Approximate Relative Value Learning for Average-reward Continuous State MDPs

Hiteshi Sharma **Mehdi Jafarnia-Jahromi** **Rahul Jain**
Electrical and Computer Engineering
University of Southern California, Los Angeles, USA
{hiteshis,mjafarni,rahul.jain}@usc.edu

Abstract

In this paper, we propose an approximate relative value learning (ARVL) algorithm for non-parametric MDPs with continuous state space and finite actions and average reward criterion. It is a sampling based algorithm combined with kernel density estimation and function approximation via nearest neighbors. The theoretical analysis is done via a random contraction operator framework and stochastic dominance argument. This is the first such algorithm for continuous state space MDPs with average reward criteria with these provable properties which does not require any discretization of state space as far as we know. We then evaluate the proposed algorithm on a benchmark problem numerically.

1 INTRODUCTION

Markov Decision Processes (MDPs) are a suitable framework for sequential decision making under uncertainty [14]. For infinite-horizon MDPs with discounted reward criterion, there are various dynamic programming algorithms available [2]. For continuous state space, different techniques like state aggregation and function approximation have been proposed [3, 11].

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]. Unlike discounted setting where the contraction parameter is the discount factor, in average reward criterion the contraction parameter depends on the dynamics of MDP. There has been an extensive study in the literature on the existence and structural properties of optimal policies. But computing such policies is generally a challenging problem when state space is uncountable. One idea is to quantize the state spaces with a finite grid and constructing a reduced discrete model with a new transition probability and reward function. For example, in [13], a meta-MDP is constructed through state-aggregation method. Similarly, [12] constructs an 'artificial' MDP using kernel averaging. In [15], a discrete MDP is constructed for MDPs with unbounded reward and continuous state and action space. But unlike them, we do not discretize state space but instead propose a sampling based algorithm. Furthermore, we consider the scenario where we do not know the transition kernel of the MDP. we only have access to samples generated by this distribution. We define an approximate Bellman operator which uses the density estimated by these samples. Another key element in our algorithm is non-parametric function approximation. Although we work with nearest neighbors in this paper, it can be extended to other non-parametric function fitting methods as long as long it is non-expansive and uniform convergence is obtained. For nearest neighbors, [5] provides a uniform convergence of under Lipschitz continuity assumption of the regression function. This requires us to have MDPs with Lipschitz continuous transition and reward function.

Our theoretical analysis is based on the idea of viewing each iteration of the algorithm as application of a random operator. The notion of *probabilistic contraction* and *probabilistic fixed points* have been introduced in [8, 6, 9, 16]. In particular, [6, 16] uses truncation which requires the knowledge of contraction coefficient. This may not be easy to compute for continuous state spaces. The convergence was argued via construction of a Markov chain that stochastically dominates the norm of the error introduced due to the approximation. Since, they were either working with discounted setting or had a truncation operator, the iterates of the algorithm were bounded. Hence, the dominant Markov chain was on a finite state space for which the invariant distribution was easy to analyze. In [7] extended this to unbounded iterates by constructing the Markov chain on the set of natural num-

bers and then analyzing the invariant distribution under some conditions. This enables us to plug in their argument for convergence of random contraction to probabilistic fixed point for ARVL.

The main contribution of this paper is to introduce an off-policy approximate, (relative) value learning algorithm for computing optimal policies for non-parametric MDPs with continuous state space and average reward criterion when the transition kernel is unknown. We do not discretize the state space and work with a reduced model. Instead, we propose a sampling-based algorithm. We also provide theoretical guarantee for the proposed algorithm under the random operator framework which can easily be extended to other regression techniques if they have non-expansive property and uniform convergence.

The rest of the paper is organized as follows. Section 2 first presents the conditions for existence of optimal policies, then introduces the approximate Bellman operator followed by definition of probabilistic contraction. The algorithm combining the approximate operator with function approximation is presented in Section 3. The theoretical analysis is then presented in Section 4 followed by numerical experiments.

2 PRELIMINARIES

2.1 Exact Bellman operator

Consider an MDP $(\mathcal{X}, \mathcal{A}, r, P)$ where $\mathcal{X}$ is the state space and $\mathcal{A}$ is the action space, $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the reward function and P is the transition kernel . We assume that $\mathcal{X}$ is a compact subset in $\mathbb{R}^d$ and $\mathcal{A}$ is finite. Let $\mathcal{B}(\mathcal{X})$ be the endowed Borel sigma-algebra on $\mathcal{X}$. Let $\mathcal{C}(\mathcal{X})$ be the set of continuous and bounded functions over $\mathcal{X}$. For each $f \in \mathcal{C}(\mathcal{X})$, define

$$\|f\|_{\text{Lip}} = \sup_{(x,y)\in\mathcal{X}\times\mathcal{X}} \frac{|f(y) - f(x)|}{\|y - x\|}.$$

