A NOTE ON CONVERGENCE OF OPTIMISTIC POLICY ITERATION FOR STOCHASTIC SHORTEST PATH PROBLEM

YUANLONG CHEN

Abstract. In this paper, we prove some convergence results of a special case of optimistic policy iteration algorithm for stochastic shortest path problem mentioned in [5] . We consider both Monte Carlo and $TD(\lambda)$ methods for the policy evaluation step under the condition that all policies are proper.

1. Introduction

In this paper we consider a Markov decision process(MDP) with a finite state set $S = \{1, 2, \ldots, n\}$. In addition, we use 0 to denote the cost-free termination state. For each state i, we assume there are only finite actions, denoted as $U(i)$. Furthermore, for each state $i \in S$ and each action $u \in U(i)$, we associate a transition probability $p_{i,j}(u)$ and an immediate cost function $g(i, u)$. A policy μ is defined as a mapping from S to U(note there are only finitely many policies since states and actions are both finite). Let's denote by X_t^μ the state at time step t under the policy μ. $\{X_t^\mu\}$ then forms a Markov chain with transition probability

$$P(X_{t+1}^\mu = j | X_t^\mu = i) = p_{i,j}(\mu(i)).$$

The total expected cost(cost-to-go) of the process starting from state i under policy μ is

$$J^\mu(i) = E\left[\sum_{t=0}^{\infty} \alpha^t g(X_t^\mu, \mu(X_t^\mu)) \big| X_0^\mu = i\right],$$

where $0 < \alpha \leq 1$ is the discouted factor. A policy μ is said to be proper if, under this policy, there is positive probability that the termination state will be reached after at most n steps, regardless of the initial state, that is, if

$$\rho_\mu = \max_{i \in S} P(X_n^\mu \neq 0 | X_0^\mu = i) < 1.$$

Proper policy basically implies that the termination state will eventually reached almost surely. To see this, note that

$$P(X_t^\mu \neq 0 | X_0^\mu = i) \leq \rho_\mu^{\lfloor t/n \rfloor}, \quad \forall i \in S.$$

The conclusion then follows from Borel-Cantelli lemma. Moreover, J_μ is finte when μ is proper, since

$$|J_\mu(i)| \leq \lim_{T \to \infty} \sum_{t=0}^{T-1} \rho_\mu^{\lfloor t/n \rfloor} \max_j |g(j, \mu(j))| < \infty, \quad \forall i \in S.$$

In this paper, we only consider the stochastic shortest path problem, in which case $\alpha = 1$. In addition, we make the following assumption

Key words and phrases. Optimistic Policy Iteration, Convergence, Stochastic Shortest Path, Dynamic Programming, Reinforcement Learning.

Assumption 1.1. *Every policy in our problem is proper.*

We denote the optimal cost-to-go function starting from i as $J^*(i)$, that is the minimal value of cost-to-go functions among all of the policies,

$$J^*(i) = \min_{\mu} J^{\mu}(i).$$

Note the minimal value can be achieved since there are only finite policies. We then define the optimal cost-to-go vector as $J^* = (J^*(1), \ldots, J^*(n))$. A policy μ is said to be optimal if $J^{\mu}(i) = J^*(i)$ for every $i \in S$.

We next introduce two dynamic programming operators. For any n dimensional vector $J = (J(1), \ldots, J(n))$, define operator $T : \mathbb{R}^n \to \mathbb{R}^n$ as

$$(TJ)(i) = \min_{u} \left\{ g(i,u) + \sum_{j=1}^{n} p_{i,j}(u) J(j) \right\}, \qquad \forall i \in S.$$

Similarly, define $T_\mu : \mathbb{R}^n \to \mathbb{R}^n$ as

$$(T_\mu J)(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{i,j}(\mu(i)) J(j), \qquad \forall i \in S.$$

In vector notation, they are equivalent to

$$(TJ)(i) = \min_{\mu} (T_\mu J)(i), \quad \forall i \in S,$$

and

$$T_\mu J = g_\mu + P_\mu J,$$

where

$$(P_\mu J)(i) = \sum_{j=1}^{n} p_{i,j}(\mu(i)) J(j), \qquad \forall i \in S.$$

These two operators associated with stochastic shortest path problem have some well-known properties, for which we summarize as the following proposition(for the proof one can refer to [1], [2], [3]).

Proposition 1.2. *Under Assumption 1.1, the following properties hold for the stochastic shortest path problem:*

(a) *The optimal cost-to-go vector J^* has finite components and it satisfies*

$$J^* = TJ^*.$$

Furthermore, J^* is the only solution for the equation above.

(b) *For every vector J, we have*

$$\lim_{k \to \infty} T^k J = J^*.$$

(c) *A policy μ is optimal if and only if*

$$T_\mu J^* = TJ^*.$$

(d) *For every proper policy μ, the associated cost-to-go vector J^μ satisfies*

$$\lim_{k \to \infty} T_\mu^k J = J^\mu,$$

for every vector J. Furthermore,

$$J^\mu = T_\mu J^\mu,$$

and J^μ is the only solution for the equation above.

