Empirical Value Iteration for Approximate Dynamic Programming

William B. Haskell, Rahul Jain and Dileep Kalathil

Abstract—We propose a simulation based algorithm, Empirical Value Iteration (EVI) algorithm, for finding the optimal value function of an MDP with infinite horizon discounted cost criteria when the transition probability kernels are unknown. Unlike simulation based algorithms using stochastic approximation techniques which give only asymptotic convergence results, we give provable, non-asymptotic performance guarantees in terms of sample complexity results: given $\epsilon > 0$ and $\delta > 0$, we specify the minimum number of simulation samples $n(\epsilon, \delta)$ needed in each iteration and the minimum number of iterations $t(\epsilon, \delta)$ that are sufficient for the EVI to yield, with a probability at least $1 - \delta$, an approximate value function that is at least $\epsilon$ close to the optimal value function.

I. INTRODUCTION

A large class of sequential decision making problems under uncertainty, like shortest path selection for transportation networks, resource allocation problems in inventory control and asset acquisition and pricing in financial markets, are generally modeled as Markov Decision Processes (MDP) [1]. The objective is to find an optimal decision making policy which minimizes the cost incurred to the decision maker. Dynamic Programming (DP) methods, Value Iteration (VI) and Policy Iteration (PI), are the standard approaches for solving such MDP problems.

Practical applications of DP are known to suffer from two major problems. First one is the well known “curse of dimensionality”. There exists a large body literature addressing this problem where the key idea is to approximate the value function using state aggregation, basis function approximation and neural networks [2]. Second, more often than not, the actual transition kernels of MDPs are not known completely. The standard way of overcoming this difficulty is by doing Monte Carlo simulations. However like all Monte Carlo methods, simulation based dynamic programming is also mostly based on heuristics or it gives only asymptotic results. Heuristic methods do not give any theoretical guarantees about the performance or may not even converge in some cases. Asymptotic methods on the other hand while giving theoretical guarantees on performance, converge at a slow pace. Real world problems demand algorithms which gives non-asymptotic solutions with provable performance guarantees. Computer simulation are done only for finite time and theoretical guarantees are thus needed to understand their approximate optimality.

In this paper, we propose a simulation based dynamic programming algorithm, Empirical Value Iteration (EVI), for finding the optimal value function of an MDP with infinite horizon discounted cost criteria when the transition kernels are unknown. We assume that the state transitions can be simulated by a simulation function which maps the current state, action and the realized value of an exogenous random variable (noise) to a next random state. In each iteration of the EVI, we generate $n$ i.i.d. noise samples to estimate the transition kernel and perform the standard value iteration step using this estimated kernel. Rather than giving a heuristic argument about the convergence, we give non-asymptotic provable performance guarantees. Specifically, we give a sample complexity result: given $\epsilon > 0$ and $\delta > 0$ we specify the minimum number of noise samples $n(\epsilon, \delta)$ needed in each iteration and the minimum number of iterations $t(\epsilon, \delta)$ that are sufficient for the EVI to yield, with a probability at least $1 - \delta$, an approximate value function that is at least $\epsilon$ close to the optimal value function.

Related Work: Many simulation-based algorithms for MDPs fall under the category of stochastic approximation methods [3]. In general, they only give asymptotic result and “do not have reliable implementable stopping rules” [4]. So this demands a large simulation time and results in a slow rate of convergence. Two particular cases of such stochastic approximation based algorithms are Q-learning algorithms and Actor-Critic algorithms which are simulation based versions of value iteration and policy iteration respectively. Though some analysis about the rate of convergence of these algorithms is available [5] [6], the rate of convergence is known to be slow like any other stochastic approximations algorithms. Also, they differ from the problem that we consider here in another important aspect. These algorithms are online learning algorithms where an optimal policy is computed ‘on the fly’. A necessary condition for such an algorithm to converge asymptotically is that each state-action pair is ‘sampled’ by the algorithm infinitely often. This condition requires some irreducibility assumption on the underlying Markov chain.

In [7] [8], authors give a simulation based method for estimating the value of a policy and give a sample complexity result, the minimum number of samples necessary to estimate the value of a policy approximately correctly with high probability uniformly over a class of policies, when the policy spaces have finite combinatorial dimension. However, they did not present any computationally tractable iterative procedure to search through the policy space for finding the optimal policy. A simulation based policy iteration for average cost MDPs is proposed in [9] but it only gives an
asymptotic result. The closest extant paper to our present work is [10] where simulated methods are used to estimate the optimal policies of finite horizon MDPs through (backward) value iteration. Though they give a sample complexity result, analysis is considerably simplified because the number of iterations are finite which makes easier to control the maximum error. Also they make a strong assumption that the simulation function is Lipschitz. The EVI algorithm does not require such an assumption on the simulation function.

Our Contributions: In this work, we propose a natural simulation based algorithm, EVI algorithm, for infinite horizon discounted MDPs and give provable non-asymptotic performance guarantees in terms of sample complexity results. Unlike stochastic approximation methods, EVI requires no irreducibility assumptions on the underlying MDP and it will work in all cases where classical value iteration works. MDPs which do not satisfy the necessary assumptions for stochastic approximation methods are typically solved by Monte Carlo algorithms based on heuristic. Our contribution is to develop a mathematical framework for analyzing simulation based algorithms for solving such general class of MDPs and giving theoretical guarantees for their performance. We also develop a new method for the convergence analysis of our simulation-based algorithm using a stochastic dominance argument.

