Empirical Dynamic Programming

William B. Haskell
ISE Department, National University of Singapore
wbhaskell@gmail.com

Rahul Jain*
EE, CS & ISE Departments, University of Southern California
rahul.jain@usc.edu

Dileep Kalathil
EECS Department, University of California, Berkeley
dileep.kalathil@berkeley.edu

We propose empirical dynamic programming algorithms for Markov decision processes (MDPs). In these algorithms, the exact expectation in the Bellman operator in classical value iteration is replaced by an empirical estimate to get 'empirical value iteration' (EVI). Policy evaluation and policy improvement in classical policy iteration are also replaced by simulation to get 'empirical policy iteration' (EPI). Thus, these empirical dynamic programming algorithms involve iteration of a random operator, the empirical Bellman operator. We introduce notions of probabilistic fixed points for such random monotone operators. We develop a stochastic dominance framework for convergence analysis of such operators. We then use this to give sample complexity bounds for both EVI and EPI. We then provide various variations and extensions to asynchronous empirical dynamic programming, the minimax empirical dynamic program, and show how this can also be used to solve the dynamic newsvendor problem. Preliminary experimental results suggest a faster rate of convergence than stochastic approximation algorithms.

Key words: dynamic programming; empirical methods; simulation; random operators; probabilistic fixed points.
MSC2000 subject classification: 49L20, 90C39, 37M05, 62C12, 47B80, 37H99.
OR/MS subject classification: TBD.
History: Submitted: November 21, 2013. Revised: March 10, 2015

---

1. Introduction Markov decision processes (MDPs) are natural models for decision making in a stochastic dynamic setting for a wide variety of applications. The 'principle of optimality' introduced by Richard Bellman in the 1950s has proved to be one of the most important ideas in stochastic control and optimization theory. It leads to dynamic programming algorithms for solving sequential stochastic optimization problems. And yet, it is well-known that it suffers from a "curse of dimensionality" [5, 33, 19], and does not scale computationally with state and action space size. In fact, the dynamic programming algorithm is known to be PSPACE-hard [34].

This realization led to the development of a wide variety of 'approximate dynamic programming' methods beginning with the early work of Bellman himself [6]. These ideas evolved independently in different fields, including the work of Werbos [48], Kushner and Clark [26] in control theory, Minsky [29], Barto, et al [4] and others in computer science, and Whitt in operations research [49, 50]. The key idea was an approximation of value functions using basis function approximation [6], state aggregation [8], and subsequently function approximation via neural networks [3]. The difficulty was universality of the methods. Different classes of problems require different approximations.

Thus, alternative model-free methods were introduced by Watkins and Dayan [47] where a Q-learning algorithm was proposed as an approximation of the value iteration procedure. It was soon noticed that this is essentially a stochastic approximations scheme introduced in the 1950s

*Corresponding author. This research was supported via NSF CAREER award CNS-0954116 and ONR Young Investigator Award N00014-12-1-0766.

by Robbins and Munro [37] and further developed by Kiefer and Wolfowitz [23] and Kushner and Clarke [26]. This led to many subsequent generalizations including the temporal differences methods [9] and actor-critic methods [24, 25]. These are summarized in [9, 43, 44, 35]. One shortcoming in this theory is that most of these algorithms require a recurrence property to hold, and in practice, often work only for finite state and action spaces. Furthermore, while many techniques for establishing convergence have been developed [11], including the o.d.e. method [28, 12], establishing rate of convergence has been quite difficult [22]. Thus, despite considerable progress, these methods are not universal, sample complexity bounds are not known, and so other directions need to be explored.

A natural thing to consider is simulation-based methods. In fact, engineers and computer scientists often do dynamic programming via Monte-Carlo simulations. This technique affords considerable reduction in computation but at the expense of uncertainty about convergence. In particular, the expectation in the Bellman operator may be expensive/impossible to compute exactly and it must be approximated by simulation. In this paper, we analyze a simple and natural method for simulation-based dynamic programming which we call 'empirical dynamic programming'. The idea behind the algorithms is quite simple and natural: In the DP algorithms, replace the expectation in the Bellman operator with a sample-average approximation. The idea is widely used in the stochastic programming literature, but mostly for single-stage problems. In our case, replacing the expectation with an empirical expectation operator, makes the classical Bellman operator a random operator. In the DP algorithm, we must find the fixed point of the Bellman operator. In the empirical DP algorithms, we must find a probabilistic fixed point of the random operator. Random operators are closely connected to random elements (see [42]). Random elements extends the notion of a random variables where outcomes are more general objects like functions or sets. Random operators can be viewed as random elements where the outcome is operator-valued.

In this paper, we first introduce two notions of probabilistic fixed points that we call 'strong' and 'weak'. We then show that asymptotically these concepts converge to the deterministic fixed point of the classical Bellman operator. The key technical idea of this paper is a novel stochastic dominance argument that is used to establish probabilistic convergence of a random operator, and in particular, of our empirical algorithms. Stochastic dominance, the notion of an order on the space of random variables, is a well developed tool (see [30, 40] for a comprehensive study).

In this paper, we develop a theory of empirical dynamic programming (EDP) for Markov decision processes (MDPs). Specifically, we make the following contributions in this paper. First, we propose both empirical value iteration and policy iteration algorithms and show that these converge. Each is an empirical variant of the classical algorithms. In empirical value iteration (EVI), the expectation in the Bellman operator is replaced with a sample-average or empirical approximation. In empirical policy iteration (EPI), both policy evaluation and the policy iteration are done via simulation, i.e., replacing the exact expectation with the simulation-derived empirical expectation. We note that the EDP class of algorithms is not a stochastic approximation scheme. Thus, we don't need a recurrence property as is commonly needed by stochastic approximation-based methods. Thus, the EDP algorithms are relevant for a larger class of problems (in fact, for any problem for which exact dynamic programming can be done.) We provide convergence and sample complexity bounds for both EVI and EPI. But we note that EDP algorithms are essentially "off-line" algorithms just as classical DP is. Moreover, we also inherit some of the problems of classical DP such as scalability issues with large state spaces. These can be overcome in the same way as one does for classical DP, i.e., via state aggregation and function approximation.

Second, since the empirical Bellman operator is a random monotone operator, it doesn't have a deterministic fixed point. Thus, we introduce new mathematical notions of probabilistic fixed points. These concepts are pertinent when we are approximating a deterministic operator with an improving sequence of random operators. Under fairly mild assumptions, we show that our two

probabilistic fixed point concepts converge to the deterministic fixed point of the classical monotone operator.

Third, since scant mathematical methods exist for convergence analysis of random operators, we develop a new technique based on stochastic dominance for convergence analysis of iteration of random operators. This technique allows for finite sample complexity bounds. We use this idea to prove convergence of the empirical Bellman operator by constructing a dominating Markov chain. We note that there is an extant theory of monotone random operators developed in the context of random dynamical systems [15] but the techniques for convergence analysis of random operators is not relevant to our context. Our stochastic dominance argument can be applied for more general random monotone operators than just the empirical Bellman operator.

We also give a number of extensions of the EDP algorithms. We show that EVI can be performed asynchronously, making a parallel implementation possible. Second, we show that a saddle point equilibrium of a zero-sum stochastic game can be computed approximately by using the minimax Bellman operator. Third, we also show how the EDP algorithm and our convergence techniques can be used even with continuous state and action spaces by solving the dynamic newsvendor problem.

Related Literature A key question is how is the empirical dynamic programming method different from other methods for simulation-based optimization of MDPs, on which there is substantial literature. We note that most of these are stochastic approximation algorithms, also called reinforcement learning in computer science [43, 44]. Within this class. there are Q learning algorithms, actor-critic algorithms, and appproximate policy iteration algorithms. Q-learning was introduced by Watkins and Dayan but its convergence as a stochastic approximation scheme was done by Bertsekas and Tsitsiklis [9]. Q-learning for the average cost case was developed in [1], and in a risk-sensitive context was developed in [10]. The learning rates for Q-learning were established in [18] and similar sample complexity bounds were given in [22]. Actor-critic algorithms as a two time-scale stochastic approximation was developed in [24]. But the most closely related work is optimistic policy iteration [46] wherein simulated trajectories are used for policy evaluation while policy improvement is done exactly. The algorithm is a stochastic approximation scheme and its almost sure convergence follows. This is true of all stochastic approximation schemes but they do require some kind of recurrence property to hold. In contrast, EDP is not a stochastic approximation scheme, hence it does not need such assumptions. However, we can only guarantee its convergence in probability.

A class of simulation-based optimization algorithms for MDPs that is not based on stochastic approximations is the adaptive sampling methods developed by Fu, Marcus and co-authors [13, 14]. These are based on the pursuit automata learning algorithms [45, 36, 32] and combine multi-armed bandit learning ideas with Monte-Carlo simulation to adaptively sample state-action pairs to approximate the value function of a finite-horizon MDP.

The closest work to this paper is by Rust [39] wherein he considers the dynamic programming problem over continuous state spaces. He introduces a simulation idea wherein the expectation in the Bellman operator is replaced by a sample average which involves sampling of the next set of states which are fixed for all iterations. He calls it a random Bellman operator but unlike our 'empirical' Bellman operator, it involves the transition probability kernel. The proof idea for convergence is also quite different from ours. Another closely related work is [31] which considers continuous state and action spaces. Finite time bounds on error are obtained but the analysis techniques are very different.

Some other closely related works are [16] which introduces simulation-based policy iteration (for average cost MDPs). It basically shows that almost sure convergence of such an algorithm can fail. Another related work is [17]. wherein a simulation-based value iteration is proposed for a finite horizon problem. Convergence in probability is established if the simulation functions corresponding to the MDP is Lipschitz continuous. Another closely related paper is [2], which considers value

iteration with error. We note that our focus is on infinite horizon discounted MDPs. Moreover, we do not require any Lipschitz continuity condition. We show that EDP algorithms will converge (probabilistically) whenever the classical DP algorithms will converge (which is always).

A survey on approximate policy iteration is provided in [7]. Approximate dynamic programming (ADP) methods are surveyed in [35]. In fact, a notion of post-decision state was also introduced by Powell whose motivation is very similar to ours. In fact, many Monte-Carlo-based dynamic programming algorithms are introduced herein (but without convergence proof.) Simulation-based uniform estimation of value functions was studied in [20, 21]. This gave PAC learning type sample complexity bounds for MDPs and this can be combined with policy improvement along the lines of optimistic policy iteration.

This paper is organized as follows. In Section 2, we discuss preliminaries and briefly talk about classical value and policy iteration. Section 3 presents empirical value and policy iteration. Section 4 introduces the notion of random operators and relevant notions of probabilistic fixed points. In this section, we also develop a stochastic dominance argument for convergence analysis of iteration of random operators when they satisfy certain assumptions. In Section 5, we show that the empirical Bellman operator satisfies the above assumptions, and present sample complexity and convergence rate estimates for the EDP algorithms. Section 6 provides various extensions including asychronous EDP, minmax EDP and EDP for the dynamic newsvendor problem. Basic numerical experiments are reported in Section 7.

2. Preliminaries We first introduce a typical representation for a discrete time MDP as the 5-tuple

$$(\mathbb{S},\, \mathbb{A},\, \{A(s)\, :\, s\in\mathbb{S}\},\, Q,\, c)\,.$$

Both the state space $\mathbb{S}$ and the action space $\mathbb{A}$ are finite. Let $\mathcal{P}(\mathbb{S})$ denote the space of probability measures over $\mathbb{S}$ and we define $\mathcal{P}(\mathbb{A})$ similarly. For each state $s\in\mathbb{S}$, the set $A(s)\subset\mathbb{A}$ is the set of feasible actions. The entire set of feasible state-action pairs is

$$\mathbb{K}\triangleq\{(s,a)\in\mathbb{S}\times\mathbb{A}\,:\,a\in A(s)\}\,.$$

The transition law Q governs the system evolution, $Q(\cdot|s,a)\in\mathcal{P}(\mathbb{A})$ for all $(s,\,a)\in\mathbb{K}$, i.e., $Q(j|s,a)$ for $j\in\mathbb{S}$ is the probability of visiting the state j next given the current state-action pair $(s,\,a)$. Finally, $c:\mathbb{K}\to\mathbb{R}$ is a cost function that depends on state-action pairs.

Let Π denote the class of *stationary deterministic Markov policies*, i.e., mappings $\pi:\mathbb{S}\to\mathbb{A}$ which only depend on history through the current state. We only consider such policies since it is well known that there is an optimal policy in this class. For a given state $s\in\mathbb{S}$, $\pi(s)\in A(s)$ is the action chosen in state s under the policy π. We assume that Π only contains feasible policies that respect the constraints $\mathbb{K}$. The state and action at time t are denoted s_t and a_t, respectively. Any policy $\pi\in\Pi$ and initial state $s\in\mathbb{S}$ determine a probability measure P_s^π and a stochastic process $\{(s_t,a_t)\,,\,t\geq 0\}$ defined on the canonical measurable space of trajectories of state-action pairs. The expectation operator with respect to P_s^π is denoted $\mathbb{E}_s^\pi[\cdot]$.

We will focus on infinite horizon discounted cost MDPs with discount factor $\alpha\in(0,\,1)$. For a given initial state $s\in\mathbb{S}$, the expected discounted cost for policy $\pi\in\Pi$ is denoted by

$$v^\pi(s)=\mathbb{E}_s^\pi\left[\sum_{t=0}^{\infty}\alpha^t c(s_t,a_t)\right].$$

The optimal cost starting from state s is denoted by

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

and $v^*\in\mathbb{R}^{|\mathbb{S}|}$ denotes the corresponding optimal value function in its entirety.

Value iteration The Bellman operator $T : \mathbb{R}^{|\mathbb{S}|} \to \mathbb{R}^{|\mathbb{S}|}$ is defined as

$$[T\,v]\,(s) \triangleq \min_{a \in A(s)} \left\{ c\,(s,a) + \alpha\,\mathbb{E}\left[v\,(\tilde{s})\,|s,a\right] \right\},\ \forall s \in \mathbb{S},$$

for any $v \in \mathbb{R}^{|\mathbb{S}|}$, where $\tilde{s}$ is the random next state visited, and

$$\mathbb{E}\left[v\,(\tilde{s})\,|s,a\right] = \sum_{j \in \mathbb{S}} v\,(j)\,Q\,(j|s,a)$$

is the explicit computation of the expected cost-to-go conditioned on state-action pair $(s,a) \in \mathbb{K}$. Value iteration amounts to iteration of the Bellman operator. We have a sequence $\left\{v^k\right\}_{k \geq 0} \subset \mathbb{R}^{|\mathbb{S}|}$ where $v^{k+1} = T\,v^k = T^{k+1}v^0$ for all $k \geq 0$ and an initial seed v^0. This is the well-known value iteration algorithm for dynamic programming.

We next state the Banach fixed point theorem which is used to prove that value iteration converges. Let U be a Banach space with norm $\|\cdot\|_U$. We call an operator $G : U \to U$ a *contraction mapping* when there exists a constant $\kappa \in [0,\,1)$ such that

$$\|G\,v_1 - G\,v_2\|_U \leq \kappa\,\|v_1 - v_2\|_U, \forall v_1,\,v_2 \in U.$$

THEOREM 2.1. *(Banach fixed point theorem) Let U be a Banach space with norm $\|\cdot\|_U$, and let $G : U \to U$ be a contraction mapping with constant $\kappa \in [0,\,1)$. Then,*
(i) there exists a unique $v^* \in U$ such that $G\,v^* = v^*$;
(ii) for arbitrary $v^0 \in U$, the sequence $v^k = G\,v^{k-1} = G^k v^0$ converges in norm to v^* as $k \to \infty$;
(iii) $\|v^{k+1} - v^*\|_U \leq \kappa\,\|v^k - v^*\|_U$ for all $k \geq 0$.

For the rest of the paper, let $\mathcal{C}$ denote the space of contraction mappings from $\mathbb{R}^{|\mathbb{S}|} \to \mathbb{R}^{|\mathbb{S}|}$. It is well known that the Bellman operator $T \in \mathcal{C}$ with constant $\kappa = \alpha$ is a contraction operator, and hence has a unique fixed point v^*. It is known that value iteration converges to v^* as $k \to \infty$. In fact, v^* is the optimal value function.