Throughout this paper, we use $\|\cdot\|$ to denote the maximum norm of an n dimensional vector J, defined by

$$\|J\| = \max_i |J(i)|.$$

For a given n dimensional vector $\xi = (\xi(1), \ldots, \xi(n))$ with all components positive, we use $\|\cdot\|_\xi$ to denote the weighted maximum norm with respect to ξ, defined by

$$\|J\|_\xi = \max_i \frac{|J(i)|}{\xi(i)}.$$

For two vectors J and $\bar{J}$, we say $J \leq \bar{J}$, if $J(i) \leq \bar{J}(i)$ for all $i \in S$. $J < \bar{J}$ has the meaning in the same manner.

We also notice the following useful monotonicity properties of T and T_μ(see Lemma 2.1 in [4]):

Proposition 1.3. *For all n dimensional vector J and $\bar{J}$, such that*

$$J \leq \bar{J},$$

for any policy μ and any positve integer k, we have

$$T^k J \leq T^k \bar{J}, \quad T_\mu^k J \leq T_\mu^k \bar{J}.$$

Let's denote by e the n dimentional vector with all components equal to 1, the following result is a direct consequence of an induction argument and Proposition 1.3:

Lemma 1.4. *For every positive scalar c and vector J, we have*

$$T^k(J + ce) \leq T^k J + ce, \qquad \forall k > 0,$$

$$T_\mu^k(J + ce) \leq T_\mu^k J + ce, \qquad \forall k > 0.$$

For T_μ, we also have the following lemma

Lemma 1.5. *Given a scalar squence $\{\lambda_l\}_{l=0}^\infty$ such that $0 < \lambda_l < 1$ and $\sum_l \lambda_l = 1$, for any bounded vector sequence $\{J_l\}_{l=0}^\infty$, we have*

$$T_\mu\left(\sum_{l=0}^\infty \lambda_l J_l\right) = \sum_{l=0}^\infty \lambda_l T_\mu J_l.$$

Proof. First note that for any positive integer L, we have

$$\begin{aligned} T_\mu\left(\sum_{l=0}^\infty \lambda_l J_l\right) &= T_\mu\left(\sum_{0\leq l\leq L} \lambda_l J_l + \sum_{l>L} \lambda_l J_l\right) \\ &= \sum_{0\leq l\leq L} \lambda_l g_\mu + \sum_{0\leq l\leq L} \lambda_l P_\mu J_l + \sum_{l>L} \lambda_l g_\mu + P_\mu\left(\sum_{l>L} \lambda_l J_l\right) \\ &= \sum_{0\leq l\leq L} \lambda_l T_\mu J_l + \sum_{l>L} \lambda_l g_\mu + P_\mu\left(\sum_{l>L} \lambda_l J_l\right). \end{aligned}$$

It's easy to see that

$$\lim_{L\to\infty} \sum_{l>L} \lambda_l = 0.$$

Note J_l is bounded and $0 < \lambda_l < 1$, therefore

$$\left| \lim_{L\to\infty} \sum_{l>L} \lambda_l J_l \right| \leq \left(\lim_{L\to\infty} \sum_{l>L} \lambda_l \right) \cdot \max_l |J_l| = 0.$$

Since g_u and P_μ are both bounded, the conclusion then follows. □

We now give a brief description of policy iteration algorithm. In the ordinary policy iteration procedure,we start with some initial policy μ, and then we do the policy evaluation, i.e. evaluate the optimal cost-to-go vector J^μ corresponding to μ. In this step, for example, onec can use learning algorithms such as Monte Carlo or $TD(\lambda)$. Once we have the cost-to-go vector J^μ, we perform policy improvement step, which updates μ as

$$\mu(i) \leftarrow \arg\min_{u\in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{i,j}(u) J^\mu(j) \right\}, \qquad \forall i \in S.$$

Such process is repeated until the algorithm converges.

One disadvantage of the algorithm described above is that, in practice, the accurate evaluation of the cost-to-go vector J^μ could be very expensive, which makes the algorithm inefficient. Optimistic policy iteration is a variation of the ordinary policy iteration to address this issue in which the policy improvement is based on an incomplete evaluation of J^μ instead of an accurate J^μ. For example, if we apply Monte Carlo method in policy evaluation step, in the ordinary policy iteration algorithm, theoretically, a large number of trajectories need to be simulated to guarantee an accurate estimation. In contrast, for optimistic policy iteration, we perform policy improvement immediately after one single trajectory sample. In [5], the convergence results have been established for discounted problems($0 < \alpha < 1$) based on both Monte Carlo and $TD(\lambda)$ methods. In the following sections, we will show that the similar convergence results can be extended to undiscounted stochastic shortest path problem($\alpha = 1$).

2. Monte Carlo based optimistic synchronous policy iteration

We first provide a precise description of the optimitic policy iteration algorithm. We start with some random vector J_0 and policy μ_0. The iteration proceeds as follows: at each time step t, for each state i, we simulate a single trajectory which starts with i under the policy μ_t(note that the termination is guaranteed since the policy is proper). The observed cumulative cost is an unbiased estimate of $J^{\mu_t}(i)$, for which we denote by $J^{\mu_t}(i) + \omega_t(i)$, where $\omega_t(i)$ is a zero-mean noise. We then update vector J_t according to the following update rule

$$J_{t+1}(i) = (1-\gamma_t) J_t(i) + \gamma_t (J^{\mu_t}(i) + \omega_t(i)), \tag{2.1}$$

where γ_t is a deterministic scalar stepsize parameter. Furthermore, we impose the well-known step-size conditions for γ_t

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \qquad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

Let $\mathcal{F}_t$ be the history of the algorithm up to and including the point where J_t has been produced, but before simulating the trajectories for the next update, based on

the argument in [5], we know that

$$E\left[\omega_t(i)|\mathcal{F}_t\right] = 0,$$

and

$$E\left[|\omega_t(i)|^2|\mathcal{F}_t\right] \leq C,$$

for some positive constant C.