II. Preliminaries

A typical representation of a discrete time MDP is the 5-tuple \((\mathcal{S}, \mathcal{A}, \{A(s) : s \in \mathcal{S}\}, Q, c)\). The state space \(\mathcal{S}\) and the action space \(\mathcal{A}\) are both finite. We define \(\mathcal{P}(\mathcal{S})\) to be the space of probability measures over \(\mathcal{S}\), and we define \(\mathcal{P}(\mathcal{A})\) similarly. For each state \(s \in \mathcal{S}\), the set \(A(s) \subseteq \mathcal{A}\) is the available set of feasible actions. The entire set of feasible state-action pairs is \(\mathcal{K} \triangleq \{(s, a) \in \mathcal{S} \times \mathcal{A} : a \in A(s)\}\). The transition law \(Q\) governs the system evolution, \(Q(\cdot|s, a) \in \mathcal{P}(\mathcal{A})\) for all \((s, a) \in \mathcal{K}\). Explicitly, \(Q(j|s, a)\) is the probability of next visiting the state \(j\) given the current state-action pair \((s, a)\). Finally, \(c : \mathcal{K} \to \mathbb{R}\) is a cost function that depends on state-action pairs.

Define \(\Pi\) to be the class of stationary deterministic Markov policies: mappings \(\pi : \mathcal{S} \to \mathcal{A}\) which only depend on history through the current state. For a given state \(s \in \mathcal{S}\), \(\pi(s) \in A(s)\) is the action chosen in state \(s\) under the policy \(\pi\). We explicitly assume that \(\Pi\) only includes feasible policies that respect the constraints \(\mathcal{K}\). Any policy \(\pi \in \Pi\) and initial state \(s \in \mathcal{S}\) determine a stochastic process \(\{(s_t, a_t), t \geq 0\}\) where \(s_t\) and \(a_t\) are the state and action at time \(t\), respectively. The expectation operator with respect to the policy \(\pi\) and initial state \(s\) is denoted \(\mathbb{E}_s[\cdot]\).

We focus on infinite horizon discounted cost MDPs with discount factor \(\alpha \in (0, 1)\). For a given initial state \(s \in \mathcal{S}\), the optimal cost starting from state \(s\) is

\[v^*(s) \triangleq \inf_{\pi \in \Pi} \mathbb{E}^s_{\pi} \left[ \sum_{t \geq 0} \alpha^t c(s_t, a_t) \right],\]

and \(v^* \in \mathbb{R}^{[\mathcal{S}]}\) denotes the corresponding optimal value function.

The Bellman operator \(T : \mathbb{R}^{[\mathcal{S}]} \to \mathbb{R}^{[\mathcal{S}]}\) will appear throughout this paper. Define \(T : \mathbb{R}^{[\mathcal{S}]} \to \mathbb{R}^{[\mathcal{S}]}\) via

\[ [Tv](s) \triangleq \min_{a \in A(s)} \left\{ c(s, a) + \alpha \sum_{j \in \mathcal{S}} Q(j|s, a) v(j) \right\}, \]

for all \(s \in \mathcal{S}\) and for any \(v \in \mathbb{R}^{[\mathcal{S}]}\). Notice the exact expectation of the cost-to-go, \(\sum_{j \in \mathcal{S}} Q(j|s, a) v(j)\), appears. We use the following helpful fact in this paper.

**Fact 1:** Let \(X\) be a set, and \(f_1 : X \to \mathbb{R}\) and \(f_2 : X \to \mathbb{R}\) be two real-valued functions on \(X\). Then:

\[
\begin{align*}
|\inf_{x \in X} f_1(x) - \inf_{x \in X} f_2(x)| & \leq \sup_{x \in X} |f_1(x) - f_2(x)|, \\
|\sup_{x \in X} f_1(x) - \sup_{x \in X} f_2(x)| & \leq \sup_{x \in X} |f_1(x) - f_2(x)|.
\end{align*}
\]

**Notation 1:** We note that the norm notation \(\| \cdot \|\) that we use in the paper always refers to sup norm.

III. The EVI Algorithm

This section presents the details of our EVI algorithm. The Bellman operator \(T\) requires the exact evaluation of the expectation

\[\mathbb{E}[v(s), a] = \sum_{j \in \mathcal{S}} Q(j|s, a) v(j).\]

To replace the exact expectation, we need a simulation model for the transition kernel \(Q\). Without loss of generality, we can assume that a sequence of independent uniform random variables drives this MDP. Let \(\psi : \mathcal{S} \times \mathcal{A} \times [0, 1] \to \mathcal{S}\) be an explicit simulation model for the state evolution such that \(\mathcal{P}(\psi(s, a, \xi) = s') = Q(s'|s, a)\) where \(\mathcal{P}\) is taken with respect to the uniform random variable \(\xi\) on \([0, 1]\). With this convention, the Bellman operator can be written as

\[ [Tv](s) \triangleq \min_{a \in A(s)} \left\{ c(s, a) + \alpha \mathbb{E}[v(\psi(s, a, \xi))] \right\}, \]

for all \(s \in \mathcal{S}\).

