2019 57th Annual Allerton Conference on Communication, Control, and Computing
(Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, September 24-27, 2019

An Approximately Optimal Relative Value Learning Algorithm for Averaged MDPs with Continuous States and Actions

Hiteshi Sharma and Rahul Jain

Abstract— **It has long been a challenging problem to design algorithms for Markov decision processes (MDPs) with continuous states and actions that are provably approximately optimal and can provide arbitrarily good approximation for any MDP. In this paper, we propose an empirical value learning algorithm for average MDPs with continuous states and actions that combines empirical value iteration with n function-parametric approximation and approximation of transition probability distribution with kernel density estimation. We view each iteration as operation of random operator and argue convergence using the probabilistic contraction analysis method that the authors (along with others) have recently developed.**

I. Introduction

Sequential decision making under uncertainty problems are often formulated as Markov Decision Processes (MDPs) [14]. Various dynamic programming algorithms are available for discrete state spaces [2]. For continuous state space, techniques like state aggregation and function approximation have to be used [3], [11]. MDPs with average reward criterion require certain restrictions on the underlying induced Markov chains to establish the existence of stationary optimal policies [1]. In the discounted setting, the Bellman operator is contractive with respect to sup-norm. For the averaged case, contractive property only holds with respect to a semi-norm.

For MDPs with finite state and action spaces, an extensive study on the existence and structural properties of optimal policies as well as algorithms to compute such policies are available [14]. But computing optimal policies is generally a challenging problem when state and action spaces are uncountable. One idea is to quantize such 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 continuous MDPs with when the rewards are not bounded.

In this paper, we consider continuous MDPs when we do not know the transition kernel. Hence unlike previous works, discretization of state and action spaces is not possible. We only have access to samples of next states and thus propose a sampling based algorithm. 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 neighbor function approximation in this paper, it can be extended to other non-parametric function fitting methods as long as it is non-expansive and uniform convergence is obtained.

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

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], [17], [18]. In particular, [6], [17] uses truncation which requires the knowledge of contraction coefficient. This may not be easy to compute for continuous state and action 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], this was extended to unbounded iterates by constructing the Markov chain on the set of natural numbers and then analyzing the invariant distribution under some conditions. This argument for convergence of random contraction to probabilistic fixed point is used in [16] for MDPs with continuous state and discrete action spaces. In this paper, we extend the empirical framework to continuous action spaces by optimization via sampling. Similar to [16], we use nearest neighbors for function approximation. For nearest neighbor function approximation, [5] provides a uniform convergence under Lipschitz continuity assumption of the regression function. This requires us to have MDPs with Lipschitz continuous transition and reward function.

The main contribution of this paper is to introduce a randomized off-policy (relative) value learning algorithm for computing optimal policies for non-parametric continuous MDPs with average reward criterion when the transition kernel is unknown. We do not discretize the spaces or 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 II presents some preliminaries. The algorithm combining the approximate operator with function approximation is presented in Section III. The theoretical analysis is then presented in Section IV.

978-1-7281-3151-1/19/$31.00 ©2019 IEEE

II. Problem Formulation

Consider an MDP $(\mathcal{X},\mathcal{U},r,P)$ where $\mathcal{X}$ is the state space and $\mathcal{U}$ is the action space, $r : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is the reward function and P is the transition kernel. 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 $P(X_{t+1} \in B|X_t = x, u_t = u)$. For a stationary and deterministic policy $\pi : \mathcal{X} \to \mathcal{U}$, we are interested in maximizing the long-run average expected reward defined as

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

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 $J^{\pi^*}(x) = J^*$. 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.,

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

We now make the following assumptions.

Assumption 1: (a) $\mathcal{X}$ and $\mathcal{U}$ are compact subsets in $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_U}$ respectively. Furthermore, $\mathcal{U}$ is convex.

(b) For every (x,u), $|r(x,u)| \le r_{\max}$ and for every u, $r(\cdot,u)$ is Lipschitz continuous.

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

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

$$|p(z|x,u) - p(z|y,u)| \le \tilde{L}_p(z)\|x - y\|$$

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

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

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

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

(f) The reward and the transition kernel are Lipschitz continuous with respect to the action i.e., there exist 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

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

Assumption 1(b) establishes that for every a, $r(\cdot,u) \in \text{Lip}(\mathcal{X})$, (c) and (d) imply that if $v \in \text{Lip}(\mathcal{X})$ then for any action u, $\int v(y)P(dy|\cdot,u) \in \text{Lip}(\mathcal{X})$; (e) 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. The last assumption is Lipschitz continuity of reward and transition kernel in action variable. The compactness of action space combined with Lipschitz continuity implies that the greedy policies do exist. Under these assumptions, there exists $(J^*, v^*) \in \mathbb{R} \times \mathcal{C}(\mathcal{X})$ such that the following optimality equation holds:

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

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

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

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 1: [16] $(\widetilde{\mathcal{C}}(\mathcal{X}), span)$ is a Banach space.