We summarize our main result as the following theorem:

Theorem 2.1. *The sequence J_t generated by the optimistic policy iteration algorithm according to (2.1) for the stochastic shortest path problem, converges to the optimal cost-to-go vector J^*, almost surely.*

Before proving Theorem 2.1, we first establish several lemmas.

Lemma 2.2. *For any given $\epsilon > 0$ and $M > 0$, there exists a positive integer $K = K(\epsilon, M)$ such that for all policy μ and vector J such that $\|J\| \leq M$, we have*

$$\|T_\mu^k J - J^\mu\| < \epsilon, \quad \forall k \geq K.$$

Proof. It suffices to prove the result for just one particular μ since there are only finite policies. For any given n dimensional vector J, by part (d) of Proposition 1.2, we have

$$\lim_{k\to\infty} T_\mu^k J = J^\mu.$$

It follows that, for any given $\epsilon > 0$, there exists a $K(J) > 0$, such that

$$\|T_\mu^k J - J^\mu\| < \epsilon/2, \quad \forall k \geq K(J). \tag{2.2}$$

Note that we have the following estimates

$$\|T_\mu J - T_\mu \bar{J}\| \leq \|J - \bar{J}\|.$$

An easy inductive argument indicates that

$$\|T_\mu^k J - T_\mu^k \bar{J}\| \leq \|J - \bar{J}\|, \quad \forall k \geq 1.$$

Thus, for this ϵ, we have $\|T_\mu^k J - T_\mu^k \bar{J}\| < \epsilon/2$ for all $k \geq 1$ and $\bar{J}$, as long as $\|\bar{J} - J\| < \epsilon/2$.

Define $B_\epsilon(J) = \{\bar{J} | \|\bar{J} - J\| < \epsilon/2\}$, then

$$\|T_\mu^k \bar{J} - J^\mu\| < \epsilon, \quad \forall \bar{J} \in B_\epsilon(J), \quad \forall k \geq K(J).$$

Let $R = \{J | \|J\| \leq M\}$. R is a compact set and $\{B_\epsilon(J)\}_{J\in R}$ form an open cover of R. By Heine-Borel theorem, there exists a finite subcover, say $B_\epsilon(J_1), \cdots, B_\epsilon(J_l)$. Set

$$K = \max_{i\in\{1,\cdots,l\}} K(J_i) < \infty,$$

the conclusion then follows. □

Lemma 2.3. *The sequence J_t generated by (2.1) is bounded almost surely.*

Proof. Since there are only finitely many possible policies, J^{μ_t} is bounded for any t . Note that the update rule is

$$J_{t+1} = (1 - \gamma_t)J_t + \gamma_t J^{\mu_t} + \gamma_t \omega_t.$$

The boundedness of sequence J_t is then a direct consequence of Proposition 4.7 on p. 159 in [4] □

Define a scalar sequence c_t by setting

$$c_t = \max_i \big((TJ_t)(i) - J_t(i)\big), \tag{2.3}$$

we have

Lemma 2.4. *For sequence c_t, the following estimate holds*

$$\limsup_{t\to\infty} c_t \leq 0.$$

Proof. The proof is essentially identical as in [5] with just a few minor modifications. Recall the vector form definition of T_{μ_t}

$$T_{\mu_t} J = g_{\mu_t} + P_{\mu_t} J, \quad \forall J.$$

By the same calculation in [5], we have

$$TJ_{t+1} - J_{t+1} \leq (1-\gamma_t)(TJ_t - J_t) + \gamma_t v_t,$$

where $v_t = P_{\mu_t}\omega_t - \omega_t$. Note that for this v_t, we still have the following properties:

$$E\left[v_t(i)|\mathcal{F}_t\right] = 0, \quad \forall i \in S,$$

and

$$E\left[v_t(i)^2|\mathcal{F}_t\right] \leq C, \quad \forall i \in S,$$

for some constant C. We can then use the identical argument in [5] to prove the result. □

Lemma 2.5. *For all $\epsilon > 0$, there exists a constant $t(\epsilon) > 0$ such that for all $t \geq t(\epsilon)$, the following estimates are true*

$$J^{\mu_t} \leq TJ_t + \epsilon e. \tag{2.4}$$

Proof. The defintion of μ_t tells us $T_{\mu_t} J_t = TJ_t$, it follows that

$$T_{\mu_t} J_t = J_t + \big(TJ_t - J_t\big) \leq J_t + c_t e. \tag{2.5}$$

Apply T_{μ_t} to both sides of inequality (2.5) $k-1$ times, an easy inductive argument together with Lemma 1.4 show that

$$T^k_{\mu_t} J_t \leq J_t + k c_t e. \tag{2.6}$$

By Lemma 2.3, there exists a constant M such that $\left|J_t\right| \leq M$ for all t almost surely. According to Lemma 2.2, for all $\epsilon > 0$, there exists $K = K(\epsilon, M)$, such that for all J_t, the following estimates are valid,