Policy iteration Policy iteration is another well known dynamic programming algorithm for solving MDPs. For a fixed policy $\pi \in \Pi$, define $T_\pi : \mathbb{R}^{|\mathbb{S}|} \to \mathbb{R}^{|\mathbb{S}|}$ as

$$[T_\pi v]\,(s) = c\,(s,\pi\,(s)) + \alpha\,\mathbb{E}\left[v\,(\tilde{s})\,|s,\pi\,(s)\right].$$

The first step is a *policy evalution* step. Compute v^π by solving $T_\pi v^\pi = v^\pi$ for v^π. Let $c^\pi \in \mathbb{R}^{|\mathbb{S}|}$ be the vector of one period costs corresponding to a policy π, $c^\pi\,(s) = c\,(s,\pi\,(s))$ and Q^π, the transition kernel corresponding to the policy π. Then, writing $T_\pi v^\pi = v^\pi$ we have the linear system

$$c^\pi + \alpha Q^\pi v^\pi = v^\pi. \quad \text{(Policy Evaluation)}$$

The second step is a *policy improvement* step. Given a value function $v \in \mathbb{R}^{|\mathbb{S}|}$, find an 'improved' policy $\pi \in \Pi$ with respect to v such that

$$T_\pi v = T\,v. \quad \text{(Policy Update)}$$

Thus, policy iteration produces a sequence of policies $\left\{\pi^k\right\}_{k \geq 0}$ and $\left\{v^k\right\}_{k \geq 0}$ as follows. At iteration $k \geq 0$, we solve the linear system $T_{\pi^k} v^{\pi^k} = v^{\pi^k}$ for v^{π^k}, and then we choose a new policy π^k satisfying

$$T_{\pi^k} v^{\pi^k} = T\,v^{\pi^k},$$

which is greedy with respect to v^{π^k}. We have a linear convergence rate for policy iteration as well. Let $v \in \mathbb{R}^{|\mathbb{S}|}$ be any value function, solve $T_\pi v = T\,v$ for π, and then compute v^π. Then, we know [9, Lemma 6.2] that

$$\|v^\pi - v^*\| \leq \alpha\,\|v - v^*\|,$$

from which convergence of policy iteration follows. Unless otherwise specified, the norm $||\cdot||$ we will use in this paper is the sup norm.

We use the following helpful fact in the paper. Proof is given in Appendix A.1.

REMARK 2.1. Let X be a given set, and $f_1 : X \to \mathbb{R}$ and $f_2 : X \to \mathbb{R}$ be two real-valued functions on X. Then,

(i) $|\inf_{x\in X} f_1(x) - \inf_{x\in X} f_2(x)| \leq \sup_{x\in X} |f_1(x) - f_2(x)|$, and

(ii) $|\sup_{x\in X} f_1(x) - \sup_{x\in X} f_2(x)| \leq \sup_{x\in X} |f_1(x) - f_2(x)|$.

3. Empirical Algorithms for Dynamic Programming We now present empirical variants of dynamic programming algorithms. Our focus will be on value and policy iteration. As the reader will see, the idea is simple and natural. In subsequent sections we will introduce the new notions and techniques to prove their convergence.

3.1. Empirical Value Iteration We introduce empirical value iteration (EVI) first.

The Bellman operator T requires exact evaluation of the expectation

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

We will simulate and replace this exact expectation with an empirical estimate in each iteration. Thus, we need a simulation model for the MDP. Let

$$\psi : \mathbb{S}\times\mathbb{A}\times[0,1] \to \mathbb{S}$$

be a simulation model for the state evolution for the MDP, i.e. ψ yields the next state given the current state, the action taken and an i.i.d. random variable. Without loss of generality, we can assume that ξ is a uniform random variable on $[0,1]$ and $(s,a) \in \mathbb{K}$. With this convention, the Bellman operator can be written as

$$[Tv](s) \triangleq \min_{a\in A(s)} \{c(s,a) + \alpha\,\mathbb{E}[v(\psi(s,a,\xi))]\},\ \forall s \in \mathbb{S}.$$

Now, we replace the expectation $\mathbb{E}[v(\psi(s,a,\xi))]$ with its sample average approximation by simulating ξ. Given n i.i.d. samples of a uniform random variable, denoted $\{\xi_i\}_{i=1}^n$, the empirical estimate of $\mathbb{E}[v(\psi(s,a,\xi))]$ is $\frac{1}{n}\sum_{i=1}^n v(\psi(s,a,\xi_i))$. We note that the samples are regenerated at each iteration. Thus, the EVI algorithm can be summarized as follows.

Algorithm 1 Empirical Value Iteration

Input: $v^0 \in \mathbb{R}^{|\mathbb{S}|}$, sample size $n \geq 1$.

Set counter $k = 0$.

1. Sample n uniformly distributed random variables $\left\{\xi_i^k\right\}_{i=1}^n$ from [0, 1], and compute

$$v_n^{k+1}(s) = \min_{a\in A(s)} \left\{ c(s,a) + \frac{\alpha}{n}\sum_{i=1}^n v_n^k\left(\psi\left(s,a,\xi_i^k\right)\right)\right\},\ \forall s\in\mathbb{S}.$$

2. Increment $k := k+1$ and return to step 1.

In each iteration, we regenerate samples and use this empirical estimate to approximate T. Now we give the sample complexity of the EVI algorithm. Proof is given in Section 5.

THEOREM 3.1. *Given $\epsilon \in (0,1)$ and $\delta \in (0,1)$, fix $\epsilon_g = \epsilon/\eta^*$ and select $\delta_1, \delta_2 > 0$ such that $\delta_1 + 2\delta_2 \leq \delta$ where $\eta^* = \lceil 2/(1-\alpha) \rceil$. Select an n such that*

$$n \geq n(\epsilon, \delta) = \frac{2\left(\kappa^*\right)^2}{\epsilon_g^2} \log \frac{2|\mathbb{K}|}{\delta_1}$$

where $\kappa^* = \max_{(s,a)\in\mathbb{K}} c(s,a)/(1-\alpha)$ and select a k such that

$$k \geq k(\epsilon, \delta) = \log\left(\frac{1}{\delta_2\, \mu_{n,\,\min}}\right),$$

where $\mu_{n,min} = \min_\eta \mu_n(\eta)$ and $\mu_n(\eta)$ is given by Lemma 4.3. Then

$$\mathcal{P}\left\{\|\hat{v}_n^k - v^*\| \geq \epsilon\right\} \leq \delta.$$

REMARK 3.1. This result says that, if we take $n \geq n(\epsilon, \delta)$ samples in each iteration of the EVI algorithm and perform $k > k(\epsilon, \delta)$ iterations then the EVI iterate $\hat{v}_n^k$ is ϵ close to the optimal value function v^* with probability greater that $1-\delta$. We note that the simulation sample complexity n has order $O\left(\frac{1}{\epsilon^2}, \log\frac{1}{\delta}, \log|\mathbb{S}|, \log|\mathbb{A}|\right)$. The total sample complexity can be obtained by multiplying this by $|\mathbb{S}|$.

The basic idea in the analysis is to frame EVI as an iteration of a random operator $\widehat{T}_n$ which we call the empirical Bellman operator. We define $\widehat{T}_n$ as

$$\left[\widehat{T}_n(\omega)\, v\right](s) \triangleq \min_{a\in A(s)} \left\{ c(s,a) + \frac{\alpha}{n}\sum_{i=1}^{n} v\left(\psi(s,a,\xi_i)\right)\right\}, \forall s \in \mathbb{S}. \tag{3.1}$$

This is a random operator because it depends on the random noise samples $\{\xi\}_{i=1}^n$. The definition and the analysis of this operator is done rigorously in Section 5.

3.2. Empirical Policy Iteration We now define EPI along the same lines by replacing exact policy improvement and evaluation with empirical estimates. For a fixed policy $\pi \in \Pi$, we can estimate $v^\pi(s)$ via simulation. Given a sequence of noise $\omega = (\xi_i)_{i\geq 0}$, we have $s_{t+1} = \psi(s_t, \pi(s_t), \xi_t)$ for all $t \geq 0$. For $\gamma > 0$, choose a finite horizon $\mathfrak{T}$ such that

$$\max_{(s,a)\in\mathbb{K}} |c(s,a)| \sum_{t=\mathfrak{T}+1}^{\infty} \alpha^t < \gamma.$$

We use the time horizon $\mathfrak{T}$ to truncate the simulation, since we must stop simulation after a finite time. Let

$$\left[\hat{v}^\pi(s)\right](\boldsymbol{\omega}) = \sum_{t=0}^{\mathfrak{T}} \alpha^t c\left(s_t(\boldsymbol{\omega}), \pi\left(s_t(\boldsymbol{\omega})\right)\right)$$

be the realization of $\sum_{t=0}^{\mathfrak{T}} \alpha^t c(s_t, a_t)$ on the sample path $\boldsymbol{\omega}$.

The next algorithm requires two input parameters, n and q, which determine sample sizes. Parameter n is the sample size for policy improvement and parameter q is the sample size for policy evaluation. We discuss the choices of these parameters in detail later. In the following algorithm, the notation $s_t(\omega_i)$ is understood as the state at time t in the simulated trajectory ω_i.

Step 2 replaces computation of $T_\pi v = T v$ (policy improvement). Step 3 replaces solution of the system $v = c^\pi + \alpha Q^\pi v$ (policy evaluation).

We now give the sample complexity result for EPI. Proof is given in Section 5.

Algorithm 2 Empirical Policy Iteration

Input: $\pi_0 \in \Pi$, $\epsilon > 0$.

1. Set counter $k = 0$.
2. For each $s \in \mathbb{S}$, draw $\omega_1, \ldots, \omega_q \in \Omega$ and compute

$$\hat{v}^{\pi_k}(s) = \frac{1}{q} \sum_{i=1}^{q} \sum_{t=0}^{\mathfrak{T}} \alpha^t c\left(s_t(\omega_i), \pi_k\left(s_t(\omega_i)\right)\right).$$

3. Draw $\xi_1, \ldots, \xi_n \in [0,1]$. Choose π_{k+1} to satisfy

$$\pi_{k+1}(s) \in \arg\min_{a \in A(s)} \left\{ c(s,a) + \frac{\alpha}{n} \sum_{i=1}^{n} \hat{v}^{\pi_k}\left(\psi(s,a,\xi_i)\right) \right\}, \forall s \in \mathbb{S}.$$

4. Increase $k := k+1$ and return to step 2.
5. Stop when $\|\hat{v}^{\pi_{k+1}} - \hat{v}^{\pi_k}\| \leq \epsilon$.

THEOREM 3.2. *Given $\epsilon \in (0,1)$, $\delta \in (0,1)$ select $\delta_1, \delta_2 > 0$ such that $\delta_1 + 2\delta_2 < \delta$. Also select $\delta_{11}, \delta_{12} > 0$ such that $\delta_{11} + \delta_{12} < \delta$. Then, select $\epsilon_1, \epsilon_2 > 0$ such that $\epsilon_g = \frac{\epsilon_2 + 2\alpha\epsilon_1}{(1-\alpha)}$ where $\epsilon_g = \epsilon/\eta^*, \eta^* = \lceil 2/(1-\alpha) \rceil$. Then, select a q and n such that*

$$q \geq q(\epsilon, \delta) = \frac{2(\kappa^*(\mathfrak{T}+1))^2}{(\epsilon_1 - \gamma)^2} \log \frac{2|\mathbb{S}|}{\delta_{11}}.$$

$$n \geq n(\epsilon, \delta) = \frac{2(\kappa^*)^2}{(\epsilon_2/\alpha)^2} \log \frac{2|\mathbb{K}|}{\delta_{12}}.$$

where $\kappa^* = \max_{(s,a) \in \mathbb{K}} c(s,a)/(1-\alpha)$, and select a k such that

$$k \geq k(\epsilon, \delta) = \log\left(\frac{1}{\delta_2 \, \mu_{n,q,\min}}\right),$$

where $\mu_{n,q,min} = \min_\eta \mu_{n,q}(\eta)$ and $\mu_{n,q}(\eta)$ is given by equation (5.6). Then,

$$\mathcal{P}\{\|\hat{v}^{\pi_k} - v^*\| \geq \epsilon\} \leq \delta.$$

REMARK 3.2. This result says that, if we do $q \geq q(\epsilon, \delta)$ simulation runs for empirical policy evaluation, $n \geq n(\epsilon, \delta)$ samples for empirical policy update and perform $k > k(\epsilon, \delta)$ iterations then the true value v^{π_k} of the policy π_k will be ϵ close to the optimal value function v^* with probability greater that $1 - \delta$. We note that q is $O\left(\frac{1}{\epsilon^2}, \log\frac{1}{\delta}, \log|\mathbb{S}|\right)$ and n is $O\left(\frac{1}{\epsilon^2}, \log\frac{1}{\delta}, \log|\mathbb{S}|, \log|\mathbb{A}|\right)$.

4. Iteration of Random Operators The empirical Bellman operator $\widehat{T}_n$ we defined in equation (3.1) is a random operator. When it operates on a vector, it yields a random vector. When $\widehat{T}_n$ is iterated, it produces a stochastic process and we are interested in the possible convergence of this stochastic process. The underlying assumption is that the random operator $\widehat{T}_n$ is an 'approximation' of a deterministic operator T such that $\widehat{T}_n$ converges to T (in a sense we will shortly make precise) as n increases. For example the empirical Bellman operator approximates the classical Bellman operator. We make this intuition mathematically rigorous in this section. The discussion in this section is not specific to the Bellman operator, but applies whenever a deterministic operator T is being approximated by an improving sequence of random operators $\{\widehat{T}_n\}_{n \geq 1}$.

4.1. Probabilistic Fixed Points of Random Operators In this subsection we formalize the definition of a random operator, denoted by $\widehat{T}_n$.

Since $\widehat{T}_n$ is a random operator, we need an appropriate probability space upon which to define $\widehat{T}_n$. So, we define the sample space $\Omega = [0,1]^\infty$, the $\sigma-$algebra $\mathcal{F} = \mathcal{B}^\infty$ where $\mathcal{B}$ is the inherited Borel $\sigma-$algebra on $[0,1]$, and the probability distribution P on Ω formed by an infinite sequence of uniform random variables. The primitive uncertainties on Ω are infinite sequences of uniform noise $\omega = (\xi_i)_{i\geq 0}$ where each ξ_i is an independent uniform random variable on $[0,1]$. We view $(\Omega, \mathcal{F}, \mathcal{P})$ as the appropriate probability space on which to define iteration of the random operators $\left\{\widehat{T}_n\right\}_{n\geq 1}$.

Next we define a composition of random operators, $\widehat{T}_n^k$, on the probability space $(\Omega^\infty, \mathcal{F}^\infty, \mathcal{P})$, for all $k \geq 0$ and all $n \geq 1$ where,

$$\widehat{T}_n^k(\boldsymbol{\omega})\, v = \widehat{T}_n(\omega_{k-1})\, \widehat{T}_n(\omega_{k-2}) \cdots \widehat{T}_n(\omega_0)\, v.$$

Note that $\boldsymbol{\omega} \in \Omega^\infty$ is an infinite sequence $(\omega_j)_{j\geq 0}$ where each $\omega_j = (\xi_{j,i})_{i\geq 0}$. Then we can define the iteration of $\widehat{T}_n$ with an initial seed $\hat{v}_n^0 \in \mathbb{R}^{|\mathbb{S}|}$ (we use the hat notation to emphasize that the iterates are random variables generated by the empirical operator) as

$$\widehat{v}_n^{k+1} = \widehat{T}_n \widehat{v}_n^k = \widehat{T}_n^k \widehat{v}_n^0 \tag{4.1}$$

Notice that we only iterate k for fixed n. The sample size n is constant in every stochastic process $\left\{\hat{v}_n^k\right\}_{k\geq 0}$, where $\hat{v}_n^k = \widehat{T}_n^k \hat{v}^0$, for all $k \geq 1$. For a fixed $\hat{v}_n^0$, we can view all $\hat{v}_n^k$ as measurable mappings from Ω^∞ to $\mathbb{R}^{|\mathbb{S}|}$ via the mapping $\hat{v}_n^k(\boldsymbol{\omega}) = \widehat{T}_n^k(\boldsymbol{\omega})\, \hat{v}_n^0$.

The relationship between the fixed points of the deterministic operator T and the probabilistic fixed points of the random operator $\left\{\widehat{T}_n\right\}_{n\geq 1}$ depends on how $\left\{\widehat{T}_n\right\}_{n\geq 1}$ approximates T. Motivated by the relationship between the classical and the empirical Bellman operator, we will make the following assumption.

ASSUMPTION 4.1. $\mathcal{P}\left(\lim_{n\to\infty} \|\widehat{T}_n v - T\, v\| \geq \epsilon\right) = 0\ \forall \epsilon > 0$ *and* $\forall v \in \mathbb{R}^{\|\mathbb{S}\|}$. *Also* T *has a (possibly non-unique) fixed point* v^* *such that* $T v^* = v^*$.

Assumption 4.1 is equivalent to $\lim_{n\to\infty} \widehat{T}_n(\omega)\, v = T\, v$ for $P-$almost all $\omega \in \Omega$. It is just strong law of large numbers when T is the Bellman operator and each $\hat{T}_n$, an empirical Bellman operator. Here, we benefit from defining all of the random operators $\left\{\widehat{T}_n\right\}_{n\geq 1}$ together on the sample space $\Omega = [0,1]^\infty$, so that the above convergence statement makes sense.

Strong probabilistic fixed point: We now introduce a natural probabilistic fixed point notion for $\left\{\widehat{T}_n\right\}_{n\geq 1}$, in analogy to the definition of a fixed point, $\|T\, v^* - v^*\| = 0$ for a deterministic operator.

DEFINITION 4.1. A vector $\hat{v} \in \mathbb{R}^{|\mathbb{S}|}$ is a strong probabilistic fixed point for the sequence $\left\{\widehat{T}_n\right\}_{n\geq 1}$ if

$$\lim_{n\to\infty} \mathcal{P}\left(\|\widehat{T}_n \hat{v} - \hat{v}\| > \epsilon\right) = 0,\ \forall \epsilon > 0.$$

We note that the above notion is defined for a sequence of random operators, rather than for a single random operator.