We will replace the expectation \(\mathbb{E}[v(\psi(s, a, \xi))]\) with an empirical estimate. Given a sample of \(n\) uniform random variables, \(\{\xi_i\}_{i=1}^n\), the empirical estimate of \(\mathbb{E}[v(\psi(s, a, \xi))]\) is \(1/n \sum_{i=1}^n v(\psi(s, a, \xi_i))\). Our algorithm is summarized next.

In each iteration, we regenerate samples and use this empirical estimate to approximate \(T\). Now we give our main result:

**Theorem 1:** Given \(\epsilon \in (0, 1)\) and \(\delta \in (0, 1)\), fix \(\epsilon_q = \epsilon/\eta^*\) and select strictly positive \(\delta_1, \delta_2\) such that \(\delta_1 + 2\delta_2 \leq \delta\) where \(\eta^* = [2/(1 - \alpha)]\). Select \(n\) such that

\[ n \geq n(\epsilon, \delta) = \frac{2(\kappa^*)^2}{\epsilon^2} \log \frac{2|\mathcal{K}|}{\delta_1}, \]

and select \(t\) such that

\[ t \geq t(\epsilon, \delta) \geq \log \left( \frac{1}{\delta_2 \mu_{n, \min}} \right). \]
Algorithm 1: Empirical Value Iteration (EVI)

Input: \( \hat{v}^0 \in \mathbb{R}^{|S|} \), sample size \( n \geq 1 \), maximum number of iteration \( N_{\text{max}} \). Set counter \( t = 0 \).

1) Sample \( n \) uniformly distributed random variables \( \{\xi_i\}_{i=1}^n \), and compute

\[
\hat{v}_{t+1}^n(s) = \min_{a \in A(s)} \left\{ c(s, a) + \frac{\alpha}{n} \sum_{i=1}^n \hat{v}_n^i(\psi(s, a, \xi_i)) \right\},
\]

for all \( s \in S \).

2) Increment \( t = t + 1 \) and return to step 1 if \( t < N_{\text{max}} \).

where \( \mu_{n, \min} = \min_{\eta} \mu_n(\eta) \) and the values of \( \mu_n(\eta) \) are given by Lemma 4.

We prove the following in section V.

This result says that, if we take \( n \geq n(\epsilon, \delta) \) samples in each iteration of the EVI algorithm and perform \( t \geq t(\epsilon, \delta) \) iterations then the EVI iterate \( \hat{v}_n^t \) is \( \epsilon \) close to the optimal value function \( v^* \) with a probability greater that \( 1 - \delta \).

IV. EMPIRICAL BELLMAN OPERATOR

We now frame our EVI algorithm in terms of iteration of an empirical Bellman operator which is a ‘random operator’. We first define the sample space \( \Omega = [0, 1]^\infty \), the \( \sigma \)-algebra \( \mathcal{F} = \mathcal{B}^\infty \) where \( \mathcal{B} \) is the inherited Borel \( \sigma \)-algebra on \([0, 1] \), and the probability distribution \( P \) on \( \Omega \) formed by an infinite sequence of uniform random variables. The primitive uncertainties on \( \Omega \) are sequences of uniform noise \( \omega = (\xi_i)_{i \geq 0} \) where each \( \xi_i \) is an independent uniform random variable on \([0, 1] \). This convention, rather than just defining \( \Omega = [0, 1]^n \) for a fixed \( n \geq 1 \), makes convergence statements with respect to \( n \) easier to make.

We formally define the empirical Bellman operator \( \hat{T}_n : \Omega \times \mathbb{R}^{|S|} \to \mathbb{R}^{|S|} \) as

\[
\hat{T}_n(\omega)(v)(s) \triangleq \min_{a \in A(s)} \left\{ c(s, a) + \frac{\alpha}{n} \sum_{i=1}^n v(\psi(s, a, \xi_i)) \right\}
\]

for all \( s \in S \). Note that for any \( \omega = (\xi_i)_{i \geq 0} \in \Omega \), \( \hat{T}_n(\omega) : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|} \) is a deterministic mapping. We prove the following important property.

Lemma 1: Choose any \( \omega = (\xi_i)_{i \geq 0} \in \Omega \), and the mapping \( \hat{T}_n(\omega) : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|} \) defined as in equation (1) is a contraction, i.e.,

\[
\| \hat{T}_n(\omega) v_1 - \hat{T}_n(\omega) v_2 \| \leq \alpha \| v_1 - v_2 \|
\]

for all \( v_1, v_2 \in \mathbb{R}^{|S|} \) and for all \( \omega \in \Omega \). Proof is omitted due to page limitation. Readers are referred to [11] for detailed proof.

We now give another important result which we will use later in our analysis.

Proposition 1: For any \( n \geq 1 \) and \( \epsilon > 0 \)

\[
P\left\{ \| \hat{T}_n v - T v \| \geq \epsilon \right\} \leq 2 |\mathbb{R}| e^{-2(\epsilon/\alpha)^2 n/(2 \kappa^2)}
\]

for all \( v \in \mathbb{R}^{|S|} \). Proof is omitted due to page limitation. Readers are referred to [11] for detailed proof.

