Some Limit Properties of Markov Chains Induced by Recursive Stochastic Algorithms[*]

Abhishek Gupta[†], Hao Chen[†], Jianzong Pi[†], and Gaurav Tendolkar[‡]

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 approximate certain contraction operators and can be viewed within the framework of iterated random operators. Accordingly, we consider iterated random operators over a Polish space that simulate an iterated contraction operator over that Polish space. Assume that the iterated random operators are indexed by certain batch sizes such that as batch sizes grow to infinity, each realization of the random operator converges (in some sense) to the contraction operator it is simulating. We show that starting from the same initial condition, the distribution of the random sequence generated by the iterated random operators converges weakly to the trajectory generated by the contraction operator. We further show that under certain conditions, the time average of the random sequence converges to the spatial mean of the invariant distribution. We then apply these results to logistic regression, empirical value iteration, and empirical Q-value iteration for finite state finite action MDPs to illustrate the general theory developed here.

Key words. stochastic gradient descent, empirical dynamic programming, constant stepsize Q learning, iterative random maps, Feller Markov chains

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

DOI. 10.1137/19M1258104

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 and generative models have 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 a reasonably good solution.

It turns out that the outputs of these recursive stochastic algorithms (RSAs) can be viewed as Markov chains. Indeed, if the parameters of the algorithm do not change with iteration,

[†]Electrical and Computer Engineering Department, The Ohio State University, Columbus, OH 43210 USA (gupta.706@osu.edu, chen.6945@osu.edu, pi.35@osu.edu).
[‡]Microsoft Corp., Sunnyvale, CA 94085 USA (gatendol@microsoft.com).

then the RSAs can be thought of as an iterated random operator acting onto certain spaces. Consider, for instance, the case of stochastic gradient descent, where the stepsize remains constant, data samples picked at every iteration are independent and identically distributed (i.i.d.), and the number of data samples remain constant. Then, each step of the stochastic gradient descent algorithm can be viewed as a random operator. To see this, let 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 stepsize $\beta > 0$ 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. We note here that under reasonable assumptions on L and β, T becomes a contraction operator under some norm on $\mathbb{R}^n$ (usually the ℓ_2 norm is used).

In contrast, in stochastic gradient descent, the operator 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 operator $\hat{T}_k^n$ 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 operator T and stochastic gradient descent operator $\hat{T}_k^n$ are related—$\hat{T}_k^n(x)$ is a consistent estimate of the $T(x)$ for any $x \in \mathbb{R}^n$.

A similar setup is considered in empirical dynamic programming using a generative model for the dynamic decision process. Consider a controlled Markov 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 and $c(s, a)$ be the corresponding cost. We use α to denote the discount parameter. In the value iteration algorithm, one needs to compute $\mathbb{E}[v(s')|s, a]$, where v is some real-valued function of the state. This leads to the usual Bellman operator T that acts on the space of value functions and is given by

$$T(v)(s, a) = \min_a \left[c(s, a) + \alpha \mathbb{E}\left[v(s')|s, a \right] \right].$$

It is not difficult to see that T is a contraction operator with the contraction coefficient α, when the space of value functions is endowed with the sup norm. If there is enough data, one can replace $\mathbb{E}[v(s')|s, a]$ with its "empirical" average, given by $\frac{1}{n} \sum_{i=1}^{n} v(s_i'(s, a))$, where

$\{s_i'(s, a)\}_{i=1}^n$ are n samples of the next state given that the current state-action pair is (s, a). Thus, the random Bellman operator $\hat{T}_k^n$ acts on the space of value functions and is given by

$$\hat{T}_k^n(v)(s, a) = \min_a \left[c(s, a) + \alpha \frac{1}{n} \sum_{i=1}^n v(s_i'(s, a)) \right].$$

In this case as well, $\hat{T}_k^n(v)$ is a consistent estimate of the $T(v)$ for any value function v. When the state or action space is continuous, then the above operator also features an additional function fitting component. The analysis of such algorithms involves understanding the error introduced due to function fitting, as well as the number of samples used at every iteration.

In fact, in both of the examples above, it is readily observed that the original operator T is a contraction map. This is no accident—the contraction mapping theorem forms the bedrock of most of the proofs of convergence algorithms used in optimization or Markov decision Process (MDP) problems. Some examples are noted below:

1. The Bellman operator for a class of stochastic shortest path problems is a contraction operator under an appropriate weighted sup norm [10, 8].
2. The Bellman operator for a class of average cost MDPs is a contraction operator under the span seminorm over the quotient space [65, 3].
3. Some variational inequality problems involve contraction operators under the usual 2 norm [31, p. 143].
4. The Bellman operator for continuous state MDP is a contraction operator over the space of continuous and bounded functions over the state space of the MDP (it requires a variety of assumptions as elucidated in [42, 43, 30]).
5. The resolvent of a strongly monotone operator is a contraction operator [67]. Many other contraction operators used in the context of optimization algorithms are discussed in section 5 of [67].

In data-driven applications, computing the exact (contraction) operator is either computationally expensive or impossible. Thus, one has to use random mappings that are drawn independent of the past operators and that simulate the effect of contraction mapping. The two examples explained above are merely instances of this more general methodology. The primary purpose of the paper is to devise sufficient conditions on the relationship between random maps $\hat{T}_k^n$ and the deterministic contraction operator T so that (a) the convergence properties of new algorithms can be readily established and (b) one can find common threads between the convergence.

1.1. Our contribution. The primary contribution of this paper is to conceptually unify the convergence analysis of certain RSAs in optimization, machine learning, and reinforcement learning using the tools from the Markov chain theory. This is achieved by leveraging several results available for convergence and stability for Feller Markov chains established in [21], [58, sect. 18.5], [18, sect. 8], and [51]. Our key contributions are summarized below.

1. Consider the deterministic sequence $y_k = T \circ \cdots \circ T(y_0)$ (k compositions of the exact operator T) and the Markov chain $\hat{z}_k^n = \hat{T}_{k-1}^n \circ \cdots \circ \hat{T}_0^n(\hat{z}_0^n)$ with $\hat{z}_0^n = y_0$. It is natural to assume that using the deterministic operator yields the best convergence guarantee. We use independent samples of data to approximate T by $\hat{T}_k^n$; then it introduces error

at every time step. How does this error accumulate? Most authors bound $\|\hat{z}_k^n - x^*\|$, where x^* is the fixed point of the operator T. Instead, we are interested in the error $\|\hat{z}_k^n - y_k\|$, noting that the triangle inequality yields

$$(1.1) \qquad \|\hat{z}_k^n - x^*\| \le \|\hat{z}_k^n - y_k\| + \|y_k - x^*\| \le \|\hat{z}_k^n - y_k\| + \alpha^k \|x_0 - x^*\|.$$

We derive a sufficient condition on the random operators $\hat{T}_k^n$ and its relationship with the exact operator T so that the sequence of distributions over the sequence $(\hat{z}_k^n)_{k\in\mathbb{N}}$ output by the RSA 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 $n \to \infty$. This further implies that $\hat{z}_k^n$ is close to y_k with high probability. Using the inequality in (1.1) above, we conclude that the error term $\|\hat{z}_k^n - x^*\|$ is less than $\alpha^k \|x_0 - x^*\|$ plus a small error with high probability.

We further show that the sufficient condition is satisfied in a sufficiently general class of problems encountered in stochastic gradient descent for strongly convex loss functions and synchronous empirical dynamic programming for MDPs with discounted cost criterion. These examples serve to illustrate how to apply the results to derive this property of an RSA.

2. Existence of invariant distributions of RSAs is an important property, as it implies some form of stability of the algorithm. A Markov chain does not admit invariant distribution if it features a cyclic behavior or diverges to infinity (there could be other reasons for the nonexistence of an invariant distribution, but these two are more common). In an RSA, we do not usually expect a cyclic behavior due to randomization. Thus, if an RSA admits an invariant distribution, then it implies that the iterates will not diverge with high probability. Thus, establishing the existence of an invariant distribution is an important problem, which we address here.

We show that the Markov chains generated by many RSAs satisfy the weak Feller property, that is, if f is a continuous and bounded function, then $\hat{z}_0^n \mapsto \mathbb{E}\left[f(\hat{z}_1^n)|\hat{z}_0^n\right]$ is also a continuous and bounded function. The existence (and in some case, the uniqueness) of an invariant measure of Feller Markov chains has been presented in [18]. 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. There has been a sustained interest in using time averages in stochastic gradient descent and deep Q learning. Particularly, references [69, 57, 47, 56, 80] propose that fixing the stepsize in stochastic gradient descent algorithms and using the average of the tail of the random sequence leads to a better performance of the trained algorithm. Within the context of reinforcement learning, [4] and [84] propose 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 the Markov chain output by RSA may be admitting a unique invariant distribution. This part leverages the law of large numbers for Markov chains, presented in [21, 58].

4. We complement the theoretical contributions with numerical simulations of two RSAs—stochastic gradient descent for logistic regression, empirical value iteration for discounted MDP with a generative model, and synchronous batch Q-value iteration for discounted MDP.

While we present complete proofs of two of our main results stated here, we admit that our proofs require minor tweaks of existing results in the literature. The need for presenting the complete proofs are twofold: Our hypotheses differ in some ways from the hypotheses presented in the standard texts, particularly in [51], [58, sect. 18.5], and [21]. Moreover, to construct the complete proofs using these texts under our hypotheses requires 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 generalizable to other settings. For instance, the convergence of stochastic gradient descent, stochastic variance reduction gradient descent, and stochastic average gradient descent follows a completely different, and often involved, sequence of arguments [20, 15, 49, 38, 69, 26]. The argument usually starts with identifying some conditions on the functions, such that for every iteration k, one can upper bound $L(\hat{z}_k^n) - L(x^*)$ (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 asymptotic convergence to the fixed point of the map. The stochastic approximation theory is one such elegant theory [17, 45, 52, 16]. There are two types of stochastic approximation algorithms—one with decreasing (also called tapering or diminishing) stepsize and the other with constant stepsize. In decreasing stepsize algorithms, the stepsize has to converge to 0 as the number of iteration goes to infinity (the stepsize is not summable but is square summable). This leads to the almost sure convergence guarantee to the fixed point in the limit. For constant stepsize, the theory says that the iterates will eventually enter a neighborhood of the fixed point and do a random walk within that neighborhood.

We now present a sample of decreasing stepsize RSAs whose convergence is ascertained using the stochastic approximation theory. Under reasonable assumptions on the loss function, stochastic gradient descent and distributed asynchronous gradient descent methods converge almost surely to the optimal solution [81, 33, 34]. The convergence of reinforcement learning algorithms usually invokes some version of the stochastic approximation theorem. References [46, 82, 1, 11] study the convergence of various types of Q learning algorithms developed for discounted cost or average cost MDPs with finite state and action spaces. The convergence of the on-policy reinforcement learning algorithm SARSA is established in [76]. More recently, the stochastic approximation theory has been used to establish the convergence of policy gradient, temporal difference, and other related methods in [78, 92, 96, 93, 90, 91]. For more information on various reinforcement learning algorithms and their convergence proofs, we refer the reader to books [3, 79, 10] and recent survey papers [94, 95].

Decreasing stepsize RSAs do not yield approximate solutions in a reasonable time frame. As a result, constant stepsize algorithms are gaining traction as a way to speed up the

computation at the cost of tolerating a small error in the final result; see, for example, [32, 60, 5, 29, 28], where constant stepsizes are used in the context of the stochastic gradient descent–type algorithms and [7, 77, 61, 70] for their usage in the reinforcement learning algorithms.

Constant stepsize stochastic approximation over finite dimensional state space has been studied in [17, 45], where the authors derive the asymptotic concentration results. It is well-understood that 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 [28]. 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. Similarly, in the Q learning algorithm, constant stepsize Q learning has been studied in [7] (both synchronous and asynchronous version are studied). Convergence of constant stepsize temporal difference methods with linear function approximation is studied in [53, 25, 12, 37, 77]. Empirical value iteration and its variants, studied in a number of works under varying assumptions [88, 89, 23, 59, 24, 48, 50, 39, 75], are also examples of constant stepsize stochastic approximation algorithms.

When the statistical properties of the i.i.d. operators used in an RSA do not change with time (e.g., constant stepsize stochastic approximation algorithms), then the output of the RSA forms a Markov chain. The goal of this paper is to study the limit properties of such a Markov chain. Since we use operator-theoretic methods, our conclusions are not algorithm-dependent. There are two ways the limit can be taken. Either the sample size n used at every time step can grow to infinity or the number of iterations k can escape to infinity. We study both limits in this paper. We note here that the generality of the model and minimal assumptions do not allow us to derive a finite time guarantee, which has significant importance in the machine learning and the reinforcement learning communities. Further, our proof approach is not algorithm-dependent. We leave these important problems for a future work.

Our work is largely motivated by the analysis of empirical dynamic programming in [39]. 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 [36]. We further relaxed some conditions on random operators assumed in [39] in our follow-up work [35]. The aim of this paper is to further expand the analysis and present conditions on random operators and their relationship to the exact operator to arrive at insightful conclusions about the random sequences generated by these RSAs.