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

Theorem 2: [10] Suppose that Assumptions 1 hold. Then, operator $T : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ is a span-contraction operator, i.e.,

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

where $v_1, v_2 \in \mathcal{C}(\mathcal{X})$ and α is defined in Assumption 1(c). Now consider a $v \in \mathcal{C}(\mathcal{X})$, and let $\tilde{v}$ be the corresponding element in $\widetilde{\mathcal{C}}(\mathcal{X})$ and $\widetilde{T} : \widetilde{\mathcal{C}}(\mathcal{X}) \to \widetilde{\mathcal{C}}(\mathcal{X})$ 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.

A. 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 $\mathcal{K} : \mathcal{X} \to \mathbb{R}$ defined as any smooth function such that $\int \mathcal{K}(x)dx = 1$, $\int x\mathcal{K}(x)dx = 0$ and $\int x^2\mathcal{K}(x)dx < \infty$. Assume that for any $(x,u) \in \mathcal{X}\times\mathcal{U}$, we have access to M independent and identically distributed samples $Y_i^{x,u} \sim P(\cdot|x,u)$, $i = 1, 2, \ldots M$. Let h_M be the bandwidth, then the kernel density estimator is defined as

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

For instance, the kernels commonly used are the Gaussian kernel, $\mathcal{K}(x) = \frac{1}{\sqrt{2\pi}}\exp(-\|x\|^2/2)$ and tophat kernel, $\mathcal{K}(x) = \frac{1}{2}\mathbb{I}(\|x\| < 1)$ 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 an approximation of Bellman operator $\widehat{T}_M : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ as follows:

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

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

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

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

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

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 IV).

Note that in the operators defined previously, we have optimization over the action space. In this paper, we consider optimization via sampling. This means for a given state $x \in \mathcal{X}$ and sample sizes M and L, we first sample L actions uniformly from $\mathcal{U}$ and then generate M samples of next state (for each sampled action). This leads us to define an approximate Bellman operator $\widehat{T}_{M,L} : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ as

$$\widehat{T}_{M,L}\, v(x) = \max_{u_1,u_2,\ldots u_L} \left[r(x,u_l) + \mathbb{E}_{x'\sim\widehat{P}_M(\cdot|x,u_l)} v(x') \right].$$

B. Nearest neighbor function approximation

Let us now define a function space $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$. 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, 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

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

Thus, it allows to reconstruct a function from some finite samples. Note that the function approximation operator $\Pi_{\mathcal{F}}(N,n)$ 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.

C. Random contraction

In this section, we introduce the definition of random contraction operator and fixed points in probabilistic sense as provided in [8], [16]. Consider a function space $\mathcal{F}$ with norm $\|\cdot\|$. Suppose there is a contraction operator $H : \mathcal{F} \to \mathcal{F}$. Let $\widehat{H}_N$ be the approximation of operator H via finite samples N.

Definition 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: 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.

III. Algorithm and the Main Result

We now present the Random Empirical Relative Value Learning (RERVaL) algorithm, a non-parametric off-policy algorithm for MDPs with continuous state and action space. It is a sampling-based algorithm combined with non-parametric density estimation and function approximation. It first samples the state and action 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. Since the optimal value function is unique up to a constant, we are choosing a non-negative optimal value function. This makes the samples for regression non-negative. Let the number of state samples be N, number of action samples be L, 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

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

where 1_N is a vector of all ones of size N. Let us denote the composed operator by $\widehat{G}(N,L,M,n) = \Pi_{\mathcal{F}}^n \Gamma_N \widehat{T}_{M,L}$ where we use the fact that the function approximation depends both on N and n. Algorithm 1 will iterate the random operator $\widehat{G}(N,L,M,n)$ (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 (probabilistic contraction).

