2015 IEEE 54th Annual Conference on Decision and Control (CDC)
December 15-18, 2015. Osaka, Japan

An Empirical Algorithm for Relative Value Iteration for Average-cost MDPs

Abhishek Gupta Rahul Jain Peter W. Glynn

Abstract— **Infinite-horizon average-cost Markov decision processes are of interest in many scenarios. A dynamic programming algorithm, called the relative value iteration, is used to compute the optimal value function. For large state spaces, this often runs into difficulties due to its computational burden. We propose a simulation-based dynamic program called empirical relative value iteration (ERVI). The idea is very simple: replace the expectation in the Bellman operator with a sample average estimate, and then use projection to ensure boundedness of the iterates. We establish that the ERVI iterates converge to the optimal value function in the span-seminorm in probability as the number of samples taken goes to infinity. Simulation results show remarkably good performance even with a small number of samples.**

I. Introduction

Markov decision processes are popular models for sequential decision-making under uncertainty [1], [2]. Using Bellman's 'principle of optimality' [3] for such settings, one can devise dynamic programming (DP) algorithms for finding optimal policies. This transforms the sequential decision-making problem into a fixed point problem, where we look for a fixed point of the Bellman operator. For the infinite-horizon discounted cost case, two examples of DP algorithms are value and policy iteration [1]. It is well-known that these and other DP algorithm suffer from the "curse of dimensionality" [3], [2], [4], i.e., the computational complexity increases exponentially in state space size. In fact, it has been shown to be PSPACE-hard [5].

This has led to a slew of work on approximate methods for dynamic programming [6], [2]. The first idea is to approximation value functions by a finite basis or kernel of functions [7], [2]. While very useful, this approach is model-dependent and not universal. The second class of approximate DP algorithms is based on stochastic approximation schemes. For example, Q-learning was introduced in [8] and has led to development of a whole class of reinforcement learning algorithms that are widely used. Unfortunately, it is well-known that such schemes have very slow convergence.

In [9], an alternative idea called 'empirical dynamic programming' (for discounted MDPs) was introduced. This involved replacing the expectation in the Bellman operator in value iteration by a sample-average approximation obtained by one-step simulation of the next state in each iteration. The idea is very simple and a natural way of doing dynamic programming by simulation. It was shown that while using a finite number of samples in each iteration may not yield convergence to the optimal value function, nevertheless, it yielded estimates having small error with high probability, even with a very small number of samples. Moreover, a formal convergence result was established that guarantees that such 'empirical' algorithms yield estimates having small error with high probability. The numerical performance is surprising, as we see that even with as few as 5-10 samples per iteration, we get to small error at almost the same rate as exact DP algorithms at a fraction of the computational cost.

In this paper, we develop similar ideas for the infinite-horizon average-cost case. The difficulty is that unlike the discounted case, the Bellman operator is non-contractive. Thus, value iteration with the 'average-Bellman operator' is not guaranteed to converge. Instead, relative value iteration is used wherein at each iteration, a normalization is done by subtracting the value iterate at a reference state from the value iterate itself. This makes it a 'relative' value iteration algorithm, which converges under a seminorm. Unfortunately, due to this last step, devising a simulation-based or 'empirical' variant of the algorithm becomes difficult.

In this paper, we take a different perspective of relative value iteration for average-cost MDPs. We view it as iteration of the average-Bellman operator on a quotient space that we introduce here. On this space, the operator is still a contraction, and by standard results about contracting operators, convergence is implied. We then introduce an 'empirical' variant of the relative value iteration algorithm. To ensure boundedness of the iterates, we apply a projection operator at each step of the iteration. We can then show that the iterates of the empirical relative value iteration (EVRI) algorithm converge to zero in the span seminorm as the number of samples and iterations goes to infinity. One difficulty with proving convergence is that each iteration of the algorithm is noisy, as we are basically iterating using a random operator. However, in [9], we developed a fairly general framework for convergence analysis of iteration of random operators via stochastic dominance argument. In particular, we construct a sequence of Markov chains that stochastically dominate the value iterates of the ERVI algorithm. As the number of samples goes to infinity, we can show that the measure of the sequence of Markov chains concentrates at zero, thus proving convergence. Numerical simulations show very competitive performance even with a small number of samples and a finite number of iterations. In fact, in practice, the projection is never needed.

Abhishek Gupta and Rahul Jain are with the USC Ming Hsieh Department of Electrical Engineering at University of Southern California, Los Angeles, CA. Emails: {guptaabh, rahul.jain}@usc.edu.

Peter W. Glynn is with the MS&E Department, Stanford University. Email: glynn@stanford.edu.