REMARK 4.1. We can give a slightly more general notion of a probabilistic fixed point which we call (ϵ, δ)-strong probablistic fixed point. For a *fixed* (ϵ, δ), we say that a vector $\hat{v} \in \mathbb{R}^{|\mathbb{S}|}$ is an (ϵ, δ)-strong probabilistic fixed point if there exists an $n_0(\epsilon, \delta)$ such that for all $n \geq n_0(\epsilon, \delta)$ we get $\mathcal{P}\left(\|\widehat{T}_n \hat{v} - \hat{v}\| > \epsilon\right) < \delta$. Note that, all strong probabilistic fixed points satisfy this condition for arbitrary (ϵ, δ) and hence are (ϵ, δ)-strong probabilistic fixed points. However the converse need

not be true. In many cases we may be looking for an ϵ-optimal 'solution' with a $1-\delta$ 'probabilistic guarantee' where (ϵ, δ) is fixed a priori. In fact, this would provide an approximation to the strong probabilistic fixed point of the sequence of operators.

Weak probabilistic fixed point: It is well known that iteration of a deterministic contraction operator converges to its fixed point. It is unclear whether a similar property would hold for random operators, and whether they would converge to the strong probabilistic fixed point of the sequence $\left\{\widehat{T}_n\right\}_{n\geq 1}$ in any way. Thus, we define an apparently weaker notion of a probabilistic fixed point that explicitly considers iteration.

DEFINITION 4.2. A vector $\hat{v} \in \mathbb{R}^{|\mathbb{S}|}$ is a weak probabilistic fixed point for $\left\{\widehat{T}_n\right\}_{n\geq 1}$ if

$$\lim_{n\to\infty} \limsup_{k\to\infty} \mathcal{P}\left(\|\widehat{T}_n^k v - \hat{v}\| > \epsilon\right) = 0, \ \forall \epsilon > 0, \ \forall v \in \mathbb{R}^{|\mathbb{S}|}$$

We use $\limsup_{k\to\infty} \mathcal{P}\left(\|\widehat{T}_n^k v - \hat{v}\| > \epsilon\right)$ instead of $\lim_{k\to\infty} \mathcal{P}\left(\|\widehat{T}_n^k v - \hat{v}\| > \epsilon\right)$ because the latter limit may not exist for any fixed $n \geq 1$.

REMARK 4.2. Similar to the definition that we gave in Remark 4.1, we can define an (ϵ, δ)-weak probablistic fixed point. For a *fixed* (ϵ, δ), we say that a vector $\hat{v} \in \mathbb{R}^{|\mathbb{S}|}$ is an (ϵ, δ)-weak probabilistic fixed point if there exists an $n_0(\epsilon, \delta)$ such that for all $n \geq n_0(\epsilon, \delta)$ we get $\limsup_{k\to\infty} \mathcal{P}\left(\|\widehat{T}_n^k v - \hat{v}\| > \epsilon\right) < \delta$. As before, all weak probabilistic fixed points are indeed (ϵ, δ) weak probabilistic fixed points, but converse need not be true.

At this point the connection between strong/weak probabilistic fixed points of the random operator $\widehat{T}_n$ and the classical fixed point of the deterministic operator T is not clear. Also it is not clear whether the random sequence $\{\hat{v}_n^k\}_{k\geq 0}$ converges to either of these two fixed point. In the following subsections we address these issues.

4.2. A Stochastic Process on $\mathbb{N}$ In this subsection, we construct a new stochastic process on $\mathbb{N}$ that will be useful in our analysis. We first start with a simple lemma.

LEMMA 4.1. *The stochastic process $\left\{\hat{v}_n^k\right\}_{k\geq 0}$ is a Markov chain on $\mathbb{R}^{|\mathbb{S}|}$.*

Proof: This follows from the fact that each iteration of $\widehat{T}_n$ is independent, and identically distributed. Thus, the next iterate $\hat{v}_n^{k+1}$ only depends on history through the current iterate $\hat{v}_n^k$. □

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

Since we are approximating T by $\widehat{T}_n$ and want to compute v^*, we should track the progress of $\left\{\hat{v}_n^k\right\}_{k\geq 0}$ to the fixed point v^* of T. Equivalently, we are interested in the real-valued stochastic process $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$. If $\|\hat{v}_n^k - v^*\|$ approaches zero then $\hat{v}_n^k$ approaches v^*, and vice versa.

The state space of the stochastic process $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$ is $\mathbb{R}$, which is simpler than the state space $\mathbb{R}^{|\mathbb{S}|}$ of $\left\{\hat{v}_n^k\right\}_{k\geq 0}$, but which is still continuous. Moreover, $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$ is a non-Markovian process in general. In fact it would be easier to study a related stochastic process on a discrete, and ideally a finite state space. In this subsection we show how this can be done.

We make a boundedness assumption next which is of course satisfied when per-stage rewards are bounded.

ASSUMPTION 4.2. *There exists a $\kappa^* < \infty$ such that $\|\hat{v}_n^k\| \leq \kappa^*$ almost surely for all $k \geq 0$, $n \geq 1$. Also, $\|v^*\| \leq \kappa^*$.*

Under this assumption we can restrict the stochastic process $\left\{\|\hat{v}^k - v^*\|\right\}_{k\geq 0}$ to the compact state space

$$\overline{B_{2\kappa^*}(0)} = \left\{v \in \mathbb{R}^{|\mathbb{S}|} : \|v\| \leq 2\kappa^*\right\}.$$

We will adopt the convention that any element v outside of $\overline{B_{\kappa^*}(0)}$ will be mapped to its projection $\kappa^* \frac{v}{\|v\|}$ onto $\overline{B_{\kappa^*}(0)}$ by any realization of $\widehat{T}_n$.

Choose a **granularity** $\epsilon_g > 0$ to be fixed for the remainder of this discussion. We will break up $\mathbb{R}$ into intervals of length ϵ_g starting at zero, and we will note which interval is occupied by $\|\hat{v}_n^k - v^*\|$ at each $k \geq 0$. We will define a new stochastic process $\left\{X_n^k\right\}_{k\geq 0}$ on $(\Omega^\infty, \mathcal{F}^\infty, \mathcal{P})$ with state space $\mathbb{N}$. The idea is that $\left\{X_n^k\right\}_{k\geq 0}$ will report which interval of $[0,\, 2\,\kappa^*]$ is occupied by $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$. Define $X_n^k : \Omega^\infty \to \mathbb{N}$ via the rule:

$$X_n^k(\boldsymbol{\omega}) = \begin{cases} 0, & \text{if } \|\hat{v}^k(\boldsymbol{\omega}) - v^*\| = 0, \\ \eta \geq 1, & \text{if } (\eta - 1)\,\epsilon_g < \|\hat{v}^k(\boldsymbol{\omega}) - v^*\| \leq \eta\,\epsilon_g, \end{cases} \tag{4.2}$$

for all $k \geq 0$. More compactly,

$$X_n^k(\boldsymbol{\omega}) = \left\lceil \|\hat{v}_n^k(\boldsymbol{\omega}) - v^*\|/\epsilon_g \right\rceil,$$

where $\lceil \chi \rceil$ denotes the smallest integer greater than or equal to $\chi \in \mathbb{R}$. Thus the stochastic process $\left\{X_n^k\right\}_{k\geq 0}$ is a report on how close the stochastic process $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$ is to zero, and in turn how close the Markov chain $\left\{\hat{v}_n^k\right\}_{k\geq 0}$ is to the true fixed point v^* of T.

Define the constant

$$N^* \triangleq \left\lceil \frac{2\,\kappa^*}{\epsilon_g} \right\rceil.$$

Notice N^* is the smallest number of intervals of length ϵ_g needed to cover the interval $[0, 2\,\kappa^*]$. By construction, the stochastic process $\left\{X_n^k\right\}_{k\geq 0}$ is restricted to the finite state space $\left\{\eta \in \mathbb{N} : 0 \leq \eta \leq N^*\right\}$.

The process $\left\{X_n^k\right\}_{k\geq 0}$ need not be a Markov chain. However, it is easier to work with than either $\left\{\hat{v}_n^k\right\}_{k\geq 0}$ or $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$ because it has a discrete state space. It is also easy to relate $\left\{X_n^k\right\}_{k\geq 0}$ back to $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$.

Recall that $X \geq_{as} Y$ denotes almost sure inequality between two random variables X and Y defined on the same probability space. The stochastic processes $\left\{X_n^k\right\}_{k\geq 0}$ and $\left\{\|\hat{v}_n^k - v^*\|/\epsilon_g\right\}_{k\geq 0}$ are defined on the same probability space, so the next lemma follows by construction of $\left\{X_n^k\right\}_{k\geq 0}$.

LEMMA 4.2. *For all $k \geq 0$, $X_n^k \geq_{as} \|\hat{v}_n^k - v^*\|/\epsilon_g$.*

To proceed, we will make the following assumptions about the deterministic operator T and the random operator $\widehat{T}_n$.

ASSUMPTION 4.3. *$\|T\,v - v^*\| \leq \alpha\,\|v - v^*\|$ for all $v \in \mathbb{R}^{|\mathbb{S}|}$.*

This is equivalent to a contraction assumption that is of course satisfied by the Bellman operator.

ASSUMPTION 4.4. *There is a sequence $\{p_n\}_{n\geq 1}$ such that*

$$P\left(\|Tv - \widehat{T}_n v\| < \epsilon\right) > p_n(\epsilon)$$

and $p_n(\epsilon) \uparrow 1$ as $n \to \infty$ for all $v \in \overline{B_{\kappa^*}(0)}$, $\forall \epsilon > 0$.

This assumption requires that the one-iteration error in the empirical operator goes to zero in probability as n goes to ∞.

We now discuss the convergence rate of $\left\{X_n^k\right\}_{k\geq 0}$. Let $X_n^k = \eta$. On the event $F = \left\{\|T\,\hat{v}_n^k - \widehat{T}_n\hat{v}_n^k\| < \epsilon_g\right\}$, we have

$$\|\hat{v}_n^{k+1} - v^*\| \leq \|\widehat{T}_n\hat{v}_n^k - T\,\hat{v}_n^k\| + \|T\,\hat{v}_n^k - v^*\| \leq (\alpha\,\eta + 1)\,\epsilon_g$$

where we used Assumption 4.3 and the definition of X_n^k. Now using Assumption 4.4 we can summarize:

$$\text{If } X_n^k = \eta, \text{ then } X_n^{k+1} \leq \lceil \alpha\,\eta + 1\rceil \text{ with a probability at least } p_n(\epsilon_g). \tag{4.3}$$

We conclude this subsection with a comment about the state space of the stochastic process of $\{X_n^k\}_{k\geq 0}$. If we start with $X_n^k = \eta$ and if $\lceil \alpha\,\eta + 1\rceil < \eta$ then we must have improvement in the proximity of $\hat{v}_n^{k+1}$ to v^*. We define a new constant

$$\eta^* = \min\{\eta \in \mathbb{N} :\ \lceil \alpha\,\eta + 1\rceil < \eta\} = \left\lceil \frac{2}{1-\alpha} \right\rceil.$$

If η is too small, then $\lceil \alpha\,\eta + 1\rceil$ may be equal to η and no improvement in the proximity of $\hat{v}_n^k$ to v^* can be detected by $\left\{X_n^k\right\}_{k\geq 0}$. For any $\eta \geq \eta^*$, $\lceil \alpha\,\eta + 1\rceil < \eta$ and strict improvement must hold. So, for the stochastic process $\left\{X_n^k\right\}_{k\geq 0}$, we can restrict our attention to the state space $\mathcal{X} := \{\eta^*, \eta^* + 1, \ldots, N^* - 1, N^*\}$.

4.3. Dominating Markov Chains If we could understand the behavior of the stochastic processes $\left\{X_n^k\right\}_{k\geq 0}$, then we could make statements about the convergence of $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$ and $\left\{\hat{v}_n^k\right\}_{k\geq 0}$. Although simpler than $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$ and $\left\{\hat{v}_n^k\right\}_{k\geq 0}$, the stochastic process $\left\{X_n^k\right\}_{k\geq 0}$ is still too complicated to work with analytically. We overcome this difficulty with a family of dominating Markov chains. We now present our dominance argument. Several technical details are expanded upon in the appendix.

We will denote our family of "dominating" Markov chains (MC) by $\left\{Y_n^k\right\}_{k\geq 0}$. We will construct these Markov chains to be tractable and to help us analyze $\left\{X_n^k\right\}_{k\geq 0}$. Notice that the family $\left\{Y_n^k\right\}_{k\geq 0}$ has explicit dependence on $n \geq 1$. We do not necessarily construct $\left\{Y_n^k\right\}_{k\geq 0}$ on the probability space $(\Omega^\infty, \mathcal{F}^\infty, \mathcal{P})$. Rather, we view $\left\{Y_n^k\right\}_{k\geq 0}$ as being defined on $(\mathbb{N}^\infty, \mathcal{N})$, the canonical measurable space of trajectories on $\mathbb{N}$, so $Y_n^k : \mathbb{N}^\infty \to \mathbb{N}$. We will use $\mathcal{Q}$ to denote the probability measure of $\left\{Y_n^k\right\}_{k\geq 0}$ on $(\mathbb{N}^\infty, \mathcal{N})$. Since $\left\{Y_n^k\right\}_{k\geq 0}$ will be a Markov chain by construction, the probability measure $\mathcal{Q}$ will be completely determined by an initial distribution on $\mathbb{N}$ and a transition kernel. We denote the transition kernel of $\left\{Y_n^k\right\}_{k\geq 0}$ as $\mathfrak{Q}_n$.

Our specific choice for $\left\{Y_n^k\right\}_{k\geq 0}$ is motivated by analytical expediency, though the reader will see that many other choices are possible. We now construct the process $\left\{Y_n^k\right\}_{k\geq 0}$ explicitly, and then compute its steady state probabilities and its mixing time. We will define the stochastic process $\left\{Y_n^k\right\}_{k\geq 0}$ on the finite state space $\mathcal{X}$, based on our observations about the boundedness of $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$ and $\left\{X_n^k\right\}_{k\geq 0}$. Now, for a fixed n and $p_n(\epsilon_g)$ (we drop the argument ϵ_g in the following for notational convenience) as assumed in Assumption 4.4 we construct the dominating Markov chain $\left\{Y_n^k\right\}_{k\geq 0}$ as:

$$Y_n^{k+1} = \begin{cases} \max\left\{Y_n^k - 1,\, \eta^*\right\}, & \text{w.p. } p_n, \\ N^*, & \text{w.p. } 1 - p_n. \end{cases} \tag{4.4}$$

The first value $Y_n^{k+1} = \max\left\{Y_n^k - 1,\, \eta^*\right\}$ corresponds to the case where the approximation error satisfies $\|T\,\hat{v}^k - \widehat{T}_n\hat{v}^k\| < \epsilon_g$, and the second value $Y_n^{k+1} = N^*$ corresponds to all other cases (giving us an extremely conservative bound in the sequel). This construction also ensures that $Y_n^k \in \mathcal{X}$ for

all k, $\mathcal{Q}-$almost surely. Informally, $\left\{Y_n^k\right\}_{k\geq 0}$ will either move one unit closer to zero until it reaches η^*, or it will move (as far away from zero as possible) to N^*.

We now summarize some key properties of $\left\{Y_n^k\right\}_{k\geq 0}$.

PROPOSITION 4.1. *For $\left\{Y_n^k\right\}_{k\geq 0}$ as defined above,*
(i) it is a Markov chain;
(ii) the steady state distribution of $\left\{Y_n^k\right\}_{k\geq 0}$, and the limit $Y_n =_d \lim_{k\to\infty} Y_n^k$, exists;
(iii) $\mathcal{Q}\left\{Y_n^k > \eta\right\} \to \mathcal{Q}\left\{Y_n > \eta\right\}$ as $k\to\infty$ for all $\eta\in\mathbb{N}$;

Proof: Parts (i) - (iii) follow by construction of $\left\{Y_n^k\right\}_{k\geq 0}$ and the fact that this family consists of irreducible Markov chains on a finite state space. □

We now describe a stochastic dominance relationship between the two stochastic processes $\left\{X_n^k\right\}_{k\geq 0}$ and $\left\{Y_n^k\right\}_{k\geq 0}$. The notion of stochastic dominance (in the usual sense) will be central to our development.

DEFINITION 4.3. Let X and Y be two real-valued random variables, then Y *stochastically dominates* X, written $X\leq_{st} Y$, when $\mathbb{E}\left[f\left(X\right)\right]\leq\mathbb{E}\left[f\left(Y\right)\right]$ for all increasing functions $f:\mathbb{R}\to\mathbb{R}$. The condition $X\leq_{st} Y$ is known to be equivalent to

$$\mathbb{E}\left[\mathbf{1}\left\{X\geq\theta\right\}\right]\leq\mathbb{E}\left[\mathbf{1}\left\{Y\geq\theta\right\}\right] \quad \text{or} \quad \Pr\left\{X\geq\theta\right\}\leq\Pr\left\{Y\geq\theta\right\},$$

for all θ in the support of Y. Notice that the relation $X\leq_{st} Y$ makes no mention of the respective probability spaces on which X and Y are defined - these spaces may be the same or different (in our case they are different).

