Some Limit Properties of Markov Chains Induced by Stochastic Recursive Algorithms∗
Abhishek Gupta, Gaurav Tendolkar, Hao Chen, Jianzong Pi†

Abstract. Recursive stochastic algorithms have gained significant attention in the recent past due to data driven applications. Examples include stochastic gradient descent for solving large-scale optimization problems and empirical dynamic programming algorithms for solving Markov decision problems. These recursive stochastic algorithms approximates certain contraction operators and can be viewed within the framework of iterated random maps. Accordingly, we consider iterated random maps over a Polish space that simulates a contraction operator over that Polish space. Assume that the iterated maps are indexed by $n$ such that as $n \to \infty$, each realization of the random map converges (in some sense) to the contraction map it is simulating. We show that starting from the same initial condition, the distribution of the random sequence generated by the iterated random maps converge weakly to the trajectory generated by the contraction operator. We further show that under certain conditions, the time average of the random sequence converge to the spatial mean of the invariant distribution. We then apply these results to logistic regression, empirical value iteration, empirical Q value iteration, and empirical relative value iteration for finite state finite action MDPs.

Key words. Stochastic Gradient Descent, Empirical Dynamic Programming, Constant Stepsize Q learning, Iterative Random Maps, Feller Markov Chains

AMS subject classifications. 93E35, 60J20, 68Q32

1. Introduction. There has been a surge of interest in using randomization to reduce computational burden in machine learning and reinforcement learning. For instance, in training neural networks with a large amount of data, stochastic gradient descent is frequently employed instead of the usual gradient descent. In data-driven Markov decision problems, empirical dynamic programming has been employed to determine approximately optimal policies and value functions. In these algorithms, instead of computing the expected value of certain functions at each step of the iteration, one computes the empirical expected value that is rather easy to compute if enough data is available. This simple trick reduces the runtime to determine reasonably good solution.

It turns out that the outputs of these stochastic recursive algorithms (SRAs) can be viewed as Markov chains. Indeed, if the parameters of the algorithm do not change with iteration, then the SRAs can be thought of as an iterated random map acting onto certain Euclidean spaces. For instance, in case of stochastic gradient descent, if the learning rate remains constant, data samples picked at every iteration are i.i.d., and the number of data samples remain constant, then each step of the stochastic gradient descent algorithm is a random map. To see this, let

∗Submitted to the editors.

Funding: The authors gratefully acknowledge support from ARPA-E NEXTCAR Program and NSF CRII Award 1565487.

†Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA. (gupta.706@osu.edu, tendolkar.3@osu.edu, chen.6945@osu.edu, pi.35@osu.edu).
us consider the problem of minimizing a sum of \( N \) functions, \( L_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, N \):

\[
\min_{x \in \mathbb{R}^n} L(x) = \frac{1}{N} \sum_{i=1}^{N} L_i(x).
\]

In usual gradient descent, one fixes a learning rate \( \beta > 0 \), and runs the iteration

\[
y_{k+1} = y_k - \beta \nabla_x L(y_k) =: T(y_k),
\]

where we used \( T : \mathbb{R}^n \to \mathbb{R}^n \) to denote the exact gradient descent map. In stochastic gradient descent, the map applied at every step of the algorithm changes. At time step \( k \) of the stochastic gradient descent algorithm, let \( \mathcal{N}_k := \{i_1, \ldots, i_n\} \) be the set of \( n \) indices that are sampled independently and uniformly from the set of all indices \( \{1, \ldots, N\} \). Then, we have

\[
\hat{z}_{k+1}^n = \hat{z}_k^n - \frac{\beta}{n} \sum_{i \in \mathcal{N}_k} \nabla_x L_i(\hat{z}_k^n) =: \hat{T}_k^n(\hat{z}_k^n).
\]

Since the set of (random) indices \( \mathcal{N}_k \) is i.i.d., the map \( \hat{T}_k^n \) is independent of the past maps and is “identically distributed”. This implies that the (random) sequence \( (\hat{z}_k^n)_{k \in \mathbb{N}} \) is a Markov chain. It should also be noted that the exact gradient descent map \( T \) and stochastic gradient descent map \( \hat{T}_k^n \) are related. We will make this relation precise in the sequel.

A similar setup is considered in empirical dynamic programming for dynamic decision process. Consider a controlled process in which \( s \) is the state of a system and \( a \) is the action of the decision maker. Let \( p(s' | s, a) \) denote the transition probability of the next state being \( s' \) given the current state \( s \) and action \( a \). In many dynamic programming equations, one needs to compute \( \mathbb{E}[v(s') | s, a] \), where \( v \) is some real-valued function of the state. If there is enough data, one can replace \( \mathbb{E}[v(s') | s, a] \) with its “empirical” average, \( \frac{1}{n} \sum_{i=1}^{n} v(s'_i) \), where \( \{s'_i\}_{i=1}^{n} \) are \( n \) samples of the next state given that the current state-action pair is \((s, a)\).

In empirical dynamic programming, if the number of samples of the next state given the current state and action remains the same throughout the runtime, then the algorithm can be viewed as an iterated random map. This idea can be applied to value iteration or Q learning algorithms for Markov decision problem (MDP), and it is called empirical value iteration or empirical Q value iteration, respectively. In the process, the sequence of value functions output by empirical value iteration, or the sequence of Q functions output by empirical Q value iteration, forms a Markov chain. Thus, in order to understand the convergence properties of such empirical processes, we can leverage the corresponding results available for Markov chains. Indeed, this yields insights that were unknown heretofore.

1.1. Our Contribution. The primary contribution of this paper is to conceptually unify the convergence analysis of certain SRAs in optimization, machine learning, and reinforcement learning using the tools from Markov chain. This is achieved by leveraging several results available for convergence and stability for Feller Markov processes. Specifically, our contributions are:

1. We derive sufficient conditions on the random operators and its relationship to the exact operator so that the distributions over the sequence \( (\hat{z}_k^n)_{k \in \mathbb{N}} \) output by the SRA
converges in weak* topology to the unit mass over the trajectory \((y_k)_{k \in \mathbb{N}}\) output by the exact algorithm as the parameter \(n \to \infty\). We show that these sufficient conditions are satisfied in sufficiently general class of problems encountered in stochastic gradient descent for strongly convex functions, empirical dynamic programming for MDPs with discounted cost and average cost criteria, and empirical Q value iteration.

2. We show that the Markov chain generated by many SRAs satisfy weak Feller property, that is, if \(f\) is a continuous and bounded function, then \(\hat{z}^n_0 \mapsto \mathbb{E}[f(\hat{z}^n_1) | \hat{z}^n_0]\) is also a continuous and bounded function. The existence (and in some case, the uniqueness) of invariant measure of Feller Markov chains has been presented in [8]. We apply these results to conclude that under some reasonable assumptions, the chains generated by stochastic gradient descent and empirical dynamic programming algorithms admit invariant distributions. In certain cases, we can show that this invariant distribution is unique.

3. The results on existence of invariant distributions is then used to show that the time average of the Markov chain converges in probability to the mean value of the invariant distribution. This part leverages the law of large numbers for Markov chains, presented in [34].

We complement the theoretical contributions with extensive numerical simulations of various SRAs. The fact that these remarkable results can be proved with little effort has not been explored in the existing literature.

While we present complete proofs of two of our main results stated here, we admit that most of the proofs follow somewhat standard arguments available in many texts. The need for presenting the complete proofs are twofolds: Our hypotheses differ in some ways from the hypotheses presented in the standard texts and papers. Moreover, to construct the complete proofs using these texts under our hypotheses require substantial effort on the part of the reader. To ease this burden, we chose to furnish the complete proofs using the notation adopted here.

1.2. Previous Work. Convergence proofs of randomized optimization and learning algorithms are usually obtained from specifically tailored arguments, which are not usually generalizable to other settings. For instance, the convergence of stochastic gradient descent, stochastic variance reduction gradient descent (SVRG), and stochastic average gradient (SAG) descent follow completely different, and often involved, sequence of arguments [9, 6, 27, 19, 36, 13]. The argument usually starts with identifying some conditions on the functions, such that for every iteration \(k\), one can upper bound \(L(\hat{z}^n_k) - L(x^*)\) (where we used the notation introduced above) by a function that decays as \(k\) grows. These tailored methods usually also yield the convergence rates specific to those algorithms. It would be conceptually elegant to determine a set of more general conditions which can be readily applied to these algorithms and many of its variants to establish the convergence guarantee.

Stochastic approximation method is one such elegant theory [29, 7]. For instance, it is well known that if the sequence of learning rate is not summable but square summable, then under reasonable assumptions on the loss function, stochastic gradient descent and distributed asynchronous gradient descent methods converge almost surely to the optimal solution [39, 15, 16]. In fact, stochastic approximation methods also yield convergence of reinforcement
learning algorithms to the optimal value function [24, 40, 4] and convergence of multi-agent reinforcement learning algorithms to the Nash equilibria [30, 23, 2, 37]. Empirical dynamic programming has also received significant attention recently to derive approximately optimal decision rules in many data-driven dynamic decision problems [42, 43, 11, 12, 26, 20].

The key challenge with stochastic approximation method is that the almost sure convergence guarantee to the point of interest is provided in the limit. This is largely due to the fact that to use stochastic approximation theory, the learning stepsize has to converge to 0 (the learning stepsize has to be not summable but square summable). In computing systems, this limit cannot be achieved in a reasonable time frame. As a result, constant step-size algorithms are gaining traction as a way to speed up the computation at the cost of tolerating a small error in the final result. For instance, in stochastic gradient descent with constant stepsize, the sequence generated by the algorithm gets closer to the optimal solution, but then does a random walk around the optimal solution. The closeness of the random walk to the optimal solution depends on the number of random samples one uses at each iteration of the algorithm.

Our work is largely motivated by the analysis of empirical dynamic programming in [20]. This work viewed empirical dynamic programs within the framework of iterated random operators. It used stochastic dominance based arguments to derive bounds on the asymptotic probability of error (between the random outputs of the algorithm and the optimal solution) being large. Inspired by this work, we extended the arguments to empirical relative value iteration in [18]. We further relaxed some conditions on random operators assumed in [20] in our follow up work [17]. The aim of this paper is to further expand the analysis and present conditions on random operators and its relationship to the exact operator to arrive at insightful conclusions about the random sequences generated by these SRAs.

1.3. Outline of the Paper. The paper is organized as follows. Section 2 presents a common mathematical framework to study the problem of convergence and stability of Markov chains induced by SRAs. We also state the three main questions we address. Section 3 presents some motivating examples where the mathematical framework we develop can be applied. Through these examples, we also illustrate certain desirable properties that the random operators enjoy. In Section 4, we show that the distributions over the trajectories generated by SRA converges to the Dirac mass over the trajectory generated by the exact algorithm. This constitutes the first main result of the paper. In Section 5, we study some sufficient conditions on the operators $\hat{T}_n^k$ such that the resulting Markov chain admits an invariant distribution. We also study conditions under which the invariant distribution is unique. Section 6 then introduces the assumptions under which the law of large numbers for Markov chains. This constitutes the second main result of the paper. The proofs of the two main results are presented in Sections 8 and 9. We finally conclude our discussion in Section 10.

1.4. Notations. Let $(A, \rho)$ be a Polish space, which is defined as complete separable metric space with metric $\rho$. We use $\varphi(A)$ to denote the set of all probability distributions over $A$. We use $\mathbf{1}_{\{a\}} \in \varphi(A)$ to denote the Dirac mass over $a \in A$. The notations $C_b(A)$ and $U_b(A)$ denote, respectively, the set of all continuous and bounded functions and uniformly continuous and bounded functions over the set $A$. We say that a sequence of measures $\{\mu_n\}_{n \in \mathbb{N}} \subset \varphi(A)$ converges to $\mu$ in weak* sense iff for every $f \in C_b(A)$, $\int f d\mu_n \to \int f d\mu$ as $n \to \infty$. This is
usually also referred to as weak convergence in probability theory literature.