$\text{Lip}(\mathcal{X})$ denotes the set of all Lipschitz continuous functions on $\mathcal{X}$, i.e.,

$$\text{Lip}(\mathcal{X}) = \{f \in \mathcal{C}(\mathcal{X}) : \|f\|_{\text{Lip}} < \infty\}.$$

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 $P(X_{t+1} \in B|X_t = x, a_t = a)$. 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 $J^*(x) = \sup_\pi J^\pi(x)$. 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 2.1. *(a) For every (x, a), $|r(x, a)| \le r_{\max}$ and for every a, $r(\cdot, a)$ is Lipschitz continuous.*

(b) For every $a \in \mathcal{A}$, transition kernel $P(\cdot|x, a)$ has a positive Radon-Nikodym derivative, $p(y|x, a)$ with respect to Lebesgue measure, λ on $\mathbb{R}^d$, for all $x, y \in \mathbb{R}^d$.

(c) The transition probability density is Lipschitz continuous in the present state, i.e, for all $a \in \mathcal{A}$ and $x, y, z \in \mathcal{X}$, there exists $L'_p(z)$ such that

$$|p(z|x, a) - p(z|y, a)| \le L'_p(z)\|x - y\|$$

where $\int_{\mathcal{X}} L'_p(z)\lambda(dz) = L_p$.

(d) There exists $\alpha < 1$ such that

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

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

Assumption 2.1 (a) establishes that for every a, $r(\cdot, a) \in \text{Lip}(\mathcal{X})$, (b) and (c) imply that if $v \in \text{Lip}(\mathcal{X})$ then for any action a, $\int v(y)P(dy|\cdot, a) \in \text{Lip}(\mathcal{X})$ and (d) 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 $(J^*, v^*) \in \mathbb{R} \times \text{Lip}(\mathcal{X})$ 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 $T : \text{Lip}(\mathcal{X}) \to \text{Lip}(\mathcal{X})$ 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.

Iteration on a Quotient Space. Let us now define the span semi-norm and the quotient space. For a function $f \in \mathcal{C}(\mathcal{X})$, $span(f) := \sup_{(x)} f(x) - \inf_x f(x)$. 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 $\mathcal{C}(\mathcal{X})$ 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 $\widetilde{\mathcal{C}}(\mathcal{X}) = \mathcal{C}(\mathcal{X})/\sim$ be the quotient space. The following then is not difficult to show for the quotient space.

Lemma 2.1. *[6] $(\widetilde{\mathcal{C}}(\mathcal{X}), span)$ is a Banach space.*

The proof is given in the appendix. Furthermore, we can show that the operator T is a contraction in the span semi-norm. The next theorem is from [10].

Theorem 2.2. *[10] Suppose that Assumptions 2.1 hold. Then, operator T : Lip$(\mathcal{X})$ $\rightarrow$ Lip$(\mathcal{X})$ is a span-contraction operator, i.e.,*

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

where $v_1, v_2 \in$ Lip$(\mathcal{X})$ and α is defined in Assumption 2.1(c).

Now consider a $v \in \text{Lip}(\mathcal{X})$, and let $\tilde{v}$ be the corresponding element in $\widetilde{\text{Lip}}(\mathcal{X})$ and $\widetilde{T} : \widetilde{\text{Lip}}(\mathcal{X}) \rightarrow \widetilde{\text{Lip}}(\mathcal{X})$ defined as $\widetilde{T}\tilde{v} = \widetilde{Tv}$. Since T is a span-contraction, then so is $\widetilde{T}$ which by Banach fixed point theorem 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 original space.

2.2 Approximate Bellman operator

In this paper, we assume that the transition kernel is unknown but for a given state-action pair, we can get samples of the next state from the generative model. Using these samples, we approximate the dynamics by non-parametric density estimation. We begin with a smoothing kernel $K : \mathcal{X} \rightarrow \mathbb{R}$ defined as any smooth function such that $\int K(x)dx = 1$, $\int xK(x)dx = 0$ and $\int x^2K(x)dx < \infty$. Assume that for any $(x, a) \in \mathcal{X} \times \mathcal{A}$, we have access to M independent and identically distributed samples $Y_i^{x,a} \sim P(\cdot|x, a)$, $i = 1, 2, \ldots M$. Let h_M be the bandwidth, then the kernel density estimator is defined as

$$\widehat{p}_M(y|x, a) = \frac{1}{Mh_M^d} \sum_{i=1}^{M} K\left(\frac{y - Y_i^{x,a}}{h_M}\right).$$