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 RSAs. 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 RSA converge 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 $\hat{T}_k^n$ 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 and establishes the weak 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. Notation. Let $(\mathcal{A}, \rho)$ be a Polish space, which is defined as a complete separable metric space with metric ρ. We use $\wp(\mathcal{A})$ to denote the set of all probability distributions over $\mathcal{A}$. We use $\mathbb{1}_{\{a\}} \in \wp(\mathcal{A})$ to denote the Dirac mass over $a \in \mathcal{A}$. The notation $C_b(\mathcal{A})$ and $U_b(\mathcal{A})$ denote, respectively, the set of all continuous and bounded functions and uniformly continuous and bounded functions over the set $\mathcal{A}$. We use $C(\mathcal{A})$ to denote the set of (possibly unbounded) continuous functions over the set $\mathcal{A}$. We say that a sequence of measures $\{\mu_n\}_{n \in \mathbb{N}} \subset \wp(\mathcal{A})$ converges to μ in weak* sense iff for every $f \in C_b(\mathcal{A})$, $\int f d\mu_n \to \int f d\mu$ as $n \to \infty$. This is usually also referred to as weak convergence in the probability theory literature.

We use $1_{\{\cdot\}}$ to denote an indicator function, which takes the value of 1 if $\{\cdot\}$ is true and 0 otherwise. By a slight abuse of notation, we also use 1_F to be the indicator function over a measurable set $F \subset \mathcal{A}$.

2. Problem formulation. Let $(\mathcal{X}, \rho)$ be a Polish space with metric ρ. Consider a contraction operator $T : \mathcal{X} \to \mathcal{X}$ with contraction coefficient $\alpha \in (0, 1)$ and the unique fixed point denoted by x^*. This means

$$\rho(T(x_1), T(x_2)) \leq \alpha \rho(x_1, x_2) \text{ and } T(x^*) = x^*.$$

Starting from any initial point $y_0 \in \mathcal{X}$, define the iterates

$$(2.1) \qquad\qquad y_k = T(y_{k-1}) \quad \text{ for } k \in \mathbb{N}.$$

By the 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 RSA rigorously.

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

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

2.1. Key questions. Consider the stochastic process that starts from $\hat{z}_0^n = y_0$ and define $\hat{z}_k^n = \hat{T}_{k-1}^n(\hat{z}_{k-1}^n)$ for all $k \in \mathbb{N}$. Due to Assumption 2.1 and the fact that n does not change with time, the stochastic process $\{\hat{z}_k^n\}_{k \in \mathbb{N}}$ is a time-homogeneous Markov chain. One can view $\hat{z}_k^n$ as an $\mathcal{X}$-valued Markov chain with the Markov transition kernel given by

$$Q_n(B|x) = \mathbb{P}\left\{\hat{z}_k^n \in B | \hat{z}_{k-1}^n = x\right\}$$

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 RSA is close to the deterministic sequence generated by exact algorithm with high probability. Let us formulate the precise mathematical problem. We let $\mu_n \in \wp(\mathcal{X}^{\mathbb{N}})$ denote the joint distribution of the sequence $(\hat{z}_0^n, \hat{z}_1^n, \hat{z}_2^n, \ldots)$. Endow $\mathcal{X}^{\mathbb{N}}$ with the product topology so that it becomes a Polish space. Then, μ_n is defined by

$$(2.2) \qquad \mu_n(B_0 \times B_1 \times B_2 \times \ldots) = \int_{B_0 \times B_1 \times B_2 \times \ldots} \mathbb{1}_{\{y_0\}}(dx_0) Q_n(dx_1|x_0) Q_n(dx_2|x_1) \ldots,$$

where $B_0, B_1, B_2, \ldots$ are Borel sets in $\mathcal{X}$.

In a 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

$$(2.3) \qquad \psi(B_0 \times B_1 \times B_2 \times \ldots) = \int_{B_0 \times B_1 \times B_2 \times \ldots} \mathbb{1}_{\{y_0\}}(dx_0) \mathbb{1}_{\{y_1\}}(dx_1) \mathbb{1}_{\{y_2\}}(dx_2) \ldots.$$

This is a Dirac mass on the sequence $(y_0, y_1, \ldots)$. Our first result, stated in section 4, proves that under a mild assumption on the random operators $\hat{T}_k^n$, the sequence of measures μ_n converges in the weak* sense to ψ.

A similar setup was considered by Karr [51]. 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 [51] 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 $(\hat{z}_k^n)_{k \in \mathbb{N}}$, one of the key questions is the existence of an invariant distribution. An invariant probability distribution of the Markov chain $(\hat{z}_k^n)_{k \in \mathbb{N}}$ is a probability measure π^n such that for any Borel set $B \subset \mathcal{X}$, we have

$$\pi^n(B) = \int_{\mathcal{X}} Q_n(B|x) \pi^n(dx).$$

A desirable property of a Markov chain is to have an invariant distribution, since it implies that the RSA satisfies a form of stability. More importantly, it implies that the Markov chain will not escape to infinity with probability 1. There is a large body of literature that studies the problem of the existence of invariant measures for Harris recurrent Markov chains that take values in continuous state spaces [18, 58]. However, the Markov chain generated by the RSAs seldom satisfies 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 RSAs satisfy certain desirable properties, which can be leveraged to not only guarantee the existence of an invariant distribution but also establish the uniqueness of the invariant distribution. These properties of the random operators are discussed in section 5. This further leads to strong conclusions about the weak law of large numbers, as we discuss next.

2.1.3. Convergence of time average of iterates. The weak law of large numbers for 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 the weak law of large numbers is also available for Feller Markov chains. This is established for Feller chains in [21] for chains residing in a compact Hausdorff space with a unique invariant measure and in [58, section 18.5] for the noncompact 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 a certain specific distribution (in fact, we do not need other technical conditions of [58, sect. 18.5]). We prove this result in section 6, the proof of which is adapted from the results from [21] and [58].

We illustrate the theoretical results using numerical simulations for batch gradient descent, empirical value iteration, and synchronous batch Q learning in section 7.

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

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

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

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

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

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

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

Consequently, each L_i is a convex function, and thus L is a convex 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 with positive eigenvalues. Consequently, L is strongly convex and therefore 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 β:

$$(3.1) \qquad x_{k+1} = T(x_k) := x_k - \beta \nabla_x L(x_k) = x_k - \frac{\beta}{N} \sum_{i=1}^N \left(f\left(u^i; x_k\right) - l^i\right) u^i,$$

where $T : \mathcal{X} \to \mathcal{X}$ is the gradient descent map (dependent on the parameter β). It can be further shown that if β is sufficiently small, then the operator T is a contraction on $\mathcal{X}$, endowed with the Euclidean norm. We note here that the above arguments would be true if $\{L_i\}_{i=1}^N$ is a collection of strongly convex and smooth loss functions.

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 the computational burden, a minibatch stochastic gradient descent is employed. In the minibatch Stochastic gradient descent, at every step k, the gradient is approximated by a small, randomly sampled, subset (of size n) of the dataset. 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

$$(3.2) \qquad \hat{z}_{k+1}^n = \hat{T}_k^n(\hat{z}_k^n) = \hat{z}_k^n - \beta \frac{1}{n} \sum_{j=1}^{n} \nabla_x(L_j(\hat{z}_k^n)).$$

Note that $\hat{T}_k^n$ is now a random operator.

Remark 3.1. Any realization of this random operator with small values of n need not be a contraction since $\nabla_x L_j$ is a rank 1 positive semidefinite matrix. One could add a regularizer to the loss function to make it strongly convex. In particular, if the loss function is chosen as

$$L_i(x) := -l^i \log f\left(u^i; x\right) - (1 - l^i) \log\left(1 - f\left(u^i; x\right)\right) + \frac{\lambda}{2} \|x\|_2^2 \quad \text{with } \lambda > 0,$$

then $\hat{T}_k^n$ is a contraction operator for a sufficiently small β irrespective of the n used.

Some obvious properties of these random operators are as follows:
1. $\hat{T}_k^n$ is continuous in $\hat{z}_k^n$.
2. For every $\epsilon > 0$ and $x \in \mathcal{X}$, we have

$$\lim_{n \to \infty} \mathbb{P}\left\{\|\hat{T}_k^n(x) - T(x)\|_2 > \epsilon\right\} = 0.$$

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 $\epsilon > 0$,

$$(3.3) \qquad \lim_{n \to \infty} \sup_{x \in \mathcal{K}} \mathbb{P}\left\{\|\hat{T}_k^n(x) - T(x)\|_2 > \epsilon\right\} = 0.$$

This statement can readily be proved using the Hoeffding inequality and the 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 the Hessian of $L_i(x)$ satisfy $0 < m \leq \lambda(\nabla_{xx}^2 L_i(x)) \leq M < \infty$ 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 $\hat{\alpha}_k^n \equiv \max\{|1 - \beta M|, |1 - \beta m|\} < 1$ for an appropriately picked $\beta > 0$.

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

3.2. Empirical value iteration for discounted cost MDP. Consider an MDP problem described by 4 tuple $(\mathcal{S}, \mathcal{A}, c, p)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, and $c : \mathcal{S} \times \mathcal{A} \longrightarrow \mathcal{R}$ is the cost function. The state transitions according to $s_{t+1} \sim p(\cdot|s_t, a_t)$. Let Γ denote the set consisting of all possible deterministic policies $\gamma : \mathcal{S} \longrightarrow \mathcal{A}$. The infinite horizon discounted cost $v^{\gamma} : \mathcal{S} \longrightarrow \mathcal{R}$ starting from state s and following policy γ is given by

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

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

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

$$(3.4) \qquad T(v)(s) = \inf_{a \in A}\left\{c(s, a) + \alpha \sum_{s' \in \mathcal{S}} v(s')p(s'|s, a)\right\}.$$

Due to the 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 the value iteration algorithm:

$$(3.5) \qquad \text{Initialize } v_0 \text{ randomly and iterate } v_{k+1}(s) = T(v_k)(s).$$

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 (3.4) to the empirical average. This algorithm is referred to as empirical dynamic programming and is written as

$$(3.6) \qquad \hat{T}_k^n(\hat{v}_k^n)(s) = \inf_{a \in \mathcal{A}}\left\{c(s, a) + \alpha \frac{1}{n}\sum_{i=1}^{n} \hat{v}_k^n\left(s_{k,i}'(s, a)\right)\right\},$$

where $s_{k,i}'(s, a)$ are n i.i.d. samples of the next state given the current state-action pair $(s.a)$, redrawn at every k independently from the past realizations. The above intuition can be turned into an algorithm to determine an approximately optimal value function and is known as the empirical value iteration algorithm:

$$(3.7) \qquad \text{Initialize } v_0 \text{ randomly and let } \hat{v}_{k+1}^n(s) = \hat{T}_k^n(\hat{v}_k^n)(s).$$

Note that $\hat{T}_k^n$ is a random operator, and its realization is dependent on the samples generated $(s_{k,i}'(s, a))_{i=1}^n$. The following properties of $\hat{T}_k^n$ are obvious:

1. $\hat{T}_k^n$ is a contraction with contraction coefficient α. 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 $\mathcal{K} \subset \mathcal{V}$, we have for every $\epsilon > 0$

$$\sup_{v \in \mathcal{K}} \mathbb{P}\left\{ \|\hat{T}_k^n(v) - T(v)\|_\infty > \epsilon \right\} \leq 2|\mathcal{S}||\mathcal{A}| \exp\left(-\frac{\epsilon n}{|\mathcal{S}|k^2} \right),$$

where $k = \max_{v \in \mathcal{K}} \|v\|_\infty$. This immediately yields

$$(3.8) \qquad \lim_{n \to \infty} \sup_{v \in \mathcal{K}} \mathbb{P}\left\{ \|\hat{T}_k^n(v) - T(v)\|_\infty > \epsilon \right\} = 0.$$

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 $v_l \uparrow v$, then $\hat{T}_k^n(v_l) \to \hat{T}_k^n(v)$ as $l \to \infty$.

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 (3.3) in the context of logistic regression is (mathematically) the same as the property stated in (3.8) 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 $\hat{T}_k^n$ satisfies certain monotonicity properties. 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. However, we will show that the realizations of the random operators in these cases have some other desirable properties that lead to the existence and uniqueness of the invariant measure.

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 μ_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 operator T in some sense. Accordingly, we make the following assumption.

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

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

We recall here that this assumption is satisfied by the logistic regression and empirical value iteration for discounted cost MDP (see (3.3) and (3.8) within the discussion at the ends 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* μ_n *converges in weak* topology to* ψ *as* $n \to \infty$, *where* ψ *is defined in* (2.3).

Proof. The proof is based on the proof by [51], except that our hypotheses are slightly different from those in [51]. 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 [62]. Generally speaking, if distributions of two random variables are close to each other in Levy–Prohorov metric, then it does not imply that the random variables will be close to each other. As an instance, the Levy–Prohorov metric between the measures of two i.i.d. random variables is zero, but the difference between the random variables themselves is not zero. If one of the random variables 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 below.