Let $\left\{\mathcal{F}^k\right\}_{k\geq 0}$ be the filtration on $(\Omega^\infty,\mathcal{F}^\infty,\mathcal{P})$ corresponding to the evolution of information about $\left\{X_n^k\right\}_{k\geq 0}$. Let $[X_n^{k+1}|\mathcal{F}^k]$ denote the conditional distribution of X_n^{k+1} given the information $\mathcal{F}^k$. The following theorem compares the marginal distributions of $\left\{X_n^k\right\}_{k\geq 0}$ and $\left\{Y_n^k\right\}_{k\geq 0}$ at all times $k\geq 0$ when the two stochastic processes $\left\{X_n^k\right\}_{k\geq 0}$ and $\left\{Y_n^k\right\}_{k\geq 0}$ start from the same state.

THEOREM 4.1. *If $X_n^0 = Y_n^0$, then $X_n^k\leq_{st} Y_n^k$ for all $k\geq 0$.*

Proof is given in Appendix A.2

The following corollary resulting from Theorem 4.1 relates the stochastic processes $\left\{\|\hat{v}_n^k - v^*\|\right\}_{k\geq 0}$, $\left\{X_n^k\right\}_{k\geq 0}$, and $\left\{Y_n^k\right\}_{k\geq 0}$ in a probabilistic sense, and summarizes our stochastic dominance argument.

COROLLARY 4.1. *For any fixed $n\geq 1$, we have*
(i)$\mathcal{P}\left\{\|\hat{v}_n^k - v^*\| > \eta\epsilon_g\right\}\leq\mathcal{P}\left\{X_n^k > \eta\right\}\leq\mathcal{Q}\left\{Y_n^k > \eta\right\}$ for all $\eta\in\mathbb{N}$ for all $k\geq 0$;
(ii) $\limsup_{k\to\infty}\mathcal{P}\left\{X_n^k > \eta\right\}\leq\mathcal{Q}\left\{Y_n > \eta\right\}$ for all $\eta\in\mathbb{N}$;
(iii) $\limsup_{k\to\infty}\mathcal{P}\left\{\|\hat{v}_n^k - v^*\| > \eta\,\epsilon_g\right\}\leq\mathcal{Q}\left\{Y_n > \eta\right\}$ for all $\eta\in\mathbb{N}$.

Proof: (i) The first inequality is true by construction of X_n^k. Then $\mathcal{P}\left\{X_n^k > \eta\right\}\leq\mathcal{Q}\left\{Y_n^k > \eta\right\}$ for all $k\geq 0$ and $\eta\in\mathbb{N}$ by Theorem 4.1.
(ii) Since $\mathcal{Q}\left\{Y_n^k > \eta\right\}$ converges (by Proposition 4.1), the result follows by taking the limit in part (i).
(iii) This again follows by taking limit in part (i) and using Proposition 4.1 □

We now compute the steady state distribution of the Markov chain $\{Y_n^k\}_{k\geq 0}$. Let μ_n denote the steady state distribution of $Y_n =_d \lim_{k\to\infty} Y_n^k$ (whose existence is guaranteed by Proposition 4.1) where $\mu_n\left(i\right) = \mathcal{Q}\left\{Y_n = i\right\}$ for all $i\in\mathcal{X}$. The next lemma follows from standard techniques (see [38] for example). Proof is given in Appendix A.3.

LEMMA 4.3. *For any fixed $n\geq 1$, the values of $\{\mu_n\left(i\right)\}_{i=\eta^*}^{N^*}$ are*

$$\mu_n\left(\eta^*\right) = p_n^{N^*-\eta^*}, \quad \mu_n\left(N^*\right) = 1-p_n, \quad \mu_n\left(i\right) = \left(1-p_n\right)p_n^{\binom{N^*-i}{}}, \quad \forall i = \eta^*+1,\ldots,N^*-1.$$

Note that an explicit expression for p_n in the case of empirical Bellman operator is given in equation (5.1).

4.4. Convergence Analysis of Random Operators We now give results on the convergence of the stochastic process $\{\hat{v}_n^k\}$, which could equivalently be written $\{\widehat{T}_n^k \hat{v}_0\}_{k\geq 0}$. Also we elaborate on the connections between our different notions of fixed points. Throughout this section, v^* denotes the fixed point of the deterministic operator T as defined in Assumption 4.1.

THEOREM 4.2. *Suppose the random operator $\widehat{T}_n$ satisfies the assumptions 4.1 - 4.4. Then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \limsup_{k\to\infty} \mathcal{P}\left(\|\hat{v}_n^k - v^*\| > \epsilon\right) = 0,$$

i.e. v^* is a weak probabilistic fixed point of $\left\{\widehat{T}_n\right\}_{n\geq 1}$.

Proof: Choose the granularity $\epsilon_g = \epsilon/\eta^*$. From Corollary 4.1 and Lemma 4.3,

$$\limsup_{k\to\infty} \mathcal{P}\left\{\|\hat{v}_n^k - v^*\| > \eta^*\epsilon_g\right\} \leq \mathcal{Q}\left\{Y_n > \eta^*\right\} = 1 - \mu_n\left(\eta^*\right) = 1 - p_n^{N^*-\eta^*}$$

Now by Assumption 4.4, $p_n \uparrow 1$ and by taking limit on both sides of the above inequality gives the desired result. □

Now we show that a strong probabilistic fixed point and the deterministic fixed point v^* coincide under Assumption 4.1.

PROPOSITION 4.2. *Suppose Assumption **4.1** holds. Then,*
(i) v^* is a strong probabilistic fixed point of the sequence $\left\{\widehat{T}_n\right\}_{n\geq 1}$.
(ii) Let $\hat{v}$ be a strong probabilistic fixed point of the sequence $\left\{\widehat{T}_n\right\}_{n\geq 1}$, then $\hat{v}$ is a fixed point of T.

Proof is given in Appendix A.4.

Thus the set of fixed points of T and the set of strong probabilistic fixed points of $\{\widehat{T}_n\}_{n\geq 1}$ coincide. This suggests that a "probabilistic" fixed point would be an "approximate" fixed point of the deterministic operator T.

We now explore the connection between weak probabilistic fixed points and classical fixed points.

PROPOSITION 4.3. *Suppose the random operator $\widehat{T}_n$ satisfies the assumptions 4.1 - 4.4. Then,*
(i) v^* is a weak probabilistic fixed point of the sequence $\left\{\widehat{T}_n\right\}_{n\geq 1}$.
(ii) Let $\hat{v}$ be a weak probabilistic fixed point of the sequence $\left\{\widehat{T}_n\right\}_{n\geq 1}$, then $\hat{v}$ is a fixed point of T.

Proof is given Appendix A.5. It is obvious that we need more assumptions to analyze the asymptotic behavior of the iterates of the random operator $\widehat{T}_n$ and establish the connection to the fixed point of the deterministic operator.

We summarize the above discussion in the following theorem.

THEOREM 4.3. *Suppose the random operator $\widehat{T}_n$ satisfies the assumptions 4.1 - 4.4. Then the following three statements are equivalent:*
(i) v is a fixed point of T,
(ii) v is a strong probabilistic fixed point of $\left\{\widehat{T}_n\right\}_{n\geq 1}$,
(iii) v is a weak probabilistic fixed point of $\left\{\widehat{T}_n\right\}_{n\geq 1}$.

This is quite remarkable because we see not only that the two notions of a probabilistic fixed point of a sequence of random operators coincide, but in fact they coincide with the fixed point of the related classical operator. Actually, it would have been disappointing if this were not the case. The above result now suggests that the iteration of a random operator a finite number k of times and for a fixed n would yield an approximation to the classical fixed point with high probability. Thus, the notions of the (ϵ, δ)-strong and weak probabilistic fixed points coincide asymptotically, however, we note that non-asymptotically they need not be the same.

5. Sample Complexity for EDP In this section we present the proofs of the sample complexity results for empirical value iteration (EVI) and policy iteration (EPI) (Theorem 3.1 and Theorem 3.2 in Section 3).

5.1. Empirical Bellman Operator Recall the definition of the empirical Bellman operator in equation (3.1). Here we give a mathematical basis for that definition which will help us to analyze the convergence behaviour of EVI (since EVI can be framed as an iteration of this operator).

The empirical Bellman operator is a random operator, because it maps random samples to operators. Recall from Section 4.1 that we define the random operator on the sample space $\Omega = [0,1]^{\infty}$ where primitive uncertainties on Ω are infinite sequences of uniform noise $\omega = (\xi_i)_{i\geq 0}$ where each ξ_i is an independent uniform random variable on $[0,1]$. This convention, rather than just defining $\Omega = [0,1]^n$ for a fixed $n \geq 1$, makes convergence statements with respect to n easier to make.

Classical value iteration is performed by iterating the Bellman operator T. Our EVI algorithm is performed by choosing n and then iterating the random operator $\widehat{T}_n$. So we follow the notations introduced in Section 4.1 and the kth iterate of EVI, $\hat{v}_n^k$ is given by $\hat{v}_n^k = \widehat{T}_n^k \hat{v}_n^0$ where $\hat{v}_n^0 \in \mathbb{R}^{|\mathbb{S}|}$ be an initial seed for EVI.

We first show that the empirical Bellman operator satisfies the Assumptions 4.1 - 4.4. Then the analysis follows the results of Section 4.

PROPOSITION 5.1. *The Bellman operator T and the empirical Bellman operator $\widehat{T}_n$ (defined in equation (3.1)) satisfy Assumptions 4.1 - 4.4*

Proof is given in Appendix A.6.

We note that we can explicitly give an expression for $p_n(\epsilon)$ (of Assumption 4.4) as below. For proof, refer to Proposition 5.1:

$$P\left\{\|\widehat{T}_n v - T v\| < \epsilon\right\} > p_n(\epsilon) := 1 - 2\,|\mathbb{K}|\, e^{-2\,(\epsilon/\alpha)^2 n/\left(2\,\kappa^*\right)^2}. \tag{5.1}$$

Also we note that we can also give an explicit expression for κ^* of in Assumption 4.2 as

$$\kappa^* \triangleq \frac{\max_{(s,a)\in\mathbb{K}} |c\,(s,a)\,|}{1-\alpha}. \tag{5.2}$$

For proof, refer to the proof of Proposition 5.1.

5.2. Empirical Value Iteration Here we use the results of Section 4 for analyzing the convergence of EVI. We first give an asymptotic result.

PROPOSITION 5.2. *For any $\delta_1 \in (0,1)$ select n such that*

$$n \geq \frac{2\,(\kappa^*)^2}{(\epsilon_g/\alpha)^2} \log \frac{2|\mathbb{K}|}{\delta_1}$$

then,

$$\limsup_{k\to\infty} \mathcal{P}\left\{\|\hat{v}_n^k - v^*\| > \eta^* \epsilon_g\right\} \leq 1 - \mu_n(\eta^*) \leq \delta_1$$

Proof: From Corollary 4.1, $\limsup_{k\to\infty} \mathcal{P}\left\{\|\hat{v}_n^k - v^*\| > \eta^* \epsilon_g\right\} \leq \mathcal{Q}\left\{Y_n > \eta^*\right\} = 1 - \mu_n\left(\eta^*\right)$. For $1 - \mu_n\left(\eta^*\right)$ to be less that δ_1, we compute n using Lemma 4.3 as,

$$1 - \delta_1 \leq \mu_n(\eta^*) = p_n^{N^* - \eta^*} \leq p_n = 1 - 2\,|\mathbb{K}|\, e^{-2\,(\epsilon_g/\alpha)^2 n/\left(2\,\kappa^*\right)^2}.$$

Thus, we get the desired result. □

We cannot iterate $\widehat{T}_n$ forever so we need a guideline for a finite choice of k. This question can be answered in terms of mixing times. The total variation distance between two probability measures μ and ν on $\mathbb{S}$ is

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

Let Q_n^k be the marginal distribution of Y_n^k on $\mathbb{N}$ at stage k and

$$d(k)=\|Q_n^k-\mu_n\|_{TV}$$

be the total variation distance between Q_n^k and the steady state distribution μ_n. For $\delta_2>0$, we define

$$t_{mix}(\delta_2)=\min\{k:d(k)\leq\delta_2\}$$

to be the minimum length of time needed for the marginal distribution of Y_n^k to be within δ_2 of the steady state distribution in total variation norm.

We now bound $t_{mix}(\delta_2)$.

LEMMA 5.1. *For any $\delta_2>0$,*

$$t_{mix}(\delta_2)\leq\log\left(\frac{1}{\epsilon\,\mu_{n,\min}}\right).$$

where $\mu_{n,\min}:=\min_\eta\mu_n(\eta)$.

Proof: Let $\mathfrak{Q}_n$ be the transition matrix of the Markov chain $\{Y_n^k\}_{k\geq0}$. Also let $\lambda_\star=\max\{|\lambda|:\lambda\text{ is an eigenvalue of }\mathfrak{Q}_n,\lambda\neq1\}$. By [27, Theorem 12.3],

$$t_{mix}(\delta_2)\leq\log\left(\frac{1}{\delta_2\,\mu_{n,\min}}\right)\frac{1}{1-\lambda_\star}$$

but $\lambda_\star=0$ by Lemma A.3 given in Appendix A.7. □

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

PROPOSITION 5.3. *For any fixed $n\geq1$, $\mathcal{P}\{\|\hat{v}_n^k-v^*\|>\eta^*\epsilon_g\}\leq2\delta_2+(1-\mu_n(\eta^*))$ for $k\geq\log\left(\frac{1}{\delta_2\mu_{n,\min}}\right)$.*

Proof: For $k\geq\log\left(\frac{1}{\delta_2\mu_{n,\min}}\right)\geq t_{\text{mix}}(\delta_2)$,

$$d(k)=\frac{1}{2}\sum_{i=\eta^*}^{N^*}|Q(Y_n^k=i)-\mu_n(i)|\leq\delta_2.$$

Then, $|Q(Y_n^k=\eta^*)-\mu_n(\eta^*)|\leq2d(t)\leq2\delta_2$. So, $\mathcal{P}\{\|\hat{v}_n^k-v^*\|>\eta^*\epsilon_g\}\leq Q(Y_n^k>\eta^*)=1-Q(Y_n^k=\eta^*)\leq2\delta_2+(1-\mu_n(\eta^*))$. □

We now combine Proposition 5.2 and 5.3 to prove Theorem 3.1.

Proof of Theorem 3.1:

Proof: Let $\epsilon_g=\epsilon/\eta^*$, and δ_1,δ_2 be positive with $\delta_1+2\delta_2\leq\delta$. By Proposition 5.2, for $n\geq n(\epsilon,\delta)$ we have

$$\limsup_{k\to\infty}\mathcal{P}\{\|\hat{v}_n^k-v^*\|>\epsilon\}=\limsup_{k\to\infty}\mathcal{P}\{\|\hat{v}_n^k-v^*\|>\eta^*\epsilon_g\}=\leq1-\mu_n(\eta^*)\leq\delta_1.$$

Now, for $k\geq k(\epsilon,\delta)$, by Proposition 5.3, $\mathcal{P}\{\|\hat{v}_n^k-v^*\|>\eta^*\epsilon_g\}=\mathcal{P}\{\|\hat{v}_n^k-v^*\|>\epsilon\}\leq2\delta_2+(1-\mu_n(\eta^*))$. Combining both we get, $\mathcal{P}\{\|\hat{v}_n^k-v^*\|>\epsilon\}\leq\delta$. □

5.3. Empirical Policy Iteration We now consider empirical policy iteration. EPI is different from EVI, and seemingly more difficult to analyze, because it does not correspond to iteration of a random operator. Furthermore, it has two simulation components, empirical policy evaluation and empirical policy update. However, we show that the convergence analysis in a manner similar to that of EVI.

We first give a sample complexity result for policy evaluation. For a policy π, let v^π be the actual value of the policy and let $\hat{v}_q^\pi$ be the empirical evaluation. Then,

PROPOSITION 5.4. *For any $\pi \in \Pi$, $\epsilon \in (0, \gamma)$ and for any $\delta > 0$*

$$P\left\{\|\hat{v}_q^\pi - v^\pi\| \geq \epsilon\right\} \leq \delta,$$

for

$$q \geq \frac{2(\kappa^*(\mathfrak{T}+1))^2}{(\epsilon-\gamma)^2} \log \frac{2|\mathbb{S}|}{\delta},$$

where $\hat{v}_q$ is evaluation of v^π by averaging q simulation runs.