2. Problem Formulation. Let \((\mathcal{X}, \rho)\) be a Polish space with metric \(\rho\). Consider a contraction map \(T : \mathcal{X} \to \mathcal{X}\) with contraction coefficient \(\alpha\) the unique fixed point denoted by \(x^*\). This means
\[
\rho(T(x_1), T(x_2)) \leq \alpha \rho(x_1, x_2) \quad \text{and} \quad T(x^*) = x^*.
\]
Starting from any initial point \(y_0 \in \mathcal{X}\), define the iterates
\[
y_k = T(y_{k-1}) \quad \text{for} \ k \in \mathbb{N}.
\]
By Banach contraction mapping theorem, this iteration converges to \(x^*\). In fact, we have
\[
\rho(y_k, x^*) \leq \alpha^k \rho(y_0, x^*).
\]
As discussed previously, in many instances, it is beneficial or required in many iterative algorithms to use randomization to evaluate an approximation of \(T(x)\) using a random operator. We now formulate a framework to analyze the output of this SRA rigorously.

Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a standard probability space, where \(\Omega\) is the set of uncertainties, \(\mathcal{F}\) is the Borel \(\sigma\)-algebra over \(\Omega\) and \(\mathbb{P}\) be the probability distribution function over \(\Omega\). Let \(\hat{T}_n : \mathcal{X} \times \Omega \to \mathcal{X}\) be a random operator that is used at \(k^{th}\) iteration and is indexed by a natural number \(n\). The index \(n\) would capture, for instance, the learning rate, the number of random samples used to approximate the operator \(T\), etcetara. Although \(\hat{T}_n\) is a function of \(\omega \in \Omega\), we will suppress this dependence for ease of exposition. Thus, \(\hat{T}_n(x) := \hat{T}_n(x; \omega)\). We make the following assumption on the independence of the sequence of operators \(\hat{T}_n\).

Assumption 2.1. For every \(x \in \mathcal{X}\), \(\hat{T}_k^n(x)\) and \(\hat{T}_{k'}^n(x)\) are statistically independent for \(k \neq k'\).

2.1. Key Questions. Consider the stochastic process that starts from \(z_0^n = y_0\) and define \(z_k^n = \hat{T}_{k-1}^n(z_{k-1}^n)\) for all \(k \in \mathbb{N}\). Due to Assumption 2.1 and the fact that \(n\) does not change with time, the stochastic process \(\{z_k^n\}_{k \in \mathbb{N}}\) is a time-homogeneous Markov chain. One can view \(z_k^n\) as an \(\mathcal{X}\)-valued Markov chain with the Markov transition kernel given by
\[
Q_n(B|x) = \mathbb{P}\{z_k^n \in B | z_{k-1}^n = x\}
\]
for any Borel set \(B \subset \mathcal{X}\). Note that \(Q_n\) does not depend on the time index \(k\), since what we have here is a time-homogeneous Markov process.

2.1.1. Convergence of Distribution of Trajectories. We are interested in deriving conditions on the random maps \(\hat{T}_k^n\) under which the random sequence generated by exact algorithm with high probability. Let us formulate the precise mathematical problem. We let \(\mu_n \in \mathcal{P}(\mathcal{X}^\mathbb{N})\) denote the joint distribution of the sequence \((z_0^n, z_1^n, z_2^n, \ldots)\). Endow \(\mathcal{X}^\mathbb{N}\) with the product topology so that it becomes a Polish space. Then, \(\mu_n\) is defined by
\[(2.2) \quad \mu_n(B_0 \times B_1 \times B_2 \times \ldots) = \int_{B_0 \times B_1 \times B_2 \times \ldots} 1_{\{y_0\}}(dx_0) Q_n(dx_1|x_0) Q_n(dx_2|x_1) \ldots,
\]
where $B_0, B_1, B_2, \ldots$ are Borel sets in $\mathcal{X}$.

In the similar vein, one can also view the iterates $(y_k)_{k \in \mathbb{N}}$ defined in (2.1) as a Markov chain on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with the distribution over this sequence defined by

\begin{equation}
\psi(B_0 \times B_1 \times B_2 \times \ldots) = \int_{B_0 \times B_1 \times B_2 \times \ldots} \mathbb{I}_{\{y_0\}}(dx_0) \mathbb{I}_{\{y_1\}}(dx_1) \mathbb{I}_{\{y_2\}}(dx_2) \ldots
\end{equation}

This is a Dirac mass on the sequence $(y_0, y_1, \ldots)$. Our first result, proved in Section 4, proves that under certain mild assumption on the random operators $\hat{T}_n$, the sequence of measures $\mu_n$ converges in the weak* sense to $\psi$.

A similar setup was considered by Karr [28]. It studies the convergence properties of a class of Feller Markov chains parametrized by $n$ such that the transition probability $Q_n$ converges in some sense to a transition probability $Q$ as $n \to \infty$. Although our assumptions are slightly different, the proof essentially imitates the one in [28] except for a couple of key steps. We also discuss numerical implication of this result in Section 4.

2.1.2. Existence of Invariant Measures for Fixed $n$. For the Markov chain $(\hat{z}_n^k)_{k \in \mathbb{N}}$, one of the key question is the existence of invariant distribution. There is a large body of literature that studies the problem of existence of invariant measures for Harris recurrent Markov chains that take values in continuous state spaces [8, 34]. However, the Markov chain generated by the SRAs seldom satisfy the strong recurrence structure required for Harris recurrent chains.

Instead, these chains satisfy the weak Feller conditions, for which there are limited results in the literature. Nonetheless, we show that many SRAs satisfy certain desirable properties, which can be leveraged to not only guarantee the existence of invariant distribution, but also establish the uniqueness of the invariant distributions. These properties of random operators are discussed in Section 5. This further leads to strong conclusions about law of large numbers, as we discuss next.

2.1.3. Convergence of Time Average of Iterates. The weak law of large numbers for independent and identically distributed (i.i.d.) random variables states that the time average of i.i.d. random variables converges to the mean of the distribution in probability under fairly mild conditions. In fact, such a version of weak law of large numbers is also available for Feller Markov chains. This is established for Feller chains in [10] for chains residing in a compact Hausdorff space with a unique invariant measure, and in [34, Section 18.5] for the non-compact case under certain technical conditions, which include existence of an invariant measure. It turns out that this result can be proved simply under the uniqueness of the invariant measure if the starting point $\hat{z}_0^n$ is chosen according to certain specific distribution (in fact, we do not need other technical conditions of [34, Section 18.5]). We prove this result in Section 6, the proof of which is adapted from the results from [10] and [34].

From applications viewpoint, there has been a sustained interest in using time-averages in stochastic gradient descent and deep Q learning. Particularly, references [36, 33, 25, 32, 38] propose that fixing the stepsize in stochastic gradient descent algorithms and using the average of tail of the random sequence leads to good performance of the trained algorithm. Within the context of reinforcement learning, [1] and [41] propose the use of averaging the deep Q function iterates to arrive at a solution with lower variance. Indeed, we show here that...
under some conditions on the random operators, the variance reduction property of time-average (or in these cases, tail average) is largely due to the fact that Markov chain output by SRA may be admitting a unique invariant distribution. It remains to be shown that the properties of random operators assumed here are satisfied by the problem settings considered in aforementioned references.

We illustrate the theoretical results using numerical simulations for batch gradient descent, empirical dynamic programming and Q learning in Section 7.

3. Motivating Examples. We introduce here two examples where we illustrate how the random operator framework can be applied.

3.1. Batch Gradient Descent in Logistic Regression. Logistic regression has been widely used in many binary or multi-class classification problems. For simplicity, we consider the logistic regression with binary classification. Let $U \subset \mathbb{R}^m$ be the set of feature vectors. Let $(u^i, l^i)_{i=1}^N \subset U \times \{0, 1\}$ denote the labeled dataset with $N$ data points and their labels. Our task is to model conditional probability distribution of label $l$ given feature vector $u \in U$. In logistic regression, we model $P\{l = 1 | u^i\}$ as $f(u^i; x)$ where $x \in X := \mathbb{R}^m$ are the parameters of $f$ to be learnt from data, where $f$ is defined below:

$$f(u; x) = \sigma(u^T x), \quad \text{where } \sigma(t) = \frac{1}{1 + e^{-t}}.$$  

Our goal is to compute the parameter $x$ that maximise the log likelihood given the labeled data, or equivalently minimize negative log likelihood. The log likelihood $L : \mathcal{X} \longrightarrow \mathbb{R}$ of i.i.d data under conditional distribution $f$ is given by

$$L(x) = \frac{1}{N} \sum_{i=1}^N L_i(x), \quad \text{where } L_i(x) := l^i \log f(u^i; x) + (1 - l^i) \log(1 - f(u^i; x))$$

It can be shown that the derivatives of $L_i$ are given by

$$\nabla_x L_i(x) = (l^i - f(u^i; x))u^i, \quad \nabla_{xx}^2 L_i(x) = -\left(f(u^i; x)(1 - f(u^i; x))\right)u^i u^T.$$ 

Consequently, each $L_i$ is a concave function (since Hessian is negative semidefinite), and thus, $L$ is a concave function over the space $\mathcal{X}$. If the matrix $[u^1 | \ldots | u^N]$ is full rank and $N > m$, then it immediately follows that $\nabla_{xx}^2 L(x)$ is a full rank matrix. Thus, the eigenvalues of the Hessian of $-L(x)$ is bounded from below by some $m > 0$. Moreover, the largest eigenvalue of the Hessian of $-L(x)$ is bounded from above by

$$\lambda_{\text{max}}(\nabla_{xx}^2(-L(x))) \leq \text{trace}\left(\nabla_{xx}^2(-L(x))\right) = \frac{1}{4N} \sum_{i=1}^N \|u^i\|_2^2 =: M.$$ 

Thus, under reasonable assumptions, we conclude that $-L$ is strongly convex and $M$-smooth. Since $-L(x)$ is convex in $x$, it has a unique minimum $x^*$. This minimum can be computed
using the usual gradient descent algorithm. The algorithm starts at $x_0$, picked arbitrarily, and proceeds in the direction of $-\nabla_x L(x_k)$ in small steps of size $\beta$:

\begin{equation}
\tag{3.1}
x_{k+1} = T(x_k) := x_k - \beta \nabla_x (-L(x_k)) = x_k - \frac{\beta}{N} \sum_{i=1}^{N} (f(u_i^1 x_k) - f_i^1) u_i^1,
\end{equation}

where $T : \mathcal{X} \rightarrow \mathcal{X}$ is the gradient descent map (dependent on the parameter $\beta$). It can be further shown that if $\beta$ is sufficiently small, then the operator $T$ is a contraction on $\mathcal{X}$, endowed with the Euclidean norm. In this case, if the eigenvalues of Hessian of $L(x)$ satisfy $0 < m \leq \lambda_i(\nabla^2_{xx}(-L(x))) \leq M < \infty$, then

\begin{align}
\|T(x) - T(x')\|_2^2 &= \|x - x' - \beta \nabla_x (-L(x)) + \beta \nabla_x (-L(x'))\|_2^2 \\
\tag{3.2}
&= \| (x - x') - \beta \nabla^2_{xx}(-L(x''))(x - x') \|_2^2 \\
&\leq \lambda_{\text{max}}(I - \beta \nabla^2_{xx}(-L(x''))) \|x - x'\|_2^2,
\end{align}

where $x''$ is a point on the line segment connecting $x$ to $x'$. Since $\left(I - \beta \nabla^2_{xx}(-L(x''))\right)$ is a symmetric matrix, its eigenvalues are real and lie between $1 - \beta M$ and $1 - \beta m$. If we choose $\beta > 0$ such that $\alpha := \max\{|1 - \beta M|, |1 - \beta m|\} < 1$, then the maximum eigenvalue of $I - \beta \nabla^2_{xx}(-L(x''))$ is less than or equal to $\alpha$. This yields

\begin{equation}
\tag{3.3}
\|T(x) - T(x')\|_2^2 \leq \alpha \|x - x'\|_2^2
\end{equation}

