2019 18th European Control Conference (ECC)
Napoli, Italy, June 25-28, 2019

An Empirical Relative Value Learning Algorithm for Non-parametric MDPs with Continuous State Space

Hiteshi Sharma[1], Rahul Jain[1], and Abhishek Gupta[2]

Abstract— **We propose an empirical relative value learning (ERVL) algorithm for non-parametric MDPs with continuous state space and finite actions and average reward criterion. The ERVL algorithm relies on function approximation via nearest neighbors, and minibatch samples for value function update. It is universal (will work for any MDP), computationally quite simple and yet provides arbitrarily good approximation with high probability in finite time. This is the first such algorithm for non-parametric (and continuous state space) MDPs with average reward criteria with these provable properties as far as we know. Numerical evaluation on a benchmark problem of optimal replacement suggests good performance.**

I. INTRODUCTION

Real-time decision making in uncertain environments are often modeled as Markov Decision Processes (MDPs) [13]. For infinite-horizon MDPs, while there are many reinforcement learning (RL) and approximate dynamic programming (ADP) algorithms available [12] for the discounted rewards criterion [3], the average rewards criterion is harder. Indeed, MDPs with average reward criterion are more difficult to analyze because establishing the existence of stationary optimal policy itself requires some restriction on the underlying Markov chains [1]. Approximate DP and RL for continuous state space MDPs is even harder, and mostly an art form. Different techniques (state space aggregation and function approximation [4]) and algorithms work for different problems but universal algorithms that work for any problem are unavailable. A popular RL algorithm for continuous MDPs is fitted value iteration (FVI) [10] which is quite effective for many problems but requires that we choose a set of basis functions appropriate to the problem for good approximation. Unfortunately, such a choice is an art-form, and given the complexity of such problems in the real world, we may not even be able to tell how well it is working.

[1] Rahul Jain and Hiteshi Sharma are with the Department of Electrical Engineering, University of Southern California. This research is supported by NSF Awards ECCS- 1611574 and CCF-1817212
(rahul.jain,hiteshis)@usc.edu
[2]Abhishek Gupta is with the Department of Electrical Engineering, Ohio State University. gupta.706@osu.edu

The goal of this paper is to introduce an off-policy empirical (or approximate), (relative) value learning algorithm for computing optimal policies for non-parametric MDPs with continuous state space and average reward criterion. Specifically, we aim to design RL algorithms that are universal (arbitrarily good approximation for any MDP), computationally simple, easy to implement and yet we would like to have non-asymptotic sample complexity bounds, even if the guarantees are probabilistic. We would also like such algorithms to work well numerically.

The algorithm we propose is inspired by the empirical dynamic programming (EDP) framework [7] - an off-policy minibatch empirical value learning algorithm for discounted MDPs with finite state and action space. The key idea is to replace the expectation in the Bellman operator with a sample average approximation obtained from a mini-batch of samples of the next state. Convergence analysis required a random operator framework and construction of a Markov chain that stochastically dominates the error introduced due to the approximation. An ERVL algorithm for average reward MDPs was proposed in [6] but for finite state space only. Our problem is non-parametric and substantially harder. The second inspiration is the work on non-parametric function approximation [5], [2]. [5] provides an uniform convergence of nearest neighbor regression under some smoothness assumption of the regression function. This enables us to plug in the function fitting part, and give guarantees in the sup-norm.

In this paper, we combine both of the above ideas to develop the ERVL algorithm for non-parametric, average-reward MDPs with continuous state spaces. This requires use of non-parametric regression techniques to generalize the value iterates on a small set of sampled states to the entire state space. We are able to provide finite time sample complexity bounds for the algorithm that is near-universal (requires only minimal asumptions on the class of MDPs we can handle) and arbitrarily good approximation with high probability. A kernel RL algorithm was introduced earlier where the expectation operator is approximated by local averagers [11].