$$\|T^K_{\mu_t} J_t - J^{\mu_t}\| < \epsilon/2. \tag{2.7}$$

We now fix K. By Lemma 2.4, for this fixed ϵ, there exists $t(\epsilon) > 0$, such that for all $t \geq t(\epsilon)$

$$Kc_t \leq \frac{\epsilon}{2} e,$$

it then follows from (2.6) that

$$T^K_{\mu_t} J_t \leq J_t + \frac{\epsilon}{2} e. \tag{2.8}$$

Combine (2.7) and (2.8), we have

$$J^{\mu_t} = J^{\mu_t} - T^K_{\mu_t} J_t + T^K_{\mu_t} J_t \leq \frac{\epsilon}{2} e + J_t + \frac{\epsilon}{2} e = J_t + \epsilon e.$$

Apply T_{μ_t} on both sides of the inequality above, using Lemma 1.4 and the fact that $T_{\mu_t} J^{\mu_t} = J^{\mu_t}$, we see that for all $t \geq t(\epsilon)$

$$J^{\mu_t} \leq T_{\mu_t} J_t + \epsilon e = T J_t + \epsilon e.$$

□

Proof of Theorem 2.1. Having established (2.4), the rest of the proof is essentially the same as the argument in Proposition 1 in [5]. First we note that for all $t \geq t(\epsilon)$

$$\begin{aligned} J_{t+1} &= (1-\gamma_t) J_t + \gamma_t J^{\mu_t} + \gamma_t \omega_t \\ &\leq (1-\gamma_t) J_t + \gamma_t T J_t + \gamma_t \epsilon e + \gamma_t \omega_t. \end{aligned}$$

For this fixed ϵ, we define a sequence Z_t that starts from time $t(\epsilon)$ by setting $Z_{t(\epsilon)} = J_{t(\epsilon)}$ and

$$Z_{t+1} = (1-\gamma_t) Z_t + \gamma_t T Z_t + \gamma_t \epsilon e + \gamma_t \omega_t, \qquad \forall t \geq t(\epsilon).$$

An easy inductive argument shows that $J_t \leq Z_t$ for all $t \geq t(\epsilon)$. Using the identical argument as in the proof of Proposition 1 in [5], we can derive

$$\limsup_{t \to \infty} J_t \leq J^*,$$

and

$$\liminf_{t \to \infty} J_t \geq J^*.$$

Thus, we have

$$\lim_{t \to \infty} J_t = J^*.$$

□

3. $TD(\lambda)$ based optimistic synchronous policy iteration

In this section, we extend the results in previous section to $TD(\lambda)$ based algorithm. The framework of $TD(\lambda)$ algorithm is essentially the same as Monte Carlo scenario except that, in each policy evaluation step, $TD(\lambda)$ based algorithm uses temporal difference method instead of Monte Carlo. Precisely, at iteration t, we have a vector J_t and the corresponding greedy policy μ_t, for each state i, we simulate a trajectory $i_0, i_1, \ldots$ that starts with i, then update $J_t(i)$ to $J_{t+1}(i)$ according to

$$J_{t+1}(i) = J_t(i) + \gamma_t \sum_{k=0}^{\infty} \lambda^k d_k, \qquad \lambda \in [0,1),$$

where d_k is called temporal difference defined as $d_k = g(i_k, \mu_t(i_k)) + J_t(i_{k+1}) - J_t(i_k)$ and γ_t is a scalar stepsize parameter. This is equivalent to

$$J_{t+1}(i) = (1-\gamma_t) J_t(i) + \gamma_t (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \left(g(i_0) + g(i_1) + \cdots + g(i_k) + J_t(i_{k+1}) \right).$$

In vector notation, we have

$$J_{t+1} = (1-\gamma_t) J_t + \gamma_t (1-\lambda) \sum_{k=0}^{\infty} \lambda^k T_{\mu_t}^{k+1} J_t + \gamma_t \omega_t, \tag{3.1}$$

where ω_t is a noise vector with zero mean reflecting the difference between the observed temporal differences and their expected values.

Before heading to the main result, let us first take a look at two extreme cases $\lambda = 1$ and $\lambda = 0$ to get some intuitions of the $TD(\lambda)$ based algorithm. If $\lambda = 1$, the update rule (3.1) becomes

$$J_{t+1} = (1 - \gamma_t)J_t + \gamma_t \sum_{k=0}^{\infty} g(i_k),$$

and this is just the Monte Carlo based method. On the other end, if $\lambda = 0$, the update rule (3.1) becomes

$$J_{t+1} = (1 - \gamma_t)J_t + \gamma_t T J_t + \gamma_t \omega_t,$$

where we use the fact that $T_{\mu_t} J_t = T J_t$. It is well known that T is a weighted maximum-norm pseudo-contraction(see Proposition 2.2 on p. 23 in [4]). General stochastic iterative algorithm result(see Proposition 4.4 on p. 156 in [4]) shows that the J_t converges to J^*.