II. Average-cost MDPs: An Overview

A. Preliminaries

Consider a Markov decision process (MDP) on a finite state space $\mathcal{S}$ and a finite action space $\mathcal{A}$. The state evolution is determined by a probability transition kernel $p(S_{t+1} = j|s_t, a_t)$ which yields the distribution over the next state S_{t+1} given action a_t is taken in current state s_t. We can write the state evolution as

$$S_{t+1} = f(S_t, A_t, W_t),$$

where $\{W_t\}$ is an i.i.d. sequence of random noise, which (without much loss of generality) we can assume to be uniformly distributed over $[0, 1]$. Let $\mathcal{A}(s)$ denote the set of permissible actions in state s. The set $\mathcal{K} \subset \mathcal{S} \times \mathcal{A}$ denotes the set of all permissible state-action pairs. We will denote a policy by $\pi_t : \mathcal{S} \to \mathcal{A}$ which yields a permissible action in each state. If this is invariant over time, we call it stationary.

Let $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denote a cost function which we will assume is bounded. For a stationary policy π, the expected average-cost can be expressed as

$$J(\pi, s) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} c(s_t, \pi(s_t))\middle| s_0 = s\right].$$

We would like to find the optimal stationary policy that minimizes the expected average-cost J given the initial state s, i.e., $J(\pi^\star, s) \leq J(\pi, s)$. We make the following assumption on the MDP.

Assumption 1 (MDP is Unichain): (i) The MDP is unichain, i.e., for every stationary policy $\pi : \mathcal{S} \to \mathcal{A}$, the resulting Markov chain $\{S_t^\pi\}_{t=1}^\infty$, defined as $S_{t+1}^\pi = f(S_t^\pi, \pi(S_t^\pi), W_t)$ is unichain. (ii) $\sum_{j\in\mathcal{S}} \min\left\{p(j|s, a), p(j|s', a')\right\} > 0$ for any $(s, a), (s', a') \in \mathcal{K}$.

The following result is then well-known.

Theorem 1 ([1]): Consider an average-cost MDP that satisfies Assumption 1. Then, there exists a value function $v^\star : \mathcal{S} \to \mathbb{R}$ and a unique gain $g^\star \in \mathbb{R}$ such that

$$v^\star(s) + g^\star = \min_{a\in\mathcal{A}(s)} \Big(c(s, a) + \mathbb{E}\left[v^\star(f(s, a, W)\right]\Big) \quad (1)$$

for all $s \in \mathcal{S}$. Moreover, the optimal stationary policy $\pi^\star$ for the MDP is given by

$$\pi^\star(s) \in \arg\min_{a\in\mathcal{A}(s)} \Big(c(s, a) + \mathbb{E}\left[v^\star(f(s, a, W)\right]\Big), \quad \forall\, s \in \mathcal{S}.$$

One can check that if $v^\star$ satisfies (1), then $v^\star + \lambda\mathbb{1}_{|\mathcal{S}|}$ also satisfies (1), where $\lambda \in \mathbb{R}$ and $\mathbb{1}_{|\mathcal{S}|}$ denotes a constant function $f : \mathcal{S} \to \{1\}$. Thus, $v^\star$ is not unique, but $g^\star$ is unique. Because of this property, the value iteration algorithm does not converge, but a variant of it, called the relative value iteration algorithm, converges to $v^\star$. We describe this algorithm in the next subsection. Furthermore, once we compute $v^\star$, then we can also compute the optimal stationary policy for the average-cost MDP.

B. Relative Value Iteration

We now present the well-known relative value iteration algorithm. We first introduce the space of value functions over the state space $\mathcal{S}$ and the classical Bellman operator over this space. Let $\mathcal{V} := \mathbb{R}^{|\mathcal{S}|}$ denote the space of value functions. Let the (exact) Bellman operator $T : \mathcal{V} \to \mathcal{V}$ be defined as

$$Tv(s) = \min_{a\in\mathcal{A}(s)} \Big(c(s, a) + \mathbb{E}\left[v(f(s, a, W)\right]\Big), \quad (2)$$

where the expectation is taken with respect to the distribution of W. It is easy to check that the operator T is not a contraction for most norms on the space $\mathcal{V}$. However, it is a contraction with respect to the span semi-norm on the space $\mathcal{V}$, which is defined as follows. Define the span of a value function $v \in \mathcal{V}$ as

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

This is a semi-norm[1]. We know from [1, Proposition 6.6.1] that the Bellman operator T satisfies

$$\text{span}(Tv_1 - Tv_2) \leq \alpha\ \text{span}(v_1 - v_2),$$

where α is given by

$$\alpha := 1 - \min_{(s,a),(s',a')\in\mathcal{K}} \sum_{j\in\mathcal{S}} \min\left\{p(j|s, a), p(j|s', a')\right\}.$$

Since Assumption 1 (ii) holds and $\mathcal{K}$ is a finite set, we naturally have $\alpha < 1$.