978-3-907144-00-8

II. PRELIMINARIES

Consider an MDP with state space $\mathcal{X}$ and action space $\mathcal{A}$. We assume that $\mathcal{X}$ is a compact subset in $\mathbb{R}^d$ and $\mathcal{A}$ is finite. Let $\mathcal{C}_B(\mathcal{X})$ be the set of continuous and bounded functions over $\mathcal{X}$. The transition probability kernel is given by $P(\cdot|x,a)$, i.e., if action a is executed in state x, the probability that the next state is in a Borel-measurable set B is $P(X_{t+1} \in B|X_t = x, a_t = a)$. The reward function is $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. For a stationary policy $\pi : \mathcal{X} \times \mathcal{A}$, we are interested in maximizing the long-run average expected reward defined as

$$J^\pi(x) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} r(x_t, a_t)\middle| x_0 = x, a_t = \pi(x_t)\right].$$

Let $J^* = \sup_\pi J^\pi(x)$. A policy π^* is said to be optimal if for all $x \in \mathcal{X}$, it satisfies $J^{\pi^*}(x) = J^*$. We make the following assumptions.

Assumption II.1. *(a) For every (x,a), $|r(x,a)| \leq r_{\max}$ and for every a, $r(\cdot,a)$ is continuous.*
(b) For every a, transition kernel $P(\cdot|x,a)$ is continuous in x.
(c) There exists $\alpha < 1$ such that

$$\sup_{(x,a),(x',a')} \|P(\cdot|x,a) - P(\cdot|x',a')\|_{TV} \leq 2\alpha$$

where $\|\cdot\|_{TV}$ denotes the total variation norm.

Assumption II.1 (a) establishes that for every a, $r(\cdot,a) \in \mathcal{C}_B(\mathcal{X})$, (b) implies that if $v \in \mathcal{C}_B(\mathcal{X})$ then for any action a, $\int v(y)P(dy|\cdot,a) \in \mathcal{C}_B(\mathcal{X})$ and (c) implies that under any stationary and deterministic policy, t-step transition probability converges to a unique invariant probability measure (over the state process $\{x_t\}$) in total variation norm, uniformly in x and at a geometric rate. Under these assumptions, there exists $(J^*, v^*) \in \mathbb{R} \times \mathcal{C}_B(\mathcal{X})$ such that the following optimality equation holds:

$$J^* + v^*(x) = \sup_{a\in\mathcal{A}} \left\{ r(x,a) + \int v^*(x')P(dx'|x,a) \right\}. \tag{1}$$

Define the Bellman operator $T : \mathcal{C}_B(\mathcal{X}) \to \mathcal{C}_B(\mathcal{X})$ as

$$Tv(x) = \max_{a\in\mathcal{A}} \left[r(x,a) + \mathbb{E}_{x'\sim P(\cdot|x,a)} v(x') \right].$$

Hence, $J^* = Tv^* - v^*$. Note that v^* is unique upto a constant. The exact computation of the expectation in the Bellman operator is computationally infeasible for continuous state space for most problems. So we replace this by a sample average approximation which leads us to define an empirical Bellman operator $\widehat{T}_M$ as

$$\widehat{T}_M\, v(x) = \max_{a\in\mathcal{A}} \left[r(x,a) + \frac{1}{M}\sum_{i=1}^{M} v(x_i') \right]$$

where $x_i' \sim P(\cdot|x,a)$ for $i = 1, 2, \ldots, M$. Note that x_i' are i.i.d. It can clearly be seen that $\widehat{T}_M\, v$ is biased estimator of Tv, i.e., $\mathbb{E}_{x'}\left[\widehat{T}_M\, v\right] \neq Tv$.

Iteration on a Quotient Space.: Let us now define the span semi-norm and the quotient space. For a function $f \in \mathcal{C}_B(\mathcal{X})$, $span(f) := \sup_{(x)} f(x) - \inf_x f(x)$. Clearly, this is a semi-norm and for the constant function f, we have $span(f) = 0$. Let us now define an equivalence relation $\sim$ on $\mathcal{C}_B(\mathcal{X})$ defined by $f \sim g$ if and only if there exists a constant c such that for all $x \in \mathcal{X}$, $f(x) - g(x) = c$. Let $\widetilde{\mathcal{C}_B}(\mathcal{X}) = \mathcal{C}_B(\mathcal{X})/\sim$ be the quotient space. The following for the quotient space then is not difficult to show.

Lemma II.1. *[6] $(\widetilde{\mathcal{C}_B}(\mathcal{X}), span)$ is a Banach space.*

The proof is available in [6] and hence omitted. Furthermore, we can show that the operator T is a contraction in the span semi-norm. The next theorem is from [8].