In practice, the exact gradient computation of loss function $-L$ is computationally expensive as it requires evaluating $N$ gradients at every time step. Therefore, to ease computational burden, mini-batch Stochastic Gradient Descent (SGD) is employed. In mini-batch SGD, at every step $k$, the gradient is approximated by a small, randomly sampled, subset (of size $n$) of the data set. To introduce this algorithm, let $\mathcal{N}_k \subset \{1, \ldots, N\}$ be the randomly sampled subset of size $n$. The state is updated as

\begin{equation}
\tag{3.4}
\hat{z}^n_{k+1} = \hat{T}^n_k(\hat{z}^n_k) = \hat{z}^n_k - \frac{\beta}{n} \sum_{j=1}^{n} \nabla x(-L_j(\hat{z}^n_k))
\end{equation}

Note that $\hat{T}^n_k$ is now a random operator. Some obvious properties of these random operators are:

1. $\hat{T}^n_k$ is continuous in $\hat{z}^n_k$.
2. For every $\epsilon > 0$ and $x \in \mathcal{X}$, we have
   \[
   \lim_{n \to \infty} \mathbb{P} \left\{ \|\hat{T}^n_k(x) - T(x)\|_2 > \epsilon \right\} = 0.
   \]
3. Suppose that for any compact set $\mathcal{K} \subset \mathcal{X}$, $\|\nabla x L_i(x)\|_2$ is uniformly bounded, that is, there exists $M_\mathcal{K} \in \mathbb{R}$ such that $\sup_{x \in \mathcal{K}} \sup_{1 \leq i \leq N} \|\nabla x L_i(x)\|_2 \leq M_\mathcal{K}$. Then, for any $\epsilon > 0$,
   \[
   \tag{3.5}
   \lim_{n \to \infty} \sup_{x \in \mathcal{K}} \mathbb{P} \left\{ \|\hat{T}^n_k(x) - T(x)\|_2 > \epsilon \right\} = 0.
   \]

This statement can readily be proved using Hoeffding inequality and union bound.
Let us depart from the specific case of logistic regression, and consider the case where \( L_i \) can be any strictly concave function for all \( i \in \{1, \ldots, N\} \). Then, \( \hat{T}_k^n \) satisfies the following property.

4. If all the eigenvalues of Hessian of \( L_i(x) \) satisfy \( 0 < m \leq \lambda(\nabla^2_x(-L_i(x))) \leq M < \infty \) for all \( i \in \{1, \ldots, N\} \), then every realization of the random operator \( \hat{T}_k^n \) is a contraction map with contraction coefficient \( \hat{\alpha}_n \equiv \max\{ |1 - \beta M|, |1 - \beta m| \} < 1 \). The proof of this result is identical to the derivation of the contraction coefficient of the exact gradient descent map \( T \) in the preceding discussions.

We now introduce the empirical dynamic programming algorithm in the context of value iteration for Markov decision problem (MDP) with discounted cost criteria.

### 3.2. Empirical Value Iteration for Discounted Cost MDP.

Consider a Markov Decision Problem (MDP) problem described by 4 tuple \( (S, A, c, g) \), where \( S \) is the state space, \( A \) is the action space, \( c: S \times A \to \mathbb{R} \) is the cost function, and \( g: S \times A \times [0, 1] \to S \) governs the state transition dynamics. The state transitions according to \( s_{t+1} = g(s_t, a_t, w_t) \), where \( w_t \) is the environment noise. Without loss of generality, we can assume \( w_t \) to be uniformly distributed in \([0, 1]\).

Let \( \Pi \) denote the set consisting of all possible deterministic policies \( \pi: S \to A \). The infinite horizon discounted cost \( v^\pi: S \to \mathbb{R} \) starting from state \( s \) and following policy \( \pi \) is given by

\[
v^\pi(s) := \mathbb{E} \left[ \sum_{k=0}^{\infty} \alpha^k c(s_k, a_k) \right] \mid s_0 = s, a_k = \pi(s_k),
\]

where \( \alpha \in (0, 1) \) is the discount factor. The goal is to compute the optimal value \( v^*(s) = \inf_{\pi \in \Pi} v^\pi(s) \). Let \( \mathcal{V} \) be the set of all \( v: S \to \mathbb{R} \); this space is isomorphic to the Euclidean space \( \mathbb{R}^{|S|} \).

It can be shown that the optimal infinite horizon discounted cost is a fixed point of a contraction map \( T: \mathcal{V} \to \mathcal{V} \), where \( T \) is the Bellman operator given by

\[
T(v)(s) = \inf_{a \in A} \left\{ c(s, a) + \alpha \mathbb{E}[v(g(s, a, W))] \right\}.
\]

Due to Banach contraction mapping theorem, \( T: \mathcal{V} \to \mathcal{V} \) admits a unique fixed point, which is equal to \( v^* \). The iterative process of finding this unique fixed point is called Value Iteration algorithm:

(3.7) Initialize \( v_0 \) randomly

\[
v_{k+1}(s) = T(v_k)(s) = \inf_{a \in A} \left\{ c(s, a) + \alpha \mathbb{E}[v_k(g(s, a, W))] \right\}
\]

In data driven applications, it is often the case that for all possible state-action pairs, multiple realizations of the next states are available. In this situation, we can replace the expectation in (3.6) to the empirical average. This algorithm is referred to as empirical dynamic programming, and is written as

(3.8) \( \hat{T}_k^n(\hat{v}_k^n)(s) = \inf_{a \in A} \left\{ c(s, a) + \frac{1}{n} \sum_{i=1}^{n} \hat{v}_k^n(g(s, a, w_i)) \right\} \)
where \( w_i \) are \( n \) independent and identically distributed samples of the noise \( W \), redrawn at every \( k \). The above intuition can be turned into an algorithm to determine approximately optimal value function, and is known as empirical value iteration:

\[
\hat{v}_{k+1}^n(s) = \hat{T}_k^n(\hat{v}_k^n)(s) = \inf_{a \in A} \left\{ c(s, a) + \alpha \frac{1}{n} \sum_{i=1}^{n} \hat{v}_k^n(g(s, a, w_i)) \right\}
\]

Note that \( \hat{T}_k^n \) is a random operator, and its realization is dependent on the noise samples \( (w_i)_{i=1}^{n} \). The following properties of \( \hat{T}_k^n \) are obvious:

1. \( \hat{T}_k^n \) is a contraction with contraction coefficient \( \alpha \). Therefore, \( \hat{T}_k^n \) is continuous.
2. For every \( \epsilon > 0 \) and \( v \in \mathcal{V} \), we have
   \[
   \lim_{n \to \infty} \mathbb{P} \left\{ \| \hat{T}_k^n(v) - T(v) \|_{\infty} > \epsilon \right\} = 0.
   \]
3. In fact, we have a stronger property here. For any compact set \( K \subset \mathcal{V} \), we have for every \( \epsilon > 0 \)
   \[
   \lim_{n \to \infty} \sup_{v \in K} \mathbb{P} \left\{ \| \hat{T}_k^n(v) - T(v) \|_{\infty} > \epsilon \right\} = 0.
   \]
4. Let \( \mathcal{V} \) be endowed with the partial order \( \preceq \) such that \( v_1 \preceq v_2 \) implies \( v_1(s) \leq v_2(s) \) for all \( s \in \mathcal{S} \). If \( c \geq 0 \), then \( \hat{T}_k^n \) satisfies
   (a) If \( v_0 = 0 \), then \( v_0 \preceq \hat{T}_k^n(v_0) \).
   (b) If \( v_1 \preceq v_2 \), then \( \hat{T}_k^n(v_1) \preceq \hat{T}_k^n(v_2) \).
   (c) If \( v_n \uparrow v \), then \( \hat{T}_k^n(v_1) \rightarrow \hat{T}_k^n(v) \).

As we observe in the rest of the paper, these properties can be exploited to yield useful insights about the convergence of the iterates (or the distributions of the iterates) in these randomized algorithms.

### 3.3. Observations

Through the two examples above, we observed that the approximate operator \( \hat{T}_k^n \) corresponding to the contraction operator \( T \) is context dependent. In the case of stochastic gradient descent, it is constructed by picking certain loss functions randomly and then averaging their gradients. In the case of empirical dynamic programming, the approximate operator involves computing the empirical average of the future expected value. Nonetheless, there are some fundamental properties that the empirical operator satisfies in both situations. For instance, the property stated in (3.5) in the context of logistic regression is (mathematically) the same as the property stated in (3.10) in the context of empirical value iteration. Similarly, every realization of the random operator \( \hat{T}_k^n \) is a contraction map under certain reasonable assumptions. We will consider more examples in Section 7, where we show that these properties (or some minor variant of these properties) are enjoyed by other empirical dynamic programming algorithms as well.

The other important observation is that every realization of random operator \( \hat{T}_k^n \) may also satisfy some other desirable properties. For instance, in the empirical value iteration example, if we endow \( \mathcal{V} \) with a partial order \( \preceq \) and the cost is nonnegative, then every realization of
\(\hat{T}_k^n\) satisfy certain monotonicity property. This is very useful in establishing the existence of unique invariant measure, as we show in Section 5. This property is, unfortunately, not satisfied by the logistic regression problem. This property is also not satisfied by the empirical relative value iteration for the average cost MDP, which we discuss in Section 7. However, we will show that the realizations of the random operators in these cases have some other desirable properties that lead to existence of a unique invariant distribution.

We now turn our attention to introducing our first main result in the next section.

4. Weak* Convergence of the Distribution of Trajectories. We now study the convergence property of the sequence of distributions \(\mu_n\), which is defined in (2.2). Before we study that, we need to ensure that the random operator \(\hat{T}_k^n\) is “close to” the map \(T\) in some sense. Accordingly, we make the following assumption.

Assumption 4.1. For every compact set \(K \subset X\), \(\epsilon > 0\), and \(\delta > 0\), there exists \(N_{\epsilon,\delta}(K) > 0\) such that

\[
\sup_{x \in K} \mathbb{P}\left\{ \rho(\hat{T}_k^n(x), T(x)) > \epsilon \right\} < \delta \quad \text{for all} \quad n \geq N_{\epsilon,\delta}(K).
\]

We recall here that this assumption is satisfied by the logistic regression and empirical value iteration for discounted cost MDP (see the discussion at the end of Subsections 3.1 and 3.2). We are now in a position to introduce our first main result.

Theorem 4.2. If Assumptions 2.1 and 4.1 hold, then \(\mu_n\) converges in weak* topology to \(\psi\) as \(n \to \infty\), where \(\psi\) is defined in (2.3).

Proof: The proof is based on the proof by [28], except that our hypotheses are slightly different from those in [28]. For completeness, we present a proof in Section 8.

Levy-Prohorov’s metric over the space of probability measures over Polish spaces metrizes the weak* topology [35]. Generally speaking, if distributions of two random variables are close to each other in Levy-Prohorov’s metric, then it does not imply that the random variables will be close to each other, as we show in the example below. However, if one of the random variable is deterministic (that is, its distribution is a Dirac mass), then the random variable must be close to the deterministic variable with high probability. This is established in the lemma following the example.

Example 4.3. Let \(X\) be a zero-mean unit-variance Gaussian random variable, and define \(Y = -X\). Let \(\mu\) denote the probability measure of \(X\) and \(\nu\) denote the probability measure of \(Y\). It is clear that \(\mu = \nu\) in weak topology (actually, it is true in any topology). However, the difference of the random variables \(X - Y = 2X\) is non-zero, and in fact, is mean-zero Gaussian random variable with variance 2. Thus, if two measures are close to each other in weak topology, then it does not imply that the underlying random variables are close to each other.

Let \(A\) be a Polish space with metric \(\rho_A\). Let \(d_P\) be the Levy-Prohorov’s metric on the space of probability measures \(\mathcal{P}(A)\) over \(A\). This metric is defined as follows. For a Borel set \(A \subset A\), let \(A^\epsilon\) be defined as

\[
A^\epsilon = \{a \in A : \rho_A(a, b) < \epsilon, b \in A\}.
\]
Let $\mu, \nu \in \wp(A)$. Then, $d_P(\mu, \nu)$ is defined by

$$d_P(\mu, \nu) = \inf \left\{ \epsilon > 0 : \mu(A) < \nu(A^c) + \epsilon, \nu(A) < \mu(A^c) + \epsilon \right\} \quad \text{for all Borel sets } A \subset A.$$ 

We are now in a position to introduce our next result. We believe that this result may not be new, but we could not locate a reference where this result is proved.