We now describe relative value iteration algorithm. Suppose that the MDP satisfies Assumption 1. Start with any v_0, say $v_0 = 0$. We compute relative value iterates $\{v_k\}$ in the following way:

$$\tilde{v}_{k+1} = Tv_k, \quad (3)$$

$$v_{k+1} = \tilde{v}_{k+1} - \Big(\min_{s\in\mathcal{S}} \tilde{v}_{k+1}(s)\Big)\mathbb{1}_{|\mathcal{S}|}. \quad (4)$$

We stop when $\text{span}(v_k)$ becomes smaller than some $\epsilon > 0$. This will happen since T is a contraction over the span semi-norm due to Assumption 1.

The above algorithm asymptotically converges to the optimal value function $v^\star$ under Assumption 1. The proof can be found in Section 8.5.5 of [1]. Moreover, it is easy to check that $\text{span}(v^\star) \leq \frac{\|c\|_\infty}{1-\alpha}$, where $\|c\|_\infty$ is defined as

$$\|c\|_\infty = \max_{(s,a)\in\mathcal{K}} |c(s, a)|.$$

One computational difficulty in the above algorithm is that in evaluating the operator T in (2), we need to evaluate the expectation for every (s, a) pair. This can be computationally expensive or even infeasible if the state space is very large. Instead, we could replace this expectation with a sample average estimate from simulation. We explore this next.

[1] In particular, it is not positive definite, that is, if $v_1 = z + v_2$ for some scalar z, then $\text{span}(v_1 - v_2) = 0$.

III. Empirical Relative Value Iteration

We now describe a simulation-based algorithm for approximate relative value iteration. The idea is simple and natural – replace the expectation in the Bellman operator by its estimate. This can be done by generating n i.i.d. sample $W_1, \ldots, W_n$, and doing one step simulation of the next state. When we plug the estimate for the expectation into the Bellman operator, we get the following *empirical Bellman operator*

$$\hat{T}_n v(s) = \min_{a \in \mathcal{A}(s)} \left(c(s,a) + \frac{1}{n} \sum_{i=1}^{n} v(f(s,a,W_i)) \right). \quad (5)$$

Note that $\hat{T}_n$ is a random operator, since it depends on the realizations of $W_1, \ldots, W_n$. The next iterate we get after applying $\hat{T}_n$ on the current iterate is therefore random. Furthermore, although the exact Bellman operator T is a contraction over the semi-normed space $(\mathcal{V}, \text{span})$, the empirical Bellman operator $\hat{T}_n$ may not be a contraction.

Thus, we introduce a projection operator $P : \mathcal{V} \to \mathcal{V}$ defined as follows:

$$Pv(s) = \begin{cases} \left(v(s) - \min_{j \in \mathcal{S}} v(j) \right) & \text{if } \text{span}(v) \leq \frac{\|c\|_\infty}{1-\alpha} \\ \frac{\left(v(s) - \min_{j \in \mathcal{S}} v(j) \right)}{\text{span}(v)} & \text{otherwise.} \end{cases}$$

This operator P first translates a value function so that the minimum of the function is zero, and then scales it so that $\text{span}(Pv) \leq \frac{\|c\|_\infty}{1-\alpha}$[2].

The ERVI algorithm consists essentially of the same steps as in exact relative value iteration, except that we plug in an estimate from simulation for the expectation in the operator T. However, since $\hat{T}_n$ is not guaranteed to be a contraction, we add a projection step using the P operator that ensures boundedness of all iterates. The algorithm is formally described above.

Algorithm 1 Empirical Relative Value Iteration (ERVI)

Input: Number of iterations $K \in \mathbb{N}$, $v^0 \in \mathbb{R}^{|\mathcal{S}|}$, sample size $n \geq 1$. Set counter $k = 0$.

1) Sample n uniformly distributed random variables $\{W_i\}_{i=1}^n$ from $[0, 1]$, and compute
$$\tilde{v}_{k+1} = \hat{T}_n \hat{v}_k$$
2) $\hat{v}_{k+1} = P\tilde{v}_{k+1}$
3) If $k \geq K$, then stop. Else $k \leftarrow k+1$ and go to Step 1.

A natural question now is how good is the solution found by the empirical RVI algorithm and whether we can provide any guarantees on the quality of the solution. Let $\hat{v}_K^n$ denote the outcome of the ERVI algorithm, where we use n samples in each iteration and do K iterations. We can establish the following.