Theorem II.2. *[8] Suppose that Assumptions II.1 hold. Then, operator $T : \mathcal{C}_B(\mathcal{X}) \to \mathcal{C}_B(\mathcal{X})$ is a span-contraction operator, i.e.,*

$$span(Tv_1 - Tv_2) \leq \alpha\, span(v_1 - v_2)$$

where $v_1, v_2 \in \mathcal{C}_B(\mathcal{X})$ and α is defined in Assumption II.1(c).

Now consider a $v \in \mathcal{C}_B(\mathcal{X})$, and let $\tilde{v}$ be the corresponding element in $\widetilde{\mathcal{C}_B}(\mathcal{X})$ and $\widetilde{T} : \widetilde{\mathcal{C}_B}(\mathcal{X}) \to \widetilde{\mathcal{C}_B}(\mathcal{X})$ defined as $\widetilde{T}\tilde{v} = \widetilde{Tv}$. Since T is a span-contraction on $\mathcal{C}_B(\mathcal{X})$, then $\widetilde{T}$ is a contraction on the Banach space $\widetilde{\mathcal{C}_B}(\mathcal{X})$, which by Banach fixed point theorem implies has a unique fixed point, which can be found by a simple iterative procedure on the quotient space that is easy to translate into an operation on the space $\mathcal{C}_B(\mathcal{X})$. Note that the empirical Bellman operator $\widehat{T}_M$ is not a span-contraction since the contraction coefficient α_M in

$$span(\widehat{T}_M v_1 - \widehat{T}_M v_2) \leq \alpha_M\, span(v_1 - v_2)$$

is random. But $\alpha_M < 1$ always, and intuitively, we can expect that $\widehat{T}_M$ *probabilistically contract* to a *probabilistic fixed point* [7]. Indeed, we introduce the following definition.

Definition II.1. *A function $v \in \mathcal{C}_B(\mathcal{X})$ is an (ϵ,δ)-weak probabilistic fixed point for a sequence of random operators $\{\widehat{G}_{N,M}\}$ with respect to a given norm $\|\cdot\|$ if there exist an $N_0(\epsilon,\delta)$, $M_0(\epsilon,\delta)$ and a $K_0(\epsilon,\delta)$ such that for all $N > N_0(\epsilon,\delta)$, $M > M_0(\epsilon,\delta)$ and all $k > K_0(\epsilon,\delta)$,*

$$\mathbb{P}(\|\widehat{G}^k_{N,M}\, v_0 - v\| > \epsilon) < \delta$$

for all $v_0 \in \mathcal{C}_B(\mathcal{X})$.

Let us now define a function space $\mathcal{G} = \{f : \mathcal{X} \to \mathbb{R}\}$. Let $\Pi_{\mathcal{G}}^N$ be the projection operator (why indexed by N will be clear later) which projects a function onto the space $\mathcal{G}$. While various non-parametric function spaces can be considered, we will choose nearest neighbors for function approximation (other non-parametric function approximation methods for e.g., kernel regression, etc. will also work). In each case, we can show that the composition operator $\Pi_{\mathcal{G}}^N \widehat{T}_M$ is a contraction with respect to the span semi-norm. We can also establish a probabilistic bound on the one-step error for each such method. Furthermore, note that for any function f, $span(f) \leq 2\|f\|_\infty$.

III. The Algorithm and Main Result

We now present the Empirical Relative Value Learning (ERVL) algorithm, a mini-batch off-policy algorithm for non-parametric MDPs with continuous state space. It is an 'empirical' or mini-batch variant of the relative value iteration algorithm for continuous state space MDPs. Note that if α is the contraction parameter, then $span(v^*) \leq \frac{r_{\max}}{1-\alpha} = v_{\max}$. Since it is not possible to argue that the iterates of relative Q-value iteration are bounded, we introduce the truncation operator Γ,

$$\Gamma v := \begin{cases} v - \min v, & \text{if} \quad span(v) \leq \dfrac{r_{\max}}{1-\alpha}, \\ \dfrac{v - \min v}{span(v)}, & \text{otherwise.} \end{cases}$$

