Probabilistic Contraction Analysis of Iterated Random Operators

Abhishek Gupta, Member, IEEE, Rahul Jain, Senior Member, IEEE, and Peter Glynn

Abstract—In many branches of engineering, Banach contraction mapping theorem is employed to establish the convergence of certain deterministic algorithms. Randomized versions of these algorithms have been developed that have proved useful in data-driven problems. In a class of randomized algorithms, in each iteration, the contraction map is approximated with an operator that uses independent and identically distributed samples of certain random variables. This leads to iterated random operators acting on an initial point in a complete metric space, and it generates a Markov chain. In this article, we develop a new stochastic dominance-based proof technique, called probabilistic contraction analysis, for establishing the convergence in probability of Markov chains generated by such iterated random operators in certain limiting regime. The methods developed in this article provide a general framework for understanding convergence of a wide variety of Monte Carlo methods, in which contractive property is present. We apply the convergence result to conclude the convergence of fitted value iteration and fitted relative value iteration in continuous state and continuous action MDPs as representative applications of the general framework developed here.

Index Terms—Convergence of numerical methods, iterative algorithms, Monte Carlo methods, optimization.

I. INTRODUCTION

Let \( (\mathcal{X}, \rho) \) be a complete metric space with the metric \( \rho \). Let \( T : \mathcal{X} \to \mathcal{X} \) be a contraction operator over this space, i.e., there exists an \( \alpha \in [0, 1) \) such that for any \( x, y \in \mathcal{X} \), we have

\[ \rho(T(x), T(y)) \leq \alpha \rho(x, y). \]

According to the Banach contraction mapping theorem [1], for any starting point \( x_0 \in \mathcal{X} \), the sequence generated by \( x_{k+1} = T(x_k), k \in \mathbb{N} \) converges to the unique fixed point \( x^* \) of \( T \). Many recursive algorithms in optimization, such as value and policy iteration algorithm, gradient descent algorithms, and primal-dual algorithms, can be viewed as iterating a contraction operator on Euclidean spaces or complete function spaces [2], [3], [4].

Often, evaluating \( T(x) \) may be computationally challenging or just not possible, particularly when it involves computing expectations of certain functions of random variables. For such cases, Monte Carlo algorithms have been devised, in which random variables are used to approximate the computation of the operator \( T \). Thus, iteration \( k \) of the algorithm is an application of a random operator \( \hat{T}_k \), where \( n \) is the number of random samples used to approximate \( T \) at the \( k \)th iteration and are assumed to be independent of the past samples. Consequently, \( \hat{T}_k \) is independent of the past random operators \( (\hat{T}_k^n)_{k=0}^{k-1} \). Examples of such algorithms are adaptive Monte Carlo algorithms [5], [6], [7], [8], sequential Monte Carlo algorithms [9], [10], temporal difference methods [11], reinforcement learning algorithms [3], [4], [12], approximate dynamic programming for risk sensitive Markov decision problems (MDPs) [13], first-step analysis, and optimal stopping problems. The methods developed in this article provide a general framework for understanding convergence of a wide variety of Monte Carlo methods, in which contractive property is present.

Intuitively speaking, such independent random operators when applied recursively to any starting point \( x_0 \) lead to a random sequence \( (\hat{X}_k^n)_{k=0}^{n-1} \) that drifts toward the fixed point \( x^* \) of \( T \). The goal of this article is to devise a novel framework for understanding asymptotic behavior (that is, consistency) of the iterates generated from such iterated random operators. We introduce the notion of a probabilistic fixed point of the random operators, which is adapted from [14].

Definition 1: A point \( \bar{x} \in \mathcal{X} \) is a probabilistic fixed point of \( (\hat{T}_k^n)_{k=0}^{n-1} \) if for any initial point \( x_0 \in \mathcal{X} \) and for every \( \kappa > 0 \), \( (\hat{X}_k^n)_{k=0}^{n-1} \) satisfies

\[ \lim_{n \to \infty} \limsup_{k \to \infty} \mathbb{P}\left\{ \rho(\hat{X}_k^n, \bar{x}) > \kappa \right\} = 0. \]

In this article, we provide sufficient conditions under which \( x^* \) is a probabilistic fixed point of the iterated random operators. The analysis of iterated random operators was first carried out in [15], [16], and [17]; see also the recent surveys on the topic [18], [19], [20]. Under the assumption that the random...
operators have a negative Lyapunov exponent and the underlying space is Polish, these papers analyzed the stability of the Markov chain generated by the iterated random operators. The authors developed a novel “backward iteration” argument and proved that the Markov chain converges to a unique invariant distribution at a geometric rate. While the rate of convergence could be inferred, it is not helpful when one is interested in sample complexity bounds for recursive stochastic algorithms and compute how far the Markov chain is from the fixed point $x^*$ of the deterministic contraction operator.

Accordingly, this article presents an analysis of iterated random operators with negative Lyapunov exponent satisfying certain consistency properties to complete metric spaces under somewhat different assumptions than in [18]. We show that $x^*$, the fixed point of the deterministic operator, is the probabilistic fixed point of the random operators. Our approach is based on a novel stochastic dominance type argument. We refer to this new approach as the probabilistic fixed point approach to distinguish it from other approaches. We demonstrate applicability of our general framework to empirical value iteration for the discounted-cost and average-cost continuous-state MDPs. Let us first consider some learning examples that can be modeled within the iterated random operator framework.

**Example 1:** Consider the infinite horizon discounted cost MDP, in which $s \in S$ is the state system, $a \in A$ is the control action, $c(s, a)$ is the one stage cost, $\alpha \in (0, 1)$ is the discount factor. The state transition function is $S' = g(S, A, Z)$, where $Z \in \mathbb{Z}$ is the exogenous noise. We assume that the state and the action spaces are finite sets.

Let $\gamma(\cdot|s)$ denote a stationary policy—which is a measure over $A$ for every $s \in S$. Let $\Gamma$ denote the set of all such stationary policies. The goal of the decision maker is to minimize the total discounted cost by solving the following minimization problem:

$$v^*(s) = \min_{\gamma \in \Gamma} \mathbb{E} \left[ \sum_{k=0}^{\infty} \alpha^k c(S_k, A_k) \mid s_0 = s, A_k \sim \gamma(\cdot|S_k) \right].$$

The optimal value function $v^*$ is a fixed point of a contraction operator $T$, which is defined as

$$[T(v)](s) = \min_{a \in A} \left\{ c(s, a) + \alpha \mathbb{E} \left[ v(g(s, a, Z)) \right] \right\}. \quad (1)$$

The operator $T$ is the Bellman operator. It is not difficult to show that $T$ is a contraction operator over the normed vector space $V := \{ v : S \to \mathbb{R} \}$ endowed with the sup norm. The optimal value function $v^*$ is the limit of the value iteration algorithm

$$v_{k+1}(s) = [T(v_k)](s) = \min_{a \in A} \left\{ c(s, a) + \alpha \mathbb{E} \left[ v_k(g(s, a, Z)) \right] \right\}. \quad (2)$$

When computing the expectation exactly is difficult or not possible, the Bellman operator is often approximated using the empirical Bellman operator, $T^n_k : V \to V$ (see, for example, [14]) defined as

$$\hat{v}_{k+1}^n(s) = [\hat{T}_k^n(\hat{v}_k^n)](s) := \min_{a \in A} \left\{ c(s, a) + \frac{1}{n} \sum_{i=1}^{n} \hat{v}_i^n(g(s, a, Z_{k,i})) \right\}. \quad (2)$$

where $\{Z_{k,i}\}_{i=1}^{n}$ are independent samples of the noise variable $Z$. This algorithm is known as empirical value iteration. Since $\{Z_{k,i}\}_{i=1}^{n}$ are independent identically distributed (i.i.d.), we conclude that $T^n_k$ is i.i.d. and $(\hat{v}_k^n)_{k \in \mathbb{N}}$ is a Markov chain.

**Example 2:** Consider the same setting as above, but instead of the discounted cost MDP, we will consider minimizing the average cost

$$\min_{\gamma \in \Gamma} \lim_{K \to \infty} \mathbb{E} \left[ \frac{1}{K+1} \sum_{k=0}^{K} c(S_k, A_k) \mid s_0 = s, A_k \sim \gamma(\cdot|S_k) \right].$$

Under some conditions on the MDP’s state transition function, one can show that the optimal solution exists [21]. In this case, the optimality condition is given by a tuple $(v^*, J^*)$, where $v^*$ is the optimal relative value function and $J^*$ is a real number called the optimal gain. The Bellman operator $T$ satisfies

$$v^*(s) + J^* = [T(v^*)](s) := \min_{a \in A} \left\{ c(s, a) + \mathbb{E} \left[ v^*(g(s, a, Z)) \right] \right\}. \quad (3)$$

Under certain conditions (see, for example, [21, Sec 8.5.2]), one can show that $T$ is a contraction operator over a quotient space with the span seminorm given by

$$\text{span}(v) = \max_{s \in S} v(s) - \min_{s \in S} v(s). \quad (3)$$

The computation of the optimal relative value function $v^*$ is the limit of the following iterative process, which is known as the relative value iteration (RVI) algorithm

$$v_{k+1}(s) = [T(v_k)](s) := \min_{a \in A} \left\{ c(s, a) + \mathbb{E} \left[ v_k(g(s, a, Z)) \right] \right\} - \min_{s \in S} v_k(s). \quad (4)$$

In this case, if the expectation operator is difficult to evaluate, then the empirical relative value iteration (ERVI) [22] is defined as

$$\hat{v}_{k+1}^n(s) = [\hat{T}_k^n(\hat{v}_k^n)](s) := \min_{a \in A} \left\{ c(s, a) + \frac{1}{n} \sum_{i=1}^{n} \hat{v}_i^n(g(s, a, Z_{k,i})) \right\} - \min_{s \in S} \hat{v}_k^n(s). \quad (5)$$

where again, $\{Z_{k,i}\}_{i=1}^{n}$ is a sequence of i.i.d. samples of the noise variable. These noise samples are generated independently from the past samples at every iteration $k$. Once again, we observe that $(\hat{T}_k^n)_{k \in \mathbb{N}}$ is an i.i.d. sequence of operators and $(\hat{v}_k^n)_{k \in \mathbb{N}}$ is a Markov chain.

Based on the two examples above, we observe some important properties of the empirical operator $T_k^n$ defined in (2) and (5). First, consider the empirical Bellman operator for the discounted case in (2). For every $v \in V$ and $\epsilon > 0$, the random operators $(\hat{T}_k^n)_{k,n \in \mathbb{N}}$ satisfy a consistency property for every $k \in \mathbb{N}$

$$\text{PCP}_1 : \lim_{n \to \infty} \mathbb{P} \left\{ \|T_k^n(v) - T(v)\| > \epsilon \right\} = 0 \quad (6)$$

which is established in [14, Prop. 5.1]. In addition to this, it is not difficult to prove that every realization of $T_k^n$ is a contraction operator over the space $V$ with contraction coefficient $\alpha$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Next, we consider the empirical Bellman operator $\hat{T}_k^n$ used in ERVI in (5). This operator satisfies the property stated in (6), where the norm is replaced with the span seminorm. Let $\hat{\alpha}_k^n$ denote the contraction coefficient of $\hat{T}_k^n$. We can show that the random operators $(\hat{T}_k^n)_{k,n \in \mathbb{N}}$ satisfy the following properties in the limit:

$$\text{PCP}_2 : \lim_{n \to \infty} \mathbb{P} \{ \hat{\alpha}_k^n > 1 - \delta \} = 0 \quad \text{for all } \delta \in (0, 1 - \alpha), \ k \in \mathbb{N}$$

and $\hat{\alpha}_k^n \leq 1$ almost surely for all $n \in \mathbb{N}, \ k \in \mathbb{N}$. (7)

We note here that for random operators, the Lyapunov exponent is defined as $\mathbb{E}[\log(\hat{\alpha}_k^n)]$. Thus, the random operators enjoy negative Lyapunov exponent property in the limit as $n \to \infty$.

