal{X})$, if

$$M \geq \frac{2v_{\max}^2}{\epsilon^2} \log\left(\frac{2L}{\delta}\right)$$

then

$$\mathbb{P}\left(|\widehat{T}_{M,L}\, v(x) - \widetilde{T}_L\, v(x)| > \epsilon\right) < \delta$$

for all $x \in \mathcal{X}$.

A. Bound on one-step error

Now, we will bound the error in one iteration of RAEVL. We will use the bounds on M and L developed in the previous section. The choice of N comes from bounding the gap between the empirical norm and expected norm by Pollard's inequality. Lastly, the bound on J comes through an application of the bounded difference concentration inequality as given in [13].

Lemma 4: Choose $v \in \mathcal{F}(\Theta)$, $\epsilon > 0$, and $\delta \in (0, 1)$. Also choose $N \geq N_0(\epsilon, \delta), M \geq M_0(\epsilon, \delta), J \geq J_0(\epsilon, \delta)$ and $L \geq L_0(\epsilon, \delta)$. Then, for $\widehat{G}\,(N,\, M,\, J,\, L,\, \mu,\, \nu)\; v$, the output of one iteration of our algorithm, we have

$$\|\widehat{G}\, v - T\, v\|_{2,\,\mu} \leq d_{2,\,\mu}\left(T\,\mathcal{F}\,(\Theta)\,,\, \mathcal{F}\,(\Theta)\right) + \epsilon$$

with probability at least $1 - \delta$.

Proof: To begin, let $\epsilon' > 0$ be arbitrary and choose $f^* \in \mathcal{F}\,(\Theta)$ such that $\|f^* - T\,v\|_{2,\,\mu} \leq \inf_{f \in \mathcal{F}(\Theta)} \|f - T\,v\|_{2,\,\mu} + \epsilon'$. Using $(x+y)^2 \geq x^2 + y^2$ for $x, y \geq 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\widehat{f} \in \widehat{\mathcal{F}}(\theta^{1:J})} \left| \|\widehat{f} - T\,v\|_{2,\mu} - \|\widehat{f} - T\,v\|_{2,\hat{\mu}} \right| > \epsilon/7\right) \\
&\leq \mathbb{P}\left(\sup_{\widehat{f} \in \widehat{\mathcal{F}}(\theta^{1:J})} \left| \|\widehat{f} - T\,v\|_{2,\mu}^2 - \|\widehat{f} - T\,v\|_{2,\hat{\mu}}^2 \right| > (\epsilon/7)^2\right) \\
&\leq 8\,e\,(J+1)\left(\frac{4\,e\,v_{\max}}{(\epsilon/7)^2}\right)^J \exp\left(\frac{-\,N\,(\epsilon/7)^4}{512\,v_{\max}^2}\right) \qquad (3)
\end{aligned}$$

where the last inequality follows from Pollard's inequality and the fact that the psuedo-dimension for the function class $\widehat{\mathcal{F}}\left(\theta^{1:J}\right)$ is J. Then, choose $\hat{f} \in \widehat{\mathcal{F}}\left(\theta^{1:J}\right)$ such that $\|\hat{f} - T\,v\|_{2,\,\mu} \leq \|f^* - T\,v\|_{2,\,\mu} + \epsilon/7$ with probability at least $1 - \delta/7$ by choosing $J \geq 1$ to satisfy

$$\frac{C}{\sqrt{J}}\left(1 + \sqrt{2\,\log\frac{1}{(\delta/7)}}\right) \leq \frac{\epsilon}{7} \Rightarrow$$

$$J \geq \left[\left(\frac{7\,C}{\epsilon}\right)\left(1 + \sqrt{2\,\log\frac{7}{\delta}}\right)\right]^2$$

by Lemma [13, Lemma 1]. Now we have the following string of inequalities, each of which hold with probability $1 - \delta/7$:

$$\begin{aligned}
\|\widehat{G}\,v - T\,v\|_{2,\mu} &\leq \|\widehat{G}\,v - T\,v\|_{2,\hat{\mu}} + \epsilon/7 &(4)\\
&\leq \|\widehat{G}\,v - \widetilde{T}_L\,v\|_{2,\hat{\mu}} + 2\epsilon/7 &(5)\\
&\leq \|\widehat{G}\,v - \widehat{T}_{M,L}\,v\|_{2,\hat{\mu}} + 3\epsilon/7 &(6)\\
&\leq \|\hat{f} - \widehat{T}_{M,L}\,v\|_{2,\hat{\mu}} + 3\epsilon/7 &(7)\\
&\leq \|f^* - \widehat{T}_{M,L}\,v\|_{2,\hat{\mu}} + 4\epsilon/7 &(8)\\
&\leq \|f^* - \widetilde{T}_L\,v\|_{2,\hat{\mu}} + 5\epsilon/7 &(9)\\
&\leq \|f^* - T\,v\|_{2,\hat{\mu}} + 6\epsilon/7 &(10)\\
&\leq \|f^* - T\,v\|_{2,\mu} + \epsilon &(11)\\
&\leq \inf_{f \in \mathcal{F}(\Theta)} \|f - T\,v\|_{2,\,\mu} + \epsilon' + \epsilon &(12)
\end{aligned}$$