The operator $\Gamma\, \widehat{T}_M$ maintains the contraction property of $\widehat{T}_M$ as Γ is just a projection operator[6]. With slight abuse of notation, we will denote $\widehat{T}_M$ as the empirical Bellman operator with truncation in the subsequent sections. Let us denote the composed operator by $\widehat{G}_{N,M} = \Pi_{\mathcal{G}}^N \widehat{T}_M$. Algorithm 1 will iterate the random operator $\widehat{G}_{N,M}$, i.e., $v_{k+1} = \widehat{G}_{N,M} v_k = \widehat{G}_{N,M}^k v_0$. Note that this operator depends on the sample sizes N and M. Note that $\widehat{T}_M$ probabilistically contracts, i.e., with high probability it is a contraction. Further, we will argue that $\Pi_{\mathcal{G}}^N$ is non-expansive, which will imply the random operator $\widehat{G}_{N,M}$ also probabilistically contracts.

Function Approximation using nearest neighbors: $k-$NN is a powerful yet simple approach in non-parametric regression. In this case, we first fix $x \in \mathcal{X}$ and reorder the samples $\{x_1, x_2, \ldots x_N\}$ according to increasing distance of x_i from x. Let the reordered samples be $\{x_{(i)}\}$ for $1 \leq i \leq N$. Now we pick k nearest neighbors and estimate the function as

$$v(s) = \Pi_{\mathcal{G}}^N \widehat{v} = \frac{1}{k} \sum_{i=1}^{k} \widehat{v}(s_{(i)}).$$

It is worthwhile to mention that for smoother average of the outputs, one could use kernel regression which also scale better with dimension.

Now, we specify the ERVL algorithm. We first sample N points from $\mathcal{X}$ uniformly (or according to another probability measure). Then, perform an 'empirical' value iteration step on these sampled points by obtaining mini-batches of next states. Then, we truncate the function at these points to ensure boundedness. Then, we do function-fitting using nearest neighbors, which gives us the next iterate of the value function.

Algorithm 1 ERVL

Input: sample sizes $N \geq 1$; $M \geq 1$; initial seed Q_0; total iterations $K \geq 1$.
For $k = 1, \ldots, K$

1) Sample $\{x_n\}_{n=1}^N$ from $\mathcal{X}$ uniformly
2) Value iteration: $\widehat{v}'_k(x_n) \leftarrow \widehat{T}_M v_{k-1}$ for $n = 1, 2, \ldots, N$
3) Truncation: $\widehat{v}_k(\cdot) = \Gamma \widehat{v}'_k(\cdot)$
4) Function fitting: $v_k \leftarrow \Pi_{\mathcal{G}}^N \widehat{v}_k$.
5) Increment $k \leftarrow k+1$ and return to Step 1.

We can now establish that the iterates of the algorithm, v_k are an (ϵ, δ)-weak probabilitic fixed of the operator $\widehat{G}_{N,M} = \Pi_{\mathcal{G}}^N \widehat{T}_M$ and hence a good approximation to v^*, the fixed point of T in the span semi-norm with high probability if N, M and k are large enough.

Theorem III.1. *Given $\epsilon, \delta > 0$, choose $\delta_1, \delta_2 > 0$ such that $\delta_1 + 2\delta_2 < \delta$. Let $\kappa^* = \left\lceil \frac{2 v_{\max}}{\epsilon} \right\rceil$ and μ_{min} is given by*

$$\mu_{min} = \min\left\{\delta_1^{\kappa^*}, (1-\delta_1), (1-\delta_1)\delta_1, \cdots, (1-\delta_1)\delta_1^{\kappa^*-1}\right\}.$$

Then for any $k \geq \log\left(\frac{1}{\delta_2 \mu_{min}}\right)$, $N \geq 2\left(\frac{16\, v_{\max} C}{\epsilon}\right)^{2d}\left[\log\frac{2}{\delta} + d\log\frac{32\, v_{\max} C}{\epsilon}\right]$ *and* $M \geq \frac{v_{\max}^2}{2(\epsilon/8)^2}\log\left(\frac{16\,|\mathbb{A}| v_{\max}}{\epsilon}\right)$, *we have*

$$P(span(v_k - v^*) > \epsilon) \leq \delta.$$

IV. Analysis: Proof of Theorem III.1

We now prove Theorem III.1. The analysis will proceed in three steps. First, we will bound the function

approximation error, next we bound the one-step error of the algorithm by viewing each iteration as a random operator, and then we use a stochastic dominance argument to argue convergence and get rate of convergence.