We refer to the two conditions in (6) and (7) as the probabilistic contraction properties of the random operators. In the next section, we show that these two properties are crucial (in addition to some other hypotheses) in establishing the following result: For any $\kappa > 0$

$$\lim_{n \to \infty} \limsup_{k \to \infty} \mathbb{P} \left\{ \rho(\hat{X}_k^n, x^*) > \kappa \right\} = 0. \quad (8)$$

This is true even when $\mathcal{X}$ is a complete metric space. We then illustrate the application of the general theory developed here to MDPs with continuous state and action spaces. Fitted value iteration with function approximation for discounted cost MDPs with compact state and action spaces are considered in Section III-B. The average cost case is treated in Section III-C. Theorems 4 and 5 establish property (8) for these cases.

A. Related Work

The convergence of random iterates generated from randomized algorithms has been a subject of intense investigation in the past 50 years. The random iterates can either converge to a deterministic quantity, admit an invariant distribution, or occupy certain open balls with high probability asymptotically. These can be construed as the stability property of the random iterates. There are three general methods to determine the stability of the random iterates: the Foster–Lyapunov method, the martingale method, or the ordinary differential equation method. The goal of this article is to develop a fourth method based on the stochastic dominance approach.

To invoke the Foster–Lyapunov method [23, 24], one requires the existence of a coercive Lyapunov function (i.e., whose sublevel sets are compact). If the random operators are independent and satisfy certain drift conditions with respect to the Lyapunov condition (called Foster–Lyapunov conditions), then it can be shown that under reasonable assumptions, the Markov chain admits an invariant distribution; we refer the reader to the texts [23, 25] where this approach is described in detail. In finite dimensional problems, such Lyapunov functions are rather easy to construct. For infinite dimensional spaces, it is rather difficult to devise a coercive Lyapunov function. Thus, this approach is difficult to apply when the underlying state space of the random iterates is a function space.

The martingale and the ordinary differential equation approaches to establishing stability of the random iterates are quite general [26, 27, 28, 29]—in these approaches, we do not require the sequence of random operators to be independent. We only need certain additive errors to be a martingale noise difference sequence with respect to some filtration. This method is applicable mostly to finite dimensional spaces [29]. We refer the reader to recent papers [30, 31] and the references therein for a discussion of this approach to Hilbert spaces. Once again, this approach is difficult to apply over general function spaces.

In contrast, the probabilistic contraction approach developed here can be used when the space of iterates is a complete metric space. However, to derive the convergence result, we assume that the random operators are i.i.d., which is more restrictive than the assumptions placed in the martingale and the ordinary differential equation approaches. We further require this sequence of operators to satisfy the consistency and the negative Lyapunov exponent property [see (6) and (7) above]. We call this approach the probabilistic contraction analysis to distinguish it from the previously studied approaches.

B. Our Contribution

The main result establishes the probabilistic fixed point of the random operators under certain reasonable assumptions. We prove this result using a stochastic dominance approach. The key benefits of the new approach we develop are as follows.

1) Our approach works when the underlying space $\mathcal{X}$ is any complete metric space without a need to construct a Lyapunov function.

2) Since we use stochastic dominance approach of the error process $\rho(\hat{X}_k^n, x^*)$ to derive the probabilistic fixed point property, we avoid certain measurability issues that would typically arise if we were working with the random operators in the original state space $\mathcal{X}$.

3) If $\mathcal{X}$ is a bounded complete metric space, then we obtain the rate of convergence and the sample complexity of getting close to the probabilistic fixed point with high probability.

The rest of this article is organized as follows. In Section II, Theorem 1 bounds $\limsup_{k \to \infty} \mathbb{P} \{ \rho(\hat{X}_k^n, x^*) \geq \kappa \}$ for $\kappa > 0$. This result in proved in Section IV, where we use the theory of stochastic dominance [32] to derive an upper bound on the probability of error being larger than $\kappa$ in the limit $k \to \infty$. We apply the results to study the convergence of the empirical value iteration algorithm for an MDP with the average cost criterion. In Section II-A, we consider the case where the contraction operator comprises of iterated composition of many nonexpansive mappings. We identify some sufficient conditions on the individual mappings and their randomized counterparts, so that the composite operator and their randomized versions satisfy the assumption of Theorem 1. In Section II-B, we present a sample complexity result to obtain a high confidence solution for probabilistic fixed point. We apply these results to determine the convergence of fitted value iteration for discounted and average cost MDPs with compact state and action spaces in Section III. We present proofs of the main results in Section IV. Finally, Section V concludes this article.
II. MAIN RESULTS

Let us now formulate the problem precisely. Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a standard probability space and \(\mathcal{X}\) be a complete metric space. Define \(\mathbb{T}_n^k : \mathcal{X} \times \Omega \to \mathcal{X}\) over this probability space such that for each \(n \in \mathbb{N}\), the operators \(\mathbb{T}_n^k (x)\) and \(\mathbb{T}_n^k (\omega)\) are independent of each other and identically distributed for \(k \neq k'\) and for all \(x \in \mathcal{X}\) (we suppress here the dependence of \(\mathbb{T}_n^k\) on \(\omega\) for easing the exposition). Let \(\hat{\alpha}_n^k : \Omega \to [0, 1]\) and \(\hat{\zeta}_n^k : \Omega \to [0, \infty)\) be measurable functions such that

\[
\rho(\mathbb{T}_n^k (x_1), \mathbb{T}_n^k (x_2)) \leq \hat{\alpha}_n^k \rho(x_1, x_2) + \hat{\zeta}_n^k \tag{9}
\]

for all \(x_1, x_2 \in \mathcal{X}\) and \(n, k \in \mathbb{N}\). Here, \(\hat{\zeta}_n^k\) could be bias and errors introduced due to certain approximations (such as function approximation or biased approximation of operators). Due to independence of the operators \((\mathbb{T}_n^k)_{k \in \mathbb{N}}\), for every \(n \in \mathbb{N}\), \((\hat{\alpha}_n^k, \hat{\zeta}_n^k)_{k \in \mathbb{N}}\) is an i.i.d. tuple.

If \(\hat{\zeta}_n^k \equiv 0\), then \(\hat{\alpha}_n^k\) denotes a measurable upper bound on the contraction coefficient of \(\mathbb{T}_n^k\). In this case, \(\hat{\alpha}_n^k\) can be defined as

\[
\hat{\alpha}_n^k = \sup_{x_1 \neq x_2} \rho(\mathbb{T}_n^k (x_1), \mathbb{T}_n^k (x_2)) / \rho(x_1, x_2).
\tag{10}
\]

However, it is not obvious that \(\hat{\alpha}_n^k\) thus defined is measurable since \(\mathcal{X}\) is uncountable. It has been shown in [18, Lemma 5.1] that if \(\mathcal{X}\) is separable and complete, then \(\hat{\alpha}_n^k\) as defined in (10) is indeed measurable. For our applications, it is generally easy to derive a measurable map \(\hat{\alpha}_n^k\) such that (9) holds.

There is no reason to believe that for any sequence of random operators, a probabilistic fixed point would exist. However, we have already seen two conditions that are met in the learning algorithms that inspired this work: the consistency condition in (6) and the negative Lyapunov exponent condition in (7). Our primary interest in this article is in identifying further restrictions on the random operators, under which probabilistic fixed point exists. We show that the probabilistic fixed point of \((\mathbb{T}_n^k)_{n,k}\) coincides with the fixed point of \(T\). First, we need to place the following conditions on the random operators.

Assumption 1: The following holds.

i) \(\mathcal{X}\) is a complete metric space.

ii) \(T : \mathcal{X} \to \mathcal{X}\) is a contraction operator with contraction coefficient \(\alpha < 1\). Let \(x^*\) denote the fixed point of \(T\).

iii) For each \(n \in \mathbb{N}\), \((\mathbb{T}_n^k)_{k \in \mathbb{N}}\) is a sequence of i.i.d. operators.

iv) For any \(k \in \mathbb{N}\) and \(\delta \in (0, 1 - \alpha)\)

\[
\lim_{n \to \infty} \mathbb{P} \{ \hat{\alpha}_n^k \geq 1 - \delta \} = 0
\]

and \(\hat{\alpha}_n^k \leq 1\) almost surely for all \(n, k \in \mathbb{N}\).

v) For any \(k \in \mathbb{N}\) and \(\epsilon > 0\)

\[
\lim_{n \to \infty} \mathbb{P} \{ \rho(\mathbb{T}_n^k (x^*), \mathbb{T}_n^k (x')) + \hat{\zeta}_n^k \geq \epsilon \} = 0.
\]

vi) There exists \(\bar{w} > 0\) such that \(\rho(\mathbb{T}_n^k (x^*), \mathbb{T}_n^k (x')) + \hat{\zeta}_n^k \leq \bar{w}\) almost surely for every \(n \in \mathbb{N}\) and \(k \in \mathbb{N}\).

In our experience, Assumption 1 can be ascertained in practice as follows. Assumption 1(i)–(iii) is routine and easy to ascertain in most cases. Assumption 1(iv) and (v) is typically established using a concentration of measures result [25], [33] or empirical processes theory [34], [35]. Computing tight bounds on \(\bar{w}\) used in Assumption 1(vi) is the one that requires some effort. Based on the assumption, we have the following result, whose proof is presented in Section IV.

Theorem 1: Suppose that Assumption 1 holds. Fix \(\kappa > 0\) and pick \(\epsilon, \delta > 0\) in the set

\[
B := \left\{ \epsilon \in \left(0, \frac{\kappa (1 - \alpha)}{2}\right), \delta \in (0, 1 - \alpha) : \frac{2}{\delta} \leq \frac{\kappa}{\epsilon} \right\}.
\]

(11)

Let \(\beta = \left[\frac{\kappa}{\epsilon}\right]\). Pick \(n\) sufficiently large so that \(p_{n, \delta}^\alpha > \frac{2 \beta}{2^n + 1}\) can be picked satisfying

\[
p_{n, \delta}^\alpha \leq \inf_{\kappa \in \mathbb{N}} \mathbb{P} \left\{ \hat{\alpha}_n^k \leq 1 - \delta, \rho(\mathbb{T}_n^k (x^*), \mathbb{T}(x^*)) + \hat{\zeta}_n^k \leq \epsilon \right\}.
\]

(12)

Then, we have

\[
\lim_{n \to \infty} \sup_{k \to \infty} \mathbb{P} \left\{ \rho(\mathbb{T}_n^k (x^*), x^*) > \kappa \right\} \leq \frac{1 - (p_{n, \delta}^\alpha)^\beta}{p_{n, \delta}^\alpha \beta}.
\]

This further implies

\[
\lim_{n \to \infty} \sup_{k \to \infty} \mathbb{P} \left\{ \rho(\mathbb{T}_n^k (x^*), x^*) > \kappa \right\} = 0.
\]

Consequently, the point \(x^*\) is the unique probabilistic fixed point of the random operators \((\mathbb{T}_n^k)_{n,k}\).

Let us illustrate the application of the theorem by applying it on Example 1.

Example 3: Recall the notation in Example 1. Here, we have \(\mathcal{X} = \mathbb{R}^\mathcal{D}\) endowed with the \(\ell_\infty\) norm, \(\hat{\alpha}_n^k \equiv \alpha\), and \(\hat{\zeta}_n^k = 0\) almost surely. We have

\[
||\mathbb{T}_n^k (v^*) - T(v^*)||_\infty = W_k^n
\]

\[
\leq \max_{s,a,Z} \left\{ \frac{1}{n} \sum_{i=1}^{n} v^*(g(s, a, Z_{i,k})) - E[v^*(g(s, a, Z))] \right\}.
\]

Using the Hoeffding inequality and the union bound, we have

\[
\mathbb{P} \{ W_k^n > \epsilon \} \leq 2 |S| |A| \exp \left( - \frac{n \epsilon^2}{2 \alpha^2 ||v^*||_\infty^2} \right).
\]

Further, \(W_k^n \leq \bar{w} := 2 \frac{\kappa^2}{\alpha^2} - \epsilon\). Pick \(\delta = 1 - \alpha\), \(\epsilon = \frac{\kappa (1 - \alpha)}{2}\), we conclude that \(v^*\) is a probabilistic fixed point of the operators \((\mathbb{T}_n^k)_{n,k}\). This result was derived in [14, Th. 3.1] by exploiting the structure of the problem. Here, we have used the general result of Theorem 1 to arrive at the conclusion.