Now, we specify the RERVaL algorithm in detail. We first sample N states from $\mathcal{X}$ uniformly followed by sampling L actions from $\mathcal{U}$. 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 RERVaL

Input: sample sizes $N, L, M, n \geq 1$; 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) Sample $\{u_{ij}\}_{j=1}^{L}$ from $\mathcal{U}$ uniformly for each $1 \leq i \leq N$
3) Kernel density estimation $\widehat{p}_M(\cdot|x_i, u_{ij})$ for each $1 \leq i \leq N$ and $1 \leq j \leq L$
4) Approximate value iteration: $\widehat{v}'_k(x_i) \leftarrow \widehat{T}_{M,L} v_{k-1}$, $\widehat{v}_k(x_i) \leftarrow \widehat{v}'_k(x_i) - \min_{x_j} \widehat{v}'_k$ for $i, j = 1, 2, \ldots, N$
5) Function approximation: $v_k \leftarrow \Pi_{\mathcal{F}} \widehat{v}_k$.
6) Increment $k \leftarrow k + 1$ and return to Step 1.

We can now establish that the iterates of the algorithm, v_k converge to a weak probabilistic fixed point of the operator $\widehat{G}(N, L, M, n) = \Pi^n_{\mathcal{F}} \Gamma_N \widehat{T}_{M,L}$ and hence a good approximation to v^*, the fixed point of T in the span semi-norm with high probability if N, L, M, n and k are large enough.

Theorem 3: Suppose that Assumptions 1 and 2 hold. Given $\epsilon, \delta > 0$, there exist constants B and C such that for any

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

$$n \geq n_0(\epsilon) = \frac{N}{2}\left(\frac{\epsilon}{4\,B\,C}\right)^{d} \text{ and}$$

$$L \geq L_0(\epsilon, \delta) = \left(\frac{(L_U = L_r + B\,L_p)\,\mathrm{diam}(\mathcal{U})}{\epsilon}\right)^{d_U} \log\frac{1}{\delta},$$

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. This can be made better by using kernel regression (e.g., Nadaraya-Watson kernel regression). Furthermore, the dependence on next state sample size M is due to asymptotic convergence of kernel density estimation.

IV. Analysis: Proof of Theorem 3

In this section, we prove Theorem 3. There are three approximations in RERVaL: first one due to sampling, second one due to density estimation and lastly due to function fitting. We first bound the error due to these approximations. As mentioned before, each iteration of RERVaL 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 optimization via sampling: 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 $v \in \mathcal{C}(\mathcal{X})$. We now argue that if the value function v is bounded by B then the Q-value function is L_U-Lipschitz continuous in action variable where $L_U = L_r + B\,L_p$. For all $(x, u, u') \in \mathcal{X} \times \mathcal{U} \times \mathcal{U}$

$$\begin{aligned}
&|Q(x,u) - Q(x,u')| \\
&\leq |r(x,u) - r(x,u')| + \int_{\mathcal{X}} |(P(dy|x,u) - P(dy|x,u'))\, v(y)| \\
&\leq L_r|u - u'| + B\int_{\mathcal{X}} |P(dy|x,u) - P(dy|x,u')| \\
&\leq (L_r + BL_p)\,\|u - u'\|
\end{aligned}$$

where the last inequalities follow from Assumption 1. Now since $v \in \mathcal{C}(\mathcal{X})$ and $\mathcal{X}$ is compact, there exists a constant B such that $|v(x)| \leq B$ for all $x \in \mathcal{X}$. Let us also define a (random) operator, $\widetilde{T}_L$ as follows

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

The next lemma bounds the error due to sampling for finding the best action. The proof is given in the appendix.

Lemma 4: 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 $v \in \mathcal{C}(\mathcal{X})$ if $L \geq L_0(\epsilon, \delta)$ then

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

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

Proof: 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 $\mathrm{vol}(\mathcal{B}(u, r)) \propto r^{d_U}$. Let $\mathcal{U}_\epsilon = \{u \in \mathcal{U} : f(u) \leq \max_u f(u) - \epsilon\}$. Moreover, let $\mathcal{U}_{\epsilon, L_U} = \{u \in \mathcal{U} : \|u^* - u\| \leq \epsilon / L_U\}$. Since f is L_U-Lipschitz, $u \notin \mathcal{U}_\epsilon \implies u \notin \mathcal{U}_{\epsilon, L_U}$. Hence,

$$\begin{aligned}
\mathbb{P}\left(u \notin \mathcal{U}_{\epsilon, L_U}\right) &= 1 - \frac{\mathrm{vol}(\mathcal{U}_{\epsilon, L_U})}{\mathrm{vol}(\mathcal{U})} \\
&\leq 1 - \left(\frac{\epsilon}{L_U\,\mathrm{diam}(\mathcal{U})}\right)^{d_U} \qquad (2)
\end{aligned}$$

where the last inequality follows from Lemma 5.2 in [19] and $\mathrm{diam}(\mathcal{U}) = \sup_{u,u'} \|u - u'\|$. Now,