For $0 < \lambda < 1$, the method can be regarded as a weighted combination of $TD(0)$ and Monte Carlo. In the rest of this section, we show that in this scenario the algorithm converges to J^* almost surely as well. We summarize our main result as follows:

Theorem 3.1. *The sequence J_t generated by the optimistic synchronous policy iteration algorithm according to update rule (3.1) for the stochastic shortest path problem, converges to the optimal cost-to-go vector J^*, almost surely.*

We first estabish several lemmas:

Lemma 3.2. *The sequence J_t generated by (3.1) is bounded almost surely.*

Proof. We first show that for all policy μ, there exist a scalar $\delta_\mu \in [0, 1)$, $G_\mu > 0$ and $K_\mu > 0$, the following estimates hold

$$\|T_\mu^{k+1} J\| \leq \delta_\mu \|J\| + G_\mu, \quad \forall k > K_\mu, \quad \forall J. \tag{3.2}$$

To prove this, note that T_μ is a contraction mapping with respect to some vector ξ_μ with all components positive, i.e. there exists $\beta_\mu \in [0, 1)$ such that

$$\|T_\mu J - T_\mu \bar{J}\|_{\xi_\mu} \leq \beta_\mu \|J - \bar{J}\|_{\xi_\mu},$$

for all vectors J and $\bar{J}$(see Proposition 2.2 on p. 23 in [4]). Thus

$$\begin{aligned} \|T_\mu J\|_{\xi_\mu} &\leq \|T_\mu J - J^\mu\|_{\xi_\mu} + \|J^\mu\|_{\xi_\mu} \\ &\leq \beta_\mu \|J - J^\mu\|_{\xi_\mu} + \|J^\mu\|_{\xi_\mu} \\ &\leq \beta_\mu \|J\|_{\xi_\mu} + D_\mu, \end{aligned} \tag{3.3}$$

where $D_\mu = (1 + \beta_\mu)\|J^\mu\|_{\xi_\mu} < \infty$. Inductively, we have

$$\|T_\mu^{k+1} J\|_{\xi_\mu} \leq \beta_\mu^{k+1} \|J\|_{\xi_\mu} + (1 + \beta_\mu + \cdots + \beta_\mu^k) D_\mu, \quad \forall k \geq 0.$$

This implies

$$\|T_\mu^{k+1} J\|_{\xi_\mu} \leq \beta_\mu^{k+1} \|J\|_{\xi_\mu} + \tilde{D}_\mu, \quad \forall k \geq 0,$$

where $\tilde{D}_\mu = (\sum_{k=0}^{\infty} \beta_\mu^k) D_\mu < \infty$.

Let us denote by $\xi_{\mu,min} = \min_i \xi_\mu(i)$, $\xi_{\mu,max} = \max_i \xi_\mu(i)$ and set $\rho_\mu = \xi_{\mu,min}/\xi_{\mu,max}$. Note that $\rho_\mu > 0$ and $\beta_\mu \in [0, 1)$, thus there exists $K_\mu > 0$ such that $\beta_\mu^{K_\mu+1} < \rho_\mu$. We

then have, for all $k > K_\mu$

$$\begin{aligned}
\|T_\mu^{k+1} J\| &= \max_i |T_\mu^{k+1} J(i)| \\
&= \xi_{\mu,max} \max_i \left\{ \frac{|T_\mu^{k+1} J(i)|}{\xi_{\mu,max}} \right\} \\
&\leq \xi_{\mu,max} \|T_\mu^{k+1} J\|_{\xi_\mu} \\
&\leq \xi_{\mu,max} \left(\beta_\mu^{k+1} \|J\|_{\xi_\mu} + \tilde{D}_\mu \right) \\
&\leq \frac{\xi_{\mu,max}}{\xi_{\mu,min}} \beta_\mu^{k+1} \|J\| + \xi_{\mu,max} \tilde{D}_\mu \\
&\leq \frac{\beta_\mu^{K_\mu+1}}{\rho_\mu} \|J\| + \xi_{\mu,max} \tilde{D}_\mu \\
&= \delta_\mu \|J\| + G_\mu,
\end{aligned}$$

where $\delta_\mu = \beta_\mu^{K_\mu+1}/\rho_\mu < 1$ and $G_\mu = \xi_{\mu,max}\tilde{D}_\mu < \infty$. Set

$$\begin{aligned}
\delta &= \max_\mu \delta_\mu \in [0, 1), \\
G &= \max_\mu G_\mu < \infty, \\
K &= \max_\mu K_\mu < \infty,
\end{aligned}$$

we then have

$$\|T_\mu^{k+1} J\| \leq \delta \|J\| + G, \quad \forall k > K, \quad \forall J, \quad \forall \mu. \tag{3.4}$$

On the other hand, it's easy to see that there exists a bounded scalar sequence $\{G_k\}_{k=0}^K$, such that

$$\|T_\mu^{k+1} J\| \leq \|J\| + G_k, \quad \forall k \leq K, \quad \forall J, \quad \forall \mu. \tag{3.5}$$