**Lemma 4.4.** Let $\mu \in \wp(A)$ and $\mathbb{I}_{\{a^*\}}$ be a unit mass at point $a^* \in A$. If $d_P(\mu, \mathbb{I}_{\{a^*\}}) < \epsilon$, then for any random variable $W$ distributed according to the law $\mu$, we have

$$P\{\rho_A(W, a^*) \geq \epsilon\} < \epsilon.$$ 

**Proof:** Let $B_\epsilon$ be an open $\epsilon$ ball around $a^*$. Let $(B_\epsilon^c)^c$ be defined as

$$(B_\epsilon^c)^c = \{ a \in A : \rho_A(a, b) < \epsilon, b \in B_\epsilon^c \}.$$ 

Note that $(B_\epsilon^c)^c = A \setminus a^*$, which implies $\mathbb{I}_{\{a^*\}}((B_\epsilon^c)^c) = 0$. Let $W$ be a random variable distributed according to the law $\mu$. Then, from the definition of Levy-Prohorov’s metric, we know that

$$\mu(B_\epsilon^c) < \mathbb{I}_{\{a^*\}}((B_\epsilon^c)^c) + \epsilon = \epsilon.$$ 

The proof then follows from noting that

$$P\{\rho_A(W, a^*) \geq \epsilon\} = \mu(B_\epsilon^c).$$

The proof is established. 

As a consequence of the lemma above, we conclude that since the distribution $\mu_n$ converges to the Dirac delta function $\psi$, it implies that for $n$ sufficiently large, the random sequence
generated by SRA lies within a small tube around the trajectory induced by the deterministic contraction map with high probability. This is illustrated in Figure 1.

To see this, let \( \hat{z}^n := (\hat{z}_0^n, \hat{z}_1^n, \ldots) \) and \( y = (y_0, y_1, \ldots) \). Endow the space \( \mathcal{X}^\mathbb{N} \) with the following metric:

\[
\rho_{\mathcal{X}^\mathbb{N}}(\hat{z}^n, y) = \sum_{k=0}^{\infty} \frac{1}{2^k} \rho(\hat{z}_k^n, y_k).
\]

It can be readily established that \( \rho_{\mathcal{X}^\mathbb{N}} \) defined above is a metric on \( \mathcal{X}^\mathbb{N} \). Then, due to Lemma 4.4, we conclude that

\[
P\{\rho_{\mathcal{X}^\mathbb{N}}(\hat{z}^n, y) < \epsilon\} \geq 1 - \epsilon.
\]

Next, note that if \( \hat{z}^n \) satisfies \( \rho_{\mathcal{X}^\mathbb{N}}(\hat{z}^n, y) < \epsilon \), then \( \hat{z}_k^n \) is within \( \epsilon \) neighborhood of \( y_k \) for most of the \( k \in \mathbb{N} \). Consequently,

5. Existence of Invariant Measures for Fixed \( n \). In this section, we identify conditions on the operators \( (T_k^n)_{k \in \mathbb{N}} \) so that the Markov chain \( \hat{z}_k^n \) admits a unique invariant distribution as \( k \to \infty \). The existence of an invariant measure is important in applications. It yields insight about “stability” of the SRA. In particular, if there is no invariant distribution, then it is likely the case that the sequence generated by SRA can blow up with positive probability. Thus, by tweaking the SRA (for instance, by changing the learning rate or increasing the number of samples), one can ensure that the sufficient conditions noted below are satisfied, thereby establishing that the SRA is stable and yields finite values with probability 1.

In case where there exists unique invariant measure under certain assumptions on the initial condition, then it means that any element in the tail of the random sequence generated by the SRA will have the law as its invariant distribution. This is a crucial step in proving that the time average of \( f(\hat{z}_k^n) \) for any \( f \in C_b(\mathcal{X}) \) converges in probability to the spatial average of the function with respect to the invariant measure. This important result is established in the next section.

To state the assumptions, we drop the subscript \( k \) wherever possible since the statistical properties of \( T^n_k \) and \( \hat{T}_k^n \) are independent and identical to each other as long as \( k \neq k' \). Below, we state three assumptions under which we can show that \( \hat{z}_k^n \) admits an invariant distribution.

Assumption 5.1. The following conditions are satisfied:

1. \( \mathcal{X} \) is a Polish space with partial order \( \preceq \). This ordering satisfies the following property: For any sequence \( (x_k)_{k \in \mathbb{N}} \) satisfying \( x_1 \preceq x_2 \preceq x_3 \preceq \ldots \), there exists a minimal element \( \bar{x} \in \mathcal{X} \) such that \( x_k \preceq \bar{x} \) for all \( k \in \mathbb{N} \). This is denoted as \( x_k \uparrow \bar{x} \).
2. The operator \( T^n \) satisfy
   (a) Monotonicity 1: exists \( x_0 \in \mathcal{X} \) such that \( x_0 \preceq T^n(x_0) \).
   (b) Monotonicity 2: If \( x_1 \preceq x_2 \), then \( T^n(x_1) \preceq T^n(x_2) \).
   (c) Continuity: If \( x_k \uparrow \bar{x} \), then \( T^n(x_k) \to T^n(\bar{x}) \) as \( k \to \infty \).

Assumption 5.1 is satisfied in Markov decision processes with non-negative cost functions. This has been noted in Subsection 3.2 in the context of empirical value iteration for discounted cost criterion. However, this is also satisfied in MDP for total cost criterion with an absorbing
state, the proof of which follows from arguments similar to the one made in Subsection 3.2; see, for example, [21, 22, 3].

**Assumption 5.2.** For $n \in \mathbb{N}$ and $m \in \mathbb{N}$, let $\hat{\alpha}_m$ denote the Lipschitz coefficient of $T^m_n \circ T^m_{n-1} \circ \ldots \circ T^1_1$. The following conditions are satisfied:
1. $\mathcal{X}$ is a compact Polish space.
2. For $n \in \mathbb{N}$, there exists $m \in \mathbb{N}$, such that for any $\epsilon > 0$, there exists a $\delta_{\epsilon} > 0$ such that
   \[ \hat{\alpha}_m \leq 1, \quad \mathbb{P} \{ \hat{\alpha}_m > 1 - \epsilon \} < \delta_{\epsilon}. \]

The notable point in Assumption 5.2 is that the assumption requires $\mathcal{X}$ to be compact. It is satisfied in empirical value iteration for MDP with average cost criterion, as long as we project the value functions outside a sufficiently large compact set back to that compact set. We adopted this approach earlier in [18] to ensure that the value functions obtained through repeated use of empirical operators don’t blow up. During simulations, however, we never needed to use projection, as the value functions were bounded.

**Assumption 5.3.** The following conditions are satisfied:
1. $\mathcal{X}$ is a Polish space.
2. There exists $a, b > 0$ such that the operator $\hat{T}^n$ satisfy
   \[ \mathbb{P} \left\{ \rho \left( \hat{T}^n(x^*), x^* \right) > \epsilon \right\} = \mathbb{P} \left\{ \rho \left( \hat{T}^n(x^*), T(x^*) \right) > \epsilon \right\} \leq \frac{a}{\epsilon^b}, \]
   where $x^*$ is the fixed point of $T$.
3. Let $\hat{\alpha}^n$ denote the Lipschitz coefficient of $(\hat{T}^n)_{k \in \mathbb{N}}$. Then,
   \[ \mathbb{E} [\hat{\alpha}^n] < \infty, \quad \mathbb{E} [\log(\hat{\alpha}^n)] < 0. \]

Assumption 5.3 is satisfied in stochastic gradient descent of strongly convex and smooth functions as noted in Subsection 3.1. This is also trivially satisfied in empirical value iteration for MDP with discounted reward, as we have noted in Subsection 3.2.

**Theorem 5.4.** Suppose that Assumption 2.1 holds. Additionally, if either one of three assumptions, Assumptions 5.1, 5.2, 5.3, holds, then there exists an invariant measure $\pi^n$ such that
\[ \mathbb{P} \left\{ \hat{T}^n(x) \in B | x \sim \pi^n \right\} = \pi^n(B). \]

Further, the invariant measure is unique under either of the following circumstances:
1. Assumption 5.1 holds, and the SRA is always initialized with $x_0$.
2. Assumption 5.2 holds.
3. Assumption 5.3 holds.

**Proof:** Under Assumption 5.1, the existence of invariant measures is proved in [8, Theorem 8.1, p. 79-81]. Under Assumption 5.2, the existence of invariant measures is proved in [5] and [8, Theorem 8.2, p. 82-83]. Under Assumption 5.3, the existence and uniqueness result is established in [14, Theorem 1.1, p. 87].
Remark 5.5. We can replace the assumption of $\mathcal{X}$ being a compact Polish space in Assumption 5.2 by making the following assumption. There exists $x_0 \in \mathcal{X}$ such that if $\hat{z}_0 \in \mathcal{X}$, then for any $\delta > 0$, there exists $N_\delta \in \mathbb{N}$ such that for any $k \geq 1$, we have

$$\mathbb{P}\left\{ \rho(\hat{z}_0^n, \hat{z}_n^k) > N_\delta \right\} < \delta. \quad (5.1)$$

If the above condition and Assumption 5.2 (2) holds, then one can show that for any initial condition $\hat{z}_0 \in \mathcal{X}$, a unique invariant distribution exists. For a proof, we refer the reader to [8, p. 179]. However, proving (5.1) is satisfied in usual SRAs appears to be difficult in our experience. □

We now turn our attention to establishing the law of large numbers for time averages of the outputs from a SRA.

6. Averaging of Iterates and The Weak Law of Large Numbers. In this section, we consider the problem of convergence of the sequence of averages of the random sequence $(\hat{z}_k^n)_{k \in \mathbb{N}}$ (or of $(f(\hat{z}_k^n))_{k \in \mathbb{N}}$ for some $f \in C_b(\mathcal{X})$). This problem is precisely the law of large numbers for Markov chains. This problem has been studied within the context of Markov chains over compact spaces in [10] and over general locally compact spaces in [34, Sec 18.5]. Let us formulate the problem precisely.

Let $\mathcal{X}$ be a Polish space, and consider a continuous function $f \in C_b(\mathcal{X})$. Let $\pi_n$ denote the invariant measure of the Markov chain $(\hat{z}_k^n)_{k \in \mathbb{N}}$. We have already studied the conditions under which such an invariant measure would exist in Theorem 5.4. In what follows, we show that under relatively mild assumptions, the sum

$$\frac{1}{K} \sum_{k=1}^{K} f(\hat{z}_k^n) \to \int f(x)\pi^n(dx) \text{ almost surely as } K \to \infty. \quad (6.1)$$

Let us define the operator $F : C_b(\mathcal{X}) \to C_b(\mathcal{X})$ and its adjoint $F^* : \wp(\mathcal{X}) \to \wp(\mathcal{X})$, where $\wp(\mathcal{X})$ is endowed with the weak* topology, as follows:

$$F(f)(x) = \int_{\mathcal{X}} f(y)Q_n(dy|x) = \mathbb{E}\left[ f(T^n(x)) \right], \quad F^*(\mu)(dy) = \int_{\mathcal{X}} Q_n(dy|x)\mu(dx).$$

Assumption 6.1. $\mathcal{X}$ is a Polish space. The distribution $\mu$ of the initial condition $\hat{z}_0^n$ is picked from a set $\mathcal{M} \subset \wp(\mathcal{X})$. Either of the following two conditions holds:

1. There exists a unique invariant measure $\pi^n$ such that for any $\mu \in \mathcal{M}$, $(F^*)^k(\mu)$ converges in weak* topology to $\pi^n$.
2. There exists a unique invariant measure $\pi^n$ such that for any $\mu \in \mathcal{M}$, the averaged operator satisfies

$$\lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} (F^*)^k(\mu) = \pi^n,$$

where the convergence is in weak* sense.
From Stolz-Cesaro theorem, it is easy to show that if Assumption 6.1(1) holds, then Assumption 6.1(2) holds as well. However, the converse may not be true. To prove our next result, we only need Assumption 6.1(2) to hold. It should be noted that Assumption 6.1(1) may be rather easy to prove using standard Markov chain theory, as is the case for all the examples considered in this paper.