Let $\mathcal{A}$ be a Polish space with metric $\rho_{\mathcal{A}}$. Let d_P be the Levy–Prohorov metric on the space of probability measures $\wp(\mathcal{A})$ over $\mathcal{A}$. This metric is defined as follows. For a Borel set $A \subset \mathcal{A}$, let A^ϵ be defined as

$$A^\epsilon = \{a \in \mathcal{A} : \rho_{\mathcal{A}}(a, b) < \epsilon, b \in A\}.$$

Let $\mu, \nu \in \wp(\mathcal{A})$. Then, $d_P(\mu, \nu)$ is defined by

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

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 4.3. *Let* $\mu \in \wp(\mathcal{A})$ *and* $\mathbb{1}_{\{a^*\}}$ *be a unit mass at point* $a^* \in \mathcal{A}$. *If* $d_P(\mu, \mathbb{1}_{\{a^*\}}) < \epsilon$, *then for any random variable* W *distributed according to the law* μ, *we have*
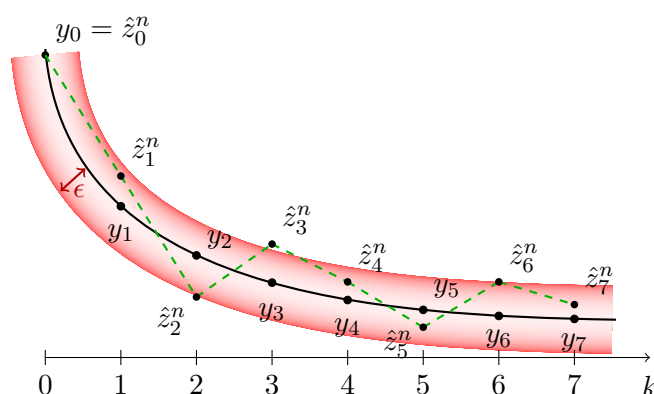
$$\mathbb{P}\left\{\rho_{\mathcal{A}}(W, a^*) \geq \epsilon\right\} < \epsilon.$$

Proof. Let B_ϵ be an open ϵ ball around a^*. Let $(B_\epsilon^\complement)^\epsilon$ be defined as

$$\left(B_\epsilon^\complement\right)^\epsilon = \left\{a \in \mathcal{A} : \rho_{\mathcal{A}}(a, b) < \epsilon, b \in B_\epsilon^\complement\right\}.$$

Note that $(B_\epsilon^\complement)^\epsilon = \mathcal{A} \setminus a^*$, which implies $\mathbb{1}_{\{a^*\}}((B_\epsilon^\complement)^\epsilon) = 0$. Let W be a random variable distributed according to the law μ. Then, from the definition of the Levy–Prohorov metric, we know that

$$\mu\left(B_\epsilon^\complement\right) < \mathbb{1}_{\{a^*\}}\left(\left(B_\epsilon^\complement\right)^\epsilon\right) + \epsilon = \epsilon.$$

Figure 1. *Illustration of the behavior of $\hat{z}_k^n$ with varying values of k. Under the hypotheses of Theorem 4.2, for every $\epsilon > 0$, there exists $N_\epsilon > 0$ such that $d_P(\mu_n, \psi) < \epsilon$ for every $n \geq N_\epsilon$. For most values of k, $\hat{z}_k^n$ stays within ϵ ball around y_k. In the illustration above, $\hat{z}_3^n$ is not within ϵ ball around y_3.*

The proof then follows from noting that

$$\mathbb{P}\{\rho_\mathcal{A}(W, a^*) \geq \epsilon\} = \mu\left(B_\epsilon^\complement\right).$$

The proof is established.   ∎

As a consequence of the lemma above, we conclude that since the distribution μ_n converges to the Dirac measure ψ, it implies that for n sufficiently large, the random sequence generated by RSA lies within a small tube around the trajectory induced by the deterministic contraction operator with high probability. This is illustrated in Figure 1.

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

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

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 4.3, we conclude that

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

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

5. Existence of invariant measures for fixed n. In this section, we identify conditions on the operators $(\hat{T}_k^n)_{k \in \mathbb{N}}$ so that the Markov chain $\hat{z}_k^n$ admits a unique invariant distribution, denoted by π^n, as $k \to \infty$. The existence of an invariant measure yields insight about "stability" of an RSA. In particular, if there is no invariant distribution, then it is likely the case that the sequence generated by RSA can blow up with positive probability. Thus, by tweaking the RSA (for instance, by changing the stepsize or increasing the number of samples), one can

ensure that the sufficient conditions noted below are satisfied, thereby establishing that the RSA is stable and yields finite values with probability 1.

In the 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 RSA will have its law as the 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 $\int f d\pi^n$ of the function with respect to the invariant measure π^n. 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 $\hat{T}_k^n$ and $\hat{T}_{k'}^n$ are independent and identical to each other as long as $k \neq k'$. Below, we list three assumptions under which we can show that $\hat{z}_k^n$ admits an invariant distribution.

To introduce our first assumption, we need to assume a partial order $\preceq$ exists on the normed space $\mathcal{X}$ (we will drop the completeness requirement on the space $\mathcal{X}$ for this assumption). For instance, $\mathcal{X}$ could be the Euclidean space with the natural partial order, wherein $x, y \in \mathbb{R}^n$, $x \preceq y$ implies $x_i \leq y_i$ for all $i \in \{1, \ldots, n\}$. The limit of an increasing sequence $(x_k)_{k \in \mathbb{N}}$ satisfying $x_1 \preceq x_2 \preceq x_3 \preceq \ldots$ under this partial order may escape to infinity. Thus, it is required to bound the space $\mathcal{X}$ to ensure that the sequences do not diverge along any of the coordinates. Another example of a normed space with partial order is the space of measurable functions from a Euclidean space $\mathbb{R}^n$ to $[-M, M]$, denoted by $\mathcal{L}_\infty(\mathbb{R}^n, [-M, M])$ (where M is a fixed positive real number). This space features a natural partial order, wherein $f_1 \preceq f_2$ iff $f_1(s) \leq f_2(s)$ for every $s \in \mathbb{R}^n$.

Assumption 5.1. The following conditions are satisfied:

1. $\mathcal{X}$ is a bounded normed space (not necessarily Polish) 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 $\hat{T}^n$ satisfies the following:
   (a) Monotonicity 1: $x_0 \in \mathcal{X}$ exists such that $x_0 \preceq \hat{T}^n(x_0)$ almost surely.
   (b) Monotonicity 2: If $x_1 \preceq x_2$, then $\hat{T}^n(x_1) \preceq \hat{T}^n(x_2)$ almost surely.
   (c) Continuity: If $x_k \uparrow x$, then $\hat{T}^n(x_k) \to \hat{T}^n(x)$ as $k \to \infty$ almost surely.

Assumption 5.1 is satisfied in Markov decision processes with nonnegative 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 and having a proper policy (that is, there is a stationary policy under which the MDP terminates in an absorbing state with probability 1). The proof of the last claim follows from arguments similar to the one made in subsection 3.2; see, for example, Chapter 1 of [10] and standard texts on MDPs with total cost criterion [42, 43, 9].

For RSAs involving MDPs, the point x_0 can be chosen easily. Assuming that the costs are nonnegative, one could take the zero function as the initial value function or Q function—and the Monotonicity 1 condition is satisfied almost surely.

Assumption 5.2. For $n \in \mathbb{N}$ and $m \in \mathbb{N}$, let $\hat{\alpha}_m$ denote the Lipschitz coefficient of $\hat{T}_m^n \circ \hat{T}_{m-1}^n \circ \cdots \circ \hat{T}_1^n$. 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 [36] to ensure that the value functions obtained through repeated use of empirical operators do not 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 $\hat{T}^n$ satisfies

$$\mathbb{P}\left\{\rho\left(\hat{T}^n(x^*), x^*\right) > \epsilon\right\} = \mathbb{P}\left\{\rho\left(\hat{T}^n(x^*), T(x^*)\right) > \epsilon\right\} \leq \frac{a}{\epsilon^b},$$

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 the empirical value iteration for an MDP with a discounted reward criterion, as we have noted in subsection 3.2.

Remark 5.4. Assumption 5.3.2 is typically proved using concentration of measures results, like the Hoeffding inequality [44, 54, 66] or the theory of empirical processes [63, 83]. Some references where such inequalities have been used in empirical dynamic programming for continuous state MDPs are [59, 40, 35].

Our next theorem summarizes the main result of this section. Let μ_k^n denote the distribution of $\hat{z}_k^n$.

Theorem 5.5. *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 π^n such that μ_k^n converges to π^n in weak* topology. Further, the invariant measure is unique under one of the following circumstances:*
1. *Assumption 5.1 holds, and the RSA is always initialized with x_0, where x_0 is defined in Assumption 5.1.*
2. *Assumption 5.2 holds.*
3. *Assumption 5.3 holds.*

Proof. Under Assumption 5.1, the existence of invariant measures is proved in [18, Theorem 8.1, pp. 79–81]. Under Assumption 5.2, the existence of invariant measures is proved in [13] and [18, Theorem 8.2, pp. 82–83]. Under Assumption 5.3, the existence and uniqueness result is established in [27, Theorem 1.1, p. 87]. ∎

Remark 5.6. 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 $\hat{z}_0^n = x_0$, then for any $\delta > 0$, there exists $N_\delta \in \mathbb{N}$ such that for any $k \geq 1$, we have

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

If the above condition and Assumption 5.2.2 holds, then one can show that for any initial condition $\hat{z}_0^n \in \mathcal{X}$, a unique invariant distribution exists. For a proof, we refer the reader to [18, p. 179]. However, proving (5.1) is satisfied in usual RSAs 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 RSA.

6. Averaging of iterates and the weak law of large numbers. RSAs with constant stepsize and averaging of iterates have received significant attention recently. Nemirovski et. al. in [60] develop an algorithm for stochastic gradient descent with averaging of the last few iterates and show that the algorithm is robust to stepsize selection and yields better results when properties of the loss functions (such as strong convexity parameter, Lipschitz coefficient of the derivative, etc.) are unknown. Along similar lines, [6, 5, 29] study stochastic gradient descent based algorithms with constant stepsize and averaging and derive the finite time guarantees on the loss achieved with averaged output. We note here that the constant stepsize algorithms with averaging are fundamentally different from Polyak–Ruppert averaging used in stochastic approximation [64, 52], where the stepsizes decrease at a rate slower than $1/k$.

Motivated by the above references, 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 $f \in C_b(\mathcal{X})$). As stated in subsection 1.1, averaging of the last few outputs of an RSA is used to reduce the variance in the final output of the algorithm. Specifically, if we terminate the iteration of RSA at k, and output $\hat{z}_k^n$, there is a small chance that this output is far from x^*. It is generally believed that if we average the last few outputs (assuming none of them are ∞), then it reduces the variance in the output. We establish this result rigorously here, under the assumption that the Markov chain has a unique invariant distribution.

We will assume throughout that the Markov chain $(\hat{z}_k^n)_{k \in \mathbb{N}}$ admits an invariant distribution (the precise assumption is stated in what follows). The time average of the Markov chain is precisely the law of large numbers for Markov chains. It has been studied within the context of Markov chains over compact spaces in [21] and over general locally compact spaces in [58, sect. 18.5]. Let us formulate the problem precisely.

Consider a continuous function $f \in C_b(\mathcal{X})$. Let π^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 5.5. In what follows, we show that under relatively mild assumptions, we have

$$(6.1) \qquad \frac{1}{K}\sum_{k=0}^{K-1} f(\hat{z}_k^n) \to \int f(x)\pi^n(dx) \text{ in probability as } K \to \infty.$$

Note that in the expression above, the term $\int f(x)\pi^n(dx)$ is the spatial average of the function f at the invariant distribution. Thus, what we show is that the time average converges in probability to the spatial average—which implies that $(f(\hat{z}_k^n))_{k\in\mathbb{N}}$ is a discrete time ergodic process.

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:

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

Assumption 6.1. The distribution μ of the initial condition $\hat{z}_0^n$ is picked from a set $\mathcal{M} \subset \wp(\mathcal{X})$, which satisfies $\mathbb{1}_{\{x\}} \in \mathcal{M}$ for all $x \in \mathcal{X}$. Either of the following two conditions holds:
1. There exists a unique invariant measure π^n such that for any $\mu \in \mathcal{M}$, $(F^*)^k(\mu)$ converges in weak* topology to π^n.
2. There exists a unique invariant measure π^n such that for any $\mu \in \mathcal{M}$, the averaged operator satisfies

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=0}^{k-1}(F^*)^k(\mu) = \pi^n,$$

where $(F^*)^k(\mu)$ denotes k compositions of F^* applied on μ and the convergence is in a weak* sense.

From the Stolz–Cesaro theorem [22, Theorem 2.7.1, p. 59], 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 Theorem 5.5.

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

Proof. The proof essentially follows the steps in [21] and [58, Theorem 18.5.1, p. 478], except that we relax the assumption on compactness of the state space as assumed in [21] and replace the hypotheses in [58] 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 $\mathcal{M} = \wp(\mathcal{X})$, that is, for every $\mu \in \wp(\mathcal{X})$, $(F^*)^k(\mu)$ converges in weak* topology to π^n. This is a very strong assumption from the applicability viewpoint in RSAs. In particular, it is possible to pick the most suitable initialization for the RSAs, 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 establish the existence of a unique invariant distribution using Theorem 5.5. Incidentally, for the law of large numbers to hold, we do not need the stronger condition of $\mathcal{M} = \wp(\mathcal{X})$.

We now have a corollary of the result above, which we capture as a theorem below due to its importance and applicability to RSAs.

Theorem 6.3. *Suppose that* $\mathcal{X} \subset \mathbb{R}^n$ *is a compact set. Suppose further that one of the following conditions holds:*

1. *Assumption* 5.1 *holds, and the RSA is always initialized with* x_0*, where* x_0 *is defined in Assumption* 5.1*.*
2. *Assumption* 5.2 *holds.*
3. *Assumption* 5.3 *holds.*

Consider the average $\hat{a}_k^n$ *of the Markov chain*

$$\hat{a}_k^n = \frac{1}{k} \sum_{t=0}^{k-1} \hat{z}_t^n.$$

If Assumption 2.1 *holds, then for any initialization* $\hat{z}_0^n \in \mathcal{X}$*, a unique invariant distribution* π^n *exists and* $\hat{a}_k^n$ *converges almost surely to the mean* $\bar{a}^n$ *of the distribution* π^n*.*

Proof. The existence of a unique invariant measure is due to Theorem 5.5. Since $\mathcal{X}$ is a compact set, we can take $f(x) = x_i$ to conclude that $(\hat{a}_k^n)_i$ converges almost surely to $\bar{a}_i^n$ by [21]. ∎

The result in (6.1) requires f to be a bounded function over $\mathcal{X}$. We now consider the case where f is a potentially unbounded continuous function. To establish essentially the same result, we need to make the following additional assumption.