For instance, the kernels commonly used are the Gaussian kernel, $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-\|x\|^2/2)$ and tophat kernel, $K(x) = \frac{1}{2}\mathbb{I}(\|x\| < 1)$ where $\mathbb{I}$ is an indicator function. In this paper, we focus on Gaussian kernels so that the Lipschitz property is preserved. The bandwidth, h_M controls the smoothness of estimation and hence, needs to be chosen carefully. Let the estimated distribution be $\widehat{P}_M$. Let us now define our approximate Bellman operator $\widehat{T}_M : \text{Lip}(\mathcal{X}) \rightarrow \text{Lip}(\mathcal{X})$ as follows:

$$\widehat{T}_M v(x) = \max_{a \in \mathcal{A}} \left[r(x, a) + \mathbb{E}_{x' \sim \widehat{P}_M(\cdot|x,a)} v(x') \right].$$

Clearly, $\widehat{T}_M$ is a *random operator*. Let $\widehat{\alpha}_M$ be the random variable defined as

$$\sup_{(x,a),(x',a')} \|\widehat{P}_M(\cdot|x, a) - \widehat{P}_M(\cdot|x', a')\|_{TV} = 2\widehat{\alpha}_M.$$

Then one can show that for all $v_1, v_2 \in \text{Lip}(\mathcal{X})$

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

We analyze *probabilistic contraction* of the approximate Bellman operator, $\widehat{T}_M$ by arguing that $\widehat{\alpha}_M < 1$ with high probability (as presented in detail in Section 4).

Let us also compare our approximate Bellman operator to empirical Bellman operator defined in [6, 9]. In their case, the estimated distribution for a state-action pair (x, a) is given by

$$\widehat{P}_M(dy|x, a) = \frac{1}{M} \sum_{i=1}^{M} \delta_{Y_i}(dy)$$

where Y_i are iid samples from the true distribution for $i = 1, \ldots M$ and δ is the delta function. It can be shown that for empirical distribution $\widehat{\alpha}_M$ will always be equal to 1. To see that, fix (x, a) and (x', a'), then

$$\|\widehat{P}_M(\cdot|x, a) - \widehat{P}_M(\cdot|x', a')\|_{TV} = 2 \sup_{B \subset \mathcal{B}(\mathcal{X})} \left| \widehat{P}_M(B|x, a) - \widehat{P}_M(B|x', a') \right|$$

Since M is finite, one can choose B such that the difference between two distributions is always 1.

2.3 Function approximation using n nearest neighbors

Let us now define a function space $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathbb{R}\}$. Let $\Pi_{\mathcal{F}}$ be the function approximation operator which maps a bunch of samples to a function in the space $\mathcal{F}$. While various non-parametric function spaces can be considered, we will choose nearest neighbors (NN) for function approximation (other non-parametric function approximation methods for e.g., kernel regression, etc. will also work). $n-$NN is a powerful yet simple approach in non-parametric regression. Suppose that we have N samples, $\{(x_i, \widehat{f}(x_i))\}_{i=1}^N$. 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 n nearest neighbors and estimate the function as

$$\left[\Pi_{\mathcal{F}}^n \widehat{f}\right](x) = \frac{1}{n} \sum_{i=1}^{n} \widehat{f}(x_{(i)}).$$

Thus, it allows to reconstruct a function from some finite samples. Note that the function approximation operator $\Pi_{\mathcal{F}}^n$ depends on both the sample size and number of nearest-neighbors used. Moreover, since this is an averaging operator, we can argue that this is non-expansive mapping with respect to sup-norm.

2.4 Random contraction

In this section, we introduce the definition of random contraction operator and fixed points in probabilistic sense as provided in [8, 7]. Consider a function space $\mathcal{F}$ with norm $\|\cdot\|$. Suppose there is an contraction operator $H : \mathcal{F} \to \mathcal{F}$. Let $\widehat{H}_N$ be the approximation of operator H via finite samples N.

Definition 2.1. *An operator $\widehat{H}_N : \mathcal{F} \to \mathcal{F}$ is said to be a random contraction operator with respect to norm $\|\cdot\|$ if there exists a random variable $\widehat{\beta}$ such that $\widehat{\beta} < 1$ with high probability and the following holds for all $f, g \in \mathcal{F}$:*

$$\|\widehat{H}_N f - \widehat{H}_N g\| \leq \widehat{\beta}\|f - g\|$$

Suppose that $\widehat{H}_N^k$ denote that the iteration of operator k times, we now define a weak probabilistic fixed point.

Definition 2.2. *A function $f \in \mathcal{F}$ is a weak probabilistic fixed point for a sequence of random operators $\{\widehat{H}_N\}$ with respect to a given norm $\|\cdot\|$ if*

$$\lim_{N\to\infty} \lim_{k\to\infty} \mathbb{P}(\|\widehat{H}_N^k f_0 - f\| > \epsilon) = 0$$

for all $f_0 \in \mathcal{F}$.

Based on the above definition, one can also define an (ϵ, δ)-weak probabilistic fixed point.