Proof: Let $v^{\pi,\mathfrak{T}} := \mathbb{E}\left[\sum_{t=0}^{\mathfrak{T}} \alpha^t c\left(s_t\left(\boldsymbol{\omega}\right), \pi\left(s_t\left(\boldsymbol{\omega}\right)\right)\right)\right]$. Then,

$$\begin{aligned}
|\hat{v}_q^\pi(s) - v^\pi(s)| &\leq |\hat{v}_q^\pi(s) - v^{\pi,\mathfrak{T}}| + |v^{\pi,\mathfrak{T}} - v^\pi| \\
&\leq |\frac{1}{q}\sum_{i=1}^{q}\sum_{t=0}^{\mathfrak{T}} \alpha^k c\left(s_t\left(\boldsymbol{\omega}_i\right), \pi\left(s_t\left(\boldsymbol{\omega}_i\right)\right)\right) - v^{\pi,\mathfrak{T}}| + \gamma \\
&\leq \sum_{t=0}^{\mathfrak{T}} \left| \frac{1}{q}\sum_{i=1}^{q} \left(c\left(s_t\left(\boldsymbol{\omega}_i\right), \pi\left(s_t\left(\boldsymbol{\omega}_i\right)\right)\right) - \mathbb{E}\left[c\left(s_t\left(\boldsymbol{\omega}\right), \pi\left(s_t\left(\boldsymbol{\omega}\right)\right)\right)\right]\right) \right| + \gamma.
\end{aligned}$$

Then, with $\tilde{\epsilon} = (\epsilon - \gamma)/(\mathfrak{T}+1)$,

$$\begin{aligned}
P\left(|\hat{v}_q^\pi(s) - v^\pi(s)| \geq \epsilon\right) &\leq P\left(\left| \frac{1}{q}\sum_{i=1}^{q} \left(c\left(s_t\left(\boldsymbol{\omega}_i\right), \pi\left(s_t\left(\boldsymbol{\omega}_i\right)\right)\right) - \mathbb{E}\left[c\left(s_t\left(\boldsymbol{\omega}_i\right), \pi\left(s_t\left(\boldsymbol{\omega}_i\right)\right)\right)\right]\right) \right| \geq \tilde{\epsilon}\right) \\
&\leq 2e^{-2q\tilde{\epsilon}^2/(2\kappa^*)^2}.
\end{aligned}$$

By applying the union bound, we get

$$P\left(\|\hat{v}_q^\pi - v^\pi\| \geq \epsilon\right) \leq 2|\mathbb{S}|e^{-2q\tilde{\epsilon}^2/(2\kappa^*)^2}.$$

For $q \geq \frac{2(\kappa^*(\mathfrak{T}+1))^2}{(\epsilon-\gamma)^2} \log \frac{2|\mathbb{S}|}{\delta}$ the above probability is less than δ. □

We define

$$P\left(\|\hat{v}_q^\pi - v^\pi\| < \epsilon\right) > r_q(\epsilon) := 1 - 2|\mathbb{S}|e^{-2q\tilde{\epsilon}^2/(2\kappa^*)^2}, \text{ with } \tilde{\epsilon} = (\epsilon-\gamma)/(\mathfrak{T}+1). \tag{5.3}$$

We say that empirical policy evaluation is ϵ_1-accurate if $\|\hat{v}_q^\pi - v^\pi\| < \epsilon_1$. Then by the above proposition empirical policy evaluation is ϵ_1-accurate with a probability greater than $r_q(\epsilon_1)$.

The accuracy of empirical policy update compared to the actual policy update indeed depends on the empirical Bellman operator $\widehat{T}_n$. We say that empirical policy update is ϵ_2-accurate if $\|\widehat{T}_n v - Tv\| < \epsilon_2$. Then, by the definition of p_n in equation (5.1), our empirical policy update is ϵ_2-accurate with a probability greater than $p_n(\epsilon_2)$

We now give an important technical lemma. Proof is essentially a probabilistic modification of Lemmas 6.1 and 6.2 in [9] and is omitted.

LEMMA 5.2. *Let $\{\pi_k\}_{k\geq 0}$ be the sequence of policies from the EPI algorithm. For a fixed k, assume that $P\left(\|v^{\pi_k} - \hat{v}_q^{\pi_k}\| < \epsilon_1\right) \geq (1-\delta_1)$ and $P\left(\|T\hat{v}_q^{\pi_k} - \widehat{T}_n\hat{v}_q^{\pi_k}\| < \epsilon_2\right) \geq (1-\delta_2)$. Then,*

$$\|v^{\pi_{k+1}} - v^*\| \leq \alpha\|v^{\pi_k} - v^*\| + \frac{\epsilon_2 + 2\alpha\epsilon_1}{(1-\alpha)}$$

with probability at least $(1-\delta_1)(1-\delta_2)$.

We now proceed as in the analysis of EVI given in the previous subsection. Here we track the sequence $\{\|v^{\pi_k} - v^*\|\}_{k\geq 0}$. Note that this being a proof technique, the fact that the value $\|v^{\pi_k} - v^*\|$ is not observable does not affect our algorithm or its convergence behavior. We define

$$X_{n,q}^k = \lceil \|\hat{v}^{\pi_k} - v^*\|/\epsilon_g \rceil$$

where the granularity ϵ_g is fixed according to the problem parameters as $\epsilon_g = \frac{\epsilon_2 + 2\alpha\epsilon_1}{(1-\alpha)}$. Then by Lemma 5.2,

$$\text{if } X_{n,q}^k = \eta, \text{ then } X_{n,q}^{k+1} \leq \lceil \alpha\,\eta + 1 \rceil \text{ with a probability at least } p_{n,q} = r_q(\epsilon_1)p_n(\epsilon_2). \tag{5.4}$$

This is equivalent to the result for EVI given in equation (4.3). Hence the analysis is the same from here onwards. However, for completeness, we explicitly give the dominating Markov chain and its steady state distribution.

For $p_{n,q}$ given in display (5.4), we construct the dominating Markov chain $\left\{Y_{n,q}^k\right\}_{k\geq 0}$ as

$$Y_{n,q}^{k+1} = \begin{cases} \max\left\{Y_{n,q}^k - 1,\, \eta^*\right\}, & \text{w.p. } p_{n,q}, \\ N^*, & \text{w.p. } 1 - p_{n,q}, \end{cases} \tag{5.5}$$

which exists on the state space $\mathcal{X}$. The family $\left\{Y_{n,q}^k\right\}_{k\geq 0}$ is identical to $\left\{Y_n^k\right\}_{k\geq 0}$ except that its transition probabilities depend on n and q rather than just n. Let $\mu_{n,q}$ denote the steady state distribution of the Markov chain $\left\{Y_{n,q}^k\right\}_{k\geq 0}$. Then by Lemma 4.3,

$$\mu_{n,q}(\eta^*) = p_{n,q}^{N^*-\eta^*}, \mu_{n,q}(N^*) = 1 - p_{n,q},\ \mu_{n,q}(i) = (1-p_{n,q})\,p_{n,q}^{(N^*-i)}, \forall i = \eta^*+1, \ldots, N^*-1. \tag{5.6}$$

Proof of Theorem 3.2:

Proof: First observe that by the given choice of n and q, we have $r_q \geq (1-\delta_{11})$ and $p_n \geq (1-\delta_{12})$. Hence $1 - p_{n,q} \leq \delta_{11} + \delta_{12} - \delta_{11}\delta_{12} < \delta_1$.

Now by Corollary 4.1,

$$\limsup_{k\to\infty} \mathcal{P}\left\{\|v^{\pi_k} - v^*\| > \epsilon\right\} = \limsup_{k\to\infty} \mathcal{P}\left\{\|v^{\pi_k} - v^*\| > \eta^*\epsilon_g\right\} \leq \mathcal{Q}\left\{Y_{n,q} > \eta^*\right\} = 1 - \mu_{n,q}(\eta^*).$$

For $1 - \mu_{n,q}(\eta^*)$ to be less than δ_1 we need $1 - \delta_1 \leq \mu_n(\eta^*) = p_{n,q}^{N^*-\eta^*} \leq p_{n,q}$ which true as verified above. Thus we get

$$\limsup_{k\to\infty} \mathcal{P}\left\{\|v^{\pi_k} - v^*\| > \epsilon\right\} \leq \delta_1,$$

similar to the result of Proposition 5.2. Selecting the number of iterations k based on the mixing time is same as given in Proposition 5.3. Combining both as in the proof of Theorem 3.1 we get the desired result. □

Remarks. Some extensions: (i) We note that when the stage costs are random, and an expected cost term appears in the Bellman operator, we can replace that with a sample average empirical estimate of the expectation in the empirical Bellman operator using the same randomness. Then, the remaining proof argument goes through without much change. (ii) Another possible extension is when the same randomness is used to drive the simulation for multiple iterations, say p. This requires no change in the proof argument either since we can redefine the operator T to be p iterations of Bellman operator, and the empirical operator is then p iterations of the empirical Bellman operator with the same ω.

6. Variations and Extensions We now consider some non-trivial and useful variations and extensions of EVI.

6.1. Asynchronous Value Iteration The EVI algorithm described above is synchronous, meaning that the value estimates for every state are updated simultaneously. Here we consider each state to be visited at least once to complete a full update cycle. We modify the earlier argument to account for the possibly random time between full update cycles.

Classical asynchronous value iteration with exact updates has already been studied. Let $(x_k)_{k\geq 0}$ be any infinite sequence of states in $\mathbb{S}$. This sequence $(x_k)_{k\geq 0}$ may be deterministic or stochastic, and it may even depend online on the value function updates. For any $x \in \mathbb{S}$, we define the asynchronous Bellman operator $T_x : \mathbb{R}^{|\mathbb{S}|} \to \mathbb{R}^{|\mathbb{S}|}$ via

$$[T_x v](s) = \begin{cases} \min_{a\in A(s)} \{c(s,a) + \alpha \mathbb{E}[v(\psi(s,a,\xi))]\}, & s = x, \\ v(s), & s \neq x. \end{cases}$$

The operator T_x only updates the estimate of the value function for state x, and leaves the estimates for all other states exactly as they are. Given an initial seed $v^0 \in \mathbb{R}^{|\mathbb{S}|}$, asynchronous value iteration produces the sequence $\{v^k\}_{k\geq 0}$ defined by $v^k = T_{x_t} T_{x_{t-1}} \cdots T_{x_0} v^0$ for $k \geq 0$.

The following key properties of T_x are well known. See, for example, Bertsekas [8].

LEMMA 6.1. *For any $x \in \mathbb{S}$:*
(i) T_x is monotonic;
(ii) $T_x[v + \eta 1] = T_x v + \alpha \eta e_x$, where $e_x \in \mathbb{R}^{|\mathbb{S}|}$ be the unit vector corresponding to $x \in \mathbb{S}$.

The proof is omitted.

The next lemma is used to show that classical asynchronous VI converges. Essentially, a cycle of updates that visits every state at least once is a contraction.

LEMMA 6.2. *Let $(x_k)_{k=1}^K$ be any finite sequence of states such that every state in $\mathbb{S}$ appears at least once, then the operator*

$$\widetilde{T} = T_{x_1} T_{x_2} \cdots T_{x_K}$$

is a contraction with constant α.

It is known that asynchronous VI converges when each state is visited infinitely often. To continue, define $K_0 = 0$ and in general, we define

$$K_{m+1} \triangleq \inf \left\{ k : k \geq K_m, (x_i)_{i=K_m+1}^k \text{ includes every state in } \mathbb{S} \right\}.$$

Time K_1 is the first time that every state in $\mathbb{S}$ is visited at least once by the sequence $(x_k)_{k\geq 0}$. Time K_2 is the first time after K_1 that every state is visited at least once again by the sequence $(x_k)_{k\geq 0}$, etc. The times $\{K_m\}_{m\geq 0}$ completely depend on $(x_k)_{k\geq 0}$. For any $m \geq 0$, if we define

$$\widetilde{T} = T_{K_{m+1}} T_{K_{m+1}-1} \cdots T_{K_m+2} T_{K_m+1},$$

then we know

$$\|\widetilde{T} v - v^*\| \leq \alpha \|v - v^*\|,$$

by the preceding lemma. It is known that asynchronous VI converges under some conditions on $(x_k)_{k\geq 0}$.

THEOREM 6.1. *[8]. Suppose each state in $\mathbb{S}$ is included infinitely often by the sequence $(x_k)_{k\geq 0}$. Then $v^k \to v^*$.*

Algorithm 3 Asynchronous empirical value iteration

Input: $v^0 \in \mathbb{R}^{|\mathbb{S}|}$, sample size $n \geq 1$, a sequence $(x_k)_{k\geq 0}$.
Set counter $k = 0$.

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

$$\hat{v}^{k+1}(s) = \begin{cases} \min_{a\in A(s)} \left\{ c(s,a) + \frac{\alpha}{n} \sum_{i=1}^n \hat{v}^k (\psi(s,a,\xi_i)) \right\}, & s = x_k, \\ v(s), & s \neq x_k. \end{cases}$$

2. Increment $k := k+1$ and return to step 1.

Next we describe an empirical version of classical asynchronous value iteration. Again, we replace exact computation of the expectation with an empirical estimate, and we regenerate the sample at each iteration.

Step 1 of this algorithm replaces the exact computation $v^{k+1} = T_{x_k} v^k$ with an empirical variant. Using our earlier notation, we let $\widehat{T}_{x,n}$ be a random operator that only updates the value function for state x using an empirical estimate with sample size $n \geq 1$:

$$\left[\widehat{T}_{x,n}(\omega) v\right](s) = \begin{cases} \min_{a\in A(s)} \left\{ c(s,a) + \frac{1}{n} \sum_{i=1}^n v(\psi(s,a,\xi_i)) \right\}, & s = x, \\ v(s), & s \neq x. \end{cases}$$

We use $\left\{\hat{v}_n^k\right\}_{k\geq 0}$ to denote the sequence of asynchronous EVI iterates,

$$\hat{v}_n^{k+1} = \widehat{T}_{x_k,n} \widehat{T}_{x_{k-1},n} \cdots \widehat{T}_{x_0,n} \hat{v}_n^0,$$

or more compactly $\hat{v}_n^{k+1} = \widehat{T}_{x_k,n} \hat{v}_n^k$ for all $k \geq 0$.

We can use a slightly modified stochastic dominance argument to show that asynchronous EVI converges in a probabilistic sense. Only now we must account for the hitting times $\{K_m\}_{m\geq 0}$ as well, since the accuracy of the overall update depends on the accuracy in $\widehat{T}_{x,n}$ as well as the length of the interval $\{K_m + 1, K_m + 2, \ldots, K_{m+1}\}$. In asynchronous EVI, we will focus on $\{\hat{v}_n^{K_m}\}_{m\geq 0}$ rather than $\left\{\hat{v}_n^k\right\}_{k\geq 0}$. We check the progress of the algorithm at the end of complete update cycles.

In the simplest update scheme, we can order the states and then update them in the same order throughout the algorithm. The set $(x_k)_{k\geq 0}$ is deterministic in this case, and the intervals $\{K_m + 1, K_m + 2, \ldots, K_{m+1}\}$ all have the same length $|\mathbb{S}|$. Consider

$$\widetilde{T} = T_{x_{K_1},n} T_{x_{K_1-1},n} \cdots T_{x_1,n} T_{x_0,n},$$

the operator $\widehat{T}_{x_0,n}$ introduces ϵ error into component x_0, the operator $\widehat{T}_{x_1,n}$ introduces ϵ error into component x_1, etc. To ensure that

$$\widehat{T} = \widehat{T}_{x_{K_1},n} \widehat{T}_{x_{K_1-1},n} \cdots \widehat{T}_{x_1,n} \widehat{T}_{x_0,n}$$

is close to $\widetilde{T}$, we require each $\widehat{T}_{x_k,n}$ to be close to T_{x_k} for $k = 0, 1, \ldots, K_1 - 1, K_1$.

The following noise driven perspective helps with our error analysis. In general, we can view asynchronous empirical value iteration as

$$v' = T_x v + \varepsilon$$

for all $k \geq 0$ where

$$\varepsilon = \widehat{T}_{x,n} v - T_x v$$

is the noise for the evaluation of T_x (and it has at most one nonzero component).

Starting with v^0, define the sequence $\{v^k\}_{k\geq 0}$ by exact asynchronous value iteration $v^{k+1} = T_{x_k} v^k$ for all $k \geq 0$. Also set $\tilde{v}_0 := v_0$ and define

$$\tilde{v}^{k+1} = T_{x_k} \tilde{v}^k + \varepsilon_k$$

for all $k \geq 0$ where $\varepsilon_k \in \mathbb{R}^{|\mathbb{S}|}$ is the noise for the evaluation of T_{x_k} on $\tilde{v}^k$. In the following proposition, we compare the sequences of value functions $\{v^k\}_{k\geq 0}$ and $\{\tilde{v}^k\}_{k\geq 0}$ under conditions on the noise $\{\varepsilon_k\}_{k\geq 0}$.

PROPOSITION 6.1. *Suppose $-\eta 1 \leq \varepsilon_i \leq \eta 1$ for all $j = 0, 1, \ldots, k$ where $\eta \geq 0$ and $1 \in \mathbb{R}^{|\mathbb{S}|}$, i.e. the error is uniformly bounded for $j = 0, 1, \ldots, k$. Then, for all $j = 0, 1, \ldots, k$:*

$$v_j - \left(\sum_{i=0}^{j} \alpha^i\right) \eta 1 \leq \tilde{v}_j \leq v_j + \left(\sum_{i=0}^{j} \alpha^i\right) \eta 1.$$

Proof is given in Appendix A.8

Now we can use the previous proposition to obtain conditions for $\|\widetilde{T} v - \widehat{T} v\| < \epsilon$ (for our deterministic update sequence). Starting with the update for state x_0, we can choose n to ensure

$$\|T_{x_0,n} v - \widehat{T}_{x_0,n} v\| < \epsilon/|\mathbb{S}|$$

similar to that in equation (5.1). However, in this case our error bound is

$$\begin{aligned} P\left\{\|\widehat{T}_{x,n} v - T_x v\| \geq \epsilon/|\mathbb{S}|\right\} &\leq P\left\{\max_{a\in A(s)} |\frac{1}{n}\sum_{i=1}^{n} v\left(\psi\left(s,a,\xi_i\right)\right) - \mathbb{E}\left[v\left(\psi\left(s,a,\xi\right)\right)\right]| \geq \epsilon/(\alpha|\mathbb{S}|)\right\} \\ &\leq 2\,|\mathbb{A}|\, e^{-2\,(\epsilon/(\alpha|\mathbb{S}|))^2 n/\left(2\,\kappa^*\right)^2}, \end{aligned}$$

for all $v \in \mathbb{R}^{|\mathbb{S}|}$ (which does not depend on x). We are only updating one state, so we are concerned with the approximation of at most $|\mathbb{A}|$ terms $c(s,a) + \alpha \mathbb{E}\left[v\left(\psi\left(s,a,\xi\right)\right)\right]$ rather than $|\mathbb{K}|$. At the next update we want

$$\|\widehat{T}_{x_1,n} \hat{v}_n^1 - T_{x_1,n} \hat{v}_n^1\| < \epsilon/|\mathbb{S}|,$$

and we get the same error bound as above.