Write

$$J_{t+1} = (1 - \gamma_t) J_t + \gamma_t H_t J_t + \gamma_t \omega_t,$$

where

$$H_t J_t = (1 - \lambda) \sum_{k=0}^{\infty} \lambda^k T_{\mu_t}^{k+1} J_t.$$

Given (3.4) and (3.5), the mapping H_t then satisfies the following estimates

$$\begin{aligned}
\|H_t J_t\| &\leq (1 - \lambda) \sum_{0 \leq k \leq K} \lambda^k \|T_{\mu_t}^{k+1} J_t\| + (1 - \lambda) \sum_{k > K} \lambda^k \|T_{\mu_t}^{k+1} J_t\| \\
&\leq (1 - \lambda) \sum_{0 \leq k \leq K} \lambda^k \left(\|J_t\| + G_k \right) + (1 - \lambda) \sum_{k > K} \lambda^k \left(\delta \|J_t\| + G \right) \\
&= \phi_\lambda \|J_t\| + G_\lambda
\end{aligned}$$

where

$$\phi_\lambda = (1 - \lambda) \sum_{0 \leq k \leq K} \lambda^k + (1 - \lambda) \sum_{k > K} \lambda^k \delta < 1,$$

and

$$G_\lambda = (1 - \lambda) \sum_{0 \leq k \leq K} \lambda^k G_k + (1 - \lambda) \sum_{k > K} \lambda^k G < \infty.$$

The boundedness of the sequence J_t then follows from Proposition 4.7 on p. 159 in [4]. □

Lemma 3.3. *For sequence c_t defined in (2.3), we have*

$$\limsup_{t\to\infty} c_t \le 0.$$

Proof. Recall that

$$T_{\mu_t} J = g_{\mu_t} + P_{\mu_t} J, \quad \forall J.$$

Using affine properties of T_{μ_t}, we have

$$\begin{aligned}
TJ_{t+1} &\le T_{\mu_t} J_{t+1} \\
&= T_{\mu_t}\left((1-\gamma_t)J_t + \gamma_t(1-\lambda)\sum_{k=0}^{\infty}\lambda^k T_{\mu_t}^{k+1}J_t + \gamma_t\omega_t\right) \\
&= g_{\mu_t} + (1-\gamma_t)P_{\mu_t}J_t + \gamma_t P_{\mu_t}(1-\lambda)\sum_{k=0}^{\infty}\lambda^k T_{\mu_t}^{k+1}J_t + \gamma_t P_{\mu_t}\omega_t \\
&= (1-\gamma_t)T_{\mu_t}J_t + \gamma_t T_{\mu_t}\left(\sum_{t=0}^{\infty}(1-\lambda)\lambda^k T_{\mu_t}^{k+1}J_t\right) + \gamma_t P_{\mu_t}\omega_t \\
&= (1-\gamma_t)\left(T_{\mu_t}J_t - J_t\right) + \left[(1-\gamma_t)J_t + \gamma_t(1-\lambda)\sum_{k=0}^{\infty}\lambda^k T_{\mu_t}^{k+1}J_t + \gamma_t\omega_t\right] \\
&\quad + \gamma_t\left[T_{\mu_t}\left(\sum_{k=0}^{\infty}(1-\lambda)\lambda^k T_{\mu_t}^{k+1}J_t\right) - \sum_{k=0}^{\infty}(1-\lambda)\lambda^k T_{\mu_t}^{k+1}J_t\right] + \gamma_t\left[P_{\mu_t}\omega_t - \omega_t\right] \\
&= (1-\gamma_t)\left(T_{\mu_t}J_t - J_t\right) + J_{t+1} + \gamma_t H_t J_t + \gamma_t v_t,
\end{aligned}$$

where

$$H_t J_t = T_{\mu_t}\left((1-\lambda)\sum_{k=0}^{\infty}\lambda^k T_{\mu_t}^{k+1}J_t\right) - \sum_{k=0}^{\infty}(1-\lambda)\lambda^k T_{\mu_t}^{k+1}J_t,$$

and

$$v_t = P_{\mu_t}\omega_t - \omega_t.$$

Equivalently, we have

$$TJ_{t+1} - J_{t+1} \le (1-\gamma_t)\left(TJ_t - J_t\right) + \gamma_t H_t J_t + \gamma_t v_t. \tag{3.6}$$

In the rest of the proof , we show that, for any $\epsilon > 0$, H_t essentially is a maximum norm contraction with a unique fixed pint ϵe. Stochastic iterative algorithm then can be applied to (3.6).

We now fix an arbitrary $\epsilon > 0$. We note that $T_{\mu_t} J^{\mu_t} = J^{\mu_t}$. Since T_{μ_t} is a continous operator and we have only finitely many policies, we see that for this fixed ϵ, there exists $\delta(\epsilon) > 0$, such that for all μ_t, and all vector J, as long as $\|J - J^{\mu_t}\| < \delta(\epsilon)$, we have

$$\|T_{\mu_t} J - J\| < \epsilon.$$

Now fix $\delta(\epsilon)$, since $\{J_t\}$ is bounded almost surely, by Lemma 2.2, there exists a positive integer $K(\epsilon)$, such that for all $k > K(\epsilon)$ and all μ_t, the following estimates hold

$$\|T_{\mu_t}^{k+1} J_t - J^{\mu_t}\| < \delta(\epsilon). \tag{3.7}$$

Now we split H_tJ_t to two parts according to $K(\epsilon)$ as

$$\begin{aligned} H_tJ_t &= (1-\lambda)\left[\sum_{k=0}^{\infty}\lambda^k T_{\mu_t}\left(T_{\mu_t}^{k+1}J_t\right) - \sum_{k=0}^{\infty}\lambda^k T_{\mu_t}^{k+1}J_t\right] \\ &= H_{t,1}J_t + H_{t,2}J_t, \end{aligned} \tag{3.8}$$

where in the first equality we apply Lemma 1.5, and

$$H_{t,1}J_t = (1-\lambda)\sum_{0\le k\le K(\epsilon)}\lambda^k\left(T_{\mu_t}(T_{\mu_t}^{k+1}J_t) - T_{\mu_t}^{k+1}J_t\right),$$

and

$$H_{t,2}J_t = (1-\lambda)\sum_{k>K(\epsilon)}\lambda^k\left(T_{\mu_t}(T_{\mu_t}^{k+1}J_t) - T_{\mu_t}^{k+1}J_t\right).$$