Classical value iteration is performed by iterating the Bellman operator \( T \). Our EVI algorithm is performed by choosing \( n \) and then iterating the random operator \( \hat{T}_n \). Each iteration of \( \hat{T}_n \) requires a new sample \( \omega \in \Omega \). We introduce an additional probability space to formalize iteration of \( \hat{T}_n \). Let \( \mathcal{P} = \mathcal{P}^\infty \) be the probability measure on \( (\Omega^\infty, \mathcal{F}^\infty) \) guaranteed by the Kolmogorov extension theorem. We view \( (\Omega^\infty, \mathcal{F}^\infty, \mathcal{P}) \) as the appropriate probability space on which to define iteration of \( \hat{T}_n \). An element in \( \Omega^\infty \) is denoted \( \omega = (\omega_i)_{i=0}^\infty \).

Let \( \hat{v}_0 \in \mathbb{R}^{|S|} \) be an initial seed for EVI. We use the ‘hat’ notation to emphasize that the iterates of the EVI algorithm are random. Next we define

\[
\hat{T}_n(\omega) v = [\hat{T}_n(\omega_{t-1})][\hat{T}_n(\omega_{t-2})] \cdots [\hat{T}_n(\omega_0)] v
\]

We will view each random operator \( \hat{T}_n \) as being defined on \( (\Omega^\infty, \mathcal{F}^\infty, \mathcal{P}) \), for all \( n \geq 1 \) and \( t \geq 0 \).

The stochastic process \( \{\hat{v}_n^t\}_{t \geq 0} \) consists of iterates produced by \( \hat{T}_n \), for fixed \( n \geq 1 \). For a fixed \( \hat{v}_n^t \), we can view all \( \hat{v}_n^t \) as measurable mappings from \( \Omega^\infty \) to \( \mathbb{R}^{|S|} \) via the mapping \( \hat{v}_n^t(\omega) = \hat{T}_n(\omega) \hat{v}_n^0 \). Thus, we consider the stochastic process \( \{\hat{v}_n^t\}_{t \geq 0} \) to be defined on the probability space \( (\Omega^\infty, \mathcal{F}^\infty, \mathcal{P}) \).

Proposition 2: The stochastic process \( \{\hat{v}_n^t\}_{t \geq 0} \) is a Markov chain on \( \mathbb{R}^{|S|} \).

Proof: Follows from the fact that each iteration of \( \hat{T}_n \) is independent, and identically distributed. Thus, the next iterate \( \hat{v}_n^{t+1} \) only depends on history through the current iterate \( \hat{v}_n^t \).

We now verify that we can restrict attention to a bounded region when iterating \( \hat{T}_n \). Define the constant

\[
\kappa^* \triangleq \max_{c(s, a) \in \mathbb{R}} \left| \frac{c(s, a)}{1 - \alpha} \right| < \infty
\]

We give the following easy lemma without proof.

Lemma 2: For all \( \pi \in \Pi \), \( \| v^\pi \| \leq \kappa^* \).

Based on Lemma 2, we can claim that \( \{\hat{v}_n^t\}_{t \geq 0} \) effectively has the compact state space \( B_{\kappa^*} \{0\} = \{ v \in \mathbb{R}^{|S|} : \| v \| \leq \kappa^* \} \).

Even though \( \{\hat{v}_n^t\}_{t \geq 0} \) is a Markov chain, its analysis is complicated by two factors. First, \( \{\hat{v}_n^t\}_{t \geq 0} \) is a Markov chain on the continuous state space \( \mathbb{R}^{|S|} \), which introduces technical complications in general when compared to a discrete state space. Second, the transition probabilities of \( \{\hat{v}_n^t\}_{t \geq 0} \) are too complicated to compute explicitly.

Since we are tracking the progress of \( \{\hat{v}_n^t\}_{t \geq 0} \) to the fixed point \( v^* \) of \( T \), we are equivalently interested in the stochastic process \( \{\| \hat{v}_n^t - v^* \| \}_{t \geq 0} \) which just measures the norm of the error \( \hat{v}_n^t - v^* \). Formally, \( \{\| \hat{v}_n^t - v^* \| \}_{t \geq 0} \) is also defined on \( (\Omega^\infty, \mathcal{F}^\infty, \mathcal{P}) \). We already know we can restrict the stochastic process \( \{\| \hat{v}_n^t - v^* \| \}_{t \geq 0} \) to the compact state...
space $[0, 2\kappa^*]$, because $\|\hat{v}^t_n - v^*\| \leq \|\hat{v}^t_n\| + \|v^*\| \leq 2\kappa^*$ by Lemma 2. If $\|\hat{v}^t_n - v^*\|$ approaches zero in some sense, then we would know that $\hat{v}^t_n$ approaches $v^*$.

We choose a granularity $\epsilon_g > 0$ to discretize our state space. The granularity $\epsilon_g > 0$ is fixed for the remainder of this discussion.