[2]We need to know the value of α in order to use the projection operator during the simulation. For some special cases, α can be bounded from above easily even when the state space is large. For example, if we know that there is a state $s_0 \in \mathcal{S}$ such that $p(s_0|s,a) \geq \eta$ for all $(s,a) \in \mathcal{K}$ for some constant $\eta > 0$, then $\alpha \leq 1 - \eta$. We thank the referee for pointing this out.

Theorem 2: Consider an MDP that satisfies Assumption 1 with an average-cost criterion. Let $\hat{v}_K^n$ be the iterate from the ERVI algorithm with n samples in each iteration and K iterations. Then, for any $\epsilon > 0$,

$$\lim_{n \to \infty} \limsup_{K \to \infty} \mathbb{P}\Big\{ \text{span}(v^\star - \hat{v}_K^n) > \epsilon \Big\} = 0.$$

This implies that as the number of samples and iterations go to infinity, the value function computed by ERVI converges in probability to the optimal value function $v^\star$ in the span seminorm. This assures us that with finite n and for finite K, ERVI will return a value function that is close (in the span seminorm) to the optimal value function with high probability.

Note that the guarantee on the quality of the solution by ERVI is weaker than what stochastic approximation-based algorithms such as Q-learning for average-cost MDPs give. However, we get a faster convergence rate and lower computational complexity than both exact and reinforcement learning algorithms.

To prove this result, we introduce a quotient space $(\mathcal{V}/\sim, \text{span})$ by defining functions in space $\mathcal{V}$ to be in the same equivalence class if they differ by a constant function: For a function $v \in \mathcal{V}$, an element $[v] \in \mathcal{V}/\sim$ is defined as

$$\begin{aligned} [v] &= \Big\{ \tilde{v} \in \mathcal{V} : \tilde{v} - v \text{ is a constant function} \Big\} \\ &= \bigcup_{\lambda \in \mathbb{R}} \{ v + \lambda \mathbb{1}_{|\mathcal{S}|} \}. \end{aligned}$$

Thus, $[v_1] = [v_2]$ if and only if $v_1 - v_2$ is a function with equal entries. Also, note that for $v_1, v_2 \in \mathcal{V}$,

$$[v_1] + [v_2] = [v_1 + v_2]. \quad (6)$$

We also have $\text{span}([v]) = \text{span}(v) = \text{span}(w)$ for any $w \in [v]$. Thus,

$$\text{span}([v_1] + [v_2]) = \text{span}([v_1 + v_2]) = \text{span}(v_1 + v_2).$$

For more information on the span seminorm and the resulting quotient space, we refer the reader to Appendix I. We extend the Bellman operator, the empirical Bellman operator, and the projection operator on space $\mathcal{V}$ to corresponding operators over the quotient space $\mathcal{V}/\sim$. Then, we study probabilistic convergence properties of the extended operators, and use a result of [9] to conclude Theorem 2.

IV. Convergence Analysis of ERVI

In this section, we present a proof of Theorem 2. For this purpose, we leverage some results about iteration of random operators over Banach spaces from [9]. There, it was shown that if a non-linear operator and corresponding sequence of random operators satisfy four simple conditions, then iterations using random operators converges to a fixed point which is close to the fixed point of the non-linear operator with high probability. We restate the result in slightly easier form to facilitate understanding our result.

Theorem 3 (Theorem 4.3, [9]): Let $\mathcal{X}$ be a complete normed vector space, $F : \mathcal{X} \to \mathcal{X}$ be a contraction operator with a fixed point $x^\star$, and $\hat{F}_n : \mathcal{X} \to \mathcal{X}$ be a random operator

approximating the operator F using n i.i.d. samples of some random variable. Suppose that the following conditions hold:

1) F is a contraction with coefficient $\alpha < 1$.
2) There exists $\kappa > 0$ such that $\|x^\star\|_{\mathcal{X}} \leq \kappa$ and $\|\hat{F}_n^k 0\|_{\mathcal{X}} \leq \kappa$ for all $k \in \mathbb{N}$ almost surely.
3) For any $\epsilon > 0$ and $n \in \mathbb{N}$, there exists constant $p_1, p_2 > 0$ such that we have

$$\mathbb{P}\left\{\|Fx - \hat{F}_n x\|_{\mathcal{X}} > \epsilon\right\} \leq p_1 \exp\left(-\frac{p_2 \epsilon^2 n}{\|x\|_{\mathcal{X}}^2}\right).$$

4) We have

$$\mathbb{P}\left\{\lim_{n\to\infty} \hat{F}_n x = Fx\right\} = 1.$$

Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \limsup_{k\to\infty} \mathbb{P}\left\{\|\hat{F}_n^k x - x^\star\|_{\mathcal{X}} > \epsilon\right\} = 0.$$

In the problem we consider here, we are approximating operator T by $\hat{T}_n$. Both operators are defined over the semi-normed space $(\mathcal{V}, \text{span})$, which, unfortunately, is not a Banach space. However, by reformulating the convergence problem over the quotient space $(\mathcal{V}/\sim, \text{span})$ (and by extending the operators), we show that the setting considered in Theorem 3 can be applied.