Now we establish estimates for $H_{t,1}$ and $H_{t,2}$ seperately.

(a) Estimiate for $H_{t,1}J_t$ term: since

$$T_{\mu_t}\left(T_{\mu_t}^{k+1}J_t\right) \le T_{\mu_t}^{k+1}(J_t + c_te) \le T_{\mu_t}^{k+1}J_t + c_te.$$

we have

$$H_{t,1}J_t \le (1-\lambda)\sum_{0\le k\le K(\epsilon)}\lambda^k c_te = \varphi_1 c_te, \tag{3.9}$$

where $\varphi_1 = (1-\lambda)\sum_{0\le k\le K(\epsilon)}\lambda^k$.

(b) Estimate for $H_{t,2}J_t$ term: since $k > K(\epsilon)$, (3.7) holds. By the choice of $\delta(\epsilon)$, we then have

$$\|T_{\mu_t}(T_{\mu_t}^{k+1}J_t) - T_{\mu_t}^{k+1}J_t\| < \epsilon,$$

this implies

$$H_{2,t}J_t \le (1-\lambda)\sum_{k>K(\epsilon)}\lambda^k\epsilon e = \varphi_2\epsilon e, \tag{3.10}$$

where $\varphi_2 = (1-\lambda)\sum_{k>K(\epsilon)}\lambda^k$.

Combine (3.8), (3.9) and (3.10), we have

$$H_tJ_t \le \varphi_1 c_te + \varphi_2\epsilon e. \tag{3.11}$$

Together with (3.6), we obtain

$$TJ_{t+1} - J_{t+1} \le (1-\gamma_t)(TJ_t - J_t) + \gamma_t(\varphi_1 c_te + \varphi_2\epsilon e) + \gamma_t v_t. \tag{3.12}$$

Set $X_t = TJ_t - J_t$, by the definition of c_t, we see that

$$X_{t+1} \le (1-\gamma_t)X_t + \gamma_t(\varphi_1 e\max_i X_t(i) + \varphi_2\epsilon e) + \gamma_t v_t.$$

We use the comparison argument as in previous section. Define a sequence of vector Y_t by setting $Y_0 = X_0$ and

$$Y_{t+1} = (1-\gamma_t)Y_t + \gamma_t(\varphi_1 e\max_i Y_t(i) + \varphi_2\epsilon e) + \gamma_t v_t.$$

An easy inductive argument shows that $X_t \le Y_t$ for all t. Note that $\varphi_1, \varphi_2 \in (0,1)$ and $\varphi_1 + \varphi_2 = 1$, it then follows that $Y \mapsto \varphi_1 e\max_i Y_t(i) + \varphi_2\epsilon e$ is a maximum norm contraction. It's a well-known fact that there exists only one fixed point for such

mapping. A straightforward calculation(using $\varphi_1 + \varphi_2 = 1$) shows that ϵe is the fixed point for this mapping.

The rest of the proof is essentially identical to the argument in [5]. Fix a positive integer l, we define the stopped process $v^l(t)$ such that it coincides with v_t as long as $E\left[|v_t|^2|\mathcal{F}_t\right] \leq l$, and is equal to 0 thereafter. Consider the iteration

$$Y_{t+1}^l = (1-\gamma_t)Y_t^l + \gamma_t(\varphi_1 e \max_i Y_t^l(i) + \varphi_2 \epsilon e) + \gamma_t v_t^l.$$

By Proposition 4.4 on p. 156 in [4], Y_t^l converges to ϵe, for every l. Since J_t is bounded, we see that $E\left[|v_t|^2|\mathcal{F}_t\right]$ is also bounded. Therefore, there exists some l such that $v_t^l = v_t$ almost surely. As a result, $Y_t^l = Y_t$ for all t. Hence Y_t also converges to ϵe, which implies that

$$\limsup_{t\to\infty} X_t \leq \epsilon e.$$

ϵ could be arbitrarily small, we can then conclude that

$$\limsup_{t\to\infty} c_t \leq 0.$$

□

Note that (3.1) only replaces J^{μ_t} by $(1-\lambda)\sum_{k=0}^{\infty}\lambda^k T_{\mu_t}^{k+1}J_t$ compared to (2.1). This indicates that we can establish the following result corresponding to Lemma 2.5,

Lemma 3.4. *For all $\epsilon > 0$, there exists $t(\epsilon) > 0$ such that for all $t \geq t(\epsilon)$, we have*

$$(1-\lambda)\sum_{k=0}^{\infty}\lambda^k T_{\mu_t}^{k+1}J_t \leq TJ_t + \epsilon e. \tag{3.13}$$

Proof. First we notice that the conclusion in Lemma 2.5 still holds for J_t generated by update rule (3.1). If we examine the proof closely there, in order to prove Lemma 2.5 for J_t, we only need the boundedness of J_t and $\limsup_{t\to\infty} c_t \leq 0$, and it does not depend on how we update J_t. Both of those facts are true for J_t generated here.