We would like to note two interesting observations here. The first one is that under this assumption, we require the consistency condition, Assumption 1(v), to hold only at \(x^*\), not over the entire space. We can exploit some specific property of \(x^*\) to establish the consistency result. For instance, in a class of MDPs, the space of value functions \(\mathcal{V}\) could be the space of bounded measurable functions, but the optimal value function \(v^*\) could be a Lipschitz continuous function (see, for example, [36] and [37]). Thus, we do not need to show the consistency condition for all bounded measurable functions, but just for one Lipschitz continuous function \(v^*\).
The second observation is that \( \bar{w} \), defined in Assumption I(vi), can be made small by an appropriate choice of the random operators \( \{T^n_k\}_{k \in \mathbb{N}} \). For instance, almost all the variance reduction algorithms we have seen in the literature, for example [38], [39], [40], and [41], attempt to make \( \bar{w} \) as small as possible by a carefully constructed random operator \( \bar{T}^n_k \). To see this, note that if \( \bar{w} = 0 \) and \( n \) is sufficiently large so that the random operators have a negative Lyapunov exponent, then \( \beta = 0 \) and for any \( \kappa > 0 \), we have

\[
\limsup_{k \to \infty} \mathbb{P} \left\{ \rho \left( \bar{T}^n_k(x^*), T(x^*) \right) \geq \frac{1 - (p_{\kappa, \delta})^\beta}{p_{\kappa, \delta}^\beta} \right\} = 0.
\]

We present a more complete discussion on this part with several examples in a recent paper of the first author [42].

Remark 1: A lower bound in (12) can be derived from Frechet–Hoeffding Copula bound [43, Th. 3.1.1]

\[
\mathbb{P}(\hat{\alpha}^n_k \leq 1 - \delta, \rho(\hat{T}^n_k(x^*), T(x^*)) + \hat{\xi}^n_k \leq \epsilon) \\
\geq \mathbb{P}(\hat{\alpha}^n_k \leq 1 - \delta) + \mathbb{P}(\rho(\hat{T}^n_k(x^*), T(x^*)) + \hat{\xi}^n_k \leq \epsilon) - 1 \\
= 1 - \mathbb{P}(\hat{\alpha}^n_k > 1 - \delta) - \mathbb{P}(\rho(\hat{T}^n_k(x^*), T(x^*)) + \hat{\xi}^n_k > \epsilon).
\]

The first term on the right-hand side can be derived from the usual concentration of measure results. The second term on the right-hand side can be derived from covering number argument frequently used in empirical process theory [35], [44] or function approximation theorems [45], [46]. These ideas have been employed in Section III of this article as well.

We now proceed to develop two other consequences of the main result. In the next section, we consider the case when the contraction operator is a composition of multiple maps, which are approximated by random maps to yield the corresponding random operators. In the subsequent section, we derive the sample complexity result for the convergence to probabilistic fixed point.

A. Composition of Operators

In many complex optimization and reinforcement learning problems, the operator \( T \) may be a composition of multiple mappings, each of which may be nonexpansive or contraction. We may need to replace each of these mappings using random approximations. Our goal in this section is to study such cases and identify sufficient conditions on the random mappings that implies Assumption 1.

To be precise, let \( \mathcal{X}_1, \ldots, \mathcal{X}_{L+1} \) be \( L + 1 \) metric spaces with \( \mathcal{X}_1 = \mathcal{X}_{L+1} = \mathcal{X} \), which is, by assumption, a complete metric space. Let \( \rho_l \) denote the metric on the space \( \mathcal{X}_l \). Let \( T_l : \mathcal{X}_l \to \mathcal{X}_{l+1}, l = 1, \ldots, L \) be a sequence of mappings such that \( T_l \) is Lipschitz with coefficient \( \alpha_l \in [0, 1] \), i.e.,

\[
\rho_{l+1}(T_l(x), T_l(y)) \leq \alpha_l \rho_l(x, y).
\]

Define \( T := T_L \circ \cdots \circ T_1 \), and let \( \alpha := \prod_{l=1}^L \alpha_l \) denote the contraction coefficient of \( T \) (we assume here that \( \alpha < 1 \)). We do not assume completeness of \( \mathcal{X}_2, \ldots, \mathcal{X}_L \). Let \( x^* \) be the fixed point of \( T \). Define \( x^*_1 = x^* \) and recursively define

\[
x^*_{l+1} = T_l(x^*), \quad l = 1, \ldots, L.
\]

It is obvious that \( x^*_{L+1} = x^* \).

Let \( \hat{T}^n_{k,l} : \mathcal{X}_l \to \mathcal{X}_{l+1} \) be the random mapping at time \( k \) that replaces the map \( T_l \) (here, \( l = 1, \ldots, L \)). Define \( \hat{T}^n_k := \hat{T}^n_{k,L} \circ \cdots \circ \hat{T}^n_{k,1} \) to the random operator that approximates \( T \). Let \( \hat{\alpha}^n_{k,l} : \Omega \to [0, 1] \) denote the measurable map such that \( \hat{\alpha}^n_{k,l} \) satisfies for \( l = 1, \ldots, L - 1 \)

\[
\rho_{l+1} \left( \hat{T}^n_{k,l}(x^*_{l+1}), \hat{T}^n_{k,l}(x^*_l) \right) \leq \hat{\alpha}^n_{k,l} \rho(x^*_{l+1}, x^*_l)
\]

\[
\rho_{L+1} \left( \hat{T}^n_{k,L}(x^*_{L+1}), \hat{T}^n_{k,L}(x^*_L) \right) \leq \hat{\alpha}^n_{k,L} \rho(x^*_{L+1}, x^*_L) + \hat{\xi}^n_k
\]

for all \( x^*_{l+1}, x^*_l \in \mathcal{X}_l, l = 1, \ldots, L \). We make the following assumption on the random mappings.

Assumption 2: The following holds.

i) For each \( l = 1, \ldots, L \), the sequence \( \{\hat{T}^n_{k,l}\}_{k \in \mathbb{N}} \) is a sequence of i.i.d. mappings. Each mapping \( \hat{T}^n_{k,l} \) is independent of all of the other mappings.

ii) For every \( n, k \in \mathbb{N}, \hat{\alpha}^n_{k,l} \leq 1 \) almost surely. Further, for every \( k \in \mathbb{N} \) and \( \delta \in (0, 1 - \alpha) \)

\[
\lim_{n \to \infty} \mathbb{P} \left\{ \min_l \hat{\alpha}^n_{k,l} \geq 1 - \delta \right\} = 0.
\]

iii) For every \( l = 1, \ldots, L - 1 \), every \( k \in \mathbb{N} \), and \( \epsilon > 0 \)

\[
\lim_{n \to \infty} \mathbb{P} \left\{ \rho \left( \hat{\alpha}^n_{k,l}(x^*_{l+1}), T_l(x^*_l) \right) \geq \epsilon \right\} = 0.
\]

iv) For each \( l = 1, \ldots, L \), there exists \( \bar{w}_l > 0 \) such that

\[
\rho(\hat{T}^n_{k,l}(x^*_l), T_l(x^*_l)) \leq \bar{w}_l \quad \text{for} \quad l = 1, \ldots, L - 1
\]

\[
\rho(\hat{T}^n_{k,L}(x^*_L), T_L(x^*_L)) + \hat{\xi}^n_k \leq \bar{w}_L \quad \text{almost surely for every} \quad n \in \mathbb{N} \text{ and} \quad k \in \mathbb{N}.
\]

Theorem 2: If Assumption 2 is satisfied, then the composite operators \( T \) and \( \hat{T}^n_k \) satisfies Assumption 1.

Proof: It is clear that by hypotheses, Assumption 1(i) and (ii) is satisfied. We refer the reader to Section IV-D, where we establish that Assumption 1(iii)–(vi) is also satisfied.

Typically, all \( \{\hat{T}^n_{k,l}\}_{k=1}^L \) may not use the same number of samples \( n_l \). In these cases, we can replace the number of samples \( n_l \) as a function of \( n \) such that \( n_l(n) \) monotonically diverges to infinity as \( n \) diverges to infinity.

B. Sample Complexity of Computing Probabilistic Fixed Point

To establish Theorem 1, we employ the theory of stochastic dominance. We construct a dominating Markov chain over the space of natural numbers whose invariant distribution dominates the probability of error being large as \( k \to \infty \). The dominating Markov chain construction can provide us with a sample complexity bound on the number of samples required to get close to the fixed point \( x^* \) with high probability. For a fixed \( \kappa > 0 \) and
confidence level \( \xi \), let \( N_{n,\xi} \) be defined as

\[
N_{n,\xi} = \inf_{(\epsilon,\delta) \in B} \left\{ n \in \mathbb{N} : \mathbb{P}\{\tilde{\alpha}_k^n \leq 1 - \delta, W_k^n \leq \epsilon \} \geq \frac{1}{1 + \xi} \right\}
\]

where \( B \) is defined in (11), and we define \( W_k^n := \rho(T_k^n(x^*), T(x^*)) + \tilde{\xi}_k^n \). Then, for any \( n \geq N_{n,\xi} \), we have

\[
\limsup_{k \to \infty} \mathbb{P}\{\rho(\tilde{X}_k^n, x^*) \geq \kappa \} \leq \xi.
\]

The proof of the above result follows immediately from Theorem 1 and Proposition 4, which is established later in Section IV. We note here that the sample complexity is only in terms of the number of samples \( n \) picked to approximate the operator at every time step; we allow the number of iterations to go to infinity, which implies that the true sample complexity is infinite.

To provide a finite sample complexity bound, we further need to bound the space as well, that is, we assume \( \mathcal{X} \) to be a bounded space satisfying

\[
D := \text{diam}(\mathcal{X}) = \sup_{x,y \in \mathcal{X}} \rho(x, y) < \infty.
\]

This condition is readily satisfied in many optimization and empirical dynamic programming algorithms, where the operator \( T_k^n \) is a contraction operator with the contraction coefficient \( \alpha \) almost surely, and one knows the region in the space \( \mathcal{X} \) where the solution lies.

This is also satisfied if we allow the random operator to project the iterates back to \( \mathcal{X} \) if the iterates become large. In this case, we have the composition of a projection operator and a random operator. It is well-known that if \( \mathcal{X} \) is a Hilbert space [say Euclidean space with the \( \ell_2 \) norm or a reproducing kernel Hilbert space (RKHS)], then the projection operation is nonexpansive. In case \( \mathcal{X} \) is not a Hilbert space, then the projection operation must be picked carefully so that it remains a nonexpansive operator. In this situation, the property of the random operators remain unchanged due to the projection operation.

In the following theorem, we present a sample complexity bound for this case, which is based on the analysis in [14] and has been used in [47], [48], [49], [50], and [51] to provide the sample complexity bounds for various empirical dynamic programming algorithms with function approximation. For \( \kappa > 0 \), pick \((\epsilon,\delta) \in B\), where \( B \) is defined in (11). Let \( d := \lceil D/\epsilon \rceil \), \( p := p_{n,\delta,\epsilon} \), which satisfies (12), and \( \eta := \frac{3}{\delta} \). For confidence level \( \xi > 0 \), define \( N_{n,\xi} \) and \( K_{n,\xi} \) as

\[
N_{n,\xi} := \inf \left\{ n \in \mathbb{N} : p \geq \max \left\{ \frac{1}{2}, \left(1 - \frac{\xi^2}{3} \right)^{(1/(d - \eta - 1))} \right\} \right\}
\]

\[
K_{n,\xi} := \log \left( \frac{3}{\xi(1 - p)pd^{\eta - 1}} \right).
\]

**Theorem 3:** Suppose that Assumption 1(i)–(v) holds with a bounded complete metric space \( \mathcal{X} \) with diameter \( D \). Then, we have

\[
\mathbb{P}\{\rho(\tilde{X}_k^n, x^*) > \kappa \} < \xi \text{ for all } k \geq K_{n,\xi} \text{ and } n \geq N_{n,\xi}.
\]

**Proof:** See Section IV-E.

We note here that for computing \( N_{n,\xi} \), one can use the lower bound stated in Remark 1 to derive the sample complexity bound. The term \( \mathbb{P}\{\tilde{\alpha}_k^n > 1 - \delta \} \) can be bounded either using Hoeffding inequality or may not need to be bounded at all if \( \tilde{\alpha}_k^n \) is a constant. On the other hand, \( \mathbb{P}\{\rho(T_k^n(x^*), T(x^*)) + \tilde{\xi}_k^n > \epsilon \} \) can be bounded by bounding the biasedness introduced due to the random operator and using the empirical process theory, which requires knowledge of the covering number of the space \( \{T_k^n(x^*)\} \). Sample complexity results for various random operators defined the context of empirical value iteration are discussed in [14], [22], [47], [48], [49], and [52], among several others.