A. Bellman and Empirical Operators over Quotient Space

We first note the following property of the projection operator.

Lemma 4: For any $v_1, v_2 \in \mathcal{V}$, $\text{span}(Pv_1 - Pv_2) \leq \text{span}(v_1 - v_2)$. Consequently, $\text{span}(Pv_1 - Pv_2) \leq 2\|v_1 - v_2\|_\infty$.

Proof: Consider the operator $\Lambda_P : \mathcal{V}/\sim \to \mathcal{V}/\sim$ obtained by defining

$$\Lambda_P[v] = [Pv].$$

Since Λ_P is simply a projection operator on $\mathcal{V}/\sim$, the result follows. ∎

We now extend the operators $PT : \mathcal{V} \to \mathcal{V}$ and $P\hat{T}_n : \mathcal{V} \to \mathcal{V}$ to the quotient space $\mathcal{V}/\sim$ as follows:

$$\Lambda[v] := [PTv], \qquad \hat{\Lambda}_n[v] = [P\hat{T}_n v].$$

We immediately conclude that for any $v \in \mathcal{V}$,

$$\begin{aligned} &\Lambda[v] - \hat{\Lambda}_n[v] = [PTv - P\hat{T}_n v] \\ \implies &\text{span}\Big(\Lambda[v] - \hat{\Lambda}_n[v]\Big) \leq 2\|Tv - \hat{T}_n v\|_\infty, \end{aligned} \tag{7}$$

where we used Lemma 4.

B. Properties of the Empirical Operator $\hat{\Lambda}_n$

We first study the relationship between T and $\hat{T}_n$. Pick any $v \in \mathcal{V}$. The strong law of large numbers [10] then implies

$$\begin{aligned} &\mathbb{P}\left\{\lim_{n\to\infty} \hat{T}_n v = Tv\right\} = 1 \\ \Leftrightarrow\ &\mathbb{P}\left\{\lim_{n\to\infty} \|\hat{T}_n v - Tv\|_\infty = 0\right\} = 1. \end{aligned}$$

Hoeffding inequality [11], [12] implies

$$\mathbb{P}\left\{\|Tv - \hat{T}_n v\|_\infty \geq \epsilon\right\} \quad \leq 2|\mathcal{K}| \exp\left(-\frac{2\epsilon^2 n}{\|v\|_\infty^2}\right).$$

With these two results, we can now state the relationship between Λ and $\hat{\Lambda}_n$.

Lemma 5: For any $v \in \mathcal{V}$ with $\text{span}(v) \leq \kappa := \frac{\|c\|_\infty}{1-\alpha}$, we have

$$\begin{aligned} &\mathbb{P}\left\{\lim_{n\to\infty} \text{span}\Big(\hat{\Lambda}_n[v] - \Lambda[v]\Big) = 0\right\} = 1, \quad \text{and} \\ &\mathbb{P}\left\{\text{span}\Big(\hat{\Lambda}_n[v] - \Lambda[v]\Big) \geq \epsilon\right\} \\ &\leq 2|\mathcal{K}| \exp\left(-\frac{2(\epsilon/2)^2 n}{\kappa^2}\right). \end{aligned}$$

Proof: Equation (7) yields both the results.

1) Let $\Omega_0 \subset [0,1]^{\mathbb{N}}$ be the set of $\{w_i\}_{i\in\mathbb{N}} \in [0,1]^{\mathbb{N}}$ such that

$$\lim_{n\to\infty} \hat{T}_n v = Tv.$$

We know from the strong law of large numbers that $\mathbb{P}\{\Omega_0\} = 1$. Since (7) holds, we know that for every $\{w_i\}_{i\in\mathbb{N}} \in \Omega_0$, we have

$$\begin{aligned} &\lim_{n\to\infty} \text{span}\Big(\Lambda[v] - \hat{\Lambda}_n[v]\Big) \\ &\leq \lim_{n\to\infty} 2\|Tv - \hat{T}_n v\|_\infty = 0. \end{aligned}$$

Thus, we get

$$\mathbb{P}\left\{\lim_{n\to\infty} \text{span}\Big(\hat{\Lambda}_n[v] - \Lambda[v]\Big) = 0\right\} \geq \mathbb{P}\{\Omega_0\} = 1.$$

This completes the proof of the first result.