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

where the second equality is due to the fact that $\{u_1, u_2 \ldots u_L\}$ are i.i.d. and the last inequality follows Lipschitz continuity of the function f. Now, using (2) we have

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

Putting $\left(\frac{\epsilon}{L_u \operatorname{diam}(\mathcal{U})}\right)^{d_U} = \frac{1}{L}\log\left(\frac{1}{\delta}\right)$ and using $1-x \leq e^{-x}$, we have $\mathbb{P}\left(f(u^*) - \max_{1\leq l\leq L} f(u_l)\right) \geq \delta$ for the choice of L. ∎

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 2:
1) Let $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$ such that $\int \mathcal{K}(x)dx = 1$ and $L(y) = \sup_{\|x\|\geq y} \mathcal{K}(x)$ for $y \geq 0$.
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 $p(\cdot|x,u)$ is almost everywhere continuous for all $(x,u) \in \mathcal{X}\times\mathcal{U}$ and $\|x\|^d\, \mathcal{K}(x) \to 0$ as $\|x\| \to \infty$.

Proposition 1: Let $\mathcal{K}$ be a smoothing kernel such that Assumption 2 holds, then the following holds with probability 1,

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

for all $(x,u) \in \mathcal{X}\times\mathcal{U}$.
This now leads to the following lemma:

Lemma 5: Assume that Assumption 2 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 Proposition 1. For any $(x,u),(x',u') \in \mathcal{X}\times\mathcal{U}$,

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

Using ergodicity of transition kernel as mentioned in assumption 1(e) and Proposition 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 $\Gamma : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ defined as $\Gamma f = f - \min f$. Let $g_M : \mathcal{X} \to \mathbb{R}$ be such that

$$g_{M,L}(x) = \left[\max_{u_1,u_2,\ldots u_L} \left\{r(x,\, a) + \mathbb{E}_{x'\sim\widehat{P}_M(\cdot|x,a)} v(x')\right\}\right]$$

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

$$f_{M,L}(x) = \mathbb{E}[g_{M,L}] \quad \text{and} \quad \widetilde{f}_{M,L}(x) = \mathbb{E}[\Gamma\, g_{M,L}]$$

$\widetilde{f}_{M,L}$ is the *regression function*. It is the expected value of our approximate estimator of Tv. As expected, $f_{M,L} \to Tv$ as $M, L \to \infty$. We note that $f_{M,L}$ 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 $f_{M,L}$ and the Bellman update $T\,v$ arbitrarily small uniformly over $x \in \mathcal{X}$ when M and L are large enough.

Lemma 6: Under Assumptions 1 and 2, if $L \geq L_0(\epsilon,\delta)$ then for any $\epsilon > 0$, following holds

$$\lim_{M\to\infty} \|f_{M,L} - T\,v\|_\infty \leq \epsilon$$

with probability at least $1-\delta$.

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

$$\begin{aligned}&|f_{M,L}(x) - Tv(x)| \\ &\leq |f_{M,L}(x) - \widetilde{T}_L v(x)| + |\widetilde{T}_L v(x) - Tv(x)|\end{aligned}$$

Let us now bound the first term:

$$\begin{aligned}&|f_{M,L}(x) - \widetilde{T}_L v(x)| \\ &\leq \mathbb{E}\Bigg[\Bigg|\max_{u_1,u_2,\ldots u_L}\left\{r(x,\, u_l) + \mathbb{E}_{x'\sim\widehat{P}_M(\cdot|x,u_l)} v(x')\right\} \\ &\qquad - \max_{u_1,u_2,\ldots u_L}\left\{r(x,\, u_l) + \mathbb{E}_{x'\sim P(\cdot\,|\,x,\,u_l)}[v(x')]\right\}\Bigg|\Bigg] \\ &\leq \mathbb{E}\left[\max_{u_1,u_2,\ldots u_L}\left|\mathbb{E}_{x'\sim\widehat{P}_M(\cdot|x,u_l)} v(x') - \mathbb{E}_{x'\sim P(\cdot\,|\,x,\,u_l)}[v(x')]\right|\right]\end{aligned}$$