### III. Application to Fitted Value Iteration

We now apply the general framework developed above to derive the probabilistic fixed point property of the fitted value iteration for continuous state continuous action MDPs. In this setting, the MDP model and the cost function are known; further, there is a generative model (a simulator) of the MDP that generates actuation noises, which can be used to generate the next state and the cost incurred given the current state-action pair. Fitted value iteration is very useful for approximately solving MDPs where the state and the action spaces are high dimensional (50 or more dimensional). This problem has been studied under different names within the reinforcement learning literature. The authors in [40], [53], [54], and [55] considered a finite-state finite-action setting, and refer to this algorithm as an MDP with a generative model (no function approximation is used here). Munos and Szepesvári [52] considered the MDPs with a continuous-state finite-action setting with function approximation and refers to this algorithm as fitted value iteration. In our previous work, this algorithm has been referred to as empirical value iteration [14], [22], [47], [48], [49], [50], [51]. In these previous works, a variety of sufficient conditions are determined under which the optimal value function of the MDP is the probabilistic fixed point of the random operators. In this section, we use the techniques developed here to show essentially the same results, albeit under different assumptions on the MDPs.

#### A. Fitted Value Iteration

We consider here the same model as introduced in Examples 1 and 2 in Section I, but with a compact state and action setting. In this setting, the value functions belong to a function space, and function approximators, such as neural networks or nonparametric regression are needed to store the value function in a computer. Thus, we need to project the value functions to appropriate function approximating class. In the sequel, we used both the transition probability matrix \( P(ds'|s,a) \) induced from the state space model \( s' = g(s, a, z) \).

Recall the Bellman operator \( T \) from (1) in Examples 1. In the context of fitted value iteration, we view the operator \( T \) as a composition of an identity operator and the Bellman operator. Since a
Let \( F \) denote an approximating class, which is a set of functions that maps the output of \( H_k^n \) to a continuous and bounded function \( \mathcal{V} \). The fitted value iteration operator is defined as \( G_k^n(v) : = \Pi_k^n \circ H_k^n(v) \). The relative fitted value iteration map is defined as \( G_k^n(v) : = \Pi_k^n \circ G_k^n(v) \). We study their probabilistic fixed point properties next.

**B. Fitted Value Iteration for Discounted MDPs**

Let \( \rho_S \) denote a metric on \( S \), \( \rho_A \) denote a metric on \( A \) and \( \rho_{S \times A} \) denote the metric on \( S \times A \) as defined in [36].

**Assumption 3:** The following holds.

i) \( S \) and \( A \) are compact subsets of the Euclidean spaces.

ii) \( g \) is Lipschitz continuous in \((s, a) \in S \times A \) for every \( z \in Z \) with Lipschitz coefficient \( L_g(z) \) and \( L_g := \int \! L_g(z)P\{dz\} < 1/\alpha \).

iii) The cost function \( c(s, a) \) is Lipschitz continuous with Lipschitz coefficient \( L_c \).

We now make the following assumption on the projection operator.

**Assumption 4:** The projection operator \( \Pi_k^n : \mathcal{L}_\infty(S) \rightarrow \mathcal{V} \) satisfies the following two conditions.

i) \( \Pi_k^n \) is nonexpansive with a bias, that is, \( \| \Pi_k^n(v_1) - \Pi_k^n(v_2) \|_{\infty} \leq \| v_1 - v_2 \|_{\infty} + \zeta_k \), where \( \{\lim_{n \to \infty} \Pi_k^n = 0\} = 1 \) for all \( k \in \mathbb{N} \) and \( \zeta_k \leq \zeta \) for all \( n, k \in \mathbb{N} \) almost surely.

ii) For any \( \epsilon > 0 \) and \( \delta > 0 \), there exists \( N_2 := N_2(\epsilon, \delta) \) such that

\[
\mathbb{P}\left\{ \| \Pi_k^n(v^*) - v^* \|_{\infty} > \epsilon \right\} < \delta \quad \text{for all} \quad n \geq N_2.
\]

Consider Assumption 4(i) and note the following inequality:

\[
\| \Pi_k^n(v_1) - \Pi_k^n(v_2) \|_{\infty} \leq \| v_1 - v_2 \|_{\infty} + \| \Pi_k^n(v_1) - v_1 \|_{\infty} + \| \Pi_k^n(v_2) - v_2 \|_{\infty}.
\]

For many function approximating class, such as a neural network, k-nearest neighbor, and RKHS, the nonexpansiveness of the projection operation holds under the limit where 1) number of samples used for function fitting goes to infinity, 2) the size of the function approximating class goes to infinity, and 3) the function approximating class contains the value functions throughout the simulation.

To establish that Assumption 4(ii) holds in a problem, we need to exploit some property of the optimal value function (such as Lipschitz or smoothness [45]). In this case, if the loss function \( L(\cdot, \cdot) \) is convex in the parameters used in the function approximating class \( F_m \), then stochastic gradient descent could yield the minimizing function in the function approximating class (sometimes using decreasing stepsizes). For situations where the loss function is nonconvex, the estimation of \( \| \Pi_k^n(v) - v \|_{\infty} \) may be difficult and remains an open problem.
With these assumptions in place, we have the following result.

**Theorem 4:** If Assumptions 3 and 4 hold, then \( v^* \) is a probabilistic fixed point of \((\hat{T}^n_k)_{n,k \in \mathbb{N}}\).

**Proof:** We show in Section III-D that \( \hat{T}^n_k \) satisfies Assumption 1. The result then immediately follows from Theorems 2 and 1.

The above result differs from [52, Th. 2, Corollary 4] and is applicable to MDPs with continuous action spaces. Munos and Szepesvári [52] assume discounted-average concentrability of future-state distributions (see [52, Assumption 2]) and conducted the analysis under \( \ell_p \) norm. This assumption does not hold for any stable deterministic control system (since the transition kernel is not absolutely continuous with respect to any base measure). In our result above, we are not making this strong assumption on the transition kernel. However, we make strong assumptions on the projection operator being biased nonexpansive operator with respect to the \( \ell_\infty \) norm, with bias term converging to 0 as the sample size diverges and function approximating class expands. We further assume \( \ell_\infty \) norm over the value function space. Consequently, Theorem 4 can be applied to deterministic dynamical systems as well (in which case the noise is deterministic). Thus, Theorem 4 and [52, Th. 2, Corollary 4] complement each-other.

**D. Proof of Theorem 4**

It is also clear that \( \mathcal{V} \), the space of continuous and bounded value functions over \( \mathcal{S} \) endowed with the sup norm, is a complete metric space and that the Bellman operator \( T \) is a contraction operator with fixed point \( v^* \). Due to the construction of the empirical Bellman operator, we conclude that \( \hat{T}^n_k \) is i.i.d. Every realization of the empirical Bellman operator \( \hat{H}^n_k \) is a contraction operator with contraction coefficient \( \alpha \) for all \( n, k \in \mathbb{N} \). This yields together with Assumption 4(i) that \( \hat{T}^n_k \) is almost surely a contraction for all \( n, k \in \mathbb{N} \). This further yields

\[
\| \hat{T}^n_k(v^*) - T(v^*) \|_\infty \leq 2\alpha \| v^* \|_\infty + \zeta.
\]

Thus, to establish Theorem 4, we only need to show that Assumption 1(\( v \)) is satisfied. We establish this fact next.

**Proposition 1:** If Assumption 3 is satisfied, then for any \( \epsilon, \delta > 0 \), there exists \( N_1 := N_1(v^*, \epsilon, \delta) \) such that

\[
\mathbb{P} \left\{ \| \hat{H}^n_k(v^*) - T(v^*) \|_\infty \geq \epsilon \right\} < \delta \quad \text{for all} \quad n \geq N_1.
\]

**Proof:** If Assumption 3 holds, then \( v^* \) is a Lipschitz continuous function with Lipschitz coefficient \( L_{v^*} \leq \frac{L_P}{\alpha} \) from [36, Th. 4.2(c), p. 13]. To establish the result, we need to introduce some notations. Let \( \mathcal{D} \) be the set of functions given by

\[
\mathcal{D} = \{ d : \mathcal{Z} \to \mathbb{R} : \exists (s, a) \text{ such that } d(z) = v^*(g(s, a, z)) \} = \bigcup_{(s, a)} \{ v^*(g(s, a, \cdot)) \}.
\]

We note that

\[
\| \hat{H}^n_k(v^*) - T(v^*) \|_\infty
\]

\[
= \max_{s \in \mathcal{S}} \min_{a \in \mathcal{A}} \left( c(s, a) + \frac{\alpha}{n} \sum_{i=1}^{n} v^*(g(s, a, Z_{k,i})) \right)
\]

\[
- \min_{a \in \mathcal{A}} \left( c(s, a) + \alpha \int v^*(g(s, a, z)) \mathbb{P} \{ dz \} \right)
\]

\[
\leq \sup_{s \in \mathcal{S}, a \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} v^*(g(s, a, Z_{k,i})) - \int v^*(g(s, a, z)) \mathbb{P} \{ dz \} \right|
\]

\[
= \sup_{d \in \mathcal{D}} \left| \frac{1}{n} \sum_{i=1}^{n} d(Z_{k,i}) - \int d(z) \mathbb{P} \{ dz \} \right|.
\]

This immediately yields

\[
\mathbb{P} \left\{ \| \hat{H}^n_k(v^*) - T(v^*) \|_\infty > \epsilon \right\}
\]
\[ \Pr \left\{ \sup_{d \in D} \left| \frac{1}{n} \sum_{i=1}^{n} d(Z_{k,i}) - \int d(z) \Pr \{ dz \} \right| > \epsilon \right\} \leq \Pr \left\{ \sup_{d \in D} \left| \frac{1}{n} \sum_{i=1}^{n} d(Z_{k,i}) - \int d(z) \Pr \{ dz \} \right| > \epsilon \right\}.

To obtain bounds on the right-hand side of the equation above, we need to show that the bracketing number of the class of functions \( D \) is finite for every \( \epsilon > 0 \). The finiteness of bracketing number is established in [35, Th. 2.7.11, p. 164]. To see this, let us write \( d_{s,a}(z) = v^*(g(s, a, z)) \). Since \( v^* \) and \( g \) are Lipschitz continuous functions, we have

\[ |d_{s,a}(z) - d_{s',a'}(z)| \leq L_{v^*} L_{g}(z) \rho_{S \times A}((s, a), (s', a')). \]

This immediately implies that the bracketing number of \( D \) is upper bounded by the covering number of \( S \times A \) by [35, Th. 2.7.11, p. 164]. Since \( S \times A \) is compact, it is totally bounded; thus, its covering number is finite for any \( \epsilon > 0 \).

We now invoke [35, Th. 2.4.1, p. 122] to conclude that for any \( k \in \mathbb{N} \)

\[ \lim_{n \to \infty} \Pr \left\{ \sup_{d \in D} \left| \frac{1}{n} \sum_{i=1}^{n} d(Z_{k,i}) - \int d(z) \Pr \{ dz \} \right| > \epsilon \right\} = 0 \]

which yields the desired result.

Remark 2: If one knows the exact covering number of \( S \times A \) under the metric \( \rho_{S \times A} \), then the sample complexity bound can also be given. For more information on covering and bracketing numbers, we refer the reader to [34, 35, and 62]. Once we know the bracketing number of the function class \( D \) under the sup norm, then we can apply Lemma 9.1 of [44] to derive the sample complexity result for the EDP algorithm.

E. Proof of Theorem 5

The proof of this result follows from minor modifications in the proof of Theorem 4 above. We define \( \mathcal{V}/ \sim \) to be the quotient space

\[ \mathcal{V}/ \sim = \left\{ [v] \in C_b(S) : v_1, v_2 \in [v] \text{ implies } v_1 - v_2 \text{ is a constant function} \right\}. \]

Under these assumptions, we establish the following result.