2) By (7), we immediately conclude

$$\begin{aligned} &\mathbb{P}\left\{\text{span}\Big(\hat{\Lambda}_n[v] - \Lambda[v]\Big) \geq \epsilon\right\} \\ &\leq \mathbb{P}\left\{2\|Tv - \hat{T}_n v\|_\infty \geq \epsilon\right\} \\ &\leq 2|\mathcal{K}| \exp\left(-\frac{2(\epsilon/2)^2 n}{\kappa^2}\right). \end{aligned}$$

∎

In the next lemma, we prove that if T is a contraction, then so is Λ.

Lemma 6: If T is a contraction with coefficient $\alpha < 1$ over the semi-normed space $(\mathcal{V}, \text{span})$, then $\Lambda : \mathcal{V}/\sim \to \mathcal{V}/\sim$ is also a contraction with coefficient α.

Proof: We have

$$\begin{aligned} \text{span}(\Lambda[v_1] - \Lambda[v_2]) &= \text{span}(PTv_1 - PTv_2), \\ &\leq \text{span}(Tv_1 - Tv_2), \\ &\leq \alpha\ \text{span}(v_1 - v_2), \\ &= \alpha\ \text{span}([v_1] - [v_2]), \end{aligned}$$

where the first inequality follows from Lemma 4 and the second inequality follows from the hypothesis. ∎

Remark 1: Recall that the projection operator is used in Step 2 of empirical relative value iteration algorithm. As a result of the projection, we always have $\text{span}(\hat{\Lambda}_n^k[0]) \leq \frac{\|c\|_\infty}{1-\alpha}$ almost surely. □

Theorem 7: Consider an average cost MDP that satisfies Assumption 1. Then, for any $\epsilon > 0$

$$\lim_{n\to\infty} \limsup_{k\to\infty} \mathbb{P}\left\{\text{span}\left(\hat{\Lambda}_n^k[0] - v^\star\right) > \epsilon\right\} = 0.$$

Proof: Notice that $\mathcal{V}/\sim$ is a complete normed vector space. Lemmas 5 and 6 and Remark 1 imply that Λ and $\hat{\Lambda}_n$ satisfy the hypotheses of Theorem 3. The result then follows immediately from Theorem 3. ∎

The result above implies Theorem 2. Thus, one can get an approximately optimal solution of the average cost MDP with empirical relative value iteration with high probability. The sample complexity for this problem is summarized in the following theorem.

Theorem 8: Given $\epsilon, \delta > 0$, pick $\delta_1, \delta_2 > 0$ such that $\delta_1 + 2\delta_2 \leq \delta$. Define

$$\eta = \left\lceil \frac{2}{1-\alpha} \right\rceil, \qquad M = \left\lceil \frac{2\eta\kappa}{\epsilon} \right\rceil,$$
$$p_n = 1 - 2|\mathcal{K}| \exp\left(-\frac{\epsilon^2 n}{2\eta^2\kappa^2}\right),$$
$$\mu_n(\eta) = p_n^{M-\eta-1}, \qquad \mu_n(M) = \frac{1-p_n}{p_n},$$
$$\mu_n(j) = (1-p_n)p_n^{M-j-1}, \qquad \eta < j < M, j \in \mathbb{N}.$$

Then, the sample complexity is given by

$$N(\epsilon,\delta) = \frac{2\kappa^2\eta^2}{\epsilon^2} \log\left(\frac{2|\mathcal{K}|}{\delta_1}\right),$$
$$K(\epsilon,\delta,n) = \log\left(\frac{1}{\delta_2 \mu_n^\star}\right),$$

where $\mu_n^\star = \min_{\eta \leq j \leq M} \mu_n(j)$. Moreover, for every $n \geq N(\epsilon,\delta)$, we get

$$\sup_{k \geq K(\epsilon,\delta,n)} \mathbb{P}\left\{\text{span}\left(\hat{\Lambda}_n^k[0] - v^\star\right) > \epsilon\right\} \leq \delta.$$

Proof: The proof follows immediately from [9, Theorem 3.1, p. 6]. ∎

V. Numerical Performance

We now present numerical performance results for the ERVI algorithm. For this purpose, we consider an MDP with 100 states and 5 actions, and generated the transition probability matrix randomly. We also ran the exact relative value iteration (RVI) algorithm on the same MDP for comparison purposes. ERVI was run with 20, 40 and 200 samples. In each case, 200 simulation runs were conducted and the normalized error

$$\frac{\|\hat{v}_k - v^\star\|_\infty}{\|v^\star\|_\infty} \times 100\%.$$

averaged over these runs.