3 ALGORITHM AND THE MAIN RESULT

We now present the Approximate Relative Value Learning (ARVL) algorithm, a non-parametric off-policy algorithm for MDPs with continuous state space. It is a sampling-based algorithm combined with non-parametric density estimation and function approximation. It first samples the state space uniformly and estimates the probability density for each sampled state and action. Then, the approximate Bellman operator gives samples of value function which are then used for regression.

Recall that in relative value iteration, there is a bias subtraction at each iteration. This does not change the span norm but keeps the iterates bounded. In our algorithm, we make our samples for regression non-negative by subtracting the minimum of the function. In other words, we are choosing a non-negative optimal value function as it is not unique. This makes the samples for regression non-negative. Let the number of state samples be N, next state samples be M and number of neighbors for function approximation be n. Let $\Gamma_N : \mathbb{R}^N \to \mathbb{R}^N$ be an operator such that

$$\Gamma_N \widehat{v}' = \widehat{v}' - \min \widehat{v}' \, \mathbb{1}_N$$

where $\mathbb{1}_N$ is a vector of all ones of size N. Let us denote the composed operator by $\widehat{G}(N, M, n) = \Pi_{\mathcal{F}}^n \Gamma_N \widehat{T}_M$ where we use the fact that the function approximation depends both on N and n. Algorithm 1 will iterate the random operator $\widehat{G}(N, M, n)$ (or just $\widehat{G}$ for compact notation), i.e., $v_{k+1} = \widehat{G}\, v_k = \widehat{G}^k\, v_0$. Using the non-expansive property of NN regression, we will establish that the composed operator $\widehat{G}$ is a contraction with high probability.

Now, we specify the ARVL algorithm in detail. We first sample N points from $\mathcal{X}$ uniformly (or according to another probability measure). Then, perform an 'approximate' value iteration step on these sampled points by estimating the density via mini-batches of next states. Then, we do function-fitting using nearest neighbors, which gives us the next iterate of the value function.

Algorithm 1 ARVL

Input: sample sizes $N \geq 1$; $M \geq 1$; $n \geq 1$; initial seed v_0; total iterations $K \geq 1$.
For $k = 1, \ldots, K$

1. Sample $\{x_i\}_{i=1}^N$ from $\mathcal{X}$ uniformly
2. Kernel density estimation $\widehat{p}_M(\cdot|x_i, a)$ for each $i = 1, 2, \ldots, N$ and $a \in \mathcal{A}$
3. Approximate value iteration: $\widehat{v}'_k(x_i) \leftarrow \widehat{T}_M v_{k-1}$, $\widehat{v}_k(x_i) \leftarrow \widehat{v}'_k(x_i) - \min_{x_j} \widehat{v}'_k$ for $i, j = 1, 2, \ldots, N$
4. Function fitting: $v_k \leftarrow \Pi_{\mathcal{F}}^n \widehat{v}_k$.
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 weak probabislitic fixed of the operator $\widehat{G}(N, M, n) = \Pi_{\mathcal{F}}^n \Gamma_N \widehat{T}_M$ 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 3.1. *Suppose that Assumptions 2.1 and 4.1 hold. Given $\epsilon, \delta > 0$, there exist constants B and C such that for any*

$$N \geq 2\left(\frac{8\,B\,C}{\epsilon}\right)^{2d} \log\frac{2}{\delta}\left(\frac{16\,B\,C}{\epsilon}\right)^{d} \quad \text{and} \quad n \geq \frac{N}{2}\left(\frac{\epsilon}{4\,B\,C}\right)^{d},$$

we have

$$\lim_{M\to\infty} \lim_{k\to\infty} P(span(v_k - v^*) > \epsilon) \leq \delta.$$

Note that the nearest neighbors scale very poorly with dimension which is reflected in our bounds. Furthermore,

the dependence on next state sample size, M, is due to asymptotic convergence of kernel density estimation.

4 ANALYSIS: PROOF OF THEOREM 3.1

We now prove Theorem 3.1. There are two approximations in ARVL: one is due to density estimation and another is due to function fitting. We first bound the error due to these approximations. As mentioned before, each iteration of ARVL can be viewed as iteration of a random operator, we then bound the error in one iteration. In the end, we use a stochastic dominance argument to argue convergence.

Error due to density estimation. We first want to establish that when M is large enough, $\widehat{\alpha}_M < 1$ with high probability. Let us now recall that L_1 distance between any two densities μ and ν over $\mathcal{X}$ is given as:

$$\|\mu - \nu\|_1 = \int_{\mathcal{X}} |\mu(x) - \nu(x)|\, dx$$

If we can bound the L_1 norm, we get a bound on total-variation norm as well since if $\int |\mu - \nu| dx < \delta$ then $|\mu(B) - \nu(B)| < \delta$ for all B. Next, we present convergence of estimated density to the true density in L_1 norm as shown in [4] which needs the following assumptions:

Assumption 4.1. *1. Let $K : \mathbb{R}^d \to \mathbb{R}$ such that $\int K(x)dx = 1$ and $L(u) = \sup_{\|x\| \geq u} K(x)$ for $u \geq 0$.*

2. h_M is a sequence of positive numbers such that $h_M \to 0$ and $M\, h_M^d \to \infty$ as $M \to \infty$.