Theorem 6.2. If Assumptions 2.1 and 6.1 hold, then (6.1) holds.

Proof: The proof essentially follows the steps in [10] and [34, Theorem 18.5.1, p. 478], except that we relax the assumption on compactness of the state space as assumed in [10] and replace the hypotheses in [34] with Assumption 6.1. For completeness, a detailed proof is presented in Section 9.

One way the presentation of Assumption 6.1 departs from the traditional Markov chain literature is as follows. It is generally assumed that \( \mathcal{M} = \mathcal{P}(\mathcal{X}) \), that is, for every \( \mu \in \mathcal{P}(\mathcal{X}) \), \( (F^*)^k(\mu) \) converges in weak* topology to \( \pi^n \). This is a very strong assumption from the applicability viewpoint in SRAs. In particular, it is possible to pick the most suitable initialization for the SRAs, which implies that \( \mathcal{M} \) can be picked appropriately. For example, in empirical dynamic programming for MDPs, one can initialize the value function to be 0. Then, we can utilize Assumption 5.1 (with \( x_0 = 0 \)) to established the existence of unique invariant distribution using Theorem 5.4. Incidentally, for the law of large number to hold, we don’t need the stronger condition of \( \mathcal{M} = \mathcal{P}(\mathcal{X}) \).

We have a slightly stronger version of the result above, which we capture in the following theorem.

Theorem 6.3. Suppose that \( \mathcal{X} \subset \mathbb{R}^n \) be a compact set. Suppose further that either of the following conditions hold:
1. Assumption 5.1 holds, and the SRA is always initialized with \( x_0 \).
2. Assumption 5.2 holds.
3. Assumption 5.3 holds.

Consider the average \( \tilde{a}_k^n \) of the Markov chain

\[
\tilde{a}_k^n = \frac{1}{k} \sum_{t=1}^k \tilde{z}_t^n.
\]

If Assumptions 2.1 holds, then for any initialization \( \tilde{z}_0^n \in \mathcal{X} \), a unique invariant distribution \( \pi^n \) exists and \( \tilde{a}_k^n \) converges almost surely to the mean \( \bar{a}_n \) of the distribution \( \pi^n \).

Proof: The existence of unique invariant measure is due to Theorem 5.4. Since \( \mathcal{X} \) is a compact set, we can take \( f(x) = x_i \) to conclude that \( (\tilde{a}_k^n)_i \) converges almost surely to \( \bar{a}_n \) by [10].

We now turn our attention to illustrating the application of Theorems 4.2 and 6.2 in various optimization and empirical dynamic programming algorithm.

7. Numerical Simulations. In this section, we complement the theoretical results proved above with extensive numerical simulations. We conduct simulations of minibatch stochastic gradient descent for logistic regression on MNIST dataset, empirical value iteration for discounted and average cost MDPs, and empirical Q value iteration.
7.1. Logistic Regression. Consider the task of classifying a subset of MNIST handwritten digits, where we consider only the images corresponding to numbers 0 and 1. Each data point is a $32 \times 32$ pixel image with the corresponding label (either 0 or 1). We use logistic regression for the classification task. We refer the reader to Subsection 3.1 for details of this problem and the notation used hereafter.

We transform each image into a vector and append 1 at the beginning of the vector. Thus, the space $\mathcal{U} = \{1\} \times [0, 1]^{1024}$. Thus, the space $\mathcal{X} = \mathbb{R}^{1025}$. As mentioned previously, the variable $n$ represents the batch size picked at every SGD iteration step. We pick $y_0$ arbitrarily in $\mathcal{X}$ and set $\hat{z}_0^n = y_0$. Then, we run the exact gradient descent and the minibatch
SGD as follows:

\[ y_{k+1} = T(y_k), \quad z_{k+1}^n = \hat{T}_k^n(z_k^n), \quad \hat{a}_k^n = \frac{1}{k+1} \sum_{i=0}^{k} z_i^n. \]

As discussed in Subsection 3.1, it is clear that the exact gradient descent is a contraction map for learning rate \( \beta \) small enough and we have large enough dataset, and therefore, \( y_k \) converges to the optimal solution \( x^* \). We can make the following claim about the operator \( \hat{T}_k^n \):

**Theorem 7.1.** The random operator \( \hat{T}_k^n \) satisfies Assumptions 2.1 and 4.1. Let \( \mu_n \) denote the distribution of \( (\hat{z}_0^n, \hat{z}_1^n, \ldots) \) and \( \psi \) be the Dirac mass on \( (y_0, y_1, \ldots) \). We have \( \mu_n \) converges to \( \psi \) in weak* topology as \( n \to \infty \).

**Proof:** The statements follow from the discussions in Subsection 3.1.

As a result of the theorem above, we conclude that \( \|y_k - \hat{z}_k^n\|_2 \) is generally small for most \( k \), and its variance converges to 0 as \( n \) increases. This is evident in Figure 2, where we show the distance \( \|y_k - \hat{z}_k^n\|_2 \) for various values of \( n \) (batch sizes) and \( k \) ranging from 0 to 10000. The black curve is the average and the red region shows one standard deviation of the distance \( \|y_k - \hat{z}_k^n\|_2 \) over hundred sample paths of the minibatch SGD iterations.

It is rather surprising to note that the random operator in the context of logistic regression does not satisfy any of the assumptions related to existence or uniqueness of the invariant measures. This is because each \( \nabla_x L_i \) is a rank 1 positive semidefinite matrix, and therefore, average of \( n \) such rank 1 matrices will be at most rank \( n \). Thus, if \( n < 1025 \), then it is impossible for each realization of \( \hat{T}_k^n \) to be a contraction map. Certainly, \( \hat{T}_k^n \) does not satisfy the monotonicity property of Assumption 5.1. As a result, we are unable to conclude the existence of an invariant measure, and we cannot resolve whether or not a unique invariant measure exists. Figure 3 plots \( \|\hat{a}_k^n - x^*\|_2 \), and we see that the variance does not go to 0 for any \( n \) as \( k \) grows large. Nonetheless, when we analyze the variance of the time average of the Markov chain, we found that the variance is reducing, albeit at a very slow rate. We leave the resolution of this mystery for a future work.

**7.2. Empirical Value Iteration for Discounted Cost MDP.** We consider here the empirical value iteration for discounted Markov decision processes (MDP) as described in Subsection 3.2. Consider value iteration algorithm applied to an MDP in which there are 20 states and 5 actions. We generate the state transition probability matrix for this MDP randomly at the beginning of the simulation.

We use here the notation introduced in Subsection 3.2. We initialize \( v_0 \) arbitrarily and set \( z_0^n = v_0 \), and define the iterates of exact value iteration and empirical value iteration as

\[ v_{k+1} = T(v_k), \quad z_{k+1}^n = \hat{T}_k^n(z_k^n), \quad \hat{a}_k^n = \frac{1}{k+1} \sum_{i=0}^{k} z_i^n. \]

We can prove the following result.

**Theorem 7.2.** The random operator \( \hat{T}_k^n \) satisfies Assumptions 2.1, 4.1, and 5.3. As a result, we conclude:
Figure 4. Plot of $\|v_k - \hat{z}^n_k\|$ for $n = 1, 25, 400$ for $k = 1, \ldots, 1000$. It is clear from the plots that as $n$ grows, the average and variance of $\|v_k - \hat{z}^n_k\|$ reduces. The mean and the variance are computed using 1000 independent runs of the iterations. There are 20 states and 5 actions in this MDP.

Figure 5. Plot of $\|v^* - \hat{a}^n_k\|$ for $n = 1, 25, 400$ for $k = 1, \ldots, 1000$. Notice that for every $n$, the variance in $\|v^* - \hat{z}^n_k\|$ reduces as $k$ increases. The plots are constructed using 1000 sample paths. The mean and the variance are computed using 1000 independent runs of the iterations.

1. Let $\mu_n$ denote the distribution of $(\hat{z}^n_0, \hat{z}^n_1, \ldots)$ and $\psi$ be the Dirac mass on $(y_0, y_1, \ldots)$. We have $\mu_n$ converges to $\psi$ in weak* topology as $n \to \infty$.
2. There exists a unique invariant distribution $\pi^n$ of the Markov chain $(\hat{z}^n_0, \hat{z}^n_1, \ldots)$.
3. The time average $\hat{a}^n_k$ converges in probability to the mean of $\pi^n$ as $k \to \infty$.

**Proof:** The proof follows from the discussions in Subsection 3.2. In particular, each realization of $\hat{T}^n_k$ is also a contraction with coefficient $\alpha$, and thus, $\hat{T}^n_k$ satisfies Assumption 5.3.

Figure 4 shows the difference $\|v_k - \hat{z}^n_k\|$ for every time step $k$ for various values of $n$. The red area shows one standard deviation of chart over hundred iterations of empirical value iteration. Figure 5 shows the time averaged infinity norm between value function of exact
value iteration and empirical value iteration at every time step.

*Figure 6.* The span norm between outputs of exact relative value iteration and empirical relative value iteration for $n = 1, 25, 400$. We observe that as $n$ grows, the probability that $\text{span}(\hat{z}_n^k, v^k)$ is large becomes smaller. The mean and the variance are computed using 1000 independent runs of the iterations. There are 20 states and 5 actions in this MDP.

*Figure 7.* Plot of $\text{span}(\hat{a}_n^k, v^*)$ for $n = 1, 25, 400$. Observe that for every $n$, the variance of $\text{span}(\hat{a}_n^k, v^*)$ converges to 0 as $k$ grows. We also observe that at $k = 1000$, the mean of $\text{span}(\hat{a}_n^k, v^*)$ becomes smaller as $n$ grows. The mean and the variance are computed using 1000 independent runs of the iterations.

**7.3. Empirical Relative Value Iteration for Average Cost MDP.** We now consider the average cost MDP in this section. The model for the MDP we consider here is the same MDP described in Subsection 3.2 — it has 20 states and 5 actions with the state transition matrix generated randomly at the beginning of the iterations. Let $P(s'|s, a)$ be the state transition probability matrix and is defined as $P(s'|s, a) = \mathbb{P}\{g(s, a, W) = s'\}$. The performance criteria here is the infinite horizon average cost, which is denoted by $v^\pi : \mathcal{S} \rightarrow \mathcal{R}$, where the agent acts according to a stationary policy $\pi$. Starting from state $s$ and following policy $\pi$ is given...
by
\[
    v^\pi(s) := \lim_{K \to \infty} \mathbb{E} \left[ \frac{1}{K+1} \sum_{k=0}^{K} c(s_k, a_k) \bigg| s_0 = s, a_k = \pi(s_k) \right].
\]

The goal is to compute the optimal value \( v^*(s) = \inf_{\pi \in \Pi} v^\pi(s) \). We again let \( V \) be the set of all \( v : S \to \mathbb{R} \). The Bellman operator \( T : V \to V \) is defined in this case as
\[
    T(v)(s) = \inf_{a \in A} \left\{ c(s, a) + \mathbb{E} [v(g(s, a, W))] \right\}
\]

The Bellman operator is not a contraction map on \( V \) under the infinity norm. However, if we endow the space \( V \) with span semi-norm, then we can prove a remarkable property of the Bellman map introduced above. The span semi-norm over the space \( V \) is defined as
\[
    \text{span}(v) = \max_{s \in S} v(s) - \min_{s \in S} v(s).
\]

It should be noted that if \( v \) is a constant map, that is, \( v(s) = v(s') \) for all \( s, s' \in S \), then \( \text{span}(v) = 0 \). The span semi-norm satisfies positivity, scalar multiplicative property, and triangle inequality properties of a norm, but it is not positive definite, as there are non-zero value functions (in particular, constant maps) that have zero span. Thus, \( (V, \text{span}) \) is not a Polish space, and the theoretical results proved in the paper cannot be applied directly onto this problem.