We now present the stochastic process $\{X^t_n\}_{t \geq 0}$: define $X^t_n : \Omega^\infty \rightarrow \mathbb{R}$ so that $X^t_n(\omega) = 0$ if $\|\hat{v}^t_n(\omega) - v^*\| = 0$ and $X^t_n(\omega) = n$ if $(\eta - 1) \epsilon_g < \|\hat{v}^t_n(\omega) - v^*\| \leq \eta \epsilon_g$ for all $t \geq 0$. More compactly,

$$X^t_n(\omega) = \left\lceil \|\hat{v}^t_n(\omega) - v^*\|/\epsilon_g \right\rceil,$$

where $\left\lceil \cdot \right\rceil$ denotes the smallest integer greater than or equal to $\chi \in \mathbb{R}$. In this fashion we view $\{X^t_n\}_{t \geq 0}$ as a stochastic process on $(\Omega^\infty, F^\infty, \mathbb{P})$ along with $\{\hat{v}^t_n\}_{t \geq 0}$.

Our main result follows by showing that $X^t_n$ tend to zero, probabilistically, as $n$ and $t$ grow.

Recall that $X \geq_{as} Y$ denotes almost sure inequality between two random variables $X$ and $Y$ defined on the same probability space. The next lemma follows from the definition of $X^t_n$.

**Lemma 3:** For all $n \geq 1$ and $t \geq 0$, $X^t_n \geq_{as} \|\hat{v}^t_n - v^*\|/\epsilon_g$.

We also define the constant $N^* = \left\lceil \frac{2\kappa^*}{\epsilon_g} \right\rceil$. Notice that $N^*$ is the number of intervals of length $\epsilon_g$ needed to cover the interval $[0, 2\kappa^*]$ and by Lemma 2, $X^t_n \leq N^*$.

V. A STOCHASTIC DOMINANCE ARGUMENT

In this section we analyze EVI using the concept of stochastic dominance. First we will construct an analytically tractable family of Markov chains on $\mathbb{N}$ that stochastically dominate $\{X^t_n\}_{t \geq 0}$ (in a sense to be made precise). We will denote our “dominating” family of Markov chains as $\{Y^t_n\}_{t \geq 0}$, for $n \geq 1$. We can view $\{Y^t_n\}_{t \geq 0}$ as being defined on $(\mathbb{N}^{\infty}, \mathcal{F}^\infty)$, the canonical measurable space of trajectories on $\mathbb{N}$, so $Y^t_n : \mathbb{N}^\infty \rightarrow \mathbb{N}$. We will use ‘Q’ to denote a probability measure on $(\mathbb{N}^{\infty}, \mathcal{F}^\infty)$.

Because we have discretized the continuous state space of $\{\|\hat{v}^t_n - v^*\|\}_{t \geq 0}$ with granularity $\epsilon_g$, and because we are approximating $T$ with $\bar{T}_n$, there is a threshold below which we cannot detect improvement in the proximity of $\hat{v}^t_n$ to $v^*$. If $X^t_n = \eta$ and $T$ is approximated by $\hat{v}^t_n$ exactly to get $\bar{T}_n\hat{v}^t_n$, then

$$\|\bar{T}_n\hat{v}^t_n - v^*\| \leq \alpha \|\hat{v}^t_n - v^*\| \leq \alpha \eta \epsilon_g.$$

Recall that we have earlier chosen and fixed the granularity $\epsilon_g > 0$. Suppose $X^t_n = \eta$ so that $(\eta - 1) \epsilon_g < \|\hat{v}^t_n - v^*\| \leq \eta \epsilon_g$. Consider the event

$$F = \left\{ \|\bar{T}_n\hat{v}^t_n - \hat{v}^{t+1}_n\| = \|\bar{T}_n\hat{v}^t_n - \bar{T}_n\hat{v}^t_n\| < \epsilon_g \right\},$$

where $\bar{T}_n\hat{v}^t_n$ is accurate to within $\epsilon_g$ of $\bar{T}_n\hat{v}^t_n$. On the event $F$, we can guarantee a worst-case improvement in $\|\hat{v}^{t+1}_n - v^*\|$ of

$$\|\hat{v}^{t+1}_n - v^*\| = \|\bar{T}_n\hat{v}^t_n + \bar{T}_n\hat{v}^t_n - T\hat{v}^t_n - v^*\| \leq \|\bar{T}_n\hat{v}^t_n - v^*\| + \|\bar{T}_n\hat{v}^t_n - T\hat{v}^t_n\| \leq (\alpha + 1) \epsilon_g.$$

The preceding error estimate follows by using the known convergence rate of $T$ (where we get the term $\alpha \eta \epsilon_g$), and the fact that $\bar{T}_n\hat{v}^t_n$ is assumed to be accurate to within $\epsilon_g$ of $T\hat{v}^t_n$ (where we get the additional term $\epsilon_g$).

In order for this worst case improvement in $\|\hat{v}^{t+1}_n - v^*\|$ to register, i.e. for $X^{t+1}_n$ to be strictly smaller than $X^t_n$, the term $\alpha \eta + 1$ must be strictly less than $\eta$. If $\eta$ is too small, then $\alpha \eta + 1$ may equal to $\eta$ and no improvement in the proximity of $\hat{v}^t_n$ to $v^*$ can be detected by $\{X^t_n\}_{t \geq 0}$.