Proposition 2: The quotient space \( \mathcal{V}/ \sim \) is a Banach space.

If Assumption 5 holds, then we have the following.

1) \( T : \mathcal{V}/ \sim \to \mathcal{V}/ \sim \) is a contraction operator with coefficient \( \alpha \) and \( \sup_{[v^*]} \mathcal{V}/ \sim \).

2) Any representative element \( \hat{v}^* \in [v^*] \) satisfies \( \text{span}(\hat{v}^*) \leq \text{span}(v^*) / \alpha \).

3) If in addition, Assumption 3 holds with \( L_P < 1 \), then any representative element \( v^* \) that lies within the unique equivalence class of the fixed points of \( T \) is Lipschitz continuous with Lipschitz coefficient \( L_P / \alpha \).

Proof: The fact that the quotient space \( \mathcal{V}/ \sim \) is a Banach space is established in [4, Th. 6.26, p. 151]. The operator \( T \) is a span norm contraction, which is established in [4, Th. 6.28, p. 154]. The only thing that needs to be established is that the value function lies in the class of continuous functions. Since our action space is compact, the continuity of the sequence of value functions generated by the value iteration algorithm follows from [63, Lemma 4.3] or from Berge’s maximum theorem [1, Th. 17.31, p. 570]. The Lipschitz continuity of \( v^* \) follows from [36, Th. 4.2(c), p. 13].

In what follows, we drop the notation \([v]\) to denote an element of \( \mathcal{V}/ \sim \) and instead just use \( v \), with the understanding that any representative element \( \hat{v} \in [v] \) can be taken to perform the operation involved. Due to Proposition 2 above, we conclude that Assumption 1(ii)–(iii) holds. Using Proposition 2(c) along with Proposition 1 allows us to conclude that Assumption 1(v) holds. Assumption 1(vi) holds since

\[ \text{span}(\hat{T}^n_k(v^*), T(v^*)) \leq \text{span}(\hat{T}^n_k(v^*), \hat{T}^n_k(0)) + \text{span}(\hat{T}^n_k(0), T(0)) + \text{span}(\text{span}(T(0), T(v^*))) \leq 2 \text{span}(v^*) \]

where we used of [4, Th. 13.1, p. 281].

We next show that Assumption 1(iv) also holds by establishing the following property of \( \hat{G}^n_k \) in the following lemma.

Lemma 6: Suppose that Assumption 5 holds for the average cost MDP. For any \( v_1, v_2 \in \mathcal{V} \), then the following holds:

\[ \text{span}(\hat{G}^n_k(v_1), \hat{G}^n_k(v_2)) \leq \alpha_k \text{span}(v_1, v_2). \]

where \( \alpha_k \) satisfies the following condition. For any \( \delta \in (0, 1 - \alpha) \), we have

\[ \lim_{n \to \infty} \Pr \{ \alpha_k^n > 1 - \delta \} = 0. \]

Further, \( \alpha_k^n \leq 1 \) almost surely.

Proof: It is clear that \( \alpha_k^n \) satisfies

\[ \alpha_k^n \leq 1 - \inf_{(s,a)} \hat{P}^n_k(s_0 | s, a). \]

The fact that \( \alpha_k^n \leq 1 \) almost surely follows in [4, Th. 4.22, p. 81]. The statement will be proved if we show that

\[ \lim_{n \to \infty} \Pr \left\{ \inf_{(s,a)} \hat{P}^n_k(s_0 | s, a) > \delta \right\} = 0. \]

Let \( K \subset S \times A \) be the countable dense subset of \( S \times A \) and define \( Z_0 \) to be the set

\[ Z_0 = \bigcap_{(s,a) \in K} \{ z \in \mathcal{Z} : g(s, a, z) = s_0 \}. \]

Then, \( Z_0 \) is measurable and \( \Pr \{ Z_0 \} = 1 - \alpha \) due to Assumption 5(ii) and (iv). This implies that \( \inf_{(s,a) \in S \times A} \hat{P}^n_k(s_0 | s, a) = \hat{P}^n_k(Z_0) \). Using Hoeffding inequality, we conclude that

\[ \Pr \left\{ \hat{P}^n_k(Z_0) - (1 - \alpha) < \alpha + \delta - 1 \right\} \leq \exp\left(-n(\alpha + \delta - 1)^2\right). \]

This readily yields the (14), which establishes the lemma.

IV. PROOFS OF THE MAIN RESULTS

We now turn our attention to proving Theorem 1. Define the error \( E_k \coloneqq \rho(X_k^n, x^*) \). We focus here on the error process.
(E_k^n)_{k\in\mathbb{N}} and find an upper bound on the probability of large error for a given n \in \mathbb{N}.

A. Proof Technique

Given two random variables X and Y defined on the same probability space, X stochastically dominates Y if \( \mathbb{P}\{X > q\} \geq \mathbb{P}\{Y > q\} \) for all \( q \in \mathbb{R} \) [32]. To prove Theorem 1, we show that the error \( E_k^n \) is stochastically dominated by a scaled version of a Markov chain constructed over the space of natural numbers. We prove that if n is sufficiently large, then the dominating Markov chain has an invariant distribution, which allows us to compute an upper bound on the probability of asymptotic error to be greater than \( \kappa \). Through this approach, we can also compute the rate of convergence as \( n \to \infty \). We now introduce some notation and proof technique in greater details below.

The error evolution can be written as

\[
E_k^n = \rho(X_k^n, x^*) = \rho(T_{k-1}^n(X_{k-1}^n), x^*) \\
\leq \rho(T_{k-1}^n(X_{k-1}^n), T_{k-1}^n(x^*)) + \rho(T_{k-1}^n(x^*), T(x^*)) \\
\leq \hat{\alpha}_k^{n-1}E_{k-1}^n + W_{k-1}^n
\]

where \( \hat{\alpha}_k^{n-1} \) denotes the contraction coefficient of \( \hat{T}_{k-1}^n \) and

\[
W_{k-1}^n := \rho(T_{k-1}^n(x^*), T(x^*)) + \tilde{\gamma}_k^n.
\]

We note here that by Assumption 1(vi), \( W_{k}^n \leq \tilde{w} \) almost surely for any \( n \in \mathbb{N} \) and \( k \in \mathbb{N} \).

Remark 3: Note that \( (\hat{\alpha}_k^n, W_k^n) \) are functions of \( \hat{T}_k^n \). Since \( T_k^n \) is not correlated with \( T_j^n \) for any \( j \neq k \), we conclude that \( (\hat{\alpha}_k^n, W_k^n)_{k\in\mathbb{N}} \) is a sequence of i.i.d. tuple of random variables. However, for every \( k \in \mathbb{N} \), \( \hat{\alpha}_k^n \) is correlated with \( W_k^n \).

We now focus on devising a Markov chain over the space of natural numbers that dominates the error process. Fix \( \kappa > 0 \) and pick \( \epsilon \in (0, (1 - \frac{\alpha}{2})] \), \( \delta \in (0, 1 - \alpha) \) such that \( \frac{\delta}{2} \leq \frac{\epsilon}{2} \). For this choice of \( \epsilon \) and \( \delta \), pick \( \hat{\alpha}_k^{n} := \left[ \frac{\delta}{2} \right] \) and \( p_{e, \delta}^{n} \) such that \( p_{e, \delta}^{n} \leq \mathbb{P}\{\hat{\alpha}_k^{n} \leq 1 - \delta, W_k^n \leq \epsilon\} \). Recall that \( \beta := \left[ \frac{\delta}{2} \right] \).

We now define a Markov chain \( (Y_k^n)_{k\in\mathbb{N}} \) on the set of natural numbers as follows. Let \( Y_k^n = \lfloor E_k^n / \epsilon \rfloor \), and assume that the chain evolves as

\[
Y_k^{n+1} = \begin{cases} 
Y_k^n - 1, & \text{with probability } p_{e, \delta}^{n} \text{ if } Y_k^n = \hat{\alpha}_k^n, \\
Y_k^n + \beta, & \text{with probability } 1 - p_{e, \delta}^{n}.
\end{cases}
\]

(15)

Fig. 1 depicts the communication structure of this Markov chain.

If \( n \) is sufficiently large such that \( p_{e, \delta}^{n} \) satisfies \( p_{e, \delta}^{n} > \beta / (\beta + 1) \), then we show that \( Y_k^n \) admits a unique invariant distribution since it is irreducible and has a negative drift. Further, we prove that at every step of the iteration \( k \), \( \epsilon Y_k^n \) stochastically dominates the error random variable \( E_k^n \), that is, for any real number \( q \in [0, \infty) \) and \( k \in \mathbb{N} \)

\[
\mathbb{P}\{\epsilon Y_k^n > q\} \geq \mathbb{P}\{E_k^n > q\}.
\]

This yields for every \( k \in \mathbb{N} \), we get

\[
\mathbb{P}\{E_k^n > \kappa\} \leq \mathbb{P}\{\epsilon Y_k^n > \kappa\} \leq \mathbb{P}\{Y_k^n > \eta_k^n\}.
\]

Let \( \pi^n \) denote the invariant distribution of the Markov chain \( (Y_k^n)_{k\in\mathbb{N}} \). Then, the above inequality implies

\[
\limsup_{k \to \infty} \mathbb{P}\{E_k^n > \kappa\} \leq \lim \mathbb{P}\{Y_k^n > \eta_k^n\} = 1 - \pi^n(\eta_k^n, \delta).
\]

Further, as \( n \) grows, we show that the invariant distribution of \( \eta_k^n, \pi^n(\eta_k^n, \delta) \), converges to 1, thereby proving the convergence of the error process \( E_k^n \) to 0 in probability as \( k \to \infty \) and \( n \to \infty \).

B. Dominating the Error With the Markov Chain \( (Y_k^n) \)

Recall that the error evolves as \( E_k^{n+1} \leq \hat{\alpha}_k^n E_k^n + W_k^n \), where \( (\hat{\alpha}_k^n, W_k^n)_{k\in\mathbb{N}} \) is an i.i.d. sequence of random variables.

Proposition 3: Let \( n \) be large such that \( \mathbb{P}\{\hat{\alpha}_k^n \leq 1 - \delta, W_k^n \leq \epsilon\} > 1/2 \). Pick \( p_{e, \delta}^{n} \in (0, 1) \) such that \( p_{e, \delta}^{n} \leq \mathbb{P}\{\hat{\alpha}_k^n \leq 1 - \delta, W_k^n \leq \epsilon\} \) and consider the Markov chain \( (Y_k^n)_{k\in\mathbb{N}} \) constructed in (15). If \( \epsilon Y_k^n \geq \hat{\alpha}_k^n \), then at every iteration \( k \), \( \epsilon Y_k^n \) stochastically dominates \( E_k^n \). In other words, for any \( k \in \mathbb{N} \) and any real number \( q \in [0, \infty) \)

\[
\mathbb{P}\{\epsilon Y_k^n > q\} \geq \mathbb{P}\{E_k^n > q\}.
\]

Proof: See Appendix A.