3. The density $p(\cdot|x, a)$ is almost everywhere continuous for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $\|x\|^d\, K(x) \to 0$ as $\|x\| \to \infty$.

Theorem 4.1. *Let K be a smoothing kernel such that Assumption 4.1 holds then the following holds with probability* 1*,*

$$\lim_{M \to \infty} \|p(\cdot|x, a) - \widehat{p}_M(\cdot|x, a)\|_1 = 0$$

for all $(x, a) \in \mathcal{X} \times \mathcal{A}$.

This now leads to the following lemma:

Lemma 4.2. *Assume that Assumption 4.1 holds then for any $\delta \in (0, 1 - \alpha)$,*

$$\lim_{M \to \infty} \mathbb{P}(\widehat{\alpha}_M \geq 1 - \delta) = 0$$

Proof. The proof is a direct application of Theorem 4.1. For any $(x, a), (x', a') \in \mathcal{X} \times \mathcal{A}$,

$$\begin{aligned}
&\|\widehat{P}_M(\cdot|x, a) - \widehat{P}_M(\cdot|x', a')\|_{TV} \\
&\qquad \leq \|\widehat{P}_M(\cdot|x, a) - P(\cdot|x, a)\|_{TV} \\
&+ \|P(\cdot|x', a') - \widehat{P}_M(\cdot|x', a')\|_{TV} + \|P(\cdot|x, a) - P(\cdot|x', a')\|_{TV}
\end{aligned}$$

Using ergodicity of transition kernel as mentioned in assumption 2.1 (d) and Theorem 4.1, we conclude the lemma. □

Error due to function approximation with nearest neighbors. In the previous section, we had defined Γ_N for vectors in $\mathbb{R}^N$ but it can be extended to $\mathcal{X}$ as $\Gamma : \text{Lip}(\mathcal{X}) \to \text{Lip}(\mathcal{X})$ defined as $\Gamma\, f = f - \min f$. Let $g_M : \mathcal{X} \to \mathbb{R}$ be such that

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

for any continuous value function $v \in \text{Lip}(\mathcal{X})$. Now, we define $f_M : \mathcal{X} \to \mathbb{R}$ and $\widetilde{f}_M : \mathcal{X} \to \mathbb{R}$ via

$$f_M\,(x) = \mathbb{E}\,[g_M] \quad \text{and} \quad \widetilde{f}_M\,(x) = \mathbb{E}\,[\Gamma\, g_M]$$

$\widetilde{f}_M$ is the *regression function*. It is the expected value of our approximate 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 $x \in \mathcal{X}$ when M is large enough.

Lemma 4.3. *Under Assumption 4.1, the following holds*

$$\lim_{M \to \infty} \|f_M - T\, v\|_\infty = 0$$

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

$$\begin{aligned}
&|f_M\,(x) - Tv\,(x)\,| \\
&\leq \mathbb{E}\left[\left| \max_{a \in \mathbb{A}} \left\{ r\,(x,\, a) + \mathbb{E}_{x' \sim \widehat{P}_M(\cdot|x,a)} v\,(x') \right\} \right.\right. \\
&\qquad \left.\left. - \max_{a \in \mathbb{A}} \left\{ r\,(x,\, a) + \mathbb{E}_{x' \sim P(\cdot\,|\,x,\,a)}\,[v\,(x')] \right\} \right|\right] \\
&\leq \mathbb{E}\left[\max_{a \in \mathbb{A}} \left| \mathbb{E}_{x' \sim \widehat{P}_M(\cdot|x,a)} v\,(x') - \mathbb{E}_{x' \sim P(\cdot\,|\,x,\,a)}\,[v\,(x')] \right| \right]
\end{aligned}$$

Note that the value function v is a continuous function on a compact set $\mathcal{X}$ hence $\sup_{x \in \mathcal{X}} v(x) = \|v\|_\infty < \infty$. Let

the action which maximizes the inner term be a_x^* then by Jenson's and Cauchy-Schwartz inequalities we have

$$\lim_{M\to\infty} |f_M(x) - Tv(x)| \leq \|v\|_\infty \lim_{M\to\infty} \mathbb{E}\Big[\int_{\mathcal{X}} |\widehat{p}_M(x'|x,a_x^*) - p(x'|x,a_x^*)|\lambda(dx')\Big]$$

Using bounded convergence theorem and Theorem 4.1, the proof concludes. □

The next lemma is from [5] which presents the rate of convergence in sup-norm for nearest neighbor regression.