As can be seen in Figure 1, exact RVI converges in 3 iterations for this MDP. ERVI with 20 samples settles down to about 15% normalized error in the same number of iterations. As we increase the number of samples, this decreases and for $n = 200$, we get less than 5% normalized error in 3 iterations.

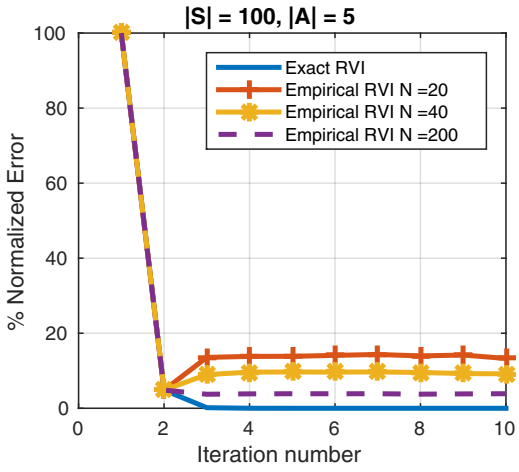

Fig. 1. Average normalized error vs. iteration number for $N = 20, 40, 200$ using empirical relative value iteration.

We also note that in these simulations, the projection operation was not needed at all. Our proof depends on the projection step in the algorithm, but we conjecture that convergence holds without it as well.

VI. Conclusion

In this paper, we introduced a simulation-based approximate dynamic programming algorithm called 'empirical relative value iteration' for average-cost Markov decision processes. Although the idea behind proposing this algorithm is simple and natural, proving convergence is rather non-trivial. We are able to cast relative value iteration as value iteration on a quotient space over which the extended Bellman operator is a contraction. In fact, we can view empirical relative value iteration as empirical value iteration on the quotient space. Thus, we are able to leverage results about the convergence analysis of iteration of random operators that was developed in [9] to prove the convergence of the relative value iterates of ERVI in probability. One of the most interesting conclusions about the empirical algorithm in this paper (and about empirical dynamic programming, in general) is that it has surprisingly good numerical performance for all practical purposes; in fact, the convergence is faster than even reinforcement learning algorithms. In the future, we will prove convergence without requiring the projection step in the algorithm. We will also extend this framework to infinite state spaces.

References

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

[2] D. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1 and 2. 4th ed., 2012.

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

[4] J. Rust, "Using randomization to break the curse of dimensionality," *Econometrica*, vol. 65, no. 3, pp. 487–516, 1997.

[5] C. H. Papadimitriou and J. Tsitsiklis, "The complexity of Markov decision processes," *Math. Oper. Res*, vol. 12, no. 3, pp. 441–450, 1987.

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

[7] R. Bellman and S. Dreyfus, "Functional approximations and dynamic programming," *Mathematical Tables and Other Aids to Computation*, vol. 13, no. 68, pp. 247–251, 1959.
[8] C. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.
[9] W. B. Haskell, R. Jain, and D. Kalathil, "Empirical Dynamic Programming," *accepted in Math. Oper. Res.*, March 2015. available online at http://arxiv.org/abs/1311.5918.
[10] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, 2010.
[11] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.
[12] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

APPENDIX I
QUOTIENT SPACE AND OPERATORS OVER THE QUOTIENT SPACE

In this section, we study the span semi-norm over the space $\mathcal{V}$. First, let us show that span is a semi-norm.

1) Let $v \in \mathcal{V} \cong \mathbb{R}^{|\mathcal{S}|}$. Then, $\text{span}(v) = \max_{s\in\mathcal{S}} v(s) - \min_{s\in\mathcal{S}} v(s) \geq 0$. Span of a function v is zero if and only if it is of the form $v = \lambda \mathbb{1}_{|\mathcal{S}|}$ for some $\lambda \in \mathbb{R}$.
2) For any constant $\lambda \in \mathbb{R}$, we have $\text{span}(\lambda v) = |\lambda|\text{span}(v)$.
3) For any $v_1, v_2 \in \mathcal{V}$, $\text{span}(v_1 + v_2) \leq \text{span}(v_1) + \text{span}(v_2)$. We can prove it as follows:

$$\begin{aligned}\text{span}(v_1+v_2) =& \max_{s\in\mathcal{S}} \Big(v_1(s)+v_2(s)\Big) \\ & - \min_{s\in\mathcal{S}} \Big(v_1(s)+v_2(s)\Big) \\ \leq& \max_{s\in\mathcal{S}} v_1(s) + \max_{s\in\mathcal{S}} v_2(s) \\ & - \min_{s\in\mathcal{S}} v_1(s) - \min_{s\in\mathcal{S}} v_2(s) \\ =& \text{span}(v_1) + \text{span}(v_2),\end{aligned}$$

where the inequality follows from the properties of minimum and maximum.