In the light of the proposition above, we need to identify a lower bound on the joint distribution \( \mathbb{P}\{\hat{\alpha}_k^n \leq 1 - \delta, W_k^n \leq \epsilon\} \) that can be used to determine \( p_{e, \delta}^{n} \). We obtain a lower bound using the Fréchet–Hoeffding theorem [43, Th. 3.1.1]: For any \( a_1, a_2 \in [0, \infty) \), we have

\[
\mathbb{P}\{\hat{\alpha}_k^n \leq a_1, W_k^n \leq a_2\} \geq \mathbb{P}\{\hat{\alpha}_k^n \leq a_1\} + \mathbb{P}\{W_k^n \leq a_2\} - 1.
\]

(16)

Suppose that we know the upper bounds \( \varphi_1(n, \delta) \) and \( \varphi_2(n, \epsilon) \) on the probability \( \mathbb{P}\{\hat{\alpha}_k^n > 1 - \delta\} \) and \( \mathbb{P}\{W_k^n > \epsilon\} \), respectively

\[
\mathbb{P}\{\hat{\alpha}_k^n > 1 - \delta\} \leq \varphi_1(n, \delta), \quad \mathbb{P}\{W_k^n > \epsilon\} \leq \varphi_2(n, \epsilon).
\]
Typically, such bounds can be obtained using a concentration of measures result, such as the Hoeffding inequality or the empirical process theory [34, 35, 44]. We let $p^{n}_{\epsilon, \delta}$ be defined as

$$
\hat{p}^{n}_{\epsilon, \delta} = 1 - \varphi_1(n, \delta) - \varphi_2(n, \epsilon).
$$

This, together with (16), implies

$$
\mathbb{P}\left\{ \hat{\alpha}^{n}_{k} \leq 1 - \delta, W^{n}_{k} \leq \epsilon \right\} \\
\geq \mathbb{P}\left\{ \hat{\alpha}^{n}_{k} \leq 1 - \delta \right\} + \mathbb{P}\left\{ W^{n}_{k} \leq \epsilon \right\} - 1 \\
\geq 1 - \varphi_1(n, \delta) - \varphi_2(n, \epsilon) = \hat{p}^{n}_{\epsilon, \delta}.
$$

We make the following observation.

**Lemma 7:** If Assumption 1 holds, then for any $\epsilon > 0$ and $\delta \in (0, 1 - \alpha)$, $\lim_{n \to \infty} p^{n}_{\epsilon, \delta} = 1$.

**Proof:** The proof essentially follows from Assumption 1. Assumption 1(iv) implies that $\varphi_{1}(n, \delta) \to 0$ as $n \to \infty$ for any $\delta \in (0, 1 - \alpha)$. Assumption 1(v) implies that $\varphi_2(n, \epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$. The proof of the lemma is complete.

According to the Lemma above, by picking $n$ sufficiently large, $p^{n}_{\epsilon, \delta}$ can be made as close to 1 as possible. As we show next, for $p^{n}_{\epsilon, \delta}$ sufficiently close to 1, the Markov chain $(Y^{n}_{k})_{k \in \mathbb{N}}$ admits an invariant distribution.

**Proposition 4:** If $p^{n}_{\epsilon, \delta} > \beta/(\beta + 1)$, then the Markov chain $(Y^{n}_{k})_{k \in \mathbb{N}}$ admits a unique invariant distribution $\pi^{n}$. If $p^{n}_{\epsilon, \delta} > 2\beta/(2\beta + 1)$, then

$$
\pi^{n}(\eta^{n}_{\epsilon, \delta}) \geq \frac{2(p^{n}_{\epsilon, \delta})^{\beta} - 1}{(p^{n}_{\epsilon, \delta})^{\beta}}.
$$

**Proof:** We first note that $(Y^{n}_{k})_{k \in \mathbb{N}}$ is an irreducible Markov chain, in which all integers greater than or equal to $\eta^{n}_{\epsilon, \delta}$ are accessible. Consider the Lyapunov function as $V(y) = y$ for all $y \in \mathbb{N}$, $y \geq \eta^{n}_{\epsilon, \delta}$. Then, $E[V(Y^{n}_{k+1})|Y^{n}_{k}] = y - V(y) < 0$ for all $y \geq \eta^{n}_{\epsilon, \delta}$. The uniqueness of the invariant distribution follows from [25, Th. 7.5.3, p.153]. See Appendix B for a derivation on the lower bound on $\pi^{n}(\eta^{n}_{\epsilon, \delta})$.

**C. Proof of Theorem 1**

We now prove Theorem 1 using Propositions 3 and 4 as follows. For $n$ sufficiently large (so that $p^{n}_{\epsilon, \delta} > 2\beta/(2\beta + 1)$), we can use Proposition 3 to conclude that for every $k \in \mathbb{N}$

$$
\mathbb{P}\left\{ E^{n}_{k} \geq \kappa \right\} \leq \mathbb{P}\left\{ \epsilon Y^{n}_{k} \geq \kappa \right\} \leq \mathbb{P}\left\{ \epsilon Y^{n}_{k} > \epsilon n_{\epsilon, \delta} \right\}.
$$

From Proposition 4, we conclude that

$$
\lim_{k \to \infty} \mathbb{P}\left\{ Y^{n}_{k} > \eta^{n}_{\epsilon, \delta} \right\} = 1 - \pi^{n}(\eta^{n}_{\epsilon, \delta}) \leq \frac{1 - (p^{n}_{\epsilon, \delta})^{\beta}}{(p^{n}_{\epsilon, \delta})^{\beta}}.
$$

Consequently, we have

$$
\limsup_{k \to \infty} \mathbb{P}\left\{ E^{n}_{k} \geq \kappa \right\} \leq \frac{1 - (p^{n}_{\epsilon, \delta})^{\beta}}{(p^{n}_{\epsilon, \delta})^{\beta}}.
$$

The proof of the theorem is complete.

**D. Proof of Theorem 2**

It is easy to observe that Assumption 2 immediately implies Assumption 1(iii) and (iv). We only need to show that Assumption 1(v) and (vi) holds.

Let us consider the case of $L = 2$ for simplicity, and the general case can then be easily deduced. For any $x^{*} \in \mathcal{X}$, define $x^{*}_{1} := x^{*}$, $x^{*}_{2} := T_{1}(x^{*}_{1})$, we have

$$
\rho(T^{n}_{k}(x^{*}), T(x^{*})) = \rho_{3}\left( \hat{H}^{n}_{k,2}\left( \hat{H}^{n}_{k,1}(x^{*}_{1}) \right), T_{2}\left( T_{1}(x^{*}_{1}) \right) \right)
$$

$$
\leq \rho_{3}\left( \hat{H}^{n}_{k,2}\left( \hat{H}^{n}_{k,1}(x^{*}_{1}) \right), \hat{H}^{n}_{k,2}\left( T_{1}(x^{*}_{1}) \right) \right)
$$

$$
\leq \rho_{3}\left( \hat{H}^{n}_{k,2}\left( T_{1}(x^{*}_{1}) \right), T_{2}\left( T_{1}(x^{*}_{1}) \right) \right)
$$

$$
\leq \hat{\alpha}^{n}_{k,\rho_{2}}\left( \hat{H}^{n}_{k,1}(x^{*}_{1}), T_{1}(x^{*}_{1}) \right) + \zeta^{n}_{k}
$$

$$
+ \rho_{3}\left( \hat{H}^{n}_{k,2}\left( x^{*}_{2} \right), T_{2}\left( x^{*}_{2} \right) \right).
$$

We now extend the same argument to the general case. Using triangle inequality and Assumption 2(ii), we get

$$
\rho(T^{n}_{k}(x^{*}), T(x^{*}))
$$

$$
= \rho L_{n+1}\left( \hat{H}^{n}_{k,L} \circ \cdots \circ \hat{H}^{n}_{k,1}(x^{*}), T_{L} \circ \cdots \circ T_{1}(x^{*}) \right)
$$

$$
\leq \sum_{l=1}^{L} \left( \prod_{m=l+1}^{L} \hat{\alpha}^{n}_{l,m} \right) \rho_{l+1}\left( \hat{H}^{n}_{k,l}(x^{*}_{l}), T_{l}(x^{*}_{l}) \right) + \zeta^{n}_{k}
$$

$$
\leq \sum_{l=1}^{L} \rho_{l+1}\left( \hat{H}^{n}_{k,l}(x^{*}_{l}), T_{l}(x^{*}_{l}) \right) + \zeta^{n}_{k}
$$

where we take $\prod_{m=L+1}^{L} \hat{\alpha}^{n}_{l,m} = 1$. The last inequality immediately yields for any $\epsilon > 0$

$$
\mathbb{P}\left\{ \rho\left( T^{n}_{k}(x^{*}), T(x^{*}) \right) \geq \epsilon \right\}
$$

$$
\leq \sum_{l=1}^{L-1} \mathbb{P}\left\{ \rho_{l+1}\left( \hat{H}^{n}_{k,l}(x^{*}_{l}), T_{l}(x^{*}_{l}) \right) \geq \frac{\epsilon}{L} \right\}
$$

$$
+ \mathbb{P}\left\{ \rho_{L+1}\left( \hat{H}^{n}_{k,L}(x^{*}_{L}), T_{L}(x^{*}_{L}) \right) + \zeta^{n}_{k} \geq \frac{\epsilon}{L} \right\}.
$$

This implies that Assumption 1(v) holds. To see that Assumption 1(vi) holds, we only need to take $\bar{w} = \sum_{l=1}^{L} \bar{w}_{l}$. The proof is hence complete.

**E. Proof of Theorem 3**

This proof is inspired from [14]. For this case, a dominating Markov chain is defined as

$$
Y^{n}_{k+1} = \begin{cases} 
\eta, & \text{with probability } p \text{ if } Y^{n}_{k} = \eta \\
Y^{n}_{k} - 1, & \text{with probability } p \text{ if } Y^{n}_{k} \geq \eta + 1 \\
d_{i}, & \text{with probability } 1 - p.
\end{cases}
$$

It can be shown that for any $k \in \mathbb{N}$, we have

$$
\mathbb{P}\left\{ \rho(X^{n}_{k+1}, x^{*}) > \kappa \right\} \leq \mathbb{P}\left\{ \epsilon Y^{n}_{k} > \kappa \right\}.
$$

In [14, Lemma 5.1, Prop. 5.3], shows that for all $k \geq K_{\kappa, \xi}$ and $n \geq N_{\kappa, \xi}, \mathbb{P}\left\{ \epsilon Y^{n}_{k} > \kappa \right\} < \xi$. The proof is hence complete.

**V. CONCLUSION**

In this article, we have introduced a novel approach to deduce the stability of iterated random operators. The approach
presented here provides an alternative to the Foster–Lyapunov method, the martingale method, and the ODE method for use in infinite dimensional spaces. We refer to this approach as the probabilistic contraction analysis. We hope that this article can accelerate the development of new recursive stochastic algorithms along with their convergence guarantees.

In our future work, we will also consider other characterizations of probabilistic fixed points (e.g., a mean square version), and explore application to other reinforcement learning problems.

**APPENDIX A**

**PROOF OF PROPOSITION 3**

Let $E_k$ denote the event $E^n_k \leq \epsilon n^p_{\delta, \delta}$ and $F_k$ denote the event \{\(W^n_k \leq \epsilon, \alpha^n_k \leq 1 - \delta\)\}. We have

$$E^n_{k+1} \mid E_k, F_k \leq \epsilon n^p_{\delta, \delta} \quad \text{almost surely}$$

$$E^n_{k+1} - E^n_k + \epsilon E^n_k, F_k \leq - \epsilon \quad \text{almost surely}$$

$$E^n_{k+1} - E^n_k \leq \bar{w} \quad \text{almost surely}.$$  \hspace{1cm} (20)

We now construct another Markov chain \(Z^n_k\) on \((\Omega, \mathcal{F}, \mathbb{P})\) as follows: Let \(Z^n_1 = [E^n_1/\epsilon]\), and the chain evolves as

$$Z^n_{k+1} = \begin{cases} \eta^n_{\epsilon, \delta}, & \text{if } Z^n_k = \eta^n_{\epsilon, \delta}, \\ Z^n_{k-1}, & \text{if } W^n_k \leq \epsilon, \alpha^n_k \leq 1 - \delta \\ Z^n_k + \epsilon, & \text{if } W^n_k \geq \epsilon \text{ or } \alpha^n_k \geq 1 - \delta. \end{cases}$$  \hspace{1cm} (21)

By construction and Lemma 8, we have \(E^n_k \leq \epsilon Z^n_k \) almost surely, which can be shown via induction. The statement is obviously true for \(k = 1\). Assume that the statement is true for some \(k\). Then, the following two cases can occur.

1. If \(E^n_k \leq \epsilon n^p_{\delta, \delta}\), which implies that \(E^n_{k+1}\) will be less than \(\epsilon n^p_{\delta, \delta}\) in the event \(F_k\) by Lemma 8. By construction, \(Z^n_{k+1} \geq \eta^n_{\epsilon, \delta}\), and thus, \(E^n_{k+1} \leq \epsilon Z^n_{k+1}\).