We define a new constant

$$\eta^* = \min \{ \eta \in \mathbb{N} \mid \eta \geq (\alpha + 1) + 1 \} = \left\lceil \frac{2}{1 - \alpha} \right\rceil,$$

and notice that for any $\eta \geq \eta^*$, $\alpha \eta + 1 < \eta$. When the current state is $X^t_n = \eta \geq \eta^*$, then

$$\|\hat{v}^{t+1}_n - v^*\| \leq \alpha \eta + 1 \epsilon_g < \eta \epsilon_g$$

and $X^{t+1}_n \leq (\alpha + 1) < X^t_n$ on event $F$.

We now formally define $\{Y^t_n\}_{t \geq 0}$. We can restrict $\{Y^t_n\}_{t \geq 0}$ to the finite state space $\{\eta^*, \eta^* + 1, \ldots, N^*, N^* + 1\}$ based on our preceding discussion. We define,

$$Y^{t+1}_n = \begin{cases} \max \{Y^t_n - 1, \eta^* \}, & \text{w.p. } p_n, \\ \eta^*, & \text{w.p. } 1 - p_n, \end{cases}$$

where,

$$p_n = 1 - 2 |\mathbb{E}| \epsilon_g \cdot \epsilon_g < 2(\alpha - 1) \epsilon_g.$$

From Proposition 1, we note that $P(F) = P\{\|\bar{T}_n\hat{v}^t_n - \bar{T}_n\bar{v}^t_n\| < \epsilon_g \} \geq p_n$. Also note that $p_n$ is the worst case probability for an ‘improvement’ (at least a one step transition from the left, closer to zero) in $\{X^t_n\}_{t \geq 0}$.

The first value $Y^{t+1}_n = \max \{Y^t_n - 1, \eta^*\}$ corresponds to the case where the approximate error satisfies $\|\bar{T}_n\hat{v}^t_n - \bar{T}_n\bar{v}^t_n\| < \epsilon_g$, and the second value $Y^{t+1}_n = \eta^*$ corresponds to all other cases (giving us an extremely conservative bound in the sequel). Informally, $\{Y^t_n\}_{t \geq 0}$ will either move closer to zero, or it will move as far away as possible. The transition matrix of $\{Y^t_n\}_{t \geq 0}$ will be denoted $\Omega$.

We now summarize some key properties of $\{Y^t_n\}_{t \geq 0}$.

**Proposition 3:** For any fixed $n \geq 1$ we have:

(i) $\{Y^t_n\}_{t \geq 0}$ is a Markov chain in the process on

(ii) The steady state distribution of $\{Y^t_n\}_{t \geq 0}$, denoted $Y^\infty = \lim_{t \rightarrow \infty} Y^t$, exists;

(iii) $Q\{Y^t_n > \eta\} \rightarrow Q\{Y^\infty > \eta\}$ as $t \rightarrow \infty, \forall \eta \in \mathbb{N};$

Proof is omitted due to page limitation. Readers are referred to [11] for detailed proof.

Now we apply the notion of stochastic dominance, the central feature of our development.

**Definition 1:** Let $X$ and $Y$ be two real-valued random variables, then $Y$ stochastically dominates $X$, written $X \leq_{st} Y$, when $\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$ for all increasing functions $f : \mathbb{R} \rightarrow \mathbb{R}$.

We now show that the Markov chain $Y^t_n$ stochastically dominates the stochastic process $X^t_n$ for any fixed $n \geq 1$.

**Theorem 2:** Assume that $X^0_n = Y^0_n$, then $X^t_n \leq_{st} Y^t_n$ for all $t \geq 0$.

Proof is given in Appendix A.
The following corollary resulting from Theorem 2 relates the stochastic processes \( \{\|\hat{v}_t^n - v^*\|\}_{t \geq 0}, \{X_t^n\}_{t \geq 0}, \) and \( \{Y_t^n\}_{t \geq 0} \) in a probabilistic sense, and summarizes our stochastic dominance argument.

**Corollary 1:** For any fixed \( n \geq 1 \), we have

(i) \( \mathbb{P}\left\{\|\hat{v}_t^n - v^*\| > \eta \varepsilon_g\right\} \leq \mathbb{P}\left\{X_t^n > \eta\right\} \leq \mathbb{Q}\left\{Y_t^n > \eta\right\} \) for all \( \eta \in \mathbb{N} \) for all \( t \geq 0 \);

(ii) \( \limsup_{t \to \infty} \mathbb{P}\left\{X_t^n > \eta\right\} \leq \mathbb{Q}\{Y_t^n > \eta\} \) for all \( \eta \in \mathbb{N} \);

(iii) \( \limsup_{t \to \infty} \mathbb{P}\left\{\|\hat{v}_t^n - v^*\| > \eta \varepsilon_g\right\} \leq \mathbb{Q}\{Y_t^n > \eta\} \) for all \( \eta \in \mathbb{N} \).

Proof is omitted due to page limitation. Readers are referred to [11] for detailed proof.

Let \( \mu_n \) denote the steady state distribution of \( Y_t^n \) (whose existence is guaranteed by Proposition 3) where \( \mu_n(i) = \mathbb{Q}\{Y_t^n = i\} \) for all \( \eta^* \leq i \leq N^* \). The next lemma follows from standard techniques (see [12] for example) and hence proof is omitted.