Assumption 6.4. There exists a random variable $C(\omega)$ such that

$$\sup_{k \geq 0} f(\hat{z}_k^n) \leq C(\omega) \text{ almost surely,}$$

and $C(\omega) < \infty$ for $\mathbb{P}$-almost all $\omega \in \Omega$.

Theorem 6.5. *Suppose that* $f : \mathcal{X} \to \mathbb{R}$ *is a continuous function (potentially unbounded). If Assumptions* 2.1, 6.1, *and* 6.4 *hold, then* (6.1) *holds.*

Proof. The proof is presented in subsection 9.1. ∎

It is easy to see in practice if Assumption 6.4 holds or not, particularly when $\mathcal{X} = \mathbb{R}^n$. If the random iterates are uniformly bounded during a single run of the RSA, then Assumption 6.4 holds along that trajectory. If almost all independent runs of the RSA are not expected to "blow up," then the time average of the iterates is likely going to have the variance reduction property—the time average converges in probability to the spatial average.

We now turn our attention to illustrating the application of Theorems 4.2, 6.2, and 6.5 in various optimization and empirical dynamic programming algorithms.

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 the MNIST dataset, empirical value iteration for discounted and average cost MDPs, and empirical Q-value iteration.

7.1. Stochastic gradient descent. Consider the task of classifying a subset of MNIST handwritten digits, where we consider only the images corresponding to the numbers 0 and 1. Each data point is a 28×28 pixel image with the corresponding label (either 0 or 1). We use logistic regression and Poisson regression with an ℓ_2 regularizer for the classification task. We refer the reader to subsection 3.1 for details of this problem for the logistic regression. The loss function for the logistic regression with the regularizer is

$$L_i(x) := -l^i \log f(u^i; x) - (1 - l^i) \log(1 - f(u^i; x)) + \frac{\lambda}{2}\|x\|_2^2, \text{ where } \lambda = 5.$$

The loss function for Poisson regression with regularizer is

$$L(x) = \frac{1}{N} \sum_{i=1}^{N} L_i(x), \text{ where } L_i(x) = \exp\left(u^{i\mathsf{T}}x\right) - l^i\left(u^{i\mathsf{T}}x\right) + \frac{\lambda}{2}\|x\|_2^2, \text{ where } \lambda = 1.$$

The first and the second derivatives of this loss function are

$$\nabla_x L_i(x) = \exp\left(u^{i\mathsf{T}}x\right)u^i - l^i u^i + \lambda x, \qquad \nabla_{xx}^2 L_i(x) = \exp\left(u^{i\mathsf{T}}x\right)u^i u^{i\mathsf{T}} + \lambda I.$$

We use here the notation introduced in subsection 3.1. We transform each image into a vector and append 1 at the beginning of the vector. Thus, the space $\mathcal{U} = \{1\} \times [0,1]^{784}$. Thus, the space $\mathcal{X} = \mathbb{R}^{785}$. As mentioned previously, the variable n represents the batch size picked at every stochastic gradient descent 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 stochastic gradient descent as follows:

$$y_{k+1} = T(y_k), \quad \hat{z}_{k+1}^n = \hat{T}_k^n(\hat{z}_k^n), \quad \hat{a}_k^n = \frac{1}{k+1}\sum_{i=0}^{k} \hat{z}_i^n.$$

As discussed in subsection 3.1, it is clear that the exact gradient descent is a contraction map for stepsize β small enough. 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. If β is sufficiently small, then $\hat{T}_k^n$ also satisfies Assumption 5.3. Let μ_n denote the distribution of $(\hat{z}_0^n, \hat{z}_1^n, \ldots)$ and let ψ be the Dirac mass on $(y_0, y_1, \ldots)$. We have $\{\mu_n\}_{n \in \mathbb{N}}$ converges to ψ in weak* topology as $n \to \infty$ and there exists a unique invariant distribution of the Markov chain $(\hat{z}_k^n)_{k \in \mathbb{N}}$ for any $n \in \mathbb{N}$.*

Proof. The first conclusion follows from the discussion leading to (3.3) in subsection 3.1. For β sufficiently small, every realization of $\hat{T}_k^n$ is a contraction as discussed in Remark 3.1. We only need to show that part 2 of Assumption 5.3 is satisfied. Since n is finite, $\hat{T}_k^n(x^*)$ can take only finitely many values, and therefore part 2 of Assumption 5.3 is trivially satisfied. The existence of invariant distribution then follows from Theorem 5.5. ∎

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 Figures 2 and 4, where we plot the distance $\|y_k - \hat{z}_k^n\|_2$ for various values of n (batch sizes) and k ranging from 0 to 5000. The black curve is the mean $\mathbb{E}\left[\|y_k - \hat{z}_k^n\|_2\right]$ and the red region shows one standard deviation of the distance $\|y_k - \hat{z}_k^n\|_2$ over 1000 sample paths of the minibatch stochastic gradient descent iterations. We picked the same initial condition $\hat{z}_0^n = y_0 = 0$ for all the sample paths to generate the figure.

Figures 3 and 5 plot $\|\hat{a}_k^n - x^*\|_2$, and we observe that the variance of time averages reduces as k grows large for any batch size n. This variance reduction of time averages can be attributed to the contraction property of the random maps $\hat{T}_k^n$, which in turn is due to the addition of a strongly convex regularizer to the empirical loss function.
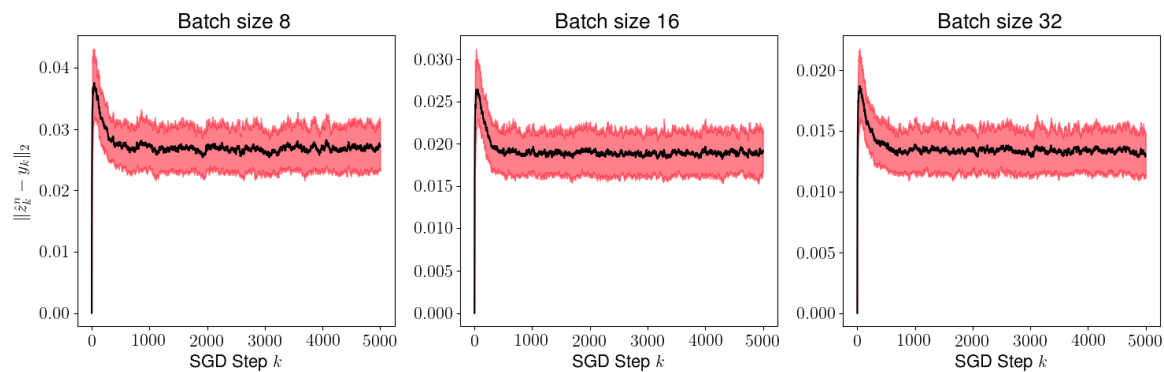
Figure 2. *The Euclidean norm between the outputs of the exact gradient descent and the stochastic gradient descent with $n = 8, 16, 32$ for the logistic regression with a regularizer. We observe that as n grows, the probability that $\|\hat{z}_k^n - y_k\|$ is large becomes smaller. The mean $\mathbb{E}\left[\|\hat{z}_k^n - y_k\|\right]$ (showed using a black line) and the variance of $\|\hat{z}_k^n - y_k\|$ are computed using 1000 independent runs of the iterations.*
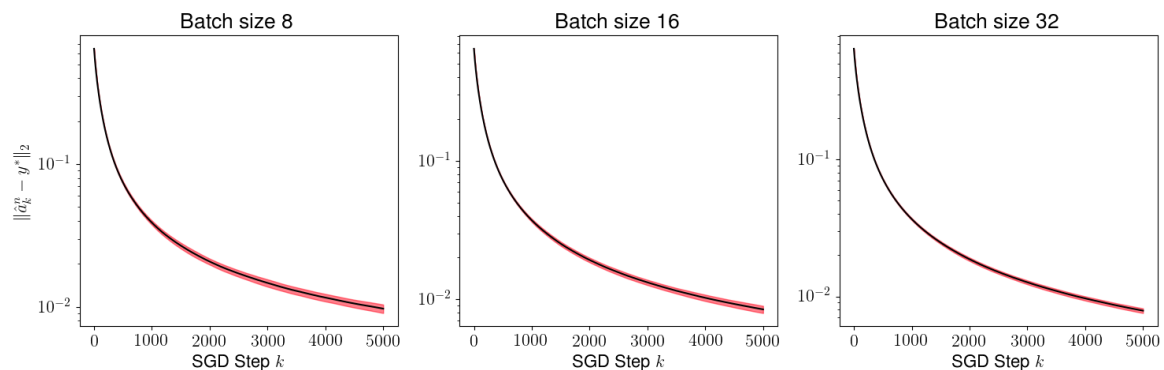
Figure 3. *Plot of $\|\hat{a}_k^n - x^*\|$ for $n = 8, 16, 32$ for the logistic regression with a regularizer. The variance of $\|\hat{a}_k^n - x^*\|$ is decreasing as k grows, which indicates that the variance reduction property of the time averages. The mean and the variance of $\|\hat{a}_k^n - x^*\|$ are computed using 1000 independent runs of the stochastic gradient descent iterations.*

7.2. Empirical value iteration for discounted cost MDP. We consider here the empirical value iteration for discounted MDP as described in subsection 3.2. Consider the 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 $\hat{z}_0^n = v_0$ and define the iterates of exact value iteration and empirical value iteration as

$$v_{k+1} = T(v_k), \qquad \hat{z}_{k+1}^n = \hat{T}_k^n(\hat{z}_k^n), \quad \hat{a}_k^n = \frac{1}{k+1}\sum_{i=0}^{k} \hat{z}_i^n.$$

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 as follows:*
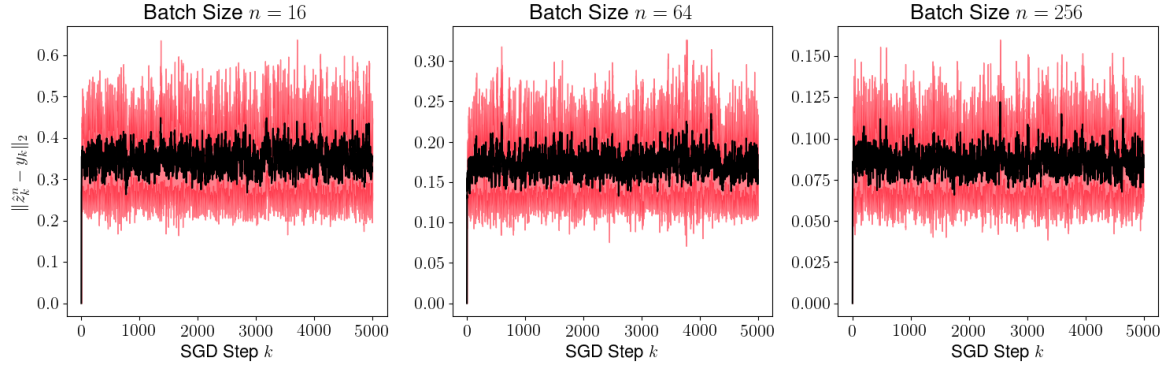
Figure 4. *The Euclidean norm between the outputs of the exact gradient descent and the stochastic gradient descent with $n = 64, 256, 1024$ for the Poisson regression with a regularizer. We observe that as n grows, the probability that $\|\hat{z}_k^n - y_k\|$ is large becomes smaller. The mean $\mathbb{E}\left[\|\hat{z}_k^n - y_k\|\right]$ (showed using a black line) and the variance of $\|\hat{z}_k^n - y_k\|$ are computed using 1000 independent runs of the iterations.*
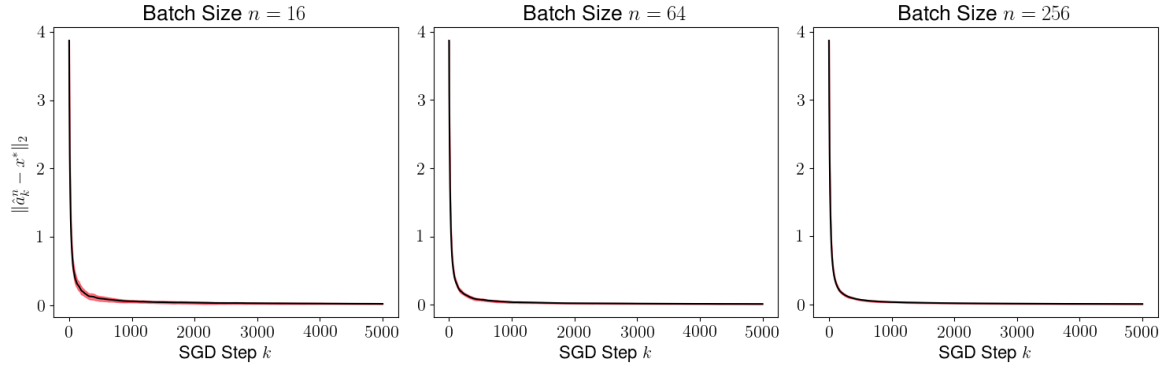
Figure 5. *Plot of $\|\hat{a}_k^n - x^*\|$ for $n = 16, 64, 256$ for the Poisson regression with a regularizer. The variance of $\|\hat{a}_k^n - x^*\|$ is decreasing as k grows, which indicates that the variance reduction property of the time averages. The mean and the variance of $\|\hat{a}_k^n - x^*\|$ are computed using 1000 independent runs of the stochastic gradient descent iterations.*

1. *Let μ_n denote the distribution of $(\hat{z}_0^n, \hat{z}_1^n, \ldots)$ and ψ be the Dirac mass on $(y_0, y_1, \ldots)$. We have $\{\mu_n\}_{n\in\mathbb{N}}$ converges to ψ in weak* topology as $n \to \infty$.*
2. *There exists a unique invariant distribution π^n of the Markov chain $(\hat{z}_0^n, \hat{z}_1^n, \ldots)$. Further, the sequence of distribution of $\hat{z}_k^n$ converges to this invariant distribution π^n.*
3. *The time average $\hat{a}_k^n$ converges in probability to the mean of π^n as $k \to \infty$.*

Proof. The proof follows from the discussions in subsection 3.2, where we show that $\hat{T}_k^n$ satisfies Assumptions 4.1 and 5.3.