Based on this reasoning, assume

$$\|T_{x_k,n} \hat{v}_n^k - \widehat{T}_{x_k,n} \hat{v}_n^k\| < \epsilon/|\mathbb{S}|$$

for all $k = 0, 1, \ldots, K_1$. In this case we will in fact get the stronger error guarantee

$$\|\widetilde{T} v - \widehat{T} v\| < \frac{\epsilon}{|\mathbb{S}|} \sum_{i=0}^{|\mathbb{S}|-1} \alpha^i < \epsilon$$

from Proposition 6.1. The complexity estimates are multiplicative, so the probability $\|T_{x_k,n} \hat{v}_n^k - \widehat{T}_{x_k,n} \hat{v}_n^k\| < \epsilon/|\mathbb{S}|$ for all $k = 0, 1, \ldots, K_1$ is bounded above by

$$p_n = 2\,|\mathbb{S}|\,|\mathbb{A}|\, e^{-2\,(\epsilon/(\alpha|\mathbb{S}|))^2 n/\left(2\,\kappa^*\right)^2}.$$

To understand this result, remember that $|\mathbb{S}|$ iterations of asynchronous EVI amount to at most $|\mathbb{S}|\,|\mathbb{A}|$ empirical estimates of $c(s,a) + \mathbb{E}\left[v\left(\psi\left(s,a,\xi\right)\right)\right]$. We require all of these estimates to be within error $\epsilon/|\mathbb{S}|$.

We can take the above value for p_n and apply our earlier stochastic dominance argument to $\{\|\hat{v}_n^{K_m} - v^*\|\}_{m\geq 0}$, without further modification. This technique extends to any deterministic sequence $(x_k)_{k\geq 0}$ where the lengths of a full update for all states $|K_{m+1} - K_m|$ are uniformly bounded for all $m \geq 0$ (with the sample complexity estimate suitably adjusted).

6.2. Minimax Value Iteration Now we consider a two player zero sum Markov game and show how an empirical min-max value iteration algorithm can be used to a compute an approximate Markov Perfect equilibrium. Let the Markov game be described by the 7-tuple

$$(\mathbb{S},\, \mathbb{A},\, \{A(s) : s \in \mathbb{S}\},\, \mathbb{B},\, \{B(s) : s \in \mathbb{S}\},\, Q,\, c).$$

The action space $\mathbb{B}$ for player 2 is finite and $B(s)$ accounts for feasible actions for player 2. We let

$$\mathbb{K} = \{(s, a, b) : s \in \mathbb{S},\, a \in A(s),\, b \in B(s)\}$$

be the set of feasible station-action pairs. The transition law Q governs the system evolution, $Q(\cdot|s, a, b) \in \mathcal{P}(\mathbb{A})$ for all $(s, a) \in \mathbb{K}$, which is the probability of next visiting the state j given (s, a, b). Finally, $c : \mathbb{K} \to \mathbb{R}$ is the cost function (say of player 1) in state s for actions a and b. Player 1 wants to minimize this quantity, and player 2 is trying to maximize this quantity.

Let the operator T be defined as $T : \mathbb{R}^{|\mathbb{S}|} \to \mathbb{R}^{|\mathbb{S}|}$ is defined as

$$[Tv](s) \triangleq \min_{a \in A(s)} \max_{b \in B(s)} \{c(s, a, b) + \alpha\, \mathbb{E}[v(\tilde{s})\,|s, a, b]\},\ \forall s \in \mathbb{S},$$

for any $v \in \mathbb{R}^{|\mathbb{S}|}$, where $\tilde{s}$ is the random next state visited and

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

is the same expected cost-to-go for player 1. We call T the Shapley operator in honor of Shapley who first introduced it [41]. We can use T to compute the optimal value function of the same which is given by

$$v^*(s) = \max_{a \in A(s)} \min_{b \in B(s)} \{c(s, a, b) + \alpha\, \mathbb{E}[v^*(\tilde{s})\,|s, a, b]\},\ \forall s \in \mathbb{S},$$

is the optimal value function for player 1.

It is well known that that the Shapley operator is a contraction mapping.

LEMMA 6.3. *The Shapley operator T is a contraction.*

Proof is given in Appendix A.9 for completeness.

To compute v^*, we can iterate T. Pick any initial seed $v^0 \in \mathbb{R}^{\mathbb{S}}$, take $v^1 = Tv^0$, $v^2 = Tv^1$, and in general $v^{k+1} = Tv^k$ for all $k \geq 0$. It is known that [41] this procedure converges to the optimal value function. We refer to this as the classical minimax value iteration.

Now, using the simulation model $\psi : \mathbb{S} \times \mathbb{A} \times \mathbb{B} \times [0, 1] \to \mathbb{S}$, the empirical Shapley operator can be written as

$$[Tv](s) \triangleq \max_{a \in A(s)} \min_{b \in B(s)} \{c(s, a, b) + \alpha\, \mathbb{E}[v(\psi(s, a, b, \xi))]\},\ \forall s \in \mathbb{S},$$

where ξ is a uniform random variable on $[0, 1]$.

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

In each iteration, we take a new set of samples and use this empirical estimate to approximate T. Since T is a contraction with a known convergence rate α, we can apply the exact same development as for empirical value iteration.

Algorithm 4 Empirical value iteration for minimax

Input: $v^0 \in \mathbb{R}^{\mathbb{S}}$, sample size $n \geq 1$.
Set counter $k = 0$.

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

$$v^{k+1}(s) = \max_{a \in A(s)} \min_{b \in B(s)} \left\{ c(s,a,b) + \frac{\alpha}{n} \sum_{i=1}^{n} v^k(\psi(s,a,b,\xi_i)) \right\}, \forall s \in \mathbb{S}.$$

2. Increment $k := k+1$ and return to step 1.

6.3. The Newsvendor Problem We now show via the newsvendor problem that the empirical dynamic programming method can sometimes work remarkably well even for continuous states and action spaces. This, of course, exploits the linear structure of the newsvendor problem.

Let D be a continuous random variable representing the stationary demand distribution. Let $\{D_k\}_{k \geq 0}$ be independent and identically distributed collection of random variables with the same distribution as D, where D_k is the demand in period k. The unit order cost is c, unit holding cost is h, and unit backorder cost is b. We let x_k be the inventory level at the beginning of period k, and we let $q_k \geq 0$ be the order quantity before demand is realized in period k.

For technical convenience, we only allow stock levels in the compact set $\mathcal{X} = [x_{\min}, x_{\max}] \subset \mathbb{R}$. This assumption is not too restrictive, since a firm would not want a large number of backorders and any real warehouse has finite capacity. Notice that since we restrict to $\mathcal{X}$, we know that no order quantity will ever exceed $q_{\max} = x_{\max} - x_{\min}$. Define the continuous function $\psi : \mathbb{R} \to \mathcal{X}$ via

$$\psi(x) = \begin{cases} x_{\max}, & \text{if } x > x_{\max}, \\ x_{\min}, & \text{if } x < x_{\min}, \\ x, & \text{otherwise,} \end{cases}$$

The function ψ accounts for the state space truncation. The system dynamic is then

$$x_{k+1} = \psi(x_k + q_k - D_k), \forall k \geq 0.$$

We want to solve

$$\inf_{\pi \in \Pi} \mathbb{E}_\nu^\pi \left[\sum_{k=0}^{\infty} \alpha^k (c\, q_k + \max\{h\, x_k, -b\, x_k\}) \right], \tag{6.1}$$

subject to the preceding system dynamic. We know that there is an optimal stationary policy for this problem which only depends on the current inventory level. The optimal cost-to-go function for this problem, v^*, satisfies

$$v^*(x) = \inf_{q \geq 0} \{c\, q + \max\{h\, x, -b\, x\} + \mathbb{E}[v^*(\psi(x + q - D))]\}, \forall x \in \mathbb{R},$$

where, the optimal value function $v^* : \mathbb{R} \to \mathbb{R}$. We will compute v^* by iterating an appropriate Bellman operator.

Classical value iteration for Problem (6.1) consists of iteration of an operator in $\mathcal{C}(\mathcal{X})$, the space of continuous functions $f : \mathcal{X} \to \mathbb{R}$. We equip $\mathcal{C}(\mathcal{X})$ with the norm

$$\|f\|_{\mathcal{C}(\mathcal{X})} = \sup_{x \in \mathcal{X}} |f(x)|.$$

Under this norm, $\mathcal{C}(\mathcal{X})$ is a Banach space.

Now, the Bellman operator $T : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ for the newsvendor problem is given by

$$[T\,v](x) = \inf_{q \geq 0} \{c\,q + \max\{h\,x, -b\,x\} + \alpha\,\mathbb{E}[v(\psi(x+q-D))]\}, \forall x \in \mathcal{X}.$$

Value iteration for the newsvendor can then be written succinctly as $v^{k+1} = T\,v^k$ for all $k \geq 0$. We confirm that T is a contraction with respect to $\|\cdot\|_{\mathcal{C}(\mathcal{X})}$ in the next lemma, and thus the Banach fixed point theorem applies.

LEMMA 6.4. *(i) T is a contraction on $\mathcal{C}(\mathcal{X})$ with constant α.*
(ii) Let $\{v^k\}_{k\geq 0}$ be the sequence produced by value iteration, then $\lim_{k\to 0} \|v^k - v^*\|_{\mathcal{C}(\mathcal{X})} \to 0$.

Proof: (i) Choose $u, v \in \mathcal{C}(\mathcal{X})$, and use Fact 2.1 to compute

$$\begin{aligned}
\|T\,u - T\,v\|_{\mathcal{C}(\mathcal{X})} &= \sup_{x\in\mathcal{X}} |\,[T\,u](x) - [T\,v](x)\,| \\
&\leq \sup_{x\in\mathcal{X},\, q\in[0,q_{\max}]} \alpha\,|\mathbb{E}[u(\psi(x+q-D))] - \mathbb{E}[v(\psi(x+q-D))]\,| \\
&\leq \alpha \sup_{x\in\mathcal{X},\, q\in[0,q_{\max}]} \mathbb{E}[|u(\psi(x+q-D)) - v(\psi(x+q-D))\,|] \\
&\leq \alpha\,\|u - v\|_{\mathcal{C}(\mathcal{X})}.
\end{aligned}$$

(ii) Since $\mathcal{C}(\mathcal{X})$ is a Banach space and T is a contraction by part (i), the Banach fixed point theorem applies. □

Choose the initial form for the optimal value function as

$$v^0(x) = \max\{h\,x, -b\,x\}, \forall x \in \mathcal{X}.$$

It is chosen to represent the terminal cost in state x when there are no further ordering decisions. Then, value iteration yields

$$v^{k+1}(x) = \inf_{q\geq 0} \left\{c\,q + \max\{h\,x, -b\,x\} + \alpha\,\mathbb{E}\left[v^k(\psi(x+q-D))\right]\right\}, \forall x \in \mathcal{X}.$$

We note some key properties of these value functions.

LEMMA 6.5. *Let $\{v^k\}_{k\geq 0}$ be the sequence produced by value iteration, then v^k is Lipschitz continuous with constant $\max\{|h|, |b|\}\sum_{i=0}^{k}\alpha^i$ for all $k \geq 0$.*

Proof: First observe that v^0 is Lipschitz continuous with constant $\max\{|h|, |b|\}$. For v^1, we choose x and x' and compute

$$\begin{aligned}
|v^1(x) - v^1(x')\,| &\leq \sup_{q\geq 0} \{\max\{h\,x, -b\,x\} - \max\{h\,x', -b\,x'\} \\
&\quad + \alpha\left(\mathbb{E}\left[v^0(\psi(x+q-D))\right] - \mathbb{E}\left[v^0(\psi(x'+q-D))\right]\right)\} \\
&\leq \max\{|h|, |b|\}\,|x - x'| + \alpha\,\max\{|h|, |b|\}\,\mathbb{E}[|\psi(x+q-D) - \psi(x'+q-D)\,|] \\
&\leq \max\{|h|, |b|\}\,|x - x'| + \alpha\,\max\{|h|, |b|\}\,|x - x'|,
\end{aligned}$$

where we use the fact that the Lipschitz constant of ψ is one. The inductive step is similar. □

From Lemma 6.5, we also conclude that the Lipschitz constant of any iterate v^k is bounded above by

$$L^* \triangleq \max\{|h|, |b|\}\sum_{i=0}^{\infty}\alpha^i = \frac{\max\{|h|, |b|\}}{1-\alpha}.$$

We acknowledge the dependence of Lemma 6.5 on the specific choice of the initial seed, v^0.

We can do empirical value iteration with the same initial seed $\hat{v}_n^0 = v^0$ as above. Now, for $k \geq 0$,

$$\hat{v}_n^{k+1}(x) = \inf_{q \geq 0} \left\{ c\,q + \max\{h\,x, -b\,x\} + \frac{\alpha}{n} \sum_{i=1}^{n} \hat{v}_n^k(\psi(x+q-D_i)) \right\}, \forall x \in \mathbb{R}.$$

Note that $\{D_1, \ldots, D_n\}$ is an i.i.d. sample from the demand distribution. It is possible to perform these value function updates exactly for finite k based on [17]. Also note that, the initial seed is piecewise linear with finitely many breakpoints. Because the demand sample is finite in each iteration, thus each iteration will take a piecewise linear function as input and then produce a piecewise linear function as output (both with finitely many breakpoints). Lemma 6.5 applies without modification to $\left\{\hat{v}_n^k\right\}_{k \geq 0}$, all of these functions are Lipschitz continuous with constants bounded above by L^*.

As earlier, we define the empirical Bellman operator $\widehat{T}_n : \Omega \to \mathcal{C}$ as

$$\left[\widehat{T}_n(\omega)\, v\right](x) = \inf_{q \geq 0} \left\{ c\,q + \max\{h\,x, -b\,x\} + \frac{\alpha}{n} \sum_{i=1}^{n} v(\psi(x+q-D_i)) \right\}, \forall x \in \mathbb{R}.$$

With the empirical Bellman operator, we write the iterates of EVI as $\hat{v}_n^{k+1} = \widehat{T}_n^k v$.

We can again apply the stochastic dominance techniques we have developed to the convergence analysis of stochastic process $\left\{\|\hat{v}_n^k - v^*\|_{\mathcal{C}(\mathcal{X})}\right\}_{k \geq 0}$. Similarly to that of equation (5.2), we get an upper bound

$$\|v\|_{\mathcal{C}(\mathcal{X})} \leq \kappa^* \triangleq \frac{c\,q_{\max} + \max\{h\,x_{\max}, b\,x_{\min}\}}{1-\alpha}$$

for the norm of the value function of any policy for Problem (6.1). By the triangle inequality,

$$\|\hat{v}_n^k - v^*\|_{\mathcal{C}(\mathcal{X})} \leq \|\hat{v}_n^k\|_{\mathcal{C}(\mathcal{X})} + \|v^*\|_{\mathcal{C}(\mathcal{X})} \leq 2\,\kappa^*.$$

We can thus restrict $\|\hat{v}_n^k - v^*\|_{\mathcal{C}(\mathcal{X})}$ to the state space $[0,\, 2\,\kappa^*]$. For a fixed granularity $\epsilon_g > 0$, we can define $\left\{X_n^k\right\}_{k \geq 0}$ and $\left\{Y_n^k\right\}_{k \geq 0}$ as in Section 4. Our upper bound on probability follows.

PROPOSITION 6.2. *For any $n \geq 1$ and $\epsilon > 0$*

$$\begin{aligned} P\left\{\|\widehat{T}_n v - T\,v\| \geq \epsilon\right\} &\leq P\left\{\alpha \sup_{x \in \mathcal{X},\, q \in [0,\, q_{\max}]} |\mathbb{E}[v(\psi(x+q-D))] - \frac{1}{n}\sum_{i=1}^{n} v(\psi(x+q-D_i))| \geq \epsilon\right\} \\ &\leq 2\left\lceil \frac{9\,(L^*)^2\, q_{\max}^2}{\epsilon^2} \right\rceil e^{-2\,(\epsilon/3)^2 n / \left(2\,\|v\|_{\mathcal{C}(\mathcal{X})}\right)^2}, \end{aligned}$$

for all $v \in \mathcal{C}(\mathcal{X})$ with Lipschitz constant bounded by L^*.