Lemma 4.4. *Suppose for a value function $v \in Lip(\mathcal{X})$, there exist constants B and C such that $\|v\|_\infty < B$ and the regression function f_M is Lipschitz with constant C for any M, then for δ, $\epsilon > 0$, $n \geq n_0(\epsilon) = \frac{N}{2}\left(\frac{\epsilon}{4\,B\,C}\right)^d$ and*

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

we have

$$\lim_{M\to\infty} \mathbb{P}(\|\widehat{G}\,v - \widetilde{f}_M\|_\infty \geq \epsilon) \leq \delta.$$

One-step error analysis of the random operator. The following lemma provides a probabilistic bound on the one-step error of the ARVL, which points out that the error in one iteration can be controlled if the samples are sufficiently large.

Lemma 4.5. *Given $v \in Lip(\mathcal{X})$, $\epsilon > 0$, and $\delta \in (0,1)$. Also choose $N \geq N_0(\epsilon,\delta)$ and $n \geq n_0(\epsilon)$, Then we have*

$$\lim_{M\to\infty} \mathbb{P}(span(\widehat{G}\,v - Tv) \geq \epsilon) \leq \delta.$$

Proof. By the triangle inequality,

$$span(\widehat{G}\,v - T\,v) \leq span(\widehat{G}\,v - \widetilde{f}_M) + span(f_M - T\,v)$$

where the last inequality follows from the fact that $span(\widetilde{f}_M - f_M) = span(\mathbb{E}[\Gamma g_M - g_M]) = 0$. From Lemma 4.4, if $n \geq n_0(\epsilon)$ and $N \geq N_0(\epsilon,\delta)$ then with probability $1-\delta$, $\|\widehat{G}\,v - \widetilde{f}_M\|_\infty < \epsilon$. Combining with Lemma 4.3 concludes the proof. □

Next we establish that it is indeed a random contraction.

Lemma 4.6. *For a given $N, M, n \geq 1$, the operator $\widehat{G}(N,M,n) = \Pi_{\mathcal{F}}^n\,\Gamma_N\,\widehat{T}_M$ is a random contraction operator, i.e, for any $v_1, v_2 \in Lip(\mathcal{X})$,*

$$span(\widehat{G}\,v_1 - \widehat{G}\,v_2) \leq \widehat{\alpha}_M\, span(v_1 - v_2)$$

where $\widehat{\alpha}_M$ is a the random contraction coefficient.

Proof. Since we use n-NN for function fitting, we can easily establish that $\Pi_{\mathcal{F}}^n$ in this case is a non-expansive mapping:

$$\|\Pi_{\mathcal{F}}^n\,\widehat{v}_1 - \Pi_{\mathcal{F}}^n\,\widehat{v}_2\|_\infty \leq \|\widehat{v}_1 - \widehat{v}_2\|_\infty.$$

Note that in the above equation if $\min\widehat{v}_1 = \min\widehat{v}_2 = 0$ then it also holds in span norm. Hence,

$$span(\Pi_{\mathcal{F}}^n\,\Gamma_N\,\widehat{v}_1 - \Pi_{\mathcal{F}}^n\,\Gamma_N\,\widehat{v}_1) = span(\widehat{v}_1 - \widehat{v}_2)$$

Since $\widehat{G}$ is a composition of a non-expansive mapping with a (random) contraction, i.e, for any $v_1, v_2 \in \text{Lip}(\mathcal{X})$

$$\begin{aligned} span(\Pi_{\mathcal{F}}^n\,\Gamma_N\,\widehat{T}_M\,v_1 - \Pi_{\mathcal{F}}^n\,\Gamma_N\,\widehat{T}_M\,v_2) \\ = span(\widehat{T}_M\,v_1 - \widehat{T}_M\,v_2) \\ \leq \widehat{\alpha}_M span(v_1 - v_2) \end{aligned}$$

□

Stochastic Dominance. The following lemma is from [7] which enables us to analyze iteration of the composed operator.

Theorem 4.7. *Assume that the following holds:*

1. *$T : Lip(\mathcal{X}) \to Lip(\mathcal{X})$ is a contraction operator in span norm with contraction coefficient $\alpha < 1$.*
2. *For any $v \in Lip(\mathcal{X})$, we have*

$$\lim_{M,N,n\to\infty} \mathbb{P}(span(\widehat{G}\,v - Tv) \geq \epsilon) = 0.$$

3. *Let $\widehat{\alpha}_M$ be the contraction coefficient of $\widehat{G}$ such that for $\delta \in (0, 1-\alpha)$,*

$$\lim_{M\to\infty} \mathbb{P}(\widehat{\alpha}_M \geq 1-\delta) = 0.$$

4. *There exists $w > 0$ such that $span(\widehat{G}\,v^* - T\,v^*) \leq w$ almost surely.*

Then, v^* is weak probabilistic fixed point of random operator $\widehat{G}(N,M,n)$.