Moreover, since every realization of $\hat{T}_k^n$ is a contraction operator with coefficient α, if $\|v\|_\infty \leq \frac{\|c\|_\infty}{1-\alpha}$, then $\|\hat{T}_k^n(v)\|_\infty \leq \frac{\|c\|_\infty}{1-\alpha}$. Thus, there is no loss of generality in restricting $\mathcal{X}$ to be in the compact set $\{v \in \mathbb{R}^{|\mathcal{S}|} : \|v\|_\infty \leq \frac{\|c\|_\infty}{1-\alpha}\}$. The fact that the time averages

converge in probability to the mean of the invariant distribution is a direct application of Theorem 6.3. ■

Figure 6 shows the mean (black line) and the variance (red region) of $\|v_k - \hat{z}_k^n\|_\infty$ for every time step $k = 1, \ldots, 10000$ for $n = 1, 25, 400$. To compute the mean and the variance, 1000 sample paths were taken. As proved in Theorem 7.2 above, the mean of $\|v_k - \hat{z}_k^n\|_\infty$ becomes smaller as n grows. Figure 7 shows the mean and the variance of $\|v^* - \hat{a}_k^n\|_\infty$ for $n = 1, 25, 400$ for $k = 1, \ldots, 10000$. For every n, since the sequence of distribution of $\hat{z}_k^n$ converges to a unique invariant distribution, $\hat{a}_k^n$ converges in probability to the mean of the invariant distribution. We observe in the figure that the variance of $\|v^* - \hat{a}_k^n\|_\infty$ reduces as k grows, implying that the average $\hat{a}_k^n$ is close but not equal to v^*. We further observe that the variance of $\|v^* - \hat{a}_k^n\|_\infty$ is vanishingly small for large values of n—a result that we have not formally proved here, and can be addressed in a future work.
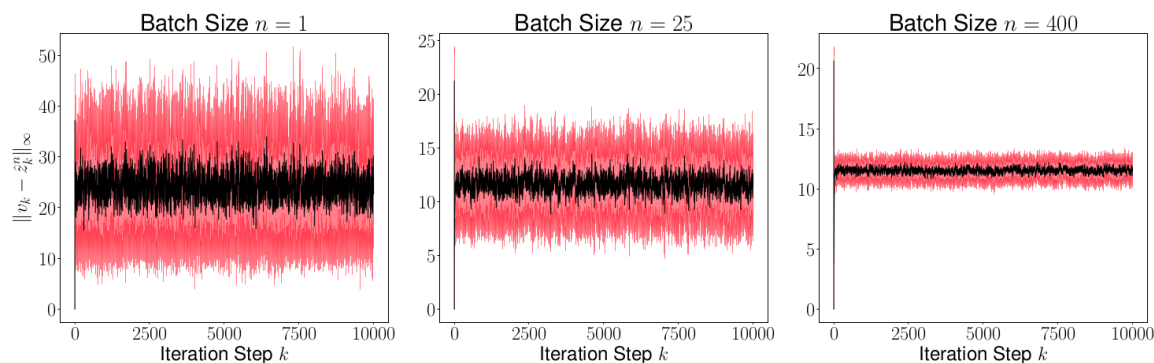
Figure 6. *Plot of $\|v_k - \hat{z}_k^n\|$ for $n = 1, 25, 400$ for $k = 1, \ldots, 1000$. It is clear from the plots that as n grows, the average and variance of $\|v_k - \hat{z}_k^n\|$ reduce. The mean and the variance of $\|v_k - \hat{z}_k^n\|$ are computed using 1000 independent runs of the iterations. There are 20 states and 5 actions in this MDP.*
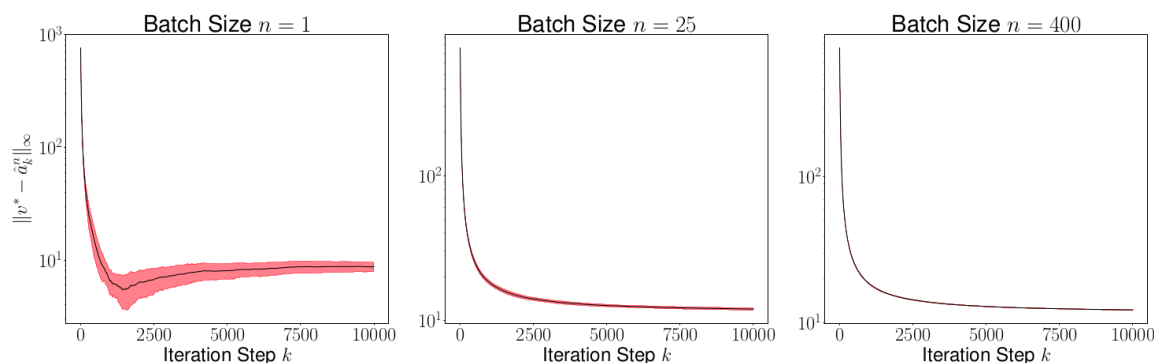
Figure 7. *Plot of $\|v^* - \hat{a}_k^n\|$ for $n = 1, 25, 400$ for $k = 1, \ldots, 1000$. Notice that for every n, the variance in $\|v^* - \hat{z}_k^n\|$ reduces as k increases. The plots are constructed using 1000 sample paths. The mean and the variance of $\|v^* - \hat{a}_k^n\|$ are computed using 1000 independent runs of the iterations.*

7.3. Synchronous batch 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 $\mathcal{Q}$ denote the set of all Q-value functions $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Similar to the Bellman operator of value iteration, we define an operator $T : \mathcal{Q} \to \mathcal{Q}$ as

$$T(q)(s,a) = c(s,a) + \alpha \sum_{s' \in \mathcal{S}} p(s'|s,a) \min_{a' \in \mathcal{A}} q(s',a').$$

Similar to the Bellman operator, T is a contraction on $(\mathcal{Q}, \|\cdot\|_\infty)$. Further, it can be shown that the fixed point of T is q^*, which is defined as $q^*(s,a) = c(s,a) + \alpha \mathbb{E}\left[v^*(s')|s,a\right]$, where $v^*(\cdot) = \min_{a \in \mathcal{A}} q(\cdot, a)$. The Q-value iteration starts with an arbitrary $q_0 \in \mathcal{Q}$ 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 $\hat{T}_k^n$:

$$\hat{T}_k^n(q)(s,a) = c(s,a) + \alpha \frac{1}{n} \sum_{i=1}^{n} \min_{a' \in \mathcal{A}} q(s'_{k,i}(s,a), a'),$$

where $\{s'_{k,i}(s,a)\}_{i=1}^n$ are n i.i.d. samples of the next state given the current state-action pair (s,a).

Let us define $\hat{z}_{k+1}^n = \hat{T}_k^n(\hat{z}_k^n)$, where $\hat{z}_0^n = q_0$. Let $\hat{a}_k^n$ be the time averaged version of $\hat{z}_k^n$. The properties of the random operator $\hat{T}_k^n$ for empirical Q-value iteration have the same properties as listed in Theorem 7.2 for the case of empirical value iteration. Thus, we omit repetition of essentially the same result here. The simulation results are plotted in Figures 8 and 9.

Figure 8 shows the mean (black line) and the variance (red region around the mean) of $\|q_k - \hat{z}_k^n\|_\infty$ for various values of n for k ranging from 1 to 10000. The mean and variance are computed using 1000 sample paths of the algorithm starting from the same initial $\hat{z}_0^n = q_0$ for all sample paths. As is evident, the mean reduces as we increase the sample size n.
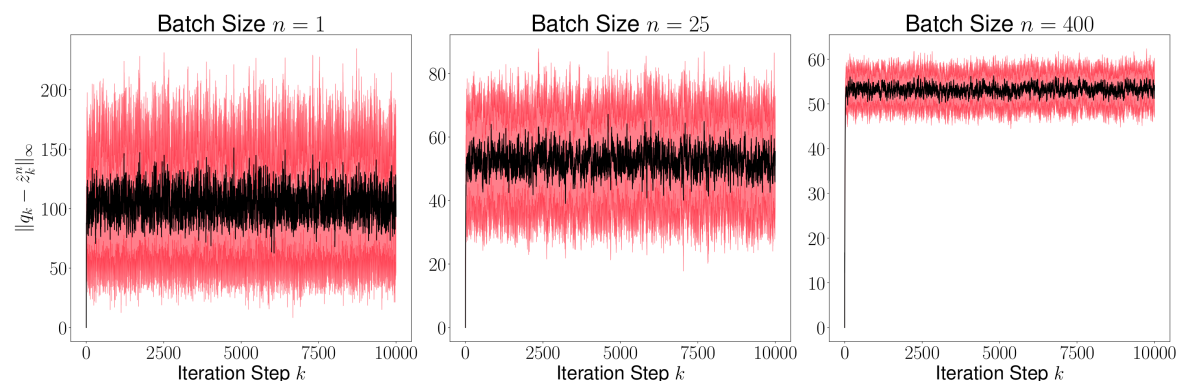
Figure 8. *Plot of the mean and the variance of $\|q_k - \hat{z}_k^n\|$ for $n = 1, 25, 400$ for $k = 1, \ldots, 1000$. It is clear from the plots that as n grows, the mean of $\|q_k - \hat{z}_k^n\|$ reduces. The mean and the variance of $\|q_k - \hat{z}_k^n\|$ are computed using 1000 independent runs of the iterations. There are 20 states and 5 actions in this MDP.*
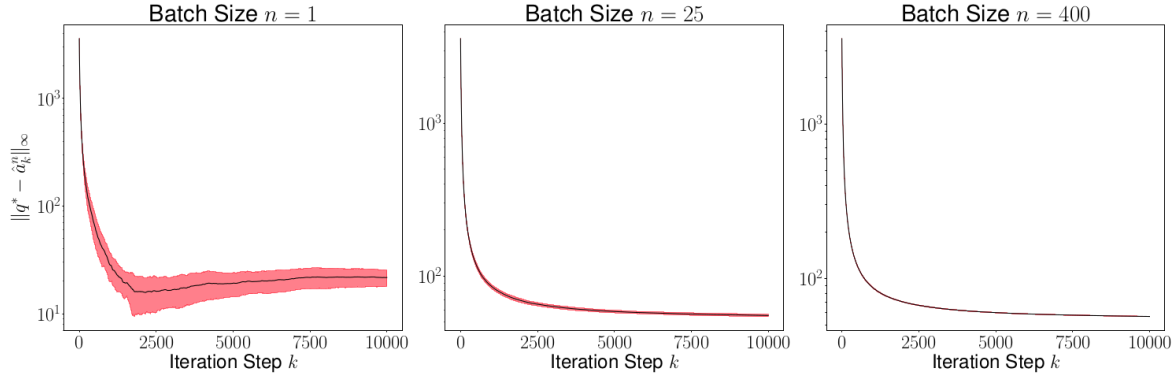
Figure 9. *Plot of the mean and the variance of $\|q^* - \hat{a}_k^n\|$ for $n = 1, 25, 400$ for $k = 1, \ldots, 1000$. Notice that for every n, the variance in $\|q^* - \hat{z}_k^n\|$ reduces as k increases. The mean and the variance of $\|q^* - \hat{a}_k^n\|$ are computed using 1000 independent runs of the iterations starting from the same initial condition $\hat{z}_0^n = q_0$.*

Figure 9 plots the mean and the variance of $\|q^* - \hat{a}_k^n\|$. Once again, we observe that the variance reduces as k increases since the Markov chain admits a unique invariant measure.

8. Proof of Theorem 4.2. To prove Theorem 4.2, we need to introduce some further notation. Define $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$ and we use $\mu_n \Rightarrow \psi$ to denote the convergence in weak topology. Let $\Pi_k : \mathcal{X}^{\mathbb{Z}_+} \to \mathcal{X}^{k+1}$ denote the projection operator that projects a sequence to its first $(k+1)$ components, that is,

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

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

Theorem 8.1 (see [14, Theorem 2.8]). *Let $(\nu_n)_{n \in \mathbb{N}} \subset \wp(\mathcal{X}^{\mathbb{Z}_+})$ be a sequence of measures. Then, $(\nu_n)_{n \in \mathbb{N}} \Rightarrow \nu_\infty$ iff $\nu_n \circ \Pi_k^{-1} \Rightarrow \nu_\infty \circ \Pi_k^{-1}$ for every $k \in \mathbb{Z}_+$.*

Recall that μ_n is a measure defined on a stochastic sequence and we would like to show that $\mu_n \Rightarrow \psi$. Due to the theorem above, all we need to establish is that $\mu_n \circ \Pi_k^{-1} \Rightarrow \psi \circ \Pi_k^{-1}$ for every k. We will establish this result via an induction argument following [51].