Therefore, span is a semi-norm on $\mathcal{V}$ and not a norm, because it does not satisfy positive definiteness. Our goal now is to prove that the normed space $(\mathcal{V}/\sim, \text{span})$ is in fact a Banach space. To prove this result, we need the following lemma.

Lemma 9: Let $\mathcal{V}_0 \subset \mathcal{V}$ be the subset of functions in which at least one of the entries is 0. For every $s \in \mathcal{S}$ and $v \in \mathcal{V}_0$, $|v(s)| \leq \text{span}(v)$.

Proof: Fix $s_0 \in \mathcal{S}$. Assume that $v(s_0) \geq 0$. We have

$$\begin{aligned}&\max_s v(s) \geq 0 \geq \min_s v(s) \quad \text{and} \quad v(s_0) \leq \max_s v(s) \\ &\implies v(s_0) \leq \max_s v(s) - \min_s v(s) = \text{span}(v).\end{aligned}$$

Now, if $v(s_0) < 0$, then apply the above approach to $-v$. We get $-v(s_0) \leq \text{span}(-v) = \text{span}(v)$. In other words, $|v(s)| \leq \text{span}(v)$ for all $s \in \mathcal{S}$. The proof is thus complete. ∎

Lemma 10: $(\mathcal{V}/\sim, \text{span})$ is a Banach space.

Proof: Consider a Cauchy sequence $\{[v_n]\}_{n\in\mathbb{N}} \subset \mathcal{V}/\sim$ such that for every $\epsilon > 0$, there exists N_ϵ such that $\text{span}([v_n] - [v_m]) < \epsilon$ for every $n, m \geq N$.

Fix $s_0 \in \mathcal{S}$. Pick $w_n \in [v_n]$ such that $w_n(s_0) = 0$ for all $n \in \mathbb{N}$. Then, for any $m, n \in \mathbb{N}$, we have $w_n(s_0) - w_m(s_0) = 0$, or $w_n - w_m \in \mathcal{V}_0$ (see Lemma 9). By Lemma 9, we get that for every $s \in \mathcal{S}$,

$$|w_n(s) - w_m(s)| \leq \text{span}(w_n - w_m) = \text{span}([v_n] - [v_m]).$$

We next claim that for every $s \in \mathcal{S}$, $\{w_n(s)\}_{n\in\mathbb{N}} \subset \mathbb{R}$ is a Cauchy sequence in $\mathbb{R}$. Fix $\epsilon > 0$. For every $m, n \geq N_\epsilon$, we get

$$|w_n(s) - w_m(s)| < \epsilon,$$

which implies that $\{w_n(s)\}_{n\in\mathbb{N}}$ is a Cauchy sequence, and therefore converges. Define $w_\infty(s) = \lim_{n\to\infty} w_n(s)$, and define $[v_\infty] = [w_\infty]$.

We next claim that $[v_\infty]$ is the limit of $\{[v_n]\}_{n\in\mathbb{N}}$, which establishes the result. Indeed,

$$\lim_{n\to\infty} \text{span}([v_n] - [v_\infty]) = \lim_{n\to\infty} \text{span}(w_n - w_\infty) = 0.$$

This completes the proof of the lemma. ∎

We now consider operators over this quotient space. Let $L : \mathcal{V} \to \mathcal{V}$ be any operator such that $L(v + \lambda \mathbb{1}_{|\mathcal{S}|}) = Lv + \lambda \mathbb{1}_{|\mathcal{S}|}$[3]. This operator can be extended to an operator $\Lambda_L : \mathcal{V}/\sim \to \mathcal{V}/\sim$ as follows:

$$\Lambda_L([v]) = [Lv]. \tag{8}$$

Note that with the above definition, if $w \in [v]$, then $\Lambda_L([w]) = \Lambda_L([v])$. Thus, it does not matter which representative element of $[v]$ we pick for computing $\Lambda_L([v])$.

Suppose that L_1 and L_2 are two operators such that $L_i(v + \lambda \mathbb{1}_{|\mathcal{S}|}) = L_i v + \lambda \mathbb{1}_{|\mathcal{S}|}$ for $i \in \{1, 2\}$. Then, for any $v \in \mathcal{V}$, the following hold:

$$\Lambda_{L_1}[v] - \Lambda_{L_2}[v] = [L_1 v - L_2 v], \tag{9}$$

$$\Lambda_{L_1}\Lambda_{L_2}[v] = \Lambda_{L_1}[L_2 v] = [L_1 L_2 v]. \tag{10}$$

[3]One can immediately recognize that Bellman operator T, empirical Bellman operator $\hat{T}_N$, and projection operator P satisfy this condition.