2. On the other hand, if \(E^n_k \geq \epsilon n^p_{\delta, \delta}\), then by Lemma 8, \(E^n_{k+1}\) is less than or equal to \(E^n_k - \epsilon\) under the event \(F_k\). Since \(Z^n_k \geq E^n_k/\epsilon\), we have \(Z^n_k \geq \eta^n_{\epsilon, \delta} + 1\) and \(Z^n_{k+1} = Z^n_k - 1\) under the event \(F_k\).

3. It is clear from Lemma 8 that \(E^n_{k+1} \leq E^n_k + \bar{w} \leq Z^n_k + \bar{w}\) in the event \(F_k\), which implies \(E^n_{k+1} \leq Z^n_{k+1}\) under the event \(F_k\).

As a result of the assertion above, \(Z^n_k\) stochastically dominates \(E^n_k\). We now have the following lemma.

**Lemma 9:** Let \(T_Y\) and \(T_Z\) denote the transition kernel of \((Y^n_k)\) and \((Z^n_k)\), respectively. Then, for any \(q \in \mathbb{N}, q \geq n^p_{\delta}\), \(T_Y(q, \cdot)\) stochastically dominates \(T_Z(q, \cdot)\). Consequently, \(Y^n_k\) stochastically dominates \(Z^n_k\).

**Proof:** Since \(\pi^n_{\epsilon, \delta} \leq \mathbb{P}\{W^n_k \leq \epsilon, \alpha^n_k \leq 1 - \delta\}, \ T_Y(q, \cdot)\) stochastically dominates \(T_Z(q, \cdot)\) for all \(q \in \mathbb{N}, q \geq n^p_{\delta}\).

Consequently, we proved that \(Y^n_{k+1}\) stochastically dominates \(\epsilon Z^n_{k+1}\), which in turn stochastically dominates \(E^n_{k+1}\). The induction step is complete and we arrive at the result.

**APPENDIX B**

**PROOF OF PROPOSITION 4**

Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a standard probability space. On this probability space, we define two different Markov chains: \((P_k)\) and \((Q_k)\). Pick \(p \in (0, 1]\). Markov chains \(P_k\) and \(Q_k, k \in \mathbb{N}\), evolves as

$$P_{k+1} = \begin{cases} 0, & \text{with probability } p \text{ if } P_k = 0 \\ P_k - 1, & \text{with probability } p \text{ if } P_k \geq 1 \\ P_k + \beta, & \text{with probability } 1 - p \end{cases}$$

$$Q_{k+1} = \begin{cases} Q_k, & \text{with probability } p \text{ if } Q_k = 0 \\ Q_k - 1, & \text{with probability } p \text{ if } Q_k \geq 1 \\ \beta ([Q_k/\alpha] + 1), & \text{with probability } 1 - p \end{cases}$$  \hspace{1cm} (22)

Fig. 2 depicts the communication structure of the Markov chains \(P_k\) and \(Q_k\). Both Markov chains thus constructed are supported over the space of nonnegative integers, are irreducible, and all nonnegative integers are accessible. It is also clear that if \(p > 2\beta/(2\beta + 1)\), then by [25, Th. 7.5.3], both Markov chains admit a unique invariant distribution. We next have the following claim.

**Claim 1:** If \(P_1 = Q_1\), then \(Q_k\) stochastically dominates \(P_k\) for every \(k \in \mathbb{N}\).

**Proof:** We show that along every sample path, \(Q_k(\bar{\omega}) \geq P_k(\bar{\omega})\). Suppose that \(P_k = Q_k = q\) for some \(q \in \mathbb{N}\). Then, for any \(\omega \in \Omega\), either \(P_{k+1} = Q_{k+1} = \max(0, q - 1)\), or \(Q_{k+1} = \beta([Q_k/\beta] + 1) \geq \beta(Q_k/\beta + 1) = Q_k + \beta = P_{k+1}\). Thus, \(Q_k+1 \geq P_k+1\). The result then holds in [32, Th. 1.A.6].

As a result of the claim above, if both Markov chains \((P_k)\) and \((Q_k)\) admit invariant distributions \(\pi^P\) and \(\pi^Q\), respectively, then \(\pi^P(0) \geq \pi^Q(0)\). We next identify certain sufficient conditions under which the two Markov chains admit invariant distributions.

**Theorem 10:** The following holds true.

1. If \(p > \beta(\beta + 1)\), then the Markov chain \((P_k)\) has an invariant distribution \((\pi^P(n))_{n=0}^\infty\).
2) If \( p > 2\beta/(2\beta + 1) \), then \( Q_k \) has an invariant distribution \( (\pi^Q(n))_{n=0}^\infty \).

**Proof:** We show that both Markov chains are weak Feller chains since they are defined over a countable state space.

Let \( \gamma_P = (\beta + 1)p - \beta > 0 \). Consider \( V_P(i) = (i + 1)/\gamma_P \) and compact set \( C_P = \{0\} \). Then, given \( P_k = i \geq 1 \), we have

\[
\mathbb{E}[V_P(P_{k+1})|P_k] - V_P(P_k) = \frac{pP_k + (1-p)(P_k + \beta + 1) - (P_k + 1)}{\gamma_P} = -1.
\]

For \( P_k = 0 \), we have

\[
\mathbb{E}[V_P(P_{k+1})|P_k = 0] - V_P(0) = \frac{p + (1-p)(\beta + 1) - 1}{\gamma_P} = -1 + \frac{p}{\gamma_P}.
\]

Thus, by [23, Th. 12.3.4], an invariant probability distribution \( \pi^P \) for the Markov chain \( (P_k) \) exists.

Let \( \gamma_Q = (\beta + 1)p - 2\beta > 0 \). Consider \( V_Q(i) = (i + 1)/\gamma_Q \) and compact set \( C_Q = \{0\} \). Then, given \( Q_k = i \geq 1 \), we have

\[
\mathbb{E}[V_Q(Q_{k+1})|Q_k] - V_Q(Q_k) = \frac{pQ_k + (1-p)(Q_k + \beta + 1) - (Q_k + 1)}{\gamma_Q} \leq \frac{pQ_k + (1-p)(Q_k + 2\beta + 1) - (Q_k + 1)}{\gamma_Q} = -1.
\]

For \( Q_k = 0 \), we have

\[
\mathbb{E}[V_Q(Q_{k+1})|Q_k = 0] - V_Q(0) = \frac{p + (1-p)(\beta + 1) - 1}{\gamma_Q} \leq \frac{p + (1-p)(2\beta + 1) - 1}{\gamma_Q} = -1 + \frac{p}{\gamma_Q}.
\]

Again, we invoke [23, Th. 12.3.4] to conclude the existence of an invariant probability distribution \( \pi^Q \) for the Markov chain \( (Q_k) \).

We characterize the invariant distribution \( \pi^Q \) in the following claim.

**Claim 2:** Assume that \( p > 2\beta/(2\beta + 1) \). Then, the invariant distribution of the Markov chain \( (Q_k) \) satisfies \( \pi^Q(0) = 2p^{\beta-1}/p^\beta \) and for all \( i \in \{1, \ldots, \beta\} \) and \( k \geq 0 \)

\[
\pi^Q(k + \beta) = \pi^Q(0)\left(\frac{1-p}{p^{\beta+1}}\right)^k.
\]

**Proof:** See Section B-A below.

**Corollary 11:** Assume that \( p > 2\beta/(2\beta + 1) \). Then, an invariant distribution \( \pi^P \) exists and \( \pi^P(0) = \pi^Q(0) = 2p^{\beta-1}/p^\beta \).

**Proof:** Note that \( 2\beta/(2\beta + 1) > \beta/(\beta + 1) \). The proof then follows immediately from Theorem 10 and Claim 2.

It is now easy to observe that if \( p = p^n_{\alpha, \delta} \), then the evolution of \( Y^n_k \) is the same as that of \( P^n_{\alpha, \delta} \). Thus, their invariant distribution is a “shifted” version of the other, that is, \( \pi^n(i) = \pi^Q(i - n_{\alpha, \delta}) \) for all \( i \geq n_{\alpha, \delta} \). Thus, \( \pi^n(n_{\alpha, \delta}) \geq 2p^{\beta-1}/(p^{\beta+1})^n \). This completes the proof of Proposition 4.

**A. Proof of Claim 2**

The invariant distribution \( \pi^Q \) exists by Theorem 10 above. It must satisfy

\[
\pi^Q(0) = p\pi^Q(0) + p\pi^Q(1) \Rightarrow \pi^Q(1) = \left(\frac{1-p}{p}\right)\pi^Q(0)
\]

\[
\pi^Q(i) = p\pi^Q(i+1) \quad \text{for all } i \in \{1, \ldots, \beta - 1\}
\]

which implies that \( \pi^Q(i) = \left(\frac{1-p}{p}\right)^i\pi^Q(0) \) for all \( i \in \{1, \ldots, \beta\} \). Thus, the statement holds for \( n = 0 \) and all \( i \in \{0, 1, \ldots, \beta\} \). For \( n = 1 \), we have

\[
\pi^Q(\beta) = (1-p)\pi^Q(0) + p\pi^Q(\beta + 1)
\]

\[
\pi^Q(\beta + i) = p\pi^Q(\beta + i + 1) \quad \text{for all } i \in \{1, \ldots, \beta - 1\}
\]
which implies
\[
\pi^Q(\beta + 1) = \frac{1}{p} \left( \pi^Q(\beta) - (1 - p)\pi^Q(0) \right) = \frac{(1 - p)}{p^\beta + 1} \pi^Q(0)(1 - p^\beta)
\]
\[
\pi^Q(\beta + i) = \frac{(1 - p)}{p^\beta + i} \pi^Q(0)(1 - p^\beta) \text{ for all } i \in \{1, \ldots, \beta\}.
\]
Consequently, the statement holds for \(n = 1\) as well. We now prove the result for arbitrary \(n \geq 2\). Suppose that the result holds for all \(m \leq n - 1\) and \(i \in \{1, \ldots, \beta\} \). Then, we have
\[
\pi^Q(m\beta + 1) + \ldots + \pi^Q(m\beta + \beta) = \frac{(1 - p)}{p^{m+1} \beta} \pi^Q(0)(1 - p^\beta)m (1 + \ldots + p^{\beta - 1}) = \frac{1}{p^{(m+1)\beta}} \pi^Q(0)(1 - p^\beta)m + 1.
\]
Next, we have
\[
\pi^Q(m\beta + 1) = (1 - p) \left( \pi^Q((n - 2)\beta + 1) + \ldots + \pi^Q((n - 2)\beta + \beta) \right) + p\pi^Q(m\beta + 1) = \frac{(1 - p)}{p^{(n - 1)\beta}} \pi^Q(0)(1 - p^\beta)^{(n - 1)\beta} + p\pi^Q(m\beta + 1)
\]
\[
\pi^Q(m\beta + i) = p\pi^Q(m\beta + i + 1) \text{ for all } i \in \{1, \ldots, \beta - 1\}.
\]
Using similar approach as for the previous cases, we have
\[
\pi^Q(m\beta + 1) = \frac{1}{p} \left[ \pi^Q((n - 1)\beta + 1) + \ldots + \pi^Q((n - 1)\beta + \beta) \right] = \frac{1}{p} \sum_{k=1}^{\beta} \pi^Q(0)(1 - p^\beta)^{(n - 1)\beta}
\]
\[
\pi^Q(m\beta + i + 1) = \frac{1}{p} \pi^Q(m\beta + i) = \frac{1}{p} \pi^Q(0)(1 - p^\beta)^{m\beta}
\]
which holds for all \(i \in \{1, \ldots, \beta - 1\}\). Thus, the statement is true for \(n\). By the principle of mathematical induction, the statement is established.

We can now compute \(\pi^Q(0)\) by noting that
\[
\pi^Q(0) + \pi^Q(0) \sum_{m=0}^{\infty} \frac{1}{p^{(m+1)\beta}}(1 - p^\beta)m + 1 = 1
\]
where we used (22). The above expression yields
\[
\pi^Q(0) \left( 1 + \frac{1 - p^\beta}{1 - p\beta} \right) = \pi^Q(0) \frac{p^\beta}{2p^\beta - 1} = 1
\]
which implies \(\pi^Q(0) = \frac{2p^\beta - 1}{p^\beta}\). The proof of the claim is complete.