Error Analysis: Function Approximation with nearest neighbors.: We first define the *regression function* $f_M : \mathcal{X} \to \mathbb{R}$ via

$$f_M(x) \triangleq \mathbb{E}\left[\max_{a\in\mathbb{A}}\left\{r(x,a) + \frac{1}{M}\sum_{m=1}^{M} v(x_m)\right\}\right] \quad (2)$$

where x_m are the next generated given x and a. It is the expected value of our empirical estimator of Tv. As expected, $f_M \to Tv$ as $M \to \infty$. We note that f_M is not necessarily equal to Tv by Jensen's inequality.

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 $x \in \mathcal{X}$ through the choice of $M \geq 1$.

Lemma IV.1. *For any $\epsilon > 0$ and $M \geq 1$,*

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

Proof. For any $x \in \mathcal{X}$, we compute

$$\begin{aligned}
&|f_M(x) - Tv(x)| \\
&\leq \mathbb{E}[|\max_{a\in\mathbb{A}}\left\{r(x,a) + \frac{1}{M}\sum_{m=1}^{M} v(x_m)\right\} \\
&\quad - \max_{a\in\mathbb{A}}\left\{r(x,a) + \mathbb{E}_{x'\sim P(\cdot\,|\,x,a)}\left[v(x')\right]\right\}|] \\
&\leq \mathbb{E}\left[\max_{a\in\mathbb{A}}|\frac{1}{M}\sum_{m=1}^{M} v(x_m^{x,a}) - \mathbb{E}_{x'\sim P(\cdot\,|\,x,a)}\left[v(x')\right]|\right] \\
&\leq \left[\epsilon + 2\,|\mathbb{A}|\,v_{\max}\,\exp\left(\frac{-2\,M\,\varepsilon^2}{{v_{\max}}^2}\right)\right]
\end{aligned}$$

where the third one is due to Hoeffding's inequality. □

Let us denote the number of nearest neighbors as k_N since k denotes the iteration number. We can easily establish that $\Pi_{\mathcal{G}}^N$ in this case is a non-expansive mapping:

$$\|\Pi_{\mathcal{G}}^N \widehat{v}_1 - \Pi_{\mathcal{G}}^N \widehat{v}_2\|_\infty \leq \|\widehat{v}_1 - \widehat{v}_2\|_\infty. \quad (3)$$

Note that in the above equation if $\min \widehat{v}_1 = \min \widehat{v}_2 = 0$ then it also holds in span norm. The next lemma is from [5] which presents the rate of convergence in sup-norm for $k-$NN with an additional assumption of Lipschitz continuity.

Lemma IV.2. *If for any $v \in \widetilde{\mathcal{C}}(\mathcal{X})$, f_M is Lipschitz with constant C, then for $\delta, \epsilon > 0$, $k_N = \frac{N}{2}\left(\frac{\epsilon}{4v_{\max}C}\right)^d$ and*

$$N \geq N_0(\epsilon,\delta) = 2\left(\frac{8\,v_{\max}\,C}{\epsilon}\right)^{2d}\log\frac{2}{\delta}\left(\frac{16\,v_{\max}\,C}{\epsilon}\right)^{d}$$

then

$$\mathbb{P}(\|\widehat{G}_{N,M}v - f_M\|_\infty \geq \epsilon) \leq \delta.$$

The following lemma provides a probabilistic bound on the function approximation error.

Lemma IV.3. *Given $v \in \mathcal{C}_B(\mathcal{X})$, $\epsilon > 0$, and $\delta \in (0,1)$. Also choose $N \geq N_0(\epsilon,\delta)$ and $M \geq M_0(\epsilon) = \frac{v_{\max}^2}{2(\epsilon/4)^2}\log\left(\frac{8\,|\mathbb{A}|v_{\max}}{\epsilon}\right)$, Then we have*

$$\mathbb{P}(\|\widehat{G}_{N,M}\,v - Tv\|_\infty \geq \epsilon) \leq \delta.$$

Proof. By the triangle inequality,

$$\|\widehat{G}_{N,M}\,v - T\,v\|_\infty \leq \|\widehat{G}_{N,M}\,v - f_M\|_\infty + \|f_M - T\,v\|_\infty.$$