There is a small trick that allows us to use the framework developed above. Let us define two value functions to be equivalent, denoted by \( v_1 \sim v_2 \), if and only if \( \text{span}(v_1 - v_2) = 0 \). This is equivalent to saying that \( v_1 \sim v_2 \) if their difference is a constant map. It can be readily shown that this is an equivalence relationship, which further implies that the quotient space \( (V/\sim, \text{span}) \) is a complete normed vector (Banach) space. It is also a separable, which it inherits from \( \mathbb{R}^{|S|} \). Thus, \( (V/\sim, \text{span}) \) is a Polish space. We use \([v] \in V/\sim\) to represent the set of all functions \( v' \in V \) that satisfy the equivalence relationship with \( v \). By a slight abuse of notation, we write \( v \in V/\sim \).

Now, we can extend the Bellman operator \( T \) which takes as input an element of \( (V/\sim, \text{span}) \) and outputs an element in \( (V/\sim, \text{span}) \), then it may be a contraction map. In particular, \( T \) satisfies
\[
    \text{span}(T(v_1) - T(v_2)) \leq \alpha \text{span}(v_1 - v_2),
\]
where
\[
    \alpha = 1 - \min_{s, s', a, a' \in A} \sum_{s \in S} \min\{P(\hat{s}|s, a), P(\hat{s}|s', a')\}
\]

Note that \( \alpha < 1 \) if for each pair of state-action tuples, there exists at least one state to which both tuples can reach with positive probability. Thus, for the purpose of this paper, we make the following assumption.

**Assumption 7.3.** \( \alpha \), as defined in (7.1), is strictly less than 1.
If we implement value iteration by picking \( v_0 \) arbitrarily and defining \( v_{k+1} = T(v_k) \), then this implementation runs into a problem as the value functions \( v_k \) can explode to infinity as \( k \) goes to \( \infty \). Thus, to determine the optimal value \( v^* \), we can use relative value iteration, in which at every time step, a constant is subtracted from the value function to ensure it remains finite at every time step. The modified Bellman map, denoted by \( \tilde{T} \), in this case is defined as:

\[
\tilde{T}(v)(s) = u(s) - \min_s u(s)
\]

The relative value iteration starts with \( v_0 \in V \) arbitrarily and then runs \( v_{k+1} = \tilde{T}(v_k) \). This iteration is guaranteed to converge to a point \( v_\infty \), which lies in the same equivalence class as \( v^* \).

In data driven applications, we can use a simulator to generate independent samples of the next state and this leads to empirical relative value iteration, which has been studied in [18]. At each step of the iteration, the map \( \hat{T}_n \) is given by

\[
\hat{T}_n(v)(s) = \hat{u}(s) - \min_s \hat{u}(s)
\]

Now pick \( v_0 \) arbitrarily and set \( \hat{z}_0^n = v_0 \). Let us define the relative value iteration and empirical relative value iteration as

\[
v_{k+1} = \tilde{T}(v_k), \quad \hat{z}_{k+1}^n = \hat{T}_n(\hat{z}_k^n).
\]

Further, define \( \hat{a}_n^k \) as the average of the iterates \( \hat{z}_0^n, \ldots, \hat{z}_k^n \):

\[
\hat{a}_n^k = \frac{1}{k+1} \sum_{i=0}^k \hat{z}_i^n.
\]

We now have the following result:

**Theorem 7.4.** The random operator \( \hat{T}_n \) satisfies Assumptions 2.1, 4.1. If Assumption 7.3 holds, then for \( n \) sufficiently large, \( \hat{T}_n \) satisfies Assumption 5.3. As a result, we conclude:

1. Let \( \mu_n \) denote the distribution of \( (\hat{z}_0^n, \hat{z}_1^n, \ldots) \) and \( \psi \) be the Dirac mass on \( (y_0, y_1, \ldots) \). We have \( \mu_n \) converges to \( \psi \) in weak* topology as \( n \to \infty \).
2. There exists a unique invariant distribution \( \pi^n \) of the Markov chain \( (\hat{z}_0^n, \hat{z}_1^n, \ldots) \).
3. The time average \( \hat{a}_n^k \) converges in probability to the mean of \( \pi^n \) as \( k \to \infty \).

**Proof:** The proof follows from the discussions in Subsection 3.2. In particular, each realization of \( \tilde{T}_k^n \) is also a contraction with coefficient \( \alpha \), and thus, \( \hat{T}_n \) satisfies Assumption 5.3.

Figure 6 shows \( \text{span}(\hat{z}_k^n - v_k) \) for various values of \( n \) for \( k \) ranging from 0 to 1000. The red area shows one standard deviation of \( \text{span}(\hat{z}_k^n - v_k) \) over hundred iterations of empirical relative value iteration. Figure 7 shows the \( \text{span}(\hat{a}_k^n - v^*) \).
7.4. Empirical Q Value Iteration for Discounted Cost MDP. Q-Value iteration is another algorithm that, like value iteration, computes the optimal value function in MDPs. Let \( Q \) denote the set of all Q-value functions \( Q : S \times A \rightarrow \mathbb{R} \). Similar to Bellman operator of value iteration, we define an operator \( G : Q \rightarrow Q \) as

\[
T(q)(s,a) = c(s,a) + \alpha E \left[ \min_{a' \in A} q(g(s,a,W),a') \right]
\]

Similar to Bellman operator, \( T \) is a contraction on \((Q, \| \cdot \|_\infty)\). Further, it can be shown that the fixed point of \( G \) is \( q^* \), which is defined as \( q^*(s,a) = c(s,a) + \alpha E [v^*(s')|s,a] \), where \( v^*(\cdot) = \min_{a \in A} q(\cdot, a) \). The Q value iteration starts with an arbitrary \( q_0 \in Q \) and generates the sequence according to \( q_{k+1} = T(q_k) \), which converges to \( q^* \) as \( k \to \infty \).

The exact operator, as in other cases considered in the paper, can be approximated by the empirical operator \( \hat{T}_n^k \):

\[
\hat{T}_n^k(q)(s,a) = c(s,a) + \alpha \frac{1}{n} \sum_{i=1}^n \min_{a' \in A} q(g(s,a,w_i),a')
\]

Let us define \( \hat{z}_n^{k+1} = \hat{T}_n^k(\hat{z}_n^k) \), where \( \hat{z}_n^0 = q_0 \). Let \( \hat{a}_n^k \) be the time averaged version of \( \hat{z}_n^k \). The properties of the random operator \( \hat{T}_n^k \) for empirical Q value iteration has the same properties as listed in Theorem 7.2 for the case of empirical value iteration. Thus, we omit repetition of the essentially the same result here. The simulation results are plotted in Figure 8 and 9.

Figure 8. Plot of \( \|q_k - \hat{z}_n^k\| \) for \( n = 1, 25, 400 \) for \( k = 1, \ldots, 1000 \). It is clear from the plots that as \( n \) grows, the average and variance of \( \|q_k - \hat{z}_n^k\| \) reduces. The mean and the variance are computed using 1000 independent runs of the iterations. There are 20 states and 5 actions in this MDP.

8. Proof of Theorem 4.2. To prove Theorem 4.2, we need to introduce some further notation. Let \( C_b(\mathcal{X}) \) and \( U_b(\mathcal{X}) \) denote, respectively, the space of all continuous and bounded functions and uniformly continuous and bounded functions over \( \mathcal{X} \). Let \( \Pi_k : \mathcal{X}^k \rightarrow \mathcal{X}^{k+1} \) denote the projection operator that projects a sequence to its first \((k+1)\) components, that
Figure 9. Plot of $\|q^* - \hat{a}_{n_k}\|$ for $n = 1, 25, 400$ for $k = 1, \ldots, 1000$. Notice that for every $n$, the variance in $\|q^* - \hat{z}_n\|$ reduces as $k$ increases. The mean and the variance are computed using 1000 independent runs of the iterations.

is,

$$\Pi_k(x_0, x_1, x_2, \ldots) = (x_0, \ldots, x_k), \quad k \in \mathbb{Z}_+.$$ 

For a measure $\mu \in \wp(\mathcal{X}^N)$, let $\mu \circ \Pi_k^{-1} \in \wp(\mathcal{X}^k)$ denote the pullback of the measure to the first $k + 1$ components. We note the following fact from probability theory.

**Theorem 8.1.** Let $(\nu_n)_{n \in \mathbb{N}} \subset \wp(\mathcal{X}^N)$ be a sequence of measures. Then, $(\nu_n)_{n \in \mathbb{N}}$ converges weakly to $\nu_\infty$ if and only if $\nu_n \circ \Pi_k^{-1}$ converges weakly to $\nu_\infty \circ \Pi_k^{-1}$ for every $k \in \mathbb{Z}_+$.

Recall that $\mu_n$ is a measure defined on a stochastic sequence and we would like to show that $\mu_n$ converges weakly to $\psi$. Due to the theorem above, all we need to establish is that $\mu_n \circ \Pi_k^{-1}$ converges weakly to $\psi \circ \Pi_k^{-1}$ for every $k$. We will establish this result via an induction argument. First, we have some preliminary results.

**Lemma 8.2.** If $g \in U_b(\mathcal{X})$, then $g \circ T \in U_b(\mathcal{X})$.

**Proof:** Fix $\epsilon > 0$. Since $g \in U_b(\mathcal{X})$, we can pick a $\delta_\epsilon > 0$ such that for any $x, x' \in \mathcal{X}$ with $\rho(x, x') < \delta_\epsilon$, $|g(x) - g(x')| < \epsilon$. Now, for any for any $y, y' \in \mathcal{X}$ with $\rho(y, y') < \delta_\epsilon$, we know that $\rho(Ty, Ty') < \alpha \delta_\epsilon < \delta_\epsilon$.

Taking $x = Ty$ and $x' = Ty'$, we conclude that $|g \circ T(y) - g \circ T(y')| < \epsilon$.

This implies that $g \circ T \in U_b(\mathcal{X})$.

We use mathematical induction to prove the result. In the induction step, we prove that if $\mu_n \circ \Pi_k^{-1} \Rightarrow \psi \circ \Pi_k^{-1}$ and for every closed set $F_1 \subset \mathcal{X}_k^{-1}$ and $F_2 \subset \mathcal{X}$, we have

$$\lim_{n \to \infty} \mu_n \circ \Pi_k^{-1}(F_1 \times F_2) \leq \psi \circ \Pi_k^{-1}(F_1 \times F_2).$$
By Portmanteau theorem, we then conclude that $\mu_n \circ \Pi_k^{-1} \to \psi \circ \Pi_k^{-1}$ in weak topology.

The proof is as follows: Fix $n \in \mathbb{N}$. For $k = 0$, let $\mu_n \circ \Pi_0^{-1}(F) = \mathbf{1}_{\{y_0\}}(F) = \psi \circ \Pi_0^{-1}(F)$. Thus, the statement is true for $k = 0$. Let us assume that the statement holds for $k = 1, \ldots, l$. Then, we make the following claim.

**Lemma 8.3.** Suppose that Assumption 4.1 holds. Then, for every $\epsilon > 0$, there exists an $N_\epsilon$ such that

$$\left| \int_{F_1} g(T(x_l))\mu_n \circ \Pi_l^{-1}(dx_{0:l}) - \int_{F_1} g(T(x_l))\psi \circ \Pi_l^{-1}(dx_{0:l}) \right| < \epsilon \quad \text{for all } n \geq N_\epsilon,$$

and

$$\int_{X} \left| \mathbb{E} \left[ g(\tilde{T}_n(x_l)) \right] - g(T(x_l)) \right| \mu_n \circ \Pi_l^{-1}(dx_{0:l}) \leq \epsilon \quad \text{for all } n \geq N_\epsilon.$$

**Proof:** Lemma 8.2 implies that $g \circ T$ is uniformly continuous. The first result is a direct consequence of the induction hypothesis. The proof of the second result is presented in Appendix A.