Fix $n \in \mathbb{N}$. For $k = 0$, let $\mu_n \circ \Pi_0^{-1}(F) = \mathbb{1}_{\{y_0\}}(F) = \psi \circ \Pi_0^{-1}(F)$ for all measurable $F \subset \mathcal{X}$. Thus, the statement is true for $k = 0$. In the induction step, we prove that if $\mu_n \circ \Pi_{k-1}^{-1} \Rightarrow \psi \circ \Pi_{k-1}^{-1}$ and for every closed set $F_1 \subset \mathcal{X}^k$ and $F_2 \subset \mathcal{X}$, we have

$$\limsup_{n \to \infty} \mu_n \circ \Pi_k^{-1}(F_1 \times F_2) \leq \psi \circ \Pi_k^{-1}(F_1 \times F_2).$$

By the portmanteau theorem [14, Theorem 2.1], we then conclude that $\mu_n \circ \Pi_k^{-1} \Rightarrow \psi \circ \Pi_k^{-1}$ for every $k \in \mathbb{Z}_+$. We divide the proof into three steps.

Step 1. We show that for any uniformly continuous function g, the map $g \circ T$ is also uniformly continuous since T is a Lipschitz operator.

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 $y, y' \in \mathcal{X}$ with $\rho(y, y') < \delta_\epsilon$, we know that $\rho(T(y), T(y')) < \alpha\delta_\epsilon < \delta_\epsilon$. Taking $x = T(y)$ and $x' = T(y')$, we conclude that $|g \circ T(y) - g \circ T(y')| < \epsilon$. This implies that $g \circ T \in U_b(\mathcal{X})$. ∎

Step 2. Next, we establish a consequence of Assumption 4.1 and Lemma 8.2. Let $x_{0:l} = (x_0, \ldots, x_l)$. We assume that $\mu_n \circ \Pi_k^{-1} \Rightarrow \psi \circ \Pi_k^{-1}$ for all $k = 0, \ldots, l$, which serves as the induction hypothesis.

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

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

Further, for any closed set $F_1 \subset \mathcal{X}^{l+1}$, we have

$$\limsup_{n \to \infty} \int_{F_1} g(T(x_l)) \mu_n \circ \Pi_l^{-1}(dx_{0:l}) \leq \int_{F_1} g(T(x_l)) \psi \circ \Pi_l^{-1}(dx_{0:l}).$$

Proof. A proof is presented in Appendix A. ∎

Step 3. We now complete the proof by establishing the induction step using the result from Lemma 8.3. We follow [51] to complete the rest of the proof. Let $F_1 \subset \mathcal{X}^{l+1}$ be any closed set. Then, using Lemma 8.3, for any $g \in U_b(\mathcal{X})$, as $n \to \infty$, we get

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

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

$$\leq \int_{\mathcal{X}^{l+1}} \left| \mathbb{E}\left[g(\hat{T}_l^n(x_l) \right] - g(T(x_l)) \right| \mu_n \circ \Pi_l^{-1}(dx_{0:l})$$

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

$$(8.1) \qquad \leq 0 \text{ as } n \to \infty,$$

where the last inequality follows from Lemma 8.3.

Now, for any $F_2 \subset \mathcal{X}$ closed, we can construct a sequence of $(g_m)_{m \in \mathbb{N}} \subset U_b(\mathcal{X})$ such that $g_{m+1} \leq g_m$ for all $m \in \mathbb{N}$ and $g_m \to 1_{F_2}$ (see, for example, Corollary 3.14 in [2, p. 79]). This leads us to the following inequality for every $m \in \mathbb{N}$:

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

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

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

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

$$(8.3) \qquad \lim_{m\to\infty} \int_{F_1} \int_{\mathcal{X}} g_m(x_{l+1}) \psi \circ \Pi_{l+1}^{-1}(dx_{0:l+1}) = \psi \circ \Pi_{l+1}^{-1}(F_1 \times F_2).$$

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. We first introduce some notation. Recall the definitions of F and F^*, which were defined before Assumption 6.1. For $f \in C_b(\mathcal{X})$, we define $g_k \in C_b(\mathcal{X})$, $k = 0, 1, \ldots$, as

$$g_0(x) = f(x), \quad g_k(x) = F^k(f)(x) = \mathbb{E}\left[f\left(\hat{T}_{k-1}^n \circ \ldots \circ \hat{T}_0^n(x) \right) \right] \text{ for } k \geq 1.$$

The following equation follows immediately from the above definitions:

$$(9.1) \qquad \mathbb{E}\left[g_k(\hat{z}_m^n) | \hat{z}_{m-1}^n \right] = g_{k+1}(\hat{z}_{m-1}^n).$$

Define $\bar{F}_k = \frac{1}{k} \sum_{i=0}^{k-1} F^i$ and $\bar{F}_k^* = \frac{1}{k} \sum_{i=0}^{k-1} (F^*)^i$. Define the constant function $c_f \in C_b(\mathcal{X})$ as

$$c_f(x) = \int f d\pi^n = \langle g_0, \bar{F}_\infty^*(\mu) \rangle = \langle \bar{F}_\infty(g_0), \mu \rangle \text{ for any } \mu \in \mathcal{M}.$$

The average of the functions g_k, denoted by $\bar{g}_k$, is

$$\bar{g}_k(x) = \frac{1}{k} \sum_{i=0}^{k-1} g_i(x) = \bar{F}_k(f)(x).$$

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 lead to the result. For the next result, let us define the measure η_K over the set $C \subset \mathcal{X}$ as

$$(9.2) \qquad \eta_K(C) = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{1}_{\{\hat{z}_k^n\}}(C) = \frac{1}{K} \sum_{i=0}^{K-1} 1_C(\hat{z}_k^n).$$

Note that $\{\eta_K\}_{K\in\mathbb{N}}$ is a sequence of random measures.

Note that since $\bar{F}_k^*(\mu) \to \pi^n$ for any $\mu \in \mathcal{M}$ by Assumption 6.1, we conclude from Prohorov's theorem that the sequence $\{\bar{F}_k^*(\mu)\}_{k\in\mathbb{N}}$ is tight. Thus, for $\epsilon > 0$, let C_ϵ be the compact set such that

$$\bar{F}_k^*(\mu)\left(C_\epsilon^\complement \right) < \epsilon^2 \text{ for all } k \in \mathbb{N}.$$

We claim the following.

Lemma 9.1. *If Assumptions* 2.1 *and* 6.1 *hold, then for every* $\epsilon > 0$,

$$\limsup_{k \to \infty} \mathbb{P} \left\{ \eta_k \left(C_\epsilon^{\complement} \right) \geq \epsilon \right\} < \epsilon.$$

Proof. We note that for any set $C \subset \mathcal{X}$, we have

$$\mathbb{E} \left[\eta_k(C) \right] = \mathbb{E} \left[\frac{1}{k} \sum_{i=0}^{k-1} \mathbb{1}_{\{\hat{z}_i^n\}}(C) \right] = \mathbb{E} \left[\frac{1}{k} \sum_{i=0}^{k-1} 1_C(\hat{z}_i^n) \right] = \bar{F}_k^*(\mu)(C).$$

Using the above identity and using Markov's inequality, we conclude that

$$\mathbb{P} \left\{ \eta_k \left(C_\epsilon^{\complement} \right) \geq \epsilon \right\} \leq \frac{\mathbb{E} \left[\eta_k \left(C_\epsilon^{\complement} \right) \right]}{\epsilon} < \epsilon.$$

The proof of the lemma is complete. ∎

Lemma 9.2. *Let* $C \subset \mathcal{X}$ *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 the Arzela–Ascoli theorem [19, Theorem 7.6.2, p. 249]. Note that as $k \to \infty$, we get

$$\bar{g}_k(x) = \left\langle \bar{F}_k(f), \mathbb{1}_{\{x\}} \right\rangle = \left\langle f, \bar{F}_k^*(\mathbb{1}_{\{x\}}) \right\rangle \to c_f(x).$$

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 establish the following result.

Lemma 9.3. *For every* $M \in \mathbb{N}$, *we have*

$$\lim_{N \to \infty} \left| \frac{1}{N} \sum_{l=0}^{N-1} \left(g_0(\hat{z}_l^n) - \bar{g}_M(\hat{z}_l^n) \right) \right| = 0 \quad \mathbb{P}\text{-almost surely.}$$

Proof. See Appendix B. ∎

The proof can now be completed easily. Fix $\epsilon > 0$ and recall the definition of the set C_ϵ from Lemma 9.1. We now note that for every $K \in \mathbb{N}$ and $M \in \mathbb{N}$, we have

$$\frac{1}{K} \sum_{k=0}^{K-1} f(\hat{z}_k^n) - \int f d\pi^n = \frac{1}{K} \sum_{k=0}^{K-1} g_0(\hat{z}_k^n) - \int f d\pi^n$$

$$\leq 1_{\{\hat{z}_0^n \in C_\epsilon, \dots, \hat{z}_{K-1}^n \in C_\epsilon\}} \left(\frac{1}{K} \sum_{k=0}^{K-1} g_0(\hat{z}_k^n) - \int f d\pi^n \right) + 2\|f\|_\infty \eta_K(C_\epsilon^{\complement}),$$

$$< 1_{\{\hat{z}_0^n \in C_\epsilon, \dots, \hat{z}_{K-1}^n \in C_\epsilon\}} \left\{ \frac{1}{K} \sum_{k=0}^{K-1} \left(g_0(\hat{z}_k^n) - \bar{g}_M(\hat{z}_k^n) \right) + \left(\frac{1}{K} \sum_{k=0}^{K-1} \bar{g}_M(\hat{z}_k^n) - \int f d\pi^n \right) \right\}$$

$$+ 2\|f\|_\infty \eta_K \left(C_\epsilon^{\complement} \right).$$

Since $\bar{g}_M(x) \to \int f d\pi^n$ uniformly on the compact set C_ϵ due to Lemma 9.2, we can pick M_ϵ sufficiently large such that for all $K \in \mathbb{N}$ and $M \geq M_\epsilon$, we get

$$1_{\{\hat{z}_0^n \in C_\epsilon, \ldots, \hat{z}_{K-1}^n \in C_\epsilon\}} \left(\frac{1}{K} \sum_{k=0}^{K-1} \bar{g}_M(\hat{z}_k^n) - \int f d\pi^n \right) < \epsilon.$$

For such M_ϵ, as $K \to \infty$, the first summand goes to 0 by Lemma 9.3. In the third summand, we know that $\eta_K(C_\epsilon^{\complement})$ is less than ϵ with probability at least $1 - \epsilon$ due to Lemma 9.1 for sufficiently large K. Collecting all these results, we conclude that

$$\limsup_{K\to\infty} \mathbb{P} \left\{ \left| \frac{1}{K} \sum_{k=0}^{K-1} f(\hat{z}_k^n) - \int f d\pi^n \right| < (2\|f\|_\infty + 2)\epsilon \right\} \geq 1 - \epsilon.$$

This completes the proof of the theorem.

9.1. Proof of Theorem 6.5. In order to prove the result, let us first consider the function $f : \mathcal{X} \to [0, \infty)$, which always takes nonnegative values. Let $f_m(x) := \min\{f(x), m\}$ be the clipped function, in which case, $f_m \in C_b(\mathcal{X})$. Define $\Omega_0 = \{\omega : C(\omega) < \infty\}$, and note that by assumption, $\mathbb{P}\{\Omega_0\} = 1$. It is obvious that for any $x \in \mathcal{X}$, if we pick $m \geq f(x)$, then $f_m(x) = f(x)$.

Note that for this case, for any $m \in \mathbb{N}$, the following inequality holds:

$$\left| \frac{1}{K} \sum_{k=0}^{K-1} f(\hat{z}_k^n) - \int f d\pi^n \right| \leq \left| \frac{1}{K} \sum_{k=0}^{K-1} \left(f(\hat{z}_k^n) - f_m(\hat{z}_k^n) \right) \right| + \left| \frac{1}{K} \sum_{k=0}^{K-1} f_m(\hat{z}_k^n) - \int f_m d\pi^n \right|$$

$$(9.3) \qquad\qquad + \left| \int f_m d\pi^n - \int f d\pi^n \right|.$$

In the next lemma, we show that each of the summands on the right is small with high probability.