Note that $\widehat{G}_{N,M}$ minimizes the empirical loss over random samples. From lemma IV.2, if $N \geq N_0(\epsilon/2,\delta)$ then with probability $1-\delta$

$$\|\widehat{G}_{N,M}\,v - f_M\|_\infty < \epsilon/2.$$

From Lemma IV.1, if

$$M \geq \frac{v_{\max}^2}{2\,(\epsilon/4)^2}\log\left(\frac{8\,|\mathbb{A}|v_{\max}}{\epsilon}\right),$$

then $\|f_M - T\,v\|_\infty \leq \epsilon/2$. Hence $\|\widehat{G}_{N,M}\,v - T\,v\|_\infty \leq \epsilon$ with probability at least $1-\delta$ if N and M are chosen appropriately. □

One-step error analysis of the random operator.: We now analyze iteration of the random operator $\widehat{G}_{N,M} = \Pi_{\mathcal{G}}^N \widehat{T}_M$ to bound the error in one iteration. In particular, we are interested in analyzing the stochastic process $\{span(v_k - v^*)\}_{k\geq 0}$.

$$\begin{aligned}
span(v_k - v^*) &\leq span(\widehat{G}_{N,M}v_{k-1} - Tv_{k-1}) \\
&\quad + span(Tv_{k-1} - Tv^*) \\
&\leq span(\widehat{G}_{N,M}v_{k-1} - Tv_{k-1}) \\
&\quad + \alpha\, span(v_{k-1} - v^*)
\end{aligned}$$

As mentioned earlier, the composed operator $\widehat{G}_{N,M}$ is a probabilistic contraction in span norm. Let

$$p(\epsilon,N,M) \triangleq \mathbb{P}\left(span\left(G_{N,M}v - Tv\right) \leq \epsilon\right),\ \forall v \in \mathcal{C}_B(\mathcal{X}), \quad (4)$$

For compact notation, let us rewrite $p(\epsilon,N,M)$ as $p_{N,M}$. Note that if $span(v_{k-1} - v^*) = \eta\,\epsilon$ and if $span(G_{N,M}v_{k-1} - Tv_{k-1}) < \epsilon$, then using the previous

decomposition, we have $span(v_k - v^*) \leq (\alpha\eta + 1)\epsilon$ with probability $p_{N,M}$. Now let us define

$$\eta^* \triangleq \min\{\eta \in \mathbb{N} : \lceil \alpha\eta + 1 \rceil < \eta\} = \left\lceil \frac{2}{1-\alpha} \right\rceil$$

Now, to track the progress of the error process, we construct a Markov chain which stochastically dominates this process. The one-step probability of this chain depends on $p_{N,M}$ as it will be clear in the following section.

Let us now define the concept of stochastic dominance.

Definition IV.1. *Let X and Y be two random variables, then Y stochastically dominates X, written $X \leq_{st} Y$, when $\mathbb{P}(X \geq \theta) \leq \mathbb{P}(Y \geq \theta)$, for all θ in the support of Y.*

The iterates of Algorithm 1 are bounded in $span$-norm by $v_{\max}$. Choose $\epsilon > 0$ and let $\kappa^* = \lceil \frac{2v_{\max}}{\epsilon} \rceil$. Define the error process $\{X_k\}_{k\geq 0}$ as follows:

$$X_k = \begin{cases} \eta & \text{if} \quad \eta\epsilon \leq span(v_k - v^*) < (\eta+1)\epsilon \\ 0 & \text{if} \quad span(v_k - v^*) = 0 \end{cases}$$

Next, we construct a Markov chain which stochastically dominates the process $\{X_k\}_{k\geq 0}$. Similar to [7], we define the Markov chain as follows:

$$Y_k = \begin{cases} \max\{Y_{k-1} - 1, \eta^*\} & \text{w.p.} \quad p_{N,M} \\ \kappa^* \quad 1 - p_{N,M} \end{cases}$$

Note that the choice of $p_{N,M}$ is governed by the underlying function space and the samples N and M. Next we establish the stochastic dominance of the Markov chain $\{Y_k\}_{k\geq 0}$ over the error process $\{X_k\}_{k\geq 0}$ provided that both of these stochastic processes have the same initial state.