We choose N from inequality (3) such that inequalities (4) and (11) hold with atleast probability $1 - \delta/7$. Now, using Lemma 2 followed by an union bound argument we choose L such that

$$\mathbb{P}\left(\max_{x_1, x_2, \ldots x_N} \left|T\,v(x_n) - \widetilde{T}_L\,v(x_n)\right| < \epsilon \,\middle|\, \{x_n\}_{n=1}^N\right) \geq 1 - \delta/7$$

Hence the empirical norm can also be bounded. This proves inequalities (5) and (10). Similarly, one can prove inequalities (6) and (9) using Lemma 3 followed by union bound argument on both the sampled, giving us a bound on M. Inequality (7) follows from the fact that $\widehat{G}$ gives the least approximation error compared to any other function $\hat{f} \in \widehat{\mathcal{F}}\left(\theta^{1:J}\right)$. The last inequality is by the choice of f^*. ■

The previous Lemma 4 gives a bound on the error in one step of RAEVL. We will now extend this result to understand the convergence of $\{\|v_k - v^*\|_{2,\,\mu}\}_{k \geq 0}$ in the next section.

B. Stochastic Dominance

Since we do not have contraction with respect to L_2 norm, we need a bound on how the errors propagate with iterations. Recall that $\widehat{G}\,v_k = T\,v_k + \epsilon_k$, we have the following by analyzing the point-wise error bounds.

Lemma 5: [12, Lemma 3] For any $K \geq 1$, and $\epsilon > 0$, suppose $\|\epsilon_k\|_{2,\,\mu} \leq \epsilon$ for all $k = 0,\, 1, \ldots,\, K-1$, then

$$\|v_K - v^*\|_{2,\,\mu} \leq 2\left(\frac{1 - \gamma^{K+1}}{1 - \gamma}\right)^{\frac{1}{2}} \left[C_\mu^{1/2}\epsilon + \gamma^{K/2}\,(2\,v_{\max})\right]. \qquad (13)$$

Now, from (13), we have

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

which gives a bound on K such that $\gamma^{K/2}\,(2\,v_{\max}) \leq C_\mu^{1/2}\epsilon$. Denote

$$K^* = \left\lceil \frac{\log\left(C_\mu^{1/2}\epsilon\right) - \log\left(2\,v_{\max}\right)}{\log \gamma} \right\rceil \qquad (14)$$

We then construct a stochastic process as follows. We call iteration k "good" if the error $\|\epsilon_k\|_{2,\,\mu}$ is within our desired tolerance ϵ and iteration k "bad" when the accuracy is greater than our desired tolerance. We then construct a stochastic process $\{X_k\}_{k\geq 0}$ with state space $\mathcal{K}$ as $\triangleq \{1,\ 2, \ldots,\ K^*\}$ 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. Whenever $X_k = 1$, it means that we just had a string of K^* "good" iterations in a row, and that $\|v_k - v^*\|_{2,\,\mu}$ is as small as desired.

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 $\{Y_k\}_{k\geq 0}$ and we let $\mathcal{Q}$ denote the probability measure of $\{Y_k\}_{k\geq 0}$. 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 now use the bound on one-step error as presented in Lemma 4 which states that when the samples are sufficiently large enough for all k,

$$\mathbb{P}\left(\|\epsilon_k\|_{2,\mu} \leq \epsilon\right) > q(N, M, J, L)$$

Let us denote this probability by q for a compact notation. Let us 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 which increases with sample sizes N, M, J and L. 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. This relationship is the key to our analysis of $\{X_k\}_{k\geq 0}$.

Definition 1: Let X and Y be two real-valued random variables, then X is *stochastically dominated* by Y, written $X \leq_{st} Y$, when $\Pr\{X \geq \theta\} \leq \Pr\{Y \geq \theta\}$ for all θ 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 $[X_{k+1}\,|\,\mathcal{F}_k]$ denote the conditional distribution of X_{k+1} given the information $\mathcal{F}_k$. We infer the following key result from the relationship between $\{X_k\}_{k\geq 0}$ and $\{Y_k\}_{k\geq 0}$.