Proof: By Fact 2.1, we know that

$$\|\widehat{T}_n v - T\,v\|_{\mathcal{C}(\mathcal{X})} \leq \alpha \sup_{x \in \mathcal{X},\, q \in [0,\, q_{\max}]} |\mathbb{E}[v(\psi(x+q-D))] - \frac{1}{n}\sum_{i=1}^{n} v(\psi(x+q-D_i))|.$$

Let $\{(x_j, q_j)\}_{j=1}^{J}$ be an $\epsilon/(3\,L^*)$ −net for $\mathcal{X} \times [0,\, q_{\max}]$. We can choose J to be the smallest integer greater than or equal to

$$\frac{x_{\max} - x_{\min}}{\epsilon/(3\,L^*)} \times \frac{q_{\max}}{\epsilon/(3\,L^*)} = \frac{9\,(L^*)^2\, q_{\max}^2}{\epsilon^2}.$$

If we have

$$|\mathbb{E}[v(\psi(x_j + q_j - D))] - \frac{1}{n}\sum_{i=1}^{n} v(\psi(x_j + q_j - D_i))| \leq \epsilon/3$$

for all $j = 1, \ldots, J$, then

$$|\mathbb{E}\left[v\left(\psi\left(x+q-D\right)\right)\right] - \frac{1}{n}\sum_{i=1}^{n} v\left(\psi\left(x+q-D_i\right)\right)| \leq \epsilon$$

for all $(x,q) \in \mathcal{X} \times [0, q_{\max}]$ by Lipschitz continuity and construction of $\{(x_j, q_j)\}_{j=1}^{J}$. Then, by Hoeffding's inequality and using union bound, we get,

$$P\left\{\alpha \sup_{j=1,\ldots,J} |\mathbb{E}\left[v\left(\psi\left(x_j+q_j-D\right)\right)\right] - \frac{1}{n}\sum_{i=1}^{n} v\left(\psi\left(x_j+q_j-D_i\right)\right)| \geq \epsilon/3\right\} \leq 2\,J\,e^{-2\,(\epsilon/3)^2 n/\left(2\,\|v\|_{\mathcal{C}(\mathcal{X})}\right)^2}.$$

□

As before, we use the preceding complexity estimate to determine p_n for the family $\left\{Y_n^k\right\}_{k\geq 0}$. The remainder of our stochastic dominance argument is exactly the same.

7. Numerical Experiments We now provide a numerical comparison of EDP methods with other methods for approximate dynamic programming via simulation. Figure 1 shows relative error ($||\hat{v}_n^k - v^*||$) of the Q-Learning algorithm, Optimistic Policy Iteration (OPI), Empirical Value Iteration (EVI), Empirical Policy Iteration (EPI) with exact Value and Policy Iteration. The MDP considered was a generic one with 1000 states, 10 actions with infinite-horizon discounted cost. For simulation-based algorithms, the number of samples in each step was $n = q = 10$. For EPI and OPI, 20 simulation runs were conducted while for EVI and QL, 50 simulation runs were conducted.The confidence intervals in each case are too small to be see in the Figure. The actor-critic algorithm was also run but it's convergence was found to be extremely slow and hence, not shown.

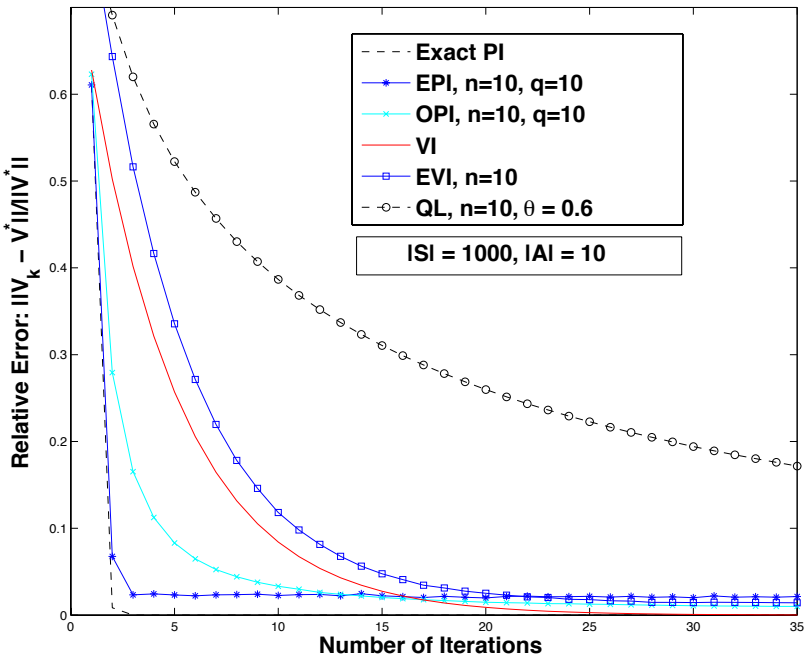

FIGURE 1. Numerical performance comparison of empirical value and policy iteration with actor-critic, Q-learning and optimistic policy iteration. Number of samples taken: $n = 10, q = 10$ (Refer EVI and EPI algorithms in Section 3 for the definition of n and q).

The experiments were peformed on a generic laptop with Intel Core i7 processor, 4GM RAM, on a 64-bit Windows 7 operating system via Matlab R2009b environment. For fair comparison, we only considered offline versions of Q-Learning, i.e., Q values for all state-action pairs were updated in a single iteration at the same time. From the figure, we see that EVI and EPI significantly outperform Q-Learning. In about 20 iterations, both EVI and EPI have less than 2% relative error, while Q-Learning still has more than 25% relative error. Optimistic policy iteration performs better than EVI since policy iteration-based algorithms are known to converge faster, but EPI outperforms OPI as well. Regarding computational load, we note that EVI performs less computational operations per iteration than exact VI, PI, EPI, OPI and Q-learning. In fact, exact PI would be the slowest of all for large state spaces since it needs to do matrix inversion for policy evaluation in each step.

These preliminary numerical results seem to suggest that EDP algorithms (in particular, EVI) outperform other ADP methods numerically, and hold good promise. In fact, in preliminary results, we notice similar numerical behavior for the asynchronous, or "online" setting. We would also like to point that EDP methods would very easily be parallelizable, and hence, they could potentially be useful for a wider variety of problem settings.

8. Conclusions In this paper, we have introduced a new class of algorithms for approximate dynamic programming. The idea isn't new, and actually quite simple and natural: just replace the expectation in the Bellman operator with an empirical estimate (or a sample average approximation, as it is often called.) The difficulty, however, is that it makes the Bellman operator a random operator. This makes its convergence analysis very challenging since (infinite horizon) dynamic programming theory is based on looking at the fixed points of the Bellman operator. However, the extant notions of 'probabilistic' fixed points for random operators are not relevant since they are akin to classical fixed points of deterministic monotone operators when ω is fixed. We introduce two notions of probabilistic fixed points - strong and weak. Furthermore, we show that these asymptotically coincide with the classical fixed point of the related deterministic operator, This is reassuring as it suggests that approximations to our probabilistic fixed points (obtained by finitely many iterations of the empirical Bellman operator) are going to be approximations to the classical fixed point of the Bellman operator as well.

In developing this theory, we also developed a mathematical framework based on stochastic dominance for convergence analysis of random operators. While our immediate goal was analysis of iteration of the empirical Bellman operator in empirical dynamic programming, the framework is likely of broader use, possibly after further development.

We have then shown that many variations and extensions of classical dynamic programming, work for empirical dynamic programming as well. In particular, empirical dynamic programming can be done asynchronously just as classical DP can be. Moreover, a zero-sum stochastic game can be solved by a minimax empirical dynamic program. We also apply the EDP method to the dynamic newsvendor problem which has continuous state and action spaces, which demonstrates the potential of EDP to solve problems over more general state and action spaces.

We have done some preliminary experimental performance analysis of EVI and EPI, and compared it to similar methods. Our numerical simulations suggest that EDP algorithms converge faster than stochastic approximation-based actor-critic, Q-learning and optimistic policy iteration algorithms. There is a need for a more extensive and careful numerical investigation of all such algorithms which will be done in the future.

We do note that EDP methods, unlike stochastic approximation methods, do not require any recurrence property to hold. In that sense, they are more universal. On the other hand, EDP algorithms would inherit some of the 'curse of dimensionality' problems associated with exact dynamic programming. Overcoming that challenge requires additional ideas, and is potentially a fruitful direction for future research. Some other directions of research are extending the EDP algorithms to the infinite-horizon average cost case, and to the partially-observed case. We will take up these issues in the future.

Acknowledgements The authors would like to thank Ugur Akyol (USC) wh o designed the original numerical experiments for this research. The authors would also like to thank Dimitris Bertsekas (MIT), Vivek Borkar (IIT Bombay), Peter Glynn (Stanford), Steve Marcus (Maryland), John Tsitsiklis (MIT), and Benjamin Van Roy (Stanford) for initial feedback on this work.

References

[1] Abounadi, J., D. Bertsekas, V. S. Borkar. 2001. Learning algorithms for markov decision processes with average cost. *SIAM J. Control Optim.* **40**(3) 681–698.

[2] Almudevar, A. 2008. Approximate fixed point iteration with an application to infinite horizon markov decision processes. *SIAM J. Control Optim.* **47**(5) 2303–2347.

[3] Anthony, M., P. Bartlett. 2009. *Neural network learning: Theoretical foundations*. cambridge university press.

[4] Barto, A., S. Sutton, P. Brouwer. 1981. Associative search network: A reinforcement learning associative memory. *Biol. Cybernet.* **40**(3) 201–211.

[5] Bellman, R. 1957. *Dynamic Programming*. Princeton University Press.

[6] Bellman, R., S. Dreyfus. 1959. Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation* **13**(68) 247–251.

[7] Bertsekas, D. 2011. Approximate policy iteration: A survey and some new methods. *J. Control Theory Appl.* **9**(3) 310–335.

[8] Bertsekas, D. 2012. *Dynamic programming and optimal control*, vol. 1 and 2. 4th ed.

[9] Bertsekas, D., J. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific.

[10] Borkar, V. S. 2002. Q-learning for risk-sensitive control. *Math. Oper. Res* **27**(2) 294–311.

[11] Borkar, V. S. 2008. *Stochastic approximation: A dynamical systems viewpoint*. Cambridge University Press, Cambridge.

[12] Borkar, V. S., S. P. Meyn. 2000. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.* **38**(2) 447–469.

[13] Chang, H. S., M. Fu, J. Hu, S. I. Marcus. 2006. *Simulation-Based Algorithms for Markov Decision Processes*. Springer, Berlin.

[14] Chang, H. S., M. Fu, J. Hu, S. I. Marcus. 2007. A survey of some simulation-based algorithms for markov decision processes. *Commun. Inf. Syst.* **7**(1) 59–92.

[15] Chueshov, I. 2002. *Monotone random systems theory and applications*, vol. 1779. Springer.

[16] Cooper, W., S. Henderson, M. Lewis. 2003. Convergence of simulation-based policy iteration. *Probab. Engrg. Inform. Sci.* **17**(02) 213–234.

[17] Cooper, W., B. Rangarajan. 2012. Performance guarantees for empirical markov decision processes with applications to multiperiod inventory models. *Oper. Res.* **60**(5) 1267–1281.

[18] Even-Dar, E., Y. Mansour. 2004. Learning rates for q-learning. *J. Mach. Learn. Res* **5** 1–25.

[19] Howard, R. 1971. *Dynamic Probabilistic Systems: Vol.: 2.: Semi-Markov and Decision Processes*. John Wiley and Sons.

[20] Jain, R., P. Varaiya. 2006. Simulation-based uniform value function estimates of markov decision processes. *SIAM J. Control Optim.* **45**(5) 1633–1656.

[21] Jain, R., P. Varaiya. 2010. Simulation-based optimization of markov decision processes: An empirical process theory approach. *Automatica* **46**(8) 1297–1304.

[22] Kakade, S. M. 2003. On the sample complexity of reinforcement learning. Ph.D. thesis, University of London.

[23] Kiefer, J., J. Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics* **23**(3) 462–466.

[24] Konda, V. R., V. S. Borkar. 1999. Actor-critic–type learning algorithms for markov decision processes. *SIAM Journal on control and Optimization* **38**(1) 94–123.

[25] Konda, V. R., J. Tsitsiklis. 2004. Convergence rate of linear two-time-scale stochastic approximation. *Ann. Appl. Probab.* 796–819.

[26] Kushner, H. J., D. S. Clark. 1978. *Stochastic approximation methods for constrained and unconstrained systems*. Springer.

[27] Levin, D., Y. Peres, E. L. Wilmer. 2009. *Markov chains and mixing times*. AMS.

[28] Ljung, L. 1977. Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Control* **22**(4) 551–575.

[29] Minsky, M. 1961. Steps toward artificial intelligence. *Proceedings of the IRE* **49**(1) 8–30.

[30] Müller, A., D. Stoyan. 2002. *Comparison methods for stochastic models and risks*, vol. 389. Wiley.

[31] Munos, R., C. Szepesvari. 2008. Finite time bounds for fitted value iteration. *J. Machine Learning Research* **1**.

[32] Narendra, K. S., M.A.L. Thathachar. 2012. *Learning automata: an introduction*. DoverPublications.

[33] Nemhauser, George L. 1966. *Introduction to dynamic programming*. Wiley.

[34] Papadimitriou, C. H., J. Tsitsiklis. 1987. The complexity of markov decision processes. *Math. Oper. Res* **12**(3) 441–450.

[35] Powell, W. B. 2011. *Approximate Dynamic Programming: Solving the curses of dimensionality*. 2nd ed.

[36] Rajaraman, K., P. S. Sastry. 1996. Finite time analysis of the pursuit algorithm for learning automata. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* **26**(4) 590–598.

[37] Robbins, H., S. Monro. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics* 400–407.

[38] Ross, Sheldon M. 1996. *Stochastic processes*. John Wiley & Sons.

[39] Rust, John. 1997. Using randomization to break the curse of dimensionality. *Econometrica* **65**(3) 487–516.

[40] Shaked, M., J. G. Shanthikumar. 2007. *Stochastic orders*. Springer.

[41] Shapley, Lloyd S. 1953. Stochastic games. *Proceedings Nat. Acad. of Sciences USA* **39**(10) 1095–1100.

[42] Shiryaev, A. 1995. *Probability*. Springer.

[43] Sutton, R. S., A. G. Barto. 1998. *Reinforcement learning: An introduction*. Cambridge Univ Press.

[44] Szepesvari, C. 2010. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan and Claypool Publishers.

[45] Thathachar, M.A. L., P. S. Sastry. 1985. A class of rapidly convergent algorithms for learning automata. *IEEE Tran. Sy. Man. Cyb.* **15** 168–175.

[46] Tsitsiklis, J. 2003. On the convergence of optimistic policy iteration. *J. Mach. Learn. Res* **3** 59–72.

[47] Watkins, C., P. Dayan. 1992. Q-learning. *Machine learning* **8**(3-4) 279–292.

[48] Werbos, P. 1974. Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. thesis.

[49] Whitt, W. 1978. Approximations of dynamic programs, i. *Math. Oper. Res* **3**(3) 231–243.

[50] Whitt, W. 1979. Approximations of dynamic programs, ii. *Math. Oper. Res* **4**(2) 179–185.

Appendix

A. Proofs of Various Lemmas, Propositions and Theorems

A.1. Proof of Fact 2.1

Proof: To verify part (i), note

$$\begin{aligned}\inf_{x\in X} f_1(x) &= \inf_{x\in X}\{f_1(x)+f_2(x)-f_2(x)\}\\ &\le \inf_{x\in X}\{f_2(x)+|f_1(x)-f_2(x)|\}\\ &\le \inf_{x\in X}\left\{f_2(x)+\sup_{y\in Y}|f_1(y)-f_2(y)|\right\}\\ &\le \inf_{x\in X} f_2(x)+\sup_{y\in Y}|f_1(y)-f_2(y)|,\end{aligned}$$

giving

$$\inf_{x\in X} f_1(x)-\inf_{x\in X} f_2(x)\le \sup_{x\in X}|f_1(x)-f_2(x)|.$$

By the same reasoning,

$$\inf_{x\in X} f_2(x)-\inf_{x\in X} f_1(x)\le \sup_{x\in X}|f_1(x)-f_2(x)|,$$

and the preceding two inequalities yield the desired result. Part (ii) follows similarly. □

A.2. Proof of Theorem 4.1 We first prove the following lemmas.