Lemma IV.4. *[7, Theorem 4.1] For all $k \geq 0$, $X_k \leq_{st} Y_k$ if $X_0 = Y_0$.*

Furthermore, process $\{Y_k\}_{k\geq 0}$ is a finite state and irreducible Markov chain and hence there exists a steady state distribution. The state space of this chain is $\{\eta^*, \eta^* + 1, \eta^* + 2, \ldots, \kappa^*\}$. Let μ denotes the steady state distribution of the Markov chain $\{Y_k\}_{k\geq 0}$. The steady state distribution can be computed as follows:

$$\mu(0) = \delta_1^{\kappa^*}, \quad \mu(\kappa^*) = 1 - \delta_1, \quad \mu(i) = (1-\delta_1)\delta_1^{\kappa^* - i}$$

$\forall \quad 0 < i < \kappa^*$ and let $\mu_{min} = \min_{0\leq i\leq \kappa^*} \mu(i)$.

We are now ready to prove the main theorem which relies on analyzing the dominating Markov chain $\{Y_k\}_{k\geq 0}$ and its mixing time.

Proof. First recall that $span(f) \leq 2\|f\|_\infty$. Since the one-step error is bounded for $N_0(\epsilon/2, \delta_1)$ and $M_0(\epsilon/2)$, we construct our Markov chain $\{Y_k\}_{k\geq 0}$ with $p_{N,M} = 1-\delta_1$. Note that in (4), we defined $p_{N,M}$ for a deterministic action-value function. But now, after iteration k, v_k is a random function. But because the samples generated are independent across iterations, one could use (4) to randomized functions. Let us denote the transition matrix for $\{Y_k\}_{k\geq 0}$ as P_Y given as follows:

$$P_Y = \begin{bmatrix} \delta_1 & 0 & 0 & \ldots & 0 & 1-\delta_1 \\ \delta_1 & 0 & 0 & \ldots & 0 & 1-\delta_1 \\ 0 & \delta_1 & 0 & \ldots & 0 & 1-\delta_1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \delta_1 & 1-\delta_1 \end{bmatrix}$$

The eigenvalues of P_Y are 0 and 1. Then, from [9, Theorem 12.3], we can bound the mixing time for any $\delta_2 > 0$ as

$$t_{mix}(\delta_2) \leq \left(\log \frac{1}{\delta_2 \mu^*} \right).$$

Using theorem 3.1 from [7], we conclude the theorem. □

V. Numerical Performance

We now show numerical performance on a benchmark problem of machine replacement. This problem has been studied for discounted setting [10]. We work out the details under average reward criterion. In this problem, the state space is non-negative real numbers and two actions are available in each state; keep the machine or replace it. Let the action of keeping the machine be denoted as 0 and replacement as 1. The transition dynamics are given as follows:

$$P(s'|s,0) = \begin{cases} \beta \exp\left(-\beta(s'-s)\right), & \text{if} \quad s' \geq s \\ 0, & \text{otherwise.} \end{cases}$$

$$P(s'|s,1) = \begin{cases} \beta \exp\left(-\beta s'\right), & \text{if} \quad s' \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

If we decide the keep the machine, we need to pay the maintenance cost which increases with state and for replacement, we need to pay a fixed amount. Hence, the reward function is given by $r(s,0) = -\alpha s$ and $r(s,1) = -C$.

The optimality equation is given as follows:

$$J^* + v^*(s) = \max\left(T_0, T_1\right)$$

where

$$T_0 = -\alpha s + \int_s^\infty \beta \exp\left(-\beta(s'-s)\right) v^*(s') ds'$$

and

$$T_1 = -C + \int_0^\infty \beta \exp\left(-\beta s'\right) v^*(s') ds'$$

One could guess that the optimal policy will be a threshold policy, i.e., there exists a $\bar{s}$ such that following holds:

$$\pi^*(s) = \begin{cases} 0, & \text{if} \quad s \leq \bar{s} \\ 1, & \text{otherwise.} \end{cases}$$

For $s \in [0, \bar{s}]$,

$$J^* + v^*(s) = -\alpha s + \int_s^\infty \beta \exp\left(-\beta(y-s)\right) v^*(y) dy$$

Differentiating both sides, we have

$$\begin{aligned}(v^*)'(s) &= -\alpha + \beta^2 \exp(\beta s) \int_s^\infty \exp\left(-\beta y\right) v^*(y) dy \\ &\qquad - \beta v^*(s) \\ &= -\alpha + \alpha\beta s + \beta J^*.\end{aligned}$$