Fix $F_1 \subset X_{l+1}$ closed. Then, using Lemma 8.3, for any $g \in U_b(X)$, there exists $N_\epsilon$ such that

$$\int_{F_1} \int_{X} g(x_{l+1})\mu_n \circ \Pi_{l+1}^{-1}(dx_{0:l+1}) - \int_{F_1} \int_{X} g(x_{l+1})\psi \circ \Pi_{l+1}^{-1}(dx_{0:l+1})$$

$$= \int_{F_1} \mathbb{E} \left[ g(\tilde{T}_n(x_l)) \right] \mu_n \circ \Pi_l^{-1}(dx_{0:l}) - \int_{F_1} g(T(x_l))\psi \circ \Pi_l^{-1}(dx_{0:l})$$

$$\leq \int_{X_{l+1}} \left| \mathbb{E} \left[ g(\tilde{T}_n(x_l)) \right] - g(T(x_l)) \right| \mu_n \circ \Pi_l^{-1}(dx_{0:l})$$

$$+ \int_{F_1} g(T(x_l))\mu_n \circ \Pi_l^{-1}(dx_{0:l}) - \int_{F_1} g(T(x_l))\psi \circ \Pi_l^{-1}(dx_{0:l})$$

$$< 2\epsilon. \quad (8.1)$$

Now, for any $F_2 \subset X$ closed, we can construct a sequence of $(g_m) \subset U_b(X)$ such that $g_m \downarrow 1_{F_2}$. This leads us to the following inequality for every $m \in \mathbb{N}$:

$$\int_{F_1} \int_{X} 1_{F_2}(x)\mu_n \circ \Pi_{l+1}^{-1}(dx_{0:l+1}) \leq \int_{F_1} \int_{X} g_m(x)\mu_n \circ \Pi_{l+1}^{-1}(dx_{0:l+1}).$$

Taking the limsup on both sides and using (8.1), we arrive at the following inequality

$$\limsup_{n \to \infty} \mu_n \circ \Pi_{l+1}^{-1}(F_1 \times F_2) \leq \int_{F_1} \int_{X} g_m(x_{l+1})\psi \circ \Pi_{l+1}^{-1}(dx_{0:l+1}). \quad (8.2)$$

Now, since the right hand side holds for every $m \in \mathbb{N}$, we take the limit $m \to \infty$ and use bounded convergence theorem to conclude

$$\lim_{m \to \infty} \int_{F_1} \int_{X} g_m(x_{l+1})\psi \circ \Pi_{l+1}^{-1}(dx_{0:l+1}) = \psi \circ \Pi_{l+1}^{-1}(F_1 \times F_2). \quad (8.3)$$
Collecting the two inequalities in (8.2) and (8.3), we conclude that
\[
\limsup_{n \to \infty} \mu_n \circ \Pi_{l+1}^{-1}(F_1 \times F_2) \leq \psi \circ \Pi_{l+1}^{-1}(F_1 \times F_2).
\]

The proof of the lemma is complete.

9. **Proof of Theorem 6.2.** For \( f \in C_b(\mathcal{X}) \), we define \( g_k \in C_b(\mathcal{X}) \), \( k = 0, 1, \ldots \) as
\[
g_0(x) = f(x), \quad g_k(x) = F^k(f)(x) = \mathbb{E} \left[ f \left( \hat{T}_n \circ \ldots \circ \hat{T}_1(x) \right) \right].
\]

The following equation follows immediately from the above definitions:
\[
\mathbb{E} \left[ g_k(\hat{z}_n^{m})|\hat{z}_m^{n-1} \right] = g_{k+1}(\hat{z}_m^{n-1}). \tag{9.1}
\]

Define the constant function \( c_f \in C_b(\mathcal{X}) \) as
\[
c_f(x) = \int f \, d\pi^n.
\]

The average of the functions \( g_k \), denoted by \( \bar{g}_k \), is
\[
\bar{g}_k(x) = \frac{1}{k} \sum_{i=1}^{k} g_{i-1}(x).
\]

For a function \( f \in C_b(\mathcal{X}) \) and a set \( C \subset \mathcal{X} \), we use \( f|_C \) to denote the restriction of the function on the set \( C \).

We now prove three lemmas that leads to the result. For the next result, let us define the occupation measure \( \eta_k \) over the set \( C \subset \mathcal{X} \) as
\[
\eta_k(C) = \frac{1}{k} \sum_{i=0}^{k-1} 1_{\{\hat{z}_i^m \in C\}}, \tag{9.2}
\]

where \( 1_{\{\cdot\}} \) is an indicator function, which takes the value of 1 if \( \{\cdot\} \) is true and 0 otherwise. We claim the following.

**Lemma 9.1.** *If Assumptions 2.1 and 6.1 holds, then for every \( \epsilon > 0 \), there exists a compact set \( C_\epsilon \subset \mathcal{X} \) such that
\[
\limsup_{k \to \infty} \mathbb{P} \left\{ \eta_k(C_\epsilon^k) \geq \epsilon \right\} < \epsilon.
\]

**Proof:** Note that since \( \tilde{F}_k^*(\mu) \to \pi^n \) for any \( \mu \in \mathcal{M} \) by Assumption 6.1, we conclude from Prohorov’s theorem that the sequence \( \{\tilde{F}_k^*(\mu)\}_{k \in \mathbb{N}} \) is tight. Thus, for \( \epsilon > 0 \), let \( C_\epsilon \) be the compact set such that
\[
\tilde{F}_k^*(\mu)(C_\epsilon^k) < \epsilon^2 \text{ for all } k \in \mathbb{N}.
\]
Further, we note that for any set \( C \subset X \), we have
\[
E[\eta_k(C)] = E\left[\frac{1}{K} \sum_{i=0}^{K-1} 1\{\hat{z}^n_k \in C\}\right] = \bar{F}^*_k(\mu)(C).
\]

Using the above identity and using Markov’s inequality, we conclude that
\[
P\{\eta_k(C^c_\epsilon) \geq \epsilon\} \leq \frac{E[\eta_k(C^c_\epsilon)]}{\epsilon} < \epsilon.
\]

The proof of the theorem is complete.

**Lemma 9.2.** Let \( C \subset X \) be a compact set. Then, the sequence of functions \( \{\bar{g}_k|C\}_{k \in \mathbb{N}} \) is uniformly bounded and equicontinuous and converges uniformly to \( c_f|C \).

**Proof:** First, we note that \( \|g_i\|_\infty \leq \|f\|_\infty \) for all \( i \in \mathbb{N} \), which implies that \( \bar{g}_k|C \) is uniformly bounded.

The proof of equicontinuity follows directly from Assumption 6.1(2) and Ascoli theorem. Note that as \( k \to \infty \), we get
\[
\bar{g}_k(x) = \langle f, \bar{F}^*_k(\delta_x) \rangle \to c_f(x).
\]

Thus, by Ascoli’s theorem, \( \{\bar{g}_k|C\}_{k \in \mathbb{N}} \) is an equicontinuous sequence of functions. The result then follows using Assumption 6.1(2).

Next, we claim that

**Lemma 9.3.** For every \( M \in \mathbb{N} \), we have
\[
\lim_{N \to \infty} \left| \frac{1}{N} \sum_{l=0}^{N-1} \left( g_0(\hat{z}^n_l) - \bar{g}_M(\hat{z}^n_l) \right) \right| = 0 \quad \text{\( \mathbb{P} \)-almost surely.}
\]

**Proof:** See Appendix B.

The proof can now be completed easily. Fix \( \epsilon > 0 \) and recall the definition of the set \( C_\epsilon \) from Lemma 9.1. We now note that for every \( K \in \mathbb{N} \) and \( M \in \mathbb{N} \), we have
\[
\frac{1}{K} \sum_{k=1}^{K} f(\hat{z}^n_k) - \int f d\pi^n = \frac{1}{K} \sum_{k=1}^{K} g_0(\hat{z}^n_k) - \int f d\pi^n
\]

\[
\leq 1\{\hat{z}^n_0 \in C_\epsilon, \ldots, \hat{z}^n_K \in C_\epsilon\} \left( \frac{1}{K} \sum_{k=1}^{K} g_0(\hat{z}^n_k) - \int f d\pi^n \right) + 2\|f\|_\infty \eta_K(C^c_\epsilon),
\]

\[
< 1\{\hat{z}^n_0 \in C_\epsilon, \ldots, \hat{z}^n_K \in C_\epsilon\} \left\{ \frac{1}{K} \sum_{k=1}^{K} \left( g_0(\hat{z}^n_k) - \bar{g}_M(\hat{z}^n_k) \right) + \left( \frac{1}{K} \sum_{k=1}^{K} \bar{g}_M(\hat{z}^n_k) - \int f d\pi^n \right) \right\}
\]

\[
+ 2\|f\|_\infty \eta_K(C^c_\epsilon).
\]
Since $\tilde{g}_M(x) \to \int f d\pi^n$ uniformly on the compact set $C_\epsilon$ due to Lemma 9.2, we can pick $M_\epsilon$ sufficiently large such that for all $K \in \mathbb{N}$ and $M \geq M_\epsilon$, we get
\[
1\{z^n_0 \in C_\epsilon, \ldots, z^n_K \in C_\epsilon\} \left(\frac{1}{K} \sum_{k=1}^{K} \tilde{g}_M(z^n_k) - \int f d\pi^n\right) < \epsilon.
\]
For such $M_\epsilon$, as $K \to \infty$, the first summand goes to 0 by Lemma 9.3. In the third summand, we know that $\eta_K(C_\epsilon)$ is less than $\epsilon$ with probability at least $1 - \epsilon$ due to Lemma 9.1 for sufficiently large $K$. Collecting all these results, we conclude that
\[
\limsup_{K \to \infty} \mathbb{P}\left\{\left|\frac{1}{K} \sum_{k=1}^{K} f(\tilde{z}^n_k) - \int f d\pi^n\right| < (2\|f\|_\infty + 2)\epsilon\right\} \geq 1 - \epsilon.
\]
This completes the proof of the theorem.

10. Conclusion. In this paper, we studied the convergence of random sequence generated from certain SRAs used in machine and reinforcement learning problems. If the randomization device used within the algorithm is independent at every iteration, and the maps don't change (for instance, the learning rate is taken as constant), then the random sequence generated can be viewed using the lens of Markov chains. We leveraged many strong results available in Markov chains literature, particularly for Feller Markov chains, and deduced many interesting characteristics of the random sequence and their distributions.

We hope that the unified framework developed in this paper will be useful for analyzing many other learning algorithms in the future. One of the problems left open for further research is the case where the number of samples or the learning rate is changed over time, which leads to non-homogeneous Markov chain. Indeed, these situations are typically studied using stochastic approximation theory, but it suffers from the disadvantage of getting only asymptotic convergence. Perhaps reformulating them as Markov chains can throw some light on their speed of convergence. We aim at studying these interesting problems in the future.

Appendix A. Proof of Lemma 8.3. Since $\mu_n \circ \Pi^{-1}_l \to \psi \circ \Pi^{-1}_l$ in weak topology as $n \to \infty$, we conclude that $\left(\mu_n \circ \Pi^{-1}_l\right)_{n \in \mathbb{N}}$ is tight. For a fixed $\epsilon > 0$, let $F_2 \subset \mathcal{X}$ be the compact set such that
\[
\mu_n \circ \Pi^{-1}_l(\mathcal{X} \setminus F_2) > 1 - \frac{\epsilon}{4\|g\|_\infty} \text{ for every } n \in \mathbb{N}.
\]
We now need the following result.

Lemma A.1. If Assumption 4.1 holds, then for any $g \in U_b(\mathcal{X})$, compact set $K \subset \mathcal{X}$ and $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that
\[
\left|\mathbb{E}\left[g(T^n_k(x))\right] - g(T(x))\right| < \epsilon \text{ for all } x \in K.
\]

Proof: Since $g$ is uniformly continuous, for every $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that for any $x, x' \in \mathcal{X}$ with $\rho(x, x') < \delta_\epsilon$, we have $|g(x) - g(x')| < \epsilon$. Since Assumption 4.1 holds,
there exists \( N_\varepsilon(g,K) \) such that
\[
\sup_{x \in K} P\left\{ \rho(\hat{T}_k^n(x), T(x)) > \delta \right\} < \frac{\varepsilon}{2\|g\|_\infty}
\]
for all \( n \geq N_\varepsilon(g,K) \).