LEMMA A.1. *$[Y_n^{k+1}\,|\,Y_n^k=\theta]$ is stochastically increasing in θ for all $k\ge 0$, i.e. $[Y_n^{k+1}\,|\,Y_n^k=\theta]\le_{st}[Y_n^{k+1}\,|\,Y_n^k=\theta']$ for all $\theta\le\theta'$.*

Proof: We see that

$$\Pr\left\{Y_n^{k+1}\ge\eta\,|\,Y_n^k=\theta\right\}$$

is increasing in θ by construction of $\left\{Y_n^k\right\}_{k\ge0}$. If $\theta>\eta$, then $\Pr\left\{Y_n^{k+1}\ge\eta\,|\,Y_n^k=\theta\right\}=1$ since $Y_n^{k+1}\ge\theta-1$ almost surely; if $\theta\le\eta$, then $\Pr\left\{Y_n^{k+1}\ge\eta\,|\,Y_n^k=\theta\right\}=1-p_n$ since the only way Y_n^{k+1} will remain larger than η is if $Y_n^{k+1}=N^*$. □

LEMMA A.2. *$[X_n^{k+1}\,|\,X_n^k=\theta,\mathcal{F}^k]\le_{st}[Y_n^{k+1}|\,Y_n^k=\theta]$ for all θ and all $\mathcal{F}^k$ for all $k\ge0$.*

Proof: Follows from construction of $\left\{Y_n^k\right\}_{k\ge0}$. For any history $\mathcal{F}^k$,

$$\mathcal{P}\left\{X_n^{k+1}\ge\theta-1\,|\,X_n^k=\theta,\mathcal{F}^k\right\}\le\mathcal{Q}\left\{Y_n^{k+1}\ge\theta-1\,|\,Y_n^k=\theta\right\}=1.$$

Now,

$$\begin{aligned}\mathcal{P}\left\{X_n^{k+1}=N^*\,|\,X_n^k=\theta,\mathcal{F}^k\right\}&\le\mathcal{P}\left\{X_n^{k+1}>\theta-1\,|\,X_n^k=\theta,\mathcal{F}^k\right\}\\ &=1-P(X_n^{k+1}\le\theta-1|X_n^k=\theta)\\ &\le 1-p_n\end{aligned}$$

the last inequality follows because p_n is the worst case probability for a one-step improvement in the Markov chain $\{X_n^k\}_{k\ge0}$ □

Proof of Theorem 4.1

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

$$\mathfrak{Y}(\theta) = \begin{cases} \max\{\theta - 1, \eta^*\}, & \text{w.p. } p_n, \\ N^*, & \text{w.p. } 1 - p_n, \end{cases}$$

as a function of θ. We see that Y_n^{k+1} has the same distribution as

$$\left[\mathfrak{Y}(\Theta) | \Theta = Y_n^k\right]$$

by definition. Since $\mathfrak{Y}(\theta)$ are stochastically increasing, we see that

$$\left[\mathfrak{Y}(\Theta) | \Theta = Y_n^k\right] \geq_{st} \left[\mathfrak{Y}(\Theta) | \Theta = X_n^k\right]$$

by [40, Theorem 1.A.6] and our induction hypothesis. Now,

$$\left[\mathfrak{Y}(\Theta) | \Theta = X_n^k\right] \geq_{st} \left[X_n^{k+1} | X_n^k, \mathcal{F}^k\right]$$

by [40, Theorem 1.A.3(d)] for all histories $\mathcal{F}^k$. It follows that $Y_n^{k+1} \geq_{st} X_n^{k+1}$ by transitivity. □

A.3. Proof of Lemma 4.3

Proof: The stationary probabilities $\{\mu_n(i)\}_{i=\eta^*}^{N^*}$ satisfy the equations:

$$\begin{aligned} \mu_n(\eta^*) &= p_n \mu_n(\eta^*) + p_n \mu_n(\eta^* + 1), \\ \mu_n(i) &= p_n \mu_n(i+1), && \forall i = \eta^* + 1, \ldots, N^* - 1, \\ \mu_n(N^*) &= (1 - p_n) \sum_{i=\eta^*}^{N^*} \mu_n(i), \\ \sum_{i=\eta^*}^{N^*} \mu_n(i) &= 1. \end{aligned}$$

We then see that

$$\mu_n(i) = p_n^{(N^* - i)} \mu_n(N^*), \forall i = \eta^* + 1, \ldots, N^* - 1,$$

and

$$\mu_n(\eta^*) = \frac{p_n}{1 - p_n} \mu_n(\eta^* + 1) = \frac{p_n^{N^* - \eta^*}}{1 - p_n} \mu(N^*).$$

We can solve for $\mu_n(N^*)$ using $\sum_{i=\eta^*}^{N^*} \mu_n(i) = 1$,

$$\begin{aligned} 1 &= \sum_{i=\eta^*}^{N^*} \mu_n(i) \\ &= \frac{p_n^{N^* - \eta^*}}{1 - p_n} \mu_n(N^*) + \sum_{i=\eta^*+1}^{N^*} p_n^{N^* - i} \mu_n(N^*) \\ &= \left[\frac{p_n^{N^* - \eta^*}}{1 - p_n} + \sum_{i=\eta^*+1}^{N^*} p_n^{N^* - i}\right] \mu_n(N^*) \\ &= \left[\frac{p_n^{N^* - \eta^*}}{1 - p_n} + \frac{1 - p_n^{N^* - \eta^*}}{1 - p_n}\right] \mu_n(N^*), \\ &= \frac{1}{1 - p_n} \mu_n(N^*), \end{aligned}$$

based on the fact that

$$\sum_{i=\eta^*+1}^{N^*} p_n^{\left(N^*-i\right)} = \sum_{i=0}^{N^*-\eta^*-1} p_n^i = \frac{1-p_n^{\left(N^*-\eta^*\right)}}{1-p_n}.$$

We conclude

$$\mu\left(N^*\right) = 1-p_n,$$

and thus

$$\mu_n\left(i\right) = \left(1-p_n\right) p_n^{\left(N^*-i\right)}, \forall i = \eta^*+1,\ldots,N^*-1,$$

and

$$\mu_n\left(\eta^*\right) = p_n^{N^*-\eta^*}.$$

□

A.4. Proof of Proposition 4.2

Proof: (i) First observe that

$$\lim_{n\to\infty} \widehat{T}_n\left(\omega\right) v^* = T\, v^*,$$

by Assumption 4.1. It follows that $\widehat{T}_n\left(\omega\right) v^*$ converges to $v^* = T\, v^*$ as $n\to\infty$, $P-$almost surely. Almost sure convergence implies convergence in probability.
(ii) Let $\hat{v}$ be a strong probabilistic fixed point. Then,

$$P(\|T\hat{v}-\hat{v}\|\geq\epsilon) \leq P(\|T\hat{v}-\widehat{T}_n\hat{v}\|\geq\epsilon/2) + P(\|\widehat{T}_n\hat{v}-\hat{v}\|\geq\epsilon/2)$$

First term on the RHS can be made arbitrarily small by Assumption 4.1. Second term on RHS can also be made arbitrarily small by the definition of strong probabilistic fixed point. So, for sufficiently large n, we get $P(\|T\hat{v}-\hat{v}\|\geq\epsilon)<1$. Since the event in the LHS is deterministic, we get $\|T\hat{v}-\hat{v}\|=0$. Hence, $\hat{v}=v^*$.

□

A.5. Proof of Proposition 4.3

Proof: (i) This statement is proved in Theorem 4.2.

(ii) Fix the initial seed $v\in\mathbb{R}^{|\mathbb{S}|}$. For a contradiction, suppose $\hat{v}$ is not a fixed point of T so that $\|v^*-\hat{v}\|=\epsilon'>0$ (we use here the fact that v^* is unique). Now

$$\|\hat{v}-v^*\| = \epsilon' \leq \|\widehat{T}_n^k v-\hat{v}\| + \|\widehat{T}_n^k v-v^*\|$$

for any n and k by the triangle inequality. For clarity, this inequality holds in the almost sure sense:

$$\mathcal{P}\left(\epsilon' \leq \|\widehat{T}_n^k v-\hat{v}\| + \|\widehat{T}_n^k v-v^*\|\right) = 1$$

for all n and k.

We already know that

$$\lim_{n\to\infty}\limsup_{k\to\infty} \mathcal{P}\left(\|\widehat{T}_n^k v-v^*\|>\epsilon'/3\right) = 0$$

by Theorem 4.2 and

$$\lim_{n\to\infty}\limsup_{k\to\infty} \mathcal{P}\left(\|\widehat{T}_n^k v-\hat{v}\|>\epsilon'/3\right) = 0$$

by assumption. Now

$$\mathcal{P}\left(\max\left\{\|\widehat{T}_n^k v-\hat{v}\|,\, \|\widehat{T}_n^k v-v^*\|\right\}>\epsilon'/3\right) \leq \mathcal{P}\left(\|\widehat{T}_n^k v-v^*\|>\epsilon'/3\right) + \mathcal{P}\left(\|\widehat{T}_n^k v-\hat{v}\|>\epsilon'/3\right),$$

so

$$\lim_{n\to\infty}\limsup_{k\to\infty} \mathcal{P}\left(\max\left\{\|\widehat{T}_n^k v-\hat{v}\|,\, \|\widehat{T}_n^k v-v^*\|\right\}>\epsilon'/3\right) = 0.$$

However, $\epsilon' \leq \|\widehat{T}_n^k v-\hat{v}\| + \|\widehat{T}_n^k v-v^*\|$ almost surely so at least one of $\|\widehat{T}_n^k v-\hat{v}\|$ or $\|\widehat{T}_n^k v-v^*\|$ must be greater than $\epsilon'/3$ for all large k. □

A.6. Proof of Proposition 5.1

Proof: (i) Assumption 4.1 :
Certainly,

$$\|T_n(\omega)(v) - Tv\| \leq \alpha \max_{(s,a)\in\mathbb{K}} |\frac{1}{n}\sum_{i=1}^{n} v(\psi(s,a,\xi_i)) - \mathbb{E}[v(\psi(s,a,\xi))]|$$

using Fact (2.1). We know that for any fixed $(s,a)\in\mathbb{K}$,

$$|\frac{1}{n}\sum_{i=1}^{n} v(\psi(s,a,\xi_i)) - \mathbb{E}[v(\psi(s,a,\xi))]| \to 0,$$

as $n\to\infty$ by the Strong Law of Large Numbers (the random variable $v(\psi(s,a,\xi))$ has finite expectation because it is essentially bounded). Recall that $\mathbb{K}$ is finite to see that the right hand side of the above inequality converges to zero as $n\to\infty$.
(ii) Assumption 4.2 :
We define the constant

$$\kappa^* \triangleq \frac{\max_{(s,a)\in\mathbb{K}}|c(s,a)|}{1-\alpha}.$$

Then it can be easily verified that the value of any policy π $v^\pi \leq \kappa^*$. Then $v^* \leq \kappa^*$ and without loss of generality we can restrict $\hat{v}_n^k$ to the set $\overline{B_{2\kappa^*}(0)}$.
(iii) Assumption 4.3:
This is the well known contraction property of the Bellman operator.
(iv) Assumption 4.4:
Using Fact 2.1, for any fixed $s\in\mathbb{S}$,

$$|\widehat{T}_n v(s) - Tv(s)| \leq \max_{a\in A(s)} \left|\frac{\alpha}{n}\sum_{i=1}^{n} v(\psi(s,a,\xi_i) - \alpha\mathbb{E}[v(\psi(s,a,\xi)]\right|$$

and hence,

$$P\left\{\|\widehat{T}_n v - Tv\| \geq \epsilon\right\} \leq P\left\{\max_{(s,a)\in\mathbb{K}}\left|\frac{\alpha}{n}\sum_{i=1}^{n} v(\psi(s,a,\xi_i) - \alpha\mathbb{E}[v(\psi(s,a,\xi)]\right| \geq \epsilon\right\}.$$

For any fixed $(s,a)\in\mathbb{K}$,

$$P\left\{\left|\frac{\alpha}{n}\sum_{i=1}^{n} v(\psi(s,a,\xi_i)) - \alpha\mathbb{E}[v(\psi(s,a,\xi))]\right| \geq \epsilon\right\} \leq 2e^{-2(\epsilon/\alpha)^2 n/(v_{\max}-v_{\min})^2} \leq 2e^{-2(\epsilon/\alpha)^2 n/(2\|v\|)^2}$$
$$\leq 2e^{-2(\epsilon/\alpha)^2 n/\left(2\kappa^*\right)^2}$$

by Hoeffding's inequality. Then, using the union bound, we have

$$P\left\{\|\widehat{T}_n v - Tv\| \geq \epsilon\right\} \leq 2|\mathbb{K}|e^{-2(\epsilon/\alpha)^2 n/\left(2\kappa^*\right)^2}.$$

By taking complements of the above event we get the desired result. □

A.7. Eigenvalues of the transition probability matrix of the dominating Markov chain

LEMMA A.3. *For any fixed $n\geq 1$, the eigenvalues of the transition matrix $\mathfrak{Q}$ of the Markov chain Y_n^k are 0 (with algebraic multiplicity $N^*-\eta^*-1$) and 1.*

Proof: In general, the transition matrix $\mathfrak{Q}_n \in \mathbb{R}^{\left(N^*-\eta^*+1\right)\times\left(N^*-\eta^*+1\right)}$ of $\left\{Y_n^k\right\}_{k\geq 0}$ looks like

$$\mathfrak{Q}_n = \begin{bmatrix} p_n & 0 & \cdots & \cdots & 0 & (1-p_n) \\ p_n & 0 & \cdots & \cdots & 0 & (1-p_n) \\ 0 & p_n & 0 & \cdots & 0 & (1-p_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & 0 & (1-p_n) \\ 0 & 0 & \cdots & \cdots & p_n & (1-p_n) \end{bmatrix}.$$

To compute the eigenvalues of $\mathfrak{Q}_n$, we want to solve $\mathfrak{Q}_n x = \lambda x$ for some $x \neq 0$ and $\lambda \in \mathbb{R}$. For $x = (x_1, x_2, \ldots, x_{N^*-\eta^*+1}) \in \mathbb{R}^{N^*-\eta^*+1}$,

$$\mathfrak{Q}_n x = \begin{pmatrix} p_n x_1 + (1-p_n)\, x_{N^*-\eta^*+1} \\ p_n x_1 + (1-p_n)\, x_{N^*-\eta^*+1} \\ p_n x_2 + (1-p_n)\, x_{N^*-\eta^*+1} \\ \vdots \\ p_n x_{N^*-\eta^*-1} + (1-p_n)\, x_{N^*-\eta^*+1} \\ p_n x_{N^*-\eta^*} + (1-p_n)\, x_{N^*-\eta^*+1} \end{pmatrix}.$$

Now, suppose $\lambda \neq 0$ and $\mathfrak{Q}\, x = \lambda\, x$ for some $x \neq 0$. By the explicit computation of $\mathfrak{Q}\, x$ above,

$$[\mathfrak{Q}\, x]_1 = p\, x_1 + (1-p)\, x_{N^*-\eta^*+1} = \lambda\, x_1$$

and

$$[\mathfrak{Q}\, x]_2 = p\, x_1 + (1-p)\, x_{N^*-\eta^*+1} = \lambda\, x_2,$$

so it must be that $x_1 = x_2$. However, then

$$[\mathfrak{Q}\, x]_2 = p\, x_1 + (1-p)\, x_{N^*-\eta^*+1} = p\, x_2 + (1-p)\, x_{N^*-\eta^*+1} = [\mathfrak{Q}\, x]_3,$$

and thus $x_2 = x_3$. Continuing this reasoning inductively shows that $x_1 = x_2 = \cdots = x_{N^*-\eta^*+1}$ for any eigenvector x of $\mathfrak{Q}$. Thus, it must be true that $\lambda = 1$. □

A.8. Proof of Proposition 6.1

Proof: Starting with v^0,

$$T_{x_0} v^0 - \eta\, 1 \leq T_{x_0} v^0 + \varepsilon_0 \leq T_{x_0} v^0 + \eta\, 1,$$

which gives

$$T_{x_0} v^0 - \eta\, 1 \leq \tilde{v}^0 \leq T_{x_0} v^0 + \eta\, 1,$$

and

$$v^1 - \eta\, 1 \leq \tilde{v}^1 \leq v^1 + \eta\, 1.$$

By monotonicity of T_{x_1},

$$T_{x_1}\left[v^1 - \eta\, 1\right] \leq T_{x_1} \tilde{v}^1 \leq T_{x_1}\left[v^1 + \eta\, 1\right],$$

and by our assumptions on the noise,

$$T_{x_1}\left[v^1 - \eta\, 1\right] - \eta\, 1 \leq \tilde{v}^2 \leq T_{x_1}\left[v^1 + \eta\, 1\right] + \eta\, 1.$$

Now

$$T_{x_1}\left[v - \eta\, 1\right] = T_{x_1} v - \alpha\, \eta\, e_x,$$

thus

$$v^2 - \eta 1 - \alpha \eta 1 \leq \tilde{v}^2 \leq v^2 + \eta 1 + \alpha \eta 1.$$

Similarly,

$$v^3 - \alpha (\eta 1 - \alpha \eta 1) - \eta 1 \leq \tilde{v}^3 \leq v^3 + \alpha (\eta 1 + \alpha \eta 1) + \eta 1,$$

and the general case follows. □

A.9. Proof of Lemma 6.3

Proof: Using Fact 2.1 twice, compute

$$\begin{aligned}
|[T v](s) - [T v'](s)| = & |\min_{a \in A(s)} \max_{b \in B(s)} \{r(s,a,b) + \alpha \mathbb{E}[v(\tilde{s})|s,a,b]\} \\
& \qquad - \min_{a \in A(s)} \max_{b \in B(s)} \{r(s,a,b) + \alpha \mathbb{E}[v'(\tilde{s})|s,a,b]\}| \\
\leq & \max_{a \in A} |\max_{b \in B(s)} \{r(s,a,b) + \alpha \mathbb{E}[v(\tilde{s})|s,a,b]\} - \max_{b \in B(s)} \{r(s,a,b) + \alpha \mathbb{E}[v'(\tilde{s})|s,a,b]\}| \\
\leq & \alpha \max_{a \in A} \max_{b \in B(s)} |\mathbb{E}[v(\tilde{s}) - v'(\tilde{s})|s,a,b]| \\
\leq & \alpha \max_{a \in A} \max_{b \in B} \mathbb{E}[|v(\tilde{s}) - v'(\tilde{s})||s,a,b] \\
\leq & \alpha \|v - v'\|.
\end{aligned}$$

□