Recall that for $s \geq \bar{s}$, the value function does not depend on s. Hence for $s = \bar{s}$, we have

$$\begin{aligned}J^* + v^*(\bar{s}) &= -\alpha\bar{s} + \beta \int_{\bar{s}}^\infty \exp\left(-\beta(y-\bar{s})\right) v^*(y) dy \\ &= -\alpha\bar{s} + v^*(\bar{s}).\end{aligned}$$

Hence, $J^* = -\alpha\bar{s}$. To compute $\bar{s}$, we need to solve the following equation:

$$\begin{aligned}&\int_0^{\bar{s}} \beta \left(\frac{\alpha\beta}{2} s^2 - \alpha s(1 + \beta\bar{s}) \right) \exp(-\beta s) ds \\ &\quad + \int_{\bar{s}}^\infty \beta \left(-\alpha\bar{s} - \frac{\alpha\beta}{2} \bar{s}^2 \right) \exp(-\beta s) ds \\ &\quad + 2\alpha\bar{s} + \frac{\alpha\beta}{2}\bar{s}^2 - C = 0\end{aligned}$$

For our experiments, we use $\beta = 2/3, \alpha = 3, C = 15$. This gives the optimality policy as $\pi^*(s) = 0$ if $s <= 2.654$,otherwise 1 and $J^* = -7.962$. Note that one could use any reference state instead of minimum of the function. For each iteration k one can compute π_k which is a greedy policy with respect to Q_k. We used Gaussian kernel $K(x, y) = \exp\left(-\gamma(x-y)^2\right)$. Fig. 1 presents the error $|J^{\pi_k} - J^*|$ against iteration kernel regression with regularization for $N = 200$.

VI. CONCLUSIONS

In this paper, we proposed an empirical relative value learning (ERVL) algorithm combined with non-parametric function approximation. In particular, we focused on nearest neighbors regression. The framework developed in this paper can be extended to any non-parametric setting as long as the function approximation is non-expansion and convergence to regression function can be established in sup-norm. Then, one can bound the one-step error and use the stochastic dominance argument to establish convergence.

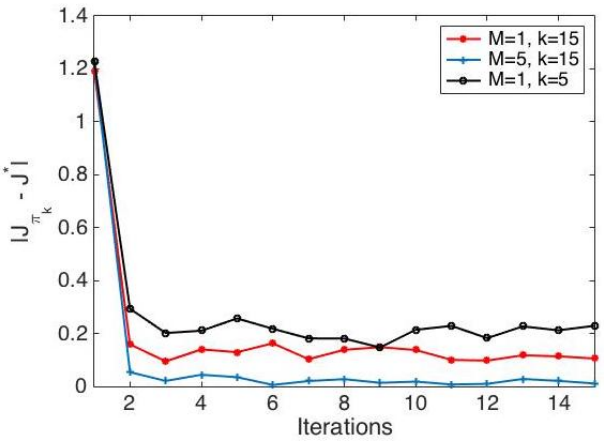

Fig. 1: Optimal replacement with nearest neighbors

REFERENCES

[1] Aristotle Arapostathis, Vivek S Borkar, Emmanuel Fernández-Gaucherand, Mrinal K Ghosh, and Steven I Marcus. Discrete-time controlled markov processes with average cost criterion: a survey. *SIAM Journal on Control and Optimization*, 31(2):282–344, 1993.

[2] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[3] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 2.

[4] Ronald A DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998.

[5] Luc Devroye. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory*, 24(2):142–151, 1978.

[6] A. Gupta, R. Jain, and P. W. Glynn. An empirical algorithm for relative value iteration for average-cost mdps. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5079–5084, Dec 2015.

[7] William B Haskell, Rahul Jain, and Dileep Kalathil. Empirical dynamic programming. *Mathematics of Operations Research*, 41(2):402–429, 2016.

[8] Onésimo Hernández-Lerma. *Adaptive Markov control processes*, volume 79. Springer Science & Business Media, 2012.

[9] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

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

[11] Dirk Ormoneit and Śaunak Sen. Kernel-based reinforcement learning. *Machine learning*, 49(2-3):161–178, 2002.

[12] Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2007.

[13] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.