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 u_x^* then by Jenson's and Cauchy-Schwartz inequalities we have

$$\begin{aligned}&\lim_{M\to\infty} |f_{M,L}(x) - \widetilde{T}_L v(x)| \\ &\qquad \leq \|v\|_\infty \lim_{M\to\infty} \mathbb{E}\Bigg[\int_{\mathcal{X}} \left|\widehat{p}_M(x'|x,u_x^*)\right. \\ &\qquad\qquad \left. - p(x'|x,u_x^*)\right|\lambda(dx')\Bigg]\end{aligned}$$

Using bounded convergence theorem and Theorem 1, this term vanishes. The second term can be bounded by Lemma 4. ∎

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

Lemma 7: Suppose for a value function $v \in \text{Lip}(\mathcal{X})$, 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 and L, then for $\delta, \epsilon > 0$, $N \geq N_0(\epsilon,\delta)$ and $n \geq n_0(\epsilon)$, we have

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

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

Lemma 8: Given $v \in \text{Lip}\,(\mathcal{X})$, $\epsilon > 0$, and $\delta \in (0, 1)$. Also choose $N \geq N_0(\epsilon, \delta)$, $n \geq n_0(\epsilon)$ and $L \geq L_0(\epsilon, \delta)$, Then, we have

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

Proof:

By the triangle inequality,

$$span(\widehat{G}\, v - T\, v) \leq span(\widehat{G}\, v - \widetilde{f}_{M,L}) + span(f_{M,L} - T\, v)$$

where the last inequality follows from the fact that $span(\widetilde{f}_M - f_M) = span(\mathbb{E}[\Gamma g_{M,L} - g_{M,L}]) = 0$. Combining with Lemma 6 and 7 concludes the proof. ∎

Next we establish that it is indeed a random contraction.

Lemma 9: For a given $N, M, n \geq 1$, the operator $\widehat{G}(N, L, M, n) = \Pi^n_{\mathcal{F}}\, \Gamma_N\, \widehat{T}_{M,L}$ is a random contraction operator, i.e, for any $v_1, v_2 \in \mathcal{C}(\mathcal{X})$,

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

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

The proof is similar to Lemma 4.6 in [16] and hence omitted.

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

Theorem 10: Assume that the following holds:

1) $T : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ is a contraction operator in span norm with contraction coefficient $\alpha < 1$.
2) For any $v \in \text{Lip}\,(\mathcal{X})$, we have
$$\lim_{M,N,L,n\to\infty} \mathbb{P}(span(\widehat{G}\, v - Tv) \geq \epsilon) = 0.$$
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 $\epsilon > 0, \delta \in (0, 1-\alpha), n, N, M \geq 1$,

$$q(\epsilon, \delta, N, M, n) \triangleq \mathbb{P}\Big(\widehat{\alpha}_M \leq 1 - \delta, \quad span\left(\widehat{G}\, v - T\, v\right) \leq \epsilon\Big), \tag{3}$$

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 3: 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 $t > 0$,

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

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: [Proof of Theorem 3] Now we apply Theorem 10. Note that the first assumption in the theorem is satisfied by the ergodicity condition assumed in Assumption 1. The second and the third assumptions are satisfied by Lemma 8 and Lemma 5 respectively. The fourth one follows from bounded rewards and the fact that v^* is a fixed point of operator T. Hence, Theorem 10 can be applied to conclude the convergence. ∎

V. Conclusion

In this paper, we presented an approximately optimal, offline batch algorithm for continuous MDPs with average reward criterion. Instead of discretizing the continuous state and action spaces, we propose a sampling based algorithm. This allows us to deal with scenarios when the MDP is unknown but there is generative model available. The samples from generative model are used to approximate the dynamics via density estimation. To generalize over state space, we use nearest neighbor function approximation. The proposed algorithm is viewed as iteration of random operator, a composition of approximate Bellman operator and nearest neighbor approximation. The analysis is through stochastic dominance argument which in turn gives us a convergence in probability.

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, Mehdi Jafarnia-Jahromi, and Rahul Jain. Approximate relative value learning for average-reward continuous state mdps.

[17] Hiteshi Sharma, Rahul Jain, and Abhishek Gupta. An empirical relative value learning algorithm for non-parametric mdps with continuous state space. In *2019 18th European Control Conference (ECC)*, pages 1368–1373. IEEE, 2019.

[18] Hiteshi Sharma, Rahul Jain, and William Haskell. Empirical algorithms for general stochastic systems with continuous states and actions. In *2019 58th Control and Decision Conference (CDC)*. IEEE, 2019.

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