Lemma 9.4. *Let $f : \mathcal{X} \to [0, \infty)$. If Assumptions 2.1, 6.1, and 6.4 hold, then (6.1) holds.*

Proof. For any $\epsilon > 0$, there exists (a random natural number) $M_\epsilon(\omega)$ such that

$$\left| \frac{1}{K} \sum_{k=0}^{K-1} \left(f(\hat{z}_k^n) - f_m(\hat{z}_k^n) \right) \right| < \epsilon \text{ for all } m \geq M_\epsilon(\omega) \text{ and } K \in \mathbb{N}.$$

Indeed, one can take $M_\epsilon(\omega) = \lceil C(\omega) \rceil$, in which case,

$$(9.4) \qquad \left| \frac{1}{K} \sum_{k=0}^{K-1} \left(f(\hat{z}_k^n) - f_m(\hat{z}_k^n) \right) \right| = 0 \text{ for all } m \geq M_\epsilon(\omega) \text{ and } K \in \mathbb{N}.$$

Further, due to the monotone convergence theorem, there exists $\bar{M}$ such that

$$\left| \int f_m d\pi^n - \int f d\pi^n \right| < \epsilon \text{ for all } m \geq \bar{M}.$$

Now, pick $\epsilon, \delta > 0$. Let us define $\Omega_{m,K} \subset \Omega$ as the set such that

$$\Omega_{m,K} = \left\{ \omega : \left| \frac{1}{K} \sum_{k=0}^{K-1} f_m(\hat{z}_k^n) - \int f_m d\pi^n \right| < \epsilon \right\}.$$

For every m, pick K_m such that $\mathbb{P}\{\Omega_{m,K_m}\} < \frac{\delta}{2^m}$; such a K_m always exists due to Theorem 6.2. Define $\check{\Omega} = \Omega_0 \bigcap \bigcap_{m=1}^{\infty} \Omega_{m,K_m}$. This immediately implies that

$$\mathbb{P}\{\check{\Omega}\} = 1 - \mathbb{P}\left\{ \Omega_0^\complement \cup \bigcup_{m=1}^{\infty} \Omega_{m,K_m}^\complement \right\} > 1 - \delta.$$

For every $\omega \in \check{\Omega}$, pick m sufficiently large such that $m \geq M_\epsilon(\omega)$, $m \geq \bar{M}$, and pick any $K = K_m$. We use (9.3) to conclude that

$$\left| \frac{1}{K} \sum_{k=0}^{K-1} f(\hat{z}_k^n) - \int f d\pi^n \right| < 3\epsilon \text{ for all } \omega \in \check{\Omega}.$$

Since $\mathbb{P}\{\omega \in \check{\Omega}\} > 1 - \delta$, the proof of the lemma is complete.   ∎

We can now complete the proof of Theorem 6.5 using the result above. Pick $f \in C(\mathcal{X})$ and define $f_+, f_- : \mathcal{X} \to [0, \infty)$ such that $f_+ = \max\{f, 0\}$ and $f_- = \max\{-f, 0\}$. This immediately yields $f = f_+ - f_-$. Further, we have

$$\frac{1}{K} \sum_{k=0}^{K-1} f(\hat{z}_k^n) = \frac{1}{K} \sum_{k=0}^{K-1} f_+(\hat{z}_k^n) - \frac{1}{K} \sum_{k=0}^{K-1} f_-(\hat{z}_k^n),$$
$$\int f d\pi^n = \int f_+ d\pi^n - \int f_- d\pi^n.$$

Now, it is easy to conclude from Lemma 9.4 that the convergence in probability result in (6.1) holds for both f_+ and f_-, due to which the convergence in probability result holds for the function f itself.

10. Conclusion. In this paper, we studied the convergence of random sequences generated from certain RSAs used in machine and reinforcement learning problems. If the randomization device used within the algorithm is independent at every iteration, and the maps do not change (for instance, the stepsize is taken as constant), then the random sequence generated can be viewed using the lens of Markov chains. We leveraged the theory of Feller Markov chains to deduce many interesting characteristics of the random sequence and their distributions. Specifically, under reasonable conditions, we showed that the entire random sequence is close to the sequence generated by the exact algorithm with high probability for n sufficiently large. We further showed that the average of the random sequence converges to the mean of the invariant distribution of the sequence of there exists a unique invariant measure of the Markov chain.

We expect that the results presented here can be applied to MDPs over continuous state and action spaces (referred to as continuous MDPs). Indeed, finite time guarantees of empirical

value iteration for continuous MDPs with function approximator have been presented in [59, 74, 73, 40, 72, 41, 35] under a variety of assumptions on the MDPs and performance criteria. Convergence of the asynchronous algorithm for continuous MDPs with nonparametric function approximation is presented in [71]. It will be interesting to investigate if the output of these algorithms satisfies the sufficient conditions for Theorems 4.2 and 6.2. It will also be interesting to apply the results presented here to variance reduced algorithms [75, 85, 68, 87, 86] that have been developed recently.

Another problem left for future research is to determine bounds on $\mathbb{P}\left\{\rho(\hat{z}_k^n, y_k) \geq \epsilon\right\}$ for any $\epsilon > 0$ for every n and k. Such bounds would unify our understanding of finite time guarantees for RSAs and allow us to improve the existing algorithms. We hope that the unified framework developed in this paper will be useful for analyzing many other learning algorithms in the future, particularly for analyzing MDPs over compact uncountable state and action spaces.

Appendix A. Proof of Lemma 8.3. Since $\mu_n \circ \Pi_l^{-1} \to \psi \circ \Pi_l^{-1}$ in weak topology as $n \to \infty$, we conclude that $(\mu_n \circ \Pi_l^{-1})_{n \in \mathbb{N}}$ is tight. For a fixed $\epsilon > 0$, let $F_2 \subset \mathcal{X}$ be the compact set such that

$$\mu_n \circ \Pi_l^{-1}(\mathcal{X}^l \times F_\epsilon) > 1 - \frac{\epsilon}{4\|g\|_\infty} \text{ for every } n \in \mathbb{N}.$$

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

$$\left| \mathbb{E}\left[g(\hat{T}_k^n(x)) \right] - g(T(x)) \right| < \epsilon \text{ for all } x \in K.$$

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 $N_\epsilon(g, K)$ such that

$$\sup_{x \in K} \mathbb{P}\left\{ \rho(\hat{T}_k^n(x), T(x)) > \delta_\epsilon \right\} < \frac{\epsilon}{2\|g\|_\infty} \text{ for all } n \geq N_\epsilon(g, K).$$

This implies

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

The proof of the lemma is complete. ■

We are now in a position to prove the first result. Consider the following expressions:

$$\int_{\mathcal{X}^{l+1}} \left| \mathbb{E}\left[g(\hat{T}_l^n(x_l))\right] - g(T(x_l)) \right| \mu_n \circ \Pi_l^{-1}(dx_{0:l})$$

$$= \int_{\mathcal{X}^l \times F_\epsilon} \left| \mathbb{E}\left[g(\hat{T}_l^n(x_l))\right] - g(T(x_l)) \right| \mu_n \circ \Pi_l^{-1}(dx_{0:l})$$

$$+ \int_{\mathcal{X}^l \times F_\epsilon^\complement} \left| \mathbb{E}\left[g(\hat{T}_l^n(x_l))\right] - g(T(x_l)) \right| \mu_n \circ \Pi_l^{-1}(dx_{0:l})$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

To establish the second result, note that we can construct a sequence of $(g_m)_{m \in \mathbb{N}} \subset U_b(\mathcal{X}^l)$ such that $g_{m+1} \leq g_m$ for all $m \in \mathbb{N}$ and $g_m \to 1_{F_1}$ (see, for example, Corollary 3.14 in [2, p. 79]). Lemma 8.2 implies that $g_m \times g \circ T$ is uniformly continuous over $\mathcal{X}^{l+1}$. This implies that

$$\limsup_{n \to \infty} \int_{F_1} g(T(x_l)) \mu_n \circ \Pi_l^{-1}(dx_{0:l}) \leq \limsup_{n \to \infty} \int_{\mathcal{X}^{l+1}} g_m(x_{0:l}) g(T(x_l)) \mu_n \circ \Pi_l^{-1}(dx_{0:l})$$

$$\leq \int_{\mathcal{X}^{l+1}} g_m(x_{0:l}) g(T(x_l)) \psi \circ \Pi_l^{-1}(dx_{0:l}).$$

Letting $m \to \infty$ on the right side and applying the dominated convergence theorem yields the second result.

The proof of the lemma is complete.

Appendix B. Proof of Lemma 9.3. This result is established in [21] for the special case when $\mathcal{X}$ is a compact Hausdorff set; our argument essentially imitates the arguments there. Define $g_p(\hat{z}_{-i}^n) := 0$ for all $i \in \mathbb{N}$ and $p \in \{0, 1, \dots\}$. Let us define random variables $u_{p,l}$ for $l, p = 0, 1, \dots$ as

$$u_{p,l} = g_p(\hat{z}_l^n) - g_{p+1}(\hat{z}_{l-1}^n).$$

Thus, $u_{p,0} = g_p(\hat{z}_0^n)$. By definition of $u_{p,l}$, we have for any $l, k \geq 0$

$$g_0(\hat{z}_l^n) - g_k(\hat{z}_{l-k}^n) = \left(g_0(\hat{z}_l^n) - g_1(\hat{z}_{l-1}^n)\right) + \cdots + \left(g_{k-1}(\hat{z}_{l-k+1}^n) - g_k(\hat{z}_{l-k}^n)\right)$$

(B.1)
$$= u_{0,l} + \cdots + u_{k-1,l} = \sum_{p=0}^{k-1} u_{p,l}.$$

We use the above expression to establish the following identity.

Lemma B.1. *For $M \leq N$, we have*

(B.2) $$\frac{1}{N} \sum_{l=0}^{N-1} \left(g_0(\hat{z}_l^n) - \bar{g}_M(\hat{z}_l^n)\right) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{p=0}^{k-1} \left(\frac{1}{N} \sum_{l=0}^{N-1} u_{p,l}\right) - \frac{1}{M} \sum_{k=0}^{M-1} \left(\frac{1}{N} \sum_{l=N-k}^{N-1} g_k(\hat{z}_l^n)\right).$$

Proof. For any $l \in \{0, 1, \dots\}$, we have

$$g_0(\hat{z}_l^n) - \bar{g}_M(\hat{z}_l^n) = g_0(\hat{z}_l^n) - \frac{1}{M} \sum_{k=0}^{M-1} g_k(\hat{z}_l^n) = \frac{1}{M} \sum_{k=0}^{M-1} (g_0(\hat{z}_l^n) - g_k(\hat{z}_l^n))$$

$$= \frac{1}{M} \sum_{k=0}^{M-1} \left(\left(g_0(\hat{z}_l^n) \right) - g_k(\hat{z}_{l-k}^n) \right) + \left(g_k(\hat{z}_{l-k}^n) - g_k(\hat{z}_l^n) \right)$$

$$= \frac{1}{M} \sum_{k=0}^{M-1} \left(\left(g_0(\hat{z}_l^n) \right) - g_k(\hat{z}_{l-k}^n) \right) + \left(g_k(\hat{z}_{l-k}^n) - g_k(\hat{z}_l^n) \right).$$

This yields

(B.3)
$$\frac{1}{N} \sum_{l=0}^{N-1} (g_0(\hat{z}_l^n) - \bar{g}_M(\hat{z}_l^n)) = \frac{1}{NM} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \left(\left(g_0(\hat{z}_l^n) \right) - g_k(\hat{z}_{l-k}^n) \right) + \left(g_k(\hat{z}_{l-k}^n) - g_k(\hat{z}_l^n) \right).$$

For any $l, k \geq 0$, (B.1) yields

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

Consider the second summand in (B.3). For any $k \in \mathbb{N}$, we have

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

These expressions immediately establish the equality in (B.2). ∎

Lemma B.2. *For a fixed $p \in \{0, 1, \ldots\}$, we have*

$$\lim_{N \to \infty} \frac{1}{N} \sum_{l=0}^{N-1} u_{p,l} = 0 \quad \mathbb{P}\text{-almost surely.}$$

Proof. For any $p \in \{0, 1, \ldots\}$ and $l \in \mathbb{N}$, let $\mathcal{F}_{p,l-1}$ denote the σ-algebra generated by $(\hat{z}_0^n, \ldots, \hat{z}_{l-1}^n, u_{p,0}, \ldots, u_{p,l-1})$. We show that $\{u_{p,l}\}_{l=1}^{\infty}$ forms a martingale with respect to the σ-algebra $\mathcal{F}_{p,l-1}$. By the definition of $u_{p,l}$, we immediately have for any $l \geq 1$,

(B.4)
$$\mathbb{E}\left[u_{p,l} | \hat{z}_{l-1}^n \right] = \mathbb{E}\left[g_p(\hat{z}_l^n) - g_{p+1}(\hat{z}_{l-1}^n) | \hat{z}_{l-1}^n \right] = 0.$$

This implies

$$\mathbb{E}\left[u_{p,l} | \mathcal{F}_{p,l-1} \right] = 0, \quad \mathbb{E}\left[(u_{p,l})^2 \right] \leq 2\|f\|_{\infty}^2 \text{ for } p \in \{0, 1, \ldots\} \text{ and } l \in \mathbb{N}.$$

The proof then is an immediate consequence of the strong law of large numbers for martingales in [55, p. 66]. ∎

Now note that for any $M \in \mathbb{N}$, as $N \to \infty$, the right side of (B.2) converges to 0 almost surely by Lemma B.2. This yields the result.

REFERENCES

[1] J. ABOUNADI, D. P. BERTSEKAS, AND V. BORKAR, *Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms*, SIAM J. Control Optim., 41 (2002), pp. 1–22.

[2] C. D. ALIPRANTIS AND K. C. BORDER, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer, New York, 2006.

[3] A. L. ALMUDEVAR, *Approximate Iterative Algorithms*, CRC Press, Boca Raton, FL, 2014.

[4] 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, PMLR 70, MLR Press, 2017, pp. 176–185.

[5] F. BACH, *Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression*, J. Mach. Learn. Res., 15 (2014), pp. 595–627.

[6] F. BACH AND E. MOULINES, *Non-strongly-convex smooth stochastic approximation with convergence rate $o\,(1/n)$*, in Advances in Neural Information Processing Systems, 2013, pp. 773–781.