Lemma 6: Choose $\epsilon > 0$, and $\delta \in (0, 1)$, and suppose N, M, J and L are chosen sufficiently large enough such that $\mathbb{P}\left(\|\epsilon_k\|_{2,\mu} \leq \epsilon\right) > q$ for all $k \geq 0$. Then for $q \geq (1/2 + \delta/2)^{1/(K^*-1)}$ and

$$K \geq \log\left(4/\left((1/2 - \delta/2)\,(1-q)\,q^{K^*-1}\right)\right),$$

we have

$$\|v_K - v^*\|_{2,\,\mu} \leq 2\left(\frac{1-\gamma^{K^*+1}}{1-\gamma}\right)^{1/2}\left[C_\mu^{1/2}\epsilon + \gamma^{K^*/2}\,(2\,v_{\max})\right] \tag{15}$$

with probability at least $1-\delta$.

Proof: This proof proceeds in three steps. First, we show that $X_k \leq_{st} Y_k$ holds for all $k \geq 0$. This stochastic dominance relation is the key to our analysis, since if we can show that Y_K is "small" with high probability, then X_K must also be small and we infer that $\|v_K - v^*\|_{2,\,\mu}$ must be close to zero. By construction, $X_k \leq_{st} Y_k$ for all $k \geq 0$ (see [5, Lemma A.1] and [5, Lemma A.2])

Second, we compute the steady state distribution of $\{Y_k\}_{k\geq 0}$ and its mixing time, in particular, we use the mixing time to choose K so that the distribution of Y_K is close to its steady state distribution. 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 [5, 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

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

$\forall q \in (0,\,1)$ and $K^* \geq 1$ appears shortly in the Markov chain mixing time bound for $\{Y_k\}_{k\geq 0}$. We note that $(1-q)\,q^{K^*-1} \leq \mu_{\min}(q;\, K^*)$ is a simple lower bound for $\mu_{\min}(q;\, K^*)$. Let Q^k be the marginal distribution of Y_k for $k \geq 0$. By a Markov chain mixing time argument, we have

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

for any $\delta' \in (0,\,1)$.

Finally, we conclude the argument by using the previous part to find the probability that $Y_K = 1$, which is an upper bound on the probability that $X_K = 1$, which is an upper bound on the probability that $\|v_K - v^*\|_{2,\,\mu}$ is below our desired error tolerance. For $K \geq \log\left(1/\left(\delta'\,(1-q)\,q^{K^*-1}\right)\right)$ we have $|\Pr\{Y_K = 1\} - \mu(1)| \leq 2\,\delta'$. Since $X_K \leq_{st} Y_K$, we have $\Pr\{X_K = 1\} \geq \Pr\{Y_K = 1\}$ and so $\Pr\{X_K = 1\} \geq q^{K^*-1} - 2\,\delta'$. Choose q and δ' 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. ■

Now, putting Lemma 4 and 6 together along with the choice of K^*, we can conclude Theorem 1.

V. Numerical Experiments

In this section, we try RAEVL on a synthetic problem for which we can compute the optimal value function. This allows us to compute error with each iteration. Let $\mathcal{X} = [0, 1]$ and $\mathcal{U} = [0, 1]$. The reward is $r(x, u) = -(x - u)^2$ and transition probability density $p(y|x, u)$. Then the optimal value function can be found by solving the fixed point equation

$$v^*(x) = \max_{0\le u\le 1}\left\{-(x-u)^2 + \frac{\gamma}{1-u}\int_u^1 v^*(y)dy\right\}$$

which gives $v^*(x) = 0$ and $\pi^*(x) = x$ for all $x \in \mathcal{X}$. For our experiments, we choose $\phi(x,\theta) = \cos(w^T x + b)$ where $\theta = (w, b)$. We sample $w \sim \mathcal{N}(0,1)$ and $b \sim$ Unif$[-1,1]$. We fix number of features, $J = 10$. Fig. 1 shows the error $\|v_k - v^*\|_\infty$ on y-axis with number of iterations k on x-axis for different choices of N, M and L. The error reduces to order of 10^{-4} when the the sample sizes are sufficiently large $N = M = L = 50$. But even for low values of sample sizes, we are able to get an error ≈ 0.1 which indicates that we can get good approximation with less computation.

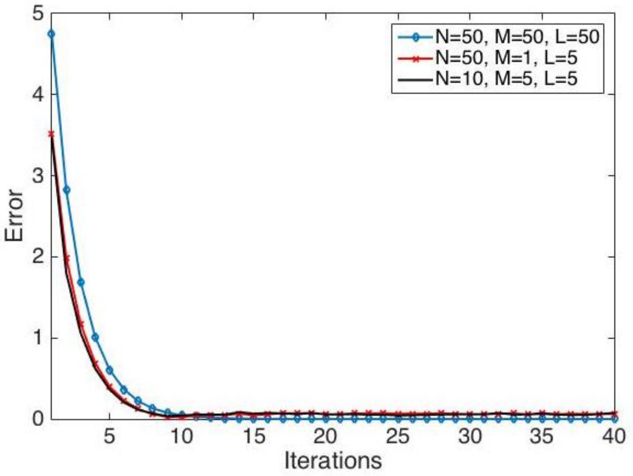

Fig. 1. Performance of RAEVL for different sample sizes

References

[1] Douglas Aberdeen. Policy-gradient methods for planning. In *Advances in Neural Information Processing Systems*, pages 9–16, 2006.