This implies
\[
\left| \mathbb{E}\left[ g(\hat{T}_k^n(x)) - g(T(x)) \right] \right| \leq \int |g(\hat{T}_k^n(x)) - g(T(x))| P\{d\omega\}
\leq \varepsilon P\left\{ \rho(\hat{T}_k^n(x), T(x)) < \delta \right\}
+ 2\|g\|_\infty P\left\{ \rho(\hat{T}_k^n(x), T(x)) \geq \varepsilon \right\}
< 2\varepsilon.
\]

The proof of the lemma is complete.

We are now in a position to prove the result. Consider the following expressions
\[
\int_{X^{l+1}} \left| \mathbb{E}\left[ g(\hat{T}_k^n(x)) - g(T(x)) \right] \right| \mu_n \circ \Pi_l^{-1}(dx_0)
= \int_{X^l \times F_n} \left| \mathbb{E}\left[ g(\hat{T}_k^n(x)) - g(T(x)) \right] \right| \mu_n \circ \Pi_l^{-1}(dx_0)
+ \int_{X^l \times F_\infty} \left| \mathbb{E}\left[ g(\hat{T}_k^n(x)) - g(T(x)) \right] \right| \mu_n \circ \Pi_l^{-1}(dx_0)
\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\]
The proof of the lemma is complete.

**Appendix B. Proof of Lemma 9.3.** Let us define random variables \( u^p_l \) for \( l, p = 0, 1, \ldots \) as
\[
u^p_l = \begin{cases} 
  g_p(\hat{z}^n_l) - g_{p+1}(\hat{z}^n_{l-1}) & p < l \\
  0 & p \geq l 
\end{cases}
\]

It is now easy to see that
\[
\mathbb{E}\left[ u^p_l | \hat{z}^n_{l-1} \right] = \begin{cases} 
  \mathbb{E}\left[ g_p(\hat{z}^n_l) - g_{p+1}(\hat{z}^n_{l-1}) \right] & p < l \\
  0 & p \geq l 
\end{cases}
= 0.
\]

**Lemma B.1.** For a fixed \( p \in \{0, 1, \ldots\} \), we have
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{l=1}^{N} u^p_l = 0 \quad \mathbb{P}\text{-almost surely.}
\]
Proof: First, note that

\[ E[u_l^p|u_{l-1}^p, \ldots, u_0^p] = E\left[ E\left[u_{l-1}^p|\hat{z}_{l-1}^n\right]|u_{l-1}^p, \ldots, u_0^p\right] = 0, \quad E[|u_l^p|^2] \leq 2\|f\|_\infty^2. \]

The proof then is an immediate consequence of Strong Law of Large numbers for martingales in [31, p. 66]. □

We immediately conclude that for any \( l, k \in \mathbb{N} \) such that \( l \geq k \), we have

\[ g_0(\hat{z}_n^l) - g_k(\hat{z}_{l-k}^n) = u_l^0 + \ldots + u_{k-1}^l = \sum_{p=0}^{k-1} u_p^l. \]

Let \( g_p(\hat{z}_n^l) := 0 \). Fix \( k \in \mathbb{N} \). This yields for any \( k \leq l \) and \( N \gg k \)

\[ \frac{1}{N} \sum_{l=0}^{N-1} g_0(\hat{z}_l^n) - \frac{1}{N} \sum_{l=k}^{N-1} g_k(\hat{z}_{l-k}^n) = \frac{1}{N} \sum_{l=0}^{N-1} \sum_{p=0}^{k-1} u_l^p = \sum_{p=0}^{k-1} \left( \frac{1}{N} \sum_{l=0}^{N-1} u_l^p \right). \]

In the above expression, we can replace \( \frac{1}{N} \sum_{l=k}^{N-1} g_k(\hat{z}_{l-k}^n) \) with

\[ \frac{1}{N} \sum_{l=k}^{N-1} g_k(\hat{z}_{l-k}^n) = \frac{1}{N} \sum_{l=0}^{N-1} g_k(\hat{z}_l^n) - \frac{1}{N} \sum_{l=N-k+1}^{N-1} g_k(\hat{z}_l^n). \]

In what follows, we ignore the second summand in the right side above since \( N \gg k \). Next, we add the terms from \( k = 1 \) to \( M \), where \( N \gg M \) to get

\[ \frac{1}{M} \sum_{k=1}^{M} \frac{1}{N} \sum_{l=0}^{N-1} \left( g_0(\hat{z}_l^n) - g_k(\hat{z}_l^n) \right) = \frac{1}{M} \sum_{k=1}^{M} \sum_{p=0}^{k-1} \left( \frac{1}{N} \sum_{l=0}^{N-1} u_l^p \right). \]

Rearranging the terms in the left side yields

\[ \frac{1}{N} \sum_{l=0}^{N-1} \left( g_0(\hat{z}_l^n) - \bar{g}_M(\hat{z}_l^n) \right) = \frac{1}{M} \sum_{k=1}^{M} \sum_{p=0}^{k-1} \left( \frac{1}{N} \sum_{l=0}^{N-1} u_l^p \right). \]

Now note that as \( N \to \infty \), the term in the brackets converge to 0 almost surely by Lemma B.1.

REFERENCES

[1] O. Anschel, N. Baram, and N. Shimkin, Averaged-DQN: Variance reduction and stabilization for deep reinforcement learning, in Proceedings of the 34th International Conference on Machine Learning—Volume 70, JMLR. org, 2017, pp. 176–185.
[2] M. Benaim, J. Hofbauer, and S. Sorin, Stochastic approximations and differential inclusions, SIAM Journal on Control and Optimization, 44 (2005), pp. 328–348.
[3] D. P. Bertsekas, Abstract dynamic programming, Athena Scientific Belmont, MA, 2013.
[4] D. P. Bertsekas and J. N. Tsitsiklis, Neuro-Dynamic Programming, Athena Scientific, 1996.
[5] R. N. Bhattacharya and O. Lee, *Ergodicity and central limit theorems for a class of Markov processes*, in Multivariate Statistics and Probability, Elsevier, 1989, pp. 80–90.

[6] S. Bonnabel, *Stochastic gradient descent on Riemannian manifolds*, IEEE Transactions on Automatic Control, 58 (2013), pp. 2217–2229.

[7] V. S. Borkar, *Stochastic approximation: A dynamical systems viewpoint*, vol. 48, Springer, 2009.

[8] A.-A. Borovkov and V. Yurinsky, *Ergodicity and Stability of Stochastic Processes*, John Wiley & Sons, Ltd., 1998.

[9] L. Bottou, *Large-scale machine learning with stochastic gradient descent*, in Proceedings of COMPSTAT’2010, Springer, 2010, pp. 177–186.

[10] L. Breiman, *The strong law of large numbers for a class of Markov chains*, The Annals of Mathematical Statistics, 31 (1960), pp. 801–803.

[11] W. L. Cooper, S. G. Henderson, and M. E. Lewis, *Convergence of simulation-based policy iteration*, Probability in the Engineering and Informational Sciences, 17 (2003), pp. 213–234.

[12] W. L. Cooper and B. Rangarajan, *Performance guarantees for empirical Markov decision processes with applications to multiperiod inventory models*, Operations Research, 60 (2012), pp. 1267–1281.

[13] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Advances in neural information processing systems, 2014, pp. 1646–1654.

[14] P. Diaconis and D. Freedman, *Iterated random functions*, SIAM review, 41 (1999), pp. 45–76.

[15] S. B. Gelfand and S. K. Mitter, *Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$*, SIAM Journal on Control and Optimization, 29 (1991), pp. 999–1018.

[16] S. B. Gelfand and S. K. Mitter, *Metropolis-type annealing algorithms for global optimization in $\mathbb{R}^d$*, SIAM Journal on Control and Optimization, 31 (1993), pp. 111–131.

[17] A. Gupta, R. Jain, and P. Glynn, *Probabilistic contraction analysis of iterated random operators*, submitted to Annals of Applied Probability, (2018). arXiv preprint arXiv:1804.01195.

[18] A. Gupta, R. Jain, and P. W. Glynn, *An empirical algorithm for relative value iteration for average-cost MDPs*, in 54th Annual Conference on Decision and Control (CDC), IEEE, 2015, pp. 5079–5084.

[19] R. Harikandeh, M. O. Ahmed, A. Virani, M. Schmidt, J. Konečný, and S. Sallinen, *Stopwasting my gradients: Practical SVRG*, in Advances in Neural Information Processing Systems, 2015, pp. 2251–2259.

[20] W. B. Haskell, R. Jain, and D. Kalathil, *Empirical dynamic programming*, Mathematics of Operations Research, 41 (2016), pp. 402–429.

[21] O. Hernández-Lerma and J. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*, vol. 30, Springer Science & Business Media, 2012.

[22] O. Hernández-Lerma and J. B. Lasserre, *Further topics on discrete-time Markov control processes*, vol. 42, Springer Science & Business Media, 2012.

[23] J. Hofbauer and W. H. Sandholm, *On the global convergence of stochastic fictitious play*, Econometrica, 70 (2002), pp. 2265–2294.

[24] T. Jaakkola, M. I. Jordan, and S. P. Singh, *Convergence of stochastic iterative dynamic programming algorithms*, in Advances in neural information processing systems, 1994, pp. 703–710.

[25] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford, *Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification*, Journal of Machine Learning Research, 18 (2017), pp. 1–42.

[26] R. Jain and P. Varaiya, *Simulation-based optimization of Markov decision processes: An empirical process theory approach*, Automatica, 46 (2010), pp. 1297–1304.

[27] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in neural information processing systems, 2013, pp. 315–323.

[28] A. F. Karr, *Weak convergence of a sequence of Markov chains*, Probability Theory and Related Fields, 33 (1975), pp. 41–48.

[29] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, vol. 35, Springer Science & Business Media, 2003.

[30] D. S. Leslie, E. Collins, et al., *Convergent multiple-timescales reinforcement learning algorithms in normal form games*, The Annals of Applied Probability, 13 (2003), pp. 1231–1251.

[31] M. Loève, *Probability Theory II*, Springer, New York, NY, 1977.
[32] S. Mandt, M. D. Hoffman, and D. M. Blei, *Stochastic gradient descent as approximate bayesian inference*, The Journal of Machine Learning Research, 18 (2017), pp. 4873–4907.

[33] S. Merity, N. S. Keskar, and R. Socher, *Regularizing and optimizing lstm language models*, arXiv preprint arXiv:1708.02182, (2017).

[34] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Cambridge University Press, 2009.

[35] K. R. Parthasarathy, *Probability measures on metric spaces*, vol. 352, American Mathematical Soc., 2005.

[36] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, Mathematical Programming, 162 (2017), pp. 83–112.

[37] J. S. Shamma and G. Arslan, *Dynamic fictitious play, dynamic gradient play, and distributed convergence to nash equilibria*, IEEE Transactions on Automatic Control, 50 (2005), pp. 312–327.

[38] N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan, *Averaging stochastic gradient descent on riemannian manifolds*, arXiv preprint arXiv:1802.09128, (2018).

[39] J. Tsitsiklis, D. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE transactions on automatic control, 31 (1986), pp. 803–812.

[40] J. N. Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Machine learning, 16 (1994), pp. 185–202.

[41] H.-T. Wai, Z. Yang, P. Z. Wang, and M. Hong, *Multi-agent reinforcement learning via double averaging primal-dual optimization*, in Advances in Neural Information Processing Systems, 2018, pp. 9649–9660.

[42] W. Whitt, *Approximations of dynamic programs, I*, Mathematics of Operations Research, 3 (1978), pp. 231–243.

[43] W. Whitt, *Approximations of dynamic programs, II*, Mathematics of Operations Research, 4 (1979), pp. 179–185.