[7] C. L. BECK AND R. SRIKANT, *Error bounds for constant step-size Q-learning*, Systems Control Lett., 61 (2012), pp. 1203–1208.

[8] D. P. BERTSEKAS, *Dynamic Programming and Optimal Control*, Athena Scientific, Belmont, MA, 2011.

[9] D. P. BERTSEKAS, *Abstract Dynamic Programming*, Athena Scientific, Belmont, MA, 2013.

[10] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Neuro-Dynamic Programming*, Vol. 5, Athena Scientific, Belmont, MA, 1996.

[11] D. P. BERTSEKAS AND H. YU, *Q-learning and enhanced policy iteration in discounted dynamic programming*, Math. Oper. Res., 37 (2012), pp. 66–94.

[12] J. BHANDARI, D. RUSSO, AND R. SINGAL, *A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation*, preprint, arXiv:1806.02450, 2018.

[13] R. N. BHATTACHARYA AND O. LEE, *Ergodicity and central limit theorems for a class of Markov processes*, in Multivariate Statistics and Probability, Elsevier, Amsterdam, 1989, pp. 80–90.

[14] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley & Sons, New York, 2013.

[15] S. BONNABEL, *Stochastic gradient descent on Riemannian manifolds*, IEEE Trans. Automat. Control, 58 (2013), pp. 2217–2229.

[16] V. S. BORKAR, *Stochastic Approximation: A Dynamical Systems Viewpoint*, Texts Read. Math. 48, Springer, New York, 2009.

[17] V. S. BORKAR AND S. P. MEYN, *The ODE method for convergence of stochastic approximation and reinforcement learning*, SIAM J. Control Optim., 38 (2000), pp. 447–469.

[18] A. A. BOROVKOV, *Ergodicity and Stability of Stochastic Processes*, John Wiley & Sons, New York, 1998.

[19] F. S. BOTELHO, *Real Analysis and Applications*, Springer, New York, 2018.

[20] L. BOTTOU, *Large-scale machine learning with stochastic gradient descent*, in Proceedings of COMPSTAT'2010, Springer, New York, 2010, pp. 177–186.

[21] L. BREIMAN, *The strong law of large numbers for a class of Markov chains*, Ann. Math. Stat., 31 (1960), pp. 801–803.

[22] A. D. R. CHOUDARY AND C. P. NICULESCU, *Real Analysis on Intervals*, Springer, New York, 2014.

[23] W. L. COOPER, S. G. HENDERSON, AND M. E. LEWIS, *Convergence of simulation-based policy iteration*, Probab. Engrg. and Inform. Sci., 17 (2003), pp. 213–234.

[24] W. L. COOPER AND B. RANGARAJAN, *Performance guarantees for empirical Markov decision processes with applications to multiperiod inventory models*, Oper. Res., 60 (2012), pp. 1267–1281.

[25] G. DALAL, B. SZÖRÉNYI, G. THOPPE, AND S. MANNOR, *Finite sample analyses for TD(0) with function approximation*, in 32nd AAAI Conference on Artificial Intelligence, 2018.

[26] 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.

[27] P. DIACONIS AND D. FREEDMAN, *Iterated random functions*, SIAM Rev., 41 (1999), pp. 45–76.

[28] A. DIEULEVEUT, A. DURMUS, AND F. BACH, *Bridging the Gap Between Constant Step Size Stochastic Gradient Descent and Markov Chains*, preprint, arXiv:1707.06386, 2017.

[29] A. DIEULEVEUT, N. FLAMMARION, AND F. BACH, *Harder, better, faster, stronger convergence rates for least-squares regression*, J. Mach. Learn. Res., 18 (2017), pp. 3520–3570.

[30] M. DUFLO, *Random Iterative Models*, Stoch. Model. Appl. Probab. 34, Springer, New York, 2013.

[31] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer, New York, 2007.

[32] J.-C. FORT AND G. PAGES, *Asymptotic behavior of a Markovian stochastic algorithm with constant step*, SIAM J. Control Optim., 37 (1999), pp. 1456–1482.

[33] S. B. GELFAND AND S. K. MITTER, *Recursive stochastic algorithms for global optimization in r^d*, SIAM J. Control Optim., 29 (1991), pp. 999–1018.

[34] S. B. GELFAND AND S. K. MITTER, *Metropolis-type annealing algorithms for global optimization in R^d*, SIAM J. Control Optim., 31 (1993), pp. 111–131.

[35] A. GUPTA, R. JAIN, AND P. GLYNN, *Probabilistic contraction analysis of iterated random operators*, SIAM J. Control Optim., submitted, arXiv:1804.01195.

[36] A. GUPTA, R. JAIN, AND P. W. GLYNN, *An empirical algorithm for relative value iteration for average-cost MDPs*, in Proceedings of the 54th IEEE Conference on Decision and Control, 2015, pp. 5079–5084.

[37] H. GUPTA, R. SRIKANT, AND L. YING, *Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning*, in Advances in Neural Information Processing Systems, 2019, pp. 4706–4715.

[38] 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.

[39] W. B. HASKELL, R. JAIN, AND D. KALATHIL, *Empirical dynamic programming*, Math. Oper. Res., 41 (2016), pp. 402–429.

[40] W. B. HASKELL, R. JAIN, H. SHARMA, AND P. YU, *A universal empirical dynamic programming algorithm for continuous state MDPs*, IEEE Trans. Automat. Control, 65 (2019), pp. 115–129.

[41] W. B. HASKELL, P. YU, H. SHARMA, AND R. JAIN, *Randomized function fitting-based empirical value iteration*, in Proceedings of the 2017 IEEE 56th Annual Conference on Decision and Control, IEEE, 2017, pp. 2467–2472.

[42] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Discrete-time Markov Control Processes: Basic Optimality Criteria*, Probab. Theory Stoch. Process. 30, Springer, New York, 1996.

[43] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Further Topics on Discrete-time Markov Control Processes*, Stoch. Model. Appl. Probab. 42, Springer, New York, 2012.

[44] W. HOEFFDING, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30.

[45] J. HUANG, I. KONTOYIANNIS, AND S. P. MEYN, *The ODE Method and Spectral Theory of Markov Operators*, in Stochastic Theory and Control, Springer, New York, 2002, pp. 205–221.

[46] 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.

[47] 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*, J. Mach. Learn. Res., 18 (2018), pp. 1–42.

[48] R. JAIN AND P. VARAIYA, *Simulation-based optimization of Markov decision processes: An empirical process theory approach*, Automatica, 46 (2010), pp. 1297–1304.

[49] R. JOHNSON AND T. ZHANG, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems, 2013, pp. 315–323.

[50] D. KALATHIL, V. S. BORKAR, AND R. JAIN, *Empirical q-value iteration*, Stoch. Syst., to appear, arXiv:1412.0180, 2014.

[51] A. F. KARR, *Weak convergence of a sequence of Markov chains*, Probab. Theory Relat. Fields, 33 (1975), pp. 41–48.

[52] H. KUSHNER AND G. G. YIN, *Stochastic Approximation and Recursive Algorithms and Applications*, Stoch. Model. Appl. Probab. 35, Springer, New York, 2003.

[53] C. LAKSHMINARAYANAN AND C. SZEPESVARI, *Linear stochastic approximation: How far does constant step-size and iterate averaging go?*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, 2018, pp. 1347–1355.

[54] M. LEDOUX, *The Concentration of Measure Phenomenon*, Math. Surveys Monogr. 89, AMS, Providence, RI, 2001.

[55] M. Loève, *Probability Theory* II, Springer, New York, 1977.

[56] S. Mandt, M. D. Hoffman, and D. M. Blei, *Stochastic gradient descent as approximate Bayesian inference*, J. Mach. Learn. Res., 18 (2017), pp. 4873–4907.

[57] S. Merity, N. S. Keskar, and R. Socher, *Regularizing and Optimizing LSTM Language Models*, preprint, arXiv:1708.02182, 2017.

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

[59] R. Munos and C. Szepesvári, *Finite-time bounds for fitted value iteration*, J. Mach. Learn. Res., 9 (2008), pp. 815–857.

[60] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

[61] D. O'Neill, M. Levorato, A. Goldsmith, and U. Mitra, *Residential demand response using reinforcement learning*, in First IEEE International Conference on Smart Grid Communications, IEEE, 2010, pp. 409–414.

[62] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, AMS, Providence, RI, 2005.

[63] D. Pollard, *Convergence of Stochastic Processes*, Springer Ser. Statist., Springer, New York, 1984, https://doi.org/10.1007/978-1-4612-5254-2.

[64] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.

[65] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, New York, 2014.

[66] M. Raginsky and I. Sason, *Concentration of measure inequalities in information theory, communications, and coding*, Found. Trends Commun. Inform. Theory, 10 (2013), pp. 1–247.

[67] E. K. Ryu and S. Boyd, *Primer on monotone operator methods*, Appl. Comput. Math, 15 (2016), pp. 3–43.

[68] H. Sato, H. Kasai, and B. Mishra, *Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport*, SIAM J. Optim., 29 (2019), pp. 1444–1472.

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

[70] R. Schoknecht and A. Merke, *Convergent combinations of reinforcement learning with linear function approximation*, in Advances in Neural Information Processing Systems, 2003, pp. 1611–1618.

[71] D. Shah and Q. Xie, *Q-learning with nearest neighbors*, in Advances in Neural Information Processing Systems, 2018, pp. 3111–3121.

[72] H. Sharma, M. Jafarnia-Jahromi, and R. Jain, *Approximate relative value learning for average-reward continuous state MDPs*, in Proceedings of UAI, 2019.

[73] H. Sharma and R. Jain, *An approximately optimal relative value learning algorithm for averaged MDPs with continuous states and actions*, in Proceedings of the 57th Annual Allerton Conference on Communication, Control, and Computing, IEEE, 2019, pp. 734–740.

[74] H. Sharma, R. Jain, and A. Gupta, *An empirical relative value learning algorithm for non-parametric MDPs with continuous state space*, in Proceedings of the 18th European Control Conference, IEEE, 2019, pp. 1368–1373.

[75] 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 Advances in Neural Information Processing Systems, 2018, pp. 5186–5196.

[76] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, *Convergence results for single-step on-policy reinforcement-learning algorithms*, Mach. Learn., 38 (2000), pp. 287–308.

[77] R. Srikant and L. Ying, *Finite-time Error Bounds for lInear Stochastic Approximation and TD Learning*, preprint, arXiv:1902.00923, 2019.

[78] W. Suttle, Z. Yang, K. Zhang, Z. Wang, T. Başar, and J. Liu, *A Multi-agent Off-policy Actor-Critic Algorithm for Distributed Reinforcement Learning*, preprint, arXiv:1903.06372, 2019.

[79] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 2018.

[80] N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan, *Averaging Stochastic Gradient Descent on Riemannian Manifolds*, preprint, arXiv:1802.09128, 2018.

[81] J. TSITSIKLIS, D. BERTSEKAS, AND M. ATHANS, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 31 (1986), pp. 803–812.

[82] J. N. TSITSIKLIS, *Asynchronous stochastic approximation and Q-learning*, Mach. Learn., 16 (1994), pp. 185–202.

[83] A. W. VAN DER VAART AND J. A. WELLNER, *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer-Verlag, New York, 1996, https://doi.org/10.1007/978-1-4757-2545-2.

[84] 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.

[85] M. J. WAINWRIGHT, *Variance-Reduced Q-Learning is Minimax Optimal*, preprint, arXiv:1906.04697, 2019.

[86] Z. WANG, K. JI, Y. ZHOU, Y. LIANG, AND V. TAROKH, *Spiderboost: A Class of Faster Variance-Reduced Algorithms for Nonconvex Optimization*, preprint, arXiv:1810.10690, 2018.

[87] Z. WANG, Y. ZHOU, Y. LIANG, AND G. LAN, *Stochastic Variance-reduced Cubic Regularization for Nonconvex Optimization*, in Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 2731–2740.

[88] W. WHITT, *Approximations of dynamic programs*, I, Math. Oper. Res., 3 (1978), pp. 231–243.

[89] W. WHITT, *Approximations of dynamic programs*, II, Math. Oper. Res., 4 (1979), pp. 179–185.

[90] H. XIONG, T. XU, Y. LIANG, AND W. ZHANG, *Non-asymptotic Convergence of Adam-type Reinforcement Learning Algorithms under Markovian Sampling*, preprint, arXiv:2002.06286, 2020.

[91] T. XU, Z. WANG, AND Y. LIANG, *Non-asymptotic Convergence Analysis of Two Time-scale (Natural) Actor-critic Algorithms*, preprint, arXiv:2005.03557, 2020.

[92] K. ZHANG, A. KOPPEL, H. ZHU, AND T. BAŞAR, *Global Convergence of Policy Gradient Methods to (Almost) Locally Optimal Policies*, preprint, arXiv:1906.08383, 2019.

[93] K. ZHANG, E. MIEHLING, AND T. BAŞAR, *Approximate Equilibrium Computation for Discrete-Time Linear-Quadratic Mean-Field Games*, preprint, arXiv:2003.13195, 2020.

[94] K. ZHANG, Z. YANG, AND T. BAŞAR, *Decentralized Multi-Agent Reinforcement Learning with Networked Agents: Recent Advances*, preprint, arXiv:1912.03821, 2019.

[95] K. ZHANG, Z. YANG, AND T. BAŞAR, *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*, preprint, arXiv:1911.10635, 2019.

[96] K. ZHANG, Z. YANG, AND T. BASAR, *Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games*, in Advances in Neural Information Processing Systems, 2019, pp. 11602–11614.