[2] András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. In *Advances in neural information processing systems*, pages 9–16, 2008.

[3] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1 and 2. Athena Scientific Belmont, 1995.

[4] Geoffrey J Gordon. Stable function approximation in dynamic programming. In *Machine Learning Proceedings 1995*, pages 261–268. Elsevier, 1995.

[5] William B. Haskell, Rahul Jain, and Dileep Kalathil. Empirical dynamic programming. *Mathematics of Operations Research*, 2016.

[6] William B. Haskell, Rahul Jain, Hiteshi Sharma, and Pengqian Yu. An Empirical Dynamic Programming Algorithm for Continuous MDPs. *IEEE Transactions on Automatic Control*, 2019.

[7] William B Haskell, Pengqian Yu, Hiteshi Sharma, and Rahul Jain. Randomized function fitting-based empirical value iteration. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2467–2472. IEEE, 2017.

[8] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), 2019.

[9] Riashat Islam, Peter Henderson, Maziar Gomrokchi, and Doina Precup. Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. *arXiv preprint arXiv:1708.04133*, 2017.

[10] Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.

[11] Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

[12] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *The Journal of Machine Learning Research*, 9:815–857, 2008.

[13] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2009.

[14] John Rust. Using randomization to break the curse of dimensionality. *Econometrica*, pages 487–516, 1997.

[15] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[16] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[17] Zelda B Zabinsky and Robert L Smith. Pure adaptive search in global optimization. *Mathematical Programming*, 53(1-3):323–338, 1992.

Appendix

Proof: [Proof of Lemma 2] Let $f(u) = r(x,u) + \gamma\mathbb{E}_{x'}v(x')$ for a given $x \in \mathcal{X}$. Let u^* be the maxima. Let the ball centered at u and radius r be $\mathcal{B}(u,r)$. Now, the volume of this d_U- dimension ball is vol$(\mathcal{B}(u,r)) \propto r^{d_U}$. Let $\mathcal{U}_\epsilon = \{u \in \mathcal{U} : f(u) \le \max_u f(u) - \epsilon\}$. Moreover, let $\mathcal{U}_{\epsilon,L_U} = \{u \in \mathcal{U} : \|u^* - u\| \le \epsilon/L_U\}$. Since f is L_U-Lipschitz, $u \notin \mathcal{U}_\epsilon \implies u \notin \mathcal{U}_{\epsilon,L_U}$. Hence,

$$\begin{aligned}\mathbb{P}\left(u \notin \mathcal{U}_{\epsilon,L_U}\right) &= 1 - \frac{\text{vol}(\mathcal{U}_{\epsilon,L_U})}{\text{vol}(\mathcal{U})}\\ &\le 1 - \left(\frac{\epsilon}{L_U\,\text{diam}(\mathcal{U})}\right)^{d_U} \qquad (16)\end{aligned}$$

where the last inequality follows from Lemma 5.2 in [17] and diam$(\mathcal{U}) = \sup_{u,u'}\|u - u'\|$. Now,

$$\begin{aligned}\mathbb{P}\left(f(u^*) - \max_{1\le l\le L} f(u_l) \le \epsilon\right) &= 1 - \mathbb{P}\left(\cap_{l=1}^L\{u_l \notin \mathcal{U}_\epsilon\}\right)\\ &= 1 - \mathbb{P}\left(\cap_{l=1}^L\{u_l \notin \mathcal{U}_\epsilon\}\right)^L\\ &\ge 1 - \mathbb{P}\left(\cap_{l=1}^L\{u_l \notin \mathcal{U}_{\epsilon,L_U}\}\right)^L\end{aligned}$$

where the second equality is due to the fact that $\{u_1, u_2 \dots u_L\}$ are i.i.d. and the last inequality follows Lipschitz continuity of the function f. Now, using (16) we have

$$\mathbb{P}\left(f(u^*) - \max_{1\le l\le L} f(u_l) \le \epsilon\right) \ge$$
$$1 - \left(1 - \left(\frac{\epsilon}{L_u\,\text{diam}(\mathcal{U})}\right)^{d_U}\right)^L$$

Putting $\left(\frac{\epsilon}{L_u\,\text{diam}(\mathcal{U})}\right)^{d_U} = \frac{1}{L}\log\left(\frac{1}{\delta}\right)$ and using $1 - x \le e^{-x}$, we have $\mathbb{P}\left(f(u^*) - \max_{1\le l\le L} f(u_l) \le \epsilon\right) \ge 1 - \delta$ for the choice of L as presented in (2). ■