Note that for \(\pi^Q(0)\) to be nonnegative, we need \(p^\beta \geq 0.5\). We show in the following remark that this is indeed true as long as \(p \geq 2\beta/(2\beta + 1)\).

**Remark 4:** We show that if \(p > 2\beta/(2\beta + 1)\), then \(p^\beta > 0.6\), which further implies \(\pi^Q(0) > 0\). To establish this result, we prove that the map \(r \mapsto (1 + \frac{1}{r})^r\) is monotonically increasing in \(r \in (1, \infty)\). Indeed
\[
\frac{d}{dr} \left( 1 + \frac{1}{r} \right)^r = \left( 1 + \frac{1}{r} \right)^r \left( \ln 1 + \frac{1}{r} - \frac{1}{r+1} \right).
\]
Now, since for \(t \in (1, 1 + \frac{1}{r})\), \(t < \frac{r+1}{r} \) or \(t > \frac{r+1}{r+1}\), we have
\[
\ln \left( 1 + \frac{1}{r} \right) - \frac{1}{r+1} = \int_1^{(1 + \frac{1}{r})^r} \left( \frac{1}{t} - \frac{1}{t+1} \right) dt > 0.
\]
Thus, \(\frac{d}{dr} \left( 1 + \frac{1}{r} \right)^r > 0\), which implies \((1 + \frac{1}{r})^r\) is monotonically increasing in \(r\) in the domain \([1, \infty)\). Thus, if \(p > 2\beta/(2\beta + 1)\), we have
\[
p^\beta \geq \frac{1}{(1 + \frac{1}{2\beta})^\beta} = \frac{1}{\sqrt{1 + \frac{1}{2\beta}}} \approx 0.606.
\]

**REFERENCES**

[1] C. D. Aliprantis and K. Border, *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Berlin, Germany: Springer, 2006.

[2] D. P. Bertsekas, *Dynamic PROGRAMMING and Optimal Control, Vol. I and II*, vol. 1, Belmont, MA, USA: Athena Scientific, 1995.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.

[4] A. L. Almudevar, *Approximate Iterative Algorithms*. Boca Raton, FL, USA: CRC Press, 2014.

[5] C. Kollman, K. Baggerly, D. Cox, and R. Picard, “Adaptive importance sampling on discrete Markov chains,” *Ann. Appl. Probability*, pp. 391–412, 1999.

[6] P. Y. Desai, *Adaptive Monte Carlo Methods for Solving Eigenvalue Problems*. Stanford, CA, USA: Stanford Univ., 2001.

[7] P. Y. Desai and P. W. Glynn, “A Markov chain perspective on adaptive Monte Carlo algorithms,” in *Proc. Winter Simul. Conf. (Cat. No 01CH37304)*, 2001, pp. 379–384.

[8] T. I. Ahamed, V. S. Borkar, and S. Juneja, “Adaptive importance sampling technique for Markov chains using stochastic approximation,” *Operations Res.*, vol. 54, no. 3, pp. 489–504, 2006.

[9] H. Jeng, A. N. Bishop, G. Deligiannidis, and A. Doucet, “Controlled sequential Monte Carlo,” *Ann. Statist.*, vol. 48, no. 5, pp. 2904–2929, 2020.

[10] A. Boustati, O. D. Akyildiz, T. Damoulas, and A. Johansen, “Generalised Bayesian filtering via sequential Monte Carlo,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 418–429, 2020.

[11] J. N. Tsitsiklis and B. Van Roy, “Analysis of temporal-difference learning with function approximation,” in *Adv. Neural Inf. Process. Syst.*, 1997, pp. 1075–1081.

[12] D. P. Bertsekas and H. Yu, “Q-learning and enhanced policy iteration in discounted dynamic programming,” *Math. Operations Res.*, vol. 37, no. 1, pp. 66–94, 2012.

[13] P. Yu, W. B. Haskell, and H. Xu, “Approximate value iteration for risk-aware Markov decision processes,” *IEEE Trans. Autom. Control*, vol. 63, no. 9, pp. 3135–3142, Sep. 2018.

[14] W. B. Haskell, R. Jain, and D. Kalathil, “Empirical dynamic programming,” *Math. Operations Res.*, vol. 41, no. 2, pp. 402–429, 2016.

[15] L. E. Dubins and D. A. Freedman, “Invariant probabilities for certain Markov processes,” *Ann. Math. Statist.*, vol. 37, no. 4, pp. 837–848, 1966.

[16] M. F. Barnsley and S. Demko, “Iterated function systems and the global construction of fractals,” *Proc. Roy. Soc. London. A. Math. Phys. Sci.*, vol. 399, no. 1817, pp. 243–275, 1985.
[17] M. F. Barnsley, J. H. Elton, and D. P. Hardin, “Recurrent iterated function systems,” Constructive Approximation, vol. 5, no. 1, pp. 3–31, 1989.
[18] P. Diaconis and D. Freedman, “Iterated random functions,” SIAM Rev., vol. 41, no. 1, pp. 45–76, 1999.
[19] O. Stenflo, “A survey of average contractive iterated function systems,” J. Difference Equ. Appl., vol. 18, no. 8, pp. 1355–1380, 2012.
[20] M. Duflo, Random Iterative Models, vol. 34. Berlin, Germany: Springer Sci. & Business Media, 2013.
[21] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2014.
[22] A. Gupta, R. J. Karandikar, and T. Parthasarathy, “An empirical algorithm for relative value iteration for average-cost MDPs,” in Proc. IEEE 54th Conf. Decis. Control, pp. 5079–5084, 2015.
[23] S. P. Meyn and R. L. Tweedie, Markov Chains and Stochastic Stability. Cambridge, UK: Cambridge Univ. Press, 2009.
[24] A.-A. Borovkov and V. Yurinski, Ergodicity and Stability of Stochastic Processes. Hoboken, NJ, USA: John Wiley & Sons, Ltd., 1998.
[25] R. Douc, E. Moulines, P. Priouret, and P. Soulier, Markov Chains. Berlin, Germany: Springer, 2018.
[26] V. S. Borkar and S. P. Meyn, “The ODE method for convergence of stochastic approximation and reinforcement learning,” SIAM J. Control Optim., vol. 38, no. 2, pp. 447–469, 2000.
[27] V. S. Borkar, Stochastic Approximation: A Dynamical Systems Viewpoint, vol. 48. Berlin, Germany: Springer, 2009.
[28] J. Huang, I. Kontoyiannis, and S. P. Meyn, “The ODE Method and Spectral Theory of Markov Operators,” in Stochastic Theory and Control. Berlin, Germany: Springer, 2002, pp. 205–221.
[29] H. Kushner and G. G. Yin, Stochastic Approximation and Recursive Algorithms and Applications, vol. 35. Berlin, Germany: Springer Sci. & Business Media, 2003.
[30] A. A. Kulkarni and V. S. Borkar, “Finite dimensional approximation and Newton-based algorithm for stochastic approximation in Hilbert space,” Automatica, vol. 45, no. 12, pp. 2815–2822, 2009.
[31] A. Dieuleveut and F. Bach, “Nonparametric stochastic approximation with large step-sizes,” Ann. Statist., vol. 44, no. 4, pp. 1363–1399, 2016.
[32] M. Ledoux, Concentration of Measure Phenomenon, vol. 89. Providence, Rhode Island, USA: American Mathematical Soc., 2001.
[33] D. Pollard, Convergence of Stochastic Processes. Springer Series in Statistics, Berlin, Germany: Springer, 1991.
[34] A. W. Van Der Vaart and J. A. Wellner, Weak Convergence and Empirical Processes With Applications to Statistics. Berlin, Germany: Springer-Verlag, 1996.
[35] K. Hinderer, “Lipschitz continuity of value functions in markovian decision processes,” Math. Methods Operations Res., vol. 62, no. 1, pp. 3–22, 2005.
[36] S. Shao, F. Harirchi, D. Dave, and A. Gupta, “Preemptive scheduling of EV charging for providing demand response services,” Sustain. Energy, Grids Netw., vol. 33, 2023, Art. no. 100986.
[37] A. Defazio, F. Bach, and S. Lacoste-Julien, “SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives,” in Adv. Neural Inf. Process. Syst., 2014, pp. 1646–1654.
[38] R. Harikande, M. O. Ahmed, A. Virani, M. Schmidt, J. Konečný, and S. Sallinen, “Stopwasting my gradients: Practical SVRG,” in Adv. Neural Inf. Process. Syst., 2015, pp. 2251–2259.
[39] A. Sidford, M. Wang, X. Wu, L. Yang, and Y. Ye, “Near-optimal time and sample complexities for solving Markov decision processes with a generative model,” in Adv. Neural Inf. Process. Syst., 2018, pp. 5186–5196.
[40] J. M. Wainwright, “Variance-reduced Q-learning is minimax optimal,” in Proc. 35th Uncertainty-Stat. Model. Conf., 2019, pp. 956–964, 1968.
[41] A. Gupta and W. B. Haskell, “Convergence of stochastic algorithms using Wasserstein divergence,” SIAM J. Math. Data Sci., vol. 3, pp. 1141–1167, 2021.
[42] S. T. Rachev and L. Rüschendorf, Mass Transportation Problems: Volume I: Theory, vol. 1. Berlin, Germany: Springer Sci. Business Media, 1998.
[43] L. Györfi, Principles of Nonparametric Learning, vol. 434. Berlin, Germany: Springer, 2002.
[44] A. Hain, “Universal function approximation by deep neural nets with bounded width and RELU activations,” Math., vol. 7, no. 10, 2019, Art. no. 992.
Rahul Jain (Senior Member, IEEE) received the B.Tech. degree in EE from the Indian Institute of Technology Kanpur, Kanpur, India, in 1997, the M.S. degree in ECE from Rice University, Houston, TX, USA, in 1999, and the M.A. degree in statistics and the Ph.D. degree in EECS from the University of California, Berkeley, Berkeley, CA, USA, in 2002 and 2004, respectively. Since 2008, he has been a Professor of electrical and computer engineering, and computer science with the University of Southern California, and a Visiting Research Scientist with Google since 2022. His research interests have current focus on AI for stochastic systems, online, statistical and reinforcement learning, queueing systems, power system economics, network economics, and game theory.

Dr. Jain was the recipient of Texas Instruments Fellowship from Rice University, James H. Zumberge Faculty Research and Innovation Award in 2009, IBM Faculty Award, NSF CAREER Award in 2010, and ONR Young Investigator Award in 2012. He has been a US Fulbright Specialist Scholar from 2017 to 2020, and was a Fulbright Visiting Professor at the Indian Institute of Science Bangalore, Bengaluru, India, in 2018.

Peter Glynn received the Ph.D. degree in operations research from Stanford University, Stanford, CA, USA, in 1982. He is currently the Thomas Ford Professor with the Department of Management Science and Engineering (MS&E), Stanford University, and also holds a Courtesy Appointment with the Department of Electrical Engineering. He then joined the Faculty with the University of Wisconsin at Madison, Madison, WI, USA, where he held a joint appointment between Industrial Engineering Department and Mathematics Research Center, and courtesy appointments in computer science and mathematics. In 1987, he returned to Stanford, where he joined the Department of Operations Research. From 1999 to 2005, he was the Deputy Chair of the Department of Management Science and Engineering. From 2006 to 2010, he was the Director of Stanford’s Institute for Computational and Mathematical Engineering. In 2012, he was elected to the National Academy of Engineering. His research interests include simulation, computational probability, queueing theory, statistical inference for stochastic processes, and stochastic modeling.

Dr. Glynn was the corecipient the Outstanding Publication Awards from the INFORMS Simulation Society in 1993, 2008, and 2018, respectively. Best (Biannual) Publication Award from the INFORMS Applied Probability Society in 2009, and John von Neumann Theory Prize from INFORMS in 2010. He delivered INFORMS Philip McCord Morse Lecture in 2020. From 2011 to 2015, he was the Chair of MS&E. He is a Fellow of INFORMS and a Fellow of the Institute of Mathematical Statistics, and was an IMS Medallion Lecturer in 1995 and INFORMS Markov Lecturer in 2014. He was Founding Editor-in-Chief of Stochastic Systems. From 2016 to 2018, he was the Editor-in-Chief of Journal of Applied Probability and Advances in Applied Probability.