Sketch of the proof: The key element in the proof is stochastic dominance of a Markov chain (over natural

numbers) on the error process $\{span(v_k - v^*)\}_{k\geq 0}$. Recall that $v_k = \widehat{G}\, v_{k-1}$, we decompose the process as

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

For for all $v \in \text{Lip}(\mathcal{X})$, let us now define for $\epsilon > 0, \delta \in (0, 1-\alpha), n, N, M \geq 1$,

$$q(\epsilon, \delta, N, M, n) \triangleq \mathbb{P}\Big(\widehat{\alpha}_M \leq 1 - \delta, \quad span\left(\widehat{G}\, v - T\, v\right) \leq \epsilon \Big), \tag{2}$$

which we will denote by q. By Hoeffding-Frechet bound,

$$q \geq \mathbb{P}\left(\widehat{\alpha}_M \leq 1 - \delta\right) + \mathbb{P}\left(span\left(\widehat{G}\, v - T\, v\right) \leq \epsilon\right) - 1$$

Fix $\kappa > 0$, $\epsilon \in (0, \kappa/2]$, $\delta \in (0, 1-\alpha)$ such that $\eta = \lceil 2/\delta \rceil \leq \kappa/\epsilon$, a Markov chain is constructed over natural numbers as follows:

$$Y_k = \begin{cases} \eta & \text{w.p.} \quad q \text{ if } Y_k = \eta \\ Y_{k-1} & \text{w.p.} \quad q \text{ if } Y_k \geq \eta + 1 \\ Y_{k-1} + \lceil w/\epsilon \rceil & \text{w.p.} \quad 1 - q \end{cases}$$

The next step is to show that this Markov chain stochastically dominates the error process. Let us first define stochastic dominance:

Definition 4.1. *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.*

This yields for any $t > 0$,

$$\mathbb{P}\left(Y_k \geq t\right) \geq \mathbb{P}\left(span(v_k - v^*) \geq t\right)$$

Now it remains to show that the Markov chain admits an invariant distribution which concentrates at state 1 when the samples are sufficiently high.

Proof of theorem 3.1. Now we apply Theorem 4.7. Note that the first assumption in the theorem is satisfied by the ergodicity condition assumed in Assumption 2.1. The second and the third assumptions are satisfied by Lemma 4.5 and Lemma 4.2 respectively. The fourth one follows from bounded rewards and the fact that v^* is a fixed point of operator T. Hence, Theorem 4.7 can be applied to conclude the convergence. □

5 NUMERICAL PERFORMANCE

We now show numerical performance on a benchmark problem of machine replacement. This problem has been studied for discounted setting [11]. 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:

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

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 $r(x,0) = -\alpha x$ and $r(x,1) = -C$.

The optimality equation is given as follows:

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

where

$$T_0 = -\alpha x + \int_x^\infty \beta \exp\left(-\beta(x'-x)\right) v^*(x') dx'$$

and

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

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

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

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

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

Differentiating both sides, we have

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

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

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

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

$$\int_0^{\bar{x}} \beta \left(\frac{\alpha\beta}{2} x^2 - \alpha x(1+\beta\bar{x}) \right) \exp(-\beta x) dx + \int_{\bar{x}}^{\infty} \beta \left(-\alpha\bar{x} - \frac{\alpha\beta}{2}\bar{x}^2 \right) \exp(-\beta x) dx + 2\alpha\bar{x} + \frac{\alpha\beta}{2}\bar{x}^2 - C = 0$$

Once we have $\bar{x}$, we can compute the optimal value function as

$$v^*(x) = \begin{cases} -\alpha(1+\beta\bar{x}) + \frac{\alpha\beta}{2}x^2 & \text{if} \quad 0 \le x \le \bar{x} \\ -\alpha\bar{x} - \frac{\alpha\beta}{2}\bar{x}^2 & \text{otherwise.} \end{cases}$$

Recall that v^* is unique upto a constant. For our experiments, we use $\beta = 2/3, \alpha = 3, C = 15$. This gives the optimality policy as $\pi^*(x) = 0$ if $x \le 2.654$,otherwise 1 and $J^* = -7.962$. Note that one could use any reference state instead of minimum of the function. Fig. 1 shows the estimation of transition probability using Gaussian and tophat kernel for $M = 50$ and $h_M = 0.2$ for a fixed state-action pair $(2, 0)$. Fig. 2 presents the optimal and

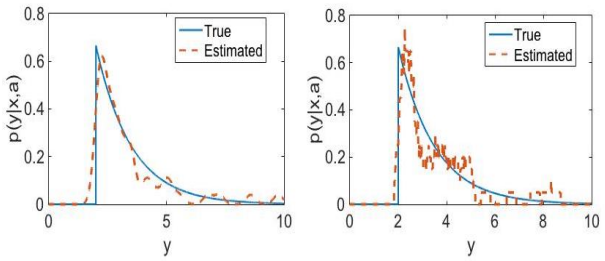

Figure 1: Kernel density estimation using Gaussian kernel (left) and tophat kernel (right)

estimated value function. We used Gaussian kernel here with a bandwidth of 0.2. The number of states sampled and number of neighbors are fixed to $N = 100$, $n = 20$ and number of next states are varied $M = 10, 50$. Larger values of M give a smoother estimation of density and hence a better estimation of value function.