**Lemma 4:** For any fixed \( n \geq 1 \), let \( p_n \) be as defined by equation (2). Then, \( \mu_n(\eta^*) = \rho_n^{N^*-\eta^*-1}, \mu_n(i) = (1 - \rho_n) \rho_n^{(N^*-i-1)}, \forall i = \eta^* + 1, \ldots, N^* - 1, \mu_n(N^*) = \frac{1 - \rho_n}{\rho_n} \).

Now, we prove our first result on sample complexity.

**Proposition 4:** For any \( \delta_1 \in (0, 1) \) select \( n \) such that

\[
n \geq \frac{2(\kappa)^2}{|\varepsilon_g/\alpha|^2} \log \left( \frac{2|\mathbb{S}|}{\delta_1} \right)
\]

then,

\[
\limsup_{t \to \infty} \mathbb{P}\left\{\|\hat{v}_t^n - v^*\| > \eta \varepsilon_g\right\} \leq 1 - \mu_n(\eta^*) \leq \delta_1.
\]

**Proof:** From Corollary 1, \( \limsup_{t \to \infty} \mathbb{P}\left\{\|\hat{v}_t^n - v^*\| > \eta \varepsilon_g\right\} \leq \mathbb{Q}\{Y_t^n > \eta\} = 1 - \mu_n(\eta^*) \). For \( 1 - \mu_n(\eta^*) \) to be less than \( \delta_1 \), we compute \( n \) using Lemma 4 as,

\[
1 - \delta_1 \leq \mu_n(\eta^*) = \rho_n^{N^*-\eta^*-1} \leq p_n = 1 - 2|\mathbb{S}|e^{-2(\varepsilon_g/\alpha)^2n/(2\kappa)^2}
\]

Thus, \( n \geq \frac{2(\kappa)^2}{|\varepsilon_g/\alpha|^2} \log \left( \frac{2|\mathbb{S}|}{\delta_1} \right) \).

We cannot iterate \( T_n \) forever so we need a guideline for a finite choice of \( t \). This question can be answered in terms of mixing times. The total variation distance between two probability measures \( \mu \) and \( \nu \) on \( \mathbb{S} \) is

\[
||\mu - \nu||_{TV} = \max_{S \subseteq \mathbb{S}} |\mu(S) - \nu(S)| = \frac{1}{2} \sum_{s \in \mathbb{S}} |\mu(s) - \nu(s)|.
\]

Let \( Q_t^n \) be the marginal distribution of \( Y_t^n \) on \( \mathbb{N} \) at time \( t \) and

\[
d(t) = ||Q_t^n - \mu_n||_{TV}
\]

be the total variation distance between \( Q_t^n \) and the steady state distribution \( \mu_n \). For \( \delta_2 > 0 \), we define

\[
t_{mix}(\delta_2) = \min \{ t : d(t) \leq \delta_2 \}
\]

to be the minimum length of time needed for the marginal distribution of \( Y_t^n \) to be within \( \delta_2 \) of the steady state distribution in total variation norm.

We first compute the eigenvalues of the transition matrix \( \Omega \) of the Markov chain \( Y_t^n \).

**Lemma 5:** For any fixed \( n \geq 1 \), the eigenvalues of the transition matrix \( \Omega \) of the Markov chain \( Y_t^n \) are 0 (with algebraic multiplicity \( N^* - \eta^* - 1 \)) and 1.

The computation is done by standard techniques and hence proof is omitted.

Define \( \mu_{n, min} := \min\eta \mu_n(\eta) \). We now bound \( t_{mix}(\delta_2) \):

**Lemma 6:** For any \( \delta_2 > 0 \), \( t_{mix}(\delta_2) \leq \log \left( \frac{1}{\delta_2 \mu_{n, min}} \right) \).

Proof is omitted due to page limitation. Readers are referred to [11] for detailed proof.

We now use the above bound on mixing time to get a non-asymptotic bound for EVI.

**Proposition 5:** For any fixed \( n \geq 1 \),

\[
\mathbb{P}\left\{\|\hat{v}_t^n - v^*\| > \eta \varepsilon_g\right\} \leq 2\delta_2 + (1 - \mu_n(\eta^*))
\]

for \( t \geq \log \left( \frac{1}{\delta_2 \mu_{n, min}} \right) \).

**Proof:** For \( t > t_{mix}(\delta_2) \), \( d(t) = \frac{1}{2} \sum_{\eta=0}^{N^*} |Q(Y_t^n = i) - \mu_n(i)| \leq \delta_2. \) Then, \( |Q(Y_t^n = \eta^*) - \mu_n(\eta^*)| \leq 2d(t) \leq 2\delta_2. \) So, \( \mathbb{P}\left\{\|\hat{v}_t^n - v^*\| > \eta \varepsilon_g\right\} \leq Q(Y_t^n > \eta^*) = 1 - Q(Y_t^n = \eta^*) \leq 2\delta_2 + (1 - \mu_n(\eta^*)). \)

We now combine Proposition 4 and 5 to prove our main theorem.