By Lemma 2.5, we know for any fixed $\epsilon > 0$, there exists a time $t_1(\epsilon)$, for all $t > t_1(\epsilon)$, we have

$$J^{\mu_t} \leq TJ_t + \frac{\epsilon}{2}e.$$

For this fixed $\epsilon > 0$, since J_t is bounded, by Lemma 2.2, there exists a positive $K(\epsilon)$, for all $k > K(\epsilon)$ and policy μ_t, we have

$$T_{\mu_t}^{k+1}J_t \leq J^{\mu_t} + \frac{\epsilon}{2}.$$

Combine the two inequalities above, we have

$$T_{\mu_t}^{k+1}J_t \leq TJ_t + \epsilon, \qquad \forall t > t_1(\epsilon), \quad \forall k > K(\epsilon). \tag{3.14}$$

Note that

$$T_{\mu_t}^{k+1}J_t \leq T_{\mu_t}J_t + kc_t e = TJ_t + kc_t e,$$

also note $\limsup_{t\to\infty} c_t \leq 0$, we see that for this fixed $K(\epsilon)$, there exists $t > t_2(\epsilon)$, such that $K(\epsilon)c_t < \epsilon$. This implies

$$T_{\mu_t}^{k+1}J_t \leq TJ_t + \epsilon, \qquad \forall t > t_2(\epsilon), \quad \forall k \leq K(\epsilon). \tag{3.15}$$

Set $t(\epsilon) = \max\{t_1(\epsilon), t_2(\epsilon)\}$, (3.14) and (3.15) then imply

$$T_{\mu_t}^{k+1}J_t \leq TJ_t + \epsilon, \qquad \forall t > t(\epsilon), \quad \forall k. \tag{3.16}$$

Given (3.16), for all $t > t(\epsilon)$, we have

$$(1-\lambda)\sum_{k=0}^{\infty}\lambda^k T_{\mu_t}^{k+1}J_t \leq \left((1-\lambda)\sum_{k=0}^{\infty}\lambda^k\right)(TJ_t+\epsilon) = TJ_t+\epsilon.$$

□

Having established all these preliminary results, let us prove our main result

Proof of Theorem 3.1. The proof is es sentially the same as the proof of Theorem 2.1. Fix some $\epsilon > 0$, by Lemma 3.4, there exists $t(\epsilon)$ such that estimates (3.13) hold. We then have

$$J_{t+1} \leq (1-\gamma_t)J_t + \gamma_t(TJ_t+\epsilon) + \gamma_t\omega_t, \quad \forall t \geq t(\epsilon).$$

Use the same argument in Theorem 2.1, we can obtain

$$\limsup_{t\to\infty} J_t \leq J^*.$$

To complete the proof, we now only need to show

$$\liminf_{t\to\infty} J_t \geq J^*. \tag{3.17}$$

To see this, note $T_{\mu_t}^k J_t \geq T^k J_t$ for all $k > 0$, we then have

$$\begin{aligned} J_{t+1} &= (1-\gamma_t)J_t + \gamma_t(1-\lambda)\sum_{k=0}^{\infty}\lambda^k T_{\mu_t}^{k+1}J_t + \gamma_t\omega_t \\ &\geq (1-\gamma_t)J_t + \gamma_t(1-\lambda)\sum_{k=0}^{\infty}\lambda^k T^{k+1}J_t + \gamma_t\omega_t \\ &= (1-\gamma_t)J_t + \gamma_t\tilde{H}_t J_t + \gamma_t\omega_t. \end{aligned}$$

T is a weighted maximum norm pseudo-contraction, so is $\tilde{H}_t$. Define $\{X_t\}$ by setting $X_0 = J_0$ and

$$X_{t+1} = (1-\gamma_t)X_t + \gamma_t\tilde{H}_t J_t + \gamma_t\omega_t.$$

By induction, it's easy to see that $J_t \geq X_t$. Note H_t is pseudo-contraction mapping with unique fixed point J^*, it follows that

$$\liminf_{t\to\infty} J_t \geq \liminf_{t\to\infty} X_t = J^*.$$

□

4. Conclusion and future work

We solved an open problem mentioned in [5]

References

1. Dimitri P. Bertsekas, *Dynamic programming and optimal control*, 2nd ed., Athena Scientific, 2000.
2. Dimitri P. Bertsekas and John N. Tsitsiklis, *Parallel and distributed computation: Numerical methods*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.
3. ———, *An analysis of stochastic shortest path problems*, Math. Oper. Res. **16** (1991), no. 3, 580–595.
4. ———, *Neuro-dynamic programming.*, Athena Scientific, Belmont, MA, 1996.
5. John N. Tsitsiklis, *On the convergence of optimistic policy iteration*, J. Mach. Learn. Res. **3** (2003), no. 1, 59–72. MR 1966053

Department of Mathematics, University of Washington, Seattle, WA, United States
Email address: ylchen88@uw.edu