6 CONCLUSIONS

In this paper, we proposed an approximate relative value learning (ARVL) algorithm for MDPs with continuous state space. It is a variant of relative value iteration for average-reward MDP, a combination of kernel density estimation and non-parametric function approximation. Although, we focused on nearest neighbors regression,

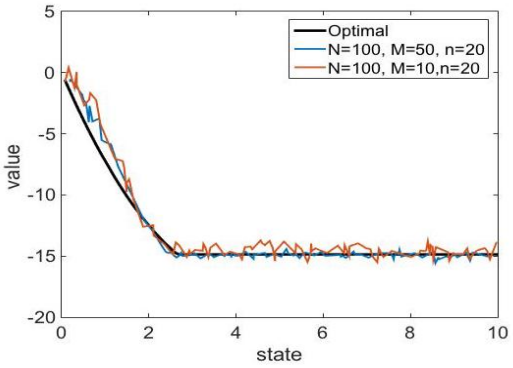

Figure 2: Optimal and estimated value function

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. The proof argument is based on probabilistic contraction and their convergence to a probabilistic fixed point. We use stochastic dominance argument to argue convergence of our algorithm with high probability.

APPENDIX

Proof of Lemma 2.1. Consider a Cauchy sequence $\{\widetilde{f}_n\}_{n\ge1}$,where $\widetilde{f}_n \in \widetilde{\mathcal{C}}(\mathcal{X})$ i.e, there exists N_ϵ such that for $m, n \ge N_\epsilon$, $span(\widetilde{f}_m - \widetilde{f}_n) < \epsilon$. Fix $x_0 \in \mathcal{X}$. Choose a function $f_n \in \widetilde{f}_n$ such that $f_n(x_0) = 0$ for all n. Hence, for $m, n \ge N_\epsilon$, $f_n(x_0) - f_m(x_0) = 0$, hence $\sup(f_n - f_m) \ge 0 \ge \inf(f_n - f_m)$. For every $x \in \mathcal{X}$,

$$|f_m(x) - f_n(x)| \le span(f_n - f_m) < \epsilon$$

Hence, $f_n(x)$ is a Cauchy sequence in $\mathbb{R}$ which implies that there exists a function f such that $\lim_{n\to\infty} f_n(x) = f(x)$. Furthermore, $\sup_x |f_m(x) - f_n(x)| < \epsilon$. Thus, we have a uniformly Cauchy sequence which implies uniform convergence. Hence $f \in \mathcal{C}_\mathcal{X}$. Let $\widetilde{f}$ be the corresponding element of f in $\widetilde{\mathcal{C}}(\mathcal{X})$ then

$$\lim_{n\to\infty} span(\widetilde{f}_n - \widetilde{f}) = \lim_{n\to\infty} span(f_n - f) \le 2 \lim_{n\to\infty} \sup(|f_n - f|) = 0$$

□

Acknowledgements

This research is supported by NSF Awards ECCS-1810447 and CCF-1817212. This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing.

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

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

[4] LP Devroye and TJ Wagner. The l1 convergence of kernel density estimates. *The Annals of Statistics*, pages 1136–1139, 1979.

[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] Abhishek Gupta, Rahul Jain, and Peter Glynn. Probabilistic Contraction Analysis of Iterated Random Operators. *arXiv e-prints*, page arXiv:1804.01195, Apr 2018.

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

[9] William B. Haskell, Rahul Jain, Hiteshi Sharma, and Pengqian Yu. An Empirical Dynamic Programming Algorithm for Continuous MDPs. *arXiv e-prints*, page arXiv:1709.07506, Sep 2017.

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

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

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

[13] Ronald Ortner. Pseudometrics for state aggregation in average reward markov decision processes. In *International Conference on Algorithmic Learning Theory*, pages 373–387. Springer, 2007.

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

[15] Naci Saldi, Serdar Yüksel, and Tamás Linder. On the asymptotic optimality of finite approximations to markov decision processes with borel spaces. *Mathematics of Operations Research*, 42(4):945–978, 2017.

[16] Hiteshi Sharma, Abhishek Gupta, and Rahul Jain. An empirical relative value learning algorithm for non-parametric mdps with continuous state space. In *2019 IEEE European Control Conference (ECC)*.