**Proof of Theorem 1:**

**Proof:** Recall that \( \varepsilon_g = \varepsilon/\alpha \), and \( \delta_1, \delta_2 \) be positive with \( \delta_1 + 2\delta_2 \leq \delta \). By Proposition 4, for \( n \geq n(\varepsilon, \delta) \)

\[
\limsup_{t \to \infty} \mathbb{P}\left\{\|\hat{v}_t^n - v^*\| > \eta \varepsilon_g\right\} = \mathbb{P}\left\{\|\hat{v}_t^n - v^*\| > \varepsilon\right\} \leq 2\delta_2 + (1 - \mu_n(\eta^*)). \]

Combining both we get,

\[
\mathbb{P}\left\{\|\hat{v}_t^n - v^*\| > \varepsilon\right\} \leq \delta
\]

**VI. CONCLUSION**

In this paper we have proposed a simulation based dynamic programming algorithm for MDPs. Unlike most of the other simulation based dynamic programming methods which are either based on heuristics or give only asymptotic result, we provide provable and non-asymptotic performance guarantees for our algorithm. Also, we have developed a new method for the convergence analysis of our simulation-based algorithm using a stochastic dominance argument. We believe that this method is quite general and can be applied to a large class of simulation based algorithms for giving non-asymptotic provable guarantees. This algorithm is outside the family of stochastic approximation algorithms, and raises interesting prospects for the development of new simulation-based algorithms. Our initial simulation studies [11] show fast convergence properties of our algorithm. In future research, we wish to extend this technique to policy iteration, asynchronous value iteration and MDPs on countable and general state spaces.
the last \( \leq \) follows because \( p_n \) is the worst case probability for a one-step improvement in \( \{X_n^t\}_{t \geq 0} \) as observed in Section V.

**Proof of Theorem 2**

*Proof:* Trivially, \( X_n^0 \leq_{st} Y_n^0 \) since \( X_n^0 = Y_n^0 \). Next, we see that \( X_n^1 \leq_{st} Y_n^1 \) by previous lemma. We prove the general case by induction. Suppose \( X_n^t \leq_{st} Y_n^t \) for \( t \geq 1 \), and for this proof define the random variable

\[
\mathcal{Q}(\theta) = \left\{ \max \{\theta - 1, \eta^*\} , \text{ w.p. } p_n, \right. \\
\left. N^*, \quad \text{ w.p. } 1 - p_n, \right. 
\]

as a function of \( \theta \). We see that \( Y_n^{t+1} \) has the same distribution as

\[
[\mathcal{Q}(\Theta) | \Theta = Y_n^t]
\]

by definition. Since \( \mathcal{Q}(\theta) \) are stochastically increasing, we see that

\[
[\mathcal{Q}(\Theta) | \Theta = Y_n^t] \geq_{st} [\mathcal{Q}(\Theta) | \Theta = X_n^t]
\]

by [13, Theorem 1.A.6] and our induction hypothesis. Now, [13, Theorem 1.A.3(d)] for all histories \( \mathcal{F}^t \). It follows that \( Y_n^{t+1} \geq_{st} X_n^{t+1} \) by transitivity.

**APPENDIX**

A. Proof of Theorem 2

Let \( \{\mathcal{F}^t\}_{t \geq 0} \) be the filtration in usual sense and let \( [\cdot | \mathcal{F}^t] \) denote the conditional distribution given the information \( \mathcal{F}^t \).

We first prove the following lemmas.

**Lemma 7:** \( [Y_n^{t+1} | Y_n^t = \theta] \) is stochastically increasing in \( \theta \) for all \( t \geq 0 \), i.e.

\[
[Y_n^{t+1} | Y_n^t = \theta] \leq_{st} [Y_n^{t+1} | Y_n^t = \theta'] \quad \text{for all } \theta \leq \theta'.
\]

*Proof:* We see that

\[
\Pr \{Y_n^{t+1} \geq \eta | Y_n^t = \theta\}
\]

is increasing in \( \theta \) by construction of \( \{Y_n^t\}_{t \geq 0} \). If \( \theta > \eta \), then \( \Pr \{Y_n^{t+1} \geq \eta | Y_n^t = \theta\} = 1 \) since \( Y_n^{t+1} \geq \theta - 1 \) almost surely; if \( \theta \leq \eta \), then \( \Pr \{Y_n^{t+1} \geq \eta | Y_n^t = \theta\} = 1 - p_n \) since the only way \( Y_n^{t+1} \) will remain larger than \( \eta \) is if \( Y_n^{t+1} = N^* \).

**Lemma 8:** \( [X_n^{t+1} | X_n^t = \theta, \mathcal{F}^t] \leq_{st} [Y_n^{t+1} | Y_n^t = \theta] \) for all \( \theta \) and all \( \mathcal{F}^t \) for all \( t \geq 0 \).

*Proof:* Follows from construction of \( \{Y_n^t\}_{t \geq 0} \). For any history \( \mathcal{F}^t \),

\[
\mathcal{P}\{X_n^{t+1} \geq \theta - 1 | X_n^t = \theta, \mathcal{F}^t\} \leq \mathcal{Q}\{Y_n^{t+1} \geq \theta - 1 | Y_n^t = \theta\} = 1
\]

Now,

\[
\mathcal{P}\{X_n^{t+1} = N^* | X_n^t = \theta, \mathcal{F}^t\} \leq \mathcal{P}\{X_n^{t+1} > \theta - 1 | X_n^t = \theta, \mathcal{F}^t\}
\]

\[
= 1 - \mathcal{P}(X_n^{t+1} \leq \theta - 1 | X_n^t = \theta, \mathcal{F}^t)
\]

\[
\leq 1 - p_n
\]